

ARE ALL PLASTIDS CREATED EQUAL? DIVERSITY, FUNCTION, AND EVOLUTION
OF PLASTID-TARGETED GENES AND CHLOROPLAST TRANSIT PEPTIDES IN
PLANTS

By

RYAN WAYNE CHRISTIAN

A dissertation submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
Molecular Plant Sciences

JULY 2019

© Copyright by RYAN WAYNE CHRISTIAN, 2019
All Rights Reserved

© Copyright by RYAN WAYNE CHRISTIAN, 2019
All Rights Reserved

To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of
RYAN WAYNE CHRISTIAN find it satisfactory and recommend that it be accepted.

Amit Dhingra, Ph.D., Chair

Helmut Kirchhoff, Ph.D.

H. Henning Kunz, Ph.D.

Eric H. Roalson, Ph.D.

ACKNOWLEDGMENT

First, I would like to thank my advisor, Dr. Amit Dhingra, for the support, instruction, supervision, and encouragement that he has provided over the past six years. He has helped me to learn and grow both in and outside of the laboratory, and his infectious passion for science has been a tremendous influence. I would also like to thank those serving on my committee, Drs. Eric Roalson, Helmut Kirchhoff, and Henning Kunz for guiding me throughout my degree.

I am grateful for all the time and hard work put in by the undergraduate student researchers that I have been fortunate to work with over the past five years including Brennan Hyden, Patrick Old, Erica Conway, Nic Conti, Jonathan Abarca, David Lorse, Justin Foslien, Grant Nelson, Sequoia Leuba, Destin Holland, Jessica Brar, Chelsea Crabb, Paola Coronel, and Emma Smith. I would also like to thank past and present members of the Dhingra lab, especially Scott Schaeffer, Deepika Minhas, Daylen Isaac, Danielle Guzman, Seanna Hewitt, Bruce Williamson-Benavides, Rick Sharpe, Rishikesh Ghogare, Karen Adams, and Nathan Tarlyn.

Many collaborators made this research possible, including Dr. Valeria Lynch-Holm and Dr. Dan Mullendore from the WSU Franceschi Microscopy and Imaging Center for their contributions to the Albino 3 study, Dr. Jonathan Lomber and Michael Apasiku at the WSU Analytical Chemistry Service Center for their expertise in ICP-MS, Magnus Wood at the WSU Phenomics Center for conducting phenotyping experiments, Dr. Jeremy Jewell and Dr. Kiwamu Tanaka for assistance with aequorin fluorescence assays, and Dr. John Vogel at DOE-JGI for providing data and assistance for the BrachyPan resources. My funding was partially provided by the WSU NIH Protein Technology Training Grant which funded me for two years and enabled me to both pursue a much more diverse education as well as an industry internship experience.

Finally, thanks to all my family and friends for their encouraging me to stick with it throughout my graduate school experience. I am extremely fortunate and thankful to have such wonderful people surrounding me and supporting me throughout a somewhat tumultuous six years. I am lucky to have had my best friend completing his Ph.D. beside me: Corey Knadler, thanks for all the late-night science discussions, and for your awesome friendship. I would especially like to acknowledge the members of my family. My father, Jim Christian, literally taught me the fundamentals of biology in high school and has inspired my interest in science throughout my life. My mother, Joanie Christian, shared her passion for gardening with me as a kid and helped to develop my sense of curiosity and wonder at the natural world, setting me on the path to study the beauty and intricacy of how plants function at the molecular level. My amazing brother, Drew Christian, is a complement to my more introverted personality traits and constantly encourages me to explore life outside the lab, learn new interests, and most importantly, to remember to have a little fun. My grandparents, Dale and Lucille Christian, have been so encouraging over the years, and I know how excited they are to finally have a doctor in the family. Finally, I'd like to acknowledge my late grandmother, Virginia "Ginny" Vaughn, who was so supportive of me pursuing a graduate degree and whom I am quite sure is looking down from heaven proudly. In her own words, this monumental achievement in my life is most certainly "Uptown, Charlie Brown."

ARE ALL PLASTIDS CREATED EQUAL? DIVERSITY, FUNCTION, AND EVOLUTION
OF PLASTID-TARGETED GENES AND CHLOROPLAST TRANSIT PEPTIDES IN
PLANTS

Abstract

by Ryan Wayne Christian, Ph.D.
Washington State University
July 2019

Chair: Amit Dhingra

Plastids are morphologically and biochemically diverse organelles which not only perform photosynthesis but are also responsible for a wide range of metabolic, storage, regulatory, and aesthetic functions. The majority of plastid proteins are encoded by the nucleus and imported post-translationally, which allows for much greater spatiotemporal control of plastid function. Most of what is known about plastids has been studied in a relatively small group of model organisms, yet the myriad functions of plastids and significant microscopy evidence suggest that plastid function may be relatively unique in each plant species. Research in this dissertation aims to expand the field of plastid biology to non-model systems by (1) Reviewing the current state of literature to explore how the plastid proteome is shaped and how this process is regulated. How nuclear-encoded plastid-targeted proteins are imported has been enigmatic since the discovery of the main import channel, but recent evidence has helped to resolve the molecular mechanism and regulatory aspects of this process; (2) Establishing

bioinformatics methods to characterize the plastid proteome quickly and efficiently in a range of plant species. A novel application of existing subcellular prediction techniques revealed that only 628-828 proteins are shared between the plastids of all assessed Angiosperms, but between 6- to 25-fold more proteins were specific to single species or taxonomic groups; (3) Analyzing mutational patterns leading to the evolution of novel chloroplast transit peptides. Insertions and deletions, particularly those that caused a shift of the transcriptional or translational start site, were responsible for the majority of novel transit peptides, implicating either exon shuffling as the dominant means of subcellular relocalization to the plastid; and, (4) Characterization of the N-terminal soluble domain of the conserved protein ALB3 when expressed from the chloroplast genome. Perturbations to chlorophyll fluorescence and ion homeostasis suggest that constitutive activation of the ALB3/SecY heterodimeric channel by substrate proteins induces permeability of the thylakoid membrane to leakage of protons and disruption of ion gradients. An overabundance of ALB3 in certain tissues may cause ion dysregulation, particularly for calcium, magnesium, and potassium.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENT.....	iii
ABSTRACT.....	v
LIST OF TABLES.....	x
LIST OF FIGURES.....	xii
ABBREVIATIONS.....	xiv
EQUATIONS.....	xix
FOREWARD.....	1
REFERENCES.....	10
CHAPTER 1: Current Status of Plastid Protein Import and Intraorganellar Trafficking	
1. Abstract.....	19
2. Introduction.....	20
3. Chloroplast Transit Peptides.....	22
4. TOC and TIC Translocons.....	27
5. Intraplastidial Sorting and Processing.....	46
6. Noncanonical Import.....	51
7. Current Status of Plastid Proteomics.....	54
8. Conclusions and Future Directions.....	58
9. Acknowledgments.....	59
10. References.....	60
11. Figures and Tables.....	111

CHAPTER 2: Genome-Scale Characterization of Predicted Plastid-Targeted Proteins in Higher Plants

1. Abstract.....	117
2. Introduction.....	118
3. Results and Discussion	123
4. Conclusions.....	144
5. Materials and Methods.....	145
6. Acknowledgments.....	151
7. References.....	153
8. Figures and Tables	173
9. Additional Files.....	198

CHAPTER 3: Plastid Transit Peptides – Where Do They Come From and Where Do They All Belong? Assessment of Chloroplast Transit Peptide Evolution in Multi-Species and Pan-Genomic Comparisons

1. Abstract.....	199
2. Introduction.....	200
3. Results and Discussion	209
4. Conclusions.....	228
5. Materials and Methods.....	230
6. Acknowledgments.....	234
7. References.....	235
8. Figures and Tables	254
9. Additional Files.....	277

CHAPTER 4: Transplastomic Expression of the N-terminal Soluble Domain of Albino 3 Causes Perturbation of Calcium Homeostasis and Photosynthetic Performance in *Nicotiana tabacum*

1. Abstract.....	278
2. Introduction.....	279
3. Results and Discussion	285
4. Conclusions and Future Directions.....	303
5. Materials and methods	304
6. Acknowledgements.....	313
7. References.....	315
8. Figures and Tables	333
9. Additional Files.....	359
Summary and Future Directions	360

LIST OF TABLES

Chapter 1

Ch2.1; Regulation Points of TOC/TIC	111
Ch2.2; Summary of Non-Green Plastid Proteomics Studies	112

Chapter 2

Ch2.1; Self-Reported Performance of Six Algorithms on Prediction of Plastid-Targeted Proteins	173
Ch2.2; Cross-Validation of Algorithms Using Experimentally-Curated Datasets	174
Ch2.3; Performance of Prediction Algorithms against GFP-validated Proteins from Monocots and Eudicots.....	175
Ch2.4; Combinatorial Prediction Approaches Ranked by Matthew's Correlation Coefficient (MCC).....	176
Ch2.5; Subcellular Targeting Prediction for Selected Genotypes	177
Ch2.6; RBH Clustering Results by Genotype.....	178
Ch2.7; UCLUST Clustering Results by Genotype	179
Ch2.8; Enriched GO Terms for Conserved Plastid-Targeted Clusters Identified using RBH	180
Ch2.9; Enriched GO Terms for Conserved Plastid-Targeted Clusters Identified using UCLUST	182

Chapter 3

Ch3.1; Residue Frequency Bias in Selected Datasets.....	249
Ch3.2; Evolutionary Patterns of Transit Peptides.....	251
Ch3.3; Sources of Transit Peptide Evolution by Dataset.....	252
Ch3.4; GO term enrichment of Nascent Chloroplast-Targeted Proteins in Taxonomically Diverse Genotypes	253

Ch3.5; GO term enrichment of Arabidopsis1001 NPTPs.....	256
Ch3.6; GO term enrichment of BrachyPan NPTPs	257

LIST OF FIGURES

Chapter 1

Ch1.1; Basic Transit Peptide Structural Model 114

Ch1.2; Model of TOC/TIC Translocation 115

Chapter 2

Ch2.1; Venn-Diagram of Combinatorial and Standalone Subcellular Prediction Algorithms 184

Ch2.2; Illustration of RBH and UCLUST Sequence Clustering Methods 185

Ch2.3; Overall Performance of RBH and UCLUST Methods 186

Ch2.4; Workflow Diagram of Sequence Clustering Methods 188

Ch2.5; RBH Visual Representation 190

Ch2.6; UCLUST Visual Representation 191

Ch2.7; Correlation of Total Proteome Size with Nascent Plastid-Targeted Proteins (NPTPs) 192

Chapter 3

Ch3.1; Amino Acid Compositional Changes in Transit Peptides 259

Ch3.2; Residue Frequency in Predicted and Experimentally-Validated Transit Peptides 260

Ch3.3; Residue Composition of Predicted Transit Peptides and Whole Proteome Sequence 261

Ch3.4; Models of Transit Peptide Evolution 263

Ch3.5; Illustration of Phylogenetics Workflow 264

Ch3.6; NPTP Mutation Sources and Characteristics in Multi-Genomic Analysis 265

Ch3.7; NPTP Mutation Sources and Characteristics in Arabidopsis1001 267

Ch3.8; Geographic Distribution of NPTPs in Arabidopsis1001 Accessions.....	268
Ch3.9; NPTP Mutation Sources and Characteristics in BrachyPan	269
Ch3.10; Geographic Distribution of NPTPs in BrachyPan Accessions	270
Ch3.11; Gene Conservation of NPTPs in BrachyPan.....	271
<u>Chapter 4</u>	
Ch4.1; Multiple Sequence Alignment of Alb3 Against YidC	328
Ch4.2; Alignment of the N-Terminal Domain of Plant Alb3 Homologs	330
Ch4.3; Models of Alb3 Topology and Interaction.....	332
Ch4.4; Subcellular Localization of Full-Length ALB3 Homologs	334
Ch4.5; Subcellular Localization of MdAlb3 N-Terminus Truncations	336
Ch4.6; Chlorophyll Fluorescence Parameters of Wildtype and Transgenic Lines	337
Ch4.7; False-Color Chlorophyll Fluorescence Raw Images.....	341
Ch4.8; Flowering Phenotype of AtAlb3 and MdAlb3 Δ N Constructs	343
Ch4.9; Chlorophyll Quantification and Chlorophyll A/B Ratio.....	345
Ch4.10; Plastid Ultrastructure of Wildtype, MdALB3 Δ N(tr), and AtALB3 Overexpression Lines.....	347
Ch4.11; Thylakoid Architecture of Wildtype and MdALB3 Δ N Overexpression Lines.....	349
Ch4.12; Bulk Tissue Ionomics.....	351
Ch4.13; Calcium Response as Measured by Stromal-Targeted Aequorin.	353

ABBREVIATIONS

ABA: Abscisic acid

ACC: Accuracy

Akr: Ankyrin repeat protein

Alb3: Albino 3 protein

AP: Amino-peptidase

APG1: albino or pale green mutant 1

ARC6: accumulation and replication of chloroplasts 6

ATP: Adenosine triphosphate

BAP: 6-benzylaminopurine

BP: Base pairs

CAB: Chlorophyll *a/b* binding protein

CAH1: α -carbonic anhydrase 1

CaMV: Cauliflower mosaic virus

CAS: Ca^{2+} Sensing receptor

CAX: $\text{H}^+/\text{Ca}^{2+}$ exchanger

ceQORH: chloroplast envelope quinone oxidoreductase homolog

CHG: Comparative Homologous Group

CKII: Casein Kinase II

Cpn: chaperonin

cropPAL: Compendium of crop proteins with annotated locations

cTP: Chloroplast transit peptide

CV: Copy number variant

DAG: Days after germination

DGD1: digalactosyldiacylglycerol synthase 1

DGDG: Digalactosyldiacylglycerol

FDX: Ferredoxin

FNR: Ferredoxin:NADP(H) oxidoreductase

Fv/Fm: Maximum quantum efficiency of photosystem II

GA: Gibberellic Acid

GAP: GTPase-activating protein

GEF: Guanine nucleotide exchange factor

GFP: Green fluorescent protein

GO: Gene ontology

GTP: Guanosine triphosphate

ER: Endoplasmic reticulum

Hsp: Heat shock protein

ICP-MS: Inductively-coupled plasma mass spectrometry

IEP: Inner envelope protein

InDel: Insertion/deletion

KB: Kilobase

kDa: Kilodalton

KOC1: Kinase of the outer chloroplast membrane 1

LHCP: Light harvesting complex proteins

MB: Megabase

MCC: Matthew's correlation coefficient

MGDG: Monogalactosyldiacylglycerol

MS: Mass spectrometry

NAA: 1-naphthaleneacetic acid

NADH: Nicotinamide adenine dinucleotide

NADPH: Nicotinamide adenine dinucleotide phosphate

NPP1: nucleotide pyrophosphatase/phosphodiesterase

NPQ: Nonphotochemical quenching

NPTP: Nascent plastid-targeted protein

OEP: Outer envelope protein

OE: Oxygen-evolving complex subunit

ORF: Open reading frame

OOP: Organellar oligopeptidase

PAV: Presence/absence variant

PIC1: Permease in chloroplasts 1

PlsP: Plastid type 1 signal peptidase

PORA: Protochlorophyllide oxidoreductase-A

POTRA: Polypeptide-transport-associated domains

PPDB: Plastid proteome database

PPF-1: *Pisum* post-floral 1

PreP: Presequence protease

Psa: Photosystem I subunit

Psb: Photosystem II subunit

PSI: Photosystem I

PSII: Photosystem II

qE: Energy-dependant quenching

qI: Photoinhibitory quenching

qL: Fraction of total open PSII centers

RBCS: RuBisCO small subunit

RBH: Reciprocal BLAST Hit

ROS: Reactive oxygen species

SE: Sensitivity

SNP: Single nucleotide polymorphis

SnRK2: Sucrose nonfermenting 1-related protein kinase 2

SP: Specificity

SP1: Suppressor of PPI1 locus 1

SPP: Signal peptide peptidase

SRP: Signal recognition particle

STY: Serine/threonine/tyrosine kinases

SUBA: Subcellular localisation database for arabidopsis proteins

TAIR: The Arabidopsis information resource

TAT: Twin arginine translocase

TEM: Transmission electron microscopy

TGD2: trigalactosyldiacylglycerol 2

TIC: Translocon at the inner chloroplast envelope

TL: Translocated loop

TMH: Transmembrane helix

TOC: Translocon at the outer chloroplast envelope

TPP: Thylakoid processing peptidase (see also PlsP1)

UTR: Untranslated region

YCF: Hypothetical chloroplast open reading frame

EQUATIONS

Parameter	Calculation	Description
Fv/Fm	$(F_m - F_o) / F_m$	Maximum quantum efficiency of PSII photochemistry. Maximum efficiency at which light absorbed by PSII is used for reduction of Qa. Healthy plants are usually between .8 and .85.
NPQ	$(F_m - F_m') / F_m'$	Non-photochemical quenching. Monitors the apparent rate constant for heat loss from PSII.
qE	$((F_m - F_m') / F_m') - ((F_m - F_m'') / F_m'')$	Energy-dependent quenching. Associated with light-induced proton transport into the thylakoid lumen. Regulates the rate of excitation of PSII reaction centers. Recovers within minutes.
qI	$(F_m - F_m'') / F_m''$	Photo-inhibitory quenching. Results from photo-inhibition of PSII photochemistry. Recovers more slowly, 20+ minutes.
Phi II	$(F_m' - F_s) / F_m'$	PSII operating efficiency. Estimates the efficiency at which light absorbed by PSII is used for Qa reduction and provides an estimate of linear electron flux through PSII.
qL	$((F_m' - F_s) / (F_m' - F_o')) * (F_o' / F_s)$	Estimates the fraction of open PSII centers (those with Qa in an oxidized state) on the basis of the lake model for PSII.

$$\text{Sensitivity}(i) = \frac{tp}{tp + fn}$$

$$\text{Specificity}(i) = \frac{tn}{tn + fp}$$

$$\text{MCC}(i) = \frac{tp \times tn - fp \times fn}{\sqrt{(tp+fn)(tp+fp)(tn+fn)(tn+fp)}}$$

$$\text{Overall Accuracy (ACC): } \frac{tp + tn}{tp + fp + tn + fn}$$

Where tp is the number of sequences correctly identified as chloroplast-targeted, tn is the number of sequences correctly categorized with other subcellular localizations, fp is the number of nonplastidial sequences incorrectly predicted as chloroplast-targeted, and fn is the number of plastidial sequences that were not detected as such by the algorithm(s). Note that these categorizations are based on the accuracy of the database annotation and any filtering that was applied to data subsets, and they may not reflect biological accuracy.

$$\text{ChlA (nmol/ml)} = 13.43 \times (A664 - A750) - 3.47 \times (A647 - A750)$$

$$\text{ChlB (nmol/ml)} = 22.90 \times (A647 - A750) - 5.38 \times (A664 - A750)$$

$$\text{ChlTotal (nmol/ml)} = 19.43 \times (A647 - A750) + 8.05 \times (A664 - A750)$$

$$\text{ChlA (}\mu\text{g/ml)} = 12.00 \times (A664 - A750) - 3.11 \times (A647 - A750)$$

$$\text{ChlB (}\mu\text{g/ml)} = 20.78 \times (A647 - A750) - 4.88 \times (A664 - A750)$$

$$\text{ChlTotal (}\mu\text{g/ml)} = 12.67 \times (A647 - A750) + 7.12 \times (A664 - A750)$$

Where A647, A664, and A750 is absorbance measured at 647, 664, and 750nm

$$\text{Circularity} = 4 \times (i \times \text{area}) / (\text{perimeter}^2)$$

FOREWARD

Life on earth is made possible by photoautotrophic organisms directly or indirectly through capture and conversion of solar energy. The first photosynthetic prokaryotes appeared on earth sometime around between 2.72 to 3.5 billion years ago (reviewed in Blankenship, 2010), but today, photosynthetic eukaryotes are dominant in both terrestrial and marine environments. Eukaryotic photoautotrophs accomplish photosynthesis using a dedicated organelle called the chloroplast which was acquired in a single endosymbiotic event in which a mitochondriate host developed symbiosis with a free-living cyanobacterium (Cavalier-Smith, 1987; McFadden and Van Dooren, 2004). The plastid became incorporated as an intracellular symbiont between 1.56-1.21 billion years ago (Falcón et al., 2010; Javaux et al., 2004, 2001; Yoon et al., 2004), but are not solely used as photosynthetic organelles. Plastids are the most dynamic organelle of the plant cell both in terms of the vast number of forms they can take as well as the incredible diversity of functions they perform. In addition to photosynthetic reactions performed by the archetypical chloroplast, modern plastids participate wholly or partially in biochemical pathways including those of fatty acids (Ohlrogge and Browse, 1995), branched-chain and aromatic amino acids (Herrmann and Weaver, 1999; Singh and Shaner, 1995), terpenoids (Lichtenthaler et al., 1997), hemes and chlorophylls (Vavilin and Vermaas, 2002), nitrate reduction and nitrogen assimilation (Lam et al., 1996), and sulfate reduction and assimilation (Leustek and Saito, 1999). Additionally, chloroplasts play crucial roles in the regulation of metal ion homeostasis, especially iron regulation via phytoferritin (reviewed in Nouet et al., 2011). Apart from the archetypical chloroplast, plastids commonly differentiate into biochemically-active leucoplasts which participate in terpene synthesis and are important in roots, secretory ducts, and trichomes (Carde, 1984; Charon et al., 1987; Cheniclet and Carde,

1985; Markus Lange and Turner, 2013), as well as into nonvacuolar storage compartments including pigment-storing chromoplasts, starch-storing amyloplasts, protein-storing proteinoplasts, and oil-rich oleoplasts (reviewed in Solymosi and Keresztes, 2013a). Minor morphotypes of plastids include gerantoplasts, which recycle nutrients during senescence (Biswal et al., 2003), xeroplasts, which prepare plastids for long-term quiescence to survive severe drought (Tuba et al., 1994), tannoplasts, which polymerize epicatechins into tannin and export them to the vacuole (Brillouet et al., 2013), and specialized undifferentiated proplastids which participate in nitrogen assimilation reactions in nitrogen-fixing nodules (Solymosi and Keresztes, 2013). This remarkable diversity is reflected in the word “plastid” itself, whose etymology stems from the Greek word πλαστος (*plastos*), meaning “formed” or “molded,” and with the word “chloroplast” deriving from χλωρος (*chloros*), meaning “green.”

Despite the prominent role that plastids play in both primary and secondary metabolite biochemistry, very little of the ancestral cyanobacterial genome remains in modern chloroplast genomes. Only roughly 90 protein-coding genes remain in the 120-160 kb genome of land plant chloroplasts, and nearly all of them are related to gene expression or are directly involved in photosynthesis (Sugiura, 1992). Of the genes left in the chloroplast genome, all plastids have their own complement of rRNA and tRNA genes, and in *Arabidopsis* and other higher plants, the plastid genome encodes for many subunits of photosystems I and II, ATP synthase, the large subunit of RuBisCO, RNA polymerase, and portions of the electron transport chain among others (Sato et al., 1999). Other photosynthetic organisms vary somewhat in terms of plastome size and gene content. On one end, parasites have much higher rates of plastome reduction; *Epifagus*, a genus of parasitic plants that parasitizes beech trees, has a plastome size of 70 kb with 21 protein-coding genes, while the Apicomplexan *Plasmodium falciparum*, the causal

organism of malaria, has just 23 genes and a plastome size of 29.4 kb (Martin and Herrmann, 1998). At the other extreme, algal plastomes can be significantly larger and more complex than higher plant plastomes, commonly coding for 120-130 proteins (Martin and Herrmann, 1998), though larger examples exist. The plastome of the red alga *Porphyra purpurea* is 191 kb and contains around 250 protein-coding genes, including some involved in amino acid and small molecule biosynthesis (Reith and Munholland, 1995), and the green alga *Chlamydomonas mouwsii* has a plastome size of 292 kb (Sugiura, 1992). Yet all plastomes are dwarfed by the genome sizes of modern cyanobacteria which represent the closest living relatives to the ancestral chloroplast. The *Synechocystis* sp. strain PCC6803 genome is 3.57 Mb and contains 3,168 predicted proteins (Kaneko et al., 1996), while *Nostoc punctiforme*, which bears a stronger phylogenetic similarity to chloroplasts (Falcón et al., 2010; Martin et al., 2002), has a genome size of 8.9 Mb with 7,432 potential genes (Meeks et al., 2001). The massive gene loss of the ancestral cyanobacteria can be partly explained by the phenomenon of Muller's ratchet, in which irreversible deleterious mutations gradually occur in asexually reproducing, nonrecombining genomes (Lynch and Blanchard, 1998; Martin and Herrmann, 1998; Muller, 1964). Horizontal transfer of these genes to the nucleus and development of a robust and dynamic import mechanism would have been necessary to complement the gene losses in the plastome to maintain plastid and whole plant viability. The chloroplast proteome of *Arabidopsis* alone is estimated to consist of between 2,000-3,500 unique proteins (Ajjawi et al., 2010; Lu et al., 2011; The Arabidopsis Genome Initiative, 2000), with some estimates being even higher. In species with more divergent plastid morphology and biochemistry than *Arabidopsis*, the number of predicted nuclear-encoded plastid-targeted proteins is likely much larger (Schaeffer et al., 2014). Unfortunately, knowledge of plastid biology is still extremely limited outside of model

organisms. Model systems research has elucidated much about the basal functions of plastids and how they integrate with the entire cell, but how plastids function differ at the molecular level in other species is still vastly underexplored. The increasing availability of whole-genome and transcriptome sequences for a variety of non-model plants offers new tools to examine variation in plastid proteomics using bioinformatics methods. This dissertation builds upon the model systems research by using experimental data validated in those organisms to develop new tools to identify and compare plastid-targeted chloroplast transit peptides in a broad range of higher plant species. Additionally, potential mechanisms of how novel chloroplast-targeted proteins evolve, and how transit peptide evolution has diverged in Monocot and Eudicot species are examined. Finally, a lumen-targeted preprotein is overexpressed from the chloroplast genome and found to indirectly impact plant physiology and development.

Chapter 1: Current Status of Plastid Protein Import and Intraorganellar Trafficking

The plastid is a dynamic organelle of the plant cell capable not only of photosynthesis, but also of complex biochemistry, storage functions, and tolerance to both environmental and biotic stresses. The morphology and function of plastids is governed at the protein level by between 2,000-3,500 unique proteins in the model plant *Arabidopsis* (Ajjawi et al., 2010; Lu et al., 2011), yet the plastid genome itself contains under 100 genes (Sato et al., 1999). The vast majority of these proteins are encoded by the cell nucleus, translated by cytoplasmic ribosomes, and imported post-translationally through several receptor complexes and independent import routes. The most important of these routes is the translocon at the outer chloroplast envelope (TOC) and translocon at the inner chloroplast envelope (TIC), which recognize substrates based on N-terminal presequences known as chloroplast transit peptides (Shi and Theg, 2013). However, several minor import routes also contribute to the chloroplast proteome, most notably

for outer envelope proteins. Once inside the chloroplast, many proteins must be further sorted to sub-organellar locations such as the thylakoid lumen, thylakoid membrane, and inner chloroplast envelope, which requires further modification to the primary transit peptide. Transit peptides have low sequence conservation but have broadly biochemical similarity in three major motifs (Bruce, 2000; Claros et al., 1997; von Heijne et al., 1989). However, much is still unclear about the structure of transit peptides and what role each motif has during import. Furthermore, the mechanism of the TOC and TIC translocons is not fully understood, and even the composition of the core TIC complex is currently under debate. This chapter will review the accumulated scientific knowledge of plastid transit peptides and each of the independent import routes to resolve some of the ambiguity in the recent literature and to arrive at a universal model for canonical import of plastid proteins. The current state of knowledge regarding regulation of protein import at the TOC and TIC translocons, and both experimental and bioinformatics methods of studying plastid proteomics will also be reviewed.

Chapter 2: Genome-Scale Characterization of Predicted Plastid-Targeted Proteins in Higher Plants

The last decade has seen an exponential increase in both the whole genome and transcriptome data as a result of advances in next-generation sequencing techniques. The greatest challenge of modern genomics is quickly, cheaply, and efficiently annotating this sequence data to assign biological significance. In plants, the plastid is a good target for improvement of annotations, as it is responsible not only for photosynthesis, but also a wide range of biosynthetic pathways including fatty acids (Ohlrogge and Browse, 1995), shikimate pathway derivatives (Herrmann and Weaver, 1999), and terpenoids (Lichtenthaler et al., 1997), and contributes strongly to plant tolerance to biotic and abiotic stresses. Almost all of the more than 3,000

estimated plastid proteins are encoded by the nuclear genome and imported post-translationally using N-terminal motifs called chloroplast transit peptides. Holistic characterization of the chloroplast proteome by wet lab methods remains a significant challenge due to low abundance, specific expression patterns, and high error rates (Jeong et al., 2012; Nesvizhskii, 2010; van Wijk and Baginsky, 2011), but *in silico* prediction techniques including the popular TargetP program (Emanuelsson et al., 2007, 2000) can more efficiently estimate plastid-targeted proteins at scale. Ways of optimizing *in silico* prediction of chloroplast transit peptides were explored, and a combination of TargetP with the recently-developed tool Localizer (Sperschneider et al., 2017) was determined to yield excellent biological accuracy and perform quickly on large datasets. To demonstrate this technique, plastid-targeted proteins were predicted for the proteomes of a diverse range of higher plants to explore the question of what proteins are required for basal chloroplast function in photosynthetic plants, and what proteins are not conserved or species-unique. This analysis found between 628-828 proteins which have conserved chloroplast targeting and likely serve critical functions, while individual species have dozens to hundreds of proteins unique to their plastid proteome. Species-unique plastid-targeted proteins may exhibit novel traits even if non-plastid targeted functions are known. For high-value crops such as apple or emerging bioenergy crops such as switchgrass, understanding how these crops differ genetically from other species presents opportunities to improve crop management for maximum yield and minimal postharvest loss. Conversely, identifying non-conserved plastidial proteins in model plants can help to more accurately apply model systems research to plants as a whole.

Chapter 3: Plastid Transit Peptides - Where Do They Come From and Where Do They All Belong? Assessment of Chloroplast Transit Peptide Evolution in Multi-Species and Pan-Genomic Comparisons

The plastid is a dynamic organelle which has high morphological variability in different plant tissues, environmental conditions, developmental time points, and even in different species or cultivars (Schaeffer et al., 2017; Solymosi and Keresztes, 2013). Experimental characterization of the plastid proteome is currently limited mostly to chloroplasts of a small number of plant species, but available reports suggest that many proteins are not conserved across evolutionary space (Suzuki et al., 2015; Wang et al., 2013). However, bioinformatics predictions suggest that a minority of the total plastid proteome is conserved across all higher plants whereas many proteins are unique to individual taxa (Richly and Leister, 2004; Schaeffer et al., 2014). Similarly high levels of non-conserved plastid proteins are reported in Chapter 3. What role these non-conserved proteins play and how they evolve is poorly understood, but the few known examples include mechanisms such as exon shuffling (Arimura et al., 1999; Gantt et al., 1991), alternative start sites (Mackenzie, 2005; Peeters and Small, 2001; Small et al., 1998), and indels (Patzoldt et al., 2006) which introduce changes at the N-terminal end of nuclear-encoded proteins to create novel chloroplast transit peptides. Subcellular relocalization is potentially a common and rapid means of evolutionary change because by changing the environment of a protein, novel functions can be achieved without altering the sequence of the mature sequence itself. In this chapter, non-conserved plastid-targeted proteins identified in Chapter 3 are examined in a multi-genomic comparison to examine mechanisms of chloroplast transit peptide evolution, while pan-genomic resources from the Arabidopsis1001 Proteomes (Cao et al., 2011; Joshi et al., 2012) and BrachyPan (Gordon et al., 2017) projects are analyzed to answer whether similar trends are

found at the subspecies level. Significant differences in the plastid transit peptides of Monocot and Eudicot species were found that may influence patterns of transit peptide evolution in these two clades. At the locus level, the evolution of differentially-targeted alleles was found to be relatively uncommon, but the use of alternative start sites in the different transcript or protein isoforms was more prominent. Insertions or deletions within the N-terminus were the most common means of transit peptide evolution, while residue substitutions were the dominant mechanism in only a third of cases. Enrichment for the functional significance of novel plastid-targeted proteins found that secondary metabolism, transcriptional regulation, and protein-protein interactions are overrepresented. Overall, the plastid proteome was found to be much more dynamic than previously realized even at the subspecies-level, and unique proteins likely have a large impact on the biochemistry, morphology, and regulation of plastids.

Chapter 4: Transplastomic Expression of the N-terminal Soluble Domain of Albino 3 Causes Perturbation of Calcium Homeostasis and Photosynthetic Performance in Nicotiana tabacum

A subset of plastid-targeted proteins undergoes secondary trafficking to the inner envelope, thylakoid membrane, or thylakoid lumen after translocation into the stroma. The thylakoid membrane contains three translocases, Alb3, SecY, and TAT, which govern the import of substrate proteins with different biochemical properties (Frain et al., 2016; Hennon et al., 2015). Membrane translocases create transient pores in the membrane which are accessible to ions, so translocase activity must be carefully controlled to avoid unrestricted loss of membrane potential and energy-generating capacity (Sachelaru et al., 2017). The Alb3 translocase is unique among these in that it does not have a physical translocation pore, but does form heterodimers with the pore-forming cpSecY (Klostermann et al., 2002; Pasch et al., 2005). Alb3 has also been linked to calcium homeostasis activity, but the mechanism of this association is unknown (Wang and

Wang, 2009; Wang et al., 2003). In this chapter, the role of Alb3 in calcium regulation and chloroplast homeostasis was investigated by overexpressing a full-length copy of Alb3 from the chloroplast genome, but a serendipitous transposon insertion truncated the exogenous protein to only the soluble N-terminal domain. However, these transgenic lines experienced significant defects in photosynthetic processing and chloroplast ultrastructure despite lacking any of the known functional domains. Therefore, portions of the Alb3 N-terminus were expressed from the chloroplast genome and showed that the causal element is likely a thylakoid transfer domain. This research provides clues to the link between Alb3 and calcium and also sheds light on previously unrecognized effects of protein trafficking within the chloroplast.

References

- Ajjawi, I., Lu, Y., Savage, L.J., Bell, S.M., Last, R.L., 2010. Large-Scale Reverse Genetics in Arabidopsis: Case Studies from the Chloroplast 2010 Project. *Plant Physiol.* 152, 529–540. <https://doi.org/10.1104/pp.109.148494>
- Arimura, S.I., Takusagawa, S., Hatano, S., Nakazono, M., Hirai, A., Tsutsumi, N., 1999. A novel plant nuclear gene encoding chloroplast ribosomal protein S9 has a transit peptide related to that of rice chloroplast ribosomal protein L12. *FEBS Lett.* 450, 231–234. [https://doi.org/10.1016/S0014-5793\(99\)00491-3](https://doi.org/10.1016/S0014-5793(99)00491-3)
- Biswal, U.C., Biswal, B., Raval, M.K., 2003. Transformation of Chloroplast to Gerontoplast, in: *Chloroplast Biogenesis*. Springer, Dordrecht, Dordrecht, pp. 155–242.
- Blankenship, R.E., 2010. Early Evolution of Photosynthesis. *Plant Physiol.* 154, 434–438. <https://doi.org/10.1104/pp.110.161687>
- Brillouet, J.M., Romieu, C., Schoefs, B., Solymosi, K., Cheynier, V., Fulcrand, H., Verdeil, J.L., Conéjéro, G., 2013. The tannosome is an organelle forming condensed tannins in the chlorophyllous organs of Tracheophyta. *Ann. Bot.* 112, 1003–1014. <https://doi.org/10.1093/aob/mct168>
- Bruce, B.D., 2000. Chloroplast transit peptides: Structure, function and evolution. *Trends Cell Biol.* 10, 440–447. [https://doi.org/10.1016/S0962-8924\(00\)01833-X](https://doi.org/10.1016/S0962-8924(00)01833-X)
- Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K.J., Weigel, D., 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.* 43, 956–963. <https://doi.org/10.1038/ng.911>
- Carde, J.P., 1984. Leucoplasts: a distinct kind of organelles lacking typical 70S ribosomes and

- free thylakoids. *Eur. J. Cell Biol.* 34, 18–26.
- Cavalier-Smith, T., 1987. The Simultaneous Symbiotic Origin of Mitochondria, Chloroplasts, and Microbodies. *Ann. N. Y. Acad. Sci.* 503, 55–71. <https://doi.org/10.1111/j.1749-6632.1987.tb40597.x>
- Charon, J., Launay, J., Carde, J.P., 1987. Spatial organization and volume density of leucoplasts in pine secretory cells. *Protoplasma* 138, 45–53. <https://doi.org/10.1007/BF01281184>
- Cheniclet, C., Carde, J.P., 1985. Presence of leucoplasts in secretory cells and of monoterpenes in the essential oil: A correlative study. *Isr. J. Bot.* 34, 219–238. <https://doi.org/10.1080/0021213X.1985.10677023>
- Claros, M.G., Brunak, S., Heijne, G. Von, 1997. Prediction of N-terminal protein sorting signals. *Curr. Opin. Struct. Biol.* 7, 394–398. [https://doi.org/10.1016/S0959-440X\(97\)80057-7](https://doi.org/10.1016/S0959-440X(97)80057-7)
- Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. <https://doi.org/10.1093/bioinformatics/btq461>
- Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H., 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* 2, 953–71. <https://doi.org/10.1038/nprot.2007.131>
- Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G., 2000. Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. *J. Mol. Biol.* 300, 1005–1016. <https://doi.org/10.1006/jmbi.2000.3903>
- Falcón, L.I., Magallón, S., Castillo, A., 2010. Dating the cyanobacterial ancestor of the chloroplast. *ISME J.* 4, 777. <https://doi.org/10.1038/ismej.2010.98>
- Frain, K.M., Gangl, D., Jones, A., Zedler, J.A.Z., Robinson, C., 2016. Protein translocation and thylakoid biogenesis in cyanobacteria. *Biochim. Biophys. Acta - Bioenerg.* 1857, 266–273.

<https://doi.org/10.1016/j.bbabi.2015.08.010>

Gantt, J.S., Baldauf, S.L., Calie, P.J., Weeden, N.F., Palmer, J.D., 1991. Transfer of *rpl22* to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron.

EMBO J. 10, 3073–3078.

Gordon, S.P., Contreras-Moreira, B., Woods, D.P., Des Marais, D.L., Burgess, D., Shu, S., Stritt, C., Roulin, A.C., Schackwitz, W., Tyler, L., Martin, J., Lipzen, A., Dochy, N., Phillips, J., Barry, K., Geuten, K., Budak, H., Juenger, T.E., Amasino, R., Caicedo, A.L., Goodstein, D., Davidson, P., Mur, L.A.J., Figueroa, M., Freeling, M., Catalan, P., Vogel, J.P., 2017.

Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. Nat. Commun. 8. <https://doi.org/10.1038/s41467-017-02292-8>

Hennon, S.W., Soman, R., Zhu, L., Dalbey, R.E., 2015. YidC/Alb3/Oxa1 Family of Insertases. J.

Biol. Chem. jbc.R115.638171. <https://doi.org/10.1074/jbc.R115.638171>

Herrmann, K.M., Weaver, L.M., 1999. the Shikimate Pathway. Annu. Rev. Plant Physiol. Plant

Mol. Biol. 50, 473–503. <https://doi.org/10.1146/annurev.arplant.50.1.473>

Javaux, E.J., Knoll, A.H., Walter, M.R., 2004. TEM evidence for eukaryotic diversity in mid-

Proterozoic oceans. Geobiology 2, 121–132. [https://doi.org/10.1111/j.1472-](https://doi.org/10.1111/j.1472-4677.2004.00027.x)

[4677.2004.00027.x](https://doi.org/10.1111/j.1472-4677.2004.00027.x)

Javaux, E.J., Knoll, A.H., Walter, M.R., 2001. Morphological and ecological ecosystems. Nature

412, 66–69.

Jeong, K., Kim, S., Bandeira, N., 2012. False discovery rates in spectral identification. BMC

Bioinformatics 13, S2. <https://doi.org/10.1186/1471-2105-13-S16-S2>

Joshi, H.J., Christiansen, K.M., Fitz, J., Cao, J., Lipzen, A., Martin, J., Smith-Moritz, A.M.,

Pennacchio, L.A., Schackwitz, W.S., Weigel, D., Heazlewood, J.L., 2012. 1001 Proteomes:

- A functional proteomics portal for the Analysis of arabidopsis thaliana accessions. *Bioinformatics* 28, 1303–1306. <https://doi.org/10.1093/bioinformatics/bts133>
- Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., Tabata, S., 1996. Sequence analysis of the genome of the unicellular cyanobacterium *synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* 3, 109–136. <https://doi.org/10.1093/dnares/3.3.109>
- Klostermann, E., Droste Gen Helling, I., Carde, J.-P., Schünemann, D., 2002. The thylakoid membrane protein ALB3 associates with the cpSecY-translocase in *Arabidopsis thaliana*. *Biochem. J.* 368, 777–781. <https://doi.org/10.1042/BJ20021291>
- Lam, H.-M., Coschigano, K.T., Oliveira, I.C., Melo-Oliveira, R., Coruzzi, G.M., 1996. The Molecular-Genetics of Nitrogen Assimilation Into Amino Acids in Higher Plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 47, 569–593. <https://doi.org/10.1146/annurev.arplant.47.1.569>
- Leustek, T., Saito, K., 1999. Sulfate Transport and Assimilation in Plants. *Plant Physiol.* 120, 637–644. <https://doi.org/10.1104/pp.120.3.637>
- Lichtenthaler, H.K., Schwender, J., Disch, A., Rohmer, M., 1997. Biosynthesis of isoprenoids in higher plant chloroplasts proceeds via a mevalonate-independent pathway. *FEBS Lett.* 400, 271–274. [https://doi.org/10.1016/S0014-5793\(96\)01404-4](https://doi.org/10.1016/S0014-5793(96)01404-4)
- Lu, Y., Savage, L.J., Larson, M.D., Wilkerson, C.G., Last, R.L., 2011. Chloroplast 2010: A Database for Large-Scale Phenotypic Screening of *Arabidopsis* Mutants. *Plant Physiol.* 155,

1589–1600. <https://doi.org/10.1104/pp.110.170118>

Lynch, M., Blanchard, J.L., 1998. Deleterious mutation accumulation in organelle genomes., in: Woodruff, R.C., Thompson, J.N. (Eds.), *Mutation and Evolution. Contemporary Issues in Genetics and Evolution*, Vol 7. Springer, Dordrecht, pp. 29–39.

<https://doi.org/10.1023/a:1017022522486>

Mackenzie, S.A., 2005. Plant organellar protein targeting: A traffic plan still under construction. *Trends Cell Biol.* 15, 548–554. <https://doi.org/10.1016/j.tcb.2005.08.007>

Markus Lange, B., Turner, G.W., 2013. Terpenoid biosynthesis in trichomes-current status and future opportunities. *Plant Biotechnol. J.* 11, 2–22. <https://doi.org/10.1111/j.1467-7652.2012.00737.x>

Martin, W., Herrmann, R.G., 1998. Gene Transfer from Organelles to the Nucleus: How Much, What Happens, and Why? *Plant Physiol.* 118, 9–17. <https://doi.org/10.1104/pp.118.1.9>

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., Penny, D., 2002. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl. Acad. Sci.* 99, 12246–12251. <https://doi.org/10.1073/pnas.182432999>

McFadden, G.I., Van Dooren, G.G., 2004. Evolution: Red algal genome affirms a common origin of all plastids. *Curr. Biol.* 14, 514–516. <https://doi.org/10.1016/j.cub.2004.06.041>

Meeks, J.C., Elhai, J., Thiel, T., Potts, M., Larimer, F., Lamerdin, J., Predki, P., Atlas, R., 2001. An overview of the Genome of *Nostoc punctiforme*, a multicellular, symbiotic Cyanobacterium. *Photosynth. Res.* 70, 85–106. <https://doi.org/10.1023/A:1013840025518>

Muller, H.J., 1964. The Relation of Recombination to Mutational Advance. *Mutat. Res.* 1, 2–9. <https://doi.org/10.1117/12.722789>

- Nesvizhskii, A.I., 2010. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* 73, 2092–2123. <https://doi.org/10.1016/j.jprot.2010.08.009>
- Nouet, C., Motte, P., Hanikenne, M., 2011. Chloroplastic and mitochondrial metal homeostasis. *Trends Plant Sci.* 16, 395–404. <https://doi.org/10.1016/j.tplants.2011.03.005>
- Ohlrogge, J., Browse, J., 1995. Lipid biosynthesis. *Plant Cell* 7, 957–70. <https://doi.org/10.1105/tpc.7.7.957>
- Pasch, J.C., Nickelsen, J., Schünemann, D., 2005. The yeast split-ubiquitin system to study chloroplast membrane protein interactions. *Appl. Microbiol. Biotechnol.* 69, 440–447. <https://doi.org/10.1007/s00253-005-0029-3>
- Patzoldt, W.L., Hager, A.G., McCormick, J.S., Tranel, P.J., 2006. A codon deletion confers resistance to herbicides inhibiting protoporphyrinogen oxidase. *Proc. Natl. Acad. Sci.* 103, 12329–12334. <https://doi.org/10.1073/pnas.0603137103>
- Peeters, N., Small, I., 2001. Dual targeting to mitochondria and chloroplasts. *Biochim. Biophys. Acta - Mol. Cell Res.* 1541, 54–63. [https://doi.org/10.1016/S0167-4889\(01\)00146-X](https://doi.org/10.1016/S0167-4889(01)00146-X)
- Reith, M., Munholland, J., 1995. Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome. *Plant Mol. Biol. Report.* 13, 333–335. <https://doi.org/10.1515/ae-2016-0033>
- Richly, E., Leister, D., 2004. An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of *Arabidopsis* and rice. *Gene* 329, 11–16. <https://doi.org/10.1016/j.gene.2004.01.008>
- Sachelaru, I., Winter, L., Knyazev, D.G., Zimmermann, M., Vogt, A., Kuttner, R., Ollinger, N., Siligan, C., Pohl, P., Koch, H.G., 2017. YidC and SecYEG form a heterotetrameric protein

- translocation channel. *Sci. Rep.* 7, 1–15. <https://doi.org/10.1038/s41598-017-00109-8>
- Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Tabata, S., 1999. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Res.* 6, 283–290. <https://doi.org/10.1093/dnares/6.5.283>
- Schaeffer, S., Harper, A., Raja, R., Jaiswal, P., Dhingra, A., 2014. Comparative analysis of predicted plastid-targeted proteomes of sequenced higher plant genomes. *PLoS One* 9, e112870. <https://doi.org/10.1371/journal.pone.0112870>
- Schaeffer, S.M., Christian, R., Castro-Velasquez, N., Hyden, B., Lynch-Holm, V., Dhingra, A., 2017. Comparative ultrastructure of fruit plastids in three genetically diverse genotypes of apple (*Malus × domestica* Borkh.) during development. *Plant Cell Rep.* 36, 1627–1640. <https://doi.org/10.1007/s00299-017-2179-z>
- Shi, L.X., Theg, S.M., 2013. The chloroplast protein import system: From algae to trees. *Biochim. Biophys. Acta - Mol. Cell Res.* 1833, 314–331. <https://doi.org/10.1016/j.bbamcr.2012.10.002>
- Singh, B.K., Shaner, D.L., 1995. Biosynthesis of Branched Chain Amino Acids: From Test Tube to Field. *Plant Cell* 7, 935–944. <https://doi.org/10.1105/tpc.7.7.935>
- Small, I., Wintz, H., Akashi, K., Mireau, H., 1998. Two birds with one stone: genes that encode products targeted to two or more compartments. *Plant Mol. Biol.* 38, 265–277. <https://doi.org/10.1023/A:1006081903354>
- Solyosi, K., Keresztes, A., 2013. Plastid Structure, Diversification and Interconversions II. *Land Plants. Curr. Chem. Biol.* 6, 187–204. <https://doi.org/10.2174/2212796811206030003>
- Sperschneider, J., Catanzariti, A.-M., DeBoer, K., Petre, B., Gardiner, D.M., Singh, K.B., Dodds, P.N., Taylor, J.M., 2017. LOCALIZER: subcellular localization prediction of both plant and

- effector proteins in the plant cell. *Sci. Rep.* 7, 44598. <https://doi.org/10.1038/srep44598>
- Sugiura, M., 1992. The chloroplast genome. *Plant Mol. Biol.* 19, 149–168.
<https://doi.org/10.1007/s00438-005-0092-6>
- Suzuki, M., Takahashi, S., Kondo, T., Dohra, H., Ito, Y., Kiriwa, Y., Hayashi, M., Kamiya, S., Kato, M., Fujiwara, M., Fukao, Y., Kobayashi, M., Nagata, N., Motohashi, R., 2015. Plastid proteomic analysis in tomato fruit development. *PLoS One* 10, 1–25.
<https://doi.org/10.1371/journal.pone.0137266>
- The Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815. <https://doi.org/10.1038/35048692>
- Tuba, Z., Lichtenthaler, H.K., Csintalan, Z., Nagy, Z., Szente, K., 1994. Reconstitution of chlorophylls and photosynthetic CO₂ assimilation upon rehydration of the desiccated poikilochlorophyllous plant *Xerophyta scabrifolia* (Pax) Key words. *Planta* 192, 414–420.
- van Wijk, K.J., Baginsky, S., 2011. Plastid Proteomics in Higher Plants: Current State and Future Goals. *Plant Physiol.* 155, 1578–1588. <https://doi.org/10.1104/pp.111.172932>
- Vavilin, D. V., Vermaas, W.F.J., 2002. Regulation of the tetrapyrrole biosynthetic pathway leading to heme and chlorophyll in plants and cyanobacteria. *Physiol. Plant.* 115, 9–24.
<https://doi.org/10.1034/j.1399-3054.2002.1150102.x>
- von Heijne, G., Steppuhn, J., Herrmann, R.G., 1989. Domain structure of mitochondrial and chloroplast targeting peptides. *Eur. J. Biochem.* 180, 535–545.
<https://doi.org/10.1111/j.1432-1033.1989.tb14679.x>
- Wang, A.X., Wang, D.Y., 2009. Regulation of the ALBINO3-mediated transition to flowering in *Arabidopsis* depends on the expression of CO and GA1. *Biol. Plant.* 53, 484–492.
<https://doi.org/10.1007/s10535-009-0089-9>

- Wang, D., Xu, Y., Li, Q., Hao, X., Cui, K., Sun, F., Zhu, Y., 2003. Transgenic expression of a putative calcium transporter affects the time of Arabidopsis flowering. *Plant J.* 33, 285–292.
<https://doi.org/10.1046/j.1365-313X.2003.01627.x>
- Wang, Y.Q., Yang, Y., Fei, Z., Yuan, H., Fish, T., Thannhauser, T.W., Mazourek, M., Kochian, L. V., Wang, X., Li, L., 2013. Proteomic analysis of chromoplasts from six crop species reveals insights into chromoplast function and development. *J. Exp. Bot.* 64, 949–961.
<https://doi.org/10.1093/jxb/ers375>
- Yoon, H.S., Hackett, J.D., Ciniglia, C., Pinto, G., Bhattacharya, D., 2004. A Molecular Timeline for the Origin of Photosynthetic Eukaryotes. *Mol. Biol. Evol.* 21, 809–818.
<https://doi.org/10.1093/molbev/msh075>

CHAPTER 1

Current Status of Plastid Protein Import and Intraorganellar Trafficking

Ryan Christian¹ and Amit Dhingra^{1§}

Target Journal: *Frontiers in Plant Science*

¹Department of Horticulture, Washington State University, Pullman, WA

[§]Corresponding author

RWC: ryan_christian@wsu.edu

AD: adhingra@wsu.edu

Abstract

Plastids are the most morphologically diverse and biochemically versatile intracellular organelles of the plant cell, but they are almost wholly dependent on import of nuclear-encoded, cytosolically-translated proteins. Protein import demands to the plastid are dynamic and change with varying environmental conditions, in different tissues, and in different plant species or cultivars. Most plastid proteins are imported post-translationally by means of a chloroplast transit peptide, or cTP, through receptor and import complexes at the chloroplast outer and inner envelopes known as TOC and TIC, respectively. The mechanisms of how these import pathways work and how chloroplast transit peptides are constructed are beginning to emerge, but much is still unknown about plastid proteome variability at the morphotype- or species-levels.

Furthermore, the current knowledge is limited mainly to model organisms and only to the chloroplast morphotype of plastids, with information on crop plants and more diverse morphotypes being extremely limited. In this review, the current state of research on the import

and processing of nuclear-encoded proteins in the plastid is assessed, including both the canonical TOC/TIC complex as well as alternative targeting pathways. Additionally, techniques for experimental and bioinformatics prediction of the plastid proteome in diverse morphotypes are presented.

Introduction

The chloroplasts of nearly all photosynthetic eukaryotes descend from an ancient cyanobacterial symbiont acquired by mitochondria harboring eukaryotic host between 1.5-1.2 billion years ago (Cavalier-Smith, 1987; Falcón et al., 2010; McFadden and Van Dooren, 2004; Yoon et al., 2004). The chloroplast retains a small genome derived from its cyanobacterial ancestor that ranges in size from 120-160 kb and 90 protein-coding genes in land plant chloroplasts, nearly all of which are related to transcription, translation, or photosynthesis (Green, 2011; Sugiura, 1992). In contrast, modern cyanobacteria have between 3,000-7,500 potential genes (Kaneko et al., 1996; Meeks et al., 2001). In higher plants, chloroplasts are estimated to contain between 2,000 to 5,000 unique proteins in the model species *Arabidopsis thaliana* (Ajjawi et al., 2010; Lu et al., 2011; The Arabidopsis Genome Initiative, 2000). Many of these genes represent horizontal gene transfers from the ancient cyanobacterium to the cell nucleus: up to 18% of the nuclear genome of *Arabidopsis thaliana*, representing roughly 4,500 genes or 1,392 gene families, can be traced to a cyanobacterial origin (Martin et al., 2002). Part of this loss is explained by the phenomenon of Muller's ratchet, in which nonrecombining, asexually-reproducing genomes gradually accumulate irreversible mutations, thus favoring horizontal transfer to the sexually-reproducing nuclear genome (Lynch and Blanchard, 1998; Martin and Herrmann, 1998; Muller, 1964). The similarly reduced plastome sizes of both plant

and algal lineages suggest this process must have occurred very quickly. Experiments assessing DNA transfer from plastids to nuclei estimate rate of transfer to be as high as 1 in 16,000 (Huang et al., 2003) to 1 in 63,291 in pollen grains (Ruf et al., 2007), or as low as 1 in 5 million in vegetative cells (Stegemann et al., 2003). Large insertions of organellar DNA into the nuclear genome have been detected in a wide range of plants, some of which have occurred relatively recently (Ayliffe et al., 1998; Richly and Leister, 2004a, 2004b). However, to maintain biochemical functionality of the organelle post massive transfer of genes from symbiont to host, a robust protein transit and import system for nuclear-encoded plastid-targeted proteins would have been necessary (Garg and Gould, 2016; Zimorski et al., 2014). About 900 genes of cyanobacterial origin are predicted to relocalize to chloroplasts *in vivo* (Richly and Leister, 2004c), with the remainder comprising both non-homologous functional replacements or completely novel eukaryotic proteins.

Presence of an import apparatus is arguably the defining feature of an organelle that separates it from an endosymbiont (Bhattacharya et al., 2007). Examples of well-characterized import apparatuses include the nuclear pores in the nuclear envelope (e.g., Dickmanns et al., 2015; Knockenhauer and Schwartz, 2016; Strambio-De-Castillia et al., 2010), the Sec translocase of the endoplasmic reticulum (e.g., Johnson et al., 2013; Park and Rapoport, 2012), and the TOM/TIM translocons of the mitochondrion (e.g., Wiedemann and Pfanner, 2017). Effective bioinformatics tools are available to predict the respective substrates of these import channels. In contrast, the composition and mechanisms of the chloroplast translocons TOC and TIC needs further characterization; however, the research so far has revealed a highly dynamic and multifaceted import system that is vastly more complex than cognate translocons of other organelles. Plastids have additional internal membrane-bound structures that subdivide protein

import and necessitate modification of the transit peptide. This review summarizes the current state of research on chloroplast transit peptides, the composition, and mechanism of major and minor import pathways, and progress in assessing protein diversity of the chloroplast proteome derived from the nuclear-encoded genes.

1. Chloroplast Transit Peptides

The term “transit peptide” was coined by Chua and Schmidt to distinguish chloroplast and mitochondrial targeting sequences from the simpler signal peptides used for trafficking to the endoplasmic reticulum and secretion pathways (Chua and Schmidt, 1979). The origin of chloroplast transit peptides, or “cTPs,” is debated, but the conservative sorting hypothesis argues that evolution from pre-existing prokaryotic mechanisms is the most parsimonious explanation for import into endosymbiotic organelles (Hartl et al., 1986). The main import channel protein of the chloroplast outer envelope, TOC75, has a homolog called SynTOC75 in the modern cyanobacterium *Synechocystis* which is hypothesized to secrete virulence factor peptides (Reumann et al., 1999; Reumann and Keegstra, 1999). Based on this homology, evolutionarily cTPs may have originally served a secretory function and were later repurposed to facilitate import into the chloroplast using the same channel protein (Bruce, 2000). Secreted virulence factors such as actin-binding proteins and activators of Rho and Rha GTPases coordinate to destabilize the cytoskeleton (reviewed in Barbieri et al., 2002). Similarly, cTPs have GTPase-modulating properties and membrane-destabilizing properties (e.g., Nicolay et al., 1994; Pinnaduwege and Bruce, 1996; Van 't Hof and de Kruijff, 1995a; Wieprecht et al., 2000), leading to a hypothesis that transit peptides evolved out of these ancient virulence factors (Bruce, 2000; Reumann et al., 1999). While speculative, these same sequences may have aided in

evading host immune responses, enabling endosymbiosis to occur in the first place (Reddick et al., 2007).

Modern chloroplast transit peptides lack sequence homology and are highly variable in length, spanning between 13 and 146 residues (Mackenzie, 2005). While no strict sequence conservation is apparent, certain residues are more abundant. Hydrophobic and hydroxylated residues are generally enriched, especially the amino acids alanine and serine (Patron and Waller, 2007; Pujol et al., 2007; von Heijne et al., 1989). In *Arabidopsis*, serine is the most abundant residue in transit peptides at 19.3%, followed by leucine (10.4%), proline (7.3%), alanine (7.1%), and threonine (6.9%) (Li and Chiu, 2010; Zhang and Glaser, 2002; Zybailov et al., 2008). Acidic residues are rare, but positively charged amino acids including arginine and lysine are common especially toward the C-terminal end of the transit peptide (Claros et al., 1997; Zhang and Glaser, 2002). The lack of sequence conservation implies that cTPs do not use a catalytic mechanism for targeting, and instead rely on loosely conserved and spatially promiscuous motifs with generally similar biochemical properties to achieve targeting.

Transit Peptide Structure:

Developing a universal model for transit peptides has proven difficult due to their sequence and length variability. The “homology block hypothesis” was the first major proposed model that described most transit peptides as containing three separate degenerate domains, sometimes encoded by separate exons (Bruce, 2001; Karlin-Neumann and Tobin, 1986; von Heijne et al., 1989). More recent analysis has shown that chloroplast transit peptides are more accurately subdivided into seven subgroups, although only half of the known transit peptides can be confidently organized into these groups (Lee et al., 2008). A unifying hypothesis, the “multi-selection, multi-order” or “M&M” model, attempts to reconcile these observations by using a

more relaxed model of transit peptide construction in which domain organization is spatially unconstrained and allows for duplicate or optional sequence motifs (Li and Teng, 2013). In this way, the organization of transit peptides is not at all dissimilar from assembly of promoter elements: a few core motifs predominate, while many potential cis-acting factors alter translocation efficiency or specificity. Constantly accumulating empirical evidence supports such a model. Earlier observations with the transit peptide of Rubisco small subunit protein, tpRBCS, suggested that insertions had little effect on import efficiency, while deletions cause translocation defects, suggesting flexibility to motif redundancy but not outright loss (Reiss et al., 1989). tpRBCS also contains partially redundant motifs with additive or synergistic effects (Lee et al., 2008, 2006). In contrast, tpCAB (chlorophyll *a/b* binding protein) has more independent motifs that are individually more critical than those of tpRBCS (Lee et al., 2015). Furthermore, hybrid transit peptides of RBCS and CAB reveal that what is a critical motif for one protein may not be for the other (Lee et al., 2015). Recently, experiments in the single-cell C4 plant *B. sinuspersici* demonstrated that the transit peptide sequence of specific proteins is sufficient to discriminate between the dimorphic chloroplasts, suggesting cis-acting transit peptide motifs with specificity for either TOC isoforms or cytosolic chaperone routes (Wimmer et al., 2017). Overall, this evidence points to the M&M model being the closest description of cTPs to date. Much work remains to characterize these cis-acting motifs completely, and they are beyond the scope of this review, however, in this section, the core motifs which correspond roughly to the tripartite homology block model will be addressed (Figure 1).

The first 20 residues at the proximal N-terminus of cTPs are typically uncharged, often beginning with a methionine-alanine residue pair and ending in either glycine or proline (Claros et al., 1997). *In vivo* import capacity is highly influenced by this region, and binding capacity to

Hsp70 has been suggested as the primary factor (Chotewutmontri and Bruce, 2015). As many as 95% of transit peptides in Arabidopsis have predicted Hsp70 binding sites, with 70% of these concentrated in the proximal region (Chotewutmontri et al., 2012; Chotewutmontri and Bruce, 2015; Ivey III et al., 2000). Furthermore, hydroxylated residues are highly abundant in this region, which may serve to provide phosphorylation sites to increase binding affinity the Hsp70/14-3-3 cytosolic chaperone complex (Lamberti et al., 2011a). Interaction of these residues with polar lipid head groups in the outer chloroplast envelope has also been proposed as an energy-independent mechanism for chloroplast-specific trafficking (Pilon et al., 1995; Van't Hof et al., 1993; Van 't Hof and de Kruijff, 1995b). Deletions affecting the proximal domain strongly impact import and binding but not processing (Pilon et al., 1995; Rensink et al., 1998).

The central third of the transit peptide is thought to be primarily a spacer region enriched in small amino acids including alanine, glycine, and proline (Claros et al., 1997; Holbrook et al., 2016). Deletions of this region do not impact binding, but strongly affect import and also block the import of a competing protein (Pilon et al., 1995; Rensink et al., 2000), suggesting that translocation of a mutated version of the transit peptide-protein is stalled and causes blockage of the channel. Notably, prolines appear to be enriched after 10th position, suggesting the need for flexibility of this region (von Heijne and Nishikawa, 1991). Up to 60% of translocation efficiency is attributed to this region (Lee et al., 2006), presumably because a sufficiently long spacer is needed to span both membranes simultaneously and to position the proximal and distal domains to interact with TIC and TOC, respectively (Chen and Li, 2017). At a minimum, transit peptides must contain 32 amino acids in addition to the TIC and TOC binding domains to span the 90 Å distance between the outer and inner membranes (Bionda et al., 2010; Chotewutmontri et al., 2012).

In the C-terminal distal end of cTPs, the most characteristic feature is an enrichment of arginine and lysine (Claros et al., 1997). Deletions of this region affect import efficiency, binding to Toc159, processing by signal peptide peptidase (SPP), and lipid insertion (Pilon et al., 1995; Rensink et al., 2000). Strong interactions with outer envelope lipids have also been noted for this domain (Pinnaduwege and Bruce, 1996). Two major functional motifs have been described for this region. A degenerate motif of “(F/W)(G/P)LK” is extremely sensitive to mutation or deletion as has been reported for preFDX (ferredoxin) (Pilon et al., 1995), and mutation of prolines and glycines around this motif in preRBCS causes loss of interaction with Toc34 and up to 90% reduction of import efficiency (Holbrook et al., 2016). The FGLK motif has been implicated in binding to Toc34 and is common enough to be a good general predictor of chloroplast transit peptides (Li and Teng, 2013). However, this motif is functional only if placed at the C-terminal end (Chotewutmontri et al., 2012). A second “R-X-R” or “di-positive charge” motif is less-well characterized but affects import efficiency (Pilon et al., 1995), and consecutive dipositive motifs may influence specificity for older, actively photosynthesizing chloroplasts (Li and Teng, 2013; Ling et al., 2012).

Some evidence suggests that transit peptide efficiency – and perhaps design – can be abrogated by the biochemical properties of the mature protein. In mitochondrial preproteins, up to 70% of the variance in import efficiency is attributed to the sequence charge and hydrophobicity of the mature protein (Doyle et al., 2013). While chloroplast transit peptides are complex, the mature protein may exhibit electrostatic repulsion or hydrophobic aggregations requiring further modification of the transit peptide. For example, the mature domains of membrane-bound proteins have significant inhibitory effects on transit peptide efficiency, and such proteins need a long spacer region in the transit peptide (Inaba and Schnell, 2008; Rolland

et al., 2016). Proteins that contain transmembrane domains also tend to have more conserved prolines in their transit peptides, suggesting that more hydrophobic mature proteins need both increased length and flexibility of the transit peptide (Lee et al., 2017b). In other cases, portions of the mature protein may be necessary to produce a fully functional transit peptide. In RBCS, FDX, CAB, DNA-J, and protochlorophyllide oxidoreductase-A (PORA), efficient targeting of C-terminal GFP fusions requires parts of the mature protein downstream of the signal peptide peptidase (SPP) cleavage site (Comai et al., 1988; Lee et al., 2008; Pilon et al., 1995; Rensink et al., 1998; Van't Hof et al., 1991). While the majority of transit peptides are between 50-55 amino acids (Bienvenut et al., 2012; Kleffmann et al., 2007), predicted cleavage sites often fall short of this length. In those shorter transit peptides, parts of the fully functional transit peptide likely fall downstream of the cleavage site. Addition of simple spacer motifs following the cleavage site enables high-efficiency translocation of a wide variety of both soluble and membrane-bound proteins (Shen et al., 2017). Overall, these reports suggest that the functional transit peptide includes the N terminal residues of the peptide to be translocated and that properties of the mature protein can influence translocation activities of an otherwise functional transit peptide.

2. TOC and TIC Translocons

There are several described routes for import of chloroplast-targeted proteins. However, the major protein complex for chloroplasts which imports almost all stromal and thylakoid proteins is known as TOC/TIC, for "Translocon at the Outer envelope of Chloroplasts" and "Translocon at the Innner envelope of Chloroplasts," respectively. TOC and TIC form a pore of at least 25.6 Å (Ganesan et al., 2018), and over 95% of preprotein transport to the chloroplast occurs through this complex, consuming a significant amount of chemical energy (L. X. Shi and

Theg, 2013). Treating the plastid as a perfect sphere, the number of TOC complexes per chloroplast was initially estimated to have as low as 7.5 receptors per square micrometer (Friedman and Keegstra, 1989; Pilon et al., 1992), but more recent estimates suggest the true number to be closer to $140/\mu^2$ due to low binding efficiency in the *in vitro* experiments (Morin and Soll, 1997). Translocation rates have been estimated at 14,000-22,000 precursors per minute per chloroplast (Pilon et al., 1992; Van 't Hof and de Kruijff, 1995b), which translocate about 2×10^7 preproteins every 24 hours (Pfisterer et al., 1982). This process requires an average of 650 ATP per preprotein, consuming 0.6% of total light-saturated energy production (L.-X. Shi and Theg, 2013). This section will discuss the current state of understanding of the essential components of each translocon, explain potential points of regulation, and propose a revised mechanistic model of translocation.

TOC Composition

The isolation of TOC was described for the first time by Waegemann and Soll, who identified an Hsp70 homolog and an 86-kDa protein along with several other lower molecular weight products (Waegemann and Soll, 1991). These proteins, as well as 75 and 34 kDa subunits, were later confirmed by Schnell and coworkers (Kessler et al., 1994; Perry and Keegstra, 1994; Schnell et al., 1994). The 34 and 86 kDa proteins, later identified as TOC34 and TOC86 (159) respectively, were found to be GTP-binding proteins that use GTP hydrolysis to stabilize early transport intermediates and function as receptors (Kessler et al., 1994; Seedorf et al., 1995; Young et al., 1999), while the 75 kDa protein was found to serve as the TOC channel protein TOC75 (Hinnah et al., 2002, 1997; Tranel et al., 1995). The core TOC complex exists in a stoichiometric ratio of between 1:3:3 (Kikuchi et al., 2006) and 1:4:4 (Schleiff et al., 2003c) between TOC159, TOC33, and TOC75, respectively. The difference in these ratios has been

suggested to be due to differences in the polymerization of whole TOC complexes caused by proteolysis of the large cytosolic acidic domain (A-domain) of TOC159 (Kikuchi et al., 2006; L. X. Shi and Theg, 2013). While the supercomplex has not been successfully imaged to date, the singlet complex that has been observed has a central finger-like domain (TOC159) surrounded by four curved translocation channels (Schleiff et al., 2003c). While these three proteins comprise the core TOC complex, an array of cytosolic and membrane-bound proteins assist in translocation, each of which is discussed in greater detail below.

The channel of the TOC complex, TOC75, is a beta-barrel protein that forms a pore of 14 Å in diameter (Hinnah et al., 2002; Schleiff et al., 2003a). TOC75 has low-level binding activity to plastid transit peptides and is also capable of discriminating against mitochondrial preproteins. Polypeptide-transport-associated (POTRA) domains of TOC75 mediate preprotein selectivity in the intermembrane space (Chen et al., 2016; Ertel et al., 2005; Hinnah et al., 2002; O'Neil et al., 2017; Paila et al., 2016). POTRA domains simultaneously bind to molecular chaperones in the intermembrane space (Paila et al., 2015) and with preproteins (O'Neil et al., 2017), and thus TOC75 serves dual roles as a channel protein and as a scaffold for intermembrane space interactions (O'Neil et al., 2017; Paila et al., 2016; Richardson et al., 2014). Arabidopsis contains two highly similar orthologs, Toc75-III and Toc75-IV (Baldwin et al., 2005; Jackson-Constan and Keegstra, 2001). Additionally, a distantly-paralogous 80-kDa outer envelope protein (OEP80) was previously described as Toc75-V (Eckart et al., 2002); it is not a true TOC subunit, but may be required for insertion of Toc75 into the outer membrane (Gentle et al., 2004; Huang et al., 2011; Inoue and Potter, 2004).

The smaller of the two GTPase receptors includes two major paralogs, TOC33, and TOC34. A single isoform is found in some species including pea and rice while two isoforms are

found in others including Arabidopsis and maize (Hirohashi and Nakai, 2000; Jelic et al., 2003). Much of the early work was done in pea, which contains a single paralog. psTOC34 is 34 kDa, but in Arabidopsis, the orthologous protein is 33 kDa (atTOC33), while a functionally distinct paralog is 34 kDa (atTOC34) (Aronsson et al., 2006; Jelic et al., 2003). Due to this nomenclature confusion, the Arabidopsis homologs will be referred to unless specified otherwise. TOC33 consists of a GTPase “G-domain” and a short C-terminal segment facing the intermembrane space (Seedorf et al., 1995). The C-terminal segment is essential for biogenesis of TOC33 (Li and Chen, 1997; Richardson et al., 2014), while the G-domain exhibits GTP hydrolysis activity and can form homodimers in its GDP-bound state (Sun et al., 2002; Yeh et al., 2007). TOC33 forms homodimers (Jelic et al., 2003; Sun et al., 2002; Weibel et al., 2003; Yeh et al., 2007) and heterodimers with TOC159 (Weibel et al., 2003; Yeh et al., 2007). Dimerization is dependent on conserved arginines in the G-domain; alanine substitution at R128 in psToc34 and R130 in atToc33 causes loss of homodimerization, heterodimerization, and also reduces GTPase activity (Reddick et al., 2007; Sun et al., 2002; Weibel et al., 2003; Yeh et al., 2007). This conserved arginine deeply embeds in the binding pocket of a dimerized partner, likely acting as a GTPase-activating protein (GAP) for the bound protein (Koenig et al., 2008; Sun et al., 2002).

TOC159, the larger GTPase receptor of the TOC complex, has multiple isoforms of differing molecular weights, including TOC90, TOC120, TOC132, and TOC159 in Arabidopsis, all of which have three major domains: a GTPase (G-) domain, a C-terminal membrane (M-) domain, and a hypervariable N-terminal acidic (A-) domain (Hiltbrunner et al., 2001a; Jackson-Constan and Keegstra, 2001). While the first two are relatively static in sequence, the A-domain is hypervariable and potentially responsible for some specificity of different TOC isoforms (Demarsy et al., 2014; Inoue et al., 2010). TOC159 was long thought to be a GTPase motor

(Schleiff et al., 2003b) but more recent evidence has confirmed that it functions instead as a GTPase-activated switch (Li and Chiu, 2010; Richardson et al., 2018), and that the force required for translocation comes instead as a pulling mechanism from stromal Hsp70 (Liu et al., 2014; Shi and Theg, 2010; Su and Li, 2010). Like TOC33, TOC159 can also form dimers using similarly conserved arginines (Yeh et al., 2007). Mutants that have defects in both binding and hydrolysis have impaired rates of translocation (Agne et al., 2009), but mutants which bind but not hydrolyze GTP increase translocation rates (Wang et al., 2008), suggesting that GTP-bound TOC159 is the translocation-active form.

The membrane-bound TOC64 is a transiently-interacting factor that associates with TOC33/34 (Qbadou et al., 2006; Schleiff et al., 2003c) and has soluble domains on both sides of the outer membrane (Sohrt and Soll, 2000). On the cytosolic face, TOC64 has clamp-type tetratricopeptide repeats which are involved in binding to cytosolic chaperones (Aronsson et al., 2007; Qbadou et al., 2006). Interestingly, absence of TOC64 does not seem to significantly affect either *in vitro* or *in vivo* translocation in *Arabidopsis* (Aronsson et al., 2007; Schleiff et al., 2003b), and causes only slight defects in *Physcomitrella patens* (Hofmann and Theg, 2003). However, another study has shown that mutant *toc64* plants have decreased import rates of preOE33 and preRBCS, changes in growth rate in high light, reductions in maximum quantum yield of photosynthesis in low light, and altered salinity tolerance (Sommer et al., 2013). Additionally, *toc33/toc64* double mutants have lower levels of TOC75 protein despite increased expression of TOC75 transcript, suggesting a potential role in stabilizing the TOC complex (Sommer et al., 2013). Thus, TOC64 may not be required under normal conditions, but more likely enhances the efficiency of import and assists in regulation during periods of high import demand or environmental stress.

An additional membrane-bound protein called TOC12 was also thought to nucleate transient intermembrane complexes with TOC64, TIC22, and Hsp70 (Becker et al., 2004; Qbadou et al., 2007). However, TOC12 has been found to be instead a fragment of a larger DnaJ-J8 protein localized to the stroma, and mutants of the gene are not impacted in translocation (Chiu et al., 2010). Additionally, no Hsp70 isoforms localize to the intermembrane space (Ratnayake et al., 2008). While TOC12 could be the true alternatively-spliced form of DnaJ-J8 (Chiu et al., 2010), it is more likely an artifact of DnaJ-J8 and TOC64 both binding Hsp70 homologs, resulting in false positives during purification. In the absence of further information, TOC12 should not be considered a *bona fide* TOC component.

In addition to the core membrane-bound TOC subunits, an array of transiently-interacting soluble proteins support preprotein trafficking to TOC. Two distinct cytosolic trafficking routes use protein chaperones, but preproteins can also travel unaided to the TOC complex, albeit with reduced import efficiency. The first chaperone-mediated route involves binding of preproteins with Hsp90 in the cytosol (Qbadou et al., 2006). TOC64 efficiently binds Hsp70.1 and Hsp90 (Schweiger et al., 2013) and is thought to serve as an intermediate receptor before passing preproteins to TOC by the interaction of its membrane domains with TOC33 (Qbadou et al., 2006; Schleiff et al., 2003c). This route appears to play a key role in delivering preproteins to the TOC complex, as Hsp90 inhibitors cause significant defects in translocation efficiency (Inoue et al., 2013). The alternative chaperone-mediated route involves a 14-3-3/Hsp70 complex, which binds cTPs that have been phosphorylated by the cytosolic kinases STY8, STY17, and STY46 at serine and threonine residues (Lamberti et al., 2011a; May and Soll, 2000). While binding to this complex increases efficiency of import up to 5-fold (May and Soll, 2000), disruption of the phosphorylation sites and 14-3-3 interaction does not cause mistargeting (Nakrieko et al., 2004).

STY kinases play an essential role in the transition from etioplasts to chloroplasts, suggesting that chaperone-assisted routes are important during periods of high protein import demand (Lamberti et al., 2011b). Interestingly, actin is also immunoprecipitated by TOC and TIC subunits, especially TOC159 (Jouhet and Gray, 2009). Though actin is not considered to be a part of the TOC complex, it is possible that preproteins could be delivered with the aid of actin filaments for some or all of the cytosolic import pathways.

Regulation of TOC

Regulation of protein import at TOC appears to be primarily based on the use of alternative GTPase isoforms to regulate selectivity of preproteins. Mutant phenotypes for the TOC33/34 (Gutensohn et al., 2004, 2000; Jarvis et al., 1998) and TOC90/120/132/159 (Bauer et al., 2000; Kubis et al., 2004) receptor gene families are nonredundant, suggesting multiple specialized TOC isoforms (Bauer et al., 2000; Ivanova et al., 2004; Jarvis et al., 1998; Kessler and Schnell, 2009). The distinct TOC complex isoforms are hypothesized to reduce competition between protein classes so that photosynthetic proteins can maintain constant import rates (Kessler and Schnell, 2006). Some preproteins import preferentially into chloroplasts while others import more efficiently into leucoplasts or equally into both morphotypes (Wan et al., 1996). Examples of leucoplast-specific proteins are rare but include at least one porin and a pyruvate kinase protein (Fischer et al., 1994; Wan et al., 1995). Preprotein selectivity is likely mediated by the hypervariable A-domain of the TOC159 GTPase family. The A-domain does not bind transit peptides directly, but exchange of the A-domains between Toc159 homologs is sufficient to transfer the respective preprotein selectively (Dutta et al., 2014; Inoue et al., 2010; Smith et al., 2004). These results suggest that the A-domain relies on an exclusion mechanism, perhaps based on steric hindrance or electrostatic repulsion (Inoue et al., 2010; Smith et al.,

2004). TOC159 is associated with photosynthetic proteins and has higher expression in leaves, while TOC132 and TOC120 are functionally redundant paralogs which import housekeeping proteins and are expressed constitutively at low levels (Bauer et al., 2000; Ivanova et al., 2004; Kubis et al., 2004; Smith et al., 2004).

Posttranslational modification helps achieve inducible control of TOC specificity. Cleavage of the A-domain of Toc159 family members is widely observed in *in vitro* experiments (Bölter et al., 1998; Chen et al., 2000), and removal of this domain removes preprotein specificity, converting the translocon into a more general importer (Dutta et al., 2014; Inoue et al., 2010). Controlled cleavage suggests to a protease activity rather than mechanical disruption or other experimental artifacts, but has this remains to demonstrated conclusively *in vivo* (Agne and Kessler, 2010; Demarsy et al., 2014). For more permanent changes in protein translocation needs, turnover of the TOC159 GTPases is mediated by Suppressor of PPI1 Locus 1 (SP1), an E3 ligase in the outer envelope (Ling et al., 2012). SP1 is activated by stress and plays a pivotal role in stress tolerance by depleting TOC under stress conditions (Ling and Jarvis, 2015). Turnover of TOC receptors may also enable a rapid transition to chromoplasts, leucoplasts, or other plastid morphotypes (Barsan et al., 2012; Reiland et al., 2011).

Both phosphorylation and formation of disulfide bridges under oxidizing conditions mediate modification of import rate through TOC. At least 12 sites on the A-domain of TOC159 can be phosphorylated (Agne et al., 2010; Demarsy et al., 2014). Sucrose nonfermenting 1-related protein kinase 2 (SnRK2) proteins phosphorylate TOC159 in an ABA-dependent manner (P. Wang et al., 2013). Interestingly, ABA mutants are impaired in chloroplast translocation (Zhong et al., 2010) and three independent ABA-deficient tomato mutants (*hp3-1*, *flacca*, and *sitens*) accumulate larger, more abundant, and more highly-pigmented chromoplasts (Egea et al.,

2010; Galpaz et al., 2008), most likely because TOC159 selectively imports photosynthetic proteins. A second integral outer membrane kinase, Kinase of the Outer Chloroplast membrane 1 (KOC1), was found to regulate phosphorylation of the A-domain in Arabidopsis (Zufferey et al., 2017), while a protein of similar mass does the same in pea (Fulgosi and Soll, 2002). Finally, a casein kinase II (CKII) has been implicated in A-domain phosphorylation *in vitro* (Agne et al., 2010). In all cases described thus far, phosphorylation of TOC159 appears to stimulate import activity. TOC33/34 are also proposed to be phosphorylated *in vivo* by soluble kinase and by a 98 kDa membrane-bound kinase, both of which remain unidentified (Fulgosi and Soll, 2002; Jelic et al., 2003, 2002; Sveshnikova et al., 2000).

Some reports find that GTP binding and preprotein binding activities of TOC33 are significantly inhibited by phosphorylation (Jelic et al., 2003; Sveshnikova et al., 2000), but phosphomimic and phosphoknockout residue mutants do not have a significant phenotype (Aronsson et al., 2006). The disparity between these observations remains unresolved, although it is possible that the partial redundancy of atToc34 can sufficiently complement the mutant phenotypes. Finally, preprotein import is highly influenced by plastid redox state, part of which is mediated by TOC (Hirohashi and Nakai, 2000; Kuchler et al., 2002; Stengel et al., 2008). Disulfide bridge formation has been observed in all the major TOC protein components in oxidizing conditions which induces supramolecular crosslinking and decreases import efficiency (Seedorf and Soll, 1995; Sjuts et al., 2017; Sohr and Soll, 2000; Stengel et al., 2009). This mechanism may serve to lock TOC in a translocation-incompetent state, preventing import into senescent or stressed chloroplasts until conditions improve (Stengel et al., 2009).

TIC Composition

In contrast to the TOC complex, the composition of TIC core subunits and even the identity of the channel protein has been subject to considerable debate. Many inner membrane proteins have been copurified with either TOC subunits or preproteins, including Tic55 (Caliebe et al., 1997), Tic56 (Köhler et al., 2015), Tic62 (Küchler et al., 2002), Tic20 (Kouranov et al., 1998), Tic21/PIC1 (Teng et al., 2006), Tic22 (Kouranov et al., 1998), Tic110 (Kessler and Blobel, 1996; Lübeck et al., 1996), Tic40 (Kessler and Blobel, 1996; Wu et al., 1994), Tic100, Tic56, and Tic214/Ycf1 (Kikuchi et al., 2013).

One of the proposed channel proteins, TIC110, is an integral membrane protein with a soluble 98 kDa stromal C-terminus and a 9 kDa N-terminal membrane anchor (Jackson et al., 1998). The soluble domain of TIC110 associates with Hsp93/ClpC and Cpn60 in the stroma (Akita et al., 1997; Kessler and Blobel, 1996; Nielsen et al., 1997), and also binds preproteins in the absence of chaperones (Inaba et al., 2003). Dominant negative mutants and downregulation of TIC110 decrease translocation efficiency and knockouts are seedling lethal (Inaba et al., 2005). TIC110/*tic110* heterozygotes have growth and protein translocation defects, while other translocon components including *ppi-1* (atToc33), *ppi-2* (atToc159), *tic40*, and *hsp93-V*, are completely recessive (Kovacheva et al., 2005). Deletions in the C-terminal stromal domain of TIC110 disrupt binding with Hsp93 and impart a dominant-negative phenotype, suggesting that Hsp93 interaction is key to the overall function of TIC110 (Inaba et al., 2005).

TIC20, the other proposed channel protein, is an integral membrane protein with four alpha helices (Kovács-Bogdán et al., 2011) and is observed to have a direct role in inner membrane translocation (Chen et al., 2002). While smaller than TIC110, TIC20 forms homo-oligomers and directly interacts with preproteins (Campbell et al., 2014; Kikuchi et al., 2009;

Kouranov et al., 1998; Kovács-Bogdán et al., 2011; Rensink et al., 2000). Arabidopsis has four isoforms including TIC20-I, TIC20-II, TIC20-IV, TIC20-V, of which TIC20-I is the major isoform and has the most severe mutant phenotypes (Hirabayashi et al., 2011; Teng et al., 2006). Knockouts of TIC20-II, TIC20-IV, and TIC20-V, in contrast, lack visible phenotypes (Hirabayashi et al., 2011; Kasmati et al., 2011). However, double mutants of TIC20-I and TIC20-IV are more severe than TIC20-I alone, while the loss of only one TIC20-I allele in a *tic20-IV* homozygous background causes growth defects (Hirabayashi et al., 2011; Kasmati et al., 2011; Kikuchi et al., 2013). These phenotypes implicate TIC20-IV as a minor isoform, perhaps involved in the import of TOC132/TOC34 substrates. In support of this, TIC20-I has the highest expression in photosynthetic tissues, while TIC20-IV is mainly expressed in roots, mirroring the expression patterns of the TOC159/TOC33 and TOC(120/132)/TOC34 isoforms (Hirabayashi et al., 2011; Kovács-Bogdán et al., 2011). Additionally, TIC20-I has higher affinity for photosynthetic precursors (Chen et al., 2002; Kasmati et al., 2011; Kikuchi et al., 2009). Of the two remaining isoforms, TIC20-II has been detected in chloroplast-containing vesicles during leaf senescence (Diaz-Mendoza et al., 2016), suggesting a specific role in gerantoplast recycling, while TIC20-V is localized to thylakoids and has a yet-unknown role (Machettira et al., 2011).

TIC22 is an intermembrane protein that is conserved from cyanobacteria to higher plants and binds specifically to the intermembrane space POTRA domains of TOC75 in plants and periplasmic POTRA domains of OMP85 in cyanobacteria (Paila et al., 2015; Tripp et al., 2012). Thus, TIC22 has been proposed to function as a scaffold for TOC-TIC supercomplexes (Becker et al., 2004; Qbadou et al., 2007; Soll and Schleiff, 2004). The crystal structure of TIC22 shows two hydrophobic funnels or grooves on either side of the molecule which allows TIC22 to function as a non-specific, ATP-independent chaperone to prevent preprotein aggregation

(Glaser et al., 2012; Tripp et al., 2012). There are two TIC22 homologs found in Arabidopsis, atTIC22-III and atTIC22-IV, both of which are expressed constitutively, although TIC22-IV is expressed at 5-fold higher levels (Kasmati et al., 2013; Rudolf et al., 2013). Double mutants have chlorotic phenotypes during germination and in high-light conditions, supporting the hypothesized role of preventing preprotein aggregation during periods of high import demand.

TIC40, a peripheral membrane protein, interacts with both TIC110 and the soluble stromal protein Hsp93, indicating more of a co-chaperone role (Chou et al., 2003; Stahl et al., 1999). TIC40 null mutations are not lethal, but are pale and slow-growing, with defects in protein import (Chou et al., 2003). TIC40 has a single N-terminal transmembrane anchor and a large C-terminal stromal domain with a tetratricopeptide repeat domain followed by Sti1p/Hop (Hsp70/Hsp90-organizing protein) and Hip (Hsp70-interacting protein) domains (Bédard et al., 2007; Chou et al., 2003; Stahl et al., 1999). Preprotein binding to TIC110 stimulates recruitment of TIC40, which subsequently activates ATPase-modifying activities in the Hip/Hop domains (Chou et al., 2006). Interestingly, TIC40 overexpression causes a 10-fold increase in S-adenosyl methionine-dependent methyltransferase (IEP37), inner envelope triose phosphate translocator (PPT), and TIC110, but does not affect abundance of outer membrane, stromal, or thylakoid membrane proteins, suggesting that TIC40 could act in the biogenesis of the inner membrane (Singh et al., 2008). Alternatively, TIC40 may also act as a molecular ratchet to prevent preprotein backsliding (Bédard and Jarvis, 2005).

In addition to the major subunits found in close association with TIC, several peripheral membrane and soluble proteins interact transiently. A triad of TIC32, TIC62, and TIC55 has been proposed to form a “redox regulon” affecting TIC110, and helping to adjust preprotein import in response to chloroplast energy status (Soll and Schleiff, 2004). TIC56 and

YCF1/TIC214 are thought to be integral to TIC20-containing complexes and could facilitate complex assembly (Kikuchi et al., 2013). Tic56 copurifies with Toc159 and is required for protein import but not chloroplast biogenesis, with knockout mutants having an albino phenotype (Köhler et al., 2015). Hsp93/ClpC (hereafter referred to as Hsp93) is an Hsp100 protein that targets hydrophobic stretches in preproteins, but it also participates separately in protein degradation as part of the Clp protease complex. During translocation, Hsp93 forms ring-like hexamers associated with Tic110 and is the only chaperone consistently found in import complexes (Akita et al., 1997; Flores-Pérez et al., 2016; Inaba et al., 2005; Kouranov et al., 1998; Nielsen et al., 1997; Rosano et al., 2011; Sjögren et al., 2014). Hsp93 directly binds preproteins during early processing stages (Huang et al., 2016), but only if the recognition motif is present at the far N-terminus of the transit peptide (Bruch et al., 2012). These data suggest that the Hsp93 hexamer is assembled on Tic110 before import begins, and incoming peptides must be threaded through the core before translocation. Hsp93 is hypothesized to provide an initial force for unfolding, which is expanded and completed by other chaperones such as Hsp70 and Hsp90C (Huang et al., 2016). Import assays suggest that Hsp93 is not the rate-limiting step for TIC translocations, and its function in translocation seems to not involve its unfoldase function (Kovacheva et al., 2005). Alternatively, Hsp93 may also serve as a protein quality control mechanism for TIC due to its association with the CLP protease (Bruch et al., 2012; Flores-Pérez et al., 2016; Sjögren et al., 2014). Comparatively less is understood about the other stromal chaperones. Cpn60 and its associated protein Cpn20 form heptameric rings that are recruited to Tic110 in an ATP-dependent manner (Akita et al., 1997; Inoue et al., 2013; Kessler and Blobel, 1996; Nielsen et al., 1997). Hsp70 is implicated in translocation for both plants and mosses (Shi and Theg, 2010; Su and Li, 2010), and is thought to provide the driving force for translocation as

an ATPase motor (Chotewutmontri and Bruce, 2015; Huang et al., 2016; Liu et al., 2014). Finally, an integral membrane protein originally termed TIC21 was described to be associated with TIC complex and was found to impact protein translocation (e.g., Kikuchi et al., 2009; Teng et al., 2006), but more recent evidence has suggested that this subunit is actually an iron permease transporter, PIC1 (Permease in chloroplasts 1) (Duy et al., 2007, 2011; Gong et al., 2015). Mutant PIC1 lines are chlorotic and experience iron anemia in the plastids with iron toxicity in the cytoplasm, which likely impacts translocation indirectly (Duy et al., 2007). Based on these data, PIC1 unlikely to be a direct participant in protein transport and is therefore not a TIC component.

Regulation of TIC

Based on the current understanding, regulation at TIC is tied to the physiological status of the individual plastid, rather than isoform composition as in TOC. Formation of disulfide bridges is common in TIC subunits, including intramolecular bridges in Tic110, Tic40, Tic55, DnaJ-J8, and supramolecular bridges between Tic40 and Tic110 (reviewed in Balsera et al., 2010). At least some of these disulfide bridges are regulated by stromal thioredoxins (Bartsch et al., 2008). As in the case of TOC, disulfide bridge formation arrests active translocation, while reducing agents and dithiols are effective at relieving this inhibition (Stengel et al., 2009).

Additionally, protein-protein interactions regulate the import rate through TIC. The redox regulon of Tic32, Tic55, and Tic62 all negatively regulate Tic110 and Tic40 based on redox status and other physiological conditions (Caliebe et al., 1997; Hörmann et al., 2004; Küchler et al., 2002). Tic32 is an NADPH-dependent dehydrogenase that binds competitively to NADPH and calmodulin, thus integrating both redox and calcium levels to fine-tune protein translocation affinity or efficiency (Chigri et al., 2005; Küchler et al., 2002). TIC activity responds strongly to

calcium levels, likely as a result of TIC32 activity (Balsera et al., 2009). TIC62 is also an NADPH-dependent dehydrogenase, but it binds with ferredoxin:NADP(H) oxidoreductase (FNR) instead of calmodulin (Küchler et al., 2002; Stengel et al., 2008). While NADPH-bound, TIC62 decreases translocation, but upon FNR binding, it dissociates into a soluble complex in the stroma (Chigri et al., 2006; Stengel et al., 2008). TIC55 is a Rieske-type monooxygenase which was initially described to have effects on translocation (Caliebe et al., 1997), but a lack of definitive phenotype in the mutants has cast doubt on that role (Boij et al., 2009; Chou et al., 2018). However, observed roles in chlorophyll breakdown and dark-induced senescence may instead indicate a specific regulatory function in senescent plastids (Chou et al., 2018; Hauenstein et al., 2016).

TIC Channel

The identity of the core TIC channel has been a subject of debate for some time due to observed channel activities for both TIC110 and TIC20. TIC110 is one of the most abundant proteins of the inner membrane (Kessler and Schnell, 2006), and is present in TOC/TIC supercomplexes (Chen and Li, 2017; Kouranov et al., 1998), so it was widely accepted as the channel protein. Furthermore, it interacts with a wide variety of other subunits and chaperones involved in TIC translocation, suggesting a core role if not acting as the channel itself. Early observations showed that purified TIC110 has preprotein-dependent channel activity (Balsera et al., 2009; Heins et al., 2002), while antisera raised against TIC110 block an inner envelope anion channel, suggesting proximity to the channel mouth (van den Wijngaard and Vredenberg, 1999). However, this view has been challenged recently with an argument for a core 1-megadalton TIC complex comprised of a TIC20 channel supported by TIC56, TIC100, and TIC214/Ycf1 subunits (Kikuchi et al., 2013, 2009). In this model, TIC40 and TIC110 function instead as chaperone-

recruiting scaffolds (Inoue et al., 2013; Kikuchi et al., 2009; Nakai, 2015a; Paila et al., 2015). In support of this new TIC hypothesis, Tic20 is similar in sequence and topology to Tim17/23, the inner membrane channel proteins of mitochondria (Inaba and Schnell, 2008; Kasmati et al., 2011), whereas the Tic110 crystal structure indicates that it is unlikely to form a channel *in vivo* (Tsai et al., 2013). Furthermore, less than 5% of Tic110 is associated with TOC complexes based on chromatography experiments, which would be unlikely for a *bona fide* channel protein (Kouranov et al., 1998), Tic110 is also absent in the apicoplasts of Apicomplexans, which although simpler than higher plant plastids, retain a functional TOC/TIC translocon (Nakai, 2015a). However, several observations render the TIC20 model incomplete. TIC20 is between 8 to 100-fold less abundant than TIC110 (Kovács-Bogdán et al., 2011), although it is still present at a ratio of 1:2.5 between TIC20 and TOC75, which could be expected if one TIC channel serves a quadruplet or sextuplet TOC channel (Kikuchi et al., 2013). The most significant problem for the Tic20 hypothesis lies in inconsistent genetic evidence for its supporting subunits. Ycf1 is absent from the plastid genome of grasses, glaucophytes, rhodophytes, and parasitic plants, while TIC56 and TIC100 are also absent outside of higher plants (de Vries et al., 2015; Nakai, 2015b). Furthermore, a high level of import of a subset of proteins is still observed when TIC56 or YCF-1 are inhibited (Bölter and Soll, 2017; Köhler et al., 2015). Due to inconsistencies in both the TIC110 and TIC20 models, many authors have instead suggested that there are two independent TIC channels. One hypothesis posits that Tic110 serves as the general translocon pore while Tic20 imports a specialized subset (Kovács-Bogdán et al., 2011), and another argues for a redox-active Tic110 channel and a redox-independent Tic20 channel (Stengel et al., 2009). Finally, others suggest that Tic110 and Tic20 operate as independent but equally important channels (Bölter and Soll, 2016; Demarsy et al., 2014). Overall, the current evidence supports a

Tic20-centered channel, but questions regarding the compositional inconsistency of the TIC channel must be addressed before it can be definitively proven.

Model of Translocase Mechanism

The overall process of translocation through TOC/TIC has been classically described in three sequential steps: 1. An energy-independent, reversible binding step to the outer membrane and translocon (Andrès et al., 2010; Bräutigam et al., 2007; Pogson et al., 2008), 2. Insertion across the outer membrane, requiring low levels of ATP and GTP (Andrès et al., 2010; Bräutigam et al., 2007; Inaba and Schnell, 2008; Jarvis, 2008; Leister, 2003; Pogson et al., 2008; Stengel et al., 2007), and 3. translocation across the inner membrane, requiring high levels of stromal ATP (Constan et al., 2004; Inaba et al., 2005). Recent work by the Schnell lab examined the topology of early translocation intermediates and placed the N-terminus of transit peptides in contact with the TIC complex, the central region in contact with TOC75, and the C-terminal portion in contact with the TOC GTPases (Richardson et al., 2018). This data, as well as recent findings in TIC translocation, has been incorporated into a revised model of chloroplast protein import that outlines what occurs at each stage of the import process (Figure 2).

During the first stage of translocation in the revised model, energy-independent binding is governed by the interaction between the transit peptide and both lipid and protein components of the chloroplast outer membrane. The chloroplast outer membrane contains high concentrations of non-bilayer and chloroplast-specific galactolipids monogalactosyldiacylglycerol (MGDG) and digalactosyldiacylglycerol (DGDG) (Pinnaduwege and Bruce, 1996) as well as unusual constituents such as sulfolipids and negatively-charged phosphatidylglycerol (Block et al., 2007; Joyard et al., 1991). Even at the higher estimates of 140 receptors per square micrometer, the outer membrane has unusually low protein content which exposes large areas of open lipid

surface area for transit peptide interaction (Block et al., 1983; Bruce, 2000). Transit peptides have been observed to change secondary structure when exposed to MGDG (Bruce, 1998; Wienk et al., 2000), and *in vitro* tests suggest that high-efficiency protein translocation is highest at physiological MGDG levels of about 20 mol% (Pinnaduwege and Bruce, 1996; Van't Hof et al., 1993; Van 't Hof and de Kruijff, 1995b). Furthermore, treatment of chloroplasts with phospholipase inhibits transfer rate of preproteins to the TOC complex but does not affect binding to the TOC complex itself (Kerber and Soll, 1992). Protein-dependent interactions may also be mediated by the POTRA domains of TOC75, which face the intermembrane space and are thus protected from proteases (Paila et al., 2016). Finally, energy-independent binding likely occurs between the highly acidic A-domain of TOC159 paralogs (e.g., Jackson-Constan and Keegstra, 2001) and the arginine/lysine-enriched C-terminal end of transit peptides.

While the second stage of chloroplast protein import has historically been defined by translocation across the outer membrane, early binding intermediates contact components of TIC before committal to full translocation (Richardson et al., 2018), and stromal chaperones including Hsp70 and Hsp93 bind at this stage (Huang et al., 2016). Because multiple TIC components are involved in early stages of transit peptide interaction, the intermediate stages likely involve “priming” of both translocons into an import-competent state. In support of this, transit peptide mutants that either lack a C-terminal GTPase binding domain or have insufficient length and flexibility can bind to TOC but not translocate (e.g., Pilon et al., 1995; Rensink et al., 2000, 1998), while transit peptides with a reversed sequence can bind but not trigger translocation (Chotewutmontri et al., 2012). Priming of the translocation reaction requires low levels of GTP and ATP suggesting interaction both with the TOC GTPases and stromal heat shock proteins. Preproteins act as GTPase nucleotide exchange factors by disrupting homo- and

heterodimers of TOC33 and TOC159, triggering GTP hydrolysis (Jelic et al., 2003; Oreb et al., 2011; Reddick et al., 2007). Therefore, transit peptides require simultaneous interaction of TIC-binding components at the proximal end and TOC-binding components at the distal end along with a sufficient spacer motif to stretch across the two membranes. This topology was recently confirmed at the earliest stages of transit peptide binding (Richardson et al., 2018).

Interaction with the TOC33 GTPase likely stabilizes early translocation intermediates (reviewed in Andrès et al., 2010; Tee, 2018), but TOC159 likely serves as a GTP-mediated switch which is translocation-competent when GTP-bound but is inactive or less competent in a GDP-bound state (Agne et al., 2009; Koenig et al., 2008; Sun et al., 2002; Wang et al., 2008). A final “primed” stage emerges when both TOC33 pairs are homodimerized in a GDP-bound state while GTP-bound TOC159 binds the preprotein (Chang et al., 2017; Richardson et al., 2018; Tee, 2018). On the TIC side of this mechanism, the requirement for low ATPase activity implies the involvement of the stromal chaperones. Although the identity of the TIC components can only be speculated at this point, it is likely a complex of Tic110/Tic40 plus their associated chaperones. This hypothesis could also explain the high molar abundance of Tic110 in comparison to Tic20: if the GTPase switch is the rate-limiting step of translocation, a high molar abundance of Tic110 could facilitate rapid translocation of preproteins. Hsp70 and Hsp93 are obvious candidates, as both bind in the early stages of import (Huang et al., 2016). The *in vivo* import capacity is strongly influenced by the first ten residues (Chotewutmontri et al., 2012), and 77% of transit peptides contain an Hsp70 binding site in the first 20 residues (Chotewutmontri and Bruce, 2015). However, Hsp93 binding is exclusively dependent on a binding site at the extreme N-terminus (Bruch et al., 2012), while Hsp70 can bind multiple regions in both the transit peptide and mature protein (Huang et al., 2016). Given this data, it is likely that both

chaperones bind simultaneously in a configuration that threads the proximal transit peptide residues through an Hsp93 hexamer while Hsp70 stabilizes this complex and prevents backsliding.

Upon successful priming, the transit peptide is in a “stretched” configuration with both the N- and C-termini bound by the TIC and TOC receptor subunits, respectively. Stimulation of a final TOC159 GTPase round by a true transit peptide triggers the final stage of translocation by releasing C-terminal transit peptide domain from TOC159, upon which uninhibited translocation can commence. Given that a minority of TIC110 is associated with the TIC complex (Kouranov et al., 1998), it is tempting to speculate that membrane diffusion of TIC40 and TIC110 provides some of the force necessary for spooling the preprotein through TIC, or otherwise facilitates rapid clearing and re-priming of TIC. High ATP consumption during this phase by Hsp70 and Hsp93 assists in translocation. Of the two, ATPase activity of Hsp70 has a greater effect on translocation rate and efficiency than Hsp93, suggesting that it is the motor protein (Huang et al., 2016; Liu et al., 2014). In support of this mechanism, an ATPase motor complex of chloroplast-encoded Ycf2 and FtsH proteins has also been recently described (Kikuchi et al., 2018). If this is the case, Hsp70 could instead function as “teeth” in a Brownian ratchet mechanism (Esaki et al., 1999; Yamano et al., 2008). There are several testable hypotheses emerging to unravel the roles of each chaperone in preprotein translocation.

3. Intraplastidial Sorting and Processing

Once successful translocation is initiated, the transit peptide must be cleaved to produce a mature, stable protein (Zhong et al., 2003). Cleavage may alternatively reveal secondary transit signals to route proteins to the inner envelope, thylakoid membrane, or lumen. In this section,

secondary targeting and processing of the cleaved transit peptide will be discussed in further detail.

N-terminal Maturation

Initial cleavage of the transit peptide is performed by signal peptide peptidase (SPP) (Richter and Lamppa, 1999, 1998), but the N-terminus of the mature protein is polished in most cases in a process called “maturation.” The N-terminal residue is a major determinant of protein stability in the plastid, following the “N-end” rule (Apel et al., 2010; Bachmair et al., 1986; Dougan et al., 2012; Gibbs et al., 2014; Nishimura et al., 2013; Rowland et al., 2015; Tasaki et al., 2012; van Wijk, 2015). Artificial peptides starting with glutamic acid, methionine, and valine are especially stable in chloroplasts, while peptides starting with asparagine, cysteine, glutamine, histidine, isoleucine, proline, and threonine are unstable (Apel et al., 2010). Maturation is controlled by stromal amino-peptidases (APs), of which seven have been identified by shotgun proteomics in *Arabidopsis* chloroplasts, with leucine-, glutamine-, and amino-APs being most abundant (Zybailov et al., 2008). Both leucine and glutamine are associated with protein instability, making the high abundance of these APs a strong indicator of a plastid N-end rule. For a more comprehensive review of N-terminal maturation, refer to van Wijk, 2015.

Transit Peptide Degradation

Free transit peptides are membrane-seeking and can penetrate membranes, potentially causing toxicity by disrupting membrane potential and decoupling redox status (Nicolay et al., 1994; Pinnaduwege and Bruce, 1996; Van 't Hof and de Kruijff, 1995a; Wieprecht et al., 2000). To prevent this, a succession of proteases acts sequentially on cleaved peptides until only free amino acids remain. First, the transit peptide undergoes secondary cleavage by SPP, with resulting fragments released into the stroma (Richter and Lamppa, 1999). Larger fragments

between 20-65 amino acids are then processed by presequence proteases 1 and 2 (PreP1/2) (Glaser et al., 2006; Ståhl et al., 2005), and shorter fragments of 11-20 amino acids are degraded by organellar oligopeptidase (OOP) (Kmiec et al., 2013). Final degradation of 3-5 residue peptides into free amino acids occurs via the metalloprotease M17-20 (Teixeira et al., 2017).

Inner Membrane Trafficking

Proteins bound for the inner envelope typically contain canonical N-terminal transit peptides that function identically to stromal transit peptides (Cline and Henry, 1996; Lee et al., 2017a); exceptions are covered in section 4 (Noncanonical import). As a result, most inner envelope proteins following the “post-import” route are TIC-dependent and can be correctly localized even if expressed from the chloroplast genome (Singh et al., 2008). An inner membrane SecAYE translocase, cpSec2, was recently identified (Li et al., 2015; Skalitzky et al., 2011) and reliable evidence for cpSec2-dependent insertion of inner membrane proteins including TIC40 and FtsH12 has been described (Li et al., 2017). Based on similar characteristics, PIC1 (Singhal and Fernandez, 2017) and Tic110 (Li and Schnell, 2006; Lübeck et al., 1997) are also Sec2 candidates. Interestingly, Sec2 depletion also affects the abundance of thylakoid membrane proteins including Alb3, SecY1, and TatC (Li et al., 2017). The thylakoid membrane is thought to originate invagination of the inner envelope (e.g., Kobayashi et al., 2007), so Sec2 may be responsible for initial insertion of these thylakoidal translocases. Targeting of preproteins to Sec2 is still poorly understood, but appears to be dependent on hydrophobic motifs or serine/proline-rich motifs in substrate proteins, whereas specificity for the thylakoid cpSec1 is influenced by signal recognition particle (SRP) signals in preprotein N-termini (Knight and Gray, 1995; Singhal and Fernandez, 2017; Tripp et al., 2007; Viana et al., 2010).

Thylakoid Trafficking

Preproteins bound for the thylakoid membrane or lumen have a secondary transit peptide called a “thylakoid transfer domain” downstream of the SPP cleavage site (de Boer and Weisbeek, 1991; Smeekens et al., 1986). The thylakoid membrane contains three translocases: SecY, Twin-Arginine Translocase (TAT), and Albino3 (Alb3), each of which has unique properties and substrates.

Among soluble luminal proteins, 50% are estimated to be substrates of cpSec, while cpTAT transports the remaining 50% (Peltier et al., 2002; Schubert et al., 2002). The cpSec channel is roughly analogous to ER and bacterial homologs, containing a SecY channel subunit, the ATPase subunit SecA, and a complex-stabilizing SecE (Frain et al., 2016). Transport of luminal proteins through cpSec is driven by ATPase activity of its SecA subunit (Aldridge et al., 2009; Dalbey and Chen, 2004). The cpTAT complex consists of Tha4, Hcf106, and cpTatC, corresponding to the *E. coli* homologs TatA, TatB, and TatC, respectively (reviewed in Frain et al., 2016). cpSec is incapable of transporting folded proteins due to a fixed-width channel, while TAT can transport proteins that either require folding in the stroma to be functional or proteins that fold too rapidly to be efficiently processed by cpSec (Aldridge et al., 2008; Hynds et al., 1998; Matos et al., 2008). TAT forms oligomeric complexes dynamically and is hypothesized to form flexible oligomers to accommodate folded protein of varying sizes (Frain et al., 2016; Gohlke et al., 2005). Additionally, the Tha4 (TatA) component of TAT has proofreading capacity and can reject misfolded proteins (Matos et al., 2008). TAT does not have inherent ATPase activity, but is instead driven by proton motive force or membrane potential, and is referred to as the ΔpH pathway in early literature (Braun et al., 2007; Theg et al., 2005). Specificity for Sec and TAT is not well-understood, although the eponymous twin-arginine motif

of TAT substrates appears necessary though insufficient for the avoidance of cpSec; additional positively-charged amino acids at the N- and C-termini appear to more effectively avoid cpSec (Bogsch et al., 1997; Zhu et al., 2013).

Translocation of membrane-bound thylakoid membrane proteins is performed by cpSec, Alb3, or a combination of the two. There are 138 integral thylakoid membrane proteins in the PPDB database of which about 50 are expressed from the plastid genome (Celedon and Cline, 2013; Sun et al., 2009). Alb3, unlike cpSec and cpTAT, is a monomeric protein that lacks a central channel and instead contains a “greasy slide” that functions to chaperone soluble lumenal loops across the membrane (Hennon et al., 2015; Kumazaki et al., 2014; Wang and Wang, 2009). Most Alb3-dependent proteins tend to have an uneven number of transmembrane domains, thus requiring translocation of at least one soluble terminal domain (Woolhead et al., 2001). Studies in the bacterial homolog YidC have demonstrated that introducing positive charges to translocated loop regions can lead to YidC dependency (Gray and Henderson-Frost, 2011; Zhu et al., 2013). Additionally, insertion of polar residues into transmembrane domains leads to YidC requirement (Price and Driessen, 2010; Zhu et al., 2013). Conversely, soluble translocated domains greater than 100 amino acids tend to be less dependent on YidC, perhaps due to dilution of unfavorable charges and hydrophobicity (Wickström et al., 2011). Similar import determinants for Alb3 are likely given the high degree of sequence conservation between Alb3 and YidC (Hennon et al., 2015). Many Alb3 and cpSec substrates are delivered by the chloroplast signal recognition particle (SRP) pathway, which recognizes substrates via 20-30 residue hydrophobic signal peptide-like sequences downstream of a positively-charged motif and upstream of a polar lumenal peptidase site (Eichacker and Henry, 2001). Interestingly, chloroplast SRP complexes lack the RNA subunit typically found in ER-type SRPs and instead

contain a novel subunit cpSRP43 that prevents aggregation of substrates (Falk and Sinning, 2010; Shuenemann et al., 1998). SRP-bound preproteins are recognized by cpFtsY, which guides SRP complexes to Alb3 or Sec and releases the protein in a GTP-dependent manner (Eichacker and Henry, 2001; Frain et al., 2016; Goforth et al., 2004; Kogata et al., 1999). For further information on plant SRPs, see Ziehe et al., 2017.

4. Noncanonical import

While the vast majority of plastid-targeted proteins appear to use the TOC/TIC translocons (Row and Gray, 2001), a small subset of proteins are targeted and inserted via alternative routes. Estimates of noncanonically-imported proteins range from just 24 outer membrane proteins (Inaba and Schnell, 2008) to 343 proteins in *Arabidopsis* (Armbruster et al., 2009). However, investigation of 28 of the proteins revealed that only four had unambiguous chloroplast localization, while an additional six had ambiguous targeting (Armbruster et al., 2009). Thus, such proteins likely represent the minority of plastid-targeted proteins. Three major noncanonical import routes have been determined to date.

Outer envelope proteins

The outer envelope proteins are a major group of noncanonically-imported proteins, and TOC75 is the only known example of this group with a canonical transit peptide (Inoue et al., 2001). Despite earlier suggestions that outer envelope proteins insert spontaneously into the membrane (Jarvis and Robinson, 2004; Schleiff and Klösgen, 2001), many lose import competency in thermolysin-treated plastids and most are likely still dependent on TOC75 (Hofmann and Theg, 2005; Tu et al., 2004). For example, OEP14 is capable of high-efficiency integration into proteoliposomes containing just TOC75 (Tu et al., 2004), while outer envelope

protein (OEP) 64 and digalactosyldiacylglycerol synthase 1 (DGD1) compete with the import of preRBCS, also implicating a role of TOC75 (Hofmann and Theg, 2005). Targeting of C-terminal tail-anchored proteins including TOC159, OEP7, OEP9, and OEP64 occurs via cytosolic ankyrin repeat proteins Akr2A and Akr2B, which bind simultaneously to cytosolic ribosomes during translation and to lipids in the chloroplast outer membrane, thus decreasing the requirement for interaction with the GTPases (Bae et al., 2008; Dhanoa et al., 2010; Kim et al., 2015, 2014). Uniquely for an outer membrane protein, TOC159 requires both TOC75 and TOC33 for insertion (Bauer et al., 2002; Hiltbrunner et al., 2001b; Smith et al., 2002; Wallas et al., 2003). Curiously, the C-terminal tail anchor domain of TOC159 from *Bienertia sinuspersici* can function as a cleavable transit peptide, suggesting that its targeting is also stabilized by TOC33 interaction (Lung and Chuong, 2012). In contrast, TOC33 appears to insert directly into protein-free liposomes, independent of chaperones or TOC75 (Dhanoa et al., 2010; Qbadou, 2003). A cytosolic loop of TOC75 has been shown to help recruit TOC33, but the channel itself may not be used in this process (Ertel et al., 2005).

Inner membrane proteins:

Several exceptions to the dominant “post-import” route for inner envelope-localized proteins appear to bypass TIC-mediated translocation and are GTP-independent. Cases such as this likely require the TOC75 channel but bypass the GTPase-mediated switch as they do not become imported to the stroma. The “stop-transfer” pathway uses a lateral insertion mechanism at TIC to insert directly without passing through a stromal intermediate stage (Brink et al., 1995; Knight and Gray, 1995). Known examples include albino or pale green mutant 1 (APG1), accumulation and replication of chloroplasts 6 (ARC6), and trigalactosyldiacylglycerol 2 (TGD2) (Froehlich and Keegstra, 2011; Motohashi et al., 2003; Viana et al., 2010). Other

candidates for this route include chloroplast envelope quinone oxidoreductase homolog (ceQORH) (Miras et al., 2007, 2002) and TIC32 (Nada and Soll, 2004), both of which lack transit peptides and require only low levels of ATP. The proposed mechanism is based on bulky hydrophobic residues of the mature transmembrane domains, but high glycine content and low proline content appear to also have a role (Froehlich and Keegstra, 2011). More unusual examples of TIC-independent import include the soluble TIC22, which does not compete with stromal preprotein for translocation yet is still ATP-dependent and requires protease-sensitive proteins of the outer membrane (Kouranov et al., 1999). Plastid type 1 signal peptidase (PlsP1) is another example but is insensitive to protease digestion of the outer membrane (Inoue et al., 2005).

Glycosylated proteins

In rare cases, chloroplast-targeted proteins that require glycosylation or other forms of specialized modification cannot use canonical import pathways. α -carbonic anhydrase 1 (CAH1) (Villarejo et al., 2005) and nucleotide pyrophosphatase/phosphodiesterase (NPP1) (Nanjo et al., 2006) use signal peptides to direct initial transport into the endoplasmic reticulum (ER), followed by TOC-independent import to chloroplasts. Treatment with brefeldin A, an inhibitor of Golgi-mediated vesicle transport, severely inhibits accumulation in the chloroplast, suggesting that vesicular fusion may deliver them to the intermembrane space. After this fusion, glycosylated proteins could enter the stroma by vesicle budding from the outer membrane, through an unknown inner membrane transporter, or by passage through the TIC translocon independent of TOC (Radhamony and Theg, 2006). If the revised TOC/TIC mechanism is correct, the latter hypothesis would be the most parsimonious route: proteins inserted into the intermembrane space could bypass the TOC159 GTPase switch and engage with the TIC import machinery

freely. Some (e.g., Bhattacharya et al., 2007) have suggested that ER-trafficked proteins could represent relicts of a more ancient targeting mechanism. In most algae, plastids are enveloped within the ER and thus require bipartite transit peptides consisting of an N-terminal signal peptide and a buried chloroplast transit peptide (reviewed in Nassoury and Morse, 2005). In plants, however, it is likely that the TOC/TIC translocons represent the primitive import mechanism, and that ER trafficking later evolved to accommodate glycosylated proteins (Bodyl et al., 2009).

5. Current Status of Plastid Proteomics

Plastid morphogenesis and differentiation is a complex and multifaceted process which alters the quantity and abundance of nuclear-encoded proteins as well as the transcription and translation rate of genes in the chloroplast genome (Liebers et al., 2017). Plastid morphotype variants are well-described by microscopy, including not only the archetypical chloroplast, but also pre-chloroplastic etioplasts, pigmented chromoplasts, biochemically-active leucoplasts, and starch-storing amyloplasts (reviewed in Solymosi and Keresztes, 2013; Wise, 2007). Within a species, plastids can rapidly change form in response to developmental or environmental cues, as exemplified in the etioplast to chloroplast transition in seedlings, and the chloroplast to chromoplast transition which is well-established in tomato (e.g., Muraki et al., 2010). New forms of plastids are still being discovered, including tannosomes, which export phenolic precursors to the vacuole (Brillouet et al., 2013), dessicoplasts (xeroplasts), which protect plastids during extreme drought stress (Ingle et al., 2008; Tuba et al., 1994), and phenyloplasts, which accumulate a single large osmiophilic vesicle that stores phenol glucosides in vanilla orchid (Brillouet et al., 2014). In developing apple peel, novel hybrid plastids displaying both

chromoplast and leucoplast characteristics arise in the epidermal cell layer, while hybrid chloroplast/amyloplasts predominate in collenchymal tissue (Schaeffer et al., 2017; Solymosi and Keresztes, 2013). These morphological and biochemical changes are mediated by regulation of plastid gene expression and differential import of nuclear-encoded plastid-targeted genes. Classification of these proteins has historically been accomplished at scale using high-throughput shotgun proteomics, with a smaller percentage using fluorescent protein chimeras. *In silico* methods are increasingly being used as their predictive power continues to improve and are widely used to supplement proteomics studies. However, the predictions need to be validated with experimental approaches. These methods and their merits and limitations will be discussed further in this section.

High-Throughput Proteomics

A wealth of experimental data exists for chloroplast-targeted proteins in Arabidopsis, rice, and maize, represented in databases including AT_CHLORO (Ferro et al., 2010), Suba4 (Heazlewood, 2005; Heazlewood et al., 2007; Hooper et al., 2017), plprot (Kleffmann et al., 2006), and PPDB (Sun et al., 2009; Van Wijk, 2004). Due to this exhaustive coverage, this review will not focus on chloroplast-targeted proteins, and will instead examine plastid proteomics in non-green plastids and in non-model species. Understandably, such research has been hampered by the difficulty of isolating different plastid morphotypes. Notable studies published so far are summarized in Table 2. Due to the biological diversity of metabolic functions carried out by non-green plastids as well as significantly different isolation, detection, analysis, and curation methods, the capture of proteomics data from a single development stage or plastid type does not provide a comprehensive picture of the plastid proteome. For instance, only 32% of the proteins identified in chromoplasts by Suzuki et al. (2015) overlapped with

those identified by Barsan et al. (2010). Commonly, chromoplasts are enriched in carotenoid storage and synthesis proteins and jasmonic acid biosynthetic enzymes (Barsan et al., 2010; Siddique et al., 2006; Suzuki et al., 2015; Y. Q. Wang et al., 2013; Zeng et al., 2011; Zhu et al., 2018). Elaioplasts (oleoplasts) of citrus peel are significantly more active in terpene synthesis compared to chromoplasts of the same tissue while having far fewer proteins involved in carotenoid metabolism (Zhu et al., 2018). Amyloplasts are most abundant in carbohydrate metabolism and hexose transporters as expected, but also contain significant lipid and amino acid biosynthesis (Andon et al., 2002; Balmer et al., 2006; Dupont, 2008). Etioplasts contain much of the photosynthetic machinery with a few exceptions, as well as abundant amino acid and lipid biosynthesis enzymes (Kanervo et al., 2008; Kleffmann et al., 2007; von Zychlinski et al., 2005). For all types of non-green plastids, enrichment of NTP translocators, hexose transporters, and carbohydrate metabolism enzymes point to heterotrophic but highly active metabolism. Similarly, abundant chaperone and heat shock proteins suggest that protein translation and import is extremely active in all plastid types, not just in chloroplasts. Finally, redox enzymes found in all plastids but especially abundant in chromoplasts allude to a need for pathogen defense, membrane protection, and reactive oxygen species detoxification. Up to 21 genes involved in the ascorbate-glutathione cycle alone were found in tomato chromoplasts (Barsan et al., 2010). This research is an encouraging step in expanding knowledge of plastid proteomics beyond the chloroplast and is expected to shed light on a more holistic view of plastid proteomics and enable greater accuracy in predicting not only plastid localization but also categorization of different import classes (van Wijk and Baginsky, 2011). However, the research in non-green plastids is still far from reaching parity with chloroplast proteomics.

Bioinformatics Predictions

An attractive alternative to high-throughput proteomics is the use of computer algorithms to predict and compare plastid-targeted proteins. This methodology is not dependent on isolation technique but is limited to identifying localization without expression level or plastid morphotype information. This method is time- and cost-efficient compared to wet lab methods, and with proper application, can approach a higher level of accuracy. Prediction software typically examines the N-terminal portion of protein models and uses either sequence and motif characteristics or annotation and sequence homology to determine localization. The lack of conserved sequence or domain structure in chloroplast transit peptides complicates prediction, but TargetP (Emanuelsson et al., 2007, 2000), the most commonly-used program for predicting chloroplast-targeted proteins, performs with 86% sensitivity and 65% specificity when compared with curated mass spectrometry data (Zybailov et al., 2008). Newer algorithms incorporating annotation and homology features as well as approaches using a combination of algorithms achieve even greater accuracy. A comprehensive analysis of six major programs using publicly-available organellar proteomics data found that a “2 of 3” combination of TargetP, MultiLoc2, and Localizer achieved better overall performance than any single predictor (Christian et al., 2019b, unpublished). Bioinformatics methods have largely been used on either small datasets or as a tool to curate mass spectrometry data, but several publications have applied them at the whole-genome level. The first such approach identified 2,261 proteins in Arabidopsis and 4,853 in rice (*Oryza sativa*) with predicted plastid localization; 880 and 817 of these proteomes are thought to originate from the cyanobacteria respectively (Richly and Leister, 2004c). This study furthermore described that the number of non-essential genes outnumber essential genes and suggested that the majority of plastid-targeted proteins are eukaryotic in origin. This analysis was

expanded to seven higher plant species, and the publication reported that only 737 proteins constituted the core, essential plastid-targeted genes (Schaeffer et al., 2014). Additionally, Schaeffer et al. reported a low of 795 species-specific plastid-targeted proteins in *Prunus persica* and a high of 4,817 in *Malus × domestica*. Arabidopsis alone had 2,154 species-specific plastid-targeted proteins. Recently, these techniques were applied to 15 plant genotypes representing a broad mixture of Angiosperm species, which found between 628-828 sequences to be shared among chloroplast proteomes of all species, and semi-conserved or species-specific plastid-targeted proteins were between six to 25 times more abundant (Christian et al., 2019b, unpublished). Additionally, almost 1,000 gene loci in the Arabidopsis pan-genome have differential use of chloroplast transit peptides, and the same is true for nearly 9,000 gene families in the *Brachypodium distachyon* pan-genome (Christian et al., 2019a, unpublished). Relatively few proteins are chloroplast-localized in all species, and most plastid-targeted proteins are likely to taxa-specific or non-essential. However, not much is known about the function of these non-essential chloroplast genes, when they are expressed, or what plastid morphotype they accumulate in. Although this work is currently only predictive, the potential impact of non-essential plastid-targeted proteins merits further investigation to determine how much of a role they play in species-specific manner.

6. Conclusions and Future Directions

The ongoing research on chloroplast transit peptides and components of the chloroplast import apparatus are drawing closer to describing a comprehensive theory of protein translocation. However, significant work remains in understanding the mechanism and core composition of the TIC complex, as well as uncovering the specifics of the many stromal

chaperones during translocation. A universal model would not only further the basic understanding of this crucial process in plants but would also facilitate more accurate modeling of the transit peptides, and therefore would increase the power and accuracy of bioinformatics tools. As the pace of high-throughput sequencing continues to accelerate in the characterization of new genomes and pan-genomes, such workflows would enable the more rapid assessment and annotation of potentially novel protein functions. The central role of the plastid makes it an ideal search ground for traits involved in alteration of photosynthetic efficiency, enhancement of nutrition and stress tolerance, and biosynthesis of novel bioactive compounds.

Authors' contributions

RC and AD prepared the manuscript, and all authors read and approved the manuscript. The authors declare no conflict of interest.

Acknowledgments

Work in the Dhingra lab was supported by Washington State University Agriculture Center Research Hatch Grant WNP00011 to AD. RC acknowledges the support received from the National Institutes of Health/National Institute of General Medical Sciences through an institutional training grant award T32-GM008336. The contents of this work are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS or NIH.

References

- Agne, B., Andres, C., Montandon, C., Christ, B., Ertan, A., Jung, F., Infanger, S., Bischof, S., Baginsky, S., Kessler, F., 2010. The Acidic A-Domain of Arabidopsis Toc159 Occurs as a Hyperphosphorylated Protein. *Plant Physiol.* 153, 1016–1030.
<https://doi.org/10.1104/pp.110.158048>
- Agne, B., Infanger, S., Wang, F., Hofstetter, V., Rahim, G., Martin, M., Lee, D.W., Hwang, I., Schnell, D., Kessler, F., 2009. A Toc159 import receptor mutant, defective in hydrolysis of GTP, supports preprotein import into chloroplasts. *J. Biol. Chem.* 284, 8670–8679.
<https://doi.org/10.1074/jbc.M804235200>
- Agne, B., Kessler, F., 2010. Modifications at the A-domain of the chloroplast import receptor Toc159. *Plant Signal. Behav.* 5. <https://doi.org/10.4161/psb.5.11.13707>
- Ajjawi, I., Lu, Y., Savage, L.J., Bell, S.M., Last, R.L., 2010. Large-Scale Reverse Genetics in Arabidopsis: Case Studies from the Chloroplast 2010 Project. *Plant Physiol.* 152, 529–540.
<https://doi.org/10.1104/pp.109.148494>
- Akita, M., Nielsen, E., Keegstra, K., 1997. Identification of protein transport complexes in the chloroplastic envelope membranes via chemical cross-linking. *J. Cell Biol.* 136, 983–994.
<https://doi.org/10.1083/jcb.136.5.983>
- Aldridge, C., Cain, P., Robinson, C., 2009. Protein transport in organelles: Protein transport into and across the thylakoid membrane. *FEBS J.* 276, 1177–1186.
<https://doi.org/10.1111/j.1742-4658.2009.06875.x>
- Aldridge, C., Spence, E., Kirkilionis, M.A., Frigerio, L., Robinson, C., 2008. Tat-dependent targeting of Rieske iron-sulphur proteins to both the plasma and thylakoid membranes in the cyanobacterium *Synechocystis* PCC6803. *Mol. Microbiol.* 70, 140–150.

<https://doi.org/10.1111/j.1365-2958.2008.06401.x>

Andon, N.L., Hollingworth, S., Koller, A., Greenland, A.J., Yates, J.R.I., Haynes, P.A., 2002.

Proteomic characterization of wheat amyloplasts using identification of proteins by tandem mass spectrometry. *Proteomics* 2, 1156–1168.

Andrès, C., Agne, B., Kessler, F., 2010. The TOC complex: Preprotein gateway to the chloroplast. *Biochim. Biophys. Acta - Mol. Cell Res.* 1803, 715–723.

<https://doi.org/10.1016/j.bbamcr.2010.03.004>

Apel, W., Schulze, W.X., Bock, R., 2010. Identification of protein stability determinants in chloroplasts. *Plant J.* 63, 636–650. <https://doi.org/10.1111/j.1365-313X.2010.04268.x>

Armbruster, U., Hertle, A., Makarenko, E., Zühlke, J., Pribil, M., Dietzmann, A., Schliebner, I., Aseeva, E., Fenino, E., Scharfenberg, M., Voigt, C., Leister, D., 2009. Chloroplast proteins without cleavable transit peptides: Rare exceptions or a major constituent of the chloroplast proteome? *Mol. Plant* 2, 1325–1335. <https://doi.org/10.1093/mp/ssp082>

Aronsson, H., Boij, P., Patel, R., Wardle, A., Töpel, M., Jarvis, P., 2007. Toc64/OEP64 is not essential for the efficient import of proteins into chloroplasts in *Arabidopsis thaliana*. *Plant J.* 52, 53–68. <https://doi.org/10.1111/j.1365-313X.2007.03207.x>

Aronsson, H., Combe, J., Patel, R., Jarvis, P., 2006. In vivo assessment of the significance of phosphorylation of the *Arabidopsis* chloroplast protein import receptor, atToc33. *FEBS Lett.* 580, 649–655. <https://doi.org/10.1016/j.febslet.2005.12.055>

Ayliffe, M.A., Scott, N.S., Timmis, J.N., 1998. Analysis of plastid DNA-like sequences within the nuclear genomes of higher plants. *Mol. Biol. Evol.* 15, 738–745.

<https://doi.org/10.1093/oxfordjournals.molbev.a025977>

Bachmair, A., Finley, D., Varshavsky, A., 1986. In vivo half-life of a protein is a function of its

- amino-terminal residue. *Science* (80-.). 234, 179–186.
<https://doi.org/10.1126/science.3018930>
- Bae, W., Lee, Y.J., Kim, D.H., Lee, J., Kim, S., Sohn, E.J., Hwang, I., 2008. AKR2A-mediated import of chloroplast outer membrane proteins is essential for chloroplast biogenesis. *Nat. Cell Biol.* 10, 220–227. <https://doi.org/10.1038/ncb1683>
- Baginsky, S., Siddique, A., Gruissem, W., 2004. Proteome analysis of tobacco bright yellow-2 (BY-2) cell culture plastids as a model for undifferentiated heterotrophic plastids. *J. Proteome Res.* 3, 1128–1137. <https://doi.org/10.1021/pr0499186>
- Baldwin, A., Wardle, A., Patel, R., Dudley, P., Park, S.K., Twell, D., Inoue, K., Jarvis, P., Le, L., B, U.K.A., 2005. A Molecular-Genetic Study of the Arabidopsis Toc75 Gene Family. *Plant Physiol.* 138, 715–733. <https://doi.org/10.1104/pp.105.063289.1>
- Balmer, Y., Vensel, W.H., DuPont, F.M., Buchanan, B.B., Hurkman, W.J., 2006. Proteome of amyloplasts isolated from developing wheat endosperm presents evidence of broad metabolic capability. *J. Exp. Bot.* 57, 1591–1602. <https://doi.org/10.1093/jxb/erj156>
- Balsera, M., Goetze, T.A., Kovács-Bogdán, E., Schürmann, P., Wagner, R., Buchanan, B.B., Soll, J., Bölder, B., 2009. Characterization of Tic110, a channel-forming protein at the inner envelope membrane of chloroplasts, unveils a response to Ca²⁺ and a stromal regulatory disulfide bridge. *J. Biol. Chem.* 284, 2603–2616. <https://doi.org/10.1074/jbc.M807134200>
- Balsera, M., Soll, J., Buchanan, B.B., 2010. Redox extends its regulatory reach to chloroplast protein import. *Trends Plant Sci.* 15, 515–521. <https://doi.org/10.1016/j.tplants.2010.06.002>
- Barbieri, J.T., Riese, M.J., Aktories, K., 2002. Bacterial Toxins That Modify The Actin Cytoskeleton. *Annu. Rev. Cell Dev. Biol.* 18, 315–344.
<https://doi.org/10.1146/annurev.cellbio.18.012502.134748>

- Barsan, C., Sanchez-Bel, P., Rombaldi, C., Egea, I., Rossignol, M., Kuntz, M., Zouine, M., Latché, A., Bouzayen, M., Pech, J.C., 2010. Characteristics of the tomato chromoplast revealed by proteomic analysis. *J. Exp. Bot.* 61, 2413–2431.
<https://doi.org/10.1093/jxb/erq070>
- Barsan, C., Zouine, M., Maza, E., Bian, W., Egea, I., Rossignol, M., Bouyssie, D., Pichereaux, C., Purgatto, E., Bouzayen, M., Latche, A., Pech, J.-C., 2012. Proteomic Analysis of Chloroplast-to-Chromoplast Transition in Tomato Reveals Metabolic Shifts Coupled with Disrupted Thylakoid Biogenesis Machinery and Elevated Energy-Production Components. *Plant Physiol.* 160, 708–725. <https://doi.org/10.1104/pp.112.203679>
- Bartsch, S., Monnet, J., Selbach, K., Quigley, F., Gray, J., von Wettstein, D., Reinbothe, S., Reinbothe, C., 2008. Three thioredoxin targets in the inner envelope membrane of chloroplasts function in protein import and chlorophyll metabolism. *Proc. Natl. Acad. Sci.* 105, 4933–4938. <https://doi.org/10.1073/pnas.0800378105>
- Bauer, J., Chen, K., Hiltbunner, A., Wehrli, E., Eugster, M., Schnell, D., Kessler, F., 2000. The major protein import receptor of plastids is essential for chloroplast biogenesis. *Nature* 403, 203–207. <https://doi.org/10.1038/35003214>
- Bauer, J., Hiltbrunner, A., Weibel, P., Vidi, P.A., Alvarez-Huerta, M., Smith, M.D., Schnell, D.J., Kessler, F., 2002. Essential role of the G-domain in targeting of the protein import receptor atToc159 to the chloroplast outer membrane. *J. Cell Biol.* 159, 845–854.
<https://doi.org/10.1083/jcb.200208018>
- Becker, T., Hritz, J., Vogel, M., Caliebe, A., Bukau, B., Soll, J., Schleiff, E., 2004. Toc12, a Novel Subunit of the Intermembrane Space Preprotein Translocon of Chloroplasts. *Mol. Biol. Cell* 15, 5130–5144. <https://doi.org/10.1091/mbc.E04>

- Bédard, J., Jarvis, P., 2005. Recognition and envelope translocation of chloroplast preproteins. *J. Exp. Bot.* 56, 2287–2320. <https://doi.org/10.1093/jxb/eri243>
- Bédard, J., Kubis, S., Bimanadham, S., Jarvis, P., 2007. Functional similarity between the chloroplast translocon component, Tic40, and the human co-chaperone, Hsp70-interacting protein (Hip). *J. Biol. Chem.* 282, 21404–21414. <https://doi.org/10.1074/jbc.M611545200>
- Bhattacharya, D., Archibald, J.M., Weber, A.P.M., Reyes-Prieto, A., 2007. How do endosymbionts become organelles? Understanding early events in plastid evolution. *BioEssays* 29, 1239–1246. <https://doi.org/10.1002/bies.20671>
- Bienvenut, W. V., Sumpton, D., Martinez, A., Lilla, S., Espagne, C., Meinel, T., Giglione, C., 2012. Comparative Large Scale Characterization of Plant *versus* Mammal Proteins Reveals Similar and Idiosyncratic *N*- α -Acetylation Features. *Mol. Cell. Proteomics* 11, M111.015131. <https://doi.org/10.1074/mcp.M111.015131>
- Bionda, T., Tillmann, B., Simm, S., Beilstein, K., Ruprecht, M., Schleiff, E., 2010. Chloroplast import signals: The length requirement for translocation in vitro and in vivo. *J. Mol. Biol.* 402, 510–523. <https://doi.org/10.1016/j.jmb.2010.07.052>
- Block, M.A., Dorne, A., Joyard, J., Douce, R., 1983. Preparation and Characterization of Membrane Fractions Enriched in Outer and Inner Envelope Membranes from Spinach Chloroplasts a procedure which includes followed by several. *J. Biol. Chem.* 258, 13281–13286.
- Block, M.A., Douce, R., Joyard, J., Rolland, N., 2007. Chloroplast envelope membranes: A dynamic interface between plastids and the cytosol. *Photosynth. Res.* 92, 225–244. <https://doi.org/10.1007/s11120-007-9195-8>
- Bodył, A., Mackiewicz, P., Stiller, J.W., 2009. Early steps in plastid evolution: Current ideas and

- controversies. *BioEssays* 31, 1219–1232. <https://doi.org/10.1002/bies.200900073>
- Bogsch, E., Brink, S., Robinson, C., 1997. Pathway specificity for a Δ pH-dependent precursor thylakoid lumen protein is governed by a “Sec-avoidance” motif in the transfer peptide and a “Sec-incompatible” mature protein. *EMBO J.* 16, 3851–3859.
<https://doi.org/10.1093/emboj/16.13.3851>
- Boij, P., Patel, R., Garcia, C., Jarvis, P., Aronsson, H., 2009. In vivo studies on the roles of Tic55-related proteins in chloroplast protein import in *Arabidopsis thaliana*. *Mol. Plant* 2, 1397–1409. <https://doi.org/10.1093/mp/ssp079>
- Bölter, B., May, T., Soll, J., 1998. A protein import receptor in pea chloroplasts, Toc86, is only a proteolytic fragment of a larger polypeptide. *FEBS Lett.* 441, 59–62.
[https://doi.org/10.1016/S0014-5793\(98\)01525-7](https://doi.org/10.1016/S0014-5793(98)01525-7)
- Bölter, B., Soll, J., 2017. Ycf1/Tic214 Is Not Essential for the Accumulation of Plastid Proteins. *Mol. Plant* 10, 219–221. <https://doi.org/10.1016/j.molp.2016.10.012>
- Bölter, B., Soll, J., 2016. Once upon a Time – Chloroplast Protein Import Research from Infancy to Future Challenges. *Mol. Plant* 9, 798–812. <https://doi.org/10.1016/j.molp.2016.04.014>
- Braun, N.A., Davis, A.W., Theg, S.M., 2007. The chloroplast tat pathway utilizes the transmembrane electric potential as an energy source. *Biophys. J.* 93, 1993–1998.
<https://doi.org/10.1529/biophysj.106.098731>
- Bräutigam, A., Weber, A.P.M., 2009. Proteomic analysis of the proplastid envelope membrane provides novel insights into small molecule and protein transport across proplastid membranes. *Mol. Plant* 2, 1247–1261. <https://doi.org/10.1093/mp/ssp070>
- Bräutigam, K., Dietzel, L., Pfannschmidt, T., 2007. Plastid-nucleus communication: Anterograde and retrograde signalling in the development and function of plastids, in: Bock, R. (Ed.),

- Topics In Current Genetics. Springer, Berlin, Heidelberg. <https://doi.org/10.1007/4735>
- Brillouet, J.M., Romieu, C., Schoefs, B., Solymosi, K., Cheynier, V., Fulcrand, H., Verdeil, J.L., Conéjéro, G., 2013. The tannosome is an organelle forming condensed tannins in the chlorophyllous organs of Tracheophyta. *Ann. Bot.* 112, 1003–1014. <https://doi.org/10.1093/aob/mct168>
- Brillouet, J.M., Verdeil, J.L., Odoux, E., Lartaud, M., Grisoni, M., Conéjéro, G., 2014. Phenol homeostasis is ensured in vanilla fruit by storage under solid form in a new chloroplast-derived organelle, the phenyloplast. *J. Exp. Bot.* 65, 2427–2435. <https://doi.org/10.1093/jxb/eru126>
- Brink, S., Fischer, K., Klosgen, R.B., Flugge, U.I., 1995. Sorting of nuclear-encoded chloroplast membrane proteins to the envelope and the thylakoid membrane. *J. Biol. Chem.* <https://doi.org/10.1074/jbc.270.35.20808>
- Bruce, B.D., 2001. The paradox of plastid transit peptides: Conservation of function despite divergence in primary structure. *Biochim. Biophys. Acta - Mol. Cell Res.* 1541, 2–21. [https://doi.org/10.1016/S0167-4889\(01\)00149-5](https://doi.org/10.1016/S0167-4889(01)00149-5)
- Bruce, B.D., 2000. Chloroplast transit peptides: Structure, function and evolution. *Trends Cell Biol.* 10, 440–447. [https://doi.org/10.1016/S0962-8924\(00\)01833-X](https://doi.org/10.1016/S0962-8924(00)01833-X)
- Bruce, B.D., 1998. The role of lipids in plastid protein transport. *Plant Mol. Biol.* 38, 223–246. <https://doi.org/10.1023/A>
- Bruch, E.M., Rosano, G.L., Ceccarelli, E.A., 2012. Chloroplastic Hsp100 chaperones ClpC2 and ClpD interact in vitro with a transit peptide only when it is located at the N-terminus of a protein. *BMC Plant Biol.* 12, 1–6. <https://doi.org/10.1186/1471-2229-12-57>
- Caliebe, A., Grimm, R., Kaiser, G., Lübeck, J., Soll, J., Heins, L., 1997. The chloroplastic

- protein import machinery contains a Rieske-type iron-sulfur cluster and a mononuclear iron-binding protein. *EMBO J.* 16, 7342–7350. <https://doi.org/10.1093/emboj/16.24.7342>
- Campbell, J.H., Hoang, T., Jelokhani-Niaraki, M., Smith, M.D., 2014. Folding and self-association of atTic20 in lipid membranes: Implications for understanding protein transport across the inner envelope membrane of chloroplasts. *BMC Biochem.* 15, 1–13. <https://doi.org/10.1186/s12858-014-0029-y>
- Cavalier-Smith, T., 1987. The Simultaneous Symbiotic Origin of Mitochondria, Chloroplasts, and Microbodies. *Ann. N. Y. Acad. Sci.* 503, 55–71. <https://doi.org/10.1111/j.1749-6632.1987.tb40597.x>
- Celedon, J.M., Cline, K., 2013. Intra-plastid protein trafficking: How plant cells adapted prokaryotic mechanisms to the eukaryotic condition. *Biochim. Biophys. Acta - Mol. Cell Res.* 1833, 341–351. <https://doi.org/10.1016/j.bbamcr.2012.06.028>
- Chang, J.-S., Chen, L.-J., Yeh, Y.-H., Hsiao, C.-D., Li, H., 2017. Chloroplast Preproteins Bind to the Dimer Interface of the Toc159 Receptor during Import. *Plant Physiol.* 173, 2148–2162. <https://doi.org/10.1104/pp.16.01952>
- Chen, K., Chen, X., Schnell, D.J., 2000. Initial binding of preproteins involving the Toc159 receptor can be bypassed during protein import into chloroplasts. *Plant Physiol.* 122, 813–22. <https://doi.org/10.1104/pp.122.3.813>
- Chen, L.J., Li, H.M., 2017. Stable megadalton TOC–TIC supercomplexes as major mediators of protein import into chloroplasts. *Plant J.* 92, 178–188. <https://doi.org/10.1111/tpj.13643>
- Chen, X., Smith, M.D., Fitzpatrick, L., Schnell, D.J., 2002. In vivo analysis of the role of atTic20 in protein import into chloroplasts. *Plant Cell* 14, 641–654. <https://doi.org/10.1105/tpc.010336>

- Chen, Y.-L., Chen, L.-J., Li, H., 2016. Polypeptide Transport-Associated Domains of the Toc75 Channel Protein Are Located in the Intermembrane Space of Chloroplasts. *Plant Physiol.* 172, 235–243. <https://doi.org/10.1104/pp.16.00919>
- Chigri, F., Hormann, F., Stamp, A., Stammers, D.K., Bolter, B., Soll, J., Vothknecht, U.C., 2006. Calcium regulation of chloroplast protein translocation is mediated by calmodulin binding to Tic32. *Proc. Natl. Acad. Sci.* 103, 16051–16056. <https://doi.org/10.1073/pnas.0607150103>
- Chigri, F., Soll, J., Vothknecht, U.C., 2005. Calcium regulation of chloroplast protein import. *Plant J.* 42, 821–831. <https://doi.org/10.1111/j.1365-313X.2005.02414.x>
- Chiu, C.-C., Chen, L.-J., Li, H. -m., 2010. Pea Chloroplast DnaJ-J8 and Toc12 Are Encoded by the Same Gene and Localized in the Stroma. *Plant Physiol.* 154, 1172–1182. <https://doi.org/10.1104/pp.110.161224>
- Chotewutmontri, P., Bruce, B.D., 2015. Non-native, N-terminal Hsp70 Molecular Motor-recognition Elements in Transit Peptides Support Plastid Protein Translocation. *J. Biol. Chem.* 290, 7602–7621. <https://doi.org/10.1074/jbc.M114.633586>
- Chotewutmontri, P., Reddick, L.E., McWilliams, D.R., Campbell, I.M., Bruce, B.D., 2012. Differential Transit Peptide Recognition during Preprotein Binding and Translocation into Flowering Plant Plastids. *Plant Cell* 24, 3040–3059. <https://doi.org/10.1105/tpc.112.098327>
- Chou, M., Fitzpatrick, L.M., Tu, S., Budziszewski, G., Potter-lewis, S., Akita, M., Levin, J.Z., Keegstra, K., Li, H., 2003. Tic40, a membrane-anchored co-chaperone homolog in the chloroplast protein translocon. *EMBO J.* 22, 2970–2980.
- Chou, M.L., Chu, C.C., Chen, L.J., Akita, M., Li, H.M., 2006. Stimulation of transit-peptide release and ATP hydrolysis by a cochaperone during protein import into chloroplasts. *J.*

- Cell Biol. 175, 893–900. <https://doi.org/10.1083/jcb.200609172>
- Chou, M.L., Liao, W.Y., Wei, W.C., Li, A.Y.S., Chu, C.Y., Wu, C.L., Liu, C.L., Fu, T.H., Lin, L.F., 2018. The direct involvement of dark-induced Tic55 protein in chlorophyll catabolism and its indirect role in the MYB108-NAC signaling pathway during leaf senescence in *Arabidopsis thaliana*. *Int. J. Mol. Sci.* 19. <https://doi.org/10.3390/ijms19071854>
- Christian, R., Hewitt, S., Nelson, G., Roalson, E., Dhingra, A., 2019a. Plastid Transit Peptides - Where Do They Come From and Where Do They All Belong? Assessment of Chloroplast Transit Peptide Evolution in Multi-Species and Pan-Genomic Comparisons.
- Christian, R., Hewitt, S., Roalson, E., Dhingra, A., 2019b. Genome-Scale Characterization of Predicted Plastid-Targeted Proteins in Higher Plants.
- Chua, N.-H., Schmidt, G.W., 1979. Transport of Proteins Into Mitochondria and Chloroplasts 81, 461–483.
- Claros, M.G., Brunak, S., Heijne, G. Von, 1997. Prediction of N-terminal protein sorting signals. *Curr. Opin. Struct. Biol.* 7, 394–398. [https://doi.org/10.1016/S0959-440X\(97\)80057-7](https://doi.org/10.1016/S0959-440X(97)80057-7)
- Cline, K., Henry, R., 1996. Import and routing of nucleus-encoded chloroplast proteins. *Annu. Rev. Cell Dev. Biol.* 12, 1–26. <https://doi.org/10.1146/annurev.cellbio.12.1.1>
- Comai, L., Larson-Kelly, N., Kiser, J., Mau, C.J.D., Pokalsky, A.R., Shewmaker, C.K., McBride, K., Jones, A., Stalker, D.M., 1988. Chloroplast Transport of a Ribulose Bisphosphate Carboxylase Small Subunit-5-Enolpyruvyl 3-Phosphoshikimate Synthase Chimeric Protein Requires Part of the Mature Small Subunit in Addition to the Transit Peptide. *J. Biol. Chem.* 263, 15104–15109.
- Constan, D., Patel, R., Keegstra, K., Jarvis, P., 2004. An outer envelope membrane component of the plastid protein import apparatus plays an essential role in *Arabidopsis*. *Plant J.* 38, 93–

106. <https://doi.org/10.1111/j.1365-313X.2004.02024.x>
- Daher, Z., Recorbet, G., Solymosi, K., Wienkoop, S., Mounier, A., Morandi, D., Lherminier, J., Wipf, D., Dumas-Gaudot, E., Schoefs, B., 2017. Changes in plastid proteome and structure in arbuscular mycorrhizal roots display a nutrient starvation signature. *Physiol. Plant.* 159, 13–29. <https://doi.org/10.1111/ppl.12505>
- Daher, Z., Recorbet, G., Valot, B., Robert, F., Balliau, T., Potin, S., Schoefs, B., Dumas-Gaudot, E., 2010. Proteomic analysis of *Medicago truncatula* root plastids. *Proteomics* 10, 2123–2137. <https://doi.org/10.1002/pmic.200900345>
- Dalbey, R.E., Chen, M., 2004. Sec-translocase mediated membrane protein biogenesis. *Biochim. Biophys. Acta (BBA)-Molecular Cell Res.* 1694, 37–53. <https://doi.org/10.1016/j.bbamcr.2004.03.009>
- de Boer, A.D., Weisbeek, P.J., 1991. Chloroplast protein topogenesis: import, sorting and assembly. *Biochim. Biophys. Acta - Rev. Biomembr.* [https://doi.org/10.1016/0304-4157\(91\)90015-O](https://doi.org/10.1016/0304-4157(91)90015-O)
- de Vries, J., Sousa, F.L., Bölter, B., Soll, J., Gould, S.B., 2015. YCF1: A Green TIC? *Plant Cell* 27, 1827–1833. <https://doi.org/10.1105/tpc.114.135541>
- Demarsy, E., Lakshmanan, A.M., Kessler, F., 2014. Border control: selectivity of chloroplast protein import and regulation at the TOC-complex. *Front. Plant Sci.* 5, 1–10. <https://doi.org/10.3389/fpls.2014.00483>
- Dhanoa, P.K., Richardson, L.G.L., Smith, M.D., Gidda, S.K., Henderson, M.P.A., Andrews, D.W., Mullen, R.T., 2010. Distinct pathways mediate the sorting of tail-anchored proteins to the plastid outer envelope. *PLoS One* 5. <https://doi.org/10.1371/journal.pone.0010098>
- Diaz-Mendoza, M., Velasco-Arroyo, B., Santamaria, M.E., González-Melendi, P., Martinez, M.,

- Diaz, I., 2016. Plant senescence and proteolysis: Two processes with one destiny. *Genet. Mol. Biol.* 39, 329–338. <https://doi.org/10.1590/1678-4685-GMB-2016-0015>
- Dickmanns, A., Kehlenbach, R.H., Fahrenkrog, B., 2015. Nuclear Pore Complexes and Nucleocytoplasmic Transport: From Structure to Function to Disease. *Int. Rev. Cell Mol. Biol.* 320, 171–233. <https://doi.org/10.1016/bs.ircmb.2015.07.010>
- Dougan, D.A., Micevski, D., Truscott, K.N., 2012. The N-end rule pathway: From recognition by N-recognins, to destruction by AAA+proteases. *Biochim. Biophys. Acta - Mol. Cell Res.* 1823, 83–91. <https://doi.org/10.1016/j.bbamcr.2011.07.002>
- Doyle, S.R., Kasinadhuni, N.R.P., Chan, C.K., Grant, W.N., 2013. Evidence of Evolutionary Constraints That Influences the Sequence Composition and Diversity of Mitochondrial Matrix Targeting Signals. *PLoS One* 8, 1–8. <https://doi.org/10.1371/journal.pone.0067938>
- Dupont, F.M., 2008. Metabolic pathways of the wheat (*Triticum aestivum*) endosperm amyloplast revealed by proteomics. *BMC Plant Biol.* 8, 1–18. <https://doi.org/10.1186/1471-2229-8-39>
- Dutta, S., Teresinski, H.J., Smith, M.D., 2014. A split-ubiquitin yeast two-hybrid screen to examine the substrate specificity of atToc159 and atToc132, two arabidopsis chloroplast preprotein import receptors. *PLoS One* 9. <https://doi.org/10.1371/journal.pone.0095026>
- Duy, D., Stube, R., Wanner, G., Philippar, K., 2011. The Chloroplast Permease PIC1 Regulates Plant Growth and Development by Directing Homeostasis and Transport of Iron. *Plant Physiol.* 155, 1709–1722. <https://doi.org/10.1104/pp.110.170233>
- Duy, D., Wanner, G., Meda, A.R., von Wiren, N., Soll, J., Philippar, K., 2007. PIC1, an Ancient Permease in Arabidopsis Chloroplasts, Mediates Iron Transport. *Plant Cell* 19, 986–1006. <https://doi.org/10.1105/tpc.106.047407>

- Eckart, K., Eichacker, L., Sohrt, K., Schleiff, E., Heins, L., Soll, J., 2002. A Toc75-like protein import channel is abundant in chloroplasts. *EMBO Rep.* 3, 557–562.
<https://doi.org/10.1093/embo-reports/kvf110>
- Egea, I., Barsan, C., Bian, W., Purgatto, E., Latché, A., Chervin, C., Bouzayen, M., Pech, J.C., 2010. Chromoplast differentiation: Current status and perspectives. *Plant Cell Physiol.* 51, 1601–1611. <https://doi.org/10.1093/pcp/pcq136>
- Eichacker, L.A., Henry, R., 2001. Function of a chloroplast SRP in thylakoid protein export. *Biochim. Biophys. Acta - Mol. Cell Res.* 1541, 120–134. [https://doi.org/10.1016/S0167-4889\(01\)00151-3](https://doi.org/10.1016/S0167-4889(01)00151-3)
- Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H., 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* 2, 953–71.
<https://doi.org/10.1038/nprot.2007.131>
- Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G., 2000. Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. *J. Mol. Biol.* 300, 1005–1016. <https://doi.org/10.1006/jmbi.2000.3903>
- Ertel, F., Mirus, O., Bredemeier, R., Moslavac, S., Becker, T., Schleiff, E., 2005. The evolutionarily related β -barrel polypeptide transporters from *Pisum sativum* and *Nostoc PCC7120* contain two distinct functional domains. *J. Biol. Chem.* 280, 28281–28289.
<https://doi.org/10.1074/jbc.M503035200>
- Esaki, M., Kanamori, T., Nishikawa, S., Endo, T., 1999. Two distinct mechanisms drive protein translocation across the mitochondrial outer membrane in the late step of the cytochrome b2 import pathway. *Proc. Natl. Acad. Sci.* 96, 11770–11775.
<https://doi.org/10.1073/pnas.96.21.11770>

- Falcón, L.I., Magallón, S., Castillo, A., 2010. Dating the cyanobacterial ancestor of the chloroplast. *ISME J.* 4, 777. <https://doi.org/10.1038/ismej.2010.98>
- Falk, S., Sinning, I., 2010. cpSRP43 is a novel chaperone specific for light-harvesting chlorophyll a,b-binding proteins. *J. Biol. Chem.* 285, 21655–21661. <https://doi.org/10.1074/jbc.C110.132746>
- Ferro, M., Brugière, S., Salvi, D., Seigneurin-Berny, D., Court, M., Moyet, L., Ramus, C., Miras, S., Mellal, M., Le Gall, S., Kieffer-Jaquinod, S., Bruley, C., Garin, J., Joyard, J., Masselon, C., Rolland, N., 2010. AT_CHLORO, a Comprehensive Chloroplast Proteome Database with Subplastidial Localization and Curated Information on Envelope Proteins. *Mol. Cell. Proteomics* 9, 1063–1084. <https://doi.org/10.1074/mcp.M900325-MCP200>
- Fischer, K., Weber, A., Brink, S., Arbinger, B., Schunemann, D., Borchert, S., Heldt, H.W., Popp, B., Benz, R., Link, T.A., Eckerskorn, C., Flugge, U.I., 1994. Porins from plants. Molecular cloning and functional characterization of two new members of the porin family. *J. Biol. Chem.* 269, 25754–25760.
- Flores-Pérez, Ú., Bédard, J., Tanabe, N., Lymperopoulos, P., Clarke, A.K., Jarvis, P., 2016. Functional Analysis of the Hsp93/ClpC Chaperone at the Chloroplast Envelope. *Plant Physiol.* 170, 147–162. <https://doi.org/10.1104/pp.15.01538>
- Frain, K.M., Gangl, D., Jones, A., Zedler, J.A.Z., Robinson, C., 2016. Protein translocation and thylakoid biogenesis in cyanobacteria. *Biochim. Biophys. Acta - Bioenerg.* 1857, 266–273. <https://doi.org/10.1016/j.bbabi.2015.08.010>
- Friedman, A.L., Keegstra, K., 1989. Chloroplast protein import : quantitative analysis of precursor binding. *Plant Physiol.* 89, 993–9. <https://doi.org/10.1104/pp.89.3.993>
- Froehlich, J.E., Keegstra, K., 2011. The role of the transmembrane domain in determining the

- targeting of membrane proteins to either the inner envelope or thylakoid membrane. *Plant J.* 68, 844–856. <https://doi.org/10.1111/j.1365-313X.2011.04735.x>
- Fulgosi, H., Soll, J., 2002. The chloroplast protein import receptors Toc34 and Toc159 are phosphorylated by distinct protein kinases. *J. Biol. Chem.* 277, 8934–8940. <https://doi.org/10.1074/jbc.M110679200>
- Galpaz, N., Wang, Q., Menda, N., Zamir, D., Hirschberg, J., 2008. Abscisic acid deficiency in the tomato mutant high-pigment 3 leading to increased plastid number and higher fruit lycopene content. *Plant J.* 53, 717–730. <https://doi.org/10.1111/j.1365-313X.2007.03362.x>
- Ganesan, I., Shi, L.-X., Labs, M., Theg, S.M., 2018. Evaluating the Functional Pore Size of Chloroplast TOC and TIC Protein Translocons: Import of Folded Proteins. *Plant Cell tpc.00427.2018.* <https://doi.org/10.1105/tpc.18.00427>
- Garg, S.G., Gould, S.B., 2016. The Role of Charge in Protein Targeting Evolution. *Trends Cell Biol.* 26, 894–905. <https://doi.org/10.1016/j.tcb.2016.07.001>
- Gentle, I., Gabriel, K., Beech, P., Waller, R., Lithgow, T., 2004. The Omp85 family of proteins is essential for outer membrane biogenesis in mitochondria and bacteria. *J. Cell Biol.* 164, 19–24. <https://doi.org/10.1083/jcb.200310092>
- Gibbs, D.J., Bacardit, J., Bachmair, A., Holdsworth, M.J., 2014. The eukaryotic N-end rule pathway: conserved mechanisms and diverse functions. *Trends Cell Biol.* 24, 603–611.
- Glaser, E., Nilsson, S., Bhushan, S., 2006. Two novel mitochondrial and chloroplastic targeting-peptide-degrading peptidasomes in *A. thaliana*, AtPreP1 and AtPreP2. *Biol. Chem.* 387, 1441–1447. <https://doi.org/10.1515/BC.2006.180>
- Glaser, S., Van Dooren, G.G., Agrawal, S., Brooks, C.F., McFadden, G.I., Striepen, B., Higgins, M.K., 2012. Tic22 is an essential chaperone required for protein import into the apicoplast.

- J. Biol. Chem. 287, 39505–39512. <https://doi.org/10.1074/jbc.M112.405100>
- Goforth, R.L., Peterson, E.C., Yuan, J., Moore, M.J., Kight, A.D., Lohse, M.B., Sakon, J., Henry, R.L., 2004. Regulation of the GTPase cycle in post-translational signal recognition particle-based protein targeting involves cpSRP43. J. Biol. Chem. 279, 43077–43084. <https://doi.org/10.1074/jbc.M401600200>
- Gohlke, U., Pullan, L., Mcdevitt, C.A., Porcelli, I., Leeuw, E. De, Palmer, T., Saibil, H.R., Berks, B.C., 2005. The TatA component of the twin-arginine protein transport system forms channel complexes of variable diameter. Proc. Natl. Acad. Sci. 102, 10482–10486.
- Gong, X., Guo, C., Terachi, T., Cai, H., Yu, D., 2015. Tobacco PIC1 Mediates Iron Transport and Regulates Chloroplast Development. Plant Mol. Biol. Report. 33, 401–413. <https://doi.org/10.1007/s11105-014-0758-5>
- Gray, A.N., Henderson-Frost, J.M., 2011. Unbalanced Charge Distribution as a Determinant for Dependence of a Subset of Escherischi coli Membrane Proteins on the Membrane Insertase YidC. MBio 2, e00238-11. <https://doi.org/10.1128/mBio.00238-11>.Editor
- Green, B.R., 2011. Chloroplast genomes of photosynthetic eukaryotes. Plant J. 66, 34–44. <https://doi.org/10.1111/j.1365-313X.2011.04541.x>
- Gutensohn, M., Pahnke, S., Kolukisaoglu, Ü., Schulz, B., Schierhorn, A., Voigt, A., Hust, B., Rollwitz, I., Stöckel, J., Geimer, S., Albrecht, V., Flügge, U.I., Klösgen, R.B., 2004. Characterization of a T-DNA insertion mutant for the protein import receptor atToc33 from chloroplasts. Mol. Genet. Genomics 272, 379–396. <https://doi.org/10.1007/s00438-004-1068-7>
- Gutensohn, M., Schulz, B., Nicolay, P., Flügge, U.I., 2000. Functional analysis of the two Arabidopsis homologues of Toc34, a component of the chloroplast protein import

- apparatus. *Plant J.* 23, 771–783. <https://doi.org/10.1046/j.1365-313X.2000.00849.x>
- Hartl, F.U., Schmidt, B., Wachter, E., Weiss, H., Neupert, W., 1986. Transport into mitochondria and intramitochondrial sorting of the Fe/S protein of ubiquinol-cytochrome c reductase. *Cell* 47, 939–951. [https://doi.org/10.1016/0092-8674\(86\)90809-3](https://doi.org/10.1016/0092-8674(86)90809-3)
- Hauenstein, M., Christ, B., Das, A., Aubry, S., Hörtensteiner, S., 2016. A Role for TIC55 as a Hydroxylase of Phyllobilins, the Products of Chlorophyll Breakdown during Plant Senescence. *Plant Cell* 28, 2510–2527. <https://doi.org/10.1105/tpc.16.00630>
- Heazlewood, J.L., 2005. Combining Experimental and Predicted Datasets for Determination of the Subcellular Location of Proteins in Arabidopsis. *Plant Physiol.* 139, 598–609. <https://doi.org/10.1104/pp.105.065532>
- Heazlewood, J.L., Verboom, R.E., Tonti-Filippini, J., Small, I., Millar, A.H., 2007. SUBA: The Arabidopsis subcellular database. *Nucleic Acids Res.* 35, 213–218. <https://doi.org/10.1093/nar/gkl863>
- Heins, L., Mehrle, A., Hemmler, R., Wagner, R., Kuchler, M., Hörmann, F., Sveshnikov, D., Soll, J., 2002. The preprotein conducting channel at the inner envelope membrane of plastids. *EMBO J.* 21, 2616–2625. <https://doi.org/10.1093/emboj/21.11.2616>
- Hennon, S.W., Soman, R., Zhu, L., Dalbey, R.E., 2015. YidC/Alb3/Oxa1 Family of Insertases. *J. Biol. Chem.* jbc.R115.638171. <https://doi.org/10.1074/jbc.R115.638171>
- Hiltbrunner, A., Bauer, J., Alvarez-Huerta, M., Kessler, F., 2001a. Protein translocation at the Arabidopsis outer chloroplast membrane. *Biochem. Cell Biol.* 79, 629–635.
- Hiltbrunner, A., Bauer, J., Vidi, P.A., Infanger, S., Weibel, P., Hohwy, M., Kessler, F., 2001b. Targeting of an abundant cytosolic form of the protein import receptor at Toc159 to the outer chloroplast membrane. *J. Cell Biol.* 154, 309–316.

<https://doi.org/10.1083/jcb.200104022>

- Hinnah, S.C., Hill, K., Wagner, R., Schlicher, T., Soll, J., 1997. Reconstitution of a chloroplast protein import channel. *EMBO J.* 16, 7351–7360. <https://doi.org/10.1093/emboj/16.24.7351>
- Hinnah, S.C., Wagner, R., Sveshnikova, N., Harrer, R., Soll, J., 2002. The chloroplast protein import channel Toc75: Pore properties and interaction with transit peptides. *Biophys. J.* 83, 899–911. [https://doi.org/10.1016/S0006-3495\(02\)75216-8](https://doi.org/10.1016/S0006-3495(02)75216-8)
- Hirabayashi, Y., Kikuchi, S., Oishi, M., Nakai, M., 2011. In vivo studies on the roles of two closely related arabidopsis Tic20 proteins, AtTic20-I and AtTic20-IV. *Plant Cell Physiol.* 52, 469–478. <https://doi.org/10.1093/pcp/pcr010>
- Hirohashi, T., Nakai, M., 2000. Molecular cloning and characterization of maize Toc34, a regulatory component of the protein import machinery of chloroplast. *Biochim. Biophys. Acta - Gene Struct. Expr.* 1491, 309–314. [https://doi.org/10.1016/S0167-4781\(00\)00043-9](https://doi.org/10.1016/S0167-4781(00)00043-9)
- Hofmann, N.R., Theg, S.M., 2005. Protein- and energy-mediated targeting of chloroplast outer envelope membrane proteins. *Plant J.* 44, 917–927. <https://doi.org/10.1111/j.1365-313X.2005.02571.x>
- Hofmann, N.R., Theg, S.M., 2003. *Physcomitrella patens* as a model for the study of chloroplast protein transport: Conserved machineries between vascular and non-vascular plants. *Plant Mol. Biol.* 53, 621–632. <https://doi.org/10.1023/B:PLAN.0000019109.01740.c6>
- Holbrook, K., Subramanian, C., Chotewutmontri, P., Reddick, L.E., Wright, S., Zhang, H., Moncrief, L., Bruce, B.D., 2016. Functional Analysis of Semi-conserved Transit Peptide Motifs and Mechanistic Implications in Precursor Targeting and Recognition. *Mol. Plant* 9, 1286–1301. <https://doi.org/10.1016/j.molp.2016.06.004>
- Hooper, C.M., Castleden, I.R., Tanz, S.K., Aryamanesh, N., Millar, A.H., 2017. SUBA4: The

- interactive data analysis centre for Arabidopsis subcellular protein locations. *Nucleic Acids Res.* 45, D1064–D1074. <https://doi.org/10.1093/nar/gkw1041>
- Hörmann, F., Kuchler, M., Sveshnikov, D., Oppermann, U., Li, Y., Soll, J., 2004. Tic32, an essential component in chloroplast biogenesis. *J. Biol. Chem.* 279, 34756–34762. <https://doi.org/10.1074/jbc.M402817200>
- Huang, C.Y., Ayliffe, M.A., Timmis, J.N., 2003. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. *Nature* 422, 72–76. <https://doi.org/10.1038/nature01435>
- Huang, P.-K., Chan, P.-T., Su, P.-H., Chen, L.-J., Li, H., 2016. Chloroplast Hsp93 Directly Binds to Transit Peptides at an Early Stage of the Preprotein Import Process. *Plant Physiol.* 170, 857–866. <https://doi.org/10.1104/pp.15.01830>
- Huang, W., Ling, Q., Bédard, J., Lilley, K., Jarvis, P., 2011. In Vivo Analyses of the Roles of Essential Omp85-Related Proteins in the Chloroplast Outer Envelope Membrane. *Plant Physiol.* 157, 147–159. <https://doi.org/10.1104/pp.111.181891>
- Hynds, P.J., Robinson, D., Robinson, C., 1998. The Sec-independent twin-arginine translocation system can transport both tightly folded and malformed proteins across the thylakoid membrane. *J. Biol. Chem.* 273, 34868–34874. <https://doi.org/10.1074/jbc.273.52.34868>
- Inaba, T., Alvarez-Huerta, M., Li, M., Bauer, J., Ewers, C., Kessler, F., Schnell, D.J., 2005. Arabidopsis tic110 is essential for the assembly and function of the protein import machinery of plastids. *Plant Cell* 17, 1482–96. <https://doi.org/10.1105/tpc.105.030700>
- Inaba, T., Li, M., Alvarez-Huerta, M., Kessler, F., Schnell, D.J., 2003. atTic110 functions as a scaffold for coordinating the stromal events of protein import into chloroplasts. *J. Biol. Chem.* 278, 34617–38627.
- Inaba, T., Schnell, D.J., 2008. Protein trafficking to plastids: one theme, many variations.

- Biochem. J. 413, 15–28. <https://doi.org/10.1042/BJ20080490>
- Ingle, R.A., Collett, H., Cooper, K., Takahashi, Y., Farrant, J.M., Illing, N., 2008. Chloroplast biogenesis during rehydration of the resurrection plant *Xerophyta humilis*: Parallels to the etioplast-chloroplast transition. *Plant, Cell Environ.* 31, 1813–1824.
<https://doi.org/10.1111/j.1365-3040.2008.01887.x>
- Inoue, H., Li, M., Schnell, D.J., 2013. An essential role for chloroplast heat shock protein 90 (Hsp90C) in protein import into chloroplasts. *Proc. Natl. Acad. Sci.* 110, 3173–3178.
<https://doi.org/10.1073/pnas.1219229110>
- Inoue, H., Rounds, C., Schnell, D.J., 2010. The Molecular Basis for Distinct Pathways for Protein Import into *Arabidopsis* Chloroplasts. *Plant Cell* 22, 1947–1960.
<https://doi.org/10.1105/tpc.110.074328>
- Inoue, K., Baldwin, A.J., Shipman, R.L., Matsui, K., Theg, S.M., Ohme-Takagi, M., 2005. Complete maturation of the plastid protein translocation channel requires a type I signal peptidase. *J. Cell Biol.* 171, 425–430. <https://doi.org/10.1083/jcb.200506171>
- Inoue, K., Demel, R., De Kruijff, B., Keegstra, K., 2001. The N-terminal portion of the preToc75 transit peptide interacts with membrane lipids and inhibits binding and import of precursor proteins into isolated chloroplasts. *Eur. J. Biochem.* 268, 4036–4043.
<https://doi.org/10.1046/j.1432-1327.2001.02316.x>
- Inoue, K., Potter, D., 2004. The chloroplastic protein translocation channel Toc75 and its paralog OEP80 represent two distinct protein families and are targeted to the chloroplastic outer envelope by different mechanisms. *Plant J.* 39, 354–365. <https://doi.org/10.1111/j.1365-313X.2004.02135.x>
- Ivanova, Y., Smith, M.D., Chen, K., Schnell, D.J., 2004. Members of the Toc159 Import

- Receptor Family Represent Distinct Pathways for Protein Targeting to Plastids. *Mol. Biol. Cell* 15, 3379–3392. <https://doi.org/10.1091/mbc.E03>
- Ivey III, R.A., Subramanian, C., Bruce, B.D., 2000. Identification of a Hsp70 Recognition Domain within the Rubisco Small Subunit Transit Peptide. *Plant Physiol.* 122, 1289–1299. <https://doi.org/10.1104/pp.122.4.1289>
- Jackson-Constan, D., Keegstra, K., 2001. Arabidopsis Genes Encoding Components of the Chloroplastic Protein Import Apparatus. *Plant Physiol.* 125, 1567–1576. <https://doi.org/10.1104/pp.125.4.1567>
- Jackson, D.T., Froehlich, J.E., Keegstra, K., 1998. The hydrophilic domain of tic110, and inner envelope membrane component of the chloroplastic protein translocation apparatus, faces stromal compartment. *J. Biol. Chem.* 273, 16583–16588. <https://doi.org/10.1074/jbc.273.26.16583>
- Jarvis, P., 2008. Targeting of nucleus-encoded proteins to chloroplasts in plants. *New Phytol.* 179, 257–285. <https://doi.org/10.1111/j.1469-8137.2008.02452.x>
- Jarvis, P., Chen, L., Li, H., Peto, C.A., Fankhauser, C., Chory, J., 1998. An Arabidopsis Mutant Defective in the Plastid General Protein Import Apparatus. *Science* (80-.). 282, 100–103.
- Jarvis, P., Robinson, C., 2004. Mechanisms of Protein Import and Routing in Chloroplasts. *Curr. Biol.* 14, R1064–R1077. <https://doi.org/10.1016/j.cub.2004.11.049>
- Jelic, M., Soll, J., Schleiff, E., 2003. Two Toc34 homologues with different properties. *Biochemistry* 42, 5906–5916. <https://doi.org/10.1021/bi034001q>
- Jelic, M., Sveshnikova, N., Motzkus, M., Hörth, P., Soll, J., Schleiff, E., 2002. The chloroplast import receptor Toc34 functions as preprotein-regulated GTPase. *Biol. Chem.* 383, 1875–1883. <https://doi.org/10.1515/BC.2002.211>

- Johnson, N., Powis, K., High, S., 2013. Post-translational translocation into the endoplasmic reticulum. *Biochim. Biophys. Acta - Mol. Cell Res.* 1833, 2403–2409.
<https://doi.org/10.1016/j.bbamcr.2012.12.008>
- Jouhet, J., Gray, J.C., 2009. Interaction of actin and the chloroplast protein import apparatus. *J. Biol. Chem.* 284, 19132–19141. <https://doi.org/10.1074/jbc.M109.012831>
- Joyard, J., Block, M.A., Douce, R., 1991. Molecular aspects of plastid envelope biochemistry. *Eur. J. Biochem.* 199, 489–509. <https://doi.org/10.1111/j.1432-1033.1991.tb16148.x>
- Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirose, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., Tabata, S., 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* 3, 109–136. <https://doi.org/10.1093/dnares/3.3.109>
- Kanervo, E., Singh, M., Suorsa, M., Paakkari, V., Aro, E., Battchikova, N., Aro, E.M., 2008. Expression of protein complexes and individual proteins upon transition of etioplasts to chloroplasts in pea (*Pisum sativum*). *Plant Cell Physiol.* 49, 396–410.
<https://doi.org/10.1093/pcp/pcn016>
- Karlin-Neumann, G.A., Tobin, E.M., 1986. Transit peptides of nuclear-encoded chloroplast proteins share a common amino acid framework. *EMBO J.* 5, 9–13.
<https://doi.org/10.1002/j.1460-2075.1986.tb04170.x>
- Kasmati, A.R., Töpel, M., Khan, N.Z., Patel, R., Ling, Q., Karim, S., Aronsson, H., Jarvis, P., 2013. Evolutionary, Molecular and Genetic Analyses of Tic22 Homologues in Arabidopsis

- thaliana Chloroplasts. PLoS One 8, e63863. <https://doi.org/10.1371/journal.pone.0063863>
- Kasmati, A.R., Töpel, M., Patel, R., Murtaza, G., Jarvis, P., 2011. Molecular and genetic analyses of Tic20 homologues in Arabidopsis thaliana chloroplasts. Plant J. 66, 877–889. <https://doi.org/10.1111/j.1365-313X.2011.04551.x>
- Kerber, B., Soll, J., 1992. Transfer of a chloroplast-bound precursor protein into the translocation apparatus is impaired after phospholipase C treatment. FEBS Lett. 306, 71–74.
- Kessler, F., Blobel, G., 1996. Interaction of the protein import and folding machineries of the chloroplast. Proc. Natl. Acad. Sci. U. S. A. 93, 7684–7689. <https://doi.org/10.1073/pnas.93.15.7684>
- Kessler, F., Blobel, G., Patel, H.A., Schnell, D.J., 1994. Identification of Two GTP-Binding Proteins in the Chloroplast Protein Import Machinery. Science (80-.). 266, 1035–1039.
- Kessler, F., Schnell, D., 2009. Chloroplast biogenesis: diversity and regulation of the protein import apparatus. Curr. Opin. Cell Biol. 21, 494–500. <https://doi.org/10.1016/j.ceb.2009.03.004>
- Kessler, F., Schnell, D.J., 2006. The function and diversity of plastid protein import pathways: A multilane GTPase highway into plastids. Traffic 7, 248–257. <https://doi.org/10.1111/j.1600-0854.2005.00382.x>
- Kikuchi, S., Asakura, Y., Imai, M., Nakahira, Y., Kotani, Y., Hashiguchi, Y., Nakai, Y., Takafuji, K., Bédard, J., Hirabayashi-Ishioka, Y., Mori, H., Shiina, T., Nakai, M., 2018. A Ycf2-FtsHi heteromeric AAA-ATPase complex is required for chloroplast protein import. Plant Cell tpc.00357.2018. <https://doi.org/10.1105/tpc.18.00357>
- Kikuchi, S., Bédard, J., Hirano, M., Hirabayashi, Y., Oishi, M., Imai, M., Takase, M., Ide, T., Nakai, M., 2013. Uncovering the protein translocon at the chloroplast inner envelope

- membrane. *Science* (80-.). 339, 571–574. <https://doi.org/10.1126/science.1229262>
- Kikuchi, S., Hirohashi, T., Nakai, M., 2006. Characterization of the preprotein translocon at the outer envelope membrane of chloroplasts by blue native PAGE. *Plant Cell Physiol.* 47, 363–371. <https://doi.org/10.1093/pcp/pcj002>
- Kikuchi, S., Oishi, M., Hirabayashi, Y., Lee, D.W., Hwang, I., Nakai, M., 2009. A 1-Megadalton Translocation Complex Containing Tic20 and Tic21 Mediates Chloroplast Protein Import at the Inner Envelope Membrane. *Plant Cell* 21, 1781–1797. <https://doi.org/10.1105/tpc.108.063552>
- Kim, D.H., Lee, J.E., Xu, Z.Y., Geem, K.R., Kwon, Y., Park, J.W., Hwang, I., 2015. Cytosolic targeting factor AKR2A captures chloroplast outer membrane-localized client proteins at the ribosome during translation. *Nat. Commun.* 6, 1–13. <https://doi.org/10.1038/ncomms7843>
- Kim, D.H., Park, M.J., Gwon, G.H., Silkov, A., Xu, Z.Y., Yang, E.C., Song, S., Song, K., Kim, Y., Yoon, H.S., Honig, B., Cho, W., Cho, Y., Hwang, I., 2014. An ankyrin repeat domain of AKR2 drives chloroplast targeting through coincident binding of two chloroplast lipids. *Dev. Cell* 30, 598–609. <https://doi.org/10.1016/j.devcel.2014.07.026>
- Kleffmann, T., Hirsch-Hoffmann, M., Gruissem, W., Baginsky, S., 2006. plprot: A comprehensive proteome database for different plastid types. *Plant Cell Physiol.* 47, 432–436. <https://doi.org/10.1093/pcp/pcj005>
- Kleffmann, T., von Zychlinski, A., Russenberger, D., Hirsch-Hoffmann, M., Gehrig, P., Gruissem, W., Baginsky, S., 2007. Proteome Dynamics during Plastid Differentiation in Rice. *Plant Physiol.* 143, 912–923. <https://doi.org/10.1104/pp.106.090738>
- Kmiec, B., Teixeira, P.F., Berntsson, R.P.-A., Murcha, M.W., Branca, R.M.M., Radomiljac, J.D.,

- Regberg, J., Svensson, L.M., Bakali, A., Langel, U., Lehtio, J., Whelan, J., Stenmark, P., Glaser, E., 2013. Organellar oligopeptidase (OOP) provides a complementary pathway for targeting peptide degradation in mitochondria and chloroplasts. *Proc. Natl. Acad. Sci.* 110, E3761–E3769. <https://doi.org/10.1073/pnas.1307637110>
- Knight, J.S., Gray, J.C., 1995. The N-Terminal Hydrophobic Region of the Mature Phosphate Translocator Is Sufficient for Targeting to the Chloroplast Inner Envelope Membrane. *Plant Cell* 7, 1421–1432. <https://doi.org/10.1105/tpc.7.9.1421>
- Knockenbauer, K.E., Schwartz, T.U., 2016. The Nuclear Pore Complex as a Flexible and Dynamic Gate. *Cell* 164, 1162–1171. <https://doi.org/10.1146/annurev.bi.64.070195>
- Kobayashi, K., Kondo, M., Fukuda, H., Nishimura, M., Ohta, H., 2007. Galactolipid synthesis in chloroplast inner envelope is essential for proper thylakoid biogenesis, photosynthesis, and embryogenesis. *Proc. Natl. Acad. Sci.* 104, 17216–17221. <https://doi.org/10.1073/pnas.0704680104>
- Koenig, P., Oreb, M., Höfle, A., Kaltofen, S., Rippe, K., Sinning, I., Schleiff, E., Tews, I., 2008. The GTPase Cycle of the Chloroplast Import Receptors Toc33/Toc34: Implications from Monomeric and Dimeric Structures. *Structure* 16, 585–596. <https://doi.org/10.1016/j.str.2008.01.008>
- Kogata, N., Nishio, K., Hirohashi, T., Kikuchi, S., Nakai, M., 1999. Involvement of a chloroplast homologue of the signal recognition particle receptor protein, FtsY, in protein targeting to thylakoids. *FEBS Lett.* 447, 329–333. [https://doi.org/10.1016/S0014-5793\(99\)00305-1](https://doi.org/10.1016/S0014-5793(99)00305-1)
- Köhler, D., Montandon, C., Hause, G., Majovsky, P., Kessler, F., Baginsky, S., Agne, B., 2015. Characterization of Chloroplast Protein Import without Tic56, a Component of the 1-Megadalton Translocon at the Inner Envelope Membrane of Chloroplasts. *Plant Physiol.*

- 167, 972–990. <https://doi.org/10.1104/pp.114.255562>
- Kouranov, A., Chen, X., Fuks, B., Schnell, D.J., 1998. Tic20 and Tic22 are new components of the protein import apparatus at the chloroplast inner envelope membrane. *J. Cell Biol.* 143, 991–1002. <https://doi.org/10.1083/jcb.143.4.991>
- Kouranov, A., Wang, H., Schnell, D.J., 1999. Tic22 is targeted to the intermembrane space of chloroplasts by a novel pathway. *J. Biol. Chem.* 274, 25181–25186. <https://doi.org/10.1074/jbc.274.35.25181>
- Kovacheva, S., Bédard, J., Patel, R., Dudley, P., Twell, D., Ríos, G., Koncz, C., Jarvis, P., 2005. In vivo studies on the roles of Tic110, Tic40, and Hsp93 during chloroplast protein import. *Plant J.* 41, 412–428. <https://doi.org/10.1093/mp/ssp079>
- Kovács-Bogdán, E., Benz, J.P., Soll, J., Bölter, B., 2011. Tic20 forms a channel independent of Tic110 in chloroplasts. *BMC Plant Biol.* 11. <https://doi.org/10.1186/1471-2229-11-133>
- Kubis, S., Patel, R., Combe, J., 2004. Functional specialization amongst the Arabidopsis Toc159 family of chloroplast protein import receptors. *Plant Cell* 16, 2059–2077. <https://doi.org/10.1105/tpc.104.023309>
- Küchler, M., Dicker, S., Hörmann, F., Soll, J., Heins, L., 2002. Protein import into chloroplasts involves redox-regulated proteins. *Curr. Opin. Plant Biol.* 21, 6136–6145.
- Kumazaki, K., Chiba, S., Takemoto, M., Furukawa, A., Nishiyama, K., Sugano, Y., Mori, T., Dohmae, N., Hirata, K., Nakada-Nakura, Y., Maturana, A.D., Tanaka, Y., Mori, H., Sugita, Y., Arisaka, F., Ito, K., Ishitani, R., Tsukazaki, T., Nureki, O., 2014. Structural basis of Sec-independent membrane protein insertion by YidC. *Nature* 509, 516–20. <https://doi.org/10.1038/nature13167>
- Lamberti, G., Drurey, C., Soll, J., Schwenkert, S., 2011a. The phosphorylation state of

- chloroplast transit peptides regulates preprotein import. *Plant Signal. Behav.* 6, 1918–1920.
<https://doi.org/10.4161/psb.6.12.18127>
- Lamberti, G., Gügel, I.L., Meurer, J., Soll, J., Schwenkert, S., 2011b. The Cytosolic Kinases STY8, STY17, and STY46 Are Involved in Chloroplast Differentiation in Arabidopsis. *Plant Physiol.* 157, 70–85. <https://doi.org/10.1104/pp.111.182774>
- Lee, D.W., Kim, J.K., Lee, S., Choi, S., Kim, S., Hwang, I., 2008. Arabidopsis Nuclear-Encoded Plastid Transit Peptides Contain Multiple Sequence Subgroups with Distinctive Chloroplast-Targeting Sequence Motifs. *Plant Cell* 20, 1603–1622.
<https://doi.org/10.1105/tpc.108.060541>
- Lee, D.W., Lee, J., Hwang, I., 2017a. Sorting of nuclear-encoded chloroplast membrane proteins. *Curr. Opin. Plant Biol.* 40, 1–7. <https://doi.org/10.1016/j.pbi.2017.06.011>
- Lee, D.W., Lee, S., Lee, G., Lee, K.H., Kim, S., Cheong, G.-W., Hwang, I., 2006. Functional Characterization of Sequence Motifs in the Transit Peptide of Arabidopsis Small Subunit of Rubisco. *Plant Physiol.* 140, 466–483. <https://doi.org/10.1104/pp.105.074575>
- Lee, D.W., Woo, S., Geem, K.R., Hwang, I., 2015. Sequence Motifs in Transit Peptides Act as Independent Functional Units and Can Be Transferred to New Sequence Contexts. *Plant Physiol.* 169, 471–484. <https://doi.org/10.1104/pp.15.00842>
- Lee, D.W., Yoo, Y.-J., Razzak, M.A., Hwang, I., 2017b. Prolines in transit peptides are crucial for efficient preprotein translocation into chloroplasts. *Plant Physiol.* pp.01553.2017.
<https://doi.org/10.1104/pp.17.01553>
- Leister, D., 2003. Chloroplast research in the genomic age. *Trends Genet.* 19, 47–56.
[https://doi.org/10.1016/S0168-9525\(02\)00003-3](https://doi.org/10.1016/S0168-9525(02)00003-3)
- Li, H., Chiu, C.-C., 2010. Protein Transport into Chloroplasts. *Annu. Rev. Plant Biol.* 61, 157–

180. <https://doi.org/10.1146/annurev-arplant-042809-112222>
- Li, H. min, Teng, Y.S., 2013. Transit peptide design and plastid import regulation. *Trends Plant Sci.* 18, 360–366. <https://doi.org/10.1016/j.tplants.2013.04.003>
- Li, H.M., Chen, L.J., 1997. A novel chloroplastic outer membrane-targeting signal that functions at both termini of passenger polypeptides. *J. Biol. Chem.* 272, 10968–10974. <https://doi.org/10.1074/jbc.272.16.10968>
- Li, M., Schnell, D.J., 2006. Reconstitution of protein targeting to the inner envelope membrane of chloroplasts. *J. Cell Biol.* 175, 249–259. <https://doi.org/10.1083/jcb.200605162>
- Li, Y., Martin, J.R., Aldama, G.A., Fernandez, D.E., Cline, K., 2017. Identification of Putative Substrates of SEC2, a Chloroplast Inner Envelope Translocase. *Plant Physiol.* 173, 2121–2137. <https://doi.org/10.1104/pp.17.00012>
- Li, Y., Singhal, R., Taylor, I.W., McMinn, P.H., Chua, X.Y., Cline, K., Fernandez, D.E., 2015. The Sec2 translocase of the chloroplast inner envelope contains a unique and dedicated SECE2 component. *Plant J.* 84, 647–658. <https://doi.org/10.1111/tpj.13028>
- Liebers, M., Grübler, B., Chevalier, F., Lerbs-Mache, S., Merendino, L., Blanvillain, R., Pfanschmidt, T., 2017. Regulatory Shifts in Plastid Transcription Play a Key Role in Morphological Conversions of Plastids during Plant Development. *Front. Plant Sci.* 8, 1–8. <https://doi.org/10.3389/fpls.2017.00023>
- Ling, Q., Huang, W., Baldwin, A.J., Jarvis, P., 2012. Chloroplast Biogenesis Is Regulated by Direct Action of the Ubiquitin-Proteasome System. *Science.* 338, 655–659. <https://doi.org/10.1126/science.1157880>
- Ling, Q., Jarvis, P., 2015. Regulation of chloroplast protein import by the ubiquitin E3 ligase SP1 is important for stress tolerance in plants. *Curr. Biol.* 25, 2527–2534.

<https://doi.org/10.1016/j.cub.2015.08.015>

- Liu, L., McNeilage, R.T., Shi, L. -x., Theg, S.M., 2014. ATP Requirement for Chloroplast Protein Import Is Set by the Km for ATP Hydrolysis of Stromal Hsp70 in *Physcomitrella patens*. *Plant Cell* 26, 1246–1255. <https://doi.org/10.1105/tpc.113.121822>
- Lu, Y., Savage, L.J., Larson, M.D., Wilkerson, C.G., Last, R.L., 2011. Chloroplast 2010: A Database for Large-Scale Phenotypic Screening of *Arabidopsis* Mutants. *Plant Physiol.* 155, 1589–1600. <https://doi.org/10.1104/pp.110.170118>
- Lübeck, J., Heins, L., Soll, J., 1997. A nuclear-coded chloroplastic inner envelope membrane protein uses a soluble sorting intermediate upon import into the organelle. *J. Cell Biol.* 137, 1279–1286. <https://doi.org/10.1083/jcb.137.6.1279>
- Lübeck, J., Soll, J., Akita, M., Nielsen, E., Keegstra, K., 1996. Topology of IEP110, a component of the chloroplastic protein import machinery present in the inner envelope membrane. *EMBO J.* 15, 4230–8. <https://doi.org/10.1126/science.292.5522.1636>
- Lung, S.-C., Chuong, S.D.X., 2012. A Transit Peptide–Like Sorting Signal at the C Terminus Directs the *Bienertia sinuspersici* Preprotein Receptor Toc159 to the Chloroplast Outer Membrane. *Plant Cell* 24, 1560–1578. <https://doi.org/10.1105/tpc.112.096248>
- Lynch, M., Blanchard, J.L., 1998. Deleterious mutation accumulation in organelle genomes., in: Woodruff, R.C., Thompson, J.N. (Eds.), *Mutation and Evolution. Contemporary Issues in Genetics and Evolution*, Vol 7. Springer, Dordrecht, pp. 29–39. <https://doi.org/10.1023/a:1017022522486>
- Machettira, A.B., Gross, L.E., Sommer, M.S., Weis, B.L., English, G., Tripp, J., Schleiff, E., 2011. The localization of Tic20 proteins in *Arabidopsis thaliana* is not restricted to the inner envelope of chloroplasts. *Plant Mol. Biol.* 77, 381.

- Mackenzie, S.A., 2005. Plant organellar protein targeting: A traffic plan still under construction. *Trends Cell Biol.* 15, 548–554. <https://doi.org/10.1016/j.tcb.2005.08.007>
- Martin, W., Herrmann, R.G., 1998. Gene Transfer from Organelles to the Nucleus: How Much, What Happens, and Why? *Plant Physiol.* 118, 9–17. <https://doi.org/10.1104/pp.118.1.9>
- Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., Penny, D., 2002. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl. Acad. Sci.* 99, 12246–12251. <https://doi.org/10.1073/pnas.182432999>
- Matos, C.F., Robinson, C., Di Cola, A., 2008. The Tat system proofreads FeS protein substrates and directly initiates the disposal of rejected molecules. *EMBO J.* 27, 2055–2063. <https://doi.org/10.1038/emboj.2008.132>
- May, T., Soll, J., 2000. 14-3-3 Proteins Form a Guidance Complex With Chloroplast Precursor Proteins in Plants. *Plant Cell* 12, 53–64. <https://doi.org/10.1105/tpc.12.1.53>
- McFadden, G.I., Van Dooren, G.G., 2004. Evolution: Red algal genome affirms a common origin of all plastids. *Curr. Biol.* 14, 514–516. <https://doi.org/10.1016/j.cub.2004.06.041>
- Meeks, J.C., Elhai, J., Thiel, T., Potts, M., Larimer, F., Lamerdin, J., Predki, P., Atlas, R., 2001. An overview of the Genome of *Nostoc punctiforme*, a multicellular, symbiotic Cyanobacterium. *Photosynth. Res.* 70, 85–106. <https://doi.org/10.1023/A:1013840025518>
- Miras, S., Salvi, D., Ferro, M., Grunwald, D., Garin, J., Joyard, J., Rolland, N., 2002. Non-canonical transit peptide for import into the chloroplast. *J. Biol. Chem.* 277, 47770–47778. <https://doi.org/10.1074/jbc.M207477200>
- Miras, S., Salvi, D., Piette, L., Seigneurin-Berny, D., Grunwald, D., Reinbothe, C., Joyard, J., Reinbothe, S., Rolland, N., 2007. Toc159- and Toc75-independent import of a transit

- sequence-less precursor into the inner envelope of chloroplasts. *J. Biol. Chem.* 282, 29482–29492. <https://doi.org/10.1074/jbc.M611112200>
- Morin, X.K., Soll, J., 1997. Immunogold labelling of cryosectioned pea chloroplasts and initial localization of the proteins associated with the protein import machinery. *Planta* 201, 119–127. <https://doi.org/10.1007/BF01007696>
- Motohashi, R., Ito, T., Kobayashi, M., Taji, T., Nagata, N., Asami, T., Yoshida, S., Yamaguchi-Shinozaki, K., Shinozaki, K., 2003. Functional analysis of the 37 kDa inner envelope membrane polypeptide in chloroplast biogenesis using a Ds-tagged *Arabidopsis* pale-green mutant. *Plant J.* 34, 719–731. <https://doi.org/10.1046/j.1365-313X.2003.01763.x>
- Muller, H.J., 1964. The Relation of Recombination to Mutational Advance. *Mutat. Res.* 1, 2–9. <https://doi.org/10.1117/12.722789>
- Muraki, N., Nomata, J., Ebata, K., Mizoguchi, T., Shiba, T., Tamiaki, H., Kurisu, G., Fujita, Y., 2010. X-ray crystal structure of the light-independent protochlorophyllide reductase. *Nature* 465, 110–114. <https://doi.org/10.1038/nature08950>
- Nada, A., Soll, J., 2004. Inner envelope protein 32 is imported into chloroplasts by a novel pathway. *J. Cell Sci.* 117, 3975–3982. <https://doi.org/10.1242/jcs.01265>
- Nakai, M., 2015a. The TIC complex uncovered: The alternative view on the molecular mechanism of protein translocation across the inner envelope membrane of chloroplasts. *Biochim. Biophys. Acta* 1847, 957–967. <https://doi.org/10.1016/j.bbabi.2015.02.011>
- Nakai, M., 2015b. YCF1: A Green TIC: Response to the de Vries et al. Commentary. *Plant Cell* 27, 1834–1838. <https://doi.org/10.1105/tpc.15.00363>
- Nakrieko, K.A., Mould, R.M., Smith, A.G., 2004. Fidelity of targeting to chloroplasts is not affected by removal of the phosphorylation site from the transit peptide. *Eur. J. Biochem.*

271, 509–516. <https://doi.org/10.1046/j.1432-1033.2003.03950.x>

Nanjo, Y., Oka, H., Ikarashi, N., Kaneko, K., Kitajima, A., Mitsui, T., Munoz, F.J., Rodriguez-Lopez, M., Baroja-Fernandez, E., Pozueta-Romero, J., 2006. Rice Plastidial N-Glycosylated Nucleotide Pyrophosphatase/Phosphodiesterase Is Transported from the ER-Golgi to the Chloroplast through the Secretory Pathway. *Plant Cell* 18, 2582–2592.

<https://doi.org/10.1105/tpc.105.039891>

Nassoury, N., Morse, D., 2005. Protein targeting to the chloroplasts of photosynthetic eukaryotes: getting there is half the fun. *Biochim. Biophys. Acta (BBA)-Molecular Cell Res.* 1743, 5–19.

Nicolay, K., Laterveer, F.D., van Heerde, W.L., 1994. Effects of amphipathic peptides, including presequences, on the functional integrity of rat liver mitochondrial membranes. *J. Bioenerg. Biomembr.* 26, 327–334.

Nielsen, E., Akita, M., Davila-Aponte, J., Keegstra, K., 1997. Stable association of chloroplastic precursors with protein translocation complexes that contain proteins from both envelope membranes and a stromal Hsp100 molecular chaperone. *EMBO J.* 16, 935–946.

<https://doi.org/10.1093/emboj/16.5.935>

Nishimura, K., Asakura, Y., Friso, G., Kim, J., Oh, S. -h., Rutschow, H., Ponnala, L., van Wijk, K.J., 2013. ClpS1 Is a Conserved Substrate Selector for the Chloroplast Clp Protease System in Arabidopsis. *Plant Cell* 25, 2276–2301. <https://doi.org/10.1105/tpc.113.112557>

O’Neil, P.K., Richardson, L.G.L., Paila, Y.D., Piszczek, G., Chakravarthy, S., Noinaj, N., Schnell, D., 2017. The POTRA domains of Toc75 exhibit chaperone-like function to facilitate import into chloroplasts. *Proc. Natl. Acad. Sci.* 114, E4868–E4876.

<https://doi.org/10.1073/pnas.1621179114>

- Oreb, M., Höfle, A., Koenig, P., Sommer, M.S., Sinning, I., Wang, F., Tews, I., Schnell, D.J., Schleiff, E., 2011. Substrate binding disrupts dimerization and induces nucleotide exchange of the chloroplast GTPase Toc33. *Biochem. J.* 436, 313–319.
<https://doi.org/10.1042/BJ20110246>
- Paila, Y.D., Richardson, L.G.L., Inoue, H., Parks, E.S., McMahon, J., Inoue, K., Schnell, D.J., 2016. Multi-functional roles for the polypeptide transport associated domains of Toc75 in chloroplast protein import. *Elife* 5, 1–29. <https://doi.org/10.7554/eLife.12631>
- Paila, Y.D., Richardson, L.G.L., Schnell, D.J., 2015. New insights into the mechanism of chloroplast protein import and its integration with protein quality control, organelle biogenesis and development. *J. Mol. Biol.* 427, 1038–1060.
<https://doi.org/10.1016/j.jmb.2014.08.016>
- Park, E., Rapoport, T.A., 2012. Mechanisms of Sec61/SecY-Mediated Protein Translocation Across Membranes. *Annu. Rev. Biophys.* 41, 21–40. <https://doi.org/10.1146/annurev-biophys-050511-102312>
- Patron, N.J., Waller, R.F., 2007. Transit peptide diversity and divergence: A global analysis of plastid targeting signals. *BioEssays* 29, 1048–1058. <https://doi.org/10.1002/bies.20638>
- Peltier, J.-B., Emanuelsson, O., Kalume, D.E., Ytterberg, J., Friso, G., Rudella, A., Liberles, D.A., Söderberg, L., Roepstorff, P., von Heijne, G., van Wijk, K.J., 2002. Central Functions of the Lumenal and Peripheral Thylakoid Proteome of Arabidopsis Determined by Experimentation and Genome-Wide Prediction. *Plant Cell* 14, 211–236.
<https://doi.org/10.1105/tpc.010304>
- Perry, S.E., Keegstra, K., 1994. Envelope Membrane Proteins That Interact with Chloroplastic Precursor Proteins. *Plant Cell* 6, 93–105. <https://doi.org/10.1105/tpc.6.1.93>

- Pfisterer, J., Lachmann, P., Kloppstech, K., 1982. Transport of proteins into chloroplasts. *Eur. J. Biochem.* 126, 143–148. <https://doi.org/10.1007/BF00047688>
- Pilon, M., Weisbeek, P.J., de Kruijff, B., 1992. Kinetic analysis of translocation into isolated chloroplasts of the purified ferredoxin precursor. *FEBS Lett.* 302, 65–68. [https://doi.org/10.1016/0014-5793\(92\)80286-P](https://doi.org/10.1016/0014-5793(92)80286-P)
- Pilon, M., Wienk, H., Sips, W., De Swaaf, M., Talboom, I., Van't Hof, R., De Korte- Kool, G., Demel, R., Weisbeek, P., De Kruijff, B., 1995. Functional domains of the ferredoxin transit sequence involved in chloroplast import. *J. Biol. Chem.* <https://doi.org/10.1074/jbc.270.8.3882>
- Pinnaduwege, P., Bruce, B.D., 1996. In vitro interaction between a chloroplast transit peptide and chloroplast outer envelope lipids is sequence-specific and lipid class-dependent. *J. Biol. Chem.* 271, 32907–32915. <https://doi.org/10.1074/jbc.271.51.32907>
- Pogson, B.J., Woo, N.S., Förster, B., Small, I.D., 2008. Plastid signalling to the nucleus and beyond. *Trends Plant Sci.* 13, 602–609. <https://doi.org/10.1016/j.tplants.2008.08.008>
- Price, C.E., Driessen, A.J.M., 2010. Conserved negative charges in the transmembrane segments of subunit K of the NADH:ubiquinone oxidoreductase determine its dependence on YidC for membrane insertion. *J. Biol. Chem.* 285, 3575–3581. <https://doi.org/10.1074/jbc.M109.051128>
- Pujol, C., Maréchal-Drouard, L., Duchêne, A.M., 2007. How Can Organellar Protein N-terminal Sequences Be Dual Targeting Signals? In silico Analysis and Mutagenesis Approach. *J. Mol. Biol.* 369, 356–367. <https://doi.org/10.1016/j.jmb.2007.03.015>
- Qbadou, S., 2003. Membrane insertion of the chloroplast outer envelope protein, Toc34: constrains for insertion and topology. *J. Cell Sci.* 116, 837–846.

<https://doi.org/10.1242/jcs.00291>

- Qbadou, S., Becker, T., Bionda, T., Reger, K., Ruprecht, M., Soll, J., Schleiff, E., 2007. Toc64 - A Preprotein-receptor at the Outer Membrane with Bipartite Function. *J. Mol. Biol.* 367, 1330–1346. <https://doi.org/10.1016/j.jmb.2007.01.047>
- Qbadou, S., Becker, T., Mirus, O., Tews, I., Soll, J., Schleiff, E., 2006. The molecular chaperone Hsp90 delivers precursor proteins to the chloroplast import receptor Toc64. *EMBO J.* 25, 1836–1847. <https://doi.org/10.1038/sj.emboj.7601091>
- Radhamony, R.N., Theg, S.M., 2006. Evidence for an ER to Golgi to chloroplast protein transport pathway. *Trends Cell Biol.* 16, 16–18.
- Ratnayake, R.M.U., Inoue, H., Nonami, H., Akita, M., 2008. Alternative Processing of Arabidopsis Hsp70 Precursors during Protein Import into Chloroplasts. *Biosci. Biotechnol. Biochem.* 72, 2926–2935. <https://doi.org/10.1271/bbb.80408>
- Reddick, L.E., Vaughn, M.D., Wright, S.J., Campbell, I.M., Bruce, B.D., 2007. In vitro comparative kinetic analysis of the chloroplast Toc GTPases. *J. Biol. Chem.* 282, 11410–11426. <https://doi.org/10.1074/jbc.M609491200>
- Reiland, S., Grossmann, J., Baerenfaller, K., Gehrig, P., Nunes-Nesi, A., Fernie, A.R., Grussem, W., Baginsky, S., 2011. Integrated proteome and metabolite analysis of the de-etiolation process in plastids from rice (*Oryza sativa* L.). *Proteomics* 11, 1751–1763.
- Reiss, B., Wasmann, C.C., Schell, J., Bohnert, H.J., 1989. Effect of mutations on the binding and translocation functions of a chloroplast transit peptide. *Proc. Natl. Acad. Sci.* 86, 886–890. <https://doi.org/https://doi.org/10.1016/B978-012435955-0/50002-9>
- Rensink, W.A., Pilon, M., Weisbeek, P., 1998. Domains of a transit sequence required for in vivo import in Arabidopsis chloroplasts. *Plant Physiol.* 118, 691–9.

<https://doi.org/10.1104/pp.118.2.691>

Rensink, W.A., Schnell, D.J., Weisbeek, P.J., 2000. The transit sequence of ferredoxin contains different domains for translocation across the outer and inner membrane of the chloroplast envelope. *J. Biol. Chem.* 275, 10265–10271. <https://doi.org/10.1074/jbc.275.14.10265>

Reumann, S., Davila-Aponte, J., Keegstra, K., 1999. The evolutionary origin of the protein-translocating channel of chloroplastic envelope membranes: Identification of a cyanobacterial homolog. *Proc. Natl. Acad. Sci.* 96, 784–789.

<https://doi.org/10.1073/pnas.96.2.784>

Reumann, S., Keegstra, K., 1999. The endosymbiotic origin of the protein import machinery of chloroplastic envelope membranes. *Trends Plant Sci.* 4, 302–307.

[https://doi.org/10.1016/S1360-1385\(99\)01449-1](https://doi.org/10.1016/S1360-1385(99)01449-1)

Richardson, L.G., Small, E.L., Inoue, H., Schnell, D.J., 2018. Molecular topology of the transit peptide during chloroplast protein import. *Plant Cell* tpc.00172.2018.

<https://doi.org/10.1105/tpc.18.00172>

Richardson, L.G.L., Paila, Y.D., Siman, S.R., Chen, Y., Smith, M.D., Schnell, D.J., 2014. Targeting and assembly of components of the TOC protein import complex at the chloroplast outer envelope membrane. *Front. Plant Sci.* 5, 1–14.

<https://doi.org/10.3389/fpls.2014.00269>

Richly, E., Leister, D., 2004a. NUMTs in sequenced eukaryotic genomes. *Mol. Biol. Evol.* 21, 1081–1084. <https://doi.org/10.1093/molbev/msh110>

Richly, E., Leister, D., 2004b. NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. *Mol. Biol. Evol.* 21, 1972–1980.

<https://doi.org/10.1093/molbev/msh210>

- Richly, E., Leister, D., 2004c. An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of Arabidopsis and rice. *Gene* 329, 11–16.
<https://doi.org/10.1016/j.gene.2004.01.008>
- Richter, S., Lamppa, G.K., 1999. Stromal Processing Peptidase Binds Transit Peptides and Initiates Their ATP-dependent Turnover in Chloroplasts. *J. Cell Biol.* 147, 33–43.
- Richter, S., Lamppa, G.K., 1998. A chloroplast processing enzyme functions as the general stromal processing peptidase. *Proc. Natl. Acad. Sci. U. S. A.* 95, 7463–7468.
<https://doi.org/10.1073/pnas.95.13.7463>
- Rolland, V., Badger, M.R., Price, G.D., 2016. Redirecting the Cyanobacterial Bicarbonate Transporters BicA and SbtA to the Chloroplast Envelope: Soluble and Membrane Cargos Need Different Chloroplast Targeting Signals in Plants. *Front. Plant Sci.* 7, 1–19.
<https://doi.org/10.3389/fpls.2016.00185>
- Rosano, G.L., Bruch, E.M., Ceccarelli, E.A., 2011. Insights into the CLP/HSP100 chaperone system from chloroplasts of Arabidopsis thaliana. *J. Biol. Chem.* 286, 29671–29680.
<https://doi.org/10.1074/jbc.M110.211946>
- Row, P.E., Gray, J.C., 2001. Chloroplast precursor proteins compete to form early import intermediates in isolated pea chloroplasts. *J. Exp. Bot.* 52, 47–56.
<https://doi.org/10.1093/jxb/52.354.47>
- Rowland, E., Kim, J., Bhuiyan, N.H., van Wijk, K.J., 2015. The Arabidopsis Chloroplast stromal N-terminome; complexities of N-terminal protein maturation and stability. *Plant Physiol.* pp.01214.2015. <https://doi.org/10.1104/pp.15.01214>
- Rudolf, M., MacHettira, A.B., Groß, L.E., Weber, K.L., Bolte, K., Bionda, T., Sommer, M.S., Maier, U.G., Weber, A.P.M., Schleiff, E., Tripp, J., 2013. In vivo function of Tic22, a

- protein import component of the intermembrane space of chloroplasts. *Mol. Plant* 6, 817–829. <https://doi.org/10.1093/mp/sss114>
- Ruf, S., Karcher, D., Bock, R., 2007. Determining the transgene containment level provided by chloroplast transformation. *Proc. Natl. Acad. Sci.* 104, 6998–7002. <https://doi.org/10.1073/pnas.0700008104>
- Schaeffer, S., Harper, A., Raja, R., Jaiswal, P., Dhingra, A., 2014. Comparative analysis of predicted plastid-targeted proteomes of sequenced higher plant genomes. *PLoS One* 9, e112870. <https://doi.org/10.1371/journal.pone.0112870>
- Schaeffer, S.M., Christian, R., Castro-Velasquez, N., Hyden, B., Lynch-Holm, V., Dhingra, A., 2017. Comparative ultrastructure of fruit plastids in three genetically diverse genotypes of apple (*Malus × domestica* Borkh.) during development. *Plant Cell Rep.* 36, 1627–1640. <https://doi.org/10.1007/s00299-017-2179-z>
- Schleiff, E., Eichacker, L.A., Eckart, K., Becker, T., Mirus, O., Stahl, T., Soll, J.Ü., 2003a. Prediction of the plant β -barrel proteome: A case study of the chloroplast outer envelope. *Protein Sci.* 12, 748–759. <https://doi.org/10.1110/ps.0237503.proteins>
- Schleiff, E., Jelic, M., Soll, J., 2003b. A GTP-driven motor moves proteins across the outer envelope of chloroplasts. *Proc. Natl. Acad. Sci. U. S. A.* 100, 4604–4609. <https://doi.org/10.1073/pnas.0730860100>
- Schleiff, E., Klösgen, R.B., 2001. Without a little help from “my” friends: Direct insertion of proteins into chloroplast membranes? *Biochim. Biophys. Acta - Mol. Cell Res.* 1541, 22–33. [https://doi.org/10.1016/S0167-4889\(01\)00152-5](https://doi.org/10.1016/S0167-4889(01)00152-5)
- Schleiff, E., Soll, J., Küchler, M., Kühlbrandt, W., Harrer, R., 2003c. Characterization of the translocon of the outer envelope of chloroplasts. *J. Cell Biol.* 160, 541–551.

<https://doi.org/10.1083/jcb.200210060>

Schnell, D.J., Kessler, F., Blobel, G., 1994. Isolation of components of the chloroplast protein import machinery. *Science* (80-.). 266, 1007–1012.

<https://doi.org/10.1126/science.7973649>

Schubert, M., Petersson, U.A., Haas, B.J., Funk, C., Schröder, W.P., Kieselbach, T., 2002.

Proteome Map of the Chloroplast Lumen of *Arabidopsis thaliana*. *J. Biol. Chem.* 277, 8354–8365. <https://doi.org/10.1074/jbc.M108575200>

Schweiger, R., Soll, J., Jung, K., Heermann, R., Schwenkert, S., 2013. Quantification of interaction strengths between chaperones and tetratricopeptide repeat domain-containing membrane proteins. *J. Biol. Chem.* 288, 30614–30625.

<https://doi.org/10.1074/jbc.M113.493015>

Seedorf, M., Soll, J., 1995. Copper chloride, an inhibitor of protein import into chloroplasts.

FEBS Lett. 367, 19–22. [https://doi.org/10.1016/0014-5793\(95\)00529-I](https://doi.org/10.1016/0014-5793(95)00529-I)

Seedorf, M., Waagemann, K., Soll, J., 1995. A constituent of the chloroplast import complex represents a new type of GTP-binding protein. *Plant J.* 7, 401–411.

<https://doi.org/10.1046/j.1365-313X.1995.7030401.x>

Shen, B.R., Zhu, C.H., Yao, Z., Cui, L.L., Zhang, J.J., Yang, C.W., He, Z.H., Peng, X.X., 2017.

An optimized transit peptide for effective targeting of diverse foreign proteins into chloroplasts in rice. *Sci. Rep.* 7, 1–12. <https://doi.org/10.1038/srep46231>

Shi, L.-X., Theg, S.M., 2013. Energetic cost of protein import across the envelope membranes of chloroplasts. *Proc. Natl. Acad. Sci.* 110, 930–935. <https://doi.org/10.1073/pnas.1115886110>

Shi, L.-X., Theg, S.M., 2010. A Stromal Heat Shock Protein 70 System Functions in Protein Import into Chloroplasts in the Moss *Physcomitrella patens*. *Plant Cell* 22, 205–220.

<https://doi.org/10.1105/tpc.109.071464>

Shi, L.X., Theg, S.M., 2013. The chloroplast protein import system: From algae to trees.

Biochim. Biophys. Acta - Mol. Cell Res. 1833, 314–331.

<https://doi.org/10.1016/j.bbamcr.2012.10.002>

Shuenemann, D., Gupta, S., Persello-Carteiaux, F., Klimyuk, V.I., Jones, J.D.G., Nussaume, L.,

Hoffman, N.E., 1998. A novel signal recognition particle targets light-harvesting proteins to the thylakoid membranes. *Proc. Natl. Acad. Sci.* 95, 10312–10316.

<https://doi.org/10.1073/pnas.95.17.10312>

Siddique, M.A., Grossmann, J., Gruissem, W., Baginsky, S., 2006. Proteome analysis of bell pepper (*Capsicum annuum* L.) chromoplasts. *Plant Cell Physiol.* 47, 1663–1673.

<https://doi.org/10.1093/pcp/pcl033>

Singh, N.D., Li, M., Lee, S.-B., Schnell, D., Daniell, H., 2008. Arabidopsis Tic40 Expression in Tobacco Chloroplasts Results in Massive Proliferation of the Inner Envelope Membrane and Upregulation of Associated Proteins. *Plant Cell* 20, 3405–3417.

<https://doi.org/10.1105/tpc.108.063172>

Singhal, R., Fernandez, D.E., 2017. Sorting of SEC translocase SCY components to different membranes in chloroplasts. *J. Exp. Bot.* 68, 5029–5043. <https://doi.org/10.1093/jxb/erx318>

Sjögren, L.L.E., Tanabe, N., Lymperopoulos, P., Khan, N.Z., Rodermel, S.R., Aronsson, H., Clarke, A.K., 2014. Quantitative analysis of the chloroplast molecular chaperone ClpC/Hsp93 in Arabidopsis reveals new insights into its localization, interaction with the Clp proteolytic core, and functional importance. *J. Biol. Chem.* 289, 11318–11330.

<https://doi.org/10.1074/jbc.M113.534552>

Sjuts, I., Soll, J., Bölter, B., 2017. Import of Soluble Proteins into Chloroplasts and Potential

Regulatory Mechanisms. *Front. Plant Sci.* 8, 1–15. <https://doi.org/10.3389/fpls.2017.00168>

Skalitzky, C.A., Martin, J.R., Harwood, J.H., Beirne, J.J., Adamczyk, B.J., Heck, G.R., Cline, K., Fernandez, D.E., 2011. Plastids Contain a Second Sec Translocase System with Essential Functions. *Plant Physiol.* 155, 354–369. <https://doi.org/10.1104/pp.110.166546>

Smeekens, S., Bauerle, C., Hageman, J., Keegstra, K., Weisbeek, P., 1986. The role of the transit peptide in the routing of precursors toward different chloroplast compartments. *Cell* 46, 365–375. [https://doi.org/10.1016/0092-8674\(86\)90657-4](https://doi.org/10.1016/0092-8674(86)90657-4)

Smith, M.D., Hiltbrunner, A., Kessler, F., Schnell, D.J., 2002. The targeting of the atToc159 preprotein receptor to the chloroplast outer membrane is mediated by its GTPase domain and is regulated by GTP. *J. Cell Biol.* 159, 833–843. <https://doi.org/10.1083/jcb.200208017>

Smith, M.D., Rounds, C.M., Wang, F., Chen, K., Afitlhile, M., Schnell, D.J., 2004. atToc159 is a selective transit peptide receptor for the import of nucleus-encoded chloroplast proteins. *J. Cell Biol.* 165, 323–334. <https://doi.org/10.1083/jcb.200311074>

Sohrt, K., Soll, J., 2000. Toc64, a new component of the protein translocon of chloroplasts. *J. Cell Biol.* 148, 1213–1221. <https://doi.org/10.1083/jcb.148.6.1213>

Soll, J., Schleiff, E., 2004. Protein import into chloroplasts. *Nat. Rev. Mol. Cell Biol.* 5, 198–208. <https://doi.org/10.1038/nrm1333>

Solyosi, K., Keresztes, A., 2013. Plastid Structure, Diversification and Interconversions II. *Land Plants. Curr. Chem. Biol.* 6, 187–204. <https://doi.org/10.2174/2212796811206030003>

Sommer, M., Rudolf, M., Tillmann, B., Tripp, J., Sommer, M.S., Schleiff, E., 2013. Toc33 and Toc64-III cooperate in precursor protein import into the chloroplasts of *Arabidopsis thaliana*. *Plant, Cell Environ.* 36, 970–983. <https://doi.org/10.1111/pce.12030>

Ståhl, A., Nilsson, S., Lundberg, P., Bhushan, S., Biverståhl, H., Moberg, P., Morisset, M.,

- Vener, A., Mäler, L., Langel, U., Glaser, E., 2005. Two novel targeting peptide degrading proteases, PrePs, in mitochondria and chloroplasts, so similar and still different. *J. Mol. Biol.* 349, 847–860. <https://doi.org/10.1016/j.jmb.2005.04.023>
- Stahl, T., Glockmann, C., Soll, J., Heins, L., 1999. Tic40, a New “Old” Subunit of the Chloroplast Protein Import Translocon. *J. Biol. Chem.* 274, 37467–72.
- Stegemann, S., Hartmann, S., Ruf, S., Bock, R., 2003. High-frequency gene transfer from the chloroplast genome to the nucleus. *Proc. Natl. Acad. Sci.* 100, 8828–8833. <https://doi.org/10.1073/pnas.1430924100>
- Stengel, A., Benz, J.P., Buchanan, B.B., Soll, J., Bölder, B., 2009. Preprotein import into chloroplasts via the toc and tic complexes is regulated by redox signals in *pisum sativum*. *Mol. Plant* 2, 1181–1197. <https://doi.org/10.1093/mp/ssp043>
- Stengel, A., Benz, P., Balsera, M., Soll, J., Bölder, B., 2008. TIC62 redox-regulated translocon composition and dynamics. *J. Biol. Chem.* 283, 6656–6667. <https://doi.org/10.1074/jbc.M706719200>
- Stengel, A., Soll, J., Bölder, B., 2007. Protein import into chloroplasts: New aspects of a well-known topic. *Biol. Chem.* 388, 765–772. <https://doi.org/10.1515/BC.2007.099>
- Strambio-De-Castillia, C., Niepel, M., Rout, M.P., 2010. The nuclear pore complex: Bridging nuclear transport and gene regulation. *Nat. Rev. Mol. Cell Biol.* 11, 490–501. <https://doi.org/10.1038/nrm2928>
- Su, P.-H., Li, H. -m., 2010. Stromal Hsp70 Is Important for Protein Translocation into Pea and Arabidopsis Chloroplasts. *Plant Cell* 22, 1516–1531. <https://doi.org/10.1105/tpc.109.071415>
- Sugiura, M., 1992. The chloroplast genome. *Plant Mol. Biol.* 19, 149–168.

<https://doi.org/10.1007/s00438-005-0092-6>

Sun, Q., Zybaylov, B., Majeran, W., Friso, G., Olinares, P.D.B., van Wijk, K.J., 2009. PPDB, the Plant Proteomics Database at Cornell. *Nucleic Acids Res.* 37, 969–974.

<https://doi.org/10.1093/nar/gkn654>

Sun, Y.J., Forouhar, F., Li, H.M., Tu, S.L., Yeh, Y.H., Kao, S., Shr, H.L., Chou, C.C., Chen, C., Hsiao, C.D., 2002. Crystal structure of pea toc34, a novel gtpase of the chloroplast protein translocon. *Nat. Struct. Biol.* 9, 95–100. <https://doi.org/10.1038/nsb744>

Suzuki, M., Takahashi, S., Kondo, T., Dohra, H., Ito, Y., Kiriiwa, Y., Hayashi, M., Kamiya, S., Kato, M., Fujiwara, M., Fukao, Y., Kobayashi, M., Nagata, N., Motohashi, R., 2015. Plastid proteomic analysis in tomato fruit development. *PLoS One* 10, 1–25.

<https://doi.org/10.1371/journal.pone.0137266>

Sveshnikova, N., Soll, J., Schleiff, E., 2000. Toc34 is a preprotein receptor regulated by GTP and phosphorylation. *Proc. Natl. Acad. Sci.* 97, 4973–4978.

<https://doi.org/10.1073/pnas.080491597>

Tasaki, T., Sriram, S.M., Park, K.S., Kwon, Y.T., 2012. The N-end Rule Pathway. *Annu. Rev. Biochem.* 81, 261–289. <https://doi.org/10.1146/annurev-biochem-051710-093308>

Tee, E.E., 2018. Tic-Tac-Toe: How TIC and TOC Coordinate Getting Proteins across the Line.

Plant Cell 30, 1666–1667. <https://doi.org/10.1105/tpc.18.00618>

Teixeira, P.F., Kmiec, B., Branca, R.M.M., Murcha, M.W., Byzia, A., Ivanova, A., Whelan, J., Drag, M., Lehtiö, J., Glaser, E., 2017. A multi-step peptidolytic cascade for amino acid recovery in chloroplasts. *Nat. Chem. Biol.* 13, 15–17.

<https://doi.org/10.1038/nchembio.2227>

Teng, Y.-S., Su, Y., Chen, L.-J., Lee, Y.J., Hwang, I., Li, H., 2006. Tic21 Is an Essential

- Translocon Component for Protein Translocation across the Chloroplast Inner Envelope Membrane. *Plant Cell* 18, 2247–2257. <https://doi.org/10.1105/tpc.106.044305>
- The Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815. <https://doi.org/10.1038/35048692>
- Theg, S.M., Cline, K., Finazzi, G., Wollman, F.A., 2005. The energetics of the chloroplast Tat protein transport pathway revisited. *Trends Plant Sci.* 10, 153–154. <https://doi.org/10.1016/j.tplants.2005.02.001>
- Tranel, P.J., Froehlich, J., Goyal, A., Keegstra, K., 1995. A component of the chloroplastic protein import apparatus is targeted to the outer envelope membrane via a novel pathway. *EMBO J.* 14, 2436–46. <https://doi.org/10.1002/J.1460-2075.1995.TB07241.X>
- Tripp, J., Hahn, A., Koenig, P., Flinner, N., Bublak, D., Brouwer, E.M., Ertel, F., Mirus, O., Sinning, I., Tews, I., Schleiff, E., 2012. Structure and conservation of the periplasmic targeting factor Tic22 protein from plants and cyanobacteria. *J. Biol. Chem.* 287, 24164–24173. <https://doi.org/10.1074/jbc.M112.341644>
- Tripp, J., Inoue, K., Keegstra, K., Froehlich, J.E., 2007. A novel serine/proline-rich domain in combination with a transmembrane domain is required for the insertion of AtTic40 into the inner envelope membrane of chloroplasts. *Plant J.* 52, 824–838. <https://doi.org/10.1111/j.1365-313X.2007.03279.x>
- Tsai, J.Y., Chu, C.C., Yeh, Y.H., Chen, L.J., Li, H.M., Hsiao, C.D., 2013. Structural characterizations of the chloroplast translocon protein Tic110. *Plant J.* 75, 847–857. <https://doi.org/10.1111/tpj.12249>
- Tu, S., Chen, L., Smith, M.D., Su, Y., Schnell, D.J., Li, H., 2004. Import Pathways of Chloroplast Interior Proteins and the Outer-Membrane Protein OEP14 Converge at Toc75.

- Plant Cell 16, 2078–2088. <https://doi.org/10.1105/tpc.104.023952>. Many
- Tuba, Z., Lichtenthaler, H.K., Csintalan, Z., Nagy, Z., Szente, K., 1994. Reconstitution of chlorophylls and photosynthetic CO₂ assimilation upon rehydration of the desiccated poikilochlorophyllous plant *Xerophyta scabrifolia* (Pax) Key words. *Planta* 192, 414–420.
- Van't Hof, R., Demel, R.A., Keegstra, K., Kruijff, B. De, 1991. Lipid-Peptide Interactions Between Fragments of the Transit Peptide of Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase and Chloroplast Membrane Lipids. *EMBO J.* 29, 0–4.
- Van't Hof, R., Van Klompenburg, W., Pilon, M., Kozubek, A., De Korte-Kool, G., Demel, R.A., Weisbeek, P.J., De Kruijff, B., 1993. The transit sequence mediates the specific interaction of the precursor of ferredoxin with chloroplast envelope membrane lipids. *J. Biol. Chem.* 268, 4037–4042.
- Van 't Hof, R., de Kruijff, B., 1995a. Transit sequence-dependent binding of the chloroplast precursor protein ferredoxin to lipid vesicles and its implications for membrane stability. *FEBS Lett.* 361, 35–40. [https://doi.org/10.1016/0014-5793\(95\)00135-V](https://doi.org/10.1016/0014-5793(95)00135-V)
- Van 't Hof, R., de Kruijff, B., 1995b. Characterization of the import process of a transit peptide into chloroplasts. *J. Biol. Chem.* 270, 22368–22373. <https://doi.org/10.1074/jbc.270.38.22368>
- van den Wijngaard, P.W.J., Vredenberg, W.J., 1999. The Envelope Anion Channel Involved in Chloroplast Protein Import is Associated with Tic110. *J. Biol. Chem.* 274, 25201–25204.
- van Wijk, K.J., 2015. Protein Maturation and Proteolysis in Plant Plastids, Mitochondria, and Peroxisomes. *Annu. Rev. Plant Biol.* 66, 75–111. <https://doi.org/10.1146/annurev-arplant-043014-115547>
- Van Wijk, K.J., 2004. Plastid proteomics. *Plant Physiol. Biochem.* 42, 963–977.

<https://doi.org/10.1016/j.plaphy.2004.10.015>

- van Wijk, K.J., Baginsky, S., 2011. Plastid Proteomics in Higher Plants: Current State and Future Goals. *Plant Physiol.* 155, 1578–1588. <https://doi.org/10.1104/pp.111.172932>
- Viana, A.A.B., Li, M., Schnell, D.J., 2010. Determinants for stop-transfer and post-import pathways for protein targeting to the chloroplast inner envelope membrane. *J. Biol. Chem.* 285, 12948–12960. <https://doi.org/10.1074/jbc.M110.109744>
- Villarejo, A., Burén, S., Larsson, S., Déjardin, A., Monné, M., Rudhe, C., Karlsson, J., Jansson, S., Lerouge, P., Rolland, N., von Heijne, G., Grebe, M., Bako, L., Samuelsson, G., 2005. Evidence for a protein transported through the secretory pathway en route to the higher plant chloroplast. *Nat. Cell Biol.* 7, 1224–1231. <https://doi.org/10.1038/ncb1330>
- von Heijne, G., Nishikawa, K., 1991. Chloroplast transit peptides the perfect random coil? *FEBS Lett.* 278, 1–3. [https://doi.org/10.1016/0014-5793\(91\)80069-F](https://doi.org/10.1016/0014-5793(91)80069-F)
- von Heijne, G., Steppuhn, J., Herrmann, R.G., 1989. Domain structure of mitochondrial and chloroplast targeting peptides. *Eur. J. Biochem.* 180, 535–545. <https://doi.org/10.1111/j.1432-1033.1989.tb14679.x>
- von Zychlinski, A., Kleffmann, T., Krishnamurthy, N., Sjölander, K., Baginsky, S., Gruissem, W., 2005. Proteome analysis of the rice etioplast: metabolic and regulatory networks and novel protein functions. *Mol. Cell. Proteomics* 4, 1072–1084.
- Waagemann, K., Soll, J., 1991. Characterization of the protein import apparatus in isolated outer envelopes of chloroplasts. *Plant J.* 1, 149–158.
- Wallas, T.R., Smith, M.D., Sanchez-Nieto, S., Schnell, D.J., 2003. The Roles of Toc34 and Toc75 in Targeting the Toc159 Preprotein Receptor to Chloroplasts. *J. Biol. Chem.* 278, 44289–44297.

- Wan, J., Blakeley, S.D., Dennis, D.T., Ko, K., 1996. Transit peptides play a major role in the preferential import of proteins into leucoplasts and chloroplasts. *J. Biol. Chem.* 271, 31227–31233. <https://doi.org/10.1074/jbc.271.49.31227>
- Wan, J., Blakeley, S.D., Dennis, D.T., Ko, K., 1995. Import Characteristics of a Leucoplast Pyruvate Kinase Are Influenced by a 19-Amino-acid Domain within the Protein. *J. Biol. Chem.* 270, 16731–16739. <https://doi.org/10.1074/jbc.270.18.10405>
- Wang, A.X., Wang, D.Y., 2009. Regulation of the ALBINO3-mediated transition to flowering in Arabidopsis depends on the expression of CO and GA1. *Biol. Plant.* 53, 484–492. <https://doi.org/10.1007/s10535-009-0089-9>
- Wang, F., Agne, B., Kessler, F., Schnell, D.J., 2008. The role of GTP binding and hydrolysis at the atToc159 preprotein receptor during protein import into chloroplasts. *J. Cell Biol.* 183, 87–99. <https://doi.org/10.1083/jcb.200803034>
- Wang, P., Xue, L., Batelli, G., Lee, S., Hou, Y.-J., Van Oosten, M.J., Zhang, H., Tao, W.A., Zhu, J.-K., 2013. Quantitative phosphoproteomics identifies SnRK2 protein kinase substrates and reveals the effectors of abscisic acid action. *Proc. Natl. Acad. Sci.* 110, 11205–11210. <https://doi.org/10.1073/pnas.1308974110>
- Wang, Y.Q., Yang, Y., Fei, Z., Yuan, H., Fish, T., Thannhauser, T.W., Mazourek, M., Kochian, L. V., Wang, X., Li, L., 2013. Proteomic analysis of chromoplasts from six crop species reveals insights into chromoplast function and development. *J. Exp. Bot.* 64, 949–961. <https://doi.org/10.1093/jxb/ers375>
- Weibel, P., Hiltbrunner, A., Brand, L., Kessler, F., 2003. Dimerization of Toc-GTPases at the chloroplast protein import machinery. *J. Biol. Chem.* 278, 37321–37329. <https://doi.org/10.1074/jbc.M305946200>

- Wickström, D., Wagner, S., Simonsson, P., Pop, O., Baars, L., Ytterberg, A.J., Van Wijk, K.J., Luirink, J., De Gier, J.W.L., 2011. Characterization of the consequences of YidC depletion on the inner membrane proteome of *E. coli* using 2D blue native/SDS-PAGE. *J. Mol. Biol.* 409, 124–135. <https://doi.org/10.1016/j.jmb.2011.03.068>
- Wiedemann, N., Pfanner, N., 2017. Mitochondrial Machineries for Protein Import and Assembly. *Annu. Rev. Biochem.* 86, 685–714. <https://doi.org/10.1146/annurev-biochem-060815-014352>
- Wienk, H.L.J., Wechselberger, R.W., Czisch, M., De Kruijff, B., 2000. Structure, dynamics, and insertion of a chloroplast targeting peptide in mixed micelles. *Biochemistry* 39, 8219–8227. <https://doi.org/10.1021/bi000110i>
- Wieprecht, T., Apostolov, O., Beyermann, M., Seelig, J., 2000. Interaction of a mitochondrial presequence with lipid membranes: Role of helix formation for membrane binding and perturbation. *Biochemistry* 39, 15297–15305. <https://doi.org/10.1021/bi001774v>
- Wimmer, D., Bohnhorst, P., Shekhar, V., Hwang, I., Offermann, S., 2017. Transit peptide elements mediate selective protein targeting to two different types of chloroplasts in the single-cell C4 species *Bienertia sinuspersici*. *Sci. Rep.* 7, 41187. <https://doi.org/10.1038/srep41187>
- Wise, R.R., 2007. The Diversity of Plastid Form and Function, in: Wise, R.R., Hooper, K.J. (Eds.), *The Structure and Function of Plastids. Advances in Photosynthesis and Respiration*, Vol. 23. Springer, Dordrecht, Dordrecht, pp. 3–26.
- Woolhead, C.A., Thompson, S.J., Moore, M., Tissier, C., Mant, A., Rodger, A., Henry, R., Robinson, C., 2001. Distinct Albino3-dependent and -independent Pathways for Thylakoid Membrane Protein Insertion. *J. Biol. Chem.* 276, 40841–40846.

<https://doi.org/10.1074/jbc.M106523200>

Wu, C., Seibert, F.S., Ko, K., 1994. Identification of chloroplast envelope proteins in close physical proximity to a partially translocated chimeric precursor protein. *J. Biol. Chem.* 269, 32264–32271. <https://doi.org/10.1021/BI00019A021>

Yamano, K., Kuroyanagi-Hasegawa, M., Esaki, M., Yokota, M., Endo, T., 2008. Step-size analyses of the mitochondrial Hsp70 import motor reveal the Brownian ratchet in operation. *J. Biol. Chem.* 283, 27325–27332. <https://doi.org/10.1074/jbc.M805249200>

Yeh, Y.H., Kesavulu, M.M., Li, H.M., Wu, S.Z., Sun, Y.J., Konozy, E.H.E., Hsiao, C.D., 2007. Dimerization is important for the GTPase activity of chloroplast translocon components atToc33 and psToc159. *J. Biol. Chem.* 282, 13845–13853. <https://doi.org/10.1074/jbc.M608385200>

Yoon, H.S., Hackett, J.D., Ciniglia, C., Pinto, G., Bhattacharya, D., 2004. A Molecular Timeline for the Origin of Photosynthetic Eukaryotes. *Mol. Biol. Evol.* 21, 809–818. <https://doi.org/10.1093/molbev/msh075>

Young, M.E., Keegstra, K., Froehlich, J.E., 1999. GTP promotes the formation of early-import intermediates but is not required during the translocation step of protein import into chloroplasts. *Plant Physiol* 121, 237–44. <https://doi.org/10.1104/pp.121.1.237>

Ytterberg, A.J., Peltier, J., Wijk, K.J. Van, 2006. Protein Profiling of Plastoglobules in Chloroplasts and Chromoplasts. A Surprising Site for Differential Accumulation of Metabolic Enzymes. *Plant Physiol.* 140, 984–997. <https://doi.org/10.1104/pp.105.076083.984>

Zeng, Y., Pan, Z., Ding, Y., Zhu, A., Cao, H., Xu, Q., Deng, X., 2011. A proteomic analysis of the chromoplasts isolated from sweet orange fruits [*Citrus sinensis* (L.) Osbeck]. *J. Exp.*

- Bot. 62, 5297–5309. <https://doi.org/10.1093/jxb/err140>
- Zeng, Y., Pan, Z., Wang, L., Ding, Y., Xu, Q., Xiao, S., Deng, X., 2014. Phosphoproteomic analysis of chloroplasts from sweet orange during fruit ripening. *Physiol. Plant.* 150, 252–270. <https://doi.org/10.1111/ppl.12080>
- Zhang, X.P., Glaser, E., 2002. Interaction of plant mitochondrial and chloroplast signal peptides with the Hsp70 molecular chaperone. *Trends Plant Sci.* 7, 14–21. [https://doi.org/10.1016/S1360-1385\(01\)02180-X](https://doi.org/10.1016/S1360-1385(01)02180-X)
- Zhong, R., Thompson, J., Ottesen, E., Lamppa, G.K., 2010. A forward genetic screen to explore chloroplast protein import in vivo identifies Moco sulfurase, pivotal for ABA and IAA biosynthesis and purine turnover. *Plant J.* 63, 44–59. <https://doi.org/10.1111/j.1365-313X.2010.04220.x>
- Zhong, R., Wan, J., Jin, R., Lamppa, G., 2003. A pea antisense gene for the chloroplast stromal processing peptidase yields seedling lethals in Arabidopsis: Survivors show defective GFP import in vivo. *Plant J.* 34, 802–812. <https://doi.org/10.1046/j.1365-313X.2003.01772.x>
- Zhu, L., Wasey, A., White, S.H., Dalbey, R.E., 2013. Charge composition features of model single-span membrane proteins that determine selection of YidC and SecYEG translocase pathways in Escherichia coli. *J. Biol. Chem.* 288, 7704–7716. <https://doi.org/10.1074/jbc.M112.429431>
- Zhu, M., Lin, J., Ye, J., Wang, R., Yang, C., Gong, J., Liu, Y., Deng, C., Liu, P., Chen, C., Cheng, Y., Deng, X., Zeng, Y., 2018. A comprehensive proteomic analysis of elaioplasts from citrus fruits reveals insights into elaioplast biogenesis and function. *Hortic. Res.* 5, 0–10. <https://doi.org/10.1038/s41438-017-0014-x>
- Ziehe, D., Dünschede, B., Schünemann, D., 2017. From bacteria to chloroplasts: evolution of the

- chloroplast SRP system. *Biol. Chem.* 398, 653–661. <https://doi.org/10.1515/hsz-2016-0292>
- Zimorski, V., Ku, C., Martin, W.F., Gould, S.B., 2014. Endosymbiotic theory for organelle origins. *Curr. Opin. Microbiol.* 22, 38–48. <https://doi.org/10.1016/j.mib.2014.09.008>
- Zufferey, M., Montandon, C., Douet, V., Demarsy, E., Agne, B., Baginsky, S., Kessler, F., 2017. The novel chloroplast outer membrane kinase KOC1 is a required component of the plastid protein import machinery. *J. Biol. Chem.* 292, 6952–6964. <https://doi.org/10.1074/jbc.M117.776468>
- Zybaïlov, B., Rutschow, H., Friso, G., Rudella, A., Emanuelsson, O., Sun, Q., van Wijk, K.J., 2008. Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. *PLoS One* 3, e1994. <https://doi.org/10.1371/journal.pone.0001994>

Table 1: Regulation Points of TOC/TIC. All current literature on regulation of protein import by formation of disulfide bridges, protein-protein interactions, proteolysis, phosphorylation, and ligand binding are summarized here. Where known, the effect of each component on import is indicated.

Category	Component	Cofactors/ Ligands	Effect on Import	References
Disulfide Bridge	Toc64		Decrease	(Sohrt and Soll, 2000)
Disulfide Bridge	Tic110	Tic40	Decrease	(Balsera et al., 2009; Stahl et al., 1999)
Disulfide Bridge	Toc75	Toc33, Toc159	Decrease	(Seedorf and Soll, 1995; Stengel et al., 2009)
Disulfide Bridge	Tic55		Unknown	(Balsera et al., 2009; Bartsch et al., 2008)
Disulfide Bridge	Tic110- Tic40		Unknown	(Stahl et al., 1999)
Protein-Protein Interaction	Tic32	Calmodulin	Decrease	(Chigri et al., 2006)
Protein-Protein Interaction	Tic62	FNR	Decrease	(Stengel et al., 2008)
Proteolysis	Toc159	Ubiquitin/SP1 E3 Ligase	Decrease	See Ling and Jarvis, 2015b and Ling and Jarvis, 2016
Phosphorylation	Toc159	KOC1	Increase	(Zufferey et al., 2017)
Phosphorylation		SnRK2		
Phosphorylation	Toc159	OEK70 (KOC1?)	Decrease (no data)	(Fulgosi and Soll, 2002)
Phosphorylation	Toc33		Decrease	(Jelic et al., 2002; Sveshnikova et al., 2000)
Phosphorylation	Toc33	OEK98	Decrease	(Fulgosi and Soll, 2002)
Ligand Binding	Tic55	Thioredoxin	Unknown	(Bartsch et al., 2008)
Ligand Binding	Tic110	Thioredoxin	Decrease	(Balsera et al., 2009)
Ligand Binding	Tic62	NADPH	Increase	(Balsera et al., 2009; Stengel et al., 2008)
Ligand Binding	Tic32	NADPH	Increase	(Balsera et al., 2009; Chigri et al., 2006)

Table 2: Summary of Non-Green Plastid Proteomics Studies. Characterization of non-green plastids are poorly represented in proteomics literature and are represented by only a handful of major studies. However, several of the listed publications have found unique proteins in non-green morphotypes that are not present in green chloroplasts. Arabidopsis: *Arabidopsis thaliana*; Cauliflower: *Brassica oleracea*; sweet orange: *Citrus sinensis*; carrot: *Daucus carota*; kumquat: *Fortunella margarita*; medicago: *Medicago truncatula*; papaya: *Carica papaya*; pea: *Pisum sativum*; pepper: *Capsicum annuum*; rice: *Oryza sativa*; tobacco: *Nicotiana tabacum*; tomato: *Solanum lycopersicum*; watermelon: *Citrullus lanatus*; wheat: *Triticum aestivum*

Publication	Species	Plastid type	Method	Number of unique plastid proteins	Overlap with Chloroplast Proteome
(Andon et al., 2002)	Wheat	Amyloplast	1-D/2-D gel, LC-MS/MS	171	N/A
(Baginsky et al., 2004)	Tobacco	Proplastid	(RP-LC)-MS/MS (reverse-phase LC); used both electrospray and nanospray ionization	168	121
(von Zychlinski et al., 2005)	Rice	Etioplast	LC-NI MS/MS	216	N/A
(Ytterberg et al., 2006)	Arabidopsis	Chloroplast (Plastoglobuli)	nLC -MS/MS (Electrospray Ionization-tandem MS))	32	N/A
(Siddique et al., 2006)	Pepper	Chromoplast	SDS-PAGE-(RP-LC)-MS/MS	151	N/A
(Balmer et al., 2006; Dupont, 2008)	Wheat	Amyloplast	2-D gel, LC-MS/MS	180	N/A
(Kleffmann et al., 2007)	Rice	Etioplast, Chloroplast	2-D PAGE	477	N/A
(Kanervo et al., 2008)	Pea	Etioplast, Chloroplast	BN-PAGE, SDS-PAGE->LC-ESI MS/MS	14	N/A
(Bräutigam and Weber, 2009)	Cauliflower	Proplastid	MS/MS	226	N/A
(Barsan et al., 2010)	Tomato	Chromoplast	LC-MS/MS	988	577
(Daher et al., 2010)	Medicago	Nodular Leucoplasts	LC-MS/MS	266	N/A
(Zeng et al., 2011)	Sweet Orange	Chromoplast	SDS-PAGE-LC-MS/MS	418	N/A
(Barsan et al., 2012)	Tomato	Chloroplast, Chromoplast		1932	N/A
(Y. Q. Wang et al., 2013)	Tomato Pepper Carrot Cauliflower Watermelon Papaya	Chromoplast	nLC-MS/MS	953, 1752, 1891, 2262, 1170, 1581	N/A
(Zeng et al., 2014)	Sweet Orange	Chromoplast	Titanium oxide affinity chromatography LC-MS/MS	109	N/A
(Suzuki et al., 2015)	Tomato	Chromoplast	GeLC-LC-MS/MS	605	82
(Daher et al., 2017)	Medicago	Nodular Leucoplasts	LC-MS/MS	490	N/A

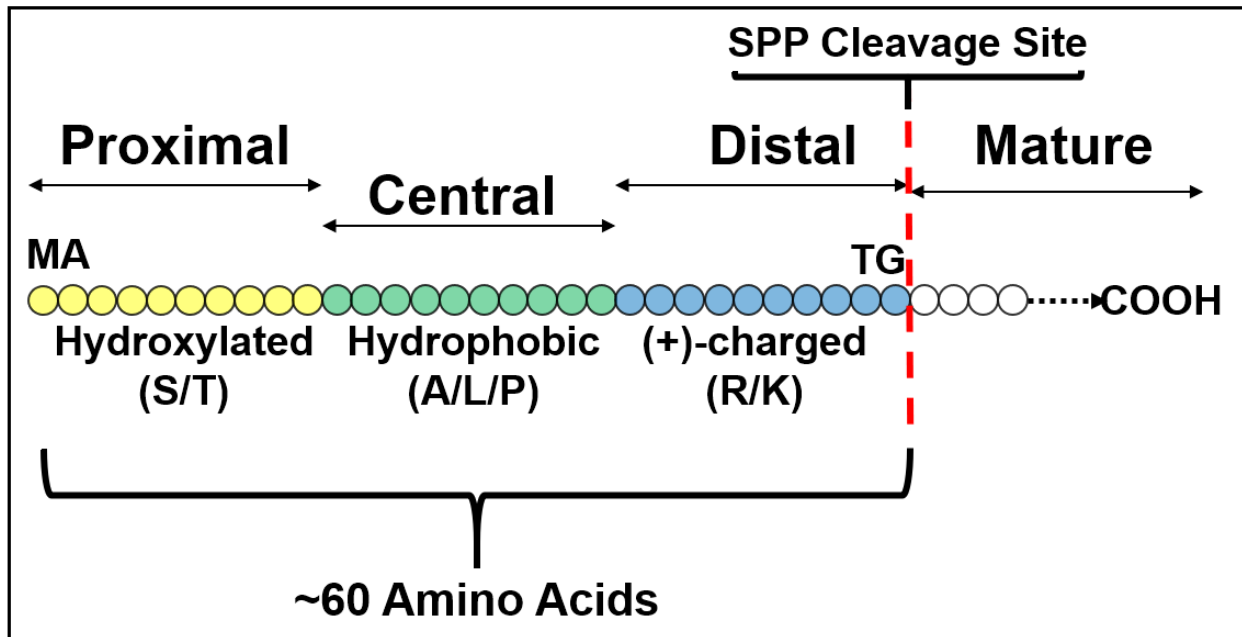


Figure 1: Basic Transit Peptide Structural Model. The three major domains required for a functional transit peptide include a hydroxylated N-terminus (N-domain) for binding to cytosolic and stromal chaperones, a hydrophobic, uncharged central domain responsible for bridging the outer and inner envelopes, and a positively-charged C-terminus that interacts with TOC GTPases to stabilize early translocation intermediates and ultimately trigger full translocation. Some elements may be repeated or be present in only some transit peptides, such as acidic residues at the C-terminus that may confer selectivity for certain TOC GTPases (Christian et al., 2019a, unpublished).

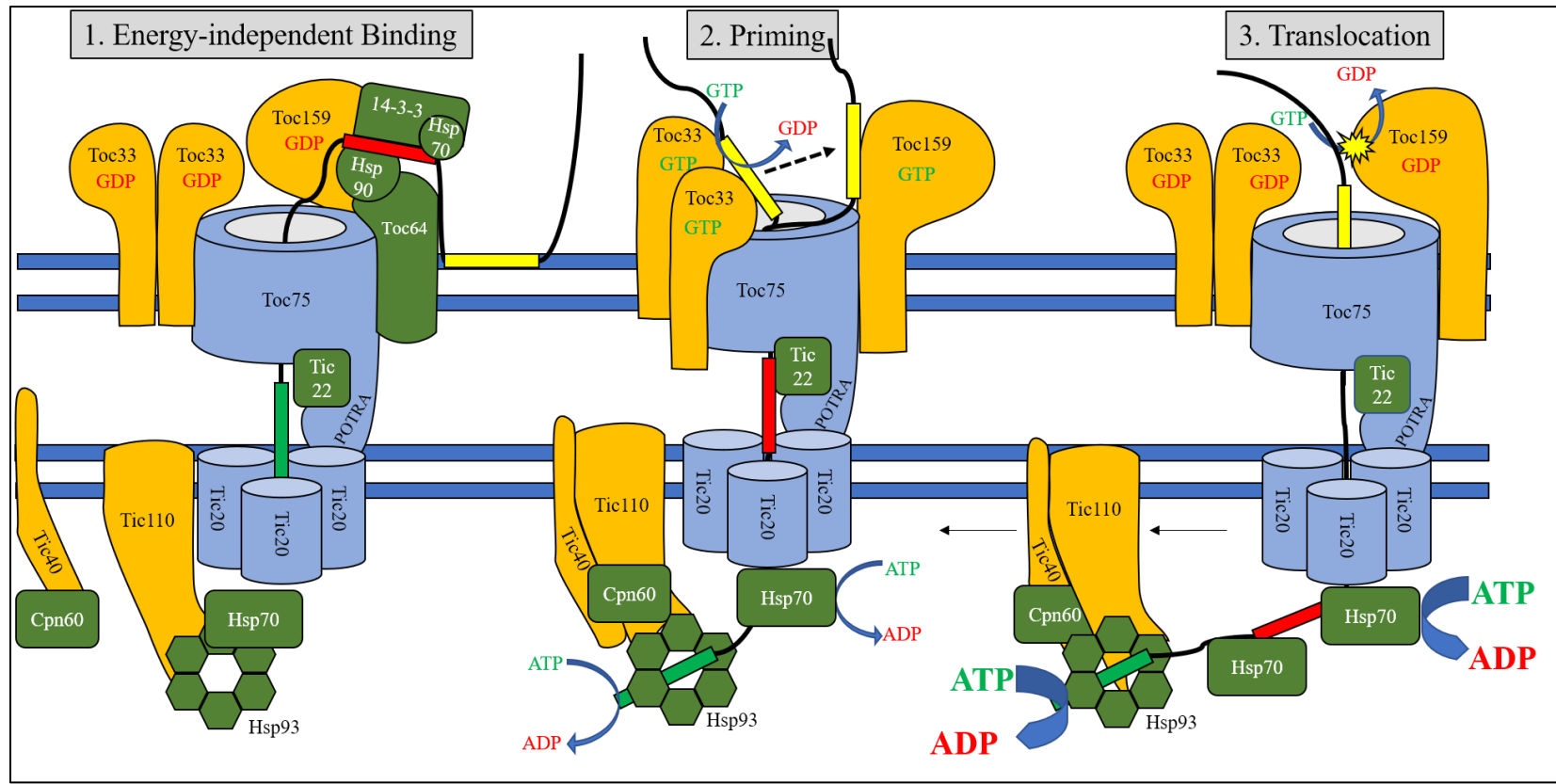


Figure 2: Model of TOC/TIC translocation. The three stages of translocation are represented in three panels (A-C), corresponding to Energy-Independent Binding, Priming, and Translocation stages. Initial stages (1) involve binding of the central transit peptide domain to cytosolic chaperones, and binding of the proximal domain to TIC22 and the TO75 POTRA domains. Transition to the early translocation intermediate stage (2) involves translocation of the proximal domain across the inner membrane and binding to stromal chaperones, while the distal domain is stabilized by TOC GTPases. Triggering the final stage of translocation (3) involves a final GTPase cycle for primed transit peptides, followed by strong ATPase activity of stromal Hsp70 and other chaperones to pull the preprotein into the stroma. The N-terminal motifs of the transit peptide are represented in green, the central motifs in red, and the C-terminal motifs in yellow. Proteins shown in blue represent pore or channel proteins, proteins shown in orange are receptor or scaffold proteins, and proteins shown in green are chaperones.

CHAPTER 2

Genome-Scale Characterization of Predicted Plastid-Targeted Proteins in Higher Plants

Ryan Christian^{1,2}, Seanna Hewitt^{1,2}, Eric Roalson^{1,2,3}, and Amit Dhingra^{1,2,§}

Target Journal: Genome Biology Special Issue

¹Department of Horticulture, Washington State University, Pullman, WA

²Molecular Plant Sciences Program, Washington State University, Pullman, WA

³School of Biological Sciences, Washington State University, Pullman, WA

[§]Corresponding author

RWC: ryan_christian@wsu.edu

AD: adhingra@wsu.edu

Abstract

Plastids are morphologically and functionally diverse organelles that are dependent on nuclear-encoded, plastid-targeted proteins for all biochemical and regulatory functions. However, how plastid proteomes vary temporally, spatially, and taxonomically has been historically difficult to analyze at genome-wide scale using experimental methods. A bioinformatics workflow was developed and evaluated using a combination of fast and user-friendly subcellular prediction programs to maximize performance and accuracy for chloroplast transit peptides and demonstrate this technique on the predicted proteomes of 15 sequenced plant genomes. Gene family grouping was then performed in parallel using modified approaches of reciprocal best BLAST hits (RBH) and UCLUST. Between 628 protein families were found to have conserved plastid targeting across angiosperm species using RBH, and 828 using UCLUST.

However, thousands of clusters were also detected where only one species had predicted plastid targeting, most notably in *Panicum virgatum* which had 1,458 proteins with species-unique targeting. An average of 45% overlap was found in plastid-targeted gene families compared with *Arabidopsis*, but an additional 20% of proteins matched against the full *Arabidopsis* proteome, indicating a unique evolution of plastid targeting. Neofunctionalization through subcellular relocation is known to impart novel biological functions but has not been described before on genome-wide scale for the plastid proteome. Further work to correlate these predicted novel plastid-targeted proteins to transcript abundance and high-throughput proteomics will uncover unique aspects of plastid biology and shed light on how the plastid proteome has evolved to change plastid morphology and biochemistry.

Introduction

Plastids represent biochemically and morphologically complex organelles in the plant cell and can change both form and function drastically in response to developmental and environmental cues. A vestigial genome of between 120-160 kb harboring 90 protein-coding genes is present in the plastids of photosynthetic higher plants (Sugiura, 1992). However, the total chloroplast proteome conservatively contains 2,000-3,500 proteins as demonstrated in *Arabidopsis* (Armbruster et al., 2011; Millar et al., 2006; Richly and Leister, 2004), but as many as 4,875 plastid-targeted proteins are estimated in eSLDB (Pierleoni et al., 2007), and 5,136 by the Chloroplast 2010 project (Ajjawi et al., 2010; Lu et al., 2011; The *Arabidopsis* Genome Initiative, 2000). In *Arabidopsis*, there is evidence to support that roughly 4,500 genes and 1,392 gene families are derived via horizontal gene transfer from the ancestral cyanobacteria, but less than 900 are predicted to be plastid-localized (Martin et al., 2002). Only 21% of plastid-targeted

rice proteins have a predicted homolog in the predicted Arabidopsis plastid proteome, while only 38% is valid for the reciprocal comparison (Richly and Leister, 2004). A similar result was obtained in a comparison of six crop plants against Arabidopsis, in which an average of 51.0% of the predicted plastid proteome of each species matched sequences to the Arabidopsis predicted plastid proteome, while 67.5% matched against the full Arabidopsis proteome (Schaeffer et al., 2014). Thus, most of the plastid pan-proteome is not conserved, but experiences fluctuation in individual species. Furthermore, as the number of shared sequences closely mirrors the number of genes of cyanobacterial origin, it is likely that non-conserved plastid-targeted proteins evolved primarily from Eukaryotic sequences. The variability in the predicted plastid proteome mirrors the observable diversity in plastid function and ultrastructure in different species and under different conditions. The diversity of plastid proteomes is evident even within the same plastid morphotype: the pigment-storing chromoplast alone has at least four described ultrastructural phenotypes across various species with unique suborganellar membrane structures that can occur either singly or mixed within individual plastids (Li and Yuan, 2013). Morphological differences in plastid shape and ultrastructure are noted even in genetically similar cultivars of the same species. Both chloroplasts and chromoplasts of developing apple peel differ significantly from tomato, which is used as a model reference for chromoplast differentiation in fruits (Barsan et al., 2012; Egea et al., 2010). Variation has also been documented between the apple cultivars and the epidermal and collenchymal plastids (Schaeffer et al., 2017).

The observed phenotypic differences could be explained based on three potential molecular factors: 1.) Differences in the expression of genes controlling the rate and total amount of protein accumulation. This aspect could lead to unique phenotypes without necessarily changing the subset of plastid-targeted proteins; 2.) Mutations within a shared group of plastid-

targeted proteins could lead to neofunctionalization, and 3.) Gain or loss of transit peptides causing subcellular mistargeting could alter the total pool of plastid-targeted proteins. These factors are not mutually exclusive, and examples of each mechanism are known. Gene expression differences, possibly caused by epigenetic DNA methylation patterns, are responsible for differential protein accumulation in mesophyll and bundle sheath cells of C4 plants, illustrating the first point (Majeran, 2005; Majeran et al., 2008; Ngernprasirtsiri et al., 1989; Stockhaus et al., 1997). In support of the second mechanism, point mutations in the active site of plastid-targeted limonene synthase change the abundance and distribution of different monoterpenoid end products in bacterial expression systems (Srividya et al., 2015), and transplastomic expression of a delta-9 desaturase gene causes changes in fatty acid concentrations and levels of unsaturation, cold tolerance, leaf senescence, and seed yield (Craig et al., 2008) are additional examples. While it is challenging to address the neofunctionalization of plastid-targeted proteins via mutation without detailed reverse genetics experiments, the other mechanisms can be evaluated with high-throughput sequencing and bioinformatics.

High-throughput proteomics using mass spectrometry (MS) has been an important means of surveying organellar proteomes and comprises the majority of current plastid proteome evidence. However, these techniques have historically been limited to the chloroplast morphotype and a restricted number of plant species. Excellent databases for high-throughput plastid proteomes based largely on mass spectrometry are accessible at AT_CHLORO (Ferro et al., 2010), PPDB (Sun et al., 2009), SUBA4 (Hooper et al., 2017), and CROPPAL (Hooper et al., 2015). However, caution should be exercised in interpreting these datasets because MS is susceptible to high false positive errors due to contamination during plastid isolation, liberal mass tolerance, and errors in peptide mapping, among other problems (Jeong et al., 2012;

Nesvizhskii, 2010; van Wijk and Baginsky, 2011). While the use of reference genomes and transcriptomes can help overcome peptide mapping issues, other technical issues are more difficult to resolve. Use of fluorescent protein chimeras (e.g., GFP – green fluorescent protein), though lower-throughput, typically have higher biological accuracy. Using these, localization of low-abundance, as well as proteins from species lacking robust plastid isolation methods, can be evaluated with higher efficiency. However, GFP techniques are not immune to experimental error either. Since the sequence of the mature protein partially influences localization (e.g., (Doyle et al., 2013; Lisenbee et al., 2003; Small et al., 1998)), GFP fused to the native protein may alter localization in some cases. Furthermore, dual-targeted mitochondrial/chloroplast proteins can be mislocalized in GFP assays (Carrie et al., 2009). Alternative transcripts or alternative protein products may also produce differential subcellular localization that are either not captured in GFP assays or give ambiguous results. Given these experimental limitations, a robust bioinformatics workflow could enable rapid and cost-effective assessment of plastid proteomes with somewhat comparable accuracy. Though wet lab validation is still necessary, these datasets could narrow the focus to smaller subsets of proteins of interest which could be more manageably targeted for wet lab validation depending on the biological question being asked.

The semi-conserved and sometimes ambiguous nature of chloroplast transit peptides makes *in silico* predictions challenging. However, sequence- and annotation-based approaches have yielded results with significant accuracy. Protein sequence-based prediction uses the amino acid content or the presence of conserved motifs in the peptide to make predictions. Use of the amino acid content alone, such as in the tool PCLR, is enough to predict many plastid-targeted proteins (Schein et al., 2001). More complex sequence-based identify conserved motifs, such as

in iPSORT (Bannai et al., 2002) and WoLF-PSORT (Horton et al., 2007), or sliding-window searching algorithm such as Localizer (Sperschneider et al., 2017), make predictions based on the sum of prediction vectors to determine transit peptide similarity. Finally, tools that use neural networks such as ChloroP (Emanuelsson et al., 1999), TargetP (Emanuelsson et al., 2007, 2000), Predotar (Small et al., 2004), PredSL (Petsalaki et al., 2006), and Protein Prowler (Bodén, 2014) use multiple layers of nodes to identify the best-scoring localization. In contrast, annotation-based methods such as CLPFD (Chou and Cai, 2002) and EpiLoc (Brady and Shatkay, 2008), or simple text-based methods based on GO annotations (Fyshe et al., 2008), use homology to proteins with known localization to designate subcellular predictions. While these methods offer advantages over sequence-based methods for proteins with annotated homologs, they perform poorly for novel proteins (Xiong et al., 2016). Hybrid approaches including MultiLoc2 (Blum et al., 2009), Sherloc2 (Briesemeister et al., 2009), Y-Loc (Briesemeister et al., 2010), and Plant-mPLoc (Chou and Shen, 2010) combine sequence- and annotation-based methods in an attempt to overcome this limitation. Unfortunately, the homology component of hybrid approaches is weighted more heavily, which can lead to the false prediction of proteins with transit peptide variation or for proteins with shared domains. Both high-throughput proteomics and bioinformatics approaches consistently indicate that the plastid proteome content is highly dynamic and likely has significant variability across the plant kingdom. With newer methods, ever-growing genomic resources, and availability of better gene annotation methods, earlier estimates of conserved and non-conserved sets of the plastid proteome warrant an update.

This study evaluated the hypothesis that bioinformatics methods could achieve similar accuracy to experimental methods by comprehensively testing previously published subcellular prediction algorithms both alone and in combination. A specific combination of methods was

found to be most efficient, which was then used to globally predict nuclear-encoded plastid-targeted proteins for fifteen higher plant species including eight eudicots, six monocots, and *Amborella trichopoda*, an early diverging species of the angiosperm clade. Two parallel approaches, Reciprocal-Best Blast Hit (RBH) and UCLUST (Edgar, 2010) were used to perform clustering, and the sub-cellular localization prediction for each cluster was analyzed to identify conserved, semi-conserved, and non-conserved plastid-targeted proteins. This approach evaluated the hypothesis that a relative minority of plastid-targeted genes are conserved among all species. It was found that natural selection and environmental influence has shaped the development of species-specific plastid proteomes.

Results and Discussion

Identification of Optimal Subcellular Prediction Workflows

To test the hypothesis that a bioinformatics workflow could reach parity with experimental methodology, the accuracy of six subcellular prediction algorithms including TargetP (Emanuelsson et al., 2000), WoLF PSORT (Horton et al., 2007), PredSL (Petsalaki et al., 2006), Localizer (Sperschneider et al., 2017), Multiloc2 (Blum et al., 2009), and PCLR (Schein et al., 2001) was first evaluated using data from the original publications. Sensitivity, specificity, accuracy, and Matthew's Correlation Coefficient (MCC) were evaluated for each program as it related to the prediction of plastid-targeted proteins (Table 1). Sensitivity, specificity, and MCC in TargetP were found to exactly match the values reported by Emanuelsson et al. (2000, 2007) and while minor differences were found for MultiLoc and PredSL, these discrepancies likely represent rounding errors. Unexpectedly, significant differences were found for PCLR and Localizer: in PCLR, sensitivity was found to be 52.1%,

which was about 5% lower than what was reported (Schein et al., 2001). In Localizer, calculated specificity was 78.9%, nearly 16% lower than the 95.7% reported (Sperschneider et al., 2017). In both cases, all other performance statistics were identical or nearly identical, so it is likely that the discrepancies in Localizer and PCLR represent either miscalculations or transcriptional errors in the data used for analysis in the original publications.

Next, cross-validation of subcellular prediction programs was performed against proteins with experimentally-determined subcellular localization retrieved from AT_CHLORO (Ferro et al., 2010), PPDB (Sun et al., 2009; Van Wijk, 2004), CropPAL and CropPAL2 (Hooper et al., 2015) and Suba4 (Heazlewood, 2005; Heazlewood et al., 2007; Hooper et al., 2017, 2014), resulting in 42,761 nonredundant sequences including 32,450 proteins validated by mass spectrometry (MS) and 3,722 validated by GFP. Most prediction algorithms were found to have lower performance against biological data than reported in the original reports, as shown in Table 2 and Figure 1. However, substantial differences were observed based on the method of experimental validation. On average among the six algorithms, sensitivity was 15.7% higher in the GFP-validated dataset while no significant change in specificity was found; this difference resulted in 10% higher overall accuracy and an increase of 0.159 in MCC for GFP-validated proteins. By further narrowing focus to a dataset of proteins validated by both methods, sensitivity increased by an additional 7.6%, and specificity increased 2.5%, on average. Due to the previously reported high false positive rates associated with shotgun proteomics of organellar proteomes (Nesvizhskii, 2010; van Wijk and Baginsky, 2011), program performance was expected to be much higher for GFP-validated proteins. While the dataset containing proteins experimentally validated by both GFP and mass spectrometry showed the highest apparent performance for the six subcellular prediction algorithms - and is likely closer to the biological

accuracy of these programs - it contains roughly a third as many proteins as the GFP-validated dataset and is heavily biased by Arabidopsis sequences. Therefore, remaining comparisons focused on the GFP-validated dataset. Similarly, MCC was used as the primary measure of biological accuracy of *in silico* approaches to avoid problems due to drastically different dataset sizes.

Overall, the highest-performing program in terms of MCC was Localizer, followed by MultiLoc2-HR, TargetP, PCLR, PredSL, WoLF PSORT, and MultiLoc2-LR. Of these, PredSL and MultiLoc2-LR performed poorly with GFP-validated proteins compared to the original reports, while other programs decreased marginally or performed similarly to the published MCC. Among the six programs that were evaluated, Localizer had the highest performance regardless of the experimental method used for validation, which is surprising since it is a simpler tool than annotation-based methods which have been at the forefront of subcellular prediction methods recently. Part of Localizer's increased accuracy may be due to its unique capacity to predict dual-targeted mitochondrial/chloroplast proteins. Over 200 dual-localized proteins have been described in Arabidopsis (Carrie and Small, 2013) and over 500 are predicted to have ambiguous transit peptides (Mitschke et al., 2009). Increased accuracy in the prediction of these sequences in Localizer could alone account for a portion of its higher performance. After Localizer, MultiLoc2 had the next-highest MCC and also had the highest specificity of any program, at 83% in GFP-validated proteins. MultiLoc is a hybrid method combining annotation and sequence analysis, so these findings support that the use of hybrid methods yields robust biological specificity. However, MultiLoc also had the worst sensitivity of any program, correctly predicting only 50% of bonafide plastid-targeted proteins validated by GFP or 31% of sequences validated by either GFP or mass spectrometry. TargetP, which has historically been

the most popular subcellular prediction program for plants since its introduction, was found to perform at lower accuracy than earlier estimates: even when using the more conservative GFP-validated data, specificity was only 59% and sensitivity was 67%. Previous experiments using high-throughput shotgun proteomics have reported that the sensitivity of TargetP is as low as 62% (Armbruster et al., 2011; Bhattacharya et al., 2007; Kleffmann et al., 2004; von Zychlinski et al., 2005). Use of strictly-curated data improves the apparent sensitivity up to 86%, but false positive rates are still problematic as a specificity of about 65% is observed (Zybaïlov et al., 2008). The results presented here suggest that the biological accuracy of TargetP is somewhat closer to the initial estimates on non-curated data. PredSL, PCLR, and WoLF-PSORT were the lowest-ranked programs by MCC for prediction of plastid-targeted proteins, in that order, but typically had higher sensitivity than Localizer or MultiLoc2.

Differences in the amino acid composition of transit peptides are observable between rice and Arabidopsis, which have an overrepresentation of alanine and serine, respectively (Zybaïlov et al., 2008). Therefore, differences in the prediction of monocot or eudicot sequences were assessed, and different programs displayed significant bias (Table 3). PCLR was the most drastically affected, with an MCC bias of +0.091 in monocots, representing a roughly 20% increase compared with eudicots. This finding is somewhat unsurprising because PCLR is the only program which uses sequence composition alone to make predictions and is, therefore, more susceptible to bias than motif- or annotation-based methods. TargetP was the only other tool that favored monocots, with an increase of 0.055 (+10.2%) in MCC. A marginal difference between monocot and eudicot prediction was observed when Localizer was used, which differed by only 0.008 in MCC, slightly favoring eudicots. Eudicot sequences were favored in the other prediction programs, with between 0.043 (+10%) higher MCC in WoLF POSRT and 0.066 in

PredSL (+14.9%). To the best of our knowledge, this is the first study to report this type of error or bias for *in silico* prediction methods. Some differences have also been described for the proposed subunits of the TIC translocon in grasses, which could result in coevolution of the transit peptide sequence composition (de Vries et al., 2015; Nakai, 2018, 2015). Choice of training and cross-validated datasets could significantly sway the predictions of sequence-based methods, while overrepresentation or prioritization of sequences for Arabidopsis and thereby eudicots could introduce bias to annotation-based methods. Although these species-specific differences are smaller than differences observed for sequences validated by mass spectrometry compared with GFP, they are still noteworthy and have consequences for whole-genome prediction. In contrast, WoLF-PSORT and Localizer were found to have insignificant if any bias, making them attractive both as standalone programs or in combinatorial approaches where they could mask biases of other programs.

Combinatorial Approaches Outperform Single Programs:

Use of multiple prediction algorithms in combination is a powerful strategy to combine the strengths and overcome the limitations of single programs. Combinatorial approaches have been used to improve the accuracy of predictions in whole-genome analyses (e.g., Richly and Leister, 2004) or to curate mass spectrometry data (e.g., Barsan et al., 2010; Zeng et al., 2014, 2011; Zhu et al., 2018). Additionally, a combinatorial workflow using 22 prediction algorithms and four experimental techniques is used in the SUBAcon algorithm implemented for the SUBA4 database of Arabidopsis proteins which reportedly yields up to 97.5% accuracy for chloroplast localization and 90% for other compartments (Hooper et al., 2017, 2014). While SUBAcon does not strictly require experimental data to perform predictions, available evidence weighs heavily on the final prediction and contributes to the reported accuracy. Even if

experimental evidence were to be ignored, the use of 22 separate subcellular prediction algorithms is not feasible for individual researchers or application to enormous datasets. Therefore, a bioinformatics-based workflow that can work efficiently would be desirable.

Calculations were performed for each possible permutation of subcellular prediction algorithms and for all possible acceptable thresholds for each combination as applied to GFP-validated proteins. For example, for the combination of TargetP, PredSL, and Localizer, three thresholds were tested in which one, two, or all three programs needed to predict plastid localization to consider that protein as having a plastid transit peptide. To simplify analyses, the poorly-performing WoLF PSORT was removed from consideration (results including WoLF PSORT and datasets including MS-validated proteins are available in Additional File 2). In total, 80 unique workflows including the five remaining standalone program workflows were evaluated against GFP-validated proteins, the results of which are graphically summarized in Figure 1, and numerically ranked by MCC in Table 4. Unequivocally, the results demonstrate that combinations of programs tend to outperform single programs for GFP-validated data: among the 25 workflows with the highest MCC, 23 were combinatorial approaches, while the standalone Localizer ranked tenth and Multiloc2-HR 22nd. Localizer was not only the best-performing standalone program but was also overrepresented in combinatorial workflows: except the standalone Multiloc2-HR workflow, Localizer appeared in all 25 top-performing workflows. It is interesting to note that combinations that rank higher tend to combine programs with high sensitivity with counterparts that have lower sensitivity but higher specificity, thus correcting for each other's deficiencies. Specifically, most of the combinations with the highest MCC and ACC tend to include Localizer most often, followed by MultiLoc2, TargetP, PCLR, and lastly PredSL. The ranking of Localizer is unsurprising given that its relatively balanced and high sensitivity

and specificity are unparalleled by any of the other programs. However, MultiLoc2's extremely high specificity makes it a valuable component of many workflows despite its low sensitivity. The best performing workflow used TargetP, Localizer, and Multiloc2 and required 2 of the three programs to predict plastid targeting to define a sequence as containing a plastid transit peptide; specificity of 78.5%, the sensitivity of 64.6%, and MCC of 0.659 was achieved with this approach. In comparison to TargetP alone, a nearly 20% increase in specificity was observed with no loss in sensitivity. However, as the annotation-based functions of MultiLoc2 make it difficult to run on extensive datasets, an alternative workflow using a "2 of 2" consensus approach for TargetP and Localizer was found which ranked 2nd and achieved a marginally higher specificity of 80.7%. Furthermore, comparing the accuracy of the best workflows to Table 2 and to prior evaluations of experimental methodology (e.g., Zybilov et al. (2008)) supported the hypothesis that bioinformatics methods could reach parity with mass spectrometry in characterizing the plastid proteome. Due to the increased simplicity and comparable performance of the TargetP/Localizer consensus approach, this workflow was selected for subsequent genome-scale prediction of plastid-targeted proteins.

Predicted Plastid Proteome Correlates with Genome Size

As a demonstration of the utility of the Localizer and TargetP workflow, subcellular prediction was performed for the whole proteomes of fifteen phylogenetically diverse species. Six monocot species, including *Anthurium amnicola*, *Brachypodium distachyon*, *Oryza sativa*, *Panicum virgatum*, *Setaria italica*, and *Sorghum bicolor* and eight eudicots, including *Arabidopsis thaliana*, *Fragaria vesca*, *Glycine max*, *Malus × domestica*, *Populus trichocarpa*, *Prunus persica*, *Solanum lycopersicum*, and *Vitis vinifera* were chosen. Additionally, *Amborella trichopoda*, a species which diverged from the rest of the angiosperms prior to the divergence of

monocots and eudicots, was also incorporated into the comparative analysis. Complete information including data version numbers, proteome sizes, and prediction of plastid-targeted proteins by Localizer and TargetP is summarized in Table 5. In Arabidopsis, 2,826 proteins were predicted to be plastid-targeted, representing 8.8% of all protein isoforms. This finding is in agreement with the conservative estimates of the Arabidopsis plastid proteome (Li and Chiu, 2010; Millar et al., 2006; Richly and Leister, 2004). Similar percentages were calculated in other species but varied from a low of 6.4% in tomato to a high of 9.3% in *A. amnicola*. As expected, the absolute number of predicted plastid-targeted genes showed a high correlation with the genome size ($R^2=0.965$) (Figure 2). This result suggests that an increase in genome size and gene content yield a similar increase in the total number of plastid-targeted proteins. Over 10,000 of the Arabidopsis sequences have experimentally-determined localization, and comparing predictions for these sequences revealed an apparent sensitivity of 55.6%, specificity of 89.8%, accuracy of 83.6%, and MCC of 0.614. Sensitivity is somewhat low in this estimation due to the use of MS data, which includes many false positives, but the high specificity suggests good prediction accuracy. With the combination of the high correlation with experimentally-validated proteins and the lack of monocot/eudicot bias imparted by Localizer, it is expected that similar levels of accuracy were achieved for the entire set of species analyzed in this study.

Clustering of Gene Families

Although the plastid is highly dependent on proteins imported from the nucleus for normal viability and function, the size and diversity of the plastid proteome across the plant kingdom remain poorly understood. The hypothesis that the plastid proteome is diverse and each species has a unique set of plastid-targeted proteins was examined by grouping sequences into homologous protein groups using two parallel clustering methods (Figure 3). Clustering method

has a significant impact on the size and accuracy of the resulting clusters, and therefore on the number and relevance of predictions. Reciprocal best BLAST Hits (RBH) using ALL-v-All BLAST comparisons of whole proteomes are a standard proxy for orthology in comparative genomics, although they are susceptible to inclusion of weakly homologous paralogs. BLAST-based approaches combined with Markov clustering or similar methods to remove paralogs are used in commonly-cited methods such as InParanoid (O'Brien et al., 2005), OrthoMCL (Li et al., 2003), and COG (Tatusov et al., 2000, 1997). However, these methods can bias single-copy genes or highly conserved families which is problematic for polyploid genomes where many-to-many gene relationships are common (Das et al., 2016; Trachana et al., 2011). For instance, the popular OrthoMCL fails to detect many homologous proteins with conserved expression patterns, and therefore with likely conserved functions, between rice and Arabidopsis (Kim et al., 2011; Van Bel et al., 2012). In contrast, more straightforward RBH methods often outperform more complicated algorithms on eukaryotic genomes (Altenhoff and Dessimoz, 2009).

A simplified RBH approach, allowing many-to-many relationships, was determined to be most appropriate for this analysis to avoid fracture of gene families with paralogs or co-orthologs. Initial homologous relationships were identified using pairwise BLAST-P comparisons of two genotypes; only sequences which are mutually the best BLAST hits for each other were utilized. Similar methods have used 40% as an appropriate identity and coverage threshold for orthologous relationships (Chiu et al., 2006; Sanderson and McMahon, 2007; Schaeffer et al., 2014; Yang and Smith, 2014). Therefore 40% was used as the initial threshold of homology. Initial clustering generated many small clusters, so a supplemental method for expansion of clusters, using reciprocal better BLAST hits of each species' proteome BLAST'ed against itself, was tested (Additional File 1). A 90% threshold was determined to be optimal for

clusters with fewer genotypes decreasing significantly in number, while clusters containing a majority of genotypes remained stable or increased. In contrast, application of between 60 and 80% expansion thresholds caused the liberal merging of clusters into extremely large clusters representing thousands of individual sequences. Additionally, GO term similarity was assessed within clusters at each population size based on the number of species in the cluster and was found to increase slightly for clusters containing few species when using a 90% expansion threshold, while more massive clusters experienced no change or slight decreases.

An alternative approach called UCLUST was implemented to complement the RBH method with a faster and more efficient technique because its semi-global algorithm detects homology in a fraction of the time required for BLAST and becomes much more efficient on enormous datasets. Initial clusters were constructed at a 40% identity and 40% coverage threshold similar to the RBH approach. However, initial clustering produced smaller clusters and resulted in cluster fragmentation. Therefore, modifications were implemented to expand initial clusters by randomly selecting sequences out of each initial cluster and iterating the UCLUST search at more stringent conditions using the selected sequences as new centroids (Additional File 1). Cluster expansion significantly increased the number of clusters with many species, which largely came from the drastic reduction of the number of single-species clusters. As with RBH, a 90% expansion threshold was found to be optimal and increased the number of clusters sharing 14-15 species roughly 4-fold, while lower thresholds resulted in the frequent grouping of nonhomologous sequences. Comparison of GO similarity for clusters containing multiple species showed that similarity increased slightly or remained stable for nearly all cluster sizes in the 90% expansion threshold compared to the initial, non-expanded UCLUST analysis. The number of iterations required to fully expand cluster space in UCLUST was also examined, and it was

found that most clusters were completely expanded by ten iterations, while further iterations yielded diminishing returns (Additional File 1). A total of 100 iterations were performed to avoid problems with the randomization of centroid sequences.

Application of the optimal clustering methods to the proteomes of the species chosen generated 170,877 clusters using RBH (Table 6) and 103,501 clusters using UCLUST (Table 7). Nearly all the additional clusters in RBH were from single-species clusters or singleton sequences (data not shown): 150,067 of the RBH clusters (87.82%) were single-species clusters of which 134,319 were singleton sequences, while UCLUST detected 74,059 single-species clusters (71.55%) including 45,033 singletons. Some of these may be orphan genes, but they are more likely to be prediction and annotation errors or pseudogenes because the lack of homology implies lack of conserved function or extreme mutation rates that are more likely to occur in non-coding sequences. A total of 20,810 and 29,442 clusters in RBH and UCLUST approach, respectively, contained sequences from multiple species; while they represented a minority of clusters, they contained the majority of initial sequences. A bimodal distribution was observed in both methods in which two clusters, the first containing 14-15 of the species and the second containing just 2-3 species, represented the majority of the clusters (Figure 3A). Comparatively fewer clusters contained between 4-13 species. Of the conserved clusters containing all 15 species, RBH detected 4,090 clusters, while UCLUST yielded 3,295. GO similarity between UCLUST and RBH was remarkably consistent, but UCLUST had somewhat better scores for conserved clusters containing plastid-targeted sequences from all species and lower scores for semi-conserved or non-conserved clusters containing few species (Figure 4B). Across both methods, GO similarity decreased with increasing cluster size. While the merging of nonhomologous sequences may be partially responsible for this decrease, the annotation methods

for all genomes are not identical, which artificially decreases the apparent similarity score regardless of clustering specificity.

Identification of Gene Families with Conserved Plastid Targeting

Genomes of endosymbiotic bacteria contain 1,500 proteins on average, and plastids are likely to contain similar numbers when accounting for both the plastid genome and core nuclear-encoded plastid-targeted genes (Hönigschmid et al., 2018). To determine the number of gene families with conserved plastid localization, clusters containing at least 13 species, of which all species contained at least one predicted plastid-targeted sequence or at least four non-plastid-targeted sequences were selected. These parameters were chosen to account for assembly and annotation errors and to correct for the 39% false negative prediction rate for bonafide plastid-targeted proteins which could eliminate many truly conserved clusters. There is a nearly 20% chance that at least one of four random sequences with non-plastid localization prediction is a false negative, but sequences that already share homology to predicted plastid-targeted sequences have a significantly higher likelihood of being false negatives. A workflow diagram representing cluster detection, filtering, processing, and categorization is represented in Figure 4. Applying this workflow, 628 conserved protein clusters were found in RBH (Table 6, Figure 5), while UCLUST detected 828 (Table 7, Figure 6). Of these, 621 clusters in RBH and 817 in UCLUST also contain sequences from *A. trichopoda*, and all have several monocot and eudicot sequences, strongly indicating that these clusters represent the fundamental core plastid-targeted gene families. Previous estimates predicted that 857-1020 sequences were shared between rice and Arabidopsis, another report projected that between 289-737 proteins were shared among the chloroplast proteomes of seven plant species (Richly and Leister, 2004; Schaeffer et al., 2014). Identification of gene families with conserved chloroplast transit peptides is an essential output

of this work, as *in silico* methods can quickly identify conserved plastid-targeted proteins that have failed to be detected by genetic screens due to embryo lethality, gene redundancy, or random chance. Several methods have validated these sequences as truly plastid-targeted and representative of conserved plastid-targeted genes. First, Arabidopsis proteins with experimentally-validated localization were examined within the conserved clusters. A total of 84.2% (183 proteins) of predicted plastid-targeted Arabidopsis sequences in conserved RBH clusters were validated by GFP and 94.5% (1,054) were validated by MS. The same was true for 80.5% (154 proteins) and 92% (855 proteins) in conserved RBH and UCLUST clusters, respectively (Additional Files 4 and 5). While these methods have yielded good overall sensitivity, small errors at initial stages of clustering can compound in larger clusters and result in unrealistically high numbers of sequences. For RBH, an average of 113.9 sequences and median of 61 were present in conserved clusters while UCLUST produced an average of 125.9 sequences and median of 84. Most sequences in these clusters come from a small set of genotypes: *G. max*, *P. virgatum*, *P. trichocarpa*, and *V. vinifera* each contributed an average of over 10 sequences each to clusters with shared plastid localization prediction, while *M. × domestica* contributed over 10 sequences on average in UCLUST (summarized in Additional Files 4 and 5). Significant gene duplication or inclusion of multiple gene isoforms especially in those species likely accounts for a portion of the larger cluster sizes, but more distantly paralogous sequences which are less likely to share biological function are also likely to be common. Thus, the list of conserved clusters reported here is not meant to be definitive and final, but rather a general guide which will require phylogenetic and experimental validation. In cases where larger clusters contain multiple paralogs or non-homologous, phylogenetic methods could resolve homology relationship with higher efficiency than the currently used RBH and UCLUST

methods. However, the biological accuracy of the predicted plastid-targeted sequences within these clusters is still high.

Next, enrichment of gene ontology (GO) annotations was performed in conserved clusters by finding GO terms shared in at least three individual sequences and for over 10% of sequences. Terms were compared to annotations extracted using the same criteria for all the clusters of the respective clustering method and GO term enrichment was performed using BLAST2GO (Conesa and Götzt, 2008). Overall, 53 terms including 29 terms associated with biological function, 23 associated with the cellular component, and one associated with the molecular process were found for RBH (Table 8). In UCLUST, a total of 33 terms were found, including 15 associated with the biological process, 17 with the cellular component, and one with the molecular process (Table 9). The most significantly enriched GO terms under the biological process ontology for both RBH and UCLUST methods were GO:0015979 (photosynthesis) and GO:0008152 (metabolic process), while a majority of the remaining highly enriched terms were associated with homeostatic processes (GO:0042592), cellular component organization (GO:0016043), single-organism biosynthetic processes (GO:0016043), generation of precursor metabolites (GO:0006091), and lipid metabolism (GO:0006629). In the RBH method, additional terms associated with amide, peptide, and organonitrogen compound biosynthesis and metabolism (GO:0043604, GO:0043603, GO:0043043, GO:0006518, GO:1901566, GO:1901564, GO:0044271, GO:0034641, GO:0006807), were enriched. UCLUST additionally had enriched GO terms associated with transport (GO:0006810), localization (GO:0051234, GO:0051179) and metabolism of carbohydrates (GO:0005975). Among cellular component ontologies, plastid (GO:0009536) was the most overrepresented term in both methods. Other highly overrepresented cellular component terms included organelle (GO:0043226), thylakoid

(GO:0009579), chloroplast (GO:0009507), and associated terms. In RBH methods, significant enrichment of ribonucleoprotein complexes (GO:1990904, GO:0030529) was found. For the molecular process ontology, structural molecule activity (GO:0005198) was enriched in RBH and catalytic activity (GO:0003824) in UCLUST. These GO terms were further compared to the results of a previous study involving intergeneric analysis that described 737 conserved plastid-targeted proteins (Schaeffer et al., 2014). In this study, 42% of enriched terms found using UCLUST overlapped with the methods reported previously (Schaeffer et al., 2014). RBH methods were somewhat lower because more enriched terms were found, but still overlapped with the previously published dataset by 24%. These results are remarkably similar given that only GO terms from Arabidopsis had been examined previously and also different methods of GO enrichment had been used in those studies. The final and perhaps the most important test of the biological significance of conserved plastid-targeted clusters is whether they contain proteins expected to be present in plastids of all higher plants. Gene names were retrieved from TAIR10 for all Arabidopsis sequences in conserved clusters, and many of the most prominent plastid proteins were confirmed to be present in clusters for both RBH and UCLUST methods. The following is not intended to be an exhaustive list but merely a representative of the types of proteins detected in conserved plastid-targeted clusters; a complete list of annotations and gene names in RBH and UCLUST clusters are available in Additional Files 4 and 5. Among genes involved in primary photosynthesis, HCEF, LhcA1, LhcA2, LhcB1, LhcB2, LhcB3, Lhcb4, LPA1, LPA3, PPDK, and RbcS were detected in both methods, while LPA66 was found in RBH only. Photosystem subunits Psa-E, Psa-F, Psa-G, Psa-H, Psa-K, Psa-N, PsbP, Psa-O, PsbQ, PsbR, PsbS, PsbW, and PsbY were also found in both methods, while PsbT-N and PsbX were found only in RBH and PsbO was found only in UCLUST. Among ribosomal proteins, Rps1,

Rps9, Rpl4, Rpl11, and Rpl12 were detected by both techniques, while Rpl9 and Rpl15 were only found using RBH and Rpl10 was found only with UCLUST. Proteins involved in translocation and chaperone functions found by both methods included ClpB, ClpC, ClpD, ClpP, ClpR, FtsH, Hsp60, Hsp70, Hsp88, Hsp90, Hsp98, Cpn10, Cpn20, Cpn60, Vipp1, Alb3, Alb4, TatC, Tic20, Tic21, Tic40, Tic55, Tic110, Toc75, and Plsp1. The Sec translocase subunits SecA, Scy1, and Scy2 were uniquely found in RBH, while organellar oligopeptidase OOP was also found in UCLUST. Finally, genes associated with primary plastid metabolism (SBPase, TPT, FRUCT5, G6PD2, and G6PD), heme biosynthesis (GUN2, GUN5, HEMA, HEMB, HEMC, HY2, PORA, PORB, and PORC), and fatty acid synthesis (ACC2, FAB2, FAD7, FAD8, FATA, FATB, lipooxygenase) were found in core clusters.

Taken together, the good correlation of protein clusters with experimentally-validated sequences, the enrichment of expected annotation terms, and the presence of expected highly-abundant proteins or proteins critical to chloroplast biology suggest that both the RBH and UCLUST methods achieved good accuracy and sensitivity for genes with conserved chloroplast targeting which are likely critical in all photosynthetic plants for minimal chloroplast function. It is noteworthy that 194 clusters in RBH and 333 core clusters in UCLUST contain at least one Arabidopsis sequence but have no associated gene synonyms available (Additional Files 4 and 5). As the sensitivity for conserved plastid-targeted proteins was found to be very high overall, many of these 194-333 clusters with missing annotation information are likely biologically accurate, in which case they are excellent candidates for understanding hitherto uncharacterized aspects of chloroplast biology.

Analysis of Semi-Conserved and Non-Conserved Plastid-Targeted Proteins

Semi-conserved plastid-targeted gene families in which predicted plastid-targeting was found for two or more sequences only in monocots, only in eudicots, or uniquely in *A. trichopoda* were identified beginning with the most diverse clades. In each case, all clusters with predicted plastid-targeted sequences or at least four predicted non-plastid-targeted sequences from the outgroup species were removed. A total of 572 gene families with plastid-targeted sequence specific to monocots and 430 to eudicots were found using RBH methods (Table 6, Figure 5), while UCLUST detected 1,054 and 885, respectively (Table 7, Figure 6). Additionally, 82 clusters with *Amborella*-specific plastid targeting were found using RBH, and 195 were found with UCLUST. These findings indicate that gene families with semi-conserved plastid-targeting outnumber core clusters by 73% in RBH and more than 150% in UCLUST. Narrowing focus to the subclade and family level revealed that semi-conserved clusters are still abundant, indicating that significant plastid proteome variation is present across all taxonomic levels. It is plausible that some of the clusters with plastid-targeting specific to either monocots or eudicots have functionally related clusters in the reciprocal group but lack sufficient homology to cluster together. Such an occurrence seems unlikely in most cases because the clustering methods used here were relatively liberal, but isolated cases may still occur. In some cases, non-orthologous or chimeric genes could also functionally replace an otherwise conserved gene and lead to loss of orthologous sequences in particular species or taxonomic groups (Koonin et al., 2000; Osterman and Overbeek, 2003).

Finally, clusters with predicted plastid targeting only present in a single species were identified in RBH (Table 6, Figure 5) and UCLUST (Table 7, Figure 6). Singletons and clusters containing only a single genotype were discarded as these likely represent gene prediction errors.

For example, predicted proteins in *Malus* which do not share homology with proteins in other species are typically poorly-supported by transcriptomics evidence: examination of over 300 such sequences revealed only one that had full coverage and was not a smaller fragment of a larger protein (data not shown). Since the chloroplast transit peptide is presumed to have arisen recently in each cluster, the term “nascent plastid-targeted proteins” (NPTPs) was coined to represent such proteins. Unsurprisingly, genotypes with large and complex genomes possessed a more significant number of NPTPs: *A. amnicola* had the least, at just 52 in RBH and 97 in UCLUST, while *P. virgatum* had the most, with 682 NPTPs found in RBH and 1,458 in UCLUST. The predicted proteome of *A. amnicola* is based on transcriptomics data rather than genome-wide prediction, while *P. virgatum* has the largest genome and most extensive predicted proteome of the species in this analysis, so these trends are consistent with expectations.

Additionally, up to 728 proteins were uniquely targeted to the plastid in *M. × domestica*, and between 300-400 proteins had species-specific plastid transit peptides in *B. distachyon*, *F. vesca*, *G. max*, *S. italica*, and *S. bicolor*. Arabidopsis had some of the lowest estimates of NPTPs, with only 74 found in RBH and 166 in UCLUST. Species-unique plastid-targeted proteins had a moderately linear correlation with the total number of sequences in each species $R^2=0.73$ in RBH and 0.72 in UCLUST, Figure 7A), but the removal of the outlier *P. virgatum* resulted in nonlinear correlation (Figure 7B). Consequently, extreme increases in genome size and complexity are hypothesized to create more opportunities for the evolution of novel transit peptides and diversification of the plastid proteome, but differences are subtler when the genomes being compared are closer in size. Previous literature (e.g., (Bennetzen and Wang, 2014; Byun and Singh, 2013; Kleine et al., 2009)) has suggested that gene duplication is a prerequisite or at least greatly encourages neofunctionalization via novel subcellular targeting,

and the generally linear correlation with proteome size suggests that this may indeed be the case. However, based on the data, the evolution of the plastid proteome is more likely to be driven by environmental adaptation and selection pressure (Christian et al., 2019).

As with the conserved plastid-targeted clusters, the accuracy of targeting prediction in NPTPs was cross-validated against experimentally-validated proteins from Arabidopsis. For the RBH clusters, 75% (4 proteins) were validated to be true plastid proteins via GFP, and 53.8% (17 proteins) validated by MS. For UCLUST, 29.4% (17 proteins) were validated by GFP, and 41.4% validated by MS. Specificity was also very high: only 6.3% of 300 predicted non-plastid-targeted proteins in RBH-generated NPTP clusters were found to actually be plastid-targeted by GFP, while the rate in MS-validated proteins was 13.4% (967 proteins). UCLUST generated similar results, with false negative error rates of 3% (493 proteins) in GFP-validated data and 12.5% (1,369 proteins) for MS-validated data. The few false negatives in predicted NPTPs may be representative of ambiguous/intermediate sequences in clusters which are already predicted to be uniquely chloroplast-targeted in Arabidopsis and therefore represent missing links. More pertinently, the GFP estimates are likely more accurate due to the experimental specificity errors inherent in mass spectrometry, and the 3-6% error rates are within an acceptable range.

Overall, these data affirm that evolution of the plastid proteome is highly dynamic at the species-level. Compared to previous reports, somewhat reduced species-unique plastid-targeted proteins are reported here (e.g., (Richly and Leister, 2004; Schaeffer et al., 2014)) due in part to the removal of singletons and single-species clusters. Homology to sequences in other species dramatically decreases the probability of pseudogenes and gene prediction errors. Remarkably, the monocot species had an average of 50-60% more species-unique plastid-targeted protein clusters than eudicot or *Amborella* counterparts. Even after removal of the outliers *P. virgatum*

and *A. amnicola*, monocots still had 40% more plastid-targeted clusters than eudicots according to RBH methods, and over 80% more clusters using UCLUST. The reasons for this could be two-fold. First, the monocot species in this analysis have larger proteomes on average, increasing the overall likelihood for both *de novo* evolution of NPTPs and for retention of orphaned singleton/species-specific proteins. Secondly, monocots, and especially grasses, have been described to have many presence/absence variants (PAV's) and copy number variants (CV's) in their genomes. Pan-genome sequencing of *B. distachyon* revealed over 7,000 pan-genes that are not present in the reference genome, and an average of 9 Mb of sequence in each accession does not align to the reference genome (Gordon et al., 2017). Similar rates of PAV's have been reported for cereal crops: only half of the pan-genome diversity of maize is present in the reference genome (Hirsch et al., 2014), over 21,000 predicted wheat genes are not represented in the reference genome (Montenegro et al., 2017), and 8,000 predicted rice genes are not represented in the Nipponbare reference genome (Yao et al., 2015). In contrast, pan-genomes of *Arabidopsis* (Alonso-Blanco et al., 2016) and tomato (Aflitos et al., 2014; Lin et al., 2014) describe variation primarily at the SNP and small insertion/deletion levels, although one report described that 14.9 Mb of the Columbia-0 genome was absent in one or more other accessions (Gan et al., 2011). In *Brassica oleracea*, less than 20% of genes were affected by presence/absence variation (Golicz et al., 2016). Somewhat higher variation is observed in legumes: 302 soybean lines including varieties, landraces, and wild accessions revealed 1,614 copy number variants and 6,388 segmental deletions, and 51.4% of gene families were dispensable (Zhou et al., 2015) while in *Medicago truncatula*, 67% of annotated genes may be dispensable (Zhou et al., 2017). It bears consideration that the pangenomes of the grasses are primarily within cultivated accessions and have already passed through a domestication filter

which already significantly reduces genomic diversity, whereas the pangenomes of most of the eudicots include wild and landrace accessions. These trends suggest that PAV's and CV's are significant drivers of plastid proteome evolution, either by retention of orphaned genes or by *de novo* evolution of transit peptides in duplicated genes. Despite the smaller number of species-unique clusters, conserved plastid-targeted proteins are still outnumbered up to 25-fold by species-unique or semi-conserved proteins. If even a fraction of these sequences is accurate and expressed *in vivo*, each could impart novel biological functions because escape from the evolutionarily established biochemical and regulatory environment could impart a different function in a new subcellular environment without changing the functional sequence of the protein. Thus, each of these is an excellent candidate for further characterization to determine if unique phenotypes are created by relocalization to the plastid.

Conversely, species-specific plastid-targeted genes in model systems could yield misleading interpretations because the same phenotypes for those genes would not be observed in species where homologs do not have plastid-specific localization. Such a situation is potentially problematic for the unique plastid-targeted proteins detected for Arabidopsis, *B. distachyon*, and rice because it is likely that some of these genes already have a described gene function that is being inaccurately ascribed to plants as a whole. Indeed, out of 113 Arabidopsis proteins with predicted species-specific plastid-targeting, 18 have a described phenotype, and 100 are cited in previous research reports (summarized in Additional File 3). In cases where the predicted localization divergence is validated, the mutant phenotypes for those sequences will have to be revised.

Conclusions

The experiments described in this study validated the hypothesis that subcellular localization prediction programs can accurately predict chloroplast transit peptides at a whole-genome scale in higher plants and can perform equally well for both Monocots and Eudicots. The best-performing method was then applied to predict chloroplast proteins globally for a diverse range of Angiosperm species and developed both a slow and accurate reciprocal best-BLAST hit method and a fast-liberal UCLUST method to cluster gene families. Though results were not identical, UCLUST yielded comparable results while performing faster and more efficiently. With the addition of more genotypes, UCLUST could be a useful tool to overcome the inefficiency of BLAST-based methods. The consensus of both methods determined that the hypothesis of extremely plastid proteome variability was correct and robust across evolutionary space. Roughly 700 genes were shared between the chloroplast proteomes in all plant species, but these were vastly outnumbered by proteins with variable plastid targeting prediction. Most of these species- or clade-specific proteins have no known function for the chloroplast and are excellent candidates for further studies. Additionally, roughly a third of conserved plastid-targeted proteins have no known function and could be targeted for future reverse genetics experiments. Biological verification of these sequences remains a significant challenge. Even if good prediction accuracy was achieved, these sequences may be poorly expressed, expressed only in particular conditions, or nonfunctional. Incorporation of transcriptomics would provide significant evidence that these genes are at least expressed, and patterns of gene expression along with co-expression information may also reveal additional information about their function. Experimental validation using mass spectrometry could also be used, but many proteins may have abundances below detection limits, and technical challenges also remain for the isolation of

non-green plastids where they may be more abundant. The decreasing costs of gene synthesis make high-throughput fluorescence protein assays an attractive alternative. In addition to increased sensitivity and specificity compared with mass spectrometry, fluorescent protein assays could also be used to simultaneously validate whether the localization of species-unique proteins is truly different from their nearest predicted non-plastid-targeted homologs. These research findings are expected to provide a foundation for further research into unique plastid biology and to understand better how diversification of the organellar proteomes contributes to important agronomic, biochemical, culinary, or even aesthetic traits.

Materials and Methods

Cross-validation of in silico techniques

Test datasets for cross-comparison of subcellular prediction algorithms were retrieved from PPDB (2012 update; current as of this writing), AT_CHLORO (January 2015 update; current as of this writing) (Ferro et al., 2010), Suba4 (30 June 2017 update; current as of this writing) (Sun et al., 2009), CropPAL version 58839ba (Hooper et al., 2015), and CropPAL2 version 74866967 (Hooper et al., 2015). Headers which could not be referenced to the most up-to-date reference proteomes were discarded. For AT_CHLORO, Suba4, and PPDB databases, all genes located on the chloroplast and mitochondrial genomes were removed, and redundant headers were merged. Subsets of data including sequences confirmed by mass spectrometry, GFP fusion, either GFP or mass spectrometry, or both were extracted from each database by filtering for the keywords “Chloroplast” or “Plastid.” All ambiguous results containing experimental evidence for both plastids and one or more additional subcellular location were removed.

Experimentally validated protein sequences were analyzed with TargetP v.1.1 (Emanuelsson et al., 2007, 2000), WoLF PSORT Command Line Version 0.2 (Horton et al., 2007), PredSL Web Server (Petsalaki et al., 2006), Localizer v.1.0.2 (Sperschneider et al., 2017), MultiLoc2 version 2-26-10-2009 (Blum et al., 2009), and PCLR update 2011-11-24 release 0.9 (Schein et al., 2001). Additionally, NLStradamus v.1.8 (Ba et al., 2009) was used as part of the Localizer algorithm, while Python v.2.7.5, LIBSVM v.2.8, BLAST v.2.2.30, and Interproscan v.5.25-64.0 were used as part of MultiLoc2. Results for each workflow were converted into binary classification and evaluated for Sensitivity (SE), Specificity (SP), Matthew's Correlation Coefficient (MCC), and accuracy (ACC) as related to plastid localization prediction based on the number of true positives, false positives, true negatives, and false negatives compared to the annotations in the corresponding experimental dataset (see equations below). Combinatorial approaches were performed for each possible combination of programs from two up to all six programs, and different thresholds were evaluated based on the number of programs in agreement for plastid localization. Complete records of individual and combinatorial workflows for each experimental dataset are available in Additional File 2. All heatmaps generated for Tables 2, 3, and 4 and Figure 1 were generated using conditional formatting in Microsoft Excel.

$$\text{Sensitivity}(i) = \frac{tp}{tp + fn}$$

$$\text{Specificity}(i) = \frac{tn}{tn + fp}$$

$$\text{MCC}(i) = \frac{tp \times tn - fp \times fn}{\sqrt{(tp+fn)(tp+fp)(tn+fn)(tn+fp)}}$$

$$\text{Overall Accuracy (ACC): } \frac{tp + tn}{tp + fp + tn + fn}$$

Where tp is the number of sequences correctly identified as plastid-targeted, tn is the number of sequences correctly predicted to be non-plastid-targeted, fp is the number of non-

plastid-targeted sequences incorrectly predicted as plastid-targeted, and fn is the number of plastid-targeted sequences that were predicted as non-plastid-targeted. Note that these categorizations are based on the accuracy of the database annotation and any filtering that was applied to data subsets, and they may not reflect biological accuracy.

Whole Proteome Analysis

Predicted proteomes for *Amborella trichopoda*, *Arabidopsis thaliana*, *Brachypodium distachyon*, *Fragaria vesca*, *Glycine max*, *Malus × domestica*, *Oryza sativa*, *Panicum virgatum*, *Populus trichocarpa*, *Prunus persica*, *Setaria italica*, *Solanum lycopersicum*, and *Sorghum bicolor* were downloaded from Phytozome (phytozome.jgi.doe.gov). The proteome of *Anthurium amnicola* was obtained by personal correspondence with Jon Suzuki in advance of the publication of (Suzuki et al., 2017). For *Vitis vinifera*, an expanded proteome version was obtained from (Vitulo et al., 2014). For *Malus × domestica*, modifications to the predicted proteome were made because over 15,000 sequences, representing over 20% of the predicted proteome, were determined to have no significant matches to proteins from other species (See Additional File 3). The predicted proteome was expanded using apple transcriptomics data that were downloaded from the NCBI SRA database under the project numbers PRJEB2506, PRJEB4314, PRJEB6212, and PRJNA231737, representing a mixture of leaf, apical meristem, fruit, and root tissues at different time points and under varying conditions (Bai et al., 2014; Gusberti et al., 2013; Krost et al., 2013, 2012; Petersen et al., 2015). These sources are described further in Additional File 1. Sequence files were processed in CLC Genomics Workbench (Qiagen Bioinformatics, Hilden, Germany), and paired Illumina read files and 454 sequencing files were indicated during import. Graphical QC reports were generated to obtain nucleotide contribution (GC content) and quality distribution (quality scores) by base position. Reads were

processed to remove ambiguous nucleotides and base quality scores lower than 0.001. Illumina reads were additionally trimmed at the 5' end until the GC content stabilized within 0.5% of the average, and reads with fewer than 34 bases remaining were discarded. All paired read files were subsequently merged using default settings. All processed read files were assembled *de novo* with default settings. Assembled contigs of >300 bp were kept and used to predict open reading frames (ORF's). Non-overlapping ORF's with at least 5x average base coverage and >300 bp were extracted and translated into protein sequences. Finally, extracted protein sequences were compared against the existing *Malus x domestica* v.1.0 predicted gene set (Velasco et al., 2010) downloaded from Rosaceae.org. All hits with greater than 98% ID and coverage (as per Bai et al., 2014) were tagged as potential duplications or alleles of the original headers but were kept in the peptide dataset in case minor mutations caused differential localization prediction. All sequences generated in this transcriptomics experiment are available in Additional File 6. In total, 36,477 sequences were obtained, of which 26,881 sequences were determined to be unique in comparison with the apple genome paper (Velasco et al., 2010). Addition of the unique genes from the *de novo* transcriptome created a final dataset of 64,680 unique proteins. Redundant sequences from the resulting transcriptome were retained in case minor differences resulted in differential targeting.

The predicted proteomes of all species were filtered to remove any sequences less than 100 residues and which did not begin with methionine. Post-analysis filtering was accomplished by removing singleton sequences that failed to find matches with both the USEARCH method and BLAST (indicated for each sequence in Additional File 3). Remaining sequences were analyzed with TargetP v.1.1 (Emanuelsson et al., 2007, 2000) and Localizer v.1.0.2 (Sperschneider et al., 2017). All sequences predicted by both methods to have a chloroplast

transit peptide were classified as plastid-targeted, and all sequences with either “1 or 2” or “0 of 2” chloroplast transit peptide predictions were classified as non-plastid-targeted.

Clustering of Gene Families

Reciprocal Best-BLAST hit clustering was performed as follows: Pairwise BLAST-P (v.2.3.0+ command line executable; Altschul et al., 1997, 1990) was performed for each species' predicted proteome set against that of every other species in both forward and reverse directions. These results were filtered for hits in which identity and coverage parameters exceeded 40%. Of these, only hits in which two sequences from different genomes were the respective best hit were kept. Next, better-BLAST hits within each species were performed by conducting pairwise BLAST-P of the predicted proteome against itself. Hits exceeding 90% coverage and identity and which was reciprocal within the first 10 hits were collected. Cluster merging was performed by iterating through each possible header and collapsing all pairwise hits containing that header.

Clustering using the UCLUST algorithm proceeded as follows: An initial run on a length-sorted FASTA file containing all sequences was performed using ‘Cluster_Fast’ function of UCLUST (v.9.2.64_win32; Edgar, 2010) with 40% identity and 40% query coverage. Next, random seeds were constructed by extracting a single random sequence from each cluster, sorting the resulting sequences by length, and appending them to a length-sorted FASTA of the full sequence list used in the initial “Cluster_Fast” analysis. 100 randomly-seeded FASTA files were then analyzed with “Cluster_Fast” set to 90% sequence identity. Target and query coverage were additionally set to 0.4 to avoid problems with small query sequences acting as centroids for much larger sequences as a result of USEARCH being performed in sequential rather than length-sorted order. Cluster merging was performed by iteratively searching through each possible sequence header and collapsing all clusters containing that header. Custom scripts were

developed for automating program workflows, referencing and translating sequences or headers, performing seed randomization for the modified UCLUST technique, performing cluster expansion, calculating statistics on clustering outputs, and referencing headers to respective clusters for both workflows. Sequence members within merged clusters from RBH and UCLUST methods were referenced to the predicted plastid targeting phenotype, and all clusters containing plastid-targeted members were extracted. Conserved plastid-targeted gene families were defined as clusters containing at least 13 species and in which all had either predicted plastid transit peptides or at least three additional sequences. Semi-conserved plastid-targeted gene families were defined as clusters containing plastid-targeted sequences from at least 2 species within each family or clade and no predicted plastid-targeted sequences from species outside that clade. Non-conserved plastid-targeted gene families were defined as all clusters containing a minimum of three species in which only one species had a plastid-targeted sequence.

Gene Ontology Enrichment

Annotations for NPTPs were retrieved from Phytozome (<https://phytozome.jgi.doe.gov>) for each of the species used in the analysis except *Anthurium amnicola* and *Vitis vinifera*, which were retrieved from (Suzuki et al., 2017) and (Vitulo et al., 2014), respectively. Non-redundant predicted proteins produced by the *de novo* transcriptome assembly of *Malus × domestica* as described in Chapter 3 were annotated using BLASTP against the NR Protein database at NCBI with BLAST2GO default parameters (Conesa and Götze, 2008) (BioBam Bioinformatics, Valencia, Spain). GO terms were converted into GOSlim annotations using BLAST2GO, and for each cluster, all terms shared by at least three species and present in over 10% of a cluster's sequences were extracted to develop query datasets. In parallel, the same methods were used to extract GO terms from the total list of clusters to serve as reference datasets. Enrichment of GO

terms in the shared plastid-targeted clusters was performed using BLAST2GO, with Fisher's Exact Test was used to calculate significance using a false discovery rate (FDR) of less than 0.05 as a minimum significance threshold (Conesa and Götze, 2008). Graphical analyses of enriched GO terms were produced in BLAST2GO.

Gene and Phenotype Identification

Full gene annotations include described gene names were downloaded for the TAIR10 Arabidopsis genome from Phytozome (phytozome.jgi.doe.gov). Gene names were referenced from the annotation file for all Arabidopsis sequences present in conserved plastid-targeted protein clusters. Phenotype information for species-unique plastid-targeted proteins was referenced on NCBI (<https://www.ncbi.nlm.nih.gov/>).

Authors' contributions

RC and AD designed the study. RC performed localization prediction, gene clustering, and data analysis. ER assisted in methods development. SH performed gene annotation analyses. AD and ER supervised the study. RC and AD prepared the manuscript. All authors read and approved the manuscript. The authors declare no conflict of interest.

Acknowledgments

Work in the Dhingra lab was supported by Washington State University Agriculture Center Research Hatch Grant WNP00011 to AD. RC and SH acknowledge the support received from the National Institutes of Health/National Institute of General Medical Sciences through an institutional training grant award T32-GM008336. The contents of this work are solely the

responsibility of the authors and do not necessarily represent the official views of the NIGMS or NIH.

References

- Aflitos, S., Schijlen, E., De Jong, H., De Ridder, D., Smit, S., Finkers, R., Wang, J., Zhang, G., Li, N., Mao, L., Bakker, F., Dirks, R., Breit, T., Gravendeel, B., Huits, H., Struss, D., Swanson-Wagner, R., van Leeuwen, H., van Ham, R.C.H.J., Fito, L., Guignier, L., Sevilla, M., Ellul, P., Ganko, E., Kapur, A., Reclus, E., de Geus, B., van de Geest, H., te Lintel Hekkert, B., van Haarst, J., Smits, L., Koops, A., Sanchez-Perez, G., van Heusden, A.W., Visser, R., Quan, Z., Min, J., Liao, L., Wang, X., Wang, G., Yue, Z., Yang, X., Xu, N., Schranz, E., Smets, E., Vos, R., Rauwerda, J., Ursem, R., Schuit, C., Kerns, M., van den Berg, J., Vriezen, W., Janssen, A., Datema, E., Jahrman, T., Moquet, F., Bonnet, J., Peters, S., 2014. Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. *Plant J.* 80, 136–148. <https://doi.org/10.1111/tpj.12616>
- Ajjawi, I., Lu, Y., Savage, L.J., Bell, S.M., Last, R.L., 2010. Large-Scale Reverse Genetics in *Arabidopsis*: Case Studies from the Chloroplast 2010 Project. *Plant Physiol.* 152, 529–540. <https://doi.org/10.1104/pp.109.148494>
- Albert, V.A., Barbazuk, W.B., Der, J.P., Leebens-Mack, J., Ma, H., Palmer, J.D., Rounsley, S., Sankoff, D., Schuster, S.C., Soltis, D.E., 2013. The Amborella Genome and the Evolution of Flowering Plants. *Science.* 342, 1241089. <https://doi.org/10.1126/science.1241089>
- Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K.M.M., Cao, J., Chae, E., Dezwaan, T.M.M., Ding, W., Ecker, J.R.R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D.G.G., Hancock, A.M.M., Henz, S.R.R., Holm, S., Horton, M., Jarsulic, M., Kerstetter, R.A.A., Korte, A., Korte, P., Lanz, C., Lee, C.R., Meng, D., Michael, T.P.P., Mott, R., Mulyati, N.W.W., Nägele, T., Nagler, M., Nizhynska, V., Nordborg, M., Novikova, P.Y.Y., Picó, F.X., Platzer, A., Rabanal, F.A.A., Rodriguez, A.,

- Rowan, B.A.A., Salomé, P.A.A., Schmid, K.J.J., Schmitz, R.J.J., Seren, Ü., Sperone, F.G.G., Sudkamp, M., Svardal, H., Tanzer, M.M.M., Todd, D., Volchenboum, S.L.L., Wang, C., Wang, G., Wang, X., Weckwerth, W., Weigel, D., Zhou, X., 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. *Cell* 166, 481–491. <https://doi.org/10.1016/j.cell.2016.05.063>
- Altenhoff, A.M., Dessimoz, C., 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput. Biol.* 5. <https://doi.org/10.1371/journal.pcbi.1000262>
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic Local Alignment Search Tool. *J. Mol. Biol.* 215, 403–410. <https://doi.org/10.1088/1126-6708/2008/04/040>
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.
- Armbruster, U., Pesaresi, P., Pribil, M., Hertle, A., Leister, D., 2011. Update on chloroplast research: New tools, new topics, and new trends. *Mol. Plant* 4, 1–16. <https://doi.org/10.1093/mp/ssq060>
- Ba, A.N.N., Pogoutse, A., Provar, N., Moses, A.M., 2009. NLStradamus: A simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinformatics* 10, 1–11. <https://doi.org/10.1186/1471-2105-10-202>
- Bai, Y., Dougherty, L., Xu, K., 2014. Towards an improved apple reference transcriptome using RNA-seq. *Mol. Genet. Genomics* 289, 427–438. <https://doi.org/10.1007/s00438-014-0819-3>
- Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., Miyano, S., 2002. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* 18, 298–305.

<https://doi.org/10.1093/bioinformatics/18.2.298>

Barsan, C., Sanchez-Bel, P., Rombaldi, C., Egea, I., Rossignol, M., Kuntz, M., Zouine, M., Latché, A., Bouzayen, M., Pech, J.C., 2010. Characteristics of the tomato chromoplast revealed by proteomic analysis. *J. Exp. Bot.* 61, 2413–2431.

<https://doi.org/10.1093/jxb/erq070>

Bennetzen, J.L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A.C., Estep, M., Feng, L., Vaughn, J.N., Grimwood, J., Jenkins, J., Barry, K., Lindquist, E., Hellsten, U., Deshpande, S., Wang, X., Wu, X., Mitros, T., Triplett, J., Yang, X., Ye, C.Y., Mauro-Herrera, M., Wang, L., Li, P., Sharma, M., Sharma, R., Ronald, P.C., Panaud, O., Kellogg, E.A., Brutnell, T.P., Doust, A.N., Tuskan, G.A., Rokhsar, D., Devos, K.M., 2012.

Reference genome sequence of the model plant *Setaria*. *Nat. Biotechnol.* 30, 555–561.

<https://doi.org/10.1038/nbt.2196>

Bennetzen, J.L., Wang, H., 2014. The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. *Annu. Rev. Plant Biol.* 65, 505–530.

<https://doi.org/10.1146/annurev-arplant-050213-035811>

Bhattacharya, D., Archibald, J.M., Weber, A.P.M., Reyes-Prieto, A., 2007. How do endosymbionts become organelles? Understanding early events in plastid evolution.

BioEssays 29, 1239–1246. <https://doi.org/10.1002/bies.20671>

Blum, T., Briesemeister, S., Kohlbacher, O., 2009. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics* 10, 274. <https://doi.org/10.1186/1471-2105-10-274>

Bodén, M., 2014. The prediction of targeting peptides is enhanced by sequentially biased recurrent networks.

- Brady, S., Shatkay, H., 2008. EpiLoc: a (working) text-based system for predicting protein subcellular location. *Pacific Symp. Biocomput.* 615, 604–615.
https://doi.org/10.1142/9789812776136_0058
- Briesemeister, S., Blum, T., Brady, S., Lam, Y., Kohlbacher, O., Shatkay, H., 2009. SherLoc2: A High-Accuracy Hybrid Method for Predicting Subcellular Localization of Proteins. *J. Proteome Res.* 8, 5363–5366.
- Briesemeister, S., Rahnenführer, J., Kohlbacher, O., 2010. YLoc-an interpretable web server for predicting subcellular localization. *Nucleic Acids Res.* 38, 497–502.
<https://doi.org/10.1093/nar/gkq477>
- Byun, S.A., Singh, S., 2013. Protein subcellular relocalization increases the retention of eukaryotic duplicate genes. *Genome Biol. Evol.* 5, 2402–2409.
<https://doi.org/10.1093/gbe/evt183>
- Carrie, C., Giraud, E., Whelan, J., 2009. Protein transport in organelles: Dual targeting of proteins to mitochondria and chloroplasts. *FEBS J.* 276, 1187–1195.
<https://doi.org/10.1111/j.1742-4658.2009.06876.x>
- Carrie, C., Small, I., 2013. A reevaluation of dual-targeting of proteins to mitochondria and chloroplasts. *Biochim. Biophys. Acta - Mol. Cell Res.* 1833, 253–259.
<https://doi.org/10.1016/j.bbamcr.2012.05.029>
- Chiu, J.C., Lee, E.K., Egan, M.G., Sarkar, I.N., Coruzzi, G.M., DeSalle, R., 2006. OrthologID: Automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics* 22, 699–707. <https://doi.org/10.1093/bioinformatics/btk040>
- Chou, K.C., Cai, Y.D., 2002. Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.* 277, 45765–45769.

<https://doi.org/10.1074/jbc.M204161200>

Chou, K.C., Shen, H. Bin, 2010. Plant-mPLoc: A top-down strategy to augment the power for predicting plant protein subcellular localization. *PLoS One* 5.

<https://doi.org/10.1371/journal.pone.0011335>

Conesa, A., Götze, S., 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics* 2008. <https://doi.org/10.1155/2008/619832>

Consortium, T.T.G., 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485, 635–641. <https://doi.org/10.1038/nature11119>.The

Craig, W., Lenzi, P., Scotti, N., De Palma, M., Saggese, P., Carbone, V., McGrath Curran, N., Magee, A.M., Medgyesy, P., Kavanagh, T.A., Dix, P.J., Grillo, S., Cardi, T., 2008.

Transplastomic tobacco plants expressing a fatty acid desaturase gene exhibit altered fatty acid profiles and improved cold tolerance. *Transgenic Res.* 17, 769–782.

<https://doi.org/10.1007/s11248-008-9164-9>

Das, M., Haberer, G., Panda, A., Laha, S. Das, Ghosh, T.C., Schäffner, A.R., 2016. Expression pattern similarities support the prediction of orthologs retaining common functions after gene duplication events. *Plant Physiol.* 171, pp.01207.2015.

<https://doi.org/10.1104/pp.15.01207>

de Vries, J., Sousa, F.L., Bölter, B., Soll, J., Gould, S.B., 2015. YCF1: A Green TIC? *Plant Cell* 27, 1827–1833. <https://doi.org/10.1105/tpc.114.135541>

Doyle, S.R., Kasinadhuni, N.R.P., Chan, C.K., Grant, W.N., 2013. Evidence of Evolutionary Constraints That Influences the Sequence Composition and Diversity of Mitochondrial Matrix Targeting Signals. *PLoS One* 8, 1–8. <https://doi.org/10.1371/journal.pone.0067938>

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*

26, 2460–2461. <https://doi.org/10.1093/bioinformatics/btq461>

Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H., 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* 2, 953–71.

<https://doi.org/10.1038/nprot.2007.131>

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

300, 1005–1016. <https://doi.org/10.1006/jmbi.2000.3903>

Emanuelsson, O., Nielsen, H., Heijne, G. Von, 1999. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.* 8, 978–984.

<https://doi.org/10.1110/ps.8.5.978>

Ferro, M., Brugière, S., Salvi, D., Seigneurin-Berny, D., Court, M., Moyet, L., Ramus, C., Miras, S., Mellal, M., Le Gall, S., Kieffer-Jaquinod, S., Bruley, C., Garin, J., Joyard, J., Masselon, C., Rolland, N., 2010. AT_CHLORO, a Comprehensive Chloroplast Proteome Database with Subplastidial Localization and Curated Information on Envelope Proteins. *Mol. Cell. Proteomics* 9, 1063–1084. <https://doi.org/10.1074/mcp.M900325-MCP200>

Fyshe, A., Liu, Y., Szafron, D., Greiner, R., Lu, P., 2008. Improving subcellular localization prediction using text classification and the gene ontology. *Bioinformatics* 24, 2512–2517.

<https://doi.org/10.1093/bioinformatics/btn463>

Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Bohnert, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T., Kahles, A., Bohnert, R., Geraldine, J., Derwent, P., Kersey, P., Belfield, E.J., Harberd, N.P., Kemen, E., Toomajian, C., Kover, P.X., Clark, R.M., Rättsch, G., Mott, R., 2011. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* 477, 419.

<https://doi.org/10.1038/nature10414>

Golicz, A.A., Bayer, P.E., Barker, G.C., Edger, P.P., Kim, H., Martinez, P.A., Kit, C., Chan, K., Severn-ellis, A., McCombie, W.R., Parkin, I.A.P., Paterson, A.H., Pires, J.C., Sharpe, A.G., Tang, H., Teakle, G.R., Town, C.D., Batley, J., Edwards, D., 2016. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat. Commun.* 7, 13390.

<https://doi.org/10.1038/ncomms13390>

Gordon, S.P., Contreras-Moreira, B., Woods, D.P., Des Marais, D.L., Burgess, D., Shu, S., Stritt, C., Roulin, A.C., Schackwitz, W., Tyler, L., Martin, J., Lipzen, A., Dochy, N., Phillips, J., Barry, K., Geuten, K., Budak, H., Juenger, T.E., Amasino, R., Caicedo, A.L., Goodstein, D., Davidson, P., Mur, L.A.J., Figueroa, M., Freeling, M., Catalan, P., Vogel, J.P., 2017.

Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.* 8. <https://doi.org/10.1038/s41467-017-02292-8>

Gusberti, M., Gessler, C., Brogini, G.A.L., 2013. RNA-seq analysis reveals candidate genes for ontogenic resistance in *Malus-Venturia* pathosystem. *PLoS One* 8.

<https://doi.org/10.1371/journal.pone.0078457>

Heazlewood, J.L., 2005. Combining Experimental and Predicted Datasets for Determination of the Subcellular Location of Proteins in *Arabidopsis*. *Plant Physiol.* 139, 598–609.

<https://doi.org/10.1104/pp.105.065532>

Heazlewood, J.L., Verboom, R.E., Tonti-Filippini, J., Small, I., Millar, A.H., 2007. SUBA: The *Arabidopsis* subcellular database. *Nucleic Acids Res.* 35, 213–218.

<https://doi.org/10.1093/nar/gkl863>

Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., Vaillancourt, B.,

Peñagaricano, F., Lindquist, E., Pedraza, A., Barry, K., Leon, N. De, Kaeppler, S.M., Buell,

- C.R., 2014. Insights into the Maize Pan-Genome and Pan-Transcriptome. *Plant Cell* 26, 121–135. <https://doi.org/10.1105/tpc.113.119982>
- Hönigschmid, P., Bykova, N., Schneider, R., Ivankov, D., Frishman, D., 2018. Evolutionary Interplay between Symbiotic Relationships and Patterns of Signal Peptide Gain and Loss. *Genome Biol. Evol.* 10, 928–938. <https://doi.org/10.1093/gbe/evy049>
- Hooper, C.M., Castleden, I.R., Aryamanesh, N., Jacoby, R.P., Millar, A.H., 2015. Finding the subcellular location of barley, wheat, rice and maize proteins: The compendium of crop proteins with annotated locations (cropPAL). *Plant Cell Physiol.* 57, e9. <https://doi.org/10.1093/pcp/pcv170>
- Hooper, C.M., Castleden, I.R., Tanz, S.K., Aryamanesh, N., Millar, A.H., 2017. SUBA4: The interactive data analysis centre for Arabidopsis subcellular protein locations. *Nucleic Acids Res.* 45, D1064–D1074. <https://doi.org/10.1093/nar/gkw1041>
- Hooper, C.M., Tanz, S.K., Castleden, I.R., Vacher, M.A., Small, I.D., Millar, A.H., 2014. SUBAcon: A consensus algorithm for unifying the subcellular localization data of the Arabidopsis proteome. *Bioinformatics* 30, 3356–3364. <https://doi.org/10.1093/bioinformatics/btu550>
- Horton, P., Park, K.J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J., Nakai, K., 2007. WoLF PSORT: Protein localization predictor. *Nucleic Acids Res.* 35, 585–587. <https://doi.org/10.1093/nar/gkm259>
- Initiative, T.I.B., 2010. Genome sequencing and analysis of the model grass *Brahyopodium distachyon*. *Nature* 463, 763–768. <https://doi.org/10.1038/nature08747>
- Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Huguene, P., Dasilva, C., Horner, D., Mica,

- E., Jublot, D., Poulain, J., Bruyère, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pè, M.E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A.F., Weissenbach, J., Quétier, F., Wincker, P., 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467. <https://doi.org/10.1038/nature06148>
- Jeong, K., Kim, S., Bandeira, N., 2012. False discovery rates in spectral identification. *BMC Bioinformatics* 13, S2. <https://doi.org/10.1186/1471-2105-13-S16-S2>
- Kim, K., Kim, W., Kim, S., 2011. ReMark: An automatic program for clustering orthologs flexibly combining a Recursive and a Markov clustering algorithms. *Bioinformatics* 27, 1731–1733. <https://doi.org/10.1093/bioinformatics/btr259>
- Kleffmann, T., Russenberger, D., Von Zychlinski, A., Christopher, W., Sjölander, K., Gruissem, W., Baginsky, S., 2004. The *Arabidopsis thaliana* chloroplast proteome reveals pathway abundance and novel protein functions. *Curr. Biol.* 14, 354–362. <https://doi.org/10.1016/j.cub.2004.02.039>
- Kleine, T., Maier, U.G., Leister, D., 2009. DNA Transfer from Organelles to the Nucleus: The Idiosyncratic Genetics of Endosymbiosis. *Annu. Rev. Plant Biol.* 60, 115–138. <https://doi.org/10.1146/annurev.arplant.043008.092119>
- Koonin, E. V., Aravind, L., Kondrashov, A.S., 2000. The Impact of Comparative Genomics on Our Understanding of Evolution. *Cell* 101, 573–576.
- Krost, C., Petersen, R., Lokan, S., Brauksiepe, B., Braun, P., Schmidt, E.R., 2013. Evaluation of

- the hormonal state of columnar apple trees (*Malus x domestica*) based on high throughput gene expression studies. *Plant Mol. Biol.* 81, 211–220. <https://doi.org/10.1007/s11103-012-9992-0>
- Krost, C., Petersen, R., Schmidt, E.R., 2012. The transcriptomes of columnar and standard type apple trees (*Malus x domestica*) - A comparative study. *Gene* 498, 223–230. <https://doi.org/10.1016/j.gene.2012.01.078>
- Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., Karthikeyan, A.S., Lee, C.H., Nelson, W.D., Ploetz, L., Singh, S., Wensel, A., Huala, E., 2012. The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res.* 40, 1202–1210. <https://doi.org/10.1093/nar/gkr1090>
- Li, H., Chiu, C.-C., 2010. Protein Transport into Chloroplasts. *Annu. Rev. Plant Biol.* 61, 157–180. <https://doi.org/10.1146/annurev-arplant-042809-112222>
- Li, L., Stoeckert, C.J.J., Roos, D.S., 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes -- Li et al. 13 (9): 2178 -- *Genome Research*. *Genome Res.* 13, 2178–2189. <https://doi.org/10.1101/gr.1224503.candidates>
- Li, L., Yuan, H., 2013. Chromoplast biogenesis and carotenoid accumulation. *Arch. Biochem. Biophys.* 539, 102–109. <https://doi.org/10.1016/j.abb.2013.07.002>
- Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S., Wang, X., Huang, Z., Li, J., Zhang, C., Wang, T., Zhang, Y., Wang, A., Zhang, Y., Lin, K., Li, C., Xiong, G., Xue, Y., Mazzucato, A., Causse, M., Fei, Z., Giovannoni, J.J., Chetelat, R.T., Zamir, D., Städler, T., Li, J., Ye, Z., Du, Y., Huang, S., 2014. Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.* 46, 1220–1226.

<https://doi.org/10.1038/ng.3117>

Lisenbee, C.S., Karnik, S.K., Trelease, R.N., 2003. Overexpression and mislocalization of a tail-anchored GFP redefines the identity of peroxisomal ER. *Traffic* 4, 491–501.

<https://doi.org/10.1034/j.1600-0854.2003.00107.x>

Lu, Y., Savage, L.J., Larson, M.D., Wilkerson, C.G., Last, R.L., 2011. Chloroplast 2010: A Database for Large-Scale Phenotypic Screening of Arabidopsis Mutants. *Plant Physiol.* 155, 1589–1600. <https://doi.org/10.1104/pp.110.170118>

Majeran, W., 2005. Functional Differentiation of Bundle Sheath and Mesophyll Maize Chloroplasts Determined by Comparative Proteomics. *Plant Cell* 17, 3111–3140.

<https://doi.org/10.1105/tpc.105.035519>

Majeran, W., Zybailov, B., Ytterberg, A.J., Dunsmore, J., Sun, Q., 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.

<https://doi.org/10.1074/mcp.M800016-MCP200>

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., Penny, D., 2002. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl. Acad. Sci.* 99, 12246–12251. <https://doi.org/10.1073/pnas.182432999>

McCormick, R.F., Truong, S.K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., Kennedy, M., Amirebrahimi, M., Weers, B.D., McKinley, B., Mattison, A., Morishige, D.T., Grimwood, J., Schmutz, J., Mullet, J.E., 2018. The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J.* 93, 338–354. <https://doi.org/10.1111/tpj.13781>

- Millar, A.H., Whelan, J., Small, I., 2006. Recent surprises in protein targeting to mitochondria and plastids. *Curr. Opin. Plant Biol.* 9, 610–615. <https://doi.org/10.1016/j.pbi.2006.09.002>
- Mitschke, J., Fuss, J., Blum, T., Höglund, A., Reski, R., Kohlbacher, O., Rensing, S.A., 2009. Prediction of dual protein targeting to plant organelles: Methods. *New Phytol.* 183, 224–236. <https://doi.org/10.1111/j.1469-8137.2009.02832.x>
- Montenegro, J.D., Golicz, A.A., Bayer, P.E., Hurgobin, B., Lee, H., Chen, C.-K.K., Visendi, P., Lai, K., Dolezel, J., Batley, J., Edwards, D., 2017. The pangenome of hexaploid bread wheat. *Plant J.* 90, 1007–1013. <https://doi.org/10.1111/tpj.13515>
- Nakai, M., 2018. New Perspectives on Chloroplast Protein Import. *Plant Cell Physiol.* 59, 1111–1119. <https://doi.org/10.1093/pcp/pcy083>
- Nakai, M., 2015. YCF1: A Green TIC: Response to the de Vries et al. Commentary. *Plant Cell* 27, 1834–1838. <https://doi.org/10.1105/tpc.15.00363>
- Nesvizhskii, A.I., 2010. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* 73, 2092–2123. <https://doi.org/10.1016/j.jprot.2010.08.009>
- Ngernprasirtsiri, J., Chollet, R., Kobayashi, H., Sugiyama, T., Akazawa, T., 1989. DNA methylation and the differential expression of C4 photosynthesis genes in mesophyll and bundle sheath cells of greening maize leaves. *J. Biol. Chem.* 264, 8241–8248.
- O'Brien, K.P., Remm, M., Sonnhammer, E.L.L., 2005. Inparanoid: A comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* 33, 476–480. <https://doi.org/10.1093/nar/gki107>
- Osterman, A., Overbeek, R., 2003. Missing genes in metabolic pathways: A comparative genomics approach. *Curr. Opin. Chem. Biol.* 7, 238–251. [https://doi.org/10.1016/S1367-5931\(03\)00027-9](https://doi.org/10.1016/S1367-5931(03)00027-9)

- Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R.L., Lee, Y., Zheng, L., Orvis, J., Haas, B., Wortman, J., Buell, R.C., 2007. The TIGR Rice Genome Annotation Resource: Improvements and new features. *Nucleic Acids Res.* 35, D883–D887. <https://doi.org/10.1093/nar/gkl976>
- Petersen, R., Djozic, H., Rieger, B., Rapp, S., Schmidt, E.R., 2015. Columnar apple primary roots share some features of the columnar-specific gene expression profile of aerial plant parts as evidenced by RNA-Seq analysis. *BMC Plant Biol.* 15, 1–16. <https://doi.org/10.1186/s12870-014-0356-6>
- Petsalaki, E.I., Bagos, P.G., Litou, Z.I., Hamodrakas, S.J., 2006. PredSL: A Tool for the N-terminal Sequence-based Prediction of Protein Subcellular Localization. *Genomics Proteomics Bioinformatics* 4, 48–55. [https://doi.org/10.1016/S1672-0229\(06\)60016-8](https://doi.org/10.1016/S1672-0229(06)60016-8)
- Pierleoni, A., Martelli, P.L., Fariselli, P., Casadio, R., 2007. eSLDB: Eukaryotic subcellular localization database. *Nucleic Acids Res.* 35, 208–212. <https://doi.org/10.1093/nar/gkl775>
- Richly, E., Leister, D., 2004. An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of Arabidopsis and rice. *Gene* 329, 11–16. <https://doi.org/10.1016/j.gene.2004.01.008>
- Sanderson, M.J., McMahon, M.M., 2007. Inferring angiosperm phylogeny from EST data with widespread gene duplication. *BMC Evol. Biol.* 7, S3. <https://doi.org/10.1186/1471-2148-7-S1-S3>
- Schaeffer, S., Harper, A., Raja, R., Jaiswal, P., Dhingra, A., 2014. Comparative analysis of predicted plastid-targeted proteomes of sequenced higher plant genomes. *PLoS One* 9, e112870. <https://doi.org/10.1371/journal.pone.0112870>
- Schaeffer, S.M., Christian, R., Castro-Velasquez, N., Hyden, B., Lynch-Holm, V., Dhingra, A.,

2017. Comparative ultrastructure of fruit plastids in three genetically diverse genotypes of apple (*Malus × domestica* Borkh.) during development. *Plant Cell Rep.* 36, 1627–1640.

<https://doi.org/10.1007/s00299-017-2179-z>

Schein, A.I., Kissinger, J.C., Ungar, L.H., 2001. Chloroplast transit peptide prediction: a peek inside the black box. *Nucleic Acids Res.* 29, E82. <https://doi.org/Artn E82>

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R.C., Jackson, S.A., 2010. Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183. <https://doi.org/10.1038/nature08670>

Shulaev, V., Sargent, D.J., Crowhurst, R.N., Mockler, T.C., Folkerts, O., Delcher, A.L., Jaiswal, P., Mockaitis, K., Liston, A., Mane, S.P., Burns, P., Davis, T.M., Slovin, J.P., Bassil, N., Hellens, R.P., Evans, C., Harkins, T., Kodira, C., Desany, B., Crasta, O.R., Jensen, R. V., Allan, A.C., Michael, T.P., Setubal, J.C., Celton, J.-M., Rees, D.J.G., Williams, K.P., Holt, S.H., Adato, A., Filichkin, S.A., Troggio, M., Viola, R., Ashman, T.-L., Wang, H., Dharmawardhana, P., Elser, J., Raja, R., Priest, H.D., Douglas, W.B.J., Fox, S.E., Givan, S.A., Wilhelm, L.J., Naithani, S., Christoffels, A., Salama, D.Y., Carter, J., Lopez Girona, E., Zdepski, A., Wang, W., Kerstetter, R.A., Schwab, W., Arus, P., Mittler, R., Flinn, B., Aharoni, A., Bennetzen, J.L., Salzberg, S.L., Dickerman, A.W., Velasco, R., Borodovsky, M., Veilleux, R.E., Folta, K.M., 2011. The genome of woodland strawberry (*Fragaria*

- vesca) Vladimir. *Nat. Genet.* 43, 109–116. <https://doi.org/10.1038/ng.740>. The
- Small, I., Peeters, N., Legeai, F., Lurin, C., 2004. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* 4, 1581–1590. <https://doi.org/10.1002/pmic.200300776>
- Small, I., Wintz, H., Akashi, K., Mireau, H., 1998. Two birds with one stone: genes that encode products targeted to two or more compartments. *Plant Mol. Biol.* 38, 265–277. <https://doi.org/10.1023/A:1006081903354>
- Sperschneider, J., Catanzariti, A.-M., DeBoer, K., Petre, B., Gardiner, D.M., Singh, K.B., Dodds, P.N., Taylor, J.M., 2017. LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. *Sci. Rep.* 7, 44598. <https://doi.org/10.1038/srep44598>
- Srividya, N., Davis, E.M., Croteau, R.B., Lange, B.M., 2015. Functional analysis of (4S)-limonene synthase mutants reveals determinants of catalytic outcome in a model monoterpene synthase. *Proc. Natl. Acad. Sci.* 112, 3332–3337. <https://doi.org/10.1073/pnas.1501203112>
- Stockhaus, J., Schlue, U., Koczor, M., Chitty, J.A., Taylor, W.C., Westhoff, P., 1997. The promoter of the gene encoding the C-4 form of phosphoenolpyruvate carboxylase directs mesophyll-specific expression in transgenic C-4 *Flaveria* spp. *Plant Cell* 9, 479–489. <https://doi.org/10.1105/tpc.9.4.479>
- Sugiura, M., 1992. The chloroplast genome. *Plant Mol. Biol.* 19, 149–168. <https://doi.org/10.1007/s00438-005-0092-6>
- Sun, Q., Zybaylov, B., Majeran, W., Friso, G., Olinares, P.D.B., van Wijk, K.J., 2009. PPDB, the Plant Proteomics Database at Cornell. *Nucleic Acids Res.* 37, 969–974. <https://doi.org/10.1093/nar/gkn654>

- Suzuki, J.Y., Amore, T.D., Calla, B., Palmer, N.A., Scully, E.D., Sattler, S.E., Sarath, G., Lichty, J.S., Myers, R.Y., Keith, L.M., Matsumoto, T.K., Geib, S.M., 2017. Organ-specific transcriptome profiling of metabolic and pigment biosynthesis pathways in the floral ornamental progenitor species *Anthurium amnicola* Dressler. *Sci. Rep.* 7, 1–15.
<https://doi.org/10.1038/s41598-017-00808-2>
- Tatusov, R.L., Galperin, M.Y., Natale, D.A., Koonin, E. V, 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28, 33–36.
<https://doi.org/10.1093/nar/28.1.33>
- Tatusov, R.L., Koonin, E. V., Lipman, D.J., 1997. A Genomic Perspective on Protein Families. *Science* (80-.). 278, 631–637. <https://doi.org/10.1126/science.278.5338.631>
- The Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815. <https://doi.org/10.1038/35048692>
- Trachana, K., Larsson, T.A., Powell, S., Chen, W.H., Doerks, T., Muller, J., Bork, P., 2011. Orthology prediction methods: A quality assessment using curated protein families. *BioEssays* 33, 769–780. <https://doi.org/10.1002/bies.201100062>
- Tuskan, G.A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, M., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R.R., Bhalerao, R.P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.L., Cooper, D., Coutinho, P.M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Djfardin, A., DePamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi,

- N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjärvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leplé, J.C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D.R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouzé, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van De Peer, Y., Rokhsar, D., 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* (80-.). 313, 1596–1604. <https://doi.org/10.1126/science.1128691>
- Van Bel, M., Proost, S., Wischnitzki, E., Movahedi, S., Scheerlinck, C., Van de Peer, Y., Vandepoele, K., 2012. Dissecting Plant Genomes with the PLAZA Comparative Genomics Platform. *Plant Physiol.* 158, 590–600. <https://doi.org/10.1104/pp.111.189514>
- Van Wijk, K.J., 2004. Plastid proteomics. *Plant Physiol. Biochem.* 42, 963–977. <https://doi.org/10.1016/j.plaphy.2004.10.015>
- van Wijk, K.J., Baginsky, S., 2011. Plastid Proteomics in Higher Plants: Current State and Future Goals. *Plant Physiol.* 155, 1578–1588. <https://doi.org/10.1104/pp.111.172932>
- Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S.K., Troggio, M., Pruss, D., Salvi, S., Pindo, M., Baldi, P., Castelletti, S., Cavaiuolo, M., Coppola, G., Costa, F., Cova, V., Dal Ri, A., Goremykin, V., Komjanc, M., Longhi, S., Magnago, P., Malacarne, G., Malnoy, M., Micheletti, D., Moretto, M., Perazzolli, M., Si-Ammour, A., Vezzulli, S., Zini, E., Eldredge, G., Fitzgerald, L.M., Gutin, N., Lanchbury, J., MacAlma, T., Mitchell, J.T., Reid, J., Wardell, B., Kodira, C., Chen, Z.,

- Desany, B., Niazi, F., Palmer, M., Koepke, T., Jiwan, D., Schaeffer, S., Krishnan, V., Wu, C., Chu, V.T., King, S.T., Vick, J., Tao, Q., Mraz, A., Stormo, A., Stormo, K., Bogden, R., Ederle, D., Stella, A., Vecchietti, A., Kater, M.M., Masiero, S., Lasserre, P., Lespinasse, Y., Allan, A.C., Bus, V., Chagné, D., Crowhurst, R.N., Gleave, A.P., Lavezzo, E., Fawcett, J.A., Proost, S., Rouzé, P., Sterck, L., Toppo, S., Lazzari, B., Hellens, R.P., Durel, C.E., Gutin, A., Bumgarner, R.E., Gardiner, S.E., Skolnick, M., Egholm, M., Van De Peer, Y., Salamini, F., Viola, R., 2010. The genome of the domesticated apple (*Malus × domestica* Borkh.). *Nat. Genet.* 42, 833–839. <https://doi.org/10.1038/ng.654>
- Verde, I., Abbott, A.G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., Zhebentyayeva, T., Dettori, M.T., Grimwood, J., Cattonaro, F., Zuccolo, A., Rossini, L., Jenkins, J., Vendramin, E., Meisel, L.A., Decroocq, V., Sosinski, B., Prochnik, S., Mitros, T., Policriti, A., Cipriani, G., Dondini, L., Ficklin, S., Goodstein, D.M., Xuan, P., Del Fabbro, C., Aramini, V., Copetti, D., Gonzalez, S., Horner, D.S., Falchi, R., Lucas, S., Mica, E., Maldonado, J., Lazzari, B., Bielenberg, D., Pirona, R., Miculan, M., Barakat, A., Testolin, R., Stella, A., Tartarini, S., Tonutti, P., Arús, P., Orellana, A., Wells, C., Main, D., Vizzotto, G., Silva, H., Salamini, F., Schmutz, J., Morgante, M., Rokhsar, D.S., 2013. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* 45, 487–494. <https://doi.org/10.1038/ng.2586>
- Vitulo, N., Forcato, C., Carpinelli, E.C., Telatin, A., Campagna, D., D'Angelo, M., Zimbello, R., Corso, M., Vannozzi, A., Bonghi, C., Lucchin, M., Valle, G., 2014. A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. *BMC Plant Biol.* 14, 20–30. <https://doi.org/10.1186/1471-2229-14-99>

- von Zychlinski, A., Kleffmann, T., Krishnamurthy, N., Sjölander, K., Baginsky, S., Gruissem, W., 2005. Proteome analysis of the rice etioplast: metabolic and regulatory networks and novel protein functions. *Mol. Cell. Proteomics* 4, 1072–1084.
- Xiong, E., Zheng, C., Wu, X., Wang, W., 2016. Protein Subcellular Location: The Gap Between Prediction and Experimentation. *Plant Mol. Biol. Report.* 34, 52–61.
<https://doi.org/10.1007/s11105-015-0898-2>
- Yang, Y., Smith, S.A., 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. *Mol. Biol. Evol.* 31, 3081–3092. <https://doi.org/10.1093/molbev/msu245>
- Yao, W., Li, G., Zhao, H., Wang, G., Lian, X., Xie, W., 2015. Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biol.* 16, 187.
<https://doi.org/10.1186/s13059-015-0757-3>
- Zeng, Y., Pan, Z., Ding, Y., Zhu, A., Cao, H., Xu, Q., Deng, X., 2011. A proteomic analysis of the chromoplasts isolated from sweet orange fruits [*Citrus sinensis* (L.) Osbeck]. *J. Exp. Bot.* 62, 5297–5309. <https://doi.org/10.1093/jxb/err140>
- Zeng, Y., Pan, Z., Wang, L., Ding, Y., Xu, Q., Xiao, S., Deng, X., 2014. Phosphoproteomic analysis of chromoplasts from sweet orange during fruit ripening. *Physiol. Plant.* 150, 252–270. <https://doi.org/10.1111/ppl.12080>
- Zhou, P., Silverstein, K.A.T., Ramaraj, T., Guhlin, J., Denny, R., Liu, J., Farmer, A.D., Steele, K.P., Stupar, R.M., Miller, J.R., Tiffin, P., Mudge, J., Young, N.D., 2017. Exploring structural variation and gene family architecture with De Novo assemblies of 15 *Medicago* genomes. *BMC Genomics* 18, 261. <https://doi.org/10.1186/s12864-017-3654-1>
- Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., Fang,

C., Shen, Y., Liu, T., Li, C., Li, Q., Wu, M., Wang, M., Wu, Y., Dong, Y., Wan, W., Wang, X., Ding, Z., Gao, Y., Xiang, H., Zhu, B., Lee, S., Wang, W., Tian, Z., 2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33, 401–414. <https://doi.org/10.1038/nbt.3096>

Zhu, M., Lin, J., Ye, J., Wang, R., Yang, C., Gong, J., Liu, Y., Deng, C., Liu, P., Chen, C., Cheng, Y., Deng, X., Zeng, Y., 2018. A comprehensive proteomic analysis of elaioplasts from citrus fruits reveals insights into elaioplast biogenesis and function. *Hortic. Res.* 5, 0–10. <https://doi.org/10.1038/s41438-017-0014-x>

Zybailov, B., Rutschow, H., Friso, G., Rudella, A., Emanuelsson, O., Sun, Q., van Wijk, K.J., 2008. Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. *PLoS One* 3, e1994. <https://doi.org/10.1371/journal.pone.0001994>

Table 1: Self-Reported Performance of Six Algorithms on Prediction of Plastid-Targeted Proteins. Self-reported values for overall and plastidial sensitivity (SE), specificity (SP), Matthew’s Correlation Coefficient (MCC), and accuracy (ACC).

Parentheses indicate values that were calculated to be different from the original paper using the same data. Programs marked with an asterisk (*) had a confusion matrix available, while those marked with a cross (†) did not, but confusion matrices were inferred by the available data; estimations were left as non-integer values, and therefore suffer from rounding errors in MCC and ACC calculations. Localizer, marked with a double cross (‡), was re-run with the original dataset provided in the publication’s supplementary information.

Algorithm	Source	Training Dataset(s)	# of Training Sequences	Plastidial SE	Plastidial SP	Plastidial MCC	Plastidial ACC
TargetP*	Emanuelsson et al., 2007, 2000	SWISS-PROT releases 36,37,38	940	0.85	0.69	0.72	N/A (0.921)
WolfPSORT	Horton et al., 2007	Uniprot version 45	2,113	0.7	0.7	N/A	N/A
PredSL†	Petsalaki et al., 2006	Various (Uniprot release 3.5)	1,002	0.9	0.91	0.88 (0.874)	N/A
Localizer‡	Sperschneider et al., 2017	CropPAL (GFP only)	410	0.725	0.957 (0.798)	0.71	0.914 (0.916)
Multiloc2 (Low-Res)†	Blum et al., 2009	BaCelLo Independent Dataset	132	0.77	0.53	0.72	N/A (0.853)
Multiloc2 (High-Res)†	Blum et al., 2009	BaCelLo Independent Dataset	132	0.53	0.94	0.51 (0.539)	N/A (0.735)
PCLR*	Schein et al., 2001	ChloroP, TargetP	847	0.87 (0.821)	0.30 (0.301)	0.372	0.720

Table 2: Cross-Validation of Algorithms Using Experimentally-Curated Datasets. For each program, SE, SP, MCC, and ACC are reported compared to *in vivo* experimental data using a conservative dataset of GFP-validated proteins, or a larger but more liberal dataset comprised of both GFP and MS data. A heatmap is applied for each column, with green values representing better-performing programs and red values indicating poor performance. Difference between observed performance statistics for each is reported with a heatmap generated by conditional formatting in Excel depicting the absolute value difference from 0. MS data adds increased error especially for observed sensitivity, indicating that a large number of MS-validated proteins are likely artefactual. Furthermore, this suggests that the overall performance of subcellular prediction methods is likely more accurate than high-throughput proteomics papers suggest.

	GFP				GFP & Mass Spectrometry				Difference			
	SE	SP	MCC	ACC	SE	SP	MCC	ACC	SE	SP	MCC	ACC
TargetP	0.67	0.59	0.54	0.86	0.46	0.55	0.32	0.73	0.21	0.04	0.22	0.13
Wolf-PSORT	0.72	0.38	0.38	0.75	0.57	0.44	0.24	0.65	0.15	-0.05	0.14	0.09
PredSL	0.57	0.53	0.45	0.84	0.37	0.52	0.26	0.71	0.19	0.01	0.20	0.12
Localizer	0.68	0.71	0.63	0.90	0.46	0.58	0.34	0.74	0.22	0.14	0.29	0.16
Multiloc2	0.50	0.83	0.59	0.89	0.31	0.63	0.30	0.74	0.18	0.20	0.28	0.15
PCLR	0.74	0.46	0.47	0.80	0.54	0.48	0.28	0.69	0.20	-0.02	0.19	0.11

Low Performance High Performance




Table 3: Performance of Prediction Algorithms against GFP-validated Proteins from Monocots and Eudicots. Dataset sizes are roughly similar in Monocot and Dicot sequences, but MCC is still preferable for comparison. Performance in monocots and eudicots is represented by a red/green heatmap, with green indicating the best values for each statistic and red indicating the worst. The difference between Monocots and Eudicots is represented by red only, normalized to the absolute value away from 0. Positive values indicate higher performance in Monocots, while negative numbers indicate higher performance in Eudicots. 161 plastid-localized proteins and 640 non-plastid-targeted proteins are included for Monocots, while Eudicots include 489 plastidial and 2,432 non-plastid-targeted proteins.

	Monocot: GFP				Eudicot: GFP				Monocot-Eudicot			
	SE	SP	MCC	ACC	SE	SP	MCC	ACC	SE	SP	MCC	ACC
TargetP	0.62	0.71	0.59	0.87	0.68	0.56	0.53	0.86	-0.06	0.15	0.05	0.02
Wolf-PSORT	0.72	0.38	0.35	0.71	0.72	0.38	0.39	0.76	0.00	-0.01	-0.04	-0.05
PredSL	0.43	0.59	0.40	0.83	0.61	0.52	0.47	0.84	-0.17	0.06	-0.07	-0.02
Localizer	0.63	0.76	0.63	0.89	0.69	0.70	0.63	0.90	-0.06	0.06	-0.01	-0.01
Multiloc2	0.40	0.89	0.54	0.87	0.53	0.81	0.60	0.90	-0.12	0.08	-0.06	-0.03
PCLR	0.75	0.56	0.54	0.83	0.73	0.43	0.45	0.80	0.01	0.12	0.09	0.03

Low Performance

High Performance



Table 4: Combinatorial Prediction Approaches Ranked by Matthew’s Correlation

Coefficient (MCC). Each performance statistic column is treated separately and colored with a heatmap in which blue cells indicate higher performance and red cells indicate lower performance. Only the best 25 workflows are represented; full results are available in Additional File 1.

Rank	Workflow	Description	SE	SP	MCC	ACC
1	125	2/3 of (TargetP, Localizer, Multiloc2)	0.646	0.785	0.659	0.907
2	167	2/2 of (TargetP, Localizer)	0.611	0.807	0.650	0.907
3	80	3/4 of (TargetP, Localizer, Multiloc2, PCLR)	0.622	0.791	0.647	0.905
4	127	3/3 of (TargetP, Localizer, PCLR)	0.588	0.822	0.644	0.906
5	152	2/3 of (PredSL, Localizer, Multiloc2)	0.597	0.803	0.639	0.904
6	68	3/4 of (TargetP, PredSL, Localizer, Multiloc2)	0.575	0.827	0.639	0.905
7	161	2/3 of (Localizer, Multiloc2, PCLR)	0.660	0.732	0.635	0.898
8	189	2/2 of (Localizer, PCLR)	0.634	0.756	0.634	0.900
9	188	1/2 of (Localizer, Multiloc2)	0.697	0.696	0.632	0.894
10	4	1 of (Localizer)	0.675	0.714	0.632	0.896
11	19	4/5 of (TargetP, PredSL, Localizer, Multiloc2, PCLR)	0.563	0.828	0.632	0.903
12	100	3/4 of (PredSL, Localizer, Multiloc2, PCLR)	0.578	0.807	0.630	0.902
13	20	3/5 of (TargetP, PredSL, Localizer, Multiloc2, PCLR)	0.660	0.688	0.606	0.888
14	72	3/4 of (TargetP, PredSL, Localizer, PCLR)	0.648	0.697	0.606	0.889
15	69	2/4 of (TargetP, PredSL, Localizer, Multiloc2)	0.678	0.656	0.595	0.882
16	187	2/2 of (Localizer, Multiloc2)	0.474	0.870	0.594	0.896
17	116	2/3 of (TargetP, PredSL, Localizer)	0.663	0.664	0.592	0.883
18	181	2/2 of (PredSL, Localizer)	0.511	0.814	0.591	0.894
19	160	3/3 of (Localizer, Multiloc2, PCLR)	0.462	0.880	0.590	0.895
20	115	3/3 of (TargetP, PredSL, Localizer)	0.491	0.835	0.588	0.894
21	155	2/3 of (PredSL, Localizer, PCLR)	0.698	0.629	0.587	0.875
22	5	1 of (Multiloc2)	0.495	0.826	0.587	0.894
23	101	2/4 of (PredSL, Localizer, Multiloc2, PCLR)	0.706	0.621	0.585	0.873
24	124	3/3 of (TargetP, Localizer, Multiloc2)	0.452	0.883	0.585	0.894
25	79	4/4 of (TargetP, Localizer, Multiloc2, PCLR)	0.445	0.895	0.585	0.894

Low Performance

High Performance



Table 5: Subcellular Targeting Prediction for Selected Genotypes. Predicted protein sequences from fifteen species representing a mixture of model organisms and crop species as well as a mixture of monocots, eudicots, and the extant species *Amborella trichopa* were downloaded from Phytozome (phytozome.jgi.doe.gov) or from the sources indicated in the table. For each genotype, the version, reference, and sequence count are provided from the original publications. *TargetP and Localizer were used to detect plastid-targeted sequences. **Indicates unpublished but publicly-available data downloaded from Phytozome for *Panicum virgatum*.

Species	Version	Source	Sequences	Chloroplast-Targeted*
<i>Amborella trichopoda</i>	1.0	(Albert et al., 2013)	26,846	1,833
<i>Anthurium amnicola</i>	1.0	(Suzuki et al., 2017)	27,959	1,324
<i>Arabidopsis thaliana</i>	TAIR10	(Lamesch et al., 2012)	35,386	2,826
<i>Brachypodium distachyon</i>	3.1	(Initiative, 2010)	52,972	4,240
<i>Fragaria vesca</i>	1.1	(Shulaev et al., 2011)	32,831	2,051
<i>Glycine max</i>	Wm82	(Schmutz et al., 2010)	73,320	5,125
<i>Malus x domestica</i>	1.0 (custom transcriptome)	(Velasco et al., 2010) (Bai et al., 2014; Gusberty et al., 2013; Krost et al., 2013, 2012)	57,386 (74,249)	4,665
<i>Oryza sativa</i>	7.0	(Ouyang et al., 2007)	49,061	3,417
<i>Panicum virgatum</i>	3.1	DOE-JGI**	133,775	10,262
<i>Populus trichocarpa</i>	3.0	(Tuskan et al., 2006)	73,013	5,741
<i>Prunus persica</i>	2.1	(Verde et al., 2013)	47,089	3,615
<i>Setaria viridis</i>	2.2	(Bennetzen et al., 2012)	43,001	3,461
<i>Solanum lycopersicum</i>	2.1	(Consortium, 2012)	47,205	1,875
<i>Sorghum bicolor</i>	3.1	(McCormick et al., 2018)	34,727	3,918
<i>Vitis vinifera</i>	2.0	(Jaillon et al., 2007) (Vitulo et al., 2014)	55,564	3,932

Table 6: RBH Clustering Results by Genotype. Clustering of gene families using 40% reciprocal Interspecies best BLAST hits and 90% reciprocal intraspecies better BLAST hits was performed, and clusters containing plastid-targeted sequences were identified for each genotype. The number of total proteomes and plastid-targeted clusters with at least one Arabidopsis sequence were identified, as well as the number of clusters containing a plastid-targeted sequence from only the selected genotype. The number clusters overlapping with Arabidopsis for all clusters and plastid-targeted clusters was identified, as well as the number of clusters containing a plastid-targeted sequence from only the selected genotype.

Species	Total Clusters	Clustered with Arabidopsis Proteome	Plastid-Targeted Clusters	Clustered with Arabidopsis Plastome	Unique Plastidial	Singleton and Single-Species Clusters	NTPs
<i>Amborella trichopoda</i>	20533	60.97%	1673	44.47%	667	585	82
<i>Anthurium amnicola</i>	7497	81.43%	937	61.26%	187	135	52
<i>Arabidopsis thaliana</i>	15817	100.00%	1796	100.00%	375	301	74
<i>Brachypodium distachyon</i>	17933	67.23%	2380	41.81%	727	498	229
<i>Fragaria vesca</i>	18328	70.63%	1798	47.66%	566	426	140
<i>Glycine max</i>	26629	63.60%	2464	43.83%	905	714	191
<i>Malus x domestica</i>	30257	49.84%	3100	32.13%	1581	1253	328
<i>Oryza sativa</i>	18657	65.83%	2204	44.01%	643	459	184
<i>Panicum virgatum</i>	43875	37.39%	5234	20.27%	3194	2512	682
<i>Populus trichocarpa</i>	20348	71.99%	2167	50.21%	580	413	167
<i>Prunus persica</i>	14375	82.64%	1838	58.65%	296	184	112
<i>Setaria italica</i>	16618	73.25%	2310	43.29%	509	241	268
<i>Solanum lycopersicum</i>	16287	87.55%	1486	66.42%	202	131	71
<i>Sorbus bicolor</i>	16201	68.65%	2351	42.28%	636	386	250
<i>Vitis vinifera</i>	16711	79.83%	1785	56.47%	353	240	113

Table 7: UCLUST Clustering Results by Genotype. Clustering of gene families was performed using an initial UCLUST iteration with 40% coverage and 40% identity followed by extraction of random sequences from each cluster to seed additional iterations performed at 90% coverage and identity. Clusters containing shared sequences were merged, followed by identification of clusters containing plastid-targeted sequences in each genotype. The number clusters overlapping with Arabidopsis for all clusters and plastid-targeted clusters was identified, as well as the number of clusters containing a plastid-targeted sequence from only the selected genotype.

Species	Total Clusters	Clustered with Arabidopsis Proteome	Plastid-Targeted Clusters	Clustered with Arabidopsis Plastome	Unique Plastidial	Singleton and Single-Species Clusters	NPTPs
<i>Amborella trichopoda</i>	19190	55.61%	1721	41.78%	736	541	195
<i>Anthurium amnicola</i>	7365	78.22%	909	58.53%	173	76	97
<i>Arabidopsis thaliana</i>	13065	100.00%	1783	100.00%	261	95	166
<i>Brachypodium distachyon</i>	16777	57.05%	2375	37.68%	623	225	398
<i>Fragaria vesca</i>	16821	65.75%	1828	46.01%	551	172	379
<i>Glycine max</i>	20157	70.78%	2320	49.31%	637	296	341
<i>Malus x domestica</i>	21427	56.25%	2846	35.45%	1197	469	728
<i>Oryza sativa</i>	18102	55.76%	2249	38.86%	564	235	329
<i>Panicum virgatum</i>	29207	34.12%	4725	20.00%	2506	1048	1458
<i>Populus trichocarpa</i>	15881	78.35%	1977	56.35%	335	95	240
<i>Prunus persica</i>	14753	79.49%	1921	57.16%	229	36	193
<i>Setaria italica</i>	17810	57.11%	2427	36.51%	480	98	382
<i>Solanum lycopersicum</i>	15675	81.13%	1574	62.26%	245	79	166
<i>Sorbus bicolor</i>	17395	54.56%	2410	36.14%	554	191	363
<i>Vitis vinifera</i>	16092	79.06%	1805	57.62%	299	102	197

Table 8: Enriched GO terms for Conserved Plastid-Targeted Clusters Identified using RBH. Clusters containing at least 13 species with predicted or likely plastid-targeted sequences were mined for common GO terms and compared against terms extracted for the total set of RBH-derived clusters using BLAST2GO. All terms enriched above $p=1.0E^{-5}$ in core plastid-targeted clusters are presented below.

	GO term	Description	Ontology	P-value	FDR
1	GO:0015979	photosynthesis	BIOLOGICAL_PROCESS	1.73E-44	3.99E-47
2	GO:0008152	metabolic process	BIOLOGICAL_PROCESS	6.40E-27	1.59E-29
3	GO:0006091	generation of precursor metabolites and energy	BIOLOGICAL_PROCESS	1.43E-24	3.82E-27
4	GO:0009058	biosynthetic process	BIOLOGICAL_PROCESS	3.73E-20	1.20E-22
5	GO:0044711	single-organism biosynthetic process	BIOLOGICAL_PROCESS	3.00E-15	1.02E-17
6	GO:0016043	cellular component organization	BIOLOGICAL_PROCESS	4.28E-12	1.64E-14
7	GO:0071840	cellular component organization or biogenesis	BIOLOGICAL_PROCESS	6.51E-12	2.66E-14
8	GO:0044710	single-organism metabolic process	BIOLOGICAL_PROCESS	1.05E-10	5.06E-13
9	GO:0006629	lipid metabolic process	BIOLOGICAL_PROCESS	3.16E-10	1.57E-12
10	GO:0043604	amide biosynthetic process	BIOLOGICAL_PROCESS	7.84E-09	4.05E-11
11	GO:0043603	cellular amide metabolic process	BIOLOGICAL_PROCESS	9.01E-09	4.81E-11
12	GO:0019725	cellular homeostasis	BIOLOGICAL_PROCESS	1.07E-08	5.93E-11
13	GO:0044699	single-organism process	BIOLOGICAL_PROCESS	1.22E-08	7.39E-11
14	GO:0009987	cellular process	BIOLOGICAL_PROCESS	1.22E-08	7.21E-11
15	GO:0065008	regulation of biological quality	BIOLOGICAL_PROCESS	1.54E-08	9.56E-11
16	GO:0006412	translation	BIOLOGICAL_PROCESS	1.67E-08	1.10E-10
17	GO:0042592	homeostatic process	BIOLOGICAL_PROCESS	1.67E-08	1.08E-10
18	GO:0043043	peptide biosynthetic process	BIOLOGICAL_PROCESS	1.73E-08	1.17E-10
19	GO:0006518	peptide metabolic process	BIOLOGICAL_PROCESS	1.80E-08	1.25E-10
20	GO:1901566	organonitrogen compound biosynthetic process	BIOLOGICAL_PROCESS	2.95E-08	2.10E-10
21	GO:1901564	organonitrogen compound metabolic process	BIOLOGICAL_PROCESS	1.04E-07	7.58E-10
22	GO:0034641	cellular nitrogen compound metabolic process	BIOLOGICAL_PROCESS	1.61E-05	1.52E-07

23	GO:0044271	cellular nitrogen compound biosynthetic process	BIOLOGICAL_PROCESS	1.93E-05	1.85E-07
24	GO:0006807	nitrogen compound metabolic process	BIOLOGICAL_PROCESS	2.26E-05	2.21E-07
25	GO:0044249	cellular biosynthetic process	BIOLOGICAL_PROCESS	2.52E-05	2.51E-07
26	GO:1901576	organic substance biosynthetic process	BIOLOGICAL_PROCESS	4.41E-05	4.55E-07
27	GO:0034645	cellular macromolecule biosynthetic process	BIOLOGICAL_PROCESS	4.47E-05	4.69E-07
28	GO:0009059	macromolecule biosynthetic process	BIOLOGICAL_PROCESS	4.74E-05	5.06E-07
29	GO:0010467	gene expression	BIOLOGICAL_PROCESS	2.96E-04	3.31E-06
30	GO:0009536	plastid	CELLULAR_COMPONENT	1.35E-279	2.41E-283
31	GO:0005622	intracellular	CELLULAR_COMPONENT	1.07E-222	3.82E-226
32	GO:0044424	intracellular part	CELLULAR_COMPONENT	1.22E-222	6.49E-226
33	GO:0044464	cell part	CELLULAR_COMPONENT	6.98E-222	6.21E-225
34	GO:0005623	cell	CELLULAR_COMPONENT	6.98E-222	5.62E-225
35	GO:0005737	cytoplasm	CELLULAR_COMPONENT	2.03E-218	2.16E-221
36	GO:0044444	cytoplasmic part	CELLULAR_COMPONENT	1.38E-217	1.72E-220
37	GO:0043229	intracellular organelle	CELLULAR_COMPONENT	1.94E-193	2.76E-196
38	GO:0043226	organelle	CELLULAR_COMPONENT	1.97E-193	3.16E-196
39	GO:0043231	intracellular membrane-bounded organelle	CELLULAR_COMPONENT	1.16E-179	2.06E-182
40	GO:0043227	membrane-bounded organelle	CELLULAR_COMPONENT	5.74E-179	1.12E-181
41	GO:0009579	thylakoid	CELLULAR_COMPONENT	2.71E-68	5.78E-71
42	GO:0016020	membrane	CELLULAR_COMPONENT	1.08E-20	3.06E-23
43	GO:0005739	mitochondrion	CELLULAR_COMPONENT	2.48E-12	8.82E-15
44	GO:0005840	ribosome	CELLULAR_COMPONENT	1.48E-11	6.31E-14
45	GO:1990904	ribonucleoprotein complex	CELLULAR_COMPONENT	3.36E-11	1.55E-13
46	GO:0030529	intracellular ribonucleoprotein complex	CELLULAR_COMPONENT	3.36E-11	1.55E-13
47	GO:0032991	macromolecular complex	CELLULAR_COMPONENT	1.22E-08	7.29E-11
48	GO:0009507	chloroplast	CELLULAR_COMPONENT	2.01E-07	1.61E-09
49	GO:0043228	non-membrane-bounded organelle	CELLULAR_COMPONENT	3.31E-06	3.06E-08
50	GO:0043232	intracellular non-membrane-bounded organelle	CELLULAR_COMPONENT	3.31E-06	3.06E-08
51	GO:0044434	chloroplast part	CELLULAR_COMPONENT	3.09E-04	3.52E-06
52	GO:0044435	plastid part	CELLULAR_COMPONENT	3.34E-04	3.86E-06
53	GO:0005198	structural molecule activity	MOLECULAR_FUNCTION	3.04E-05	3.08E-07

Table 9: Enriched GO terms for Conserved Plastid-Targeted Clusters identified using UCLUST. Clusters containing at least 13 species with predicted or likely plastid-targeted sequences were mined for common GO terms and compared against terms extracted for the total set of UCLUST -derived clusters using BLAST2GO. All terms enriched above $p=1.0E^{-5}$ in core plastid-targeted clusters are presented below.

	GO term	Description	Ontology	P-value	FDR
1	GO:0008152	metabolic process	BIOLOGICAL_PROCESS	3.36E-32	9.19E-35
2	GO:0015979	photosynthesis	BIOLOGICAL_PROCESS	1.24E-29	3.62E-32
3	GO:0044710	single-organism metabolic process	BIOLOGICAL_PROCESS	1.38E-21	4.57E-24
4	GO:0044711	single-organism biosynthetic process	BIOLOGICAL_PROCESS	5.52E-16	2.15E-18
5	GO:0044699	single-organism process	BIOLOGICAL_PROCESS	2.89E-15	1.18E-17
6	GO:0006091	generation of precursor metabolites and energy	BIOLOGICAL_PROCESS	1.15E-13	4.93E-16
7	GO:0005975	carbohydrate metabolic process	BIOLOGICAL_PROCESS	1.20E-10	5.84E-13
8	GO:0006629	lipid metabolic process	BIOLOGICAL_PROCESS	1.33E-06	7.01E-09
9	GO:0051234	establishment of localization	BIOLOGICAL_PROCESS	1.85E-04	1.23E-06
10	GO:0006810	transport	BIOLOGICAL_PROCESS	1.85E-04	1.20E-06
11	GO:0051179	localization	BIOLOGICAL_PROCESS	2.65E-04	1.81E-06
12	GO:0016043	cellular component organization	BIOLOGICAL_PROCESS	2.98E-04	2.10E-06
13	GO:0044723	single-organism carbohydrate metabolic process	BIOLOGICAL_PROCESS	3.01E-04	2.23E-06
14	GO:0071840	cellular component organization or biogenesis	BIOLOGICAL_PROCESS	4.29E-04	3.26E-06
15	GO:0042592	homeostatic process	BIOLOGICAL_PROCESS	8.59E-04	6.87E-06
16	GO:0009536	plastid	CELLULAR_COMPONENT	1.01E-165	1.97E-169
17	GO:0044464	cell part	CELLULAR_COMPONENT	1.54E-140	6.02E-144
18	GO:0005623	cell	CELLULAR_COMPONENT	1.52E-139	8.87E-143
19	GO:0044444	cytoplasmic part	CELLULAR_COMPONENT	8.76E-120	6.83E-123
20	GO:0005737	cytoplasm	CELLULAR_COMPONENT	3.57E-119	3.49E-122
21	GO:0044424	intracellular part	CELLULAR_COMPONENT	2.06E-110	2.41E-113

22	GO:0005622	intracellular	CELLULAR_COMPONENT	1.27E-104	1.73E-107
23	GO:0043229	intracellular organelle	CELLULAR_COMPONENT	6.39E-93	1.05E-95
24	GO:0043226	organelle	CELLULAR_COMPONENT	6.39E-93	1.12E-95
25	GO:0043231	intracellular membrane-bounded organelle	CELLULAR_COMPONENT	6.87E-82	1.34E-84
26	GO:0043227	membrane-bounded organelle	CELLULAR_COMPONENT	1.90E-81	4.07E-84
27	GO:0009579	thylakoid	CELLULAR_COMPONENT	4.05E-39	9.47E-42
28	GO:0016020	membrane	CELLULAR_COMPONENT	4.66E-36	1.18E-38
29	GO:0071944	cell periphery	CELLULAR_COMPONENT	7.58E-11	3.55E-13
30	GO:0005886	plasma membrane	CELLULAR_COMPONENT	1.32E-07	6.69E-10
31	GO:0009507	chloroplast	CELLULAR_COMPONENT	3.01E-04	2.18E-06
32	GO:0005840	ribosome	CELLULAR_COMPONENT	5.00E-04	3.90E-06
33	GO:0003824	catalytic activity	MOLECULAR_FUNCTION	9.90E-19	3.67E-21

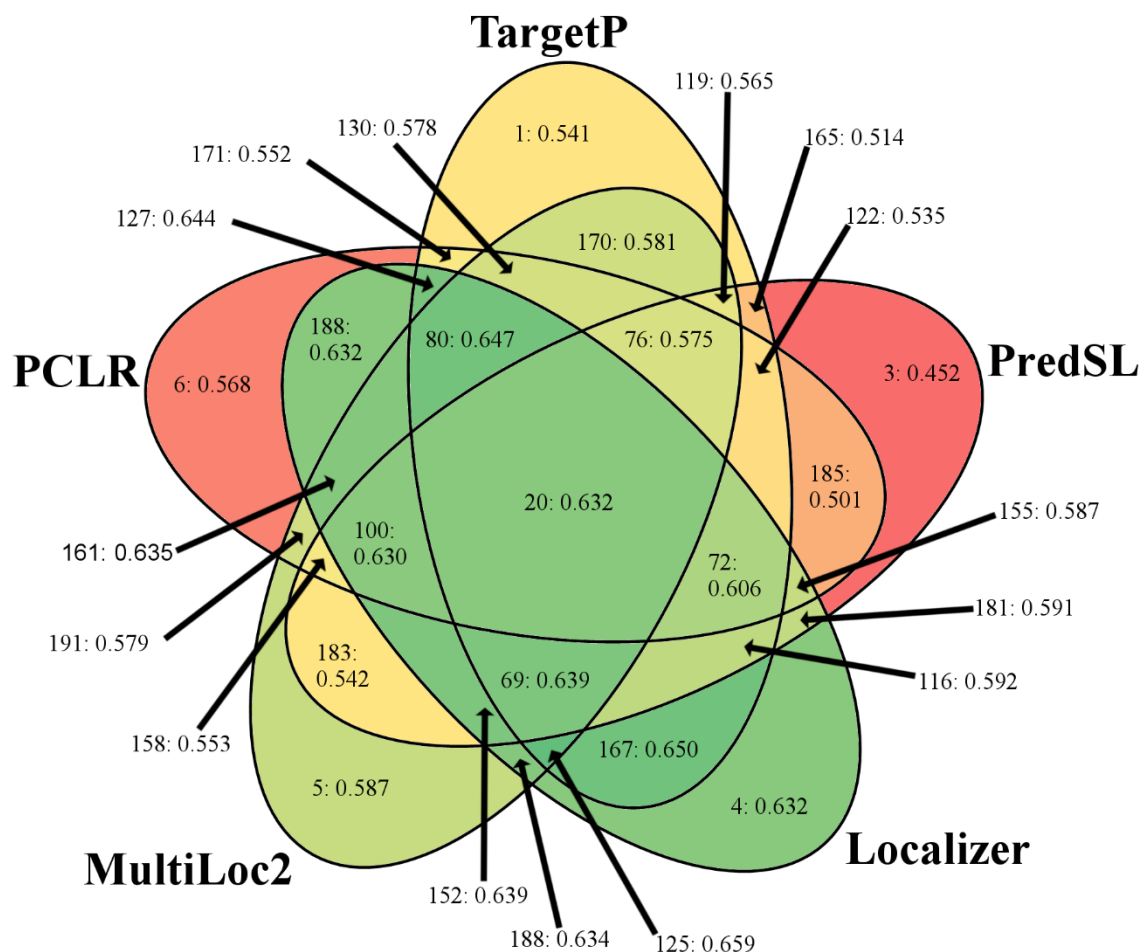


Figure 1: Venn-Diagram of Combinatorial and Standalone Subcellular Prediction

Algorithms. Performance measured by MCC on proteins with subcellular localization validated by GFP is represented as a heatmap with high values in green and low values in red. For each intersection, only the best accept threshold is represented. Numbers indicate workflow number followed by the calculated MCC.

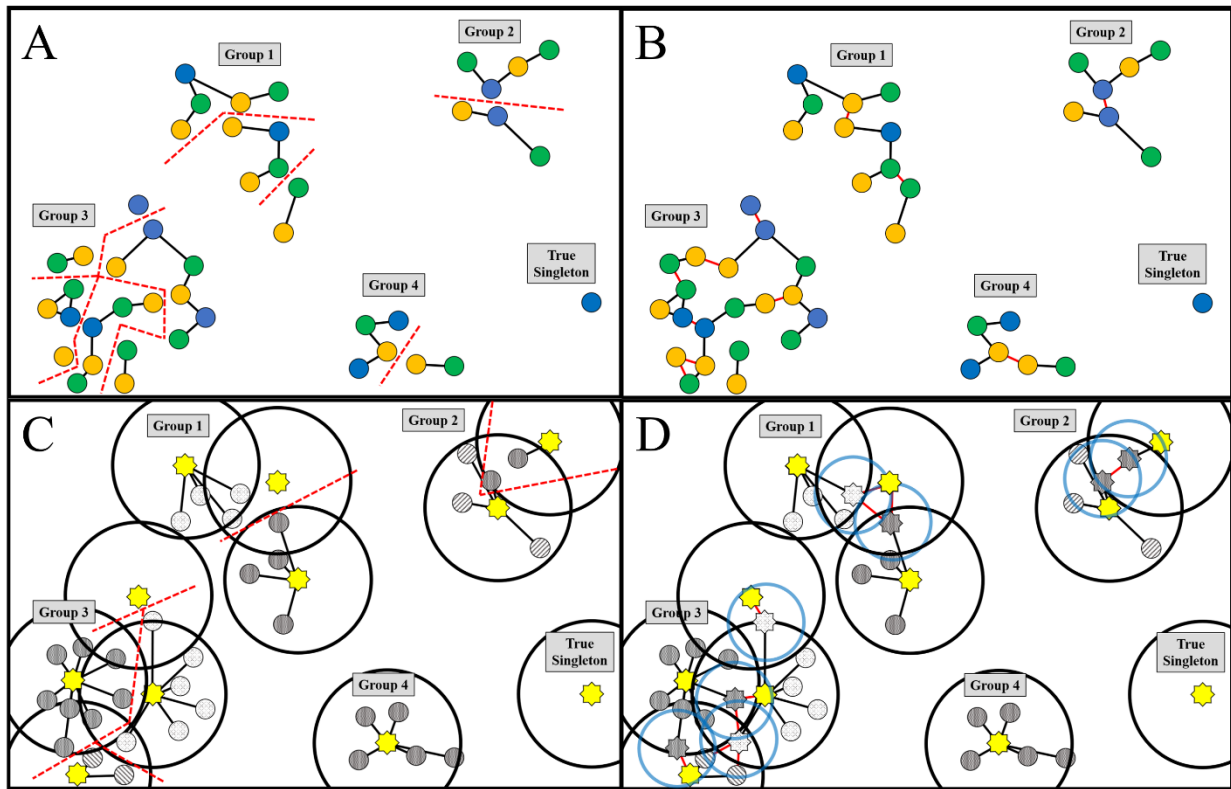


Figure 2: Illustration of RBH and UCLUST Sequence Clustering Methods. Initial (A) and expanded (B) RBH figures indicate clustering between genotypes 1 (blue circles), 2 (green circles), and 3 (orange circles). Bidirectional best BLAST hits between sequences from different genotypes are indicated with black lines; bidirectional better BLAST hits between sequences within the same genotype with red lines and fragments with dotted red lines. For UCLUST, the initial length-sorted (C) run is illustrated with yellow stars indicating centroids, small gray patterned circles indicating non-centroid sequences, large black circles indicating the range for initial centroids, and black lines indicating initial cluster connections. For clarity, sequences are patterned to indicate belonging to each initial cluster, and red dotted lines indicate cluster fragmentation. Centroid randomization (D) mitigates this problem; gray patterned stars indicate randomly-selected centroids, light blue circles indicate the match range for randomly-seeded centroids, and red lines indicate new matches found with red lines. Distances not drawn to scale.

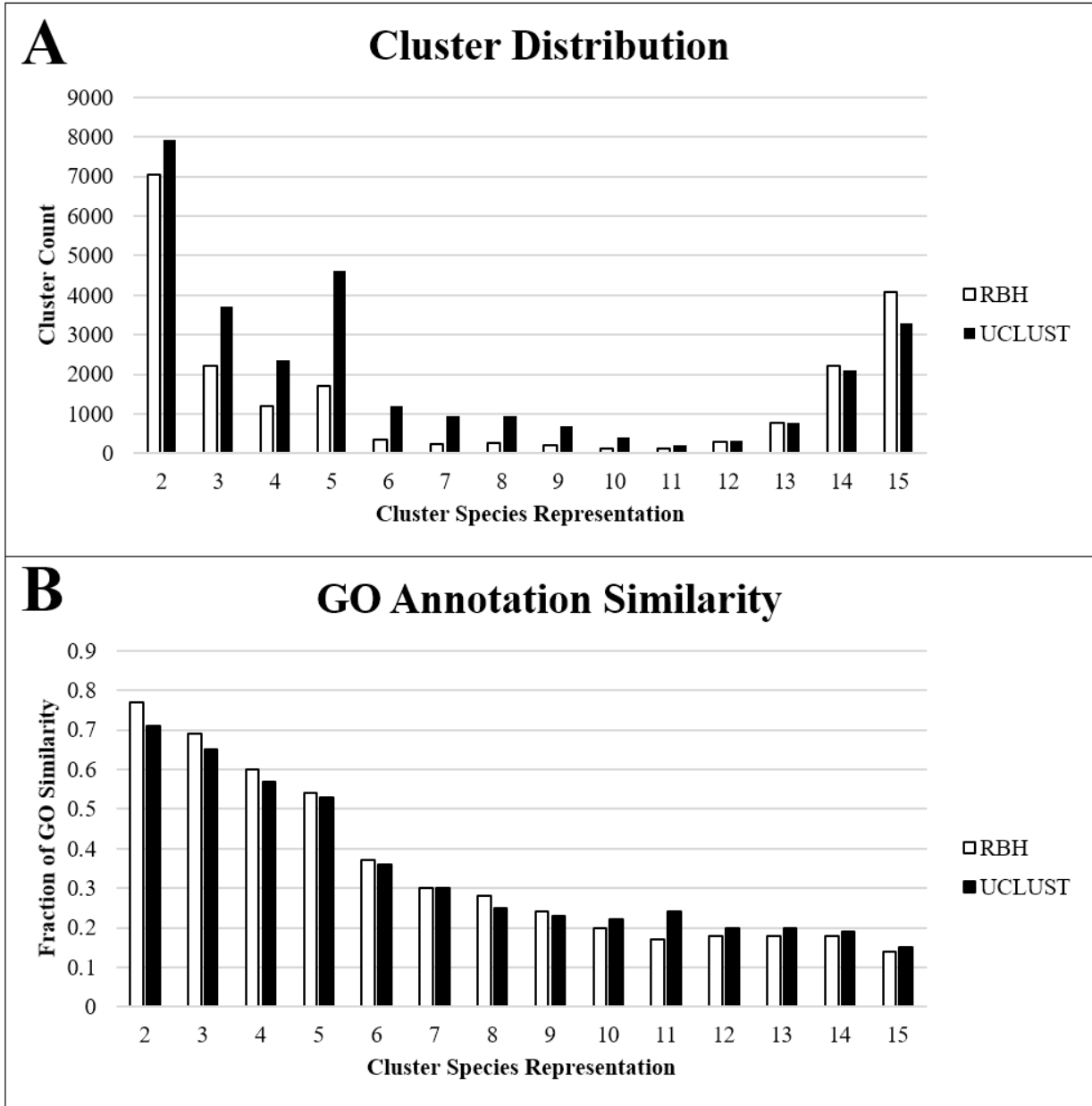
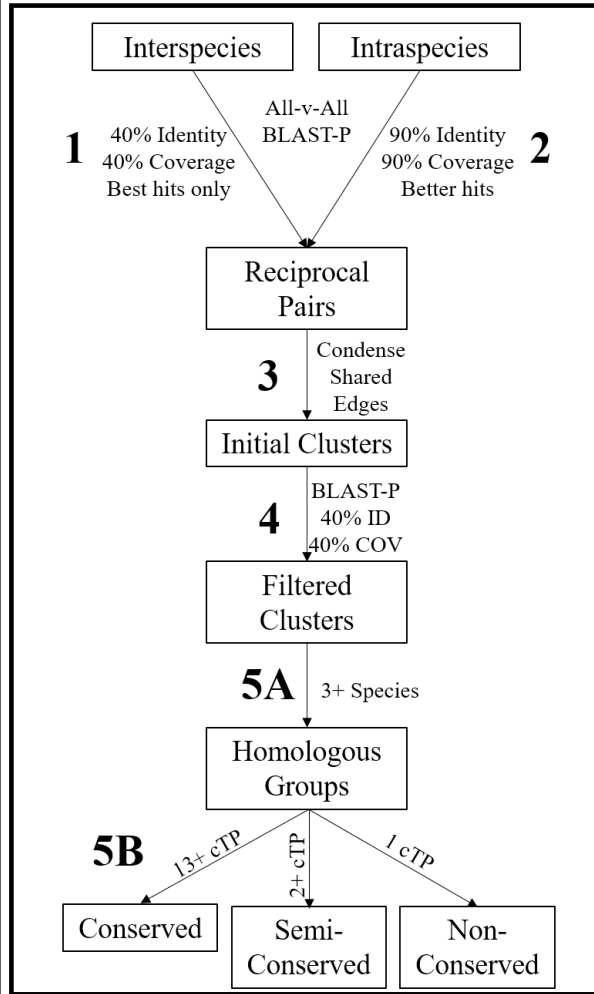


Figure 3: Overall Performance of RBH and UCLUST methods. (A) Cluster distribution in RBH and UCLUST. Both methods resulted in similar distributions of clusters, although RBH resulted in slightly more clusters with 13-15 species and UCLUST resulted in more clusters from 2-12 species. The slight increase in clusters with five species is interesting and may result from sequences with homology within the Poaceae family or within Rosids but with no significant homologs outside those groups. (B) GO annotation similarity in RBH and UCLUST clusters. Lower similarity scores in higher-order clusters are partially due to different annotation methods and thresholds used for different species. Annotation similarity was generally higher in RBH at smaller cluster sizes and higher in UCLUST for larger clusters. Similarity decreased with the increasing representation of species, which may be partially caused by different annotation methods used for different genome sequencing projects or may alternatively be caused by decreased homology within large clusters.

RBH Workflow



UCLUST Workflow

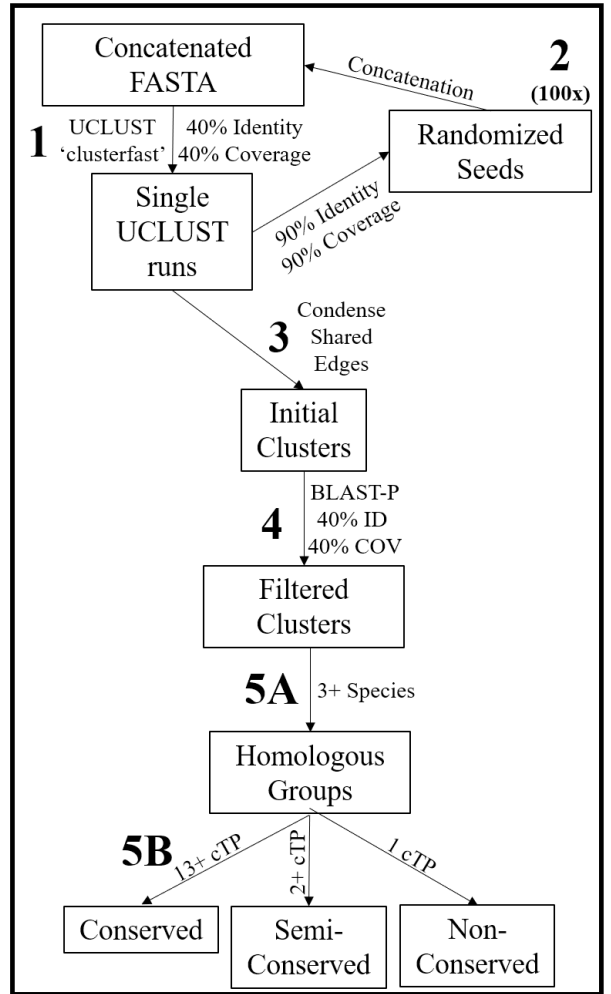


Figure 4: Workflow Diagram of Sequence Clustering Methods. For RBH (left panel), 1. initial cluster edges were generated by finding all reciprocal best-BLAST hits in all-v-all comparisons of proteomes from two separate species at a 40% identity, 40% coverage threshold, and 2. Secondary cluster edges were generated by finding all reciprocal better-BLAST hits in all-v-all comparisons of each proteome against itself at a 90% identity, 90% coverage threshold. For UCLUST (right panel), 1. An initial run was performed at 40% identity and 40% coverage threshold on a FASTA file containing sequences from every species in length-sorted order, and 2. Random sequences of at least 90% identity and 90% coverage were extracted from each cluster, this subset was length-sorted, and then the original length-sorted FASTA file was concatenated to the new seed sequences. This process was iterated 100 times, and a separate UCLUST run was performed for each iteration. Downstream processes for RBH and UCLUST were identical: 3. All clusters/pairs with a shared sequence were condensed into single clusters, 4. All sequences that failed to have at least 40% identity and 40% coverage based on BLAST-P analysis to any of the predicted plastid-targeted sequences in the cluster were trimmed out, 5A. all clusters with at least three species were extracted, and 5B. Clusters containing plastid-targeted sequences were sorted into “conserved,” “semi-conserved,” and “non-conserved” groups according to the number of species with predicted plastid targeting and the taxonomic grouping of those species.

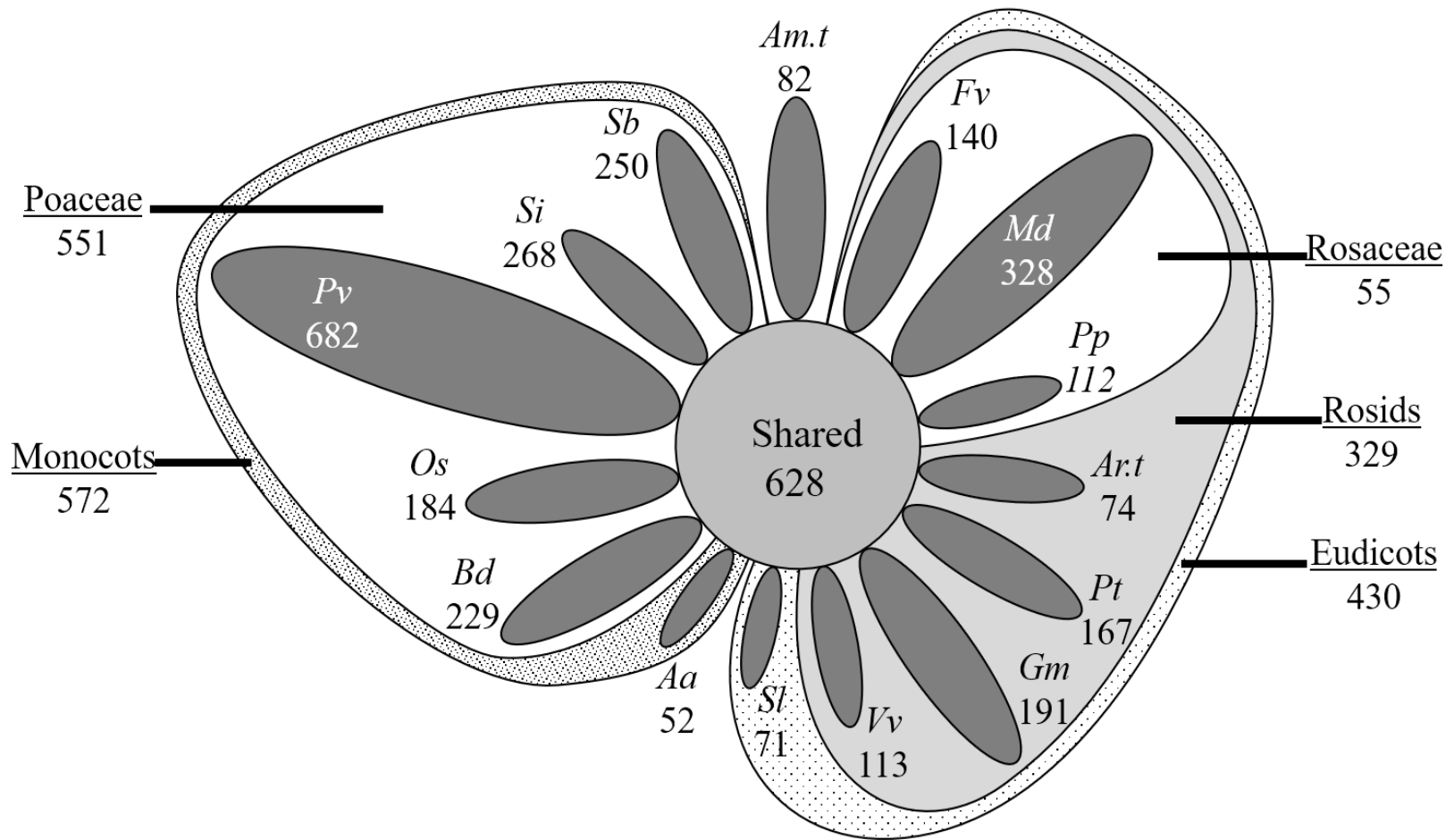


Figure 5: RBH Visual Representation. For “unique” clusters, single-species and singleton clusters are not represented, leaving only clusters with non-targeted homologs present in other species. The relative size of these unique clusters is represented by the area of the respective circle. Shared protein groups at the kingdom, clade, subclade, and family levels are not represented by figure size. Overall, 628 protein clusters were shared between all 15 species, 1,002 had plastid-targeting specific to either Monocots or Eudicots, and 2,943 had plastid-targeting specific to only a single species.

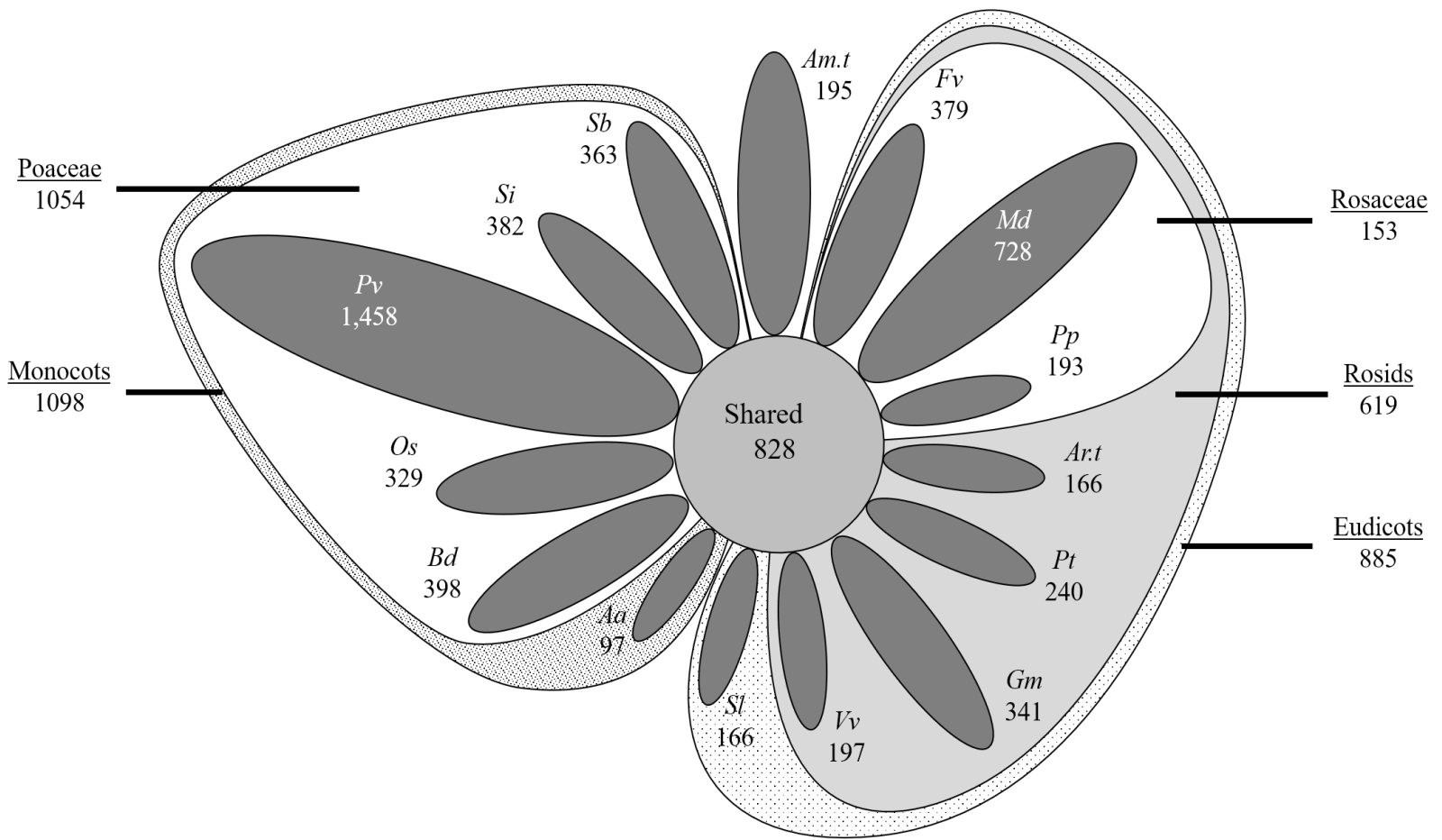


Figure 6: UCLUST Visual Representation. For “unique” clusters, single-species and singleton clusters are not represented, leaving only clusters with non-targeted homologs present in other species. The relative size of these unique clusters is represented by the area of the respective circle. Shared protein groups at the kingdom, clade, subclade, and family levels are not represented by figure size. Overall, 828 protein clusters included plastid-targeted sequences from all 15 species, 1,983 had plastid-targeting specific to Monocots or Eudicots, and 5,632 had plastid-targeting specific to a single species.

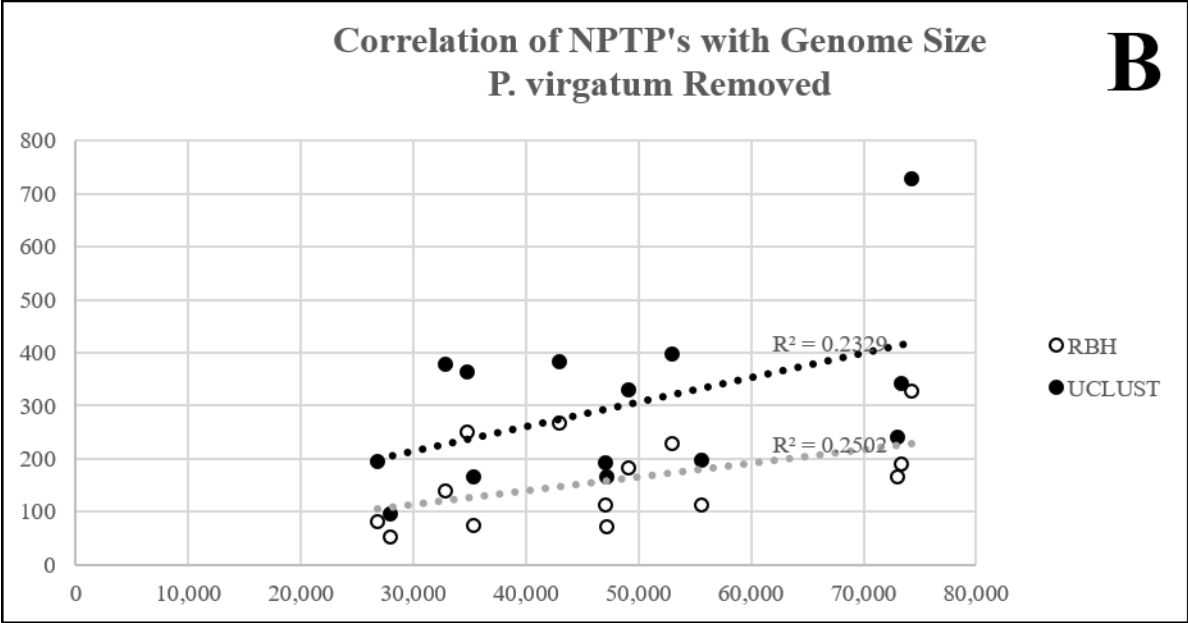
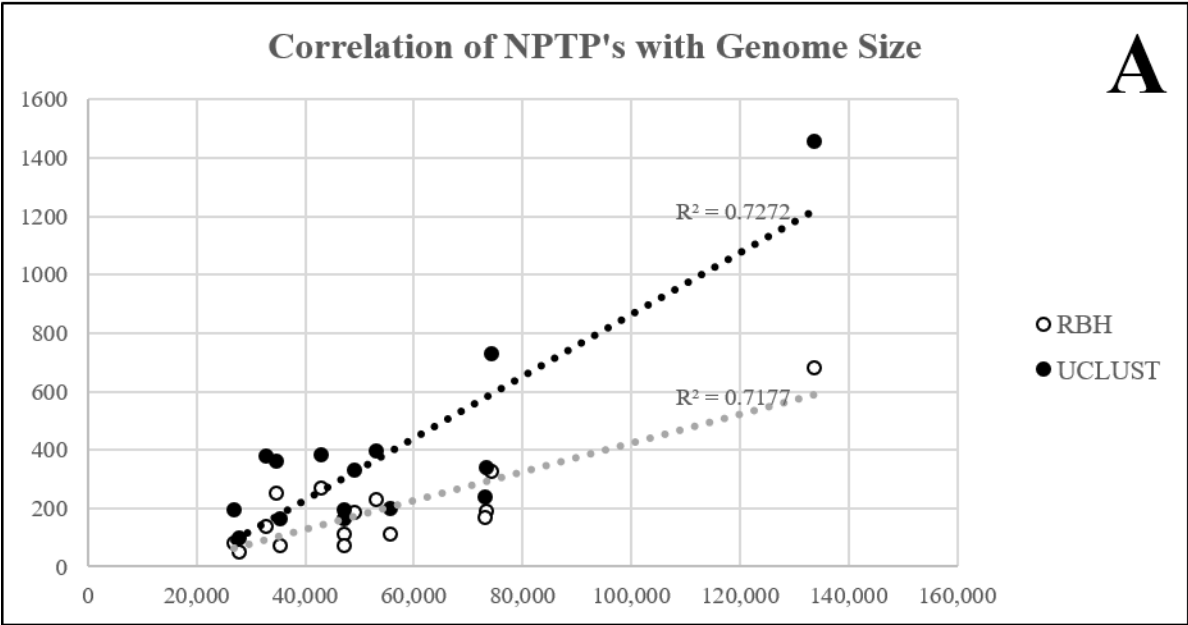


Figure 7: Correlation of Total Proteome Size with Nascent Plastid-Targeted Proteins

(NPTPs). Clusters containing at least three species and with predicted plastid-targeted proteins in only one species were compared to the total genome size for both RBH and UCLUST clustering methods. Although the correlation was moderately linear when *P. virgatum* was included (A), its extremely large genome skewed results. Removal of this genotype resulted in only weakly linear correlation (B), indicating that the evolution of novel transit peptides is a random process.

Additional Files

Additional File 1 – Supplementary Tables and Figures Describing Data Sources and Methods Development of RBH and UCLUST Modifications

Additional File 2 – Combinatorial Subcellular Prediction Analysis

Compressed ZIP file containing Excel documents which summarize bioinformatics prediction of Plastidial and Nonplastidial sequences and calculate performance of single and combinatorial approaches.

Additional File 3 – Proteome Data for 15 Phylogenetically Diverse Angiosperm Species

Compressed ZIP file containing Excel files for each species used in the subcellular prediction and clustering analysis. Full information including protein sequence, targeting prediction, and mapping information to RBH and UCLUST clusters is provided for each protein header.

Additional File 4 – RBH clustering results

Excel file containing all clusters with at least one predicted plastid-targeted sequence detected using RBH methods. Conserved, non-conserved, and semi-conserved clusters are provided along with gene names and synonyms for Arabidopsis sequences in conserved, shared clusters.

Additional File 5 – UCLUST clustering results

Excel file containing all clusters with at least one predicted plastid-targeted sequence detected using UCLUST methods. Conserved, non-conserved, and semi-conserved clusters are provided along with gene names and synonyms for Arabidopsis sequences in conserved, shared clusters.

Additional File 6 – *Malus de novo* transcriptome

Excel File containing the *de novo* transcriptome assembly including sequence information, translated protein sequence, and homology analysis against the Velasco et al. (2010) predicted gene set.

CHAPTER 3

Plastid Transit Peptides - Where Do They Come From and Where Do They All Belong? Assessment of Chloroplast Transit Peptide Evolution in Multi-Species and Pan-Genomic Comparisons

Ryan Christian^{1,2}, Seanna Hewitt^{1,2}, Grant Nelson^{1,2}, Eric Roalson^{1,2,3} and Amit Dhingra^{1,2}

Target Journal: Genome Biology

¹Department of Horticulture, Washington State University, Pullman, WA

²Molecular Plant Sciences, Washington State University, Pullman, WA

³School of Biological Sciences, Washington State University, Pullman, WA

[§]Corresponding author

RWC: ryan_christian@wsu.edu

AD: adhingra@wsu.edu

Abstract

Subcellular relocalization of proteins can significantly affect the function of genes. It is expected to impact ecological adaptation, but the phenomenon is not yet well understood. In plants, the plastid is a major recipient of nuclear-encoded proteins and exhibits significant biochemical, morphological, and spatiotemporal variation. Evidence from bioinformatics and high-throughput shotgun proteomics has suggested that only a relative minority of plastid-targeted proteins are shared by all higher plants, while non-conserved or semi-conserved plastid-targeted proteins comprise a larger share of the plastid proteome in any individual plant species. However, not much is known about the molecular mechanisms or evolutionary forces behind the

evolution of novel plastid-targeted proteins. To better understand the properties of plastid transit peptides, the distribution of amino acids in the transit peptides of known and predicted chloroplast-targeted proteins was examined, revealing several enriched amino acids as well as distinct patterns in residue position that were different for monocot and eudicot sequences. The mutational changes responsible for *de novo* evolution of chloroplast transit peptides were examined to study how novel chloroplast transit peptides evolve *in vivo* using three datasets: A Multi-genome dataset of fifteen phylogenetically diverse species, the Arabidopsis pan-genome, and the *Brachypodium* pan-genome. While gene duplication is not strictly required for the evolution of novel subcellular targeting, 40% of novel transit peptides were found to be most closely related to a sequence within the same genome, and of these, 30.5% result from alternative transcription or translation initiation sites. Finally, insertions or deletions were found to be the dominant mechanism of acquiring novel transit peptides, with more than half being due to alternative start sites.

Introduction

Minor changes in the genome or the epigenome provide the raw genetic material and drive evolution through natural selection, ultimately compounding to observable phenotypes for the whole organism (Nei, 2007). Three primary mechanisms are considered as the critical factors of gene evolution. First, primary effects caused by a mutation to the catalytic and interaction sites of proteins, or changes in the folding pattern of the protein can occur through the introduction of single nucleotide polymorphisms (SNPs), insertions or deletions (indels), or splice site alterations. Second, transcript-level cis-mutations or epigenetic modifications can change the expression of the gene without changing the coding sequence. Finally, protein-level

mutations which affect the rate of translation and post-translational stability including codon usage, ribosomal binding affinity, and protein stability can change the steady-state levels of a protein (Chen, 2007; Nei, 2007; Tokuriki and Tawfik, 2009; Wang et al., 2004; Xia and Levitt, 2002). A fourth and a less emphasized mechanism in the evolution of genes is a change in the subcellular localization of the gene product (Byun-McKay and Geeta, 2007; Davis et al., 2006; McKay et al., 2009). Relocalization to different organelles can occur by modifications to the properties of the mature protein, for instance by addition or removal of transmembrane domains, or by gain or loss of specific targeting motifs including nuclear localization signals (NLSs), endoplasmic reticulum and secretory pathway signal peptides (SPs), and mitochondria or chloroplast transit peptides (mTPs and cTPs, respectively). Signal and transit peptides are typically appended to the N-terminus of preproteins and consist of loosely conserved motifs lacking distinct catalytic sites. However, they possess broad biochemical properties that facilitate interactions with chaperones and import receptors to facilitate translocation. Because of this loose conservation, transit peptides are more likely to evolve *de novo* compared with functional domains of catalytic proteins (Tonkin et al., 2008). Alteration to these sequences can cause mistargeting or ‘subcellular relocalization’ which alters the functional environment of the mature protein and can lead to novel effects even without changing the mature protein sequence (Byun-McKay and Geeta, 2007; McKay et al., 2009). Plants and other photosynthetic eukaryotes (algae and protists) may be more affected by changes to subcellular localization due to the presence of chloroplasts, cell wall, and phragmoplasts as distinct subcellular organelles, as well as from partial overlap in the specificity of mitochondria and chloroplast transit peptides (Pujol et al., 2007). Newly acquired localization patterns may, therefore, contribute actively to the evolution

of plant species, and the plastid is potentially the most dynamic organelle in terms of its proteomic diversity (Christian et al., 2019, unpublished; Schaeffer et al., 2014).

Plastids are highly variable organelles which can assume myriad forms and functions including the archetypical chloroplast. These changes follow a branched family of plastid classes originating with the "proplastid" found in the embryonic and meristematic tissues, which is non-specialized and assumes different morphological forms based on the fate of the tissue during cellular differentiation (Solymosi and Keresztes, 2013). Among the differentiated forms of plastids, the green, photosynthetic chloroplasts are the dominant and most complex form. Many other morphotypes including etioplasts, chromoplasts, leucoplasts, gerantoplasts, and tannosomes perform nonphotosynthetic functions in specialized tissues or at certain stages of development (reviewed in Solymosi and Keresztes, 2013). There is significant diversity even within individual plastid morphotypes: chromoplasts, for example, have five distinct morphotypes in different fruits and flowers (Li and Yuan, 2013). A recent analysis of plastids in fruit peel of apple (*Malus × domestica*) has shown that epidermal plastids adopt a distinct morphology during development which is distinct from collenchymal plastids (Schaeffer et al., 2017; Solymosi and Keresztes, 2013). Plastids adopt distinct morphology in different cultivars, at different points during fruit development, or between different cell layers in the fruit peel, pointing to dynamic and specialized functions in apple fruits (Schaeffer et al., 2017). Similar work has not yet been described in fruits of other species, but various distinctive plastid morphotypes have been described in fruit or other specialized tissues of other species (Solymosi and Keresztes, 2013; Wang et al., 2013). New plastids which play important roles in plant biology continue to be discovered, such as the recently-described phenyloplast which accumulates phenylglucosides in vanilla orchid (Brillouet et al., 2014).

Differences in plastid morphology are accompanied by highly dynamic biochemistry: the highly reductive environment of the plastid facilitates the synthesis of many compounds with agronomic, culinary, medicinal, or therapeutic value. Most major biochemical classes including alkaloids, polyphenols, flavonoids, and terpenoids utilize precursors generated in the plastids, notably isoprene precursors from the MEP pathway (Rodriguez-Concepcion and Boronat, 2002), and aromatic rings from the shikimate pathway (Herrmann and Weaver, 1999). Flux through these pathways can significantly influence the concentration of downstream compounds. Enzymes involved in the biosynthesis of downstream secondary metabolites are also often plastid-localized. For instance, straight-chain esters derived from fatty acids in the plastids are responsible for the characteristic flavors and aromas of many temperate fruits (Dixon and Hewett, 2000; El Hadi et al., 2013; Song and Bangerth, 2003). Monoterpenes and sesquiterpenes including aromatic and essential oils such as pinene and limonene are produced primarily in leucoplasts of secretory glands, glandular trichomes, and resin ducts (Carde, 1984; Cheniclet and Carde, 1985; Markus Lange and Turner, 2013). In hops, humulone and its resulting bitter acids are similarly produced in plastids through the terpenoid metabolic pathway (Clark et al., 2013; Wang et al., 2008). Critical steps in the biosynthesis of taxol (Walker and Croteau, 2001) and vincristine (De Luca and Cutler, 1987) occur in plastids. CCD2, an enzyme which participates in the biosynthesis of safranal and crocin in saffron crocus, is also localized to the plastid (Demurtas et al., 2018). For many essential metabolites, the biosynthetic pathways are not fully understood, but a central role of plastids is likely responsible for metabolites which accumulate in glandular trichomes, such as terpenes in the case of cannabidiol (CBD) production and mint essential oils, or secretory ducts in the case of conifer resins (Duke et al., 2000; Mahlberg and Kim, 2004).

The nucleus encodes over 95% of the estimated 2,000-3,050 plastid proteins (Armbruster et al., 2011; Millar et al., 2006; Richly and Leister, 2004). *In silico* prediction methods have reported that gain or loss of transit peptides in gene homologs is widespread (Richly and Leister, 2004; Schaeffer et al., 2014), and high-throughput proteomics experiments have reported a high variability in the proteomes of plastids both between different plastid morphotypes and across species (Suzuki et al., 2015; Wang et al., 2013). An optimized bioinformatics pipeline was recently utilized to confirm that non-conserved, species- or taxa-specific proteins in the plastid proteome are more abundant than conserved plastid-targeted proteins. Altogether, conserved proteins accounted for less than 15% of the total number of plastid-targeted pan-proteome, and within individual species, they are outnumbered by semi- to non-conserved proteins by a factor of 2-to-1 (Christian et al., 2019, unpublished). The impact of differentially-targeted proteins is almost wholly unexplored, but individual cases can cause significant phenotypes. For instance, loss of a chloroplast transit peptide in the CA3 isoform of carbonic anhydrase is partially responsible for the development of C4 photosynthesis in the *Flaveria* genus of sedges (Clayton et al., 2017; Tanz et al., 2009). In *Amaranthus tuberculatus*, a 3-bp deletion in PPX2L causes dual-localization to both mitochondria and chloroplasts, conferring resistance to herbicides targeting protoporphyrinogen oxygenase (Patzoldt et al., 2006). In the Brassicaceae family, a homomeric acetyl-CoA carboxylase that is typically found as a cytosolic protein has evolved a plastid-targeted isoform that partially duplicates the function of heteromeric ACCase that is partially encoded by the plastid-encoded gene *accD* (Roesler et al., 1997; Yu et al., 2017). This relocalization of ACCase is involved in lipid biosynthesis of many important oilseed crops. Additionally, this retargeted protein makes Brassicaceae plants partially resistant to antibiotics

which target plastid translation activity such as spectinomycin, hampering plastid transformation efforts (Yu et al., 2017).

A variety of mechanisms at both the DNA and transcript level could result in acquisition or loss of subcellular localization peptides through various mechanisms that have been discussed comprehensively in published literature (Byun-McKay and Geeta, 2007; Davis et al., 2006; McKay et al., 2009). At the DNA level, alterations to the nucleotide sequence through the introduction of single nucleotide polymorphisms (SNPs), in-frame indels, or frameshift mutations result in residue substitutions, small peptide insertions or deletions, or modification to downstream sequence, all of which could alter subcellular localization. Substitution mutation within signal peptides occur 3-4 times faster than in the mature protein domains, but half as fast as neutral DNA sequence in taxa as diverse as bacteria, yeast, and mammals (Li et al., 2009; Williams et al., 2000). Even single substitutions could change localization patterns to different organelles (Byun-McKay and Geeta, 2007). In one report, multiple spontaneous mutations restored mitochondrial targeting of an N-terminally truncated F1-ATPase B-subunit (Vassarotti et al., 1987). *De novo* evolution of transit peptides seems unlikely to be a significant mechanism because most transit peptides can be categorized into subgroups based on homolog and domain structure (Lee et al., 2008). However, sequences pre-disposed to subcellular redirection could experience changes in localization relatively easily. For instance, transit peptides of plastids and mitochondria are similar in composition and structure, and many known proteins including aminoacyl-tRNA synthetases, transcription factors, and RNA polymerases are dual-targeted using a single “ambiguous” transit peptide recognized by the translocons of both organelles (Mackenzie, 2005; Pujol et al., 2007). Substitutions in mitochondria transit peptide sequences

could thereby elicit novel plastid-targeting more easily compared with mutation of random N-terminal sequence.

Another strong possibility is that plastid transit peptides are acquired from heterologous loci via exon shuffling either through unequal recombination or movement of retrotransposons. In retrotransposon-mediated exon shuffling, a spliced intron attaches to the spliceosome, reinserts into another mRNA at a protosplice site, and the resulting hybrid product is reverse-transcribed and inserted back into the genome (Vibranovski et al., 2006). This mechanism results in an intron (often class I) between the transit peptide and downstream coding sequence. Class I introns are overrepresented near signal peptide cleavage sites of proteins in both mammals and plants (Nielsen and Wernersson, 2006; Tordai and Patthy, 2004). Furthermore, many plastid transit peptides are encoded by distinct exons (Bruce, 2000). Retrotransposon-facilitated exon shuffling has been implicated in the acquisition of transit peptides in several recently horizontally-transferred mitochondria-targeted genes (Adams et al., 2001; Kadowaki et al., 1996; Nugent and Palmer, 1991; Sandoval et al., 2004; Ueda et al., 2006; Wischmann and Schuster, 1995) and plastid-targeted genes (Arimura et al., 1999; Gantt et al., 1991). Several nuclear-encoded genes have apparently gained a transit peptide via exon shuffling in cases more likely involving unequal recombination rather than the movement of retrotransposons (Long et al., 1996; Schlute et al., 1997; Wolter et al., 1988). Nuclear-encoded proteins involved in mitochondrial DNA and RNA processing are clustered in the genome of *Arabidopsis*, and significant homology in the presequence regions suggests that recombination-based exon shuffling could propagate transit peptides locally (Elo et al., 2003; Kadowaki et al., 1996). Most described cases of exon shuffling involve the transfer of a functional transit peptide to a heterologous gene, but the use of random sequence may also engender similar results, albeit at

reduced efficiency. For instance, 20% of random short DNA sequences from the human genome are capable of functioning as signal peptides when N-terminally fused to a cargo protein (Kaiser et al., 1987). Additionally, 2.7% of clones derived from random *E. coli* sequence and >5% of clones from a cytosolic dihydrofolate reductase gene restored mitochondrial targeting to an N-terminally truncated COX-IV (Baker and Schatz, 1987). Successful clones come from all six reading frames, indicating that functional transit peptides do not need to be derived from the coding sequence. Similar results have been reported *in silico* for the apicoplast, a plastid found in Apicomplexans, in which almost 30% of randomly-generated 24mer peptide sequences with similar residue composition to known apicoplast transit peptides were predicted to be functional, and several of these were validated *in vivo* (Tonkin et al., 2008).

In addition to sequence-level variation, gain or loss of a transit peptide could also occur within a transcriptional unit. Alternative transcriptional or translational start sites could either skip or unmask a buried transit peptide, resulting in dual localization of the two isoforms. The use of alternative start sites in “twinned presequences” is widespread among proteins which are dual-targeted to the chloroplast and mitochondria, (Mackenzie, 2005; Peeters and Small, 2001; Small et al., 1998). In these cases, the most proximal N-terminal transit peptide is active and masks transit peptides downstream of it. A similar mechanism that relies on the use of alternative first exons creates alternative start sites through the use of differential splicing rather than different RNA polymerase or ribosomal binding sites. In mice, alternative first exons are the most common mechanism for differential use of transit peptides in transcriptional units (Davis et al., 2006). However, this may not be consistent across evolutionary space: one report in plants found that only 46 out of 1,159 transcriptional start site variants were generated by the use of alternative first exons (Kitagawa et al., 2005).

Finally, gene duplication could be a significant contributor to the evolution of novel plastid-targeted proteins, especially when the original gene is of critical importance in its native subcellular location. Modern land plant genomes are either polyploid or built on the remnants of older polyploidy events (paleopolyploid), and gene family size is extremely variable as a result (Flagel and Wendel, 2009; Soltis et al., 2009). An earlier report in *Arabidopsis* suggested that 239 gene families had paralogs with divergent subcellular localization (Heilmann et al., 2004), and this figure is likely to be higher in larger genomes. Interestingly, duplicated genes are more likely to be retained if the predicted subcellular localization is different, indicating that many of these re-localized genes are functional and evolutionarily advantageous (Byun and Singh, 2013).

Despite the many anecdotal reports of novel plastid-targeted proteins and the apparent abundance of dual-targeted plastid/mitochondrial proteins, how proteins acquire transit peptides is not well understood. This gap in knowledge was first addressed by examining experimentally validated and predicted plastid transit peptides to identify whether patterns of amino acid frequency and distribution are unique in different species. Transit peptides may evolve via simple substitutions, insertions and deletions, or alternative start sites. To test the hypothesis that transit peptides evolve via these predictable patterns, nascent plastid-targeted proteins (NPTPs) were identified in a diverse range of angiosperms and potential evolutionary mechanisms were tested to determine which, if any, is the most common one. ‘Nascent’ in this case implies an evolutionarily recent origin of the transit peptide which, if validated, may better define the evolutionary moment of transit peptide acquisition and more accurately identify the responsible mutation. To test whether these mechanistic trends also applied to accessions within a single species and to determine the scale of variation at the infraspecies level, the pan-genomes of

Arabidopsis thaliana and *Brachypodium distachyon* were also evaluated for nascent plastid-targeted proteins.

Results and Discussion

Residue Analysis of Experimentally-validated and Predicted Transit Peptides

Although chloroplast transit peptides lack significant homology or functional motifs, they are well-documented to be enriched in certain amino acids and biochemical classes and deficient in others. A comprehensive analysis was conducted using the amino acid content and sequence of a set of 10,868 non-redundant sequences validated by mass spectrometry available in the PPDB (Sun et al., 2009), AT_CHLORO (Ferro et al., 2010), SUBA4 (Hooper et al., 2017), and CROPPAL2 (Hooper et al., 2015) databases. These sequences were primarily derived from *Arabidopsis*, but also contained significant numbers of proteins from rice and maize and a smaller percentage of proteins from other plant species. The transit peptide was defined as the first 60 amino acids of each experimentally-validated or predicted plastid-targeted protein. Most cleaved transit peptides are 41-70 residues long with an average of 51-60 residues (Bienvenut et al., 2012; Huang et al., 2009; Kleffmann et al., 2007; Teixeira and Glaser, 2013), but measuring the length of a transit peptide based on its cleavage site is not a good determinant of its functionality. For one, the mature protein can exhibit constraints on an otherwise functional transit peptide, causing it to lose translocation functionality or efficiency (Rolland et al., 2016). Additionally, the transit peptide residues upstream of the signal peptide peptidase (SPP) cleavage site in many cases is insufficient to target GFP efficiently to the chloroplast, but addition of some downstream sequence restores plastid translocation (Comai et al., 1988; Lee et al., 2008; Pilon et al., 1995; Rensink et al., 1998; Shen et al., 2017; Van't Hof et al., 1991). Finally, transit peptides

must be long enough to both span the length of the membranes in early binding intermediates- a minimum distance of 90 Å which would require at least 32 residues in addition to the terminal TOC and TIC binding elements (Bionda et al., 2010; Chotewutmontri et al., 2012; Richardson et al., 2018). A transit peptide of sixty residues in length enables high-efficiency import in most plastid-targeted proteins, so this size was used as the standard length of transit peptides for this experiment.

The average residue composition of experimentally-validated transit peptides was compared to the average residue content of the whole TAIR10 Arabidopsis proteome, revealing multiple residues with altered abundance (Table 1, Figure 1). Transit peptides were highly enriched in alanine (+64.1%), proline (+52.4%), arginine (+41.6%), serine (+35.5%), and threonine (+11.1%), a composition well established in the literature (e.g. Bruce, 2001; Zybailov et al., 2008). Depletion of the negatively-charged glutamic acid (-46.3%) and aspartic acid (-43.4%) and of the aromatic amino acids tryptophan (-36.8%), and tyrosine (-38.6%) were also observed, but strikingly, phenylalanine was not highly different (-7.9%). Less severe depletion of isoleucine (-24.3%), asparagine (-28.6%), lysine (-22.4%), and glutamine (-16.1%) were seen relative to the average TAIR residue composition, none of which have been documented before in plastid transit peptides. The decrease in average lysine content is unusual, as early reports suggested that the C-terminal or distal end of transit peptides is enriched in positively-charged residues including both arginine and lysine (Bruce, 2000; Zhang and Glaser, 2002). Additionally, glycine and valine- although abundant in transit peptides at about 6% of residues each- were underrepresented in comparison to average proteome sequence, indicating that they are relatively neutral to the function of plastid transit peptides. Finally, methionine was increased (+21.8%), although most if not all of this increase is likely from the enrichment of methionine as the

starting codon. Overall, these data suggest that bulky R groups are generally not favored in transit peptides unless they serve a special purpose, such as the highly-enriched arginine which is proposed to interact with the TOC GTPases TOC33 and TOC159 (Jelic et al., 2003; Vetter and Wittinghofer, 2001). The relatively unchanged proportion of phenylalanine is therefore puzzling; it comprises from 4-5% of the amino acids at positions 4-42 and decreases to an average frequency of 3.6% after that position. At this point, the role of phenylalanine is not clear given its higher abundance in transit peptides.

Plastid transit peptide composition is not homogenous, but is instead organized in three major domains as described by the homology block (Karlin-Neumann and Tobin, 1986; Quigley et al., 1988), modular domain (Bruce, 2001, 2000) and Multi-Order Multi-Selection (M&M) (Li and Teng, 2013) models of transit peptide structure. These three domains include an uncharged N-terminal proximal region, a central domain rich in hydroxylated residues and lacking acidic residues, and a C-terminal distal region enriched in arginine (Bruce, 2001), corresponding roughly to regions of the transit peptide which interact primarily with the TIC translocase and HSP motor complex, the TOC75 POTRA domains, and the TOC33/159 GTPases, respectively (Richardson et al., 2018). The positional bias of each amino acid was therefore observed for both experimentally validated transit peptides confirmed using mass spectrometry, and in putative transit peptides predicted by a combination of TargetP and Localizer (Table 1, Figure 2). For all residues, nearly identical distribution patterns were observed between experimental (solid lines) and predicted plastid transit peptides (dotted lines), thus confirming that the prediction methods selected in this study corresponded well with experimentally validated residues. However, the actual frequency was often somewhat different, as predicted transit peptides had higher frequency of the more abundant amino acids compared with experimentally-validated proteins,

especially serine, arginine, and proline. Serine was nearly 5% more frequent (absolute frequency) in predicted transit peptides despite the higher proportion of monocot sequences that should dilute its frequency. Proline was also overestimated by 2-4% depending on the position, and arginine was overrepresented by about 2% after position 15. Conversely, predicted transit peptides were underrepresented in rare/neutral or distally-enriched amino acids such as aspartic acid, glutamic acid, glycine, valine, and tyrosine. These observations point to a systemic bias in the software, which may be biased to focus on only the most essential components. However, the patterns and generally close abundances indicate that prediction tools achieve results that are similar to the experimental methods.

Several amino acids including alanine, leucine, serine, threonine, and methionine were more abundant at the proximal end of transit peptides but declined significantly in frequency over its length (Figure 2). Of these, all except methionine were found at initial frequencies greater than 5%. Serine alone is initially present at about 17% frequency, while the small nonpolar amino acids alanine and leucine together comprise 25% of residues in the proximal third of the transit peptide. A second set of centrally-enriched amino acids including proline, arginine, phenylalanine, and histidine are initially low in abundance, rise to a peak value between positions 10-30, then decrease in frequency across the remaining length. Arginine reaches a peak value somewhat later than the other three, peaking instead between positions 20-40. Histidine was rare among this group at only 1.5% at the proximal end and increasing to a high of only 3% at position 20, though its trend matched that of other centrally-enriched residues. A third group of residues which were initially rare but increased in frequency across the length of the transit peptide and reached a peak in the distal end included lysine, aspartic acid, glycine, and glutamic acid. Both glutamic and aspartic acids followed a small but consistent trend, starting at about

3.5% each for every position at the beginning of the cTP and increasing gradually to a high of 5% for aspartic acid, and 6% for glutamic acid. To the best of our knowledge, such a trend has not been observed before, but a 2-fold increase of these residues between the proximal and distal ends of the transit peptide is evident in both experimentally-validated and predicted sequences. From what is known of the TOC GTPases, the primary means of selectivity is due to the hypervariable acidic A-domain of the TOC159 family (Inoue et al., 2010; Richardson et al., 2009; Smith et al., 2004). It is possible that the pattern that was observed for negatively-charged amino acids in the distal GTPase-interacting domains of transit peptides is evidence of exclusion motifs that alter import efficiency, perhaps by charge repulsion against the acidic TOC GTPase A-domain (Smith et al., 2004). A final group of residues were either very rare (<2%) or exhibited more moderate fluctuations in frequency across the length of the transit peptide, and included valine, glutamine, cysteine, asparagine, isoleucine, tyrosine, and tryptophan.

Transit peptides of different plant taxa may have differences due to genetic drift, to the binding affinity of the TOC and TIC translocon receptors, or expansion or contraction of gene families for translocon and chaperone subunits. For instance, transit peptides of monocots and eudicots have previously been reported to be enriched in alanine and serine, respectively (Zybailov et al., 2008). Therefore, the amino acid content of sequences from predicted plastid transit peptides of six monocots (*Anthurium amnicola*, *Brachypodium distacyon*, *Oryza sativa*, *Panicum virgatum*, *Setaria italica*, and *Sorghum bicolor*), eight eudicots (*Arabidopsis thaliana*, *Fragaria vesca*, *Glycine max*, *Malus × domestica*, *Populus trichocarpa*, *Prunus persica*, *Solanum lycopersicum*, and *Vitis vinifera*), and the early-diverging angiosperm species *Amborella trichopoda* were compared to determine if these trends held true, or if other amino acid biases occurred in certain taxa (Table 1, Figure 3A). In all genotypes, serine was found to be

overrepresented by between 50-100% compared with the whole proteome, but in eudicots, serine was more abundant with 30% more serine on average compared with monocot transit peptides. In contrast, alanine was marginally enriched in the transit peptides over whole proteome sequence in eudicots but was extremely enriched in monocots, with a minimum of +50.7% enrichment in *A. amnicola* up to a high of +81.1% in *O. sativa* compared to the respective whole proteome. In eudicots, by contrast, a maximum of only +16.7% enrichment was found in *Malus × domestica* (+16.7%), while transit peptides of *Populus trichocarpa* were underrepresented in alanine compared to the whole proteome (-1.0%). Alanine was also the most abundant amino acid of monocot transit peptides for all genotypes except *A. amnicola*, and overall, was enriched by 111.6% compared with eudicot transit peptides, whereas serine was the most abundant amino acid in all eudicot genotypes. While alanine is somewhat enriched in the whole proteome of monocot species, this is not significant enough to explain these results (Figure 3B). Alanine enrichment in monocots was counterbalanced by underrepresentation in phenylalanine (-91.3%), isoleucine (-105.4%), leucine (-121.5%), asparagine (-169.8%), glutamine (-35.6%), and threonine (-34.2%) compared with eudicots. Tyrosine was also significantly underrepresented (-60.1%) in monocots, although its frequency is extremely low in both clades. Eudicots, in contrast, had somewhat lower glycine, proline, and arginine compared with monocots. Overall, transit peptides of monocots contain more small, nonpolar amino acids including glycine, valine, and proline in comparison with eudicots, which have a more flexible amino acid composition. Furthermore, arginine, which is essential in binding and interaction with the TOC GTPases, was relatively higher in monocot sequences (Pilon et al., 1995; Rensink et al., 2000). One possibility that explains these differences is that changes to the translocons of monocots select for more conserved transit peptides. Monocots lack all but one isoform of the core TIC subunit TIC20,

and also lack the suspected TIC component Ycf1 compared with Eudicots (Bölter and Soll, 2017; de Vries et al., 2015; Nakai, 2015). If these components impact import efficiency or selectivity, their loss may favor small, uncharged amino acids in the transit peptides of monocots to minimize steric hindrance and necessitate a higher arginine content to ensure high import efficiency. As research progresses on the non-essential components of TIC such as TIC100, TIC56, and Ycf1, it will be interesting to see if they have a role in increasing import efficiency for transit peptides with unfavorable amino acids.

In *A. trichopoda*, it was expected that the transit peptides would have intermediate residue composition between monocots and eudicots because it is the sister lineage to the combined monocot/eudicot lineage (Albert et al., 2013; Soltis et al., 2009). Surprisingly, however, predicted transit peptides in *A. trichopoda* were nearly identical to eudicots for almost all residues. Slightly intermediate values were observed for some residues (e.g., lysine, asparagine, proline, glutamine, and serine), but even in these cases, *Amborella* was closer to eudicot than monocot sequences. The only residues with a significantly different trend were valine, which was somewhat decreased compared to both monocots and eudicots, and glutamic acid, which was slightly higher than both. This result seems to indicate that monocots have experienced changes to the protein translocation machinery that has selected for different amino acid content in transit peptides.

Evolution of Novel Plastid Transit Peptides in Diverse Angiosperm Species

To test the hypothesis that transit peptides evolve in predictable patterns, NPTPs were first examined among comparative homologous protein clusters detected using either Reciprocal-Best-BLAST hits (RBH) or UCLUST in Christian et al. (2019, unpublished). Plastid targeting prediction was performed using a consensus approach of TargetP and Localizer, which have

been shown recently to be highly efficient at predicting plastid targeting (Christian et al., 2019, unpublished). Clusters containing at least three species and in which only one species had predicted plastid-targeted proteins were identified as “NPTP’s” for each method. These candidate clusters were then compared with the reciprocal method: if both methods found that the NPTP sequence(s) belonged to a cluster in which no other species had a predicted chloroplast transit peptide, that cluster was determined to be robustly-supported. If the reciprocal method found that this sequence was not uniquely plastid-targeted, it was determined to be moderately-supported by only a single method. In total, 1,328 clusters were supported by both methods, 618 clusters were detected using RBH only, and 1,443 clusters in UCLUST only. Phylogenetic trees were constructed for each cluster using MAFFT, Phyutility, and RAxML, and the resulting alignments were examined for nascent chloroplast transit peptides. All clusters containing multiple unlinked branches of chloroplast transit peptides, losses of transit peptides within a branch, or predicted chloroplast-targeted sequences at the root of the phylogenetic tree were removed in order to focus on single, recent transit peptide acquisitions in single genes (Figure 5, Table 2). For each monophyletic tree, the node in which localization prediction diverged was identified and the neighboring sequences analyzed for the responsible mutations (Table 3, Figure 6). Because the phylogenetic techniques used in this study resolve any problems arising from nonhomologous sequence contamination, all three datasets were pooled for comparisons. However, the results from each group were also generated and are summarized in Table 3.

Residue substitutions were the primary evolutionary factor for 31.4% of NPTPs, with an average of 12.3 substitutions per divergent pair (Figure 6A). Substitutions were somewhat concentrated at the proximal N-terminal third of the transit peptides (Figure 6B). Just 34.8% of residue substitutions conserved the same biochemical properties (overall charge, size, and

polarity). Of the remainder, nonpolar to polar substitutions and polar to nonpolar substitutions were most common, at 23.7% and 16.2% of the total, respectively. *De novo* evolution of transit peptides by single or multiple residue substitutions has been suggested as a primary mechanism of transit peptide evolution (Byun-McKay and Geeta, 2007). Signal peptides evolve two times faster than mature proteins on average, and up to 5-6 times faster than random sequence (Williams et al., 2000). Among the sequence pairs in the multi-genome dataset, the transit peptide region shared just 37.9% identity compared with 65.6% identity in the downstream mature protein. The first ten residues of transit peptides are known to strongly influence import efficiency (Chotewutmontri et al., 2012; Chotewutmontri and Bruce, 2015), so simple substitutions in this region could impart novel plastid targeting of the cargo protein as long as positively-charged amino acids are optimally positioned downstream. Similar trends in amino acid enrichment and depletion were observed for both substitutions and insertions/deletions (Figure 6C). Surprisingly, the absolute abundance and distribution of small insertions and deletions was relatively similar, but insertions became much more prevalent after the 20-residue range (Figure 6D). For length mutations affecting the entire 60-residue window of the putative transit peptide, insertions were over three times more common than deletions. On average, 25.6 positions were inserted, and 12.2 positions were deleted for each NPTP sequence alignment, leading to slightly less than half (49.8 %) of all NPTPs being caused by insertions and 29.4% by alternative start sites introduced by an insertion. Alternative start sites caused by a deletion at the proximal end, *i.e.*, where a cTP begins at a downstream position, accounted for only 8.7% of the total. Both alternative start sites (Davis et al., 2006) and exon shuffling (Long et al., 1996; Vibranovski et al., 2006) have been suggested as primary drivers of subcellular relocalization, and evidence supporting both mechanisms was found in this analysis. Most insertions and

deletion mutations occurred at the beginning and aligned on an initial methionine in the shorter sequence, suggesting that they are alternative start sites (Table 3). But alternative start sites did not account for all of cases involving 5' insertions or deletions, suggesting that exon shuffling may also play a significant role. The potential impact of alternative first exons was also examined, but no candidates were detected. However, as the analysis focused on the first 60 amino acids, it is possible that instances in which the first exon was longer were missed. Additionally, only exons of at least ten residues in length were examined for possible alternative exons, which may have excluded microexons (Guo and Liu, 2015).

In 48.6% of clusters, the most closely related predicted non-plastid-targeted sequence was from within the same genome, strongly implying either gene duplication or an alternative transcript in the evolution of a novel transit peptide. Overwhelmingly, novel plastid-targeted sequences came from gene duplication events: only 27.9% of aligned pairs arising from the same species, or 13.6% of the NPTP total, were due to alternative gene products or alleles of the same locus. Because relocalization of the protein from its native environment into the chloroplast would create a *de facto* knockout phenotype, gene duplication or alternative isoforms may be necessary to maintain the evolutionary function of the original gene or transcript while giving flexibility for the duplicated copy to evolve a new function. However, as the majority of clusters were not duplicated, either the current proteomes are not fully annotated, or duplication is not strictly required for subcellular relocalization.

First, to confirm that terms associated with plastids are underrepresented or neutral, and second to find any overrepresented terms to discover functions that are broadly selected for in novel plastid-targeted genes, GO enrichment of NPTPs was conducted. A custom dataset consisting of the full proteomes from each of the included species was used as a reference

dataset. As expected for non-conserved plastid-targeted proteins, terms associated with plastid (GO:0009536), thylakoid (GO:0009579), localization (GO:0051179, GO:0051234), and transport (GO:0006810), were significantly underrepresented, as shown in Table 4. A total of 49 terms were overrepresented in this dataset, almost all of which were associated with metabolism and biosynthesis, regulation of gene expression, and protein binding or regulation. Most biosynthetic terms were associated with primary metabolism, but several terms involved in heterocyclic and aromatic compounds were also identified. The enrichment of these terms confirms an assumption that novel chloroplast-targeted proteins likely contribute to species-specific biochemistry. However, the high enrichment of terms associated with transcriptional processes and protein binding suggests that an equally important function of novel plastid-targeted genes may be in the regulation of plastid gene expression and protein function.

Mechanisms of NPTP evolution in the Arabidopsis Pan Genome

The analysis workflow was applied to pan-genomes to test the hypothesis that the broader trends in transit peptide evolution observed in the multi-genome dataset hold true at smaller evolutionary scales. The Arabidopsis1001 Project (Alonso-Blanco et al., 2016; Cao et al., 2011; Joshi et al., 2012) has generated a wealth of protein allelic variants for 256 diverse *A. thaliana* accessions; these sequences were accessed from the Arabidopsis1001 web portal, and targeting prediction was performed as for the multi-genome sequences. Out of a total of 35,176 gene isoform groups, 928 proteins were detected that contained both predicted plastid-targeted and non-plastid-targeted sequence variants. A monophyletic origin of a plastid transit peptide was found in 180 genes. Surprisingly, only a single deletion was found which deleted a single residue in AT3G06180.1, while no insertions were detected. A second gene, AT3G13820.1, had a C-terminal 4-residue deletion that changed the predicted targeting of the protein, but this appears to

be the result of a nonsense mutation in the predicted NPTP which likely disrupts the functional domains of the original protein. The remaining mutations were all due to residue substitutions (Figure 7A). While these results are significantly different from what was observed in the multi-genome dataset, it should be stressed that the Arabidopsis1001 project contains only allelic variants, and genetic diversity is relatively low in coding sequence: an average of slightly more than 439,000 SNPs, or roughly 1 SNP in every 1 kb, are found on average between any two Arabidopsis accession genomes (Alonso-Blanco et al., 2016). Indeed, only 1.34 substitutions occurred between each pair of sequences observed in the current experiment, and single substitutions were responsible for 71% of predicted NPTPs. Substitutions were heavily concentrated in the proximal third of the transit peptide: 126 substitutions occurred in the first 20 positions, followed by 90 in positions 21-40, and only 37 in positions 41-60 (Figure 7B). Nearly 75% of the substitutions were nonconservative, with nonpolar to polar transitions accounting for over 21% of nonconservative substitutions. Basic to polar substitutions occurred 15.3% of the time, followed by acidic to nonpolar (9.5%), aromatic to polar (8.5%), and nonpolar to basic (7.9%) (Figure 7C). Overall, this pattern follows what had been observed for transit peptide composition (Figures 1, 2), with increases particularly in serine, proline, alanine, and arginine, and decreases in acidic, aromatic, and long polar amino acids.

In addition, the Arabidopsis1001 project identified several major taxonomic sub-groups of accessions that enable tracking of NPTP evolution across both taxonomic and geographic space. Distribution of NPTPs was examined in accessions of different taxonomic groups to test whether the evolution of plastid transit peptides follows similar trends to the genomic diversity of each taxonomic group (Figure 8). It was observed that relict accessions native to the Iberian Peninsula had the most NPTPs per accession, at an average of 10 each, but accounted for a small

number of the total due to relatively few Relict accessions being included in the 256 currently available accessions. Non-Relict Spanish accessions and accessions from the Italian/Balkan/Caucasus group were also diverse, but in terms of average diversity they were closer to other groups than to the relict group. The pan-genome sequencing project suggested that these taxonomic groups are more variable because they survived the last glaciation period, while the other taxonomic groups developed after the species radiated back out from glacial refuges. In contrast, the Asian accessions are at the easternmost edge of the Arabidopsis native range and are the least diverse, with the least NPTPs both as a whole and expressed as average NPTP content per accession. Surprisingly, the admixed group, which contains intermediate characteristics between two or more distinct taxonomic groups, had the largest share of NPTPs overall. In keeping with admixed populations sharing genetics with more distinct taxonomic groups, many NPTPs were not necessarily unique to a single accession but rather shared between a mixture of accessions. Although most NPTP clusters contained only a single accession with a predicted novel plastid-targeted protein, several contained more accessions, up to a maximum of 38 observed for one cluster. Most NPTPs in the admixed accessions derives from these examples where a plastid-targeted allele has propagated through many accessions, rather than from truly unique NPTPs.

GO enrichment of NPTPs in the Arabidopsis1001 demonstrated that significant underrepresentation ($p < 0.05$) of membrane-bound organelle (GO:0043227; GO:0043231), organic substance metabolic process (GO:0071704), and primary metabolic process (GO:0044238) were underrepresented, which is expected for genes with variable plastid targeting (Table 5). In contrast, the majority of overrepresented terms were involved in secondary metabolic and redox processes, such as flavin-containing compound metabolism and

biosynthesis (GO:0042726; GO:0042727), adenylyltransferase activity (GO:0003919; GO:0070566) glycosyl or hexosyl transferase activity (GO:0016757; GO:0016758), and glycerol ether metabolism (GO:0006662). Additionally, processes involved in G-coupled receptor signaling (GO:0007205; GO:0007186) were overrepresented.

Mechanisms of NPTP Evolution in the Brachypodium Pan Genome

Pan-genome sequencing in many species has revealed large variation in duplicated and expendable genes: in some cases, the core genes represent a minority of the total genes for a given species. The BrachyPan project for *Brachypodium distachyon* conducted deep re-sequencing of 54 diverse accessions of *Brachypodium distachyon* to characterize presence/absence variants (PAVs) and copy number variants (CVs), in addition to allelic and isoform variants (Gordon et al., 2017). Predicted proteomes of 56 different *B. distachyon* ecotypes including two internal Bd21-3 controls were accessed from BrachyPan (<https://brachypan.jgi.doe.gov/>) and sequences were arranged into clusters according to a reference matrix file provided by John Vogel, DOE, USA (Personal communication). Gene clusters were analyzed using the same localization prediction process as described for the Arabidopsis1001 dataset, resulting in 8,990 orthologous pan-gene clusters that had at least one predicted plastid-targeted gene and one non-plastid-targeted gene. RAxML was performed on 7,551 of the candidate gene clusters. The most recent common ancestor was likely plastid-targeted in 4,616 of these clusters, indicated by a plastid transit peptide at the root of the phylogenetic tree. Multiple points of transit peptide evolution were found in 1,616 clusters, while 116 clusters had a single point of origin but had a loss of the transit peptide within the same branch. Of the remaining 2,272 clusters, the node corresponding to the evolutionary gain of a transit peptide was extracted, and MUSCLE was used to realign the paired sequences. It is

interesting to note that in both BrachyPan and Arabidopsis1001, plastid transit peptides were more likely to be lost than gained, indicated by a predicted plastid transit peptide at the root of the tree. This pattern has been observed before in signal peptides, where the loss of the signal peptide prevailed over gains by a factor of almost 4-fold (Hönigschmid et al., 2018). This ratio was similar for both pan-genomes. For the remaining sequences, gene variants were much more divergent than the Arabidopsis1001 clusters and more closely mirrored the multi-genome dataset (Figure 9). An average of 12.7 substitutions were found in each divergent pair of sequences, 69.4% of which were non-conservative. Length variants were also common, with an average of 0.80 insertions and 1.08 deletions occurring in each alignment. Despite a large number of substitutions per aligned pair, substitutions were the dominant means of transit peptide acquisition in only 28.3% of clusters, while insertions and deletions were responsible for the remaining 71.7%. Alternative start sites resulting in an insertion for the plastid-targeted protein were found in 20.1% of cases, while alternative start sites resulting in a deletion were found in 11.4% of cases (Figure 9A). Insertions and deletions that did not align with an in-frame methionine and therefore did not represent alternative start sites accounted for 19.6% and 20.6% of cases, respectively. The most frequent size of both insertions and deletions were those that covered the full 60 amino acids of the putative transit peptide. Of the remainder, the most abundant size range in both cases was between 1-5 residues, and frequency declined as gap size increased. More substitutions were found at the beginning of the aligned sequence, although the difference was not nearly as pronounced as observed in Arabidopsis: 38.2% of substitutions were found in the first 20 positions, 34.2% between positions 21-40, and 27.6% between positions 41-60 (Figure 9B). Large increases in proline, arginine, serine, and threonine were observed along with decreases in aspartic acid, glutamic acid, glycine, leucine, valine, and tyrosine, although the

magnitude of these changes was not nearly as drastic as observed for Arabidopsis1001 sequences (Figure 9C). Most residues were not substantially different between modes of mutation, although it is interesting to note that both serine and threonine were somewhat more likely to be caused by substitutions rather than length variants: the net change of serine was +4.4% for (insertions-deletions) and +6.5% for substitutions, while the net changes in threonine were +1.0% and +2.2%, respectively (Figure 9D). Introductions of proline were conversely more likely to occur due to insertions or deletions (+6.4%) compared to residue substitutions (+5.1%). All other amino acids differed by a less significant margin between mutational modes. Among all BrachyPan clusters, less than 5% of variant sequence pairs were from the same accession, indicating that variants caused by either isoforms or gene duplications within a single accession are rare. The frequency of NPTPs correlated well with the taxonomic groups reported previously (Gordon et al., 2017) (Figure 10). When measured as an average number of NPTPs per accession in each group, the extremely-delayed flowering (EDF+) accessions had nearly twice the diversity as the Turkish (T+) and Spanish (S+) accessions. The geographic distribution of accession diversity (Figure 10C) clearly shows that the most diverse accessions are generally found in or near Turkey. Even the reference Bd21-3 genome, which was collected in Iraq, has a greater number of NPTPs than many of the Spanish accessions. These results indicate that NPTP evolution followed a similar pattern to the Arabidopsis1001 dataset, and further implies that novel plastid-targeted proteins evolve in response to environmental pressures followed by natural selection.

Within the Brachypodium pan-genome, clusters are categorized into “core” (56 genomes; 100%), “softcore” (53-55 genomes; 95-98%), “shell” (3-52 lines; 5-94%), and “cloud” (1-2 genomes; 2-5%) categories (Gordon et al., 2017). Over half (56.6%) of NPTPs occurred in core

clusters, 25.2% occurred in softcore clusters, 17.9% occurred in shell clusters, and cloud clusters accounted for only 2 NPTPs, or 0.2% of the total (Figure 11).

Previously published GO annotation information (Gordon et al., 2017) was converted into GO slim categories in BLAST2GO and compared to the Bd21-3 reference genome to find under- and over-represented terms (Table 6). Using $p\text{-Value} < 0.05$ as a significance threshold, significant underrepresentation of cytoplasmic and ribosomal terms was found for cellular component ontologies, while structural molecule activity, ribosomal components, and cyclic compound binding terms were underrepresented for molecular function ontologies. Overrepresented terms involved in biological processes included nitrogen metabolism, cell wall organization, and lipid metabolism, suggesting that secondary metabolic processes are significantly more likely to have differential targeting. Terms of the molecular function ontology included ion binding, kinase activity, transferase activity, and catalytic activity. Overrepresented cellular component ontology terms were somewhat scattered, reflecting the selection of differentially-targeted genes.

Interestingly, several terms associated with inclusion bodies (GO:0090083, GO:0090084, GO:0070841) were overrepresented. Inclusion bodies are associated with viruses, so the enrichment of these terms suggests that not all NPTPs may be endogenous proteins. Sequences from pathogens and endosymbionts are common contaminants of high-throughput sequencing data, and eukaryotic genomes often have dormant retroviral elements scattered throughout their genomes (e.g., Sabot and Schulman, 2006). Although these sequences may at first glance appear to be false positives, effector proteins from multiple pathogenic species have been observed to translocate to the chloroplast (Dodds and Rathjen, 2010; Win et al., 2012). For instance, bioinformatic analysis of *Pseudomonas syringae* effector proteins predicts many to be

chloroplast-targeted (Guttman et al., 2002), and at least four have been confirmed *in vivo* (Jelenska et al., 2007; Li et al., 2014; Rodríguez-Herva et al., 2012). Plastid-targeted effectors also appear to be highly abundant in rust fungi (reviewed in Lorrain et al., 2018). Effector proteins from both bacteria and fungi suppress hypersensitive responses by targeting protein folding, salicylic and jasmonic acid production, photosystem II, and ROS signaling pathways. Furthermore, coat proteins of cucumber necrosis virus and *Lolium lentivirus* have also been described to have plastid localization, which may promote virus coat disassembly as well as target host immune pathways (Hui et al., 2010; Vaira et al., 2018).

Transposon-based origin of NTPs

Transposable element sequences for all Viridiplantae species were downloaded from GIRI REPBASE (<https://www.girinst.org/repbase/>) release 23.03. All possible open reading frames of at least 300 bp were mined from this dataset, translated to protein sequences, and analyzed with TargetP and Localizer. A total of 19,848 sequences with a consensus plastid targeting prediction were extracted and collected into a BLAST database for analysis against potential evolutionarily emergent chloroplast transit peptides. Each pair of diverged sequences was compared to this database of transposon sequences to see if the same transposon sequence was a match in both or unique to one sequence. Using an E-value cutoff of e^{-4} transposons were not found to be a significant source of transit peptide acquisition. No examples were found in the Arabidopsis1001 dataset, although this is unsurprising given that almost all pairs differed by only 1-2 substitutions. In BrachyPan, 33 potential candidates were identified, while in the multi-genome dataset, a total of 12 candidates were found. However, only a small fraction of these candidates were high-scoring matches covering a majority of the transit peptide, so many initial hits were the result of random sequence alignment. Although transposons may donate functional

transit peptides in a minority of cases, the evidence suggests that they are insignificant in the evolution of the plastid proteome.

Gaps in Orthologous Protein Prediction

It is likely that the total number of NPTPs has been underestimated in all datasets because relatively stringent criteria were implemented in this study. In the Arabidopsis1001 dataset, sequences seldom differ by more than a few residues throughout the whole gene sequence, and as a result, the phylogenetic trees have extremely short branch lengths. Many poorly-resolved trees were likely discarded due to this problem alone. In contrast, the BrachyPan and multi-genome gene clusters represent broader orthologous or homologous gene families and are far more likely to have relatively divergent branches with nascent plastid transit peptides. Yet, the inclusion of broader sequence variants also introduces the potential for error. In BrachyPan, many in-paralogs had poor sequence alignment and are likely to be unrelated, while in the multi-genome analysis, most of the smaller clusters were orthologous, but larger clusters often included paralogs. However, these trees were resolved with maximum likelihood methods, poorly aligned or nonhomologous sequences are unlikely to affect the analysis of these clusters. Even so, it is many larger clusters in which plastid targeting arose independently in multiple paralogs within that cluster were probably rejected. In these cases, resolution of independently evolving transit peptides would require more stringent clustering methods. Finally, the prediction approach using TargetP and Localizer achieves excellent correlation with experimentally validated results but has a significant sensitivity gap which may underrepresent NPTPs and lead to inaccurate prediction of the point of targeting divergence. It warrants the in-depth spatiotemporal study of plastid proteomes via transcriptomics or more aptly by using high throughput proteomics methods first to better understand the variation of the plastid proteome in

specific tissues or conditions as well as across the plant kingdom, and second to validate the localization predictions made in this study.

Conclusions

The plastid proteome is in a state of constant flux across evolutionary space in the plant kingdom due to gains and losses of plastid transit peptides in nuclear-encoded genes. Anecdotal reports which influence taxon-specific biochemistry, herbicide tolerance, and photosynthesis are known, but this study describes that transit peptide variation is widespread and likely responsible for significant phenotypic changes. Based on gene ontology enrichment, such localization variants are related to secondary metabolism, transcriptional regulation, and protein regulation. A better understanding of these unique plastid-targeted proteins could lead to improvements in yield, nutrient content, environmental stress tolerance, and production of valuable medicinal or aromatic compounds. The hypothesis that transit peptides evolve in predictable patterns was tested both at broad evolutionary scales and at the single-species level. Overall, it was found that in both the *Arabidopsis* and *Brachypodium* pan-genomes, loss of transit peptides occurs roughly four times more frequently than the gain of a novel transit peptide within orthologous protein clusters. However, gain events still occur regularly both in pan-genomes and across a wider phylogenetic landscape. Surprisingly, the primary sequence composition of monocot and eudicot transit peptides is divergent, possibly due to differences in the composition of the TIC translocon at the inner plastid envelope. Residue substitutions were less important to novel transit peptide evolution than were small insertions and deletions. A majority of these length variants represent probable alternative start sites, but internal insertions or deletions suggest that indels and alternative splicing are also major factors equaling or surpassing residue substitutions in

importance. Finally, it was found that gene duplications and alternative protein isoforms are more important factors in the evolution of novel plastid-targeted proteins than allelic variants. These results validate the hypothesis that transit peptides do evolve in predictable patterns of insertion, deletion, or use of alternative start sites, rather than from random substitution events. Less than 10% of relocalization examples involved alternative transcripts or translation products; deep transcriptome sequencing would provide much-needed evidence of the veracity and relative abundance of alternative transcripts. An additional 33% of cases involved gene duplication within the same genome. Both categories of transit peptide evolution are excellent candidates to examine very recent evolutionary changes which have resulted in the gain of a transit peptide. Moving forward, GFP fusion or *in vitro* translocation studies of both alternative transcript variants and gene duplication variants will be crucial to test the biological accuracy of predictions and to validate the predicted shift in subcellular localization. Although the focus was on chloroplast transit peptides within otherwise non-targeted gene families, functional divergence could additionally occur within conserved or semi-conserved plastid-targeted genes by mutation of the functional protein domains. It should be cautioned that without examining the genomic context of each potential localization variant, it is impossible to determine the exact evolutionary mechanism, such as whether length variants represent independent exons, shifts in exon/intron boundaries, or intraexonic indels. The outcome of these analyses, however, has built a foundational collection of candidate proteins to begin exploring localization and functional roles *in vivo*.

Materials and Methods

Clustering gene families and subcellular prediction

For the Arabidopsis1001 dataset, protein sequence files from 246 accessions cataloged in the Arabidopsis1001 proteomes project (Joshi et al., 2012) were downloaded from <http://1001data.masc-proteomics.org/index.html>. Sequences were sorted into single gene files using the reference Columbia-0 gene ID. In the BrachyPan dataset, protein sequences files were downloaded from <https://brachypan.jgi.doe.gov/> and grouped into orthologous protein clusters using resources by Dr. John Vogel (Gordon et al., 2017). For the multi-genome dataset, predicted proteomes from *Amborella trichopoda*, *Arabidopsis thaliana*, *Brachypodium distachyon*, *Fragaria vesca*, *Glycine max*, *Malus × domestica*, *Oryza sativa*, *Panicum virgatum*, *Populus trichoarpa*, *Prunus persica*, *Setaria italica*, *Solanum lycopersicum*, and *Sorghum bicolor* were downloaded from Phytozome (<https://phytozome.jgi.doe.gov>). Sequences for *Anthurium amnicola* were retrieved from Suzuki et al., 2017, and sequences for *Vitis vinifera* were retrieved from Vitulo et al., 2014. For *Malus × domestica*, a supplementary transcriptomics-based predicted proteome was created using the SRA datasets from (Bai et al., 2014; Gusberti et al., 2013; Krost et al., 2013, 2012; Petersen et al., 2015) and assembled using CLC Genomics Workbench v.8 (Qiagen Bioinformatics, Hilden, Germany). For all species, poorly annotated sequences were removed if no BLAST hits above 40% identity and 40% coverage were found for other sequences within the same dataset. Clustering of homologous proteins was performed using two parallel methods. First, reciprocal-best-BLAST hits (RBH) were generated by performing ALL-v-ALL BLASTP comparisons of the predicted proteome of each species against those of every other species. Sequences for each genome pair which were the mutual best hits above 40% identity and 40% coverage were kept as initial cluster connections. Initial

clusters were expanded using reciprocal better-BLAST hits within each genome, performed using ALL-v-ALL BLASTP comparisons of the predicted proteome of each species against itself. Sequences which had mutual hits above 90% identity and 90% coverage were kept. All cluster edges were collapsed to form clusters. In the second method, the UCLUST algorithm (Edgar, 2010) was executed on a concatenated, length-sorted sequence file of the predicted proteomes of all fifteen species using a 40% identity and 40% coverage threshold. From these initial clusters, random sequences from within each cluster, sorting these sequences by length, concatenating them to the beginning of the length-sorted initial sequence file, and re-running using a 90% identity threshold, minimum of 40% coverage, and maximum target length of 2.5x the query length. This randomized seed method was iterated 100 times, and all clusters from the initial run and subsequent iterations were condensed if they shared at least one sequence. From both RBH and UCLUST methods, clusters were selected which contained at least three species and only a single species with a predicted plastid-targeted sequence. Protein sequences from each dataset were analyzed with TargetP v.1.1 (Emanuelsson et al., 2007, 2000) and Localizer v.1.0.2 (Sperschneider et al., 2017). All sequences predicted by both methods to have chloroplast localization were determined to be plastid-localized, and sequences predicted by one or neither method were determined to be non-plastid-targeted.

Transit Peptide Analysis

Experimentally-validated proteomics data were retrieved from PPDB (Sun et al., 2009), AT_CHLORO (Ferro et al., 2010), SUBA4 (Hooper et al., 2017), CropPAL, and CropPAL2 (Hooper et al., 2015). Non-redundant sequences validated by mass spectrometry and with unambiguous localization were extracted and residue composition and positional frequency within the 60 residues comprising the transit peptide and all downstream residues comprising the

mature protein were analyzed. For the analysis of *in silico* prediction methods, all sequences with a predicted plastid transit peptide from each species used in the multi-genome dataset were similarly analyzed for residue composition and positional frequency.

Phylogenetics

All clusters derived for Arabidopsis1001, BrachyPan, and the multi-genome datasets were trimmed of poorly-aligned sequences before phylogenetic analysis using a BLAST filter to remove any sequences with less than 40% identity and 40% coverage to any of the predicted plastid-targeted sequences. Maximum likelihood phylogenetic trees were constructed for each cluster using MAFFT v. 7.407 (Katoh et al., 2002; Katoh and Standley, 2013), and trimmed with Phyutility v2.2.6 (Smith and Dunn, 2008). Pasta v.1.0 (Mirarab et al., 2015) and FastTree 2.1.10 (Price et al., 2010) were used for alignments of trimmed files, and phylogenetic trees were constructed using RAxML v. 8.2.31 (Stamatakis, 2014, 2006). The point of divergence between plastid-targeted and non-plastid-targeted sequences for each gene cluster was performed using a custom Perl script (Additional file 1) and determined by examination of branch lengths in Newick-formatted maximum likelihood trees. Clusters in which predicted plastid-targeted sequences arose more than once were discarded as probable examples of the polyphyletic origin or ambiguous targeting sequence.

Additionally, clusters in which a transit peptide was gained and then lost or in which a plastid-targeted sequence rooted the tree were discarded. Second sequence alignment was performed on the divergent pairs of sequences using MUSCLE v.3.8.31. Mutations within the transit peptide region were defined as the part of the alignment corresponding to the first 60 residues of the plastid-predicted sequence, and the mature region was considered to be the remainder of the alignment. Frequency of substitutions and residue classes was collected for all

sequences. Alignments which started with a gap but in which the first aligned residue was a methionine in both sequences were classified as alternative start sites. Alignments with insertion of at least ten residues followed by a deletion of at least ten residues, or vice versa, were categorized as alternative first exons. Finally, alignments in which more substitutions occurred than gaps were classified as substitution-dominant, while alignments fulfilling the reverse criteria were classified as insertion or deletion-dominant.

Gene Ontology Analysis

Annotations for NPTPs were retrieved from Phytozome (<https://phytozome.jgi.doe.gov>) for each of the species used in the analysis except *Anthurium amnicola* and *Vitis vinifera*, which were retrieved from (Suzuki et al., 2017) and (Vitulo et al., 2014), respectively. Annotations for BrachyPan were retrieved from the supplementary materials of Gordon et al. (2017). Non-redundant predicted proteins produced by the *de novo* transcriptome assembly of *Malus × domestica* as described in Chapter 3 were annotated using BLASTP against the NR Protein database at NCBI with BLAST2GO default parameters (Conesa et al., 2005; Conesa and Götz, 2008) (BioBam Bioinformatics, Valencia, Spain). GOslim annotations were retrieved using BLAST2GO. Over- and under-represented GO terms for each dataset was performed using BLAST2GO. Fisher's Exact Test was used to calculate significance, and all terms below a false discovery rate (FDR) significance threshold of 0.05 or p-value threshold of 0.05 were extracted.

Authors' contributions

RC and AD designed the study. RC performed localization prediction, gene clustering, and data analysis. ER provided phylogenetics expertise. GN wrote scripts for automating phylogenetics workflows and performed phylogenetic comparisons. AD and ER supervised the study. SH performed gene annotation analyses. RC and AD prepared the manuscript. All authors read and approved the manuscript. The authors declare no conflict of interest.

Acknowledgments

Work in the Dhingra lab was supported by Washington State University Agriculture Center Research Hatch Grant WNP00011 to AD. RC and SLH acknowledge the support received from the National Institutes of Health/National Institute of General Medical Sciences through an institutional training grant award T32-GM008336. The contents of this work are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS or NIH. SLH acknowledges the support received from ARCS Seattle Chapter.

References

- Adams, K.L., Rosenblueth, M., Qiu, Y.L., Palmer, J.D., 2001. Multiple losses and transfers to the nucleus of two mitochondrial succinate dehydrogenase genes during angiosperm evolution. *Genetics* 158, 1289–1300.
- Albert, V.A., Barbazuk, W.B., Der, J.P., Leebens-Mack, J., Ma, H., Palmer, J.D., Rounsley, S., Sankoff, D., Schuster, S.C., Soltis, D.E., 2013. The Amborella Genome and the Evolution of Flowering Plants. *Science* (80-.). 342, 1241089. <https://doi.org/10.1126/science.1241089>
- Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K.M.M., Cao, J., Chae, E., Dezwaan, T.M.M., Ding, W., Ecker, J.R.R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D.G.G., Hancock, A.M.M., Henz, S.R.R., Holm, S., Horton, M., Jarsulic, M., Kerstetter, R.A.A., Korte, A., Korte, P., Lanz, C., Lee, C.R., Meng, D., Michael, T.P.P., Mott, R., Mulyati, N.W.W., Nägele, T., Nagler, M., Nizhynska, V., Nordborg, M., Novikova, P.Y.Y., Picó, F.X., Platzer, A., Rabanal, F.A.A., Rodriguez, A., Rowan, B.A.A., Salomé, P.A.A., Schmid, K.J.J., Schmitz, R.J.J., Seren, Ü., Sperone, F.G.G., Sudkamp, M., Svardal, H., Tanzer, M.M.M., Todd, D., Volchenboum, S.L.L., Wang, C., Wang, G., Wang, X., Weckwerth, W., Weigel, D., Zhou, X., 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. *Cell* 166, 481–491. <https://doi.org/10.1016/j.cell.2016.05.063>
- Arimura, S.I., Takusagawa, S., Hatano, S., Nakazono, M., Hirai, A., Tsutsumi, N., 1999. A novel plant nuclear gene encoding chloroplast ribosomal protein S9 has a transit peptide related to that of rice chloroplast ribosomal protein L12. *FEBS Lett.* 450, 231–234. [https://doi.org/10.1016/S0014-5793\(99\)00491-3](https://doi.org/10.1016/S0014-5793(99)00491-3)
- Armbruster, U., Pesaresi, P., Pribil, M., Hertle, A., Leister, D., 2011. Update on chloroplast

- research: New tools, new topics, and new trends. *Mol. Plant* 4, 1–16.
<https://doi.org/10.1093/mp/ssq060>
- Bai, Y., Dougherty, L., Xu, K., 2014. Towards an improved apple reference transcriptome using RNA-seq. *Mol. Genet. Genomics* 289, 427–438. <https://doi.org/10.1007/s00438-014-0819-3>
- Baker, A., Schatz, G., 1987. Sequences from a prokaryotic genome or the mouse dihydrofolate reductase gene can restore the import of a truncated precursor protein into yeast mitochondria. *Proc. Natl. Acad. Sci.* 84, 3117–3121.
<https://doi.org/10.1073/pnas.84.10.3117>
- Bienvenut, W. V., Sumpton, D., Martinez, A., Lilla, S., Espagne, C., Meinel, T., Giglione, C., 2012. Comparative Large Scale Characterization of Plant *versus* Mammal Proteins Reveals Similar and Idiosyncratic *N*- α -Acetylation Features. *Mol. Cell. Proteomics* 11, M111.015131. <https://doi.org/10.1074/mcp.M111.015131>
- Bionda, T., Tillmann, B., Simm, S., Beilstein, K., Ruprecht, M., Schleiff, E., 2010. Chloroplast import signals: The length requirement for translocation in vitro and in vivo. *J. Mol. Biol.* 402, 510–523. <https://doi.org/10.1016/j.jmb.2010.07.052>
- Bölter, B., Soll, J., 2017. Ycf1/Tic214 Is Not Essential for the Accumulation of Plastid Proteins. *Mol. Plant* 10, 219–221. <https://doi.org/10.1016/j.molp.2016.10.012>
- Brillouet, J.M., Verdeil, J.L., Odoux, E., Lartaud, M., Grisoni, M., Conéjéro, G., 2014. Phenol homeostasis is ensured in vanilla fruit by storage under solid form in a new chloroplast-derived organelle, the phenyloplast. *J. Exp. Bot.* 65, 2427–2435.
<https://doi.org/10.1093/jxb/eru126>
- Bruce, B.D., 2001. The paradox of plastid transit peptides: Conservation of function despite divergence in primary structure. *Biochim. Biophys. Acta - Mol. Cell Res.* 1541, 2–21.

[https://doi.org/10.1016/S0167-4889\(01\)00149-5](https://doi.org/10.1016/S0167-4889(01)00149-5)

Bruce, B.D., 2000. Chloroplast transit peptides: Structure, function and evolution. *Trends Cell Biol.* 10, 440–447. [https://doi.org/10.1016/S0962-8924\(00\)01833-X](https://doi.org/10.1016/S0962-8924(00)01833-X)

Byun-McKay, S.A., Geeta, R., 2007. Protein subcellular relocalization: a new perspective on the origin of novel genes. *Trends Ecol. Evol.* 22, 338–344.
<https://doi.org/10.1016/j.tree.2007.05.002>

Byun, S.A., Singh, S., 2013. Protein subcellular relocalization increases the retention of eukaryotic duplicate genes. *Genome Biol. Evol.* 5, 2402–2409.
<https://doi.org/10.1093/gbe/evt183>

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K.J., Weigel, D., 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* 43, 956–963. <https://doi.org/10.1038/ng.911>

Carde, J.P., 1984. Leucoplasts: a distinct kind of organelles lacking typical 70S ribosomes and free thylakoids. *Eur. J. Cell Biol.* 34, 18–26.

Chen, Z.J., 2007. Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. *Annu. Rev. Plant Biol.* 58, 377–406.
<https://doi.org/10.1146/annurev.arplant.58.032806.103835>

Cheniclet, C., Carde, J.P., 1985. Presence of leucoplasts in secretory cells and of monoterpenes in the essential oil: A correlative study. *Isr. J. Bot.* 34, 219–238.
<https://doi.org/10.1080/0021213X.1985.10677023>

Chotewutmontri, P., Bruce, B.D., 2015. Non-native, N-terminal Hsp70 Molecular Motor-recognition Elements in Transit Peptides Support Plastid Protein Translocation. *J. Biol.*

- Chem. 290, 7602–7621. <https://doi.org/10.1074/jbc.M114.633586>
- Chotewutmontri, P., Reddick, L.E., McWilliams, D.R., Campbell, I.M., Bruce, B.D., 2012. Differential Transit Peptide Recognition during Preprotein Binding and Translocation into Flowering Plant Plastids. *Plant Cell* 24, 3040–3059. <https://doi.org/10.1105/tpc.112.098327>
- Christian, R., Hewitt, S., Roalson, E., Dhingra, A., 2019. Genome-Scale Characterization of Predicted Plastid-Targeted Proteins in Higher Plants.
- Clark, S.M., Vaitheeswaran, V., Ambrose, S.J., Purves, R.W., Page, J.E., 2013. Transcriptome analysis of bitter acid biosynthesis and precursor pathways in hop (*Humulus lupulus*). *BMC Plant Biol.* 13. <https://doi.org/10.1186/1471-2229-13-12>
- Clayton, H., Saladié, M., Rolland, V., Sharwood, R., Macfarlane, T., Ludwig, M., 2017. Loss of the Chloroplast Transit Peptide from an Ancestral C₃ Carbonic Anhydrase Is Associated with C₄ Evolution in the Grass Genus *Neurachne*. *Plant Physiol.* 173, 1648–1658. <https://doi.org/10.1104/pp.16.01893>
- Comai, L., Larson-Kelly, N., Kiser, J., Mau, C.J.D., Pokalsky, A.R., Shewmaker, C.K., McBride, K., Jones, A., Stalker, D.M., 1988. Chloroplast Transport of a Ribulose Bisphosphate Carboxylase Small Subunit-5-Enolpyruvyl 3-Phosphoshikimate Synthase Chimeric Protein Requires Part of the Mature Small Subunit in Addition to the Transit Peptide. *J. Biol. Chem.* 263, 15104–15109.
- Conesa, A., Götz, S., 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics* 2008. <https://doi.org/10.1155/2008/619832>
- Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., Robles, M., 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. <https://doi.org/10.1093/bioinformatics/bti610>

- Davis, M.J., Hanson, K.A., Clark, F., Fink, J.L., Zhang, F., Kasukawa, T., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Teasdale, R.D., 2006. Differential use of signal peptides and membrane domains is a common occurrence in the protein output of transcriptional units. *PLoS Genet.* 2, 554–563. <https://doi.org/10.1371/journal.pgen.0020046>
- De Luca, V., Cutler, A.J., 1987. Subcellular Localization of Enzymes Involved in Indole Alkaloid Biosynthesis in *Catharanthus*. *Plant Physiol.* 85, 1099–1102.
- de Vries, J., Sousa, F.L., Bölter, B., Soll, J., Gould, S.B., 2015. YCF1: A Green TIC? *Plant Cell* 27, 1827–1833. <https://doi.org/10.1105/tpc.114.135541>
- Demurtas, O.C., Frusciante, S., Ferrante, P., Diretto, G., Azad, N.H., Pietrella, M., Aprea, G., Taddei, A.R., Romano, E., Mi, J., Al-Babili, S., Frigerio, L., Giuliano, G., 2018. Candidate Enzymes for Saffron Crocin Biosynthesis Are Localized in Multiple Cellular Compartments. *Plant Physiol.* 177, 990–1006. <https://doi.org/10.1104/pp.17.01815>
- Dixon, J., Hewett, E.W., 2000. Factors affecting apple aroma/flavour volatile concentration: A review. *New Zeal. J. Crop Hortic. Sci.* 28, 155–173. <https://doi.org/10.1080/01140671.2000.9514136>
- Dodds, P.N., Rathjen, J.P., 2010. Plant immunity: towards an integrated view of plant–pathogen interactions. *Nat. Rev. Genet.* 11, 539–548. <https://doi.org/10.1038/nrg2812>
- Duke, S.O., Canel, C., Rimando, A.M., Telle, M.R., Duke, M. V, Paul, R.N., 2000. Current and potential exploitation of plant glandular trichome productivity.
- Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. <https://doi.org/10.1093/bioinformatics/btq461>
- El Hadi, M.A.M., Zhang, F.J., Wu, F.F., Zhou, C.H., Tao, J., 2013. Advances in fruit aroma volatile research. *Molecules* 18, 8200–8229. <https://doi.org/10.3390/molecules18078200>

- Elo, A., Lyznik, A., Gonzalez, D.O., Kachman, S.D., Mackenzie, S.A., 2003. Nuclear Genes That Encode Mitochondrial Proteins for DNA and RNA Metabolism Are Clustered in the Arabidopsis Genome. *Plant Cell* 15, 1619–1631. <https://doi.org/10.1105/tpc.010009>
- Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H., 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* 2, 953–71. <https://doi.org/10.1038/nprot.2007.131>
- Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G., 2000. Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. *J. Mol. Biol.* 300, 1005–1016. <https://doi.org/10.1006/jmbi.2000.3903>
- Ferro, M., Brugière, S., Salvi, D., Seigneurin-Berny, D., Court, M., Moyet, L., Ramus, C., Miras, S., Mellal, M., Le Gall, S., Kieffer-Jaquinod, S., Bruley, C., Garin, J., Joyard, J., Masselon, C., Rolland, N., 2010. AT_CHLORO, a Comprehensive Chloroplast Proteome Database with Subplastidial Localization and Curated Information on Envelope Proteins. *Mol. Cell. Proteomics* 9, 1063–1084. <https://doi.org/10.1074/mcp.M900325-MCP200>
- Flagel, L.E., Wendel, J.E., 2009. Gene duplication and evolutionary novelty in plants. *New Phytol.* 183, 557–564.
- Gantt, J.S., Baldauf, S.L., Calie, P.J., Weeden, N.F., Palmer, J.D., 1991. Transfer of rpl22 to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. *EMBO J.* 10, 3073–3078.
- Gordon, S.P., Contreras-Moreira, B., Woods, D.P., Des Marais, D.L., Burgess, D., Shu, S., Stritt, C., Roulin, A.C., Schackwitz, W., Tyler, L., Martin, J., Lipzen, A., Dochy, N., Phillips, J., Barry, K., Geuten, K., Budak, H., Juenger, T.E., Amasino, R., Caicedo, A.L., Goodstein, D., Davidson, P., Mur, L.A.J., Figueroa, M., Freeling, M., Catalan, P., Vogel, J.P., 2017.

- Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.* 8. <https://doi.org/10.1038/s41467-017-02292-8>
- Guo, L., Liu, C.M., 2015. A single-nucleotide exon found in *Arabidopsis*. *Sci. Rep.* 5, 1–5. <https://doi.org/10.1038/srep18087>
- Gusberti, M., Gessler, C., Broggin, G.A.L., 2013. RNA-seq analysis reveals candidate genes for ontogenic resistance in *Malus-Venturia* pathosystem. *PLoS One* 8. <https://doi.org/10.1371/journal.pone.0078457>
- Guttman, D.S., Vinatzer, B.A., Sarkar, S.F., Ranall, M. V, Kettler, G., Greenberg, J.T., 2002. A Functional Screen for the Type III (Hrp) Secretome of the Plant Pathogen *Pseudomonas syringae* Linked references are available on JSTOR for this article : Type III (Hrp) Secretome of the Plant Pathogen *Pseudomonas syringae*. *Science* (80-.). 295, 1722–1726.
- Heilmann, I., Pidkowich, M.S., Girke, T., Shanklin, J., 2004. Switching desaturase enzyme specificity by alternate subcellular targeting. *Proc. Natl. Acad. Sci.* 101, 10266–10271. <https://doi.org/10.1073/pnas.0402200101>
- Herrmann, K.M., Weaver, L.M., 1999. the Shikimate Pathway. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 50, 473–503. <https://doi.org/10.1146/annurev.arplant.50.1.473>
- Hönigsmid, P., Bykova, N., Schneider, R., Ivankov, D., Frishman, D., 2018. Evolutionary Interplay between Symbiotic Relationships and Patterns of Signal Peptide Gain and Loss. *Genome Biol. Evol.* 10, 928–938. <https://doi.org/10.1093/gbe/evy049>
- Hooper, C.M., Castleden, I.R., Aryamanesh, N., Jacoby, R.P., Millar, A.H., 2015. Finding the subcellular location of barley, wheat, rice and maize proteins: The compendium of crop proteins with annotated locations (cropPAL). *Plant Cell Physiol.* 57, e9. <https://doi.org/10.1093/pcp/pcv170>

- Hooper, C.M., Castleden, I.R., Tanz, S.K., Aryamanesh, N., Millar, A.H., 2017. SUBA4: The interactive data analysis centre for Arabidopsis subcellular protein locations. *Nucleic Acids Res.* 45, D1064–D1074. <https://doi.org/10.1093/nar/gkw1041>
- Huang, S., Taylor, N.L., Whelan, J., Millar, A.H., 2009. Refining the Definition of Plant Mitochondrial Presequences through Analysis of Sorting Signals, N-Terminal Modifications, and Cleavage Motifs. *Plant Physiol.* 150, 1272–1285. <https://doi.org/10.1104/pp.109.137885>
- Hui, E., Xiang, Y., Rochon, D., 2010. Distinct regions at the N-terminus of the Cucumber necrosis virus coat protein target chloroplasts and mitochondria. *Virus Res.* 153, 8–19.
- Inoue, H., Rounds, C., Schnell, D.J., 2010. The Molecular Basis for Distinct Pathways for Protein Import into *Arabidopsis* Chloroplasts. *Plant Cell* 22, 1947–1960. <https://doi.org/10.1105/tpc.110.074328>
- Jelenska, J., Yao, N., Vinatzer, B.A., Wright, C.M., Brodsky, J.L., Greenberg, J.T., 2007. A J Domain Virulence Effector of *Pseudomonas syringae* Remodels Host Chloroplasts and Suppresses Defenses. *Curr. Biol.* 17, 499–508. <https://doi.org/10.1016/j.cub.2007.02.028>
- Jelic, M., Soll, J., Schleiff, E., 2003. Two Toc34 homologues with different properties. *Biochemistry* 42, 5906–5916. <https://doi.org/10.1021/bi034001q>
- Joshi, H.J., Christiansen, K.M., Fitz, J., Cao, J., Lipzen, A., Martin, J., Smith-Moritz, A.M., Pennacchio, L.A., Schackwitz, W.S., Weigel, D., Heazlewood, J.L., 2012. 1001 Proteomes: A functional proteomics portal for the Analysis of *arabidopsis thaliana* accessions. *Bioinformatics* 28, 1303–1306. <https://doi.org/10.1093/bioinformatics/bts133>
- Kadowaki, K., Kubo, N., Ozawa, K., Hirai, A., 1996. Targeting presequence acquisition after mitochondrial gene transfer to the nucleus occurs by duplication of existing targeting

signals. *EMBO J.* 15, 6652–6661. <https://doi.org/10.1002/j.1460-2075.1996.tb01055.x>

Kaiser, C.A., Preuss, D., Grisafi, P., Botstein, D., 1987. Many random sequences functionally replace the secretion signal sequence of yeast invertase. *Science* (80-.). 235, 312–317.

Karlin-Neumann, G.A., Tobin, E.M., 1986. Transit peptides of nuclear-encoded chloroplast proteins share a common amino acid framework. *EMBO J.* 5, 9–13.
<https://doi.org/10.1002/j.1460-2075.1986.tb04170.x>

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066.
<https://doi.org/10.1093/nar/gkf436>

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780.
<https://doi.org/10.1093/molbev/mst010>

Kitagawa, N., Washio, T., Kosugi, S., Yamashita, T., Higashi, K., Yanagawa, H., Higo, K., Satoh, K., Ohtomo, Y., Sunako, T., Murakami, K., Matsubara, K., Kawai, J., Carninci, P., Hayashizaki, Y., Kikuchi, S., Tomita, M., 2005. Computational analysis suggests that alternative first exons are involved in tissue-specific transcription in rice (*Oryza sativa*). *Bioinformatics* 21, 1758–1763. <https://doi.org/10.1093/bioinformatics/bti253>

Kleffmann, T., von Zychlinski, A., Russenberger, D., Hirsch-Hoffmann, M., Gehrig, P., Gruissem, W., Baginsky, S., 2007. Proteome Dynamics during Plastid Differentiation in Rice. *Plant Physiol.* 143, 912–923. <https://doi.org/10.1104/pp.106.090738>

Krost, C., Petersen, R., Lokan, S., Brauksiepe, B., Braun, P., Schmidt, E.R., 2013. Evaluation of the hormonal state of columnar apple trees (*Malus x domestica*) based on high throughput gene expression studies. *Plant Mol. Biol.* 81, 211–220. <https://doi.org/10.1007/s11103-012->

9992-0

- Krost, C., Petersen, R., Schmidt, E.R., 2012. The transcriptomes of columnar and standard type apple trees (*Malus x domestica*) - A comparative study. *Gene* 498, 223–230.
<https://doi.org/10.1016/j.gene.2012.01.078>
- Lee, D.W., Kim, J.K., Lee, S., Choi, S., Kim, S., Hwang, I., 2008. Arabidopsis Nuclear-Encoded Plastid Transit Peptides Contain Multiple Sequence Subgroups with Distinctive Chloroplast-Targeting Sequence Motifs. *Plant Cell* 20, 1603–1622.
<https://doi.org/10.1105/tpc.108.060541>
- Li, G., Froehlich, J.E., Elowsky, C., Msanne, J., Ostosh, A.C., Zhang, C., Awada, T., Alfano, J.R., 2014. Distinct *Pseudomonas* type-III effectors use a cleavable transit peptide to target chloroplasts. *Plant J.* 77, 310–321. <https://doi.org/10.1111/tpj.12396>
- Li, H. min, Teng, Y.S., 2013. Transit peptide design and plastid import regulation. *Trends Plant Sci.* 18, 360–366. <https://doi.org/10.1016/j.tplants.2013.04.003>
- Li, L., Yuan, H., 2013. Chromoplast biogenesis and carotenoid accumulation. *Arch. Biochem. Biophys.* 539, 102–109. <https://doi.org/10.1016/j.abb.2013.07.002>
- Li, Y.D., Xie, Z.Y., Du, Y.L., Zhou, Z., Mao, X.M., Lv, L.X., Li, Y.Q., 2009. The rapid evolution of signal peptides is mainly caused by relaxed selection on non-synonymous and synonymous sites. *Gene* 436, 8–11. <https://doi.org/10.1016/j.gene.2009.01.015>
- Long, M., de Souza, S.J., Rosenberg, C., Gilbert, W., 1996. Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor. *Proc. Natl. Acad. Sci.* 93, 7727–7731.
- Lorrain, C., Petre, B., Duplessis, S., 2018. Show me the way: rust effector targets in heterologous plant systems. *Curr. Opin. Microbiol.* 46, 19–25. <https://doi.org/10.1016/j.mib.2018.01.016>

- Mackenzie, S.A., 2005. Plant organellar protein targeting: A traffic plan still under construction. *Trends Cell Biol.* 15, 548–554. <https://doi.org/10.1016/j.tcb.2005.08.007>
- Mahlberg, P.G., Kim, E.S., 2004. Accumulation of cannabinoids in glandular trichomes of *Cannabis* (Cannabaceae). *J. Ind. Hemp* 9, 15–36.
- Markus Lange, B., Turner, G.W., 2013. Terpenoid biosynthesis in trichomes-current status and future opportunities. *Plant Biotechnol. J.* 11, 2–22. <https://doi.org/10.1111/j.1467-7652.2012.00737.x>
- McKay, S.A.B., Geeta, R., Duggan, R., Carroll, B., McKay, S.J., 2009. Missing the subcellular target: a mechanism of eukaryotic gene evolution, in: *Evolutionary Biology*. Springer, pp. 175–183.
- Millar, A.H., Whelan, J., Small, I., 2006. Recent surprises in protein targeting to mitochondria and plastids. *Curr. Opin. Plant Biol.* 9, 610–615. <https://doi.org/10.1016/j.pbi.2006.09.002>
- Mirarab, S., Nguyen, N., Guo, S., Wang, L.-S., Kim, J., Warnow, T., 2015. PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences. *J. Comput. Biol.* 22, 377–386. <https://doi.org/10.1089/cmb.2014.0156>
- Nakai, M., 2015. YCF1: A Green TIC: Response to the de Vries et al. Commentary. *Plant Cell* 27, 1834–1838. <https://doi.org/10.1105/tpc.15.00363>
- Nei, M., 2007. The new mutation theory of phenotypic evolution. *Proc. Natl. Acad. Sci.* 104, 12235–12242. <https://doi.org/10.1073/pnas.0703349104>
- Nielsen, H., Wernersson, R., 2006. An overabundance of phase 0 introns immediately after the start codon in eukaryotic genes. *BMC Genomics* 7, 1–15. <https://doi.org/10.1186/1471-2164-7-256>
- Nugent, J.M., Palmer, J.D., 1991. RNA-mediated transfer of the gene *coxII* from the

- mitochondrion to the nucleus during flowering plant evolution. *Cell* 66, 473–481.
[https://doi.org/10.1016/0092-8674\(81\)90011-8](https://doi.org/10.1016/0092-8674(81)90011-8)
- Patzoldt, W.L., Hager, A.G., McCormick, J.S., Tranel, P.J., 2006. A codon deletion confers resistance to herbicides inhibiting protoporphyrinogen oxidase. *Proc. Natl. Acad. Sci.* 103, 12329–12334. <https://doi.org/10.1073/pnas.0603137103>
- Peeters, N., Small, I., 2001. Dual targeting to mitochondria and chloroplasts. *Biochim. Biophys. Acta - Mol. Cell Res.* 1541, 54–63. [https://doi.org/10.1016/S0167-4889\(01\)00146-X](https://doi.org/10.1016/S0167-4889(01)00146-X)
- Petersen, R., Djozgic, H., Rieger, B., Rapp, S., Schmidt, E.R., 2015. Columnar apple primary roots share some features of the columnar-specific gene expression profile of aerial plant parts as evidenced by RNA-Seq analysis. *BMC Plant Biol.* 15, 1–16.
<https://doi.org/10.1186/s12870-014-0356-6>
- Pilon, M., Wienk, H., Sips, W., De Swaaf, M., Talboom, I., Van't Hof, R., De Korte- Kool, G., Demel, R., Weisbeek, P., De Kruijff, B., 1995. Functional domains of the ferredoxin transit sequence involved in chloroplast import. *J. Biol. Chem.*
<https://doi.org/10.1074/jbc.270.8.3882>
- Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One* 5. <https://doi.org/10.1371/journal.pone.0009490>
- Pujol, C., Maréchal-Drouard, L., Duchêne, A.M., 2007. How Can Organellar Protein N-terminal Sequences Be Dual Targeting Signals? In silico Analysis and Mutagenesis Approach. *J. Mol. Biol.* 369, 356–367. <https://doi.org/10.1016/j.jmb.2007.03.015>
- Quigley, F., Martin, W.F., Cerff, R., 1988. Intron conservation across the prokaryote-eukaryote boundary: structure of the nuclear gene for chloroplast glyceraldehyde-3-phosphate dehydrogenase from maize. *Proc. Natl. Acad. Sci.* 85, 2672–2676.

<https://doi.org/10.1073/pnas.85.8.2672>

Rensink, W.A., Pilon, M., Weisbeek, P., 1998. Domains of a transit sequence required for in vivo import in Arabidopsis chloroplasts. *Plant Physiol.* 118, 691–9.

<https://doi.org/10.1104/pp.118.2.691>

Rensink, W.A., Schnell, D.J., Weisbeek, P.J., 2000. The transit sequence of ferredoxin contains different domains for translocation across the outer and inner membrane of the chloroplast envelope. *J. Biol. Chem.* 275, 10265–10271. <https://doi.org/10.1074/jbc.275.14.10265>

Richardson, L.G., Jelokhani-Niaraki, M., Smith, M.D., 2009. The acidic domains of the Toc159 chloroplast preprotein receptor family are intrinsically disordered protein domains. *BMC Biochem.* 10, 10–12. <https://doi.org/10.1186/1471-2091-10-35>

Richardson, L.G., Small, E.L., Inoue, H., Schnell, D.J., 2018. Molecular topology of the transit peptide during chloroplast protein import. *Plant Cell tpc.00172.2018.*

<https://doi.org/10.1105/tpc.18.00172>

Richly, E., Leister, D., 2004. An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of Arabidopsis and rice. *Gene* 329, 11–16.

<https://doi.org/10.1016/j.gene.2004.01.008>

Rodríguez-Concepcion, M., Boronat, A., 2002. Elucidation of the Methylerythritol Phosphate Pathway for Isoprenoid Biosynthesis in Bacteria and Plastids. A Metabolic Milestone Achieved through Genomics. *Plant Physiol.* 130, 1079–1089.

<https://doi.org/10.1104/pp.007138.ISOPRENOID>

Rodríguez-Herva, J.J., González-Melendi, P., Cuartas-Lanza, R., Antúnez-Lamas, M., Río-Alvarez, I., Li, Z., López-Torrejón, G., Díaz, I., Del Pozo, J.C., Chakravarthy, S., Collmer, A., Rodríguez-Palenzuela, P., López-Solanilla, E., 2012. A bacterial cysteine protease

- effector protein interferes with photosynthesis to suppress plant innate immune responses. *Cell. Microbiol.* 14, 669–681. <https://doi.org/10.1111/j.1462-5822.2012.01749.x>
- Roesler, K., Shintani, D., Savage, L., Boddupalli, S., Ohlrogge, J., 1997. Targeting of the Arabidopsis homomeric acetyl-coenzyme A carboxylase to plastids of rapeseeds. *Plant Physiol.* 113, 75–81. <https://doi.org/10.1104/pp.113.1.75>
- Rolland, V., Badger, M.R., Price, G.D., 2016. Redirecting the Cyanobacterial Bicarbonate Transporters BicA and SbtA to the Chloroplast Envelope: Soluble and Membrane Cargos Need Different Chloroplast Targeting Signals in Plants. *Front. Plant Sci.* 7, 1–19. <https://doi.org/10.3389/fpls.2016.00185>
- Sabot, F., Schulman, A.H., 2006. Parasitism and the retrotransposon life cycle in plants: A hitchhiker’s guide to the genome. *Heredity (Edinb).* 97, 381–388. <https://doi.org/10.1038/sj.hdy.6800903>
- Sandoval, P., León, G., Gómez, I., Carmona, R., Figueroa, P., Holuigue, L., Araya, A., Jordana, X., 2004. Transfer of RPS14 and RPL5 from the mitochondrion to the nucleus in grasses. *Gene* 324, 139–147.
- Schaeffer, S., Harper, A., Raja, R., Jaiswal, P., Dhingra, A., 2014. Comparative analysis of predicted plastid-targeted proteomes of sequenced higher plant genomes. *PLoS One* 9, e112870. <https://doi.org/10.1371/journal.pone.0112870>
- Schaeffer, S.M., Christian, R., Castro-Velasquez, N., Hyden, B., Lynch-Holm, V., Dhingra, A., 2017. Comparative ultrastructure of fruit plastids in three genetically diverse genotypes of apple (*Malus × domestica* Borkh.) during development. *Plant Cell Rep.* 36, 1627–1640. <https://doi.org/10.1007/s00299-017-2179-z>
- Schlute, W., Töpfer, R., Stracke, R., Schell, J., Martini, N., 1997. Multi-functional acetyl-CoA

- carboxylase from *Brassica napus* is encoded by a multi-gene family: Indication for plastidic localization of at least one isoform. *Proc. Natl. Acad. Sci.* 94, 3465–3470.
- Shen, B.R., Zhu, C.H., Yao, Z., Cui, L.L., Zhang, J.J., Yang, C.W., He, Z.H., Peng, X.X., 2017. An optimized transit peptide for effective targeting of diverse foreign proteins into chloroplasts in rice. *Sci. Rep.* 7, 1–12. <https://doi.org/10.1038/srep46231>
- Small, I., Wintz, H., Akashi, K., Mireau, H., 1998. Two birds with one stone: genes that encode products targeted to two or more compartments. *Plant Mol. Biol.* 38, 265–277. <https://doi.org/10.1023/A:1006081903354>
- Smith, M.D., Rounds, C.M., Wang, F., Chen, K., Afithile, M., Schnell, D.J., 2004. atToc159 is a selective transit peptide receptor for the import of nucleus-encoded chloroplast proteins. *J. Cell Biol.* 165, 323–334. <https://doi.org/10.1083/jcb.200311074>
- Smith, S.A., Dunn, C.W., 2008. Phyutility: A phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* 24, 715–716. <https://doi.org/10.1093/bioinformatics/btm619>
- Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., DePamphilis, C.W., Wall, P.K., Soltis, P.S., 2009. Polyploidy and angiosperm diversification. *Am. J. Bot.* 96, 336–348. <https://doi.org/10.3732/ajb.0800079>
- Solyosi, K., Keresztes, A., 2013. Plastid Structure, Diversification and Interconversions II. *Land Plants. Curr. Chem. Biol.* 6, 187–204. <https://doi.org/10.2174/2212796811206030003>
- Song, J., Bangerth, F., 2003. Fatty acids as precursors for aroma volatile biosynthesis in pre-climacteric and climacteric apple fruit. *Postharvest Biol. Technol.* 30, 113–121. [https://doi.org/10.1016/S0925-5214\(03\)00098-X](https://doi.org/10.1016/S0925-5214(03)00098-X)
- Sperschneider, J., Catanzariti, A.-M., DeBoer, K., Petre, B., Gardiner, D.M., Singh, K.B., Dodds, P.N., Taylor, J.M., 2017. LOCALIZER: subcellular localization prediction of both plant and

- effector proteins in the plant cell. *Sci. Rep.* 7, 44598. <https://doi.org/10.1038/srep44598>
- Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>
- Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690. <https://doi.org/10.1093/bioinformatics/btl446>
- Sun, Q., Zybaylov, B., Majeran, W., Friso, G., Olinares, P.D.B., van Wijk, K.J., 2009. PPDB, the Plant Proteomics Database at Cornell. *Nucleic Acids Res.* 37, 969–974. <https://doi.org/10.1093/nar/gkn654>
- Suzuki, J.Y., Amore, T.D., Calla, B., Palmer, N.A., Scully, E.D., Sattler, S.E., Sarath, G., Lichty, J.S., Myers, R.Y., Keith, L.M., Matsumoto, T.K., Geib, S.M., 2017. Organ-specific transcriptome profiling of metabolic and pigment biosynthesis pathways in the floral ornamental progenitor species *Anthurium amnicola* Dressler. *Sci. Rep.* 7, 1–15. <https://doi.org/10.1038/s41598-017-00808-2>
- Suzuki, M., Takahashi, S., Kondo, T., Dohra, H., Ito, Y., Kiriiwa, Y., Hayashi, M., Kamiya, S., Kato, M., Fujiwara, M., Fukao, Y., Kobayashi, M., Nagata, N., Motohashi, R., 2015. Plastid proteomic analysis in tomato fruit development. *PLoS One* 10, 1–25. <https://doi.org/10.1371/journal.pone.0137266>
- Tanz, S.K., Tetu, S.G., Vella, N.G.F., Ludwig, M., 2009. Loss of the Transit Peptide and an Increase in Gene Expression of an Ancestral Chloroplastic Carbonic Anhydrase Were Instrumental in the Evolution of the Cytosolic C4 Carbonic Anhydrase in *Flaveria*. *Plant Physiol.* 150, 1515–1529. <https://doi.org/10.1104/pp.109.137513>

- Teixeira, P.F., Glaser, E., 2013. Processing peptidases in mitochondria and chloroplasts. *Biochim. Biophys. Acta - Mol. Cell Res.* 1833, 360–370.
<https://doi.org/10.1016/j.bbamcr.2012.03.012>
- Tokuriki, N., Tawfik, D.S., 2009. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* 19, 596–604. <https://doi.org/10.1016/j.sbi.2009.08.003>
- Tonkin, C.J., Foth, B.J., Ralph, S. a, Struck, N., Cowman, A.F., McFadden, G.I., 2008. Evolution of malaria parasite plastid targeting sequences. *Proc. Natl. Acad. Sci.* 105, 4781–4785.
<https://doi.org/10.1073/pnas.0707827105>
- Tordai, H., Patthy, L., 2004. Insertion of spliceosomal introns in proto-splice sites: The case of secretory signal peptides. *FEBS Lett.* 575, 109–111.
<https://doi.org/10.1016/j.febslet.2004.08.045>
- Ueda, M., Fujimoto, M., Arimura, S.I., Tsutsumi, N., Kadowaki, K.I., 2006. Evidence for transit peptide acquisition through duplication and subsequent frameshift mutation of a preexisting protein gene in rice. *Mol. Biol. Evol.* 23, 2405–2412.
<https://doi.org/10.1093/molbev/msl112>
- Vaira, A.M., Lim, H.S., Bauchan, G., Gulbranson, C.J., Miozzi, L., Vinals, N., Natilla, A., Hammond, J., 2018. The interaction of lolium latent virus major coat protein with ankyrin repeat protein NbANKr redirects it to chloroplasts and modulates virus infection. *J. Gen. Virol.* 99, 730–742. <https://doi.org/10.1099/jgv.0.001043>
- Van't Hof, R., Demel, R.A., Keegstra, K., Kruijff, B. De, 1991. Lipid-Peptide Interactions Between Fragments of the Transit Peptide of Ribulase-1,5-Bisphosphate Carboxylase/Oxygenase and Chloroplast Membrane Lipids. *EMBO J.* 29, 0–4.
- Vassarotti, a, Stroud, R., Douglas, M., 1987. Independent mutations at the amino terminus of a

- protein act as surrogate signals for mitochondrial import. *EMBO J.* 6, 705–11.
<https://doi.org/10.1002/j.1460-2075.1987.tb04811.x>
- Vetter, I.R., Wittinghofer, A., 2001. The Guanine in Switch Three Dimensions. *Science* (80-.).
294, 1299–1304. <https://doi.org/10.1126/science.1062023>
- Vibrantovski, M.D., Sakabe, N.J., De Souza, S.J., 2006. A possible role of exon-shuffling in the evolution of signal peptides of human proteins. *FEBS Lett.* 580, 1621–1624.
<https://doi.org/10.1016/j.febslet.2006.01.094>
- Vitulo, N., Forcato, C., Carpinelli, E.C., Telatin, A., Campagna, D., D'Angelo, M., Zimbello, R., Corso, M., Vannozzi, A., Bonghi, C., Lucchin, M., Valle, G., 2014. A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. *BMC Plant Biol.* 14, 20–30. <https://doi.org/10.1186/1471-2229-14-99>
- Walker, K., Croteau, R., 2001. Taxol biosynthetic genes. *Phytochemistry* 58, 1–7.
- Wang, G., Tian, L., Aziz, N., Broun, P., Dai, X., He, J., King, A., Zhao, P.X., Dixon, R.A., 2008. Terpene Biosynthesis in Glandular Trichomes of Hop. *Plant Physiol.* 148, 1254–1266.
<https://doi.org/10.1104/pp.108.125187>
- Wang, J., Tian, L., Madlung, A., Lee, H.S., Chen, M., Lee, J.J., Watson, B., Kagochi, T., Comai, L., Chen, Z.J., 2004. Stochastic and epigenetic changes of gene expression in Arabidopsis polyploids. *Genetics* 167, 1961–1973. <https://doi.org/10.1534/genetics.104.027896>
- Wang, Y.Q., Yang, Y., Fei, Z., Yuan, H., Fish, T., Thannhauser, T.W., Mazourek, M., Kochian, L. V., Wang, X., Li, L., 2013. Proteomic analysis of chromoplasts from six crop species reveals insights into chromoplast function and development. *J. Exp. Bot.* 64, 949–961.
<https://doi.org/10.1093/jxb/ers375>

- Williams, E.J.B., Pal, C., Hurst, L.D., Li, W.-H., 2000. The molecular evolution of signal peptides. *Gene* 253, 313–322.
- Win, J., Chaparro-Garcia, A., Belhaj, K., Saunders, D.G.O., Yoshida, K., Dong, S., Schornack, S., Zipfel, C., Robatzek, S., Hogenhout, S.A., Kamoun, S., 2012. Effector biology of plant-associated organisms: Concepts and perspectives. *Cold Spring Harb. Symp. Quant. Biol.* 77, 235–247. <https://doi.org/10.1101/sqb.2012.77.015933>
- Wischmann, C., Schuster, W., 1995. Transfer of rps10 from the mitochondrion to the nucleus in *Arabidopsis thaliana*: evidence for RNA-mediated transfer and exon shuffling at the integration site. *FEBS Lett.* 374, 152–156. [https://doi.org/10.1016/0014-5793\(95\)01100-S](https://doi.org/10.1016/0014-5793(95)01100-S)
- Wolter, F.P., Fritz, C.C., Willmitzer, L., Schell, J., Schreier, P.H., 1988. rbcS genes in *Solanum tuberosum*: conservation of transit peptide and exon shuffling during evolution. *Proc. Natl. Acad. Sci.* 85, 846–850. <https://doi.org/10.1073/pnas.85.3.846>
- Xia, Y., Levitt, M., 2002. Roles of mutation and recombination in the evolution of protein thermodynamics. *Proc. Natl. Acad. Sci.* 99, 10382–10387. <https://doi.org/10.1073/pnas.162097799>
- Yu, Q., Lutz, K.A., Maliga, P., 2017. Efficient Plastid Transformation in *Arabidopsis*. *Plant Physiol.* 175, 186–193. <https://doi.org/10.1104/pp.17.00857>
- Zhang, X.P., Glaser, E., 2002. Interaction of plant mitochondrial and chloroplast signal peptides with the Hsp70 molecular chaperone. *Trends Plant Sci.* 7, 14–21. [https://doi.org/10.1016/S1360-1385\(01\)02180-X](https://doi.org/10.1016/S1360-1385(01)02180-X)
- Zybailov, B., Rutschow, H., Friso, G., Rudella, A., Emanuelsson, O., Sun, Q., van Wijk, K.J., 2008. Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. *PLoS One* 3, e1994. <https://doi.org/10.1371/journal.pone.0001994>

Table 1: Residue Frequency Bias in Selected Datasets. Three sources of bias in transit peptide sequences were examined, comparing transit peptides to whole proteome sequence (column 2), comparing predicted to experimentally-validated transit peptides (column 3), and comparing sequences derived from predicted transit peptides or proteins from different taxa (columns 4-7). For each column, the bias is reported as the percentage change of the residue frequency in the subject dataset divided by the frequency in the query dataset. Heatmaps illustrating the magnitude of differences were generated using conditional formatting in Microsoft Excel, and are represented with white representing no change, blue representing 100% or higher changes, and red representing -100% or lower changes. Extreme frequency bias was observed for transit peptides of experimentally-validated proteins as compared with whole proteome sequence. Smaller but significant biases were observed for comparing predicted transit peptides to transit peptides validated by mass spectrometry. Finally, major bias was found for several residues when comparing predicted transit peptides of monocot species to transit peptides either eudicot species or *Amborella trichopoda*. Minor differences were noted for the whole proteomes of these same taxa, and little difference was observed between transit peptides of eudicot species and *Amborella*.

	Transit Peptide Bias	Prediction Bias	Taxa Bias (Predicted)			
Subject	cTP (MS-Validated)	Predicted cTP	Monocot CTP	Monocot Whole Proteome	Monocot CTP	Eudicot CTP
Query	TAIR10 Proteins	MS-Validated CTP	Eudicot CTP	Eudicot Whole Proteome	Amborella CTP	Amborella CTP
Ala	43.4	3.4	111.6	20.1	112.1	0.2
Cys	13.5	25.1	4.1	2.0	-4.7	-8.5
Asp	-68.9	-45.8	-2.2	1.1	5.6	8.0
Glu	-70.3	-48.1	-10.6	-4.5	-25.2	-16.3
Phe	-4.0	-3.5	-47.7	-7.0	-45.4	4.5
Gly	-30.1	-20.3	55.7	6.6	37.5	-11.7
His	10.2	27.1	-7.4	1.8	-12.9	-5.9
Ile	-37.1	-22.6	-51.3	-7.7	-51.8	-1.1
Lys	-32.4	-22.1	-54.8	-11.4	-52.0	6.3
Leu	-1.3	1.5	-4.8	-1.2	-11.1	-6.6
Met	14.7	-6.8	-1.4	0.6	-5.5	-4.2
Asn	-26.3	-3.2	-62.9	-11.7	-61.9	2.8
Pro	100.8	42.2	44.6	7.2	30.8	-9.5
Gln	-24.3	-6.9	-26.3	-3.8	-25.2	1.5
Arg	51.6	13.2	55.6	11.1	57.4	1.2
Ser	76.3	25.9	-23.1	-4.8	-17.9	6.8
Thr	25.4	7.8	-25.5	-2.5	-16.2	12.4
Val	-27.9	-22.4	4.5	1.2	17.1	12.1
Trp	-47.2	-13.4	22.1	0.6	-6.1	-23.2
Tyr	-63.5	-43.4	-37.7	-2.8	-41.9	-6.7

Table 2: Evolutionary Patterns of Transit Peptides. RAxML maximum likelihood software was used to resolve phylogenetic relationships of sequences within each candidate cluster, and the distribution of predicted transit peptides in each cluster was analyzed to determine whether the cluster had a single, monophyletic origin of the transit peptide, if multiple origins were detected, if a transit peptide was acquired and then lost, or if the most recent common ancestor of all sequences was likely to be plastid-targeted. Note that multiple scenarios can apply to the same cluster, so numbers do not add to 100%.

Dataset	Candidate Clusters	Polyphyletic Clusters	Rooted/Basal Clusters	Gain/Loss Clusters	Monophyletic Clusters
Arabidopsis1001	928	99	527	6	180
BrachyPan	7,551	1,616	4,616	116	2,272
Multi-Genome-RBH only	618	15	108	2	430
Multi-Genome-UCLUST only	1,443	38	293	11	1,061
Multi-Genome-Consensus	1,328	44	180	13	1,101

Table 3: Sources of Transit Peptide Evolution by Dataset. For each cluster, the dominant mechanism for transit peptide evolution was determined. Alterations that resulted in an apparent shift of the start site were prioritized, regardless of insertion or deletion size. If an alternative start site was not present, the mechanism responsible for the highest number of changes was selected as the dominant mechanism. Italicized datasets represent subsets of the multi-genome-merged dataset, and are presented to demonstrate that the proportion of each mechanism was similar in each subset.

Dataset	NTPs	Substitution	Alternative Start Site (Insertion)	Alternative Start Site (Deletion)	Independent Insertion	Independent Deletion
Arabidopsis1001	181	179 (98.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (1.1%)
BrachyPan	2,272	643 (28.3%)	457 (20.1%)	259 (11.4%)	446 (19.6%)	468 (20.6%)
Multi-Genome-Merged	2,592	832 (31.4%)	780 (29.4%)	230 (8.7%)	539 (20.3%)	269 (10.2%)
Multi-Genome-RBH only	<i>430</i>	<i>158 (36.2%)</i>	<i>103 (23.6%)</i>	<i>27 (6.2%)</i>	<i>100 (22.9%)</i>	<i>49 (11.2%)</i>
Multi-Genome-UCLUST only	<i>1,061</i>	<i>361 (33.4%)</i>	<i>277 (25.6%)</i>	<i>109 (10.1%)</i>	<i>209 (19.3%)</i>	<i>126 (11.6%)</i>
Multi-Genome-Consensus	<i>1,101</i>	<i>313 (27.7%)</i>	<i>400 (35.4%)</i>	<i>94 (8.3%)</i>	<i>230 (20.3%)</i>	<i>94 (8.3%)</i>

Table 4: GO term enrichment of Nascent Chloroplast-Targeted Proteins in Taxonomically Diverse Genotypes. Significance at $FDR < 0.05$ is shown for under- and over-represented terms. Terms associated with chloroplasts and localization were rare in NPTPs, while terms associated with metabolism, biosynthesis, gene expression, and protein interactions were highly enriched.

Tags	GO ID	GO Name	GO Category	FDR	P-Value
UNDER	GO:0051179	localization	BIOLOGICAL_PROCESS	5.06E-03	1.77E-05
UNDER	GO:0051234	establishment of localization	BIOLOGICAL_PROCESS	5.67E-03	2.04E-05
UNDER	GO:0006810	transport	BIOLOGICAL_PROCESS	6.37E-03	2.35E-05
UNDER	GO:0009607	response to biotic stimulus	BIOLOGICAL_PROCESS	2.28E-02	9.61E-05
UNDER	GO:0016020	membrane	CELLULAR_COMPONENT	2.12E-06	2.47E-09
UNDER	GO:0009579	thylakoid	CELLULAR_COMPONENT	1.29E-02	5.22E-05
UNDER	GO:0009536	plastid	CELLULAR_COMPONENT	3.36E-02	1.51E-04
OVER	GO:0031323	regulation of cellular metabolic process	BIOLOGICAL_PROCESS	1.46E-06	1.35E-09
OVER	GO:0032774	RNA biosynthetic process	BIOLOGICAL_PROCESS	1.46E-06	1.05E-09
OVER	GO:0097659	nucleic acid-templated transcription	BIOLOGICAL_PROCESS	1.46E-06	9.99E-10
OVER	GO:0051171	regulation of nitrogen compound metabolic process	BIOLOGICAL_PROCESS	1.46E-06	1.16E-09
OVER	GO:0034654	nucleobase-containing compound biosynthetic process	BIOLOGICAL_PROCESS	1.46E-06	8.66E-10
OVER	GO:0006351	transcription, DNA-templated	BIOLOGICAL_PROCESS	1.46E-06	9.99E-10
OVER	GO:0019219	regulation of nucleobase-containing compound metabolic process	BIOLOGICAL_PROCESS	1.46E-06	1.58E-09
OVER	GO:0080090	regulation of primary metabolic process	BIOLOGICAL_PROCESS	1.46E-06	1.55E-09
OVER	GO:2000112	regulation of cellular macromolecule biosynthetic process	BIOLOGICAL_PROCESS	2.43E-06	3.22E-09
OVER	GO:0031326	regulation of cellular biosynthetic process	BIOLOGICAL_PROCESS	2.43E-06	4.25E-09
OVER	GO:1903506	regulation of nucleic acid-templated transcription	BIOLOGICAL_PROCESS	2.43E-06	4.56E-09
OVER	GO:2001141	regulation of RNA biosynthetic process	BIOLOGICAL_PROCESS	2.43E-06	4.61E-09
OVER	GO:0009889	regulation of biosynthetic process	BIOLOGICAL_PROCESS	2.43E-06	4.80E-09
OVER	GO:0010556	regulation of macromolecule biosynthetic process	BIOLOGICAL_PROCESS	2.43E-06	3.29E-09
OVER	GO:0018130	heterocycle biosynthetic process	BIOLOGICAL_PROCESS	2.43E-06	4.77E-09
OVER	GO:0006355	regulation of transcription, DNA-templated	BIOLOGICAL_PROCESS	2.43E-06	4.56E-09
OVER	GO:1901362	organic cyclic compound biosynthetic process	BIOLOGICAL_PROCESS	3.00E-06	6.23E-09
OVER	GO:0019438	aromatic compound biosynthetic process	BIOLOGICAL_PROCESS	3.00E-06	6.47E-09
OVER	GO:0051252	regulation of RNA metabolic process	BIOLOGICAL_PROCESS	3.64E-06	8.17E-09
OVER	GO:0090304	nucleic acid metabolic process	BIOLOGICAL_PROCESS	3.35E-05	7.82E-08
OVER	GO:0016070	RNA metabolic process	BIOLOGICAL_PROCESS	1.49E-04	4.02E-07
OVER	GO:0060255	regulation of macromolecule metabolic process	BIOLOGICAL_PROCESS	0.005061	1.77E-05

OVER	GO:0019222	regulation of metabolic process	BIOLOGICAL_PROCESS	0.009434	3.65E-05
OVER	GO:0006357	regulation of transcription by RNA polymerase II	BIOLOGICAL_PROCESS	0.010222	4.04E-05
OVER	GO:0034641	cellular nitrogen compound metabolic process	BIOLOGICAL_PROCESS	0.02888	1.27E-04
OVER	GO:0010468	regulation of gene expression	BIOLOGICAL_PROCESS	0.036366	1.67E-04
OVER	GO:0090083	regulation of inclusion body assembly	BIOLOGICAL_PROCESS	0.036923	1.86E-04
OVER	GO:0090084	negative regulation of inclusion body assembly	BIOLOGICAL_PROCESS	0.036923	1.86E-04
OVER	GO:0070841	inclusion body assembly	BIOLOGICAL_PROCESS	0.036923	1.86E-04
OVER	GO:0005667	transcription factor complex	CELLULAR_COMPONENT	2.49E-08	2.23E-12
OVER	GO:0032777	Piccolo NuA4 histone acetyltransferase complex	CELLULAR_COMPONENT	1.28E-06	2.31E-10
OVER	GO:0035267	NuA4 histone acetyltransferase complex	CELLULAR_COMPONENT	1.46E-06	1.53E-09
OVER	GO:0043189	H4/H2A histone acetyltransferase complex	CELLULAR_COMPONENT	1.46E-06	1.53E-09
OVER	GO:1902562	H4 histone acetyltransferase complex	CELLULAR_COMPONENT	2.43E-06	3.73E-09
OVER	GO:0000123	histone acetyltransferase complex	CELLULAR_COMPONENT	1.42E-04	3.44E-07
OVER	GO:1902493	acetyltransferase complex	CELLULAR_COMPONENT	1.49E-04	3.89E-07
OVER	GO:0031248	protein acetyltransferase complex	CELLULAR_COMPONENT	1.49E-04	3.89E-07
OVER	GO:0044451	nucleoplasm part	CELLULAR_COMPONENT	5.93E-04	1.65E-06
OVER	GO:0090575	RNA polymerase II transcription factor complex	CELLULAR_COMPONENT	0.001407	4.05E-06
OVER	GO:0044798	nuclear transcription factor complex	CELLULAR_COMPONENT	0.00427	1.34E-05
OVER	GO:0005669	transcription factor TFIID complex	CELLULAR_COMPONENT	0.004629	1.54E-05
OVER	GO:1990234	transferase complex	CELLULAR_COMPONENT	0.00769	2.90E-05
OVER	GO:0016864	intramolecular oxidoreductase activity, transposing S-S bonds	MOLECULAR_FUNCTION	0.001435	4.38E-06
OVER	GO:0003756	protein disulfide isomerase activity	MOLECULAR_FUNCTION	0.001435	4.38E-06
OVER	GO:0005515	protein binding	MOLECULAR_FUNCTION	0.004535	1.47E-05
OVER	GO:0016901	oxidoreductase activity, acting on the CH-OH group of donors, quinone or similar compound as acceptor	MOLECULAR_FUNCTION	0.022345	9.24E-05
OVER	GO:0016671	oxidoreductase activity, acting on a sulfur group of donors, disulfide as acceptor	MOLECULAR_FUNCTION	0.02337	1.01E-04
OVER	GO:0047405	pyrimidine-5'-nucleotide nucleosidase activity	MOLECULAR_FUNCTION	0.036923	1.86E-04
OVER	GO:0044183	protein folding chaperone	MOLECULAR_FUNCTION	0.036923	1.86E-04

Table 5: GO term enrichment of Arabidopsis1001 NPTs. Due to the small number of final NPTs in this dataset, no results were significant at an FDR < 0.05 significance threshold. Therefore, data significant at p-value < 0.05 is presented.

Tags	GO ID	GO Name	GO Category	FDR	P-Value
UNDER	GO:0043227	membrane-bounded organelle	CELLULAR_COMPONENT	1	0.014882
UNDER	GO:0043231	intracellular membrane-bounded organelle	CELLULAR_COMPONENT	1	0.014882
UNDER	GO:0071704	organic substance metabolic process	BIOLOGICAL_PROCESS	1	0.033193
UNDER	GO:0044238	primary metabolic process	BIOLOGICAL_PROCESS	1	0.010442
OVER	GO:0042726	flavin-containing compound metabolic process	BIOLOGICAL_PROCESS	1	2.15E-02
OVER	GO:0042727	flavin-containing compound biosynthetic process	BIOLOGICAL_PROCESS	1	2.15E-02
OVER	GO:0006662	glycerol ether metabolic process	BIOLOGICAL_PROCESS	1	3.21E-02
OVER	GO:0009231	riboflavin biosynthetic process	BIOLOGICAL_PROCESS	1	2.15E-02
OVER	GO:0003919	FMN adenyltransferase activity	MOLECULAR_FUNCTION	1	1.62E-02
OVER	GO:0006766	vitamin metabolic process	BIOLOGICAL_PROCESS	1	4.78E-02
OVER	GO:0006771	riboflavin metabolic process	BIOLOGICAL_PROCESS	1	2.15E-02
OVER	GO:0006767	water-soluble vitamin metabolic process	BIOLOGICAL_PROCESS	1	4.78E-02
OVER	GO:0004143	diacylglycerol kinase activity	MOLECULAR_FUNCTION	1	1.62E-02
OVER	GO:0007205	protein kinase C-activating G protein-coupled receptor signaling pathway	BIOLOGICAL_PROCESS	1	1.62E-02
OVER	GO:0007186	G protein-coupled receptor signaling pathway	BIOLOGICAL_PROCESS	1	2.68E-02
OVER	GO:0042364	water-soluble vitamin biosynthetic process	BIOLOGICAL_PROCESS	1	4.78E-02
OVER	GO:0070566	adenyltransferase activity	MOLECULAR_FUNCTION	1	2.68E-02
OVER	GO:0016757	transferase activity, transferring glycosyl groups	MOLECULAR_FUNCTION	1	2.80E-02
OVER	GO:0016758	transferase activity, transferring hexosyl groups	MOLECULAR_FUNCTION	1	1.99E-02
OVER	GO:0009110	vitamin biosynthetic process	BIOLOGICAL_PROCESS	1	4.78E-02
OVER	GO:0018904	ether metabolic process	BIOLOGICAL_PROCESS	1	0.03212

Table 6: GO term enrichment of BrachyPan NPTPs. Due to the small number of final NPTPs in this dataset, few results were found using $FDR < 0.05$ as a significance threshold. Therefore, data significant at $p\text{-value} < 0.05$ is presented.

Tags	GO ID	GO Name	GO Category	FDR	P-Value
UNDER	GO:0005737	cytoplasm	CELLULAR_COMPONENT	0.026818	2.10E-04
UNDER	GO:0044444	cytoplasmic part	CELLULAR_COMPONENT	0.045423	5.32E-04
UNDER	GO:0005198	structural molecule activity	MOLECULAR_FUNCTION	0.106841	0.002523
UNDER	GO:0003735	structural constituent of ribosome	MOLECULAR_FUNCTION	0.155505	0.005467
UNDER	GO:1990904	ribonucleoprotein complex	CELLULAR_COMPONENT	0.178068	0.007651
UNDER	GO:0005840	ribosome	CELLULAR_COMPONENT	0.178068	0.007651
UNDER	GO:1901363	heterocyclic compound binding	MOLECULAR_FUNCTION	0.37075	0.036206
UNDER	GO:0003676	nucleic acid binding	MOLECULAR_FUNCTION	0.37075	0.036206
UNDER	GO:0097159	organic cyclic compound binding	MOLECULAR_FUNCTION	0.37075	0.036206
OVER	GO:0043167	ion binding	MOLECULAR_FUNCTION	8.38E-05	3.27E-07
OVER	GO:0005488	binding	MOLECULAR_FUNCTION	0.064876	1.01E-03
OVER	GO:0071941	nitrogen cycle metabolic process	BIOLOGICAL_PROCESS	0.106841	2.38E-03
OVER	GO:0071554	cell wall organization or biogenesis	BIOLOGICAL_PROCESS	0.106841	0.002921
OVER	GO:0051276	chromosome organization	BIOLOGICAL_PROCESS	0.117522	0.003673
OVER	GO:0043233	organelle lumen	CELLULAR_COMPONENT	0.269756	0.015806
OVER	GO:0070013	intracellular organelle lumen	CELLULAR_COMPONENT	0.269756	0.015806
OVER	GO:0031981	nuclear lumen	CELLULAR_COMPONENT	0.269756	0.015806
OVER	GO:0031974	membrane-enclosed lumen	CELLULAR_COMPONENT	0.269756	0.015806
OVER	GO:0006629	lipid metabolic process	BIOLOGICAL_PROCESS	0.279342	0.017459
OVER	GO:0005654	nucleoplasm	CELLULAR_COMPONENT	0.348322	0.028573
OVER	GO:0016301	kinase activity	MOLECULAR_FUNCTION	0.348322	0.023589
OVER	GO:0120025	plasma membrane bounded cell projection	CELLULAR_COMPONENT	0.348322	0.028467
OVER	GO:0042995	cell projection	CELLULAR_COMPONENT	0.348322	0.028467
OVER	GO:0005929	cilium	CELLULAR_COMPONENT	0.348322	0.028467
OVER	GO:0016772	transferase activity, transferring phosphorus-containing groups	MOLECULAR_FUNCTION	0.37075	0.034041
OVER	GO:0065007	biological regulation	BIOLOGICAL_PROCESS	0.375382	0.038125
OVER	GO:0003824	catalytic activity	MOLECULAR_FUNCTION	0.401607	0.042357
OVER	GO:0044428	nuclear part	CELLULAR_COMPONENT	0.424632	0.046444

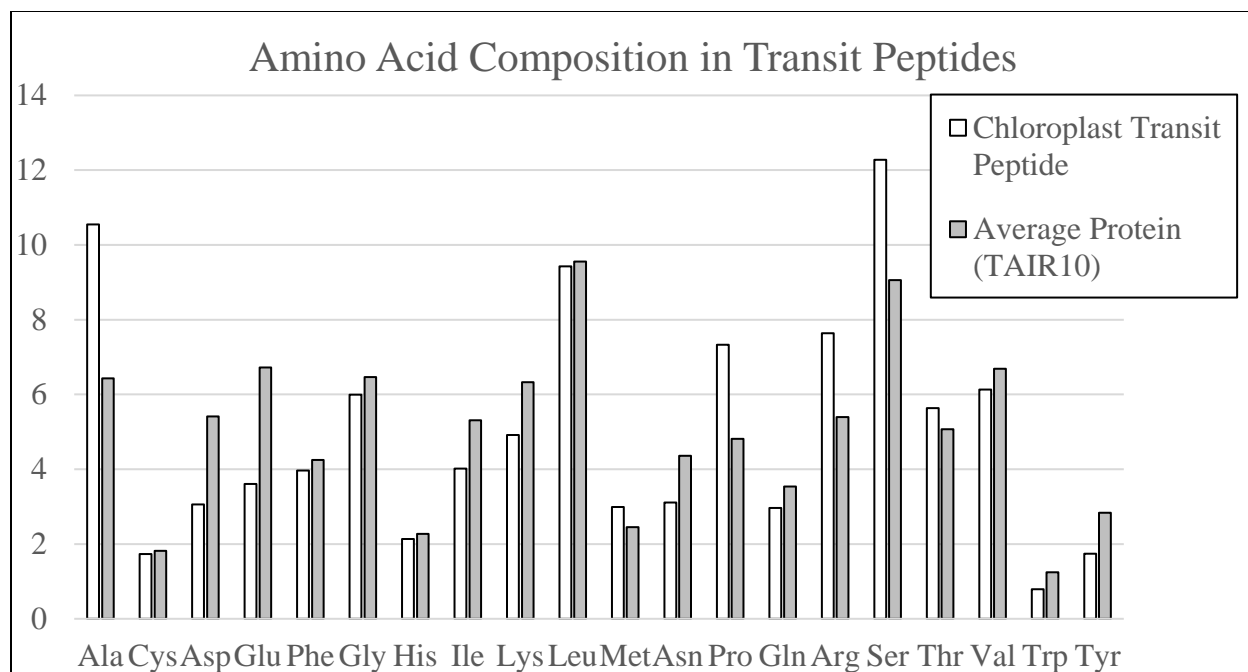


Figure 1: Amino Acid Compositional Changes in Transit Peptides. The first 60 residues of Arabidopsis proteins validated by mass spectrometry were analyzed for residue composition and compared with the average residue composition of all Arabidopsis proteins. Extreme enrichment was observed for alanine, proline, arginine, and serine, while significant depletion was found for aspartic acid, glutamine, isoleucine, lysine, asparagine, and tyrosine. Although glycine, leucine, threonine, and valine were abundant, they did not differ significantly from the average residue content in Arabidopsis proteins.

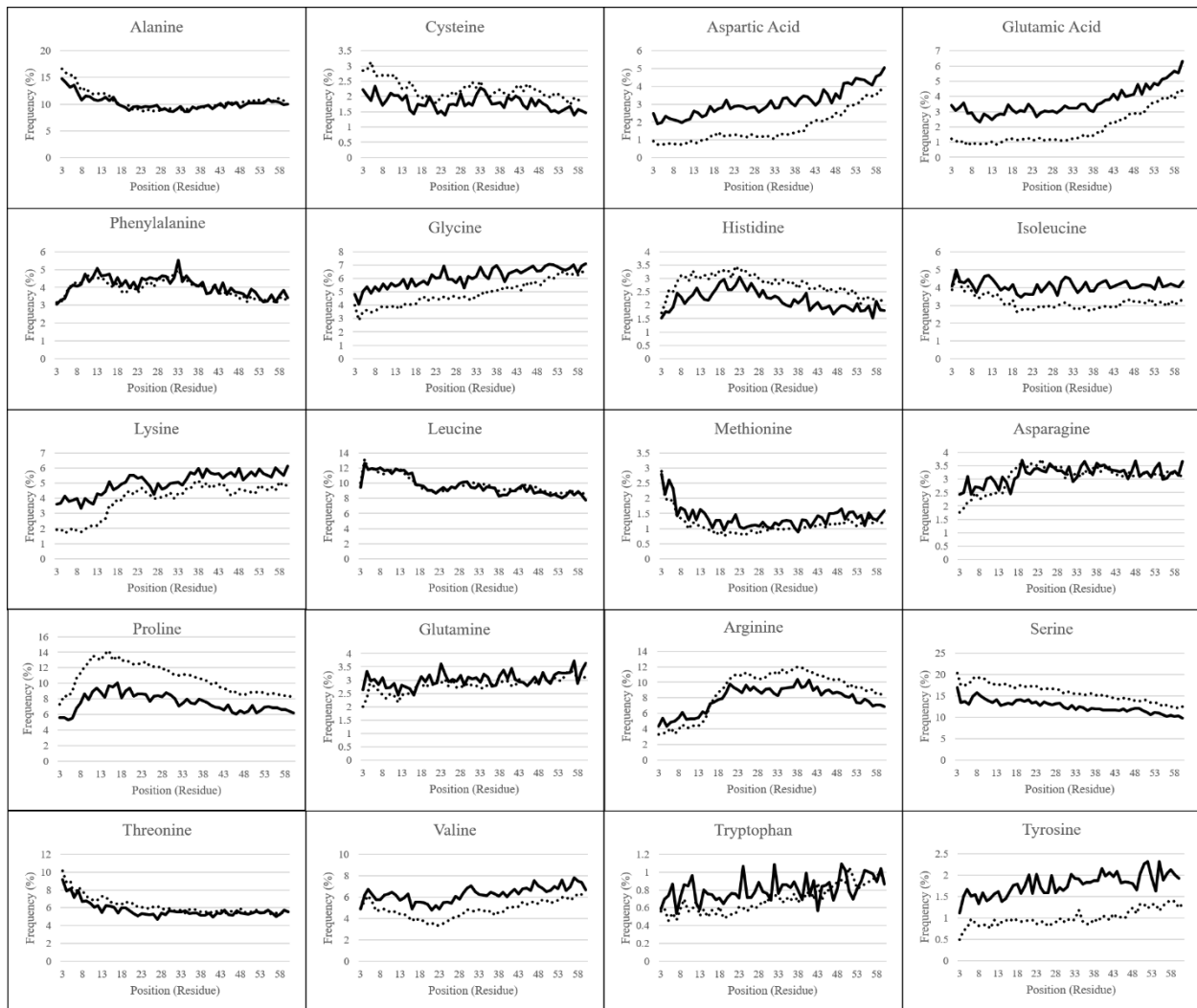


Figure 2: Residue Frequency in Predicted and Experimentally-Validated Transit Peptides.

The first 60 residues of sequences validated to be plastid-targeted by mass spectrometry methods (solid lines) and of predicted plastid-targeted proteins in 15 plant genomes (dotted lines) were collected, and residue composition was assessed for each position. Positions 1 and 2 are omitted from each graph due to skew by methionine and alanine, respectively. Frequency patterns match almost exactly between experimentally-validated and predicted plastid-targeted proteins, but there are differences in the absolute frequency for many amino acids, in particular, highly-enriched and highly-depleted residues.

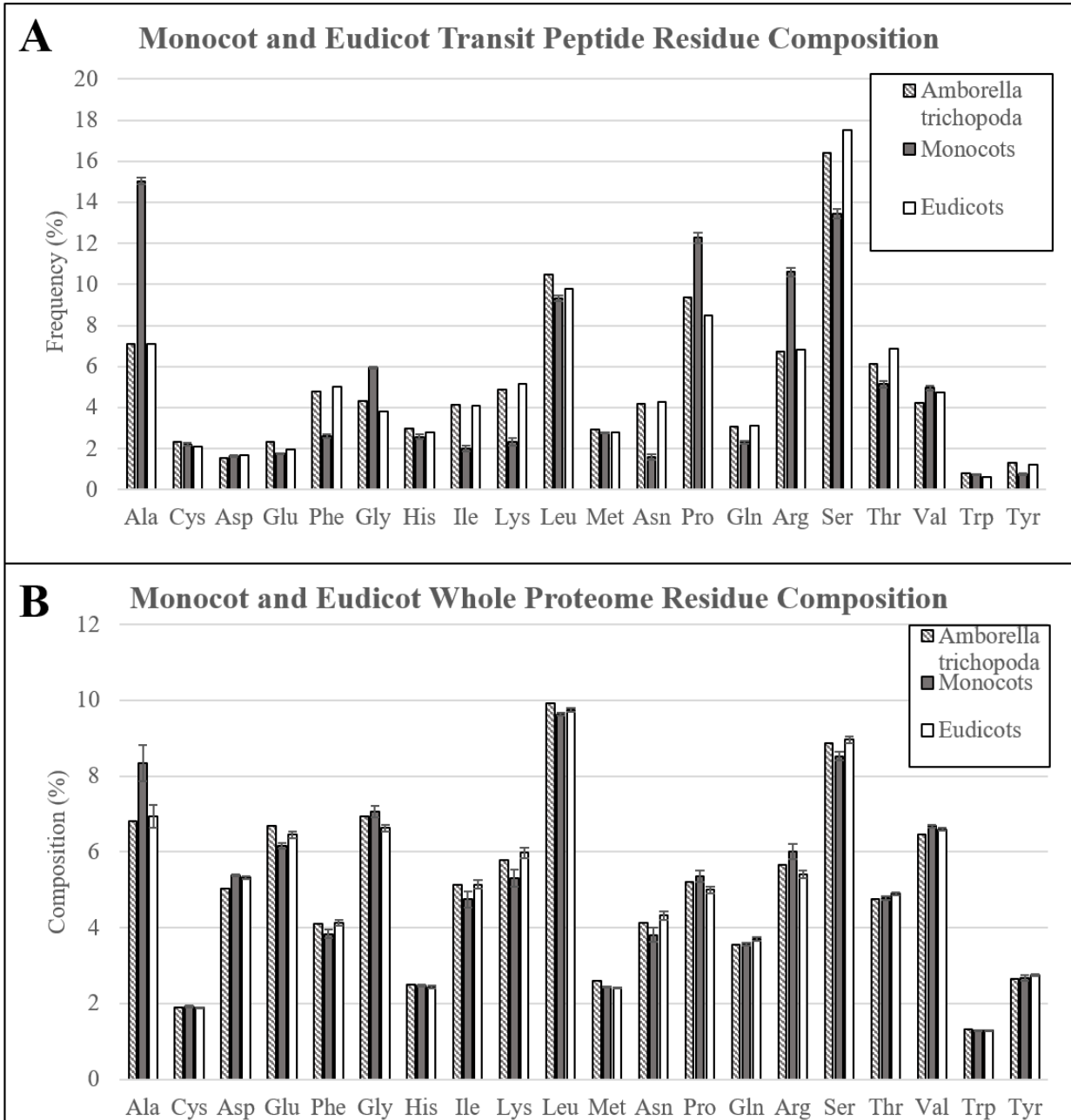


Figure 3: Residue Composition of Predicted Transit Peptides and Whole Proteome

Sequence. (A) Residue composition of predicted transit peptides shows significant enrichment in alanine, leucine, proline, arginine, and serine in all assessed organisms, but predicted transit peptides of monocots were highly overrepresented in alanine, proline, and arginine.

Corresponding decreases in serine and many of the minor amino acids including phenylalanine, isoleucine, lysine, and asparagine were also found in monocot sequences. *A. trichopoda* sequences closely matched eudicot sequences. (B) Residue composition of whole proteomes found that alanine was about 2% higher in monocot sequences, but this does not explain the extreme differences found for predicted transit peptides.

Mechanism	Illustration	Phenotype
Substitution		Plastid-Targeted Non-Plastidial
Alternative Initiation Exon		Plastid-Targeted Non-Plastidial
Alternative Splice Variant		Plastid-Targeted Non-Plastidial
Alternative Start Site		Plastid-Targeted Non-Plastidial

Figure 4: Models of Transit Peptide Evolution. In each panel, exons are indicated with shaded boxes and introns with black lines. The initial RNA molecule is indicated in black, and protein isoforms are indicated in light green. Substitution variants are the only mechanism requiring a change to the DNA sequence, although sequence variants may also promote or hinder the other mechanisms. Adapted from Davis et al., 2006.

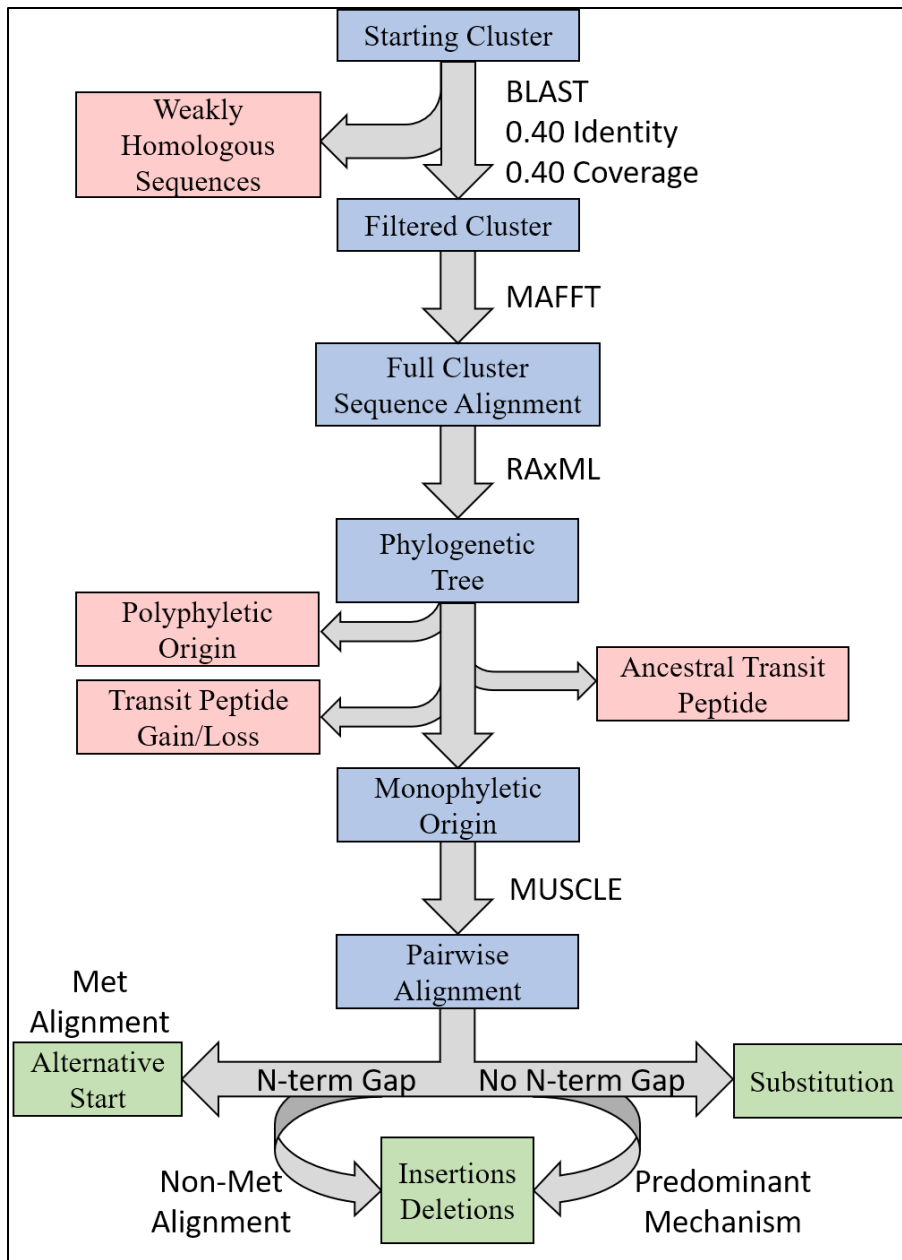


Figure 5: Illustration of Phylogenetics Workflow. For each candidate cluster representing a potential NPTP, the steps for filtering, alignment, tree prediction, and mutation analysis are depicted. Blue boxes indicate the path of candidate clusters, red boxes indicate sequences or clusters that are filtered out, green boxes indicate the potential mutational categories, and programs or conditions used in the workflow are indicated to the right of arrows.

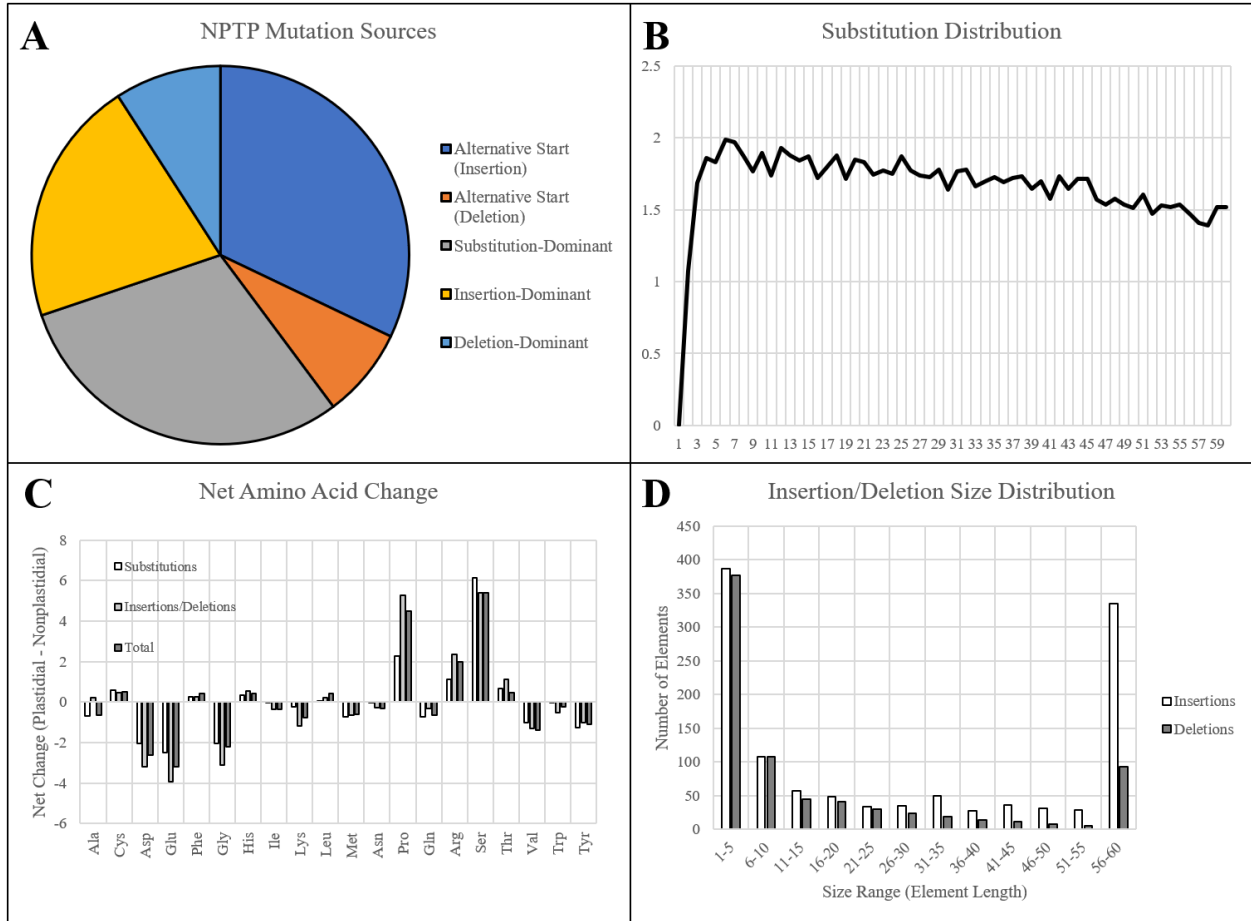


Figure 6: NPTP Mutation Sources and Characteristics in Multi-Genomic Analysis. Merged results for RBH, UCLUST, and Consensus datasets are presented here. (A) Clusters in which an upstream alternative start site caused the acquisition of a transit peptide were most abundant, followed by substitutions and insertions. Deletions were comparatively rare. (B) A slight but significant linear decrease in the amino acid substitution frequency at each position was observed, indicating that positions at the proximal end of the N-terminus have a greater effect on transit peptide prediction strength. (C) Arginine, proline, and serine experienced significant net increases, while aspartic acid, glutamic acid, and glycine all experienced more than 2% net decrease when nascent transit peptides were compared to the closest non-targeted neighbor sequence. Net change of amino acids was generally similar between substitutions and insertions/deletions, but proline and arginine were far less likely to be acquired due to substitutions (D) Most insertions and deletions were between 1-5 amino acids in length and generally decreased in frequency as the size increased. However, elements that covered the entirely 60-residue length of the transit peptide region were extremely abundant, especially those caused by insertions.

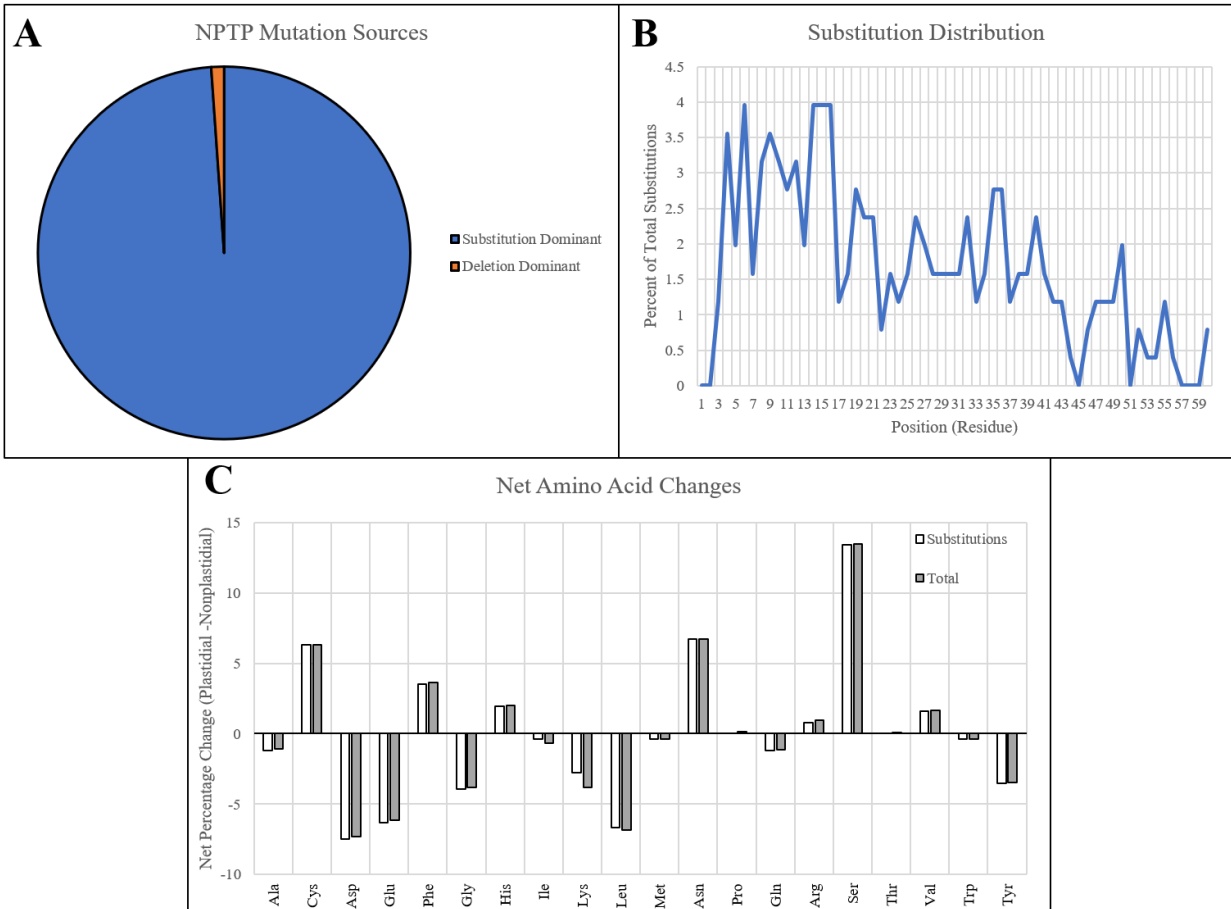


Figure 7: NPTP Mutation Sources and Characteristics in Arabidopsis1001. (A) The proportion NPTP mechanisms in Arabidopsis1001 showed only one instance of an internal deletion and one instance of a C-terminal deletion causing a difference in targeting prediction, while the remaining instances were caused by substitutions. (B) A significant negative trend in substitution frequency was observed for Arabidopsis1001 sequences, with up to 4% of substitutions occurring at positions in the proximal end, and 0-1% of substitutions occurring at positions in the distal end. (C) significant increases in cysteine, asparagine, and serine were observed in the substitutions, while aspartic acid, glutamic acid, and leucine had significant decreases.

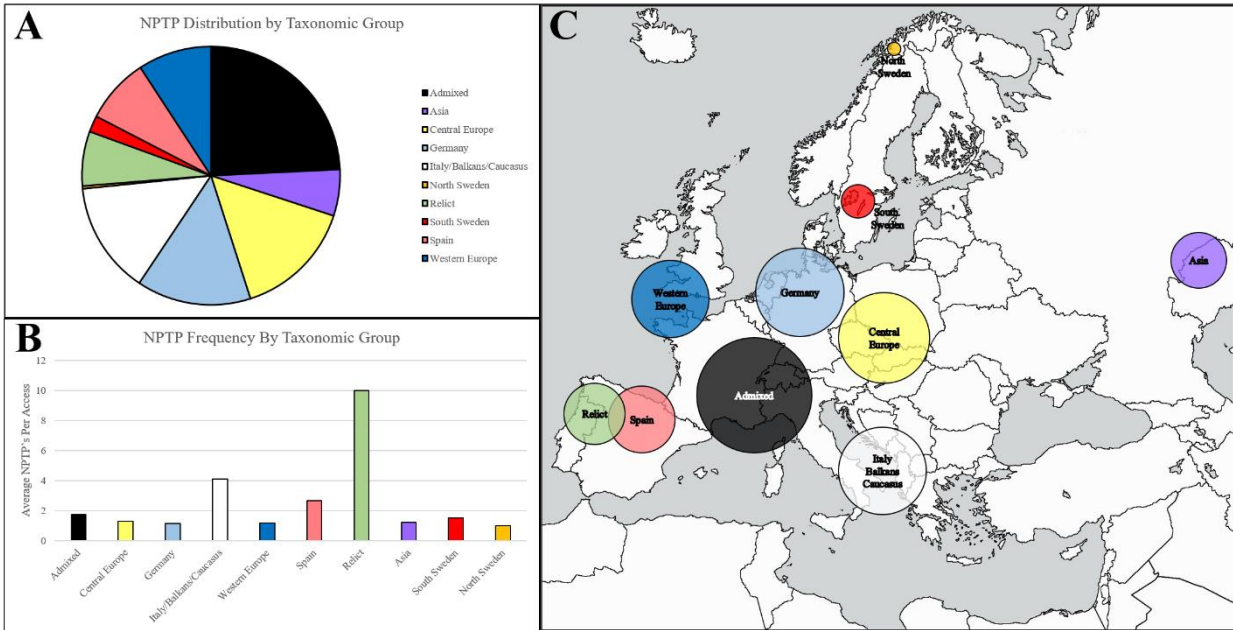


Figure 8: Geographic Distribution of NPTPs in Arabidopsis1001 Accessions. Taxonomic groups were referenced according to the Arabidopsis1001 Proteomes project, and the number of NPTP's for each accession were added to the respective taxonomic group. The admixed group, of which Columbia-0 is a part of, accounted for the most NPTP's (A) but was overall one of the least diverse taxonomic groups. Relict, Italian/Balkan/Caucasus, and Spanish accessions contained the most NPTPs per genotype, while Asian, Swedish, Germanic, and Central European accessions contained the fewest (B). Geographic distribution is indicated in (C), where shaded circular areas are indicative of magnitude. Taxonomic groups are color-coded according to Joshi et al., 2012.

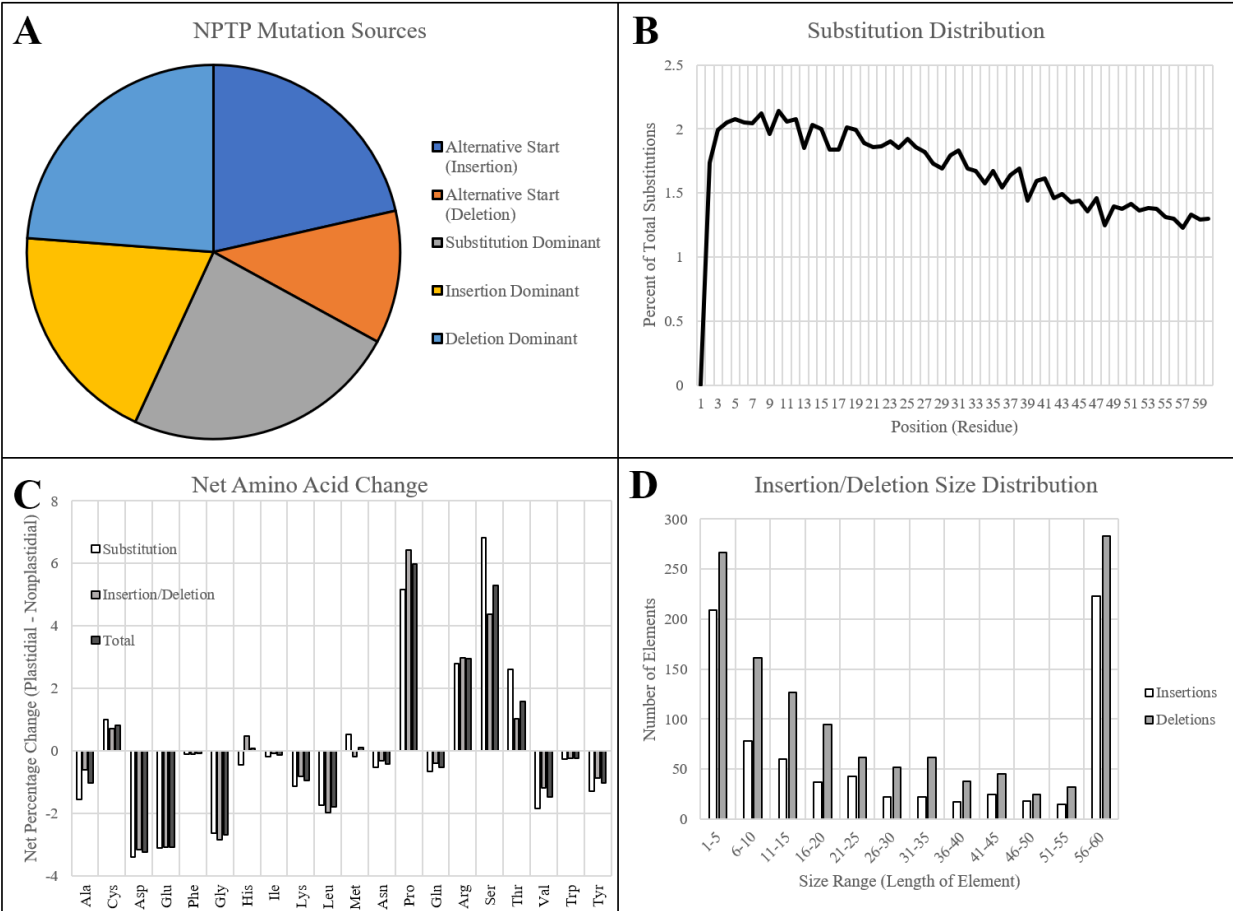


Figure 9: NPTP Mutation Sources and Characteristics in BrachyPan. (A) Substitutions were most dominant at 28.3%, followed by independent deletions, upstream alternative start sites, and independent deletions. (B) Substitution frequency was greatest at the proximal end and decreased linearly to the distal end. (C) the net change in residue composition favored increases in proline, arginine, serine, and threonine in transit peptides, while aspartic and glutamic acids, glycine, and leucine had 2% or greater decreases. (D) Insertions or deletions which covered the entirety of the transit peptide region were most abundant, indicating possible exon swapping. Of the remaining elements, smaller elements were most abundant and decreased in frequency with increasing size.

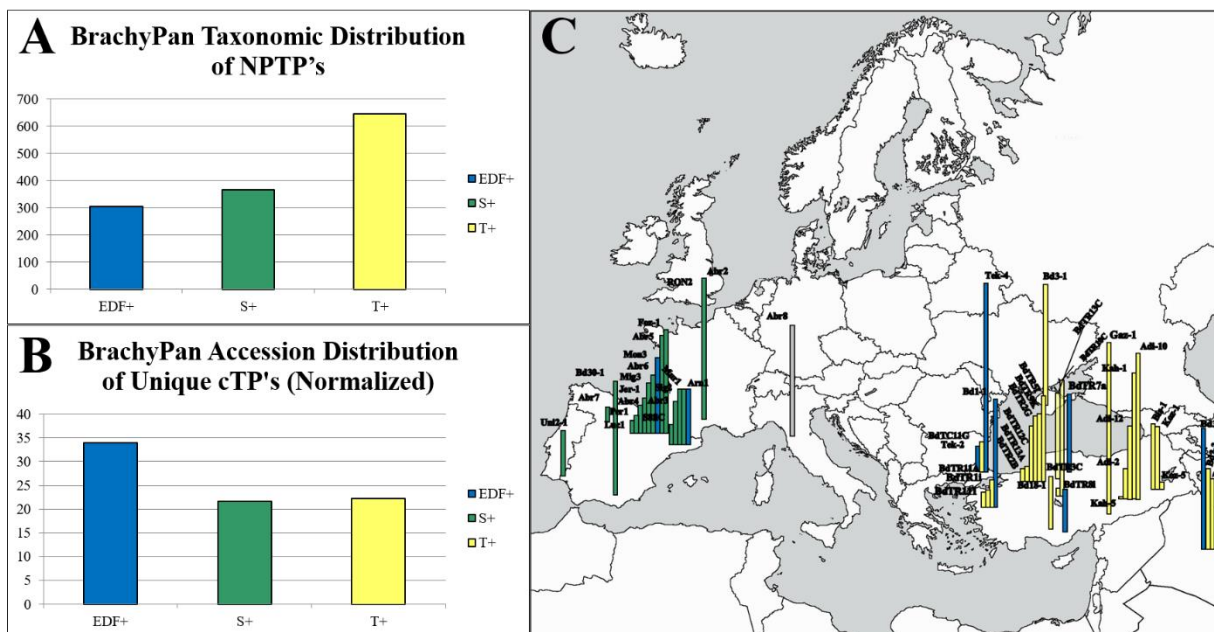


Figure 10: Geographic Distribution of NPTPs in BrachyPan Accessions. The Turkish (T+) taxonomic group had the most NPTPs in absolute numbers (A), but when measured per accession, the extremely-delayed flowering taxonomic group (EDF+) was nearly twice as divergent (B). Geographic distribution of these lines is indicated in (C), where the length of each bar corresponds the number of NPTPs for each labeled accession. Taxonomic groups are color-coded according to Gordon et al., 2017.

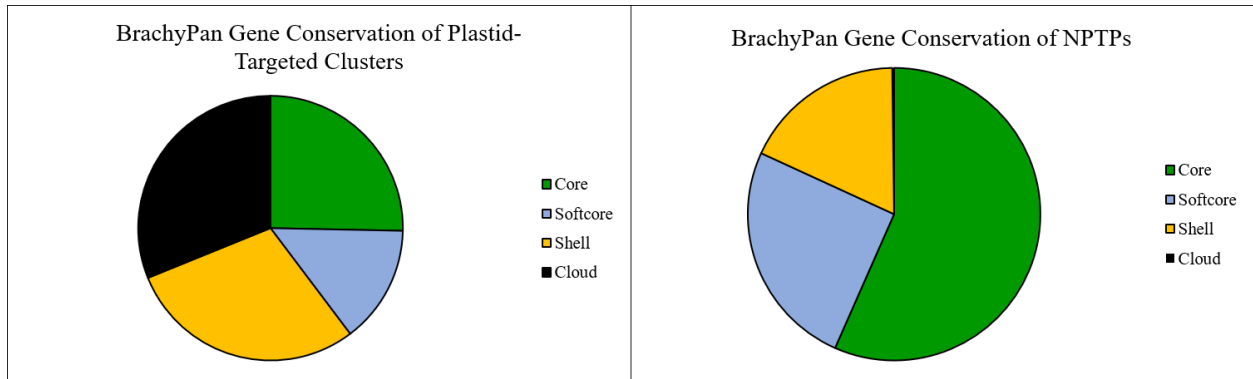


Figure 11: Gene Conservation of NPTPs in BrachyPan. “Core” genes are those shared by all 56 accessions, “Softcore” genes are shared by 53-55 accessions, “Shell” genes are shared by 3-52 accessions, “Cloud” genes are found in 1-2 accessions. While most gene clusters reported by Gordon et al. are in the “Shell” category, the “Core” and “Softcore” categories accounted for the most NPTPs. Categories are color-coded according to Gordon et al., 2017.

Additional Files

Additional File 1 – Perl scripts used for cluster analysis, phylogenetics, and mutation detection

Compressed ZIP file containing all custom scripts used to generate results.

Additional File 2 – Residue Frequency Analysis in Predicted and Experimentally-Validated Plastid-Targeted Proteins

Excel file containing residue frequency and comparisons between datasets.

Additional File 3 – Multi-Genome Mutation Analysis

Excel File containing RAxML best trees, full mutation analysis and categorization of each cluster in the multi-genome comparison. Sequences are also provided.

Additional File 4 – Arabidopsis1001 Mutation Analysis

Excel File containing RAxML best trees, full mutation analysis and categorization of each cluster in the Arabidopsis1001 dataset. Geographic and taxonomic analysis is also provided.

Additional File 5 - BrachyPan Mutation Analysis

Excel File containing RAxML best trees, full mutation analysis and categorization of each cluster in the BrachyPan dataset. Geographic and taxonomic analysis is also provided.

CHAPTER 4

Transplastomic Expression of the N-terminal Soluble Domain of Albino 3 Causes Perturbation of Calcium Homeostasis and Photosynthetic Performance in *Nicotiana tabacum*

Ryan Christian^{1,2}, Scott Schaeffer^{1,3}, Deepika Minhas², Seanna Hewitt^{1,2}, Bruce Williamson Benavides^{1,2}, Henning Kunz⁴, Magnus Wood⁵, Helmut Kirchhoff⁵, and Amit Dhingra^{1,2§}

Target Journal: Plant Physiology

¹Molecular Plant Sciences Program, Washington State University, Pullman, WA

²Department of Horticulture, Washington State University, Pullman, WA

³Indigo LLC, Boston, MA

⁴School of Biological Sciences, Washington State University, Pullman, WA

⁵Institute of Biological Chemistry, Washington State University, Pullman, WA

[§]Corresponding author

RWC: ryan_christian@wsu.edu

AD: adhingra@wsu.edu

Abstract

Albino3 (Alb3) is a membrane integrase protein of the thylakoid membrane where it is required for biogenesis of the thylakoid membrane and assembly of the photosystems and electron transport chain complexes. Significant evidence from the Alb3 homolog PPF-1 has implicated the protein in an additional calcium homeostasis role, which could have profound implications for chloroplast response to environmental signals and chloroplast senescence.

However, only indirect evidence for an Alb3 role in calcium transport has been reported to date. The hypothesis that the calcium-linked PPF-1 phenotype is dependent on membrane depolarization was tested by creating a truncation series of the N-terminal soluble domain of Alb3 and overexpressed them in the chloroplasts of *Nicotiana tabacum*. Fluorescent protein experiments demonstrated that residues 1-49 comprise the primary transit peptide signal, while residues 50-82 contain a likely thylakoid transfer domain. When these truncated proteins were overexpressed in the chloroplast, major changes to senescence, thylakoid ultrastructure, and chlorophyll fluorescence were observed that were dependent on presence of the Alb3 thylakoid transfer domain. Furthermore, the potassium, calcium, and magnesium composition of leaf tissue in transplastomic lines was significantly affected, and preliminary evidence that *in vivo* calcium dynamics is impacted is presented. These results have led to a hypothesis that preAlb3 causes constitutive use of the Alb3/SecY heterodimeric channel, inducing transient membrane permeability to ions and disruptions to thylakoid membrane potential and ion homeostasis.

Introduction

The myriad biochemical processes in the plastid are enabled by proteins that are imported from the cytosol and a small subset of protein that are synthesized within the organelle. Protein translocation across the plastid outer and inner envelope is a complex process requiring both post-translational and co-translational processes for nuclear-encoded and plastid-encoded gene products, respectively. At least three semi-independent pathways coordinate to transport proteins to the thylakoid membrane or lumen. Soluble thylakoid proteins can be transported through either the Δ pH-dependent twin arginine transport (TAT) pathway or ATP-dependent cpSec translocase pathway (Robinson and Bolhuis, 2001). Each of these translocases transport roughly

half of the more than 100 proteins residing in the thylakoid lumen (Peltier et al., 2002; Schubert et al., 2002), but TAT can transport pre-folded proteins, while the cpSec has a fixed pore size and can only process unfolded proteins (DeLisa et al., 2003; Hynds et al., 1998; Matos et al., 2008; Robinson et al., 2011). Membrane-bound, insoluble proteins can be inserted by the energy-independent Albino3 (Alb3), by the cpSec translocase, or by a heterodimer complex of these two proteins (reviewed in Hennon et al., 2015; Kuhn et al., 2003). Albino3 is a member of the YidC/Oxa1/Alb3 superfamily of proteins that is found in prokaryotes, mitochondria, and chloroplasts, respectively, all of which contain a highly conserved integrase domain consisting of five transmembrane helices (Kuhn et al., 2003). The degree of conservation is high enough that Oxa1 and Alb3 can functionally replace YidC *in vivo* (Benz et al., 2009; Jiang et al., 2002; Wang and Wang, 2009). Unlike TAT and Sec translocons which have physical pores for protein translocation, YidC and its paralogs have a unique fold that forms a hydrophilic, positively-charged groove that decreases the energetic cost of translocation of soluble loop domains (Jiang et al., 2003; Kumazaki et al., 2014b, 2014a). During co-dependent translocation by the heterodimer of YidC and SecY, YidC has been suggested to chaperone membrane proteins (Jong et al., 2010; Nagamori et al., 2004), facilitate release of transmembrane helices from SecY (Beck et al., 2001; Houben et al., 2004; Urbanus et al., 2001), promote formation of membrane protein complexes (Wagner et al., 2008) and control protein quality (van Bloois et al., 2008). In addition to the conserved integrase domain, the N- and C-termini contribute to the function of these proteins. The C-terminus, which faces the cytosol in case of YidC and the stroma in case of Alb3, contains binding sites for ribosomes (Chen et al., 2014; Gruschke et al., 2010; Haque et al., 2010a, 2010b; Jia et al., 2003; Kohler et al., 2009; Szyrach et al., 2003), signal recognition particle (SRP) subunits cpSRP54 and cpSRP43 (Bals et al., 2010; Dünschede et al., 2011; Falk et

al., 2010; Horn et al., 2015; Moore et al., 2003), and FtsY (Horn et al., 2015; Moore et al., 2003). At the N-terminus, which faces the periplasm in case of YidC and the thylakoid lumen in the case of Alb3, an amphipathic helix upstream of the first transmembrane domain lies parallel to the membrane and forms part of the mature protein (Kumazaki et al., 2014b, 2014a; Ravaud et al., 2008). The YidC N-terminal domain is substantially larger than that of Alb3, although up to 92% can be deleted without impairing protein activity (Jiang et al., 2003). Portions of the N-terminal soluble domain in Alb3 also participate in subcellular targeting.

In plants, Alb3 is required for photoautotrophic growth (Long et al., 1993; Sundberg et al., 1997). Alb3 can facilitate insertion of some substrates by itself but cooperates with chloroplast Sec translocase (cpSec) for others (Hennon et al., 2015; Klostermann et al., 2002; Pasch et al., 2005). Insertion of photosystem I subunits A1 (PsaA) and A2 (PsaB) (Pasch et al., 2005; Schöttler et al., 2011) photosystem II subunits D1 (PsbA) and D2 (PsbD) (Pasch et al., 2005), light harvesting complex proteins (LHCP) (Bals et al., 2010; Klostermann et al., 2002; Li et al., 2017; Moore et al., 2003, 2000), cytochrome b6f (Króliczewski et al., 2017; Trösch et al., 2015), ATP synthase (Benz et al., 2009; Pasch et al., 2005), VIPP1 (Trösch et al., 2015), and CP43 CP43 (Schneider et al., 2014) are all Alb3-dependent. Insertion of at least CP43 involves a second insertion factor, TerC (Schneider et al., 2014). Alb3 also interacts with itself and may be responsible for self-biogenesis, as mutant ALB3 proteins lacking the C-terminal tail containing binding elements for SecY, SRP, and ribosomes are more unstable unless grown in low-light conditions (Pasch et al., 2005; Urbischek et al., 2015). Additionally, protein-protein interactions have been observed between Alb3 and the photosystem II assembly factors LPA3 and LPA2 (Cai et al., 2010). A homolog of Alb3 known as Alb4 in Arabidopsis is responsible for the insertion of a specialized subset of membrane proteins, most notably ATP synthase subunits CF₁β and CF₀II

(Benz et al., 2009; Gerdes et al., 2006). In addition to its role as a membrane integrase protein, a moonlighting role of Alb3 has also been documented. *Pisum* post-floral gene 1 (PPF-1), the major Alb3 homolog in garden pea, has an additional hypothesized role in regulating chloroplast calcium homeostasis (Li et al., 2004; Wang et al., 2003). PPF-1 was first discovered in short-day-grown plants of the 'G2' pea genetic line and is believed to contribute to their never-senescent phenotype (Li et al., 1998; Zhu et al., 1998). Flowering time is correlated to the expression level of PPF-1 and is also directly correlated to both the amount of calcium in isolated chloroplasts and to the timing of a cytoplasmic calcium spike (Li et al., 2004; Wang et al., 2003; Xu et al., 2002). Interestingly, expression of PPF-1 in human Novikoff cell lines, which lack native calcium transport activity with the extracellular media, causes an inward flux of calcium into the transgenic cells (Wang et al., 2003). Despite the significance of calcium to plant cellular signaling, key details about how it is transported and stored in chloroplasts are missing.

Chloroplasts are one of the largest subcellular stores of calcium, containing an average of 15mM Ca^{2+} (Kreimer et al., 1987; Portis and Heldt, 1976) whereas resting calcium levels are 100-150nM in the cytosol and stroma (Loro et al., 2016; Nomura et al., 2012). Nearly all calcium in chloroplasts is thought to be bound to either proteins or lipids in the thylakoid membrane (Nomura and Shiina, 2014). Calcium also accumulates in thylakoids in a light- and ATP-dependent manner (Ettinger et al., 1999) and is released to the stroma in the dark in a non-circadian mechanism (Sai and Johnson, 2002). This transition is greater than 1000-fold, with free stromal Ca^{2+} levels spiking up to 5-10 μM in the dark (Ettinger et al., 1999; Johnson et al., 1995). Despite this drastic phenotype, the identity of the apparent $\text{Ca}^{2+}/\text{H}^{+}$ antiporter has not yet been elucidated definitively (Rocha and Vothknecht, 2012). Recent reviews have proposed PPF-1 to be a putative $\text{Ca}^{2+}/\text{H}^{+}$ antiporter, but connections to the Arabidopsis Alb3 homolog have not been

suggested (Hochmal et al., 2015; Nomura and Shiina, 2014). Alb3 may, therefore, represent this missing link for calcium transport between the stroma and thylakoid lumen.

Several other calcium transporter candidates have been suggested at the thylakoid membrane. POLLUX and CASTOR were previously reported to be chloroplast-localized transporters with permeability to calcium (Imaizumi-Anraku et al., 2004; Weinl et al., 2008), but a later study showed that the localization reported was likely artefactual and that these proteins are nuclear-localized (Charpentier et al., 2008). CAS (Ca²⁺ Sensing receptor), was also thought to localize to thylakoid membranes and hypothesized to be a transporter (Nomura et al., 2008), but is not an actual transporter and instead is central to chloroplast calcium sensing and signaling (Huang et al., 2012; Nomura et al., 2012; Weinl et al., 2008). Several H⁺/Ca²⁺ exchanger (CAX) genes including AtCax1, AtCax3, and AtCax4 have predicted chloroplast targeting, but current evidence shows vacuolar localization for these homologs (Carraretto et al., 2016). CCHA1 (PAM71) is also hypothesized to be a thylakoid calcium transporter, although mutants accumulate higher levels of Ca²⁺ in chloroplasts (Schneider et al., 2016). CCHA1 has a higher affinity for manganese, and thus its calcium transport phenotypes may be the result of leaky affinity for divalent cations; indeed, Mn²⁺ is lower in *cchal* chloroplasts (Schneider et al., 2016). *cchal* plants are also sensitive to high external calcium but are rescued by EGTA, further suggesting that they hyper-accumulate calcium (Wang et al., 2016). However, CCHA1 is also able to rescue yeast mutants deficient in Ca²⁺/H⁺ antiport activity, suggesting at least some bidirectional transport (Wang et al., 2016). It will be interesting to see how CCHA1 plays into chloroplast calcium dynamics and especially if it is responsible for the release of calcium into the stroma, but the available evidence suggests that CCHA1 does not act alone.

A serendipitous insight into the potential link between Alb3 and calcium homeostasis was uncovered when Alb3 homologs from Arabidopsis (AtALB3) and *Malus × domestica* (MdALB3.1) were expressed from the chloroplast genome of *Nicotiana tabacum*. The transplastomic plants exhibited altered chlorophyll fluorescence, changes in development, and unusual chloroplast ultrastructure with poorly stacked and inflated thylakoids, however, the phenotype was much more severe in case of MdALB3.1-expressing chloroplast transgenic plants. Sequencing of the chloroplast transgenic region revealed that while the AtALB3 insertion had the expected sequence, a chance insertion of an *E. coli* transposon created a nonsense mutation in MdALB3.1, truncating it into a short 119-amino acid segment comprising only of the soluble N-terminal domain. This observation prompted further investigation into how a domain with no previously described function could result in a phenotype at all, let alone one more severe than expressing the full protein. Working with the hypothesis that the N-terminus creates transient membrane permeability during translocation by ALB3/SecY heterodimers, a series of truncations for this domain were created and transplastomic plants were generated with each. These truncations were analyzed to determine which sequence motifs were responsible for plastid localization, and when integrated into the chloroplast genome, for the chloroplast ultrastructural and physiological phenotypes. Several lines of evidence point to a physiological role of the second half of the N-terminal soluble domain, while preliminary evidence from *in vivo* aequorin kinetics and bulk tissue ICP-MS show that calcium may play a role in this phenotype. These results add further support to the hypothesized link between PPF-1/ALB3 and regulation of thylakoid ion homeostasis.

Results and Discussion

Identification of Conserved Structural Domains of Alb3

Crystal structures available for the YidC homologs from *E. coli* (EcYidC) and *B. halodurans* (BhYidC), which Alb3 is known to be able to functionally complement *in vivo* (Jiang et al., 2002). However, the sequence and function of all homologs within the YidC/Oxa1/Alb3 superfamily are highly conserved, and structure is likely to be similar. Multiple sequence alignment of Alb3 homologs from *A. thaliana* (AtALB3), *Malus × domestica* (MdALB3.1), and *N. tabacum* (NtALB3) was conducted against the *E. coli* YidC (EcYidC) sequence as a reference (Figure 1). Using the crystal structures for EcYidC and BhYidC as references, secondary structural elements were highlighted. A high level of sequence conservation was found between YidC and Alb3 throughout the central integrase domains of the protein. In contrast, the soluble N- and C-termini were poorly conserved with a few notable exceptions. Unlike both Alb3 and Oxa1, EcYidC has a very large N-terminal tail localized to the periplasm (Ravaud et al., 2008), so it is unsurprising that the Alb3 homologs have large deletions in the alignment. However, significant homology is observed for the region immediately upstream of the first transmembrane helix (yellow highlighted region, Figure 1). This motif has been described to form an amphipathic, soluble helix (Kumazaki et al., 2014a, 2014b). The C-terminus of YidC is much shorter than that of Alb3, but homology was found for a highly positively-charged motif after a 31-residue insertion in the Alb3 homologs. This motif is likely to be a ribosomal binding site, as described in both Oxa1 and YidC (Haque et al., 2010a; Jia et al., 2003; Kedrov et al., 2013; Kohler et al., 2009; Palmer et al., 2012). Interestingly, the R366 residue of EcYidC which is highly conserved among prokaryotic YidC orthologs is replaced by lysine in all three Alb3 sequences. Replacement of this arginine by lysine decreases integrase activity in BhYidC but is

otherwise tolerated (Kumazaki et al., 2014a). R/K366 is strictly required and conserved in all Alb3/YidC/Oxa1 homologs and possibly helps escort acidic residues that are commonly found in translocated soluble loops (Chen et al., 2014).

Alb3 Multiple Sequence Alignment and Topology

A total of 20 proteins homologous to AtALB3 (NP_001189626.1; AAM64642.1) were identified using BLAST, and the sequences were aligned using T-COFFEE to determine sequence conservation of the Alb3 N-terminus (Figure 2). Alignment of these Alb3 orthologs revealed four relatively conserved regions in the N-terminal soluble domain. The first region, corresponding to residues 1-25 of AtALB3, is highly enriched in serines, small nonpolar residues, and prolines as is typical for most chloroplast transit peptides (Bruce, 2000; Christian et al., 2019, unpublished; Zybailov et al., 2008). The second region, corresponding to residues 38-63 of AtALB3, is highly positively-charged, and terminates in a nearly-invariant “IPP” motif that is possibly involved in cleavage and processing by signal peptide peptidase (SPP). A third region corresponding to residues 69-97 of AtALB3, contains several strictly conserved acidic residues, a highly conserved arginine, and a nearly-invariant “LLYTLADAAVA” motif. Finally, the fourth conserved region, corresponding to residues 107-138 of AtALB3, contains three conserved basic residues, three conserved acidic residues, a conserved histidine, and four conserved aromatic residues. The central portion of this domain is poor in glycine and contains no prolines but has an abundance of the nonpolar aliphatic residues isoleucine, leucine, and valine. This region aligned closely to the periplasmic helix of YidC and contains five invariant positions at the core of this proposed helix, so it is likely that this region serves a similar role in Alb3. The mixture of polar and aromatic amino acids and scarcity of helix-breaking residues suggests that this region also forms an amphipathic helix (LH1) in that inserts parallel to the

thylakoid membrane. The amphipathic helix in YidC is thought to insert partially into the membrane and lie parallel with it, helping in overall folding and helix packing (Kumazaki et al., 2014b, 2014a).

Using the alignment against the known crystal structure for YidC (Figure 1) and conserved features present in all Alb3 sequences (Figure 2), a revised model of the Alb3 topology and domain charges is proposed Figure 3A. Transmembrane helices 1 and 3 are longer and likely cross the membrane at an angle as in YidC, while transmembrane helices 2, 4, and 5 are much shorter and either cross the membrane at a perpendicular angle or do not fully traverse the membrane.

Impact of N-terminal Sub-Domains on Subcellular Localization

ALB3 in Arabidopsis and many other species is undisputedly chloroplast-localized (Gerdes et al., 2006; Sundberg et al., 1997). *Malus × domestica* contains two ALB3 homologs (MdALB3.1 and MdALB3.2) which are highly similar but differ at several positions in the N- and C-termini which could change their subcellular localization (Figure 2 and Additional File 1). Therefore, localization of both *Malus* homologs and NtALB3 homolog was tested using confocal microscopy of transiently-expressed constructs introduced via Agrobacterium infiltration in *Nicotiana benthamiana*. In contrast to the clearly cytoplasmic localization of the smGFP without N-terminal modification (Figure 4A), all three ALB3 homologs were localized to the plastid based on an overlay of the GFP signal with chlorophyll autofluorescence (Figures 4B-D).

After plastid-localization of the full MdALB3.1 protein was confirmed, four stepwise truncation mutants of the MdAlb3.1 gene were constructed based on the sites indicated in Figure 2 to test what effect each conserved motif had on protein translocation. The first truncation, comprising residues 1-22 of MdAlb3, corresponds to the first homologous motif; this truncation

was designated to be MdALB3 Δ N(1-1). The second, MdALB3 Δ N(1-2), consists of residues 1-49. The third motif including the hypothesized thylakoid transfer domain is comprised of residues 1-82 and is termed MdALB3 Δ N(1-3). Finally, the fourth motif including residues 1-117 and the possible amphipathic luminal helix, is called MdALB3 Δ N(1-4). The first homologous region (MdALB3 Δ N-(1-1), was found to be insufficient to drive localization of GFP into the chloroplast, as evidenced by a diffuse signal in the cytoplasm, shown in Figure 5A. However, each of the longer truncation constructs was sufficient to drive plastid localization of GFP, indicated in Figure 5B-D. These results are consistent with the amino acid content: the first conserved region is enriched in serine/threonine and uncharged amino acids as typical for transit peptides, while the second domain contains several conserved basic residues and is preceded by an abundance of prolines, which are necessary for interaction with TOC GTPases and transit peptide flexibility, respectively.

Although the MdALB3 Δ N(1-3) and (1-4) truncations are expected to be localized to the lumen, the corresponding micrographs in Figure 4 are stroma-localized. The most likely case is that GFP folds too rapidly in the cytoplasm/stroma, and cannot be transported by either cpSec or ALB3 (Gray and Henderson-Frost, 2011). TAT is the only translocase capable of transporting folded smGFP, but preALB3 lacks the twin-arginine motif that is required for TAT translocation (Marques et al., 2004; Thomas et al., 2001). In contrast, the full ALB3 proteins in Figure 3 exhibit thylakoid localization. As the ALB3 C-terminal domain is stroma-localized, GFP in these constructs does not need to cross the membrane and thus does not disrupt localization. Nevertheless, the third homologous region in ALB-N contains a highly invariant alanine-rich region that is likely the cleavage site for plastid-localized signal peptidase PlsP1 (e.g., TPP), which favors alanine at -1 and -3 positions (Paetzel et al., 2002; Shipman-Roston et al., 2010).

Chlorophyll Fluorescence Analysis of Chloroplast-Transformed Lines

Chloroplast transgenic lines of *Nicotiana tabacum* ‘Petit Havana’ were generated resulting in full-length AtALB3, transposon-truncated MdALB3 Δ N(tr), MdALB3 Δ N(1-1), MdALB3 Δ N(1-3), and MdALB3 Δ N(1-4) to test the hypothesis that only certain motifs are responsible for the phenotypic and physiological effects of the ALB3 N-terminal domain. The chloroplast does not have the same transcriptional and translational regulation as the nucleus, resulting in higher protein accumulation, and therefore leads to more pronounced phenotypes than are possible with nuclear overexpression (Clarke and Daniell, 2011). Additionally, expression of ALB3 from the chloroplast genome allows the study of the protein in its ancient evolutionary context. Chloroplast transgenic and wildtype plants were grown to the mid-juvenile stage (approximately 50 days post-germination) and analyzed for the chlorophyll fluorescence phenotype at the Washington State University Phenomics Core. The results of this analysis are summarized in Figure 6, and representative false-color images are presented in Figure 7. Across all metrics, the fluorescence phenotype of MdALB3 Δ N(1-1) plants was identical to wildtype and empty vector lines, indicating that expression of an incomplete transit peptide alone does not affect photosynthetic efficiency. In contrast, the AtALB3 expressing lines and all of the remaining MdALB3 Δ N lines had altered phenomics values across most of the parameters. Wildtype and empty vector plants exhibited a maximum quantum yield (Fv/Fm) of 0.8, which is typical for non-stressed plants (Björkman and Demmig, 1987). However, MdALB3 Δ N(1-3) and MdALB3 Δ N(1-4) exhibited a 25% lower maximum quantum yield than wildtype, while lines expressing the transposon-truncated MdALB3 (MdALB3 Δ N(tr)) were 37.5% lower.

Minimum fluorescence, Fo, was consistent in wildtype at about 40-50. MdALB3 Δ N(1-3) and MdALB3 Δ N(1-4) were significantly higher, ranging between 70-90. MdALB3 Δ N(tr) were

even higher; both lines averaging above 100. NPQ in all wildtype-like lines was consistently between 1.0-1.2, but MdALB3 Δ N(1-3), MdALB3 Δ N(1-4), MdALB Δ N(tr), and AtALB3 lines were higher, between 1.4-1.5. The energy-dependent quenching component qE was extremely variable, but wildtype, empty vector, and MdALB3 Δ N(1-1) lines were generally stable at a value of 1.2. MdALB3 Δ N(1-3), MdALB3 Δ N(1-4), and AtALB3 lines trended lower at between 0.9-1.2 but were generally not statistically different. MdALB Δ N(tr) had less than half the qE of WT (0.55-0.59), and this difference was highly significant. Finally, a significant factor distinguishing AtALB3 lines from the severe MdALB3 Δ N lines was PhiII, defined as the operating efficiency of PSII. Wildtype and AtALB3 lines averaged 0.35, but MdALB3 Δ N(1-3) and MdALB3 Δ N(1-4) were much lower at between 0.25-0.30, and MdALB3 Δ N(tr) lines were consistently at 0.20. Interestingly, plants overexpressing the AtALB3 protein did not match the pattern observed in the MdALB3 Δ N lines. Notably, quantum yield (Fv/Fm) and minimal fluorescence (Fo) were closer to wildtype levels, and the fraction of open PSII centers (PhiII) was statistically indistinguishable from wildtype. In contrast, MdALB3 Δ N lines, and especially the MdALB3 Δ N(tr) lines, had significantly altered fluorescence phenotypes for those parameters as well as for NPQ, qI, and qL. These differences are unusual, as plants for all AtALB3 lines were generally weaker and grew more slowly than any of the MdALB3 Δ N lines. The energy-dependent quenching component qE was also lower in the severely-affected MdALB3 Δ N(tr) lines but is not strongly affected in other lines. qE is dependent on the thylakoid Δ pH gradient, suggesting that ALB3-transgenic lines have a lower Δ pH gradient, and thus either photosynthesis is defective, or proton leakage is responsible for the change in pH.

Chloroplast-Transgenic Lines Display Delayed Senescence

PPF1/Alb3 overexpression has been shown to cause delayed senescence in divergent plant taxa (Li et al., 2007; Wang et al., 2003, 2008). Chloroplast transgenic lines in this study exhibited similar delays in senescence and were taller on average upon first flowering (Figure 8). After completion of the phenomics experiments in this study, most plants with wildtype levels of chlorophyll fluorescence grew larger and flowered earlier, whereas lines expressing AtALB3 and MdALB3 Δ N(tr) were delayed in flowering by an average of 10 days, but the flower spikes were also somewhat taller when the plants did flower (Figure 8A,B). MdALB3 Δ N(1-3) and MdALB3 Δ N(1-4) were somewhat intermediate. Graphing these flowering height and days to flowering together revealed that the phenotypes of individual plants cluster into distinct groups: wildtype and MdALB3 Δ N(1-1) group together with a faster, shorter-flowering phenotype that normal senescence (Figure 8C,D). AtALB3 and MdALB3 Δ N(tr) group together in a slower, taller-flowering phenotype, and the MdALB3 Δ N(1-3) and MdALB3 Δ N(1-4) lines formed an intermediate group. Though delayed senescence can be partly explained by lower photosynthetic efficiency, the increase in flowering height suggests that the plants still accumulate more carbon by the time of flowering, potentially by extending the growing season. In plants that have a determinate flowering habit and experience leaf senescence upon fruit or seed maturation, an extension of the growing season could result in significant improvements to yield, particularly in legumes and cereals (Li et al., 2007; Wang et al., 2003, 2008).

Chlorophyll Content

ALB3 overexpression could elicit physiological effects due to disruption of the translocase activity of the native protein. This null hypothesis was first tested by measuring chlorophyll content of leaf tissue in wildtype and transgenic plants. Chlorophyll content is

reduced in *alb3* mutants (Sundberg et al., 1997; Wang et al., 2008) because Alb3 mediates insertion of chlorophyll binding proteins (Moore et al., 2000; Woolhead et al., 2001) and is also responsible for insertion of core subunits for photosystems I and II (Pasch et al., 2005; Schöttler et al., 2011). Removal of the Alb3 C-terminus also affects chlorophyll levels (Urbischek et al., 2015), indicating that the SRP, FtsY, cpSec, and ribosomal binding domains are crucial for maintaining efficient insertion of chlorophyll-binding proteins. Because ALB3 is involved in the insertion of photosystem subunits and light-harvesting complexes, DMF extractions of chlorophyll were performed to measure chlorophyll-a, chlorophyll-b, total chlorophyll, and the chlorophyll(a/b) ratio of wildtype and transgenic lines to detect if the expression of Alb3 variants altered chlorophyll accumulation (Figure 9). No decreases in chlorophyll were observed for most of the MdALBΔN truncations; however AtALB3 had lower chlorophyll-a, -b, and total chlorophyll when measured by $\mu\text{g}/\text{cm}^2$. For MdALB3, only MdALB3ΔN(tr)b, and MdALB3ΔN(1-4)b showed at least one altered chlorophyll measurement, and these were only significant at the $p=0.05$ level. Other lines were statistically indistinguishable from wildtype and empty vector lines. Only a few lines had statistically significant differences in the ratio of chlorophylls(a/b). All AtALB3 lines had a higher ratio, at about 2.0 compared to the typical 1.80 for wildtype and empty vector lines. Of these, only AtALB3b and c were statistically significant, with the other lines having too much variation to be significant. Higher ChlA:ChlB ratios in the AtALB3 plants strongly suggests that ALB3-dependent integration of the light-harvesting complexes is impaired in those lines.

Chloroplast Ultrastructural Changes

As a membrane integrase, Alb3 is responsible for the biogenesis of thylakoid membrane proteins and therefore can influence chloroplast ultrastructure. Furthermore, if Alb3 impacts

chloroplast ionomics, the change in osmolarity of the stroma or thylakoid lumen could disrupt membrane organization. The chloroplast ultrastructure of MdALB3 Δ N(tr) and AtALB3 plants was examined. In contrast to the well-stacked thylakoid grana with tightly-appressed membranes in wildtype, both MdALB3 Δ N(tr) and AtALB3 had severe but distinct phenotypes (Figure 10). In MdALB3 Δ N(tr) plants, thylakoid membranes were tightly-appressed and stacked to similar depths as wildtype, but the thylakoid lumens appeared to be inflated and rippled especially when exposed to the stroma, adopting an almost Golgi-like appearance. Furthermore, some isolated membranes formed circular vesicle-like structures appearing to be derived from the thylakoids and stromal lamellae rather than plastoglobules. Stromal lamellae were poorly visible, and when present were severely thickened. AtALB3 plastids had extremely elongated grana stacks that were swollen, but not to the same severity as MdALB Δ N(tr). Grana stacks were more organized and seemed similar to wildtype except stromal lamellae, which were the most swollen and had distinctive triangular vesicles (see marked structures in Figure 10).

TEM was also performed on the additional MdALB Δ N truncations to observe at what point a similar phenotype to MdALB3 Δ N(tr) appeared (Figure 11). The chloroplast ultrastructure of MdALB Δ N(1-1) lines was identical to wildtype plants, but in contrast, MdALB Δ N(1-3) and to a lesser degree MdALB Δ N(1-4) exhibited the same type of thylakoid and stromal lamellae inflation as MdALB3 Δ N(tr). Of these, the most severe was MdALB Δ N(1-3)b, which exhibited nearly-identical chloroplast ultrastructure compared with MdALB Δ N(tr). Inflated grana and lamellae appeared in 90% of plastids for the MdALB Δ N(1-3) lines and in 55% of MdALB Δ N(1-4) plastids, but in only 10% of plastids from MdALB Δ N(1-1). Wildtype, by comparison, had no plastids exhibiting abnormal thylakoid structure. A similar phenotype has been observed in literature for mutants of calcium-sensing receptor (CAS) grown with high calcium

supplementation and in high light (Huang et al., 2012). Thylakoid swelling and grana destabilization are associated with osmotic stress, but may also occur as a photoprotective mechanism to promote repair of photosystem II (Kirchhoff, 2013; Zhang et al., 2012; Zhao et al., 2017). Indeed, the microscopy phenotypes somewhat correlate to but do not fully match the photoinhibitory quenching component (qI) observed in the chloroplast transgenic lines (Figure 6). Plastoglobule proliferation is also commonly observed in mutants with defective thylakoid membrane biogenesis, and are correlated with oxidative damage (Bédard et al., 2017; Bréhélin and Kessler, 2008; Takechi et al., 2000). Plastoglobule density for the lines in this study was calculated as the average number of plastoglobuli per μm^2 to test for membrane destabilization and degradation. Plastoglobuli were often highly clustered but averaged out to similar values in each line. The MdALB Δ N(1-3) lines had 2-fold higher plastoglobuli per μm^2 , but this difference was not significant in either line.

Perturbations in Ionic Composition

To test the hypothesis that ALB3 overexpression affects ion homeostasis in chloroplasts and in leaves, bulk leaf tissue was collected from transgenic and wildtype plants and levels of Na^+ , K^+ , Ca^{2+} , and Mg^{2+} were quantified using inductively-coupled plasma mass spectrometry (ICP-MS) (Figure 12). The sodium content of the leaf tissue was largely stable, suggesting that their ion exclusion was normal. Wildtype and empty vector expression plants had Na^+ levels close to 1mM, and almost all assessed lines were between 1-2 mg/g dry weight. Potassium levels exhibited much more variation in the chloroplast transgenic lines. Wildtype and empty vector expressing plants were found to be near 10 mg/g K^+ , while nearly 3-fold higher levels were found in MdALB3 Δ N(tr) plants, and between 2-fold to 5-fold higher in the AtALB3 lines. The MdALB3 Δ N(1-3) lines showed a slightly higher potassium content. Divalent cations were also

significantly affected. Magnesium content was found to be between 3.5-4 mg/g in wildtype and empty vector lines. Slightly higher levels of Mg^{2+} were found in MdALB3 Δ N(1-3) and MdALB3 Δ N(tr), with higher levels found in only one AtALB3 line and neither of the MdALB3 Δ N(1-4) lines. The highest Mg^{2+} concentrations were measured in MdALB3 Δ N(1-3)a, which averaged about 5.5 mg/g. Calcium also showed a general increase in chloroplast transgenic lines. Ca^{2+} levels had a baseline concentration of 10-14 mg/g in wildtype and empty vector lines but reached above 20 mg/g for MdALB3 Δ N(tr)a and AtALB3. MdALB3 Δ N(1-3) and MdALB3 Δ N(1-4) trended slightly higher than wildtype but were not significantly different except in MdALB3 Δ N(1-3)b. The lack of a distinct Na^+ phenotype suggests that transgenic lines are not experiencing a systemic loss of ion homeostasis, but rather a specific perturbation to ions known to be enriched in the chloroplast thylakoid lumen. As the N-terminus of ALB3 translocates to the lumen *in vivo*, the observed increases in K^+ , Ca^{2+} , and Mg^{2+} content of leaf tissue suggests that chloroplasts overexpressing ALB3 are accumulating abnormal levels of these ions, leading to higher overall levels in the leaf.

Aequorin

Chloroplast-transformed lines were pollinated with an *N. tabacum* line with nuclear-expressed stromal-targeted aequorin. Seeds of double-transformed lines were germinated on media containing low $CaCl_2$ (100 μ M) and a Ca^{2+} response was stimulated by exogenous application of NaCl to a final concentration of 200mM (Figure 13). Both of the MdALB3 Δ N(1-3) lines displayed aberrant Ca^{2+} signatures. MdALB3 Δ N(1-3)-a showed a nearly 100% increase in peak height, increasing from average measurement height of 268nM to 384nM whereas baseline was 150nM in both lines. MdALB3 Δ N(1-3)-b was slightly elevated at 312nM, but this was not significant. However, its baseline was higher at 196nM. This observation agrees with the

ionomics data, which demonstrated line ‘b’ to have significantly higher total Ca^{2+} in comparison to wildtype, while line ‘a’ was higher, but not significantly so. Why the stimulation peak heights differ is not clear, but as calcium response phenotype appears to be extremely dependent on the specific line, there may be differences in the developmental stage or initial growth rate. The MdALB3 Δ N(1-4) lines did not have a significantly higher baseline or peak Ca^{2+} levels, although this result is consistent with what was observed in TEM.

MdALB3 Δ N Phenotype is Linked to the Thylakoid-Transfer Domain

The overexpression constructs could also affect the folding and activity of mature NtALB3. The only known portion of the ALB3 N-terminus likely to be part of the mature protein structure is the LH-1 amphipathic helix in the fourth conserved region of Alb3. However, both MdALB3 Δ N(1-3) and MdALB3 Δ N(1-4) had similar phenotypes, suggesting this helix is not responsible for the phenotypes observed in this study. This result is surprising given that the third conserved region is hypothesized to contain the thylakoid-transfer domain of ALB3 and is probably cleaved and degraded *in vivo*. Multiple mechanisms could be at play. First, overexpression of the N-terminal domain could inhibit endogenous ALB3 or SecY translocase activity. If this was the case, the integration of LHCP and core photosystem proteins and the resulting levels of chlorophyll would be expected to be impacted (e.g., Bellafiore, 2002; Urbischek et al., 2015). However, the only changes observed in transgenic lines were found for lines overexpressing the AtALB3 full protein, whereas overexpression of the MdALB3 N-terminus did not significantly affect chlorophyll levels or the ratio of chlorophyll-a to chlorophyll-b (Figure 9). Alteration to chlorophyll in the AtALB3 lines suggests that the full protein impacts translocase activity when overexpressed, while expression of the truncated protein does not. As a second test of whether protein translocation was affected in the

overexpression lines, total soluble protein of leaf tissue was evaluated by SDS-PAGE gel electrophoresis. The concentration of major proteins was not affected relative to total leaf protein (Additional File 1), and thus it is unlikely that the chloroplast transgenes affected folding or abundance of thylakoid membrane proteins. Furthermore, the possibility that overexpression of transgenes from the chloroplast genome disrupted expression of endogenous NtALB3 in the nuclear genome was disproven with qRT-PCR using primers specific to NtAlb3; expression of NtALB3 remains stable in all transgenic lines and equivalent to wildtype expression (Additional File 2).

Another potential explanation is that overexpression of the MdALB3 transit peptide causes cytotoxicity. Transit peptides form toxic aggregates which can destabilize membranes if they are not proteolyzed, thereby disrupting membrane gradients and uncoupling redox enzymes (Kmiec et al., 2014; Zhang and Glaser, 2002). This is deemed unlikely, as lines expressing MdALB3 Δ N(1-1), comprising much of the transit peptide, are indistinguishable from wildtype in all tests. Interestingly, no transgenic events were obtained for MdALB3(1-2) despite several attempts. It was shown that MdALB3(1-2) is the shortest sequence capable of targeting GFP to the chloroplast (Figure 5). The SPP cleavage site may be somewhat downstream of the second domain, in which case overexpressed protein may not be processed correctly and could accumulate to toxic levels. In contrast, the shorter MdALB3 Δ N(1-1) is small enough to be efficiently processed by presequence protease or organellar oligopeptidases without initial processing by SPP (Kmiec et al., 2014; Teixeira et al., 2017). A third potential mechanism is that overexpression of MdALB3 Δ N constructs disrupts thylakoid membrane permeability, causing osmotic or redox stress. The swollen thylakoid phenotype that was observed in TEM is characteristic of osmotic stress in the thylakoids (Kirchhoff, 2013; Zhang et al., 2012). Whole-

tissue ICP-MS uncovered moderately increased levels of K^+ in MdALB3 Δ N(tr) lines, and slightly increased levels in the MdALB3 Δ N(1-3) and MdALB3 Δ N(1-4) lines (Figure 12). K^+ is thought to provide a counterbalance for H^+ ions in the thylakoid lumen by dissipating trans-thylakoid membrane potential and balancing osmotic stress (Checchetto et al., 2016; Kunz et al., 2014), but the energy-dependent quenching component qE was lower in MdALB3 Δ N lines, suggesting a decreased Δ pH gradient, while the photoinhibitory component qI was much higher. Furthermore, Na^+ levels were not significantly different in most of the transgenic lines, suggesting that Na^+ -inducible K^+ transporters such as CHX17, KUP6, KUP11, KOR, and AKT2/3 are not responsible for the increase in leaf K^+ content (Adams and Shin, 2014; Cellier et al., 2004; Kunz et al., 2014; Maathuis, 2013). Though the current data support the hypothesis that the Δ pH is lower in lines expressing MdALB3 Δ N, such an observation could also be due to loss of membrane integrity. Confirming the Δ pH change with electrochromic shift measurements would more definitely link MdALB3 Δ N with disruption membrane permeability, while immunodetection of soluble lumenal proteins such as plastocyanin could determine whether competition at the Sec translocase is indirectly impacting protein abundance and electron transport capacity.

Based on the available evidence, the Alb3 Δ N phenotype is hypothesized to be caused by disruption to thylakoid ion regulation in a mechanism dependent on thylakoid localization, rather than on disruption to native translocase activity or interaction with ion channels. If this is the case, the more severe phenotype of MdALB3 Δ N(tr) compared to MdALB3 Δ N(1-4) is curious because the protein coding sequences are identical. qRT-PCR analysis using primers specific for MdALB3 was conducted, and it was found that MdALB3 Δ N(tr) displayed much higher expression levels than either MdALB3 Δ N(1-3) or MdALB3 Δ N(1-4), likely due to increased

transcript stability (Additional File 2). These results indicate that the mechanism is likely the same between these lines, but MdALB3 Δ N(tr) has an exacerbated phenotype due to its higher expression.

Chloroplast transformations with the complete MdALB3 protein were attempted but no successful transgenic events were obtained. Additionally, several chloroplast-transgenic lines expressing the full NtALB3 protein were generated, but all were extremely weak, had a bleached phenotype in moderate light, required an external carbon source, and failed to set seed. These results could indicate problems with overcrowding of the thylakoid membrane space or disruption of endogenous binding partners that could have prevented the capture of stable transplastomic plants. For instance, overly abundant ALB3 could lead to increased binding to ribosomes, SecY, and components of the SRP pathway, which could disrupt native translocase activity enough to cause lethality. Interestingly, expression of the full AtALB3 protein was successful, demonstrating that this is likely not the case. Furthermore, it was possible to successfully express both MdALB3 and NtALB3 if the C-terminal domains were truncated, but none of the plants thus generated exhibited significant phenotypes (Additional File 1). Of the three proteins, NtAlb3 and MdAlb3 are more closely related to each other (~72% sequence identity) than either is to AtALB3 (~63% sequence identity), and AtALB3 is more unique at the N- and C-termini (Additional File 1). Because of the higher homology between MdAlb3 and NtAlb3 proteins, both may interact strongly with endogenous proteins, which could cause toxicity when overexpressed. In contrast, the binding sites at the C-terminus of AtALB3 may be different enough to have weaker interaction with endogenous proteins, but its overexpression still causes significant phenotypes. While speculative at this point, it would be interesting to test

this hypothesis by constructing chimeras of either MdALB3 or NtALB3 with the C-terminus of AtALB3.

Alb3 and SecY likely form an ion-permeable channel in the thylakoid membrane

The data suggest that osmotic stress and membrane destabilization is occurring in MdALB3 Δ N chloroplast transgenic lines and that this phenotype is dependent on the putative thylakoid transfer domain of ALB3. This implicates that translocation at the thylakoid membrane, rather than any functional or structural role in the MdALB3 N-terminus, is responsible for the observed phenotypes. Preventing ion leakage at the translocases is crucial, especially for energy-generating membranes such as the thylakoid. Translocases have pores large enough to permit amino acids, causing smaller ions to rapidly diffuse without compensatory mechanisms. The functional channel of the Twin Arginine Translocase (TAT) only oligomerizes transiently to transport its substrates, thus preventing constitutive permeability to ions (reviewed in Frain et al., 2016). SecY is arranged in two lobes which surround the central pore, but two cooperative countermeasures prevent major ion permeability: a ring of hydrophobic residues on the interior surface of the channel acts as a gasket to seal against ion leakage during translocation, while a short helical segment of SecY plugs the channel while it is inactive (Park and Rapoport, 2011; Van den Berg et al., 2004). While Alb3 lacks an intrinsic channel, Alb3, Oxa1, and YidC have all been observed to be capable of dimerization upon interaction with ribosomes and could be ion-permeable in this state (Dünschede et al., 2011; Kohler et al., 2009).

Interestingly, Oxa1 in mitochondria has voltage-gated potassium transport activity which is stimulated by a substrate peptide, preCox2 (Krüger et al., 2012). In bacteria, a heterotetrameric channel of SecY and YidC has voltage-gated ion channel activity caused by YidC reorienting in a hinge-like mechanism to allow transmembrane domains to escape the SecY pore (Figure 3B)

(Sachelaru et al., 2017, 2013). Alb3 is capable of functionally complementing YidC (Jiang et al., 2002) and the cpSecE component of cpSec translocase is capable of functionally complementing EcSecE (Fröderberg et al., 2001), so it is likely that this voltage-gated channel activity in chloroplasts is similar to bacteria. Several observations were made that lead to a hypothesis that preALB3 is a substrate of the Alb3/SecY heterodimer and generates the MdALB3 Δ N phenotype by the constitutive opening of the cpSecY/Alb3 heteromeric channel.

First, the bacterial homolog YidC is dependent on a heteromeric YidC/SecAYEG translocon for its biogenesis, and its high homology to ALB3 suggests that a similar mechanism could lead to ALB3 biogenesis (Koch et al., 2002; Urbanus et al., 2002). Secondly, most Alb3-independent proteins have N_{in}-C_{in} topology, and proteins with odd numbers of transmembrane domains tend to have ALB3-dependency (Woolhead et al., 2001). ALB3 has five transmembrane domains, and thus fits the overall trend for ALB3-dependent insertion. Length of soluble translocated domains has also been suggested as a determinant of YidC/Alb3-dependency, as preproteins with soluble translocated domains greater than 100 residues are less sensitive to YidC depletion in *E. coli* (Wickström et al., 2011). After removal of the primary transit peptide in the stroma, all translocated segments are shorter than 100 amino acids. Insertion of charged residues, especially positively charged residues, in translocated domains requires YidC-SecY heterodimers (Gray and Henderson-Frost, 2011; Zhu et al., 2013). Two of the soluble luminal domains of Alb3 are highly charged, including the N-terminal tail with an overall charge of -5 not including the cleaved stromal peptide, and luminal loop 1, which has an overall charge of -1 (Figure 3A). The luminal N-terminus also contains ten basic residues, while luminal loop 1 has one basic residue. Finally, charged residues in transmembrane helices (TMH's) require YidC-SecY heterodimers in bacteria (Price and Driessen, 2010; Zhu et al., 2013). TMH1, THM3, and

THM4 of ALB3 all have conserved charged residues, including the essential lysine/arginine in TMH1. It is important to note that all soluble loops and transmembrane domains had conserved charge between AtALB3 and MdALB3 sequences. However, a minor difference was noted in LH-1, which had an overall charge of -1 in MdALB3 but was neutral in AtALB3. Based on Alb3's fulfillment of all these criteria, it is most likely inserted by a heteromeric Alb3/SecY channel and thus can cause ion-permeability of the thylakoid membrane during translocation consistent with the mechanism observed previously (Sachelaru et al., 2017). Interaction with an overexpressed substrate such as preALB3 could increase the frequency and duration of translocon opening, causing somewhat constitutive membrane permeability. In turn, increased permeability of the thylakoid membrane may allow for unrestricted movement of ions including calcium, which would be attracted to neutralize negative charges at the luminal face of the bilayer and in luminal proteins (Anderson et al., 2008). Outwardly, this mechanism may appear to be an active antiport mechanism, thus explaining the apparent $\text{Ca}^{2+}/\text{H}^{+}$ antiport activity found in thylakoid membranes (Enz et al., 1993; Ettinger et al., 1999; Sai and Johnson, 2002). Sec-dependent translocation is driven by the ATPase activity of SecA (Andersson and von Heijne, 1993), and thus membrane permeability would be tied to ATP consumption even without ALB3 performing as a true antiporter. It was also observed that the qI component of NPQ was substantially lower than wildtype, indicating a higher lumen pH. This observation suggests that proton leakage through empty ALB3/cpSec heterodimers could also contribute to the observed $\text{Ca}^{2+}/\text{H}^{+}$ antiport activity.

Conclusions and Future Directions

The role of PPF1/ALB3 in calcium homeostasis has been previously documented in the literature but has lacked a proposed mechanism. ALB3 is homologous to PPF1, so chloroplast transgenic overexpression lines of MdALB3 and AtALB3 were generated to determine if calcium-dependent phenotypes were observed. Chloroplast transgenic lines expressing full-length AtALB3 were generated, but MdALB3 was serendipitously truncated by a transposon, eliminating the conserved integrase domains yet generating a more severe phenotype. As the endogenous NtALB3 is unaffected by this event, it was determined that an unidentified physiological role of the soluble N-terminus containing the transit peptide, thylakoid transfer domain, and an amphipathic luminal helix must be responsible for the unique phenotype. A series of stepwise truncation mutants representing four regions of conserved homology in the ALB3 N-terminus of N terminus were expressed via the chloroplast genome to characterize this further. It was found that regions 1 and 2 comprised the stromal transit peptide sequence, while regions 3 and 4 resulted in a similar phenotype to the transposon-truncated MdALB3. Using ICP-MS and stromal-targeted aequorin, it was shown that calcium levels are impacted in these plants, linking calcium homeostasis to thylakoid translocation activity in general. These results suggest that the observed phenotype is caused by the thylakoid transfer domain and is thus tied to the translocation of the overexpressed N-terminus fragment into the thylakoid lumen, rather than to folding or regulation of mature ALB3. Translocation of soluble proteins or loop regions of transmembrane proteins creates transient ion permeability, and constitutive expression of a translocated protein may cause membrane depolarization as a result. To test the hypothesis that a SecY/Alb3 heterodimer is responsible for the phenotypes observed in MdALB3 Δ N lines, subsequent work could test additional presequences containing thylakoid transfer domains and

specificity for Alb3, SecY, or SecY/Alb3. Additionally, the use of improved aequorin constructs, calcium-sensitive dyes, or ionomic analysis of isolated chloroplasts would greatly improve the hypothesized role of Alb3 in calcium homeostasis.

Further research is necessary to resolve the lack of viable plants expressing full-length MdALB3 and NtALB3, as well as the second truncated N-terminal domain of MdALB3. It is possible that chloroplast transformation with these fragments proved lethal. One way would be to ectopically express these from the nuclear genome or under control of an inducible promoter to obtain viable plants. Additionally, it would be beneficial to generate knockdown or knockout mutants of the endogenous NtALB3 to determine if heterologous expression from the chloroplast can functionally complement the loss of the nuclear copy.

Materials and methods

Multiple sequence alignment

Sequences used for the multiple sequence alignments included EcYidC (NC_000913.3) and ALB3 sequences from *Arabidopsis thaliana* (NP_001189626.1, *Brassica oleracea* (CDX77209.1), *Glycine max* (XP_003537999.1, XP_003539682.1), *Medicago truncatula* (KEH30130.1), *Pisum sativum* (Q9FY06.2), *Fragaria vesca* (XP_004294570.1), *Prunus persica*, *Malus x domestica* (XP_008357463.1, XP_008385136.1), *Cucumis sativus* (XP_004150182.1), *Cucumis melo* (XP_008457492.1), *Solanum lycopersicum* (XP_004250966.1), *Citrus trifoliata* (ABU75304.1), *Populus trichocarpa* (XP_002313443.2, XP_002298365.2), *Theobroma cacao* (XP_007016461.1), *Vitis vinifera* (XP_002284077.1), *Nicotiana tabacum* (XP_016481228.1), and *Ricinus communis* (XP_002531362.1). All sequenced PCR products for AtALB3 showed an insertion compared to the inferred mRNA sequence for locus NP_001189626, and the translated

protein sequence for the cloned cDNA instead matched the alternative protein product AAM64642.1. Sequences were aligned in T-Coffee Version_11.00.d625267 (2016-01-11 15:25:41 – Revision d625267 - Build 507 (Di Tommaso et al., 2011; Notredame et al., 2000) and visualized with Boxshade 3.21 (available online at embnet.vital-it.ch/software/BOX_form.html)

GFP fusion localization

MdALB3 truncations and full-length MdALB3.1, MdALB3.2, AtALB3, and NtALB3 were cloned from cDNA from the appropriate species, incorporating *NcoI* and *ApaI* restriction enzyme sites into the primer 5' termini (see Additional File 1 for primer sequences). PCR products were digested with these enzymes (New England Biolabs, Ipswich, MA) and ligated into 'pRWC3' containing an in-frame C-terminal soluble modified GFP (Davis and Vierstra, 1998), a 2x35S-CaMV promoter, and a nopaline synthase terminator. For MdALB3 and MdALB3 Δ N(1-4), a hetero-stagger technique (Liu, 1996; Xie and Xie, 2011) was used in an enzyme-free cloning strategy to avoid internal restriction sites and enable the use of the same restriction sites. Briefly, two complementary PCR's with staggered 5' and 3' ends were conducted, then mixed, heat-denatured, and gradually cooled to generate a mixture of products. Only 25% of annealed products with compatible cohesive ends to the restriction sites used in the plasmid backbone ligate with high efficiency.

Sequence of cloned fragments was confirmed with Sanger sequencing using M13 primers (SMS_103 and SMS_104). Plasmids lacking mutations were digested with *HindIII* (New England Biolabs) and inserted into *HindIII*-digested pCambia2300 binary vector treated with FastAP thermosensitive alkaline phosphatase (Thermo Fisher Scientific, Waltham, MA). An unmodified smGFP was also cloned to serve as a localization control. Each vector was then transformed into *Agrobacterium tumefaciens* GV3101-PMP90 using electroporation at 2500V,

25 μ F, 200 Ω with a GenePulser Xcell™ (Bio-Rad Laboratories, Hercules, CA) and 1mm Gene Pulser®/MicroPulser™ cuvettes. Electroporated cells were incubated for 3 hours in LB medium without selectable markers before plating on YEP (10g Bacto Peptone, 5g NaCl, 10g Bacto Yeast Extract, 15g Bacto Agar, pH 7.0) media containing 25 mg/L rifampicin, 50 mg/L gentamycin, and 100 mg/L kanamycin. Colonies were verified by PCR and bacterial cultures were infiltrated into *Nicotiana benthamiana* leaves via the protocol of Li, 2011. Positive colonies were grown overnight in 5 mL of YEP containing 25 mg/L rifampicin and 50 mg/L kanamycin. Bacteria were spun down and resuspended in 10 mM MgCl₂ to a final O.D.₆₀₀ of 0.2. Acetosyringone was added to bacterial solutions to a final concentration of 100 μ M and left to incubate at room temperature for 2 hours. *Nicotiana tabacum* ‘Petit Havana’ leaves were infiltrated with *Agrobacteria*. After 48 hours, fluorescence and GFP localization was inspected by confocal microscopy. Full protein chimeras of MdALB3.1, MdALB3.2, and AtALB3 were imaged using a Zeiss Confocal LSM 510 Meta Laser Scanning Microscope, while imaging of MdALB3 Δ N lines was done using a Leica TCS SP-8 X Confocal Laser Scanning Microscope.

Chloroplast transformation

MdALB3 truncations and full-length AtALB3 were cloned from cDNA prepared from *Malus \times domestica* and *Arabidopsis thaliana* tissue, respectively, and restriction sites for *NcoI* and *XbaI* were cloned into the 5' ends. NtALB3, MdALB3.1, MdALB3.2, AtALB3, and MdALB3 truncations were initially cloned into the shuttle vector pADct50a to incorporate the psbA promoter and psbA 5'UTR. As with GFP constructs, hetero-stagger techniques were used for MdALB3 and MdALB3 Δ N(1-4), while restriction enzyme cloning was used for the other constructs. The sequence of the cloned fragments was confirmed with Sanger sequencing (Eurofins MWG Operon, Eurofins Genomics, Louisville, Kentucky) using M13 forward and

reverse primers (SMS_103, SMS_104). Sequences were aligned using SeqManPro (DNASTAR, Madison, Wisconsin) to detect any mutations. Correct cloned fragments were digested with *KpnI* and *XbaI* and cloned into the chloroplast expression vector pADct44 containing *psbA* 3'UTR, spectinomycin resistance marker *aadA* (Prrn promoter, *TrbcL* 3'UTR), and right and left flanking regions homologous to the inverted repeat region of the tobacco chloroplast genome between *trnV* and *rrn23*. Chloroplast transformation was performed using biolistic bombardment of leaf tissue as described previously (Dhingra et al., 2004) with a PDS-1000/He™ system (Bio-Rad, Hercules, California). Briefly, leaves from *Nicotiana tabacum* 'Petit Havana' were bombarded with each construct using particle bombardment and subsequently transferred to filter paper placed on RMOP media. After incubation in the dark for 48 hours, leaves were cut into small ~1.5 cm sections and placed on RMOP (MS basal media plus 1.5 mg/L 6-Benzylaminopurine (BAP) and 0.05mg/L 1-Naphthaleneacetic acid (NAA)) media containing 500mg/L of spectinomycin. Primary transgenic events were excised from the original leaf and maintained on selection. Leaves from these were excised for a second round of regeneration on selective media from each line to obtain homoplasmic lines. Transgene presence was confirmed using ADctP36 and ADctP39 to detect chloroplast integration at the site-specific locus, and gene-specific primers for sequencing. Tissue culture ingredients and antibiotics were purchased from Phytotech Laboratories, Shawnee Mission, KS

Phenomics

Wildtype and transplastomic *N. tabacum* were germinated on the respective selection antibiotic media and transferred to 4" pots in the soil after 30 days. Plants were grown at greenhouse conditions (16hr daylength, 25°C day/20°C night) for 21 days, then transferred to the Phenomics facility for three days of acclimation. Phenomics measurements were recorded for

one week by a FluorCam XYZ outfitted on a movable scaffold to measure fluorescence at multiple points throughout the growth chamber. Excitation was performed using 455nm and 618nm LEDs, and fluorescence was captured by a FluorCam 2701 LU camera with a fluorescence filter. Chamber conditions were kept at 12-hour daylength with 200 μ mol/m²/s illumination, with temperature held at 21°C in the dark and 23°C in the light. Measurements were performed once daily after at least 5 hours of darkness. Images were analyzed using FluorCam Version 7 (Photon Systems Instruments, Drasov, Czech Republic) and parameters including Fv/Fm, Phi II, qL, NPQ, qI, and qE were calculated. Organization of data and data analysis were performed using custom R and Python scripts.

Physiology: Time to Flowering and Chlorophyll Content

After completion of phenomics, plants were moved back to greenhouse conditions. Leaf punches of 1.825cm² were collected from the youngest fully-expanded leaf (typically 3rd position from the apical tip) of 65-day old plants and flash-frozen in liquid nitrogen for chlorophyll analysis and stored at -80°C. Days to flowering and stem height at flowering were recorded for each plant. Bulk leaf tissue was collected and flash frozen in liquid nitrogen. Samples were processed in a 6770 Freezer/Mill® (Spex SamplePrep, Metuchen, New Jersey) and stored at -80°C.

Chlorophyll quantification

N,N-Dimethylformamide (DMF)-extractions of chlorophyll were performed (Porra et al., 1989). Two flash-frozen leaf discs stored at -80°C with a combined area of 3.65cm² were placed in disposable 15 mL Falcon tubes and 5mL of DMF was added. Samples were incubated at 4°C overnight with shaking, then 1 mL of the supernatant was measured in quartz

cuvettes at 647, 664, and 750 nm in a Genesys 20 Spectrophotometer. Chlorophyll contents were calculated as follows:

$$\text{ChlA (nmol/ml)} = 13.43(\text{A664-750}) - 3.47(\text{A647-750})$$

$$\text{ChlB (nmol/ml)} = 22.90(\text{A647-750}) - 5.38(\text{A664-750})$$

$$\text{ChlTotal (nmol/ml)} = 19.43(\text{A647-750}) + 8.05(\text{A664-750})$$

$$\text{ChlA (ug/ml)} = 12.00(\text{A664-750}) - 3.11(\text{A647-750})$$

$$\text{ChlB (ug/ml)} = 20.78(\text{A647-750}) - 4.88(\text{A664-750})$$

$$\text{ChlTotal(ug/ml)} = 12.67(\text{A647-750}) + 7.12(\text{A664-750})$$

Protein Analysis

Total leaf protein was extracted from milled tissue using a modified Laemmli buffer extraction protocol. Briefly, frozen tissue was weighed and 2 volumes of 2x Laemmli buffer (4% SDS, 20% glycerol, 0.125 M Tris HCl pH 6.8 plus 50 μ l/ml 2-mercaptoethanol (5%) and 10 μ l/ml Sigma Plant Protease Inhibitor Cocktail (Product P9599, MilliporeSigma, St. Louis, MO)) were added immediately and vortexed. Samples were incubated at 95°C for 5 minutes, then centrifuged at 13,900 x G for 5 minutes at room temperature. Protein concentrations were measured twice independently using the RC DC™ Protein Assay kit (Product #5000121, Bio-Rad, Hercules, CA) and a Genesys 20 Spectrophotometer (Thermo Fisher Scientific, Waltham, Massachusetts); protein quantity was calculated as the average of these measurements. Protein samples were diluted to 2.5mg/ml and stored at -80°C.

ICP-MS

Ionomics was performed at the Analytical Chemistry lab at Washington State University using inductively-coupled plasma mass spectrometry. Bulk leaf tissue was collected following completion of the chlorophyll phenomics experiments, dried at 60°C for 24 hours, and

pulverized using a mortar and pestle. Samples were treated according to (Grusak, 1994), in which dried and pulverized samples are wet-digested with nitric acid and heated to 150-200°C for 1 hour, or until dry. Digests were resuspended in 0.5mL of 12M HCl, incubated for 1 hour, and volume was adjusted to 10mL with MilliQ H₂O. Samples were analyzed for K⁺, Na⁺, Ca²⁺, and Mg²⁺ using an Agilent LC/GC-ICP-MS (Santa Clara, California).

Aequorin

pBINU-CHYA(K), a vector expressing stromal-localized aequorin (Mehlmer et al., 2012) was used to transform wildtype *N. tabacum* ‘Petit Havana’ using agrobacterium cocultivation. Transgenic plants were generated using agrobacterium-mediated transformation of wildtype ‘Petit Havana’ tobacco as per (Gallois and Marinho, 1995). Agrobacterium cultures grown to OD₆₀₀ = 0.4 were pelleted and resuspended in 2MS media (MS basal salts, 20g/L sucrose), then incubated with 1cm² *Nicotiana tabacum* ‘Petit Havana’ leaf explants for 48 hours, rinsed with sterile water 3x, with a final rinse of sterile water plus 200mg/L cefotaxime. Explants were selected on RMOP plus 200mg/L cefotaxime, 100mg/L timentin, and 100mg/L kanamycin. Plants were rooted in the equivalent media without added hormones, then transitioned to soil. DNA was extracted from leaves using CTAB/Phenol-Chloroform-Isoamyl alcohol (25:24:1, v/v) extractions and tested for the transgene using PCR with pRWC348/pRWC350 for pBINU-CHY(K). Expression level and plastid localization was confirmed using a Zeiss LSM 510 META confocal microscope to detect C-terminally fused YFP (Additional File 1). The line with the highest YFP signal was selected as a pollen donor for outcrossing to each of the chloroplast-transformed MdALB3 truncation lines to create dual chloroplast-nuclear transgenic lines. Seeds of confirmed dual-transgenic lines were sterilized (3 minutes of 70% ethyl alcohol followed by 10 minutes of 20% bleach (1.25% sodium hypochlorite), and 3 rinses with sterile water), then

plated on MS media with appropriate selection antibiotic (200mg/L kanamycin for aequorin parent lines, 200mg/L kanamycin and 500mg/L spectinomycin for doubled-transformed aequorin x transplastomic lines). Fourteen days after sowing, seedlings were transferred to a 96-well CELLSTAR® white fluorescence plate (Greiner Bio-One, Kremsmünster, Austria). 50µl of pretreatment media (5mM MES, 1.4mM CaCl₂, 20mM KCl, 10 µM native coelenterazine (NanoLight Technologies, Pinetop, Arizona), pH 5.8) was added to each well, and plates were incubated in total darkness for 24 hours. Aequorin kinetics were measured on a GloMax® Navigator (Promega, Madison, Wisconsin) for 12 seconds (5 readings per second) following the addition of 50ul of stimulant solution (400mM NaCl, 0.1% Silwet-77). 100ul of discharge solution (2M CaCl₂ dissolved in 20% (v/v) ethanol) was then added and measured for 2 minutes to quantify total coelenterazine and data was normalized to nM using the calculations of (Knight et al., 1996; Stephan et al., 2016).

Microscopy

Plants were grown in greenhouse conditions at the WSU Plant Growth Facilities, and fresh, young leaves were collected from the distal end of the first fully-expanded leaf (typically position 3 from the meristem). Samples were fixed in 2% formaldehyde/2% glutaraldehyde in 50mM cacodylate buffer overnight, washed three times with 50mM cacodylate buffer for 10 minutes each, and postfixed with 1% osmium tetroxide at 4°C overnight. Resin exchange was performed using a gradient of 30-100% ethanol, then infiltrated in propylene oxide and embedded in Spurr's resin. Thin sections of roughly 50nm thickness were taken on a Leica Reichert Ultracut R microtome. Sections were stained with 5 µl potassium permanganate and 2mL of 4% uranyl acetate (aq), then stained with Reynold's lead citrate buffered with carbonate-free sodium hydroxide. Finally, grids were stained with 1% uranyl acetate and rinsed three times

with sterile water. Sections were imaged from the palisade cell layer on an FEI Tecnai G2 20 Twin Transmission Electron Microscope.

PCR confirmation

Transgene presence was initially performed via PCR amplification using RWCP_322 and RWCP_323 (1st), RWCP_324 (2nd), RWCP_325 (3rd), RWCP_345 (4th), and RWCP_359 (Full). Integration into the chloroplast genome was confirmed using ADctP_36/ADctP_39 (Left flank) and ADctP_35/ADctP_38 (right flank). Insert size was determined using ADctP36/ADctP38.

qRT-PCR Analysis

RNA was extracted from milled leaf tissue of wildtype and transgenic plants using Plant RNA Isolation Kit (Agilent Technologies, Agilent, CA). Extracted RNA was treated with Turbo DNA-free™ Kit (Invitrogen, Carlsbad, CA). Quality was checked on 0.8% agarose gel, and RNA was quantified using Qubit™ RNA HS Assay Kit (Thermo Fisher Scientific, Waltham, MA). First-strand cDNA synthesis was performed on 200ng of RNA from each sample using SuperScript™ VILO™ III (Invitrogen, Carlsbad, CA). cDNA was quantified using Qubit™ ssDNA Assay Kit (Thermo Fisher Scientific, Waltham, MA) and diluted to 25ng/μl. Real-time quantitative PCR was performed on 25ng of cDNA for selected genes using iTaq™ Universal SYBR® Green Mastermix (Bio-Rad Laboratories, Hercules, CA) and run on a Mx3005® real-time PCR machine (Stratagene California, San Diego, California). PCR conditions were performed using a 2.5 minute second denaturation at 94°C followed by 50 cycles of 20-second denaturation at 94°C, 20-second annealing at 60°C, and 20-second extension at 72°C. A melt curve was performed with denaturation at 95°C for 30 seconds, annealing at 57°C for 30 seconds, and a final denaturation step at 95°C for 30 seconds. Isoform-specific SNP-anchored primers were used for each ALB3 homolog, with SNPs at the 3' end of the primer. AtALB3 was

tested using RWCP_604 and RWCP_605, NtALB3 using RWCP_608 and RWCP_609, and MdALB3.1 using RWCP_610 and RWCP_611. Elongation factor 1- α was used as a reference gene, and was amplified using RWCP_612 and RWCP_613.

Authors' contributions

RC, SS, and AD designed the study. DM and AD designed and constructed the basic chloroplast transformation vector pADct44, and SS and RC constructed ALB3 transformation vectors. RC cloned Alb3 genes and generated chloroplast transformation and fusion protein chimera constructs. SS and RC performed PCR analysis of transgenic events and performed growth performance assays on chloroplast transformants. HK, MW, HK, RC, and AD designed and analyzed phenomics experiment. BW and SH contributed to electron microscopy and confocal microscopy. AD supervised the study. SS, RC, and AD prepared the manuscript. All authors read and approved the manuscript.

Acknowledgments

Work in the Dhingra lab was supported by Washington State University Agriculture Center Research Hatch Grant WNP00011 to AD. RC, SMS, and SLH acknowledge the support received from National Institutes of Health/National Institute of General Medical Sciences through an institutional training grant award T32-GM008336. The contents of this work are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS or NIH. SLH acknowledges the support received from ARCS Seattle Chapter. The authors would like to thank the staff of the Franceschi Microscopy and Imaging Center at

Washington State University, especially Daniel Mullendore for performing confocal microscopy and Valerie Lynch-Holm for performing transmission electron microscopy.

Additionally, the authors would like to thank the staff of the Analytical Chemistry Lab at Washington State University, particular Jonathan Lomber and Michael Apasiku for preparing and analyzing samples via ICP-MS. For the performance of phenomics experiments, the authors thank Magnus Wood at the Phenomics Core at Washington State University, and for aequorin experiments, the authors thank Kiwamu Tanaka and Jeremy Jewell for provision of the pBINU-CHYA(K) vector, training, and troubleshooting. Finally, the authors would like to acknowledge the following undergraduate students that contributed to this project: Bernadette Quint, Nohely Castro-Velasquez, Chelsea Crabb, Grant Nelson, Paola Coronel, and Paula Aubrey.

References

- Adams, E., Shin, R., 2014. Transport , signaling , and homeostasis of potassium and sodium in plants. *J. Integr. Plant Biol.* 56, 231–249. <https://doi.org/10.1111/jipb.12159>
- Anderson, J.M., Chow, W.S., De Las Rivas, J., 2008. Dynamic flexibility in the structure and function of photosystem II in higher plant thylakoid membranes: The grana enigma. *Photosynth. Res.* 98, 575–587. <https://doi.org/10.1007/s11120-008-9381-3>
- Andersson, H., von Heijne, G., 1993. Sec dependent and sec independent assembly of E . coli inner membrane proteins : the topological rules depend on chain length. *EMBO J.* 12, 683–691. <https://doi.org/10.1002/j.1460-2075.1993.tb05702.x>
- Bals, T., Dünschede, B., Funke, S., Schünemann, D., 2010. Interplay between the cpSRP pathway components, the substrate LHCP and the translocase Alb3: An in vivo and in vitro study. *FEBS Lett.* 584, 4138–4144. <https://doi.org/10.1016/j.febslet.2010.08.053>
- Beck, K., Eisner, G., Trescher, D., Dalbey, R.E., Brunner, J., Müller, M., 2001. YidC, an assembly site for polytopic Escherichia coli membrane proteins located in immediate proximity to the SecYE translocon and lipids. *EMBO Rep.* 2, 709–714. <https://doi.org/10.1093/embo-reports/kve154>
- Bédard, J., Trösch, R., Wu, F., Ling, Q., Flores-Pérez, Ú., Töpel, M., Nawaz, F., Jarvis, P., 2017. New Suppressors of the Chloroplast Protein Import Mutant tic40 Reveal a Genetic Link between Protein Import and Thylakoid Biogenesis. *Plant Cell* 29, tpc.00962.2016. <https://doi.org/10.1105/tpc.16.00962>
- Bellafiore, S., 2002. Loss of Albino3 Leads to the Specific Depletion of the Light-Harvesting System. *Plant Cell* 14, 2303–2314. <https://doi.org/10.1105/tpc.003442>
- Benz, M., Bals, T., Gügel, I.L., Piotrowski, M., Kuhn, A., Schünemann, D., Soll, J., Ankele, E.,

2009. Alb4 of arabidopsis promotes assembly and stabilization of a non chlorophyll-binding photosynthetic complex, the CF1CF0-ATP synthase. *Mol. Plant* 2, 1410–1424.
<https://doi.org/10.1093/mp/ssp095>
- Björkman, O., Demmig, B., 1987. Photon yield of O₂ evolution and chlorophyll fluorescence characteristics at 77 K among vascular plants of diverse origins. *Planta* 170, 489–504.
- Bréhélin, C., Kessler, F., 2008. The Plastoglobule: A Bag Full of Lipid Biochemistry Tricks. *Photochem. Photobiol.* 84, 1388–1394.
- Bruce, B.D., 2000. Chloroplast transit peptides: Structure, function and evolution. *Trends Cell Biol.* 10, 440–447. [https://doi.org/10.1016/S0962-8924\(00\)01833-X](https://doi.org/10.1016/S0962-8924(00)01833-X)
- Cai, W., Ma, J., Chi, W., Zou, M., Guo, J., Lu, C., Zhang, L., 2010. Cooperation of LPA3 and LPA2 is essential for photosystem II assembly in Arabidopsis. *Plant Physiol.* 154, 109–120.
<https://doi.org/10.1104/pp.110.159558>
- Carraretto, L., Teardo, E., Checchetto, V., Finazzi, G., Uozumi, N., Szabo, I., 2016. Ion Channels in Plant Bioenergetic Organelles, Chloroplasts and Mitochondria: From Molecular Identification to Function. *Mol. Plant* 9, 371–395.
<https://doi.org/10.1016/j.molp.2015.12.004>
- Cellier, F., Conéjéro, G., Ricaud, L., Luu, D.T., Lepetit, M., Casse, F., 2004. Characterization of AtCHX17, a member of the cation / H⁺ exchangers, CHX family, from Arabidopsis thaliana suggests a role in K⁺ homeostasis. *Plant J.* 39, 834–846.
<https://doi.org/10.1111/j.1365-313X.2004.02177.x>
- Charpentier, M., Bredemeier, R., Wanner, G., Takeda, N., Schleiff, E., Parniske, M., 2008. Lotus japonicus CASTOR and POLLUX Are Ion Channels Essential for Perinuclear Calcium Spiking in Legume Root Endosymbiosis. *Plant Cell Online* 20, 3467–3479.

<https://doi.org/10.1105/tpc.108.063255>

Checchetto, V., Teardo, E., Carraretto, L., Leanza, L., Szabo, I., 2016. Biochimica et Biophysica Acta Physiology of intracellular potassium channels : A unifying role as mediators of counterion fluxes ? ☆. *Biochim. Biophys. Acta* 1857, 1258–1266.

<https://doi.org/10.1016/j.bbabi.2016.03.011>

Chen, Y., Soman, R., Shanmugam, S.K., Kuhn, A., Dalbey, R.E., 2014. The role of the strictly conserved positively charged residue differs among the gram-positive, gram-negative, and chloroplast YidC homologs. *J. Biol. Chem.* 289, 35656–35667.

<https://doi.org/10.1074/jbc.M114.595082>

Christian, R., Hewitt, S., Roalson, E., Dhingra, A., 2019. Genome-Scale Characterization of Predicted Plastid-Targeted Proteins in Higher Plants. Unpublished Work.

Clarke, J.L., Daniell, H., 2011. Plastid biotechnology for crop production: present status and future perspectives. *Plant Mol. Biol.* 76, 211–20. <https://doi.org/10.1007/s11103-011-9767-z>

Davis, S.J., Vierstra, R.D., 1998. Soluble, highly fluorescent variants of green fluorescent protein (GFP) for use in higher plants. *Plant Mol. Biol.* 36, 521–528.

<https://doi.org/10.1023/A:1005991617182>

DeLisa, M.P., Tullman, D., Georgiou, G., 2003. Folding quality control in the export of proteins by the bacterial twin-arginine translocation pathway. *Proc. Natl. Acad. Sci.* 100, 6115–6120. <https://doi.org/10.1073/pnas.0937838100>

Dhingra, A., Portis, A.R., Daniell, H., 2004. Enhanced translation of a chloroplast-expressed RbcS gene restores small subunit levels and photosynthesis in nuclear RbcS antisense plants. *Proc. Natl. Acad. Sci.* 101, 6315–6320. <https://doi.org/10.1073/pnas.0400981101>

- Di Tommaso, P., Moretti, S., Xenarios, I., Orobittg, M., Montanyola, A., Chang, J.M., Taly, J.F., Notredame, C., 2011. T-Coffee: A web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.* 39, 13–17. <https://doi.org/10.1093/nar/gkr245>
- Dünschede, B., Bals, T., Funke, S., Schünemann, D., 2011. Interaction studies between the chloroplast signal recognition particle subunit cpSRP43 and the full-length translocase Alb3 reveal a membrane-embedded binding region in Alb3 protein. *J. Biol. Chem.* 286, 35187–35195. <https://doi.org/10.1074/jbc.M111.250746>
- Enz, C., Steinkamp, T., Wagner, R., 1993. Ion channels in the thylakoid membrane (a patch-clamp study). *Biochim. Biophys. Acta (BBA)-Bioenergetics* 1143, 67–76.
- Ettinger, W.F., Clear, A.M., Fanning, K.J., Peck, M. Lou, 1999. Identification of a Ca²⁺/H⁺ Antiport in the Plant Chloroplast Thylakoid Membrane¹. *Plant Physiol.* 119, 1379–1386. <https://doi.org/10.1104/pp.119.4.1379>
- Falk, S., Ravaud, S., Koch, J., Sinning, I., 2010. The C terminus of the Alb3 membrane insertase recruits cpSRP43 to the thylakoid membrane. *J. Biol. Chem.* 285, 5954–5962. <https://doi.org/10.1074/jbc.M109.084996>
- Frain, K.M., Gangl, D., Jones, A., Zedler, J.A.Z., Robinson, C., 2016. Protein translocation and thylakoid biogenesis in cyanobacteria. *Biochim. Biophys. Acta - Bioenerg.* 1857, 266–273. <https://doi.org/10.1016/j.bbabi.2015.08.010>
- Fröderberg, L., Röhl, T., Van Wijk, K.J., De Gier, J.W.L., 2001. Complementation of bacterial SecE by a chloroplastic homologue. *FEBS Lett.* 498, 52–56. [https://doi.org/10.1016/S0014-5793\(01\)02494-2](https://doi.org/10.1016/S0014-5793(01)02494-2)
- Gallois, P., Marinho, P., 1995. Leaf disk transformation usin *Agrobacterium tumefasciens* -

- expression of heterologous genes in tobacco., in: Jones, H. (Ed.), *Methods in Molecular Biology*, Vol 49. *Plant Gene Transfer and Expression Protocols*. Humana Press, Inc, Totowa, New Jersey, pp. 38–48. <https://doi.org/10.1385/0-89603-321-X:39>
- Gerdes, L., Bals, T., Klostermann, E., Karl, M., Philippar, K., Hünken, M., Soll, J., Schünemann, D., 2006. A second thylakoid membrane-localized Alb3/OxaI/YidC homologue is involved in proper chloroplast biogenesis in *Arabidopsis thaliana*. *J. Biol. Chem.* 281, 16632–16642. <https://doi.org/10.1074/jbc.M513623200>
- Gray, A.N., Henderson-Frost, J.M., 2011. Unbalanced Charge Distribution as a Determinant for Dependence of a Subset of *Escherichia coli* Membrane Proteins on the Membrane Insertase YidC. *MBio* 2, e00238-11. <https://doi.org/10.1128/mBio.00238-11>. Editor
- Grusak, M.A., 1994. Iron Transport to Developing Ovules of *Pisum sativum*. *Plant Physiol.* 205, 649–655.
- Gruschke, S., Gröne, K., Heublein, M., Hölz, S., Israel, L., Imhof, A., Herrmann, J.M., Ott, M., 2010. Proteins at the polypeptide tunnel exit of the yeast mitochondrial ribosome. *J. Biol. Chem.* 285, 19022–19028. <https://doi.org/10.1074/jbc.M110.113837>
- Haque, M.E., Elmore, K.B., Tripathy, A., Koc, H., Koc, E.C., Spremulli, L.L., 2010a. Properties of the C-terminal tail of human mitochondrial inner membrane protein Oxa1L and its interactions with mammalian mitochondrial ribosomes. *J. Biol. Chem.* 285, 28353–28362. <https://doi.org/10.1074/jbc.M110.148262>
- Haque, M.E., Spremulli, L.L., Fecko, C.J., 2010b. Identification of protein-protein and protein-ribosome interacting regions of the C-terminal tail of human mitochondrial inner membrane protein Oxa1L. *J. Biol. Chem.* 285, 34991–34998. <https://doi.org/10.1074/jbc.M110.163808>
- Hennon, S.W., Soman, R., Zhu, L., Dalbey, R.E., 2015. YidC/Alb3/Oxa1 Family of Insertases. *J.*

- Biol. Chem. jbc.R115.638171. <https://doi.org/10.1074/jbc.R115.638171>
- Hochmal, A.K., Schulze, S., Trompelt, K., Hippler, M., 2015. Calcium-dependent regulation of photosynthesis. *Biochim. Biophys. Acta - Bioenerg.* 1847, 993–1003. <https://doi.org/10.1016/j.bbabi.2015.02.010>
- Horn, A., Hennig, J., Ahmed, Y.L., Stier, G., Wild, K., Sattler, M., Sinning, I., 2015. Structural basis for cpSRP43 chromodomain selectivity and dynamics in Alb3 insertase interaction. *Nat. Commun.* 6, 8875. <https://doi.org/10.1038/ncomms9875>
- Houben, E.N.G., ten Hagen-Jongman, C.M., Brunner, J., Oudega, B., Luirink, J., 2004. The two membrane segments of leader peptidase partition one by one into the lipid bilayer via a Sec/YidC interface. *EMBO Rep.* 5, 970–975. <https://doi.org/10.1038/sj.embor.7400261>
- Huang, S.S., Chen, J., Dong, X.J., Patton, J., Pei, Z.M., Zheng, H.L., 2012. Calcium and calcium receptor CAS promote *Arabidopsis thaliana* de-etiolation. *Physiol. Plant.* 144, 73–82. <https://doi.org/10.1111/j.1399-3054.2011.01523.x>
- Hynds, P.J., Robinson, D., Robinson, C., 1998. The Sec-independent twin-arginine translocation system can transport both tightly folded and malfolded proteins across the thylakoid membrane. *J. Biol. Chem.* 273, 34868–34874. <https://doi.org/10.1074/jbc.273.52.34868>
- Imaizumi-Anraku, H., Takeda, N., Charpentier, M., 2004. Plastid proteins crucial for symbiotic fungal and bacterial entry into plant roots. *Nature* 428, 527–531.
- Jia, L., Dienhart, M., Schrap, M., McCauley, M., Hell, K., Stuart, R.A., 2003. Yeast Oxal1 interacts with mitochondrial ribosomes: The importance of the C-terminal region of Oxal1. *EMBO J.* 22, 6438–6447. <https://doi.org/10.1093/emboj/cdg624>
- Jiang, F., Chen, M., Yi, L., De Gier, J.W., Kuhn, A., Dalbey, R.E., 2003. Defining the regions of *Escherichia coli* YidC that contribute to activity. *J. Biol. Chem.* 278, 48965–48972.

<https://doi.org/10.1074/jbc.M307362200>

Jiang, F., Yi, L., Moore, M., Chen, M., Rohl, T., Van Wijk, K.J., De Gier, J.W.L., Henry, R., Dalbey, R.E., 2002. Chloroplast YidC homolog Albino3 can functionally complement the bacterial YidC depletion strain and promote membrane insertion of both bacterial and chloroplast thylakoid proteins. *J. Biol. Chem.* 277, 19281–19288.

<https://doi.org/10.1074/jbc.M110857200>

Johnson, C.H., Knight, M.R., Kondo, T., Masson, P., Sedbrook, J., Haley, A., Trewavas, A., 1995. Circadian oscillations of cytosolic and chloroplastic free calcium in plants. *Science* (80-.). 269, 1863–1865. <https://doi.org/10.1126/science.7569925>

Jong, W.S.P., Ten Hagen-Jongman, C.M., Ruijter, E., Orru, R.V.A., Genevaux, P., Luirink, J., 2010. YidC is involved in the biogenesis of the secreted autotransporter hemoglobin protease. *J. Biol. Chem.* 285, 39682–39690. <https://doi.org/10.1074/jbc.M110.167650>

Kedrov, A., Sustarsic, M., De Keyzer, J., Caumanns, J.J., Wu, Z.C., Driessen, A.J.M., 2013. Elucidating the native architecture of the YidC: Ribosome complex. *J. Mol. Biol.* 425, 4112–4124. <https://doi.org/10.1016/j.jmb.2013.07.042>

Kirchhoff, H., 2013. Architectural switches in plant thylakoid membranes. *Photosynth. Res.* 116, 481–487. <https://doi.org/10.1007/s11120-013-9843-0>

Klostermann, E., Droste Gen Helling, I., Carde, J.-P., Schünemann, D., 2002. The thylakoid membrane protein ALB3 associates with the cpSecY-translocase in *Arabidopsis thaliana*. *Biochem. J.* 368, 777–781. <https://doi.org/10.1042/BJ20021291>

Kmiec, B., Teixeira, P.F., Glaser, E., 2014. Shredding the signal: Targeting peptide degradation in mitochondria and chloroplasts. *Trends Plant Sci.* 19, 771–778. <https://doi.org/10.1016/j.tplants.2014.09.004>

- Knight, H., Trewavas, A.J., Knight, M.R., 1996. Cold Calcium Signaling in Arabidopsis Involves Two Cellular Pools and a Change in Calcium Signature after Acclimation. *Plant Cell* 8, 489–503. <https://doi.org/10.1105/tpc.8.3.489>
- Koch, H.G., Moser, M., Schimz, K.L., Müller, M., 2002. The integration of YidC into the cytoplasmic membrane of *Escherichia coli* requires the signal recognition particle, SecA and SecYEG. *J. Biol. Chem.* 277, 5715–5718. <https://doi.org/10.1074/jbc.C100683200>
- Kohler, R., Boehringer, D., Greber, B., Bingel-Erlenmeyer, R., Collinson, I., Schaffitzel, C., Ban, N., 2009. YidC and Oxa1 Form Dimeric Insertion Pores on the Translating Ribosome. *Mol. Cell* 34, 344–353. <https://doi.org/10.1016/j.molcel.2009.04.019>
- Kreimer, G., Surek, B., Woodrow, I.E., Lutzko, E., 1987. Calcium binding by spinach stromal proteins. *Planta* 171, 259–265. <https://doi.org/10.1007/BF00391103>
- Króliczewski, J., Bartoszewski, R., Króliczewska, B., 2017. Chloroplast PetD protein: Evidence for SRP/Alb3-dependent insertion into the thylakoid membrane. *BMC Plant Biol.* 17, 1–15. <https://doi.org/10.1186/s12870-017-1176-2>
- Krüger, V., Deckers, M., Hildenbeutel, M., Van Der Laan, M., Hellmers, M., Dreker, C., Preuss, M., Herrmann, J.M., Rehling, P., Wagner, R., Meinecke, M., 2012. The mitochondrial oxidase assembly protein1 (Oxa1) insertase forms a membrane pore in lipid bilayers. *J. Biol. Chem.* 287, 33314–33326. <https://doi.org/10.1074/jbc.M112.387563>
- Kuhn, A., Stuart, R., Henry, R., Dalbey, R.E., 2003. The Alb3/Oxa1/YidC protein family: Membrane-localized chaperones facilitating membrane protein insertion? *Trends Cell Biol.* 13, 510–516. <https://doi.org/10.1016/j.tcb.2003.08.005>
- Kumazaki, K., Chiba, S., Takemoto, M., Furukawa, A., Nishiyama, K., Sugano, Y., Mori, T., Dohmae, N., Hirata, K., Nakada-Nakura, Y., Maturana, A.D., Tanaka, Y., Mori, H., Sugita,

- Y., Arisaka, F., Ito, K., Ishitani, R., Tsukazaki, T., Nureki, O., 2014a. Structural basis of Sec-independent membrane protein insertion by YidC. *Nature* 509, 516–20.
<https://doi.org/10.1038/nature13167>
- Kumazaki, K., Kishimoto, T., Furukawa, A., Mori, H., Tanaka, Y., Dohmae, N., Ishitani, R., Tsukazaki, T., Nureki, O., 2014b. Crystal structure of *Escherichia coli* YidC, a membrane protein chaperone and insertase. *Sci. Rep.* 4, 1–6. <https://doi.org/10.1038/srep07299>
- Kunz, H.-H., Gierth, M., Herdean, A., Satoh-Cruz, M., Kramer, D.M., Spetea, C., Schroeder, J.I., 2014. Plastidial transporters KEA1, -2, and -3 are essential for chloroplast osmoregulation, integrity, and pH regulation in *Arabidopsis*. *Proc. Natl. Acad. Sci.* 111, 7480–7485.
<https://doi.org/10.1073/pnas.1323899111>
- Li, H.Y., Guo, Z.F., Zhu, Y.X., 1998. Molecular cloning and analysis of a pea cDNA that is expressed in darkness and very rapidly induced by gibberellic acid. *Mol. Gen. Genet.* 259, 393–397. <https://doi.org/10.1007/s004380050828>
- Li, J., Wang, D.Y., Li, Q., Xu, Y.J., Cui, K.M., Zhu, Y.X., 2004. PPF1 inhibits programmed cell death in apical meristems of both G2 pea and transgenic *Arabidopsis* plants possibly by delaying cytosolic Ca²⁺ elevation. *Cell Calcium* 35, 71–77.
<https://doi.org/10.1016/j.ceca.2003.07.003>
- Li, S., Yin, D., Wu, F., Wang, S., Deng, Q., Tang, Y., Zhou, H., Li, P., 2007. Introduction of the PPF1 gene into rice (*Oryza sativa* L.) results in delayed leaf senescence. *Euphytica* 153, 257–265. <https://doi.org/10.1007/s10681-006-9261-x>
- Li, X., 2011. Infiltration of *Nicotiana benthamiana* Protocol for Transient Expression via *Agrobacterium*. *Bio-Protocol* 1, e95. <https://doi.org/10.1017/CBO9781107415324.004>
- Li, Y., Martin, J.R., Aldama, G.A., Fernandez, D.E., Cline, K., 2017. Identification of Putative

- Substrates of SEC2, a Chloroplast Inner Envelope Translocase. *Plant Physiol.* 173, 2121–2137. <https://doi.org/10.1104/pp.17.00012>
- Liu, Z., 1996. Hetero-stagger cloning: Efficient and rapid cloning of PCR products. *Nucleic Acids Res.* 24, 2458–2459. <https://doi.org/10.1093/nar/24.12.2458>
- Long, D., Martin, M., Sundberg, E., Swinburne, J., Puangsomlee, P., Coupland, G., 1993. The maize transposable element system Ac/Ds as a mutagen in Arabidopsis: identification of an albino mutation induced by Ds insertion. *Proc. Natl. Acad. Sci.* 90, 10370–10374. <https://doi.org/10.1073/pnas.90.21.10370>
- Loro, G., Wagner, S., Doccula, F.G., Behera, S., Weinl, S., Kudla, J., Schwarzländer, M., Costa, A., Zottini, M., 2016. Chloroplast-specific in vivo Ca²⁺ imaging using Yellow Cameleon fluorescent protein sensors reveals organelle-autonomous Ca²⁺ signatures in the stroma. *Plant Physiol.* 171, 2317–2330. <https://doi.org/10.1104/pp.16.00652>
- Maathuis, F.J.M., 2013. Sodium in plants: perception, signalling, and regulation of sodium fluxes. *J. Exp. Bot.* 65, 849–858. <https://doi.org/10.1093/jxb/ert326>
- Marques, J.P., Schattet, M.H., Hause, G., Dudeck, I., Klösgen, R.B., 2004. In vivo transport of folded EGFP by the Δ pH/TAT-dependent pathway in chloroplasts of Arabidopsis thaliana. *J. Exp. Bot.* 55, 1697–1706. <https://doi.org/10.1093/jxb/erh191>
- Matos, C.F., Robinson, C., Di Cola, A., 2008. The Tat system proofreads FeS protein substrates and directly initiates the disposal of rejected molecules. *EMBO J.* 27, 2055–2063. <https://doi.org/10.1038/emboj.2008.132>
- Mehlmer, N., Parvin, N., Hurst, C.H., Knight, M.R., Teige, M., Vothknecht, U.C., 2012. A toolset of aequorin expression vectors for in planta studies of subcellular calcium concentrations in Arabidopsis thaliana. *J. Exp. Bot.* 63, 1751–1761.

<https://doi.org/10.1093/jxb/err406>

- Moore, M., Goforth, R.L., Mori, H., Henry, R., 2003. Functional interaction of chloroplast SRP/FtsY with the ALB3 translocase in thylakoids: Substrate not required. *J. Cell Biol.* 162, 1245–1254. <https://doi.org/10.1083/jcb.200307067>
- Moore, M., Harrison, M.S., Peterson, E.C., Henry, R., 2000. Chloroplast Oxa1p homolog albino3 is required for post-translational integration of the light harvesting chlorophyll-binding protein into thylakoid membranes. *J. Biol. Chem.* 275, 1529–1532. <https://doi.org/10.1074/jbc.275.3.1529>
- Nagamori, S., Smirnova, I.N., Kaback, H.R., 2004. Role of YidC in folding of polytopic membrane proteins. *J. Cell Biol.* 165, 53–62. <https://doi.org/10.1083/jcb.200402067>
- Nomura, H., Komori, T., Kobori, M., Nakahira, Y., Shiina, T., 2008. Evidence for chloroplast control of external Ca²⁺-induced cytosolic Ca²⁺ transients and stomatal closure. *Plant J.* 53, 988–998. <https://doi.org/10.1111/j.1365-313X.2007.03390.x>
- Nomura, H., Komori, T., Uemura, S., Kanda, Y., Shimotani, K., Nakai, K., Furuichi, T., Takebayashi, K., Sugimoto, T., Sano, S., Suwastika, I.N., Fukusaki, E., Yoshioka, H., Nakahira, Y., Shiina, T., 2012. Chloroplast-mediated activation of plant immune signalling in Arabidopsis. *Nat. Commun.* 3, 910–926. <https://doi.org/10.1038/ncomms1926>
- Nomura, H., Shiina, T., 2014. Calcium signaling in plant endosymbiotic organelles: Mechanism and role in physiology. *Mol. Plant* 7, 1094–1104. <https://doi.org/10.1093/mp/ssu020>
- Notredame, C., Higgins, D.G., Heringa, J., 2000. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302, 205–217. <https://doi.org/10.1006/jmbi.2000.4042>
- Paetzel, M., Karla, A., Strynadka, N.C.J., Dalbey, R.E., 2002. Signal peptidases. *Chem. Rev.*

102, 4549–4579. <https://doi.org/10.1021/cr010166y>

- Palmer, S.R., Crowley, P.J., Oli, M.W., Adam Ruelf, M., Michalek, S.M., Jeannine Brady, L., 2012. YidC1 and YidC2 are functionally distinct proteins involved in protein secretion, biofilm formation and cariogenicity of *Streptococcus mutans*. *Microbiol. (United Kingdom)* 158, 1702–1712. <https://doi.org/10.1099/mic.0.059139-0>
- Park, E., Rapoport, T.A., 2011. Preserving the membrane barrier for small molecules during bacterial protein translocation. *Nature* 473, 239–424.
- Pasch, J.C., Nickelsen, J., Schünemann, D., 2005. The yeast split-ubiquitin system to study chloroplast membrane protein interactions. *Appl. Microbiol. Biotechnol.* 69, 440–447. <https://doi.org/10.1007/s00253-005-0029-3>
- Peltier, J.-B., Emanuelsson, O., Kalume, D.E., Ytterberg, J., Friso, G., Rudella, A., Liberles, D.A., Söderberg, L., Roepstorff, P., von Heijne, G., van Wijk, K.J., 2002. Central Functions of the Lumenal and Peripheral Thylakoid Proteome of *Arabidopsis* Determined by Experimentation and Genome-Wide Prediction. *Plant Cell* 14, 211–236. <https://doi.org/10.1105/tpc.010304>
- Porra, R.J., Thompson, W.A., Kriedemann, P.E., 1989. Determination of accurate extinction coefficients and simultaneous equations for assaying chlorophylls a and b extracted with four different solvents: verification of the concentration of chlorophyll standards by atomic absorption spectoscopy. *Biochim. Biophys. Acta - Bioenerg.* 975, 384–394.
- Portis, A.R., Heldt, H.W., 1976. Light-dependent changes of the Mg²⁺ concentration in the stroma in relation to the Mg²⁺-dependency of CO₂ fixation in intact chloroplasts. *BBA - Bioenerg.* 449, 434–446. [https://doi.org/10.1016/0005-2728\(76\)90154-7](https://doi.org/10.1016/0005-2728(76)90154-7)
- Price, C.E., Driessen, A.J.M., 2010. Conserved negative charges in the transmembrane segments

- of subunit K of the NADH:ubiquinone oxidoreductase determine its dependence on YidC for membrane insertion. *J. Biol. Chem.* 285, 3575–3581.
<https://doi.org/10.1074/jbc.M109.051128>
- Ravaud, S., Stjepanovic, G., Wild, K., Sinning, I., 2008. The crystal structure of the periplasmic domain of the *Escherichia coli* membrane protein insertase YidC contains a substrate binding cleft. *J. Biol. Chem.* 283, 9350–9358. <https://doi.org/10.1074/jbc.M710493200>
- Robinson, C., Bolhuis, A., 2001. Protein targeting by the twin-arginine translocation pathway. *Nat. Rev. Mol. Cell Biol.* 2, 350–356. <https://doi.org/10.1038/35073038>
- Robinson, C., Matos, C.F.R.O., Beck, D., Ren, C., Lawrence, J., Vasisht, N., Mendel, S., 2011. Transport and proofreading of proteins by the twin-arginine translocation (Tat) system in bacteria. *Biochim. Biophys. Acta - Biomembr.* 1808, 876–884.
<https://doi.org/10.1016/j.bbamem.2010.11.023>
- Rocha, A.G., Vothknecht, U.C., 2012. The role of calcium in chloroplasts-an intriguing and unresolved puzzle. *Protoplasma* 249, 957–966. <https://doi.org/10.1007/s00709-011-0373-3>
- Sachelaru, I., Petriman, N.A., Kudva, R., Kuhn, P., Welte, T., Knapp, B., Drepper, F., Warscheid, B., Koch, H.G., 2013. YidC occupies the lateral gate of the SecYEG translocon and is sequentially displaced by a nascent membrane protein. *J. Biol. Chem.* 288, 16295–16307. <https://doi.org/10.1074/jbc.M112.446583>
- Sachelaru, I., Winter, L., Knyazev, D.G., Zimmermann, M., Vogt, A., Kuttner, R., Ollinger, N., Siligan, C., Pohl, P., Koch, H.G., 2017. YidC and SecYEG form a heterotetrameric protein translocation channel. *Sci. Rep.* 7, 1–15. <https://doi.org/10.1038/s41598-017-00109-8>
- Sai, J., Johnson, C.H., 2002. Dark-Stimulated Calcium Ion Fluxes in the Chloroplast Stroma and Cytosol. *Plant Cell* 14, 1279–1291. <https://doi.org/10.1105/tpc.000653.1280>

- Schneider, A., Steinberger, I., Herdean, A., Gandini, C., Eisenhut, M., Kurz, S., Morper, A., Hoecker, N., Rühle, T., Labs, M., Flüge, U.I., Geimer, S., Schmidt, S.B., Husted, S., Weber, A.P.M., Spetea, C., Leister, D., 2016. The Evolutionarily Conserved Protein PHOTOSYNTHESIS AFFECTED MUTANT71 is Required for Efficient Manganese Uptake at the Thylakoid Membrane in Arabidopsis. *Plant Cell* 28, tpc.00812.2015. <https://doi.org/10.1105/tpc.15.00812>
- Schneider, A., Steinberger, I., Strissel, H., Kunz, H.H., Manavski, N., Meurer, J., Burkhard, G., Jarzombski, S., Schünemann, D., Geimer, S., Flüge, U.I., Leister, D., 2014. The Arabidopsis Tellurite resistance C protein together with ALB3 is involved in photosystem II protein synthesis. *Plant J.* 78, 344–356. <https://doi.org/10.1111/tpj.12474>
- Schöttler, M.A., Albus, C.A., Bock, R., 2011. Photosystem I: Its biogenesis and function in higher plants. *J. Plant Physiol.* 168, 1452–1461. <https://doi.org/10.1016/j.jplph.2010.12.009>
- Schubert, M., Petersson, U.A., Haas, B.J., Funk, C., Schröder, W.P., Kieselbach, T., 2002. Proteome Map of the Chloroplast Lumen of Arabidopsis thaliana. *J. Biol. Chem.* 277, 8354–8365. <https://doi.org/10.1074/jbc.M108575200>
- Shipman-Roston, R.L., Ruppel, N.J., Damoc, C., Phinney, B.S., Inoue, K., 2010. The Significance of Protein Maturation by Plastidic Type I Signal Peptidase 1 for Thylakoid Development in Arabidopsis Chloroplasts. *Plant Physiol.* 152, 1297–1308. <https://doi.org/10.1104/pp.109.151977>
- Stephan, A.B., Kunz, H.-H., Yang, E., Schroeder, J.I., 2016. Rapid hyperosmotic-induced Ca²⁺ responses in *Arabidopsis thaliana* exhibit sensory potentiation and involvement of plastidial KEA transporters. *Proc. Natl. Acad. Sci.* 113, E5242–E5249. <https://doi.org/10.1073/pnas.1519555113>

- Sundberg, E., Slagter, J.G., Fridborg, I., Cleary, S.P., Robinson, C., Coupland, G., 1997. ALBINO3, an Arabidopsis nuclear gene essential for chloroplast differentiation, encodes a chloroplast protein that shows homology to proteins present in bacterial membranes and yeast mitochondria. *Plant Cell* 9, 717–730. <https://doi.org/10.1105/tpc.9.5.717>
- Szyrach, G., Ott, M., Bonnefoy, N., Neupert, W., Herrmann, J.M., 2003. Ribosome binding to the Oxa1 complex facilitates co-translational protein insertion in mitochondria. *EMBO J.* 22, 6448–6457. <https://doi.org/10.1093/emboj/cdg623>
- Takechi, K., Sodmergen, Murata, M., Motoyoshi, F., Sakamoto, W., 2000. The yellow variegated (*var2*) locus encodes a homologue of FtsH, an ATP-dependent protease in arabidopsis. *Plant Cell Physiol.* 41, 1334–1346. <https://doi.org/10.1093/pcp/pcd067>
- Teixeira, P.F., Kmiec, B., Branca, R.M.M., Murcha, M.W., Byzia, A., Ivanova, A., Whelan, J., Drag, M., Lehtiö, J., Glaser, E., 2017. A multi-step peptidolytic cascade for amino acid recovery in chloroplasts. *Nat. Chem. Biol.* 13, 15–17. <https://doi.org/10.1038/nchembio.2227>
- Thomas, J.D., Daniel, R.A., Errington, J., Robinson, C., 2001. Export of active green fluorescent protein to the periplasm by the twin-arginine translocase (Tat) pathway in *Escherichia coli*. *Mol. Microbiol.* 39, 47–53. <https://doi.org/10.1046/j.1365-2958.2001.02253.x>
- Trösch, R., Töpel, M., Flores-Pérez, Ú., Jarvis, P., 2015. Genetic and Physical Interaction Studies Reveal Functional Similarities between ALBINO3 and ALBINO4 in Arabidopsis. *Plant Physiol.* 169, 1292–1306. <https://doi.org/10.1104/pp.15.00376>
- Urbanus, M.L., Fröderberg, L., Drew, D., Björk, P., De Gier, J.W.L., Brunner, J., Oudega, B., Luirink, J., 2002. Targeting, insertion, and localization of *Escherichia coli* YidC. *J. Biol. Chem.* 277, 12718–12723. <https://doi.org/10.1074/jbc.M200311200>

- Urbanus, M.L., Scotti, P.A., Fröderberg, L., Sääf, A., De Gier, J.W.L., Brunner, J., Samuelson, J.C., Dalbey, R.E., Oudega, B., Luirink, J., 2001. Sec-dependent membrane protein insertion: Sequential interaction of nascent FtsQ with SecY and YidC. *EMBO Rep.* 2, 524–529. <https://doi.org/10.1093/embo-reports/kve108>
- Urbischek, M., Nick von Braun, S., Brylok, T., Gügel, I.L., Richter, A., Koskela, M., Grimm, B., Mulo, P., Bölter, B., Soll, J., Ankele, E., Schwenkert, S., 2015. The extreme Albino3 (Alb3) C terminus is required for Alb3 stability and function in *Arabidopsis thaliana*. *Planta* 242, 733–746. <https://doi.org/10.1007/s00425-015-2352-y>
- van Bloois, E., Dekker, H.L., Fröderberg, L., Houben, E.N.G., Urbanus, M.L., de Koster, C.G., de Gier, J.W., Luirink, J., 2008. Detection of cross-links between FtsH, YidC, HflIK/C suggests a linked role for these proteins in quality control upon insertion of bacterial inner membrane proteins. *FEBS Lett.* 582, 1419–1424. <https://doi.org/10.1016/j.febslet.2008.02.082>
- Van den Berg, B., Clemons, W.M., Collinson, I., Modis, Y., Hartmann, E., Harrison, S.C., Rapoport, T.A., 2004. X-ray structure of a protein-conducting channel. *Nature* 427, 36–44. <https://doi.org/10.1038/nature02218>
- Wagner, S., Pop, O., Haan, G.J., Baars, L., Koningstein, G., Klepsch, M.M., Genevaux, P., Luirink, J., De Gier, J.W., 2008. Biogenesis of MalF and the MalFGK2 maltose transport complex in *Escherichia coli* requires YidC. *J. Biol. Chem.* 283, 17881–17890. <https://doi.org/10.1074/jbc.M801481200>
- Wang, a. X., Wang, D.Y., 2009. Regulation of the ALBINO3-mediated transition to flowering in *Arabidopsis* depends on the expression of CO and GA1. *Biol. Plant.* 53, 484–492. <https://doi.org/10.1007/s10535-009-0089-9>

- Wang, C., Xu, W., Jin, H., Zhang, T., Lai, J., Zhou, X., Zhang, S., Liu, S., Duan, X., Wang, H., Peng, C., Yang, C., 2016. A Putative Chloroplast-Localized Ca²⁺/H⁺ Antiporter CCHA1 Is Involved in Calcium and pH Homeostasis and Required for PSII Function in Arabidopsis. *Mol. Plant* 9, 1183–1196. <https://doi.org/10.1016/j.molp.2016.05.015>
- Wang, D., Xu, Y., Li, Q., Hao, X., Cui, K., Sun, F., Zhu, Y., 2003. Transgenic expression of a putative calcium transporter affects the time of Arabidopsis flowering. *Plant J.* 33, 285–292. <https://doi.org/10.1046/j.1365-313X.2003.01627.x>
- Wang, D.Y., Li, Q., Cui, K.M., Zhu, Y.X., 2008. PPF1 may suppress plant senescence via activating TFL1 in transgenic Arabidopsis plants. *J. Integr. Plant Biol.* 50, 475–483. <https://doi.org/10.1111/j.1744-7909.2008.00643.x>
- Weinl, S., Held, K., Schlücking, K., Steinhorst, L., Kuhlger, S., Hippler, M., Kudla, J., 2008. A plastid protein crucial for Ca²⁺-regulated stomatal responses. *New Phytol.* 179, 675–686. <https://doi.org/10.1111/j.1469-8137.2008.02492.x>
- Wickström, D., Wagner, S., Simonsson, P., Pop, O., Baars, L., Ytterberg, A.J., Van Wijk, K.J., Luirink, J., De Gier, J.W.L., 2011. Characterization of the consequences of YidC depletion on the inner membrane proteome of *E. coli* using 2D blue native/SDS-PAGE. *J. Mol. Biol.* 409, 124–135. <https://doi.org/10.1016/j.jmb.2011.03.068>
- Woolhead, C.A., Thompson, S.J., Moore, M., Tissier, C., Mant, A., Rodger, A., Henry, R., Robinson, C., 2001. Distinct Albino3-dependent and -independent Pathways for Thylakoid Membrane Protein Insertion. *J. Biol. Chem.* 276, 40841–40846. <https://doi.org/10.1074/jbc.M106523200>
- Xie, B., Xie, Y., 2011. Twin PCRs: a simple and efficient method for directional cloning of PCR products. *World J. Microbiol. Biotechnol.* 27, 2223–2225.

- Xu, Y.-J., Wang, D.-Y., Zhu, Y.-X., 2002. Expression of the Thylakoid Membrane Localized PPF1 in Transgenic Arabidopsis Affects Chloroplast Development. *Acta Bot. Sin.* 44, 1314–1320.
- Zhang, L., Kato, Y., Otters, S., Vothknecht, U.C., Sakamoto, W., 2012. Essential Role of VIPP1 in Chloroplast Envelope Maintenance in Arabidopsis. *Plant Cell* 24, 3695–3707.
<https://doi.org/10.1105/tpc.112.103606>
- Zhang, X.P., Glaser, E., 2002. Interaction of plant mitochondrial and chloroplast signal peptides with the Hsp70 molecular chaperone. *Trends Plant Sci.* 7, 14–21.
[https://doi.org/10.1016/S1360-1385\(01\)02180-X](https://doi.org/10.1016/S1360-1385(01)02180-X)
- Zhao, X., Chen, T., Feng, B., Zhang, C., Peng, S., Zhang, X., Fu, G., Tao, L., 2017. Non-photochemical Quenching Plays a Key Role in Light Acclimation of Rice Plants Differing in Leaf Color. *Front. Plant Sci.* 7, 1–17. <https://doi.org/10.3389/fpls.2016.01968>
- Zhu, L., Wasey, A., White, S.H., Dalbey, R.E., 2013. Charge composition features of model single-span membrane proteins that determine selection of YidC and SecYEG translocase pathways in Escherichia coli. *J. Biol. Chem.* 288, 7704–7716.
<https://doi.org/10.1074/jbc.M112.429431>
- Zhu, Y., Zhang, Y., Luo, J., Davies, P.J., Ho, D.T.H., 1998. PPF-1, a post-floral-specific gene expressed in short-day-grown G2 pea, may be important for its never-senescing phenotype. *Gene* 208, 1–6. [https://doi.org/10.1016/S0378-1119\(97\)00613-6](https://doi.org/10.1016/S0378-1119(97)00613-6)
- Zybailov, B., Rutschow, H., Friso, G., Rudella, A., Emanuelsson, O., Sun, Q., van Wijk, K.J., 2008. Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. *PLoS One* 3, e1994. <https://doi.org/10.1371/journal.pone.0001994>

EcYidC 1 MDSQRNLTIVIALLFVSMIWAQWEQDKNLPQPAQQITQTITTAAGSADQVPA-SGQCK
AtAlb3_Arabidop 1 MAK--VLMV--SPS-----SFFGSPLIKPSSSLRHSQ-VGGGTAQFLPYRSMNKK
MdAlb3.1_Malus 1 MAK--TLIS--ST-----PFVGTSL--PSVSLRGS-----YTLPYRS--NG
NtAlb3_Nicotian 1 MAK--TLIS--SP-----SFFGTPIL--PSFSRH-----IIPH--RRR

EcYidC 60 LISVKTIVLLDLTINTRGGDVEQALLPAYPKELNS-TQPFQILF--TSQFIYQAQSGLTGR
AtAlb3_Arabidop 47 LPTTSTTVR-----FSLNR--LPPFHGLD--SSVD-----
MdAlb3.1_Malus 34 --LAVTRIR-----FSLHESVPPINPDEHSEVD-----
NtAlb3_Nicotian 29 --LISARVK-----PSFHE--LPPHOSIH--SSVD-----

EcYidC 118 DGPDPNPANGPRPLYNVEKDAYVLAEGQNELQVPMTYTDAAGNTFHTFVFLKRGDYAVNVN
AtAlb3_Arabidop 73 -----IGSIFTRAP-----
MdAlb3.1_Malus 60 -----VASTLSRAP-----
NtAlb3_Nicotian 53 -----FNAVISRAP-----

EcYidC 178 YNVQAGEKPLEISSFGQLKQSITLPPHLDTGSSNFALHTFFCAAYSTPEEYKDYKFDT
AtAlb3_Arabidop 82 -----SLLYTIADA AV--VGADS-----
MdAlb3.1_Malus 69 -----GLLYTIADA AV--A--VDE-----
NtAlb3_Nicotian 62 -----GLLYTIADA AV--A--ADP-----

EcYidC 238 IADNENL--N--SSKGGWVAMLQQYFATAWI PHNDGTNNFYTANLGNIAAIGYKSQPVLV
AtAlb3_Arabidop 98 VVTDSS--AV-----
MdAlb3.1_Malus 84 SSA--DA--AA-----
NtAlb3_Nicotian 77 QVADAAIAATP-----

Periplasmic/Lumenal Helix

EcYidC 296 QPGQTGAMNSTLWVGPEIQDKMAAVAPHLDLTVDYGLWLFISQPIFKLKLWHSF----
AtAlb3_Arabidop 107 --QRSGGWFGFISDAMEVVLKILKDGLSAVH-----
MdAlb3.1_Malus 91 --QRNGGWFGFISDAMEVVLKILKCGLDVAVH-----
NtAlb3_Nicotian 88 --QRSGGWFGFISDAMEVVLKVMKDGLA AVH-----

Transmembrane Helix 1

EcYidC 351 --VGMGSESTIITFIVRGIMYPLTRACYTSMARKRMIPKIQAVRERLGDDKQRISQEMM
AtAlb3_Arabidop 136 VPYRYGFATILLTVKAAATYPLTKQOVESTLAMONLQPKIKAIQORYAGNQERIOLETS
MdAlb3.1_Malus 120 VPKSYGFATILLTVIVKATLPLTKQOVESTLAMONLQPKIKAIQORYAGNQERIOLETS
NtAlb3_Nicotian 117 VPKSYGFATILLTVIVKATLPLTKQOVESTLAMONLQPKIKAIQORYAGNQERIOLETS

Transmembrane Helix 2

EcYidC 410 ALYKREKVNPLGCGFFLIQVPIFLALYVMGSGV--ELRQAPFALWHDLSA-----
AtAlb3_Arabidop 196 RLYKQASVNPPLAGCPPTLATIPVWIGLYQALSNVANEGLLTEGFFWIPSLGGPTSAARQ
MdAlb3.1_Malus 180 RLYKQAGVNPPLAGCPPTLATIPVWIGLYQALSNVANEGLLTEGFFWIPSLGGPTSAARQ
NtAlb3_Nicotian 177 RLYKQAGVNPPLAGCPPTLATIPVWIGLYQALSNVANEGLLTEGFFWIPSLGGPTSAARQ

Transmembrane Helix 3 TMH4

EcYidC 461 -----QDFVYLPILGVTMFP--IQKMSFTTVIDPQOKI---
AtAlb3_Arabidop 256 SGSGISWLPFFVDGHPPLGWHDTVAYLVLVLLIASQYVSMEMKPPQTD DPAQKNTLLV
MdAlb3.1_Malus 240 SGSGISWLPFFVDGHPPLGWHDTVAYLVLVLLIASQYVSMEMKPPQTD DPAQKNTLLV
NtAlb3_Nicotian 237 SGSGISWLPFFVDGHPPLGWHDTVAYLVLVLLIASQYVSMEMKPPQTD DPAQKNTLLV

TMH4 Transmembrane Helix 5

EcYidC 495 MTFMPVIFTVERLWFPSGLVLYIVSNLVIITQQQLIYE-----
AtAlb3_Arabidop 316 PKFLPLMIGYFSLVSPSGLSIYWFNTNVLTAQVWLRKLGGAKEVFNENASGIISAGRA
MdAlb3.1_Malus 300 PKFLPLMIGYFSLVSPSGLSIYWFNTNVLTAQVWLRKLGGAKEVFNENASGIISAGRA
NtAlb3_Nicotian 297 PKFLPLMIGYFSLVSPSGLSIYWFNTNVLTAQVWLRKLGGAKEV VSGDASGIISAGRA

EcYidC 534 -----GLERKGLHSREKSKS-----
AtAlb3_Arabidop 376 KRSLAQPDAGETPROLKECEKRSKKNKAVAKTVEVVEESQSSSE-----GSDDEEFA
MdAlb3.1_Malus 360 KRSLSQVVEAGSRFRKLEERKSKKQLSKALTNBEVQT-----SDSEVGPNEESNDKGEV
NtAlb3_Nicotian 357 KRSLSQSVEAGERFRQLKEDEKSKKSKALPTDVEEII--SASTSDSELEADDEIKKDEEV

EcYidC -----
AtAlb3_Arabidop 432 REGLASSITSKPLPE--VGGRRSKRSKRKRIV--
MdAlb3.1_Malus 415 LEEVYGSV--GKELPNDPFRRSKRKRKRADGP
NtAlb3_Nicotian 416 LEEYASSS--SKEVPNYSQPRKRSKRKRRAV--

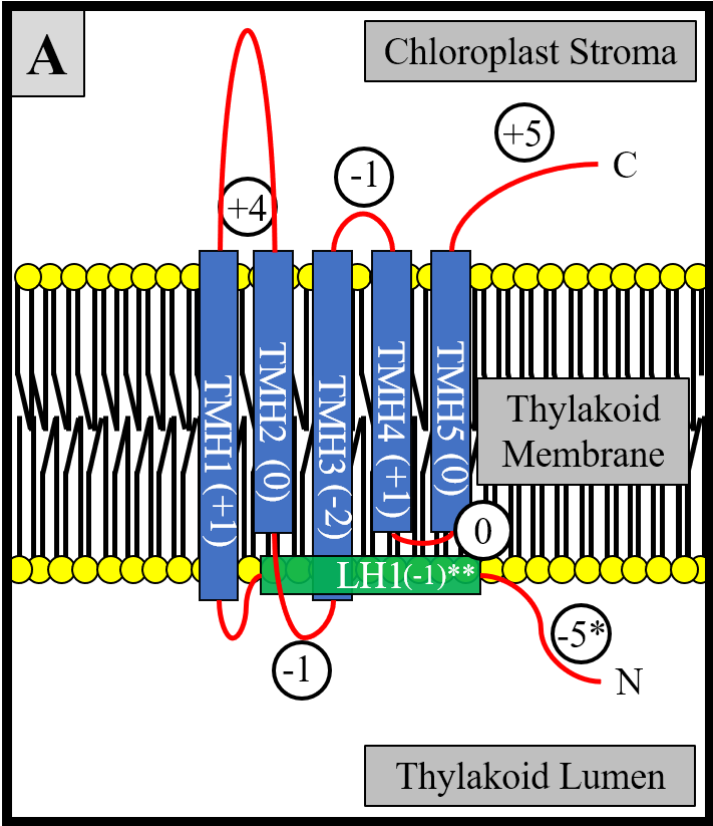
Figure 1: Multiple Sequence Alignment of Alb3 against YidC. The position of the periplasmic/lumenal amphipathic helix is indicated in yellow, while transmembrane helices are indicated in red. Both types of secondary structure are inferred by alignment against EcYidC and BhYidC, for which crystal structures have been solved. TMH: transmembrane helix.

		Conserved Region	↓ 1	Conserved Region
AtAlb3_Arabidop	1	MARVLISSPSSFFGSPILKPPSS	LRHSGVGGGGTAQ	PIPYRSNNKLFETSTTVR
BnAlb3_Brassica	1	MARVLISPPSSFFGSPILKPPSS	SRHVA	-----G---GSLQFRN-KHLV
MdAlb3.1_Malus	1	MAKTLISSTPFVGTSLPSVS	LRGS	-----YTLPYRS---NGLA-TTRLR
MdAlb3.2_Malus	1	MAKTLISSTPFVGTSLPSVS	RRGS	-----YALPYRS---XGLA-TTRLR
PmAlb3_Prunus	1	MAKTLISSTPFVGTSLPTIS	RRGA	-----YTLPYRS---GNGLV-STRLS
FvAlb3_Fragaria	1	MAKTLISSTPFVGTSLPSLS	RRVS	-----YTLPYHRA-GNGLV-ATRLK
VvAlb3_Vitis	1	MARTLISSP-PFIGKPL-PSLS	SRHGL	-----LHSL-PHR-RL-T-STRVK
PtAlb3.1_Populu	1	MARTLISSP-PFIGTPT-PSLS	RHA	-----LFTNR-RE-T-STRLK
PtAlb3.2_Populu	1	MARTLISSP-PFIATSL-PSLS	RHT	-----LFTNR-RE-T-STRLK
RcAlb3_Ricinus	1	MAKTLISSP-SFIGAPT-PSSS	SRHGL	-----QHSLPSSR-RE-ISTTKVK
TcAlb3_Theobrom	1	MARTLISSQ-PFIGTPL-PSK	ISHHG	-----LYTL-PHR-RL-V-STRVK
CsAlb3_Cucumis	1	MAKTLISSP-PFVGSST-PSLS	RHLP	-----LHTL-PHR-RH-T-TTRVN
CmAlb3_Cucumis	1	MAKTLISSP-PFVGSST-PSLS	RH	-----V-PLR-RH-T-TTRVN
CtAlb3_Citrus	1	MAKTLISSP-PFVGTPL-PSFSSLSRHG	LHP	-----LHPT-PNR-RLA-STRVK
SlAlb3_Solanum	1	MAKTLISSPSSFIGTPL-PSLS	RHV	-----F-HRR-RL-T-STRVK
MtAlb3_Medicago	1	MAKTLISSP-SFIGTPT-PSIH	RNF	-----S-PNR-TR-T-YTKVH
GmAlb3.1_Glycin	1	MAKTLISSQ-SFIGTPL-PSLP	RHH	-----L-PHRT-RL-V-ATKVI
GmAlb3.2_Glycin	1	MAKTLISSP-SFIGTPL-PSLP	RHH	-----L-PHRT-RE-V-ITKVK
PsAlb3_Pisum	1	MAKTLISSP-SFIGTPL-PSIH	RTE	-----S-PNR-TR-T-FTKVC
NtAlb3_Nicotian	1	MAKTLISSP-SFIGTPL-PSFS	RHI	-----F-HRR-RL-T-SARVK

		CR2	↓ 2	Conserved Region	↓ 3
AtAlb3_Arabidop	56	FSLNE-IPPPHG-LD	SSVDIGSITRAESLLYT	LADAADV	GVGADSV-----VTI---DS
BnAlb3_Brassica	42	FSLNE-IPPPHH----	GSVDIGAILTRAESLLYT	VADAADV	SGAADS-----AVS---TD
MdAlb3.1_Malus	41	FSLHDSVPPINP-FDHS	PVDVASLISRAEGLLYT	LADAAVA	VDPSS-----S---AD
MdAlb3.2_Malus	41	LSLHDSVPPINP-FDHS	AVDVASLISRAEGLLYT	LADAAVA	VDPSS-----S---TD
PmAlb3_Prunus	42	FSLHDVPPINP-FD	SGSDVASLISRAEGLLYT	LADAAVA	VDPAAASG-----S---AD
FvAlb3_Fragaria	43	FSLHD-VPPINP-FD	SGVDLSALYTRAEGLLYT	LADAAVA	VDPESVSG-----A---GD
VvAlb3_Vitis	41	FSLHD-IPPPHS-LD	SSSIDFAGIVSRAESLLYT	LADAAVS	ADPAAGP-ASGT---AD
PtAlb3.1_Populu	37	LSLHDNIPPIHHHLH	SSVDENTISRAEGLLYT	LADAAVA	VDSAAST-----ISSDIA
PtAlb3.2_Populu	37	LSLHDNIPPIHHHLH	SSIDENISISRAEGLLYT	LADAAVA	VDSAAST-----T---ST
RcAlb3_Ricinus	43	FSLHE-IPPPHH-LD	SSVDENISISRAESLLYT	LADAAVA	VDSAA-----TD
TcAlb3_Theobrom	42	LSFNE-IPPPHS-FD	SSDFEQALYKAESELYT	LADAAVA	ADPA-----GS---ID
CsAlb3_Cucumis	40	FSEHO-IPPPHH-FH	SSDFEQALYKAESELYT	LADAAVA	VDSLAAATST---PD
CmAlb3_Cucumis	35	FSEHO-IPPPHS-FH	SSDFEQALYKAESELYT	LADAAVA	VDSVSGGATST---PD
CtAlb3_Citrus	43	LSFOE-IPPPHS-LD	SSIDLNSVSRTESELYT	LADAAVS	LDSAGGA-ASTS---AD
SlAlb3_Solanum	37	FSEHO-IPPPHS-FH	SNDFEQALYKAESELYT	LADAAVA	ADPGVAP---DVT---AA
MtAlb3_Medicago	36	FSEHO-IPPPHS-FH	HSIDVAVFARAEGLLYT	LADAAVT	ADAVT-----ST---TD
GmAlb3.1_Glycin	37	VSLHE-IPPPHS-FH	RNIDFAGIVTRAEGLLYT	LADAAVA	ADPAVAADSAAS---TD
GmAlb3.2_Glycin	37	VSLHE-IPPPHS-FH	HSIDFAGIVTRAEGLLYT	LADAAVA	ADPAVAADTAAS---TD
PsAlb3_Pisum	36	FSEHO-IPPPHS-FH	HSVDLSGIFARAEGLLYT	LADATVA	ADAAA-----S---TD
NtAlb3_Nicotian	36	FSEHO-IPPPHS-FH	SSDFENAVISRAEGLLYT	LADAAVA	ADPGVAS---D-A---IA

		Conserved Region	↓ 4	TMH
AtAlb3_Arabidop	104	S---AVQKSGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
BnAlb3_Brassica	88	P---AVQKSGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
MdAlb3.1_Malus	88	A---AAQKNGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
MdAlb3.2_Malus	88	A---TAQKNGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
PmAlb3_Prunus	91	A---AAQKNGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
FvAlb3_Fragaria	91	AAVDVQKSGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
VvAlb3_Vitis	94	A---AVQKNGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
PtAlb3.1_Populu	90	D---AAQKNGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
PtAlb3.2_Populu	87	D---TAQKNGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
RcAlb3_Ricinus	88	T---AVQKNGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
TcAlb3_Theobrom	89	A---TPQKNGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
CsAlb3_Cucumis	93	T---AVQKNGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
CmAlb3_Cucumis	88	I---TVQKNGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
CtAlb3_Citrus	95	G---ATQKNGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
SlAlb3_Solanum	87	G---TAQKNGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
MtAlb3_Medicago	84	V---TVQKNGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
GmAlb3.1_Glycin	90	A---AVQKSGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
GmAlb3.2_Glycin	90	A---AVQKSGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
PsAlb3_Pisum	82	V---AAQKNGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK
NtAlb3_Nicotian	85	A---TPQKNGGWFGFISDAMEFV	LKLDGLSAVHVPYA	YGFAILLTIVKAAATPLTK

Figure 2: Alignment of the N-Terminal Domain of Plant Alb3 Homologs. Major regions of homology are indicated in gray, while the first transmembrane helix is indicated in red. Sites for each of the truncation constructs generated in our study are indicated with white arrows. Note that for the fourth conserved domain, the truncation construct is slightly before the transmembrane helix because the original transposon-generated MdAlb3 Δ N(tr) caused a nonsense mutation slightly upstream of the first transmembrane helix. CR: conserved region, TMH: transmembrane helix.



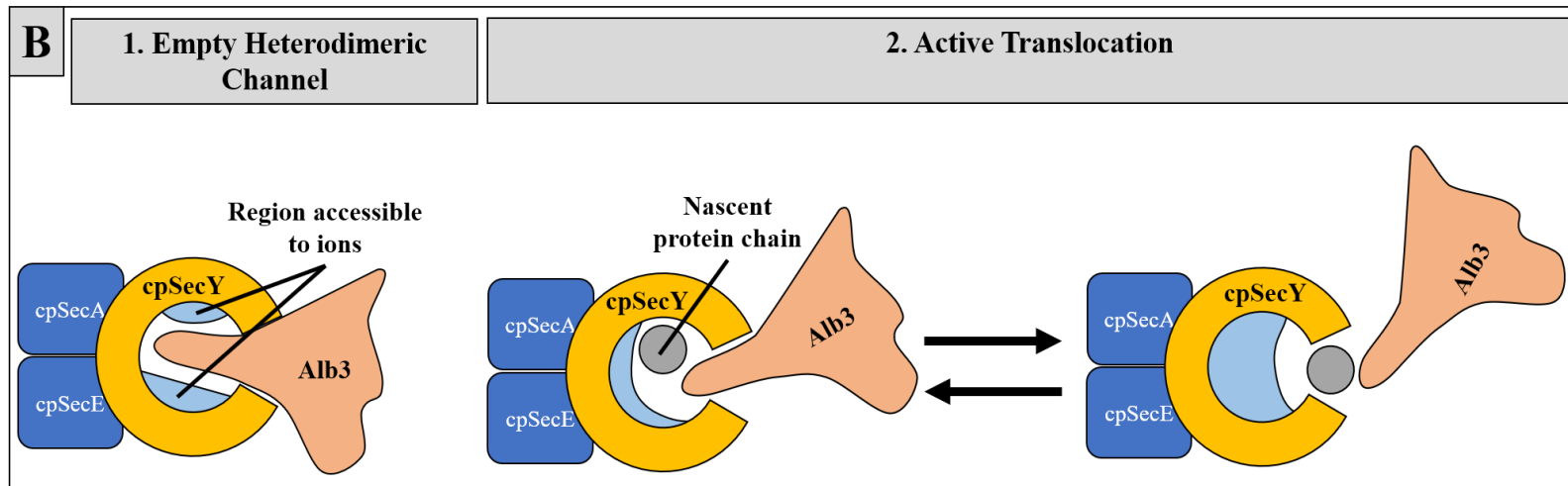


Figure 3: Models of Alb3 Topology and Interaction. (A) Topology as determined by alignment against the bacterial YidC crystal structure and known plant Alb3 homologs. Overall charges of each segment are indicated to support that Alb3 is inserted by heterodimerized Alb3 and cpSecY. *The charge of the N-terminal domain is downstream of the putative SPP cleavage site (position 40 in MdALB3, position 55 in AtALB3), and includes the LH1 charge. **The charge of LH1 is -1 in MdALB3 and neutral in AtALB3. (B) Model of ALB3 and cpSec interaction in thylakoids. When the channel is empty or actively loading a nascent protein chain, the SecY channel is relatively closed to ion permeability. However, as ALB3 moves outward to chaperone a transmembrane segment into the membrane, the channel is left exposed and permeable to ion flux. Adapted from Sachelaru et al., 2017.

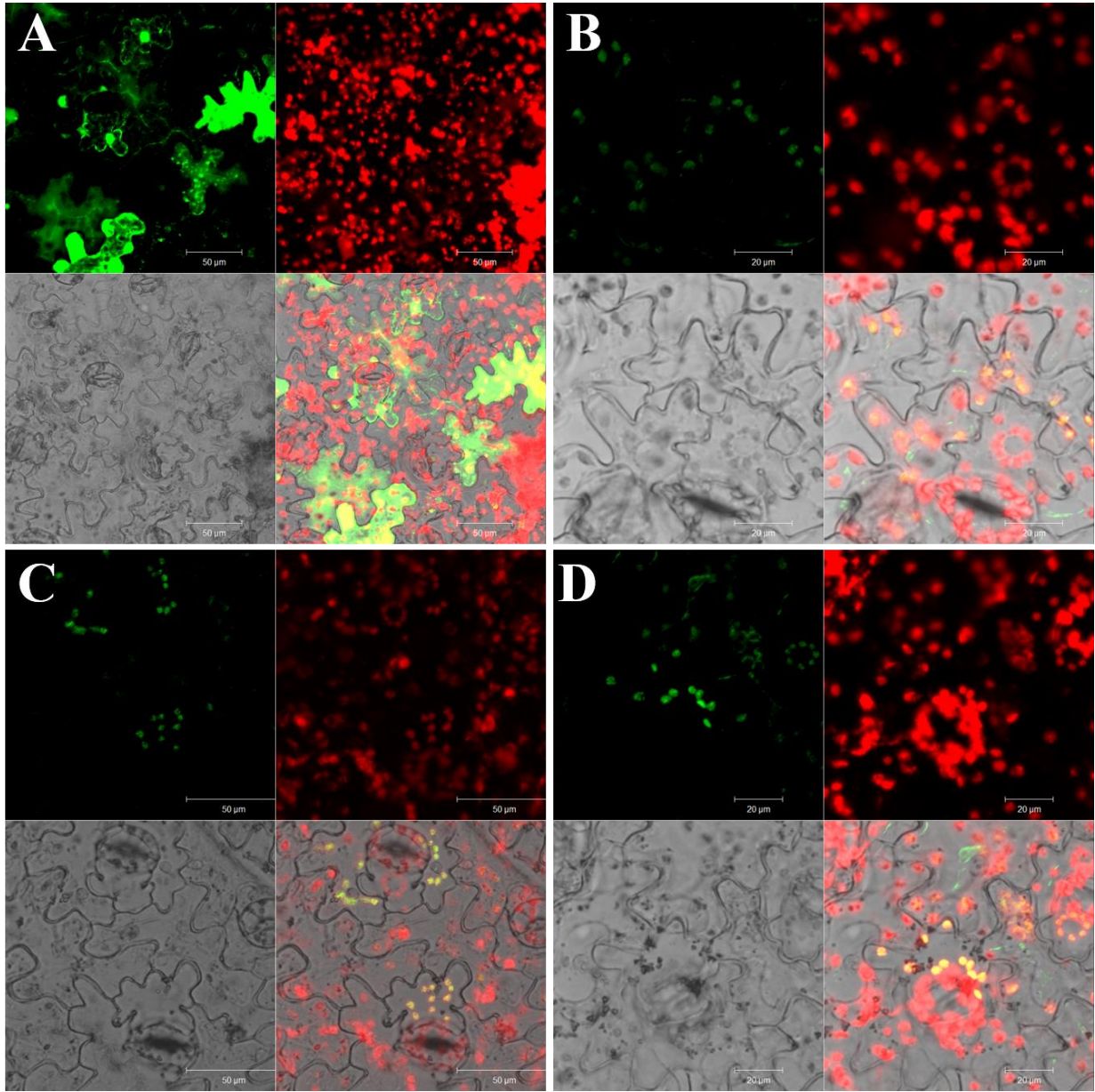


Figure 4: Subcellular Localization of Full-Length ALB3 Homologs. smGFP control (A), MdAlb3.1 (B), MdAlb3.2 (C), and NtAlb3 (D) fusion proteins after agroinfiltration. Quadrants in each inset figure depict (clockwise from top left): GFP fluorescence, autofluorescence of chloroplasts, a hybrid stack, and a brightfield image. Merged images reveal that all three Alb3 homologs are plastid-localized in *Nicotiana benthamiana*, while unmodified smGFP control has signal throughout the cytoplasm and does not accumulate in chloroplasts.

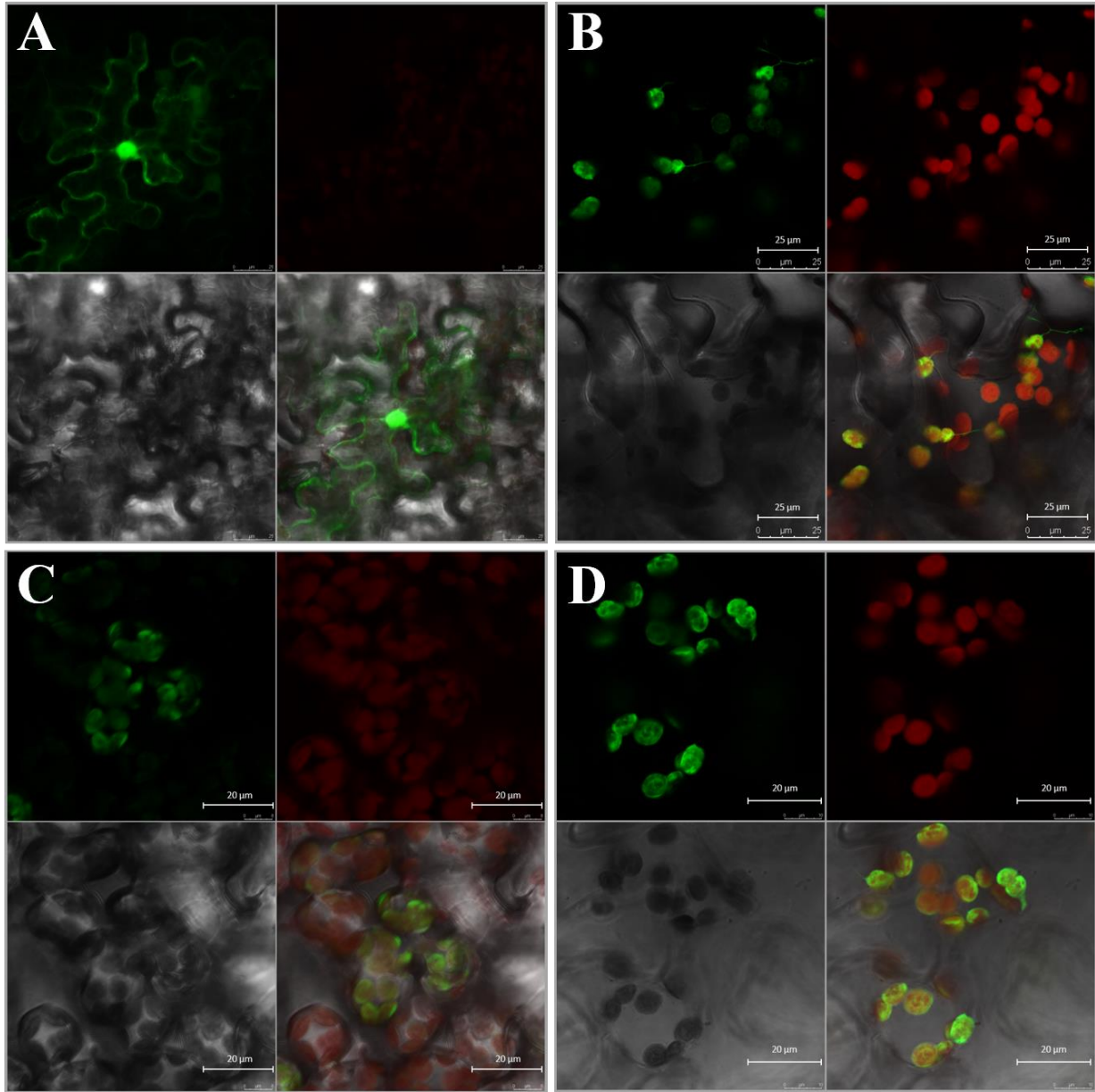
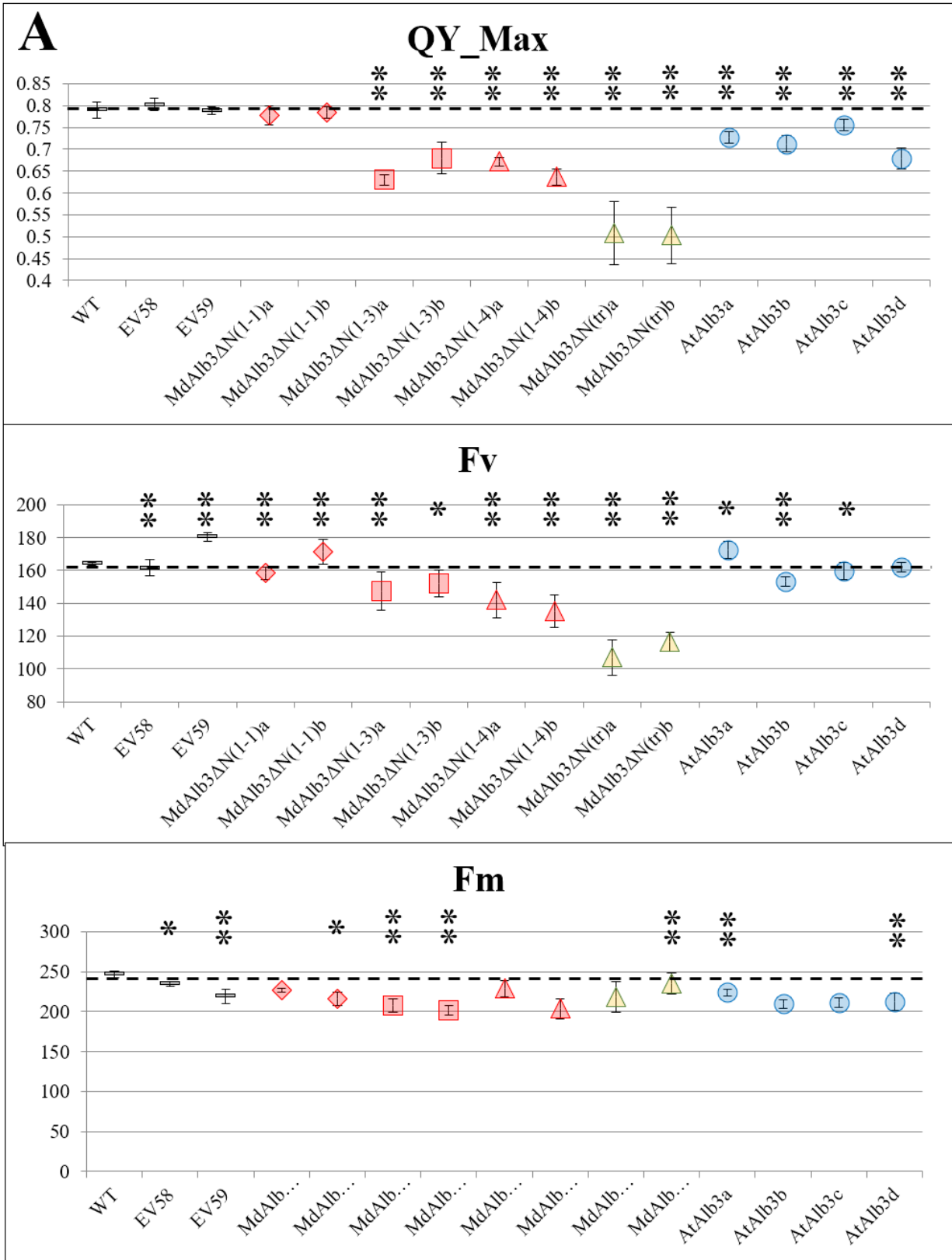
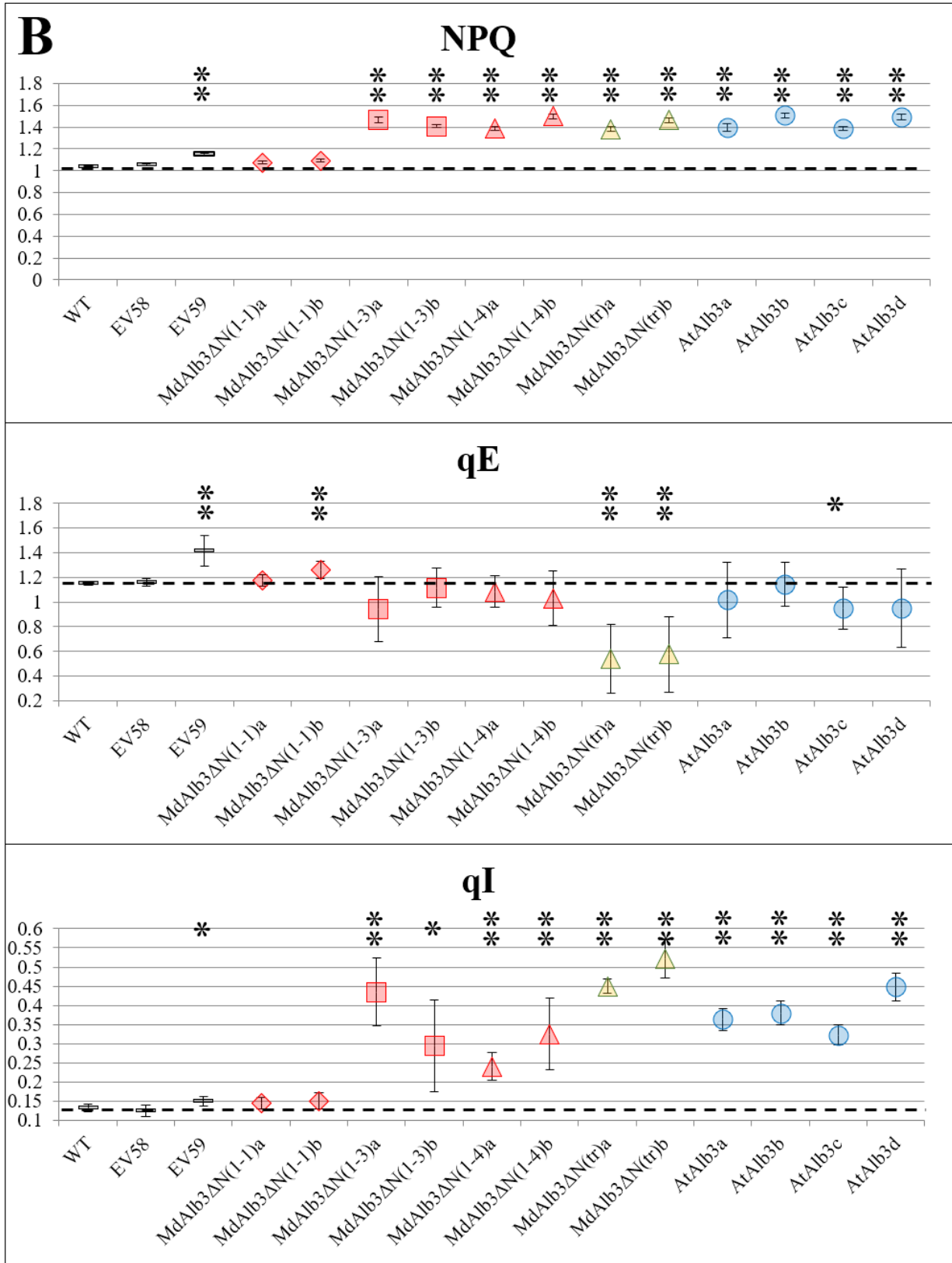


Figure 5: Subcellular Localization of MdAlb3 N-Terminus Truncations. Localization patterns of MdAlb3 Δ N(1-1) (A), MdAlb3 Δ N(1-2) (B), MdAlb3 Δ N(1-3) (C), and MdAlb3 Δ N(1-4) (D) fusion proteins after agroinfiltration. Quadrants in each inset figure depict (clockwise from top left): GFP fluorescence, autofluorescence of chloroplasts, a hybrid stack, and a brightfield image. Diffuse localization in A indicates that MdAlb3 Δ N(1-1) is insufficient for chloroplast targeting, while longer truncation constructs restore chloroplast targeting to GFP.





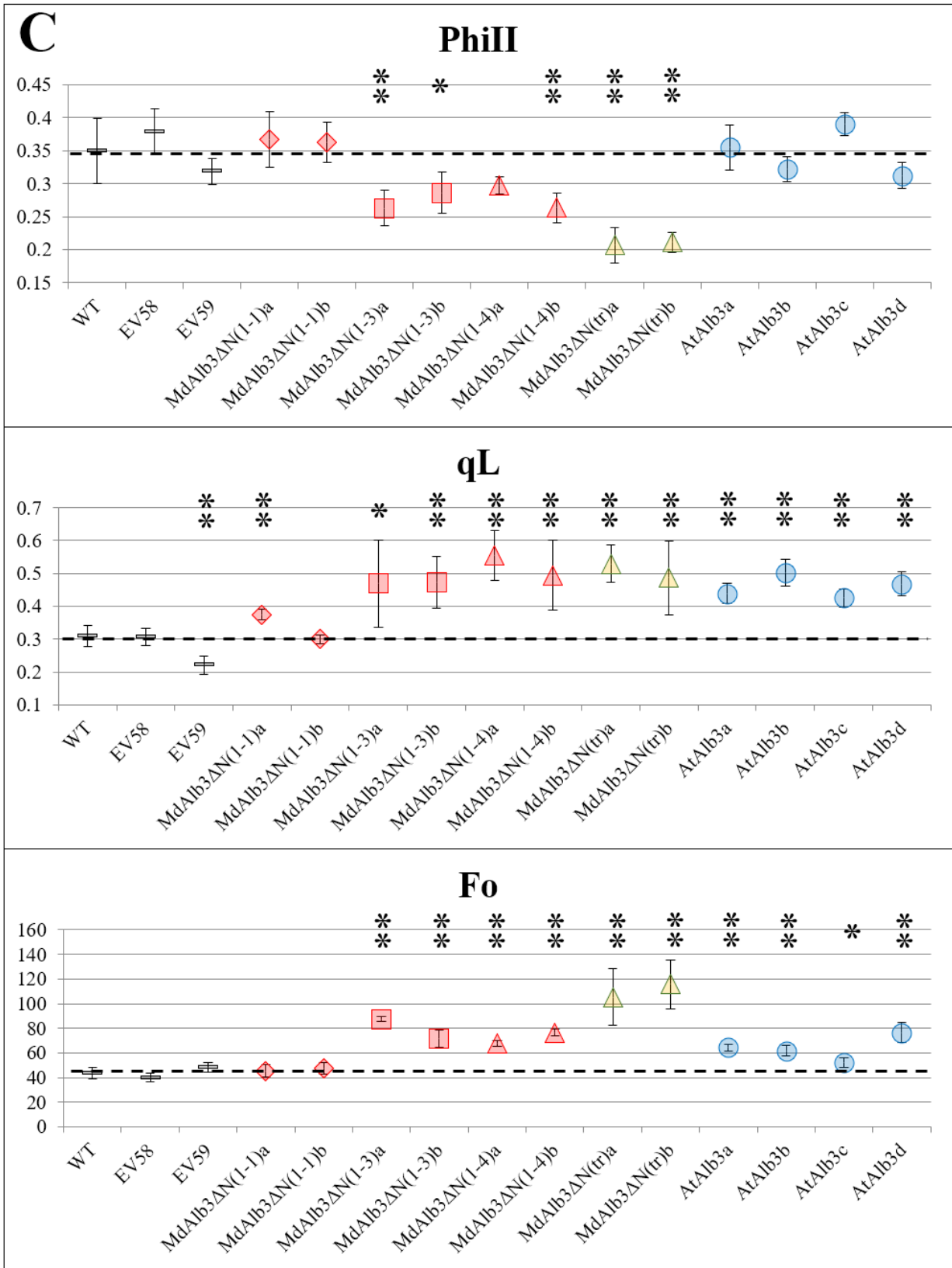


Figure 6: Chlorophyll Fluorescence Parameters of Wildtype and Transgenic Lines. Five plants of each line were monitored for one week and images were analyzed in FluorCam. Each parameter was averaged across all timepoints for each individual plant. Error bars represent standard deviation. MdALB3 Δ N(1-1) showed generally wildtype-like levels of chlorophyll fluorescence, but downstream truncations and full-length AtALB3 exhibited abnormal fluorescence. (A) includes maximum quantum yield of photosynthesis and its components, Fv and Fm. (B) represents nonphotochemical quenching (NPQ) as well as energy-dependent quenching (qE) and photoinhibitory quenching (qI). (C) represents photosystem II operating efficiency (PhiII), fraction of open photosystem II centers (qL), and fluorescence in a dark-adapted state (Fo). Both full-length AtALB3 and truncated MdALB3 lines had altered fluorescence phenotypes in comparison to wildtype, empty vector, and MdAlb3 Δ N(1-1). However, phenotypes were not identical between full-length AtALB3 and truncated MdALB3 lines.

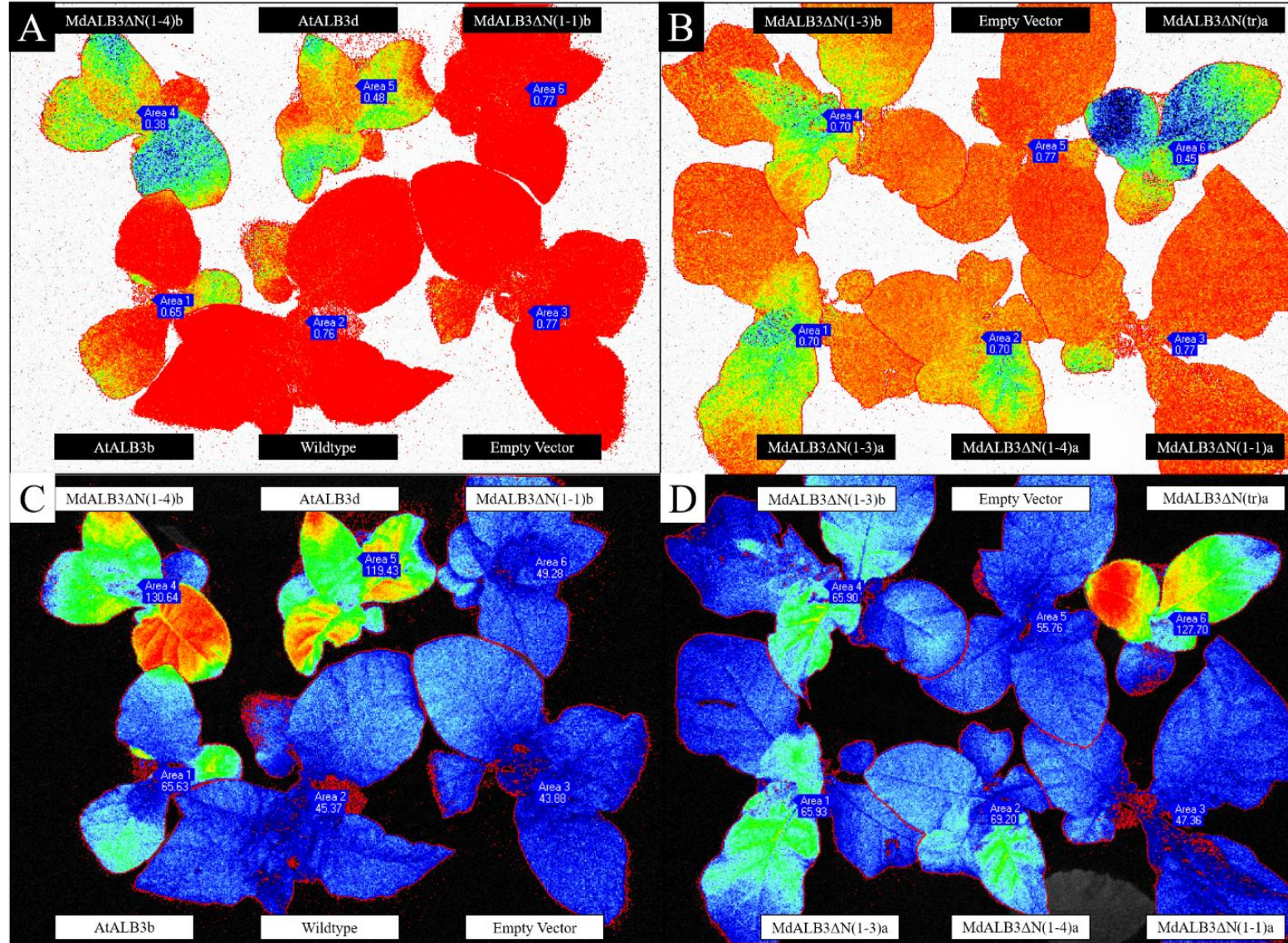


Figure 7: False-Color Chlorophyll Fluorescence Raw Images. Images representing maximum quantum yield F_v/F_m (A-B) and minimal fluorescence in a dark-adapted state F_o (C-D) are represented with 12 individual plants each. Significantly altered fluorescence in AtALB3, MdALB3 Δ N(1-3), MdALB3 Δ N(1-4), and MdALB3 Δ N(tr) were discovered, but empty vector and MdALB3 Δ N(1-1) were indistinguishable from wildtype. Insets A & C and B & D are separate positions within the growth chamber and colorization appears slightly different because images are normalized individually.

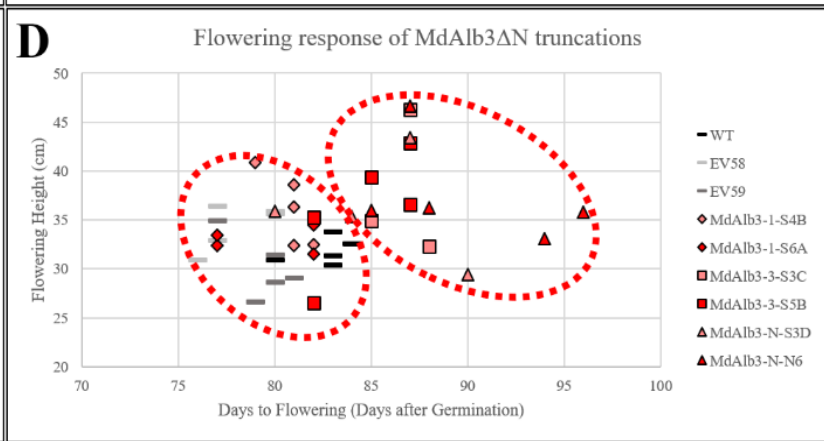
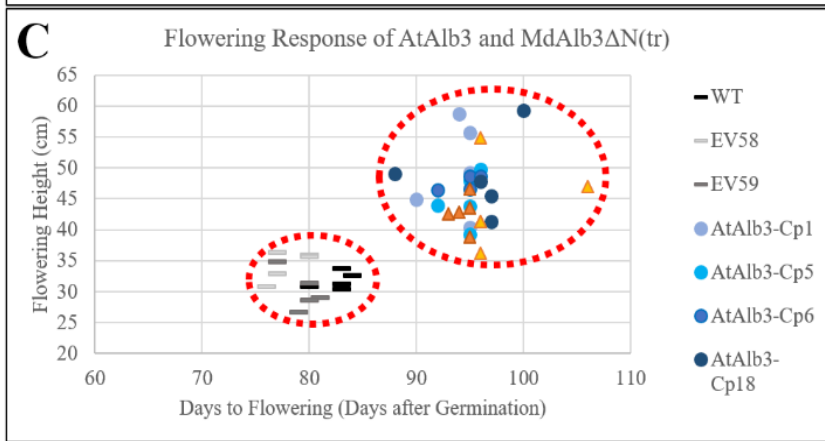
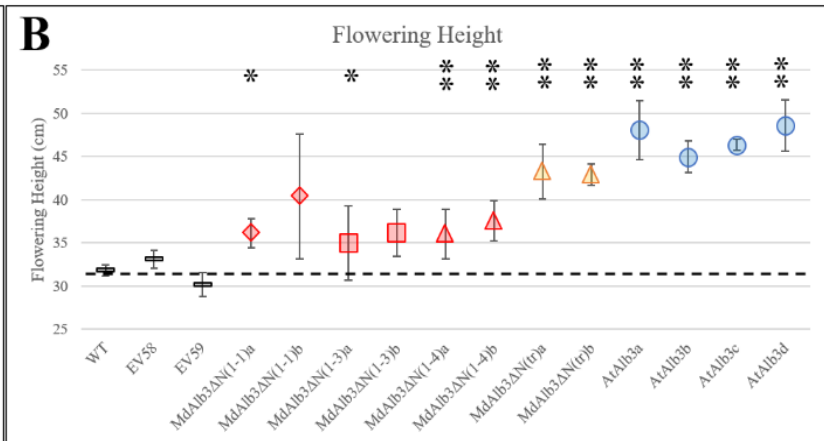
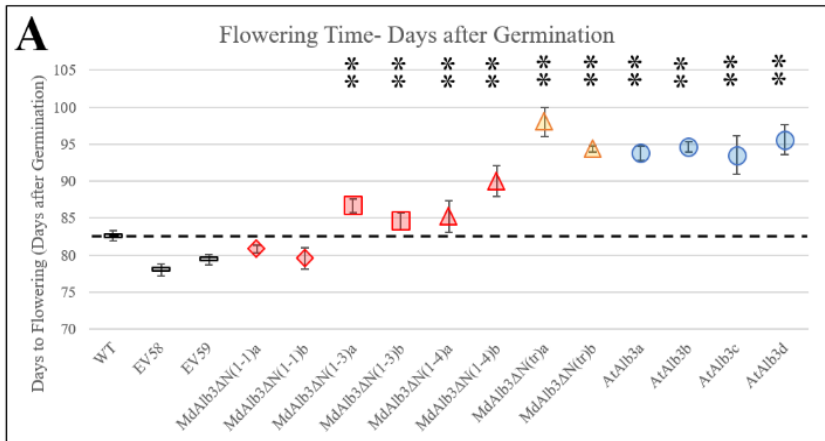


Figure 8: Flowering Phenotype of AtAlb3 and MdAlb3ΔN Constructs. Time to first flowering is depicted in graph A, height at first flowering in graph B, and composites between these variables in graphs C and D. For clarity, the AtAlb3 full protein overexpression lines and MdAlb3ΔN(tr) lines are separated in graph C, while the synthetically-constructed MdAlb3 ΔN lines are represented in graph D. Wildtype-like lines and transgenic lines with aberrant senescence phenotypes are emphasized with red circles. Time and height of flowering is increased by an average of 15 days and nearly 20 cm in the most severe lines (graph C), but MdAlb3ΔN(1-3) and MdAlb3ΔN(1-4) mutations still show significant changes (graph D).

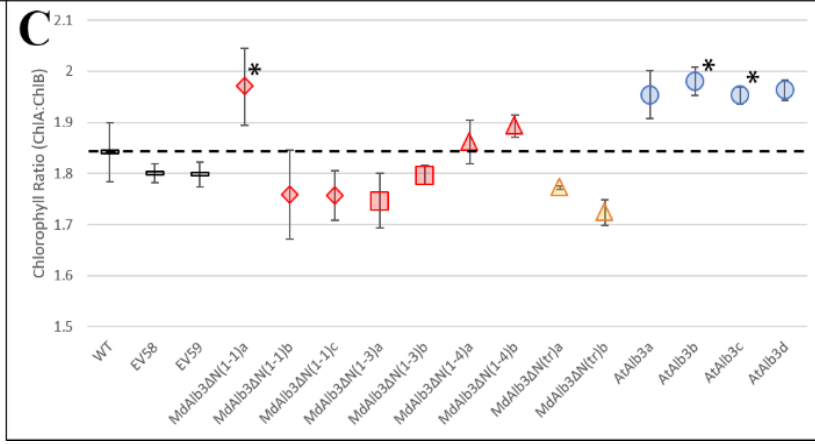
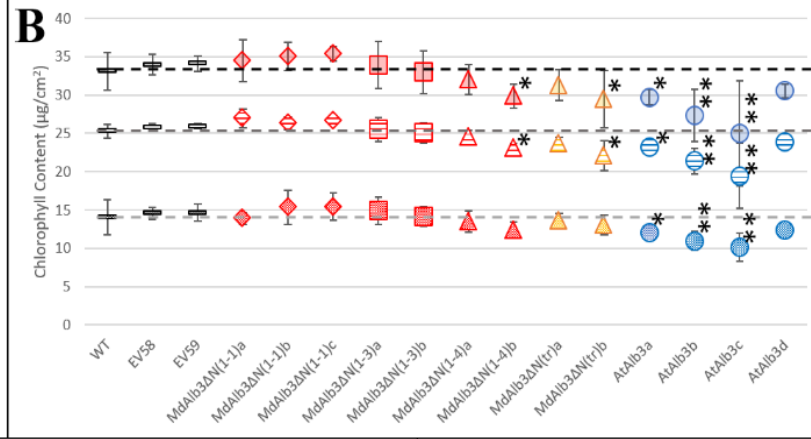
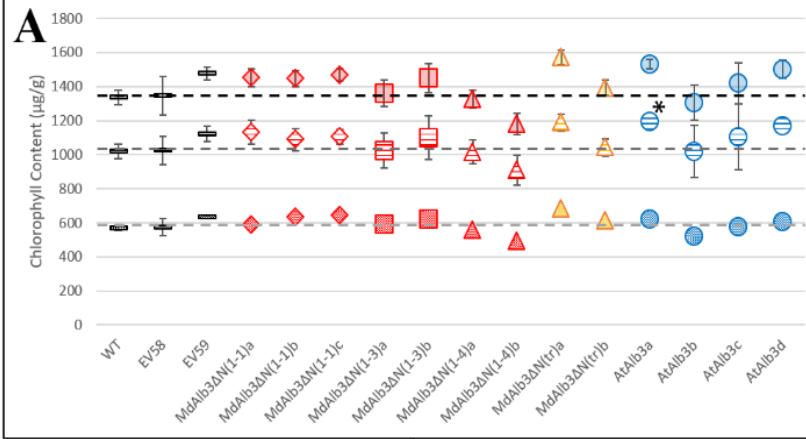


Figure 9: Chlorophyll Quantification and Chlorophyll A/B Ratio. Leaf samples were collected after completion of chlorophyll phenomics experiments and quantified via DMF extraction and spectrometry. Graph A summarizes chlorophyll-A, chlorophyll-B, and total chlorophyll as measured by $\mu\text{g/g}$, graph B summarizes these variables as measured by $\mu\text{g/cm}^2$, and graph C indicates ratio of chlorophyll-A to chlorophyll-B. In Graphs A and B, solid shading (top series) indicates total chlorophyll, while striped shading (middle series) indicates chlorophyll B and dotted shading (bottom series) indicates chlorophyll A. Chlorophyll was largely unchanged in MdAlb3 Δ N truncation lines but was significantly lower in AtAlb3 lines. Furthermore, AtAlb3 overexpression resulted in higher chlorophyll-b, indicating an overabundance of PSI and possibly indicating defects in insertion of LHCB proteins.

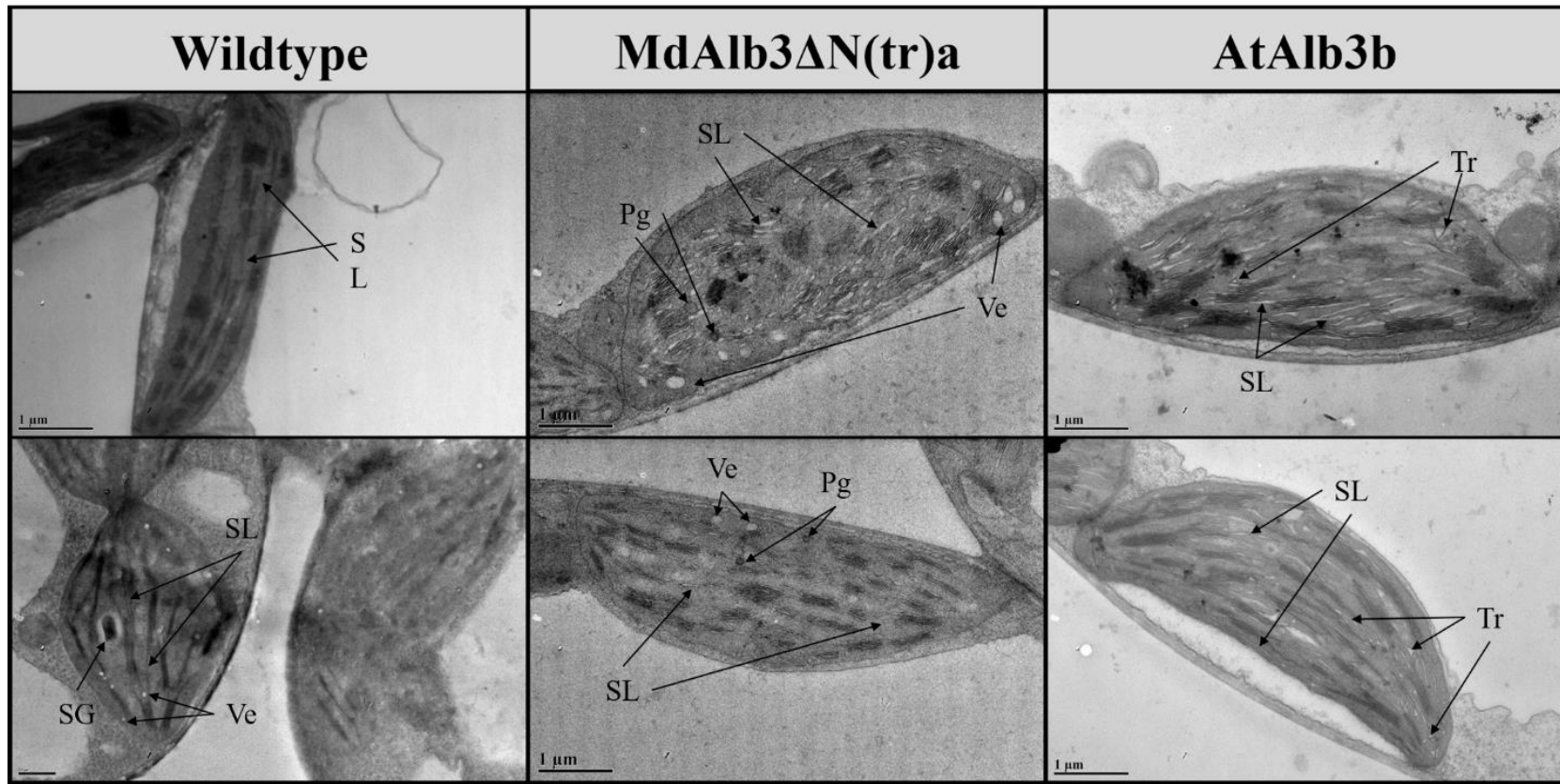


Figure 10: Plastid Ultrastructure of Wildtype, MdALB3 Δ N(tr), and AtALB3

Overexpression Lines. Plastid ultrastructure of palisade mesophyll chloroplasts in young leaves was observed using transmission electron microscopy. Starch granules (SG), stromal lamellae (SL), plastoglobuli (Pg), vesicles (Ve), and triangular thylakoid structures (Tr) are emphasized with arrows. In contrast to the flattened and tightly-appressed thylakoid grana in wildtype, MdALB3 Δ N(tr) exhibited significant swelling of granal and lamellar thylakoids, while AtALB3 exhibited unusually swollen stromal lamellae and blunt triangular structures in thylakoids exposed to the stroma.

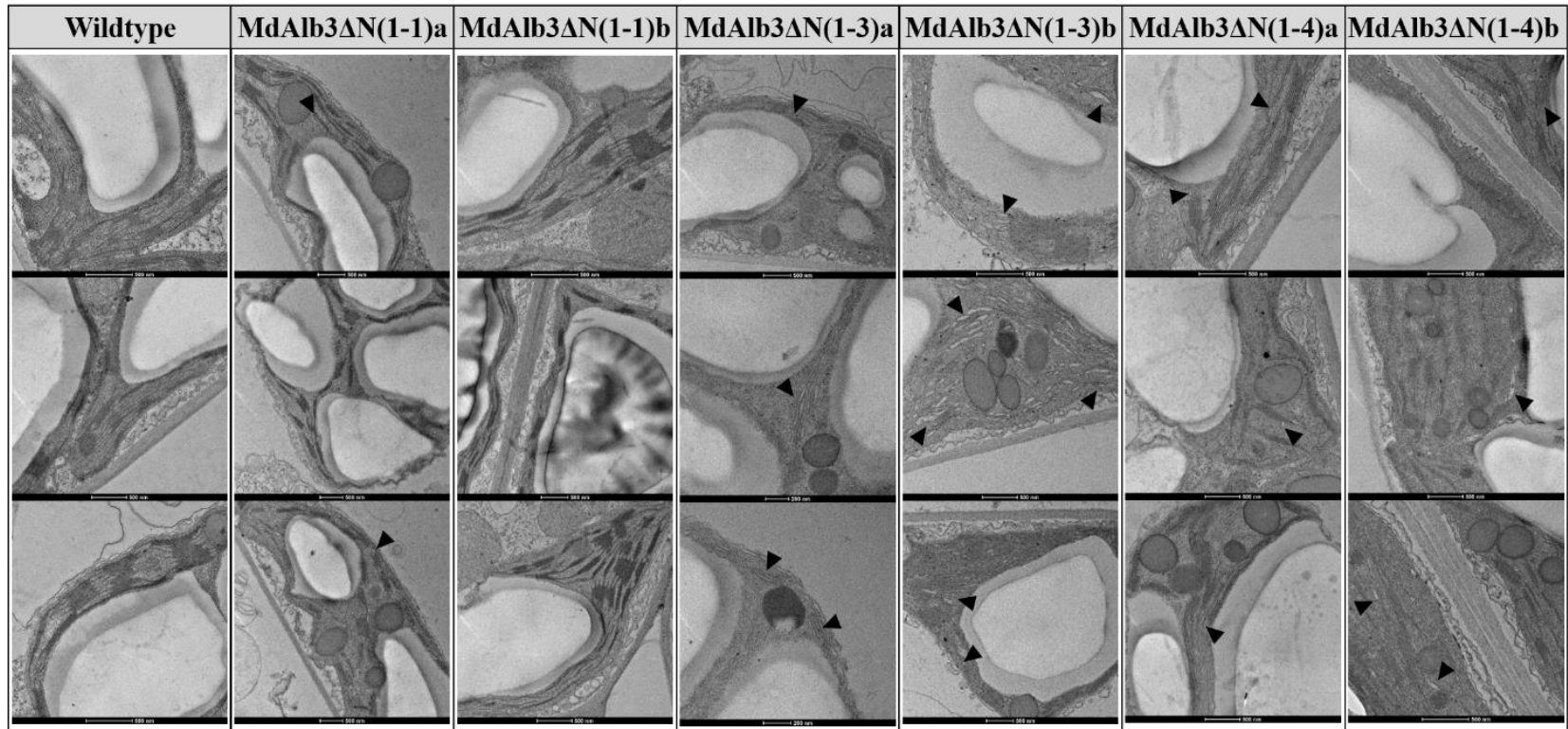


Figure 11: Thylakoid Architecture of Wildtype and MdALB3ΔN Overexpression Lines.

Plastid ultrastructure of palisade mesophyll chloroplasts in young leaves was observed using transmission electron microscopy. Arrows indicate areas of thylakoid swelling consistent with photoinhibition and osmotic stress. In wildtype and MdALB3ΔN(1-1) lines, thylakoid grana were well-stacked and clearly differentiated, with little to no swelling of the thylakoid lumen. In contrast, MdALB3ΔN(1-3) and MdALB3ΔN(1-4) had significantly swollen thylakoid lumens and disaggregation and distortion of both granal thylakoids and stromal lamellae. Even when well-stacked, grana stacks were thinner and comprised fewer layers than in wildtype.

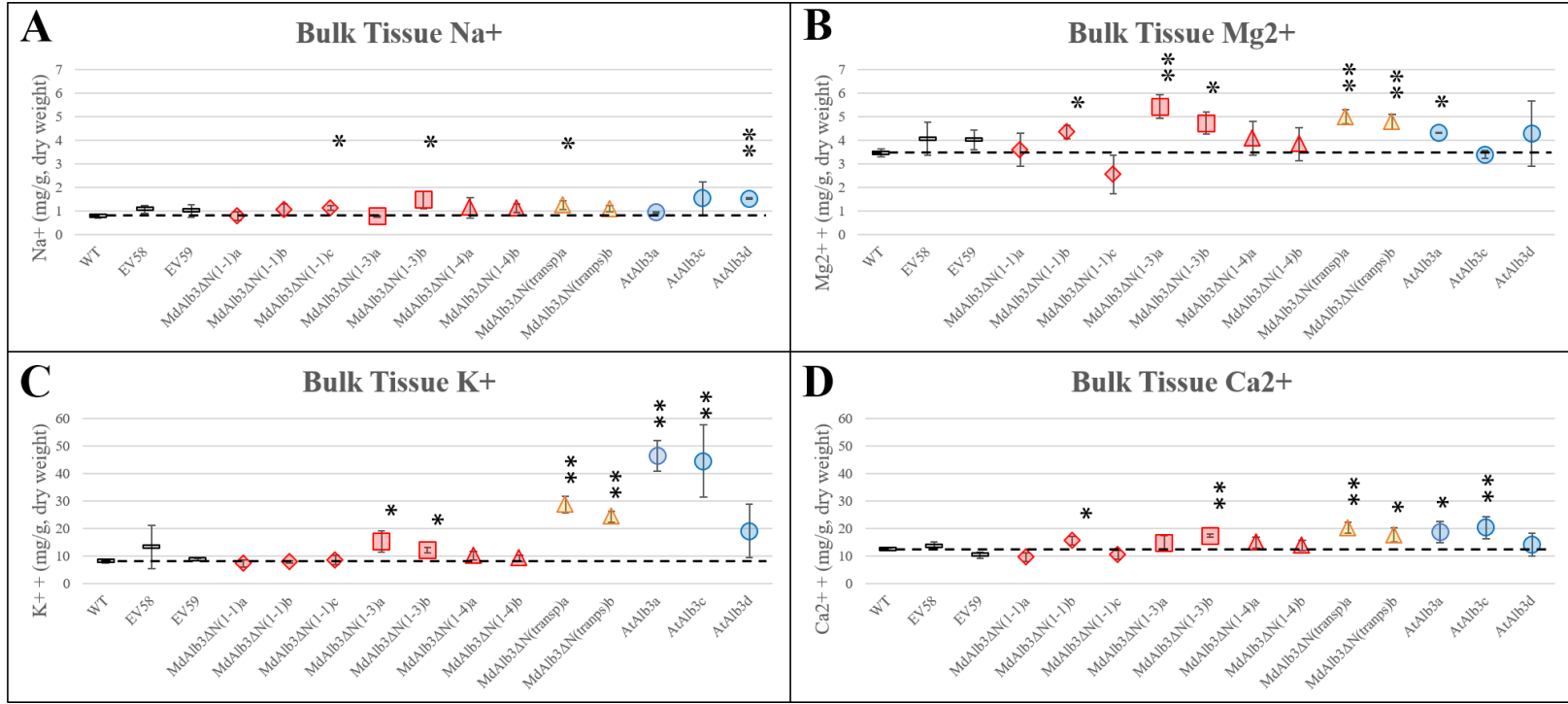


Figure 12: Bulk Tissue Ionomics. Levels of Na⁺ (A), K⁺ (B), Mg²⁺ (C), and Ca²⁺ (D) were determined by ICP-MS analysis of dried, powdered leaf tissue collected after completion of chlorophyll phenomics experiments. Na⁺ was significant in several lines, but the absolute difference was small; all lines were between 1-2 mg/g (dry weight), suggesting that transgenic lines were not experiencing a major defect in ion regulation. Mg²⁺, K⁺, and Ca²⁺ in contrast all experienced significant increases in affected transgenic lines, most notably in MdAlb3ΔN(1-3)a, MdAlb3ΔN(tr), and AtAlb3. Note that scales for graphs A and B are the same, while scales for graphs C and D are approximately 10-fold higher.

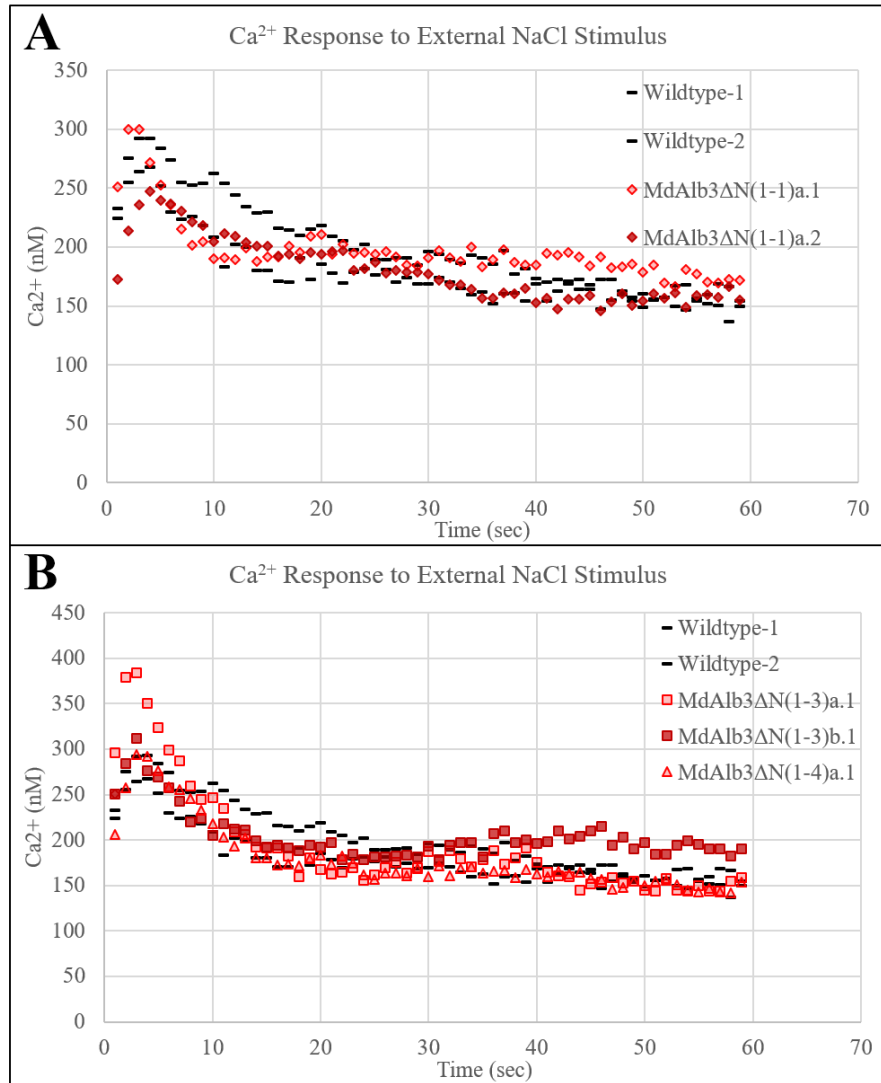


Figure 13: Calcium Response as Measured by Stromal-Targeted Aequorin. (A) wildtype compared to MdAlb3ΔN(1-1), and (B) wildtype compared to MdAlb3ΔN(1-3) and MdAlb3ΔN(1-4). All lines were crossed with the same nuclear-transformed stromal aequorin line. The first conserved domain in the MdALB3 N-terminus did not substantially change calcium dynamics, but peak Ca²⁺ was up to 20mM higher in MdAlb3ΔN(1-3)b and up to 100mM higher in MdAlb3ΔN(1-3)a. Additionally, resting Ca²⁺ was 25-50mM higher in MdAlb3ΔN(1-3)b.

Additional Files

Additional File 1 – Supplemental Figures and Tables

Word document containing additional information. Supplementary Table 1 includes primer sequences utilized for cloning, sequencing, and qRT-PCR. Supplementary figures include a full ALB3 sequence alignment, PCR genotyping, SDS-PAGE gels of protein extracted from wildtype and transgenic lines, stomatal imaging and analysis, microscopy image and analysis, and phenomic analysis of additional transgenic lines.

Additional File 2 – qRT-PCR Analysis of ALB3 Homologs

Excel file summarizing qRT-PCR results of AtALB3, MdALB3.1, and NtALB3 run against representative samples of wildtype, empty-vector, and transgenic lines. Elongation factor-1 α is used as a reference gene control.

SUMMARY AND FUTURE DIRECTIONS

This dissertation establishes a foundation work for expanding the knowledge of plastid biology beyond chloroplasts and which is unconstrained by technical experimental constraints. We describe an updated description of the TOC and TIC translocases of the chloroplast outer and inner envelope membranes, summarize the literature research on both canonical and noncanonical chloroplast transit peptides, and establish a working model for translocation through the canonical route which is responsible for the majority of imported proteins. Comprehensive testing of subcellular localization prediction algorithms revealed a wide range in accuracy for biologically-validated sequences, and also revealed that a combination of TargetP and Localizer achieves 80.7% specificity and 61.1% sensitivity, a significant improvement compared to the use of TargetP alone. These observations are immediately useful to plant researchers aiming to analyze large datasets for plastid localization, or to more accurately predict plastid targeting for single sequences. We also developed two contrasting techniques for gene family prediction, including both an accurate prediction method based on reciprocal best-BLAST hit approaches, and a modified version of UCLUST (Edgar, 2010) which is more efficient and faster at analyzing larger datasets. Application of these methods to the predicted proteomes of 15 higher plant genotypes revealed that roughly 10% of proteins are likely to be plastid-targeted in most plants, and between 628-828 protein families maintain conserved plastid targeting. However, at least 6-fold more plastid-targeted sequences had only semi-conserved or species-unique chloroplast targeting, which likely influences plastid morphology, gene expression, and biochemistry in different species. However, significant work remains to functionally characterize these proteins, starting with biological validation. A current Ph.D. student is currently working on validating the gene expression of potential semi- and non-conserved plastid-targeted proteins

in *Malus × domestica* throughout the growing season, and this work will link gene expression to stages of plastid ultrastructural changes observed in our previous publication “Comparative ultrastructure of fruit plastids in three genetically diverse genotypes of apple (*Malus × domestica* Borkh.) during development” published in Plant Cell Reports. Furthermore, methods development is in progress to isolate chromoplasts and amyloplasts from apple fruit peel with the goal of conducting high-throughput proteomics to biologically validate both protein abundance and plastidial localization. The investigation of evolutionary mechanisms of chloroplast transit peptide evolution in Chapter 4 provides a much-improved model of transit peptide structure and sequence, as well as providing newfound insights into differences between Monocot and Eudicot transit peptides. We found that sequence insertions and deletions were responsible for a majority of novel chloroplast transit peptides, and many of these were the result of alternative start sites. This work provides a framework to begin searching for novel chloroplast-targeted proteins in both in-paralogs and in alternative gene isoforms. Finally, our work in overexpressing portions of the ALB3 translocase from the chloroplast genome in Chapter 5 provided preliminary evidence that thylakoid translocation rate has indirect effects on ion homeostasis, photosynthetic efficiency, and chloroplast ultrastructure. Ongoing work aims to optimize Western blot procedures to more accurately determine the levels of endogenous ALB3. Furthermore, future experiments using electrochromic shift to determine membrane potential and integrity, and Western blotting to examine abundance of thylakoid lumen proteins will continue to test the hypothesis that we have proposed. As we work to replicate this work using other thylakoid transfer domains, we expect to solidify the proposed link between ALB3/PPF-1 and calcium homeostasis as a membrane depolarization phenotype, which could lead to development of crop varieties with delayed senescence or resistance to calcium-related physiological disorders.

Altogether, this dissertation improves on current understanding of plastid proteomics and provides novel tools for improving annotation of high-throughput sequencing data. We also hope that our results will bring greater awareness to the scale of variation of the plastid proteome and lead to renewed research efforts on the plastids of non-model species, as well as more careful interpretation of model systems research.