A Comparison of Orthologous Proteins in Homotypic Vacuole Fusion

---

A Thesis

Presented to

The Division of Mathematics and Natural Sciences

Reed College

In Partial Fulfillment

of the Requirements for the Degree

Bachelor of Arts

Aden J. O'Brien

May 2025

Approved for the Division
(Biology)

Anna Ritz

Acknowledgements

I want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

List of Abbreviations

PPI	Protein Protein Interactions

Table of Contents

Introduction **1**

Chapter 1: Yeast Homotypic Vacuole Fusion **3**

1.1 Components 3

1.2 Stages 6

1.2.1 Priming 6

1.2.2 Tethering & Docking 7

1.2.3 Fusion 8

Chapter 2: Guard Cell Homotypic Vacuole Fusion **11**

Chapter 3: Computational Methods **13**

3.1 Data Collection 13

3.2 Thresholding and Pruning 13

3.3 Network Importance Analysis 16

3.4 Network Visualization 17

Chapter 4: Ortholog Comparisons **19**

4.1 Layer 0 Homologs and Properties 19

4.2 Layer 1 Homologs and Properties 20

4.3 Layer 2 Properties 23

4.4 Discussion 24

4.4.1 Data Availability and Quality 24

4.4.2 Layer Based Analysis 25

4.4.3 Node Importance Metrics 26

4.4.4 Protein Protein Interaction Networks 27

4.4.5 Smith-Waterman Alignment 27

4.4.6 Layer 0 28

4.4.7 Layer 1 28

4.4.8 Layer 2 29

Conclusion **31**

Appendix **33**

Protein Tables 33

S. Cerevisiae 33

A. Thaliana 33

References **35**

List of Tables

List of Figures

3.1 S. Cerevisiae layers 15
3.2 A. Thaliana layers 16
3.3 Homologous edges between A. Thaliana and S. Cervisiae 18

4.1 Layer 0 top 20 Network Importance 20
4.2 Layer 1 top 20 Network Importance 22
4.3 Layer 1 top 20 Network Importance 23
4.4 S. Cerevisiae layer statistics 30
4.5 A. Thaliana layer statistics 30

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

Dedication

You can have a dedication here if you wish.

Introduction

At a cursory glance, vacuoles may seem to be relatively simple organelles, storing water, nutrients, and waste products while lacking intra-organelle machinery in comparison to many of the more 'glamorous' organelles. However, despite their simplicity, vacuoles play crucial roles in various cellular processes across all eukaryotes. These functions include, but are not limited to, waste disposal, maintaining internal cell pressure, structural support, storage for nutrients, ions, pigments, and water, endocytosis, exocytosis, osmoregulation, autophagy, and defense against pathogens.

One such role of vacuoles is the regulation of stomatal opening and closing in the vast majority of terrestrial plants. The intake of carbon dioxide and other atmospheric gasses is essential for photosynthesis in all plants. However, to absorb atmospheric gasses plants must expose their internal structures to the environment which causes water stored within the plant to evaporate. Therefore, plants must carefully balance their CO2 absorption with water availability. Stomata are small mouth shaped pores present on the epidermis of nearly all land plants and are responsible for regulating gas exchange. The term stomata refers to a two part structure consisting of the stomatal aperture and two guard cells which surround the stomatal aperture creating a mouth like appearance. In the stomata's open state, both guard cells are deflated and the stomatal aperture is exposed allowing for the free exchange of atmospheric gasses and water between the plant and its environment. In the stomata's closed state, the guard cells inflate, covering the stomatal aperture and preventing gasses from entering or leaving the plant. In the deflated state, guard cells contain many small vacuoles filled primarily with water. In the inflated state, the vacuoles within guard cells fuse together to form much larger vacuoles, causing the guard cells to swell and cover the stomatal aperture. (write something about how fusion is critical to the functioning of vacuoles)

In contrast to their simplistic internal structure, the process of vacuole fusion is remarkably complex. Consequently, researchers have turned to the study of model organisms in an attempt to better understand the components and signaling mech-

anisms involved in vacuole fusion. *S. Cerevisiae*, also known as brewers yeast, has proven to be an ideal candidate for studying vacuole fusion (specifically homotypic vacuole fusion) due to its extensively studied genome, easy to visualize vacuoles, and our ability to isolate, purify, and store its vacuoles in high quantities. Furthermore, nearly 50 years of extensive research has been devoted to the study of yeast vacuole fusion providing researchers with a strong foundation of research to draw from.

Chapter 1

Yeast Homotypic Vacuole Fusion

1.1 Components

- SNARE Complex
  - SNARE proteins have two separate classification systems based on functionality and structural features respectively. Functional classification splits SNAREs into two groups: v-SNAREs which are localized to the vesicle, and t-SNAREs which are localized to the target membrane. Likewise, structural classification also splits SNAREs into two groups: Q-SNAREs which contain the central amino acid residue glutamine, and R-SNAREs which contain the central amino acid residue arginine. While it is true that R-SNAREs are often also v-SNAREs and Q-SNAREs are often also t-SNAREs, this is not always the case. There are four distinct SNARE proteins present on the yeast vacuole membrane, three Q-SNAREs: Qa (Vam3), Qb (Vti1), Qc (Vam7), and one R-SNARE: R (Nyv1). In their inactive state, the four SNARE proteins form tightly bound bundles known as cis-SNARE complexes along the vacuole membrane.
- Vps1
  - Vps1 is responsible for sequestering individual Qa SNAREs, preventing them from forming stable cis-SNARE complexes along with the R SNARE and the other two Q SNAREs. It is thought that Vps1 polymerizes around the Qa SNARE domain, sequestering a population of Qa SNAREs which are able to undergo HOPS tethering but are unable to bind to other SNARE proteins. Vps1 must then be released prior to trans-SNARE assembly in a process involving Sec18 GTP hydrolysis independent from

Sec17.

- LMA1
  - While Sec18 is bound to the cis-SNARE complex, LMA1 is bound to Sec18. Upon Sec18 mediated ATP hydrolysis, LMA1 is released from Sec18 and is bound to the Qa SNARE, stabilizing the protein.
- Sec17
  - Sec17 is responsible for the recruitment of Sec18 to cis-SNARE complexes. It functions as an intermediary by binding to both Sec18 and the cis-SNARE complex.
- Sec18
  - Sec18 is responsible for ATP hydrolysis leading to the disassembly of the cis-SNARE complex and the transfer of LMA1 to the Qa SNARE.
- Phosphoinositides
  - The primary role of phosphoinositides during the priming stage is the release of monomeric Sec18 from the vacuole membrane. Phosphatidic acid (PA) is responsible for binding and sequestering Sec18 along the vacuole membrane. The release of Sec18 from the membrane is catalyzed by the phosphatidic acid phosphatase Pah1p which converts PA to diacylglycerol (DAG). Once released from the membrane the Sec18 monomers assemble into the Sec18 hexamer which is the active form responsible for cis-SNARE disassembly.
- Pah1
  - A phosphatidic acid phosphatase responsible for the conversion of PA to DAG
- Ypt7
  - A Ras-like GTPase which, in conjunction with the HOPS complex, acts as a tether holding the two opposing membranes in close proximity. Ypt7 is present in both its GTP and GDP forms along the vacuole membrane. While Ypt7 is able to associate with the non-phosphorylated HOPS complex in both its GTP and GDP forms, it can only associate with the phosphorylated HOPS complex in its GTP form.

- HOPS Complex
  - A hexameric protein complex consisting of Vps39,Vps41,Vps33, Vps11, Vps16, and Vps18. The HOPS complex plays two main roles in yeast homotypic vacuole fusion: Vacuole tethering via Ypt7 binding, and organization of SNARE proteins.
- Ccz1–Mon1 Complex
  - A protein complex responsible for the conversion of Ypt7-GDP to Ypt7-GTP. Mon1 is released from the vacuole in an ATP-dependant fashion at which point it dimerizes with Ccz1 throug a currently uncharacterized mechanism.
- PI3P
  - A regulatory lipid which recruits the Ccz1–Mon1 Complex to Ypt7 in its GDP form. PI3P is also responsible for binding the Qc SNARE Vam7 to the membrane and recruiting the HOPS complex.
- PI
  - Phosphatidylinositol is a glycerophospholipid present in the cytosolic leaflet of the vacuole membrane. While it serves no direct function in vacuole fusion, it serves as a precursor lipid for a variety of Phosphoinositides such as PI3P and PI(4,5)P.
- Vps34/Vps15 Complex
  - Vps34 is a PI3 kinase responsible for the conversion of membrane bound Phosphatidylinositol (PI) to PI3P. It forms complexes with Vps15 which acts as a membrane anchor. Once bound to the membrane, Vps34 is activated by its effector protein Gpa1 and converts PI into PI3P.
- Vam7
  - Origially part of the cis-SNARE complex, Vam7 is expelled into the cytosol following cis-SNARE dissassembly. Vam7 is then recruited to the vacuole membrane by PI3P and is partially responsible for HOPS complex tethering.
- V-ATPase Complex

  - A 13 member protein complex split into two sectors. The V0 sector is comprised of VPH1, VMA3, VMA11, VMA16, and VMA6. The V1 sector is comprised of VMA1, VMA2, VMA4, VMA5, VMA7, VMA8, VMA10, AND VMA13. While the V-ATPase complex is commonly implicated in the active transport of protons across the vacuolar membrane, evidence points to a separate role of the V-ATPase complex in homotypic vacuole fusion. Analysis of V-ATPase activity during vacuole fusion has shown that the V0 sectors of V-ATPase complexes on the opposing membranes form inter-membrane V0 complexes. It has been proposed that these inter-membrane V0 complexes act as the initial fusion pores during the process of vertex ring hemifusion.

- Cadmodulin
  - A regulatory protein which responds to Ca+ concentration and is bound to the membrane in a Ca+ dependant fashion. Cadmodulin binds to the V0 sector of the V-ATPase complex and, in response to Ca+, induces the formation of a V0 protein pore.

- VTC Complex
  - A 4 member protein complex comprised of VTC1, VTC2, VTC3, and VTC4. Co-immunoprecipitation studies have shown that the VTC complex binds to the VPH1 protein present in the V0 sector of V-ATPase. This binding activity is theorized to induce conformational changes in the V0 sector and induce the opening of the fusion pore.

1.2 Stages

The process of homotypic vacuole fusion in *S. Cerevisiae* is commonly divided into four general stages: Priming, Tethering, Docking, and Fusion.

1.2.1 Priming

It is important to note that the protein complexes present on the vacuole membrane prior to fusion are often the result of prior fusion reactions. For example, the cis-SNARE complex is assembled during the final step of the fusion process and remains bound in its cis conformation until the next fusion event. Likewise Sec17 is bound to the cis-SNARE complex in the same fashion. Prior to the initation of priming,

Sec18 is bound to the membrane by Phosphatidic acid (PA) and acts as a membrane receptor for the protein LMA1. To initiate priming, PA is converted to Diacylglycerol (DAG) by the phosphatidic acid phosphatase Pah1. This conversion releases monomeric Sec18 from the membrane. The Sec18 monomers then assemble into hexameric complexes and are recruited to cis-SNARE complexes by Sec17. Sec18 binds to Sec17 and hydrolyzes ATP, facilitating the disassembly of the cis-SNARE complex into individual SNARE proteins. During this reaction, LMA1 dissociates from Sec18 and binds to the Qa SNARE (Vam3), providing stabilization. While three of the four SNARE proteins are membrane-anchored, the Qc SNARE (Vam7) is bound only to the other SNARE proteins within the cis-SNARE complex. Upon Sec18 mediated disassembly, the Qc SNARE and Sec17 are released into the cytosol. There is also evidence that bundles of Qa SNAREs exist along the membrane during the priming process. These Qa SNARE groups are sequestered by polymerized Vps1 proteins which prevent Qa snares from associating with other SNARE proteins. Additionally, several phosphoinositides such as Ergesterol and PI(4,5)P2 have been implicated in the priming stage of homotypic vacuole fusion although their functions have yet to be defined.

1.2.2 Tethering & Docking

In the initial stage of the tethering process, both Vps41 (a HOPS complex subunit) and Ypt7 must be phosphorylated by Yck3 and the Ccz1-Mon1 complex respectively. during this step, Vps34 is recruited to the membrane by the membrane bound Vps15. This Vps34/15 complex is then activated by Gpa1 to synthesize PI3P from PI. Following these phosphorylation events, the HOPS complex binds to Ypt7 proteins on the two opposing membranes tethering the two vesicles in close proximity. The HOPS complex shares a binding affinity for both Vam7 and the PI3P which promotes the localization of both components along the vertex ring at which point HOPS-Vam7 binding initiates trans-SNARE assembly. It is important to note that this process not only tethers the two membranes together but also correctly positions the SNARE proteins on the opposing membranes as antiparallel bound SNAREs often lead to fusion inactive complexes. In addition to the numerous regulatory mechanisms above, tethering is also mediated by the dynamin-like Vps1 protein. As discussed in the priming stage, Vps1 sequesters a population of individual Qa SNARE proteins which remain separate from cis-SNARE complexes. Qa SNARE bundles associated with Vps1 can undergo tethering upon contact with the HOPS complex; however, to form

trans-SNARE complexes, Vps1 must be depolymerized by Sec18 to release the individual Qa-SNAREs. While tethering and docking are generally defined as separate processes in yeast homotypic vacuole fusion, the distinction between when tethering ends and docking begins is not clearly delinated.

Docking is said to be completed once trans-SNARE complexes have been established, bridging the two opposing membranes. This process is primarily mediated by the HOPS complex. During the docking process, the HOPS complex protects trans SNARE complexes from Sec18 mediated dissassembly, initiates trans SNARE complex assembly by binding to the Qa SNARE and R SNARE on opposing membranes, proofreads SNARE complexes allowing only correctly assembled SNARES to undergo trans-complex formation, and recruits the remaining Q-SNAREs to its binding site to complete trans SNARE complex assembly. While it has been shown that Qa and R SNARE HOPS association precedes the recruitment of the remaining Q-SNAREs, the mechanism behind the recruitment of the remaining Q-SNAREs and the order of their assembly is currently unknown. (add section about luminal acidification/ Ca+)

1.2.3 Fusion

The final fusion event is perhaps the least understood stage of *S. Cerevisiae* vacuole fusion. Although a number of components have been linked to the final fusion process through genetic deletion screens and co-immunoprecipitation, the mechanisms by which many components regulate the fusion event have yet to be revealed. The final fusion event begins once trans-SNARE complexes have been established between two vacuoles. The assembly of trans-SNARE complexes induces the formation of the vertex ring, a region in which the two vacuole membranes are pressed tightly together forming a flattened boundary. Bilayer and Lumenal Content Mixing is thought to occur via vertex ring hemifusion. Whereas the classical model of hemifusion involves the zippering of membranes proceeding outward from a single fusion pore, the vertex ring hemifusion model proceeds via two separate fusion pores resulting in a fused vacuole containing smaller intralumenal membrane vesicles. These intralumenal membrane vesicles can be observed in *S. Cerevisiae* post fusion which provides strong evidence for the vertex ring hemifusion model as the classical model of hemifusion cannot account for the formation of such vesicles. The activity of the V-ATP complex is one of the several proposed mechanisms by which hemifusion occurs. In response to luminal acidification, the V1 subunit is released from the membrane bound V0 subunit. Concomitantly, in response to Ca+ efflux triggered by trans-SNARE forma-

tion, cadmodulin binds to and triggers conformational change in the V0 subunit. The VTC complex is thought to bind to, and stabilize these primed V0 subunits. Following priming, two V0 subunits on the opposing membranes form trans-V0 complexes which are thought to act as fusion pores.

Chapter 2

Guard Cell Homotypic Vacuole Fusion

While *S. Cerevisiae* homotypic vacuole fusion has been extensively studied, the process is not well characterized in *A. Thaliana* or other plant models. It is known that homotypic vacuole fusion is highly conserved among eukaryotes. However, recent studies have shown that *A. Thaliana* guard cell vacuole fusion may occur through an unknown mechanism, unlike those observed in *S. Cerevisiae* homotypic vacuole fusion. In both *A. Thaliana* and *S. Cerevisiae*, the phosphoinositide PI3P is required for the recruitment of the HOPS complex to the vacuole membrane. However, chemical depletion of PI3P via wortmanin treatment in *A. Thaliana* has been shown to induce rapid guard cell vacuole fusion while the same treatment applied to *S. Cerevisiae* significantly inhibited vacuole fusion. The mechanism underlying this disparity is still being investigated with current research pointing to phosphoproteomic regulation as a possible factor.

Despite the relatively understudied nature of *A. Thaliana* homotypic vacuole fusion, and the observed disparities between *A. Thaliana* and *S. Cerevisiae* vacuole fusion processes, there are a number of conserved elements which have been confirmed to perform similar roles in each organism. One such element is the HOPS complex, a hexamer composed of 6 homologous proteins (VPS 11,16,18,33,39, and 41). Additionally, the YPT7 homolog in *A. Thaliana*, RABG3f, has been shown to be mediated by the MON1-CCZ1 complex and, in turn, mediates downstream HOPS complex interactions.

Homotypic vacuole fusion is highly conserved among eukaryotes. However, while vacuole fusion has been extensively studied in *S. Cerevisiae* models, *A. Thaliana* guard cell vacuole fusion has received comparatively little research. Furthermore,

continuing research into the process of *A. Thaliana* guard cell vacuole fusion indicates a number of inconsistencies between the processes of guard cell vacuole fusion and *S. Cerevisiae* homotypic vacuole fusion. For example, chemical depletion of PI3P via wortmanin treatment in *A. Thaliana* has been shown to induce rapid guard cell vacuole fusion while the same treatment applied to *S. Cerevisiae* significantly inhibited vacuole fusion. PI3P is an essential regulatory lipid in *S. Cerevisiae*, and in both organisms, is responsible for the recruitment of the HOPS complex to the vacuole membrane. This disparity exemplifies the fact that, while proteins may share homologous structure, their regulation and participation in signaling pathways may be significantly different.

Chapter 3

Computational Methods

3.1 Data Collection

All protein-protein interaction data for both *A. Thaliana* and *S. Cerevisiae* were downloaded from the STRING database (`https://string-db.org/`). For each organism, two files were downloaded: protein-info and protein-links. The protein-info files contain the STRING ID of each protein, the proteins common name, the proteins size, and any annotations attached to the protein. The protein-links files contain protein-protein pairs representing individual protein interactions, a number of scoring parameters, and a combined score value which represents the certainty of the interaction. The *A. Thaliana* dataset consisted of 27,463 proteins and 12,557,495 protein-protein interactions while the *S. Cerevisiae* dataset contained 6601 proteins and 2,824,843 protein-protein interactions.

3.2 Thresholding and Pruning

The STRING database uses a number of distinct evidence souces to calculate the confidence score of each protein protein interaction. Several of these sources use homologous protein interactions as evidence to support the possiblily of identical protein protein interactions occuring in the organism of interest. Due to the nature of this study, interaction evidence obtained through the comparison of homologs must be excluded to prevent the possiblilty of data contamination. Fortunatly, the STRING database provides their edge scoring algorithm as a python script which allowed for evidence channels obtained through homology to be removed and for the combined scores to be re-calculated. The homolog-free edgelist was then thresholded, allowing

only PPI's with a combined confidence score greater than 900 to be included in the dataset. This step reduced the *S. Cerevisiae* edge count to 69,784 and the *A. Thaliana* edge count to 49,604. To further isolate proteins which are likely to implicated in *S. Cerevisiae* vacuole fusion, a literature review was conducted with the goal of creating a list of core homotypic vacuole fusion proteins. A total of 45 proteins were identified in the literature review. 43 of the 45 proteins names were successfully converted to STRING IDs, however, PBI2 has no recorded PPI's with a confidence score greater than 900. Proteins without recorded PPI's cause errors when integrating the identified proteins of interest into the thresholded network and PBI2 has thus been removed from the core protein list. It is impractical to expect every PPI invovled in homotypic vacuole fusion to be identified through literature review, therfore, a K-hop algorithm was employed to create a subgraph which likely includes the majority of PPIs involved in the fusion process. The K-hop algorithm requires a set of initial nodes (in this case the 42 core proteins) and a k value (k=2 for the *S. Cerevisiae* dataset) which dictates the number of 'hops' the algorithm performs. This algorithm produces a subgraph which contains the initial nodes and all initial node neighbors that are within k edges from the initial nodes. The *A. Thaliana* dataset was pruned through a similar mechanism. However, due to the lack of literature focusing on *A. Thaliana* guard cell homotypic vacuole fusion, a literature review was insufficient for the purpose of idenitfying a set of core proteins. Therefore, a set of *S. Cerevisiae* homologous proteins was used as the basis for constructing an *A. Thaliana* core protein set. Using the STRING API's 'homology_best?' method, 37 of the 42 *S. Cerevisiae* core proteins were matched with homologous proteins from the *A. Thaliana* dataset. STRING's 'homology_best?' method employs the Smith–Waterman algorithm to perform local sequence alignment and identify likely homologous proteins. This method produces a score which describes the similarity of the two proteins where a higher score denotes a greater similarity between the two proteins. The list of *A. Thaliana* core homologous proteins was thresholded to include only proteins with SW scores greater than 550 and pruned with the K-hop algorithm using parameters identical to those used in *S. Cerevisiae* pruning.

Figure 3.1: S. Cerevisiae layers

Figure 3.2: A. Thaliana layers

3.3 Network Importance Analysis

The steps above resulted in two datasets, one for each organism, each containing three 'layers' consisting of a set of core proteins, the immidiate neighbors of the core proteins, and the neighbors of the immediate neighbors (henceforth referred to as layer 0, layer 1 and layer 2 respectively). All layers were analyzed to determine the degree centrality and clustering coefficient of their constituent nodes. Degree centrality measures the number of edges connecting the node of interest to other nodes within the graph (ie. if a node is connected to 5 neighbors in the graph that node will have a degree centrality of 5). Nodes with higher degree are often

thought to be more 'important' to the process modeled by the network and the distribution of node degrees in a network often occurs in a scale free fashion where a small number of nodes have very high degree centrailty while the majority of the nodes have low degree centrality. This is thought to be somewhat of an evolutionary defense mechanism against genetic mutation as loss of function mutations in proteins with low degree centrality are both more likely to happen and are less likely to cause major disruptions in signaling pathways. Clustering coefficinet measures the interconnectedness of nodes in a network. Individual node clustering coefficients are calculated by counting the number of edges between a nodes neighbors and dividing that number by the number of possible connnections between those neighbors. These individual clustering coefficients can be used as a metric for node importance similar to degree or averaged to determine the overal clustering coefficent of the network. Layer 0 of each organism was also analyzed to determine the proportion of homologous edges (ie. edges in which both proteins are homologous and the edge between them appears in both organisms).

3.4 Network Visualization

Two python packages were used to create visual representations of the *S. Cerevisiae* and *A. Thaliana* PPI networks. Cosmograph was used to create two network visualizations for each organism, one which displays a subgraph containing the initial core proteins and their k-neighbors, and one which displays the aforementioned subgraph within the context of the full STRING PPI network. Pymnet was used to construct a multi-layer network visualization in which each of the two layers represents the PPI's between the core proteins of each organism, and edges between the layers represent homologous proteins present in each organim.

Figure 3.3: Homologous edges between A. Thaliana and S. Cervisiae

Chapter 4

Ortholog Comparisons

4.1 Layer 0 Homologs and Properties

SC Layer 0 represents 0.63% (42/6,600) of the total nodes and 0.47% (166/34,892) of the total edges in the *S. Cerevisiae* dataset. AT Layer 0 represents 0.10% (28/27,462) of the total nodes and 0.14% (35/24,802) of the total edges in the *A. Thaliana* dataset. Layer 0 displays the largest discrepancy in average clustering values with AT layer 0 displaying an average clustering value of 0.35 while SC layer 0 displayed an average clustering value of 0.69. A total of 33 homologous edges were identified between the two layer 0's representing 19.87% (33/166) of the total *S. Cerevisiae* layer 0 edges, and 94.28% (33/35) of the total *A. Thaliana* layer 0 edges.

Analysis of the top 20 layer 0 *A. Thaliana* nodes by combined network importance score shows that the majority of the top scoring nodes in *A. Thaliana* layer 0 consist of proteins belonging to the HOPS complex, SNARE complex, and V-ATPase complex. All 6 HOPS proteins, 3 out of the 4 SNARE proteins, and 9 out of the 13 V-ATPase proteins were represented in the top 20 nodes of *A. Thaliana* layer 0. A comparison of each set of 20 proteins in the respective layer 0's shows that 13 out of the top 20 proteins in *A. Thaliana* layer 0 are homologous to proteins in the top 20 proteins in *S. Cerevisiae* layer 0. Out of the 20 proteins in the *A. Thaliana* layer 0, two proteins were not members of the HOPS, SNARE, or V-ATPase complex. These two proteins (Vps34 and Vps15) are homologous to *S. Cerevisiae* Vps34 and Vps15 which have be shown to form complexes responsible for the conversion of membrane bound PI to PI3P.

SC Combined Scores

	node	Combined Index Score
0	VPS41	4
1	VAM3	6
2	YPT7	6
3	VMA1	6
4	VAM7	8
5	VMA7	9
6	VMA6	9
7	VMA4	9
8	VMA8	9
9	VMA3	9
10	VMA5	9
11	VMA16	9
12	VPS33	9
13	VTI1	9
14	VMA2	9
15	VMA11	9
16	VAM6	9
17	VMA10	9
18	VMA13	9
19	VPH1	9

AT Combined Scores

	node	Combined Index Score
0	VCL1	4
1	VPS11	4
2	VPS39	4
3	VPS33	4
4	VPS18	4
5	VHA-D	4
6	VAMP713	4
7	VPS41	4
8	VHA-H	6
9	VHA-A	6
10	VHA-C	6
11	VHA-F	7
12	VHA-B1	9
13	VHA-d2	9
14	VHA-a3	10
15	VHA-c2	10
16	SYP21	11
17	VTI13	11
18	VPS15	11
19	VPS34	11

Figure 4.1: Layer 0 top 20 Network Importance

4.2 Layer 1 Homologs and Properties

SC Layer 1 represents 3.5% (236/6,600) of the total nodes and 3.5% (1,250/34,892) of the total edges in the *S. Cerevisiae* dataset. AT Layer 1 represents 0.44% (121/27,462) of the total nodes and 1.64% (407/24,902) of the total edges in the *A. Thaliana* dataset. The average clustering coefficients for each organisms layer 1 were similar with *A. Thaliana* displaying a 0.59 clustering coefficient and *S. Cerevisiae* displaying a 0.64 clustring coefficient. A total of 101 homologous edges were identified between

the two layer 1's representing 8.08% (101/1,250) of the total *S. Cerevisiae* layer 1 edges, and 24.81% (101/407) of the total *A. Thaliana* layer 1 edges.

Analysis of the top 20 layer 1 *A. Thaliana* nodes shows a marked difference in comparison to layer 0. All *A. Thaliana* layer 0 nodes in the top 20 are involved in lipid synthesis or modification (ie. phosphorylation or transferase activity). A single protein, Vps34, in shared between the top 20 nodes of each layer 1. Out of the top 20 *A. Thaliana* proteins, 10 have no identified homologs in yeast.

SC Combined Scores

	node	Combined Index Score
0	VAM3	5
1	CDC42	9
2	YKT6	10
3	NYV1	10
4	YPT7	11
5	VTI1	11
6	VPS34	13
7	SEC9	14
8	VPS33	17
9	VPS41	18
10	SEC22	19
11	VAM6	19
12	SED5	20
13	ATG15	22
14	SSO1	23
15	SNC1	24
16	VPS16	24
17	RHO1	25
18	SEC17	25
19	BET1	26

AT Combined Scores

	node	Combined Index Score
0	VPS34	4
1	PAH1	6
2	PAH2	6
3	SDP1L	8
4	SDP1	8
5	PDAT1	10
6	DGAT1	10
7	DGAT3	11
8	LPP2	11
9	LPP1	11
10	PI4KB1	12
11	DGAT2	13
12	FAB1B	14
13	FAB1C	14
14	PDAT2	14
15	FAB1D	14
16	FAB1A	14
17	NPC1	15
18	AAPT1	16
19	SAC7	16

Figure 4.2: Layer 1 top 20 Network Importance

4.3 Layer 2 Properties

SC Layer 2 represents 10.66% (704/6,600) of the total nodes and 9.58% (3,345/34,892) of the total edges in the *S. Cerevisiae* dataset. AT Layer 2 represents 1.46% (401/27,462) of the total nodes and 7.99% (1,983/24,802) of the total edges in the *A. Thaliana* dataset. The average clustering coefficients for each organisms layer 2 were nearly identical with *A. Thaliana* displaying a 0.59 clustering coefficient and *S. Cerevisiae* displaying a 0.58 clustering coefficient. A total of 158 homologous edges were identified between the two layer 2's representing 4.72% (158/3,345) of the total *S. Cerevisiae* layer 1 edges, and 7.96% (158/1,983) of the total *A. Thaliana* layer 2 edges.

AT Combined Scores			SC Combined Scores		
	node	**Combined Index Score**		**node**	**Combined Index Score**
0	SDP1	7	0	ACT1	4
1	PIS1	10	1	CDC42	10
2	NPC4	15	2	ATG1	12
3	NPC5	15	3	LAS17	17
4	SDP1L	15	4	YPT1	18
5	NPC1	16	5	VAM3	20
6	PIS2	18	6	YPT7	21
7	NPC2	18	7	MYO2	21
8	NPC6	20	8	SEC9	22
9	AAPT2	21	9	ATG8	22
10	AAPT1	21	10	RHO1	23
11	VPS34	22	11	ATG11	23
12	PI4KB1	23	12	ATG15	23
13	LPLAT1	25	13	VPS34	24
14	HXK1	28	14	CDC24	27
15	SAC7	29	15	NYV1	27
16	LPLAT2	29	16	ARP4	28
17	PDAT1	29	17	SEC4	28
18	ATG7	30	18	BEM1	29
19	T2K12.9	30	19	SEC22	30

Figure 4.3: Layer 1 top 20 Network Importance

4.4 Discussion

4.4.1 Data Availability and Quality

As with any computation based study, the quality and quantity of data is extremely important. All data used in this study was downloaded from the STRING database which collects both known and predicted protein protein interactions (PPI's). STRING draws from a number of data sources when collecting PPI's. These sources include automated text mining of the scientific literature, computational interaction predictions from co-expression, conserved genomic context, databases of interaction experiments and known complexes/pathways from curated sources. String assigns each PPI a confidence score with higher scores signaling a higher confidence that the two proteins are interacting in vivo. As a single PPI may have multiple sources of evidence, the confidence score is calculated as an average of all evidence channels. However, to account for the probabililty that any two random proteins may interact, a prior value of 0.41 is integrated into each evidence channel score and must be removed and re-applied when calculating the overall confidence score of an interaction. As mentioned above, one of the evidence sources used by STRING to calculate confidence scores is conserved genomic context. Out of the 14 individual evidence channels used to calculate the confidence score, 6 consist of interaction data inferred from homologous PPI's observed in other species. While evidence obtained through homology is by no means inaccurate or invalid, these evidence channels were excluded from the data used in this analysis. As the goal of this experiment is to compare homolgous PPI's present in both *A. Thaliana* and *S. Cerevisiae*, including PPI's inferred through homology introduces potentially confounding data. For example, If a PPI in *S. Cerevisiae* was inferred through an observed interaction between homologous proteins in *D. Melanogaster* , the STRING data would include that interaction in the *S. Cerevisiae* PPI dataset. If those two proteins also interact in the *A. Thaliana*, this interaction would be classified as homologous between *A. Thaliana* and *S. Cerevisiae* despite the data actually being representative of a homologous PPI between *A. Thaliana* and *D. Melanogaster*. Fortunately, STRING provides the script used for calculating confidence scores from individual evidence channels and evidence inferred through homology was easily removed. In addition to removing potentially confounding evidence channels, a thresholding mechanism was employed during the data cleaning process. The threshold was set at a somewhat arbitrary value of 900, meaning that only PPI's with a total confidence score of 900 or greater would be included in the dataset. A threshold of 900 is relatively high in

the context of STRING confidence scores and was chosen with the intention that all PPI's included in the dataset would have strong evidence backing their interaction in vivo. While removing homology derived evidence and thresholding by confidence scores reduce confounding data and increase data quality respectively, these two measures also reduce the total number of PPI's in the dataset. This potentially means removing PPI's that do occur in vivo, yet don't have sufficient supporing evidnece in the literature. This is unfortunately unavoidable, yet must be taken into consideration when comparing the two PPI networks.

4.4.2 Layer Based Analysis

The decision to perform independant analyses of each layer was a natural consequence of the data pruning process. After re-calculating confidence scores and thresholding the STRING data, additional processing was needed to isolate the subnetwork in each oraganisms dataset containing proteins likely to be implicated in the homotypic vacuole fusion process. While there are multiple methods for isolating subgraphs, a neighborhood-based approach seemed to be the best fit for the purpose of this experiment. By expanding the subgraph based on the neighbors of the nodes in the current subgraph, all edges between all nodes in the subgraph are preserved which increases the potential number of observed homologous edges. Additionally, by confining the subgraph to a maximum of two 'hops', the longest distance between any periphrial node and a layer 0 node is two edges, which increases the likelyhood that periphrial nodes are in some way invloved in the homotypic vacuole fusion process. Initially, the process of determining the initial protein set for each organim was identical, perform a literature review and collate a list of proteins which have been experimentally verified to participate in the process of homotypic vacuole fusion. However, after completing *S. Cerevisiae* literature review, it became apparent that a similar process was not feasible for *A. Thaliana*. *S. Cerevisiae* is considered to be the model organism for studing the process of homotypic vaculole fusion and a large body of work is devoted to the study of this particular process in both isolated *S. Cerevisiae* vacuoles and living *S. Cerevisiae* models. Conversely, there is little research devoted to the process of homotypic vacuole fusion in *A. Thaliana* and even less research focused on the precise mechanism of *A. Thaliana* guard cell homotypic vacuole fusion. Therefore, in lieu of compiling a initial protein set for *A. Thaliana* through literature review, an initial set of *A. Thaliana* proteins was inferred by compiling all proteins in the *S. Cerevisiae* initial protein set and selecting those with homologous proteins in *A. Thaliana*. This

decision is justified by substantial evidence pointing towards the highly conserved nature of homotypic vacuole fusion accross all eukaryotes. In theory, if the homotypic vacuole fusion process is highly conserved, then the core homologous proteins in *S. Cerevisiae* should, to an extent, represent proteins highly involved in the *A. Thaliana* homotypic vacuole fusion process.

4.4.3 Node Importance Metrics

While a host of node importance metrics are commonly used in network analysis, several considerations were taken into accound when selecting the three importance metrics used in this experiment. The three selected importance metrics represent three general methods for assessing node importance, degree, clustering, and centrality. Degree centrality represents the number of neighbors connected to each node. Nodes with higher degree are considered to be more important in the context of the network as they represent a higher proportion of the PPI's than nodes with lower degree. To assess the clustering coefficient of each node a standard clustering coefficient algorithm was employed. The clustering coefficient algorithm counts the number of neighbors connected to any one node and uses that number to calculate the maximum number of edges that could exist between each neighbor. The maximum number of edges betweeen neighbors is then divided by the observed number of edges between neighbors to compute the clustering coefficient. Clustering coefficient measures how interconnected a node is within its immediate neighborhood. This metric is useful for identifying biologically relevant subnetworks such as protein complexes as each protein within the complex interacts with most if not all of the other proteins in the complex, increasing the clustering coeffient for all proteins in the complex. Computing node centrality was slightly more challenging due to the *A. Thaliana* network being composed of multiple disconnected components. While degree and clustering calculations are not confined to fully connected networks, many centrality measures assume a fully connected network. Therefore, the harmonic centrality algorithm was used to compute node centrality. The harmonic centrality algorithm assigns a centrality value to a node equal to the sum of the reciprocals of each shortest path between a node and every other node in the graph. If there is no path between a node and another node in the graph, the reciprocal value of that path is set to 0 and does not affect the calculation. Centrality is used to identify nodes which are likely important to the flow of information throughout the network. While each individual node importance metric provides important insight, analyzing each of the several thousand nodes in

each network by each individual importance metric is not feasible. Therefore, an average importance metric was calculated for the purposes of assessing a nodes overall importance in the network. However, a simple average of scores does not suffice in this case, as each importance metric score is represented by a different range of values (ex. clustering coefficient is always a number between 0 and 1 while harmonic centrality values are unconstrained). Therfore each network importance metric for each node was assigned a score based on the nodes position in a list of nodes sorted by the importance network score (ex. if a node has the highest clustering coefficient in its layer, the clustering coefficient score of that node will be 1). These relative node importance scores were then added to create an overall network importance score for each node where lower scores represent a higher total network importance (with a score of 3 being the highest possible overall importance).

4.4.4 Protein Protein Interaction Networks

The PPI networks used in this experiment represent a subset of the data collected from the analysis of the signaling pathways governing homotypic vacuole fusion. Both networks are undirected graphs meaning that each PPI represents the interaction of two proteins, not the way in which one protein affects another. Additionally, the type of interaction is not included in the dataset, meaning that an interaction in which one protein phosphorylates another is indistinguisable from two proteins binding to form a complex. Furthermore, any non-protein regulatory element such as lipids are not represented in the PPI dataset. These factors limit the conclusions that can be drawn from PPI networks. However, for the purposes of this experiment, PPI networks provide sufficient data for the identification of broad potential similarities between the signaling networks involved in homotypic vacuole fusion between organisms.

4.4.5 Smith-Waterman Alignment

The Smith-Waterman alignment algorithm was employed to assess the likelihood that a protein from one organism was homologous to a protein from another organism. The algorithm compares segments of each proteins nucleotide sequence, taking into account gaps between aligned sequences, and calculates an overall alignment score which represents the similarity between each proteins nucleotide sequence. However, comparing SW scores between proteins is complicated as the SW score is dependant on the length of the proteins being compared. This means that a pair of relatively long proteins with a SW score of 200 are less similar than a pair of relatively short

proteins with the same score. It may be possible to normalize the SW score of each protein pair based on length, however due to the time constraints of this project, this step was not performed and all proteins identified via SW alignment were designated as homologous regardless of SW score.

4.4.6 Layer 0

Due to the methodology of curating *A. Thaliana* layer 0 proteins from the list of *S. Cerevisiae* layer 0 proteins, a comparison betweeen the two layer's should be treated as a control case as opposed to experimental results. One of the most obvious discrepancies between the two layers is a marked difference between edge counts. While *S. Cerevisiae* layer 0 contains 166 edges, *A. Thaliana* layer 0 contains only 35. There are several possible explanations for this discrepancy. It is possible that edges in the *A. Thaliana* dataset were highly dependent on homology data which would have reduced the nubmer of edges included when homology data was removed from the edge confidence calculation. Likewise, it is also possible that *A. Thaliana* edges had a lower confidence scores overall and, combined with a further reduction of confidence due to homology data removal, were greatly reduced in number by the 900 confidence score threshold. Another possible explanation is that the nodes in *A. Thaliana* layer 0 simply have less edges connecting them to other layer 0 nodes. This explanation is supported by the relatively low clustering coefficient displayed by *A. Thaliana* layer 0. The low clustering coefficient and relatively disconnected nodes in *A. Thaliana* layer 0 indicate that the individual connected components are not directly interacting with each other as they are in the *S. Cerevisiae* layer 0 and that additional regulatory elements are required to bridge the connected components. Conversely, the highly connected nature of the *S. Cerevisiae* layer 0 subgraph indicates that the literature review process was successful in identifying core proteins involved in the *S. Cerevisiae* homotypic vacuole fusion process.

4.4.7 Layer 1

Layer 1 is the most experimentally relevant layer out of the three. While layer 0 is moreso representative of choices in methology, layer 1 serves as somewhat of a middleground where each node is likely involved in homotypic vacuole fusion yet was not curated directly through the literature review process. Comparing the top 20 proteins of each layer 1 subgraph by combined network importance score reveals a striking difference in protein function between the two organisms. While the majority

of the high importance nodes in the *S. Cerevisiae* subgraph are subcomponents of the various protein complexes known to be essential for homotypic vacuole fusion, all of the top 20 hight importance nodes in the *A. Thaliana* layer 0 subgraph are involved in lipid synthesis or modification. This incongruity is likely representative of an outsized role of regulatory lipids in the process of *A. Thaliana* homotypic vacuol fusion in comparison to *S. Cerevisiae* homotypic vacuole fusion. This conclusion is supported by the fact that there are fundamental differences between the phospholipid biosynthetic pathways of *A. Thaliana* and *S. Cerevisiae*. Furthermore depletion of PI3P, an essential regulatory lipid in the *S. Cerevisiae* homotypic vacuole fusion process, has been shown to rapidly induce *A. Thaliana* guard cell vacuole fusion.

4.4.8 Layer 2

Layer 2 was determined to not be experimentally relevant as the majority of proteins added in the conversion between layer 1 and layer 2 were not implicated in the vacuole fusion process. This conclusion is further supported by the low combined index scores of the top 20 proteins in each layer 2 subgraph. These scores show that nodes in the top 20 importance metric score relatively high in one metric but not others, indicating that the network lacks true 'hubs' and that the proteins in each subgraph are more likely to be representative of multiple distinct signaling pathways rather than the process of homotypic vacuole fusion.

Figure 4.4: S. Cerevisiae layer statistics

Figure 4.5: A. Thaliana layer statistics

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Appendix

Protein Tables

S. Cerevisiae

A. Thaliana

References

Seals, Darren F., Gary Eitzen, Nathan Margolis, William T. Wickner, and Albert Price. 2000. "A Ypt/Rab Effector Complex Containing the Sec1 Homolog Vps33p Is Required for Homotypic Vacuole Fusion." Proceedings of the NationalAcademy of Sciences of the United States of America 97(17):9402–7. doi:10.1073/pnas.97.17.9402.