A Spatial Proteome of Paramecium tetraurelia

by

Timothy J. Licknack

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2022 by the
Graduate Supervisory Committee:

Michael Lynch, Chair
Jeremy Wideman
Jay Taylor
Julian Chen

ARIZONA STATE UNIVERSITY

December 2022

ABSTRACT

I studied the evolution and cell biology of *Paramecium tetraurelia*—a model ciliate with over 40,000 distinct protein-coding genes resulting from as many as three ancient whole-genome duplication events. I was interested in the functional diversification of these gene duplicates at the level of protein localization, but the commonly used tools to study this were tedious. I instead applied a protein-correlation profiling approach to this system by way of generating a dozen sub-cellular fractions with different protein constituents due to the density of their resident organelle and then assayed these fractions using quantitative mass spectrometry. Each protein's unique abundance profile provided evidence for its subcellular localization, and I used both supervised and unsupervised classification algorithms to cluster proteins together based on the similarity of these profiles to several hundred "marker proteins" which I manually curated. After expanding the protein inventory for numerous organelles by as many as a thousand proteins, I determined many features not previously understood or appreciated such as mosaic biochemical pathways, evidence for differential sorting mechanisms, and the abnormal evolutionary patterns of the mitochondrial proteome of ciliates. I developed a simple bioinformatic tool to probe spatial proteomics datasets more easily for proteins of interest. I demonstrate its applicability using a handful of well-characterized proteins in the budding yeast *Saccharomyces cerevisiae* as well as interesting proteins in less well-studied model systems like *P. tetraurelia* and the apicomplexan *Toxoplasma gondii* to both recapitulate known interactions and discover new ones. Finally, I look for large-scale evidence of gene duplicates relocalizing to new cellular compartments in *P. tetraurelia* and *S. cerevisiae* using this new dataset and a previously generated one, respectively. I find thousands of pairs of duplicates which are differentially identified and display evidence for subcellular divergence, and this seems to be largely decoupled from large changes in protein sequence but are instead associated with indels in their N-terminal peptide. These findings support the use of high-throughput proteomic techniques to determine evidence of functional divergence of gene duplicates. Taken together, this works provides a deep characterization of one of the largest unicellular proteomes in nature.

DEDICATION

This dissertation is singularly dedicated to my mom, Shelley Licknack. I would not be here without your unconditional love and support. Thank you very much for everything you do.

ACKNOWLEDGMENTS

I want to acknowledge my family, friends, and mentors over the years who socially and mentally helped me get to this point. None of any of ‘this’ matters without people in your life with whom you can share successes and failures.

I have been lucky to have been well-funded over my time at ASU, and this has allowed me to both perform interesting research, travel to conferences to present that work, and attend workshops to improve my practical skills. Thank you to the NSF, Biodesign Institute, SOLS (ASU), and the Graduate College (ASU).

To my collegiate science mentors who inspired me into research: Drs Ellis Benjamin, Guy Barbato, Dan Mulkey, and John Karanicolas— thank you for your role in the advancement of my scientific career. Dr. B—Thank you in particular for being one of the best people I’ve ever met. Your kindness, generosity, and support were the major reason I went this route.

To my graduate committee: Drs Julian Chen and Jay Taylor. Thank you for your support and guidance throughout this process. And Jay— thank you in particular for chairing my Comprehensive Exams and volunteering so much of your time to make that a competent document.

To my advisor Dr Michael Lynch—thank you for giving me a chance after our brief interview at Indiana University all those years ago. I still clearly remember the feeling of walking into your office, surrounded by books, and feeling undeserving of your full attention. You said you wanted students who didn’t confine themselves to a single box, and if nothing else, I hope I was able to live up to that expectation.

To my advisor Dr Jeremy Wideman—thank you as well for pushing me and taking the time to help graft my graduate research project. Although our mutualist relationship began on shaky footing, I think it’s clear now that we made a good choice in working together. I am a much better scientist and communicator because of your input.

To the Lynch, Wideman, Geiler-Samerotte, Hu, and McCutcheon Labs in the CME— thank you for providing me a warm, intense, and stimulating environment in which I was able to learn about diverse topics. There was always someone I could go to for help or conversation. It’s

been a real pleasure watching this Center grow over the years, and I will always consider myself an Evolutionary Cell Biologist because of my time and training here.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

The History and Evolution of the Paramecium aurelia spp. Complex and its Whole-Genome Duplications

Ciliated protozoa represent one of the most diverse eukaryotic lineages in nature whose numerous model organisms have played center stage to some of the most important discoveries in biology. Although obscured through the foggy lens of the early history of science, a representative of the genus *Paramecium* was likely amongst the first microbes to be observed by Antonie van Leeuwenhoek in one of the world's first microscopes in the late 17th century. *Paramecia* reappear in the scientific limelight when Herbert Spencer Jennings recognized their utility in modeling both Mendelian and non-Mendelian patterns of genetic inheritance (Jennings 1906), the latter being taken up with a healthy spirit by his student Tracey Sonneborn who established the germ of the field of transgenerational epigenetic inheritance (Beisson and Sonneborn 1965). Many forget that the success of H.S. Jennings's pioneering work of ciliates allowed him to win the coveted job as the Director of the Zoological Laboratory at Johns Hopkins University over the likes of the scientific giant Thomas Hunt Morgan (Kingsland 1987). Although Tracey Sonneborn would establish a long academic legacy of researchers interested in *Paramecium* cell biology and genetics through his work at Indiana University, the middle of the 20th century saw a decreased interest in *Paramecium* relative to its relative *Tetrahymena*. Amongst the numerous discoveries facilitated through *Tetrahymena* were Nobel-prize winning experiments done in regard to catalytic RNA (Cech, Zaug, and Grabowski 1981; Kruger et al. 1982) and the characterization of the telomerase enzyme (Greider and Blackburn 1985). Relevant to this dissertation, *Tetrahymena* played a critical role in the discovery and characterization of the peroxisome by Christian De Duve (C de Duve 1969). *Paramecium* continued to enjoy some interest in biology, being the first organism to demonstrate a deviation from the "universal" genetic code (Caron and Meyer 1985), but it has lagged far behind more

favored model systems in almost all areas of study. A “Renaissance” in *Paramecium* genetics has brought new interest in not only its cell biology but evolutionary biology as well (Beisson et al. 2010). As investigators continue to build from preexisting models of how *Paramecium* are structured and how they function, the development of new cellular and molecular tools must proliferate into this system to better catch it up to that of more commonly studied microbes like *Saccharomyces cerevisiae* and *Chlamydomonas reinhardtii*. This dissertation outlines steps taken towards further developing *Paramecium tetraurelia* as a model organism in both evolutionary and cellular biology through the application of the emerging toolkit of spatial proteomics to better provide a comprehensive understanding of the structure and history of *P. tetraurelia* in particular and ciliates in general.

The ciliate phylum contains a remarkable diversity of unicellular eukaryotes who range from spherical cells of less than a dozen microns in diameter to elongated cells of more than 4mm (Lynn 2008). Ciliates are cosmopolitan, meaning that some representative can be found in almost every major body of water in the world. From marine nanociliates (Sherr et al. 1986) to larger, freshwater “macro” ciliates (Beaver and Crisman 1989), these organisms occupy a variety of ecological niches. A few cell biological features make ciliates distinct from other lineages. First, they contain cilia— from which they get their name— which are modified flagella that extend through the plasma membrane and are either evenly or unevenly distributed across the cell surface for the purpose of motility. These cilia are often called “somatic” to distinguish them from the “oral” cilia concentrated near the oral groove which facilitates the filtration of prey organisms into their intricate digestive system. Adjacent to the cilia are an intricate cortical network whose organization is central to ciliary function (Aubusson-Fleury et al. 2013). One component of the cell cortex are the flattened, alveolar sacs which are the distinguishing feature of the Alveolate superphylum (Stelly et al. 1991).

Another key ciliate innovation is the existence of two types of nuclei: a larger, transcriptionally active macronucleus (MAC) and smaller, largely silent micronucleus (MIC) which hosts the germline genome to be passed on after sexual reproduction (Hausmann, Bradbury, and others 1996). While other lineages physically separate their germline and soma in entirely

different cells (e.g., sperm/eggs), ciliates manage to do this within the confines of a single cell. The MAC performs all the gene expression needed for growth and maintenance of the cell during its vegetative life cycle, while the MIC only becomes active when meiosis is initiated. MAC chromosomes are highly developmentally regulated, as they all form from some version of a MIC chromosome subject to an extensive splicing-like process in which internally eliminated sequences (IESs) are developmentally excised from interspersed MAC-destined sequences (MDSs) which go onto form the entirety of the MAC genome (Duharcourt, Lepère, and Meyer 2009). The degree to which IES removal and MSD ligation is efficient varies across organisms and between genomic regions of the same organism (Vitali, Hagen, and Catania 2019; Catania et al. 2013). The somatic MAC genome is often highly polyploid due to a complex series of genome amplification events coinciding with the IES elimination and MDS ligation. *Tetrahymena thermophila* contains ~45 copies of each MAC chromosome with substantial variation in smaller, mini chromosomes (Eisen et al. 2006), while the massive *Stentor coeruleus* has ~50,000 copies of each MAC chromosome, possibly related to its scrambled MIC genome (Slabodnick et al. 2017). Ploidy does not scale directly with size, as can be seen in Paramecium caudatum (~400x; (McGrath et al. 2014)) and P. tetraurelia (~800x; (Aury et al. 2006)), but it is thought to be functionally related to the above IES genome structure and unusual feature of MAC amitosis in which MAC chromosomes do not align along the metaphase plate which results in noisier chromosome inheritance patterns (Vitali, Hagen, and Catania 2019; Catania et al. 2013). This has been linked to accelerated protein evolution (Zufall et al. 2006).

P. tetraurelia is a member of the *P. aurelia spp*. complex containing at least 14 unique species (Sonneborn 1975) all having experienced two whole-genome duplication (WGD) events preceding their speciation (Gout et al. 2019) and at least one more ancient event preceding the split between the *Paramecium* and *Tetrahymena* genuses. While small-scale duplications (SSDs) can occur for a variety of reasons, WGDs typically occur by polyploidy events during chromosome segregation in which one daughter cell inherits two copies of its parental genome. WGDs are quite common across the tree of eukaryotes, with the most famous events occurring at the root of vertebrates (Dehal and Boore 2005). Paralogous genes resulting from WGD events

are often called “ohnologs” as an homage to the late Susumu Ohno. Gene retention from SSD events are quite low, as is the case in the budding yeast *Saccharomyces cerevisiae* who retain only ~10% of ohnologs generated 100+mya (Guan, Dunham, and Troyanskaya 2007). Comparably, *P. aurelia spp.* have still retained a large proportion of them ancient ohnologs (McGrath et al. 2014). 40+% were retained in the various *P. aurelia spp.* from its most recent event ~300-350mya: ~50% for *P. tetraurelia*. This timing is comparable to the WGDs experienced by the flowering plant *Arabidopsis thaliana* in which roughly 29% of ohnologs have been retained (Thomas, Pedersen, and Freeling 2006).

A number of features dictate whether a gene will be retained, but the strongest is likely its mRNA expression level (Gout et al. 2010). More highly expressed genes are under stronger purifying selection which prevents their accumulation of mutations affecting their function. Dosage of gene duplicates must be optimized to maintain proper functioning of the cell, especially when that protein is involved in some pathway or protein complex with stoichiometric constraints (Veitia, Bottani, and Birchler 2008). This is thought to play a major role in repressing SSDs in *Paramecium aurelia spp*. which have only a few hundred instances of post-WGD SSD events (Gout et al. 2019). Freed of these constraints, ohnologs have many paths for functional diversification. One path is neofunctionalization (Ohno 1970), a classic view in which one copy is free to adopt a wholly new function while its ohnolog retains the ancestral function. Another path is subfunctionalization in which each copy retains part of its ancestral function (Force et al. 1999). This typically occurs through the accumulation of complementary mutations in each ohnolog which degrades a different ancestral function such that the combination of the two genes fulfills the role of the single-copy ancestor. Subfunctionalization can be either qualitative or quantitative, the former exemplified by differential expression of one copy in one tissue and the other copy in another, and the latter represented by the sum of each gene’s mRNA expression being relatively equal to the pre-duplicate state (Gout and Lynch 2015). This can occur entirely in the absence of equivalent, functional changes to the duplicates, although this has only been assessed at the level of protein sequence.

Cell Biology of Paramecium

Though gene duplication plays no major role in the organization of the *Paramecium* cell, it certainly plays a role in the proteomic composition of *Paramecium*'s numerous organelles. Indeed, one of the many attractive features of *P. tetraurelia* is its numerous families of paralogs, some of which display evidence for functional diversification. Here, I'll explore what we know about protein localization in *Paramecium* in general and focus on duplicates in *P. tetraurelia* when appropriate.

Paramecium hosts a fairly standard suite of eukaryotic organelles as well as some unique ones. A "standard" eukaryotic cell is characterized by its compartmentalization of numerous biochemical pathways into distinct membrane-bound organelles and specialized cellular compartments. These organelles typically include one double-membrane nucleus, numerous double-membrane mitochondria, and single membrane peroxisomes, lysosomes (sometimes called vacuoles), and endosomes which interactively make-up its membrane trafficking system in conjunction with the endoplasmic reticulum (ER) and Golgi apparatus. Numerous smaller vesicles dynamically interact with these organelles and play key roles in shuttling proteins to both the endomembrane system and to the plasma membrane and extracellular space via constitutive secretion. Ribosome complexes can be seen in both the cytoplasm and rough-ER actively translating proteins as well as in the nucleolus in which they are assembled.

The nuclear dimorphism of ciliates accompanies a continuous endoplasmic reticulum (ER) which is directly adjacent Golgi stacks (Allen and Fok 2000) containing a well-developed trans-Golgi network observed to have clathrin-coated vesicles being released into a network of filaments. *Paramecium* contains both mitochondria and peroxisomes with roles in metabolic functioning, the former being concentrated near the cell cortex. Mitochondria are the site of oxidative phosphorylation which ends in a divergent ATP synthase whose conformation likely dictates the distinctly tubular, not lamellar, cristae; a synapomorphy of the ciliate phylum (Balabaskaran Nina et al. 2010). The TCA cycle and possibly glycolysis occur in ciliate mitochondria, but the latter needs more investigation (Smith et al. 2007). The two organelles are

thought to interactively regulate lipid metabolism in *Tetrahymena* (Krueger et al. 2022), but this is unknown in *Paramecium*. Constitutive, or so-called receptor-mediated, endocytosis occurs at coated pits in the plasma membrane characterized by the presence of parasomal sacs (Allen, Schroeder, and Fok 1992) which results in the formation of early endosomes, although the direct connection with lysosomes is more tenuous than that of “higher” eukaryotes (Reuter, Stuermer, and Plattner 2013). Indeed, endosomes certainly share protein machinery with lysosomes, but the latter is more commonly studied in the context of their fusion with large phagosomes (Fok and Allen 1990). The interplay between these two pathways has motivated its joint terminology as the phagolysosomal pathway. Phagosomes form at the oral groove of *Paramecium* and are rapidly bound by non-lysosomal acidosomes (Allen and Fok 1983), which lowers the pH substantially, and then by lysosomes which kill and digest prey as the phagosome contracts and expands (Fok, Lee, and Allen 1982).

Perhaps the most striking feature of *Paramecium* is its large contractile vacuole complex (CVC) made up of a large, pulsating central vacuole and surrounding radial arms which act to relentlessly expel water and cellular material out of a small pore in the plasma membrane (Plattner 2013). This organelle is a unique solution to the problem of maintaining water balance and homeostasis of ions like calcium (Plattner 2020). As Paramecium ingests water from the environment at both its oral groove, when filtering prey, and across its large surface area, water dilutes intracellular materials and must be removed. Anterior and posterior CVCs periodically pulsate as they fill with water and expels it. Many protists have some version of a CVC, but none is quite as impressive structurally as that of *Paramecium* due to its radial arms (Allen and Naitoh 2002). The relationship between these morphologically different but functionally related CVCs in diverse lineages has not been subject to deep investigation.

In addition to the constitutive exocytosis in which proteinous material is shuttled out of the cell (e.g., surface antigen proteins) (Preer Jr 1986), *Paramecium* hosts large trichocysts with spear-like shapes used primarily for defense (Plattner 2017). Again, many protists have some type of extruding organelle (called extrusomes) (Rosati and Modeo 2003), but that of *Paramecium* is remarkable for being perhaps the fastest secretion process in nature. While

impossible to measure due to its speed being faster than the shutter rate of most cameras, H. Plattner estimates that trichocysts are discharged at a rate faster than 24um per millisecond (Plattner 2017). This is achieved by their protein contents forming a metastable crystalline structure which rapidly decondenses and elongates when stimulated by calcium. Adjacent to trichocysts are the numerous cortical organelles which define the complexity of the plasma membrane of alveolates such as alveolar sacs (Stelly et al. 1991), basal bodies (Tassin, Lemullois, and Aubusson-Fleury 2015), and cilia (Dute and Kung 1978). Alveolar sacs are a synapomorphy of the alveolate superphylum, from which they get their name, and act mainly as calcium stores regulating various cortical activities in different lineages (Gould et al. 2011). Cilia are the defining morphological feature of the ciliate phylum, also from which they get their name, whose role in motility is critical. Different ciliate lineages achieve this through very different orientations—*Paramecium* being a holitrich ciliate are uniformly covered in somatic cilia (Lynn 2008), while other ciliate lineages contain bundles of cilia called cirri. This former organization is the case for hypotrich ciliates like *Oxytricha*. “Oral” cilia exist near the oral groove and act to pull in prey for phagocytosis. All cilia in this lineage are structurally constituted of a microtubular axoneme in a classic 9+2 orientation which can be observed in organisms as diverse as humans and algae (Ishikawa 2017). Cilia are anchored to cortical basal bodies which connect them to a vast cytoskeletal network (Tassin, Lemullois, and Aubusson-Fleury 2015). Basal bodies have been of particular interest due to their central role in ciliary function and numerous protein constituents conserved across eukaryotes implicated in human disease (Valentine and van Houten 2021).

The major tools for studying protein localization in Paramecium have been GFP-fusion and antibody staining (Hauser, Haynes, et al. 2000b). While the sophisticated genetic manipulations available in other systems are not amenable in *Paramecium*, gene knockdowns can be performed using RNAi introduced by feeding (Galvani and Sperling 2002). GFP-fusion proteins were first introduced in *P. tetraurelia* in two studies, the first demonstrating GFPs ability to be expressed (Hauser, Haynes, et al. 2000b), and the second using the ER-residents ptSERCA1 and ptSERCA2 which highlighted immediately some issues with this technique

(Hauser, Pavlovic, et al. 2000). The two proteins displayed dual localization to the ER and alveolar sacs only when the GFP molecule was inserted into the C-terminus, but alveolar localization was abolished when moved to a cytoplasmic loop containing a preserved KKIQ motif. While this finding was important for establishing the role of both organelles as major calcium stores in *Paramecium*, it also highlighted the problems with GFP-fusion proteins well-known in all model systems. Despite this early warning, very little effort has been made to systematically study differential localization in *Paramecium* due to GFP placement, as has been done in *S. cerevisiae* (Weill et al. 2018; Huh et al. 2003). These findings from yeast suggest that as many as ~40% of localization assignments from GFP fusion proteins differ between N- and C-terminal proteins, with 11% being entirely different. Despite this, orthogonal approaches like RNAi provide ways to probe the removal of a protein of interest on the structure and function of some organelle in which it localizes, and these results provide independent support for the findings hereafter.

There is a certain survivorship bias in the literature with respect to which proteins are chosen for direct assaying of protein localization. This bias comes from the types of proteins in which investigators are interested, and thus many 'classic' residents of organelles are never studied due to the assumption that their localization is conserved across species. Sometimes, these proteins are assayed as indirect expectations for what protein localization in that region looks like. One example is the ubiquitous ER chaperone protein disulfide-isomerase (PDI) which acts to synthesize disulfide bonds at cysteine residues in developing proteins (Wilkinson and Gilbert 2004). PDI is often used as a standard for the ER, as was the case in a study of a large family of calcium-release channels (Ladenburger and Plattner 2011). The authors described 34 CRC genes and raised polyclonal antibodies specific for some of the six major families which each contained two to five individual genes (ohnologs). CRC-I stained the ER, as was evidenced by its complete overlap with both PDI (anti-mouse) signal and the ER-stain DiOC6. The other families localized to diverse structures like the phagosome, recycling vesicles, contractile vacuole, both nuclei, trichocyst tip, and various areas of the cortex. Many more proteins were localized to the ER in *Paramecium*, but none of them were ubiquitous ER markers as was PDI. Another example is the syntaxin PtSyx8-2, a member of a large family of at least 26 genes with a

similarly abundant number of ohnologs (Kissmehl et al. 2007). While this gene localized to the ER, its sister gene, PtSyx9-1, localizes to small acidic vesicles in the cytoplasm, and the large syntaxin family spanned structures like the plasma membrane, contractile vacuole, cortical vesicles, Golgi, discoidal vesicles, “streaming” vesicles, and patches on phagosomes. The synaptobrevin PtSyb1-1 had a similar ER localization while its diverse family members stained equally diverse sites (Schilde et al. 2006). It is interesting to note that syntaxin and synaptobrevin genes, specific types of SNAREs, are named in accordance with their role in neurons, but in *Paramecium* they serve similar intracellular functions, i.e., the regulation of vesicular binding (Plattner 2010). Many more examples of diversified protein families in *P. tetraurelia* have been investigated in proteins as diverse as actin (Sehring, Reiner, and Plattner 2010), Rab GTPase (Bright, Gout, and Lynch 2017), stomatin (Reuter, Stuermer, and Plattner 2013), V-ATPase (Wassmer et al. 2006), and calcineurin (Fraga et al. 2010).

These results, taken together, have aided in building a model of the *Paramecium* cell: its major systems, and the protein families spanning those systems. One such system is the phagolysosome system introduced previously as the means by which *Paramecium* breaks down prey into usable building blocks (Fok and Allen 1990; 1998). This process can generally be summarized as such: phagosomes form at the cytostome (oral groove) before binding with first acidosomes then lysosomes before being ejected from the cytoproct as “spent” vacuoles and forming discoidal vesicles which then provide recycled membrane for the next phagosome. This process involves a few major protein families: H^+-ATPases, SNAREs, Rab GTPases, and CRCs (Plattner 2022). Proton pumps have been localized to coated pits from which early endosomes form and are thought to precede acidosomes (Wassmer et al. 2006). The fusion of acidisomes with phaogosomes is an important step in acidifying the vacuolar environment for the proper functioning of lysosomal enzymes which will subsequently break down various macromolecules within the phagolysosome. These “true” lysosomes contain inactive enzymes coated with traditional lysosome membrane proteins like LAMP (Huynh et al. 2007), and these enzymes are recycled after digestion is complete so as to not eject them from the cell. Phagolysosomes are

propelled through parts of the cytoplasm via specific actin isoforms which appear as a “steam” behind the vacuole (Sehring, Mansfeld, et al. 2007).

Another system is the osmoregulatory system centered around the CVC and its numerous ion-mediated processes (Plattner 2013). Remarkably, the CVC contains much of the same vesicle trafficking machinery as phagosomes minus actin in exchange for tubulin. While protein localization in the phagosome is typically uniform, CVC localization is typically more specific due to its multi-component structure. For example, the synaptobrevin Syb2 and many NSF genes stain both the central vacuole and radial arms (Schilde et al. 2010; Kissmehl et al. 2002), while the F-subunit of the V1 proton pump stains only the decorated spongiome of the radial arms (Wassmer et al. 2006). Vesicle binding through a variety of t- and v-SNAREs is suggested along both the distal, decorated and proximal, sooth spongiomes which may regulate CVC contraction (Plattner 2022). The pumping of protons modulates, in some way, how water and ions are brought into the CVC through various types of ion channels observed in other species (Plattner 2013). In *P. multimicronucleatum*, a specialized aquaporin protein aids in the influx of water to the CVC (Ishida et al. 2021).

The final system I will discuss is the dense core-secretory exocytic pathway centered around the biogenesis of the *Paramecium* trichocyst (Plattner 2017). Trichocysts are a type of extrusome found in various ciliates and other lineages like dinoflagellates (Rosati and Modeo 2003). It is homologous to the better-studied *Tetrahymena* mucocyst, although the two structures are morphologically and functionally quite different. As mentioned previously, trichocyst discharge is thought to be the faster secretion process in nature, and it is achieved through the coordinated elongation of its protein constituents called trichocyst matrix proteins (TMPs). TMPs are so concentrated that they form crystals tenuously held together with repulsive negative charges—essentially loading a ”thermodynamic trap” from which trichocyst discharge gets its energy (Vayssié 2000). This process is modulated by positively charged calcium ions which either comes from the external environment or alveolar sacs (Plattner 2022) and can be stimulated in the laboratory by dextran and its derivatives. Ca^{2+} binds to these negative charges and causes an immediate relaxation of the proteins to a lower energy state. Trichocyst maturation is multi-

stepped and involves both the ER and Golgi. TMPs are synthesized as ~40KDa precursor proteins subjected to translocation to the ER through putative signal peptides (SPs) on their N-terminus (Arnaiz, Meyer, and Sperling 2020). After translocation to the ER, the inactive protein undergoes proteolytic cleavage into a smaller ~20KDa active protein capable of crystallizing outside of the Ca^{2+}-rich environment of the ER (Adoutte, de Loubresse, and Beisson 1984). TMPs are then passed to the Golgi apparatus and often subjected to heavy glycosylation before moving into smaller vesicles in which crystallization occurs (Richard D Allen 1988). These vesicles are often called pre-trichocysts and mature at the cell cortex into functioning trichocysts. This is the route of travel for TMPs, but non-TMP trichocyst proteins are not well-understood. However, the delivery of pre-trichocysts is mediated by actin similar to phagolysosomes (Plattner 2022). The synaptobrevin ptSyb5 localized to these vesicles and may bind to the syntaxin Syx1 (Schilde et al. 2010; Kissmehl et al. 2007). Again, these examples highlight the overlapping protein machinery of many of these distinct important cellular systems.

The Development of Spatial Proteomics as a Tool in Cell Biology

A burgeoning toolkit in the field of spatial proteomics offers the potential to localize thousands of proteins to organelles and subcellular compartments with a single experiment (Christopher et al. 2021). This would certainly aid the understanding of ciliate complexity, and I will outline the history of this field and the details of these techniques.

The development of spatial proteomics coincides largely with the development of the field of cell biology as a whole. The 1974 Nobel Prize in Medicine and Physiology went to three researchers credited largely with laying the groundwork for the field of cell biology to become a recognized field of biology: George Palade, Albert Claude, and Christian De Duve. Palade's discoveries are not of direct importance in terms of background, but his work on the discovery of ribosomes and the structural organization of cells cannot be understated. However, it was Claude that pioneered the "pulverization" of cells and structural assaying of subsequent subcellular fractions to better understand their spatial organization inside the cell (Claude 1975). And then it

was De Duve who ‘perfected’ these techniques and introduced biochemical assays that aided in the discovery of a number of organelles like the lysosome and peroxisome, the former of which was the main cause of his Nobel Prize (Sabatini and Adesnik 2013). De Duve’s group made use of enzyme activity assays of different fractions to determine the differential protein constituents of those fractions that correspond to some subcellular structure. After much experimenting, De Duve et *al*. (1955) found that a non-mitochondrial “L; Light” fraction contained a number of acid hydrolases with the ability to break down a number of different macromolecules, and he named this ‘lytic microsome’ the lysosome (Holtzman 2013). These findings were crucial for establishing what has sometimes been called “De Duve’s principle” in which the subsequent centrifugation of a cell lysate will result in organelles adopting a distribution of abundances across those fractions. Intact organelles should contain the entirety of their protein constituents under certain conditions, especially those absent of harsh detergents.

The development of cell biology by ultracentrifugation has coincided with the development of a number of techniques for probing macromolecular structure and composition. One of these techniques is mass spectrometry, and it is the central method of spatial proteomics. Mass spectrometry comes in many flavors, but in studying proteomics, a workflow typically involves the digestion of proteins into peptides, the ionization of those peptides into precursor ions, the measurement of retention time, intensity, and mass to charge ratio (m/z) of those ions, and then the optional step to fragment those precursors into smaller ions for one to two more rounds of intensity and m/z measurements (Aebersold and Mann 2003). A number of Nobel prizes were awarded for methods underlying mass spectrometry, most recently in 2002 for the development of the wildly popular electrospray method of peptide ionization (Fenn et al. 1989). Modern mass spectrometers can easily identify thousands of proteins from a complex mixture regardless of its source.

The combination of cell fractionation and quantitative mass spectrometry was first performed by the group of Matias Mann, who sought to study the human centrosome using classical affinity purification methods but found this method to pick-up too much background (Mann 2020). Instead, the protein correlation profiling was developed in which numerous

subcellular fractions were subjected to LC-MS/MS with label-free quantification (LFQ). Using previously known centrosomal proteins, they could predict a number of new components which expanded that amembranous organelle's known proteome (Andersen et al. 2003). LFQ has its disadvantages due to the requirement that each fraction be assayed separated in the mass spectrometer. Kathryn Lilley's group modified this method using peptide labeling, called Localization of Organellar Proteins by Isotopic Tagging (LOPIT), and showed in the flowering plant *Arabidopsis thaliana* that four subcellular fractions could provide clear localization profiles for hundreds of proteins simultaneously (Dunkley et al. 2004). Labeled quantification has the advantage of comparing peptide abundances in the same exact run (i.e., head-to-head) which reduces noise from missing values. In the past ~20 years, these methods have been applied to a number of different cell/tissue types, organisms, and cell states (Borner 2020). LOPIT has undergone a few iterations, including hyperLOPIT, which utilizes ten (and now 16) isobaric TMT labels, and LOPIT-DC, which fractionates with differential centrifugation instead of density gradient centrifugation (Geladaki et al. 2019). Density-gradient centrifugation provides slightly better resolution to organellar maps but requires a greater degree of technical expertise and equipment, while differential centrifugation is far simpler and quicker. TMT-isobaric tags have enabled a higher degree of quantitative accuracy through the ability to multiplex different fractions into the same mass spectrometry run without the need for post-hoc retention time alignment (Ong 2003). TMT-labeling can cause 'ion stacking' in which MS2 profiles cannot resolve different isobaric tags, and MS3-based quantification is required. However, this is still highly accurate, while label-free methods provide modestly noisier quantification of a deeper proteomics characterization for a lower cost. Metabolic labeling of cell cultures using the SILAC method provides an alternative approach that is very powerful for comparing two or few conditions. The modularity of this approach has aided in its adoption in numerous systems.

Regardless of the methodology being used, spatial proteomics techniques involve the generation of unique profile of protein abundance that should be more similar for proteins of the same cellular compartment than biophysically unrelated proteins. Simply, a protein found in X organelle should be relatively more abundant in a fraction enriched in X organelle. Across many

diverse fractions, a number of relative abundance profiles should be present for a whole manner of organelles and non-organellar compartments. For example, the large and dense mitochondria should pellet early in a differential centrifugation experiment or segregate further along a density gradient relative to some lighter organelle like the lysosome or peroxisome. Thus, one expects true mitochondrial residents (like the F1 ATP synthase subunits) to display their highest relative abundance in the fraction in which mitochondria are enriched. In addition to the normalization methods inherent in all mass spectrometry-based proteomics (i.e., normalizing a peptide abundance relative to all other peptide abundances), spatial proteomics analyses typically involve some protein-level normalization to get all proteins onto the same scale, for e.g., by summing all protein abundances and making each a ratio of that sum (Callister et al. 2006). The de novo clustering of abundance profiles should group proteins residing in the same or similar regions of the cell with respect to their steady-state protein localization (Barylyuk et al. 2020). The application of *a priori* biological knowledge onto this data structure allows for the use of supervised classification algorithms that can use labeled data (i.e., marker proteins) to make predictions about the localization pattern of unlabeled data (i.e., unknown proteins). This was first done using the so-called “chi-square” measurement which was calculated using the sum of squared differences between peptide abundance values in each fraction (Andersen et al. 2003). Using numerous peptides unique to centrosomal proteins, an empirical cut-off value was set after which peptides did not clearly correspond to centrosomal proteins. Since this classic approach, a number of machine learning techniques have been used, perhaps the most popular being support vector machines (SVMs) (Boser, Guyon, and Vapnik 1992)and recently Bayesian mixture modeling (Gatto, Breckels, Wieczorek, et al. 2014; Crook et al. 2019; 2018; Gatto, Breckels, Burger, et al. 2014). These classification methods are often called “hard” and “soft”, respectively (Liu, Zhang, and Wu 2011). Simply, hard classifiers like SVMs involve the generation of a decision boundary (e.g., a line in 2D space) which is iteratively trained on labeled data in order to maximize the number of true positives and minimize the number of false negatives on each side of the boundary. For spatial proteomics, the training set is made up of marker protein abundance profiles, and dimensionality is determined by the number of subcellular fractions generated. The

closer a protein is to a given decision boundary, the weaker its classification score is. For example, a protein on the mitochondrial side of a decision boundary will have a higher membership likelihood (i.e., larger SVM score) the further it is from its decision boundary with non-mitochondrial marker proteins. In contrast, soft classifiers like Bayesian mixture modeling measure conditional (or posterior) probabilities for each protein profile belonging to each marker class to measure global uncertainty in classification to all compartments (Crook et al. 2018). A protein with high probabilities for multiple compartments is more uncertain than one matching only a single compartment. Both approaches have value in this domain and provide similar levels of organellar resolution (Crook et al. 2018). Irrespective of the analytical method, the output is that each protein in the dataset gets classified to a single compartment, and the distribution of either classification scores or global uncertainty can be used to filter out lower-confidence classifications to make predictions with fewer false positives.

Recent years have seen these methods applied to a number of diverse organisms necessitating high-throughput techniques of this kind. This has been done in unicellular and multicellular organisms in many different conditions and can likely be applied to any organism capable of basic cell biological manipulations. Within a single organism, one can probe changes to global protein localization before and after some treatment. The simplest application is the global determination of steady-state protein localization patterns in a single organism in a constant environment. Applications to mouse (Christoforou et al. 2016; Foster et al. 2006), rat (Jadot et al. 2017), yeast (Nightingale, Oliver, and Lilley 2019), and human (Geladaki et al. 2019; Thul et al. 2017) have yielded a massive expansion to their known organellar proteomes. Assayed as well were the cyanobacteria *Synechocystis* (Baers et al. 2019) and apicomplexan *Toxoplasma gondii* (Barylyuk et al. 2020) with cell biologies unique from classic model systems.

Additionally, a few experiments were done in different conditions to highlight dynamic changes to protein localization within the same species. For example, human HeLa cells treated with epidermal growth factor were shown to exhibit large changes to protein localization including the translocation of numerous transcription factors to the nucleus (Itzhak et al. 2016). Mouse liver cells from individuals with alcohol-induced hepatic disease were shown to exhibit large changes

to protein localization within and around the Golgi apparatus and involving lipid metabolism (Krahmer et al. 2018). Human cell lines exposed to proinflammatory lipid polysaccharides displayed stark differences in both protein abundance and localization in immune and signaling proteins (Mulvey et al. 2021). Human HaCaT skin cells exposed to UV light displayed a disproportionate number of mitochondria to secretory translocations (Valerio et al. 2022). One experiment managed to relocalize proteins to the mitochondria by modifying a key Golgin protein which directs cargo from the trans-Golgi (Shin et al. 2020).

Taken together, the experimental and bioinformatic toolkit of spatial proteomics-based organellar mapping provides a powerful way to study protein localization and cell biology. The ability to apply this toolkit to any organism capable of basic cell biological manipulations opens up a number of avenues for studying the evolution of protein localization and cellular organization across diverse lineages.

Dissertation Structure

This dissertation will take steps towards the establishment of a spatial proteomics toolkit in the ciliate *Paramecium* with its numerous gene duplicates. To facilitate more discoveries in non-model systems, I developed a simple bioinformatic infrastructure to empower researchers in the future to make more discoveries and provide more insight into the nascent field of evolutionary cell biology. This introductory chapter provided important background information necessary for framing the context of the remainder which can be summarized simply as: *P. tetraurelia* is a large, complex cell with a unique evolutionary history in a poorly studied lineage of unicellular eukaryotes. The second chapter will outline my contribution to the study of the evolution and cell biology of *P. tetraurelia* through the generation of a novel spatial proteomics dataset in which over 9,000 unique proteins are predicted to a subcellular localization. The third chapter will be my contribution of a new bioinformatic tool which provides a simple way for researchers to make use of these types of datasets without the need for high-level programming expertise. The fourth chapter will be a thorough dive into the subject of gene duplication via the identification of

hundreds of pairs of ohnologs both differentially identified in this deep proteomic survey and differentially abundant in subcellular fractions. These findings broadly support the use of spatial proteomics data as a high-throughput assay for protein relocalization. A final conclusion chapter will wrap up the dissertation, outline its shortcomings, and highlight take-home messages.

CHAPTER 2

SPATIAL PROTEOMICS OF PARAMECIUM TETRAURELIA REVEALS THE PERVADING NATURE OF MEMBRANE TRAFFICKING

Highlights:

- Spatial proteomics generates multi-dimensional protein abundance profiles for over 9,000 unique proteins in *Paramecium tetraurelia*
- The known protein inventories for dozens of organelles and compartments are expanded by as many as a thousand new proteins
- Some metabolic pathways, like glycolysis, are spatially distributed between membrane-bound organelles and the cytoplasm
- Membranous trafficking proteins predicted to different compartments show evidence for different sorting mechanisms

Summary:

Paramecium tetraurelia is a model ciliate with over 40,000 distinct protein-coding genes resulting from as many as three ancient whole-genome duplication events. This had led to the expansion of many gene families and their subsequent, functional diversification, but we know virtually nothing about most of these genes – whether they actually produce proteins, and if so, where those proteins localize after synthesis. When protein localization is assayed, the results are often ambiguous due to the immeasurable complexity of membrane trafficking systems and observed promiscuity of many labeled proteins. Here, we take a protein-correlation profiling approach to cluster proteins based on their relative abundance in biochemically distinct fractions after gentle cell lysis. We use supervised and unsupervised learning models to leverage known biological information and make predictions about the localization patterns of thousands of unknown proteins. Our findings largely recapitulate the expected properties of organellar proteomes and allow us to expand their protein inventory. In some cases, biochemical pathways contain enzymes differentially localized to more than one organelle, and we highlight the case of upstream glycolytic enzymes displaying a cytoplasmic pattern and downstream enzymes appearing mitochondrial. We also describe new biological properties like targeting sequences, regulatory elements and protein sorting pathways. With 901 proteins containing a predicted signal peptide, we discover a hydrophobic motif found in ~95% of proteins differentially positioned in

proteins predicted to different organellar compartments and associated with transmembrane domain presence. We did not resolve the macronucleus and micronucleus of *P. tetraurelia* but show that the nuclear proteins we do predict represent many core nuclear protein complexes and cover important functions like transcription and DNA replication. We end by identifying high confidence orthologs across eukaryotic diversity to determine the evolutionary pattern underlying each compartment and carefully compare our dataset to a similar one from the apicomplexan parasite *Toxoplasma gondii*. Taken together, this work provides the deep characterization of one of the largest microbial proteomes in nature as well as a resource for community-driven discoveries resulting from these data.

Introduction:

The hallmark of eukaryogenesis is the evolution of new organelles and the accompanying expansion of their protein repertoire. The ciliate phylum is one lineage with a plethora of unique organelles whilst also containing some of the largest cell sizes and largest gene families of all extant eukaryotes (Maurer-Alcalá and Nowacki 2019; Lynch et al. 2022a). The *Paramecium aureli*a species complex takes this to a new extreme with roughly 40,000 protein coding genes produced as a result of as many as three ancient, whole-genome duplication (WGD) events (Aury et al. 2006; Gout et al. 2019). *Paramecium tetraurelia* is the best-studied species in this complex, having served as a historical model system in genetics and cell biology and an emerging system in genomics and evolutionary biology (Beisson et al. 2010). Despite this, relatively few protein-coding genes have been functionally investigated or localized within the cell. Investigations into *P. tetraurelia* using RNAi knockdown or GFP-fusion proteins have revealed some of the molecular mechanisms underlying ciliate cell biology (Hauser, Haynes, et al. 2000a; Plattner 2018), but direct cell biological analysis of all these proteins is unrealistic, and novel approaches must be used to better understand *P. tetraurelia*'s macromolecular composition.

In addition to their gene expansions, ciliates host a number of unique organelles and pathways alongside their suite of standard eukaryotic features (Lynch et al. 2022b). Perhaps most conspicuous is the phenomenon of nuclear dimorphism whereby ciliates have two types of

nuclei: a large transcriptionally active macronucleus (MAC) and a smaller, silent micronucleus (MIC) acting analogously to its germline (Hausmann, Bradbury, and others 1996). A complex series of RNA-mediated genome rearrangement events accompanies meiosis, and each sexual generation sees the destruction of the old MAC and the creation of a new nuclei (Nowacki, Shetty, and Landweber 2011). Continuous with these nuclei is a vast endoplasmic reticulum (ER) accompanying Golgi stacks through which protein trafficking is conducted (Guerrier et al. 2017). Perhaps the most striking features of their endomembrane system are the osmoregulatory contractile vacuole complexes (CVC), phagosomes, and cortical organelles like alveolar and parasomal sacs (Plattner 2020; 2013; 2010; Plattner and Kissmehl 2003; Stelly et al. 1991). As a holotrich ciliate, *Paramecium* are uniformly covered in cilia anchored to the cell via cortical basal bodies and their vast cytoskeletal network (Lynn 2008; Tassin et *al.* 2015). The sheer complexity of this system requires the development of novel approaches to understand it as a whole. Recent developments in spatial proteomics offer a solution to the problem of studying ciliate cell biology, bringing the potential to localize thousands of proteins to organelles and subcellular compartments simultaneously (Christopher et al. 2021). Briefly, all experiments under the umbrella of spatial proteomics utilize cell lysis, fractionation, and quantitative proteomics to identify proteins with similar abundances due to their shared organellar environment. High confidence “marker proteins” are used to indicate the expected pattern of all proteins within an organelle, and various statistical and machine-learning techniques then classify unknown proteins based on how they compare to distributions of the marker proteins themselves. So far, these methods have only been applied to a limited number of model systems (Gatto, Breckels, Wieczorek, et al. 2014), but their design is applicable to any organism capable of basic cell biological manipulations.

Here, we combine cell lysis and fractionation with label-free quantitative (LFQ) proteomics to produce a multi-dimensional dataset of protein abundance across biochemically distinct fractions in *P. tetraurelia*. We first predict the localization pattern of thousands of unknown proteins using a supervised classification algorithm trained on hundreds of manually curated marker proteins and found good overlap between them and proteins grouped through

unsupervised clustering. We then use the shared cell biological properties of these subcellular compartments to predict new features like regulatory motifs and biochemical modules. Surprisingly, we discover metabolic pathways with mosaic organellar composition such as glycolytic enzymes with either cytosolic or mitochondrial localizations. We demonstrate the pervading nature of the ER-Golgi system in influencing the steady-state protein abundance behavior of many proteins acting at the cell cortex and secreted to the cell surface and discover a hydrophobic peptide motif whose position may influence protein trafficking. We then discuss the irreducible complexity of the MAC and MIC proteomes whose localization patterns could not be resolved in this study. Taken together, we provide a deep characterization of one of the largest unicellular eukaryotic proteomes in nature.

Organellar Compartment	Description of Marker Proteins	Marker Proteins (N)	Classified Proteins (N)	SVM Score (Median)	Predicted Proteins (N)	Predicted TMD (%)	Predicted SP (%)	Predicted NLS (%)	Predicted mTS (%)	DE after Trichocyst Discharge (%)	DE during Reciliation (%)
Axoneme	Ciliary microtubules, dynein, radial spokes; Cortical vesicles	14	334	0.36	59	1.8	1.8	10.9	0.0	0.0	**43.8**
Basal Body Associated	Striated Rootlet; Infraciliary lattice	14	137	0.33	30	0.0	0.0	0.0	0.0	3.2	0.0
Basal Body Core	Basal body; Centriole	13	653	0.6	387	10.6	2.1	3.6	0.3	2.1	5.0
Cytosol	Cytoplasmic enzymes	16	621	0.55	343	1.7	0.0	2.0	0.0	0.9	3.5
ER	ER membrane, import, chaperones; Exocytotic vesicles	16	647	0.54	342	**66.0**	**31.4**	5.8	0.3	**9.6**	1.8
Lysosome	Lysosomal peptidases, Sphingol. glycosidases, membrane	14	300	0.52	146	16.4	**57.5**	0.7	0.0	1.4	2.8
Membrane Trafficking Insoluble	Ciliary/plasma membrane; Vesicle organization/fusion; Contractile Vacuole	23	1,221	0.5	540	**61.9**	11.7	1.7	0.0	2.8	3.0
Membrane Trafficking Soluble	Cytoskeleton; Vesicle sorting; Endocytotic vesicles; Ciliary trafficking	26	1,434	0.5	625	2.3	1.7	4.2	0.0	2.3	1.5
Mitochondria	Mitochondrial matrix, IMS, inner membrane	34	1,313	0.81	1,089	17.0	2.3	1.7	**19.8**	0.9	2.7
Mitochondrial Outer Membrane	Porins; Amidase; Membranous	13	102	0.46	40	38.1	7.1	0.0	0.0	2.5	7.5
Nuclei Insoluble	Chromatin; Chromosome segregation; DNA replication machinery	13	233	0.34	42	0.0	0.0	**18.6**	0.0	2.3	2.3
Nuclei Soluble	Transcription and DNA replication machinery; Nucleoli	22	637	0.45	251	0.8	0.0	**20.7**	0.0	0.4	8.9
Peroxisome	Peroxisomal membrane, import, and enzymes	16	227	0.45	89	13.3	1.1	1.1	2.2	0.0	8.4
Proteasome	20S and 26S proteasome	16	175	0.42	66	0.0	0.0	1.5	0.0	0.0	0.0
Ribosome	40S and 60S ribosome	14	508	0.48	181	2.2	1.1	**28.3**	0.0	0.0	1.1
Surface Antigen	Glycosylation; Environmental signaling	13	139	0.42	57	13.8	**34.5**	5.2	0.0	3.6	7.3
Trichocyst Matrix	Trichocyst matrix proteins	14	345	0.6	220	0.9	**79.2**	2.3	0.0	4.8	0.5
Total*, Median**, Unknown***	-----	291*	9,026*	0.53**	4,513*	18.2***	7.9***	6.6***	1.0***	2.6***	5.5***

Table 1. Summary of Protein Predictions and their Various Properties
Unknown proteins were predicted to one of 17 organellar compartments using an SVM algorithm trained on 291 high-confidence marker proteins. The names of these compartments are displayed here with a brief description of the marker proteins descriptions from which they were created. The number of marker proteins and of unknown proteins classified to each respective compartment are then displayed as well as the median SVM scores which informed that classification. The global, median SVM score was used as a cut-off for 'predicting' proteins to a certain compartment, and the number of each compartment's predicted proteins are then shown. We then determine the percentage of proteins with each of the follow protein properties: transmembrane domains (TMD), signal peptides (SP), nuclear localization signal (NLS), and mitochondrial targeting sequence (mTS). Two functional genomic datasets were then used to determine the percentage of genes differentially expressed (DE) after trichocyst discharge and during reciliation (Arnaiz, Meyer, and Sperling 2020). The mTS peptides were predicted using TargetP 2.0 (Armenteros et al. 2019), NLS using NLStradamus with a prediction cutoff of posterior prediction cutoff of 0.6 (Nguyen Ba et al. 2009), and the remainder of features were downloaded from the ParameciumDB and were previously described (Arnaiz, Meyer, and Sperling 2020). Statistical significance was assessed using a chi-square test corrected for multiple testing ($p < 0.003$), and boldened values are those which are significantly larger than expected. The first column's order will serve as that of all subsequent figures unless specified otherwise.

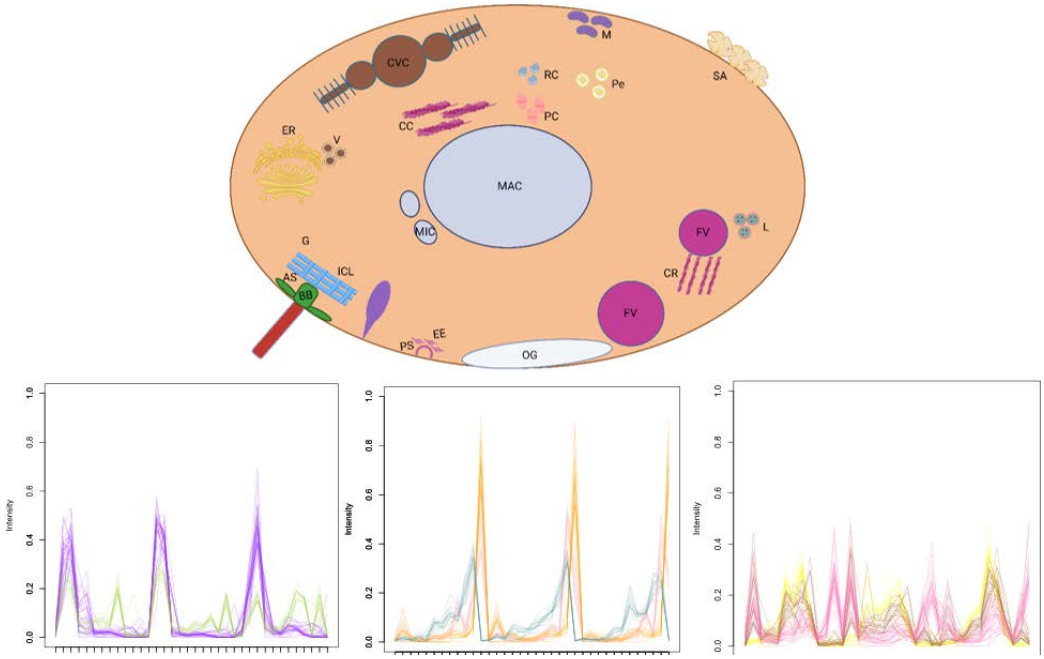

Figure 2.1. Experimental overview and the Generation of Protein Abundance Profiles
Well-fed batch cultures of P. tetraurelia cells were lysed and fractionated in two different ways to generate twelve biochemically distinct, subcellular fractions (top). After quantitative proteomics on three biological replicates, we observe unique distribution profiles for proteins corresponding to their steady-state localization pattern within the cell (bottom). The Y-axis of distribution profiles are intensity (normalized from 0 to 1), and the X-axis contains the triplicate centrifugal fraction names described in the Results section (i.e., MAC-1, 300g-1, … Sup-3). Using these profiles, we predicted localizations for almost 9,000 unique proteins covering a wide range of cellular structures. Our work greatly expands a growing model of the membrane-trafficking and organellar composition of Paramecium first illustrated by Allen and Fok (2000) by describing its cell biology in the light of that and subsequent knowledge. Created with BioRender.com.

Results:

Thousands of unnamed or ambiguously annotated proteins are confidently predicted to specific organelles in P. tetraurelia

To define the spatial proteome of *P. tetraurelia*, we used a protein correlation profiling (PCP) approach modified from the LOPIT-DC protocol described previously (Geladaki et al. 2019) with label-free quantification (LFQ). After optimizing cell culturing, lysis, and fractionation conditions, we performed LFQ analysis using an Orbitrap Fusion Lumos Tribrid Mass Spectrometer analyzed in ProteomeDiscoverer (Thermo) and then using the proloc package in the R programming language(Team and others 2013; Breckels et al. 2016). We detected over 11,000 total proteins whose coverage and peptide evidence were comparable to similar studies (Supp Text). We ran BUSCO and found these represented ~62% of core eukaryotic proteins and ~94% of core alveolate proteins, while the entire assembly of 40,460 proteins represented ~71% and ~99% of core eukaryotic and alveolate proteins, respectively. After processing and filtering, our proteomics dataset contained 9,026 proteins (Table 1).

Proteins that localize to a particular organelle or intracellular compartment are expected to have similar relative abundance profiles across subcellular fractions (Figure 2.1). We manually curated a set of 291 marker proteins to predict the localization of the remaining 8,735 proteins based on their shared protein abundance profiles representing seventeen diverse cellular compartments (Figure 2.2; Figure 2.10). Marker proteins are either experimentally validated to localize to their respective compartment or share homology with genes known to do so in other systems. We trained a support-vector machine (SVM) classifier with these marker protein profiles and classified all 9,026 proteins to one of the seventeen compartments before filtering low-scoring classifications to end up with 4,513 predictions (Supplement). To confirm that our application of biological information onto the data structure was appropriate, we compared SVM predictions with an equal number of clusters generated by the k-Means (KM) clustering algorithm (Likas, Vlassis, and Verbeek 2003). Qualitatively, KM clusters overlapped well with organellar

predictions (Figure 2.2; Figure 2.12). While only the cytosol, proteasome, and ribosome compartments were made up of a single KM cluster, most were overrepresented by one or few clusters. This was more so the case for compartments characterized by a single peak of abundance (Figure 2.10) as opposed to multiple, disconnected peaks. Indeed, the nuclear and membrane trafficking compartments were spready across three to seven KM clusters likely due to the heterogeneous nature of their abundance profile.

Proteins predicted to the same compartment were characteristic of that compartment (Table 1). Both nuclear compartments and the ribosomal predictions were enriched with nuclear localization signals (NLSs), as were mitochondrial proteins for mitochondrial targeting sequences (MTSs). Transmembrane domains (TMDs) were present in ~38% of predicted mitochondrial outer membrane (MOM) proteins, 66% of ER proteins, and ~62% of the insoluble membrane trafficking compartment. Proteins of the lysosome, ER, trichocyst matrix, or cell surface (i.e., surface antigens) were enriched with signal peptides (SPs) supporting their need to be translocated to the ER for processing and sorting. Since the range of expression values was often quite small for genes of the same predicted compartment, we performed de novo motif prediction on their putative promoter region using MEME (Bailey et al. 2009). We identified nineteen significantly enriched motifs, six of which were highly enriched near the annotated start codon (Figure 2.13). Of these six, two were specific for the trichocyst matrix, two for the proteasome, one for the ER, and one for the lysosome. These findings suggest coregulation between some components of the same organellar compartment as has been demonstrated in other contexts (Tsypin and Turkewitz 2017).

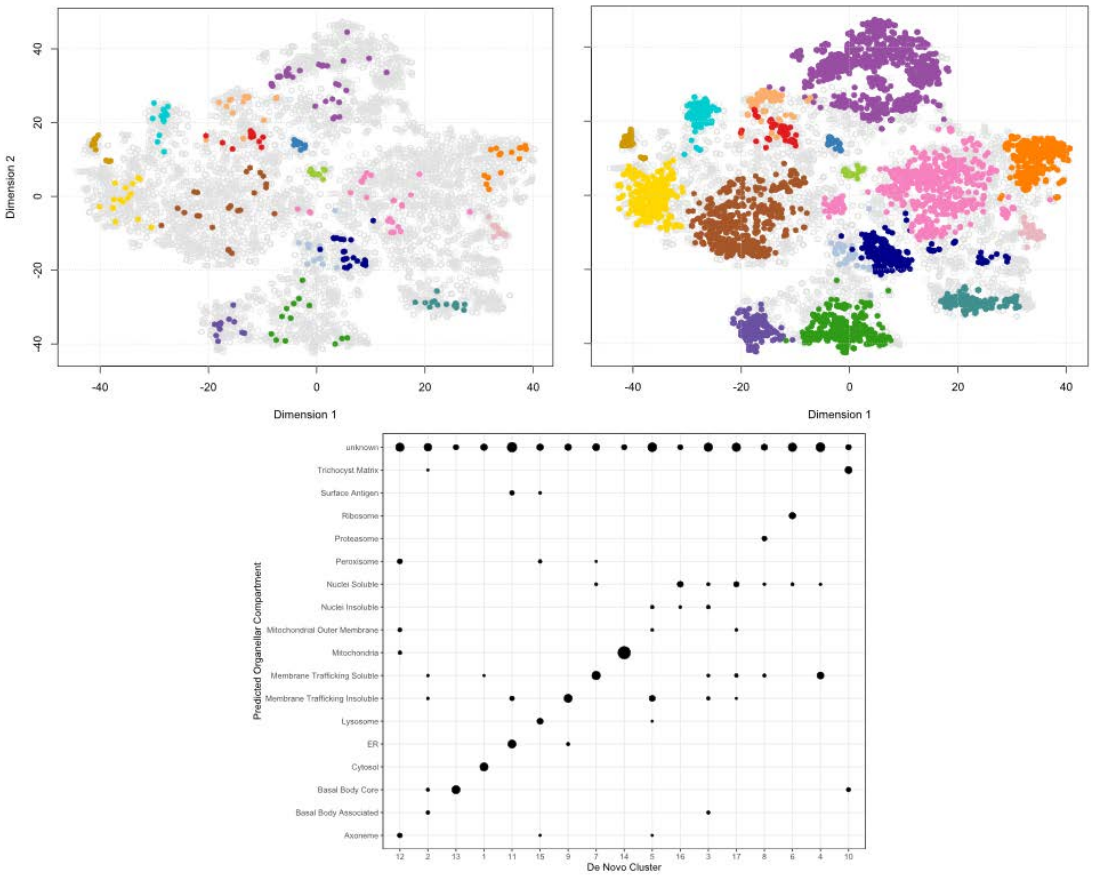

Figure 2.2. Organellar Maps of P. tetraurelia Overlap De Novo Clusters
The multidimensional protein abundance data were visualized using the t-SNE algorithm and overlayed with 291 manually curated marker proteins representing 17 organellar compartments (top left). An SVM algorithm was trained on these marker protein profiles and used to predict the localization status of >9,000 unknown proteins. Each classification was associated with an SVM score, and the median score across the entire dataset was used as a cut-off to determine the predicted protein constituents of each organellar compartment (top right). Predicted proteins were compared to de novo clusters generated using the k-means algorithm with an equal number of clusters as compartments. Most compartments correspond to one or a few clusters, while others were dispersed amongst many clusters (bottom). In no cases did a de novo cluster contain proteins not represented in one of the 17 organellar compartments.
Color code: Axoneme (red), Basal Body Association (navy blue), Basal Body Core (dark green), Cytosol (orange), ER (yellow), Lysosome (sky blue), Membrane Trafficking Insoluble (brown), Membrane Trafficking Soluble (pink), Mitochondria (violet), Mitochondrial Outer Membrane (light green), Nuclei Insoluble (blue/gray), Nuclei Soluble (dark blue), Peroxisome (tan), Proteasome (light pink), Ribosome (teal), Surface Antigen (gold), Trichocyst Matrix (indigo).

The cellular degradation machinery formed tight clusters around the lysosome and proteasome

Lysosomes are specialized organelles which play a variety of roles in breaking down cellular materials and defending against pathogens (Holtzman 2013). The methods underlying their discovery were key to the development of cell biology and served as the direct precursor to all spatial proteomics experiments (Christian de Duve et al. 1955; Mann 2020). Our lysosomal markers were a combination of peptidases, glycosidases, and annotated lysosomal membrane proteins which had a simple abundance profile characterized by high abundance in the 3K/5K fractions exclusively (Figure 2.10). We predicted 146 proteins to this organelle with enriched terms relating to digestive processes like protein and sugar degradation with a significantly lower isoelectric point (pI) than expected by chance (Figure 2.14). All 300 classified proteins share these same properties and are thus likely lysosomal. By contrast, the proteasome operates in a more targeted way to degrade individual proteins via a ubiquitin-dependent process (Tanaka 2009). In our study, the homologous components of the proteasome had more complex abundance profiles whose peak was always in the 120K fraction with a smaller peak in the 30K fraction (Figure 2.10). We predicted 66 proteins to this complex and found the singular representation of the “Proteasome” KEGG pathway with every 20S proteasome subunit and numerous components of the lid and base. The inclusion of all 175 predicted proteins did not expand any component of this complex but did include two modules of the “Ubiquitin mediated proteolysis” pathway (APC8: PTET.51.1.P1060100; UBLE1B: PTET.51.1.P0480203), however the remainder included spliceosomal and ribosomal subunits which raises doubts that all classified proteins are relevant in this context.

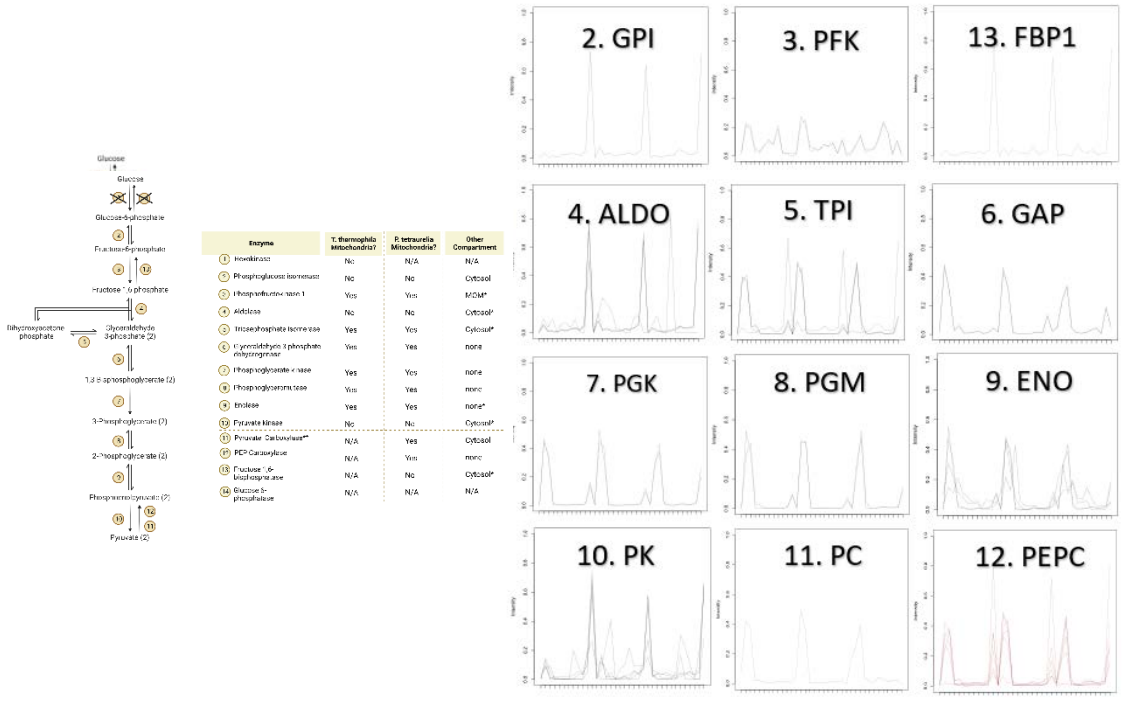

Figure 2.3. Differential Localization of Glycoytic/Gluconeogenic Enzymes in *P. tetraurelia* is Indicative of Substrate Channeling

Based on work from *Tetrahymena thermophila* (Smith et al. 2007), the glycolytic pathway of ciliates is thought take place in the mitochondria, despite most model eukaryotes performing cytosolic glycolysis and importing pyruvate and electron carriers to mitochondria.

(Left) Here, we show the glycolytic pathway and its associated enzymes, numbered one through ten, as well as the enzymes of the reverse reactions. In our study, we identified nine of ten enzymes and confirmed the sole mitochondrial location of enzymes 7-9. Enzyme 2 conflicts with *T. thermophila*, while the two copies of enzyme 3 were surprisingly predicted to the MOM. Enzymes 5 and 10 were present in multiple copies which were predicted to different locations. The first gluconeogenic enzyme, Pyruvate Carboxylase, has both mitochondrial and cytosolic variants while the next step is entirely mitochondrial.

Adapted from “Glycolysis and Glycolytic Enzymes” by BioRender.com (2020). Retrieved from https://app.biorender.com/biorender-templates.

(Right) The distribution profiles associated with these predictions are shown, numbered in accordance with their position in the pathway with gluconeogenic enzyme counts beginning after glycolytic enzymes. The former is also colored red. In cases where multiple paralogs were present, all were plotted simultaneously. The x-axis is fraction name as is the case in Figure 2.1, and the y-axis is normalized abundance.

Metabolic enzymes in the mitochondria, cytosol, peroxisome is indicative of mosaic biochemical pathways

Mitochondria in ciliates are large, double-membrane organelles with tubular cristae concentrated at the cell cortex and dispersed throughout the cell containing the machinery for oxidative phosphorylation and ATP synthesis (Powers, Ehret, and Roth 1955). The mitochondrial outer membrane (MOM) is the site of contact with other cellular organelles like the ER, and porin proteins act to translocate cytoplasmic proteins to the mitochondrial matrix, inner membrane (MIM), and intermembrane space (IMS) proteins in the presence of a signaling tag like the mitochondrial targeting sequence; mTS (Hartl et al. 1989). While most mitochondrial proteins are encoded by the nuclear genome, many are retained on its own genome and are universally functional within the mitochondria— making these mitochondrial ORFs ideal marker proteins. We observed a singular of abundance profiles characterized by high abundance in the 300g and 1K fractions (Figure 2.10) which was shared with a variety of annotated mitochondrial proteins. We saw another separate cluster of porins identified previously (Wideman et al. 2013) and used those as the basis for the mitochondrial outer membrane (MOM). The MOM compartment proteins had similar high abundance in the 300g/1K fractions but an additional 30K peak (Figure 2.10), and the marker TOM40 (PTET.51.1.P0280026) was orthologous to TOM40 in *T. gondii* and *S. cerevisiae* whose pattern was similar in their respective hyperLOPIT datasets (Barylyuk et al. 2020; Nightingale, Oliver, and Lilley 2019) (Figure 2.15). Only 40 proteins were predicted to the MOM, and most with detectable orthologs were either peptidases (e.g., pepN: PTET.51.1.P0240282; LTA4H: PTET.51.1.P0970095) or Rab GTPases (e.g., TBC1D20: PTET.51.1.P0090350) – the latter of which may regulate mitophagy (Sidjanin et al. 2016). The MOM compartment had an altogether different composition because the membrane-bound porin markers led to several probable-ER proteins and mitochondrial biogenesis factors being predicted here. These included the GTPase RHOT1/GEM1 (PTET.51.1.P0870070), DNAJC11 chaperone (PTET.51.1.P0320066) and two copies of the acyl-Coa reductase HSD17B12 (PTET.51.1.P0760003 and PTET.51.1.P0520130). GEM1 is a component of the ER-

mitochondrial encounter structure (ERMES) which physically links the two organelles and may explain other ER proteins (Kornmann, Osman, and Walter 2011).

The predicted mitochondrial proteome contains over 1,000 proteins and hosts the highest SVM scores of any compartment in this study (Table 1). This suggests the mitochondria are well resolved in this study, and indeed most genes fall within the large “Metabolism” pathway. All but one component of the TCA cycle (2-oxoglutarate dehydrogenase E2 component) was identified as were numerous components of the pyruvate and glyoxylate metabolic pathways (Figure 2.16). Each component of the electron transport chain had multiple predicted proteins represented except Complex III. These results support a massive expansion of the known protein inventory of mitochondria in *P. tetraurelia* and highlight the relationship between the MOM and ER.

Two mitochondrial pathways show clear mosaicism with other cellular compartments. The first is glycolysis (Figure 2.3), which typically occurs in the cytoplasm of model eukaryotes, but in the ciliate *T. thermophila* is thought to be mitochondrial due to the identification of six of ten glycolytic enzymes in purified mitochondria subjected to mass spectrometry (Smith et al. 2007). Our study confirms the mitochondrial localization of all these enzymes; however, many are paralogs with diverse localizations. Three of four phosphoglucose isomerase paralogs are mitochondrial while the other is cytosolic, and both paralogs of phosphofructokinase were predicted to the MOM compartment. Association with the MOM/IMS was observed in four glycolytic enzymes of *Arabidopsis thaliana* (Giegé et al. 2003), however this enzyme was not amongst them. From these data, we see evidence for substrate channeling, which describes the biased spatial distribution of enzymes for the purpose of directing its final product, e.g., preventing pyruvate from being used for non-respiratory processes (Sweetlove and Fernie 2018). The unambiguous mitochondrial localization for four of the last five glycolytic enzymes is contrasted with cytosolic and noisy nature of four of the first five. Why pyruvate kinase did not follow this pattern is unclear, but the enzyme was not detected in *T. thermophila* mitochondria and was found in our study in at least five copies with two cytosolic copies and three more ambiguously assigned copies.

The second compartment which biochemically overlaps with the mitochondria is the peroxisome. Despite the importance of ciliates in the early characterization of the peroxisome, very little work has been done on its proteomic composition in this lineage, and no proteins have been localized here in *Paramecium* (Müller, Hogg, and de Duve 1968; C de Duve 1969). Using marker proteins like the enzymes thiolase, isocitrate lyase, and many PEX and PMP membrane proteins (Supp Table), we predicted 89 proteins to this organelle including a handful of orthologs of *P. caudatum* genes described previously (Richardson and Dacks 2022) as having a putative peroxisomal targeting signal (PTS)*:* PTET.51.1.P1610060, PTET.51.1.P0250064, and PTET.51.1.P0150057. The entirety of putative peroxisomal membrane proteins were predicted to be peroxisomal, but many enzymes traditionally of the peroxisomal matrix were predicted to the mitochondria. Enzymes traditionally mediated by the PTS1 import pathway into the peroxisome included many predicted mitochondrial proteins in this study. Others, like isocitrate dehydrogenase, had paralogs with mitochondrial, peroxisomal, and cytosolic predictions; as seen in plants (Corpas et al. 1999). The antioxidant system displays this pattern as well, such as the mitochondrial superoxide dismutase (SOD2: PTET.51.1.P1060137) and cytosolic PRDX1 (Peroxiredoxin: PTET.51.1.P3140006). While the overlapping roles of the peroxisome and mitochondria is known in ciliates in the context of lipid metabolism (Krueger et al. 2022), these data suggest many shared biochemical modules linking the two organelles.

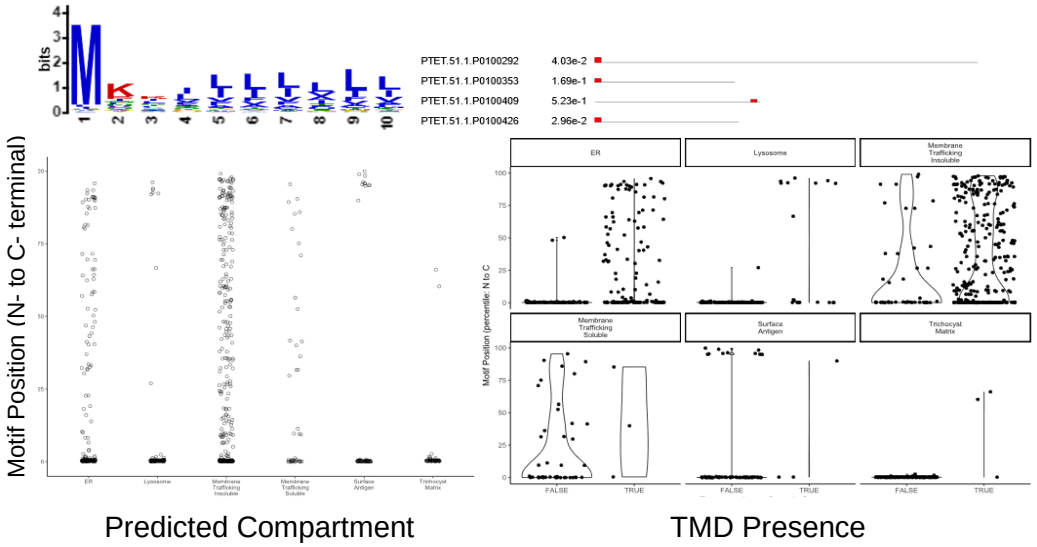

Figure 2.4. A Hydrophobic Motif Found in Nearly All Signal-Peptide Containing Proteins is Associated with Organellar Compartments and Transmembrane Domain Presence
De novo motif discovery of all signal peptide (SP) containing proteins yielded a highly hydrophobic motif present ~95% of the time. The sequence logo shows a high-information content methionine corresponding to the universal first position of all proteins (top left), however, the distribution of the motif across SP-containing proteins showed many C-terminal variants (example; top right).

We plotted the distribution of N- vs C-terminal variants (bottom left) and found stark differences between numerous compartments like the trichocyst matrix and insoluble membrane trafficking compartments. The trichocyst matrix had a similar pattern to that of the lysosome in that nearly all motif variants were N-terminal, but the membrane trafficking and ER compartments appeared to have a wider spread of motif positions. From left to right, the compartments read: ER, Lysosome, Membrane Trafficking Insoluble, Membrane Trafficking Soluble, Surface Antigen, Trichocyst Matrix. The surface antigens were intermediate to these patterns.

We then compared the TMD-presence with motif position and saw clear differences between TMD-absent proteins with N-terminal variants and TMD-containing proteins with more diverse positions. Since SP-containing proteins are likely to be translocated to the ER and processed before being shuttled to their cellular destination, these patterns support a different sorting mechanism for those proteins destined for membranes and those not.

Cortical and membrane trafficking proteins exhibit different abundance profiles based on their relationship with the endoplasmic reticulum and Golgi apparatus

The traditional view of membrane trafficking involves the processing and shuttling of proteins from their origin of synthesis to their cellular and extracellular destinations (Schekman 1985). These general activities are common across all eukaryotes, but some protists perform highly regulated versions of these processes (Plattner and Kissmehl 2003).

Many ciliates contain subcortical organelles broadly called extrusomes (Rosati and Modeo 2003) which in *Paramecium* are the spear-shaped trichocysts used primarily for defense (Plattner 2017). The body of trichocysts are made up of trichocyst matrix proteins (TMPs), which are highly abundant and post-translationally processed through the Golgi apparatus. In our study, TMPs formed a tight cluster due to their singular abundance only in the 300g fraction (Figure 2.10). Using fourteen annotated TMPs, we predicted 220 proteins to the trichocyst matrix compartment, most of which were annotated TMPs which were both highly expressed and highly acidic (Figure 2.14). Adjacent to trichocysts are basal bodies which physically anchor cilia to the rest of the cell. Basal bodies are themselves embedded in several networks of cytoskeletal proteins like the infraciliary lattice (Aubusson-Fleury et al. 2013; Garreau DE Loubresse et al. 1988). In our study, we constructed core and peripheral basal body compartments built around epiplasmins of the epiplasm and SF-assemblins of the striated rootlet, respectively (Supplement). The basal body core contained 387 proteins, many of which have known roles in basal body and ciliary function (Jerka-Dziadosz et al. 2010; Gogendeau et al. 2020) like SAS6 (PTET.51.1.P0200220) and NPH4 (PTET.51.1.P0220034). While not directly a part of the basal body, alveolar sacs are intracellular calcium stores conserved across alveolata and contain a family of conserved alveolins (Gould et al. 2008) (PTET.51.1.P0130289, PTET.51.1.P0190258, and PTET.51.1.P0660154) predicted here as well. The basal body associated compartment had only 30 predictions, and these were mainly the additional centrins and SF-assemblins related to the marker proteins as well as a handful of signaling enzymes like the bi-functional DHFR-TS (PTET.51.1.P0620252). The final of these cortical organelles is the axoneme itself which makes

up the structural components of the cilia-proper grounded in numerous dynein and tubulin marker proteins. The abundance profile of this and the basal body associated compartment were highly similar in the position of their ‘peaks’ of abundance but differed starkly in the ‘valleys’ (Figure 2.10). Using numerous axonemal dynein and tubulin marker proteins, we predicted 59 proteins here, and roughly half were differentially expressed (DE) after reciliation (Arnaiz et al. 2010) (Table 1). In addition to the structural proteins, we also see many signaling enzymes like adenylate kinase (e.g., PTET.51.1.P0730129) and calcyphosin (e.g., PTET.51.1.P0100152) predicted here as well as the golgin PTET.51.1.P1260140. These non-structural proteins suggest a tight association between the structural components of the cilia and the signaling molecules modulating its activity.

The ER compartment in this study was constructed around three experimentally validated proteins (PDI1-1: PTET.51.1.P0980088; ptSERCA: PTET.51.1.P0640022; HSP70Pt08: PTET.51.1.P1020045) as well as a number of putative chaperone and ER membrane proteins. ER proteins peak in abundance between the 9K and 30K fractions with little signal elsewhere (Figure 2.10). We then predicted 342 proteins highly enriched in SPs and overwhelmingly represented the KEGG pathway “Protein processing in endoplasmic reticulum” (Figure 2.17) including the signal peptidases SPCS3 (PTET.51.1.P0880138) and SEC11 (PTET.51.1.P1080019). The predicted ER proteins contained the highest proportion of genes differential expressed after trichocyst discharge, further supporting previous observations that trichocyst maturation is regulated by ER-to-Golgi transport (Arnaiz et al. 2010). A compartment spatially distinct but similar in its abundance profile was that of the surface antigen proteins (sAGs) whose processing and secretion is well-understood (Baranasic et al. 2014). These large proteins coat the cell surface and play important roles in signaling, and both traditional sAGs and mini “mAGs” formed a tight cluster containing 57 predicted proteins. Most of these were annotated ‘Paramecium surface, but others included the glycosyl carrier protein DOLPP1 (PTET.51.1.P0100142 and PTET.51.1.P0590073), and the glycoproteins LRP2 (PTET.51.1.P0120163) and MPDU1 (PTET.51.1.P0790038). Only the MPDU1 ortholog (PTET.51.1.P0790038), had both an SP and TMD, and in humans, this is an ER-resident

responsible for glycosylating many target proteins (Kranz et al. 2001). These associations suggest a deep connection between the ER's glycosylating machinery and secretion system facilitating sAG transport to the cell surface (Fiedler and Simons 1995).

We created two abstract 'Membrane Trafficking' compartments using membrane trafficking proteins identified computationally (Richardson and Dacks 2022), ciliary proteins identified experimentally by mass spectrometry (Yano et al. 2013), and experimentally validated endomembrane proteins. These proteins shared the ER's high abundance between 9K to 30K but differed in one or two other fractions (Figure 2.10). One compartment was 'insoluble' in that it had high abundance in the 300g and none whatsoever in the Sup fraction. This contained ciliary membrane proteins (e.g., PMCAs), rabs, coatomers, and known residents of the contractile vacuole and Golgi. Of the 540 proteins predicted here, ~62% had predicted TMDs (Table 1). "Membrane Trafficking" and "Exosome" were the best represented KEGG BRITE classes, while the pathway "Protein processing in endoplasmic reticulum" was well represented with different components than the ER proper (Figure 2.18). Two markers localized to the osmoregulatory contractile vacuole complex (CVC), PtSTO1c (PTET.51.1.P1670084) and NSF2 (PTET.51.1.P0410185) aided in predicting other CVC proteins, e.g., VATA2_1 (PTET.51.1.P0380139), Rab11c (PTET.51.1.P0430208), and PtSYB2-2 (PTET.51.1.P0670153). We also see the phagosome proteins VATA6 and VATA9 (both ohnologs) predicted here alongside the ortholog of TtVPS13A (PTET.51.1.P0160355) localized to the phagosome membrane in *T. thermophila* (Samaranayake, Cowan, and Klobutcher 2011). The prediction of the ER-localized VATA7_1 (PTET.51.1.P0580140) and Golgi-localized VATA8_2 (PTET.51.1.P0080391) further supports a broad, membrane-associated trafficking compartment.

The soluble membrane trafficking compartment had a similar profile to the insoluble compartment but exhibited high abundance in the Sup fraction. We included different membrane trafficking proteins (Richardson and Dacks 2022), SNAREs (Kaur et al. 2022), and intraciliary transport proteins (IFTs). Of the 625 proteins predicted here, only ~2.3% had TMDs. Notable in these predictions are many components of endocytosis such as AP2M1 (PTET.51.1.P0260049), EEA1 (PTET.51.1.P1650038), and HSPA1 (also called HSP70Pt01: PTET.51.1.P0330220). Of

the cytoplasmic or phagosome-associated actins described previously (Sehring, Mansfeld, et al. 2007), we found ACT1_1 (PTET.51.1.P0130204), ACT1_5 (PTET.51.1.P0850133), and ACT5_1 (PTET.51.1.P1560086). Again, the pathway “Protein processing in the endoplasmic reticulum” was highly represented, but we see largely different modules than either the insoluble membrane trafficking compartment or ER (Figure 2.17). Taken together, the differences in abundance profiles allow us to spatially resolve the ER proper from the membranous and cytoskeletal components of the ER-Golgi related trafficking systems.

We performed *de novo* motif discovery on all 901 SP-containing proteins and found the highly hydrophobic motif MKKJJJJLLJ present ~95% of the time (Figure 2.4). In about 80% of cases, this motif began with the N-terminal Methionine residue immediately upstream of the predicted SP, but many were C-terminal. The trichocyst matrix and lysosomal predicted proteins had mostly N-terminal variants while the ER and both membrane trafficking compartments had a wider spread of motif positions, and surface antigens were either N- or C-terminal with no intermediate position. This raises the possibility of multiple sorting pathways present in *Paramecium*. Indeed, protein sorting in model eukaryotes is done through both signal recognition particle (SRP)-dependent or independent mechanisms depending on the presence of a hydrophobic N-terminal sequence and C-terminal TMD (Gemmer and Förster 2020). SRP directly facilitates the translocation of proteins with N-terminal hydrophobic stretches, while “tail-anchored” proteins are processed via their C-terminal TMDs (Borgese, Colombo, and Pedrazzini 2003). While only ~27% of proteins with an N-terminal MKKJJJJLLJ had TMDs, ~84% with the C-terminal variant did (Figure 2.4). This pattern was far stronger for the trichocyst matrix and lysosome compartments, which may highlight the former’s role as a lysosome-related organelle (Kuppannan et al. 2022). These findings suggest that the predominant sorting mechanism for non-membranous proteins in *Paramecium* involves a hydrophobic N-terminus upstream of a hydrophobic SP. TMD-containing proteins with SPs do not require a hydrophobic N-terminus but instead are processed through a different mechanism. This former may represent the ‘cargo’ proteins being trafficked, while the latter represents the proteins responsible for trafficking.

Supervised classification can distinguish between the ribosomes and nucleoli, chromatin and nucleoplasm, but not the MAC and MIC proteomes

A unique feature of ciliates is their nuclear dimorphism in which a large MAC performs nearly all the gene expression from a highly polyploid and developmentally excised genome, while a transcriptionally silent MIC acts analogously to the germline genome (Prescott 1994). To ensure a consistent pelleting behavior of the MIC and MAC, we used standard culturing methods to maintain our cells in vegetative (VEG) growth after initiating the culture with starvation-induced autogamy (Sonneborn 1970). In optimizing our lysis and fractionation protocol, we used an antibody raised against a conserved region of Histone H3 and demonstrated that no histone proteins were present in the Sup fraction (Figure 2.7), but DAPI staining confirmed no intact MACs in any fractions except the MAC fraction itself. After proteomic analysis, we saw a variety of distribution profiles for annotated nuclear proteins such as histones, transcription factors, RNA polymerase subunits, and nucleoporins, but none which clearly distinguished the MAC and MIC proteomes. We used these profiles and a handful of MAC-localized proteins to build insoluble and soluble nuclear compartments based on high abundance in the 300g or Sup fractions. A total of 293 proteins (42 insoluble, 251 soluble) were predicted to be nuclear after filtering low SVM scores (Table 1). This is almost certainly an underestimate, but the complexity of this compartment necessitated a conservation approach to characterizing it.

The soluble nuclear compartment included the transcriptional and DNA replication machinery, while the insoluble compartment included histones and components of chromosome segregation machinery. In contrast to the membrane trafficking compartments, insoluble here does not necessarily mean TMD-containing but instead is more deeply connected with chromatin and large macromolecular complexes. Despite only 42 predictions, this compartment contains four of five condensin components necessary for cell division SMC2 (PTET.51.1.P0330075), SMC4 (PTET.51.1.P0410063), YCS4 (PTET.51.1.P0050262), and BM1 (PTET.51.1.P0480269). The 251 soluble nuclear proteins have excellent representation of the spliceosome, RNA polymerase I/II/III, nuclear exosome, and 5 of 7 mini-chromosome maintenance complex

members; MCM2 (PTET.51.1.P1180147), MCM3 (PTET.51.1.P0380097), MCM4 (PTET.51.1.P0330339), MCM5 (PTET.51.1.P0880073), and MCM6 (PTET.51.1.P0570197).

We noticed that distribution profiles of different components of the same nuclear complexes were often quite different. For example, the core helicase of the basal transcription factor TFIIH (ERCC2: PTET.51.1.P0110310 and ERCC3: PTET.51.1.P1610014) were both predicted to the insoluble nuclear compartment, while the TFIID proteins TAF5 (PTET.51.1.P0020465) was a soluble marker and aided in predicting its paralog TAF1 (PTET.51.1.P0390139). We identified three of four paralogs of the TATA-binding protein TBP but saw that PTET.51.1.P0250306 was predicted to the soluble nuclear compartment while PTET.51.1.P0210036 and PTET.51.1.P0360255 were classified to the insoluble nuclear compartment with low SVM scores. These conflicts arise through differential abundance in the Sup fraction which reflects its decoupling from the insoluble chromatin. The basal transcription machinery contains proteins with many roles, such as ERCC's dual role in DNA repair (Wood et al. 2001), and this diversity may obscure a more consistent profile as seen in RNA Pol subunits. We also used putative nucleolar proteins as soluble nuclear markers and predicted many new rRNA maturation/biogenesis factors such as RRP5 (PTET.51.1.P0080157), UTP14 (PTET.51.1.P0470044) and NOM1 (PTET.51.1.P0450089), RNA Pol I subunits like RPA2 (PTET.51.1.P1360088) and RPC2 (PTET.51.1.P0370108), and others with ribosomal descriptions. Our cytoplasmic ribosome contained 14 ribosomal *S. cerevisiae* orthologs as markers and predicted 181 proteins in total representing almost every eukaryotic, ribosomal subunit (Table 1). Combined with our identification of mitochondrial ribosomes, this study confidently aides in the resolution of annotated ribosomal subunits between the cytoplasm, mitochondria, and nucleolus.

The nuclear membrane separates the nucleoplasm from the cytoplasm in a double-layer structure containing nuclear pore complexes (NPCs) made up of cytoplasmic and transmembrane components, a disordered FG NUP basket, inner and outer ring, and linker proteins (Strambio-De-Castillia, Niepel, and Rout 2010). In *T. thermophila*, the cytoplasmic fibril protein NUP98 has MAC- and MIC-specific variants with GLFG and NIFN repeats, respectively, thought to

differentially regulate export/import between each nuclei and the cytoplasm (Iwamoto et al. 2009). We did not detect the *T. thermophila* MAC-Nup98 (TTHERM_00071070) ortholog (PTET.51.1.P0490206) in this study, but we did detect three of the five copies of the MIC-Nup98 (TTHERM_00530720): Mic-NUP98A (PTET.51.1.P0010578), Mic-NUP98C (PTET.51.1.P0950099) and Mic-NUP98D (PTET.51.1.P0750016). Mic-NUP98B had a noisy profile likely caused by its low number of PSMs, but both Mic-NUP98A/C were classified to the soluble nuclear compartments with low SVM scores. Only the transmembrane NUP210 had identifiable copies in *P. tetraurelia* (PTET.51.1.P1240016 and PTET.51.1.P0880039), but its abundance profile was different than NUP98. The other identifiable NUP proteins in *P. tetraurelia* (NUP37: PTET.51.1.P1740028; NUP42: PTET.51.1.P0780138; NUP43: PTET.51.1.P0110417) are homologous to coatomers and were predicted nuclear, axonemal, and membrane trafficking proteins, while the putative nuclear membrane lamin receptor (LBR: PTET.51.1.P0020334) was ER-like. It is tempting to speculate that the nuclear membrane is pelleting differentially from other nuclear components due to its physical association with the ER, but the relationship between ciliary and nuclear trafficking may explain the overlapping protein machinery seen here (Kee and Verhey 2013). Indeed, cortical proteins like IFT57 often display dual localization with the cilia and MAC (Shi et al. 2018).

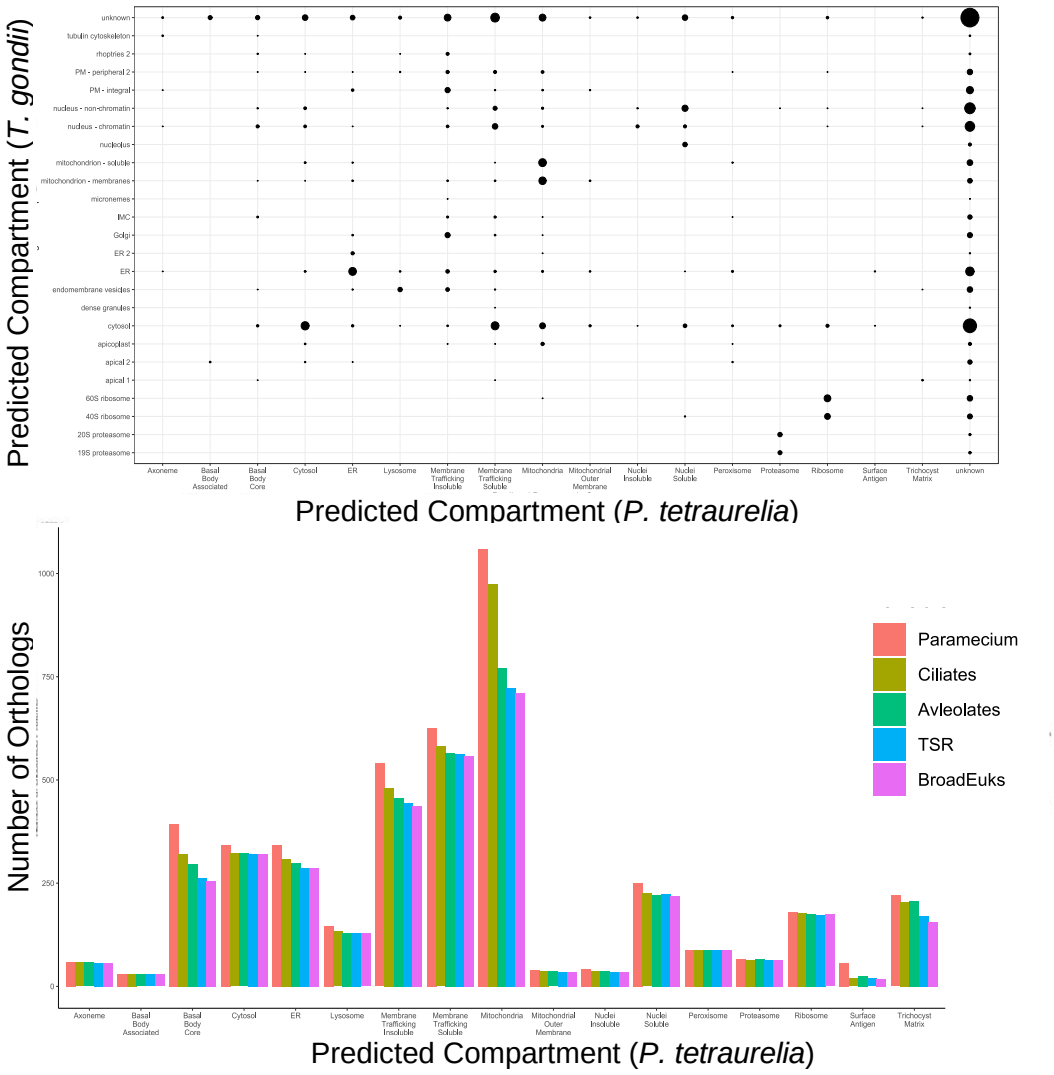

Figure 2.5. The Evolution of Organellar Compartments and Comparison with *Toxoplasma gondii*
We compared organellar predictions in our study to those made in Barylyuk et al. (2019) of Toxoplasma gondii using many-to-one orthologs present in both datasets (top). Many compartments are only described in one or another study due to the lack of the existence of that cellular structure in the organism or the lack of resolution in compiling that structure's marker proteins. However, those compartments present in both studies have a high degree of overlap. This comparison highlights the cellular regions underrepresented in each study, such as our nuclear compartments and their membrane trafficking compartments.

We determined orthology relationships across the tree of eukaryotes using the EukProt 3.0 database (Richter et al. 2022). We search iteratively beginning within the *Paramecium* genus (which was 100% represented in all protein prediction) followed by all ciliates except for *Paramecium*, then all alveolates except ciliates, all of TSAR except alveolates, and all other eukaryotes except TSAR. We observed a large number of ciliate-specific mitochondrial proteins which are entirely absent from the rest of eukaryotic diversity. This finding suggests either a number of highly divergent mitochondrial genes or the emergence of de novo genes private to ciliates and functional within the mitochondria.

Poorly localized P. tetraurelia proteins are orthologous to poorly localized proteins in the apicomplexan Toxoplasma gondii

We compared our PCP dataset to a *T. gondii* hyperLOPIT dataset (Barylyuk et al. 2020). We generated many-to-one relationships between 3,851 *P. tetraurelia* orthologous to 1,377 *T. gondii* proteins and immediately noticed strong overlap between the same compartments in both organisms (e.g., ribosomal and proteasomal). The consolidation of many types of cytoplasmic proteins (e.g., cytoskeleton) into the cytosol compartment of *T. gondii* explains their spreading across the cytosol, soluble membrane trafficking, mitochondrial, and unknown compartments in our study. The three nuclear *T. gondii* compartments were large, while our two nuclear compartments were smaller, but they overlapped well. The relationship between *T. gondii* nuclear compartments and our soluble membrane trafficking compartment may be caused by the behavior of the nuclear membrane in our study, but chromatin components remain unclear. Strikingly, most unknown *T. gondii* proteins are orthologous to unknown proteins in our study, and these represent a unique class of genes with ambiguous functions in both organisms. These may be enriched in dynamically localizing proteins, like nuclear import gene (TNPO1: PTET.51.1.P0010163) and nuclear export gene (PCID2: PTET.51.1.P0030147), but functional investigation will be needed to illuminate their role in alveolate biology.

A quarter of P. tetraurelia's mitochondrial proteins have no orthologs beyond ciliates

We then determined the evolutionary history of the proteins predicted to each organellar compartment using the EukProtv3 database (Richter et al. 2022). Briefly, this approach determines high confidence orthologs in species of various phylogenetic lineages using both fully assembled genomes and incomplete transcriptomic data to better characterize the representation of sequences across eukaryotic diversity. We confined our query to only the proteins in our organelle dataset and used all 9,026 classified proteins and then iteratively searched the TSAR supergroup (telonemids, stramenopiles, alveolates, and Rhizaria) first within the *Paramecium*

genus (excluding *P. tetraurelia*), then across ciliates (excluding *Paramecium*), alveolates (excluding ciliates), all TSR (excluding alveolates), and all other eukaryotic lineages for which we have data (excluding TSR) (Figure 2.5). Our null expectation was that each compartment would be similarly depleted of orthologs the more diverged from *P. tetraurelia* the lineage was. This seemed to be the case for most compartments, however the classified mitochondrial proteins displayed the starkest deviation from this pattern due to the significant underrepresentation outside of the ciliate phylum. While ~92% of mitochondria-predicted proteins had ciliate orthologs, only 75% had alveolate orthologs compared with ~89% and ~85% of all proteins in this study, respectively. This is noteworthy because of the previously discussed biochemical peculiarities of ciliate mitochondria, but differentially localized proteins alone cannot explain entirely new mitochondrial genes private to ciliates. This may be due to a combination of ciliate-specific or highly divergent mitochondrial genes as is seen in ATP synthase (Balabaskaran Nina et al. 2010; Dudkina et al. 2010).

Discussion

Our study is the first to examine global patterns of protein localization in any ciliated protozoa– a highly diverse lineage of eukaryotes whose complexity rivals that of multicellular organisms but within the confines of a single cell. So much is known about protein localization in a small handful of model organisms while most eukaryotic lineages remain poorly explored. Spatial proteomics offers the ability to leverage *a priori* knowledge to produce bursts of understanding using robust experimental and bioinformatic procedures (Breckels et al. 2016; Arslan et al. 2022). Central to this effort is the careful curation of marker proteins whose distribution profiles would provide the expectation for their resident organelle, protein complex, or broad cellular compartment. After many rounds of assembly, we continued the analysis with 291 marker proteins spanning 17 compartments each capturing some unique element of ciliate biology. However, as our knowledge of *Paramecium* grows, and more proteins are subject to cell biological interrogation, this dataset will provide a deep resource for reanalysis and new interpretation in that light. We

expect the nuclear and membrane trafficking compartments to contain numerous sub-compartments whose cell biological role is not currently clear from our work.

These steady-state protein abundance measurements provide a unique perspective on the highly dynamic nature of membrane trafficking, which we show pervades all but a few components of the plasma/ciliary membrane. It is interesting to speculate as to why the TMPs of the trichocyst matrix and epiplasmins of the core basal body appear as singular spikes in abundance while the SF-assemblins and centrins of the associated basal body compartment appear as a heterogeneous mixture of peaks and valleys with experiment-to-experiment inconsistencies (Figure 2.10). In the former case, post-Golgi vesicles containing immature TMPs have been shown to exist during trichocyst biogenesis (Laurence Vayssié, de Loubresse, and Sperling 2001), but we see no signs of that dynamic behavior even though our cells are actively growing and dividing and presumably rebuilding their trichocysts. In comparison, large sAG proteins display two to three peaks in abundance (depending on the experiment) which indicates discrete populations of proteins present in different fractions, possibly before and after proteolytic processing. Both are enriched in SPs, depleted of TMDs, glycosylated, and present at the cell cortex; but their patterns are highly dissimilar. This raises interesting questions about the mechanisms behind protein sorting underlying diverse compartments in *Paramecium*, and here we provide clues as to the nature of that sorting (Figure 2.4).

An unexpected finding from this work was the mosaic glycolytic pathway (Figure 3). We highlight the mitochondrial nature of six of ten of these enzymes, but many systems perform mitochondrial glycolysis, and this may be ancestral state for the pathway in the TSAR supergroup (Rio Bártulos et al. 2018). Even in classic model systems, glycolytic enzymes may be spatially arranged in a way to better direct pyruvate (and tRNAs) to the mitochondria even if the enzymes are not localized there per se, as seems to be the case with the MOM-associated late glycolytic enzymes in *S. cerevisiae* (Brandina et al. 2006). An example of so-called “substrate channeling” is thought to prevent pyruvate from being used for non-respiratory processes like amino acid biosynthesis (Sweetlove and Fernie 2018), however no such phenomenon has been observed in

a ciliate to our knowledge. While we provide a few key examples of these processes, further interrogation of our dataset will surely yield more discoveries of this kind.

Supplemental Text:

Protein Search Database:

The protein database from which predicted peptide spectra were searched against observed spectra came from three sources. First, the most recent genome annotation for *Paramecium tetraurelia* strain 51 (https://paramecium.i2bc.paris-saclay.fr) formed the base with 40,460 protein-coding genes. We used clustering software CdHit (Li and Godzik 2006) to combine proteins that were greater than 99% similar at the sequence level and were either identical in length or differed by 1 amino acid. This process resulted in 318 protein clusters which contained 735 proteins with each cluster made up of 2-9 proteins per cluster as well as 39,725 individual proteins, summing to 40,043 proteins representing the nuclear-encoded proteome. The predicted mitochondrial proteome for *P. tetraurelia* contains 46 annotated proteins, however none of these were identified in preliminary surveys of the mass spectrometry data, motivating us to generate 283 ORFs using ORFinder (Rombel et al. 2002). Of these ORFs, we identified 31 in nine subcellular fractions (three technical replicates of three experiments) enriched with mitochondria (1K) using both ProteomeDiscoverer and MaxQuant (Figure 2.6). These 31 were appended to the nuclear-encoded proteins. Our cell culture is monoxenic between *P. tetraurelia* and its prey bacterium *Klebsiella pneumoniae* (also called *Enterobacter aerogenes*), and we included its Uniprot database as well. Finally, to remove common laboratory contaminants, we included the cRaP database (http://www.theGPM.org/crap).

Data Structure and Summary

We identified 12,579 proteins in our dataset, of which 11,856 were *P. tetraurelia*, 659 were *K. pneumoniae,* 34 were cRaP, and 31 were mitochondrial-ORFs. Confidence for each protein was determined using a decoy-based FDR measurement performed in ProteomeDiscoverer (Orsburn 2021). We removed low confidence (333 proteins) as well as those with fewer than 10 PSMs (2317 proteins) determined after iterative visualization of the data structure. Despite originally hoping that *K. pneumoniae* proteins could serve as markers of phagosomes, we found a wide spread of distribution patterns and thus removed them from the analysis (659 proteins). Many *K. pneumoniae* proteins had abundance profiles that were either high only in the MAC and 300g fractions (contamination) or spread across fractions 9K-30K (possible digested proteins), but the former prevented our use of the latter for identifying phagosomal proteins. We finally removed the only protein lacking a unique peptide underlying its identification. The median protein in our dataset ~388 amino acids long and was identified by ~7 peptides (4 of which were unique) covering ~27% of the protein and supported by ~90 PSMs. We compared this to recent hyperLOPIT experiment (Barylyuk et al. 2020) and saw similar or better values based on the range of values from their three experiments: 16-23% coverage, 8-10 peptides per protein, 10-20 PSMs per peptide.

The data structure was assessed using principal component analysis (PCA) (Figure 2.8) and the t-stochastic neighbor embedding (t-SNE) algorithm (Figure 2.9). The data were well structured allowing the visualization of these 36th dimensional data across a few PCs wherein more than half of the variation was explained by PC1 and PC2. The t-SNE projection is purely for visualization, and we used this to assess the extent to which imputation affected our data structure.

Data Imputation

Clustering and classification algorithms require each observation (i.e., protein) to have a value for each measured variable (i.e., abundance in centrifugal fractions) or else that observation is removed from the analysis. Our decision to produce twelve fractions for each of three

experiments likely led to a higher number of missing values than had we generated fewer fractions per experiment. The general pattern was that missing values were influenced by the number of PSMs from which each protein was identified, but many proteins had both high PSMs and were missing quantification values. We noticed that many highly important proteins across diverse cellular functions had at least one missing value. The most extreme of this group were the epiplasmins and of the basal body which were highly abundant in only the MAC and 300g fractions (discussed below) only and typically had missing values for most other fractions. To include these proteins and other proteins, we performed two types of data imputation. The first step involved the averaging of neighboring fractions to recover observations with one or few missing values in between fractions of higher abundance which may be caused by randomness. We kept the three experimental datasets separate before imputing to prevent a Sup fraction from informing the imputation of the MAC fraction and vice versa. This was implemented in the proloc R package (Breckels et al. 2016) which also imputes the global minimum into fractions adjacent to either the MAC or Sup fractions if they are the only fractions with abundant proteins measured. This step increased the number of 'complete' proteins from 2345 to 4353. The next step simply imputed zeros to remaining fractions assuming that these zeros were caused by non-random bias (i.e., the protein was lowly abundant in those fractions) which increased our dataset to 9026 proteins.

Marker Protein Curation

Marker proteins serve a pivotal role in spatial proteomics experiments due to their behavior (i.e., relative abundance profile) being used as the expectation for all proteins with which it colocalizes. Incorrectly assigning marker proteins will result in the inappropriate interpretation of the resulting data, and this problem is amplified in non-model systems in which few (if any) proteins have been directly studied *in vivo* for their subcellular localization. *P. tetraurelia* has had many dozens (perhaps hundreds) of proteins studied using a variety of molecular methods like GFP-tagging, immunostaining, and western blotting of enriched cellular fractions. We thus had a core of

experimentally validated proteins around which we could build robust compartments of at least 13 proteins in accordance with previous guidelines (Breckels et al. 2016). This core was combined with proteins of similar distribution profiles and properties such as GO terms, PFAM/INTERPRO domains, and homology to proteins of known functions. In order to negate the artificial clustering observed for many pairs of highly similar gene duplicates, we did not include WGD1 ohnologs in the same compartment except for the mitochondrial outer membrane (MOM) compartment discussed below. When possible, compartments were built with unrelated marker proteins. We will go organelle-by-organelle justifying the creation of each compartment and the inclusion of its constituent marker proteins. Most qualitative descriptions of profiles will focus on that of experiment 1 with mention to experiment 2 and 3 when they differ in key ways.

First, the cilia are responsible for motility and feeding in Paramecium which has long served as an important model system in understanding its structure and function across eukaryota. Cilia have a membranous and non-membranous component, the former consisting largely of mainly ion pumps/channels and signaling enzymes, and the latter consisting of more structural components of the axoneme and its associated motor proteins (Yano et al. 2013). Our first ciliary compartment was dubbed the axoneme due to the clustering BUG22p and DHC-6 both localized directly to the cilium in *P. tetraurelia* (Laligné et al. 2010; Asai et al. 1994; Kandl, Forney, and Asai 1995). We also included the calmodulin binding protein PCM1 localized to small cortical vesicles (Chan, Saimi, and Kung 1999) suggesting that this compartment does have some degree of feedback between trafficked proteins and the structural constituents of ciliary axoneme. We included seven unrelated dynein genes with the INTERPRO IDs IPR026983 (Dynein heavy chain) or IPR026975 (Dynein heavy chain 1, axonemal) as well as three tubulins with GO:0005874 (microtubule). Two more genes were orthologs of the *Chlamydomonas reinhardtii* flagellar radial spokes proteins each playing important roles in the flagellar beating and axoneme assembly (Yang et al. 2006). The distribution profiles generally had high abundance in the 300g fraction with subsequent drops and rises of abundance peaking in fractions 3K, 9K, and 15K with dips in 1K, 5K, 12K, and 30K with virtually no abundance in fractions. This pattern was strongest in experiment 1, but in experiment 2, much higher abundance in the 300g fraction

depleted some abundance in the subsequent rises and falls, and in experiment 3 the pattern is significantly noisier with higher peaks of abundance in the 3K fractions.

Anchoring cilia to the cytoskeleton and regulating its activity are the subcortical basal bodies. The basal body is embedded in a superficial cytoskeletal network called the epiplasm which is delineated from the core basal body by the transition zone (Tassin, Lemullois, and Aubusson-Fleury 2015). The core basal body and its associated appendages form the kinetid(Lynn 1981). We built the core basal body compartment from proteins exclusively directly localized to the basal body or had a paralog which was. This included SAS6 (Jerka-Dziadosz et al. 2010), PtCen2a (Ruiz et al. 2005), FOR20a (Aubusson-Fleury et al. 2013), and several epiplasmins (Aubusson-Fleury et al. 2013). The epiplasmins all stained the basal body directly and ranged across sub-structures like the ring, rim, terminal plate and plasma membrane. A recent review (Valentine and van Houten 2021) described the NPHP module of the cilium associated with the transition zone with the basal body, and we included NPHP4 as a marker as well. Finally, two more *C. reinhardtii* orthologs of the centriole proteins FAP45 and POC1 were included due to their role in regulating microtubule structure and centriole duplication (Owa et al. 2019; Keller et al. 2009). All of these proteins had abundance profiles consisting of a single peak in abundance in the MAC/300g fractions with higher abundance in 300g and then no abundance elsewhere. Associated with the basal body is a network of numerous cytoskeletal proteins, two of which we combined here: SF-assemblins (SFAs) of the striated rootlet and centrins of the infraciliary lattice (ICL). The striated rootlet (also called the kinetodesmal fibre) forms a physically connection between the basal body and the anterior pole of the cell (Tassin, Lemullois, and Aubusson-Fleury 2015), while the ICL is a mesh-like network spanning the entire cell surface contacting the proximal end of basal bodies (Garreau De Loubresse et al. 1988). The abundance profile of this compartment was similar to the axoneme in that it had high abundance in 300g, 3K and 9K fractions, but the intermediate fractions were far lower in abundance than those of the axoneme which were smoother and less extreme. We included eight SFAs localized directly to the striated rootlet (Nabi et al. 2019) as well as paralogs of three others which were. We included two centrins localized to the ICL (Gogendeau et al. 2008).

Another cortical compartment was built from a tight cluster of trichocyst matrix proteins (TMPs), a few dozen of which were localized directly (Madeddu et al. 1995). We included three from this study and eleven annotated TMPs found via direct search on the ParameciumDB (Arnaiz, Meyer, and Sperling 2020). All TMPs in this compartment were characterized by high abundance in the 300g fraction only, but there were considerably few missing values contrasting it with the core basal body compartment with a similar profile.

Many ciliary proteins identified by mass spectrometry (Yano et al. 2013) had a distribution profile that appeared as a mixture of our ER compartment and the cortical compartments which motivated the creation of two 'membrane trafficking' compartments. First, the ER in this study was formed around two of the only proteins to be directly localized to the ER in *P. tetraurelia*: PDI1_1 and ptSERCA1 (Ladenburger and Plattner 2011; Hauser, Pavlovic, et al. 2000). PDI is a ubiquitous ER chaperone, while ptSERCA displayed dual localization to the ER and alveoli; interpreted as overlapping protein machinery in two calcium-storage compartments. Despite this, a large number of annotated ER chaperones clustered with these two and supported this was the ER in particular. Two chaperones are orthologs of the *S. cerevisiae* ER markers: DNAJC25 and ALG11. Six more chaperones had either the GO term GO:0005783 or PFAM domains PF00012 (HSP70) or PF00226 (DnaJ), while six orthologs of the *S. cerevisiae* ER membrane proteins were included to make this compartment more so the ER proper. However, we did see the syntaxin PtSYX1-1, localized to exocytic vesicles (Kissmehl et al. 2007), with this same profile making the ER compartment partially involved in membrane trafficking outside of the ER proper. The distribution profiles of these proteins had key differences between experiment 1 and experiments 2 and 3 wherein the former contained a double peak in the 9K and 15K fractions with a marked drop in the 12K fraction and in the latter 9K abundance was lower than 12K. This was true of all ER chaperones we investigated. ptSyx1-1 also peaked in the 15K fraction, but its profile lacked the double-peak with the 9K fraction and instead was lower in 9K than 12K making it similar to that of the ER chaperone's experiment 3 pattern. This inclusion of ptSyx1-1 makes this compartment a combination of the ER proper with some inclusion of trafficking machinery.

The two membrane trafficking compartments shared the ER's high abundance between fractions 9K and 30K but contain high abundance elsewhere depending on the membranous nature of the protein. The soluble membrane trafficking compartment shifted abundance towards the 300g and Sup fractions with a modest drop in abundance in fractions 1K -5K. Three well-studied proteins formed the core of this compartment: the cytoplasmic dynein DHC-8 (Asai et al. 1994), the parasomal sac protein CaNA4a (Momayezi et al. 1986), and the ciliary/basal body protein IFT57a (Shi et al. 2018). IFT57a is also thought to localize to VEG MACs, but its clustering with cortical proteins here suggests that its role in ciliary signaling predominates that function. This combination suggested to us a compartment responsible for shuttling proteins to the cell cortex, and the combination of high abundance in the 300g fraction (heavy) and Sup (soluble) supported this. We also included the endosomal dynamin DRPD (Wiejak, Surmacz, and Wyroba 2003). This combination of endosomal and parasomal sac proteins again supports a generalized trafficking compartment. Nine other ciliary IFT proteins (Yano et al. 2013) were included as were 13 key trafficking proteins (Richardson and Dacks 2022) which made up core components of the COPI/II and AP Complex. Taken together, this combination of trafficking proteins suggests this compartment is responsible for endocytosis and shuttling proteins to the cell cortex through ER/Golgi-mediated processes.

The insoluble membrane trafficking compartment originated from two contractile vacuole proteins: NSF2 and PtSo1c (Kissmehl et al. 2002; Reuter, Stuermer, and Plattner 2013). The former has promiscuous localization across the ER, lysosome, and small vesicles, while the latter also localizes to small microdomains beneath the plasma membrane. Both had high abundance in 300g and 9K-30K fractions but lacked any abundance in the Sup and had a small peak of abundance in experiment 3's 120K fraction. Clustering with these two is the SNARE PtSec22 which localizes directly to the Golgi (Kissmehl et al. 2007). The overlap between the ER, Golgi, and endomembrane systems here suggested a generalized membrane trafficking compartment that is distinct from the first due to its lack of Sup abundance. Eight ciliary membrane proteins (Yano et al. 2013) shared this pattern and as did three PMCAs not identified. Four Rab GTPases

with PF00071 (Ras) were combined with a handful of components of the COPII and TRAPPI complexes (Richardson and Dacks 2022) and Q-SNARES (Kaur et al. 2022).

Surface antigens (sAGs) are large, heavily glycosylated proteins at the cell surface responsible for a variety of signaling pathways between the cell and environment (Preer Jr 1986). Despite their localization to the cortex, we observed here another ER-like pattern in most annotated sAGs and “mini” mAGs had higher peaks of abundance in 5K and 12K/15K in Experiment 3 but variable behavior in Experiment 1 and 2 which sometimes matches this pattern and sometimes appears more like the ER compartment. Three of these markers were described previously (Breuer et al. 1996) and the remainder had the PFAM domain PF01508 (‘Paramecium Surface Antigen’).

The final of the so-called “membrane trafficking” compartments was the well-resolved organelle: the lysosome. In *P. tetraurelia*, no protein has been clearly localized to the lysosome, however many were localized to ‘phagolysosomes’, but there was no consensus pattern from these. Instead, we saw the same abundance profile for 14 *S. cerevisiae* orthologs of lysosomal transporters, peptidases, phosphatases, RNases and glycosidases. All lysosomal proteins had high abundance in the 3K/5K fraction, but interestingly in Experiment 3 there was an additional ‘12K’ peak. Some matrix proteins had an additional bump of abundance in the Sup fractions. This combination of matrix and membrane proteins supports a clear lysosome-proper in this experiment.

Two types of metabolic compartments were described in this study. The first was the peroxisome, constituted of proteins with high abundance in 1K-5K with peaks often in either 1K or 3K. We used 15 *S. cerevisiae* orthologs of various peroxisomal membrane/importer proteins as well as enzymes like thiolase, isocitrate lyase, and acyl-CoA Reductase. Only a PEX11 homolog was included without bona fide orthology. This again supports a combined membranous and matrix component to the peroxisome compartment.

The second metabolic organelle was the mitochondria—which to our knowledge—no protein has ever been localized experimentally in *P. tetraurelia*. We used the mitochondrial ORFs described above as marker proteins as well as a four *S. cerevisiae* orthologs of TCA Cycle

enzymes. These ORFs covered the mitochondrial matrix, inner membrane space (IMS) and inner membrane (MIM). All these had high abundance in the 300g/1K fractions with some having moderate bumps of abundance in either the 5K or Sup fractions; the latter being typical of matrix/IMS proteins. A tight cluster of porins putatively of the mitochondrial outer membrane (MOM) described previously (Wideman et al. 2013) formed the based of the MOM compartment with seven *S. cerevisiae* orthologs of the MOM proteins: HFD1, FAAH, TOM22, and RHOT1. The abundance distribution of this compartment was similar to the mitochondria proper with high 300g/1K abundance but also had a peak in the 30K fraction possibly corresponding to its association with the ER.

A less well-resolved organelle in this study was the macronucleus (MAC) and micronucleus (MIC) collectively combined in the nuclear compartments based on their pelleting behavior. The soluble nuclear compartment was based around the experimentally localized RNA Pol II subunits RPB1 and RPB2 (Owsian et al. 2022; Drews et al. 2022) another twenty *S. cerevisiae* orthologs of RNA pol subunits, spliceosome subunits, transcription factors, DNA replication/repair factors, and nucleolar proteins. The abundance distribution of these proteins was characterized by high relative abundance in the MAC fraction which was often higher than the 300g and Sup fractions but was highly heterogeneous between the 12K-30K fractions. The insoluble nuclear compartment was very similar to this but more often had higher abundance in the 300g fractions than the MAC fraction and rarely had any abundance in the Sup fraction. Two experimentally studied proteins formed the base of this compartment: the histone H3P3 and actin-like protein ALP1-1, although it was their closely related paralogs which were studied directly (Lhuillier-Akakpo et al. 2016; Sehring, Reiner, et al. 2007). Accompanying these are ten *S. cerevisiae* orthologs of genes involved in chromatin structure, chromosome segregation, and DNA replication as well as an unannotated PADR1 gene with the GO term GO:0005634 (nucleus).

Finally, we created three types of cytoplasmic compartments based on high abundance in either the 79K and 120K, 120K and Sup, or just the Sup fractions. In the first category are the ribosomes made up of 14 *S. cerevisiae* orthologs of both the 40S and 60S ribosomal subunits

which formed a tight cluster due to low abundance in MAC-5K followed by a steady increase peaking in the 120K fraction. In experiment 3, both the 79K and 120K abundance values with approximately the same. The proteasome, by contrast, had high abundance in the 30K, 120K, and Sup fractions peaking in the 120K for most proteins. The markers included 15 *S. cerevisiae* orthologs of proteins of both the 20S and 26S subunits as well as one homolog with the GO term GO:0005839 (proteasome core complex). Finally, the cytosol proper was constituted of 16 *S. cerevisiae* orthologs of numerous enzymes involved in processes like metabolism, tRNA processing, and glycolysis. These proteins had uniquely high abundance in the Sup fractions with a small bump in the 300g and variable low abundance between 9K and 120K.

Marker Protein Resolution

Our marker protein resolution was compared with those from the *T. gondii* hyperLOPIT experiment (Barylyuk et al. 2020). The Qsep metric (Gatto, Breckels, and Lilley 2019) to quantify marker resolution by measuring the average Euclidean distance between marker proteins of different compartments and normalizes by the within-compartment average Euclidean distance such that large values are associated with marker proteins whose spatial distance is further from the other marker proteins in the dataset. The median Qsep score in our experiment was ~3.1, while that of *T. gondii* was ~3.4. On the low end, our insoluble membrane trafficking compartment had a mean Qsep score of ~1.96 while our highest was the trichocyst matrix compartment of ~7.34. Comparably, the dense granules (analogous to granules observed during trichocyst maturation) had a mean Qsep score of ~2.17 on the low end and the 20S proteasome was ~7.89 on the high end. These observations support that our marker protein resolution is comparable to a similar study of this kind.

Supplemental Methods:

Resource availability

Lead contact

Further information and requests for resources, data, and code should be directed to and will be fulfilled by the lead contact, Timothy J. Licknack: licknacktim (at) gmail (dot) com.

Materials availability

This study did not generate new unique reagents.

Experimental model and subject details

Cell lines

Cultures of *Paramecium tetraurelia* strain 51 were a generous gift from Sascha Krenek (German Federal Institute of Hydrology).

Cell husbandry

Cells were cultured using standard husbandry techniques described in Sonneborn (1975). Briefly, three flasks containing ~1000 cells each were subjected to multiple days of starvation before being inoculated with fresh wheat-grass medium (Cerophyl, yeast extract, stigmasterol) bacterialized with stationary phase *Klebsiella pneumoniae* in order to induce autogamy. DAPI staining was used to assess the macronuclear (MAC) state of each population, such that autogamous cells had fragmented MACs, while vegetative (VEG) cells had intact macronuclei. When a flask reach ~100% VEG cells, their culture volume was doubled in this fashion until 4L of culture was obtained at a cell concentration of ~1000 cells/ml. Culture volume never exceeded one half of the vessel volume in order to ensure adequate aeration.

Method details

Cell lysis and fractionation- VEG *P. tetraurelia* cells were harvested using a 10µm diameter, nylon mesh sieve after removing bacterial biofilm and debris using cheesecloth. Cells were washed on the nylon mesh with Dryl's Solution before being decanted into multiple 50ml tubes. Cells were gently spun (1000g x 10min) three times to replace Dryl's Solution (2mM Na Citrate, 1mM NaH2PO4, 1mM Na2HPO4, 1mM CaCl2) with either detergent-present (DP: 0.25 M sucrose, 10 mM HEPES pH 7.4, 2 mM EDTA, 2 mM magnesium acetate, Halt™ Protease and Phosphatase Inhibitor Cocktail) or detergent-free (DF: 1% Triton-X, 10mM Tris, 0.25M Sucrose, 3mM $CaCl_2$, 8mM $MgCl_2$, Halt™ Protease and Phosphatase Inhibitor Cocktail) lysis Buffer. Cells in DP Buffer were lysed using a Dounce homogenizer to 100% efficiency (20-30 strokes) before

being washed thrice with DP Buffer and thrice with DF Buffer (300g X 5min each wash). This pellet was then stored at -80C. Cells in DF Buffer were lysed using a nitrogen cavitation bomb pressurized to 250psi and incubated for 10min before slowly releasing the lysate, mixing the foam and liquid with a pipet, and incubating at 250psi for 5min before slowly releasing (Simpson 2010). This lysate was then differentially centrifuged in accordance with Geladaki et al. (2019): 300g x 5min, 1000g x 10min, 3000g x 10min, 5000g x 10min, 9000g x 15min, 12000g x 15min, 15000g x 15min, 30000g x 20min, 79000g x 43min, 120000g x 45min. Each dry pellet was stored at -80C as was the remaining supernatant.

Sample preparation and LC-MS Analysis- Twelve fractions were resultant from three separate experiments, each named in accordance with its centrifugation speed and experiment number (i.e., 300g-1, 1K-1, 3K-1, … Sup-3), while the DP lysis pellet was called the MAC fraction (i.e., MAC-1, MAC-2, MAC-3) due to its enrichment of intact MACs. All pellets were resuspended in 100µl Resuspension Buffer before being vortexed, boiled at 95C for 10min, and then centrifuged at 15000g x 10min. If a large pellet remained, then an additional 100µl of Resuspension Buffer was added, boiled, and centrifuged again as many as four additional times. The liquid Sup fractions were diluted in 2X Resuspension Buffer and subjected to the same procedure as the pellets.

Solubilized, reduced and heat-treated samples were quantified using EZQ Protein Quantitation Kit (https://www.thermofisher.com/order/catalog/product/R33200). Samples were then alkylated by addition of iodoacetamide (Pierce) to 40mM final concentration for 30 minutes in the dark at room temperature. 5.0ug total protein across the 12 samples were then processed using the Protifi S-trap Micro Columns and instructions provided in the S-trap Ultra High Recovery Protocol (https://protifi.com/pages/s-trap). Briefly, samples were acidified by addition of 12% phosphoric acid to a final concentration of ~1.2% phosphoric acid. Proteins were digested by addition of 2.0 µg of porcine trypsin (MS grade, Pierce) and incubated at 30ºC for 2 hours. S-trap buffer (90% methanol, 100 mM TEAB final) was also added in volumes 7X our total sample volume. Acidified sample and the S-trap buffer was filtered through columns. Columns were washed 3X with S-trap buffer. An additional 0.5 µg of trypsin and 25 µL of 50 mM TEAB was

added to the top of each column and incubated for 1 hour at 47ºC. Samples were eluted off the S-trap columns using three elution buffers: 50 mM TEAB, 0.2% formic acid in water, and 50% acetonitrile/50% water + 0.2% formic acid. Samples were dried down via speed vac and resuspended in 20-30 µL of 0.1% formic acid.

Liquid-chromatography tandem mass spectrometry- All LC-MS analyses were performed at the Biosciences Mass Spectrometry Core Facility (https://cores.research.asu.edu/mass-spec/) at Arizona State University. All data-dependent mass spectra were collected in positive mode using an Orbitrap Fusion Lumos mass spectrometer (Thermo Scientific) coupled with an UltiMate 3000 UHPLC (Thermo Scientific). One µL of peptides were fractionated using an Easy-Spray LC column (50 cm × 75 µm ID, PepMap C18, 2 µm particles, 100 Å pore size, Thermo Scientific) equipped with an upstream 300um x 5mm trap column. Electrospray potential was set to 1.6 kV and the ion transfer tube temperature to 300ºC. The mass spectra were collected using the "Universal" method optimized for peptide analysis provided by Thermo Scientific. Full MS scans (375–1500 m/z range) were acquired in profile mode with the Orbitrap set to a resolution of 120,000 (at 200 m/z), cycle time set to 3 seconds and mass range set to "Normal". The RF lens was set to 30% and the AGC set to "Standard". Maximum ion accumulation time was set to "Auto". Monoisotopic peak determination (MIPS) was set to "peptide" and included charge states 2-7. Dynamic exclusion was set to 60s with a mass tolerance of 10ppm and the intensity threshold set to 5.0e3. MS/MS spectra were acquired in a centroid mode using quadrupole isolation window set to 1.6 (m/z). Collision-induced fragmentation (CID) energy was set to 35% with an activation time of 10 milliseconds. Peptides were eluted during a 240-minute gradient at a flow rate of 0.250 uL/min containing 2-80% acetonitrile/water as follows: 0-3 minutes at 2%, 3-75 minutes 2-15%, 75-180 minutes at 15-30%, 180-220 minutes at 30-35%, 220-225 minutes at 35-80% 225-230 at 80% and 230-240 at 80-5%.

Label-free quantification (LFQ)- Four protein databases were used for the search: P. tetraurelia strain 51's predicted proteome (downloaded from https://paramecium.i2bc.paris-saclay.fr/), 31 mitochondrial ORFs described in the Supp Text, K. aerogenes's predicted

proteome (https://www.uniprot.org/), and common laboratory contaminants (cRAP; common Repository of Adventitious Proteins: https://www.thegpm.org/crap/). Protein identification was performed using all combined fractions, while normalized abundance values were calculated for each protein within each replicate fraction.

LFQ was performed using Proteome Discover 2.4 (Thermo Scientific). Raw files were searched using SequestHT that included Trypsin as enzyme, maximum missed cleavage site 3, min/max peptide length 6/144, precursor ion (MS1) mass tolerance set to 20 ppm, fragment mass tolerance set to 0.5 Da and a minimum of 1 peptide identified. Carbamidomethyl (C) was specified as fixed modification, and dynamic modifications set to Acetyl and Met-loss at the N-terminus, and oxidation of Met. A concatenated target/decoy strategy and a false-discovery rate (FDR) set to 1.0% was calculated using Percolator. The data was imported into Proteome Discoverer 2.4, and accurate mass and retention time of detected ions (features) using Minora Feature Detector algorithm. The identified Minora features were then used to determine area-under-the-curve (AUC) of the selected ion chromatograms of the aligned features across all runs and relative abundances calculated.

Data analysis was performed using the R Bioconductor packages MSnbase (v 2.20.1) and pRoloc (v 1.34.0) as described in Breckels et al. (2016). Briefly, a protein-level csv file containing 12,579 proteins was filtered such that proteins were removed under the following criteria: they were from cRAP or K. aerogenes, had low FDR confidence, had fewer than ten total PSMs, had no unique peptides, or were identified in only the MAC or Sup fractions. Technical triplicates (e.g., three replicates of 12K-1) were averaged to generate a 36th dimensional dataset of relative protein abundance. The datasets were split into their respective experiments (i.e., 1-12. 13-24, 25-36) to perform hybrid imputation described in Supp Text and sum-normalization across rows. The 36 fractions were concatenated together and used for downstream analyses.

Supervised and Unsupervised Classification

291 manually curated marker proteins (whose curation is described in Supp Text) were used to classify all 9,026 proteins to one of seventeen cellular compartments. Supervised classification was done using a support vector machine (SVM) model using the svmOptimization and

svmClassification functions in pRoloc. Briefly, 100 rounds of five-fold cross-validation were performed to optimize the SVM parameters, sigma and cost, using the marker protein profiles. Each fold is stratified 80/20 for training/testing, respectively, wherein parameter values determined in the training set are 'tested' with the test set. Macro F1 scores were used to assess the classifier accuracy, and this is the harmonic mean of precision (True Positive / True Positive + False Positive) and recall (True Positive / True Positive + False Negative). The optimal parameters for the SVM classifier were then applied to all proteins in the dataset with a corresponding SVM score whose range is 0-1 with 1 being the score of marker proteins. The SVM classifier was then applied to unlabeled data (i.e., non-marker proteins) with corresponding weights applied to each marker class on the basis of its size. Each protein was thus classified to one compartment, and any protein whose classification fell below the global median SVM score was reset to 'unknown' while the other half of the dataset was considered "predicted" to its corresponding compartment due to their higher SVM scores.

Unsupervised clustering was performed using the K-means (KM) algorithm implemented in the MLearn function from the MLInterfaces package in R (). Briefly, KM generated k random centroids and includes surrounding datapoints iteratively such that all data points are included in one of the k clusters and the size of each centroid is minimized. We generated 17 KM clusters and compared them to the 17 SVM-predicted compartments.

Characterization of Organellar Compartments - Properties for predicted proteins were obtained through several sources. First, all protein IDs were submitted to the Sherlock tool (formerly BioMart) from https://paramecium.i2bc.paris-saclay.fr/ for the following characteristics: protein size, isoelectric point, INTERPRO/PFAM domains, GO terms, mRNA expression level VEG growth (Arnaiz et al. 2017), differential expression after trichocyst discharge and ciliary shedding (Arnaiz et al. 2017), transmembrane domain presence (via TMHMM; Möller et al. 2001), and signal peptide presence (via SignalP 3.0; Bendtsen et al. 2004). Target peptides were predicted using the TargetP tool (Armenteros et al. 2019). Orthology relationships were determined in one of two ways: 1) using GhostKOALA (Kanehisa et al, 2016) to identify KEGG Ortholog (KO)-based gene names, 2) using EukProt v3 (Richter et al. 2022) to determine

evolutionary relationships across the tree of eukaryotes. Instances of specific orthologs for named species are validated using reciprocal best BLASTp hits. KEGG pathways were visualized using the KEGG Mapper- Reconstruct tool (https://www.kegg.jp/kegg/). Motif discovery for promoter regions was done using the MEME algorithm implemented in the MEME software suite (Bailey et al. 2009). Five motifs of size 6-12nts were searched within a single fasta file containing 200nt upstream of all protein-coding genes predicted to the same compartment.

Data and code availability

Raw mass spectrometry data will be deposited to the ProteomeXchange Consortium (http://www.proteomexchange.org/) and intermediate files can be assessed by request to the corresponding author. All code is available on GitHub (https://github.com/Tlicknack/Paramecium_Spatial-Proteomics).

Figure 2.6. Mitochondrial Genome Annotation
The mitochondrial genome of Paramecium tetraurelia was reanalyzed using mass spectra from an enriched mitochondrial fraction (three technical replicates of three biological replicates of 1K). We generated 283 mitochondrial ORFs and found protein evidence 42 of these, but 31 were identified confidently using both ProteomeDiscoverer (Thermo) and MaxQuant (Tyanova et al. 2016). Mito-ORFs (yellow) overlapped with many annotated proteins (blue) except in regions of predicted rRNA genes.

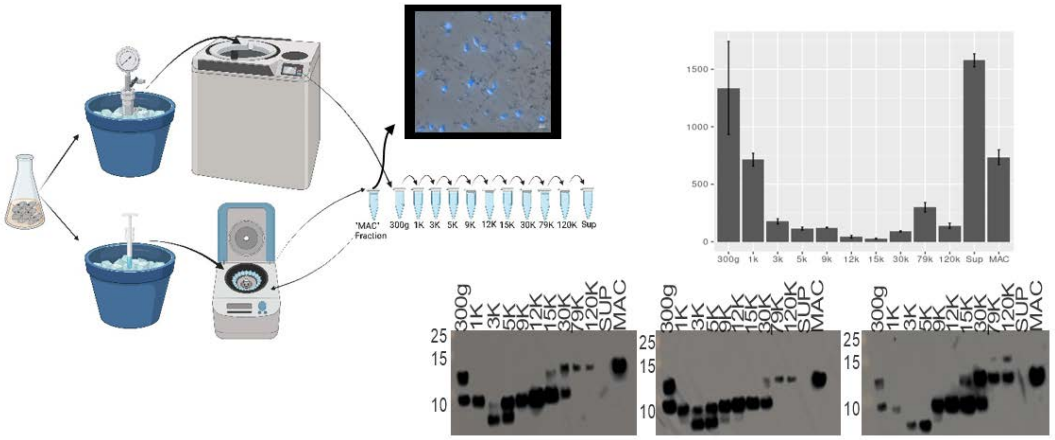

Figure 2.7. Cell Lysis and Fractionation
Our experimental design was undertaken as a modification of the fractionation scheme from Geladaki et al. (2019) described in the SI Methods (Top left). The MAC fraction was assayed with the nuclear stain DAPI to confirm the existence of enriched macronuclei. Protein yields were plotted with their standard error (top right) which are similar to that observed in Geladaki et al. (2019) with most proteins in the Sup fraction; in their study, the 300g (or 200g) fraction was discarded. Protein fractions were assayed with an anti-Histone antibody which reacted strongly with many fractions—importantly not the Sup fraction. The highest band is private to the late spins, MAC fraction, and 300g fraction, while the middle band is present in most fractions between 300g and 30K, and the lowest band is private to 3K/5K. This fractionation pattern suggested a clear biochemical difference between fractions, especially with respect to the nuclear proteomes.

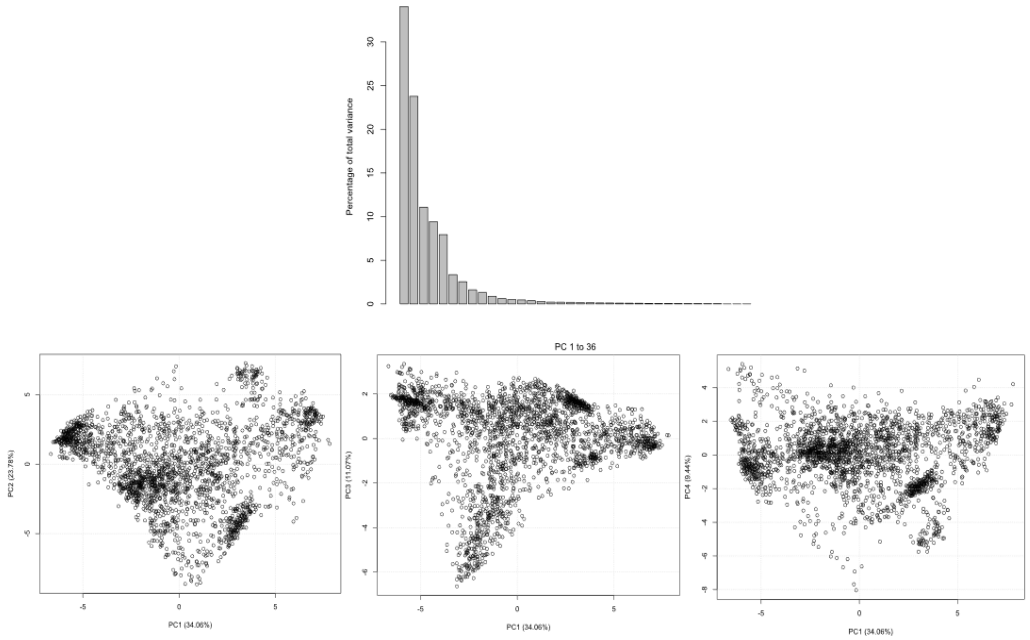

Figure 2.8. Principal Component Analysis (PCA) Reveals Data Structure
PCA was done to project our non-imputed multidimensional dataset across a smaller number of principal components (PCs) which each explain some percentage of the variance present in the dataset. PC1 and PC2 explain more than half of the variance while the next three PCs explain ~10% each (top). Clusters are apparent in plots of PC1 vs PC2, PC1 vs PC3, and PC1 vs PC4. This analysis supports that the data are non-randomly structured and can be visualized in a few PCs despite containing 36 dimensions.

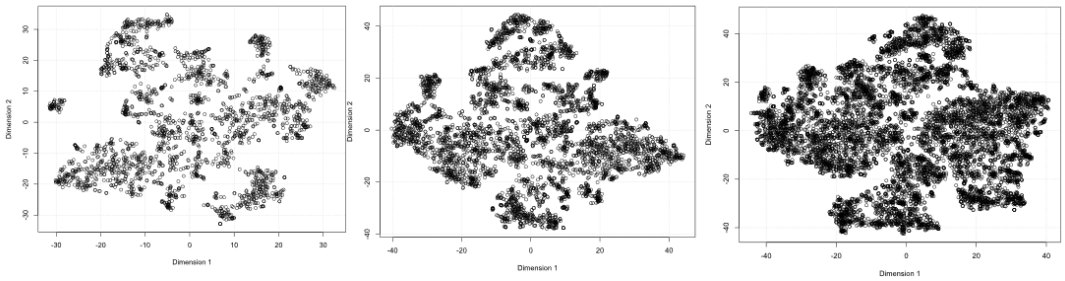

Figure 2.9. t-distributed stochastic neighbor embedding (t-SNE) and Data Imputation
We used the t-SNE algorithm to better visualize our data on two dimensions than PCA could achieve. Here, we show the effects of imputation on our data structure, first imputing neighboring fractions (left to middle) then imputing zeros to the remaining fractions (middle to right). In general, we see that as more proteins are imputed and therefore included in each plot, the data structure “collapses” towards the center of the plot. Since the data structure is determined by the relationship between unique values across each dimension, this loss of structure via the imputation of the same value is greatest when zeros are included (middle to right). However, the maintenance of large clusters in the hybrid imputed dataset supported its use in downstream organelle-prediction analyses. The x- and y-axes are t-SNE dimensions and contain arbitrary coordinates purely for the purpose of data visualization.

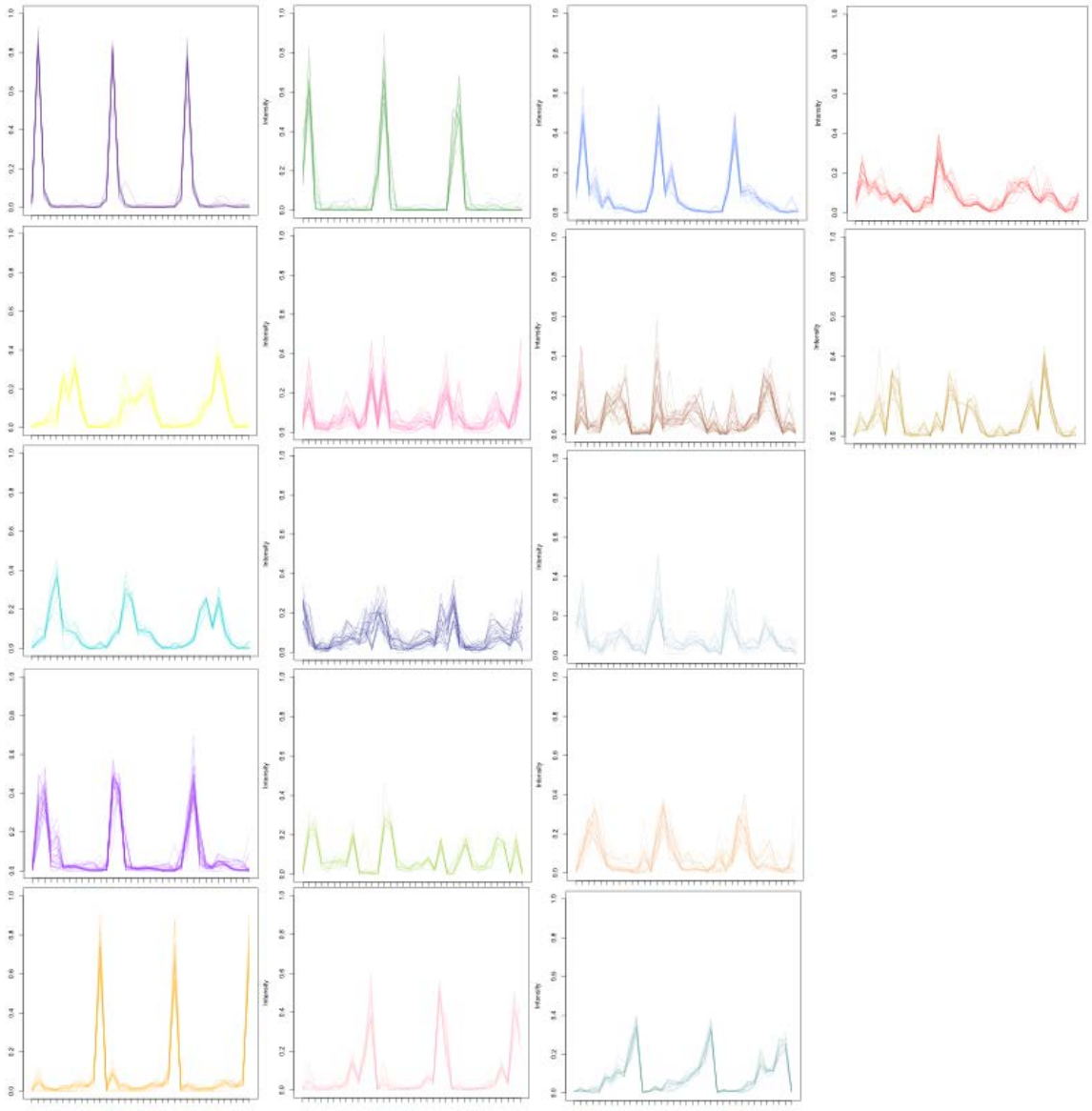

Figure 2.10. Marker Protein Abundance Distribution Profiles
After manually curating a list a of 291 marker proteins, we plotted their abundance distribution profile across all 36 fractions (three experiments each containing 12 fractions, concatenated). The proteins are arranged roughly according to their spatial distribution in the cell as seen in Figure 2.1. The top-most row are the four cortical compartments, the next row is the membrane trafficking (ER and ER-like) compartments, the next row are the lysosome and nuclear compartments, the second-to-last row are the mitochondrial and peroxisomal compartments, and the bottom row are cytoplasmic compartments. The color code is shown in Figure 2.2., and all x- and y-axes are consistent for those in Figure 1.1.

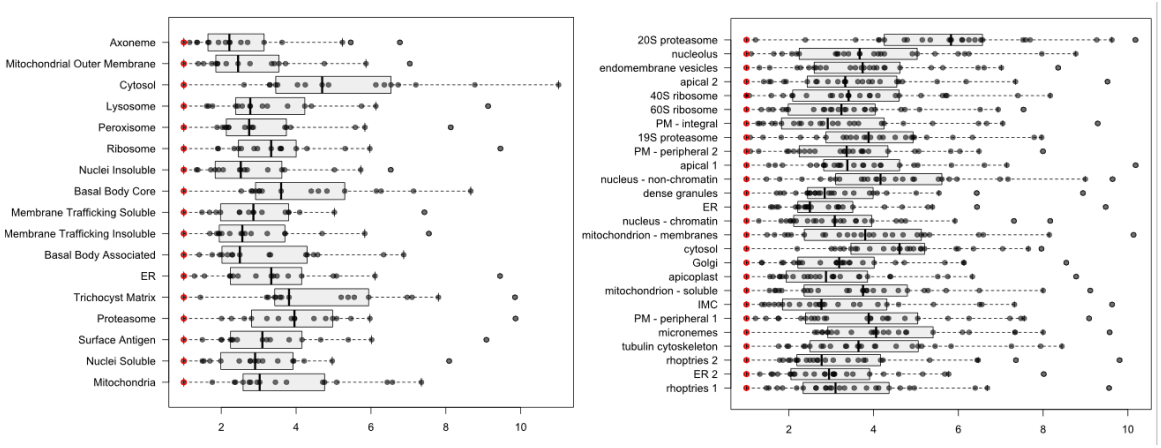

Figure 2.11. Marker Protein Resolution
Marker protein resolution was assessed using the Qsep metric introduced by Gatto et al. (2019). Large Qsep values translate to better resolved marker classes. We computed this for all marker proteins in our dataset (left) before computing it for the marker proteins used in a hyperLOPT study of *Toxoplasma gondii* from Barylyuk et al. (2020). We see a similar range of values in both experiments with median values all greater than two meaning that each marker class is twice as distant from another marker class as it is from a given protein within the same marker class.

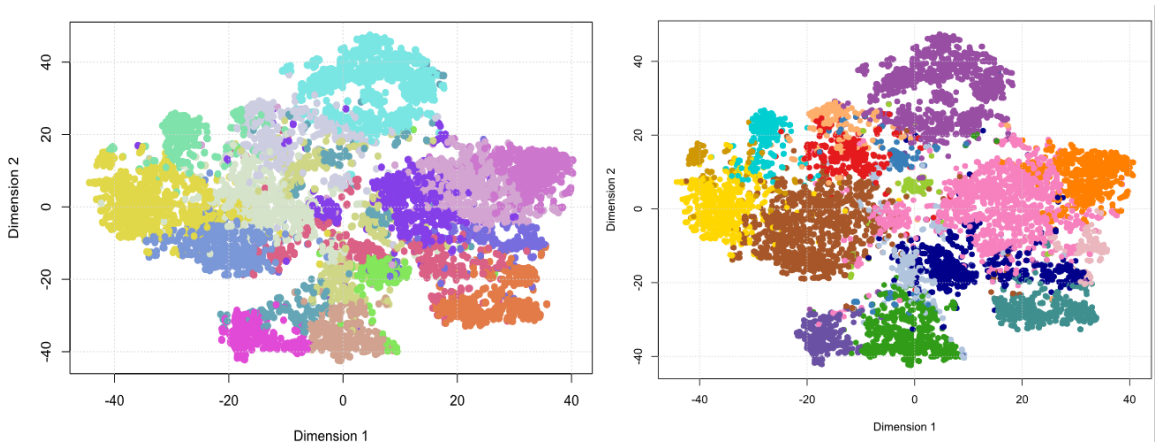

Figure 2.12. K-Means (KM) Clusters versus Organellar Classifications
We generated 17 de novo clusters using the KM algorithm to compare with our biologically relevant SVM classifications. Here, we projected KM clusters and organellar classifications for all 9,026 proteins onto a t-SNE plot to visually inspect the overlap quantitatively assessed in Figure 2.2 with the same color scheme corresponding to the rightmost plot. The KM cluster names for each protein can be found in Supp Table. In general, we see a high degree of spatial overlap between both methods. In some cases, additional structure within each organellar compartment can be seen when KM clusters split them as is the case for the soluble (right, pink) and insoluble (left: brown) membrane trafficking compartments known to consistent of many discrete components.

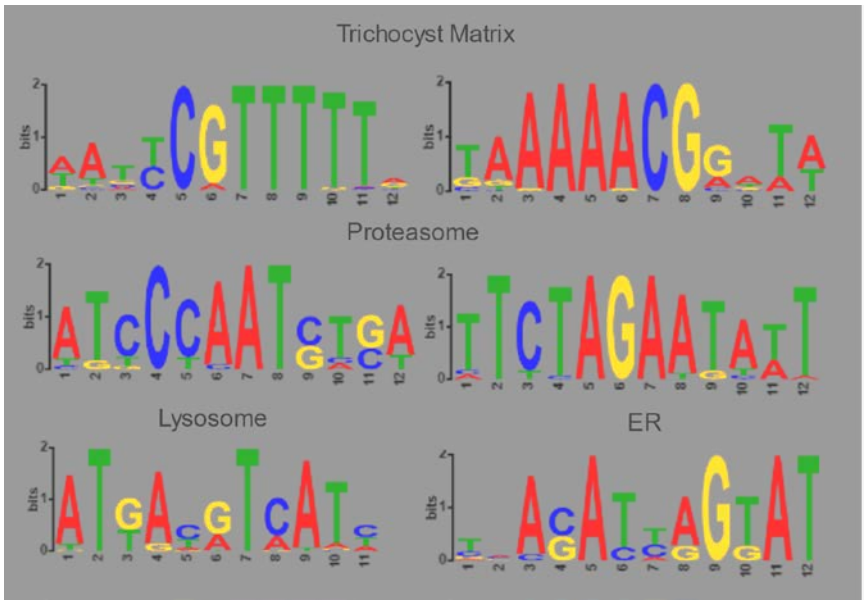

Figure 2.13. Promoter Motifs with Conserved Positions
We performed de novo motif discovery within 200nt upstream of each predicted protein-coding gene's annotated start codon to assess the presence of putative regulatory elements enriched within each organellar class. In addition to the six shown here with relatively conserved motif positions, 13 more were found to be significantly enriched using a hypergeometric test implemented with MEME (Bailey and Elkan 1994). The trichocyst matrix motifs are palindromic and found upstream of ~53% (left) and ~19% (right) of all its predicted genes but in only four of 154 instances are they found in the same promoter. Similarly, the two proteasomal motifs were upstream of ~27% (left) and ~40% (right) of predicted proteasomal genes and were found together only once. The other two motifs were found upstream of ~36% and ~8% of all predicted lysosomal and ER predicted genes, respectively. In all cases, motif variants were found within ~50nts of the start codon.

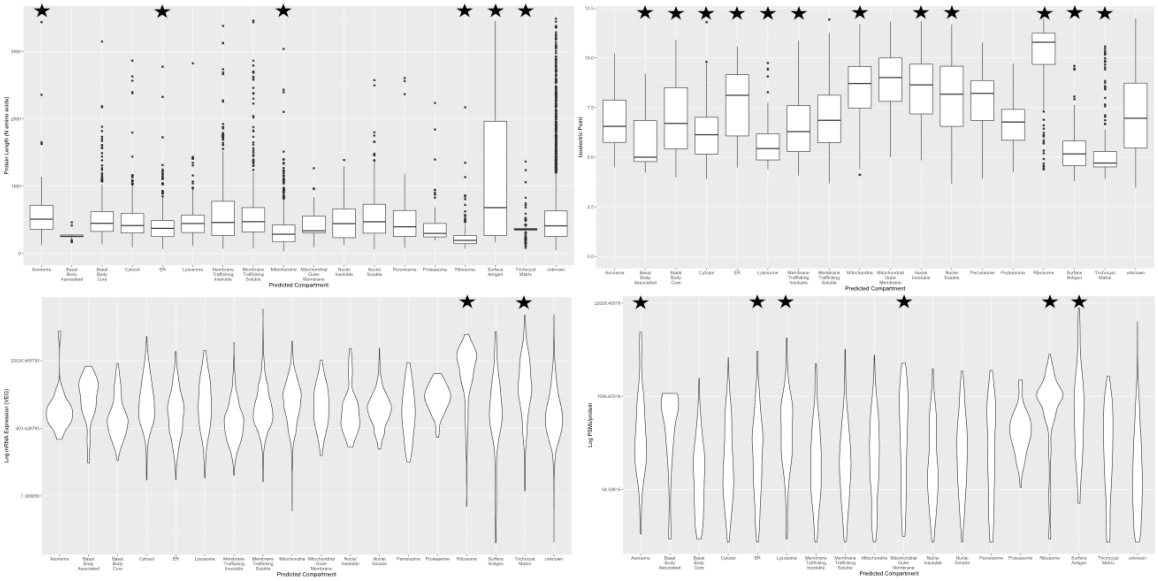

Figure 2.14. Protein and Genic Properties of Predicted Compartments
We computed the protein size (top left), isoelectric point (top right), mRNA expression level (bottom left), and number of peptide spectral matches (bottom right) for each protein/gene of each predicted class and compared them using ANOVA, denoting those significantly different from the unknown with a star after correcting for multiple tests ($\rho < 0.0029$). Protein size, isoelectric point, and mRNA expression values were obtained from the ParameciumDB (Arnaiz et al. 2019), while the peptide spectral matches were computed in this study.

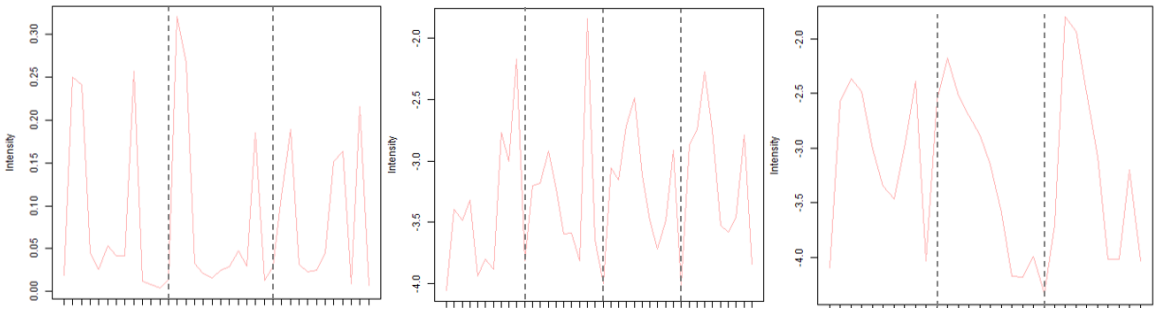

Figure 2.15. TOM40 Distribution Profiles in Paramecium tetraurelia, Saccharomyces cerevisiae, and Toxoplasma gondii

We compared the protein abundance profiles for TOM40 in P. tetraurelia (left: PTET.51.1.P0280026), S. cerevisiae (middle: P23644), and T. gondii (right: TGME49_218280) using their respective hyperLOPIT datasets (Barylyuk et al. 2019; Nightingale et al. 2019). The x-axes correspond to each experiment's fractions. Our dataset contains sum-normalized data plotted across three experiments with 12 fractions per experiment in accordance with the fraction names in Figure 2.1 and Figure S2. In *S. cerevisiae*, there are four experiments with ten fractions per experiment, sum normalized and log transformed. These fractions were generated from a density gradient centrifugation fractionation (compared with differential centrifugation in our study) and thus lighter cellular material will be present in 'early' fractions (i.e., 1-5) while heavier material will be in 'later fractions' (i.e., 6-10). The same is true of *T. gondii* as is *S. cerevisiae*, except in the former there being three experiments. Dotted lines denote the end of each experiment to better visualize their shared pattern. In all three organisms from each study, there is a shared 'heavy' and 'light' peak presumably associated with the mitochondria proper and the MOM specifically, respectively. The relative heights of these peaks do vary between experiments, and in our case, the existence of a third peak, associated with membrane trafficking, is seen in the 120K fraction (Figure S5).

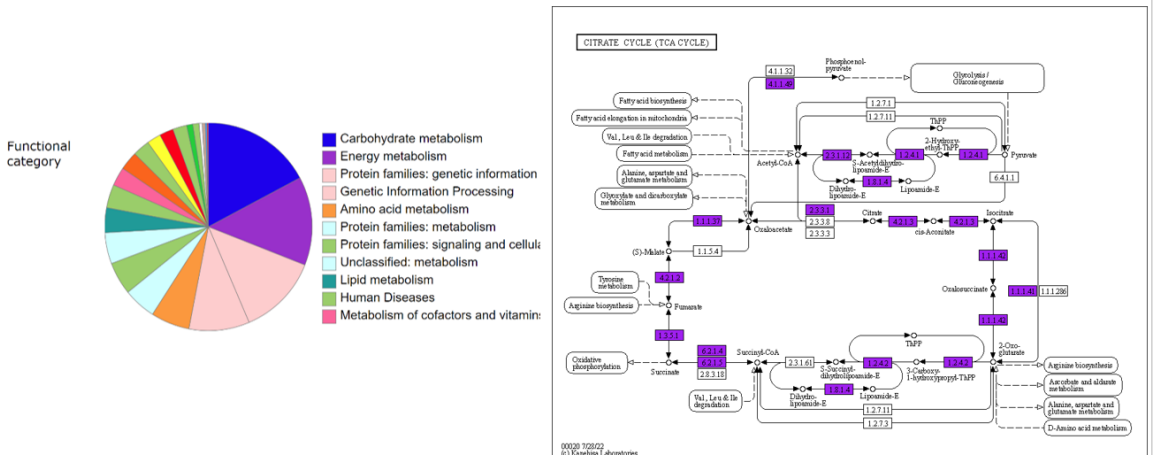

Figure 2.16. Functional Categories and the TCA Cycle for Mitochondrial Predictions
Of the 1,056 predicted mitochondrial proteins, 484 were given KO terms via the ghostKOALA software (Kanehisa et al. 2016) and mapped in the KEGG database. Most of these proteins have putative roles in metabolism or genetic information processing (left) with the latter relating to the presence of mitochondrial ribosomes. The metabolic TCA cycle (right) contains the vast majority of its expected enzymatic components in these mitochondrial predictions.

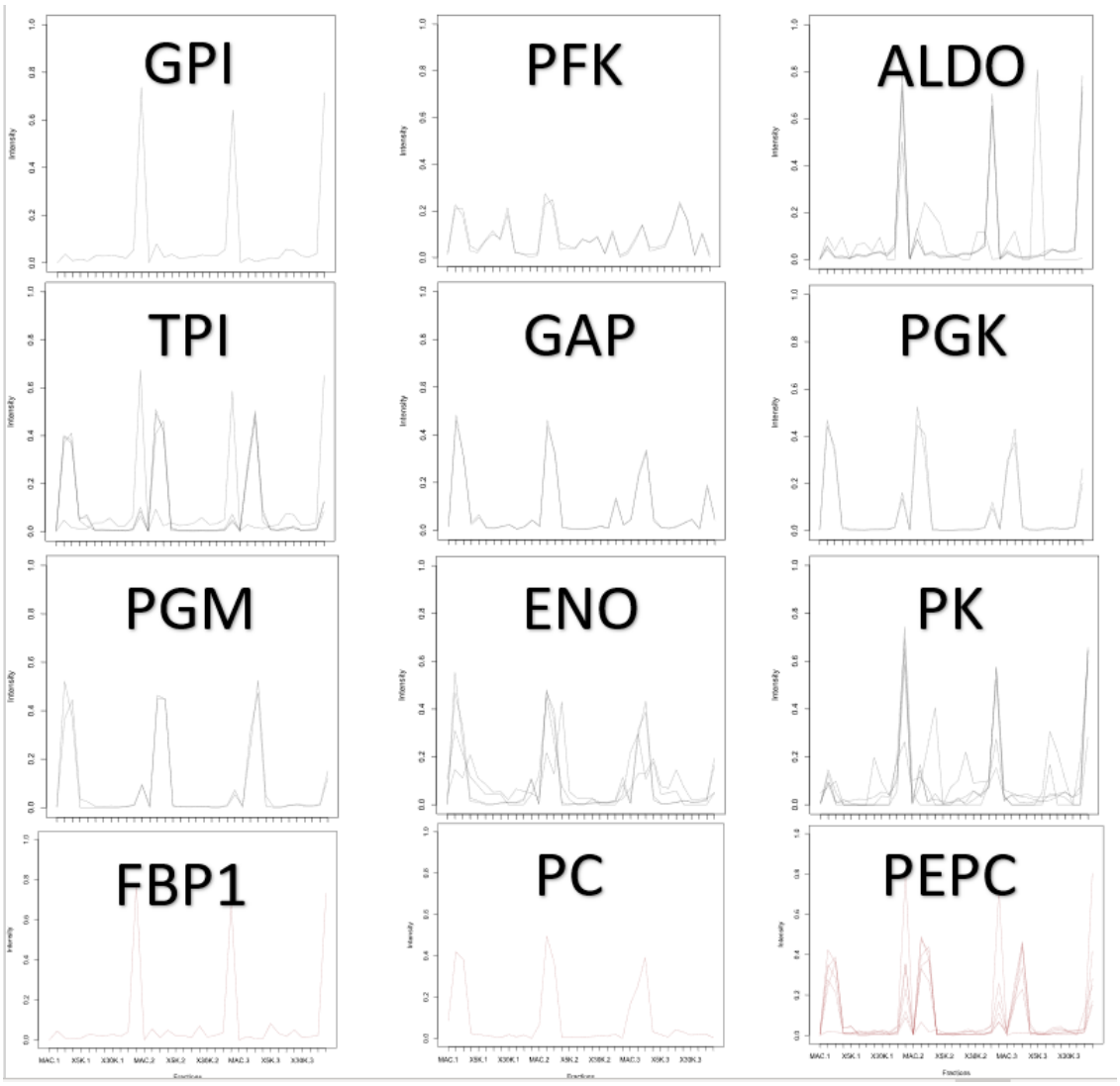

Figure 2.17. The Mosaic Glycolytic and Gluconeogenic Pathway
We plotted the distribution profile for each glycolytic enzyme described in Figure 3. Hexokinase is thought to be lost in ciliates, but we did identify the other nine enzymes. Six were localized to either the cytosol (GPI: PTET.51.1.P0300257; ALDO: PTET.51.1.P0770173, PTET.51.1.P0940159, PTET.51.1.P0810050, PTET.51.1.P0900034), MOM (PFK: PTET.51.1.P0480063, PTET.51.1.P0670026) or mitochondria (GAP: PTET.51.1.P0380195, PTET.51.1.P0500184; PGK: PTET.51.1.P0700046, PTET.51.1.P1180061; PGM: PTET.51.1.P0890070, PTET.51.1.P1190051). The other three enzymes contained multiple copies whose paralogs were predicted to different compartments: TPI (PTET.51.1.P1550028, PTET.51.1.P1070138, PTET.51.1.P1550027, PTET.51.1.P1370031) with three mitochondrial and one cytosolic copy; ENO (PTET.51.1.P0870049, PTET.51.1.P0100278, PTET.51.1.P0590214, PTET.51.1.P0040146) with two mitochondrial and two unknown copies; and PK (PTET.51.1.P0360217, PTET.51.1.P0110153, PTET.51.1.P0210069, PTET.51.1.P0070160, PTET.51.1.P0100415) with one cytosolic, one soluble membrane trafficking, and three unknown copies. The gluconeogenic enzymes catalyzing the reverse of the final reaction are shown in the final row with a dark-red color. FBP1 (PTET.51.1.P0730140) catalyzes the reverse of PFK's and is cytosolic. PC (PTET.51.1.P0530213) is orthologous to H. sapiens MCCA, however it is a strong BLASTp hit from H. sapiens PC and clearly mitochondrial. PEPC (PTET.51.1.P0180052, PTET.51.1.P1010170, PTET.51.1.P0460194, PTET.51.1.P0360202, PTET.51.1.P1250149, PTET.51.1.P0440244) has six copies, five of which are mitochondrial, and one is cytosolic.

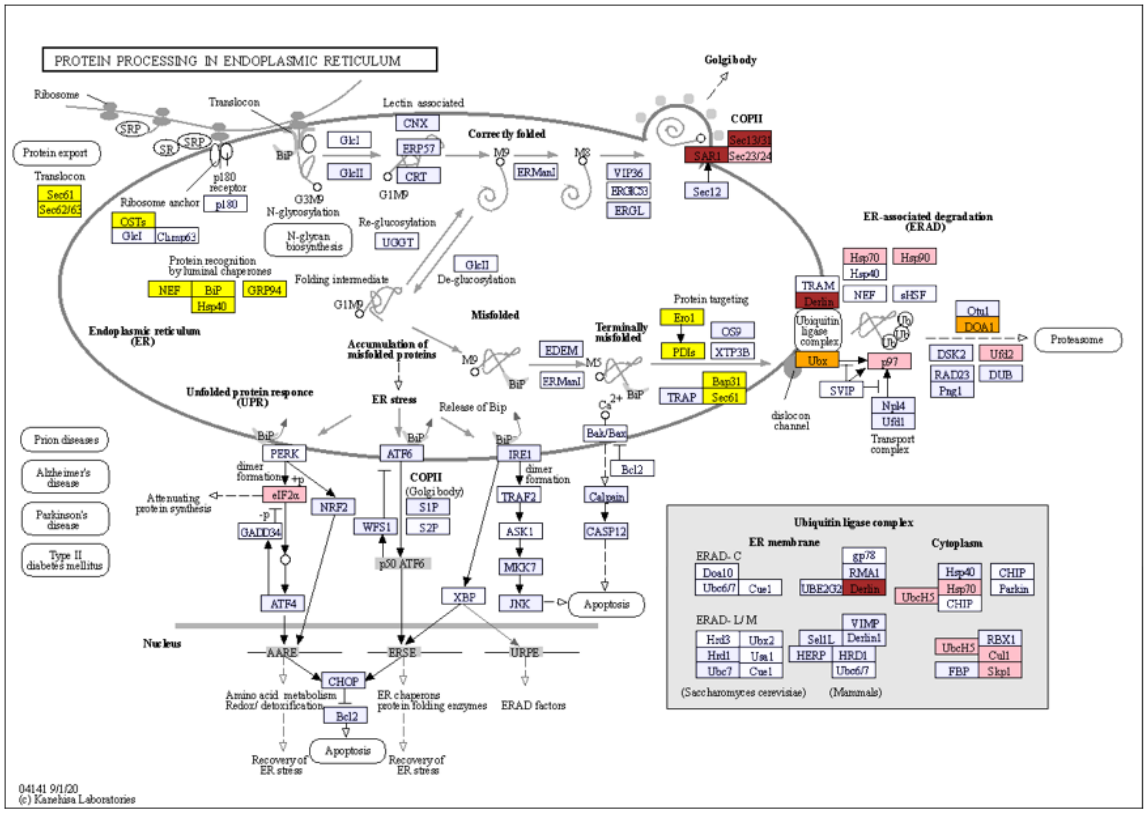

Figure 2.18. Partitioning of the Endoplasmic Reticulum and Membrane Trafficking
The processing proteins of the endoplasmic reticulum (ER) contain proteins with diverse organellar predictions: ER proper (yellow), insoluble membrane trafficking (red), soluble membrane trafficking (pink) or cytosol (orange). Only the ER predicted proteins represent the luminal chaperones, translocation machinery, and protein targeting pathways while the two membrane trafficking compartments play a role in ER-associated degradation (ERAD), transport to the Golgi, and ubiquitination; all of which involve the removal of material from the ER instead of import and retention. Two cytosol predicted proteins are also involved in ERAD.

CHAPTER 3

PCPULLDOWN: A SIMPLE TOOL TO PROBE HIGH-DIMENSIONAL PROTEOMICS EXPERIMENTS FOR A PROTEIN OF INTEREST

Abstract:

Background: Spatial proteomics allows researchers to map global patterns of protein localization simultaneously by grouping proteins based on their shared abundance profiles across diverse subcellular fractions. While these techniques employ complex mathematical algorithms which directly make use of the higher-dimensional distance between pairs of proteins, this information is largely ignored after classification is performed. Here, I introduce a computational pipeline whose sole purpose is to simply identify the most spatially similar proteins in the dataset. I show how this method recapitulates known biological interactions and predicts new ones. Indirectly, I discover that proteins predicted to the same organellar compartment have shared relationship to all other proteins in the proteome likely a product of their intraorganellar or cytosolic nature. Taken together, this study provides a simple framework for studying the complex problem of protein localization.

Methods: Using both published and unpublished spatial proteomics datasets, I measure all Protein Profile Similarity Scores (PPSSs) using a modified Euclidean distance metric. This is stored as an S3 object in R which can easily be assessed by a number of custom functions all available for download on: https://github.com/Tlicknack/Paramecium_Spatial-Proteomics.

Results: I demonstrated the efficacy of this approach using a handful of well-studied proteins from *Saccharomyces cerevisiae, Toxoplasma gondii,* and *Paramecium tetraurelia*; three very different unicellular eukaryotes. Nearly all proteins which are members of well-known large protein complexes identify most other members of their shared complex as well as proteins with close association to that complex (e.g., nucleolar proteins with RNA Pol. I). Cytosolic proteins— not a part of larger complex—have seemingly random combinations of closely related proteins, although those involved in biochemical pathways associated with some organelle are spatially closer to proteins of that organelle (e.g., glycolysis in yeast). Finally, I show how global PPSS distributions for individual proteins may represent well its entire spatial interactome. This is

demonstrated using mitochondrial and cytosolic compartments in numerous model organisms including *Homo sapiens* and *Mus musculus*.

Conclusions: My functions are immediately impactful in providing an orthologonal approach to analyzing spatial proteomics data. Additionally, I show hints of new cell biological phenomena illuminated by the metrics underlying this technique. The ability to identify candidate protein-protein interactions from already existent datasets should provide a valuable resource to cell biologists working in those model systems.

Introduction:

Knowledge of where a protein localizes and with whom it interacts is paramount to understanding cell biology. A burgeoning toolkit has developed under the umbrella of 'spatial proteomics' whose experimental designs all share a few key properties: gentle cell lysis, fractionation, and quantitative proteomic analysis (Lundberg and Borner 2019). The modularity of this approach has aided its success, allowing researching to use a variety of lysis and fractionation methods (Geladaki et al. 2019) as well as either labeled (Dunkley et al. 2004) or label-free (Foster et al. 2006) quantification of proteins— the former called Localization of Organellar Proteins by Isotopic Tagging (LOPIT) and the latter called protein correlation profiling (PCP). Resulting from all of these is a higher dimensional dataset in which each protein has a unique abundance profile which reflects its steady-state abundance within the cell.

The analysis of these data is non-trivial, and interpretation requires a broad expertise of both computational and cell biology. A few groups have provided robust workflows and tools for this (Gatto and Lilley 2012; Breckels et al. 2016; Gatto, Breckels, Wieczorek, et al. 2014; Gatto, Breckels, Burger, et al. 2014), but these focus mainly on uncovering broad patterns of protein localization without an easy way of querying proteins of interest. While other experimental methods exist for the sole purpose of understanding the cell biological environment of individual proteins, such as affinity purification (Dunham, Mullin, and Gingras 2012) and proximity biotinylation (Roux et al. 2012), both PCP and LOPIT produce spatial maps in which this

information is preserved in the form of ‘microclusters’ within larger organellar clusters (Lundberg and Borner 2019). These microclusters reflect the nonuniform distance between proteins within the same organelle due to their physical association. Despite this qualitative observation, little work has gone towards investigating whether these patterns hold bona fide cell biological information.

In this article, I introduce a simple, new R (Team and others 2013) package called PCpulldown whose main purpose is to query individual proteins and uncover their relationship to all other proteins within diverse spatial proteomics datasets. Through a modified Euclidean distance measurement, I am able to recapitulate known biological interactions for a handful of well-studied proteins as well as make predictions about new putative interactions. I also describe how proteins residing in the same organelle have highly similar patterns to their relationship to all other proteins in the dataset. Taken together, this package provides a simple way of accessing spatial proteomics datasets without the need for high levels of expertise in the R programming language.

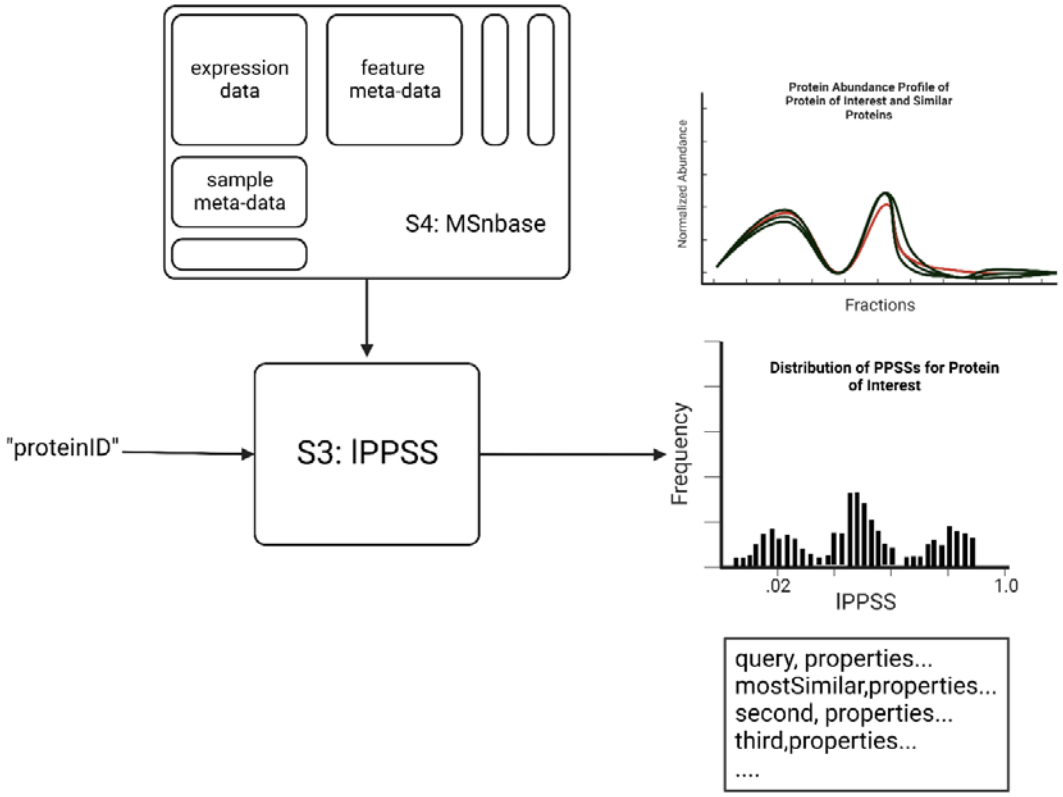

Figure 3.1. PCpulldown Overview

An MSnbase object is converted into an S3 list using the makePPSS function which quickly measures all pairwise Elucidean distances and converts them into normalized Protein Profile Similarity Scores (PPSSs). An S3 object stores these PPSSs and is easily queried using base R indexing, i.e., lPPSS[[“protein”]]. The PCpulldown function uses base R functions to produce three types of outputs: 1) the protein abundance profile for the protein of interest (red) and highly similar proteins (black), 2) the distribution of all PPSSs for that protein, and 3) a csv file with proteins and their associated properties ranked by PPSS. This set-up allows for easy and quick access to higher-dimensional proteomics datasets acting akin to an in silico “pulldown” of a protein of interest. Created with BioRender.com.

Results:

PCpulldown: A simple way to probe spatial proteomics datasets

To easily access spatial proteomics datasets, I made use of the preexisting infrastructure developed in the MSnbase and pRoloc R packages available through Bioconductor (Gatto and Lilley 2012; Gatto, Breckels, Wieczorek, et al. 2014; Crook et al. 2019). Briefly, MSnbase serves to process and store mass spectrometry data irrespective of its source and application through the generation of an S4 object in which both expression and feature data are stored. These objects can be generated through csv files produced by popular mass spectrometry library searching software like MaxQuant or ProteomeDiscoverer (Tyanova, Temu, and Cox 2016; Orsburn 2021). MSnbase is also capable of handling raw xml-based file formats directly from the Mass Spectrometer, but I assume most users will obtain protein or peptide-level csv files from some intermediate program. The pRoloc package makes use of MSnbase but for the specific application of subcellular/organellar proteomics. The utility and value of these packages are immense to any researcher in the field, and I only introduce this tool as a complement to these. An interactive application implemented in the pRolocGUI package, built on the shinydashboardPlus infrastructure, provides the best way to directly interact with these data (Gatto et al. 2015). PCpulldown provides an intermediate to the purely programmatic and purely UI-based methods of data analysis.

The PCpulldown workflow is simple: 1) make protein profile similar score database using the makePPSS function, and 2) query using the PCpulldown function with a protein of interest (Figure 3.1). The query can be loaded with a file containing diverse properties of all proteins in the dataset, e.g., by pulling information from the organism's genome browser, but by default, the feature data stored in the MSnbase object is returned. The makePPSS function takes as an input the MSnbase object and performs three tasks: 1) measures all pairwise Euclidean distances with a custom C++ function, 2) normalizes all Euclidean distances to the largest and smallest global values, and 3) subtracts this from one. The resulting S3 object contains all protein names in their

first layer, followed by a second layer with all protein names (excluding themselves) with a corresponding PPSS. The PCpulldown function takes five potential inputs: 1) the protein of interest, 2) the original MSN object, 3) the S3 PPSS database, 4) the output directory, and 5) an option to provide properties of all proteins in the dataset. This function returns three files: 1) a csv file containing proteins listed from the most similar (i.e., highest PPSS) to least similar with their corresponding properties or feature data, 2) the distribution profile of the protein of interest with the highest 95th percentile of similar proteins, and 3) a histogram of all PPSSs for that protein of interest. The workflow is accessible here: https://github.com/Tlicknack/Paramecium_Spatial-Proteomics.

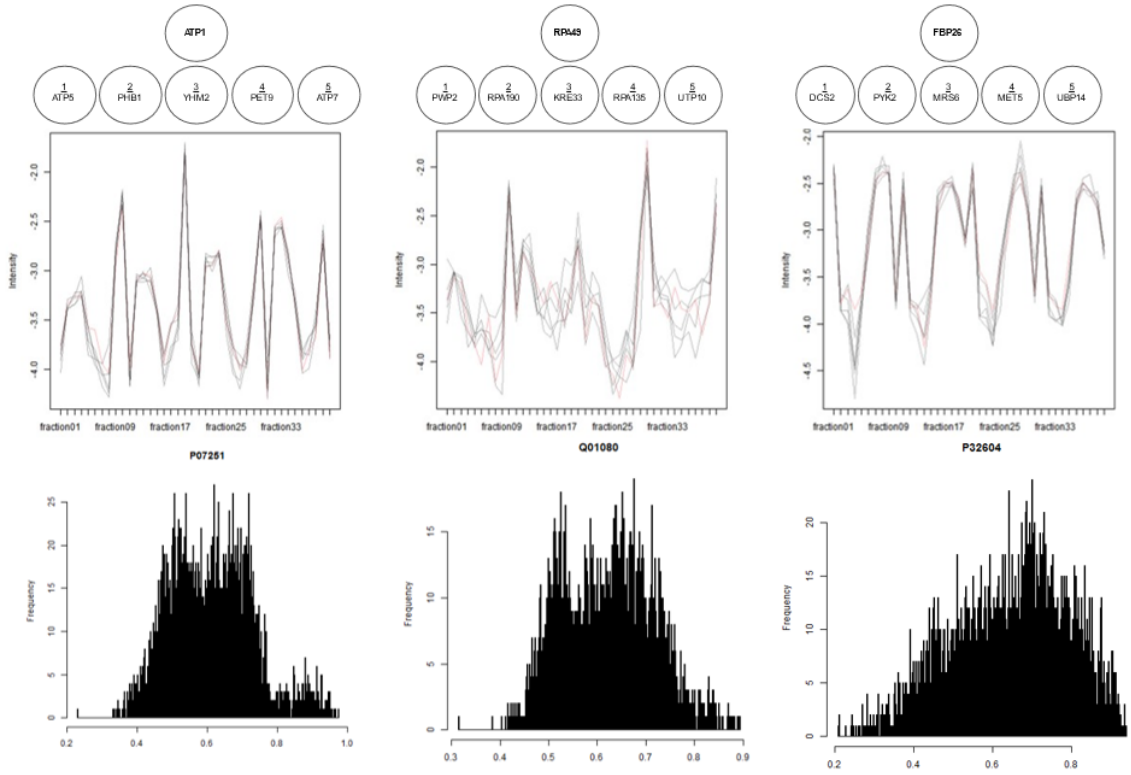

Figure 3.2. PCpulldown of *S. cerevisiae* Proteins Demonstrate Strengths and Weaknesses of Approach

I selected two S. cerevisiae proteins which were members of an intraorganellar protein complex and one cytosolic protein to demonstrate when this approach is appropriate. The ATP synthase subunit ATP1 (P07251) has two of its five most similar proteins as fellow ATP synthase subunits (left). The RNA Polymerase A subunit RPA49 (Q01080) similarly had two of its five most similar proteins as fellow RPA genes (middle). In contrast to these two, the cytosolic FBP26 regulates glycolysis and contains a semi-random mixture of other cytosolic proteins with which it is similar, however many in its top five play some role in the mitochondria or glycolysis.

Protein *Profile Similarity Scores Provide Useful Information about the Breadth of Cell Biological Interactions for Individual Proteins*

PPSSs range from zero to one such that proteins with lower PPSSs are more dissimilar from the query. We first benchmarked this approach using several different types of proteins from diverse organisms.

A hyperLOPIT experiment in the budding yeast *Saccharomyces cerevisiae* yielded protein abundance profiles for 2,847 proteins (Nightingale, Oliver, and Lilley 2019). PCpulldown of the F1 ATP synthase subunit ATP1 (P07251) recovers 5 ATP synthase subunits within the top 99th percentile range: ATP5, ATP3, ATP7, ATP2. The remainder of annotated ATP synthase subunits appeared as positions: 68th (ATP4), 76th (ATP16), 99th (ATP20), 192nd (ATP15). Intermixed with these complex members are additional mitochondrial proteins, and it is not until 185th most similar protein that a non-mitochondrial prediction appears, but that is ER-predicted gene HMG1 which does function in the mitochondria as well (Diffley and Stillman 1991). PCpulldown of the nucleolar, Pol I subunit RPA49 recovers several subunit A complex members in the 2nd (RPA190), 4th (RPA135), and 7th (RPA34) positions with five other non-specific RNA pol subunits appearing intermediate to these and the last subunit A member in the dataset at the 155th position (RPA43). RPA49's most similar protein was the pre-rRNA processing protein PWP2 (Dosil and Bustelo 2004). The first non-nuclear protein is the mitochondrial biotin synthase enzyme BIO2 at the 73rd position and may serve as an empirical cut-off for biological relevance. These two show examples of when this approach is useful: for intraorganellar protein complexes. Cytosolic proteins represent a different class with more ambiguous results. The glycolysis mediating FBP26 pulls down the purely cytosolic decapping protein DCS2 with no relationship whatsoever to glycolysis. Interestingly though, the 2nd nearest neighbor was the glycolytic PYK2, the third a mitochondrial splicing factor (MRS6), and the 12th was the glycolytic ENO2. This does suggest a tighter association between protein components of the same cytoplasmic pathways,

but the intermixing proteins are not as contained as those of proteins within the same organellar environment.

We previously mapped the spatial proteome of the ciliate *Paramecium tetraurelia*. I used the same method as above using a database of 2,345 proteins that did not require imputation. ATP1 BLASTs to a large family of proteins, the best hit in our data being PTET.51.1.P0330225. This was confidently predicted to be mitochondrial, and it is not until the 333rd position that a non-mitochondrial prediction– the peroxisomal thiolase enzyme– appears. The two most similar proteins are the inner membrane (MIM) proteins NNT (PTET.51.1.P0060269) and SDHB (PTET.51.1.P1270083). However, the next handful of similar proteins are mitochondrial ribosomes, and there is no clear demarcation between the MIM and other mitochondrial components. The 15th most similar protein is another ATP synthase subunit ATPeF1B (PTET.51.1.P0230128). RPA49's ortholog (PTET.51.1.P0540090) pulled down, as its most similar proteins, the ribosome biogenesis factor BRX1 (PTET.51.1.P0070153), and nucleolar proteins make up the vast majority of similar proteins with the first non-nuclear protein appearing as a ribosomal subunit at the 102nd position. These results support the utility of PCpulldown in describing the immediate proteomic neighborhood of a protein of interest.

PCpulldown Identifies New Interactions in Non-Model Eukaryotes

The apicomplexan parasite *T. gondii* was subjected to a hyperLOPIT experiment which revealed the localization pattern for 3,832 proteins (Barylyuk et al. 2020). The specialized cell invasion machinery of Apicomplexa is centered around the cortical microneme and rhoptry; organelles which perform regulated exocytosis and share numerous properties with other exocytotic structures across Alveolates (Gubbels and Duraisingh 2012). In one study, a group of Paramecium "non-discharge" mutants, TgND6 (TGME49_248640) and TgND9 (TGME49_249730), were shown to be conserved across Alveolata and important for regulated exocytosis in *T. gondii* (Aquilini et al. 2021). They performed immunoprecipitation-MS/MS to identify binding partners for these proteins and identified three candidates significantly

overrepresented in abundance: TgNdP1 (TGGT1_222660), TgNdP2 (TGGT1_316730, TGME49_249730), and TgFER2 (TGGT1_260470, TGME49_260470). Both TgND6 and TgND9 were identified in hyperLOPIT, although the former was classified to the Golgi and the latter to the peripheral plasma membrane compartments. Encouragingly though, PCpulldown of TgND6 identifies TgNdP2 as its 4th nearest neighbor and TgFER2 as its 46th. Conversely, TgND9 identifies TgNdP1 as its 1st nearest neighbor. Two named calcium dependent protein kinases are tightly associated with each ND gene: CDPK4 (11th: TGME49_237890) and CDPK7 (5th: TGME49_228750) for TgND9. These types of signaling genes are thought to play a major role in microneme function, and the plethora of other proteins identified here may aid in better elucidating this mechanism.

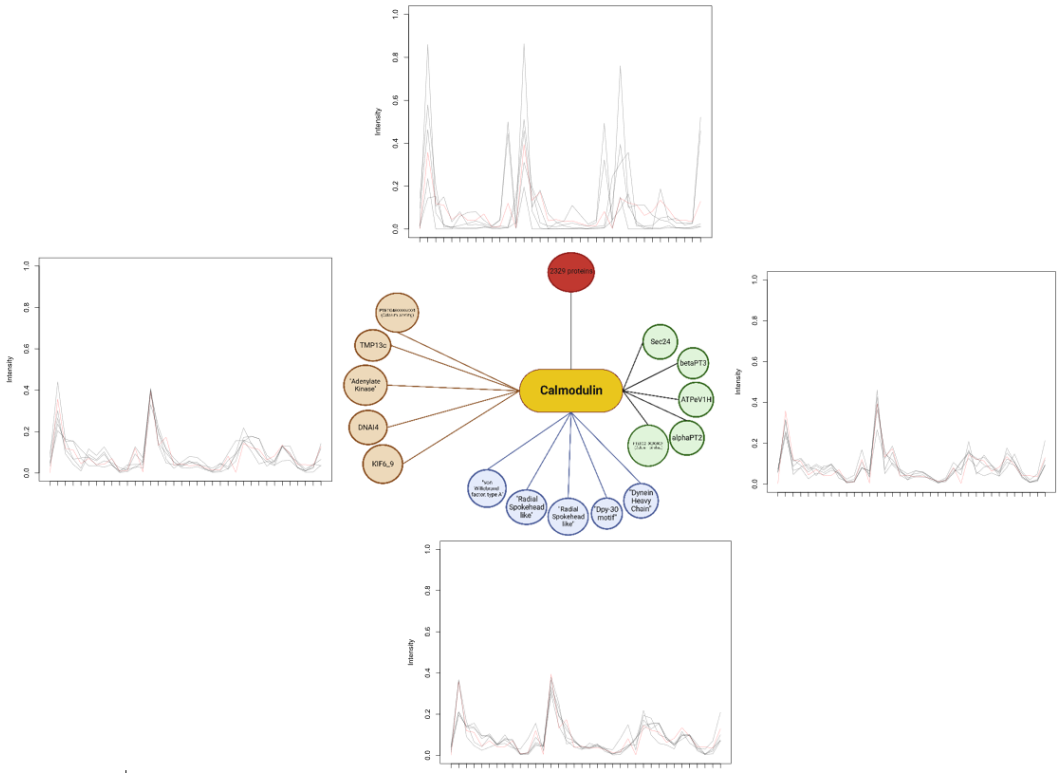

Figure 3.3. Calmodulin in P. tetraurelia Has a Complex Role in the Cell Cortex

The P. tetraurelia CAM1 protein abundance profile is shown in red with its five most similar proteins (right), 6-10 (bottom), 11-15 (left), and five randomly chosen proteins (top) from the non-imputed dataset. This approach recapitulates previously known interactions (e.g., between CAM and tubulin) and identifies a number of previously unknown interactions worthy of follow-up. Some of these have little to no annotation, although nearly all with annotated domains suggest some functional role in calcium signaling. Created with BioRender.com.

I then followed this up by studying the calcium signaling protein calmodulin (CAM) found promiscuously in the cell cortex as well as numerous membranous structures like food vacuoles and the contractile vacuole (Plattner 2013). CAM is highly conserved across eukaryotes, and in humans it binds directly to several hundred targets (Yap et al. 2000). In *Paramecium*, no gene has been subject to more biological scrutiny with dozens of well-characterized mutants. In our dataset, CAM1 (PTET.51.1.P0460139) was discretely classified to the axoneme and had its two most similar protein neighbors as the tubulins: alphaPT2 and betaPT3 (Figure 3.3). This relationship has never been observed in *Paramecium*, however the interplay between calmodulin-dependent processes and tubulin phosphorylation has been long known(Means and Dedman 1980). Several homologous components of the membrane trafficking system appeared as well, such as VPS33A (PTET.51.1.P0030390) at the 16th position. Both CAM and various VPSs are well-known to regulate endosome formation (Colombo, Beron, and Stahl 1997; Babst et al. 1998), and this link connects the complex membrane trafficking machinery to the regulatory and structural components of the cell cortex. The ciliate-specific gene PTET.51.1.P0060450 is the first with a TMD at the 31st position, and it is annotated with the GO CC term “metal ion binding”. Nearly all of CAM’s most similar proteins have some expected role in the functioning in the cortex, and many are likely involved in calcium signaling. This pattern likely reflects the underlying and broad role that CAM plays in regulating calcium signaling in order to modulate the activity of these diverse cellular membranes. Those unnamed genes are excellent candidates for follow-up, and this approach can easily be applied to any protein of interest.

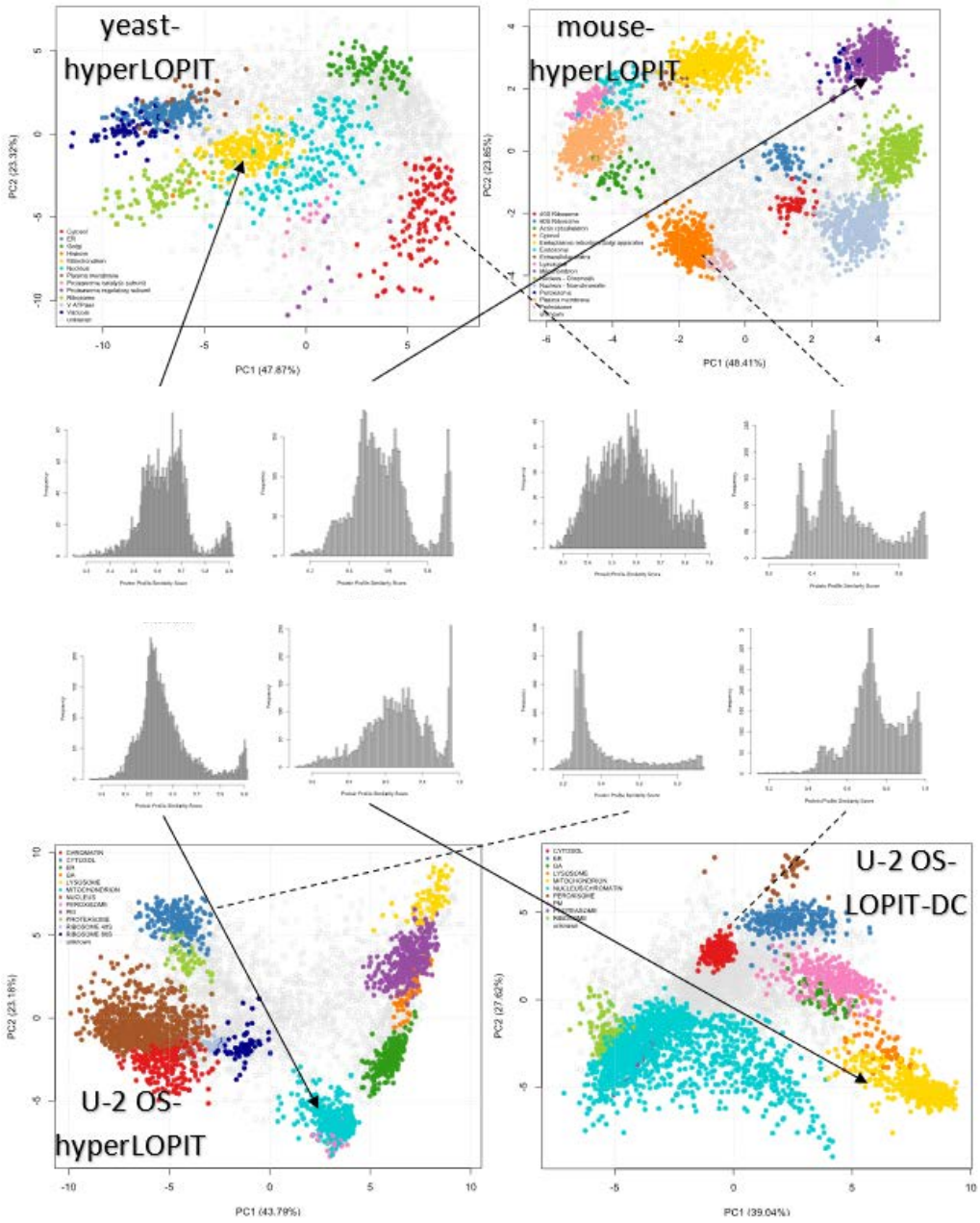

Figure 3.4. Differential Resolution of the Cytosol and Mitochondria Across LOPIT Datasets

PPSS distributions display a wide range of shapes, although some are constant for the same organelle in different types of experimental set-up. The mitochondrial shapes (left) are typically multimodal, with a far-right peak of highly similar proteins followed by a broader, left peak. This behavior supports the hypothesis that membrane bound organelles constrain their proteins from being highly similar to proteins not in their organelle. In contrast, cytosolic proteins (right) are not bound by these constraints and adopt often broader PPSS distributions depending on the organism and experimental design. For example, the LOPIT-DC data for U-2 OS cells has a far different pattern than hyperLOPIT data from the same sample, the former being multimodal and the latter having a long right tail. The left versus right skewness of these distributions reflects whether the mean organellar protein profile is less or more similar, respectively, to all other protein in the dataset.

The Topology of Protein Profile Similarity Scores Differs Across Organelles, Organisms, and Experiments

I noticed that proteins predicted to the same subcellular compartment often had similar PPSS distributions, and I wondered if this was a general feature which reflected their global associations in the cell. Many organellar proteins had PPSS distributions with multiple peaks, and it is tempting to imagine this is caused by gradients of decoupling between proteins of the same pathway/complex, within the same organelle, and those associated purely by chance. Cytosolic proteins, however, should not be as clearly demarcated in this manner due to the lack of a membrane-bound barrier between them and other proteins.

To determine the generality of this phenomenon, I created a mean PPSS distribution for each compartment in both *S. cerevisiae* and *T. gondii* hyperLOPT data (Nightingale, Oliver, and Lilley 2019; Barylyuk et al. 2020) as well as both hyperLOPIT and LOPIT-DC data from human U2O2 cell lines (Geladaki et al. 2019). One constant across these datasets was the mitochondria, which represented a singular compartment characterized by a tight cluster when projected onto principal components 1 and 2 (Figure 3.4). The three hyperLOPIT PPSS distributions appeared more similar in that the rightmost peak (most similar protein profiles) and middle peak (moderately similar profiles) had a large gap between them, while the gap in the LOPIT-DC data was smaller, but the right peak was higher. In general, the mitochondria are well-resolved in any study of this kind, but the higher relative proportion of points in the rightmost peak from the LOPIT-DC data suggests a possibly better resolution than in hyperLOPIT. In contrast, cytosolic proteins adopted many shapes to their distributions, with huge differences observed within the same U2-OS cell line depending on whether the data came from hyperLOPIT or LOPIT-DC design (Figure 3.4). LOPIT-DC was already known to better resolve the cytosol and proteosome (Geladaki et al. 2019). These observations give hints into the topological organization of proteins in terms of their global interaction profiles as reflected in spatial proteomics data.

Discussion:

PCpulldown is a useful tool when querying spatial proteomics data for individual proteins. The data produced from these studies are massive, but the analyses performed in the articles producing them are typically superficial and focus on expanding the protein inventory for a handful of cellular compartments. To make these data more useful for a broad community, there must be an easy way to access them. This is particularly true on non-model systems in which these techniques are finding new life such as the alveolates *P. tetraurelia* and *T. gondii* (Barylyuk et al. 2020). I show an example in *P. tetraurelia* using the promiscuous but critically important calcium-signaling protein, CAM1, and reveal new properties about its cell biological context which may aide researchers attempting to better understand its role in the cortex. A molecular biologist with any level of R coding background should be able to source and execute the two functions described in this study, although more will be produced.

CHAPTER 4

THE SUBCELLULAR FATE OF DUPLICATE GENES IN PARAMECIUM TETRAURELIA

Abstract:

Gene duplications are thought to underly much of the observed phenotypic diversity in eukaryotes by providing the raw materials for the functional diversification of resulting paralogous genes. Despite this, the ciliate lineage *Paramecium aurelia* has experienced two whole genome duplication (WGD) events preceding its speciation into 14 species that are morphologically identical. Much work has gone towards understanding the parallel patterns of gene loss in different *P. aurelia spp.* and the functional diversification of retained paralogs (called ohnologs) using tools like mRNA sequencing. However, little is known about the expression and localization of the protein products of these duplications. Recently, we mapped the spatial proteome of *P. tetraurelia* using a combination of cell fractionation and quantitative proteomics (Licknack et al. in prep). The protein abundance data underlying protein localization prediction hold clues as to changes in protein function due to reallocation of protein material to different regions of the cell. Of all possible pairs of ohnologs from the three most recent WGD events, we discovered ~4,500 instances where just a single protein was identified and another ~4,200 instance where both protein pairs were identified. Protein pairs in which only one copy was identified were significantly more divergent at the sequence level than those pairs in which both copies were. Protein pairs with significantly dissimilar abundance profiles tended to have N-terminal indels relative to their ancestral state, but there was a weak relationship between sequence divergence as a whole and quantitative, subcellular divergence. We observed a plethora of enriched gene duplicates predicted to be either ribosomal, proteasomal, or of the trichocyst matrix of *P. tetraurelia*, possibly providing a cell biological explanation for the previously reported relationship between mRNA expression and ohnolog retention. Taken together, we provide a novel approach to studying the functional diversification of gene duplicates at the level of cell biology.

Introduction:

Gene duplication has played an undeniable role in the evolutionary process by providing the raw material for the expansion of the protein repertoire for the lineage in which it occurs (Ohno 1970). There has now been a deep literature produced which describes both how evolutionary forces act to determine if a gene duplicate is retained as well as how functional consequence of that retention at the level of mRNA expression, protein sequence, and organismal fitness (Force et al. 1999; McGrath et al. 2014; Innan and Kondrashov 2010; Gout and Lynch 2015; Ohno 1970; Lynch and Force 2000). Although the most common fate for a duplicate gene is loss, retention can occur if both copies are fixed and preserved over evolutionary time. While fixation simply describes the state in which the entire population contains both duplicated gene copies, perseveration of those duplicates then describes their symmetric or asymmetric rate of sequence evolution. Fate-determining mutations can arise before the fixation of the post-duplicate state in the population, and these events can seed more mutations leading to the divergence between the two copies. The extent to which these mutations alter gene function is constrained by the cell-biological environment in which those gene products must act.

The constraints on the evolutionary fate of duplicate genes are alleviated somewhat when whole-genome duplication (WGD) events result in the systematic duplication of every gene in the genome alongside its entire suite of regulatory elements and genomic contexts(Birchler et al. 2001). Paralogs resulting from such events are called ohnologs. Evidence of their occurrence can be found across the tree of eukaryotic life from animals, plants, fungi, and a variety of protists. In this ciliate lineage *Paramecium aurelia*, a cryptic species complex has evolved after two subsequent WGD events with a possible ancient duplication preceding the split between *Paramecium* and *Tetrahymena* (Aury et al. 2006). While much effort has gone towards characterizing the patterns of gene loss in each *P. aurelia* spp. (Gout et al. 2019), less effort has been made on the functional impacts of WGD events at the cell biological level. One clear trend to emerge is the relationship between mRNA expression level and ohnolog retention such that more highly expressed genes are more likely to be retained (McGrath et al. 2014). Investigations

into protein expression and function are lacking a similar high-throughput nature and typically resort to manually tagging and assaying individual proteins and observing their localization via microscopy (Hauser, Pavlovic, et al. 2000). The emergent techniques of spatial proteomics have been applied to a handful of microbes (Nightingale, Oliver, and Lilley 2019; Barylyuk et al. 2020) and illuminated the broad localization patterns for thousands of proteins simultaneously. This is achieved through the gentle lysis and fractionation of a cell culture followed by quantitative mass spectrometry in order to generate unique profiles for each identified protein corresponding to its subcellular localization. In a recent study, we applied this technique to the model ciliate *Paramecium tetraurelia* and determined protein localization for over 9,000 proteins. We predicted protein localization to one of seventeen compartments spanning diverse cell biological structures like the cortex (two basal body, one trichocyst matrix, one axoneme, one surface antigen cluster), nuclei (soluble and insoluble), mitochondria (outer membrane and remainder), peroxisome, lysosome, membrane trafficking (soluble and insoluble), ER, cytosol, and the ribosome and proteasome complexes. The strict cut-offs used in that study largely ignored any differentially identified ohnologs which may be functionally divergent, so here we pay special attention to the thousands of pairs of duplicates and at multiple levels of phylogenetic divergence.

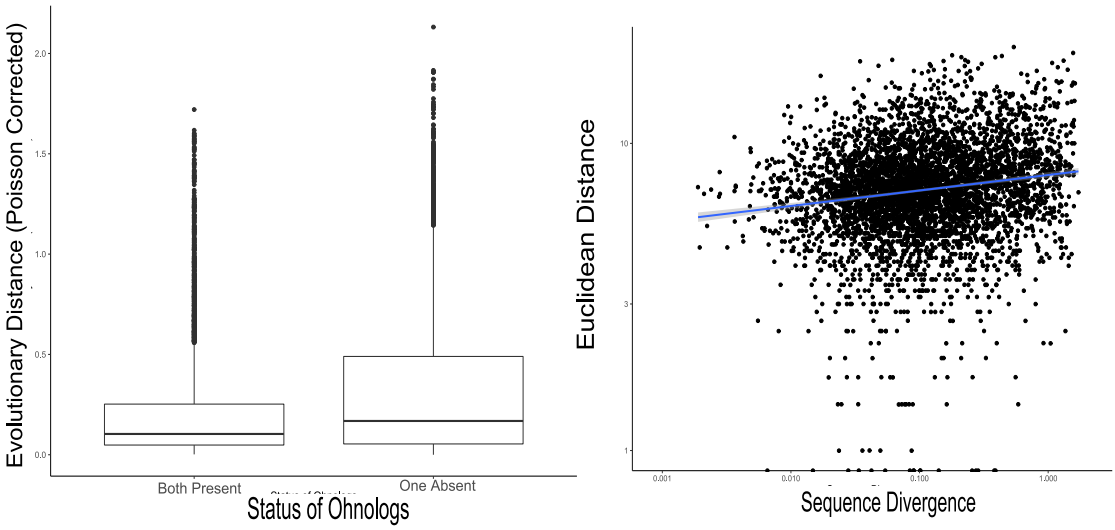

Figure 4.1. Changes in Protein Sequence Affect the Differential Identification of Ohnologs but Do Not Determine the Extent of their Functional Diversification
We used quantitative proteomics data to identify WGD ohnologs differentially present in biochemically distinct, subcellular fractions. Proteins without any peptide evidence in any fraction were considered absent (0), and present proteins were coded as either lowly (1) or highly (2) abundant based on the presence of one or many peptide peaks corresponding to that protein.
Left: Ohnolog pairs in which both copies were present were compared to those in which only one was present based on the pairwise evolutionary distance between their amino acid sequence calculated using MEGAX; Poisson Correction (Kumar, Tamura, and Nei 1994). Pairs in which both proteins were present were significantly more similar to each other at the sequence-level than pairs in which only a single copy was identified and the other was absent ($\rho \sim= 0$).
Right: Pairwise Euclidean distances were computed between coded abundance profiles of all ohnolog pairs and regressed against their evolutionary distance in a log-log plot. Heteroscedasticity was observed with a narrowing of points as protein pairs became more dissimilar at the protein sequence level. This weak trend suggests a weak relationship between sequence and subcellular divergence as seen in a spatial proteomics experiment.

Results:

Thousands of gene duplicates are differentially detected in a deep proteomic survey of Paramecium tetraurelia

In a previous study, we reported the protein localization patterns for over 9,000 proteins of *P. tetraurelia* using a combination of cell fractionation and quantitative proteomics (Licknack et al. *in prep*). While we identified 11,856 *P. tetraurelia* proteins in total, we excluded many whose low number of peptide spectral matches (PSMs) resulted in noisy data not reflective of their localization pattern. This indirectly resulted in the systematic removal of several hundred proteins which may contain clues corresponding to the diversification of gene duplicates which are differentially present or absent from different fractions. Due to a positive relationship between the number of PSMs identified in the proteomics data and mRNA expression data from the underlying gene (Arnaiz et al. 2010), this biased our coverage towards more highly expressed genes (Figure 4.4). We thus included all proteins with any level of proteomic support. Of all possible pairs of WGD1 (young), WGD2 (intermediate), and WGD3 (ancient) ohnologs, 4,490 pairs had only one copy identified in the proteomics data while another 4,214 pairs were both. Across all phylogenetic levels, ohnolog pairs with smaller evolutionary distances were more likely to both be identified than more distant pairs (Figure 4.1). This is expected due to the nature of proteomics experiments in which shared peptides aide in the identification of proteins despite unique peptides being a prerequisite (Zhang et al. 2010). But the possibility that one ohnolog is not identified due to its lack of expression opens a potential avenue for studying changes in gene function. In all but a single case, both gene copies are expressed at the mRNA level.

For those ohnolog pairs in which both copies were present, we measured the Euclidean distance (i.e., the “norm”) between protein abundance values measured across 108 fractions (three technical triplicates of three biological triplicates across twelve subcellular fractions) using coded values of zero for absent, one for detectable, and two for highly abundant based on standard protein identification techniques implemented in ProteomeDiscoverer (Orsburn 2021).

This coding system removes the need to impute missing values while capturing large-scale trends in differential protein abundance across subcellular fractions. A protein entirely missing from all fractions compared with a protein highly abundant in all fractions would yield the largest Euclidean distances and thus would be the most dissimilar from each other at the level of protein localization. The scaling relationship between pairwise Euclidean distances (i.e., localization divergence) and evolutionary distances (i.e., sequence divergence) was very weak and exhibited heteroscedasticity (Figure 4.1). In lieu of an alternative metric to measure quantitative changes to protein localization, these data suggest a decoupling between sequence evolution and protein localization.

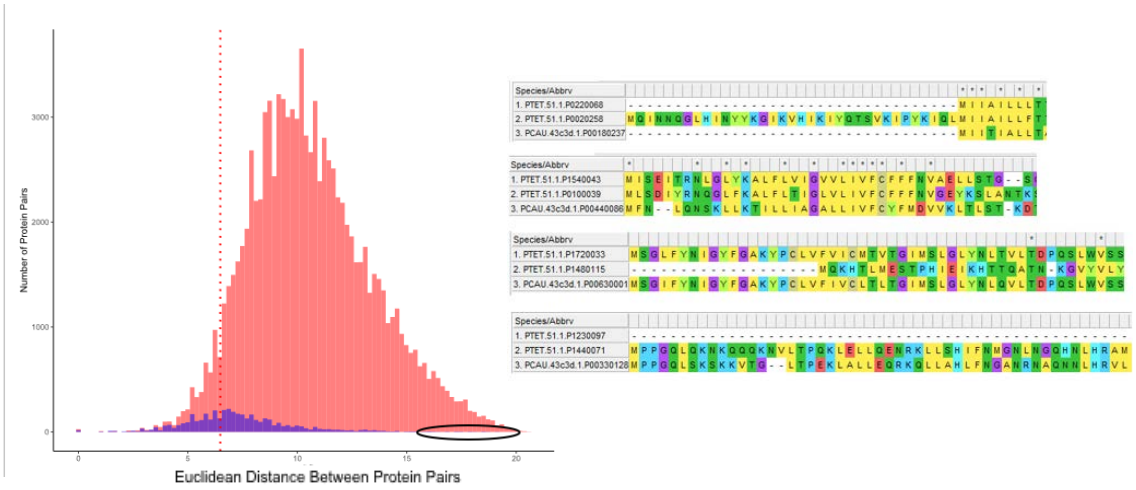

Figure 4.2. Functionally Divergent Proteins Often have N-terminal Changes
Left: The pairwise Euclidean distances between ohnolog pairs (purple) was compared with randomly chosen proteins (red) to determine which ohnologs were statistically dissimilar from one another. The dotted red line denotes the bottom 5th percentile of the random pairs and a possible cut off between divergent and non-divergent ohnolog pairs. Circled points denotes those whose Z-scores indicated with 99% confidence (17.9-20.6) were identical to randomly chosen protein pairs.
Right: Protein alignments for the four most divergent ohnolog pairs were made with ClustalW and are shown in ascending order of their Euclidean distance. Three of four of these proteins have large, N-terminal extensions or deletions when compared to their single-copy *P. caudatum* ortholog. The only pair without a large-scale N-terminal change does contain two smaller N-terminal insertions and deletions. These observations support the importance of N-terminal peptides in the evolution of protein localization.

We then wished to determine highly divergent ohnolog pairs by comparing their pairwise Euclidean distances to randomly chosen proteins from the same dataset (Figure 4.2). This latter dataset should serve as a null expectation under the assumption that only a minority of unrelated proteins should have similar patterns (Figure 4.1). Only 1,552 ohnolog pairs had a Euclidean distance within the highest 95th percentile of random proteins, while the other 2,662 were less than this cut-off. These 1,552 pairs constitute ohnologs which are as different from one another as most randomly chosen pairs of proteins. We then calculated a Z-score using this random expectation to identify 18 ohnolog pairs that are significantly different in their differential abundance profile with 99% confidence. Of the four highest Euclidean distances, all contain some either a large indel or many small indels in their N-terminal peptide (NTP); one being a large extension of the NTP, two large deletions, and one with many smaller changes. These results validate expectations that NTPs play a predominant role in driving protein relocalization (Byun and Singh 2013).

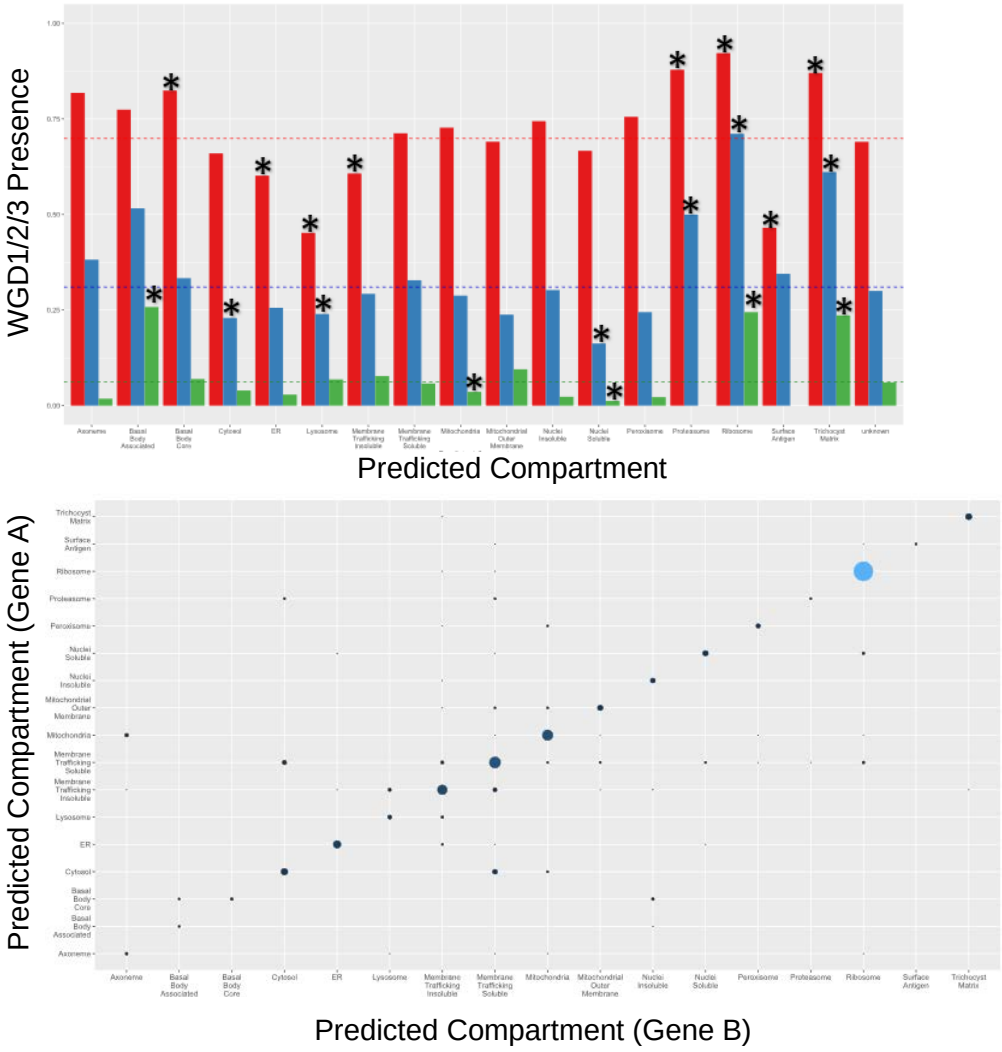

Figure 4.3. Ohnolog Retention and Relocalization is Associated with Predicted Subcellular Compartments

Top: Using predictions from Licknack et al. (in prep) across over 9,000 protein-coding genes, ohnolog presence was compared across three levels of phylogenetic divergence: young- WGD1 (red), intermediate- WGD2 (blue), old- WGD3 (green). Means for the dataset plotted as dotted lines with the following values: ~70% (WGD1), ~31% (WGD2), and ~6% (WGD3). Stars indicate statistically significant enrichment or depletion of ohnolog presence relative to the uknown category which contains a semi-random mixture of proteins across the proteome with weak support for their predicted localization to any compartment. Ribosomal and trichocyst matrix predictions were disproportionately enriched in all three types of ohnologs possibly owing to their high mRNA expression levels (Supp Figure). Proteasomal predictions were enriched in both WGD1 and WGD2 ohnologs but notably contained no proteins with WGD3 still retained. Conversely, the associated basal body predictions were only significantly enriched at the WGD3 level.

Bottom: Organellar predictions were compared between all WGD1/2/3 ohnolog pairs such that larger, lighter dots are indicative of more proteins in that class. Most protein pairs were either both 'unknown' or contained one unknown copy and another predicted to some compartment like the ribosome or mitochondria. The latter case implies that one pair is more ambiguously assigned than the other which may reflect meaningful subcellular divergence, but a few other cases appear to be more straightforward changes to new subcellular locations. The broad class of soluble membrane trafficking proteins was often involved in relocalization events between confidently predicted protein pairs. This was a similar trend to that seen in hyperLOPT data in S. cerevisiae (Figure 4.5).

Numerous gene duplicates have entirely different predicted protein localizations in both P. tetraurelia and S. cerevisiae

We reanalyzed the previously discussed *P. tetraurelia* spatial proteomics dataset to determine the relationship between ohnolog retention and discrete changes in protein localization. We first determined the rate of retention for WGD1, WGD2, and WGD3 ohnologs and compared them by their predicted subcellular compartment (Figure 4.3). WGD1 ohnologs were present in ~70% of proteins in our dataset, but several compartments were either significantly enriched or depleted. In the former category were the basal-body core proteins, proteasomal, ribosomal, and trichocyst matrix compartments. In the latter are the ER, lysosomal, surface antigen, and insoluble membrane trafficking compartments. WGD2 ohnologs were present in far fewer proteins– averaging ~31% across the dataset and only being significantly enriched in proteasomal, ribosomal, and trichocyst matrix compartments while being depleted in cytosolic and soluble nuclear predictions. A minority of genes still retained ancient duplicates (~6%), but there were a handful of compartments highly enriched with WGD3 ohnologs: basal body associated, ribosome, and trichocyst matrix. Conversely, the mitochondrial and soluble nuclear compartment were significantly depleted, and the proteasome compartment contained no genes with retained WGD3 copies. This leaves both the trichocyst matrix and ribosomal compartments containing the largest numbers of young, intermediate, and ancient duplicates. These two compartments have the highest mRNA expression, as does the basal body network, so this observation tracks well with the positive relationship between gene expression and gene duplicate retention (Gout et al. 2010).

Most WGD ohnologs in this study were classified to the same compartment, either both being ribosomal, mitochondrial, or in one of the membrane trafficking compartments (Figure 4.3). In total, 96 pairs of ohnologs were classified to a different compartment and may be true relocalization events. This commonly occurred between the cytosol and soluble membrane

trafficking compartment, the former being made up of most cytoskeletal elements and protein complexes involved in protein transport. These changes could occur simply through modest changes in relative abundance away from the Sup (supernatant) fraction and towards the protein processing (12K-30K) fractions overlapping with ER chaperones. More surprisingly are cases involving relocalization to entirely different compartments, although there are very few of these. One example is the WGD1 pair, PTET.51.1.P0480068 and PTET.51.1.P0340227, the former of which is confidently ribosomal, and the latter is confidently predicted to the soluble membrane trafficking compartment. Both are putative nucleosome assembly factors (NAP1L1) with homologs across eukaryotes acting as chaperones of histone proteins. However, the NAP1L1 gene in *Arabidopsis thaliana* has been shown to directly interact with ribosomes (Son et al. 2015), so this may be the case in other species. When we aligned and built a tree of both *P. tetraurelia* protein sequences with that of *P. caudatum*, we found much longer branch lengths in the ribosomal copy (Figure 4.6). Upon inspection of the alignment, this copy contains a 5 bp deletion immediately after the position 140 glutamine (Q), which appears to be substituted from an ancestral lysine (K) overlapping its expected dimerization domain (Zhou et al. 2015). How the abolition of its ancestral chaperone roles could occur from such changes is unknown, as is the true ancestral state of the *Paramecium* NAP1 genes. A similar study conducted on a pre-duplicate outgroup, like *P. caudatum,* would aid in clarifying this and other examples.

We then reanalyzed a hyperLOPIT dataset that reported protein localization predictions for 2847 proteins in *S. cerevisiae* (Nightingale, Oliver, and Lilley 2019) using the prolocdata R package (Gatto, Breckels, Wieczorek, et al. 2014). While a small fraction of its ancient WGD duplicates have been retained, 546 pairs were, and of these, 175 pairs had both copies identified and predicted in this proteomics dataset. The clearest observation was that most ohnologs were classified to the same compartment (Figure 4.5). Of those with different classifications, relocalizations between the cytosol and nucleus made up the largest class, possibly owing to their classifications making up roughly 61% of the assayed proteome. Eight ohnolog pairs contained a nuclear and cytosolic copy. Changes between the cytosol and nucleus can be achieved easily through the loss of a nuclear localization signal (NLS), although not all the nucleus-cytosol pairs

contained a predicted NLS via NLStradamus (Nguyen Ba et al. 2009). One pair that did were the cytosol-predicted TCD1/YHR003C and nucleus-predicted TCD2/YKL027W, the latter containing the putative NLS: 240-RRKLKKR-246. The role of tRNA threonylcarbamoyladenosine dehydratases in the cytoplasm, nucleus, and mitochondria has been of considerable interest, and this observation may shine light into one possible mechanism of splitting bidirectional pathways with specialized ohnologs (Chatterjee et al. 2018). Another pair was the cytosolic FKS1 and nuclear GSC2, wherein both contained roughly the same putative NLS: 244-GKLSRKARKAKKKNKK-259 in the former and 261-KLGKLSRKARKAKKKNKK-278 in the latter. Another ohnolog pair was the ER-classified CPR5/CYPD and cytosol-classified CPR3/CYPB (both cydophilins), the former containing the classical ER-retention signal HDEL (Pelham 1990) and the latter missing that region of its C-terminus (Figure 4.7). A simple string search identified that non-*Saccharomyces* yeast orthologs had some version of a C-terminal DEL motif, and this is likely the ancestral state. These and other findings support the use of high-throughput protein-localization techniques to discover widescale evidence of protein relocalization in diverse eukaryotes.

Discussion:

Here, we provide a preliminary survey into the use of spatial proteomics to detect large-scale changes in protein localization between gene duplicates. The value of this approach to cell biology is undeniable— the ability to survey protein localization of nearly every expressed protein in the proteome with high resolution (Christopher et al. 2021; Gatto, Breckels, Wieczorek, et al. 2014; Lundberg and Borner 2019). However, little to no effort has been made into the ability of this technique to detect divergent paralogous genes, despite the development of the original LOPIT technique in the polyploid *Arabidopsis thaliana* (Dunkley et al. 2004). By using only four subcellular fractions, this study was able to spatially resolve hundreds of proteins and assign them to specific subcellular niches. Now, TMT16-plex (Thermo) offers the possibility of assaying dozens of subcellular fractions simultaneously, and the continued development of label-free

methods of protein quantification provides additional flexibility to these workflows. As more subcellular fractions are included, the better resolution can be attained between gene duplicates for subtle changes in steady-state protein localization behavior.

In our study, we used two approaches to identify divergent paralogs/ohnologs. One somewhat clear-cut method involved the determination of duplicate copies predicted to different compartments (Figure 4.3; Figure 4.5). Here, we rely on the 'traditional' techniques of spatial proteomics to make use of supervised classification algorithms in order to determine copies that are more similar to a pre-defined class of marker proteins chosen *a priori* as residents of a particular organelle or compartment. The benefit of this approach is the simplicity in its output: copies are either both predicted to the same compartment or each to a different compartment. An additional level of complexity can be added by including some cut-off value under which proteins are predicted as 'unknown', as is commonly done using SVM scores or Bayesian measures of uncertainty (Breckels et al. 2016; Crook et al. 2018; Gatto, Breckels, Burger, et al. 2014). One concern with this approach is its reliance on a prior knowledge of the system which motivates the choice of one or another marker protein to serve as the expected behavior of the organelle. This is even more important in non-model systems in which few— if any— proteins have been subjected to direct assays of protein localization. An alternative approach is to simply measure the higher-dimensional distance between two proteins agnostic of the true localization of either copy (Figure 4.2). In this study, we measured the Euclidean Distance between ohnologs in *P. tetraurelia* using their coded absence, presence, or high abundance and found many divergent pairs with clear changes to the NTP. Interestingly though, large-scale changes in sequence alone could not explain the observed divergence in changes to the pair's spatial pattern (Figure 4.1). Since NTPs are typically quite small (roughly a dozen amino acids), we do not expect major structural changes to a protein to be a prerequisite for changing localization patterns. However, one may expect that such changes should institute a selective regime in which subsequent changes are more likely. An alteration to an NTP regulating protein sorting may relax selection on functional domains only active within the originally destined location, and vice versa. Since NTPs

evolve more rapidly than the rest of the protein (Williams, Pal, and Hurst 2000), one may expect the former scenario to happen quite quickly.

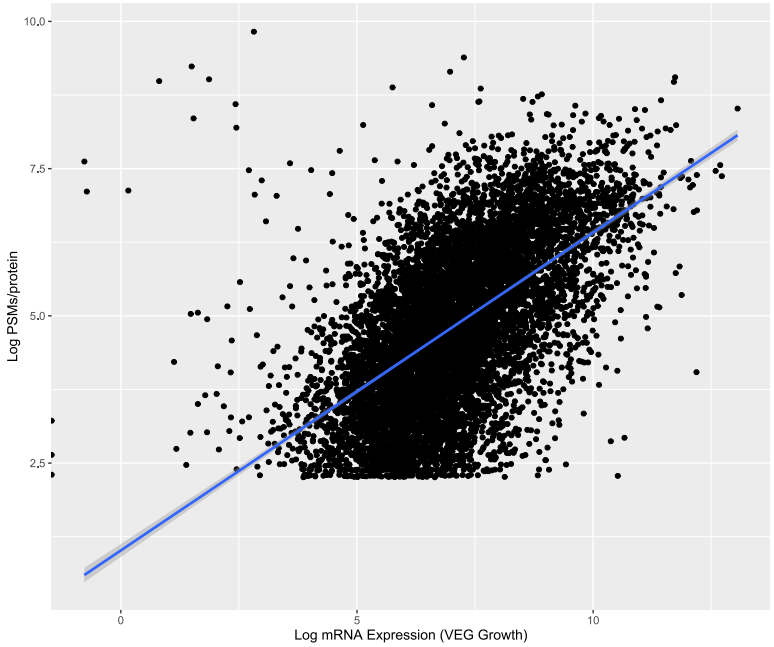

Figure 4.4. More Highly Expressed Genes at the mRNA level Produce Proteins with a Higher Number of Peptide Spectral Matches
Proteins identified and quantified in Licknack et al. (in prep) were done using peptide spectral matching implemented in ProteomeDiscoverer. Each match (PSM) reflects the number of events in which the mass spectrometer encounters the peptide underlying the protein of interest. The summed PSMs per protein were regressed against the mRNA expression level of the corresponding gene, pulled from the ParameciumDB, on a log-log plot. The adjusted R-squared of this fit was ~0.32 and the p-value was highly significant ($p < 2.2e-16$), meaning that this relationship is quite strong.

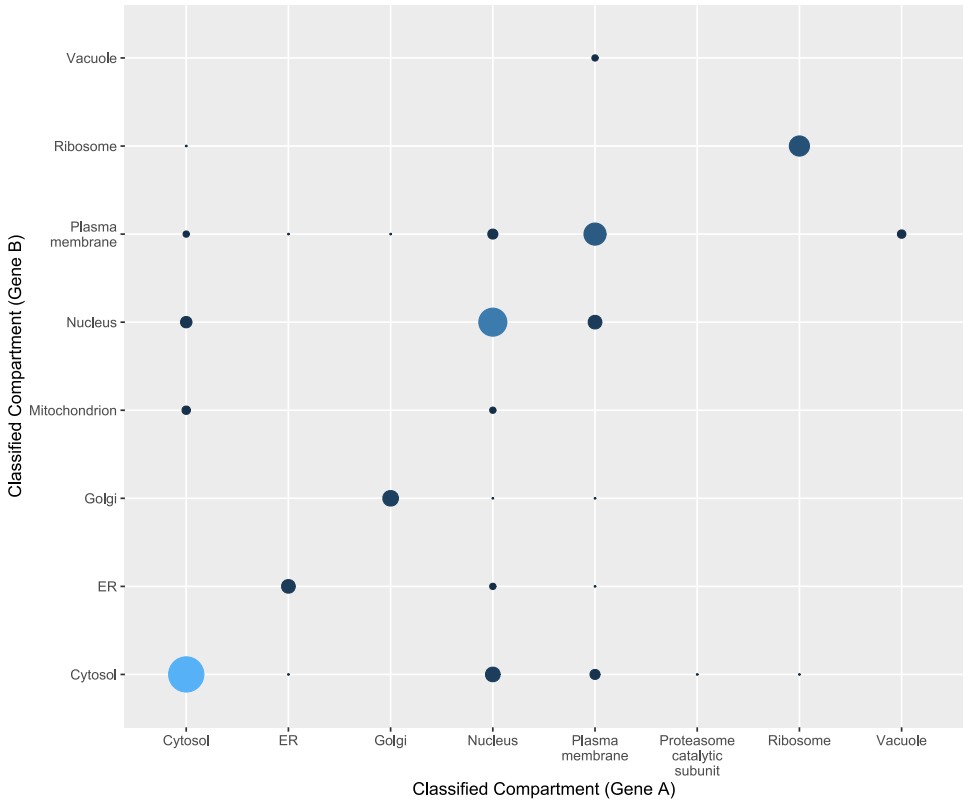

Figure 4.5. Ohnolog Relocalization in S. cerevisiae
We detected 175 instances in which both copies of ohnologs were classified in S. cerevisiae hyperLOPIT data and plotted the localization of both copies. In most cases, both copies were classified to the same compartment, the most common fate being two cytosolic ohnologs, but there are many instances of differences in one copy, the most common being between the nucleus and cytosol. Mitochondrial, vacuolar, and— to a less degree— proteasomal ohnologs never had both copies classified to them.

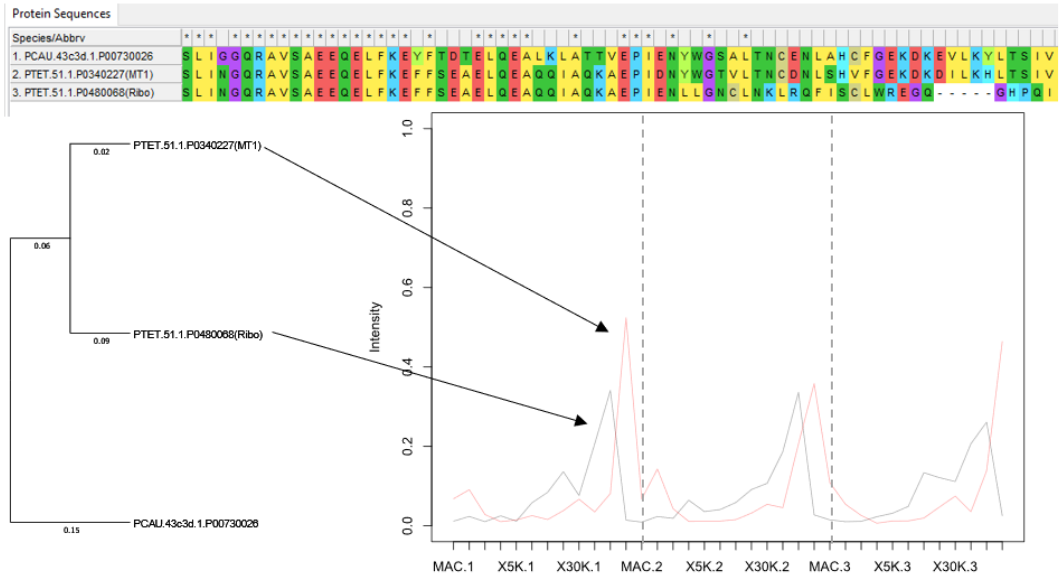

Figure 4.6. A NAP1 Homolog in *P. tetraurelia* May Relocalize to the Ribosome After the Deletion of its Putative Dimerization Domain

Ohnologs with different classifications are candidates for bona fide protein relocation events, and here we show an example of one such case. The pair PTET.51.1.P0480068 and PTET.51.1.P0340227 are both homologous to the nucleosome assembly factor NAP1. This protein's probable ancestral state is cytoplasmic— which is partially explained by the soluble membrane trafficking compartment. However, one copy appeared clearly ribosomal (bottom right). The x-axis represents each subcellular fraction across three biological replicates (i.e., different experiments), while the y-axis is the protein's normalized, relative abundance in that fraction. This copy had longer branch lengths than its ohnolog when aligned with its orthologous *P. caudatum* copy (left) and contained a small deletion which overlaps with its putative dimerization domain (top). This example shows how spatial proteomics data can identify relocalized ohnologs in a non-model system.

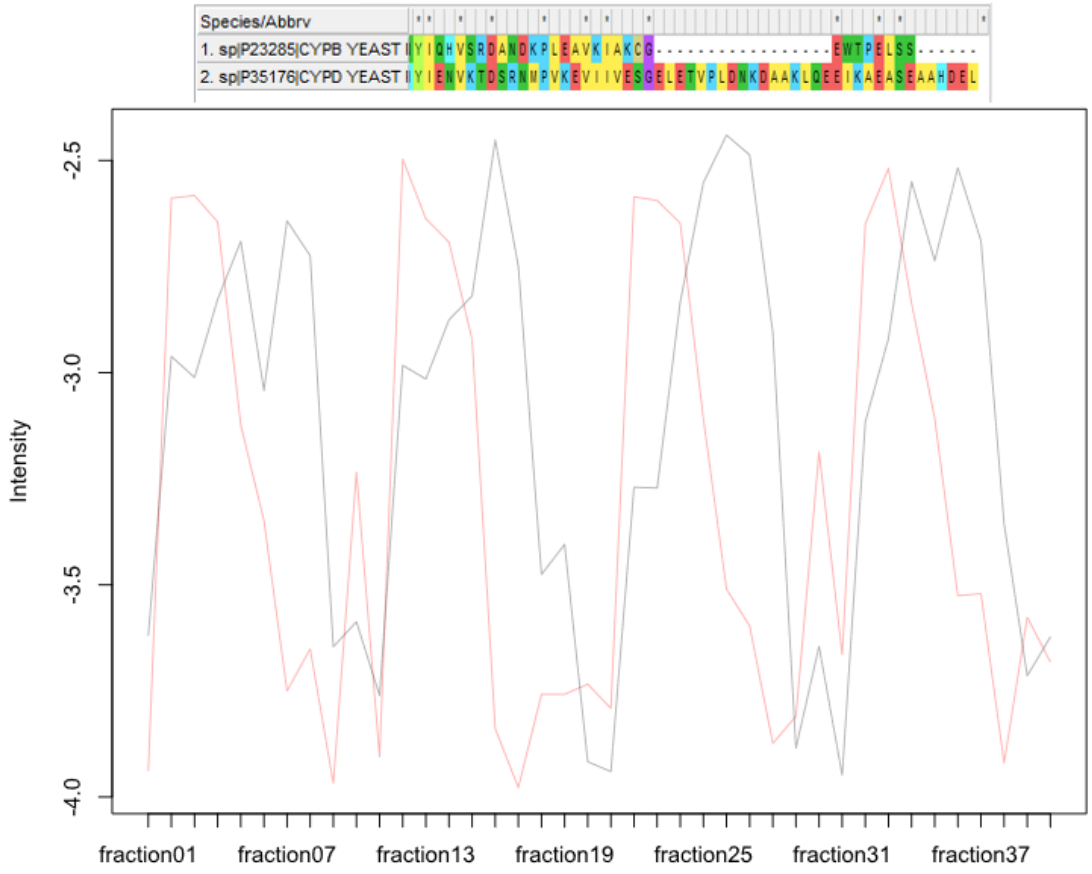

Figure 4.7. A S. cerevisiae Relocalization Event Associated with the Loss of HDEL
We identified CYPB and CYPD as ohnologs classified to different compartments in S. cerevisiae. CYPB (bottom, black) is cytosolic, while CYPD (bottom, red) localizes to the ER. Their abundance profiles displayed stark differences driven by differential abundance in different subcellular compartments (bottom). The axes are the same as that of Figure 4.6, but here four biological replicate experiments were done resulting in 40 fractions. The ancestral state of this gene is likely of the ER due to the persistence of the HDEL motif on the C-terminus associated with retention in the ER, however CYPB contains a large deletion of this region (top) which might have caused its lack of ER-association.

CONCLUSION

Protein localization is a key step in gene expression and is particularly important in the context of large, compartmentalized eukaryotic cells with numerous membrane-bound and amembranous organelles. Emergent tools in spatial proteomics have enabled the simultaneous assaying of protein localization patterns for thousands of proteins without the need for sophisticated molecular manipulation (Borner 2020). This is particularly useful for non-model systems. However the reliance on *a priori* knowledge may limit the generality of each predicted organellar compartment—not so much that predictions are erroneous but more so that they are biased. To this end, I not only generated and analyzed a deep spatial proteomics dataset in the ciliate *P. tetraurelia*, but I also developed a simple tool to allow researchers to glean cell biological knowledge from this and other datasets without the need to rely entirely on the marker-protein based predictions. I described only a small handful of the numerous cell biological stories one could extract from these data, but I hope the interpretation and tools provided will find a useful place within the Paramecium community (Arnaiz, Meyer, and Sperling 2020) as well as the broader fields of cell and evolutionary biology.

While I consider this work to be rich, there are a handful of regrets and disappointments that have persisted. The first is the lack of resolution between the MAC and MIC of *P. tetraurelia*, which one may expect to be simple to resolve due to their massive size difference (Cummings 1977). I took numerous steps to increase the likelihood that these nuclei pellet in different fractions, such as retaining the first differential centrifugation fraction (300g) which is normally discarded and including an enriched MAC fractions generated through an entirely separate lysis/fractionation scheme (Figure 2.7). Western Blotting data confirmed the existence of numerous Histone protein isoforms in different fractions, and Histone H3 is known to have MAC- and MIC-specific isoforms due to the absence and presence of centromeres, respectively (Lhuillier-Akakpo et al. 2016). Despite this, those centromeric, MIC-specific Histone H3 proteins were absent from the data, and other MIC-specific proteins like NUP98 (Iwamoto et al. 2009) did not differ from other nuclear proteins. While it is unclear why MIC-H3 proteins were absent, the stochastic nature of mass spectrometry-based proteomics results leads to many false negative

protein identifications. The takeaway from this finding was that the resolution of organelles requires not only an appropriate fractionation scheme but also the identification of “diagnostic” proteins to serve as a differentiator of those organelles. Without those proteins, differential pelleting is irrelevant.

Similarly, the contractile vacuole complex (CVC) of *P. tetraurelia* is large and distinct and should be amenable to these techniques, but there was no unique pattern corresponding to its known protein inventory (Plattner 2013). This is likely due to the overlap between most components of the ciliate membrane trafficking system such as phagolysosome and post-Golgi vesicles (Plattner 2022). Most proteins that decorate the CVC are a part of large gene families with members localizing numerous components of the endomembrane system. The shared peptides between members of these gene families will undoubtably affect mass spectrometry-based proteomics identification despite the universality of unique peptides underlying all protein identifications. For example, calmodulin (CAM1) strongly stains the CVC as well as phagolysosomes and components of the cell cortex (Momayezi et al. 1986). Despite this, I show that CAM1’s abundance profile is indicative of a strong cortical role predominating its cell biological functioning due to similarities with axonemal dyneins and tubulins (Figure 3.3). The dynamic nature of GFP-tagging allows for less prominent localizations to be determined, while this approach is weighed towards the localization patterns adopted by most proteins at steady state. As new analytical methods are introduced, these problems around multi-localization will be better resolved (Crook et al. 2018). However, the use of PCpulldown, or some equivalent nearest-neighbor sorting tool, will provide immediate assistance to understanding these promiscuous or otherwise unclear proteins.

Despite the above concerns, these data broadly support the expansion of numerous membrane-bound organelles and protein complexes in *P. tetruarelia*, in particular the mitochondria whose ~1,000 proteins are close to the expectation set by mitochondrial proteomes in other species. Remarkably, many genes predicted here had no orthologs outside of ciliates, which is a finding reminiscent of the rapid mitochondrial divergence of the apicomplexan *T. gondii* after the split between its common ancestor and that of Dinoflagellates (Barylyuk et al. 2020). The

emergence of novel genes in these two Alveolate lineages suggests a broader diversification ongoing in the superphylum as a whole, and more investigation will illuminate the extent to which this is occurring broadly. The verification of mitochondrial glycolysis, suggested previously for the ciliate *T. thermophila*, further opens up new questions about how the metabolic machinery of ciliates— needed to power large, motile cells— is being remodeled through both the relocalization of existing gene products and the emergence of novel genes.

REFERENCES

Allen, Richard D., and Yutaka Naitoh. 2002. “Osmoregulation and Contractile Vacuoles of Protozoa.” In , 351–94. https://doi.org/10.1016/S0074-7696(02)15015-7.

Adoutte, André, Nicole Garreau de Loubresse, and Janine Beisson. 1984. “Proteolytic Cleavage and Maturation of the Crystalline Secretion Products of Paramecium.” *Journal of Molecular Biology* 180 (4): 1065–81.

Aebersold, Ruedi, and Matthias Mann. 2003. “Mass Spectrometry-Based Proteomics.” *Nature* 422 (6928): 198–207. https://doi.org/10.1038/nature01511.

Allen, Richard D. 1988. “Cytology.” In *Paramecium*, 4–40. Springer.

Allen, Richard D, and Agnes K Fok. 1983. “Nonlysosomal Vesicles (Acidosomes) Are Involved in Phagosome Acidification in Paramecium.” *The Journal of Cell Biology* 97 (2): 566–70.

Allen, Richard D, and Agnes K Fokt. 2000. “Membrane Trafficking and Processing in Paramecium.” *International Review of Cytology* 198: 277–318.

Allen, RICHARD D, CHRISTOPHER C Schroeder, and AGNES K Fok. 1992. “Endosomal System of Paramecium: Coated Pits to Early Endosomes.” *Journal of Cell Science* 101 (2): 449–61.

Andersen, Jens S, Christopher J Wilkinson, Thibault Mayor, Peter Mortensen, Erich A Nigg, and Matthias Mann. 2003. “Proteomic Characterization of the Human Centrosome by Protein Correlation Profiling.” *Nature* 426 (6966): 570–74.

Aquilini, Eleonora, Marta Mendonça Cova, Shrawan Kumar Mageswaran, Nicolas dos Santos Pacheco, Daniela Sparvoli, Diana Marcela Penarete-Vargas, Rania Najm, et al. 2021. “An Alveolata Secretory Machinery Adapted to Parasite Host Cell Invasion.” *Nature Microbiology* 6 (4): 425–34. https://doi.org/10.1038/s41564-020-00854-z.

Armenteros, Jose Juan Almagro, Marco Salvatore, Olof Emanuelsson, Ole Winther, Gunnar von Heijne, Arne Elofsson, and Henrik Nielsen. 2019. “Detecting Sequence Signals in Targeting Peptides Using Deep Learning.” *Life Science Alliance* 2 (5).

Arnaiz, Olivier, Jean-François Goût, Mireille Bétermier, Khaled Bouhouche, Jean Cohen, Laurent Duret, Aurélie Kapusta, Eric Meyer, and Linda Sperling. 2010. “Gene Expression in a Paleopolyploid: A Transcriptome Resource for the Ciliate Paramecium Tetraurelia.” *BMC Genomics* 11 (1): 1–13.

Arnaiz, Olivier, Eric Meyer, and Linda Sperling. 2020. “ParameciumDB 2019: Integrating Genomic Data across the Genus for Functional and Evolutionary Biology.” *Nucleic Acids Research* 48 (D1): D599–D605.

Arslan, Taner, Yanbo Pan, Georgios Mermelekas, Mattias Vesterlund, Lukas M Orre, and Janne Lehtiö. 2022. “SubCellBarCode: Integrated Workflow for Robust Spatial Proteomics by Mass Spectrometry.” *Nature Protocols*, 1–41.

Asai, David J, Susan M Beckwith, Kimberly A Kandl, Heather H Keating, Hendri Tjandra, and James D Forney. 1994. “The Dynein Genes of Paramecium Tetraurelia. Sequences Adjacent to the Catalytic P-Loop Identify Cytoplasmic and Axonemal Heavy Chain Isoforms.” *Journal of Cell Science* 107 (4): 839–47.

Aubusson-Fleury, Anne, Geneviève Bricheux, Raghida Damaj, Michel Lemullois, Gérard Coffe, Florence Donnadieu, France Koll, Bernard Viguès, and Philippe Bouchard. 2013. “Epiplasmins and Epiplasm in Paramecium: The Building of a Submembraneous Cytoskeleton.” *Protist* 164 (4): 451–69.

Aury, Jean-Marc, Olivier Jaillon, Laurent Duret, Benjamin Noel, Claire Jubin, Betina M Porcel, Béatrice Ségurens, et al. 2006. “Global Trends of Whole-Genome Duplications Revealed by the Ciliate Paramecium Tetraurelia.” *Nature* 444 (7116): 171–78.

Babst, Markus, Beverly Wendland, Eden J Estepa, and Scott D Emr. 1998. “The Vps4p AAA ATPase Regulates Membrane Association of a Vps Protein Complex Required for Normal Endosome Function.” *The EMBO Journal* 17 (11): 2982–93.

Baers, Laura L., Lisa M. Breckels, Lauren A. Mills, Laurent Gatto, Michael J. Deery, Tim J. Stevens, Christopher J. Howe, Kathryn S. Lilley, and David J. Lea-Smith. 2019. “Proteome Mapping of a Cyanobacterium Reveals Distinct Compartment Organization and Cell-Dispersed Metabolism.” *Plant Physiology* 181 (4): 1721–38. https://doi.org/10.1104/pp.19.00897.

Bailey, Timothy L, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. 2009. “MEME SUITE: Tools for Motif Discovery and Searching.” *Nucleic Acids Research* 37 (suppl_2): W202–W208.

Balabaskaran Nina, Praveen, Natalya v Dudkina, Lesley A Kane, Jennifer E van Eyk, Egbert J Boekema, Michael W Mather, and Akhil B Vaidya. 2010. “Highly Divergent Mitochondrial ATP Synthase Complexes in Tetrahymena Thermophila.” *PLoS Biology* 8 (7): e1000418.

Baranasic, Damir, Timo Oppermann, Miriam Cheaib, John Cullum, Helmut Schmidt, and Martin Simon. 2014. “Genomic Characterization of Variable Surface Antigens Reveals a Telomere Position Effect as a Prerequisite for RNA Interference-Mediated Silencing in Paramecium Tetraurelia.” *MBio* 5 (6): e01328–14.

Barylyuk, Konstantin, Ludek Koreny, Huiling Ke, Simon Butterworth, Oliver M Crook, Imen Lassadi, Vipul Gupta, et al. 2020. “A Comprehensive Subcellular Atlas of the Toxoplasma Proteome via HyperLOPIT Provides Spatial Context for Protein Functions.” *Cell Host & Microbe* 28 (5): 752–66.

Beaver, John R, and Thomas L Crisman. 1989. “The Role of Ciliated Protozoa in Pelagic Freshwater Ecosystems.” *Microbial Ecology* 17 (2): 111–36.

Beisson, Janine, Mireille Bétermier, Marie-Hélène Bré, Jean Cohen, Sandra Duharcourt, Laurent Duret, Ching Kung, et al. 2010. “Paramecium Tetraurelia: The Renaissance of an Early Unicellular Model.” *Cold Spring Harbor Protocols* 2010 (1): pdb–emo140.

Beisson, Janine, and Tracy M Sonneborn. 1965. “Cytoplasmic Inheritance of the Organization of the Cell Cortex in Paramecium Aurelia.” *Proceedings of the National Academy of Sciences* 53 (2): 275–82.

Birchler, James A, Utpal Bhadra, Manika Pal Bhadra, and Donald L Auger. 2001. “Dosage-Dependent Gene Regulation in Multicellular Eukaryotes: Implications for Dosage Compensation, Aneuploid Syndromes, and Quantitative Traits.” *Developmental Biology* 234 (2): 275–88.

Borgese, Nica, Sara Colombo, and Emanuela Pedrazzini. 2003. “The Tale of Tail-Anchored Proteins.” *Journal of Cell Biology* 161 (6): 1013–19. https://doi.org/10.1083/jcb.200303069.

Borner, Georg H.H. 2020. “Organellar Maps Through Proteomic Profiling – A Conceptual Guide.” *Molecular & Cellular Proteomics* 19 (7): 1076–87. https://doi.org/10.1074/mcp.R120.001971.

Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. “A Training Algorithm for Optimal Margin Classifiers.” In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory - COLT ’92*, 144–52. New York, New York, USA: ACM Press. https://doi.org/10.1145/130385.130401.

Brandina, Irina, James Graham, Christelle Lemaitre-Guillier, Nina Entelis, Igor Krasheninnikov, Lee Sweetlove, Ivan Tarassov, and Robert P Martin. 2006. “Enolase Takes Part in a Macromolecular Complex Associated to Mitochondria in Yeast.” *Biochimica et Biophysica Acta (BBA)-Bioenergetics* 1757 (9–10): 1217–28.

Breckels, Lisa M, Claire M Mulvey, Kathryn S Lilley, and Laurent Gatto. 2016. “A Bioconductor Workflow for Processing and Analysing Spatial Proteomics Data.” *F1000Research* 5.

Breuer, Marion, Gerald Schulte, Klaus J. Schwegmann, and Helmut J. Schmidt. 1996. “Molecular Characterization of the D Surface Protein Gene Subfamily in Paramecium Tetraurelia.” *The Journal of Eukaryotic Microbiology* 43 (4): 314–22. https://doi.org/10.1111/j.1550-7408.1996.tb03994.x.

Bright, Lydia J, Jean-Francois Gout, and Michael Lynch. 2017. “Early Stages of Functional Diversification in the Rab GTPase Gene Family Revealed by Genomic and Localization Studies in Paramecium Species.” *Molecular Biology of the Cell* 28 (8): 1101–10.

Byun, S Ashley, and Sarabdeep Singh. 2013. “Protein Subcellular Relocalization Increases the Retention of Eukaryotic Duplicate Genes.” *Genome Biology and Evolution* 5 (12): 2402–9.

Callister, Stephen J., Richard C. Barry, Joshua N. Adkins, Ethan T. Johnson, Wei-jun Qian, Bobbie-Jo M. Webb-Robertson, Richard D. Smith, and Mary S. Lipton. 2006. “Normalization Approaches for Removing Systematic Biases Associated with Mass Spectrometry and Label-Free Proteomics.” *Journal of Proteome Research* 5 (2): 277–86. https://doi.org/10.1021/pr050300l.

Caron, François, and Eric Meyer. 1985. “Does Paramecium Primaurelia Use a Different Genetic Code in Its Macronucleus?” *Nature* 314 (6007): 185–88.

Catania, Francesco, Casey L. McGrath, Thomas G. Doak, and Michael Lynch. 2013. “Spliced DNA Sequences in the Paramecium Germline: Their Properties and Evolutionary Potential.” *Genome Biology and Evolution* 5 (6): 1200–1211. https://doi.org/10.1093/gbe/evt087.

Cech, Thomas R, Arthur J Zaug, and Paula J Grabowski. 1981. “In Vitro Splicing of the Ribosomal RNA Precursor of Tetrahymena: Involvement of a Guanosine Nucleotide in the Excision of the Intervening Sequence.” *Cell* 27 (3): 487–96.

Chan, Catherine W M, Yoshiro Saimi, and Ching Kung. 1999. “A New Multigene Family Encoding Calcium-Dependent Calmodulin-Binding Membrane Proteins of Paramecium Tetraurelia.” *Gene* 231 (1–2): 21–32.

Chatterjee, Kunal, Regina T. Nostramo, Yao Wan, and Anita K. Hopper. 2018. “TRNA Dynamics between the Nucleus, Cytoplasm and Mitochondrial Surface: Location, Location, Location.”

Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1861 (4): 373–86. https://doi.org/10.1016/j.bbagrm.2017.11.007.

Christoforou, Andy, Claire M. Mulvey, Lisa M. Breckels, Aikaterini Geladaki, Tracey Hurrell, Penelope C. Hayward, Thomas Naake, et al. 2016. “A Draft Map of the Mouse Pluripotent Stem Cell Spatial Proteome.” *Nature Communications* 7 (1): 9992. https://doi.org/10.1038/ncomms9992.

Christopher, Josie A, Charlotte Stadler, Claire E Martin, Marcel Morgenstern, Yanbo Pan, Cora N Betsinger, David G Rattray, et al. 2021. “Subcellular Proteomics.” *Nature Reviews Methods Primers* 1 (1): 1–24.

Claude, Albert. 1975. “The Coming of Age of the Cell.” *Science* 189 (4201): 433–35. https://doi.org/10.1126/science.1098146.

Colombo, Maria I, Walter Beron, and Philip D Stahl. 1997. “Calmodulin Regulates Endosome Fusion.” *Journal of Biological Chemistry* 272 (12): 7707–12.

Corpas, Francisco J, Juan B Barroso, Luisa M Sandalio, José M Palma, José A Lupiánez, and Luis A del Ro. 1999. “Peroxisomal NADP-Dependent Isocitrate Dehydrogenase. Characterization and Activity Regulation during Natural Senescence.” *Plant Physiology* 121 (3): 921–28.

Crook, Oliver M, Lisa M Breckels, Kathryn S Lilley, Paul D W Kirk, and Laurent Gatto. 2019. “A Bioconductor Workflow for the Bayesian Analysis of Spatial Proteomics.” *F1000Research* 8.

Crook, Oliver M, Claire M Mulvey, Paul D W Kirk, Kathryn S Lilley, and Laurent Gatto. 2018. “A Bayesian Mixture Modelling Approach for Spatial Proteomics.” *PLoS Computational Biology* 14 (11): e1006516.

Cummings, Donald J. 1977. “Methods for the Isolation of Nuclei from Ciliated Protozoans.” In *Methods in Cell Biology*, 16:97–112. Elsevier.

Dehal, Paramvir, and Jeffrey L Boore. 2005. “Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate.” *PLoS Biology* 3 (10): e314. https://doi.org/10.1371/journal.pbio.0030314.

Diffley, J F, and Bruce Stillman. 1991. “A Close Relative of the Nuclear, Chromosomal High-Mobility Group Protein HMG1 in Yeast Mitochondria.” *Proceedings of the National Academy of Sciences* 88 (17): 7864–68.

Dosil, Mercedes, and Xosé R Bustelo. 2004. “Functional Characterization of Pwp2, a WD Family Protein Essential for the Assembly of the 90 S Pre-Ribosomal Particle*?” *Journal of Biological Chemistry* 279 (36): 37385–97.

Drews, Franziska, Abdulrahman Salhab, Sivarajan Karunanithi, Miriam Cheaib, Martin Jung, Marcel H Schulz, and Martin Simon. 2022. “Broad Domains of Histone Marks in the Highly Compact Paramecium Macronuclear Genome.” *Genome Research* 32 (4): 710–25.

Dudkina, N v, L A Kane, J E van Eyk, E J Boekema, M W Mather, A B Vaidya, and others. 2010. “Highly Divergent Mitochondrial ATP Synthase Complexes in Tetrahymena Thermophila.” *PLos Biology* 8 (7): e1000418–e1000418.

Duharcourt, Sandra, Gersende Lepère, and Eric Meyer. 2009. “Developmental Genome Rearrangements in Ciliates: A Natural Genomic Subtraction Mediated by Non-Coding Transcripts.” *Trends in Genetics* 25 (8): 344–50.

Dunham, Wade H, Michael Mullin, and Anne-Claude Gingras. 2012. “Affinity-Purification Coupled to Mass Spectrometry: Basic Principles and Strategies.” *Proteomics* 12 (10): 1576–90.

Dunkley, T.P.J., R. Watson, J.L. Griffin, P. Dupree, and K.S. Lilley. 2004. “Localization of Organelle Proteins by Isotope Tagging (LOPIT).” *Molecular & Cellular Proteomics* 3 (11): 1128–34. https://doi.org/10.1074/mcp.T400009-MCP200.

Dute, Roland, and Ching Kung. 1978. “Ultrastructure of the Proximal Region of Somatic Cilia in Paramecium Tetraurelia.” *The Journal of Cell Biology* 78 (2): 451–64.

Duve, C de. 1969. “The Peroxisome: A New Cytoplasmic Organelle.” *Proceedings of the Royal Society of London. Series B. Biological Sciences* 173 (1030): 71–83.

Duve, Christian de, B C Pressman, R Gianetto, R Wattiaux, and Françoise Appelmans. 1955. “Tissue Fractionation Studies. 6. Intracellular Distribution Patterns of Enzymes in Rat-Liver Tissue.” *Biochemical Journal* 60 (4): 604.

Eisen, Jonathan A, Robert S Coyne, Martin Wu, Dongying Wu, Mathangi Thiagarajan, Jennifer R Wortman, Jonathan H Badger, et al. 2006. “Macronuclear Genome Sequence of the Ciliate Tetrahymena Thermophila, a Model Eukaryote.” *PLoS Biology* 4 (9): e286. https://doi.org/10.1371/journal.pbio.0040286.

Fenn, John B., Matthias Mann, Chin Kai Meng, Shek Fu Wong, and Craig M. Whitehouse. 1989. “Electrospray Ionization for Mass Spectrometry of Large Biomolecules.” *Science* 246 (4926): 64–71.

Fiedler, Klaus, and Kai Simons. 1995. “The Role of N-Glycans in the Secretory Pathway.” *Cell* 81 (3): 309–12.

Fok, Agnes K, and Richard D Allen. 1990. “The Phagosome-Lysosome Membrane System and Its Regulation in Paramecium.” *International Review of Cytology* 123: 61–94.

Fok, Agnes K and Richard D Allen. 1998. “The Lysosome System.” In *Paramecium*, 301–24. Springer.

Fok, Agnes K, Yeng Lee, and Richard D Allen. 1982. “The Correlation of Digestive Vacuole PH and Size with the Digestive Cycle in Paramecium Caudatum 1.” *The Journal of Protozoology* 29 (3): 409–14.

Force, Allan, Michael Lynch, F Bryan Pickett, Angel Amores, Yi-lin Yan, and John Postlethwait. 1999. “Preservation of Duplicate Genes by Complementary, Degenerative Mutations.” *Genetics* 151 (4): 1531–45. https://doi.org/10.1093/genetics/151.4.1531.

Foster, Leonard J, Carmen L de Hoog, Yanling Zhang, Yong Zhang, Xiaohui Xie, Vamsi K Mootha, and Matthias Mann. 2006. “A Mammalian Organelle Map by Protein Correlation Profiling.” *Cell* 125 (1): 187–99.

Fraga, D., I. M. Sehring, R. Kissmehl, M. Reiss, R. Gaines, R. Hinrichsen, and H. Plattner. 2010. “Protein Phosphatase 2B (PP2B, Calcineurin) in Paramecium: Partial Characterization Reveals That Two Members of the Unusually Large Catalytic Subunit Family Have Distinct

Roles in Calcium-Dependent Processes." *Eukaryotic Cell* 9 (7): 1049–63. https://doi.org/10.1128/EC.00322-09.

Galvani, Angélique, and Linda Sperling. 2002. "RNA Interference by Feeding in Paramecium." *Trends in Genetics* 18 (1): 11–12. https://doi.org/10.1016/S0168-9525(01)02548-3.

Garreau De Loubresse, Nicole, Guy Keryer, Bernard Viguès, and JANINE BEISSON. 1988. "A Contractile Cytoskeletal Network of Paramecium: The Infraciliary Lattice." *Journal of Cell Science* 90 (3): 351–64.

Gatto, Laurent, Lisa M Breckels, Thomas Burger, Daniel J H Nightingale, Arnoud J Groen, Callum Campbell, Nino Nikolovski, et al. 2014. "A Foundation for Reliable Spatial Proteomics Data Analysis." *Molecular & Cellular Proteomics* 13 (8): 1937–52.

Gatto, Laurent, Lisa M Breckels, and Kathryn S Lilley. 2019. "Assessing Sub-Cellular Resolution in Spatial Proteomics Experiments." *Current Opinion in Chemical Biology* 48: 123–49.

Gatto, Laurent, Lisa M. Breckels, Thomas Naake, and Sebastian Gibb. 2015. "Visualization of Proteomics Data Using R and Bioconductor." *PROTEOMICS* 15 (8): 1375–89. https://doi.org/10.1002/pmic.201400392.

Gatto, Laurent, Lisa M Breckels, Samuel Wieczorek, Thomas Burger, and Kathryn S Lilley. 2014. "Mass-Spectrometry-Based Spatial Proteomics Data Analysis Using PRoloc and PRolocdata." *Bioinformatics* 30 (9): 1322–24.

Gatto, Laurent, and Kathryn S Lilley. 2012. "MSnbase-an R/Bioconductor Package for Isobaric Tagged Mass Spectrometry Data Visualization, Processing and Quantitation." *Bioinformatics* 28 (2): 288–89.

Geladaki, Aikaterini, Nina Kočevar Britovšek, Lisa M Breckels, Tom S Smith, Owen L Vennard, Claire M Mulvey, Oliver M Crook, Laurent Gatto, and Kathryn S Lilley. 2019. "Combining LOPIT with Differential Ultracentrifugation for High-Resolution Spatial Proteomics." *Nature Communications* 10 (1): 1–15.

Gemmer, Max, and Friedrich Förster. 2020. "A Clearer Picture of the ER Translocon Complex." *Journal of Cell Science* 133 (3): jcs231340.

Giegé, Philippe, Joshua L Heazlewood, Ute Roessner-Tunali, A Harvey Millar, Alisdair R Fernie, Christopher J Leaver, and Lee J Sweetlove. 2003. "Enzymes of Glycolysis Are Functionally Associated with the Mitochondrion in Arabidopsis Cells." *The Plant Cell* 15 (9): 2140–51.

Gogendeau, Delphine, Catherine Klotz, Olivier Arnaiz, Agata Malinowska, Michal Dadlez, Nicole Garreau de Loubresse, Françoise Ruiz, France Koll, and Janine Beisson. 2008. "Functional Diversification of Centrins and Cell Morphological Complexity." *Journal of Cell Science* 121 (1): 65–74.

Gogendeau, Delphine, Michel Lemullois, Pierrick le Borgne, Manon Castelli, Anne Aubusson-Fleury, Olivier Arnaiz, Jean Cohen, et al. 2020. "MKS-NPHP Module Proteins Control Ciliary Shedding at the Transition Zone." *PLoS Biology* 18 (3): e3000640.

Gould, S. B., L. G. K. Kraft, G. G. van Dooren, C. D. Goodman, K. L. Ford, A. M. Cassin, A. Bacic, G. I. McFadden, and R. F. Waller. 2011. "Ciliate Pellicular Proteome Identifies Novel Protein Families with Characteristic Repeat Motifs That Are Common to Alveolates." *Molecular Biology and Evolution* 28 (3): 1319–31. https://doi.org/10.1093/molbev/msq321.

Gould, Sven B, Wai-Hong Tham, Alan F Cowman, Geoffrey I McFadden, and Ross F Waller. 2008. “Alveolins, a New Family of Cortical Proteins That Define the Protist Infrakingdom Alveolata.” *Molecular Biology and Evolution* 25 (6): 1219–30.

Gout, Jean-Francois, Parul Johri, Olivier Arnaiz, Thomas G Doak, Simran Bhullar, Arnaud Couloux, Fréderic Guérin, et al. 2019. “Universal Trends of Post-Duplication Evolution Revealed by the Genomes of 13 Paramecium Species Sharing an Ancestral Whole-Genome Duplication.” *BioRXiv*, 573576.

Gout, Jean-François, Daniel Kahn, Laurent Duret, and Paramecium Post-Genomics Consortium. 2010. “The Relationship among Gene Expression, the Evolution of Gene Dosage, and the Rate of Protein Evolution.” *PLoS Genetics* 6 (5): e1000944.

Gout, Jean-Francois, and Michael Lynch. 2015. “Maintenance and Loss of Duplicated Genes by Dosage Subfunctionalization.” *Molecular Biology and Evolution* 32 (8): 2141–48. https://doi.org/10.1093/molbev/msv095.

Greider, Carol W, and Elizabeth H Blackburn. 1985. “Identification of a Specific Telomere Terminal Transferase Activity in Tetrahymena Extracts.” *Cell* 43 (2): 405–13.

Guan, Yuanfang, Maitreya J Dunham, and Olga G Troyanskaya. 2007. “Functional Analysis of Gene Duplications in Saccharomyces Cerevisiae.” *Genetics* 175 (2): 933–43.

Gubbels, Marc-Jan, and Manoj T. Duraisingh. 2012. “Evolution of Apicomplexan Secretory Organelles.” *International Journal for Parasitology* 42 (12): 1071–81. https://doi.org/10.1016/j.ijpara.2012.09.009.

Guerrier, Sabrice, Helmut Plattner, Elisabeth Richardson, Joel B Dacks, and Aaron P Turkewitz. 2017. “An Evolutionary Balance: Conservation vs Innovation in Ciliate Membrane Trafficking.” *Traffic* 18 (1): 18–28.

Hartl, Franz-Ulrich, Nikolaus Pfanner, Donald W Nicholson, and Walter Neupert. 1989. “Mitochondrial Protein Import.” *Biochimica et Biophysica Acta (BBA)-Reviews on Biomembranes* 988 (1): 1–45.

Hauser, Karin, W John Haynes, Ching Kung, Helmut Plattner, and Roland Kissmehl. 2000a. “Expression of the Green Fluorescent Protein in Paramecium Tetraurelia.” *European Journal of Cell Biology* 79 (2): 144–49.

Hauser, Karin, W.John Haynes, Ching Kung, Helmut Plattner, and Roland Kissmehl. 2000b. “Expression of the Green Fluorescent Protein in Paramecium Tetraurelia.” *European Journal of Cell Biology* 79 (2): 144–49. https://doi.org/10.1078/S0171-9335(04)70016-3.

Hauser, Karin, Nada Pavlovic, Norbert Klauke, Deisy Geissinger, and Helmut Plattner. 2000. “Green Fluorescent Protein-Tagged Sarco (Endo) Plasmic Reticulum Ca2+-ATPase Overexpression in Paramecium Cells: Isoforms, Subcellular Localization, Biogenesis of Cortical Calcium Stores and Functional Aspects.” *Molecular Microbiology* 37 (4): 773–87.

Hausmann, Klaus, Phyllis Clarke Bradbury, and others. 1996. *Ciliates: Cells as Organisms*. Gustav Fischer Stuttgart.

Holtzman, Eric. 2013. *Lysosomes*. Springer Science & Business Media.

Huh, Won-Ki, James v Falvo, Luke C Gerke, Adam S Carroll, Russell W Howson, Jonathan S Weissman, and Erin K O’Shea. 2003. “Global Analysis of Protein Localization in Budding Yeast.” *Nature* 425 (6959): 686–91.

Huynh, Kassidy K, Eeva-Liisa Eskelinen, Cameron C Scott, Anatoly Malevanets, Paul Saftig, and Sergio Grinstein. 2007. “LAMP Proteins Are Required for Fusion of Lysosomes with Phagosomes.” *The EMBO Journal* 26 (2): 313–24. https://doi.org/10.1038/sj.emboj.7601511.

Innan, Hideki, and Fyodor Kondrashov. 2010. “The Evolution of Gene Duplications: Classifying and Distinguishing between Models.” *Nature Reviews Genetics* 11 (2): 97–108.

Ishida, Masaki, Manabu Hori, Yui Ooba, Masako Kinoshita, Tsuyoshi Matsutani, Musumi Naito, Taeko Hagimoto, et al. 2021. “A Functional *Aqp1* Gene Product Localizes on The Contractile Vacuole Complex in *Paramecium Multimicronucleatum*.” *Journal of Eukaryotic Microbiology* 68 (3). https://doi.org/10.1111/jeu.12843.

Ishikawa, Takashi. 2017. “Axoneme Structure from Motile Cilia.” *Cold Spring Harbor Perspectives in Biology* 9 (1): a028076.

Itzhak, Daniel N, Stefka Tyanova, Jürgen Cox, and Georg H H Borner. 2016. “Global, Quantitative and Dynamic Mapping of Protein Subcellular Localization.” *Elife* 5: e16950.

Iwamoto, Masaaki, Chie Mori, Tomoko Kojidani, Fumihide Bunai, Tetsuya Hori, Tatsuo Fukagawa, Yasushi Hiraoka, and Tokuko Haraguchi. 2009. “Two Distinct Repeat Sequences of Nup98 Nucleoporins Characterize Dual Nuclei in the Binucleated Ciliate Tetrahymena.” *Current Biology* 19 (10): 843–47.

Jadot, Michel, Marielle Boonen, Jaqueline Thirion, Nan Wang, Jinchuan Xing, Caifeng Zhao, Abla Tannous, et al. 2017. “Accounting for Protein Subcellular Localization: A Compartmental Map of the Rat Liver Proteome.” *Molecular & Cellular Proteomics* 16 (2): 194–212. https://doi.org/10.1074/mcp.M116.064527.

Jennings, Herbert Spencer. 1906. *Behavior of the Lower Organisms*. Columbia University Press, The Macmillan Company, agents.

Jerka-Dziadosz, Maria, Delphine Gogendeau, Catherine Klotz, Jean Cohen, Janine Beisson, and France Koll. 2010. “Basal Body Duplication in Paramecium: The Key Role of Bld10 in Assembly and Stability of the Cartwheel.” *Cytoskeleton* 67 (3): 161–71.

Kandl, K A, J D Forney, and D J Asai. 1995. “The Dynein Genes of Paramecium Tetraurelia: The Structure and Expression of the Ciliary Beta and Cytoplasmic Heavy Chains.” *Molecular Biology of the Cell* 6 (11): 1549–62.

Kaur, Harpreet, Elisabeth Richardson, Komal Kamra, and Joel B. Dacks. 2022. “Molecular Evolutionary Analysis of the SM and SNARE Vesicle Fusion Machinery in Ciliates Shows Concurrent Expansions in Late Secretory Machinery.” *Journal of Eukaryotic Microbiology* 69 (4). https://doi.org/10.1111/jeu.12919.

Kee, H Lynn, and Kristen J Verhey. 2013. “Molecular Connections between Nuclear and Ciliary Import Processes.” *Cilia* 2 (1): 1–10.

Keller, Lani C, Stefan Geimer, Edwin Romijn, John Yates III, Ivan Zamora, and Wallace F Marshall. 2009. “Molecular Architecture of the Centriole Proteome: The Conserved WD40 Domain Protein POC1 Is Required for Centriole Duplication and Length Control.” *Molecular Biology of the Cell* 20 (4): 1150–66.

Kingsland, Sharon. 1987. “A Man out of Place: Herbert Spencer Jennings at Johns Hopkins, 1906–1938.” *American Zoologist* 27 (3): 807–17.

Kissmehl, Roland, Marine Froissard, Helmut Plattner, Massoud Momayezi, and Jean Cohen. 2002. “NSF Regulates Membrane Traffic along Multiple Pathways in Paramecium.” *Journal of Cell Science* 115 (20): 3935–46.

Kissmehl, Roland, Christina Schilde, Thomas Wassmer, Carsten Danzer, Kathrin Nuehse, Kaya Lutter, and Helmut Plattner. 2007. “Molecular Identification of 26 Syntaxin Genes and Their Assignment to the Different Trafficking Pathways in Paramecium.” *Traffic* 8 (5): 523–42.

Kornmann, Benoît, Christof Osman, and Peter Walter. 2011. “The Conserved GTPase Gem1 Regulates Endoplasmic Reticulum–Mitochondria Connections.” *Proceedings of the National Academy of Sciences* 108 (34): 14151–56. https://doi.org/10.1073/pnas.1111314108.

Krahmer, Natalie, Bahar Najafi, Florian Schueder, Fabiana Quagliarini, Martin Steger, Susanne Seitz, Robert Kasper, et al. 2018. “Organellar Proteomics and Phospho-Proteomics Reveal Subcellular Reorganization in Diet-Induced Hepatic Steatosis.” *Developmental Cell* 47 (2): 205-221.e7.

Kranz, Christian, Jonas Denecke, Mark A Lehrman, Sutapa Ray, Petra Kienz, Gunilla Kreissel, Dijana Sagi, et al. 2001. “A Mutation in the Human MPDU1 Gene Causes Congenital Disorder of Glycosylation Type If (CDG-If).” *The Journal of Clinical Investigation* 108 (11): 1613–19.

Krueger, Vivienne, Daaé Ransom, Abbie Williams, Emily Carson, Byunghyun Ahn, Kim Kandl, and Laura Listenberger. 2022. “Unraveling the Roles of Mitochondria and Peroxisomes in Lipid Droplet Utilization in Tetrahymena Thermophila.” *The FASEB Journal* 36.

Kruger, Kelly, Paula J Grabowski, Arthur J Zaug, Julie Sands, Daniel E Gottschling, and Thomas R Cech. 1982. “Self-Splicing RNA: Autoexcision and Autocyclization of the Ribosomal RNA Intervening Sequence of Tetrahymena.” *Cell* 31 (1): 147–57.

Kumar, Sudhir, Koichiro Tamura, and Masatoshi Nei. 1994. “MEGA: Molecular Evolutionary Genetics Analysis Software for Microcomputers.” *Bioinformatics* 10 (2): 189–91.

Kuppannan, Aarthi, Yu-Yang Jiang, Wolfgang Maier, Chang Liu, Charles F Lang, Chao-Yin Cheng, Mark C Field, Minglei Zhao, Martin Zoltner, and Aaron P Turkewitz. 2022. “A Novel Membrane Complex Is Required for Docking and Regulated Exocytosis of Lysosome-Related Organelles in Tetrahymena Thermophila.” *PLoS Genetics* 18 (5): e1010194.

Ladenburger, Eva-Maria, and Helmut Plattner. 2011. “Calcium-Release Channels in Paramecium. Genomic Expansion, Differential Positioning and Partial Transcriptional Elimination.” *PLoS ONE* 6 (11): e27111. https://doi.org/10.1371/journal.pone.0027111.

Laligné, C, C Klotz, N de Loubresse, M Lemullois, M Hori, F X Laurent, J F Papon, B Louis, J Cohen, and F Koll. 2010. “Bug22p, a Conserved Centrosomal/Ciliary Protein Also Present in Higher Plants, Is Required for an Effective Ciliary Stroke in Paramecium.” *Eukaryotic Cell* 9 (4): 645–55.

Lhuillier-Akakpo, Maoussi, Frédéric Guérin, Andrea Frapporti, and Sandra Duharcourt. 2016. “DNA Deletion as a Mechanism for Developmentally Programmed Centromere Loss.” *Nucleic Acids Research* 44 (4): 1553–65.

Li, Weizhong, and Adam Godzik. 2006. “Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences.” *Bioinformatics* 22 (13): 1658–59.

Likas, Aristidis, Nikos Vlassis, and Jakob J Verbeek. 2003. “The Global K-Means Clustering Algorithm.” *Pattern Recognition* 36 (2): 451–61.

Liu, Yufeng, Hao Helen Zhang, and Yichao Wu. 2011. “Hard or Soft Classification? Large-Margin Unified Machines.” *Journal of the American Statistical Association* 106 (493): 166–77. https://doi.org/10.1198/jasa.2011.tm10319.

Lundberg, Emma, and Georg H H Borner. 2019. “Spatial Proteomics: A Powerful Discovery Tool for Cell Biology.” *Nature Reviews Molecular Cell Biology* 20 (5): 285–302.

Lynch, Michael, and Allan Force. 2000. “The Probability of Duplicate Gene Preservation by Subfunctionalization.” *Genetics* 154 (1): 459–73.

Lynch, Michael, Paul E Schavemaker, Timothy J Licknack, Yue Hao, and Arianna Pezzano. 2022a. “Evolutionary Bioenergetics of Ciliates.” *Journal of Eukaryotic Microbiology*, e12934.

Lynn, Denis H. 1981. “The Organization and Evolution of Microtubular Organelles in Ciliated Protozoa.” *Biological Reviews* 56 (2): 243–92.

Lynn, Denis H. 2008. “The Ciliated Protozoa: Characterization, Classification, and Guide to the Literature.”

Madeddu, Luisa, Marie-Christine Gautier, L Vayssié, A Houari, and L Sperling. 1995. “A Large Multigene Family Codes for the Polypeptides of the Crystalline Trichocyst Matrix in Paramecium.” *Molecular Biology of the Cell* 6 (6): 649–59.

Mann, Matthias. 2020. “The Origins of Organellar Mapping by Protein Correlation Profiling.” *Proteomics* 20 (23): 1900330.

Maurer-Alcalá, Xyrus X, and Mariusz Nowacki. 2019. “Evolutionary Origins and Impacts of Genome Architecture in Ciliates.” *Annals of the New York Academy of Sciences* 1447 (1): 110–18.

McGrath, Casey L, Jean-Francois Gout, Parul Johri, Thomas G Doak, and Michael Lynch. 2014. “Differential Retention and Divergent Resolution of Duplicate Genes Following Whole-Genome Duplication.” *Genome Research* 24 (10): 1665–75.

Means, Anthony R, and John R Dedman. 1980. “Calmodulin—an Intracellular Calcium Receptor.” *Nature* 285 (5760): 73–77.

Momayezi, M, H Kersken, U Gras, J Vilmart-Seuwen, and H Plattner. 1986. “Calmodulin in Paramecium Tetraurelia: Localization from the in Vivo to the Ultrastructural Level.” *Journal of Histochemistry & Cytochemistry* 34 (12): 1621–38.

Müller, Miklós, James F Hogg, and Christian de Duve. 1968. “Distribution of Tricarboxylic Acid Cycle Enzymes and Glyoxylate Cycle Enzymes between Mitochondria and Peroxisomes in Tetrahymena Pyriformis.” *Journal of Biological Chemistry* 243 (20): 5385–95.

Mulvey, Claire M., Lisa M. Breckels, Oliver M. Crook, David J. Sanders, Andre L. R. Ribeiro, Aikaterini Geladaki, Andy Christoforou, et al. 2021. “Spatiotemporal Proteomic Profiling of the Pro-Inflammatory Response to Lipopolysaccharide in the THP-1 Human Leukaemia Cell Line.” *Nature Communications* 12 (1): 5773. https://doi.org/10.1038/s41467-021-26000-9.

Nabi, Ashikun, Junji Yano, Megan S Valentine, Tyler Picariello, and Judith L van Houten. 2019. “SF-Assemblin Genes in Paramecium: Phylogeny and Phenotypes of RNAi Silencing on the Ciliary-Striated Rootlets and Surface Organization.” *Cilia* 8 (1): 1–21.

Nguyen Ba, Alex N, Anastassia Pogoutse, Nicholas Provart, and Alan M Moses. 2009. “NLStradamus: A Simple Hidden Markov Model for Nuclear Localization Signal Prediction.” *BMC Bioinformatics* 10 (1): 1–11.

Nightingale, Daniel J H, Stephen G Oliver, and Kathryn S Lilley. 2019. “Mapping the Saccharomyces Cerevisiae Spatial Proteome with High Resolution Using HyperLOPIT.” In *Yeast Systems Biology*, 165–90. Springer.

Nowacki, Mariusz, Keerthi Shetty, and Laura F Landweber. 2011. “RNA-Mediated Epigenetic Programming of Genome Rearrangements.” *Annual Review of Genomics and Human Genetics* 12: 367.

Ohno, Susumu. 1970. *Evolution by Gene Duplication*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-86659-3.

Ong, S. 2003. “Mass Spectrometric-Based Approaches in Quantitative Proteomics.” *Methods* 29 (2): 124–30. https://doi.org/10.1016/S1046-2023(02)00303-1.

Orsburn, Benjamin C. 2021. “Proteome Discoverer—A Community Enhanced Data Processing Suite for Protein Informatics.” *Proteomes* 9 (1): 15.

Owa, Mikito, Takayuki Uchihashi, Haru-aki Yanagisawa, Takashi Yamano, Hiro Iguchi, Hideya Fukuzawa, Ken-ichi Wakabayashi, Toshio Ando, and Masahide Kikkawa. 2019. “Inner Lumen Proteins Stabilize Doublet Microtubules in Cilia and Flagella.” *Nature Communications* 10 (1): 1–10.

Owsian, Dawid, Julita Gruchota, Olivier Arnaiz, and Jacek K Nowak. 2022. “The Transient Spt4-Spt5 Complex as an Upstream Regulator of Non-Coding RNAs during Development.” *Nucleic Acids Research* 50 (5): 2603–20.

Pelham, Hugh R.B. 1990. “The Retention Signal for Soluble Proteins of the Endoplasmic Reticulum.” *Trends in Biochemical Sciences* 15 (12): 483–86. https://doi.org/10.1016/0968-0004(90)90303-S.

Plattner, Helmut. 2010. “Membrane Trafficking in Protozoa: SNARE Proteins, H+-ATPase, Actin, and Other Key Players in Ciliates.” *International Review of Cell and Molecular Biology* 280: 79–184.

Plattner, Helmut. 2013. “Contractile Vacuole Complex: Its Expanding Protein Inventory.” *International Review of Cell and Molecular Biology* 306: 371–416.

Plattner, Helmut. 2017. “Trichocysts: Paramecium’s Projectile-like Secretory Organelles: Reappraisal of Their Biogenesis, Composition, Intracellular Transport, and Possible Functions.” *Journal of Eukaryotic Microbiology* 64 (1): 106–33.

Plattner, Helmut. 2018. “Evolutionary Cell Biology of Proteins from Protists to Humans and Plants.” *Journal of Eukaryotic Microbiology* 65 (2): 255–89.

Plattner, Helmut. 2020. “Secretory Mechanisms in Paramecium.” In *Neurosecretion: Secretory Mechanisms*, 271–90. Springer.

Plattner, Helmut. 2022. “Membrane Traffic and Ca2+ Signals in Ciliates.” *Journal of Eukaryotic Microbiology*, e12895.

Plattner, Helmut, and Roland Kissmehl. 2003. “Molecular Aspects of Membrane Trafficking in Paramecium.” *International Review of Cytology* 232: 185–216.

Powers, E L, C F Ehret, and L E Roth. 1955. “Mitochondrial Structure in Paramecium as Revealed by Electron Microscopy.” *The Biological Bulletin* 108 (2): 182–95.

Preer Jr, John R. 1986. “Surface Antigens of Paramecium.” *The Molecular Biology of Ciliated Protozoa*, 301–39.

Prescott, David M. 1994. “The DNA of Ciliated Protozoa.” *Microbiological Reviews* 58 (2): 233–67.

Reuter, Alexander T, Claudia A O Stuermer, and Helmut Plattner. 2013. “Identification, Localization, and Functional Implications of the Microdomain-Forming Stomatin Family in the Ciliated Protozoan Paramecium Tetraurelia.” *Eukaryotic Cell* 12 (4): 529–44.

Richardson, Elisabeth, and Joel B. Dacks. 2022. “Distribution of Membrane Trafficking System Components across Ciliate Diversity Highlights Heterogenous organelle-associated Machinery.” *Traffic*, March. https://doi.org/10.1111/tra.12834.

Richter, Daniel J, Cédric Berney, Jürgen F H Strassert, Yu-Ping Poh, Emily K Herman, Sergio A Muñoz-Gómez, Jeremy G Wideman, Fabien Burki, and Colomban de Vargas. 2022. “EukProt: A Database of Genome-Scale Predicted Proteins across the Diversity of Eukaryotes.” *Peer Community Journal* 2.

Rio Bártulos, Carolina, Matthew B Rogers, Tom A Williams, Eleni Gentekaki, Henner Brinkmann, Rüdiger Cerff, Marie-Françoise Liaud, et al. 2018. “Mitochondrial Glycolysis in a Major Lineage of Eukaryotes.” *Genome Biology and Evolution* 10 (9): 2310–25.

Rombel, Irene T, Kathryn F Sykes, Simon Rayner, and Stephen Albert Johnston. 2002. “ORF-FINDER: A Vector for High-Throughput Gene Identification.” *Gene* 282 (1–2): 33–41.

Rosati, Giovanna, and Letizia Modeo. 2003. “Extrusomes in Ciliates: Diversification, Distribution, and Phylogenetic Implications.” *Journal of Eukaryotic Microbiology* 50 (6): 383–402.

Roux, Kyle J, Dae In Kim, Manfred Raida, and Brian Burke. 2012. “A Promiscuous Biotin Ligase Fusion Protein Identifies Proximal and Interacting Proteins in Mammalian Cells.” *Journal of Cell Biology* 196 (6): 801–10.

Ruiz, Françoise, Nicole Garreau de Loubresse, Catherine Klotz, Janine Beisson, and France Koll. 2005. “Centrin Deficiency in Paramecium Affects the Geometry of Basal-Body Duplication.” *Current Biology* 15 (23): 2097–2106.

Sabatini, David D., and Milton Adesnik. 2013. “Christian de Duve: Explorer of the Cell Who Discovered New Organelles by Using a Centrifuge.” *Proceedings of the National Academy of Sciences* 110 (33): 13234–35.

Samaranayake, Haresha S, Ann E Cowan, and Lawrence A Klobutcher. 2011. “Vacuolar Protein Sorting Protein 13A, TtVPS13A, Localizes to the Tetrahymena Thermophila Phagosome Membrane and Is Required for Efficient Phagocytosis.” *Eukaryotic Cell* 10 (9): 1207–18.

Schekman, R. 1985. “Protein Localization and Membrane Traffic in Yeast.” *Annual Review of Cell Biology* 1: 115–43. https://doi.org/10.1146/annurev.cb.01.110185.000555.

Schilde, Christina, Barbara Schönemann, Ivonne M Sehring, and Helmut Plattner. 2010. “Distinct Subcellular Localization of a Group of Synaptobrevin-like SNAREs in Paramecium Tetraurelia and Effects of Silencing SNARE-Specific Chaperone NSF.” *Eukaryotic Cell* 9 (2): 288–305.

Schilde, Christina, Thomas Wassmer, Joerg Mansfeld, Helmut Plattner, and Roland KissmehI. 2006. “A Multigene Family Encoding R-SNAREs in the Ciliate Paramecium Tetraurelia.” *Traffic* 7 (4): 440–55.

Sehring, Ivonne M, Jörg Mansfeld, Christoph Reiner, Erika Wagner, Helmut Plattner, and Roland Kissmehl. 2007. “The Actin Multigene Family of Paramecium Tetraurelia.” *BMC Genomics* 8 (1): 1–16.

Sehring, Ivonne M, Christoph Reiner, Jorg Mansfeld, Helmut Plattner, and Roland Kissmehl. 2007. “A Broad Spectrum of Actin Paralogs in Paramecium Tetraurelia Cells Display Differential Localization and Function.” *Journal of Cell Science* 120 (1): 177–90.

Sehring, Ivonne M, Christoph Reiner, and Helmut Plattner. 2010. “The Actin Subfamily PtAct4, out of Many Subfamilies, Is Differentially Localized for Specific Local Functions in Paramecium Tetraurelia Cells.” *European Journal of Cell Biology* 89 (7): 509–24.

Sherr, Evelyn B, Barry F Sherr, Robert D Fallon, and Steven Y Newell. 1986. “Small, Aloricate Ciliates as a Major Component of the Marine Heterotrophic Nanoplankton 1.” *Limnology and Oceanography* 31 (1): 177–83.

Shi, Lei, France Koll, Olivier Arnaiz, and Jean Cohen. 2018. “The Ciliary Protein IFT57 in the Macronucleus of *Paramecium*.” *Journal of Eukaryotic Microbiology* 65 (1): 12–27. https://doi.org/10.1111/jeu.12423.

Shin, John J. H., Oliver M. Crook, Alicia C. Borgeaud, Jérôme Cattin-Ortolá, Sew Y. Peak-Chew, Lisa M. Breckels, Alison K. Gillingham, Jessica Chadwick, Kathryn S. Lilley, and Sean Munro. 2020. “Spatial Proteomics Defines the Content of Trafficking Vesicles Captured by Golgin Tethers.” *Nature Communications* 11 (1): 5987. https://doi.org/10.1038/s41467-020-19840-4.

Sidjanin, D J, Anna K Park, Adam Ronchetti, Jamaria Martins, and William T Jackson. 2016. “TBC1D20 Mediates Autophagy as a Key Regulator of Autophagosome Maturation.” *Autophagy* 12 (10): 1759–75.

Slabodnick, Mark M., J. Graham Ruby, Sarah B. Reiff, Estienne C. Swart, Sager Gosai, Sudhakaran Prabakaran, Ewa Witkowska, et al. 2017. “The Macronuclear Genome of Stentor Coeruleus Reveals Tiny Introns in a Giant Cell.” *Current Biology* 27 (4): 569–75. https://doi.org/10.1016/j.cub.2016.12.057.

Smith, Daryl G S, Ryan M R Gawryluk, David F Spencer, Ronald E Pearlman, K W Michael Siu, and Michael W Gray. 2007. “Exploring the Mitochondrial Proteome of the Ciliate Protozoon Tetrahymena Thermophila: Direct Analysis by Tandem Mass Spectrometry.” *Journal of Molecular Biology* 374 (3): 837–63.

Son, Ora, Sunghan Kim, Yun-jeong Shin, Woo-Young Kim, Hee-Jong Koh, and Choong-Ill Cheon. 2015. “Identification of Nucleosome Assembly Protein 1 (NAP1) as an Interacting

Partner of Plant Ribosomal Protein S6 (RPS6) and a Positive Regulator of RDNA Transcription." *Biochemical and Biophysical Research Communications* 465 (2): 200–205.

Sonneborn, T M. 1970. "Methods in Paramecium Research." In *Methods in Cell Biology*, 4:241–339. Elsevier.

Sonneborn, T M. 1975. "The Paramecium Aurelia Complex of Fourteen Sibling Species." *Transactions of the American Microscopical Society*, 155–78.

Stelly, Nicole, Jean-Pierre Mauger, Michel Claret, and André Adoutte. 1991. "Cortical Alveoli of Paramecium: A Vast Submembranous Calcium Storage Compartment." *The Journal of Cell Biology* 113 (1): 103–12.

Strambio-De-Castillia, Caterina, Mario Niepel, and Michael P Rout. 2010. "The Nuclear Pore Complex: Bridging Nuclear Transport and Gene Regulation." *Nature Reviews Molecular Cell Biology* 11 (7): 490–501.

Sweetlove, Lee J, and Alisdair R Fernie. 2018. "The Role of Dynamic Enzyme Assemblies and Substrate Channelling in Metabolic Regulation." *Nature Communications* 9 (1): 1–12.

Tanaka, Keiji. 2009. "The Proteasome: Overview of Structure and Functions." *Proceedings of the Japan Academy, Series B* 85 (1): 12–36.

Tassin, Anne-Marie, Michel Lemullois, and Anne Aubusson-Fleury. 2015. "Paramecium Tetraurelia Basal Body Structure." *Cilia* 5 (1): 1–6.

Team, R Core, and others. 2013. "R: A Language and Environment for Statistical Computing."

Thomas, Brian C., Brent Pedersen, and Michael Freeling. 2006. "Following Tetraploidy in an *Arabidopsis* Ancestor, Genes Were Removed Preferentially from One Homeolog Leaving Clusters Enriched in Dose-Sensitive Genes." *Genome Research* 16 (7): 934–46. https://doi.org/10.1101/gr.4708406.

Thul, Peter J., Lovisa Åkesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, et al. 2017. "A Subcellular Map of the Human Proteome." *Science* 356 (6340). https://doi.org/10.1126/science.aal3321.

Tsypin, Lev M, and Aaron P Turkewitz. 2017. "The Co-Regulation Data Harvester: Automating Gene Annotation Starting from a Transcriptome Database." *SoftwareX* 6: 165–71.

Tyanova, Stefka, Tikira Temu, and Juergen Cox. 2016. "The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics." *Nature Protocols* 11 (12): 2301–19.

Valentine, Megan, and Judith van Houten. 2021. "Using Paramecium as a Model for Ciliopathies." *Genes* 12 (10): 1493. https://doi.org/10.3390/genes12101493.

Valerio, Hellen Paula, Felipe Gustavo Ravagnani, Angela Paola Yaya Candela, Bruna Dias Carvalho da Costa, Graziella Eliza Ronsein, and Paolo di Mascio. 2022. "Spatial Proteomics Reveals Subcellular Reorganization in Human Keratinocytes Exposed to UVA Light." *IScience* 25 (4): 104093.

Vayssié, L. 2000. "Molecular Genetics of Regulated Secretion in Paramecium." *Biochimie* 82 (4): 269–88. https://doi.org/10.1016/S0300-9084(00)00201-7.

Vayssié, Laurence, N de Loubresse, and Linda Sperling. 2001. “Growth and Form of Secretory Granules Involves Stepwise Assembly but Not Differential Sorting of a Family of Secretory Proteins in Paramecium.” *Journal of Cell Science* 114 (5): 875–86.

Veitia, Reiner A., Samuel Bottani, and James A. Birchler. 2008. “Cellular Reactions to Gene Dosage Imbalance: Genomic, Transcriptomic and Proteomic Effects.” *Trends in Genetics* 24 (8): 390–97. https://doi.org/10.1016/j.tig.2008.05.005.

Vitali, Valerio, Rebecca Hagen, and Francesco Catania. 2019. “Environmentally Induced Plasticity of Programmed DNA Elimination Boosts Somatic Variability in *Paramecium Tetraurelia*.” *Genome Research* 29 (10): 1693–1704. https://doi.org/10.1101/gr.245332.118.

Wassmer, Thomas, Roland Kissmehl, Jean Cohen, and Helmut Plattner. 2006. “Seventeen A-Subunit Isoforms of Paramecium V-ATPase Provide High Specialization in Localization and Function.” *Molecular Biology of the Cell* 17 (2): 917–30.

Weill, Uri, Ido Yofe, Ehud Sass, Bram Stynen, Dan Davidi, Janani Natarajan, Reut Ben-Menachem, et al. 2018. “Genome-Wide SWAp-Tag Yeast Libraries for Proteome Exploration.” *Nature Methods* 15 (8): 617–22.

Wideman, Jeremy G, Ryan M R Gawryluk, Michael W Gray, and Joel B Dacks. 2013. “The Ancient and Widespread Nature of the ER–Mitochondria Encounter Structure.” *Molecular Biology and Evolution* 30 (9): 2044–49.

Wiejak, Jolanta, Liliana Surmacz, and Elzbieta Wyroba. 2003. “Dynamin Involvement in Paramecium Phagocytosis.” *European Journal of Protistology* 39 (4): 416–22.

Wilkinson, Bonney, and Hiram F Gilbert. 2004. “Protein Disulfide Isomerase.” *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1699 (1–2): 35–44. https://doi.org/10.1016/j.bbapap.2004.02.017.

Williams, Elizabeth J.B., Csaba Pal, and Laurence D. Hurst. 2000. “The Molecular Evolution of Signal Peptides.” *Gene* 253 (2): 313–22.

Wood, Richard D, Michael Mitchell, John Sgouros, and Tomas Lindahl. 2001. “Human DNA Repair Genes.” *Science* 291 (5507): 1284–89.

Yang, Pinfen, Dennis R Diener, Chun Yang, Takahiro Kohno, Gregory J Pazour, Jennifer M Dienes, Nathan S Agrin, et al. 2006. “Radial Spoke Proteins of Chlamydomonas Flagella.” *Journal of Cell Science* 119 (6): 1165–74.

Yano, Junji, Anbazhagan Rajendran, Megan S. Valentine, Madhurima Saha, Bryan A. Ballif, and Judith L. van Houten. 2013. “Proteomic Analysis of the Cilia Membrane of Paramecium Tetraurelia.” *Journal of Proteomics* 78 (January): 113–22. https://doi.org/10.1016/j.jprot.2012.09.040.

Yap, Kyoko L, Justin Kim, Kevin Truong, Marc Sherman, Tao Yuan, and Mitsuhiko Ikura. 2000. “Calmodulin Target Database.” *Journal of Structural and Functional Genomics* 1 (1): 8–14.

Zhang, Ying, Zhihui Wen, Michael P Washburn, and Laurence Florens. 2010. “Refinements to Label Free Proteome Quantitation: How to Deal with Peptides Shared by Multiple Proteins.” *Analytical Chemistry* 82 (6): 2272–81.

Zhou, Wangbin, Yan Zhu, Aiwu Dong, and Wen-Hui Shen. 2015. “Histone H2A/H2B Chaperones: From Molecules to Chromatin-Based Functions in Plant Growth and Development.” *The Plant Journal* 83 (1): 78–95.

Zufall, Rebecca A., Casey L. McGrath, Spencer v. Muse, and Laura A. Katz. 2006. “Genome Architecture Drives Protein Evolution in Ciliates.” *Molecular Biology and Evolution* 23 (9): 1681–87. https://doi.org/10.1093/molbev/msl032.