



# VCU

Virginia Commonwealth University  
VCU Scholars Compass

---

Theses and Dissertations

Graduate School


---

2022

## Identifying the human homologs of yeast Rab proteins Ypt10 & Ypt11 and a global-scale louse endosymbiont genome variation

Nathaniel P. Smith  
*Virginia Commonwealth University*

Follow this and additional works at: <https://scholarscompass.vcu.edu/etd>

 Part of the [Biodiversity Commons](#), [Bioinformatics Commons](#), [Cell Biology Commons](#), [Computational Biology Commons](#), [Evolution Commons](#), and the [Molecular Genetics Commons](#)

© The Author

---

Downloaded from

<https://scholarscompass.vcu.edu/etd/7167>

This Thesis is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact [libcompass@vcu.edu](mailto:libcompass@vcu.edu).

©Nathaniel Smith 2022

All rights reserved

**Identifying the human homologs of yeast Rab proteins Ypt10 & Ypt11 and a global-scale  
louse endosymbiont genome variation**

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science  
at Virginia Commonwealth University.

By

Nathaniel Smith

B.S. Bioinformatics – Virginia Commonwealth University

Co-Mentors:

Derek C. Prosser, Ph.D.

Assistant Professor, Department of Biology

Bret M. Boyd, Ph.D.

Assistant Professor, Center for Biological Data Science

Virginia Commonwealth University

Richmond, Virginia

December 2022

## Acknowledgements

I'd like to thank:

- Dr. Allison Johnson, my undergraduate and graduate advisor, for encouraging me to pursue a Master's degree in Bioinformatics, and her continued support, motivation, and encouragement throughout my tenure at VCU
- Dr. Bret Boyd, my bioinformatics mentor, for his guidance on the phylogenetic approach, the opportunity to practice my programming skills, as well as his friendship and mentorship
- Dr. Derek Prosser, my wet-lab mentor, for providing me with the opportunity to learn and practice mammalian cell culturing in his lab, and the opportunity to teach and do research
- My mom, for her unwavering love and encouragement to follow my dreams
- My dad, for his keen interest and support in pursuing a scientific career

## Table of Contents

Acknowledgements .....	3
List of Abbreviations .....	7
List of Figures and Tables .....	8
Chapter 1: Identifying the human homologs of yeast Rab proteins Ypt10 & Ypt11 .....	9
Abstract .....	9
Introduction .....	10
ALS8 and the VAPB mutation .....	10
Genetic screening of an ALS8 model to identify suppressors of ER stress sensitivity ....	11
Rab GTPases and the application of budding yeast as a disease model .....	13
Rab GTPases act as molecular switches .....	14
The application of phylogenetics to uncover distantly related proteins .....	15
Previous studies attempting to identify the human homologs of <i>YPT10</i> and <i>YPT11</i> .....	16
Objective .....	16
Methods .....	16
Data Sources .....	16
Taxa selection .....	17
Multiple sequence alignment .....	18

Phylogenetic inference .....	18
Removal of rogue taxa .....	19
Results .....	19
Discussion .....	36
Rab22a, candidate homolog of Ypt10 in humans .....	36
Rab31, also known as Rab22b, as a candidate homolog of Ypt10 .....	37
Rab20 as a candidate homolog of Ypt10 .....	37
Rab34 as a candidate homolog of Ypt11 .....	38
Rab36 as a candidate homolog of Ypt11 .....	38
Insights on potential functions of uncharacterized proteins .....	39
Limitations of this study .....	40
Future directions .....	41
Conclusion .....	42
Chapter 2: Global-scale louse endosymbiont genome variation study .....	43
Abstract .....	43
Introduction .....	43
The human head louse and its obligate intracellular endosymbiont's genome reduction.	43
Objective .....	45

Methods .....	45
Sequence data pipeline .....	45
Parsing VCF files for variant positions .....	45
Parsing FASTA file of USDA genome to plot intergenic vs intragenic variation .....	46
Results .....	46
Discussion .....	49
Conclusion .....	50
Supplementary materials .....	51
References .....	53

## List of Abbreviations

Abbreviation	Meaning
ADL	Adenylosuccinate lyase
ALS	Amyotrophic lateral sclerosis
ALS8	Amyotrophic lateral sclerosis subtype 8
AMP	Adenosine monophosphate
CHO	Chinese hamster ovary
ER	Endoplasmic reticulum
FFAT	Two phenylalanines in an acidic tract
GAP	GTPase-activating protein
GEF	Guanine nucleotide exchange factor
GTPases	Guanosine triphosphatases
IF1	Translation initiation factor 1
ML	Maximum Likelihood
MSP	Major sperm protein
MUSCLE	Multiple Sequence Comparison by Log-Expectation
NGS	Next Generation Sequencing
ORPs	OSBP-related proteins
OSBP	Oxysterol-binding protein
RAxML	Randomized-Axelerated Maximum Likelihood
SAM	Sequence alignment map
SREBP	Sterol-responsive element-binding protein
UPR	Unfolded protein response
USDA	<i>Candidatus Riesia</i> subtype
VAMP	Vesicle-associated membrane protein
VAPB	VAMP-associated protein B
VCF	Variant Call Format



## List of Figures

■ Figure 1: Number of Rabs vs estimated time of divergence from humans .....	13
■ Figure 2: The Rab cycle .....	15
■ Table 1: Phylogenetic results table .....	19
■ Figure 3: Initial phylogenetic tree .....	23
■ Figure 4: Initial phylogenetic tree close-up .....	24
■ Figure 5: Gblocks aligned tree .....	25
■ Figure 6: Gblocks aligned tree close-up .....	26
■ Figure 7: HMM bootstrap tree .....	27
■ Figure 8: HMM bootstrap tree close-up .....	28
■ Figure 9: MUSCLE bootstrap tree .....	29
■ Figure 10: MUSCLE bootstrap tree close-up .....	30
■ Figure 11: Revised taxa phylogenetic tree .....	31
■ Figure 12: Revised taxa phylogenetic tree close-up .....	32
■ Figure 13: Gene of interest clade phylogenetic tree .....	33
■ Figure 14: Rogue taxa removal phylogenetic tree .....	34
■ Figure 15: Rogue taxa removal phylogenetic tree close-up .....	35
■ Figure 16: Histogram of variation by position for USDA genome .....	47
■ Figure 17: Histogram of intragenic variation by position of USDA genome .....	48
■ Figure 18: Histogram of intergenic variation by position of USDA genome .....	49
■ Figure S1: Alignment of Ypt10 and candidate human homologs .....	51
■ Figure S2: Alignment of Ypt11 and candidate human homologs .....	52

## Chapter 1: Identifying the human homologs of yeast Rab proteins Ypt10 & Ypt11

### Abstract

Amyotrophic lateral sclerosis (ALS) is a late-onset fatal neurodegenerative disease that causes loss of upper and/or lower motor neurons, and currently has no treatment or cure available. Over 90% of cases occur spontaneously with unknown causes, highlighting the complexity of the disease, and only 10% of cases are linked to heritable genetic mutations. Numerous ALS-linked genes are conserved through evolution, and model organisms may therefore provide opportunities to understand disease pathology at a molecular or cellular level, proving instrumental in identifying therapeutic targets. ALS subtype 8 (ALS8) is caused by an autosomal dominant P56S mutation in the *VAPB* gene that alters morphology and function of the endoplasmic reticulum (ER), leading to ER stress sensitivity. In a budding yeast (*Saccharomyces cerevisiae*) model of ALS8 that recapitulates these phenotypes, we identified Rab GTPases and their regulators involved in membrane traffic as a class of genes whose overexpression improved tolerance to ER stress. Yeast possesses 11 Rab genes, and while the majority of these are characterized and have clear homologs in mammals, the function of both *YPT10* and *YPT11* remain poorly understood. Notably, *YPT10* was isolated as a possible suppressor of ALS8 phenotypes in the yeast model.

The goal of this study was to obtain genetic information about Ypt10 and Ypt11 function and phylogeny using bioinformatic approaches. By identifying the human homologs of yeast Rabs, we can potentially study their function, and identify targets for ALS treatments. This study narrowed down the potential human homologs for Ypt10 to *Homo sapiens* Rab20, Rab22a, and Rab31, as well as for Ypt11 to *H. sapiens* Rab34 and Rab36.

## Introduction

### ALS8 and the *VAPB*<sup>P56S</sup> mutation

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease in humans that causes death of upper and lower motor neurons, progressively resulting in paralysis and respiratory failure, with a median survival of three to five years after diagnosis (Yamashita and Ando, 2015; Brotman *et al.*, 2021). The etiological factors of ALS are poorly understood, as over 90% of cases occur sporadically; however, the remaining 10% of cases can be attributed to a family history of ALS (Yamashita and Ando, 2015; Brotman *et al.*, 2021). This suggests heritable mutations can cause the onset of the disease. To date, mutations in over 30 genes have been attributed to different subtypes of ALS (Yamashita and Ando, 2015; Brotman *et al.*, 2021). Prosser *et al.* (2008) sought to understand the molecular pathology and relieve the phenotypic symptoms of ALS8 in mammalian cell lines (Chinese hamster ovary cells; CHO). ALS8 is an autosomal dominant subtype of ALS caused by a P56S mutation in vesicle-associated membrane protein (VAMP)-associated protein B (VAPB), which plays important roles in maintaining the structure and function of the endoplasmic reticulum (ER) in motor neurons (Nishimura *et al.*, 2005; Teuling *et al.*, 2007; Prosser *et al.*, 2008; Aliaga *et al.*, 2013; Kabashi *et al.*, 2013). The VAPB protein has a single C-terminal transmembrane domain that anchors it in the ER, with the majority of the protein residing within the cytoplasm. Mutation P56S in the *VAPB* gene (*VAPB*<sup>P56S</sup>) leads to misfolding and protein aggregation of the cytoplasmic major sperm protein (MSP) domain of VAPB, and the transmembrane domain causes ER membrane to be incorporated into inclusions (Kanekura *et al.*, 2006; Teuling *et al.*, 2007). As a result, *VAPB*<sup>P56S</sup> causes collapse of the ER, disrupting vital cellular processes including membrane traffic, which may contribute to cell death in motor neurons through unknown mechanisms (Teuling *et al.*,

2007; Prosser *et al.*, 2008). Notably, VAPB<sup>P56S</sup> expression leads to aberrant regulation of the unfolded protein response (UPR), which is a critical stress response during accumulation of misfolded proteins, in which the cell pauses translation and upregulates chaperone function (Kanekura *et al.*, 2006; Suzuki *et al.*, 2009; Aliaga *et al.*, 2013; Tokutake *et al.*, 2015). Inability to correctly respond to ER stress may be causative to, or contribute to, motor neuron loss in ALS8. The major sperm protein (MSP) domain of VAPB and its closely-related homolog VAPA bind to FFAT (two phenylalanines in an acidic tract) motifs that are found in numerous cytoplasmic proteins involved in lipid transfer, including oxysterol-binding protein (OSBP), OSBP-related proteins (ORPs), and sterol-responsive element-binding protein (SREBP) (Kanekura *et al.*, 2006). It was subsequently discovered that overexpression of an FFAT motif in VAPB<sup>P56S</sup>-expressing CHO cells reduced ER aggregation and restored exit of transmembrane cargos from the ER (Prosser *et al.*, 2008).

### **Genetic screening of an ALS8 model to identify suppressors of ER stress sensitivity**

Based on the findings of Prosser *et al.* (2008), a follow-up study was conducted to identify genes capable of suppressing the ER stress sensitivity that is characteristic of VAPB<sup>P56S</sup> expression (D. Prosser, unpublished results); they used the budding yeast *Saccharomyces cerevisiae* as a model system for genetics, and because the *VAPB* gene is largely conserved between humans and yeast. In yeast, there are two homologs (*SCS2* and *SCS22*) that are similarly involved in ER structure and function. Moreover, the P56 amino acid residue that is mutated in ALS8 is present within a highly conserved region of all three proteins.

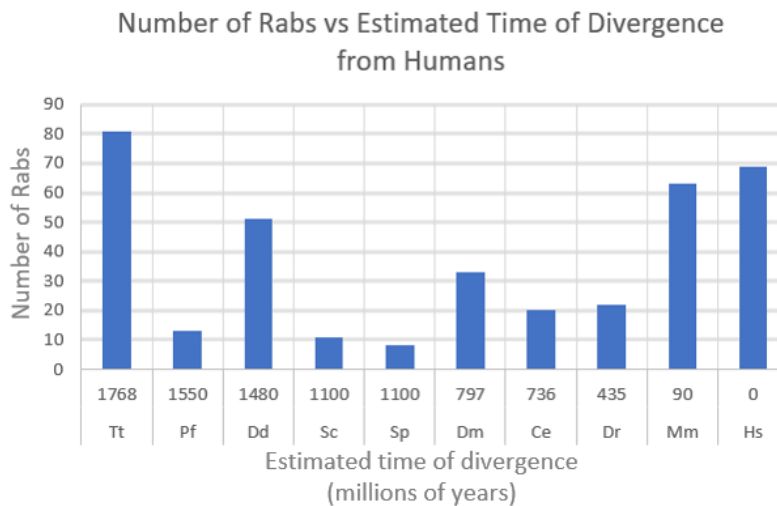
Prosser *et al.* created a mutant yeast strain (*scs2*Δ *scs22*Δ + *scs2*<sup>P51S, P58S</sup>), hereafter referred to as yALS8, in which the chromosomal copies of *SCS2* and *SCS22* were deleted and a

mutant *scs2*<sup>P51S, P58S</sup> allele of *SCS2*, which is equivalent to the human *VAPB*<sup>P56S</sup> mutation (Nakamichi *et al.*, 2011), was reintroduced at a heterologous gene locus (*LEU2*). The yALS8 strain shows several hallmarks of ALS that are seen in mammalian cells expressing *VAPB*<sup>P56S</sup>, including collapse of the ER, appearance of membrane-containing inclusions in the perinuclear region, and hypersensitivity to ER stress-inducing drugs such as tunicamycin, which blocks glycosylation of newly-synthesized proteins (Leavitt *et al.*, 1977; Merlie *et al.*, 1982). yALS8 cells were transformed with a plasmid-based library that overexpressed short regions of the yeast genome (~10 kb genomic intervals containing an average of 2-3 complete genes; Carlson and Botstein, 1982) and grown on plates containing tunicamycin at a concentration (1 mg/ml) that was lethal in yALS8, but not wild-type cells. Colonies that grew had presumably acquired a suppressor gene; thus, plasmids from each colony were isolated and sequenced to identify candidate genomic intervals. Subsequently, individual genes from each interval were subcloned and independently tested for suppression of stress sensitivity to validate the suppressor gene.

One of the genes found in this genetic screen was *AVL9*, a putative guanine nucleotide exchange factor (GEF) that activates Rab GTPases but whose substrate(s) has not yet been identified. Remarkably, other Rab GTPases and Rab regulatory proteins were also isolated as potential suppressor genes from this screen: the secretory Rab *SEC4*, a Rab of unknown function (*YPT10*), and the Rab6/Ypt6-inactivating gene *GYP6*. Of these, *SEC4* and *GYP6* have been verified as suppressors, while *YPT10* remains a potential suppressor. Since Rabs and their regulators are highly overrepresented in the screen, we reason that they may form an important class of genes capable of ameliorating ER stress responses in yALS8.

## Rab GTPases and the application of budding yeast as a disease model

Rab GTPases are specialized GTP-binding and -hydrolyzing proteins that are involved in organelle identity and in transport/tethering of lipid- and protein-containing vesicles to their appropriate target organelles and membranes (Guadagno and Progida, 2019). The dysfunction of Rabs may lead to a number of fatal and chronic diseases in humans, including ALS, Alzheimer's disease, Parkinson's disease, and Huntington's disease (Zahraoui *et al.*, 1989). While over 60 Rabs have been identified in humans, the budding yeast *Saccharomyces cerevisiae* has only 11 Rab genes, suggesting that the Rab family has undergone expansion during vertebrate evolution, as seen in Figure 1.



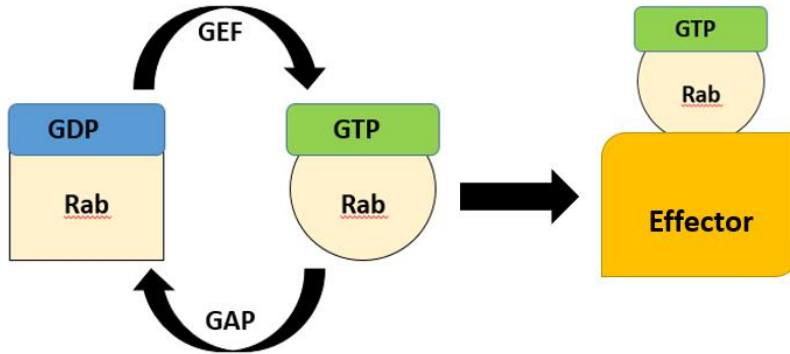
**Figure 1: Number of Rabs compared to estimated pairwise divergence time for each species compared to humans.** Estimated median time of divergence from humans. Information obtained from TimeTree (Kumar *et al.*, 2017); graph produced in Excel. Tt = *Tetrahymena thermophila*; Pf = *Plasmodium falciparum*; Dd = *Dictyostelium discoideum*; Sc = *Saccharomyces cerevisiae*; Sp = *Saccharomyces pombe*; Dm = *Drosophila melanogaster*; Ce = *Caenorhabditis elegans*; Dr = *Danio rerio*; Mm = *Mus musculus*; Hs = *Homo sapiens*.

Of the 11 yeast Rabs, 9 of them are highly conserved and have known orthologues across many diverse species (Buvelot *et al.*, 2006); however, some members of this protein family

remain poorly characterized (Zahraoui *et al.*, 1989, Buvelot *et al.*, 2006; Li and Marlin, 2015; Homma *et al.*, 2021). The feasibility of studying yeast compared to mammalian Rabs, especially when considering that these proteins are highly conserved across species, makes yeast a prime candidate for studying Rabs, their functions, and their genetic interactions due to the reduced number of Rabs required for cellular function in yeast.

### **Rab GTPases act as molecular switches**

The localization and function of Rabs depends on whether the protein is active or inactive, which is due to the Rab GTPase alternating between GDP- (inactive) and GTP-bound (active) states (Hutagalung and Novick, 2011). During activation, a guanine nucleotide exchange factor (GEF) binds to an inactive, GDP-associated Rab and causes a conformational change, leading to the release of GDP and association with GTP, which is present at higher concentration in the cytoplasm (Figure 2). During inactivation, a GTPase-activating protein (GAP) binds to the active, GTP-associated Rab, stimulating the inherent GTPase activity of the Rab and thereby leading to hydrolysis of GTP into GDP. The cycling between activation and inactivation of Rabs dictates the transport of lipid- and protein-containing vesicles to their appropriate membrane-bounded organelles, and plays roles in establishing and/or maintaining organelle identity (Stenmark and Olkkonen, 2001; Hutagalung and Novick, 2011). For Rabs to perform their functions in membrane trafficking, the active, GTP-bound Rab must bind to an effector protein; effector proteins have various key functions, such as selecting cargo, promoting vesicle transport, and tethering vesicles to their appropriate target membrane (Grosshans *et al.*, 2006; Hutagalung and Novick, 2011). Understanding the physical and genetic interactions of Rabs with GEFs, GAPs, and effector proteins remains a key area of inquiry for diseases that have mutations in these proteins or that impair membrane trafficking.



**Figure 2: The Rab cycle.** An inactive, GDP-bound Rab is activated by a guanine nucleotide exchange factor (GEF) to become the active GTP-bound Rab, which is then inactivated by a GTPase activating protein (GAP) to return to the inactive GDP-bound state. In the active GTP-bound state, the Rab binds to effector proteins, which tethers vesicles to their appropriate target membranes prior to fusion.

### The application of phylogenetics to uncover distantly related proteins

A way to identify the potential human homolog of our yeast ALS candidate suppressor gene, *YPT10*, is using a bioinformatics method called phylogenetics, which uses DNA or protein sequences and a variety of statistical algorithms to construct a dendrogram showing evolutionary relationships among species, entire groups, or individual genes. The application of phylogenetics in this study utilized amino acid sequences to identify the potential human homologs of the budding yeast Ypt10 and Ypt11 proteins. By inferring the evolutionary relationships between yeast proteins in non-humans and their human counterparts, an informed starting point can be carried out in wet-lab analyses which can identify potential therapeutic approaches with a narrower, more accurate set of target genes. Uncovering the genetic interactions and evolutionary relationships among Rabs in this study could help us further understand this disease and provide a template for future ALS research



## **Previous studies attempting to identify the human homologs of *YPT10* and *YPT11***

To date, multiple phylogenetic studies have attempted to elucidate the evolutionary relationships between yeast and human Rabs (Stenmark and Olkkonen, 2001; Frei *et al.*, 2006; Klöpffer *et al.*, 2012); however, these studies have not definitively pinpointed some homologs. Frei *et al.*, (2006) suggest the candidate human homolog for *YPT10* may be Rab20, and the potential homolog of *YPT11* is unclear; despite this, they fail to go into detail about their phylogenetic analysis or the methodology; the accuracy of these results are thus unclear. Klöpffer *et al.*, (2012) and Stenmark and Olkkonen (2001) performed similar studies to uncover the evolutionary relationships between yeast and human Rabs, but do not have results for our two yeast genes of interest, *YPT10* and *YPT11*. Therefore, additional research is warranted to identify candidate homologs of *YPT10* and *YPT11* in humans.

## **Objective**

My objective in this study is to use phylogenetic methods to identify candidate human homologs for two uncharacterized yeast Rabs, *YPT10* and *YPT11*.

## **Methods**

### **Data sources**

The majority of protein sequences were obtained from the UniProt database; only entries given the “reviewed” status were accepted, as we have confidence that they were correctly annotated (Uniprot [Computer software]. Retrieved from [www.uniprot.org](http://www.uniprot.org)). Additional Rab sequences were collected from model-organism specific websites, including the *Saccharomyces* Genome Database (Cherry *et al.*, 2012), WormBook (Eisenmann, 2005), or FlyBase (Larkin *et al.*, 2021). The Rabs from these model organism-specific websites were cross-referenced with

literature to confirm the accuracy of identified Rabs included in this study (Pereira-Leal and Seabra, 2001; Zhang *et al.*, 2007; Gallegos *et al.*, 2012).

### **Taxa selection**

Taxa consisted of single-celled organisms, multicellular organisms, invertebrates, and finally vertebrates – representing diverse organisms across the evolutionary tree of life.

Additionally, I evaluated the inclusion of some specific taxa as follows:

First, to facilitate evaluation of the phylogenetic results, fission yeast was included since the homology between its Rabs and Rabs in baker's yeast have been previously established (Pereira-Leal and Seabra, 2001).

Second, *Plasmodium* and *Tetrahymena* were included as they were similar in complexity to baker's yeast. *Tetrahymena* had an extensive phylogenetic study on Rabs within the species, of which the authors sent FASTA files of their protein sequences to include in our study (Bright *et al.*, 2010).

Third, the social amoeba *D. discoideum*, although having a rather similar number of Rabs to humans and many more than yeast, shares between 8,000 and 10,000 genes with vertebrates. It is a valuable taxon in the dataset because it can act as a reference between yeast (similar in cellular complexity and evolutionary divergence) and humans (similar in genomic composition) (Sunderland, 2009).

Fourth, *C. elegans*, *D. melanogaster*, and *D. rerio* were included in other phylogenetic studies of Rabs (Pereira-Leal and Seabra, 2001; Mackiewicz and Wyroba, 2009; Klöpper *et al.*, 2012), and due to their increasing cellular complexity, would be the final taxa added to begin the study.

Fifth, Rabs from *T. thermophila* were initially included, but often produced long branches, and clustered with Ypt11, and were therefore excluded; unlike organisms of similar complexity (fission yeast, baker's yeast, *Plasmodium*), it had more Rabs than significantly more complex multi-cellular organisms, such as zebrafish, roundworm, fruit fly, even humans. Moving forward, *T. thermophila* was excluded from the dataset.

Sixth, a species that would have been ideal to include was tunicates, as they are the closest relative to humans compared to other invertebrate animals. However, Rab sequences from any tunicate or species from the Chordata family proved difficult to identify. Instead, we included *M. musculus* because of the evolutionary proximity to humans.

### **Multiple sequence alignment**

Alignments of Rab proteins were created using MUSCLE (Edgar, 2004) and HMM (Johnson *et al.*, 2010). Gblocks (Castresana, 2000) and TrimAl (Capella-Gutiérrez *et al.*, 2009) were used to modify the alignments created by MUSCLE and HMM, by removing poorly aligned or gap heavy regions.

### **Phylogenetic inference**

Phylogenetic trees were inferred by Randomized Accelerated Maximum Likelihood (RAxML), using the PROTGAMMAAUTO model of protein evolution (“PROT” referring to amino acids, “GAMMA” referring to a gamma model of rate heterogeneity, “AUTO” referring to an automatic protein model selection; Mayrose *et al.*, 2005; Stamatakis, 2014). 100 bootstrap replicates were obtained using RAxML's rapid bootstrapping function. All trees were rooted on *S. cerevisiae RAS1*, as it is the founding member of the Ras superfamily, is most closely related to Rab sequences, and serves as an outgroup to base our analyses on.

## Removal of rogue taxa

Previous studies (Aberer *et al.*, 2013, Goloboff and Szumik, 2015) have shown that pruning rogue taxa can improve phylogenetic signal, therefore the tool *RogueNaRok* (Aberer *et al.*, 2011) was used for rogue taxa removal. The algorithm uses an optimization technique to identify rogue taxa, such that when certain tips (representing individual Rab genes) are removed from a clade and exceed a majority rule consensus threshold, indicate removal of the tips had the most impact on disrupting tree arrangements and branch support values (Aberer *et al.*, (2013).

## Results

**Table 1:** A synopsis of different phylogenies produced, candidate sister genes of *YPT10* and *YPT11*, and recovery of known homologs in humans and yeast.

Alignment type and edits	Tree inference program	Ypt10 homolog	Ypt11 homolog	Do yeast Rabs pair with their known human homolog?					
				Ypt1/ Rab1	Ypt6/ Rab6	Ypt7/ Rab7	Vps21, Ypt52, Ypt53/ Rab5	Sec4/ Rab3, Rab8	Ypt31, Ypt32/ Rab11
MUSCLE	RAxML	Rab22a, Rab31	Rab34, Rab36	Yes	Yes	Yes	Yes	Yes	Yes
Gblocks	RAxML	Rab34, Rab36	Rab8, Rab10	Yes	Yes	Yes	Yes	No	Yes
HMM	RAxML	Rab22a, Rab31	Rab34, Rab36	Yes	Yes	Yes	Yes	Yes	Yes
MUSCLE*	RAxML	Rab22a, Rab31	Rab34, Rab36	Yes	Yes	Yes	Yes	Yes	Yes
MUSCLE*, taxa**	RAxML	Rab22a, Rab31	Rab34, Rab36	Yes	Yes	Yes	Yes	Yes	Yes
MUSCLE*, taxa**, rogue taxa***	RAxML	Rab20	Rab34, Rab36	Yes	Yes	Yes	Yes	Yes	Yes

\* = alignment trimmed of 75% of columns with gap characters. \*\* = removal of *T. thermophila*, addition of *M. musculus*. \*\*\* = Rogue taxa removed.

An initial tree (Figure 3) inferred using RAxML and an alignment created using the tool MUSCLE, that included all Rabs, yielded a tree with low bootstrap support (<75%) at nodes

joining Ypt10 and Ypt11 with candidate homologs. Ypt10 was found to be most closely related to human Rab22a and Rab31, while Ypt11 was mostly closely related to Rab34 and Rab36 (Table 1; Figure 3, Figure 4). Both Rab22a and Rab31, as well as Rab34 and Rab36, are considered paralogs, sharing high sequence similarity with each other; as the clade of our yeast protein of interest splits with our candidate human homologs, these two human proteins diverge, resulting in two candidate human homologs.

Following modification of the original MUSCLE alignment using Gblocks, which included all Rabs, all positions from the alignment that contained over 50% gap characters were removed. A newly-inferred tree using RAxML yielded an alternative arrangement with different Rabs being most closely related to Ypt10 and Ypt11 (Table 1 and Figure 5). The tree created by Gblocks (Figure 5, Figure 6) displayed a significantly different configuration than the tree aligned by MUSCLE, with the prior homologs of Ypt11 now belonging to Ypt10. Ypt10 was found to be most closely related to Rab34 and Rab36, while Ypt11 was found to be most closely related to Rab8 and Rab10. After a literature review, Talvera and Castresana (2007) found that for short genes (400-800 amino acids), the regions removed by Gblocks added more noise than signal compared to complete alignments, and with most of this dataset containing short genes (<400 amino acids), Gblocks modification of the alignment may yield questionable results.

Based on a previous study by Pereira-Leal and Seabra (2001) that utilized a Hidden Markov Model (HMM) as their alignment tool, we decided to infer phylogenies based on alignments created using both MUSCLE and an HMM alignment program to test if various alignments methods had a significant impact on the resulting phylogeny. An analysis that included all Rabs (Figure 7) was inferred using RAxML and an alignment created using HMM yielded a tree with low bootstrap support (<75%) at nodes joining Ypt10 and Ypt11 with

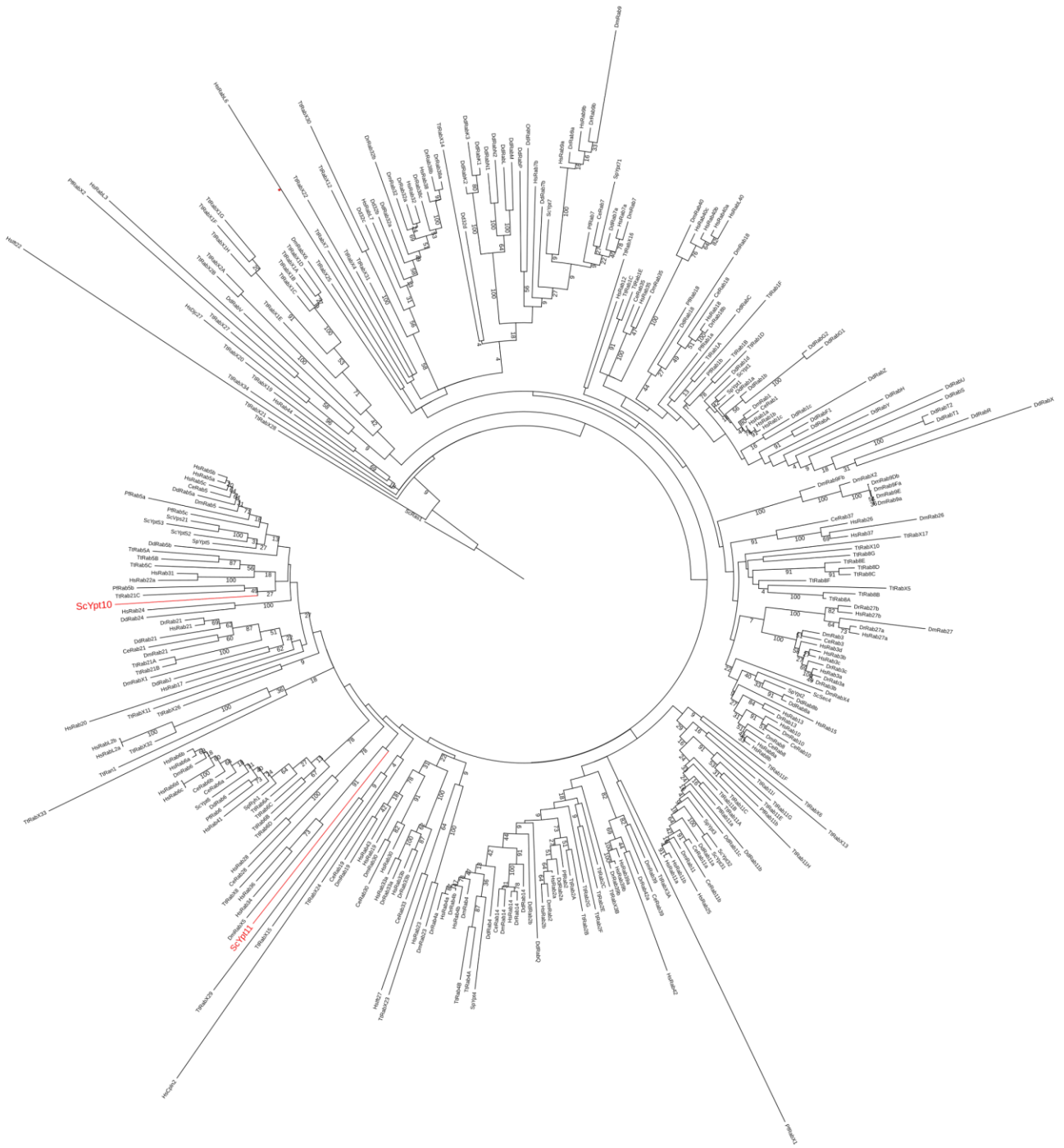
candidate homologs. Ypt10 was found to be most closely related to human Rab22a and Rab31, while Ypt11 was mostly closely related to Rab34 and Rab36 (Table 1; Figure 7, Figure 8). There was no difference in the pairings between Ypt10 and Ypt11 and their closest human Rabs; however, the MUSCLE tree resulted in slightly higher bootstrap support values, and was used in subsequent analyses.

Building off the original tree (Figure 3), there were concerns about low bootstrap support for our two yeast proteins of interest, and TrimAl was used to remove positions in the alignment where 75% of columns contained gaps (Figure 9) in the original alignment generated using MUSCLE. Inference from the reduced alignment yielded a tree with low bootstrap support (<75%) at nodes joining Ypt10 and Ypt11 with candidate homologs. Ypt10 was found to be most closely related to human Rab22a and Rab31, while Ypt11 was mostly closely related to Rab34 and Rab36 (Table 1; Figure 9, Figure 10).

Despite the alignment trimming, our results continued to yield low bootstrap support for our two proteins of interest. One particular concern was that *T. thermophila*, which shows an unexpected expansion of Rab genes (Figure 1), might introduce noise in our dataset, and that there needed to be another mammal in the dataset. Thus, the *T. thermophila* Rabs were removed, and Rabs from *M. musculus* were included instead (Table 1; Figure 11, Figure 12). A tree inferred using RAxML, an alignment using the tool MUSCLE with the updated list of Rabs, and using TrimAl to remove positions in the alignment where 75% of columns contained gaps, continued to indicate that Ypt10 is most closely related to Rab22a and Rab31, and Ypt11 is most closely related to Rab34 and Rab36.

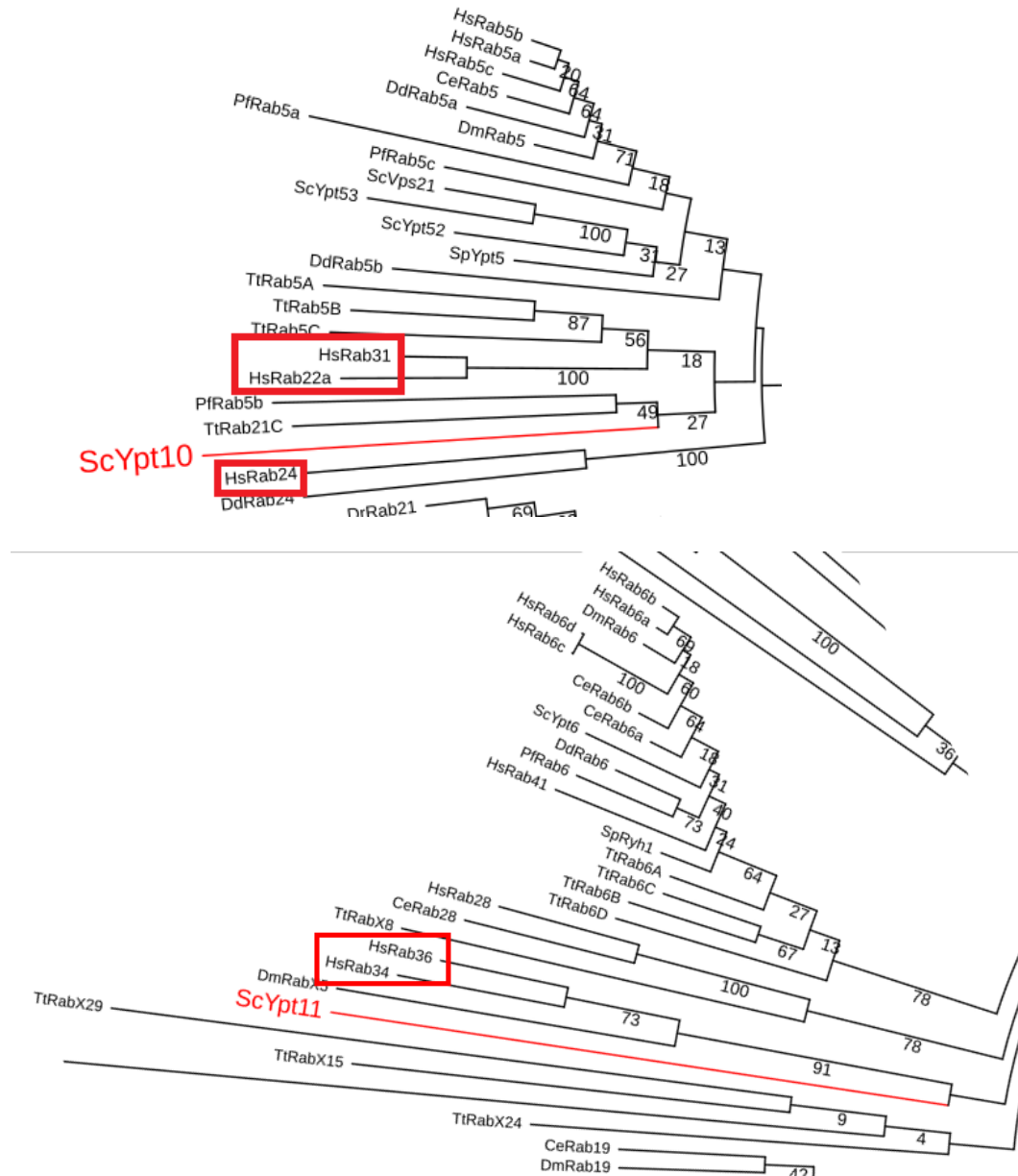
Throughout all iterations there were a few Rabs in our dataset that continued to have extremely long branches, across the different methodologies used. The concern was that these

consistently obscure Rabs, or rogue taxa, could be affecting the resolution of Ypt10 and Ypt11, as well as introducing noise into the study, and was subsequently addressed by removing these rogue taxa. Inference using RAxML and an alignment using the tool MUSCLE, with the updated list of Rabs, using TrimAl to remove positions in the alignment where 75% of columns contained gaps, and removal of rogue taxa, yielded a tree with low bootstrap support (<75%) at nodes joining Ypt10 and Ypt11 with candidate homologs. Ypt10 was found to be most closely related to human Rab20, while Ypt11 was mostly closely related to Rab34 and Rab36 (Table 1; Figure 14, Figure 15).

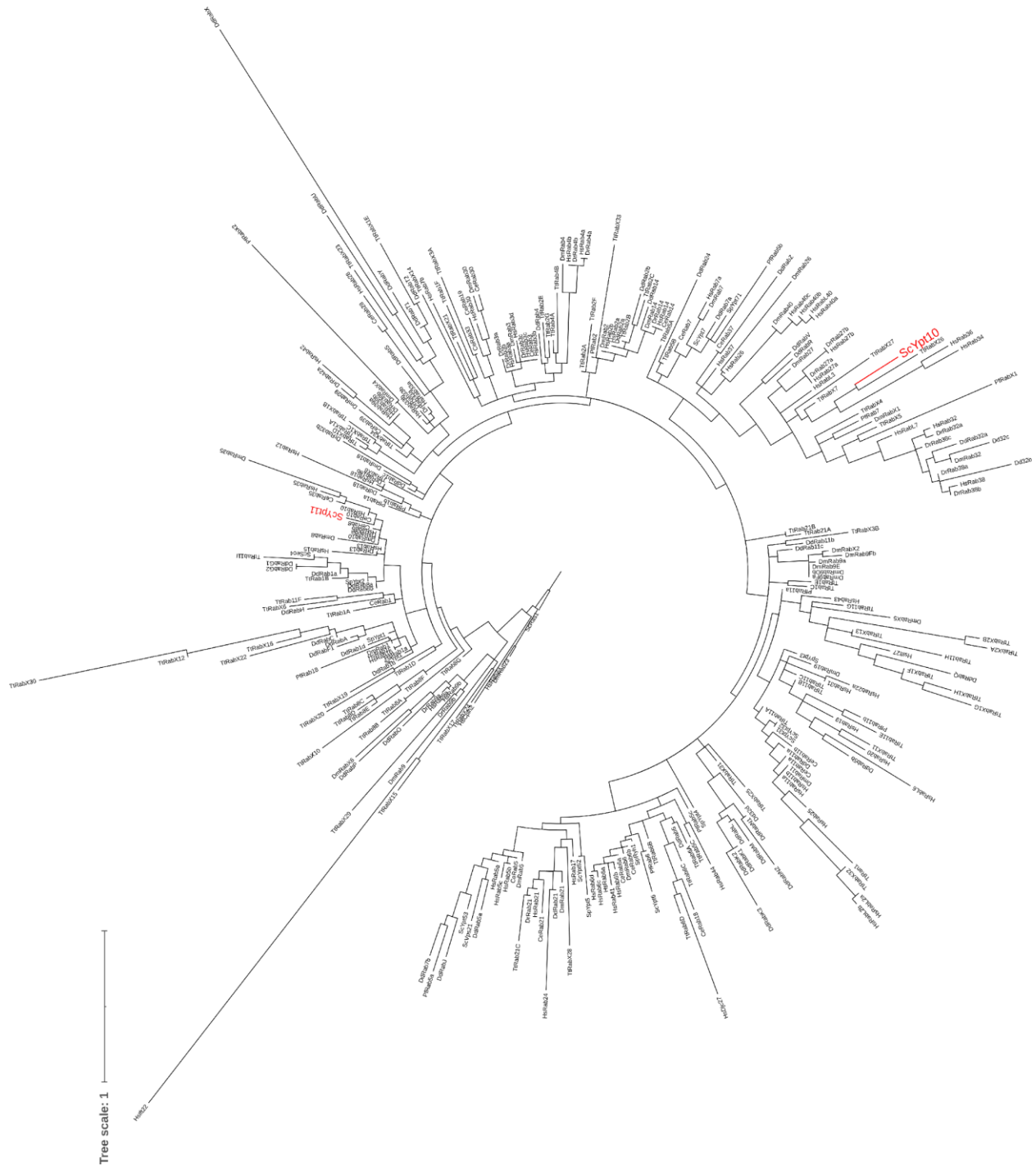


**Figure 3: Maximum-likelihood phylogram depicting the evolutionary relationships of Ras1 from *S. cerevisiae* and other Rabs.** The tree was inferred using all sites in the aligned amino acid sequences using RAxML. Proteins were aligned using MUSCLE. Numbers at nodes represent percent of 100 bootstrap replicates that recovered the same node.

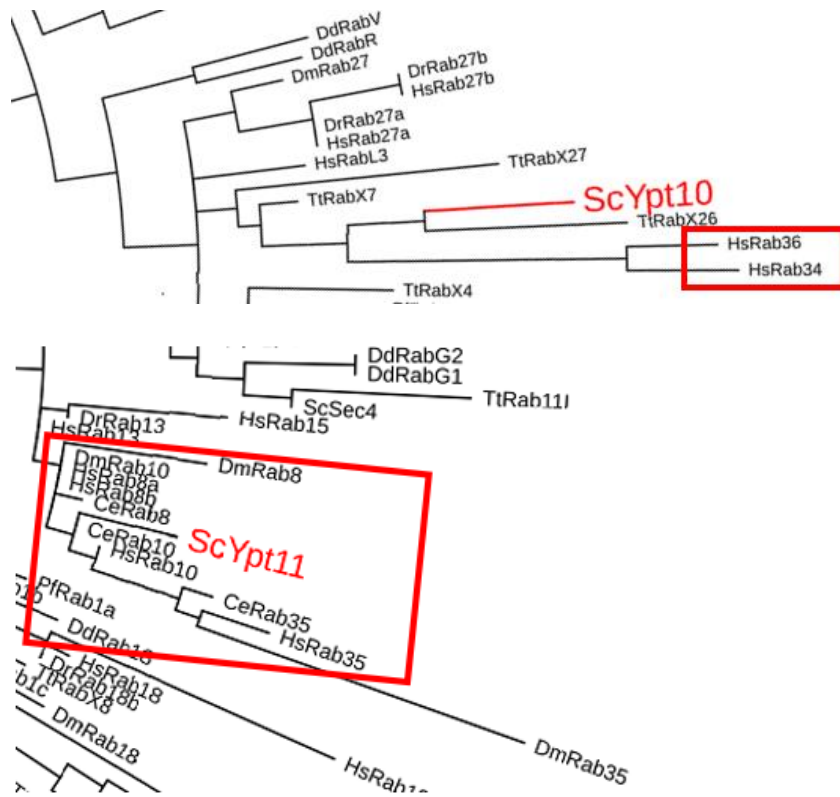




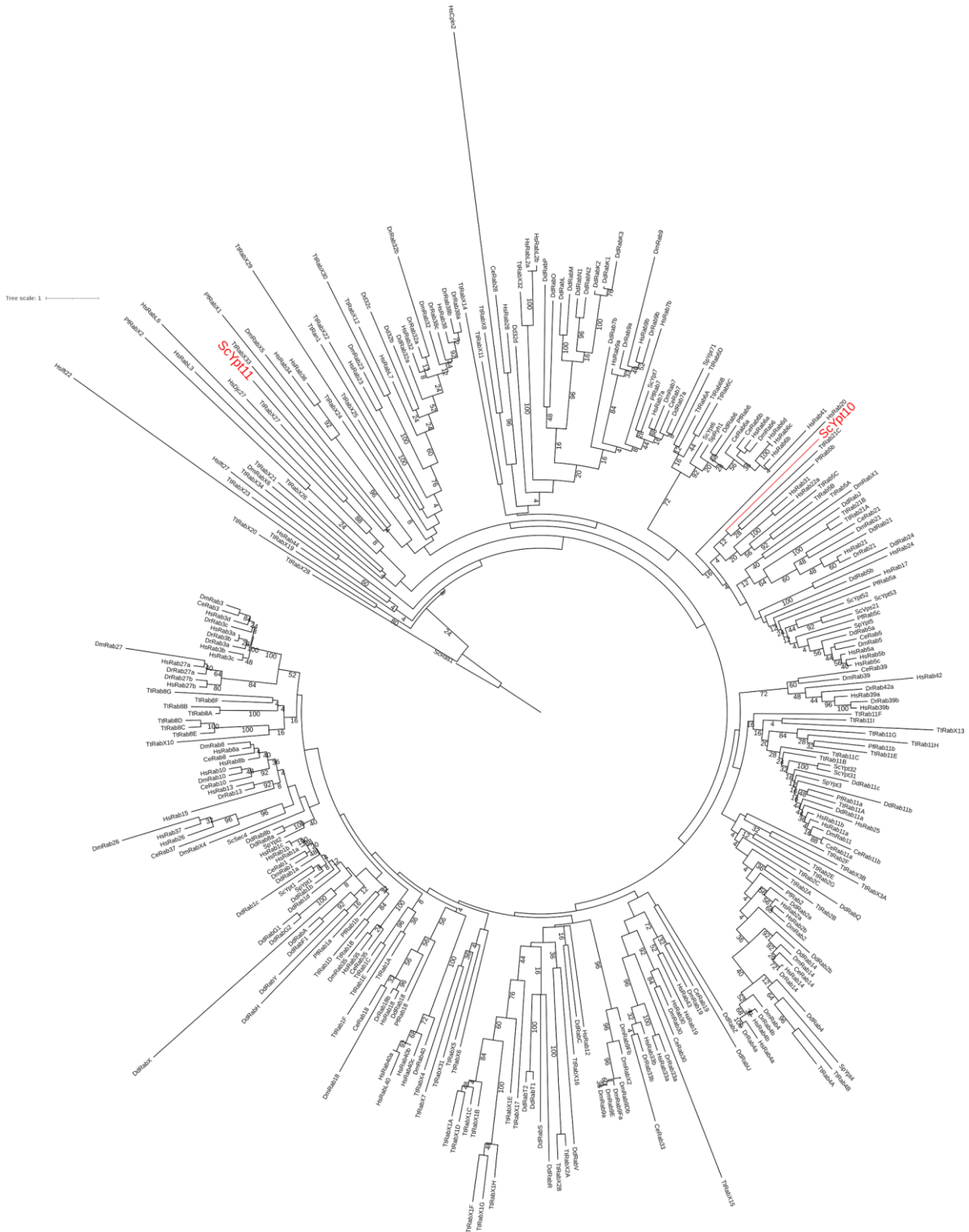
**Figure 4:** Enlarged region of the maximum-likelihood phylogram shown in Figure 3, highlighting the relationships of Rabs Ypt10 (top) and Ypt11 (bottom) along with their candidate human Rab homologs: Ypt10 – Rab22a, Rab31, Rab24; Ypt11: Rab34, Rab36.



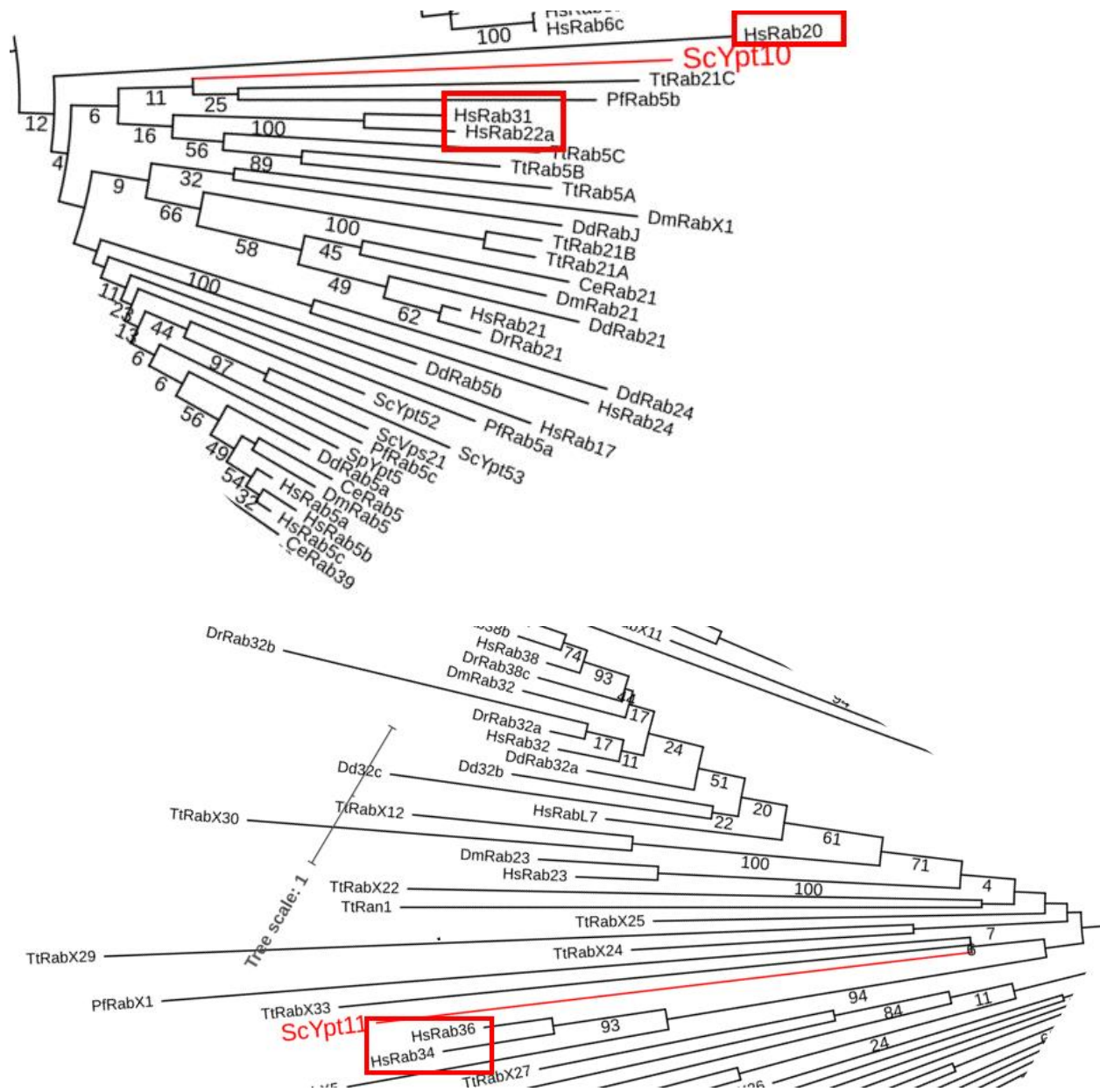
**Figure 5: Maximum-likelihood phylogram depicting the evolutionary relationships of Ras1 from *S. cerevisiae* and other Rabs.** Tree inference was performed using the same alignment as the results presented in Figure 3, but following elimination of poorly aligned regions using GBLOCKS.



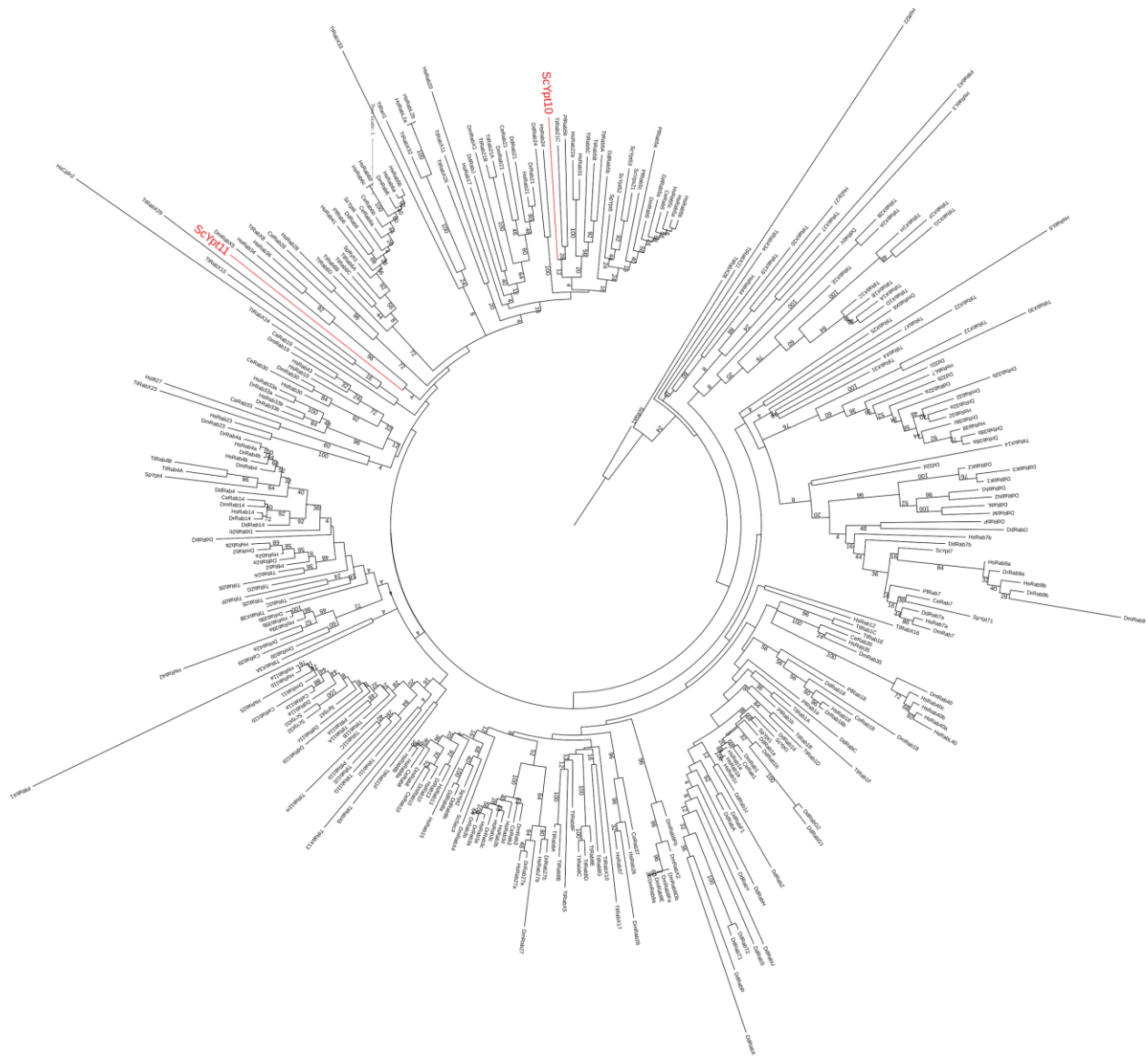
**Figure 6:** Enlarged region of the maximum-likelihood phylogram shown in Figure 5, highlighting the relationships of Rabs Ypt10 (top) and Ypt11 (bottom) along with their candidate human Rab homologs: Ypt10 – Rab34, Rab36; Ypt11: Rab8a/b, Rab10, Rab35.



**Figure 7: Maximum-likelihood phylogram depicting the evolutionary relationships of Ras1 from *S. cerevisiae* and other Rabs.** Tree inference based on HMM alignment, rooted on Ras1 (*S. cerevisiae*). Columns in the alignment with greater than 75% gap characters were removed. Numbers at nodes represent percent of 100 bootstrap replicates that recovered the same node.

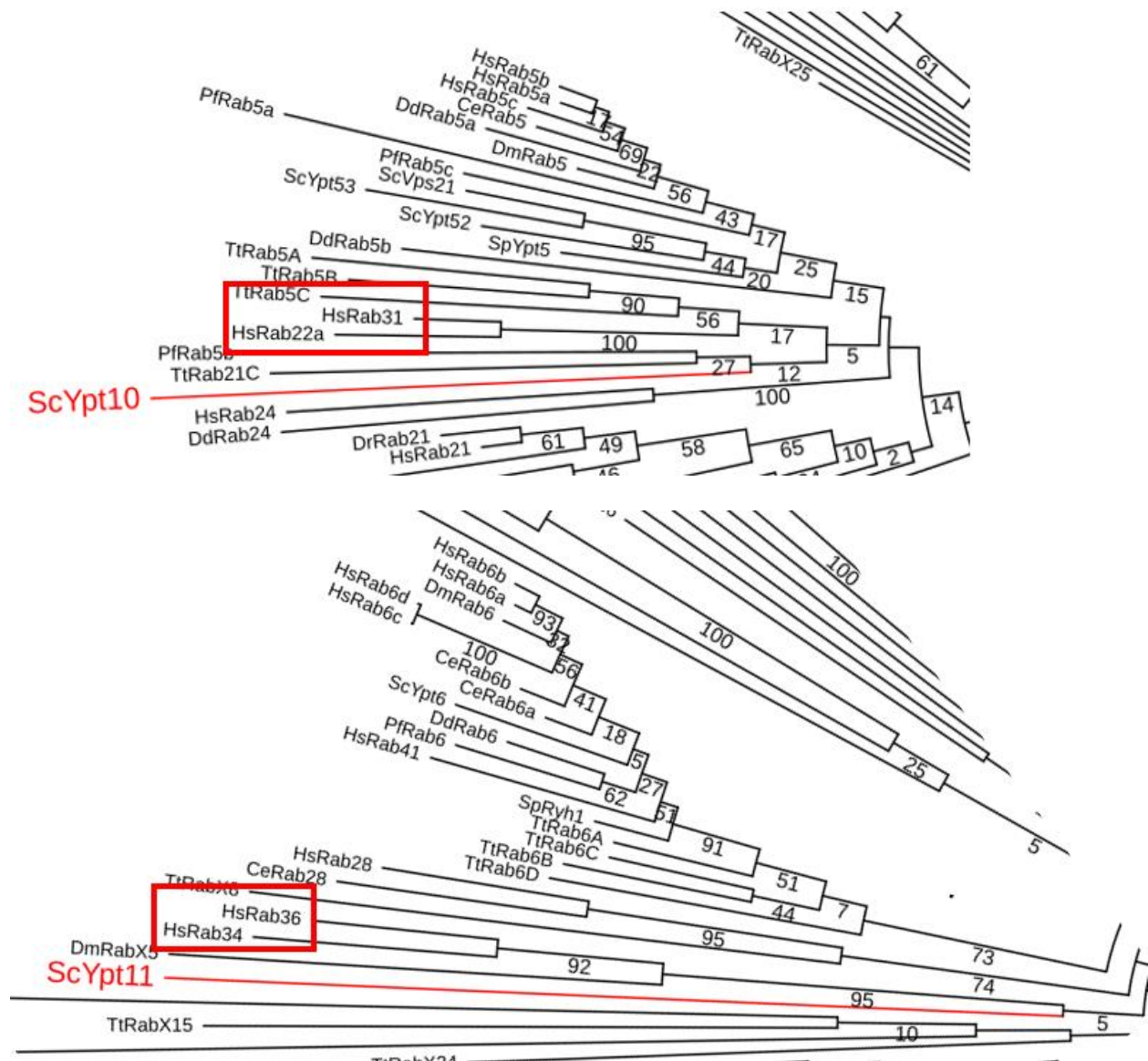


**Figure 8:** Enlarged region of the maximum-likelihood phylogram shown in Figure 7, highlighting the relationships of Rabs Ypt10 (top) and Ypt11 (bottom) along with their candidate human Rab homologs: Ypt10 – Rab22a, Rab31; Ypt11: Rab34, Rab36.

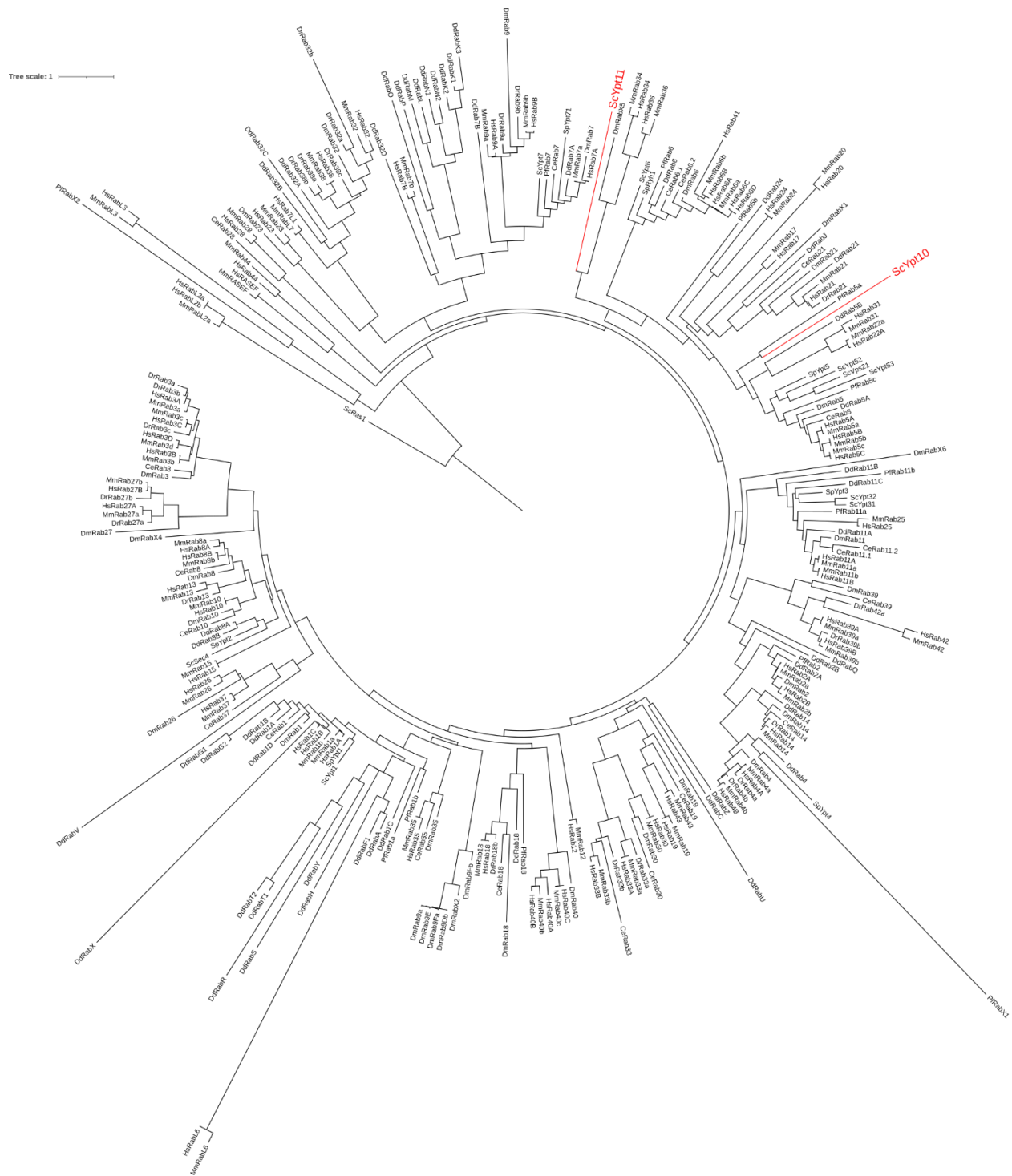


**Figure 9: Maximum-likelihood phylogram depicting the evolutionary relationships of Ras1 from *S. cerevisiae* and other Rabs.** Tree inference was based on MUSCLE alignment, rooted on Ras1 (*S. cerevisiae*). Columns in the alignment with greater than 75% gap characters were removed. Numbers at nodes represent percent of 100 bootstrap replicates that recovered the same node.



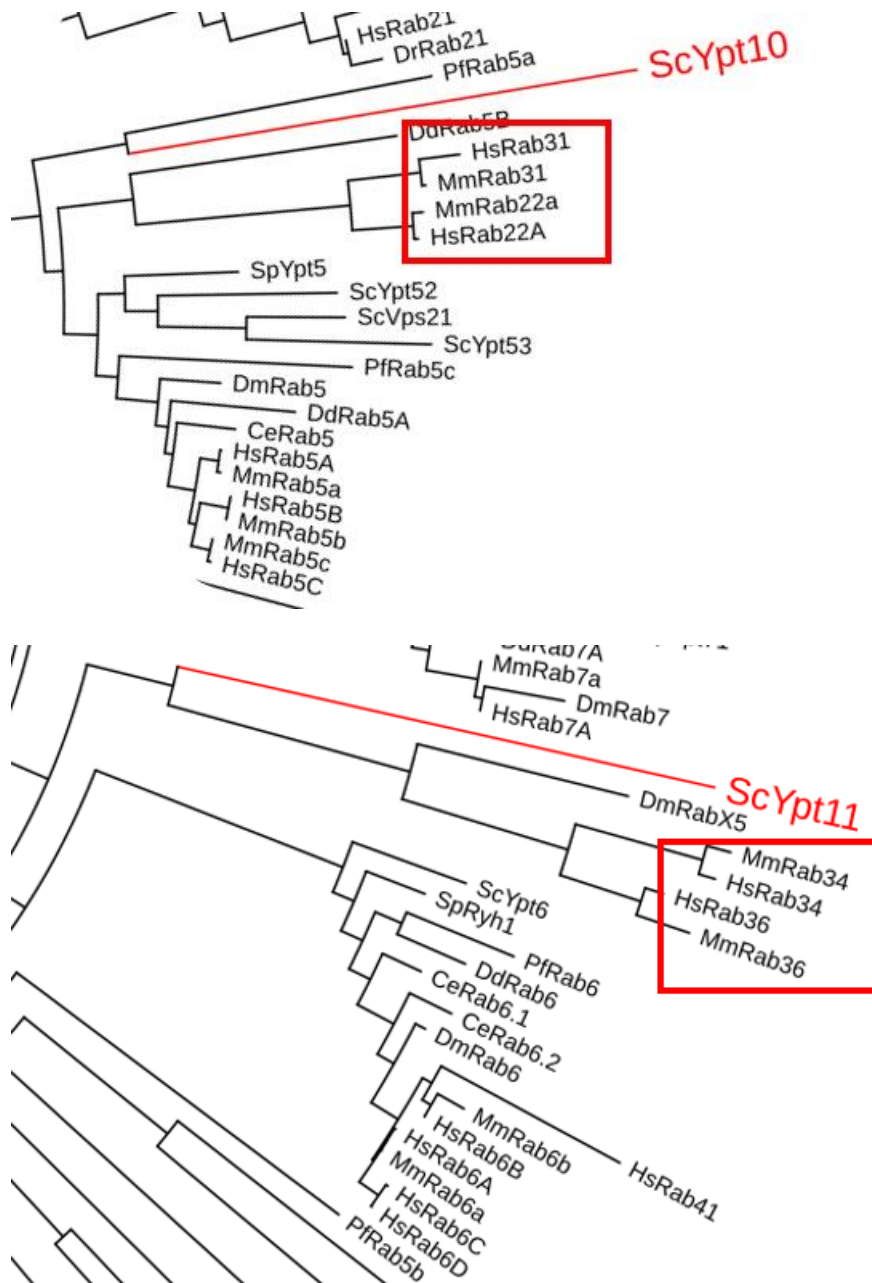


**Figure 10:** Enlarged region of the maximum-likelihood phylogram shown in Figure 9, highlighting the relationships of Rabs Ypt10 (top) and Ypt11 (bottom) along with their candidate human Rab homologs: Ypt10 – Rab22a, Rab31; Ypt11: Rab34, Rab36.



**Figure 11: Maximum-likelihood phylogram depicting the evolutionary relationships of Ras1 from *S. cerevisiae* and other Rabs.** Tree inference was based on MUSCLE alignment, rooted on Ras1 (*S. cerevisiae*). Columns in the alignment with greater than 75% gap characters were removed. Rab sequences from *T. thermophila* were removed, and *M. musculus* was added.





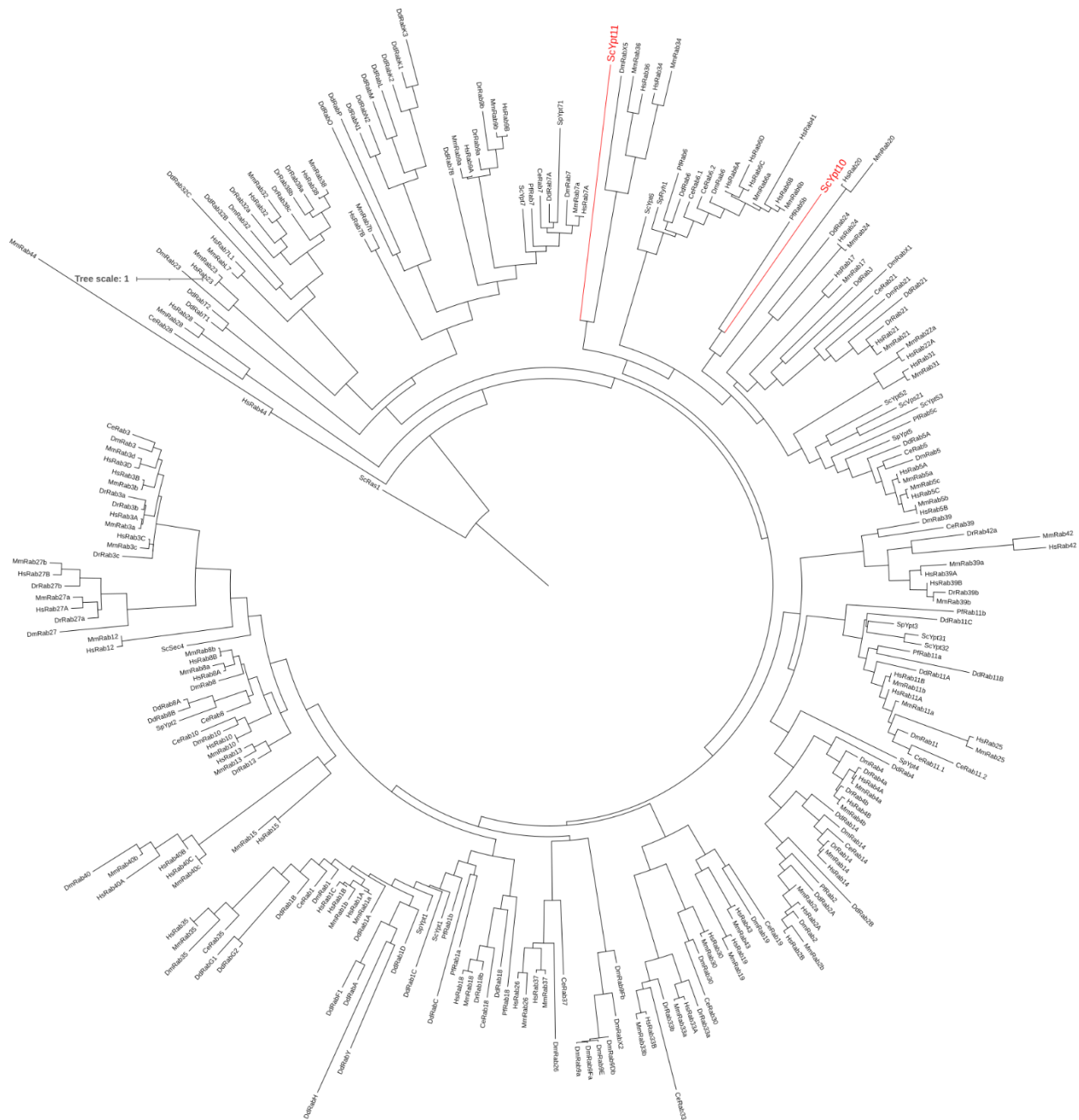
**Figure 12:** Enlarged region of the maximum-likelihood phylogram shown in Figure 11, highlighting the relationships of yeast Rabs Ypt10 (top) and Ypt11 (bottom) along with their candidate human Rab homologs: Ypt10 – Rab22a, Rab31; Ypt11: Rab34, Rab36.

The continuance of the long-isolated branches of our two proteins of interest raised concerns that perhaps they were not resolving well in the tree, thus we took a closer look at the clade

containing both Ypt10 and Ypt11, along with the other genes belonging to that clade (Figure 13) to determine if there were any distinct differences in the phylogeny with a significant reduction of taxa in the dataset. In the following tree (Figure 13), Ypt10 pairs closest to the Rab5 family, as well as Rab17. Ypt11 appears to be the most divergent taxa, but pairs closest to Rab34 and Rab36.



**Figure 13:** Tree created from a previous iteration (Figure 7) by selecting the entire clade that hosted our two genes of interest, with yeast Rabs Ypt10 and Ypt11 along with their candidate human Rab homologs: Ypt10 – Rab17, Rab5; Ypt11: Rab34, Rab36.



**Figure 14: Maximum-likelihood phylogram depicting the evolutionary relationships of Ras1 from *S. cerevisiae* and other Rabs.** Tree inference based on MUSCLE alignment, rooted on Ras1 (*S. cerevisiae*). Columns in the alignment with greater than 75% gap characters were removed, and rogue taxa removed using RogueNaRok



**Figure 15:** Enlarged region of the maximum-likelihood phylogram shown in Figure 14, highlighting the relationships of Rabs Ypt10 (top) and Ypt11 (bottom) along with their candidate human Rab homologs: Ypt10 – Rab20; Ypt11: Rab34, Rab36.

## Discussion

Five different phylogenetic analyses identified the same candidate homologs for Ypt10 and Ypt11, the exception being when the alignment was modified using Gblocks, which may be due to loss of phylogenetic signal from the significant alignment reduction by Gblocks (Table 1). In addition, inferences based on the alignments with “Rogue-taxa” removed also supported alternative arrangements (Table 1). Of the 9 out of 11 yeast Rabs with a previously known human homolog, all 9 known homologous pairs were recovered in each phylogenetic analyses. This was consistent compared to other studies with different methodology and taxa selection, suggesting our study was consistent with previous studies with regards to these Rabs (Stenmark and Olkkonen, 2001; Buvelot *et al.*, 2006; Klöpper *et al.*, 2012). This suggests that the robust taxon selection and phylogenetic inference method used here is appropriate for identifying yeast-human Rab homologs. Despite the different arrangement in the final analysis, I believe that the human homologs identified there are also worthy of consideration, as Aberer *et al.* (2013) have shown that their rogue-taxa removal algorithm, *RogueNaRok* (Aberer *et al.*, 2011), in conjunction with simulated data with known homologies, yielded more accurate phylogenies when rogue taxa were removed. The consistency of these results, as well as the recovery of known homologous pairs, leads to the conclusion that the candidate human homologs for Ypt10 are the common ancestor of human paralogs Rab22a and Rab31 or Rab24, and similarly for Ypt11, the candidate human homologs are Rab34 or Rab36.

### **Rab22a, candidate homolog of Ypt10 in humans**

Rab22a plays a role in endocytic and intracellular transport of proteins, localizing on early endosomes, regulating early endosomal sorting, and on recycling endosomes (Wang *et al.*, 2011; Patel *et al.*, 2021). Additionally, Wang *et al.* (2011) found that Rab22a was essential for

nerve growth factor (NGF)-induced neurite (axon-like extensions) outgrowth in PC12 cells (a cell line derived from rats that exhibits similar morphology as mature dopaminergic neurons), signal transduction, cell differentiation, and cell survival (Wiatrak *et al.*, 2020). This places an important role of Rab22a in the biogenesis and function of NGF-signaling endosomes that are important for development and survival of nerve cells (neurons) (Wang *et al.*, 2011).

### **Rab31, also known as Rab22b, as a candidate homolog of Ypt10**

Rab31, also known as Rab22b, shares over 70% sequence identity with the aforementioned Rab22a and is even referred to in the literature as Rab22b (Ng *et al.*, 2007; Ng *et al.*, 2009). Despite the high percentage identity shared between these two proteins, Ng *et al.* (2007) demonstrated that the two are functionally distinct, and that Rab31 localizes to and plays a role in function of the *trans*-Golgi network, while Rab22a localizes to early endosomes (Wang *et al.*, 2011). Multiple studies have shown that Rab31 plays a role in tumor development and progression (Pan *et al.*, 2015; Zhang *et al.*, 2018; Li *et al.*, 2021; Soelch *et al.*, 2021). Additionally, Ng *et al.* (2009) show that Rab31 is associated with epidermal growth factor receptor (EGFR) trafficking in some neuronal cell types. These roles imply Rab31 has many functions that are not isolated to just one specific type of cancer or mutation, but plays a broad role in the development of a disease phenotype.

### **Rab20 as a candidate homolog of Ypt10**

Using fluorescence microscopy, Sarma *et al.* (2008) showed that Rab20 localizes to the perinuclear region of the cell as well as the endoplasmic reticulum, the organelle affected by the disease mechanism in our ALS8 disease model (Nishimura *et al.*, 2005; Kabashi *et al.*, 2013). Some basic roles of Rab20, described by Seto *et al.* (2011), show that it is involved in

phagosome maturation and acidification of phagosomes that engulf pathogens. Rab20 has associations with various types of cancers and diseases; Schnettger *et al.* (2017) showed that a Rab20-dependent membrane trafficking pathway regulates *M. tuberculosis* replication, while Liu *et al.* (2021) showed that Rab20 mediates extracellular vesicles that are responsible for promoting hepatocarcinogenesis. Moreover, Amillet *et al.* (2006) showed that Rab20 is overexpressed in exocrine pancreatic carcinoma, and Torri *et al.* (2010) showed that Rab20 is one of a myriad of genes able to predict inflammatory signatures in dendritic cells.

### **Rab34 as a candidate homolog of Ypt11**

Rab34 is responsible for regulating the distribution of lysosomes, maturation of phagosomes, is involved in intra-Golgi protein transport, and localizes to the Golgi (Wang and Hong, 2002; Goldenberg *et al.*, 2007; Starling *et al.*, 2016). Additionally, Wang *et al.*, (2015) analyzed Rab34 expression in patients with low- and high-grade gliomas and found that Rab34 expression levels were related to glioma progression and had a significant effect on patient survival.

### **Rab36 as a candidate homolog of Ypt11**

Rab34 and Rab36 share an amino acid identity of 56% and have also been found to exhibit analogous functions (Chen *et al.*, 2010). To verify this identified homology, Chen and colleagues performed fluorescence microscopy using HeLa cells to determine if Rab34 and Rab36 shared similar localization and function and showed that Rab36 also localizes to and associates with the Golgi apparatus, and observed that late endosomes and lysosomes were distributed by Rab36, similar to Rab34. Similar to other Rabs involvement in cancer, Zhu *et al.* (2018) utilized a microRNA (miR-1247) that was reported to suppress tumors in multiple cancer

types, and found that downregulation of Rab36 mimicked the tumor suppressive effects of this microRNA. Upon further investigation, they found that miR-1247, the tumor suppressive microRNA, was found to target the untranslated region of Rab36, inhibiting Rab36 expression.

### **Insights on potential functions of uncharacterized proteins**

Ypt10 may be functioning similarly to Rab20, namely by localizing to the ER and through involvement in phagosomes. In light of the results that identified overexpression of *YPT10* as a possible suppressor of ALS8 disease phenotypes in a yeast model (D. Prosser, unpublished results), this makes sense: the phagosomes and closely-related endosomal structures may participate in lysosomal delivery of aggregated proteins, while the ER is the organelle affected in ALS8. Ypt11, whose function remains unclear, may be responsible for distribution of late endosomes and lysosomes in light of these results and the known functions of its candidate human homologs, Rab34 and Rab36 (Chen *et al.*, 2010; Zhu *et al.*, 2018).

A study conducted by Buvelot Frei *et al.*, (2006) observed the localization of Ypt10 and Ypt11 in yeast and HeLa cells using fluorescence microscopy which, in conjunction with our results, further narrows down the potential human homologs of Ypt10 and Ypt11. Buvelot Frei *et al.*, (2006) found that Ypt10 localized to endosomes, may be related to endocytic and vacuolar functions, and suggest homology with human Rab20. A multiple sequence alignment of Ypt10 and our candidate human homologs (Supplementary Figure S1) shows Rab20 has the lowest amino acid percent identity compared to Ypt10 (23.6%), versus Rab22a (31.7%) and Rab31 (33.2%).

Additionally, Buvelot Frei *et al.*, (2006) found that Ypt11 localized to the endoplasmic reticulum as well as the Golgi, which coincide with known localizations for Rab34 and Rab36 (Chen *et al.*,



2010). A multiple sequence alignment of Ypt11 and our candidate human homologs Rab34 and Rab36 (Supplementary Figure S2) shows both homologs have a very low amino acid percent identity compared to Ypt11 (16.9% for Rab34, 14.3% for Rab36). Despite this, the known functions of these human Rabs combined with the localization study conducted by Frei *et al.*, (2006) indicate that Ypt11 may be behaving similarly to our identified candidate human homologs.

### **Limitations of this study**

One limitation of this study is that some of the genes included in the phylogenetic analysis may not actually be Rabs. This was addressed in the methodology by removing genes that differed from other Rabs in their sequence and lacked formalized nomenclature (e.g. “RabX” or “Ift22”). Another limitation is that my study may have excluded Rabs that have not been formally classified. With the rise of sequencing technology, the number of Rabs discovered in humans has increased over the years, as Stenmark and Olkkonen (2001) identified 60 human Rab proteins; Colicelli (2004) describes there being 71 human Rabs; Korbeel and Freson (2008) state there are more than 60 human Rabs, but fail to state an exact number; Li and Marlin (2015) describe there being 66 Rabs in the human genome; Pfeffer (2017) states that there are at least 63 human Rabs, but also fail to state an exact number. Due to this uncertainty and dynamic evolution of identified human Rabs, an effort to consolidate and create a centralized repository for Rabs would be beneficial, especially considering the many implications of Rabs in diseases and cancers. A consensus for the identification of Rabs has been identified previously (Pereira-Leal and Seabra, 2001; Quevillon *et al.*, 2003) using RabF and RabSF motifs, sequences that are hallmark identifiers of Rabs; however, both studies state there are outliers that defy this motif

requirement in a marginal number Rabs, which may be a contributing factor to the dynamic evolution of the number of Rabs identified in humans.

### **Future directions**

Subsequent phylogenetic hypothesis testing would be a next approach to identify the best candidate human homolog, starting by altering tree topology to pair each candidate homolog with our gene of interest. This, in conjunction with implementation of an approximately unbiased test (following methods described in McCutcheon *et al.* (2019)), can determine which homologous pair and the corresponding phylogenetic arrangement has the highest likelihood score, thus narrowing down our candidate human homolog from three Rabs (in the instance of Ypt10) to just one (Huelsenbeck and Bull, 1996).

Wet-lab verification, using complementation studies, of the proposed homologs would be the next logical direction. One method would be to observe conservation of gene function: do our candidate human homologs recover the function of *ypt10Δ* and *ypt11Δ* in yeast? Does overexpression of the candidate human homolog to Ypt10 reduce ER sensitivity in a mammalian model of ALS8? The difficulty in this scenario is that *ypt10Δ* and *ypt11Δ* do not have obvious phenotypes, which renders a complementation assay useless unless a measurable phenotype is discovered. One potential route would be to analyze RNA-seq data and gene expression levels in response to each of the knockouts to see if there is correlation between the yeast and human genes. Another direction would be to follow in the footsteps of Oguchi *et al.*, (2018), who used PC12 cells, a cell line that mimics mature dopaminergic neurons (such as those affected by ALS), which are derived from a rat adrenal gland carcinoma. They measured the outgrowth of neurites (axon-line extensions from the soma) in response to overexpression and knockout of various Rab proteins, and determined whether specific Rabs had a negative, positive, or neutral

effect on neurite outgrowth, a simple yet measurable phenotype. A future step could entail applying a similar approach and measuring neurite outgrowth in response to overexpression and knockout of our candidate human homologs, and if expressing plasmids containing *YPT10* or *YPT11* show similar results. A similar response between potential homologs could indicate similar function, serving as further validation of homology.

If a viable complementation assay is discovered, or a neurite outgrowth assay or similar study confirms homology between the yeast and human genes, the next approach would be to replicate both the yeast (yALS8) and the mammalian (CHO cells or another cell line) disease models testing the ability of candidate human homologs to suppress disease phenotypes.

In the event that a Rab being overexpressed in a mammalian cell model reduces the disease phenotype, further research and development for ALS8 treatment could begin. However, one important implication should be considered, given that Rabs and their expression levels are involved in many different complex diseases and disorders: are there any unexpected ill effects of overexpressing or activating our candidate Rabs? Gene expression data in both a control and disease cell line before and after treatment with the suppressor Rab could tell us if the treatment may be worth pursuing or not.

## **Conclusion**

In conclusion, this study attempts to uncover the potential human homologs of *YTP10* and *YPT11*, in order to further our understanding of the suppressor gene of yALS8 with hopes of discovering novel routes of therapy, as well as to further characterize two of the most elusive yeast Rabs. By performing this phylogenetic analysis, it serves as a foundation for future research and uncovers the potential human homologs of *YTP10* and *YPT11*.

## Chapter 2: Global-scale louse endosymbiont genome variation

### Abstract

Human head lice rely on bacteria, *Candidatus* RIESIA, that are intracellular, heritable, and beneficial for survival and reproduction. The genome of *Ca.* RIESIA is small in size when compared to closely related bacteria, due a process of genome and proteome reduction, which was facilitated by metabolic complementation with lice. To understand the extent and the location of variation in the genome of *Ca.* RIESIA resulting from DNA substitutions, I identified sites in the genome that varied across 76 *Ca.* RIESIA samples. Data was stored in Variant Call Format files and custom Python scripts to quantify overall difference and examine the distribution of single base pair substitutions across the genome. High levels of variation were found at 190,000 bp from the 5' end of the genome, as well as highly conserved regions at 40,000 and 340,000 bp from the 5' end of the genome.

### Introduction

#### **The human head louse and its obligate intracellular endosymbiont's genome reduction**

Blood feeding lice rely on an intracellular symbiotic and heritable bacterium classified as *Candidatus* RIESIA pediculicola, herein endosymbiont, to provide metabolites required for louse development and reproduction (Perotti *et al.*, 2007; Kirkness *et al.*, 2010; Boyd *et al.*, 2017). The endosymbiont found in the human head louse possesses a small and AT-rich genome (genome size: 0.5Mb; AT%: 65), when compared to closely related non-endosymbiont bacterial species, such as *Escherichia coli* str. K-12 substr. MG1655 (Genome size: 4.6 Mb, AT %: 49.5; Rode *et al.*, 1999; Kirkness *et al.* 2010; Boyd *et al.*, 2014; Boyd *et al.* 2017). It is generally accepted that endosymbionts of insects start with a larger genome that is reduced as the insect

and bacteria coevolve (Kirkness *et al.*, 2010; McCutcheon and Moran, 2010; McCutcheon *et al.*, 2019). The Black Queen hypothesis, introduced by Morris *et al.* (2012), suggests that as the host louse provides numerous different metabolites, selective pressure for the endosymbiont to maintain many biosynthetic pathways is relaxed and the underlying genes are inactivated from the endosymbiont genome. The removal of inactivated genes reduces the endosymbiont genome size. This coincides with the proteome constraint theory, particularly in endosymbionts such as *Ca. Riesia*, which have relatively small and tightly packed genomes; as the selective pressure to maintain genes required for DNA replication and repair are reduced (Massey, 2008). Additionally, endosymbionts are asexual, have no opportunity for horizontal gene exchange, and undergo population bottlenecks at each host generation, a process that accelerates the fixation of new, potentially slightly deleterious substitutions in a phenomenon known as Muller's ratchet (Muller, 1964; Moran, 1996). Due to the reduction in genome space devoted to DNA replication and repair, along with Muller's ratchet, endosymbionts are expected to have a high rate of DNA substitution compared to closely related bacteria (Moran *et al.*, 2008). All these factors contribute to *Ca. Riesia* undergoing significant genome reduction compared to free-living bacteria. Given a worldwide sampling of endosymbiont genomes, the aim of this study was to determine how many sites in the genome varied relative to the reference genome, if the total number of differences were evenly distributed, or if there were areas of the genome that had seen more changes than other regions, including whether the variations were located within intra- or inter-genic regions of the genome.

## **Objective**

My objective is to identify differences in *Ca. Riesia* genome from whole genome sequence data using a reference genome, and to determine whether the genetic variation was within intergenic or intragenic regions of the genome.

## **Methods**

### **Sequence data pipeline**

Whole lice were collected and DNA was extracted and prepared for Next Generation Sequencing (NGS) by the Reed lab (University of Florida). Samples were sequenced using the short-read NovaSeq platform. Sequence reads were aligned to a reference genome (*Ca. Riesia pediculicola* str. USDA NCBI identifiers ASM9308v1, NC\_014109.1).

Alignment information was stored as a Sequence Alignment Map (SAM) file and transmitted to Virginia Commonwealth University. The SAM files were filtered for unaligned reads and converted to a binary format BAM. Single nucleotide differences were identified using the BAM file and used to create a Variant Call Format (VCF) file.

### **Parsing VCF files for variant positions**

I created a Python program to read in VCF files, converting the data into a 2D array before parsing the data. In VCF file data, a column contains each position in the genome along with an exact match to the reference genome (denoted as a “.”) or the actual nucleotide change. This “.” was set as the “match” variable and was used to parse the VCF file for columns that did not contain the match identifier.

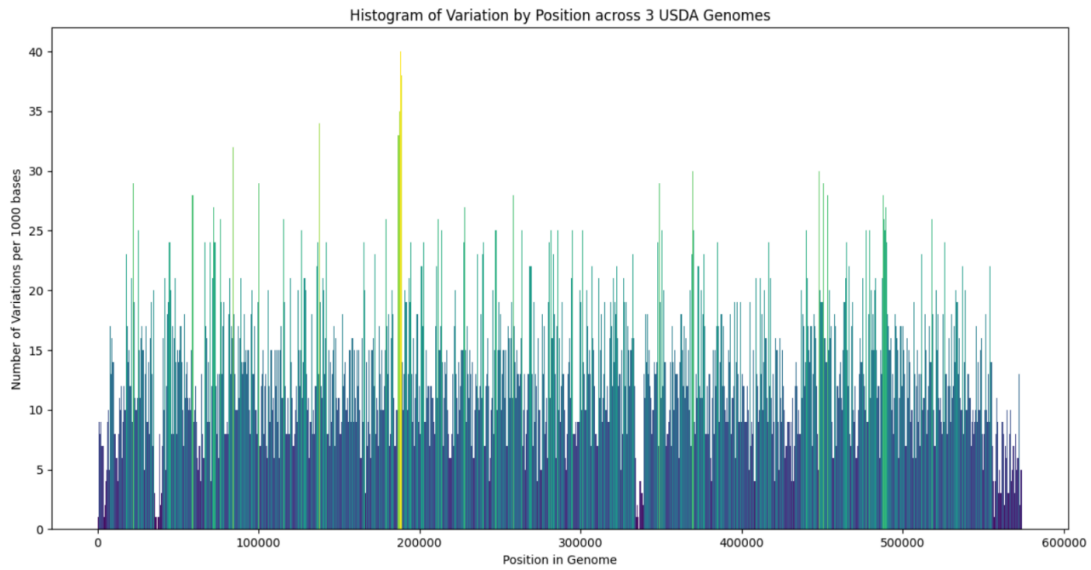
A dictionary data structure was selected to store both the positions (as the dictionary keys) and nucleotide variations (as the dictionary values); a list data structure was used to store the total number of positions of variance from the list of values in the dictionary. This list was used to create a position vs variation histogram to visualize the distribution of variation across the entire USDA genome (plotting the position in the genome (x-axis) against the number of variations per 100 nucleotide base pairs (y-axis)).

### **Parsing FASTA file of USDA genome to plot intergenic vs intragenic variation**

To determine whether the variations collected in the previous step belonged to intragenic or intergenic regions of the genome, I downloaded a FASTA file containing the complete genome of *Ca. Riesia*. I then constructed a regular expression search identifier to parse through the FASTA file of the complete genome to identify the numerical ranges of stop and start codons denoting the coding sequences (e.g. 4985..5470). The list of variants among the complete USDA genome (collected in the previous step) was then cross-referenced against this new list of coding sequences. This was used to create a position vs. variation histogram to visualize the distribution of variation among the intragenic and the intergenic regions of the genome (plotting the position in the genome (x-axis) against the number of variations per 100 nucleotide base pairs (y-axis)).

## **Results**

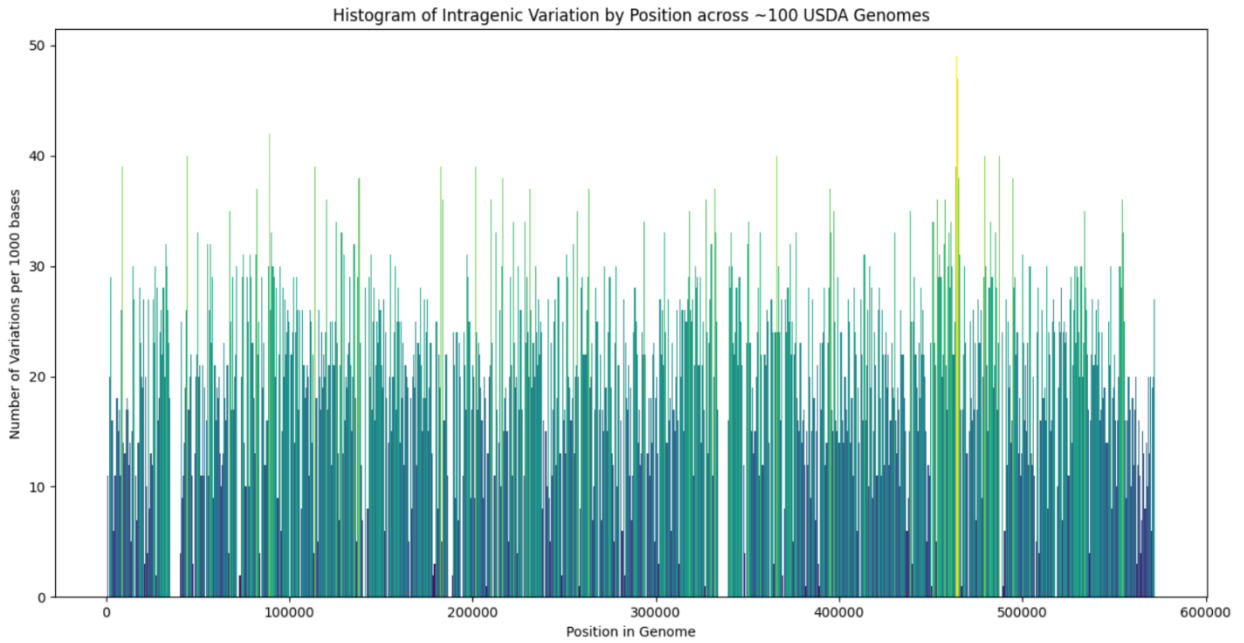
In a sample of 76 genomes, single nucleotide differences were noted throughout the genome when compared to the reference (Figure 16). There was a higher frequency of changes around ~190,000 bp from the 5' end, with fewer changes occurring in regions around ~40,000 and ~340,000 bp (Figure 16).



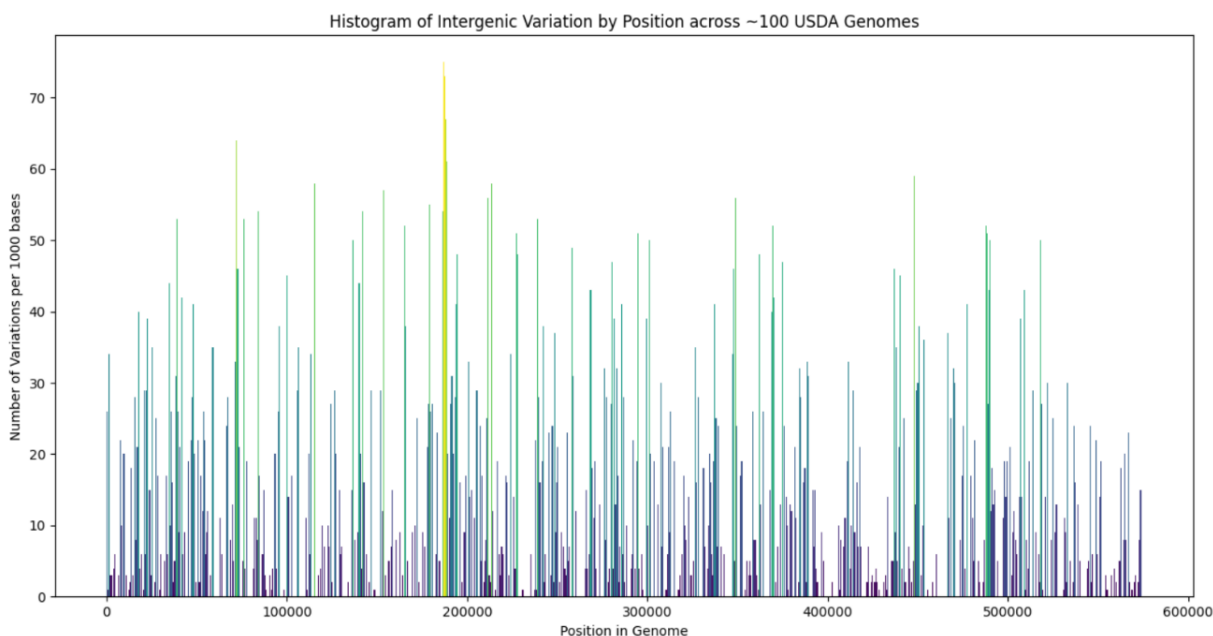
**Figure 16: Histogram of genetic variation by position for the USDA genome, using three VCF files.** The bars are heat-mapped to visually represent regions of high variation (yellow, green, light blue) versus regions of low variation (purple, dark blue).

Subsequent histograms were created to determine if the variations were primarily located within intragenic regions or intergenic regions (Figures 17 and 18, respectively). The intragenic regions display an increased level of genomic diversity at ~460,000 bp from the 5' end of the genome (Figure 17). In the intergenic regions, there was a spike in variation at ~190,000 bp from the 5' end of the genome.





**Figure 17: Histogram of intragenic variation by position for the USDA genome using over 100 VCF files.** The bars are heat-mapped to visually represent regions of high variation (yellow, green, light blue) versus regions of low variation (purple, dark blue).



**Figure 18: Histogram of intergenic variation by position for the USDA genome using over 100 VCF files.** The bars are heat-mapped to visually represent regions of high variation (yellow, green, light blue) versus regions of low variation (purple, dark blue).

## Discussion

Upon observation, we can see little to no variation in certain sites. According to the reference sequence of *Ca. Riesia*, there are two areas of the genome that each contain a copy of 16S rRNA (38,674-40,233 bp; 337,291-338,850bp) (figure 16), both of which fall within the highly conserved regions in the results. One region with high variation, around the 190,000 bp mark, falls within the *purB* gene (184835-186238bp), which encodes for adenylosuccinate lyase (ADL). ADL is necessary for the synthesis of adenosine monophosphate (AMP) and fumarate from adenylosuccinate and inactivation of the gene *purB* resulted in reduced growth in the closely related *E. coli* (Tsai *et al.*, 2007; Fyfe *et al.*, 2010; Jung *et al.*, 2010). Another region with high variation contained the gene *infA* (191989-192207bp), which encodes for Initiation

factor 1 (IF1), a protein that is required for initiating translation and has been found to be essential in *E. coli* (Dahlquist and Puglisi, 2000; Ko *et al.*, 2006). These results relate to Muller's ratchet as the isolated population size contributes to an acquisition and increase in deleterious mutations in a gene that is not essential and another that is essential in *E. coli* (both genes are present in the genomes of other sequenced *Ca. Riesia* species) (Muller, 1964; Moran, 1996), suggesting no gene is impervious to the impacts of Muller's ratchet. These results also suggest some regions of the genome are impacted more by DNA substitutions than others.

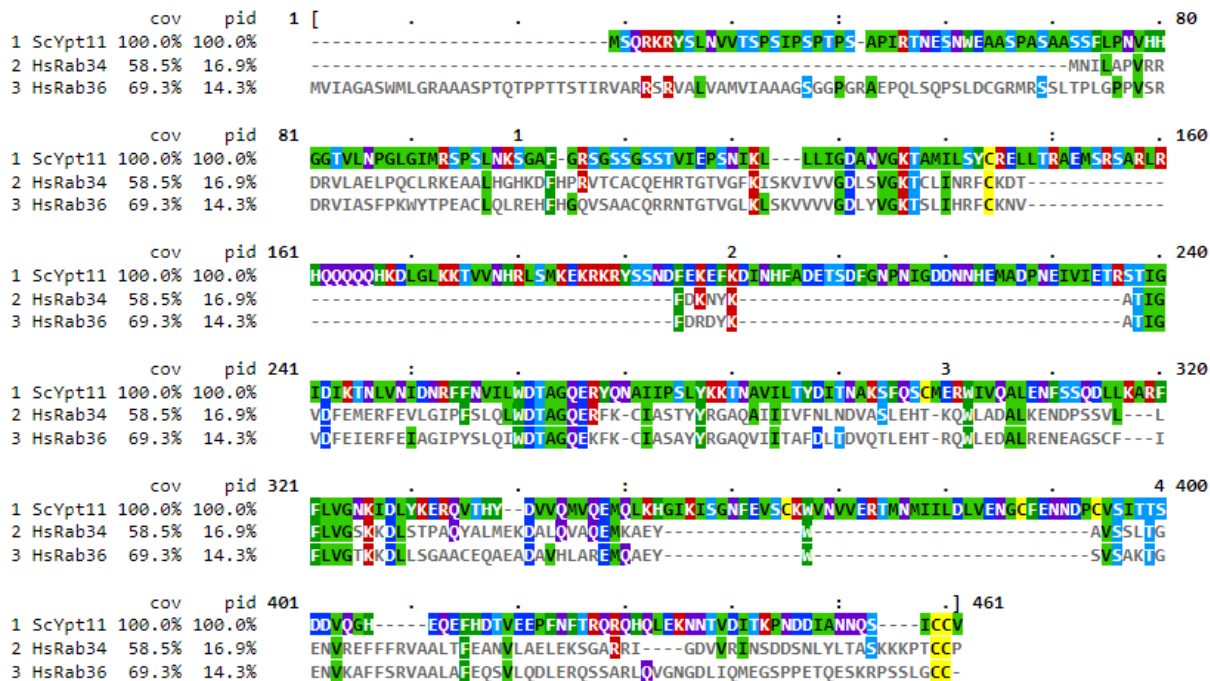
Another level of high variation occurs ~460,000bp within an intragenic region, with genes *murC* (454520-455977bp), *murG* (456015-457088bp), and *murD* (458224-459543bp); all of these protein-encoding genes involved in cell membrane synthesis which may be impacted due to being isolated within the host (Mengin-Lecreux *et al.*, 1991; Wachi *et al.*, 1999; Sink *et al.*, 2013). The high level of variation within these seemingly essential cell-membrane genes suggests they are nascent pseudogenes and the Black Queen hypothesis may provide some insights; since the endosymbiont lives comfortably within the cytoplasm of the louse host, it may no longer require a protective cell-wall as much as its free-living bacterial relatives. Thus, the rate of mutation of these genes increases in *Ca. Riesia*, which may lead to the eventual loss of these genes (Morris *et al.*, 2012, Derilus *et al.*, 2020).

## **Conclusion**

This global-scale study observing the variations of the endosymbiont *Ca. Riesia* reinforces the Black Queen hypothesis, Muller's ratchet, and the proteome constraint theory. Additionally, it serves as a foundation for looking into the specific genes undergoing high rates of mutation observed, and may lead to novel predictions on future gene loss.



Reference sequence (1): ScYpt11  
 Identities normalised by aligned length.  
 Colored by: identity



**Figure S2. Multiple sequence alignment of the Yeast protein Ypt11 and candidate human homologs Rab34 and Rab36.** Alignment created using MUSCLE (Edgar, 2004) and visualized using MView (Brown *et al.* 1998), with amino acid identity highlighted by color.

## References

- Aberer, A. J., Krompass, D., & Stamatakis, A. (2013). Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Systematic biology*, 62(1), 162–166. <https://doi.org/10.1093/sysbio/sys078>
- Aberer, Andre & Krompass, Denis & Stamatakis, Alexandros. (2011). RogueNaRok: An Efficient and Exact Algorithm for Rogue Taxon Identification. Heidelberg Institute for Theoretical Studies: Exelixis-RRDR-2011--10. 201.
- Aliaga L, Lai C, Yu J, Chub N, Shim H, Sun L, Xie C, Yang WJ, Lin X, O'Donovan MJ, Cai H. Amyotrophic lateral sclerosis-related VAPB P56S mutation differentially affects the function and survival of corticospinal and spinal motor neurons. *Hum Mol Genet*. 2013 Nov 1;22(21):4293-305. doi: 10.1093/hmg/ddt279. Epub 2013 Jun 13. Erratum in: *Hum Mol Genet*. 2014 Jun 1;23(11):3069. PMID: 23771029; PMCID: PMC3792689.
- Allen JM, Light JE, Perotti MA, Braig HR, Reed DL. Mutational meltdown in primary endosymbionts: selection limits Muller's ratchet. *PLoS One*. 2009;4(3):e4969. doi: 10.1371/journal.pone.0004969. Epub 2009 Mar 23. PMID: 19305500; PMCID: PMC2654755.
- Amillet, J.-M., Ferbus, D., Real, F. X., Antony, C., Muleris, M., Gress, T. M., & Goubin, G. (2006, January 24). *Characterization of human rab20 overexpressed in exocrine pancreatic carcinoma*. *Human Pathology*. Retrieved October 13, 2022, from <https://www.sciencedirect.com/science/article/pii/S0046817705006350?via%3Dihub>
- Andersson, S. G. E., & Kurland, C. G. (1998, September 21). *Reductive evolution of resident genomes*. *Trends in Microbiology*. Retrieved November 13, 2022, from <https://www.sciencedirect.com/science/article/abs/pii/S0966842X98013122?via%3Dihub>
- Boyd BM, Allen JM, Nguyen NP, Vachaspati P, Quicksall ZS, Warnow T, Mugisha L, Johnson KP, Reed DL. Primates, Lice and Bacteria: Speciation and Genome Evolution in the Symbionts of Hominid Lice. *Mol Biol Evol*. 2017 Jul 1;34(7):1743-1757. doi: 10.1093/molbev/msx117. PMID: 28419279; PMCID: PMC5455983.
- Bright, L. J., Kambesis, N., Nelson, S. B., Jeong, B., & Turkewitz, A. P. (2010). Comprehensive analysis reveals dynamic and evolutionary plasticity of Rab GTPases and membrane traffic in *Tetrahymena thermophila*. *PLoS genetics*, 6(10), e1001155. <https://doi.org/10.1371/journal.pgen.1001155>
- Brotman RG, Moreno-Escobar MC, Joseph J, et al. Amyotrophic Lateral Sclerosis. [Updated 2021 Aug 29]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2021 Jan-. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK556151/>
- Brown NP, Leroy C, Sander C. (1998). MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics*. 14(4):380-1. doi: 10.1093/bioinformatics/14.4.380. PMID: 9632837.

Buvelot Frei, S., Rahl, P. B., Nussbaum, M., Briggs, B. J., Calero, M., Janeczko, S., Regan, A. D., Chen, C. Z., Barral, Y., Whittaker, G. R., & Collins, R. N. (2006). Bioinformatic and comparative localization of Rab proteins reveals functional insights into the uncharacterized GTPases Ypt10p and Ypt11p. *Molecular and cellular biology*, 26(19), 7299–7317. <https://doi.org/10.1128/MCB.02405-05>

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, 25(15), 1972–1973. <https://doi.org/10.1093/bioinformatics/btp348>

Carlson M, Botstein D. (1982). Two differentially regulated mRNAs with different 5' ends encode secreted with intracellular forms of yeast invertase. *Cell*. 28(1):145-54. doi: 10.1016/0092-8674(82)90384-1. PMID: 7039847.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res. Jan;40(Database issue):D700-5. [PMID: 22110037]*

Colicelli J. Human RAS superfamily proteins and related GTPases. *Sci STKE*. 2004 Sep 7;2004(250):RE13. doi: 10.1126/stke.2502004re13. PMID: 15367757; PMCID: PMC2828947.

Corbeel L, Freson K. Rab proteins and Rab-associated proteins: major actors in the mechanism of protein-trafficking disorders. *Eur J Pediatr*. 2008 Jul;167(7):723-9. doi: 10.1007/s00431-008-0740-z. Epub 2008 May 8. PMID: 18463892; PMCID: PMC2413085.

Dahlquist KD, Puglisi JD. (2000). Interaction of translation initiation factor IF1 with the E. coli ribosomal A site. *J Mol Biol*. 299(1):1-15. doi: 10.1006/jmbi.2000.3672. PMID: 10860719.

Das Sarma J, Kaplan BE, Willemsen D, Koval M. Identification of rab20 as a potential regulator of connexin 43 trafficking. *Cell Commun Adhes*. 2008 May;15(1):65-74. doi: 10.1080/15419060802014305. PMID: 18649179.

Delwiche, C. (n.d.). Analytical methods: Maximum Likelihood. Retrieved January 9, 2022, from <https://science.umd.edu/labs/delwiche/MSyst/lec/Likelihood-1.html>

Derilus, D., Rahman, M.Z., Pinero, F. *et al.* Synergism between the Black Queen effect and the proteomic constraint on genome size reduction in the photosynthetic picoeukaryotes. *Sci Rep* 10, 8918 (2020). <https://doi.org/10.1038/s41598-020-65476-1>

Diekmann, Y., Seixas, E., Gouw, M., Tavares-Cadete, F., Seabra, M. C., & Pereira-Leal, J. B. (2011). Thousands of rab GTPases for the cell biologist. *PLoS computational biology*, 7(10), e1002217. <https://doi.org/10.1371/journal.pcbi.1002217>

Edgar R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5), 1792–1797. <https://doi.org/10.1093/nar/gkh340>

Edor Kabashi, Hajer El Oussini, Valérie Bercier, François Gros-Louis, Paul N. Valdmanis, Jonathan McDearmid, Inge A. Meijer, Patrick A. Dion, Nicolas Dupre, David Hollinger, Jérôme Sinniger, Sylvie Dirrig-Grosch, William Camu, Vincent Meininger, Jean-Philippe Loeffler, Frédérique René, Pierre Drapeau, Guy A. Rouleau, Luc Dupuis, Investigating the contribution of *VAPB/ALS8* loss of function in amyotrophic lateral sclerosis, *Human Molecular Genetics*, Volume 22, Issue 12, 15 June 2013, Pages 2350–2360

Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci U S A*. 1996 Nov 12;93(23):13429–34. doi: 10.1073/pnas.93.23.13429. PMID: 8917608; PMCID: PMC24110.

Eisenmann, D. M., Wnt signaling (June 25, 2005), *WormBook*, ed. The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1.7.1, <http://www.wormbook.org>.

Felsenstein J. CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. *Evolution*. 1985 Jul;39(4):783–791. doi: 10.1111/j.1558-5646.1985.tb00420.x. PMID: 28561359.

Felsenstein J. Phylogenies from molecular sequences: inference and reliability. *Annu Rev Genet*. 1988;22:521–65. doi: 10.1146/annurev.ge.22.120188.002513. PMID: 3071258.

Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol* 17, 368–376 (1981). <https://doi.org/10.1007/BF01734359>

Fyfe PK, Dawson A, Hutchison MT, Cameron S, Hunter WN. (2010). Structure of *Staphylococcus aureus* adenylosuccinate lyase (PurB) and assessment of its potential as a target for structure-based inhibitor discovery. *Acta Crystallogr D Biol Crystallogr*. 2010 Aug;66(Pt 8):881–8. doi: 10.1107/S0907444910020081. PMID: 20693687; PMCID: PMC2917274.

Gaut BS, Lewis PO. Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol Biol Evol*. 1995 Jan;12(1):152–62. doi: 10.1093/oxfordjournals.molbev.a040183. PMID: 7877489.

Goldenberg, N. M., Grinstein, S., & Silverman, M. (2007, December). *Golgi-bound rab34 is a novel member of the secretory pathway*. *Molecular biology of the cell*. Retrieved October 19, 2022, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2096593/>  
Goloboff PA, Szumik CA. Identifying unstable taxa: Efficient implementation of triplet-based measures of stability, and comparison with Phyutility and RogueNaRok. *Mol Phylogenet Evol*. 2015 Jul;88:93–104. doi: 10.1016/j.ympev.2015.04.003. Epub 2015 Apr 10. PMID: 25865266.

Grosshans, B., Ortiz, D., Novick, P. (2006). Rabs and their effectors: Achieving specificity in membrane traffic. *Proceedings of the National Academy of Sciences of the United States of America*, 103 (32) 11821–11827. <https://doi.org/10.1073/pnas.0601617103>



Guadagno, N. A., & Progida, C. (2019). Rab GTPases: Switching to Human Diseases. *Cells*, 8(8), 909. <https://doi.org/10.3390/cells8080909>

Hillis, D. M., & Bull, J. J. (1993, June 1). *Empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis*. OUP Academic. Retrieved October 27, 2022, from <https://academic.oup.com/sysbio/article-abstract/42/2/182/1730933?login=true>

Ho, S. (2008) The molecular clock and estimating species divergence. *Nature Education* 1(1):168

Homma Y, Hiragi S, Fukuda M. Rab family of small GTPases: an updated view on their regulation and functions. *FEBS J.* 2021 Jan;288(1):36-55. doi: 10.1111/febs.15453. Epub 2020 Jul 1. PMID: 32542850; PMCID: PMC7818423.

Hutagalung, A. H., & Novick, P. J. (2011). Role of Rab GTPases in membrane traffic and cell physiology. *Physiological reviews*, 91(1), 119–149. <https://doi.org/10.1152/physrev.00059.2009>

Itay Mayrose, Nir Friedman, Tal Pupko, A Gamma mixture model better accounts for among site rate heterogeneity, *Bioinformatics*, Volume 21, Issue suppl\_2, , Pages ii151–ii158, <https://doi.org/10.1093/bioinformatics/bti1125>

J. Castresana, Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis, *Molecular Biology and Evolution*, Volume 17, Issue 4, April 2000, Pages 540–552, <https://doi.org/10.1093/oxfordjournals.molbev.a026334>

John P. Huelsenbeck, J. J. Bull, A Likelihood Ratio Test to Detect Conflicting Phylogenetic Signal, *Systematic Biology*, Volume 45, Issue 1, March 1996, Pages 92–98, <https://doi.org/10.1093/sysbio/45.1.92>

Johnson, L.S., Eddy, S.R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* 11, 431 (2010). <https://doi.org/10.1186/1471-2105-11-431>

José B. Pereira-Leal, Miguel C. Seabra (2001). Evolution of the rab family of small GTP-binding proteins 11 Edited by J. Thornton, *Journal of Molecular Biology*, Volume 313, Issue 4, 2001, Pages 889-901, ISSN 0022-2836, <https://doi.org/10.1006/jmbi.2001.5072>

Jung SC, Smith CL, Lee KS, Hong ME, Kweon DH, Stephanopoulos G, Jin YS. (2010). Restoration of growth phenotypes of Escherichia coli DH5alpha in minimal media through reversal of a point mutation in purB. *Appl Environ Microbiol.* 2010 Sep;76(18):6307-9. doi: 10.1128/AEM.01210-10. PMID: 20675450; PMCID: PMC2937491.

Kanekura, K., Nishimoto, I., Aiso, S., & Matsuoka, M. (2006, August 4). *Characterization of amyotrophic lateral sclerosis-linked P56s mutation of vesicle-associated membrane protein-associated protein B (VAPB/ALS8)*. *Journal of Biological Chemistry*. Retrieved October 26,

2022, from

<https://www.sciencedirect.com/science/article/pii/S0021925819339274?via%3Dihub>

Kiral FR, Kohrs FE, Jin EJ, Hiesinger PR. (2018). Rab GTPases and Membrane Trafficking in Neurodegeneration. *Curr Biol*. 28(8):R471-R486. doi: 10.1016/j.cub.2018.02.010. PMID: 29689231; PMCID: PMC5965285.

Kirkness, E. F., Haas, B. J., Sun, W., & Pittendrigh, B. R. (2010, April 14). Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *pnas.org*. Retrieved November 10, 2022, from <https://pnas.org/doi/full/10.1073/pnas.1000699107>

Klöpffer, T.H., Kienle, N., Fasshauer, D. *et al.* Untangling the evolution of Rab G proteins: implications of a comprehensive genomic analysis. *BMC Biol* 10, 71 (2012). <https://doi.org/10.1186/1741-7007-10-71>

Ko JH, Lee SJ, Cho B, Lee Y. (2006). Differential promoter usage of *infA* in response to cold shock in *Escherichia coli*. *FEBS Lett*. 580(2):539-44. doi: 10.1016/j.febslet.2005.12.066. Epub 2005 Dec 28. PMID: 16405963.

Larkin A, Marygold SJ, Antonazzo G, Attrill H, dos Santos G, Garapati PV, Goodman JL, Gramates LS, Millburn G, Strelets VB, Tabone CJ, Thurmond J and the FlyBase Consortium (2021) FlyBase: updates to the *Drosophila melanogaster* knowledge base. [Nucleic Acids Res. 49\(D1\) D899–D907](https://doi.org/10.1093/nar/nkab100)

Leavitt R, Schlesinger S, Kornfeld S. Tunicamycin inhibits glycosylation and multiplication of Sindbis and vesicular stomatitis viruses. *J Virol*. 1977 Jan;21(1):375-85. doi: 10.1128/JVI.21.1.375-385.1977. PMID: 189071; PMCID: PMC353824.

Li Chen, Jingjie Hu, Ye Yun & Tuanlao Wang (2010) Rab36 regulates the spatial distribution of late endosomes and lysosomes through a similar mechanism to Rab34, *Molecular Membrane Biology*, 27:1, 23-30, DOI: [10.3109/09687680903417470](https://doi.org/10.3109/09687680903417470)

Li G. (2011). Rab GTPases, membrane trafficking and diseases. *Curr Drug Targets*. 12(8):1188-93. doi: 10.2174/138945011795906561. PMID: 21561417; PMCID: PMC4260923.

Li, G., & Marlin, M. C. (2015). Rab family of GTPases. *Methods in molecular biology (Clifton, N.J.)*, 1298, 1–15. [https://doi.org/10.1007/978-1-4939-2569-8\\_1](https://doi.org/10.1007/978-1-4939-2569-8_1)

Li, X., Zhu, F., Liu, Z., Tang, X., Han, Y., Jiang, J., Ma, C., & He, Y. (2021, March). *High expression of rab31 confers a poor prognosis and enhances cell proliferation and invasion in oral squamous cell carcinoma*. National Library of Medicine. Retrieved October 13, 2022, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7859975/>

Liu, B. H. M., Tey, S. K., Mao, X., Ma, A. P. Y., Yeung, C. L. S., Wong, S. W. K., Ng, T. H., Xu, Y., Yao, Y., Fung, E. Y. M., Tan, K. V., Khong, P.-L., Ho, D. W.-H., Ng, I. O.-L., Tang, A. H. N., Cai, S. H., Yun, J. P., & Yam, J. W. P. (2021, August). *TPII-reduced*

*extracellular vesicles mediated by RAB20 downregulation promotes aerobic glycolysis to drive hepatocarcinogenesis*. Journal of extracellular vesicles. Retrieved October 13, 2022, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8357635/>

Mackiewicz, P., Wyroba, E. Phylogeny and evolution of Rab7 and Rab9 proteins. *BMC Evol Biol* 9, 101 (2009). <https://doi.org/10.1186/1471-2148-9-101>

Mai U, Mirarab S. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics*. 2018 May 8;19(Suppl 5):272. doi: 10.1186/s12864-018-4620-2. PMID: 29745847; PMCID: PMC5998883.

McCutcheon JP, Moran NA. Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. *Genome Biol Evol*. 2010;2:708-18. doi: 10.1093/gbe/evq055. Epub 2010 Sep 9. PMID: 20829280; PMCID: PMC2953269.

McCutcheon, J. P., Boyd, B. M., & Dale, C. (2019, June 3). *The life of an insect endosymbiont from the cradle to the grave*. *Current Biology*. Retrieved October 28, 2022, from <https://www.sciencedirect.com/science/article/pii/S0960982219303306#app2>

Mengin-Lecreulx D, Texier L, Rousseau M, van Heijenoort J. The murG gene of *Escherichia coli* codes for the UDP-N-acetylglucosamine: N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase involved in the membrane steps of peptidoglycan synthesis. *J Bacteriol*. 1991 Aug;173(15):4625-36. doi: 10.1128/jb.173.15.4625-4636.1991. PMID: 1649817; PMCID: PMC208138.

Merlie JP, Sebbane R, Tzartos S, Lindstrom J. Inhibition of glycosylation with tunicamycin blocks assembly of newly synthesized acetylcholine receptor subunits in muscle cells. *J Biol Chem*. 1982 Mar 10;257(5):2694-701. PMID: 7061443.

Michael P. Cummings, Scott A. Handley, Daniel S. Myers, David L. Reed, Antonis Rokas, Katarina Winka, Comparing Bootstrap and Posterior Probability Values in the Four-Taxon Case, *Systematic Biology*, Volume 52, Issue 4, 1 August 2003, Pages 477–487, <https://doi.org/10.1080/10635150390218213>

Moran NA. Microbial minimalism: genome reduction in bacterial pathogens. *Cell*. 2002 Mar 8;108(5):583-6. doi: 10.1016/s0092-8674(02)00665-7. PMID: 11893328.

Moran NA. Tracing the evolution of gene loss in obligate bacterial symbionts. *Curr Opin Microbiol*. 2003 Oct;6(5):512-8. doi: 10.1016/j.mib.2003.08.001. PMID: 14572545.

Moran, N. A., McCutcheon, J. P., & Nakabachi, A. (2008, December). *Genomics and evolution of heritable bacterial symbionts*. *Annualreviews.org*. Retrieved November 15, 2022, from <https://www.annualreviews.org/doi/10.1146/annurev.genet.41.110306.130119>  
Moran, N. Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *PNAS*, Volume 93, Issue 7, April 1996, Pages 2873-2878, <https://doi.org/10.1073/pnas.93.7.2873>

Morris JJ, Lenski RE, Zinser ER. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *mBio*. 2012 May 2;3(2):e00036-12. doi: 10.1128/mBio.00036-12. PMID: 22448042; PMCID: PMC3315703.

MULLER HJ. THE RELATION OF RECOMBINATION TO MUTATIONAL ADVANCE. *Mutat Res.* 1964 May;106:2-9. doi: 10.1016/0027-5107(64)90047-8. PMID: 14195748.

Munjal, G., Hanmandlu, M., & Srivastava, S. (2018). Phylogenetics Algorithms and Applications. *Ambient Communications and Computer Systems: RACCCS-2018*, 904, 187–194. [https://doi.org/10.1007/978-981-13-5934-7\\_17](https://doi.org/10.1007/978-981-13-5934-7_17)

N Saitou, M Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees., *Molecular Biology and Evolution*, Volume 4, Issue 4, Jul 1987, Pages 406–425, <https://doi.org/10.1093/oxfordjournals.molbev.a040454>

Nakamichi S, Yamanaka K, Suzuki M, Watanabe T, Kagiwada S. Human VAPA and the yeast VAP Scs2p with an altered proline distribution can phenocopy amyotrophic lateral sclerosis-associated VAPB(P56S). *Biochem Biophys Res Commun.* 2011 Jan 14;404(2):605-9. doi: 10.1016/j.bbrc.2010.12.011. Epub 2010 Dec 7. PMID: 21144830.

Ng EL, Ng JJ, Liang F, Tang BL. Rab22B is expressed in the CNS astroglia lineage and plays a role in epidermal growth factor receptor trafficking in A431 cells. *J Cell Physiol.* 2009 Dec;221(3):716-28. doi: 10.1002/jcp.21911. PMID: 19725050.

Ng, E. L., Wang, Y., & Tang, B. L. (2007, July 30). *Rab22b's role in trans-golgi network membrane dynamics*. *Biochemical and Biophysical Research Communications*. Retrieved October 13, 2022, from <https://www.sciencedirect.com/science/article/pii/S0006291X07015367?via%3Dihub>

Nishimura, A.L., Al-Chalabi, A. & Zatz, M. A common founder for amyotrophic lateral sclerosis type 8 (ALS8) in the Brazilian population. *Hum Genet* 118, 499–500 (2005). <https://doi.org/10.1007/s00439-005-0031-y>

Oguchi ME, Etoh K, Fukuda M. Rab20, a novel Rab small GTPase that negatively regulates neurite outgrowth of PC12 cells. *Neurosci Lett.* 2018 Jan 1;662:324-330. doi: 10.1016/j.neulet.2017.10.056. Epub 2017 Oct 28. PMID: 29107708.

Pan, Y., Zhang, Y., Chen, L. *et al.* The Critical Role of Rab31 in Cell Proliferation and Apoptosis in Cancer Progression. *Mol Neurobiol* 53, 4431–4437 (2016). <https://doi.org/10.1007/s12035-015-9378-9>

Patel NM, Siva MSA, Kumari R, Shewale DJ, Rai A, Ritt M, Sharma P, Setty SRG, Sivaramakrishnan S, Soppina V. KIF13A motors are regulated by Rab22A to function as weak dimers inside the cell. *Sci Adv.* 2021 Feb 3;7(6):eabd2054. doi: 10.1126/sciadv.abd2054. PMID: 33536208; PMCID: PMC7857691.

Perotti, M.A., Allen, J.M., Reed, D.L. and Braig, H.R. (2007), Host-symbiont interactions of the primary endosymbiont of human head and body lice. *FASEB J*, 21: 1058-1066. <https://doi.org/10.1096/fj.06-6808com>

Pervez, M. T., Babar, M. E., Nadeem, A., Aslam, M., Awan, A. R., Aslam, N., Hussain, T., Naveed, N., Qadri, S., Waheed, U., & Shoaib, M. (2014). Evaluating the accuracy and efficiency of multiple sequence alignment methods. *Evolutionary bioinformatics online*, 10,

205–217. <https://doi.org/10.4137/EBO.S19199>

Pfeffer SR. Rab GTPases: master regulators that establish the secretory and endocytic pathways. *Mol Biol Cell*. 2017 Mar 15;28(6):712-715. doi: 10.1091/mbc.E16-10-0737. PMID: 28292916; PMCID: PMC5349778.

Posada, D., & Crandall, K. A. (2021). Felsenstein Phylogenetic Likelihood. *Journal of molecular evolution*, 89(3), 134–145. <https://doi.org/10.1007/s00239-020-09982-w>

Prosser, D. C., Tran, D., Gougeon, P. Y., Verly, C., & Ngsee, J. K. (2008). FFAT rescues VAPA-mediated inhibition of ER-to-Golgi transport and VAPB-mediated ER aggregation. *Journal of cell science*, 121(Pt 18), 3052–3061. <https://doi.org/10.1242/jcs.028696>

Quevillon E, Spielmann T, Brahimi K, Chattopadhyay D, Yeramian E, Langsley G. The Plasmodium falciparum family of Rab GTPases. *Gene*. 2003 Mar 13;306:13-25. doi: 10.1016/s0378-1119(03)00381-0. PMID: 12657463.

Rode CK, Melkerson-Watson LJ, Johnson AT, Bloch CA. Type-specific contributions to chromosome size differences in Escherichia coli. *Infect Immun*. 1999 Jan;67(1):230-6. doi: 10.1128/IAI.67.1.230-236.1999. PMID: 9864220; PMCID: PMC96301.

S. Kumar, G. Stecher, M. Suleski, and S.B. Hedges, 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution* 34: 1812-1819, DOI: 10.1093/molbev/msx116.

Schnettger, L., Rodgers, A., Repnik, U., Lai, R. P., Pei, G., Verdoes, M., Wilkinson, R. J., Young, D. B., & Gutierrez, M. G. (2017, May 10). *A rab20-dependent membrane trafficking pathway controls M. tuberculosis replication by regulating phagosome spaciousness and integrity*. *Cell host & microbe*. Retrieved October 13, 2022, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5432432/>

Seto S, Tsujimura K, Koide Y. Rab GTPases regulating phagosome maturation are differentially recruited to mycobacterial phagosomes. *Traffic*. 2011 Apr;12(4):407-20. doi: 10.1111/j.1600-0854.2011.01165.x. Epub 2011 Feb 21. PMID: 21255211.

Šink, Roman, Barreteau, Hélène, Patin, Delphine, Mengin-Lecreulx, Dominique, Gobec, Stanislav and Blantot, Didier. "MurD enzymes: some recent developments" *BioMolecular Concepts*, vol. 4, no. 6, 2013, pp. 539-556. <https://doi.org/10.1515/bmc-2013-0024>

Soelch S, Beaufort N, Loessner D, Kotzsch M, Reuning U, Luther T, Kirchner T, Magdolen V. Rab31-dependent regulation of transforming growth factor  $\beta$  expression in breast cancer cells. *Mol Med*. 2021 Dec 14;27(1):158. doi: 10.1186/s10020-021-00419-8. PMID: 34906074; PMCID: PMC8670132.

Soltis, D. E., & Soltis, P. S. (2003). The role of phylogenetics in comparative genetics. *Plant physiology*, 132(4), 1790–1800. <https://doi.org/10.1104/pp.103.022509>

Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, 30(9), 1312–1313.



<https://doi.org/10.1093/bioinformatics/btu033>

Starling GP, Yip YY, Sanger A, Morton PE, Eden ER, Dodding MP. Folliculin directs the formation of a Rab34-RILP complex to control the nutrient-dependent dynamic distribution of lysosomes. *EMBO Rep.* 2016 Jun;17(6):823-41. doi: 10.15252/embr.201541382. Epub 2016 Apr 13. PMID: 27113757; PMCID: PMC4893818.

Stenmark, H., & Olkkonen, V. M. (2001). The Rab GTPase family. *Genome biology*, 2(5), REVIEWS3007.

<https://doi.org/10.1186/gb-2001-2-5-reviews3007>

Steven E. Massey, The Proteomic Constraint and Its Role in Molecular Evolution, *Molecular Biology and Evolution*, Volume 25, Issue 12, December 2008, Pages 2557–2565,

<https://doi.org/10.1093/molbev/msn210>

Sunderland, Mary E., "Dictyostelium discoideum". *Embryo Project Encyclopedia* (2009-06-10). ISSN: 1940-5030 <http://embryo.asu.edu/handle/10776/1792>.

Suzuki, H., Kanekura, K., Levine, T.P., Kohno, K., Olkkonen, V.M., Aiso, S. and Matsuoka, M. (2009), ALS-linked P56S-VAPB, an aggregated loss-of-function mutant of VAPB, predisposes motor neurons to ER stress-related death by inducing aggregation of co-expressed wild-type VAPB. *Journal of Neurochemistry*, 108: 973-985.

<https://doi.org/10.1111/j.1471-4159.2008.05857.x>

Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol.* 2007 Aug;56(4):564-77. doi: 10.1080/10635150701472164. PMID: 17654362.

Teuling E, Ahmed S, Haasdijk E, Demmers J, Steinmetz MO, Akhmanova A, Jaarsma D, Hoogenraad CC. Motor neuron disease-associated mutant vesicle-associated membrane protein-associated protein (VAP) B recruits wild-type VAPs into endoplasmic reticulum-derived tubular aggregates. *J Neurosci.* 2007 Sep 5;27(36):9801-15. doi: 10.1523/JNEUROSCI.2661-07.2007. PMID: 17804640; PMCID: PMC6672975.

The UniProt Consortium

**UniProt: the universal protein knowledgebase in 2021**

[Nucleic Acids Res. 49:D1 \(2021\)](https://doi.org/10.1093/nar/nkab036)

Tokutake Y, Yamada K, Ohata M, Obayashi Y, Tsuchiya M, Yonekura S. ALS-Linked P56S-VAPB Mutation Impairs the Formation of Multinuclear Myotube in C2C12 Cells. *Int J Mol Sci.* 2015 Aug 10;16(8):18628-41. doi: 10.3390/ijms160818628. PMID: 26266407; PMCID: PMC4581263.

Torri A, Beretta O, Ranghetti A, Granucci F, Ricciardi-Castagnoli P, Foti M. Gene expression profiles identify inflammatory signatures in dendritic cells. *PLoS One.* 2010 Feb 24;5(2):e9404. doi: 10.1371/journal.pone.0009404. Erratum in: *PLoS One.* 2010;5(6). doi: 10.1371/annotation/53736770-ad30-4c6b-8279-d344a1232cc6. PMID: 20195376; PMCID: PMC2827557.

Tsai M, Koo J, Yip P, Colman RF, Segall ML, Howell PL. (2007). Substrate and product complexes of Escherichia coli adenylosuccinate lyase provide new insights into the enzymatic mechanism. *J Mol Biol.* 370(3):541-54. doi: 10.1016/j.jmb.2007.04.052. Epub 2007 May 4. PMID: 17531264; PMCID: PMC4113493.

Wachi M, Wijayarathna CD, Teraoka H, Nagai K. A murC gene from coryneform bacteria. *Appl Microbiol Biotechnol.* 1999 Feb;51(2):223-8. doi: 10.1007/s002530051385. PMID: 10091329.

Wang L, Liang Z, Li G. Rab22 controls NGF signaling and neurite outgrowth in PC12 cells. *Mol Biol Cell.* 2011 Oct;22(20):3853-60. doi: 10.1091/mbc.E11-03-0277. Epub 2011 Aug 17. PMID: 21849477; PMCID: PMC3192864.

Wang, HJ., Gao, Y., Chen, L. *et al.* RAB34 was a progression- and prognosis-associated biomarker in gliomas. *Tumor Biol.* 36, 1573–1578 (2015). <https://doi.org/10.1007/s13277-014-2732-0>

Wang, T., & Hong, W. (2002, December). *Interorganellar regulation of lysosome positioning by the golgi apparatus through rab34 interaction with rab-interacting lysosomal protein.* *Molecular biology of the cell.* Retrieved October 19, 2022, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC138636/>

Wernegreen JJ. Endosymbiosis: lessons in conflict resolution. *PLoS Biol.* 2004 Mar;2(3):E68. doi: 10.1371/journal.pbio.0020068. Epub 2004 Mar 16. PMID: 15024418; PMCID: PMC368163.

Wiatrak, B., Kubis-Kubiak, A., Piwowar, A., & Barg, E. (2020). PC12 Cell Line: Cell Types, Coating of Culture Vessels, Differentiation and Other Culture Conditions. *Cells*, 9(4), 958. <https://doi.org/10.3390/cells9040958>

Wolfe, K., Shields, D. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708–713 (1997). <https://doi.org/10.1038/42711>

Woollard A. Gene duplications and genetic redundancy in *C. elegans*. 2005 Jun 25. In: *WormBook: The Online Review of C. elegans Biology* [Internet]. Pasadena (CA): WormBook; 2005-2018. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK19659/>

Zahraoui A, Touchot N, Chardin P, Tavitian A. The human Rab genes encode a family of GTP-binding proteins related to yeast YPT1 and SEC4 products involved in secretion. *J Biol Chem.* 1989 Jul 25;264(21):12394-401. PMID: 2501306.

Zhang Y, Yang B, Cheng X, Liu L, Zhu Y, Gong Y, Yang Y, Tian J, Peng X, Zou D, Yang L, Mei S, Wang X, Lou J, Ke J, Li J, Gong J, Chang J, Yuan P, Zhong R. Integrative functional genomics identifies regulatory genetic variant modulating RAB31 expression and altering susceptibility to breast cancer. *Mol Carcinog.* 2018 Dec;57(12):1845-1854. doi: 10.1002/mc.22902. Epub 2018 Sep 19. PMID: 30182384.

Zhu Y, Liang S, Pan H, Cheng Z, Rui X. Inhibition of miR-1247 on cell proliferation and invasion in bladder cancer through its downstream target of RAB36. *J Biosci.* 2018 Jun;43(2):365-373. PMID: 29872024.

ZUCKERKANDL, E. (1965). Evolutionary Divergence and Convergence in Proteins. In *Evolving Genes and Proteins* (pp. 97–166). <https://doi.org/10.1016/B978-1-4832-2734-4.50017-6>