Misannotations and Data Sources

Amelia Harrison

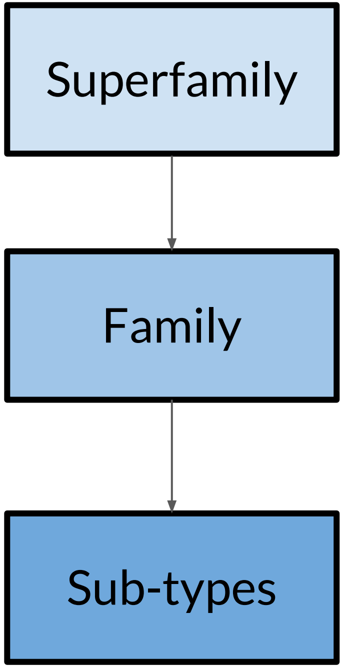
Presented: April 21, 2022

What is a misannotation?

- A misannotation is an incorrect annotation
    - Can occur in any dataset at any level
- Most common type of misannotation is overprediction
    - Annotation of a protein at a level that is not supported by the data
- How do we usually annotate proteins or genes?

Misannotations are common in databases

- A 2009 study estimated the misannotation of 37 protein families (6 superfamilies) in SwissProt, GenBank, TrEMBL, and KEGG
    - Misannotation estimates for the superfamilies ranged from 5% to 63%
    - 10 of the 37 families had misannotation rates > 80%
    - SwissProt consistently had the lowest misannotation rates
- How does SwissProt differ from the other databases?



Example protein hierarchy

Two types of databases or database entries

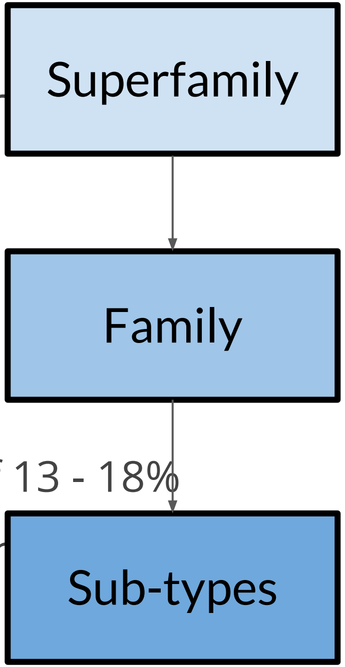
- Reviewed
  - Database entries are manually reviewed and curated by experts
  - SwissProt, NCBI RefSeq, SILVA, PDB, ChEMBL
- Unreviewed
  - Annotations provided by researchers and are not reviewed
  - TrEMBL, GenBank, KEGG

Terms used to describe protein annotations

- Biochemically characterized
  - Annotation(s) based on biochemical experiments
- Automatic annotation
  - Annotation based the result of a bioinformatic method alone
  - Usually refers to annotations based on sequence similarity
- Manual annotation
  - An annotation made by a person
    - Generally involves the use of bioinformatic tools/methods
      - Sequence similarity, HMMs, alignments, structural similarity

Another study of misannotation in a database

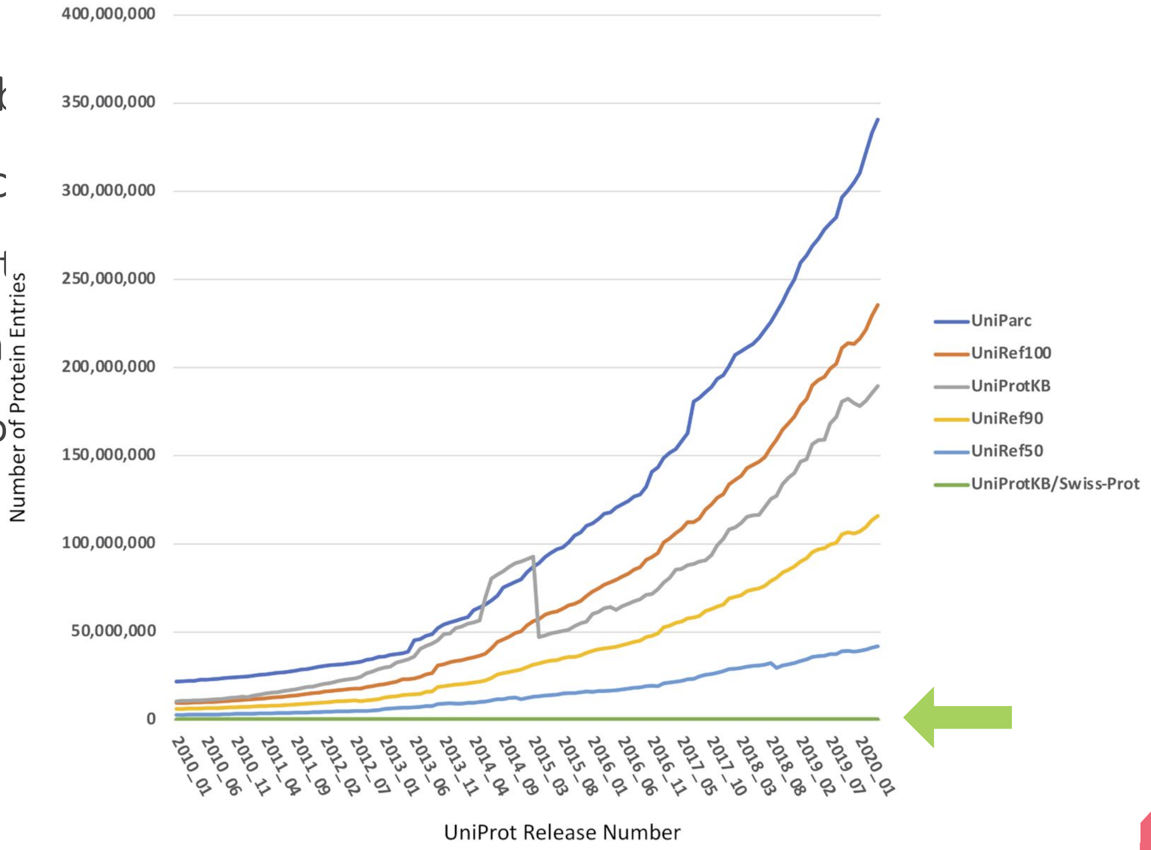
- Another study estimated error in Gene Ontology (GO) sequen
annotations
    - Total error was 28% - 30%
    - Sequences annotated based on sequence similarity alone had a
49%
    - Sequences annotated using other methods had an error rate of 13 - 18%
- Remember, this does not mean that 49% of sequences are an
the wrong protein entirely
    - Most of these are overannotations



Example protein hierarchy

Why are misannotation rates so high?

- The number of proteins sub since the databases were c
  - Lower cost and larger out
- More proteins are being an
  - Used to characterize all p

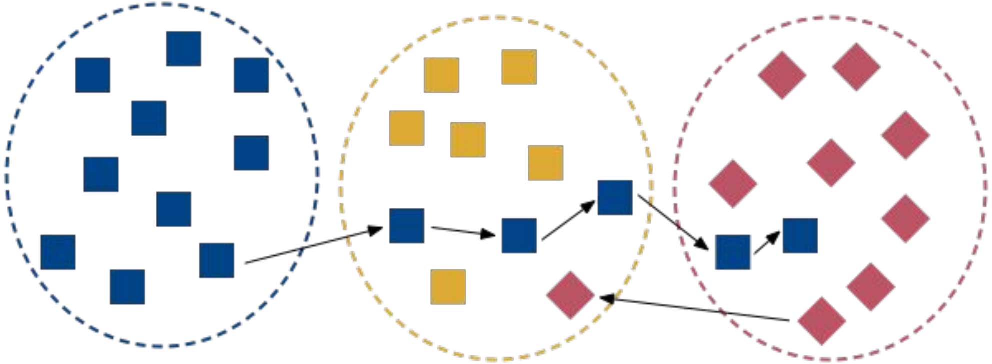


Why else are misannotations so high?

- Individual datasets are large
  - Databases are not the only ones overwhelmed by the amount of data
- Annotating proteins can be confusing
  - Interpretation of ontologies, naming schemes, etc. is tricky for non-experts
- Trouble tracking provenance
  - Databases do not know which annotations rely on others
  - Therefore, a single misannotation can lead to many misannotations, but it’s correction does not necessarily lead to many corrections
- Conflicting advice regarding annotation of sequences
  - Some recommend considering surrounding genes, others advise against it

Dangers of misannotation

- Cause researchers to draw false conclusions, or hides discoveries
- Error percolation (chains of misannotation)
    - In a modeled database with annotations made based on sequence similarity alone, the database eventually lost all ability to differentiate proteins

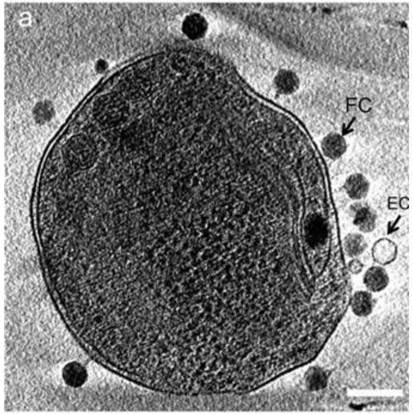


Adapted from Gilks et al. 2002

A misannotation in a cyanophage

Annotation of a viral genome

- *Prochlorococcus* phage P-SSP7
  - dsDNA virus first described in 2005
  - Infects the marine cyanobacterium *Prochlorococcus marinus*
- One of the first marine virus genomes to be sequenced and annotated
  - Only 9 marine viral genomes available at the time
- Marine cyanobacteria are B_{12} producers



Murata et al. 2017

B_{12} synthesis

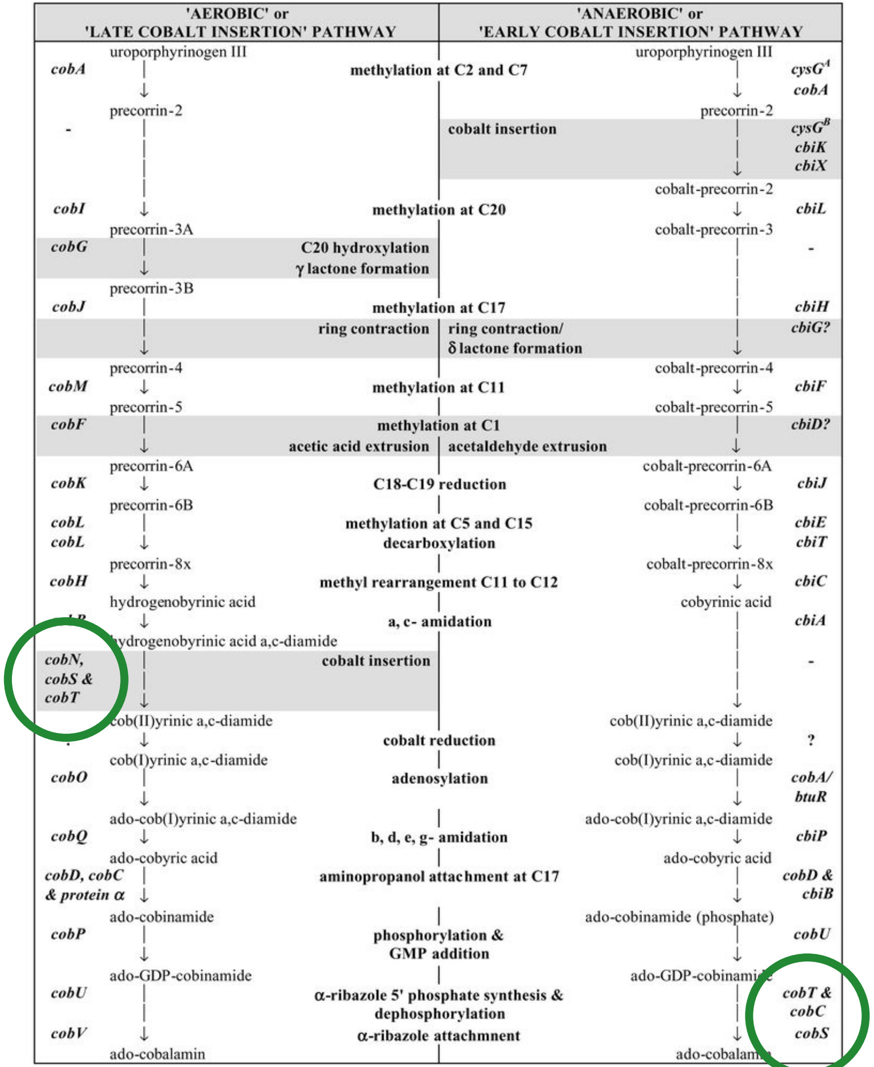
- There are only a handful of B_{12} producers in the oceans
  - Cyanobacteria, *Thaumarchaeota*
  - Everyone else imports B_{12}
    - Some do contain partial pathways

A B_{12} synthesis gene in the P-SSP7 genome

- One of the protein sequences showed similarity to the gene *cobS*
    - Encodes an enzyme involved in B_{12} synthesis
- Both a surprising and unsurprising discovery:
    - **Unsurprising** because the host cyanobacteria produces B_{12}
    - **Unsurprising** because they also found a B_{12}-dependent protein in the genome
        - Class II RNR (ribonucleotide reductase)
    - **Surprising** because the viral *cobS* was dissimilar to the cyanobacterial *cobS*

B_{12} synthesis

- There are only a handful of B_{12} producers in the oceans
    - Cyanobacteria, *Thaumarchaeota*
    - Everyone else imports B_{12}
        - Some do contain partial pathways
- Two B_{12} production pathways
    - Aerobic (late cobalt insertion) pathway
    - Anaerobic (early cobalt insertion) pathway
        - Cyanobacteria use this one (cyanobacteria pre-date the oxygenated atmosphere)



'AEROBIC' or 'LATE COBALT INSERTION' PATHWAY			'ANAEROBIC' or 'EARLY COBALT INSERTION' PATHWAY	
	uroporphyrinogen III		uroporphyrinogen III	
cobA	↓	methylation at C2 and C7	↓	*cysG*^{A} *cobA*
	precorrin-2		precorrin-2	
-		cobalt insertion	↓	*cysG*^{B} *cbiK* *cbiX*
			cobalt-precorrin-2	
cobI	↓	methylation at C20	↓	*cbiL*
	precorrin-3A		cobalt-precorrin-3	
cobG	↓	C20 hydroxylation γ lactone formation		-
	precorrin-3B			
cobJ		methylation at C17		*cbiH*
	↓	ring contraction ring contraction/ δ lactone formation	↓	*cbiG?*
	precorrin-4		cobalt-precorrin-4	
cobM	↓	methylation at C11	↓	*cbiF*
	precorrin-5		cobalt-precorrin-5	
cobF	↓	methylation at C1 acetic acid extrusion acetaldehyde extrusion	↓	*cbiD?*
	precorrin-6A		cobalt-precorrin-6A	
cobK	↓	C18-C19 reduction	↓	*cbiJ*
	precorrin-6B		cobalt-precorrin-6B	
cobL *cobL*	↓	methylation at C5 and C15 decarboxylation	↓	*cbiE* *cbiT*
	precorrin-8x		cobalt-precorrin-8x	
cobH	↓	methyl rearrangement C11 to C12	↓	*cbiC*
	hydrogenobyrinic acid		cobyrinic acid	
cobB	↓	a, c- amidation		*cbiA*
	hydrogenobyrinic acid a,c-diamide			
cobN, *cobS &* *cobT*	↓	cobalt insertion	↓	-
	cob(II)yrinic a,c-diamide		cob(II)yrinic a,c-diamide	
-	↓	cobalt reduction	↓	?
	cob(I)yrinic a,c-diamide		cob(I)yrinic a,c-diamide	
cobO	↓	adenosylation	↓	*cobA/* *btuR*
	ado-cob(I)yrinic a,c-diamide		ado-cob(I)yrinic a,c-diamide	
cobQ	↓	b, d, e, g- amidation	↓	*cbiP*
	ado-cobyric acid		ado-cobyric acid	
cobD, cobC *& protein α*	↓	aminopropanol attachment at C17	↓	*cobD &* *cbiB*
	ado-cobinamide		ado-cobinamide (phosphate)	
cobP	↓	phosphorylation & GMP addition	↓	*cobU*
	ado-GDP-cobinamide		ado-GDP-cobinamide	
cobU		α-ribazole 5' phosphate synthesis & dephosphorylation		*cobT &* *cobC*
cobV	↓	α-ribazole attachmnent	↓	*cobS*
	ado-cobalamin		ado-cobalamin	

B$_{12}$ pathways

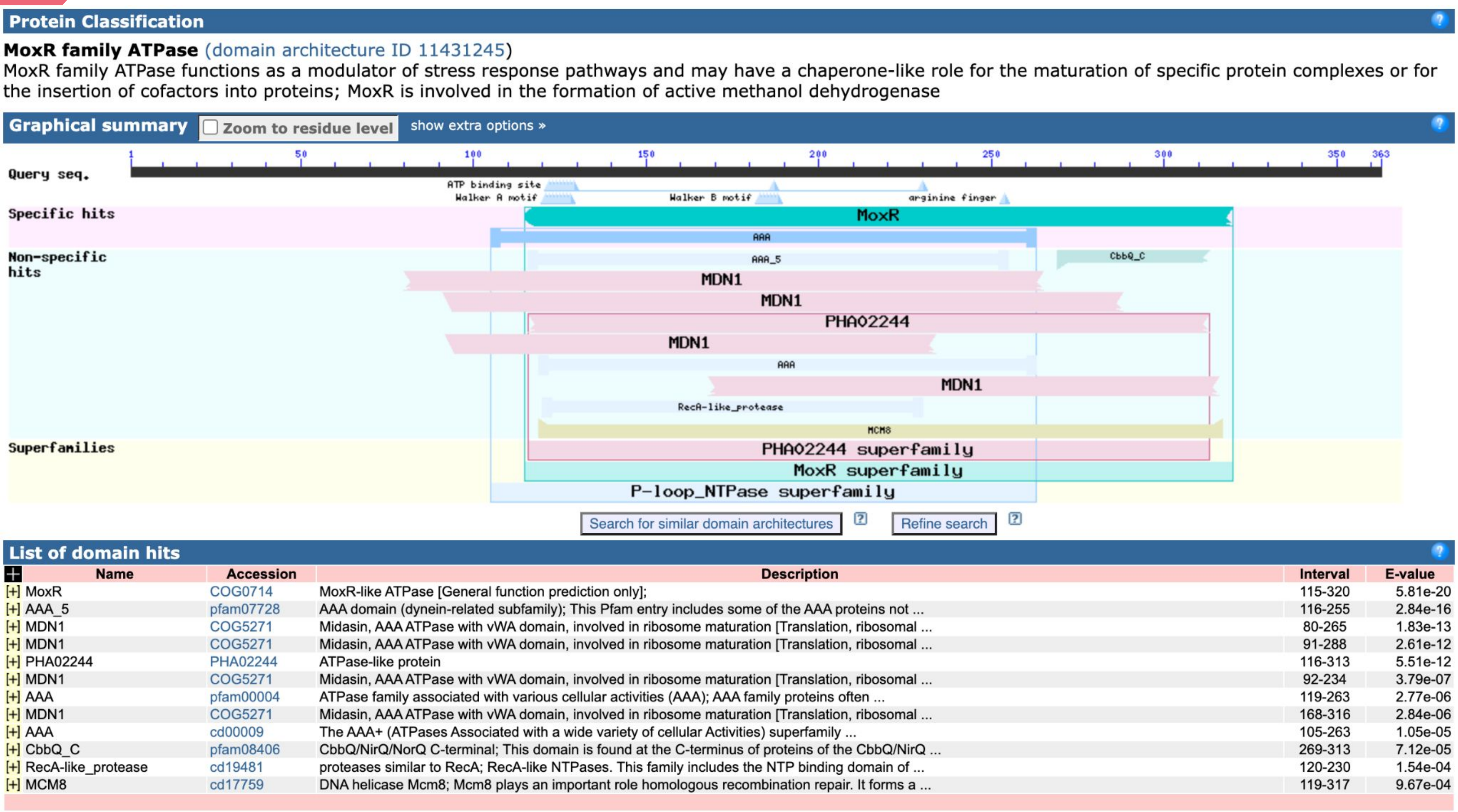
Warren et al. 2002

cobS

- There is a gene called *cobS* in each of the B_{12} production pathways
- They are not in any way related (not homologous) and the enzymes they encode have different functions
    - *cobS* in the aerobic pathway encodes a cobaltochelatase subunit
        - Catalyzes cobalt insertion
        - This is the *cobS* that showed similarity to the P-SSP7 sequence
    - *cobS* in the anaerobic pathway encodes cobalamin-5-phosphate synthase
        - Catalyzes the conversion of ado-GDP-cobamide to ado-cobalamin (last step)
        - This is the *cobS* present in cyanobacteria

How did the *cobS* annotation come to be?

- The researchers did not realize there are two genes named *cobS*
- Reason to believe that encoding *cobS* would increase fitness
- Cobaltochelatase subunit *cobS* has a very common domain
    - AAA+ ATPase (~230 amino acids)
        - Found in all organisms
        - Involved in many, diverse activities
        - Use energy from ATP to exert mechanical force

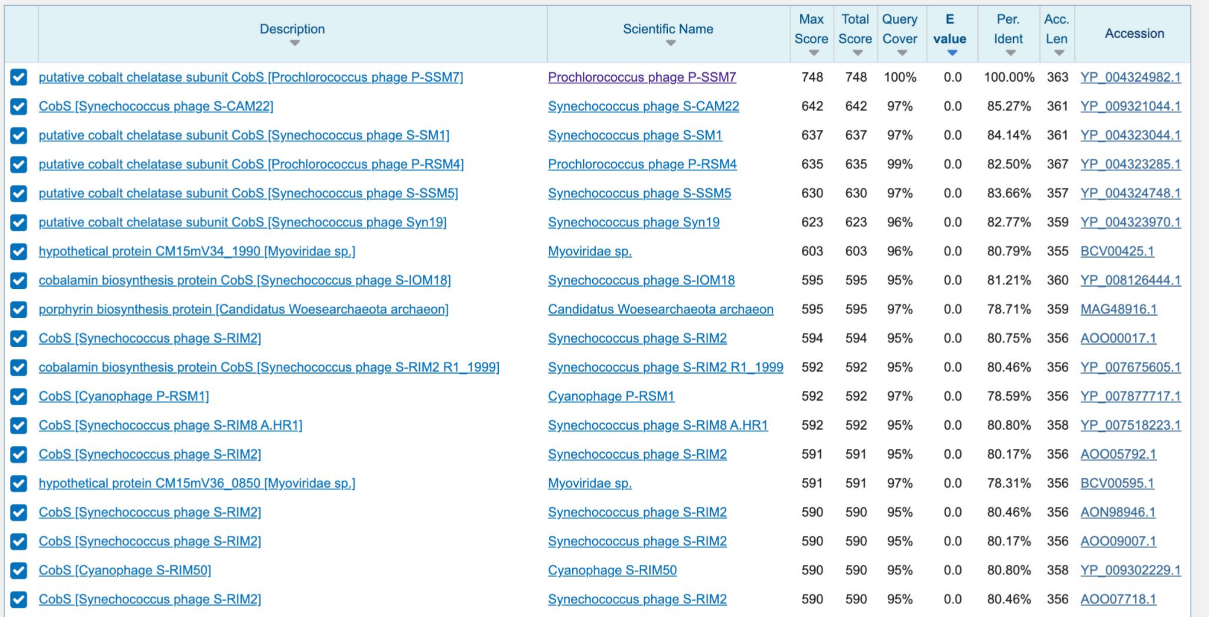


MoxR family ATPase (domain architecture ID 11431245)

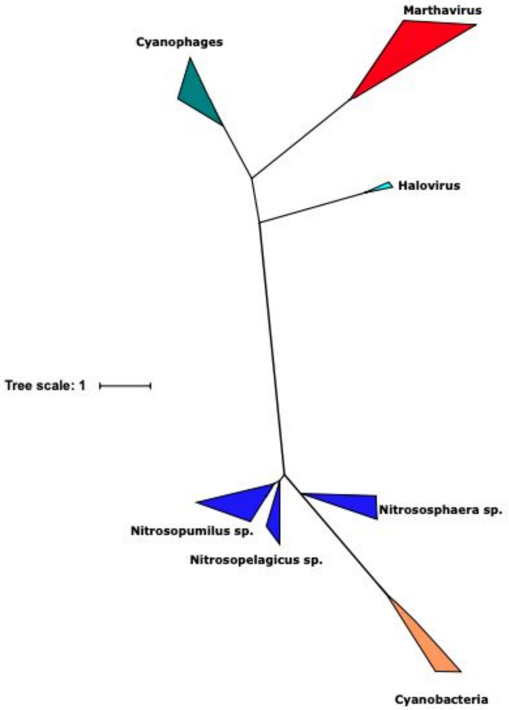
MoxR family ATPase functions as a modulator of stress response pathways and may have a chaperone-like role for the maturation of specific protein complexes or for the insertion of cofactors into proteins; MoxR is involved in the formation of active methanol dehydrogenase

Name	Accession	Description	Interval	E-value
[+] MoxR	COG0714	MoxR-like ATPase [General function prediction only];	115-320	5.81e-20
[+] AAA_5	pfam07728	AAA domain (dynein-related subfamily); This Pfam entry includes some of the AAA proteins not ...	116-255	2.84e-16
[+] MDN1	COG5271	Midasin, AAA ATPase with vWA domain, involved in ribosome maturation [Translation, ribosomal ...	80-265	1.83e-13
[+] MDN1	COG5271	Midasin, AAA ATPase with vWA domain, involved in ribosome maturation [Translation, ribosomal ...	91-288	2.61e-12
[+] PHA02244	PHA02244	ATPase-like protein	116-313	5.51e-12
[+] MDN1	COG5271	Midasin, AAA ATPase with vWA domain, involved in ribosome maturation [Translation, ribosomal ...	92-234	3.79e-07
[+] AAA	pfam00004	ATPase family associated with various cellular activities (AAA); AAA family proteins often ...	119-263	2.77e-06
[+] MDN1	COG5271	Midasin, AAA ATPase with vWA domain, involved in ribosome maturation [Translation, ribosomal ...	168-316	2.84e-06
[+] AAA	cd00009	The AAA+ (ATPases Associated with a wide variety of cellular Activities) superfamily ...	105-263	1.05e-05
[+] CbbQ_C	pfam08406	CbbQ/NirQ/NorQ C-terminal; This domain is found at the C-terminus of proteins of the CbbQ/NirQ ...	269-313	7.12e-05
[+] RecA-like_protease	cd19481	proteases similar to RecA; RecA-like NTPases. This family includes the NTP binding domain of ...	120-230	1.54e-04
[+] MCM8	cd17759	DNA helicase Mcm8; Mcm8 plays an important role homologous recombination repair. It forms a ...	119-317	9.67e-04

The misannotation has been fixed . . . right?



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	putative cobalt chelatase subunit CobS [Prochlorococcus phage P-SSM7]	Prochlorococcus phage P-SSM7	748	748	100%	0.0	100.00%	363	YP_004324982.1
☑	CobS [Synechococcus phage S-CAM22]	Synechococcus phage S-CAM22	642	642	97%	0.0	85.27%	361	YP_009321044.1
☑	putative cobalt chelatase subunit CobS [Synechococcus phage S-SM1]	Synechococcus phage S-SM1	637	637	97%	0.0	84.14%	361	YP_004323044.1
☑	putative cobalt chelatase subunit CobS [Prochlorococcus phage P-RSM4]	Prochlorococcus phage P-RSM4	635	635	99%	0.0	82.50%	367	YP_004323285.1
☑	putative cobalt chelatase subunit CobS [Synechococcus phage S-SSM5]	Synechococcus phage S-SSM5	630	630	97%	0.0	83.66%	357	YP_004324748.1
☑	putative cobalt chelatase subunit CobS [Synechococcus phage Syn19]	Synechococcus phage Syn19	623	623	96%	0.0	82.77%	359	YP_004323970.1
☑	hypothetical protein CM15mV34_1990 [Myoviridae sp.]	Myoviridae sp.	603	603	96%	0.0	80.79%	355	BCV00425.1
☑	cobalamin biosynthesis protein CobS [Synechococcus phage S-IOM18]	Synechococcus phage S-IOM18	595	595	95%	0.0	81.21%	360	YP_008126444.1
☑	porphyrin biosynthesis protein [Candidatus Woesearchaeota archaeon]	Candidatus Woesearchaeota archaeon	595	595	97%	0.0	78.71%	359	MAG48916.1
☑	CobS [Synechococcus phage S-RIM2]	Synechococcus phage S-RIM2	594	594	95%	0.0	80.75%	356	AOO00017.1
☑	cobalamin biosynthesis protein CobS [Synechococcus phage S-RIM2 R1_1999]	Synechococcus phage S-RIM2 R1_1999	592	592	95%	0.0	80.46%	356	YP_007675605.1
☑	CobS [Cyanophage P-RSM1]	Cyanophage P-RSM1	592	592	97%	0.0	78.59%	356	YP_007877717.1
☑	CobS [Synechococcus phage S-RIM8 A.HR1]	Synechococcus phage S-RIM8 A.HR1	592	592	95%	0.0	80.80%	358	YP_007518223.1
☑	CobS [Synechococcus phage S-RIM2]	Synechococcus phage S-RIM2	591	591	95%	0.0	80.17%	356	AOO05792.1
☑	hypothetical protein CM15mV36_0850 [Myoviridae sp.]	Myoviridae sp.	591	591	97%	0.0	78.31%	356	BCV00595.1
☑	CobS [Synechococcus phage S-RIM2]	Synechococcus phage S-RIM2	590	590	95%	0.0	80.46%	356	AON98946.1
☑	CobS [Synechococcus phage S-RIM2]	Synechococcus phage S-RIM2	590	590	95%	0.0	80.17%	356	AOO09007.1
☑	CobS [Cyanophage S-RIM50]	Cyanophage S-RIM50	590	590	95%	0.0	80.80%	358	YP_009302229.1
☑	CobS [Synechococcus phage S-RIM2]	Synechococcus phage S-RIM2	590	590	95%	0.0	80.46%	356	AOO07718.1



Supplementary Fig. S4. Unrooted Maximum Likelihood phylogenetic tree of the CobS protein.

Why did this misannotation snowball?

1. P-SSP7 genome was “the first of its kind” to be annotated
2. A “just-so story”

What can I do about misannotations?

- Check your sequences (when you can)
- Tips for annotating proteins
  - Check your annotations manually
  - Use multiple approaches to inform annotations
  - Check for active or catalytic sites
  - Check for protein domains and their order
  - Use the most recent literature
- If you see a misannotation in a database, let the database know
  - UniProt and NCBI have official means for reporting misannotations

Questions?