A Guide to Manual Genome Annotation (v. 1.6) with examples from the *Porphyra umbilicalis* genome project

Mine Berg, Juliet Brodie, Jay Kim, Simon Prochnik and John Stiller (Members of the NSF *Porphyra* Genome Research Collaboration Network, see end of document)

This version of the manual is a draft (v. 1.6) to which a few additional tutorials based on other, public genomes will be added by the editorial team later in 2016; please see the end of the document for information on how to get the update and for conditional permission to distribute v. 1.6.

Introduction

Genome sequencing, assembly and annotation is a large task that typically involves automated processes as well as manual curation. Genomic DNA (gDNA) is isolated, sequenced and assembled into scaffolds to generate a genome assembly. The genes are located on the assembly and annotated with putative functions and possibly names. Concurrent with gDNA sequencing, sequences are generated from mRNA to capture sequence information from the genes that are expressed in the organism. This involves making complementary DNA (cDNA) from the RNA and sequencing the cDNA as ESTs or in RNA-Seq experiments which also quantify changes in expression under different conditions. Following gDNA and cDNA sequencing, gene prediction is performed automatically using dedicated software. Gene start, stop, and exon splice sites are identified. Automated gene prediction is aided by mapping the cDNA sequences onto the genome assembly. In addition, protein sequences from previously sequenced, related organisms are blasted against the genome assembly. Manual curation may also add extra information not captured in the automatic gene prediction process. Here we describe how to use the reciprocal best hits (RBH) approach to manually curate putative functions of automatically predicted genes in a genome using the genome of the red alga *Porphyra umbilicalis* (Pumb) as an example.

We have used Pumb because this manual was produced by members of the *Porphyra umbilicalis* genome project. However, the annotation steps will be very similar when working with other organisms' genomes. We have developed this manual in the hope that it will speed up the annotation process for anyone working on other genome projects.

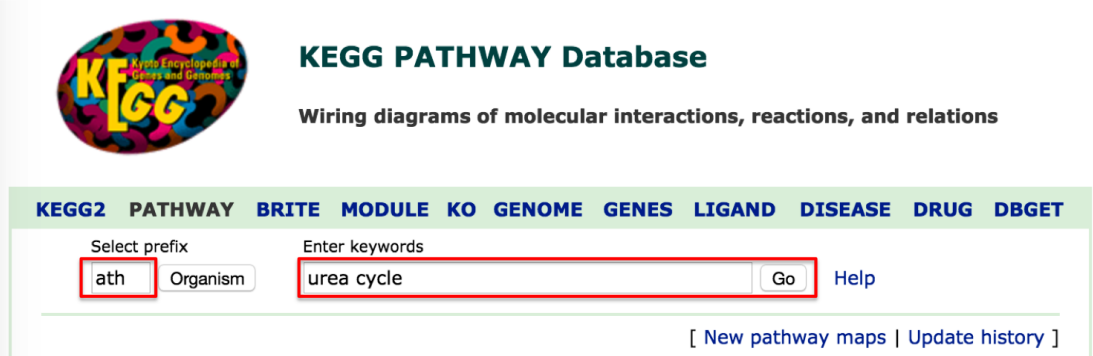
1. Identification of candidate genes to search the Pumb genome

In some instances, you may know the identity of only one protein in a metabolic or developmental pathway but you would like to search all the potentially occurring proteins and their encoding genes in the pathway to which your protein-of-interest belongs. Information on protein pathways can be found at online databases such as the Kyoto Encyclopedia of Genes

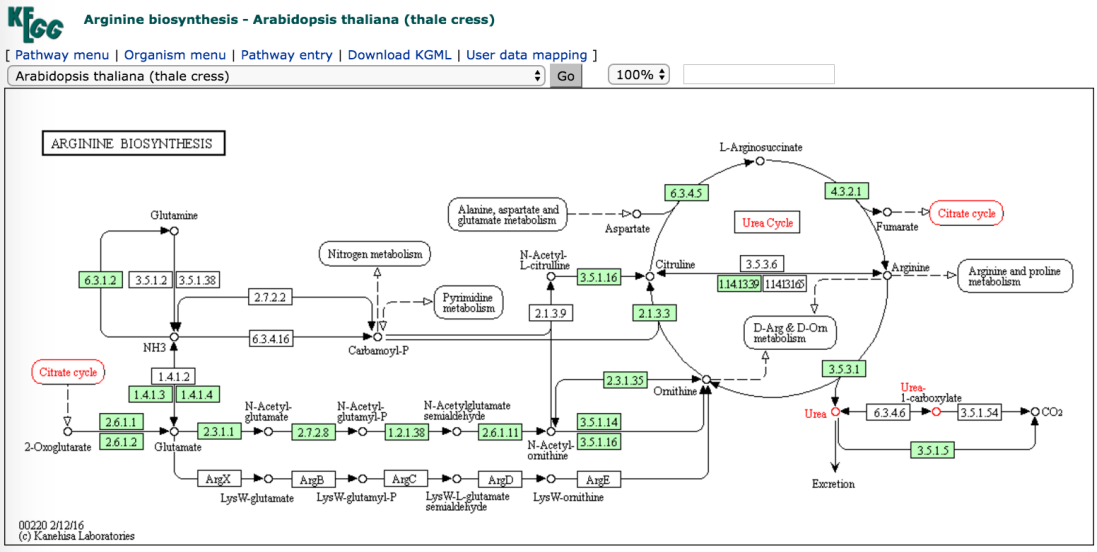
and Genomes (KEGG: www.genome.jp/kegg/pathway.html) or BioCyc (http://biocyc.org). **Tutorial 1** (below) illustrates how to use KEGG for selecting specific enzymes/proteins from the urea cycle of *Arabidopsis thaliana*.

Tutorial 1:

- Navigate to the KEGG homepage in your web browser (www.genome.jp/kegg/pathway.html). In order to retrieve a pathway for a specific organism, you need to provide an organism code and keywords. Type the organism prefix for *Arabidopsis thaliana* ("ath") into the form labeled "Select prefix". Type the words "urea cycle" into the keywords' form and click "Go".



- You will be presented with several pathway map choices. Select the arginine biosynthesis pathway by clicking on it. Squares on the map represent the enzymes in the pathway and circles represent products of the reactions; detailed information on enzymes and their products can be obtained by clicking directly on the square or the circle. Choose the enzymes/proteins from this pathway that you want to query in your genome, and make a list of them.

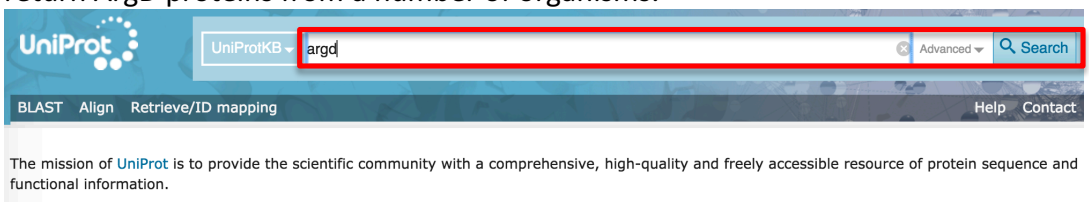


2. Obtaining a FASTA sequence to query the Pumb database

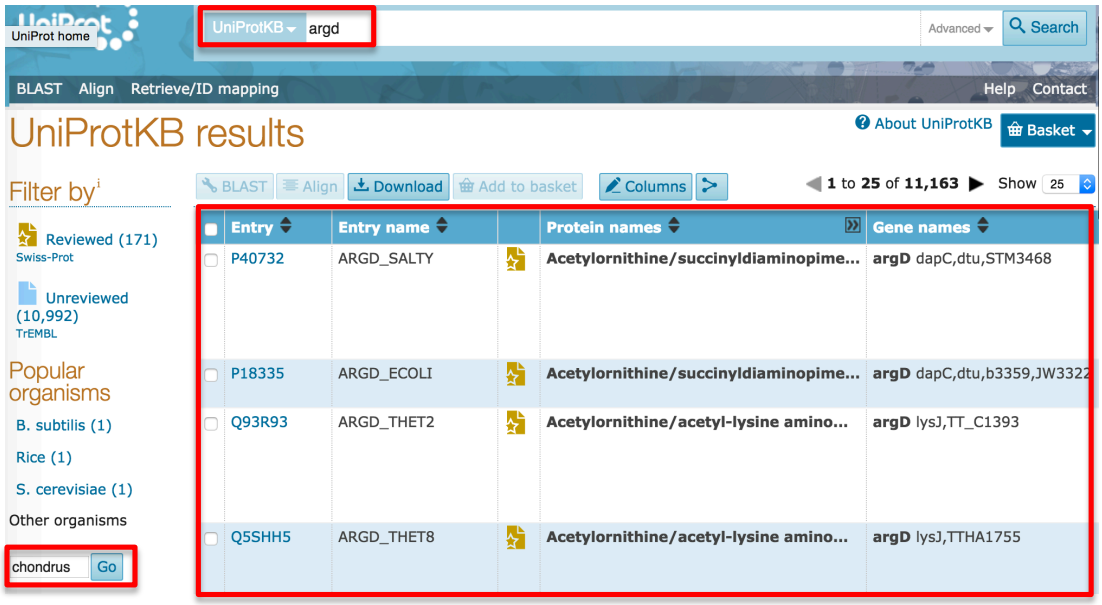
In order to identify whether genes encoding your protein-of-interest are in the *Porphyra umbilicalis* (Pumb) genome, you need a **FASTA** protein sequence to use in a blast search. If you already have a sequence, go to #3 below. Otherwise, **Tutorial 2** (below) illustrates how to retrieve the protein FASTA sequence for acetylornithine aminotransferase (ArgD) from the Uniprot database (http://www.uniprot.org/).

Tutorial 2:

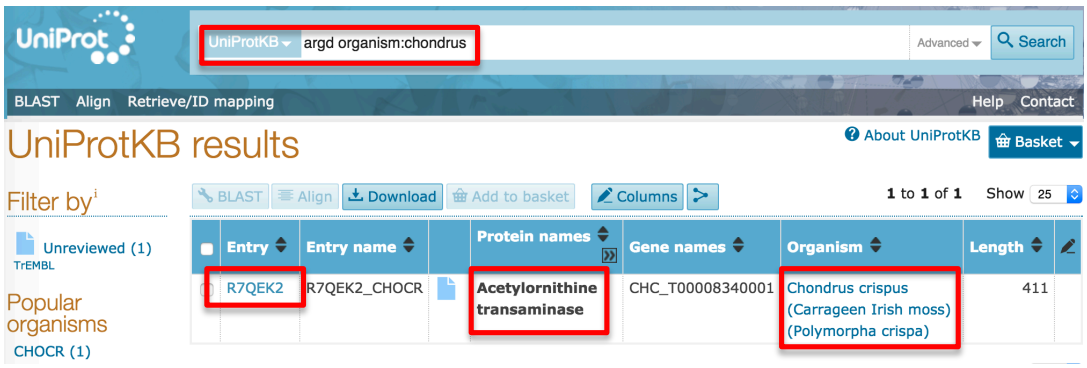
- Go to the **Uniprot** web page (http://www.uniprot.org/). Type in the name of your protein-of-interest in the top dialog box (i.e. "ArgD") and hit return. The search will return ArgD proteins from a number of organisms.



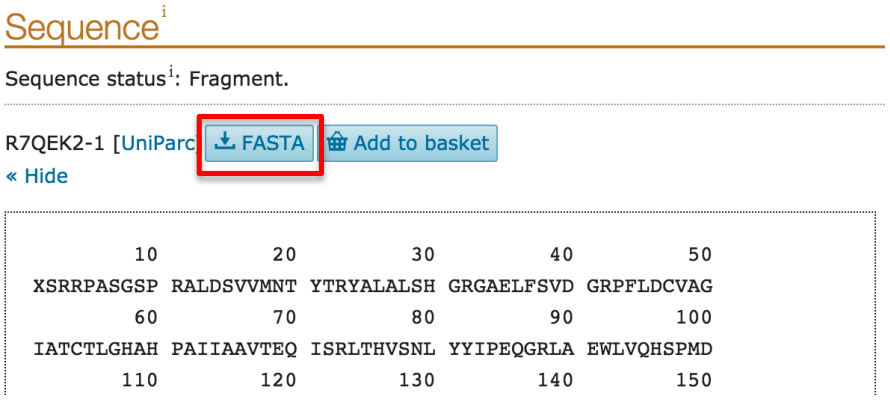
- To obtain an ArgD protein sequence from an organism that is closely related to your organism, go to the left margin and under "Popular Organisms" scroll down to "other". Type in "Chondrus" and/or another organism name; it is important to look for the most closely related organism(s) available. Hit return or "Go".



- You should now see a *Chondrus crispus* ArgD protein listed. Click on the Entry ID (i.e. R7QEK2); this will take you to a page that describes that protein (e.g. sequence, structure, function, etc.).



- At the bottom of this page under the "Sequence" subsection, click on the 'FASTA' button to download the FASTA sequence of the protein. Check your browser's download directory for the FASTA file.



3. Reciprocal (mutual) best hits approach

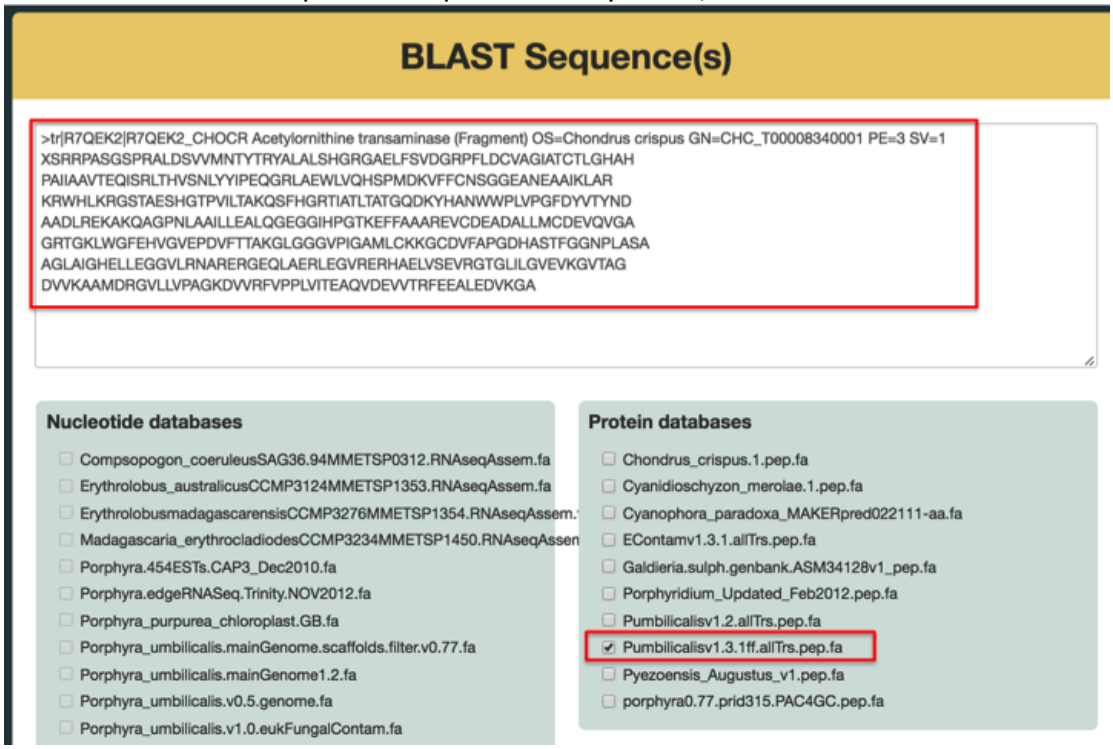
The next step is to perform a BLAST search with your candidate FASTA protein sequence against your genome, and if positive hits are obtained, take these and blast them back against the genome of your candidate protein sequence. This procedure is the reciprocal best hits (RBH) approach, which ensures greater support for orthology (i.e. that genes in different species evolved from a common ancestor by speciation) between your protein of interest and the best BLAST hit in a related genome, whereas best hits from unidirectional BLAST searches can more often be paralogs, or different genes containing similar protein domains.

BLAST searches against your genome require a custom BLAST sequence database containing your genome; generating and executing search queries against this custom BLAST database are tasks requiring knowledge of the command line (e.g. navigating and executing commands in a Unix environment). However, expert members of your genome project can setup a BLAST search engine that is accessible and operable using a web interface.

Tutorial 3 below illustrates how to use a custom BLAST search engine (web interface) to find best hits corresponding to your protein of interest in a closely related organism; it then shows how to run a reciprocal BLAST search using the NCBI BLAST engine. Please proceed with **Tutorial 3a** if you are a member of the *Porphyra* RCN with access to the Pumb BLAST server. Otherwise, please proceed with **Tutorial 3b**, which illustrates the same procedure using publically available tools.

Tutorial 3a:

- Go to the webpage (e.g., that of the *Porphyra* project site if you are a member of this project) where the Pumb BLAST server can be accessed.
- Enter your user name and password in the dialog box.
- Copy and paste the protein FASTA sequence of *Chondrus crispus* ArgD from Uniprot (from **Tutorial 2**) into the query sequence text form. Under "Protein Database", select the most recent protein sequence library. Then, click the blue "BLAST" button.



- Pumb hits will be arranged from best to worst; consider only significant hits; for instance, we can decide to set a significance threshold at E-values of 10^{-4} or less.

Query= tr|R7QEK2|R7QEK2_CHOCR Acetylornithine transaminase (Fragment)

```
OS=Chondrus crispus GN=CHC_T00008340001 PE=3 SV=1

Length=411
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

lcl|Poumbv13000603m  polypeptide=Poumbv13000603m.p locus=Poumbv13...   498    6e-174
lcl|Poumbv13006747m  polypeptide=Poumbv13006747m.p locus=Poumbv13...   202    2e-60
lcl|Poumbv13001914m  polypeptide=Poumbv13001914m.p locus=Poumbv13...  82.8    2e-17
lcl|Poumbv13008953m  polypeptide=Poumbv13008953m.p locus=Poumbv13...  29.3    2.6
lcl|Poumbv13009108m  polypeptide=Poumbv13009108m.p locus=Poumbv13...  29.3    2.7
lcl|Poumbv13009659m  polypeptide=Poumbv13009659m.p locus=Poumbv13...  27.7    9.2
```

- You should be able to click on a particular hit to see the alignment and to obtain the FASTA sequence used for that alignment. Note that although significant, E-values greater than 10^{-10} (e.g., 10^{-9}) may not be biologically informative, depending on the quality of the alignment and the sequence of the protein of interest. Biologically informative alignments generally consist of nearly continuous runs of conserved amino acids.

```
>lcl|Poumbv13000603m polypeptide=Poumbv13000603m.p locus=Poumbv13000603m.g annot-version=v1.3.1
Length=489
```

Click blue link to obtain FASTA

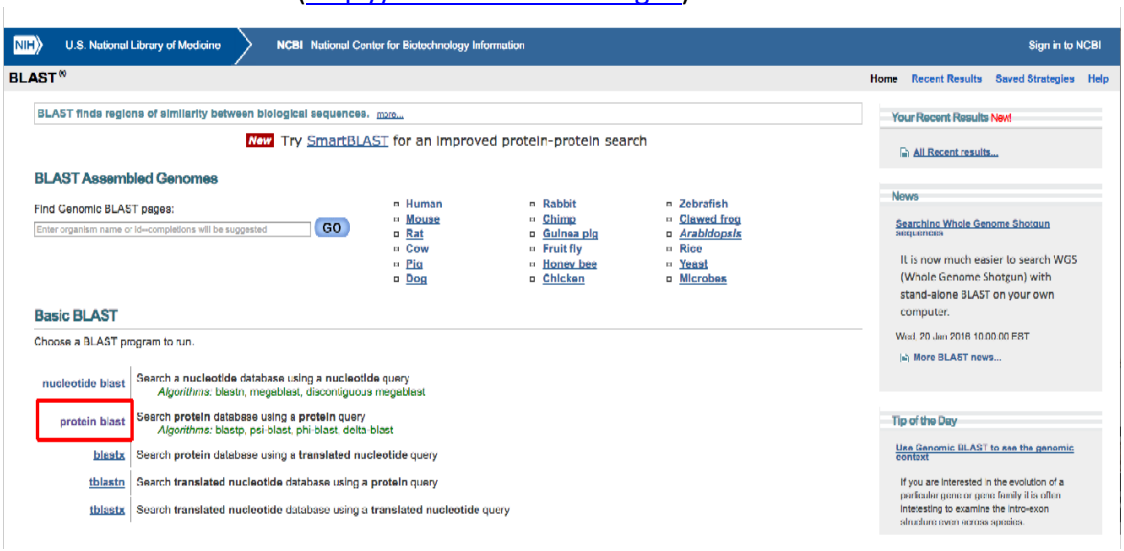
```
 Score =   498 bits (1283),  Expect = 6e-174, Method: Compositional matrix adjust.
 Identities = 256/416 (62%), Positives = 303/416 (73%), Gaps = 9/416 (2%)

Query  2    SRRPASGSPRALDSVVMNTYTRYALALSHGRGAELFSVDGRPFLDCVAGIATCTLGHAHP  61
            S  PA  + +  DSVVM+TY RY L L  G+GA ++  +G+ ++DCVAGIATC+LGHAHP
Sbjct  68   SAAPAGDTIKDFDSVVMHTYGRYPLVLDKGQGASVWDSEGKHYIDCVAGIATCSLGHAHP  127

Query  62   AIIAAVTEQISRLTHVSNLYYIPEQGRLAEWLVQHSPMDKVFFCNSGGEANEAAIKLARK  121
             +IAAVTEQ+ R+ HVSNLYY   QG+LA WLV+HSP DK FFCNSG EANEAAIKLARK
Sbjct  128  GLIAAVTEQLGRVHHVSNLYYTTPQGQLAAWLVEHSPADKAFFCNSGAEANEAAIKLARK  187

Query  122  RWHLKRGSTAESHGTPVILTAKQSFHGRTIATLTATGQDKYHANWWPLVPGFDYVTYNDA  181
             WH K   +  +  +PVI+TA+ SFHGRT+A +TATGQ KYH  + PLVPGF Y  YND
Sbjct  188  HWHAKHDFSPPAGASPVIITAEASFHGRTLAAITATGQPKYHKGFHPLVPGFVYTPYNDP  247

Query  182  ADLREKAKQAGP-NLAAILLEALQGEGGIHPGTKEFFAAAREVCDEADALLMCDEVQVGA  240
              L +     G  N+AAILLE LQGEGG++PGTK FF AARE+CD A ALLM DEVQ G
Sbjct  248  EALAKTVAAVGEGNVAAILLEPLQGEGGVNPGTKAFFGAARELCDSAGALLMVDEVQTGM  307

Query  241  GRTGKLWGFEHVGVEPDVFTTAKGLGGGVPIGAMLCKKGCDVFAPGDHASTFGGNPLASA  300
            GRTG LWG EH+GV+PDV TTAKGLGGG+PIGAMLC   CDVFAPGDHASTFGGNPLASA
Sbjct  308  GRTGLLWGHEHLGVKPDVVTTAKGLGGGIPIGAMLCSAHCDVFAPGDHASTFGGNPLASA  367
```

Long, continuous runs of conserved amino

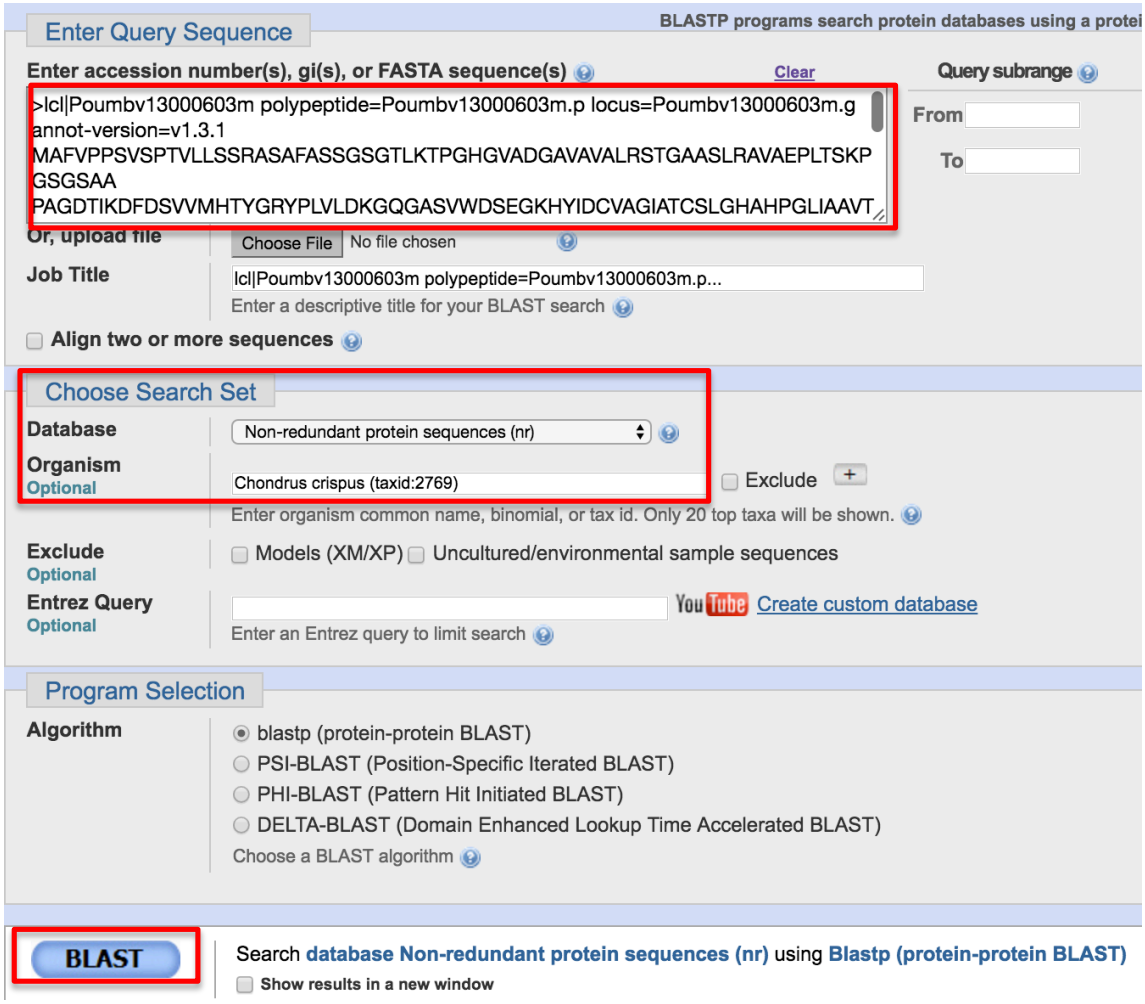
- Copy the FASTA sequence of the putative Pumb ArgD protein. Notice that this protein sequence has a unique identifier: Poumbv13000603m.

```
>lcl|Poumbv13000603m polypeptide=Poumbv13000603m.p locus=Poumbv13000603m.g annot-version=v1.3.1
MAFVPPSVSPTVLLSSRASAFASSGSGTLKTPGHGVADGAVAVALRSTGAASLRAVAEPLTSKPGSGSAAPAGDTIKDFD
SVVMHTYGRYPLVLDKGQGASVWDSEGKHYIDCVAGIATCSLGHAHPGLIAAVTEQLGRVHHVSNLYYTTPQGQLAAWLV
EHSPADKAFFCNSGAEANEAAIKLARKHWHAKHDFSPPAGASPVIITAEASFHGRTLAAITATGQPKYHKGFHPLVPGFV
YTPYNDPEALAKTVAAVGEGNVAAILLEPLQGEGGVNPGTKAFFGAARELCDSAGALLMVDEVQTGMGRTGLLWGHEHLG
VKPDVVTTAKGLGGGIPIGAMLCSAHCDVFAPGDHASTFGGNPLASAAGLAVADALAVDDAEGGVMANVNARGEQLMALL
GEVAAKYGPGVVSEVRGWGLLVGVELSEDAPFNAGEVVAACMASGLLLVPAGPRVVRFVPPLVISETEVATAVSYMDAAL
GDLLAGAK*
```

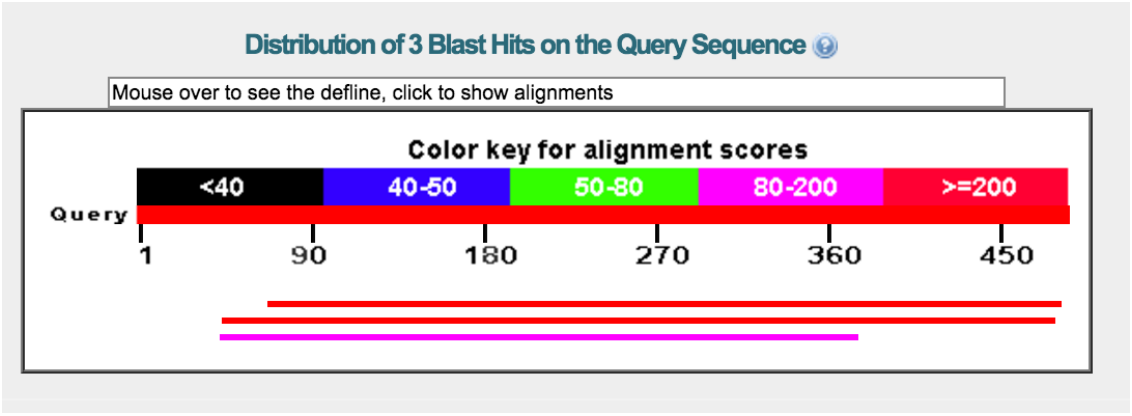
- Now we will use the sequence of Poumbv13000603m to perform a reciprocal BLAST. Go to the NCBI website (http://blast.ncbi.nlm.nih.gov) and select “Protein blast”.



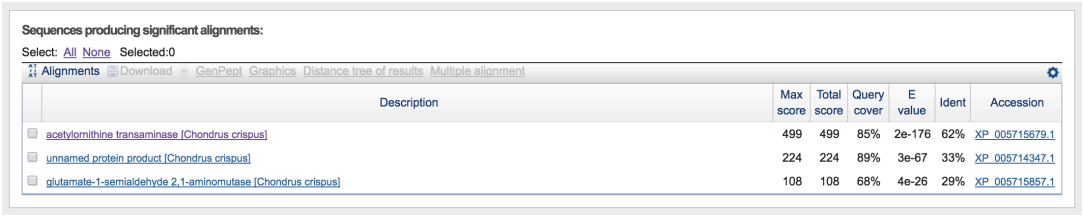
- Paste your Pumb FASTA sequence into the query sequence text form. Under “Choose Search Set” make sure the “Database” selected is “Non-redundant protein database (nr)” and specify the “Organism” as *Chondrus crispus* in order to limit the BLAST search to this species. Note that as you type the organism name, choices for auto-completion will be provided. Hit the BLAST button at the bottom of the screen.



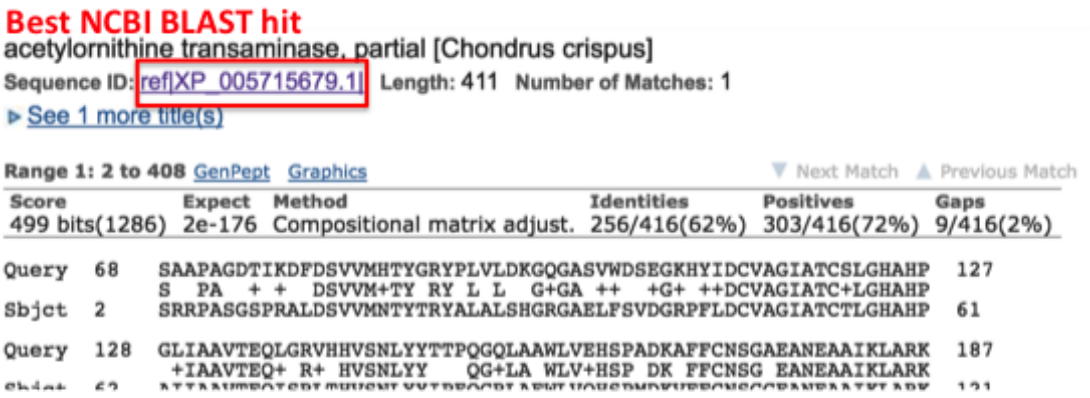
- Your results will be arranged from best hit to worst hit represented by colored bars. Red bars indicate a very good alignment, whereas pink and green represent successively poorer alignments.

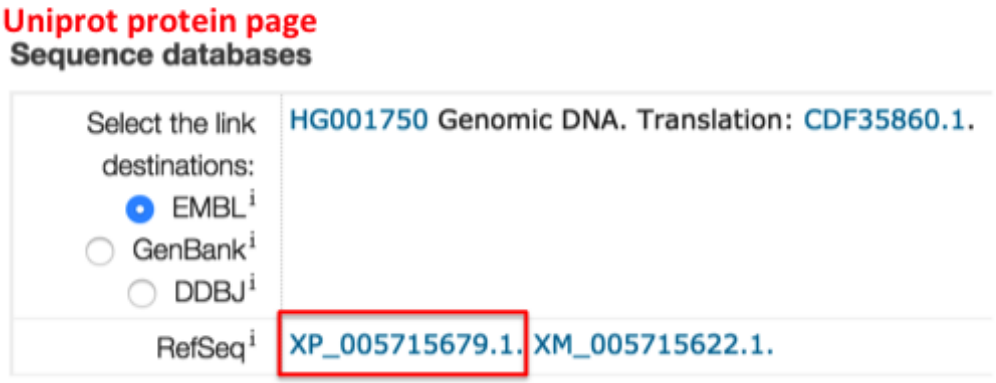


- Check the top alignment result by scrolling down to the “Descriptions” part of the page and clicking on the first line in blue. This will bring you to the alignment page. From here, you can obtain more information, including sequence ID – this line will contain the gene symbol, sequence ID and/or protein name of the top hit generated by your reciprocal BLAST.



- Note that the RefSeq protein identifier for your best hit matches the RefSeq ID provided on the Uniprot page where we downloaded the FASTA sequence for *Chondrus crispus* ArgD (**Tutorial 2**). Thus these two protein sequences from *C. crispus* and Pumb are reciprocal best hits.



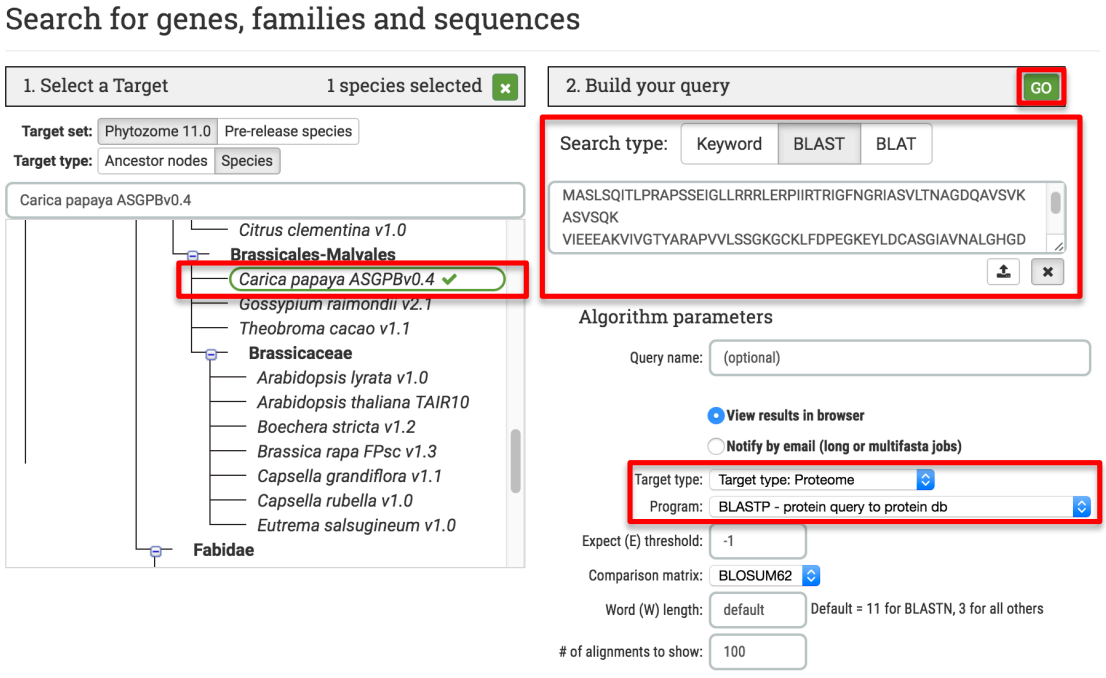


Tutorial 3b:

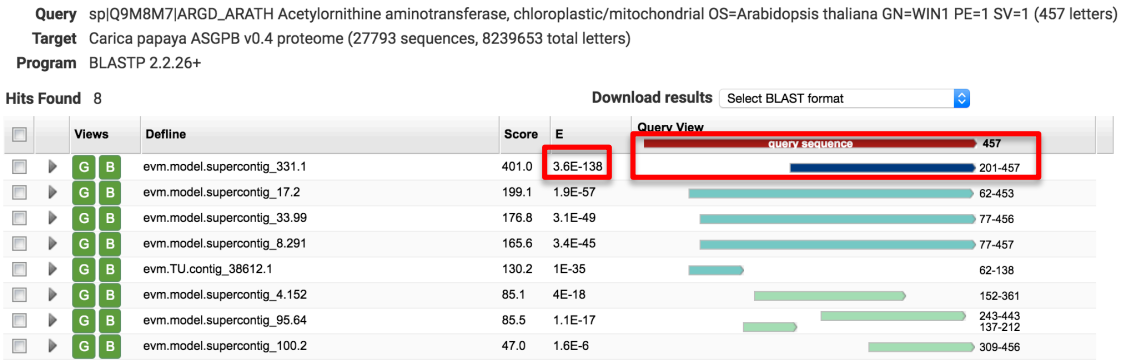
- Obtain the protein FASTA sequence for *Arabidopsis thaliana* ArgD using Uniprot (see **Tutorial 2**). The Uniprot identifier for this protein sequence is Q9M8M7. We will search for the orthologous gene/protein in Papaya (*Carica papaya*).

- Go to the Phytozome 11 website (https://phytozome.jgi.doe.gov/pz/portal.html). Under the "Tools" tab, select "BLAST" to navigate to the BLAST query submission page.

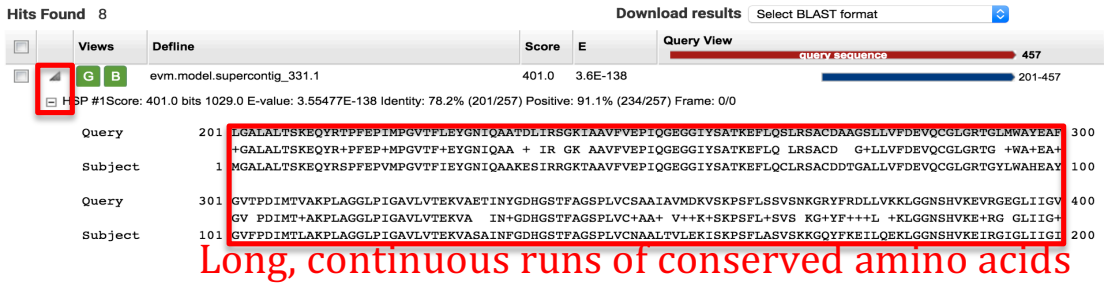
- Select *Carica papaya* ASGPBv0.4 as your target. Paste the *A. thaliana* ArgD protein sequence into the sequence text form and make sure that "BLAST" is selected as the search type. Under "Algorithm Parameters", change the target type to "Proteome" and make sure that BLASTP is the program selected. Now click the green "GO" button on the top-right.



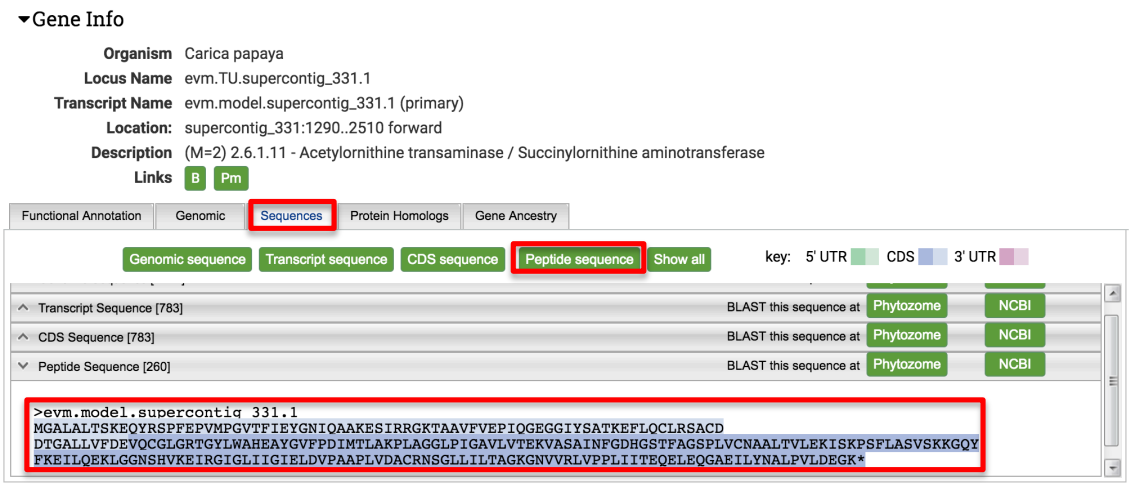
- The results page should come up in a few seconds. Hits will be arranged from best to worst; consider only significant hits; for instance, we can decide to set a significance threshold at E-values of 10^{-4} or less. Notice that the top hit has a good E-value, aligning to roughly half of the *A. thaliana* ArgD protein sequence (residues 201-457).



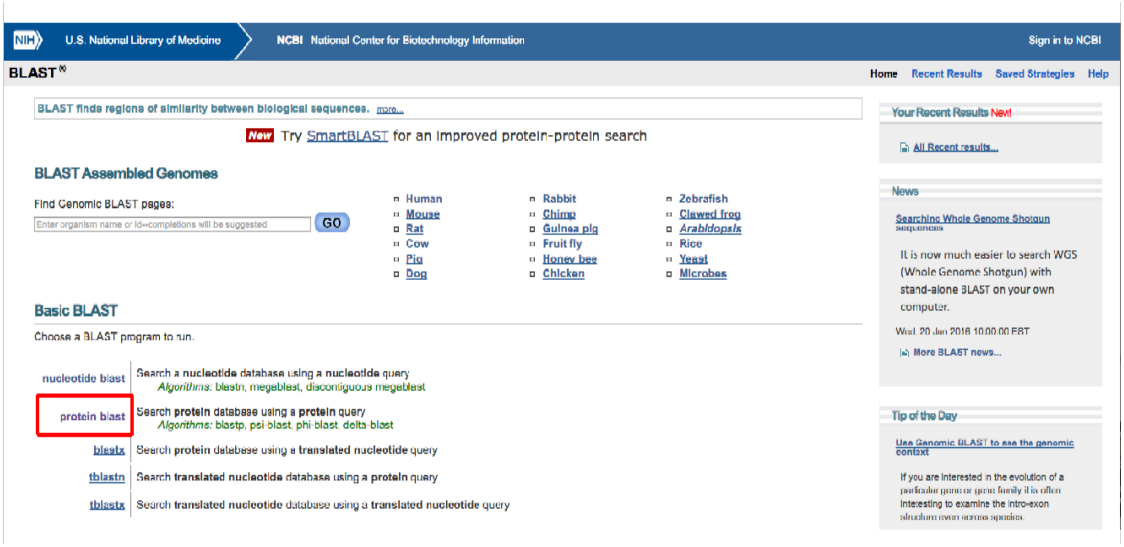
- Click on the right-arrow for the top hit and click on the "+" sign to see the actual alignment. Note that although significant, E-values greater than 10^{-10} (e.g., 10^{-9}) may not be biologically informative, depending on the quality of the alignment and the sequence of the protein of interest. Biologically informative alignments generally consist of nearly continuous runs of conserved amino acids. Notice that the alignment is highly conserved, although the *C. papaya* gene appears to encode only half of the ArgD protein. It may be that we have found a pseudogene, or this may be an incomplete gene model (i.e. the first methionine is actually not the first amino acid) arising from errors in sequencing or automated gene model prediction. We will investigate this further in section 4.



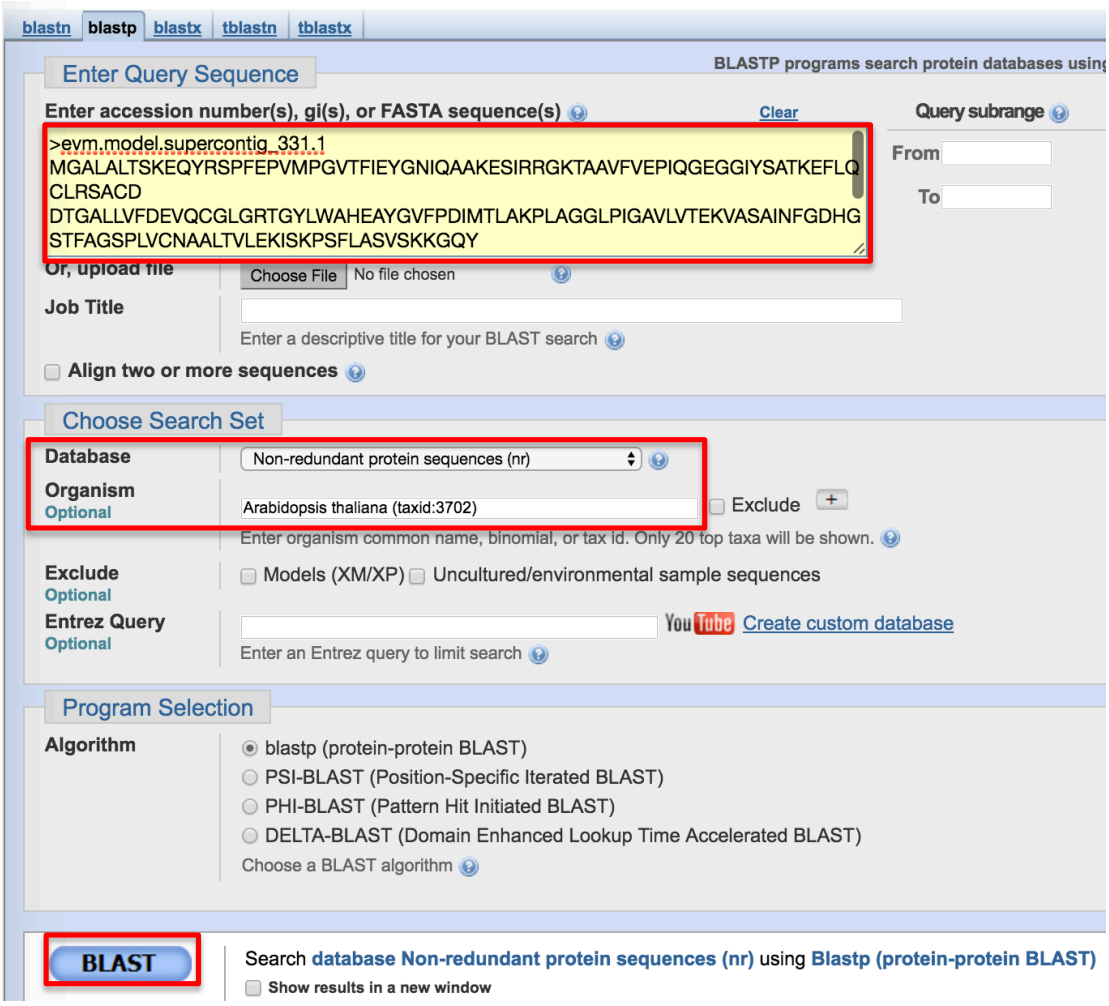
- Now click on the green "G" button to see detailed gene information and obtain the protein FASTA sequence corresponding to this best hit. Under the "Sequences" tab, click "Peptide sequence". We will need to copy and paste this putative *C. papaya* ArgD sequence to perform a reciprocal BLAST.



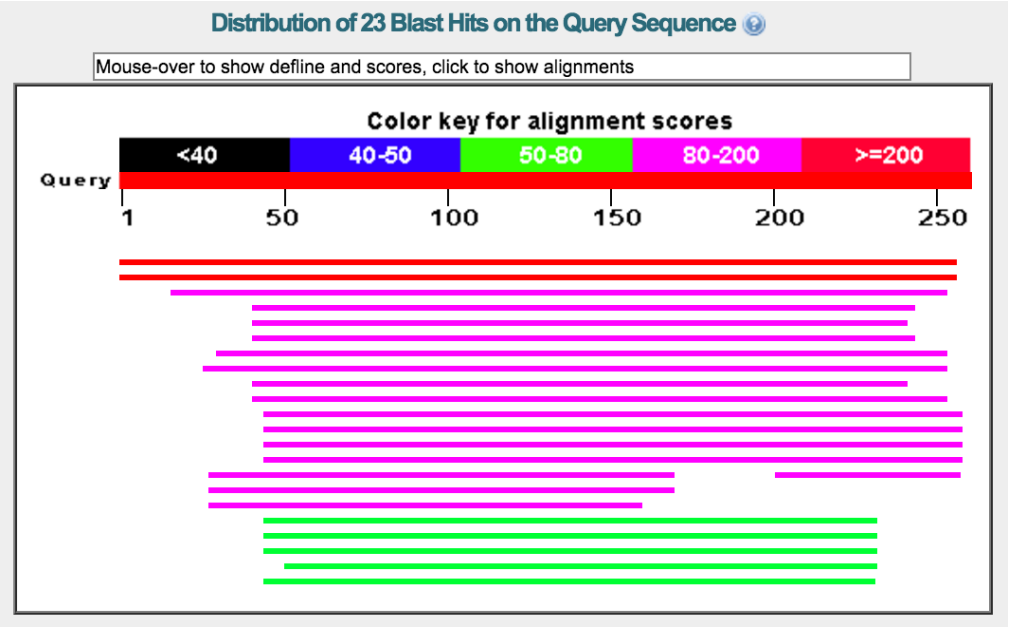
- Now we will use this sequence to perform a reciprocal BLAST. Go to the NCBI website (http://blast.ncbi.nlm.nih.gov) and select “Protein blast”.



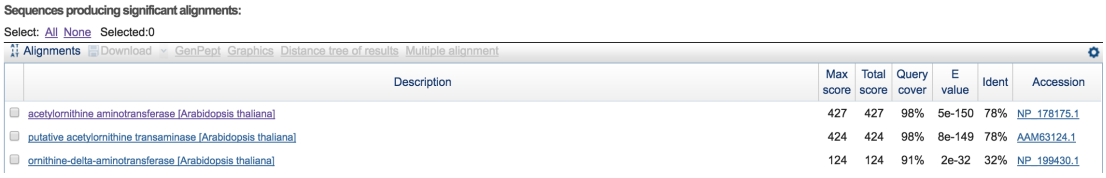
- Paste your *C. papaya* best hit FASTA sequence into the query sequence text form. Under “Choose Search Set” make sure the “Database” selected is “Non-redundant protein database (nr)” and specify the “Organism” as *Arabidopsis thaliana* in order to limit the BLAST search to this species. Note that as you type the organism name, choices for auto-completion will be provided. Hit the BLAST button at the bottom of the screen.



- Your results will be arranged from best hit to worst hit represented by colored bars. Red bars indicate a very good alignment, whereas pink and green represent successively poorer alignments. Notice that there are two very good hits (red).



- Check the top alignment result by scrolling down to the "Descriptions" part of the page and clicking on the first line in blue. This will bring you to the alignment page. From here, you can obtain more information, including sequence ID – this line will contain the gene symbol, sequence ID and/or protein name of the top hit generated by your reciprocal BLAST. Also investigate the second best hit. It turns out that the second best hit is also *A. thaliana* ArgD with a slightly different sequence (two amino acid substitutions); this sequence was deposited into Genbank from a study conducted to improve genome annotation using full-length mRNA sequencing. So the first and second best hits are essentially the same gene, and we can ignore the second best hit.



- Note that the RefSeq protein identifier for your best hit matches the RefSeq ID provided on the Uniprot page where we downloaded the FASTA sequence for *Chondrus crispus* ArgD (**Tutorial 2**). Thus these two protein sequences from *C. crispus* and Pumb are reciprocal best hits.

Best NCBI BLAST hit

acetylornithine aminotransferase [Arabidopsis thaliana]

Sequence ID: ref|NP_178175.1| Length: 457 Number of Matches: 1

See 6 more title(s)

Range 1: 201 to 457 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
427 bits(1098)	5e-150	Compositional matrix adjust.	201/257(78%)	234/257(91%)	0/257(0%)

```
Query  1    MGALALTSKEQYRSPFEPVMPGVTFIEYGNIQAAKESIRRGKTAAVFVEPIQGEGGIYSA  60
            +GALALTSKEQYR+PFEP+MPGVTF+EYGNIQAA + IR GK AAVFVEPIQGEGGIYSA
Sbjct  201  LGALALTSKEQYRTPFEPIMPGVTFLEYGNIQAATDLIRSGKIAAVFVEPIQGEGGIYSA  260

Query  61   TKEFLQCLRSACDDTGALLVFDEVQCGLGRTGYLWAHEAYGVFPDIMTLAKPLAGGLPIG  120
```

Uniprot protein page

Sequence databases

Select the link destinations: EMBL[i] (•) GenBank[i] ( ) DDBJ[i] ( )	EU214908 mRNA. Translation: ABW84224.1. AC018849 Genomic DNA. Translation: AAF27117.1. CP002684 Genomic DNA. Translation: AEE36426.1. AY054594 mRNA. Translation: AAK96785.1. BT002584 mRNA. Translation: AAO00944.1. AY085912 mRNA. Translation: AAM63124.1. AK220871 mRNA. Translation: BAD94250.1.
PIR[i]	B96838.
RefSeq[i]	NP_178175.1. NM_106708.3.
UniGene[i]	At.46389. At.75567.

- At this point, we have found a *Carica papaya* gene putatively encoding ArgD. Unfortunately, this *C. papaya* gene is roughly half the length of the *A. thaliana* gene encoding ArgD. It may be that the current *C. papaya* gene model doesn't include the full protein, or perhaps this is a pseudogene. Please proceed to section 4 and to **Tutorial 4b**, where we will look for EST evidence.

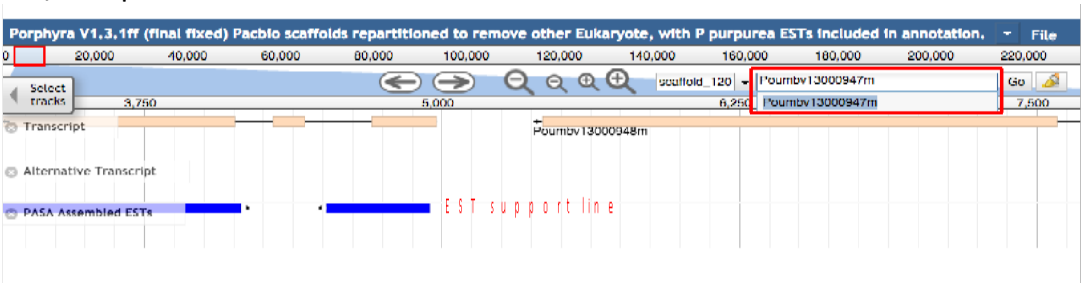
4. Check the EST and other support of your gene on the scaffold

After you have obtained a solid hit with your genome's protein sequence, check whether there is good support for this sequence on your genome scaffold assembly. This can be done using a genome browser to visually confirm various genetic elements (e.g. valid start/stop codons, splice sites, promoters, etc.) and/or by confirming gene expression using ESTs/RNA-Seq, which are short subsequences of cDNA transcripts. Note that a custom genome browser containing your genome must be setup by expert members of your genome project. Also, various annotated features including gene models, repeat elements, ESTs, etc. need to be uploaded to

the genome browser by an expert, as data become available. **Tutorial 4a** (below) shows how to check for EST support using the Pumb genome browser. **Tutorial 4b** shows how to check for EST support using the Phytozome 11 genome browser (Jbrowse).

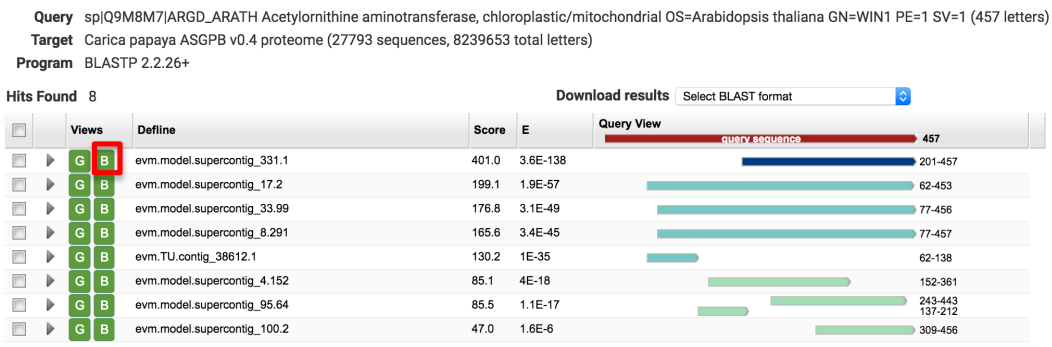
Tutorial 4a:

- Go to the "Genome Browser" in your genome's project; the rest of this example shows this for the *Porphyra* site.
- Click on the link for the current version of your genome's draft assembly.
- Enter the gene/proteinID (e.g. Poumbv13000947m) at the top and hit "Go". The EST support will be marked in blue in this genome project. If your EST contains an intron, or comes from a stranded sequencing protocol, it will be placed on the strand it was expressed from. Examine the direction of the Pumb sequence of interest compared with the EST sequence; they should be the same. (The EST sequencing from Pumb was not stranded, so unless an EST includes an intron, the direction it is displayed in the browser is not necessarily the same as the direction of transcription.) If there is no EST support, examine the placement of other closely related Pumb sequences on the scaffold. If those also do not have EST support, then make a decision based on the protein evidence from the RBH/blastp.

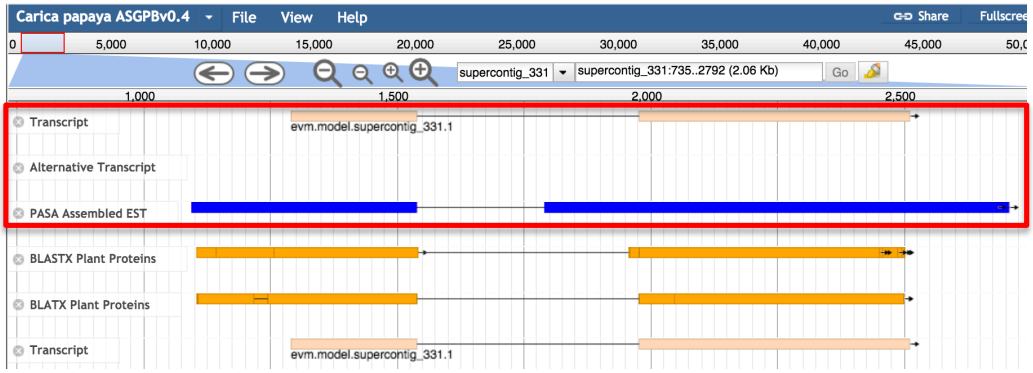


Tutorial 4b:

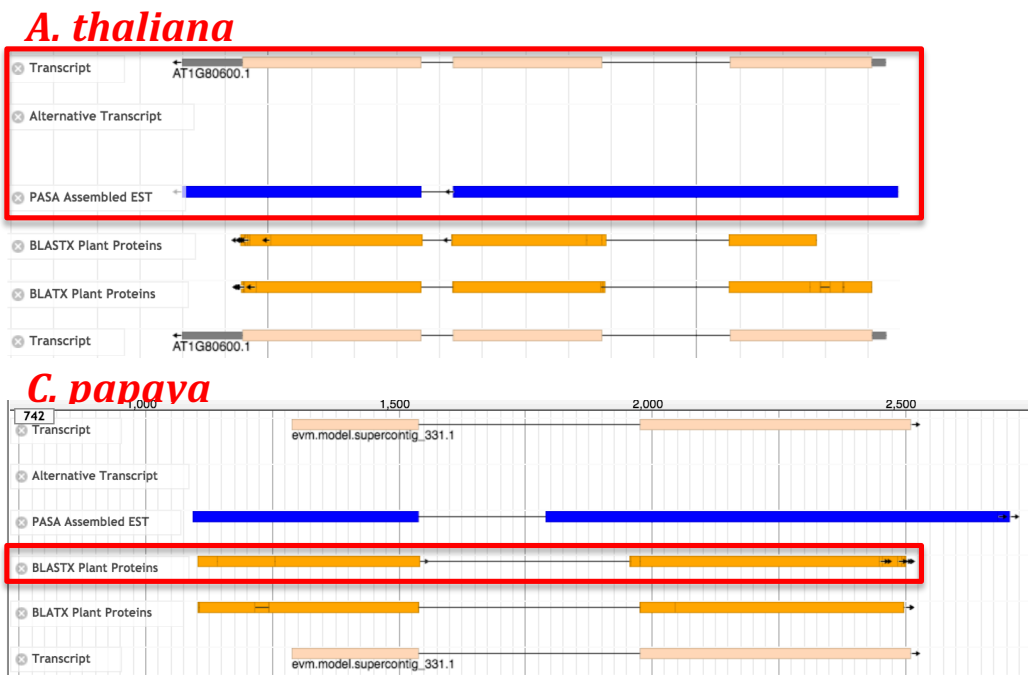
- Go back to the Phytozome 11 BLAST results page from when we blasted *A. thaliana* ArgD protein sequence against the *C. papaya* proteome (see **Tutorial 3b**).
- Click on the green "B" button corresponding to the top hit.



- The EST support will appear in blue on the track called “PASA assembled EST”. If your EST contains an intron, or comes from a stranded sequencing protocol, it will be placed on the strand it was expressed from. Examine the direction of the *C. papaya* sequence of interest (on the “Transcript” track) compared with the EST sequence; they should be the same. (The EST sequencing from Pumb was not stranded, so unless an EST includes an intron, the direction it is displayed in the browser is not necessarily the same as the direction of transcription.) In the event that there is no EST support, examine the placement of other closely related *C. papaya* sequences on the scaffold. If those also do not have EST support, then make a decision based on the protein evidence from the RBH/blastp.



- Great! Our *C. papaya* gene has EST support. However, note that our gene does not fully cover the EST evidence (i.e. the EST extends further in both the 5’ and 3’ directions). This is not surprising since our BLAST (**Tutorial 3b**) revealed that the *C. papaya* gene is much shorter than the full-length gene encoding ArgD on *A. thaliana*. Based on the evidence presented here and **Tutorial 3b**, it should be fairly safe to conclude that our *C. papaya* gene does encode ArgD, and that it is the ortholog of the *A. thaliana* gene encoding ArgD (sequence obtained from Uniprot in **Tutorial 2**). However, the *C. papaya* gene model appears to be incomplete, especially on the 5’ end. Compare the gene models for *C. papaya* and *A. thaliana* below. Notice that the *A. thaliana* gene model corroborates its EST evidence much better. Also, notice that for *C. papaya*, there is a BLASTX Plant Proteins hit that extends out into the 5’ region along with the EST.



5. Choose gene symbol/name and produce gene identification information

When you have resolved the identity of the Pumb gene that you are annotating, choose a gene symbol and complete a table with gene identification information. This information can be gleaned from the gene symbol of the best hit of the reciprocal blast with the Pumb protein FASTA.

- Choose a gene symbol encoding the protein represented by your sequence. This symbol should be 3-5 letters long, capitalized, that reflects the name of the protein/enzyme encoded by the Pumb sequence. Examine the gene symbol chosen by the annotator of the top hit generated by your sequence (and other hits) to get an idea for an appropriate gene symbol.
- Some genes were given different names by different researchers. If this is the case, reading the literature will help determine which is the most commonly used gene symbol. Alternative gene symbols should be recorded as aliases or synonyms.
- In addition to a gene symbol, gather information on the protein name (Defline) encoded by the gene and the function of the protein (Description); arrange this information into a table that includes the Pumb Gene ID:

Gene ID	Gene symbol	Defline	Description
Poumbv0770011580	ARP5	Actin-related protein 5	Component of chromatin-modulating complex (PMID:16195354)

6. Examining expanded gene families for possible pseudogenes.

It is important to examine gene models carefully when drawing biological conclusions about gene function. This is a good rule of thumb when interpreting gene models in any newly sequenced genome, but is of particular importance when a number of gene models match a given gene or gene family; that is, when there appears to an expanded gene family present. In multiple cases in the *P. umbilicalis* genome, careful applications of the methods described in this manual have uncovered evidence that is consistent with relatively recent duplications and/or amplifications of genes. In some cases, a number of these copies may not have EST support on the genome browser, raising the possibility that pseudogenes are included among annotated gene models. In genomes with higher G+C content, which is the case for *P. umbilicalis*, automated identification of pseudogenes can be more complicated. This is because stop codons are high A+T (TAA, TAG, TGA) and, in any genome with a relatively high G+C content, stops are not found as often as expected from random mutations. This means that pseudogenes may continue to appear as extended open reading frames long after they have lost function.

As mentioned above, one potential way to determine whether a called gene model likely represents a functional gene is to look for EST support on the genome browser. Clearly EST support across the model is good evidence that the gene is expressed and probably has a function. It is important to be aware, however, that the absence of EST support, by itself, does not indicate that a given gene model represents a pseudogene. Some genes might be expressed at very low levels, or under conditions from which mRNA was not isolated and, therefore, may not be present in EST data mapped to a genome browser. It is also important to consider that a pseudogene that still retains wild type regulatory sequences should be expected to have an expression pattern that is comparable to that of the non-mutated version, and therefore could also have EST support. Consequently, a more holistic examination should be pursued to determine whether each gene model likely represents a reliably annotated, functional gene for understanding the biology of the organism in question. Tutorial 5 and Tutorial 6 (below) provide examples of such approaches from examination of the *P. umbilicalis* genome.

Tutorial 5

Duplicated MADS-box genes in *Porphyra*

A BlastP search of *Porphyra* gene models, using a MADS protein from *Cyanidioschyzon merolae* as the query, returned five gene models with significant similarity to the region representing the "MADS domain". The results of this Blast search are pictured on the next page.

Sequences producing significant alignments:		Score (Bits)	E Value
lcl|Poumbv13005514m	polypeptide=Poumbv13005514m.p locus=Poumbv13...	84.7	2e-17
lcl|Poumbv13000844m	polypeptide=Poumbv13000844m.p locus=Poumbv13...	82.8	8e-17
lcl|Poumbv13014540m	polypeptide=Poumbv13014540m.p locus=Poumbv13...	60.8	2e-11
lcl|Poumbv13012585m	polypeptide=Poumbv13012585m.p locus=Poumbv13...	56.2	1e-09
lcl|Poumbv13003807m	polypeptide=Poumbv13003807m.p locus=Poumbv13...	45.8	2e-06
lcl|Poumbv13011231m	polypeptide=Poumbv13011231m.p locus=Poumbv13...	28.9	1.8
lcl|Poumbv13004224m	polypeptide=Poumbv13004224m.p locus=Poumbv13...	28.9	4.9

Blast P results when *C. merolae* MADS gene is used as the query

The two sequences highlighted in yellow have significant similarity to the query protein, and also have EST support on the genome browser. Together, these observations provide good support for annotating these two gene models as *bona fide* MADS-family genes. Although three other sequences have similarity at a frequently accepted E-value cut-off of 1e-05, they all have much lower similarity to the query sequence. In addition, none of them have EST support on the browser. Their weaker similarity to known MADS homologs, combined with the lack of EST support, suggest these three gene models could represent unexpressed pseudogenes. If so, they are most likely explained by recent duplications of functional genes, followed by sequence degeneration to form pseudogenes, because they are no longer under purifying selection. Therefore, it is worthwhile exploring possible relationships among all five of the gene models identified by the initial Blast search.

Poumbv13005514m is a long gene model of 1,438 inferred amino acids. Only a short stretch of this represents the conserved MADS domain, the only region that shows similarity to either the *C. merolae* gene, or to any of the other four *Porphyra* gene models identified. When Poumbv13005514m was used as a BlastP query against the *Porphyra* gene model dataset, all of the original four other MADS domains were returned, but so were a number of other sequences not found when another red algal gene was used as the query. This is because there are additional gene models present that are similar to unique regions to the Poumbv13005514m gene model. Further examination showed that these regions are present on a number of scaffolds. When they are on the same scaffolds, they do not appear as contiguous sequences. This indicates they represent duplicated copies of the

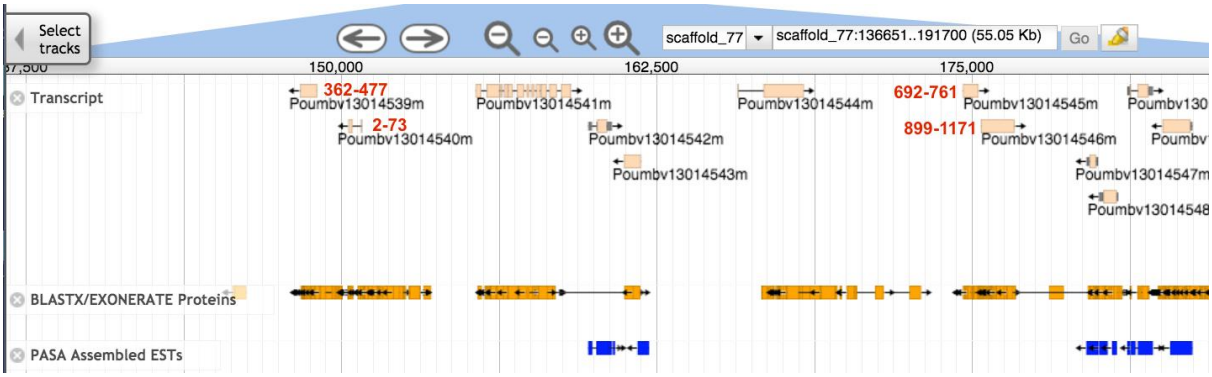
Poumbv13005514m sequence that have been broken apart into discontinuous fragments. None of these fragmented gene models have EST support on the browser. The gene model Poumbv13014540m, highlighted in red in the figure above, is the fragment from one duplication event of Poumbv13005514m that contains sequence similar to the MADS domain. Inferring how it relates to other fragments from that duplication is shown below.

Sequences producing significant alignments:		Score (Bits)	E Value
lcl|Poumbv13005514m	polypeptide=Poumbv13005514m.p locus=Poumbv13...	2620	0.0
lcl|Poumbv13013539m	polypeptide=Poumbv13013539m.p locus=Poumbv13...	274	2e-80
lcl|Poumbv13000920m	polypeptide=Poumbv13000920m.p locus=Poumbv13...	261	1e-74
lcl|Poumbv13000176m	polypeptide=Poumbv13000176m.p locus=Poumbv13...	231	1e-65
lcl|Poumbv13014546m	polypeptide=Poumbv13014546m.p locus=Poumbv13...	203	4e-56
lcl|Poumbv13012585m	polypeptide=Poumbv13012585m.p locus=Poumbv13...	139	1e-37
lcl|Poumbv13014540m	polypeptide=Poumbv13014540m.p locus=Poumbv13...	126	2e-33
lcl|Poumbv13014539m	polypeptide=Poumbv13014539m.p locus=Poumbv13...	107	2e-25
lcl|Poumbv13000844m	polypeptide=Poumbv13000844m.p locus=Poumbv13...	112	5e-25
lcl|Poumbv13003807m	polypeptide=Poumbv13003807m.p locus=Poumbv13...	100	2e-24
lcl|Poumbv13003806m	polypeptide=Poumbv13003806m.p locus=Poumbv13...	99.8	3e-22
lcl|Poumbv13000183m	polypeptide=Poumbv13000183m.p locus=Poumbv13...	100	5e-22
lcl|Poumbv13014545m	polypeptide=Poumbv13014545m.p locus=Poumbv13...	92.8	1e-20

Results of a BlastP search using Poumbv13005514m as the query

The four gene models highlighted in red are all near one another on the same scaffold, and each shows significant similarity to a different region of the original Poumbv13005514m sequence that was duplicated and fragmented. As indicated by the gene model numbers (4539/4540 and 4545/4546), these sequences occur as two adjacent pairs, but the pairs are separated from each other by four other gene models (4541-4544). An examination of the browser (see figure below) also shows that they occur in inverted orientations. The figure below is a snapshot of the Pumb genome browser in the region containing the four gene models highlighted in the figure above. The portions of the corresponding regions of similarity with the Poumbv13005514m gene model are shown in red numbers next to each of the four gene models on scaffold 77 of the Pumb draft genome assembly.

Four duplicated fragments of Poumbv13005514m on scaffold 77



Regions that are similar to the originally duplicated gene can be found by examining alignments associated with the BlastP search using Poumbv13005514m as the query. For example, the figure below shows how the gene model Poumbv13014539m matches and aligns with the originally duplicated Poumbv13005514m gene. Note how the coordinates from the Poumbv13005514m query sequence, highlighted in red below, correspond to the numbers next to the Poumbv13014539m gene model in the image of the genome browser above.

```
>lcl|Poumbv13014539m polypeptide=Poumbv13014539m.p locus=Poumbv13014539m.g annot-version=v1.3.1
Length=224

 Score =   107 bits (267),  Expect = 2e-25, Method: Compositional matrix adjust.
 Identities = 86/116 (74%), Positives = 93/116 (80%), Gaps = 3/116 (3%)

Query  362  TSDVAGGGAPPPLPRARSAVGVRDDAWRRHGAGGGASAGGRADPSGSSRTRGHGPPRLAP  421
            DV+GG APP   R R+AV V++DAWRR GA GGA AG  ADPSGSS TRGHG PRL P
Sbjct  1    MEDVSGGDAPP---RRRNAVAVKEDAWRRPGAAGGALAGALADPSGSSHTRGHGSPRLCP  57

Query  422  HQAVQTRDPRLPAPPVARLVQGLGAARAARGAAGADGRPWGGSFSTDASGDPWAPP  477
            HQ VQT+D  LPAPPVARLVQGLGAARAARGAAGADGRPWG S +T+A GDPW PP
Sbjct  58   HQKVQTQDLLLPAPPVARLVQGLGAARAARGAAGADGRPWGVSPATNAPGDPWKPP  113
```

When the other *Porphyra* MADS gene model with EST support, Poumbv13000844m, is used as the query, it recovers three of the other four MADS domains found in the original search, but does not return any other sequences in the gene model database that are similar to the rest of its sequence. The most reasonable overall conclusions are: 1) Poumbv13005514m and Poumbv13000844m are *bona fide* MADS genes, and 2) the other three domains identified in the initial search, using the red algal homolog from *C. merolae* as the query, are from fragmented pseudogenes from past duplications of Poumbv13005514m.

Tutorial 6
Examining possible pseudogenes in a large gene family expansion

Four histone proteins, H2A, H2B, H3 and H4, combine to form the eukaryotic core nucleosome in an octamer composed of two H2A/H2B dimers and a tetramer of H3/H4. Although clearly distinct families, all four core-histones have similar overall structures and generally are among the most conserved proteins across eukaryotes. In *Porphyra*, the H2A, H2B and H4 gene families look typical. They range from two to five copies, and all gene models show EST support on the browser.

Unlike the other three core histones, there are 62 gene models predicted from the histone H3 family. Only four of the H3 paralogs have EST support, and only two of those are as strongly conserved, as expected, from comparisons across eukaryotic homologs. Understanding the nature of this kind of large gene expansion requires additional approaches beyond Blast searches and examinations of the genome browser. For example, aligning inferred amino acid sequences from all gene models, along with validated homologs from diverse eukaryotes can show which gene models are conserved relative to known functional sequences. Once an alignment is in place, additional approaches can be used, such as clustering or phylogenetic analyses, to assess relationships among the various gene models present in the genome. The figure below provides a summary of such a set of approaches applied to *Porphyra* histone H3 gene models.

A multiple sequence alignment of *Porphyra* histone H3 gene models, and a tree from UPGMA cluster analysis based on the alignment.

The clade above the red line is composed of H3 homologs from other red algae, as well as sequences from animals, yeast, and green plants that have been validated experimentally. It

also contains the two *Porphyra* gene models with the most significant similarity scores when queried with H3 sequences from *Chondrus*. These two sequences both have EST support on the browser (EST support is indicated by yellow highlighting on the tree for a given Pumb gene model). In contrast, all of the other 60 *Porphyra* H3 gene models are excluded from this strongly conserved cluster. This reflects their varying degrees of divergence from canonical H3 sequences, suggesting they are not under the same purifying selection as functional genes. In addition, only two of these more diverged sequences have EST support on the browser.

When confronted with a complicated case like this, it is important to examine the totality of evidence when drawing conclusions about whether gene models represent functional genes or pseudogenes. Below are suggested questions to address using the case of *Porphyra* H3 gene models as an example.

- ***Does the gene expansion make sense biologically?*** In the case of *Porphyra* histones, H3 and H4 should combine in stoichiometry in the nucleosome. There are two copies of H4 in *Porphyra* and two conserved, expressed copies of H3. A reasonable explanation for 60 additional H3 copies based on biological function is not apparent.
- ***Do sequences appear to conform to known functional requirements from experimentally validated sequences?*** Multiple sequence alignment and cluster analyses demonstrate that only two of the 62 *Porphyra* H3 sequences are highly conserved relative to experimentally validated homologs from other eukaryotes.
- ***Does a multiple sequence alignment suggest multiple rounds of gene amplification?*** The alignment of H3 gene models shows groups of sequences that have similar shared insertions, suggesting multiple, independent rounds of amplification of the same sequences.
- ***Is EST support more consistent with functional genes or pseudogenes?*** In this case, the presence of EST support for the two H3 sequences that cluster with known functional homologs, combined with the absence of EST support for all but two of the additional 60 gene models, is reasonable evidence that at least some of the divergent copies are pseudogenes. The fact that two of these divergent gene models have EST support could indicate pseudogenes with intact regulatory sequences, or may indicate the evolution of a novel function. These possibilities would need to be investigated experimentally.

The above discussions of *Porphyra* MADS and histone H3 genes highlight the need for careful examination of expanded gene families in genome investigations, and provide examples of how the bioinformatic resources described in this manual can be applied to the problem.

Acknowledgements and Conditional Permission to Distribute:

We will provide permission for this manual's free distribution, but please send Susan Brawley (brawley@maine.edu) a communication that requests permission to distribute it. We would like to track its use for grant-reporting purposes. This version is a draft (v. 1.6) that is still being refined. Additionally, a few additional topics will be added by the editorial team later in 2016 that will be based upon publicly available genomes.

This manual is based on an earlier tutorial by Simon Prochnik (JGI Staff Scientist and *Porphyra* RCN member; seprochnik@lbl.gov) that was refined and enlarged by RCN members Jay Kim (University of California- Santa Cruz; jay.wj.kim@gmail.com), Mine Berg (Applied Marine Sciences, Inc., Santa Cruz; berg@amarine.com), Juliet Brodie (Natural History Museum of London; j.brodie@nhm.ac.uk), Simon Prochnik, and John Stiller (East Carolina University, stillerj@ecu.edu). Authors are listed alphabetically. We welcome comments and suggestions from users.

Preparation of this manual was supported by the National Science Foundation *Porphyra* Research Collaboration Network (NSF IOS 074097 to S. H. Brawley, E. Gantt, A. Grossman, J. Stiller). The work conducted by the U.S. Department of Energy Joint Genome Institute was supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 to S. H. Brawley, E. Gantt, A. Grossman, J. Stiller. The writing team prepared this work at an April (2016) meeting of the *Porphyra* genome project in Cambridge, England, which was co-sponsored by the NSF *Porphyra* RCN and by NERC Globalseaweed to Claire Gachon (Scottish Association for Marine Science) from the NERC IOF Pump-priming + scheme, number NE/L013223/1.