

Supplemental File S1: How to use PubSEED

In order to use PubSEED to its full capabilities it is important to apply for a RAST ID using the following link: <http://rast.nmpdr.org/?page=Register>

The AT1G78620 gene product is annotated as “Putative uncharacterized protein At1g78620” in Uniprot (<http://www.uniprot.org/uniprot/Q9SYM0>) and as “Protein of unknown function DUF92, transmembrane” in TAIR (<https://www.arabidopsis.org/servlets/TairObject?type=locus&name=AT1G78620>)


First step: finding your gene of interest in PubSEED

Your gene of interest can be found by different ways in PubSEED <http://pubseed.theseed.org/>

1) *Find your gene of interest by gene identifier (ID) from the PubSEED entry page:*

In the “Search String Box” in the entry page, type in the gene ID (here At1g78620) and click the search Button.

You will be brought to the result page. This gene has been annotated with its COG number (COG1836) in PubSEED.



Feature ID ▲▼	Genome ▲▼	Function ▲▼	In set
fig 3702.7.peg.25649	Arabidopsis thaliana	Predicted protein	<input type="checkbox"/>
fig 3702.7.peg.25650	Arabidopsis thaliana	COG1836	<input type="checkbox"/>
fig 3702.11.peg.27174	Arabidopsis thaliana	COG1836	<input type="checkbox"/>

Click on the feature ID shown by the arrow and you will be brought to the gene page.

Note: The fig (Fellowship for Interpretation of Genomes) number is the genome ID. For the purpose of this exercise, use the 3702.7 genome. The peg (protein encoding gene) number is the identifier of the gene itself.

2) *Find your gene of interest by ID from any page*

Your gene of interest can also be found from any PubSEED page via the blank box on the upper right next to the “find” button. Enter the gene ID (here At1g78620) and press enter or click on “find”. This will lead you to the same results page as above. Click on [fig|3702.7.peg.25650](#) to get to the gene page.

3) *By sequence similarity search*

Using the “>>Navigate” tab, click on “BLAST search”. By default, it is set to protein search. Paste the sequence of your input protein and choose the target genome (here: Arabidopsis thaliana; use the Arabidopsis genome that comes up in the second position in the list). You can scroll down the list of the genomes or start typing in the box, then click on the “BLAST” button.

At1g78620 sequence:

```
MATISSTLLLNSSRSALPLRFPKFSGFSSSSPFARSYRFGRNLEPLSNGMLSSGSRADG
ATAAAASMEGVMTEAMKLIQSASPTWKSAVANNLLIFVLGSPLLVTGLSASGIAAAFLLG
TLTWRAYGSAGFLLVAAYFVIVSAFVINLNGTAATKVKMTQKEAQGVAEKRRGRRGPRSV
IGSSAAGCVCAFLSIYQVGGAAFSFLFRLGFVSSFCTKVS DTVSSEIGKAYGKTTYLATT
FKIVPRGTEGAMSLEGLTAGLLASFFLASVGCFLGQITPPEAAVCVLASQIANLGESIIG
ASFQDKEGFKWLNNDVVNVINISLGSIVA ILMQQF I LQNWVK
```

The BLAST search will result in two top hits (splice variants). Click on the second hit from the top ([fig|3702.7.peg.25650](http://pubseed.theseed.org/?page=Annotation&feature=fig|3702.7.peg.25650)), which will bring you to the gene page.

Second step: exploring the genome neighborhood

The gene page (<http://pubseed.theseed.org/?page=Annotation&feature=fig|3702.7.peg.25650>) has several useful features in the top section, including (c)DNA and protein sequences, pre-computed phylogenetic trees and alignments, and links to other databases such as Conserved Domain Database (CDD). The “compare regions” tool is found in the bottom section of the gene page. Genes are depicted as arrows, which point toward the direction of transcription/translation. Genes of the same color are homologous to each other. The genomes displayed are those with homologs of the query gene, arranged in descending order of sequence similarity. To change the number of genomes shown to 100, alter the number in the “number of regions” box to 100. Then click on the “Advanced” button, then click on the “Collapse identical tax-ids” button. To find genes with lower similarity, change the “Evaluate cutoffs” to 1e-10. After making these changes, click on the “draw” button:

Compare Regions For fig|3702.7.peg.25650

The chromosomal region of the focus gene (top) is compared with four similar organisms. The graphic is centered on the focus gene, which is red and numbered 1. Sets of genes with similar sequence are grouped with the same number and color. Genes whose relative position is conserved in at least four other species are functionally coupled and share gray background boxes. The size of the region and the number of genomes may be reset. Click on any arrow in the display to refocus the comparison on that gene. The focus gene always points to the right, even if it is located on the minus strand.

Display options	<input checked="" type="radio"/> Regular <input type="radio"/> Advanced
Region Size (bp)	16000
Number of Regions	15

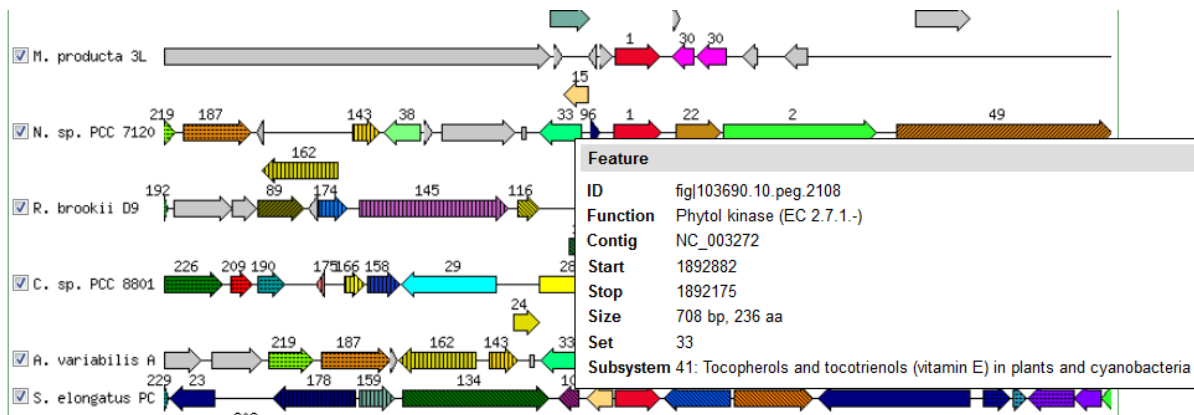


Compare Regions For fig|3702.7.peg.25650

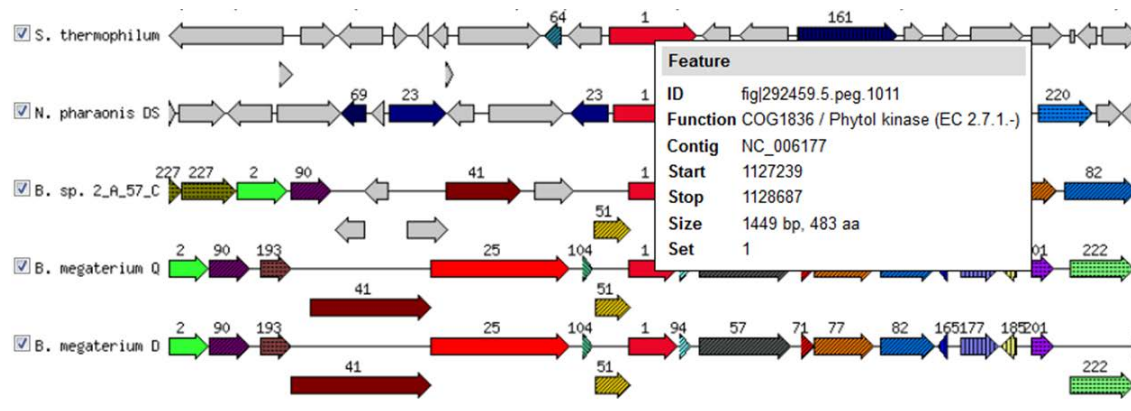
The chromosomal region of the focus gene (top) is compared with four similar organisms. The graphic is centered on the focus gene, which is red and numbered 1. Sets of genes with similar sequence are grouped with the same number and color. Genes whose relative position is conserved in at least four other species are functionally coupled and share gray background boxes. The size of the region and the number of genomes may be reset. Click on any arrow in the display to refocus the comparison on that gene. The focus gene always points to the right, even if it is located on the minus strand.

Display options	Regular	Advanced
Region Size (bp)	16000	
Number of Regions	100	
Pinned CDS selection	<input checked="" type="radio"/> Similarity <input type="radio"/> Kmer <input type="radio"/> PCH pin	
Genome selection	<input type="radio"/> Collapse close genomes <input checked="" type="radio"/> Collapse identical tax-ids <input type="radio"/> Show all	
Sort genomes by	<input checked="" type="radio"/> Similarity to input CDS <input type="radio"/> Phylogenetic distance to input CDS <input type="radio"/> Phylogeny	
Value cutoff for selection of pinned CDS	1e-10	
Value cutoff for coloring CDS sets	1e-10	
Coloring algorithm	<input checked="" type="radio"/> Fast <input type="radio"/> Slower (but exact)	

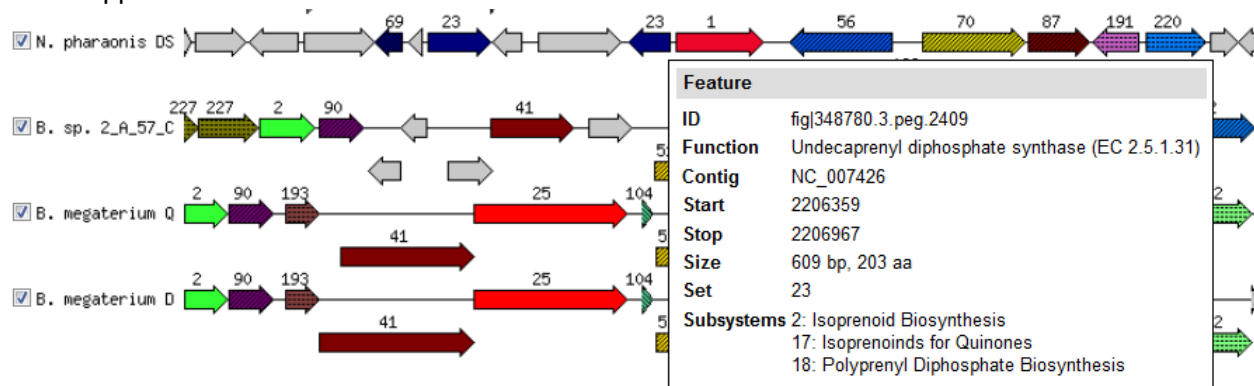
Hovering over any gene (arrow) in the compare regions tool will produce a pop-up with this gene's annotation information. For example, in the "*N. sp PCC 7120*" genome, a phytol kinase gene in the opposite direction gene is just upstream of your focus gene:



More strikingly, COG1836 and phytol kinase are fused in the “*S. thermophilum*” genome:



Further evidence for a general connection of COG1836 with isoprenoid metabolism is found in the “*N. pharaonis* DS” genome; an undecaprenyl diphosphate synthase gene is upstream of COG1836 in the opposite orientation:



More examples like the connections between COG1836 and polyprenyl metabolism highlighted above can easily be detected with the compare regions tool for any gene of interest. PubSEED offers a lot more functions, and more information on how to use these can be obtained from <http://www.hos.ufl.edu/meteng/HansonWebpagecontents/workshop/2014-CGW-UF-Day1&2.html>.