Where’s my old gene ID?

Contents

Introduction

As genome sequencing technology improves, reference genomes need to be updated. These new references genomes generally have better quality sequences, but may not always include annotations from older, more established reference genomes.

If you have one particular gene of interest that you want to find on the new genome there are several methods you can use to get there.

If you can’t find what you’re looking for or need guidance, contact us at help@veupathdb.org. Further resources are available at https://veupathdb.org/veupathdb/app/static-content/faq.html

Search methods

1. Site search

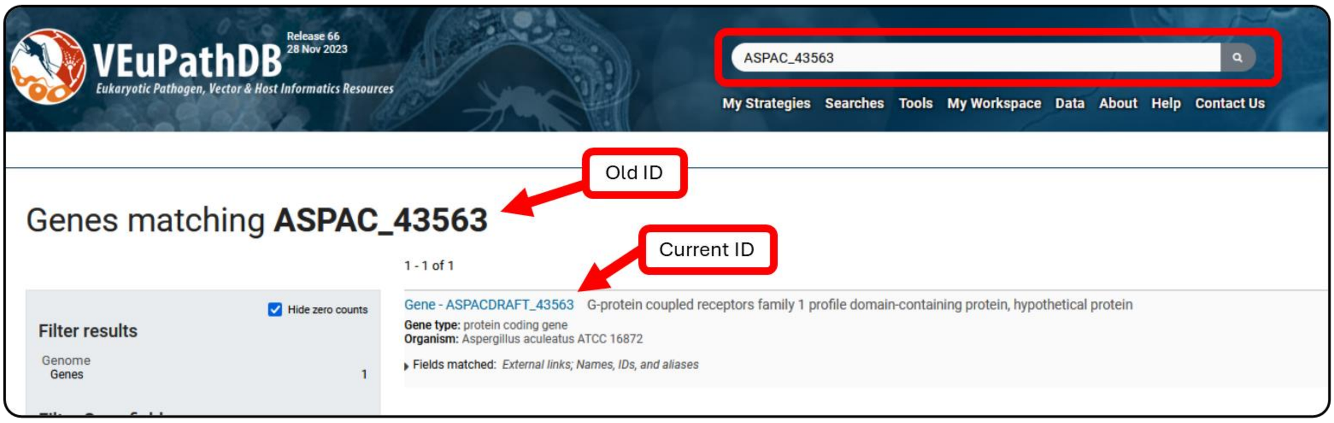
Sometimes your gene may have an alias on the VEuPathDB sites. Aliases are made when gene models and assemblies are updated, and new gene identifiers are created or imported. Aliases are searchable on VEuPathDB and you can do this with single or multiple genes.

The following methods may not work for all genomes as aliases may be unavailabe for various reasons. If you cannot find your gene identifier this way you may have to try one of the other methods listed in this document.

If you want to help improve alias mappings, VEuPathDB does accept files from the community. Files can be sent to help@veupathdb.org or submitted though this form: https://veupathdb.org/veupathdb/app/contact-us.

For single genes

You can use any of these past identifiers in the search bar and it will return the gene with the current name like this:



Genes matching **ASPAC_43563**

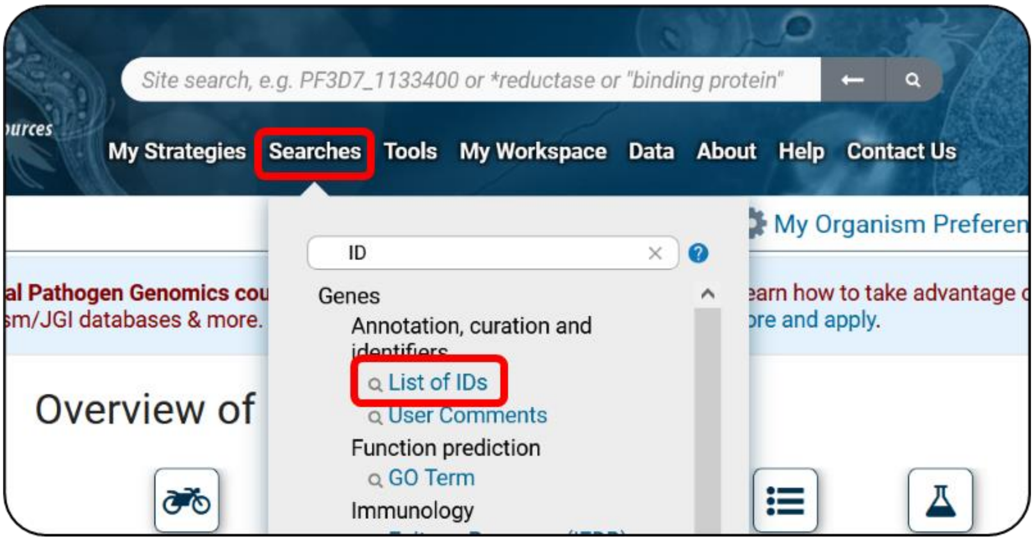
1 - 1 of 1

Gene - ASPACDRAFT_43563 G-protein coupled receptors family 1 profile domain-containing protein, hypothetical protein
Gene type: protein coding gene
Organism: Aspergillus aculeatus ATCC 16872
Fields matched: *External links; Names, IDs, and aliases*

For multiple genes

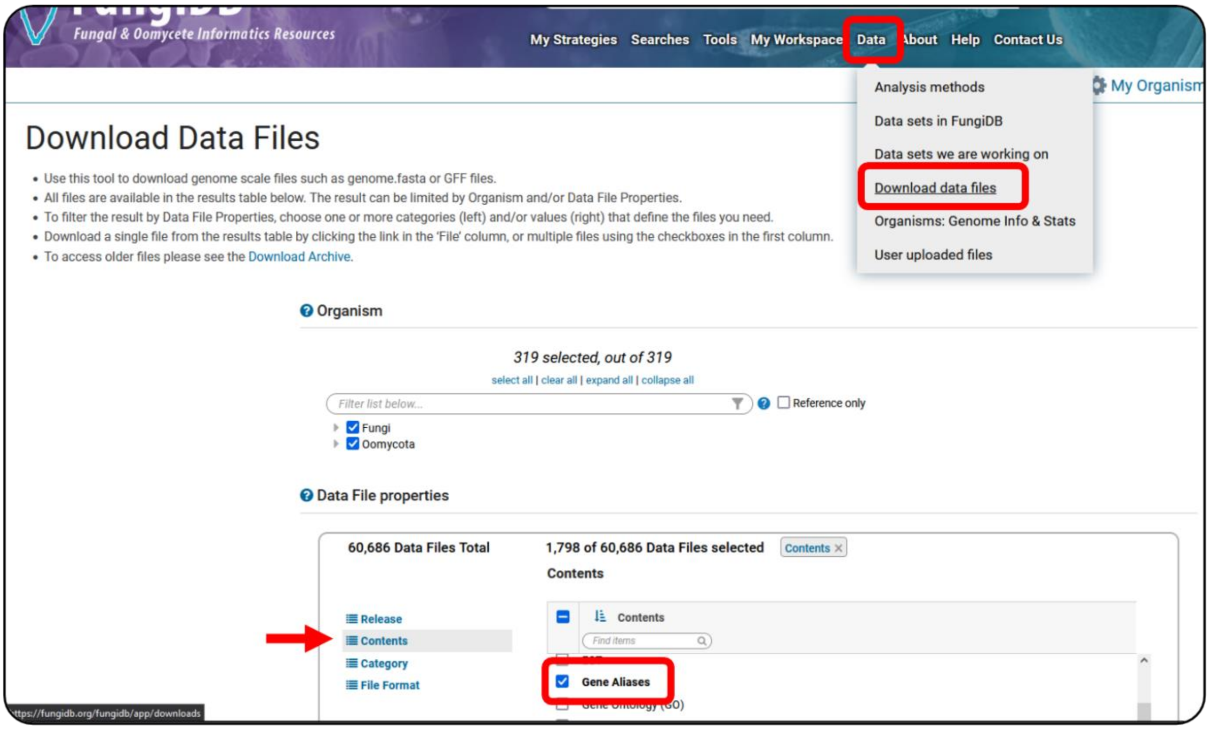
Use the “List of IDs” search to find aliases for multiple gene IDs.

https://veupathdb.org/veupathdb/app/search/transcript/GeneByLocusTag



Finding lists of aliases

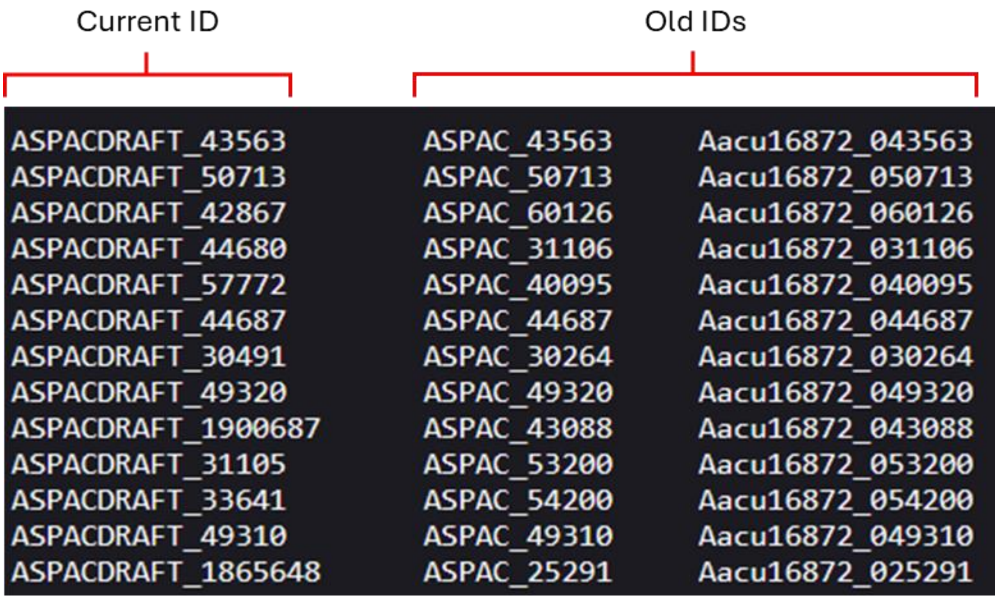
You can find lists of aliases for organisms on any VEuPathDB child site, such as FungiDB, Vectobase, or ToxoDB under the “Download data files” tab. For example, https://fungidb.org/fungidb/app/downloads



Download Data Files

- Use this tool to download genome scale files such as genome.fasta or GFF files.
- All files are available in the results table below. The result can be limited by Organism and/or Data File Properties.
- To filter the result by Data File Properties, choose one or more categories (left) and/or values (right) that define the files you need.
- Download a single file from the results table by clicking the link in the ‘File’ column, or multiple files using the checkboxes in the first column.
- To access older files please see the Download Archive.

This will give you a tab delimitted list of aliases used for genes in a given organism that looks like this:



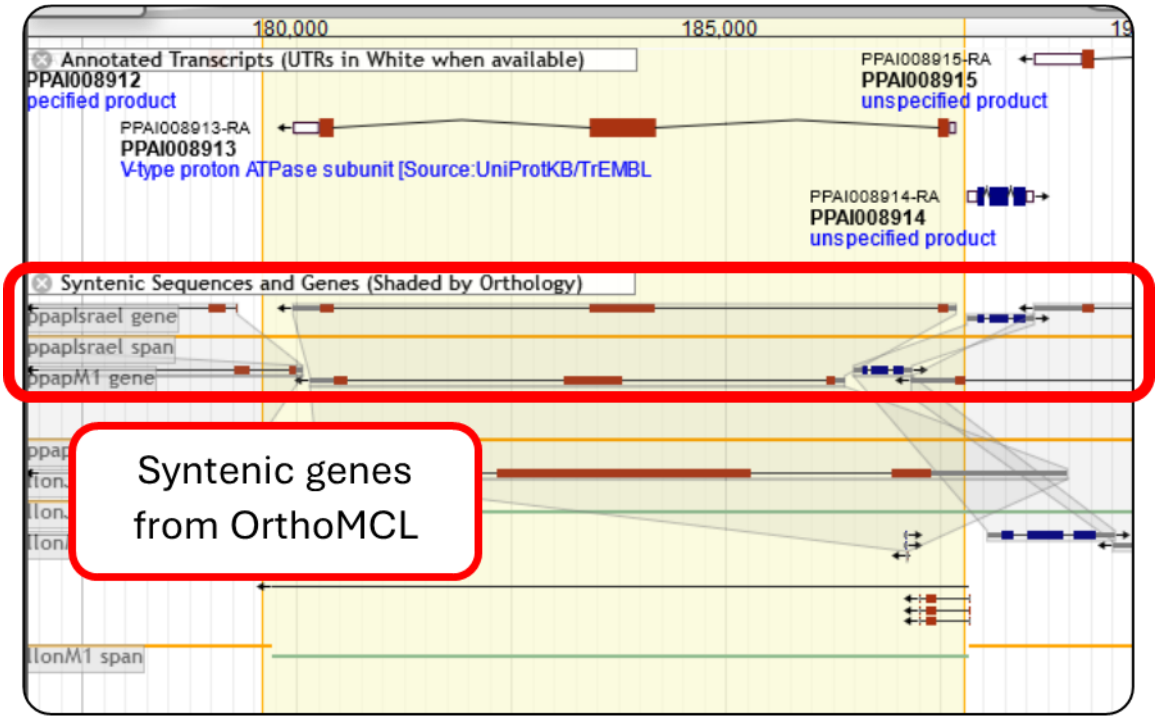
```
ASPACDRAFT_43563      ASPAC_43563     Aacu16872_043563
ASPACDRAFT_50713      ASPAC_50713     Aacu16872_050713
ASPACDRAFT_42867      ASPAC_60126     Aacu16872_060126
ASPACDRAFT_44680      ASPAC_31106     Aacu16872_031106
ASPACDRAFT_57772      ASPAC_40095     Aacu16872_040095
ASPACDRAFT_44687      ASPAC_44687     Aacu16872_044687
ASPACDRAFT_30491      ASPAC_30264     Aacu16872_030264
ASPACDRAFT_49320      ASPAC_49320     Aacu16872_049320
ASPACDRAFT_1900687    ASPAC_43088     Aacu16872_043088
ASPACDRAFT_31105      ASPAC_53200     Aacu16872_053200
ASPACDRAFT_33641      ASPAC_54200     Aacu16872_054200
ASPACDRAFT_49310      ASPAC_49310     Aacu16872_049310
ASPACDRAFT_1865648    ASPAC_25291     Aacu16872_025291
```

On the left is the current gene identifier, and any column to the right are past identifiers.

2. Orthology and Synteny

For a single gene

You can compare your old genes to the ortholog and synteny track on on its gene page or on JBrowse to find potential matches on the new reference genome.

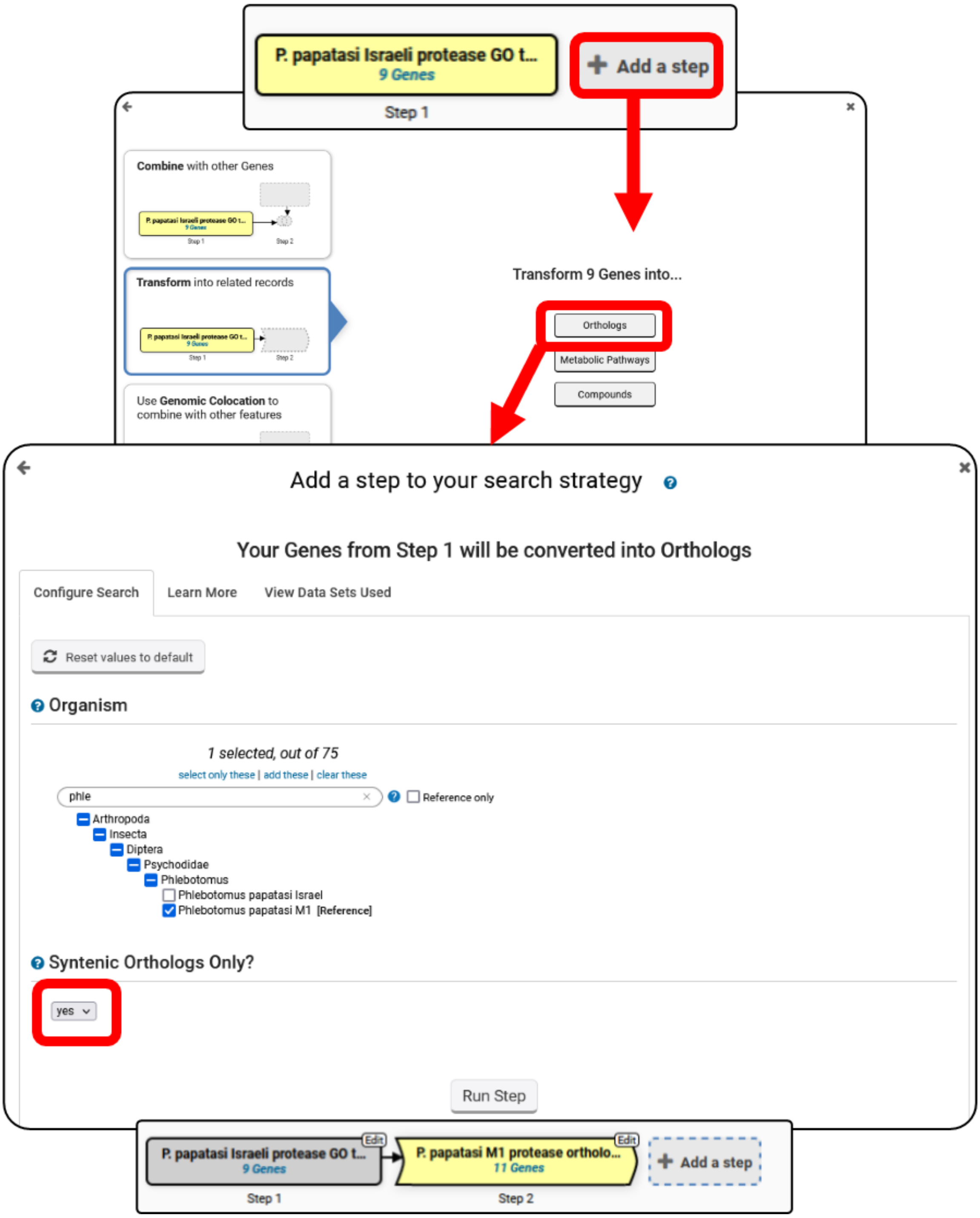


Syntenic genes from OrthoMCL

For multiple genes

You can make use of search strategies to transform multiple gene records from the old reference genome into syntenic orthologs in the new reference genome. (see https://veupathdb.org/veupathdb/app/#search-strategies)

You can add a step onto the end of a search strategy:



Add a step to your search strategy

Your Genes from Step 1 will be converted into Orthologs

Organism

1 selected, out of 75

phle

☐ Reference only

- Arthropoda
  - Insecta
    - Diptera
      - Psychodidae
        - Phlebotomus
          - ☐ Phlebotomus papatasi Israel
          - ☑ Phlebotomus papatasi M1 [Reference]

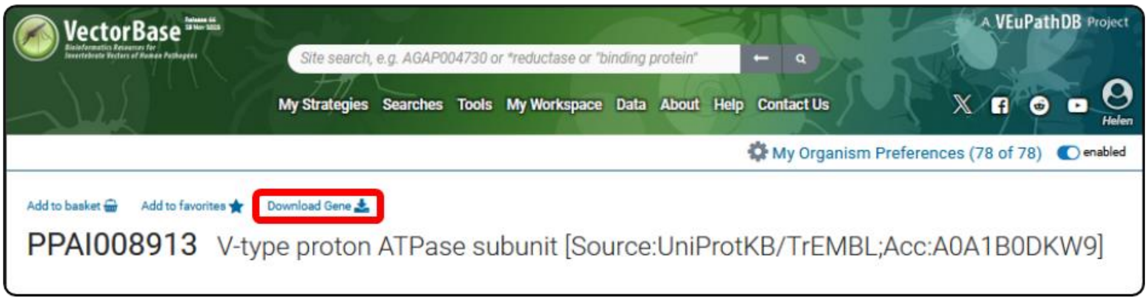
Syntenic Orthologs Only?

yes

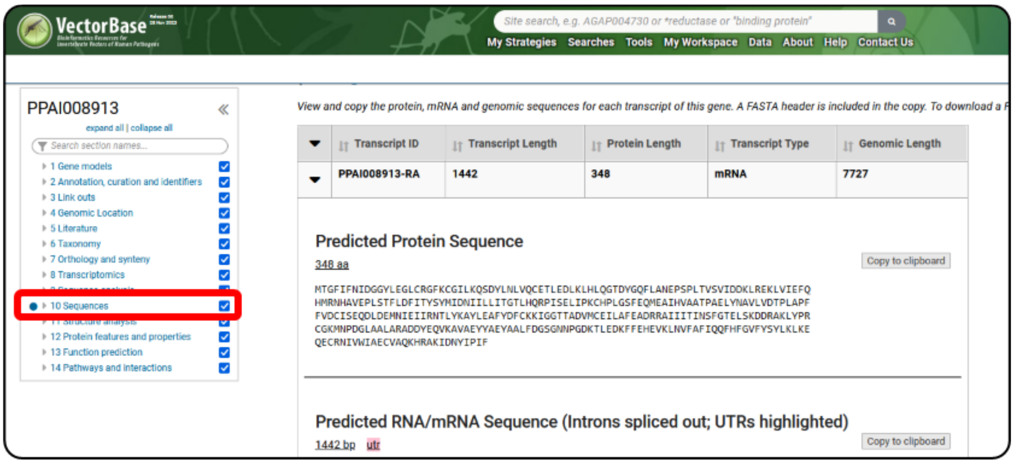
3. Sequence searches

Find sequence matches for your sequence of interest with either the BLASTP multiblast tool or a BLAT search on Apollo. BLASTP will give you a more in-depth summary of the match and take you to the gene page, from which you can navigate to a genome view. BLAT on Apollo will allow you to immediately see the gene in genome context.

The sequence can be obtained from the gene record page where it can be exported with the button at the top of the page or copied and pasted from the “Sequences” section



PPAI008913 V-type proton ATPase subunit [Source:UniProtKB/TrEMBL;Acc:A0A1B0DKW9]



View and copy the protein, mRNA and genomic sequences for each transcript of this gene. A FASTA header is included in the copy. To download a F

Transcript ID	Transcript Length	Protein Length	Transcript Type	Genomic Length
PPAI008913-RA	1442	348	mRNA	7727

Predicted Protein Sequence

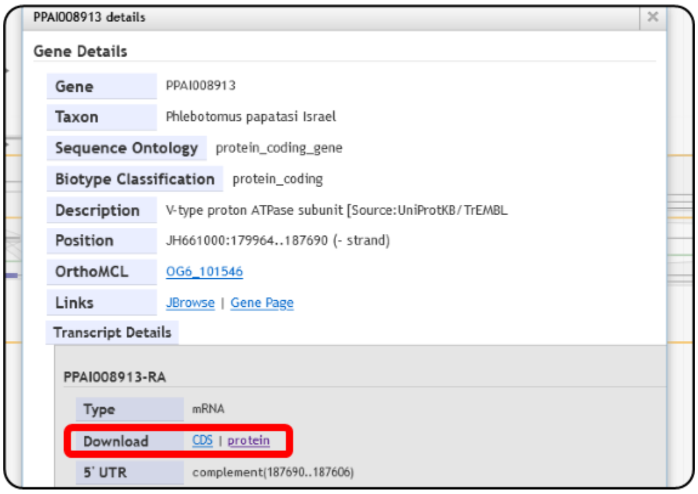
348 aa

```
MTGFIFNIDGGYLEGLCRGFKCGILKQSDYLNLVQCETLEDLKLHLQGTDYGQFLANEPSPLTVSVIDDKLREKLVIEFQ
HMRNHAVEPLSTFLDFITYSYMIDNIILLITGTLHQRPISELIPKCHPLGSFEQMEAIHVAATPAELYNAVLVDTPLAPF
FVDCISEQDLDEMNIEIIRNTLYKAYLEAFYDFCKKIGGTTADVMCEILAFEADRRAIIITINSFGTELSKDDRAKLYPR
CGKMNPDGLAALARADDYEQVKAVAEYYAEYAALFDGSGNNPGDKTLEDKFFEHEVKLNVFAFIQQFHFGVFYSYLKLKE
QECRNIVWIAECVAQKHRAKIDNYIPIF
```

Predicted RNA/mRNA Sequence (Introns spliced out; UTRs highlighted)

1442 bp utr

Or the JBrowse window:



PPAI008913 details

Gene Details

Gene	PPAI008913
Taxon	Phlebotomus papatasi Israel
Sequence Ontology	protein_coding_gene
Biotype Classification	protein_coding
Description	V-type proton ATPase subunit [Source:UniProtKB/TrEMBL
Position	JH661000:179964..187690 (- strand)
OrthoMCL	OG6_101546
Links	JBrowse | Gene Page

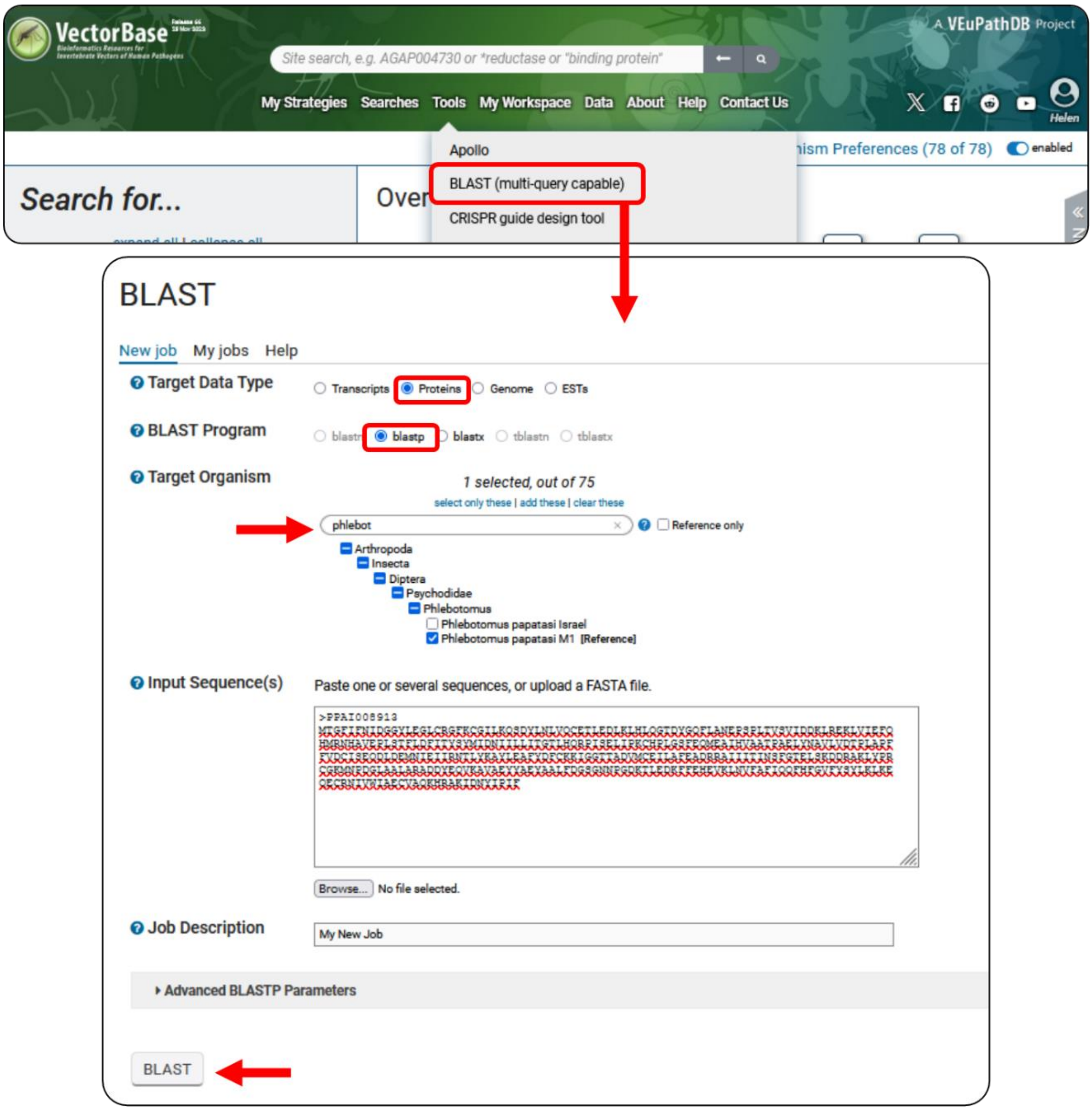
Transcript Details

PPAI008913-RA

Type	mRNA
Download	CDS | protein
5' UTR	complement(187690..187606)

With BLASTP

Navigate to the BLAST tool on the webpage an paste your gene, or genes, of interest into the text box. Then select the appropriate data type, BLAST tool, and organism you want to search against.



Target Data Type: ○ Transcripts ● Proteins ○ Genome ○ ESTs

BLAST Program: ○ blastn ● blastp ○ blastx ○ tblastn ○ tblastx

Target Organism: 1 selected, out of 75

select only these | add these | clear these

phlebot ☐ Reference only

- Arthropoda
  - Insecta
    - Diptera
      - Psychodidae
        - Phlebotomus
          - ☐ Phlebotomus papatasi Israel
          - ☑ Phlebotomus papatasi M1 [Reference]

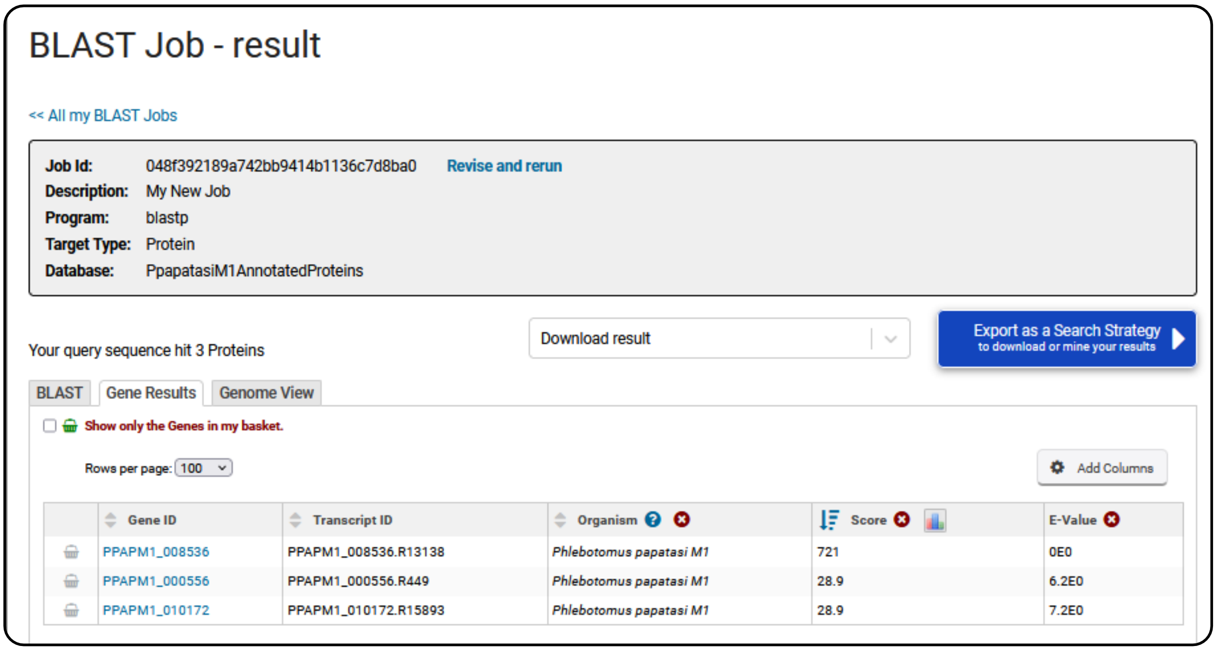
Input Sequence(s): Paste one or several sequences, or upload a FASTA file.

```
>PPAI008913
MTGFIFNIDGGYLEGLCRGFKCGILKQSDYLNLVQCETLEDLKLHLQGTDYGQFLANERSPLTVSVIDDKLREKLVIEFQ
HMRNHAVEPLSTFLDFITYSYMIDNIILLITGTLHQRRISELIPKCHPLGSFEQMEAIHVAATPAELYNAVLVDTPLAPF
FVDCISEQDLDEMNIEIIRNTLYKAYLEAFYDFCKKIGGTTADVMCEILAFEADRRAIIITINSFGTELSKDDRAKLYPR
CGKMNPDGLAALARADDYEQVKAVAEYYAEYAALFDGSGNNPGDKTLEDKFFEHEVKLNVFAFIQQFHFGVFYSYLKLKE
QECRNIVWIAECVAQKHRAKIDNYIRIF
```

Browse... No file selected.

Job Description: My New Job

This will return the blast result, a list of matching genes, and where those genes appear on the genome. Links to the gene pages of the matching genes can be found on any of the results tabs.



BLAST Job - result

<< All my BLAST Jobs

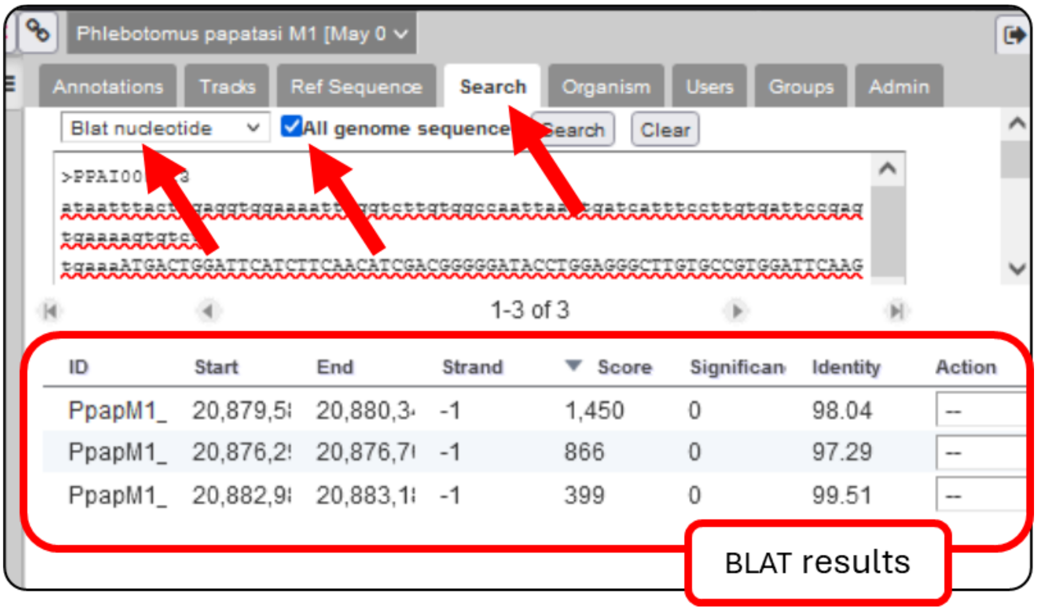
Job Id: 048f392189a742bb9414b1136c7d8ba0 Revise and rerun
Description: My New Job
Program: blastp
Target Type: Protein
Database: PpapatasiM1AnnotatedProteins

Your query sequence hit 3 Proteins

Gene ID	Transcript ID	Organism	Score	E-Value
PPAPM1_008536	PPAPM1_008536.R13138	*Phlebotomus papatasi M1*	721	0E0
PPAPM1_000556	PPAPM1_000556.R449	*Phlebotomus papatasi M1*	28.9	6.2E0
PPAPM1_010172	PPAPM1_010172.R15893	*Phlebotomus papatasi M1*	28.9	7.2E0

With BLAT

Copy the protein or nucleotide sequence to the search text box on the reference genome on Apollo and run BLAT. Select the correct type of BLAT search and check "all genome sequences".



ID	Start	End	Strand	Score	Significan	Identity	Action
PpapM1_	20,879,5	20,880,3	-1	1,450	0	98.04	--
PpapM1_	20,876,2	20,876,7	-1	866	0	97.29	--
PpapM1_	20,882,9	20,883,1	-1	399	0	99.51	--

4. Liftoff track

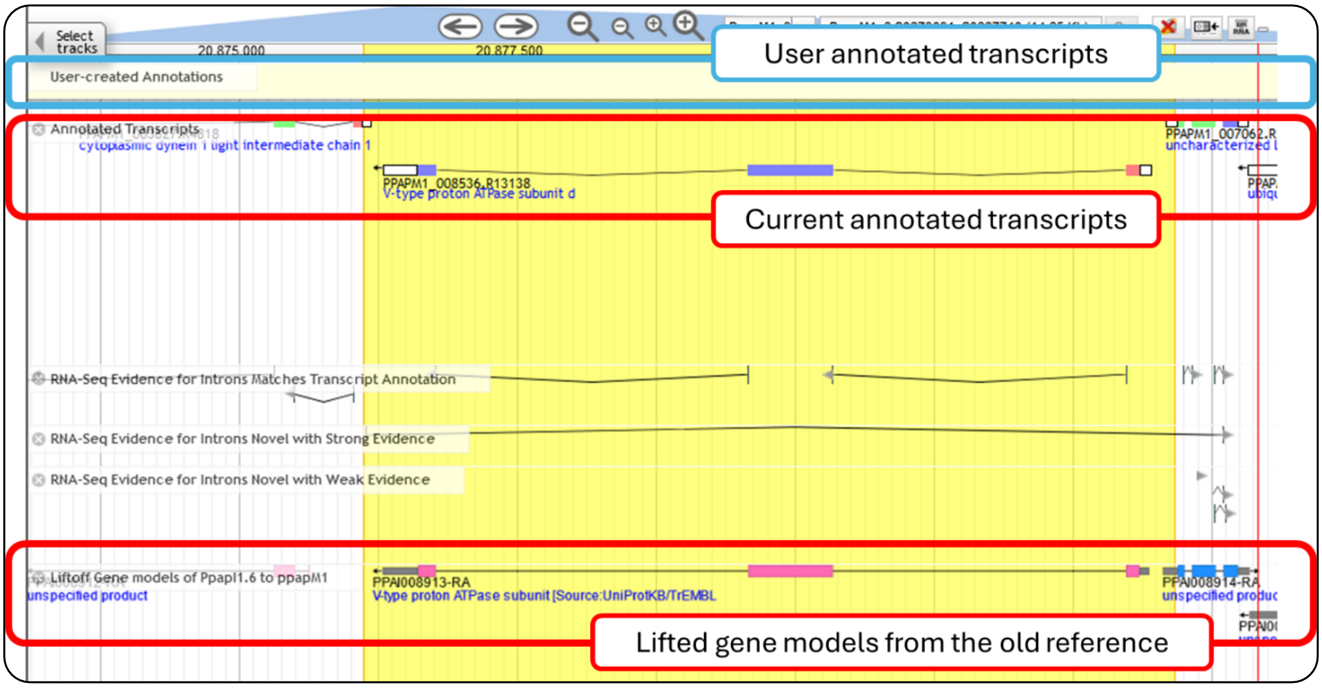
Liftoff has been used for some genomes to locate exactly where old gene annotations might be located on a new reference genome. Liftoff uses minimap2 to map gene models between assemblies with high accuracy. Approximately 90% to 99% of gene models can be transferred from one genome to another in this way. However, due to differences in assembly sequence these lifted gene models may contain errors and must be cross referenced with other available evidence.

The Liftoff track on Apollo or Jbrowse can be used to compare current annotations with gene models from the old reference genome. If your genome has a Liftoff track, you can find it in the “select track tab” to add it to the display.

Links to the source gene (the gene model on the original genome) are accessible by clicking on the gene model in the liftoff track.

In JBrowse, it looks like this:

In Apollo, it looks like this:

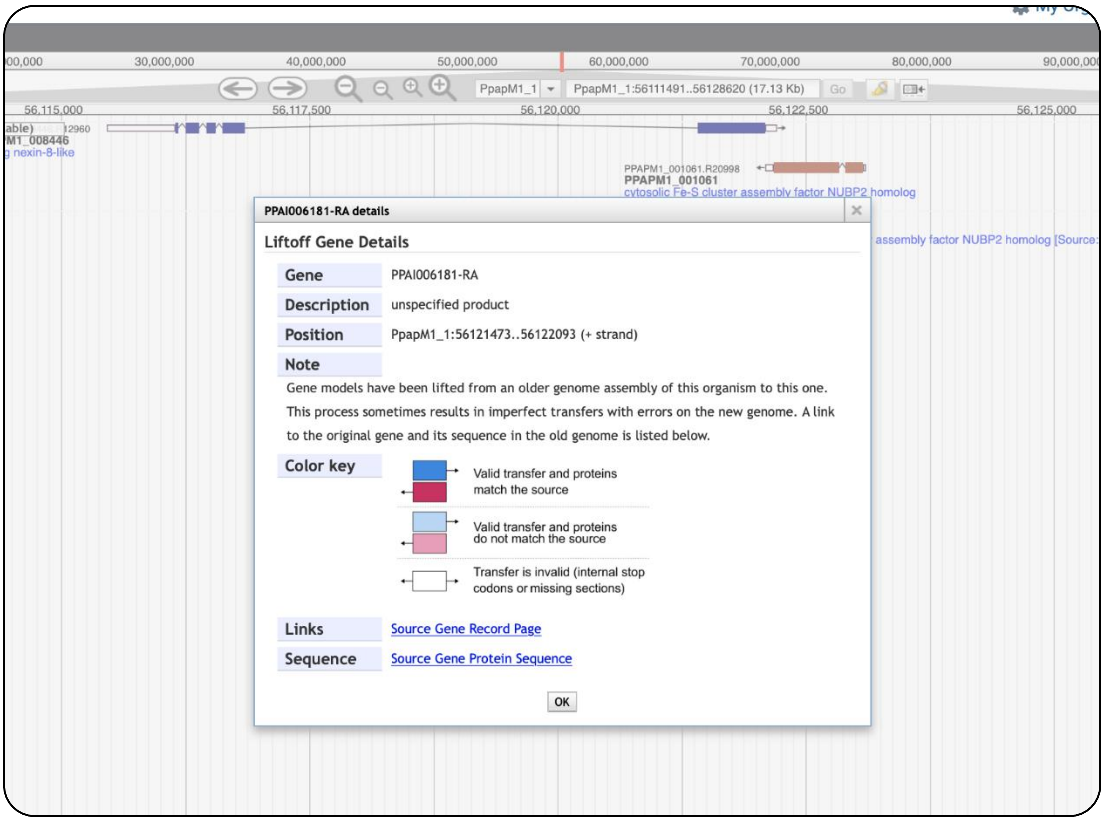


User annotated transcripts

Current annotated transcripts

Lifted gene models from the old reference

From release 68, the accuracy of Lifted genes will be displayed as colour coded glyphs as follows:



PPAI006181-RA details

Liftoff Gene Details

Gene	PPAI006181-RA
Description	unspecified product
Position	PpapM1_1:56121473..56122093 (+ strand)
Note	

Gene models have been lifted from an older genome assembly of this organism to this one. This process sometimes results in imperfect transfers with errors on the new genome. A link to the original gene and its sequence in the old genome is listed below.

Color key

- Valid transfer and proteins match the source
- Valid transfer and proteins do not match the source
- Transfer is invalid (internal stop codons or missing sections)

Links	Source Gene Record Page
Sequence	Source Gene Protein Sequence

OK