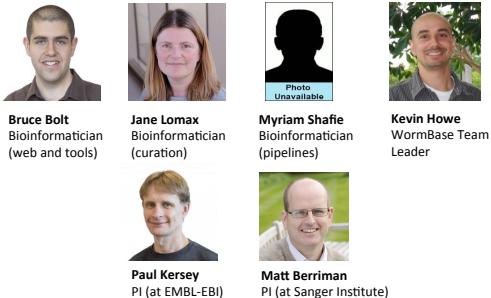


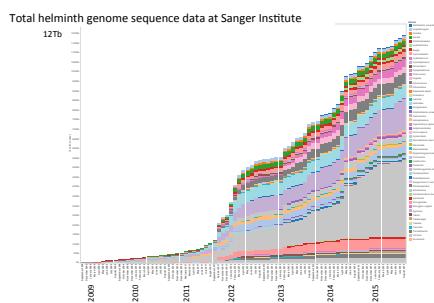
WormBase ParaSite Workshop

Glasgow
24th February 2016

WormBase ParaSite Team



An explosion of parasitic worm genomes



Introduction to WormBase ParaSite

- Collaboration between EMBL-EBI and Sanger Institute
- Funded by BBSRC for three years
- Launched September 2014
- Features both nematodes (roundworms) and platyhelminthes (flatworms) genomes
- No additional curation for most genomes
- Focus on rapid availability of new data
- Automated pipelines run over all genomes

Current release

- Release 5
 - 2,070,948 genes
 - 108 genomes
 - 99 species

(Including nine free living nematodes from WormBase for comparative purposes)



The Data

- All genomes are shown “as supplied” by the submitter (except WormBase “core” genomes)
- Varying levels of coverage and quality
- Transcriptomic data annotated and displayed on browser
- We welcome new data submissions (genomic, transcriptomic and variation data)

WormBase “Core” Parasite Genomes

- These are:
 - *Brugia malayi*
 - *Onchocerca volvulus*
 - *Pristionchus pacificus*
 - *Strongyloides ratti*
- Receive more care and attention
- Community driven manual curation
- Displayed in both WormBase and WormBase ParaSite

The Website

- Genome Browser
- Transcriptomic Data Display
- Gene, transcript and protein information pages
- Comparative Genomics
- Sequence Similarity Search (BLAST)
- Variant Effect Predictor (VEP) *
- Advanced Search Tool (BioMart)
- Access to BioMart data using R *
- Programmatic Access (REST API) *

* = Not covered today – speak to us for more information

WormBase and WormBase ParaSite

- wormbase.org is the home for highly curated data from *C. elegans* and other related nematodes
- Genes from “core” parasites also displayed here
- More genomic data for parasites available from parasite.wormbase.org



This afternoon’s agenda...

- 13:00 – 13:10 Introduction to WormBase ParaSite
- 13:10 – 13:50 Using the website
- 13:50 – 14:30 Sequence search with BLAST
- 14:30 – 15:00 Coffee Break
- 15:00 – 15:15 Comparative Genomics
- 15:15 – 15:50 Data Mining with BioMart
- 15:50 – 16:00 Opportunity to ask questions

Workshop Feedback

- Feedback form located on last page of workshop booklet
- Your feedback helps tailor future workshops
- We would be very grateful if you could complete this before leaving

Post-workshop Feedback

We would be grateful if you could spend a few moments giving us some feedback about today's workshop. Your feedback will help us to improve future workshops.

1. Did you see WormBase ParaSite before the workshop?
2. What do you think WormBase ParaSite needs after this workshop?
3. Would you recommend this workshop to your colleagues?
4. How would we rank the usefulness of the workshop?

Very useful	Useful	Not useful	Useless (not useful, if possible)
Very useful	Useful	Not useful	Useless (not useful, if possible)
5. How happy were you with each of the following?

Very satisfied	Satisfied	Not satisfied	Very unsatisfied
Very satisfied	Satisfied	Not satisfied	Very unsatisfied
6. Do you have any other comments or feedback?

Thank you!

Part 1: Browsing and searching

Part 1: summary

1. Front page
2. Locating genomes
3. Searching
4. Navigating genes, transcripts and scaffolds
5. Adding your data
6. User accounts

Front page

Front page

Front page

Front page

Front page: browse genomes

Genomes list						
Contents	Provider	Assembly	Build/Project ID	Taxonomy ID		
Nematoda (21)						
Panthermidae (2)						
Nematodes (19)						
Nematoda						
Species Name	Provider	Assembly	Build/Project ID	Taxonomy ID		
<i>Acantholaimus vittata</i>	University of Edinburgh	A.vittata_1.0	PRJNA20409	62270		
<i>Angylopygidae caninum</i>	Genome Institute at Washington University	A.caninum_3.0;qc-pg	PRJNA25939	90170		
<i>Angylopygidae oxytropis</i>	Genome Institute	A.oxytropis_1.0	PRJNA25940	93335		
<i>Angylopygidae pyriformis</i>	Genome Institute at Washington University	A.pyriformis_1.0;qc-pg	PRJNA25941	93345		
<i>Angylopygidae duodenale</i>	Genome Institute at Washington University	A.duodenale_3.2;qc-pg	PRJNA25941	93612		
<i>Angylopygidae heterodactylus</i>	Genome Institute at Washington University	A.heterodactylus_1.0;qc-pg	PRJNA25942	93613		
<i>Angylopygidae costaricensis</i>	Wellcome Trust Sanger Institute	A.costaricensis	PRJEB486	325428		
<i>Ascaris simplex</i>	Wellcome Trust Sanger Institute	A.simplex_1.0;qc-pg	PRJEB489	6281		
<i>Ascaris suum</i>	University of Edinburgh	A.suum_1.0;qc-pg	PRJNA20410	6281		
<i>Ascaris suum</i>	University of Melbourne	A.suum_1.0;qc-pg	PRJNA20410	6281		
<i>Bracon malayi</i>	Wellcome Trust Sanger Institute	B.malayi_3.1	PRJNA15209	6293		
<i>Bracon malayi</i>	Wellcome Trust Sanger Institute	B.malayi_3.1;qc-pg	PRJNA15209	6293		
<i>Bracon pachangi</i>	Wellcome Trust Sanger Institute	B.pachangi_3.0;qc-pg	PRJNA26127	6296		
<i>Bracon pachangi</i>	Wellcome Trust Sanger Institute	B.pachangi_3.0;qc-pg	PRJNA26127	6296		
<i>Braconophthorinae xylophaga</i>	Wellcome Trust Sanger Institute	AM2011h1; assembled	PRJEB4460	6298		
<i>Diplogastridae virgo</i>	Wellcome Trust Sanger Institute	D.virgo_1.0;qc-pg	PRJNA25116	91712		
<i>Diplogastridae virgo</i>	University of Edinburgh	D.virgo_1.0;qc-pg	PRJNA25116	91712		
<i>Diplogastridae virgo</i>	Genome Institute at Washington University	D.virgo_3.0;qc-pg	PRJNA25120	91712		
<i>Diplogastridae virgo</i>	University of Edinburgh	D.virgo_3.0;qc-pg	PRJNA25120	91712		
<i>Dicranolaimus medenensis</i>	Wellcome Trust Sanger Institute	D.medenensis_3.0;qc-pg	PRJEB600	91809		
<i>Elephora egypti</i>	Wellcome Trust Sanger Institute	E.egypti_1.0;qc-pg	PRJNA20302	116721		
<i>Elmironiella tenuis</i>	Wellcome Trust Sanger Institute	E.tenuis_1.0;qc-pg	PRJNA20303	116741		
<i>Globodera pallida</i>	Wellcome Trust Sanger Institute	G.pallida_1.0	PRJEB123	36000		
<i>Globodera pallida</i>	Genome Institute at Washington University	G.pallida_1.0;qc-pg	PRJNA25123	91713		
<i>Haemonchus contortus</i>	Wellcome Trust Sanger Institute	H.contortus_1.0;qc-pg	PRJNA20306	62980		
<i>Haemonchus contortus</i>	University of Melbourne	H.contortus_1.0;qc-pg	PRJNA20302	62980		
<i>Haemonchus contortus</i>	Genome Institute at Washington University	H.contortus_1.0;qc-pg	PRJNA20302	62980		
<i>Heligmosomidae polygyra</i>	Wellcome Trust Sanger Institute	H.polygyra_3.0;qc-pg	PRJEB103	97903		

Genome pages

Search Worldwide Parasites

[Species List](#)

[BLAST](#)

[REST API](#)

[Downloads](#)

[Wormbase](#)

[WormBase Parasite](#)

[Parasite Metabolism \(PRIM\)](#)

[PRIM](#)

[PRIM API](#)

[Search Worldwide Parasites](#)

[Search Worldwide Metabolism](#)

[Search Worldwide PRIM](#)

[Search](#)

[Home](#)

[Documentation](#)

Echinococcus multilocularis

OrthoDB 15.0.1 • One search interface for 1000+ genomes | Taxonomy | BLAST

Search Echinococcus multilocularis (PRIM) (22) | [A](#)

About Echinococcus multilocularis

The cestode *Echinococcus multilocularis* is a member of the Cyclophyllidae, which comprises the majority of tapeworms that are of medical importance. The disease caused by *Echinococcus*, which is caused by the larval stage of *E. multilocularis*, is called alveolar echinococcosis and is a zoonosis that has been reported in Europe, Asia, and North America. It is the most prevalent form of human tapeworm disease in the world.

What are PRIM? Protein-coding and non-coding genes, splice variants, cRNA and protein sequence.

[View information and statistics](#)

Genome assembly: ENA/PRJEB3406

The *E. multilocularis* reference genome was sequenced by the Parasite Genomics group at the Institute of Parasitology, University of Bern, Switzerland. The genome assembly was published in the final version of the genome was described in [Tug et al. \(2022\)](#). The assembly has been updated to include the latest version of the genome and was last updated on January 01, 2019.

 Example: region

Comparative genomics

What can I benefit? Homologs, gene trees, whole genome alignments across multiple species.

 Example: gene tree

Tools

What can I benefit? Comparative genomics, BLAST, PRIM, PRIM API.

- [BLAST](#) with links to data using the Worldwide Parasite database stored in PRIM
- [PRIM](#) (PRIM) with links to data using the PRIM API
- [PRIM API](#) (PRIM API) with links to data using the PRIM API

 Example: PRIM API

Downloads

- [Genomic Resource \(PRIM\)](#)
- [Genomic Resource \(PRIM API\)](#)
- [Sub-metabolic Genome \(PRIM\)](#)
- [PRIM](#)
- [Full-length transcript \(PRIM\)](#)
- [CDS transcript \(PRIM\)](#)

 Example: transcript

Search results	
One-eat-4 [WBGene0058445]	
Description	Protein vesicular glutamate transporter eat-4 [Source:UniProtKB/Swiss-Prot;Acc:P31461] (projected from <i>Caenorhabditis elegans</i> ortholog eat-4)
Gene ID	WBGene0058445
Species	<i>Onthophaga rufa</i> (Lytta/Blattidae)
Location	Q9QDC_OMA16437501:16443875
Gene trees	WBGT0080000116156
C. elegans orthologue	eat-4
 Bma-eat-4 [WBGene0027744]	
Description	Protein Bma-EAT-4 ; isoform b [Source:UniProtKB/TriEMBL;Acc:ABNM81]
Gene ID	WBGene0027744
Species	<i>Bacillus malayi</i> (P59410/298)
Location	Bma_01_scaf00025_2208-5498
Gene trees	WBGT0080000116165
C. elegans orthologue	eat-4

Filtering search results

Gene pages

Gene pages

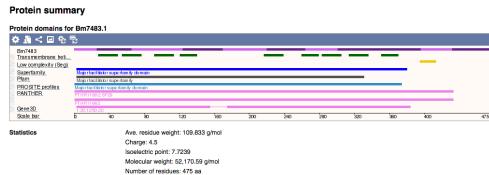
GO terms

Transcript pages: summary

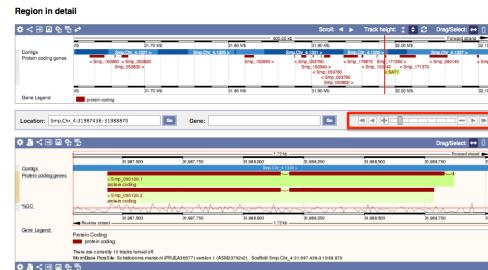
Transcript pages: navigating

5

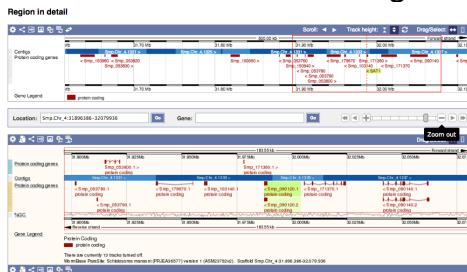
Transcript pages: protein domains



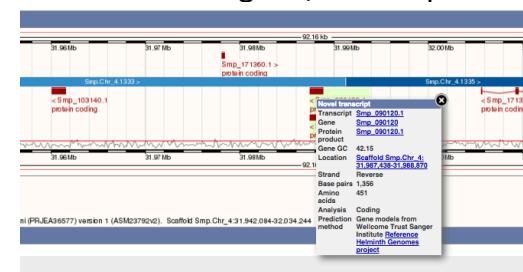
Location view: zooming



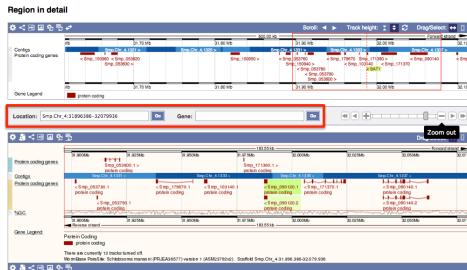
Location view: zooming



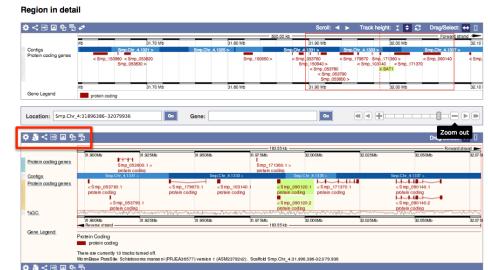
Location view: gene/transcript info



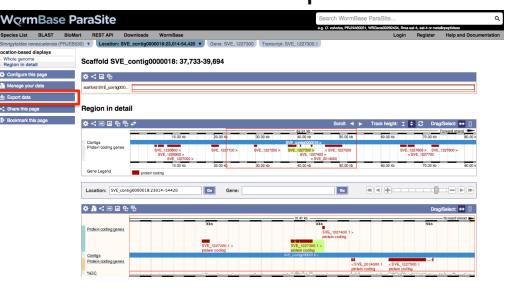
Location view: jump to...



Location view: configure



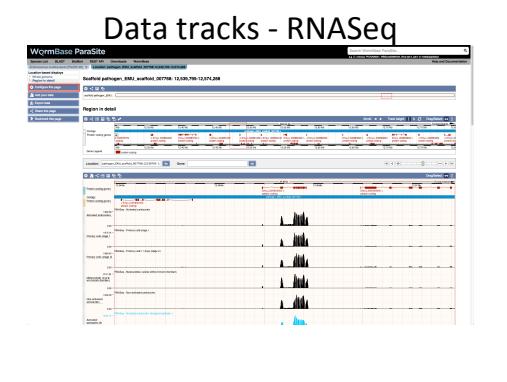
Location view: export data



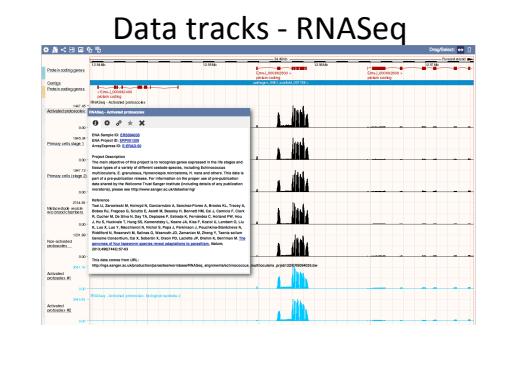
Location view: export data



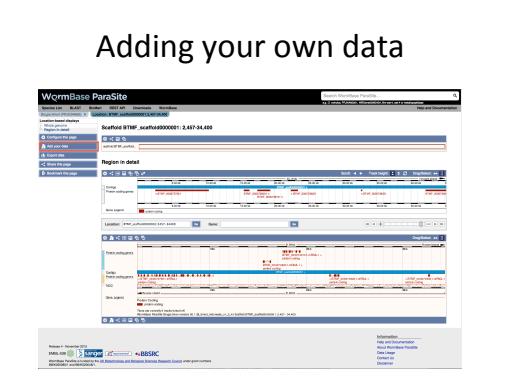
Data tracks - RNASeq



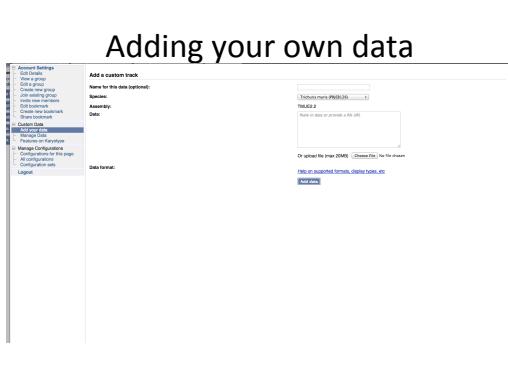
Data tracks - RNASeq



Adding your own data



Adding your own data



Adding your own data

User accounts

- Saving attached data tracks
- Sharing data tracks with collaborators
- Saving configuration settings

User accounts

User accounts: registering

User accounts

Part 2: Comparative Genomics in WormBase ParaSite

Introduction

- During each release, we compute phylogenetic trees
- Every gene is included from 120 species:
 - 99 helminths
 - 9 free-living nematodes
 - 12 comparator species (e.g. human, mouse, etc)
- Determine orthologues and paralogues

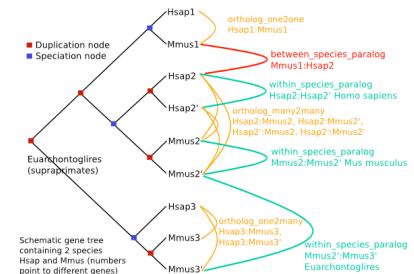
A word of caution...

- Trees are re-calculated between each release
- Homologies which are poorly defined may not be defined in next release
- Always check the %ID of each alignment

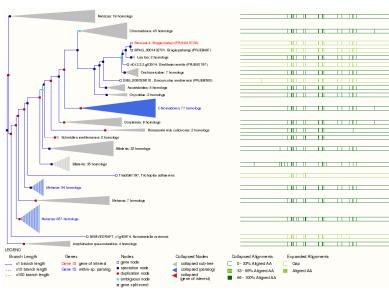
Homology types

- Orthologues: any gene pairwise relation where the ancestor node is a speciation event
 - 1-to-1 orthologue
 - 1-to-many orthologue
 - Many-to-many orthologue
- Paralogues: any pairwise relation where the ancestor node is a duplication event

Understanding the gene tree



Visual access to the trees



Tabular access to tree data

Selected orthologues						
View protein alignments of all orthologues						
Group	Type	Ortho	Status ID & gene name	Compare	Location	Filter
Arachnophoroproteins (21303)	1-to-1	n/a	Abu.1.0.20136	• Alignment present • Alignment 100% • Gene Tree (inferred)	Ortho.1.0.ac00000730007-07666.1	73 76
Arachnophoroproteins (21303)	Many-to-many	n/a	Abu.2.0.10136	• Alignment present • Alignment 100% • Gene Tree (inferred)	Ortho.2.0.ac00000730007-07666.1	27 28
Arachnophoroproteins (21303)	Many-to-many	n/a	Abu.2.0.10136	No description	Ortho.2.0.ac00000730007-07666.1	26 29
Arachnophoroproteins (21303)	Many-to-many	n/a	Abu.2.0.10136	No description	Ortho.2.0.ac00000730007-07666.1	28 29
Arachnophoroproteins (21303)	Many-to-many	n/a	Abu.2.0.10136	No description	Ortho.2.0.ac00000730007-07666.1	28 30
Arachnophoroproteins (21303)	Many-to-many	n/a	Abu.2.0.10136	No description	Ortho.2.0.ac00000730007-07666.1	28 30
Arachnophoroproteins (21303)	1-to-1	n/a	ANCAN_09056	• Alignment present • Alignment 100% • Gene Tree (inferred)	Ortho.1.0.ac00000730007-07666.1	98 60
Arachnophoroproteins (21303)	1-to-1	n/a	Abcb11_00001	• Alignment present • Alignment 100% • Gene Tree (inferred)	Ortho.1.0.ac00000730007-07666.1	98 67

Part 3: Sequence Similarity Search using BLAST

What is BLAST?

- BLAST = Basic Local Alignment Search Tool
- Sequence similarity tool
- Allows comparison of a **query** sequence, against a **database** of sequences
- Query = your nucleotide or protein sequence
- Database = the genome or proteome of any species

What is BLAST?

- Input:
Nucleotide or protein sequence
Search Parameters
- Output:
List of all hits ranked in order of statistical significance

Types of BLAST

BLAST Type	Query Sequence	Target Database
BLASTN	Nucleotide	Genome (nucleotide)
BLASTP	Peptide	Proteome (peptide)
BLASTX	Six frame translation of a nucleotide sequence	Proteome (peptide)
TBLASTX (slowest)	Six frame translation of a nucleotide sequence	Six frame translation of genome
TBLASTN	Peptide	Six frame translation of genome

Using the ParaSite BLAST

The screenshot shows the WormBase ParaSite interface. At the top, the species is set to **B. malayi** (strain **INRA1295**). The main navigation bar includes **BLAST**, **blast**, **blast**, **REST API**, **Downloads**, and **WormBase**. The location is specified as **Bm147_v9_scaffold11_3,207,430-3,210,542**. The gene identifier is **Gene: Bm147**. The gene display section for **Bm147** (WBGene00222408) includes a summary, splice variants, aliases, external references, ontologies (GO: Molecular function, GO: Cellular component), and a forward-facing link. Below this, the gene detail section for **Bm147** (WBGene00222408) shows the location (Bm147_v9_scaffold11_3,207,430-3,210,542), a summary (SuperConfig Bm147_v9_scaffold11_3,207,430-3,210,542 forward), and details about the gene (transcripts, orthologs, protein coding).

Using the ParaSite BLAST

The figure shows a screenshot of the ParaSite BLAST interface. At the top, the title "Using the ParaSite BLAST" is displayed. Below the title, there are sections for "Gene type", "Annotation Method", "Protein coding model imported from WormBase®", and a "Show transcript table" link. A large section titled "Marked-up sequence" contains a sequence viewer with a "Download sequence" and "BLAST this sequence" button. The sequence viewer highlights a specific region with a red box, and a "Marked selected sequence" link is located below the highlighted region. The sequence itself is a long string of DNA bases.

Making sense of the results

Making sense of the results

Part 4: Data-mining with BioMart

Data-mining with BioMart

Search WormBase ParaSite

WormBase ParaSite

Wormbase Home | ParaSite Home

New Search Results

All Species (WP54)

[None selected]

Attributes

Gene/protein ID

GeneOntology

Gene/protein ID

Please restrict your query using criteria below
(If the values are truncated in any lists, hover over the list item to see the full text)

Species

REGION

GENE

PROTEIN

GENE ONTOLOGY

MULTI-SPECIES COMPARISONS

PROTEIN DOMAINS

- **SPECIES:** Use this filter to select either individual genomes or nematode clades.
 - Multiple genomes can be selected by holding down the ctrl key or the option key on a Mac.

- **REGION:** Restrict to a particular genomic region.
 - Should only be used where a single genome has been selected, as it is possible that a particular region is present in multiple genomes.
 - If start/end co-ordinates are being specified, a scaffold or chromosome id is always required.
 - Where multiple regions are specified, the format is 'Scaffold/Chr:Start:End:Strand' e.g. AG00032:411187:446321:1.
 - If no strand is specified, both strands are selected.
 - Regions should be separated by a comma or new line.

- **GENE:** Specify a list of genes with WormBase IDs, or one of the other ID types listed.
 - IDs should be separated by a new line.

- **GENE ONTOLOGY:** Restrict by one or more Gene Ontology (GO) terms for functional descriptions.
 - Paste or upload a list of GO IDs or use the autocomplete box to populate the list.
 - Alternatively restrict to a particular GO evidence type e.g. Inferred by Electronic Annotation (IEA).
 - Multiple codes can be selected by holding down the ctrl key, or option key on a Mac.
- **PROTEIN DOMAINS:** Allows you to restrict your query based on the presence or absence of protein domains.
 - **Limit to genes...** lets you choose a particular database feature set to include or exclude e.g. "restrict to all proteins containing any feature found in Pfam".
 - **Limit to genes with these family or domain IDs:** allows you to restrict to one or more protein domains/families.
 - Accepts IDs from several databases including InterPro, Pfam and Panther. IDs should be separated by a new line.

“I’d like to extract all *C. elegans* orthologs for *Nippostrongylus* genes involved in a particular process.”

1. In the SPECIES menu select *Nippostrongylus*
2. In the MULTI-SPECIES COMPARISONS menu select **Orthologous *C. elegans* genes -> Only**
3. Further refine this list by function, process or location by choosing one or more categories from the GENE ONTOLOGY list.
 - Start typing in the upper box and choose your terms of interest from the autocomplete, they will be added to the box beneath.
4. Click the **Results** button (top left) to see your results. By default a two-column file is returned that contains gene ID and Genome Project. To configure different options for the output, select **Attributes** in the left menu.

"I have a list of genes from *Ascaris suum* and would like to know which ones have orthologs in humans and mammals and which ones might be nematode-specific."

- In the GENE menu paste in your gene list
- in the MULTI-SPECIES COMPARISONS select **Orthologous human genes -> Excluded**
- You can also run this query against against mouse orthologs by selecting **Orthologous mouse genes -> Excluded** (the results are the same in this case)
- Click the Results button (top left) to see your results. By default a two-column file is returned that contains gene ID and Genome Project. To configure different options for the output, select **Attributes** in the left menu.

"I need the sequences for a set of *Schistosoma mansoni* genes. I have the chromosome, start, and stop for each."

- From the **SPECIES** filter choose *Schistosoma mansoni*.
- Open the **REGION** section and enter the list of co-ordinates under 'Multiple regions' separated by commas or new lines.
- In **Attributes**, check the **Sequences** option, then in the **SEQUENCES** section choose **Unspliced (genes)**.
- Click the **Results** button

"I need a list of genes with predicted signal peptide that are present in *Brugia malayi* a given organism but not present in *C. elegans*."

- In the **SPECIES** section choose *Brugia malayi*, then in the **MULTI-SPECIES COMPARISONS** select **Orthologous C. elegans** genes -> Excluded
- In the **PROTEIN DOMAINS** section check **Limit to genes...**
- From the menu select **with signal P protein features -> Only**
- Click the **Results** button (top left) to see your results. By default a two-column file is returned that contains gene ID and Genome Project. To configure different options for the output, select **Attributes** in the left menu.