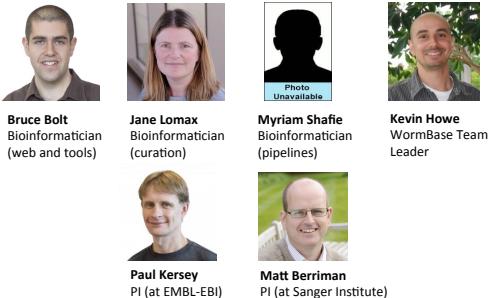


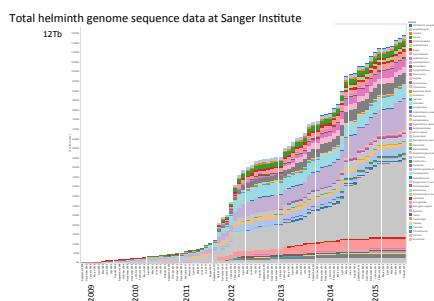
WormBase ParaSite Workshop

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 comments will help us to improve future workshops.			
1. Did you see WormBase ParaSite before the workshop? <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No			
2. What do you think WormBase ParaSite needs after this workshop? (check all that apply)			
3. Would you recommend this workshop to your colleagues? (check all that apply)			
4. How useful are each section of the workshop?			
Browsing the website	Very Useful	Useful	Not Useful
BLAST	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Comparative Genomics	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Data mining	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
5. How happy were you with each of the following?			
Topics Covered	Very Happy	Happy	Not Happy
Level	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Interactivity	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Quality of presentations	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Exercises	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Balance of presentations and exercises	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
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

The screenshot shows the WormBase ParaSite front page. The top navigation bar includes links for Species List, BLAST, BioMart, REST API, Downloads, and WormBase. The main content area features a "Find a genome" search bar and a "Species" section with icons for Species, BLAST, BioMart, API, Downloads, and WormBase. Below this is an "Announcements" box about the release of WWP54, followed by a "Blog" section with a post about the Nobel Prize for work on parasitic nematodes. A "Statistics" box shows version information (Version: WWP54 (November 2015), WormBase Version: WS230, 99 genomes, representing 90 species). On the right, there is a "Twitter" sidebar with several tweets from the official WormBase account.

Front page

This screenshot is identical to the one above, but with a red box highlighting the "Find a genome" section. This section contains a search bar and dropdown menus for selecting species like Nematoda (Roundworms) or Platyhelminthes (Flatworms).

Front page

This screenshot is identical to the one above, but with a red box highlighting the "Find a genome" section. This section contains a search bar and dropdown menus for selecting species like Nematoda (Roundworms) or Platyhelminthes (Flatworms).

Front page

This screenshot is identical to the one above, but with a red box highlighting the "Find a genome" section. This section contains a search bar and dropdown menus for selecting species like Nematoda (Roundworms) or Platyhelminthes (Flatworms).

Front page: browse genomes

This screenshot is identical to the one above, but with a red box highlighting the "Find a genome" section. This section contains a search bar and dropdown menus for selecting species like Nematoda (Roundworms) or Platyhelminthes (Flatworms).

Locating genomes

Genomes list

Genome pages

Searching

Search results

Search results

Filtering search results

Gene pages

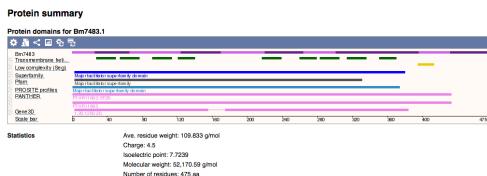
Gene pages

GO terms

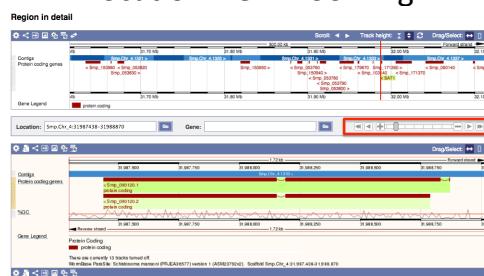
Transcript pages: summary

Transcript pages: navigating

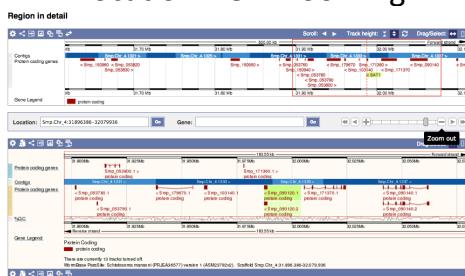
Transcript pages: protein domains



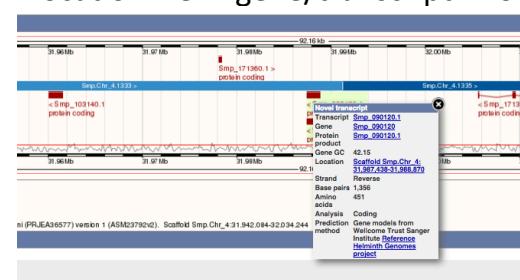
Location view: zooming



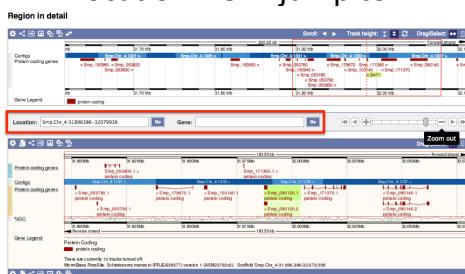
Location view: zooming



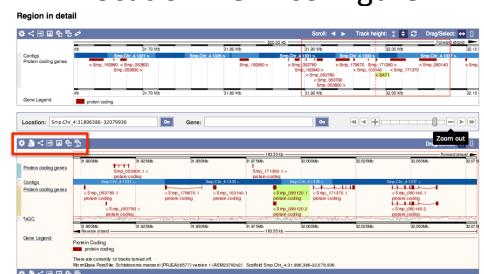
Location view: gene/transcript info



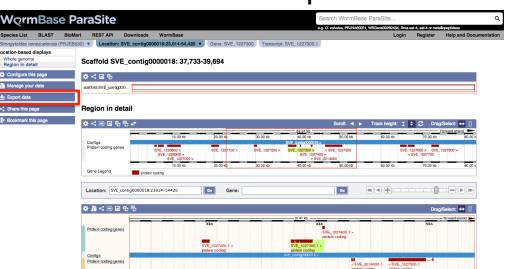
Location view: jump to...



Location view: configure



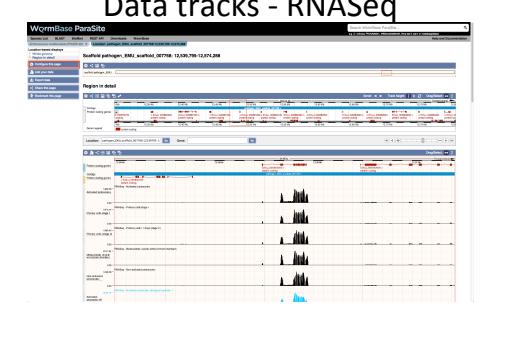
Location view: export data



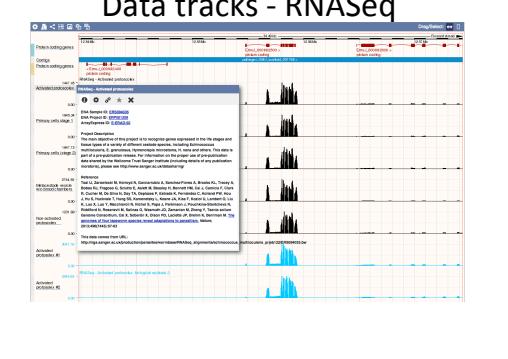
Location view: export data



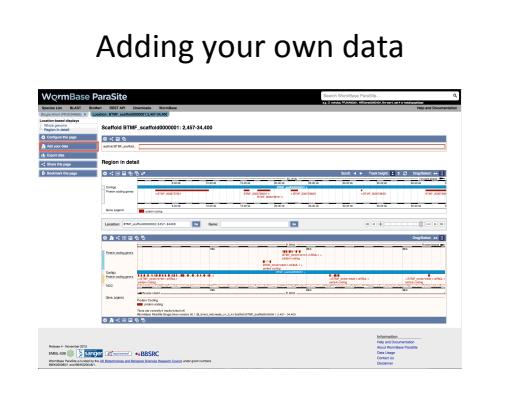
Data tracks - RNASeq



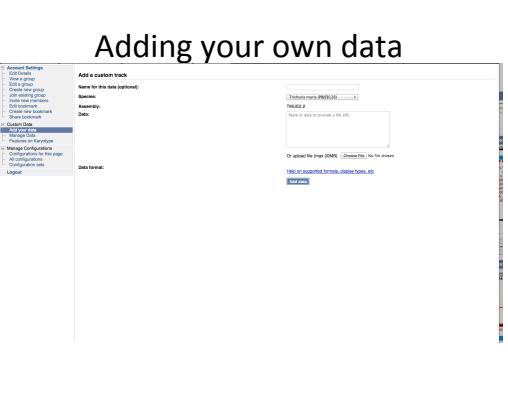
Data tracks - RNASeq



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

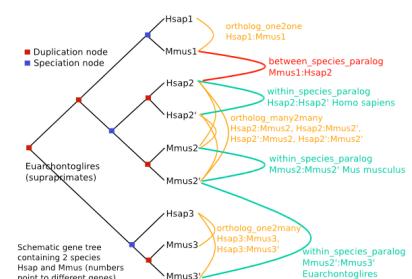
A word of caution...

- 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

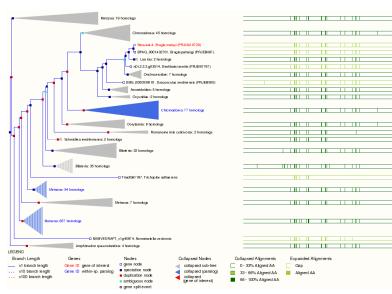
Homology types

- 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		Showcase columns				Filter	
Species	Type	dRNA	Status ID & gene name	Comments	Location	Target	Query
<i>Anthonomus grandis</i> (Strain: 17142306)	1-to-1	n/a	Adult-17142306	Probable nuclear gapase homolog; contains 4 repeats before ortholog to C. elegans gapase. Contains 1 repeat after gapase.	(Adult-1.0) Adut-17142306-17142306-1	73	76
<i>Anthonomus grandis</i> (Strain: 17142306)	Many-to-many	n/a	Adut-17142306		Contact170119-17019600-1	27	26
<i>Aztecinae</i> (Strain: 17142306)	Many-to-many	n/a	Adut-17142306	No description	Contact170119-17019600-1	26	29
<i>Aztecinae</i> (Strain: 17142306)	Many-to-many	n/a	Adut-17142306	No description	Contact170119-17019600-1	28	23
<i>Aztecinae</i> (Strain: 17142306)	Many-to-many	n/a	Adut-17142306	No description	Contact170119-17019600-1	28	30
<i>Anthonomus grandis</i> (Strain: 17142306)	1-to-1	n/a	ANCON_00055	Probable nuclear gapase homolog; contains 4 repeats before ortholog to C. elegans gapase. Contains 1 repeat after gapase.	(ANCON_0.0) ANCON_00055-17142306-1	98	60
<i>Anthonomus grandis</i> (Strain: 17142306)	1-to-1	n/a	ANCON_01370	Probable nuclear gapase homolog; contains 4 repeats before ortholog to C. elegans gapase. Contains 1 repeat after gapase.	(ANCON_0.0) ANCON_01370-17142306-1	99	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 homepage. The top navigation bar includes links for Species List, BLAST, BioMart, REST API, Downloads, and WormBase. The BLAST link is highlighted with a red box. Below the navigation, there's a search bar with the text "Brugia malayi" and a dropdown menu showing "Bm1A10290". The main content area displays a gene record for "Gene: Bm2147 WBGen00222408". It shows the location as "SuperContig Bmal_v3_scaffold1:3,207,430-3,210,542" and the gene type as "Protein coding". A note indicates "This gene has 2 transcripts (splice variants) and 112 orthologues". The sidebar on the left contains sections for Gene-based displays (Splice variants, Sequence, External references), Other genes (GO: Molecular function, GO: Cellular component), and a "Species List" section.

Using the ParaSite BLAST

The screenshot shows the results of a BLAST search. At the top, it says "Sequence: Bmal_v3_scaffold1:3,207,430-3,210,542 forward strand." and "This gene has 2 transcripts (splice variants) and 112 orthologues". Below this, there's a "Marked-up sequence" section with a "Download sequence" button and a checked "BLAST this sequence" button. A "Run BLAST" button is also present. The sequence itself is a long string of DNA bases. On the right side, there's a "Location" panel with tabs for "About this gene", "Gene type", "Annotation Method", and "Transcripts". The "About this gene" tab is active, showing details like "Protein coding" and "Protein-coding model imported from WormBase".

Using the ParaSite BLAST

ParaSite Home

Gene type: Protein coding
Annotation Method: Protein-coding model imported from WormBase®
Transcripts: Show transcript table

Marked-up sequence:

Exons: Bm2147 exon All exons in this region

Download sequence BLAST this sequence

```
>supercontig1B_malayi-3.1:Real_v3_scaffold1:32048830:3211142:1
ATTTACTCTGATTTTCAGTACCTACTACCCAAACATTATTTTGACTCTTGTG
CTACTCTGATTTTCAGTACCTACTACCCAAACATTATTTTGACTCTTGTG
TGAGCTACTATTTTGAGTACCTACTACCCAAACATTATTTTGACTCTTGTG
TTAGCTACTATTTTGAGTACCTACTACCCAAACATTATTTTGACTCTTGTG
TTAGACTCTTGTGTCAGTGGCAAAATTATTTAGAACATAGAATTTCGATCAGTATTTAAAAA
AAATGACCTGCAATTTCGATCAGTGGCAAAATTATTTAGAACATAGAATTTCGATCAGTATTTAAAAA
ATAGACCTGCAATTTCGATCAGTGGCAAAATTATTTAGAACATAGAATTTCGATCAGTATTTAAAAA
ATAAGATGATTCGATCAGTGGCAAAATTATTTAGAACATAGAATTTCGATCAGTATTTAAAAA
ATAAGATGATTCGATCAGTGGCAAAATTATTTAGAACATAGAATTTCGATCAGTATTTAAAAA
```

WormBase ParaSite

Gene: Bm-a-eat-4 wbeGene00227744

Summary: Gene summary
Details: Gene details
Sequence: Gene sequence
Annotations: Gene annotations

- GO: Molecular function
- GO: Cellular component
- GO: Biological process
- Protein domains
- Gene tree
- Paralogous

Annotation Method: Protein-coding model imported from WormBase®

Transcripts: Show transcript table

Marked-up sequence:

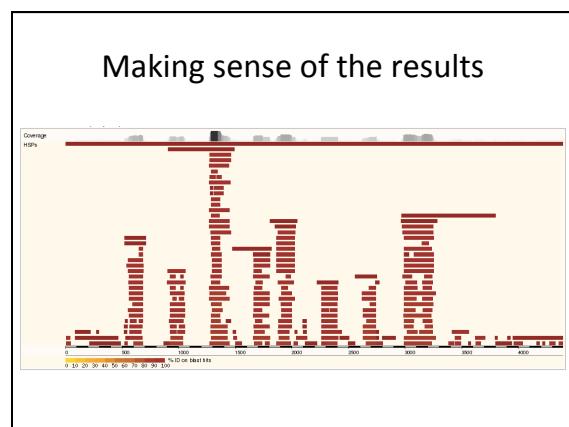
Download sequence BLAST this sequence

Exons: Bm-a-eat-4 exon All exons in this region

```
>supercontig1B_malayi-3.1:Real_v3_scaffold1:32048830:3211142:1
ATTTACTCTGATTTTCAGTACCTACTACCCAAACATTATTTTGACTCTTGTG
CTACTCTGATTTTCAGTACCTACTACCCAAACATTATTTTGACTCTTGTG
TGAGCTACTATTTTGAGTACCTACTACCCAAACATTATTTTGACTCTTGTG
TTAGCTACTATTTTGAGTACCTACTACCCAAACATTATTTTGACTCTTGTG
TTAGACTCTTGTGTCAGTGGCAAAATTATTTAGAACATAGAATTTCGATCAGTATTTAAAAA
AAATGACCTGCAATTTCGATCAGTGGCAAAATTATTTAGAACATAGAATTTCGATCAGTATTTAAAAA
ATAGACCTGCAATTTCGATCAGTGGCAAAATTATTTAGAACATAGAATTTCGATCAGTATTTAAAAA
ATAGACCTGCAATTTCGATCAGTGGCAAAATTATTTAGAACATAGAATTTCGATCAGTATTTAAAAA
ATAAGATGATTCGATCAGTGGCAAAATTATTTAGAACATAGAATTTCGATCAGTATTTAAAAA
ATAAGATGATTCGATCAGTGGCAAAATTATTTAGAACATAGAATTTCGATCAGTATTTAAAAA
```

Making sense of the results

- Score**
Used to assess the biological relevance by describing the alignment quality
Higher score = higher similarity
- E-value**
Probability that event occurred by chance (in short, a p-value that has been corrected for multiple testing)
Lower E-value = more significant result
- %ID**
Percentage of your query sequence that matches the genome/proteome database



Part 4: Data-mining with BioMart

Data-mining with BioMart

WormBase ParaSite

Dataset: All Species (WBPS4)

Attributes:

- [None selected]
- Gene
- Gene product
- Gene stable ID

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

Species:

- GENE
- GENE ONTOLOGY
- MULTI-SPECIES COMPARISONS
- PROTEIN DOMAINS

The screenshot shows the WormBase ParaSite search interface. At the top, there's a navigation bar with links for 'Home', 'All Genomes', 'Parasite', 'Search', 'Help', and 'Downloads'. Below this is a search bar with placeholder text 'Search Wormbase results...'. Underneath the search bar is a 'Dataset' dropdown set to 'All Species (WBPS4)'. To the right of the dataset dropdown is a 'Filter' dropdown with several options: 'SPECIES', 'REGION', 'GENE', 'GENE ONTOLOGY', 'MULTI-SPECIES COMPARISONS', and 'PROTEIN DOMAINS'. The 'PROTEIN DOMAINS' option is highlighted with a red box.

- **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.

BioMart output

The screenshot shows the BioMart search results for the dataset "All Species (WPS4)". The search term "mRNA" has been entered. The results table lists numerous entries, each containing a "Gene stable ID" and a "Gene name". The columns include "Gene stable ID", "Gene name", "Organism", "Assembly accession", "Protein stable ID", and "Homologene links". A "View" link is present at the bottom of the table.

Setting Attributes (output): features

This screenshot shows the "Setting Attributes (output)" feature for the "Features" category in WormBase ParaSite. It displays a list of attributes such as "Gene stable ID", "Gene name", "Organism", "Assembly accession", "Protein stable ID", and "Homologene links". On the right, there is a panel titled "Please select columns to be included in the output and hit 'Results'" which contains checkboxes for selecting specific columns.

Setting Attributes (output): structures

This screenshot shows the "Setting Attributes (output)" feature for the "Structures & Sequences" category in WormBase ParaSite. It displays a list of attributes including "Gene stable ID", "Gene name", "Organism", "Assembly accession", "Protein stable ID", and "Homologene links". The right panel for selecting output columns is also visible.

Setting Attributes (output): homologues

This screenshot shows the "Setting Attributes (output)" feature for the "Homologues" category in WormBase ParaSite. It displays a list of attributes including "Gene stable ID", "Gene name", "Organism", "Assembly accession", "Protein stable ID", and "Homologene links". The right panel for selecting output columns is also visible.

Setting Attributes (output): sequence

This screenshot shows the "Setting Attributes (output)" feature for the "Sequence" category in WormBase ParaSite. It displays a list of attributes including "Gene stable ID", "Gene name", "Organism", "Assembly accession", "Protein stable ID", and "Homologene links". The right panel for selecting output columns is also visible.

"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 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

The screenshot shows the WormBase ParaSite search interface. In the 'Dataset' dropdown, 'Schistosoma mansoni' is selected. Under 'Species', 'Parasites' is chosen. In the 'REGION' section, 'Multiple regions' is selected. The 'SEQUENCES' section has 'Unspliced (genes)' checked. On the right, there are options for output formats: 'FASTA' (selected), 'GFF', 'CDS', 'Gene feature', 'CDS sequence', 'CDS transcript', and 'Cgi'. At the bottom, there are 'Output' and 'Overviews' buttons.

"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.

The screenshot shows the WormBase ParaSite search interface. In the 'Dataset' dropdown, 'Brugia malayi' is selected. Under 'Species', 'Parasites' is chosen. In the 'MULTI-SPECIES COMPARISONS' section, 'Orthologous C. elegans genes -> Excluded' is selected. In the 'PROTEIN DOMAINS' section, 'Limit to genes' is checked. At the bottom, there are 'Output' and 'Overviews' buttons.