WormBase ParaSite Workshop

Glasgow

24th February 2016

WormBase ParaSite Team

Bruce Bolt
Bioinformatician
(web and tools)

Jane Lomax
Bioinformatician
(curation)

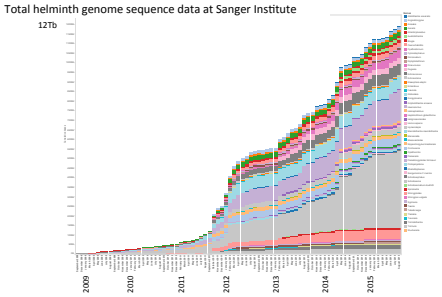
Myriam Shafie
Bioinformatician
(pipelines)

Kevin Howe
WormBase Team
Leader

Paul Kersey
PI (at EMBL-EBI)

Matt Berriman
PI (at Sanger Institute)

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

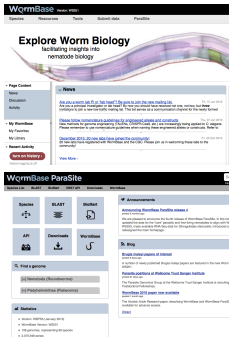
The Website

- 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

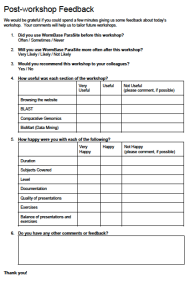


This afternoon's agenda…

- 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 minutes giving us some feedback about today's workshop. Your comments will help us to tailor future workshops.

1. Did you use WormBase ParaSite before this workshop?
   Often / Sometimes / Never
2. Will you use WormBase ParaSite more often after this workshop?
   Very Likely / Likely / Not Likely
3. Would you recommend this workshop to your colleagues?
   Yes / No
4. How useful was each section of the workshop?

	Very Useful	Useful	Not Useful (please comment, if possible)
Browsing the website			
BLAST			
Comparative Genomics			
BioMart (Data Mining)			

5. How happy were you with each of the following?

	Very Happy	Happy	Not Happy (please comment, if possible)
Duration			
Subjects Covered			
Level			
Documentation			
Quality of presentations			
Exercises			
Balance of presentations and exercises			

6. Do you have any other comments or feedback?

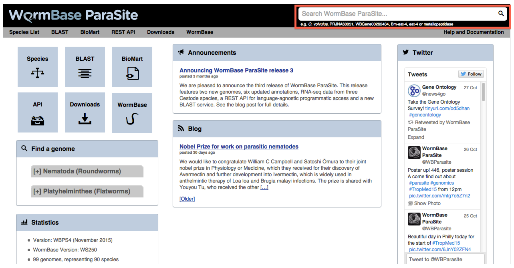
Thank you!

Part 1: Browsing and searching

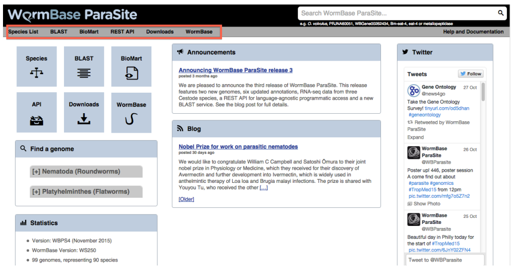
Part 1: summary

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

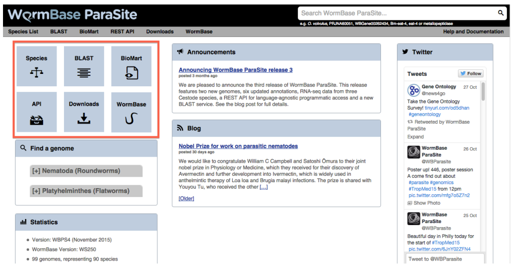
Front page



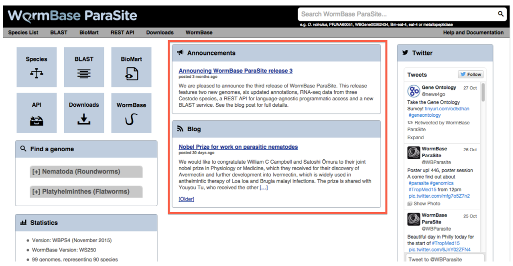
Front page



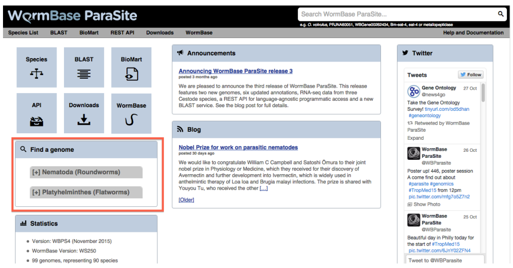
Front page



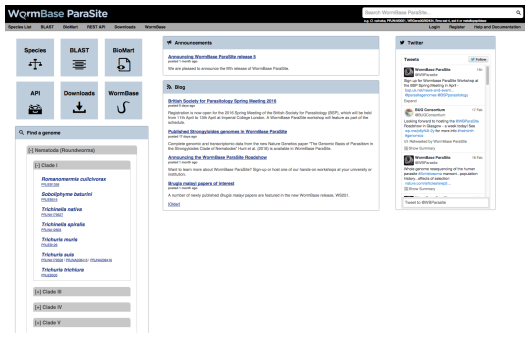
Front page



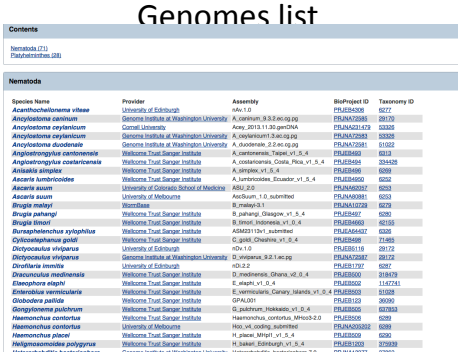
Front page: browse genomes



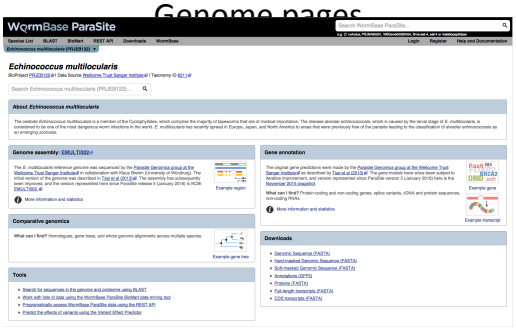
Locating genomes



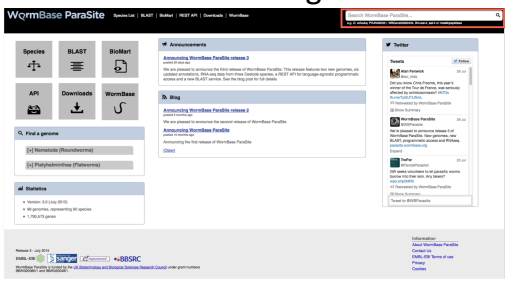
Genomes list



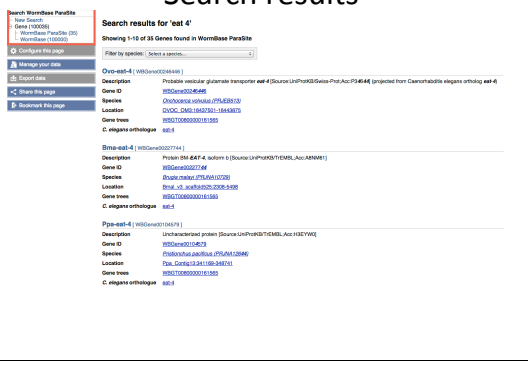
Genome pages



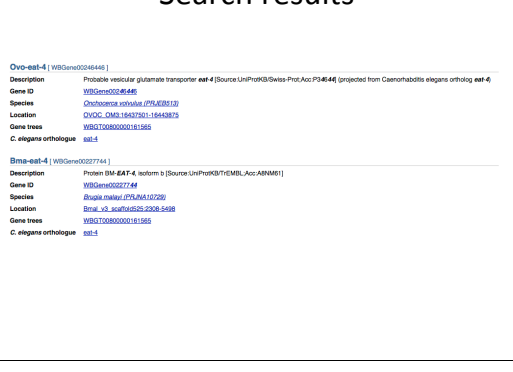
Searching



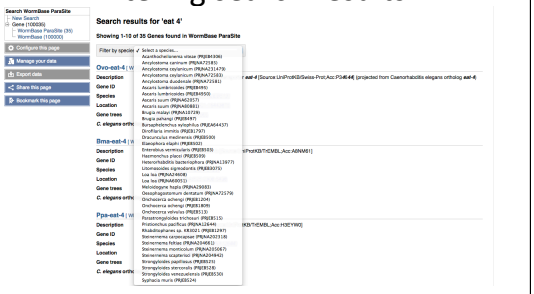
Search results



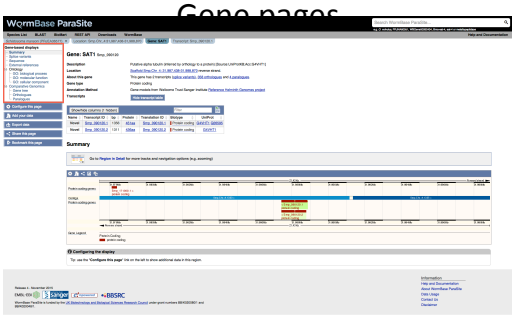
Search results



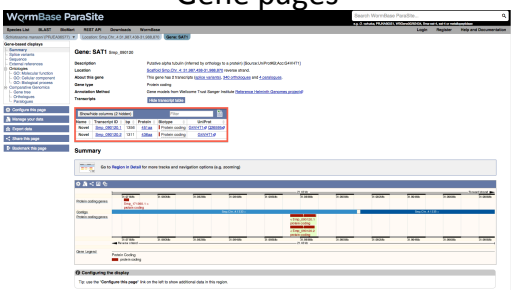
Filtering search results



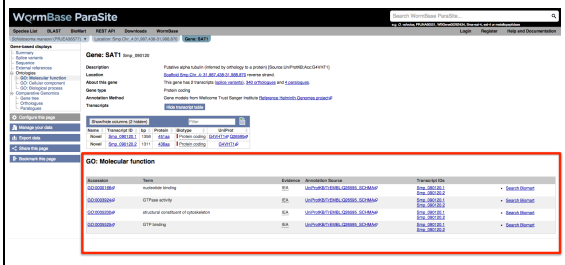
Gene pages



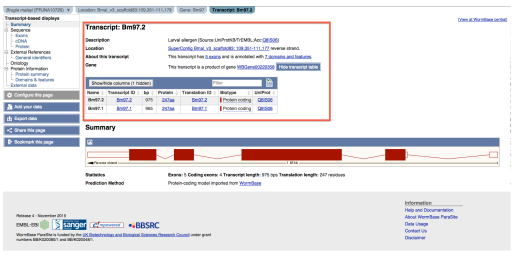
Gene pages



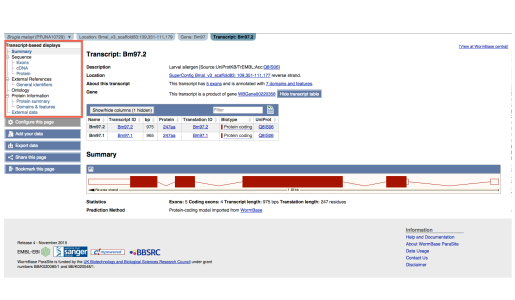
GO terms



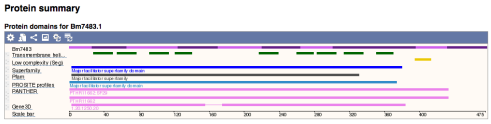
Transcript pages: summmary



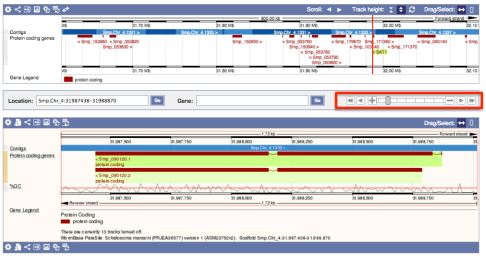
Transcript pages: navigating



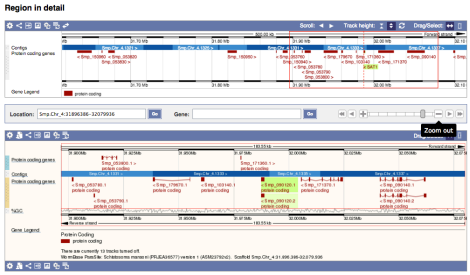
Transcript pages: protein domains



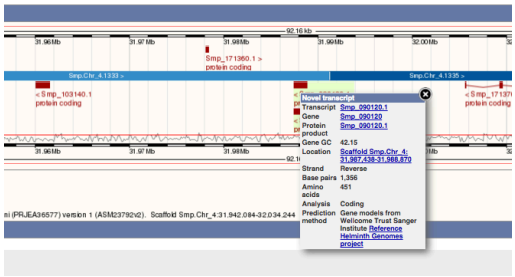
Location view: zooming



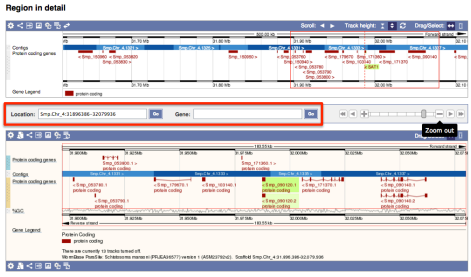
Location view: zooming



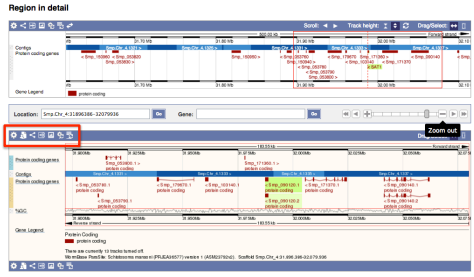
Location view: gene/transcript info



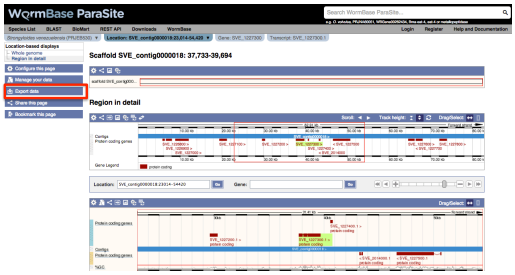
Location view: jump to…



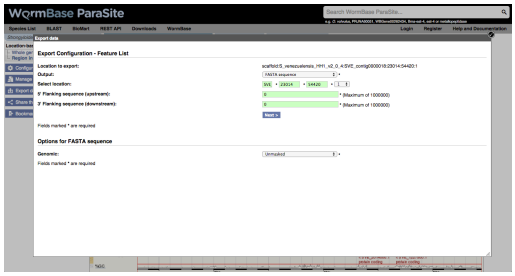
Location view: configure



Location view: export data

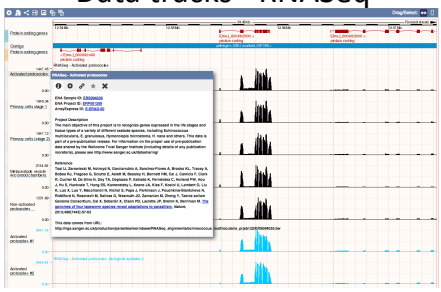


Location view: export data

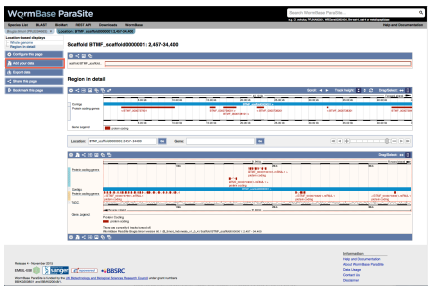


Data tracks - RNASeq

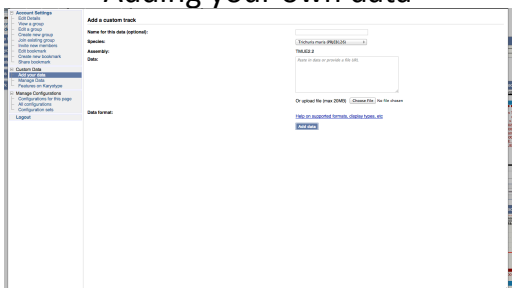
Data tracks - RNASeq



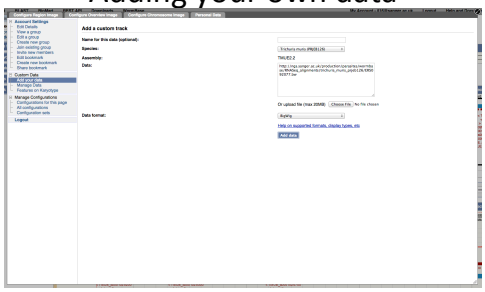
Adding your own data



Adding your own data



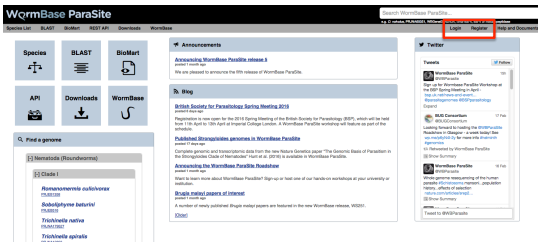
Adding your own data



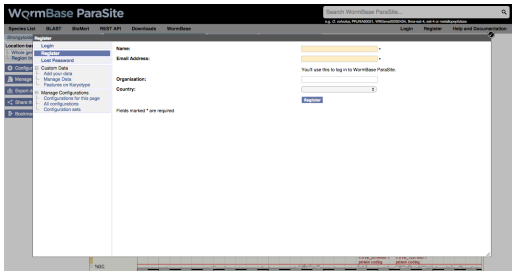
User accounts

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

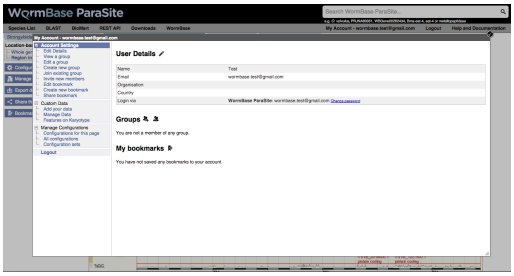
User accounts



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

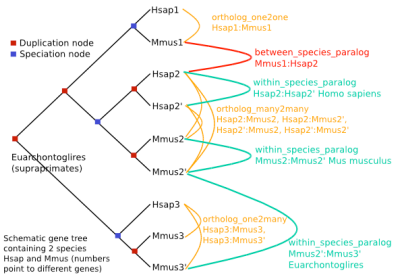
A word of caution…

- 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

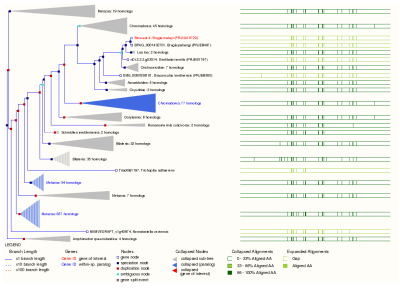
Homology types

- 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

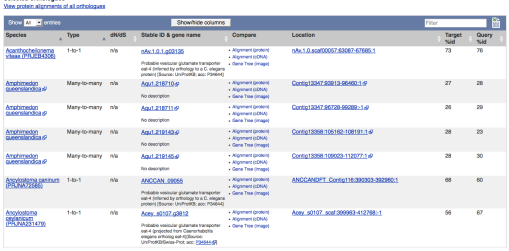
Understanding the gene tree



Visual access to the trees



Tabular access to tree data



Part 3: Sequence Similarity Search using BLAST

What is BLAST?

- BLAST = **B**asic **L**ocal **A**lignment **S**earch **T**ool
- 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

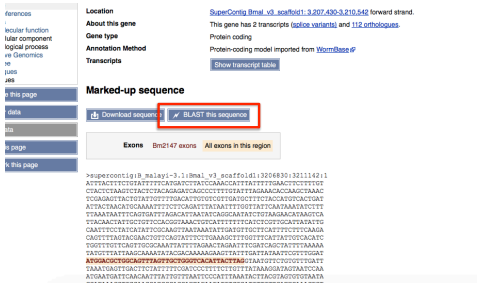
Types of BLAST

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

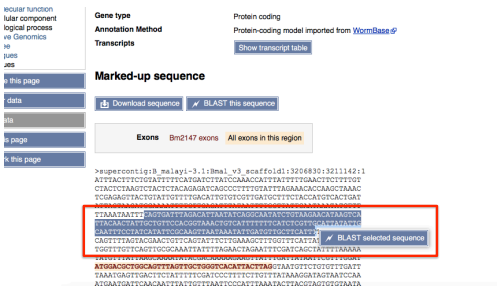
Using the ParaSite BLAST

Defaults to the species you are currently browsing

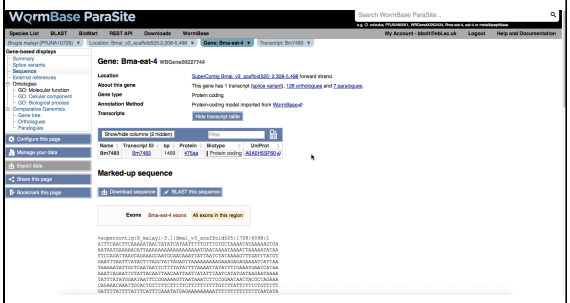
Using the ParaSite BLAST



Using the ParaSite BLAST



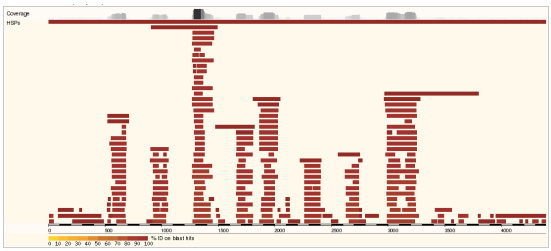
Using the ParaSite BLAST



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

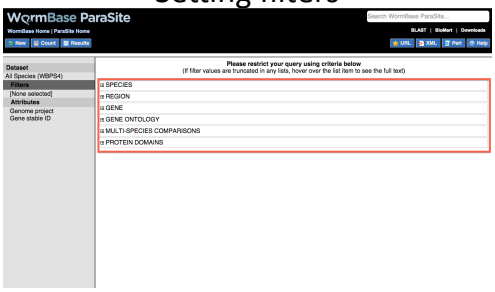
Making sense of the results



Part 4: Data-mining with BioMart

Data-mining with BioMart

Setting filters



- 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.

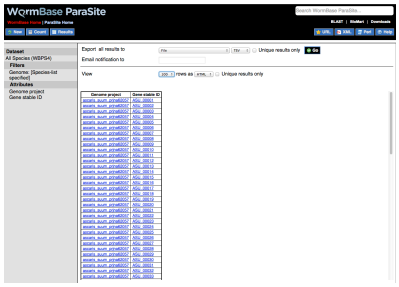
- REGION: Restrict to a particular genomic region.
  - Should only be used where a single genome has been selected, as it is possible that a particular region is present in multiple genomes.
  - If start/end co-ordinates are being specified, a scaffold or chromosome id is always required.
  - Where multiple 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.

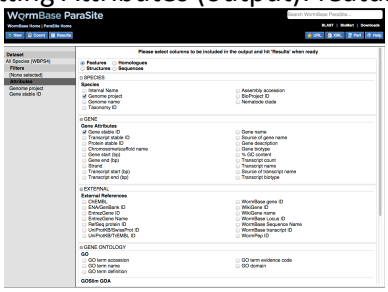
- GENE ONTOLOGY: Restrict by one or more Gene Ontology (GO) terms for functional descriptions.
  - Paste or upload a list of GO IDs or use the autocomplete box to populate the list.
- Alternatively restrict to a particular GO evidence type e.g. Inferred by Electronic Annotation (IEA).
  - Multiple codes 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 in 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.

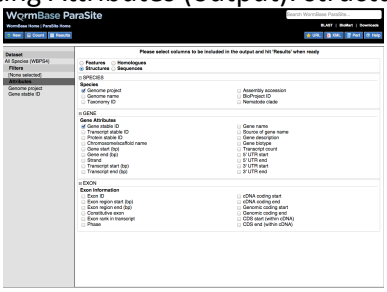
BioMart output



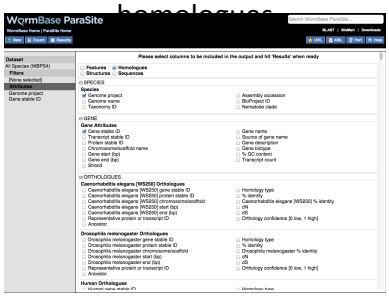
Setting Attributes (output): features



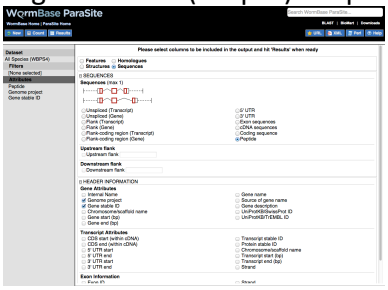
Setting Attributes (output): structures



Setting Attributes (output): homologues

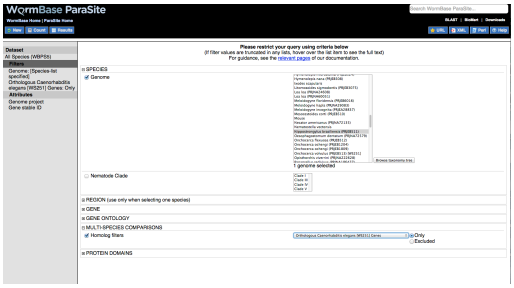


Setting Attributes (output): sequence



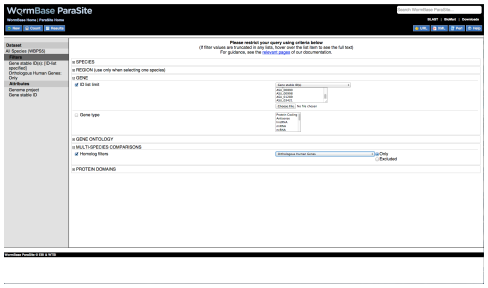
"I'd like to extract all *C. elegans* orthologs for *Nippostrongylus* genes involved in a particular process."

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.



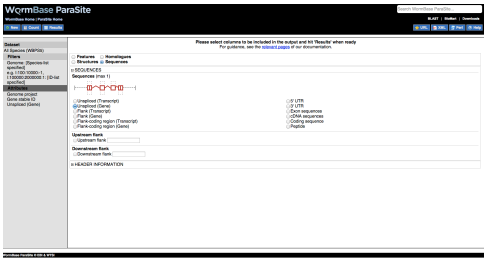
"I have a list of genes from *Ascaris suum* and would like to know which ones have orthologs in humans and mammals and which ones might be nematode-specific."

- 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.