

## WormBase ParaSite Workshop

Glasgow  
24<sup>th</sup> February 2016

## WormBase ParaSite Team



**Bruce Bolt**  
Bioinformatician  
(web and tools)



**Jane Lomax**  
Bioinformatician  
(curation)



**Myriam Shaffe**  
Bioinformatician  
(pipelines)



**Kevin Howe**  
WormBase Team  
Leader



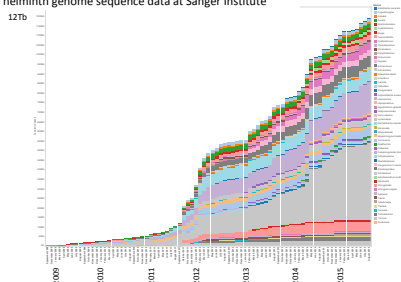
**Paul Kersey**  
PI (at EMBL-EBI)



**Matt Berriman**  
PI (at Sanger Institute)

## An explosion of parasitic worm genomes

Total helminth genome sequence data at Sanger Institute



## 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 completing this form before you leave the workshop. Your feedback will help us to improve future workshops.

OR: Complete Online

1. Did you find the workshop useful? (Please tick one box)

2. Did you find the workshop helpful? (Please tick one box)

3. How useful was the workshop to your colleagues?

4. How useful was each section of the workshop?

	Very Useful	Useful	Not Useful	Not Answered
Introduction to Parasites				
BLAST				
Comparative Genomics				
BioMart				

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

	Very Helpful	Helpful	Not Helpful	Not Answered
Genome Browser				
Transcript Display				
Protein Display				
Comparative Genomics				
Quality of presentations				
Content				
Balance of presentations and content				

6. Do you have any other comments or feedback?

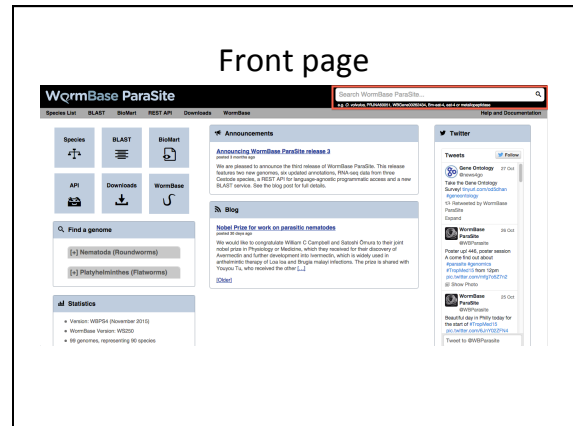
Thank you!

## Part 1: Browsing and searching

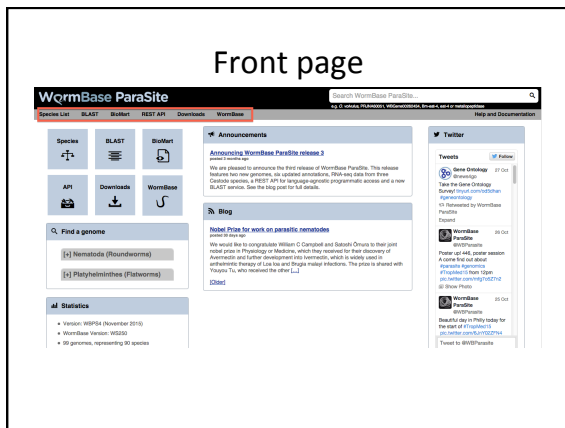
## Part 1: summary

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

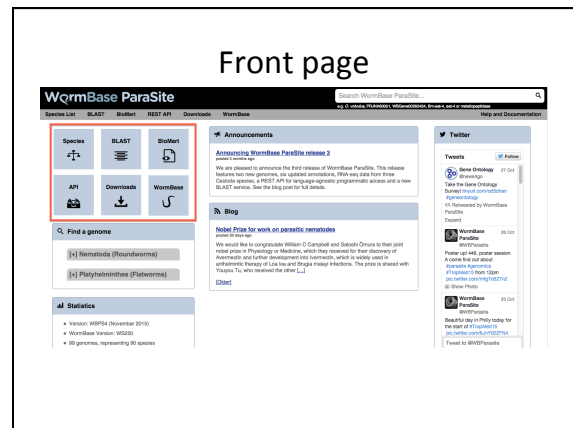
## Front page



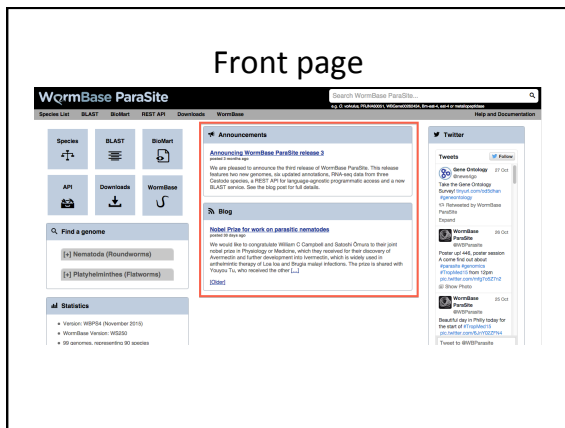
## Front page



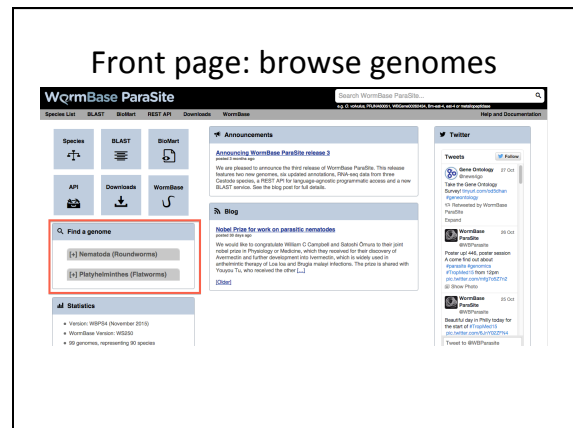
## Front page



## Front page



## Front page: browse genomes



## Locating genomes

## Genomes list

Contents

Nematoda (71)  
Diplobranchia (26)

Species Name	Provider	Assembly	Bioproject ID	Taxonomy ID
<i>Acanthochoyromyxa vishae</i>	University of Edinburgh	rs.1.0	PRJNA5250	6277
<i>Angiostrongylus cantoniensis</i>	Genome Institute at Washington University	A. cantoniensis_3.2 PacBio pg	PRJNA72981	19179
<i>Angiostrongylus ceylanicus</i>	Genome Institute at Washington University	Ang. 2013.11.20 genOMA	PRJNA42123	53302
<i>Angiostrongylus japonicus</i>	Genome Institute at Washington University	A. japonicus_3.2 PacBio pg	PRJNA72981	19179
<i>Angiostrongylus duodenalis</i>	Genome Institute at Washington University	A. duodenalis_3.2 PacBio pg	PRJNA72981	19179
<i>Angiostrongylus cantoniensis</i>	Wellcome Trust Sanger Institute	A. cantoniensis_Pepin_V1_3_4	PRJNA5485	6812
<i>Angiostrongylus costaricensis</i>	Wellcome Trust Sanger Institute	A. costaricensis_Costa_Pico_V1_3_4	PRJNA5484	53423
<i>Aspilota simplex</i>	Wellcome Trust Sanger Institute	A. simplex_V1_3_4	PRJNA5485	6812
<i>Ascaris lumbricoidea</i>	Wellcome Trust Sanger Institute	A. lumbricoidea_Coxsack_V1_3_4	PRJNA5486	6813
<i>Ascaris suum</i>	University of Colorado School of Mines	ASU_2.0	PRJNA00721	5263
<i>Brugia malayi</i>	University of Michigan	Br. malayi_1.0 subverted	PRJNA0081	5033
<i>Brugia pahangi</i>	Wellcome Trust Sanger Institute	B. pahangi_Otago_V1_3_4	PRJNA12729	6029
<i>Brugia imrayi</i>	Wellcome Trust Sanger Institute	B. imrayi_Spokane_V1_3_4	PRJNA5486	6813
<i>Brugia malayi</i>	Wellcome Trust Sanger Institute	AGM331.1b1 subverted	PRJNA5487	6812
<i>Cyathostephanos goldi</i>	Wellcome Trust Sanger Institute	C. goldi_Chenhai_V1_3_4	PRJNA5488	51485
<i>Cyathostephanos ophryostoma</i>	Wellcome Trust Sanger Institute	CMO331.1b1 subverted	PRJNA5487	6812
<i>Cyathostephanos vespertinus</i>	Genome Institute at Washington University	D. vespertinus_3.2 PacBio pg	PRJNA72982	19179
<i>Dicrocoelium viviparum</i>	University of Edinburgh	dv.2.2	PRJNA5118	60173
<i>Dicrocoelium viviparum</i>	Wellcome Trust Sanger Institute	D. viviparum_China_V0_3_4	PRJNA5500	53672
<i>Echinochlamys niphi</i>	Wellcome Trust Sanger Institute	E. niphi_V1_3_4	PRJNA5502	51672
<i>Echinochlamys niphi</i>	Wellcome Trust Sanger Institute	E. niphi_Canary_Islands_V1_3_4	PRJNA5503	51673
<i>Echinochlamys niphi</i>	Wellcome Trust Sanger Institute	OPAL01	PRJNA123	50309
<i>Empoasca fabae</i>	Wellcome Trust Sanger Institute	E. fabae_Paris_1844_V1_3_4	PRJNA5502	51673
<i>Haemonchus contortus</i>	Wellcome Trust Sanger Institute	Haemonchus contortus_MP1005.2.0	PRJNA5508	5369
<i>Haemonchus contortus</i>	Wellcome Trust Sanger Institute	Ha. contortus_Canary_Islands	PRJNA5509	5370
<i>Haemonchus contortus</i>	Wellcome Trust Sanger Institute	H. contortus_MP101_V1_3_4	PRJNA5509	5369
<i>Heliconia pascalis</i>	Wellcome Trust Sanger Institute	H. pascalis_Schlegel_V1_3_4	PRJNA5510	5370
<i>Heliconia pascalis</i>	Wellcome Trust Sanger Institute	H. pascalis_Schlegel_V1_3_4	PRJNA5510	5370

## Genome pages

## Searching

## Search results

## Search results

## Filtering search results

**Search WormBase Parasite**

Search results for 'test 4'

Showing 1-11 of 18 Genes found in WormBase Parasite

Filter by species:  All species

Filter by gene type:  All gene types

Filter by location:  All locations

Filter by date:  All dates

Filter by status:  All statuses

Filter by gene name:  All gene names

Gene ID: [WormBase Parasite: WPM000001](#)

Description: [WormBase Parasite: WPM000001](#)

Location: [WormBase Parasite: WPM000001](#)

Species: [WormBase Parasite: WPM000001](#)

Gene name: [WormBase Parasite: WPM000001](#)

## Gene pages

**WormBase Parasite**

Gene SAT1

Description: [WormBase Parasite: SAT1](#)

Location: [WormBase Parasite: SAT1](#)

Species: [WormBase Parasite: SAT1](#)

Gene name: [WormBase Parasite: SAT1](#)

Summary

GO Molecular function

## Gene pages

**WormBase Parasite**

Gene SAT1

Gene ID: [WormBase Parasite: SAT1](#)

Description: [WormBase Parasite: SAT1](#)

Location: [WormBase Parasite: SAT1](#)

Species: [WormBase Parasite: SAT1](#)

Gene name: [WormBase Parasite: SAT1](#)

## GO terms

**WormBase Parasite**

Gene SAT1

GO Molecular function

Accession	Term	Evidence	Associated Score	Score by file	Search Status
GO:0003674	transcription	EA	UNIPROT:WPM000001:G0000000000	0.000000	Search Status
GO:0003675	transcription factor activity	EA	UNIPROT:WPM000001:G0000000000	0.000000	Search Status
GO:0003676	transcription factor activity, sequence-specific	EA	UNIPROT:WPM000001:G0000000000	0.000000	Search Status
GO:0003677	transcription factor activity, sequence-specific, DNA-binding	EA	UNIPROT:WPM000001:G0000000000	0.000000	Search Status
GO:0003678	transcription factor activity, sequence-specific, RNA-binding	EA	UNIPROT:WPM000001:G0000000000	0.000000	Search Status

## Transcript pages: summary

**WormBase Parasite**

Transcript Bmp7.2

Description: [WormBase Parasite: Bmp7.2](#)

Location: [WormBase Parasite: Bmp7.2](#)

Species: [WormBase Parasite: Bmp7.2](#)

Gene name: [WormBase Parasite: Bmp7.2](#)

Summary

GO Molecular function

## Transcript pages: navigating

**WormBase Parasite**

Transcript Bmp7.2

Transcript ID: [WormBase Parasite: Bmp7.2](#)

Description: [WormBase Parasite: Bmp7.2](#)

Location: [WormBase Parasite: Bmp7.2](#)

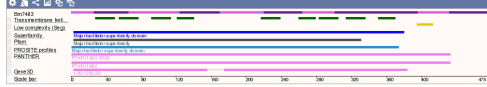
Species: [WormBase Parasite: Bmp7.2](#)

Gene name: [WormBase Parasite: Bmp7.2](#)

## Transcript pages: protein domains

### Protein summary

#### Protein domains for Bm7483.1



Statistics  
 Ave. residue weight: 108.853 g/mol  
 Charge: 4.5  
 Isoelectric point: 7.7253  
 Molecular weight: 52,179.59 g/mol  
 Number of residues: 473 aa

## Location view: zooming

### Region in detail

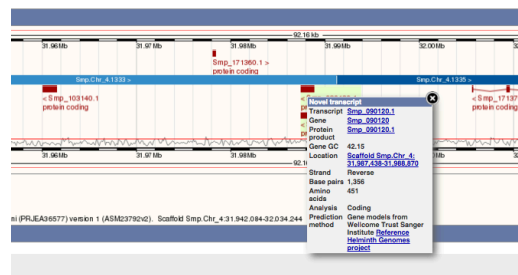


## Location view: zooming

### Region in detail

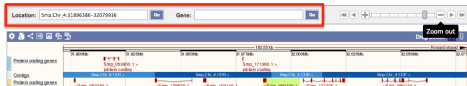


## Location view: gene/transcript info



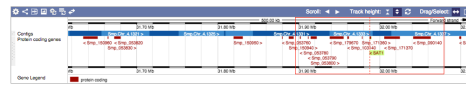
## Location view: jump to...

### Region in detail



## Location view: configure

### Region in detail



### Location view: export data

WormBase ParaSite

Scaffold SVE\_contig000018:37,733-39,694

Region in detail

Export data

### Location view: export data

WormBase ParaSite

Export Configuration - Feature List

Location to export: scaffold\_svecontig000018\_37733-39694

Search features: [SVE] + [SVE] + [SVE]

FASTA options: [FASTA] [FASTA]

Options for FASTA sequence

Export

### Data tracks - RNASeq

WormBase ParaSite

Region in detail

RNASeq

### Data tracks - RNASeq

WormBase ParaSite

Region in detail

RNASeq

### Adding your own data

WormBase ParaSite

Adding your own data

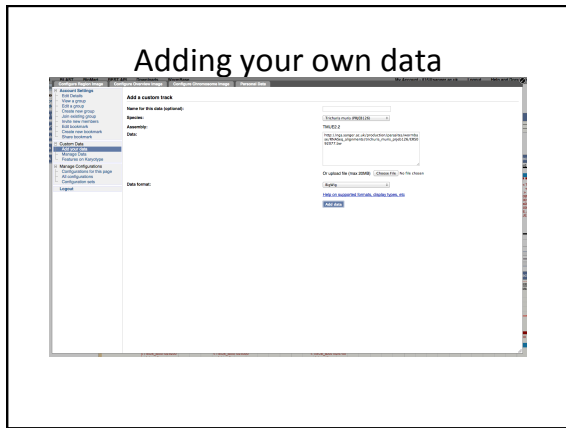
Add your own data

### Adding your own data

WormBase ParaSite

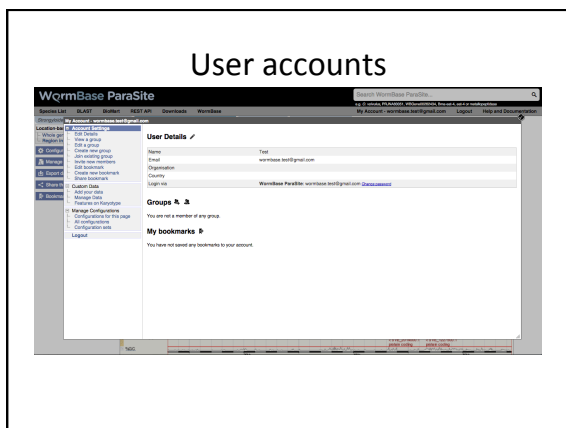
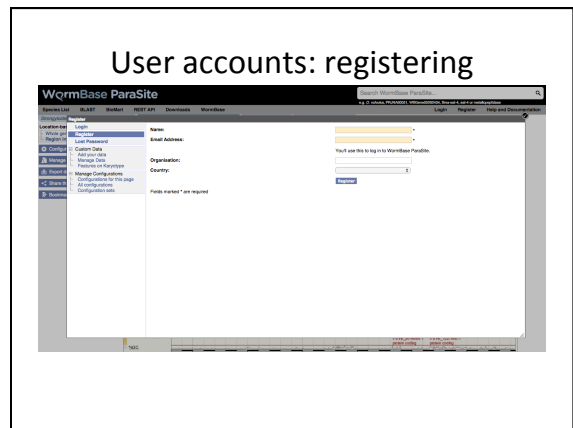
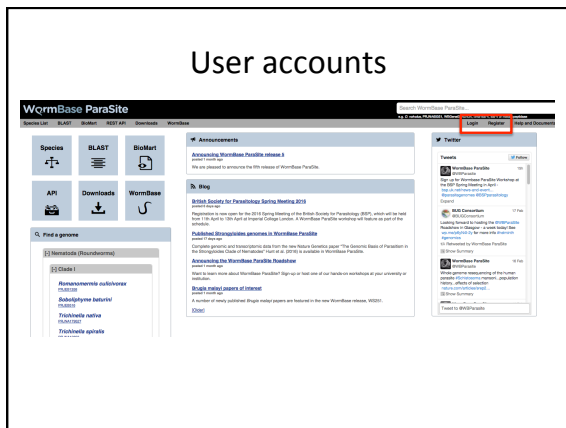
Adding your own data

Add your own data



### User accounts

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



### 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

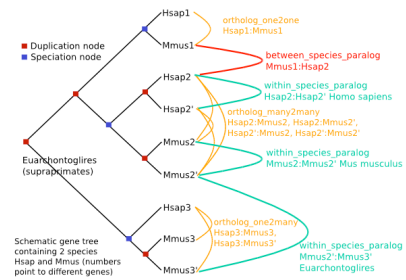
### A word of caution...

- 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

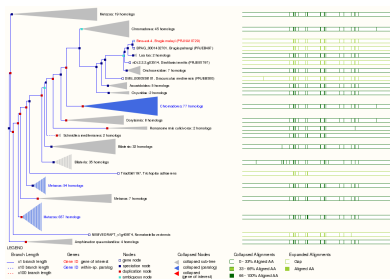
### Homology types

- 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 details: [alignments](#) of all orthologues

Species	Type	dN/dS	Statistics	Gene name	Comments	Location	Target	Query
<i>Ascaris suum</i> ASCA011200	1-to-1	0.06	0.06:0.12:0.12:0.12	ASCA011200	Agarose polymerase (Agarose polymerase) Agarose polymerase (Agarose polymerase) Agarose polymerase (Agarose polymerase)	SLC1A5:ASCA011200:1-1000	73	76
<i>Ascaris suum</i> ASCA011200	Many-to-many	0.06	0.06:0.12:0.12:0.12	Agp_131210	Agarose polymerase (Agarose polymerase) Agarose polymerase (Agarose polymerase) Agarose polymerase (Agarose polymerase)	Contig1247:ASCA011200:1-1000	27	28
<i>Ascaris suum</i> ASCA011200	Many-to-many	0.06	0.06:0.12:0.12:0.12	Agp_131211	Agarose polymerase (Agarose polymerase) Agarose polymerase (Agarose polymerase) Agarose polymerase (Agarose polymerase)	Contig1247:ASCA011200:1-1000	26	29
<i>Ascaris suum</i> ASCA011200	Many-to-many	0.06	0.06:0.12:0.12:0.12	Agp_131212	Agarose polymerase (Agarose polymerase) Agarose polymerase (Agarose polymerase) Agarose polymerase (Agarose polymerase)	Contig1248:ASCA011200:1-1000	28	29
<i>Ascaris suum</i> ASCA011200	Many-to-many	0.06	0.06:0.12:0.12:0.12	Agp_131213	Agarose polymerase (Agarose polymerase) Agarose polymerase (Agarose polymerase) Agarose polymerase (Agarose polymerase)	Contig1248:ASCA011200:1-1000	28	30
<i>Ascaris suum</i> ASCA011200	1-to-1	0.06	0.06:0.12:0.12:0.12	ASCA011200	Agarose polymerase (Agarose polymerase) Agarose polymerase (Agarose polymerase) Agarose polymerase (Agarose polymerase)	ASCA011200:Contig1248:ASCA011200:1-1000	68	60
<i>Ascaris suum</i> ASCA011200	1-to-1	0.06	0.06:0.12:0.12:0.12	Agp_131214	Agarose polymerase (Agarose polymerase) Agarose polymerase (Agarose polymerase) Agarose polymerase (Agarose polymerase)	Agp_131214:ASCA011200:1-1000	66	67

## 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

Equal function  
 Similar component  
 Logical process  
 Molecular function  
 Part  
 Pathway  
 Similarity

**Gene type**  
 Annotation Method  
 Transcripts

**Protein coding**  
 Protein-coding model imported from WormBase

Marked-up sequence  
 Download sequence  
 BLAST this sequence

Exons: Bm12147 exons. All exons in this region.

```

>suprec001318_malayi-3.118m1_v3_scaf101|3204830|321142|1
ATTACTTCGGATTTTCAGATGCTTACGAGACATTAATTTCAGCCTGTTT
CTACTCTAGCTCTACAGAGATCAACGCTTTGTTATTAAGAAACAGCCTAAC
TCGAGATCTACTGATTTGTCATCTGCTTTCTGATGCTTCTGACATGTC
TTAAATATTTCTGGATTTAGCATTAATCTGCGCATATCTGATGATATGCT
GAAATGCTATATCTGGTCCGATGATATGATATATTTTCAGTCTGATGAT
CACTTATAGACAGATTCAGATCTTCGACAGCTTCAGTCTGATGAT
TGGTTTCTGCTTCTGCAATTTTTCAGACATGATTTCTGCTGATTTTAAAG
ATGACAGCTTCAGATTTGATGCTTCAGTCTGATGATTTCTGCTGATTT
TAAATGATTCAGATTTGATGCTTCAGTCTGATGATTTCTGCTGATTT
AAGATATTCAGATTTGATGCTTCAGTCTGATGATTTCTGCTGATTT
    
```

### Using the ParaSite BLAST

WormBase ParaSite

Species List  
 BLAST  
 BLAST API  
 Downloads  
 WormBase

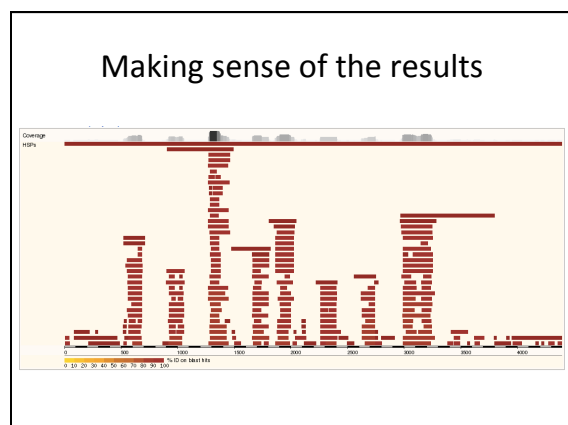
Gene: Bm12147-4  
 Location: Bm12147-4  
 Gene type: Protein coding  
 Annotation Method: Protein-coding model imported from WormBase

Marked-up sequence  
 Download sequence  
 BLAST this sequence

Exons: Bm12147-4 exons. All exons in this region.

### 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

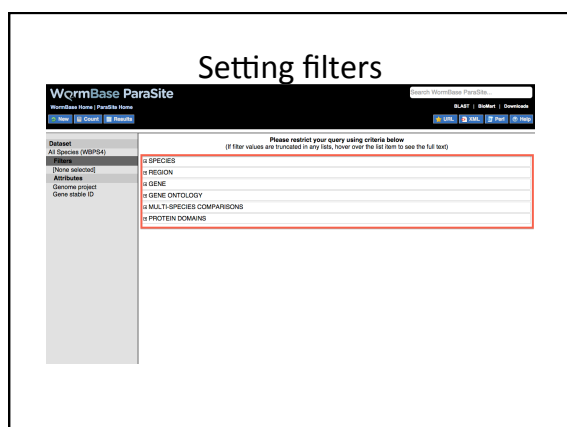
WormBase ParaSite

Home  
 BLAST  
 BLAST API  
 Downloads

Dataset: All Species (WormBase)

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

- SPECIES
- REGION
- GENE
- 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.

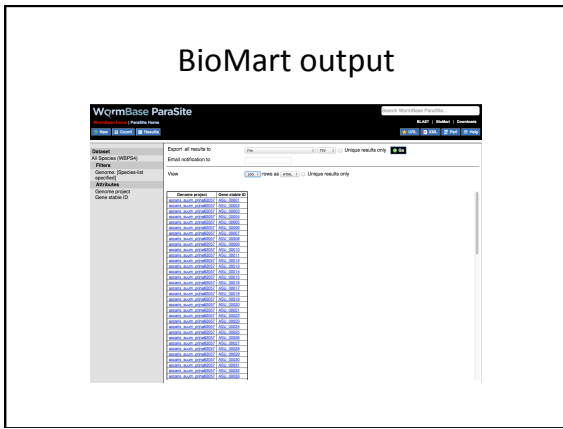
- **REGION:** Restrict to a particular genomic region.
  - Should only be used where a single genome has been selected, as it is possible that a particular region is present in multiple genomes.
  - If start/end co-ordinates are being specified, a scaffold or chromosome id is always required.
  - Where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

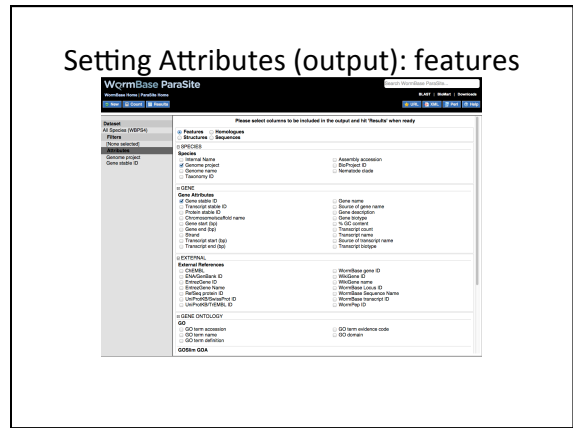
- **GENE ONTOLOGY:** Restrict by one or more Gene Ontology (GO) terms for functional descriptions.
  - Paste or upload a list of GO IDs or use the autocomplete box to populate the list.
- Alternatively restrict to a particular GO evidence type e.g. Inferred by Electronic Annotation (IEA).
  - Multiple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

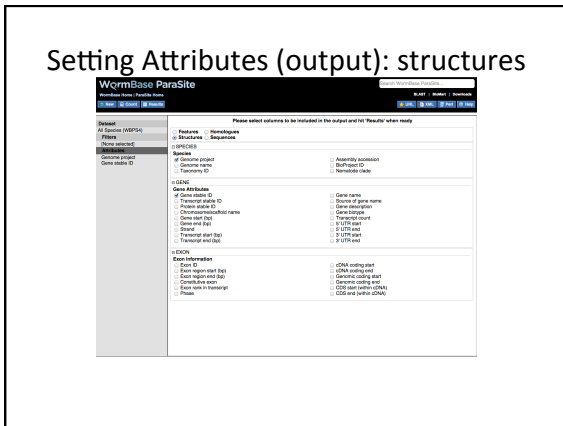
## BioMart output



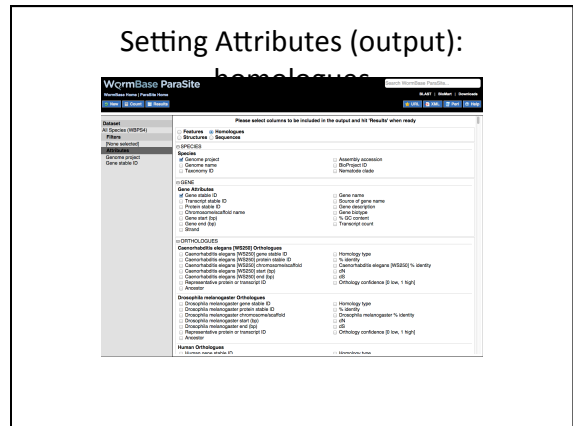
## Setting Attributes (output): features



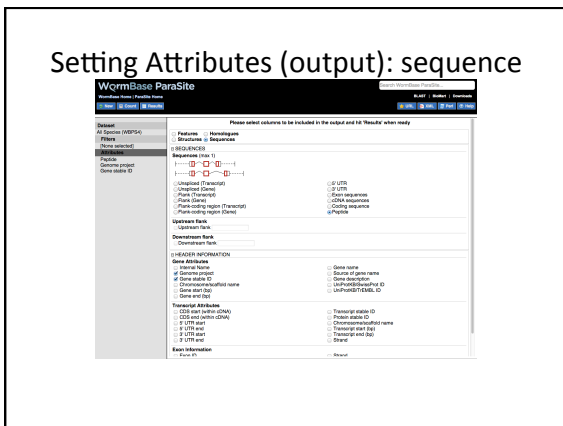
## Setting Attributes (output): structures



## Setting Attributes (output): homologues

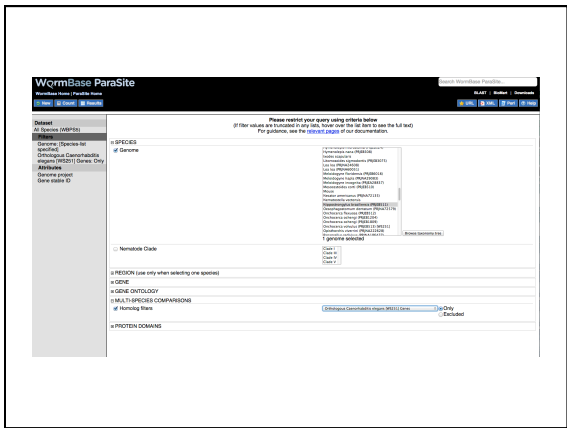


## Setting Attributes (output): sequence



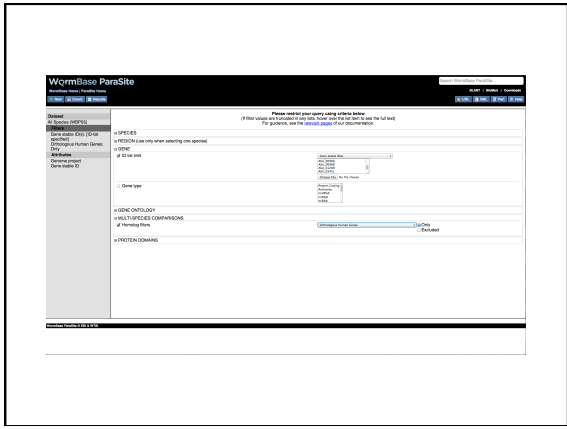
"I'd like to extract all *C. elegans* orthologs for *Nippostrongylus* genes involved in a particular process."

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.



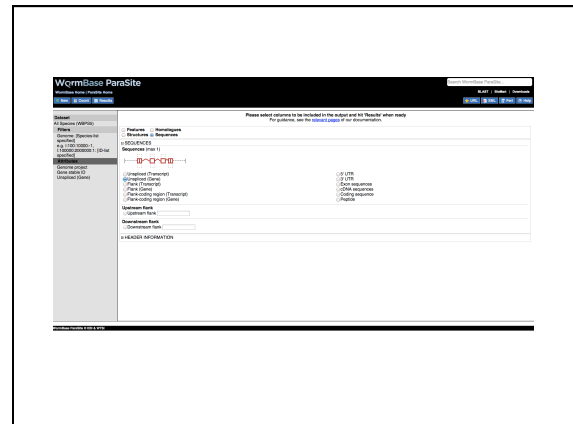
“I have a list of genes from *Ascaris suum* and would like to know which ones have orthologs in humans and mammals and which ones might be nematode-specific.”

- 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 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.

