


# WormBase ParaSite Workshop

Edinburgh  
9<sup>th</sup> March 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)

## parasite.wormbase.org

- 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

WormBase ParaSite

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


## 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

## The Data

- All genomes are shown “as supplied” by the submitter (except WormBase “core” genomes)
- Varying levels of coverage and quality
- Details of assembly and annotation displayed on information page
- “Core” parasitic genomes: *Brugia malayi*, *Onchocerca volvulus*, *Pristionchus pacificus* and *Strongyloides ratti*
- Receive more care and attention
- Community driven manual curation



## Your Data

- Publicly available transcriptomic data annotated and displayed on browser
- Website supports ad-hoc visualisation of your own data (e.g. RNA-Seq alignments, variations)
- We welcome submissions of your own data to display on genome browser – allow readers of your papers to easily visualise your data
- Please contact us (link at bottom of website) to discuss requirements

## 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](http://parasite.wormbase.org)

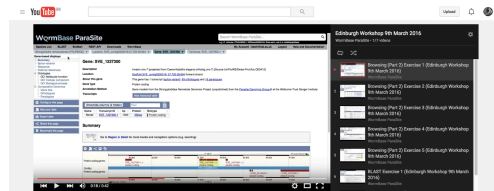


## This afternoon's agenda...

- 13:00 – 13:15  
Introduction to WormBase ParaSite
- 13:15 – 13:45  
Using the website (Part 1)
- 13:45 – 14:15  
Using the website (Part 2)
- 14:15 – 15:00  
Sequence Search with BLAST
- 15:00 – 15:30  
Coffee Break
- 15:30 – 16:30  
Data Mining with BioMart
- 16:30 – 16:45  
Bulk downloads and programmatic access

## After this workshop...

- Please contact us with any questions (contact form link at bottom of every page)
- Solutions to exercises on YouTube: [parasite.wormbase.org/workshop](http://parasite.wormbase.org/workshop)



## Workshop Feedback

- 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 feedback form at the end of the workshop. Your comments will help us to improve future workshops.

1. Did you find WormBase ParaSite useful for your research?  
 Omit "Sometimes" row

2. Did you find WormBase ParaSite useful when using the website?  
 Omit "Sometimes" row

3. How useful was the workshop to your colleagues?  
 Omit "Not at all"

4. How useful was each section of the workshop?

	Very useful	Useful	Not useful	Not completed, if possible
Overview of the website				
BLAST				
Comparative Genomics				
BioMart				

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

	Very helpful	Helpful	Not helpful	Not completed, if possible
Overview				
Sequence Search				
BLAST				
Comparative Genomics				
BioMart				
Quality of presentation				
Content				
Balance of presentation and activities				

6. Do you have any other comments or feedback?

Thank you!

## Part 1: Using the website

## Part 1: Summary

1. Front page
2. Locating genomes
3. Navigating genes, transcripts and scaffolds
4. RNASeq tracks
5. Adding your own data

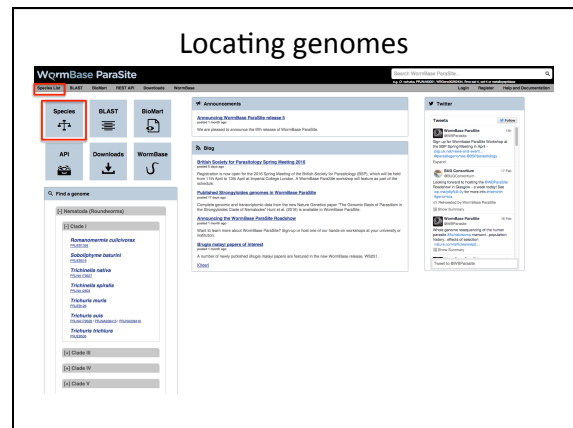
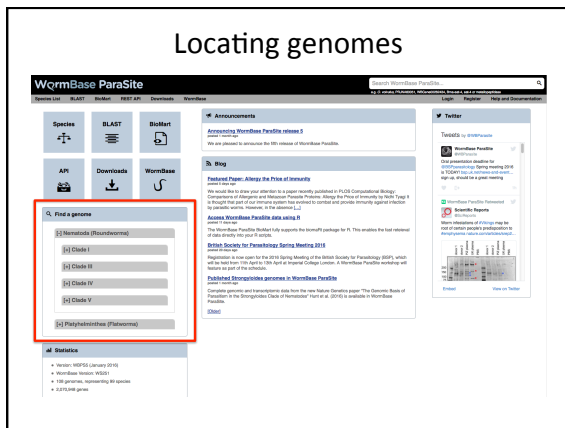
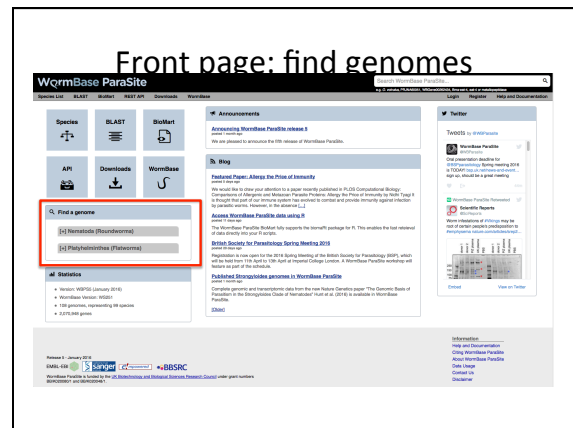
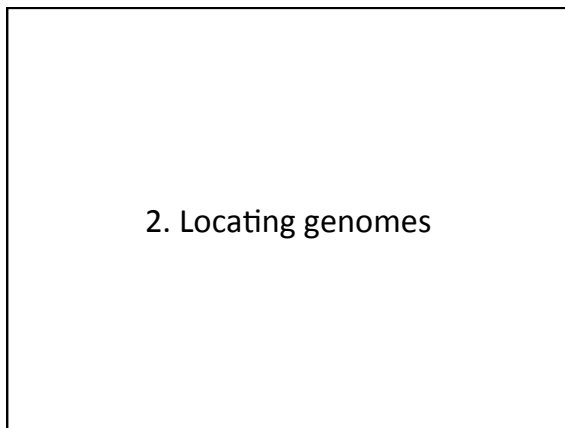
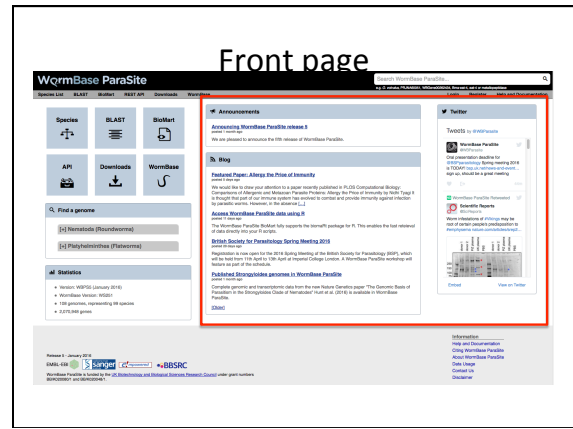
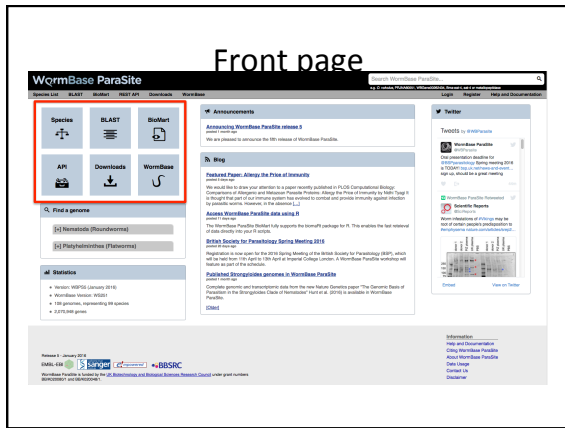
## 1. Front page

The screenshot shows the WormBase ParaSite front page. At the top, there is a search bar with the text "Search WormBase ParaSite" and a red highlight around it. Below the search bar is a navigation menu with links for Species, BLAST, BioMart, Downloads, and WormBase. The main content area is divided into several sections: "Announcements" with a link to "Advancing WormBase ParaSite release 3", "Blog" with a link to "Published Paper: Always Be First at Immunity", "Find a genome" with search filters for Nematodes and Platyhelminths, and "Statistics" showing metrics for WormBase Parasites (2016), WormBase Parasites (2015), and 138 genomes representing 88 species. The footer includes the date "Release 3 - January 2016" and logos for the Wellcome Trust, Sanger Institute, and BBSRC.

This screenshot is identical to the one on the left, showing the WormBase ParaSite front page with the search bar highlighted in red. It displays the same navigation menu, announcements, blog, search filters, and statistics.

This screenshot is identical to the previous ones, showing the WormBase ParaSite front page with the search bar highlighted in red. It displays the same navigation menu, announcements, blog, search filters, and statistics.

This screenshot is identical to the previous ones, showing the WormBase ParaSite front page with the search bar highlighted in red. It displays the same navigation menu, announcements, blog, search filters, and statistics.



### Genomes list

Species Name	Provider	Assembly	RefSeq ID	Taxonomy ID
<i>Acanthocheilium virens</i>	University of Colorado	WU-1	FS:WU1	5277
<i>Ancylostoma caninum</i>	Genome Institute at Washington University	A.caninum_3.2.30 (sp.9)	FS:WU1096	5373
<i>Ancylostoma ceylanicum</i>	Genome Institute at Washington University	Acny_2013.11.20 (sp.9)	FS:WU1219	5359
<i>Ancylostoma pretense</i>	Genome Institute at Washington University	A.pretense_3.0 (sp.9)	FS:WU1082	5358
<i>Ancylostoma duodenale</i>	Genome Institute at Washington University	A.duodenale_3.0 (sp.9)	FS:WU1081	5352
<i>Angiostrongylus cantonensis</i>	Wellcome Trust Sanger Institute	A.cantonensis_Tsukaguchi_1.5.4	FS:WU1461	5123
<i>Angiostrongylus costaricensis</i>	Wellcome Trust Sanger Institute	A.costaricensis_Costa_Rica_1.3.4	FS:WU1461	5145
<i>Anisakis simplex</i>	Wellcome Trust Sanger Institute	A.simplex_1.3.4	FS:WU1461	5126
<i>Ascaris lumbricoide</i>	Wellcome Trust Sanger Institute	A.lumbricoide_Ecuador_1.1.5.4	FS:WU1461	5125
<i>Ascaris suum</i>	University of Colorado School of Medicine	ASU-2.0	FS:WU1461	5125
<i>Ascaris suum</i>	University of Colorado School of Medicine	Ascaris_su_1.5 (submitted)	FS:WU1461	5125
<i>Brugia malayi</i>	Wellcome Trust Sanger Institute	B.malayi-3.1	FS:WU1461	5129
<i>Brugia pahangi</i>	Wellcome Trust Sanger Institute	B.pahangi_Thailand_1.1.5.4	FS:WU1461	5129
<i>Brugia timori</i>	Wellcome Trust Sanger Institute	B.timori_Indonesia_1.1.5.4	FS:WU1461	5129
<i>Brugia malayi</i>	Wellcome Trust Sanger Institute	B.malayi_Thailand_1.1.5.4	FS:WU1461	5129
<i>Cyathostomum pygmaeum</i>	Wellcome Trust Sanger Institute	C.pygmaeum_1.1.5.4	FS:WU1461	5129
<i>Cyathostomum sp.</i>	Wellcome Trust Sanger Institute	C.sp._China_1.1.5.4	FS:WU1461	5129
<i>Cyathostomum sp.</i>	Wellcome Trust Sanger Institute	C.sp._China_1.1.5.4	FS:WU1461	5129
<i>Cyathostomum sp.</i>	Wellcome Trust Sanger Institute	C.sp._China_1.1.5.4	FS:WU1461	5129
<i>Dirofilaria immitis</i>	University of Colorado	DI-2.2	FS:WU1461	5129
<i>Dracunculus medintensis</i>	Wellcome Trust Sanger Institute	D.medintensis_China_1.1.5.4	FS:WU1461	5129
<i>Elaeophora alpestris</i>	Wellcome Trust Sanger Institute	E.alpestris_1.1.5.4	FS:WU1461	5129
<i>Elaeophora leucophaea</i>	Wellcome Trust Sanger Institute	E.leucophaea_Canary_Islands_1.1.5.4	FS:WU1461	5129
<i>Gyrodontia papillosa</i>	Wellcome Trust Sanger Institute	GPAL01	FS:WU1461	5129
<i>Gongylonema pinnatum</i>	Wellcome Trust Sanger Institute	G.pinnatum_Hawaii_1.1.5.4	FS:WU1461	5129
<i>Haemonchus contortus</i>	Wellcome Trust Sanger Institute	Haemonchus_contortus_MP003.2.0	FS:WU1461	5129
<i>Haemonchus contortus</i>	Wellcome Trust Sanger Institute	H.contortus_1.1.5.4	FS:WU1461	5129
<i>Haemonchus placei</i>	Wellcome Trust Sanger Institute	H.placei_Mexico_1.1.5.4	FS:WU1461	5129
<i>Heligmosomoides polygyrus</i>	Wellcome Trust Sanger Institute	H.polygyrus_Canary_Islands_1.1.5.4	FS:WU1461	5129

### Genome pages

**WormBase ParaSite** Home | About | Help | Contact Us

**Echinococcus multilocularis**  
 WormBase ID: [Echinococcus multilocularis \(PM262122\)](#)

**Gene overview**

This E. multilocularis gene is located on the [Chromosome 1](#) of the genome. The gene structure is shown below. The gene structure is shown below. The gene structure is shown below.

**Gene structure**

The gene structure is shown below. The gene structure is shown below. The gene structure is shown below.

**Downloads**

- GenBank (FASTA)
- EMBL (FASTA)
- GenBank (FASTA)
- EMBL (FASTA)
- GenBank (FASTA)
- EMBL (FASTA)

## 3. Navigating genes, transcripts and scaffolds

### Gene pages

**WormBase ParaSite** Home | About | Help | Contact Us

**Gene: SAT1**  
 WormBase ID: [SAT1 \(PM262122\)](#)

**Summary**

This gene is located on the [Chromosome 1](#) of the genome. The gene structure is shown below. The gene structure is shown below. The gene structure is shown below.

**Gene structure**

The gene structure is shown below. The gene structure is shown below. The gene structure is shown below.

**Downloads**

- GenBank (FASTA)
- EMBL (FASTA)
- GenBank (FASTA)
- EMBL (FASTA)
- GenBank (FASTA)
- EMBL (FASTA)

### Gene pages: exons

**WormBase ParaSite** Home | About | Help | Contact Us

**Gene: SAT1**  
 WormBase ID: [SAT1 \(PM262122\)](#)

**Gene overview**

This gene is located on the [Chromosome 1](#) of the genome. The gene structure is shown below. The gene structure is shown below. The gene structure is shown below.

**Gene structure**

The gene structure is shown below. The gene structure is shown below. The gene structure is shown below.

**Downloads**

- GenBank (FASTA)
- EMBL (FASTA)
- GenBank (FASTA)
- EMBL (FASTA)
- GenBank (FASTA)
- EMBL (FASTA)

### Gene pages: exons

**WormBase ParaSite** Home | About | Help | Contact Us

**Gene: SAT1**  
 WormBase ID: [SAT1 \(PM262122\)](#)

**Gene overview**

This gene is located on the [Chromosome 1](#) of the genome. The gene structure is shown below. The gene structure is shown below. The gene structure is shown below.

**Gene structure**

The gene structure is shown below. The gene structure is shown below. The gene structure is shown below.

**Downloads**

- GenBank (FASTA)
- EMBL (FASTA)
- GenBank (FASTA)
- EMBL (FASTA)
- GenBank (FASTA)
- EMBL (FASTA)

## GO terms

**WormBase ParaSite**

Gene: **SAT1** *smg-100.1*

**GO: Molecular function**

Accession	Term	Evidence	Annotation Source	Transcript ID(s)	Search
GO:0005387	membrane binding	IS	UniProt/Swiss-Prot/EMBL/GenBank/RefSeq	smg-100.1	Search
GO:0005207	CTP binding	IS	UniProt/Swiss-Prot/EMBL/GenBank/RefSeq	smg-100.1	Search
GO:0005208	ATPase activity	IS	UniProt/Swiss-Prot/EMBL/GenBank/RefSeq	smg-100.1	Search
GO:0005209	GTP binding	IS	UniProt/Swiss-Prot/EMBL/GenBank/RefSeq	smg-100.1	Search

## Transcript pages: summary

**Transcript: Bm97.2**

**Description:** Larval allergen protein (Larval Allergen Protein, LAP) (Bm97.2) (Bm97.2) (Bm97.2)

**Location:** [Chromosome: Bm97.2](#) (1,308,301-1,311,177) (view in genome)

**Gene:** [Bm97.2](#) (1,308,301-1,311,177) (view in genome)

**Statistics:**

Transcript ID	Seq. Protein	Transcript ID	Seq. Protein
Bm97.2	976	Bm97.2	170,559
Bm97.2	486	Bm97.2	170,559

**Statistics:** Exons: 1, Coding exons: 1, Transcript length: 171 bp, Translated length: 247 residues. Protein coding model imported from [Bm97.2](#)

## Transcript pages: navigating

**Transcript: Bm97.2**

**Description:** Larval allergen protein (Larval Allergen Protein, LAP) (Bm97.2) (Bm97.2) (Bm97.2)

**Location:** [Chromosome: Bm97.2](#) (1,308,301-1,311,177) (view in genome)

**Gene:** [Bm97.2](#) (1,308,301-1,311,177) (view in genome)

**Statistics:**

Transcript ID	Seq. Protein	Transcript ID	Seq. Protein
Bm97.2	976	Bm97.2	170,559
Bm97.2	486	Bm97.2	170,559

**Statistics:** Exons: 1, Coding exons: 1, Transcript length: 171 bp, Translated length: 247 residues. Protein coding model imported from [Bm97.2](#)

## Transcript pages: protein domains

**Protein summary**

**Protein domains for Bm97.2**

**Statistics:**

Stat	Value
Avg. residue weight:	109.833 g/mol
Change:	4.5
Isoelectric point:	7.7259
Molecular weight:	52,170.59 g/mol
Number of residues:	475 aa

## Navigating: tabs

**WormBase ParaSite**

Transcript: **Bm97.2**

**Description:** Larval allergen protein (Larval Allergen Protein, LAP) (Bm97.2) (Bm97.2) (Bm97.2)

**Location:** [Chromosome: Bm97.2](#) (1,308,301-1,311,177) (view in genome)

**Gene:** [Bm97.2](#) (1,308,301-1,311,177) (view in genome)

**Statistics:**

Transcript ID	Seq. Protein	Transcript ID	Seq. Protein
Bm97.2	976	Bm97.2	170,559
Bm97.2	486	Bm97.2	170,559

**Statistics:** Exons: 1, Coding exons: 1, Transcript length: 171 bp, Translated length: 247 residues. Protein coding model imported from [Bm97.2](#)

## Location view: zooming

**SuperContig Bm97\_v3.acefa0082: 74,678-76,964**

**Region in detail**

**Location:** [Chromosome: Bm97.2](#) (1,308,301-1,311,177) (view in genome)

**Gene:** [Bm97.2](#) (1,308,301-1,311,177) (view in genome)

**Statistics:**

Stat	Value
Avg. residue weight:	109.833 g/mol
Change:	4.5
Isoelectric point:	7.7259
Molecular weight:	52,170.59 g/mol
Number of residues:	475 aa

### Location view: gene/transcript info

This screenshot shows the 'Region in detail' view in WormBase ParaSite. The top section displays the 'SuperContig Bmal\_v3\_scaffold8: 109,291-111,179'. Below this, there are several tracks including 'Gene' and 'Transcript' information. The 'Gene' track shows the gene structure with exons and introns. The 'Transcript' track shows the transcript structure with exons and introns. The 'Region in detail' section provides a summary of the region, including the coordinates and the gene/transcript names.

### Location view: jump to...

This screenshot shows the 'Region in detail' view in WormBase ParaSite. The 'Location' field is highlighted with a red box, and a 'Jump to...' button is visible next to it. The 'Region in detail' section provides a summary of the region, including the coordinates and the gene/transcript names.

### Location view: configure

This screenshot shows the 'Region in detail' view in WormBase ParaSite. The 'Configure' button is highlighted with a red box. The 'Region in detail' section provides a summary of the region, including the coordinates and the gene/transcript names.

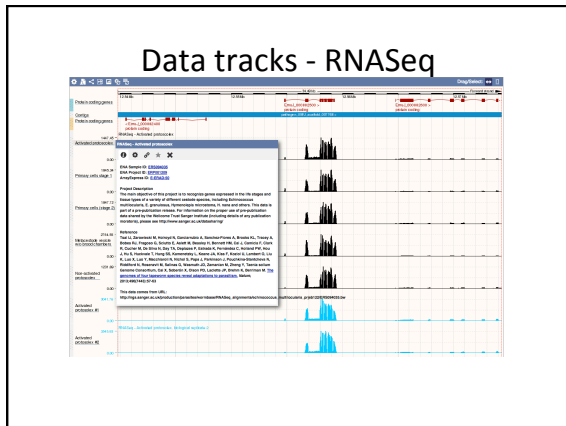
### Location view: export data

This screenshot shows the 'Region in detail' view in WormBase ParaSite. The 'Export data' button is highlighted with a red box. The 'Region in detail' section provides a summary of the region, including the coordinates and the gene/transcript names.

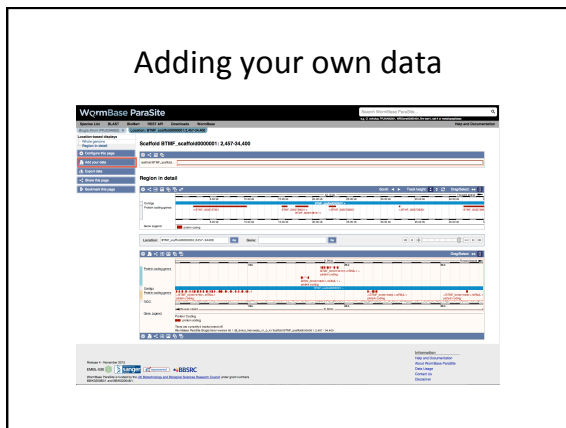
4. RNASeq tracks

### Data tracks - RNASeq

This screenshot shows the 'Data tracks - RNASeq' view in WormBase ParaSite. The tracks display RNASeq data for the region SuperContig Bmal\_v3\_scaffold8: 110,710-110,710. The tracks include 'RNASeq' and 'RNASeq' data, showing signal intensity across the region.



## 5. Adding your own data



### Adding your own data

The figure shows the 'Add a custom track' form in the WormBase ParaSite website. The form includes fields for 'Name for the data track', 'Species', 'Assembly', 'Date', and a description. A 'Data format' dropdown is also visible. The form is titled 'Add a custom track' and has a 'Log out' button at the bottom.

### Adding your own data

The figure shows the 'Add a custom track' form in the WormBase ParaSite website, with the 'Data format' dropdown menu open. The dropdown menu lists several options: 'BigWig', 'BigBed', 'BigBed (with tracks)', 'BigWig (with tracks)', 'BigWig (with tracks and tracks)', 'BigWig (with tracks and tracks and tracks)', and 'BigWig (with tracks and tracks and tracks and tracks)'. The form is titled 'Add a custom track' and has a 'Log out' button at the bottom.

## Part 1b: Browsing the website

- Searching the website
- Comparative genomics
- User accounts



## Searching

## Search results

## Filtering search results

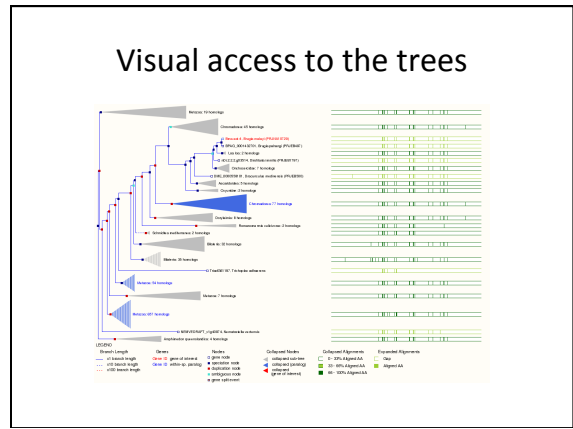
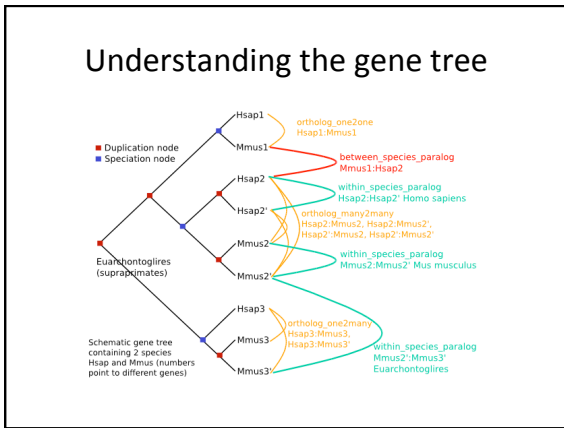
## Comparative Genomics

## 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

## 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



### Tabular access to tree data

Selected orthologues  
View protein alignments of all orthologues

Species	Type	dN/dS	StatID & gene name	Compare	Location	Target Size	Query Size
<i>Aspergillus nidulans</i> G000000000	1 to 1	n/a	AN1.1.1.g0135	Alignment (protein) Alignment (cds) Gene Tree (img)	Mus1.1.8000007.02087.07865.1	73	76
<i>Aspergillus nidulans</i> G000000000	Many-to-many	n/a	Asp1.118110.p	Alignment (protein) Alignment (cds) Gene Tree (img)	Contig13047.8391.26490.1.p	27	28
<i>Aspergillus nidulans</i> G000000000	Many-to-many	n/a	Asp1.118111.p	Alignment (protein) Alignment (cds) Gene Tree (img)	Contig13047.8578.86086.1.p	26	29
<i>Aspergillus nidulans</i> G000000000	Many-to-many	n/a	Asp1.1191.0.p	Alignment (protein) Alignment (cds) Gene Tree (img)	Contig13088.198198.19819.1.p	28	23
<i>Aspergillus nidulans</i> G000000000	Many-to-many	n/a	Asp1.1191.0.p	Alignment (protein) Alignment (cds) Gene Tree (img)	Contig13088.198053.110077.1.p	28	30
<i>Aspergillus nidulans</i> G000000000	1 to 1	n/a	ANCAN_0055	Alignment (protein) Alignment (cds) Gene Tree (img)	ANCAN0011_Cover116.993032.30990.1	68	60
<i>Aspergillus nidulans</i> G000000000	1 to 1	n/a	Asp1.1017.g0112	Alignment (protein) Alignment (cds) Gene Tree (img)	Asp1.1017.001.500983.412788.1	66	67

### User Accounts

### User accounts

- Saving and sharing attached data tracks
- Saving configuration settings
- Saving and sharing BLAST results

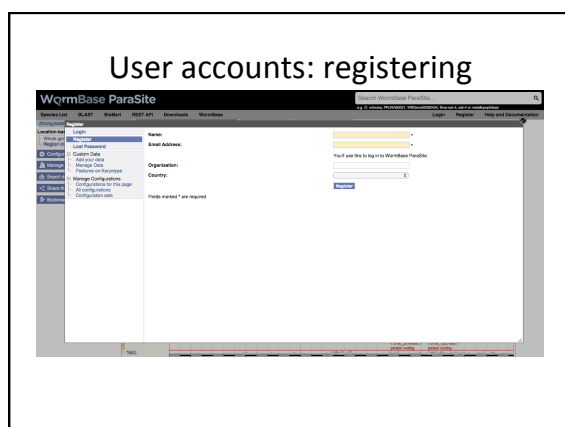
### User accounts

WormBase ParaSite

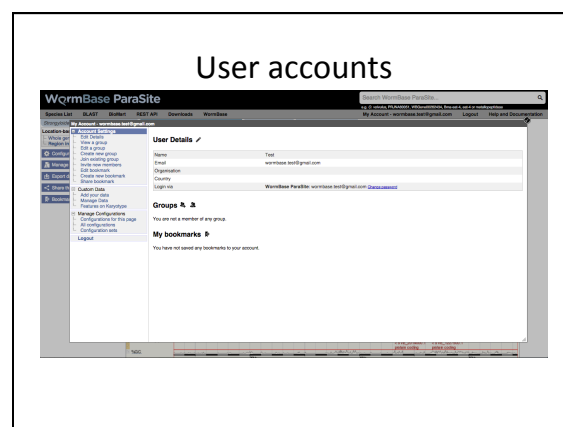
Navigation: BLAST, Download, WormBase

Log Out (highlighted)

## User accounts: registering



## User accounts



## Part 2: 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

WormBase ParaSite

Species List **BLAST** Blast! REST API Downloads WormBase

Species: *Bm2147* Location: *Bmal\_v3\_scaffolds1.3.207.430-3.210.542* Gene: *Bm2147*

Gene-based displays

- Summary
- Splice variants
- Sequence
- External references
- Ontologies
  - GO: Molecular function
  - GO: Cellular component

Gene: *Bm2147* WBGene0222408

Location: *SuperContig Bmal\_v3\_scaffolds1.3.207.430-3.210.542* forward

About this gene: This gene has 2 transcripts ([splice variants](#)) and 112 orthologues

Gene type: Protein coding

Defaults to the species you are currently browsing

## Using the ParaSite BLAST

WormBase ParaSite

Species List **BLAST** Blast! REST API Downloads WormBase

Species: *Bm2147* Location: *Bmal\_v3\_scaffolds1.3.207.430-3.210.542* Gene: *Bm2147*

Gene-based displays

- Summary
- Splice variants
- Sequence
- External references
- Ontologies
  - GO: Molecular function
  - GO: Cellular component

Gene: *Bm2147* WBGene0222408

Location: *SuperContig Bmal\_v3\_scaffolds1.3.207.430-3.210.542* forward

About this gene: This gene has 2 transcripts ([splice variants](#)) and 112 orthologues

Gene type: Protein coding

Marked-up sequence

Download sequence **BLAST this sequence**

Exons: Bm2147 exons All exons in this region

```

>supercontigB_malayi-3.118mal_v3_scaffolds1.3206830:3211142:1
ATTACTCTTCCTATTTTCAGACATCTTCCAGACACCTTTATTTTAACTCTTCTTCT
CTACTTAACTTAACTTACAGACAGCCTTTATTTAGACACCAATATAC
TTCAGATTAAGTATTTTTCAGATTTCTTATCTTTCAGACATCTTCTT
ATTTACAGACAGCCTTTTTCAGATTTCTTATCTTTCAGACATCTTCTT
TTAAATATTTCTATTTTACATATACTACAGACATCTTTCAGACACATCTC
TTCAGACTTCTTCTTTCAGACATCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT
CAATTCATCAGATTTTCAGATTTTAAATATTTATTTCTTCTTCTTCTTCTTCTT
CGCTTTTTCAGACATCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT
TGCTTTCTCAGTTTCAGACATTTTTCAGACATGATTTTCAGACATTTTAAAA
TTCTTTATTTAGACAAATTTTCAGACAAAGATTTTATTTAAATTTCTTCTTCTT
ATGAGACCTTCAGATTTTTCAGACATCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT
TAAATCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT
ATGATGATTCAGACATTTATTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT
    
```

## Using the ParaSite BLAST

WormBase ParaSite

Species List **BLAST** Blast! REST API Downloads WormBase

Species: *Bm2147* Location: *Bmal\_v3\_scaffolds1.3.207.430-3.210.542* Gene: *Bm2147*

Gene-based displays

- Summary
- Splice variants
- Sequence
- External references
- Ontologies
  - GO: Molecular function
  - GO: Cellular component

Gene: *Bm2147* WBGene0222408

Location: *SuperContig Bmal\_v3\_scaffolds1.3.207.430-3.210.542* forward

About this gene: This gene has 2 transcripts ([splice variants](#)) and 112 orthologues

Gene type: Protein coding

Marked-up sequence

Download sequence **BLAST this sequence**

Exons: Bm2147 exons All exons in this region

```

>supercontigB_malayi-3.118mal_v3_scaffolds1.3206830:3211142:1
ATTACTCTTCCTATTTTCAGACATCTTCCAGACACCTTTATTTTAACTCTTCTTCT
CTACTTAACTTAACTTACAGACAGCCTTTTATTTAGACACCAATATAC
TTCAGATTAAGTATTTTTCAGATTTCTTATCTTTCAGACATCTTCTT
ATTTACAGACAGCCTTTTTCAGATTTCTTATCTTTCAGACATCTTCTT
TTAAATATTTCTATTTTACATATACTACAGACATCTTTCAGACACATCTC
TTCAGACTTCTTCTTTCAGACATCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT
CAATTCATCAGATTTTCAGATTTTAAATATTTATTTCTTCTTCTTCTTCTTCTT
CGCTTTTTCAGACATCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT
TGCTTTCTCAGTTTCAGACATTTTTCAGACATGATTTTCAGACATTTTAAAA
TTCTTTATTTAGACAAATTTTCAGACAAAGATTTTATTTAAATTTCTTCTTCTT
ATGAGACCTTCAGATTTTTCAGACATCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT
TAAATCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT
ATGATGATTCAGACATTTATTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT
    
```

BLAST selected sequence

## Using the ParaSite BLAST

WormBase ParaSite

Species List **BLAST** Blast! REST API Downloads WormBase

Species: *Bm2147* Location: *Bmal\_v3\_scaffolds1.3.207.430-3.210.542* Gene: *Bm2147*

Gene-based displays

- Summary
- Splice variants
- Sequence
- External references
- Ontologies
  - GO: Molecular function
  - GO: Cellular component

Gene: *Bm2147* WBGene0222408

Location: *SuperContig Bmal\_v3\_scaffolds1.3.207.430-3.210.542* forward

About this gene: This gene has 2 transcripts ([splice variants](#)) and 112 orthologues

Gene type: Protein coding

Marked-up sequence

Download sequence **BLAST this sequence**

Exons: Bm2147 exons All exons in this region

```

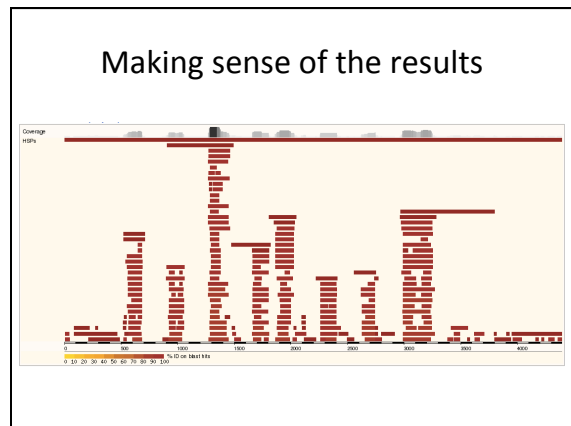
>supercontigB_malayi-3.118mal_v3_scaffolds1.3206830:3211142:1
ATTACTCTTCCTATTTTCAGACATCTTCCAGACACCTTTATTTTAACTCTTCTTCT
CTACTTAACTTAACTTACAGACAGCCTTTTATTTAGACACCAATATAC
TTCAGATTAAGTATTTTTCAGATTTCTTATCTTTCAGACATCTTCTT
ATTTACAGACAGCCTTTTTCAGATTTCTTATCTTTCAGACATCTTCTT
TTAAATATTTCTATTTTACATATACTACAGACATCTTTCAGACACATCTC
TTCAGACTTCTTCTTTCAGACATCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT
CAATTCATCAGATTTTCAGATTTTAAATATTTATTTCTTCTTCTTCTTCTTCTT
CGCTTTTTCAGACATCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT
TGCTTTCTCAGTTTCAGACATTTTTCAGACATGATTTTCAGACATTTTAAAA
TTCTTTATTTAGACAAATTTTCAGACAAAGATTTTATTTAAATTTCTTCTTCTT
ATGAGACCTTCAGATTTTTCAGACATCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT
TAAATCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT
ATGATGATTCAGACATTTATTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT
    
```

BLAST selected sequence

## Making sense of the results

- Score  
Used to assess the biological relevance by describing the alignment quality  
Higher score = higher similarity
- E-value  
Similar to (but not the same as) a *p*-value that has been corrected for multiple testing - decreases exponentially as the score increases  
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

WormBase ParaSite

WormBase Home | ParaSite Home

Search WormBase ParaSite

BLAST | BLAST | Download

Home | About | Help

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)

Dataset	All Species (WSPS4)	Species	Species
Filters	None selected	Region	Region
Attributes	Genome project	Gene	Gene
Gene stable ID		Gene Ontology	Gene Ontology
		Multi-species Comparisons	Multi-species Comparisons
		Protein Domains	Protein Domains

### Setting filters

WormBase ParaSite

WormBase Home | ParaSite Home

Search WormBase ParaSite

BLAST | BLAST | Download

Home | About | Help

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)

Dataset	All Species (WSPS4)	Species	Species
Filters	None selected	Region	Region
Attributes	Genome project	Gene	Gene
Gene stable ID		Gene Ontology	Gene Ontology
		Multi-species Comparisons	Multi-species Comparisons
		Protein Domains	Protein Domains

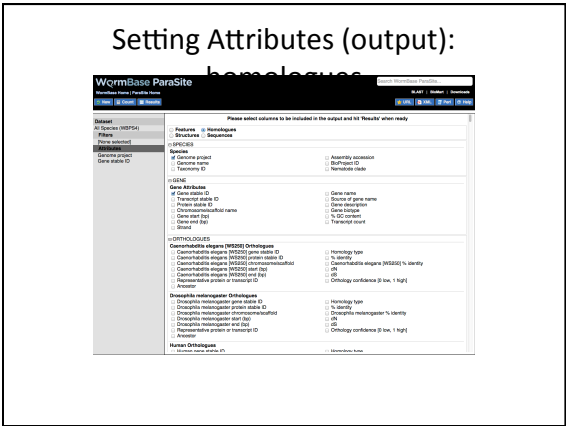
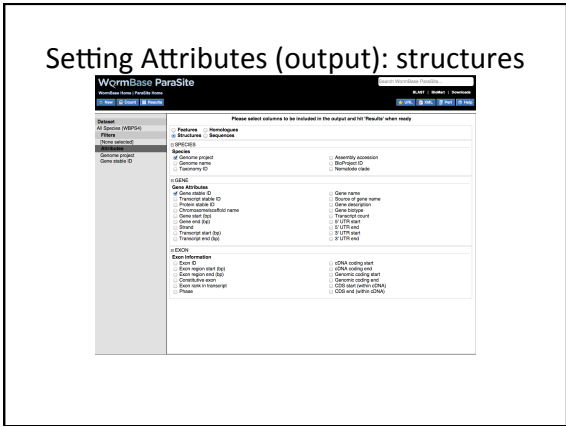
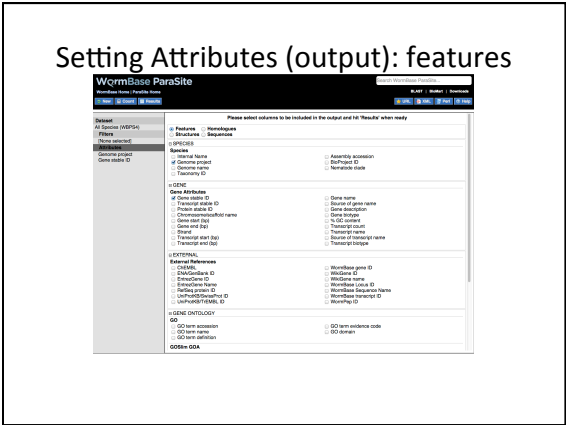
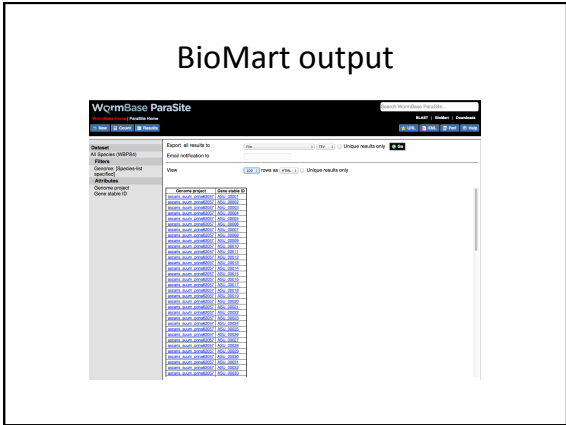
- **SPECIES:** Use this filter to select either individual genomes or nematode clades.
  - Multiple genomes can be selected by holding down the ctrl key or the option key on a Mac.

- **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.

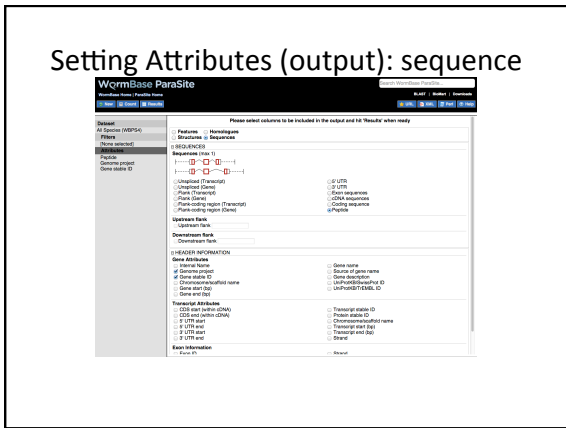
- **GENE:** Specify a list of genes with WormBase IDs, or one of the other ID types listed.
  - IDs should be separated by a new line.

- 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.



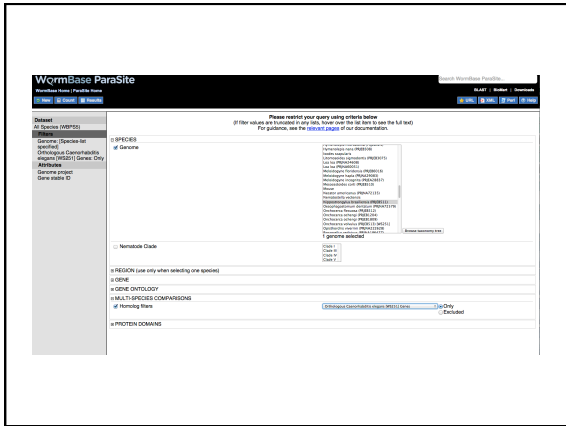
### Setting Attributes (output): sequence



Practical exercises: part 1

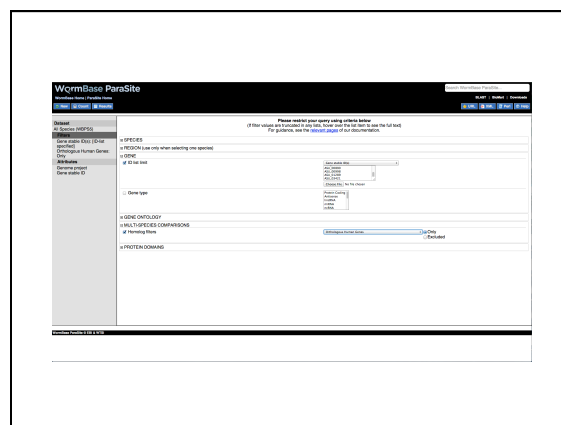
"I'd like to extract all *C. elegans* orthologs for *Nippostrongylus* genes involved in a particular process."

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.



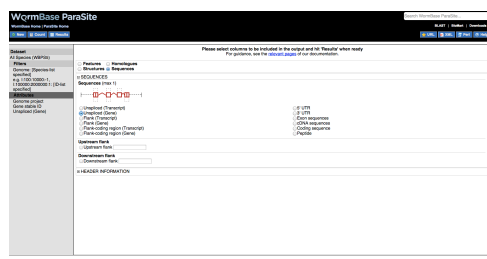
"I have a list of genes from *Ascaris suum* and would like to know which ones have orthologs in humans and mammals and which ones might be nematode-specific."

- 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.”

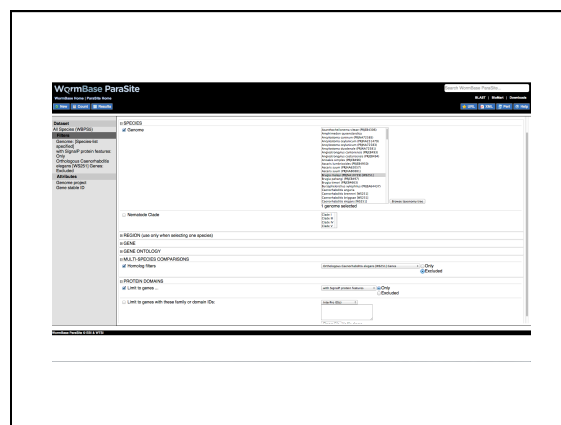
- From the **SPECIES** filter choose *Schistosoma mansoni*.
- Open the **REGION** section and enter the list of co-ordinates under 'Multiple regions' separated by commas or new lines.
- In **Attributes**, check the **Sequences** option, then in the SEQUENCES section choose **Unspliced (genes)**.
- Click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



## Part 4: Bulk downloads and programmatic access

## Downloads

- All genomes, proteomes and annotations available to download as compressed flat files
- Ideal for use with alignment software, etc.
- Data from all previous releases available to download
- Please remember to cite the genome provider and WormBase ParaSite

## Downloads – File Formats

Genomic	Raw FASTA genome file
Masked Genomic	Genome FASTA with repeat regions hard-masked
Soft-masked Genomic	Genome FASTA with repeat regions soft-masked
Annotations	GFF3 file containing all annotations
Proteins	FASTA protein file
mRNA Transcripts	FASTA of the spliced full-length transcripts
CDS Transcripts	FASTA of the spliced CDS-portion of the protein coding transcripts

## Access using R

- Access our database directly from R, via the biomaRt package
- Syntax identical to Ensembl
- Very quick access to large amounts of data
- Please don't use excessively (i.e. download the results once then store them locally for processing)

## WormBase ParaSite in R

- Install the biomaRt package:

```
source("http://bioconductor.org/biocLite.R")
biocLite("biomaRt")
```

- Install the biomaRt package:

```
library(biomaRt)
```

## WormBase ParaSite in R

- Establish a connection to WormBase ParaSite

```
mart <- useMart("parasite_mart",
               dataset = "wbps_eg_gene",
               host = "parasite.wormbase.org")
```

## WormBase ParaSite in R

- Example: get all the *Schistosoma mansoni* genes with a *C. elegans* orthologue:

```
genes <- getBM(mart = mart,
               filters = c("species_id_1010",
                           "with_celegans_eg_homologue"),
               value = list("prj36577", TRUE),
               attributes = c("ensembl_gene_id",
                              "celegans_eg_gene"))
head(genes)
```

	ensembl_gene_id	celegans_eg_gene
1	Smp_078570	WBGene00009448
2	Smp_063300	WBGene00004450
3	Smp_210640	WBGene00009305
4	Smp_049930	WBGene00010465
5	Smp_132740	WBGene00001395
6	Smp_132740	WBGene00001396

## Language neutral queries

- REST API allows access using any programming language
- For processing large amounts of data: consider whether making one query to BioMart may be more suitable
- Examples provided in Perl, Python, Ruby, Java, Curl and Wget

## Endpoint Catalogue

### Comparative Genomics

Resource	Description
<a href="#">GET /rest/genetree/id/:id</a>	Retrieves a gene tree dump for a gene tree stable identifier
<a href="#">GET /rest/genetree/member/id/:id</a>	Retrieves a gene tree that contains the stable identifier
<a href="#">GET /rest/genetree/member/:symbol/:species/:symbol</a>	Retrieves a gene tree containing the gene identified by a symbol
<a href="#">GET /rest/homology/id/:id</a>	Retrieves homology information (orthologues) by gene id
<a href="#">GET /rest/homology/symbol/:species/:symbol</a>	Retrieves homology information (orthologues) by symbol

## Endpoint Specifics

### GET genetree/member/id/:id

Retrieves a gene tree that contains the stable identifier

#### Parameters

##### Required

Name	Type	Description	Default	Example Values
id	String	A stable ID	-	WBGene00225050

## Endpoint Examples

### Example Requests

[/rest/genetree/member/symbol/brugia\\_malayi\\_prjna10729/Bma-unc-1?content-type=text/x-phyloxml%2Bxml](#)

Example output [Perl](#) [Python2](#) [Python3](#) [Ruby](#) [Java](#) [Curl](#) [Wget](#)

```
<?xml version="1.0" encoding="UTF-8"?>
<phyloxml xmlns:schemaLocation="http://www.phyloxml.org http://www.phyloxml.org/1.0/phyloxml.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.phyloxml.org">
  <phylogeny rootedge="true" type="gene tree">
    <clade branchLength="0">
      <confidence type="duplication_confidence_score">0.7311</confidence>
      <taxonomy>
        <id>3388</id>
        <scientificName>Metazoa/scientific_name</scientificName>
        <rank>0</rank>
        <events>
          <typeSelectionOrDuplicationType>
            <duplicationOrDuplicationType>
              <events>
                <clade branchLength="0.003861">
                  <confidence type="duplication_confidence_score">0.1384</confidence>
                </clade>
              </events>
            </duplicationOrDuplicationType>
          </typeSelectionOrDuplicationType>
        </events>
      </clade>
    </clade branchLength="0">
      <confidence type="duplication_confidence_score">0.1384</confidence>
    </clade>
  </phylogeny>
</phyloxml>
```

## Code Examples

### Example Requests

[/rest/genetree/member/symbol/brugia\\_malayi\\_prjna10729/Bma-unc-1?content-type=text/x-phyloxml%2Bxml](#)

Example output [Perl](#) [Python2](#) [Python3](#) [Ruby](#) [Java](#) [Curl](#) [Wget](#)

```
1. use strict;
2. use warnings;
3.
4. use HTTP::Tiny;
5.
6. my $http = HTTP::Tiny->new();
7.
8. my $server = "http://parasite.wormbase.org";
9. my $url = "/genetree/member/symbol/brugia_malayi_prjna10729/Bma-unc-1?";
10. my $response = $http->get($server.$url, {
11.     headers => { 'Content-type' => 'text/x-phyloxml+xml' }
12. });
13.
14. die "Failed!\n" unless $response->{success};
15.
16. print "$response->{status} $response->{reason}\n";
17.
18.
```