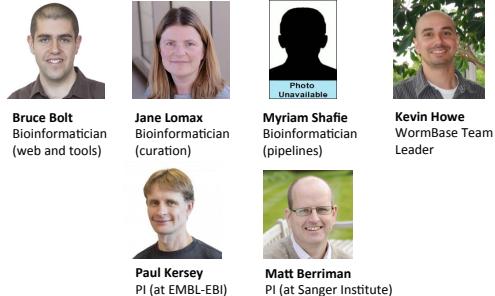


## WormBase ParaSite Workshop

Edinburgh  
9<sup>th</sup> March 2016

## WormBase ParaSite Team



## 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



## 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

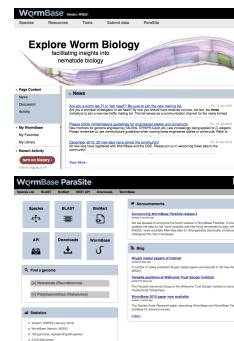
The screenshot shows the genome assembly page for *Trichuris muris* (PRJEB120). It includes sections for Gene annotation, Comparative genomics, and Downloads. The gene annotation section shows a genome browser track for *TM00232*.

## 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



## 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 moments giving us some feedback about today's workshop. Your feedback will help us to improve future workshops.

1. Did you see WormBase ParaSite before the workshop?
2. Will you use WormBase ParaSite more often after this workshop?
3. Would you recommend this workshop to your colleagues?
4. How useful are each section of the workshop?
 

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

Very satisfied	Satisfied	Not happy	Unsatisfied, if possible
Very satisfied	Satisfied	Not happy	Unsatisfied, if possible
6. Do you have any other comments or feedback?

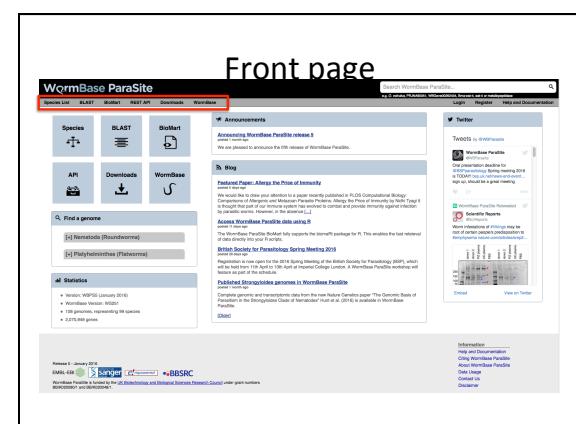
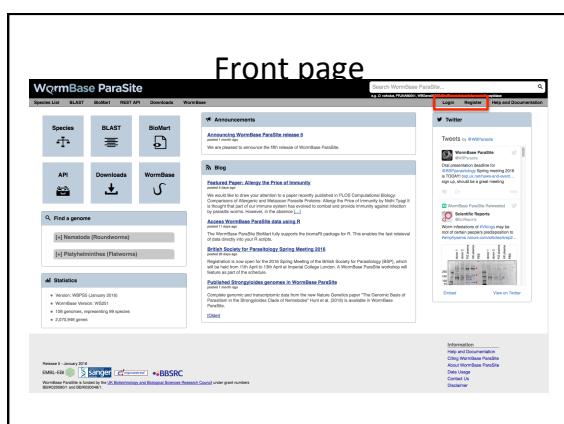
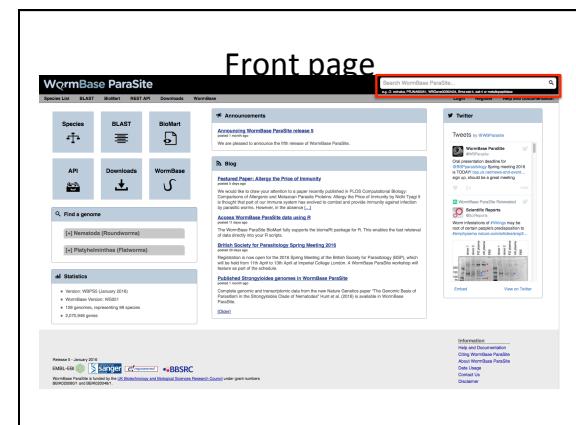
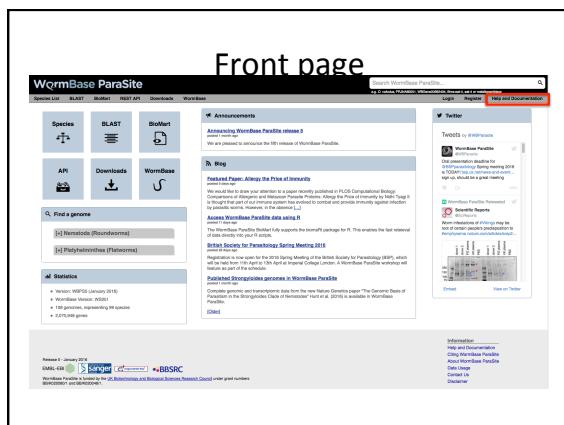
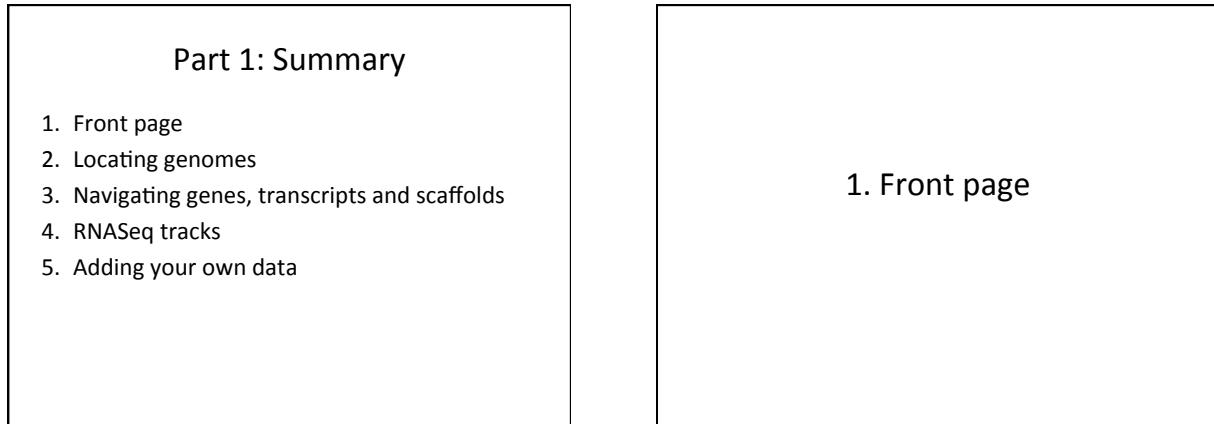
Thank you!

## 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



## 2. Locating genomes

Genomes list					
Contents	Provider	Assembly	BioProject ID	Taxonomy ID	
<a href="#">Nematoda (21)</a>					
<a href="#">Platyhelminthes (26)</a>					
<b>Nematoda</b>					
<i>Acetosphaera vires</i>	University of Edinburgh	Alpha-1	PRJNA24200	39170	
<i>Acynthosoma caninum</i>	Genome Institute at Washington University	A_canninum_3.0-rc0-09	PRJNA25986	39170	
<i>Acynthosoma oxytropis</i>	Genome Institute at Washington University	A_oxytropis_3.0-rc0-09	PRJNA25987	33335	
<i>Acynthosoma punctatum</i>	Genome Institute at Washington University	A_punctatum_3.0-rc0-09	PRJNA25988	33335	
<i>Acynthosoma duodenale</i>	Genome Institute at Washington University	A_duodenale_2.0-rc0-09	PRJNA25981	31682	
<i>Angiostrongylus cantonensis</i>	Wellcome Trust Sanger Institute	A_cantonensis_3.0-rc0-09	PRJNA25441	394528	
<i>Angiostrongylus cantonensis</i>	Wellcome Trust Sanger Institute	A_cantonensis_Costa_Rica_1-3	PRJNA25442	394528	
<i>Ascaris simplex</i>	Wellcome Trust Sanger Institute	A_simplex_v1_5	PRJEB898	8288	
<i>Ascaris suum</i>	University of Colorado School of Medicine	ASU_v1.0	PRJNA26502	8288	
<i>Ascaris suum</i>	University of Melbourne	A_suum_v1.0	PRJNA26503	8288	
<i>Brigia matyi</i>	Wellcome Trust Sanger Institute	B_matyi_1	PRJNA31209	8279	
<i>Brigia matyi</i>	Wellcome Trust Sanger Institute	B_matyi_2	PRJNA31210	8279	
<i>Brigia matyi</i>	Wellcome Trust Sanger Institute	B_matyi_3	PRJNA31211	8279	
<i>Buaphelomys oxyphilus</i>	Wellcome Trust Sanger Institute	B_oxyphilus_1-1	PRJNA26442	32365	
<i>Oxytelus jacobsoni</i>	University of Edinburgh	O_jacobsoni_Chester_V1.0	ASW001191-001	32365	
<i>Dirofilaria immitis</i>	Wellcome Trust Sanger Institute	D_immitis_1.0	PRJNA259118	28152	
<i>Dirofilaria immitis</i>	Genome Institute at Washington University	D_immitis_2.0-rc0-09	PRJNA25924	28152	
<i>Dirofilaria immitis</i>	University of Edinburgh	D_immitis_3.0-rc0-09	PRJNA25925	28152	
<i>Dirofilaria medinensis</i>	Wellcome Trust Sanger Institute	D_medinensis_Ghana_v1_0	PRJNA5600	31878	
<i>Elasmophora espebi</i>	Wellcome Trust Sanger Institute	E_espebi_v1.0	PRJNA26002	147241	
<i>Elasmophora espebi</i>	Wellcome Trust Sanger Institute	E_espebi_v1.0_Costa_Rica_v1.0	PRJNA26003	147241	
<i>Globodera pallida</i>	Wellcome Trust Sanger Institute	G_pallida_1.0	PRJNA123	36500	
<i>Globodera pallida</i>	Wellcome Trust Sanger Institute	G_pallida_2.0-rc0-09	PRJNA25920	36500	
<i>Haemonchus contortus</i>	Wellcome Trust Sanger Institute	Haemonchus_contortus_MV0-2	PRJNA25909	8288	
<i>Haemonchus contortus</i>	University of Melbourne	Hc_v1_coding_assembled	PRJNA26002	8288	
<i>Haemonchus contortus</i>	Wellcome Trust Sanger Institute	Hc_v1_coding_assembled	PRJNA26003	8288	
<i>Heligmosoma polygyra</i>	Wellcome Trust Sanger Institute	H_bakeri_Escherichia_v1_5	PRJNA25920	37398	

### 3. Navigating genes, transcripts and scaffolds

## Gene pages: exons

# Transcript pages: protein domains

**Protein summary**

Protein domain for gene 7483.1

Statistics

- Ave. residue weight: 109.833 g/mol
- Charge: 4.5
- Isoelectric point: 7.7295
- Molecular weight: 55,170.591 g/mol
- Number of residues: 475 aa

Location view: zooming

**Location view: gene/transcript info**

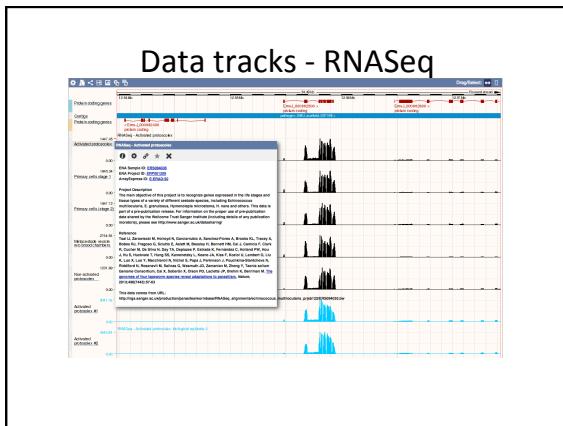
**Location view: jump to...**

**Location view: configure**

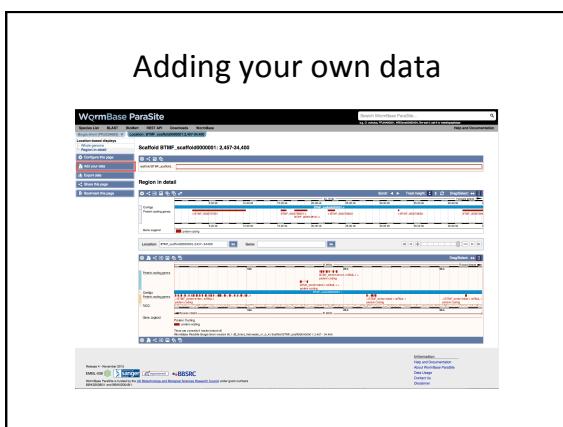
**Location view: export data**

**4. RNASeq tracks**

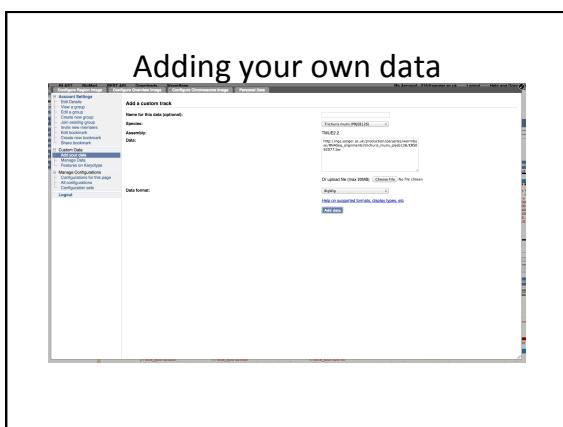
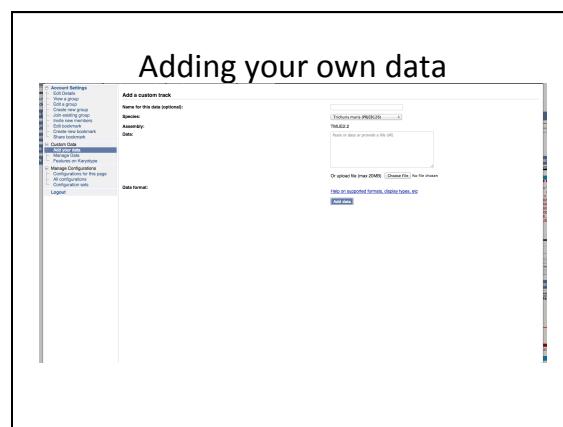
**Data tracks - RNASeq**



## 5. Adding your own data



## Adding your own data



## Part 1b: Browsing the website

Searching the website  
Comparative genomics  
User accounts

# Searching

[Search WormBase ParSite](#)

Species List
BLAST
BLAST
REST API
Downloads
Wormbase

---



Species  
List



BLAST



BLAST



REST API



Downloads



Wormbase

---

**Annotations**

Annotation of *WormBase* ParSite release 3

We are pleased to announce the 11th release of WormBase ParSite.

Search WormBase ParSite

---

**B: Blog**

*Feature Paper: Always in the Price of Integrity*

We would like to draw your attention to a paper published in PLoS Computational Biology. The paper, titled "Feature Paper: Always in the Price of Integrity", was published on 10th April 2014. It describes how the use of an immune system has evolved to inhibit and provide immunity against infection by viruses.

**Access WormBase ParSite data using R**

The WormBase ParSite R package fully supports the R/Bioconductor package R. This enables the fast retrieval of WormBase ParSite data using R. The R/Bioconductor package is available on the WormBase ParSite website.

---

**British Society for Parapsychology Spring Meeting 2014**

The WormBase ParSite team will be attending the British Society for Parapsychology (BSP) meeting, which will be held from 11th to 13th April at Imperial College London. A WormBase workshop will feature on the 12th April.

**Published Paralogous genes in *WormBase ParSite***

Complete genomic and proteomic data from the new *WormBase* Genomic (WormBase ParSite) is available on the WormBase ParSite website. WormBase ParSite (WormBase ParSite 2014) is a version of WormBase ParSite.

---

**Information**

WormBase ParSite is a knowledge base of the *WormBase* and *Bioassay Research Group* under prior numbers

**Information**

Very few documents  
about WormBase ParSite  
about Wormbase ParSite  
about Wormbase

The screenshot shows the WormBase ParaSite search results for the query "eat-4". The results are divided into two main sections: "Search results for 'eat-4'" and "Cross-references for 'eat-4'".

**Search results for 'eat-4' (1 of 83 genes found in WormBase ParSite)**

**eat-4** (WormBase ID:CB1001161)

**Description:** Probable vesicular glutamate transporter eat-4 [Source:UniProtKB/Swiss-Prot;Acc:P34844]

**Gene ID:** WBGene0000120

**Species:** *C. elegans*

**Location:** 8.23951-9.11382  
*Unc-54 locus; WormBase*

**Cross-references for 'eat-4' (1 of 1)**

**eat-4** (WormBase ID:CB1001161)

**Description:** CDP-4,4'-cyclic di-GMP [Source:UniProtKB/Swiss-Prot;Acc:Q8W4V6]

**Gene ID:** WBGene0000120

**Species:** *C. elegans*

**Location:** CDP-4,4'-cyclic di-GMP  
*Unc-54 locus; WormBase*

**CyP-eat-4** (WormBase ID:CB1001161)

**Description:** Cytochrome P450 eat-4 [Source:UniProtKB/Swiss-Prot;Acc:P34844]

**Gene ID:** WBGene0000120

**Species:** *C. elegans* (adult; female)

**Location:** CYP-1 locus; WormBase

**Pyp-eat-4** (WormBase ID:CB1001161)

**Description:** n/a

**Gene ID:** WBGene0000120

**Species:** *Prionace glauca* (PBLN-2044)

**Location:** Pyp-1 locus; WormBase

**Gene tree:** *C. elegans* orthologues

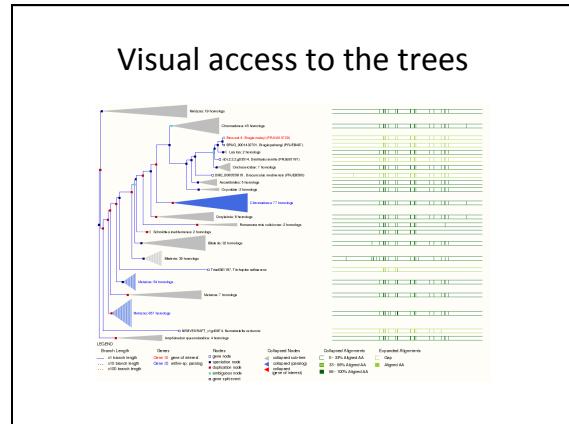
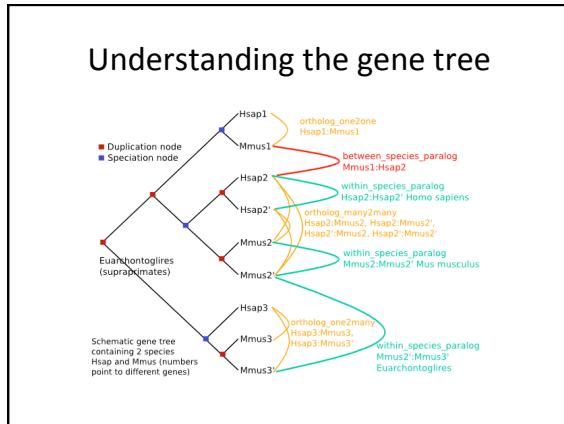
**WormBase**  
*eat-4*

# Comparative Genomics

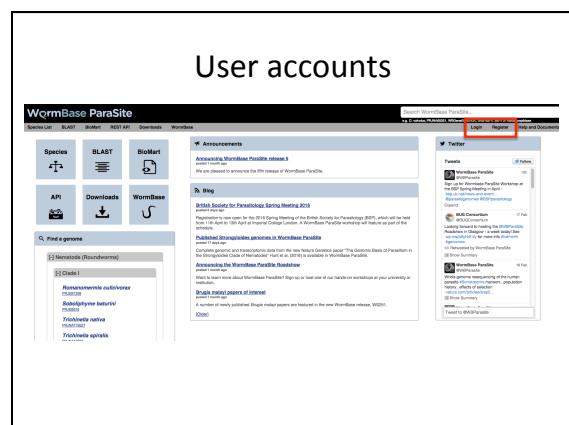
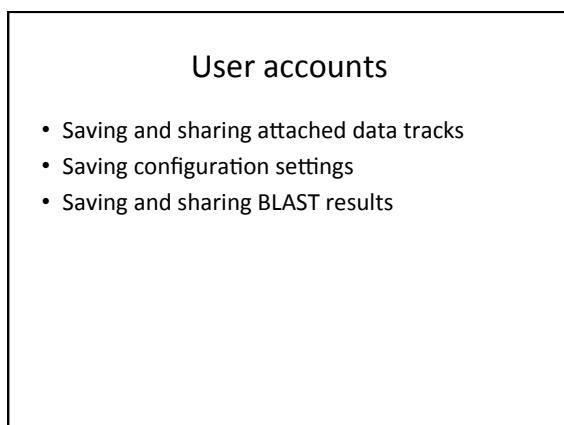
# Introduction

## 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

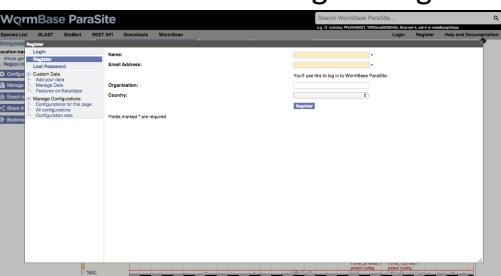


## User Accounts

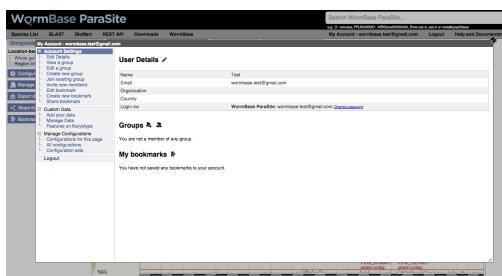


- Saving and sharing attached data tracks
- Saving configuration settings
- Saving and sharing BLAST results

### User accounts: registering



### User accounts



## Part 2: 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

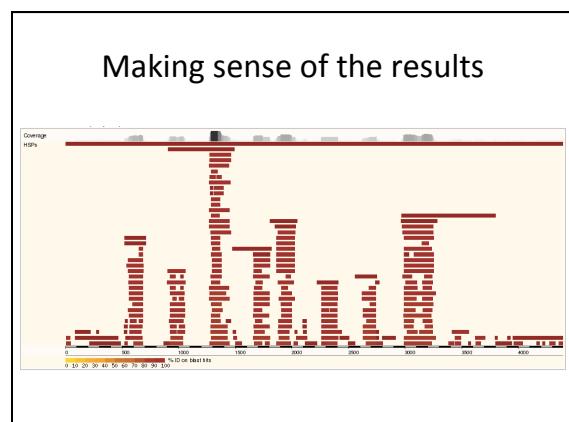
### Types of BLAST

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

## 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



## 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

## 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 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.

The screenshot shows the WormBase ParaSite search interface. In the 'SPECIES' section, the 'Orthologous human genes -> Excluded' option is selected. Other options like 'Orthologous mouse genes -> Excluded' and 'Orthologous genes -> Excluded' are also listed. The 'REGION' section is empty. The 'SEQUENCES' section has a 'Sequences' dropdown set to 'Unspliced (genes)'. The 'Attributes' dropdown is set to 'Only'. The 'Results' button is visible at the bottom left.

"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 'SEQUENCES' section, the 'Sequences' dropdown is set to 'Unspliced (genes)'. The 'Attributes' dropdown is set to 'Only'. The 'Results' button is visible at the bottom left.

"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("prjea36577", 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  WBGene000010465
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
GET /rest/genetree/id/:id	Retrieves a gene tree dump for a gene tree stable identifier
GET /rest/genetree/member/id/:id	Retrieves a gene tree that contains the stable identifier
GET /rest/genetree/member/:symbol/:species/:symbol	Retrieves a gene tree containing the gene identified by a symbol
GET /rest/homology/id/:id	Retrieves homology information (orthologues) by gene id
GET /rest/homology/symbol/:symbol	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-phylxml%2Bxml

Example output | Perl | Python2 | Python3 | Ruby | Java | Curi | Wget

```
<?xml version="1.0" encoding="UTF-8"?>
<phylxml xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.phylxml.org/1.10/phylxml.xsd"
  xmlns="http://www.phylxml.org/1.10/phylxml">
  <phylogeney rooted="true" type="phylogenetic tree">
    <clade branch_length="0">
      <confidence type="duplication_confidence_score">0.731</confidence>
      <taxon>
        <id>3324</id>
        <scientific_name>brugia_malayi</scientific_name>
      </taxon>
      <events>
        <speciation_or_duplication type="speciation_or_duplication">
          <duplication>
            <confidence type="duplication_confidence_score">0.003661</confidence>
            <confidence type="duplication_confidence_score">0.1584</confidence>
          </duplication>
        </speciation_or_duplication>
      </events>
    </clade>
  </phylogeney>

```

## Code Examples

### Example Requests

/rest/genetree/member/symbol/brugia\_malayi\_prjna10729  
/Bma-unc-1?content-type=text/x-phylxml%2Bxml

Example output | Perl | Python2 | Python3 | Ruby | Java | Curi | Wget

```
use strict;
use warnings;
use HTTP::Tiny;
my $http = HTTP::Tiny->new();
my $server = "http://genetree.ncbi.nlm.nih.gov";
my $response = $http->get($server . "/rest/genetree/member/symbol/brugia_malayi_prjna10729/Bma-unc-1?";
                           headers => { "Content-type" => 'text/x-phylxml+xml' });
die "Failed!\n" unless $response->(success);
my $status = $response->(status);
my $reason = $response->(reason);
print "$status $reason\n";

```