WormBase ParaSite Workshop

Edinburgh

9th March 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)

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

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

The Website

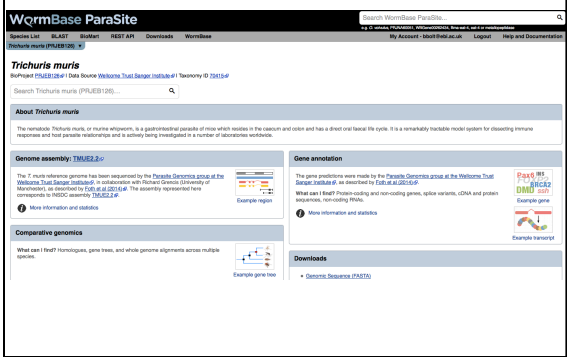
- Genome Browser
- Transcriptomic Data Display
- Gene, transcript and protein information pages
- Comparative Genomics
- Sequence Similarity Search (BLAST)
- Variant Effect Predictor (VEP) *
- Advanced Search Tool (BioMart)
- Access to BioMart data using R
- Programmatic Access (REST API)

* = 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

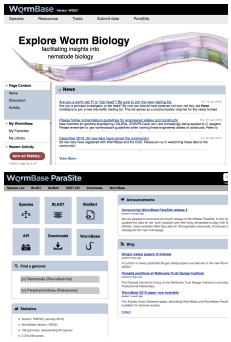


Your Data

- 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

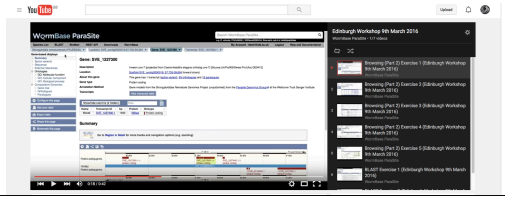


This afternoon's agenda…

- 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

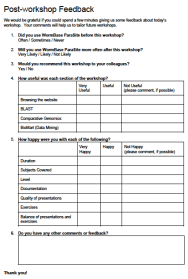
After this workshop…

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



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 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!

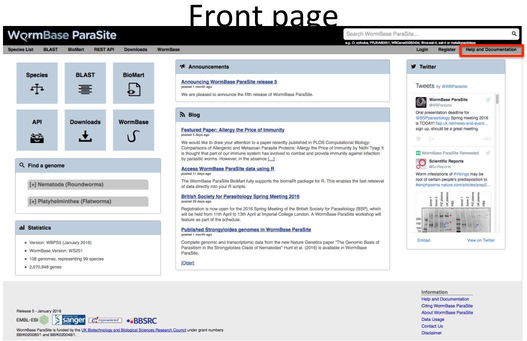
Part 1: Using the website

Part 1: Summary

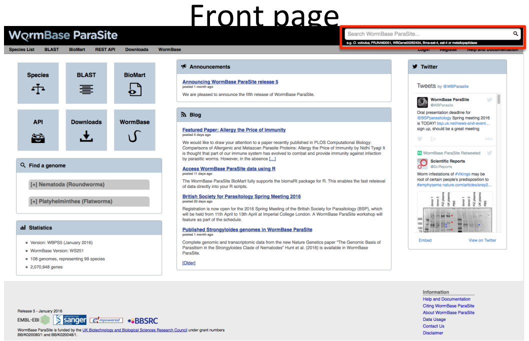
1. Front page
2. Locating genomes
3. Navigating genes, transcripts and scaffolds
4. RNASeq tracks
5. Adding your own data

1. Front page

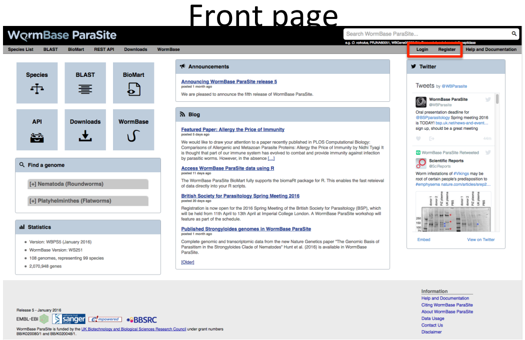
Front page



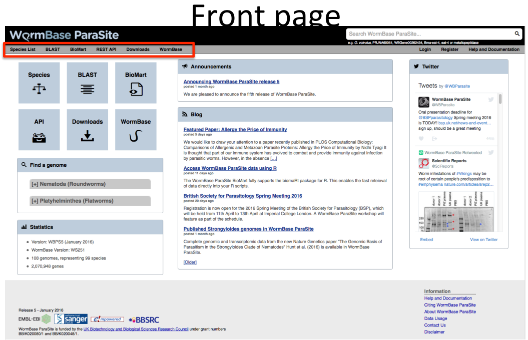
Front page



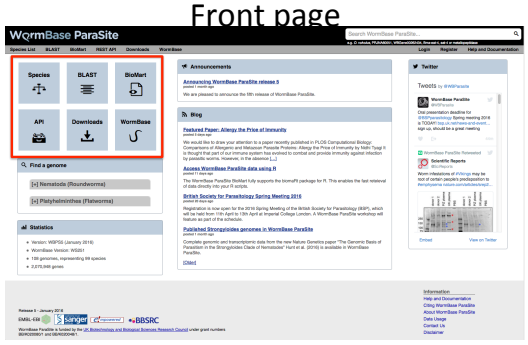
Front page



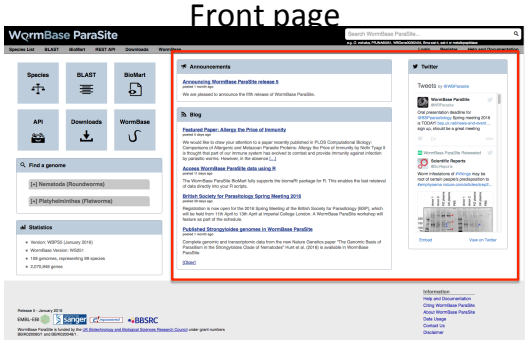
Front page



Front page

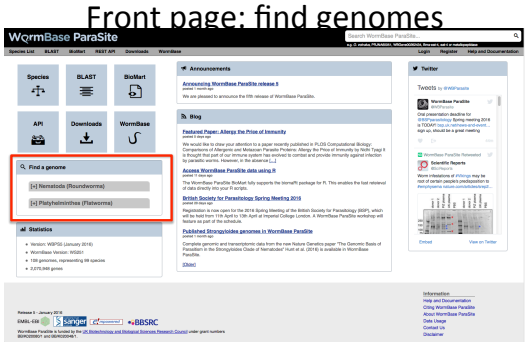


Front page

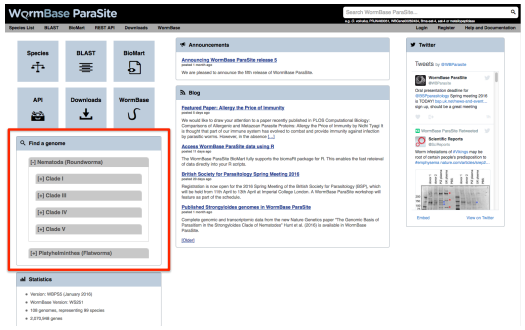


2. Locating genomes

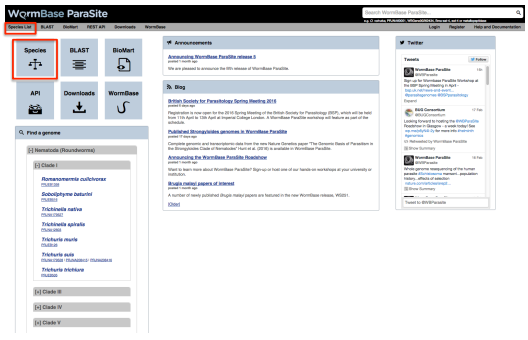
Front page: find genomes



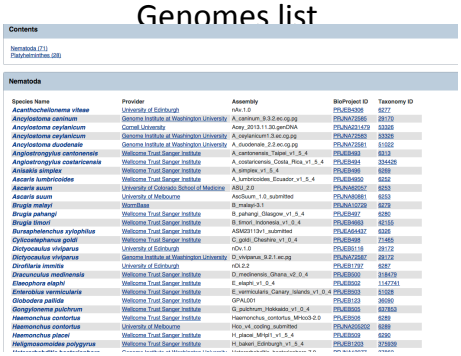
Locating genomes



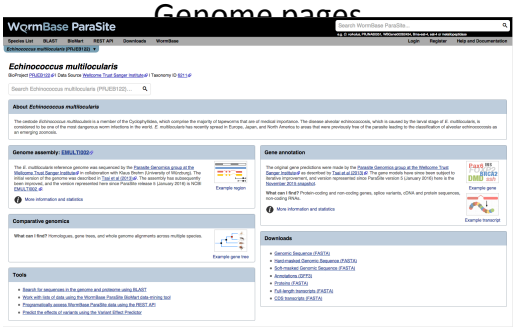
Locating genomes



Genomes list

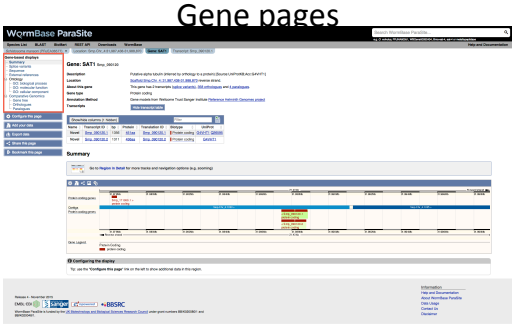


Genome pages

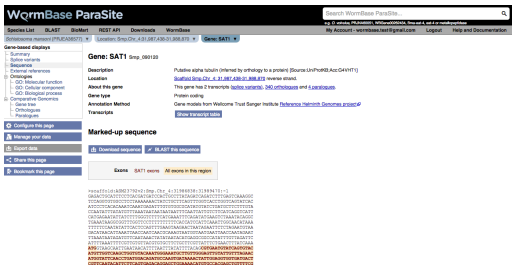


3. Navigating genes, transcripts and scaffolds

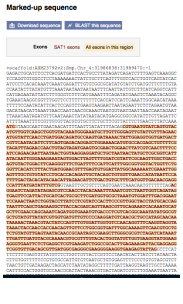
Gene pages



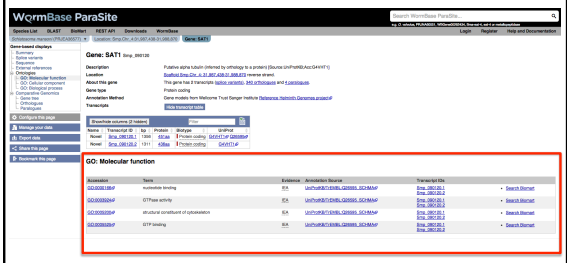
Gene pages: exons



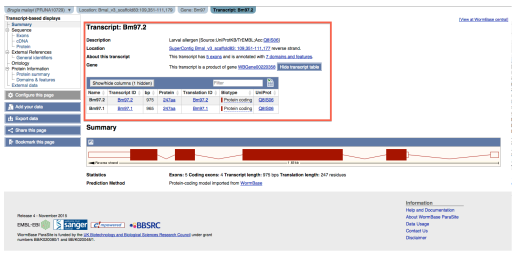
Gene pages: exons



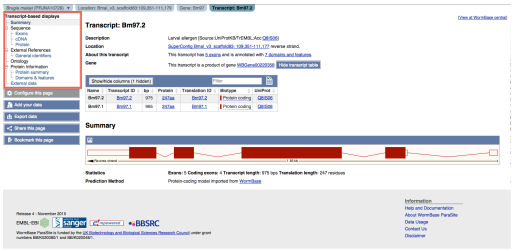
GO terms



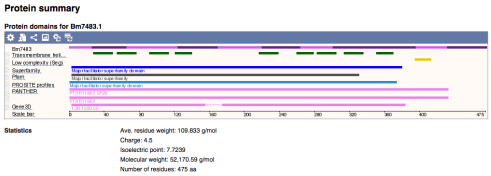
Transcript pages: summary



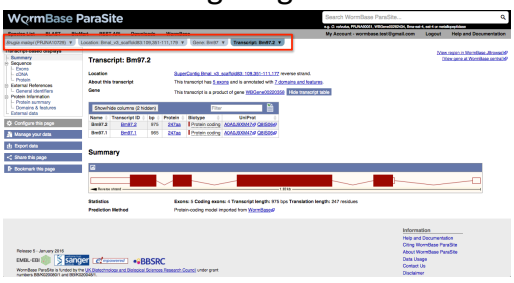
Transcript pages: navigating



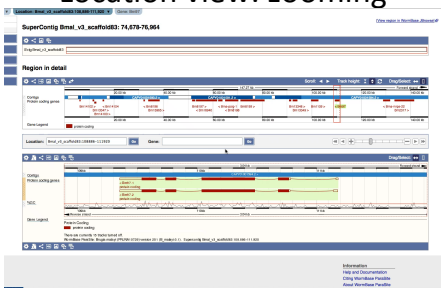
Transcript pages: protein domains



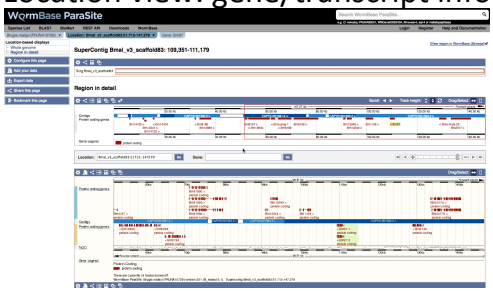
Navigating: tabs



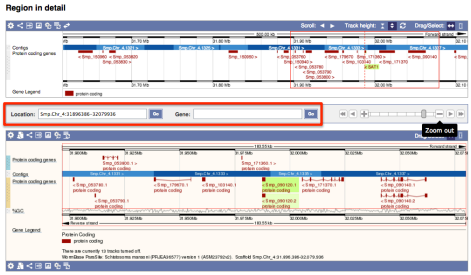
Location view: zooming



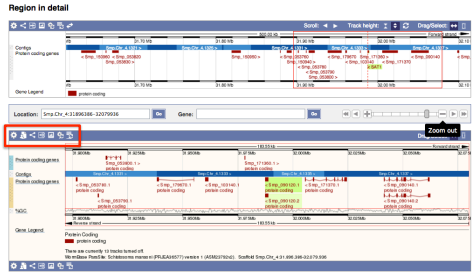
Location view: gene/transcript info



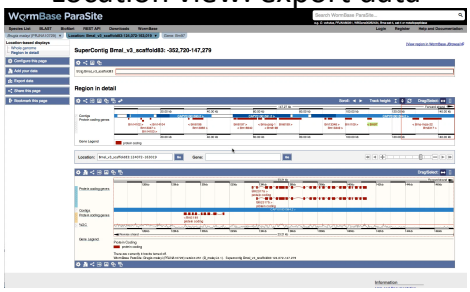
Location view: jump to…



Location view: configure



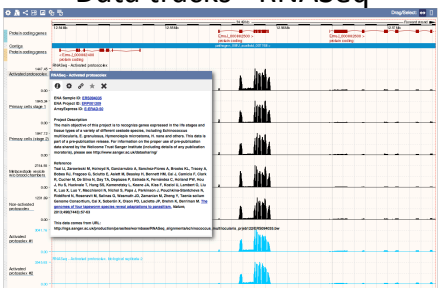
Location view: export data



4. RNASeq tracks

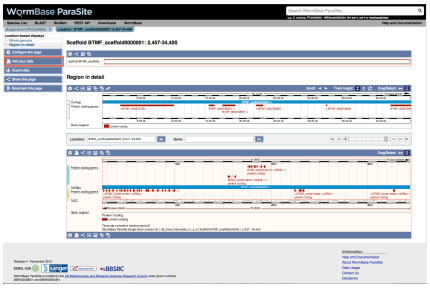
Data tracks - RNASeq

Data tracks - RNASeq

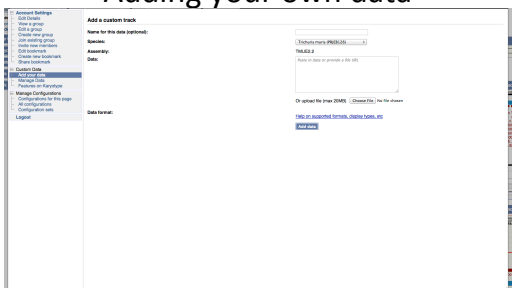


5. Adding your own data

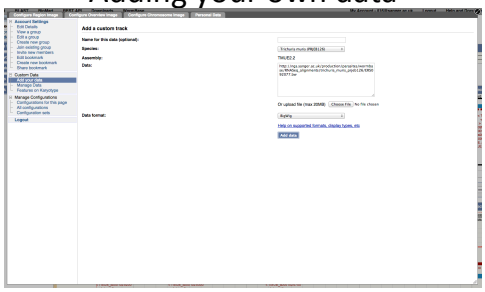
Adding your own data



Adding your own data



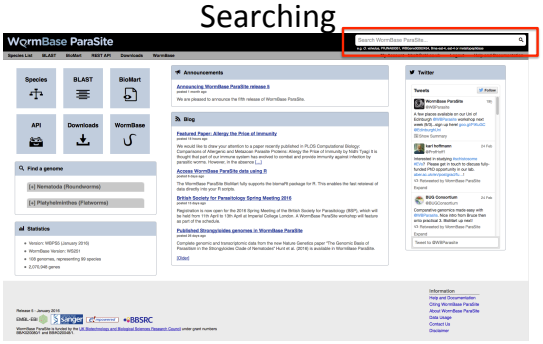
Adding your own data



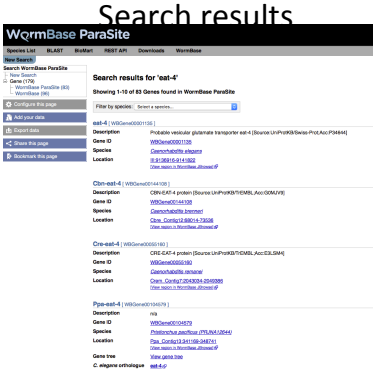
Part 1b: Browsing the website

Searching the website
Comparative genomics
User accounts

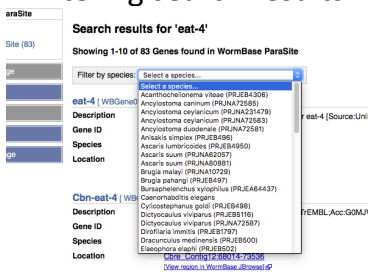
Searching



Search results



Filtering search results



Comparative Genomics

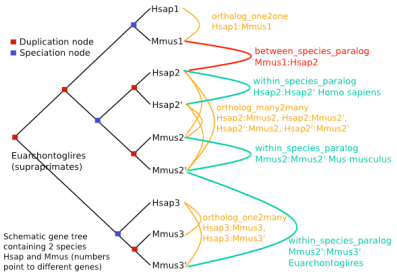
Introduction

- 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

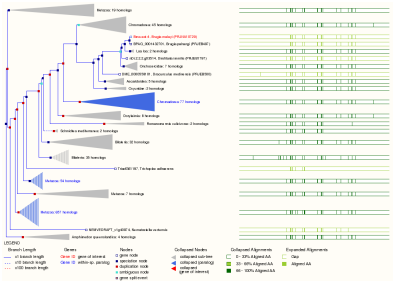
Homology types

- 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

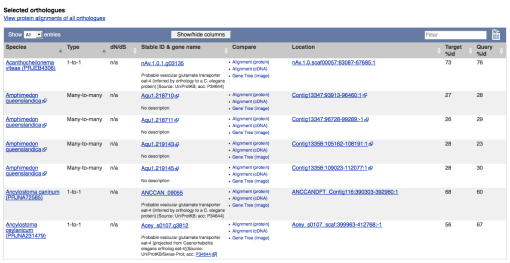
Understanding the gene tree



Visual access to the trees



Tabular access to tree data

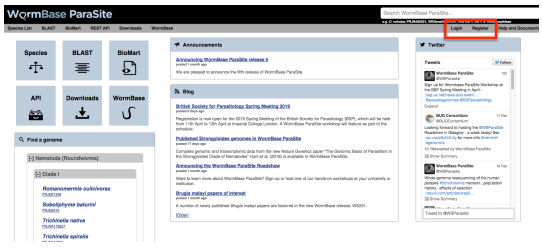


User Accounts

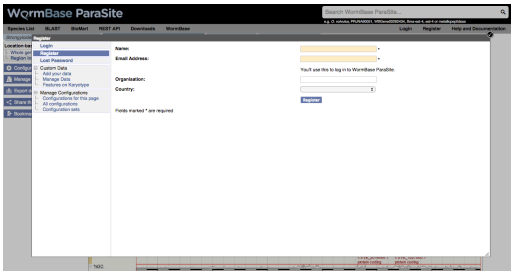
User accounts

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

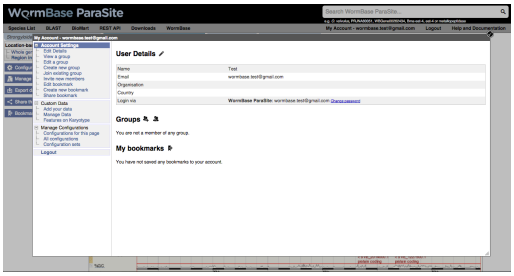
User accounts



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

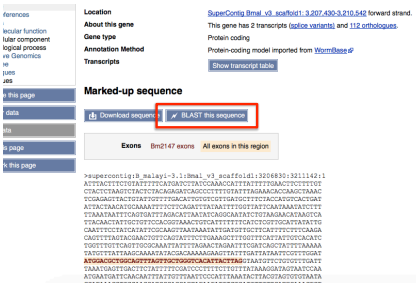
Types of BLAST

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

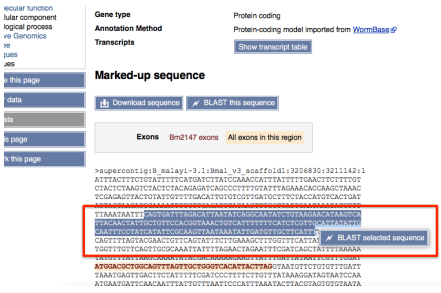
Using the ParaSite BLAST

Defaults to the species you are currently browsing

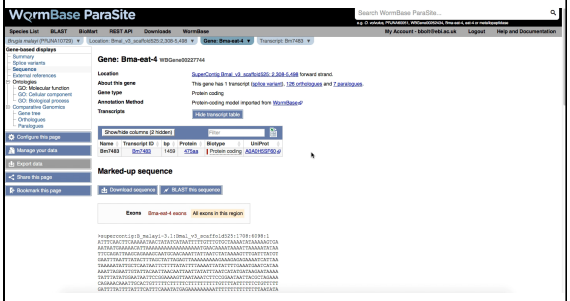
Using the ParaSite BLAST



Using the ParaSite BLAST



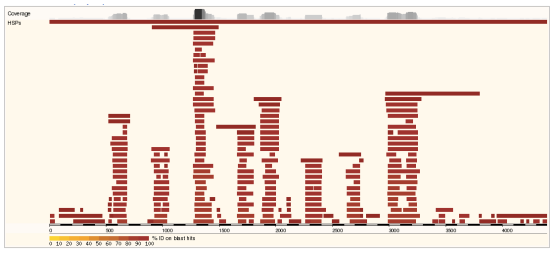
Using the ParaSite BLAST



Making sense of the results

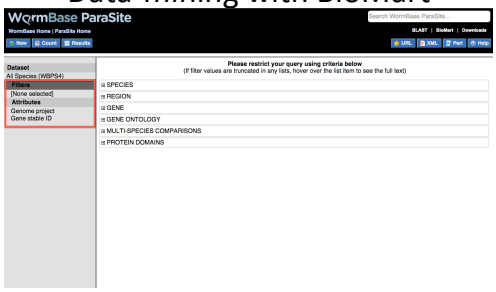
- Score
  Used to assess the biological relevance by describing the alignment quality
  Higher score = higher similarity
- *E*-value
  Similar to (but not the same as) a *p*-value that has been corrected for multiple testing - decreases exponentially as the score increases
  Lower *E*-value = more significant result
- %ID
  Percentage of your query sequence that matches the genome/proteome database

Making sense of the results

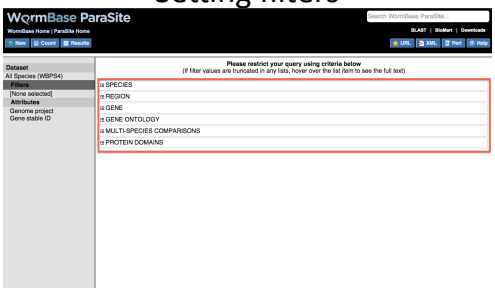


Part 4: Data-mining with BioMart

Data-mining with BioMart



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.

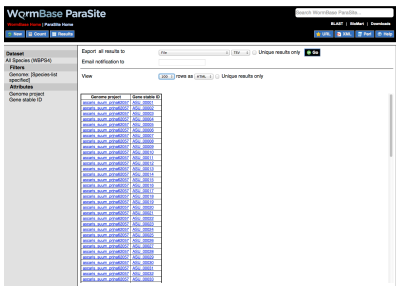
- REGION: Restrict to a particular genomic region.
  - Should only be used where a single genome has been selected, as it is possible that a particular region is present in multiple genomes.
  - If start/end co-ordinates are being specified, a scaffold or chromosome id is always required.
  - Where multiple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

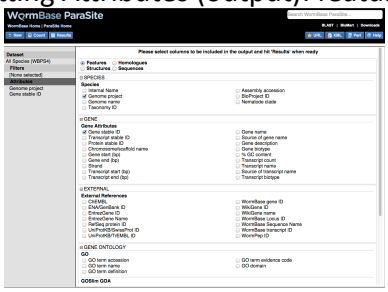
- GENE ONTOLOGY: Restrict by one or more Gene Ontology (GO) terms for functional descriptions.
  - Paste or upload a list of GO IDs or use the autocomplete box to populate the list.
- Alternatively restrict to a particular GO evidence type e.g. Inferred by Electronic Annotation (IEA).
  - Multiple codes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

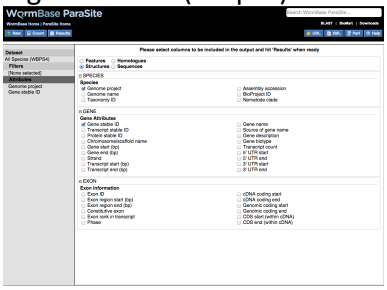
BioMart output



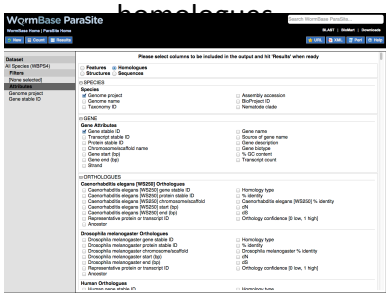
Setting Attributes (output): features



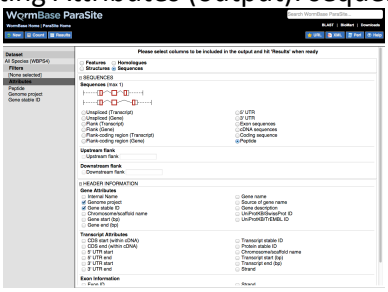
Setting Attributes (output): structures



Setting Attributes (output): homologues



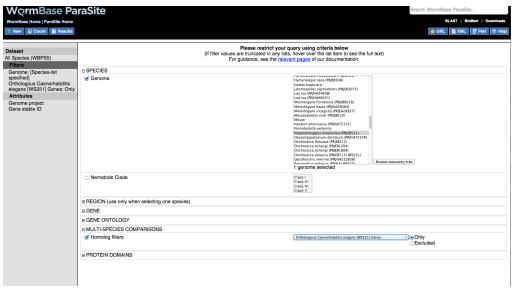
Setting Attributes (output): sequence



Practical exercises: part 1

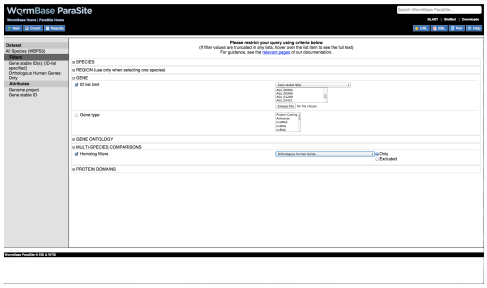
"I'd like to extract all *C. elegans* orthologs for *Nippostrongylus* genes involved in a particular process."

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.



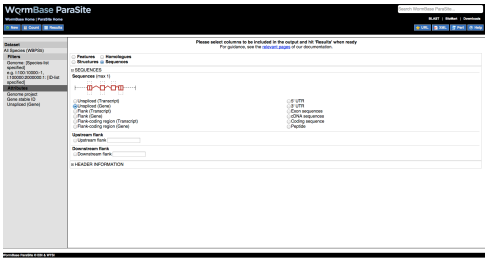
"I have a list of genes from *Ascaris suum* and would like to know which ones have orthologs in humans and mammals and which ones might be nematode-specific."

- 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.



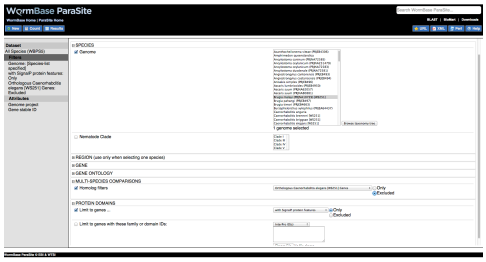
"I need the sequences for a set of *Schistosoma mansoni* genes. I have the chromosome, start, and stop for each."

- 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("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    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
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/:species/: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-phyloxml%2Bxml

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

```
<?xml version="1.0" encoding="UTF-8"?>
<phyloxml xsi:schemaLocation="http://www.phyloxml.org http://www.phyloxml.org/1.10/phyloxml.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.phyloxml.org">
  <phylogeny rooted="true" type="gene tree">
    <clade branch_length="0">
      <confidence type="duplication_confidence_score">0.7311</confidence>
      <taxonomy>
        <id>33208</id>
        <scientific_name>Metazoa</scientific_name>
      </taxonomy>
      <events>
        <type>speciation_or_duplication</type>
        <duplications>1</duplications>
      </events>
      <clade branch_length="0.003861">
        <confidence type="duplication_confidence_score">0.1584</confidence>
        <taxonomy>
```

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

```
use strict;
use warnings;

use HTTP::Tiny;

my $http = HTTP::Tiny->new();

my $server = 'http://parasite.wormbase.org';
my $ext = '/genetree/member/symbol/brugia_malayi_prjna10729/Bma-unc-1';
my $response = $http->get($server.$ext, {
  headers => { 'Content-type' => 'text/x-phyloxml+xml' }
});

die "Failed!\n" unless $response->{success};


print "$response->{status} $response->{reason}\n";

```