WormBase ParaSite

On-going development of an open access helminth genomics resource

Bruce Bolt
bbolt@ebi.ac.uk

WormBase ParaSite sanger EMBL-EBI

Overview

- Introduction to WormBase ParaSite
- Website interface update
- New and updated data
- Features and tools
- Q&A session

Poster #14
Monday evening

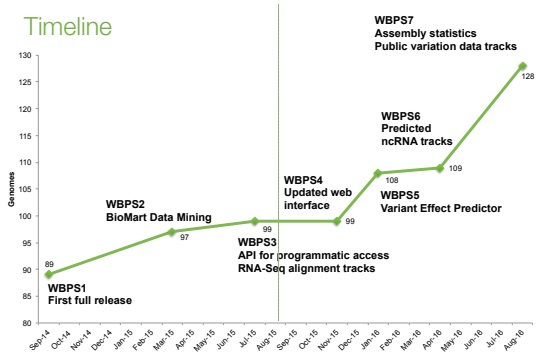
WormBase ParaSite

Introduction to WormBase ParaSite

- Portal to collect and display genomes of helminths (both nematodes and platyhelminthes)
- Integrated with other genomic databases
- Data annotated (mainly) using computational pipelines
- Updated approximately three times per year
- Community driven resource (new data and feature requests are always welcome!)
- Free and open-access at **parasite.wormbase.org**

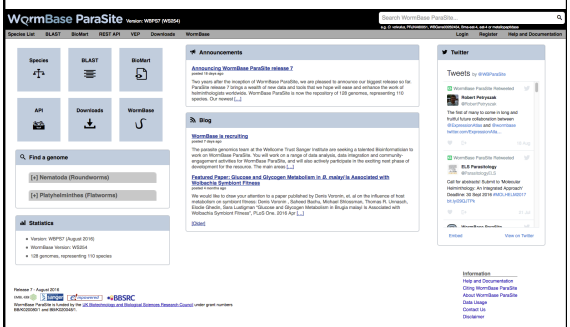
WormBase ParaSite

Timeline



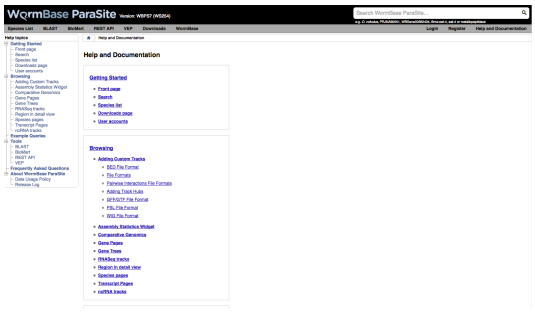
WormBase ParaSite

WormBase ParaSite Release 7



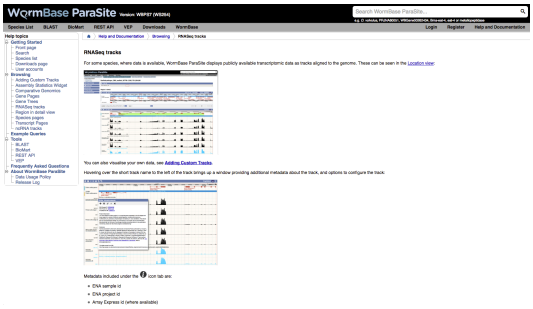
WormBase ParaSite

Help & Documentation



WormBase ParaSite

Help & Documentation



WormBase ParaSite

New and updated genomes

- 21 new genomes:
  - 16 new *Trichinella* genomes
  - Alternate assemblies for: *Echinococcus granulosus*, *Onchocerca ochengi*, *Toxocara canis* and *Wuchereria bancrofti*
  - Free-living flatworm *Macrostomum lignano*
- Updated genomes:
  - New *Brugia malayi* assembly and annotation
  - Updated annotations: *Echinococcus multilocularis*, *Hymenolepis microstoma* and *Onchocerca ochengi*

WormBase ParaSite

Features and tools

- Comparative Genomics
- Quality Statistics
- BLAST
- Variant Effect Predictor (VEP)
- BioMart
- Genome Browser Tracks (RNA-Seq, ncRNA, Variation)
- Programmatic Access (REST API)

- Red = not covered in this session, speak to us at poster

WormBase ParaSite

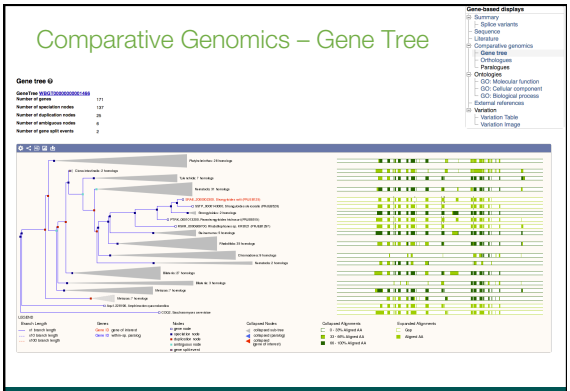
Comparative Genomics

- All genes are aligned against all others
- Gene family trees generated
- Orthologues/paralogues predicted
- Use this data to produce gene coverage and quality statistics

- Re-calculated at each release to incorporate new or updated genomes

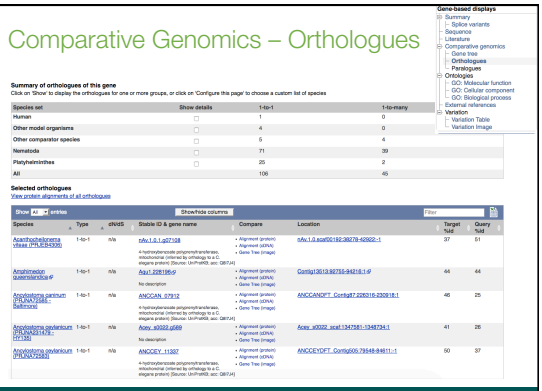
WormBase ParaSite

Comparative Genomics – Gene Tree



WormBase ParaSite

Comparative Genomics – Orthologues



WormBase ParaSite

Frequently asked question

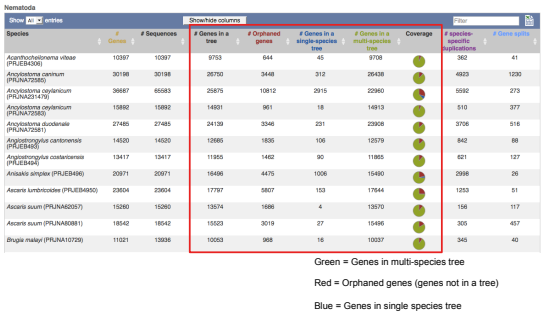
"How do I know whether a genome is high or low quality?"

We now approach this in two ways:

- Statistics based on comparative genomics data
- BUSCO and CEGMA scores for gene set coverage

WormBase ParaSite

Comparative Genomics – Summary Statistics



WormBase ParaSite

Comparative Genomics – Summary Statistics

Yellow = Speciation (orthologues)

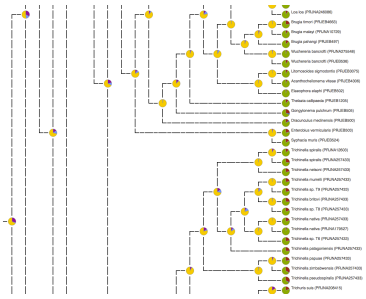
Purple = Duplications (paralogues)

Blue = Dubious

Green = Genes in multi-species tree

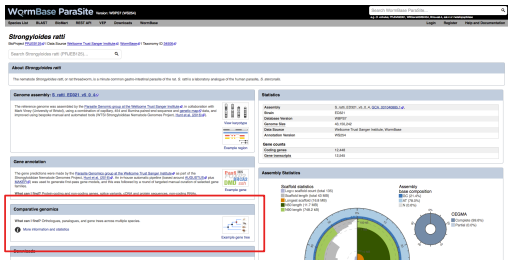
Red = Orphaned genes (tree of their own)

Blue = Genes in single species tree



WormBase ParaSite

Comparative Genomics – Summary Statistics



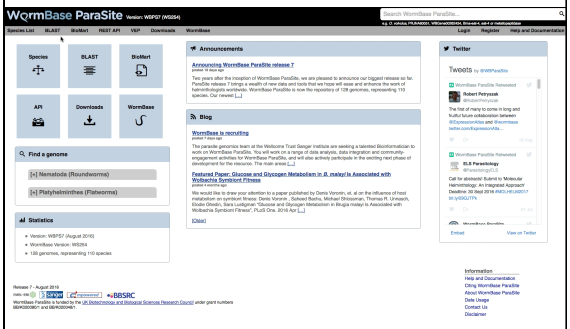
WormBase ParaSite

Assembly quality statistics

- We additionally calculate statistics which show the quality of the assembly and gene set prediction
- CEGMA and BUSCO scores calculated for each genome:
  - CEGMA looks for a set of highly conserved genes found in most eukaryotes
  - BUSCO looks for single-copy orthologues found in more than 90% of animal species
  - In both cases, higher score generally represents a higher quality genome assembly and gene set annotation

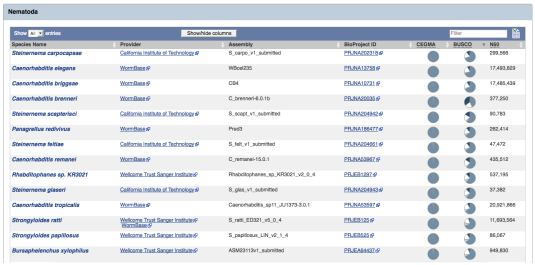
WormBase ParaSite

Quality metrics



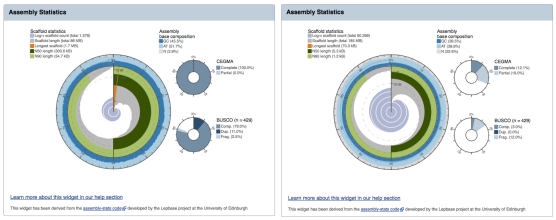
WormBase ParaSite

Quality metrics



WormBase ParaSite

Assembly quality widget



Adapted from code developed by Richard Challis for use in LepBase

WormBase ParaSite

Variant Effect Predictor (VEP)

- Determine the effect of your variants:
  - SNPs
  - Insertions
  - Deletions
  - CNVs
  - Structural variants

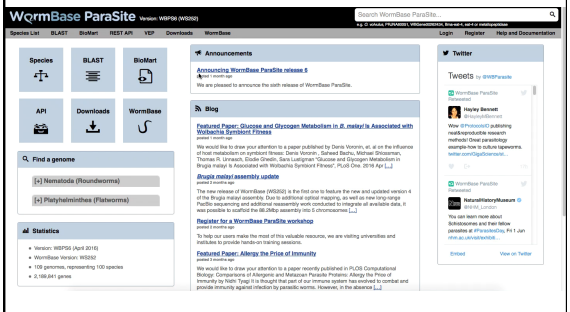
WormBase ParaSite

Variant Effect Predictor (VEP)

- Input:
  - Coordinates of the variant and nucleotide change (or VCF file)
- Output:
  - Genes and transcripts affected by the variants
  - Location of the variants (e.g. upstream of a transcript, in coding sequence, etc)
  - Consequence of the variant on protein sequence (e.g. stop gained/lost, missense, frameshift, synonymous, non-synonymous, etc)
- View results in table and on genome browser

WormBase ParaSite

Using the VEP



WormBase ParaSite

Display of public variation data

- From release 7, we have begun importing data from the European Variation Archive (EVA)
- Make your data public by depositing in EVA
- Variation track will become available in WormBase ParaSite
- Speak to us for advice before depositing in EVA

European Variation Archive → WormBase ParaSite

ENA European Nucleotide Archive

BioSamples

WormBase ParaSite

BioMart

- Advanced search and data export tool
- Produces tables of data or files containing sequence
- Table contents are entirely customisable
- Ideal for working with, or generating, lists of data

Gene stable ID	Chromosome/scaffold name	Gene start (bp)	Gene end (bp)	Exon ID	Exon region start (bp)	Exon region end (bp)	cDNA coding start	cDNA coding end
A01321	Scaffold1052	8861	10795	A01321.a1	8861	9057	1	197
A01321	Scaffold1052	8861	10795	A01321.a2	9712	9353	198	339
A01321	Scaffold1052	8861	10795	A01321.a3	10646	10795	340	489
A01322	Scaffold1052	19019	19861	A01322.a1	19019	19286	1	268
A01322	Scaffold1052	19019	19861	A01322.a2	19663	19861	249	447
A02773	Scaffold1159	9064	10222	A02773.a1	10184	10222	1	39
A02773	Scaffold1159	9064	10222	A02773.a2	9064	9354	40	330
A03307	Scaffold118	185300	190808	A03307.a1	190768	190808	1	41
A03307	Scaffold118	185300	190808	A03307.a2	190413	190656	42	285
A03307	Scaffold118	185300	190808	A03307.a3	189719	188879	286	445
A03307	Scaffold118	185300	190808	A03307.a4	189038	188160	446	588
A03307	Scaffold118	185300	190808	A03307.a5	187378	187413	589	624
A03307	Scaffold118	185300	190808	A03307.a6	185300	185401	625	726
A03308	Scaffold118	191713	191954	A03308.a1	191817	191954	1	138
A03308	Scaffold118	191713	191954	A03308.a2	191713	191775	139	201
A04991	Scaffold127	310799	318815	A04991.a1	318652	318815	1	164
A04991	Scaffold127	310799	318815	A04991.a2	315979	316050	165	236
A04991	Scaffold127	310799	318815	A04991.a3	314409	314432	237	260
A04991	Scaffold127	310799	318815	A04991.a4	313695	313832	361	498
A04991	Scaffold127	310799	318815	A04991.a5	310799	311017	499	717

WormBase ParaSite

Three simple steps…

- Three simple steps:
  1. Filter the entire database to include only the genes you are interested in (query can be specific or vague)
  2. Choose the data you would like to include in the output file
  3. View or download the results (including direct export to Excel, or CSV for import to R)
- No programming or database knowledge required!

WormBase ParaSite

Data available for export

- Sequences (genomic, cDNA, UTR, flanking, cDNA, peptide)
- Gene IDs, names and descriptions
- Identifiers for data from external databases (e.g. UniProt)
- Gene structure (e.g. exons)
- Protein domains and function (e.g. InterPro, Gene3D, PANTHER etc.)
- Gene ontology terms
- Orthologues and paralogues (in all nematodes, flatworms and a number of non-worm comparators, e.g. human, mouse and rat)

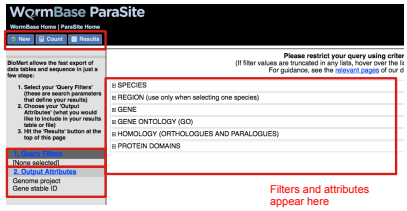
WormBase ParaSite

BioMart Interface

Control buttons:
New = reset form
Count = count results
Results = preview results

Query filters = search terms to restrict the query

Output attributes = select data to add into results (i.e. which columns would you like to appear in your table)



Filters and attributes appear here

WormBase ParaSite

BioMart Example 1 – Working with a list

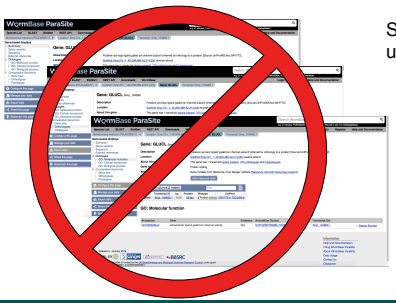
I have a list of *Schistosoma mansoni* genes and would like to find:

1. The gene name and gene description
2. Which of these genes have orthologues in human and mouse
3. The Pfam protein domains (and their positions) for the genes with a human and mouse orthologue

	A
1	Smp_158080
2	Smp_078570
3	Smp_063300
4	Smp_204760
5	Smp_145060
6	Smp_135070
7	Smp_210640
8	Smp_160900
9	Smp_049930
10	Smp_126600
11	Smp_132740
12	Smp_139350
13	Smp_055760
14	Smp_141410
15	Smp_051410
16	Smp_175210
17	Smp_169250
18	Smp_129000
19	Smp_128010
20	Smp_079640
21	Smp_038870
22	Smp_213140
23	Smp_161540
24	Smp_151280
25	Smp_012010
26	Smp_181360
27	Smp_010260
28	Smp_199660

WormBase ParaSite

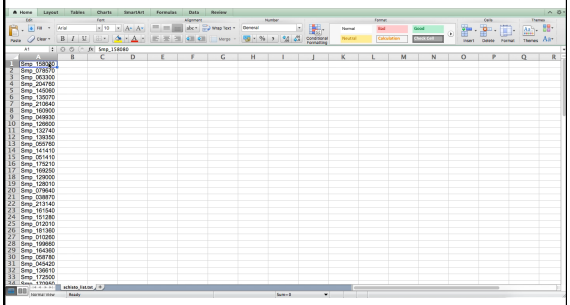
BioMart Example 1 – Working with a list



Save time by using BioMart!

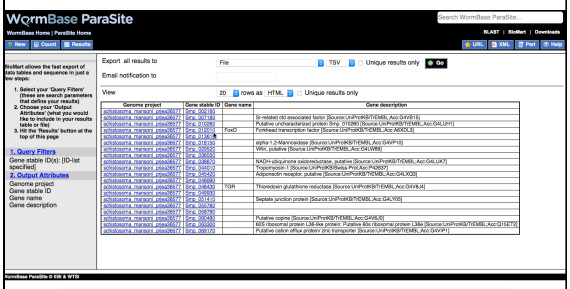
WormBase ParaSite

I have a list of *Schistosoma mansoni* genes and would like to find:
1. The gene name and gene description



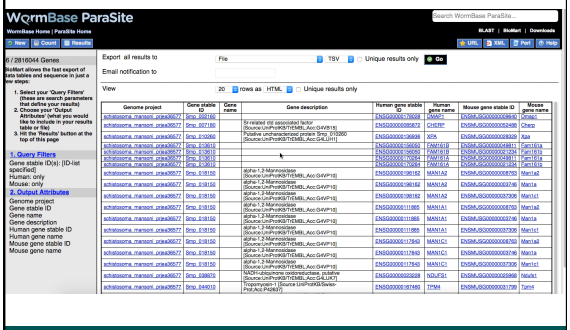
WormBase ParaSite

I have a list of *Schistosoma mansoni* genes and would like to find:
2. Which of these genes have an orthologue in human and mouse



WormBase ParaSite

I have a list of *Schistosoma mansoni* genes and would like to find:
3. The Pfam protein domains (and their positions) for the genes with a human and mouse orthologue



WormBase ParaSite

BioMart Example 2 – Generating a list

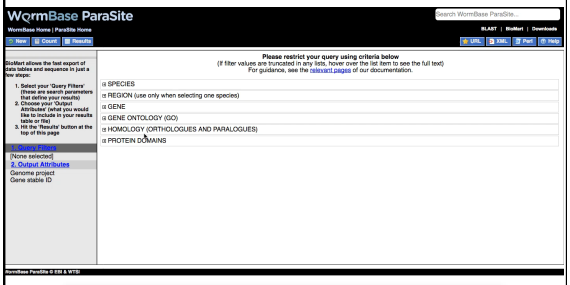
I think a novel drug targets transmembrane signalling receptor activity in *Brugia* species. I would like to model this in *C. elegans*.

Therefore I want to generate a list of *Brugia* genes, which:

- Have an orthologue in *C. elegans*
- Do not have an orthologue in humans
- Are associated with transmembrane signalling receptor activity

WormBase ParaSite

I want to generate a list of *Brugia* genes, which have an orthologue in *C. elegans*,
do not have an orthologue in humans and are associated with transmembrane signalling receptor activity



WormBase ParaSite

BioMart Summary

- Use when working with lists of data, for generating lists of genes and for retrieving sequence
- All data from the website is available
- No programming knowledge required
- Export directly to Excel or CSV for import to R
- Contact us if you are unsure how to construct a query (contact link at bottom of website)

WormBase ParaSite

Contacting us

- Please get in touch with us to report:
  - Feature and data requests
  - Bug reports
  - Questions and comments
- E-mail: link at bottom of website
- Twitter: @WBParaSite

WormBase ParaSite

Acknowledgements

- WormBase ParaSite
  - Kevin Howe
  - Myriam Shafie
  - Jane Lomax
  - Michael Paulini
  - Matt Berriman
  - Paul Kersey
- Ensembl and Ensembl Genomes
- Data Providers
- WormBase Funding: BBSRC, MRC, NHGRI

WormBase ParaSite