

WormBase ParaSite

On-going development of an open access helminth genomics resource



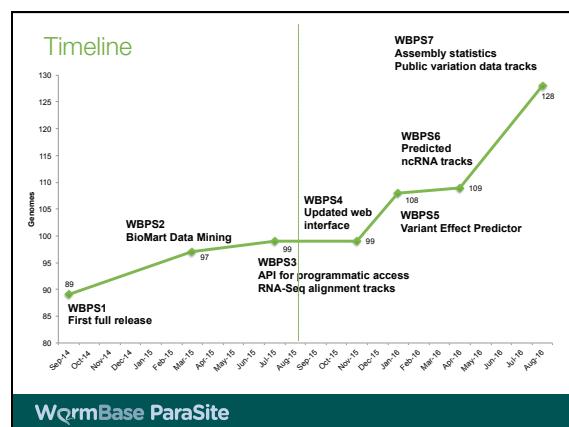
Overview

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



Poster #14
Monday evening

- 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



Features and tools

- Comparative Genomics
- Quality Statistics
- BLAST
- Variant Effect Predictor (VEP)
- BioMart
- Genome Browser Tracks (RNA-Seq, ncRNA, Variation)
- Programmatic Access (REST API)
- Bed = not covered in this session, speak to us at poster

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

- 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

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 – 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

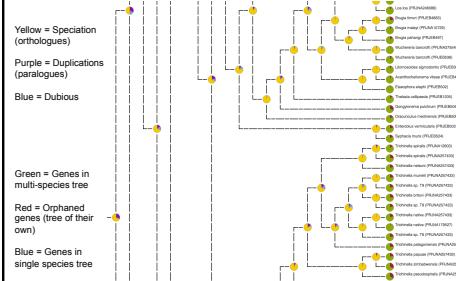
Comparative Genomics – Summary Statistics

Nematoda	Show □	Entries	Genes	Sequences	Show hide columns		# genes in a species tree	# genes in multi-species tree	# genes in a multi-species tree	# genes in a single-species tree	Coverage	# species in multi-species tree	# genes in multi-species tree
					Genes	Sequences							
<i>Ascaris suum</i> (PRJEB400)	10297	10297	9753	644	46	9708	469	41	493	1230	493	41	493
<i>Ascaris suum</i> (PRJNA2089)	30188	30188	26750	3448	312	25408	5992	273	5992	273	5992	273	5992
<i>Ascaris suum</i> (PRJNA2089)	36687	66583	26879	10812	2915	22980	510	377	510	377	510	377	510
<i>Ascaris suum</i> (PRJEB400)	15862	15862	14801	961	18	14913	3706	518	3706	518	3706	518	3706
<i>Ascaris suum</i> (PRJEB400)	27485	27485	24139	3346	231	23928	621	127	621	127	621	127	621
<i>Ascaris suum</i> (PRJEB400)	14030	14030	12665	1635	106	12779	842	88	842	88	842	88	842
<i>Ascaris suum</i> (PRJEB400)	13417	13417	11955	1462	90	11885	2998	26	2998	26	2998	26	2998
<i>Ascaris suum</i> (PRJEB400)	20971	20971	16469	4475	1006	15490	1253	51	1253	51	1253	51	1253
<i>Ascaris suum</i> (PRJNA2089)	23804	23804	17797	587	153	17544	156	117	156	117	156	117	156
<i>Ascaris suum</i> (PRJNA2089)	15260	15260	13574	1686	4	13570	305	457	305	457	305	457	305
<i>Brugia malayi</i> (PRJNA10729)	18542	18542	15523	3019	27	15496	349	40	349	40	349	40	349

Green = Genes in multi-species tree
Red = Orphaned genes (genes not in a tree)
Blue = Genes in single species tree

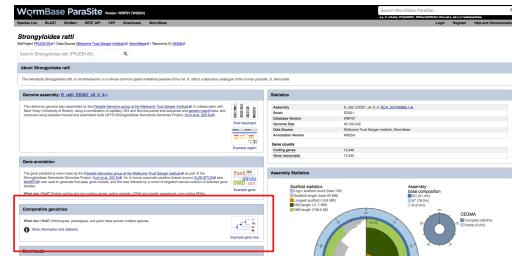
WormBase ParaSite

Comparative Genomics – Summary Statistics



WormBase ParaSite

Comparative Genomics – Summary Statistics



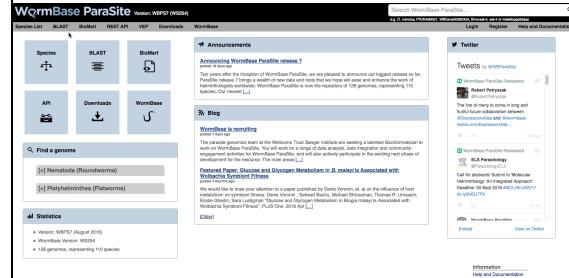
WormBase ParaSite

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						
Data type		Provider	Source column	Bioproject ID	CSDB ID	DISCO ID
Genotype	WormBase	California Institute of Technology	S_cope_v1_submitted	PRJNA202174#	299,595	
Stenotomus carpocapsae	WormBase	Wibed235	PRJNA313259#		17,450,823	
Cannaphthidium elegans	WormBase	CS4	PRJNA10123#		17,486,438	
Cannaphthidium briggsae	WormBase	C. (cannaphthidium) 0.1b	PRJNA200523#		37,350	
Cannaphthidium brenneri	WormBase		PRJNA101245#		90,793	
Stenotomus escherichi	California Institute of Technology	S_cope_v1_submitted	PRJNA202162#		292,414	
Paragnathus endovius	WormBase	Predict	PRJNA186727#		292,414	
Stenotomus fitchii	California Institute of Technology	S_fitch_v1_submitted	PRJNA202149#		47,412	
Cannaphthidium remanei	WormBase	C. remanei 0.1.1	PRJNA202602#		435,512	
Aeloschistus sp. KR9001	Wellcome Trust Sanger Institute	Neurodeposits, NC_KR9001_v2_3.4	PRJNA38267#		537,195	
Stenotomus glaseri	California Institute of Technology	S_gles_v1_submitted	PRJNA202169#		37,382	
Cannaphthidium tropicalis	WormBase		PRJNA202157#		20,371,385	
Worchesia	WormBase		PRJNA202151#		11,803,964	
Strongyloides ratti	Wellcome Trust Sanger Institute	S_ratti_E2001_v3_0.4	PRJNA38251#		86,007	
Strongyloides papillae	Wellcome Trust Sanger Institute	S_papillae_LN_v1_1.4	PRJNA38260#		86,007	
Buaphelopeltis ypsiloniphila	Wellcome Trust Sanger Institute	ADM959191_submitted	PRJNA44621#		649,830	

Assembly Statistics

Graphical statistics

Assembly	Complete	Partial	Unknown
CEGMA	102 (2%)	104 (2%)	894 (96%)
BL2GO	14 (1%)	19 (2%)	867 (97%)

Assembly Statistics

Graphical statistics

Assembly	Complete	Partial	Unknown
CEGMA	130 (2%)	10 (1%)	960 (97%)
BL2GO	14 (1%)	19 (2%)	867 (97%)

Learn more about this widget in our help section

This widget has been derived from the [assembly_stats.widget](#) developed by the LepBase project at the University of Edinburgh

Learn more about this widget in our help section

This widget has been derived from the [assembly_stats.widget](#) developed by the LepBase project at the University of Edinburgh

Variant Effect Predictor (VEP)

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

- 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

Using the VEP

WormBase ParaSite Version: WormPf35 (WS18)

Search WormBase ParaSite
e.g. `GWAS`, `transcript`, `gene`, `allele`, `variant`

Log in Register Help and Documentation

Twitter

Announcements

Announcing WormBase Pf35 release 5

We are pleased to announce the sixth release of WormBase Pf35.

Blog

Featured Paper: Glucose and Glycogen Metabolism in *C. elegans* is Associated with Wormbase Symbiont Fitness

We are pleased to draw your attention to a paper published by Daniel Winters et. al. in the *Journal of Biological Chemistry* (Volume 294, Issue 12, April 2019, Pages 4433-4443). The paper describes the relationship between *C. elegans* fitness and the presence of the *Wolbachia* symbiont. The *Wolbachia* genome is associated with Wormbase Symbiont Fitness. Plus! One. 2019 Apr 1; 2019:101.

Brugia malayi assembly update

The new release of WormBase (WS18) is the first to include the new and updated version 4 of the *Brugia malayi* genome assembly. The new assembly includes new long-range BAC sequencing and additional genome-wide constraints to integrate data available at this time. It was developed by the *Brugia malayi* community and is available here.

Register for a WormBase Pf35 workshop

Are you or your team using the most of this valuable resource, we are inviting universities and institutions to register for a WormBase Pf35 workshop.

Featured Paper: Allergy to the Prion of Immunity

We would like to draw your attention to a paper recently published in *PLOS Computational Biology*: Generation of Allergens and Majorana Protein Peptides: Allergy to the Prion of Immunity. The paper describes the generation of majorana protein peptides and their potential and possible immune response by parasitic worms. However, in the absence of...

Statistics

- WormBase (2018-04-20)
- Wormbase release: WS18
- 109 genomes, representing 100 species
- 2,168,641 genes

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

The diagram illustrates the data flow from the European Variation Archive (EVA) to WormBase ParaSite, ENA (European Nucleotide Archive), and BioSamples. On the left, the European Variation Archive logo is shown with a green arrow pointing to a central hub. From this hub, three arrows branch out to the WormBase ParaSite logo, the ENA logo, and the BioSamples logo.

- 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!

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)

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)

WormBase ParaSite

Member Home / Parallel Home

Search **Create** **Results**

Species allows the fast export of data to a selected species in just a few steps:

1. Select your "Query Filter" (these are search parameters that will be used for every query)
2. Choose your "Output" (this is the data that you would like to include in your results)
3. Hit the "Results" button at the bottom of the page

Query Filter

None selected

WormBase

WormBase Parallel

WormBase Genome project

Gene stable ID

Output

None selected

WormBase

WormBase Parallel

WormBase Genome project

Gene stable ID

Please restrict your query using criteria
 (If filter values are truncated in any lists, hover over the list for guidance, see the [relevant pages](#) of our documentation)

eSPECIES
 REGION (use only when selecting one species)

GENE

GENE ONTOLOGY (OID)

HOMOLOGY (ORTHOLOGUES AND PARALOGUES)

PROTEIN DOMAINS

Filters and attributes appear here

A
1 Smp_010000
2 Smp_010570
3 Smp_063300
4 Smp_204760
5 Smp_145060
6 Smp_107000
7 Smp_210640
8 Smp_169000
9 Smp_049830
10 Smp_100000
11 Smp_132740
12 Smp_139350
13 Smp_055760
14 Smp_100000
15 Smp_261410
16 Smp_175210
17 Smp_169250
18 Smp_128000
19 Smp_176100
20 Smp_079640
21 Smp_038670
22 Smp_213140
23 Smp_155540
24 Smp_151280
25 Smp_012010
26 Smp_181360
27 Smp_010380
28 Smp_199690

Save time by using BioMart!

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