**WormBase ParaSite Workshop**
**9th March 2016**

**Venue:**
Room 4325D, James Clerk Maxwell Building
University of Edinburgh, EH9 3FD

**Instructors:**
Bruce Bolt (bbolt@ebi.ac.uk)
Jane Lomax (jl16@sanger.ac.uk)

**12:45 – 13:00**
Arrival & Registration

**13:00 – 13:15 (Bruce)**
Introduction to WormBase ParaSite

**13:15 – 13:45 (Jane)**
Using the website 1: Locating genome of interest, genome browser, attaching files and viewing RNA-Seq tracks
Practical exercises part 1 [20 mins]

**13:45 - 14:15 (Bruce)**
Using the website 2: Searching, comparative genomics and user accounts
Practical exercises part 2 [20 mins]

**14:15 – 15:00 (Bruce)**
Sequence searching with BLAST
Practical exercises [35 mins]

**15:00 - 15:30**
Coffee Break

**15:30 – 16:30 (Jane)**
Data mining with BioMart
Practical exercises part 1 [25 mins]
Practical exercises part 2 [20 mins]

**16:30 – 16:45 (Bruce)**
Bulk downloads and programmatic access (including access with R)
Note: this section is just an overview of the services we have available, no prior programming or coding knowledge is required

# Contents



Please note: this booklet has been compiled based on release 5 (January 2016) of WormBase ParaSite. Should any features or tools change in future releases, we will endeavour to keep our online documentation up-to-date: http://parasite.wormbase.org/info/.

After the workshop, solutions to exercises will be made available on YouTube. These are accessible using the link located at http://parasite.wormbase.org/workshop.

# Introduction

WormBase-ParaSite is a sub-portal of WormBase dedicated to parasitic worms (helminths). It encompasses both the nematodes (roundworms) and platyhelminthes (flatworms).

Release 5 (January 2016) contains 108 genomes, representing 99 species. Some species have been sequenced by multiple groups, at varying levels of coverage and quality, which gives rise to the discrepancy between genomes and species.

## Social Media

Twitter: twitter.com/wbparasite
Blog: wbparasite.wordpress.com

If you would like to list an upcoming meeting on our blog or Twitter, please let us know by emailing parasite-help@sanger.ac.uk.

## Citing WormBase ParaSite

Where you have used the data or tools provided by WormBase, you are asked to cite the following article:

Kevin L. Howe, Bruce J. Bolt, Scott Cain, Juancarlos Chan, Wen J. Chen, Paul Davis, James Done, Thomas Down, Sibyl Gao, Christian Grove, Todd W. Harris, Ranjana Kishore, Raymond Lee, Jane Lomax, Yuling Li, Hans-Michael Muller, Cecilia Nakamura, Paulo Nuin, Michael Paulini, Daniela Raciti, Gary Schindelman, Eleanor Stanley, Mary Ann Tuli, Kimberly Van Auken, Daniel Wang, Xiaodong Wang, Gary Williams, Adam Wright, Karen Yook, Matthew Berriman, Paul Kersey, Tim Schedl, Lincoln Stein, and Paul W. Sternberg.
**WormBase 2016: expanding to enable helminth genomic research**
*Nucleic Acids Research* 2016 44(D1) D774-D780

Additionally, you must credit the authors of the genome(s) you use in your work. Where a genome has been published, a link to this publication is available on the WormBase ParaSite website. For unpublished genomes, please contact the principal investigator of the relevant sequencing project before undertaking any genome wide analyses.

## Funding

WormBase ParaSite is funded by the UK Biotechnology and Biological Sciences Research Council. WormBase is funded by the US National Institutes of Health and the UK Medical Research Council.

# Browsing

## Practical Exercises Part 1

1. Navigate to the page for *Trichuris suis* (Hint: it is a Clade I nematode)
    a. How many genomes for this species are in WormBase ParaSite?
    b. Select the Washington assembly (PRJNA179528)
    c. How many genes have been predicted in this genome?
    d. What is the length of the genome?

2. Navigate to the region 16,599,624 to 16,631,134 on *Onchocerca volvulus* scaffold OVOC_OM1b
    a. How many genes are predicted within this region?
    b. Zoom in on gene OVOC2189
    c. What are the genomic coordinates of this gene?
    d. Create a 'share link' for this display
    e. Export the sequence of the region you are viewing in FASTA format

3. Scroll down the page you are on to see the RNASeq tracks aligned to this sequence
    a. How many studies are being displayed for this species?
    b. Identify the study ID and follow the link to see the ENA project page
    c. Locate the configuration for this page and turn OFF visualization of study ERP001350
    d. Identify the publication for study SRP056861 and navigate to the full text.

4. Navigate to any *Trichuris muris* genome browser page
    a. Open up the 'Add your Data' window
    b. Add the BigWig file located at: http://www.ebi.ac.uk/~jane/testdata/ERS092077.bw
    c. Share the track by generating a unique URL
    d. Navigate to gene TMUE_s0016004100. How would you judge the existing gene model? (Hint: the 'Region in Detail' or 'location view' allows zooming)
    e. In the 'Configure this page' menu, turn on 'translated sequence' and 'start/stop codons'. Does this data confirm your opinions about the gene model? (Hint: you need to be on maximum zoom to view the translated sequence).

# Practical Exercises Part 2

1. Locate the gene SVE_1227300
    a. In which species is this gene found?
    b. What is the length of the protein product of this gene?
    c. How many Gene Ontology (GO) terms are assigned to this gene?

2. Move onto the transcript view for the single transcript of SVE_1227300
    a. How many exons does the transcript of this gene have?
    b. Which Pfam domain has been assigned to the protein product of this gene?

3. Navigate to the gene page for the *Necator americanus* gene NECAME_00069.
    a. How many orthologues have been predicted in flatworm genomes?
    b. Which species has an orthologue with the highest percentage identity?
    c. View the alignment between this gene's protein product and the protein of the *Ancylostoma duodenale* orthologue.

4. Paralogues are also predicted. These are caused by duplication events. How many paralogues are predicted for the *Necator americanus* gene NECAME_00069? Look at the percent identity for this alignment - would you call this as a paralogue?

5. Locate the *Fasciola hepatica* (PRJNA179522) orthologue of the human gene BRCA2. Using the gene trees:
    a. How close in evolutionary history is this gene located to its orthologue?
    b. Are there any duplication events in the evolution of this gene and its homologues?

    e. Click on the 'Register' link in the horizontal toolbar under the search box

    a. Enter your details and create an account (if you don't want to create an account, skip to c. and use the test account: email: wormbase.test@gmail.com password: W0rmbase)
    b. Verify your account via your email address
    c. Log in to your account via the 'Login' link
    d. Go to 'Manage your data' and save the data track you added to your account (Hint: use the floppy disk icon under 'Actions')
    e. Try logging out and logging back in again. Does your data track persist?

# BLAST

## 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 |

## 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.

## Practical Exercises

1. Locate the peptide sequence for the *Brugia malayi* gene Bma-eat-4. Using BLAST, find which other nematode(s) this peptide sequence occurs in at 100% similarity?

2. Locate the cDNA sequence for the *Clonorchis sinensis* gene csin111107. Using BLAST, find which other genomes the six-frame translation of this sequence occurs in with a %ID of more than 90%. Which BLAST tool is most appropriate here?

3. Use the ‘Edit’ button to populate the input form with the sequence from (2). There are many advanced options that can be set. In most cases, these can be left at the default values. Decrease the maximum number of alignments displayed to 5 and run the query. Why are there more than five results in the table?

4. Find a gene of choice from your favourite species. BLAST the sequence of the first exon against the database of all helminth transcriptomes.

# Data-mining with BioMart

## Practical Exercises Part 1

1. Setting Filters:
    a. In ‘Filters’, select ‘Brugia malayi’ from the SPECIES menu.
    b. In the GENE ONTOLOGY menu, enter ‘potassium channel activity (GO:0005267)’ under ‘GO term(s)’.
    c. In the MULTI-SPECIES COMPARISONS menu, select ‘Orthologous human genes’ and check ‘Excluded’.
    d. Hit ‘Results’. You should get a gene list for of *Brugia malayi* K+ channels without a human ortholog

2. BioMart is a very handy tool for converting between IDs from different databases:

    a. Click ‘New’
    b. In the GENE menu, enter the following *Schistosoma mansoni* ids: `Smp_198150, Smp_202490, Smp_008770, Smp_005010, Smp_142730`
    c. Now switch to ‘Attributes’ in the left menu
    d. Under the EXTERNAL menu select ‘EntrezGene ID’
    e. Hit ‘Results’. You should get your gene list with the corresponding EntrezGene IDs.

3. BioMart can also return sequence, instead of a gene-based results table:

    a. In Filters, enter the gene list above in the GENE menu.
    b. Switch to ‘Attributes’ in the left menu.
    c. Check ‘Sequences’
    d. In the SEQUENCES menu, select ‘Unspliced (Transcript)’
    e. Hit ‘Results’.

4. BioMart does not require a species to be specified: it is possible to mine the data in a “species neutral” way.
    a. In the SPECIES menu, select ‘Nematode Clade’, ‘Clade I’
    b. In the PROTEIN DOMAINS menu, enter InterPro ID ‘IPR001806’ (Small GTPase superfamily)
    c. Hit ‘Results’.

# Practical Exercises Part 2

1. For the following list of *Onchocerca volvulus* gene IDs, generate a list showing the WormBase ParaSite gene ID (stable ID), gene name and the RefSeq Protein ID.
   `WBGene00249437, WBGene00249438, WBGene00249440, WBGene00249441, WBGene00249442, WBGene00249445, WBGene00249446, WBGene00249820, WBGene00249821`

2. Using the same list of genes as in (1), generate a list showing the WormBase ParaSite gene ID (stable ID), *O. volvulus* gene name, *C. elegans* orthologue gene name and human orthologue gene name. How many of these genes have an orthologue defined in both *C. elegans* and human?

3. Using the same list of *O. volvulus* genes in (1), get the sequence for the region 500bp upstream of each gene. Export this as a FASTA file.

4. Create a list of genes, from all the nematodes, which have a predicted signal peptide (SignalP) domain, do not have a human orthologue and are associated with 'transmembrane signaling receptor activity (GO:0004888)'.

5. How many genes in *Brugia malayi* have been annotated as non-coding RNA (ncRNA) genes?