WormBase ParaSite Workshop
20th June 2016

Venue:
Room B23, Llandinam Building, Penglais Campus
Aberystwyth University, SY23 3DB

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

10:00 - 10:15
Arrival & Registration

10:15 - 10:30 (Jane)
Introduction to WormBase ParaSite

10:30 - 12:00 (Jane)
Using the website
Practical exercises

12:00 - 13:00
Lunch Break

13:00 - 13:45 (Bruce)
Sequence searching with BLAST
Practical exercises

13:45 - 14:30 (Bruce)
Data export with BioMart (Part 1)
Practical exercises

14:30 - 14:50
Tea & Coffee Break

14:50 - 15:20 (Jane)
Data export with BioMart (Part 2)
Practical exercises

15:20 - 16:00 (Bruce)
Variant Effect Predictor
Practical exercises

16:00 - 16:30
Q&A Session

Contents

Please note: this booklet has been compiled based on release 6 (April 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 6 (April 2016) contains 109 genomes, representing 100 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.

Release 7 (due August 2016) will have 128 genomes, representing 110 species. This release will include new variation tracks for some species, gene expression data and a bulk import of new *Trichinella* genomes.

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.

Using the website

Practical Exercises

1. Navigate to the page for *Schistosoma mansoni*
    a. How many coding genes have been predicted in this genome?
    b. What is the length of the genome?
    c. Which institute sequenced this genome?

2. Navigate to gene OVOC2189 from *Onchocerca volvulus*, then click on the 'Region in detail' link to get to the interactive browser page
    a. What are the genomic coordinates of OVOC2189?
    b. Create a 'share link' for this display
    c. Zoom out in the lower browser so that you can see more than one gene
    d. Export the sequence of the region you are viewing in FASTA format (hint: look for the 'Export data' button in the sidebar)

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? (Hint: studies are shown in different colours)
    b. Identify the study ID for one of the studies and follow the link to see the ENA project page
    c. Locate the configuration for this page and turn OFF visualization of study ERP001350 (Hint: look for the 'Configure this page' option in the sidebar)
    d. Identify the publication for study SRP056861 and navigate to the full text.

4. Navigate to the *Trichuris muris* genome page, and click on the 'Example region' link in the Genome assembly information box:

    a. Open up the 'Add your Data' window by clicking the link in the sidebar
    b. Attach one of the BigWig files located at: http://www.ebi.ac.uk/~jane/testdata/ by pasting the URL in to the Data box (Hint: to copy the URL, right-click file name and 'Copy link address')
    c. Navigate to gene TMUE_s0016004100 and have a look at the RNASeq track. How would you judge the existing gene model? (Hint: go to the 'Region in Detail' view to see the tracks and zoom in)

5. 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?

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

7. 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.

8. 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?

9. 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?

10. 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.

Advanced search and data export with BioMart

Practical Exercises Part 1

For the first set of exercises, we will look at some worked examples to create customised data tables.

1. Create a list of *Brugia malayi* genes which are associated with potassium channel activity:
    a. In ‘Query 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. Hit ‘Results’ and you will see a preview of a gene list for of *Brugia malayi* K+ channels

2. BIoMart is very useful for quickly converting between IDs from different databases. Using a short list of *Schistosoma mansoni* gene IDs, get the matching UniProt database entry:
    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 ‘Output Attributes’ in the left menu
    d. Under the EXTERNAL DATABASE REFERENCES AND ID CONVERSION menu select ‘UniProtKB/SwissProt ID’ and ‘UniProtKB/TrEMBL ID’
    e. Hit ‘Results’

3. BioMart does not require a species to be specified: it is possible to mine the data in a “species neutral” way. In this example, create a list of all Clade I nematode genes that have a small GTPase superfamily domain:
    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’

It is also possible to use BioMart to return sequence, instead of data tables. In these examples, we will look at the filters and attributes that must be set to create sequence files.

4. Produce a file containing the 5’ UTR sequence of each *Brugia malayi* gene:
    a. In ‘Query Filters’, select ‘Brugia malayi’ from the SPECIES menu
    b. Switch to ‘Output Attributes’ in the left menu
    c. Check ‘Retrieve Sequences’
    d. In the SEQUENCES menu, select ‘5’ UTR’
    e. Hit ‘Results’

5. Retrieve a file containing the sequence for all *Clonorchis sinensis* proteins. Annotate this file with the gene ID, protein ID and gene name:
    a. In ‘Query Filters’, select ‘Clonorchis sinensis’ from the SPECIES menu
    b. Switch to ‘Output Attributes’ in the left menu
    c. Check ‘Retrieve Sequences’

d. In the SEQUENCES menu, select ‘Peptide’
e. In the HEADER INFORMATION menu, select ‘Gene Stable ID’, ‘Protein Stable ID’ and ‘Gene Name’
f. Hit ‘Results’
g. Export this to a downloadable file using the ‘Export all results to…’ buttons at the top of the results preview

Following the coffee break, we will take a look at some real-life scenarios of where you may wish to use BioMart.

Practical Exercises Part 2

In these exercises, we will look at some more advanced examples of BioMart usage, demonstrating that complex data mining tasks can be completed with little effort and in minimal time.

Scenario:

You have been doing an experiment to look at differential gene expression for different life-cycle stages in *Onchocerca volvulus*, and have come up with a short list of gene IDs that you want to know more about. The list is here: http://www.ebi.ac.uk/~bbolt/aber/onc.html

1. Using BioMart and your gene list, generate a table that contains: i. WormBase ParaSite gene ID (stable ID), ii. gene name and iii. RefSeq Protein ID.

You decide to further investigate these genes as potential drug targets. You need to find out whether they have any orthologs in a model system (*C. elegans*) or in humans where they might cause adverse effects:

2. Using BioMart, generate a table showing i. the WormBase ParaSite gene ID (stable ID), ii. *O. volvulus* gene name, iii. *C. elegans* orthologue gene stable ID and iv. human orthologue gene stable ID. How many of these genes have an orthologue defined in both *C. elegans* and human?

In some follow-up work, you are interested in looking for miRNA binding sites in the 5’ UTRs of your genes. To do this you need to download the sequences of your genes, with some upstream sequence as input for your analysis program:

3. Using BioMart, get the sequence for the region 500bp upstream of each gene in your list. Export this as a FASTA file.

Finally, you’re interested in finding out which protein domains your genes encode.

4. Retrieve a table that contains: i. WormBase ParaSite gene ID (stable ID), ii. gene name and gene description, iii. InterPro ID and iv. short InterPro description.

Variant Effect Predictor (VEP)

Practical Exercises

Output from variant calling is often provided in VCF files. These can be compressed (gzipped) to reduce the size. The WormBase ParaSite VEP can handle both compressed and uncompressed files. When working with files on the website, there is a 50MB file size limit, therefore compressed files are suggested. To work with larger files, it is highly recommended to run the VEP locally. Pre-computed "VEP caches" can be downloaded from our FTP site.

For these exercises, we have provided a small file containing 499 variant calls for *Schistosoma mansoni* at `http://www.ebi.ac.uk/~bbolt/aber/mansoni.vcf.gz`

1. Download this file to your computer. Navigate to the VEP (there is a link in the main website header bar). Upload the file and submit the VEP job.

2. View the results of this job:
    a. How many genes are overlapped by these variations?
    b. How many of the variants would result in an upstream gene variation?
    c. Of the coding consequences, how many would give a mis-sense mutation?

3. Now we will look at the variants that have consequences on a specific gene: Smp_160490. In the 'Filters' section, select 'Gene' 'is' 'Smp_160490' then press 'Add'. The results table will update to only show variants that have consequences on this gene. How many variants affect this gene?