WormBase ParaSite – a comprehensive resource for helminth genomics

Kevin L. Howe a,*, Bruce J. Bolt b, Myriam Shafie b, Paul Kersey a, Matthew Berriman b

a European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
b Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

ARTICLE INFO

Article history:
Received 7 October 2016
Received in revised form 24 November 2016
Accepted 25 November 2016
Available online xxx

Keywords:
Genome browser
Comparative genomics
Functional genomics
Helminths
WormBase
Ensembl

ABSTRACT

The number of publicly available parasitic worm genome sequences has increased dramatically in the past three years, and research interest in helminth functional genomics is now quickly gathering pace in response to the foundation that has been laid by these collective efforts. A systematic approach to the organisation, curation, analysis and presentation of these data is clearly vital for maximising the utility of these data to researchers. We have developed a portal called WormBase ParaSite (http://parasite.wormbase.org) for interrogating helminth genomes on a large scale. Data from over 100 nematode and platyhelminth species are integrated, adding value by way of systematic and consistent functional annotation (e.g. protein domains and Gene Ontology terms), gene expression analysis (e.g. alignment of life-stage specific transcriptome data sets), and comparative analysis (e.g. orthologues and paralogues). We provide several ways of exploring the data, including genome browsers, genome and gene summary pages, text search, sequence search, a query wizard, bulk downloads, and programmatic interfaces. In this review, we provide an overview of the back-end infrastructure and analysis behind WormBase ParaSite, and the displays and tools available to users for interrogating helminth genomic data.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The WormBase project (http://www.wormbase.org, [1]) was initiated to facilitate and accelerate biological research that uses the model nematode Caenorhabditis elegans, by making the collected outputs of scientists accessible from a single resource. This enables the transfer of this wealth of knowledge to the study of other metazoans, from nematodes to humans. C. elegans data described in the research literature, deposited in the archives, or submitted directly is placed into context via a combination of detailed manual curation and semi-automatic data integration. In addition, the project curates the reference genome sequence, gene structures and other genomic features for C. elegans, thereby providing a high quality foundation for downstream studies. The WormBase mission also extends to free-living relatives of C. elegans, which were the focus of early post-genomic research in the areas of evolution and comparative biology [2].

In recent years, WormBase has begun to expand its mission to include plant and animal parasitic nematodes which, while more distantly related to C. elegans, have direct biomedical and agricultural importance and therefore attract research interest in their own right. The first parasitic nematode to have its genome fully sequenced was Brugia malayi [3], and since then, many have followed [4–37]. Of the WormBase “core” species (the ones for which the reference genome sequence and annotation are curated), three are parasitic worms: Brugia malayi, Onchocerca volvulus and Strongyloides ratti (http://www.wormbase.org/species). However, there are a number of challenges associated with expanding the remit to helminths. Firstly, the primary research goal of parasitologists is to identify ways of controlling the parasite, and as such their desired entry points and common use-cases for WormBase are often distinct from those of scientists doing basic science using C. elegans as a model. Secondly, the wider community of worm parasitologists includes those studying platyhelminths (flatworms), which are outside the taxonomic scope of WormBase, which is a nematode resource. Thirdly, there have been recent concerted efforts to sequence the genomes of many helminths (nematodes and platyhelminths) (e.g. the 50 Helminth Genomes Initiative [38]), resulting in a flood of new draft genomes that vary considerably in contiguity.

In response to these challenges, we have created WormBase ParaSite, a comprehensive new resource for parasitic worm genomes, aiming to serve parasitologists working on helminths who use genomics as an investigatory tool. WormBase ParaSite leverages the
2. Data integration and analysis

2.1. Genomes

Our mission is to include all publicly available nematode and platyhelminth genomes in WormBase ParaSite. Where multiple genome assemblies exist for the same species (e.g., different genome projects for *Haemonomus contortus* [18,19], and male and female isolates for *Trichuris suis* [24]), we have included all. Release 7 of the resource (August 2016) included genomes from 82 nematode species (98 genomes) and 28 platyhelminth species (30 genomes). We maintain an up-to-date list of genomes for ease of reference (http://parasite.wormbase.org/species.html).

Our primary source for genome sequences is the International Nucleotide Sequence Database Collaboration (INSDC) resources [39]. In rare cases, we collect genomes from project-specific FTP sites or direct engagement with genome project scientists. However, it is our strong preference to use genomes deposited with INSDC as these have been formally checked, processed and versioned by an authoritative sequence archive resource. As well as allowing us to formally disambiguate between different genomes for the same species, by way of the INSDC BioProject identifier (http://www.ncbi.nlm.gov/bioproject), it also simplifies the process of integrating and displaying functional genomics data that has been submitted to the archives. For the species in common between WormBase and WormBase ParaSite (C. elegans and other free-living nematodes, *Brugia malayi*, *Strongyloides ratti* and *Onchocerca volvulus*), we synchronise the data with a specific release of WormBase. For example, WormBase ParaSite release 7 was synchronised with WormBase release WS254, meaning that data for species in common were identical in both resources.

2.2. Genome annotation

2.2.1. Gene structures

The definition of the intron/exon structure of protein-coding genes is an important foundation for interpretation of genome function. WormBase curates gene structures for a *C. elegans* and a small set of nematode species with high-quality reference genomes [40]. For others, we import annotations from the acknowledged authority for that genome (e.g. GeneDB [41] for *Schistosoma mansoni*), or from the group that sequenced and published the genome [4–37].

In certain circumstances, we collaborate with genome projects to provide first pass annotation of protein-coding gene structures. We have co-developed a pipeline that produces high-quality gene predictions using MAKER [42] to integrate evidence from multiple sources: *ab initio* gene predictions from AUGUSTUS [43], GeneMark-ES [44], and SNAP [45]; projected annotations from *C. elegans* and the taxonomically nearest previously-annotated helminth using GenBlatG [46] and RATT [47]; and alignments of ESTs, mRNAs and proteins from related organisms. The pipeline was used to annotate the majority of the genomes sequenced as part of the 50 Helminth Genomes Project [38].

A small number of genomes have curated structures for non-coding RNAs, which we also import into WormBase ParaSite. To supplement these, we have a pipeline that uses RNAmer [48], tRNAscan-SE [49] and RFam [50] to predict structures for (respectively) ribosomal RNAs, transfer RNAs, and other non-coding RNAs.

2.2.2. Functional annotation

Literature describing gene function is sparse for most of the species in WormBase ParaSite, although we anticipate that the availability of the genomes will stimulate new research. In lieu of annotations based on experimental evidence, we have used established automated methods to predict the function of as many gene products as possible. Firstly, we broker the submission of all protein sequences in WormBase ParaSite to the UniProt Knowledgebase [51], and import the product names assigned by them. These are defined by a combination of manual curation and automatic annotation. For more detailed annotation, we use InterProScan [52] from the InterPro project [53]. As well as predicting protein domains (e.g. from the Pfam [54] database), it also assigns terms from the Gene Ontology (GO) [55]. We run the latest version of the pipeline and data across the complete proteome for every genome each release, so that we are always up-to-date with the latest InterPro domain and Gene Ontology annotations.

2.2.3. Gene expression

High-throughput sequencing of RNA has become the standard assay for measuring gene expression, and numerous studies conducting “RNA-Seq” experiments in helminth species have now been performed and deposited in the sequence archives. We ultimately aim to align all helminth RNA-Seq data to the corresponding reference genome using a standard pipeline, in collaboration with the Gene Expression Atlas project [56] (see Section 4). In the meantime, we have processed data for a small number of species using our own pipeline, as a pilot study (Table 1). Briefly, we use STAR [57] to align each experiment to the reference genome, merging resulting BAM files when they are technical replicates of the same sample. We have labelled each sample with the appropriate descriptive terms, using ontological terms where possible (e.g. from the WormBase life-stage ontology). The alignments can be viewed on the genome browser (see Section 3.2).

2.3. Comparative genomics

Having genomes, genes and proteins for many helminth species provides the opportunity to study the evolution of helminths at the molecular level. We use the Ensembl Compare system [68] to infer the orthology and paralogy relationships between all nematode and platyhelminth genomes, supplemented by genes from a number of other comparator species, including human, mouse and yeast.

The method can be briefly summarised as follows: genes are first organised into homologous clusters. Traditionally, this has been done using NCBI BLAST+ [69] and hclusterg (H. Li, unpublished) but more recent versions of the system support the use of Hidden Markov Models from the PANTHER protein classification system [70]. A protein multiple alignment for each cluster is then constructed using M-Coffee [71] or MAFFT [72]. The choice of program

### Table 1

<table>
<thead>
<tr>
<th>Species</th>
<th>Studies</th>
<th>Samples (tracks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nematode</td>
<td>Brugia malayi</td>
<td>3 [58,59]</td>
</tr>
<tr>
<td></td>
<td>Onchocerca volvulus</td>
<td>2 [60,61]</td>
</tr>
<tr>
<td></td>
<td>Strongyloides ratti</td>
<td>2 [34]</td>
</tr>
<tr>
<td></td>
<td>Strongyloides stercoralis</td>
<td>2 [62]</td>
</tr>
<tr>
<td></td>
<td>Trichuris muris</td>
<td>1 [25]</td>
</tr>
<tr>
<td>Platyhelminth</td>
<td>Echinococcus granulosus</td>
<td>2 [15]</td>
</tr>
<tr>
<td></td>
<td>Echinococcus multilocularis</td>
<td>2 [15]</td>
</tr>
<tr>
<td></td>
<td>Fasciola hepatica</td>
<td>2 [31]</td>
</tr>
<tr>
<td></td>
<td>Hymenolepis microstoma</td>
<td>2 [15]</td>
</tr>
<tr>
<td></td>
<td>Schistosoma mansoni</td>
<td>6 [63–67]</td>
</tr>
</tbody>
</table>

Please cite this article in press as: K.L. Howe et al., WormBase ParaSite – a comprehensive resource for helminth genomics, Mol Biochem Parasitol (2016). http://dx.doi.org/10.1016/j.molbiopara.2016.11.005
is made dynamically by the pipeline using a set of empirically-derived rules regarding certain properties of the input data (for example, MAFFT is generally favoured for larger clusters). Finally, TreeBest [68] is used to produce a gene tree, combining a number of methods which reconcile the sequence-based gene phylogeny with the species phylogeny. The resulting tree represents an evolutionary history of the gene family, which can be used to infer true orthologues (genes in different species related by a speciation event) and paralogues (genes in the same species related by a duplication event).

2.4. Infrastructure

The Ensembl infrastructure [73] is the basis for much of the management and analysis of the data in WormBase ParaSite. The main components we use are (a) the MySQL database schema and Application Programming Interface for loading and storing the data [74]; (b) the genome analysis pipelines and workflow management tools [75,76]; and (c) the website displays and tools [73]. We have customised some of the Ensembl tools for use in WormBase ParaSite. For example, we have modified Ensembl code to provide sequence search and data mining services that allow the interrogation of all species, or large sub-groups of species (e.g. all nematodes) at once in a single query (see Section 3.3).

3. Website – displays and tools

3.1. General navigation

The most common entry point to WormBase ParaSite is via the home page (http://parasite.wormbase.org), which collects together the latest news from the project, some basic statistics, and a tool for finding genomes of interest. The header bar is present on all pages in WormBase ParaSite. As well as providing short-cut links to a variety of tools and documentation pages (http://parasite.wormbase.org/info), it has a search-box which allows searching for genes by a vari-

Please cite this article in press as: K.L. Howe, et al., WormBase ParaSite — a comprehensive resource for helminth genomics, Mol Biochem Parasitol (2016), http://dx.doi.org/10.1016/j.molbiopara.2016.11.005
Fig. 2. Compare tree showing the evolution of fox-1 in *Trichinella spiralis*, a gene involved in sex determination. In this view, the *Trichinella* sub-tree has been expanded for additional detail. Other parts of the tree are collapsed for brevity. A schematic of a protein multiple alignment of the homologs is shown in the right, with regions of conservation represented as coloured blocks.

ety of criteria, including gene name, product name, protein domain names/accessions, and Gene Ontology terms. The search box has an auto-complete function, allowing searches to be performed using only partial information.

We also maintain a page for each genome and gene in WormBase ParaSite. The genome pages collect together a description of the species, attribution/references for the genome sequencing and annotation, and example entry points to data for that genome (e.g. gene, genomic region). Also shown are some basic statistics about the quality of the genome, for example CEGMA [77] and BUSCO [78] scores. Recent releases have displayed these graphically in an Assembly Statistics panel, using code developed by the LepBase project [79]. We also provide these data as columns in the table of all genomes (http://parasite.wormbase.org/species.html), which can be sorted by each of the criteria.

The gene pages act a starting place for exploring various types of information associated with the genes, including transcript models, functional annotations, orthologues and paralogues, cross-references to other resources, and the genomic context of the gene via the genome browser. A recent addition has been variation data from re-sequencing of isolates from selected species. These data can be viewed directly in WormBase ParaSite, via custom pages and tables (Fig. 1), and the genome browser (see Section 2.2).

The principal entry point for the comparative genomics data (Section 2.3) is from the gene pages. The left-hand side-bar has links to tables of orthologues and paralogues for the gene, which can be filtered in a variety of ways. It is also possible to view the full tree for the family of which the gene is a member, collapsing and expanding sub-trees according to interest (Fig. 2). Alongside this, protein multiple alignments for the family, or selected sub-trees, can be viewed and downloaded in a variety of formats.

It is also possible to create a personal user account in WormBase ParaSite, and to store configuration and tool results under that account. Attaching a custom genome browser track whilst “logged in” will result in that track becoming available on any computer where the user has logged into their account. Additionally, results from the online tools (see Sections 3.3–3.5) are stored indefinitely for retrieval at a later date.

3.2. Genome browser

We provide a fully interactive genome browser for every genome. By default, this shows the assembly and annotated gene models. Additional tracks can be switched on using the left-hand navigation menu of the genome browser. Some tracks are available for every genome: DNA sequence, 6-frame protein translation, %GC content, repetitive elements, protein-coding gene structures, and non-coding RNAs. Other tracks are available only for genomes for which data is currently available. For example, we have RNA-Seq tracks for 8 genomes (see Section 2.2). The RNA-Seq tracks can be used obtain a graphical overview of the gene expression landscape for a gene, or genomic region (Fig. 3).

The browser allows users to view their own data in the context of the reference data. For small files (less than 20MB), these can be uploaded directly through the web interface. Larger files must be stored externally (e.g. on a FTP server), where they will be retrieved

Please cite this article in press as: K.L. Howe et al., WormBase ParaSite – a comprehensive resource for helminth genomics, Mol Biochem Parasitol (2016), http://dx.doi.org/10.1016/j.molbiopara.2016.11.005
on-demand by the WormBase ParaSite website. A variety of file formats can be attached, including BigWig, BAM, CRAM and UCSC Track Hubs (http://parasite.wormbase.org/info/Browsing/Upload).

3.3. Sequence searching

Our BLAST service (using NCBI-BLAST+ [69]) can be used to search a nucleotide or peptide sequence against the complete proteome or genome of any combination of species in WormBase ParaSite. We have short-cut buttons for searching common groups of species (e.g. all nematodes, or all platyhelminths), and also provide a graphical widget for defining a fully customised list. By default, BLAST results are saved on the WormBase ParaSite servers for seven days. However, by saving a BLAST submission to a user account, results are retained indefinitely. BLAST hits can be viewed and downloaded in a variety of formats (http://parasite.wormbase.org/info/Tools/blast.html).

3.4. Advanced querying

Advanced search and custom data export is available through our BioMart tool (http://parasite.wormbase.org/biomart). BioMart is designed to allow the simple and intuitive creation of custom tables and sequence files from sets of genes that meet a defined set of criteria [80].

A typical use of WormBase ParaSite BioMart involves three basic steps:

1. Define query filters: these are a set of criteria that must be met for genes to be included in the output. Filters can be narrow (e.g. a specific list of gene identifiers) or broad (e.g. all genes for a species or entire clade of species).
2. Define output attributes: these are the columns that will appear in the output table or, for sequences, the information that will be included in the entry headers.
3. View results. Initially, a preview of the first 10 lines of output are shown in the browser. Once happy that the query is working as intended, the user can then export the full result set to a file.

Please cite this article in press as: K.L. Howe, et al., WormBase ParaSite – a comprehensive resource for helminth genomics, Mol Biochem Parasitol (2016), http://dx.doi.org/10.1016/j.molbiopara.2016.11.005
The types of data that be used as filters and attributes include:

- Gene IDs, names, and descriptions
- Identifiers for data from external databases (e.g. UniProt, RefSeq)
- Gene structure (e.g. exons, introns)
- Protein domains and function (e.g. InterPro, Pfam)
- Gene Ontology annotations
- Orthologues and paralogues
- Sequences (e.g. genomic, transcript, peptide)

A typical use-case for BioMart might be to find all membrane proteins in filarial nematodes that do not have a human orthologue. This query relies on the fact that we have flagged potential transmembrane proteins using the TMHMM [81] software. The query is performed by (a) selecting the “Filarioidea” taxon in the SPECIES filter; (b) selecting “Restrict to genes without orthologues in Human” in the HOMOLOGY filter; (c) selecting “Limit to genes with TMHMM protein features” in the PROTEIN DOMAINS filter; and (d) clicking the “Results” button. Another example use-case is shown in Fig. 4, and more detailed instructions and worked examples are available from our help pages (http://parasite.wormbase.org/info/Tools/biomart.html).

3.5. Other tools

We provide a collection of files for each genome, including the genome sequence itself (both plain and repeat-masked), and the annotations in a variety of formats. Individual files can be downloaded from our browser (http://parasite.wormbase.org/ftp.html), or bulk downloads can be performed by FTP (ftp://ftp.wormbase.org/pub/wormbase/parasite). The structure of the site is fairly self-explanatory, with folders nested by (a) releases, (b) species in that release, (c) genomes for that species (identified by NCBI BioProject), and (d) data files for that genome.

Another tool we make available is the Ensembl Variant Effect Predictor (VEP) [82]. This annotates genomic variants from re-sequencing or population genomics studies with predictions of the how the reference annotations are affected by the variants (e.g. whether they coincide with protein-coding genes, and the putative effect they have on the corresponding protein sequence). The user can upload a list of variants in some standard formats, e.g. the Variant Call Format (VCF). Results are fully exportable, including as annotated VCF files. Full documentation and examples are available on the website (http://parasite.wormbase.org/info/Tools/vep.html).

For bioinformaticians wishing to develop applications using the data in WormBase ParaSite, direct programmatic access is available via two routes. Firstly, we provide a “RESTful” application programming interface (API) [83] that can be used with any programming language. A documented catalogue of “endpoints” and example code is available through the RESTful API section of the website (http://parasite.wormbase.org/rest). Secondly, data be retrieved from our BioMart service using the R programming language and BioMart Bioconductor package (http://parasite.wormbase.org/info/Tools/biomart.html). These services allow software developers and analysts to retrieve and manipulate the latest data in WormBase ParaSite on demand, without having to download complete data sets in bulk.

4. Challenges and future directions

The availability of many helminth reference genomes opens up the possibility a wide variety of functional genomics investigations, from life-stage specific gene expression to population genetics.
Making the data from these studies available in a way that is useful to users has traditionally required inefficient and redundant interactions between data providers, Wormbase curators and numerous archive and specialist resources. In order to reduce these costs as data volumes continue to grow rapidly, we are increasingly moving towards a model where our website pulls data directly from the archives and other complementary resources, using their existing programming interfaces. As a pilot for this model, we have brokered the submission of genome variation data from Strongyloides ratti (Mark Viney, pers. comm.) and Schistosoma mansoni [84] to the European Variation Archive (EVA) [85]. Our website code has been extended to extract these data from the EVA and display them directly (see Section 3.1).

For the past year, we have been aligning selected helminth RNA-Seq data sets to the corresponding reference genome, and making the alignments available as Track Hubs [86] for visualization in Wormbase-ParaSite and other compliant browsers (Fig. 4). We are now collaborating with the Gene Expression Atlas (GxA) [56] to provide curation, analysis, display and querying of helminth gene expression data. Specific high-value data-sets are curated within the GxA framework, and these can be interrogated by the GxA tools (for example querying for genes with a similar expression profile to a given gene). To simplify the user experience, we plan to embed access to these tools directly in the Wormbase ParaSite. We will also continue the curation to expression data sets in other species.

We will continue to extend Wormbase ParaSite with functionality geared towards the use-cases of helminth parasitologists. One example of this is the identification of candidate genes for therapeutics. In the near future, we will develop a pipeline to identify reliable homology between helminth genomes and known drug targets in the ChEMBL database [87].

Funding

Wormbase ParaSite is funded by the UK Biotechnology and Biological Sciences Research Council [BB/K020080].

Acknowledgements

We are indebted to Alessandra Traini, Eleanor Stanley and Jane Lomax for their work on Wormbase ParaSite. We would also like to thank Michael Paulin, Avril Cochran, and other members of the Wormbase and WTSI Parasite genomics groups for their advice and feedback.

References


Please cite this article in press as: K.L. Howe et al., WormBase ParaSite—a comprehensive resource for helmint genomics, Mol Biochem Parasitol (2016), http://dx.doi.org/10.1016/j.molbiopara.2016.11.005