International Journal for Parasitology 42 (2012) 161–169

Contents lists available at SciVerse ScienceDirect

International Journal for Parasitology

journal homepage: www.elsevier.com/locate/ijpara

CattleTickBase: An integrated Internet-based bioinformatics resource for *Rhipicephalus* (*Boophilus*) *microplus* ☆

Matthew I. Bellgard [a,b,1], Paula M. Moolhuijzen [a,b,1], Felix D. Guerrero [c,*], David Schibeci [a], Manuel Rodriguez-Valle [b,d], Daniel G. Peterson [e], Scot E. Dowd [f], Roberto Barrero [a], Adam Hunter [a], Robert J. Miller [g], Ala E. Lew-Tabor [a,b,d]

[a] *Centre for Comparative Genomics, Murdoch University, Perth, WA 6150, Australia*
[b] *Cooperative Research Centre for Beef Genetic Technologies, Armidale, NSW 2350, Australia*
[c] *USDA-ARS Knipling Bushland US Livestock Insect Research Laboratory, 2700 Fredericksburg Road., Kerrville, TX 78028, USA*
[d] *Queensland Alliance for Agriculture, Food & Innovation, The University of Queensland, Department of Employment, Economic Development & Innovation, P.O. Box 6097, St. Lucia, 4067 QLD, Australia*
[e] *Department of Plant & Soil Sciences, Life Sciences & Biotechnology Institute, Mississippi State University, 117 Dorman Hall, Box 9555, Mississippi State, MS 39762, USA*
[f] *Molecular Research, 503 Clovis Road, Shallowater, TX 79363, USA*
[g] *USDA-ARS Cattle Fever Tick Research Laboratory, 22675 North Moorefield Road, Building 6419, Edinburg, TX 78541, USA*

ARTICLE INFO

Article history:
Received 9 September 2011
Received in revised form 16 November 2011
Accepted 17 November 2011
Available online 1 December 2011

Keywords:
Cattle tick
Transcriptome
Genome project
Bioinformatics

ABSTRACT

The *Rhipicephalus microplus* genome is large and complex in structure, making it difficult to assemble a genome sequence and costly to resource the required bioinformatics. In light of this, a consortium of international collaborators was formed to pool resources to begin sequencing this genome. We have acquired and assembled genomic DNA into contigs that represent over 1.8 Gigabase pairs of DNA from gene-enriched regions of the *R. microplus* genome. We also have several datasets containing transcript sequences from a number of gene expression experiments conducted by the consortium. A web-based resource was developed to enable the scientific community to access our datasets and conduct analysis through a web-based bioinformatics environment called YABI. The collective bioinformatics resource is termed CattleTickBase. Our consortium has acquired genomic and transcriptomic sequence data at approximately 0.9X coverage of the gene-coding regions of the *R. microplus* genome. The YABI tool will facilitate access and manipulation of cattle tick genome sequence data as the genome sequencing of *R. microplus* proceeds. During this process the CattleTickBase resource will continue to be updated.

Published by Elsevier Ltd. on behalf of Australian Society for Parasitology Inc.

1. Introduction

The global cattle population is estimated at approximately 1 billion. Of this population, 80% inhabit areas that have been considered suitable habitat for ticks and tick-borne diseases (Snelson, 1975). The cattle tick *Rhipicephalus* (*Boophilus*) *microplus* is considered the most significant cattle parasite in the world, having established populations in most of the world's tropical and subtropical countries. This tick causes blood loss and physical damage to hides of infested animals. In addition, *R. microplus* is the vector for several bovine diseases, including babesiosis (caused by protozoan species *Babesia bovis* and *Babesia bigemina*) and anaplasmosis (caused by the rickettsia *Anaplasma marginale*), with severe impact on agricultural systems globally (de Castro, 1997). Economic losses to cattle producers from ticks and tick-borne diseases are US$13–18 billion globally on an annual basis (de Castro, 1997). Annual losses attributable to *R. microplus* in Brazil and Australia alone are estimated at US$2 billion (Grisi et al., 2002) and AUS$175 million (Playford et al., 2005), respectively.

Ticks are believed to be among the most ancient terrestrial arachnids and possibly the earliest organisms to have evolved blood-feeding capabilities (Mans and Neitz, 2004). *Rhipicephalus microplus* is a single-host species and has evolved such that it must maintain sustained contact with its host during the life stages, from the attached and feeding larva through to the fully engorged female. This period of attachment typically lasts approximately 3 weeks with some variation depending on environmental conditions. The species has developed a unique means of avoiding the host animal's immune responses during infestation (Wikel, 1999) and *R. microplus* salivary gland extracts have been shown to have an immunosuppressive effect on the bovine host (Turni et al.,

☆ *Note:* Nucleotide sequence data reported in this paper are available in GenBank under Accession Nos. HN108288–HN118367, HM748958–HM748967, HN108288–HN118367.

* Corresponding author. Tel.: +1 830 792 0327; fax: +1 830 792 0314.
E-mail address: Felix.Guerrero@ars.usda.gov (F.D. Guerrero).

[1] These authors contributed equally to this work.

0020-7519/$36.00 Published by Elsevier Ltd. on behalf of Australian Society for Parasitology Inc.
doi:10.1016/j.ijpara.2011.11.006

2004, 2007). The tick must also respond to many microorganisms, both symbiotic and parasitic, from the external environment or those ingested through feeding-associated activities (Andreotti et al., 2011).

With this interplay between bovine host, tick and microbiota, determining the whole genome sequence of *R. microplus* will greatly advance tick gene discovery, enable a better understanding of tick-host-pathogen immunology and provide insight on how the cattle tick responds to environmental perturbations, including pressures from moisture and temperature extremes and acaricidal applications. The *R. microplus* genome size is estimated to be 7.1 Gigabase pairs (Gbp), more than twice the size of the human genome, and consists of greater than 70% repetitive DNA (Ullmann et al., 2005). It is therefore a challenge for de novo assembly, even with contemporary DNA sequencing technologies. A 4X shotgun coverage genome sequence for the blacklegged tick, *Ixodes scapularis*, is available (Lawson et al., 2009) and is the only reported tick genome sequence to date. The version 1.1 sequence assembly consists of 369,492 supercontigs, totalling 1.76 Gbp with a supercontig N50 size of 72 kb.

From a taxonomic perspective, both *R. microplus* and *I. scapularis* are classified as hard ticks. There are two lineages of hard ticks, the Prostriata, which consists of the single genus *Ixodes* containing approximately 250 species, and the Metastriata, which consists of approximately 464 species from several genera including *R. microplus* (Barker and Murrell, 2004). Given the sequence divergence between *R. microplus* and *I. scapularis* (Guerrero et al., 2006), gene discovery efforts solely using *I. scapularis* as the model tick genome would prove limiting for *R. microplus* research efforts.

Towards a goal of generating a genetic resource for this economically important tick species, efforts have focused on a combination of sequencing strategies. The goal of this project was to maximise the utility of the data that could be generated with the resources available. To date, these strategies include Cot-filtered genomic DNA sequencing, bacterial artificial chromosome (BAC)-end sequencing (BES), targeted whole BAC sequencing, whole transcriptome sequencing and small RNA sequencing. Presently, we have acquired, assembled and annotated over 2 Gb of sequence data. This is comprised of 1.7 Gb of assembled contigs from three Cot reassociation experiments. These Cot experiments utilised methodologies to select randomly sheared genomic DNA for fractions depleted in highly repetitive sequences and enriched for putative gene coding regions (Guerrero et al., 2010). Also available are three transcriptome library assemblies (21 Mb) representing over 33,000 transcripts. Integrating data generated from these various approaches is already starting to provide new insights into the very large and complex cattle tick genome. This paper provides an overview of the coordinated cattle tick genome sequence resource as well as a new Internet-based bioinformatics resource that is designed to integrate our various genomic and transcriptomic datasets. This enables the cattle tick research community to access and analyse genomic and transcriptomic data at a single online resource, similar to approaches used by insect and worm researchers e.g. FlyBase, WormBASE (Harris and Stein, 2006; Drysdale, 2008). Examples are provided of these new genomic analysis tools and how they can be utilised by the research community to understand the genomic structure, organisation and content of the cattle tick genome.

2. Materials and methods

2.1. Source of tick materials

For the USA ticks, genomic and Cot DNA were extracted from eggs of the *R. microplus* Deutsch strain, f7, f10, f11 and f12 generations. These were pooled and a total of 10 g was used to purify very high molecular weight genomic DNA (Guerrero et al., 2010). This strain was started from only a few individual engorged females collected from a 2001 tick outbreak in South Texas. Although the strain has been inbred since its creation in 2001, it is not genetically homogeneous. For the Australian ticks, the larvae and fully engorged N strain of Australian *R. microplus* were utilised in these analyses. The N strain is maintained by the Biosecurity laboratories at the Department of Employment, Economic Development & Innovation (DEEDI), Queensland, under controlled conditions of 28 °C and 80% relative humidity prior to bovine infestation (Stewart et al., 1982).

2.2. Sequencing and assembly

For the BAC library synthesis, approximately 2 g of larvae from the f8 generation of the Deutsch strain were used by Amplicon Express Inc. (Pullman, WA, USA) to isolate genomic DNA partially digested with *Mbo*I to synthesise a BAC library of approximately 0.8X coverage (Guerrero et al., 2010). Subsequently, a second library of 2.4X coverage was synthesised from genomic DNA partially digested with *Hind*III. Five BAC assemblies, BM-074-Random-F12, BM-077-Random-J09, BM-129-CzEst9-N14, BM-066-M07, BM-077-G20, are as described by Guerrero et al. (2010). The remaining 10 BAC sequences were trimmed for vector and bacterial contamination by phred-phrap software (Ewing and Green, 1998) cross_match, with options set at minmatch 12 and minscore 20. Contig order and orientation were based on Phrapview paired end reads.

Total *R. microplus* genomic DNA was prepared and processed by three Cot filtration experiments to enrich for single/low-copy and moderately repetitive DNAs. Cot-filtered DNA was sequenced using 454 FLX and Titanium pyrosequencing (Research and Testing Laboratory, Lubbock, TX, USA). Methods are as described in Guerrero et al. (2010).

The filtered genomic DNA (total number of reads 7,289,230 and total number of bases 1,798,400,445) was de novo assembled using the Newbler assembler for 454 reads (Margulies et al., 2005) with default settings. All contigs (745,975 sequences) and BES (GenBank Accession Nos. HN108288–HN118367) were then assembled with Cap3 (Huang and Madan, 1999) default settings. Whole Genome Shotgun (WGS) project ADMZ02000000 is the result of this two-step assembly.

2.3. BAC and Cot read alignment

BAC and Cot read alignments were carried out with BWA-SW (Li and Durbin, 2010) for long reads as our average read length was 245 bp. A mapping accuracy >99% was expected with a mapping quality (MapQ) of 10 and sensitivity (*Z*) of 100.

2.4. Gene predictions

Gene predictions for the BAC sequences and WGS were made with GenScan version 1.0 (Burge and Karlin, 1997) for default 'optimal exons', parameters for human/vertebrates and coding sequences (CDS) option. BAC predicted gene and AutoFACT (Koski et al., 2005) annotation can be found in Supplementary Table S1 and Supplementary Fig. S1.

2.5. Repeat analysis

Repeat sequences and rRNA were identified using RepeatMasker version 3.2.6 (Smit, A.F.A., Hubley, R., Green, P., 2004. RepeatMasker Open-3.0. 1996–2010 <http://www.repeatmasker.org>) with parameters set up for the arthropod clade of input sequences.