International Journal for Parasitology xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

International Journal for Parasitology

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

Gene-enriched draft genome of the cattle tick *Rhipicephalus microplus*: assembly by the hybrid Pacific Biosciences/Illumina approach enabled analysis of the highly repetitive genome

Roberto A. Barrero [a], Felix D. Guerrero [b], Michael Black [a], John McCooke [a], Brett Chapman [a], Faye Schilkey [c], Adalberto A. Pérez de León [b], Robert J. Miller [d], Sara Bruns [e], Jason Dobry [e], Galina Mikhaylenko [e], Keith Stormo [e], Callum Bell [c], Quanzhou Tao [e], Robert Bogden [e], Paula M. Moolhuijzen [f], Adam Hunter [a], Matthew I. Bellgard [a,*]

[a] *Centre for Comparative Genomics, Murdoch University, WA 6151, Australia*
[b] *USDA-ARS Knipling-Bushland US Livestock Insects Research Laboratory and Veterinary Pest Genomics Center, 2700 Fredericksburg Rd., Kerrville, TX 78028, USA*
[c] *National Center for Genome Resources, Santa Fe, NM, USA*
[d] *USDA-ARS Cattle Fever Tick Research Laboratory, 22675 North Moorefield Rd., Edinburg, TX 78541, USA*
[e] *Amplicon Express, Pullman, WA, USA*
[f] *Centre for Crop Disease and Management, Curtin University, Bentley, WA 6102, Australia*

ARTICLE INFO

Article history:
Received 13 June 2016
Received in revised form 16 March 2017
Accepted 16 March 2017
Available online xxxx
Note: The *R. microplus* v2.0 genome assembly has been deposited at GenBank/DDBJ/ENA under the accession LYUQ00000000. Raw Illumina and PacBio reads were submitted to the Short Read Archive (SRA) database under the BioProject PRJNA312025.

Keywords:
Cattle tick
Low-Cot enrichment
MicroRNAs
Tick DNA repeats
PacBio error correction
Complex genome

ABSTRACT

The genome of the cattle tick *Rhipicephalus microplus*, an ectoparasite with global distribution, is estimated to be 7.1 Gbp in length and consists of approximately 70% repetitive DNA. We report the draft assembly of a tick genome that utilized a hybrid sequencing and assembly approach to capture the repetitive fractions of the genome. Our hybrid approach produced an assembly consisting of 2.0 Gbp represented in 195,170 scaffolds with a N50 of 60,284 bp. The Rmi v2.0 assembly is 51.46% repetitive with a large fraction of unclassified repeats, short interspersed elements, long interspersed elements and long terminal repeats. We identified 38,827 putative *R. microplus* gene loci, of which 24,758 were protein coding genes (≥100 amino acids). OrthoMCL comparative analysis against 11 selected species including insects and vertebrates identified 10,835 and 3,423 protein coding gene loci that are unique to *R. microplus* or common to both *R. microplus* and *Ixodes scapularis* ticks, respectively. We identified 191 microRNA loci, of which 168 have similarity to known miRNAs and 23 represent novel miRNA families. We identified the genomic loci of several highly divergent *R. microplus* esterases with sequence similarity to acetylcholinesterase. Additionally we report the finding of a novel cytochrome P450 CYP41 homolog that shows similar protein folding structures to known CYP41 proteins known to be involved in acaricide resistance.

1. Introduction

The cattle tick, *Rhipicephalus microplus*, is a tick that parasitizes cattle in tropical and subtropical countries. This tick is a vector for several bovine diseases, harboring such infectious pathogens as *Babesia bovis*, *Babesia bigemina*, and *Anaplasma marginale*. The economic burdens caused by this parasite are enormous, impacting at levels from family farmers up to large cattle production operations (de Castro, 1998). Annual losses attributed to this tick have been estimated to be over USD 2 billion and AUD 100 million for Brazil and Australia, respectively (Angus, 1996; Grisi et al., 2002). The United States eradicated the cattle tick in the 20th century and the annual savings attributable to this eradication project have been estimated at USD 3 billion in 2015 dollar value (Graham and Hourrigan, 1977). Global climate change has exacerbated the threat of the cattle tick reinfesting the United States and expansion of its range in other regions of the world.

New countermeasures are needed to protect and enhance the productivity of livestock affected by the cattle tick and the diseases it transmits. The primary method of control implemented against the cattle tick centers upon applications of chemical acaricide to

* Corresponding author.
E-mail address: mbellgard@ccg.murdoch.edu.au (M.I. Bellgard).

http://dx.doi.org/10.1016/j.ijpara.2017.03.007
0020-7519/

infested herds of cattle. However, *R. microplus* has developed economically significant levels of resistance to all the commercially available acaricides (Andreotti et al., 2011; Rodriguez-Vivas et al., 2014) and there is a great need for the development of novel effective tick control technologies. Tick vaccines are an option for cattle tick control and in some cases the tick control efficacy of vaccination exceeded 99% (Canales et al., 2009). However, the only commercially available tick vaccine suffers from variable efficacy against *R. microplus* and the search for new vaccines is ongoing.

This need for novel tick control technology drove the initiation of a cattle tick genome sequencing project in 2005, beginning with acquisition of expressed sequence tags (ESTs) using Sanger protocols (Guerrero et al., 2005) and determination of genome size by a reassociation kinetics-based approach (Ullmann et al., 2005). The estimated genome size of 7.1 Gbp and the highly repetitive nature of the cattle tick *R. microplus* genome precluded full genome sequencing until the commercial maturation of second (e.g. Illumina Inc, San Diego, CA, USA) and third generation (e.g. Pacific Biosciences (PacBio), USA) sequencing technologies. In the meantime, the cattle tick *R. microplus* genome sequencing project focused upon elucidating sequences from the transcriptome (Wang et al., 2007; Lew-Tabor et al., 2010) and unique low copy fraction of the genome (Guerrero et al., 2010). This incipient genome project enabled several reverse vaccinology approaches aimed at identification of target antigens in the cattle tick for tick vaccine development (Guerrero et al., 2012; Maritz-Olivier et al., 2012).

Ticks are believed to be among the earliest terrestrial arachnids, perhaps the first to develop blood-feeding capabilities (Mans and Neitz, 2004). The Prostriata lineage of hard ticks is composed of a single genus, *Ixodes*, containing 243 species. Assembled tick genome sequences are currently available only for the Prostriate ticks, *Ixodes scapularis* (Gulia-Nuss et al., 2016; https://www.vectorbase.org/organisms/ixodes-scapularis) and *Ixodes ricinus* (Cramaro et al., 2015). *Ixodes scapularis* was sequenced using a Sanger whole genome shotgun approach and the *I. ricinus* genome was sequenced using Illumina 100 nucleotide (nt) paired-end reads. Both of these assemblies contain high numbers of scaffolds that could likely be further assembled with the aid of long reads. The Metastriata line of hard ticks consists of 13 genera and over 459 species, including many species of medical and veterinary importance across the genera *Rhipicephalus*, *Hyalomma*, *Hemaphysalis, Amblyomma*, and *Dermacentor* (Guglielmone et al., 2010). The evolutionary distance between *I. scapularis* and the Metastriate ticks results in significant sequence divergence between orthologous genes, impeding molecular studies of the Metastriates. The persisting scientific and applied agricultural need for a Metastriate genome assembly drove the design and implementation of a hybrid genome sequencing/assembly approach for the *R. microplus* project. Initially, we acquired an Illumina- and 454-based blended draft-level genome assembly. This assembly composed primarily of contigs derived from the low-Cot unique DNA fraction was curated and published as part of the resources provided by the CattleTickBase (Bellgard et al., 2012). However, the commercial introduction of the PacBio platform, offering single molecule real-time sequencing with long reads (Eid et al., 2009), facilitated movement of the cattle tick *R. microplus* genome sequencing to the final phase tackling the complex repetitive regions of the genome.

Our study reports the assembly and annotation of the 7.1 Gbp *R. microplus* genome. We generated long reads of very high molecular weight genomic DNA by PacBio protocols. A subset of these reads was error-corrected by an assembled set of Illumina-generated contigs sourced from genomic DNA purified by reassociation kinetics protocols to select for the unique low-copy genome fraction. Assembly programs were customized to take optimal advantage of Cloud-based computational resources, as the huge scope of the error-correction process exceeded the available super-computer resources in Australia. The genome was searched for microRNAs (miRNA) and the expansion of the numbers of known candidate miRNAs was significant. The transcriptome of *R. microplus* was mapped to the genome assembly and functional annotation identified metabolic pathway members and gene ontologies (GO).

2. Materials and methods

2.1. Source of tick materials

Genomic DNA was extracted from pooled collections of eggs from the f7, f10, f11, and f12 generation of the *R. microplus* Deutsch strain. The Deutsch strain was started from a few individual engorged females collected from a 2001 tick outbreak in Webb County, TX, USA. Although the strain has been inbred since its collection and creation in 2001, it is not genetically homogeneous. A total of 10 g of eggs was used in a protocol from Sambrook et al. (1989) to purify very high molecular weight genomic DNA (Guerrero et al., 2010). The protocol consisted of pulverizing frozen material in a liquid nitrogen-cooled mortar and pestle, addition to an aqueous buffer, followed by RNAse treatment, digestion by proteinase K, phenol extraction, and dialysis in 50 mM Tris, 10 mM EDTA, pH 8.0. The resultant DNA was determined by agarose gel electrophoresis to be >200 kb. An aliquot of this genomic DNA was processed by Cot filtration to enrich for single, low copy, and moderately repetitive genomic DNA (Guerrero et al., 2010).

2.2. Preparation of a bacterial artificial chromosome (BAC) library and sequencing of random BAC clones

A genomic BAC library of *R. microplus* was constructed as previously described (Guerrero et al., 2010) and 18,432 BAC clones were randomly selected and sequenced using Illumina pair-end technology (described in Section 2.4) by Amplicon Express Inc. (Pullman, WA, USA). The *R. microplus* BAC library was constructed from High Molecular Weight (HMW) genomic DNA processed at Amplicon Express, Inc. as previously described (Tao and Zhang, 1998). HMW DNA was partially digested with the restriction enzyme *Bam*HI and size selected prior to ligation of fragments into the pECBAC1 vector and transformation of DH10B *Escherichia coli* host cells, which were then plated on Luria–Bertani (LB) agar with chloramphenicol (12.5 μg/ml), X-gal and isopropyl β-D-1-thiogalactopyranoside (IPTG) at appropriate concentrations. Clones were robotically picked with a Genetix QPIX (Molecular Devices, Sunnyvale, CA, USA) into 120 × 384-well plates containing LB freezing media. Plates were incubated for 16 h, replicated and then frozen at −80 °C. DNA from 28 random BAC clones was digested with 5 U of *Not*I enzyme for 3 h at 37 °C. The digestion products were separated by pulsed-field gel electrophoresis (CHEF-DRIII system, Bio-Rad, Hercules, CA, USA) in a 1% agarose gel in TBE. Insert sizes were compared with those of the Lambda Ladder MidRange I PFG Marker (New England Biolabs, Ipswich, MA, USA). Electrophoresis was carried out for 18 h at 14 °C with an initial switch time of 5 s, a final switch time of 15 s, in a voltage gradient of 6 V/cm. The average BAC clone insert size for the library was found to be 118 kb.

2.3. Focused Genome Sequencing (FGS)

Focused Genome Sequencing (FGS) was used to sequence 18,432 randomly selected *R. microplus* BAC clones. FGS is a Next Generation Sequencing (NGS) method developed at Amplicon Express that allows high quality assembly and scaffolding of BAC clone sequence data generated on the Illumina HiSeq platform (Illumina, Inc.). Individual BAC clones were made into Pools and Superpools according to US Patent 8301388 (Amplicon Express,

Please cite this article in press as: Barrero, R.A., et al. Gene-enriched draft genome of the cattle tick *Rhipicephalus microplus*: assembly by the hybrid Pacific Biosciences/Illumina approach enabled analysis of the highly repetitive genome. Int. J. Parasitol. (2017), http://dx.doi.org/10.1016/j.ijpara.2017.03.007