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

Inc., Pullman,; www.google.com/patents/US8301388). BAC DNA was prepared using a standard alkaline lysis procedure (Sambrook and Russell, 2001). Pooled BAC DNA was resuspended in 10 mM Tris, 1 mM EDTA, pH 8.0 at a high concentration and mechanically sheared to fragment sizes of 170 bp, 400 bp, 800 bp, 2,000 bp and 4,000 bp using a GeneMachines Hydroshear (San Carlos, CA, USA) and SonicMan (Spokane, WA, USA) technologies to achieve the respective sizes. Illumina libraries were prepared according to the manufacturer's instructions, and sequenced on Illumina HiSeq to an average depth of 100× coverage. The gapped contiguous sequences were ordered and orientated using Mira mapping software in conjunction with ALLPATH-LG taking into account various Illumina library sizes. Illumina libraries of insert sizes 170 bp, 400 bp, 800 bp, 2,000 bp and 4,000 bp were used on each pool of BAC DNA. The sequence reads were made into contigs and scaffolds using Mira and ALLPATHS-LG. Data from the assembled pools were deconvoluted using lookup tables based on the Pool and Superpool Matrix scheme. Unique sequence reads were deconvoluted with specific Illumina tags and mapped back to specific BAC clone addresses. Repeat elements were traced to BAC clone "sets" that share common repeats and sizes of the gaps caused by repeats were determined using ALLPATHS-LG output files. The FGS process makes NGS tagged libraries of BAC clones and generates a consensus sequence of the BAC clones with all reads assembled at ~80 bp overlap and ~98% identity.

2.4. Library preparation and sequencing of Illumina and PacBio datasets

Rhipicephalus microplus genomic DNA was sequenced at the National Center for Genomic Resources (Santa Fe, NM, USA) as described in McCooke et al. (2015). Briefly, the Illumina-based sequencing of the Cot-selected genomic DNA made use of the standard Illumina DNA library preparation protocol and the TruSeq DNA Sample Preparation V2 kit (Illumina Inc.). This library was sequenced as 100 nt-paired ends on three lanes in a flowcell using the HiSeq2000. The resulting raw reads were quality processed by the Illumina pipeline and the contaminant-filtering pipeline developed at the National Center for Genomic Resources. The PacBio sequencing was performed on five libraries prepared according to the Pacific Biosciences low-input 10 kb library preparation and sequencing protocol. C2 chemistry and XL polymerase were used with 178 SMRT cells.

2.5. PacBio error correction

Raw SMRT PacBio reads were error corrected as described by Au et al. (2012) using the Illumina reads generated from the Cot-selected genomic DNA. Large single copy (LSC) scripts for PacBio error correction were modified to reduce the size of intermediate files, which is a significant bottleneck when using large input datasets. The PacBio error correction was found to be very computationally intensive. Initially, we utilized the EPIC supercomputer resource at Murdoch University, Australia, which is an 87.2 TeraFLOPS system comprised of 9600 processors and 500 Tb of storage. However, small-scale error correction jobs allowed us to project that the entire error correction process would require over 4 months with the entire computer resource dedicated to this project and would be cost-prohibitive. Subsequently, it became obvious that cloud resources were required and a prioritizing of the PacBio sequences for error correction was also necessary. We benchmarked and designed a massive parallel approach to error correct up to 2,000 PacBio reads per job via an Amazon cloud computing service. Owing to the computational demand for the error correction process, only PacBio reads ≥5,000 nt were error corrected as this subset of the longest PacBio reads would provide the most valuable information to close gaps and expand the Rmi v2.0 genome assembly. Of note is that error correction of PacBio reads using LSC is limited by the alignment of short Illumina reads. PacBio reads with no aligned Illumina reads are not error corrected, while those with limited Illumina read coverage, the error correction takes place only between the foremost 5′-end to the foremost 3′-end mapped reads on the PacBio sequence. Either 5′-end and/or 3′-end terminal regions of a PacBio read without mapped Illumina reads will be trimmed during the error correction process (Au et al., 2012).

2.6. De novo genome assembly

The assembly strategy is depicted in Fig. 1. We drew from four sources of genome sequence data: (i) the existing assembled contigs of 454-sequenced Cot-selected genomic DNA contained in Rmi v1.0 (Guerrero et al., 2010), (ii) HiSeq reads from the Cot-selected DNA, (iii) reads from the Illumina-based shotgun sequencing of the 18,432 randomly selected BACs, and (iv) PacBio reads of unselected genomic DNA (Supplementary Table S1). The Cot-selected Illumina raw reads were quality trimmed using ConDeTri version 2.2 (Smeds and Kunstner, 2011), then de novo assembled using SOAPdenovo2 version 2.04-r240 (Luo et al., 2012) and Ray version 2.3.1 (Boisvert et al., 2010). Redundancy of contigs from the data derived from the BACs, Cot-selected genomic DNA, and Rmi v1.0 was removed using CD-HIT (Fu et al., 2012) and/or BLAT (Kent, 2002) sequence clustering and comparisons, respectively (Fig. 1). Resulting Illumina scaffolds were aligned to individual error-corrected PacBio reads using BLASTN. Error-corrected PacBio reads with alignment ≥200 bp and identity ≥90% to scaffolds were used to join Illumina scaffolds using PBJelly2 (PBSuite_15.8.24; English et al., 2012). Finally, error-corrected PacBio reads were de novo assembled using CANU (Berlin et al., 2015), a new PacBio assembler based on the Celera assembler (Myers et al., 2000). CANU utilises the MinHash Alignment Process (MHAP) for overlapping noisy, long reads using probabilistic, locality-sensitive hashing. In assembling the cleaned reads, a high MHAP sensitivity and low minimum coverage of 2× was set along with a high error rate of 0.12. We ran a shortened CANU pipeline, only running at the 'assemble option', as the reads have already been error-corrected.

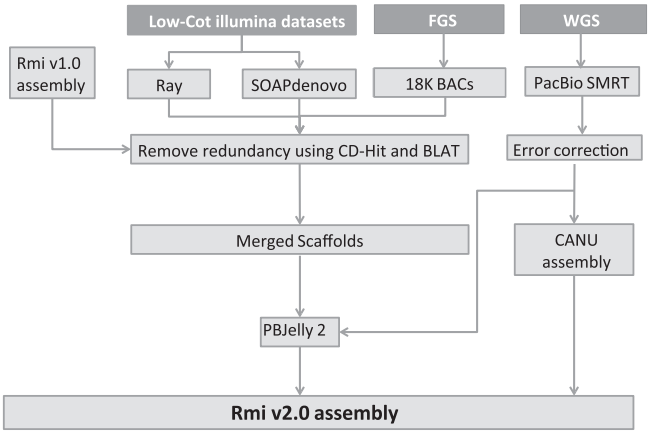

Fig. 1. Overall strategy for the de novo assembly of the cattle tick *Rhipicephalus microplus* Rmi v2.0 genome. Illumina low-Cot, Focused Genome Sequencing and whole genome shotgun datasets were de novo assembled and then merged with Rmi v1.0 (Guerrero et al., 2010). Error corrected PacBio reads were then used to close gaps and join Illumina scaffolds using PBJelly2 (PBSuite_15.8.24; English et al., 2012). Additionally, error corrected PacBio reads were de novo assembled using CANU (Berlin et al., 2015).

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

The process depicted in Fig. 1 produced the *R. microplus* assembly designated as Rmi v2.0.

The *R. microplus* v2.0 genome assembly has been deposited at GenBank/DDBJ/ENA under the accession LYUQ00000000. The version described in this paper is version LYUQ01000000. Raw Illumina and PacBio reads were submitted to the Short Read Archive (SRA) database under the BioProject PRJNA312025.

2.7. Discovery and annotation of *R. microplus* repeat families

Illumina scaffolds derived from sequencing of the Cot-selected DNA and the BAC clones were used as templates to compile *R. microplus* repeat models using RepeatModeler v1.0.4 (http://www.repeatmasker.org/RepeatModeler.html). RepeatModeler uses RECON (Bao and Eddy, 2002) and RepeatScout v1 (Price et al., 2005) to perform de novo predictions of repeat families. The Repbase library 20150807 (Jurka et al., 2005) was used for reference-based repeat element searches. Tandem repeat elements were predicted using trf (Benson, 1999). RepeatMasker version 3.2.9 (http://www.repeatmasker.org/) was then used to mask these predicted repeat elements in the Rmi v2.0 assembly (described in Section 2.6) using the predicted *R. microplus* custom repeat model. Additionally, contigs and scaffolds from the BAC, 454 (Rmi v1.0), Cot-selected Illumina and PacBio-derived datasets were masked as separate datasets to compare the repeat content detected by the different technologies and/or enrichment approaches.

2.8. Gene prediction and genome annotation

Gene predictions were performed using two approaches: (i) ab initio prediction using MAKER version 2.31.8 (Cantarel et al., 2008), SNAP version 2006-07-28 (Korf, 2004) and Augustus version 3.01 (Stanke et al., 2006), and (ii) mapping non-redundant *R. microplus* transcripts onto the Rmi v2.0 assembly using BLAT (Kent, 2002) (Supplementary Fig. S1).

Public and unpublished *R. microplus* transcriptome datasets were clustered using Cd-hit (Li and Godzik, 2006) with a minimal 95% sequence identity as previously reported (Ma et al., 2014) yielding a 63,416 non-redundant *R. microplus* transcript dataset (Supplementary Table S2). These non-redundant sequences were then parsed using RepeatMasker v4.0.5 (http://www.repeatmasker.org/) to identify repeat containing sequences. Transcripts encoding repeats were discarded, leaving 43,874 non-redundant *R. microplus* transcripts for ab initio predictions. These filtered transcripts were provided to MAKER (Cantarel et al., 2008) as 'EST' evidence to generate a *R. microplus* 'training set' for ab initio programs. SNAP (Korf, 2004) or AUGUSTUS (Stanke et al., 2006) were then trained and their resulting predicted gene models were used as a 'refined gene model training set' to re-run these tools. The predicted refined gene models by both SNAP (Korf, 2004) and AUGUSTUS (Stanke et al., 2006) were then combined and used to re-run MAKER to identify *R. microplus* gene loci. MAKER identified 67,982 putative gene loci that were then filtered to < 0.5 Annotation Edit Distance (AED), yielding 8,084 gene loci supported by 8,645 gene isoforms. The gene isoform with the lowest AED value was selected as the representative sequence for each locus (Supplementary Fig. S1).

To identify additional *R. microplus* gene loci that may have been missed by the ab initio pipelines, we mapped the 63,416 non-redundant *R. microplus* transcripts onto the Rmi v2.0 genome assembly using BLAT (Kent, 2002) with at least 95% sequence identity and over 50% sequence coverage. A total of 30,301 non-redundant *R. microplus* transcripts mapped onto the Rmi v2.0 genome assembly and then parsed using RepeatMasker as described above (Supplementary Fig. S1). Three filtering strategies were applied to mapped transcripts: (i) transcripts with no repeat content were re-aligned onto the Rmi v2.0 genome assembly using BLAT with ≥95% nt sequence identity and ≥50% sequence coverage; (ii) transcripts with ≥50% repeat content were discarded as these are not typical mRNA sequences and owing to the ambiguous mapping of their repeat regions; and (iii) mapped transcripts containing less than 50% repeats were evaluated for their protein coding potential using TransDecoder v3.0 (Tang et al., 2015). Transcripts with Open Reading Frames (ORFs) of at least 100 amino acids (aa) were re-aligned onto the Rmi v2.0 genome assembly with ≥95% sequence identity and ≥50% sequence coverage. To increase the mapping accuracy of transcripts lacking an ORF ≥100 aa, these were re-aligned onto the Rmi v2.0 genome assembly with ≥99% sequence identity and ≥90% sequence coverage. The genomic sequences of mapped transcripts passing the above filtering steps were concatenated and converted to GFF format using Blat2GFF (http://iubio.bio.indiana.edu:8081/gmod/tandy/blat2gff.pl). Subsequently these transcripts were clustered and merged using Gffread v2.2.1 (Cufflinks suite v2.2.1). A representative sequence for each locus was then selected using the following filtering criteria: (i) largest number of exons; (ii) highest BLAT bit score; and (iii) longest alignment length (Supplementary Fig. S1).

Finally, representative sequences from both ab initio predicted gene loci and transcript-mapping evidence supported gene loci were merged using Gffread v2.2.1. A representative gene sequence for each merged and non-redundant gene locus was then selected based on the following criteria: (a) encodes the largest number of exons; (b) has the lowest AED score; and/or (c) has the highest BLAT bit score. Representative gene sequences were then rerun using Gffread to define their genomic coordinates. A fraction of the RNA-seq data used in this study was not stranded, resulting in possible isoforms mapping to both sense and antisense strands on the same genomic region. To remove duplicates encoding the same ORF we used BEDTools intersect (v2.26.0) (Quinlan and Hall, 2010) to identify sense-antisense pairs and then their encoded ORFs were evaluated. Sense-antisense pairs encoding the same ORF were further assessed to remove 'duplicated' loci with a conflicting ORF orientation (Supplementary Fig. S1). As a final step, representative gene loci sequences were subjected to TransDecoder analysis to identify ORF ≥100 aa and classify identified protein coding genes into 'complete', '5prime partial', 'internal' or '3prime partial' proteins. Gene loci with an ORF shorter than 100 aa may encode non-coding RNAs or represent partial fragments of protein coding genes.

2.9. Discovery of *R. microplus* miRNA loci

Nearly 22 million raw *R. microplus* small RNA reads were obtained from Barrero et al. (2011a). Adaptor sequences were clipped using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Unclipped reads and/or clipped reads shorter than 18 nts were excluded from downstream analyses. Redundant reads were then collapsed using mapper.pl (Friedlander et al., 2008). Error-corrected PacBio reads and PacBio scaffolds were excluded from this analysis due to their anticipated ~5% to 10% sequencing error rate. Non-redundant small RNAs were mapped onto the Rmi v2.0 assembly using Bowtie with zero mismatches (Langmead et al., 2009). Candidate *R. microplus* miRNA loci were predicted using miRDeep as previously described (Friedlander et al., 2008). Predicted miRNA loci were retained if the locus: (i) has an overall miRDeep score greater than 4.0, (ii) does not contain Ns (ambiguous base calls) overlapping the loop region of the miRNA hairpin, and (iii) candidate precursors have no sequence similarity to the *R. microplus* mitochondrial genome (McCooke et al., 2015), *R. microplus* protein-coding genes, or repeat elements.

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

2.10. Functional annotation of protein-coding and RNA genes

Representative sequences for 38,827 *R. microplus* gene loci were annotated using AutoFACT (Koski et al., 2005). Top 10 hits were utilised to assign functional information derived from searches against the nr, uniref100, cog, pfam, smart databases, large subunit (LSU) and short subunit (SSU) ribosomal sequences (Koski et al., 2005). The AutoFACT pipeline was used to assign a 'gene description' to representative *R. microplus* gene sequences along with the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway information, cluster of orthologous genes (COG) function and/or gene ontology (GO).

To identify protein-coding genes, representative gene sequences for each *R. microplus* locus were subjected to TransDecoder v3.0 (Tang et al., 2015). Representative gene sequences encoding an ORF ≥100 aa were annotated as protein coding genes. The completeness of the identified protein coding genes was further assessed using TransDecoder and classified into 'complete', '5prime partial', '3prime partial' or 'internal' (Tang et al., 2015).

RNA genes were identified using two approaches: (i) AutoFACT top hit to LSU or SSU ribosomal sequences, and (ii) by screening representative gene loci sequences lacking an ORF ≥100 aa against RFAM database (Griffiths-Jones et al., 2003) using BLASTN (Altschul et al., 1990).

Post-translational modifications were predicted using NetNGlyc v1.0 (http://www.cbs.dtu.dk/services/NetNGlyc/) and NetPhos v3.1 (http://www.cbs.dtu.dk/services/NetPhos/) that identify putative N-glycosylation and phosphorylation sites, respectively.

Cytochrome P450 (CYP) coding genes were identified based on sequence similarity screening of *R. microplus* protein coding genes against three databases, namely, Cytochrome P450 Engineering Database (Cyped) V6.0 (Fischer et al., 2007), Arthropod P450 sequences fetched from the National Center for Biotechnology Information (NCBI), USA, gene database (https://www.ncbi.nlm.nih.gov/gene/?term=P450+and+arthopoda), and AutoFACT reference databases. P450 annotation was assigned based on the best match (highest bit score) against the three above databases sources. Positive CYP matches from across AutoFACT, Cyped and NCBI Arthropoda databases with the best match, were broken down into family, subfamily and member, according to the HUGO Gene Nomenclature Committee (http://www.genenames.org/genefamilies/CYP). Assignment of Cyp identifiers required >40% identity for family, >55% identity for sub-family, and >90% for member when identified within Arthropoda, all with a coverage threshold of ≥50%, as similarly done in Parvez et al. (2016) (REF http://www.nature.com/articles/srep33099). Regardless of whether the family, subfamily and/or member were identified, the assigned CYP identifiers were prefixed with 'putative' if the match was <80% identity.

All identified CYPs that were 'complete' coding ORFs according to TransDecoder (Tang et al., 2015) were then assessed through phylogenetic analysis to validate their assigned CYP. This was performed using MrBayes as previously described (Baldwin et al., 2009), using 11,265,000 generations to achieve a topological convergence of <0.01, with CYP sequences from four different species: *R. microplus* (Rmi), *Daphnia pulex* (Dpu), *Apis mellifera* (Ame) and *Drosophila melanogaster* (Dme). FigTree v1.4.3 was used to depict the final phylogenetic tree using a midpoint rooting.

To ascertain differences between the other CYP3s and the CYP41 sequences identified in the phylogenetic analysis, alignments of CYP3A and CYP41 sequences were performed using MUSCLE (Edgar, 2004). To identify pocket regions, binding sites and catalytic sites, the sequences used in the multiple sequence alignment were then converted to PDB format using Swiss-Model (https://swissmodel.expasy.org/) prior to further analysis. Catalytic sites were identified using Catalytic Site Identification (http://catsid.llnl.gov/), pocket regions were identified using CastP (http://sts.bioe.uic.edu/castp/), and binding sites within the pocket regions were identified using MetaPocket 2.0 (http://projects.biotec.tu-dresden.de/metapocket/).

2.11. Evolutionary analyses

To compare the evolutionary relationship of esterases or other protein families a BLASTP (*E*-value ≤1e−10) screening against the NCBI non-redundant database (https://blast.ncbi.nlm.nih.gov/) was conducted and top hits were selected. The evolutionary history was inferred using the Neighbor-Joining method (Saitou and Nei, 1987). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) was calculated as previously described (Felsenstein, 1985). The evolutionary distances were computed using the Poisson correction method (Zuckerkandl and Pauling, 1965) and are in the units of the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016).

2.12. Assembly and gene prediction assessment

The completeness of the *R. microplus* Rmi v2.0 genome assembly was evaluated using BUSCO (Simao et al., 2015). Predicted genes were assessed for completeness using both TransDecoder (Tang et al., 2015), which assessed completeness of predicted ORFs, and BUSCO (Simao et al. http://busco.ezlab.org/files/BUSCO-Simao-Waterhouse-Bioinformatics-2015.pdf) which assessed the predictions in context with known ancestral Arthropoda proteins.

Once protein-coding ORFs were predicted by TransDecoder (Tang et al., 2015), those were further processed by BUSCO for gene assessment, which uses hmmer3 to assess orthologous groups with single-copy orthologs from ancestral arthropoda proteins. The results are then interpreted as complete single-copy genes, complete duplicate copy genes, fragmented genes, and missing genes (ancestral proteins not accounted for).

2.13. Comparative analyses

To evaluate the conservation of identified *R. microplus* protein-coding genes in the Rmi v2.0 assembly, reference protein datasets for the following 11 species were selected: *Anopheles gambiae* (GCF_000005575.2_AgamP3_protein.faa), *A. mellifera* (GCF_000002195.4_Amel_4.5_protein.faa), *Bos taurus* (GCF_000003055.6_Bos_taurus_UMD_3.1.1_protein.faa), *Caenorhabditis elegans* (GCF_000002985.6_WBcel235_protein.faa), *D. melanogaster* (GCF_000001215.4_Release_6_plus_ISO1_MT_protein.faa), *D. pulex* (GCA_000187875.1_V1.0_protein.faa), *Danio rerio* (GCF_000002035.5_GRCz10_protein.faa), *Gallus gallus* (GCF_000002315.3_Gallus_gallus-4.0_protein.faa), *Homo sapiens* (GCF_000001405.31_GRCh38.p5_protein.faa), *I. scapularis* (GCF_000208615.1_JCVI_ISG_i3_1.0_protein.faa), and *Tribolium castaneum* (GCF_000002335.2_Tcas_3.0_protein.faa). Protein coding genes from all the species were then clustered using OrthoMCL as previously described (Li et al., 2003; Barrero et al., 2011b) to identify protein clusters common to all species ('core proteins') and those unique to subsets of species and/or unique to a single species lineage.

2.14. Data accessibility

Supplementary Tables S5, S7–S12 are available at Mendeley Data, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1.

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

3. Results and discussion

3.1. Rhipicephalus microplus genome assembly

The large size and the repetitive nature of the *R. microplus* genome imposed significant difficulties upon the assembly process as we sought to incorporate the repetitive fraction of the genome into the assembly. The first *R. microplus* genome assembly, Rmi v1.0 (Guerrero et al., 2010), focused only upon the low-copy/unique fraction of the genome, as it utilized Cot-selected (to remove the repetitive DNA fraction) genomic DNA for sequencing upon the 454 platform. Sequencing the *R. microplus* genome wholly by the long read PacBio platform was cost-prohibitive within our available resources, thus we adopted a hybrid approach. Koren et al. (2012) reported the sequencing of the 1.2 Gbp parrot genome using a 3.8× PacBio coverage strategy supplemented with a 15.4 × 454-based coverage and error correction of the PacBio reads using a 54× coverage Illumina paired end read data set. We set these parameters as our goals for coverage of the *R. microplus* genome. To obtain sequence information over the entire genome, we utilized two approaches. First, we sequenced unselected very high molecular weight *R. microplus* genomic DNA with PacBio to a coverage of 4.6×. Second, we sequenced 18,432 BAC clones from a BAC library prepared from unselected *R. microplus* genomic DNA using an Illumina-based pooling and reassembly method (FGS). This BAC-based data set produced 0.24× coverage of the genome. To acquire the data set for error correcting the PacBio reads, we sequenced Cot-selected genomic DNA with Illumina HiSeq to a coverage of 15.84×. We wished to focus resources available for error correction upon the gene-rich regions of the genome. This drove our decision to use the HiSeq technology on genomic DNA that had been selected with the Cot technique, as it is enriched for the unique and less repetitive DNA fraction, rather than whole genome-derived DNA. Supplementary Table S1 contains the raw statistics for these sequencing datasets.

Fig. 2 shows the read length distributions for the 13,909,582 PacBio reads. There are 1,446,530 reads ≥5,000 nt and these contain over 10.3 Gbp of sequence data comprising 1.46× genome coverage (Supplementary Table S1). We selected the PacBio reads ≥5,000 nt for error correction using the Cot-selected genomic DNA Illumina sequence data set for this purpose. As the error correction process relies upon alignment to the Illumina short reads, PacBio reads that have no alignment to the short reads will be excluded from the error correction. Trimming also occurs during the error correction, as described below. Losses due to trimming and lack of alignment to short reads resulted in a final output from the error correction process of 1,389,498 reads encoding a total of 7,633,231,856 bases representing a genome coverage of 1.08× (Supplementary Table S1).

During the error correction process with LSC (Au et al., 2012), 5′ and 3′ ends of PacBio reads that are not covered by Illumina reads are trimmed and the output error corrected PacBio reads have lost sequence information. In our error correction process, approximately 78% of the error corrected PacBio reads retained at least 70% of their original length (Supplementary Fig. S2). It is possible that a higher percentage of PacBio read length could have been preserved had the Illumina data been generated from unselected genomic DNA rather than the Cot-selected fraction. However, the unselected genomic DNA consists of ~70% repetitive DNA and it is not clear how the LSC algorithm would handle this type of short read data. Au et al. (2012) reported on LSC in the context of isoform assembly of RNASeq data, which does not typically contain large, highly repetitive sequence regions. This would be an interesting topic for future studies. Another possible route to an assembled genome would be through obtaining a high genome coverage solely from PacBio sequencing. PacBio errors are reported to be randomly distributed (Eid et al., 2009) and sufficient coverage should allow PacBio read redundancy to "self-correct" and produce a highly accurate assembled genome. At the time we acquired our sequence data sets, the yield of PacBio SMRT II technology was not sufficient to allow this approach to be cost-effective.

Following error correction, the 1,389,498 error corrected PacBio reads were screened with the merged scaffolds dataset (Fig. 1) using BLASTN, generating a set of 1,254,669 reads with a total of 7,609,790,717 nt (1.07× coverage) that had significant sequence similarity to the merged scaffolds. The Rmi v1.0 scaffolds, Illumina scaffolds, and these PacBio reads were assembled with the merged scaffold dataset using PBJelly2 (Fig. 1 and Supplementary Table S3). Subsequent to the PBJelly2 assembly, there were 448,651 error corrected PacBio reads that did not assemble. Those remaining 448,651 error corrected PacBio reads were assembled with CANU (Berlin et al., 2014). The resulting CANU scaffolds and singletons were combined with the PBJelly2 scaffolds to create the resulting *R. microplus* assembly, designated Rmi v2.0. This genome assembly is composed of 280,135 contigs in 195,170 scaffolds (N50 = 60,284 bp) representing 2.0 Gbp, including 3.01% of Ns (61.9 Mbp) incorporated into the assembled scaffolds (Table 1). This assembly is a substantial improvement over the first *R. microplus* genome assembly encoding a total of 144.6 Mbp with an N50 of 825 bp (Guerrero et al., 2010). In our present study, we aimed at preferentially enriching and sequencing gene-rich regions of the *R. microplus* genome. The remaining ~5Gbp of unassembled regions of the *R. microplus* genome are anticipated to consist of 60–80% fraction of repeat elements and portions of gene loci that have been partially assembled or missing in the Rmi v2.0 assembly (see Section 3.2).

3.2. Identification of transposable elements and other repeats

Similar to *I. scapularis*, the *R. microplus* genome has been determined to consist of ~70% repetitive DNA (Ullmann et al., 2005). Table 2 shows that the overall repeat content of the Rmi v2.0 genome assembly was 51.44%. This fraction represents 14.56% of the entire *R. microplus* genome (see Section 3.4). Class I transposable elements represented the largest fraction of repeat elements accounting for 28.37% of our genome assembly, while Unclassified, simple repeats and Class II transposable elements encoded 18.16%, 2.21% and 2.12% of the Rmi v2.0 assembly, respectively. The most abundant Class I repeat elements are long interspersed nuclear repeats (LINEs) (11.64%) and long terminal repeats (LTRs) (9.74), while short interspersed nuclear repeats (SINEs) and Class II DNA elements account for 7.00% and 1.66% of our *R. microplus* genome assembly, respectively. Among the SINE elements, we found that the Ruka small interspersed element comprised 69% of the SINE content. The Ruka element was previously found to be specific to ixodid tick species including *Rhipicephalus appendiculatus*, *R. microplus, Amblyomma variegatum* and *I. scapularis* (Sunter et al., 2008). The longest repetitive element in the *R. microplus* assembly was a 55,684 bp simple repeat sequence, $(CTAT)_n$, which was contained in positions 3,396 to 59,680 in a 82,162 bp long genomic scaffold (RMI_v2_068324). The repeat content of subsets of datasets generated for this study are shown in Supplementary Table S4.

Our three Cot-selected datasets, derived from genomic DNA selected for the unique and low copy fractions of the genome, contain 24.6% (454 low-Cot data; Rmi v1.0), 28.4% (Illumina low-Cot; Ray assembly) and 32.2% (Illumina low-Cot; SOAPdenovo assembly), while our two whole genome shotgun datasets, the BAC and PacBio CANU assembled datasets, contain 59.75% and 60.0% repeat elements, respectively (Supplementary Table S4). Our findings that the genome, as reflected by the BAC and PacBio datasets, contained 54% to 60% repeat element content is generally consistent with the

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

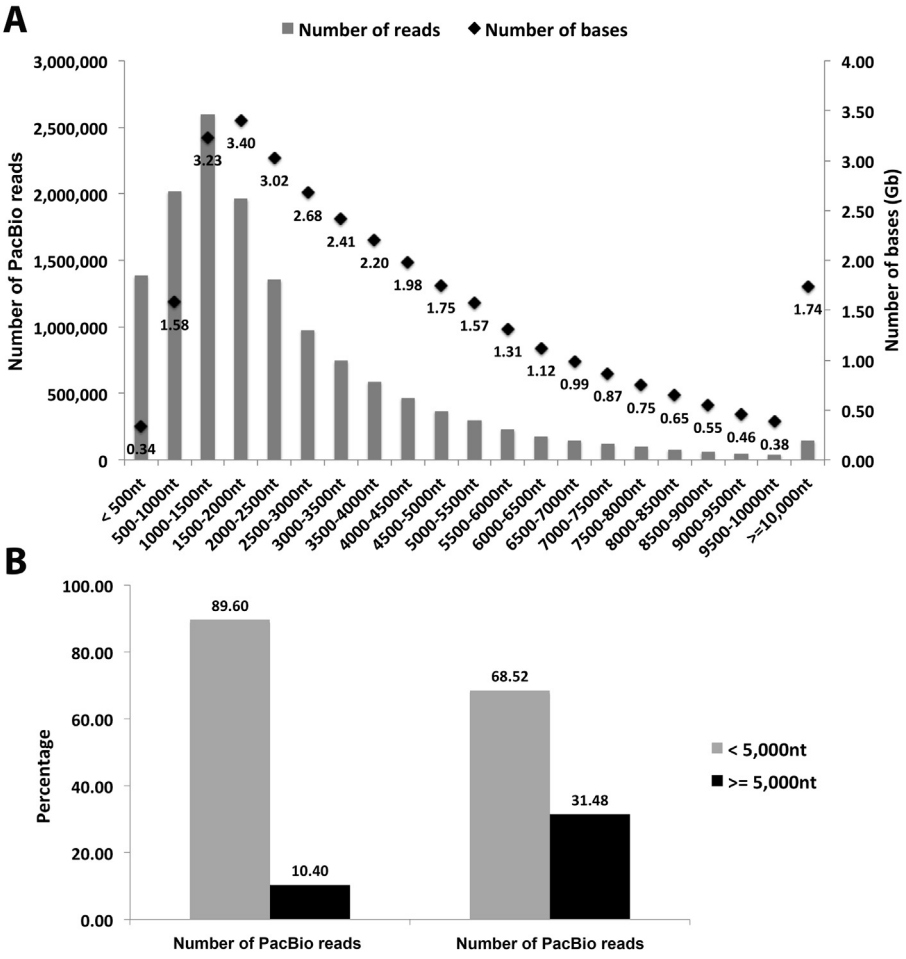

Fig. 2. Length distribution and base content of 13,909,582 PacBio reads. (A) Number of PacBio reads and their encoded bases are shown for each length interval. (B) PacBio reads were divided into two pools (<5,000 nucleotides (nt) and ≥5,000 nt). The percentages of PacBio reads and their corresponding total percentage of bases are shown for each pool.

Table 1
Genome assembly statistics of the cattle tick Rmi_v2.0 genome.

Statistics	Rmi v2.0
Number of contigs used in assembly	280,135
Number of resulting scaffolds	195,170
Total nt in scaffolds	2,009,073,859
Estimated fold coverage of the assembly	0.28×
Longest scaffold	432,897
Shortest scaffold	500
Number of scaffolds >1 K nt	191,944
Number of scaffolds >10 K nt	33,972
Number of scaffolds >100 K nt	5,390
Number of scaffolds >1 M nt	0
Mean scaffold size	10,294
Median scaffold size	2,574
N50 scaffold length	60,284
L50 scaffold count	9,005
Scaffold% A	26.66
Scaffold% C	21.76
Scaffold% G	21.8
Scaffold% T	26.67
Scaffold% N	3.08
Scaffold% non-ACGTN	0.02
Number of scaffold non-ACGTN nt	449,760

re-association kinetics-based result of ~70% repetitive DNA in the *R. microplus* genome. There is the possibility that repeat-rich regions may comprise a larger fraction of the *R. microplus* genome than our data indicates, as these elements may have been collapsed into fewer highly similar repeat elements during the assembly process. Additionally, our PacBio dataset used to produce Rmi v2.0 is biased toward the less repetitive fraction of the genome, since our error correction Illumina data was derived from Cot-selected DNA. Thus, our estimate of 54–60% repeat element content for the *R. microplus* genome is likely a conservative figure (Table 2).

Gulia-Nuss et al. (2016) reported that 16.7% of the *I. scapularis* genome consisted of transposable elements, including Non-LTR retrotransposons, DNA transposons, and LTR retrotransposons that comprised 6.7%, 3.06%, and 0.64% of that tick's genome, respectively. Comparing the repeats found in the *R. microplus* genome with those reported in the *I. scapularis* genome, we found there is a larger fraction of the *R. microplus* genome as LTR retrotransposons (2.76%), and a lower fraction of DNA transposons (0.47%). The LTR retrotransposon *Gypsy* has over 233,000 copies in the *R. microplus* Rmi v2.0 assembly while the *I. scapularis* genome contains approximately 29,000 copies. As *Gypsy* is known to be infectious in invertebrates (Kim et al., 1994), the role of the retrotransposon in *R. microplus* genome evolution would be an interesting topic for further study. Nystedt et al. (2013) reported the Norway spruce genome size estimate of 19.6 Gbp, with 58% of the genome consisting of LTR retrotransposons and 35% of this consisting of *Gypsy* repeat elements. The large genome size of the Norway spruce and other conifers was proposed to be due to expansions of LTR elements such as *Gypsy* combined with conifers' inefficient mechanisms to inactivate and remove transposable elements from their genomes.

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

Table 2
Distribution of repeat elements in the cattle tick Rmi v2.0.

Repeat class	Classification	*Sub-classification*	No. of eements	Total bases	Assembly coverage (%)	Genome coverage (%)
Class I transposable element						
	SINE		662,258	140,632,871	7.00	1.98
		RUKA	*415,351*	*97,142,944*	*4.84*	*1.37*
		Other/Unknown	*246,907*	*43,489,927*	*2.16*	*0.61*
	LINE		469,463	233,777,102	11.64	3.29
		I	*136,757*	*88,721,046*	*4.42*	*1.25*
		L1	*29,189*	*14,574,490*	*0.73*	*0.21*
		L2	*41,671*	*15,392,611*	*0.77*	*0.22*
		L3/CR1	*91,827*	*35,517,458*	*1.77*	*0.50*
		Penelope	*45,428*	*14,836,907*	*0.74*	*0.21*
		Jockey	*11,056*	*6,450,861*	*0.32*	*0.09*
		Other/Unknown	*113,535*	*58,283,729*	*2.90*	*0.82*
	LTR		258,860	195,585,817	9.74	2.76
		Gyspy	*233,185*	*185,724,131*	*9.24*	*2.62*
		Pao	*17,014*	*7,838,699*	*0.39*	*0.11*
		Other/Unknown	*8,661*	*2,022,987*	*0.10*	*0.03*
Class II transposable element						
	DNA Elements		94,243	33,351,492	1.66	0.47
		hAT	*21,883*	*5,438,102*	*0.27*	*0.08*
		Mariner	*1,715*	*890,553*	*0.04*	*0.01*
		Harbinger	*3,194*	*2,285,817*	*0.11*	*0.03*
		PiggyBac	*652*	*151,162*	*0.01*	*0.00*
		Other/Unknown	*66,799*	*24,585,858*	*1.22*	*0.35*
	Rolling-circle/Helitron		38,128	9,281,718	0.46	0.13
Unclassified			1,586,243	364,944,019	18.16	5.14
Small RNA			12,786	2,351,195	0.12	0.03
Satellite			19,650	8,158,536	0.41	0.11
Simple repeat			437,017	44,342,020	2.21	0.62
Low complexity			21,420	1,110,781	0.06	0.02
Total				1,033,535,551	51.44	14.56

SINE, short interspersed nuclear elements; LINE, long interspersed nuclear elements; LTR, long terminal repeat.

3.3. Identification of *R. microplus* gene loci

We identified 38,827 putative gene loci in the Rmi v2.0 assembly derived from both mapping non-redundant *R. microplus* transcripts onto the genome assembly and ab initio gene prediction using MAKER (Cantarel et al., 2008), SNAP (Korf, 2004) and AUGUSTUS (Stanke et al., 2006) (Supplementary Fig. S1). We identified 24,758 protein coding gene loci (≥100aa), of these 9,265 (37.42%), 8,867 (35.81%), 2,301 (9.29%) and 4,325 (17.47%) were classified by TransDecoder (Tang et al., 2015) as 'complete', '5 prime partial', '3 prime partial' and 'internal', respectively (Supplementary Table S5, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1).

Overall we found 17,297 (44.55%) and 21,530 (55.45%) gene loci encoded by a single-exon or multiple exons, respectively (Supplementary Table S6). Gene loci classified as 'complete' showed the least proportion of single-exon loci (24.73%), while more than half (54.38%) of the gene loci classified as 'internal' were single-exon (Supplementary Table S6). The average length of exons was similar for gene loci classified as 'complete', '5 prime', 'internal' and '3 prime' ranging from 277.6 bp to 410 bp indicating a consistent exon structure across all gene loci (Supplementary Table S6 and Fig. S3).

Nucleotide sequences for identified *R. microplus* gene loci were annotated using AutoFACT annotation pipeline (Koski et al., 2005) and by BLASTN comparison to the RFAM database. Table 3 summarises the classification of 38,827 *R. microplus* gene loci into sequences with similarity to 'known genes' (47.84%), 'domain-containing proteins' (3.04%), 'conserved hypothetical genes' (4.90%) with hits to sequences in public databases with an unknown function, and 'hypothetical genes' (44.22%) encompassing sequences only found in the Rmi v2.0 genome assembly (Supplementary Table S5, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1).

Table 3
Annotation of protein coding and RNA genes in the cattle tick Rmi v2.0 assembly.

Classification	All loci (% loci)	ORF≥100aa	RNA genes
Known genes	18,575 (47.8%)	16,520	506
Domain-containing protein	1,181 (3.0%)	600	0
Conserved hypothetical genes	1,901 (4.9%)	1,457	0
Hypothetical genes	17,170 (44.2%)	6,181	0
Total	38,827 (100%)	24,758	506

3.4. Evaluation of completeness of *R. microplus* genome

To estimate genome completeness, we examined the 38,827 putative Rmi v2.0 gene loci described above for highly conserved single-copy protein coding genes that are found in nearly all arthropods using BUSCO analysis software (Simao et al., 2015). BUSCO indicated 40.1% genome completeness based on 2,675 ancestral proteins used in the analysis (Supplementary Table S5, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). Among the identified ancestral arthropod genes found by BUSCO there were 794 (29.7%) and 278 (10.4%) *R. microplus* gene loci classified as complete and fragments, respectively (Supplementary Table S5, doi: http://dx.doi.org/10.17632/s7jrdzfmb7.1). These findings generally correlate with the current length of the Rmi v2.0 assembly (2.0 Gbp) that represents ~28% of the estimated 7.1 Gbp *R. microplus* genome.

We refined our BUSCO analysis to attempt to find more of the missing 1,603 (59.9% of the total BUSCO genes) ancestral arthropod genes in the Rmi v2.0 assembly. It is possible that Metastriate ticks such as *R. microplus* possess genes with distant sequence similarity to the BUSCO ancestral set. It is also possible that the selective low-Cot DNA sequencing approach may result in partial assembly of

gene loci such that our BUSCO analysis missed those. Additionally, the RNA-seq datasets used in this study may not encompass the expression of all *R. microplus* counterparts to ancestral proteins and/or the stringency of the computational pipeline resulted in the identification of only a limited fraction of the corresponding ancestral gene loci (Supplementary Fig. S1). To evaluate whether any of the missing ancestral arthropod proteins could be found in our RNA-seq dataset, we subjected our 63,416 non-redundant *R. microplus* transcriptome dataset, described earlier and in Supplementary Table S2, to BUSCO analysis. We identified hits to 2,274 ancestral proteins (85.01% of the total BUSCO ancestral arthropod gene dataset) with only 401 ancestral proteins (~15% of the total BUSCO dataset) missing in our *R. microplus* transcriptome dataset (Supplementary Fig. S4). This result suggests that *R. microplus* counterparts to ancestral proteins are not highly divergent in sequence, but more likely the 1,603 ancestral proteins missing in our Rmi v2.0 genome assembly might be due to a lack of scaffolds encoding these ancestral proteins and/or existing scaffolds accounting for less than 50% coverage of the BUSCO gene sequences. We re-aligned 1,919 RNA-seq transcripts that had not aligned onto the Rmi v2.0 genome assembly and had hits to 1,125 of the 1,603 missing BUSCO ancestral proteins (Supplementary Fig. S4). We also lowered the transcript alignment coverage threshold to ≥30%, to help identify partial gene loci on the Rmi v2.0 assembly for these transcripts. We found partial mapping positions for 507 of these *R. microplus* transcripts onto 445 putative genomic locations on the Rmi v2.0 assembly (Supplementary Table S7, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). These partially mapped transcripts have hits to 349 missing ancestral arthropods genes (Supplementary Fig. S4). Thus, combining results from both BUSCO analyses, we have identified 1421 hits to BUSCO arthropods ancestral proteins (53.1% of the total BUSCO dataset) based on 24,758 protein coding gene loci identified in Rmi v2.0 and 445 from our partial mapping strategy. This finding suggests that ~47% of the arthropods ancestral proteins are missing in the ~2.0 Gbp Rmi v2.0 genome assembly.

3.5. Genome-wide annotation

Representative sequences identified for 38,827 *R. microplus* gene loci (Supplementary Fig. S1) were annotated using AutoFACT (Koski et al., 2005). Of that total 18,575 (47.84%), 1,181 (3.04%) and 1,901 (4.90%) gene loci had significant sequence similarity to known genes, domain-containing proteins, and conserved hypothetical proteins, respectively (Table 3 and Supplementary Table S5, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). We also identified 17,170 (44.22%) hypothetical gene loci with no sequence similarity to available gene sequences from any species (Supplemental Table S5; Table 3).

Functional annotation of representative *R. microplus* sequences identified 1,153 gene loci associated with 136 KEGG pathways (Supplementary Table S5, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). Comparing the KEGG pathways found in *R. microplus* to those in *I. scapularis*, 970 of these gene loci assigned to 91 KEGG pathways are shared with *I. scapularis* (Supplemental Fig. S5A), while 45 and 14 KEGG pathways were unique to *R. microplus* or *I. scapularis*, respectively. Among the top overrepresented KEGG pathways in *R. microplus* we found a 12-fold, 11-fold and eightfold increase in genes assigned to the 'Neuroactive ligand-receptor interaction', TGF-beta signalling pathway and ubiquitin mediated proteolysis, while *I. scapularis* showed 1.6-fold increase in genes assigned to the Notch signalling pathway (Supplementary Fig. S5B). Comparison of genes involved in neuroactive ligand-receptor identified two and a single putative 'Partitioning-defective 3-like beta' (*PAR3B*) gene in the Rmi v2.0 (Rmi_v2_LOC_023216.1 and Rmi_v2_LOC_023217.1) and IscaW1 (XP_002404429.1) genome assemblies, respectively. We also found two putative PAR3 isoform X11 genes in *R. microplus* (Rmi_v2_LOC_023009.1 and Rmi_v2_LOC_023010.1) encoded in the same genomic region as sense-antisense transcripts. Owing to the lack of strand information in the RNA-seq datasets used in the gene loci prediction approaches, caution needs to be exercised when evaluating sense-antisense genes with non-identical ORFs. Assessment of annotated putative *R. microplus PAR3B* and *PAR3X11* genes for both gene locus and their encoded ORF orientation shows that there is likely a single *PAR3B* (Rmi_v2_LOC_023217.1) and *PARX11* (Rmi_v2_LOC_023010.1) gene in Rmi v2.0 genome assembly. The *I. scapularis* IscaW1 genome assembly was not reported to encode a *PAR3X11* gene. The diverse *PAR3* gene family has been shown to play a role in asymmetrical cell division and direct polarized cell growth from nematodes to vertebrates (Kohjima et al., 2002).

Another key difference found is the lack of a putative 'Glutamate receptor' in *I. scapularis*, while in the Rmi v2.0 genome assembly we identified four putative metabotropic glutamate receptors (Rmi_v2_LOC_007400.1, Rmi_v2_LOC_019725.1, Rmi_v2_LOC_019727.1 and Rmi_v2_LOC_034524.1) with most of them being classified as 'Complete' proteins by TransDecoder. Glutamate receptors are synaptic receptors located primarily on the membranes of neuronal cells including the larval neuromuscular junction (Thomas and Sigrist, 2012). The use of glutamate receptor inhibitors against insect skeletal muscle glutamate receptors has been suggested as a potential insecticide. It remains to be elucidated whether one or more of the putative *R. microplus* glutamate receptors identified in this study is uniquely expressed in skeletal muscle, providing new opportunities for the controlling *R. microplus* populations.

The AutoFACT program assigned cluster of orthologous genes (COG) annotation to 4,143 transcripts that mapped onto the Rmi v2.0 assembly (Supplementary Table S5, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). Comparison of *R. microplus* COG annotation against that of *I. scapularis* identified a 13.7 and 4.3-fold increase in 'Replication, recombination and repair' and 'Chromatin structure and dynamics', respectively (Supplementary Fig. S5C). Among the gene loci with COG annotation of 'Replication, recombination and repair function', we found 305, 147 and 132 genes annotated as 'gag-pol fusion proteins', 'putative reverse transcriptase' and 'rve-domain containing protein', respectively. Inspection of gag-pol fusion protein gene loci show that most of them represent partial sequences, while only 33 gag-pol fusion proteins were classified as 'complete' by TransDecoder (Supplementary Table S5, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). BLASTP comparison of the complete gap-pol fusion proteins against a reference *Rhizophagus irregularis* gag-pol fusion protein (EXX59955.1) revealed that 25 putative *R. microplus* gag-pol proteins had varying degrees of similarity to the reference EXX59955.1 gag-pol protein, highlighting that these were divergent and likely to represent independent genome-insertion events of tick viral sequences (Supplementary Table S8, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). The abundant repertoire of viral-like sequences in *R. microplus* may facilitate its innate antiviral response via the RNA interference (RNAi) machinery that is common to insects and plants (Gammon and Mello, 2015; Barrero et al., 2017).

Supplementary Fig. S6 shows a substantial increase in *R. microplus* gene loci with Gene Ontology (GO) annotation in Rmi v2.0 compared with the Rmi v1.0 release (Guerrero et al., 2010). Rmi v1.0 had 338 genes with GO annotation while Rmi v2.0 contains 9,169 genes with GO annotation, showing a significant increase in nearly all annotated cellular component, molecular function and biological process GO categories (Supplementary Fig. S6 and Table S9, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). Supplementary Fig. S6 depicts a comparison of *R. microplus* GO genome-wide annotations against that of *I. scapularis* which contains GO codes assigned to 6,142 protein coding genes. Notably we found

a 5.47, 5.0, 4.13 and 3.96-fold increase in 'protein-DNA complex', 'DNA packing', 'chromatin binding' and 'nucleic acid binding' GO codes in *R. microplus* Rmi v2.0 compared with *I. scapularis* IscaW1 (Supplementary Tables S9, S10, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). One of the key families of proteins accounting for the observed differences between *R. microplus* and *I. scapularis* is histone proteins (Supplementary Table S10, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). We identified 58 histone proteins in *R. microplus* including 45 predicted to encode a complete ORF by TransDecoder (Tang et al., 2015) (Supplementary Table S10, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). Among the complete histone proteins we found eight, eight and 29 annotated as H2A, H3 and H4, respectively, while in *I. scapularis* were identified one, four, three and three proteins annotated as H1, H2A, H2B and H3, respectively (Supplementary Table S10, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). Evolutionary relationship analysis of H2 and H3 proteins commonly found in the two tick species, showed that *R. microplus* encodes three clusters of histone H2 proteins, one (Rmi_v2_LOC_036667.1, Rmi_v2_LOC_036900, Rmi_v2_LOC_023192.1, Rmi_v2_LOC_03694.1) clustering with *I. scapularis* H2A (XP_002403551.1) and another (Rmi_v2_LOC_001119.1) with *I. scapularius* H2B (XP_002436057.1 and XP_002400137.1) proteins, while the third cluster (Rmi_v2_LOC_036898.1, Rmi_v2_LOC_012238.1 and Rmi_v2_LOC_017065.1) is more divergent (Supplementary Fig. S6A). Similarly, annotated histone 3 (H3) proteins also show three phylogenetic branches, where Rmi_v2_LOC_036654.1 is the more divergent H3 protein (Supplementary Fig. S7). Variants of nuclear core histones H2A and H3 have been reported to have specific functions in the regulation of gene expression and genome stability (Yukawa et al., 2014). Our findings in *R. microplus* correlates with the sequence diversity of H2A and H3 found in other species.

Absence of an *R. microplus* histone H1, a gene family known to be highly variable across species that is involved in the establishment of pericentric heterochromatin and normal chromatic structure in *D. melanogaster* (Lu et al., 2009), maybe due to the partial Rmi v2.0 genome assembly. Interestingly, we found a large expansion of the histone H4 family in *R. microplus*, which is histone family that is lost in the *I. scapularis* lineage (Gulia-Nuss et al., 2016). Mutations in the yeast histone H4 in which all four tail lysines are replaced by glutamines cause a pronounced defect in genome integrity owing to failure to repair damaged DNA (Bird et al., 2002). The significantly larger genome of *R. microplus* (~7.1 Gbp) compared with *I. scapularis* (2.1 Gbp) may potentially explain the need for a large repertoire of histone H4 proteins.

3.6. Discovery of novel R. microplus miRNAs

Previously we reported the finding of 87 *R. microplus* miRNA loci (Barrero et al., 2011a) using the genomes of *I. scapularis*, *D. melanogaster* and the Rmi v1.0 assembly derived solely from Cot-selected *R. microplus* genome sequences. Only 24 of the precursor sequences for the 87 predicted *R. microplus* miRNAs could be found in the Rmi v1.0 assembly. With the significant increase in the content of the *R. microplus* assembly from Rmi v1.0 (0.15 Gbp) to Rmi v2.0 (2.0 Gbp), we reanalyzed the *R. microplus* genome using Rmi v2.0 to identify miRNA loci, using small RNA libraries derived from *R. australis* (formerly known as *R. microplus*) life stages (egg, larvae, frustrated larvae exposed to the host but not allowed to feed, adult females and males) and selected adult female ticks organs (salivary glands, midgut, and ovaries). miRNAs are highly conserved throughout evolution facilitating the use of genomes from related species and/or small RNA datasets to discover miRNA loci in a species of interest (Barrero et al., 2011a). We found 191 non-redundant *R. microplus* miRNAs in the Rmi v2.0 assembly (Supplementary Table S11, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). Of these, 169 (88.5%) and 22 (11.5%) miRNAs have sequence similarity to known miRNAs (Barrero et al., 2011a) or represent novel tick microRNA loci, respectively. Previously we identified 24 *R. microplus* miRNA precursors in the Rmi v1.0 genome assembly (Barrero et al., 2011a). Additionally, the mature sequences for other 63 miRNAs were found by using the *D. melanogaster* and *I. scapularis* genomes (Barrero et al., 2011a). Thus, the finding of 191 miRNA loci in the Rmi v2.0 genome assembly represents a substantial improvement. Evaluation of the expression profiles of these microRNAs discovered that 143 and 93 were expressed in our life stage and adult tick samples, respectively (Fig. 3). The majority of the identified microRNAs show specific expression restricted to either one specific life stage (Fig. 3A), sex or adult female organ (Fig. 3B). This contrasts with our previous report (Barrero et al., 2011a), where most of the identified *R. microplus* miRNAs were found expressed in multiple life stages and/or organs. In that previous study, owing to the small and incomplete nature of the Rmi v1.0 *R. microplus* genome assembly, we relied upon the *I. scapularis* (IscaW1.0) and *D. melanogaster* genomes for miRNA prediction in *R. microplus*. We had also imposed the requirement to only report

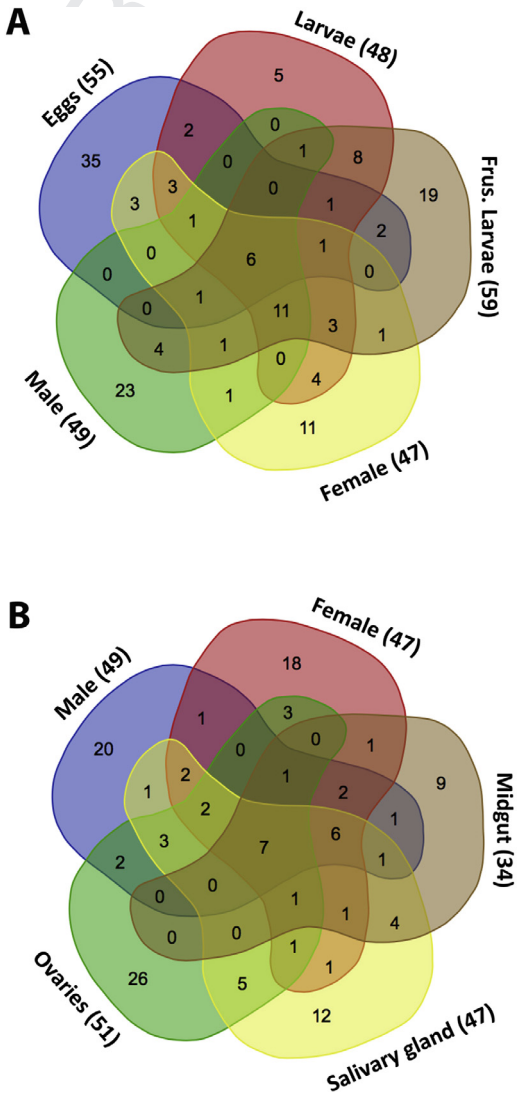

Fig. 3. Global expression of cattle tick microRNAs. Expression of miRNAs among (A) life stages, and (B) adult tick samples are shown. Frus. Larvae, 'frustrated larvae' exposed to the cattle for 6 h but not allowed to feed on them. Organs (salivary glands, midgut and ovaries) were collected from semi-engorged adult female cattle ticks (Barrero et al., 2011).

miRNAs that were expressed in at least two independent small RNA libraries. Thus, our previous findings were likely biased towards evolutionarily conserved miRNAs. The identification of life stage-specific and/or organ-specific cattle tick miRNAs provides new opportunities for investigating the role of miRNAs in tick gene expression and development.

3.7. Conservation of cattle tick protein-coding genes

To evaluate the conservation of *R. microplus* protein-coding genes (ORF ≥100 amino acids in length) compared with their counterparts in *I. scapularis*, insects, crustacean and vertebrates species, the proteomes of 12 species were selected including *I. scapularis*, *Anopheles gambiae*, *A. mellifera*, *B. taurus*, *C. elegans*, *Drosophila melanogaster*, *T. castaneum*, *Drosophila pulex*, *D. rerio*, *Gallus gallus* and *H. sapiens*. The various proteome sizes are listed in Table 4. Comparison of shared protein families was conducted using OrthoMCL (Li et al., 2003). This analysis clustered 415,232 protein-coding genes into 36,405 protein families (Table 5). We found a core set of 1,254 protein families shared amongst all species included in our analysis (Table 5). The core set of protein families contains 47,031 proteins. Within this core protein set, there were 2,405 and 1,604 *R. microplus* and *I. scapularis* proteins representing 9.43% and 7.84% of their proteome content, respectively (Table 4). Interestingly, the species with the largest proportion of its proteome as part of the core set was *A. mellifera* with 15.70% of its protein sequences. Our results suggest that 8% - 9% of tick genes are conserved since the Nephrozoan (640 million years ago) ancestor (Barrero et al., 2011a).

In the OrthoMCL analysis we identified 1697, 647 and 810 protein families unique to *R. microplus*, *I. scapularis* or to both tick species (Table 5; Supplementary Table S12, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). The unique protein families found in *R. microplus* comprised 10,835 proteins representing 43.76% of its proteome of 24,758 proteins. In contrast, we identified 2014 unique *I. scapularis* proteins accounting for 9.84% of its proteome set (Supplementary Table S12, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). These subsets of tick specific genes will facilitate functional genomic analyses aiming to understand the unique biology of ticks.

3.8. Diverse genomic loci of R. microplus esterases

Increased esterase activity in arthropods is considered an important biochemical marker of pesticide resistance. In the *R. microplus*, esterases were found associated with resistance to both organophosphate (OP) and pyrethroid pesticides (Rosario-Cruz et al., 2009). Temeyer and colleagues reported the limited aa sequence similarities among the three characterized acetylcholinesterases (AChEs) from *R. microplus* (Temeyer et al., 2004). More recently, 27 esterase-like transcripts were found in *R. microplus* transcriptome datasets (Bendele et al., 2015). We found 31 *R. microplus* gene loci with similarity to AChE as indicated by the AutoFACT description (Supplementary Table S5, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1); of these 13 gene loci were assessed by TransDecoder as encoding complete proteins (Supplementary Table S5, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). AChEs are large proteins encoding an average ~600 aa; the shortest annotated AChE in *I. scapularis* encodes 464 aa (XP_002409706.1). Of the 13 *R. microplus* gene loci predicted to encode 'complete' AChEs, only eight have an ORF larger than 500 aa, while the remaining five gene loci encoded AChEs with an ORF shorter than 370 aa. These findings highlight the need to take with caution the ORF classification assigned using TransDecoder. To provide insights into the exon–intron structures and functional domains of *R. microplus* AChEs four divergent gene loci (Rmi_v2_LOC_008474.1, Rmi_v2_LOC_024806.1, Rmi_v2_LOC_000817.1 and Rmi_v2_LOC_014526.1) were selected and evaluated. Our results show significant gene structure diversity in terms of location of the ORF and size of both exon and introns (Fig. 4). Despite such gene structure divergence, these esterases are predicted to have similar

Table 5
OrthoMCL summary of shared protein clusters amongst 12 species.

	Clusters/protein families	Total No. of proteins
Core protein set conserved in all 12 species	1254	47031
R. microplus and *I. scapularis* only	810	4634
R. microplus only	1697	10835
I. scapularis only	647	2014
Insects only	371	3170
Insects and ticks only	25	312
Insects and *I. scapularis* only	33	323
Insects and *R. microplus* only	1	6
Insects + *D. pulex* only	220	2205
Insects + *C. elegans* only	12	129
Insects + *D. pulex* + *C. elegans* only	28	455
Mammals only	2796	18452
Mammals + *G. gallus* only	1025	9339
Mammals + *G. gallus* + *D. rerio* only	3795	51224
Mammals + *G. gallus* + *D. rerio* + *D. pulex* only	112	1905
Mammals + *D. rerio*	672	6293
Human + *I. scapularis* only	22	78
Human + *R. microplus* only	7	31
Human + ticks only	1	5
Cattle + *R. microplus* only	0	0
Cattle + *I. scapularis*only	3	8
Cattle + ticks only	0	0

Table 4
Proteomes analysed by OrthoMCL.

Species	Total No. of proteins (No. of proteins ≥100aa)	Core Conserved proteins (No.)[a]	Core conserved proteins (% of total proteome)
Anopheles gambiae	14,099 (13,371)	1872	13.28
Apis mellifera	21,772 (21,373)	3418	15.70
Bos taurus	51,914 (50,971)	5565	10.72
Caenorhabditis elegans	27,876 (25,738)	2754	9.88
Drosophila melanogaster	30,277 (28,944)	4473	14.77
Daphnia pulex	30,611 (26,331)	2011	6.57
Danio rerio	47,504 (46,827)	5432	11.43
Gallus gallus	32,134 (31,560)	3924	12.21
Homo sapiens	98,125 (96,747)	10979	11.19
Ixodes scapularis	20,467 (16,642)	1604	7.84
Rhipicephalus microplus	24,758 (24,758)[b]	2405	9.43
Tribolium castaneum	18,076 (17,688)	2710	14.99

[a] Number of proteins in this species' proteome that occur in the set of 1,254 core conserved protein families.
[b] Number of proteins-coding genes with a predicted ORF of at least 100 amino acids.

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

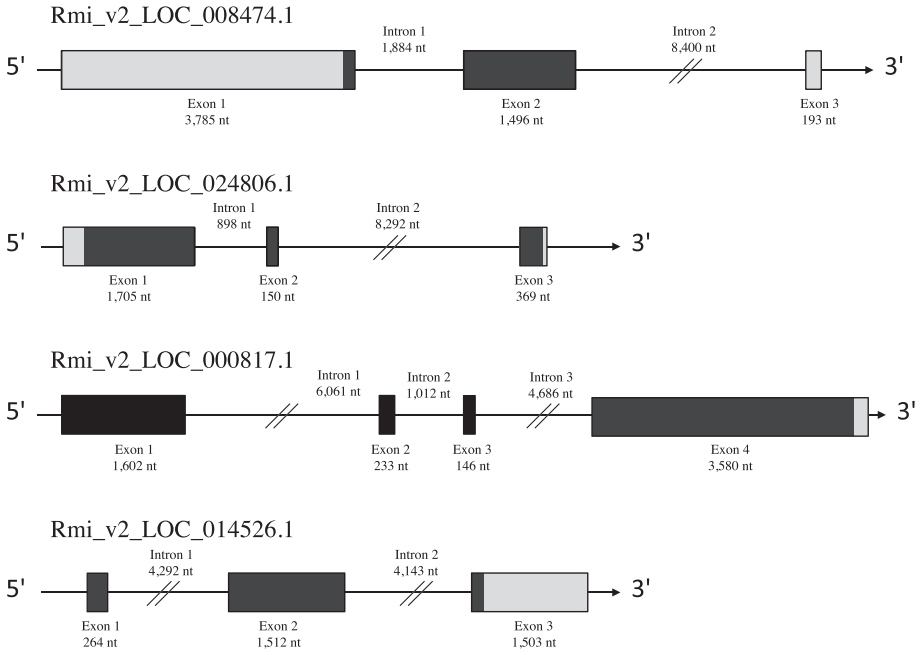

Fig. 4. Gene structure of *Rhipicephalus microplus* transcripts encoding predicted full-length protein with top BLASTX hit to acetylcholinesterase. Exons and introns are not drawn to scale. Size of exons and introns are indicated in nucleotides. The 5′and 3′untranslated regions are denoted as grey boxes. The open reading frames are denoted as black boxes.

protein secondary and tertiary structures (Supplementary Fig. S8, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). We also evaluated the conservation of functional domains in these putative esterases by comparing them against published AChE protein sequences (Baxter and Barker, 1999; Bendele et al., 2015). We found the active site serine (S), glutamate (E) and histidine (H) that form the AChE catalytic triad are conserved in Rmi_v2_LOC_008474.1, Rmi_v2_LOC_014526.1 and Rmi_v2_LOC_000817.1, but not in Rmi_v2_LOC_024806.1 (Supplementary Fig. S9). Additionally, aas predicted to form part of the substrate-binding pocket and post-translational modification motifs showed variability among all sequences suggesting that these esterases may undergo distinct post-translational modification events.

Interestingly, we found that the proteins encoded by Rmi_v2_LOC_008474.1 and Rmi_v2_LOC_000816.1 harbour a signal peptide in their N-terminal region, while no such domain was found in the ORF encoded by Rmi_v2_LOC_014526.1 and Rmi_v2_LOC_024806.1 (Supplementary Fig. S10). To gain insight into the possible function of Rmi_v2_LOC_024806.1, we conducted a BLASTP screening against a manually curated Swiss-Prot human database. We identified a Butyrylcholine esterase (P06276) as its top hit with 32% sequence identity and 81% query coverage. This finding contrasts with the top match found in the Swiss-Protein database for proteins encoded by the Rmi_v2_LOC_008474.1 and Rmi_v2_LOC_000817.1 representative gene sequences. The Rmi_v2_LOC_008474.1 protein showed 34% protein sequence similarity and 95% query coverage to a human AChE protein (accession number P04058), while the Rmi_v2_LOC_000817.1 encoded-protein had 30% protein sequence similarity and 90% query coverage to a different the human AChE (GenBank accession number Q27677). Thus, both Rmi_v2_LOC_008474.1 and Rmi_v2_LOC_000816.1 transcripts and their corresponding gene loci appear to encode secreted *R. microplus* AChEs.

A Neighbour-Joining phylogenetic analysis was used to estimate the evolutionary relationship of the Rmi_v2_LOC_008474.1, Rmi_v2_LOC_014526.1, Rmi_v2_LOC_024806.1 and Rmi_v2_LOC_000817.1 encoded esterases (Supplementary Fig. S11). We found that the AChE encoded by Rmi_v2_LOC_008474.1 is part of a sister branch to *R. microplus* AChE1 (Bendele et al., 2015). Curiously, the esterase encoded by Rmi_v2_LOC_024806.1 is placed as an outgroup for the AChE protein encoded by Rmi_v2_LOC_000817.1 and other *I. scapularis* proteins annotated as acetylcholinesterases (Supplementary Fig. S11). The putative AChE encoded by Rmi_v2_LOC_014526.1 showed closer similarity to an *I. scapularis* AChE (XP_002402742.1) protein and a cluster of *R. microplus* AChE3 proteins. Overall our findings highlight the significant diversity of AChE gene-structures that can be utilised to design targeted control strategies to impair tick OP resistance.

3.9. Identification of *R. microplus* CYP superfamily genes

CYPs in animals fall into two categories, namely, those that synthetize or metabolize endogenous molecules and those that interact with exogenous chemicals from the diet or environment (Baldwin et al., 2009). The latter form a critical component of detoxification systems including resistance to OP coumaphos, which is the only acaracide approved for use in the U.S. Department of Agriculture (USDA) Animal and Plant Health Inspection Service (APHIS). Veterinary Services (VS) quarantine dipping vats and spray boxes along the Texas (USA)-Mexico border (Guerrero et al., 2007).

We found 56 *R. microplus* gene loci with similarity to CYP (Supplementary Table S5, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). Of these, 20 were classified as encoding 'complete' protein sequences by TransDecoder analysis (https://github.com/TransDecoder/TransDecoder/releases). We identified two, 14 and four Cyp2, Cyp3 and Cyp4 clan protein coding genes, respectively (Supplementary Table S13, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). Interestingly, among the identified *R. microplus* CYP genes, we found two neighbouring gene loci (Rmi_v2_LOC_001286.1 and Rmi_v2_LOC_001287.1) with top similarity to *CYP41*, a cyp gene reported to be associated with OP resistance (Crampton et al., 1999; Guerrero et al., 2006). Despite having similar gene locus

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

sizes, both putative *R. microplus CYP41* genes show distinct exon–intron structures (Supplementary Fig. S12), suggesting that they were not acquired via a tandem segmental duplication (Gotoh, 1993). Phylogenetic analysis of *R. microplus CYP* genes together with known counterparts in *Daphnia*, honeybee and *Drosophila*, showed agreement with previous studies reporting four major *CYP* clans for arthropod species (Baldwin et al., 2009) (Fig. 5 and Supplementary Fig. S13, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). Both *R. microplus* CYP41 proteins clustered with CYP3 clan proteins including Rmi_v2_LOC_001284.1, Rmi_v2_LOC_006171.1, Rmi_v2_LOC_035583.1, Rmi_v2_LOC_0021156.1, Rmi_v2_LOC_007826.1, and Rmi_v2_LOC_021651.1 (Supplementary Fig. S13, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1).

Multiple sequence alignment of *R. microplus* CYP3 proteins including CYP41 identified five basic and relatively conserved motifs that were arranged from the N-terminal to the C-terminal as follows: helix C, helix I, helix K, PERF and heme-binding motifs (Supplementary Fig. S14). These motifs are consistent with domains found in *Plutella xylostella* cytochrome P450 proteins (Yu et al., 2015). Interestingly, we found that one of the *R. microplus* CYP41 (Rmi_v2_LOC_001287.1) and one *R. microplus* CYP3A (Rmi_v2_LOC_035583.1) protein lack the PERF domain (Supplementary Fig. S15). The absence of this motif may impact the overall folding structure of the CYP protein and its interaction with either endogenous and/or exogenous compounds.

Binding sites and active sites of proteins and DNAs are often associated with structural pockets and cavities. We used CASTp (http://sts.bioe.uic.edu/castp/) to estimate the area and volume of predicted structural pockets in selected CY3A and CYP41 proteins. We found that proteins lacking the PERF motif (Rmi_v2_LOC_001287 and Rmi_v2_LOC_035583) and/or substrate-biding sites in the N-terminal region of the CYP protein (Rmi_v2_-

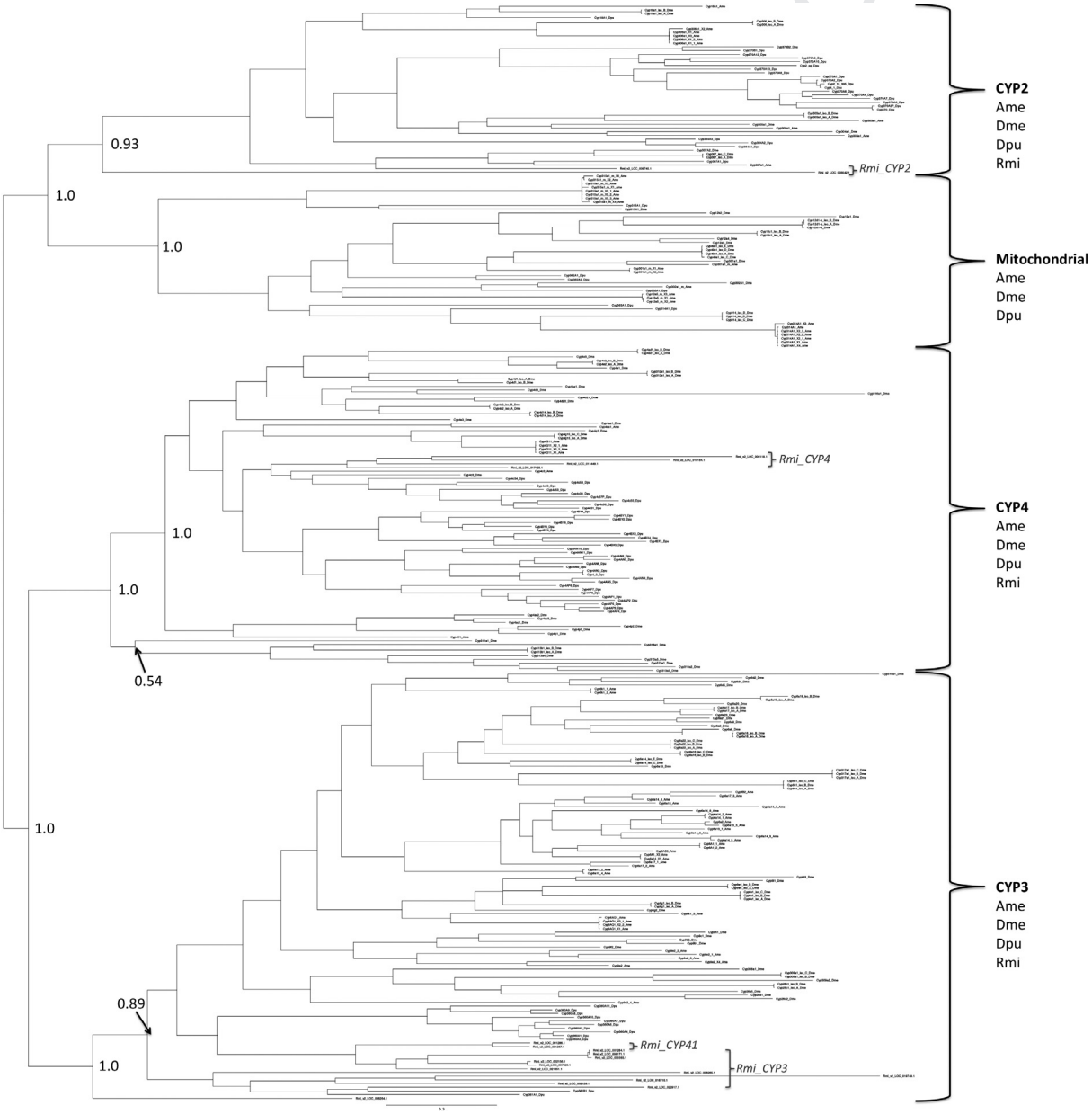

Fig. 5. Phylogenetic relationship of the different cytochrome P450 clans. cytochrome P450 protein coding genes of four species were subjected to phylogenetic comparisons using MrBayes as previously described (Baldwin et al., 2009). These species are *Rhipicephalus microplus*, *Daphnia pulex*, *Apis mellifera* and *Drosophila melanogaster*. Numbers at nodes are posterior probabilities from the Bayesian analysis. *Note:* Fig. 5 is also available in an easily readable, expandable pdf format as a Supplementary Fig. S13 that allows recognition of individual cytochrome P450 isoforms within each clade.

LOC_002156, Rmi_v2_LOC_007826 and Rmi_v2_LOC_35583) were predicted to have a large solvent accessible surface (SA mean volume = 3,249 Å^3 ± 345.3; $n = 4$) and molecular surface (MS mean volume = 7,751.9 Å^2 ± 107.45; $n = 4$) (Supplementary Table S14, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1, and Supplementary Fig. S15). In contrast, Rmi_v2_LOC_001286 (CYP41) and other *R. microplus* CYP3A proteins harbouring a PERF motif and substrate binding sites in their N-terminal region have a substantially smaller solvent accessible surface (SA mean volume = 800 Å^3 ± 75.1; $n = 6$) and molecular surface (MS mean volume = 3,578.6 Å^2 ± 328.1; $n = 6$) (Supplementary Table S14, doi: http://dx.doi.org/10.17632/s7jrdzfmb7.1, and Supplementary Fig. S15). We also found that the predicted mean solvent accessible surface for CYP41 proteins (Rmi_v2_LOC_001286, AAD5400 and AAD91667) was 49.3 Å^3 larger than that of CYP3A proteins (Rmi_v2_LOC_ 001284, Rmi_v2_LOC_006171, and Rmi_v2_-LOC_021651) (Supplementary Table S14, doi:http://dx.doi.org/10.17632/s7jrdzfmb7.1). The larger predicted cavity in CYP41 proteins may be required to accommodate coumaphos ($C_{14}H_{16}ClO_5PS$) that has an average monoisotopic mass of 362.014465 Da and an estimated volume of ~602 Å^3. It remains to be elucidated if the new *R. microplus* CYP41 (Rmi_v2_LOC_001286.1) gene found in this study with an estimated solvent accessible surface volume of 897.3 Å^3 could bind coumaphos and mediate OP resistance.

In conclusion, we report the gene-enriched draft assembly of ~2 Gbp of the cattle tick *R. microplus* genome that is one of the most significant pests of cattle production worldwide. The *R. microplus* genome assembly represents the first large-scale genomic resource for the diverse lineage of metastriate ticks. We envisage that this resource will facilitate a number of applications including understanding its unique biology, transmission of pathogens and designing novel strategies to overcome the *R. microplus* resistance to acaricides that pose a significant threat to diary and beef industries across the globe.

Acknowledgements

This project was supported by funding (FDG) from the U. S. Department of Agriculture, Agricultural Research Service (USDA-ARS) Project Nos. 6205-32000-024-00D, 6205-32000-026-00D, 6205-32000-031-00D, and 3094-32000-036-00D. Funding from the Organization for Economic Co-operation and Development's Co-operative Research Programme: Biological Resource Management for Sustainable Agriculture Systems, Washington, USA (FDG 2009, MIB 2012) was instrumental in allowing this research to continue. We are grateful for the efforts of Drs. Ron Rosenburg, Steve Kappes, Dan Strickman, and John George from USDA-ARS to secure administrative funding for genome sequencing. The guidance provided by the late Dr. Ernie Retzel and staff at the National Center for Genome Resources, USA, especially Andrew Farmer, Nico Devitt and Patricia Mena (Santa Fe, NM, USA) was important for the PacBio sequencing phase of this project. During the startup phase, Dr. Vishvanath Nene (then at The Institute for Genomic Research, Rockville, MD, USA, now at International Livestock Research Institute, Nairobi, Kenya) provided invaluable technical guidance. USDA is an equal opportunity employer. We also would like to acknowledge BioPlatforms Australia for enabling access to analytical workflows.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ijpara.2017.03.007.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
Andreotti, R., Guerrero, F.D., Soares, M.A., Barros, J.C., Miller, R.J., de Leon, A.P., 2011. Acaricide resistance of *Rhipicephalus* (*Boophilus*) *microplus* in State of Mato Grosso do Sul. Brazil. Rev. Bras. Parasitol. Vet. 20, 127–133.
Angus, B.M., 1996. The history of the cattle tick *Boophilus microplus* in Australia and achievements in its control. Int. J. Parasitol. 26, 1341–1355.
Au, K.F., Underwood, J.G., Lee, L., Wong, W.H., 2012. Improving PacBio long read accuracy by short read alignment. PLoS One 7, e46679.
Baldwin, W.S., Marko, P.B., Nelson, D.R., 2009. The cytochrome P450 (CYP) gene superfamily in *Daphnia pulex*. BMC Genomics 10, 169.
Bao, Z., Eddy, S.R., 2002. Automated *de novo* identification of repeat sequence families in sequenced genomes. Gen. Res. 12, 1269–1276.
Barrero, R.A., Keeble-Gagnere, G., Zhang, B., Moolhuijzen, P., Ikeo, K., Tateno, Y., Gojobori, T., Guerrero, F.D., Lew-Tabor, A., Bellgard, M., 2011a. Evolutionary conserved microRNAs are ubiquitously expressed compared to tick-specific miRNAs in the cattle tick *Rhipicephalus* (*Boophilus*) *microplus*. BMC Genomics 12, 328.
Barrero, R.A., Chapman, B., Yang, Y.F., Moolhuijzen, P., Keeble-Gagnere, G., Zhang, N., Tang, Q., Bellgard, M.I., Qiu, D.Y., 2011b. De novo assembly of *Euphorbia fischeriana* root transcriptome identifies prostratin pathway related genes. BMC Genomics 12, 600.
Barrero, R.A., Napier, K.R., Cunnington, J., Liefting, L., Keenan, S., Frampton, R.A., Szabo, T., Bulman, S., Hunter, A., Ward, L., Whattam, M., Bellgard, M.I., 2017. An internet-based bioinformatics toolkit for plant biosecurity diagnosis and surveillance of viruses and viroids. BMC Bioinform. 18, 26.
Baxter, G.D., Barker, S.C., 1999. Comparison of acetylcholinesterase genes from cattle ticks. Int. J. Parasitol. 29, 1765–1774.
Bellgard, M.I., Moolhuijzen, P.M., Guerrero, F.D., Schibeci, D., Rodriguez-Valle, M., Peterson, D.G., Dowd, S.E., Barrero, R., Hunter, A., Miller, R.J., Lew-Tabor, A.E., 2012. CattleTickBase: an integrated Internet-based bioinformatics resource for *Rhipicephalus* (*Boophilus*) *microplus*. Int. J. Parasitol. 42, 161–169.
Bendele, K.G., Guerrero, F.D., Miller, R.J., Li, A.Y., Barrero, R.A., Moolhuijzen, P.M., Black, M., McCooke, J.K., Meyer, J., Hill, C.A., Bellgard, M.I., 2015. Acetylcholinesterase 1 in populations of organophosphate-resistant North American strains of the cattle tick, *Rhipicephalus microplus* (Acari: Ixodidae). Parasitol. Res. 114, 3027–3040.
Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580.
Berlin, K., Koren, S., Chin, C.S., Drake, J.P., Landolin, J.M., Phillippy, A.M., 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat. Biotechnol. 33, 623–630.
Bird, A.W., Yu, D.Y., Pray-Grant, M.G., Qiu, Q.F., Harmon, K.E., Megee, P.C., Grant, P.A., Smith, M.M., Christman, M.F., 2002. Acetylation of histone H4 by Esa1 is required for DNA double-strand break repair. Nature 419, 411–415.
Boisvert, S., Laviolette, F., Corbeil, J., 2010. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J. Comput. Biol. 17, 1519–1533.
Canales, M., Almazan, C., Naranjo, V., Jongejan, F., de la Fuente, J., 2009. Vaccination with recombinant *Boophilus annulatus* Bm86 ortholog protein, Ba86, protects cattle against *B. annulatus* and *B. microplus* infestations. BMC Biotechnol. 9, 29.
Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A. S., Yandell, M., 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18, 188–196.
Cramaro, W.J., Revets, D., Hunewald, O.E., Sinner, R., Reye, A.L., Muller, C.P., 2015. Integration of *Ixodes ricinus* genome sequencing with transcriptome and proteome annotation of the naïve midgut. BMC Genomics 16, 871.
Crampton, A.L., Baxter, C.D., Barker, S.C., 1999. A new family of cytochrome P450 genes (CYP41) from the cattle tick, *Boophilus microplus*. Insect Biochem. Mol. 29, 829–834.
de Castro, J.J., 1998. Sustainable tick and tickborne disease control in livestock improvement in developing countries. Vet. Parasitol. 77, 213–215.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.
Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., Turner, S., 2009. Real-time DNA sequencing from single polymerase molecules. Science 323, 133–138.
English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C., Gibbs, R.A., 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One 7, e47768.
Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.
Fischer, M., Knoll, M., Sirim, D., Wagner, F., Funke, S., Pleiss, J., 2007. The cytochrome P450 engineering database: a navigation and prediction tool for the cytochrome P450 protein family. Bioinformatics 23, 2015–2017.

Friedlander, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., Rajewsky, N., 2008. Discovering microRNAs from deep sequencing data using miRDeep. Nat. Biotechnol. 26, 407–415.

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152.

Gammon, D.B., Mello, C.C., 2015. RNA interference-mediated antiviral defense in insects. Curr. Opin. Insect Sci. 8, 111–120.

Gotoh, O., 1993. Evolution and differentiation of P450 genes Cytochrome P450. In: Omura, T., Ishimura, Y., Fujii-Kuriyama, Y. (Eds.), Cytochrome P450. second ed. Kodansha, Tokyo, Japan, pp. 255–272.

Graham, O.H., Hourrigan, J.L., 1977. Eradication programs for the arthropod parasites of livestock. J. Med. Entomol. 13, 629–658.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S.R., 2003. Rfam: an RNA family database. Nucleic Acids Res. 31, 439–441.

Grisi, L., Massard, C.L., Moya Borha, G.E., Pereira, J.B., 2002. Impacto, economico das principais ectoparitoses em bovinos no Brasil. Hora. Vet. 21, 3.

Guerrero, F.D., Miller, R.J., Rousseau, M.E., Sunkara, S., Quackenbush, J., Lee, Y., Nene, V., 2005. BmiGI: A database of cDNAs expressed in *Boophilus microplus*, the tropical/southern cattle tick. Insect Biochem. Mol. Biol. 35, 585–595.

Guerrero, F.D., Bendele, K.G., Chen, A.C., Li, A.Y., Miller, R.J., Pleasance, E., Varhol, R., Rousseau, M.E., Nene, V.M., 2007. Serial analysis of gene expression in the southern cattle tick following acaricide treatment of larvae from organophosphate resistant and susceptible strains. Insect Mol. Biol. 16, 49–60.

Guerrero, F.D., Moolhuijzen, P., Peterson, D.G., Bidwell, S., Caler, E., Bellgard, M., Nene, V.M., Djikeng, A., 2010. Reassociation kinetics-based approach for partial genome sequencing of the cattle tick, *Rhipicephalus* (*Boophilus*) *microplus*. BMC Genomics 11, 374.

Guerrero, F.D., Miller, R.J., de Leon, A.A.P., 2012. Cattle tick vaccines: many candidate antigens, but will a commercially viable product emerge? Int. J. Parasitol. 42, 421–427.

Guglielmone, A.A., Robbins, R.G., Apanaskevich, D.A., Petney, T.N., Estrada-Pena, A., Horak, I.G., Shao, R.F., Barker, S.C., 2010. The argasidae, ixodidae and nuttalliellidae (Acari: Ixodida) of the world: a list of valid species names. Zootaxa 2528, 1–28.

Gulia-Nuss, M., Nuss, A.B., Meyer, J.M., Sonenshine, D.E., Roe, R.M., Waterhouse, R. M., Sattelle, D.B., de la Fuente, J., Ribeiro, J.M., Megy, K., Thimmapuram, J., Miller, J.R., Walenz, B.P., Koren, S., et al., 2016. Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. Nature Comm. 7, 10507.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogen. Gen. Res. 110, 462–467.

Kent, W.J., 2002. BLAT-the BLAST-like alignment tool. Gen. Res. 12, 656–664.

Kim, A., Terzian, C., Santamaria, P., Pelisson, A., Prud'homme, N., Bucheton, A., 1994. Retroviruses in invertegrates: The *gypsy* retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U.S.A. 91, 1285–1289.

Kohjima, M., Noda, Y., Takeya, R., Saito, N., Takeuchi, K., Sumimoto, H., 2002. PAR3beta, a novel homologue of the cell polarity protein PAR3, localizes to tight junctions. Biochem. Biophys. Res. Commun. 299, 641–646.

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D., Phillippy, A.M., 2012. Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. Nat. Biotechol. 30, 693–700.

Korf, I., 2004. Gene finding in novel genomes. BMC Bioinform. 5, 59.

Koski, L.B., Gray, M.W., Lang, B.F., Burger, G., 2005. AutoFACT: an automatic functional annotation and classification tool. BMC Bioinform. 6, 151.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Gen. Biol. 10, R25.

Lew-Tabor, A.E., Moolhuijzen, P.M., Vance, M.E., Kurscheid, S., Valle, M.R., Jarrett, S., Minchin, C.M., Jackson, L.A., Jonsson, N.N., Bellgard, M.I., Guerrero, F.D., 2010. Suppressive subtractive hybridization analysis of *Rhipicephalus* (*Boophilus*) *microplus* larval and adult transcript expression during attachment and feeding. Vet. Parasitol. 167, 304–320.

Li, W.Z., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.

Li, L., Stoeckert, C.J., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178–2189.

Lu, X.W., Wontakal, S.N., Emelyanov, A.V., Morcillo, P., Konev, A.Y., Fyodorov, D.V., Skoultchi, A.I., 2009. Linker histone H1 is essential for *Drosophila* development, the establishment of pericentric heterochromatin, and a normal polytene chromosome structure. Genes Dev. 23, 452–465.

Luo, R.B., Liu, B.H., Xie, Y.L., Li, Z.Y., Huang, W.H., Yuan, J.Y., He, G.Z., Chen, Y.X., Pan, Q., Liu, Y.J., et al., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1, 18.

Ma, H.Y., Ma, C.Y., Li, S.J., Jiang, W., Li, X.C., Liu, Y.X., Ma, L.B., 2014. Transcriptome analysis of the mud crab (*Scylla paramamosain*) by 454 deep sequencing: assembly, annotation, and marker discovery. PLoS One 9, e102668.

Mans, B.J., Neitz, A.W.H., 2004. Adaptation of ticks to a blood-feeding environment: evolution from a functional perspective. Insect Biochem. Mol. Biol. 34, 1–17.

Maritz-Olivier, C., van Zyl, W., Stutzer, C., 2012. A systematic, functional genomics, and reverse vaccinology approach to the identification of vaccine candidates in the cattle tick, *Rhipicephalus microplus*. Ticks Tick-Borne Dis. 3, 179–187.

McCooke, J.K., Guerrero, F.D., Barrero, R.A., Black, M., Hunter, A., Bell, C., Schilkey, F., Miller, R.J., Bellgard, M.I., 2015. The mitochondrial genome of a Texas outbreak strain of the cattle tick, *Rhipicephalus* (*Boophilus*) *microplus*, derived from whole genome sequencing Pacific Biosciences and Illumina reads. Gene 571, 135–141.

Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H., Remington, K.A., et al., 2000. A whole-genome assembly of *Drosophila*. Science 287, 2196–2204.

Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D.G., Vezzi, F., Delhomme, N., et al., 2013. The Norway spruce genome sequence and conifer genome evolution. Nature 497, 579–584.

Price, A.L., Jones, N.C., Pevzner, P.A., 2005. *De novo* identification of repeat families in large genomes. Bioinformatics Suppl 1, i351–i358.

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842.

Rodriguez-Vivas, R.I., Perez-Cogollo, L.C., Rosado-Aguilar, J.A., Ojeda-Chi, M.M., Trinidad-Martinez, I., Miller, R.J., Li, A.Y., de Leon, A.P., Guerrero, F., Klafke, G., 2014. *Rhipicephalus* (*Boophilus*) *microplus* resistant to acaricides and ivermectin in cattle farms of Mexico. Rev. Bras. Parasitol. Vet. 23, 113–122.

Rosario-Cruz, R., Guerrero, F.D., Miller, R.J., Rodriguez-Vivas, R.I., Tijerina, M., Dominguez-Garcia, D.I., Hernandez-Ortiz, R., Cornel, A.J., McAbee, R.D., Alonso-Diaz, M.A., 2009. Molecular survey of pyrethroid resistance mechanisms in Mexican field populations of *Rhipicephalus (Boophilus) microplus*. Parasitol. Res. 105, 1145–1153.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Sambrook, J., Russell, D.W., 2001. Molecular Cloning. A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA.

Sambrook, J., Fritsch, E.F., Maniatis, T., 1989. Molecular Cloning. A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA, pp. 9.17–9.19.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212.

Smeds, L., Kunstner, A., 2011. CONDETRI – a content dependent read trimmer for illumina data. PLoS One 6, e26314.

Stanke, M., Schoffmann, O., Morgenstern, B., Waack, S., 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinform. 7, 62.

Sunter, J.D., Patel, S.P., Skilton, R.A., Githaka, N., Knowles, D.P., Scoles, D.P., Nene, V., de Villiers, E., Bishop, R.P., 2008. A novel SINE family occurs frequently in both genomic DNA and transcribed sequences in ixodid ticks of the arthropod sub-phylum Chelicerata. Gene 415, 13–22.

Tang, S., Lomsadze, A., Borodovsky, M., 2015. Identification of protein coding regions in RNA transcripts. Nucleic Acids Res. 43, e78.

Tao, Q., Zhang, H.B., 1998. Cloning and stable maintenance of DNA fragments over 300 kb in *Escherichia coli* with conventional plasmid-based vectors. Nucleic Acids Res. 26, 4901–4909.

Temeyer, K.B., Davey, R.B., Chen, A.C., 2004. Identification of a third *Boophilus microplus* (Acari: Ixodidae) cDNA presumptively encoding an acetylcholinesterase. J. Med. Entomol. 41, 259–268.

Ullmann, A.J., Lima, C.M.R., Guerrero, F.D., Piesman, J., Black, W.C., 2005. Genome size and organization in the blacklegged tick, *Ixodes scapularis* and the Southern cattle tick, *Boophilus microplus*. Insect Mol. Biol. 14, 217–222.

Wang, M.H., Guerrero, F.D., Pertea, G., Nene, V.M., 2007. Global comparative analysis of ESTs from the southern cattle tick, *Rhipicephalus* (*Boophilus*) *microplus*. BMC Genomics 8, 368.

Yu, L.Y., Tang, W.Q., He, W.Y., Ma, X.L., Vasseur, L., Baxter, S.W., Yang, G., Huang, S.G., Song, F.Q., You, M.S., 2015. Characterization and expression of the cytochrome P450 gene family in diamondback moth, *Plutella xylostella* (L.). Sci. Rep. 5.

Zuckerkandl, E., Pauling, L., 1965. Evolutionary divergence and convergence in proteins. In: Bryson, V., Vogel, H.J. (Eds.), Evolving Genes and Proteins. Academic Press, New York, pp. 97–166.

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