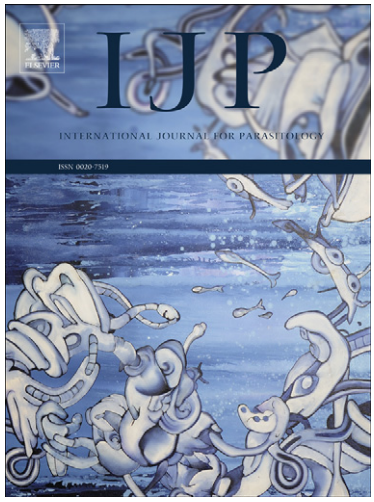


**Bellgard, M.I., Moolhuijzen, P.M., Guerrero, F.D., Schibeci, D., Rodriguez-Valle, M., Peterson, D.G., Dowd, S.E., Barrero, R., Hunter, A., Miller, R.J. and Lew-Tabor, A.E. (2011)** ***CattleTickBase: An integrated Internet-based bioinformatics resource for Rhipicephalus (Boophilus) microplus.*** **International Journal for Parasitology, 42 (2). pp. 161-169.**




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

Matthew I. Bellgard, Paula M. Moolhuijzen, Felix D. Guerrero, David Schibeci, Manuel Rodriguez-Valle, Daniel G. Peterson, Scot E. Dowd, Roberto Barrero, Adam Hunter, Robert J. Miller, Ala E. Lew-Tabor

PII: S0020-7519(11)00281-5
DOI: 10.1016/j.ijpara.2011.11.006
Reference: PARA 3344

To appear in: *International Journal for Parasitology*

Received Date: 9 September 2011
Revised Date: 16 November 2011
Accepted Date: 17 November 2011

Please cite this article as: 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., CattleTickBase: An integrated Internet-based bioinformatics resource for *Rhipicephalus* (*Boophilus*) *microplus*, *International Journal for Parasitology* (2011), doi: 10.1016/j.ijpara.2011.11.006

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

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

[a] *Centre for Comparative Genomics, Murdoch University, Perth, WA 6150, Australia*

[b] *Cooperative Research Centre for Beef Genetic Technologies, Armidale, NSW, Australia*

[c] *USDA-ARS Knipling Bushland US Livestock Insect Research Laboratory, 2700 Fredericksburg Rd., Kerrville, TX 78028, USA*

[d] *Queensland Alliance for Agriculture, Food & Innovation, The University of Queensland and Dept. of Employment, Economic Development & Innovation, P. O. Box 6097, St. Lucia 4067 QLD, Australia*

[e] *Department of Plant & Soil Sciences and Life Sciences & Biotechnology Institute, Mississippi State University, 117 Dorman Hall, Box 9555, Mississippi State, MS 39762, USA*

[f] *Molecular Research, 503 Clovis Road, Shallowater, TX, 79363, USA*

[g] *USDA-ARS Cattle Fever Tick Research Laboratory, 22675 North Moorefield Road, Building 6419, Edinburg, TX 78541, USA*

[1]These authors contributed equally.

*Corresponding author. Felix D. Guerrero, USDA-ARS Knipling Bushland US Livestock Insect Research Laboratory, 2700 Fredericksburg Rd., Kerrville, TX 78028, USA. Tel.: +1-830-792-0327; fax: +1-830-792-0314.

*E-mail address*: Felix.Guerrero@ars.usda.gov

✯Note: Nucleotide sequence data reported in this paper are available in GenBank under accession numbers **HN108288**-**HN118367**, **HM748958**-**HM748967**, **HN108288**-**HN118367**.

Note: Supplementary data associated with this article.

**ABSTRACT**

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

*Keywords*: Cattle tick; Transcriptome; Genome project; Bioinformatics

## 1. Introduction

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

Ticks are believed to be among the most ancient terrestrial arachnids and possibly the earliest organisms to have evolved blood-feeding capabilities (Mans and Neitz, 2004). *Rhipicephalus microplus* is a single-host species and has evolved such that it must maintain sustained contact with its host during the life stages, from the attached and feeding larva through to the fully engorged female. This period of attachment typically lasts approximately 3 weeks with some variation depending on environmental conditions. The species has developed a unique means of avoiding the host animal's immune responses during infestation (Wikel, 1999) and *R. microplus* salivary gland extracts have been shown to have an immunosuppressive effect on the bovine host (Turni et al., 2004, 2007). The tick must also respond to many

79 microorganisms, both symbiotic and parasitic, from the external environment or those
80 ingested through feeding-associated activities (Andreotti et al., 2011).

81 With this interplay between bovine host, tick and microbiota, determining the
82 whole genome sequence of *R. microplus* will greatly advance tick gene discovery,
83 enable a better understanding of tick-host-pathogen immunology and provide insight
84 on how the cattle tick responds to environmental perturbations, including pressures
85 from moisture and temperature extremes and acaricidal applications. The *R. microplus*
86 genome size is estimated to be 7.1 Gigabase pairs (Gbp), more than twice the size of
87 the human genome, and consists of greater than 70% repetitive DNA (Ullmann et al.,
88 2005). It is therefore a challenge for de novo assembly, even with contemporary DNA
89 sequencing technologies. A 4X shotgun coverage genome sequence for the
90 blacklegged tick, *Ixodes scapularis*, is available (Lawson et al., 2009) and is the only
91 reported tick genome sequence to date. The version 1.1 sequence assembly consists of
92 369,492 supercontigs, totalling 1.76 Gbp with a supercontig N50 size of 72kb.

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

101 Towards a goal of generating a genetic resource for this economically
102 important tick species, efforts have focused on a combination of sequencing

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

## 2. Materials and methods

### *2.1. Source of tick materials*

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

### *2.2. Sequencing and assembly*

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

cross_match, with options set at minmatch 12 and minscore 20. Contig order and orientation were based on Phrapview paired end reads.

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

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

*2.3. BAC and Cot read alignment*

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

*2.4. Gene predictions*

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

*2.5. Repeat analysis*

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

*2.6. RNA Searches: tRNA and rRNA*

Transfer RNA (tRNA) searches were conducted with trnascan version 1.23 (Schattner et al., 2005) and rRNA using rnammer version 1.2 (Lagesen et al., 2007).

*2.7. Sequence comparative analysis*

The genomic data set *Ixodes scapularis* SUPERCONTIGS-Wikel.IscaW1.fa.gz (dataset downloaded from VectorBase (Lawson et al., 2009) Date: Sept. 21, 2010) was aligned to *R. microplus* assembled Cot DNA (GenBank WGS division second version, ADMZ02000000) using BLASTn (Altschul et al., 1990). Homologous regions of interest were selected at an expected value < 1e-50.

The *R. microplus* BAC sequences submitted to GenBank: **HM748958**-**HM748967** and five BACs as described in Guerrero et al. (2010) were aligned to the BES submitted to the Genome Survey Sequence (GSS) division of GenBank: **HN108288**-**HN118367** using BLAT with 70% identity, a length greater than 100 bp, and the option 'fastMap'.

The *R. microplus* BES, predicted BAC gene content and new transcript sequences were comparatively aligned to Dana Farber Cancer Institute (DFCI) gene indices, IsGI version 3.0 and *Bmi*GI version 2.1 (Quackenbush et al., 2001), NCBI RefSeq mRNA, and the Subtraction Library clones as described previously (Lew-Tabor et al., 2009) using BLASTn at an expected value <1e-10, and to NCBI RefSeq Protein and GenPeptide, iscapularis.PEPTIDES-IscaW1.1.fa (VectorBase Date: Sep 21, 2010, Lawson et al., 2009) using BLASTx at an expected value <1e-10. The collective *R. microplus* transcriptome (*Rmi*Tr Version 1.0, Table 1) and *I. scapularis* (DFCI IscGI) sequences were searched using tBLASTx at an expected value of 1e-05 to uniref100, and orthologous sequences were determined for those alignments that had greater than or equal to 60% *R. microplus* sequence coverage and an amino acid conservation greater than or equal to 30%. Transcript alignments were also made using BLAT (Kent, 2002) to a comparative species with 70% identity and a length greater than 100 bp.

### *2.8. Transcriptome sequencing*

For transcriptome sequencing, female adult dissected gut and 'frustrated' larvae were prepared as described previously from Australian N strain ticks (Lew-Tabor et al., 2009). The frustrated larval sample contains larvae placed in a gas-

permeable bag taped directly to the host animal feed source. Thus, the larvae are able to sense the presence of the host but the bag presents a barrier that prevents attachment and feeding. Approximately 30 mg of total RNA from each of these samples were collected for high-throughput sequencing of the tick transcriptome using the Illumina/GA single-end reads format as described previously (Mortazavi et al., 2008).

### *2.9. Transcriptome assembly and clustering*

The de novo transcriptome assembly of the adult female gut and frustrated larvae transcriptomes using 60 bp single-end Illumina/GA reads was conducted using Abyss (Birol et al., 2009) with *k*-mer sizes ranging from 36 to 64. The assembled contigs were then clustered using cap3 (Huang and Madan, 1999) with a 98% sequence identity threshold and an overlap region of at least 30 bases to remove transcript redundancy. Non-redundant sets of transcripts for the two libraries can be found on the CattleTickBase website.

The *Rmi*Tr version 1.0 data set contains sequences from DFCI_bmigi.V2.1, adult female gut transcriptome, frustrated larvae transcriptome and *R. microplus* subtraction library (Lew-Tabor et al., 2009), which were clustered into contigs using cap3 (Huang and Madan, 1999) with following the options: -p 99.99999 -m 1 -n -10 -g 1 -b 16 -y 6.

### *2.10. YABI*

The YABI application consists of a front-end web application responsible for the user interface. Users create a secure account and are free to access the datasets and analysis tools available within the system. Users can create bioinformatics pipelines from the available tools. The tools currently available for sequence analysis include: similarity/homology searches, feature prediction, high throughput downstream analysis, assembly and annotation. Datasets (including other tick-related GenBank bioprojects) in YABI are updated from GenBank/EMBL/DDBJ and VectorBase at regular intervals. Individual dataset contributors can deposit and update data sets by contacting yabi@ccg.murodch.edu.au with an option to have a secure account to conduct their analysis within their own teams. The Centre for Comparative Genomics (CCG), Murdoch University, Australia is committed to supporting the bioinformatics aspects of the *R. microplus* project. The CCG houses a supercomputer (currently ranked number 87 in the world - http://www.top500.org/list/2010/11/100) and provides support to national and international bioinformatics and other high-end science-based activities. The CCG develops and deploys sophisticated software solutions, supports and conducts a diverse range of bioinformatics analysis. Requests and suggestions can be made by contacting info@ccg.murdoch.edu.au. The *R. microplus* datasets currently available for sequence similarity searching and other bioinformatics analyses are summarised in Table 1.

## 3. Results

### *3.1. Rhipicephalus microplus datasets currently available*

The nine datasets that are available for access and further analysis on the CattleTickBase website are summarised in Table 1 and described below. For Dataset

1, total *R. microplus* genomic DNA was prepared and processed by three Cot filtration experiments to enrich for single/low-copy and moderately repetitive DNA. It was anticipated that the DNA obtained via this process would remove a significant portion of highly repetitive DNA and contain predominantly gene rich regions. The resulting DNA fragments ranged in size from 250 to 600 bp and were sequenced in three separate sequencing experiments with one experiment using six runs of 454 FLX pyrosequencing (Guerrero et al., 2010) and two experiments each using three Titanium 454 runs. The data from the latter two experiments have been deposited in GenBank SRA, submission: **SRA012677.4/SID00001**. Approximately 1.8 Gbp of sequence were generated and assembled. This assembly (Dataset 2) was submitted to GenBank Whole Genome Sequencing Project; under the United States Department of Agriculture, Agricultural Research Service (USDA-ARS) *R. microplus* Project ID 46685 assigned Project accession **ADMZ00000000**. The most recent version for this project reported in this paper has the accession number **ADMZ02000000**; this submission consists of 175,208 contig sequences of average contig size 825 bp and a maximum contig length of 9,681 bp. These sequences have been submitted under GenBank Accession Numbers **ADMZ02000001**-**ADMZ02175208**. Based on the estimate that the *R. microplus* genome size is approximately 7.1 Gbp, the assembled Cot DNA dataset represents approximately 2% of the *R. microplus* genome. We undertook a comparative analysis of this Cot DNA dataset with the *Bmi*GI Version 2.1 Gene Index ( http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=b_microplus; Quackenbush et al., 2001; Guerrero et al., 2005) to confirm that the Cot DNA filtration process filtered for gene rich regions. When we aligned the 175,208 Cot DNA contigs to the 14,586 *Bmi*GI transcript contigs, 52% of the *Bmi*GI entries had a match (>85% identity) to at least one contig in the Cot

dataset. Thus, even though the coverage of the Cot DNA over the entire 7.1 Gbp *R. microplus* genome was very low (2%), the coverage of the *Bmi*GI transcriptome dataset was high (52%). This also indicates that deeper sequencing of the Cot-selected DNA would be warranted to obtain fuller coverage of the gene-encoding regions, as 48% of the *Bmi*GI entries did not have a match in the Cot DNA dataset.

For Dataset 3, a 3.2x coverage BAC library was used to sequence 10,582 BES resulting in 7,290,530 bp of sequence. The amount of BES data generated represents approximately 0.1% of the *R. microplus* genome. BES greater than 500 bp in size have been deposited in GenBank GSS under Accession Numbers **HN108288**-**HN118367**. Approximately 70% matched with the Cot DNA Dataset 2. Similarity sequence analysis between BES and the *Bmi*GI Version 2.0 reveals that a total of 502 BES (4.7%) aligned, at a length greater than 100 nucleotides and at greater than 90% identity (percent identity, PID), to 224 *Bmi*GI entries or approximately 1.6% of the *Bmi*GI.

The 10,582 BES were compared with the *I. scapularis* genome sequence, the nearest taxonomic whole genome sequencing project, comprising 369,492 scaffolds (supercontigs). Only 58 BES aligned with greater than 80% BES coverage and 80% PID. The number of BES found in comparative searches at specified thresholds were: 2,418 (23%) at an expected alignment value (e-value) of 1e-05 to the NCBI protein non-redundant (nr) database, 416 (4%) at an e-value 1e-20 to *I. scapularis* proteins (Lawson et al., 2009; Megy et al., 2009), 2,559 (24%) BES at an e-value 1e-20 to *R. microplus* gene indices (Guerrero et al., 2005), and 134 (1.2%) to the *I. scapularis* gene index (Ribeiro et al., 2006) at an e-value of 1e-20. The protein functional analyses by Gene Ontology (GO) classification are presented in Supplementary Table S2 and Supplementary Fig. S2.

For Dataset 4, 15 BAC clones were selected and sequenced to completion; 13 were based on hybridization of BAC library filter arrays to probes from known transcripts of interest involved in tick feeding and acaricide resistance and two were randomly selected BACs. These 15 BACs were subjected to sequencing and de-novo assembly. The sequencing of five of the BACs has been reported (Guerrero et al., 2010), and the sequences from the remaining 10 BACs have been deposited in GenBank (Accession Numbers **HM748958**-**HM748967**). Gene prediction analysis found 180 predicted genes with 919 predicted exons with an average exon size of 311 bp (Supplementary Fig. S2). From these predicted genes, there are 166 full-length genes comprised of 145 multiple exon genes containing both the initial and terminal predicted exons and 21 are single exon genes. Genes of particular interest to our research group that were identified in these BACs include cytochrome P450 (Guerrero et al., 2010), the permethrin-degrading carboxylesterase CzEst9 (Guerrero et al., 2010), papilin (Moolhuijzen et al., 2011), transmembrane protein 215-like, transpanin and serpin (Moolhuijzen et al., 2011). The Cot DNA dataset was also mapped to the BACs. Table 2 provides a summary of all 15 BACs, including the length of each assembled BAC, the number of Cot-selected DNA reads from Dataset 1 that mapped to each BAC using the next generation alignment tool BWA (Li and Durbin, 2010) with an alignment score > 10, the percentage of coverage of BAC sequence provided by the Cot reads, and the number of predicted genes and exons. As shown in Table 2, the percentage of Cot DNA coverage over these BACs ranged from 6% (BM-012-E08) to as high as 82% (BM-005-B21). Thirteen BACs had an average GC content between 44-48% (Table 2). One BAC (BM-012-E08) had a GC content of 58% and another BAC (BM-074-Random-F12) had a GC content of 40%. BAC BM-012-E08 contains rDNA and intergenic spacer and a significant proportion of highly

repetitive DNA. Present in the 15 BACs are the Ruka SINE elements identified previously in *Rhipicephalus appendiculatus* (Sunter et al., 2008) and a range of other interspersed repeats such as LINEs L2 and R1, LTR Gypsy and DNA transposons as found by RepeatMasker (Smit, A.F.A., Hubley, R., Green, P., 2004. RepeatMasker Open-3.0. 1996-2010 <http://www.repeatmasker.org>.). BAC BM-005-G14 contained a number of Ruka elements, which are a novel SINE element that was reported to occur frequently in both genomic and transcribed ixodid ticks (Sunter et al., 2008). A single 195 bp length Ruka element found in BM-004-G14 had 91% sequence identity to the138 bp *R. appendiculatus* (*Rap*) Ruka (Genbank: **EU018139.1** complement (9,947-10,084 bp)). A total of 70 Ruka (> 100 bp) elements were found in the *Rmi*Tr Version 1.0 transcriptome (Table 1 Dataset 8), while 409 were found in the assembled Cot DNA.

The remaining datasets are expression related. Dataset 5 consists of the *R. microplus* Gene Index *Bmi*GI Version 2.1 (Quackenbush et al., 2001) entries that were extended by comparative analysis with the assembled Cot-selected genomic DNA (Guerrero et al., 2010). Datasets 6 and 7 consist of transcriptome sequence from adult female tick gut and 'frustrated' larvae libraries, respectively. These datasets were de novo assembled using Illumina short reads. We wanted to compare the transcriptome data contained in our datasets 6 and 7 with the *Bmi*GI Version 2.1 to determine whetehr our data presented new transcripts that could be added to the current *Bmi*GI to create a more comprehensive resource. Fig. 1 provides an overview of the CAP3 sequence clustering between datasets 6 and 7 and *Bmi*GI Version 2.1 at greater that 98% identity and default settings. The Venn diagram shows that datasets 6 and 7 (Aus *Rmi* as noted in Fig. 1) contain 9,652 contigs not found in the *Bmi*GI version 2.1. Dataset 8 is an updated *R. microplus* transcript data set made up of the

combination of the original *Bmi*GI version 2.1 and Datasets 6 and 7, plus the Subtraction Library data (GenBank: **GO253189.1**- **GO253184.1**, **GE650059.1**-**GE650181.1**) and is referred to as *Rmi*Tr version 1.0, containing 28,893 sequences. Finally, Dataset 9 consists of data from microarray experiments using Nimblegen arrays (Rodriguez-Valle et al., 2010; Saldivar et al., 2008) to characterize the *R. microplus* transcriptome responses to attaching and feeding upon *Bos indicus* and *Bos taurus* cattle (Rodriguez-Valle et al., 2010).

### *3.2. CattleTickBase web resource*

The CattleTickBase web resource provides a central point for the cattle tick research community to access genome and transcriptome information on *R. microplus*. The home page is shown in Fig. 2. This website contains a summary of the datasets that have been generated to date and are available for download. It also contains links to precomputed results, which are either figures from supplementary material from publications or the datasets precomputed into the GBrowse genome browser (Stein et al., 2002). To date, precomputed results include a summary of the protein hits from similarity searches contained within the sequenced BACs (Dataset 4) as well as genes that were extended by the Cot DNA contigs that can be viewed within GBrowse (Dataset 5).

A unique feature of our genome project web resource is the ability for the research community to conduct their own sophisticated bioinformatics analysis online. The open source YABI system (http://ccg.murdoch.edu.au/yabi) consists of a front-end web application responsible for the user interface. Users create a secure account and are free to access the datasets and analysis tools available within the

system. Users can create customized bioinformatics pipelines from the available tools and scripts that capture provenance information of the tools used, such as parameters used for each tool, and outputs of tools generated at each step. Fig. 3 shows the layout of the Design tab within YABI for constructing analysis workflows. The tools available are presented in the menu on the left. In the centre, dragging and dropping tools from the menu onto the workflow panel will construct a workflow. The example in Fig. 3 shows a screen shot of the creation of an automated workflow to search for G-Protein Coupled Receptors (GPCRs) in frustrated larvae transcripts that make up Dataset 7 (Table 1). GPCRs are an interesting family of membrane proteins that perform a range of critical biological functions in eukaryotes. Approximately 40% of the prescription pharmaceuticals target GPCRs (Filmore, 2004) and they are attractive as potential targets for acaricidal product development. Predicted open reading frames (ORFs) from the 6,082 assembled contig transcriptome from Dataset 7 were determined and the resulting sequences were searched using GPCRHMM (Wistrand et al., 2006). From a total of 17, 230 predicted ORFs (>= 200 bp), six ORFs were predicted to encode GPCRs (data not shown). A screen cast of this workflow can be found at (http://ccg.murdoch.edu.au/yabi). In CattleTickBase, the web-based analysis workflow environment (YABI) enables researchers to create sophisticated bioinformatics analysis pipelines from a diverse range of tools. For example, tools are available for multiple sequence alignment and assembly (including data from next-generation sequencing technologies), sequence searches, predictions for genes, rRNAs, tRNAs and small RNAs, repeat masking and various annotations. Also, the calculation of the Codon Adaptation Index for a given nucleotide sequence, given a reference codon usage table, can be used for predicting the level of expression of a

given gene and for making comparisons among codon usage in different organisms. *R. microplus* codon usage tables will be available for use in YABI.

## 4. Discussion

This paper provides an overview of the status of our collaborative efforts towards sequencing the *R. microplus* genome. Currently there is a scarcity of tick genome sequence and that limits progress in projects designed to address the significant worldwide impacts that *R. microplus* presents to cattle producers, both large and small. With the difficulties associated with sequencing the large and complex genome of *R. microplus*, an international collaboration has been formed to commence a coordinated effort of targeted sequencing of both genome and transcriptome with the longer-term view to obtain resources to sequence the complete genome of *R. microplus*.

As described in the results section, nine datasets have been generated and these contain the cumulative data our consortium has obtained which is relevant to *R. microplus* genome and transcriptome sequences. In broad terms, these are the Cot-filtration genomic DNA sequences, the BAC-associated sequences and the transcriptome data. The Cot-filtration process filters out highly repetitive DNA and allows the focus of resources on regions of the genome that are enriched for gene coding sequences. The result from our mapping of the Cot-selected DNA to each of the BACs is consistent with the Cot DNA experiments filtering out the highly repetitive DNA fraction. Our analysis demonstrated that Cot-filtration is an effective strategy to focus on gene-rich regions, because our Cot-filtered assembled contigs found matches with 50% of the *Bmi*GI Version 2.1 Gene Index despite the assembled

Cot sequences (Table 1 Dataset 2) only representing 2% of the *R. microplus* genome. The availability of contigs from the Cot genomic DNA data set to align to corresponding transcript sequences in *Bmi*GI will enable the determination of intronic and promoter regions associated with the exonic regions contained in *Bmi*GI.

The acquisition of BES data from *R. microplus* represents the first project of its type for this species. The 502 genes that were matched to the BES data represent a 16-fold increase over the expected number of matches based on the genome coverage of the BES. The BES represents 0.1% of the genome, yet matches to 1.6% of the *Bmi*GI version 2.1 sequences. Further, the repetitive nature of the *R. microplus* genome has been determined by reassociation kinetics analysis and was shown to consist of 0.8% foldback, 30% unique DNA, 38% moderately repetitive DNA and 31% highly repetitive DNA (Ullmann et al., 2005). As the Cot DNA was selected to remove the highly repetitive sequence fraction (32% of the genome), this is consistent with 70% of BES alignment to the assembled Cot DNA contigs. This implies that the BES is high in gene content and accounted for a high percentage alignment to the gene-enriched Cot-selected DNA contigs. One possible reason for this is that the BAC libraries were made with *Mbo*I partially digested genomic DNA. *Mbo*I has GATC as the recognition site, and as GC-rich areas tend to be in gene-rich areas of genomes, *Mbo*I digestion which forms the BES may lend bias to these gene-rich areas.

The 15 BACs provide an insight into the genomic structure and organisation of *R. microplus*. The 15 BAC sequences are composed of 4.9% low complexity sequence and known retrotransposable elements, for example GYPSY and LTR. There is an average of 12 genes per BAC and this was a higher than expected ratio, given the estimated gene space. It is possible that these BACs, 13 of which were selected to be sequenced based on hybridization to known genes, occur in regions of

gene clusters. One particular BAC (BM-012-E08) had 6% of the BAC sequence length mapped by Cot DNA and the remaining unmapped BAC sequence contained highly repetitive intergenic regions nested between rDNA genes. In these selected 15 BACs a total of 97 genes out of 180 had significant alignments to known proteins. These included 28 hypothetical proteins of which 20 were found in *Ixodes scapularis*, five alignments to cytochrome P450 (*I. scapularis*), a receptor for egg jelly protein, a papilin (*Pediculus humanus corporis*), a zinc finger, five unknown *Tribolium* proteins, a serpin, an esterase and a transpanin. The remainder were comprised of polyproteins, helicases and transcriptases (transposable element-like), (gene predictions and annotation can be found in Supplementary Table S1 and Supplementary Fig. S1).

The content of the newly assembled *R. microplus* transcriptome for the cattle tick (Table 1 Dataset 8: *Rmi*Tr Version 1.0) contains 2,379 full-length sequences in common with the previous *R. microplus* gene index and an additional 9,652 novel full-length transcript sequences (Fig. 1). The *Bmi*GI version 2.1 dataset consists of 9,851 contigs and 4,735 singletons. In our experience working with *Bmi*GI, approximately one-third to one-half of the contigs encode full-length proteins while a very low percentage of the singletons encode full-length proteins. Thus, Dataset 8 should contain approximately 14,000 full-length transcripts. Further, the newly assembled transcripts for the cattle tick contained 3,829 sequences with orthologues to *I. scapularis* transcript sequences version 3.0 while there were 4,948 sequences in *I. scapularis* Gene Index Version 3.0 that had orthologues to sequences in *Rmi*TR Version 1.0 (Table 3). As an example of the utility of CattleTickBase and *Rmi*Tr Version 1.0, we searched both *Rmi*Tr Version 1.0 and *I. scapularis* Gene Index Version 3.0 for sequences with similarity to esterases, cytochrome P450, and

485 glutathione S-transferases, metabolic proteins that often play roles in tick resistance to
486 acaricides. The results are shown in Table 3 and Supplementary Data S1-S3.

487 CattleTickBase is designed to become the comprehensive bioinformatics web
488 resource for *R. microplus*. CattleTickBase currently contains over 1.8 Gbp of genomic
489 sequence from *R. microplus*. Most of this sequence was derived from the Cot-selected
490 genomic DNA. This DNA was selected to yield the unique gene-enriched fraction of
491 the genomic DNA from *R. microplus* (Guerrero et al., 2010). Thus, if we assume the
492 30% unique fraction value from Ullmann et al. (2005), and also assume that our 1.8
493 Gbp of Cot-selected data primarily sources from the unique fraction (Table 1 Dataset
494 1), then with the *R. microplus* genome size of 7.1 Gbp, 30% of unique fraction would
495 represent 2.1 Gbp. Thus continuing the assumptions, CattleTickBase can be
496 considered to represent 0.86X coverage of the unique fraction (gene-enriched
497 fraction) of the *R. microplus* genome. However our Cot filtration dataset, although
498 effective in the recovery of cattle tick genomic DNA regions enriched in gene coding
499 sequences, has a drawback in that the low coverage has resulted in a somewhat
500 fragmented sequence assembly. This needs to be rectified by deeper sequencing of the
501 Cot-selected DNA or whole genome sequencing. Additionally, we expect genetic
502 variations amongst *R. microplus* populations from different countries or ecological
503 regions and this is necessarily an important consideration when defining consensus
504 reference data sets for *R. microplus*. Most of the data in CattleTickBase is derived
505 from North American *R. microplus*, although datasets 6 and 7 in Table 1 are from
506 Australian ticks. An example of a well-studied *R. microplus* transcript with sequence
507 that varies among different population is *Bm86*. Freeman et al. (2010) found 8.3%
508 sequence variation between *Bm86* isolated from *R. microplus* ticks from South Texas,

USA and from the Australian Yeerongpilly strain. Andreotti et al. (2008) reported similar geographical differences in *Bm86* isolated from South American populations.

Nevertheless, we believe CattleTickBase is a significant resource for the tick research community that should facilitate tick research in a number of areas. Not only does CattleTickBase provide standard features of a genome sequencing project browser and a BLAST interface; it also contains an online web-based analysis environment (YABI) that enables the scientific community to conduct sophisticated bioinformatics analysis. YABI's ease of use and facility for results to be easily downloaded and used in other bioinformatics analysis environments makes it a flexible web-based bioinformatics environment. To our knowledge, this type of environment is not offered elsewhere. With the advent of third generation sequencing technologies that promise longer reads than current 454 and Illumina technologies, we expect to focus upon whole genome sequencing and targeted transcriptomic analysis to further advance the *R. microplus* genome sequencing project. Naturally, progress will be dependent on obtaining the necessary resources (both financial and scientific) to further sequence these more difficult regions of the cattle tick genome. As obtained, this information will be integrated into CattleTickBase to provide the scientific community with access to the latest information from the *R. microplus* genome-sequencing project.

In summary, we have acquired and assembled genomic DNA and transcriptomic sequence data into contigs which represent over 1.8 Gbp of DNA from gene-enriched regions of the *R. microplus* genome. The data was compiled into a resource called CattleTickBase and a web-based YABI resource was developed to enable the scientific community to access our databases. The YABI tool will facilitate

533 access and manipulation of cattle tick genome sequence data and, as the genome of *R.*
534 *microplus* proceeds to completion, the CattleTickBase resource will be updated.

535

536 **Acknowledgements**

537 The authors acknowledge funding support from the Beef CRC and the
538 National and International Research Alliances Program (DEEDI, Queensland
539 Government Smart Futures Funds, Australia) and U.S. Department of Agriculture
540 (USDA)-ARS Knipling Bushland US Livestock Insect Research Laboratory CRIS
541 Project Number 6205-32000-031-00 for the tick BAC and BES. We further
542 acknowledge technical assistance from Ms. Kylie G. Bendele, USDA-ARS Kerrville,
543 TX, USA and Dr. Zenaida Magbanua, Mississippi State University, USA and Dr.
544 Louise Jackson from Biosecurity, DEEDI, Australia for providing Australian ticks.
545 Thanks to Dr. Rudi Appels for helpful discussions during data analysis and
546 manuscript preparation. USDA is an equal opportunity employer.

547

548 **References**

549 Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local
550 alignment search tool. J. Mol. Biol. 215, 403-410.
551 Andreotti, R., Perez de Leon, A.A., Dowd, S.E., Guerrero, F.D., Bendele, K.G.,
552 Scoles, G.A., 2011. Assessment of bacterial diversity in the cattle tick
553 *Rhipicephalus (Boophilus) microplus* through tag-encoded pyrosequencing.
554 BMC Microbiol. 11, 6.
555 Andreotti, R., Pedroso, M.S., Caetano, A.R., Martins, N.F., 2008. Comparison of
556 predicted binders in *Rhipicephalus (Boophilus) microplus* intestine protein
557 variants Bm86 Campo Grande strain, Bm86 and Bm95. Rev. Brasileira
558 Parasitol. Vet. 17, 93-98.
559 Barker, S.C., Murrell, A., 2004. Systematics and evolution of ticks with a list of valid
560 genus and species names. Parasitol. 129 Suppl., S15-S36.
561 Birol, I., Jackman, S.D., Nielsen, C.B., Qian, J.Q., Varhol, R., Stazyk, G., Morin,
562 R.D., Zhao, Y., Hirst, M., Schein, J.E., Horsman, D.E., Connors, J.M.,
563 Gascoyne, R.D., Marra, M.A., Jones, S.J., 2009. De novo transcriptome
564 assembly with ABySS. Bioinformatics 25, 2872-2877.
565 Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic
566 DNA. J. Mol. Biol. 268, 78-94.
567 de Castro, J.J., 1997. Sustainable tick and tickborne disease control in livestock
568 improvement in developing countries. Vet. Parasitol. 71, 77-97.
569 Drysdale, R., 2008. FlyBase : a database for the *Drosophila* research community.
570 Methods Mol. Biol. 420, 45-59.
571 Ewing, B., Green, P., 1998. Base-calling of automated sequencer traces using phred.
572 II. Error probabilities. Genome Res. 8, 186-194.

573 Filmore, D., 2004. It's a GPCR world. Modern Drug Disc. 7, 24-28.
574 Freeman, J.M., Davey, R.B., Kappmeyer, L.S., Kammlah, D.M., Olafson, P.U., 2010.
575 Bm86midgut protein sequence variation in South Texas cattle fever ticks.
576 Parasites Vect. 3, 101.
577 Grisi, L., Massard, C.L., Moya Borja, G.E., Pereira, J.B., 2002. Impacto, economico
578 das principais ectoparasitoses em bovinos no Brasil. Hora Vet. 21, 8-10.
579 Guerrero, F.D., Miller, R.J., Rousseau, M.E., Sunkara, S., Quackenbush, J., Lee, Y.,
580 Nene, V., 2005. BmiGI: a database of cDNAs expressed in *Boophilus*
581 *microplus*, the tropical/southern cattle tick. Insect Biochem. Mol. Biol. 35,
582 585-595.
583 Guerrero, F.D., Nene, V.M., George, J.E., Barker, S.C., Willadsen, P., 2006.
584 Sequencing a new target genome: the *Boophilus microplus* (Acari: Ixodidae)
585 genome project. J. Med. Entomol. 43, 9-16.
586 Guerrero, F.D., Moolhuijzen, P.M., Peterson, D.G., Bidwell, S., Caler, E., Appels, R.,
587 Bellgard, M., Nene, V.M., Djikeng, A., 2010. Reassociation kinetics-based
588 approach for partial genome sequencing of the cattle tick, *Rhipicephalus*
589 *(Boophilus) microplus*. BMC Genomics 11, 374.
590 Harris, T.W., Stein, L.D., 2006. WormBase: methods for data mining and
591 comparative genomics. Methods Mol. Biol. 351, 31-50.
592 Huang, X., Madan, A., 1999. CAP3: A DNA sequence assembly program. Genome
593 Res. 9, 868-877.
594 Kent, W.J., 2002. BLAT--the BLAST-like alignment tool. Genome Res. 12, 656-664.
595 Koski, L.B., Gray, M.W., Lang, B.F., Burger, G., 2005. AutoFACT: an automatic
596 functional annotation and classification tool. BMC Bioinformatics 6, 151.

597 Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., Ussery, D.W.,
598 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes.
599 Nucleic Acids Res. 35, 3100-3108.
600 Lawson, D., Arensburger, P., Atkinson, P., Besansky, N.J., Bruggner, R.V., Butler,
601 R., Campbell, K.S., Christophides, G.K., Christley, S., Dialynas, E., Emmert,
602 D., Hammond, M., Hill, C.A., Kennedy, R.C., Lobo, N.F., MacCallum, M.R.,
603 Madey, G., Megy, K., Redmond, S., Russo, S., Severson, D.W., Stinson, E.O.,
604 Topalis, P., Zdobnov, E.M., Birney, E., Gelbart, W.M., Kafatos, F.C., Louis,
605 C., Collins, F.H., 2007. VectorBase: a home for invertebrate vectors of human
606 pathogens. Nucleic Acids Res. 35, D503-D505.
607 Lawson, D., Arensburger, P., Atkinson, P., Besansky, N.J., Bruggner, R.V., Butler,
608 R., Campbell, K.S., Christophides, G.K., Christley, S., Dialynas, E.,
609 Hammond, M., Hill, C.A., Konopinski, N., Lobo, N.F., MacCallum, R.M.,
610 Madey, G., Megy, K., Meyer, J., Redmond, S., Severson, D.W., Stinson, E.O.,
611 Topalis, P., Birney, E., Gelbart, W.M., Kafatos, F.C., Louis, C., Collins, F.H.,
612 2009. VectorBase: a data resource for invertebrate vector genomics. Nucleic
613 Acids Res. 37, D583-D587.
614 Lew-Tabor, A.E., Moolhuijzen, P.M., Vance, M.E., Kurscheid, S., Valle, M.R.,
615 Jarrett, S., Minchin, C.M., Jackson, L.A., Jonsson, N.N., Bellgard, M.I.,
616 Guerrero, F.D., 2009. Suppressive subtractive hybridization analysis of
617 *Rhipicephalus (Boophilus) microplus* larval and adult transcript expression
618 during attachment and feeding. Vet. Parasitol. 167, 304-320.
619 Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-
620 Wheeler transform. Bioinformatics 26, 589-595.

621 Mans, B.J., Neitz, A.W., 2004. Adaptation of ticks to a blood-feeding environment:
622 evolution from a functional perspective. Insect Biochem. Mol. Biol. 34, 1-17.
623 Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A.,
624 Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro,
625 J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk,
626 G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B.,
627 Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J.,
628 Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P.,
629 Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T.,
630 Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro,
631 K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y.,
632 Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M., 2005. Genome sequencing
633 in microfabricated high-density picolitre reactors. Nature 437, 376-380.
634 Megy, K., Hammond, M., Lawson, D., Bruggner, R.V., Birney, E., Collins, F.H.,
635 2009. Genomic resources for invertebrate vectors of human pathogens, and the
636 role of VectorBase. Infect. Genet. Evol. 9, 308-313.
637 Moolhuijzen, P., Lew-Tabor, A., Morgan, A.T.J., Rodriguez Valle, M., Peterson,
638 D.G., Dowd, S.E., Guerrero, F.D., Bellgard, M.I., Appels, R., 2011. The
639 complexity of *Rhipicephalus (Boophilus) microplus* genome characterised
640 through detailed analysis of two BAC clones. BMC Res. Notes 4, 254.
641 Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping
642 and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5,
643 621-628.

644 Playford, M., Rabiee, A.R., Lean, I.J., Ritchie, M., 2005. Review of research needs
645 for cattle tick control, Phases I and II. In, Meat & Livestock Australia Ltd,
646 Sydney.
647 Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B.,
648 Pertea, G., Sultana, R., White, J., 2001. The TIGR Gene Indices: analysis of
649 gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids
650 Res. 29, 159-164.
651 Ribeiro, J.M., Alarcon-Chaidez, F., Francischetti, I.M., Mans, B.J., Mather, T.N.,
652 Valenzuela, J.G., Wikel, S.K., 2006. An annotated catalog of salivary gland
653 transcripts from *Ixodes scapularis* ticks. Insect Biochem. Mol. Biol. 36, 111-
654 129.
655 Rodriguez-Valle, M., Lew-Tabor, A., Gondro, C., Moolhuijzen, P., Vance, M.,
656 Guerrero, F.D., Bellgard, M., Jorgensen, W., 2010. Comparative microarray
657 analysis of *Rhipicephalus (Boophilus) microplus* expression profiles of larvae
658 pre-attachment and feeding adult female stages on *Bos indicus* and *Bos taurus*
659 cattle. BMC Genomics 11, 437.
660 Saldivar, L., Guerrero, F.D., Miller, R.J., Bendele, K.G., Gondro, C., Brayton, K.A.,
661 2008. Microarray analysis of acaricide-inducible gene expression in the
662 southern cattle tick, *Rhipicephalus (Boophilus) microplus*. Insect Mol. Biol.
663 17, 597-606.
664 Schattner, P., Brooks, A.N., Lowe, T.M., 2005. The tRNAscan-SE, snoscan and
665 snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids
666 Res. 33, W686-689.
667 Snelson, J.T., 1975. Animal ectoparasites and disease vectors causing major
668 reductions in world food supplies. FAO Plant Protect. Bull. 23, 103-117.

669 Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E.,
670 Stajich, J.E., Harris, T.W., Arva, A., Lewis, S., 2002. The generic genome
671 browser: a building block for a model organism system database. Genome
672 Res. 12, 1599-1610.
673 Stewart, N.P., Callow, L.L., Duncalfe, F., 1982. Biological comparisons between a
674 laboratory-maintained and a recently isolated field strain of *Boophilus*
675 *microplus*. Parasitology 68, 691-694.
676 Sunter, J.D., Patel, S.P., Skilton, R.A., Githaka, N., Knowles, D.P., Scoles, G.A.,
677 Nene, V., de Villiers, E., Bishop, R.P., 2008. A novel SINE family occurs
678 frequently in both genomic DNA and transcribed sequences in ixodid ticks of
679 the arthropod sub-phylum Chelicerata. Gene 415, 13-22.
680 Turni, C., Lee, R.P., Jackson, L.A., 2004. A comparison of the immunosuppressive
681 effects of salivary gland extracts from two laboratory strains of *Boophilus*
682 *microplus*. Int. J. Parasitol. 34, 833-838.
683 Turni, C., Lee, R.P., Jackson, L.A., 2007. The effects of salivary gland extracts from
684 *Boophilus microplus* ticks on mitogen-stimulated bovine lymphocytes. Vet.
685 Res. Commun. 31, 545-552.
686 Ullmann, A.J., Lima, C.M., Guerrero, F.D., Piesman, J., Black IV, W.C.., 2005.
687 Genome size and organization in the blacklegged tick, *Ixodes scapularis* and
688 the Southern cattle tick, *Boophilus microplus*. Insect Mol. Biol. 14, 217-222.
689 Wikel, S.K., 1999. Tick modulation of host immunity: an important factor in
690 pathogen transmission. Int. J. Parasitol. 29, 851-859.
691 Wistrand, M., Kall, L., Sonnhammer, E.L.L., 2006. A general model of G protein-
692 coupled receptor sequences and its application to detect remote homologs.
693 Protein Sci. 15, 509-521.
694

## Legends to Figures

**Fig. 1.** Venn diagram summarizing the overlap between tick transcripts (S1 and S2) and the *Rhipicephalus microplus* (*Rmi)* Gene Index (*Bmi*GI Version 2.1). Transcript sequences were clustered between Australian datasets 6 and 7 (Table 1) and the USA BmiGI Version 2.1 gene index (http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=b_microplus) at greater that 98 percent identity (PID). The Venn diagram shows that the Australian datasets 6 and 7 contain 9,652 transcripts not found in the BmiGI.

**Fig. 2.** Screen shot of the CattleTickBase home page. The CattleTickBase main web page is a major resource for *Rhipicephalus microplus* data and analysis. On the CattleTickBase home page pre-computed analyses can be viewed under the "View Analyses" box; this includes genome browsers and full transcript annotations. *Rhipicephalus microplus* sequence datasets are freely available for download from the "Download Data" box. In the "Analysis Tools" box *R. microplus* datasets are available for researchers to conduct their own analysis through a number of listed bioinformatics applications and create customized bioinformatics pipelines. Users are free to access the datasets and analysis tools available within the system, or can create a secure account.

**Fig. 3.** YABI workflow analysis example for the identification of G protein-coupled receptors (GPCRs) from transcript sequences in Dataset 7 of Table 1. The flow diagram on the left hand side shows a simplified three-step bioinformatics workflow, Step 1. Select a file of nucleotide sequences, Step 2. Translate all available open reading frames (ORF) to amino acid sequences, Step 3. Search for GPCRs in the translated ORF. On the right hand side of the figure is a screen capture of the YABI

719 online application; in the "Jobs" tab the automated three-step analysis pipeline
720 labelled "GPCR Rmi FL" is running. Steps 1 and 2 have completed successfully
721 (green circle with white mark) and step 3 is running, waiting for completion indicated
722 by the green bar. The remaining tabs include a "design" tab to create workflows and a
723 "files" tab where all files and results are stored.

**Table 1.** Cattle Tick genome sequence and other relevant datasets available on CattleTickBase.

| Dataset | Description | No. of sequences | No. of bp | References |
|---|---|---|---|---|
| 1 | Cot-selected genomic DNA seqs. | 7,289,230 | 1,798,400,445 | Guerrero et al., 2010 |
| 2 | Cot-selected genomic DNA assembly | 175,226 | 144,709,321 | |
| 3 | BAC[a] end sequences | 10,582 | 7,290,530 | |
| 4 | Full-length sequenced BACs (15 BAC clones) | 15 | 1,502,117 | Guerrero et al., 2010, Moolhuijzen et al., 2011 |
| 5 | BmiGI[b] contigs extended by Cot-selected seqs. | 3,913 | 4,240,351 | Guerrero et al., 2010 |
| 6 | Tick gut transcriptome | 11,333 | 9,228,737 | |
| 7 | Frustrated larval transcriptome | 6,082 | 3,617,080 | |
| 8 | Assembled transcriptome *Rmi*Tr[c] Version 1.0 | 28,893 | 24,673,517 | |
| 9 | NimbleGen microarray analyses GEO dataset GSE20605 (10 arrays) | | | Rodriguez-Valle et al., 2010 |

735 [a]Bacterial Artifical Chromosome

736 [b]*Boophilus microplus* Gene Index at http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=b_microplus

737 [c]*Rhipicephalus microplus* transcriptome

738

739 **Table 2**. Statistics of the 15 sequenced Bacterial Artificial Chromosomes (BACs) compared with Cot-selected genomic DNA.

| BAC | BAC length (bp) | # Mapped Cot DNA reads[a] | BAC Cot % coverage[b] | No. of Genes/ No. of Exons[c] | GC content % |
|---|---|---|---|---|---|
| BM-005-B21 | 92,305 | 4,068 | 82 | 12/42 | 45 |
| BM-005-G14 | 135,319 | 3,138 | 58 | 5/76 | 44 |
| BM-013-M17 | 90,249 | 2,286 | 64 | 9/44 | 46 |
| BM-026-P08 | 108,580 | 2,649 | 70 | 11/62 | 45 |
| BM-031-L02 | 125,915 | 3,115 | 65 | 14/64 | 46 |
| BM-118-H10 | 92,057 | 3,658 | 72 | 10/50 | 46 |
| BM-004-A11 | 103,837 | 5,085 | 79 | 15/66 | 45 |
| BM-006-B07 | 102,433 | 4,416 | 79 | 12/53 | 46 |
| BM-010-J12 | 172,065 | 3,014 | 53 | 24/124 | 48 |
| BM-012-E08 | 51,489 | 146 | 6 | 3/14 | 58 |
| BM-074-Random-F12 | 95,687 | 1,534 | 61 | 15/40 | 40 |
| BM-077-Random-J09 | 103,645 | 2,777 | 59 | 12/52 | 47 |
| BM-129-CzEst9-N14 | 126,498 | 3,919 | 64 | 14/71 | 46 |
| BM-066-M07 | 151,523 | 3,608 | 51 | 16/70 | 46 |
| BM-077-G20 | 94,838 | 1,692 | 54 | 8/64 | 46 |

740

741 [a]The number of Cot DNA reads mapped onto BAC sequence

742 [b]The percentage coverage of BAC sequence length by aligned Cot DNA reads

743 [c]Count of predicted genes and total number of exons for each BAC sequence

744

745 **Table 3.** *Ixodes scapularis* (*Isc)* transcript orthologous sequences in the *Rhipicephalus microplus* (*Rmi)*Tr Version 1.0.

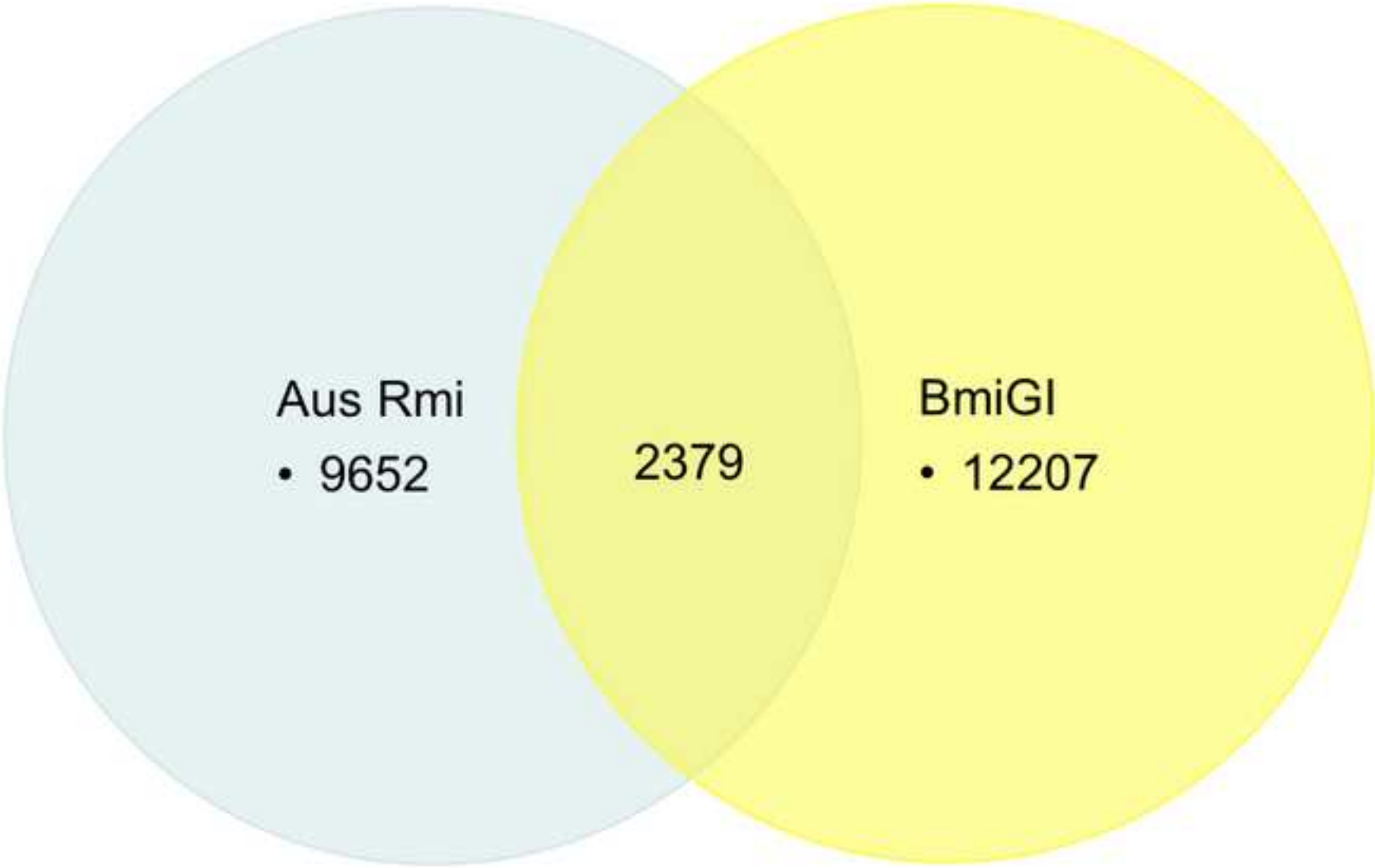
746

747

| Data | *Rmi*Tr1.0 | | *Isc* Gene Index Version 3.0 | |
|---|---|---|---|---|
| | Total Seqs. | Orthologues with *Isc* | Total Seqs. | Orthologues with *Rmi*Tr |
| Overall Statistics | 28,893 | 3,829 | 38,392 | 4,948 |
| Cytochrome P450 | 115 | 55 | 357 | 235 |
| Esterases | 81 | 14 | 133 | 33 |
| Glutathione S-transferases | 39 | 28 | 96 | 60 |

Figure 1 colour





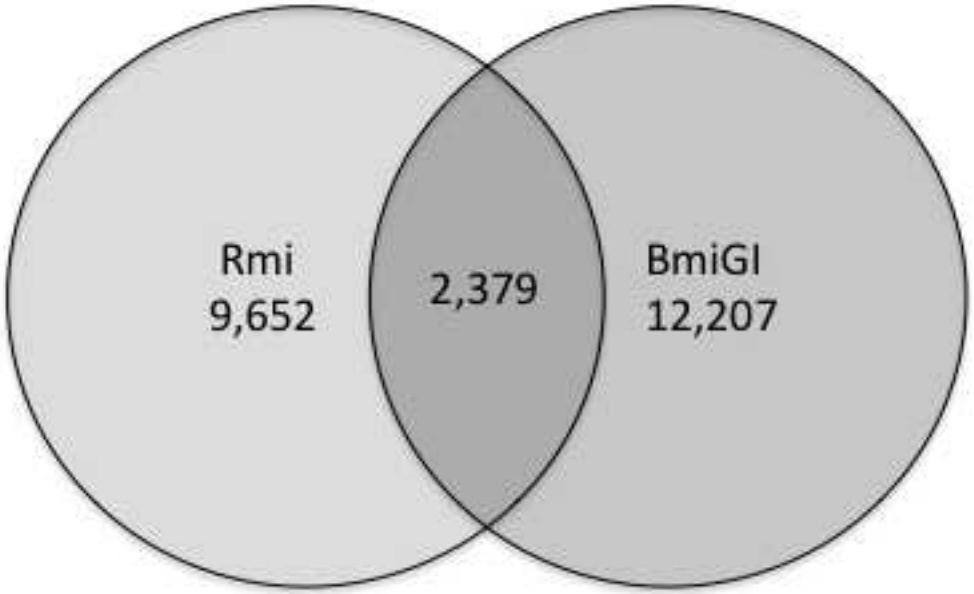



Figure 1

Figure 2



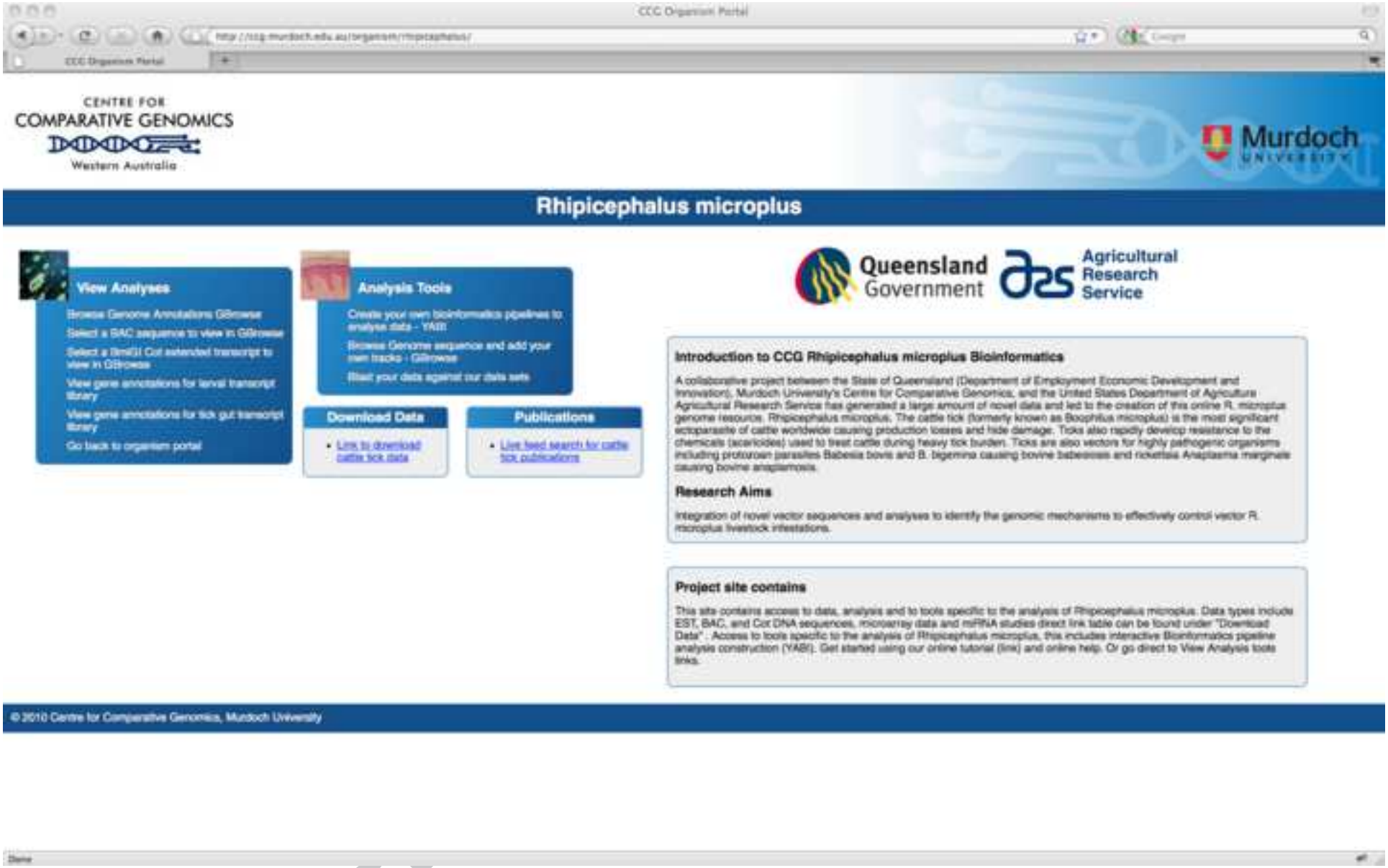
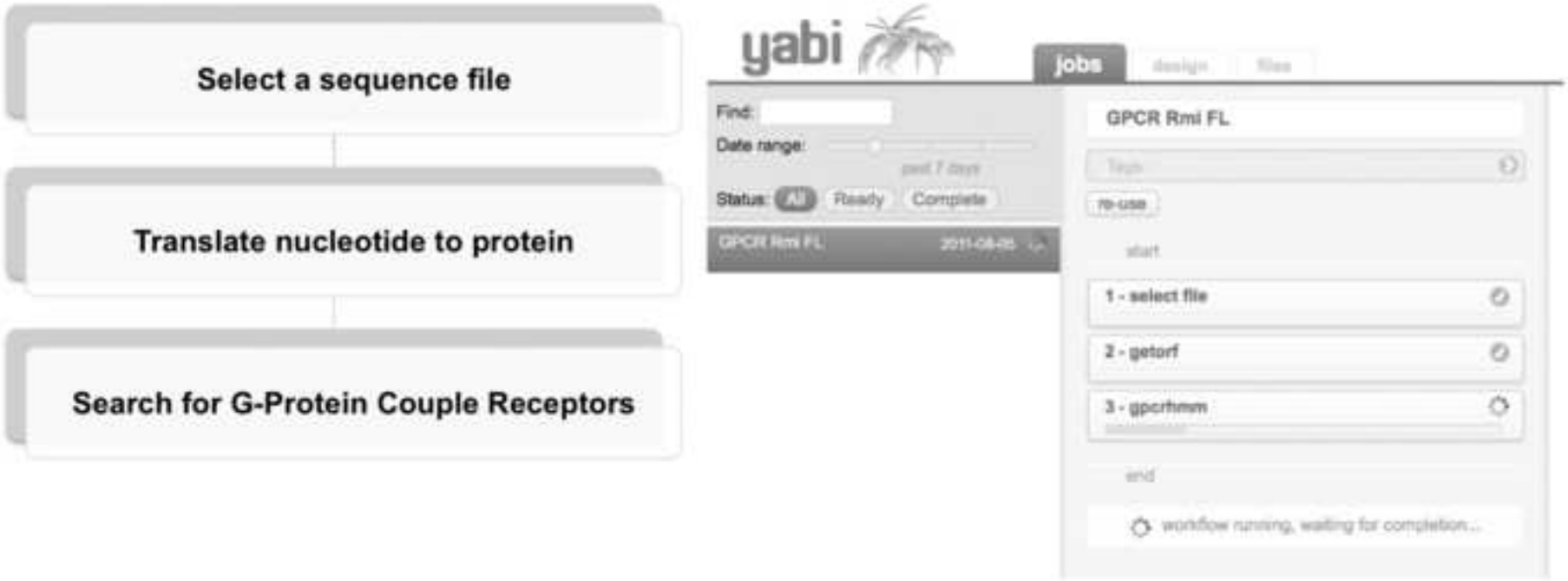


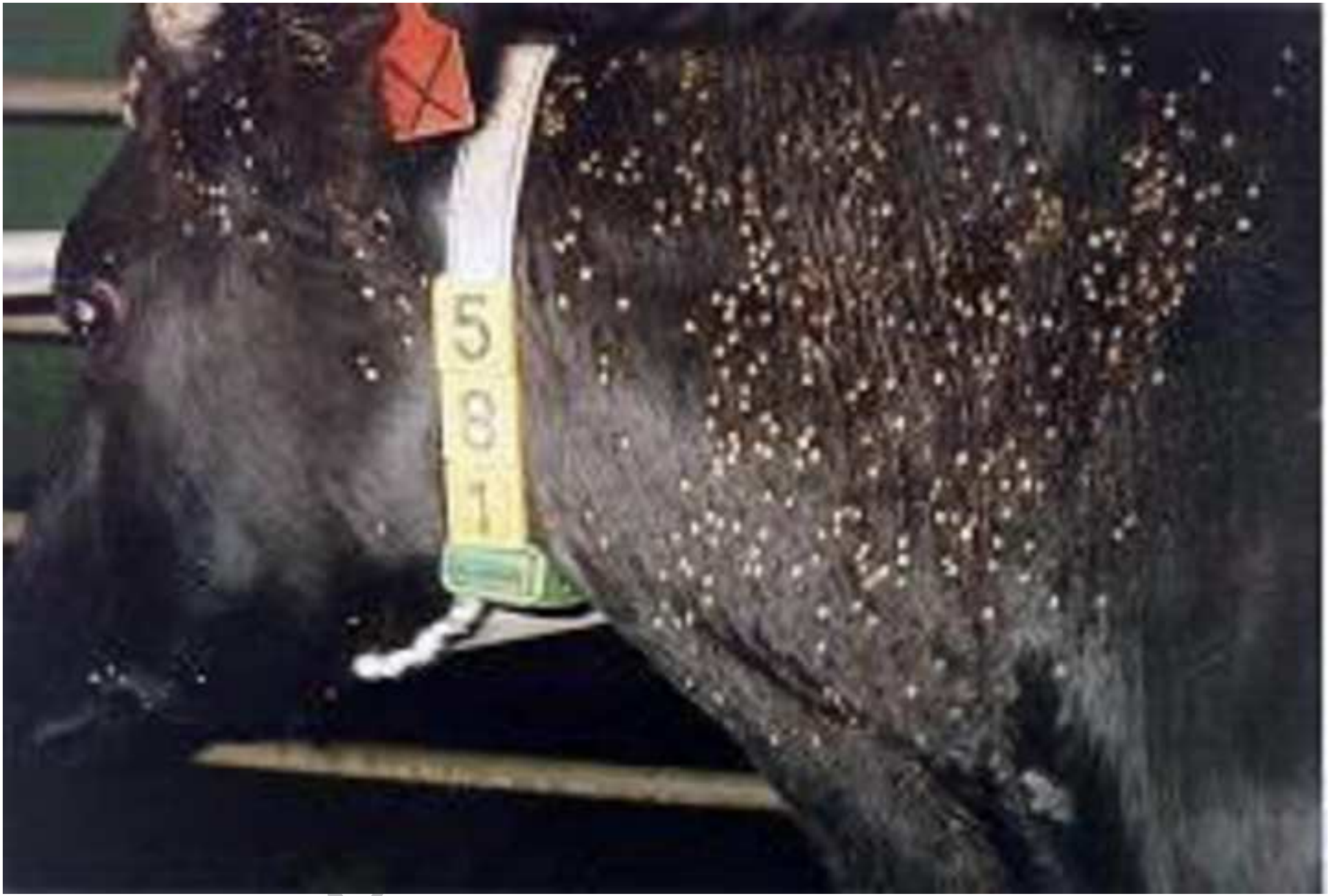
Figure 3 colour



**Figure 3 Black and white**







*Highlights



>Consortium formed to begin sequencing cattle tick genome. >Acquired and assembled over 1.8 Gbp from gene-enriched regions of genome. >Web resource developed for access to data through web-based environment.