Ticks and Tick-borne Diseases 6 (2015) 450–462

Contents lists available at ScienceDirect

Ticks and Tick-borne Diseases

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

Original article

Next-generation sequencing as means to retrieve tick systematic markers, with the focus on *Nuttalliella namaqua* (Ixodoidea: Nuttalliellidae)

Ben J. Mans[a,c,d,*], Daniel de Klerk[a], Ronel Pienaar[a], Minique H. de Castro[a,b], Abdalla A. Latif[a,c]

[a] Parasites, Vectors and Vector-borne Diseases, Agricultural Research Council-Onderstepoort Veterinary Institute, Onderstepoort 0110, South Africa
[b] The Biotechnology Platform, Agricultural Research Council-Onderstepoort Veterinary Institute, Onderstepoort 0110, South Africa
[c] Department of Veterinary Tropical Diseases, University of Pretoria, Pretoria, South Africa
[d] Department of Life and Consumer Sciences, University of South Africa, South Africa

ARTICLE INFO

Article history:
Received 21 November 2014
Received in revised form 6 March 2015
Accepted 8 March 2015
Available online 22 April 2015

Keywords:
Genomics
Phylogenetics
Molecular systematics
Next-generation sequencing

ABSTRACT

Nuclear ribosomal RNA (18S and 28S rRNA) and mitochondrial genomes are commonly used in tick systematics. The ability to retrieve these markers using next-generation sequencing was investigated using the tick *Nuttalliella namaqua*. Issues related to nuclear markers may be resolved using this approach, notably, the monotypic status of *N. namaqua* and its basal relationship to other tick families. Four different Illumina datasets (~55 million, 100 bp reads each) were generated from a single tick specimen and assembled to give 350k–390k contigs. A genome size of ~1 Gbp was estimated with low levels of repetitive elements. Contigs (>1000 bp, >50-fold coverage) present in most assemblies ($n = 69$), included host-derived 18S and 28S rRNA, tick and host-derived transposable elements, full-length tick 18S and 28 rRNA, the mitochondrial genome in single contig assemblies and the histone cassette. Coverage for the nuclear rRNA genes was above 1000-fold confirming previous sequencing errors in the 18S rRNA gene, thereby maintaining the monotypic status of this tick. Nuclear markers for the soft tick *Argas africolumbae* were also retrieved from next-generation data. Phylogenetic analysis of a concatenated 18S–28S rRNA dataset supported the grouping of *N. namaqua* at the base of the tick tree and the two main tick families in separate clades. This study confirmed the monotypic status of *N. namaqua* and its basal relationship to other tick families. Next-generation sequencing of genomic material to retrieve high quality nuclear and mitochondrial systematic markers for ticks is viable and may resolve issues around conventional sequencing errors when comparing closely related tick species.

Introduction

Molecular systematics of arthropods makes use of nuclear and mitochondrial genes as markers. Most common markers include the full mitochondrial genome, mitochondrial 16S ribosomal RNA, nuclear 18S and 28S ribosomal RNA and their ITS regions (Giribet and Edgecombe, 2012). These genes are generally abundant and amenable to samples with limited available biological material; they have no introns and are highly conserved allowing PCR amplification using universal primers. In most cases several independent PCR amplification steps need to be performed and several clones need to be sequenced, making targeting of independent genes cumbersome and increasing the probability for sequencing errors. For tick systematics these basic approaches have been extensively used (Black and Piesman, 1994; Black et al., 1997; Black and Roehrdanz, 1998; Klompen et al., 2000; Shao et al., 2004, 2005; Klompen et al., 2007). Next-generation sequencing has been used to sequence amplified nuclear and mitochondrial PCR products (Burger et al., 2012, 2013, 2014; Xiong et al., 2013). However, recently next generation sequencing of total genomic DNA without amplification of PCR products was used to assemble the mitochondrial genomes from the ticks *Nuttalliella namaqua* and *Argas africolumbae* (Mans et al., 2012). The possibility of using this approach to find other markers for tick systematics were investigated in the current study, focusing specifically on *N. namaqua*, since some biological questions regarding its monotypic status and its relationship to the

* Corresponding author at: Parasites, Vectors and Vector-borne Diseases, Agricultural Research Council-Onderstepoort Veterinary Institute, Onderstepoort 0110, South Africa. Tel.: +27 125299200.
E-mail address: mansb@arc.agric.za (B.J. Mans).

http://dx.doi.org/10.1016/j.ttbdis.2015.03.013
1877-959X/

other tick families exist that may be amenable to next generation sequencing.

Ticks constitute three families, Argasidae (soft ticks ~200 species), Ixodidae (hard ticks ~700 species) and the Nuttalliellidae (Guglielmone et al., 2010). The latter family comprise one species, *N. namaqua*, and has been considered to be the "missing link" between the hard and soft tick families as well as being a living fossil (Mans et al., 2011). Differences in the 18S rRNA gene were detected for geographically isolated *N. namaqua* populations (Horak et al., 2012). This is of interest, since it would negate the monotypic status of *N. namaqua* and could suggest that many undiscovered species of this family still exist, given its general host preference (skinks, geckos, girdled lizards, hyrax, meerkat, murid rodents and possibly birds), and its wide distribution from southern Africa (Namibia and South Africa) to Tanzania in East Africa (Keirans et al., 1976; Mans et al., 2014).

Analysis of the nuclear 18S rRNA and mitochondrial genes indicated that *N. namaqua* group basal to the Argasidae and Ixodidae (Mans et al., 2011, 2012; Gu et al., 2014; Chen et al., 2014). Other studies using concatenated 18S–28S rRNA gene sets suggested that *N. namaqua* does not group basal to the other tick families, but shows a closer affinity to the Argasidae (Burger et al., 2013), or that its relationship to the other families was unresolved (Burger et al., 2014), or that it shows a closer relationship to the Ixodidae using mitochondrial data (Burger et al., 2014). No 28S data were included for *N. namaqua* or Argasidae in these latter studies and raised the question whether its inclusion would affect the systematic analysis (Wiens and Morrill, 2011; Roure et al., 2013), since its placement at the base of the tick tree or as part of the main tick families affects conclusions regarding the ancestral tick lineage (Mans et al., 2012).

The current study shows that a next-generation sequencing approach is useful to retrieve the major nuclear and mitochondrial systematic markers from a single tick specimen. This includes full-length 28S rRNA sequences for *N. namaqua* and *A. africolumbae*, histone and transposable element sequences not previously reported. Differences in 18S rRNA sequences could be attributed to conventional sequencing errors in previous reported sequences, thereby maintaining the monotypic status of *N. namaqua*, while inclusion of the 28S rRNA sequences in a nuclear phylogenetic analysis support the basal position of *N. namaqua* in the tick tree. Host markers were also retrieved by next generation sequencing.

Materials and methods

Next generation sequencing of tick genomic DNA

Genomic DNA from a single *N. namaqua* (40 ng) was submitted to the Biotechnology Platform Next Generation Sequencing Service of the Agricultural Research Council (South Africa). Samples were processed using the Nextera DNA sample preparation kit (Epicenter) and sequenced using the Illumina HiScanSQ (Illumina). Four different datasets of ~5 Gbp (~55 million reads) paired-end reads with lengths of 100 bp were analyzed and these were unique datasets from that previously reported (Mans et al., 2012). Data were processed using the CLC Genomics Workbench v5.1 software package, imported using a range of 100–500 bp, quality trimmed (0.05 quality limit), Nextera adapters removed and the last 19 bp trimmed to give reads with an average length of 78 bp. Reads were de novo assembled using assembly parameters: mismatch cost-2, insertion cost-3, deletion cost-3, length fraction-0.9, similarity-0.9, minimum contig length-200, word size-23, bubble size-50. Data were filtered to give contigs with molecular sizes >1 kb and average coverage >50. Contigs with multiple regions with no coverage (multiple *N*'s due to paired read assembly) were discarded. Reciprocal BLAST analysis were used to determine best orthologous hits between the datasets and contigs were retained if they found hits in at least three of the databases. Consensus sequences were derived by multiple alignment of the reciprocal best BLAST hits using ClustalX alignment (Jeanmougin et al., 1998) and Genedoc (Nicholas et al., 1997). Consensus sequences were trimmed to only include consensus regions from all contigs and were analyzed using BLASTN or BLASTX analysis (Altschul et al., 1990), and hits with *E*-values below $E-5$ were considered significant. To obtain final mapping statistics, data from the different assemblies were mapped to the consensus sequences using the same parameters previously used for assembly. Consensus sequences were deposited in Genbank (KF925832–KF925880). A dataset for *A. africolumbae* (Mans et al., 2012) were mined in a similar manner to obtain full-length 18S rRNA (JQ731646) and 28S rRNA genes (KF984488).

Estimating repetitive elements

Repetitive element content was estimated using RepeatMasker (Smit et al., 1996–2004), by analyzing 700 000 reads (1.2%) from each dataset. The option "–species all" was used to screen all possible repeat structures present in the RepeatMasker databases (RepeatMasker and RepBase version 20120418).

Estimating genome size

(A) Using the average coverage peak obtained for the different contigs (3.7 coverage), which resemble coverage of unique genes and dividing the total size of the reads used in each assembly by this number (Fig. 1). (B) The combined datasets (~16.5 Gbp) were analyzed in Kmergenie (Chikhi and Medvedev, 2014) using a variety of kmers (15, 20, 25, 30, 35, 40, 45) and assuming a diploid model, to estimate the unique haploid average coverage, which was used to estimate genome size by dividing the total length of the reads (16552689327 bp) by the unique average coverage (peak height ranging from 15 to 20). (C) Mapping the combined datasets to 18737 open reading frames derived from a salivary gland transcriptome (manuscript in preparation) using CLC Genomics Workbench. The frequency distribution for the average coverage was plotted to determine the coverage depth with highest frequency (peak height = 15). Genome size was then determined using the formula: Genome size = Genome reads × Read length/Highest frequency coverage depth (Hu et al., 2011).

Phylogenetic analysis

Acarine 28S and 18S rRNA sequences were retrieved from the non-redundant database by BLASTN analysis (Altschul et al., 1990). Only sequences from species with representatives of both sequences were used for downstream analysis. Sequences for the different datasets were aligned separately using ClustalX (Jeanmougin et al., 1998). Alignments were manually inspected, adjusted and trimmed and then concatenated to yield the final super-matrix used for phylogenetic analysis.

Bayesian analysis was performed using MrBayes 3.1.2 (Ronquist and Huelsenbeck, 2003). A general time reversible (GTR) model of nucleotide substitution with a proportion of invariant sites and a gamma distribution of among site heterogeneity using the nst = 6 rates = ingamma command was used. Four categories were used to approximate the gamma distribution and two runs were performed simultaneously, each with four Markov chains (one cold, three heated) which ran for 5,000,000 generations. The first 2,000,000 generations were discarded from the analysis (burnin) and every 100th tree was sampled to calculate a 50% majority-rule