Infection, Genetics and Evolution 10 (2010) 304–310

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Deep resequencing of *Trichinella spiralis* reveals previously un-described single nucleotide polymorphisms and intra-isolate variation within the mitochondrial genome

Kristen M. Webb, Benjamin M. Rosenthal *

Animal Parasitic Diseases Laboratory, Agricultural Research Service, United States Department of Agriculture, Building 1180, BARC-East, Beltsville, MD 20705, USA

ARTICLE INFO

Article history:
Received 28 October 2009
Received in revised form 23 December 2009
Accepted 11 January 2010
Available online 18 January 2010

Keywords:
Trichinella spiralis
Mitochondrial genome
Next generation sequencing
Intra-isolate variability

ABSTRACT

The phylogeny and historical dispersal of *Trichinella spp*. have been studied, in part, by sequencing portions of the mitochondrial genome. Such studies rely on two untested beliefs: that variation in a portion is representative of the entire mitochondrial genome, and that each isolate is characterized by only one mitochondrial haplotype. We have used next generation DNA sequencing technology to obtain the complete mitochondrial genome sequence from a second isolate of *T. spiralis*. By aligning it to the only previously sequenced genome, we sought to establish whether the exceptionally deep sequencing coverage provided by such an approach could detect regions of the genome which had been misassembled, or nucleotide positions which may vary within an isolate. The new data broadly confirm the gene order and sequence assembly for protein-coding regions. However, in the repetitive non-coding region, alignment to the previously published genome sequence proved difficult. Such discrepancies may represent true biological variation, but may rather result from methodological or algorithmic sources. Within the 13,902 bp protein-coding region, 7 polymorphisms were identified. Six of these polymorphisms occurred within protein-coding genes and three alter an amino acid sequence, one occurred in a tRNA-Ile sequence, and four were found to vary within our isolate. Thus, comparing only two isolates of *T. spiralis* has enabled the discovery of previously unrecognized variation within the species. Characterizing diversity within and among the mitochondrial genomes of additional species of *Trichinella* would undoubtedly yield further insights into the diversification history of the genus. Our study affirms that next generation DNA sequencing technology can reliably characterize a complete mitochondrial genome.

1. Introduction

Variation in the mitochondrial genome provides an important means to understand the relationships among closely related, recently diverged animal taxa and to characterize the migratory history of populations (Avise et al., 1987; Moritz et al., 1987; Simon et al., 2006). Substitutions generally accumulate within the mitochondrial genome faster than within most nuclear genomic regions, and the mitochondrion is inherited as a non-recombinant, maternal haplotype making it a suitable marker to study population dynamics. Mitochondrial sequence data are typically represented as though only a single genome had been sequenced, assuming uniformity among the upwards of 10,000 mitochondrial genomes that may occupy just a single cell.

The first complete nucleotide sequence of the *Trichinella spiralis* mitochondrial genome provided valuable insights regarding gene order and genome composition (Lavrov and Brown, 2001). Typical of the mitochondria of metazoans, that of *T. spiralis* was found to encode 37 genes within its first 13,902 base pairs (bps). Two large, non-coding repeats (1232 bp each) were characterized within the region between the *nad1* and *nad2* genes, the largest non-coding region of the genome. It was not then possible to obtain the complete sequence of the non-coding region, and it was presumed that additional repeat units may lie between those already identified, perhaps resulting in size variation among mitochondrial genomes within and/or between individuals (Lavrov and Brown, 2001).

The extent and geographic pattern of mitochondrial sequence variation from a 3100 bp portion of the organellar genome has contributed to reconstructing the dispersal history of this zoonotic parasite (Rosenthal et al., 2008). This fragment, representing 18.5% of the mitochondrial genome, was characterized in 14 isolates of *T. spiralis* from Europe and North America (collectively termed as "western" isolates), as well as 8 isolates derived from Asia. Among the western *T. spiralis* sequences, only a single polymorphism was identified, whereas eight polymorphisms occurred among the

* Corresponding author. Tel.: +1 301 504 5408; fax: +1 301 504 8979.
E-mail address: Benjamin.Rosenthal@ars.usda.gov (B.M. Rosenthal).

1567-1348/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2010.01.003

Asian isolates. The minimal differentiation evident among western isolates, obtained using traditional Sanger sequencing of PCR products, was interpreted as suggesting they shared an especially close evolutionary relationship. Reinforced by scant variation in microsatellite markers, and interpreted within the history of swine domestication in Europe, it was proposed that a population bottleneck may have occurred in Central Europe roughly 16,000–6000 years before present (ybp) (Rosenthal et al., 2008). To date, this study represents the broadest survey of intra-specific genetic variability for this zoonotic parasite.

Greater precision in such estimates might be attainable from complete mitochondrial resequencing efforts. 'Next generation' sequencing, via the Illumina Genome Analyzer II (GAII) platform, characterizes a library of short DNA fragments flanked (by means of ligation) with adapters that facilitate their simultaneous sequencing, en masse. Millions of raw sequencing reads may result from such a procedure, providing the possibility that an organellar genome can be sequenced to far greater coverage than can be expected from traditional Sanger methods. Variation can thereby be identified both within and between isolates.

We have sequenced a second mitochondrial genome from an isolate of *T. spiralis*, using such methods capable of resolving instances of intra-isolate variability. The goals of this study were (a) to resequence the *T. spiralis* mitochondrial genome and compare to the results of Lavrov and Brown (2001), (b) to use next generation technologies to perform this sequencing effort, demonstrating the utility of this method, and (c) to define the extent of uniformity across the mitochondrial genome and among the multiple mitochondrial genomes comprising an isolate of *T. spiralis*.

2. Materials and methods

We characterized the complete mitochondrial genome of the 'Beltsville pig' isolate of *T. spiralis*, first isolated in the United States in 1936 (as described by Murrell et al. (1985) and deposited in the International Trichinella Reference Center as ISS #31) (www.iss.it/site/Trichinella/). This isolate had been subjected to serial passage in rats until 1991, when DNA was extracted using conventional phenol/chloroform procedures. Typically, such a DNA preparation would be presumed to derive from hundreds to thousands of individual larvae, which (owing to the long period of serial passage) would be expected to be full siblings of a highly inbred line.

Complete genomic DNA of isolate ISS #31 was sequenced using Next Generation technology on an Illumina GAII. The library preparation, DNA sequencing, and raw data processing via the Illumina Genome Analyzer Analysis Pipeline were carried out in accordance with the manufacture's protocols for single-ended, 36 bp reads. The only exception involved fractionation of the genomic DNA via sonication rather than via nebulization.

To align the resultant 36 bp short reads against a reference sequence of the mitochondrial genome (Genbank accession # NC_002681.1) (Lavrov and Brown, 2001), we used the program "Mapping and Assembly with Quality" (MAQ Version 0.7.1) (maq.sourceforge.net). This procedure separated reads corresponding to the mitochondrial genome from those corresponding to the nuclear genome based on sequence similarity. Reads were mapped to the reference genome sequence provided that doing so presumed no more than two mismatches.

Additionally, reconstruction of the mitochondrial genome sequence from our short reads without reference to prior information was done using a *de novo* assembler, Velvet 0.7.53 (Zerbino and Birney, 2008). The *de novo* assembly was derived from the entire collection of raw sequence reads (without prior *in silico* identification to likely constituents of the mitochondrial genome) and used a hash length of 29. Only those resultant contigs greater than 100 bp in length were evaluated. The mitochondrial candidate contigs and their spatial orientation were determined by means of pairwise alignment to the reference mitochondrial genome using Sequencher 4.7 (www.genecodes.com).

Candidate single nucleotide polymorphisms (SNPs) were identified using the MAQ algorithms and the pairwise alignment of the *de novo* contigs and reference genome. Dr. Dennis Lavrov graciously provided the genomic DNA that had been used to derive the reference genome NC_002681.1, which together with genomic DNA from ISS #31 was used to validate candidate SNPs via PCR and traditional Sanger sequencing. For each region in question, success was defined as bi-directional sequence with clean, high quality chromatograms, resulting in 2× coverage. A complete list of all primers is presented in supplemental Table 1. Sequences were edited and aligned using Sequencher Version 4.7. The location and any potential alterations to the translation from DNA to amino acid caused by each SNP were established. Finally, the frequency of discordance between a given base and the consensus base (derived from a deep collection of such bases) was used to estimate a sequencing error rate at those coding region sites not identified as polymorphic by MAQ.

Signatures of selection were estimated using MEGA 4 (Tamura et al., 2007) from rate differences at synonymous and non-synonymous sites via Nei-Gojobori method (Nei and Gojobori, 1986), and 4-fold degenerate and non-degenerate sites via the Kumar method (Nei and Kumar, 2000). The distribution of variable sites across the coding region was evaluated by aligning the newly assembled genome and the reference genome sequence. Evidence of selection was evaluated using the codon-based *Z*-test also available within MEGA 4. Pairwise sequence divergences were then estimated for those 3100 bp, and for the entire 13,902 bp coding region.

3. Results

Total genomic DNA of *T. spiralis* isolate ISS #31 was sequenced using massively parallel technologies. From a single lane of an 8 lane flowcell, 241,844,760 bases of raw sequence were generated. This would represent an equivalent of just over four nuclear genomes of *T. spiralis*, estimated at 56.8 million bases (Mitreva and Jasmer, 2006) or the equivalent of almost 15,000 mitochondrial genomes, estimated as 16,706 bases (Lavrov and Brown, 2001).

A total of 3,025,810 nucleotides comprised the sequence reads attributed to the coding region of mitochondrial genome. Only 16,799 (0.55%) of these disagreed with the consensus base derived from the population of aligned reads. Deep sequencing coverage (averaging 217.8 reads per site within the unambiguously assembled coding region, defined below) essentially precluded any possibility of erroneous consensus base calls attributable to sequencing error.

The scaffold assembly (generated by MAQ) of our small sequence reads necessarily corresponded precisely in length to that of the reference genome sequence (16,706 bp) (as this algorithm does not allow for the insertion of alignment gaps). The *de novo* assembly of all short reads resulted in 1005 contigs. An alignment of the *de novo* assembled contigs to the reference sequence revealed that seven of these contigs were mitochondrial in origin and covered 15,371 bp (92%) of the genome, with the coding region (bases 1–13,902) sequenced in its entirety (Fig. 1). Two of the *de novo* assembled contigs did not overlap with the others but rather aligned to immediately adjacent positions in the reference sequence (Fig. 1). These junctions occurred at positions 4335/4336 and 9650/9651. By using NCBI BLAST to align the mitochondrial genome reference sequence to itself, the presence of a small repeat was discovered spanning positions 4303–4363 and