522–533 Nucleic Acids Research, 2010, Vol. 38, No. 2 Published online 5 November 2009
doi:10.1093/nar/gkp883

An integrated pipeline for next-generation sequencing and annotation of mitochondrial genomes

Aaron R. Jex[1,*], Ross S. Hall[1], D. Timothy J. Littlewood[2] and Robin B. Gasser[1]

[1]Department of Veterinary Science, The University of Melbourne, Victoria 3030, Australia and [2]Department of Zoology, The Natural History Museum, London SW7 5BD, UK

Received August 5, 2009; Revised September 9, 2009; Accepted October 2, 2009

ABSTRACT

Mitochondrial (mt) genomics represents an understudied but important field of molecular biology. Increasingly, mt dysfunction is being linked to a range of human diseases, including neurodegenerative disorders, diabetes and impairment of childhood development. In addition, mt genomes provide important markers for systematic, evolutionary and population genetic studies. Some technological limitations have prevented the expanded generation and utilization of mt genomic data for some groups of organisms. These obstacles most acutely impede, but are not limited to, studies requiring the determination of complete mt genomic data from minute amounts of material (e.g. biopsy samples or microscopic organisms). Furthermore, post-sequencing bioinformatic annotation and analyses of mt genomes are time consuming and inefficient. Herein, we describe a high-throughput sequencing and bioinformatic pipeline for mt genomics, which will have implications for the annotation and analysis of other organellar (e.g. plastid or apicoplast genomes) and virus genomes as well as long, contiguous regions in nuclear genomes. We utilize this pipeline to sequence and annotate the complete mt genomes of 12 species of parasitic nematode (order Strongylida) simultaneously, each from an individual organism. These mt genomic data provide a rich source of markers for studies of the systematics and population genetics of a group of socioeconomically important pathogens of humans and other animals.

INTRODUCTION

Mitochondrial (mt) genomics has received increased attention in recent years (1). In human medicine, mt dysfunction is being explored as a contributor to neurodegenerative disorders, such as Parkinson's disease (2,3), multiple sclerosis (4) and Alzheimer's disease (5), disabilities linked to childhood (6) or ageing (1), complications associated with diabetes (7) and even decreased efficacy of anti-retroviral treatments (8). Many of these hypothesized links are in the early stages of investigation and are the subject of some controversy (9); thus, the contribution that mutations in the mt genome make to mt dysfunction is in need of extensive study. In addition to potential links to disorders, mt genes and genomes have proven utility as molecular markers for systematic and population genetic studies across a broad range of animal groups (10–16). Among pathogenic organisms, mt genes and genomes have demonstrated utility in epidemiological studies, often allowing the investigation of links the genetics of pathogens (e.g. population or 'strain' differentiation and speciation) and the characteristics of diseases (e.g. pathogenesis, host affiliations, virulence and drug resistance) (10–12,17–20).

Despite the utility of mt genomes, technical obstacles have limited the full potential of mt markers. Usually, mt genome sequencing has relied on the isolation of mtDNA from the organism under study. From large organisms, such as vertebrates, the purification of mtDNA in sufficient quantities to allow direct sequencing (without amplification and/or cloning) is achievable from fresh material (21,22). However, direct mt sequencing is not possible from specific tissues (e.g. neurons) or biopsy samples, may not be possible from material preserved for long periods, and cannot be applied directly to small invertebrates, such as parasitic worms. Although an increased availability of material can be achieved through the 'pooling' of multiple samples, often this is not possible due to the need to specifically examine one sample (e.g. from biopsy), the high level of sequence polymorphism occurring between or among multiple individuals (e.g. parasitic nematodes) or unavailability of additional material. Recent advances in long-range polymerase chain reaction (long-PCR) amplification

*To whom correspondence should be addressed. Tel: +61 3 9731 2294; Fax: +61 3 9731 2366; Email: ajex@unimelb.edu.au

(23,24) and the subsequent sequencing of mt amplicons by 'primer-walking' have overcome many of these obstacles. For species for which the mt genome is well characterized, this process is rapid and relatively inexpensive. However, for uncharacterized species (e.g. most parasites), *de novo* sequencing by the approach of 'primer-walking' can be laborious, costly and inefficient. For highly AT-rich templates, such as the mt genomes of nematodes (10,11), primer-walking is significantly hampered by short sequence reads (~100 bp) and the limited availability of suitable regions for primer design. For any species, the depth of coverage (DOC) achieved using a standard bi-directional primer-walking approach is low (usually two times) (25,26), limiting opportunities for the detection of point mutations.

Advances in high-throughput sequencing technologies, such as massively parallel picolitre reactor sequencing [='454 technology'; (27)] provide a means for the rapid and effective characterization of large numbers of mt genomes, and might be suitable for overcoming many of the obstacles associated with methods based on primer-walking. Recently, this sequencing approach was evaluated for selected mt genomes of parasitic nematodes (26,28), preserved, prehistoric human tissue (29) and fish (30). Although the use of 454 technology for mt genomic sequencing represents an important advance, a major limitation still remains the ability to rapidly process such sequence data. In the present study, (i) we used PCR-coupled 454 technology to sequence the complete mt genomes from 12 socioeconomically important parasitic nematodes from tiny amounts of material from individual adult worms, and (ii) constructed a prototypical bioinformatic pipeline for the automated annotation and analysis of the mt genomic sequence datasets produced.

METHODS

Collection and procurement of parasitic nematodes

Worms were collected upon necropsy of their definitive hosts under the Scientific Procedures Premises License for the Faculty of Veterinary Science, The University of Melbourne (SPPL045). Individuals representing adults of *Cylicocyclus insignis* (Strongyloidea: Strongylidae) and *Strongylus vulgaris* (Strongyloidea: Strongylidae) were each collected at necropsy from the large intestine of a horse from Victoria, Australia. Individual adults of *Chabertia ovina* (Strongyloidea: Chabertiidae), *Cooperia oncophora* (Trichostrongyloidea: Cooperiidae), *Teladorsagia circumcincta* (Trichostrongyloidea: Haemonchidae), *Trichostrongylus axei* and *Trichostrongylus vitrinus* (both Trichostrongyloidea: Trichostrongylidae), were collected from the gastrointestinal tracts from sheep in Victoria, Australia. An adult of *Mecistocirrus digitatus* (Trichostrongyloidea: Haemonchidae) was collected from the intestine of a sheep from China. Individual adult males of *Metastrongylus pudendotectus* and *Metastronglyus salmi* (both Metastrongyloidea: Metastrongylidae) were each collected from the terminal bronchi in the lungs from a pig from Estonia. An individual adult male of *Oesophagostomum dentatum* (Strongyloidea: Chabertiidae) was collected from the large intestine of an experimentally infected pig in Denmark. Lastly, an individual representing an adult male of *Syngamus trachea* (Metastrongyloidea: Syngamidae) was collected from the trachea of an adult Australian Magpie (*Gymnorhina tibicen*) in Victoria, Australia. Each worm was transferred to a sterile, screw-top cryogenic tube (Nunc) and frozen (−70°C) in a minimum amount of physiological saline. After thawing, total genomic DNA was isolated from a mid-body section (usually 5–15 mm in length) of each worm using a standard sodium dodecylsulphate/proteinase K treatment (31), followed by purification over a mini-column (Wizard CleanUp, Promega). The specific identity of each specimen was verified by PCR-based amplification of the second internal transcribed spacer (ITS-2) of nuclear ribosomal DNA using an established method, followed by mini-column purification (Wizard PCR-Preps, Promega) of the amplicon and subsequent automated sequencing (BigDye chemistry v.3.1) (32).

Next-generation sequencing of mt genomes

The complete mt genome of each worm/species was amplified by long-PCR (BD Advantage 2; BD Biosciences) as two overlapping amplicons ('large' and 'small'), using the protocol described by Hu *et al*. (23), with appropriate positive and negative (i.e. no template) controls. Amplicons were consistently produced from the positive control samples and in no case was a product detected for any of the negative controls. Amplicons were then purified over a mini-column (Wizard, Promega) and quantified spectrophotometrically using a ND-1000 UV-VIS spectrophotometer v.3.2.1 (NanoDrop Technologies). Following electrophoretic verification of their quality, the two amplicons (~5 and ~10 kb; 2.5 µg of each), spanning the mt genome of each species, were pooled and subsequently sequenced using the 454 Genome Sequencer FLX (Roche) according to the protocol provided (27). The consensus mt genome sequences (GenBank accession numbers GQ888711–GQ888722) were each assembled automatically (using the Newbler Program, Roche) from thousands of individual ~300 bp 'reads' based on a majority rule-threshold among all reads representing each contig.

Custom-built bioinformatic pipeline for the automated annotation and analysis of sequence data

Following the assembly, the genes and features of each mt genome from each worm were annotated using an 'in-house' automated annotation pipeline (ARJ and RH). Briefly, each protein coding mt gene was identified by local alignment comparison (performed in all six reading frames) using amino acid sequences conceptually translated from corresponding genes from the mt genome of a reference species [e.g. *Necator americanus*; accession number: NC_003416; (33)]. The large and small ribosomal RNA genes (*rrn*S and *rrn*L) were identified by local alignment (i.e. using nucleotide sequence data) using the same approach. All transfer RNA (tRNA) genes were detected

and identified in a three-part process. Initially, all possible tRNA genes present in each consensus sequence were predicted (from both strands) based on a folding structure, using scalable models based on the standard nematode mt tRNAs (11). Employing this approach ~20 000–40 000 potential tRNA genes were predicted from each mt genome sequence. All predicted tRNA genes were then clustered into groups based on their anti-codon sequence and provisionally identified based on the amino acid encoded by this anti-codon. Two separate tRNA gene groups were predicted each for leucine (one each for the anticodons CUN and UUR, respectively) and for serine (one each for the anticodons AGN and UCN, respectively), as these tRNA genes have been shown to be duplicated in most invertebrate mt genomes, including those of nematodes (11). All predicted tRNAs within each amino acid group were ranked based on structural 'strength' (as inferred by the number of mismatched nt pairs in each stem), and the 100 best-scoring structures for each group were compared by BLASTn alignment against a database comprising all tRNA gene sequences for each amino acid of all published nematode mt genome sequences [available via http://drake.physics.mcmaster.ca/ogre/; (34)]. All tRNA genes of each mt genome were then identified and annotated based on having the highest sequence identity to known nematode tRNAs. Annotated sequence data were imported into the program SEQUIN (available via http://www.ncbi.nlm.nih.gov/Sequin/) for final verification of the mt genomic structure and subsequent, direct submission to GenBank.

Confirmatory PCR-based sequencing of short mt DNA tracts

Conventional PCR was used to amplify short mt DNA regions (~250–600 bp) representing ambiguously assembled sequence data (as determined by comparative alignment against all available mt genomic sequence data for the Nematoda). In brief, amplicons were produced from ~10–20 ng of total genomic DNA by PCR, conducted in 50 µl volumes using 25 pmol of each of two oligonucleotide primers, 250 µM of each dNTP, 3 mM $MgCl_2$ and 1 U *Taq* polymerase (Promega) in a 480 thermocycler (Perkin Elmer Cetus) under the following conditions: 94°C for 5 min (initial denaturation) followed by denaturation (94°C for 30 s), annealing (50°C for 30 s) and extension (65°C for 1 min) for 35 cycles, followed by a final extension (65°C for 5 min). The following, custom-designed primer sets used were: CYND1F (5′-CCAGGAGCCAGAGTGTTTCTTAT-3′) and CYND1R (5′-TTCCGCAAAATCAAAAGGTGCAC-3′) for *Cy. insignis*; ND1F (5′-CAGGGGAGTAAGTTGTAGTAAAG-3′) and TCND1R (5′-CTTTAGTTGGGCCTAAACGATTTTG-3′) for *Te. circumcincta*; ND1F and TAND1R (5′-CTTTTGTCGGACCTAAACGATTTTG-3′) for *Tr. axei*; ND1F and TVND1R (5′-GACACCTTAACTGGACCTAAAACG-3′) for *Tr. vitrinus*; MECX1F (5′-TTGGACATAGTTATCAAAGGGAAAT-3′) and MECX2R (5′-ACACCCCTAACAATAAACTACAATT-3′) for *Me. digitatus*, and CORNLF (5′-GCTTTGGAAGTAACTTTGTTAGACAA-3′) and CORNLR (5′-TCCTCACGCTAAGACTGCCATTT-3′) for *Ch. ovina*. A known-positive and a no-template controls were included in each PCR run. Amplicons were column-purified (Wizard PCR Preps, Promega) and subjected to conventional (Sanger) sequencing (BigDye Chemistry v.3.1, Applied Biosystems).

Alignment and phylogenetic analysis of concatenated nucleotide or amino acid sequence data

The phylogenetic analysis of amino acid sequence data was conducted using Bayesian inference (BI) (28) employing the software package MrBayes v.3.1.2 (http://mrbayes.csit.fsu.edu/index.php) and maximum likelihood (ML) using GARLI (36) (http://www.bio.utexas.edu/faculty/antisense/garli/Garli.html), each running on a four dual-core Opteron-based Unix cluster (http://pug.nhm.ac.uk). For individual species, the amino acid sequences inferred from all protein-coding mt genes were concatenated. A selection of published mt genomes from ascaridoids were used as outgroups (*Ascaris suum*, accession number: NC_003127; *Anisakis simplex*, NC_007934; *Toxocara canis*, EU730761), and alignments included additional (rhabditid) ingroup taxa (*Heterorhabditis bacteriophora*, accession number: NC_008534; *Caenorhabditis elegans*, NC_001328; *Ca. briggsae*, NC_009885). Amino acid sequences were aligned using MUSCLE (37). Ambiguously alignable positions were excluded prior to phylogenetic analysis.

RESULTS

Twelve circular mt genomes representing distinct species of strongylid (=bursate) nematodes of major socio-economic importance as parasites of animals (Table 1) were sequenced simultaneously by automated, massively parallel picolitre reactor sequencing (454 Life Sciences) and annotated using our custom-built bioinformatic pipeline (Figure 1). The 12 mt genomes sequenced herein varied in total length from ~13.7 (for *Tr. axei* and *Te. circumcincta*) to ~15.2 kb (for *Me. digitatus*) and were assembled from 6300 (*Cy. insignis*) to 15 100 (*O. dentatum*) individual sequence reads (Table 1). The mean read length per mt genome was 216 (*Tr. vitrinus*) to 270 nt (*Co. oncophora*), and the total sequence data contributing to the assembled mt genomes ranged from 1.6 (*Cy. insignis*) to 3.6 Mb (*Tr. vitrinus*). The DOC (i.e. the number of reads contributing to each nt position of the consensus sequence) was assessed for each species (representative result provided in Figure 2A); the mean DOC within each mt genome ranged from 117.8- (*Cy. insignis*) to 255.8-fold (*Co. oncophora*). A frequency histogram for a combined dataset representing the DOC at each nt position for each mt genome showed that 79.4% of all nt positions among all consensus sequences were supported by ≥50 individual sequence reads (Figure 2B). Sequence heterogeneity among the raw reads was assessed at each nt position of each consensus sequence (representative result provided in Figure 2C); the percentage of raw reads with the same sequence as the consensus (defined here as 'percent-read agreement') at each nt

Table 1. Consensus sequence length and read statistics for each nematode mt genome sequenced by 454 technology

Species	Total sequence length (nt)	Reads	Total sequence output (nt)	Read length (nt)	Mean length (SD)	DOC/nt	Mean DOC (SD)	Percentage of read Agr./nt	Mean percentage of read Agr. (SD)
Chabertia ovina	14147	14339	2197156	45–400	253 (66)	1–863	155 (138)	26.4–100.0	53.6 (11.4)
Cooperia oncophora	13919	13440	1639442	41–369	270 (73)	25–454	118 (79)	25.8–92.2	51.9 (10.6)
Cylicocyclus insignis	13674	6346	3498444	44–372	258 (84)	1–1457	256 (323)	26.2–100.0	52.6 (11.3)
Mecistocirrus digitatus	15221	12465	3220781	4–389	260 (68)	6–1197	206 (165)	26.7–92.9	54.1 (10.8)
Metastrongylus pudendotectus	13804	10023	2403274	43–414	267 (81)	87–1405	174 (101)	26.0–94.1	51.0 (9.9)
Metastrongylus salmi	13775	11771	2798458	46–417	253 (81)	54–1433	203 (171)	26.7–88.9	47.4 (9.1)
Oesophagostomum dentatum	13871	15054	2810287	36–374	244 (71)	25–2929	203 (307)	25.8–96.7	53.6 (11.0)
Strongylus vulgaris	14301	11049	2560992	40–372	236 (86)	6–1136	179 (171)	25.8–100.0	49.4 (9.8)
Syngamus trachea	14661	7498	2017991	31–409	247 (83)	1–991	138 (193)	25.1–100.0	55.4 (14.8)
Teladorsagia circumcincta	13731	11346	2603970	41–396	242 (90)	1–708	190 (120)	0.0[a]–96.4	50.6 (10.2)
Trichostrongylus axei	13731	8512	1883186	41–375	228 (92)	5–697	137 (72)	27.3–95.5	55.4 (11.5)
Trichostrongylus vitrinus	14036	14721	3588580	42–377	216 (89)	15–5914	256 (598)	25.9–100.0	53.8 (11.5)

Seq.: sequence; DOC: read depth of coverage at each nucleotide (nt) position of each consensus sequence; read Agr. = the percentage of reads at each nt position of each consensus sequence that displayed the same nt sequence as the consensus; SD: standard deviation.
[a]Representing a single nucleotide position wherein a single read had an 'N' in the sequence. This position was corrected upon resequencing.

position ranged from 25.1 to 100.0% [mean value of 47.4% (*M. salmi*) to 55.4% (*Sy. trachea* and *Tr. axei*)]. A frequency histogram of the read agreement (in %) for all mt genomes, as a combined dataset, showed that read-to-consensus sequence agreement was >40% at 85.7% of all nt positions (Figure 2D).

Following 454 sequencing and an assessment of sequence quality as well as DOC, each mt genome was annotated using a novel prototypical bioinformatic pipeline (Figure 1). To verify the accuracy of this automated annotation process, unannotated sequences representing the published mt genomes of *Ancylostoma duodenale* [accession number: NC_003415; ref. (33)], *Ascaris suum* [accession number: NC_003127; ref. (35)] and *N. americanus* were subjected to a blinded test. For each of the three mt genomes, the sequence annotated using our bioinformatic pipeline matched (with 100% accuracy) that published previously. Consistent with all species of Secernentea sequenced to date (11), the mt genomes of the 12 strongylid nematodes sequenced herein contained 12 protein coding genes [adenosine triphosphatase subunit 6 (*atp6*), cytochrome *b* subunit (*cyt*b), cytochrome *c* subunits 1–3 (*cox*1–3), and the nicotinamide dehydrogenase subunits 1–6 and 4l (*nad*1–6 and 4l) genes], two ribosomal subunits [large (*rrn*L) and small (*rrn*S)] and 22 transfer RNA (tRNA) genes (including two leucine and two serine tRNA genes). All genes were transcribed from the forward strand, and all protein coding genes had open reading frames (ORFs).

Following mt genome annotation, the error rate was estimated for each sequence (Table 2) by pairwise alignment against all published mt genome sequences available for nematodes (see 'Materials and Methods' section). Because most of the mt genomes determined herein represent species for which no complete mt genomic sequence was available previously, it was not possible to directly assess error rates for all regions of the mt genome. As a result, we adopted the cautious approach of estimating overall sequence error rates by comparing the coding regions of the genome (which are more conserved than the non-coding regions) to available sequence data from other species of nematodes. Using this approach, we were able to infer single indels that resulted in frame-shifts in a coding gene, resulting in an interruption of the ORF. Because the protein-coding mt genes are involved directly in cellular respiration, it was presumed that all 12 genes are essential for life and, therefore, an interrupted ORF for any of these genes would be deleterious to the organism and hence must represent a sequencing error. Based on these assumptions, comparative alignment revealed five (*Tr. axei* and *Tr. vitrinus*) to 28 (*M. salmi*) single nt (indel) sequencing errors in protein-coding regions of the mt genomes. Of these errors ($n = 163$), all but 31 were determined to be associated with homopolymeric sequence regions of five or more nucleotides in length. Furthermore, of the predicted indel errors associated with homopolymeric regions of the mt genome, all but three (~2.3%) were inferred to be the result of a single nucleotide deletion rather than an insertion. Based on these observations, we estimated a mean error rate of 6 indel errors per 100 homopolymers

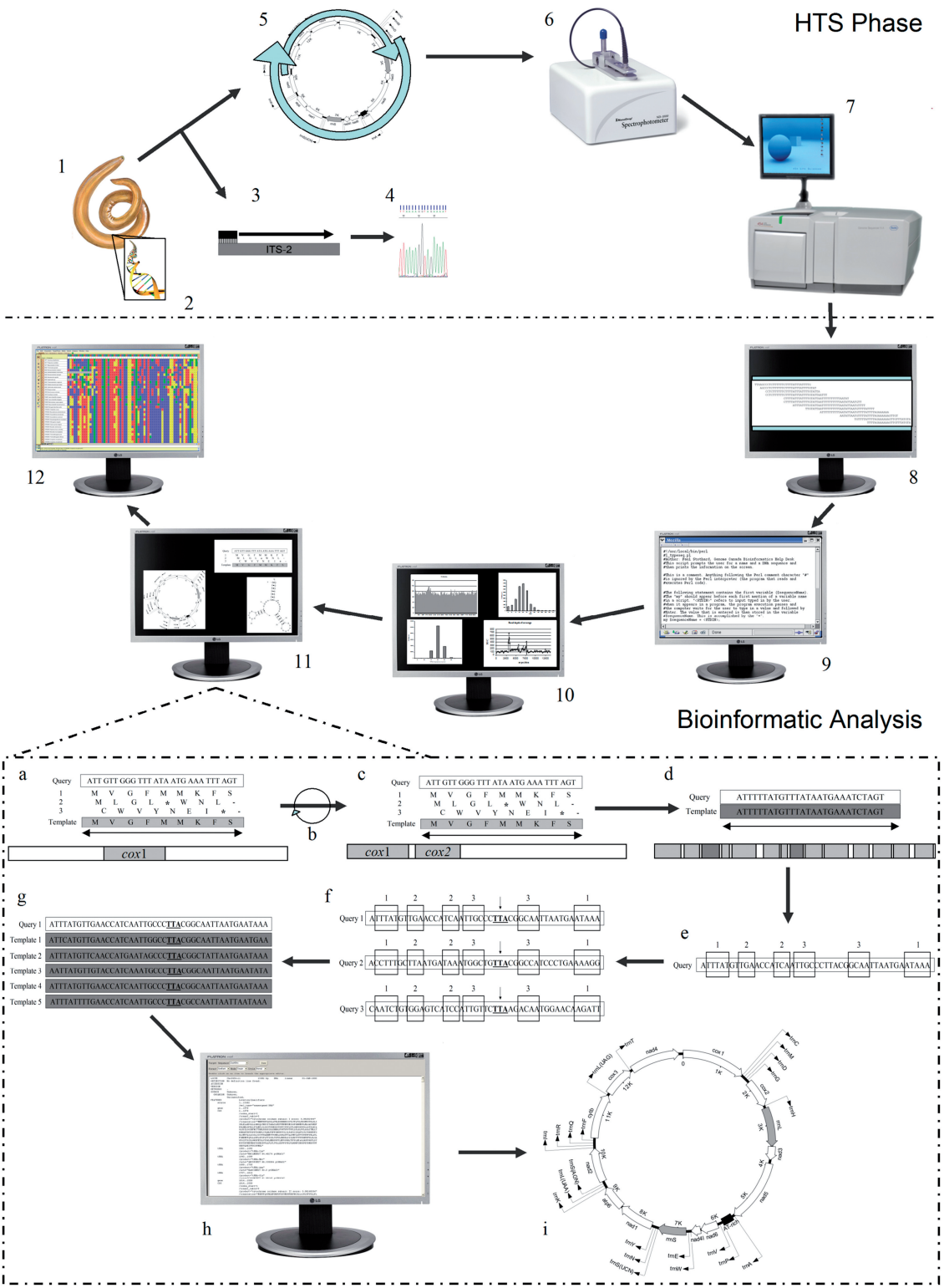

Figure 1. Flow-diagram of the 454 technology based high-throughput mt genomic sequencing and bioinformatic pipeline. The pipeline is divided into two main phases: the high-throughput sequencing (HTS) and annotation/analysis phases. Individual stages of the high-throughput pipeline are numbered 1–12 as follows: 1 = morphological identification of an individual nematode, 2 = total genomic extraction, 3 = PCR-amplification of the second internal transcribed spacer (ITS-2) of the nuclear ribosomal DNA, 4 = Direct sequencing of the ITS-2 (allowing specific identification of the nematode), 5 = long-PCR amplification of the complete mt genome as two overlapping fragments (see 23,24), 6 = quantification of each long-PCR amplicon by spectrophotometry (NanoDrop), 7 = simultaneous sequencing of the complete mt genome of each specimen by 454 technology (max. $n \geq 16$), 8 = read assembly (automated: Newbler) generating a majority rule consensus sequence for the mt genome of each specimen sequenced (see 26,28), 9 = bioinformatic analysis of raw read data (using Perl, Python and Java script), 10 = analysis of read length, as well as, assessment of read depth of coverage and read nucleotide (nt) diversity at each nt position of each consensus sequence generated, 11 = automated annotation of each mt

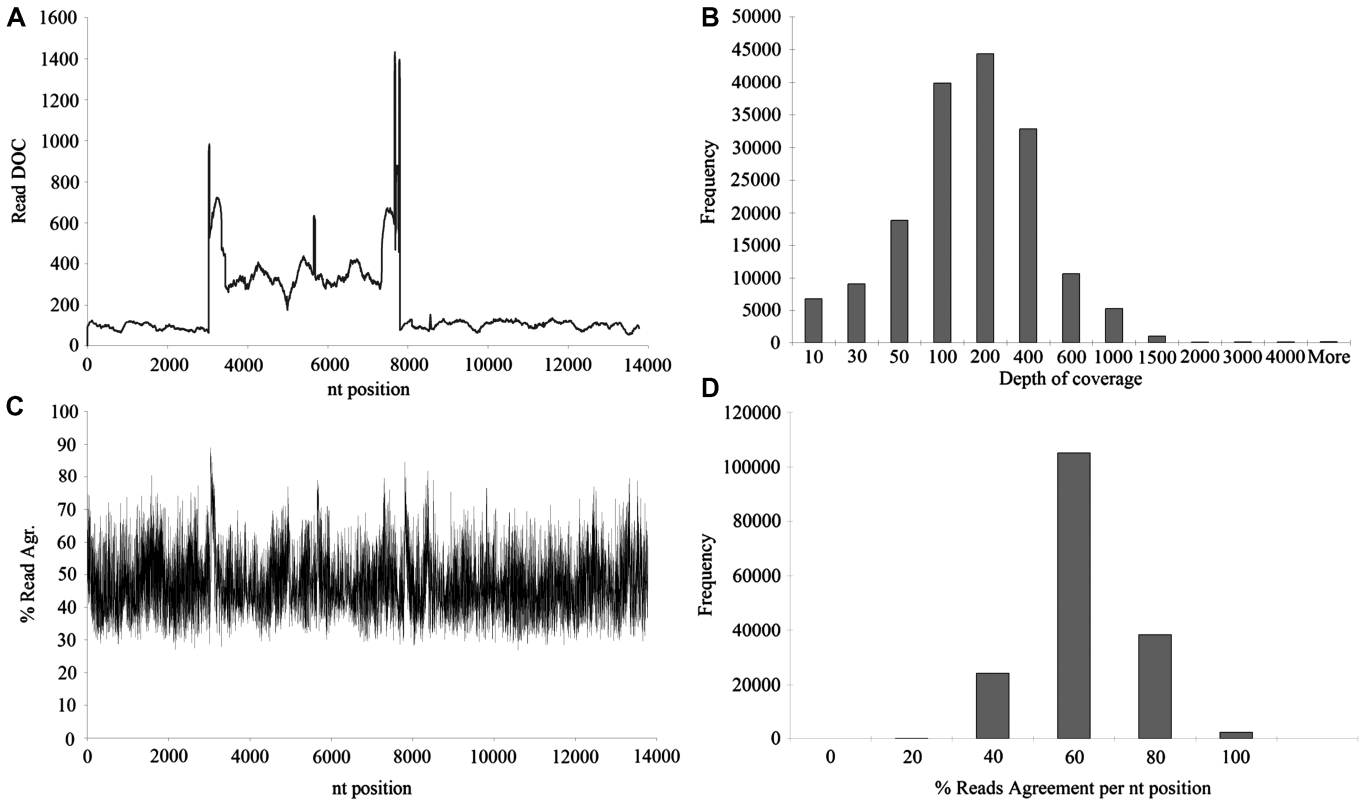

Figure 2. Summary of 454 read statistics displaying DOC and percent read agreement (Read Agr.) data for each position of each consensus sequence representing the complete mt genome of each of 12 species of Strongylida nematode sequenced from individual worms. (**A**) representative line graph of the range in depth of coverage across a complete mt genome (data displayed represents *Metastrongylus salmi*). (**B**) Frequency histogram displaying the range in depth of coverage at each nt position among all 12 sequencing runs. (**C**) Representative line graph displaying the range in % read agr. at each nt position across a complete mt genome (data displayed represents *Metastrongylus salmi*). (**D**) Frequency histogram displaying the percentage of read sequences at each nt position of the consensus sequence that were consistent with the consensus (i.e. had the nucleotide identity representing the majority at each nt position of the consensus sequence).

in the protein-coding regions of all 12 mt genomes sequenced. The number of homopolymers within the non-coding regions for each mt genome was counted and, using the error rate predicted for individual coding regions, the number of indel errors associated with these regions was estimated to range from three to eight per mt genome. Thus, the total estimated error rate for the consensus sequences representing all 12 mt genomes determined here ranged from 8 nt (for *Tr. vitrinus*) to 33 nt (for *M. salmi*), indicating an overall sequencing accuracy rate of 99.8–99.9% (excluding potential substitution errors). Among all 12 mt genomes, sequence ambiguity was detected in six short DNA tracts (~50–500 nt), requiring confirmatory (conventional, bi-directional) sequencing. These tracts were within *nad*1 of *Cy. insignis*, *Te. circumcincta*, *Tr. axei* and *Tr. vitrinus*, between *cox*1 and *cox*2 of *Me. digitatus* and within *rrn*L of *Ch. ovina*. Specifically, the four ambiguous regions that distrupted the ORF for *nad*1 for *Cy. insignis*, *Te. circumcincta*, *Tr. axei* and *Tr. vitrinus* were 82, 96, 72 and 228 nt, respectively. For *Me. digitatus*, the ambiguous region of 396 nt resulted in an apparent duplication of the

Figure 1. Continued
genome consensus sequence (see a–i below), 12 = estimation of sequence error rate, base-calling and selective re-sequencing of small regions by conventional means, and comparative analysis against published reference sequence data (from the GenBank database). All mt genomes were annotated using the novel bioinformatic pipeline developed for this study. The individual stages (boxed in diagram) of this annotation process (a–i) are: a = identification of the *cox*1 gene by comparative alignment using the translated amino acid sequence from a published ('template') sequence (user-defined), b = rotation of the 'query' consensus sequence relative to *cox*1, c = identification of all other coding mt genes by comparative alignment as per 'a', d = identification of the small and large ribosomal subunit genes (non-coding) by comparative alignment using nucleotide sequence data from the published template sequence, e = prediction of all possible tRNA genes based on inferred secondary structure, f = clustering of all predicted tRNA genes into amino acid groups based on their anticodon, g = comparative alignment of all predicted tRNA gene sequence data, employing a purpose built database containing all sequences of all published mt tRNA genes for nematodes (each predicted tRNA gene with the highest sequence identity score relative to the template database) and the highest bonding score (fewest mismatched pairs in the stem regions of the secondary structure) is selected for each amino acid), h = exportation of all annotated data for each mt genome sequence to a Sequin table for import into the program 'Sequin' (http://www.ncbi.nlm.nih.gov), allowing direct uploading of the annotated mt genome sequence to a public databases, such as GenBank (http://www.ncbi.nlm.nih.gov).

Table 2. Estimation of consensus sequence error rates based on observed frameshift deletion errors in the protein-coding mt genes for each nematode mt genome sequenced by 454 technology

Species	Homopoly (coding)	Homopoly (non-coding)	Indel errors (coding)	Error rate (per 100 homopoly)[a]	Predicted errors (non-coding)	Total errors	Percentage of error (per genome)	Percentage of accuracy (per genome)
Chabertia ovina	214	55	27	13	3	30	0.22	99.78
Cooperia oncophora	201	54	15	7	3	18	0.13	99.87
Cylicocyclus insignis	194	49	16	8	3	19	0.14	99.86
Mecistocirrus digitatus	207	120	9	4	8	17	0.11	99.89
Metastrongylus pudendotectus	291	79	17	6	5	22	0.16	99.84
Metastrongylus salmi	282	81	28	10	5	33	0.24	99.76
Oesophagostomum dentatum	192	52	14	7	3	17	0.12	99.88
Strongylus vulgaris	199	74	13	7	5	18	0.12	99.88
Syngamus trachea	210	75	7	3	5	12	0.08	99.92
Teladorsagia circumcincta	194	56	5	3	4	9	0.06	99.94
Trichostrongylus axei	221	82	7	3	5	12	0.09	99.91
Trichostrongylus vitrinus	194	50	5	3	3	8	0.06	99.94
Complete dataset	2599	827	163	6	52	215	0.13	99.87

Homopoly: homopolymeric regions of five or more nucleotides in length; Indel: insertion or deletion frameshift error. Predicted errors within the non-coding regions of each mt genome were estimated using the error rate calculated for the complete dataset. Total errors estimated for each mt genome are the sum of the predicted errors in the non-coding region and the observed errors in the coding region. All error estimates are based on observed insertion or deletion errors and do not include substitutions.
[a]Based on homopolymers detected in the coding regions of each mt genome only.

methionine and cysteine tRNA genes. For *Ch. ovina*, the ambiguous region of 470 nt appeared as a repeat within the *rrnL* gene. All of these regions were shown to be artefacts using the confirmatory sequencing approach.

A phylogeny for the Strongylida was reconstructed using Bayesian inference (BI) and maximum likelihood (ML) algorithms, employing mt data for 12 species sequenced in the present study as well as representative nematodes for which complete mtDNAs have been sequenced to date [Strongylida: *Anc. caninum* (33), *Anc. duodenale* (28), *B. phlebotomum* (28), *Haemonchus contortus* (26) and *N. americanus* (33); Rhabditida: *Caenorhabditis elegans* (35), *Ca. briggsae* (38) and *Heterorhabditis bacteriophora* (unpublished; GenBank NC_008534); Ascaridida: *Anisakis simplex* (39), *Ascaris suum* (35) and *Toxocara canis* (25) being included as the outgroups]. Trees for BI (nucleotides and amino acids) and ML (nucleotides) were the same in topology; Figure 3 shows the BI (amino acid) tree with nodal support from each of the analyses.

DISCUSSION

454 sequencing, consensus sequence generation, read coverage and raw read statistics

In the present study, we demonstrate the utility of a high-throughput platform for the sequencing and analysis of multiple mt genomes amplified from individual adult parasitic nematodes. In total, the 12 mt genomes sequenced ranged in size from ~13.7 to 15.2 kb. The abundance of sequence data produced by 454 technology provided high (mean ~118- to 256-fold) coverage of each genome, producing a consensus sequence with strong statistical support. The analysis of the variation in DOC across the genome (Figure 2A) indicated that there was a bi-modal spike for each mt genome and, generally, the region of the mt genome encompassed within these two peaks was more deeply covered than regions external to them. For all 12 species, these bi-modal peaks were ~5 kb apart and represented the boundaries of the small (~5 kb) and large (~10 kb) long-PCR amplicons synthesized from the total genomic DNA from each individual worm. The higher DOC found in the region between these peaks may be expected, considering that the mt genome was sequenced from a 'pool' of two amplicons of ~2.5 μg each. Given the size difference between the two amplicons, the number of copies of the small (5 kb) amplicon sequenced was approximately twice that of the large (10 kb) amplicon. The bimodal peaks (Figure 2A) are consistent with the regions of the mt genome in which there is an overlap by the ends of the large and small amplicons. The coverage of the regions represented by these peaks may be even deeper due to the premature termination of extension in long-PCR, which is predicted to be high, given the A + T richness of the template. Thus, these observations support the hypothesis that (at least for relatively short templates) 454 technology provides an unbiased sequence coverage (40) and that the DOC appears to be determined by the abundance of template in the sample being sequenced. This point may have relevance also for transcriptomic analyses. Future studies utilizing the approach proposed herein might benefit also from undertaking titrations of the concentration of amplicons based on their length to ensure a relatively uniform DOC. However, although the range in DOC across each mt genome was substantial in the present study, examination of the DOC data for all 12 mt genomes indicated that short regions (1–5 nt) of low coverage (1–5 reads) were rare and that the vast majority (~80%) of individual nt positions representing the consensus sequences were supported by ≥50 individual sequence reads. Where coverage was low, conservation was demonstrated at individual positions by comparative

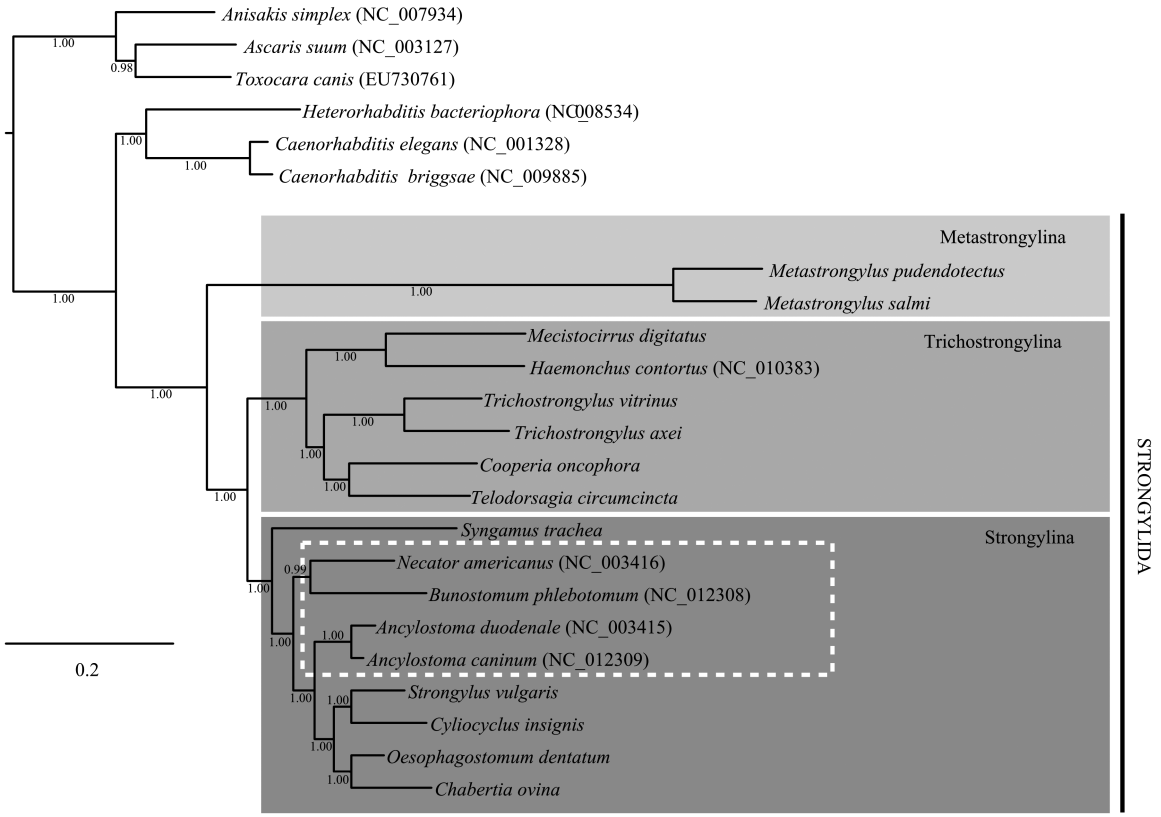

Figure 3. Phylogenetic analysis (Bayesian Inference) of the concatenated amino acid sequence data for all 12 protein coding mt genes for species representing the Strongylida, Rhabditida and Ascaridida (as an outgroup). Shaded boxes indicate the three major suborders of the Strongylida (Metastrongylina, Trichostrongylina and Strongylina). A dash-bordered box highlights paraphyletic clustering of the hookworms. The numbers above the midpoint of each tree branch represent the statistical support for each node (based on posterior probability score). The phylogram provided is presented to scale (scale bar = 0.2 substitutions per site). GenBank accession numbers are provided (in parentheses) for all reference sequences. An identical topology was found with maximum likelihood; all nodes supported by >99% bootstrap re-sampling ($n = 100$).

sequence alignments. Thus, it is highly unlikely that the variation in DOC had an effect on the quality of the consensus sequences determined herein.

A principal strategy of 454 sequencing is the generation of consensus data by a 'majority rule' analysis of the raw read output for each sample (27). However, an in-depth analysis of the raw read data generated herein indicates that the proportion of reads constituting a 'majority' at each nt position can range substantially across the length of the consensus sequence (from 25.1 to 100%). This variation may represent errors in sequencing or amplification, or genuine sequence polymorphism. We contend that the variation measured herein largely represents the latter. We base this contention on three observations. First, the error rate of the enzyme used for the long-PCR amplification is low (~25–30 bp per 100 kb of amplicon synthesized—manufacturer's estimate). Secondly, comparative analyses of the protein-coding regions within the mt genome sequences generated herein estimated an error rate for the consensus sequence of ~0.1–0.2%, which is consistent with that reported previously for 454 based sequencing (27,41). If the majority of the sequence variation detected in the raw read-output is linked to sequence error, the mean error rate for the raw reads would be ~44.6 to ~52.6% (i.e. 100% minus the mean read agreement scores for each mt genome). From a probabilistic viewpoint, it is considered highly unlikely that the raw reads have such high error rates as, collectively, they produced a quality consensus sequence (as indicated by comparative analyses and ORFs). Thus, nucleotide polymorphism rather than error is proposed to be the likely explanation for this variability among individual reads making up the consensus. Although this hypothesis is supported by high mutation rates in the mt genomes of nematodes (42–44), it would require experimental testing (e.g. through the isolation of different mt subpopulations and subsequent sequencing).

Error rate for 454 sequencing from mt genomic template

Most of the errors ($n = 131$ among all 12 mt genomes) detected in the protein-coding regions of the consensus sequences generated in the present study were associated with homopolymeric regions of >5 nt. Interestingly, all but three of these homopolymeric sequencing errors appeared to be the result of a single nucleotide deletion rather than an insertion (i.e. our data suggest that 454 technology tends to underestimate the length of homopolymers by one nucleotide, if there is an error). In the present study, where possible, ambiguous positions

were called, based on comparative alignment (with reference sequences), resulting in a ~80% reduction of identifiable sequencing errors in each sequence. Errors occurring in the non-coding regions of the mt genome cannot be readily identified for species for which no sequence data are available. Because most of the errors that we inferred in the coding regions of the mt genomes were associated with homopolymers, we could estimate that the error rate in the non-coding regions would be approximately proportional to the number of homopolymers that they contain. Based on this assumption, the estimated error rate for each mt genome was ~0.1–0.2%. Because it was not possible to detect substitution errors in the present dataset, it is possible that this figure represents an underestimation. However, considering the high DOC levels achieved in the present analysis and the 'majority-rule' nature of the construction of the consensus sequence for each mt genome, it is unlikely that substitution errors have contributed significantly to the consensus sequences determined herein.

The sequence ambiguity detected within a small number of mt genomic tracts of ~50–500 nt (in *nad*1, between *cox*1 and *cox*2, and within *rrn*L for some species; in some cases, repetitive elements) was interpreted to be associated with the automated *in silico*-assembly of the raw sequence data rather than sequence heterogeneity within individual nematodes, because 'clean' sequence was determined using confirmatory, conventional sequencing of short amplicons (150–600 bp) spanning these tracts. Nonetheless, artefacts need to be considered when sequencing using any platform, and confirmatory sequencing (e.g. using the Sanger method) is recommended in cases of sequence ambiguity (e.g. apparent gene duplications or the presence of large insert/repeat regions) to ensure an optimum, final sequence. Additional studies, in which multiple complete mt genome sequence replicates are generated from the same individual, should be useful in further assessing the accuracy of 454 sequencing.

The mt genomes of the Strongylida

Although the high-throughput pipeline reported here is considered to represent a significant scientific and technological advance in itself, the sequence data from this study also has major implications in relation to strongylid nematodes and nematodes in general. The Strongylida has undergone extensive species radiation; species of this order have successfully adapted to a range of ecological niches as parasites of (mainly terrestrial) vertebrates (45). Many of these species represent some of the most damaging, widespread and prevalent metazoan parasites of humans, wildlife and domestic animals worldwide. Amongst the most significant strongylids are *Anc. duodenale* and *N. americanus*, which cause hookworm disease, affecting hundreds of millions of people globally (46). In addition, species of this order [e.g. *Haemonchus contortus* (a blood-feeding parasite of sheep and other small ruminants), *Trichostrongylus* spp. (gastrointestinal parasites of ruminants) and *Ch. ovina* (an large intestinal parasite of sheep)] are pathogens of livestock animals, causing substantial economic losses to the agricultural and livestock industries globally. Prior to the present study, mt genomes were available for only five strongylids, namely *Anc. caninum* (canine hookworm) (28), *Anc. duodenale* (33), *B. phlebotomum* (bovine hookworm) (28), *H. contortus* (26) and *N. americanus* (33). Thus, the present study provides a significant enhancement in available mt genomic data for this important group of parasites.

Examination of the 12 mt genomes sequenced and the five mt genomes available previously for members of this order of nematodes (26,28,33) suggests that the size and structure of the mt genomes of species of Strongylida is largely conserved. As with all species of Strongylida sequenced to date, those studied herein displayed the gene arrangement 'GA2', reported previously by Hu *et al.* (10,47). The compact size of the mt genomes described is consistent with all other members of this order (26,28,33), with all 12 mt genomes displaying minimal intergenic spaces, primarily confined to one large, highly AT-rich region for each genome (previously hypothesized to represent the origin of replication; ref. (10). Examination of the nucleotide variability occurring within and among the protein-coding regions of each mt genome (data not shown) was consistent with previous studies of the mt genomes of strongylid species, with the most conserved genes being *cox*1, *cox*2 and *cox*3, and the most variable being *nad*2, *nad*4 and *nad*6. These data support the use of the 'conserved' mt genes, such as *cox*1, for exploring the systematics and potentially for population genetics and/or cryptic speciation events in nematodes (10–12,17,18). However, this study further supports the hypothesis (11,25,26,28) that the more variable mt genes (e.g. *nad*2 and *nad*6) are worthy of consideration as population genetic markers for epidemiological studies. In addition to being able to rapidly determine the mt genomes of a wide range of parasites, the genomic-bioinformatic approach established herein provides a high-throughput platform to generate complete mt genomes as barcodes for individual organisms or populations of organisms. Although the focus here was on parasitic nematodes, the platform developed has far-reaching implications for mt genomics of other groups of organisms, and for the automated sequencing and annotation of other small to medium-sized templates, including other organellar genomes (e.g. plastids and apicoplasts), large nuclear gene operons and virus genomes.

Phylogenetic relationships of the Strongylida

In the present study, a concatenated amino acid sequence dataset for all 17 available mt genomic sequences for species of Strongylida (including the 12 determined herein) was subjected to phylogenetic analyses (using the BI and ML methods). The trees constructed were largely consistent in topology with those proposed by Chilton *et al.* (48) using nuclear ribosomal gene data and supported the finding that the Strongylida represents a monophyletic group. In the present study, the major suborders within the Strongylida (e.g. the Metastrongylina,

Strongylina and Trichostrongylina) were each resolved as distinct, monophyletic clades with maximum statistical and nodal support (posterior probability = 1.00; bootstrap = 100). This is in contrast to the phylogeny constructed by Chilton *et al*. (48), in which no monophyletic clades were formed for species classified within the Strongylina or the Trichostrongylina (although species from each suborder generally clustered together within the tree). This difference could relate to the greater lengths of complete mt sequences and higher numbers of informative characters compared with the nuclear ribosomal sequence data used by Chilton *et al*. (48) or, alternatively, might reflect the use of different phylogenetic algorithms (Bayesian inference and maximum likelihood) in the present study. Importantly, the present analyses supported the distinct placement of the lungworms (represented herein by *M. pudendotectus* and *M. salmi*) relative to all other (suborders of the) Strongylida, consistent with the hypothesis (48) that the most basal evolutionary dichotomy within the Strongylida divides the order according to predilection site in the host [i.e. those species which infect the pulmonary tissues (i.e. lungworms) and those which infect different parts of the gastrointestinal tract]. However, presently there are no complete mt genome sequences available for species of *Dictyocaulus*, which infect the lungs at the adult stage. Traditionally, species of *Dictyocaulus* were placed within the Trichostrongylina (49), but were found to group with the Metastrongylina using nuclear ribosomal gene data (48). An analysis of complete mt genomic datasets, including members of the Dictyocaulidae, is needed. The present analyses suggest also that the hookworms do not represent a monophyletic clade. Consistent with a previous study employing nuclear ribosomal DNA data (48), *N. americanus* and *B. phlebotomum* (Bunostomatidae) group relatively closely with *Sy. trachea*, whereas the two *Ancylostoma* species (*Anc. caninum* and *Anc. duodenale*) appear to represent an evolutionarily distinct lineage. This finding highlights the need for an expanded study of the superfamily Ancylostomatoidea, a superfamily which contains numerous major pathogens of human, veterinary and agricultural importance (45).

Overall, the phylogenetic hypotheses formulated based on mt genomic data for the Strongylida did not differ significantly from that proposed for this order using nuclear ribosomal DNA data (other than providing increased resolution of some clades, which is to be expected given the larger size of the mt genomic dataset). This is an important finding. Previous phylogenetic analyses of the Nematoda using whole mt genomic datasets (11,25,26,47,50) were distinct from the current phylogenetic classification of this phylum based on data for the small subunit of the nuclear ribosomal RNA gene (51). However, it was not possible to determine whether these differences simply related to differences in mt *versus* nuclear genomic evolution. Based on the present study, the comparative appraisal of the phylogenetic relationships of the Strongylida proposed using mt genomic data or nuclear ribosomal DNA data (48) does not provide a substantive indication that the mt genomes and nuclear genomes (at least as indicated by the ribosomal gene) of species within this order have evolved along different paths. Therefore, our current findings provide additional support for the reappraisal of the phylogeny of the Nematoda using whole mt genomic datasets.

Concluding remarks

In the present study, we developed and evaluated an integrated approach for the sequencing and automated annotation of mt genomes using a practical, custom-built bioinformatic pipeline. Although whole mt genomes are relatively small compared with nuclear genomes, we have demonstrated that the high-throughput pipeline established represents an efficient method with broad applicability to any group of organisms. Importantly, our analyses show that the DOC, provided by the substantial data output of next-generation sequencing platforms (e.g. the 454 FLX genome sequencer), results in robust genome coverage, giving strong statistical support for each consensus sequence. The prototypical annotation pipeline, built for the present study, provides a rapid and efficient method for data analysis, to keep pace with the increased sequence output provided by new sequencing technologies. Together, these advances provide a high-throughput and efficient means for the sequencing and annotation of organellar (e.g. plastid and apicoplast) genomes and other 'medium-sized' genomes (e.g. of viruses) or large nuclear genomic operons from any source. The efficiency and throughput of this process has the potential to be further enhanced to great effect using indexing technologies (52,53), allowing the multiplexing of tens or hundreds of mt genome templates in a single high-throughput sequencing reaction. In addition, when coupled with whole genome amplification (WGA) systems (54) and/or laser micro-dissection technology (55), this approach might be applicable to assessing sequence variation or single nucleotide polymorphisms (SNPs) in mt genes among tissue types, providing insights into the heterogeneity within an individual, with possible implications for understanding mutation rates and inheritance.

For the Strongylida, the present study elucidates the evolutionary relationships of key parasites of major socio-economic importance. The phylogenetic analyses conducted herein re-enforce previous findings, indicating a major, early divergence of the Strongylida according to predilection site in the host, with a major lineage (the 'lungworms') infecting the lungs and the other infecting (primarily) parts of the gastrointestinal tract. We provide strong molecular support for the monophyly of the Strongylina and Trichostrongylina (not achieved previously with nuclear ribosomal gene data). In addition, we reinforce the finding that the group commonly referred to as 'the hookworms' (e.g. *Anc. caninum*, *Anc. duodenale*, *B. phlebotomum* and *N. americanus*), which represent a group of blood-feeding (haematophagous) pathogens of major importance in humans and other animals, might represent a paraphyletic

group of species with disparate evolutionary lineages. Taken together, these findings have important implications for our understanding of the evolution of a major radiation of species within the Nematoda and a group that arguably includes some of the most significant metazoan pathogens, in terms of the their impact on human and animal health worldwide. Furthermore, our findings provide support for the utility of whole mt genomic data for investigations into the phylogenetic history of the Nematoda, and, when coupled with the high-throughput sequencing and analytical pipeline proposed herein, paves the way for expanded investigations of this major phylum of invertebrates, with implications for systematic and population genetic studies of a broad range of other organisms and a range of small- to medium-sized genomes.

ACKNOWLEDGEMENTS

The authors thank Ian Beveridge, Chris Morrow, Toivo Jaervis, Jan VanWyk, Duncan Bucknell and the late Peter Nansen for the provision of some specimens used in the present study and to Min Hu for the preparation of some amplicons.

FUNDING

Australian Research Council (grant numbers LX0882215 and LX0775848).

Conflict of interest statement. None declared.

REFERENCES

1. Crimi,M. and Rigolio,R. (2008) The mitochondrial genome, a growing interest inside an organelle. *Int. J. Nanomed.*, **3**, 51–57.
2. Vanitallie,T.B. (2008) Parkinson disease: primacy of age as a risk factor for mitochondrial dysfunction. *Metabolism*, **57(Suppl. 2)**, S50–S55.
3. Gandhi,S. and Wood,N.W. (2005) Molecular pathogenesis of Parkinson's disease. *Hum. Mol. Genet.*, **14(Spec No. 2)**, 2749–2755.
4. Geurts,J.J. and Barkhof,F. (2008) Grey matter pathology in multiple sclerosis. *Lancet Neurol.*, **7**, 841–851.
5. Reeve,A.K., Krishnan,K.J. and Turnbull,D.M. (2008) Age related mitochondrial degenerative disorders in humans. *Biotechnol. J.*, **3**, 750–756.
6. Debray,F.G., Lambert,M. and Mitchell,G.A. (2008) Disorders of mitochondrial function. *Curr. Opin. Pediatr.*, **20**, 471–482.
7. Forbes,J.M., Coughlan,M.T. and Cooper,M.E. (2008) Oxidative stress as a major culprit in kidney disease in diabetes. *Diabetes*, **57**, 1446–1454.
8. Tarr,P.E. and Telenti,A. (2007) Toxicogenetics of antiretroviral therapy: genetic factors that contribute to metabolic complications. *Antivir. Ther.*, **12**, 999–1013.
9. Fukui,H. and Moraes,C.T. (2008) The mitochondrial impairment, oxidative stress and neurodegeneration connection: reality or just an attractive hypothesis? *Trends Neurosci.*, **31**, 251–256.
10. Hu,M., Chilton,N.B. and Gasser,R.B. (2004) The mitochondrial genomics of parasitic nematodes of socio-economic importance: recent progress, and implications for population genetics and systematics. *Adv. Parasitol.*, **56**, 133–212.
11. Hu,M. and Gasser,R.B. (2006) Mitochondrial genomes of parasitic nematodes – progress and perspectives. *Trends Parasitol.*, **22**, 78–84.
12. Le,T.H., Blair,D. and McManus,D.P. (2002) Mitochondrial genomes of parasitic flatworms. *Trends Parasitol.*, **18**, 206–213.
13. Whitehead,A. (2009) Comparative mitochondrial genomics within and among species of killifish. *BMC Evol. Biol.*, **13**, 11.
14. Piganeau,G. and Eyre-Walker,A. (2009) Evidence for variation in the effective population size of animal mitochondrial DNA. *PLoS ONE*, **4**, e4396.
15. Dowton,M., Cameron,S.L., Dowavic,J.I., Austin,A.D. and Whiting,M.F. (2009) Characterization of 67 mitochondrial tRNA gene rearrangements in the Hymenoptera suggests that mitochondrial tRNA gene position is selectively neutral. *Mol. Biol. Evol.*, **26**, 1607–1617.
16. Mueller,R.L. and Boore,J.L. (2005) Molecular mechanisms for extensive mitochondrial gene rearrangement in plethodontid salamanders. *Mol. Biol. Evol.*, **22**, 2104–2112.
17. Zarowiecki,M.Z., Huyse,T. and Littlewood,D.T. (2007) Making the most of mitochondrial genomes – markers for phylogeny, molecular ecology and barcodes in Schistosoma (Platyhelminthes: Digenea). *Int. J. Parasitol.*, **37**, 1401–1418.
18. Littlewood,D.T. (2008) Platyhelminth systematics and the emergence of new characters. *Parasite*, **15**, 333–341.
19. Oliveira,D.C., Raychoudhury,R., Lavrov,D.V. and Werren,J.H. (2008) Rapidly evolving mitochondrial genome and directional selection in mitochondrial genes in the parasitic wasp *Nasonia* (Hymenoptera: Pteromalidae). *Mol. Biol. Evol.*, **25**, 2167–2180.
20. Jongwutiwes,S., Putaporntip,C., Iwasaki,T., Ferreira,M.U., Kanbara,H. and Hughes,A.L. (2005) Mitochondrial genome sequences support ancient population expansion in *Plasmodium vivax*. *Mol. Biol. Evol.*, **22**, 1733–1739.
21. Burger,G., Lavrov,D.V., Forget,L. and Lang,B.F. (2007) Sequencing complete mitochondrial and plastid genomes. *Nat. Protocols*, **2**, 603–614.
22. Lang,B.F. and Burger,G. (2007) Purification of mitochondrial and plastid DNA. *Nat. Protocols*, **2**, 652–660.
23. Hu,M., Chilton,N.B. and Gasser,R.B. (2002) Long PCR-based amplification of the entire mitochondrial genome from single parasitic nematodes. *Mol. Cell Probes*, **16**, 261–267.
24. Hu,M., Jex,A.R., Campbell,B.E. and Gasser,R.B. (2007) Long PCR amplification of the entire mitochondrial genome from individual helminths for direct sequencing. *Nat. Protocols*, **2**, 2339–2344.
25. Jex,A.R., Waeschenbach,A., Littlewood,D.T., Hu,M. and Gasser,R.B. (2008) The Mitochondrial Genome of *Toxocara canis*. *PLoS Negl. Trop. Dis.*, **2**, e273.
26. Jex,A.R., Hu,M., Littlewood,D.T., Waeschenbach,A. and Gasser,R.B. (2008) Using 454 technology for long-PCR based sequencing of the complete mitochondrial genome from single *Haemonchus contortus* (Nematoda). *BMC Genomics*, **9**, 11.
27. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
28. Jex,A.R., Waeschenbach,A., Hu,M., van Wyk,J.A., Beveridge,I., Littlewood,D.T.J. and Gasser,R.B. (2009) The mitochondrial genomes of *Ancylostoma caninum* and *Bunostomum phlebotomum* – two hookworms of animal health and zoonotic importance. *BMC Genomics*, **10**, 79.
29. Ermini,L., Olivieri,C., Rizzi,E., Corti,G., Bonnal,R., Soares,P., Luciani,S., Marota,I., De Bellis,G., Richards,M.B. *et al.* (2008) Complete mitochondrial genome sequence of the Tyrolean Iceman. *Curr. Biol.*, **18**, 1687–1693.
30. Cui,Z., Liu,Y., Li,C.P., You,F. and Chu,K.H. (2009) The complete mitochondrial genome of the large yellow croaker, *Larimichthys crocea* (Perciformes, Sciaenidae): unusual features of its control region and the phylogenetic position of the Sciaenidae. *Gene*, **432**, 33–43.
31. Gasser,R.B., Chilton,N.B., Hoste,H. and Beveridge,I. (1993) Rapid sequencing of rDNA from single worms and eggs of parasitic helminths. *Nucleic Acids Res.*, **21**, 2525–2526.
32. Schindler,A.R., de Gruijter,J.M., Polderman,A.M. and Gasser,R.B. (2005) Definition of genetic markers in nuclear ribosomal DNA for a neglected parasite of primates, *Ternidens deminutus* (Nematoda: Strongylida) – diagnostic and epidemiological implications. *Parasitology*, **131**, 539–546.
33. Hu,M., Chilton,N.B. and Gasser,R.B. (2002) The mitochondrial genomes of the human hookworms, *Ancylostoma duodenale* and

Necator americanus (Nematoda: Secernentea). *Int. J. Parasitol.*, **32**, 145–158.
34. Jameson,D., Gibson,A.P., Hudelot,C. and Higgs,P.G. (2003) OGRe: a relational database for comparative analysis of mitochondrial genomes. *Nucleic Acids Res.*, **31**, 202–206.
35. Okimoto,R., Macfarlane,J.L., Clary,D.O. and Wolstenholme,D.R. (1992) The mitochondrial genomes of two nematodes, *Caenorhabditis elegans* and *Ascaris suum*. *Genetics*, **130**, 471–498.
36. Zwickl,D.J. (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. *PhD Thesis*. University of Texas, Austin.
37. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
38. Stein,L.D., Bao,Z., Blasiar,D., Blumenthal,T., Brent,M.R., Chen,N., Chinwalla,A., Clarke,L., Clee,C., Coghlan,A. *et al.* (2003) The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol.*, **1**, E45.
39. Kim,K.H., Eom,K.S. and Park,J.K. (2006) The complete mitochondrial genome of *Anisakis simplex* (Ascaridida: Nematoda) and phylogenetic implications. *Int. J. Parasitol.*, **36**, 319–328.
40. Blow,M.J., Zhang,T., Woyke,T., Speller,C.F., Krivoshapkin,A., Yang,D.Y., Derevianko,A. and Rubin,E.M. (2008) Identification of ancient remains through genomic sequencing. *Genome Res.*, **18**, 1347–1353.
41. Moore,M.J., Dhingra,A., Soltis,P.S., Shaw,R., Farmerie,W.G., Folta,K.M. and Soltis,D.E. (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biol.*, **6**, 17.
42. Anderson,T.C., Blouin,M.S. and Beech,R.N. (1998) Population biology of parasitic nematodes: applications of genetic markers. *Adv. Parasitol.*, **41**, 219–283.
43. Avise,J.C., Arnold,J., Ball,R.M., Berminham,E., Lamb,T., Neigel,J.E., Carol,A.R. and Saunder,N.C. (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Ann. Rev. Ecol. Syst.*, **18**, 489–522.
44. Blouin,M.S. (1998) Mitochondrial DNA diversity in nematodes. *J. Helminthol.*, **72**, 285–289.
45. Skryabin,K.I., Shikhobalova,N.P., Schulz,R.S., Popova,T.I., Boev,S.N. and Delyamure,S.L. (1992) *Key to Parasitic Nematoda: Volume 3 Strongylata*. E. J. Brill, Leiden, The Netherlands.
46. Hotez,P.J., Bethony,J., Bottazzi,M.E., Brooker,S. and Buss,P. (2005) Hookworm: 'the great infection of mankind'. *PLoS Med.*, **2**, e67.
47. Hu,M., Chilton,N.B. and Gasser,R.B. (2003) The mitochondrial genome of *Strongyloides stercoralis* (Nematoda) – idiosyncratic gene order and evolutionary implications. *Int. J. Parasitol.*, **33**, 1393–1408.
48. Chilton,N.B., Huby-Chilton,F., Gasser,R.B. and Beveridge,I. (2006) The evolutionary origins of nematodes within the order Strongylida are related to predilection sites within hosts. *Mol. Phylogenet. Evol.*, **40**, 893–899.
49. Durette-Desset,M.-C. (1985) Trichostrongyloid nematodes and their vertebrate hosts: Reconstruction of the phylogeny of a parasitic group. *Adv. Parasitol.*, **24**, 239–306.
50. Kang,S., Sultana,T., Eom,K.S., Park,Y.C., Soonthornpong,N., Nadler,S.A. and Park,J. (2008) The mitochondrial genome sequence of *Enterobius vermicularis* (Nematoda: Oxyurida)—an idiosyncratic gene order and phylogenetic information for chromadorean nematodes. *Gene*, **429**, 87–97.
51. Blaxter,M.L., De Ley,P., Garey,J.R., Liu,L.X., Scheldeman,P., Vierstraete,A., Vanfleteren,J.R., Mackey,L.Y., Dorris,M., Frisse,L.M. *et al.* (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature*, **392**, 71–75.
52. Meyer,M., Stenzel,U., Myles,S., Prufer,K. and Hofreiter,M. (2007) Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res.*, **35**, e97.
53. Parameswaran,P., Jalili,R., Tao,L., Shokralla,S., Gharizadeh,B., Ronaghi,M. and Fire,A.Z. (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res.*, **35**, e130.
54. Pinard,R., de Winter,A., Sarkis,G.J., Gerstein,M.B., Tartaro,K.R., Plant,R.N., Egholm,M., Rothberg,J.M. and Leamon,J.H. (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics*, **7**, 216.
55. Ranjit,N., Jones,M.K., Stenzel,D.J., Gasser,R.B. and Loukas,A. (2006) A survey of the intestinal transcriptomes of the hookworms, *Necator americanus* and *Ancylostoma caninum*, using tissues isolated by laser microdissection microscopy. *Int. J. Parasitol.*, **36**, 701–710.