Queensland University of Technology
Brisbane Australia

This may be the author's version of a work that was submitted/accepted for publication in the following source:

Cameron, Stephen
(2014)
How to sequence and annotate insect mitochondrial genomes for systematic and comparative genomics research.
Systematic Entomology, *39*(3), pp. 400-411.

This file was downloaded from: https://eprints.qut.edu.au/73195/

© Consult author(s) regarding copyright matters

This work is covered by copyright. Unless the document is being made available under a Creative Commons Licence, you must assume that re-use is limited to personal use and that permission from the copyright owner must be obtained for all other uses. If the document is available under a Creative Commons License (or other specified license) then refer to the Licence for details of permitted re-use. It is a condition of access that users recognise and abide by the legal requirements associated with these rights. If you believe that this work infringes copyright please provide details by email to qut.copyright@qut.edu.au

Notice: *Please note that this document may not be the Version of Record (i.e. published version) of the work. Author manuscript versions (as Submitted for peer review or as Accepted for publication after peer review) can be identified by an absence of publisher branding and/or typeset appearance. If there is any doubt, please refer to the published source.*

https://doi.org/10.1111/syen.12071

How to sequence and annotate insect mitochondrial genomes for systematic and comparative genomics research

Submitted to *Systematic Entomology* as a review.

Stephen L. Cameron

Earth, Environmental & Biological Sciences School, Science & Engineering Faculty, Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, AUSTRALIA.

Phone: (+617) 3138 2869; FAX: (+617) 3138 2330; Email: sl.cameron@qut.edu.au

Running Title: Sequencing insect mt genomes

Introduction

Over the past decade the mitochondrial (mt) genome has become the most widely used genomic resource available for systematic entomology. While the availability of other types of '–omics' data, in particular transcriptomes, are increasing rapidly, mt genomes are still vastly cheaper to sequence and are far less demanding of high quality templates. Furthermore, almost all other '–omics' approaches also sequence the mt genome and so it can form a bridge between legacy and contemporary datasets. Mitochondrial genomes have now been sequenced for all insect orders, and in many instances representatives of each major lineage within orders (suborders, series or superfamilies depending on the group). They have also been applied to systematic questions at all taxonomic scales from resolving interordinal relationships (e.g. Cameron *et al*., 2009; Wan *et al*., 2012; Wang *et al*., 2012), through many intraordinal (e.g. Dowton *et al.*, 2009; Timmermans *et al*., 2010; Zhao *et al*. 2013a) and family level studies (e.g. Nelson *et al*., 2012; Zhao *et al*., 2013b) to population / biogeographic studies (e.g. Ma *et al*., 2012). Methodological issues around the use of mt genomes in insect phylogenetic analyses and the empirical results found to date have recently been reviewed by Cameron (2014), however, the technical aspects of sequencing and annotating mt genomes were not covered. Most papers which generate new mt genome report their methods in a simplified form which can be difficult to replicate without specific knowledge of the field. Published studies utilize a wide enough range of approaches, usually without justification for the one chosen, that confusion about commonly used jargon such as 'long-PCR' and 'primer walking' could be a serious barrier to entry. Furthermore, sequenced mt genomes have been annotated (gene locations defined) to wildly varying standards and improving data quality through consistent annotation procedures will benefit all downstream users of these datasets.

The aims of this review are therefore to:

1. Describe in detail the various sequencing methods used on insect mt genomes;
2. Explore the strengths/weakness of different approaches;
3. Outline the procedures and software used for insect mt genome annotation; and
4. Highlight quality control steps used for new annotations, and to improve the re-annotation of previously sequenced mt genomes used in systematic or comparative research.

Mitochondria Basics

The mt genome of most animals is an extremely conserved and constrained molecule. It is descended from the genome of the alpha-proteobacterial symbiont that became the mitochondrion in the ancestor of all eukaryotes, and retains many bacterial-type features. Like most bacterial genomes it is usually a circular molecule, the only exceptions being non-insects such as cnidarians (Burger et al. 2003). It has undergone massive reductive evolution with many genes either moved to the nuclear genome or their

function replaced by nuclear encoded orthologs. The gene-set of bilaterian animals (i.e. all metazoans excluding cnidarians, ctenophores, poriferans and placozoans) is fixed at just 37 genes: 13 protein-coding genes (PCGs) which form part of the electron transport chain, plus 2 ribosomal RNA (rRNAs) and 22 transfer RNA (rRNA) genes which are responsible for translating the mt PCGs (Osigus *et al*., 2013). Very few bilaterian animals have less than 37 genes, and the few which have more have duplicate copies of one or more of these core 37 genes. In addition to its genic content, the mt genome also includes one or more non-coding regions that function as binding sites for proteins involved in genome replication such as the control-region (CR) and transcription. In most animals mt genes are transcribed on both strands; the stand with the most genes is termed the 'majority' strand and the other the 'minority' stand. Other terms used include the H (heavy) and L (light) strands, a reference to difference in G+T content between the two stands that arises due to asymmetric replication of the two strands (Reyes *et al.*, 1998). In most insects the majority strand corresponds to the H strand and minority to the L, however as each naming convention has independent basis one cannot assume that they are interchangeable. The arrangement of genes (both gene order and transcription direction) within the mt genome varies widely across bilaterians, however, sufficient conservation between different groups has allowed the recognition of conserved gene blocks (Bernt *et al.*, 2013a) as well as ancestral genome arrangements for the Ecdysozoa (Braband *et al.*, 2010), Pancrustacea and Insecta (Boore *et al.*, 1998). While there are many insects that have mt genome arrangements derived relative to this ancestral insect genome (Fig. 1), the majority of insect species share this arrangement (see Cameron, 2014 for a full discussion of genome rearrangements found in insects). Naming conventions for mt genomes were established by Boore *et al.* (2006), however a variety of alternative names are used e.g. *nad1*, nd1, nad1, NADH1 all describe the same gene.

Mitochondrial Genome Sequencing

Methods for sequencing mt genomes have improved vastly over the last decade and these improvements are largely responsible for the rapid increase in the numbers of available genomes over this time (Boore *et al*., 2005). The first mt genomes were sequenced using the direct isolation of mtDNA either by differential centrifugation to separate mtDNA from nuclear DNA using caesium chloride or of tissue lysate to separate whole mitochondria from other cell components using sucrose (Clary & Wolstenholme, 1985; Crozier & Crozier, 1993). Purified mtDNA was then digested using restriction enzymes, cloned and the clone library sequenced. Mt genomes for only 8 insect species were sequenced using these methods between 1985 (*Drosophila yakuba* Burla: Diptera: Drosophilidae) and 2000 (*Cochliomyia hominovorax* Coquerel: Diptera: Calliphoridae), highlighting the technical demands of this approach. The remaining 98% of insect mt genomes have been sequenced by one of the four methods outlined

below: Long PCR plus primer walking; long PCR plus next-generation sequencing (NGS); RNA sequencing (RNAseq) plus gap filling; and direct shotgun sequencing (Fig. 2, 3).

The introduction of PCR revolutionised mt genomics as it has virtually every other area of molecular biology. Of most relevance to mt genomics is the application of long PCR (sometimes termed long-range PCR), the targeting of amplicons that span multiple genes. It was first applied to insect mt genomes by Roehrdanz (1995) to assess population-level variability in mtDNA via restriction fragment length polymoprhisms (RFLP) and *Triatoma dimidiate* Latreille (Hemiptera: Reduviidae) was the first mt genome to be sequenced using this method (Dotson & Beard, 2001). Long PCR has been used in virtually every insect mt genome sequenced since. From a technical perspective, long PCR doesn't differ greatly from regular PCR. Primers are used to delimit the target amplicon, and the same unmodified oligonucleotide primers can be used as in other PCRs. While it is common to design species-specific primers for long PCR, it is not necessary and primer sets conserved at various taxonomic scales e.g. all animals (Simon *et al.*, 2006), arthropods (Yamauchi *et al.*, 2004), Dictyoptera (Cameron *et al.*, 2012), Coleoptera (Song pers. comm.) have been identified. Long PCRs can also be run on standard PCR machines. Amplification conditions should be changed to reflect the longer amplicons typically by increasing the extension and run-out steps; most commercial enzyme mixes include formulae for calculating required extension times for a range of expected amplicon lengths. Annealing temperatures are defined by primer base composition, additionally, it is useful to reduce the extension temperature by 4°C from manufacturer recommendations due to the high A+T nucleotide bias of insect mt genomes. Many commercial polymerases are suitable for long PCR, however, formulations which include error-checking enzymes such as *Pfu* or have ultra-low error rates are preferred due to possibility of errors accumulating over long target regions.

The advantages of long PCR over direct isolation are enormous; far less tissue is required, preserved insects can be studied and the ability to amplify the entire mt genome in as little as two overlapping PCR fragments is many times faster than mtDNA isolation. Due to the circular nature of mt genomes long PCRs anchored in any gene can be used to amplify the entire genome, it is thus quite flexible with respect to where one starts amplifying a genome. Highly variable gene regions that fail to amplify by short PCRs can be bypassed and amplified through by long PCRs. The weaknesses of the technique include a requirement for high quality templates, susceptibility to changes in genome structure and non-target amplifications. While long PCR's requirement for intact DNA templates covering the entire target region means that high quality preservation is preferred, in practice even relatively poorly preserved tissue can still yield successful amplicons. Standard DNA preservation in 96% ethanol is almost always sufficient and mt genomes successfully amplified from samples preserved in isopropanol or even air dried. Finally,

while mtDNA isolation as described above is usually unnecessary, in practice, most studies target mtDNA rich tissues such as muscle and avoid tissues such as the gut or cuticle which may have high levels of PCR inhibitory metabolites. Tissue specification may not be possible for extremely small insects resulting in unavoidably suboptimal DNA templates.

Failure of long PCR is usually attributable to sequence variation at the primer sites or changes to genome structure due to rearrangements or deletions (e.g. in lice, Cameron *et al*., 2011). Heteroplasmic DNA templates (two or more DNA sequence types in a given specimen) can lead to PCR bias, when the templates differ in size the smallest will be consistently and preferentially amplified. Long PCR also occasionally yields false positives by amplifying numts, nuclear pseudogene copies of mitochondrial genes (Benasson *et al*., 2001). As numts are non-functional and lack any mutational constraint, they are classically distinguished from functional, mt genome copies by the presence of in-frame stop codons. Frame-shift mutations, block deletions and equal substitution rates across all three codon positions, however, are also likely outcomes of incorporation of mtDNA into the nuclear genome and the absence of an in-frame stop codon should not be taken as definitive proof that a particular amplicon is truly mitochondrial. Short PCRs of mt genes are also susceptible to equal or even preferential amplification of numts (Song *et al.*, 2008). While long PCR has been invoked as a solution, there are examples of long PCR generated numts in multiple insect groups; the largest almost 9.5 kb and spanned 28 genes, in a mirid hemipteran (unpublished data). Preprocessing of template DNA to enrich for mtDNA, either via alkaline lysis (Tamura & Aotsuka, 1988) or rolling-cycle amplification (RCA) (Wolff *et al.*, 2012), prior to long PCR have been used to avoid numts but the utility of these methods across a broad range of insect taxa has not been tested.

Sequencing of long PCR amplicons has most often been via Sanger sequencing with primer walking, although NGS methods are rapidly replacing the former method. In primer walking, the ends of each amplicon are sequenced using the amplification primers, the resulting sequence is then used to design novel primers 650 – 800 bp downstream of the initial primers. This second set of primers is used to sequence a further 650+ bp further into the amplicon. This cycle of 'sequence – design new primers – sequence again' is repeated until the entire amplicon has been sequenced; 40 – 50 primers are required for a typical insect mt genome. Consistent with other forms of Sanger sequencing, complete sequencing of the genome in both directions is necessary to avoid sequencing errors. Minor variations include sequencing one species by primer walking and then reusing the resulting primer set on related species (e.g. termites, Cameron & Whiting, 2007; blowflies, Nelson *et al.*, 2012). The principle advantage of primer walking is specificity to the target species that avoids failures due to sequence variability at 'universal' primer sites. The disadvantages are that it is relatively slow and costly. Mitochondrial

genomes can only be sequenced as rapidly as the total number of amplicons, and the speed of each 'step' depends on turnaround times for sequencing and primer purchase. The costs of novel primer design are also significant, typically at least twice the cost of the Sanger sequencing, for what is often a single use primer. Degenerate sequencing primers sets have been designed for broad taxonomic groups (e.g. Lepidoptera, Park *et al.*, 2012) but have yet to be broadly adopted. Finally, the sequencing of the control region by primer walking is often impossible due to sequence simplicity (i.e. insufficient G's and C's to design useful primers), homopolymer runs (e.g. poly A or poly T) and tandem repeats (e.g. Cameron *et al.*, 2012). For this reason a sizable number of the insect mt genomes available on GenBank have not been completely sequenced, these 'near complete' mt genomes have been completely sequenced through the coding regions but the control region is incomplete.

The desire to overcome the limitations of primer walking, has led to enthusiastic application of NGS methods to mt genomics. First used by Jex *et al.* (2008) for parasitic nematodes, the simplest approach involves processing long-PCR amplicons for NGS thus removing the need for primer walking. Comparison with expressed sequence tag (EST) sequences has demonstrated that the method is highly accurate, better capable of detecting nucleotide polymorphisms than Sanger sequencing yet no more susceptible to errors when sequencing homopolymer regions (Jex *et al.*, 2010). Unit costs of most NGS platforms are, however, considerably more than primer walking (Glenn, 2011) and so attention has focused on approaches to multiplexing such that multiple mt genomes can be sequenced from a single NGS run. Libraries constructed from long-PCR products can be labelled with coded DNA-reference tags, termed barcodes (Parameswaran *et al.*, 2007), which allows reads from a single sample to be separately pooled prior to assembly of a contiguous sequence (contig). Timmermans *et al.* (2010), however, have demonstrated that mt genomes can be reassembled without the need for barcoding using Sanger generated 'bait' sequences of short mt genes to match contigs to species identifications. The taxonomic limits of this approach are presently unknown; Timmermans *et al.* (2010) sequenced mixtures of up to 15 beetle species, however all species were from different families. Subsequent studies have focused on a single beetle series (Timmermans *et al.*, 2012: Elateriformia) or superfamily (Haran *et al.*, 2013: Curculionoidea) that pooled multiple representatives at the family and subfamily levels respectively. Studies at finer taxonomic scales run the risk of assembling heterospecific contigs, however the sensitivity of assembler software has yet to be tested in this way.

One limitation of most current applications of NGS to mt genomics is their continuing dependence on long-PCR. Transcriptome datasets generated by RNAseq typically include all of the mt PCG and rRNA genes at high coverage (Nabholz *et al.*, 2010). tRNAs are typically not well represented and transcript mapping against the mt genome typically show peaks towards the middle of the PCGs/rRNAs and very

low/no read depth for tRNA regions (e.g. Margam *et al.*, 2011; Wang *et al.*, 2013). This pattern reflects the balance between the initially multigene (polycistronic) mt transcripts and mature mRNAs which are formed by the excision of tRNAs by endonucleases (see below). Mature mRNAs are captured by RNAseq methods, tRNAs are usually excluded, and polycistronic transcripts are greatly outnumbered by mature mRNA species. No study to date, has reported a complete mt genome assembly from RNAseq, however, this may simply be a factor of sequencing depth; with ever larger transcriptomes being sequenced the coverage of rarer, polycistronic RNA species is likely to improve.

While transcriptome assemblies reliably provide the mt gene sequences typically used in phylogenetic analyses of mt genomes, it is possible to use these sequences as templates to complete sequencing of the genome (e.g Oliveira *et al.*, 2008; Wang *et al.*, 2013). Designing primers based on each mt gene-containing fragment allows the gaps between contigs to be amplified by short-PCRs and sequenced by Sanger methods. While this approach still involves PCR and as such is susceptible to PCR failures, it requires much shorter stretches of intact DNA and usually involves less than half the number of species-specific primers as a full primer walking approach. Given the costs involved in generating a high coverage transcriptome, it is not more economical than primer walking, but rather is a way of deriving extra value from existing transcriptome datasets.

Finally, direct shotgun sequencing of genomic DNA extracts allows the recovery of mt genomes without any amplification or enrichment protocols at all. The first insect mt genome to be sequenced de novo from shotgun sequencing was the human body louse, *Pediculus humanus* Linnaeus, which was assembled from Sanger reads generated as part of the nuclear genome sequencing project (Shao *et al.*, 2009; Kirkness *et al.*, 2010). The unique genome architecture of some louse species including *Pediculus*, i.e. multiple, minicircular chromosomes each with 1-3 genes (Cameron *et al.*, 2011; Wei *et al.*, 2012), had previously defeated long PCR based attempts at sequencing (e.g. Covacin *et al.*, 2006) as target amplicons tried to link protein-coding genes that in actuality were on different chromosomes. Nuclear genome sequencing projects, however, often use demitochondriated samples from which mitochondria have been removed (e.g. pea-aphid genome project; International Aphid Genomics Consortium, 2010), leaving just nuclei for DNA extraction and largely eliminating mt genomic DNA. Additionally, certain assembler programs such as SOAPdenovo (Luo *et al.*, 2012), 'expect' target genome sequences to be present at similar coverage and contigs with significantly higher coverage are treated as repetitious or contaminants and excluded. Due to their higher copy number within the cell, mt genomes can in this way be eliminated from the reported assembly. The precise methods used are thus very relevant to the chance of success in mining mt genomes as a by-product of nuclear genome projects.

More recent studies have focused directly on recovering mt genomes from low-pass NGS runs while treating any resulting nuclear reads as contaminants. No special preparation is used to target mt genomes, whole genomic DNA extractions are fractionated, size-selected and sequenced using any of the standard NGS platforms. Software has been developed to automate assembling mt genomes from NGS reads using either a previously sequenced, close relative as a reference genome, or using individual mt genes from the target species as 'seeds' for iterative assembly (Hahn *et al.*, 2013). Examples of this approach from insects (e.g. Lorenzo-Carballa *et al.*, *in press*; Elbrecht *et al.*, *in press*) are short on detail, however studies from other invertebrate taxa have recorded the entire process (e.g. Groenenberg *et al.*, 2012; Williams *et al.*, 2014). The use of short reads by NGS technologies lends itself to application on degraded tissues (e.g. museum or sub-fossilized specimens) for which long-PCR is impossible. Hung *et al.* (2013) were able to sequence the mt genome of the extinct passenger pigeon (*Ectopistes migratorius*, (Linnaeus)) based on museum specimens 130 years old and a tissue sample 5x2x2 mm in size – smaller than many pinned insects – suggesting that a significant expansion of mt genomic data could be achieved within existing collections. None of the low-pass NGS studies to date, however, have successfully sequenced mt genomes from multiple species indexed onto a single NGS run (*cf.* Williams *et al.*, 2013), making this approach much more expensive than either the primer walking or long-PCR plus multiplexed NGS approaches.

There are thus four viable approaches to sequencing insect mt genomes at the present time. Each have their advantages and disadvantages in terms of cost, speed, reliability and applicability to difficult templates (see Table 1) which should be considered prior to the design of any mt genome sequencing project. Collectively, however, these methods are sufficient to sequence virtually any insect mt genome.

Genome Annotation

Regardless of sequencing method, accurate annotations of mt genomes are then necessary for all downstream analyses. Annotation refers to the process of determining where genes start and finish plus their transcription strand (H or L), the location of repeat regions, and of any other structural features such as the origins of transcription and replication. Several online mt genome annotation pipelines have been developed which use BLAST searches to identify protein-coding genes, covariance analyses to identify tRNAs and output annotated files for GenBank submission. DOGMA (Wyman *et al.*, 2004) was the first package developed, however its internal database of curated mt genomes is now extremely out of date; no new mt genomes have been added since mid-2004 and just 25 insect species are included. MOSAS (Sheffield *et al.*, 2010) used refined tRNA inference methods and a larger, insect focused internal database, however, the program is no longer web hosted at the time of writing. MITOS (Bernt *et al.*, 2013b), is the most advanced annotation pipeline yet produced, however its annotations of protein-coding

genes are wildly unreliable (to the extent of clearly not applying the chosen genetic code correctly). Automated annotation methods have not been widely adopted and majority of insect mt genomes sequenced to date have been hand annotated. The need to validate automated annotations by comparison with hand annotations will likely persist for some time. For these reasons and to highlight annotation issues specific to insects, an outline of the mt genome annotation approach is provided below and conceptually mapped in Figure 4.

Mitochondrial genes are transcribed polycistronically (multiple genes on a single mRNA molecule), then cleaved by an endonuclease at the sites of tRNA secondary structures, liberating mature mRNAs; this is referred to as the tRNA-punctuated model (Ojala *et al.*, 1981). Thus conceptually, the first step in mt genome annotation involves identifying tRNA genes, usually via secondary structure covariation models. Online implementations such as tRNAScan-SE (Lowe & Eddy, 1997) and ARWEN (Laslett & Canback, 2008), predict the presence of tRNAs by identifying sequences with the potential to form the canonical tRNA cloverleaf secondary structure by detecting covariation between complementary stem base positions. tRNA isotype is determined by the sequence at positions 3 – 5 of the anticodon loop. Prediction based on secondary structure, however, misses tRNA isotypes that depart from the cloverleaf structure, e.g. *trnS1* in almost all animals and multiple tRNA isotypes in groups such as gall midges (Beckenback & Joy, 2009) and chelicerates (Domes *et al.*, 2008; Ovchinnikov & Masta, 2012). Isotype specific covarion models have recently been developed (e.g. MiTFi, Juhling *et al.*, 2012, implemented in MITOS which for tRNAs works perfectly), but missing tRNAs are typically annotated by eye. For non-rearranged genomes comparison of sequence at 'expected' tRNA locations with the published mt genomes of close relatives is usually sufficient to identify tRNAs not inferred by automated methods. For rearranged genomes, any regions not assigned to other genes can be searched using generalised RNA secondary structure prediction software such as Mfold (Zuker 2003), to identify potential anticodon stem-loops followed by comparison with the tRNA sequences of other species to test candidate regions. Only a small number of insect species, such as some lice, have genuinely lost one or more tRNA genes from the mt genome. The absence of a particular tRNA from an annotation is usually due to either annotation error or failure to sequence a portion of the genome, especially for genes located near the control region, the most frequently missed portion of 'mostly-complete' mt genomes. Conversely, it is common to find additional tRNA copies beyond the expected 22 genes. All of the inference methods give COVE scores which measure how well a particular region of DNA fits the covariation model for a tRNA, in cases where there are multiple possible copies of the a given isotype the one with the highest COVE score is likely to be the actual, functional copy of the gene. Sequence comparisons with the homologous gene from related species also usually will quickly confirm which of several possibilities, is the real tRNA gene. Additional copies of a tRNA isotype that are inferred to fall within open-reading frames (step 2 below), even if they are encoded on the opposite strand, are almost certainly spurious. tRNA copies that

297 are found in the control region (step 4 below) may represent duplication events, however, the high degree
298 of sequence variation between these copies, the originals and homologues from related species suggests
299 that they are likely non-functional (Cameron *et al*., 2007).

300

301 Following identification of tRNAs, protein-coding genes can be predicted by finding open reading frames
302 between tRNAs (Step 2). Proteins can be identified by BLAST, most reliably using peptide searches such
303 as blastp, blastx or tblastx (Altschul *et al.*, 1997). Note that translation, and thus reading frames is
304 relative to the direction of translation and both the forward and reverse reading frames should be assessed
305 for the potential PCGs. Once PCGs containing regions are identified, the first inframe start codon
306 downstream of its flanking tRNA is typically taken to form the N-terminal end of each gene. There is,
307 however, considerable variability in start codon usage. In addition to the canonical start codons ATN,
308 encoding methionine (M) and isoleucine (I), NTG start codons, encoding lysine (L) and valine (V) are
309 also used across a range of insect taxa (Stewart & Beckenbach, 2009). The tRNA punctuation model also
310 affects the annotation of stop codons. Partial stop codons, a T or TA codon immediately preceding a
311 tRNA, are a common feature of mt protein coding genes. Partial codons are converted to complete TAA
312 stop codons by polyadenylation (Ojala *et al.*, 1981; Stewart & Beckenbach, 2009).

313

314 The annotation of *cox1* is a special case in that it often lacks either a canonical or other potential start
315 codon and its annotation across insects has been wildly inconsistent. In the first insect mt genome to be
316 sequenced, *D. yakuba*, 41 bp separate the preceding tRNA, *trnY*, from the first inframe ATN codon which
317 would encode a peptide 13 amino acids shorter than orthologues. Clary & Wolstenholme (1983) thus
318 proposed a 4 bp start codon, ATAA, for *cox1* in *D. yakuba* that functions as an ATA codon due to either
319 ribosomal frame-shifting or a *trnM* which could read ATAA as a single codon. It should be noted that no
320 evidence for this ATAA start codon was even presented; it was simply a hypothesis to avoid proposing a
321 *cox1* peptide substantially shorter than was found in other species. Furthermore, the 4-bp start codon is
322 not well conserved across Diptera, let alone across insects, for example ATAA, GTAA and TTAA are all
323 found within different *Drosophila* Fallén species (Ballard, 2000). Conversely the *cox1* gene itself is the
324 most highly conserved mt gene at the amino acid level and comparisons across orders led to the proposal
325 of highly conserved sites as start codons for different groups e.g. TCG (S) in Diptera (Beard *et al.*, 1993),
326 CGA (R) in Lepidoptera (Cameron & Whiting, 2008) and CAA (Q), CGA (R) or AAN (N) at a conserved
327 position in Coleoptera (Sheffield *et al.*, 2008). Transcript studies, although only examining a limited
328 number of species e.g. Stewart & Beckenbach (2009), Margam *et al.*, (2011), Neira-Oviedo *et al.*, (2011),
329 have validated the comparative approach predicting the same start codons and finding that the
330 tetranucleotide positions are cleaved from mature *cox1* mRNA. These studies also demonstrate that *cox1*
331 transcripts do not overlap with the upstream tRNA, as has been proposed for several insect species (c.f.

Sheffield *et al.*, 2008, for examples within beetles). Annotation of *cox1* start codons can be justifiably conducted on the basis of comparative amino acid alignments, aiming to identify conserved sites downstream of the flanking tRNA. There is thus no justification for continued speculation about polynucleotide start codons, for proposing annotations that significantly overlap with flanking tRNAs or are significantly longer or shorter than close relatives.

Most of the remaining inconsistencies in protein-coding gene annotations concern those not flanked by tRNAs. In the ancestral insect mt genome there are 4 PCG-PCG gene boundaries resulting in genes for which the mature mRNA transcript is not defined by flanking tRNAs: *atp8-atp6, atp6-cox3, nad4l-nad4* and *nad6-cob*. Two of these, *atp6-cox3* and *nad6-cob* usually, but not universally, overlap by a single base, with the terminal A of the first gene's TAA stop codon forming the first base of the second gene's ATG start codon. Conversely, *atp8-atp6* and *nad4l-nad4* almost always overlap by 7bp with a -1 frame shift (AGA TGA TAA → ATG ATA A). Several instances have, however, been reported of PCG-PCG gene boundaries which lack stop codons due to single base indels within the stop codon of the first gene (Kim *et al.*, 2006; Fenn *et al.*, 2007). Hairpin-loop RNA secondary structures at the 3' end of each gene have been proposed to function like tRNA secondary structures as cleavage sites between PCG-PCG gene boundaries (de Bruijn, 1983; Clary & Woolstenholme, 1985); in such instances polyadenylation would complete the apparently missing stop codons (Kim *et al.*, 2006; Fenn *et al.*, 2007). The secondary structures of the inferred hair-pin loops are, however, highly variable between different insect groups (see Fenn *et al.*, 2007), unlike tRNA secondary structures which are highly uniform. The RNase enzymes responsible for tRNA cleavage are known to be sensitive to tRNA base substitutions (Levinger *et al.*, 1998; Dubrovsky *et al.*, 2004), suggesting that any cleavage at PCG-PCG boundaries is due to other, and as yet unidentified, RNase-like enzymes. The extension of the tRNA-punctuation model to include cleavage at PCG-PCG boundaries is further undermined by transcript studies which suggest that at least some of these gene pairs are co-translated e.g. *atp8-atp6* and *nad4l-nad4* in *Drosophila* (Stewart & Beckenbach, 2009), *atp8-atp6-cox3* in *Maruca* Walker, (Margam *et al.*, 2011). Transcript studies are required from a much broader range of insect taxa so that protein-gene annotations can reflect functional reality. In the meantime, the amino acid sequences at the C- and N-terminal portions of these genes are highly conserved at broad taxonomic scales (e.g. within orders), and thus, as with *cox1*, comparative alignments allow consistent annotations of gene boundaries even in rare instances where stop codons are absent.

With high levels of length variability, the ribosomal RNA genes are perhaps the most difficult mt genes to annotate (Step 3). In the ancestral insect mt genome, *rrnL* is located between two tRNAs (*trnV* and *trnL1*), and this gene has been consistently annotated to occupy every base between these two flanking

genes. While sequencing transcript cDNA has confirmed this for *Drosophila* (Stewart & Beckenbach, 2009), no other insects have been examined despite enormous size variability in this gene, from 868 bp in the wasp *Venturia* Saccardo, (Dowton *et al.*, 2009) to 1514 bp in the flat bug *Neuroctenus* Stål (Hua *et al.*, 2008). Some size variability can be accounted for by expansion regions within the gene, e.g. two vespid wasps *Abispa* Mitchell, and *Polistes* Latreille, differ in size by 100 bp despite high similarity at both 5' and 3' ends (Cameron *et al.*, 2008). Others are due to microsatellite sequences either within the gene (e.g. *Adoxophyes* Meyrick, Lee *et al.*, 2006) or between *rrnL* and flanking tRNAs (e.g. *Helicoverpa* Hardwick, Yin *et al.*, 2010). Secondary structure models of *rrnL* have been proposed (e.g. Gillespie *et al.*, 2006; Niehuis *et al.*, 2006; Cameron & Whiting, 2007), however the 5' end of the molecular, domain I, is poorly conserved across even closely related insects; the 3' end, domain VI has several conserved stems but includes a large, poorly conserved loop and a length variable trailing sequence. Accordingly, secondary structure models have not significantly improved our annotation of homologous regions for this gene. *rrnS* has similarly been very inconsistently annotated, particularly as the 5' end of the gene is not flanked by another gene but rather by the control region. In contrast to *rrnL*, however, the secondary structure of the 5' end of *rrnS* has a high degree of conservation forming part of two pseudoknots that are located between domains II and III. Recognition of this conserved motif (e.g. Song *et al.*, 2010) has resulted in much more consistent annotation of *rrnS*, however GenBank entries for some mt genome submissions still reflect earlier 'guestimate' approaches to delimiting this gene. Software for implementing covariance modelling of rRNA secondary structures has recently been released (e.g. Infernal, Nawrocki *et al.*, 2009, implemented in MITOS), which could potentially result in more consistent annotations of not just gene boundaries but also functional features such as individual domains, stems and loops, within each rRNA.

The non-coding, regulatory features of the mt genome have also not been consistently annotated (Step 4). The origin of replication is typically located in the largest non-coding region and is between *rrnS* and *trnI* in the insect ground plan genome. Rather than identify specific features within it, this entire region is typically annotated as the 'control region' or the 'A+T rich region'. Zhang & Hewitt (1997) proposed a series of five conserved structural elements within the insect control region based on the limited mt genomes available at the time. While Zhang & Hewitt's (1997) structure has proven to be highly descriptive of some groups such as Lepidoptera, overall few of the elements identified are conserved across insects. This is in contrast to the mt genomes of other groups such as vertebrates with highly conserved control region sub-structures (Saccone *et al.*, 1997). The origin of heavy-strand replication (O_H) has been experimentally mapped to a long poly-Thymine stretch that is found in most insects, although its location within the control region varies enormously (Saito *et al.*, 2005). The origin of light-strand replication (O_L) has not been mapped for any insect other than *Drosophila* where it also occurs in the control region and is associated with a second poly-T stretch (Saito *et al.*, 2005). The only other

regulatory element that has been consistently identified is the binding site of mtTERM, a transcription termination peptide, which is located in a non-coding region between *nad1* and *trnS2* in the insect ground plan mt genome. This site has a highly conserved 7bp motif that is conserved across insects (Cameron & Whiting, 2007), even in species such as *Rhagophthalmus* Motschoulsky, where a frame shift mutation results in a longer *nad1* peptide which overlaps the binding site (Sheffield *et al.*, 2008). mtTERM functions to control over-expression of the rRNA genes relative to the protein-coding genes (Taanman, 1999; Roberti *et al.*, 2003), and the mtTERM binding site is lost in rearranged mt genomes where *nad1* is no longer downstream of the rRNA cluster e.g. some hymenopterans (Dowton *et al.*, 2009) and lice (Cameron *et al.*, 2011). The origins of transcription units, of which four are typically inferred (Torres *et al.*, 2009; Beckenbach, 2011), have yet to be mapped for any insect.

Following a first-pass annotation as described above (tRNAs, then PCGs and rRNAs, finally non-coding elements), there is a need for quality control i.e. have the steps followed resulted in a reasonable annotation. Again, the key quality control questions are outlined in Figure 4. Conceptually these are all about whether the mt genome annotation conforms to our 'expectations' – the expected number and type or genes, their transcription direction and size. While it is usual scientific practice to limit *a priori* expectations, in the case of mt genome annotation it is justified due to the demonstrated high level of constraint on this molecule within insects. Departures from the expected number of genes need to be thoroughly investigated to exclude the possibility of mis-annotations or sequencing errors. As outlined above certain tRNA isotypes are only poorly picked up by annotation software and their absence needs to be investigated not blindly accepted. Similarly frame shift mutations resulting in significant extension or truncation of PCGs are far more likely to be due to sequencing errors than real and are best picked up by the primary sequencing lab by examination of their trace files. The sequencing of both genome strands (for Sanger based studies) or with deep coverage (NGS studies), while often not reported is vital to confidence in the reported sequence. Once on GenBank sequence errors are virtually impossible to definitively clear up. Clearly variation is real and there are insect species whose mt genome annotations genuinely depart from one or more of the quality control questions, however, these step serve to narrow our attention on mt genome 'oddities' which have the highest chance of being real rather than simply trusting software outputs.

Finally it also very advisable to check the annotations of previously published mt genomes before using them in phylogenetic or comparative analyses. GenBank doesn't make consistent distinctions between complete, 'near complete' (part of the CR unsequenced) or even 'mostly complete' (one or more genic regions unsequenced) mt genomes and subsequent analyses need to recognize what is actually being compared (e.g. missing genes vs unsequenced genes). Furthermore, the GenBank submission process

includes only limited error checking. Protein-coding gene annotations resulting in frameshifts are flagged (but can be retained by use of the <Exception> function), however other features such as tRNA and rRNA boundaries are not checked and clear errors exist. For example, in a recent analysis of Lepidopteran mt genomes (*unpublished data*), 132 incorrect annotations across 36 species, 3.6 per genome were found or roughly 1 in 20 of the gene boundaries was incorrectly reported in GenBank. While many of these may seem minor, e.g. tRNAs annotated to be 1bp too long or too short, they still result in inaccurate homology statements when aligning genes for phylogenetic analysis. Others, however, are quite substantial and radically change gene alignments with other species e.g. the *rrnS* gene of *Phalera* Hübner was annotated to be 190bp too short due to an unrecognized 225bp repeat in the middle of the gene (Sun *et al.*, 2012). Some are due to errors in earlier publications being propagated into later mt genome annotations. The first published lepidopteran mt genome, *Bombyx mori* (Linnaeus) (Yukihiro *et al.*, 2002) contains many errors that have been followed in the annotation of other species. Similarly due to unrecognised T/TA partial stop codons (as discussed above), large overlaps between *nad4* and *trnH* as well as *nad5* and *trnF* were annotated in the first tortricid mt genome sequenced, *Adoxophryes honmai* Yasuda, (Lee *et al.*, 2006), and these have been followed in other tortricid mt genomes e.g. *Spilonota* Stephens (Zhao *et al.*, 2011), *Grapholita* Treitschke, (Gong *et al.*, 2012). Third party, curated mt genome databases such as MitoZOA (Lupi *et al.*, 2010) have identified many such errors in GenBank submissions, however, these databases are not the usual source for downloading mt genome sequences for analysis, GenBank is. All users of mt genome data should check the accuracy of underlying data in their studies. It is also true that each new genome expands our understanding of what is conserved/variable in insect mt genomes and thus is an opportunity to refine annotations. Of the 126 incorrect boundaries identified above, nine were in species whose mt genomes were published by the author (*Manduca* Hübner: Cameron & Whiting, 2008; *Acraea* Fabricius: Hu *et al.*, 2010; *Spilonota*: Zhou *et al.*, 2011) and with additional data from other species the most probable annotation has changed. Annotation is ultimately our best opinion about gene boundaries which can be produced at a given time, accordingly re-annotation should form a part of all analyses that use mt genome data and any differences from published annotations noted as part of resulting publications.

Conclusions

Whole mt genomes are a useful data source for a wide variety of population genetic, phylogenetic and comparative genomic analyses. Methods for acquiring whole mt genome data have developed rapidly over the last decade and depending upon the scale, budget, time frame and type of templates targeted, different sequencing methods may be most appropriate. Mt genomes can be sequenced reliably, cheaply and rapidly for almost all insect groups and ‘sledgehammer’ NGS based approaches can be applied to those groups that aren’t easy, cheap or timely to sequence. Mt genome annotation requires care and

despite advances in automation it is still advisable that workers in this field be competent in hand-annotation, if only to understand what automated methods are actually doing and the guiding principles behind previous annotations. A functional understanding of how mt genomes are transcribed and how the polycistronic transcripts mature is essential to accurate annotations. A comparative approach to mt genome annotations whereby features conserved across insects or across orders are most likely to represent gene boundaries, especially in the case of non-standard start codons, has been verified by transcript mapping studies. There is no evidence for the existence of polynucleotide codons in mt genomes and there is no excuse for continuing to hypothesize such codons for newly sequenced mt genomes given that transcript studies have disproven their existence. For legacy data, there has been a wide variety in annotation competence between different labs but our understanding of annotations has also evolved over time. Accordingly studies that use mt genomes deposited on GenBank should be re-annotated as part of alignment or comparative analyses to ensure homologous gene comparisons are being applied.

Acknowledgements

Thanks to the students and mentors with whom I have worked on insect mt genomes over the past decade, in particular Stephen Barker, Renfu Shao, Michael Whiting, Mark Dowton and Daniel Fenn. This work has been supported by the US National Science Foundation (DEB0444972, EF0531665), CSIRO Julius Career Awards, QUT Vice Chancellor's Research Fellowship scheme and the Australian Research Council Future Fellowships scheme (FT120100746).

References

Altschul, S.F., Madden, T.L. Schäffer, A.A. Zhang, J. Zhang, Z. Miller, W. & Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389-3402.

Ballard, J.W.O. (2000) Comparative genomics of mitochondrial DNA in members of the *Drosophila melanogaster* subgroup. *Journal of Molecular Evolution*, **51**, 48-63.

Beard, C.B., Hamm, D.M. & Collins, F.H. (1993) The mitochondrial genome of the mosquito *Anopheles gambiae*: DNA sequence, genome organization and comparisons with mitochondrial sequences of other insects. *Insect Molecular Biology*, **2**, 103-24.

Beckenbach, A.T. (2011) Mitochondrial genome sequences of Nematocera (Lower Diptera): Evidence of rearrangement following a complete genome duplication in a winter crane fly. *Genome Biology & Evolution*, **4**, 89-101.

Beckenbach, A.T. & Joy, J.B. (2009) Evolution of the mitochondrial genomes of gall midges (Diptera: Cecidomyiidae): Rearrangement and severe truncation of tRNA genes. *Genome Biology & Evolution*, **1**, 278-87.

Benasson, D., Zhang, D., Hartl, D.L., Hewitt, G.M. (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology & Evolution*, **16**, 314-321.

Bernt, M., Bleidorn, C., Braband, A., Dambach, J., Donath, A., Fritzch, G., Golombek, A., Hadrys, H., Jühling, F., Meusemann, K., Middendorf, M., Misof, B., Perseke, M., Podsiadlowski, L., von Reumont, B., Shierwater, B., Schlegel, M., Schrödl, M., Simon, S., Stafler, P.F., Stöger, I. & Struck, T.H. (2013a) A comprehensive analysis of bilaterian mitochondrial genomes and phylogeny. *Molecular Phylogenetics & Evolution*, **69**, 352-364.

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Putz, J., Middendorf, M. & Stadler, P.F. (2013b) MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics & Evolution*, **69**, 313-319.

Boore, J.L., Lavrov, D.V. & Brown, W.M. (1998) Gene translocation links insects and crustaceans. *Nature*, **392**, 667-8.

Boore, J.L., Macey, J.R. & Medina, M. (2005) Sequencing and comparing whole mitochondrial genomes of animals. *Methods in Enzymology*, **395**, 311-348.

Boore, J.L. (2006) Requirements and standards for organelle genome databases. *OMICS*, **10**, 119-126.

Braband, A., Cameron, S.L., Podsiadlowski, L., Daniels, S.R. & Mayer, G. (2010) The mitochondrial genome of the onychophoran *Opisthopatus cinctipes* (Peripatopsidae) reflects the ancestral mitochondrial gene arrangement of Panarthropoda and Ecdysozoa. *Molecular Phylogenetics & Evolution*, **57**, 285-292.

Burger, G., Gray, M.W. & Lang, B.F. (2003) Mitochondrial genomes: anything goes. *Trends in Genetics*, **19**, 709-716.

Cameron, S.L. (2014) Insect mitochondrial genomics: Implications for evolution and phylogeny. *Annual Review of Entomology*, **59**, 95-117.

Cameron, S.L. & Whiting, M.F. (2008) The complete mitochondrial genome of the tobacco hornworm, *Manduca sexta*, (Insecta: Lepidoptera: Sphingidae), and an examination of mitochondrial gene variability within butterflies and moths. *Gene*, **408**, 112-123.

Cameron, S.L., Lambkin, C.L., Barker, S.C. & Whiting, M.F. (2007) A mitochondrial genome phylogeny of Diptera: Whole genome sequence data accurately resolve relationships over broad timescales with high precision. *Systematic Entomology*, **32**, 40-59.

Cameron, S.L., Dowton, M., Castro, L.R., Ruberu, K., Whiting, M.F., Austin, A.D., Diement, K. & Stevens, J. (2008) The sequence of the mitochondrial genomes of two vespid wasps reveals a number of derived tRNA gene rearrangements. *Genome*, **51**, 800-808.

Cameron, S.L., Sullivan, J., Song, H., Miller, K.B., & Whiting, M.F. (2009) A mitochondrial genome phylogeny of the Neuropterida (lace-wings, alderflies and snakeflies) and their relationship to the other holometabolous insect orders. *Zoologica Scripta*, **38**, 575-590.

Cameron, S.L., Yoshizawa, K., Mizukoshi, A., Whiting, M.F. & Johnson, K.P. (2011) Mitochondrial genome deletions and minicircles are common in lice (Insecta: Phthiraptera). *BMC Genomics*, **12**, 394.

Cameron, S.L., Lo, N., Bourguignon, T., Svenson, G.J. & Evans, T.A. (2012) A mitochondrial genome phylogeny of termites (Blattodea: Termitoidae): Robust support for interfamilial relationships and molecular synapomorphies define major clades. *Molecular Phylogenetics & Evolution*, **65**, 162-173.

Clary, D.O. & Wolstenholme, D.R. (1983) Genes for cytochrome *c* oxidase subunit I, URF2 and three tRNAs in *Drosophila* mitochondrial DNA. *Nucleic Acids Research*, **11**, 6859-6872.

Clary, D.O. & Wolstenholme, D.R. (1985) The mitochondrial DNA molecular of *Drosophila yakuba*: Nucleotide sequence, gene organization, and genetic code. *Journal of Molecular Evolution*, **22**, 252-271.

Covacin, C., Shao, R., Cameron, S.L. & Barker, S.C. (2006) Extraordinary amounts of gene rearrangement in the mitochondrial genomes of lice (Insecta: Phthiraptera). *Insect Molecular Biology*, **15**, 63-68.

Crozier, R.H. & Crozier, Y.C. (1993) The mitochondrial genome of the honeybee *Apis mellifera*: Complete sequence and genome organization. *Genetics*, **133**, 97-117.

de Bruijn, M.H.L. (1983) *Drosophila melanogaster* mitochondrial DNA, a novel organization and genetic code. *Nature*, **304**, 234-241.

Domes, K., Maraun, M., Scheu, S. & Cameron, S.L. (2008) The complete mitochondrial genome of the sexual oribatid mite *Steganacarus magnus*: genome rearrangements and loss of tRNAs. *BMC Genomics*, **9**, 532.

Dotson, E.M. & Beard, C.B. (2001) Sequence and organization of the mitochondrial genome of the Chagas disease vector, *Triatoma dimidiate*. *Insect Molecular Biology*, **10**, 205-215.

Dowton, M., Cameron, S.L, Austin, A.D. & Whiting, M.F. (2009) Phylogenetic approaches for the analysis of mitochondrial genome sequence data in the Hymenoptera – a lineage with both rapidly and slowly evolving mitochondrial genomes. *Molecular Phylogenetics & Evolution*, **52**, 512-519.

Dubrovsky, E.B., Dubrovskaya, V.A., Levinger, L., Schiffer, S. & Marchfelder, A. (2004) *Drosophila* RNase Z processes mitochondrial and nuclear pre-tRNA 3' ends *in vivo*. *Nucleic Acids Research*, **32**, 255-262.

Elbrecht, V., Poettker, L., John, U. & Leese, F. *in press*. The complete mitochondrial genome of the stonefly *Dinocras cephalotes* (Plecoptera, Perlidae). *Mitochondrial DNA*.

Fenn, J.D., Cameron, S.L. & Whiting, M.F. (2007) The complete mitochondrial genome of the Mormon cricket (*Anabrus simplex*: Tettigoniidae: Orthoptera) and an analysis of control region variability. *Insect Molecular Biology*, **16**, 239-252.

Gillespie, J.J., Johnston, J.S., Cannone, J.J., Gutell, R.R. (2006) Characteristics of the nuclear (18S, 5.8S, 28S and 5S) and mitochondrial (12S and 16S) Rrna genes of *Apis mellifera* (Insecta: Hymenoptera): Structure, organization and retrotransposable elements. *Insect Molecular Biology*, **15**, 657-686.

583 Glenn, T.C. (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**,
584 759-769.

585 Gong, Y.-J., Shi, B.-C., Kang, Z.-J., Zhang, F. & Wei, S.-J. (2012) The complete mitochondrial genome
586 of the oriental fruit moth *Grapholita molesta* (Busck) (Lepidoptera: Tortricidae). *Molecular*
587 *Biology Reports*, **39**, 2893-2900.

588 Groenenberg, D.S.J., Pirovano, W., Gittenberger, E. & Schilthuizen, M. (2012) The complete
589 mitogenomne of *Cylindrus obtusus* (Helicidae, Ariantinae) using Illumina next generation
590 sequencing. *BMC Genomics*, **13**, 114.

591 Hahn, C., Bachmann, L. & Chevreux, B. (2013) Reconstructing mitochondrial genomes directly from
592 genomic next-generation sequencing reads – a baiting and iterative mapping approach. *Nucleic*
593 *Acids Research*, **41**, e129.

594 Haran, J., Timmermans, M.J.T.N. & Vogler, A.P. (2013) Mitogenome sequences stabilize the
595 phylogenetics of weevils (Curculionoidea) and establish the monophyly of larval ectophagy.
596 *Molecular Phylogenetics & Evolution*, **67**, 156-66.

597 Hu, J., Zhang, D., Hao, J., Huang, D., Cameron, S.L. & Zhu C.D. (2010) The complete mitochondrial
598 genome of the yellow coaster, *Acraea issoria* (Lepidoptera: Nymphalidae: Heliconiinae:
599 Acraeini): sequence, gene organization and a unique tRNA translocation event. *Molecular Biology*
600 *Reports*, **37**, 3431-3438.

601 Hua, J., Li, M., Dong, P., Cui, Y., Xie, Q. & Bu, W. (2008) Comparative and phylogenomics studies on
602 the mitochondrial genomes of Pentatomorpha (Insecta: Hemiptera: Heteroptera). *BMC Genomics*,
603 **9**, 610.

604 Hung, C.-M., Lin, R.-C., Chu, J.-H., Yeh, C.-F., Yao, C.-J. & Li, S.-H. (2013) The *de novo* assembly of
605 mitochondrial genomes of the extinct passenger pigeon (*Ectopistes migratorius*) with next
606 generation sequencing. *PLoS One*, **8**, e56301.

607 International Aphid Genome Consortium (2010) Genome sequence of the pea aphid *Acyrthosiphon*
608 *pisum*. *PLoS Biology*, **8**, e1000313.

609 Jex, A.R., Hu, M., Littlewood, D.T.J., Waeschenbach, A. & Gasser, R.B. (2008) Using 454 technology
610 for long-PCR base sequencing of the complete mitochondrial genome from single *Haemonchus*
611 *contortus* (Nematoda). *BMC Genomics*, **9**, 11.

612 Jühling, F., Putz, J., Bernt, M., Donath, A., Middendorf, M., Florentz, C. & Stadler, P.F. (2012) Improved
613 systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA
614 structues and into the mechanisms of mitochondrial genome rearrangements. *Nucleic Acids*
615 *Research*, **40**, 2833-2845.

616 Kim, I., Lee, E.M., Seol, K.Y., Yun, E.Y., Lee, Y.B., Hwang, J.S. & Jin, B.R. (2006) The mitochondrial
617 genome of the Korean hairstreak, *Coreana raphaelis* (Lepidoptera: Lycaenidae). *Insect*
618 *Molecular Biology*, **15**, 217-225.

619 Kirkness, E.F., Haas, B.J., Sun, W., Braig, H.R., Perotti, M.A., Clark, J.M., Lee, S.H., Robertson, H.M.,
620 Kennedy, R.C., Elhaik, E., Gerlach, D., Kriventseva, E.V., Elsik, C.G., Graur, D., Hill, C.A.,
621 Veenstra, J.A., Walenz, B., Tubio, J.M.C., Ribeiro, J.M.C., Rozas, J., Johnston, J.S., Reese, J.T.,
622 Popadic, A., Tomoyasu, Y., Tojo, M., Raoult, D., Reed, D.L., Kraus, E., Mittapalli, O., Margam,
623 V.M., Li, H.-M., Meyer, J.M., Johnson, R.M., Romero-Severson, J., Pagel VanZee, J., Alvarez-
624 Ponce, D., Vieira, F.G., Aguade, M., Guirao-Rico, S., Anzola, J.M., Yoon, K.S., Strycharz, J.P.,
625 Unger, M.F., Christley, S., Lobo, N.F., Seufferheld, M.J., Wang, N.K., Dasch, G.A., Struchiner
626 C.J., Madey, G., Hannick, L.I., Bidwell, S., Joardar, V., Caler, E., Shao, R. Barker, S.C.,
627 Cameron, S.L., Bruggner, R.V., Regier, A., Johnson, J., Viswanathan, L., Utterback, T.R., Sutton,
628 G.G., Lawson, D., Waterhouse, R.M., Venter, J.C., Strausberg, R.L., Berenbaum, M., Collins,
629 F.H., Zdobnov, E.M. & Pittendrigh, B.R. (2010) Genome sequences of the human body louse and

its primary endosymbiont: Insights into the permanent parasitic lifestyle. *Proceedings of the National Academy of Sciences USA*, **107**, 12168-12173.

Laslett, D. & Canbeck, B. (2008) ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. *Bioinformatics*, **24**, 172-175.

Lee, E.-S., Shin, K.S., Kim, M.-S., Park, H., Cho, S., Kim, C.-B. (2006) The mitochondrial genome of the smaller tea tortrix *Adoxophyes honmai* (Lepidoptera: Tortricidae). *Gene*, **373**, 52-57.

Levinger, L., Vasisht, V., Greene, V., Bourne, R., Birk, A. & Kolla, S. (1995) Sequence and structure requirements for *Drosophila* tRNA 5'- and 3'- end processing. *Journal of Biological Chemistry*, **270**, 18903-18909.

Lorenzo-Carballa, M.O., Thompson, D.J., Cordero-Rivera, A. & Watts, P.C. *in press*. Next generation sequencing yields the complete mitochondrial genome of the scarce blue-tailed damselfly, *Ischnura pumilio*. *Mitochondrial DNA*.

Lowe, T.M. & Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, **25**, 955-964.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu., G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T.-W. & Wang, J. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience*, **1**, 18.

Lupi, R., D'Onorio de Meo, P., Picardi, E., D'Antonio, M., Paoletti D., Castignano, T., Pesole, G. & Gissi, C. (2010) MitoZoa: A curated mitochondrial genome database of metazoans for comparative genomics studies. *Mitochondrion*, **10**, 192-199.

Ma, C., Yang, P.C., Jiang, F., Chapuis, M.-P., Shall, Y., Sword, G.A. & Kang, L. (2012) Mitochondrial genomes reveal the global phylogeography and dispersal routes of the migratory locust. *Molecular Ecology*, **21**, 4344-4358.

Margam, V.M., Coates, B.S., Hellmich, R.L., Agunbiade, T., Seufferheld, M.J., Sun, W., Ba, M.N., Sanon, A., Binso-Dabire, C.L., Baoua, I., Ishiyaku, M.F., Covas, F.G., Srinivasan, R., Armstrong, J., Murdock, L.L. & Pittendrigh, B.R. (2011) Mitochondrial genome sequences and expression profiling for the Legume Pod Borer *Maruca vitrata* (Lepidoptera: Crambidae). *PLoS One*, **6**, e16444.

Nabholz, B., Jarvis, E.D. & Ellegren, H. (2010) Obtaining mtDNA genomes from next-generation transcriptome sequencing: A case study on the basal Passerida (Aves: Passeriformes) phylogeny. *Molecular Phylogenetics & Evolution*, **57**, 466-470.

Nawrocki, E.P., Kolbem D.L. & Eddy, S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335-1337.

Neira-Oviedo, M., Tsyganov-Bodounov, A., Lycett, G.J., Kokoza, V., Raikhel, A.S. & Krzywinski, J. (2011) The RNA-seq approach to studying the expression of mosquito mitochondrial genes. *Insect Molecular Biology*, **20**, 141-152.

Nelson, L.A., Lambkin, C.L., Batterham, P., Wallman, J.F., Dowton, M., Whiting, M.F., Yeates, D.K. & Cameron, S.L. (2012) Beyond Barcoding: Genomic approaches to molecular diagnostics in blowflies (Diptera: Calliphoridae). *Gene*, **511**, 131-142.

Niehuis, O., Naumann, C.M. & Misof, B. (2006) Identification of evolutionary conserved structural elements in the mt SSU Rrna of Zygaenoidea (Lepidoptera): A comparative sequence analysis. *Organisms, Diversity & Evolution*, **6**, 17-32.

674 Ojala, D., Montoyo, J. & Attardi, G. 1981. Trna punctuation model of RNA processing in human
675 mitochondria. *Nature*, **290**, 470-474.

676 Oliveira, D.C.S.G., Raychoudhury, R., Lavrov, D.V. & Werren, J.H. (2008) Rapidly evolving
677 mitochondrial genome and directional selection in mitochondrial genes in the parasitic wasp
678 *Nasonia* (Hymenoptera: Pteromalidae). *Molecular Biology & Evolution*, **25**, 2167-2180.

679 Osigus, H.-J., Eitel, M., Bernt, M., Donath, A. & Schierwater, B. (2013) Mitogenomics at the base of
680 Metazoa. *Molecular Phylogenetics & Evolution*, **69**, 339-351.

681 Ovchinnikov, S. & Masta, S.E. (2012) Pseudoscorpion mitochondria show rearranged genes and genome-
682 wide reductions of RNA gene sizes and inferred structures, yet typical nucleotide composition
683 bias. *BMC Evolutionary Biology*, **12**, 31.

684 Park, J.S., Cho, Y., Kim, M.J., Nam, S.-H. & Kim, I. (2012) Description of the complete mitochondrial
685 genome of the black-veined white, *Aporia crataegi* (Lepidoptera: Papilionoidea), and a
686 comparison to papilionoid species. *Journal of Asia-Pacific Entomology*, **15**, 331-341.

687 Parameswaran, P., Jalili, R., Tao, L., Shokralla, S., Gharizadeh, B., Ronaghi, M. & Fire, A.Z. (2007) A
688 pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample
689 multiplexing. *Nucleic Acids Research*, **35**, e130.

690 Reyes, A., Gissi, C., Pesole, G. & Saccone, C. (1998) Asymmetrical directional mutation pressure in the
691 mitochondrial genome of mammals. *Molecular Biology & Evolution*, **15**, 957-966.

692 Roberti, M., Polosa, P.L., Bruni, F., Musicco, C., Gadaleta, M.N. & Cantatore, P. (2003) DmTTF, a novel
693 mitochondrial transcription factor that recognizes two sequences of *Drosophila melanogaster*
694 mitochondrial DNA. *Nucleic Acids Research*, **31**, 1597-1604.

695 Roehrdanz, R.L. (1995) Amplification of complete insect mitochondrial genomes in two easy pieces.
696 *Insect Molecular Biology*, **4**, 169-172.

697 Saccone, C., Arrimonelli, M. & Sbisa, E. (1987) Structural elements highly preserved during the
698 evolution of the D-Loop containing region in vertebrate mitochondrial DNA. *Journal of*
699 *Molecular Evolution*, **26**, 205-211.

700 Saito, S., Tamura, K. & Aotsuka, T. (2005) Replication origin of mitochondrial DNA in insects.
701 *Genetics*, **171**, 1695-1705.

702 Shao, R., Kirkness, E.F. & Barker, S.C. (2009) The single mitochondrial chromosome typical of animals
703 has evolved into 18 minichromosomes in the human body louse, *Pediculus humanus*. *Genome*
704 *Research*, **19**, 904-12.

705 Sheffield, N.C., Song, H., Cameron, S.L. & Whiting, M.F. (2008) A comparative analysis of
706 mitochondrial genomes in Coleoptera (Arthropoda: Insecta) and genome descriptions of six new
707 beetles. *Molecular Biology & Evolution*, **25**, 2499-2509.

708 Sheffield, N.C., Hiatt, K.D., Valentine, M.C., Song, H. & Whiting, M.F. (2010) Mitochondrial genomics
709 in Orthoptera using MOSAS. *Mitochondrial DNA*, **21**, 87-104.

710 Simon, C., Buckley, T.R., Frati, F., Stewart, J.B. & Beckenbach, A.T. (2006) Incorporating molecular
711 evolution into phylogenetic analysis and a new compilation of conserved polymerase chain
712 reaction primers for animal mitochondrial DNA. *Annual Review of Ecology, Evolution and*
713 *Systematics*, **37**, 545-579.

714 Song, H., Buhay, J.E., Whiting, M.F. & Crandall, K.A. (2008) Many species in one: DNA barcoding
715 overestimates the number of species when nuclear mitochondrial pseudogenes are co-amplified.
716 *Proceedings of the National Academy of Sciences, USA*, **105**, 13486-13491.

Song, H., Sheffield, N.C. Cameron, S.L., Miller, K.B. & Whiting, M.F. (2010) What happens when the phylogenetic assumptions are violated?: The effect of base compositional heterogeneity and among-site rate heterogeneity in beetle mitochondrial phylogenomics. *Systematic Entomology*, **35**, 429-448.

Stewart, J.B. & Beckenbach, A.T. (2009) Characterization of mature mitochondrial transcripts in *Drosophila* and the implications for the tRNA punctuation model in arthropods. *Gene*, **445**, 49-57.

Sun, Q.-Q., Sun, X.-Y., Wang, X.-C., Gai, Y.-H., Hu, J., Zhu, C.D. & Hao, J.-S. (2012) Complete sequence of the mitochondrial genome of the Japanese buff-tip moth, *Phalera flavescens* (Lepidoptera: Notodontidae). *Genetics & Molecular Research*, **11**, 4213-4225.

Taanman, J.-W. (1999) The mitochondrial genome: Structure, transcription, translation and replication. *Biochimica et Biophysica Acta*, **1410**, 103-123.

Tamura, K. & Aotsuka, T. (1988) Rapid isolation method of animal mitochondrial DNA by the alkaline lysis procedure. *Biochemical Genetics*, **26**, 815-819.

Torres, T.T., Dolezal, M., Schlotterer, C. & Ottenwalder, B. (2009) Expression profiling of *Drosophila* mitochondrial genes via deep mRNA sequencing. *Nucleic Acids Research*, **37**, 7509-7518.

Timmermans, M.J.T.N., Dodsworth, S., Culverwell, C.L., Bocak, L. Ahrens D, Littlewood, D.T.J., Pons, J. & Vogler, A.P. (2010) Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Research*, **38**, e197.

Timmermans, M.J.T.N. & Vogler, A.P. (2012) Phylogenetically informative rearrangements in mitochondrial genomes of Coleoptera, and monophyly of aquatic elateriform beetles (Dryopoidea). *Molecular Phylogenetics & Evolution*, **63**, 299-304.

Wan, X., Kim, M.I., Kim, M.J. & Kim, I. (2012) Complete mitochondrial genome of the free-living earwig, *Challia fletcheri* (Dermaptera: Pygidicranidae) and phylogeny of Polyneoptera. *PLoS One*, **7**, e42056.

Wang, H.-L., Yang, J., Boykin, L.M., Zhao, Q.-Y., Wang, X.-W. & Liu, S.-S. (2013) The characteristics and expression profile of the mitochondrial genome for the Mediterranean species of *Bemisia tabaci* complex. *BMC Genomics*, **14**, 401.

Wang, Y., Liu, X., Winterton, S.L. & Yang, D. (2012) The first mitochondrial genome for the fishfly subfamily Chauliodinae and implications for the higher phylogeny of Megaloptera. *PLoS One*, **7**, e47302.

Wei, D.D., Shao, R., Yuan, M.-L., Dou, W., Barker, S.C. & Wang, J.-J. (2012) The multipartite mitochondrial genome of *Liposcelis bostrychophila*: Insights into the evolution of mitochondrial genome in bilaterian animals. *PLoS One*, **7**, e33973.

Wolff, J.N., Shearman, D.C.A., Brooks, R.C. & Ballard, J.W.O. 2012. Selective enrichment and sequencing of whole mitochondrial genomes in the presence of nuclear encoded mitochondrial pseudogenes (Numts). *PLoS One*, **7**, e37142.

Williams, S.T., Foster, P.G. & Littlewood, D.T.J. (2014) The complete mitochondrial genome of a turbinid vetigastropod from MiSeq Illumina sequencing of genomic DNA and steps towards a resolved gastropod phylogeny. *Gene*, **533**, 38-47.

Wyman, S.K., Jensen, R.K. & Boore, J.L. (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, **20**, 3252-3255.

Yamauchi, M.M., Miya, M. & Nishida, M. (2004) Use of a PCR-based approach for sequencing whole mitochondrial genomes of insects: two examples (cockroach and dragonfly) based on the method developed for decapod crustaceans. *Insect Molecular Biology*, **13**, 435-442.

762 Yin, J., Hong, G.-Y., Wang, A.-M., Cao, Y.-Z. & Wei, Z.-J. (2010) Mitochondrial genome of the cotton
763 bollworm *Helicoverpa armigera* (Lepidoptera: Noctuidae) and comparison with other
764 Lepidoptera. *Mitochondrial DNA*, **21**, 160-169.

765 Yukuhiro, K., Sezutsu, H., Itoh, H., Shimizu, K. & Banno, Y. (2002) Significant levels of sequence
766 divergence and gene rearrangements have occurred between the mitochondrial genomes of the
767 wild mulberry silkmoth, *Bombyx mandarina*, and its close relative, the domesticated silkmoth,
768 *Bombyx mori*. *Molecular Biology & Evolution*, **19**, 1385-1389.

769 Zhang, D.X. & Hweitt, G.M. (1997) Insect mitochondrial control region: A review of its structure,
770 evolution and usefulness in evolutionary studies. *Biochemical Systematics & Ecology*, **25**, 99-
771 120.

772 Zhao, J. Winterton, S.L. & Liu, Z. (2013a) Ancestral gene organization in the mitochondrial genome of
773 *Thyridosmylus langii* (McLachlan, 1970) (Neuroptera: Osmylidae) and implications for lacewing
774 evolution. *PLoS One*, **8**, e62943.

775 Zhao, F., Huang, D.-Y., Sun, X.-Y., Shi, Q.-H., Hao, J.-S., Zhang, L.-L. & Yang, Q. (2013b) The first
776 mitochondrial genome for the butterfly family Riodinidae (*Abisara fylloides*) and its systematic
777 implications. *Zoological Research*, **34**, E109-E119.

778 Zhao, J.-L., Zhang, Y.-Y., Luo, A.-R., Jiang, G.-F., Cameron, S.L. & Zhu, C.-D. (2011) The complete
779 mitochondrial genome of *Spilonota lechriaspis* Meyrick (Lepidoptera: Tortricidae). *Molecular*
780 *Biology Reports*, **38**, 3757-3764.

781 Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids*
782 *Research*, **31**, 3406-3415.

783

784 Table 1. Advantages and disadvantages of different mt genome sequencing methods.

785

	Long PCR plus primer walking	Long PCR plus Next-gen sequencing	RNAseq plus gap filling	Direct shotgun sequencing
Speed	Slow, 2-3 months	Fast, 1-2 weeks	Fast, 1-2 weeks	Very fast, 2-3 days
Cost *	Moderate, US$500	Low, < US$100	High, US$1000 (inc. RNAseq run)	High, US$750+
Acceptable template quality	Broad, ethanol or dried specimens successful	Broad, ethanol or dried specimens successful	Narrow, RNA extracts needed	Broad, ethanol or dried specimens successful
Ease of lab. Procedures	Very easy, standard PCR methods	Moderate, NGS template prep/ library indexing	High, RNA extraction and sequencing	Moderate, NGS template prep/ library indexing
Multiplexing	No	Yes	No	Yes
Specialised equipment	None.	NGS platform	NGS platform and RNA extraction facilities	NGS platform
Assembly complexity	Low, any contig assembly software	Low, any contig assembly software	High, de novo transcriptome assembly required	High, de novo genome assembly required

786

787 * Precise costs depends on local sequencing centre, for NGS applications it depends on platform and how
788 many samples are multiplexed into a single run, but the relative pricing is the key point. NGS Prices after
789 Glenn (2011).

790

Figure Legends

Figure 1. Map of the ancestral insect mt genome, linearized between the control region (CR) and *trnI*. The length of each gene is approximately proportional to its DNA length. Protein-coding genes are coloured coded by OXPHOS complex (*cox*: Blue; *nad*: Green; *atp*: Orange; *cob*: Yellow); tRNAs: White; rRNAs: Grey; and control region: Black. Gene names are the standard abbreviations used in this paper; tRNA genes are indicated by the single letter IUPAC-IUB abbreviation for their corresponding amino acid; genes transcribed on the minority strand are underlined.

Figure 2. Mitochondrial genome sequencing procedures. Short PCRs: Green; Long PCRs/Long PCR fragments: Light Blue.

Figure 3. Mitochondrial genome sequencing procedures (continued). Short PCRs: Green; RNAseq Contigs: Yellow; Genomic DNA: Pink.

Figure 4. Flowchart for annotation procedures for mt genomes plus quality control questions to resolve conflicts in first pass annotations.

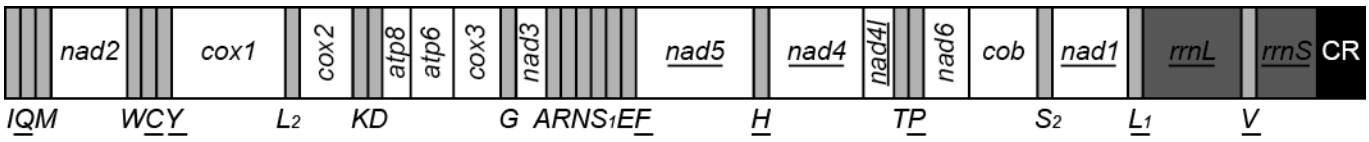
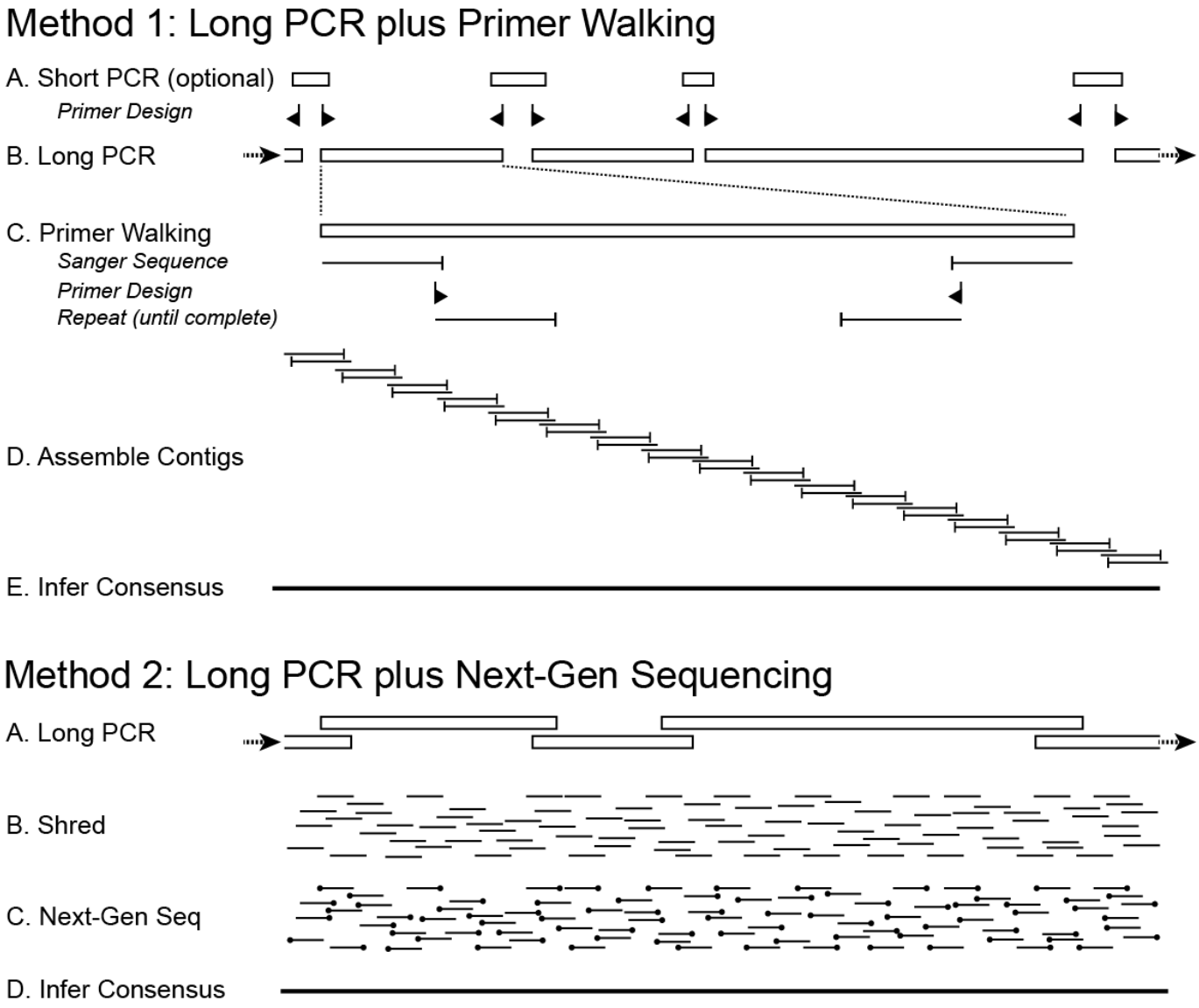
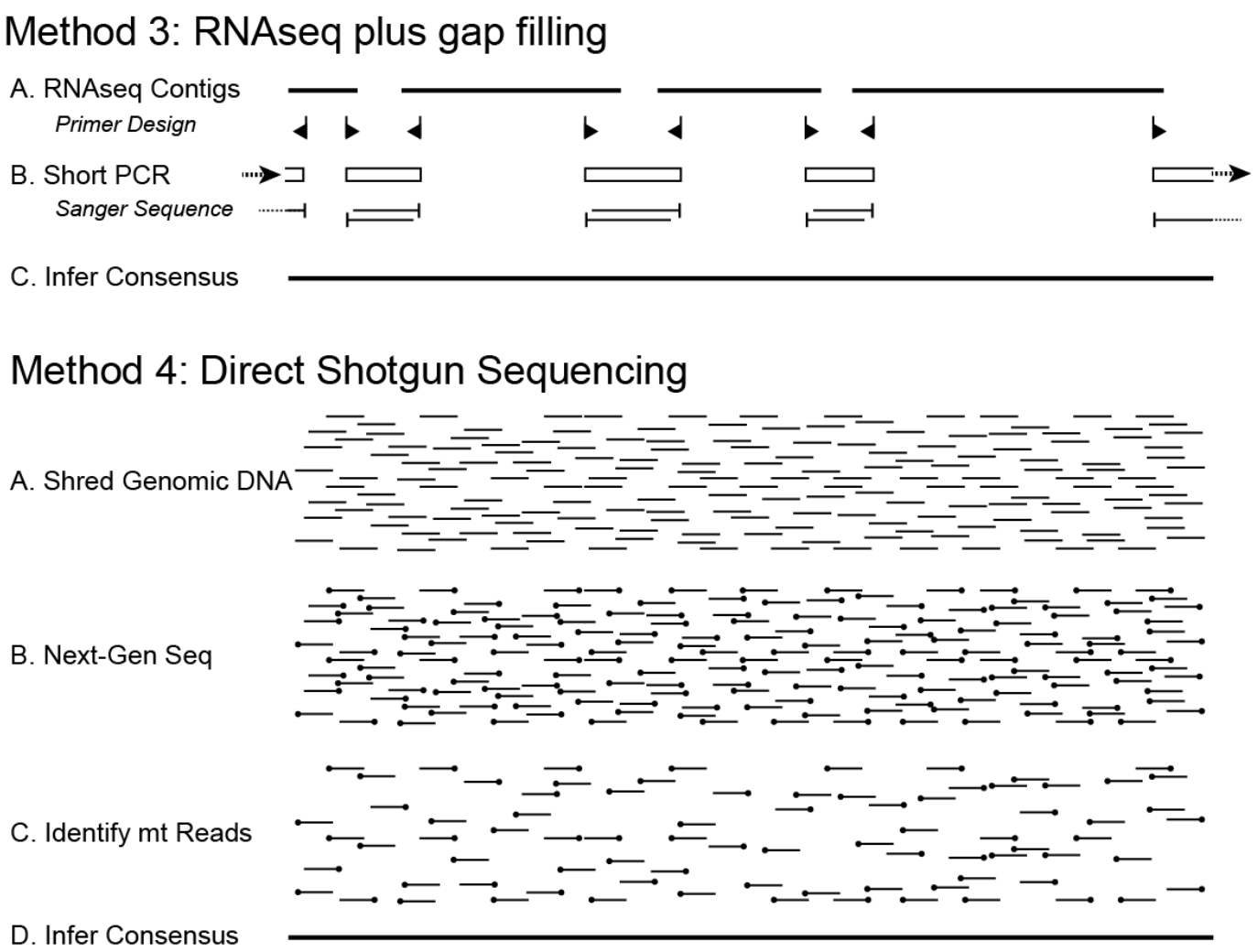

nad2
cox1
cox2
atp8
atp6
cox3
nad3
nad5
nad4
nad4l
nad6
cob
nad1
rrnL
rrnS
CR
IQM
WCY
L2
KD
G
ARNS1EF
H
TP
S2
L1
V

Fig 1.


Method 1: Long PCR plus Primer Walking
A. Short PCR (optional)
Primer Design
B. Long PCR
C. Primer Walking
Sanger Sequence
Primer Design
Repeat (until complete)
D. Assemble Contigs
E. Infer Consensus
Method 2: Long PCR plus Next-Gen Sequencing
A. Long PCR
B. Shred
C. Next-Gen Seq
D. Infer Consensus

Fig, 2

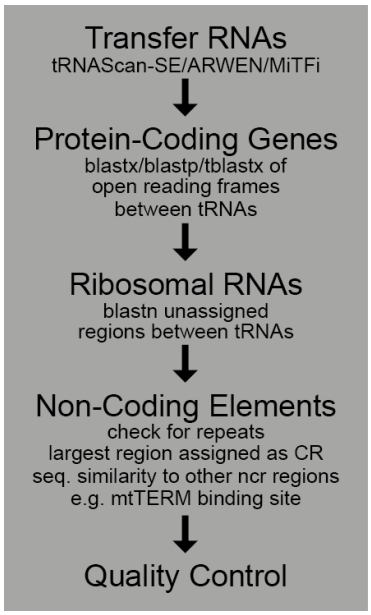


816

817

818 Fig 3.

819



Quality Control

1. *Are the expected 22 tRNA isotypes present ?*

YES	Great, continue!
NO, SOME MISSING	Check unassigned regions, c.f. tRNA sequences of relatives.
NO, EXTRA tRNAs	Found within an ORF? Probably spurious.
	Found in a non-coding region? Compare COVE scores.

2. *Are the expected 13 Protein-Coding Genes present ?*

YES	Great, continue!
NO, SOME MISSING	Check for frame-shift mutations/ check reads for seq. error.
	If no unassigned regions, PCG may be genuinely missing.
NO, EXTRA PCGs	Check for duplicated genes e.g. sequence repeats.
	Non-mt PCG with low blast score? Probably spurious.

3. *Are the PCGs of expected length ?*

YES	Great, continue!
NO	Check for frame-shift mutations/ check reads for seq. error.
	Compare start codon to related species. Possible non-standard codons.
	Compare stop codon to related species. Unrecognised tRNA editing?

4. *Are the rRNAs of expected length ?*

YES	Great, continue!
NO	Compare to rRNA secondary structure models. Look for indels in conserved regions/loop expansions/repeat units.

820

821

822 Fig. 4.