MITOS: Improved *de novo* Metazoan Mitochondrial Genome Annotation

Matthias Bernt[a,*], Alexander Donath[b,c,*], Frank Jühling[b,h], Fabian Externbrink[a], Catherine Florentz[h], Guido Fritzsch[b], Joern Pütz[h], Martin Middendorf[a], Peter F. Stadler[b,d,e,f,g]

[a] *Parallel Computing and Complex Systems Group, Department of Computer Science, University Leipzig, Johannisgasse 26, 04103 Leipzig, Germany*
[b] *Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, Universität Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany*
[c] *Zentrum für Molekulare Biodiversitätsforschung, Zoologisches Forschungsmuseum Alexander Koenig, Adenauerallee 160, 53113 Bonn, Germany*
[d] *Max-Planck-Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany*
[e] *Fraunhofer Institut für Zelltherapie und Immunologie Perlickstraße 1, 04103 Leipzig, Germany*
[f] *Department of Theoretical Chemistry University of Vienna, Währingerstraße 17, 1090 Wien, Austria*
[g] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*
[h] *Architecture et Réactivité de l'ARN, Université de Strasbourg, CNRS, IBMC, 15 rue René Descartes, 67084 Strasbourg, France*

Abstract

About 2000 completely sequenced mitochondrial genomes are available from the NCBI `RefSeq` data base together with manually curated annotations of their protein-coding genes, rRNAs, and tRNAs. This annotation information, which has accumulated over two decades, has been obtained with a diverse set of computational tools and annotation strategies. Despite all efforts of manual curation it is still plagued by misassignments of reading directions, erroneous gene names, and missing as well as false positive annotations in particular for the RNA genes. Taken together, this causes substantial problems for fully automatic pipelines that aim to use these data comprehensively for studies of animal phylogenetics

*Corresponding author
Email addresses: `bernt@informatik.uni-leipzig.de` (Matthias Bernt), `a.donath@zfmk.de` (Alexander Donath), `frank@bioinf.uni-leipzig.de` (Frank Jühling), `fabian@bioinf.uni-leipzig.de` (Fabian Externbrink), `c.florentz@ibmc-cnrs.unistra.fr` (Catherine Florentz), `guido@bioinf.uni-leipzig.de` (Guido Fritzsch), `j.puetz@unistra.fr` (Joern Pütz), `middendorf@informatik.uni-leipzig.de` (Martin Middendorf), `studla@bioinf.uni-leipzig.de` (Peter F. Stadler)

and the molecular evolution of mitogenomes. The MITOS pipeline is designed to compute a consistent *de novo* annotation of the mitogenomic sequences. We show that the results of MITOS match RefSeq and MitoZoa in terms of annotation coverage and quality. At the same time we avoid biases, inconsistencies of nomenclature, and typos originating from manual curation strategies. The MITOS pipeline is accessible online at http://mitos.bioinf.uni-leipzig.de.
Keywords: Metazoa, Mitochondria, Genome, Annotation, Server

1. Introduction

A reliable and standardised genome annotation is an indispensable prerequisite for a systematic comparative analysis of genomic sequence data. This is true in particular for phylogenetic reconstruction, studies of the mechanisms of genome rearrangements, and the investigation of the effects of sequence variation. The need for accurate and unbiased annotations becomes even more pressing when automatised pipelines are employed to process the increasingly large amounts of data that are becoming available in the wake of new sequencing technologies.

At present, complete sequences of mitochondrial genomes are available for more than 2000 metazoan species from a wide variety of taxonomic groups. Metazoan mitogenomes are (with few exceptions) circular molecules with an average length of approximately 16 500 nt with extreme length values such as 11 423 nt (*Paraspadella gotoi* NC_006083) and 43 079 nt (*Trichoplax adhaerens* NC_008151). Mitochondrial genomes have a well preserved gene content usually comprising 13 protein coding genes, 22 tRNAs, two rRNAs, and one non-coding region containing most of the regulatory elements (Wolstenholme, 1992). This simple structure makes animal mitogenomes an attractive target for large-scale comparative studies.

Mitochondrial genes usually consist of a single continuous exon, although in some clades exceptions have been reported in protein coding genes as well as in rRNAs (Beagley et al., 1996; Dellaporta et al., 2006; Wang and Lavrov, 2008)

and conserved frameshifts exist in some sauropsid groups (Mindell et al., 1998). In several cases there is also evidence for some duplication and deletion events (e.g. San Mauro et al., 2006; Fujita et al., 2007). A peculiarity of mitogenomes is their use of deviant genetic codes and the presence of overlapping genes and incomplete stop codons (Wolstenholme, 1992; Jühling et al., 2011), see Bernt et al. (2012b) (in this special issue) for a more detailed overview. Taken together, all these issues complicate the task of genome annotation and made extensive manual "expert curation" indispensable. In this process a multitude of different tools have been used by different curators. As discussed e.g. by Boore (2006), this entails a number of problems: a) tools used in older annotation may be outdated, i.e. improved methods are already available, b) sequences used as basis for homology annotation can be either wrong or incomplete, and c) no generally accepted guidelines exist for the annotation.

The most comprehensive and up-to-date resource for mitochondrial genomes and their annotation is NCBI `RefSeq` (Pruitt et al., 2007). Despite substantial efforts by the curators of `RefSeq` to improve the quality of the data several inconsistencies and errors in the annotations have remained that cause problems for automatised analysis pipelines. This includes missing or incorrect information of the reading direction (strand), erroneous gene designations, missing gene annotations, mistaken identity of *trnL1*/*trnL2* and *trnS1*/*trnS2* tRNAs, and inconsistencies in gene names (see Supplement 1 for selected examples).

Boore (2006) suggested a number of possible solutions to overcome these problems: Systematic error screening, standardisation of gene names, anticodon labelling of tRNAs, standards for gene and gene boundaries designation, and standards for accepting the reality of a gene assignment. Several data bases, reviewed in more detail in Bernt et al. (2012a) (in this special issue), aim at providing improved annotations for `RefSeq` mitogenomes along these lines. `METAMiGA` (Feijao et al., 2006) and `OGRe` (Jameson et al., 2003) incorporate manual improvements of the data based on expert knowledge. Systematic semi-automatic error screening using a list of rules based on `tRNAscan-SE` (Lowe and Eddy, 1997), `ARWEN` (Laslett and Canback, 2008), and `BLAST` (Altschul et al., 1990)

searches as well as expert knowledge is used for MitoZoa (Lupi et al., 2010), a recently released new data base.

De novo annotation with a consistent set or pipeline of methods is a promising alternative to evaluating and improving existing annotations. DOGMA (Wyman et al., 2004) is a semi-automated pipeline of methods dealing with both mitochondrial and chloroplast genomes. It uses BLAST to identify coding and non-coding genes. COVE (Eddy and Durbin, 1994) is employed by DOGMA to identify tRNAs candidates based on secondary structure. MOSAS (Sheffield et al., 2010) is a set of methods that has its focus on the organisation of sequence data and annotation and was originally intended for insect mitogenomes. It employs ARWEN and tRNAscan-SE for tRNA prediction. BLAST is used by MOSAS to search for open reading frames and rRNAs based on a local data base of query sequences (currently from insects only). The need for user-defined cutoff values and manual improvements of the predictions makes this approach difficult to apply to large data sets and limits the comparability of the predictions.

The MITOchondrial genome annotation Server (MITOS) provides access to a fully automated pipeline for the *de novo* annotation of metazoan mitochondrial genomes. It uses a novel strategy based on aggregating BLAST searches with previously annotated protein sequences to identify protein coding genes (Section 2.1), thereby avoiding the need for a built-in data base of specifically curated protein models. Both tRNAs and rRNAs are annotated using specific covariance models for each of the structured RNAs (Section 2.2). In this contribution we apply MITOS for the *de novo* annotation of all animal mitogenomes contained in RefSeq 39, focusing on a careful evaluation of the quality of the results (Section 3).

2. Materials and Methods

MITOS requires only a sequence file in FASTA format and the corresponding genetic code as input. The pipeline proceeds in two stages, first identifying candidate sequences for each gene, then reconciling these to derive a final an-

notation. In the following we provide a detailed description of the individual components of MITOS.

2.1. Protein Homology Search

The annotation of each protein coding gene starts with a BLASTX-based similarity search using as queries the amino acid sequences of previously annotated orthologs. These are taken from a data base which we assume to contain some inconsistencies, misannotations, incomplete sequences, or other errors. We therefore consider aggregations of matches at the same locations of the input mitogenome. This is motivated by the assumption that aggregates generated by correctly annotated genes will dominate those generated by a moderate number of erroneous queries. The aggregation process thus serves as an automatic filter of the query data base and relieves us of the necessity to first manually curate the queries. In the following we describe the technical details of this strategy.

2.1.1. Extraction of reference amino acid sequences

The basis for the annotation is the collection of all amino acid sequences of all proteins annotated in a complete metazoan mitochondrial genome. These sequences were extracted directly from the *CDS* feature of the GenBank files available for all RefSeq 39 mitogenomes. We implemented specialised parser for mitogenome annotations based on biopython (Cock et al., 2009) for this task. Protein sequences are compiled separately for each query gene.

2.1.2. Similarity Search and Aggregation

For a concise description of the pipeline we need to introduce a bit of notation. A BLASTX *hit* is characterised by its start and end position in the query (s_q, e_q) and target sequence (s_t, e_t) and its *quality* measured as $-\log_{10}$(E-value). For E-values of 0 we set the quality to 100. For a given hit, the *relative query position* p_q of a position p_t in the target sequence is the corresponding position in the query sequence, i.e. a position with $\frac{p_q - s_q}{e_q - p_q} = \frac{p_t - s_t}{e_t - p_t}$ (see Figure 1). For a given position p and a set of BLAST hits, we define the *quality of position* p as the sum of the quality values of all hits that include p. Analogously, the *relative*

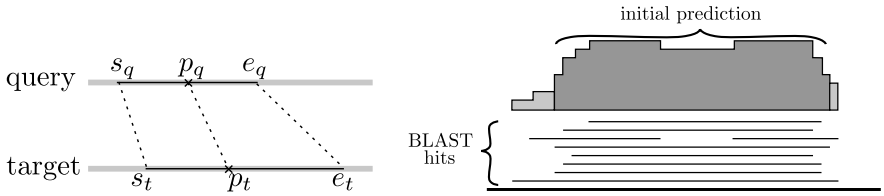

Figure 1: Left: A `BLAST` hit, given by the start and end of the hit in query (s_q, e_q) and target (s_t, e_t) sequence and the relative query position p_q and its corresponding position in the target p_t; right: generation of initial predictions: lines on the bottom depict the `BLAST` hits (the region in the target sequence), the shaded area shows the sum of `BLAST` hits per position; parts removed due to the cutoff are shown in light gray.

query position at target position p is defined as the relative query positions of p averaged over all hits that cover p.

Hits are treated separately for each of the six possible reading frames with respect to the target sequence. We reject `BLASTX` hits with a quality less than 2.0 to keep the number of spurious hits at bay. For each query gene and reading frame, the hits are aggregated separately to obtain *predictions* as consecutive stretches of positions that have a quality value of at least 50 % of the maximum value over all positions for the currently considered reading frame. Each prediction is represented by its *start* and *end* in target and query (i.e. the first and last positions of the prediction in the target sequence and the corresponding average relative query positions) and its *quality* given by the sum of the quality values at the included positions.

2.1.3. Overlapping predictions

Since the predictions are obtained separately for each gene and reading frame we have to expect conflicts. We employ a greedy strategy for conflict resolution, processing the predictions in the order of decreasing quality values. If one of the predictions overlaps more than 20 % with a prediction that has already been processed then they are either clipped, if they are predictions of the same

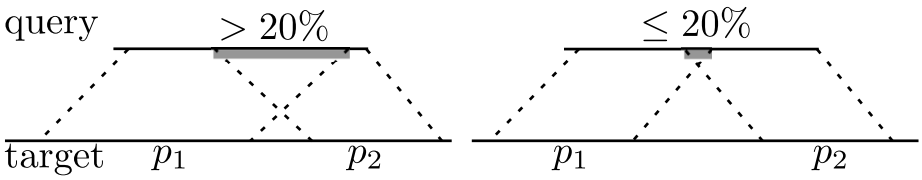

Figure 2: Illustration of the decision criterion used for determining if two predictions p_1 and p_2 are copies or parts of the same transcript; the shaded area is the common part of the query covered by both prediction; left: p_1 and p_2 are duplicates; right: p_1 and p_2 constitute fragments of the same transcript.

gene whose quality values differ by a factor of less than 10, or otherwise it is discarded entirely. The amount of overlap is measured as the fraction of the shorter sequence that is covered by the longer one. Overlapping predictions of the same gene are clipped by simultaneously modifying the start and end position in 1 nt steps, i.e. the start (end) position of the right (left) prediction is increased (decreased) as long as the quality values of the respective positions are better than those of the left (right) prediction.

2.1.4. Discrimination of duplicates and gene fragments

The resulting set of predictions for each gene is then checked for duplicates and the presence of multiple parts belonging to the same transcript. The latter occurs for instance in the case of frameshifts or splicing. `MITOS` assumes that two predictions are parts of the same transcript if their respective query ranges (i.e. the ranges given by the relative query positions of start and stop) overlap by at most 20 % and their quality differs by less than a factor of 10; otherwise they are treated as parts of different transcripts, i.e. as paralogous copies (see Figure 2). Again, a greedy strategy is used to iteratively add additional parts to split transcripts in the order of decreasing quality values. Finally, fragments within a split transcript are arranged according to the order of start positions in the query.

2.1.5. Improvement of start and end position

The local `BLAST` alignments are often inaccurate at their ends. `MITOS` thus scans the vicinity of each prediction for the presence of in-frame start and stop codons. `MITOS` investigates up to six amino acids upstream and downstream. Extensive testing showed that this yields very good results. If a stop codon is found to be within the region examined for a start codon, then the search for a start codon is limited to the region downstream of this stop codon.

2.2. Non-coding RNA annotation

Both tRNAs and rRNAs are highly structured, with large parts of the molecules exhibiting strong conservation in their base pairing patterns. Their primary sequence, however, shows high levels of variability. Therefore `MITOS` employs covariance models capturing the similarity of primary sequence as well as secondary structure for the identification of tRNA and rRNA coding genes. For a plausible structural annotation we use `Infernal` 1.0.2 (Nawrocki et al., 2009) in "glocal" search mode (i.e. global with respect to the model and local for the target sequence) and calibrated covariance models so that E- and p-values are computed. The generation of the covariance models for mitochondrial tRNAs and rRNAs and specific details of the search procedure is described in the following.

2.2.1. tRNA annotation

For the annotation of the tRNAs we use the strategy presented in Jühling et al. (2011). In brief, structure annotated covariance models were created for each mitochondrial tRNA gene in an iterative process improving the prediction rates of the models with every step.

2.2.2. rRNA annotation

The covariance models of the two rRNAs were constructed by a strategy similar to the one applied to the tRNAs. Initial covariance models based on manually curated and structurally annotated alignments of well-known metazoan rRNA sequences from the European Ribosomal RNA data base (Wuyts

et al., 2004) were created for each of the two rRNA genes. To include more sequence variation, the models were enhanced by adding all rRNA sequences annotated in the metazoan mitogenomes included in `RefSeq` 39 to the initial alignments. The resulting alignments were extensively manually curated. This was necessary in particular because the ends of most of the rRNA genes (almost 80 %) in `RefSeq` mitogenomes are determined by the flanking genes. After realigning the cleaned sequences with `Infernal`, we selected a seed alignment from which very similar sequences were omitted to avoid overfitting. To this end, an auxiliary graph representing each of the sequences as a node is used. Nodes are connect by an edge if their corresponding sequences differ by less than 5%, i.e. 47 nt and 81 nt for the *rrnS* and *rrnL*, respectively. From this graph we iteratively delete the neighbours of the vertex with the largest degree, until no edges are left. From the resulting seed alignments the final *rrnS* and *rrnL* rRNA covariance models were built.

`MITOS` falls back to searching rRNAs with `Infernal`'s local search mode if a "glocal" search remains unsuccessful.

2.3. Final annotation procedure

After candidates for protein, tRNA, and rRNA encoding genes have been determined according to the methods described above possible conflicts between these sets of predictions need to be resolved in order to obtain the final annotation. Therefore from each pair of genes of different type (protein, tRNA, or rRNA) that overlap by more than 35 nt one is removed. Because tRNAs often are adjacent to an rRNA, and rRNAs were observed in some cases to be annotated too long, we allow tRNAs to be overlapped to a larger extent by rRNAs as long as the tRNA is not included in the rRNA.

In a first round the best candidate is chosen from each gene, prioritising proteins, tRNAs (with E-value $\leq 10^{-3}$), and finally rRNAs. This is motivated by the fact that metazoan mitochondrial genomes usually have a single copy of each gene. In case no "glocal" rRNA hit is found that is compatible with the other predictions local rRNA hits are annotated. Thus, the method is able

to annotate also the rare cases of fragmented rRNAs and rRNAs with highly diverged secondary structure not covered by the complete model.

In a second round potential gene copies are determined. Therefore, the remaining candidate genes are added to the final annotation if they fulfil the overlap constraints. The order of the addition prioritises candidates with a large quotient of the quality and the quality of the best representative per gene (for non-coding RNAs the reciprocal of the E-value is used instead of the quality).

In a final step, local rRNA hits, if any, are merged if their respective position is in agreement with the location of the hits in the query covariance models.

2.4. Nomenclature

For the assignment of gene names we follow the guideline suggested by Boore (2006) (Supplemental Table 2). The tRNA-encoding genes are named in lower case with the one letter code for the corresponding amino acid and the anticodon appended in parentheses, e.g. *trnF*(gaa) for phenylalanine tRNA with anticodon gaa. Likewise, protein-coding genes and ribosomal RNAs are named in lower case with the ribosomal subunit indicated by a single upper case letter, i.e. *rrnS* and *rrnL* for small and large subunit ribosomal RNA, respectively. The Serine and Leucine tRNAs are distinguished by the recognised codon: *trnS1* for agn or agy, *trnS2* for ucn, *trnL1* for cun, and *trnL2* for uur (Boore, 2001). The protein-coding genes are named in lower case but otherwise according to the human gene nomenclature (Wain et al., 2002). In case of gene copies, the gene name is followed by a hyphen and an Arabic numerical, starting from 0 (e.g. *cox1*-0 for the first copy and *cox1*-1 for the second copy of a duplicated *cox1* gene). Gene parts are indicated by the addition of an underscore and a lower case Latin letter, starting from "a" for the most 5' part of the gene (e.g. *cox*1_*c* for the 3rd part of the *cox1* transcript).

2.5. Data sets

For an evaluation, the `MITOS` predictions have been determined for all 1878 mitogenome sequences contained in `RefSeq` 39 and all 203 mitogenome se-

quences of RefSeq 44, which were added to the RefSeq collection since release 39. The MITOS predictions have been compared to: (i) the annotations of the 1878 metazoan mitogenomes in RefSeq 39, (ii) the annotations of 203 mitogenomes, which were newly added in RefSeq 44, and (iii) the annotations for the mitogenome sequences of RefSeq 39 given in MitoZoa 9.1 (which excludes the four Placozoa and *Hemidactylus frenatus* (NC_012902)).

In order to develop means to quantify the quality of the prediction, in particular the quality values of the proteins, the predictions including all conflicting ones have been determined for (i) the 203 sequences of the mitogenomes newly added to RefSeq 44 and (ii) permutations of these genomic sequences (ten dinucleotide and ten trinucleotide frequency preserving shufflings).

3. Results and Discussion

In order to assess the quality of the MITOS predictions we employed our pipeline for a *de novo* annotation of RefSeq 39. By showing that the default parameters chosen in MITOS are suitable for the entire metazoan data set, we hope to relieve the user of the tedious empirical work of choosing appropriate cutoff values and parameters.

In the following we will refer to RefSeq or MITOS annotations as "genes".

3.1. Overlapping predictions

Since neighbouring genes can overlap in mitochondrial genomes allowing overlaps is a necessity for mitochondrial genome annotation. Furthermore, determining exact gene boundaries is often difficult, e.g. because of incomplete stop codons. On the other hand, we do not expect large overlaps between adjacent genes. The overlap of gene predictions thus has been evaluated in detail (see also Supplemental Table 3 and Supplemental Figure 4). Out of the 69 917 neighbouring pairs of MITOS predictions 17 152 ($\approx 25\%$) overlap. The situation is qualitatively similar in RefSeq 39: 8540 out of 68 874 annotations overlap. The additional overlaps among the MITOS predictions are typically small (average

length 7.23 nt). The average overlap involving tRNAs (1.68 nt for tRNA-tRNA overlaps) is smaller than the overlaps found for proteins or rRNAs. The maximum overlap in the data set is 68 nt found for a tRNA and rRNA. Most likely, the *rrnL* prediction is too long in this case.

3.2. Fragmented predictions

The protein search implemented in `MITOS` allows the identification of gene copies and fragmented genes including frameshifts and cases where the coding sequence is not contiguous. Split genes are for instance caused by the insertion of self-splicing introns, which are not infrequent in certain metazoan lineages, see Beagley et al. (1996); Dellaporta et al. (2006); Wang and Lavrov (2008); Bernt et al. (2012b). `MITOS` predicts 220 protein coding genes in fragments, most frequently *nad3* (109), *nad5* (38), and *cox1* (20). These three cases are discussed in more detail, see also Supplemental Tables 5, 6, and 7.

Most of the fragmented *nad3* genes are found in mitogenomes of Aves (81) and Testudines (25), i.e. taxa where frameshifts have been reported frequently for *nad3* (Mindell et al., 1998). Compared with the *nad3* frameshifts reported in `RefSeq` 39 for birds and turtles `MITOS` predicts all but one (NC_003712) and adds three previously unannotated ones (NC_001947, NC_009509, and NC_011516).

Most of the cases of fragmented *nad5* predictions coincide with known cases where these genes are separated on multiple exons, e.g. in Placozoa and Cnidaria (Bernt et al., 2012b, in this special issue). In 31 of the 38 cases where *nad5* was predicted in fragments compatible annotations are found in `RefSeq` 39, i.e. same fragments are reported although there may be differences in the precise start and stop locations. Only one of these cases was reported as frameshift in `RefSeq` and all other cases correspond to known cases of fragmented genes in Cnidaria (27 cases) and Placozoa (3 cases).

All but two of the 20 mitogenomes where `MITOS` predicts *cox1* in multiple parts are also annotated as multiple fragments in `RefSeq` 39, but in a few cases usually very short fragments are missed or additional fragments are predicted by `MITOS`. Again, most of these cases affect Cnidaria (8) and Placozoa (4). The

case of *Hexamermis agrotis* (NC_008828) is particularly interesting, since MITOS predicts three fragmented copies of *cox1*. All but two of these fragments are predicted within unannotated regions and the quality values ($> 10^5$) suggest that they might be pseudogenes (see also Supplementary Figure 8). Interestingly, also duplicates of other genes are annotated in RefSeq 39 or predicted by MITOS in this mitogenome (see also Yatawara et al., 2010).

MITOS predicts a similar number of fragmented genes as are annotated in RefSeq 39 also for other proteins (see Supplementary Table 5). This demonstrates that MITOS is able to annotate even complicated cases including frameshifts and genes with fragmented coding sequences.

MITOS also predicts several fragmented rRNAs. The local search mode, employed in cases where the "glocal" mode fails, leads to 161 and 22 fragmented predictions for *rrnL* and *rrnS*, respectively. This can be caused by known fragmented genes in Metazoa (e.g. Dellaporta et al., 2006) as well as highly variable, inserted or deleted domains that are overlooked because no homologous sequence is contained in the seed alignment.

3.3. MITOS vs. RefSeq 39

For each of the genes predicted by MITOS we define the corresponding RefSeq 39 annotation as the gene that shares the most positions with the MITOS prediction provided the MITOS prediction shares at least 75 % of its position with the corresponding RefSeq annotation. Pairs of MITOS predictions and corresponding RefSeq annotations identified by this definition are differentiated in the following classes: they are *equal* if both annotate the same gene and are located on the same strand. If they only annotate the same gene but are found on the opposite strand, the gene is marked as having a *strand difference*. Corresponding pairs annotating different genes are marked as *different*. We consider MITOS predictions without corresponding RefSeq annotation to be *false positives* (FP) and RefSeq annotations without corresponding MITOS prediction to be *false negatives* (FN).

Table 1: Comparison of the MITOS predictions with the annotations found in RefSeq 39, RefSeq 44 without RefSeq 39, and MitoZoa 9.1; given are the number of cases and fraction in parentheses where MITOS predicts: the same gene on the same strand (equal), the same gene on the opposite strand ($\Delta\pm$), a false positive (FP), a false negative (FN), a gene with different name (different); see text for details; a gene wise overview is given in Supplemental Tables 14, 18, and 21.

		equal	$\Delta\pm$	FN	FP	different
RefSeq 39	Protein	24 533 (0.97)	0 (0.00)	107 (0.00)	493 (0.02)	226 (0.01)
	rRNA	4087 (0.96)	24 (0.01)	57 (0.01)	84 (0.02)	14 (0.00)
	tRNA	39 000 (0.95)	355 (0.01)	698 (0.02)	367 (0.01)	709 (0.02)
RefSeq 44 without RefSeq 39	Protein	2632 (0.96)	2 (0.00)	2 (0.00)	69 (0.03)	27 (0.01)
	rRNA	430 (0.97)	1 (0.00)	11 (0.02)	1 (0.00)	1 (0.00)
	tRNA	4188 (0.93)	20 (0.00)	71 (0.02)	66 (0.01)	163 (0.04)
MitoZoa 9.1	Protein	24 515 (0.97)	1 (0.00)	40 (0.00)	429 (0.02)	243 (0.01)
	rRNA	4114 (0.97)	3 (0.00)	38 (0.01)	63 (0.01)	6 (0.00)
	tRNA	39 674 (0.97)	5 (0.00)	557 (0.01)	283 (0.01)	352 (0.01)

Table 1 shows that the results obtained from MITOS are in excellent agreement with the annotations found in RefSeq 39. Nevertheless a number of discrepancies are obvious.

Strand differences to RefSeq annotations are found for 355 tRNAs and 24 rRNAs, but none were found for protein coding genes. In nearly all of these cases (330 of 379) MITOS predicts the respective gene on the "minus" strand, whereas the RefSeq annotation is located on the "plus" strand. Since RefSeq entries are per default on the "plus" strand and genes on the "minus" strand need to be marked with a "complement" statement, this discrepancy can be explained simply by annotation errors in RefSeq resulting from forgetting the "complement" statement.

A large fraction of the MITOS predictions that are classified as different refer to Serine or Leucine tRNAs for which either the distinction between the two anticodon types is not annotated (166 cases) or interchanged (*trnL1*/*trnL2* in 78 cases and *trnS1*/*trnS2* in 142 cases). Despite the fact that there is no standard nomenclature for the Leucine and Serine tRNAs, the naming scheme given in Boore (2001) seems to be generally accepted. That is, these cases are inconsistencies of RefSeq. We emphasise that a consistent naming is indispensable for many studies and gene arrangement analysis in particular.

A closer inspection shows that quite a few of our "false positive" and "false negative" predictions are in fact false negatives and false positives in the RefSeq annotation. This is discussed in the following.

Many of the MITOS predictions that are classified as false positives (resp. different gene) are very well supported (quality $> 10^6$ or E-value $< 10^{-4}$). There is little doubt that these MITOS-predictions, comprising 64 (resp. 58) protein coding genes, 23 (resp. 10) rRNA, and 193 (resp. 211) tRNA, are correct. This list excludes the *trnL* and *trnS* cases discussed above. See Supplementary Tables 16 and 17 for details. The most prevalent source of false positives is MITOS' strategy to accept very low scoring predictions as long as they are not conflicting with high scoring predictions. This leads to predictions in the large control region in 246 cases, but in most cases these predictions can be rejected

based on their score, i.e. only 48 of these have a noteworthy quality ($> 10^4$) or E-value $< 10^{-4}$.

A substantial fraction of "false negative" `MITOS`-predictions is explained by genes for which RefSeq annotates a much too long region. This accounts for 147 false negatives and 150 annotations classified as different as well as 4 false positives. An additional set of false negatives is due to an overlap smaller than the 75% threshold. The false negative predictions have been analysed manually for the protein and rRNA coding genes (see Supplementary Table 15). We have identified 66 protein coding and 17 rRNA `MITOS` predictions that are actually correct. For the remaining false negatives of the protein coding genes (17 protein and 24 rRNA fragments <100 nt) no support could be found via `BLAST`. For the remaining false negative rRNA genes corresponding predictions are made by `Infernal` but could not be placed due to conflicts caused by overlap with other predictions.

Since the predictive power of the mt-tRNA covariance models (Jühling et al., 2011) employed by `MITOS` outperforms `tRNAscan-SE`, which is the standard tool for tRNA detection in mitogenomes, it is likely that `MITOS` tRNA results are an improvement of the `RefSeq` annotations. Therefore, the ∼700 tRNA annotations that have not been confirmed by `MITOS` and 330 previously unannotated tRNAs which are identified by our method have not been checked manually.

To conclude the comparison with `RefSeq` 39 we show that most predictions are much more accurate than the threshold of 75 % might suggests (see also Supplemental Tables 9, 10 and Supplemental Figures 11 and 13). The average fraction of positions of `MITOS` predictions that are shared with `RefSeq` 39 annotations exceeds 99 % for each type of gene. Conversely, the average percentage of the `RefSeq` 39 annotations that are shared by the `MITOS` predictions is larger than 99 % for tRNAs, larger than 95 % for protein coding genes and *rrnS*. The mean value is only 77 % for *rrnL*. One reason for this are *rrnL* genes predicted with the local search mode.

`MITOS` also predicts start and stop positions quite well in comparison to the `RefSeq` 39 annotations. For more than 64 % of the predictions start and

stop position of the prediction are identical and for nearly 80 % the difference is less than 5 nt. Note that differences of 1 nt may occur in cases where the boundary designation in RefSeq annotations does not follow the formal GenBank guidelines, i.e. start and stop are included and counting starts at 1. For the start and stop position MITOS also achieves better results for the tRNAs than for the proteins and rRNAs. The smaller coverage of RefSeq annotations as well as the inferior precision of the boundaries for the rRNA genes might be based on the fragmented prediction when the local search mode is applied. Furthermore, the start and stop positions of the RefSeq annotations are often chosen so that they touch the adjacent genes (Boore et al., 2005). Contrary to that MITOS adopts a more conservative strategy. This presents an additional reason for the lower coverage of RefSeq *rrnL* annotations by MITOS predictions. Large positional differences may also occur due to typing errors in RefSeq. More than two dozen putative cases where permutations of more than two digits explain the differences are given in Supplemental Table 12.

3.4. MITOS' results on new genomes

We have validated the predictive power of MITOS on all 203 genomes found in RefSeq 44 that were not included in RefSeq 39 (see Table 1). The large number of genes with a different type is again in most cases (116) caused by the Serine and Leucine tRNAs. The RefSeq annotation of 31 Leucine and 5 Serine tRNAs is interchanged when compared to the MITOS prediction and in 80 cases the sub-classification of these tRNAs is missing in RefSeq.

Two proteins were found to be located on a different strand, i.e. *nad3* and *atp8* of *Platevindex mortoni* (NC_013934). A closer inspection indicated that both are correctly annotated by MITOS (see Supplemental Figure 19). While *nad3* has been corrected in later RefSeq versions, *atp8* is still annotated on the wrong strand. We remark that RefSeq also annotates the *rrnS* of *P. mortoni* on the opposite strand.

Overall, the results show that MITOS also yields very high quality annotations of mitogenome sequences which have not contributed to the data used for the

annotation procedure or for the development of the covariance models.

3.5. MITOS vs. MitoZoa

MitoZoa aims to provide annotations of higher quality than RefSeq (Lupi et al., 2010). The results of the comparison of the MITOS predictions with the annotations provided in MitoZoa 9.1 for the RefSeq 39 data are shown in Table 1. Given that the modified annotations provided by MitoZoa are improvements the results clearly support the predictive power of MITOS. Strand differences are almost completely removed. The numbers of false negatives, false positives, and predictions with different names (except for protein coding genes) are clearly reduced.

3.6. Duplicated genes

MITOS treats putative duplicates of a gene differently than the best scoring copy, see Section 2.3. By adding putative paralogs greedily in the remaining gaps, MITOS is able to also annotate duplicated genes quite reliably. In the mitogenomes of Porifera and Placozoa several tRNAs, in particular *trnI*, *trnM*, and *trnR*, are reported as duplicates. With the exception of one copy of *trnA* in *Axinella corrugata* (NC_006894) all tRNAs that are reported as duplicates in RefSeq 39 are matched by a tRNA prediction of MITOS. In most cases the MITOS prediction is of the same type as the RefSeq annotation, but there are noteworthy exceptions. Most notably, the prediction of *trnM* instead of an annotated *trnI* (see Supplementary Material 22). For these cases the anticodon cau (*trnM*) is post-transcriptionally modified to gau (*trnI*) as reported in (Lavrov et al., 2005) for Porifera.

3.7. Limitations

The current implementation of MITOS does not explicitly account for a variety of lineage-specific deviations. Most importantly, no annotation is provided for additional genes as observed in Placozoa (Dellaporta et al., 2006) or the FORFs in Unionida (Breton et al., 2010). In some cases, protein coding genes

feature additional extensions, such as the 3' extension of *cox2* in male unionid mussels (Breton et al., 2010). Unless the extensions are well represented in the set of input proteins, they will not be included in the aggregation procedure, so that the ends of these genes will not be annotated correctly. The organisation of the `MITOS` pipeline can be adapted to such cases; this will require additional filtering rules for conflict resolution and has been left for future releases.

A potential problem may arise from the strategy employed by `MITOS` and all other currently available tRNA annotation tools to determine the identity of the genes based on their anticodon sequence. This strategy is reasonable but it might fail or be misleading in a few cases. One example is a post-transcriptional modification of the anticodon, e.g. the case reported in Section 3.6. Furthermore, gene recruitment might be misleading, i.e. when tRNAs change their identity by point mutations of the anticodon (Saks et al., 1998). This has been reported for metazoan mitogenomes (Rawlings et al., 2003) and in particular for Porifera (Lavrov and Lang, 2005; Wang and Lavrov, 2011), where gene recruitment seems to be more frequent than in other metazoan groups. Since for the Porifera all reported cases are annotated by their anticodon in `RefSeq` 39 as well as in `MitoZoa` no discrepancy to the `MITOS` annotation is found. In order to resolve this issue, we plan for future versions of `MITOS` that the name indicates the evolutionary origin, i.e. homology, while the current function of the tRNA will be indicated by the anticodon.

3.8. `MITOS` on randomised validation data sets and interpretation of protein quality values

`MITOS`' results on the permuted validation data sets show that scores of correct annotations are clearly distinguishable from random hits, except for *atp8* (see Supplementary data), which is the shortest protein coding gene found in mitogenomes (average length 161 nt in the 201 *atp8* genes in the positive validation set and 172 nt in `RefSeq` 39). The randomised data still yields 126 (dinucleotide shuffling) and 133 (trinucleotide shuffling) *atp8* genes with an average size of 122 nt and 118 nt, respectively. The low number of false positive hits for other

genes (for instance, MITOS never predicts *cox1* or *cox2* from shuffled sequences) in combination with the annotation strategy of MITOS, reduces the possibility of high scoring random hits in the usually densely packed mitogenomes.

Quality values differ strongly not only between the different genes but sometimes even within the same gene. Therefore, we do not apply a general cutoff value, in order to enable the detection of degenerated duplication fragments. This, in turn, could increase the possibility of false positive hits. We provide quality value distributions for a) the initial hits found for the 13 protein coding genes in the mitogenomes of RefSeq 39 and b) the di- and trinucleotide frequency preserving permutations of the complete mitogenome sequences. These plots enable the user to assess the protein annotations by MITOS, see Supplementary Figures 23 and 24.

3.9. MITOS web server

We have set up a web server that implements the presented pipeline and allows the *de novo* annotation of whole metazoan mitochondrial genomes. A major focus in the development of MITOS is to minimise the necessary amount of manual interaction by the user – but an advanced mode is provided where parameters can be modified. After uploading a mitogenome sequence in FASTA format, the user simply has to select the appropriate genomic translation code. If all computing resources are already in use, the job will be put in a queue. Queued jobs can be cancelled. Once the genome is being processed the user will be redirected to a web page where the final results can be found and a notification including a link to the final results will be sent to the users email-address, if provided. The results page gives a tabular overview of the predictions, a visual representation, that is also available for download, and links to a variety of commonly used file formats containing the annotation: BED, GFF, FASTA, and Sequin format. Furthermore all raw data of BLAST and Infernal, a graphical representation of the structure of the ncRNAs predicted by the covariance models, and a file containing the gene order are available (lacking the anticodon information for tRNA-encoding genes). The MITOS web server features a help

page and a tutorial that guides the user through the annotation of an example sequence.

With the current soft- and hardware infrastructure (Intel® Xeon™ CPU 3.20 GHz) the annotation of a single mitogenome needs about 1.5 h. Note that most of the time is spent on searching rRNAs with `Infernal`. To keep waiting times short the `MITOS` web server can process several requests in parallel. For the newly added genomes in `RefSeq` 44 the average time required to annotate one mitogenome was 1.3 h using an AMD Opteron™ Processor (2.6 GHz) and 2 GB of main memory.

4. Conclusion

`MITOS` is an automated pipeline that tackles the problem of reliable metazoan mitochondrial genome annotation, using state of the art methods. Protein coding genes are annotated by means of a sophisticated aggregation procedure based on `BLAST` searches, which allows for the detection of frameshifts, duplication events, and split genes. Structural conservation is utilised for non-coding RNA annotation by employing novel covariance models. `MITOS` allows for a systematic error screening, the standardisation of gene name and gene boundary designation, anticodon labelling of tRNAs, and provides the means for the assessment of the validity of a gene assignment. Using `MITOS` for *de novo* annotation yields high-quality data for a variety of subsequent analyses, such as genome rearrangement studies and phylogenetic analyses.

5. Acknowledgements

We thank the Center for Information Services and High Performance Computing (ZIH) of the TU Dresden (`http://tu-dresden.de/zih/`) for providing computational facilities.

This work was supported by the Deutsche Forschungsgemeinschaft [SPP-1174 - *Deep Metazoan Phylogeny* projects STA 850/2 and STA 850/3-2]; Centre

National de la Recherche Scientifique (CNRS); Université de Strasbourg; Association Française contre les Myopathies (MNM1 2009); ANR MITOMOT (ANR-09-BLAN-0091-01); French-German PROCOPE program (DAAD D/0628236, EGIDE PHC 14770PJ); French-German University (DFH-UFA, Cotutelle de thèse CT-08-10); doctoral fellowship of the German Academic Exchange Service (DAAD D/10/43622); bridge scholarship of the Collège Doctoral Européen (CDE), Université de Strasbourg; and the an NSC Taiwan Fellowship of the Alexander von Humboldt Foundation.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., Oct 1990. Basic local alignment search tool. J. Mol. Biol. 215 (3), 403–410.

Beagley, C. T., Okada, N. A., Wolstenholme, D. R., 1996. Two mitochondrial group I introns in a metazoan, the sea anemone *Metridium senile*: one intron contains genes for subunits 1 and 3 of NADH dehydrogenase. Proc Natl Acad Sci USA 93, 5619–5623.

Bernt, M., Braband, A., Middendorf, M., Misof, B., Rota-Stabelli, O., Stadler, P. F., 2012a. Bioinformatics methods for the comparative analysis of metazoan mitochondrial genome sequences. Mol Phyl Evol.

Bernt, M., Braband, A., Schierwater, B., Stadler, P. F., 2012b. Genetic aspects of mitochondrial genome evolution. Mol Phyl Evol.

Boore, J. L., 2001. Mitochondrial genome rearrangement guide. Version 6.1.

Boore, J. L., 2006. Requirements and standards for organelle genome databases. OMICS 10, 119–126.

Boore, J. L., R., M. J., Medina, M., 2005. Sequencing and comparing whole mitochondrial genomes of animals. Method Enzymol 395, 311–348.

Breton, S., Stewart, D. T., Shepardson, S., Trdan, R. J., Bogan, A. E., Chapman, E. G., Ruminas, A. J., Piontkivska, H., Hoeh, W. R., 2010. Novel protein genes in animal mtDNA: a new sex determination system in freshwater mussels (Bivalvia: Unionoida)? Mol Biol Evol 28, 1645–1659.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M. J. L., 2009. Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423.

Dellaporta, S. L., Xu, A., Sagasser, S., Jakob, W., Moreno, M. A., Buss, L. W., Schierwater, B., 2006. Mitochondrial genome of *Trichoplax adhaerens* supports placozoa as the basal lower metazoan phylum. Proc Natl Acad Sci USA 103, 8751–8756.

Eddy, S. R., Durbin, R., 1994. RNA sequence analysis using covariance models. Nucleic Acids Res 22, 2079–2088.

Feijao, P. C., Neiva, L. S., de Azeredo-Espin, A. M., Lessinger, A. C., 2006. AMiGA: the arthropodan mitochondrial genomes accessible database. Bioinformatics 22, 902–903.

Fujita, M. K., Boore, J. L., Moritz, C., 2007. Multiple origins and rapid evolution of duplicated mitochondrial genes in parthenogenetic geckos (*Heteronotia binoei*; Squamata, Gekkonidae). Mol Biol Evol 24, 2775–2786.

Jameson, D., Gibson, A. P., Hudelot, C., Higgs, P. G., 2003. OGRe: a relational database for comparative analysis of mitochondrial genomes. Nucleic Acids Res 31, 202–206.

Jühling, F., Pütz, J., Bernt, M., Donath, A., Middendorf, M., Florentz, C., Stadler, P. F., 2011. Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. Nucleic Acids Res.

Laslett, D., Canback, B., 2008. ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics 24, 172–175.

Lavrov, D. V., Forget, L., Kelly, M., Lang, B. F., 2005. Mitochondrial genomes of two demosponges provide insights into an early stage of animal evolution. Mol Biol Evol 22 (5), 1231–1239.

Lavrov, D. V., Lang, B. F., 2005. Transfer RNA gene recruitment in mitochondrial DNA. Trends in Genetics 21 (3), 129 – 133.

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

Lupi, R., de Meo, P. D., Picardi, E., D'Antonio, M., Paoletti, D., Castrignano, T., Pesole, G., Gissi, C., 2010. MitoZoa: a curated mitochondrial genome database of metazoans for comparative genomics studies. Mitochondrion 10, 192–199.

Mindell, D. P., Sorenson, M. D., Dimcheff, D. E., 1998. An extra nucleotide is not translated in mitochondrial ND3 of some birds and turtles. Mol Biol Evol 15, 1568–1571.

Nawrocki, E. P., Kolbe, D. L., Eddy, S. R., 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics 25, 1335–1337.

Pruitt, K. D., Tatusova, T., Maglott, D. R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35, D61–65.

Rawlings, T. A., Collins, T. M., Bieler, R., 2003. Changing identities: tRNA duplication and remolding within animal mitochondrial genomes. Proc Natl Acad Sci USA 100 (26), 15700–15705.

Saks, M. E., Sampson, J. R., Abelson, J., 1998. Evolution of a transfer RNA gene through a point mutation in the anticodon. Science 279 (5357), 1665–1670.

San Mauro, D., Gower, D. J., Zardoya, R., Wilkinson, M., 2006. A hotspot of gene order rearrangement by tandem duplication and random loss in the vertebrate mitochondrial genome. Mol Biol Evol 23, 227–234.

Sheffield, N. C., Hiatt, K. D., Valentine, M. C., Song, H., Whiting, M. F., 2010. Mitochondrial genomics in Orthoptera using MOSAS. Mitochondr DNA 21, 87–104.

Wain, H. M., Bruford, E. A., Lovering, R. C., Lush, M. J., Wright, M. W., Povey, S., 2002. Guidelines for human gene nomenclature. Genomics 79, 464–470.

Wang, X., Lavrov, D. V., 2008. Seventeen new complete mtDNA sequences reveal extensive mitochondrial genome evolution within the Demospongiae. PLoS ONE 3, e2723.

Wang, X., Lavrov, D. V., 2011. Gene recruitment a common mechanism in the evolution of transfer rna gene families. Gene 475 (1), 22 – 29.

Wolstenholme, D. R., 1992. Animal mitochondrial DNA: Structure and evolution. Int Rev Cytol 141, 173–216.

Wuyts, J., Perriere, G., Van De Peer, Y., 2004. The European ribosomal RNA database. Nucleic Acids Res 32, D101–103.

Wyman, S. K., Jansen, R. K., Boore, J. L., 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20, 3252–3255.

Yatawara, L., Wickramasinghe, S., Rajapakse, R. P. V. J., Agatsuma, T., 2010. The complete mitochondrial genome of *Setaria digitata* (Nematoda: Filarioidea): Mitochondrial gene content, arrangement and composition compared with other nematodes. Molecular and Biochemical Parasitology 173 (1), 32–38.