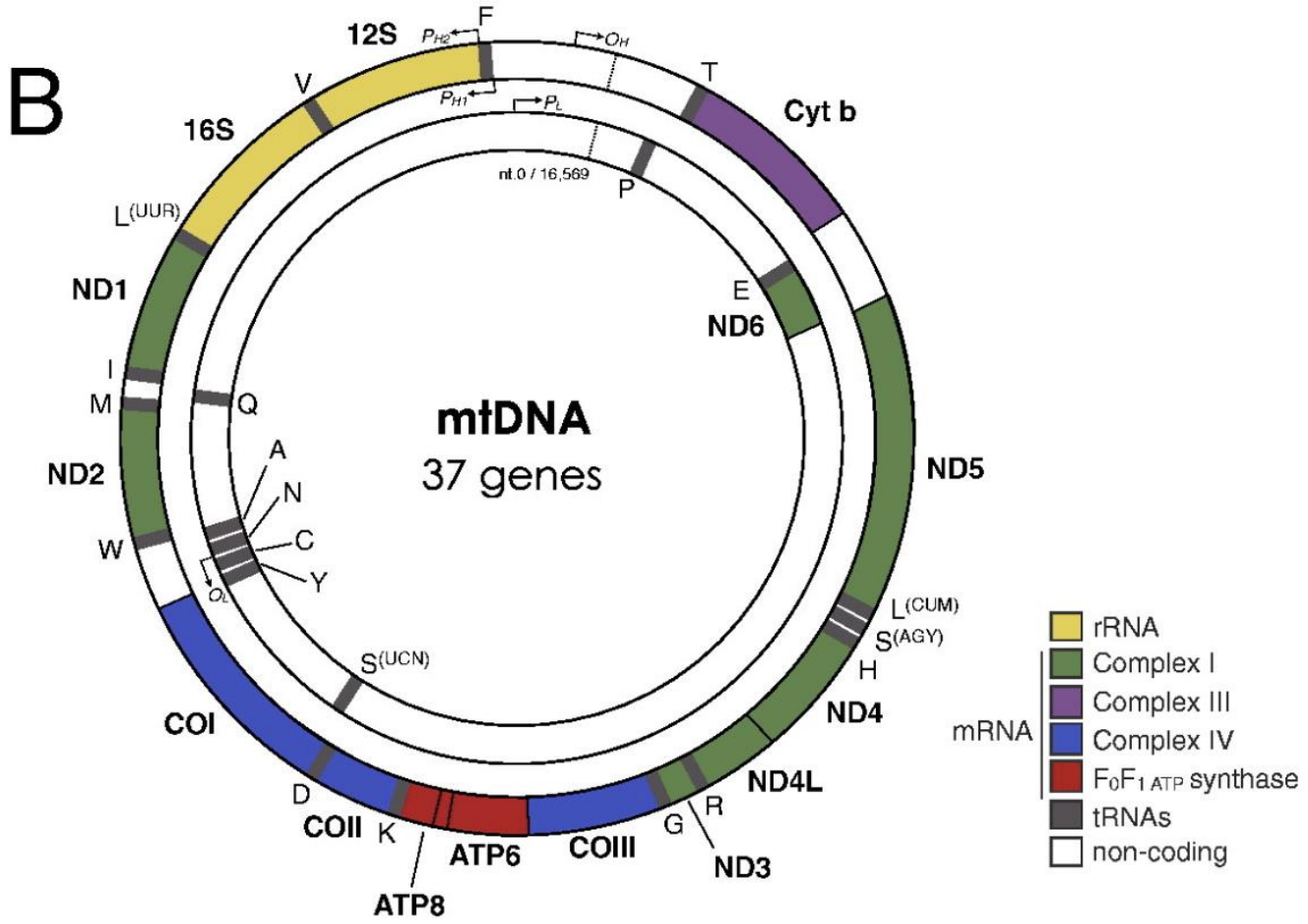
A Brief Primer on Reference Mitochondrial Genomes

Dan MacGuigan
NOAA National Systematics Lab

What is a mitochondrial genome?

- Usually (but not always) a single circular DNA molecule
- Usually (but not always) contains only genes with few non-coding regions
- Usually (but not always) maternal inheritance with no recombination
- Contains many important DNA barcoding genes: *CO1, 12s, 16s, CytB, MutS*…



human mitochondrial genome

Picard et al. 2016, Mitochondrion

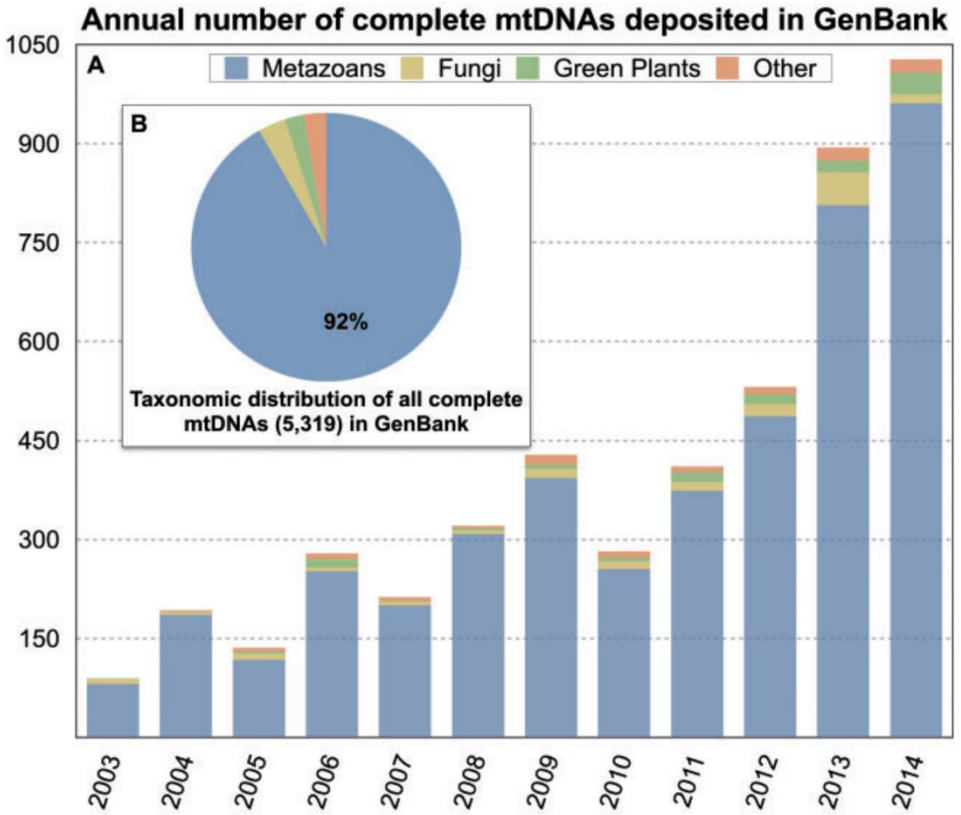
Do we need more mitochondrial genomes?

The past, present and future of mitochondrial genomics: have we sequenced enough mtDNAs?

David Roy Smith

Briefings in Functional Genomics, 15(1), 2016, 47–54

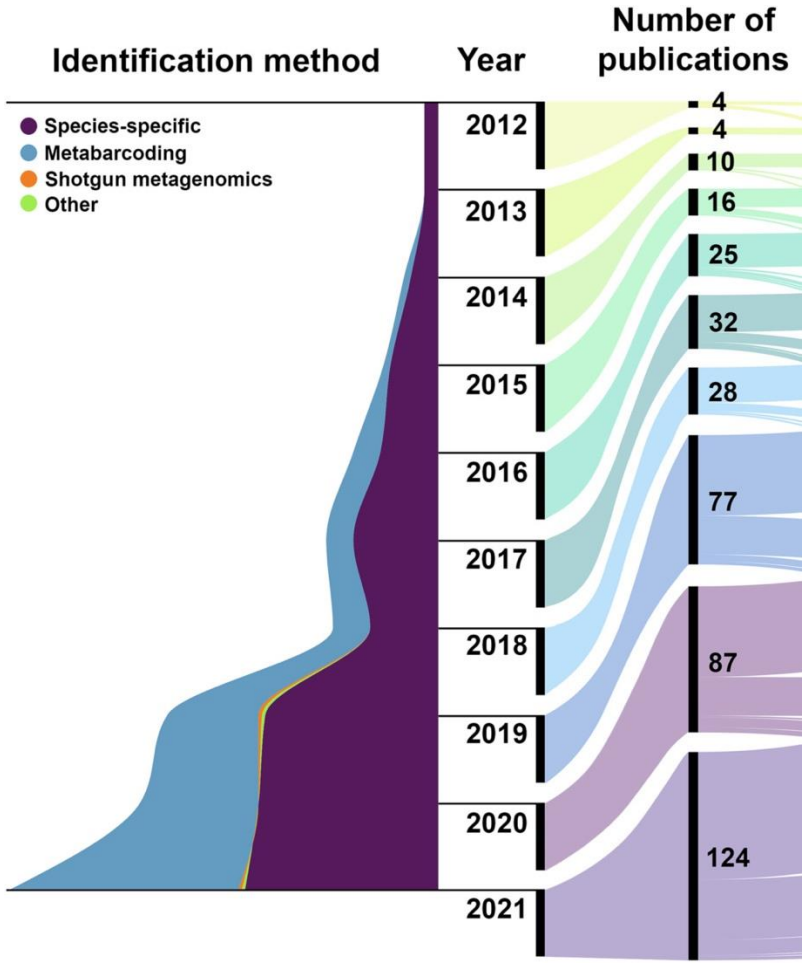
doi: 10.1093/bfgp/elv027
Advance Access Publication Date: 27 June 2015
Review paper



"For many groups…mtDNAs are so well sampled that newly published genomes are ***no longer contributing to the progression of science****…"*

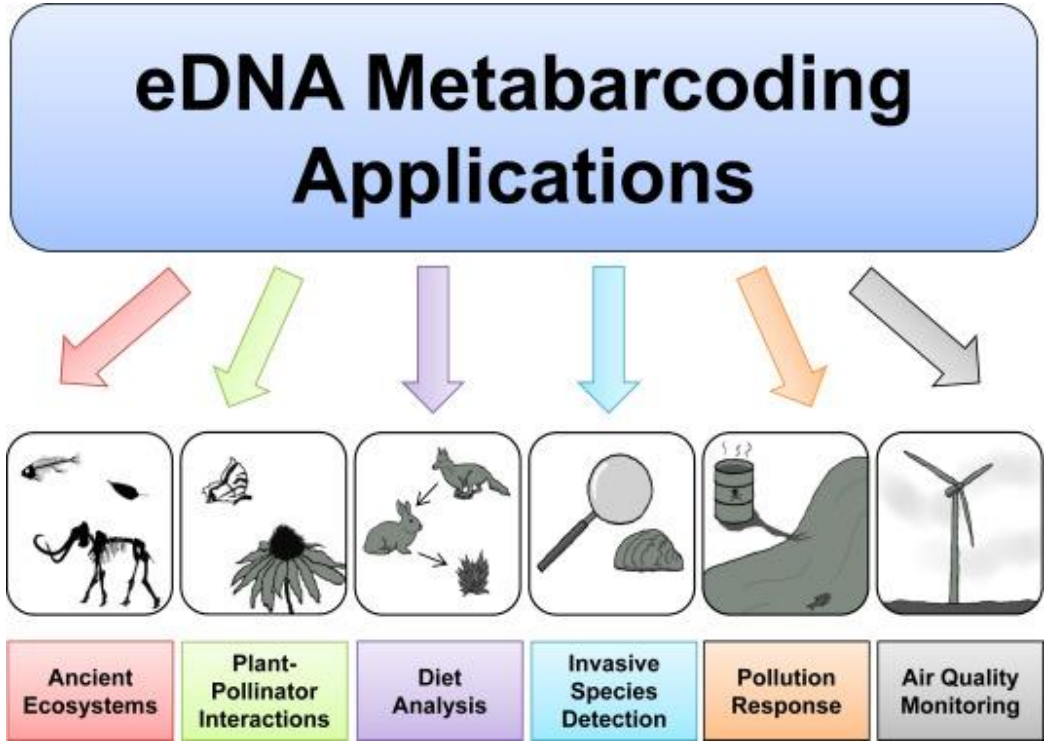
I would argue that, even a decade later, this is simply not true.

The rise of eDNA metabarcoding makes mitochondrial reference libraries more important than ever



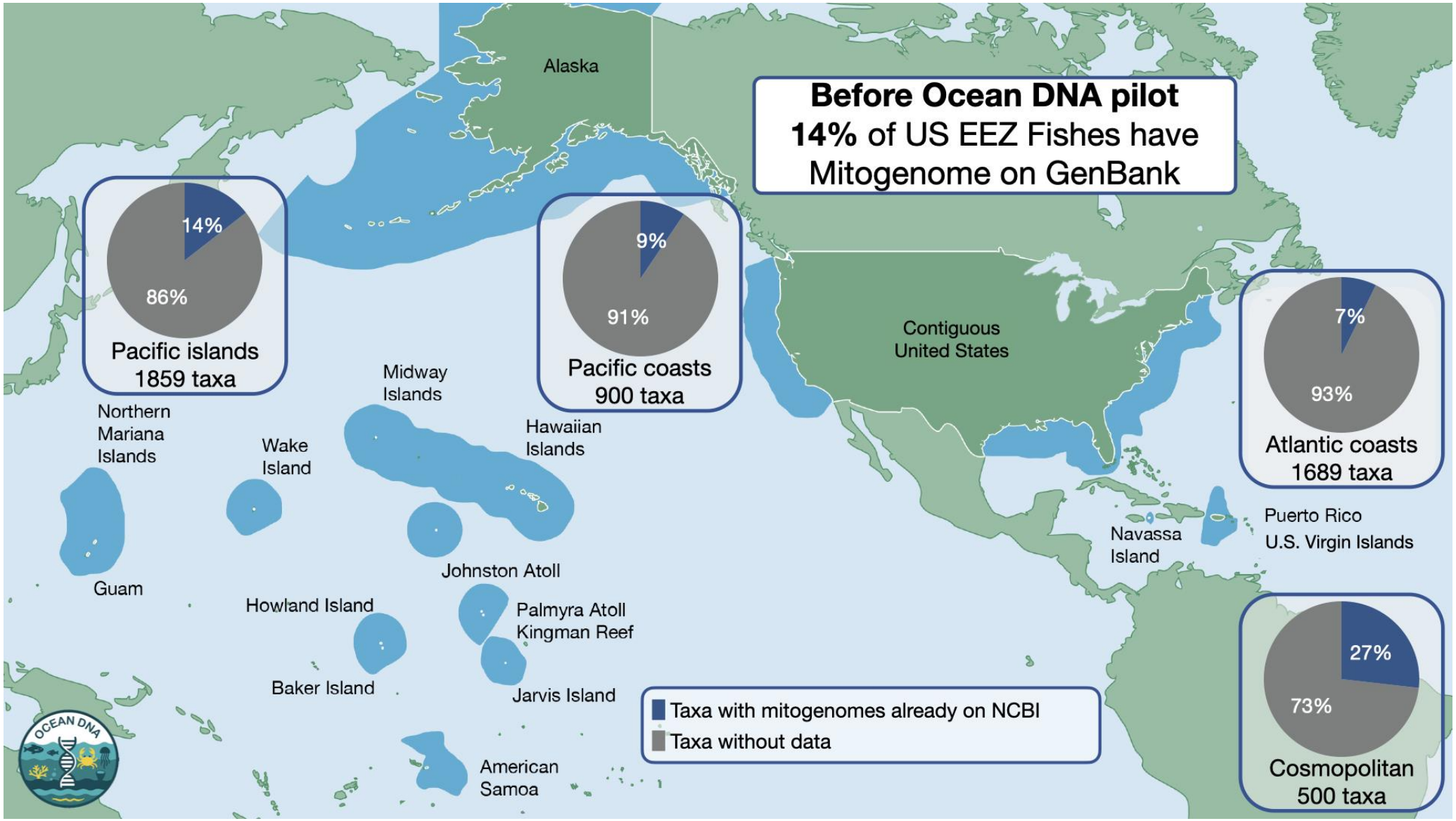
Aquatic eDNA publications

Takahashi et al. 2023, Science of the Total Environment



Ruppert et al. 2019, Global Ecology and Conservation

Much of biodiversity is still underrepresented, even for well-studied taxa like fishes



The importance of reference-quality mitochondrial genomes

- Problems with mitogenomes in public repositories (like GenBank)
  - Poor quality annotations
  - Assembly errors
  - Misidentification
  - Lack of specimen voucher information
- An ideal reference-quality mitochondrial genome should have none of those issues

Do we need more mitochondrial genomes?

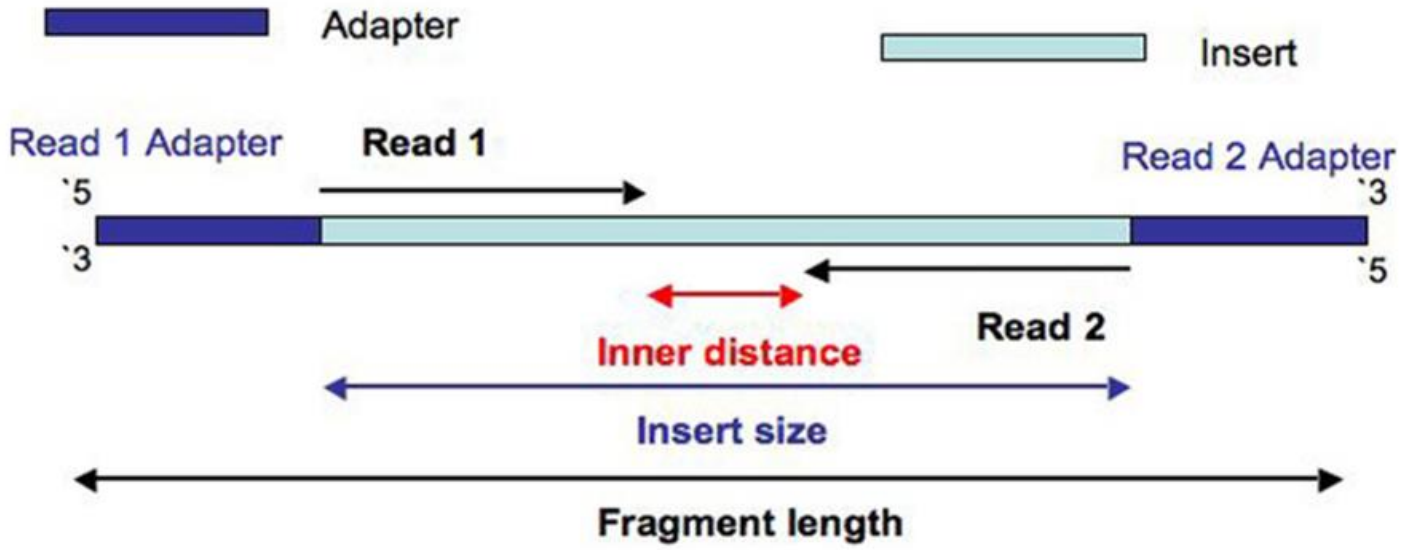
YES!
Especially voucher-based, reference-quality mitogenomes

How to build a reference-quality mitogenome

- Step 1: sequencing
- Step 2: sequence data QC and filtering
- Step 3: assembly
- Step 4: annotation
- Step 5: curation
- Step 6: submission to GenBank

Step 1 – sequencing

- Illumina short-read sequencing is the most common strategy



- Long-read sequencing (Oxford Nanopore, PacBio) may be the future

Step 2 – sequence data QC and filtering

- Filter sequence data to remove
  - Multiplexing indexes/barcodes
  - Adapter contamination
  - Low quality regions and reads
  - And possibly more
- Popular software:
  - FastQC and MultiQC for basic quality checks – Hydra script here
  - fastp
  - cutadapt
  - Trimmomatic
  - Trim Galore
  - And many more

Step 3 – assembly

- Reference-based assembly
  - “Map to reference” method in Geneious
  - MITObim (also has a de novo mode)
- De novo assembly specifically for mitogenomes
  - GetOrganelle
  - MitoZ
  - MitoFinder
  - mtGrasp
  - MitoHiFi - specifically for PacBio long reads
- Or any general de novo genome assembler, followed by manual identification of the mitochondrial contig(s)

Step 4 – annotation

- Homology-based annotation
  - BLAST
  - MitoFinder
- De novo annotation
  - MITOS2
  - DeGeCI
- tRNA annotation
  - MiTFi
  - tRNAscan-SE 2.0
  - ARWEN
- Taxon-specific annotation
  - MitoAnnotator

"The most important thing is…annotating [the mitogenome] correctly."

Smith 2016, Briefings in Functional Genomics

Step 5 – curation

- Most annotation methods produce decent gene models
- But manual curation is often still required
- To my knowledge, there is no software designed specifically for manual curation of mitochondrial genome annotations

"The most important thing is…annotating [the mitogenome] correctly."

Smith 2016, Briefings in Functional Genomics

Step 6 – submission to GenBank

- NCBI Submission Portal
- Requires two files: the mitogenome assembly (FASTA) and the annotation information (feature table)
- Every submission is manually reviewed by GenBank staff
  - No clear rules on what makes a submission acceptable
  - A singe problematic sample may lead to rejection of the entire submission batch
  - Often becomes an iterative process of correcting errors and resubmitting

*"The most important thing is **depositing the mtDNA into GenBank** and annotating it correctly."*

Smith 2016, Briefings in Functional Genomics