Don’t believe the RefSeq!

or: How to annotate a genome.

Alexander Donath

Professur für Bioinformatik
Institut für Informatik
Universität Leipzig

Bled, 16. Februar 2010

If you use the RefSeq, you will get (among other things):

- Current knowledge of sequence data and biology
- Data validation and format consistency
- Ongoing curation by NCBI staff and collaborators, with reviewed records indicated
- Curation in accordance with established nomenclature

- [..] an increasing number of tests [..] to evaluate the quality of sequence and annotation [..]
- [..] a critical foundation for integrating sequence, genetic and functional information [..]
- [..] if whole genome sequencing projects are submitted [..] without annotation then calculate [..] and provide annotated proteins in the RefSeq

If you use the RefSeq, you will get (among other things):

- Current knowledge of sequence data and biology
- Data validation and format consistency
- Ongoing curation by NCBI staff and collaborators, with reviewed records indicated
- Curation in accordance with established nomenclature

- [..] an increasing number of tests [..] to evaluate the quality of sequence and annotation [..]
- [..] a critical foundation for integrating sequence, genetic and functional information [..]
- [..] if whole genome sequencing projects are submitted [..] without annotation then calculate [..] and provide annotated proteins in the RefSeq

If you use the RefSeq, you will get (among other things):

- Current knowledge of sequence data and biology
- Data validation and format consistency
- Ongoing curation by NCBI staff and collaborators, with reviewed records indicated
- Curation in accordance with established nomenclature

- [..] an increasing number of tests [..] to evaluate the quality of sequence and annotation [..]
- [..] a critical foundation for integrating sequence, genetic and functional information [..]
- [..] if whole genome sequencing projects are submitted [..] without annotation then NCBI may calculate [..] and provide annotated proteins in the RefSeq

- Phylogenetic analyses based on sequence and structure
- Genome rearrangements
- Homology search
- ...

Curated set of RefSeq:

- Annotations on wrong strand
- Missing strand annotation
- Obviously erroneous gene designations
- Missing gene annotations
- L1/L2 and S1/S2 identities erroneously switched
- No consistency in gene names

Non-curated set:

- Nonsensical gene annotation (tRNA-Asx: tRNA-Asn and tRNA-Asp)
- No attention in any case paid to conventions regarding upper vs. lower case lettering for gene names
- No gene designations for tRNA-encoding genes and naming for their products (e.g., tRNA-cys) instead of using actual gene names (e.g., trnC)

Some Examples:

NC_010300 tRNA **3916..13847** /product="tRNA-Met"
15-APR-2009
REVIEWED REFSEQ: This record has been curated by NCBI staff.

NC_012453 rRNA **1224..1244** /product="s-rRNA" (12S rRNA)
`GCACTAATCCACCTGCGGGCA`
20-APR-2009
REVIEWED REFSEQ: This record has been curated by NCBI staff.

NC_006893 rRNA **1..17202** /product="l-rRNA" (16S rRNA)
01-FEB-2010
REVIEWED REFSEQ: This record has been curated by NCBI staff.

Some Examples:

NC_010300 tRNA **3916..13847** /product=”tRNA-Met”
15-APR-2009
REVIEWED REFSEQ: This record has been curated by NCBI staff.

NC_012453 rRNA **1224..1244** /product=”s-rRNA” (12S rRNA)
`GCACTAATCCACCTGCGGGCA`
20-APR-2009
REVIEWED REFSEQ: This record has been curated by NCBI staff.

NC_006893 rRNA **1..17202** /product=”l-rRNA” (16S rRNA)
01-FEB-2010
REVIEWED REFSEQ: This record has been curated by NCBI staff.

Some Examples:

NC_010300 tRNA **3916..13847** /product=”tRNA-Met”
15-APR-2009
REVIEWED REFSEQ: This record has been curated by NCBI staff.

NC_012453 rRNA **1224..1244** /product=”s-rRNA” (12S rRNA)
`GCACTAATCCACCTGCGGGCA`
20-APR-2009
REVIEWED REFSEQ: This record has been curated by NCBI staff.

NC_006893 rRNA **1..17202** /product=”l-rRNA” (16S rRNA)
01-FEB-2010
REVIEWED REFSEQ: This record has been curated by NCBI staff.

Some Examples:

NC_010300 tRNA **3916..13847** /product="tRNA-Met"
15-APR-2009
REVIEWED REFSEQ: This record has been curated by NCBI staff.

NC_012453 rRNA **1224..1244** /product="s-rRNA" (12S rRNA)
`GCACTAATCCACCTGCGGGCA`
20-APR-2009
REVIEWED REFSEQ: This record has been curated by NCBI staff.

NC_006893 rRNA **1..17202** /product="l-rRNA" (16S rRNA)
01-FEB-2010
REVIEWED REFSEQ: This record has been curated by NCBI staff.

Curated set of RefSeq:

- Annotations on wrong strand
- Missing strand annotation
- Obviously erroneous gene designations
- Missing gene annotations
- L1/L2 and S1/S2 identities erroneously switched
- No consistency in gene names

Non-curated set:

- Nonsensical gene annotation (tRNA-Asx: tRNA-Asn and tRNA-Asp)
- No attention in any case paid to conventions regarding upper vs. lower case lettering for gene names
- No gene designations for tRNA-encoding genes and naming for their products (e.g., tRNA-cys) instead of using actual gene names (e.g., trnC)

Curated set of RefSeq:

- Annotations on wrong strand
- Missing strand annotation
- Obviously erroneous gene designations
- Missing gene annotations
- L1/L2 and S1/S2 identities erroneously switched
- No consistency in gene names

Non-curated set:

- Nonsensical gene annotation (tRNA-Asx: tRNA-Asn and tRNA-Asp)
- No attention in any case paid to conventions regarding upper vs. lower case lettering for gene names
- No gene designations for tRNA-encoding genes and naming for their products (e.g., tRNA-cys) instead of using actual gene names (e.g., trnC)

ATP6/8: atp6/8, ATP 6/8, ATPase 6/8, *apt6/8*

COX1: CO1, cox1, COI, CoxI, COX1_C

COX2: CO2, cox2, COII, CoxII, Cox2, cob

COX3: CO3, cox3, COIII, CoxIII

CYTB: cytb, cytB, Cyt b, Cytb, Cyt B, CytB, *CTYB*

ND1/2/3/4/4L/5: nadX, NADHX

12S rRNA: rns, rrns, rrnS, 12S ribosomal RNA, srRNA, rnr1, Rnr1, RNR1, ssu

16S rRNA: rnl, rrnl, rrnL, 16S ribosomal RNA, lrRNA, rnr2, Rnr2, RNR2, lsu-rRNA

tRNAs: tRNA-Cys, trnC, trnC-gca, trnC(gca), trnC(GCA), Trnc, TrnC, TRNC

Curated set of RefSeq:

- Annotations on wrong strand
- Missing strand annotation
- Obviously erroneous gene designations
- Missing gene annotations
- L1/L2 and S1/S2 identities erroneously switched
- No consistency in gene names

Non-curated set:

- Nonsensical gene annotation (tRNA-Asx: tRNA-Asn and tRNA-Asp)
- No attention in any case paid to conventions regarding upper vs. lower case lettering for gene names
- No gene designations for tRNA-encoding genes and naming for their products (e.g., tRNA-cys) instead of using actual gene names (e.g., trnC)

Curated set of RefSeq:

- Annotations on wrong strand
- Missing strand annotation
- Obviously erroneous gene designations
- Missing gene annotations
- L1/L2 and S1/S2 identities erroneously switched
- No consistency in gene names

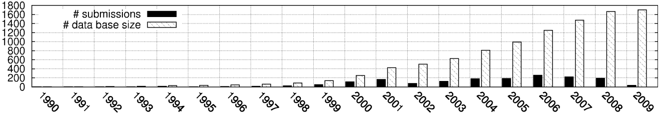
Non-curated set:

- Nonsensical gene annotation (tRNA-Asx: tRNA-Asn and tRNA-Asp)
- No attention in any case paid to conventions regarding upper vs. lower case lettering for gene names
- No gene designations for tRNA-encoding genes and naming for their products (e.g., tRNA-cys) instead of using actual gene names (e.g., trnC)

Jeffrey Boore (2006):

- Systematic error screening
- Standardization of gene names
- Anticodon labeling of tRNAs
- Standardize the format for designating genes
- Standards for gene boundaries designation
- Standards for accepting the reality of a gene assignment
- Include information on RNA editing
- Automate sequence alignments and phylogenetic analyses
- Fully integrate data from prokaryotic genomes

- The RefSeq Situation:



  - Version 1 (June 30, 2003): 417 mito-genomes
  - Version 39 (Jan 23, 2010): 2069 mito-genomes

MITOS:

- Library for each protein family
- Sequence-structure-profiles for tRNAs and rRNAs

1. `cmsearch` tRNAs
2. `blastx` proteins
   1. Initial protein hits
   2. Removal of overlapping hits
   3. Combination of fragments and detection of duplicates
   4. Improvement of start & stop
3. `cmsearch` rRNAs

MITOS:

- Library for each protein family
- Sequence-structure-profiles for tRNAs and rRNAs

1. `cmsearch` tRNAs
2. `blastx` proteins
    - 2.1 Initial protein hits
    - 2.2 Removal of overlapping hits
    - 2.3 Combination of fragments and detection of duplicates
    - 2.4 Improvement of start & stop
3. `cmsearch` rRNAs

MITOS:

- Library for each protein family
- Sequence-structure-profiles for tRNAs and rRNAs

1. `cmsearch` tRNAs
2. `blastx` proteins
   2.1 Initial protein hits
   2.2 Removal of overlapping hits
   2.3 Combination of fragments and detection of duplicates
   2.4 Improvement of start & stop
3. `cmsearch` rRNAs

MITOS:

- Library for each protein family
- Sequence-structure-profiles for tRNAs and rRNAs

1. `cmsearch` tRNAs
2. `blastx` proteins
    2.1 Initial protein hits
    2.2 Removal of overlapping hits
    2.3 Combination of fragments and detection of duplicates
    2.4 Improvement of start & stop
3. `cmsearch` rRNAs

MITOS:

- Library for each protein family
- Sequence-structure-profiles for tRNAs and rRNAs

1. `cmsearch` tRNAs
2. `blastx` proteins
    - 2.1 Initial protein hits
    - 2.2 Removal of overlapping hits
    - 2.3 Combination of fragments and detection of duplicates
    - 2.4 Improvement of start & stop
3. `cmsearch` rRNAs

MITOS:

- Library for each protein family
- Sequence-structure-profiles for tRNAs and rRNAs

1. `cmsearch` tRNAs
2. `blastx` proteins
    2.1 Initial protein hits
    2.2 Removal of overlapping hits
    2.3 Combination of fragments and detection of duplicates
    2.4 Improvement of start & stop
3. `cmsearch` rRNAs

- ~1000 **tRNAs** wrongly annotated, missing, or annotated too often
  - Correction of 203 tRNA families in 73 species (L1/L2 and S1/S2)
  - Correction of multiplicity of 166 tRNAs in 107 species
  - Identification of 108 missing tRNA genes
  - Removal of up to 4 duplicated tRNAs in 58 cases
  - Correction of the strand of 327 tRNAs in 195 species
- **Proteins**
  - One species with wrong gene code
  - ...
- **rRNAs**
  - Ongoing analysis

		e-value			
		5	**10**	**15**	**20**
cut-off	**5**		121 (125/184)		
	10	71 (77/212)	79 (83/155)	82 (85/142)	94 (97/149)
	15		34 (38/120)		
	20	24 (30/182)	27 (31/116)	61 (64/126)	81 (84/140)
	25		22 (26/114)		
	30	19 (25/182)	19 (23/107)	53 (56/115)	72 (75/133)
	35		15 (19/105)		
	40	19 (26/192)	**13** (18/106)	47 (51/109)	68 (72/126)
	45		**11** (16/101)		
	50		**11** (16/103)		

after 3rd cleaning (after 2nd/1st cleaning)

Let's recap!
If you use the RefSeq, you will get (among other things):

- Current knowledge of sequence data and biology
- Data validation and format consistency
- Ongoing curation by NCBI staff and collaborators, with reviewed records indicated
- Curation in accordance with established nomenclature

- [..] an increasing number of tests [..] to evaluate the quality of sequence and annotation [..]
- [..] a critical foundation for integrating sequence, genetic and functional information [..]
- [..] if whole genome sequencing projects are submitted [..] without annotation then NCBI may calculate [..] and provide annotated proteins in the RefSeq

Let's recap!
If you use the RefSeq, you will get (among other things):

- Current knowledge of sequence data and biology ✗
- Data validation and format consistency
- Ongoing curation by NCBI staff and collaborators, with reviewed records indicated
- Curation in accordance with established nomenclature

- [..] an increasing number of tests [..] to evaluate the quality of sequence and annotation [..]
- [..] a critical foundation for integrating sequence, genetic and functional information [..]
- [..] if whole genome sequencing projects are submitted [..] without annotation then NCBI may calculate [..] and provide annotated proteins in the RefSeq

Let's recap!
If you use the RefSeq, you will get (among other things):

- Current knowledge of sequence data and biology ✗
- Data validation and format consistency ✗
- Ongoing curation by NCBI staff and collaborators, with reviewed records indicated
- Curation in accordance with established nomenclature

- [..] an increasing number of tests [..] to evaluate the quality of sequence and annotation [..]
- [..] a critical foundation for integrating sequence, genetic and functional information [..]
- [..] if whole genome sequencing projects are submitted [..] without annotation then NCBI may calculate [..] and provide annotated proteins in the RefSeq

Let's recap!
If you use the RefSeq, you will get (among other things):

- Current knowledge of sequence data and biology ✗
- Data validation and format consistency ✗
- Ongoing curation by NCBI staff and collaborators, with reviewed records indicated ✗
- Curation in accordance with established nomenclature

- [..] an increasing number of tests [..] to evaluate the quality of sequence and annotation [..]
- [..] a critical foundation for integrating sequence, genetic and functional information [..]
- [..] if whole genome sequencing projects are submitted [..] without annotation then NCBI may calculate [..] and provide annotated proteins in the RefSeq

Let's recap!
If you use the RefSeq, you will get (among other things):

- Current knowledge of sequence data and biology ✗
- Data validation and format consistency ✗
- Ongoing curation by NCBI staff and collaborators, with reviewed records indicated ✗
- Curation in accordance with established nomenclature ✗

- [..] an increasing number of tests [..] to evaluate the quality of sequence and annotation [..]
- [..] a critical foundation for integrating sequence, genetic and functional information [..]
- [..] if whole genome sequencing projects are submitted [..] without annotation then NCBI may calculate [..] and provide annotated proteins in the RefSeq

Let’s recap!

If you use the RefSeq, you will get (among other things):

- Current knowledge of sequence data and biology ✗
- Data validation and format consistency ✗
- Ongoing curation by NCBI staff and collaborators, with reviewed records indicated ✗
- Curation in accordance with established nomenclature ✗

- [..] an increasing number of tests [..] to evaluate the quality of sequence and annotation [..] ✗
- [..] a critical foundation for integrating sequence, genetic and functional information [..]
- [..] if whole genome sequencing projects are submitted [..] without annotation then NCBI may calculate [..] and provide annotated proteins in the RefSeq

Let's recap!
If you use the RefSeq, you will get (among other things):

- Current knowledge of sequence data and biology ✗
- Data validation and format consistency ✗
- Ongoing curation by NCBI staff and collaborators, with reviewed records indicated ✗
- Curation in accordance with established nomenclature ✗

- [..] an increasing number of tests [..] to evaluate the quality of sequence and annotation [..] ✗
- [..] a critical foundation for integrating sequence, genetic and functional information [..] ✗
- [..] if whole genome sequencing projects are submitted [..] without annotation then NCBI may calculate [..] and provide annotated proteins in the RefSeq

Let's recap!
If you use the RefSeq, you will get (among other things):

- Current knowledge of sequence data and biology ✗
- Data validation and format consistency ✗
- Ongoing curation by NCBI staff and collaborators, with reviewed records indicated ✗
- Curation in accordance with established nomenclature ✗
- [..] an increasing number of tests [..] to evaluate the quality of sequence and annotation [..] ✗
- [..] a critical foundation for integrating sequence, genetic and functional information [..] ✗
- [..] if whole genome sequencing projects are submitted [..] without annotation then MITOS should calculate [..] and provide annotated proteins in the RefSeq

Let's recap!

Jeffrey Boore (2006):

- Systematic error screening
- Standardization of gene names
- Anticodon labeling of tRNAs
- Standardize the format for designating genes
- Standards for gene boundaries designation
- Standards for accepting the reality of a gene assignment
- Include information on RNA editing
- Automate sequence alignments and phylogenetic analyses
- Fully integrate data from prokaryotic genomes

Let's recap!

Jeffrey Boore (2006):

- Systematic error screening ✔
- Standardization of gene names
- Anticodon labeling of tRNAs
- Standardize the format for designating genes
- Standards for gene boundaries designation
- Standards for accepting the reality of a gene assignment
- Include information on RNA editing
- Automate sequence alignments and phylogenetic analyses
- Fully integrate data from prokaryotic genomes

Let’s recap!

Jeffrey Boore (2006):

- Systematic error screening ✔
- Standardization of gene names ✔
- Anticodon labeling of tRNAs
- Standardize the format for designating genes
- Standards for gene boundaries designation
- Standards for accepting the reality of a gene assignment
- Include information on RNA editing
- Automate sequence alignments and phylogenetic analyses
- Fully integrate data from prokaryotic genomes

Let’s recap!

Jeffrey Boore (2006):

- Systematic error screening ✓
- Standardization of gene names ✓
- Anticodon labeling of tRNAs ✓
- Standardize the format for designating genes
- Standards for gene boundaries designation
- Standards for accepting the reality of a gene assignment
- Include information on RNA editing
- Automate sequence alignments and phylogenetic analyses
- Fully integrate data from prokaryotic genomes

Let's recap!

Jeffrey Boore (2006):

- Systematic error screening ✔
- Standardization of gene names ✔
- Anticodon labeling of tRNAs ✔
- Standardize the format for designating genes ✔
- Standards for gene boundaries designation
- Standards for accepting the reality of a gene assignment
- Include information on RNA editing
- Automate sequence alignments and phylogenetic analyses
- Fully integrate data from prokaryotic genomes

Let’s recap!

Jeffrey Boore (2006):

- Systematic error screening ✔
- Standardization of gene names ✔
- Anticodon labeling of tRNAs ✔
- Standardize the format for designating genes ✔
- Standards for gene boundaries designation ✔
- Standards for accepting the reality of a gene assignment
- Include information on RNA editing
- Automate sequence alignments and phylogenetic analyses
- Fully integrate data from prokaryotic genomes

Let's recap!

Jeffrey Boore (2006):

- Systematic error screening ✔
- Standardization of gene names ✔
- Anticodon labeling of tRNAs ✔
- Standardize the format for designating genes ✔
- Standards for gene boundaries designation ✔
- Standards for accepting the reality of a gene assignment ✔
- Include information on RNA editing
- Automate sequence alignments and phylogenetic analyses
- Fully integrate data from prokaryotic genomes

Let’s recap!

Jeffrey Boore (2006):

- Systematic error screening ✓
- Standardization of gene names ✓
- Anticodon labeling of tRNAs ✓
- Standardize the format for designating genes ✓
- Standards for gene boundaries designation ✓
- Standards for accepting the reality of a gene assignment ✓
- Include information on RNA editing
- Automate sequence alignments and phylogenetic analyses
- Fully integrate data from prokaryotic genomes

Let’s recap!

Jeffrey Boore (2006):

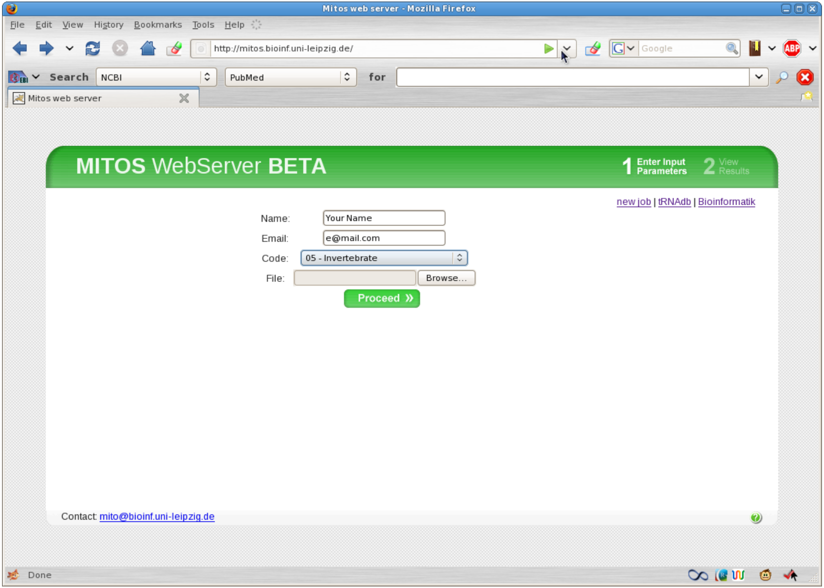
- Systematic error screening ✔
- Standardization of gene names ✔
- Anticodon labeling of tRNAs ✔
- Standardize the format for designating genes ✔
- Standards for gene boundaries designation ✔
- Standards for accepting the reality of a gene assignment ✔
- Include information on RNA editing
- Improved data set for phylogenetic analyses and homology search
- Fully integrate data from prokaryotic genomes

Let's recap!

Jeffrey Boore (2006):

- Systematic error screening ✔
- Standardization of gene names ✔
- Anticodon labeling of tRNAs ✔
- Standardize the format for designating genes ✔
- Standards for gene boundaries designation ✔
- Standards for accepting the reality of a gene assignment ✔
- Include information on RNA editing
- Improved data set for phylogenetic analyses and homology search
- Future: Fully integrate data from all eukaryotic genomes

- Matthias Bernt, Guido Fritzsch, Frank Jühling
- Fabian Externbrink
- HPC TU-Dresden
- DFG SPP 1174 DMP



Name: Your Name

Email: e@mail.com

Code: 05 - Invertebrate

File:

Contact: mito@bioinf.uni-leipzig.de

`mitos.bioinf.uni-leipzig.de`

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

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

		e-value			
		5	**10**	**15**	**20**
cut-off	**5**		121 (125/184)		
	10	71 (77/212)	79 (83/155)	82 (85/142)	94 (97/149)
	15		34 (38/120)		
	20	24 (30/182)	27 (31/116)	61 (64/126)	81 (84/140)
	25		22 (26/114)		
	30	19 (25/182)	19 (23/107)	53 (56/115)	72 (75/133)
	35		15 (19/105)		
	40	19 (26/192)	**13** (18/106)	47 (51/109)	68 (72/126)
	45		**11** (16/101)		
	50		**11** (16/103)		

after 3rd cleaning (after 2nd/1st cleaning)

e = 10	**MITOS**	**RefSeq**
c = 40	2	11
c = 45	2	9
c = 50	2	9

e=10,c=45	**MITOS**	**RefSeq**
NC_013253	1,0,1,1,**1**,1,1,**2**,1,1,1,1,1	1,0,1,1,**0**,1,1,**1**,1,1,1,1,1
NC_009081	1,0,1,**1**,1,1,1,1,1,1,1,1,1	1,0,1,**0**,1,1,1,1,1,1,1,1,1
NC_008572	1,**0**,1,1,1,1,1,1,1,1,1,1,1	1,**1**,1,1,1,1,1,1,1,1,1,1,1
NC_012893	1,**0**,1,1,1,1,1,1,1,1,1,1,1	1,**1**,1,1,1,1,1,1,1,1,1,1,1
NC_012739	1,**0**,1,1,1,1,1,1,1,1,1,1,1	1,**1**,1,1,1,1,1,1,1,1,1,1,1
NC_012727	1,**0**,1,1,1,1,1,**2**,1,1,1,1,1	1,**1**,1,1,1,1,1,**1**,1,1,1,1,1
NC_005928	1,0,1,1,1,1,1,1,1,1,1,1,**0**	1,0,1,1,1,1,1,1,1,1,1,1,**1**
NC_002651	1,1,1,1,1,1,1,1,1,1,**0**,1,1	1,1,1,1,1,1,1,1,1,1,**1**,1,1
NC_008070	1,1,1,1,1,1,1,1,1,1,**0**,1,1	1,1,1,1,1,1,1,1,1,1,**1**,1,1
NC_005939	1,**0**,1,1,1,1,1,1,1,1,1,**2**,1	1,**1**,1,1,1,1,1,1,1,1,1,**1**,1
NC_010568	1,**0**,1,1,1,1,1,1,1,1,1,1,1	1,**1**,1,1,1,1,1,1,1,1,1,1,1
Genes	**ATP6 | ATP8 | COX1 | COX2 | COX3 | CYTB | ND1 | ND2 | ND3 | ND4 | ND4L | ND5 | ND6**	**ATP6 | ATP8 | COX1 | COX2 | COX3 | CYTB | ND1 | ND2 | ND3 | ND4 | ND4L | ND5 | ND6**