

MinION sequencing provides new insight on the evolutionary history of seabird mitochondrial genomes

Lucas Torres, Andreanna Welch, Catherine Zanchetta, Vincent Bretagnolle & Eric Pante



Procellariiformes are most of 200 seabird species, presenting a diversity of morphologies and history life traits. They show high dispersion abilities, theoretically facilitating gene flows. Many genetic studies focus on Procellariiformes.



© JJ Harrison



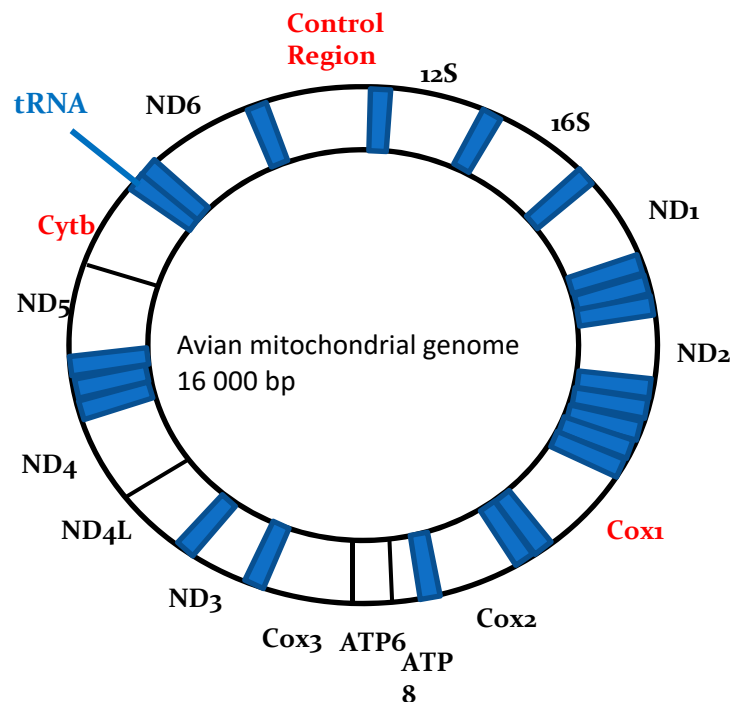
© P. Blevin



© P. Blevin



© N. Dean



Most of these studies use mitochondrial markers, principally:

Cox1 (cytochrome oxidase 1),
Cyt-b (cytochrome-b)
and CR (control region)



© JJ Harrison

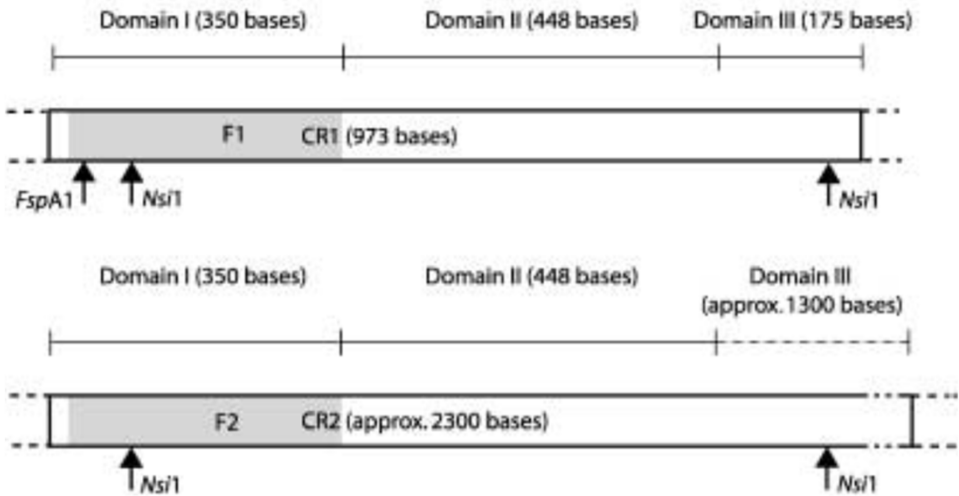
Standard avian mitochondrial genome composition

Cytb	ND6	CR	12S
------	-----	----	-----

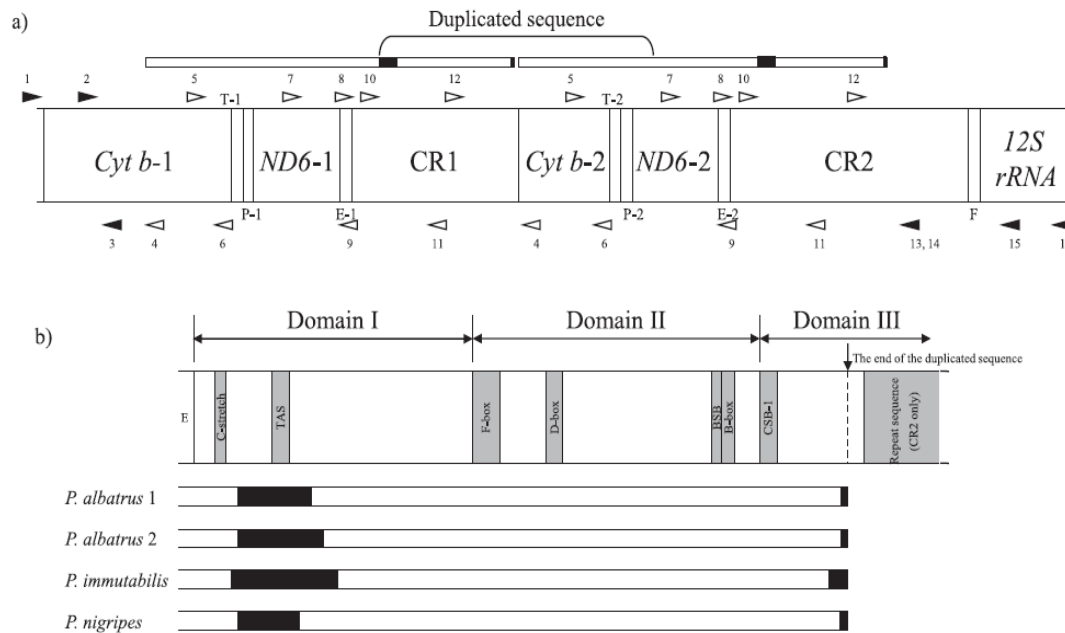
Thalassarche albatross mitochondrial genome composition

Cytb-1	ND6-1	CR1	pCytb	ND6-2	CR2	12S
--------	-------	-----	-------	-------	-----	-----

Using primer-walking, Abbott et al. 2005 found a duplicated region in the *Thalassarche melanophris* mito-genome.



The two copies of ND6 were identical, only a part of Cytb was copied and the two copies of CR were divergent at the extremities



Three other studies found the same results on *T. melanophrys* and three species of *Phoebastria* albatrosses, using primer-walking or Illumina sequencing

Eda et al. 2010 (Fig),
Gibb et al. 2006,
Lounsberry et al. 2015



Thalassarche

© JJ Harrison



Phoebastria

© Forest &
Kim Starr

The duplication raises a methodological problem

Albatross mitochondrial genome composition

Cytb-1	ND6-1	CR1	pCytb	ND6-2	CR2	12S
--------	-------	-----	-------	-------	-----	-----



© JJ Harrison

CR1 GCGTAAAT

CR2 GCGTGAAT



Amplification by PCR



Preferential amplification of one copy

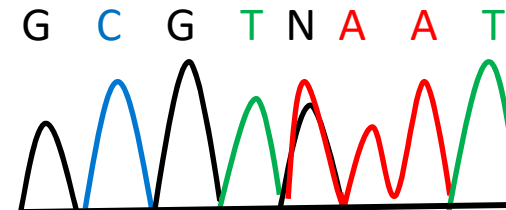
Individual 1 GCGTAAAT

Individual 2 GCGTGAAT

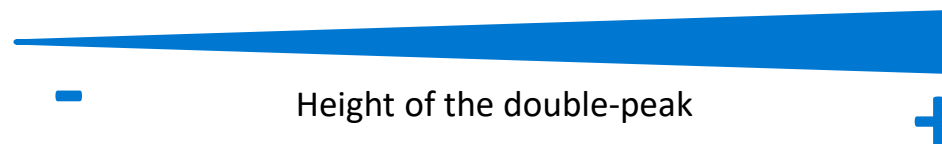
False information of divergence between individuals

Albatrosses 10% divergence (= 2 different species)

Equal amplification of both copies



Ambiguity in sequencing,
loss of the information for this base

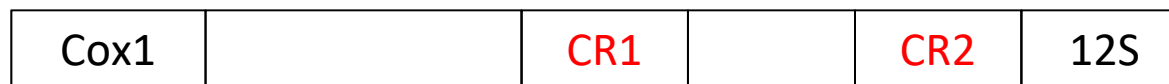




© V. Lemoine

We want to see if the duplication is present on another Procellariiformes species:
Audubon shearwater *Puffinus lherminieri*

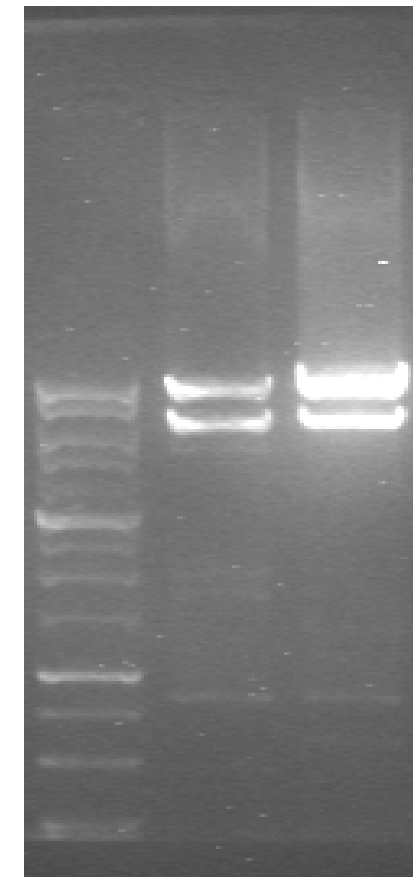
We carried out a Long Range PCR between Cox1 and CR



7 kb



10 kb



10 kb
7 kb

Two products were amplified, of size corresponding to the expected size in case of duplication



Puffinus mitochondrial genome composition

Cytb	?	12S
------	---	-----

Albatross mitochondrial genome composition

Cytb-1	ND6-1	CR1	pCytb	ND6-2	CR2	12S
--------	-------	-----	-------	-------	-----	-----



DNA fragments used in Abbot et al. 2005

Sequencing was made by several short DNA fragments, hence no fragment contained the entire duplicated region => indirect evidence

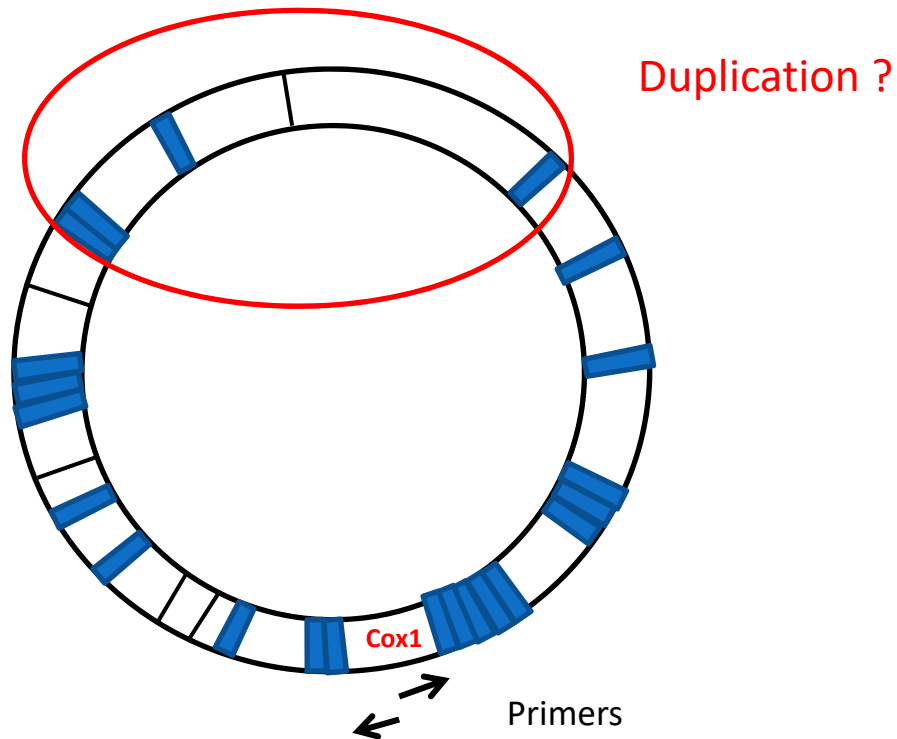
We will try to obtain DNA fragments long enough to show a direct evidence of the putative duplicated region in the *Puffinus* genome



We choose to use MinION ONT sequencing


MinION allows to sequence long DNA fragments (reads over 45 kb long) but with a local error rate of 1%. It is acceptable, since we are only interested by the genome composition. MinION will allow us to sequence the complete putative duplicated region.

To enrich our sample in mitochondrial DNA, we carried out a long range PCR (18 kb)
TaKaRa LA Taq® DNA Polymerase Hot-Start Version



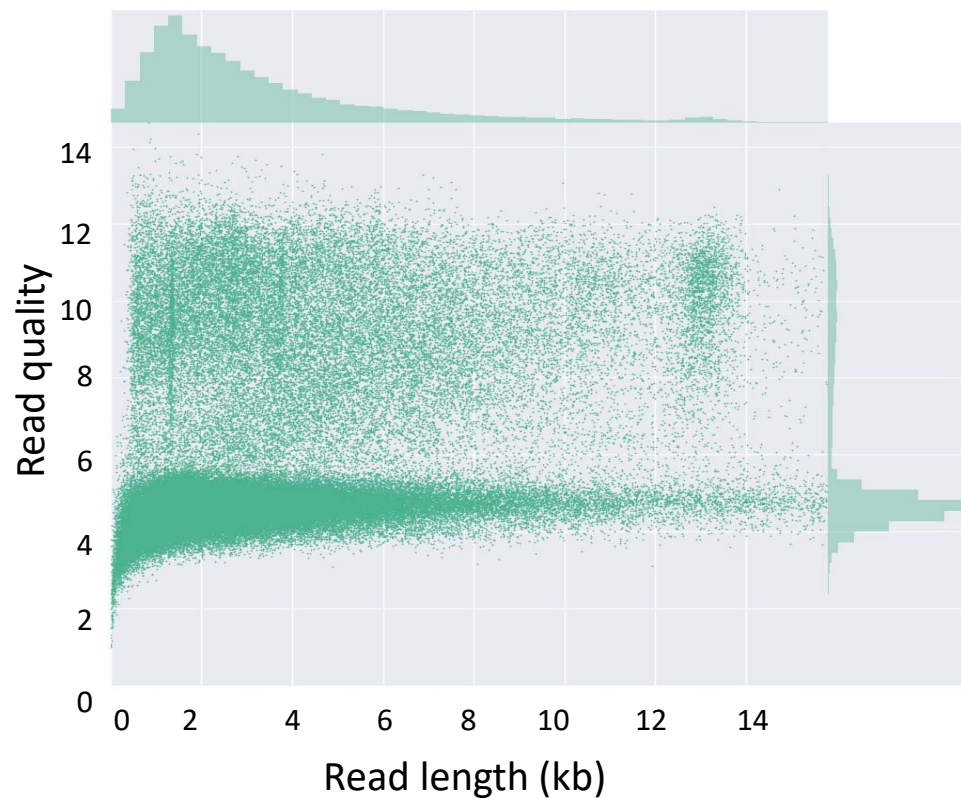
Three PCR products of 50 µL were pooled, and a final purification was performed using AMPure beads, to obtain 13 µg of DNA.

DNA was prepared and sequenced by MinION at the Genotoul with Catherine Zanchetta, using a protocol modified from ONT. Libraries were prepared using the ONT 1D ligation sequencing kit. DNA was not sheared to maximize sequencing read length.

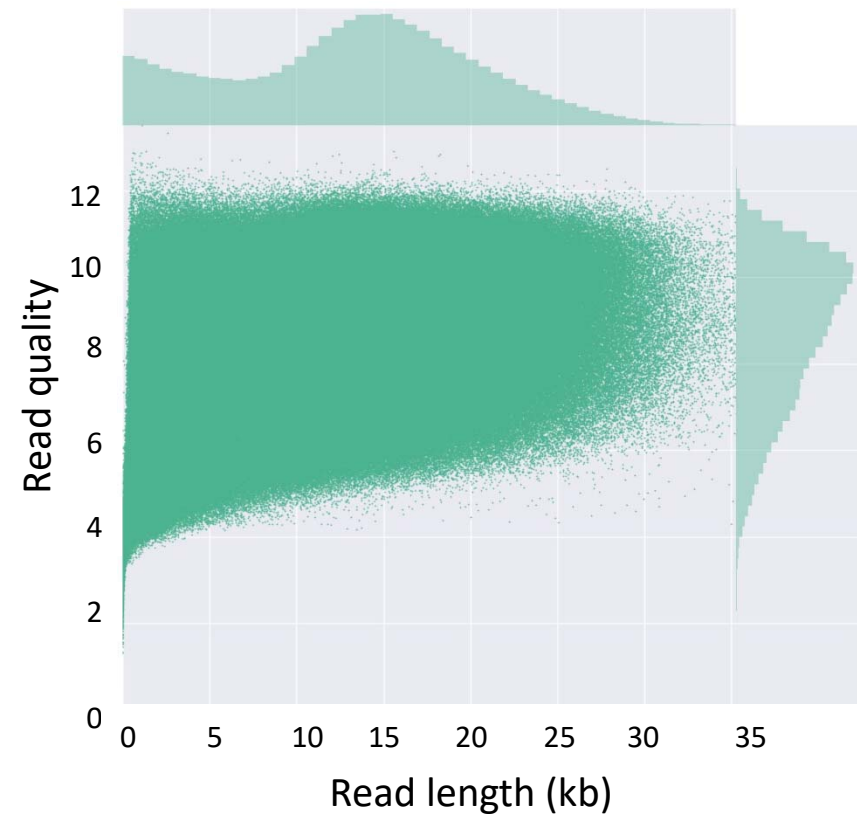



We obtained reads between 1 kb and 100 kb,
as we expected a genome of 18 kb length we decided to filter reads longer than 20 kb.
Adapters were deleted using Porechop
Reads with MinION quality less than 10 were filtered with NanoFilt, to increase the accuracy.

Our run



A more typical run

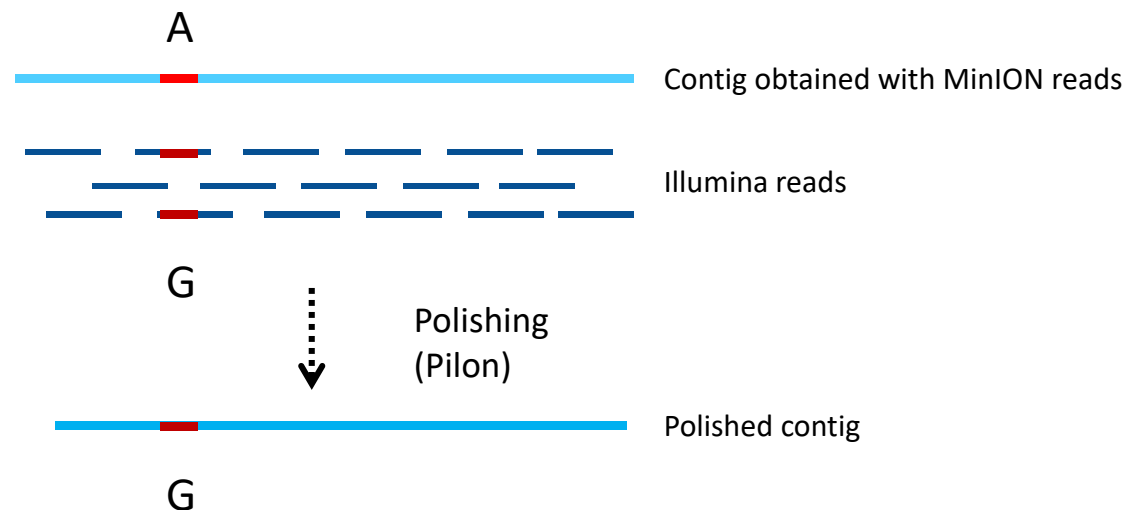




Remaining reads were assembled with Canu.
Canu set assembly with a targeted coverage of 100X all along the 18 kb long genome,
to increase the accuracy.

The resulting sequence was annotated with MITOS and compared to available
Albatrosses and Shearwaters sequences with BLASTN

To obtain clean sequences we used Illumina reads, provided by Andreanna Welch.
Contrarily to MinION, Illumina allows to sequence a large amount of short (100 bp)
but accurate DNA reads. The combination of the two methods will allow us to
obtain accurate sequences of the putative duplicated region.



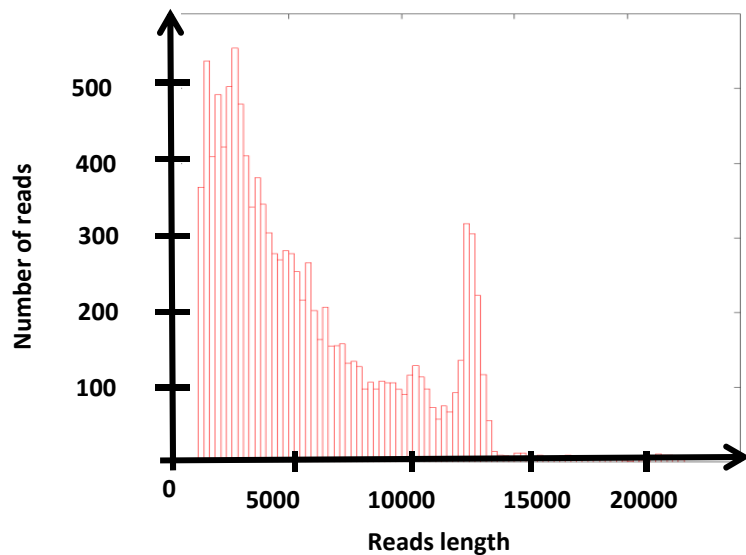


Results and Discussion

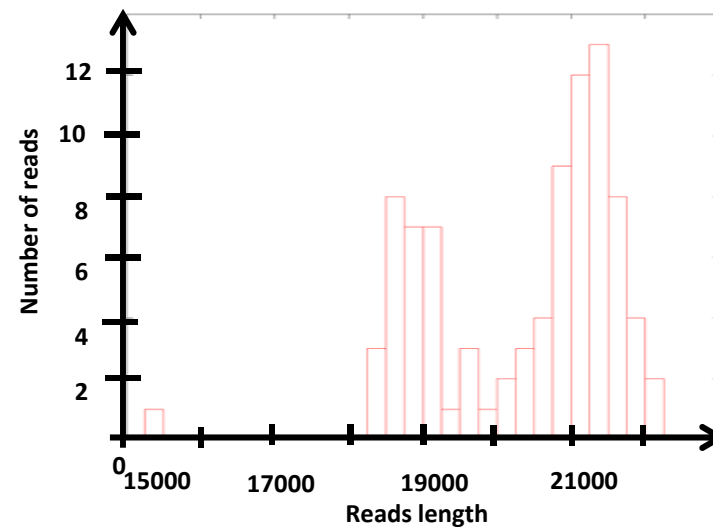


We have obtained 148 644 raw MinION reads.

After filtering we ran Canu with 12 764 reads and 87 reads were retained for the final assembly. The median length of the retained reads was 21 203 bp.



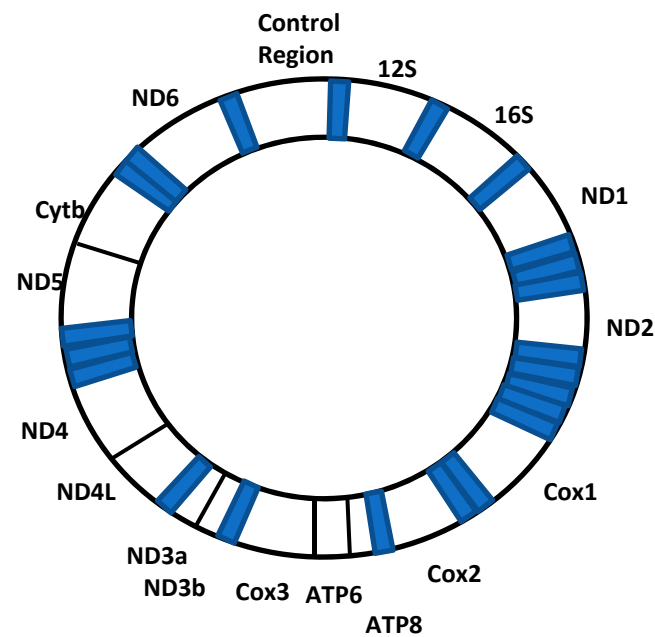
Length distribution of the filtered input reads



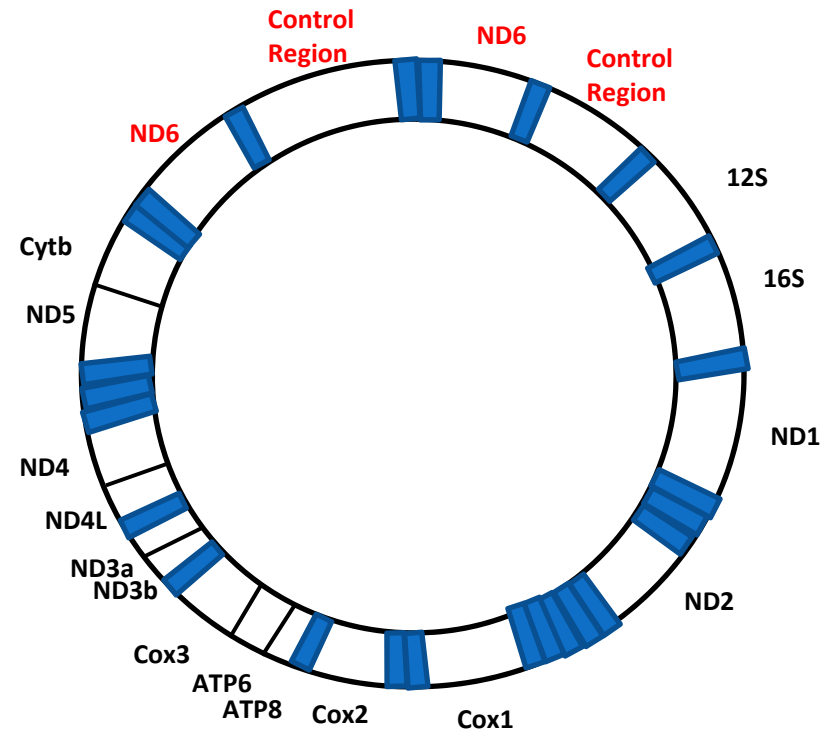
Length distribution of the reads used in the assembly

We mapped 277 693 Illumina reads on this assembly that served for the polishing

We have obtained two complete mitochondrial genome, 18 884 bp and 21 144 bp long. and composed of 13-14 protein coding genes, 25 tRNA genes and 2 rRNA genes.




$12/87 = 15\%$ of the reads



$75/87 = 85\%$ of the reads

Artefact or heteroplasmy ?



We have obtained MinION reads containing the entire duplicated region, bringing a direct evidence of its presence in the Procellariiformes genome

Puffinus mitochondrial genome composition

Cytb	ND6-1	CR1	ND6-2	CR2	12S
------	-------	-----	-------	-----	-----

Albatross mitochondrial genome composition

Cytb-1	ND6-1	CR1	pCytb	ND6-2	CR2	12S
--------	-------	-----	-------	-------	-----	-----

This duplicated region was different from the Albatrosses, since no copy of Cyt-b was found in the mitochondrial genome of this *Puffinus* individual



Similarly to Albatrosses species:

The two copies of ND6 are identical.

ND6-1

ND6-2

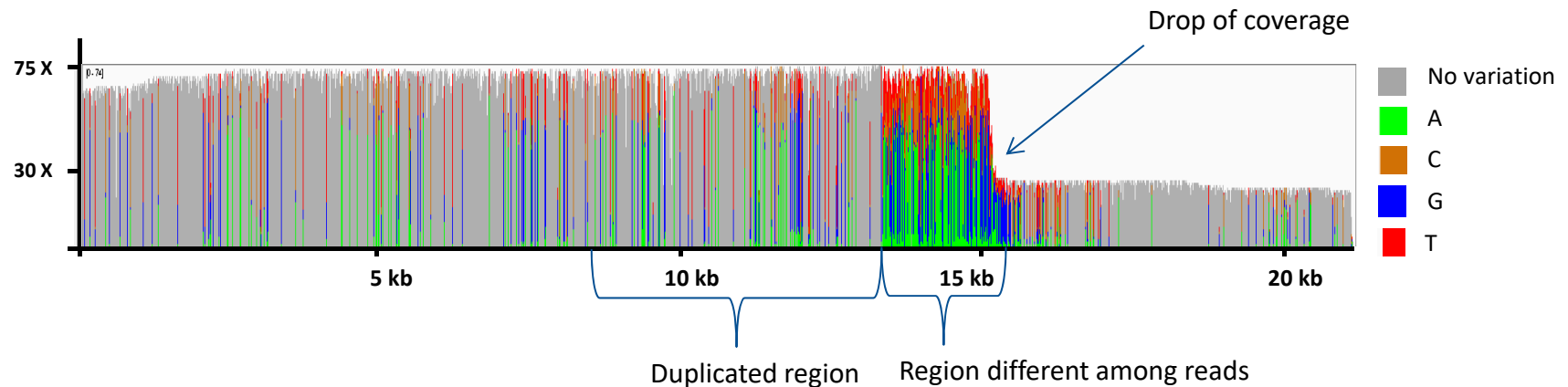
The two CRs show 24 SNP in the first 150 bp, followed by 1100 bp identical in the two sequences, then CR2 show 1805 supplementary bp.

CR1

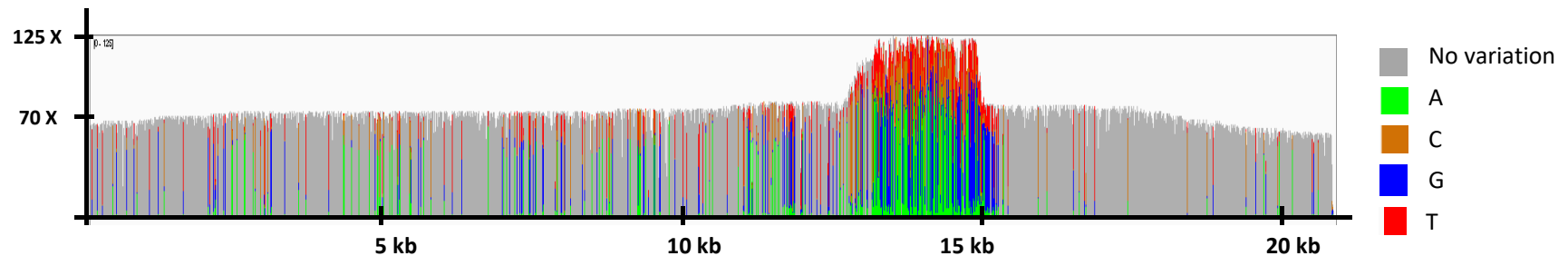
CR2



MinION reads used in the assembly were mapped against the complete mitogenome with the duplication. All the 75 reads mapped again.



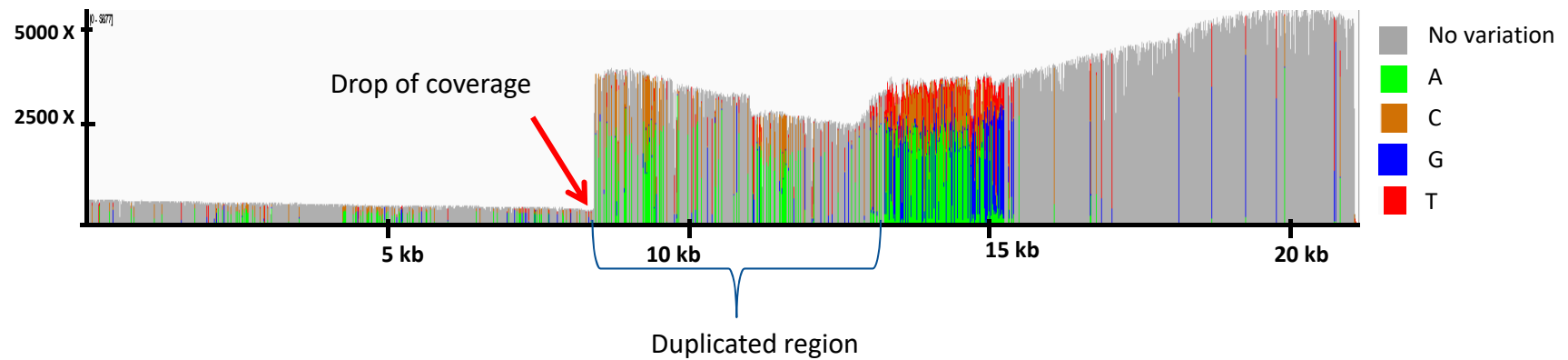
With the IGV2 option « supplementary alignment » if two alignments of one read are possible, both of them are shown.



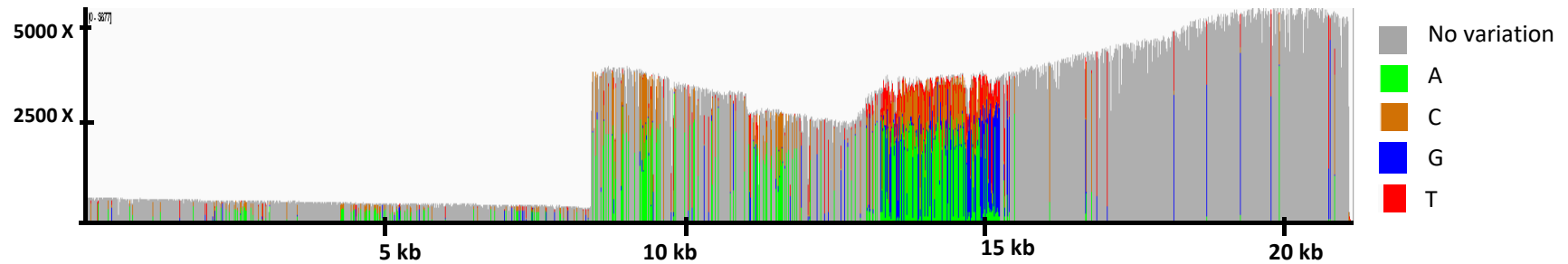
Numerous read can be placed in two emplacements, along the complex region, which indicates probably a repeated region.



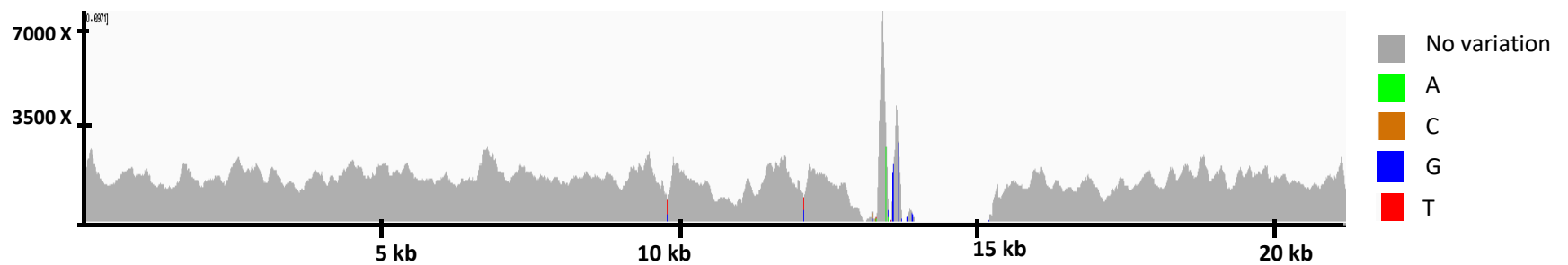
The 12 000 MinION input reads were mapped against the complete mitogenome.



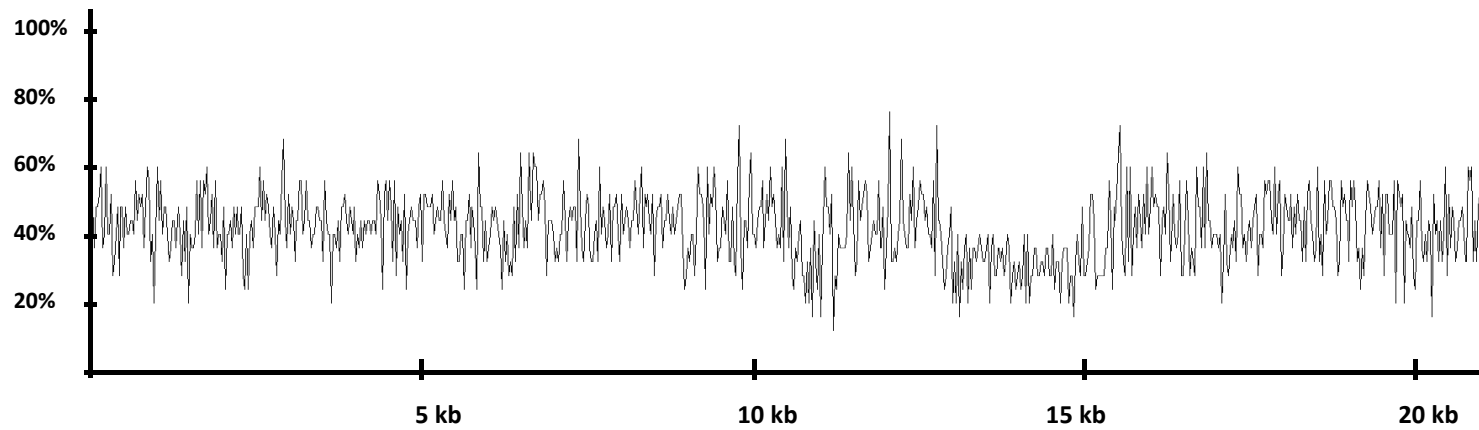
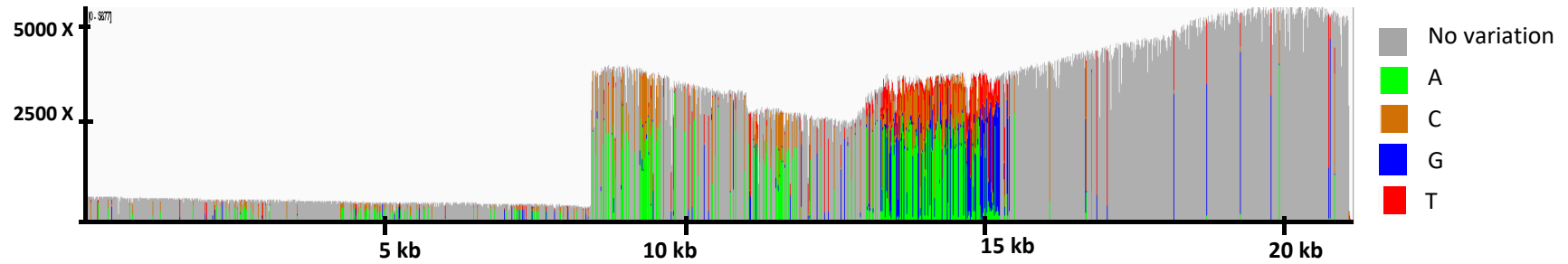
This showed that MinION reads did not cover all the mitogenome equally and confirmed that the complex region is present in all MinION reads



Illumina reads were mapped against the mitogenome too.



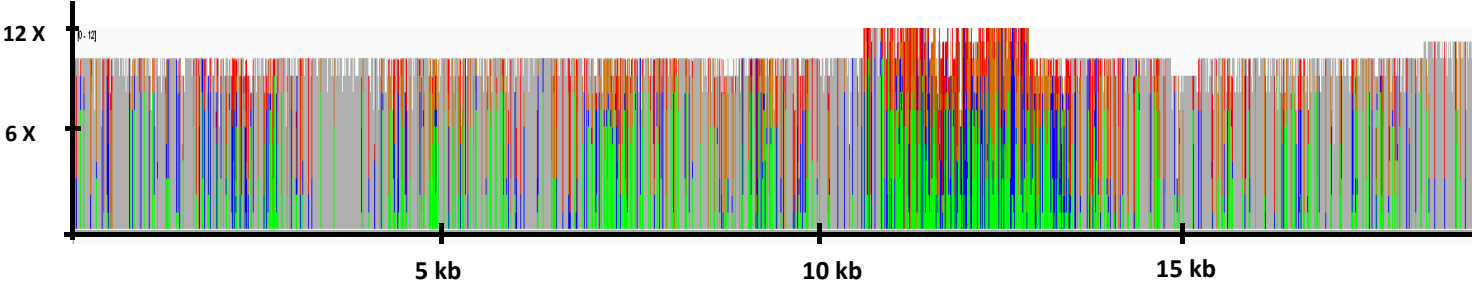
The complex region shows a peak of coverage, suggesting a repeated region, badly linked to the rest of the genome.



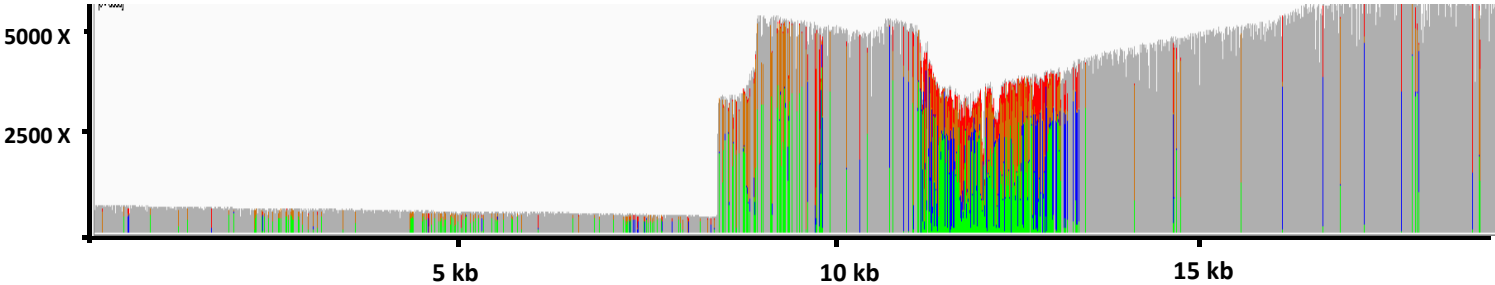
GC content confirms a repeated region GC poor



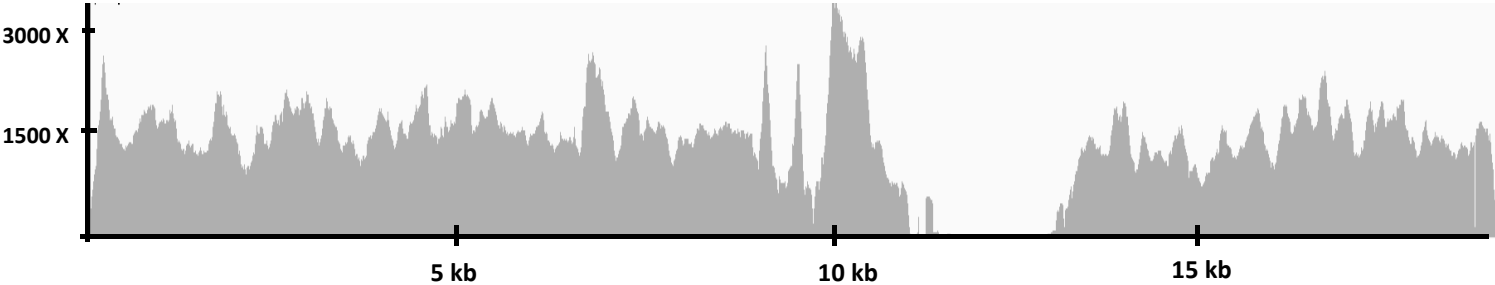
The same things are shown on the shortest contig, with peak of coverage on the duplicated region




Coverage of mapped 12 reads used in the assembly.



Coverage of mapped 12000 input reads



Coverage of mapped Illumina reads

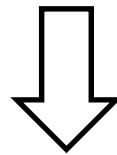


The assembly contains a repeated sequence of 40-60 bp, repeated 60 times, covering 2700 bp.

MinION reads all show this region in the complete genome, with variation among reads probably since MinION have trouble sequencing it.

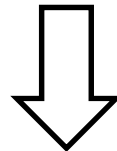
Illumina reads show that this region is probably composed of the same repetition, but do not link it with the rest of the mitogenome

This region appears between two complex motifs: the control region AT rich, and a repeated AC motif in 12S.
Submitted to BlastN, the repeated region does not match certainly for anything



H0: The repeated region is due to an artefact, for example to a chimeric amplification in the early stages of the PCR.

H1: The repeated region is biologically present in the mitogenome of *Puffinus lherminieri*



To test these hypotheses, we are planning to:

1. Shotgun MinION sequencing of genomic DNA
2. Carry out a PCR from the control region to 12S




We have spotlighted a duplicated region in the mitochondrial genome

Cytb	ND6-1	CR1	ND6-2	CR2	Repeated region	12S
------	-------	-----	-------	-----	-----------------	-----

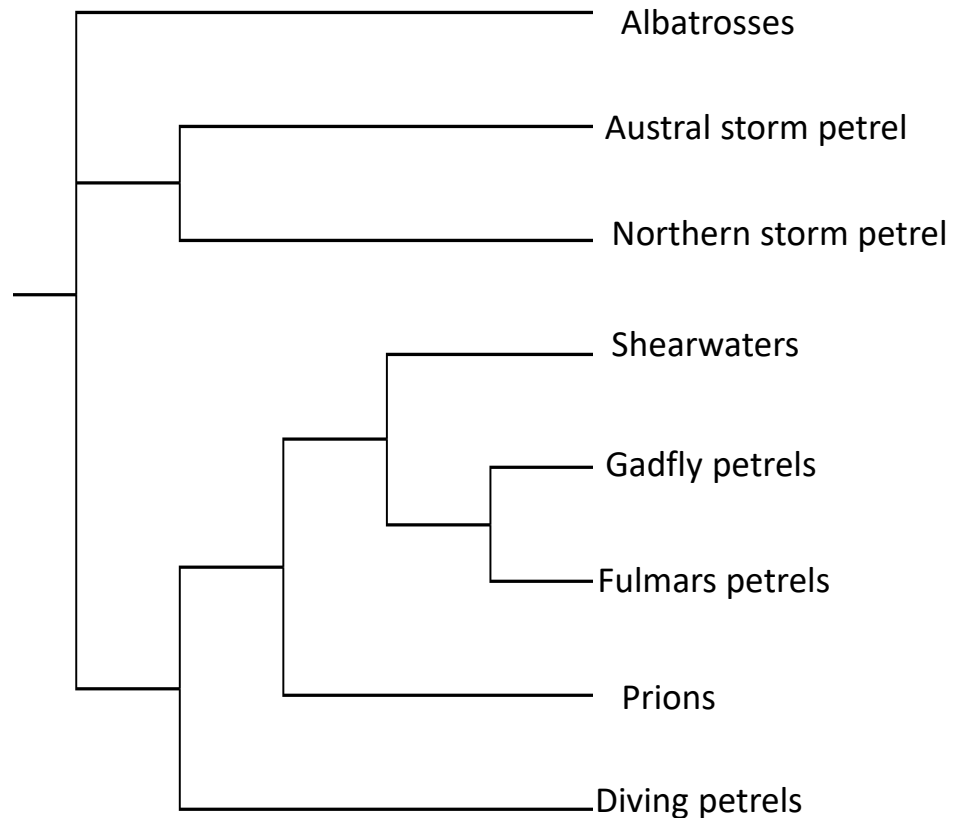
This region is similar to the one found in Albatrosses mitogenome, except that Cytb seems to be not duplicated

MinION reads allowed us to find the gene order but have to be used in complement with more accurate reads to obtain the exact sequence.

Our protocol show the presence of a repeated region, which we cannot resolve with this only method.



What is the evolutionary scenario of the duplicated region within the phylogeny of Procellariiformes?



We have Illumina reads, obtained as a by-catch of a targeted enrichment project directed by Andreanna Welch.

We will complete the taxonomic sampling Including one species by genus and an outgroup

We will try to determinate the ancestral state and rebuild the evolutionary scenario of this duplicated region



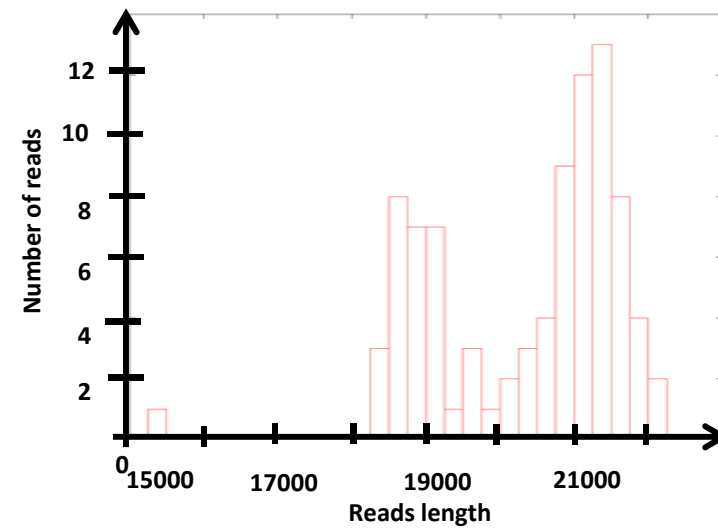
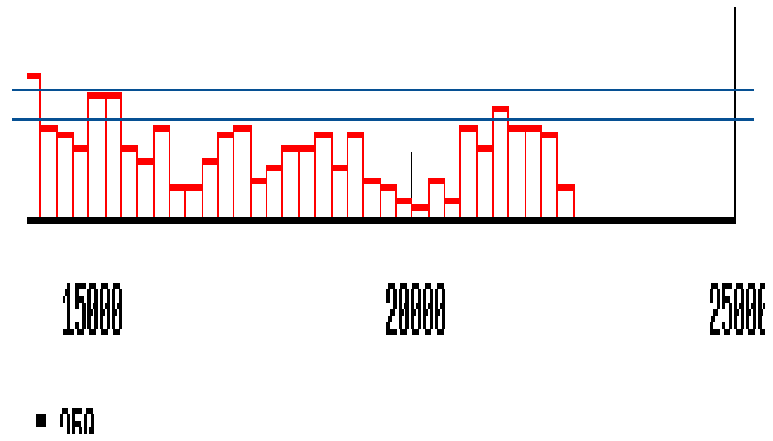
© V. Lemoine

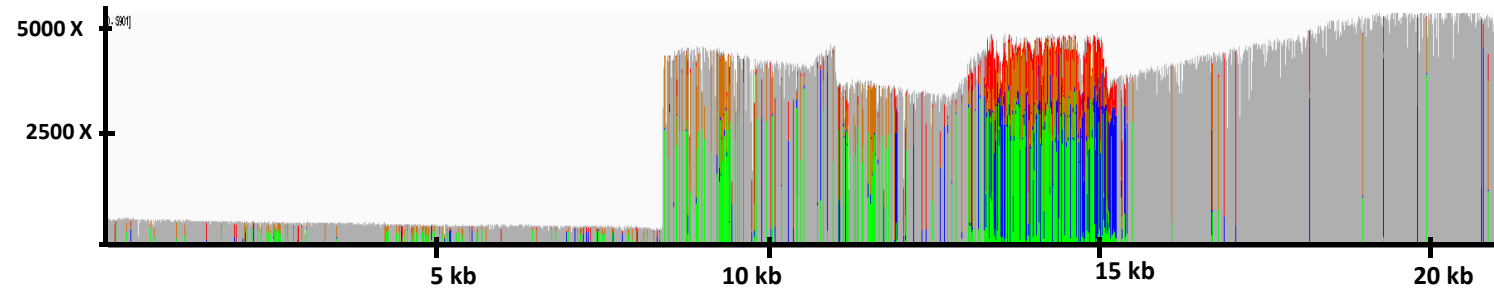
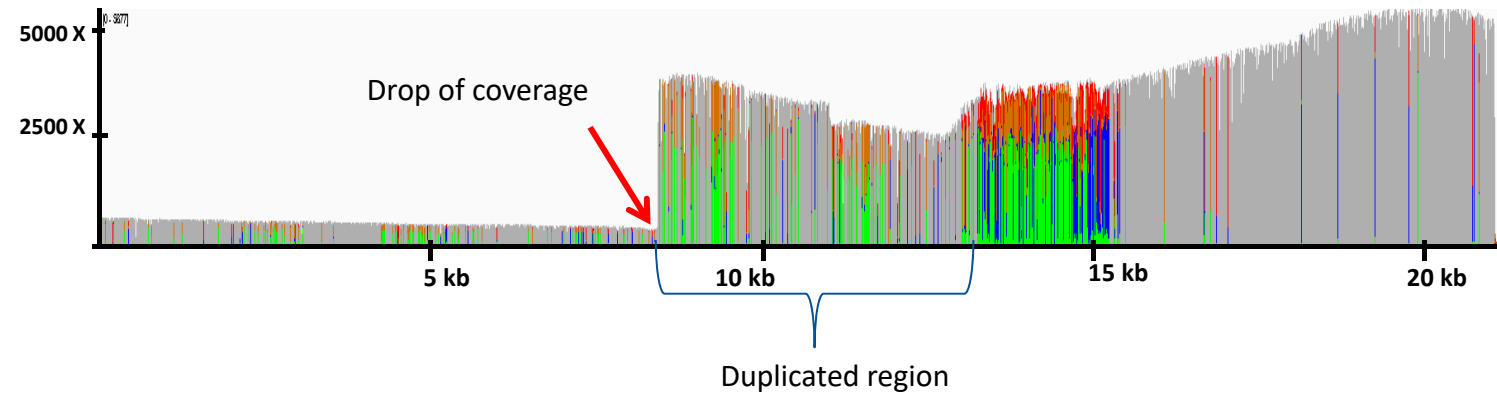


© S. Garvie

Thank you







Procellariiformes are most of 200 seabird species, presenting a variety of morphologies and history life traits. Moreover several species are considered as endangered, which implies conservation matters.



© P. Blevin



© R. Johansen



© N. Dean



© P. Blevin



© LaTourrette



© F. Pelsy



© JJ Harrison



© P. Blevin



© T. Hardaker



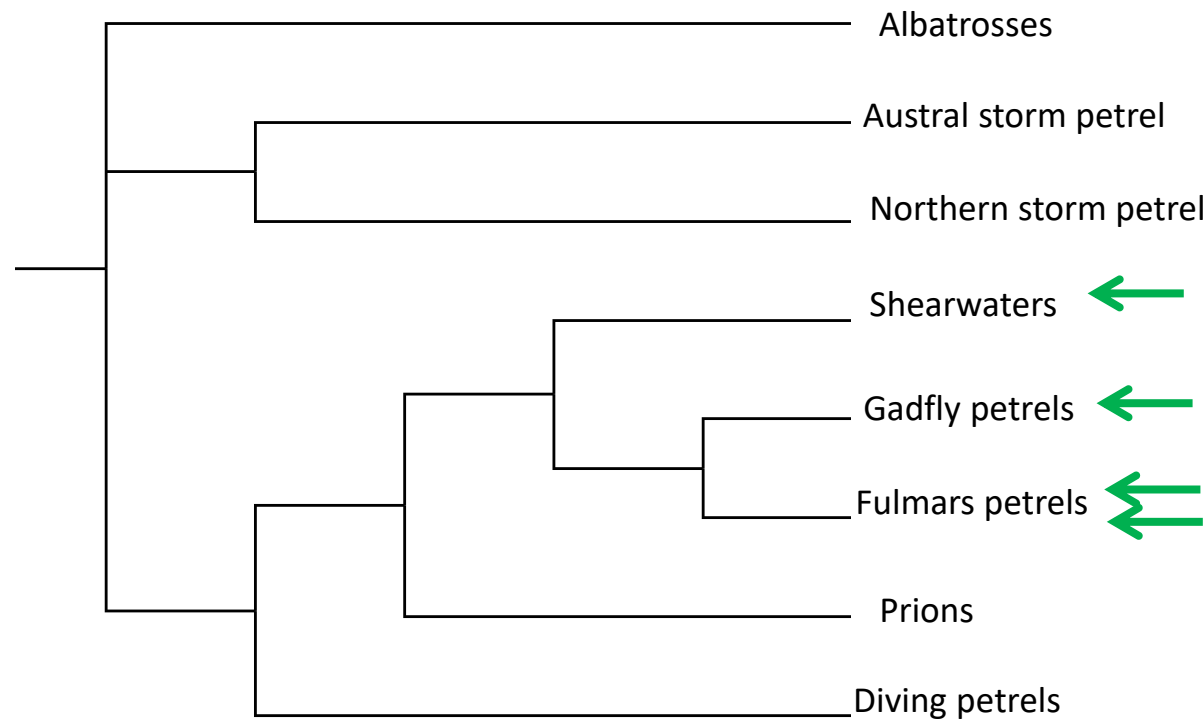
© JJ Harrison



Cox1 (800bp) allows to distinguish related species (molecular barcode) but can also be used to rebuild order systematics
Cytb (1100 bp), faster, allows to infer relationships among closely related taxa
RC (1000 bp) non-coding and hypervariable, is often used for population analysis

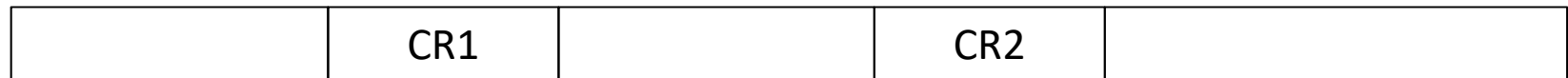


To go further Andreanna Welch provided us Illumina reads for four other Procellariiformes species





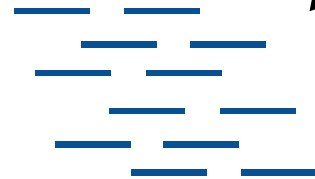
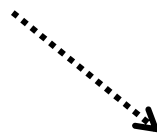
If the genome contains two similar copies of CR



Illumina reads
corresponding to CR1



Illumina reads
corresponding to CR2



Illumina reads corresponding to
CR1 and CR2



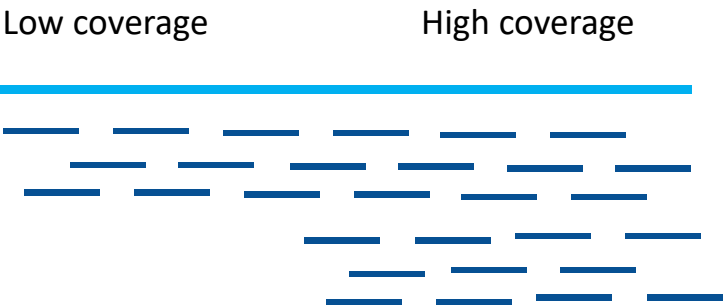
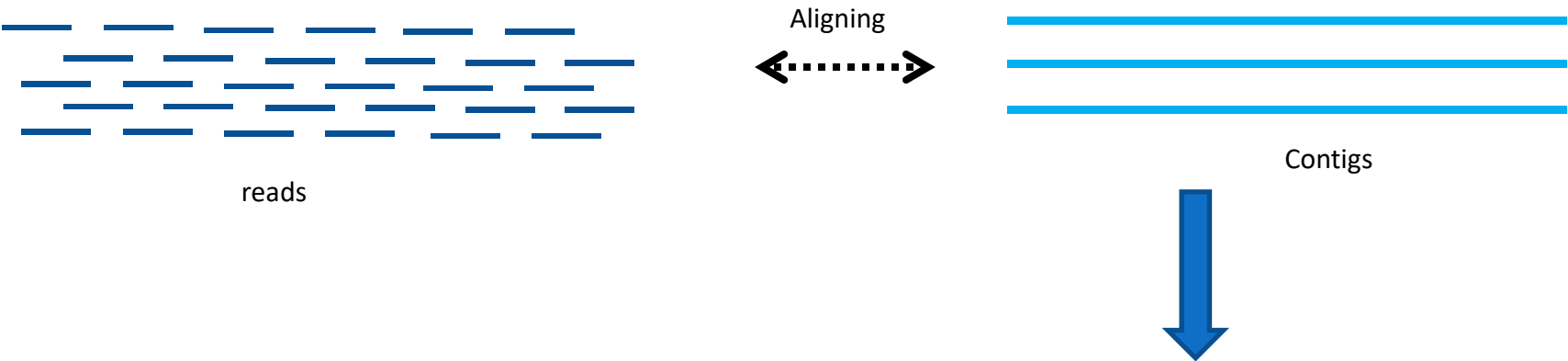
Assembling



During the assembling, all the reads will be assembled together indicating only one CR

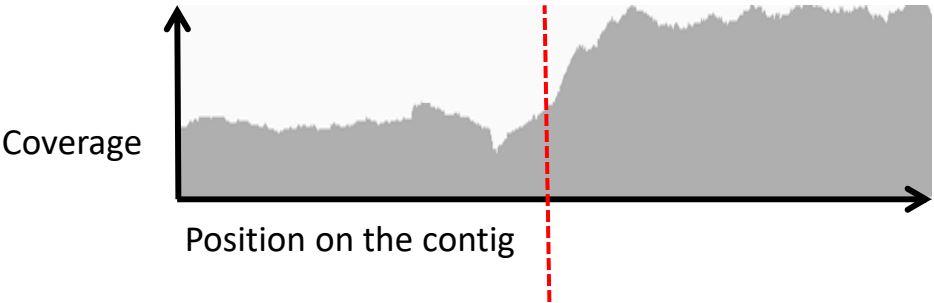


To address this issue, we aligned all the reads on the inferred contigs

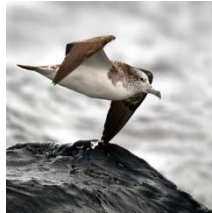


Coverage (i.e. the number of reads aligned) along the contigs was graphically measured

A peak of coverage is a strong hint of a duplicated region

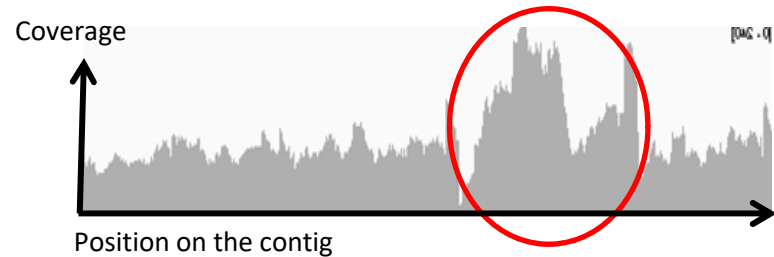


Calonectris leucomelas

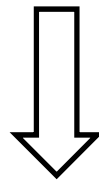


© Robin Newlin

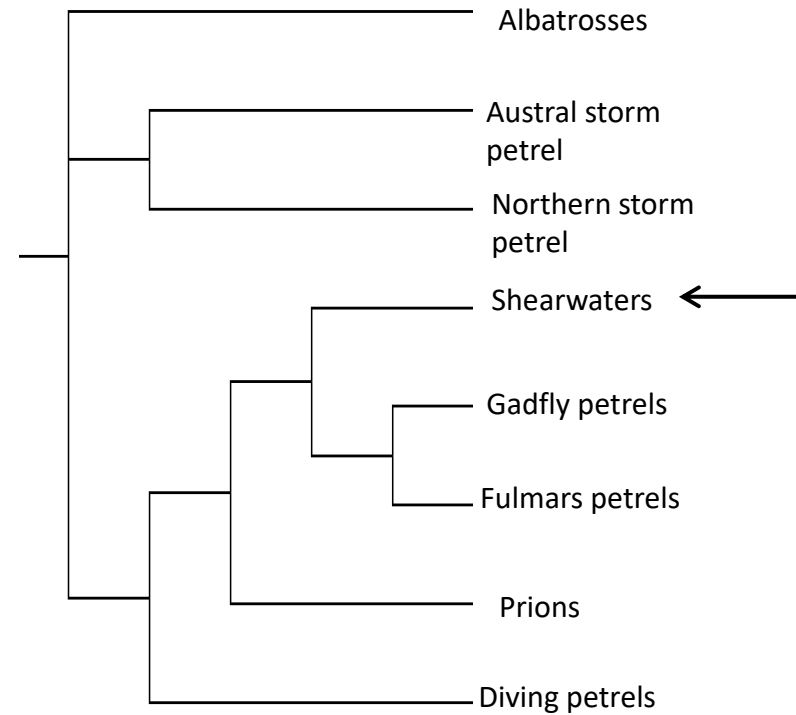
Peak of coverage on CR



Cytb	ND6-1	CR1	ND6-2
------	-------	-----	-------



Cytb	ND6-1	CR1	ND6-2	CR2	12S
------	-------	-----	-------	-----	-----



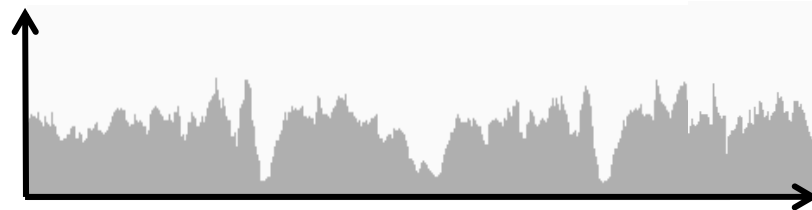
The duplicated region pattern is similar between the two Shearwaters genera

Fulmarus glacialis



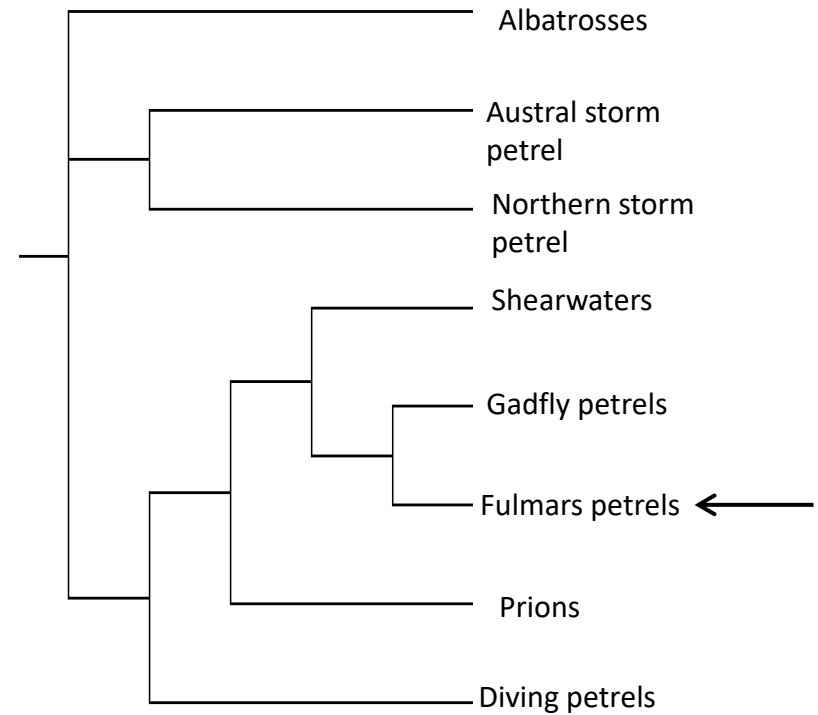
© S. Garvie

Coverage



Position on the contig

Cytb	ND6-1	CR1	ND6-2	CR2	12S
------	-------	-----	-------	-----	-----



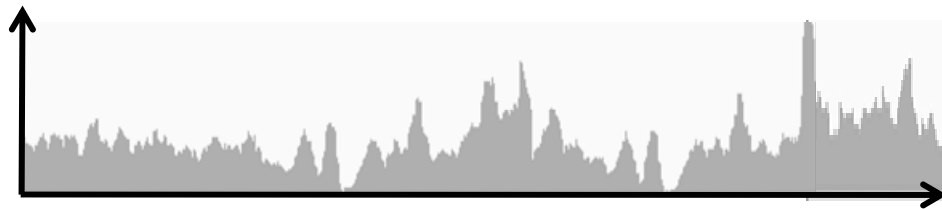
The duplicated region pattern is similar between *Puffinus* and *Fulmarus*

Daption capense



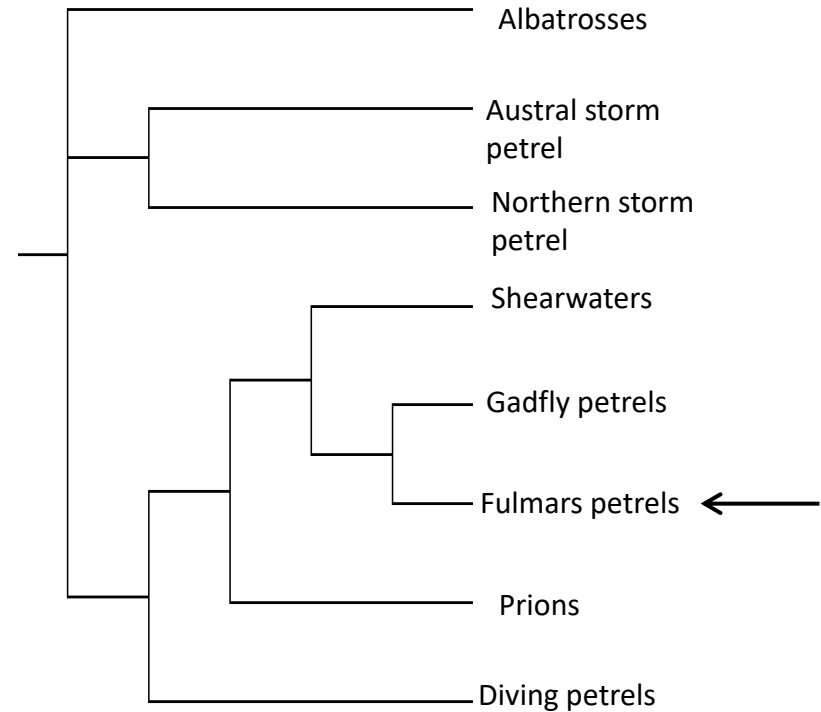
© P. Blevin

Coverage



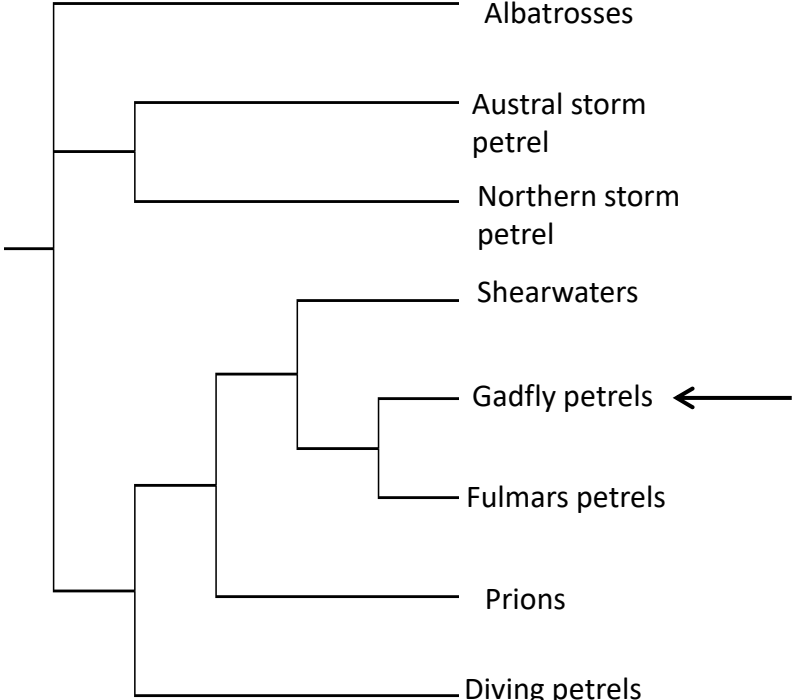
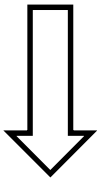
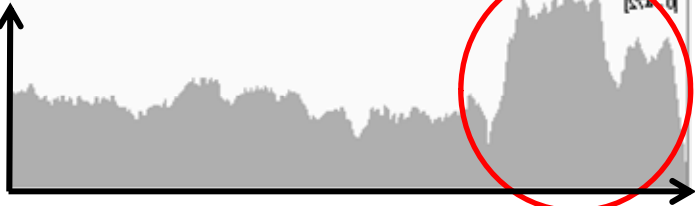
Position on the contig

Cytb-1	ND6-1	CR1	pCytb	ND6-2	CR2	12S
--------	-------	-----	-------	-------	-----	-----



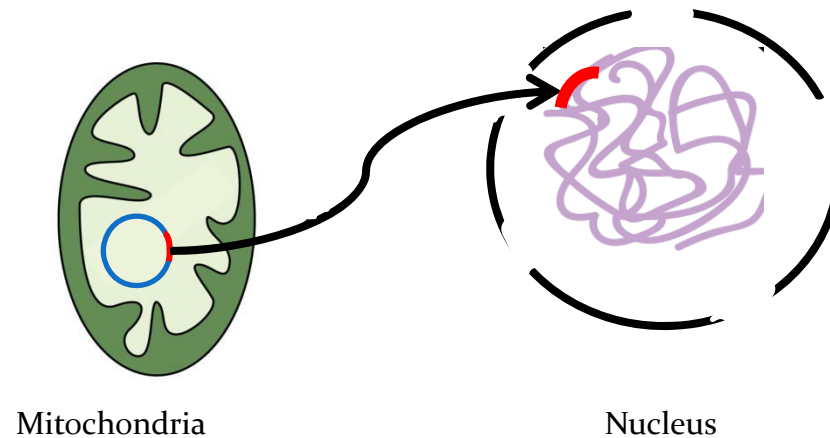
The duplicated region pattern is different between the two Fulmarine genera
 The pattern of *Daption capense* is closer of the pattern of the Albatrosses
 than of the pattern of *Fulmarus*

Pterodroma brevirostris



The duplicated region pattern of *Pterodroma* is different from all the others Procellariiformes, only the Control Region is duplicated

Numts (nuclear mitochondrial DNA segments) are copies of mitochondrial DNA fragments in the nuclear DNA. Sequencing of these two copies could cause the same double-peaks pattern.



We have used the Exonuclease V (RED by NEB) to digest non-linear DNA but to keep circular DNA, then we sequenced mitochondrial markers again
-> double-peaks were still present on CR sequences