unikseq

unique & **conserved** sequence identification
using **multiple**, **complete** genomes

René L Warren
2022

COMPARATIVE GENOMICS

COMPARISON OF COMPLETE GENETIC M.

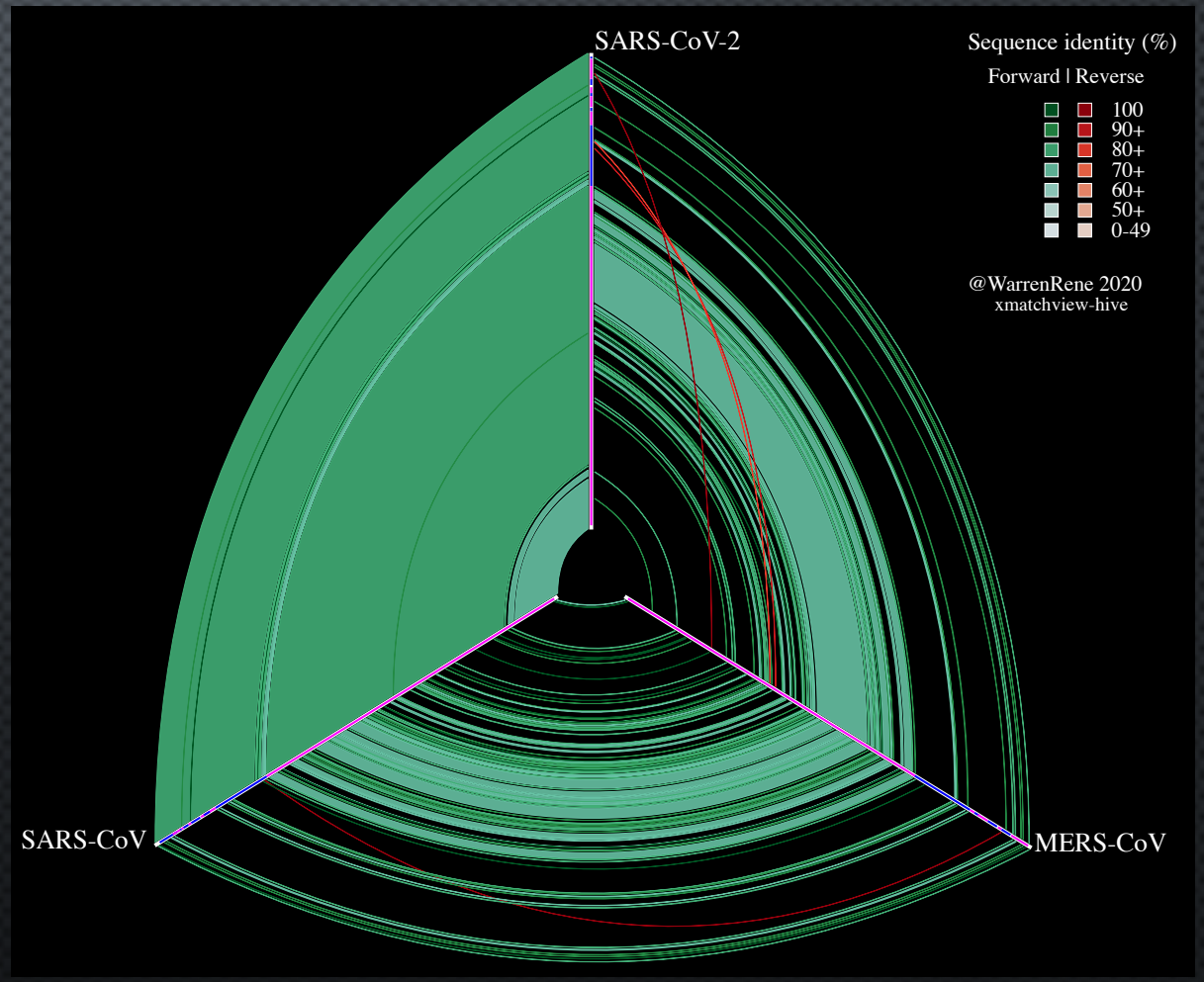
- SPECIES EVOLUTION / GENE FUNCTION
- BASE-LEVEL RESOLUTION

VISUALIZED TO GAIN INSIGHTS

- REGION CONSERVATION / UNIQUENESS
- CO-LINEARITY / STRUCTURE

BIOINFORMATICS AT THE HEART

- SEQUENCE ALIGNMENTS
- PAIRWISE / MULTIPLE SEQUENCE ALIGNMENTS (MSA)



MSA

COMPUTATIONALLY MORE COMPLEX VS. PAIR-WISE ALIGNMENTS + COSTLY + NOT SCALABLE

MSA **KEY TO COMPARATIVE GENOMICS**

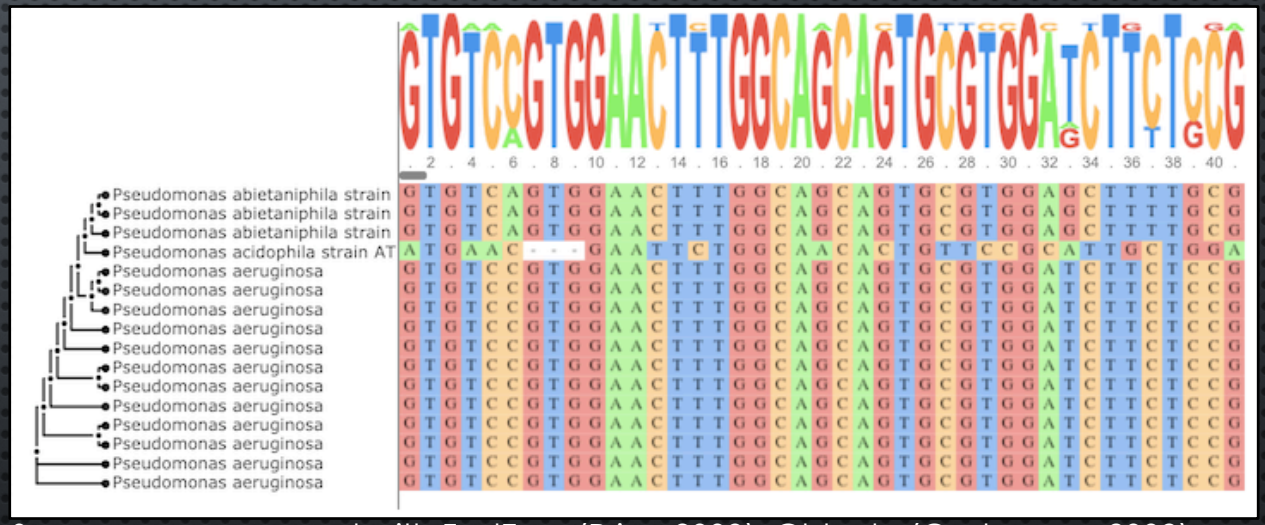
- CLUSTAL, MUSCLE, MAFFT
- USEFUL FOR PCR ASSAY DESIGN

WITH **PCR, NEED VISUAL INSPECTION**

- IDENTIFY UNIQUE / CONSERVED REGIONS

I/O ISSUES

- FINICKY INPUT : STRAND / SEQUENCE START
- OUTPUT INTERPRETABILITY ON KB-STRETCHES, >10S ENTRIES



Sequences compared with FastTree (Price 2009), Gblocks (Castresana 2002), Muscle (Edgar 2004) visualized with MSAviewer/PATRIC (https://www.bv-brc.org/)

POLYMERASE CHAIN REACTION

REVOLUTIONARY

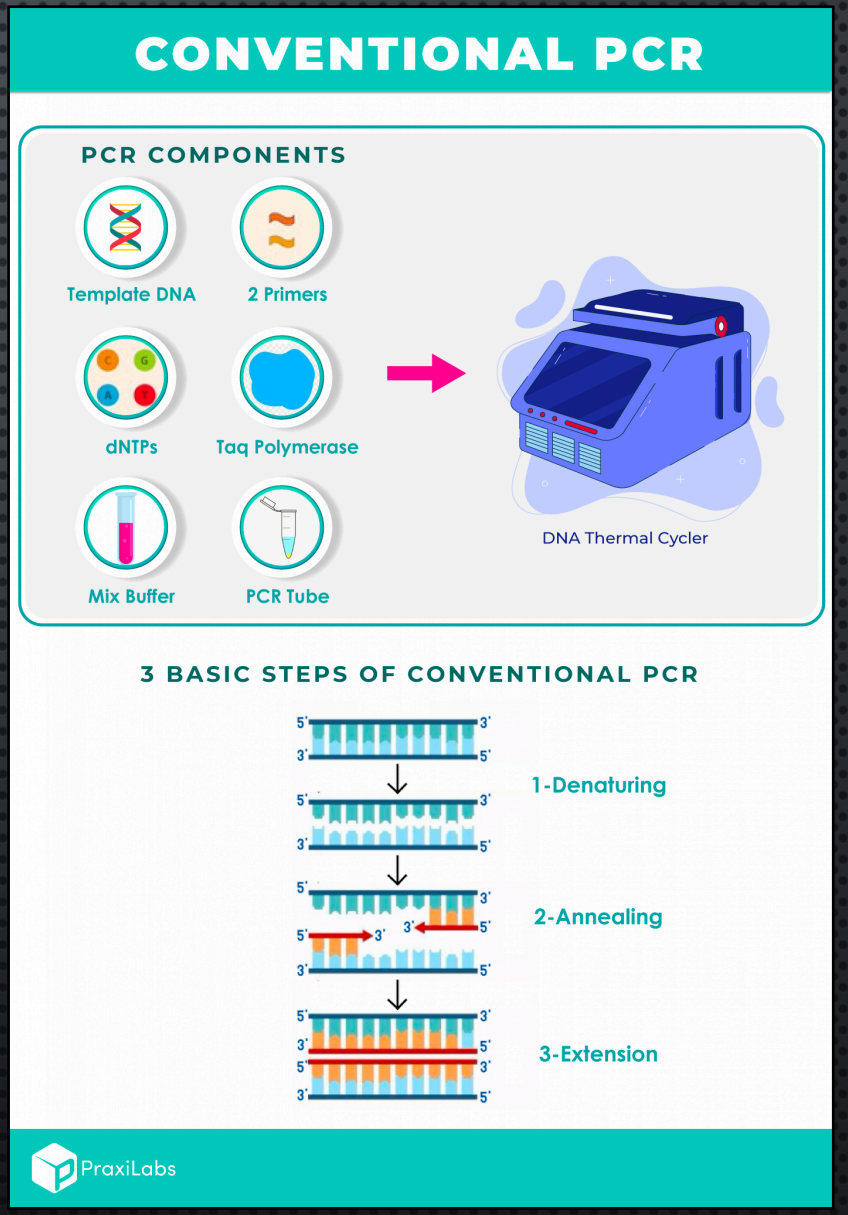
- MOLEC BIO LAB TECHNIQUES FOR
- NUCLEIC ACID AMPLIFICATION, DETECTION, QUANT.

BROAD APPLICATIONS

- FORENSICS, RESEARCH, CLINICAL SETTINGS, BIOLOGY
- SPECIES MONITORING, CONSERVATION, ETC.

EFFECTIVE **PCR : SENSITIVE & SPECIFIC**

- DESIGNING PCR PRIMERS (&QUANTITATIVE QPCR PROBE)
- CHALLENGING + LABOUR + TIME



BIODIVERSITY LOSS

ANIMALS AND PLANTS DESTRUCTION

environmental infographic

- 226 species of mammals
- 181 species of birds
- 168 species of fish
- 77 species of reptiles
- 35 species of amphibians

RED LIST

annual losses from **10** to **100** species

there are about 2000 individuals of the GIANT PANDA

UNDER THE THREAT OF EXTINCTION

PLANTS 18,700 species 7.4 %

REPTILES 1,355 species 21.5 %

BIRDS 924 species 10 %

MAMMALS 414 species 10 %

AMPHIBIANS 48 species 1.1 %

FISH 320 species 1.6 %

EXTINCT ANIMALS

DODO
1 m · 15 kg
Mauritius, Madagascar
1662

PASSENGER PIGEON
40 cm
east of North America
1901

GREAT AUK
80 cm · 5 kg
North Atlantic
1852

STELLER'S SEA COW
7 m · 5 tn
Commander Islands
1768

AUROCHS
1.8 m · 800 kg
Europe, Caucasus
1627

GALAPAGOS TORTOISE
1.8 m · 400 kg
Galapagos Islands
2012 (restored)

QUAGGA
180 cm
South Africa
1883

THYLACINE
Tasmanian wolf
1.3 m · 25 kg
Australia, Tasmania
1936

Moa
3.6 m · 250 kg
New Zealand
1500

CAUSES

- swamp drying
- deforestation
- steppes plowing
- poaching
- dams consruction
- hunting
- grazing
- urban construction
- pollution
- overfishing
- animal skins
- tropical fish trade
- climate change
- death from vehicles
- wild animals trade

credit : Olha1981/Shutterstock

iTrackDNA

non-destructive precision genomics for environmental impact tracking in a global climate change era

University of Victoria

INRS

UQO

Université Laval

Adapted from C. Helbing

iTrackDNA

- 4 years $12M Large Scale Applied Research Project
- To fill critical knowledge gaps & promote eDNA uptake
- Build genomic resources & eDNA capacity (100s kits)
- Develop national standard for targeted eDNA assays
- Promote training / certification
- Influence natural resource management / policy

GenomeCanada

Genome BritishColumbia
Leading ▸ Investing ▸ Connecting

GenomeQuébec

Adapted from C. Helbing

ENVIRONMENTAL DNA / RNA (eDNA / eRNA)

GENETIC MATERIAL EXTRACTED FROM ENVIRONMENTAL SAMPLES

RAPID
ACCURATE
COST-EFFECTIVE
NON-INVASIVE
ACTIONABLE

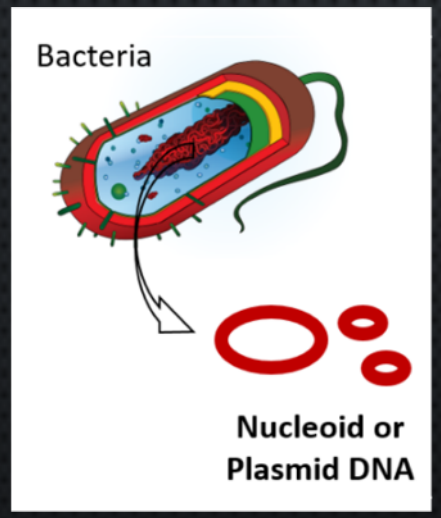
BARRIERS TO UPTAKE
NEED STANDARDS
NEED GENOMIC RESOURCES

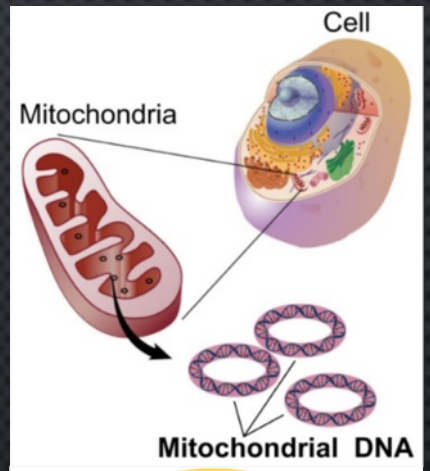
Adapted from C. Helbing

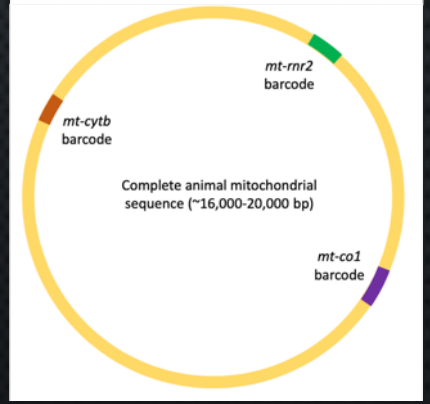
eDNA

microbial

bulk sample
macrobial
(shed)







ANIMAL MITOGENOMES

ABUNDANT, TAXA SPECIFICITY & SEQUENCE CONSERVATION

Adapted from C. Helbing

UTILITY OF eDNA

DETECTION / ENFORCEMENT OF TRAFFICKED SPECIES

CITES : CONVENTION ON INTERNATIONAL TRADE IN ENDANGERED SPECIES OF WILD FAUNA AND FLORA

WAPTR : WILD ANIMAL AND PLANT TRADE REGULATIONS

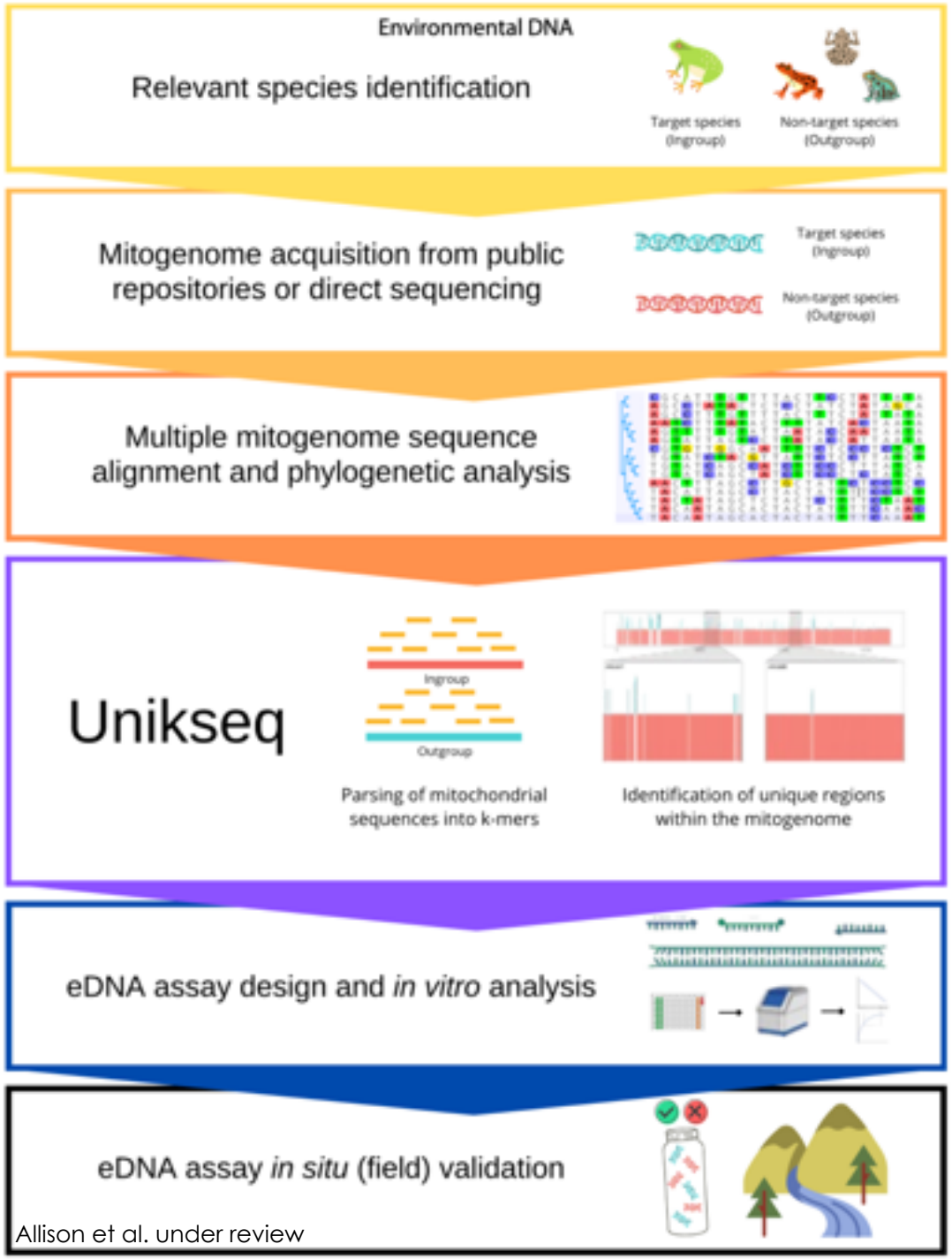
ENVIRONMENTAL/ECOSYSTEM MONITORING (E.G. IMPACT OF OIL SANDS, CLIMATE, ETC.)





CITES, wikipedia

Adapted from C. Helbing



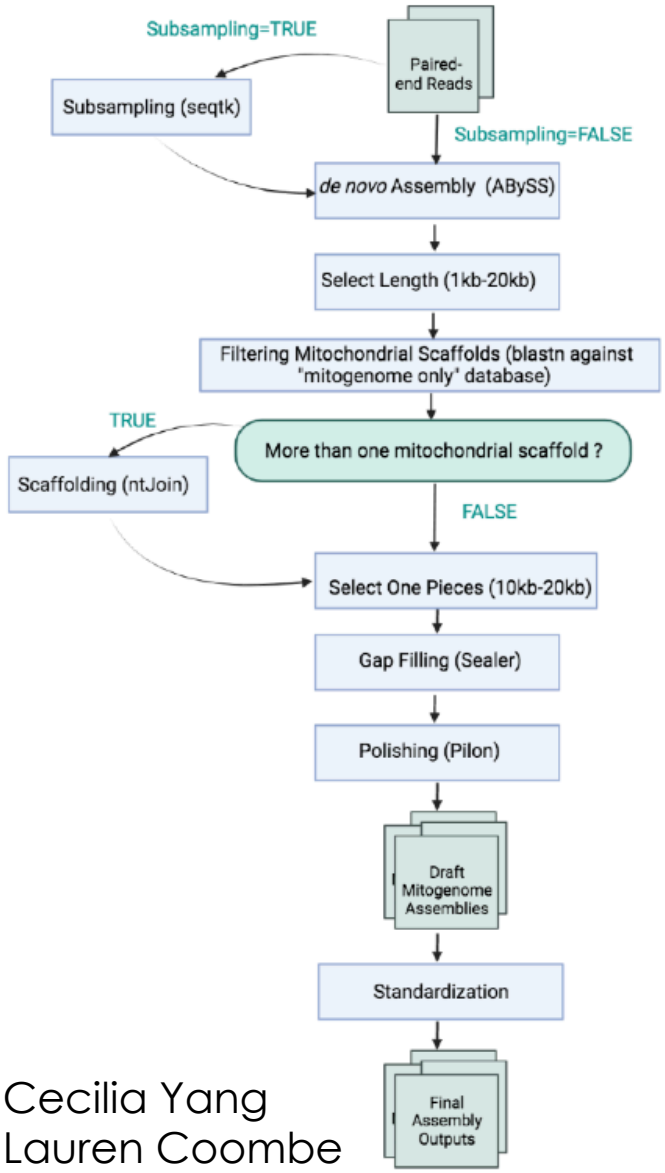
eDNA ASSAY

multi-step workflow

+collections of complete mitogenomes [sequences]

develop high-quality eDNA assays / kits

BUILDING MITOGENOME RESOURCES



Cecilia Yang
Lauren Coombe


Sablefish (4)


White sharks (9)


Rockfish (57)


Thornyhead (4)


Greater sturgeons (9)

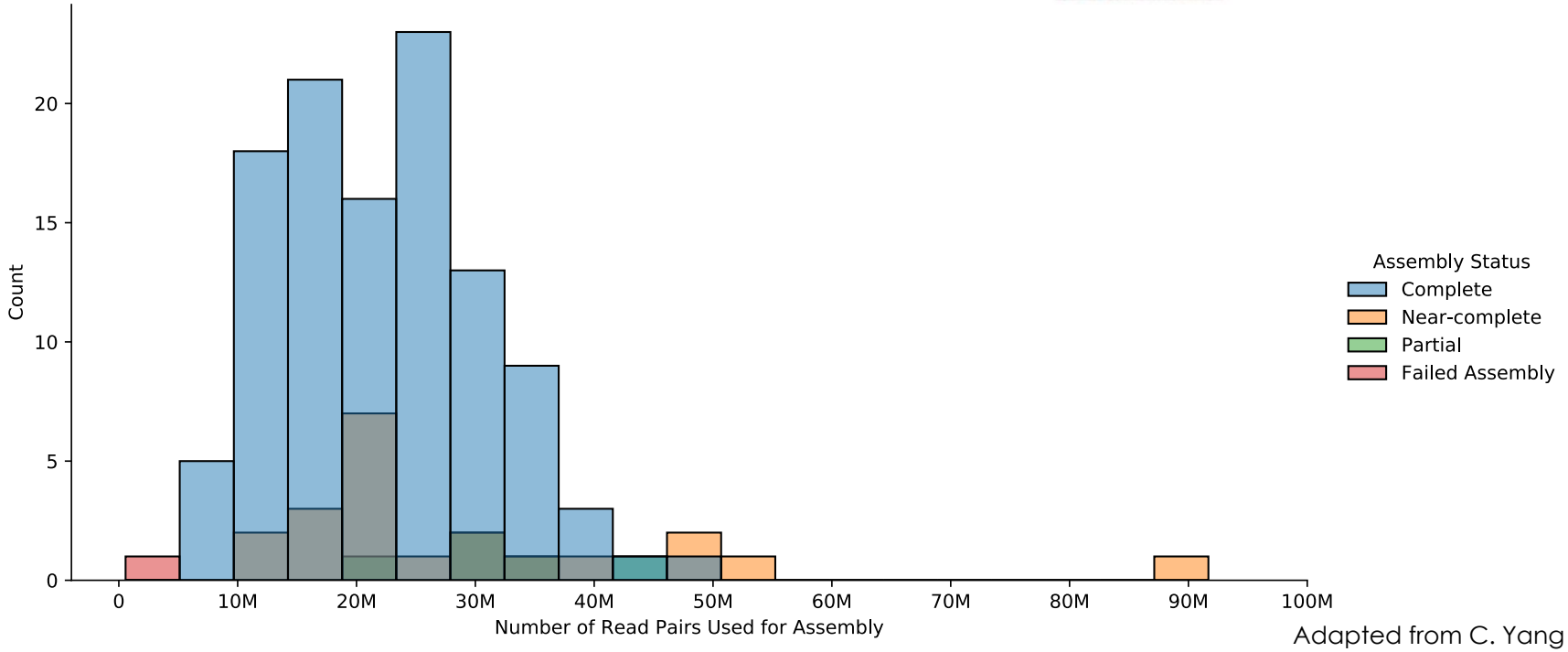

Longfin smelt (9)


Oyster (10)

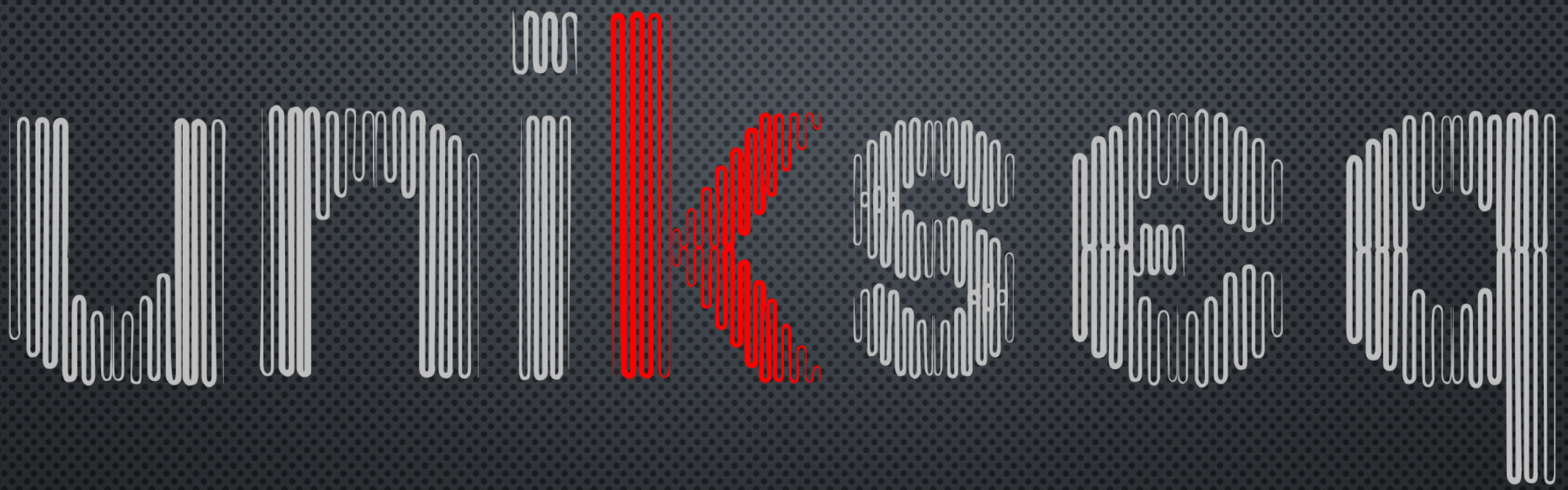

Brown frog (2)


Long-tailed shrew (24)


Worm (9)



Adapted from C. Yang

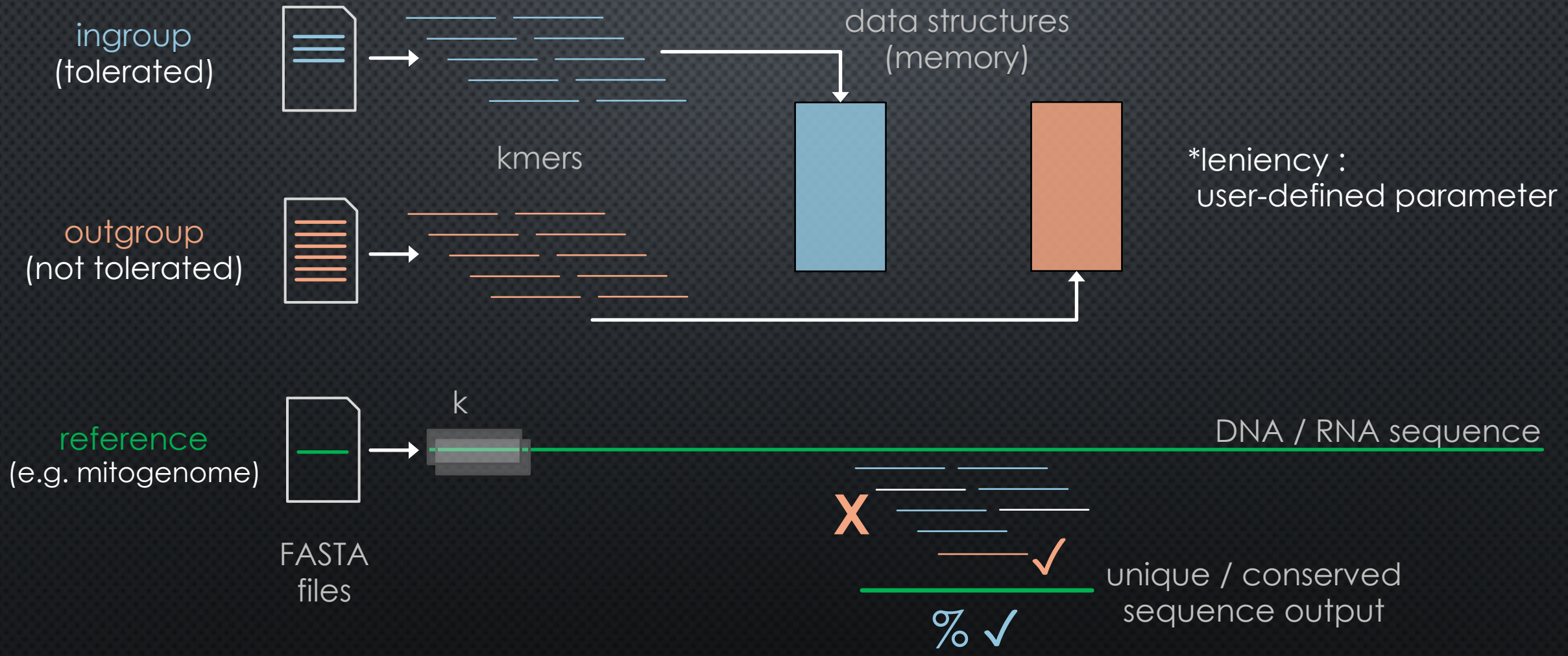


unique / conserved region identification in
DNA / RNA sequences using a kmer-based approach

https://github.com/bcgsc/unikseq

UNIQUE REGION IDENTIFICATION

IN A REFERENCE, TOLERATED IN INGROUP, NOT* IN OUTGROUP



GREAT WHITE SHARK : USE CASE

- VULNERABLE SPECIES

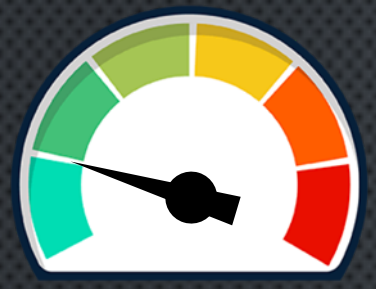
  ENVIRONMENT MONITORING : CONSERVATION EFFORT INSIGHTS

- FIND MITOGENOME REGIONS UNIQUE &

  HIGHLY (100%) CONSERVED IN GREAT WHITES

- SPECIFIC qPCR-BASED eDNA ASSAY DESIGN

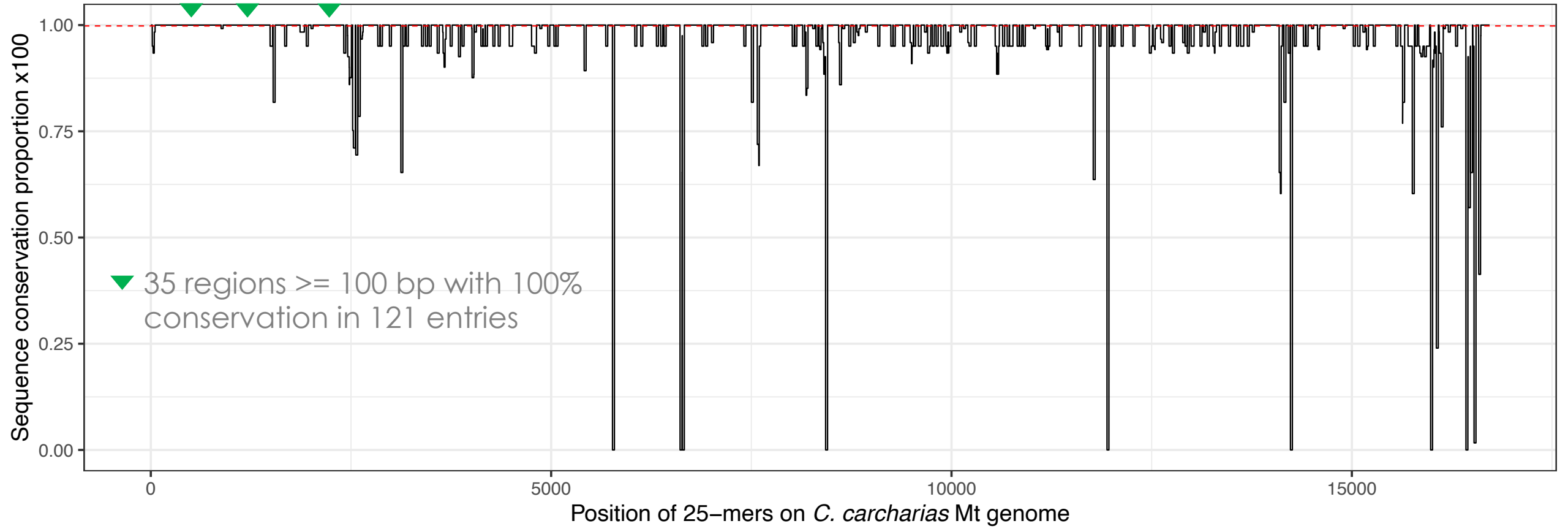
121 great white shark mitogenome [isolates] (ingroup)
856 fish mitogenomes (outgroup)



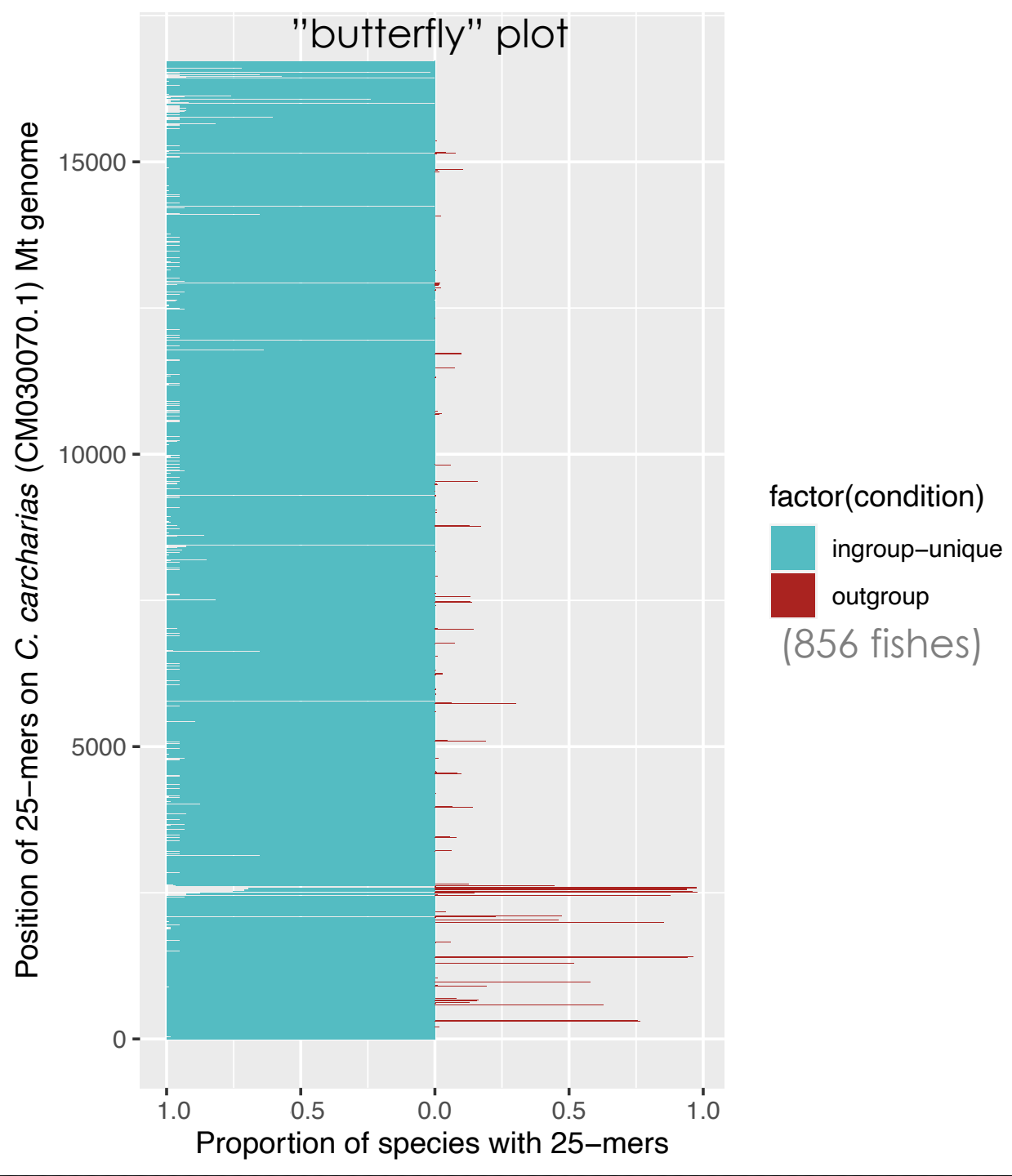
SEQUENCE CONSERVATION

ACROSS 121 WHITE SHARK MTG (ISOLATES)

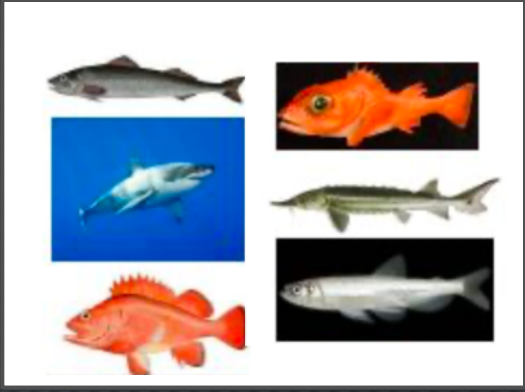
comparing 2,000,000 bp



clustal 24h52m MUSCLE 28m27s MAFFT 14s **unikseq 3.2s**



VS



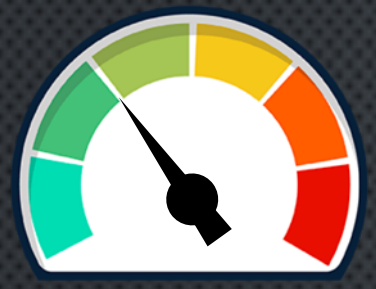
UNIQUE REGIONS

- MANY REGIONS UNIQUE IN GREAT WHITE REF
  43 UNIQUE REGIONS (100 BP+), 12.9 KB
- IDEAL FOR DESIGNING **SPECIFIC** qPCR

left (blue) : sequence uniqueness / conservation
right (red) : reference kmers in outgroup

UNIQUE + CONSERVED REGIONS

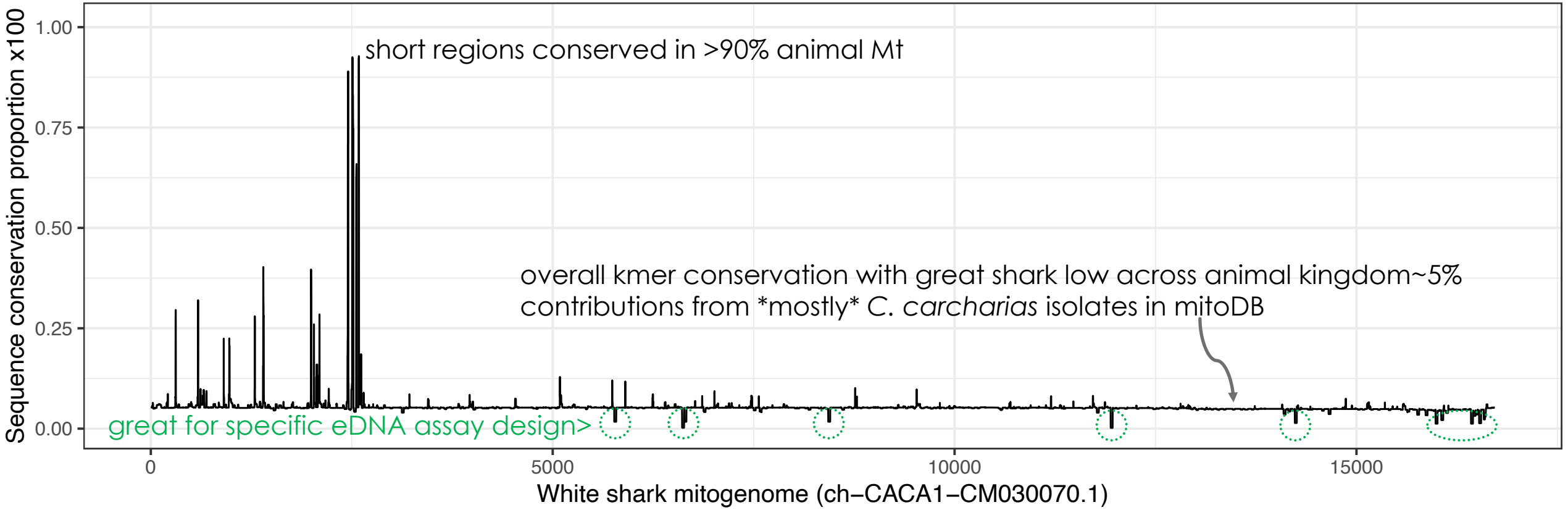
>ch-CACA1-CM030070.1region11205-11314_size110_propspcIN98.9_propunivsOUT100.0_avgOUTentries0.0
ccgtaagccacataggcctggtagcaggagcaatcctaatccaaacaccatgaagctttgcaggagcaattacactaataattgcccacggcctgatttcatccgcccta
>ch-CACA1-CM030070.1region11317-11476_size160_propspcIN98.9_propunivsOUT100.0_avgOUTentries0.0
ctgcttagccaacactaactacgagcgaatccacagccgaacaatactcctagcccgaggcATACAAATTATTTTTCCACTAATAGCTACCTGATGATTCCTTGCCATCCTAGCCAACCTTGCCCTTCCC
CCCTCTCCCAATTTTATAGGAGAACTCCTT
>ch-CACA1-CM030070.1region11480-11719_size240_propspcIN99.3_propunivsOUT100.0_avgOUTentries0.0
ATTACCTCCTTATTTAACTGATCCAACTGAACTATTACCCTCTCGGGCCTTGGAGTATTAATCACCGCTTCCTACTCCCTTTATATATTCCTAATAATTCAACGCGGACCAACccccctaccatatc
ttatcattaaatccaagccacACACGAGAACACCTCCTCCTGAGCCTCCACCTCATGCCCGTCCTACTTCTAATATTTAAACCAGAACTTATCTGAGGCTGAACACTCTGTGCT
>ch-CACA1-CM030070.1region11724-12333_size610_propspcIN93.4_propunivsOUT100.0_avgOUTentries0.0
GTTTAACCAAAACATTAGATTGTGGTTCTAAAAATAAAAGTTAAAAccttttttaaccaccgagagaggtcggggacatgaagaactgctaattcttctcatcatggctcaaatccatgactcactcagcttctgaaa
gatatcagtaatctattggtcttaggaaccaaaaaccccttggtgcaactccaagcaaaagctatgaatactatctttaactcatccttcctcctaatctttgttatcctcctcttcccactaataacctcattaaatcccaaagaact
cactcctaactgggcctcctcctacgcaaaaacagctgtaaaaatctccttcttcattagccttatcccactatccatttttctagaccaaggtttagagtcaatcataaccaactacaactgaatcaacattggaCCTTTCGA
TATTAATATAAGCTTCAAATTTGATACATACTCTGTTCTATTTACCCCTGTGGCCCTCTACGTTACTTGATCTATCCTTGAATTTGCCCTATGATACATATACTCCGACCCAAAcatcaaccgcttcttt
aaatatcttCTACTCTTCCTAATCTCAATAATCATTTTAGTGACTGCCAAC
>ch-CACA1-CM030070.1region12339-12756_size418_propspcIN98.4_propunivsOUT100.0_avgOUTentries0.0
ATTCCAACTGTTCATTGGCTGAGAGGGAGTTGGAATCATATCATTTCTCCTCATTGGTTGATGACATAGTCGAACAGACGCCAACACAGCCGCCCTACAAGCTGTAATCTATAACCGAG
TAGGAGATATTGGcctaattcttagcatagcttgactggctataaatttaaactcctgagaaattcaacaactattcatcttatccaaagacataaacttaaccctacctctccttggtcttgtcctagccgcagctggaaaat
ccgcacaattcggccttcacccatgacttccatcggccatagaaggacccacaccagtctccgccttactccactccagcacaatagttgtagccggtatcttcctcctaattcgccttcacccattaatccacaataatcaa
ttaattctaaca
>ch-CACA1-CM030070.1region12936-13134_size199_propspcIN97.5_propunivsOUT100.0_avgOUTentries0.0
taaagcaatacttttcctctgttccggatccattatccatagcctcaatgacgaacaagatatccgcaaaataggaggtcttcacaaactcctgccattcacctcctcctccttaactattggaagcctagccctcataggcatg
cccttcttatcaggcttcttttcaaaagatgccatcatcgaagccataaacact
>ch-CACA1-CM030070.1region13136-13248_size113_propspcIN98.5_propunivsOUT100.0_avgOUTentries0.0
ctcacctcaacgcctgagcccttatccttaccctaattgcaacatcatttacagccatttacagcctccgcctcatttttttcacattaataaattttccacggtttaattca
>ch-CACA1-CM030070.1region13267-14066_size800_propspcIN98.9_propunivsOUT100.0_avgOUTentries0.0
aacaaccccctaattattaaccccatcaaacgcctagcttacggaagtatcctgtctggcctcattatcacatccaacataatccctacaaaaacccaaattatgactataattcccctactaaaactctccgccctactaatt
actattgctggccttctactagccttagaactagcaaacctaactaactctcagcttaaaataaccccccaccctctatccccatcacttctcaaatatattgggatactttccccaaattatccaccgcctcctgcccaaaatta
acctatcctgagcccaacatgtttccacccacctaattgaccaaacatggtctgaaaaaattggaccaaaaagtgccttcattcaacaaacccctctaattaaattatccacccaacctcaacaaggctatattaaagttta
cctaacactacttttcctcacactagccttagccatactcactacattaacctaaccacacgtaatgtccctcatgctagacctCGAGTCAACTCCAACACCACAAACAAAGTCAATAATAACACCCACCC
ACTTAAAACTAACAACCACCCCCCATCCTCATAAAGCAAAGCCACCCCCACAAAATCCCCACGAGTTATCTCCATATTGCTCAACTCCTCTACCCCTGACCAATCCAACTCAAATC
ACTCTACCATAAAATAATTACCAACAAAAAATAACATTACTAAATAAAAACCAACATACAACAAAACAGATCAATTACCCCATGACTCAGGATATGGCTCAGCAGCAAGAGCTGCCG



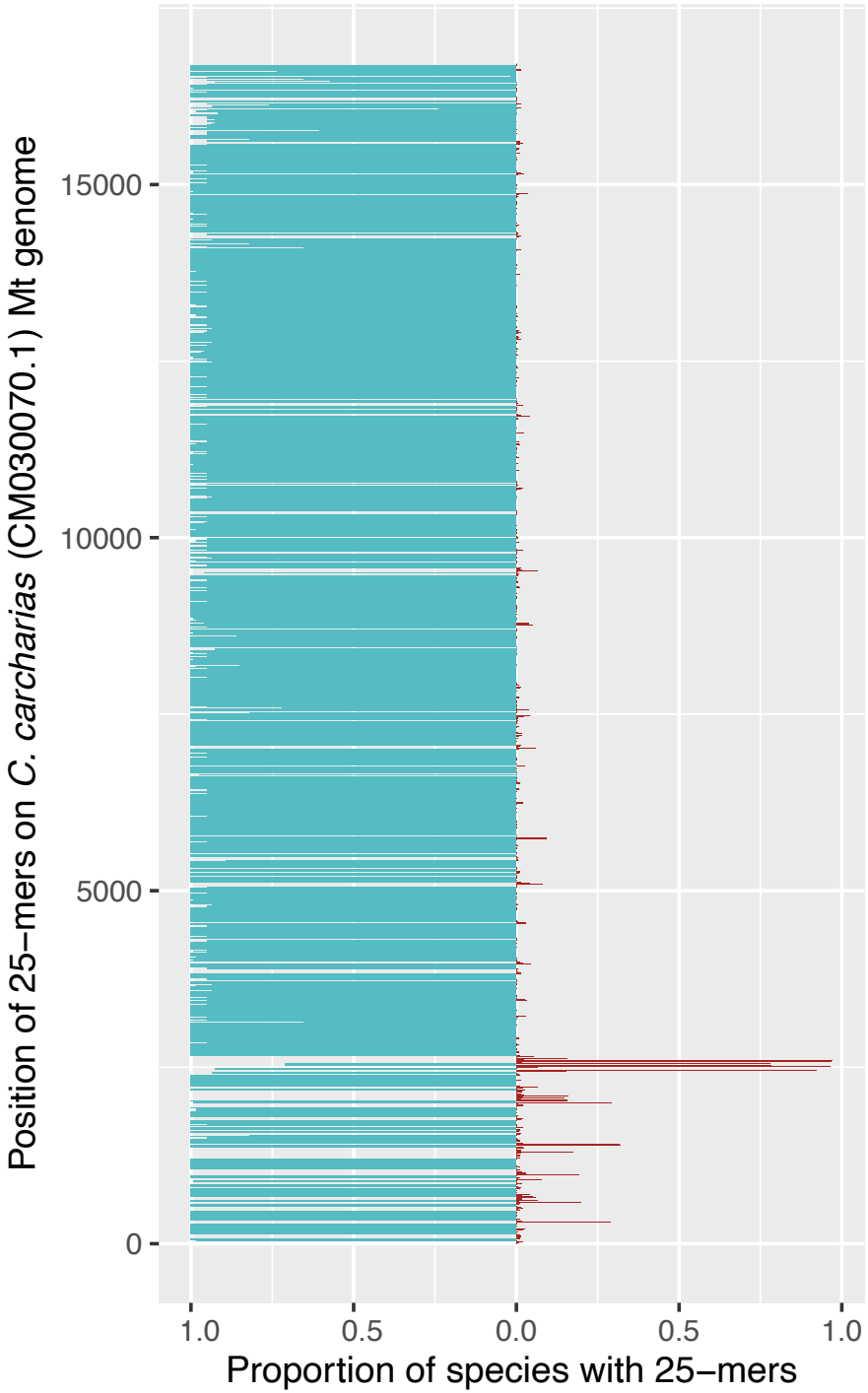
SEQUENCE CONSERVATION

ACROSS 4,000 ANIMAL MITOGENOMES

comparing 65,000,000 bp



clustal >3weeks MUSCLE error MAFFT 9h50m, **4.4GB RAM** **unikseq 2m**, 9.0GB RAM

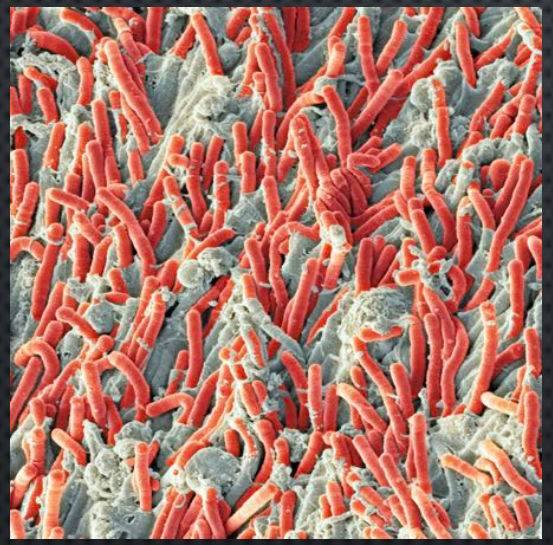


VS

UNIQUE REGIONS

- LARGER OUTGROUP ~ RESTRICTIONS FOR UNIQUE REGION IDENTIFICATION
- 29 UNIQUE REGIONS (100 BP+), 5.2 KB
  3 : 100% AVG INGROUP CONSERVATION, 461 BP
- GREAT WHITE SHARK MITOGENOMES
  HARBOR MANY UNIQUE REGIONS
  CONSERVED ACROSS ISOLATES

FUSOBACTERIUM NUCLEATUM



"Dental plaque bacteria hitch a ride in blood to fortify colon cancers" — *Cosmos mag.*

› Genome Res. 2012 Feb;22(2):299-306. doi: 10.1101/gr.126516.111. Epub 2011 Oct 18.

Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma

Mauro Castellarin [1], René L Warren, J Douglas Freeman, Lisa Dreolini, Martin Krzywinski, Jaclyn Strauss, Rebecca Barnes, Peter Watson, Emma Allen-Vercoe, Richard A Moore, Robert A Holt

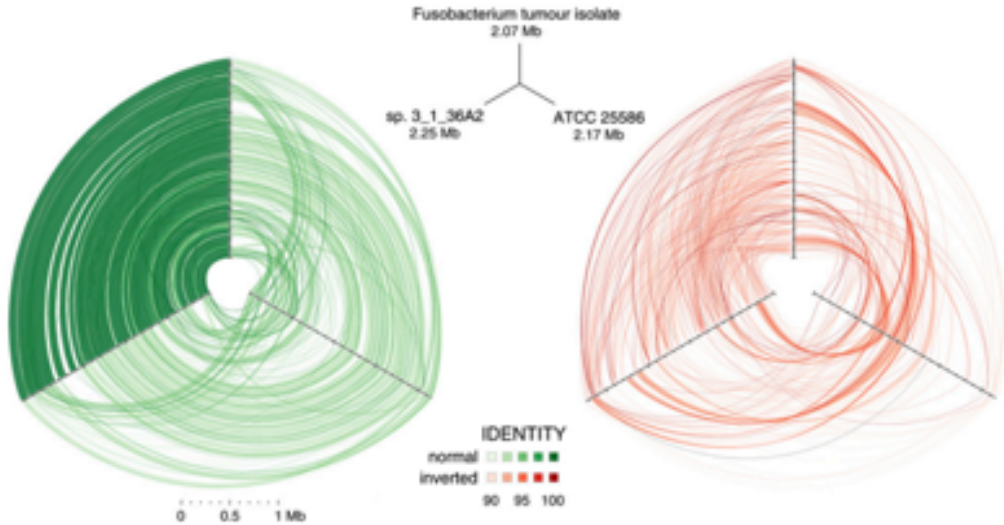


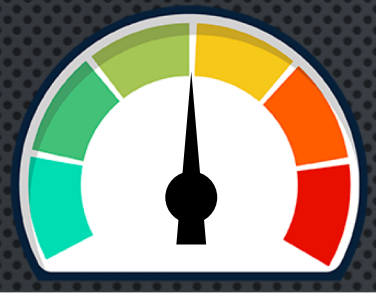
Figure 3. Hive plots showing alignment of three *Fusobacterium* genomes. Approximately 32 million high-quality WGS Illumina HiSeq reads (≥99 consecutive Q30 bases) from *Fusobacterium* tumor isolate CC53 were assembled with SSAKE (v3.7, default options) into 379 contigs. The contigs were aligned using cross_match (-minmatch 29 -minscore 59 -masklevel 101) to the complete *F. nucleatum* susb. nucleatum ATCC 25586 genome and, independently, to the 12-contig HMP *Fusobacterium* sp. 3_1_36A2 assembly, respectively; and ordered/oriented based on the highest identity to the latter sequence. Three-way cross_match (http://www.phrap.org) alignments between each *Fusobacterium* genome were performed and represented visually using hive plots (http://www.hiveplot.com). For each, the *top*, *left*, and *right* axes are proportional to genome size and represent the *Fusobacterium* tumor

2.2 MBP CIRCULAR GENOME

ASSOCIATED WITH CRC POLYPS

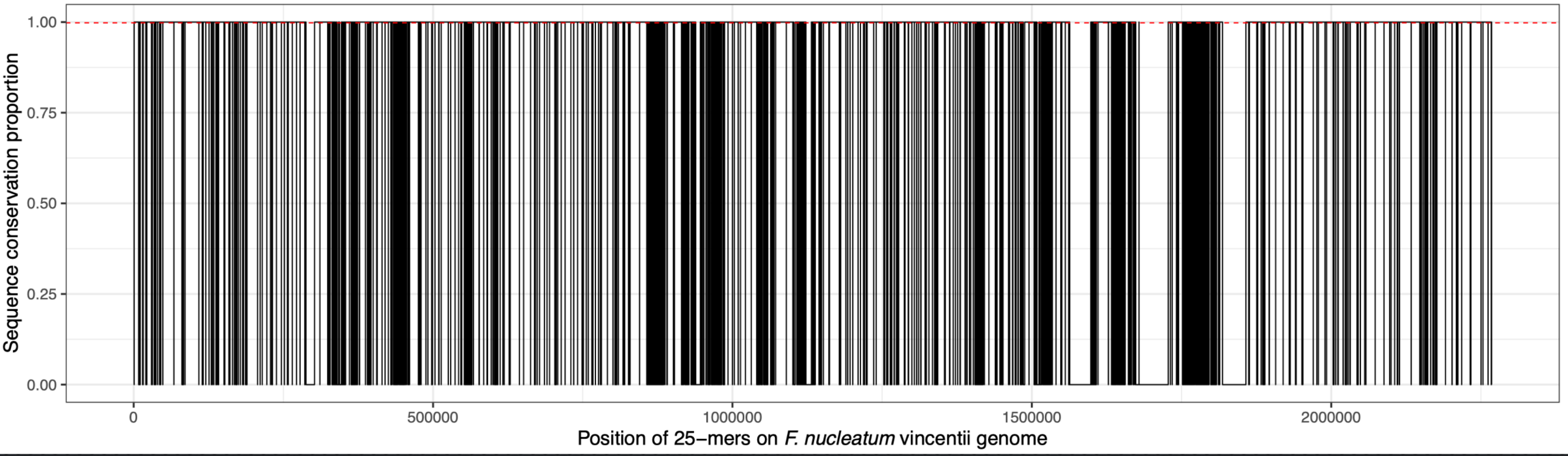
TUMOUR ISOLATE SEQUENCED 2012 ASSEMBLED WITH SSAKE : **CC53**

SIMILAR TO *F. N.* VINCENTII AND ATCC (ORAL STRAIN)



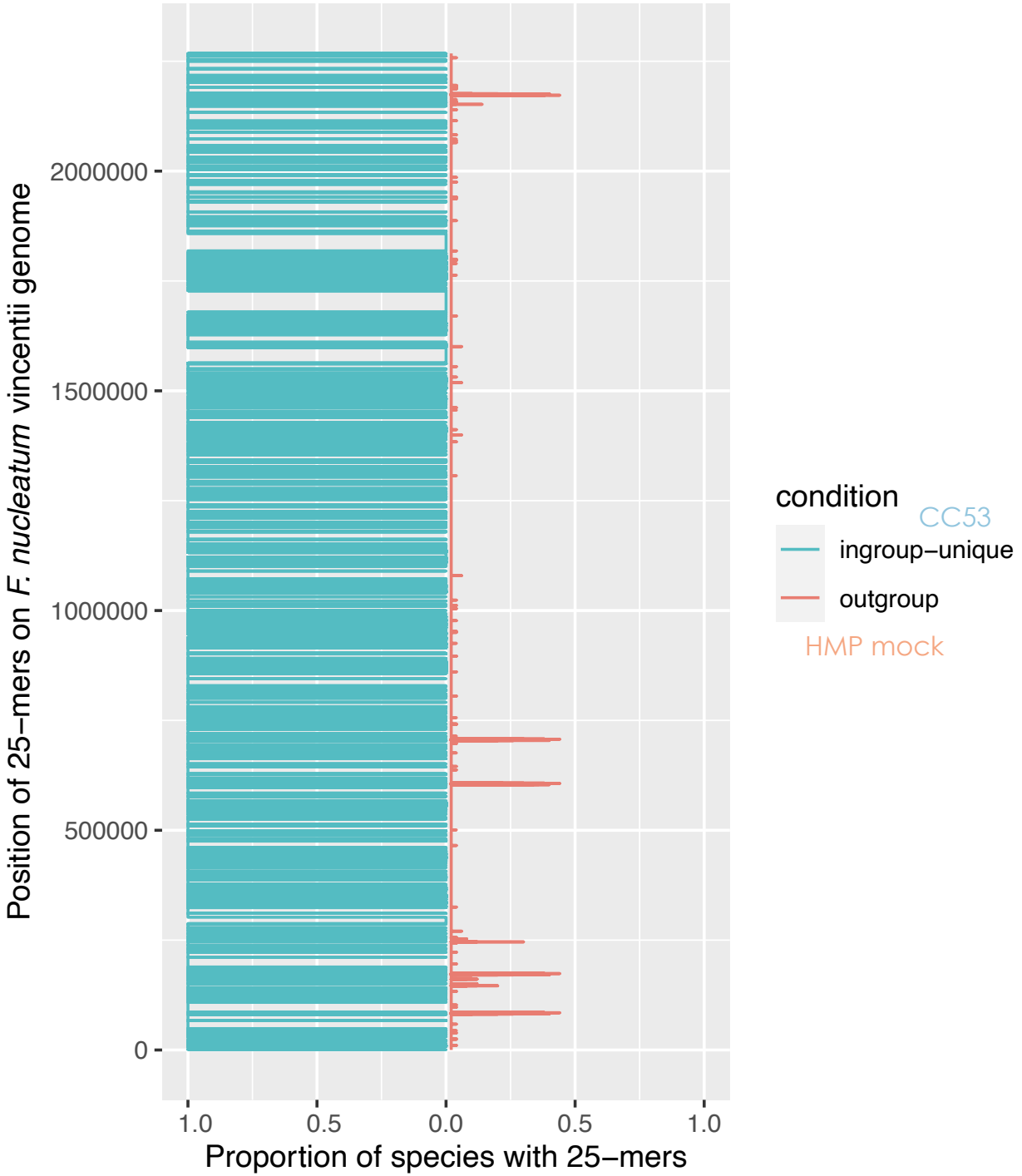
CONSERVATION *FN. VINCENTII* VS CC53

comparing 90,000,000 bp



1 REFERENCE (*F. NUCLEATUM* VINCENTII FROM HMP)
1 INGROUP ENTRY (*F. NUCLEATUM* CC53 CRC ISOLATE)
50 OUTGROUP ENTRIES (HMP MOCK + FUSOBACTERIUM ATCC25586)

clustal >3weeks MUSCLE error MAFFT >3weeks **unikseq 5m, 54GB RAM**



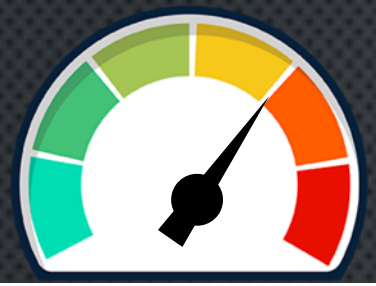
UNIQUE REGIONS

UNIQUE+CONSERVED WITH VINCENTII

n	Min	N50	Max	Sum
3,357	100	188	6,503	646,774

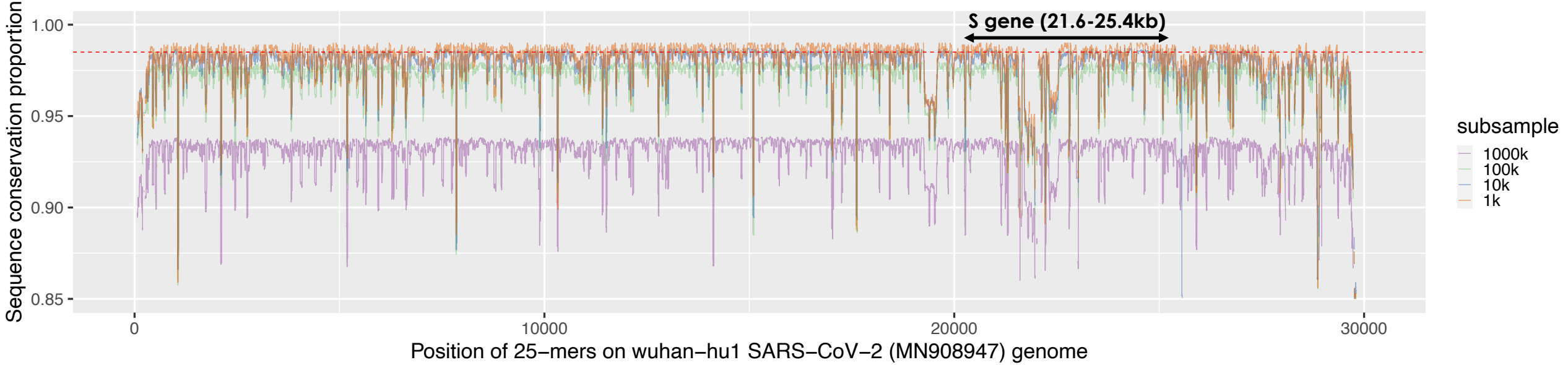
UNIQUE TO CC53

n	Min	N50	Max	Sum
81	100	457	3,570	21,720



1 MILLION SARS-CoV-2 GENOMES

comparing 30,000,000,000 bp

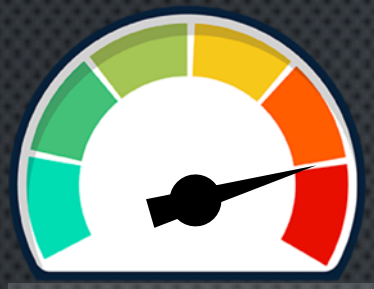


1 CPU @ hpcg02

Subsample (k)	Run time (hh:mm:ss)	Memory (GB)
1	00:00:55	3.2
10	00:12:28	35.9
100	02:02:11	323.5
1000	11:27:02	1238.5

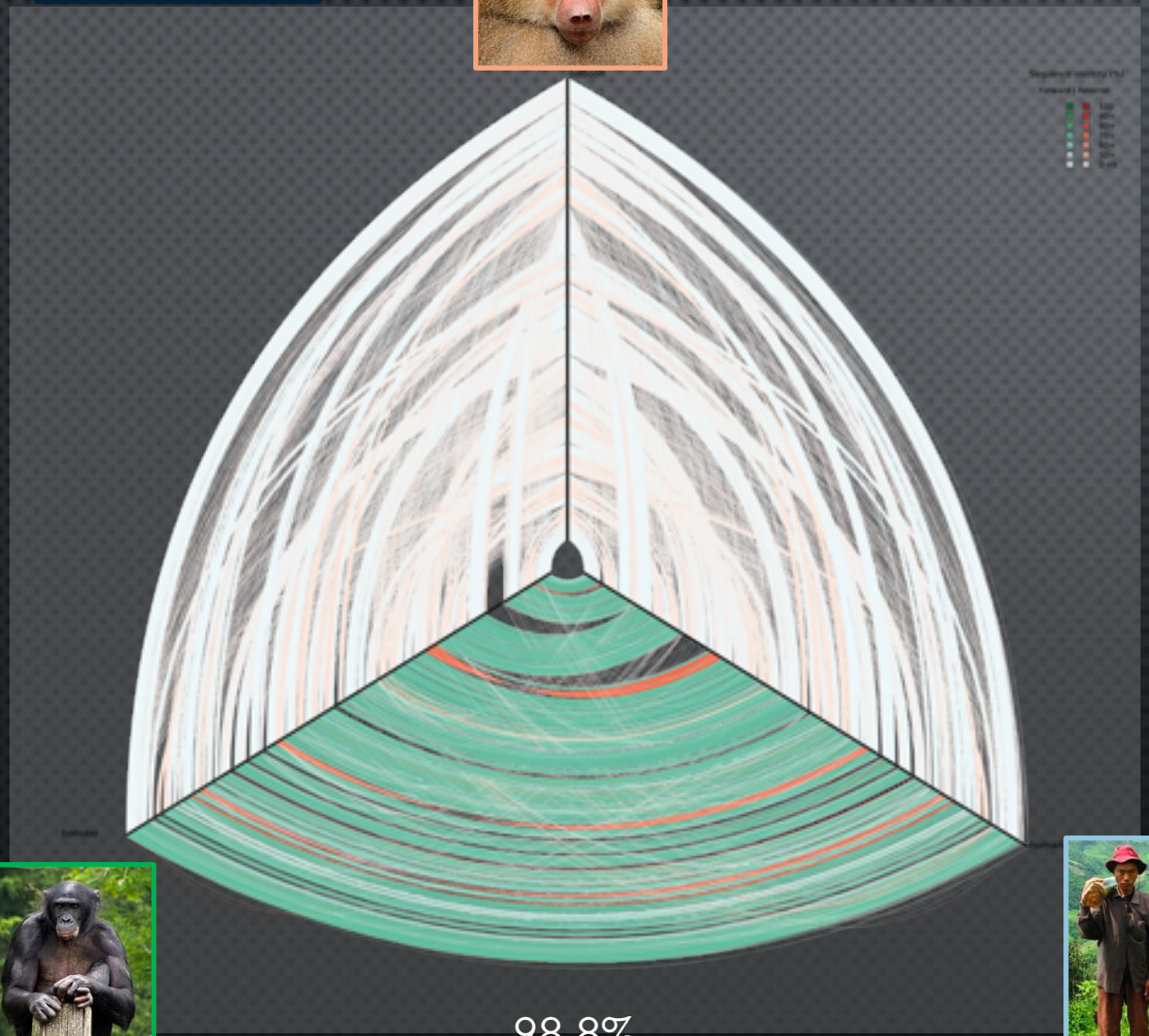
FURTHER BENEFIT FROM :

1) COUNTING BLOOM FILTERS
2) PARALLELISM



BONOBO (& HUMAN) VS BABOON GENOMES

comparing 9,000,000,000 bp



98.8%

USING TWO (INGROUP/OUTGROUP) BLOOM FILTERS

UNIQUE+CONSERVED

n	Min	N50	Max	Sum
415,629	100	123	913	52.9e6 (1.8%)

Bloom filters : 45' (16GB RAM) unikseq : 2h wall clock (21 GB RAM)

NEANDERTHAL (& DENISOVAN) VS HUMAN GENOMES

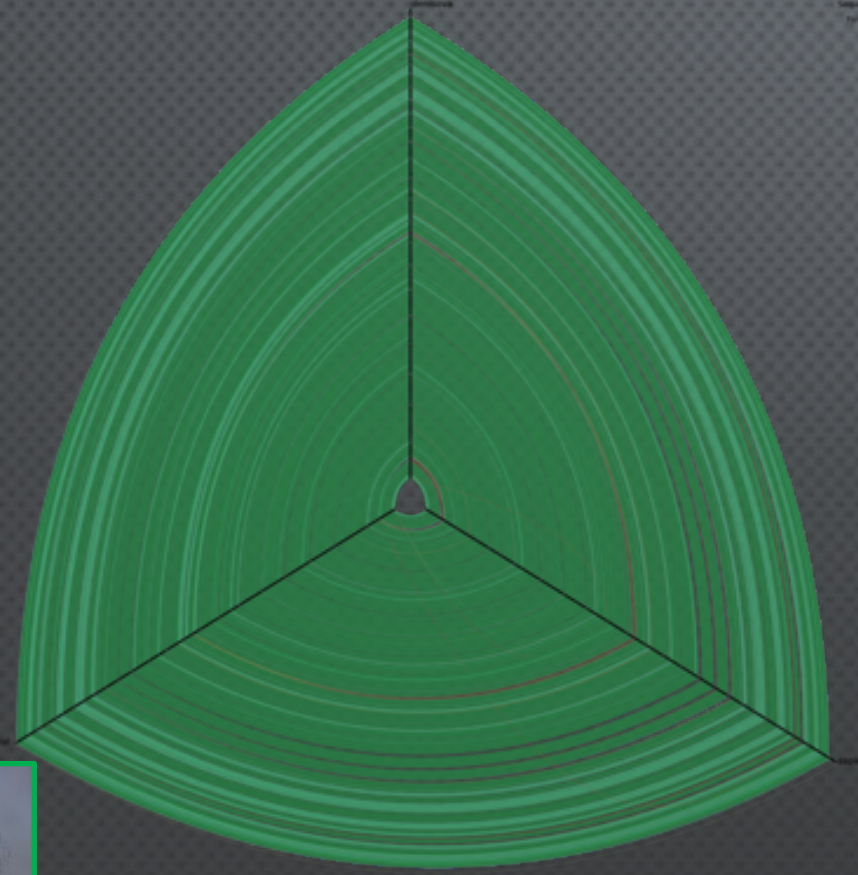
comparing 15,000,000,000 bp

UNIQUE+CONSERVED (100%)

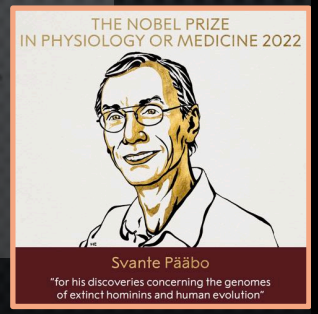
n	Min	N50	Max	Sum
6,314	100	682	5,407	2.5e6 (0.09%)

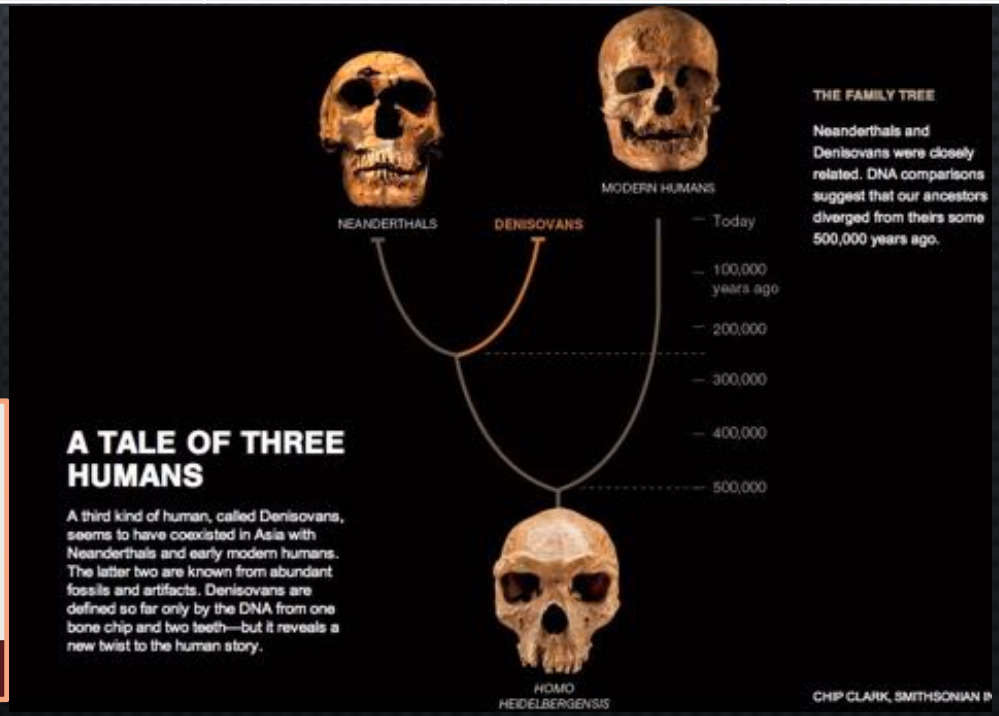
Bloom filters : 44' (16GB RAM) unikseq : 2h wall clock (23 GB RAM)

%	HG02055	HG01243	NA24385
Neanderthal	0.02	0.01	0.02
Denisovan	0.01	0.01	0.01



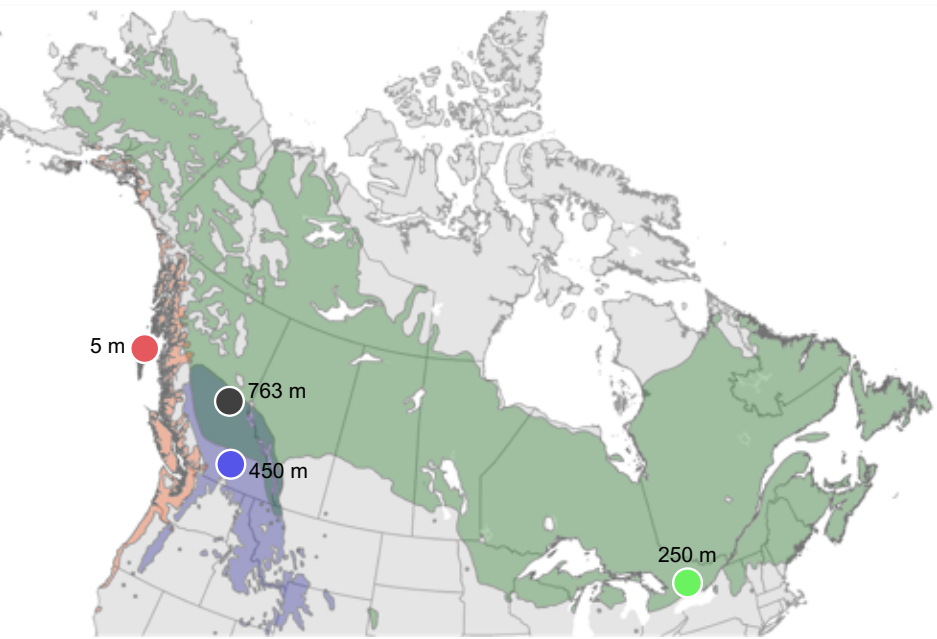
99.5%

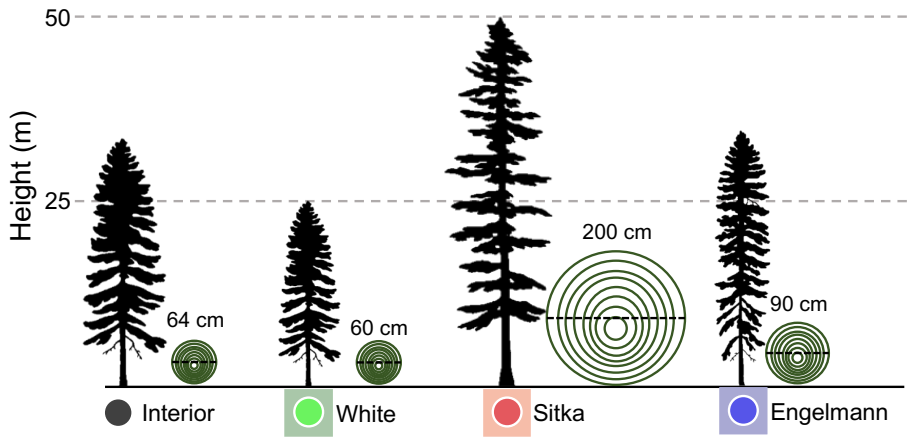




White (& BC interior) vs Sitka+Engelmann spruce genomes

comparing 80,000,000,000 bp

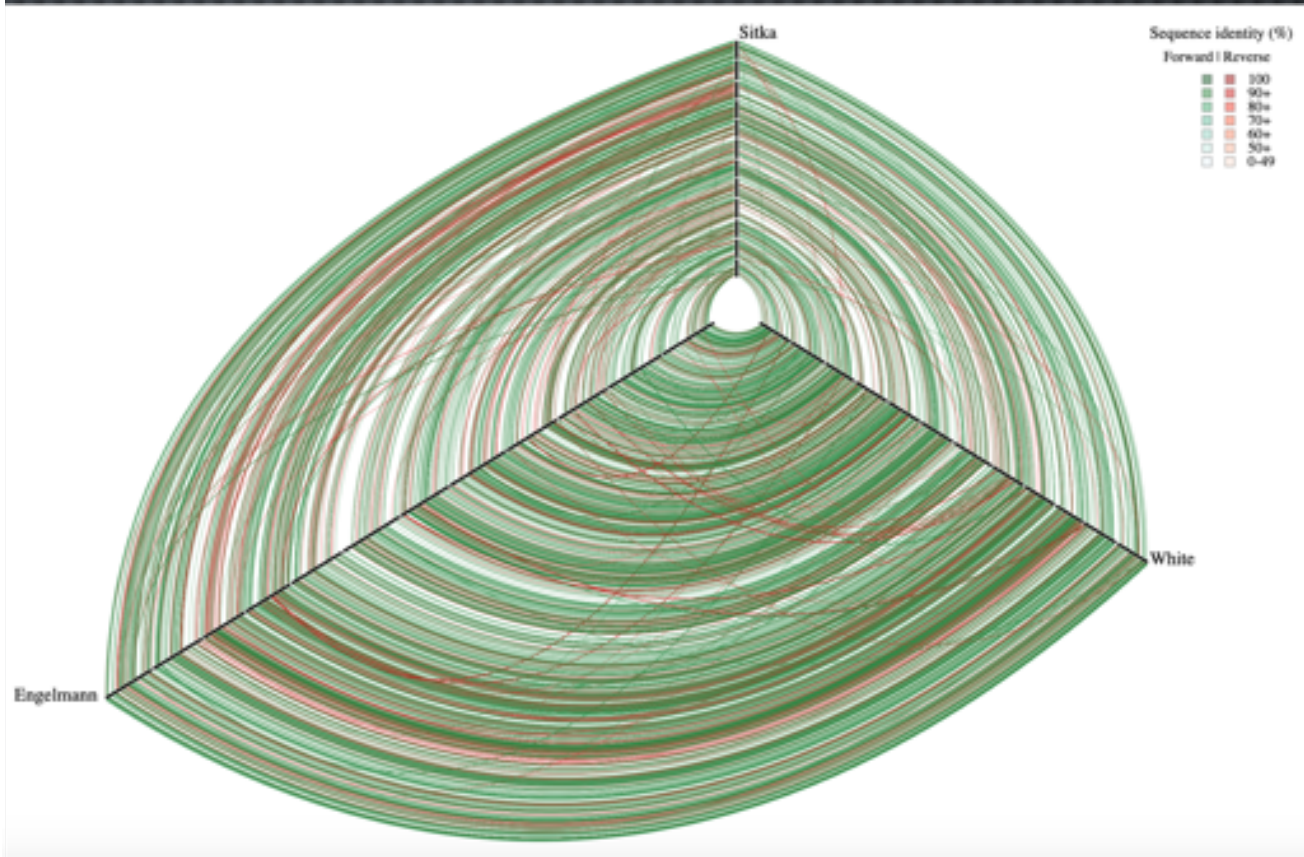




UNIQUE+CONSERVED (100%)

n	Min	N50	Max	Sum
59,565	100	129	1,642	8.3e6 (0.04%)

Bloom filters : 9h (86GB RAM) unikseq : 14h07 (124GB RAM)



SUMMARY

NOVEL UTILITY FOR COMPARATIVE GENOMICS

- IDENTIFY UNIQUE / CONSERVED DNA / RNA REGIONS
- K-MER BASED : NO MSA / PAIR-WISE ALIGNMENTS REQUIRED
- FAST & SCALABLE
- NO STRICT INPUT : GENOMES / CONTIGS / READS

BROAD UTILITY

- SPECIFIC eDNA ASSAY DESIGN
- CLINICAL / DIAGNOSTIC TESTS

github.com/bcgsc/unikseq

ACKNOWLEDGEMENTS

GSC
Inanc Birol
Lauren Coombe
Cecilia Yang

UVic
Caren Helbing
Michael Allison
Neha Acharya-Patel
Louie Lopez

GenomeCanada

Genome BritishColumbia
Leading ▸ Investing ▸ Connecting

GenomeQuébec

unikseq

github.com/bcgsc/unikseq

QUESTIONS?

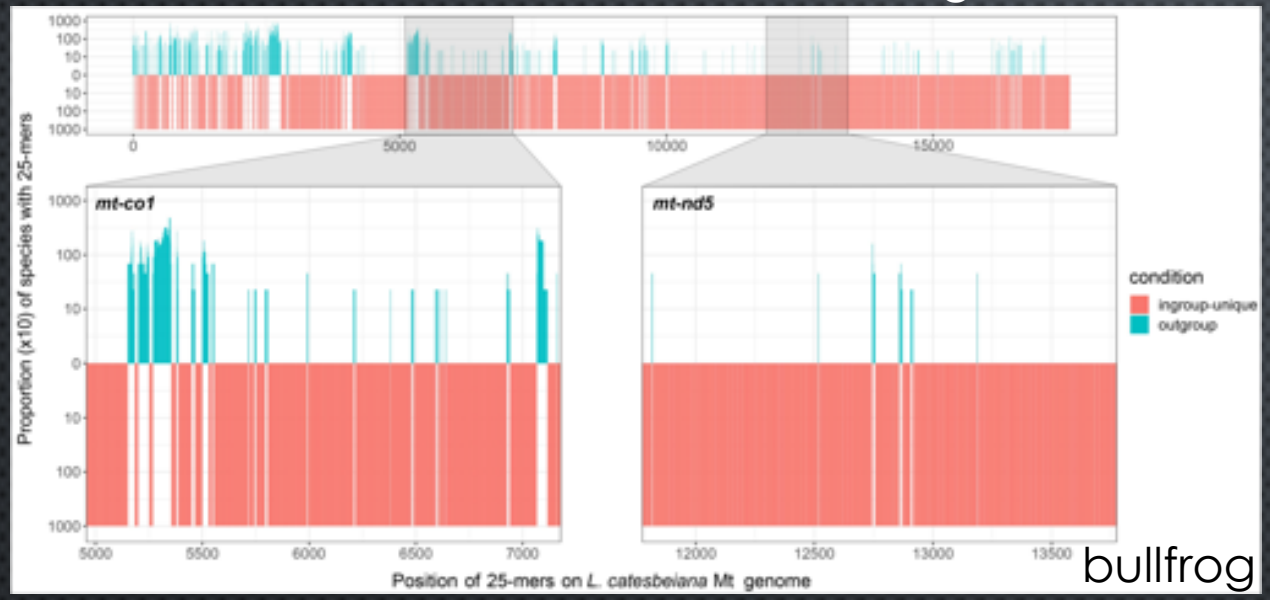
UNIKSEQ v1.2.2

- OUTPUTS CONSERVED REGIONS
- ROBUST TO DOS/WINDOWS AND UNIX-FORMATTED RNA/DNA FASTA FILES
- HANDLES TRADITIONAL MULTI-LINE FASTA RECORDS (SEQUENCE ON MULTI-LINES)
- VERSION CONTROLLED HTTPS://GITHUB.COM/BCGSC/UNIKSEQ
- WORKS ON VARIED INPUT, NON-CONTIGUOUS GENOMES / READ SETS

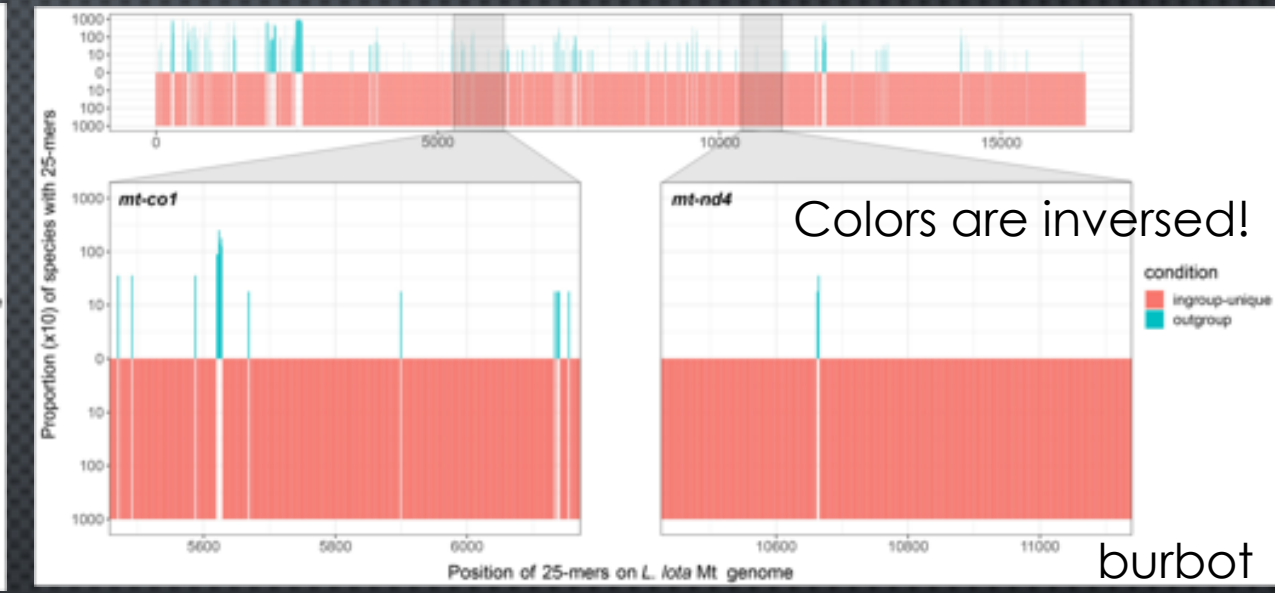
```
Usage: ./unikseq v1.2.2
-----input files-----
 -r reference FASTA (required)
 -i ingroup FASTA (required)
 -o outgroup FASTA (required)
-----kmer uniqueness filters-----
 -k length (option, default: -k 25)
 -l [leniency] min. non-unique consecutive kmers allowed in outgroup (option, default: -l 0)
 -m max. [% entries] in outgroup tolerated to have a reference kmer (option, default: -m 0 %
[original behaviour])
-----output filters-----
 -t print only first t bases in tsv output (option, default: -t 25)
 -c output conserved FASTA regions between reference and ingroup entries (option, -c 1==yes -c
0==no, [default, original unikseq behaviour])
 -s min. reference FASTA region [size] (bp) to output (option, default: -s 100 bp)
 -p min. [-c 0:region average /-c 1: per position] proportion of ingroup entries (option, default:
-p 25 %)
 -u min. [% unique] kmers in regions (option, default: -u 90 %)
```

eDNA ASSAY DESIGN : CASE STUDY

american bullfrog and burbot : strong assay design potential



bullfrog



Colors are inversed!

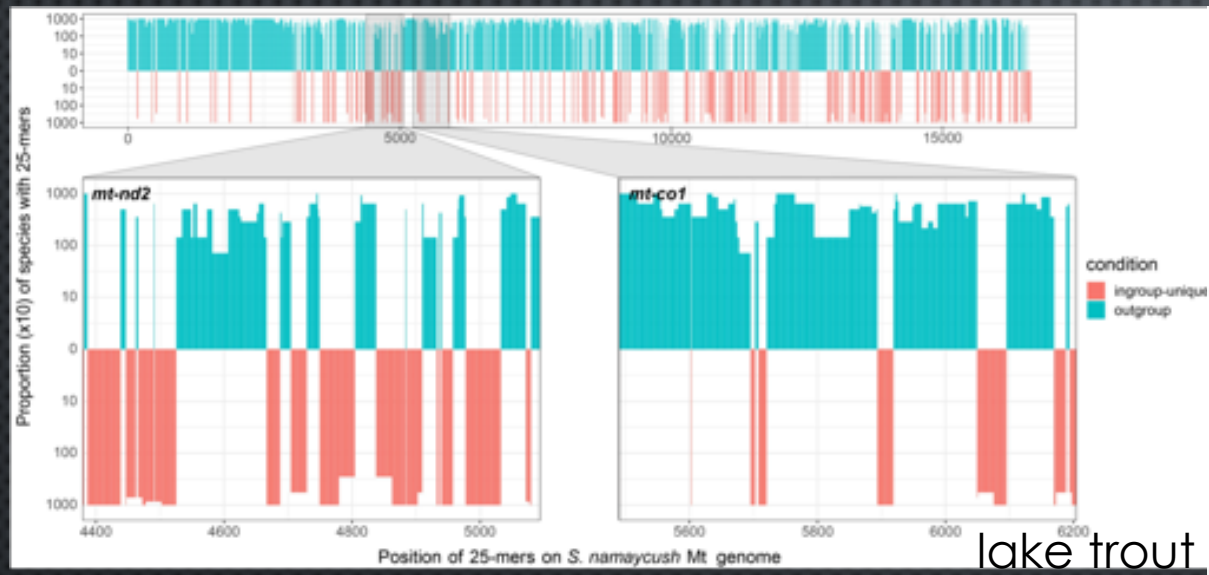
burbot

mt-coi1 barcode

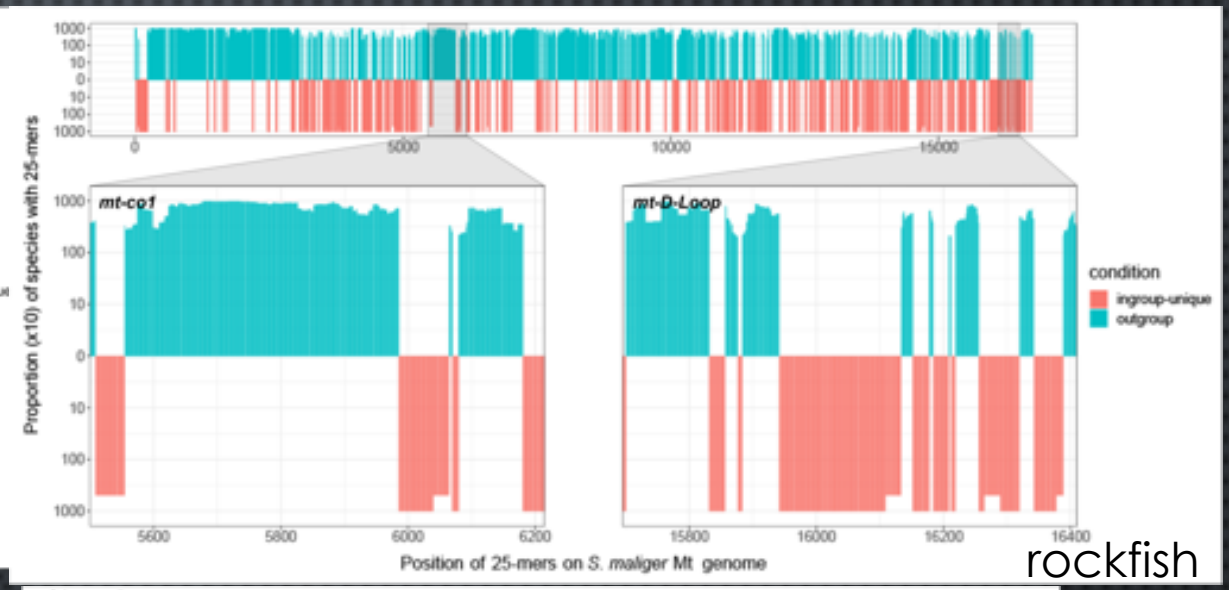
unikseq

eDNA ASSAY DESIGN : CASE STUDY

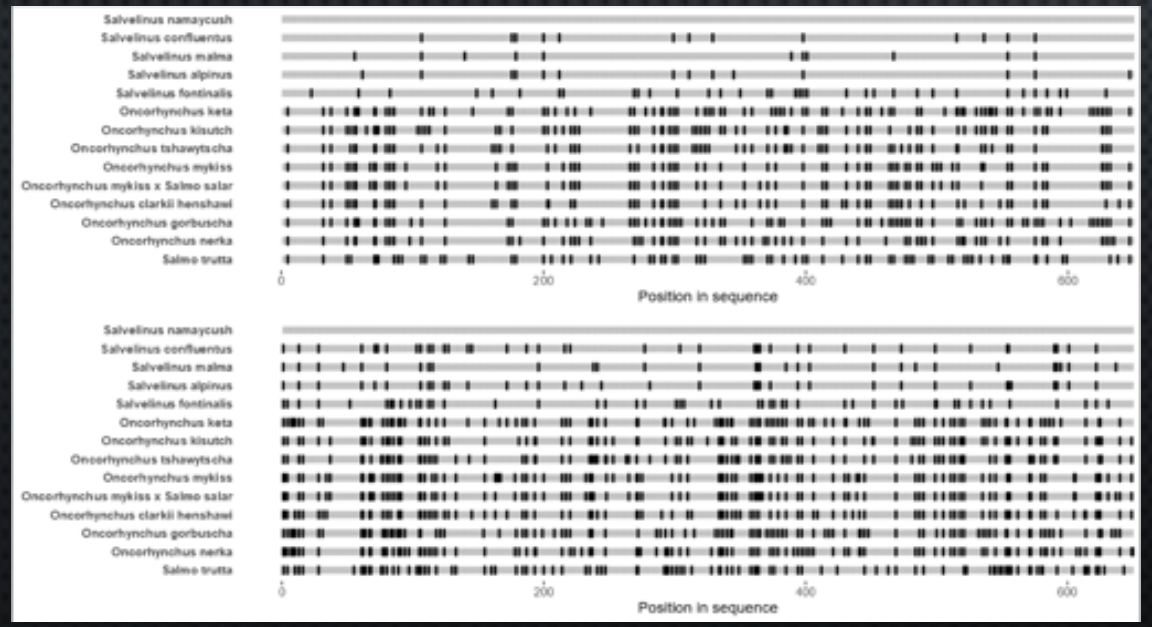
trout and rockfish : closely related taxa limit mitogenome qPCR design potential



lake trout



rockfish



mt-coi1 barcode

unikseq