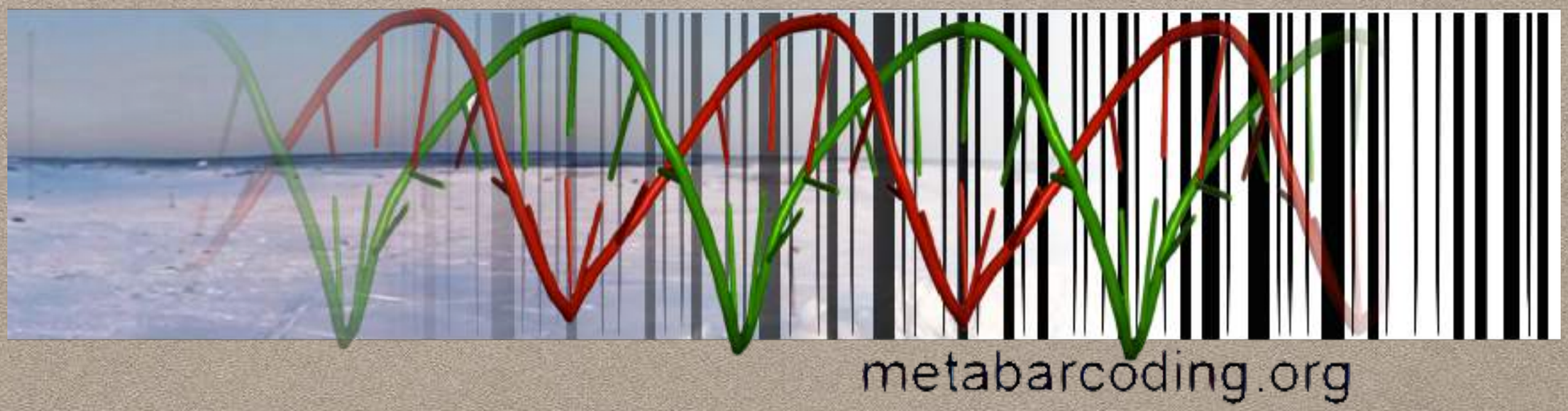
DESIGNING AND TESTING IN SILICO NEW METABARCODES. INI

ERIC COISSAC

PORTO DNA METABARCODING SPRING SCHOOL

MAI 1ST, 2017



WHAT IS A GOOD BARCODE ?

DNA BARCODING STAND VIEW

THE ROYAL SOCIETY

Received 29 July 2002
Accepted 30 September 2002
Published online 8 January 2003

Biological identifications through DNA barcodes

Paul D. N. Hebert*, Alina Cywinska, Shelley L. Ball and Jeremy R. deWaard

Department of Zoology, University of Guelph, Guelph, Ontario N1G 2W1, Canada

- *Animals : COI*
- *Plants : RbcL/MatK/ITS2*
- *Fungi : ITS1*

BMC Genomics

BioMed Central

Research article

Open Access

A universal DNA mini-barcode for biodiversity analysis

Isabelle Meusnier[1], Gregory AC Singer[2], Jean-François Landry[3], Donal A Hickey[4], Paul DN Hebert[1] and Mehrdad Hajibabaei*[1]

Published: 12 May 2008

BMC Genomics 2008, **9**:214 doi:10.1186/1471-2164-9-214

DNA METABARCODING STAND VIEW

biology letters

rsbl.royalsocietypublishing.org

Opinion piece

Population genetics

DNA metabarcoding and the cytochrome *c* oxidase subunit I marker: not a perfect match

Bruce E. Deagle[1], Simon N. Jarman[1], Eric Coissac[2,3], François Pompanon[2,3] and Pierre Taberlet[2,3]

Accepted: 19 August 2014

WHAT IS A GOOD BARCODE ?

It depends what you want to do with it ?

AMPLIFYING FROM A MIX OF DNA

- We want :
  - A representative amplification of all present DNA molecules
  - In term of diversity
  - Potentially, with respect of abundances
- Capacity of a DNA molecule to be amplified depends
  - DNA molecule integrity
  - of the primer affinity to it (related to the mismatches between both the molecules).

WE MUST SELECT HIGLY CONSERVED PRIMERS and **SHORT MARKERS**

TOO MANY CONSTRAINTS

- IDEAL MARKER DOESN'T EXIST -

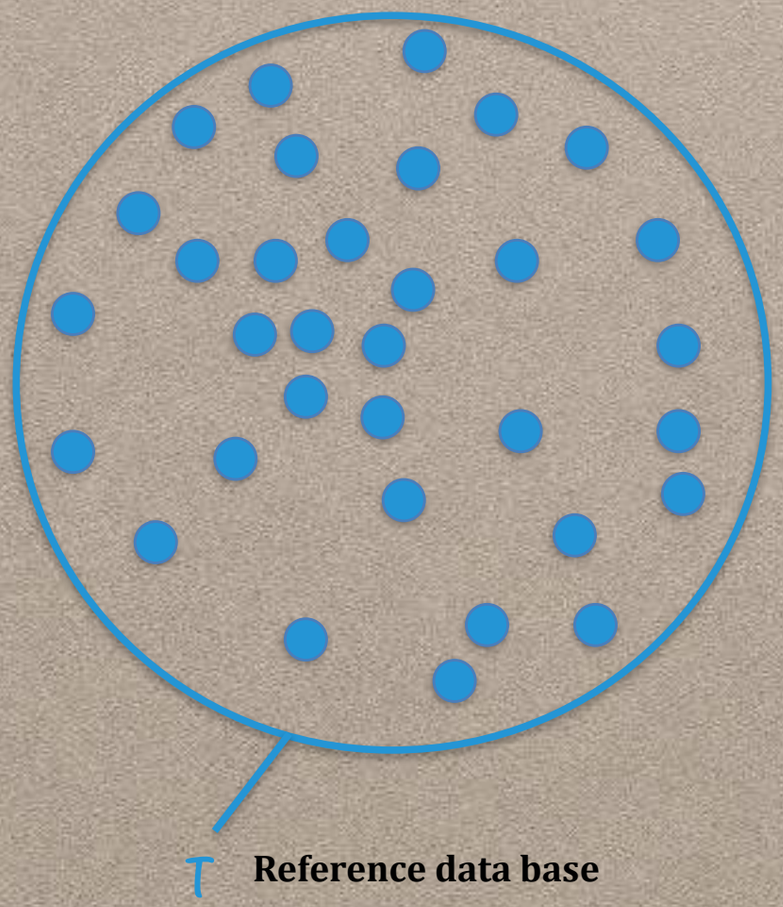
We want :

- ***Small** barcodes*
- *Highly **conserved** primers*
- ***Broad** taxonomy range*

SOMETIME WE HAVE TO SELECT MARKERS
WITH LOWER TAXONOMIC RESOLUTION

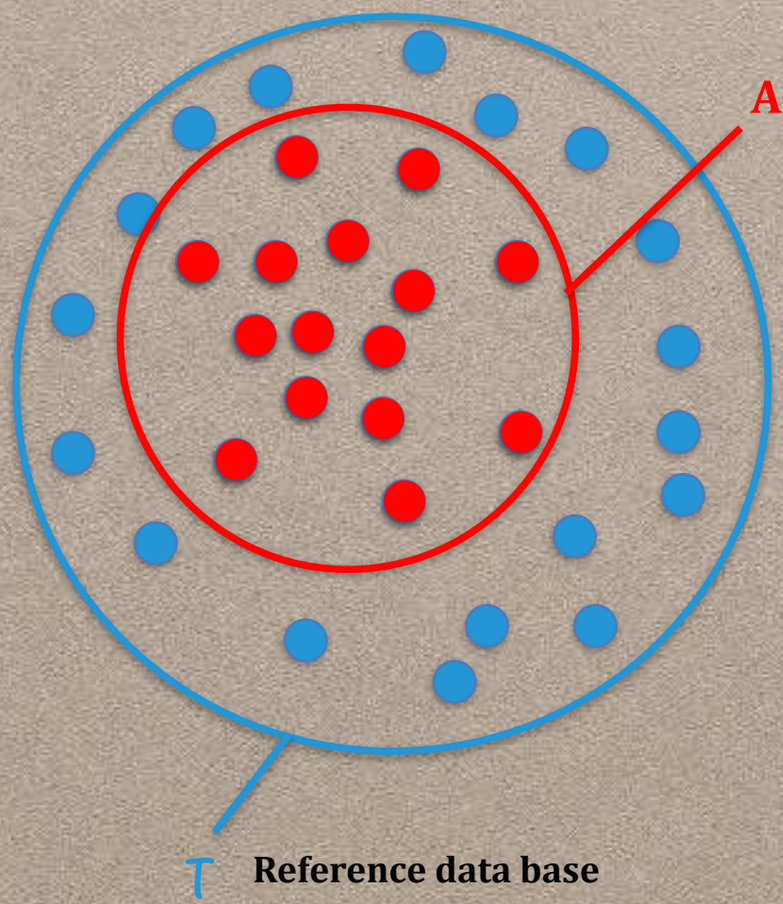
QUALIFYING BARCODES USING BOTH INDICES BC & BS

QUALIFYING BARCODES USING BOTH INDICES BC & BS



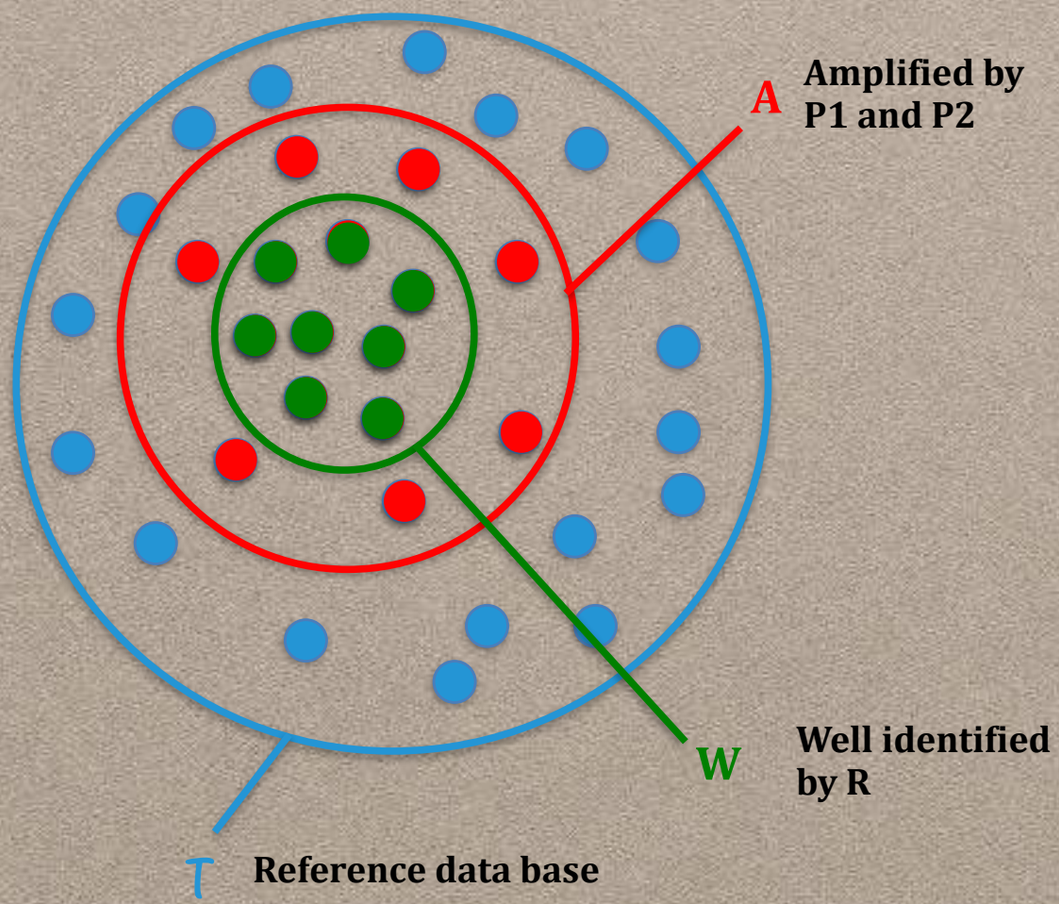
T **Reference data base**

QUALIFYING BARCODES USING BOTH INDICES BC & BS



$$B_c = \frac{\text{Amplified taxa.}}{\text{Total taxa.}} = \frac{|A|}{|T|}$$

QUALIFYING BARCODES USING BOTH INDICES BC & BS



$$B_c = \frac{\text{Amplified taxa.}}{\text{Total taxa.}} = \frac{|A|}{|T|}$$

$$B_s = \frac{\text{Well identified taxa}}{\text{Amplified taxa}} = \frac{|W|}{|A|}$$

TO COMPARE METABARCODES : ECOPCR

http://metbarcoding.org/ecoPCR

Ficetola *et al. BMC Genomics* 2010, **11**:434
http://www.biomedcentral.com/1471-2164/11/434

BMC Genomics

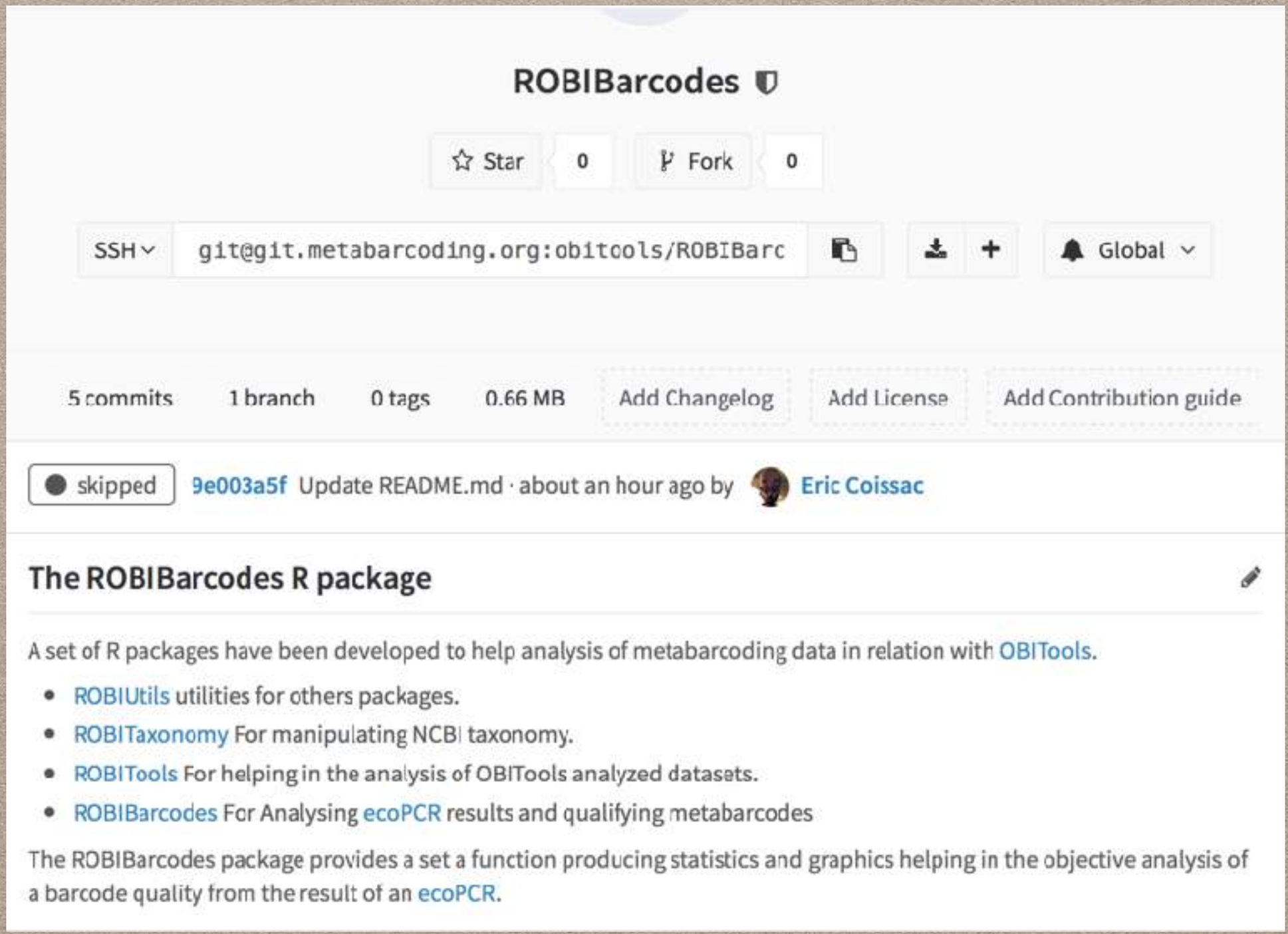
METHODOLOGY ARTICLE Open Access

An *In silico* approach for the evaluation of DNA barcodes

Gentile Francesco Ficetola[1,2,3*†], Eric Coissac[1*†], Stéphanie Zundel[1], Tiayyba Riaz[1], Wasim Shehzad[1], Julien Bessière[1], Pierre Taberlet[1], François Pompanon[1]

ANALYSING ECOPCR RESULTS

https://git.metabarcoding.org/obitools/ROBIBarcodes



TESTING FISH PRIMERS

```
Forward primer : ACACCGCCCGTCACTCTC
Reverse primer : CCAAGTGCACCTTCCGG
```

```
ecoPCR -d mito.vert \
       -e 5 \
       -l 30 -L 150 \
       -c \
        ACACCGCCCGTCACTCTC CCAAGTGCACCTTCCGGT > Teleostei.ecoprimers
```

WHAT CONTENTS AN ECOPCR RESULT FILE

```
> library(ROBIBarcodes)
> ecopcr = read.ecopcr.result("Teleostei.04.vert.ecopcr")
> head(ecopcr,2)
         AC seq_length  taxid    rank species              species_name
1 NC_013146      16960 100858 species  100858 Threskiornis aethiopicus
2 NC_016427      16379  82464 species   82464           Myodes regulus
   genus   genus_name family       family_name superkingdom
1 100857 Threskiornis  33574 Threskiornithidae         2759
2 447134       Myodes 337677        Cricetidae         2759
  superkingdom_name strand      forward_match forward_mismatch forward_tm
1         Eukaryota      D ATACCGCCCGTCACCCTC                2      45.64
2         Eukaryota      D ACACCGCCCGTCACCCTC                1      53.84
      reverse_match reverse_mismatch reverse_tm amplicon_length
1 CTAAGTGCACATTCCGGT                2      45.55              74
2 CCAAGCACACTTTCCAGT                4      16.28              79
                                                            sequence
1     CTCATAAGCTACTGACTCCCATAACTAATACCCTAATTAGCCGAAGATGAGGTAAGTCGTAACAAGGTAAGTGT
2 CTCAAACTAAATAAATGAGATCTATACATAATTACATCAAACTTTTACGAGAGGAGATAAGTCGTAACAAGGTAAGCAT
                                      definition
1 Threskiornis aethiopicus mitochondrion, complete genome
2          Myodes regulus mitochondrion, complete genome
```

NCBI TAXONOMY

D136–D143 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 1 December 2011*
doi:10.1093/nar/gkr1178

The NCBI Taxonomy database

Scott Federhen*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received October 20, 2011; Revised November 10, 2011; Accepted November 11, 2011

*Provides **unambiguous identifiers** for taxa and relationships among these taxa*

ftp://ftp.ncbi.nlm.nih.gov/pub/taxdump.tar.gz

LOAD A TAXONOMY INTO R

```
> library(ROBITaxonomy)

Attachement du package : 'ROBITaxonomy'

The following object is masked from 'package:stats':

    family

> taxonomy = read.taxonomy("ncbi20150518")
Readind 1283820 taxa...
No local taxon
Computing longest branches...
```

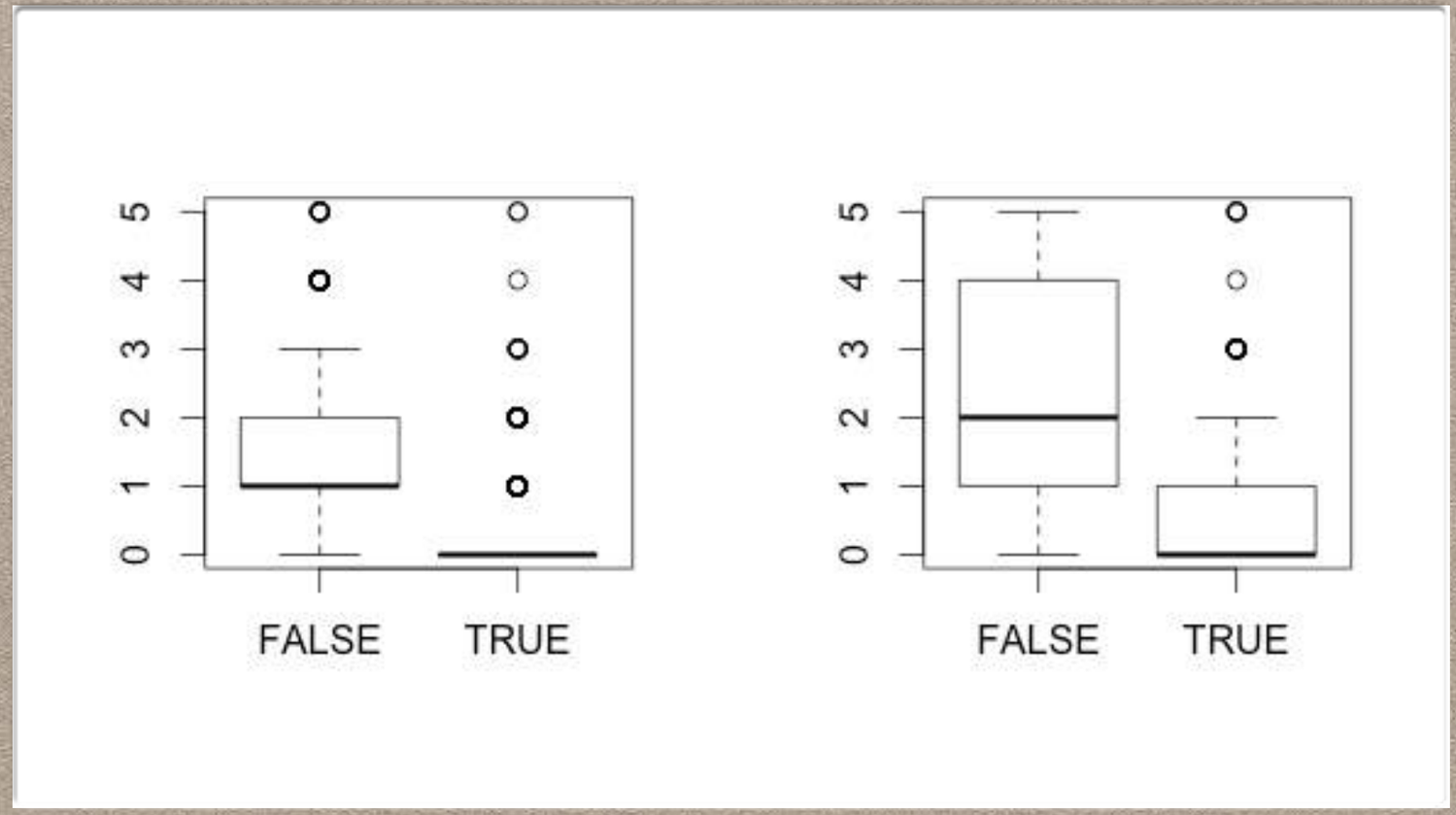
HOW MANY FISHES DO WE AMPLIFY ?

```
> teleostei = ecofind(taxonomy,"^teleostei$")
> teleostei
[1] 32443
> is.a.fish = is.subcladeof(taxonomy,
                            ecopcr$taxid,
                            teleostei)
```

```
> table(is.a.fish)
is.a.fish
FALSE  TRUE
 1673  1577
```

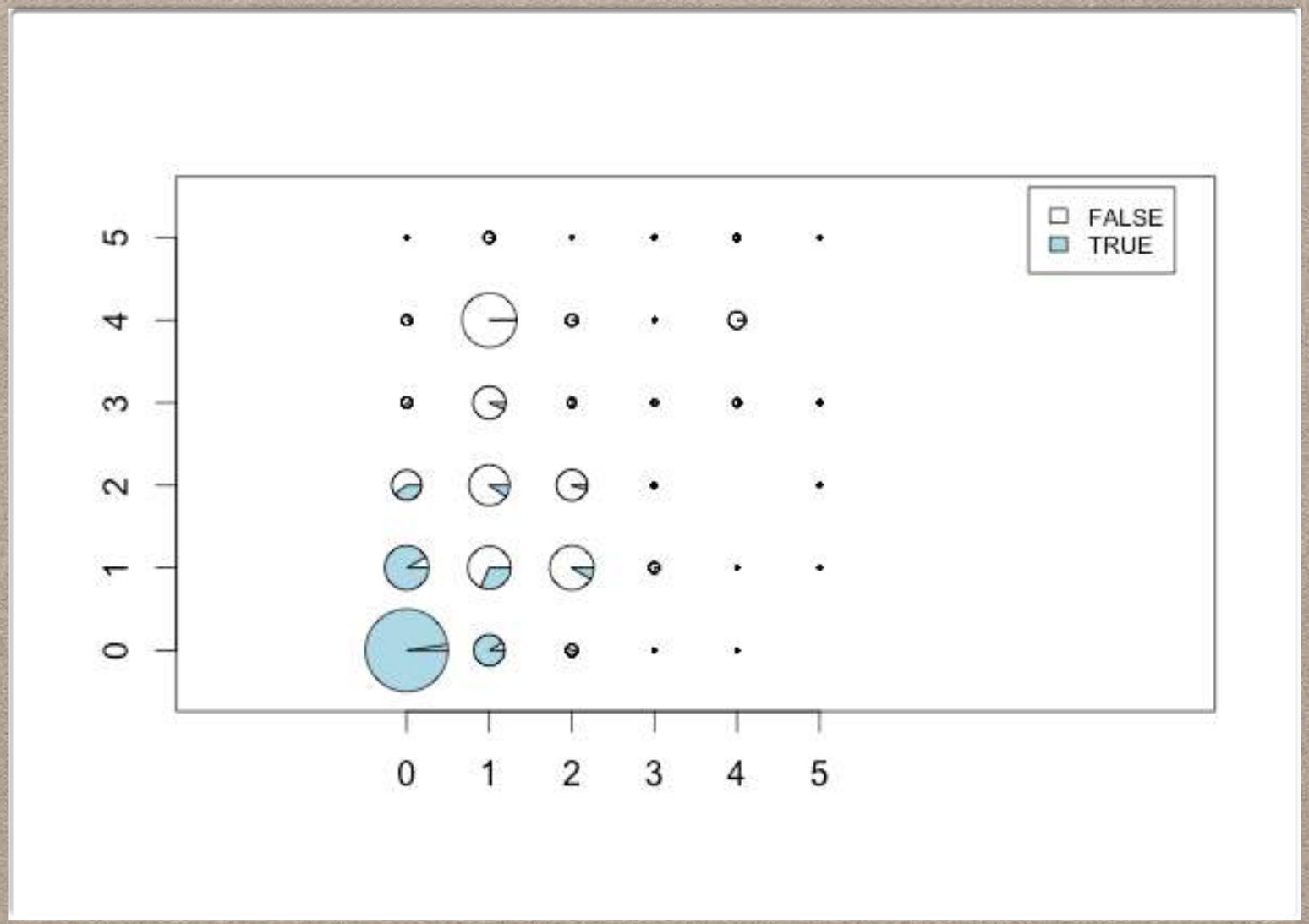
IS FISH AND NON-FISH AMPLIFICATION EQUIPROBABLE ?

```
> par(mfrow=c(1,2))
> boxplot(ecopcr$forward_mismatch ~ is.a.fish)
> boxplot(ecopcr$reverse_mismatch ~ is.a.fish)
```



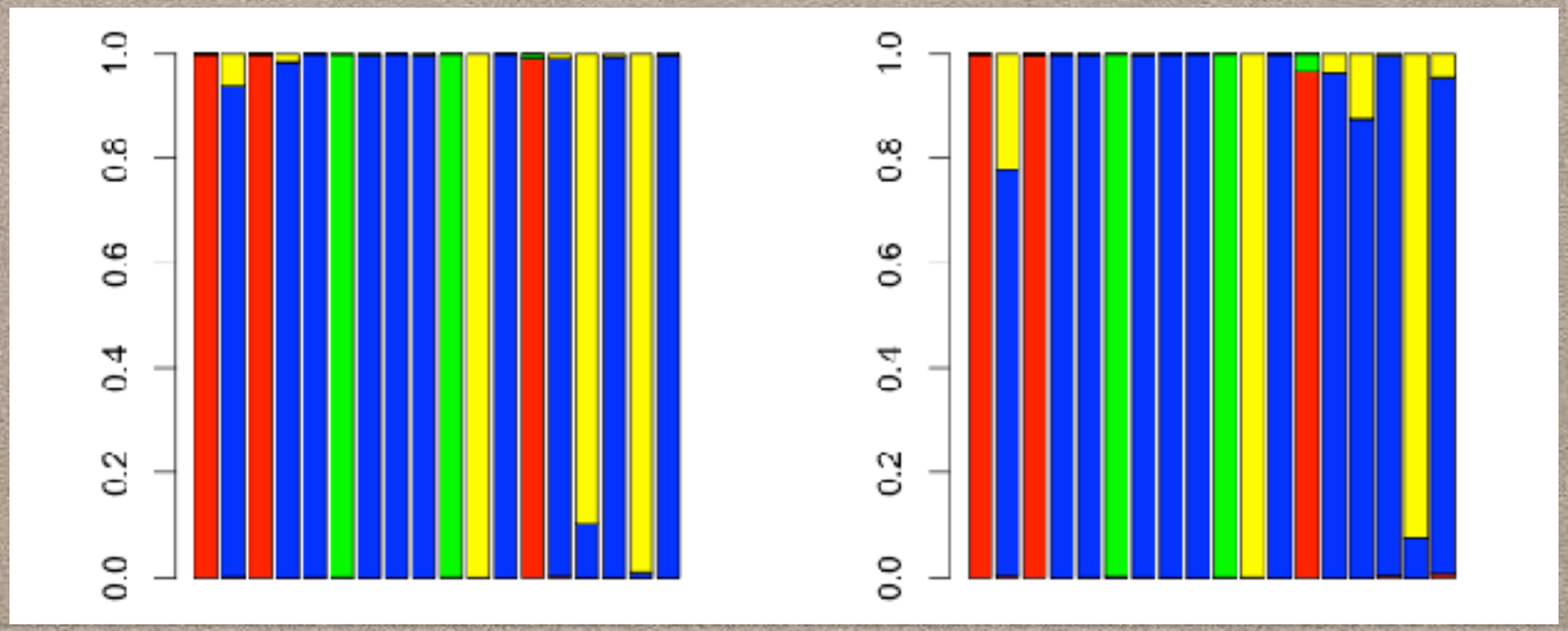
CAN WE DO A BETTER JOB ?

> mismatchplot(ecopcr,group=is.a.fish)



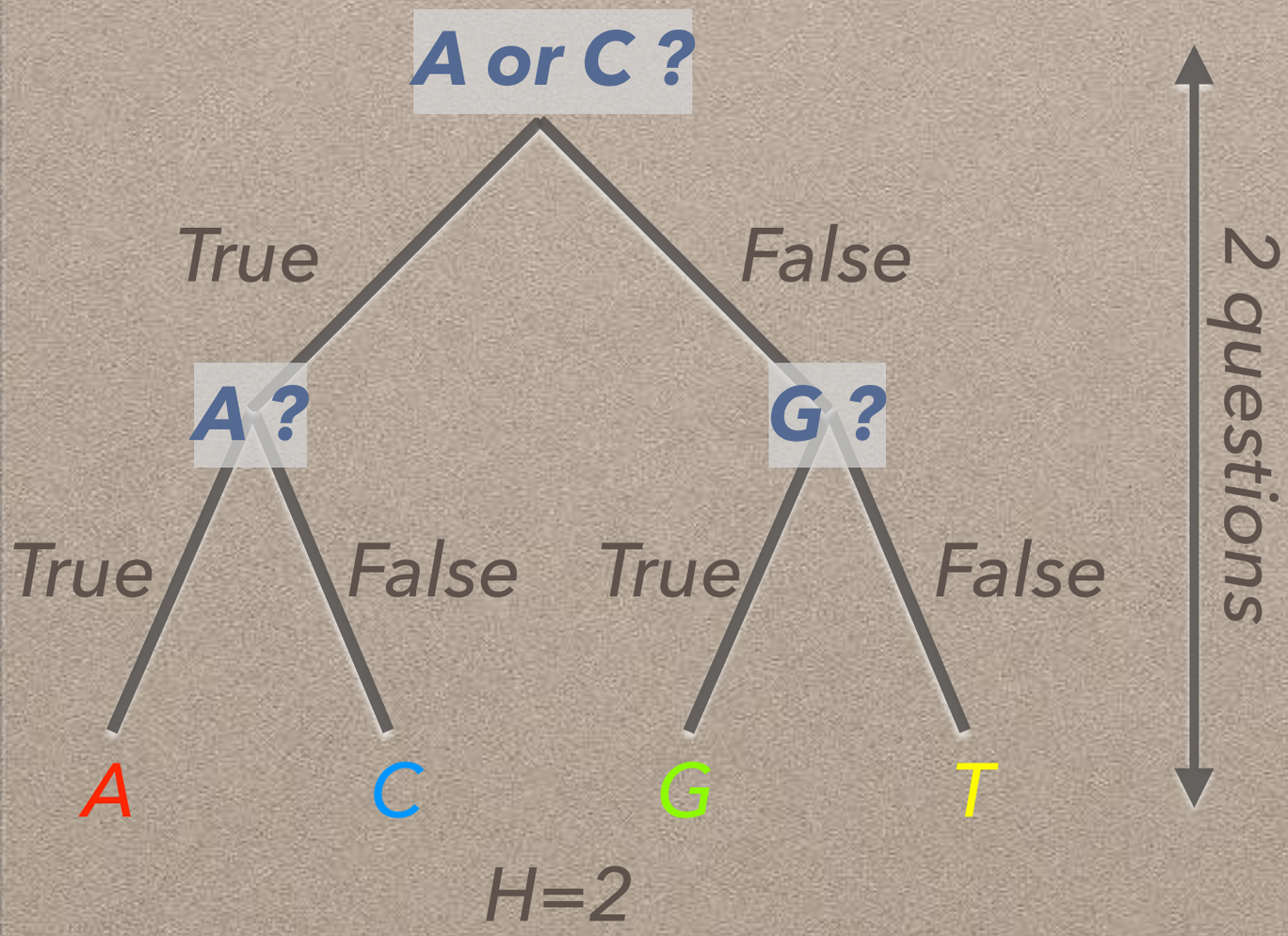
WHAT IS DIFFERENT BETWEEN FISHES AND OTHERS ?

```
> f.freq=ecopcr.forward.frequencies(ecopcr,group = is.a.fish)
> round(f.freq$'TRUE',2)
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17]
A    1 0.00    1 0.00    0    0    0    0    0     0     0     0  0.99  0.00   0.0  0.00  0.00
C    0 0.94    0 0.98    1    0    1    1    1     0     0     1  0.00  0.99   0.1  0.99  0.01
G    0 0.00    0 0.00    0    1    0    0    0     1     0     0  0.01  0.00   0.0  0.00  0.00
T    0 0.06    0 0.02    0    0    0    0    0     0     1     0  0.00  0.01   0.9  0.00  0.99
  [,18]
A     0
C     1
G     0
T     0
attr(,"count")
[1] 28386
> par(mfrow=c(1,2))
> barplot(f.freq$'TRUE',col=c("red","blue","green","yellow"))
> barplot(f.freq$'FALSE',col=c("red","blue","green","yellow"))
```



SHANON INDEX

$$H = -\sum_{i=\{A,C,G,T\}} P_i . \log_2(P_i)$$



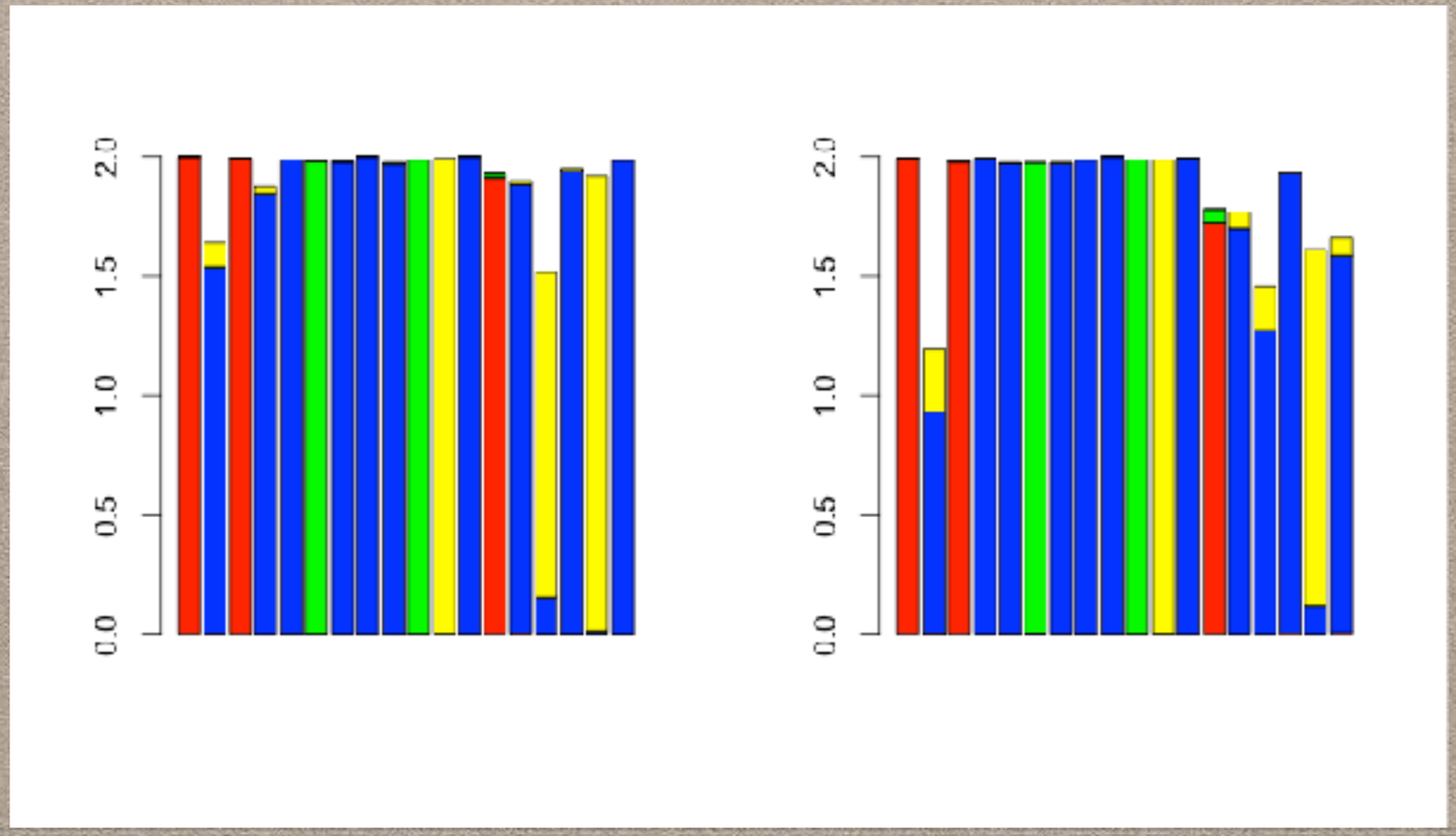
No question needed

G

H=0

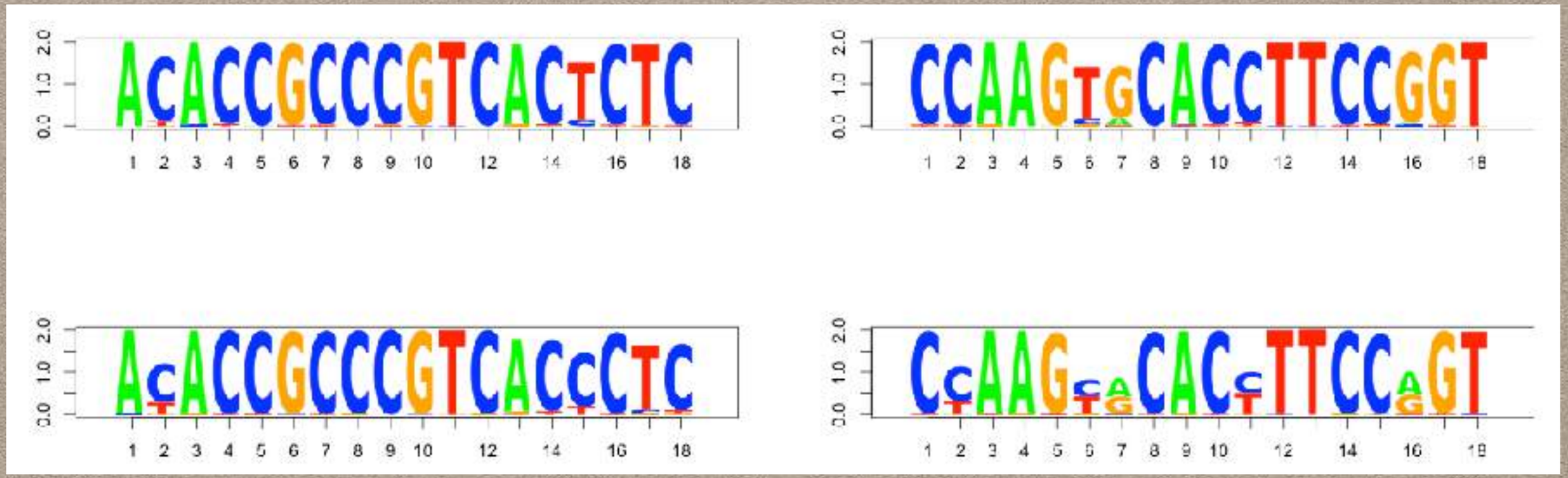
SHANON INDEX PER POSITION

```
> s.freq=ecopcr.forward.shanon(ecopcr,group = is.a.fish)
> barplot(s.freq$'TRUE',
          col=c("red","blue","green","yellow"))
> barplot(s.freq$'FALSE',
          col=c("red","blue","green","yellow"))
```



DNA LOGO TO SUMMARIZE INFORMATION

```
> s.forward=ecopcr.forward.shanon(ecopcr,group = is.a.fish)
> s.reverse=ecopcr.reverse.shanon(ecopcr,group = is.a.fish)
> par(mfrow=c(2,2))
> dnalogoplot(s.forward$'TRUE')
> dnalogoplot(s.reverse$'TRUE')
> dnalogoplot(s.forward$'FALSE')
> dnalogoplot(s.reverse$'FALSE')
```



DNA METABARCODING STAND VIEW

biology letters

rsbl.royalsocietypublishing.org

Opinion piece

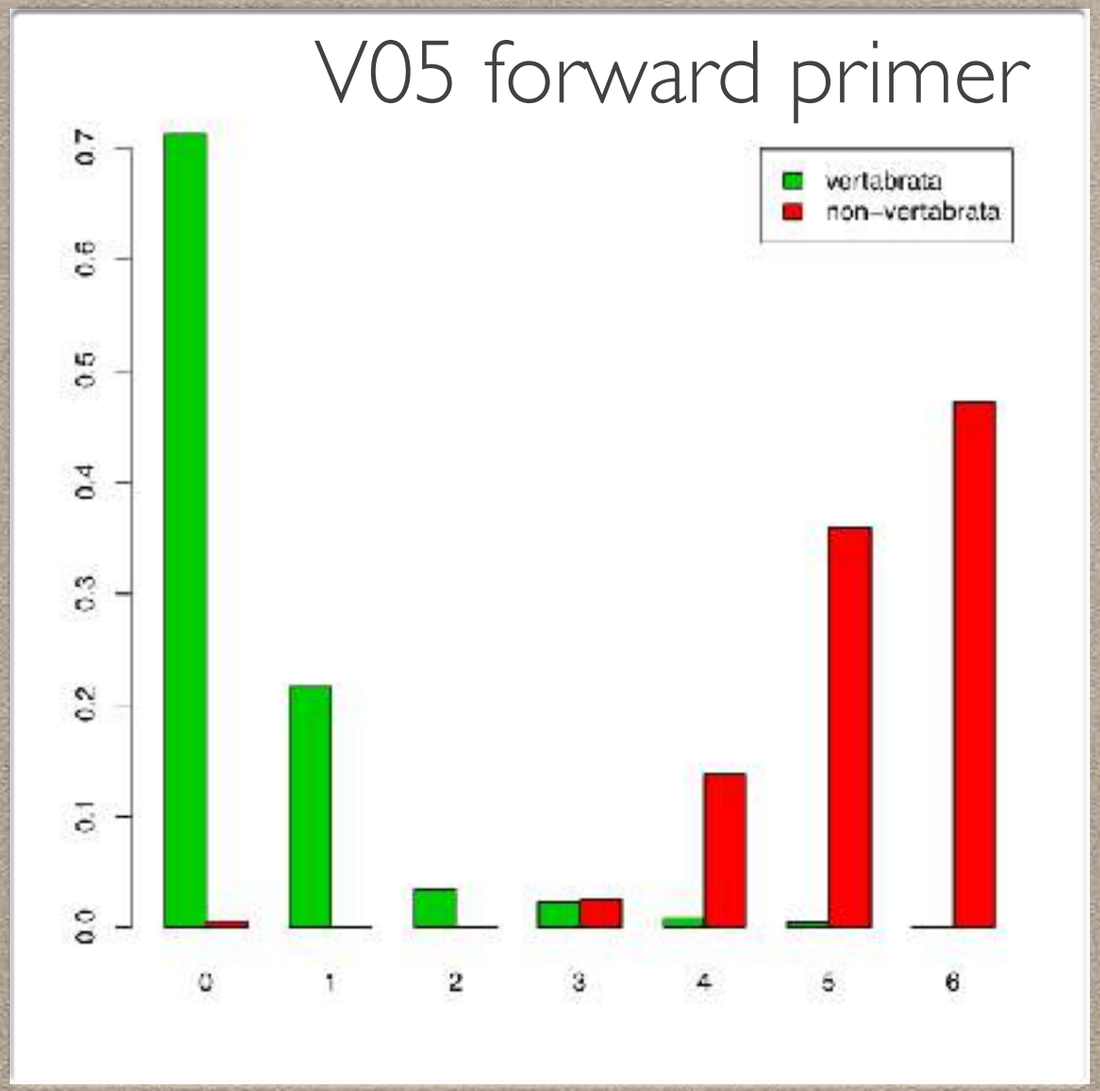
Population genetics

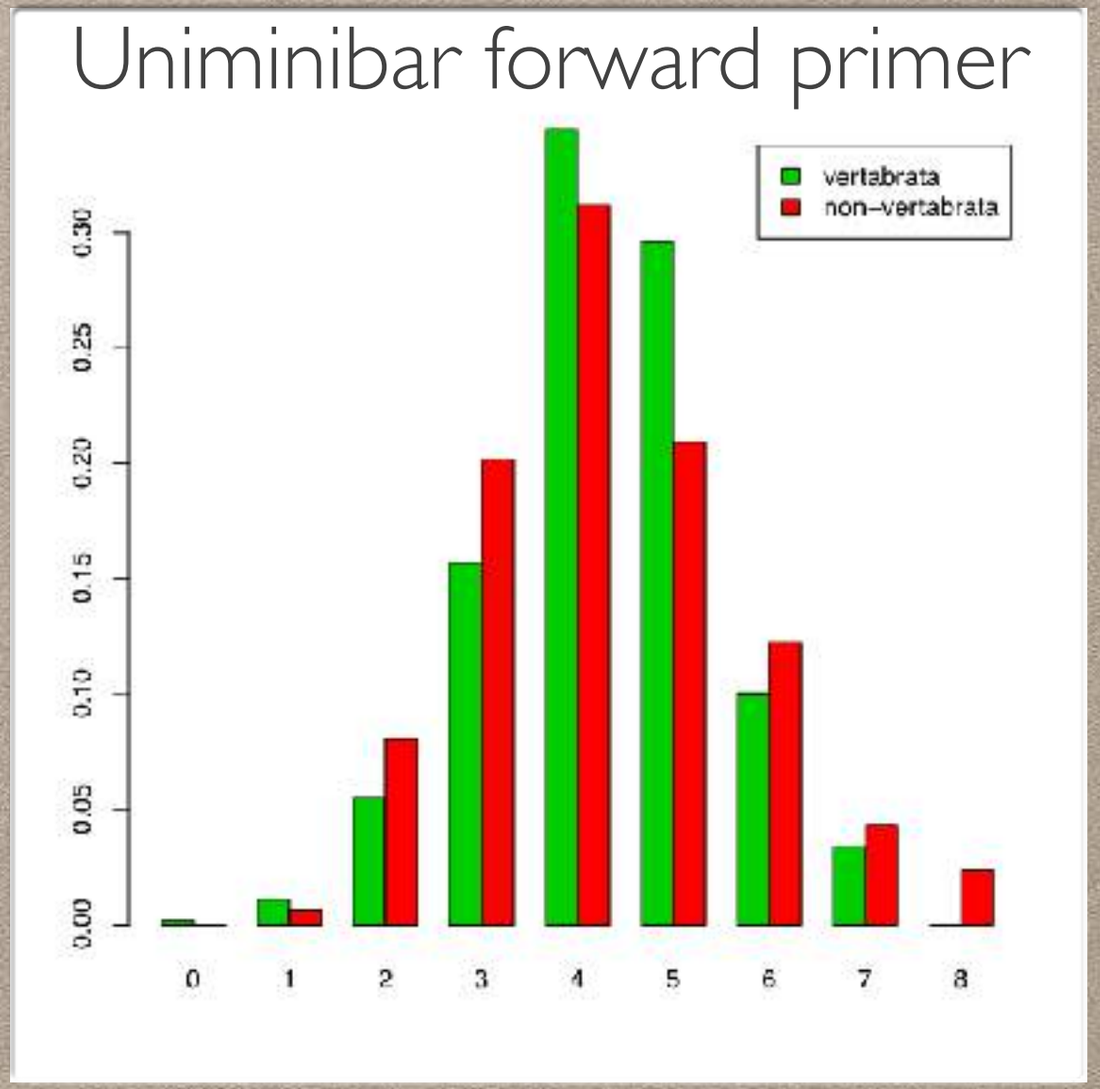
DNA metabarcoding and the cytochrome *c* oxidase subunit I marker: not a perfect match

Bruce E. Deagle[1], Simon N. Jarman[1], Eric Coissac[2,3], François Pompanon[2,3] and Pierre Taberlet[2,3]

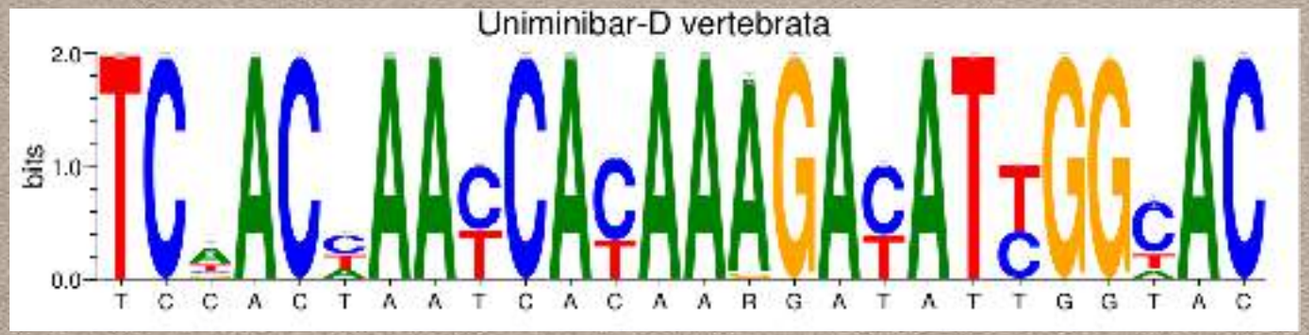
Accepted: 19 August 2014

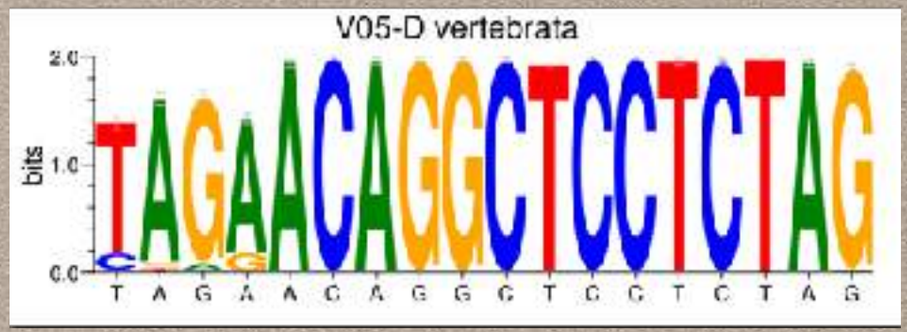
UNIVERSAL COI PRIMER VERSUS VERTEBRATE 12S RNA PRIMER

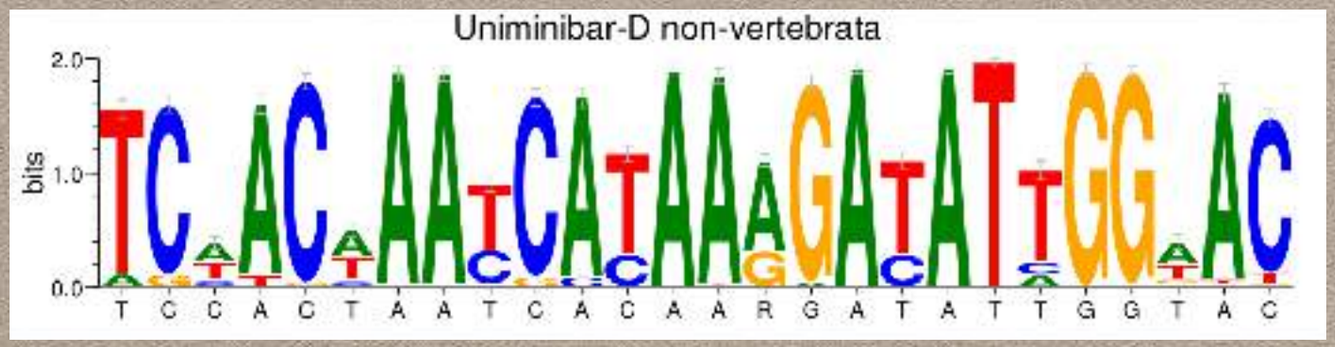


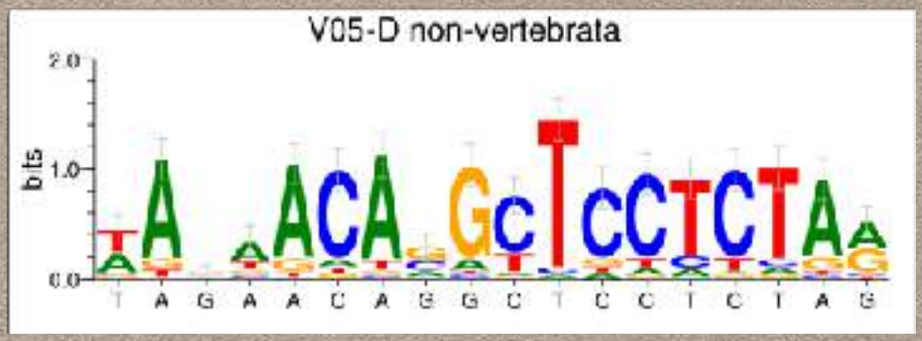


UNIVERSAL COI PRIMER
VERSUS
VERTEBRATE 12S RNA PRIMER

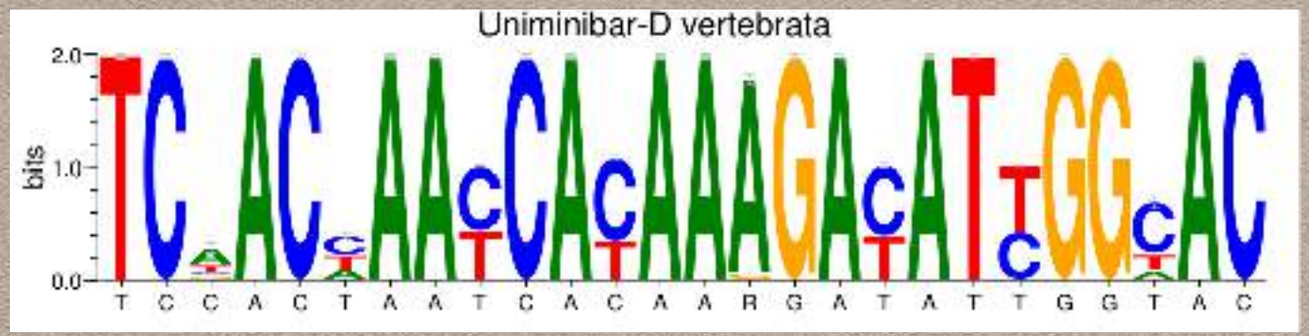




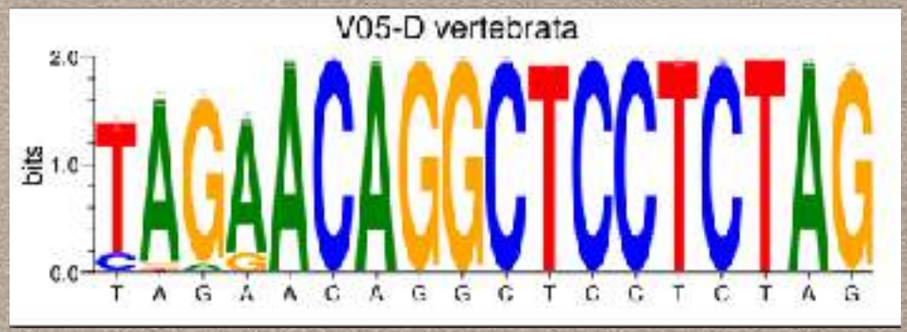


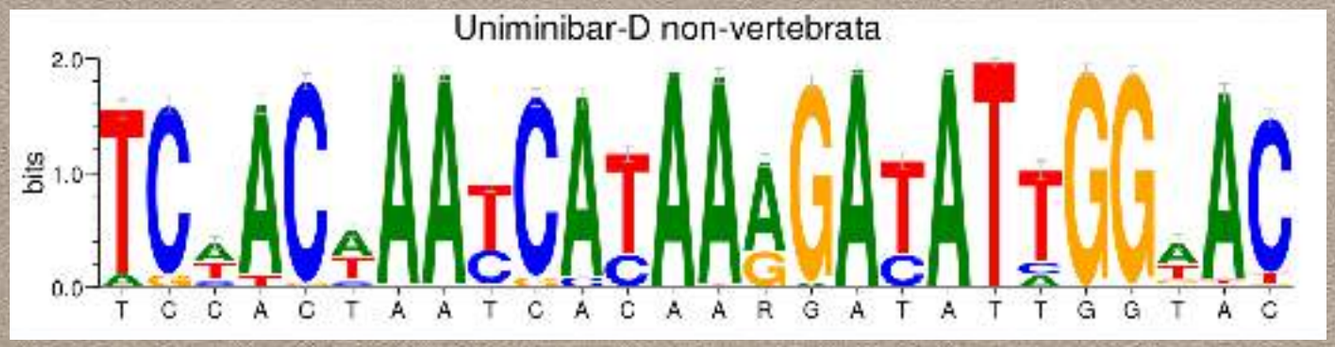


UNIVERSAL COI PRIMER
VERSUS
VERTEBRATE 12S RNA PRIMER

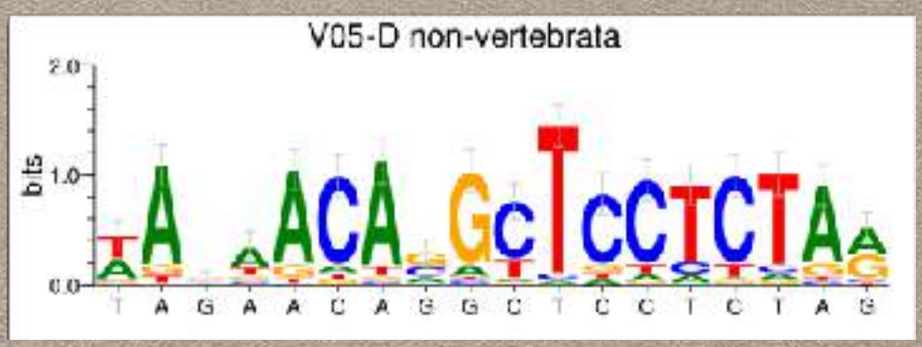


S T N H K D I G



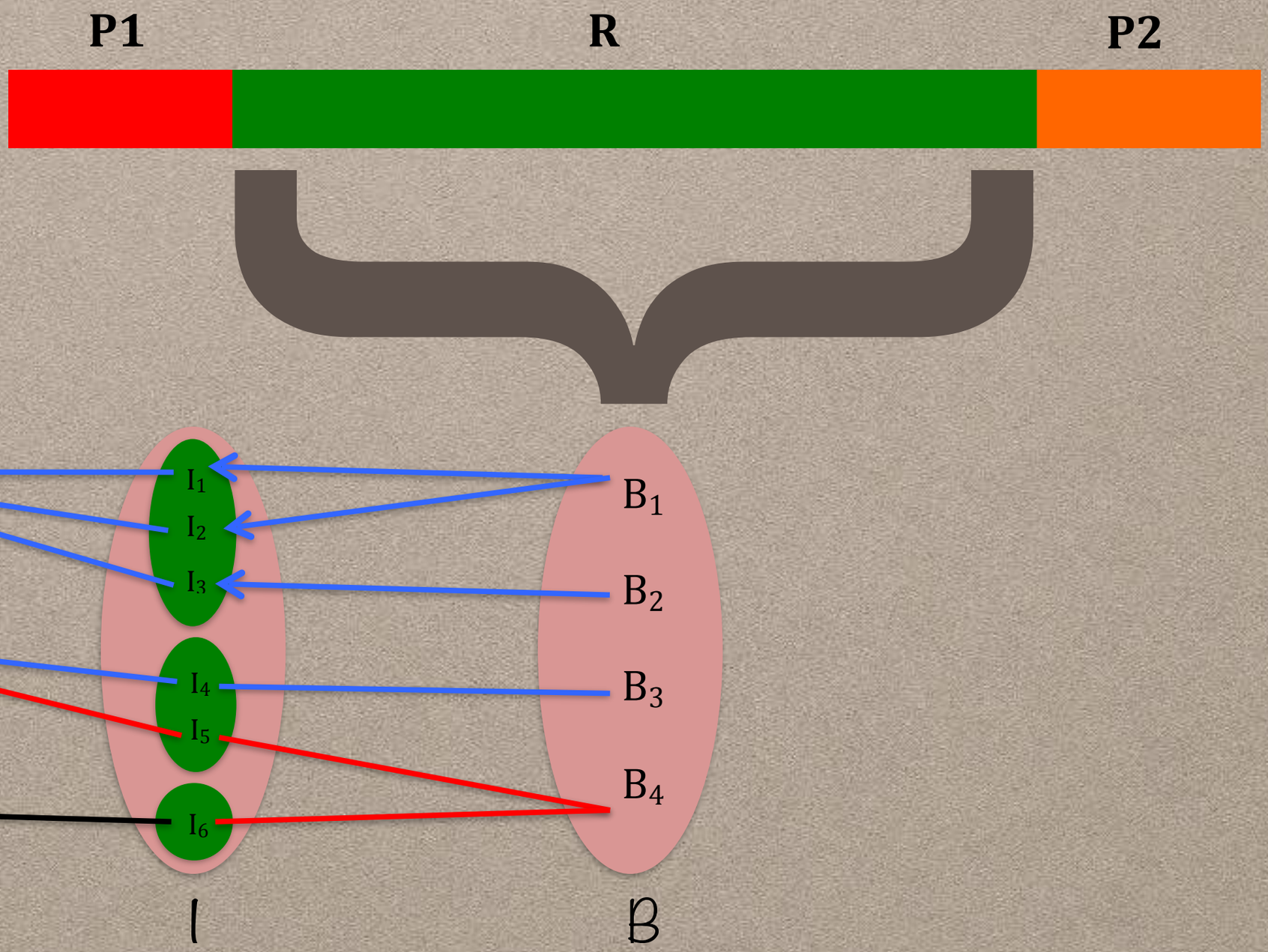


S T N H K D I G



S T N H K D I G

IS THE METABARCODE ABLE TO DISTINGUISH TAXA ?



STATISTICS ABOUT TAXONOMIC RESOLUTION

```
> r = resolution(taxonomy,ecopcr)
> round(table(r)/sum(table(r)) * 100)
r
    family      genus    no rank    species  subfamily   subgenus subspecies      tribe
         4         10          1         82          1          0          2          0

> r.fish = resolution(taxonomy,ecopcr[is.a.fish,])
> round(table(r.fish)/sum(table(r.fish)) * 100)
r.fish
    family      genus    no rank    species  subfamily subspecies      tribe
         7         12          1         78          1          1          1

> r.not.fish = resolution(taxonomy,ecopcr[!is.a.fish,])
> round(table(r.not.fish)/sum(table(r.not.fish)) * 100)
r.not.fish
    family      genus    no rank    species  subfamily   subgenus subspecies
         2          9          0         86          1          0          3

> unique(ecopcr[is.a.fish,][r.fish=='family',]$family_name)
 [1] Cyprinidae      Diceratiidae    Monacanthidae   Gadidae         Sinipercidae
 [6] Cetomimidae     Catostomidae    Channichthyidae Mormyridae      Balitoridae
[11] Istiophoridae   Sciaenidae
```

STATISTICS ABOUT TAXONOMIC RESOLUTION

```
> r = resolution(taxonomy,ecopcr)
> round(table(r)/sum(table(r)) * 100)
r
    family      genus    no rank    species  subfamily   subgenus subspecies      tribe
         4         10          1         82          1          0          2          0

> r.fish = resolution(taxonomy,ecopcr[is.a.fish,])
> round(table(r.fish)/sum(table(r.fish)) * 100)
r.fish
    family      genus    no rank    species  subfamily subspecies      tribe
         7         12          1         78          1          1          1

> r.not.fish = resolution(taxonomy,ecopcr[!is.a.fish,])
> round(table(r.not.fish)/sum(table(r.not.fish)) * 100)
r.not.fish
    family      genus    no rank    species  subfamily   subgenus subspecies
         2          9          0         86          1          0          3

> unique(ecopcr[is.a.fish,][r.fish=='family',]$family_name)
 [1] Cyprinidae      Diceratiidae    Monacanthidae   Gadidae         Sinipercidae
 [6] Cetomimidae     Catostomidae    Channichthyidae Mormyridae      Balitoridae
[11] Istiophoridae   Sciaenidae
```

HOW MANY CYPRINIDAE SPECIES WILL BY IDENTIFIED ?

```
> table(aggregate(r.cyprinidae,
        by=list(species=ecopcr[ecopc$family_name=="Cyprinidae",
                               "species_name"]),
        function(x) all(x=='species'))$x)

FALSE  TRUE
  102   210
```

ECOPRIMERS :
http://metabarcoding.org/ecoPrimers

Nucleic Acids Research, 2011, 1–11
doi:10.1093/nar/gkr732

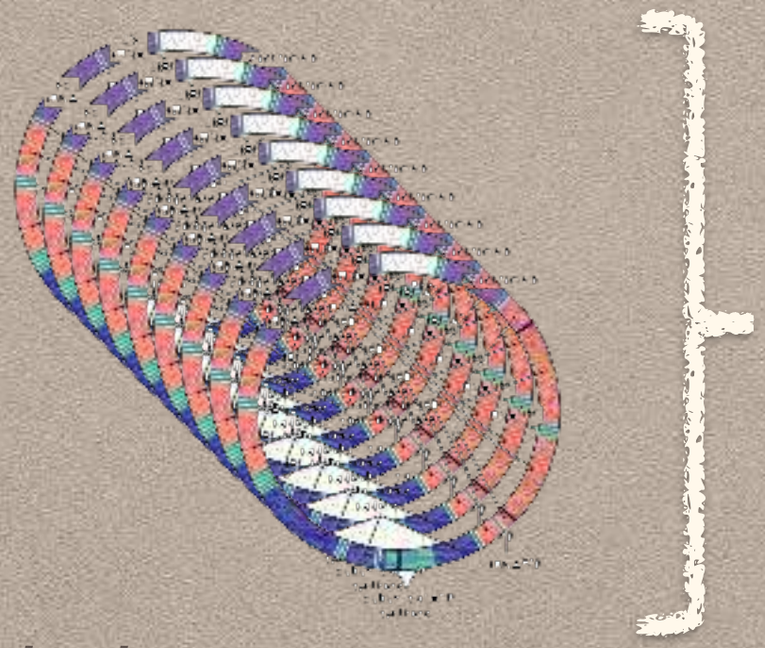
ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis

Tiayyba Riaz[1], Wasim Shehzad[1], Alain Viari[2], François Pompanon[1], Pierre Taberlet[1] and Eric Coissac[1,*]

[1]Laboratoire d'Ecologie Alpine (LECA) CNRS UMR 5553 2233, Université Joseph Fourrier, BP 53, 38041 Grenoble Cedex-9 and [2]INRIA Rhône-Alpes – Projet Bamboo, ZIRST-655 Avenue de l'Europe, 38334 Montbonnot Cedex, France

Received June 9, 2011; Revised and Accepted August 23, 2011

ECOPRIMER IS AN ALIGNMENT FREE ALGORITHM



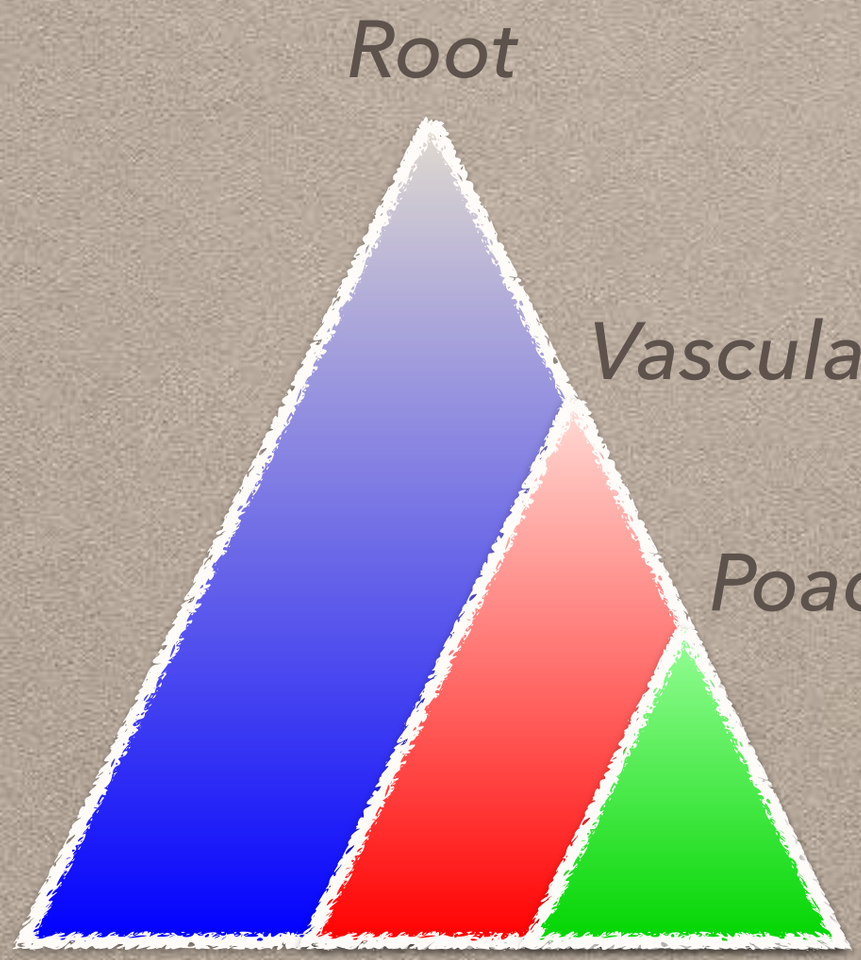
many whole genome sequences

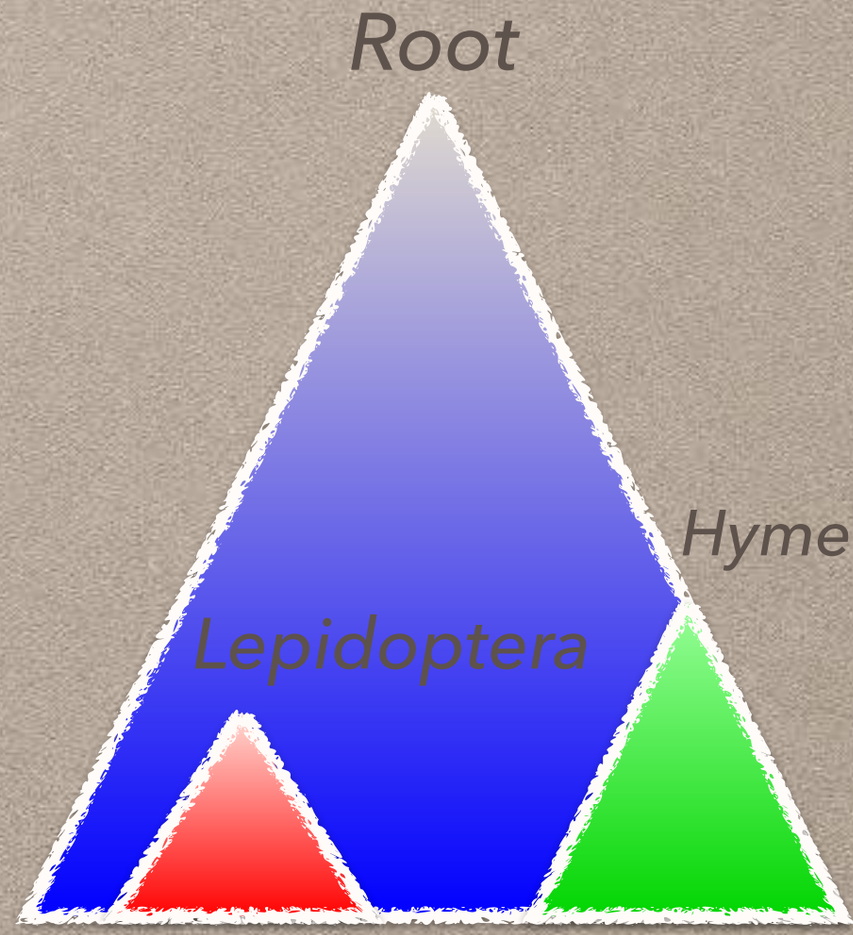
Look for conserved regions that flank variable regions

And look for primer pairs that optimize both

- *conservation of the priming sites*
- *taxonomic resolution of the barcode*

RESTRICT PRIMERS TO A CLADE





HOW MANY CYPRINIDAE SEQUENCES DO WE HAVE ?

```
> cyprinidae = ecofind(taxonomy,"^cyprinidae$")
> cyprinidae
[1] 7953
> is.a.cyprinidae = is.subcladeof(taxonomy,
                                  ecopcr$taxid,
                                  cyprinidae)
> table(is.a.cyprinidae)
is.a.cyprinidae
FALSE  TRUE
 2927   323
> table(is.a.fish)
is.a.fish
FALSE  TRUE
 1673  1577
```

RUNNING ECOPRIMERS...

```
ecoPrimers -d mito.vert \
           -e 3 -3 2 \
           -l 30 -L 150 \
           -r 7953 -c > Cyprinidae.ecoprimers
```

```
0   ACGGCGTAAAGGGTGGTT  CATAGTGGGGTATCTAAT  59.849.047.639.510  7   GG
          314 0   0.972   303 0   0.971   247 0.815   123 138 129.66

1   ACGGCGTAAAGGGTGGTT  GCATAGTGGGGTATCTAA  59.849.050.743.010  8   GG
          314 0   0.972   303 0   0.971   247 0.815   124 139 130.66

2   ACGGCGTAAAGGGTGGTT  AGCATAGTGGGGTATCTA  59.849.051.443.810  8   GG
          314 0   0.972   303 0   0.971   247 0.815   125 140 131.66

3   ACGGCGTAAAGGGTGGTT  GAGCATAGTGGGGTATCT  59.849.053.141.510  9   GG
          314 0   0.972   303 0   0.971   247 0.815   126 141 132.66
                                                                  …
14  ACGGCGTAAAGGGTGGTT  TATCTAATCCCAGTTTGT  59.849.047.535.310  6   GG
          314 0   0.972   303 0   0.971   247 0.815   113 128 119.66
```

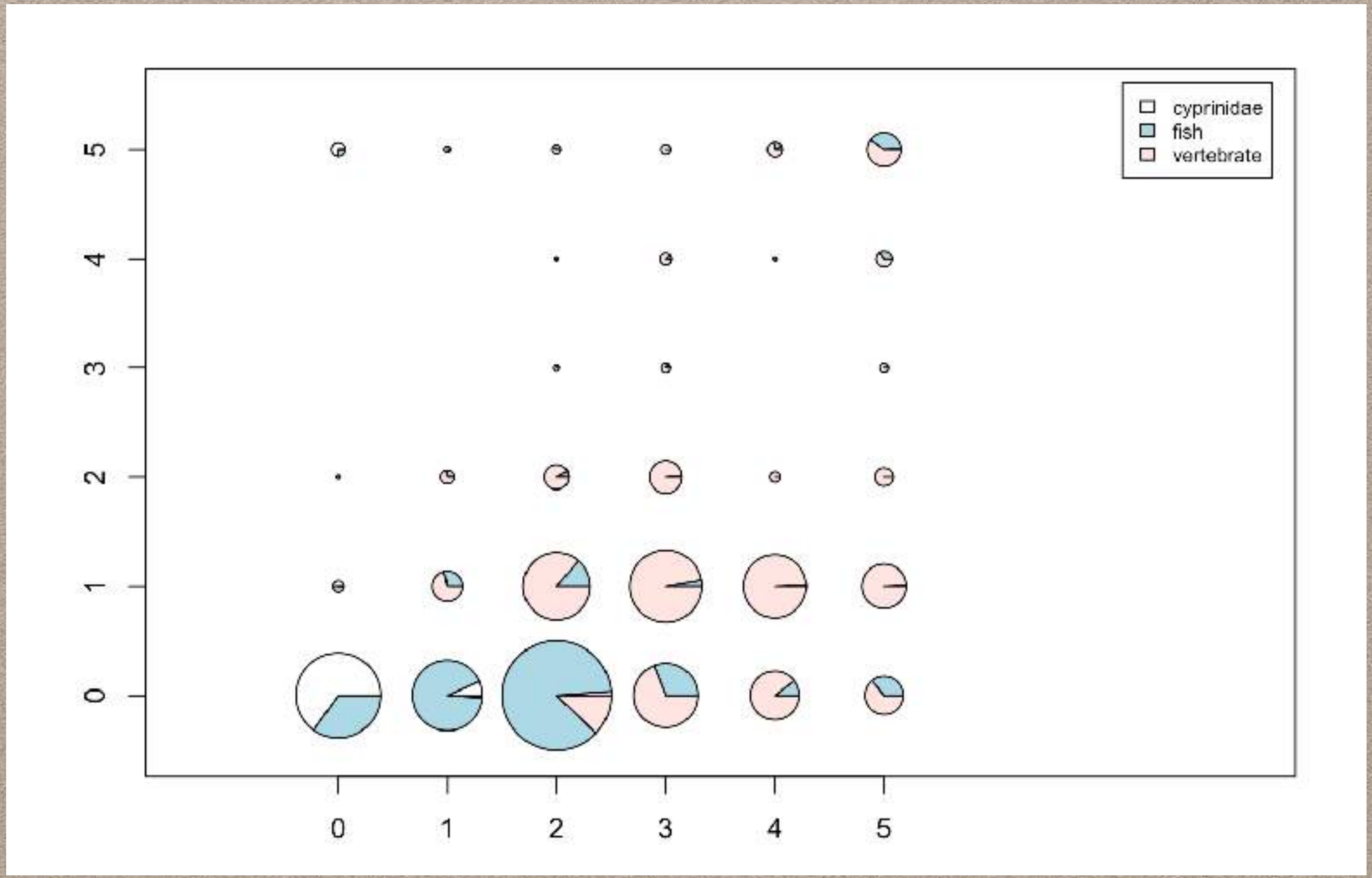
RUNNING AGAIN ECOPCR WITH THE NEW PRIMERS...

```
ecoPCR -d mito.vert \
       -e 5 \
       -l 30 -L 500 \
       -c \
ACGGCGTAAAGGGTGGTT TATCTAATCCCAGTTTGT \
> Cyprinidae.14.vert.ecopcr
```

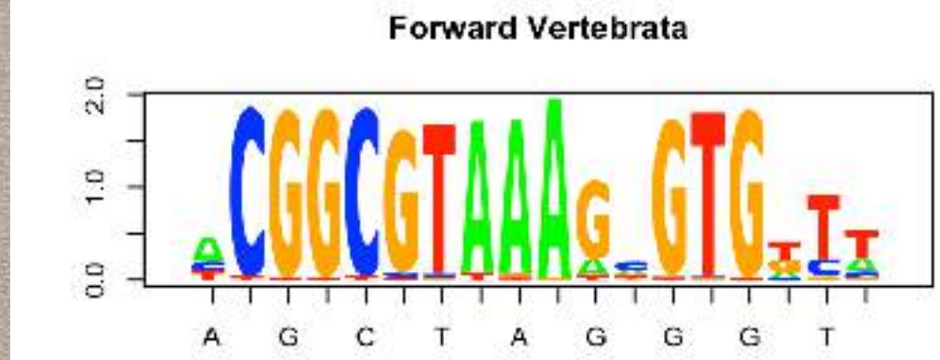
SOME BASIC STATISTICS

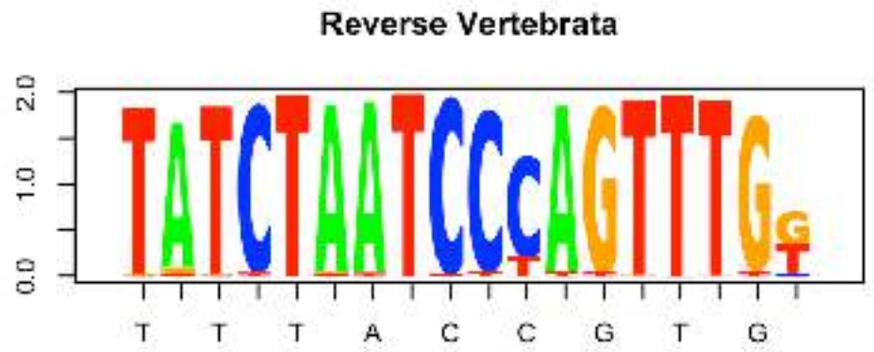
```
cyprinidae = read.ecopcr.result('Cyprinidae.14.vert.ecopcr')

is_a_fish=is.subcladeof(taxo,cyprinidae$taxid,32443)
is_a_cyprinidae = is.subcladeof(taxo,cyprinidae$taxid,7953)

group = rep('vertebrate',length(is_a_fish))
group[is_a_fish]='fish'
group[is_a_cyprinidae]='cyprinidae'
group=as.factor(group)

table(group)
group
cyprinidae       fish vertebrate
       333       1318       1626
```

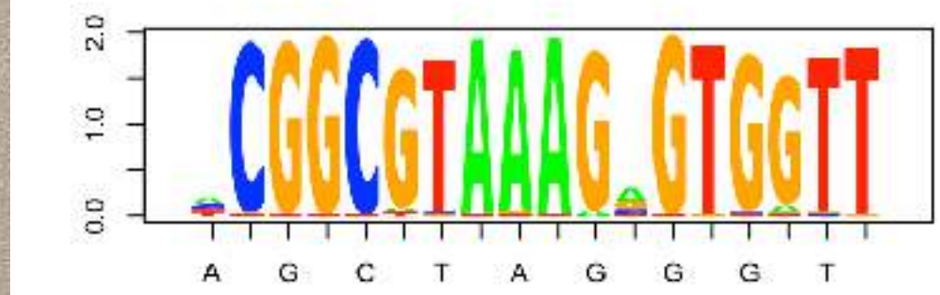
HOW PRIMERS WILL BE SELECTIVE ?

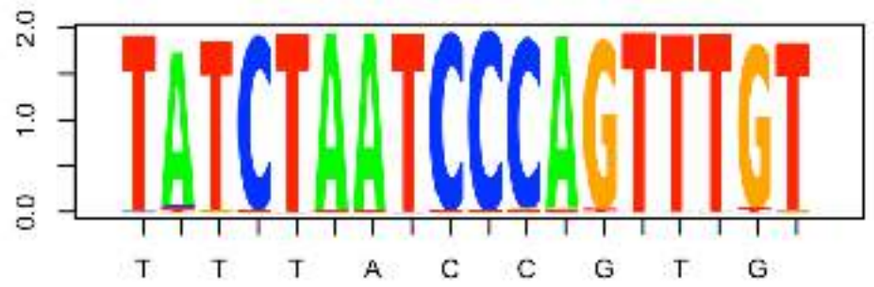


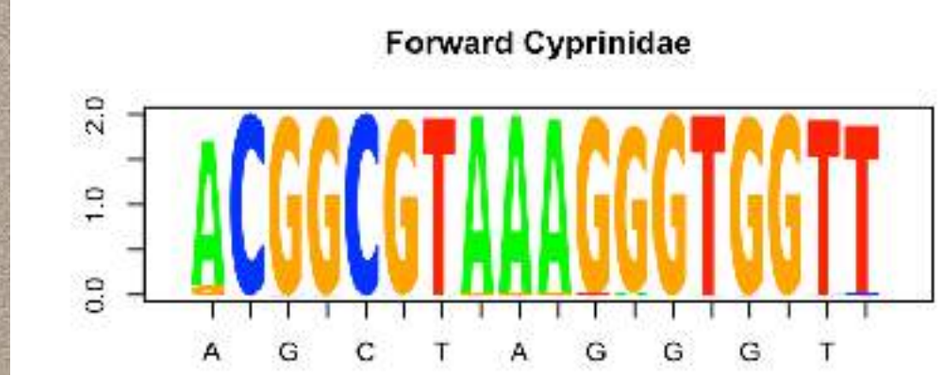
HOW PRIMING SITES ARE CONSERVED ?

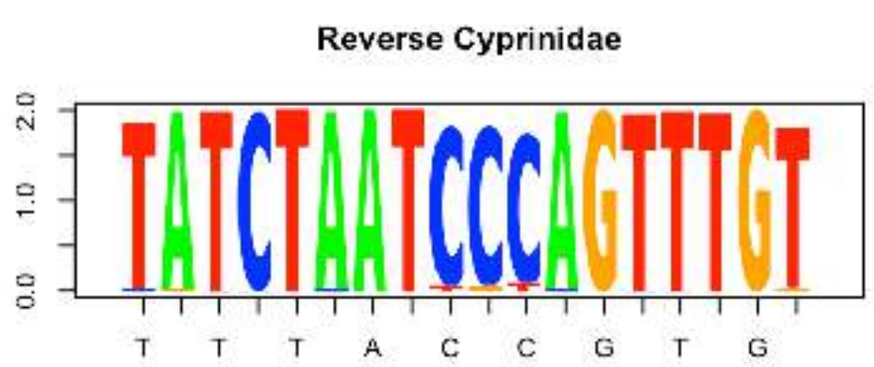










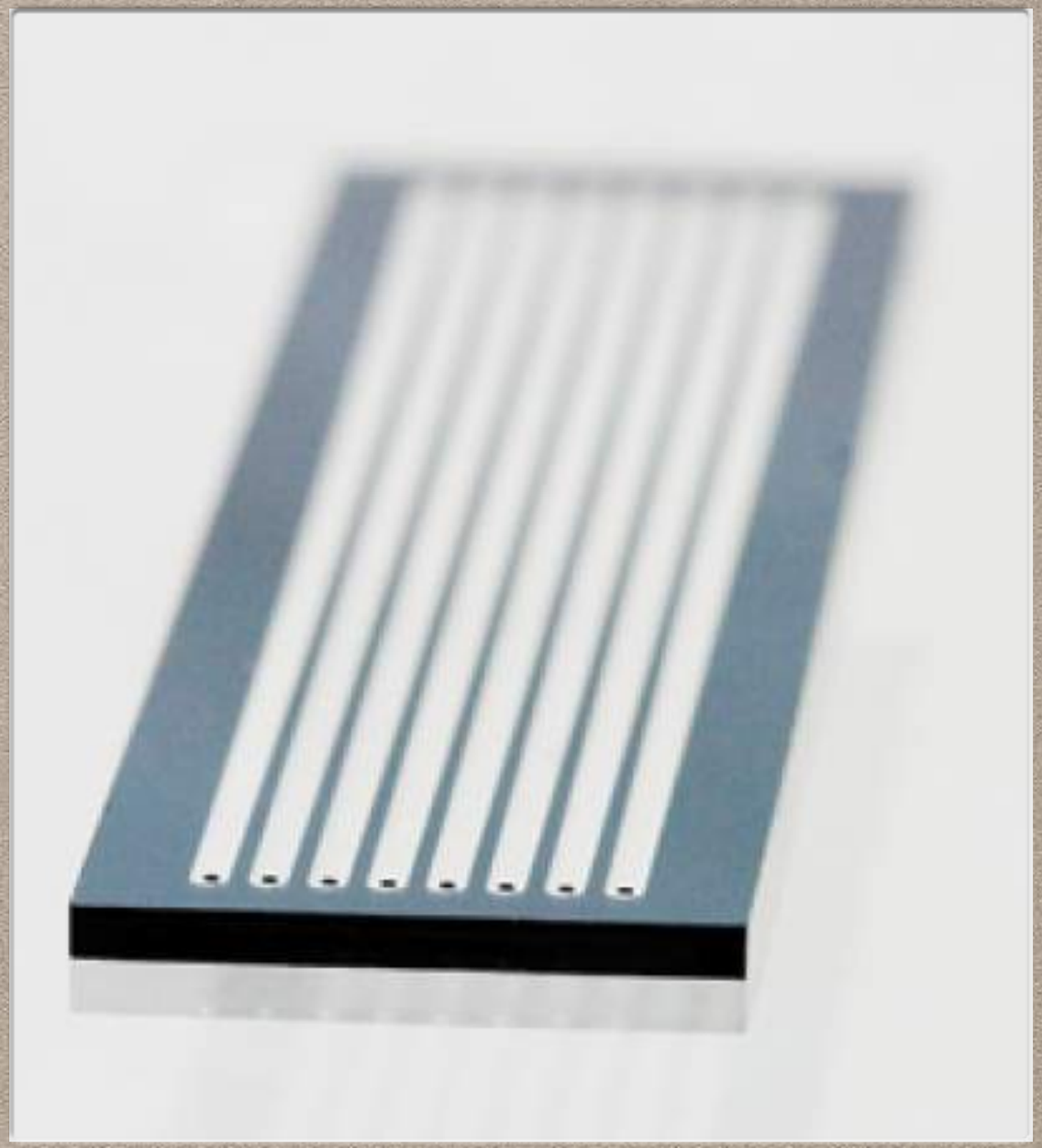


HOW MUCH BARCODE IS DISCRIMINENT ?

	cyprinidae	fish.primers
species	0.78947368	0.66563467
family	0.11455108	0.26006192
genus	0.06191950	0.06811146
subspecies	0.01547988	0.00619195

AN ILLUMINA FLOW CELL

2 X 8 LANES
ONLY 16 SAMPLES ?



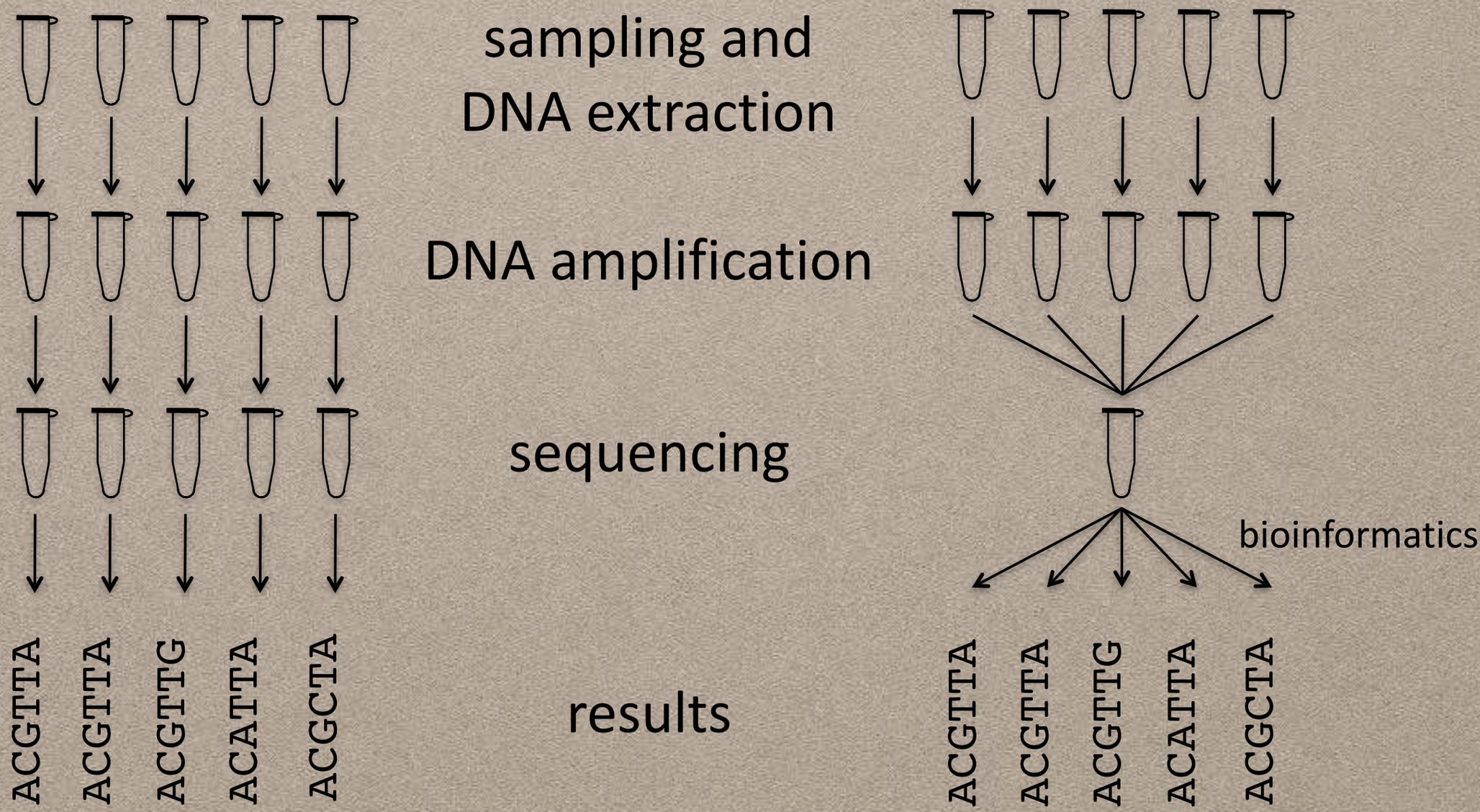
AN IDEA OF THE HISEQ 2000 PRODUCTION PER RUN

- 6 billions of reads of 100 bp
- 6 lines per read
- 55 lines per page (time 11)
- 654 545 454 pages
- 194 400 km long,
- 70.5 km high
- more than 3000 tons of paper

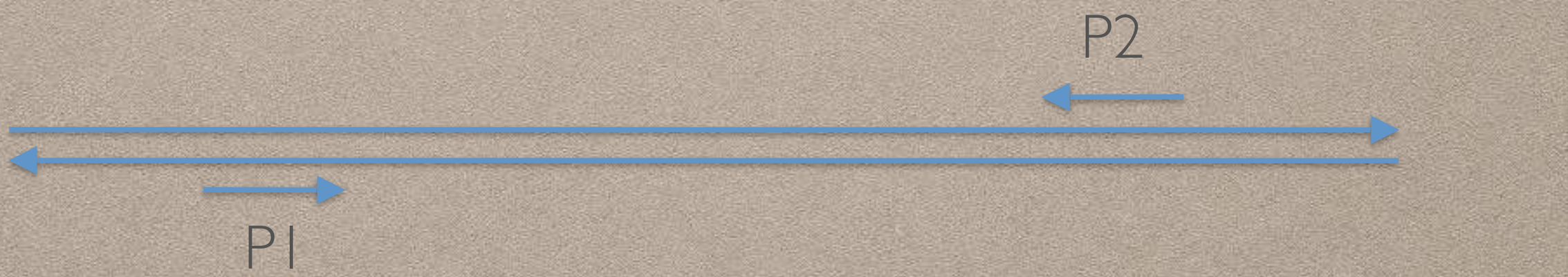
SAMPLE MULTIPLEXING

traditional sequencing

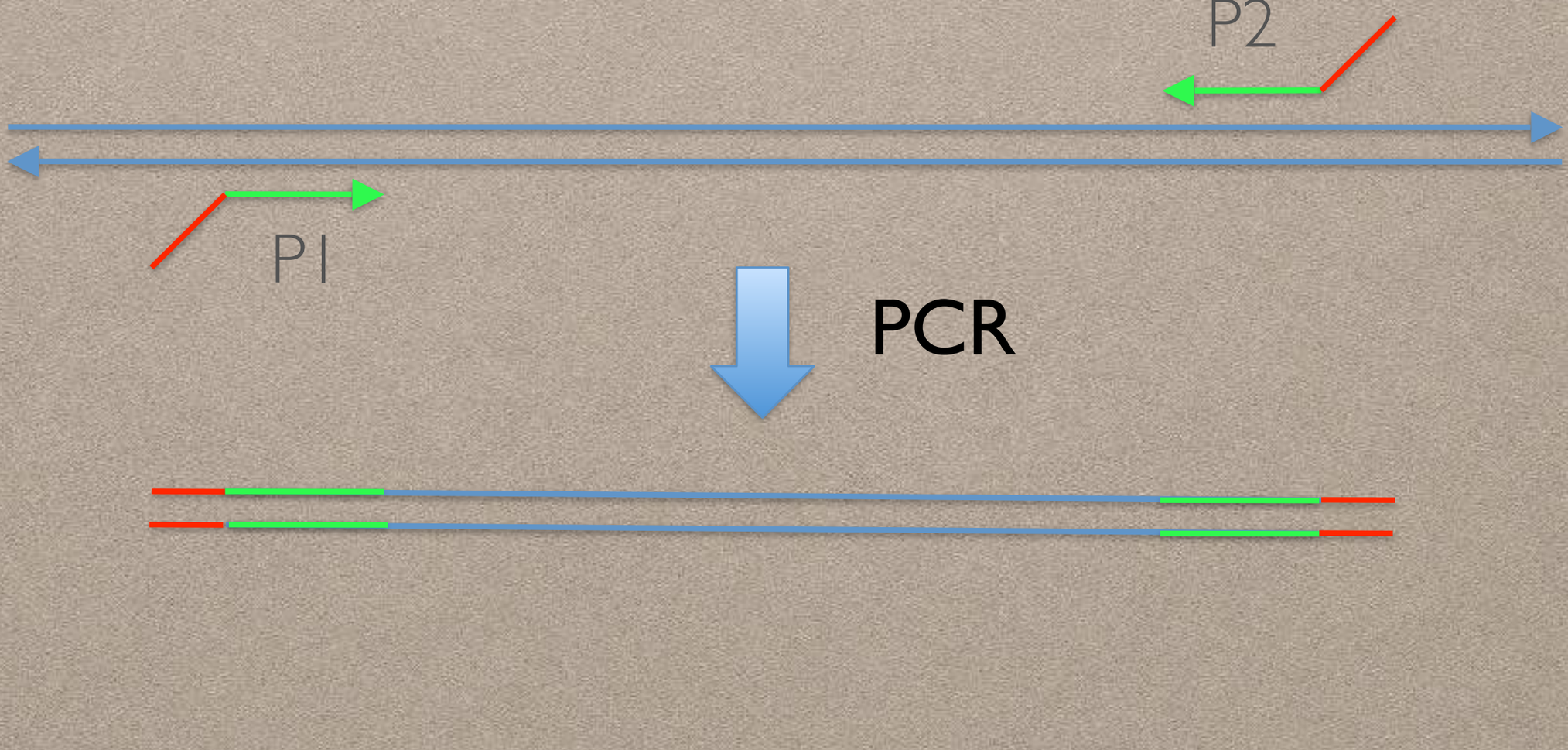
next generation sequencing



ADD TAGS TO PRIMERS



ADD TAGS TO PRIMERS



MULTIPLEXING SAMPLES

Add small tag in front of barcode primer for identifying samples

TAATGC

MULTIPLEXING SAMPLES

Add small tag in front of barcode primer for identifying samples

TAATGC → TAAGGC

MULTIPLEXING SAMPLES

Add small tag in front of barcode primer for identifying samples

TAATGC → TAAGGC

TAATGC → TAAGGC → TTAGGC

MULTIPLEXING SAMPLES

Add small tag in front of barcode primer for identifying samples

TAATGC → TAAGGC

TAATGC → TAAGGC → TTAGGC

TAATGC → TAAGGC → TTAGGC → TTAGCC

MULTIPLEXING SAMPLES

Add small tag in front of barcode primer for identifying samples

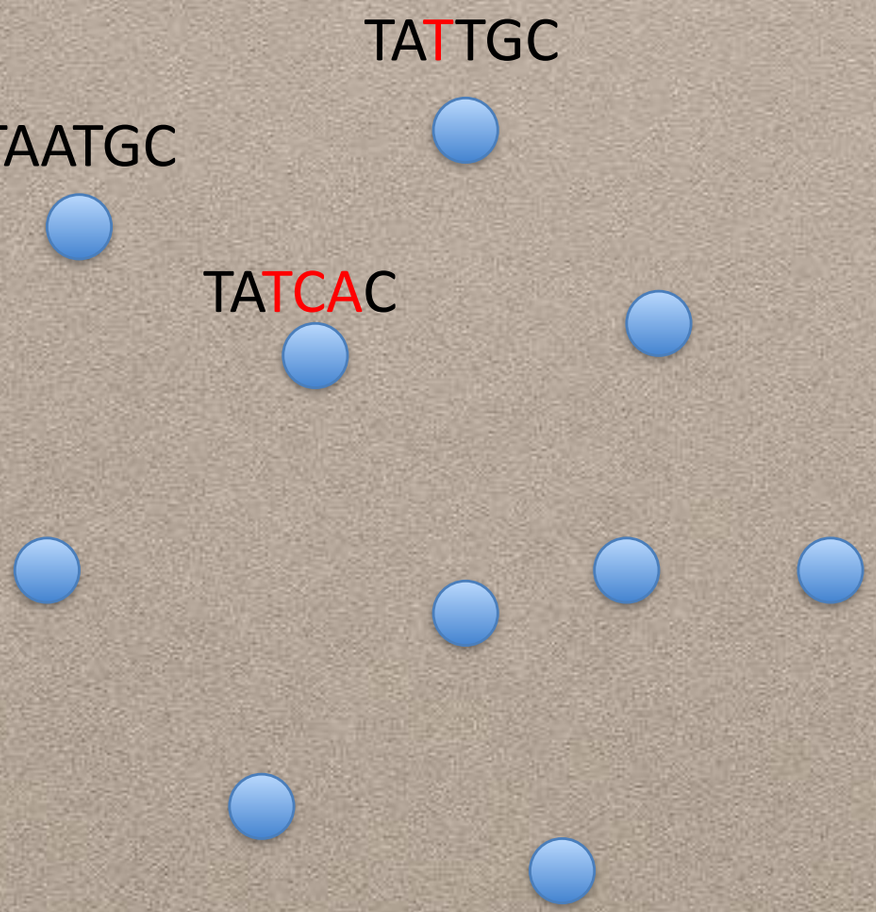
TAATGC → TAAGGC

TAATGC → TAAGGC → TTAGGC

TAATGC ⇄ TAAGGC → TTAGGC ⇄ TTAGCC

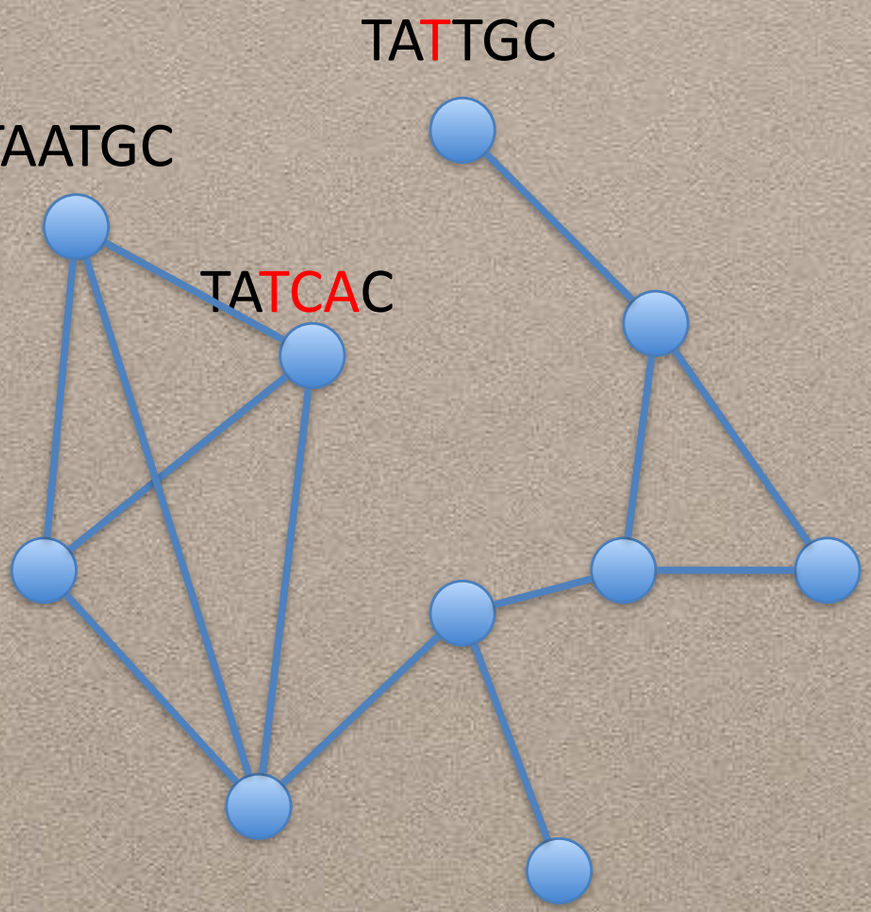
HOW DESIGN TAG SETS ?

oligoTag : part of OBITools



DESIGN OF TAGS

oligoTag : part of OBITools

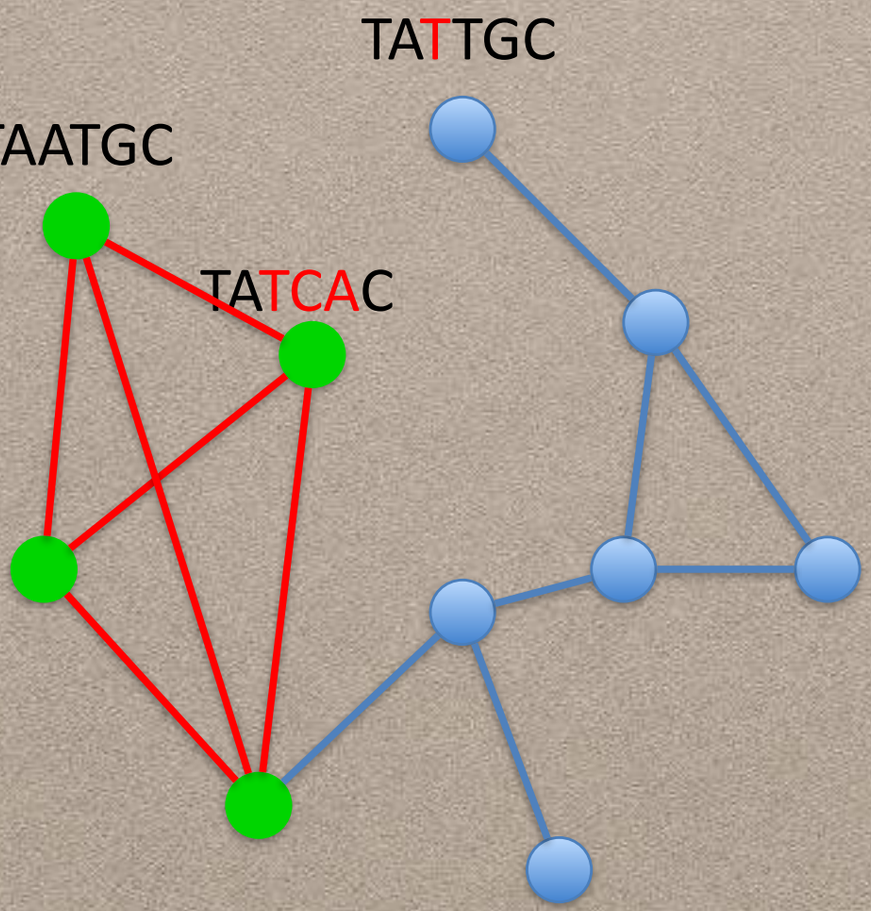


d_{min}=3

Link node when
hamming distance ≥ Threshold

DESIGN OF TAGS

oligoTag : part of OBITools



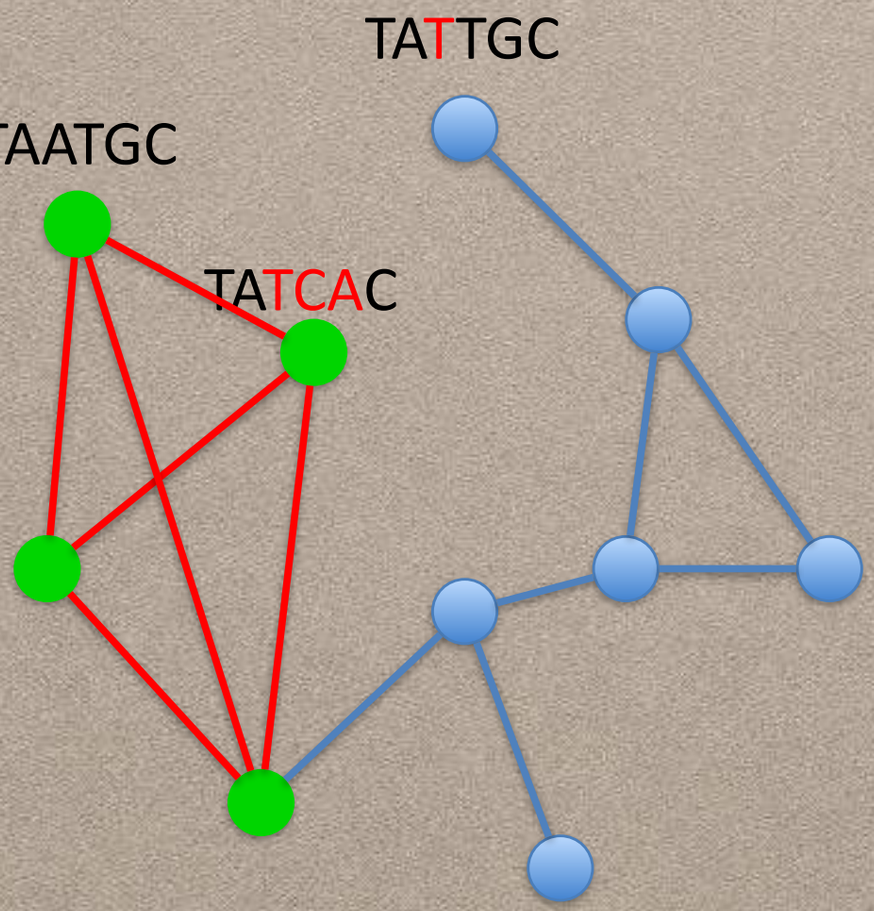
d_{min}=3

Link node when
hamming distance ≥ Threshold

Identifying larger complete subgraph or the clique max

DESIGN OF TAGS

oligoTag : part of OBITools



d_{min}=3

With $d \geqslant 3$
And $l_{homopolymer} \leqslant 2$

d_{min}	l	clique size
2	4	55
	5	173
	6	606
3	4	7
	5	33
	6	97
4	8	100

Thank you to :

Frédéric Boyer
Aurélie Bonin
Céline Mercier
Tiayyba Riaz
François Pompanon
Francesco Ficetola
Lucie Zinger
Pierre Taberlet

Waiting for satellite
biodiversity assessment...
Why not DNA metabarcoding ?