Design and evaluation of PCR primers for metabarcoding

A Coruña DNA Metabarcoding Workshop
A Coruña, Spain, 27th & 28th June 2018

Daniel Marquina

Department of Bioinformatics and Genetics, Swedish Museum of Natural History
Department of Zoology, Stockholms Universitet

OUTLINE

- HOW TO BE A GOOD METABARCODING MARKER?
- WHERE TO LOOK FOR PRIMERS?
- MISMATCHES AND HOW TO OVERCOME THEM
- PRIMER DESIGN SOFTWARE
- *IN SILICO* EVALUATION OF PRIMERS AND MARKERS
- MULTIPLEXING

How to be a good metabarcoding marker?

PRIMERS FOR BARCODING

- Amplifies DNA from a single specimen.
- Can be species-specific.
- Amplification success high-low.
- Amplicon sequence must be very variable between closely related species.
- Short to long amplicon size.
- Length between 18-30 bp.

PRIMERS FOR METABARCODING

- Amplifies DNA from many specimens.
- Must be universal (within target group).
- Amplification success equally high for all taxa.
- Amplicon sequence must be very variable between closely related species, but flanked by conserved regions.
- Amplicon length limited by resolution capacity (> 100 bp) and sequencing technologies + eDNA degradation (<400).
- The longer, the lower probability of been conserved (18-22 bp).

How to be a good metabarcoding marker?

SOME COMMON METABARCODING MARKERS

- PROKARYOTA — rRNA 16S
- EUKARYOTA — rRNA 18S (N)
- Metazoa — COI (M), rRNA 18S (N)
  - “Invertebrates” — COI, rRNA 16S (M), rRNA 18S (N)
  - Arthropoda — COI, rRNA 16S (M)
  - Vertebrates — rRNA 16S (M)
  - Fishes — rRNA 16S, rRNA 12S, CytB (M)
- Fungi — ITS (N)
- Plants — rbcL, matK (P), ITS (N)

How to be a good metabarcoding marker?

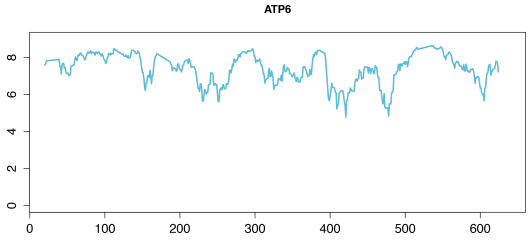
THINGS TO KEEP IN MIND BEFORE CHOOSING A MARKER

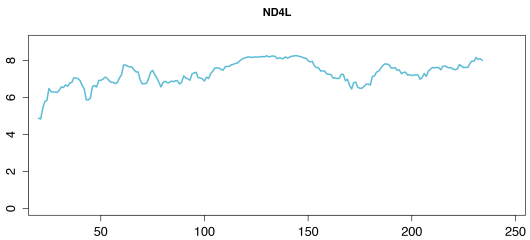
- Copy number:
  - MANY mitochondrion vs ONE nucleus.
  - In the nucleus: many copies of rRNA genes.
- Degradation in eDNA samples:
  - Mitochondria has an extra layer of protection & circular genome.
- Ribosomal vs protein-coding genes
  - Very high variability in length / Very high variability in the 3rd position
- Reference databases:
  - Species list or ecological indices

Where to look for primers?

COVERAGE – RESOLUTION TRADE-OFF: ENTROPY

Very high mutation rate genes

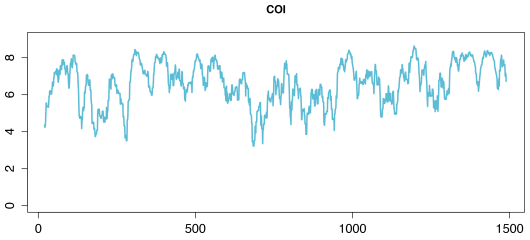


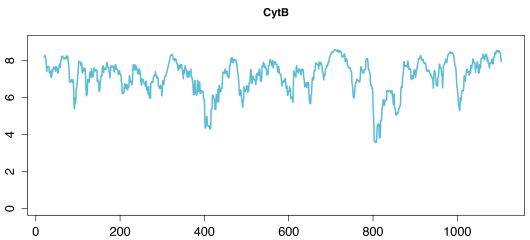


Too variable to find good primers

✗

High mutation rate genes

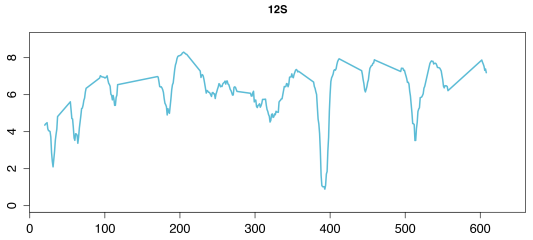


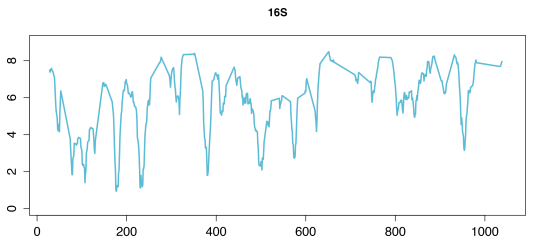


Difficult to find good primers, but possible. Good resolution with short amplicons.

✓

Lower mutation rate genes





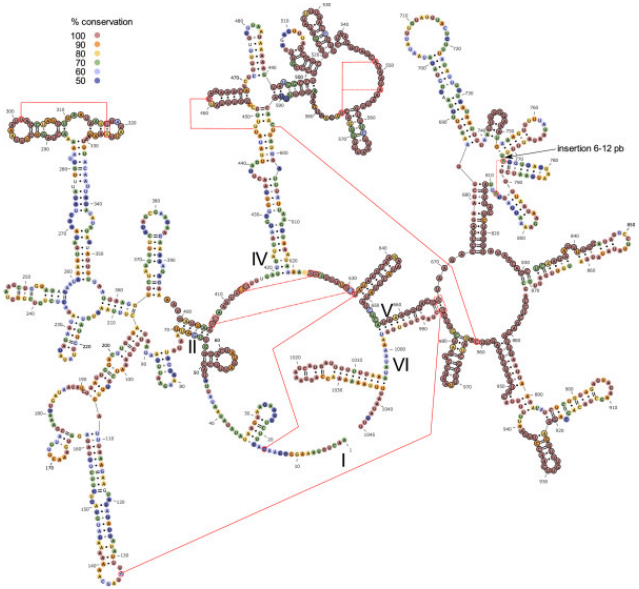
Easier to find good primers, but lower resolution requires longer amplicons.

✓

Where to look for primers?

COVERAGE – RESOLUTION TRADE-OFF: Protein-coding vs Ribosomal

Ribosomal genes



- Very conserved regions
- High variability in length

Protein-coding genes

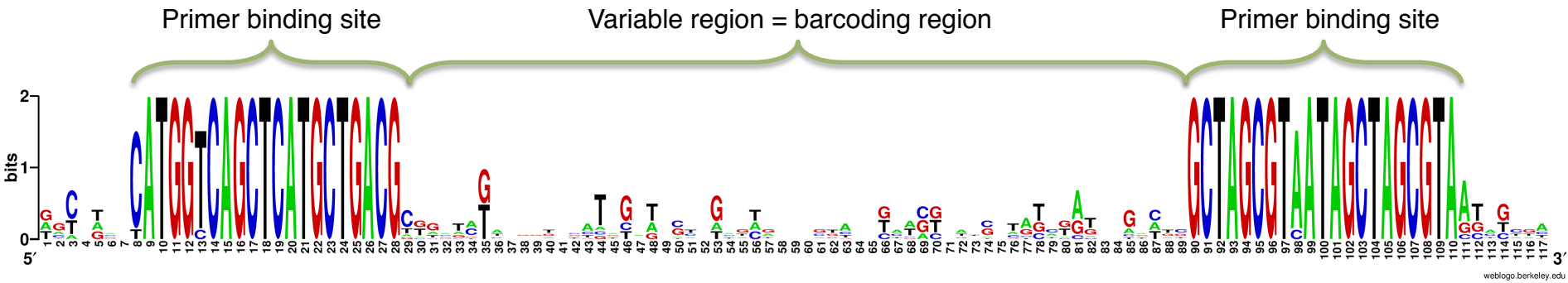
Second Letter

First Letter	T	C	A	G	Third Letter
T	TTT Phe, TTC Phe, TTA Leu, TTG Leu	TCT Ser, TCC Ser, TCA Ser, TCG Ser	TAT Tyr, TAC Tyr, TAA Stop, TAG Stop	TGT Cys, TGC Cys, TGA Stop, TGG Trp	T C A G
C	CTT Leu, CTC Leu, CTA Leu, CTG Leu	CCT Pro, CCC Pro, CCA Pro, CCG Pro	CAT His, CAC His, CAA Gln, CAG Gln	CGT Arg, CGC Arg, CGA Arg, CGG Arg	T C A G
A	ATT Ile, ATC Ile, ATA Ile, ATG Met	ACT Thr, ACC Thr, ACA Thr, ACG Thr	AAT Asn, AAC Asn, AAA Lys, AAG Lys	AGT Ser, AGC Ser, AGA Arg, AGG Arg	T C A G
G	GTT Val, GTC Val, GTA Val, GTG Val	GCT Ala, GCC Ala, GCA Ala, GCG Ala	GAT Asp, GAC Asp, GAA Glu, GAG Glu	GGT Gly, GGC Gly, GGA Gly, GGG Gly	T C A G

- Variability even in āā conserved regions
- Very constant in lenght

16S: Pons, J., Bauzà-Ribot, M. M., Jaume, D., & Juan, C. (2014). Next-generation sequencing, phylogenetic signal and comparative mitogenomic analyses in Metacrangonyctidae (Amphipoda: Crustacea). BMC genomics, 15(1), 566.

Where to look for primers?



- No homopolymers
- CG content 40–60 % evenly distributed
- 3’ end codon with C & G

THIS WILL NEVER HAPPEN!!

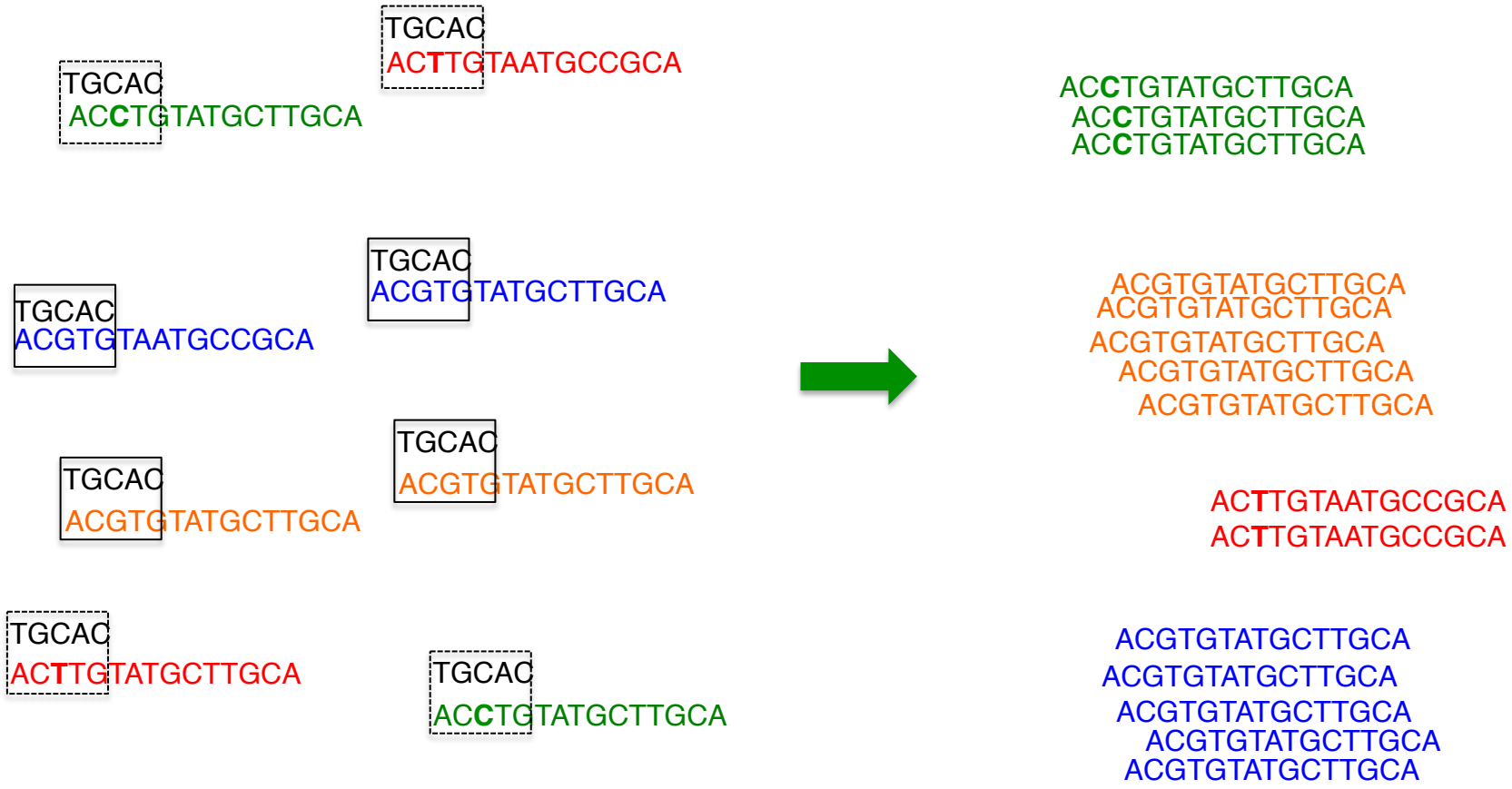
Mismatches and how to overcome them

We have found a conserved region with the following sequence being the most abundant:

But this is what the alignment really looks like:

Mismatches and how to overcome them

Primer: TGCAC; Diptera; Hymenoptera; Coleoptera; Lepidoptera



Mismatches and how to overcome them

DEGENERACY

- Use of ‘wobbly bases’ in the primer sequence that indicate a mix of sequences
- IUPAC code
  - Two bases **R** (A/G), **Y** (C/T), **M** (A/C), **K** (G/T), **S** (C/G), **W** (A/T)
  - Three bases **H** (A/C/T), **B** (C/G/T), **V** (A/C/G), **D** (A/G/T)
  - Four bases **N** (A/C/G/T)
- Calculated as no. possible sequences in the mix

CTA**B**GTA**D**CTA**S**GT**W**GC

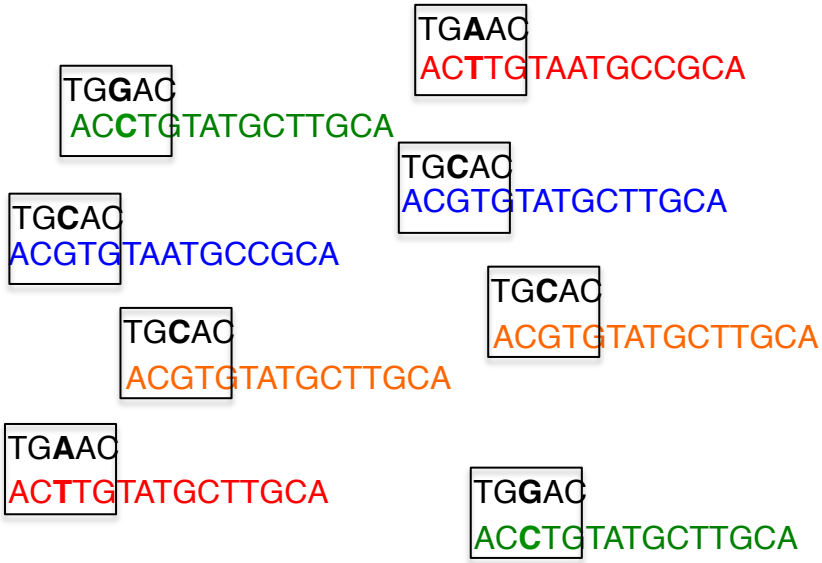
3 x 3 x 2 x 2 = 12

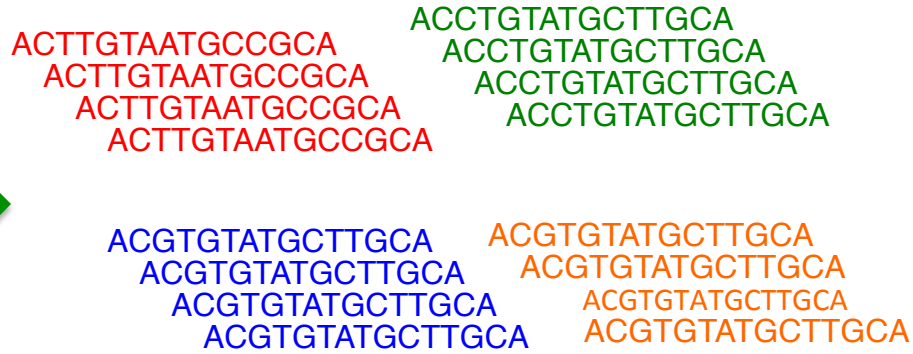
- Position of degenerate bases is very important!

Mismatches and how to overcome them

DEGENERACY

Primer: TGVAC; Diptera; Hymenoptera; Coleoptera; Lepidoptera





Mismatches and how to overcome them

DEGENERACY: Too much of something good can be bad

- With every new wobbly base the concentration of unique sequences diminishes. The proportion of different sequences in the primer does not match the proportion in the template.

- Very high degeneracy has risks:
  - Binding off-site in the genome
  - Binding to non-target taxa (rRNA!)
  - Tm temperatures very different between unique sequences
  - Higher probability of homopolymers and dimer formation

Primer design software

- ‘Non-computational’: custom download of sequences, custom alignment, visual search for primers.
  - Bioedit
  - Geneious

- ‘Visual-computational’
  - PrimerMiner (Elbrecht & Leese 2016): download sequences, custom alignment, assisted visual search for primers with entropy guidance. R package. YouTube tutorials.

- ‘Computational’
  - ecoPrimers (Riaz *et al.* 2011): custom download of sequences, alignment-free, automated search for primers. Combined with OBITools package (python).
  - DegePrime (Hugerth *et al*. 2014): custom download of sequences, custom alignment, automated search for **degenerate** primers. Perl scripts.

Primer design software

ecoPrimers: Strict Primer Algorithm (SPA)

E

ATTCGGCTACTAACT
ATACGGCTACTAACT
ATACGGCTACTAACT
ATACGGCTAGTAACT
ATTCGGCTACTAAGT
ATTCGGCTACTAAGT
ATTCGGCTACTAAGT
ATTCGGCTACTAAGT

Words of length *L* present in at least *S* sequences of **E**
L: number (18-21)
S: percentage (defaut=70)

ATACGGCTACTAACT
ATTCGGCTACTAAGT

Words of length *L* present in at least *S* sequences of **E**, and present in *T* sequences of **E** with no more than *m* mismatches.
T: percentage (default=90)
m: number (1-3)

*Lp'(**E**)*

ATACGGCTACTAACT
ATACGGCTAGTAACT
ATTCGGCTACTAAGT

Finds a space **D** within the interval of amplified sequence length [l_{min}-l_{max}] and creates Lp'(**D**). Pairs Lp'(**E**)-Lp'(**D**).

ATACGGCTACTAACT - ATACGGCTACTAACT
ATACGGCTAGTAACT - ATACGGCTAGTAACT
ATTCGGCTACTAAGT - ATTCGGCTACTAAGT

Positive:
Evaluates primers inmediately.
Constrains no mismatches in 3'-end of the primer.
Pairs primers within an interval of barcode length.
Considers 'countersequences'.

Negative:
No degeneracy allowed. Mismatches are mismatches.
Very format-constrained.
Not very flexible, impossible to consider alternative primers (blackbox).

Primer design software

ecoPrimers: Strict Primer Algorithm (SPA)

$ ecoPrimers -d sixlegs -e 3 -l 50 -L 500 -r 6960 -3 2 -c > insects_primers.ecoprimers

```
#
# ecoPrimer version 0.3
# Rank level optimisation : species
# max error count by oligonucleotide : 3
#
# Restricted to taxon:
#     6960 : Hexapoda (superclass)
#
# strict primer quorum  : 0.70
# example quorum        : 0.90
# counterexample quorum : 0.10
#
# database : sixlegs
# Database is constituted of  1602 examples        corresponding to  1115 species
#                         and     0 counterexamples corresponding to     0 species
#
# amplifiat length between [50,500] bp
# DB sequences are considered as circular
# Pairs having specificity less than 0.60 will be ignored
#
     0  ATAGAAACCAACCTGGCT      TTACCTTAGGGATAACAG      53.6    1.7     47.7    27.0    8       7       GG      1514    0       0.945   1059    0       0.950   835     0.788   138     217     142.67
     1  ATAGAAACCAACCTGGCT      TACCTTAGGGATAACAGC      53.6    1.7     50.6    30.9    8       8       GG      1502    0       0.938   1048    0       0.940   824     0.786   137     216     141.67
     2  GATAGAAACCAACCTGGC      TACCTTAGGGATAACAGC      53.6    2.0     50.6    30.9    9       8       GG      1499    0       0.936   1046    0       0.938   822     0.786   138     217     142.67
     3  ATAGAAACCAACCTGGCT      GACCTCGATGTTGGATTA      53.6    11.4    51.8    38.1    8       8       GG      1498    0       0.935   1045    0       0.937   671     0.642   79      158     83.67
     4  GATAGAAACCAACCTGGC      TTACCTTAGGGATAACAG      53.6    2.0     47.7    27.0    9       7       GG      1511    0       0.943   1057    0       0.948   654     0.619   139     218     143.67
```

Primer sequences

Tm, A temperatures

GC #

Sequence & taxa statistics

Max, min & average size of amplicon

Primer design software

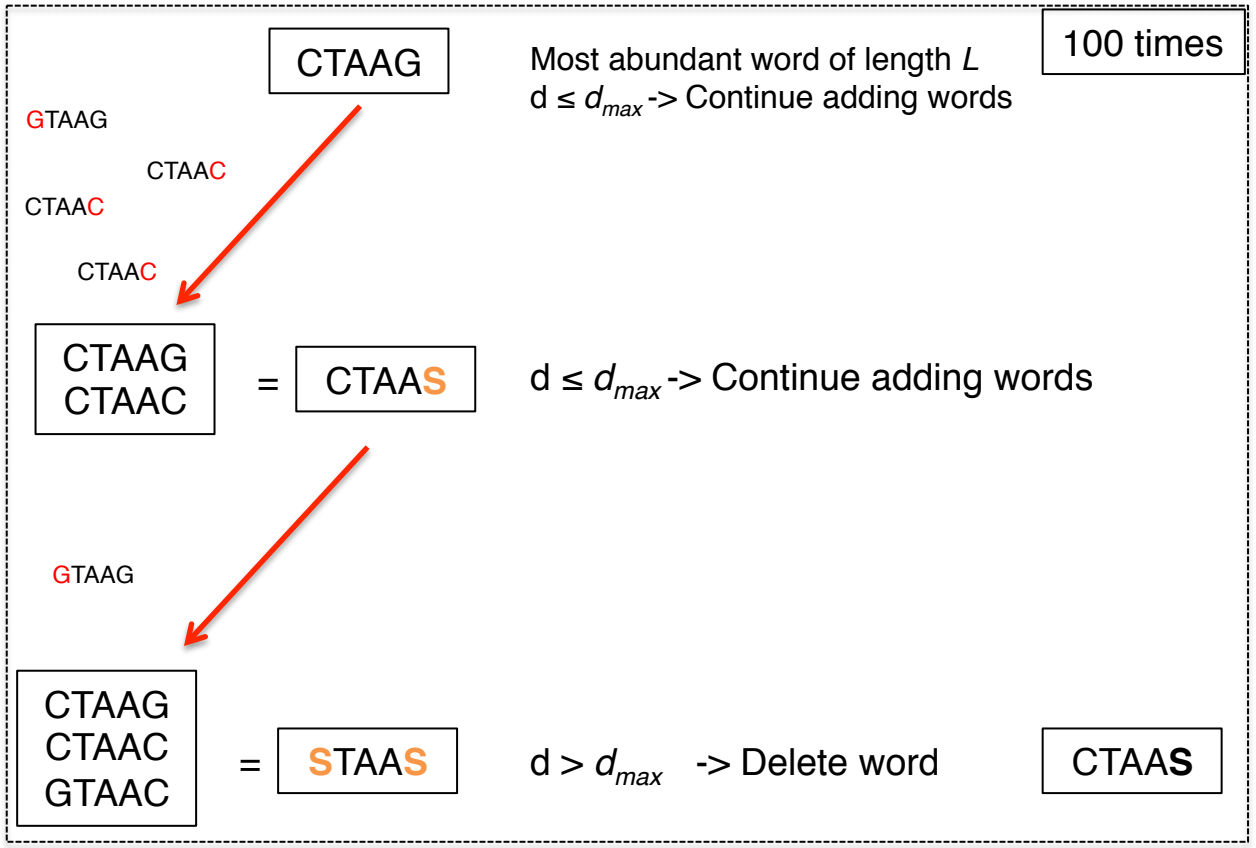
$d_{max} = 2$

ATTCGGCTACTAACT
ATACGGCTACTAACT
ATACGGCTACTAACT
ATACGGCTAGTAACT
ATTCGGCTACTAAGT
ATTCGGCTACTAAGT
ATTCGGCTACTAAGT
ATTCGGCTACTAAGT

Positive: Allows degeneracy (not mismatches).
Gives measure of sequence diversity.
Much more flexible, user decides which primer is best.

Negative: No 3'-end constrain.
No pairing at length interval.
Too much degeneracy when not needed.

DegePrime: Weighed Randomized Combination



Primer design software

DegePrime: Weighed Randomized Combination

```
$ perl TrimAlignment.pl -i COI_aligned.fasta -min 0.9 -o COI_trimmed
$ perl DegePrime.pl -i COI_trimmed -d 12 -l 18 -o COI_12_primer
```

Pos	TotalSeq		UniqueMers	Entropy	PrimerDeg		PrimerMatching	PrimerSeq
28	1102	268	7.48379621854699		8	104	ATATWATTATTTTAATRY	
29	1102	339	7.38153610376529		12	133	AAAWTTAWTTTHAATATT	
30	1102	351	7.4692352932659	12	117	AAWTTAWTTHAAATATTT		
31	1102	386	7.54756592401369		12	101	AATTATTWHWAATATTTG	
32	1102	388	7.58315528709506		12	101	AATTATTYWVATATTTGG	
33	1103	401	7.53849495330789		8	104	ATTATTTTAATRYYTGGT	
34	1103	393	7.4448370074104	12	112	TAWTTTAAATATTTRDTC		
35	1103	391	7.43047410793722		12	112	AWTTTAAATATTTRDTCC	
36	1104	378	7.38105179555872		12	176	WTTTHAATATTTGRTCCT	
37	1108	365	7.24742751747064		12	225	TTTWAATATTTRDTCCTT	
38	1109	351	7.16808385113982		12	239	TTWAATATTTRDTCCTTT	
39	1109	329	7.05284872213714		12	257	HWAATATTTGRTCCTTTC	
40	1109	269	6.55154354374301		12	374	WAATATTTRDTCCTTTCG	
41	1109	225	5.98699172403398		12	451	AATATTWRDTCCTTTCGT	
42	1109	196	5.4783962275216	12	548	ATATTHRRTCCTTTCGTA		
43	1109	187	5.22167661484077		12	565	TATTHRRTCCTTTCGTAC	
44	1123	145	5.05480291515133		12	570	ATWTRDTCCTTTCGTACT	
45	1123	108	4.89196676712661		12	580	THTRRTCCTTTCGTACTA	
46	1123	86	4.3190556715473	12	739	HTRRTCCTTTCGTACTAA		
47	1123	81	4.21005427431276		12	733	TRRTCCTTTCGTACTAAD	
48	1137	94	4.38778629742602		12	708	RDTCCTTTCGTACTAAWA	
49	1138	123	4.56233579370246		12	704	DTCCTTTCGTACTRAWAT	
50	1138	117	4.1634201719062	12	709	TCCTTTCGTACTRAWDTA		
51	1142	46	5.25901778596935		12	358	CCTYTCGTACTAARDTAT	
52	1143	53	5.81979600727008		12	319	CTTTCGTACTAAANTATH	
53	1145	25	6.36448501434196		12	22	TYTCGTACTAARDTATAA	
74	1147	104	3.78304221579834		12	744	HAARGATAGAAACCRACC	
75	1147	84	3.20851598204967		12	905	DARGATAGAAACCRACCT	
76	1147	64	2.73932404768507		12	998	DRGATAGAAACCRACCTG	
77	1147	53	2.31697916635107		12	1029	RGATAGAAACCVAYCTGG	
78	1147	51	1.84232151630854		12	1051	GATAGAAACCVAYCTGRC	
79	1147	50	1.83770042648698		12	1052	ATAGAAACCVAYCTGRCT	
80	1147	60	2.70634555623366		12	1030	TAGAAACCVAYCTGGCTY	
81	1147	69	2.78725147850408		12	1018	AGAAACCVAYCTGGCTYA	
82	1147	70	2.84960237485124		12	1005	GAAACCVAYCTGGCTYAC	
83	1147	74	3.535267851515	12	986	AAACCRACCTGGCTYACV		
84	1147	73	3.52730560949005		12	987	AACCRACCTGGCTYACVC	
85	1147	72	3.51933729111953		12	988	ACCRACCTGGCTYACVCC	
86	1147	63	3.48334974098986		12	987	CCRACCTGGCTYACVCCG	

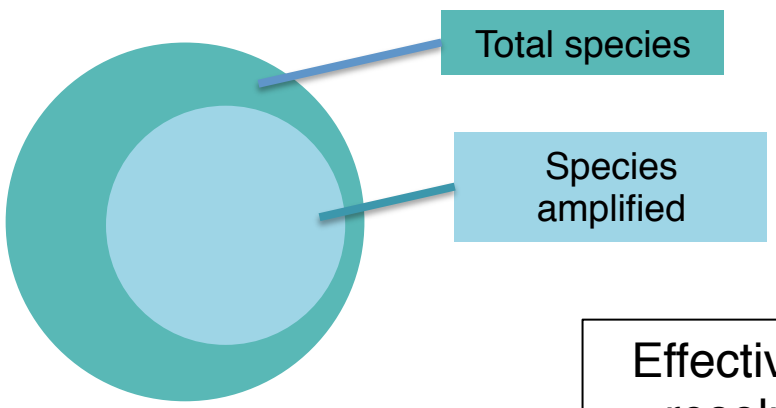
Primer design software

DegePrime: Weighed Randomized Combination

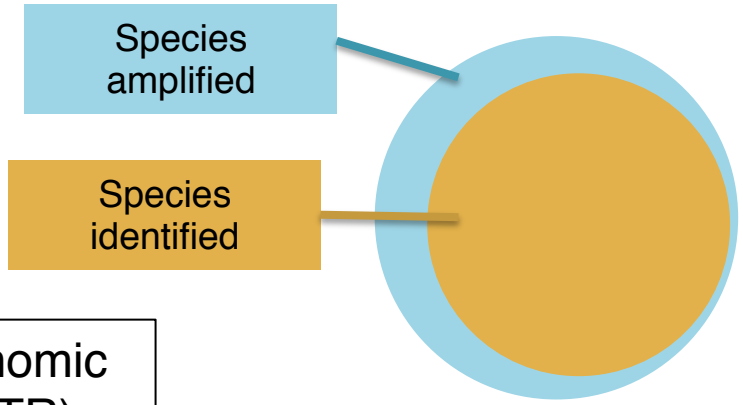
SeqMatched;	Position;	Sequence;	Entropy;	Degneracy
1124	235	GCTGTTATCCCYDARGTA	1.18472930134388	12
1119	236	CTGTTATCCCTDARGTAW	1.30281588929785	12
1119	177	TARTYCAACATCGRGGTC	0.943419523590396	8
1118	176	HTARTYCAACATCGAGGT	0.993034683932689	12
1113	234	YGCTGTTATCCCTDARGT	1.41155527609562	12
1113	178	ARTYCAACATCGAGGTCD	1.14169299463664	12
1110	181	YCAACATCGAGGTCRHAA	1.18575168118205	12
1110	180	TYCAACATCGAGGTCDYA	1.18575168118205	12
1105	106	YTDAACTCARATCATGTA	2.19358211227913	12
1103	107	TDARCTCARATCATGTAA	1.40288819616529	12
1102	233	AYGCTGTTATCCCTDARG	1.46427498638947	12
1101	105	TYTDAACTCARATCATGT	2.22363483361992	12
1100	179	ATYCAACATCGAGGTCRH	1.2520525289237	12
1096	231	TWAYGCTGTTATCCCTDA	1.12484399581344	12
1095	502	AYTATGCTACCTTHGYAC	2.09393069325809	12
1094	380	TYTAHAGGGTCTTMTCGT	1.79765322876033	12
1094	238	GTTATCCCTAARGTADYT	2.14920382960784	12
1094	237	TGTTATCCCTAARGTADY	2.14920382960784	12
1093	182	CAACATCGAGGTCRYAAH	1.94081389862307	12
1092	381	YTAHAGGGTCTTMTCGTC	1.81166015943193	12
1089	230	ATWAYGCTGTTATCCCTD	1.14625815990182	12
1088	504	TATGCTACCTTHGYACRG	2.64470359914178	12
1085	505	ATGCTACCTTHGYACRGT	2.6632459891823	12
1085	382	TAHAGGGTCTTMTCGTCY	2.02068856068754	12
1084	232	HAYGCTGTTATCCCTAAR	1.79593567327562	12
1084	108	RAACTCARATCATGTAAD	1.90729194561761	12
1082	109	AACTCARATCATGTAARD	2.00429613209112	12
1080	501	KATTATGCTACCTTHGYA	2.5048795153959	12
1080	229	DATWAYGCTGTTATCCCT	2.13741160103578	12
1075	500	TKATTATGCTACCTTHGY	2.54351291138962	12
1073	503	TTATGCTACCTTHGYACR	2.7279768141556	12
1070	183	AACATCGAGGTCRYAAHC	2.0787451154341	12

In silico evaluation of primers and markers

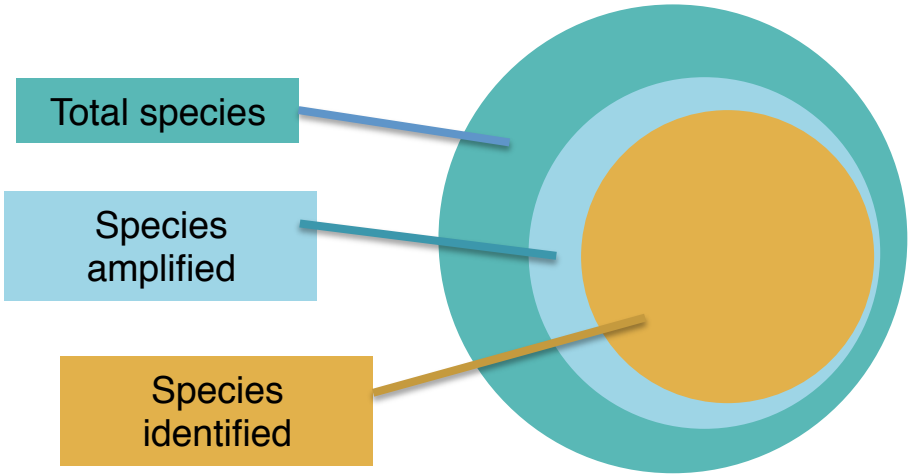
Taxonomic coverage (B_C)



Exclusive taxonomic resolution (B_E)



Effective taxonomic resolution (ETR)

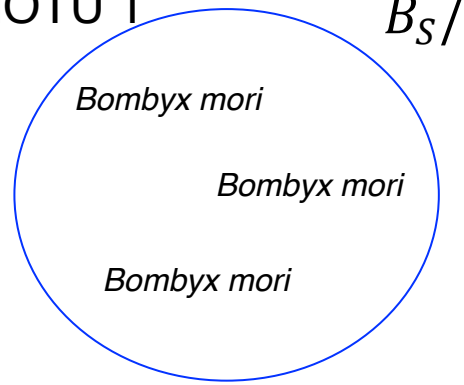


In silico evaluation of primers and markers

Taxonomic resolution (B_S) vs Exclusive taxonomic resolution (B_E)

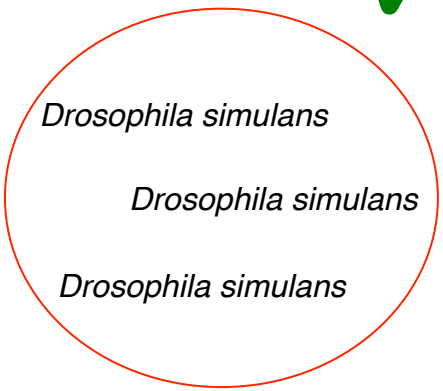
$$B_S/B_E = \frac{no.\ species\ unambiguously\ identified(*)}{no.\ species\ amplified} = [0\text{-}1]$$

MOTU 1

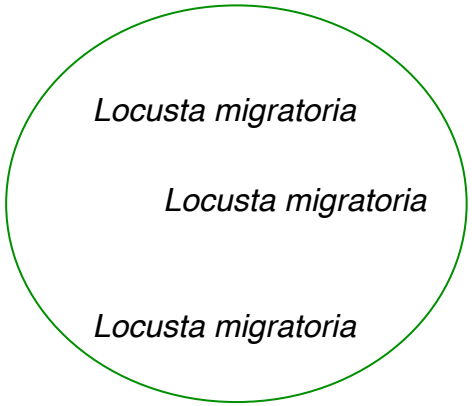


$B_S = 3/4 = 0.75$

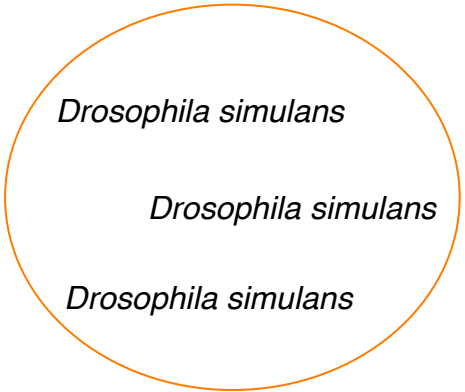
MOTU 2



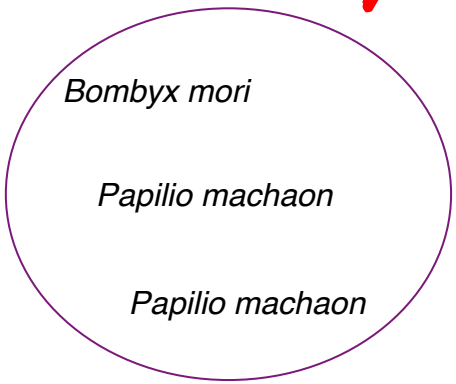
MOTU 3



MOTU 4



MOTU 5



In silico evaluation of primers and markers

Taxonomic resolution (B_S) vs Exclusive taxonomic resolution (B_E)

$$B_S/B_E = \frac{no.\ species\ unambiguously\ identified(*)}{no.\ species\ amplified} = [0\text{-}1]$$

MOTU 1

✗

Bombyx mori

Bombyx mori

Bombyx mori

$B_E = 1/4 = 0.25$

MOTU 2

✗

Drosophila simulans

Drosophila simulans

Drosophila simulans

MOTU 3

✓

Locusta migratoria

Locusta migratoria

Locusta migratoria

MOTU 4

✗

Drosophila simulans

Drosophila simulans

Drosophila simulans

MOTU 5

✗

Bombyx mori

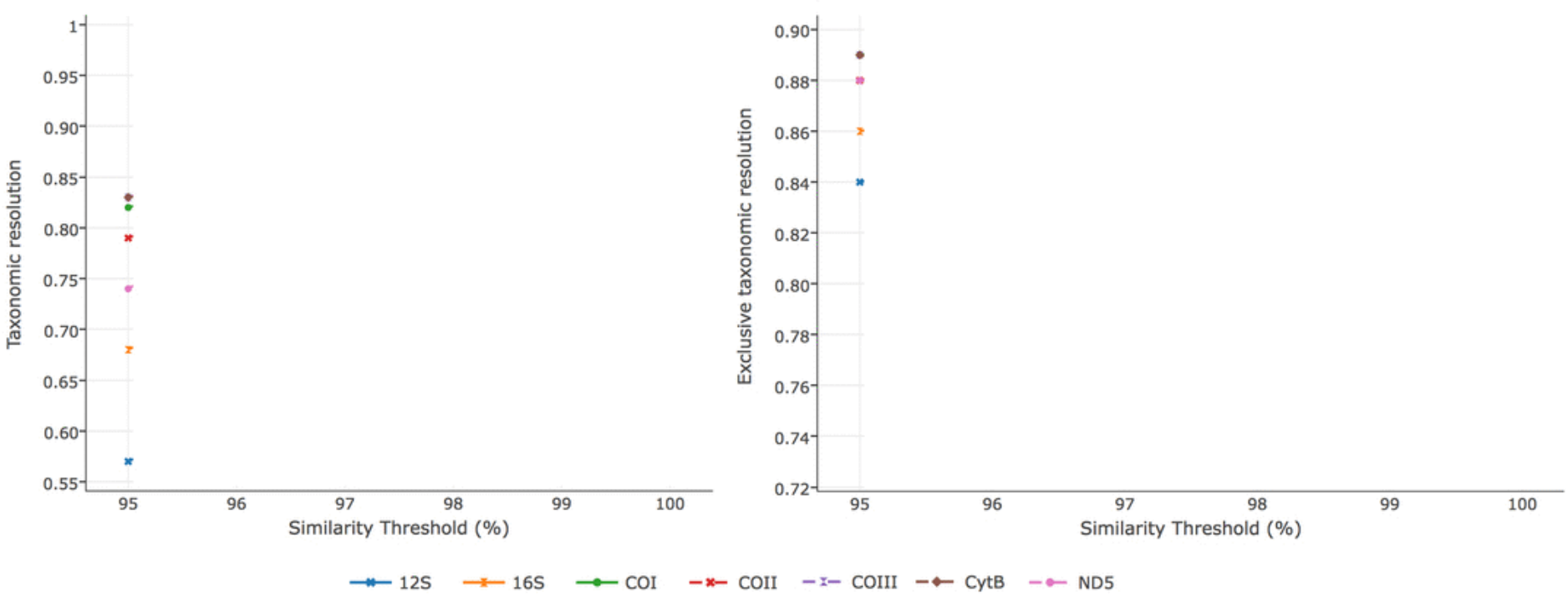
Papilio machaon

Papilio machaon

In silico evaluation of primers and markers

Taxonomic resolution (B_S) vs Exclusive taxonomic resolution (B_E)

B_E provides a more biologically accurate measure of the taxonomic resolution of a metabarcoding marker than B_S.



Multiplexing

- Sample tags attached at the 5' end of the Forward and Reverse primers:

F_tag1	NNNNCCATGCAGCTABGTADCTASGTWGC
F_tag2	NNCATGCATGCTABGTADCTASGTWGC
F_tag3	NNNNNNCCTAGCTGCTABGTADCTASGTWGC

	R_tag1	R_tag2	R_tag3
F_tag1	Sample 1	Sample 2	Sample 3
F_tag2	Sample 4	Sample 5	Sample 6
F_tag3	Sample 7	Sample 8	Sample 9

Thank you for your attention!
Questions?

Long presentation and questions:
daniel.marquina@nrm.se