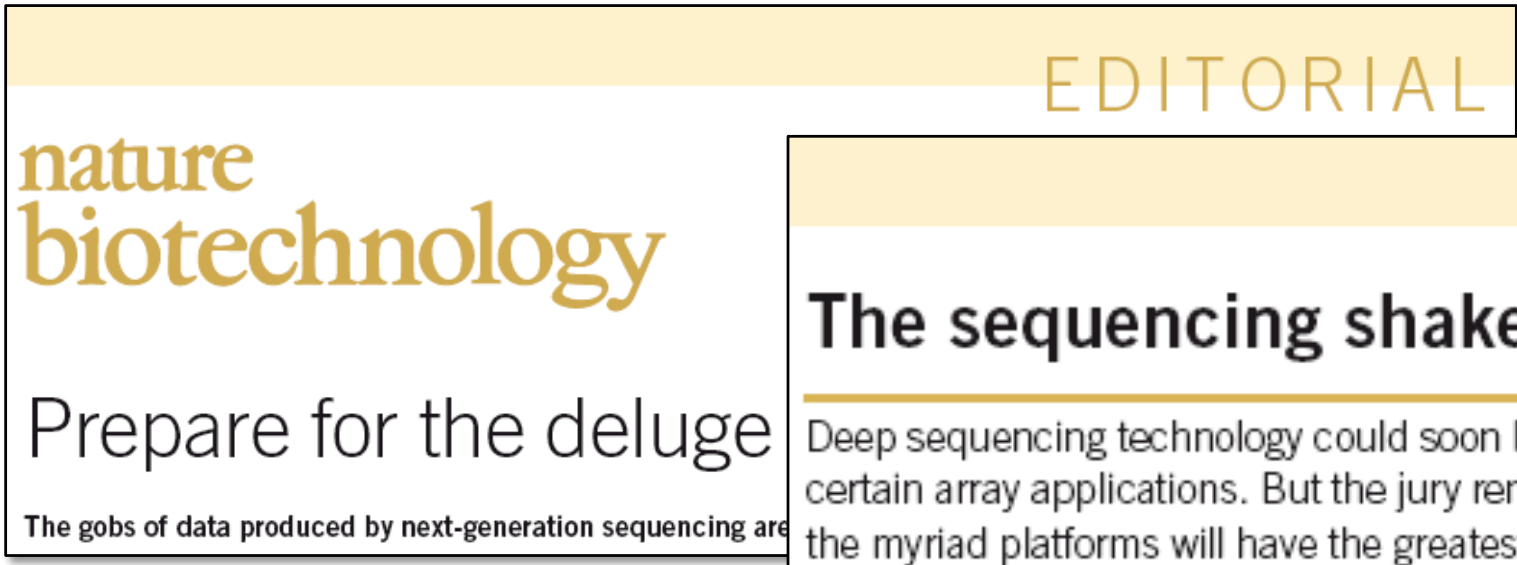
Biodiversity and DNA barcoding

Eric Coissac, Pierre Taberlet

Laboratoire d'Ecologie Alpine, CNRS UMR 5553
Université Joseph Fourier, Grenoble, France

Biodiversity and DNA barcoding

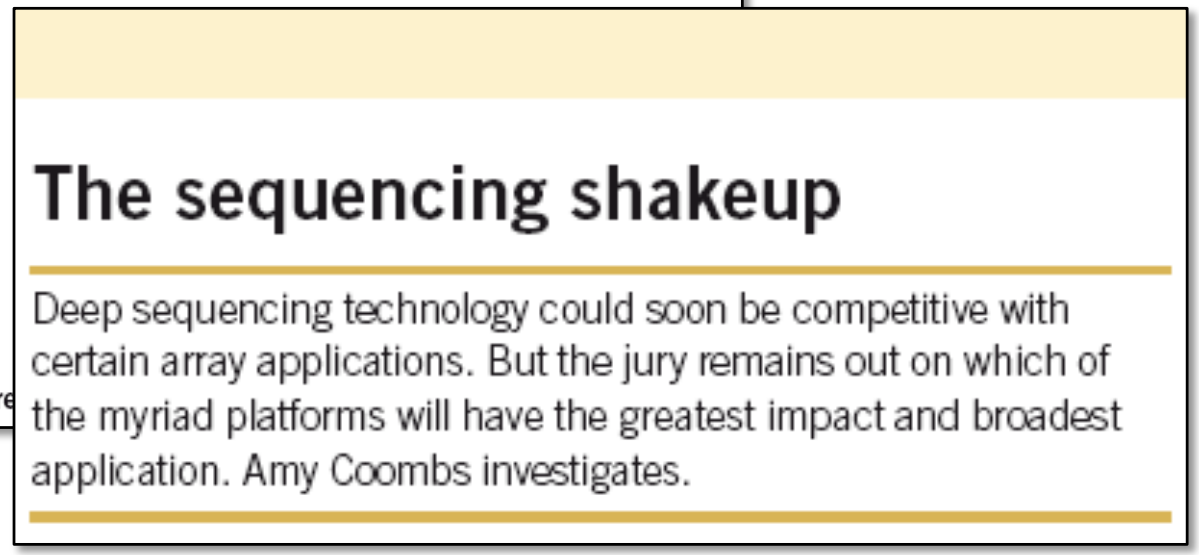
- The DNA barcoding concept
- Our scientific objectives
- New sequencing technologies
- Bioinformatic aspects
  - The challenges
  - Available programs for preparing the experiments
  - Available programs for analyzing the output of next generation sequencers
- High throughput plant identification
- High throughput animal identification
- Conclusion


EDITORIAL

nature biotechnology

Prepare for the deluge

The gobs of data produced by next-generation sequencing are


The sequencing shakeup

Deep sequencing technology could soon be competitive with certain array applications. But the jury remains out on which of the myriad platforms will have the greatest impact and broadest application. Amy Coombs investigates.

What would you do if you could sequence everything?

Nature Biotechnology, October 2008: special issue on next generation sequencing


How to get genomes at one ten-thousandth the cost

Jeffery A Schloss

The NHGRI's Advanced DNA Sequencing Technology program is spearheading the development of platforms that will bring routine whole-genome sequencing closer to reality.

Biodiversity and DNA barcoding

- The DNA barcoding concept
- Our scientific objectives
- New sequencing technologies
- Bioinformatic aspects
  - The challenges
  - Available programs for preparing the experiments
  - Available programs for analyzing the output of next generation sequencers
- High throughput plant identification
- High throughput animal identification
- Conclusion

Molecular barcodes for soil nematode identification

ROBIN FLOYD, EYUALEM ABEBE, ARTEMIS PAPERT and MARK BLAXTER
Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, UK

Abstract

Using a molecular barcode, derived from single-specimen polymerase chain reaction (PCR) and sequencing of the 5′ segment of the small subunit ribosomal RNA (SSU) gene, we have developed a molecular operational taxonomic unit (MOTU) scheme for soil nematodes. Individual specimens were considered to belong to the same MOTU when the sequenced segment of 450 bases was > 99.5% identical. A Scottish upland *Agrostis-Festuca* grassland soil was sampled, using both culture-based and random selection methods. One hundred and sixty-six cultured isolates were sequenced, and clustered into five MOTU. From 74 randomly sampled individuals across the study site, 19 MOTU were defined. A subsequent sample of 18 individuals from a single subplot contained eight MOTU, four of which were unique to the single subplot sample. Interestingly, seven of these MOTU were not present in the culture-independent sampling. Overall, a total of 23 MOTU were defined from only 240 sequences. Many MOTU could readily be assigned to classical, morphologically defined taxonomic units using a database of SSU sequences from named nematode species. The MOTU technique allows a rapid assessment of nematode taxon diversity in soils. Correlation with a database of sequences from known species offers a route to application of the technique in ecological surveys addressing biological as well as genetic diversity.

Keywords: biodiversity assessment, DNA sequence, nematodes, 18S ribosomal RNA (SSU)

Received 17 August 2001; revision received 7 December 2001; accepted 7 December 2001

Floyd *et al.* (2002) *Molecular Ecology*, **11**, 839-850.

THE ROYAL SOCIETY

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

Biological identifications through DNA barcodes

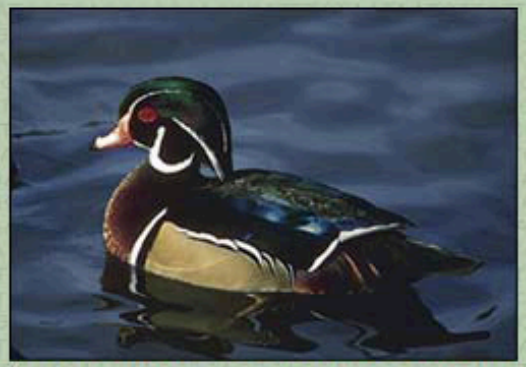
Paul D. N. Hebert*, Alina Cywinska, Shelley L. Ball and Jeremy R. deWaard

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

Although much biological research depends upon species diagnoses, taxonomic expertise is collapsing. We are convinced that the sole prospect for a sustainable identification capability lies in the construction of systems that employ DNA sequences as taxon 'barcodes'. We establish that the mitochondrial gene cytochrome *c* oxidase I (COI) can serve as the core of a global bioidentification system for animals. First, we demonstrate that COI profiles, derived from the low-density sampling of higher taxonomic categories, ordinarily assign newly analysed taxa to the appropriate phylum or order. Second, we demonstrate that species-level assignments can be obtained by creating comprehensive COI profiles. A model COI profile, based upon the analysis of a single individual from each of 200 closely allied species of lepidopterans, was 100% successful in correctly identifying subsequent specimens. When fully developed, a COI identification system will provide a reliable, cost-effective and accessible solution to the current problem of species identification. Its assembly will also generate important new insights into the diversification of life and the rules of molecular evolution.

Keywords: molecular taxonomy; mitochondrial DNA; animals; insects; sequence diversity; evolution

Hebert *et al.* (2003) *Proc. Roy. Soc. London B*, **270**, 313-321.



Aix sponsa

Taxonomy			
Phylum:	Chordata	Specimen Accession:	1B-334
Class:	Aves	Specimen Label:	Aix sponsa
Order:	Anseriformes	Museum Holding:	ROM
Family:	Anatidae	Museum Accession:	1B-334x
Genus:	*Aix*	Country:	Canada
Species:	*sponsa*	Province/State:	Ontario
		Region:	Hamilton
Sex:	Male	GPS Latitude:	43.250
Reproduction:	Sexual	GPS Longitude:	-79.833
Life Stage:	Adult		

Aix sponsa

DNA Barcodes and Biodiversity

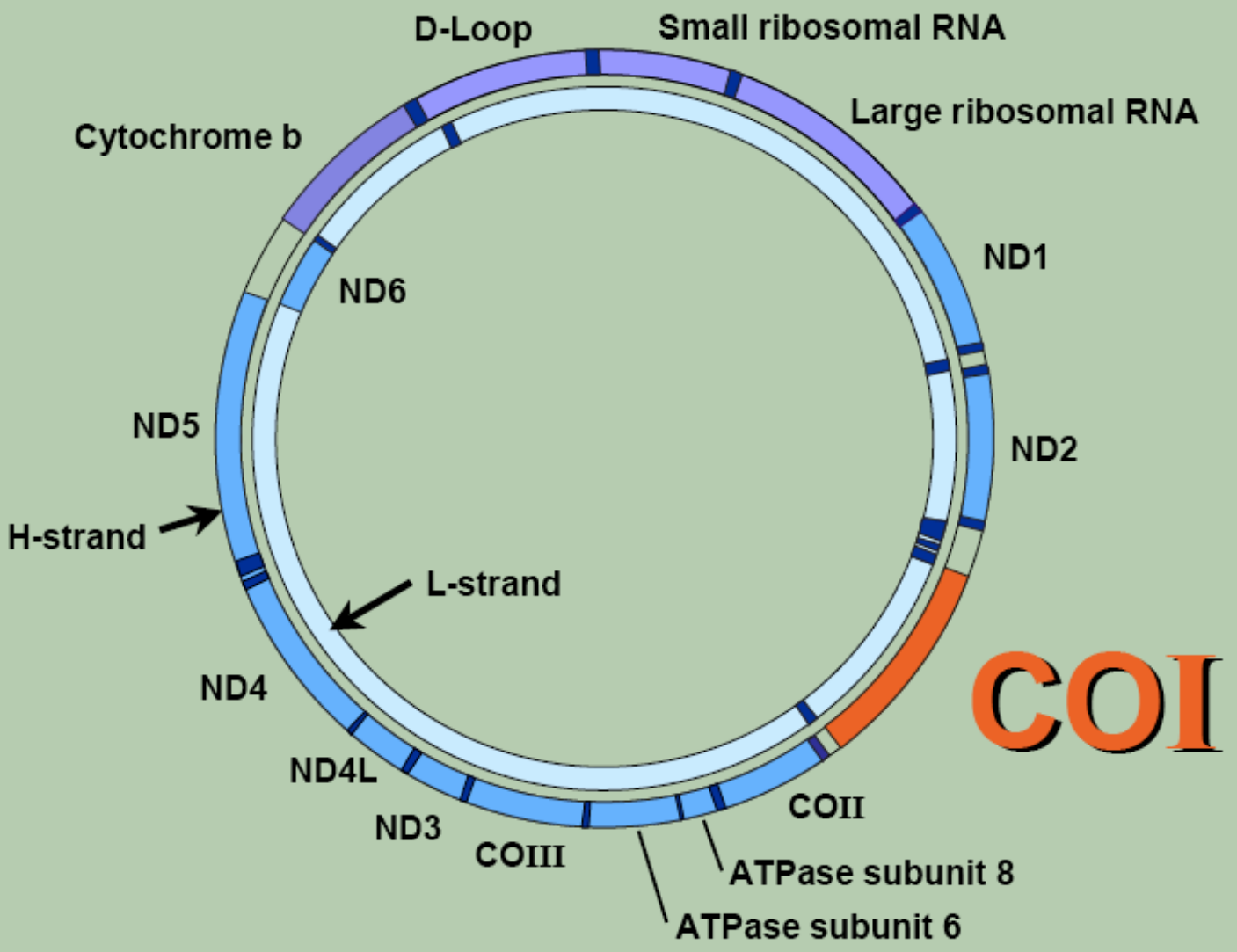
The "Consortium for the Barcode of Life" (CBoL)

Pandalus danae

Pandalus danae

Taxonomy			
Phylum:	Arthropoda	Specimen Accession:	43-SCF-26
Class:	Malacostraca	Specimen Label:	Pandalus
Order:	Decapoda	Museum Holding:	U of G
Family:	Pandalidae	Museum Accession:	43-SCF-26
Genus:	*Pandalus*	Country:	Canada
Species:	*danae*	Province/State:	BC
		Region:	Nanaimo
Sex:	Male	GPS Latitude:	52.511
Reproduction:	Sexual	GPS Longitude:	-132.143
Life Stage:	Adult		

The Mitochondrial Genome



Barcoding - The Future

Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification

Kipling W. Will[a,*] and Daniel Rubinoff[b]

[a]*University of California, Department Environmental Science, Policy and Management, Division of Insect Biology, Berkeley, CA 94720, USA*
[b]*University of Hawaii, Department of Plant and Environmental Protection Sciences, 3050 Maile Way, 310 Gilmore Hall, Honolulu, HI 96822, USA*

Accepted 8 December 2003

Abstract

So-called DNA barcodes have recently been proposed to answer the problem of specimen identification and to quantify global biodiversity. We show that this proposition is wanting in terms of rationale, methodology and interpretation of results. In addition to falling short of all its stated goals, the method abandons the benefits of morphological studies in favor of a limited molecular identification system that would ultimately impede our understanding of biodiversity.

Will & Rubinoff (2004) *Cladistics*, **20**, 47-55.

New . . but not good.

- **unjustified emphasis/reliance on DNA data**

- **the incorrect analogy to Universal Product Codes (UPC)**: typological thinking is not really new.

- **the concerted effort**, but largely to do the wrong thing

Will & Rubinoff (2004) *Cladistics*, **20**, 47-55.

DNA barcoding: definitions

- DNA barcoding *sensu stricto*: identification of the species level using a single standardized DNA fragment; (= CBoL view)
- DNA barcoding *sensu lato*: identification of any taxonomical level using any DNA fragment

Valentini A, Pompanon F, Taberlet P (2009) DNA barcoding for ecologists. *Trends in Ecology & Evolution*, **24**, 110-117.

The ideal DNA barcoding marker

- Variability: the gene region sequenced should be nearly identical among individuals of the same species, but different between species
- Standardization: the same DNA region used for different taxonomic groups
- Phylogenetic information: to easily assign unknown or not yet "barcoded" species to their taxonomic group (genus, family, etc.)
- Robustness: highly conserved priming sites, and highly reliable DNA amplifications and sequencing
- Shortness: amplification of degraded DNA

The ideal DNA barcoding marker

- Variability: the gene region sequenced should be nearly identical among individuals of the same species, but different between species
- Standardization: the same DNA region used for different taxonomic groups
- Phylogenetic information: to easily assign unknown or not yet "barcoded" species to their taxonomic group (genus, family, etc.)
- Robustness: highly conserved priming sites, and highly reliable DNA amplifications and sequencing
- Shortness: amplification of degraded DNA

The ideal DNA barcoding marker

- Variability: the gene region sequenced should be nearly identical among individuals of the same species, but different between species
- Standardization: the same DNA region used for different taxonomic groups
- Phylogenetic information: to easily assign unknown or not yet "barcoded" species to their taxonomic group (genus, family, etc.)
- Robustness: highly conserved priming sites, and highly reliable DNA amplifications and sequencing
- Shortness: amplification of degraded DNA

Next generation sequencers and biodiversity: the rare biosphere…

- Routinely used for bacteria, fungi, and nematodes
- Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 12115-12120.
- Buee M, Reich M, Murat C, Morin E, Nilsson RH, Uroz S, Martin F (2009) 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist*, **184**, 449-456.
- Porazinska DL, Giblin-Davis RM, Faller L, Farmerie W, Kanzaki N, Morris K, Powers TO, Tucker AE, Sung W, Thomas WK (2009) Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. *Molecular Ecology Resources*, **9**, 1439-1450.

Biodiversity and DNA barcoding

- The DNA barcoding concept
- Our scientific objectives
- New sequencing technologies
- Bioinformatic aspects
  - The challenges
  - Available programs for preparing the experiments
  - Available programs for analyzing the output of next generation sequencers
- High throughput plant identification
- High throughput animal identification
- Conclusion

Our scientific objectives

- Understanding the mechanisms underlying biodiversity (with an integrated approach)
- Biodiversity monitoring using standardized protocols
- DNA-based biodiversity assessment using environmental samples
- For plants and animals
- Many "by-products": diet analysis, biotechnological applications

The metabarcoding approach

- Sampling in the field (soil, water, feces, etc.)
- DNA extraction
- DNA amplification with barcode primers
- Sequencing of the PCR products on next generation sequencers
- Species identification using a reference database (or identification of MOTUs)

Biodiversity and DNA barcoding

- The DNA barcoding concept
- Our scientific objectives
- New sequencing technologies
- Bioinformatic aspects
  - The challenges
  - Available programs for preparing the experiments
  - Available programs for analyzing the output of next generation sequencers
- High throughput plant identification
- High throughput animal identification
- Conclusion

DNA sequencing

- Capillary electrophoresis
  - 500-1000 bp per sequencing reaction
  - 12 x 96 reactions per day (≈ 1 Mb per day)
- Next generation sequencers
  - Roche 454: ≈ 0.4 Gb per day
  - HiSeq 2000: ≈ 25 Gb per day

454 GS FLX™

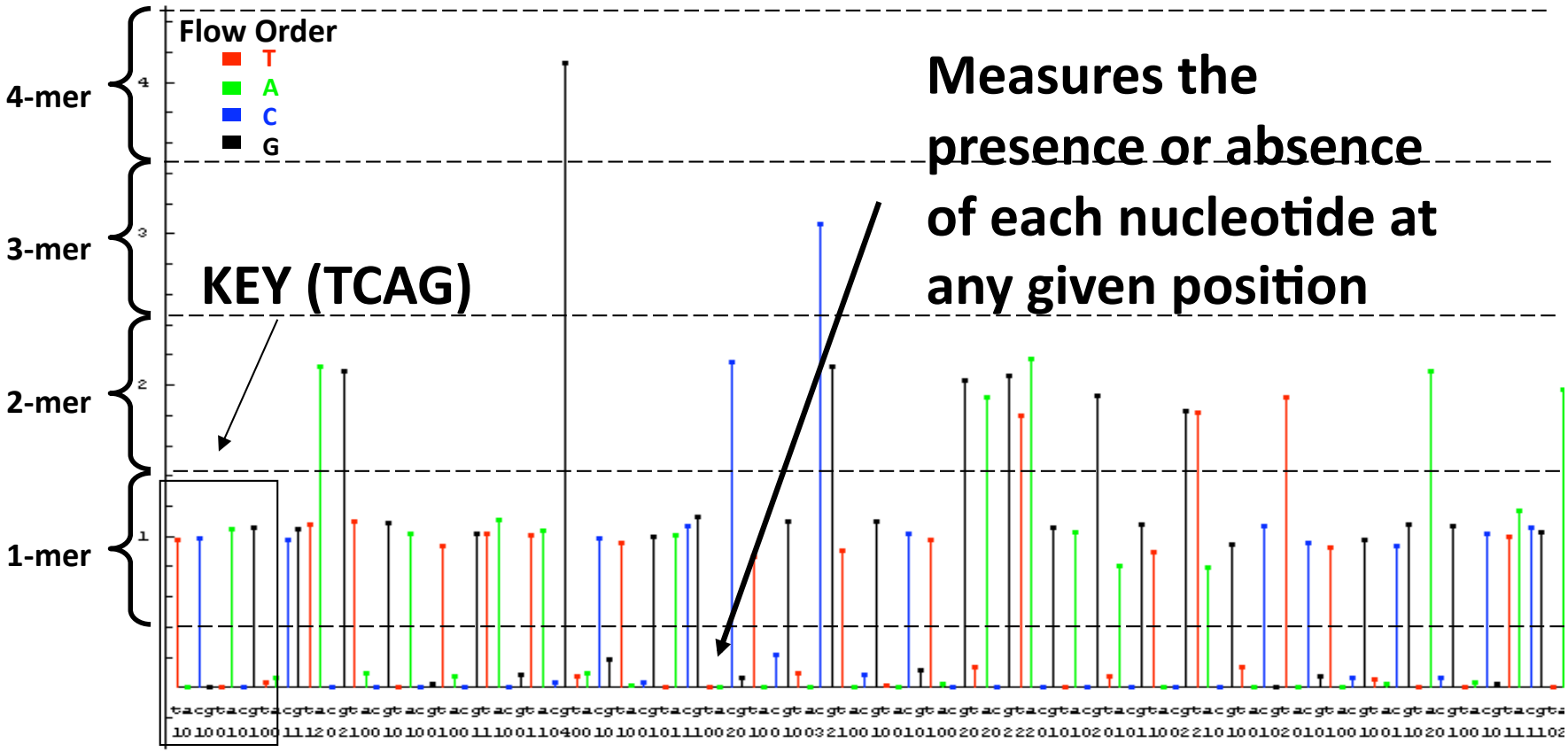
- Company: Roche Diagnostic®
- Website: https://roche-applied-science.com/sis/sequencing/flx/index/jsp
- Fragment length: 400 bases
- Number of reads per run: 1 10^6
- Total output per run: 0.4 Gb per run
- Time per run: 8 hours

GS Junior

- Company: Roche Diagnostic®
- Website: http://www.gsjunior.com/
- Fragment length: 400 bases
- Number of reads per run: 0.1 10^6
- Total output per run: 0.04 Gb per run
- Time per run: 12 hours

454 Sequencing: BaseCalling

- Count the photons generated for each “flow”
- Base call using signal thresholds
- Delivery of one nucleotide per flow ensures accurate base calling



Genetic Analyzer™/Solexa™ IIx

- Company: Illumina®
- Website: http://www.illumina.com/systems/genome_analyzer_iix.ilmn
- Fragment length: 35-105 bases
  (possibility of "paired-end" sequencing)
- Number of reads per run: 350 10^6
- Total output per run: 33 Gb
- Time per run: 6 days

HiSeq 2000

- Company: Illumina®
- Website: http://www.illumina.com/systems/hiseq_2000.ilmn
- Fragment length: 100 bases (2x100 paired-ends)
- Number of reads per run: 2 10^9
- Total output per run: 200 Gb
- Time per run: 9 days

An idea of the HiSeq 2000 production per run

- 2 billions of reads of 100 bp
- 6 lines per read
- 55 lines per page (time 11)
- 218 181 818 pages (64 800 km long, or 26 km high)
- more than 1000 tonnes of paper

Biodiversity and DNA barcoding

- The DNA barcoding concept
- Our scientific objectives
- New sequencing technologies
- Bioinformatic aspects
  - The challenges
  - Available programs for preparing the experiments
  - Available programs for analyzing the output of next generation sequencers
- High throughput plant identification
- High throughput animal identification
- Conclusion

Four barcode challenges

- Evaluate the quality of a barcode
- Infer new barcode region and their corresponding primers
- Prepare samples: the multiplex sample tagging system
- Analyze sequence data…
  - How to deal with large sequence files?
  - How to assign sequences to a taxon, with or without reference sequences?
  - How to deal with amplification/sequencing errors?

Barcode quality indices

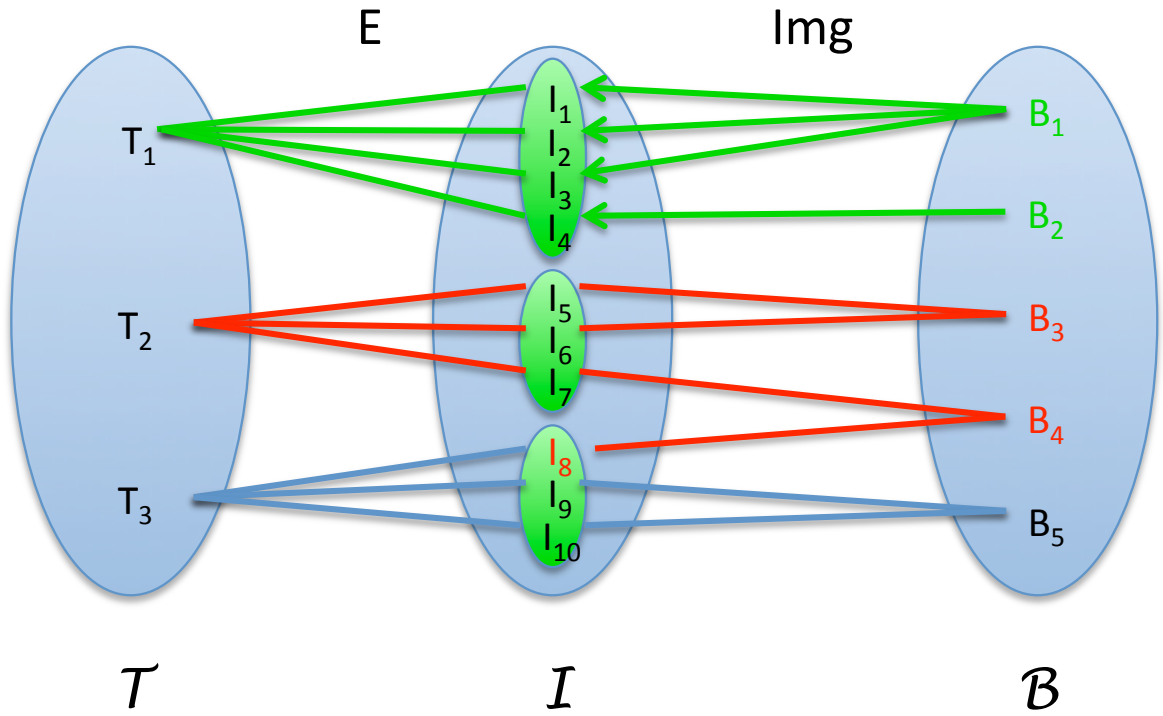
- Barcode coverage

$$Bc = \frac{amplified\ \ taxa}{total\ \ number\ \ of\ \ taxa}$$

- Barcode specificity

$$Bs = \frac{unambiguously\ \ identified\ \ taxa}{total\ \ number\ \ of\ \ taxa}$$

An unambiguously identified taxon



Testing a barcode region

EcoPCR : http://www.grenoble.prabi.fr/trac/ecoPCR

```
ecoPCR -d ebpln96 -l5 -L300 -k -e3 GGGCAATCCTGAGCCAA CCATTGAGTCTCTGCACCTATC
```

```
#
# ecoPCR version 0.1
# direct  strand oligo1 : GGGCAATCCTGAGCCAA              ; oligo2c :          GATAGGTGCAGAGACTCAATGG
# reverse strand oligo2 : CCATTGAGTCTCTGCACCTATC         ; oligo1c :               TTGGCTCAGGATTGCCC
# max error count by oligonucleotide : 3
# database : arctic_01_02_2008
# amplifiat length between [5,300] bp
# output in kingdom mode
#
0240g           |       495 | 10000726 | subspecies          |   282718 | Achillea alpina               |
1043o           |       496 | 10000724 | subspecies          |   282718 | Achillea alpina               |
0239g           |       495 | 10000001 | subspecies          |    13329 | Achillea millefolium          |
1042o           |       496 | 10000725 | subspecies          |    13329 | Achillea millefolium          |
0722g           |       567 | 10000639 | subspecies          | 10000110 | Aconitum delphinifolium       |
1474o           |       567 | 10000639 | subspecies          | 10000110 | Aconitum delphinifolium       |
0818g           |       567 | 10000002 | subspecies          |   112589 | Aconitum lycoctonum           |
1649o           |       567 | 10000002 | subspecies          |   112589 | Aconitum lycoctonum           |
0282g           |       464 | 10000111 | species             | 10000111 | Aconogonon alaskanum          |
```

One amplified sequence description

SeqId	:	0240g
Seq length	:	495
TaxId	:	10000726
Taxon rank	:	subspecies
Sp taxid	:	282718
Sp name	:	Achillea alpina
Genus taxid	:	13328
Genus name	:	Achillea
Family taxId	:	4210
Family name	:	Asteraceae
Kgdom taxid	:	33090
Kgdom name	:	Viridiplantae
Match strand	:	D
Direct match	:	GGGCAATCCTGAGCCAA
Direct error	:	0
Rev match	:	CCATCGAGTCTCTGCACCTATC
Rev error	:	1
Amp length	:	90
Amp sequence	:	ATCACGTTTTCCGAAAACAAACAAAGGTTCAGAAAGCGAAAAGAAAAAAAA

Coverage assessment

```
>ecoTaxStat.py –d mitovert P09.3.ecopcr
```

rank	ecopcr	db	percent
class	5	5	100.00
family	347	365	95.07
genus	630	663	95.02
infraclass	2	2	100.00
infraorder	9	9	100.00
kingdom	1	1	100.00
order	102	105	97.14
parvorder	4	4	100.00
phylum	1	1	100.00
species	777	814	95.45
species group	1	1	100.00
subclass	3	3	100.00
subfamily	120	127	94.49
subgenus	7	7	100.00
suborder	75	76	98.68
subphylum	1	1	100.00
subspecies	43	45	95.56
superclass	1	1	100.00
superfamily	38	41	92.68
superkingdom	1	1	100.00
superorder	18	18	100.00
tribe	22	23	95.65

Specificity assessment

```
>ecoTaxSpecificity.py –d mitovert P09.3.ecopcr
rank                        taxon_ok      taxon_total     percent
order                            102              102      100.00
infraclass                         2                2      100.00
superfamily                       38               38      100.00
parvorder                          4                4      100.00
species group                      1                1      100.00
superkingdom                       1                1      100.00
kingdom                            1                1      100.00
phylum                             1                1      100.00
infraorder                         9                9      100.00
subfamily                        118              120       98.33
class                              5                5      100.00
species                          685              777       88.16
superorder                        18               18      100.00
suborder                          75               75      100.00
subclass                           3                3      100.00
subgenus                           7                7      100.00
genus                            600              630       95.24
subspecies                        17               43       39.53
superclass                         1                1      100.00
family                           347              347      100.00
tribe                             21               22       95.45
subphylum                          1                1      100.00
```

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]

(2010) *BMC Genomics*, **11**, 434.

Abstract

Background: DNA barcoding is a key tool for assessing biodiversity in both taxonomic and environmental studies. Essential features of barcodes include their applicability to a wide spectrum of taxa and their ability to identify even closely related species. Several DNA regions have been proposed as barcodes and the region selected strongly influences the output of a study. However, formal comparisons between barcodes remained limited until now. Here we present a standard method for evaluating barcode quality, based on the use of a new bioinformatic tool that performs *in silico* PCR over large databases. We illustrate this approach by comparing the taxonomic coverage and the resolution of several DNA regions already proposed for the barcoding of vertebrates. To assess the relationship between *in silico* and *in vitro* PCR, we also developed specific primers amplifying different species of Felidae, and we tested them using both kinds of PCR

Results: Tests on specific primers confirmed the correspondence between *in silico* and *in vitro* PCR. Nevertheless, results of *in silico* and *in vitro* PCRs can be somehow different, also because tuning PCR conditions can increase the performance of primers with limited taxonomic coverage. The *in silico* evaluation of DNA barcodes showed a strong variation of taxonomic coverage (i.e., universality): barcodes based on highly degenerated primers and those corresponding to the conserved region of the *Cyt*-b showed the highest coverage. As expected, longer barcodes had a better resolution than shorter ones, which are however more convenient for ecological studies analysing environmental samples.

Conclusions: *In silico* PCR could be used to improve the performance of a study, by allowing the preliminary comparison of several DNA regions in order to identify the most appropriate barcode depending on the study aims.

RESEARCH ARTICLE Open Access

ITS as an environmental DNA barcode for fungi: an *in silico* approach reveals potential PCR biases

Eva Bellemain*[1], Tor Carlsen[2], Christian Brochmann[1], Eric Coissac[3], Pierre Taberlet[3] and Håvard Kauserud[2]

(2010) *BMC Microbiology*, **10**, 189.

Abstract

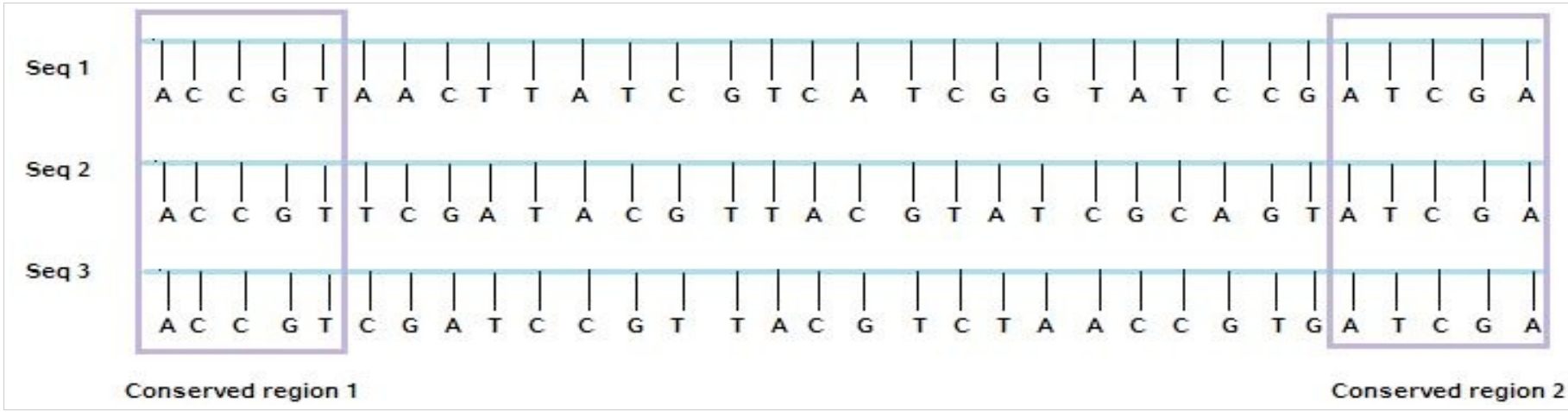
Background: During the last 15 years the internal transcribed spacer (ITS) of nuclear DNA has been used as a target for analyzing fungal diversity in environmental samples, and has recently been selected as the standard marker for fungal DNA barcoding. In this study we explored the potential amplification biases that various commonly utilized ITS primers might introduce during amplification of different parts of the ITS region in samples containing mixed templates ('environmental barcoding'). We performed *in silico* PCR analyses with commonly used primer combinations using various ITS datasets obtained from public databases as templates.

Results: Some of the ITS primers, such as ITS1-F, were hampered with a high proportion of mismatches relative to the target sequences, and most of them appeared to introduce taxonomic biases during PCR. Some primers, e.g. ITS1-F, ITS1 and ITS5, were biased towards amplification of basidiomycetes, whereas others, e.g. ITS2, ITS3 and ITS4, were biased towards ascomycetes. The assumed basidiomycete-specific primer ITS4-B only amplified a minor proportion of basidiomycete ITS sequences, even under relaxed PCR conditions. Due to systematic length differences in the ITS2 region as well as the entire ITS, we found that ascomycetes will more easily amplify than basidiomycetes using these regions as targets. This bias can be avoided by using primers amplifying ITS1 only, but this would imply preferential amplification of 'non-dikarya' fungi.

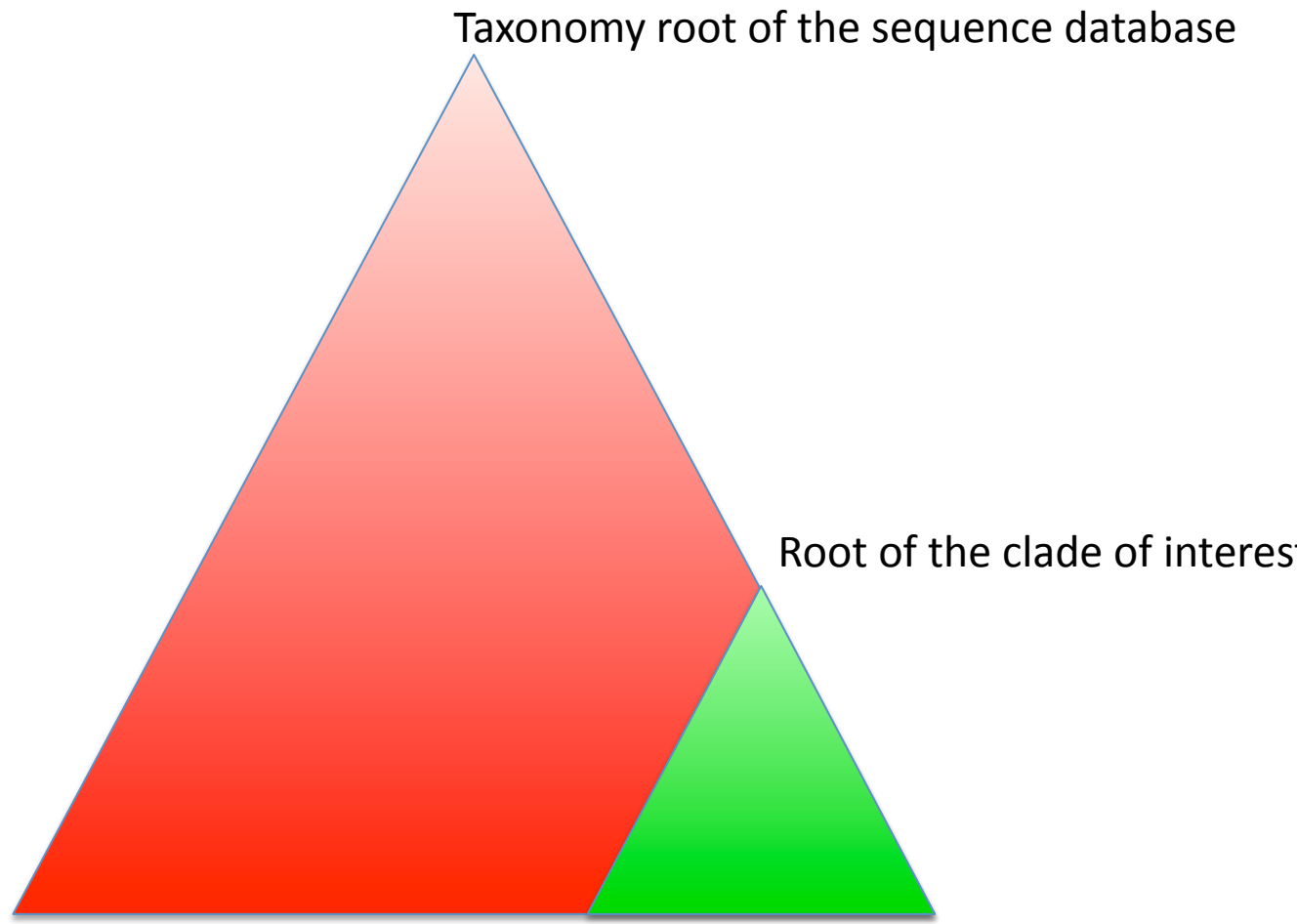
Conclusions: We conclude that ITS primers have to be selected carefully, especially when used for high-throughput sequencing of environmental samples. We suggest that different primer combinations or different parts of the ITS region should be analyzed in parallel, or that alternative ITS primers should be searched for.

Infer new barcode region

ecoPrimer : http://www.grenoble.prabi.fr/trac/ecoPrimer



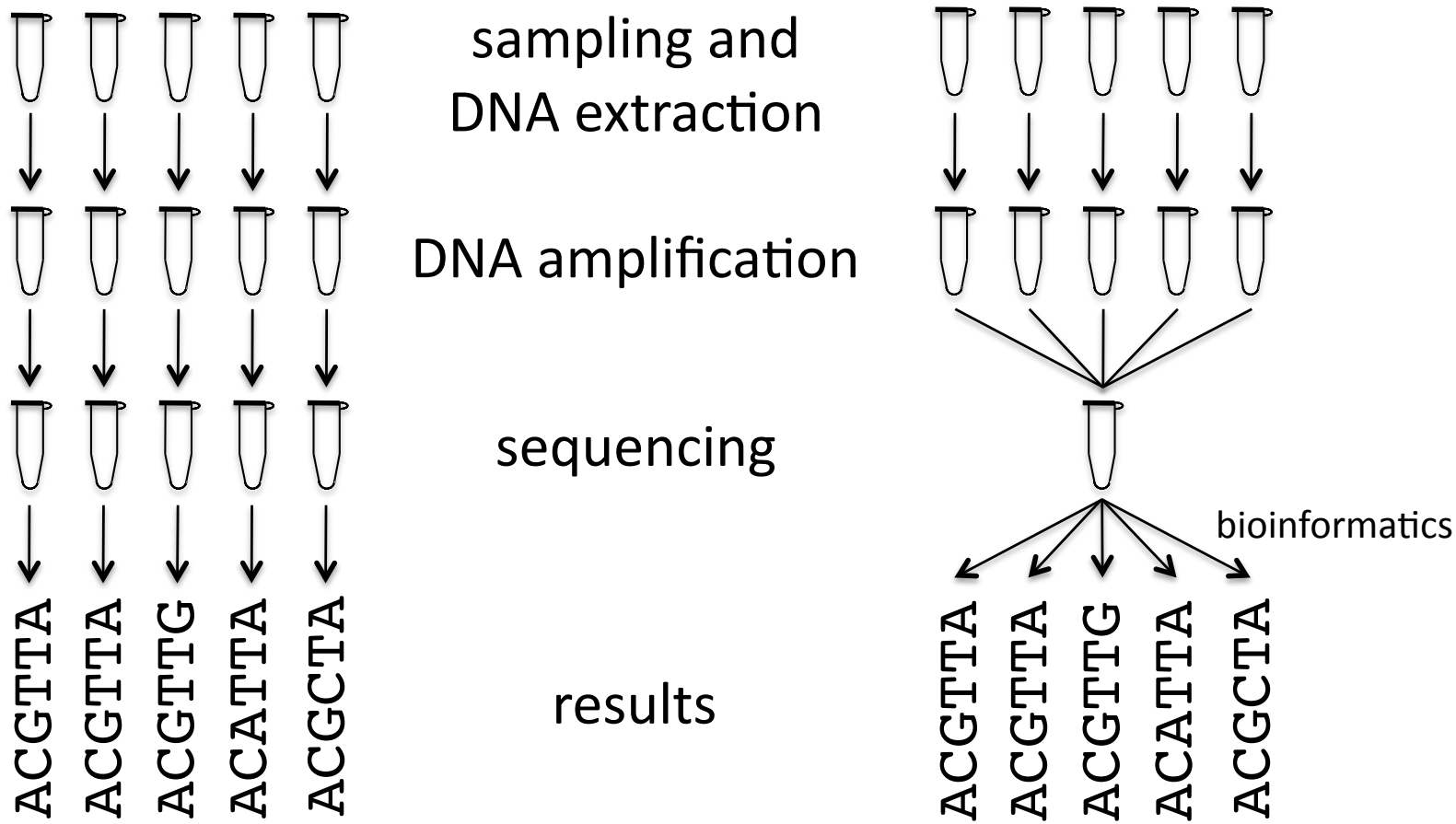
Restrict primers to a clade



Traditional *versus* next generation sequencing

traditional sequencing

next generation sequencing



Multiplexing samples

Add small tags on 5' end of barcode primer for identifying samples

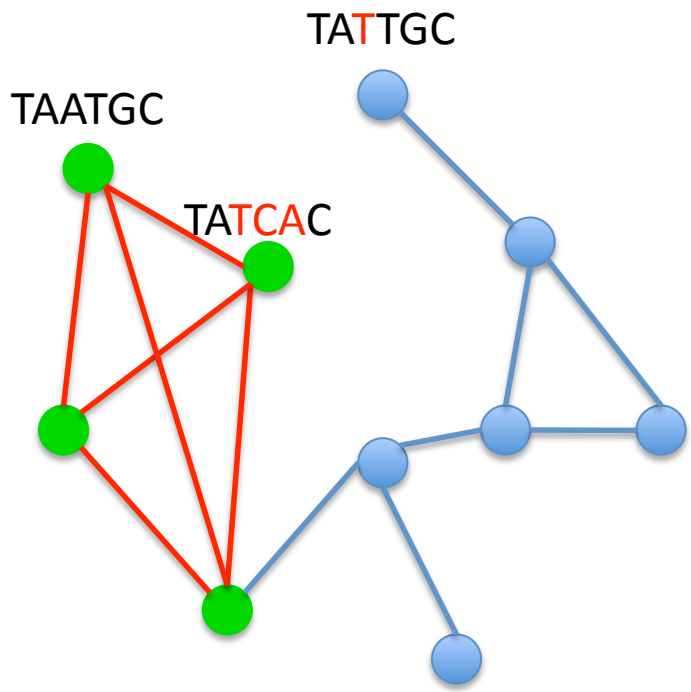
No error tolerant code

One error tolerant code

One error autocorrective code

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

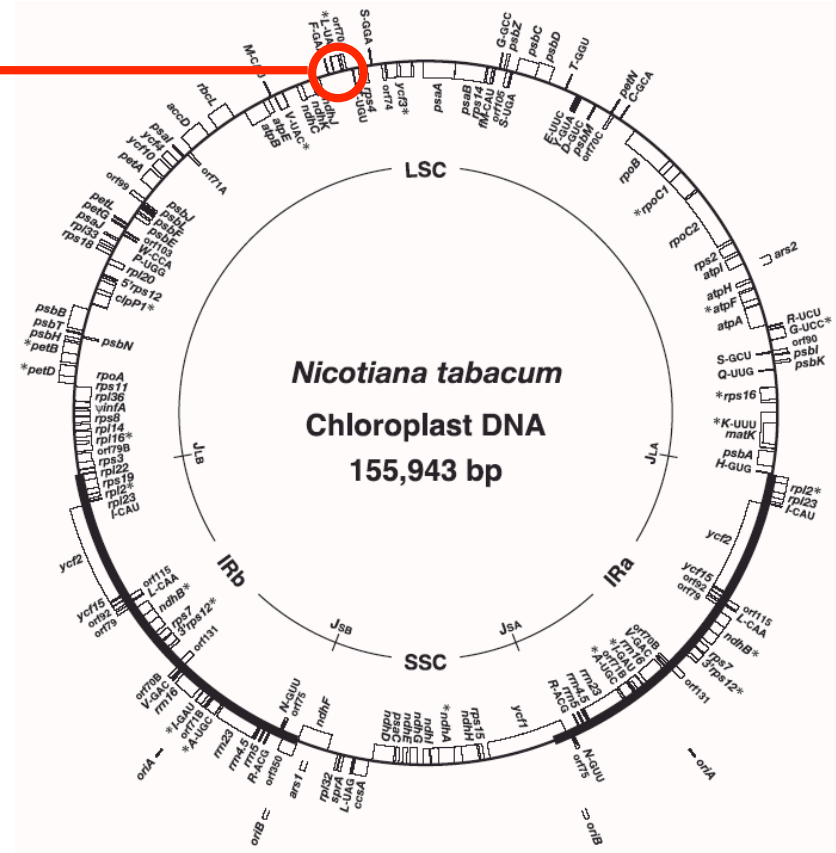
Analyze high throughput sequence data

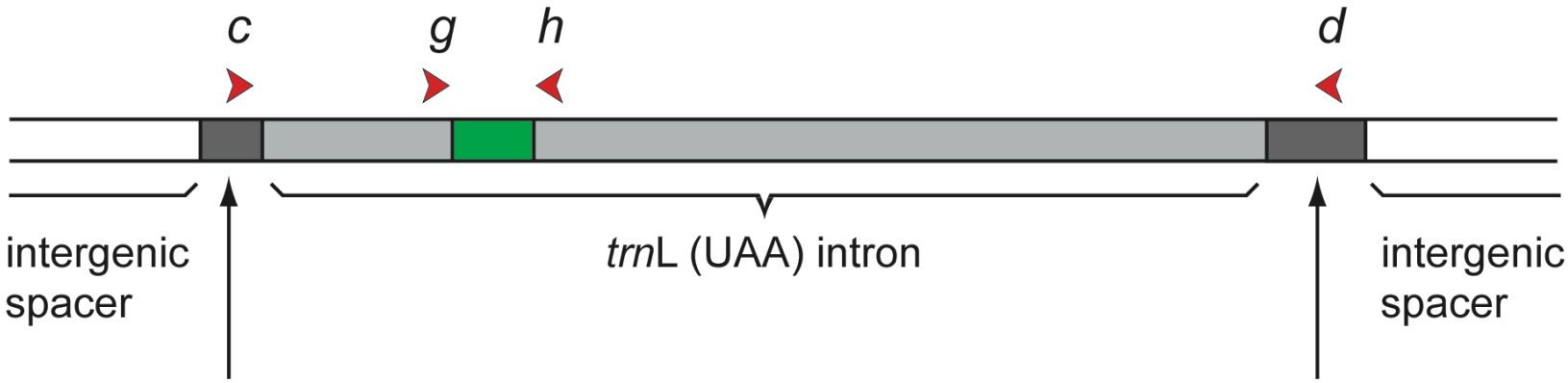
- Manipulating large amount of data
  - OBITools: www.grenoble.prabi.fr/trac/OBITools
- Identifying species
  - ecoTag (part of OBITools)
- Dealing with errors
  - We just start to working on

Biodiversity and DNA barcoding

- The DNA barcoding concept
- Our scientific objectives
- New sequencing technologies
- Bioinformatic aspects
  - The challenges
  - Available programs for preparing the experiments
  - Available programs for analyzing the output of next generation sequencers
- High throughput plant identification
- High throughput animal identification
- Conclusion

The chloroplast *trn*L(UAA) intron

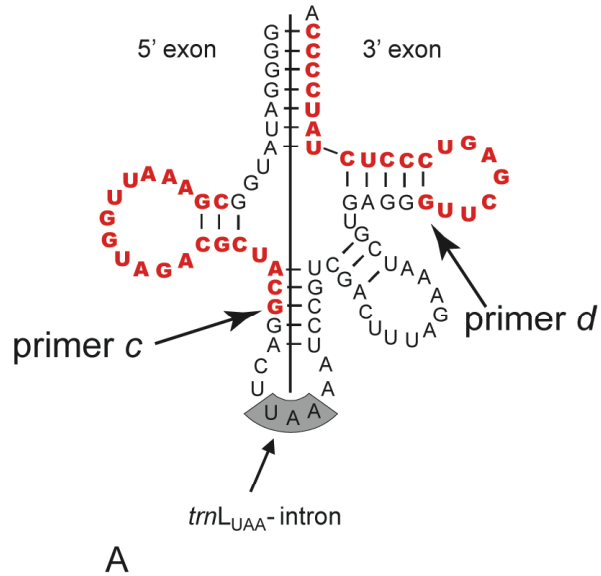


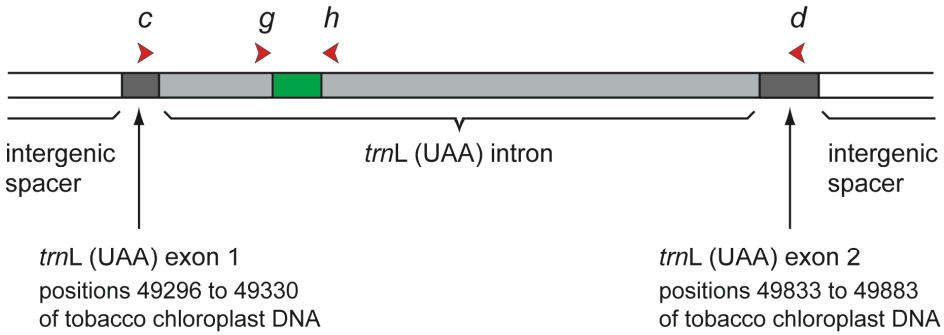


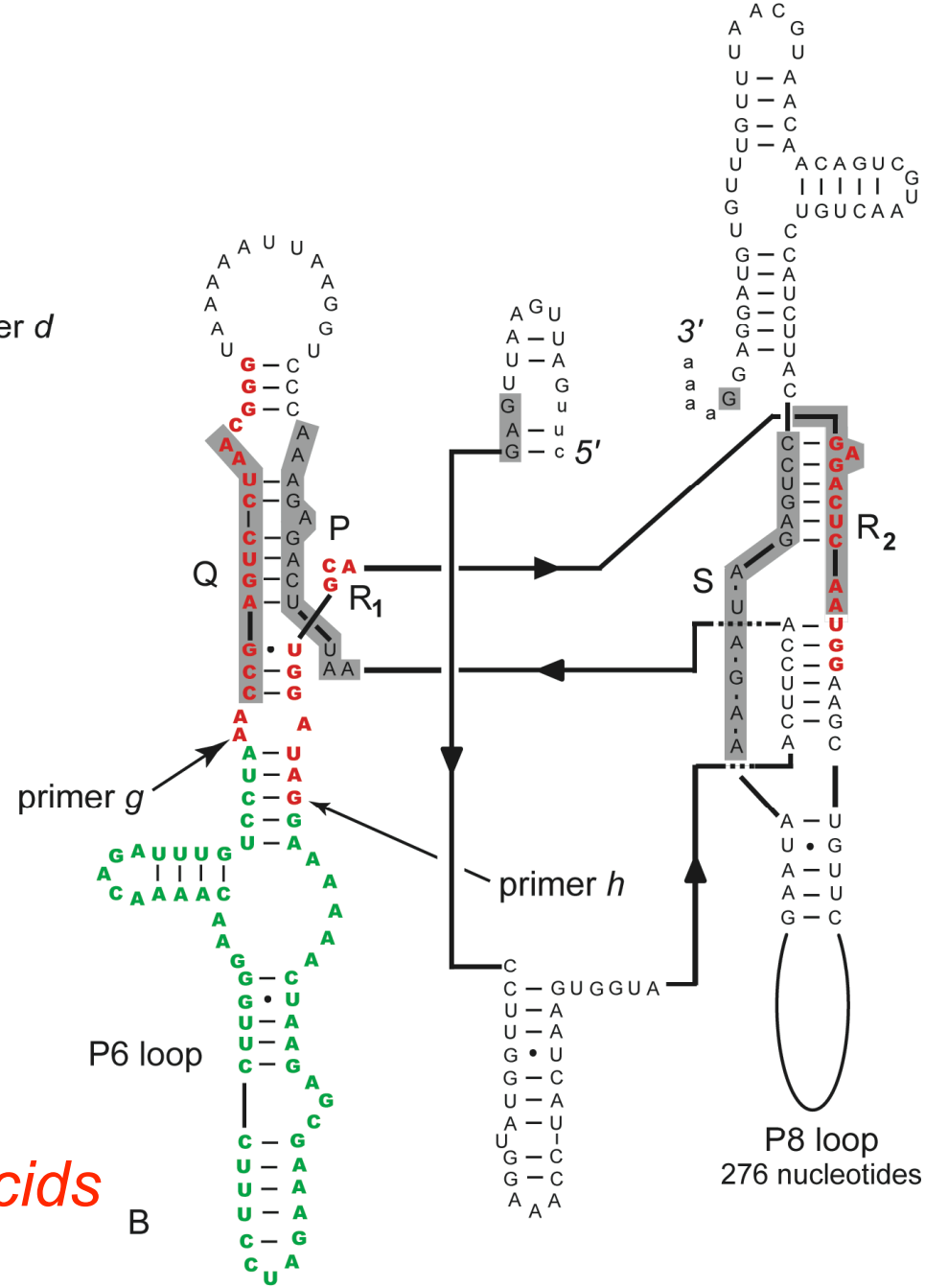
*trn*L (UAA) exon 1
positions 49296 to 49330
of tobacco chloroplast DNA

*trn*L (UAA) exon 2
positions 49833 to 49883
of tobacco chloroplast DNA

The chloroplast *trn*L(UAA) intron







Taberlet *et al.* (2007) *Nucleic Acids Research*, **35**, e14.


Theobroma_cacao*
Beta_vulgaris*
Castanea_sativa
Cannabis_sativa
Cicer_arietinum
Saccharum_officinarum
Asparagus_officinalis
Triticum_aestivum
Secale_cereale
Oryza_sativa
Panicum_miliaceum
Ribes_aureum
Fragaria_vesca*
Triphasia_trifolia
Vitis_vinifera*
Prunus_persica
Prunus_armeriana*
Prunus_cerasus*
Actinidia_chinensis
Zea_mais
Pisum_sativum*
Phaseolus_vulgaris
Sorghum_halepense
Cynara_cardunculus
Arctium_lappa
Lactuca_sativa
Helianthus_annuus
Ficus_carica*
Humulus_lupulus
Avena_sativa
Nasturtium_officinale
Armoracia_rusticana
Hordeum_vulgare
Anthriscus_cerefolium
Allium_cepa*
Allium_porum*
Carum_petroselinum*
Solanum_lycopersicum
Solanum_melongena
Solanum_tuberosum*
Raphanus_sativus
Brassica_oleracea_capita
Brassica_rapa_rapa
Brassica_nigra
Sinapis_alba
Olea_europaea*
Urtica_dioica
Rumex_acetosa
1 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120

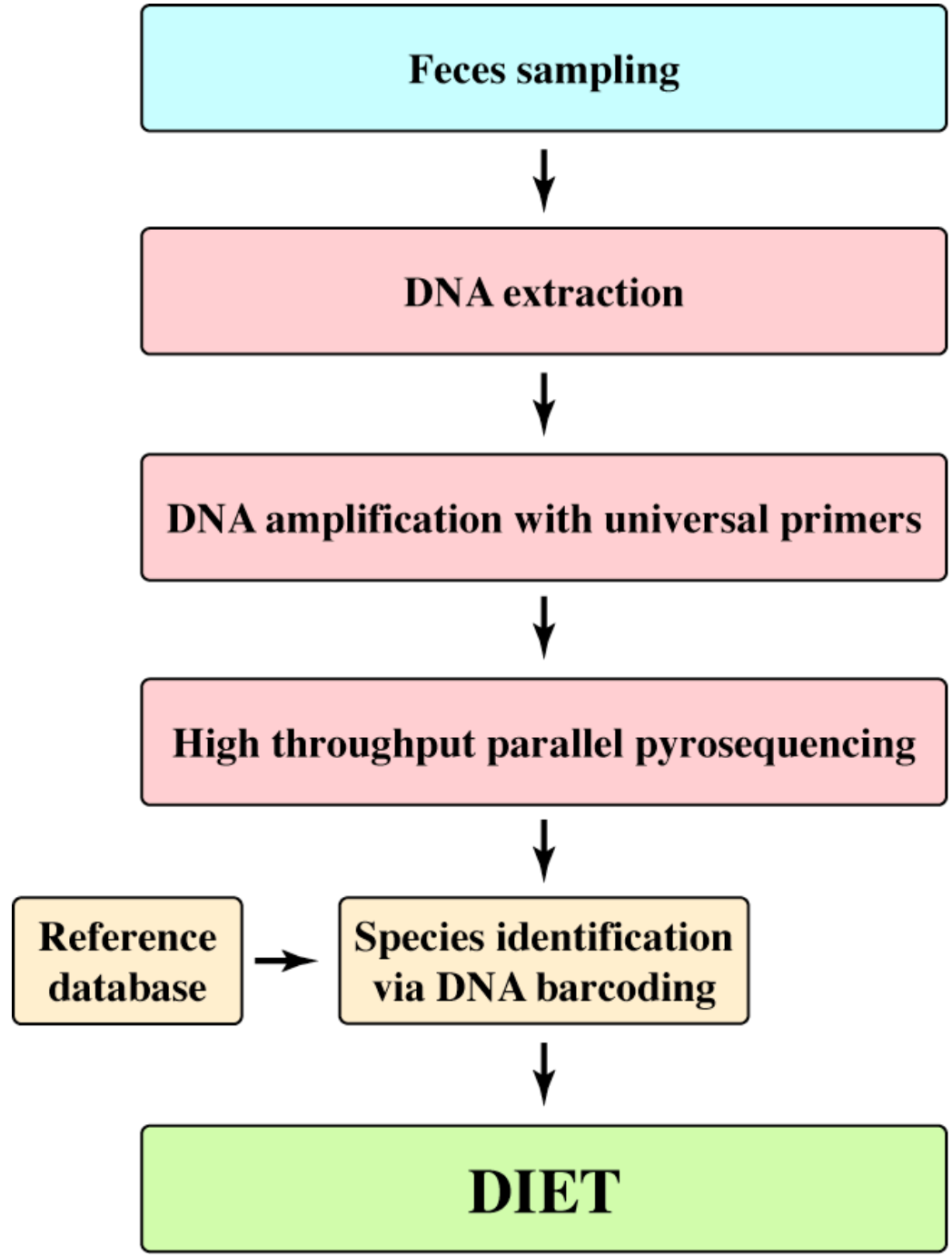
Is it possible to find better than the chloroplast *trn*L P6 loop primers with ecoPrimer ?

- Dataset of 133 full chloroplast genomes
  - with 110 Streptophyta species
  - and 23 non Streptophyta species

G : GGGCAATCCTGAGCCAAatc

H : CCATTGAGTCTCTGCACCTATCc

DNA-based diet analysis



TECHNICAL ADVANCES

New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trn*L approach

ALICE VALENTINI,*† CHRISTIAN MIQUEL,* MUHAMMAD ALI NAWAZ,‡§ EVA BELLEMAIN,* ERIC COISSAC,* FRANÇOIS POMPANON,* LUDOVIC GIELLY,* CORINNE CRUAUD,¶ GIUSEPPE NASCETTI,† PATRICK WINCKER,¶ JON E. SWENSON‡** and PIERRE TABERLET*

**Laboratoire d'Ecologie Alpine, CNRS UMR 5553, Université Joseph Fourier, BP 53, F-38041 Grenoble cedex 9, France, †Dipartimento di Ecologia e Sviluppo Economico Sostenibile, Università degli Studi della Tuscia, via S. Giovanni Decollato 1, I-01100 Viterbo, Italy, ‡Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, Post Box 5003, NO-1432 Ås, Norway, §Himalayan Wildlife Foundation, 01, Park Road, Sector F-8/1 Islamabad 44000, Pakistan, ¶Genoscope — CNS, 2 rue Gaston Crémieux, BP 5706, F-91057 Evry cedex, France, **Norwegian Institute for Nature Research, NO-7485 Trondheim, Norway*

Abstract

The development of DNA barcoding (species identification using a standardized DNA sequence), and the availability of recent DNA sequencing techniques offer new possibilities in diet analysis. DNA fragments shorter than 100–150 bp remain in a much higher proportion in degraded DNA samples and can be recovered from faeces. As a consequence, by using universal primers that amplify a very short but informative DNA fragment, it is possible to reliably identify the plant taxon that has been eaten. According to our experience and using this identification system, about 50% of the taxa can be identified to species using the *trn*L approach, that is, using the P6 loop of the chloroplast *trn*L (UAA) intron. We demonstrated that this new method is fast, simple to implement, and very robust. It can be applied for diet analyses of a wide range of phytophagous species at large scales. We also demonstrated that our approach is efficient for mammals, birds, insects and molluscs. This method opens new perspectives in ecology, not only by allowing large-scale studies on diet, but also by enhancing studies on resource partitioning among competing species, and describing food webs in ecosystems.

Valentini *et al.* (2009) *Molecular Ecology Resources*, **9**, 51-60.



Mollusks

Deroceras reticulatum (1)
Arion rufus (1)
Helix aspera (1)


Insects

Gonfophocerippus rufus (2)

Chorthippus biguttulus (1)

Birds

Tetrao urogallus aquitanicus (4)

Tetrao urogallus major (2)


Mammals

Ursus arctos (12)

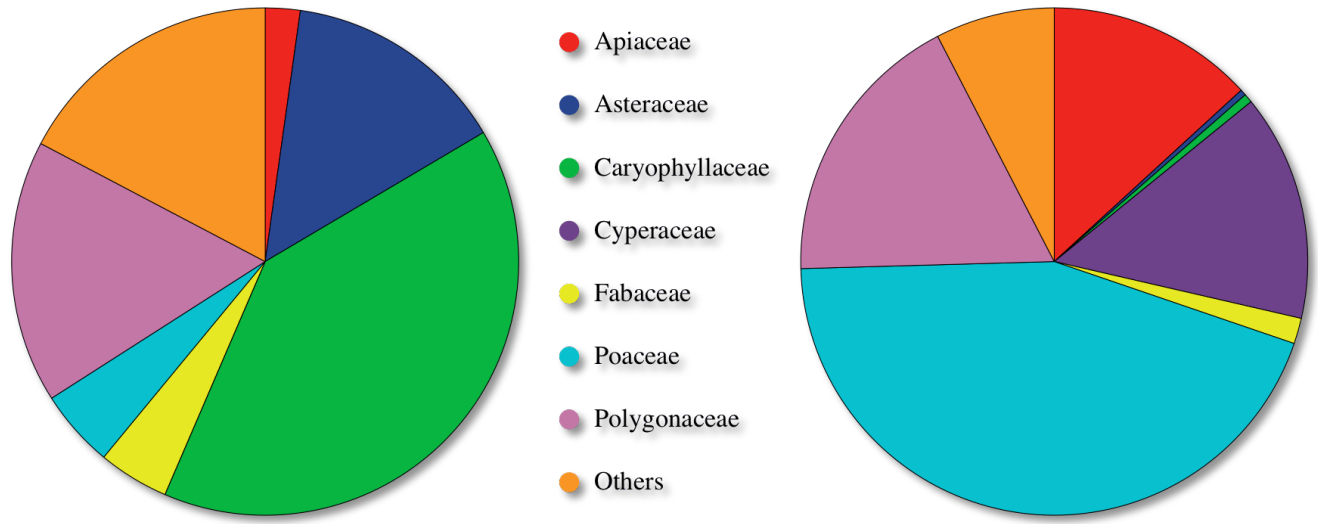
Marmota caudata (12)

Marmot and bear diets

Deosai National Park, Pakistan

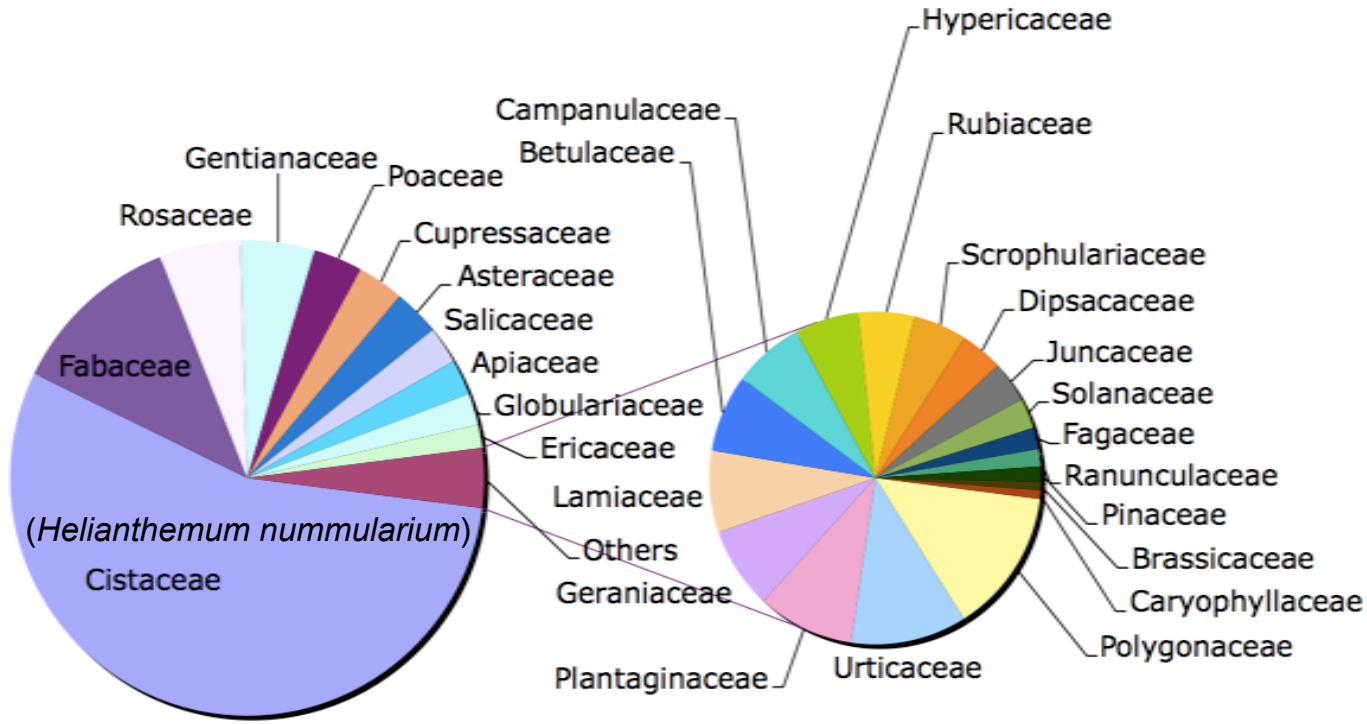
Golden marmot

Brown bear

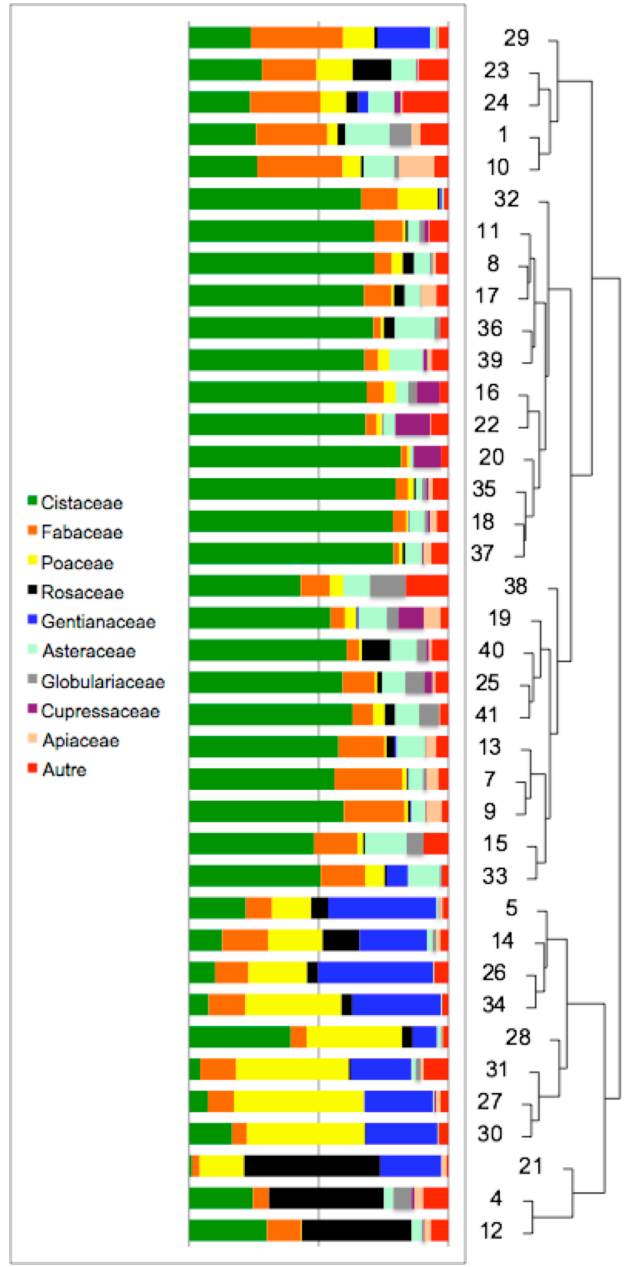


Bauges massif, France

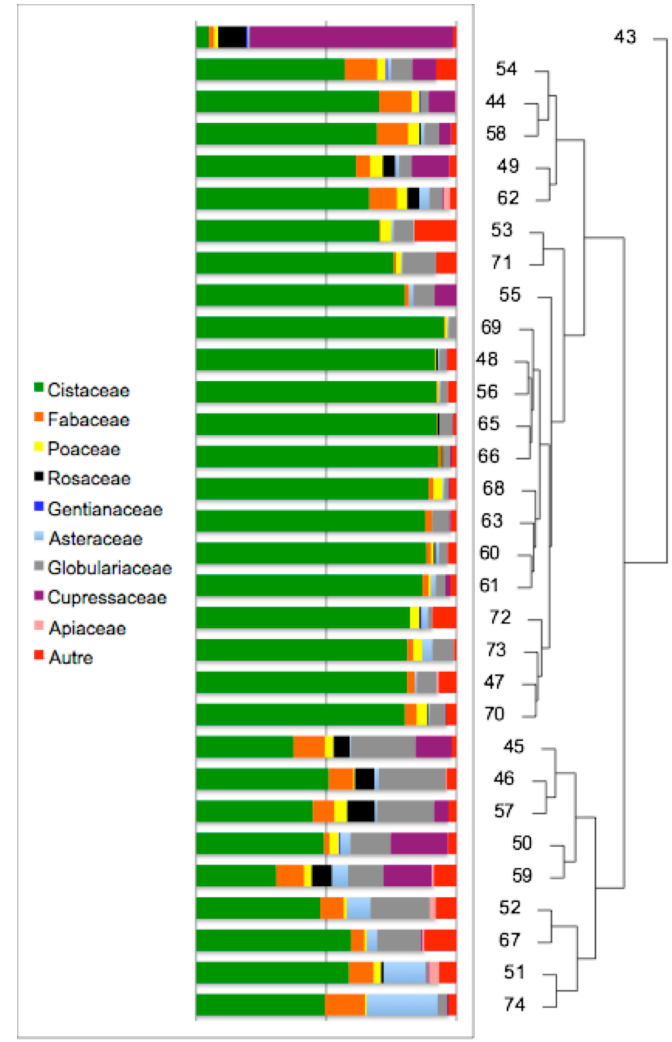
Chamois diet



Chamois diet



October

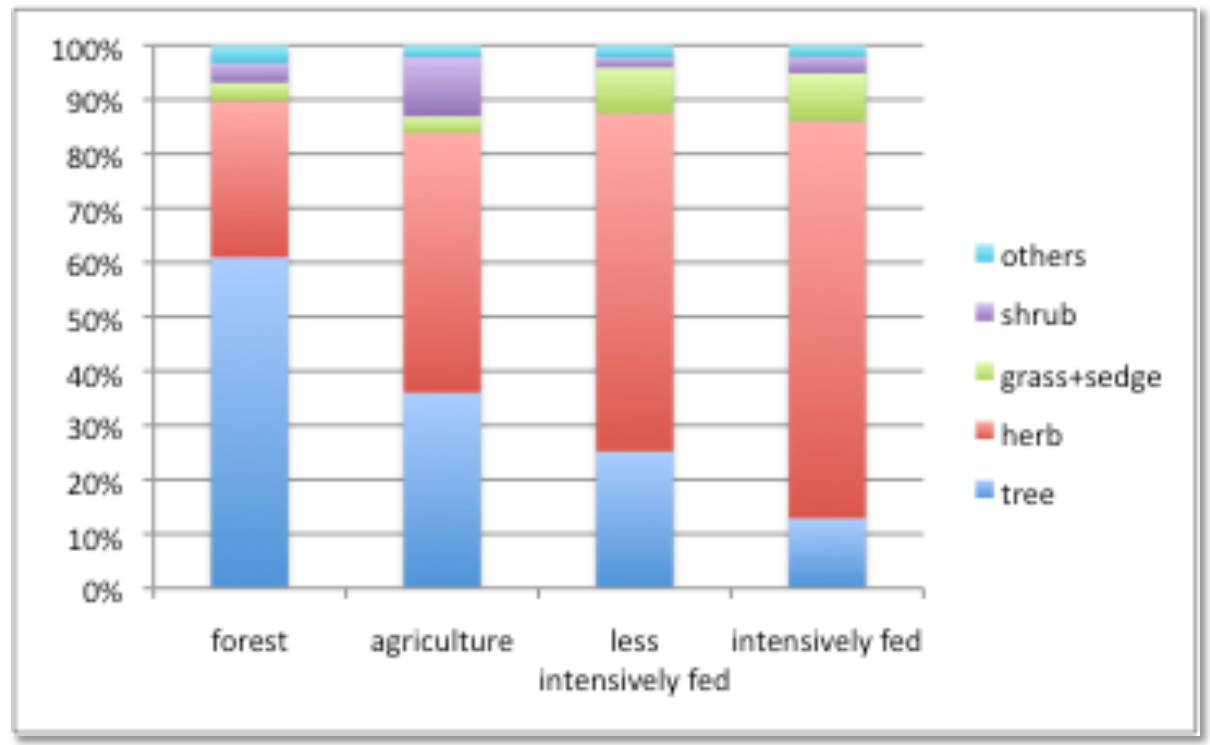


November

Bison diet

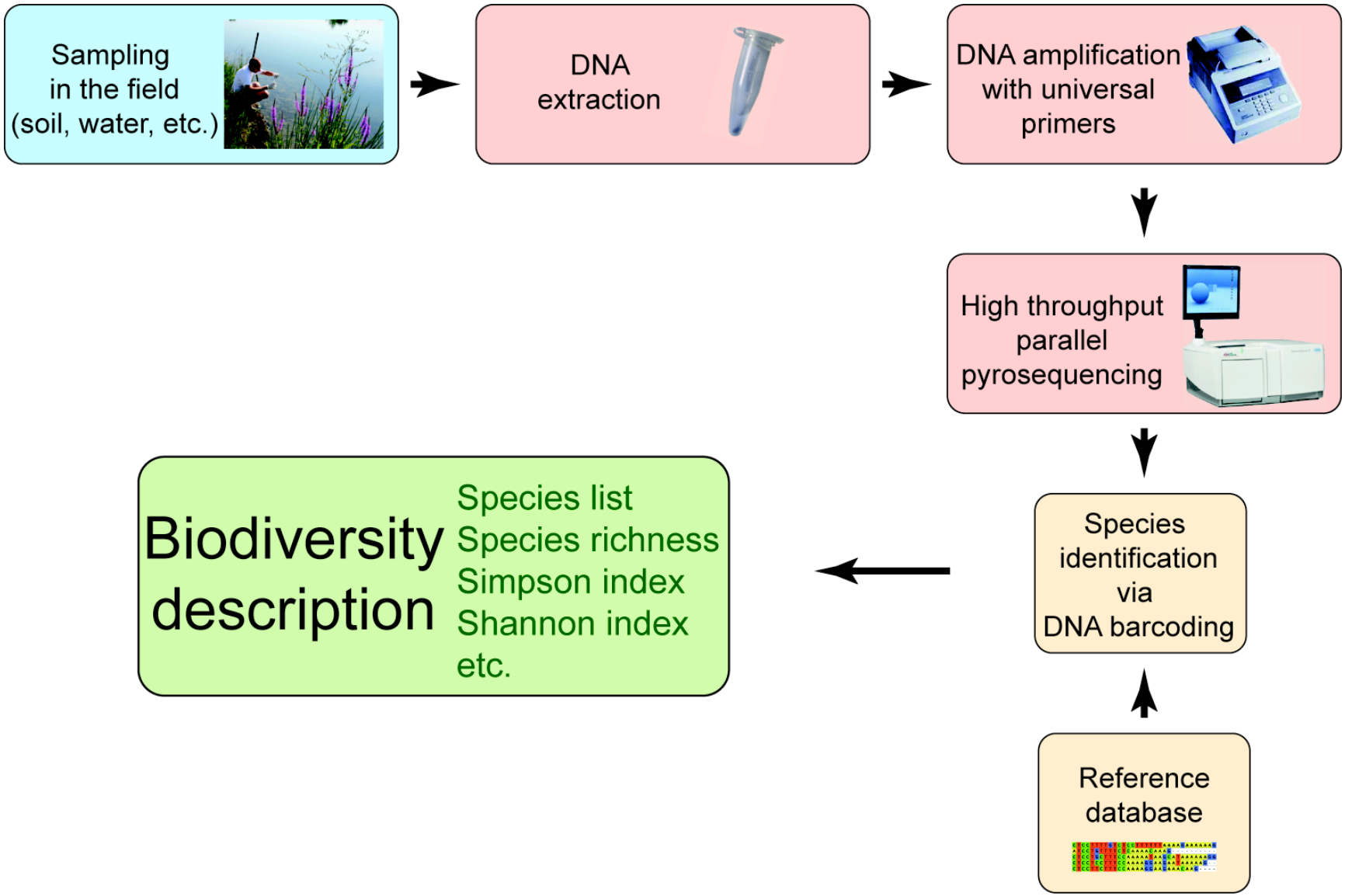
Bison diet in four different management schemes

(i) forest (unmanaged animals in Bialowieza reserve)
(ii) animals in agricultural areas
(iii) animals not intensively fed with hay
(iv) animals intensively fed with hay

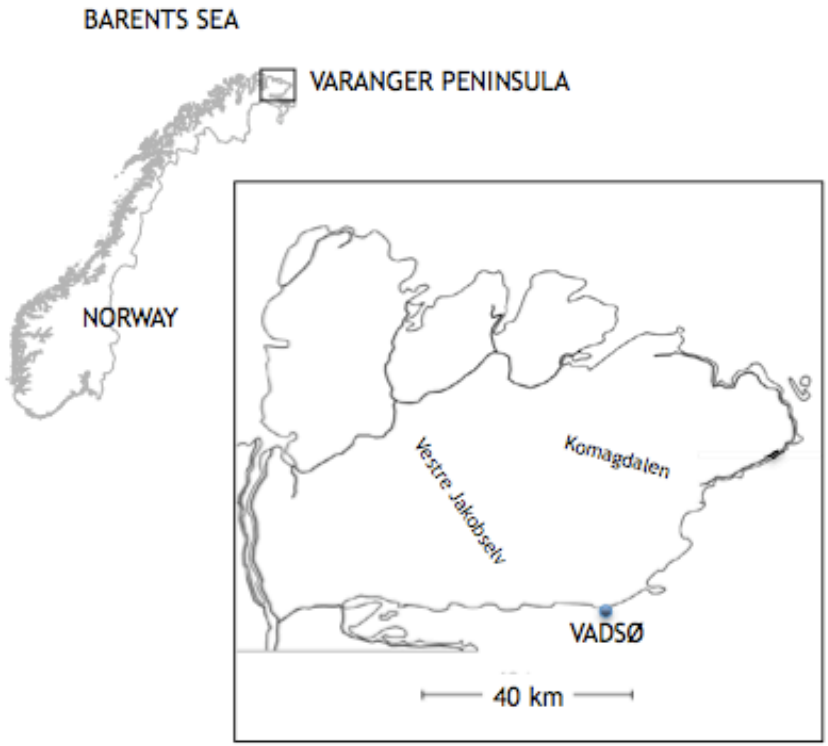


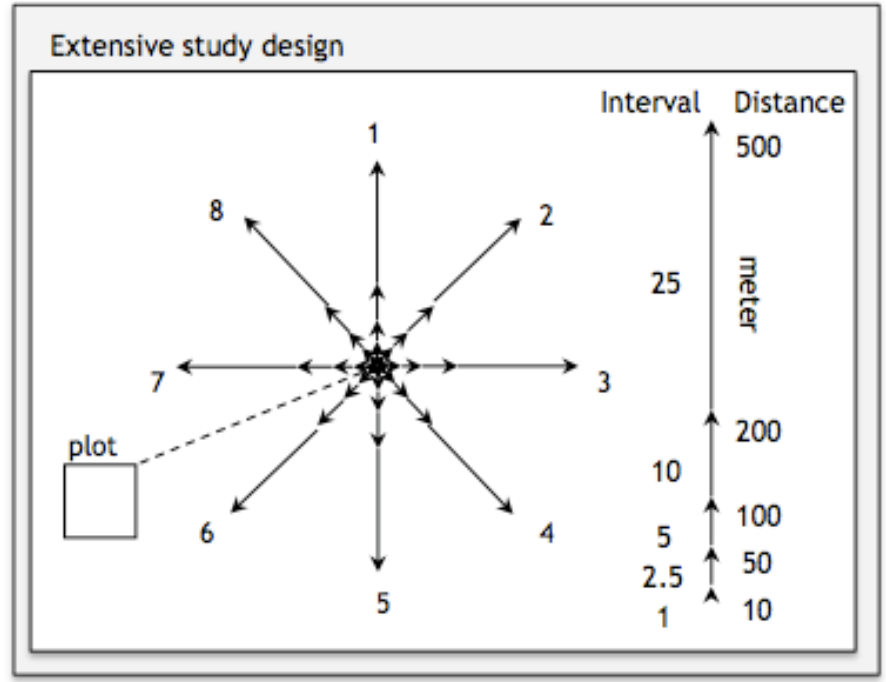
Collaboration with Jan Wojcik and Rafal Kowalczyk, Mammal Research Institute, Polish Academy of Sciences, Bialowieza, Poland

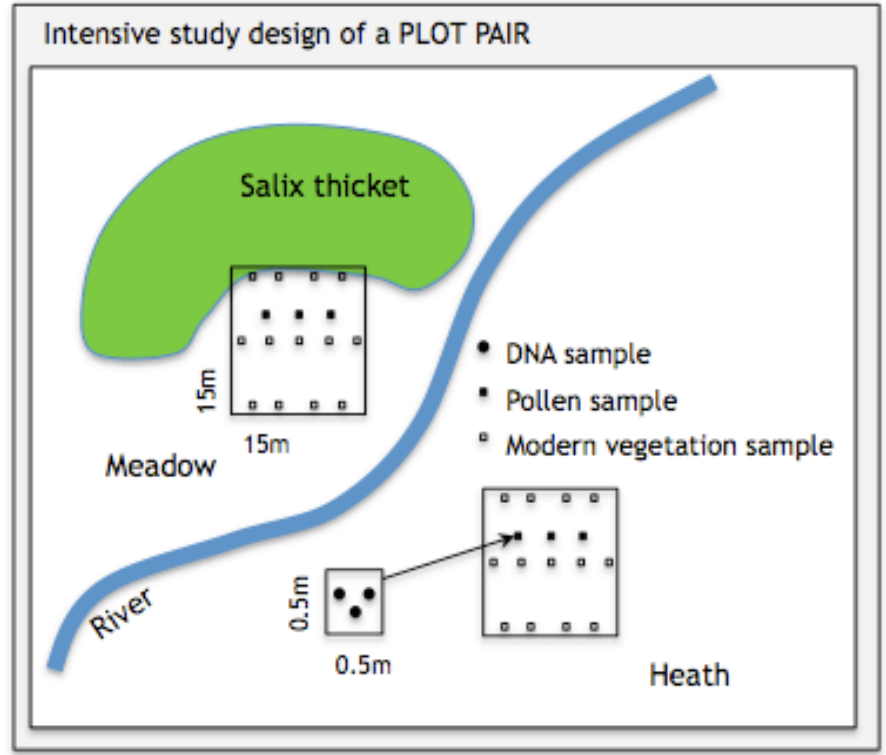
DNA-based biodiversity assessment



Calibration of soil DNA with aboveground species at Varanger, Northern Norway







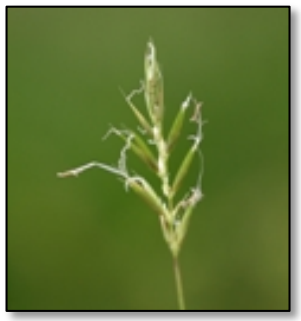
Above ground analysis

Avenella flexuosa

Poa sp.

Taraxacum sp.

Anthoxanthum nipponicum



Carex sp.

Festuca sp.

Deschampsia sp.

Rumex sp.

Calamagrostris sp.

DNA-based soil analysis

Bistorta vivipara

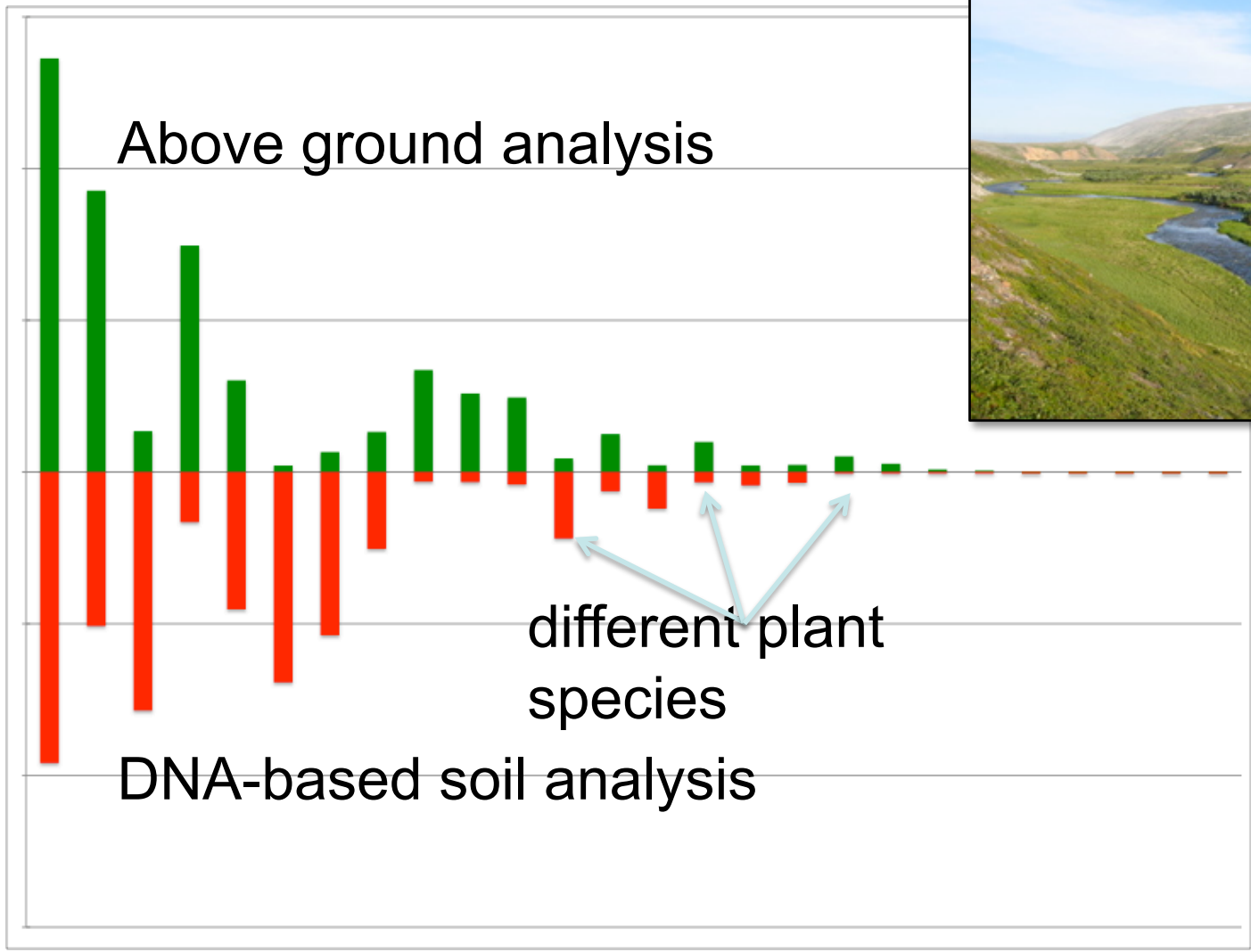
Salix sp.

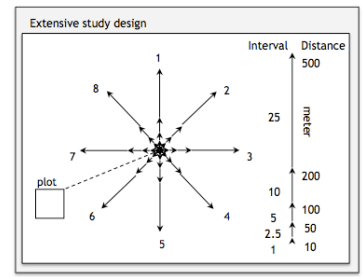
Alchemilla sp.

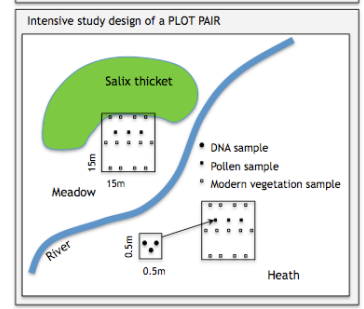
Viola biflora

Equisetum sp.

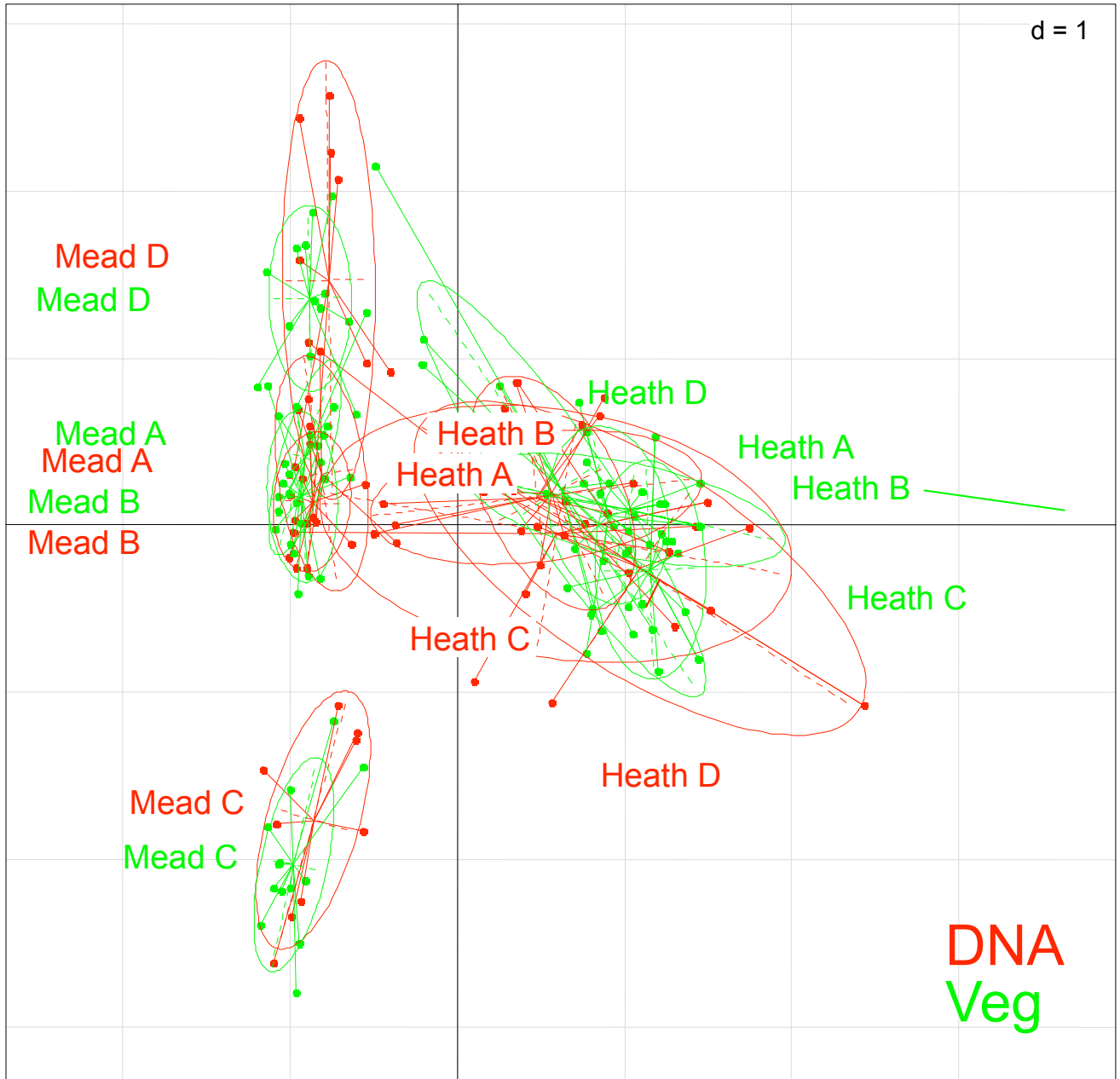
DNA in soil mirrors above ground plant diversity



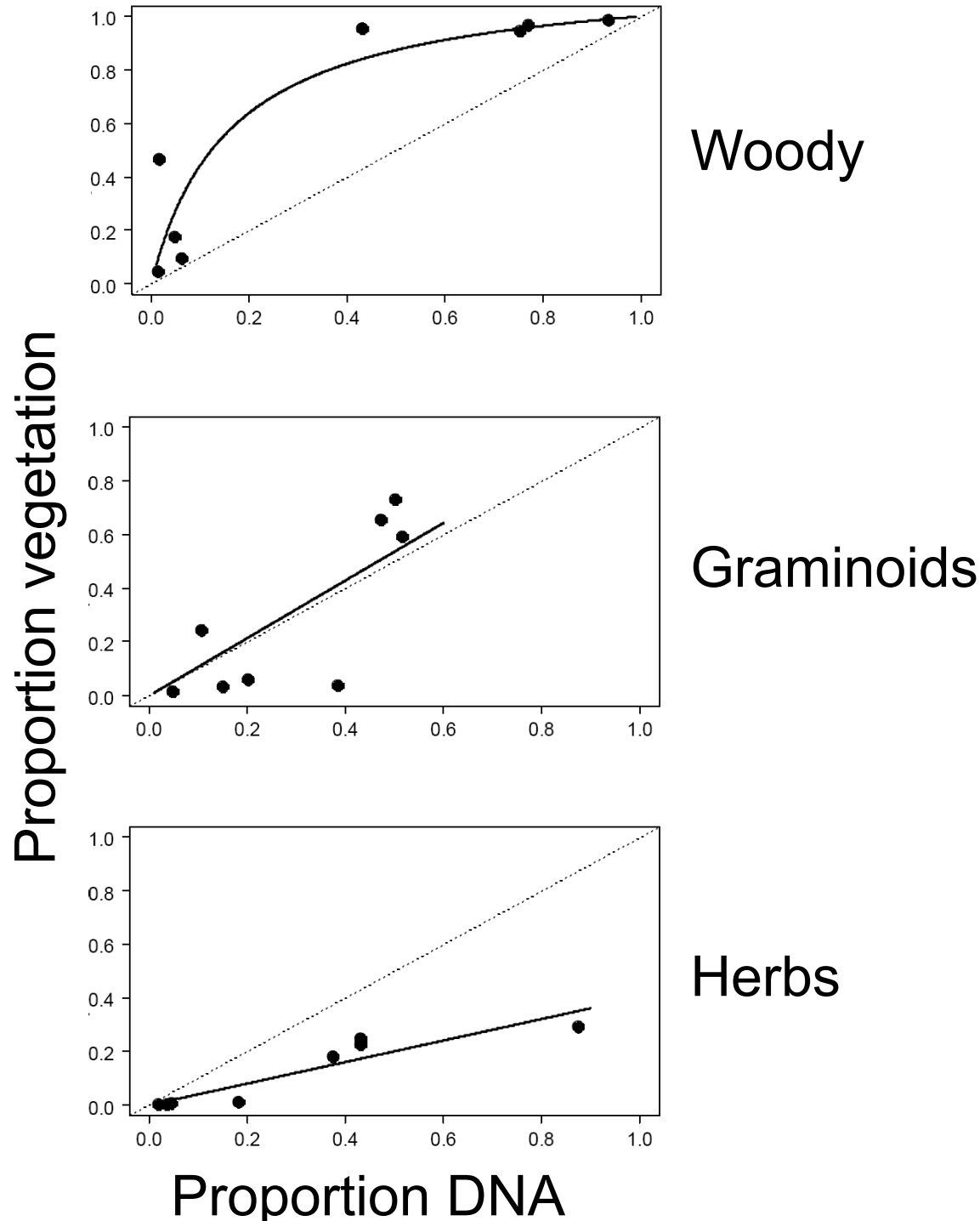




DNA in soil mirrors above ground plant diversity

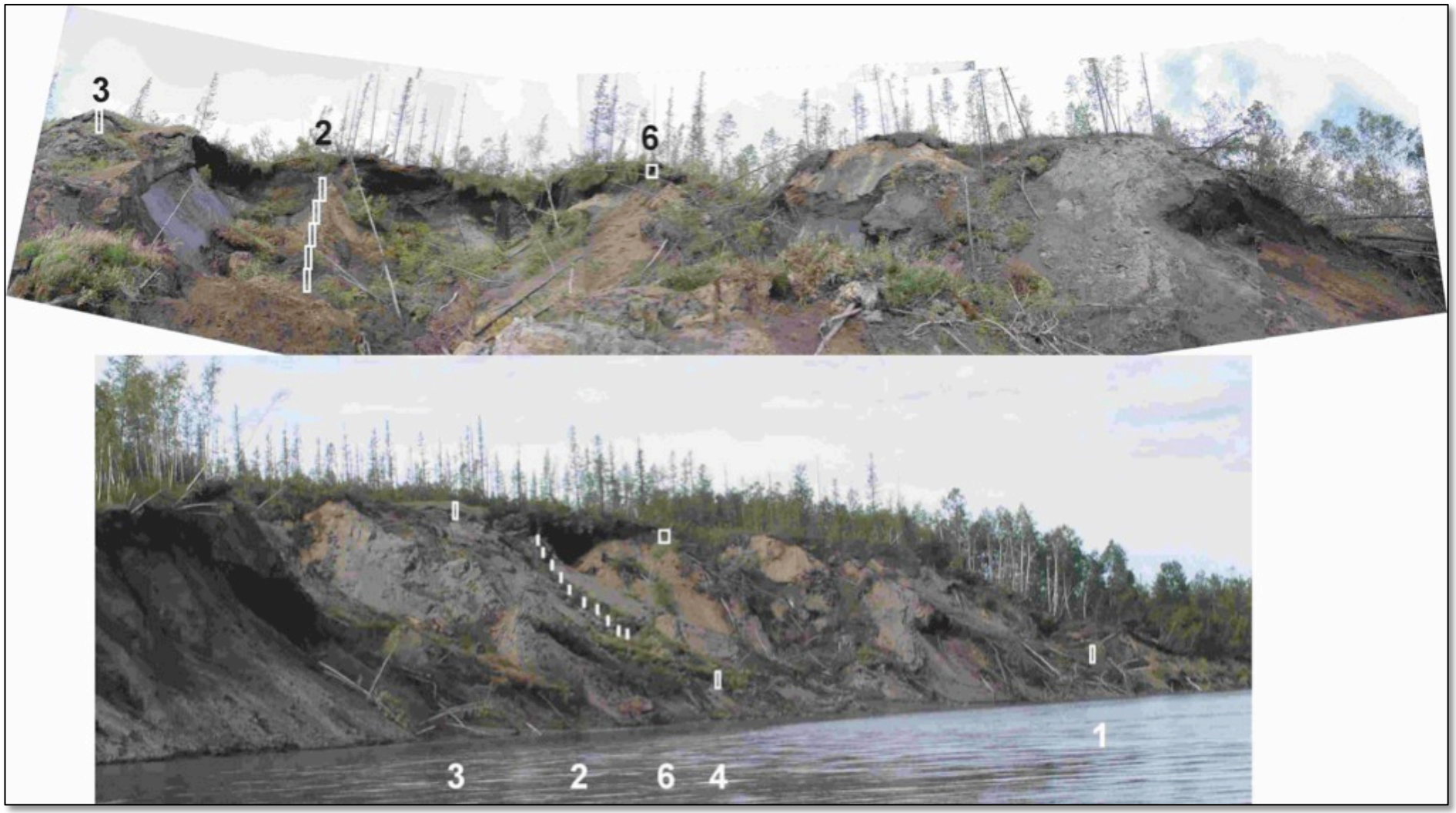


DNA-based soil analysis: functionnal groups

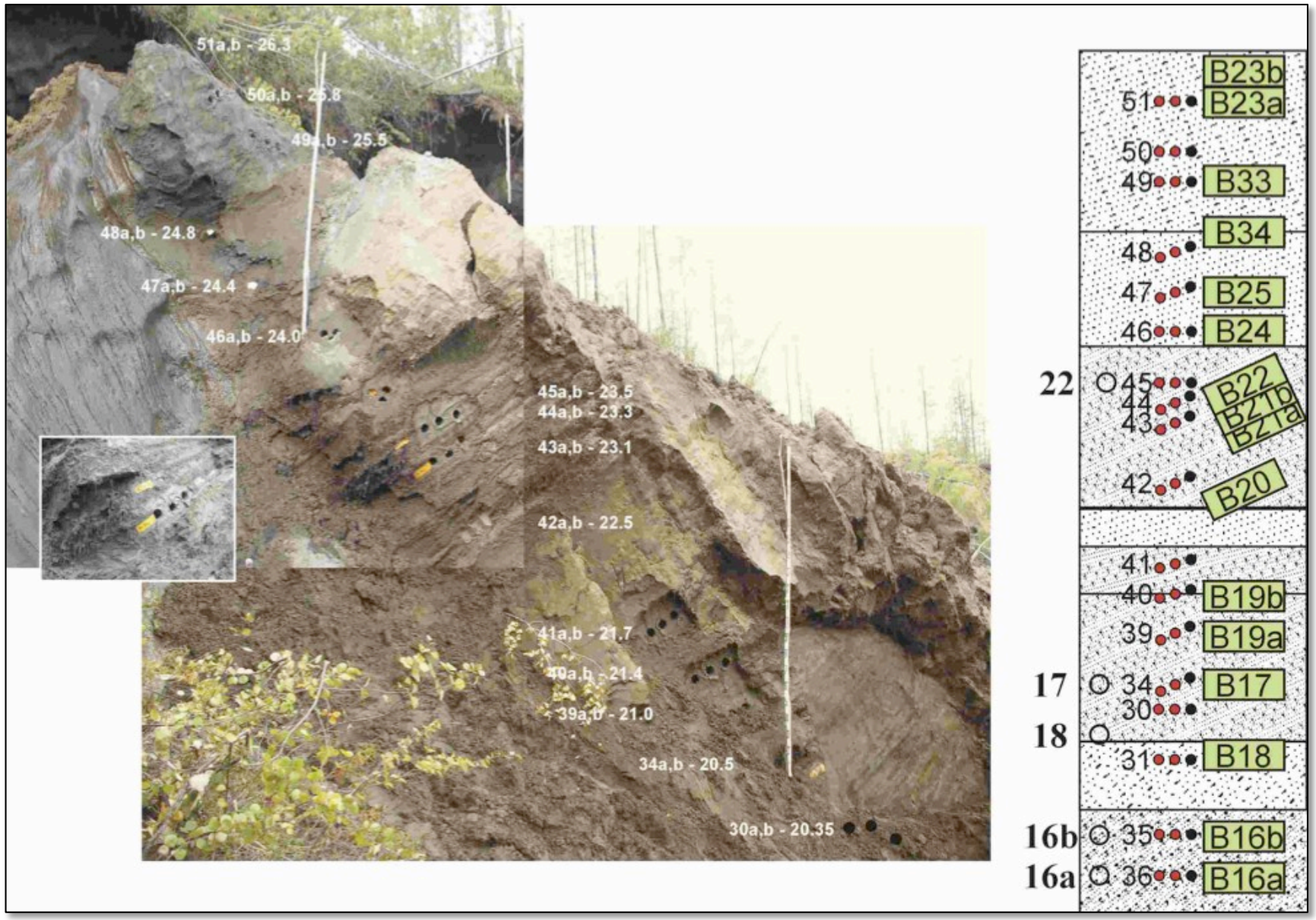


Permafrost analysis (1)

Permafrost analysis (2)



Permafrost analysis (3)



22,960 years old

Nb Seq	Family	Genus	Species
937			*Bistorta vivipara*
482		*Equisetum*	*arvense / fluviatile / sylvaticum*
94	Salicaceae	3	*Salix sp. / Chosenia arbutifolia / Populus balsamifera*
60			*Armeria scabra*
55			*Thymus oxyodontus*
43			*Lagotis glauca*
37	Asteraceae	3	Asteraceae 4
35			*Avenella flexuosa*
27		*Aconogonon*	*alaskanum / ocreatum / tripterospermum*
26		*Rumex*	5
19	Asteraceae	2	*Packera* sp. / *Senecio* sp.
19	Poaceae		Poaceae 2
16		*Ranunculus*	*acris / subborealis / turneri*
15		*Festuca*	9
13			*Hulteniella integrifolia*
11			*Saxifraga hirculus*
9			*Trientalis europaea*
8	Asteraceae	4	Asteraceae 2
7		*Valeriana*	*capitata / sambucifolia / transjenisensis*
6			*Myosotis alpestris*
6	Asteraceae	5	Asteraceae 3
6		*Empetrum*	*sibiricum / subholarcticum*
5			*Anthoxanthum nipponicum*
5			*Crepis chrysantha*
4		*Saxifraga*	5
3		*Papaver*	19
3	Poaceae	2	*Elymus* sp. / *Leymus* sp.
3			*Trollius europaeus*

The *trn*L database for Arctic plant species

- 842 species was successfully sequenced for the *trn*L intron
- 709 species were sequenced twice or more
- 33.5% of the species and 77.1% of the genera could be identified by the P6 loop
- Low resolution in Salicaceae (*Salix* sp.), in Cyperaceae (*Carex* sp.), in Asteraceae, and in Poaceae.

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2010)
doi: 10.1111/j.1755-0998.2010.02855.x

TECHNICAL ADVANCES

Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate

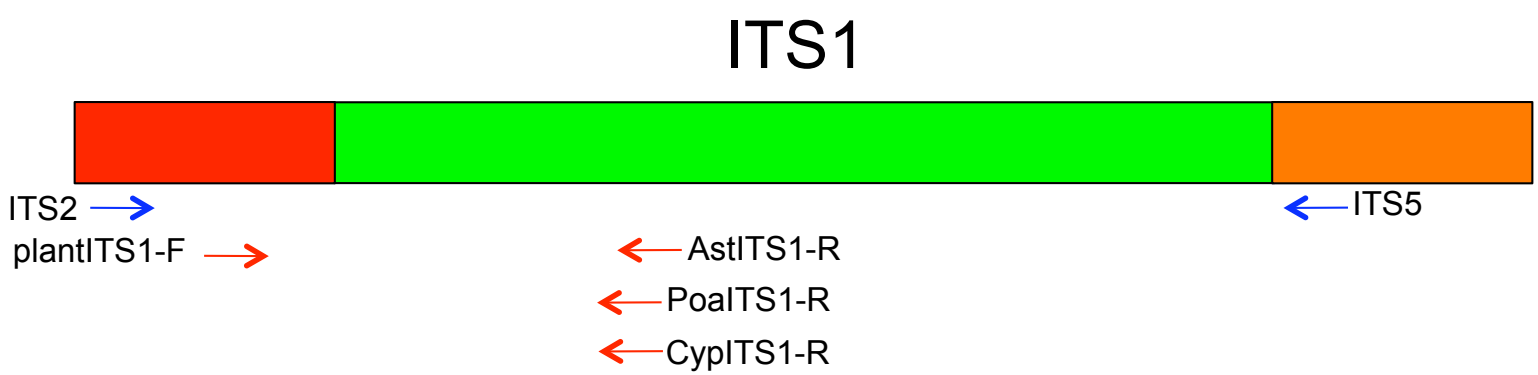
J. H. SØNSTEBØ,* L. GIELLY,† A. K. BRYSTING,‡ R. ELVEN,* M. EDWARDS,§ J. HAILE,¶** E. WILLERSLEV,¶ E. COISSAC,† D. RIOUX,† J. SANNIER,* P. TABERLET† and C. BROCHMANN*

**National Centre for Biosystematics, Natural History Museum, University of Oslo, P.O.Box 1172 Blindern, NO-0318 Oslo, Norway, †Laboratoire d'Ecologie Alpine, UMR UJF-CNRS 5553, Université Joseph Fourier, 38041 Grenoble, France, ‡Centre for Ecological and Evolutionary Synthesis, Department of Biology, University of Oslo, P.O.Box 1066 Blindern, NO-0316 Oslo, Norway, §School of Geography, University of Southampton, Room 2019 Shackleton Building, Highfield, Southampton, SO17 1BJ, UK, ¶Centre for Ancient Genetics, University of Copenhagen, Universitetsparken 15, DK-2100, Denmark, **Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK*

Sønstebø *et al.* (2010) *Molecular Ecology Resources*, in press.

New ITS1 barcode primers for Asteraceae, Poaceae, and Cyperaceae

- Asteraceae
  - plantITS1-F: 5'-GATATCCGTTGCCGAGAGTC-3'; AstITS1-R: 5'-CGGCACGGCATGTGCCAAGG-3'
  - Length: less than 90bp; 68% Asteraceae with a maximum of 2 mismatches
- Poaceae
  - plantITS1-F: 5'-GATATCCGTTGCCGAGAGTC-3'; PoaITS1-R: 5'-CCGAAGGCGTCAAGGAACAC-3'
  - Length: 54bp to 88bp; 98.5% Poaceae with a maximum of 2 mismatches
- Cyperaceae
  - plantITS1-F: 5'-GATATCCGTTGCCGAGAGTC-3'; CypITS1-R: 5'-GGATGACGCCAAGGAACAC-3'
  - Length: 46bp to 78bp; 93% Cyperaceae with a maximum of 2 mismatches



Biodiversity and DNA barcoding

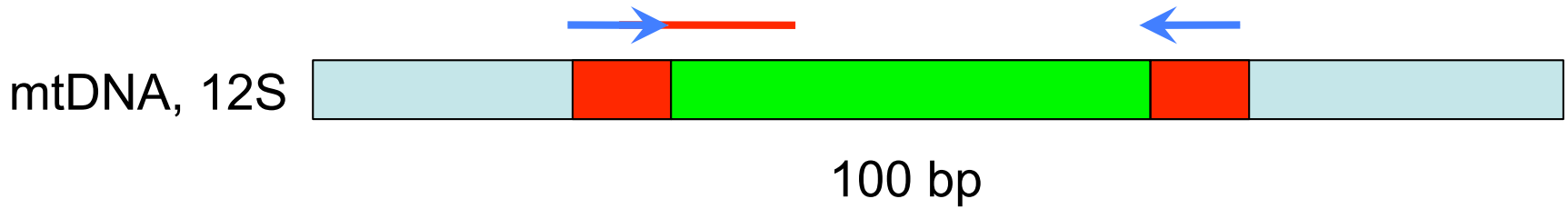
- The DNA barcoding concept
- Our scientific objectives
- New sequencing technologies
- Bioinformatic aspects
  - The challenges
  - Available programs for preparing the experiments
  - Available programs for analyzing the output of next generation sequencers
- High throughput plant identification
- High throughput animal identification
- Conclusion

Diet analysis of carnivores

- Technically more difficult than when analyzing herbivore diet
- How to amplify the prey DNA, without amplifying host DNA when using universal primers?
- Possibility of using a blocking primer for host DNA during the amplification

Experimental protocol

- Sampling in the field
- DNA extraction
- DNA amplification
  - with mtDNA 12S universal primers
  - with and without blocking primer



- DNA sequencing on a next generation sequencer (either 454 or Solexa)
- Analysis of the sequencer output

Snow leopard diet

Influence of the blocking primer during the DNA amplification

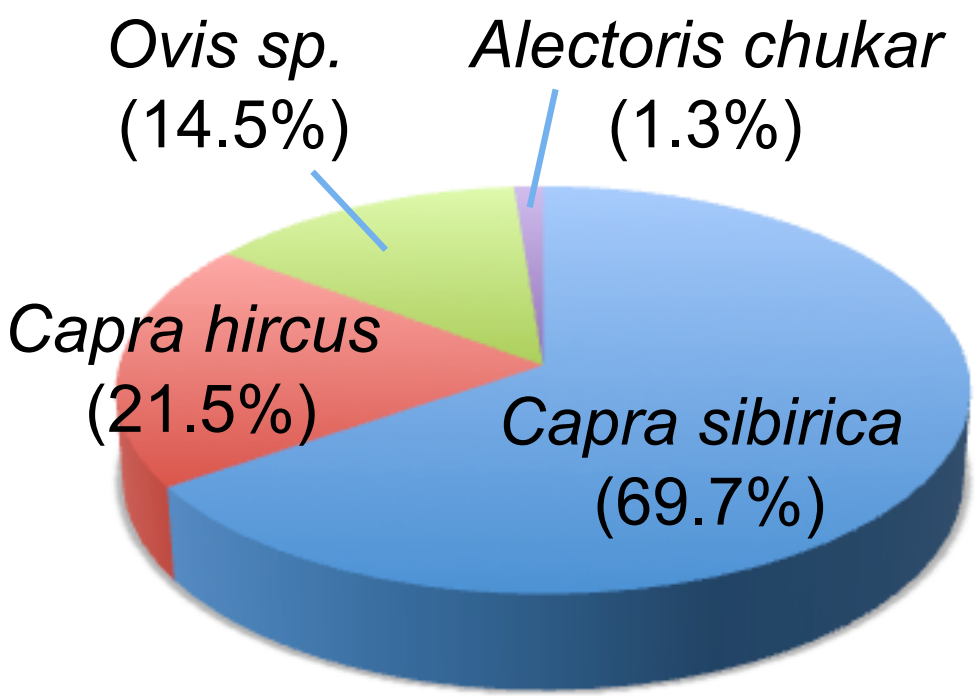
	feces 1		feces 2		feces 3	
Uncia uncia	74%	1%	97%	43%	100%	100%
Capra sibirica	26%	99%	0%	0%	0%	0%
Capra hircus	0%	0%	0%	0%	0%	0%
Ovis sp.	0%	0%	3%	57%	0%	0%

forward primer + reverse primer

forward primer + reverse primer + blocking primer

Snow leopard diet in Mongolia

89 feces analyzed,
results obtained for 70 (79%)

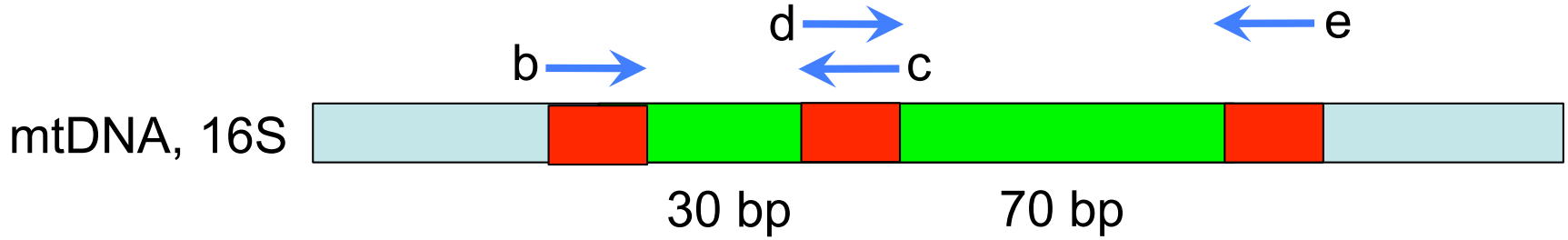


Animal biodiversity from extracellular DNA in soil

- Soil sampling in the field
- DNA extraction
- DNA amplification with universal primers
- DNA sequencing using next generation sequencers
- Data analysis…

Earthworms from soil DNA

- Eight soil samples collected per plot
- Universal short barcodes for earthworms



Earthworms from soil DNA: results

		Chartreuse		Grenoble	
Species	Barcode	Plot 1	Plot 2	Plot 1	Plot 2
Aporrectodea icterica	catcttaatgaagactaaaacttcactaaa	836954	649677	834031	1359355
Aporrectodea longa	tattttaacaaaaacccaaaaattttcaataaa	2	6	244463	271829
Aporrectodea sp	cattttaataaaaattataaattttactaaa	0	0	236024	236678
Octolasion cyaneum	cattttaatagaagcttactattctaataaa	468462	3823	0	2
Unidentified Oligochaeta	tattttaataaaatagtaaattttactaaa	334804	96337	0	1
Unidentified Oligochaeta	tattataaatcaattaataattgagcata	0	372828	0	0
Lumbricus terrestris	aatttaaataaatataaaaaatttactaaa	0	0	174286	143682
Octolasion tyrtaeum	cattttaatagaaaaataatatcctaataaa	306476	0	0	2
Unidentified Oligochaeta	tatcacaatatttatacaataaatattatg	183116	68615	0	0
Unidentified Oligochaeta	tatttttcttatactttagtaaacaaaaa	96924	42148	0	0
Lumbricus castaneus	aatttaaataaatataaaaaaatttactaaa	0	0	56	131001
Aporrectodea longa	tattttaacaaaacccaaaaattttcaataaa	2469	105312	159	145
Allobophora chlorotica	cattttaataaagatataaactttactaaa	0	0	51953	43196
Unidentified Oligochaeta	tattttatttacctaaaacagtaacaaaa	0	0	62901	0
Unidentified Oligochaeta	tatttttcttatactttagtaaataaaaa	592	61802	0	0
Unidentified Oligochaeta	taccttaacaaatattattatttcgaag	30571	0	0	0
Aporrectodea caliginosa	tattttaataaaaaaatataaatttttaataa	0	23005	0	0

Next generation sequencers and biodiversity

- Next generation sequencers
- The DNA barcoding concept
- High throughput plant identification
  - The *trn*L approach
  - Diet analysis of herbivores
  - Biodiversity assessment from environmental samples
- High throughput animal identification
  - Diet analysis of carnivores
  - Biodiversity assessment from environmental samples
- Conclusion

The "solexa" experiment

- A single run on the illumina/Solexa GS IIx sequencer
- Six months of preparation at the bench (more than 12,500 PCR products)
- Many experiments
  - Herbivore diet analysis (16 species)
  - Carnivore diet analysis (4 species)
  - Soil samples
    - Permafrost (plants)
    - Alpine soils (plants, earthworms)
    - Tropical soils (plants, earthworms, termites)
- 480,000,000 sequence reads of 108 nucleotides

DNA metabarcoding

- Works extremely well for diet analysis using feces as a source of DNA
  - For herbivores
  - For carnivores
  - Can be adjusted for other type of diet
- Works well for plant biodiversity using soil samples
  - In the Arctic
  - In the temperate region
  - In the tropical region
- Need to be further adjusted for animal biodiversity using soil samples
  - Optimistic for animals with high biomass (earthworms, etc.)
  - Might be problematic for animals with low biomass

Number of occurrences	P6 loop (chloroplast *trn*L (UAA) intron) sequences	Identification
3103	ATCCGGTTCATGGAGAC-AATAGTTT-CTT-CTTTTATTCTCCTAAGATA-GGAAGGG	*Picea*
45	. . . . . - . . . . . - . . . - . . . . . -	*Picea* variant
42	. . . . . - . . . . . - . . . - . . . . . - . . . . . -	*Picea* variant
13	. . . . . - . . . - . . . . . - . . . . . A	*Picea* variant
9	. . . . . - . . . . . - . . . T . . . . . -	*Picea* variant
9	. . . . . - . . . . . - . . . - . . . . . C . . . . . -	*Picea* variant
6	. . . . . - . . . . . - . . C - . . . . . -	*Picea* variant
6	. . . . . - . . . . . - . . . - . . . . . C . . . . . -	*Picea* variant
6	. . . . . - . . . . . - . . . C . . . . . -	*Picea* variant
5	. . . . . A . . . - . . . . . - . . . - . . . . . -	*Picea* variant
5	. . . . . - . . . . . - . . . - . . . . . T . . . . . -	*Picea* variant
5	. . . . . T . . . . . - . . . - . . . . . -	*Picea* variant
5	. . . . . - G . . . . . - . . . - . . . . . -	*Picea* variant
5	. . . . . - . . . . . - . . . - . . . . . A	*Picea* variant
5	. . . . . - . . . . . T . . . - . . . . . -	*Picea* variant
5	. . . . . A T . . . . . - . . . - . . . . . -	*Picea* variant
4	. . . . . - . . . . . - . . . - . . . . . - . . . A . .	*Picea* variant
4	- . . . . . - . . . . . - . . . - . . . . . -	*Picea* variant
4	. . . . . T - . . . . . - . . . - . . . . . -	*Picea* variant
4	. . . . . - . . . . . - . . . - . . . . . G	*Picea* variant
4	. . . . . C . . . . . - . . . . . - . . . -	*Picea* variant
4	. . . . . - . . . . . - . . . - . . . . . - . . . . . -	*Picea* variant
4	. . . . . - . . . . . - . . . - . . . . . - . . . G . . .	*Picea* variant
4	. . . . . A . . - . . . . . - . . . - . . . . . -	*Picea* variant
236	ATCCGGTTCATAGAGAAAAGGGTTTCTCTCCTTCTCCTAAGGAAAGG	*Abies*
4	. . . . . -	*Abies* variant

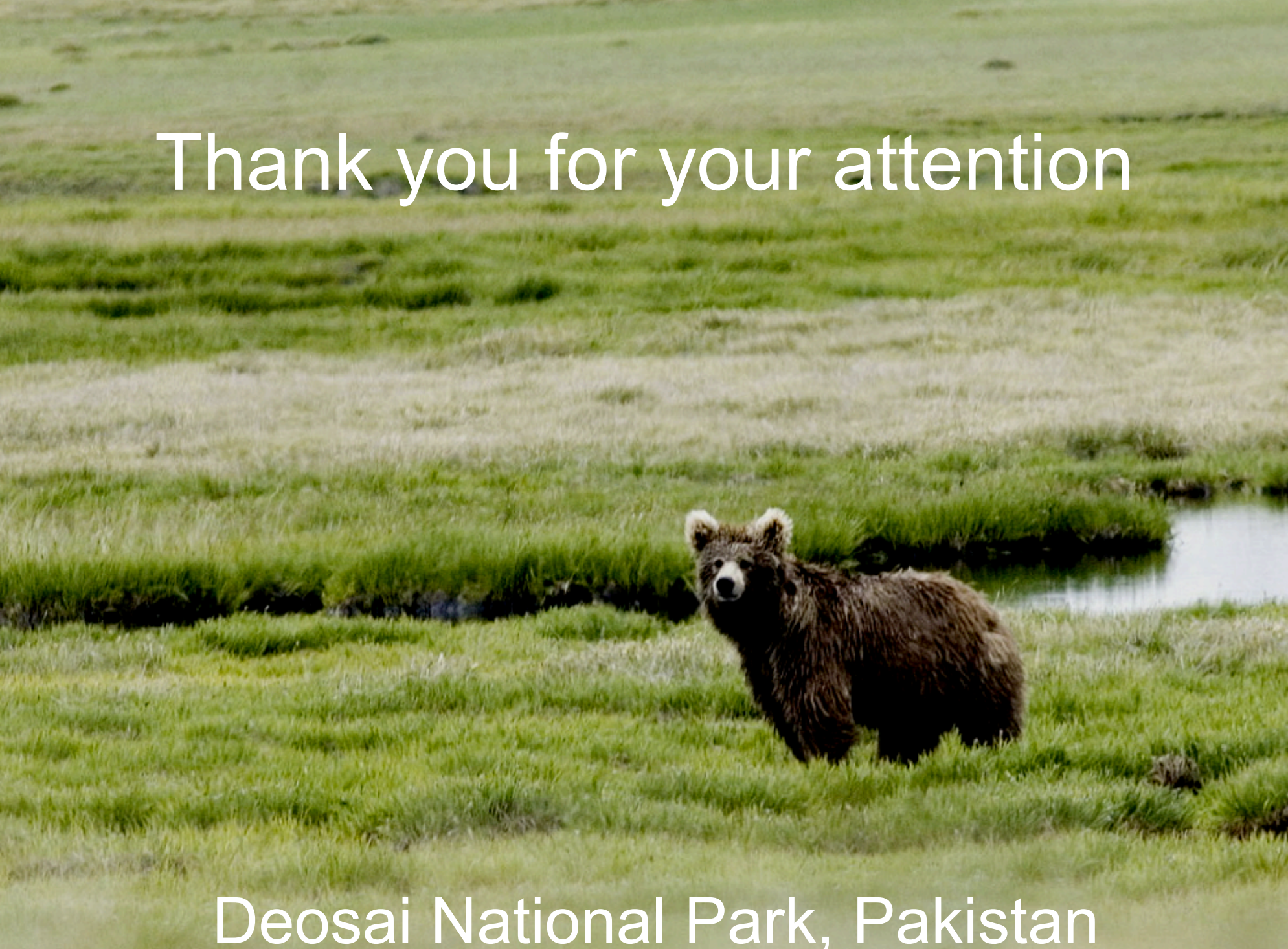
P6 loop (chloroplast *trn*L (UAA) intron) sequences obtained after high throughput pyrosequencing for a bird faeces sample (*Tetrao urogallus major*). A total of 3546 sequences were obtained with an occurrence higher than three.

Acknowledgement

Eric.Coissac@inrialpes.fr

pierre.taberlet@ujf-grenoble.fr

Rike Bienert
Kari Anne Bråthen
Christian Brochmann
Anne Krag Brysting
Corinne Cruaud
Mary Edwards
Francesco Ficetola
Ludovic Gielly
James Haile
Christelle Melo de Lima
Christian Miquel
François Pompanon
Delphine Rioux
Jorn Henrik Sønstebø
Alice Valentini
Eske Willerslev
Patrick Wincker
Nigel Yoccoz
Stéphanie Pellier-Cuit


Thank you for your attention
Deosai National Park, Pakistan