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.

Keywords: chloroplast DNA, diet analysis, DNA barcoding, faeces, pyrosequencing, *trn*L (UAA) intron, universal primers

Received 16 March 2008, accepted 24 March 2008

Introduction

Trophic relationships are of prime importance for understanding ecosystem functioning (e.g. Duffy *et al*. 2007). They can only be properly assessed by integrating the diets of animal species present in the ecosystem. Furthermore, the precise knowledge of the diet of an endangered species might be of special interest for designing a sound conservation strategy (e.g. Marrero *et al*. 2004; Cristóbal-Azkarate & Arroyo-Rodríguez 2007).

Several methods have been developed to evaluate the composition of animal diets. The simplest approach is the direct observation of foraging behaviour. However, in many circumstances, direct observation is difficult or even impossible to carry out. It is often very time-consuming or even impracticable when dealing with elusive or nocturnal animals, or when an herbivore feeds in complex environments, with many plant species that are not separated spatially. The analysis of gut contents has also been widely used to assess the diet composition of wild herbivores foraging in complex environments (Norbury & Sanson 1992). Such an approach can be implemented either after slaughtering the animals, or by obtaining the stomach extrusa after anaesthesia.

Faeces analysis represents an alternative, non-invasive, and attractive approach. Up to now, four main faeces-based techniques have been used. First, for herbivores, microscope examination of plant cuticle fragments in faecal samples has been the most widely employed technique (Holechek

Correspondence: P. Taberlet, Fax: +33(0)4 76 51 42 79;
E-mail: pierre.taberlet@ujf-grenoble.fr

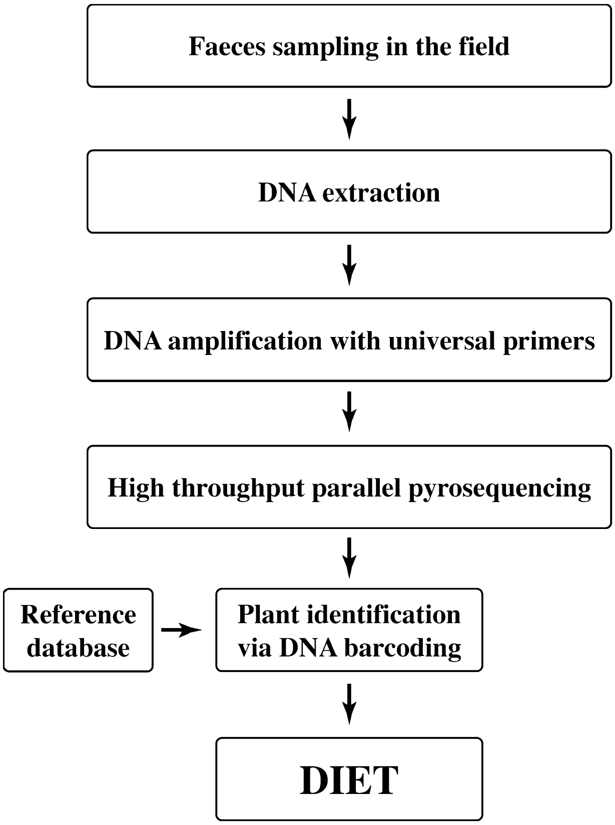
Faeces sampling in the field
DNA extraction
DNA amplification with universal primers
High throughput parallel pyrosequencing
Reference database
Plant identification via DNA barcoding
DIET

Fig. 1 Flowchart diagram showing the main steps of the *trn*L approach for assessing diet composition using faeces.

et al. 1982; McInnis *et al.* 1983). This method is very tedious to perform, and requires a considerable amount of training and a variable proportion of plant fragments remains unidentifiable. However, some herbivores do not masticate their food into small fragments, allowing some of the plants present in the faeces to be identified visually (Dahle *et al.* 1998).

The second technique is based on the analysis of the natural alkanes of plant cuticular wax (Dove & Mayes 1996). This wax is a complex chemical mixture containing *n*-alkanes (saturated hydrocarbons) with chain lengths ranging from 21 to 35 carbons, and with the odd-numbered molecules largely predominating over the even-numbered ones. There are marked differences in alkane composition among plant taxa (families, genera, species), and thus the alkane fingerprints represent a chemical approach for estimating the species composition. The approach is limited when the animal feeds in complex environment. In this case, it may be extremely difficult or impossible to have alkane concentrations in the samples that are representative of those present in the diet of the animal (Dove & Mayes 1996).

The third approach corresponds to near infrared reflectance spectroscopy (NIRS; e.g. Foley *et al.* 1998; Kaneko & Lawler 2006). Near infrared spectra depend on the number and type of chemical bonds (C-H, N-H and O-H) present in the material being analysed. After an appropriate calibration, the spectral features are used to predict the composition of new or unknown samples. The most common use of NIRS for diet analysis is the estimation of nutritional components in animal feeds, including total nitrogen, moisture, fibre, starch, etc. However, this technique has several limitations. Particle size and particle homogeneity can bias the analysis. The calibration model is a crucial and challenging step, specific to the animal under study and to the species eaten.

The fourth method is based on DNA analysis by using either specific primers for a prey group or universal primers. The former procedure has been implemented by Deagle *et al.* (2007) for analysing the diet of the Macaroni penguin (*Eudyptes chrysolophus*) using faeces as a source of DNA. The presence/absence of the different prey was detected by carrying out five different polymerase chain reaction (PCR) assays using group-specific primers. Additionally, they also tested an approach involving universal 16S rDNA primers and subsequent cloning of the PCR products. These primers were designed to amplify DNA from fish, cephalopods and crustaceans, but to prevent the amplification of bird DNA. A good concordance was found between the diet deduced from DNA-based analyses of stomach contents and of faeces. After the initial experiment of Höss *et al.* (1992), universal primers targeting the chloroplast *rbc*L gene and subsequent cloning have been used to analyse the diet of herbivorous species, either extinct species using coprolithes as a source of DNA (Poinar *et al.* 1998, 2001; Hofreiter *et al.* 2000, 2003), or living primates using fresh faeces (Bradley *et al.* 2007). The same type of DNA-based approaches was also performed for analysing gut content in insects (see review in Symondson 2002) and in birds and mammals (e.g. Jarman *et al.* 2004).

In this study, we expand the DNA-based approach by combining the plant barcoding concept (Chase *et al.* 2005, 2007) with the new highly parallel sequencing systems (Margulies *et al.* 2005). More specifically, our goal is to describe a universal method for diet analysis of herbivorous animals by amplifying the P6 loop of the chloroplast *trn*L (UAA) intron (Taberlet *et al.* 2007) via PCR (Mullis & Faloona 1987) and by subsequently sequencing individual molecules of this PCR product on the 454 automated sequencer (Roche Diagnostic). We targeted very short DNA fragments that remain in degraded DNA samples (Deagle *et al.* 2006). We demonstrate the efficiency of this new approach by analysing the diet of various herbivorous species, including mammals, birds, molluscs and insects.

Materials and methods

General strategy

Figure 1 gives an overview of the main steps necessary to estimate the diet of herbivorous species. After collecting

Table 1 Sequences of the primer pairs used for building the DNA barcoding database for the Deosai National Park [primers *c* and *d*, amplification of the whole chloroplast *trn*L (UAA) intron], and used for assessing diet from faecal samples [primers *g* and *h*, amplification of the P6 loop of the *trn*L (UAA) intron]. The code denotes the 3′-most base pair in the published tobacco cpDNA sequence (Shinozaki *et al*. 1986). The length of the amplified fragment with primer pairs *c–d* and *g–h* in tobacco is 456 bp and 40 bp, respectively

Name	Code	Sequence 5′–3′	Reference
c	A49325	CGAAATCGGTAGACGCTACG	Taberlet *et al*. (1991)
d	B49863	GGGGATAGAGGGACTTGAAC	Taberlet *et al*. (1991)
g	A49425	GGGCAATCCTGAGCCAA	Taberlet *et al*. (2007)
h	B49466	CCATTGAGTCTCTGCACCTATC	Taberlet *et al*. (2007)

faeces in the field and extracting DNA, variable and short fragments of chloroplast DNA of the eaten plant species are amplified using universal primers. These fragments are subsequently sequenced. The plant taxa they come from are then identified using the DNA barcoding concept, by comparing the sequences obtained either with public databases (GenBank, EMBL, etc.) and/or with a database made for this purpose.

Faeces sampling

A total of 36 faeces samples were collected for analysis. For mammals, we sampled 12 faeces from golden marmots (*Marmota caudata*) in Deosai National Park (Pakistan), with no more than one faeces per marmot colony. The marmot faeces were air-dried and preserved at room temperature in paper envelopes. We also analysed 12 faeces from brown bears (*Ursus arctos*) collected in the same area, and previously used in another study for identifying individual bears (Bellemain *et al*. 2007). Brown bears are mainly vegetarian in this area, and the knowledge of its diet might have some conservation implications. Brown bear faeces were preserved in alcohol. For birds, we used six capercaillie (*Tetrao urogallus*) samples previously analysed in Duriez *et al*. (2007), four from the French Pyrenees (*Tetrao urogallus aquitanus*) and two from the Corinthian Alps in Austria (*Tetrao urogallus major*). Capercaillie faeces were preserved dry in silica gel. For the invertebrates, we collected three grasshopper faeces (two from *Chorthippus biguttulus*, and one from *Gomphocerippus rufus*) and three mollusc faeces (from the snail *Helix aspersa*, and from the slugs *Deroceras reticulatum* and *Arion ater*). Insect and mollusc faeces were also preserved dry in silica gel.

DNA extraction from faeces

Total DNA was extracted from about 10 mg of sample with the DNeasy Tissue Kit (QIAGEN GmbH), following the manufacturer's instructions, except for the three grasshopper samples where the whole faeces were used. The DNA extracts were recovered in a total volume of 300 μL. Mock extractions without samples were systematically performed to monitor possible contaminations.

DNA amplification

DNA amplifications were carried out in a final volume of 25 μL, using 2.5 μL of DNA extract as template. The amplification mixture contained 1 U of Ampli*Taq* Gold DNA Polymerase (Applied Biosystems), 10 mM Tris-HCl, 50 mM KCl, 2 mM of $MgCl_2$, 0.2 mM of each dNTPs, 0.1 μM of each primer, and 0.005 mg of bovine serum albumin (BSA, Roche Diagnostics). After 10 min at 95 °C (*Taq* activation), the PCR cycles were as follows: 35 cycles of 30 s at 95 °C, 30 s at 55 °C; the elongation was removed in order to reduce the + A artefact (Brownstein *et al*. 1996; Magnuson *et al*. 1996) that might decrease the efficiency of the first step of the sequencing process (blunt-end ligation). Each sample was amplified with primers *g* and *h* (Table 1; Taberlet *et al*. 2007), modified by the addition of a specific tag on the 5′ end in order to allow the recognition of the sequences after the pyrosequencing, where all the PCR products from the different samples are mixed together. These tags were composed of six nucleotides, always starting with CC on the 5′ end, followed by four variable nucleotides that were specific to each sample. Our system was different from the tagging approach from Binladen *et al*. (2007) that proposed two variable nucleotides on the 5′ end of the primers.

DNA sequencing

PCR products were purified using the MinElute PCR purification kit (QIAGEN GmbH). DNA quantification was carried out using the NanoDrop ND-1000 UV-Vis Spectrophotometer (NanoDrop Technologies). Then, a mix was made taking into account these DNA concentrations in order to obtain roughly the same number of molecules per PCR product corresponding to the different faeces samples. Large-scale pyrosequencing was carried out on the 454 sequencing system (Roche) following manufacturer's instructions, and using the GS 20 for marmot and bear, and the GS FLX for other samples.

DNA barcoding database for the Deosai National Park

In order to more precisely assess the diets of brown bears and golden marmots in Deosai National Park, leaves of the most common plant species occurring in this alpine environment were collected and identified by three botanists (Dr Muhammad Qaiser, Dr Muqarrab Shah and Dr Mir Ajab Khan). The database was elaborated by sequencing the whole chloroplast *trn*L (UAA) intron of these species using the *c–d* primer pair (Table 1; Taberlet *et al*. 1991), and following the protocol described in Taberlet *et al*. (2007). This database will be further called DNPDB.

Data analysis for estimating diet composition

The first step of analysing the mix of sequences obtained after the pyrosequencing consisted of dispatching the different sequences according to the tag present on the 5′ end of the primers. Thus, for each sample (each faeces), a data set was generated, containing all the sequences having the relevant tag. Then, these sequences were analysed to determine the diet. To limit the influence of sequence errors (Huse *et al*. 2007), only sequences that were present more than three times were considered in the subsequent analyses. The taxon was assigned to each sequence in a data set by similarity assessment using MEGABLAST (Zhang *et al*. 2000) or FASTA (Pearson & Lipman 1988) algorithm. Reference databases used are DNPDB using FASTA algorithm on local computers for bear and marmot samples and GenBank using MEGABLAST on the National Center for Biotechnology Information (NCBI) web site (www.ncbi.nlm.nih.gov/blast) for all other data sets and for the bear and marmot sequences that were not fully identified using the DNPDB. We have verified that with the high similarity threshold used in this assignation step (98% of identity and 100% of query coverage for species level identification), the FASTA and MEGABLAST approaches yielded similar results. If two or more taxa could be assigned with the same score for a given sequence, we assigned this sequence to the higher taxonomic level that included both taxa. This method results in some sequence taxa being assigned to the rank of genus or family.

Results

DNA barcoding database for the Deosai National Park

The chloroplast *trn*L (UAA) intron was sequenced for 91 plant species belonging to 69 genera and 32 families. Seventy-five per cent of the species analysed have a unique P6 loop sequence (i.e. the sequence amplified with the *g–h* primer pair) and thus can be identified to species. Of the remaining 25, 20% could be identified to genus, and 5% to family. All these sequences have been deposited in European Molecular Biology Laboratory (EMBL) database, under accession nos EU326032–EU326103.

Pyrosequencing results

For the analysis of the 36 faeces, we obtained a total of 97 737 P6 loop sequences, corresponding to an average of 2715 ± 1130 sequences per sample (range from 1049 to 5368 per sample). The size range of the PCR product (excluding primers) was 20–85 bp. We obtained 100% of full-length reads of the amplified region on both the GS 20 and the GS FLX, but we missed a part of the reverse primer and the reverse tag only in the GS 20 experiments (in 13.9% and 1.2% of the sequences in bears and marmots, respectively). In each sample, a few sequences were found hundreds of times, whereas some other sequences are only represented either once or by very few occurrences (see details in Table 2). The sequences occurring only once, twice, or three times were not taken into account in the subsequent analysis. They were almost always very close to a highly represented sequence, and thus considered to be the result of sequencing errors in the P6 loop. Sequences occurring more than three times, but very close to a highly represented sequence were also considered to be sequencing errors. For example, the sequence found 45 times in Table 2 clearly corresponds to a variant in a T stretch of the most common sequence found 3103 times and identified as *Picea*. In rare cases, we also found sequences represented only once, that were not close to a highly represented sequence. Such sequences most likely correspond to a sequencing error within the tag, leading to an assignment to a wrong sample. This observation led us to modify our tagging system (see Discussion).

DNA-based diet analysis

The DNA-based diet analyses of marmots and bears are summarized in Table 3 and Fig. 2. Sixty-four per cent and 31% of the different P6 loop sequences obtained in their diet were identified to species for marmots and bears, respectively. Overall, the marmot has a much more eclectic diet, with 28 species identified (out of the 779 different P6 loop sequences), belonging to 15 families. Only 557 different P6 loop sequences were identified in the brown bear diet, which is composed mainly of Poaceae and Polygonaceae, with a significant contribution of Cyperaceae and Apiaceae.

Table 4 gives the results obtained for the birds, molluscs and insects. All these results are consistent with what we know about the diet of these animals, particularly for capercaillie, which eat mainly conifer needles in winter, and grasshoppers, which eat mainly grasses.

Discussion

Using faeces as a source of DNA, and by combining universal primers that amplify a very short but informative fragment of chloroplast DNA and large-scale pyrosequencing, we were able to successfully assess the diet composition of several herbivorous species. This DNA-based method is broadly applicable to potentially all herbivorous species eating angiosperms and gymnosperms, including mammals, insects, birds and molluscs.

Table 2 P6 loop [chloroplast *trn*L (UAA) intron] sequences obtained after high throughput pyrosequencing for the bird faeces sample n°5 (*Tetrao urogallus major*). A total of 4602 sequences were obtained, containing 3546 sequences with an occurrence higher than three. The diet was composed of two plant taxa: *Picea* and *Abies*. Besides the most common sequences for each of these two taxa, it is interesting to note the presence of sequence variants due to errors originating from the degradation of the template DNA in faeces, from nucleotide misincorporation during DNA amplification, or from the sequencing process on the 454 sequencer. Out of the 4602 sequences, 27.4% corresponded to sequence variants. By removing the sequences occurring only once, twice, or three times, the percentage of sequence variants decreased to 4.5%

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

Such an approach has many advantages over previous methods used for diet analysis (i.e. microscope examination of plant cuticle fragments, chemical analysis of alkanes, NIRS). Our approach is robust and reliable, in relation to the very short length of the amplified region. The primers target highly conserved regions in angiosperms and gymnosperms, preventing strong bias due to primer mismatch in the efficiency of amplifications among species (Taberlet *et al*. 2007). The two highly conserved regions targeted by these primers flank a short and variable region that allows the identification of the plant taxa. The results obtained in marmots show clearly that the system is particularly well adapted for analysing complex situations, when the diet is composed of many different species. This approach can be coupled with individual identification using microsatellite polymorphism (Taberlet & Luikart 1999), allowing diet comparisons among individuals, even without observing the animals. An alternative and very inexpensive approach could involve the pooling of many faeces in the same DNA extraction in order to obtain the average diet composition directly, but this strategy would prevent the analysis of individual diets.

The *trn*L approach represents a significant progress in plant identification when using faecal material. The same standardized method is easy to implement and can be applied to a wide range of animal species. It is particularly well suited for large-scale analyses, with the possibility to analyse several hundreds of samples in the same 454 GS FLX sequencing run and to automate the sequence analysis by implementing bioinformatic tools. This offers the prospect of following the diet composition over seasons and of comparing among age classes, individuals and sexes. Within the same species, it also allows the analysis of diet shifts according to plant availability and food preferences.

However, this method still has some limitations, and it is clear that the resolution does not reach the species level in all cases. However, by building a comprehensive database of *trn*L (UAA) introns for the majority of the plant species

Table 3 Plant taxa identified in the diet of the Himalayan brown bear (*Ursus arctos*) and of the golden marmot (*Marmota caudata*) in Deosai National Park (Pakistan), based on sequence variation of the P6 loop of the chloroplast *trn*L (UAA) intron using faeces as a source of DNA

			Ursus arctos													*Marmota caudata*												
			Faeces sample													Faeces sample												
Family	Plant taxon	Level of identification	1	2	3	4	5	6	7	8	9	10	11	12	Total	1	2	3	4	5	6	7	8	9	10	11	12	Total
Apiaceae	Apoideae	subfamily				x									1													—
	Heracleum candicans	species		x		x					x	x		x	5		x					x		x				3
	Pleurospermum hookeri	species													—				x	x	x			x				4
Araceae	Araceae*	family													—			x										1
Asteraceae	*Anaphalis nepalensis*	species													—										x			1
	Anthemideae_1*	tribe				x									1		x		x	x		x	x	x	x	x		8
	Anthemideae_2*	tribe													—				x	x		x			x			4
	Aster falconeri	species													—				x			x	x		x		x	5
	Asteraceae_1*	family													—				x									1
	Asteraceae_2*	family		x										x	2				x	x	x	x			x	x		6
	Asteraceae_3*	family													—			x	x									2
	Asteraceae_4*	family													—						x				x			2
	Asteraceae_5*	family													—						x				x			2
	Asteraceae_6*	family													—											x		1
	Asteroideae_1*	subfamily													—		x	x		x	x	x	x	x		x		8
	Asteroideae_2*	subfamily													—					x		x			x	x		4
	Asteroideae_3*	subfamily													—					x								1
	Asteroideae_4*	subfamily													—				x									1
	Coreopsideae*	tribe													—			x		x	x							3
	Gnaphalieae*	tribe													—					x								1
	Inuleae*	tribe									x				1			x		x			x			x		4
	Leontopodium brachyactis	species													—							x						1
Brassicaceae	Brassicaceae	family													—										x			1
	Draba oreades	species													—			x			x							2
	Thlaspi andersonii	species													—	x					x							2
Cannabaceae	*Cannabis sativa**	species													—									x				1
Caryophyllaceae	*Cerastium*	genus					x								1	x		x	x		x	x	x		x	x	x	9
	Cerastium cerastoides	species										x		x	2	x		x	x		x	x	x	x	x	x	x	10
	Cerastium pusillum	species					x								1	x		x							x	x	x	5
	*Silene**	genus													—	x		x										2
	Silene tenuis	species													—	x								x		x		3
Crassulaceae	Crassulaceae	family													—				x	x		x		x				4
	Rhodiola	genus													—		x											1
Cyperaceae	*Carex*	genus		x		x	x		x		x	x		x	7													—
	Carex diluta	species		x		x	x				x	x		x	6													—
Fabaceae	*Astragalus rhizanthus*	species		x											1	x	x	x	x	x		x	x	x	x			9
	Galegeae	tribe		x											1	x			x			x						3
	Oxytropis cachemiriana	species													—	x		x	x		x			x		x	x	7
Juncaceae	*Juncus**	genus									x				1													—
Lamiaceae	*Dracocephalum nutans*	species													—		x	x										2
	Mentheae	tribe			x	x									2	x	x	x	x	x			x	x		x		8
Onagraceae	*Chamerion latifolium*	species													—			x										1
Orobanchaceae	*Pedicularis*	genus		x											1													—
	Pedicularis albida	species		x											1													—
Papaveraceae	*Papaver nudicaule*	species													—	x	x											2
Pinaceae	*Cedrus**	genus			x										1													—
	*Picea**	genus													—			x										1
Plantaginaceae	*Lagotis kunawurensis*	species													—											x		1
	*Plantago**	genus													—											x		1
Poaceae	*Agrostis vinealis*	species	x			x		x	x	x		x			6								x					1
	Elymus longiaristatus	species													—						x		x		x			3
	Poa alpina	species													—						x							1

Table 3 *Continued*

			Ursus arctos													*Marmota caudata*												
			Faeces sample													Faeces sample												
Family	Plant taxon	Level of identification	1	2	3	4	5	6	7	8	9	10	11	12	Total	1	2	3	4	5	6	7	8	9	10	11	12	Total
	Poa	genus												x	1													—
	Poa supina	species													—					x	x			x			x	4
	Pooideae*	subfamily	x	x	x	x	x	x	x	x	x	x	x	x	12	x		x			x	x	x	x		x		7
Polygonaceae	*Aconogonon rumicifolium*	species			x							x	x		3		x		x	x								3
	Bistorta affinis	species				x		x				x		x	4													—
	Polygonaceae	family													—							x		x	x			3
	Polygonum cognatum	species													—	x		x			x							3
	*Rumex**	genus									x				1	x		x	x	x	x	x		x	x			8
	Rumex nepalensis	species									x				1	x		x	x	x	x	x			x			7
Ranunculaceae	*Aconitum violaceum*	species										x			1													—
Rosaceae	*Cotoneaster affinis*	species													—											x		1
	Potentilla argyrophylla	species													—	x	x	x			x			x				5
	Rosoideae	subfamily												x	1	x	x			x		x		x				5
Rubiaceae	*Galium boreale*	species												x	1		x											1
Saxifragaceae	*Saxifraga hirculus*	species											x		1												x	1
Solanacee	*Solanum**	genus													—						x	x						2
Total number of plant species per faeces			2	9	4	9	5	3	3	2	8	9	3	10		17	12	21	18	18	20	19	11	17	17	16	7	

*Plants identified by comparing the sequence with sequence data in public databases.

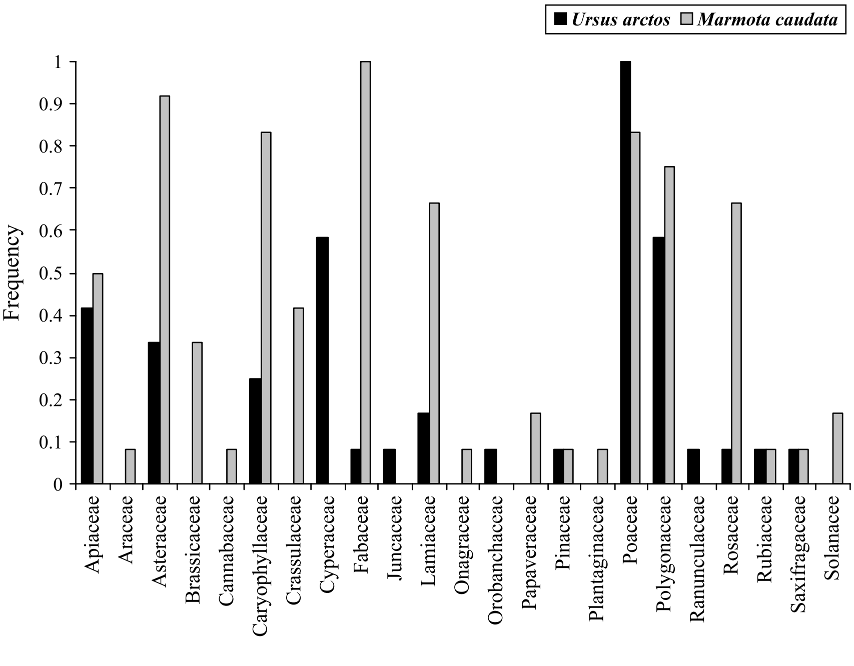

Fig. 2 Comparison of the diet compositions of the golden marmot (*Marmota caudata*) and of the brown bear (*Ursus arctos*) in Deosai National Park (Pakistan). See Table 3 for the plant taxa identified within each of these families. The *Y*-axis corresponds to the frequency of presence of taxa from the same family in the twelve samples of each mammal species.

that occur in a particular area, usually about 50% of the different species should be identified to species, and 90% to genus. The percentage of identification to the species level is lower when the sequences obtained were compared to public databases (cases of insect, molluscs and birds) instead of local databases (cases of marmots and bears). Such a difference is due to the higher occurrence of closely related species that exhibit the same P6 loop sequence in public databases. It is interesting to note that some genera exhibit a limited variation (e.g. *Carex*) or almost no variation (e.g. *Salix*, *Pinus*) on this P6 loop. When it is important to determine the species, we suggest to complement the

Table 4 Plant taxa identified in the diet of birds, molluscs and insects based on sequence variation of the P6 loop of the chloroplast *trn*L (UAA) intron using faeces as a source of DNA

Family	Plant taxon	Level of identification	B1	B2	B3	B4	B5	B6	M1	M2	M3	I1	I2	I3
Apoideae	Apoideae	family											x	
Asteraceae	Asteraceae	family								x	x			
Brassicaceae	Brassicaceae	family							x					
Ericaceae	Rhodoreae	tribe	x											
Fagaceae	Fagaceae	family						x						
Lamiaceae	Nepetoideae	subfamily									x			
Linnaeaceae	Linnaeaceae	family	x											
Oleaceae	Oleaceae	family								x				
Pinaceae	*Abies*	genus					x							
	Picea	genus					x	x						
	Pinaceae	family						x						
	Pinus	genus	x	x	x	x		x						
Plantaginaceae	*Veronica*	genus									x			
	Veroniceae	tribe								x				
Poaceae	*Bromus*	genus											x	x
	Holcus lanatus	species												x
	Hordeum	genus											x	
	Poae	tribe											x	
	Pooideae	subfamily										x	x	x
Ranunculaceae	Ranunculus	genus				x								
Rosaceae	Maloideae	subfamily								x	x			
	Prunus	genus									x			
Total number of plants per faeces			3	1	1	2	2	4	1	4	5	1	5	3

B1, *Tetrao urogallus aquitanus* Sample 1; B2, *T. u. aquitanus* Sample 2; B3, *T. u. aquitanus* Sample 3; B4, *T. u. aquitanus* Sample 4; B5, *T. u. major* Sample 1; B6, *T. u. major* Sample 2; M1, *Helix aspersa*; M2, *Deroceras reticulatum*; M3, *Arion ater*; I1, *Chorthippus biguttulus* Sample 1 (male); I2, *C. biguttulus* Sample 2 (female); I3, *Gonphocerippus rufus*.

universal *trn*L approach by one or several additional systems, specially designed for amplifying a short and variable region in these genera. According to the availability of more and more DNA sequences in databases, primer pairs can be designed that are specific to these problematic genera. These primers might target other more variable parts of the chloroplast DNA, or the nuclear ribosomal DNA, such as the internal transcribed spacers (as in Bradley *et al*. 2007).

We would like to highlight two potential difficulties of our approach, linked to the sequencing strategy using a huge mix of DNA molecules, and to the sequencing errors observed with the 454 sequencer. The 454 sequencer produces several hundreds of thousands of sequences per run, in a single file containing unsorted sequences corresponding to the mix of DNA molecules. The only way to reduce costs, while still producing many sequences per sample, is to pool many PCR products before the sequencing step. As a consequence, we tagged each sample differently in order to find the corresponding sequences in the sequencer output. Our first tagging system added a 5′-CCNNNN-3′ tag to the 5′ end of the primers. However, due to the occurrence of sequencing errors within the tags, either substitutions or indels (insertions/deletions), we suggest to improve the tagging system by using the following sequence: 5′-CCDNNNN-3′ (D = A or G or T), with at least two differences among tags and avoiding stretches of the same nucleotide longer than two. Using this latter tagging system and 96 different tags, we are currently able to pool 96 samples per region, each region producing about 200 000 reads, and each run being composed of two regions. This corresponds to a total of 192 samples per GS FLX run. The second difficulty comes from the sequencing errors within the P6 loop itself. Such errors can come from the degradation of the template DNA in faeces, from nucleotide misincorporation during DNA amplification, or from the sequencing process itself. The 454 sequencer is known for having difficulty in counting the exact number of repeats of the same nucleotide, even in short stretches of three or four nucleotides. We also observed many substitutions, and indels not linked to stretches (see Table 2). All these errors make the species identification more complex. Nevertheless, the exact sequences are usually present in a high copy number, whereas those containing errors occur at a low frequency (see Table 2). In this first study, we only considered sequences present at least four times. It is clear that the method can be improved significantly by a better knowledge of the types of sequencing errors and of their associated probabilities. The availability of a *trn*L (UAA) intron database with the plant species available in the study area greatly facilitates plant identification when using the *trn*L approach for diet analyses.

Another potential difficulty is the risk of contamination, from the sampling step in the field to the sequencing step. The g–h primer pair is highly efficient, and we do not recommend carrying out more than 35 amplification cycles, except if strong measures are taken to avoid potential contaminations, as in ancient DNA studies. During a pilot experiment, we noticed that samples extracted with the QIAGEN Stool Kit (QIAGEN GmbH) systematically contained potato DNA, most likely coming from the 'inhibitex' pill used during the extraction process. QIAGEN technical support confirmed that 'it cannot be ruled out that Inhibitex may contain DNA from plants'. As a consequence, we recommend to avoid the QIAGEN Stool Kit when amplifying plant DNA.

An important aspect in diet analysis is the absolute or relative quantification of the different plant species that have been eaten. The *trn*L approach provides the number of molecules after DNA amplification. However, at the moment these numbers cannot be interpreted as quantitative for several reasons. First, the preferential amplification of some species when analysing a mixture of templates is well known (Polz & Cavanaugh 1998). The fact that the *g–h* primer pair targets highly conserved regions, with almost no variation (Taberlet *et al*. 2007), should limit preferential amplification due to primer mismatch. Additionally, new technologies, such as emulsion PCR, can minimize this problem and at the same time should enable the quantification of DNA fragments in a mix (Williams *et al*. 2006). Second, the amount of template DNA (chloroplast DNA) clearly varies among the types of tissue eaten. Leaves will undoubtedly provide more chloroplast DNA than roots fruit, or seeds, and the *trn*L approach cannot determine the tissue that has been eaten. Knowing the species eaten, the NIRS method has the potential of providing information about the tissue eaten. Third, the *trn*L approach alone cannot assess the absolute quantity of the different plant species eaten. Thus, it provides an estimate of the frequency of occurrence of a food item in the faeces, but not an estimate of the volume eaten. In simple conditions, that is, when the animal is eating only a few species and is additionally feed with a known amount of even-numbered alkane molecules, the alkane approach can supply estimates of the absolute quantity of plant eaten (Dove & Mayes 1996). Consequently, the *trn*L, the NIRS, and the alkane approaches should be considered as complementary.

Non-invasive genetic studies are very attractive and now extensively used, especially when dealing with endangered species. With this new *trn*L approach for diet analysis, we widen the field of non-invasive analysis using faeces as a source of information. This opens new perspectives in conservation biology and more generally in ecological studies by enhancing research on resource partitioning among competing species, and describing food webs in ecosystems.

Acknowledgements

We thank five anonymous reviewers for useful suggestions on the initial version of the manuscript. This work has been supported by the French Agence Nationale de la Recherche (ANR-06-PNRA-024–06) and the International Bear Association (John Sheldon Bevins Memorial Foundation). AV and MAN were supported by PhD scholarships from the Italian and the Norwegian governments, respectively. FP was supported by the French Institut National de la Recherche Agronomique. Noor Kamal Khan assisted in collection of plants from the Deosai National Park, and for identification of these plants we thank Dr Muhammad Qaiser (University of Karachi, Karachi), Dr Muqarrab Shah (Pakistan Museum of Natural History, Islamabad), and Dr Mir Ajab Khan (Quaid-e-Azam University, Islamabad). We also thank Delphine Rioux for her technical help when building the reference *trn*L intron database.

References

Bellemain E, Nawaz MA, Valentini A, Swenson JE, Taberlet P (2007) Genetic tracking of the brown bear in northern Pakistan and implications for conservation. *Biological Conservation*, **134**, 537–547.

Binladen J, Gilbert MTP, Bollback JP *et al*. (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE*, February 2007, e197.

Bradley BJ, Stiller M, Doran-Sheehy DM *et al*. (2007) Plant DNA sequences from feces: Potential means for assessing diets of wild primates. *American Journal of Primatology*, **69**, 699–705.

Brownstein MJ, Carpten JD, Smith JR (1996) Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping. *BioTechniques*, **20**, 1008–1010.

Chase MW, Salamin N, Wilkinson M *et al*. (2005) Land plants and DNA barcodes: short-term and long-term goals. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **360**, 1889–1895.

Chase MW, Cowan RS, Hollingsworth PM *et al*. (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon*, **56**, 295–299.

Cristóbal-Azkarate J, Arroyo-Rodríguez V (2007) Diet and activity pattern of howler monkeys (*Alouatta palliata*) in Los Tuxtlas, Mexico: effects of habitat fragmentation and implications for conservation. *American Journal of Primatology*, **69**, 1013–1029.

Dahle B, Sørensen OJ, Wedul EH, Swenson JE, Sandegren F (1998) The diet of brown bears *Ursus arctos* in central Scandinavia: effect of access to free-ranging domestic sheep *Ovis aries*. *Wildlife Biology*, **4**, 147–158.

Deagle BE, Eveson JP, Jarman SN (2006) Quantification of damage in DNA recovered from highly degraded samples – a case study on DNA in faeces. *Frontiers in Zoology*, **3**, 11.

Deagle BE, Gales NJ, Evans K *et al*. (2007) Studying seabird diet through genetic analysis of faeces: a case study on macaroni penguins (*Eudyptes chrysolophus*). *PLoS ONE*, **2**, e831.

Dove H, Mayes RW (1996) Plant wax components: a new approach to estimating intake and diet composition in herbivores. *Journal of Nutrition*, **126**, 13–26.

Duffy JE, Carinale BJ, France KE, McIntyre PB, Thebault E, Loreau M (2007) The functional role of biodiversity in ecosystems: incorporating trophic complexity. *Ecology Letters*, **10**, 522–538.

Duriez O, Sachet JM, Ménoni E *et al*. (2007) Phylogeography of the capercaillie in Eurasia: what is the conservation status in the Pyrenees and Cantabrian Mounts? *Conservation Genetics*, **8**, 513–526.

Foley WJ, McIlwee A, Lawler I *et al*. (1998) Ecological applications of near infrared reflectance spectroscopy – a tool for rapid, cost-effective prediction of the composition of plant and animal tissues and aspects of animal performance. *Oecologia*, **116**, 293.

Hofreiter M, Poinar HN, Spaulding WG *et al*. (2000) A molecular analysis of ground sloth diet through the last glaciation. *Molecular Ecology*, **9**, 1975–1984.

Hofreiter M, Betancourt JL, Sbriller AP, Markgraf V, McDonald HG (2003) Phylogeny, diet, and habitat of an extinct ground sloth from Cuchillo Cura, Neuquen Province, southwest Argentina. *Quaternary Research*, **59**, 364–378.

Holechek JL, Vavra M, Pieper RD (1982) Botanical composition determination of range diets: a review. *Journal of Range Management*, **35**, 309–315.

Höss M, Kohn M, Pääbo S, Knauer F, Schröder W (1992) Excrement analysis by PCR. *Nature*, **359**, 199.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively-parallel DNA pyrosequencing. *Genome Biology*, **8**, R143.

Jarman SN, Deagle BE, Gales NJ (2004) Group-specific polymerase chain reaction for DNA-based analysis of species diversity and identity in dietary samples. *Molecular Ecology*, **13**, 1313–1322.

Kaneko H, Lawler IR (2006) Can near infrared spectroscopy be used to improve assessment of marine mammal diets via fecal analysis? *Marine Mammal Science*, **22**, 261–275.

Magnuson VL, Ally DS, Nylund SJ *et al*. (1996) Substrate nucleotide-determined non-templated addition of adenine by Taq DNA polymerase: implications for PCR-based genotyping and cloning. *BioTechniques*, **21**, 700–709.

Margulies M, Egholm M, Altman WE *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Marrero P, Oliveira P, Nogales M (2004) Diet of the endemic Madeira Laurel Pigeon *Columba trocaz* in agricultural and forest areas: implications for conservation. *Bird Conservation International*, **14**, 165–172.

McInnis ML, Vavra M, Krueger WC (1983) A comparison of 4 methods used to determine the diets of large herbivores. *Journal of Range Management*, **36**, 700–709.

Mullis KB, Faloona F (1987) Specific synthesis of DNA *in vitro* via a polymerase-catalysed chain reaction. *Methods in Enzymology*, **155**, 335–350.

Norbury GL, Sanson GD (1992) Problems with measuring diet selection of terrestrial, mammalian herbivores. *Australian Journal of Ecology*, **17**, 1–7.

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences, USA*, **85**, 2444–2448.

Poinar HN, Hofreiter M, Spaulding WG *et al*. (1998) Molecular coproscopy: dung and diet of the extinct ground sloth *Nothrotheriops shastensis*. *Science*, **281**, 402–406.

Poinar HN, Kuch M, Sobolik KD *et al*. (2001) A molecular analysis of dietary diversity for three archaic Native Americans. *Proceedings of the National Academy of Sciences, USA*, **98**, 4317–4322.

Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology*, **64**, 3724–3730.

Shinozaki K, Ohme M, Tanaka M *et al*. (1986) The complete nucleotide sequence of tobacco chloroplast genome: its gene organization and expression. *EMBO Journal*, **5**, 2043–2049.

Symondson WOC (2002) Molecular identification of prey in predator diets. *Molecular Ecology*, **11**, 627–641.

Taberlet P, Luikart G (1999) Non-invasive genetic sampling and individual identification. *Biological Journal of the Linnean Society*, **68**, 41–55.

Taberlet P, Gielly L, Pautou G, Bouvet J (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology*, **17**, 1105–1109.

Taberlet P, Coissac E, Pompanon F *et al*. (2007) Power and limitations of the chloroplast *trn*L (UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, **35**, e14.

Williams R, Peisajovich SG, Miller OJ *et al*. (2006) Amplification of complex gene libraries by emulsion PCR. *Nature Methods*, **3**, 545–550.

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, **7**, 203–214.