UNIVERSITÀ DEGLI STUDI DELLA TUSCIA DI VITERBO

DIPARTIMENTO DI ECOLOGIA E SVILUPPO SOSTENIBILE

CORSO DI DOTTORATO DI RICERCA

ECOLOGIA E GESTIONE DELLE RISORSE BIOLOGICHE - XX CICLO.

Non-invasive diet analysis based on DNA Barcoding: the Himalayan Brown Bears (*Ursus arctos isabellinus*) as a case study.

BIO/07

Coordinatore: Prof. Roberta Cimmaruta

Tutor: Prof. Giuseppe Nascetti

Tutor: Dott. Pierre Teberlet

Dottoranda: Alice Valentini

Abstract

The study of food webs and their dynamics is fundamental to understand how the feeding habits of the different species can affect the community, thus improving our understanding of the functioning of the ecosystem as a whole. Furthermore, the study of feeding ecology becomes crucial when it concerns endangered species since a precise knowledge of their diet is to be gathered when designing reliable conservation strategies. A wide range of methodologies have been proposed for diet analysis, including simple ones, as visual observation of foraging behavior, and more complex ones such as Near Infrared Reflectance Spectroscopy and DNA based methods.

DNA barcoding, i.e. species identification using a standardized DNA region or markers, has recently received much attention and is being further developed through an international initiative called "Consortium for the Barcode of Life". When using DNA barcoding for diet analysis, the choice of the markers is crucial. The ideal DNA barcoding marker should meet several criteria. It should be variable among species, standardized, with enough phylogenetic information, extremely robust, and short enough to allow amplification of degraded DNA.

In this study we propose the *trn*L (UAA) intron as marker for plant DNA barcoding. The power and the limitations of this system were evaluated as well as the possibility of species identification with highly degraded DNA. The main limitation of this system is its relatively low resolution in discriminating closely related species. Despite the relatively low resolution, it has many advantages: the primers are highly conserved, the amplification system is very robust and it is able to work with much degraded DNA samples. This system has been coupled with massively parallel pyrosequencing technique. We demonstrate the efficiency of this new approach by analyzing the diet of various herbivorous species. The whole chloroplast *trn*L (UAA) intron (254–767 bp) and a shorter fragment of this intron (the P6 loop, 10–143 bp) were used in this study. For the whole *trn*L intron 67.3% of the species retrieved from GenBank were unambiguously identified and 19.5% for the P6 loop. The resolution is much higher after calibration of specific contexts using species originating from the same ecosystem.

Furthermore, the *trn*L approach was coupled with individual and sex identification using microsatellites polymorphism in the Himalayan brown bear (*Ursus arctos isabellinus*). Among world brown bears populations; those in Asia are the most

endangered and least studied. Here, populations have declined by more than half in the past century owing to habitat loss and fragmentation and human activity. Presently in Pakistan brown bear occur sparsely in seven small populations, with the largest isolate in the Deosai National Park. We examined this population using a combination of fecal DNA analysis and field data for which geographical location and date of sampling were available, with the aim to study individual and sexual differentiation in the diet, and also temporal and geographical variations. Twenty-eight individuals (16 male, 10 females and 2 unknown sex) were identified in this study with microsatellites markers. Only eight plant species were found represented in more than 50% of individual feces. Temporal differences were found with more energetic food detected before the hibernation periods.

Resumé

L'étude des réseaux trophiques et leur dynamique est fondamentale pour comprendre comment les habitudes alimentaires des différentes espèces peuvent influencer la communauté, afin d'améliorer notre compréhension du fonctionnement de l'écosystème dans son ensemble. En outre, l'étude de l'écologie alimentaire devient cruciale lorsqu'il s'agit des espèces en voie de disparition, une connaissance précise de leur alimentation doit être acquise lors de la conception des stratégies de conservation. Un large éventail de méthodes a été proposé pour l'analyse du régime alimentaire, y compris les plus simples, comme l'observation visuelle du comportement d'alimentation, et les plus complexes, comme la spectrométrie dans le proche infrarouge et les méthodes basées sur l'ADN.

"DNA barcoding" (Code barre d'ADN), c'est-à-dire l'identification des espèces en utilisant une région standardisée d'ADN ou des marqueurs standardisés, a récemment reçu beaucoup d'attention et est actuellement développé grâce à une initiative internationale appelée "Consortium for The Barcoding of Life". Lors de l'utilisation du "DNA barcoding" pour le régime alimentaire, le choix des marqueurs est crucial. Le marqueur idéal pour le "DNA barcoding" doit satisfaire plusieurs critères. Il doit être variable entre les espèces, standardisé, avec suffisamment d'informations phylogénétiques, très robuste, et suffisamment court pour permettre l'amplification de l'ADN dégradé.

Dans cette étude, nous proposons l'intron *trn*L (UAA) en tant que marqueur pour le "DNA barcoding" des plantes. Le pouvoir et les limites de ce système ont été évalués, ainsi que la possibilité d'identification des espèces avec de l'ADN fortement dégradé. La principale limitation de ce système est la relative faible résolution de discrimination des espèces très proches. En dépit de la résolution relativement faible, elle présente de nombreux avantages: les amorces sont hautement conservées, le système d'amplification est très robuste et il est capable de travailler avec des échantillons d'ADN très dégradés. Ce système a été couplé avec la technique du pyroséquençage. Nous avons démontré l'efficacité de cette nouvelle approche par l'analyse de l'alimentation de différentes espèces herbivores. L'ensemble des introns chloroplastique *trn*L (UAA) (254-767 pb) et d'un court fragment de cet intron (P6 boucle, 10-143 pb) ont été utilisés dans cette étude. Pour l'ensemble de l'intron *trn*L, 67,3% des espèces récupérées à partir de GenBank ont été identifiées sans ambiguïté et 19,5% pour la P6 boucle. La résolution

est beaucoup plus élevée après calibration sur des contextes spécifiques en utilisant des espèces originaires d'un même écosystème.

En outre, l'approche par le *trn*L a été associée à l'identification individuelle et le sexe en utilisant le polymorphisme de microsatellites dans l'ours brun himalayen (*Ursus arctos isabellinus*). Parmi les populations mondiales d'ours bruns, celles d'Asie sont les plus menacées et les moins étudiées. Ici, les populations ont diminué de plus de moitié au cours du siècle dernier en raison de la perte d'habitats, de sa fragmentation et de l'activité humaine. Actuellement, au Pakistan l'ours brun existe dans sept petites populations isolées, dont la plus grande est située dans le Parc National du Deosai. Nous avons examiné cette population au moyen d'une combinaison de l'analyse d'ADN des fèces et de données de terrain pour lesquelles les coordonnées géographiques et la date de prélèvement étaient disponibles, pour étudier la différenciation sexuelle et individuelle du régime alimentaire, ainsi que les variations temporelles et géographiques. Vingt-huit individus (16 mâles, 10 femelles et 2 de sexe inconnu), ont été identifiés dans cette étude avec les marqueurs microsatellites. Seulement huit espèces de plantes ont été trouvées représentées dans plus de 50% des fèces des individus. Des différences temporelles ont été trouvées, avec une alimentation plus énergétique avant la période d'hibernation.

Riassunto

Lo studio delle reti trofiche e della loro dinamica è fondamentale per comprendere come le abitudini alimentari delle diverse specie possono incidere sulla comunità, migliorando in tal modo la nostra comprensione sul funzionamento dell'ecosistema nel suo complesso. Inoltre, lo studio dell'ecologia dell'alimentazione diventa cruciale quando si tratta di specie in via d'estinzione, nelle quali una precisa conoscenza della loro dieta deve essere acquisita per definire una strategia di conservazione di successo. Una vasta gamma di metodi é stata proposta per l'analisi della dieta, che vanno da quelli più semplici, come osservazione visiva dell'animale durante il pasto, a quelli più complessi, come la Near Infrared Reflectance Spectroscopy e i metodi basati sul DNA.

Il DNA barcoding, cioè l'identificazione di attraverso una regione standardizzata di DNA o attraverso marcatori, ha recentemente ricevuto molta attenzione e si è ulteriormente sviluppato attraverso un'iniziativa del consorzio internazionale denominato "Consortium for Barcoding of Life". Quando si usa il DNA barcoding per l'analisi della dieta, la scelta dei marcatori è cruciale. Il marcatore ideale per il DNA barcoding deve soddisfare diversi criteri. Deve essere variabile tra specie, standardizzato, avente una sufficiente informazione filogenetica, molto robusto, e abbastanza corto da consentire l'amplificazione di DNA degradato.

In questo studio si propone l'introne *trn*L (UAA) come marcatore per il DNA barcoding delle piante. I vantaggi e gli svantaggi di questo sistema sono stati valutati come pure la possibilità di identificare di specie da DNA molto degradato. Il limite principale di questo sistema è la sua relativamente bassa risoluzione in discriminare specie filogeneticamente molto simili. Nonostante la relativa bassa risoluzione, il sistema ha molti vantaggi: i primers sono molto conservati, il sistema di amplificazione è molto robusto ed è in grado di funzionare con campioni di DNA molto degradati. Questo sistema è stato accoppiato con la tecnica di pirosequenziamento parallelo. Abbiamo dimostrato l'efficacia di questo nuovo approccio analizzando la dieta di varie specie d'erbivori. L'intero introne *trn*L (UAA) del cloroplasto (254-767 pb) e un breve frammento di questo introne (il P6 loop, 10-143 pb) sono stati utilizzati in questo studio. Per l'intero introne *trn*L, il 67,3% delle specie, recuperate da GenBank, sono statie indentificate in modo inequivocabile e 19,5% per il P6 loop. La risoluzione è molto più elevata dopo la calibrazione in contesti specifici utilizzando specie originarie dello stesso ecosistema.

Inoltre, il *trn*L approach è stato condotto in parallelo con identificazione degli individui e del sesso dell'animale tramite microsatelliti nell'orso bruno imalaiano (*Ursus arctos isabellinus*). Tra le popolazioni d'orso bruno al mondo, quelle in Asia sono le più minacciate e meno studiate. Qui, le popolazioni sono diminuite di oltre la metà nel secolo passato, a causa della frammentazione, la perdita di habitat e le attività umane. Attualmente in Pakistan, l'orso bruno è presente in sette piccole popolazioni isolate, con la più grande nel Parco Nazionale del Deosai. Abbiamo esaminato questa popolazione utilizzando una combinazione d'analisi di DNA da feci e di dati di campo per i quali la localizzazione geografica e la data di campionamento erano disponibili, con l'obiettivo di studiare differenze nella dieta a livello individuale tra i due sessi, e anche variazioni geografiche e temporali. Ventotto individui (16 maschi, 10 femmine e 2 di sesso sconosciuto) sono stati identificati in questo studio con i marcatori microsatelliti. Solo otto specie di piante sono state trovate rappresentate per oltre il 50% degli individui. Differenze temporali sono state riscontrate, con un consumo di cibo più energtico prima del periodo del letargo.

List of papers

PAPER I

Eva Bellemain, Muhammad Ali Nawaz, Alice Valentini, Jon E. Swenson, Pierre Taberlet. 2007. **Genetic tracking of the brown bear in northern Pakistan and implications for conservation**. Biological Conservation 134: 537 –547.

PAPER II

Pierre Taberlet, Eric Coissac, François Pompanon, Ludovic Gielly, Christian Miquel, Alice Valentini, Thierry Vermat, Gérard Corthier, Christian Brochmann and Eske Willerslev. 2007. **Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding**. Nucleic Acids Research 35, No. 3 e14.

PAPER III

Alice Valentini, Christian Miquel, Muhammad Ali Nawaz, Eva Bellemain, Eric Coissac, François Pompanon, Ludovic Gielly, Corinne Cruaud, Giuseppe Nascetti, Patrick Winker, Jon E. Swenson, Pierre Taberlet. **New perspective in diet analysis based on DNA Barcoding and large scale pyrosequencing**. Molecular Ecology (submitted).

PAPER IV

Alice Valentini, François Pompanon, Pierre Taberlet. **DNA Barcoding for ecologists.** Trend in Ecology and Evolution (submitted).

Introduction

Diet analysis

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 sound conservation strategies (e.g. Marrero *et al.* 2004; Cristóbal-Azkarate & Arroyo-Rodrígez 2007).

Several methods have been developed to evaluate the composition of animal diets. The simplest approach is the direct observation of foraging behavior. 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 a complex environment, with many plant species that are not spatially separated. 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 anesthesia. Feces analysis represents an alternative, non-invasive, and attractive approach. Up to now, four main feces-based techniques have been used.

First, for herbivores, microscope examination of plant cuticle fragments in fecal samples has been the most widely employed technique (Holechek *et al.* 1982; McInnis *et al.* 1983). Some herbivores do not masticate their food into small fragments, allowing plants present in the feces to be identified visually (Dahle *et al.* 1998). However, this method is very tedious to perform, and requires a considerable amount of training while a variable proportion of plant fragments remains unidentifiable.

Another method is stable isotope analysis. This approach is based on the fact that stable isotopes ratios in tissue and feces are related to the organism diet (DeNiro & Epstein 1978, 1981). Stable carbon isotopes can distinguish marine from terrestrial dietary protein, and C_3 plants that fix CO_2 by Calvin cycle (most grasses, trees, roots, and tubers) from C_4 plants (such as maize), which use dicarboxylic acid pathway. Nitrogen isotopes can successfully distinguish plant from animal protein and thus define the trophic level and the position an organism occupies in the food chain (DeNiro & Epstein 1981). This method was used for inferring the diet of several species, including

black and brown bears (Hobson *et al.* 2000), red-backed voles (Sare *et al.* 2005), blue and black wildebeest (Codron & Brink 2007), etc. This method can be used as a simple tool for investigating the passage rate of plants in the digestive track (Sponheimer *et al.* 2003). The main advantage of this technique is that diet can be also inferred from hairs and bones, and surveys a very large time span. This method can be used to infer diet of ancient remains (Feranec & MacFadden 2000) or mummies (Wilson *et al.* 2007). The main disadvantage is that is not possible to perform identification at species level.

The third 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, with the odd-numbered molecules largely predominating the even-numbered ones. There are marked differences in alkane composition and concentrations among plant taxa (families, genera, species), and thus the alkane fingerprints represent another chemical approach for estimating the species composition. This method is very common for study ruminant diet (e.g. Ferreira *et al.* 2007a, b; Piasentier *et al.* 2007). However, the approach is limited when the animal feeds in complex environment, because composition and concentration of the chemical markers are confounded. In this case it may be extremely difficult or impossible to have a discrimination of the eaten species (Dove & Mayes 1996).

The fourth approach is the Near Infrared Reflectance Spectroscopy (NIRS) (e.g. Foley *et al.* 1998; Kaneko & Lawler 2006). Near infrared spectra depend on the number and type of H chemical bonds (C-H, N-H and O-H) present in the material being analyzed. 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, fiber, starch, etc. However this technique has several limitations. Particle size variation and non 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.

A quite recent approach is DNA barcoding.

DNA Barcoding (Paper IV)

The term DNA barcoding is of recent use in the literature (Floyd *et al.* 2002; Hebert *et al.* 2003). It relies on the use of a standardized DNA region as a tag for rapid, accurate

and automatable species identification (Hebert & Gregory 2005). However, DNA barcoding is not a new concept. The term "DNA barcodes" was first used in 1993 (Arnot *et al.* 1993) in a paper that did not receive very much attention from the scientific community. Actually, the concept of species identification using molecular tools is even older, and came before the invention of the Sanger sequencing technique (Sanger *et al.* 1977). However, the “gold age” of DNA barcoding began in 2003 (Hebert *et al.* 2003) and the number of publications on the subject has grown exponentially, with now more than 250 articles published.

The now well-established Consortium for the Barcode of Life (CBOL; http://barcoding.si.edu/), an international initiative supporting the development of DNA barcoding, aims to promote global standards and to coordinate research in DNA barcoding. For animal, the gene region that is proposed as the standard barcode is a 650 base-pair region in the mitochondrial (mt) cytochrome c oxidase 1 gene (“COI”) (Hebert *et al.* 2003). For plants, the situation is still controversial, but recently it has been proposed to use three coding chloroplast DNA regions that together would represent the standard barcode: *rpoC1*, *matK*, and either *rpoB* or *psbA-trnH* (Chase *et al.* 2007).

As pointed out by Chase *et al.* (2005), taxonomists are not the only potential users of DNA barcode, since it may be helpful for scientists from other fields (e.g. forensic science, biotechnology and food industry, animal diet). Taxonomists are concerned in DNA barcoding “*sensu stricto*". Other scientists will be more interested in DNA barcoding “*sensu lato*” i.e. by DNA-based taxon identification using diverse techniques than can lies outside the CBOL approach (such as RFLP, AFLP, SSCP, etc). The difference between the two approaches mainly relies on different priorities given to the criteria used for the choice of the ideal barcoding system. It should be sufficiently variable to discriminate among all species, but conserved enough to be less variable within than between species; it should be standardized with the same DNA region used for different taxonomic groups; the target DNA region should contain enough phylogenetic information to easily assign species to its taxonomic group (genus, family, etc.); it should be extremely robust, with highly conserved priming sites, and highly reliable DNA amplification and sequencing. This is particularly important when using environmental samples where each extract contains a mixture of many species to be identified at the same time. The target DNA region should be short enough to allow

amplification of degraded DNA as usually DNA regions longer than 150 bp are difficult to amplify from degraded DNA.

Thus, the ideal DNA marker should be variable, standardized, with enough phylogenetic information, extremely robust and short. Unfortunately, such an ideal DNA marker does not exist (or at least has not been found up to now). As a consequence, according to the scientific and technical context, the different categories of users (e.g., taxonomists, ecologists, etc.) will not give the same priority to the five criteria listed above. Taxonomists are more interested in standardized markers that express a high level of variation with sufficient phylogenetic information, following the CBOL strategy, while other scientists may favor highly robust procedures even if the identification to species level is not always possible.

But when an ecologist needs to use DNA barcoding for species identification? When the use of non-invasive samples is necessary (i. e. only traces of the organism are present, the animal should not be disturbed or the species is endangered) or when the species identification is not possible or easy on morphological criteria. Barcoding has the advantage that it can be used as a non-invasive technique. It will be useful as tool when only traces of an organism are present in nature, for example (Valiere & Taberlet 2000) utilized mtDNA control region for identifying species (in their case wolf and dog) from urine traces left by the animals on the snow.

The study of endangered species is one of the central topics of most of the ecologists, and in this case the use of non-invasive molecular tools can be vital for the species studied. In some case the capture of the animal can lead to the injury or the death of it, and in the cases of endangered species this loss will have a huge cost for the species. In their article (Sugimoto *et al.* 2006) describe a non-invasive technique to identify two endangered species that live in sympatry from their feces samples: Amur leopard *Panthera pardus orientalis*, the world most endangered species of leopard, and Siberian tiger *Panthera tigris altaica*.

DNA Barcoding became fundamental when the species identification is not possible or easy on morphological criteria. In many cases the species cannot be identified in all life stages or only one sex has the keys characters for the identification.

In **Paper IV** we review some studies that have applied DNA barcoding from an ecological point of view.

Diet analysis using DNA barcoding.

DNA barcoding is a very useful tool to establish the diet of an individual from its feces or stomach contents. This is really helpful when the food is not identifiable by morphological criteria, such as in liquid feeders like spiders (Augusti *et al.* 2003). This technique also provides valuable information when eating behavior is not directly observable, as in the case of krill eating diatoms (Passmore *et al.* 2006), giant squid (*Architeuthis* sp.) in the sea abyss (Deagle *et al.* 2005), or deep sea invertebrates . Most of the studies that use DNA markers for diet analysis are based on carnivorous animals (e.g., insects (Pons 2006, Symondson 2002), whale and Adelie penguin (*Pygoscelis adeliae*) (Jarman *et al.* 2004)). Fewer studies were carried on herbivorous animals (e.g. Bradley *et al.* 2007). DNA barcoding approach was also successfully applied to study the diet from ancient coprolites (Hofreiter *et al.* 2000) and human mummies (Rollo *et al.* 2002; Poinar *et al.* 2001).

There are two different strategies when using molecular tools for diet analysis: the use of group-specific primers (Nystrom *et al.* 2006) or the use of universal primer. When analyzing the diet of the Macaroni penguin (*Eudyptes chrysolophus*) using feces as a source of DNA, Deagle *et al.* 2007 applied both group-specific and universal primers. The results obtained with five different groups of specific primers were similar to those involving universal (for fish, cephalopods and crustaceans) 16S rDNA primers and subsequent cloning of the PCR products. In general, the use of specific primers requires an *a priori* knowledge of the animal's diet. This is not possible in most cases and makes the "universal" approach more appropriate.

Himalayan brown bear

Brown bear is one of the eight different species of bears in the word, and it is widely distributed on the northern hemisphere, and it is found in Europe, North America and Asia. In Asia the brown bear (*Ursus arctos*) is widely distributed from the tundra and boreal forests of Russia in the north to the Himalayas in the south (Servheen 1990). Among world brown bears populations, those in Asia are the most endangered and least studied. Here, populations have declined by more than half in the past century (Servheen, 1990; Servheen *et al*., 1999).

Himalayan brown bear (*Ursus arctos isabellinus*) is a subspecies of brown bear distributed in small populations in Afghanistan, China, India, Kazakhstan, Kirghizstan, Nepal, Uzbekistan, Pakistan, and Tajikistan. This bear subspecies is threatened with

extinction and for this is listed in the Appendix I of CITES (Convention on International Trade in Endangered Species of Wild Fauna and Flora). Historically, it occupied the western Himalaya, the Karakoram, the Hindu Kush, the Pamir, the western Kunlun Shan and the Tian Shan range in southern Asia, but today its geographical distribution has been strongly reduced, compared with its historically range. In Pakistan brown bear are found in sub-alpine and alpine areas (2600-5000m) and its primary habitat are alpine meadows (51,000 km²) and blue pine forest (19,000 km²) (Nawaz 2007). Approximately 150-200 bears survive in seven isolated (or with limited connections) populations, Himalayan, Karakoram, Hindu Kush, Kalam, Indus Kohistan, Kaghan, Neelam Valley. Himalayan, Karakoram, Hindu Kush, and Neelam Valley are divided in sub-population each. Deosai National Park, Minimerg and Nanga Parbat are sub-populations of Himalayan population. Karakoram host Central Karakoram National Park and Khunjerab National Park, a Hindu Kush host Ghizer, Karambar, Tirch Mir sub-populations. Gumot, Shontar Valley and Gurez Valley are the subpopulations of Neelam Valley (Nawaz 2007). Except for the Deosai National Park subpopulation, that is increasing (Nawaz *et al.* unpubblish), all the subpopulations and populations are declining and they have a very small size, with only 5 bears recorded in some cases. The Deosai National Park supports the largest population of brown bears in the country (with 40-50 bears recorded (**Paper I** and Nawaz 2007). The brown bear population in this park has been protected and closely monitored since 1993, when bear population was composed only by 19 individual, after that the population started to recover gradually (Himalayan Wildlife foundation 1999a)

Brown bears in Pakistan are declining for habitat loss and fragmentation, and human activity, which include commercial poaching of cubs and body parts, bear baiting and hunting (Nawaz 2007). The most used habitat for brown bear are alpine meadows in Northern Area of Pakistan, but those areas are now used as grazing areas due to the expansion of nomadic and transhumance grazing because of the deficiency of natural grazing areas after the nearly doubling of livestock population (Ehlers & Kreutzmann 2000). Bear in the region is hunted, and poached in protected areas, such as Deosai National Park, for sport (mostly by military officers), by villagers, that feel brown bears as a danger for their livestock, and for commercial purpose (Nawaz 2007). Climate change will influence brown bear population, by the reduction of alpine tundra, and a northward and upward shift of coniferous biome (Hagler Bailly Pakistan 1999).

The Himalayan brown bears are mainly vegetarian with very low dietary meat (Nawaz *et al.* in preparation), and this characteristic gave it the name of *spang drenmo* (vegetarian bear) in the Balti language (the dialect of the Northern Area region), for distinguishing it from Asiatic black bear, *shai drenmo* (carnivorous bear) (Nawaz 2007). The Deosai population has a very low reproductive capacity, with smaller litter size and longer maternal care than others brown bears populations (Nawaz *et al.* unpublished), probably due to its diet. In fact it was demonstrated that the reproductive success in bear is linked to the amount of meat ingested (Bunnel & Tait 1981, Hilderbrand *et al.* 1999). Due to its particular diet Deosai brown bears spend most of its daily activity foraging (67%, mainly grazing) (Nawaz & Kok 2004). Therefore, the study of its diet will be fundamental for assessing good conservational plans for this population.

Objectives of the thesis

In this thesis we will use molecular tools for better understand the biology and ecology of Himalayan brown bear. The thesis has two main objectives.

1. Determinate the genetic status and the size of the brown bear population in Deosai National Park (**Paper I**).
2. supply a new tool for assessing the diet of the brown bear population

To achieve the second objective the DNA barcoding approach is proposed. We review the possible applications of this approach for ecologists (**Paper IV**). After, we propose a new system for plant barcoding (**Paper II**). Finally we test this approach for study the diet of different herbivorous animals (**Paper III**)

Material and Methods

The study area

The study for population analysis was conducted in the Deosai National Park, Northern Areas, Pakistan. Deosai National Park is a plateau in the alpine ecological zone encompassing about 20,000 km², situated 30 km south of Skardu and 80 km east of the Nanga Parbat Peak in Pakistan. Elevations range from 3500 to 5200m and about 60% of the area lies between 4000 and 4500 m. The Deosai Plateau is situated between two of the world's major mountain ranges, the Karakoram and Himalaya. The area receives abundant snow fall and rain, with annual precipitation in Deosai in the range of 510–750 mm, which falls mostly as snow (Himalayan Wildlife Foundation, 1999b). Water percolates in the soil and emerges during spring along ravines and in open grassy valley. Where water emerges, the areas are covered by deep grassland and numerous flowering plants. Recorded mean daily temperatures range from -20 C° to 12 C°. The Deosai plains are covered by snow during winter months between November and May, and life on the plateau is confined to a window of five months.

Different habitats are present in Deosai Plateau: open sunny sites, rock slops, steppes and marshy places. The flora can be divided in three categories: weeds, desert type native plants and high alpine plants. The firsts are found close to cultivated fields, the seconds on cliff, sandy soil and on the streams and last type are found near melting snow and glaciers along moraines. The plants most represented are herbs and small shrubs. The only trees are birches, junipers and conifers, and they occur in the valley of the lower limits of Deosai, but they are very rare. The high elevation and the strong wind prevent the growth of trees on higher areas of the plateau. The plants are generally dwarf and tufted, owing to severe wind and frost, and are perennials, having a brief growing period. (Woods *et al.*1997).

The biota includes plants and animals from Karakoram, Himalaya and Indus Valley. As a result, Deosai is a centre of unique biota in northern Pakistan. The documented biota of Deosai National Park includes 342 species of plants, 18 of mammals, 208 of birds, three of fishes, one of amphibian, and two of reptiles (Woods *et al.* 1997).

Genetic methods

Non-invasive genotyping of brown bears in Deosai National Park

One hundred thirty six feces were collected in the field and used as the source of DNA. All samples were preserved in 95% alcohol until extraction. The extraction was performed using Qiamp DNA Stool Kit (Qiagen GmbH, Hilden, Germany). This study was divided in two parts: the first one was focused on the identification of the differents individuals from feces samples and the second was focused on the population genetic study. For individual identification six microsatellites loci were analysed. The number of loci studied is a compromise between the probability of identity and the probability of genotyping errors. As the number of loci increase, the probability of identity decrease, but the genotyping error rate increases (Pompanon *et al*. 2005). Because of the poor DNA conditions, we decided to follow the protocol already successfully used for brown bear individual identification from feces samples in the Scandinavian population (Piggott *et al.* 2004, Bellemain & Taberlet 2004). Four primer pairs were already described in Bellemain & Taberlet (2004) (Mu23, Mu50, Mu51, and Mu59) and 2 microsatellite primer pairs were specially designed for this study (G10H, G10J, from Paetkau & Strobeck (1994) and Paekau *et al.* (1995)), in order to obtain a probability of identity low enough for discriminate among individuals. For sex identification the SRY-primers (Bellemain & Taberlet 2004) were used. Amplification was carried using the protocol described in Taberlet *et al*. 1996. Quality index (Miquel *et al.* 2006) was calculated for each sample. Only the samples with a quality index above 0.5 were retained for the population genetic analysis. For this second part of the study the number of loci analysed was increased, because the probability of error rate was reduced after the selection based the quality index. Others 12 loci were added: G1A, G1D, G10B, G10C, G10L, G10P, G10X, G10O (Paetkau *et al.* 1995; Paetkau & Strobeck 1994) and Mu05, Mu10, Mu15, Mu61 (Taberlet *et al.* 1997). The amplifications were carried out using a modified protocol from Waits *at al.* 2000. One primer of each pair was synthesized with a fluorescent dye group (6-FAM, TET or HEX) on the 5' end to allow detection and sizing of fragments on ABI Prism 3100 automatic sequencer. The gels were analyzed with GeneMapper version 3.0 when using the ABI Prism 3100. A new quality index for this second analysis was calculated and 3 microsatellites loci were discarded, because their QI was below 0.6.

The probabilities of identity, i.e. the probability to obtain two identical genotypes by chance, (PI; Paetkau & Strobeck (1994); PIsib, for siblings; Waits *et al.* (2001)) were low: PI=1.881e-05 and PIsib=1.206e-02 for the 6 microsatellites loci set, 5.827e-10 and 1.329e-04 respectively for the 15 microsatellite loci set. This allowed to perform reliable relatedness analysis.

Population size was estimated, as in Bellemain *et al.* (2005), using two different rarefaction indices, the one proposed by Kohn *et al*. 1999 and the one proposed by Eggert *et al.* 2003. In the Kohn methods the population size is estimated as the asymptote of the relationship between the cumulative number of unique genotype and the number of samples typed. The estimates are made using the equation y=ax/(b+x), where *a* is the asymptote, *x* the number of feces sampled, *y* the number of unique genotypes, and *b* the rate of decline in the value of slope. The Eggert methods is based on the equation $y = a(1 - e^{bx})$. The small sample size and small number of recaptures not allow performing the analysis with the MARK method (White & Burnham 1999), which was suggested for population size analysis from fecal samples (Bellemain *et al.* 2005). The genetic diversity was calculated for this population and compared with other brown bear populations in Europe and North America (Taberlet *et al.* 1997; Peatkau *et al.* 1998; Waits *et al.* 1998; Waits *et al.* 2000). Hardy-Weinberg equilibrium, linkage disequilibrium were analyzed in the Deosai bear population, based on the 15 loci genotypes, we ran population genetic analyses using the software GENEPOP version 3.4 (Raymond and Rousset, 1995) and GENETIX version 4.02 (Belkhir *et al.*, 1996–2004). For detect a signature of bottleneck and date of this potential bottleneck in the Deosai bear population we used a bayesian approach, implemented in the MSVAR program (Beaumont, 1999).

Diet analysis

The diet study using DNA barcoding from fecal samples coupling universal primer approach and next generation sequencing technique was first implemented in this thesis (**Papers II** and **Paper III**).

The next-generation sequencing systems

Recently several new techniques were implemented, all based on a massively parallel approach, and sequencing individual molecules (with or without an amplification step) (e. g. Solexa™, SOLiD™ DNA Sequencer, Heliscope™, 454 GS FLX™, but see Box 3

in **Paper IV**). All new sequencers but one produce very short fragments (25-35 bp). The only system that allows sequencing longer fragments is the 454 GS FLX (Roche) that currently deliver 200-300 bp fragments (an upgrade of the system is already announced, multiplying by about ten the total output, with fragments of 400 bp). This new method is a combination of an emulsion-based method to isolate and amplify DNA fragments, and pyrosequencing in picolitre-sized wells. Single strand DNA is generated by fragmentation of the genome, or amplification by PCR. Subsequently each fragment is capture on its own beads and, within the droplets of an emulsion, clonally amplified. This part is defined as emulsion PCR (emPCR). Once the clonally reaction had finish the emulsion is broken, the DNA strands are denatured, and beads carrying single-stranded DNA clones are loaded into wells of a fiber-optic slide with beads carrying immobilized enzymes required for sequencing. The slide is loaded in the sequencer and cyclically nucleotides in a fixed order (TACG) flow perpendicularly and simultaneously to all the wells. If the nucleotide that flow in the well is complementary to the template strand is added to the strand generating a chemiluminescent signal that is recorded by the CCD camera in the instruments. The intensity of the signal is proportional at the number of nucleotides added to the strand. The results are shown in a flowgram that gives the sequence of the fragments analyzed (Margulies *et al.* 2005). Using this method c.a. 400,000 sequences are obtained per run. The enormous amount of sequences that are product without cloning step make this new technique suitable for environmental barcoding studies where there is the need to deal with samples composed by mixed species (e.g. of deep sea biodiversity (Sogin *et al.* 2006)).

Test of primer universality

We have chosen to amplify and sequence *trn*L locus on chloroplast for several reason. Universal primers for this region were designed more than15 years ago (Taberlet *et al.* 1991), and subsequently extensively used, mainly in phylogenetic studies among closely related genera and species (Gielly & Taberlet 1996). The evolution of the *trn*L (UAA) intron has been thoroughly analyzed and is well understood (Quandt & Stech 2005; Quandt *et al.* 2004). Furthermore, this region has an alternation of conserved and variable regions (Quandt *et al.* 2004), as a consequence, new versatile primers that will amplify can be easily designed in the conserved region.

The power and the robustness of the *trn*L intron for DNA barcoding were first evaluated with the data available in public sequences databases. PCR were simulated on

the full plant division found on GenBank download from NCBI server on December 14, 2005 (ftp://www.ncbi.nlm.nih.gov/genbank), that correspond to 731,531 entries, using ePCR (electronic PCR) software, specifically developed (**Paper II**). This software allows use very short sequences as a query, to specify maximum mismatch count, minimum and maximum length of the amplified region and takes care also to retrieve taxonomic data from the analyzed entries. ePCR was applied on GenBank data, first with the *c* and *d* primers (Taberlet *et al.* 1991) that amplify the entire intron, then with the primers designed for the amplification of a shorter internal region of this intron, the P6 loop (*g* and *h*, Figure 1 and **Paper II**), then on a short *rbc*L fragment with the *h1aF* and *h2aR* primers, used in study with very degraded DNA (Poinar *et a.l* 1998), and finally with eight primer pairs found in Shaw *et al.* 2005 (*psb*B-*psb*H, *rpo*B-*trn*C (GCA), *rp*S16 intron, *trn*D (GUC)-*trn*T (GGU), *trn*H (GUG)-*psb*A and *trn*S (UGA)-*trn*fM (CAU)), that were previously suggested for phylogenetic studies on plants.

The power and the robustness were evaluated also using two specific datasets. The first dataset was implemented by sequencing the whole intron (using *c* and *d* primers) for 132 artic plant samples (GenBank accession numbers DQ860511- DQ860642). The second one was build retrieving the sequences of the 72 plants used in the food industry from GenBank, and sequencing 7 plants species (cacao, beet, strawberry, apricot, sour cherry, garden pea, potato) (GenBank accession numbers EF010967- EF010973). The universality of the four primers *c*, *d*, *g* and *h* was examined by comparing their sequences with homologous sequences, either from GenBank (for primers *c*, *d*, *g* and *h*) or produced in this study (for primers *g* and *h*). Finally, the robustness of the new pair of internal primers (*g* and *h*) was tested applying this approach to different substrates supposed to contain highly degraded DNA: processed food (four samples: brown sugar from sugar cane, cooked potatoes, cooked pasta and lyophilized potage), human feces (two samples) and permafrost samples (four samples).

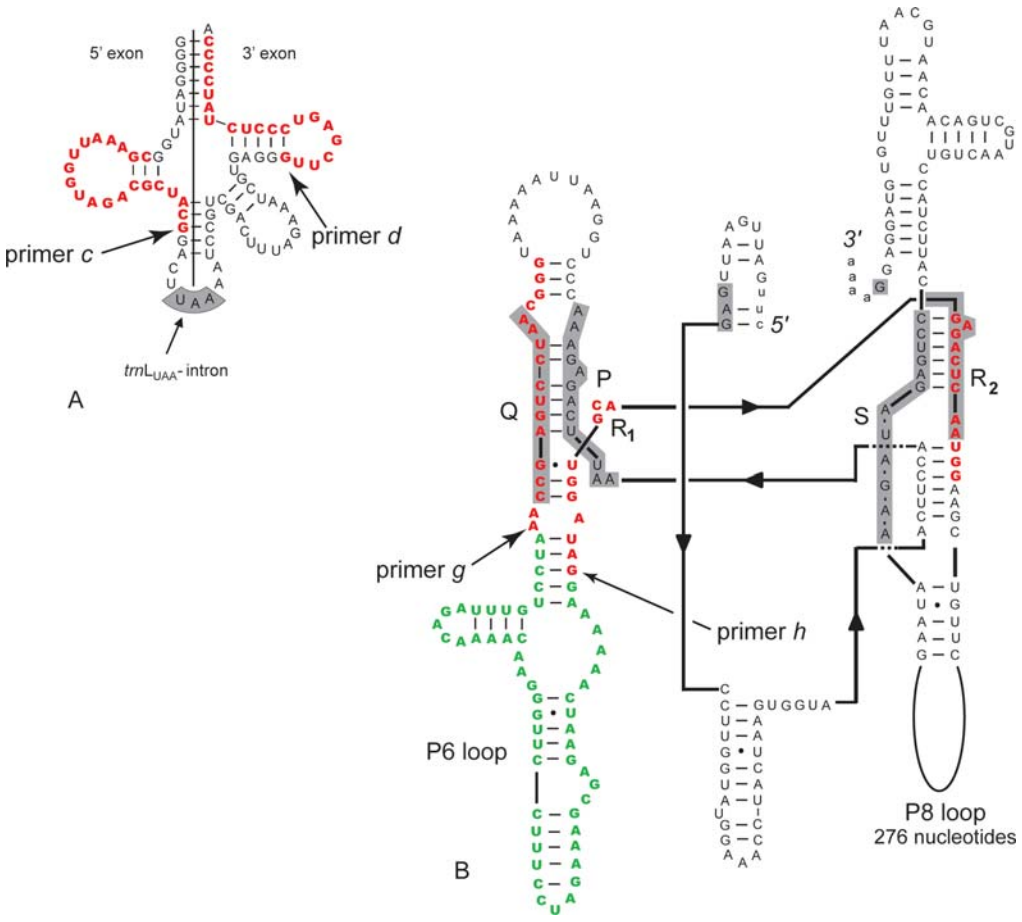


Figure 1 Positions of the primers *c* and *d* on the secondary structure of the *trn*L (UAA) exon (**A**) and of the primers *g* and *h* on the secondary structure of the *trn*L (UAA) intron (**B**) for *Nymphaea odorata* (modified from Borsch *et al.* (2003)).

The *trn*L approach: primer universality and parallel pyrosequencing for diet analysis

In this second part of the diet analysis we coupled the universal amplification method using the P6 loop of the chloroplast *trn*L (UAA) intron (as described in **Paper II)** with the new highly parallel sequencing systems (Margulies *et al.* 2005) for describe a universal method for diet analysis of herbivorous animals. A total of 36 feces samples, form different herbivorous species (mammals, birds, mollusks and insects), were collected for the analysis as described in **Paper III**. For mammals, we sampled 12 feces from golden marmots (*Marmota longicauda*) in the Deosai National Park (Pakistan), with no more than one feces per marmot colony. We also analyzed 12 faeces from brown bears (*Ursus arctos*) collected in the same area, and previously used for genotyping analysis in **Paper I**. For birds, we used six capercaillie (*Tetrao urogallus*) samples previously analyzed in Duriez *et al.* (2007), four from the French Pyrenees (*T.*

u. aquitanus) and two from the Corinthian Alps in Austria (*T. u. major*). For the invertebrates, we collected three grasshopper feces (two from *Chorthippus biguttulus*, one male and one female, and one from *Gomphocerippus rufus*) and three mollusc faeces (from the snail *Helix aspersa*, and from the slugs *Deroceras reticulatum* and *Arion ater*). DNA from feces was extracted using DNeasy Tissue Kit (Qiagen GmbH, Hilden, Germany), and not Qiagen Stool Kit, as in **Paper I**, because during a pilot experiment, we noticed that samples extracted with this kit 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".

In order to more precisely assess the diet of brown bears and golden marmots in Deosai National Park, a sequences database specific of the plants of this environment was constructed. Leaves of 91 plant species, that represent the most common species in Deosai plains, 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 (Taberlet *et al.* 1991).

The amplification from fecal samples was carried using universal primers for plants (*g* and *h*) using a modified PCR protocol, were the elongation step was removed for avoiding +A artefact (Brownstein *et al.* 1996; Magnuson *et al.* 1996). Each sample was amplified with primers *g* and *h* (**Paper II**), modified by the addition of a specific tag (5'-CCNNNN-3') on the 5' end in order to allow the recognition of the sequences after the pyrosequencing. The first two base of the tag were added, because previous study on 454 tagging system had shown an overrepresentation of sequence with 5'-CN tags (Binladen *et al.* 2007). Large-scale pyrosequencing was carried out on the 454 life sciences® technique (Margulies *et al.* 2005) following manufacturer's instructions, and using the GS 20 (Roche, Basel, Switzerland) for marmot and bear, and the GS FLX (Roche, Basel, Switzerland) for other samples. The plant taxa were then identified by comparing the sequences obtained either with public databases (GenBank, EMBL, etc.), using MEGABLAST algorithm (Zhang *et al.* 2000), and/or with a database made for this purpose.

The *trn*L approach applied to Himalayan brown bear

For brown bears all samples that were successfully amplified in **Paper I** (63 samples) were analysed using the *trn*L approach, Of the total of brown bear samples 12 samples were used for describing the *trn*L approach applied to diet analysis in **Paper III**.

Difference at sample and individual level in diet was tested using a correspondence analysis (Benzécri 1973) with the function dudi.coa implemented in ade4 package in R software version 2.4.0 (http://www.r-project.org/) using plants species, families and groups as variables. Plants were divided in five different groups: Graminoid, Forbs, Fruits, Browse, Tree and Other plants.

For investigating sex preferences in major diet groups, we arranged data into sex x presence/absence x k table, with k= number of plant regular species. Regular plant species (with ≥10% overall frequency) in diet were set as Z variable. Table analysis was run using PROC FREQ in SAS and Breslow-Day statistics was computed to determine if there was a homogenous relationship among sexes. We also computed Cohran-Mantel-Haenzel (CMH) statistics to investigate conditional independence between X and Y at each level of Z (Agresti 1996).

Data were grouped according to months of sampling (July through September) to determine the temporal trend in diet selection. We had too few samples for October, which were included in September. Locations of fecal samples were plotted on a vegetation map in Arc GIS (ESRI Inc., 2006) to determine their habitat types (marshy, grassy, stony, rocky, valley). Habitat differences in diet contents were investigated counting the number of species and families in each group.

Main Results and Discussion

The genetic status of brown bear population in Deosai National Park (Paper I)

Totally, 136 fecal samples were collected and 63 (46%) of those samples were successfully amplified for 4–7 loci (including the SRY sex locus), and 28 individual genotypes were obtained (16 males, 10 females and 2 individuals of unknown sex). The amplification success was correlated negatively with the age of fecal samples. Amplification success was relatively good (58%) for feces that were less than 2–3 days old while samples older than one week had a poor amplification success.

Population size estimates provided by the two rarefaction indices are in the same order of magnitude as the numbers derived from field censuses, which gives us confidence that those results are realistic (Kohn's estimate of 47 bears (95% CI: 33–102), Eggert's estimate of 32 bears (95% CI: 28–58), and 38 bears from the census 2004 (Nawaz *et al.* 2006)). Usually field methods give underestimates of wild populations, particularly for elusive animals (Solberg *et al.* 2006), but in our case several factors contributed to the realistic estimates of visual census: the open terrain of the Deosai plateau, the small population size, the presence of distinctive marks on many bears, and the expertise of the field staff. We conclude that approximately 40–50 bears were present in the park in 2004.

The results from the bottleneck analysis suggested that a decline in the Deosai population occurred approximately 80–100 generations ago. This period approximately corresponds to 800– 1000 years ago, assuming a generation time of 10 years. The ancestral population (before the decline; N1) was estimated to 10,000–12,500 individuals, which gives a density of about 55 bears per 1000 km². These results are consistent with previous knowledge on ancient population distributions (Nawaz 2007). The 200–300-fold decrease during the last thousand years was probably due to both natural (climatic and geological) and socio-political factors, such as the ‘‘little ice age’’ (1180–1840 AD; Kuhle 1997; Esper *et al.* 2002; Mackay *et al.* 2005), the influence of growing human population, the large deforestation in the Middle Ages (Bertrand *et al.* 2002), political unrest, and the spread of firearms in the late 19th century.

The population genetics analyses revealed that the level of nuclear genetic diversity of the Deosai population is globally lower than brown bear populations

considered to have a good conservation status, such those in Scandinavia or North America (Taberlet *et al.* 1997; Peatkau *et al.* 1998; Waits *et al.* 1998; Waits *et al.* 2001). However, this population is in Hardy Weinberg equilibrium and its level of relatedness is similar to that of the Scandinavian brown bear population. Therefore, the Deosai bear population does not appear to be at immediate risk of inbreeding depression. Its level of genetic diversity is comparable to the brown bear population in the Yellowstone area, USA, which has become an isolated remnant, separated from other brown bears for nearly a century (Paetkau *et al.* 1998). Furthermore four individuals in our genetic dataset showed private alleles at two different loci, suggesting that they could be migrants (or descendants from migrants) from outside of the study area. This result was supported by field observations.

Reliability of the trn*L approach for barcoding (Paper II)*

Via ePCR with primers *c* and *d* we retrieved 1308 sequences from GenBank, corresponding to 706 species, 366 genera and 119 families (excluding all sequences with at least one ambiguous nucleotide, and excluding genera with a single species and families with single genera). With primers *g* and *h*, we retrieved 18 200 sequences, corresponding to 11 404 species, 4215 genera and 410 families. The *c–d* primer pairs had a much lower number of hits because recorded sequences in GenBank often do not contain both primer sequences.

Globally, on the GenBank dataset, the entire *trn*L (UAA) intron and the P6 loop allow the identification of 67.3% (for *c-d* primer pair) and 19.5% (for *g-h* primer pair) of the species without taking into account single species within a genus. However, these values are probably underestimates, because of the possibility of misidentification of the species whose sequence has been submitted in public databases (Harris 2003). When analyzing the artic plant dataset the species identification was possible for the 85.44 % of the cases with *c-d* primers and 47.17% for *g-h*. For the food dataset analyzed with *g-h* primers the species were identified in 77.78% of the cases. The ePCR using other primer pairs found in Shaw *et al.* 2005 never retrieved more than 100 sequences, and were not taken into account in the present study. For *rbc*L fragment amplified using *h1a*F and *h2a*R, the resolution at species level is even lower, with only 15% of species identified.

It is clear that the *trn*L intron does not identify all plant species and cannot distinguish among closely related species, but this limitation is compensated by several

advantages. First, the primers used to amplify both the entire region (*c* and *d*) and the P6 loop (*g* and *h*) are extremely well conserved, from Bryophytes to Angiosperms for the first primer pair, from Gymnosperms to Angiosperms for the second one. The primers *g* and *h* are much more conserved than primers *h1a*F and *h2a*R (Poinar *et al.* 1998) which target a protein sequence and thus have much more variable positions. This advantage is particularly important when amplifying multiple species within the same PCR. Second, the number of *trn*L (UAA) intron sequences available in databases is already very high (more than 15,000 sequences), by far the most numerous among non-coding chloroplast DNA sequences, allowing in many cases the identification of the species or the genus. Finally, the robustness of both systems also represents an important advantage, allowing the standardization and automation of the system. In many situations, the number of possible plant species is restricted, reducing the impact of the relatively low resolution, as in the case of arctic plant dataset.

The amplification of much degraded samples as processed food or permafrost sample that was between 21050 and 25440 years old was possible. So P6 loop has the potential to be extensively used in food industry, in forensic science, in diet studies based on feces, and in permafrost analyses for reconstructing past plant communities.

Diet analysis (Paper III)

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

For the analysis of the 36 feces, we obtained a total of 97,737 P6 loop sequences, corresponding to an average of 2,715 ± 1130 sequences per sample. In each sample, a few sequences were found hundreds of time, whereas some other sequences were only represented either once or by very few occurrences. The sequences showing only up to three times were not taken into account in the subsequent analysis. They were almost always very close to a highly represented sequence, and thus we considered to be the result of sequencing errors in the P6 loop. 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.

Plants were identified at several taxonomic levels from the species to the family, with a different rate for the different animal species. The percentage of discrimination is presented on Table 1. When a specific database was used the discrimination at species and genus level became more precise, such as for marmots were 64% of the sequences were unambiguously assigned at species level.

Table 1. Percentage of identified plants per level of identification per animal species studied

Level of identification	I1	I2	M1	M2	M3	B1	B2	Ma1	Ma2
Species	-	0.33	-	-	-	-	-	0.31	0.64
Genus	0.40	0.67	-	-	0.40	0.67	0.75	0.51	0.77
Tribe	0.80	1	-	0.67	0.80	0.67	0.75	0.59	0.82
Subfamily	0.80	1	-	1	0.80	0.67	0.75	0.90	0.89
Family	1	1	1	1	1	1	1	1	1

I1= *Chorthippus biguttulus,* I2= *Gomphocerus rufus* , M1= *Helix aspera,* M2= *Deroceras reticulatum,* M3= *Arion ater,* B1= *Tetrao urogallus aquitanus,* B2= *Tetrao urogallus major,* Ma1= *Ursus arctos,* Ma2= *Marmota caudata*

All the results are consistent with the known diet of the animals, particularly for capercaillie, which eat mainly conifers in winter, and grasshoppers, which eat mainly grasses. The second slug was sampled in a compost box containing known plants, which were correctly identified by their sequences (C. Miquel, personal communication).

Mammals' diet is more complex than that of birds and invertebrates, where a maximum 5 plant taxa were found. For bears an average of 5.6 different plant taxa were found in each scat (2-10). In the bear feces a total of 14 different plant families were identified.. For marmots an average of 16.5 different plant taxa were found in each scat (7-21) and 20 different plant families were identified. The results obtained in marmots show clearly that the system is particularly well adapted for analyzing complex situations, where the diet is composed of many different species. Blumstein & Foggin (1997) analyzed the diet of golden marmot population in Dhee Shar (Northern Pakistan). Identification of plants was carried out using feces microhistological analysis,

unfortunately species identification from feces was not possible, so plants were divided in 4 broad groups: graminoids (7 species), legumes (6 species), scrubs (8 species) and other herbs (71 species). The Authors demonstrated that marmot have forage preferences for legumes, that composed 71% of their diet. The *trn*L approach applied to golden marmot population in the Deosai National Park allowed us to describe their diet more precisely (arriving to identify 64% of the plant species eaten). In total 58 plant taxa were detected in marmots diet, of those 11 were present in more than 50% of the samples. Those plants belong to six different families (Asteraceae, Caryophyllaceae, Fabaceae, Lamiaceae, Poaceae, Polygonaceae). The most eaten species by marmots is *Cerastium cerastoides*, in fact it was found in 10 samples of 12 (Table 2). The results presented in **Paper III** demonstrate that golden marmots are more eclectic grazers than previously though.

Table 2 Plant taxa identified in the diet of the golden marmot (*Marmota caudata*) in Deosai National Park (Pakistan), based on sequence variation of the P6 loop of the chloroplast trnL (UAA) intron using feces as a source of DNA.

			Faeces sample												
Family	**Plant taxon**	**Level of identification**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total**
Apiaceae	*Heracleum candicans*	Species		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		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	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		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

			Faeces sample												
Family	**Plant taxon**	**Level of identification**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total**
Cannabaceae	*Cannabis sativa**	Species									x				1
Caryophyllaceae	*Cerastium*	Genus	x		x	x		x	x	x		x	x	x	9
	Cerastium cerastoides	Species	x		x	x		x	x	x	x	x	x	x	10
	Cerastium pusillum	Species	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
Fabaceae	*Astragalus rhizanthus*	Species	x	x	x	x	x		x	x	x	x			9
	Galegeae	Tribe	x			x			x						3
	Oxytropis cachemiriana	Species	x		x	x		x			x		x	x	7
Lamiaceae	*Dracocephalum nutans*	Species		x	x										2
	Mentheae	Tribe	x	x	x	x	x			x	x		x		8
Onagraceae	*Chamerion latifolium*	Species			x										1
Papaveraceae	*Papaver nudicaule*	Species	x	x											2
Pinaceae	*Picea**	Genus			x										1
Plantaginaceae	*Lagotis kunawurensis*	Species											x		1
	*Plantago**	Genus											x		1
Poaceae	*Agrostis vinealis*	Species								x					1
	Elymus longi-aristatus	Species						x		x		x			3
	Poa alpina	Species						x							1
	Poa supina	Species					x	x			x			x	4
	*Pooideae**	Subfamily	x		x			x	x	x	x		x		7
Polygonaceae	*Aconogonon rumicifolium*	Species		x		x	x								3
	Polygonaceae	Family							x		x	x			3
	Polygonum cognatum	Species	x		x			x							3
	*Rumex**	Genus	x		x	x	x	x	x		x	x			8

			Faeces sample												
Family	**Plant taxon**	**Level of identification**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total**
	Rumex nepalensis	Species	x		x	x	x	x	x			x			7
Rosaceae	*Cotoneaster affinis*	Species											x		1
	Potentilla argyrophylla	Species	x	x	x			x			x				5
	Rosoideae	Subfamily	x	x			x		x		x				5
Rubiaceae	*Galium boreale*	Species		x											1
Saxifragaceae	*Saxifraga hirculus*	Species												x	1
Solanacee	*Solanum**	Genus						x	x						2
Total number of plant species per faeces			17	12	21	18	18	20	19	11	17	17	16	7	

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

Diet analysis of Himalayan brown bears (Nawaz et al. *in preparation)*

The *trn*L approach was coupled with individual and sex identification using microsatellites polymorphism (**Paper I**). Moreover, for each fecal sample the sampling date and the geographical coordinates were recorded by a GPS receiver (Garmin 12XL). This gives the opportunity to assess individual and sexual differentiation in the diet, and also to study temporal and geographical variations.

The 63 fecal samples that were successfully typed by microsatellites (**Paper I**) were also typed at the *trn*L locus. All samples gave consistent results but one, also when PCR cycles were increased to 45, conditions that favor the amplification of trace DNA molecules.

For all 62 samples totally 142030 sequences were obtained, with an average of 2,328.36 ± 921.09 sequences per sample. As found in **Paper III**, for each sample a few sequences were found hundreds of times, and only sequences that were repeated at least 4 times were taken into account in the subsequent analysis.

In total 57 plant taxa were found in bear feces, belonging to 50 genera and 29 families. Forty-seven percent of plants were identified at species level, 74% at genera, 77% at tribe, 82% at subfamily, and 100% at their family level. These results are much higher than those obtained for the 12 test samples in **Paper III**.

About 70% of identified taxa were present in ≤ 3 samples, and 27 taxa were present only in one sample. There were only four taxa with occurrence in more than 50% samples; one unidentified species of Poaceae, two of Cyperaceae (*Carex diluta, Carex* sp.), and one of Apiaceae (*Heracleum candicans*). The unidentified Poaceae species (subfamily Pooideae) had the highest frequency (92%). Among the 29 identified families, 14 were only present in one sample. Regular plant diet (≥ 10% occurrence) of brown bears consists of only eight families; Poaceae, Polygonaceae, Cyperaceae, Apiaceae, Asteraceae, Caryophyllaceae, Lamiaceae, Rubiaceae. The first four families make the preferred diet with more than 50% occurrence. Those results suggest that the plant families that do not belong to the preferred diet enter occasionally in the bear diet by chance during grazing.

The diet per individual is shown in Table 3. Graminoids (Poaceae and Cyperaceae) are presents in all individuals. Poeaceae were eaten by 27 individuals and

Cyperaceae by 21. Polygonaceae species are also an important food source for Himalayan brown bears, found in the diet of 26 of the 28 total individuals studied.

Variation among individuals in terms of total taxa in diet was not significant at species (χ^2: 38.06, P = 0.09) and family level (χ^2: 24.12, P = 0.67). The frequency of plants across individual bears ranged from 3 to 97%, the majority (72%) of plants were occasionally eaten (represented in <10% individuals). There were only eight taxa (*Agrostis vinealis, Asteraceae sp., Bistorta affinis, Carex diluta, Carex sp., Heracleum candicans, Poa supina, Poa sp*.) present in > 50% individuals. Four taxa Poaceae, Polygonaceae, Cyperaceae, Apiaceae were present in at least half of the individuals. Graminoids and forbs were eaten by all individuals, and browse plants occurred only in 10% of cases. Correspondence analysis did not show major difference in the individual level for plant species, family and groups. Again this results support the hypothesis that bears have few preferred plants while the majority of plants species found by barcoding enter in the their diet by chance.

Among the 62 fecal samples analyzed with the *trn*L method; 21 belonged to females, 37 to males, and four to individuals of unidentified sex. . In females, 34 plants txa were identified and 43 taxa from male samples. The ratio of graminoids to forbs did not differ significantly (χ^2: 0.24, P = 0.63) among sexes. The analysis of odds shows that females prefer *Aconogonon rumicifolium*, *Agrostis vinealis*, *Heracleum candicans*, and Menteae (*Nepeta linearis* or *Thymus linearis*). Instead males preference was for *Bistorta affinis*, two *Carex* species (*Carex diluta, C.* sp.), and one unidentified species of Polygonaceae. However those differences were not statistically significant, except for preferences of males for *Bistorta affinis* and *Carex* sp.. Female with cubs may show difference in their diet due also to different habitat selection, related to the avoidance of males. It will be interesting to test this hypothesis, but it was not possible in the present study.

No temporal difference was found in number of taxa (χ^2: 2.54, P = 0.77) and families (χ^2: 2.2, P = 0.82). However the ratio of graminoid forage to forbs changed significantly over three months (Spearman's r: -0.82, P = 0.04), favoring forbs in later season. At family level four plant families showed a temporal trend; Asteraceae and Poaceae declined in late season, while Polygonaceae and Fabaceae showed an increasing trend. This increase is linked to seed growth in the late season. Protein-rich food is preferred by bears for structural growth, instead calories-rich food is preferred

during hyperphagic period (period characterized by and high food intake that just precede hibernation period) (Gilbert & Lanner 1995; Brody & Pelton 1988). Brown bears hibernate during winter so they have to accumulate fat layer to maintain basic metabolic rate during this period. For female, storage of energetic food is even more important, because cub birth happens during hibernation. Other brown bear populations such as the Cantabrian (Naves *et al.* 2006) and Scandinavian (Dahle *et al.* 1998; Persson *et al.* 2001) also show a temporal trend in the feeding behavior. Energy-rich food is favored in the late season, for example the quantity in the total fecal volume of berries increase for Scandinavian brown bears from spring to autumn (from 31.9 to 64.7 of the fecal volume). Similar to other brown bears population, the Deosai bears prefer seed and caloric food in late season.

Among 62 fecal samples used with the *trn*L method, 15 were collected from marshy habitats, 16 from grassy and 13, 7 from stony, and rocky within the park. Ten were from surrounding valleys and location of 1 sample was not known (missing GPS coordinates). Neither number of taxa (χ^2:1.52, P = 0.82) and number of families (χ^2:1.85, P = 0.76) varied significantly across habitat types. However four families; Adoxaceae, Araliaceae, Ephedraceae and Orobanchaceae were represented only in samples from valleys. Pinaceae and Cupressaceae were found in samples collected in the plateau, although these families are present only in valleys outside the park. Unfortunately, the passage rate of food in the gut is unknown for this species, so we can't deduce when the bear have eaten a particular plant. We can only hypothesize that after having foraged in the valley they came back to the Deosai plain. Brown bear hibernate in the valley, and before hibernation they eat coniferous leaves (Nawaz personal communication). The feces in which we found conifers DNA, was collected close to a valley and at the end of September, therefore we hypothesize that this individual was preparing for the hibernation period.

Table 3. Plant taxa identified in the diet of the Himalayan brown bear individuals (*Ursus arctos*) in Deosai National Park (Pakistan), based on sequence variation of the P6 loop of the chloroplast trnL (UAA) intron using faeces as a source of DNA

Family	Plant founded	Level of identification	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	Total
Actinidiaceae	*Actinidia*	genus														x															1
Adoxaceae	Adoxaceae	family	x																												1
Apiaceae	Apioideae	subfamily	x																						x						2
	Heracleum candicans	species	x		x	x	x	X	x	x	x			x		x	x	X	x	x	x	x			x	x					18
Araliaceae	Araliaceae	family	x																												1
Asteraceae	Asteraceae	family	x			x		X	x	x				x			x	X				x	x	x	x				x	13
	Leontopodium brachyactis	species																X													1
Brassicaceae	*Thlaspi andersonii*	species																	x												1
Caryophyllaceae	*Cerastium*	genus										x								x	x										3
	Cerastium cerastoides	species				x				x	x	x						x			x		x								7
	Cerastium pusillum	species							x			x								x	x					x					5
Crassulaceae	*Rhodiola*	genus											x																		1
Cupressaceae	Cupressaceae	family								x						x															2
Cyperaceae	*Carex*	genus	x			x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x			x					19
	Carex diluta	species		x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x			x					20
Ephedraceae	*Ephedra gerardiana*	species						X																							1
Euphorbiaceae	*Euphorbia*	genus																												x	1
	Euphorbiaceae	family																												x	1
Fabaceae	*Astragalus rhizanthus*	species				x		X								x															3
	Galegeae	tribe				x		X																							2
	Glycine	genus					x																								1
	Oxytropis cachemiriana	species														x															1
Juncaceae	*Juncus*	genus												x																	1
Lamiaceae	Mentheae	tribe			x			X	x					x	x												x			x	7
Lycopodiaceae	Lycopodiaceae	family								x																					1
Orobanchaceae	*Pedicularis*	genus				x																									1
	Pedicularis albida	species				x																									1

Family	Plant founded	Level of identification	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	Total
Papaveraceae	*Papaver nudicaule*	species																		x											1
Pinaceae	*Cedrus*	genus																									x				1
Plantaginaceae	Plantaginaceae	family														x															1
Poaceae	*Agrostis vinealis*	species	x	x				X	x	x	x		x		x				x	x	x			x						x	13
	Elymus longi-aristatus	species		x	x			X	x						x	x			x	x	x		x								10
	Koeleria macrantha	species							x									x					x								3
	Poa	genus																					x								1
	Poa alpine	species																		x											1
	Poa	genus								x																					1
	Poa supine	species	x	x	x	x	x	X	x	x		x		x	x	x	x	x	x	x			x			x					18
	Pooideae	subfamily	x	x	x	x	x	X	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	27
	Stipeae	tribe						X																							1
Polygonaceae	Aconogonon rumicifolium	species		x	x		x				x		x		x	x			x		x			x			x	x		x	13
	Bistorta affinis	species			x	x			x	x	x	x	x	x		x	x	x		x		x	x	x				x	x		17
	Polygonaceae	family		x	x		x		x				x	x		x	x			x				x	x			x	x		13
	Polygonum cognatum	species						X								x															2
	Rumex nepalensis	species			x		x		x				x	x			x			x			x	x	x						10
Polyosmaceae	Polyosma	genus																							x					x	2
Ranunculaceae	Aconitum violaceum	species									x					x	x														3
	Thalictrum	genus	x																												1
Rosaceae	Alchemilla	genus																					x								1
	Cotoneaster affinis	species							x																						1
	Rosoideae	subfamily	x					X		x																					3
Rubiaceae	Galium	genus						X									x														2
	Galium boreale	species						X		x							x				x				x	x						6
	Rubiaceae	family						X																							1
Rutaceae	Rutaceae	family															x														1
Salicaceae	Salix	genus														x															1
Saxifragaceae	Saxifraga flagellaris	species											x																		1
	Saxifraga hirculus	species				x	x									x															3

Conclusions and perspectives

The genetic status of Himalayan brown bear population in Deosai National Park

The first goal of this thesis was to evaluate the genetic status of Himalayan brown bear population in Deosai National Park. We have documented that, in contrast with the expectations, this population while showing moderate levels of diversity it is not at immediate risk of inbreeding. The population probably began to lose genetic diversity about 1000 years ago, when it began to decline from a single large population throughout northern Pakistan, with population fragmentations and consequent loss of connectivity. The population decline stopped in Deosai about 15 years, ago when the population received increased protection, with the creation of the park. We also documented a gene flow between this population and neighbor populations, that probably maintained the moderate gene diversity recorded in this population. However, it will be essential, for the viability of that population, to improve connectivity with adjacent populations. Otherwise, the population will continue to lose genetic diversity over time. This population is the biggest recorded in Pakistan (Nawaz 2007), and also for its genetic status its role in the future will be probably as source population for the other ones. We suggest that future studies will continue to carefully monitor the population, both with field observations and genetic analyses. Concrete management actions should aim at maintaining and improving connectivity with other populations to maintain or improve levels of genetic diversity. Increasing the size and range of fecal sampling would not only allow a more precise estimate of the population size, but also give a better estimate of incoming gene flow.

DNA barcoding applied to diet analysis

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 mollusks.

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 in the efficiency of amplifications among species. 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 analyzing complex situations, when the diet is composed of many different species. When coupled with individual identification using microsatellite polymorphism, it will be possible to assess and compare diet among individuals and sexes, even without observing the animals.

The method is particularly well suited for large-scale analyses, with the possibility to analyze 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 individuals, and sexes, as in the case of Himalayan brown bears. 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: the level of identification, sequencing errors, and the quantification of plant species in diet.

The first limitation can be avoided by building a comprehensive database of *trn*L (UAA) introns for the majority of the plant species that occur in a particular area, and by complement this approach by one or several additional systems, specially designed for amplifying a short and variable region in problematic genera, such as *Carex, Salix, Pinus*, etc.. These primers might target other more variable parts of the chloroplast DNA, or the nuclear ribosomal DNA, such as the internal transcribed spacers.

Sequencing errors occur in every sequencer, but the specific strategy we used might induce some errors. 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. In order to assign sequences to individuals, we tagged each sample by a short oligonucleotide sequence. 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), the tagging system was changed 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 (Gielly *et al.* in preparation). 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 feces, 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. 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 only at a low frequency. 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 type of the different sequencing errors and of their associated probabilities.

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, these numbers cannot be interpreted as quantitative at the moment for several reasons.

First, the preferential amplification of some species when analyzing 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 (**Paper II**), should limit such preferential amplification. 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 type of tissue eaten. Leaves will undoubtedly provide more chloroplast DNA than roots, and the *trn*L approach cannot determine the tissue that has been eaten.

In simple conditions, i.e. when the animal is eating only a few species and is additionally fed 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.

In future studies (e.g. Nawaz *et al.* in preparation) our method will be coupled with classical morphological fecal analysis and isotopes approach on the same Himalayan brown bears samples to test the reliability of the technique. The *trn*L approach is suitable only for study herbivorous diet. For this reason it is necessary to couple it with other complementary method for study carnivorous or omnivorous animals. Further investigation will help to develop a similar approach that we propose in this study, i. e. DNA barcoding via universal primers and parallel pyrosequencing, for diet analysis of carnivorous animals. Preliminary studies on this approach have been already initiated in Laboratoire d'Ecologie Alpine, Grenoble, France. The work on Macaroni penguin by Deagle *et al.* 2007, has already shown that the use of universal primers for carnivorous animals is possible.

Acknowledgements

Several people helped me during those three years, and made this thesis possible. I would like to thank them all, both for their material and emotional support. First of all, I owe my sincere gratitude to Giuseppe Nascetti who followed me since my first steps in science, when I was a graduate student, and who encouraged me to continue my PhD thesis in France (not following the usual Italian academic habits…). I wish to express my warmest thanks to Pierre Taberlet who received me in his laboratory (and Marie-Odile in her house) and helped and supported me in these years. I would like to particularly thank him for his enormous enthusiasm for science, that he tried to convey to me. He let me participate in the newest advances in his lab and gave me unique opportunities. Jon Swenson accepted me in the SBBRP I would like to thank him for allowing me be a "bear girl", for his enthusiasm in all things and for the innumerous and fruitful discussions we had together. I am grateful to Christian who helped me with the experiments and, most important, for making the life in the lab less stressful and more funny. Carole, "mon amorce" Ludo, Delphine and Stéphanie always there to answer all my questions and for supporting me. I would like to thank "Les filles", Margot, Lucie, Bénédicte, for the constructive discussions with me at work, for the less constructive discussions in the evening, and for been there in all the difficult moments. Eva, who introduced me in the bear genetics world and to Grenoble life. All the Bureau 313 people, Jean-Marie, Aurélie, Ameline, Claire and especially Francesco for assisting me to grow up in a scientific context. Thanks to François for his discussions, for his help in the mysterious world of 454 data analysis and for his very pertinent comments on my articles. A special thank to Eric and his "magic software" that saved me from blindness, and to Ali. He has worked in parallel with me on Himalayan brown bear and has completed my genetic knowledge, I would like to thank him for his pertinent questions that pushed me to improve myself. I would like to thank all the SBBRP people, Andreas, Andrés, Åsa, Jonas, Jonna, Jody, .Øle-Gunnar, Richard, and Sven for making the project meetings so interesting, for having always good ideas, to be a stimulus, but at the same time to be so friendly and funny. I would like to thank Gilles that helped me in my first steps in the statistical word and for his hospitality in the cold Norwegians winter. Finally I would like to thank Olive, who supported with patience all the up and down of my humor and for been always there for cheer me up during the "down periods".

Grazie a tutti voi-Merci à vous tous-Thanks all of you

References

Agresti A. (1996) An Introduction to Categorical Data Analysis. Wiley, New York

Agusti N., Shayler S.P., Harwood J.D., Vaughan I.P., Sunderland K.D. & Symondson W.O.C. (2003) Collembola as alternative prey sustaining spiders in arable ecosystems: prey detection within predators using molecular markers. *Molecular Ecology*, 12, 3467-3475

Arnot D.E., Roper C. & Bayoumi R.A.L. (1993) Digital codes from hypervariable tandemly repeated DNA sequences in the *Plasmodium falciparum* circumsporozoite gene can genetically barcode isolates. *Molecular and Biochemical Parasitology*, 61, 15-24

Beaumont, M. (1999) Detecting population expansion and decline using microsatellites. Genetics 153, 2013–2029.

Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N., & Bonhomme, F. (1996–2004) GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier (France).

Bellemain E. & Taberlet P. (2004) Improved noninvasive genotyping method: application to brown bear (Ursus arctos) faeces. *Molecular Ecology Notes*, 4, 519-522

Bellemain E., Swenson J.E., Tallmon O., Brunberg S. & Taberlet P. (2005) Estimating population size of elusive animals with DNA from hunter-collected feces: Four methods for brown bears. *Conservation Biology*, 19, 150-161

Benzécri, J.P. (1973) L'analyse des données. II L'analyse des correspondances, Bordas, Paris

Bertrand, C.D., Loutre, M.F., Crucifix, M., & Berger, A. (2002.) Climate of the last millennium: a sensitivity study. *Tellus – Series A – Dynamic Meteorology and Oceanography*, 54, 221–245.

Binladen, J., Gilbert, M.T.P., Bollback, J.P., Panitz, F., Bendixen, C., Nielsen, R., & Willerslev, E. (2007) The use of coded PCR primers enables high-troughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *Plos One*, 2, e197.

Blumstein D.T. & Foggin J.M. (1997) Effects of vegetative variation on weaning success, overwinter survival, and social group density in golden marmots (Marmota caudata aurea). *Journal of Zoology*, 243, 57-69

Borsch,T., Hilu,K.W., Quandt,D., Wilde,V., Neinhuis,C. & Barthlott,W. (2003) Noncoding plastid trnT-trnF sequences reveal a well resolved phylogeny of basal angiosperms. *Journal of Evolutionary Biology*, 16, 558–576.

Bradley B.J., Stiller M., Doran-Sheehy D.M., Harris T., Chapman C.A., Vigilant L. & Poinar H. (2007) Plant DNA sequences from feces: Potential means for assessing diets of wild primates. *American Journal Of Primatology*, 69, 699-705

Brody A.J. & Pelton M.R. (1988) Seasonal changes in digestion in black bears. *Canadian Journal of Zoology-Revue Canadienne de Zoologie*, 66, 1482-1484

Brownstein M.J., Carpten J.D. & Smith J.R. (1996) Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping. *Biotechniques*, 20, 1004-6, 1008-10

Bunnell F.L. & Tait D.E.N. (1981) Population dynamics of bears implications. Dynamics of large mammal populations, 75-98, Fowler C. W. and Smith T. D. (eds.), John Wiley and Sons, New York

Chase M.W., Salamin N., Wilkinson M., Dunwell J.M., Kesanakurthi R.P., Haidar N. & Savolainen V. (2005) Land plants and DNA barcodes: short-term and long-term goals. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 360, 1889-1895

Chase M.W., Cowan R.S., Hollingsworth P.M., van den Berg C., Madrinan S., Petersen G., Seberg O., Jorgsensen T., Cameron K.M., Carine M., Pedersen N., Hedderson T.A.J., Conrad F., Salazar G.A., Richardson J.E., Hollingsworth

M.L., Barraclough T.G., Kelly L. & Wilkinson M. (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon*, 56, 295-299

Codron D. & Brink J.S. (2007) Trophic ecology of two savanna grazers, blue wildebeest *Connochaetes taurinus* and black wildebeest *Connochaetes gnou*. *European Journal of Wildlife Research*, 53, 90-99

Cristobal-Azkarate J. & Arroyo-Rodriguez 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 O.J., Wedul E.H., Swenson J.E., 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 B.E., Jarman S.N., Pemberton D. & Gales N.J. (2005) Genetic screening for prey in the gut contents from a giant squid (Architeuthis sp.). *Journal of Heredity*, 96, 417-423

Deagle B.E., Gales N.J., Evans K., Jarman S.N., Robinson S., Trebilco R. & Hindel M.A. (2007) Studying seabird diet through genetic analysis of faeces: A case study on macaroni penguins (*Eudyptes chrysolophus*). *PLoS ONE*, 2, e831

DeNiro M.J. & Epstein S. (1978) Influence of diet on distribution of carbon isotopes in animals. *Geochimica et Cosmochimica Acta*, 42, 495-506

DeNiro M.J. & Epstein S. (1981) Influence of diet on the distribution of nitrogen isotopes in animals. *Geochimica et Cosmochimica Acta*, 45, 341-351

Dove H. & Mayes R.W. (1996) Plant wax components: A new approach to estimating intake and diet composition in herbivores. *Journal Of Nutrition*, 126, 13-26

Duffy J.E., Carinale B.J., France K.E., McIntyre P.B., Thebault E. & Loreau M. (2007) The functional role of biodiversity in ecosystems: incorporating trophic complexity. *Ecology Letters*, 10, 522-538

Duriez O., Sachet J.M., Menoni E., Miquel C. & Taberlet P. (2006) Phylogeography of the Capercaillie in Eurasia: what is the status of the Pyrenean-Cantabrian population? *Journal Of Ornithology*, 147, 160-161

Eggert L.S., Eggert J.A. & Woodruff D.S. (2003) Estimating population sizes for elusive animals: the forest elephants of Kakum National Park, Ghana. *Molecular Ecology*, 12, 1389-1402

Ehlers, E., & Kreutzmall, H. (2000) High mountain ecology and economy potential and constraints. *High mountain pastoralism in Northern Pakistan*. Erdkundliches Wissen 132, 9–36. Ehlers E., & Kreutzmann, H. (eds). Franz Steiner Verlag, Stuttgart, Germany

Esper, J., Schweingruber, F.H., & Winiger, M. (2002) 1300 years of climatic history for Western Central Asia inferred from tree-rings. *Holocene*, 12, 267–278.

Feranec R.S. & MacFadden K.B.J. (2000) Evolution of the grazing niche in Pleistocene mammals from Florida: evidence from stable isotopes. *Palaeogeography Palaeoclimatology Palaeoecology*, 162, 155-169

Ferreira L.M.M., Garcia U., Rodrigues M.A.M., Celaya R., Dias-Da-Silva A. & Osoro K. (2007a) Estimation of feed intake and apparent digestibility of equines and cattle grazing on heathland vegetation communities using the n-alkane markers. *Livestock Science*, 110, 46-56

Ferreira L.M.M., Olivan M., Celaya R., Garcia U., Rodrigues M.A.M. & Osoro K. (2007b) The use of the alkane technique to estimate diet selection of sheep grazing grass-clover/heather-gorse vegetation communities. *Journal Of The Science of Food and Agriculture*, 87, 274-285

Floyd R., Abebe E., Papert A. & Blaxter M. (2002) Molecular barcodes for soil nematode identification. *Molecular Ecology*, 11, 839-850

Foley W.J., McIlwee A., Lawler I., Aragones L., Woolnough A.P. & Berding N. (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

Gielly L. & Taberlet P. (1996) A phylogeny of the European gentians inferred from chloroplast trnL (UAA) intron sequences. *Botanical Journal of the Linnean Society,* 120, 57-75

Gielly L, Coissac E, Valentini A, Miquel C, Paris M, Pompanon F, & Taberlet P in preparation. A new tagging system for parallel pyrosequencing of multiple homologous PCR products.

Gilbert, B. K. & R. M. Lanner. 1995. Energy, diet selection and restoration of brown bear populations. *Proceedings of the 9th International Conference of Bear Research and Management*, 231-240

Hagler Bailly Pakistan (1999) Pakistan's National Communication to the UNFCC. Report prepared for the Ministry of Environment, Islamabad, Pakistan

Harris D.J. (2003) Can you bank on GenBank? *Trends in Ecology & Evolution*, 18, 317-319

Hebert P.D.N., Ratnasingham S. & deWaard J.R. (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London Series B-Biological Sciences,* 270, S96-S99

Hilderbrand G.V., Jacoby M.E., Schwartz C.C., Arthur S.M., Robbins C.T., Hanley T.A. & Servheen C. (1999) The importance of meat, particularly salmon, to body size, population productivity, and conservation of North American brown bears. *Canadian Journal of Zoolo*gy 77, 132-138

Himalayan Wildlife Foundation (1999a) Deosai brown bear project, 1998. Final Report, Islamabad. Himalayan Wildlife Foundation, Islamabad

Himalayan Wildlife Foundation (1999b) Management plan for Deosai National Park Northern Areas Pakistan. Himalayan Wildlife Foundation, Islamabad

Hobson K.A., McLellan B.N. & Woods J.G. (2000) Using stable carbon ($d^{13}C$) and nitrogen ($d^{15}N$) isotopes to infer trophic relationships among black and grizzly bears in the upper Columbia River basin, British Columbia. *Canadian Journal of Zoology-Revue Canadienne de Zoologie*, 78, 1332-1339

Hofreiter M., Poinar H.N., Spaulding W.G., Bauer K., Martin P.S., Possnert G. & Paabo S. (2000) A molecular analysis of ground sloth diet through the last glaciation. *Molecular Ecology*, 9, 1975-1984

Holechek J.L., Vavra M., Pieper R.D. (1982) Botanical composition determination of range diets: a review. *Journal of Range Management*, 35, 309-315.

Jarman S.N., Deagle B.E. & Gales N.J. (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 I.R. (2006) Can near infrared spectroscopy be used to improve assessment of marine mammal diets via fecal analysis? *Marine Mammal Science*, 22, 261-275

Kohn M.H., York E.C., Kamradt D.A., Haugt G., Sauvajot R.M. & Wayne R.K. (1999) Estimating population size by genotyping faeces. *Proceedings of the Royal Society of London Series B-Biological Sciences*, 266, 657-663

Kuhle, M. (1997) New findings concerning the ice age (last glacial maximum) glacier cover of the East-Pamir, of the Nanga Parbat up to the Central Himalaya and of Tibet, as well as the age of the Tibetan inland ice. *GeoJournal*, 42, 87–257.

Mackay, A.W., Ryves, D.B., Battarbee, R.W., Flower, R.J., Jewson, D., Rioual, P., & Sturm, M. (2005) 1000 years of climate variability in central Asia: assessing the evidence using Lake Baikal (Russia) diatom assemblages and the application of a diatom-inferred model of snow cover on the lake. *Global and Planetary Change*, 46, 281–298.

Magnuson V.L., Ally D.S., Nylund S.J., Karanjawala Z.E., Rayman J.B., Knapp J.I., Lowe A.L., Ghosh S. & Collins F.S. (1996) Substrate nucleotide-determined

non-templated addition of adenine by Taq DNA polymerase: implications for PCR-based genotyping and cloning. *Biotechniques*, 21, 700-9

Margulies M., Egholm M., Altman W.E., Attiya S., Bader J.S., Bemben L.A., Berka J., Braverman M.S., Chen Y.-J., Chen Z., Dewell S.B., Du L., Fierro J.M., Gomes X.V., Godwin B.C., He W., Helgesen S., Ho C.H., Irzyk G.P., Jando S.C., Alenquer M.L.I., Jarvie T.P., Jirage K.B., Kim J.-B., Knight J.R., Lanza J.R., Leamon J.H., Lefkowitz S.M., Lei M., Li J., Lohman K.L., Lu H., Makhijani V.B., McDade K.E., McKenna M.P., Myers E.W., Nickerson E., Nobile J.R., Plant R., Puc B.P., Ronan M.T., Roth G.T., Sarkis G.J., Simons J.F., Simpson J.W., Srinivasan M., Tartaro K.R., Tomasz A., Vogt K.A., Volkmer G.A., Wang S.H., Wang Y., Weiner M.P., Yu P., Begley R.F. & Rothberg J.M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437, 376

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 M.L., Vavra M. & Krueger W.C. (1983) A Comparison Of 4 Methods Used To Determine The Diets Of Large Herbivores. *Journal of Range Management*, 36, 302-306

Miquel C., Bellemain E., Poillot C., Bessiere J., Durand A. & Taberlet P. (2006) Quality indexes to assess the reliability of genotypes in studies using noninvasive sampling and multiple-tube approach. *Molecular Ecology Notes*, 6, 985-988

Naves J., Fernandez-Gil A., Rodriguez C. & Delibes M. (2006) Brown bear food habits at the border of its range: A long-term study. *Journal of Mammalogy*, 87, 899-908

Nawaz M.A. & Kok O.B. (2004) Aktiwiteitspatrone van bruinbere (*Ursus arctos*) op die Deosaiplato, noordelike Pakistan (Activity patterns of brown bears (*Ursus arctos*) on the Deosai Plateau, Northern Pakistan). Suid Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie 23, 61-63.

Nawaz, M.A., Shah, M., & Zakaria, V. (2006) Environmental Baseline of Deosai National Park Draft Report. Himalayan Wildlife Foundation, Islamabad.

Nawaz M.A. (2007) Status of the brown bear in Pakistan. *Ursus*, 18, 89-100

Nawaz M.A., Swenson J.E. & Zakaria V. unpublished An increasing low-productive, high-altitude brown bear population in South Asia; a successful case of national park management

Nawaz, M.A., Valentini, A., Noor Kamal Khan, N.K., Taberlet, P., & Swenson, J.E. in preparation Diet of Himalayan brown bear

Norbury G.L. & Sanson G.D. (1992) Problems with measuring diet selection of terrestrial, mammalian herbivores. *Australian Journal of Ecology*, 17, 1-7

Nystrom J., Dalen L., Hellstrom P., Ekenstedt J., Angleby H. & Angerbjorn A. (2006) Effect of local prey availability on gyrfalcon diet: DNA analysis on ptarmigan remains at nest sites. *Journal of Zoology*, 269, 57-64

Paetkau D. & Strobeck C. (1994) Microsatellite analysis of genetic variation in black bear populations. *Molecular Ecology*, 3, 489-495

Paetkau D., Calvert W., Stirling I. & Strobeck C. (1995) Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology*, 4, 347-354

Paetkau, D., Waits, L.P., Clarkson, P., Craighead, L., Vyse, E., Ward, R., & Strobeck, C. (1998) Dramatic variation in genetic diversity across the range of North American brown bears. *Conservation Biology*, 12, 418–429

Passmore A.J., Jarman S.N., Swadling K.M., Kawaguchi S., McMinn A. & Nicol S. (2006) DNA as a Dietary Biomarker in Antarctic Krill, Euphausia superba. *Marine Biotechnology*, V8, 686

Persson I.L., Wikan S., Swenson J.E. & Mysterud I. (2001) The diet of the brown bear *Ursus arctos* in the Pasvik Valley, northeastern Norway. *Wildlife Biology*, 7, 27-37

Piasentier E., Sacca E. & Bovolenta S. (2007) Dietary selection and ingestive behaviour of fallow deer and sheep grazing on adjacent monocultures of white clover and tall fescue. *Small Ruminant Research*, 71, 222-233

Piggott M.P., Bellemain E., Taberlet P. & Taylor A.C. (2004) A multiplex pre-amplification method that significantly improves microsatellite amplification and error rates for faecal DNA in limiting conditions. *Conservation Genetics*, 5, 417-420

Poinar H.N., Hofreiter M., Spaulding W.G., Martin P.S., Stankiewicz B.A., Bland H., Evershed R.P., Possnert G. & Paabo S. (1998) Molecular coproscopy: Dung and diet of the extinct ground sloth Nothrotheriops shastensis. *Science*, 281, 402-406

Poinar H.N., Kuch M., Sobolik K.D., Barnes I., Stankiewicz A.B., Kuder T., Spaulding W.G., Bryant V.M., Cooper A. & Paabo S. (2001) A molecular analysis of dietary diversity for three archaic Native Americans. *Proceedings of the National Academy of Sciences of the United States Of America*, 98, 4317-4322

Polz M.F. & Cavanaugh C.M. (1998) Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology*, 64, 3724-3730

Pompanon F., Bonin A., Bellemain E. & Taberlet P. (2005) Genotyping errors: Causes, consequences and solutions. *Nature Reviews Genetics*, 6, 847-859

Pons J. (2006) DNA-based identification of preys from non-destructive, total DNA extractions of predators using arthropod universal primers. *Molecular Ecology Notes*, 6, 623-626

Quandt,D., Mu¨ller,K., Stech,M., Frahm,J.P., Frey,W., Hilu,K.W. & Borsch,T. (2004) Molecular evolution of the chloroplast *trn*L-F region in land plants. *Monographs in Systematic Botany from the Missouri Botanical Garden*, 98, 13–37.

Quandt D. & Stech M. (2005) Molecular evolution of the *trn*L(UAA) intron in bryophytes. *Molecular Phylogenetics And Evolution*, 36, 429-443

Raymond, M., Rousset, F. (1995) Genepop (version 1.2), population genetics software for exact tests and ecumenicism. *Journal of Heredity*, 86, 248–249.

Rollo F., Ubaldi M., Ermini L. & Marota I. (2002) Otzi's last meals: DNA analysis of the intestinal content of the Neolithic glacier mummy from the Alps. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 12594-12599

Sanger F., Nicklen S. & Coulson A.R. (1977) Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74, 5463-5467

Sare D.T.J., Millar J.S. & Longstaffe F.J. (2005) Tracing dietary protein in red-backed voles (*Clethrionomys gapperi*) using stable isotopes of nitrogen and carbon. *Canadian Journal of Zoology-Revue Canadienne de Zoologie*, 83, 717-725

Servheen, C. (1990) The status and conservation of the bears of the world. *International Association for Bear Research and Management Monograph Series*, 2

Servheen, C., Herrero, S., & Peyton, B. (1999) Status Survey and Conservation Action Plan for BearsIUCN/SSC Bear and Polar Bear Specialist Groups. IUCN, Gland, Switzerland and Cambridge, UK

Shaw,J., Lickey,E.B., Beck,J.T., Farmer,S.B., Liu,W., Miller,J., Siripun,K.C., Winder,C.T., Schilling,E.E. & Small,R.L. (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany*, 92, 142–166.

Sogin M.L., Morrison H.G., Huber J.A., Welch D.M., Huse S.M., Neal P.R., Arrieta J.M. & Herndl G.J. (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

Solberg, H., Bellemain, E., Drageset, O.M., Taberlet, P., & Swenson, J.E. (2006) An evaluation of field and genetic methods to estimate brown bear (*Ursus arctos*) population size. *Biological Conservation* 128, 158–168.

Sponheimer M., Robinson T., Ayliffe L., Passey B., Roeder B., Shipley L., Lopez E., Cerling T., Dearing D. & Ehleringer J. (2003) An experimental study of carbon-

isotope fractionation between diet, hair, and feces of mammalian herbivores. *Canadian Journal of Zoology-Revue Canadienne de Zoologie*, 81, 871-876

Sugimoto T., Nagata J., Aramilev V., Belozor A., Higashi S. & McCullough D. (2006) Species and sex identification from faecal samples of sympatric carnivores, Amur leopard and Siberian tiger, in the Russian Far East. *Conservation Genetics*, 7, 799

Symondson W.O.C. (2002) Molecular identification of prey in predator diets. *Molecular Ecology*, 11, 627-641

Taberlet P., Gielly L., Pautou G. & Bouvet J. (1991) Universal Primers For Amplification Of 3 Noncoding Regions Of Chloroplast Dna. *Plant Molecular Biology*, 17, 1105-1109

Taberlet P., Griffin S., Goossens B., Questiau S., Manceau V., Escaravage N., Waits L.P. & Bouvet J. (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*, 24, 3189-3194

Taberlet P., Camarra J.J., Griffin S., Uhres E., Hanotte O., Waits L.P., DuboisPaganon C., Burke T. & Bouvet J. (1997) Noninvasive genetic tracking of the endangered Pyrenean brown bear population. *Molecular Ecology*, 6, 869-876

Valiere N. & Taberlet P. (2000) Urine collected in the field as a source of DNA for species and individual identification. *Molecular Ecology*, 9, 2150-2152

Waits, L.P., Talbot, S., Ward, R.H., & Shields, G. (1998) Phylogeography of the North American brown bear and implications for conservation. *Conservation Biology*, 12, 408–417.

Waits L., Taberlet P., Swenson J.E., Sandegren F. & Franzen R. (2000) Nuclear DNA microsatellite analysis of genetic diversity and gene flow in the Scandinavian brown bear (Ursus arctos). *Mol Ecol*, 9, 421-31

Waits L.P., Luikart G. & Taberlet P. (2001) Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Molecular Ecology*, 10, 249-56

White G.C. & Burnham K.P. (1999) Program MARK: survival estimation from populations of marked animals. *Bird Study*, 46, 120-139

Williams R., Peisajovich S.G., Miller O.J., Magdassi S., Tawfik D.S. & Griffiths A.D. (2006) Amplification of complex gene libraries by emulsion PCR. *Nature Methods*, 3, 545

Wilson A.S., Taylor T., Ceruti M.C., Chavez J.A., Reinhard J., Grimes V., Meier-Augenstein W., Cartmell L., Stern B., Richards M.P., Worobey M., Barnes I. & Gilbert M.T.P. (2007) Stable isotope and DNA evidence for ritual sequences in Inca child sacrifice. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 16456-16461

Woods, C.A., Kalpatrick, W.C., Rafique, M., Shah, M., Khan, W. (1997) Biodiversity and conservation of the Deosai Plateau, Northern areas, Pakistan. *Biodiversity of Pakistan*, 33-62, Mufti, S.A., Woods, C.A., Hasan, S.A. (Eds.), Pakistan Museum of Natural History, Islamabad and Florida Museum of Natural History, Gainesville

Zhang D.X. & Hewitt G.M. (1996) Highly conserved nuclear copies of the mitochondrial control region in the desert locust Schistocerca gregaria: Some implications for population studies. *Molecular Ecology*, 5, 295-300

Paper I

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/biocon

Genetic tracking of the brown bear in northern Pakistan and implications for conservation

Eva Bellemain[a,*], *Muhammad Ali Nawaz*[b,e], *Alice Valentini*[a,c], *Jon E. Swenson*[b,d], *Pierre Taberlet*[a]

[a]*Laboratoire d'Ecologie Alpine (LECA), CNRS UMR 5553, Université Joseph Fourier, BP 53, F-38041 Grenoble Cedex 9, France*
[b]*Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, Postbox 5003, No-1432 Ås, Norway*
[c]*Dipartimento di Ecologia e Sviluppo Economico Sostenibile, Università degli Studi della Tuscia, via S. Giovanni Decollato 1, 01100 Viterbo, Italy*
[d]*Norwegian Institute for Nature Research, Tungasletta 2, NO-7485 Trondheim, Norway*
[e]*Himalayan Wildlife Foundation, Islamabad, Pakistan*

ARTICLE INFO

Article history:
Received 4 April 2006
Received in revised form
23 August 2006
Accepted 3 September 2006
Available online 30 October 2006

Keywords:
Bottleneck
Feces
Individual identification
Pakistan
Ursus arctos

ABSTRACT

Asian bears face major threats due to the impact of human activities as well as a critical lack of knowledge about their status, distribution and needs for survival. Once abundant in northern Pakistan, the Himalayan brown bear (*Ursus arctos isabellinus*) has been exterminated in most of its former distribution range. It presently occurs sparsely, in small populations, the Deosai National Park supporting the largest isolate. This decline might imply a reduction in genetic diversity, compromising the survival of the population. Using a combination of fecal DNA analysis and field data, our study aimed at assessing the size and genetic status of the Deosai population and give guidelines for its conservation and management. Using fecal genetic analysis, we estimated the population to be 40–50 bears, which compares well with the field census of 38 bears. The northern Pakistani brown bear population may have undergone an approximate 200–300-fold decrease during the last thousand years, probably due to glaciations and the influence of growing human population. However, in spite of the presence of a bottleneck genetic signature, the Deosai population has a moderate level of genetic diversity and is not at immediate risk of inbreeding depression. Gene flow might exist with adjacent populations. We recommend careful monitoring of this population in the future both with field observations and genetic analyses, including sampling of adjacent populations to assess incoming gene flow. The connectivity with adjacent populations in Pakistan and India will be of prime importance for the long-term survival of Deosai bears.

1. Introduction

Brown bears (*Ursus arctos*) are the most endangered and least studied in Asia, where populations have declined by more than half in the past century (Servheen, 1990; Servheen et al., 1999). Asian bears face threats due to the impact of human activities and there is a critical lack of knowledge concerning their status, distribution and requirements for

* *Corresponding author:* Present address: Smithonian Tropical Research Institute, Apdo 2072, Balboa, Panama. Tel.: +507 212 8832; fax: +507 212 8790.
E-mail addresses: evabellemain@gmail.com (E. Bellemain), ali.nawaz@umb.no (M.A. Nawaz), alice.valentini@e.ujf-grenoble.fr (A. Valentini), jon.swenson@umb.no (J.E. Swenson), pierre.taberlet@ujf-grenoble.fr (P. Taberlet).
0006-3207/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.biocon.2006.09.004

survival (Servheen et al., 1999). The Himalayan brown bear (*U.a.isabellinus*), a highly threatened subspecies, is distributed in small isolated populations over the Himalaya, Karakoram, Hindu Kush, Pamir, western Kun Lun Shan, and Tian Shan ranges in southern Asia.

This bear has been exterminated in most of its former distribution range in Pakistan, and occurs very sparsely in small populations with limited connectivity in northern mountainous areas. Deosai National Park is the main stronghold of the brown bear population in Pakistan (Schaller, 1977; Roberts, 1997). Once abundant in Deosai, bear numbers declined drastically to as low as 19 in 1993 (Himalayan Wildlife Project, 1994). Although the population in Deosai has been recovering gradually since 1993 due to strict protection and conservation efforts, the decline could have reduced the genetic variability considerably. As a consequence, this population might suffer from inbreeding, and its survival might be compromised. Small population size is a great concern in conservation biology because small populations are more vulnerable to genetic factors, demographic and environmental stochasticity, genetic drift and inbreeding and have increased probability of extinction (Soulé, 1987). Evolutionary processes such as mutations, migration, selection and stochasticity are also fundamentally different than those in large populations. In small populations the role of stochasticity increases and the impact of selection is limited (Frankham et al., 2002). The loss of genetic diversity as a result of a bottleneck or continued small populations has been documented in many endangered species such as the northern elephant seal (*Mirounga angustirostris*) (Bonnell and Selander, 1974), Mauritius kestrel (*Falco punctatus*) (Groombridge et al., 2000), Indian rhinoceros (*Rhinoceros unicornis*) and Siberian tiger (*Panthera tigris*) (Hedrick, 1992). Fragmented populations are prone to many subtle threats, such as limited dispersal and colonization and restricted access to food and mates (Primack, 2002).

Documenting the status and distribution of Asian bears has been identified as a priority action for conservation by the IUCN/SSC Bear Specialist Group (Servheen et al., 1999). A comprehensive action plan is required for the long-term management of Himalayan brown bears. In order to be effective, an action plan should be based on reliable biological data, such as trustworthy estimates of population size, population genetic status and connectivity with other populations. Population size estimates are difficult to obtain for rare and elusive animals like brown bears (Bellemain et al., 2005). Field methods based on observations of recognizable individual bears have been used to estimate the size of the Deosai population, but these methods have not been compared with censuses using independent methods in order to evaluate their consistency.

To assess the genetic status and size of the Deosai population and give guidelines for the conservation and management of this population, we used the increasingly popular non-invasive genetic technique (Taberlet et al., 1996, 1999), in combination with field data. Using DNA analyses of fecal sampling, we aimed to answer the following questions: (i) Is the population size estimated from field data consistent with genetic censuses? (ii) Did the population suffer from a bottleneck at the genetic level and how long ago did it begin to decline? (iii) Are Deosai bears at risk of inbreeding depression? (iv) Is the population genetically isolated?

2. Material and methods

2.1. Study area and studied populations

The study was conducted in the Deosai National Park, Northern Areas, Pakistan. Deosai National Park is a plateau in the alpine ecological zone encompassing about 20,000 km^2, situated 30 km south of Skardu and 80 km east of the Nanga Parbat Peak. Elevations range from 3500 to 5200 m and about 60% of the area lies between 4000 and 4500 m. Recorded mean daily temperatures range from −20 °C to 12 °C. The annual precipitation in Deosai is 510–750 mm, and falls mostly as snow (Himalayan Wildlife Foundation, 1999a). The Deosai plains are covered by snow during winter months between November and May, and life on the plateau is confined to a window of five months.

The Deosai Plateau is situated between two of the world's major mountain ranges, the Karakoram and Himalaya. The biota includes plants and animals from Karakoram, Himalaya and Indus Valley. As a result Deosai is a center of unique biota in northern Pakistan. The documented biota of Deosai National Park includes 342 species of plants, 18 of mammals, 208 of birds, three of fishes, one of amphibian, and two of reptiles (Woods et al., 1997). Most of the plant species are herbaceous perennials, and cushions forming and tufted plants are common growth forms. Plains present a mosaic of plant communities according to the availability of water. The low lying areas usually consist of bogs and pools with associated flora consisting predominantly of grasses and sedges and plants such as *Saxifraga hircus*, *Swetia perfoliata* and *Aconitum violaceum*.

Deosai National Park supports the largest population of brown bears in Pakistan (unpublished data). The brown bear population in this park has been protected and closely monitored since 1993, and primary data on population size, behavior and ecology have been gathered (Himalayan Wildlife Foundation, 1999b). Field personnel were able to recognize dominant bears from their physical characteristics, coloration and well defined home ranges on this open plateau (Himalayan Wildlife Foundation, 1999a,b; Nawaz et al., 2006). Based on this, they estimated the number of bears annually, the approximate age of some males and females, as well as their reproductive behavior and, in some cases, relatedness (mothers and their young).

2.2. Fecal sampling

The study area was searched for bear feces from July to early October 2004. We divided the study area into five blocks, and each block was searched for bear feces in order to cover most of Deosai National Park (Fig. 1). Transects of 40–60 km length were placed in each block, and walked by a team of 2–3 people. The transect routes were planned in a way that these covered the maximum extent of the block and passed through areas known for frequent bear sightings. Transect routes usually resembled a loop, starting from the central road, progressing towards periphery of the park, and ended at the starting point. The team walked along opposite borders of a block while going towards the periphery of the park and coming back to the road. Each transect was completed in 2–3 days,

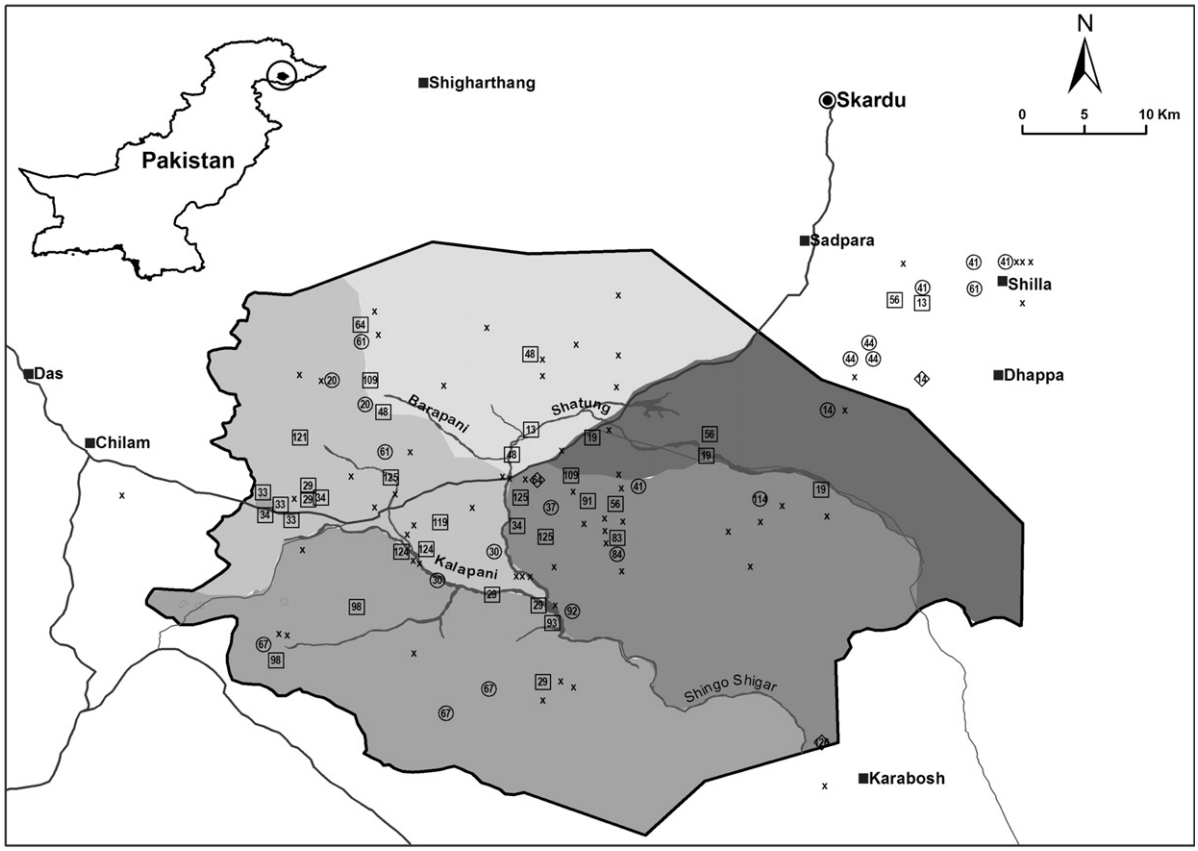


Fig. 1 – Map of the study area in the Deosai National Park, Northern Areas, Pakistan. Spatial distribution of brown bear genotypes is represented with squares for males, circles for females, and diamonds for unknown sex. Numbers within squares or circles represent individuals' identification numbers. Samples with negative/poor amplification are shown as "x". Five survey blocks are represented by different shades of grey.

with night stays made in portable tents. Apart from this planned collection, the field staff of Deosai National Park collected samples during their normal patrolling of the park.

Brown bears exhibit altitudinal migration in Deosai, and spend part of their life in surrounding valleys. We therefore collected feces from valleys connected to the park. When we found many feces together, usually at a bedding site, we collected one sample from the freshest feces. However if several feces were found at a food source (e.g. carcass) or we could differentiate different sizes, we took multiple samples. We picked up each fecal sample with a stick of wood and put 1 cm^3 of it in a 20-ml bottle. For each fecal sample, a sampling date, a geographical location and coordinates (latitude/longitude) were recorded using a GPS receiver (Garmin 12XL). Bottles were then filled with 95% alcohol to preserve the samples until DNA extraction.

Approximate ages of fecal samples were evaluated on the field and categorized into five classes; (1) fresh feces of the same day, (2) two–three days old, (3) one week old, (4) feces of the same month, and (5) feces older than one month.

2.3. *DNA extractions and typing*

2.3.1. *Extraction*

For every collected fecal sample, DNA extractions were performed using the Qiamp DNA Stool Kit (Qiagen, Hilden, Germany), developed especially for this type of material and following the manufacturer's instructions. All extractions occurred in a room dedicated to processing hairs and feces. Tubes containing samples and tubes without feces were treated identically to check for exogenous DNA contaminations.

2.3.2. *Genotyping for individual identification*

The extracted DNA was amplified using the six microsatellite primers described in Bellemain and Taberlet (2004) on a set of 16 feces to test for their polymorphism. The number of alleles per locus ranged from one to eight. The two primers showing only one or two alleles (Mu10 and G10L) were discarded for this analysis (but included later, see below) and the four others (Mu23, Mu50, Mu51, and Mu59) were kept. In order to obtain a probability of identity low enough to differentiate among all individuals, we redesigned two other microsatellite primer pairs, namely G10J and G10H (from Paetkau and Strobeck, 1994; Paetkau et al., 1995):

G10HFIpak: GGAGGAAGAAAGATGGAAAAC
G10HRpak: AAAAGGCCTAAGCTACATCG
G10JFpak: GCTTTTGTGTGTGTTTTTGC
G10JRIpak: GGTATAACCCCTCACACTCC

For sex identification, we used the SRY-primers described in Bellemain and Taberlet (2004).

We simultaneously amplified the following loci: Mu23 with Mu50; SRY with Mu51 and Mu59; G10Jpak with G10Hpak, using the internal fluorescent primers together with the appropriate external primers. We repeated each amplification eight times following the multi-tube approach (Taberlet et al., 1996). The fluorescent PCR products were loaded together on the single electrophoresis (ABI Prism 3100 DNA sequencer; Applied Biosystems, Foster City, California). The gels were analyzed using Genemapper (version 3.0) software package (Applied Biosystems, Foster City, California). We typed samples as heterozygous at one locus if both alleles appeared at least twice among the eight replicates and as homozygous if all the replicates showed identical homozygous profiles. If neither of those cases occurred, the alleles were treated as missing data.

We calculated a quality index for each sample following the rules defined in Miquel et al. (2006). To be conservative, we discarded the samples that had a quality index below 0.5.

2.3.3. *Genotyping for population genetics analyses*

To estimate population genetics parameters and relatedness, we increased the number of loci for each genetically identified individual. The highest quality sample per individual was selected, based on quality indices when the individual was represented by several samples. We amplified the following 12 additional microsatellites: G1A, G1D, G10B, G10C, G10L, G10P, G10X, G10O (Paetkau and Strobeck, 1994; Paetkau et al., 1995) and Mu05, Mu10, Mu15, Mu61 (Taberlet et al., 1997), using a modified protocol from Waits et al. (2000). The amplifications were performed using five combinations of loci: (1) G10B, G10C (2) G10X, G10P; (3) Mu61, Mu05; (4) G10O, G10L (5) G1D, Mu15; loci Mu10 and G1A were amplified separately. PCR reactions of 12.5 μL containing 2 μL template DNA, 0.1 mM each dNTP, 0.5 μM of each primer, 3 mM MgCl 2, 0.5 U AmpliTaq Gold Polymerase (Applied Biosystems) and 1 × Taq buffer (containing 100 mm Tris–HCl, pH 8.3, 500 mm KCl, according to the manufacturer's specifications; Applied Biosystems). Amplifications were performed in a GeneAmp PCR system 9700 (Applied Biosystems) with the following conditions: 10 min at 95 °C, 35 cycles composed of 30 s denaturing at 95 °C, 30 s annealing at 57 °C for combination 1, 45 °C for combination 2, 48 °C for combination 3, 52 °C for combination 4, 55 °C for combination 5, 52 °C for Mu10 and 55 °C for G1A, 1-min extension at 72 °C, and as a final extension step, 7 min at 72 °C. We repeated each amplification four times. The PCR products were mixed in three multiplexes (1st: 2 μL G1A, 3 μL G10B/G10C, 5 μL Mu61/Mu05; 2nd: 3 μL G1D/Mu15, 7 μL G10P/G10X; 3rd: 5 μL Mu10, 5 μL G10O/G10L). One μL of this multiplex was added to a 10 μL mix of formamide and ROX 350 (10:0.2), and then loaded on an automatic sequencer ABI3100 (Applied Biosystems, Foster City, California). The gels were analyzed using Genemapper (version 3.0) software package (Applied Biosystems, Foster City, California). The same rules as described above were applied for defining homozygous and heterozygous loci.

A new quality index Miquel et al. (2006) was calculated for each sample and locus. The loci G10P, Mu05 and Mu61 were discarded from the analysis because of their low quality indices (below 0.6). Finally, genotypes were obtained based on 15 loci.

2.3.4. *Calculating the probability of identity*

Using the software GIMLET version 1.3.1 (Valière, 2002), and both datasets (6 and 15 loci), we computed the probability of identity, i.e. the overall probability that two individuals drawn at random from a given population share identical genotypes at all typed loci (Paetkau and Strobeck, 1994). We also computed the probability of identity among siblings (Waits et al., 2001).

2.3.5. *Estimating current population size using rarefaction indices*

Following the method described in Kohn et al. (1999), we compared the 6-loci genotype of each sample with all those drawn previously and calculated the population size as the asymptote of the relationship between the cumulative number of unique genotypes and the number of samples typed. This curve is defined by the equation $y = (ax)/(b + x)$, where a is the asymptote, x the number of feces sampled, y the number of unique genotypes, and b the rate of decline in the value of slope. Eggert et al. (2003) derived another estimator with a similar equation; $y = a(1 - e^{bx})$. These are referred to as the Kohn and Eggert methods, respectively. We analyzed data with the software package GIMLET version 1.3.1 (Valière, 2002), with 1000 random iterations of the genotype sampling order. Rarefaction equations were run using R software (version 1.7.1; available at http://www.r-project.org). Confidence intervals were calculated using the iterative approach, which is usually employed for rarefaction curves. However, this gives an indication of only the sampling variance and not the estimator variance.

2.3.6. *Investigating the genetic signature of the bottleneck*

At selectively neutral loci, populations that have experienced a recent reduction of their effective population size exhibit a characteristic mode-shift distortion in the distribution of allele frequencies (alleles at low frequency (<0.1) becoming less abundant; Luikart et al., 1998) and develop heterozygosity excess (Cornuet and Luikart, 1996). We used a Bayesian approach to detect and date a potential bottleneck in the Deosai bear population. This method is implemented in the MSVAR program (Beaumont, 1999) available at http://www.rubic.rdg.ac.uk/~mab. MSVAR calculates the Bayesian posterior distribution of demographic and mutational parameters, using a Markov Chain Monte Carlo approach. Mutations are assumed to occur under a stepwise mutation model with a rate $\theta = 2N_0\mu$, where μ is the locus mutation rate; the change in population size is assumed linear or exponential. The model assumes demographic history in a single stable population that was of size N_1 t_a generations ago and subsequently changed gradually in size to N_0 over the period from t to the current time. The program estimates two demographic parameters $t_f = t_a/N_0$ and $r = N_0/N_1$, where r indicates the population trend (population expansion if $r > 1$; population decline if $r < 1$).

For calculations we used the exponential growth models with the default parameters, as it is more suitable than the linear growth model for modeling population changes over a shorter time scale (Beaumont, 1999). For each population, 2×10^8 updates were calculated and only the last 90% of the chains were used. The model was run twice to test the general

stability of the solution from the Markov chain. In addition, we estimated the time since the population had started to decline (t_a) with $t_a = t_f * N_0$ and N_0 corresponding to the estimated population size, as well as the ancestral population size (before the decline), with $N_1 = N_0/r$.

2.3.7. Estimating nuclear DNA diversity, Hardy Weinberg equilibrium and linkage disequilibrium

Based on the 15 loci genotypes, we ran population genetic analyses using the softwares GENEPOP version 3.4 (Raymond and Rousset, 1995) and GENETIX version 4.02 (Belkhir et al., 1996–2004). Nuclear genetic diversity was measured as the number of alleles per locus (A), the observed heterozygosity (Ho), as well as Nei's unbiased expected heterozygosity (He) (Nei, 1978). Deviations from Hardy–Weinberg equilibrium were tested using an exact test. For loci with more than four alleles, a Markov chain was used to obtain an unbiased estimate of the exact probability. The Markov chain was set to 100 batches, with 5000 iterations per batch and 10 000 steps of dememorization. Global tests across loci for heterozygote deficiency and heterozygote excess and pairwise tests for linkage disequilibrium were performed using Fisher's method (Sokal and Rohlf, 1994) with 10,000 batches and 10,000 iterations per batch.

2.3.8. Comparing genetic diversity with other brown bear populations

We compared the genetic diversity of the Deosai population with the one from other documented bear populations in Europe and North America (A, Ho and He when available). However the values given in the literature cannot be compared directly with our data as they do not represent the same number of individuals and the same set of loci. Consequently, we took the opportunity of having the whole dataset from the Scandinavian brown bear population (Bellemain, 2004) for a comparison based on the same number of individuals and the same loci. A random selection of 28 bears, in each of the 3 subpopulations of the Scandinavian genetic dataset (M, N and S; Waits et al., 2000), was repeated 1000 times to estimate genetic diversity (A, He, Ho) and compare it with the corresponding values in the Deosai population.

2.3.9. Assessing relatedness

Based on the 15 loci genotypes of the different individuals identified in the population, we calculated pairwise genetic relatedness between pairs of individuals using Wang's estimator (Wang, 2002) and the software SPAGeDi version 1.0 (Hardy and Vekemans, 2002). This estimator includes (1) low sensitivity to the sampling error that results from estimating population allele frequencies; and (2) a low sampling variance that decreases asymptotically to the theoretical minimum with increasing numbers of loci and alleles per locus (Blouin, 2003). Relatedness values range from 1 to −1, indicating the percentage of alleles shared among individuals. Theoretically, a value of 1 means that genotypes are identical; a value of 0.5 indicates that 50% of the alleles are shared (e.g. parent/offspring or siblings relationship). Unrelated individuals have relatedness values ranging from 0 to −1 with the more negative values indicating greater differences in the genotypes of the individuals. We also used the genetic dataset for the Scandinavian subpopulations (M, N and S) to compare the level of pairwise relatedness between the Deosai population and those 3 subpopulations (using the same loci).

3. Results

3.1. Individual identification, probability of identity and reliability of the data

Totally, 136 samples were collected and 63 (~46%) of those samples were successfully amplified for 4–7 loci (including the sex locus). Twenty-three samples were from females, 37 from males and the sex could not be determined for three samples.

The data were judged to be reliable due to a high global quality index among successfully amplified samples (Fig. 2). Nine samples were discarded for further analysis due to their low quality index (below 0.5; Fig. 2). Finally, 54 samples typed for 6–7 loci were considered. Among those 54 samples, 28 individual genotypes were obtained (16 males, 10 females and 2 individuals of unknown sex). Each multilocus genotype was found from 1 to 5 times, with a mean of 2.22 ± 1.08 (SE) times. One sample for each of the 28 genetically identified individuals was further typed with 9 more microsatellites. The mean quality index per sample was 0.85 ± 0.13 for the 54 samples typed using 6 microsatellite loci and 0.91 ± 0.10 for the samples typed using 15 microsatellite loci.

Age of the feces was estimated for all but 11 samples. There was a significant negative correlation between the age of fecal samples and the proportion of positive amplification (Spearman's $\rho = -0.279$; $p = 0.01$) (Fig. 3) as well as between the age of fecal samples and the quality index (Spearman's $\rho = -0.271$; $p = 0.02$).

The probability of identity among the six amplified microsatellite loci for unrelated individuals was 1.881e−05 and 1.206e−02 for related individuals (sibs), thus we could identify each individual reliably. The probability of identity among the 15 amplified microsatellite loci unrelated individuals was 5.827e−10 and 1.329e−04 for related individuals. This allowed us to perform reliable parentage and relatedness analyses.

3.2. Estimating current population size

The population size estimates varied depending on the rarefaction equation used. The Kohn's estimate yielded a population size of 47 bears (95% CI: 33–102), whereas the Eggert's estimate gave a size of 32 bears (95% CI: 28–58).

3.3. Investigating the signature and age of the bottleneck

The analyses of the population's expansion and decline using MSVAR, based on the exponential growth model (Beaumont, 1999) gave the following values: $\log_{10}(r) = -2.423$, $\log_{10}(t_f) = 0.297$, $\log_{10}(\theta) = -1.410$. The low r value ($r < 1$) implies that the original population size declined to current population size. Considering the mean population size estimates for each rarefaction equation (see above), the number of generations

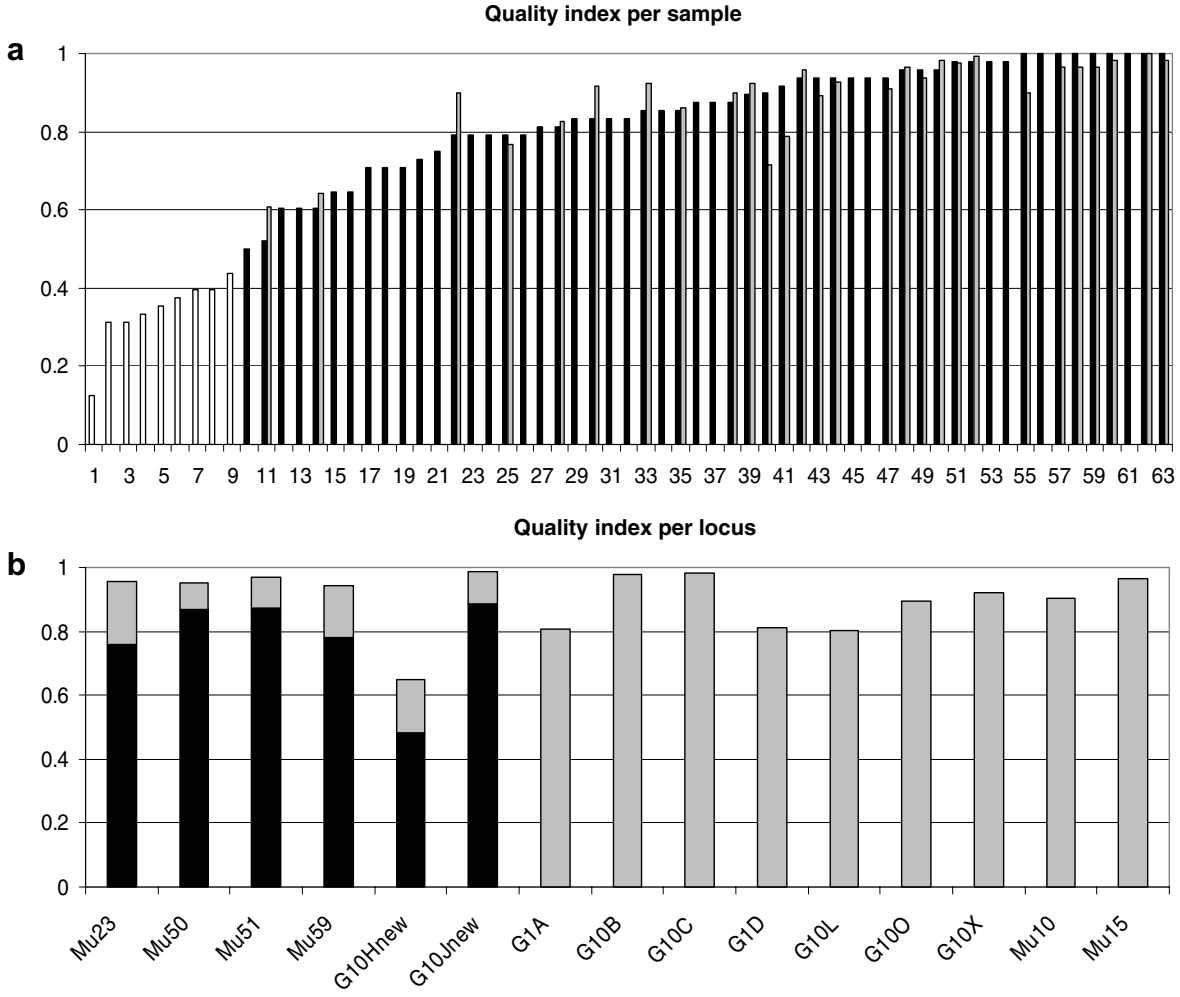


Fig. 2 – Quality indices (QI) per sample (a) and per locus (b) for successfully amplified genetic samples from brown bears in Deosai National Park, Pakistan. Black bars indicate QI for samples typed with 6 loci (for individual identification), grey bars indicate QI for samples typed with 15 loci (further analysis) and white bars indicate samples discarded from the analysis (because of their low QI).

since the population started to decline (t_a) was estimated to be between 63 and 93 and the ancestral population size ($N1$) ranged from 8000 to 11,750 individuals.

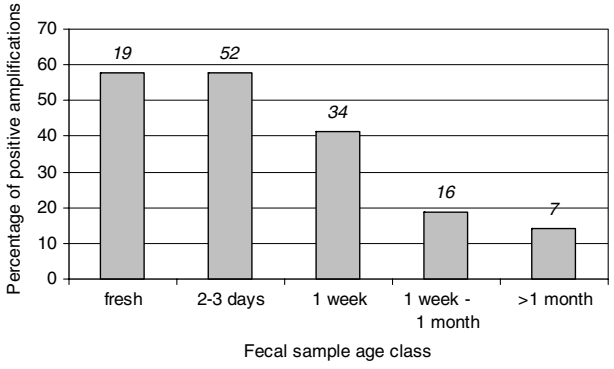


Fig. 3 – Success of brown bear fecal DNA amplifications from Deosai National Park, Pakistan, according to the age class of the fecal samples. Numbers above the bars, represent the sample size of each age class.

3.4. *Nuclear DNA diversity, Hardy–Weinberg equilibrium and linkage disequilibrium*

The number of alleles per locus among the 28 individual genotypes ranged from 2 to 7, with an average of 3.87 ± 1.36 (Table 1). The mean observed heterozygosity was 0.557, a value not significantly different from the unbiased expected heterozygosity (0.548). Global tests showed that the population is in Hardy–Weinberg equilibrium, although three loci (G10L, G10O, Mu10) had a significant deficiency in heterozygotes at the $p < 0.05$ level (Table 1). The overall multilocus Fis value was −0.016. Statistical tests for linkage disequilibrium were computed for all pairs of loci, and 15 of 105 tests revealed significant results ($p < 0.05$).

3.5. *Comparing genetic diversity with other bear populations*

The level of heterozygosity in the Deosai bear population ($Ho = 0.557$) was lower than in other bear populations in North America that are considered to have a good conservation sta-

Table 1 – Nei's unbiased expected (He) and observed (Ho) heterozygosities, and deviation from Hardy Weinberg equilibrium by locus from fecal samples of brown bears from Deosai National Park, Pakistan

Locus	Alleles	Allelic frequencies	He	Ho	P
Mu23	136	0.232	0.770	0.893	
	140	0.339			
	144	0.161			
	146	0.054			
	150	0.214			
Mu50	92	0.643	0.541	0.571	
	94	0.125			
	96	0.036			
	100	0.196			
G10B	136	0.382	0.466	0.518	
	150	0.618			
Mu59	95	0.25	0.830	0.857	
	109	0.196			
	111	0.054			
	113	0.089			
	115	0.036			
	117	0.214			
	119	0.161			
G10Jpak	80	0.518	0.656	0.678	
	84	0.089			
	86	0.232			
	88	0.161			
G1D	171	0.17	0.642	0.679	
	175	0.038			
	177	0.302			
	179	0.491			
Mu51	119	0.714	0.425	0.50	
	121	0.268			
	127	0.018			
G10Hpak	241	0.442	0.602	0.76	
	243	0.115			
	245	0.423			
	249	0.019			
G1A	189	0.593	0.496	0.5	
	191	0.019			
	193	0.389			
G10C	104	0.4	0.492	0.518	
	108	0.6			
G10L	143	0.204	0.773	0.583	0.009
	155	0.224			
	157	0.286			
	159	0.265			
	163	0.02			
G10O	193	0.019	0.037	0.037	
	195	0.981			
G10X	142	0.849	0.281	0.115	0.023
	154	0.057			
	156	0.057			
	158	0.038			
Mu10	140	0.094	0.656	0.5	0.0002
	142	0.057			
	150	0.019			
	152	0.434			
	154	0.396			

(*continued on next page*)

Table 1 – *continued*

Locus	Alleles	Allelic frequencies	He	Ho	P
Mu15	137	0.018	0.527	0.556	
	139	0.473			
	141	0.509			
Average			0.548	0.557	

Only significant *P*-values are shown ($P < 0.05$).

tus (*Ho* = 0.78 in North America; Paetkau et al., 1998 and *Ho* = 0.66–0.76 in different regions of Canada and USA; Waits et al., 1998). However, it is comparable to the level of heterozygosity in the Yellowstone area (*Ho* = 0.55; Paetkau et al., 1998) and higher than the level observed in some isolated populations such as the Kodiak Islands in Alaska (*Ho* = 0.26; Paetkau et al., 1998) or the Pyrenees in France (*Ho* = 0.39; Taberlet et al., 1997).

In comparison with each of the three subpopulations in Scandinavian bears, Deosai bears had a significantly lower number of alleles and observed and unbiased expected heterozygosity (for the same number of individuals and loci subsampled; Table 2). When compared to the mean genetic characteristics in the entire Scandinavia, the expected heterozygosity in the Deosai population is reduced by 17.5% and the number of alleles per locus by 44%.

3.6. *Assessing relatedness*

The average pairwise relatedness in the Deosai bear population was 0.0265 ± 0.292 (SE). This was not significantly different from the average pairwise relatedness in the subpopulations of the Scandinavian bears for the same loci (paired *t*-tests for each subpopulation: N: $r = -0.0232 \pm 0.044$; $p = 0.231$; S: $r = 0.015 \pm 0.044$; $p = 0.206$; M: $r = -0.001 \pm 0.032$; $p = 0.052$).

4. Discussion

4.1. *Quality of the genetic data*

We ensured a high reliability of the genetic data by repeating amplifications (multi-tubes approach) and selecting samples with high quality indices. The probability of misidentification was low, allowing us to identify unambiguously each individual. Therefore, we are confident that we have not overestimated the number of individuals in the fecal sampling.

The amplification success was correlated negatively with the age of fecal samples. Amplification success was relatively good (~58%) for fresh feces or feces that were only 2–3 days old and dropped to 41% for 1 week old samples, but this rate might still be acceptable. However, samples older than one week had a poor amplification success. We recommend, for future studies in Deosai, that fecal samples older than one

Table 2 – Comparison of the genetic diversity of brown bears between the Deosai population in Pakistan and the three subpopulations in the Scandinavian genetic dataset (mean over 28 randomly and repeatedly chosen individual bears)

	Pakistan			Scandinavia South			Scandinavia Middle			Scandinavia North		
	A	He	Ho	A	He	Ho	A	He	Ho	A	He	Ho
Mu23	5	0.77	0.89	7	0.70	0.73	7	0.82	0.83	6	0.72	0.70
Mu50	4	0.54	0.57	7	0.74	0.72	7	0.79	0.76	9	0.71	0.69
Mu51	3	0.43	0.50	7	0.78	0.80	8	0.77	0.75	8	0.76	0.74
Mu59	7	0.83	0.86	10	0.76	0.77	11	0.83	0.86	11	0.83	0.83
G10Jnew	4	0.66	0.68	6	0.57	0.58	6	0.66	0.66	7	0.75	0.75
G10Hnew	4	0.61	0.76	8	0.59	0.58	8	0.53	0.47	11	0.74	0.74
G1A	3	0.51	0.50	6	0.63	0.69	5	0.71	0.70	7	0.67	0.63
G1D	4	0.64	0.77	7	0.61	0.59	5	0.66	0.65	8	0.74	0.79
G10B	2	0.48	0.52	5	0.69	0.68	8	0.64	0.69	8	0.74	0.70
G10C	2	0.49	0.52	5	0.69	0.66	5	0.67	0.69	6	0.68	0.68
G10L	5	0.77	0.58	7	0.77	0.79	7	0.69	0.63	8	0.81	0.74
G10O	2	0.04	0.04	3	0.38	0.38	3	0.36	0.36	3	0.12	0.12
G10X	4	0.28	0.12	4	0.54	0.56	5	0.65	0.62	7	0.54	0.53
Mu10	5	0.66	0.50	8	0.80	0.79	8	0.74	0.75	8	0.78	0.75
Mu15	3	0.53	0.56	4	0.66	0.66	4	0.53	0.50	5	0.51	0.52
Mean	3.80	0.55	0.56	6.27	0.66	0.67	6.47	0.67	0.66	7.47	0.67	0.66
SD	1.37	0.20	0.24	2.07	0.11	0.11	2.07	0.13	0.13	2.07	0.18	0.17
P-values				6.82e−07	0.0121	0.059	1.02e−05	0.008	0.065	6.98e−07	0.0006	0.0161

P-values represent the significance of paired *t*-tests performed between the Pakistan population and each of the three Scandinavian subpopulations.

week not be collected in order to optimize the cost and benefit of the genetic analyses.

Brown bears in Deosai are mainly vegetarians (Schaller, 1977; unpublished data of fecal analysis). Previous studies have suggested that plant secondary compounds can inhibit PCRs (Huber et al., 2002). However, this study demonstrated that reasonable brown bear DNA amplification can be obtained from fecal samples composed mainly of plants (Murphy et al., 2003).

4.2. *The genetic status of the brown bear population in Deosai*

The analyses performed from the fecal DNA dataset allowed us to answer important questions regarding the management and conservation of bears in the Deosai population. First, the population size estimates provided by the two rarefaction indices are in the same order of magnitude as the numbers derived from the field censuses, which gives us confidence that those results are realistic. The census carried out during summer 2004 recorded 38 bears from the Deosai National Park, with a density of 19 bears per 1000 km^2 area (Nawaz et al., 2006). Based on this, the Eggert method seemed to underestimate the population size, whereas Kohn's method seemed to be more realistic, although the upper limit of the confidence intervals seems to be an overestimate. Unfortunately, the small sample size and small number of recaptures prevented us from using the MARK method, which is thought to give better estimates of population sizes (Bellemain et al., 2005). Considering the minimum number of individuals captured from the fecal samples (28) and the rarefaction method estimates, the field estimates appear to be conservative, though they fall within the range of the other estimates. Field methods usually give underestimates of wild populations, particularly for elusive animals (Solberg et al., 2006). The open terrain of the Deosai plateau, which allows bears to be observed, the small population size, distinctive marks on many bears, and the expertise that the field staff had gained over a period of 12 years from observing bears, probably contributed to the realistic observation-based estimates in Deosai National Park. We conclude that approximately 40–50 bears were present in the park in 2004.

The results from the analysis using the program MSVAR suggested that a decline in the Deosai population occurred approximately 63–93 generations ago using the mean estimates given by the rarefaction analysis and 80–100 generations ago, using a more realistic population size of 40–50 individuals. This period approximately corresponds to 800–1000 years ago, with a generation time of 10 years (calculated using the software RAMAS, Ferson and Akçakaya, 1990 and considering an age of first reproduction of 6 years old). The ancestral population (before the decline; N1) was estimated to contain 8000–11,750 individuals using rarefaction estimates or 10,000–12,500 individuals using a more realistic population size of 40–50 individuals. This estimate seems realistic considering an approximate area of 200,000 km^2 of bear distribution range in northern Pakistan and Kashmir, which gives a density of about 55 bears per 1000 km^2. These results suggest that the brown bear population in northern Pakistan might have undergone an approximate 200–300-fold decrease during the last thousand years. This decline cannot be linked to a single event or phenomenon. It was probably affected by both natural (climatic and geological) and socio-political factors. In the medieval warm period (1000–1200 AD), the bears certainly formed a single, large population, with a contiguous habitat in Hindu Kush, Karakoram and Western Himalaya ranges. The historic phase of glaciations in High Asia identified as a "little ice age" (1180–1840 AD; Kuhle, 1997; Esper et al., 2002; Mackay et al., 2005) is considered to have been similar in extent to the Neogeological stages (Meiners, 1997) and may have acted as a proximal cause of decline, destroying part of the population and fragmenting the rest. The influence of a growing human population, including large deforestation in the Middle Ages (Bertrand et al., 2002), political unrest due to presence of the Tibetan army in the area and its clashes with local people and China (Sheikh, 1998; Rashid S, personal communication) and the spread of firearms in the late 19th century, probably contributed further to the population decline and did not allow bears to colonize in a natural way.

Third, we assessed whether the Deosai population is currently at risk of inbreeding depression. The population genetics analyses revealed that the level of nuclear genetic diversity of the Deosai population is globally lower than brown bear populations considered to have a good conservation status, such as in Scandinavia or North America. In addition, and for the first time, we made an unbiased comparison of nuclear diversity between two populations, based on the same loci and same number of individuals. This analysis supports the conclusion that the Deosai population harbors significantly less heterozygosity and a smaller number of alleles per locus than any of the three subpopulations in Scandinavia. However, this population is in Hardy Weinberg equilibrium and its level of relatedness is similar to that in the Scandinavian brown bear population. Therefore, the Deosai bear population does not appear to be at immediate risk of inbreeding depression. Its level of genetic diversity is comparable to the brown bear population in the Yellowstone area, USA, which has become an isolated remnant, separated from other brown bears for nearly a century (Paetkau et al., 1998). A similar scenario could be envisaged for the Deosai brown bear, which probably lost genetic diversity due to isolation and genetic drift in the last centuries and due to the currently small population size.

Our final goal was to examine the degree of isolation of the Deosai population. Four individuals in our genetic dataset showed private alleles at two different loci, suggesting that they could be migrants (or descendants from migrants) from outside of the study area. Field observations support this hypothesis. Brown bears also exist in the Minimerg and Astore valleys, which are adjacent to Deosai National Park. Movements of bears have been observed between these areas during recent surveys, and the Deosai population may have interchanged not only with bears in these valleys, but also with the bear populations in the Neelam Valley and in Indian Kashmir through these valleys (unpublished data). When we began our studies of the Deosai brown bear population, we had expected to find genetic loss due to isolation and a small population; however, we documented a moderate level of genetic diversity. This strongly suggests that connectivity exists

between the Deosai population and the neighboring populations through movements of individuals.

4.3. *Conclusions and recommendations*

We have documented that the Deosai brown bear population shows moderate levels of diversity and is not at immediate risk of inbreeding. The population probably began to lose genetic diversity about 1000 years ago, when it began to decline from a single large population throughout northern Pakistan. This resulted in fragmentation of the population into smaller units that lost connectivity during the course of time. The population decline stopped in Deosai about 15 years, ago when the population received increased protection. Under a scenario of an isolated population, the population would probably suffer from inbreeding today. Therefore, we believe that the moderate level of genetic diversity observed has been maintained by gene flow with adjacent populations in Pakistan and India. Nevertheless, this level of genetic diversity is lower than in healthy populations in Europe or North America. Maintaining and improving the connectivity with adjacent populations in Pakistan and India will be of paramount importance for the long-term survival of this small population in future.

We suggest that future studies continue to monitor the population carefully, both with field observations and genetic analyses. Concrete management actions should aim at maintaining and improving connectivity with other populations to maintain or improve levels of genetic diversity. Otherwise, the population will continue to lose genetic diversity over time. Increasing the size and range of fecal sampling would not only allow a more precise estimate of the population size, but also give a better estimate of incoming gene flow.

Acknowledgements

The fieldwork for this study was supported by the Himalayan Wildlife Foundation (HWF), Islamabad. All field staff of the HWF, particularly M. Yunus, G. Murtaza and A. Khan, helped in collection of fecal samples. The cost of lab analysis was funded by the International Bear Association (John Sheldon Bevins Memorial Foundation), Norwegian University of Life Sciences, and Laboratoire d'Ecologie Alpine (LECA), University Joseph Fourier, France. We are grateful to L. Gielly, D. Rioux, C. Miquel and C. Poillot for their help in the laboratory work. Special thanks to Vaqar Zakaria and Dr. Anis ur Rahman, coordinators of HWF, for their support and suggestions during the study.

REFERENCES

Beaumont, M., 1999. Detecting population expansion and decline using microsatellites. Genetics 153, 2013–2029.

Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N., Bonhomme, F., 1996–2004. GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier (France).

Bellemain, E., 2004. Genetics of the Scandinavian brown bears in Scandinavia, implications for biology and management. Ph.D. thesis, Department of Biology and Natural Resource Management, As, Norway and Laboratoire d'Ecologie Alpine, Grenoble, France. 243pp. ISBN 82-575-0624-9.

Bellemain, E., Taberlet, P., 2004. Improved non-invasive genotyping method: application to brown bear (*Ursus arctos*) feces. Molecular Ecology Notes 4, 519–522.

Bellemain, E., Swenson, J.E., Tallmon, D.A., Brunberg, S., Taberlet, P., 2005. Estimating population size of elusive animals using DNA from hunter-collected feces: comparing four methods for brown bears. Conservation Biology 19, 150–161.

Bertrand, C.D., Loutre, M.F., Crucifix, M., Berger, A., 2002. Climate of the last millennium: a sensitivity study. Tellus – Series A – Dynamic Meteorology and Oceanography 54, 221–245.

Blouin, M.S., 2003. DNA-based methods for pedigree reconstructions and kinship analysis in natural populations. Trends in Ecology and Evolution 18, 503–511.

Bonnell, M.L., Selander, R.K., 1974. Elephant seals: genetic variation and near extinction. Science 184, 908–909.

Cornuet, J.M., Luikart, G., 1996. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics 144, 2001–2014.

Eggert, L.S., Eggert, J.A., Woodruff, D.S., 2003. Estimating population sizes for elusive animals: the forest elephants of Kalum National Park, Guana. Molecular Ecology 12, 1389–1402.

Esper, J., Schweingruber, F.H., Winiger, M., 2002. 1300 years of climatic history for Western Central Asia inferred from tree-rings. Holocene 12, 267–278.

Ferson, S., Akçakaya, H.R., 1990. RAMAS/Age, User Manual, Modeling Fluctuations in Age-Structured Populations. Exeter Software, Setauket, New York.

Frankham, R., Ballou, J.D., Briscoe, D.A., 2002. Introduction to Conservation Genetics. Cambridge University Press, Cambridge, UK.

Groombridge, J.J., Jones, G.G., Bruford, M.W., Nichols, R.A., 2000. 'Ghost' alleles of the Mauritius kestrel. Nature 403, 616.

Hardy, O.J., Vekemans, X., 2002. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Notes 2, 218–620.

Hedrick, P.W., 1992. Genetic conservation in captive populations and endangered species. In: Jain, S.K., Botsford, L.W. (Eds.), Applied Population Biology. Kluwer, Dordrecht, Netherlands, pp. 45–68.

Himalayan Wildlife Foundation, 1999a. Management Plan for Deosai National Park Northern Areas Pakistan. Himalayan Wildlife Foundation, Islamabad.

Himalayan Wildlife Foundation. 1999b. Deosai brown bear project, 1998. Final Report, Islamabad. Himalayan Wildlife Foundation, Islamabad.

Himalayan Wildlife Project. 1994. Deosai Brown Bear Project, 1994. Annual Report, Islamabad. Himalayan Wildlife Foundation, Islamabad.

Huber, S., Bruns, U., Arnold, W., 2002. Sex determination of red deer using polymerase chain reaction of DNA from feces. Wildlife Society Bulletin 30, 208–212.

Kohn, M.H., York, E.C., Kamradt, D.A., Haught, G., Sauvajot, R.M., Wayne, R.K., 1999. Estimating population size by genotyping feces. Proceedings of the Royal Society of London B Biological Sciences 266, 657–663.

Kuhle, M., 1997. New findings concerning the ice age (last glacial maximum) glacier cover of the East-Pamir, of the Nanga Parbat up to the Central Himalaya and of Tibet, as well as the age of the Tibetan inland ice. GeoJournal 42, 87–257.

Luikart, G., Allendorf, F.W., Cornuet, J.M., Sherwin, W.B., 1998. Distortion of allele frequency distributions provides a test for recent population bottlenecks. Journal of Heredity 89, 238–247.

Mackay, A.W., Ryves, D.B., Battarbee, R.W., Flower, R.J., Jewson, D., Rioual, P., Sturm, M., 2005. 1000 years of climate variability in central Asia: assessing the evidence using Lake Baikal (Russia) diatom assemblages and the application of a diatom-inferred model of snow cover on the lake. Global and Planetary Change 46, 281–298.

Meiners, S., 1997. Historical to post glacial glaciation and their differentiation from the Late Glacial Period on examples of the Tian Shan and the N.W. Karakorum. GeoJournal 42, 259–302.

Miquel, C., Bellemain, E., Poillot, C., Bessière, J., Durand, A., Taberlet, P., 2006. Quality indices to assess genotypes reliability for studies using non-invasive sampling and multiple-tube approach. Molecular Ecology Notes. doi:10.1111/j.1471-8286.2006.01413.

Murphy, M., Waits, L., Kendall, K., 2003. Impact of diet on faecal DNA amplification and sex identification brown bears (*Ursus arctos*). Molecular Ecology 12, 2261–2265.

Nawaz, M.A., Shah, M., Zakaria, V., 2006. Environmental Baseline of Deosai National Park Draft Report. Himalayan Wildlife Foundation, Islamabad.

Nei, M., 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89, 583–590.

Paetkau, D., Strobeck, C., 1994. Microsatellite analysis of genetic variation in black bear populations. Molecular Ecology 3, 489–495.

Paetkau, D., Calvert, W., Stirling, I., Strobeck, C., 1995. Microsatellite analysis structure of population structure in Canadian polar bears. Molecular Ecology 4, 347–354.

Paetkau, D., Waits, L.P., Clarkson, P., Craighead, L., Vyse, E., Ward, R., Strobeck, C., 1998. Dramatic variation in genetic diversity across the range of North American brown bears. Conservation Biology 12, 418–429.

Primack, R.B., 2002. Essentials of Conservation Biology. Sinauer Associates, Sunderland, Mass.

Raymond, M., Rousset, F., 1995. Genepop (version 1.2), population genetics software for exact tests and ecumenicism. Journal of Heredity 86, 248–249.

Roberts, T.J., 1997. The Mammals of Pakistan. Oxford University Press, New York.

Schaller, G.B., 1977. Mountain Monarchs: Wild Sheep and Goats of the Himalaya. The University of Chicago Press, Chicago and London.

Servheen, C., 1990. The status and conservation of the bears of the world. International Association for Bear Research and Management Monograph Series No. 2.

Servheen, C., Herrero, S., Peyton, B., 1999. Status Survey and Conservation Action Plan for BearsIUCN/SSC Bear and Polar Bear Specialist Groups. IUCN, Gland, Switzerland and Cambridge, UK.

Sheikh, A.G., 1998. Ladakh and Baltistan through the Ages. In: Stellrecht, I. (Ed.), Karakpram–Hindukush–Himalaya: Dynamics of Change. Rudiger Koppe Verlag Koln, Germany.

Sokal, R.R., Rohlf, F.J., 1994. Biometry: The Principles and Practice of Statistics in Biological Research, third ed. W.H. Freeman, New York, NY.

Solberg, H., Bellemain, E., Drageset, O.M., Taberlet, P., Swenson, J.E., 2006. An evaluation of field and genetic methods to estimate brown bear (*Ursus arctos*) population size. Biological Conservation 128, 158–168.

Soulé, M.E. (Ed.), 1987. Viable Populations for Conservation. Cambridge University Press, Cambridge, UK.

Taberlet, P., Griffin, S., Goossens, B., Questiau, S., Manceau, V., Escaravage, N., Waits, L.P., Bouvet, J., 1996. Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Research 24, 3189–3194.

Taberlet, P., Camarra, J.J., Griffin, S., Hanotte, O., Waits, L.P., Dubois-Paganon, C., Burke, T., Bouvet, J., 1997. Noninvasive genetic tracking of the endangered Pyrenean brown bear population. Molecular Ecology 6, 869–876.

Taberlet, P., Waits, L.P., Luikart, G., 1999. Noninvasive genetic sampling: look before you leap. Trends in Ecology and Evolution 14, 323–327.

Valière, N., 2002. Gimlet: a computer program for analyzing genetic individual identification data. Molecular Ecology Notes 2, 377.

Waits, L.P., Talbot, S., Ward, R.H., Shields, G., 1998. Phylogeography of the North American brown bear and implications for conservation. Conservation Biology 12, 408–417.

Waits, L.P., Taberlet, P., Swenson, J.E., Sandegren, F., 2000. Nuclear DNA microsatellite analysis of genetic diversity and gene flow in the Scandinavian brown bear (*Ursus arctos*). Molecular Ecology 9, 421–431.

Waits, L.P., Luikart, G., Taberlet, P., 2001. Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. Molecular Ecology 10, 249–256.

Wang, J., 2002. An estimator for pairwise relatedness using molecular markers. Genetics 160, 1203–1215.

Woods, C.A., Kalpatrick, W.C., Rafique, M., Shah, M., Khan, W., 1997. Biodiversity and conservation of the Deosai Plateau, Northern areas, Pakistan. In: Mufti, S.A., Woods, C.A., Hasan, S.A. (Eds.), Biodiversity of Pakistan. Pakistan Museum of Natural History, Islamabad and Florida Museum of Natural History, Gainesville, pp. 33–62.

Paper II

Published online 14 December 2006 *Nucleic Acids Research, 2007, Vol. 35, No. 3* ***e14***
doi:10.1093/nar/gkl938

Power and limitations of the chloroplast *trn*L (UAA) intron for plant DNA barcoding

Pierre Taberlet[1,*], Eric Coissac[2,3], François Pompanon[1], Ludovic Gielly[1], Christian Miquel[1], Alice Valentini[1,4,5], Thierry Vermat[6], Gérard Corthier[7], Christian Brochmann[8] and Eske Willerslev[9]

[1]Laboratoire d'Ecologie Alpine, CNRS UMR 5553, Université Joseph Fourier, BP 53, 38041 Grenoble Cedex 9, France, [2]Laboratoire Adaptation et Pathogénie des Microorganismes, CNRS UMR 5163, Université Joseph Fourier, BP 170, 38042 Grenoble Cedex 9, France, [3]INRIA Rhône-Alpes, Hélix Project, 655 Avenue de l'Europe, 38334 Montbonnot Cedex, France, [4]Dipartimento di Ecologia e Sviluppo Economico Sostenibile, Università degli Studi della Tuscia, via S. Giovanni Decollato 1, 01100 Viterbo, Italy, [5]Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, PO Box 5003, No-1432 Ås, Norway, [6]Bioinformatics, GENOME Express, 11 Chemin des Prés, 38944 Meylan, France, [7]UR 910 Ecologie et Physiologie du Système Digestif, INRA Domaine de Vilvert, 78352 Jouy-en-Josas Cedex, France, [8]National Centre for Biosystematics, Natural History Museum, University of Oslo, PO Box 1172 Blindern, NO-0318 Oslo, Norway and [9]Center for Ancient Genetics, Niels Bohr Institute & Biological Institutes, University of Copenhagen, Juliane Maries vej 30, DK-2100 Copenhagen, Denmark

Received June 29, 2006; Revised September 21, 2006; Accepted October 16, 2006

ABSTRACT

DNA barcoding should provide rapid, accurate and automatable species identifications by using a standardized DNA region as a tag. Based on sequences available in GenBank and sequences produced for this study, we evaluated the resolution power of the whole chloroplast *trn*L (UAA) intron (254–767 bp) and of a shorter fragment of this intron (the P6 loop, 10–143 bp) amplified with highly conserved primers. The main limitation of the whole *trn*L intron for DNA barcoding remains its relatively low resolution (67.3% of the species from GenBank unambiguously identified). The resolution of the P6 loop is lower (19.5% identified) but remains higher than those of existing alternative systems. The resolution is much higher in specific contexts such as species originating from a single ecosystem, or commonly eaten plants. Despite the relatively low resolution, the whole *trn*L intron and its P6 loop have many advantages: the primers are highly conserved, and the amplification system is very robust. The P6 loop can even be amplified when using highly degraded DNA from processed food or from permafrost samples, and has the potential to be extensively used in food industry, in forensic science, in diet analyses based on feces and in ancient DNA studies.

INTRODUCTION

DNA barcoding is a relatively new concept (1,2), aiming to provide rapid, accurate and automatable species identifications by using a standardized DNA region as a tag (3). As recently pointed out by Chase *et al*. (4), there are two categories of potential DNA barcode users: taxonomists and scientists in other fields (e.g. forensic science, biotechnology and food industry, animal diet).

According to the current technology, the ideal DNA barcoding system should meet the following criteria. First, it should be sufficiently variable to discriminate among all species, but conserved enough to be less variable within than between species. Second, it should be standardized, with the same DNA region as far as possible used for different taxonomic groups. Third, the target DNA region should contain enough phylogenetic information to easily assign species to its taxonomic group (genus, family, etc.). Fourth, it should be extremely robust, with highly conserved priming sites, and highly reliable DNA amplifications and sequencing. This is particularly important when using environmental DNA where each extract contains a mixture of many species to be identified at the same time. Fifth, the target DNA region should be short enough to allow amplification of degraded DNA. Unfortunately, such an ideal DNA marker does not exist. However, for different category of users (i.e. taxonomists versus scientists in other fields), the five criteria listed above will not be equally important. For example, a high level of variation with sufficient phylogenetic information will be most important for taxonomists. In contrast, the levels

*To whom correspondence should be addressed. Tel: +33 476 51 45 24; Fax: +33 476 51 42 79; Email: pierre.taberlet@ujf-grenoble.fr

of standardization and robustness will be most important in forensics or when analyzing processed food.

So far, methodological papers published on DNA barcoding have typically dealt with the most suitable region of the genome according to the taxonomists' point of view [e.g. Ref. (5–7)]. In animals, the 5′ fragment of the mitochondrial gene for the cytochrome oxidase subunit I (*COI* or *COXI*) represents a good candidate [e.g. Ref. (5,8,9)]. However, there is no consensus in the scientific community, and 16S rRNA, another mitochondrial gene, or the nuclear ribosomal DNA have also been proposed as useful barcoding markers (7,10). In plants, the situation is much more difficult, because both the mitochondrial and chloroplast genomes are evolving too slowly to provide enough variation. For taxonomists, the current strategy is to sequence several DNA regions (4), including both nuclear and chloroplast fragments such as the internal transcribed spacer (ITS) region of the 18S–5.8S–26S nuclear ribosomal cistron (11) or the chloroplast *trn*H–*psb*A region (6).

In this study, we approach the plant DNA barcoding problem in another way, by emphasizing the point of view of scientists other than taxonomists, looking for standardized and robust methodologies. For this purpose, we must find a genome region as variable as possible, but bearing the possibility of designing highly conserved PCR primers that amplify a very short DNA region, of no more than 100–150 bp. Such a short region should allow reliable amplifications of even highly degraded DNA found in processed food or in fossil remains. Up to now, when working with substrates such as ancient DNA, the strategy has been to use primers based on the chloroplast *rbc*L gene (12), but this system only allows in most cases the identification of families, not genera or species.

The chloroplast *trn*L (UAA) intron may represent a good target region for our purpose. Its sequences have been widely used for reconstructing phylogenies between closely related species (13–15) or for identifying plant species (16,17). Nevertheless, it is widely recognized that it does not represent the most variable non-coding region of chloroplast DNA (18), but it bears some unique advantages. Universal primers for this region were designed ~15 years ago (19), and subsequently extensively used, mainly in phylogenetic studies among closely related genera and species (20). The evolution of the *trn*L (UAA) intron has been thoroughly analyzed and is well understood (21,22). Furthermore, this region is the only Group I intron in chloroplast DNA (23,24). This means that it has a conserved secondary structure (25,26) with alternation of conserved and variable regions (22). As a consequence, the alignment of diverse *trn*L intron sequences might allow the design of new versatile primers embedded in conserved regions and amplifying the short variable region in between.

More specifically, our objective in this paper is to evaluate the power and the limitations of the chloroplast *trn*L (UAA) intron for plant DNA barcoding, and to assess the possibility for designing a new system allowing species identification with highly degraded DNA.

MATERIALS AND METHODS

General strategy

The power and the robustness of the *trn*L intron for DNA barcoding were first evaluated with the data available in GenBank. Then, they were evaluated on two specific datasets by sequencing the whole intron for more than 100 plant species originating from the same environment, and by compiling sequences of the main plants used in the food industry. Finally, we tested the robustness of a new pair of internal primers applied on different substrates supposed to contain highly degraded DNA.

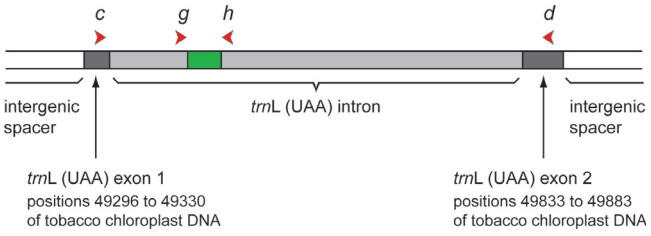


Figure 1. Position of the primers *c, d, g* and *h* on the chloroplast *trn*L (UAA) gene. The P6 loop amplified with primer *g* and *h* is indicated in green.

Table 1. Sequences of the two universal primer pairs amplifying the *trn*L (UAA) intron

Name	Code	Sequence 5′–3′
c	A49325	CGAAATCGGTAGACGCTACG
d	B49863	GGGGATAGAGGGACTTGAAC
g	A49425	GGGCAATCCTGAGCCAA
h	B49466	CCATTGAGTCTCTGCACCTATC

Length of the amplified fragment with primers *c*–*d* in tobacco: 456 bp. Length of the amplified fragment with primers *g*–*h* in tobacco: 40 bp. The code denotes the 3′-most base pairs in the published tobacco cpDNA sequence (23). Primers *c* and *d* are from Taberlet *et al*. (19). Primer *g* and *h* were designed for this study (France patent no 2 876 378; April 14, 2006).

Primer used

Figure 1 presents the location of the primers in the chloroplast *trn*L (UAA) gene, and Table 1 gives their sequences. The primers *c* and *d* are from Taberlet *et al*. (19). This fragment encompasses the entire *trn*L (UAA) intron plus a few base pairs on each side belonging to the *trn*L (UAA) gene itself. The primers *g* and *h* were designed for this study on two highly conserved regions after aligning various sequences, either from GenBank or produced earlier in the Grenoble laboratory.

The Arctic plant dataset

We analyzed 123 arctic plant samples collected between 1998 and 2003, partly taken from herbarium specimens and partly from field-collected, silica-dried leaf samples deposited at the Natural History Museum in Oslo. Total DNA was extracted from around 10 mg of dried leaf tissue with the DNeasy 96 Plant Kit (Qiagen), following the manufacturer's protocol. Double-stranded DNA amplifications were performed in volumes of 25 µl containing 2.5 mM $MgCl_2$, 200 µM of each dNTP, 1 µM of each primer and 1 U of AmpliTaq Gold® DNA polymerase (Applied Biosystems). The *trn*L (UAA) intron was amplified with primers *c* and *d* (19). Following an activation step of 10 min at 95°C for the enzyme (Applied Biosystems specification), the PCR mixture underwent 35 cycles of 30 s at 95°C, 30 s at 50°C and 2 min at 72°C on a GeneAmp PCR system 2720 (Applied Biosystems).

To remove excess primers and deoxynucleotide triphosphates after amplification, PCR products were purified on QIAquick PCR Purification Kit columns (Qiagen), according to the manufacturer's instructions. Sequencing was performed, on both strands, using the BigDye® Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems) in volumes of 20 µl containing 20 ng of purified DNA and 4 pmol of amplification primer, according to the manufacturer's specifications. Sequencing reactions underwent 25 cycles of 30 s at 96°C, 30 s at 50°C and 4 min at 60°C. Excess dye terminators were removed by a spin-column purification. Sequencing reactions were electrophoresed for 45 min on an ABI PRISM®3100 Genetic Analyzer (Applied Biosystems) using 36 cm capillaries and POP-4™ polymer.

The Food dataset

Seventy-two sequences of the main plants used in the food industry were retrieved from GenBank or sequenced following the previous protocol. For this analysis, we restricted our investigations to the short fragment amplified with the *g*–*h* primer pair.

Bioinformatic approach

PCR were simulated on the full plant division of GenBank download from NCBI server on the December 14, 2005 (ftp://ftp.ncbi.nlm.nih.gov/genbank). This release corresponds to 731 531 entries. The electronic PCR software (ePCR) was specially developed for this study. It is based on the agrep algorithm (27) that allows identifying occurrences of a small pattern (corresponding to a PCR primer) on a large text (genomic sequence) with a fixed maximum mismatch count. This strategy is more relevant than simple blast queries, which are not suitable to identify similarity on nucleic sequences when the query sequence (here oligonucleotide sequence) is too short. Our ePCR software allows specifying maximum mismatch count, minimum and maximum length of the amplified region and takes care to also retrieve taxonomic data from analyzed entries. It works on Genbank, EMBL or fasta formatted sequence files (in the latter case, taxonomic data must be encoded in a special format on the title line). The ePCR software is available for academic users upon e-mail request to Eric Coissac (eric.coissac@inrialpes.fr).

ePCR was realized on GenBank data, first with the *c* and *d* primers, second with the *g* and *h* primers, third on a short *rbc*L fragment with the *h1aF* and *h2aR* primers (12), and finally with eight primer pairs found in Shaw *et al*. (18). ePCR was also realized on the arctic plant dataset with the *c* and *d* primers (after adding the *c* and *d* sequences on each side of the sequenced PCR product), and with the *g* and *h* primers.

Next, amplicon databases constructed by the ePCR software were analyzed to extract taxonomic specificities of the amplified sequences. This analysis used the taxonomic classification provided by NCBI to assess taxonomic relationships between sequences. The main goal of this analysis was to determine the proportion of the species, genera and families unambiguously identified by the sequences amplified via ePCR. A taxon (species, genus or family) was defined as 'unambiguously identified' if all the sequences associated with this taxon are not found in any other taxa. To limit the influence of the taxonomic coverage of the GenBank database, we discarded genera represented by only one species and families represented by only one genus. The same measure of specificity was applied to the arctic plant dataset described above. We also assessed the intraspecific variation of the whole *trn*L intron and of the short P6 loop fragment by extracting, from the GenBank amplicon database constructed by the ePCR software, all the species represented by more than one entry.

Primer 'universality'

The universality of the four primers *c, d, g* and *h* was examined by comparing their sequences with homologous sequences, either from GenBank (for primers *c, d, g* and *h*) or produced in this study (for primers *g* and *h*).

Robustness of the system for biotechnological applications

To illustrate the possibility of using the *g*–*h* primer pair in biotechnology, we retrieved from GenBank some sequences corresponding to common plant species frequently used in food industry. To demonstrate the robustness of the system using the *g* and *h* primers, we tried to amplify this fragment in several highly degraded templates, such as processed food (four samples: brown sugar from sugar cane, cooked potatoes, cooked pasta and lyophilized potage), human feces (two samples) and permafrost samples (four samples). Appropriate criteria for the retrieval of highly degraded DNA were followed (28). This included DNA extraction and PCR setup in dedicated and isolated ancient DNA facilities in Grenoble and Copenhagen, and the use of multiple extraction and PCR blank controls. Importantly, the permafrost sample had been drilled spiking the drilling apparatus with a recognizable bacterial vector (pCR4-TOPO; Stratagene) to test for contamination during drilling and handling. After arrival (frozen) in the laboratory, ~2–3 cm of the core surfaces was removed. The outer scrape and the interior core material were subjected to DNA extractions followed by 40 cycles of PCR using vector-specific primers T3/T7. No vector contaminants were detected in the inner core extracts used for the plant DNA studies. For processed food, total DNA was extracted from 50 mg of dried material using the DNeasy Tissue Kit (Qiagen) following the manufacturer's instructions. The DNA extract was recovered in a volume of 200 µl. Total DNA was extracted according to Godon *et al.* (29) and to Willerslev *et al.* (30) for the human feces and the permafrost sample, respectively. DNA amplifications were carried out using the primers *g* and *h* in final volume of 25 µl, using 2.5 µl of DNA extract as template. The amplification mixture contained 1 U of AmpliTaq® Gold DNA Polymerase (Applied Biosystems), 10 mM Tris–HCl, 50 mM KCl, 2 mM of $MgCl_2$, 0.2 mM of each dNTPs, 1 µM of each primer (for some experiments, the *g* primer was labeled with the HEX fluorochrome, or the *h* primer was labeled with the FAM fluorochrome), and 200 mg/ml of BSA (Roche). 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 and 30 s at 72°C, except for the sugar extract for which we performed 50 cycles, and for the amplifications

with the fluorescent *g* primer for which we removed the elongation time in order to reduce the +A artefact (31,32). PCR products obtained with the fluorescent *g* or *h* primers were electrophoresed for 35 min on an ABI PRISM® 3100 Genetic Analyzer (Applied Biosystems) using 36 cm capillaries and POP-4™ polymer. PCR products obtained with non-fluorescent primers were either directly sequenced, or cloned (except for the permafrost samples) if the sequences obtained with direct sequencing were not readable (i.e. a mixture of different sequences).

RESULTS

The three datasets

Via the ePCR with primers *c* and *d* we retrieved 1308 sequences from GenBank, corresponding to 706 species, 366 genera and 119 families (excluding all sequences with at least one ambiguous nucleotide, and excluding genera with a single species and families with a single genera). With primers *g* and *h*, we retrieved 18 200 sequences, corresponding to 11 404 species, 4215 genera and 410 families. These 18 200 sequences give a good evaluation of the number of chloroplast *trn*L (UAA) intron sequences in GenBank. The much lower number obtained for the *c–d* ePCR is simply due to the fact that the recorded sequences do not contain the primer sequences, and thus are not 'amplified' via our ePCR approach. The arctic plant dataset produced for this study consists of 132 species, 58 genera and 28 families (GenBank accession nos DQ860511–DQ860642). The food dataset analyzed for primers *g* and *h*, consists of 72 species, 64 genera and 37 families retrieved from GenBank, or produced for this study (GenBank accession numbers of species sequenced for this study: EF010967–EF010973).

For all datasets, the length of the sequences amplified with *c* and *d* varies from 254 to 767 bp, and the length of the P6 loop amplified with *g* and *h* varies from 10 bp in *Cuscuta indecora* to 143 bp in *Schoenoplectus littoralis*.

Universality of primer sites

Table 1 presents the sequences of the two primer pairs *c–d*, and *g–h*. Figure 2 shows the exact positions of the four

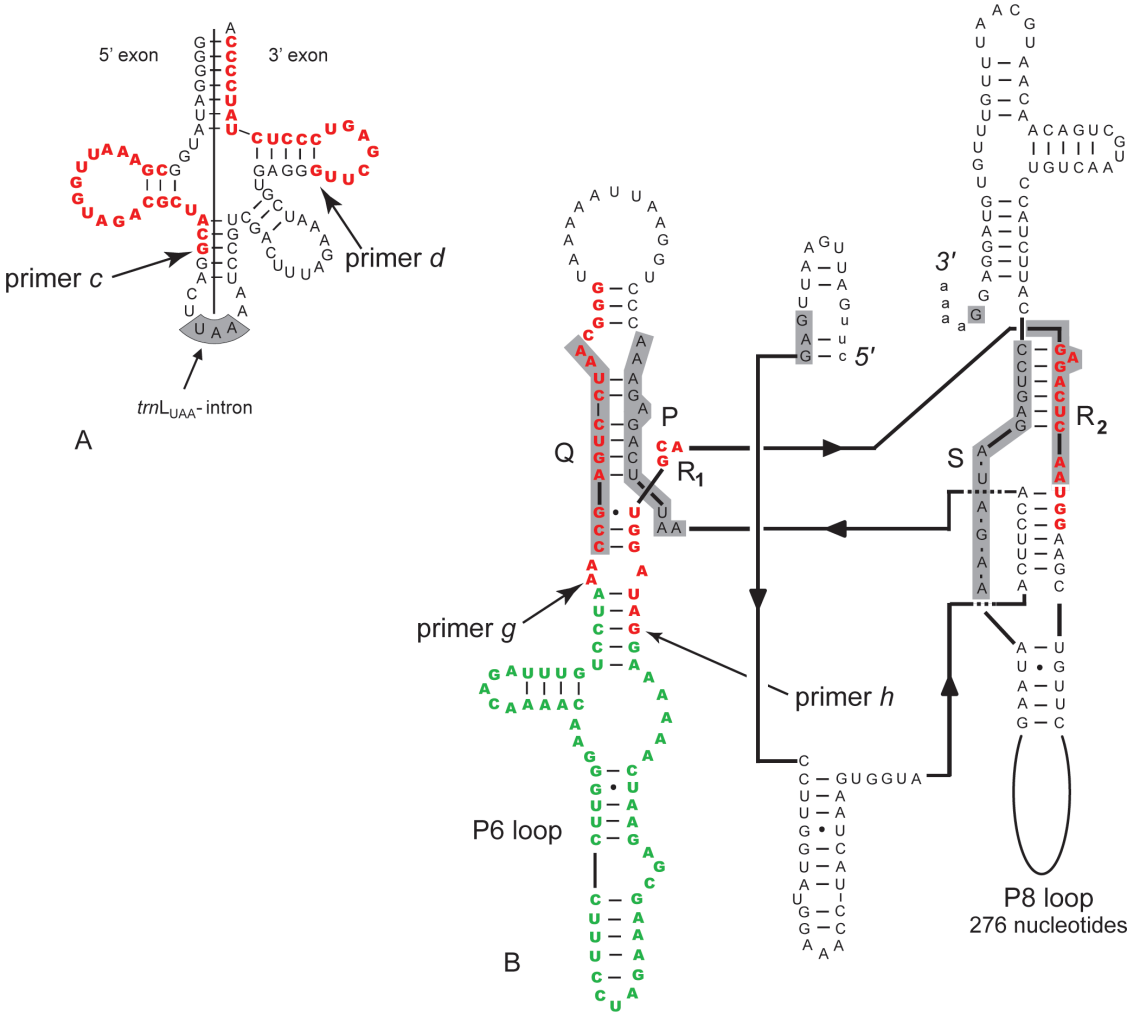


Figure 2. Positions of the primers *c* and *d* on the secondary structure of the *trn*L (UAA) exon (**A**) and of the primers *g* and *h* on the secondary structure of the *trn*L (UAA) intron (**B**) for *Nymphaea odorata* [modified from Ref. (33)]. Highly conserved elements of the catalytic core (P, Q, R1, R2 and S) are located in grey boxes. The P6 loop, amplified with primers *g* and *h*, is identified by green letters. The 3′ ends of each of the four primers *c*, *d*, *g* and *h* are marked out by an arrow and their positions are identified by red letters.

Table 2. Sequence variation of priming site for primer *c, d, g* and *h*

Primer	Sequence 5′–3′	%	Species	Acc. no.
c	CGAAATCGGTAGACGCTACG	76.65	*Nicotiana tabacum*	M16898
	......T..............	17.46	*Carex phacota*	AB079396
	......T..........G..	2.86	*Angelica archangelica*	AF444007
	.........C..........	2.07	*Manulea annua*	AJ550529
	..G..................	0.69	*Luzula rufa*	AY437945
d	GGGGATAGAGGGACTTGAAC	94.18	*N.tabacum*	M16898
	............T......	2.76	*Elegia cuspidata*	AF148735
	..............C....	1.08	*Nymphaea alba*	AJ627251
	...A................	0.89	*Cephalanthus natalensis*	AJ414549
g	GGGCAATCCTGAGCCAA	92.55	*N.tabacum*	M16898
	...T.............	3.78	*Picea abies*	AB045065
	.......T.........	1.27	*Apteranthes europaea*	AJ488313
	....G.......T....	0.51	*Lamium purpureum*	AJ608588
h	CCATTGAGTCTCTGCACCTATC	65.60	*N.tabacum*	M16898
	..G..................	16.15	*Sedum clavatum*	AY540575
	....C................	9.74	*Veronica davisii*	AY540871
	..G.C................	4.28	*Stapeliopsis pillansii*	AY780507
	..T..................	1.55	*Cinnamomum zeylanicum*	AB040085
	..T..............T..	0.60	*Corryocactus brevistylus*	AY015393

Only variants at a frequency higher than 0.005 are indicated. A total of 1014 and 14 145 GenBank entries were used for the primer pairs *c–d* and *g–h*, respectively. %: percentage of sequence variants found in GenBank. Species: Example of species corresponding to the sequence variant. Acc. no.: accession number in GenBank.

primers used in the secondary structure RNAs produced by both the *trn*L (UAA) exon and the *trn*L (UAA) intron. Primers *g* and *h* are located on highly conserved catalytic parts of the intron, leading to the amplification of the short P6 loop.

Table 2 shows the variation at the priming sites. Only sequence variants with a frequency of more than 0.005 were listed. Primers *c* and *d* are highly conserved among land plants, from Angiosperms to Bryophytes. Even in some algae, this primer pair has the potential to produce PCR products. The very large number of *trn*L (UAA) intron sequence retrieved as well as those produced for this study allowed an extensive evaluation of the universality of primers *g* and *h*. These new primers are highly conserved in Angiosperms and Gymnosperms.

Proportions of species, genera and families identified

Table 3 shows the percentage of species, genera and families properly identified using the primer pairs *c–d* and *g–h* in both the GenBank and arctic plant datasets, and the primer pair *h1aF–h2aR* (12). Globally, on the GenBank dataset, the entire *trn*L (UAA) intron and the P6 loop amplified with primers *g* and *h* allow the identification of 67.3 and 19.5% of the species without taking into account single species within a genus, respectively. However, these values are probably underestimates, because of the possibility of misidentification in GenBank (i.e. a wrong species assignment, either by misidentification of the specimen, by problems of synonymy or by PCR contamination). The ePCR using other primer pairs found in Shaw *et al*. (18), which amplify *psb*B-*psb*H, *rpo*B-*trn*C (GCA), *rp*S16 intron, *trn*D (GUC)-*trn*T (GGU), *trn*H (GUG)-*psb*A and *trn*S (UGA)-*trn*fM (CAU), never retrieved more than 100 sequences, and were not taken into account. Table 4 illustrates the sequence variation of *g-h* amplicons for commonly eaten plant species.

Among all the amplicons retrieved from GenBank by using the ePCR software, the percentage of species represented by more than a single entry was 11% for the whole *trn*L intron and 14% for the P6 loop. This subset of sequences allowed to estimate the lower and upper limits of the intraspecific variability. The lower limit was estimated assuming no variation in species represented by a single entry in GenBank, and the upper limit by taking into account only species represented by more than one entry in GenBank. The intraspecific variability lies between 5.9 and 55.0% for the whole intron, and 3.4 and 24.1% for the P6 loop. However, the upper values certainly represent a large overestimation of the real values, because a single entry in GenBank might correspond to many analyzed individuals from the same species. Furthermore, for the P6 loop, the intraspecific polymorphism does not compromise the species identification in 85 cases out of 481.

Robustness of the system using the *g* and *h* primers

We obtained PCR products with 35 cycles for all the samples analyzed, except for the sugar sample, for which 50 cycles were necessary. After electrophoresis of the fluorescent PCR products, some samples gave a single peak (data not shown; sugar, cooked potatoes, cooked pasta) while all the other samples gave a multi-peak profile. The sequences obtained after direct sequencing for the three samples that gave a single peak correspond to sugarcane (*Saccharum officinarum*), potato (*Solanum tuberosum*) and wheat (*Triticum vulgare*). Figure 3 illustrates the multi-peak profiles obtained after electrophoresis of the fluorescent PCR products for more than 20 000 years old permafrost sample, and for a human fecal sample. The PCR products of the lyophilized potage and of the human feces were cloned and sequenced. Table 5 shows the sequences obtained after cloning the PCR product obtained from the lyophilized potage. Twenty-three clones were sequenced, and three species were unambiguously identified: leek (*Allium porum*), potato (*S.tuberosum*) and onion (*Allium cepa*). The same approach was used for the human feces, and the plant species identified are banana (*Musa acuminata*), lettuce (*Lactuca sativa*) and cacao (*Theobroma cacao*).

Table 3. Percentages of species, genera and families identified using the chloroplast *trn*L (UAA) intron, the P6 loop of this intron and comparison with another primer pairs

cpDNA gene and dataset	Length variation (bp)[a]	No. of species/genera/families analyzed[b]	Species (%)	Genus (%)	Family (%)
Chloroplast *trn*L (UAA) intron amplified with primers *c* and *d*. GenBank dataset	254–767	706/366/119	67.28	86.34	100.00
Chloroplast *trn*L (UAA) intron amplified with primers *c* and *d*. Arctic plant dataset	355–653	103/47/24	85.44	100.00	100.00
P6 loop of *trn*L intron amplified with primers *g* and *h*. GenBank dataset	10–143	11 404/4225/310	19.48	41.40	79.35
P6 loop of *trn*L intron amplified with primers *g* and *h*. Arctic plant dataset	22–83	106/48/25	47.17	89.58	100.00
P6 loop of *trn*L intron amplified with primers *g* and *h*. Food dataset	22–65	72/64/37	77.78	87.50	100.00
P6 loop of *trn*L intron amplified with primers *g* and *h*. Subset of the GenBank dataset[c]	10–127	1524/1525/244	24.02	59.48	90.57
*rbc*L amplified with primers *h1aF* and *h2aR* (12). Subset of the GenBank dataset[c]	91–98	1524/1525/244	15.09	37.51	68.03

Note that these estimates were made by taking into account genera with more than two species for the species identification, families with more than two genera for genus identification, and orders with more than two families for family identification.
[a]Length in base pairs excluding primers.
[b]Excluding families with a single genera, genera with a single species and species alone in a genus except for food dataset.
[c]Based on species in common between the *g*–*h* and the *h1aF*–*h2aR* datasets.

Table 4. Example of P6 loop [*trn*L (UAA)] sequences of commonly eaten plant species amplified with primers *g* and *h*

Common name	Scientific name	P6 loop sequence amplified with primers *g* and *h*	Acc. no.
Cacao	*Theobroma cacao*	ATCCTATTATTTTATTATTTTACGAAACTAAACAAAGGTTCAGCAAG-CGAGAATAATAAAAAAAG	EF010969
Beet	*Beta vulgaris*	CTCCTTTTTTCAAAAGAAAAAAAATAAGGATTCCGAAAACAAGAATAAAAAAAAAG	EF010967
Sugarcane	*Saccharum officinarum*	ATCCCCTTTTTTGAAAAAACAAGTGGTTCTCAAACTAGAACCCAAAGGAAAAG	AY116253
Wheat	*Triticum aestivum*	ATCCGTGTTTTGAGAAAACAAGGGGTTCTCGAACTAGAATACAAAGGAAAAG	AB042240
Rye	*Secale cereale*	ATCCGTGTTTTGAGAAAACAAGGGGTTCTCGAACTAGAATACAAAGGAAAAG	AF519162
Rice	*Oryza sativa*	ATCCATGTTTTGAGAAAACAAGCGGTTCTCGAACTAGAACCCAAAGGAAAAG	X15901
Millet	*Panicum miliaceum*	ATCCCTTTTTTGAAAAAACAAGTGGTTCTCAAACTAGAACCCAAAGGAAAAG	AY142738
Strawberry	*Fragaria vesca*	ATCCCGTTTTATGAAAACAAACAAGGGTTTCAGAAAGCGAGAATAAATAAAG	EF010971
Apricot	*Prunus armeniaca*	ATCCTGTTTTATTAAAACAAACAAGGGTTTCATAAACCGAGAATAAAAAAG	EF010968
Sour cherry	*Prunus cerasus*	ATCCTGTTTTATTAAAACAAACAAGGGTTTCATAAACCGAGAATAAAAAAG	EF010970
Maize	*Zea mais*	ATCCCTTTTTTGAAAAACAAGTGGTTCTCAAACTAGAACCCAAAGGAAAAG	NC_001666
Garden pea	*Pisum sativum*	ATCCTTCTTTCTGAAAACAAATAAAAGTTCAGAAAGTGAAAATCAAAAAAG	EF010972
Common bean	*Phaseolus vulgaris*	ATCCCGTTTTCTGAAAAAAAGAAAAATTCAGAAAGTGATAATAAAAAAGG	AY077945
Johnson grass	*Sorghum halepense*	ATCCACTTTTTTCAAAAAAGTGGTTCTCAAACTAGAACCCAAAGGAAAAG	AY116244
Lettuce	*Lactuca sativa*	ATCACGTTTTCCGAAAACAAACAACGGTTCAGAAAGCGAAAATCAAAAAG	U82042
Sunflower	*Helianthus annuus*	ATCACGTTTTCCGAAAACAAACAAAGGTTCAGAAAGCGAAAATAAAAAAG	U82038
Wild oat	*Avena sativa*	ATCCGTGTTTTTGAGAGGGGGGTTCTCGAACTAGAATACAAAGGAAAAG	X75695
Barley	*Hordeum vulgare*	ATCCGTGTTTTGAGAAGGGATTCTCGAACTAGAATACAAAGGAAAAG	X74574
Potato	*Solanum tuberosum*	ATCCTGTTTTCTGAAAACAAACAAAGGTTCAGAAAAAAAG	EF010973
Tomato	*Solanum lycopersicum*	ATCCTGTTTTCTGAAAACAAACCAAGGTTCAGAAAAAAAG	AY098703
Egg plant	*Solanum melongena*	ATCCTGTTTTCTCAAAACAAACAAAGGTTCAGAAAAAAAG	AY266240
Radish	*Raphanus sativus*	ATCCTGAGTTACGCGAACAAACCAGAGTTTAGAAAGCGG	AF451576
Cabbage	*Brassica oleracea*	ATCCTGGGTTACGCGAACAAAACAGAGTTTAGAAAGCGG	AF451574

DISCUSSION

DNA barcoding concerns two categories of scientists: taxonomists and scientists in fields other than taxonomy (4). The goal of this paper was to evaluate the potential use of the chloroplast DNA *trn*L (UAA) intron for plant DNA barcoding in areas other than taxonomy. We will first discuss the drawbacks of this molecular marker, and then its advantages.

The main, and maybe the only but extremely important drawback is the relatively low resolution of the *trn*L (UAA) intron compared with several other noncoding chloroplast regions. This has already been pointed out in several studies (6,18). It is clear that the *trn*L intron does not represent the best choice for characterizing plant species and for phylogenetic studies among closely related species. Obviously, this drawback is even more dramatic when using the very short P6 loop (amplified with primers *g* and *h*), but on the same subset of species, the short P6 loop performs significantly better than the alternative system used to date when analyzing highly degraded DNA [*rbc*L fragment amplified with *h1aF* and *h2aR* (12)]. Finally, even if the proportion of species unambiguously identified with the P6 loop seems low (around 20%), usually only closely related species are not resolved.

It is interesting to note that the relatively low resolution of the *trn*L (UAA) intron is logically linked to a lower intraspecific variation, compared with other noncoding regions of

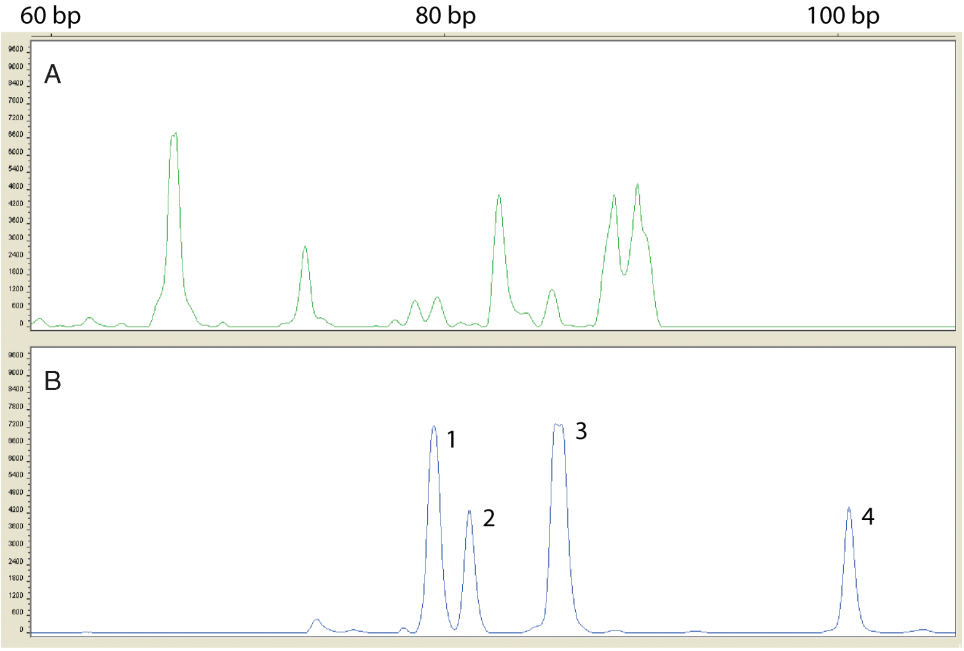


Figure 3. Example of multi-peak profiles obtained after capillary electrophoresis of the fluorescent PCR products obtained using the *g* and *h* primers. (**A**) Permafrost sample drilled from Main River Ice Bluff (N.E. Siberia, 64.06N, 171.11E), between 21 050 and 25 440 years old (uncalibrated ^{14}C years, based on AMS dating of plant macrofossils from the section); *g* fluorescent primer; each peak represents at least one arctic plant species. (**B**) Human feces sample; *h* fluorescent primer; three of the four main peaks have been identified after cloning and sequencing: peak 1, nonidentified; peak 2, banana (*Musa acuminata*); peak 3, lettuce (*Lactuca sativa*); and peak 4, cacao (*Theobroma cacao*).

Table 5. Sequences obtained after cloning the PCR product from the lyophilized potage

Sequence obtained 5′–3′	Species	Number of clones
ATCTTTATTTTTTGAAAAACAA-GGGTTTAAAAAAGAGAAT-AAAAAAG	Leek (*Allium porum*)	19
ATCCTGTTTTCTGAAAACAAA-CAAAGGTTCAGAAAAAAAG	Potato (*Solanum tuberosum*)	3
ATCTTTCTTTTTTGAAAAACAA-GGGTTTAAAAAAGAGAAT AAAAAAG	Onion (*Allium cepa*)	1

Note that onion and leek belong to the same genus *Allium*, and that their sequences differ by a single substitution.

chloroplast DNA (18). Nevertheless, even the short P6 loop can present some intraspecific variation, due in 21.2% of the cases to the presence of a T (or A) stretch of >10 bp long.

However, the strong drawback posed by the relatively low resolution is compensated by several advantages. First, the primers used to amplify both the entire region (*c* and *d*) and the P6 loop (*g* and *h*) are extremely well conserved (Table 2), from Bryophytes to Angiosperms for the *c–d* primer pair, from Gymnosperms to Angiosperms for the *g–h* pair. The primers *g* and *h* are much more conserved than the primers *h1aF* and *h2aR* (12) targeting a protein sequence, and thus having much more variable positions. This advantage is particularly important when amplifying multiple species within the same PCR. Second, the number of *trn*L (UAA) intron sequences available in databases is already very high, by far the most numerous among noncoding chloroplast DNA sequences, allowing in many cases the identification of the species or the genus. Finally, the robustness of both systems (the entire intron and the P6 loop) also represents an important advantage. This last advantage might be linked to the two previous ones, because a robust system will incite scientists to use this region, increasing the number of sequences in databases, and the robustness mainly comes from the primer universality.

Actually, in some situations, the relatively low resolution of the *trn*L intron can be largely compensated by the possibilities of standardization. In many situations, the number of possible plant species is restricted, reducing the impact of the relatively low resolution. In our arctic plant dataset, the number of species unambiguously identified among 123 is close to 50% for the P6 loop, and close to 85% for the entire intron. In the same way, the eaten plant species are few and taxonomically diverse, and can be identified in most cases. Even the short P6 loop allows the identification of the three commonly eaten species of the genus *Solanum* (potato, tomato and eggplant), which differ by a single mutation (see Table 4). However, the P6 loop does not allow the identification of the different cultivars of the same species [specifically, of *Brassica oleracea* (Brussels sprouts, Kohl rabi, Broccoli, etc.) or of *Phaseolus vulgaris* (different cultivated varieties)]. In addition, the P6 loop cannot distinguish most of the species of the genus *Prunus* (apricot, peach, cherry, etc.).

To conclude, the *trn*L (UAA) intron, despite its relatively low resolution, provide a unique opportunity for plant DNA barcoding in the biotechnology area, because of the universality of the *c–d* and *g–h* primers, of the robustness of the amplification process, and of the possibility of developing highly standardized procedures. Furthermore, the

low-intraspecific variation represents an important advantage if the amplicons are detected by hybridization. Even the short P6 loop allows to gather valuable information about plant identification and will undoubtedly become the marker of choice for highly degraded template DNA. This P6 loop has the potential to be extensively used in food industry, in forensic science, in diet studies based on feces, and in permafrost analyses for reconstructing past plant communities.

ACKNOWLEDGEMENTS

This study has been financially supported by an ECLIPSE II grant (CNRS). We thank Dietmar Quandt for help with Figure 2, and Jean-Pierre Furet for extracting the DNA from human fecal samples. E.W. wants to thank Andrei Sher and James Haile for helping with sample collection, Tina Brand for assisting the lab work and The Wellcome Trust, UK and the National Science Foundation, DK for financial support. F.P. is supported by the French 'Institut National de la Recherche Agronomique'. C.B. thanks Reidar Elven and Hanne H. Grundt for help with the arctic plant sample collection and the Research Council of Norway (grant 146 515/420) for funding. Funding to pay the Open Access publication charges for this article was provided by CNRS.

Conflict of interest statement. None declared.

REFERENCES

1. Floyd,R., Abebe,E., Papert,A. and Blaxter,M. (2002) Molecular barcodes for soil nematode identification. *Mol. Ecol*., **11**, 839–850.
2. Hebert,P.D.N., Cywinska,A., Ball,S.L. and de Waard,J.R. (2003) Biological identification through DNA barcodes. *Proc. R. Soc. Lond., B. Biol. Sci*., **270**, 313–321.
3. Hebert,P.D.N. and Gregory,T.R. (2005) The promise of DNA barcoding for taxonomy. *Syst. Biol*., **54**, 852–859.
4. Chase,M.W., Salamin,N., Wilkinson,M., Dunwell,J.M., Kesanakurthi,R.P., Haidar,N. and Savolainen,V. (2005) Land plants and DNA barcodes: short-term and long-term goals. *Philos. Trans. R. Soc. B Biol. Sci*., **360**, 1889–1895.
5. Hebert,P.D.N., Ratnasingham,S. and de Waard,J.R. (2003) Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Lond. B Biol. Sci*., **270**, S96–S99.
6. Kress,W.J., Wurdack,K.J., Zimmer,E.A., Weigt,L.A. and Janzen,D.H. (2005) Use of DNA barcodes to identify flowering plants. *Proc. Natl Acad. Sci. USA*, **102**, 8369–8374.
7. Vences,M., Thomas,M., van der Meijden,A., Chiari,Y. and Vieites,D. (2005) Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. *Front. Zool*., **2**, 5.
8. Hebert,P.D.N., Penton,E.H., Burns,J.M., Janzen,D.H. and Hallwachs,W. (2004) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proc. Natl Acad. Sci. USA*, **101**, 14812–14817.
9. Hebert,P.D.N., Stoeckle,M.Y., Zemlak,T.S. and Francis,C.M. (2004) Identification of birds through DNA barcodes. *PLoS Biol*., **2**, e312.
10. Tautz,D., Arctander,P., Minelli,A., Thomas,R.H. and Vogler,A.P. (2003) A plea for DNA taxonomy. *Trends Ecol. Evol*., **18**, 70–74.
11. Álvarez,I. and Wendel,J.F. (2003) Ribosomal ITS sequences and plant phylogenetic inference. *Mol. Phylogenet. Evol*., **29**, 417–434.
12. Poinar,H.N., Hofreiter,M., Spaulding,W.G., Martin,P.S., Stankiewicz,B.A., Bland,H., Evershed,R.P., Possnert,G. and Pääbo,S. (1998) Molecular coproscopy: Dung and diet of the extinct ground sloth Nothrotheriops shastensis. *Science*, **281**, 402–406.
13. Scharaschklin,T. and Doyle,J.A. (2005) Phylogeny and historical biogeography of Anaxagorea (Annonaceae) using morphology and noncoding chloroplast sequence data. *Syst. Bot*., **30**, 712–735.
14. McDade,L.A., Daniel,T.F., Kiel,C.A. and Vollesen,K. (2005) Phylogenetic relationships among Acantheae (Acanthaceae): major lineages present contrasting patterns of molecular evolution and morphological differentiation. *Syst. Bot*., **30**, 834–862.
15. Chen,S.Y., Xia,T., Wang,Y.J., Liu,J.Q. and Chen,S.L. (2005) Molecular systematics and biogeography of Crawfurdia, Metagentiana and Tripterospermum (Gentianaceae) based on nuclear ribosomal and plastid DNA sequences. *Ann. Bot*., **96**, 413–424.
16. Ronning,S.B., Rudi,K., Berdal,K.G. and Holst-Jensen,A. (2005) Differentiation of important and closely related cereal plant species (Poaceae) in food by hybridization to an oligonucleotide array. *J. Agric. Food Chem*., **53**, 8874–8880.
17. Ward,J., Peakall,R., Gilmore,S.R. and Robertson,J. (2005) A molecular identification system for grasses: a novel technology for forensic botany. *Forensic Sci. Int*., **152**, 121–131.
18. Shaw,J., Lickey,E.B., Beck,J.T., Farmer,S.B., Liu,W., Miller,J., Siripun,K.C., Winder,C.T., Schilling,E.E. and Small,R.L. (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *Am. J. Bot*., **92**, 142–166.
19. Taberlet,P., Gielly,L., Pautou,G. and Bouvet,J. (1991) Universal primers for amplification of three noncoding regions of chloroplast DNA. *Plant Mol. Biol*., **17**, 1105–1109.
20. Gielly,L. and Taberlet,P. (1996) A phylogeny of the European gentians inferred from chloroplast *trn*L (UAA) intron sequences. *Bot. J. Linn. Soc*., **120**, 57–75.
21. Quandt,D. and Stech,M. (2005) Molecular evolution of the *trn*L (UAA) intron in bryophytes. *Mol. Phylogenet. Evol*., **36**, 429–443.
22. Quandt,D., Müller,K., Stech,M., Frahm,J.P., Frey,W., Hilu,K.W. and Borsch,T. (2004) Molecular evolution of the chloroplast *trn*L-F region in land plants. *Monogr. Syst. Bot. Missouri Botanic Garden*, **98**, 13–37.
23. Shinozaki,K., Ohme,M., Tanaka,M., Wakasugi,T., Hayashida,N., Matsubayashi,T., Zaita,N., Chunwongse,J., Obokata,J., Yamaguchi-Shinozaki,K. *et al*. (1986) The complete nucleotide sequence of tobacco chloroplast genome: its gene organization and expression. *EMBO J*., **5**, 2043–2049.
24. Palmer,J.D. (1991) Plastid chromosomes: structure and evolution. *Cell Cult. Som. Cell Genet. Plants*, **7A**, 5–53.
25. Michel,F., Jacquier,A. and Dujon,B. (1982) Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. *Biochimie*, **64**, 867–881.
26. Davies,R.W., Waring,R.B., Ray,J.A., Brown,T.A. and Scazzocchio,C. (1982) Making ends meet—a model for RNA splicing in fungal mitochondria. *Nature*, **300**, 719–724.
27. Wu,S. and Manber,U. (1992) Agrep-a fast approximate pattern-matching tool. In *Proceedings of the USENIX Winter 1992 Technical Conference,* USENIX Association, Berkeley, CA, pp. 153–162.
28. Willerslev,E. and Cooper,A. (2005) Ancient DNA. *Proc. R. Soc. Lond. B*, **272**, 3–16.
29. Godon,J.J., Zumstein,E., Dabert,P., Habouzit,F. and Moletta,R. (1997) Molecular microbial diversity of an anaerobic digestor as determined by small-subunit rDNA sequence analysis. *Appl. Environ. Microbiol*., **63**, 2802–2813.
30. Willerslev,E., Hansen,A.J., Binladen,J., Brand,T.B., Gilbert,M.T.P., Shapiro,B., Bunce,M., Wiuf,C., Gilichinsky,D.A. and Cooper,A. (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science*, **300**, 791–795.
31. Brownstein,M.J., Carpten,J.D. and Smith,J.R. (1996) Modulation of non-templated nucleotide addition by *Taq* polymerase: primer modification that facilitate genotyping. *BioTechniques*, **20**, 1004–1010.
32. Magnuson,V.L., Ally,D.S., Nylund,S.J., Karanjawala,Z.E., Rayman,J.B., Knapp,J.I., Lowe,A.L., Ghosh,S. and Collins,F.S. (1996) Substrate nucleotide-determinated non-templated addition of adenine by *Taq* DNA polymerase: implications for PCR-based genotyping and cloning. *BioTechniques*, **21**, 700–709.
33. Borsch,T., Hilu,K.W., Quandt,D., Wilde,V., Neinhuis,C. and Barthlott,W. (2003) Noncoding plastid *trn*T-*trn*F sequences reveal a well resolved phylogeny of basal angiosperms. *J. Evol. Biol*., **16**, 558–576.

Paper III

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

Alice Valentini[1, 2], Christian Miquel[1], Muhammad Ali Nawaz[3, 4], Eva Bellemain[1], Eric Coissac[1], François Pompanon[1], Ludovic Gielly[1], Corinne Cruaud[5], Giuseppe Nascetti[2], Patrick Winker[5], Jon E. Swenson[3, 6], Pierre Taberlet[1]

[1] Laboratoire d'Ecologie Alpine, CNRS UMR 5553, Université Joseph Fourier, BP 53, F-38041 Grenoble Cedex 9, France.

[2] Dipartimento di Ecologia e Sviluppo Economico Sostenibile, Università degli Studi della Tuscia, via S. Giovanni Decollato 1, I-01100 Viterbo, Italy.

[3] Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, Postbox 5003, NO-1432 Ås, Norway.

[4] Himalayan Wildlife Foundation, Islamabad, Pakistan.

[5] Genoscope - CNS, 2 rue Gaston Crémieux, BP 5706, F-91057 Evry Cedex, France.

[6] Norwegian Institute for Nature Research, NO-7485 Trondheim, Norway.

Received:

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

Corresponding author: P. Taberlet, Laboratoire d'Ecologie Alpine, CNRS-UMR 5553, Université Joseph Fourier, BP 53, F-38041 Grenoble Cedex 09, France; Tel: +33(0)4 76 51 45 24; Fax: +33(0)4 76 51 42 79; E-mail: pierre.taberlet@ujf-grenoble.fr

Running title: Non-invasive DNA-based diet analysis

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 150 base pairs are usually degraded very slowly 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, i.e. 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. Undoubtedly, 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.

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-Rodrígez 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 a complex environment, 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 *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. Some herbivores do not masticate their food into small fragments, allowing 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 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 feed 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 analyzed. 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 analyzing the diet of the Macaroni penguin (*Eudyptes chrysolophus*) using faeces as a source of DNA. The presence/absence of the different prey were detected by carrying out five different 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. Universal primers targeting the chloroplast *rbc*L gene and subsequent cloning have been used to analyze 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 analyzing gut content in insects (see review in Symondson 2002) and in birds and mammals (e.g. Jarman *et al.* 2004).

In this paper 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 *trnL* (UAA) intron (Taberlet *et al.* 2007) via the polymerase chain reaction (PCR; Mullis & Faloona 1987) and by subsequently sequencing individual molecules of this PCR product on the 454 automated sequencer (Roche Diagnostic, Basel, Switzerland). We demonstrate the efficiency of this new approach by analyzing the diet of various herbivorous species, including mammals, birds, molluscs, and insects.

Materials and methods

General strategy

Fig. 1 gives an overview of the main steps necessary to estimate the diet of herbivorous species. After collecting 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 longicauda*) in the 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 analyzed 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 (*T. u. aquitanus*) and two from the Corinthian Alps in Austria (*T. u. 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, Hilden, Germany), 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 AmpliTaq® Gold DNA Polymerase (Applied Biosystems, Foster City, CA), 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 Diagnostic, Basel, Switzerland). 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). Each sample was amplified with primers *g* and *h* (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.

DNA sequencing

PCR products were purified using the MinElute PCR purification kit (Qiagen GmbH, Hilden, Germany). DNA quantification was carried out using the NanoDrop® ND-1000 UV-Vis Spectrophotometer (NanoDrop Technologies® Wilmington, DE). 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, Basel, Switzerland) 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 (Taberlet *et al.* 1991), and following the protocol described in Taberlet *et al.* (2007).

Data analysis for estimating diet composition

Out of the mix of sequences obtained after the pyrosequencing, the first step of the data analysis 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 file was generated, containing all the sequences having the relevant tag on its 5' end. Then, these sequences were analyzed to determine the diet. Only sequences present more than three times were taken into account in the subsequent analyses. The diet was then determined by comparing these sequences to the homologous sequences available in databases. In the case of the brown bear and marmot, the sequences were first compared to the database generated for the Deosai National Park and then, if no match was found, to public databases. For all other species, the sequences were directly compared to public databases to find their closest match using the MEGABLAST algorithm (Zhang *et al.* 2000).

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 percent of the species analyzed 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 EMBL database, under accession numbers XXXXX-XXXXX.

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. In each samples, a few sequences were found hundreds of time, whereas some other sequences are only represented either once or by very few occurrences (Table 1). The sequences showing up 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. 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 2 and Fig. 2. Sixty-four percent and 31% of the different P6 loop sequences obtained in their diet was 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 3 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 conifers 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.

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 in the efficiency of amplifications among species. 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 analyzing 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 analyze 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 that occur in a particular area, usually about 50% of the different species should be identified to species, and 90% to genus. It is interesting to note that some genera exhibit a limited variation (e.g. *Carex*) or almost no variation (e.g. *Salix*, *Pinus*, etc.) on this P6 loop. When it is important to determine the species, we suggest to complement the 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.

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 (Gielly *et al.* in preparation). 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 1). 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 1). 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 type of the different 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, Hilden, Germany) 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, these numbers cannot be interpreted as quantitative at the moment for several reasons. First, the preferential amplification of some species when analyzing 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 such preferential amplification. 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 type of tissue eaten. Leaves will undoubtedly provide more chloroplast DNA than roots, 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, i.e. 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.

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.

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**, 1004-1006, 1008-1010.

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

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 B-Biological Sciences*, **360**, 1889-1895.

Cristóbal-Azkarate J, Arroyo-Rodrígez 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, 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.

Gielly L, Coissac E, Valentini A, Miquel C, Paris M, Pompanon F, Taberlet P (200X) A new tagging system for parallel pyrosequencing of multiple homologous PCR products. *Molecular Ecology Notes*, in preparation.

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.

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.

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 of the United States of America*, **98**, 4317-4322.

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

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

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

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.

Acknowledgements

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.

415 **Figure legends**
416
417 **Fig. 1** Flowchart diagram showing the main steps of the *trn*L approach for assessing diet
418 composition using faeces.
419
420 **Fig. 2** Comparison of the diet compositions of the golden marmot (*Marmota caudata*) and of
421 the brown bear (*Ursus arctos*) in the Deosai National Park (Pakistan). See Table 2 for the
422 plant taxa identified within each of these families. The Y-axis corresponds to the frequency of
423 presence of taxa from the same family in the twelve samples of each mammal species.

Tables and Figures

Table 1 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 3546 sequences were obtained 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.

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

Table 2 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

			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**
	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 longi-aristatus	species													-						x		x		x			3
	Poa alpina	species													-						x							1
	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

			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**
	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	

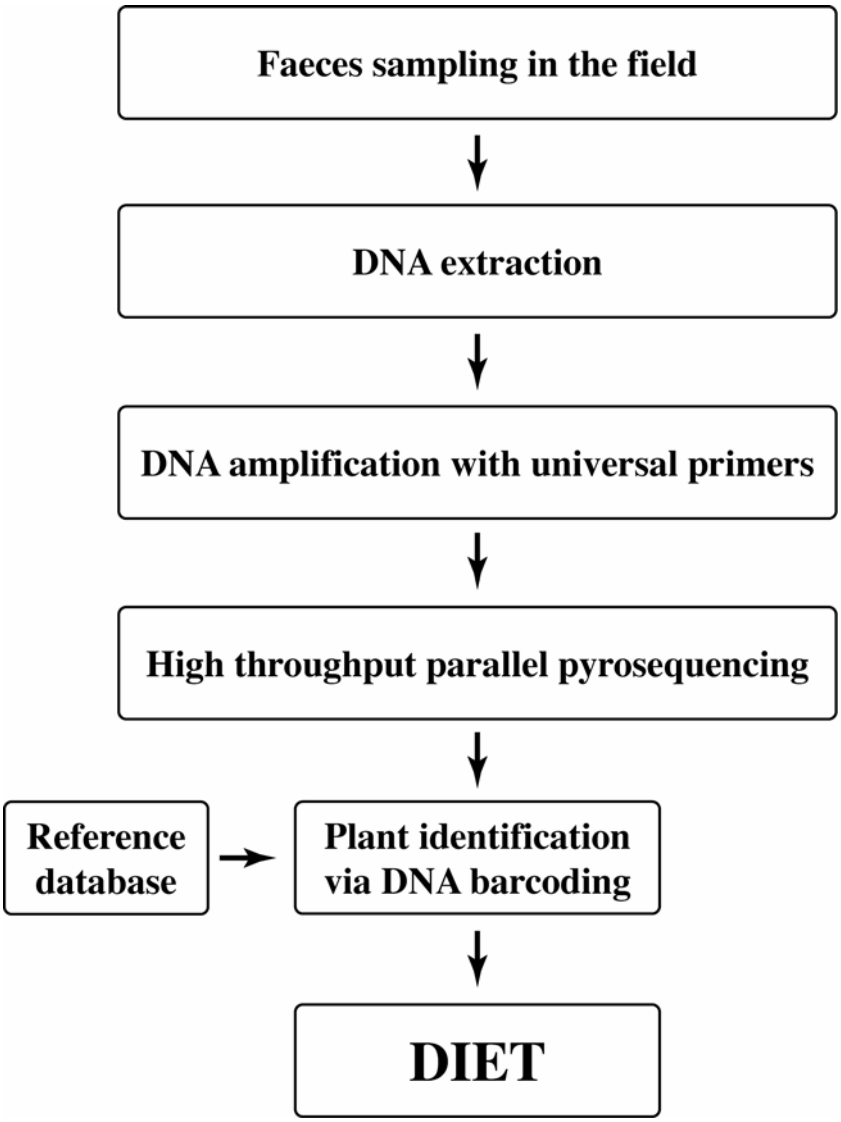
433 * Plants identified by comparing the sequence with sequence data in public databases.

Table 3 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

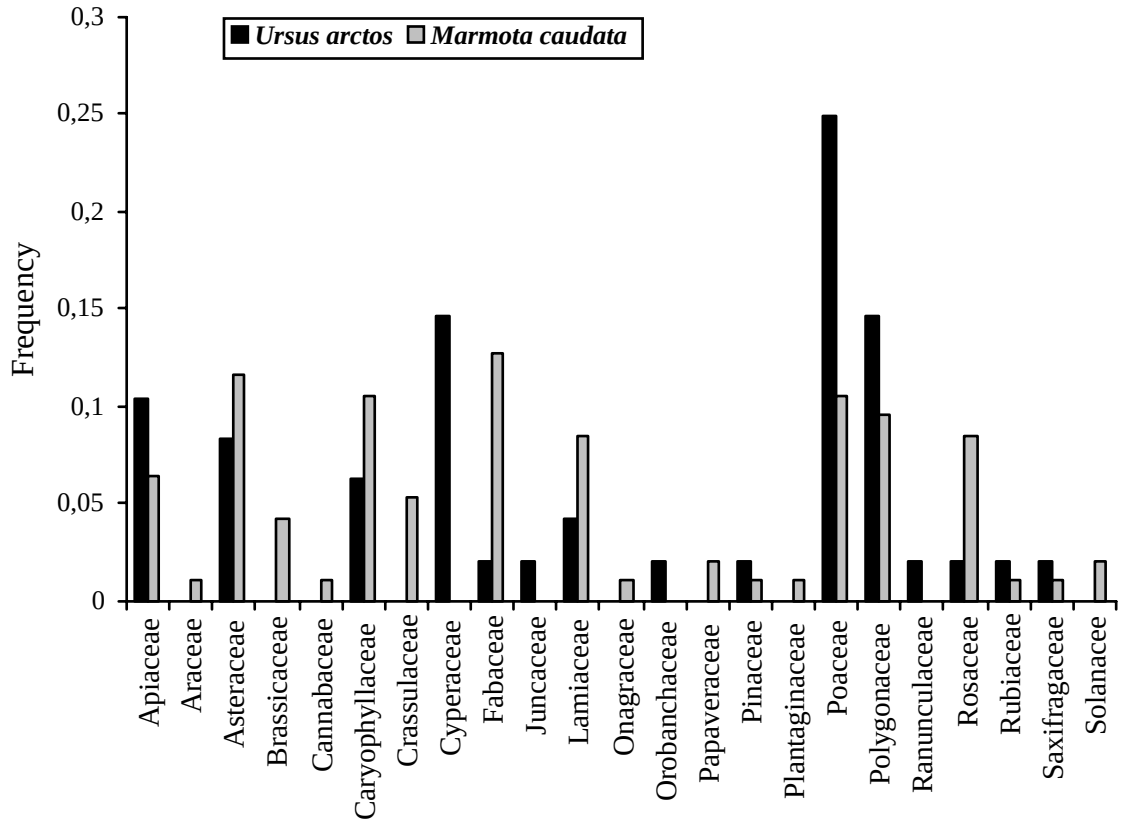
B1 = *Tetrao urogallus aquitanus* Sample 1; B2 = *T. u. aquitanus* Sample 2; B3 = *T. u. aquitanus* Sample 3; B4 = *T. u. aquitanus* Sample 4; B5 = *T. u. major* Sample 1; B6 = *T. u. major* Sample 2; M1 = *Helix aspera*; M2 = *Deroceras reticulatum*; M3 = *Arion rufus*; I1 = *Chorthippus biguttulus*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

440 Figure 1
441
442
443
444
445



446

Figure 2


Ursus arctos
Marmota caudata
Frequency
0,3
0,25
0,2
0,15
0,1
0,05
0
Apiaceae
Araceae
Asteraceae
Brassicaceae
Cannabaceae
Caryophyllaceae
Crassulaceae
Cyperaceae
Fabaceae
Juncaceae
Lamiaceae
Onagraceae
Orobanchaceae
Papaveraceae
Pinaceae
Plantaginaceae
Poaceae
Polygonaceae
Ranunculaceae
Rosaceae
Rubiaceae
Saxifragaceae
Solanacee

Paper IV

DNA barcoding for ecologists

Alice Valentini [1, 2], François Pompanon [1], Pierre Taberlet [1]

[1] Laboratoire d'Ecologie Alpine, CNRS UMR 5553, Université Joseph Fourier, BP 53, 38041 Grenoble Cedex 9, FRANCE

[2] Dipartimento di Ecologia e Sviluppo Economico Sostenibile, Università della Tuscia, Via S.Giovanni Decollato 1, 01100 Viterbo, ITALY

Running title: DNA Barcoding for ecologists

Keywords: DNA barcoding, species identification, DNA sequencer, environmental samples, DNA-based diet analysis

Word counts (ex abstract, boxes, tables, legends, references): 3000

*Corresponding author: Taberlet, P (pierre.taberlet@ujf-grenoble.fr)

DNA barcoding, i.e. species identification using a standardized DNA region, has received much attention of late, and is being further developed through an international initiative called "Consortium for the Barcode of Life". With more and more DNA sequences allowing species identification accessible in databases and new sequencing technologies dramatically expanding available sequencing power, we anticipate that DNA barcoding techniques will be increasingly used by ecologists. They will be able to not only identify a single species from a specimen or an organism's remains using the standardized approach, but also determine the composition of complex source material. For example, the use of very short DNA fragments that persist in the environment will allow an assessment of local biodiversity from soil or water, or establishing diet composition from feces.

A new name for an old concept

The term DNA barcoding is of recent use in the literature [1, 2]. It relies on the use of a standardized DNA region as a tag for rapid and accurate species identification [3]. Nevertheless, DNA barcoding is not a new concept. The term "DNA barcodes" was first used in 1993 [4], in a paper that did not receive very much attention from the scientific community. The concept of species identification using molecular tools is older still, and came before the invention of the Sanger sequencing technique. However, the "golden age" of DNA barcoding began in 2003 [2] and the number of publications on the subject has grown exponentially, and now surpasses 250 articles (Figure 1).

The now well-established Consortium for the Barcode of Life (CBOL; http://barcoding.si.edu/), an international initiative supporting the development of DNA barcoding, aims to both promote global standards and coordinate research in DNA barcoding. For animals, the gene region proposed for the standard barcode is a 650

base-pair region in the mitochondrial (mt) cytochrome c oxidase 1 gene ("COI") [2]. For plants, the situation is still controversial but recently it has been proposed to use three coding chloroplast DNA regions that together would represent the standard barcode: *rpoC1*, *matK*, and either *rpoB* or *psbA-trnH* [5].

As pointed out by Chase *et al.* [6], taxonomists are not the only potential users of DNA barcoding, as the technique may be useful for scientists from other fields (e.g. forensic science, biotechnology and food industry, animal diet). Taxonomists are concerned in DNA barcoding "*sensu stricto*". Other scientists will be more interested in DNA barcoding "*sensu lato*" i.e. by DNA-based taxon identification using diverse techniques than may lie outside the CBOL approach (Table 1). The difference between the two approaches derives mainly from different priorities given to the criteria used for designing the molecular markers (Box 1). Taxonomists tend to prefer standardized markers that express a high level of variation with sufficient phylogenetic information, following the CBOL strategy, while other scientists may favor highly robust procedures even if identification to species level is not always possible.

Here, our purpose is to discuss the main applications of DNA barcoding in ecology. First, we will present the new tools that make the barcoding approach for ecologists easier. Then, we will focus on single species identification, which is the historical fundament of DNA barcoding. Finally, we will discuss the use of DNA barcoding for simultaneous multiple species identification from a single sample, for biodiversity assessment and for diet analysis from feces.

New tools for new prospects

In the past twenty years, the technology of DNA sequencing has greatly improved, from manual sequencing to automated sequencers. A single automated 96-capillary sequencer

66 can provide more than 1000 sequences of 1000 base pair (bp) per day. Even non-
67 geneticists now have easy access to sequencing via companies that offer this service at a
68 competitive price. Clearly, the development of DNA barcoding is linked to these
69 improvements.

70 When using the classical sequencing approach via capillary electrophoresis,
71 environmental samples (Box 2) require an additional step of cloning the different
72 amplified DNA fragments into bacteria, followed by sequencing many clones in order
73 to reveal the full complexity of those samples. Such a cloning step is both expensive
74 and time-consuming, thus limiting large-scale applications. New DNA sequencing
75 technologies bypassing the cloning step have recently been developed (Box 3), opening
76 the way to applying large-scale DNA barcoding studies to environmental samples
77 (Figure 2).

78 More and more sequence data are becoming available in public databases
79 (GenBank, EMBL, DDBJ) as sequencing facilities improve. This greatly stimulates the
80 development of species identification via DNA barcoding, and enhances the design of
81 standardized methods by allowing a better design of "universal" primers. However, the
82 quality of the sequence data in GenBank, EMBL, or DDBJ is not always perfect [7],
83 either due to sequencing errors, contaminations, sample misidentifications, or
84 taxonomic problems. CBOL's recent initiative to build a new database specially
85 dedicated to DNA barcoding will change this situation, and will provide an efficient and
86 accurate tool for species identification (Barcoding Of Life Database, BOLD,
87 http://www.boldsystems.org/). BOLD has been designed to record not only DNA
88 sequences from several individuals per species (including primer sets, electropherogram
89 trace files, and translations), but also complete taxonomic information, place and date of

collection, and specimen images [8]. It seems obvious that ecologists will take advantage of such a high quality database.

Single species identification

The classical use of morphological traits for species identification has several limitations. They include, for example, the misidentification of a taxon due to the phenotypic plasticity of the trait studied, or the existence of cryptic taxa. Moreover, morphological keys are sometimes only effective for a particular life stage or gender. Thus, a high level of expertise is often required to correctly identify species with the accuracy, required in ecological studies. There is no doubt that the DNA barcoding approach currently represents the best solution for identifying species when their morphology is of limited use, even if DNA barcoding itself also presents some limitations (see e.g. Ref. [9]). The following examples illustrate some situations where DNA barcoding greatly helped ecologists.

Nematodes play an important role in the ecosystem either as free-living species or as parasites. They can contribute to nitrogen mineralization and distribution of biomass within plants in soil ecosystems [10]. Despite this important role, their identification is still extremely difficult. Nematode species are morphologically similar and most identification keys are based on adult characteristics. Due to these difficulties their biodiversity is greatly underestimated. For example it is estimated that global marine nematode species richness exceeds one million, while only a few thousand of them are described [11]. In order to better understand which species play which role, ecologists studying this group took the opportunity to use DNA barcoding in both marine and terrestrial environments [1, 12, 13].

In addition, ecologists can take advantage of DNA tools when only hair, feces, or urine left behind by animals are used for species identification. Such an approach is now widely used, and is particularly useful for detecting the presence of elusive or endangered species (e.g. Refs [14, 15]).

Identifying species through DNA barcoding is also helpful for understanding inter-species interactions. For example, the barcoding approach has already shown that the existence of cryptic species could mask the specialization of a parasite to a single host. Smith *et al.* found that several morphospecies of tachinid parasitoid flies thought to be generalist actually corresponded to many different cryptic species that were less generalist or even specialist [16, 17].

Furthermore, DNA barcoding can be advantageous for monitoring illegal trade in animal byproducts. When such products are sold, identification through morphological characteristics may no longer be possible. Sometimes, only hairs are available for species identification, and it is very difficult or even impossible to visually determine whether a hair came from an endangered or from a legally sold species. DNA methodology has been successfully implemented to identify Eurasian badger (*Meles meles*) hairs in luxury shaving brushes [18] and Tibetan Antelope (*Pantholops hodgsonii*) in shahtoosh [19]. Many other examples can be found in the food market (see e.g. Ref. [20]). For instance, scientists using this technique revealed that twenty three percent of black caviar samples purchased were with a wrong species name and in some cases the commercial species (Russian sturgeon, *Acipenser gueldenstaedtii*) was replaced with an endangered one (ship sturgeon, *A. nudiventris*) [21].

In the field of biosecurity, the reliable and fast identification of exotic species is fundamental. In many cases, especially in insects, a pest at the egg or larval stage might not recognizable without DNA identification. For example, stem borer larvae of the

genus *Busseola* leaving on sugarcane in Ethiopia were identified using COI sequences [22]. Molecular markers can even allow the possibility of identifing populations within species, and to eventually identify the source population of invasive species. *Caulerpa taxifolia* is the most famous case of invasion in marine environments. Sequences of this algae's ITS region proved that its invasion of California [23] and Australia [24] was due to a clone similar to the one released by error from the Monaco Oceanography Museum in 1984.

Diet analysis

The study of food webs and their dynamics is fundamental to understand how the feeding habits of the different species can affect the community. Thus, diet analysis of the animal species present in a given environment can improve our understanding of the functioning of the ecosystem as a whole (e.g. Ref. [25]). Furthermore, the study of feeding ecology becomes crucial when it concerns endangered species. A precise knowledge of the diet of those species could be considered when designing reliable conservation strategies (e.g. Refs [26, 27]). DNA barcoding makes it possible to establish the diet of an individual from its feces or stomach contents. This is helpful when the food is not identifiable by morphological criteria, such as in liquid feeders like spiders [28]. This technique also provides valuable information when eating behavior is not directly observable, as in the case of krill-eating diatoms [29], giant squid (*Architeuthis* sp.) in the sea abyss [30], or deep sea invertebrates [31]. Most of the studies that use DNA markers for diet analysis are based on carnivorous animals (e.g., insects Ref. [32, 33], whale and Adelie penguin (*Pygoscelis adeliae*）, (Ref. [34]). Fewer studies were carried on herbivorous animals (e.g. Ref. [35].

There are two different strategies when using molecular tools for diet analysis:: the use of group-specific primers [36] (Nystrom 2006) or the use of universal primer. When analyzing the diet of the Macaroni penguin (*Eudyptes chrysolophus*) using feces as a source of DNA, Deagle *et al.* [37] applied both group-specific and universal primers. The results obtained with five different groups of specific primers were similar to those involving universal (for fish, cephalopods and crustaceans) 16S rDNA primers and subsequent cloning of the PCR products. In general, the use of specific primers requires an *a priori* knowledge of the animal's diet. This is not possible in most cases and makes the "universal" approach more appropriate. Recently, Taberlet *et al.* [38] have described the possible use of a very short fragment (10-140 bp) of the chloroplast *trn*L (UAA) intron as a target for plant barcoding in situations where only degraded DNA is available. The *trn*L approach is suitable for diet studies of herbivorous animals because the primers are universal (for Gymnosperms and Angiosperms) and because it works for feces that contain degraded DNA. This barcoding system, combined with high-throughput parallel pyrosequencing, was successfully used to amplify and analyze the diet of mammals (Figure 3), birds, mollusks, and insects [39].

Biodiversity assessment

Current biodiversity

Even if morphological identification of a species is possible, DNA barcoding may enhance biodiversity inventories by being faster and cheaper. It could allow biodiversity assessment through the identification of taxa from the traces of DNA present in environmental samples such as soil or water. The cost for identifying a sample via barcoding, including DNA extraction, DNA amplification, purification of the PCR product, and sequencing using capillary electrophoresis, was estimated to range from

185 2.5$ to 8$ per sample, depending on laboratory facilities and consumable equipment
186 [40-42]. A technician in DNA barcoding could replace dozens of taxonomists for
187 routine identification, allowing taxonomists to concentrate on identifying reference
188 specimens for establishing reliable databases. The use of DNA barcoding will not be
189 necessary for assessing the biodiversity of well-known ecosystems such as that of
190 Alpine prairie plants. However, in ecosystems showing high species richness like those
191 in tropical environments, it is unrealistic, within a limited time period, to identify all
192 animals and plants by morphology alone. The biodiversity of environments with low
193 accessibility can also be estimated with this new technique, as demonstrated by Sogin *et*
194 *al.* [43] who applied the technique to the microbial biodiversity in deep sea. Large-scale
195 studies also become possible, because the barcoding approach allows the simultaneous
196 identification of most species from a given biotope (e.g. Ref. [44], Global Ocean
197 Sampling project).

198 The use of classical biodiversity indices could be complemented by new indices
199 developed to exploit the information contained in whole sequence sets obtained from a
200 single environment. Thus, the estimation of biodiversity indices can be based on
201 MOTUs (Molecular Operational Taxonomic Units) detected with the barcoding
202 approach, where the relative abundances of each type of DNA sequences (MOTUs)
203 replace the classical relative abundance of each species estimated from the number of
204 individuals. This approach is now common in environmental microbiology for
205 estimating different diversity indices (species richness, Shannon's or Simpson's indices;
206 e.g. Refs [45-47]), but presents some bias when the number of species is very large [48].
207 To our knowledge, DNA-based biodiversity indices have not yet been used for plants or
208 animals using environmental samples such as water or soil. With the recent

development of parallel pyrosequencing, this type of approach might also be applied to eukaryotes in the near future.

Paleoecology

Generally, reconstructing the ecosystems of the past from fossil data is very difficult because of the nature of the samples and their low preservation. In most cases, morphological identification is very difficult or even impossible. In this case molecular tools and DNA barcoding could help scientists to successfully describe past plant and animal communities. Only a few sites are suitable for DNA-based paleoecology, where molecules are well preserved in dry or cold environments.

Past communities have been analyzed from samples collected in Siberian permafrost sediment from the Pleistocene and Holocene periods, revealing a change in plant composition between these two periods, and identifying eight different species of mammals, including mammoth (*Mammuthus primigenius*), musk ox (*Ovibos moschatus*), reindeer (*Rangifer tarandus*), and lemming (*Lemmus lemmus)* [49]. The same type of study has been recently carried out on samples taken from 450,000-year old silty ice extracted from the bottom of the Greenland ice cap. The results showed that Southern Greenland was covered by a forest at that time, composed by trees of the genera *Picea*, *Pinus*, and *Alnus* as in Southern Scandinavia today [50].

Rodent middens represent another interesting sources of information for studying past communities. Middens are a mix of plant macrofossils, pollen, rodent feces, bones and insects, coming from an area within a radius of about 100 m from the den, and agglomerated by rodent's urine salts. Using animal mitochondrial and chloroplast DNA genes, it is possible to identify the rodent species and to deduce the nature of their environment from the flora and fauna present. DNA analyses carried out on 11,700-year-old middens from the Atacama Desert in Chile [51] revealed that this past

environment was more productive, more diverse, and much more humid, with a fivefold decrease in precipitation since that time.

A molecular approach was also used for studying the diet of extinct animals. Diet of the ground sloth *Nothrotheriops shastensis* was studied using the chloroplast *rbcL* gene from five coprolites in Gypsum Cave, Nevada, and dated approximately to 11,000, 20,000, and 28,500 years BP. Additionally it was possible to study the vegetation changes due to different climate condition in those three periods. According to the vegetation identified, it appeared that the climate was much drier about 11,000 years ago [52]. Another interesting study dealt with the analysis of the last meals of a Neolithic glacier mummy [53], which DNA sequences revealed to be red deer (*Cervus elaphus*) and alpine ibex (*Capra ibex*).

Limitations and Perspectives

The current DNA barcoding approach relies on a single mitochondrial or chloroplast gene, even if several regions from these organelle DNAs are sequenced because such regions are linked. The main limitation of this technique comes from this single-gene identification system. It is well known that identical mitochondrial or chloroplast DNA sequences can be present in different related species due to introgression, or due to incomplete lineage sorting since the time of speciation. Furthermore, mitochondrial or chloroplast DNA nuclear copies are common and can be preferentially amplified in some circumstances [54], leading to potential identification errors. Finally, heteroplasmy can also confuse the identification system. These potential problems have been extensively discussed (e.g. Ref. [9]). Nevertheless, these caveats do not seem to seriously compromise species identification. Even if the system is not 100% reliable,

257 from a practical point of view, it works in most cases and stimulates the building a high
258 quality database (CBOLD) that will be extremely useful for ecologists.

259 Another limitation of the current DNA barcoding approach *sensu stricto* lies in
260 the length of the sequences used, usually more than 500 bp, which prevents the
261 amplification of degraded DNA. Unfortunately, many potential DNA barcoding
262 applications can only be based on degraded DNA. This is the case for all environmental
263 samples where the target is DNA from dead animals, or dead parts of plants. It is
264 usually difficult to amplify DNA fragments longer than 150 bp from such samples. As a
265 consequence, there is a need for shorter barcoding markers. Clearly, the DNA regions
266 currently chosen as standard cannot work for those constrained to using degraded DNA,
267 because of the difficulty or even the impossibility of designing conserved internal
268 primers.

269 There is no doubt that ecologists will increasingly turn to the DNA barcoding
270 approach, because in many circumstances it represents the only easy way to identify
271 species. This trend will be further enhanced by the availability of reliable databases now
272 under construction. However, as suggested by the diverse DNA-based taxon
273 identification used up to now, ecologists will probably continue to develop diverse
274 approaches beside the standardized DNA-barcoding. There are some circumstances
275 where the number of species to distinguish is limited, and where the sequencing of long
276 DNA fragments is not necessary.

277 The possibility of analyzing complex mixtures of DNA fragments, either by DNA
278 hybridization on microarrays [20, 55] or in relation with the new sequencing
279 technologies (Box 3), opens new horizons for ecologists. We can anticipate that DNA-
280 based biodiversity assessment using environmental samples will be implemented for
281 plants and animals, as is already the usual approach for microorganisms. The same

method will be used for diet analysis. Undoubtedly DNA-based taxon identification techniques will soon fully integrate the ecologist's toolbox.

References

1. Floyd, R., *et al.* (2002) Molecular barcodes for soil nematode identification. *Mol. Ecol.* 11, 839-850
2. Hebert, P.D.N., *et al.* (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Lond. B. Biol. Sci.* 270, S96-S99
3. Hebert, P.D.N., and Gregory, T.R. (2005) The Promise of DNA barcoding for taxonomy. *Syst. Biol.* 54, 852
4. Arnot, D.E., *et al.* (1993) Digital codes from hypervariable tandemly repeated DNA sequences in the *Plasmodium falciparum* circumsporozoite gene can genetically barcode isolates. *Mol. Biochem. Parasit.* 61, 15-24
5. Chase, M.W., *et al.* (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon* 56, 295-299
6. Chase, M.W., *et al.* (2005) Land plants and DNA barcodes: short-term and long-term goals. *Philos. Trans. R. Soc. London Ser. B* 360, 1889-1895
7. Harris, D.J. (2003) Can you bank on GenBank? *Trends Ecol. Evol.* 18, 317-319
8. Ratnasingham, S., and Hebert, P.D.N. (2007) BOLD: the barcode of life data system (www.barcodinglife.org). *Mol. Ecol. Notes* 7, 355-364
9. Rubinoff, D. (2006) Utility of mitochondrial DNA barcodes in species conservation. *Conserv. Biol.* 20, 1026-1033
10. Neher, D.A. (2001) Role of nematodes in soil health and their use as indicators. *J. Nematol.* 33, 161-168
11. Lambshead, P. (2004) Marine nematode biodiversity. In *Nematology: advances and perspectives* (Chen ZX, C.Y., Chen S.Y., Dickson D.W., eds), pp. 436–467, CABI Publishing
12. Bhadury, P., *et al.* (2006) Development and evaluation of a DNA-barcoding approach for the rapid identification of nematodes. *Mar. Ecol.-Prog. Ser.* 320, 1-9

13. Powers, T. (2004) Nematode molecular diagnostics: from bands to barcodes. *Annu. Rev. Phytopathol.* 42, 367-383
14. Valiere, N., and Taberlet, P. (2000) Urine collected in the field as a source of DNA for species and individual identification. *Mol. Ecol.* 9, 2150-2152
15. Sugimoto, T., *et al.* (2006) Species and sex identification from faecal samples of sympatric carnivores, Amur leopard and Siberian tiger, in the Russian Far East. *Conserv. Genet.* 7, 799
16. Smith, M.A., *et al.* (2006) DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). *Proc. Natl. Acad. Sci. USA* 103, 3657-3662
17. Smith, M.A., *et al.* (2007) DNA barcodes affirm that 16 species of apparently generalist tropical parasitoid flies (Diptera, Tachinidae) are not all generalists. *Proc. Natl. Acad. Sci. USA* 104, 4967-4972
18. Domingo-Roura, X., *et al.* (2006) Badger hair in shaving brushes comes from protected Eurasian badgers. *Biol. Conserv.* 128, 425
19. Lee, J.C.I., *et al.* (2006) DNA profiling of shahtoosh. *Electrophoresis* 27, 3359-3362
20. Teletcha, F., *et al.* Molecular identification of vertebrate species by oligonucleotide microarray in food and forensic samples. *J. Appl. Ecol.* (in press)
21. Birstein, V.J., *et al.* (2000) Polyphyly of mtDNA lineages in the Russian sturgeon, *Acipenser gueldenstaedtii*: forensic and evolutionary implications. *Conserv. Genet.* 1, 81
22. Assefa, Y., *et al.* (2007) DNA identification of *Busseola* (Lepidoptera: Noctuidae) larvae in Ethiopian sugarcane. *Afr. Entomol.* 15, 375-379
23. Jousson, O., *et al.* (2000) Invasive alga reaches California - The alga has been identified that threatens to smother Californian coastal ecosystems. *Nature* 408, 157-158
24. Schaffelke, B., *et al.* (2002) Using genetic techniques to investigate the sources of the invasive alga *Caulerpa taxifolia* in three new locations in Australia. *Mar. Pollut. Bull.* 44, 204
25. Duffy, J.E., *et al.* (2007) The functional role of biodiversity in ecosystems: incorporating trophic complexity. *Ecol. Lett.* 10, 522-538

26. Marrero, P., *et al.* (2004) Diet of the endemic Madeira Laurel Pigeon Columba trocaz in agricultural and forest areas: implications for conservation. *Bird Conserv. Int.* 14, 165-172
27. Cristobal-Azkarate, J., and Arroyo-Rodriguez, V. (2007) Diet and activity pattern of howler monkeys (*Alouatta palliata*) in Los Tuxtlas, Mexico: effects of habitat fragmentation and implications for conservation. *Am. J. Primatol.* 69, 1013-1029
28. Agusti, N., *et al.* (2003) Collembola as alternative prey sustaining spiders in arable ecosystems: prey detection within predators using molecular markers. *Mol. Ecol.* 12, 3467-3475
29. Passmore, A.J., *et al.* (2006) DNA as a dietary biomarker in Antarctic krill, Euphausia superba. *Mar. Biotechnol.* V8, 686
30. Deagle, B.E., *et al.* (2005) Genetic screening for prey in the gut contents from a giant squid (Architeuthis sp.). *J. Hered.* 96, 417-423
31. Blankenship, L.E., and Yayanos, A.A. (2005) Universal primers and PCR of gut contents to study marine invertebrate diets. *Mol. Ecol.* 14, 891-899
32. Pons, J. (2006) DNA-based identification of preys from non-destructive, total DNA extractions of predators using arthropod universal primers. *Mol. Ecol. Notes* 6, 623-626
33. Symondson, W.O.C. (2002) Molecular identification of prey in predator diets. *Mol. Ecol.* 11, 627-641
34. Jarman, S.N., *et al.* (2004) Group-specific polymerase chain reaction for DNA-based analysis of species diversity and identity in dietary samples. *Mol. Ecol.* 13, 1313-1322
35. Bradley, B.J., *et al.* (2007) Plant DNA sequences from feces: Potential means for assessing diets of wild primates. *Am. J. Primatol.* 69, 699-705
36. Nystrom, J., *et al.* (2006) Effect of local prey availability on gyrfalcon diet: DNA analysis on ptarmigan remains at nest sites. *J. Zool.* 269, 57-64
37. Deagle, B.E., *et al.* (2007) Studying seabird diet through genetic analysis of feces: a case study on Macaroni Penguins (*Eudyptes chrysolophus*). *PLoS ONE* 2, e831
38. Taberlet, P., *et al.* (2007) Power and limitations of the chloroplast *trn*L (UAA) intron for plant DNA barcoding. *Nucl. Acids Res.* 35, e14-
39. Valentini, A., *et al.* New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trn*L approach. *Mol. Ecol.* (submitted)

40. Hajibabaei, M., *et al.* (2005) Critical factors for assembling a high volume of DNA barcodes. *Philos. Trans. R. Soc. London Ser. B* 360, 1959-1967
41. Cameron, S., *et al.* (2006) Who will actually use DNA barcoding and what will et cost? *Syst. Biol.* 55, 844
42. Janzen, D.H., *et al.* (2005) Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. *Philos. Trans. R. Soc. London Ser. B* 360, 1835
43. Sogin, M.L., *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. U. S. A.* 103, 12115-12120
44. Rusch, D.B., *et al.* (2007) The Sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern tropical Pacific. *PLoS Biol.* 5, e77
45. Herrera, A., *et al.* (2007) Species richness and phylogenetic diversity comparisons of soil microbial communities affected by nickel-mining and revegetation efforts in New Caledonia. *Eur. J. Soil Biol.* 43, 130
46. Oline, D.K. (2006) Phylogenetic comparisons of bacterial communities from serpentine and nonserpentine soils. *Appl. Environ. Microb.* 72, 6965-6971
47. Gomez-Alvarez, V., *et al.* (2007) Comparative bacterial diversity in recent Hawaiian volcanic deposits of different ages. *FEMS Microb. Ecol.* 60, 60-73
48. Blackwood, C.B., *et al.* (2007) Interpreting ecological diversity indices applied to terminal restriction fragment length polymorphism data: insights from simulated microbial communities. *Appl. Environ. Microb.* 73, 5276-5283
49. Willerslev, E., *et al.* (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science* 300, 791-795
50. Willerslev, E., *et al.* (2007) Ancient biomolecules from deep ice cores reveal a forested Southern Greenland. *Science* 317, 111-114
51. Kuch, M., *et al.* (2002) Molecular analysis of a 11 700-year-old rodent midden from the Atacama Desert, Chile. *Mol. Ecol.* 11, 913-924
52. Hofreiter, M., *et al.* (2000) A molecular analysis of ground sloth diet through the last glaciation. *Mol. Ecol.* 9, 1975-1984
53. Rollo, F., *et al.* (2002) Otzi's last meals: DNA analysis of the intestinal content of the Neolithic glacier mummy from the Alps. *Proc. Natl. Acad. Sci. USA* 99, 12594-12599

54. Zhang, D.X., and Hewitt, G.M. (1996) Highly conserved nuclear copies of the mitochondrial control region in the desert locust *Schistocerca gregaria*: some implications for population studies. *Mol. Ecol.* 5, 295-300
55. Hajibabaei, M., *et al.* (2007) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends Genet.* 23, 167
56. Lucentini, L., *et al.* (2007) A PCR-RFLP method on faecal samples to distinguish *Martes martes*, *Martes foina*, *Mustela putorius* and *Vulpes vulpes*. *Conserv. Genet.* 8, 757
57. Yamazaki, Y., *et al.* (2005) Occurrence of larval Pacific lamprey *Entosphenus tridentatus* from Japan, detected by random amplified polymorphic DNA (RAPD) analysis. *Ichthyol. Res.* 52, 297-301
58. Zinger *et al.* CE-SSCP and CE-FLA, simple and high throughput alternatives for fungal diversity studies. *J. Microb. Meth.* (in press)
59. Maldini, M., *et al.* (2006) Fish and seafood traceability based on AFLP markers: elaboration of a species database. *Aquaculture* 261, 487
60. Belfiore, N.M., *et al.* (2003) The use of nuclear and mitochondrial single nucleotide polymorphisms to identify cryptic species. *Mol. Ecol.* 12, 2011-2017
61. Herke, S.W., *et al.* (2007) A SINE-based dichotomous key for primate identification. *Gene* 390, 39-51

Boxes

Box 1. The ideal DNA barcoding system

The ideal DNA barcoding system should meet the following criteria.

(i) it should be sufficiently variable to discriminate among all species, but conserved enough to be less variable within than between species.

(ii) it should be standardized with the same DNA region used for different taxonomic groups.

(iii) the target DNA region should contain enough phylogenetic information to easily assign species to its taxonomic group (genus, family, etc.).

(iv) it should be extremely robust, with highly conserved priming sites, and highly reliable DNA amplifications and sequencing. This is particularly important when using environmental samples where each extract contains a mixture of many species to be identified at the same time.

(v) the target DNA region should be short enough to allow amplification of degraded DNA. Usually, DNA regions longer than 150 bp are difficult to amplify from degraded DNA.

Thus, the ideal DNA marker should be variable, standardized, with enough phylogenetic information, extremely robust and short. Unfortunately, such an ideal DNA marker does not exist (or at least has not been found up to now). As a consequence, according to the scientific and technical context, the different categories of users (e.g., taxonomists, ecologists, etc.) will not give the same priority to the five criteria listed above. Criteria (i), (ii), and (iii) are the most important for taxonomists, while ecologists working with environmental samples will favor criteria (iv) and (v).

Box 2. Environmental samples

An environmental sample is a mix of organic and eventually inorganic materials taken from the environment (e.g., soil or feces). It may contain live individuals (i.e., microorganisms or small macroorganisms such as nematodes or collemboles) and remains of macroorganisms present around the sampling site. Until now, environmental samples have been used mainly for studying microbial communities, using 16S rDNA or ITS as barcode. In this case, DNA sequences of several hundreds of bp can be retrieved because DNA of good quality is extracted from live microorganisms. However, environmental samples should be useful for characterising the diversity of macroorganic species (such as plant or animals) in an ecosystem. Here, the DNA is in most cases highly degraded and only short sequences can be amplified. Therefore, the new sequencing techniques (Box 3) coupled with universal primers amplifying short fragments make it possible to identify the different taxa. A few studies were already carried out using different types of environmental samples:

- Water [44]
- Soil [45-47,49]
- Ice cores [50]
- Middens [51]
- Feces samples for diet analysis [37, 39]

Box 3. The next-generation sequencing systems

The current barcoding system (CBOL approach) has been designed to fit with DNA sequencers based on capillary electrophoresis, with typical read length of 500-1000 bp. Recently, next-generation sequencing systems have become available. Several new techniques have been implemented, all based on a massively parallel approach, and sequencing individual molecules (with or without an amplification step) (see Table below). All new sequencers but one produce very short fragments. The only system that allows sequencing fragments longer than 25-35 bp is the 454 GS FLX (Roche) that currently deliver 200-300 bp fragments (an upgrade of the system is already announced, multiplying by about ten the total output, with fragments of 400 bp). The enormous amount of relatively long sequences produced make this new sequencer suitable for environmental barcoding studies where there is the need to deal with complex samples composed of a mix of many species (e.g. deep sea biodiversity [43] and diet analysis [39]).

Sequencer	Genetic Analyzer™/Solexa™	SOLiD™ DNA Sequencer	Heliscope™	454 GS FLX™
Company	Illumina®	Applied Biosystems®	Helicos®	Roche Diagnostics®
website	www.solexa.com	solid.appliedbiosystems.com	http://helicosbio.com/	https://www.roche-applied-science.com/sis/sequencing/flx/index.jsp
Fragments length	25-50 bases	25-35 bases	25-50 bases	250 bases
Number of reads in	60 000 000 reads per run	85 000 000 reads per run	100 000 000 bases per hr	400 000 reads per run
Total output (Gb = giga bases)	2 Gb per run	3 Gb per run	2 Gb bases per day	0.1 Gb per run
Time per run	6.5 days	6 days	------	7.5 hours

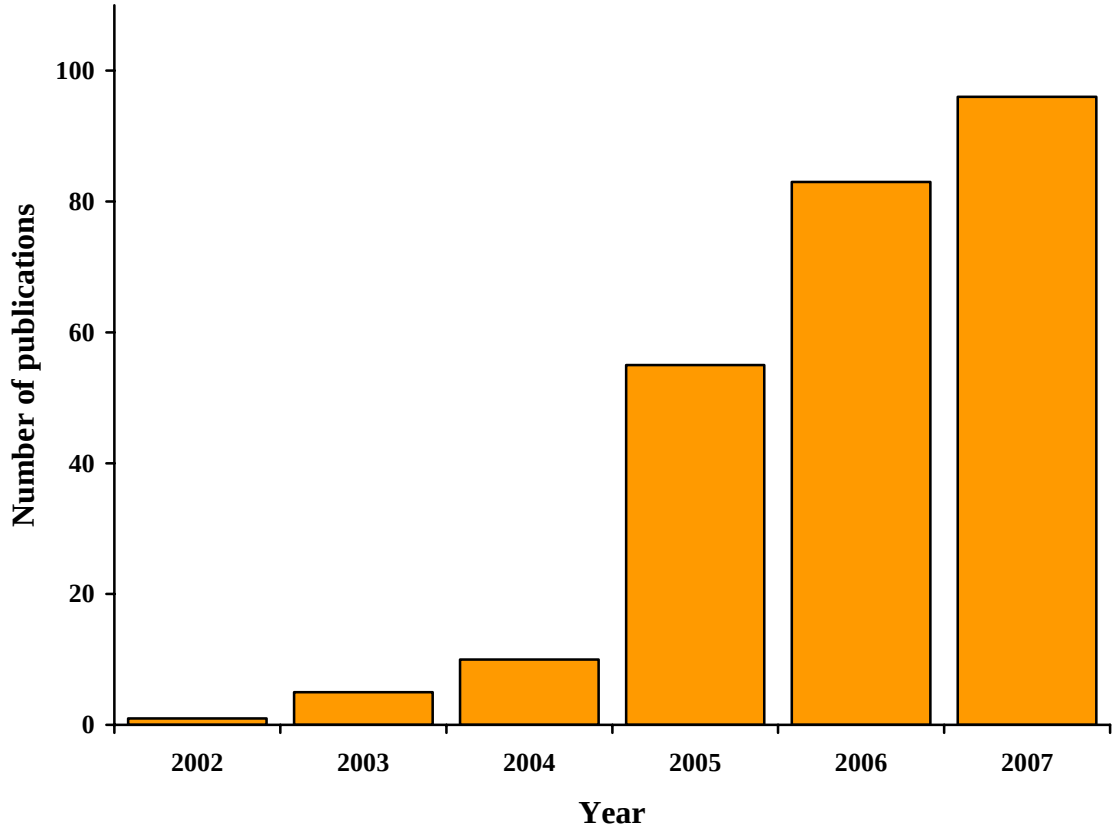
485 **Table 1. Examples of DNA-based molecular techniques used by ecologists for**
486 **species identification.**
487

Technique	Method	Application	Refs
Restriction Fragments Length Polymorphism (RFLP)	Variation in length of restriction fragments between individuals due to point mutation in the restriction enzyme site	Identification using feces samples of three different species of Mustelids and one species of Canids that live in sympatry	[56]
Random Amplified Polymorphic DNA (RAPD)	Difference in amplification from genomic DNA with a single short (usually 10-bases) primer	Identification of four different species of lamprey larvae	[57]
Single Strand Conformational Polymorphism (SSCP)	Change in the conformation of a single-stranded DNA depending on its nucleotide sequence	Determination of 19 fugal species using SSCP profile	[58]
Amplified Fragments Length Polymorphism	Selective amplification of restriction fragments from digested genomic	Identification of 32 different species of fish, molluscs and	[59]

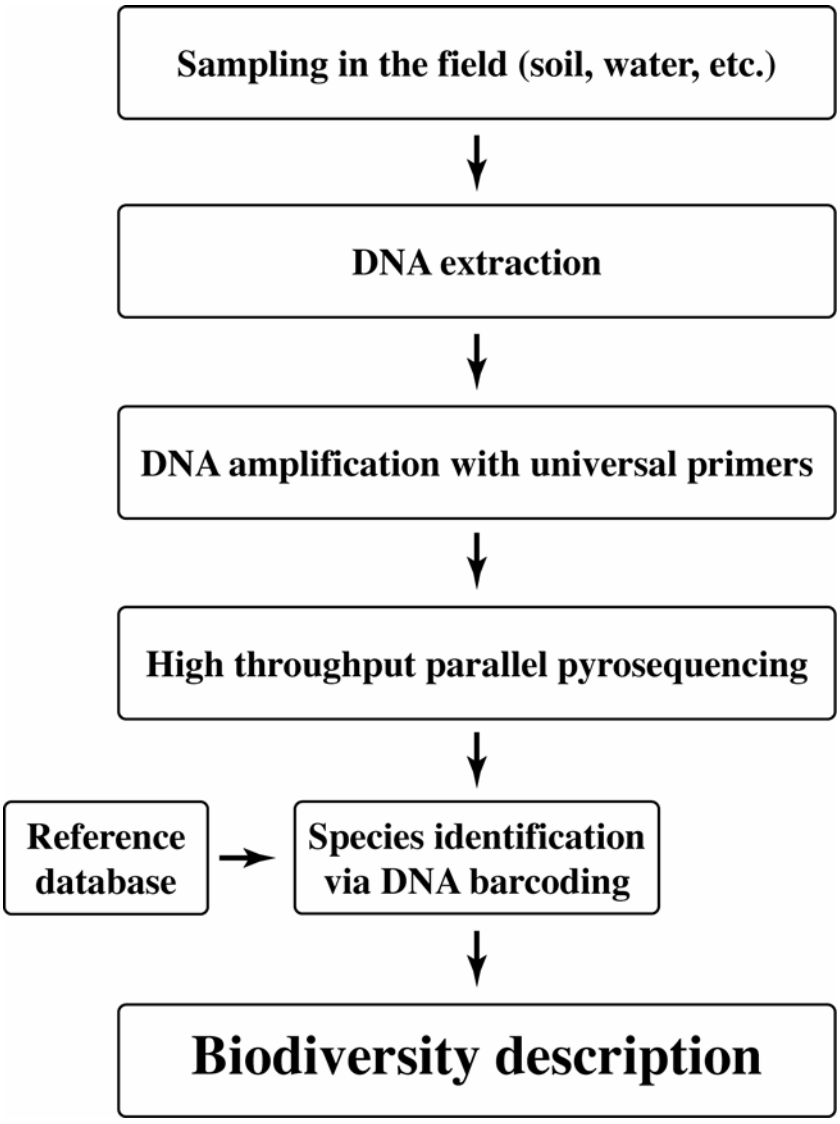
(AFLP)	DNA	crustaceans from processed food	
Single Nucleotide Polymorphism (SNP)	The variation of a single nucleotide in DNA sequences	Identification of four cryptic species of *Microtus*	[60]
short interspersed repetitive elements (SINE)	Repetitive sequences of DNA of about 300 base pairs in length that occur about every 3000-5000 bp in the genome	56 primates species identification using *Alu* elements	[61]
Microarrays	Amplification of a mtDNA cytochrome *b* fragment and hybridization on a microarray containing diagnostic oligonucleotides	Identification of 71 vertebrate species, including birds, mammals, actinopterygians and chondrichtyans	[20]
DNA Sequences	The precise ordering of nucleotides of which the DNA is composed	Identification of 200 closely species of lepidopterans	[2]

488

Figure 1. Number of DNA barcoding publications since 2003. Source: ISI web of science, with the topic "DNA barcod*" OR "molecular barcod*" followed by a manual filtering (12 publications removed). The number of DNA barcoding publications greatly increased in the last three years.



496 **Figure 2.** Methodology for analyzing biodiversity from environmental samples based
497 on next generation DNA sequencers.
498
499
500



501

Figure 3. Example of feces analysis for estimating diet composition using a DNA barcoding approach (from Ref. [39]). Comparison of brown bear (*Ursus arctos*) and golden marmot (*Marmota caudata*) diets in Himalayan environment. Twelve feces of each species were analyzed. After DNA amplification with universal primers, and sequencing on the 454 GS FLX sequencer, plant taxa were identified by comparison with available reference sequences.