UNIVERSITA' DEGLI STUDI DI PADOVA

DOCTORATE SCHOOL OF CROP SCIENCE

CURRICULUM AGROBIOTECNOLOGY – CYCLE XXII

Department of Enviromental Agronomy and Crop Science

BIODIVERSITY ANALYSIS THROUGH DNA BARCODING

Applications in agrifood and seafood products

Director of the school: Ch.mo Prof. Andrea Battisti

Supervisor: Ch.mo Prof. Gianni Barcaccia

PhD student : Silvia Nicolè

DATE OF THESIS SUBMISSION

February 01st, 2010

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Silvia Nicolè / February 1st, 2010

A copy of the thesis will be available at http://paduaresearch.cab.unipd.it/

Dichiarazione

Con la presente affermo che questa tesi è frutto del mio lavoro e che, per quanto io ne sia a conoscenza, non contiene materiale precedentemente pubblicato o scritto da un'altra persona né materiale che è stato utilizzato per l’ottenimento di qualunque altro titolo o diploma dell'università o altro istituto di apprendimento, a eccezione del caso in cui ciò venga riconosciuto nel testo.

Silvia Nicolè, 1 Febbario 2010

Una copia della tesi sarà disponibile presso http://paduaresearch.cab.unipd.it/

Table of contents

Riassunto

Capitolo 1 - Negli ultimi decenni, l’impiego del DNA ribosomale per la ricostruzione delle relazioni evolutive tra specie è stato gradualemente sostituito da approcci di analisi di DNA mitocondriale per studi di biodiversità. La valutazione del polimorfismo genetico a livello di DNA è stata estensivamente usata per comprendere la tassonomia di diversi gruppi di organismi e per identificare singoli organismi. Sebbene l’identificazione delle specie tramite DNA *fingerprinting* non sia un concetto nuovo, solo adesso l’approccio, con il nome di “DNA *barcoding*”, sta riscuotendo un notevole successo e sta rivoluzionando il sistema di indagine tassonomico. Paul Hebert dell’Università di Guelph, in Canada, ha proposto di utilizzare la variabilità presente nella sequenza nucleotidica di un gene target come “firma molecolare” unica per catalogare la biodiversità. Una breve porzione del gene mitocondriale cox1, codificante per l’enzima citocromo c ossidasi subunità I, è stata proposta come “barcode” potenziale. Il concetto chiave alla base del DNA *barcoding* è l’esistenza del “*barcoding gap*”, una discontinuità tra la variabilità intra ed interspecifica, e precisamente è stato sperimentalmente dimostrato che la variazione nucleotidica all’interno di una specie è generalmente 10 volte inferiore alla variabilità nucleotidica riscontrata tra specie. Al momento sono attivi numerosi progetti di DNA *barcoding* che hanno dimostrato l’efficacia di questa tecnica in diversi gruppi animali. Nel 2004 è stato fondato il Consortium for the Barcode of Life (CBOL) che riunisce molte organizzazioni come musei zoologici, erbari, centri di ricerca pubblici e diversi enti privati, con l’obiettivo di promuovere lo sviluppo di un sistema tassonomico universale per le specie eucariotiche, una sorta di “inventario della vita” (Barcode of Life Initiative), e la creazione di un database pubblico costituito da sequenze di riferimento ottenute da campioni di identità certa. La metodologia proposta potrebbe rivelarsi utile in numerosi settori scientifici, quali la biologia evoluzionistica, l’ecologia, la biogeografia e la biologia della conservazione, ed avere numerosi riscontri pratici. Interessanti applicazioni riguardano le scienze forensi, il monitoraggio del commercio internazionale di prodotti di origine animale e vegetale (regolamentazioni CITES, convenzione sul commercio internazionale delle specie di flora e fauna minacciate di estinzione), la biosicurezza e la sicurezza alimentare. In quest’ultimo settore, il DNA *barcoding* potrebbe venir sfruttato per il riconoscimento dei prodotti derivanti dall’impiego di specie protette e in via di estinzione e per prevenire casi di falsificazione alimentare.

Capitolo 2 – La frequente sostituzione di tranci o filetti di specie ittiche pregiate con carni di esemplari di minor valore o l'utilizzo di nomi generici usati per etichettare i prodotti della pesca ha messo in luce la necessità di sviluppare un sistema di tracciabilità molecolare degli alimenti di origine animale. L'impossibiità di ricorrere al riconoscimento morfologico quando il pesce è sottoposto a "toelettatura" richiede lo sviluppo di nuovi approcci analitici, basati sullo studio del DNA e il DNA barcoding si è rivelato un promettente strumento diagnostico alternativo ai tradizionali metodi di indagine e a quelli basati sull'analisi delle proteine. Dal momento che tale ricerca era finalizzata all'identificazione delle specie utilizzate per la preparazione degli alimenti e all'individuazione di eventuali casi di falsificazione, si è proceduto ad una estesa indagine di mercato al fine di scoprire le specie maggiormente coinvolte in casi di sostituzione fraudolenta. Una volta ottenute queste informazioni, sì è proceduto con il reperimento di 37 campioni da analizzare, freschi, congelati o processati, appartenenti a tre diversi gruppi tassonomici, pesci, molluschi e crostacei.

La procedura sperimentale ha previsto l'adozione di un approccio multi-locus basato sull'amplificazione, con primer universali, e il sequenziamento di tre regioni mitocondriali, i geni cox1, cob e 16S-rDNA. Successivamente, sono state condotte un'analisi di similarità di sequenza, usando BOLD and GenBank come database di riferimento, il calcolo delle matrici di distanza e la costruzione di un albero Neighbor-Joining per attribuire un'identità ai nostri campioni. In generale, il DNA *barcoding* ha dimostrato di essere un efficiente strumento per identificare campioni di origine sconosciuta e quindi per controllare le informazioni fornite nelle etichette dei prodotti. Infatti, l'analisi ha confermato, sulla base almeno di una regione mitocondriale, la specie dichiarata nell'etichetta in 32 casi tra quelli analizzati. In contrapposizione, il 13% dei campioni è risultato frutto di un probabile evento di sostituzione, volontaria o accidentale, con un individuo appartenente ad una specie differente.

Capitolo 3 – L'impiego del DNA *barcoding* potrebbe rivelarsi utile, non solo per il riconoscimento di specie vegetali di interesse agronomico, ma anche per la tracciabilità genetica delle varietà e dei loro derivati alimentari, senza la valutazione dei tratti morfologici. Invece di usare il genoma mitocondriale, per il DNA barcoding delle piante il miglior candidato è risultato il DNA cloroplastico che possiede gli stessi attributi di quello mitocondriale.

Per quanto riguarda il materiale vegetale, sono state campionate diverse linee pure di fagiolo (*Phaseolus vulgaris* L.), appartenenti a landrace selvatiche e domesticate e a varietà moderne coltivate, insieme ad alcune accessioni di *P. coccineus*, *P. lunatus* and *Vigna unguiculata*, usate come fuori-gruppo. Un approccio multi-locus ha previsto l'amplificazione di sette regioni cloroplastiche, tre codificanti (*rbcL, trnL* e *matK*) e quattro spaziatori intergenici (*rpoB-trnC, atpBrbcL, trnT-trnL* e *psbA-trnH*), e due nucleari, ITS1 e ITS2. I principale obiettivi della ricerca erano individuare i marker e gli SNP con la miglior capacità discriminante a livello di varietà, testare due distinti metodi analitici (uno basato sulle distanze genetiche e uno sulla condivisione dei caratteri diagnostici) per indagini di biodiversità e studi di tracciabilità genetica e infine valutare l'utilità del genoma cloroplastico in generale per la ricostruzione dell'origine delle moderne varietà di fagiolo in relazione ai due pool genici principali, quello Mesoamericano e quello Andino. La caratterizzazione molecolare ha previsto: I) l'amplificazione e il sequenziamento di distinte regioni cloroplastiche e nucleari; II) l'*editing* e l'allineamento delle regioni nucleotidiche; III) la stima delle distanze genetiche e la costruzione del NJ; IV) l'impiego dell'approcio basato sull'individuazione dei caratteri diagnostici informativi, SNP e In/Del, associati ad uno o più gruppi tassonomici. L'approccio fenetico ha confermato di essere un efficace strumento per l'identificazione delle specie perché ha separato membri appartenenti a specie diverse e ha raggruppato accessioni corrispondenti a membri della stessa specie. A livello di varietà, invece, il metodo si è rivelato scarsamente informativo per discriminare i due diversi pool genici e infatti tutte le accessioni afferenti alla specie *P. vulgaris* sono state raggruppate in pochi sottogruppi con bassi valori di *bootstrap*. Perciò si è ricorsi ad un sistema basato sulla condivisione dei caratteri diagnostici e tale approccio si è rivelato utile per definire 16 aplotipi all'interno della specie *P. vulgaris,* sulla base delle regioni cloroplastiche analizzate, corrispondenti ad altrettanti sottogruppi, ognuno costituito da accessioni Mesoamericane o Andine. Le accessioni italiane, invece, tendevano a clusterizzare prevalentemente con il pool genico Andino confermando l'origine Andina dei fagioli comuni italiani. A differenza delle regioni cloroplastiche, le regioni nucleari sono risultate scarsamente informative e la maggior parte dei genotipi hanno formato un unico aplotipo, eccetto per le accessioni corrispondenti agli ancestrali che hanno formato un gruppo separato.

Capitolo 4 – Un terzo caso di studio è rappresentato da *V. vinifera*, la più importante specie della famiglia delle Vitaceae conosciuta per il suo impiego nella produzione di vino. La ricerca è stata finalizzata allo studio delle potenzialità del DNA *barcoding* per la distinzione delle più comuni varietà di vite destiante alla tavola o alla produzione enologica. Si è proceduto con la selezione di 144 genotipi insieme con altre 5 accessioni appartenenti a diverse specie di *Vitis*, adottate come fuori-gruppo. Dopo lo studio pilota condotto in fagiolo, l'applicazione della tecnica si è focalizzata inizialmente in un'indagine preliminare del genoma cloroplastico, ma è parso subito evidente l'insufficiente grado di variablità genetica di tale DNA per distinguere le varietà. Infatti le sette regioni cloroplastiche testate sono risultate monomorfiche non solo tra varietà, ma anche tra le sei specie di *Vitis*. Da qui la decisione di passare allo studio del genoma nucleare: sono state amplificate quattro regioni EST, precedentemente impiegate per la valutazione della variabilità genetica di *V. vinifera*, e il gene GAI1, usato per la ricostruzione della filogenesi nella famiglia delle Vitaceae. L'analisi è ancora in corso, ma risultati preliminari indicano che numerosi SNP esistono tra cultivar, sia allo stato omozigote che eterozigote. Infatti, un problema sollevato dall'impiego di regioni nucleari risiede proprio nella rilevazione di casi di additività, attribuibili sia alla natura altamente eterozigote della specie, sia come conseguenza di eventi di ibridazione. Dall'analisi iniziale di tre delle cinque regioni nucleari amplificate, sembrano confermate le potenzialità della tecnica per identificare specie diverse, mentre a livello varietale la variabilità genetica e quindi la distinguibilità dei genotipi è meno marcata. Precisamente, tra i 149 genotipi studiati, è stato possibile ricostruire 63 aplotipi di cui 38 cultivar- specifici, mentre nei restanti casi più cultivar venivano raggruppate insieme. La definizione degli aplotipi ha permesso di definire non solo alcuni SNP sfruttabili per il riconoscimento delle cultivar, ma anche di confermare alcune ipotesi avanzate circa l'origine di alcune cultivar, come per esempio eventuali casi di sinonimia e omonimia. I dati ottenuti fino ad ora dimostrano che il DNA barcoding applicato al genoma nucleare potrebbe essere uno strumento utile per il *fingerprinting* di cultivar di vite sia per studi di biodiversità che per scopi di tracciabilità alimentare, applicata anche a prodotti derivati, come i vini.

Capitolo 5 – La strategia del DNA barcoding potrebbero rivelarsi estremamente utile per la vita quotidiana in quanto potrebbe contribuire all'identificazione univoca di specie in tutte quelle situazioni in cui i tratti morfologici sono di valore limitato. In sintesi, tali ricerca ha permesso di:

- testare il potere diagnostico del gene mitocondriale cox1 come marcatore genetico specie-specifico e dimostrare la sua utilità per la tracciabilità genetico-molecolare applicata a prodotti alimentari di origine marina;
- spingere la tecnica del DNA *barcoding* fino al caso limite della SNP *detection* per distinguere entità genetiche infra-specie (varietà) all'interno di due specie coltivate ed economicamente rilevanti, quali *P. vulgaris* e *V. vinifera*, rivelando la sua abilità nella definizione di aplotipi cultivar-specifici;
- porre le basi per il futuro sviluppo di saggi diagnostici più rapidi ed affidabili, basati sulla costruzione di una piattaforma *microarray*, che consentiranno il riconoscimento genetico di materiali animali e vegetali e derivati trasformati di carne, semi e frutti.

Summary

Chapter 1 - In the last decades, the employment of ribosomal DNA to infer the phylogentic relationships among organisms was gradually substituted by the analysis of mitochondrial DNA for biodiversity studies and molecular systematics. The detection of nucleotide polymorphisms was extensively used to understand the taxonomy of several taxa and to identify single organisms. Although the species identification through DNA typing is an old concept, only now the approach under the label of “DNA barcoding” is gaining an incredible success and is revolutionizing the way to practice taxonomy. Paul Hebert of the University of Guelph, in Canada, proposed the use of this term to describe the technique that exploits a short DNA sequence, a barcode, from a standardized region of the mitochondrial genome, precisely citochrome oxydase I (cox1), as a universal and unique identification marker for animal species. The core idea of DNA barcoding is the existence of “barcoding gap”, a discontinuity between the intra- and interspecific divergence values, precisely the variation of the nucleotide sequences within species is proved to be usually 10 fold less than the differences among species. Several projects have demonstrated the effectiveness of this approach in many groups of animals.

In 2004 Consortium for the Barcode of Life (CBOL) was launched and joined several organizations as natural history museums, herbaria, research centres and private patterns with the purpose of promoting the development of universal system for eukaryotic species inventory (Barcode of Life Initiative) and the creation of a public database of documented and vouchered reference sequences.

DNA barcoding can turn out of great support for many aspects of the life because it can facilitate rapid and large-scale biodiversity surveys, both for several research fields, such as evolutionary biology, ecology, biogeography and conservation biology, and also for many practical uses. These applications range from forensic science, international trade monitoring (CITES regulations), biosecurity, e.g. for surveillance of disease vectors, to the food traceability. In the food sector, DNA barcoding could be valuable for recognizing products prepared from protected and threatened species and for preventing the mislabelling of commercial species.

Chapter 2 - The seafood certification is gaining particular attention because it was demonstrated that mislabeling of fish products, fraudulent or not, and the use of vernacular or generic labels for fisheries that contain both sustainable and non-sustainable fished species are known to occur. The lack of morphological features, lost when the fish is filleted or processed, makes the traditional authenticity tests impossible to carry out. Therefore the species identification demands the development of new analytical methods and molecular techniques based on DNA analysis, in particular DNA barcoding, have proven to be an promising tool, alternative to the traditional methods and those based on protein analysis.

Since the research purpose was to assay the potentials of DNA barcoding technique as tool of diagnosis for the identification of seafood components to detect cases of fish substitution, an intensive search of the most common species, involved in mislabeling and substitution events, were conducted. Once completed, we proceeded with the collection of 37 samples to analyze, including raw, frozen and processed commercial seafood, from three different taxonomic groups, fishes, molluscs and crustaceans. The experimental procedure adopted was a multi-locus approach based on the amplification and sequencing of three mitochondrial markers, cox1, cob and 16S-rDNA genes, using universal primer pairs. After that, a sequence similarity search, using BOLD and GenBank as reference databases, and the computation of distance matrices and building of NJ tree to assign the identity of the specimens were performed. Overall, the technique proved to be an efficient tool to ensure the correct detection of food composition and thus to control the label information. In fact, 32 samples were correctly identified and, on the basis of at least one region, it was possible to confirm the origin of the meat declared on the label. On the opposite, about 13% of the analyzed samples were shown to be most likely substituted, voluntary or by accident, with different species.

Chapter 3 - The employ of DNA barcoding to crop plants could turn out valuable to accurately identify species and also for genetic traceability of varieties and food derivates, without scoring morphological traits. Instead of using the mitochondrial genome, for DNA barcoding of plants the best candidate genome is represented by the chloroplast one that owns the same attributes of the mtDNA. The technique was applied to several pure lines of

Phaseolus vulgaris belonging to wild, domesticated and cultivated common beans, along with a few *P. coccineus*, *P. lunatus* and *Vigna unguiculata* accessions. A multilocus approach was exploited using three chloroplast genic regions (*rbcL, trnL and matK*) and four intergenic spacers (*rpoB-trnC*, *atpBrbcL, trnT-trnL* and *psbA-trnH*) together with the nuclear ITS1 and ITS2. The main goals were to provide the markers and SNPs showing the best discriminant power at variety level in common bean germplasm, to test two distinct methods (*i.e*. tree-based versus character-based) for biodiversity analysis and traceability assays and to evaluate the overall utility of plastidial DNA barcodes for reconstructing the origin of modern Italian varieties in relation to the two main gene pools, Mesoamerican and Andean ones. The experimental strategy included the following steps: i) amplifying and sequencing of the distinct cpDNA regions along with the ITS1-ITS2 for rDNA regions; ii) editing and alignment of sequences; iii) clustering of sequences by NJ method supported by bootstrapping analysis; iv) character-based method that consists in the identification of taxonomic groups through the sharing of specific informative character states, SNPs or In/Dels, narrowed to one nucleotide position or extended to multiple positions. Our results indicated that the phenetic approach, based on the computation of a distance matrix and the derived NJ tree, confirmed to be a powerful technique to correctly separate different species and to cluster together accessions corresponding to members of the same species. At the varietal level, on the opposite, this method revealed to be scarcely informative to discriminate gene pools and to identify varieties within *P. vulgaris* since all the accessions tend to group in few subgroups with low bootstrapping values. Thus a second approach, the character-based system, was tested and it revealed to be useful to detect within *P. vulgaris* species a total of 16 haplotypes, over all cpDNA regions, corresponding to as many subgroups, each one made up by Mesoamerican or Andean accessions. Instead, the Italian accessions tended to cluster with one or the other gene pool, even if most of the Italian commercial varieties grouped with the Andean pool confirming the Andean origin of the Italian common beans. Differently from chloroplast DNA regions, as expected, the nuclear ITS data set of *P. vulgaris* resulted poorly informative and almost all accessions were clustered together in one single group, except for the ancestral entries that clustered apart.

Chapter 4 - A third study case is represented by *V. vinifera*, the most important species of the Vitaceae family, known for its employment for the production of wine. The study aimed at investigating the potentials of DNA barcoding to distinguish the most common grapevine cultivars destinated to table consumption of to the production of wines. We proceeded with the selection of 144 grapevine genotypes along with other 5 accessions of *Vitis* spp. adopted as reference standards and out-types. After the pilot study conducted in bean, the application of the technique in grapevine was initially focused on the use of chloroplast DNA, but from a preliminary analysis of the cpDNA, it was evident that this genome was not enough variable to distinguish grapevine cultivars. In fact all the seven chloroplast markers tested resulted to be monomorphic not only among varieties, but also among the six species within the genus *Vitis*. Thus we moved beyond to the nuclear genome and amplified precisely four ESTs, previously employed for SNP detection in grapevine, and the GAI1 gene, already used for the construction of phylogeny of Vitaceae family. The analysis is still ongoing, but the preliminary results indicate that several SNPs exist among cultivars, both at homozygous and heterozygous status. The problem of using nuclear regions relies on the detection of additive patterns that may be symptom of hybridization event. From the initial analysis of three out of the five markers, it seems confirmed the potentials of the technique to identify different species, while at sub-species level the genetic variability and thus the distinctiveness of the genotypes seem less marked. Precisely, among the 149 genotypes studied, it was possible to define 63 haplotypes of which 38 were cultivar-specific, while the other cases grouped several varieties at the same time. The haplotype reconstruction allowed not only to define some SNP markers exploitable for cultivar recognition, but also to corroborate some hypothesis, regarding the origin of some local cultivars, thought to be involved in misidentification events (synonymy/homonymy). The obtained data proved that a SNP detection technique applied to the nuclear genome could be a suitable tool for grapevine fingerprinting useful for biodiversity and food traceability aims.

Chapter 5 –The DNA barcoding assay could be of great support to the everyday life because it can provide valuable information to unequivocally distinguish species in all those situations where morphological characters are of limited or null value. Overall, the present research allowed to:

- testing the diagnostic power of the mitochondrial cox1 as genetic species-specific tag and proving its utility for the molecular traceability applied to seafood derivates;
- pushing the barcoding technique toward the limit case of SNP detection to identify genetic entities below the species level (variety) for two important crop species, such as *P. vulgaris* and *V. vinifera*, demonstrating its ability for the definition of cultivar-specific haplotypes;
- putting the basis for the future development of faster and reliable diagnostic assays, based on microaray technology, suitable for the genetic recognition of animal and plant materials and marine, seed and fruit-derived products.

Chapter 1

General introduction

”Biodiversity and taxonomic crises”

The biodiversity, intended as “the biological diversity among living organisms from all sources, including terrestrial, marine and other aquatic ecosystems, and the ecological complexes of which they are part” (International Convention on Biological Diversity, 1992; http://www.cbd.int), has emerged in the nineties as a topic of growing concern for sustainable development. Taxonomy is the science that deals with the definition, diagnosis, description and naming of organisms and the subsequent organization of this information into systems of classification (Lipscomb, 2003). Species identification is essential for large-scale biodiversity monitoring and conservation and the measuring of species richness is the most useful indicator of biodiversity. Initially, most species were differentiated by their adult morphology but more sophisticated approaches have been added over the generations. Electron microscopy, behavioural traits and biochemical markers became all tools that taxonomists have acquired to improve the science of taxonomy (http://www.barcoding.si.edu).

The first system of cataloguing of species was founded more than 250 years ago by the Swedish naturalist Carl von Linné (1707-1778) who began the formal taxonomy by means of the introduction of the binomial species nomenclature (including the genus and species name), relied mainly on morphology, to describe the biodiversity (Linneus, 1756). His pioneer work represented a milestone toward a classification system of the species, even if he underestimated the real biological diversity on the Earth.

Currently taxonomic knowledge is far from complete. Up to now, using morphological and behavioural observations and more recently biochemical markers, taxonomists were able to identify, describe and classify just a fraction of the estimated species. Although approximately 1.7 million species have been described, the majority of species on the Earth remains still unknown and it is estimated to vary widely, from 5 millions to more than 100 millions (Hawksworth and Kalin-Arroyo, 1995; http://tolweb.org/tree/). The gap in our knowledge can be split into two types: whereas above the generic level, discovery of new families, orders and phyla is rare, at the species and genus level we ignore most of the diversity in many taxa. Furthermore, there is a clear bias of focus on particular groups, mainly larger eukaryotes, such as vertebrates or flowering plants, while for smaller taxa that require expert skills for correct identification, such as nematodes, insects and microorganisms, the percentage of known diversity is definitely lower (Blaxter, 2003). It is estimated that less than 10% of vertebrates remain to be described, but more than 50% of terrestrial arthropods and up to 95% of protozoa are undescribed (www.cbd.int).

Unfortunately the global biodiversity is being lost at an unprecedented rate, 50-100 times the natural rate, as result of human activities that are responsible for an increase of extinction rates of many species (www.cbd.int; Newmaster *et al*., 2006). At the same time, we are assisting to a "taxonomic crisis": part of the biodiversity will remain unknown because the work of cataloguing species with traditional morphological methods is long, laborious and demands high level of expertise, not common (Hebert *et al*., 2003a). In addition, the "morphological taxonomy" revealed to be inadequate to account the Earth's biodiversity because of other three limitations. First, omoplasy (Vences *et al.*, 2005) and phenotypic plasticity to environmental factors (Saunders, 2005) of a given diagnostic character employed for species recognition can lead to an incorrect identification. Second, this approach overlooks morphologically cryptic taxa, such as sibling species (*i.e.* morphologically identical species, but genetically different) that are common in many groups (Knowlton, 1993; van Velzen *et al.*, 2007). Third, since morphological keys are often effective only for a particular life stage or gender, many individuals, mainly in their juvenile stages, cannot be identified (Pegg *et al*., 2006). Therefore, even if the binomial Linnaean naming system is well established and broadly used, its incapacity to solve these crisis, caused by the combination of the erosion of Earth's biodiversity and severe impediments to taxonomic research, has led to seek new adequate species identification instruments for cataloguing the biodiversity. DNA-based taxonomy could reveal a valuable support to the classic taxonomy allowing to cope with the growing need of accurate and accessible taxonomic information (Tautz *et al*., 2003).

The answer of DNA-based taxonomy

A taxonomic character is defined as "any feature of a subject of a taxon that marks the difference with the subject of another taxon" (Ayala, 1983). It has long been recognized that DNA sequence diversity, whether assessed directly or indirectly through protein analysis, can be used to discriminate species because the nucleotide composition of the genome is specific of a given species (Manwell and Baker, 1963). Microgenomic identification systems permit life's discrimination through the analysis of the nucleotide polymorphisms of a small segment of the genome (Hebert *et al*., 2003a). The advantage to use directly DNA, rather than proteins, is that this molecule is relatively stable allowing its extraction from many different types of samples, including museum specimens with damaged DNA, and from all stages of life (Blaxter, 2004). Furthermore, DNA analyses are independent of the tissue origins (*e.g*. muscle, gonad, bone, etc.) because all cell types contain identical genetic information and the DNA information content is higher compared to that of proteins, because of the degeneracy of the genetic code (Civera 2003).

The employment of a DNA-based system to investigate evolutionary relationships was first applied by Carl Woese who recognized the existence of the Archea domain by using the highly conserved 16S-rDNA gene coding for the small ribosomal subunit (Woese and Fox, 1977). Subsequently, this approach was further exploited in several taxonomic groups with few morphological diagnostic characters as viruses, protests and bacteria (Nanney, 1982; Pace, 1997; Allander *et al*., 2001). This approach, known as "DNA taxonomy", differs from DNA barcoding because it does not aim to link the genetic entities recognized through sequence analysis with Linnean species and thus it is most useful for groups of organisms that lack detailed taxonomic systems (Blaxter, 2004). In this case, the development of an universal system led to the introduction of the term "Molecular Operational Taxonomic Unit" (MOTU) (Floyd *et al*., 2002; Blaxter *et al*., 2005). For those organisms, such as meiofauna (Markmann and Tautz, 2005) or microorganisms, the concept of MOTU was largely applied to describe clusters of genetic entities that are recognized exclusively on the basis of the sequence similarity without any reference to the species name imposed with Linnaean binomial classification.

According to Tautz's idea, instead, the DNA-based taxonomy system by means of detection of nucleotide sequence differences in a single gene for the identification of the organisms, would represent just an additional tool for assigning taxonomic status, through matching the DNA sequence to a species already labelled with Linnaean name, without giving to it a central role (Godfray, 2002; Tautz *et al*., 2003). This approach considers DNA-based system as a "new scaffold for the accumulated taxonomic knowledge" and does not want to be a replacement, but only a plea for the conventional taxonomy. Infact, as none would use a single morphological character to define or identify an organism, DNA sequence alone would not be sufficient to characterize a species (Ferguson *et al*., 2002), except for some character-poor organisms, such as soil nematodes, but an integrative approach, combining broad range of data from phenotypic traits to molecular markers, could add robustness to the species recognition (Dunn, 2003; Will *et al*., 2005; Padial and De La Riva, 2009; Smith *et al*., 2007). The introduction of DNA-based taxonomy system, integrating the traditional taxonomy, was proposed in 2002 in Munich, Germany, during the DNA Taxonomy Workshop where it was discussed the idea to use the DNA as a new character for a taxonomic reference system and which markers could be the most suitable for this purpose.

DNA barcoding: a new name for an old concept

The first time that the term “DNA barcoding” appeared was in 1993 to designate an universal DNA typing system. The group led by Arnot developed a molecular approach in parasitology based on the detection of allelic sequence variation of a specific target locus (Arnot *et al*, 1993). However this concept did not gain much attention until 2002, date of the first DNA barcoding publication. Paul Hebert of the University of Guelph, Ontario, Canada, proposed the use of this term to describe the technique that exploits a short DNA sequence, a barcode, from a standardized region of the genome as a universal and unique identification marker for animal species (Hebert *et al.*, 2003a). The system entails detecting nucleotide polymorphisms of a nucleotide snippet, 648 bp in length, from the 5’ end of the mitochondrial locus coding for the cytochrome c oxidase subunit 1 (cox1), from ideally all metazoans.. This sequence should contain enough unique information, in terms of SNPs (Single Nucleotide Polymorphisms) and In/Dels (Insertion/Deletions), shared among individuals of a species with slight variations, but specific of one species. The core idea of DNA barcoding is the existence of “barcoding gap” (**Figure 1**) that means that the variation of the nucleotide sequences within species is much less than the differences among species (Hebert *et al.,* 2003a).

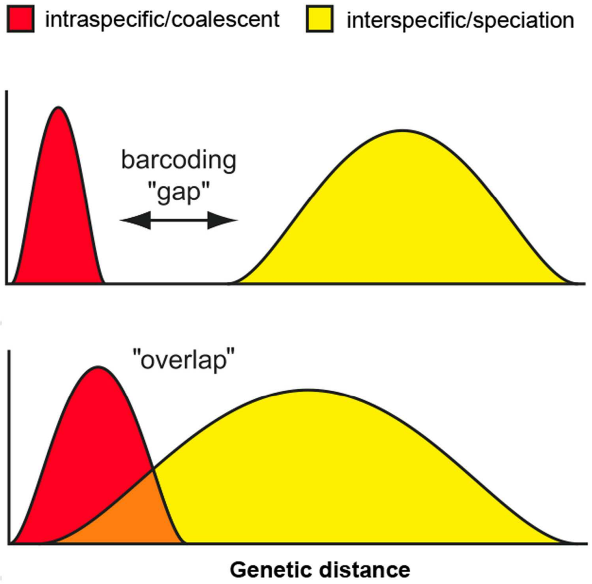


Figura 1: Schematic representation of the inferred barcoding gap (from Meyer and Paulay, 2005).

DNA barcoding aims to provide a rapid and reliable tool for species-level identification by comparing a short DNA sequence from an unknown specimen to a comprehensive library of reference ortologhous sequences related to verified and vouchered specimens of established identity (Hajibabaei *et al*., 2006a). The two essential components for an effective DNA barcode system are the standardization on an uniform barcode sequence, such as cox1 gene, and a library of sequences linked to named voucher specimens (Hebert *et al*., 2004a). Thus, the sequence of the target gene has been likened to the Universal Product Codes of manufactured products employed in the markets to identify all products sold, but instead of 10 alternate numbers at 11 positions, genomic barcodes have only four alternate nucleotides at each position with a huge string of sites available (Hebert *et al*., 2003a). It is calculated that 15 variable sites in cox1 gene provide one billion different nucleotide combinations corresponding to as many DNA barcode patterns, even if only a relatively few of them could actually result in synonymous mutations, thereby reducing the actual amount of information afforded by cox1 (DeSalle *et al*., 2005).

The DNA project was proposed as a standard global system for fast and accurate identification of organisms exploitable from a wider group of users, without any expertise, than is possible at present. The main ambitions of DNA barcoding are: i) to assembly a database of reference sequences which can be used as a tool to assign unknown specimens to species (Hebert *et al*., 2004a), and ii) to facilitate the discovery of new species, particularly in cryptic, microscopic and other understudied taxonomic groups because of their complex or inaccessible morphology. Its utility is evident for associating the sexes in dimorphic species (Sheffield *et al*., 2009) or the larval and adult forms (Kohler, 2007) and for the identification of fragmentary remains (Wong and Hanner, 2008). Current studies suggest that in several taxa species can be delineated by a particular sequence or by a tight cluster of very similar sequences (Hebert *et al*., 2004b; DeSalle *et al*., 2005). It was also advocated that the information contained in the cox1 sequence could have some phylogenetic value and it could contribute to draw the Tree of Life (Ward *et al*., 2005), but this is still one of the more controversial issues concerning the technique and many scientists agree that any sequence does not contain enough information to reliably infer phylogenetic relationships among organisms (Hajibabaei *et al*., 2006b).

Several projects have demonstrated the effectiveness of this approach, based on cox1 gene, in many groups of animals, such as birds (Hebert *et al*., 2004a; Kerr *et al*., 2007), fish (Ward *et al*., 2005), gastropods (Remigio and Hebert, 2003), crustacea (Costa *et al*., 2007), cowries (Meyer and Paulay, 2005), spiders (Barrett and Hebert, 2005; (Greenstone *et al*., 2005), ants (Smith, 2005), springtails (Hogg and Hebert, 2004), mayflies (Ball *et al*., 2005) and several arrays of Lepidoptera (Hebert *et al*., 2003a, 2004b; Janzen *et al*., 2005; Hajibabaei *et al*., 2006a). In addition many campaigns have been launched in order to construct libraries of cox1 sequences of pest insects, disease vectors and other economically important groups (**Table 1**) (Miller, 2007). Finally other studies are underway with the object to extend DNA barcoding to other taxonomic groups, such as plants (Kress *et al*., 2005), fungi (Seifert *et al*., 2007; Geiser *et al*., 2007), macroalgae (Saunders *et al*., 2005) and protests (Scicluna *et al*., 2006).

Table 1. Major barcoding project launched by the principal organizations involved in the barcoding of the Earth's life

Campaign	Goal	Website
FISH-BOL (Fish Barcode of Life Initiative)	cox1library for 30,000 species of marine, freshwater fish of the world	http://www.fishbol.org
ABBI (All Birds Barcoding Initiative)	cox1 barcode data for 10,000 known species of world birds	http://www.barcodingbirds.org
All-Leps (All Leps Barcoding Initiative)	cox1 barcode library for 160,000 known Lepidpetra species	http://www.lepbarcoding.org
BIOCODE (Moorea Biocode Project)	inventory of all non-microbial life on the French Polynesian island	http://www.mooreabiocode.org
PolarBol (Canadian Arctic Initiative)	barcoding the northern biota of Canada and other circumpolar countries	http://www.polarbarcoding.org
CMarZ (Census of Marine Zooplankton)	inventory of the marine biota, around 6800 species representing 15 phyla	http://www.cmarz.org
TBI (Tephritid Barcode Initiative)	cox1 barcode database of 2000 species of all tephritid fruit flies	http://www.dnabarcodes.org
MBI (Mosquito Barcoding Initiative)	identifying 26000 known mosquito species (mainly the disease-bearing)	http://www.dnabarcodes.org

DNA barcoding theory

The gold standard for any taxonomic system is its ability to deliver accurate species identifications. At this regard, it is important to verify the capacity of the approach to aid the initial delineation of a species, by means of defining clusters of individuals species-specific. Hebert *et al.* (2004b) proposed that the validation of the DNA barcoding technique should be performed by evaluating genetic distances within and between species and by a clustering method, such as distance-based neighbour-joined (NJ) tree.

The ability of DNA barcoding system to identify an unknown organism should rely on a divergence–threshold, *i.e.* exploiting the barcoding gap between variability intra- and interspecies. The standard divergence threshold value advised to flag a species using the cox1 gene is so far 10 times the mean intraspecific variation (’10-fold rule’). In the first paper published by Hebert *et al.* (2003a) it was reported that cox1 species profile was 100% successful in identifying species within the Lepidoptera, that is one of the most taxonomically differentiated order of animals, even if with low sequence divergence (Janzen *et al.*, 2005). The divergence values between species were ordinarily greater than 3%, with the exception of only four cases, congeneric species genetically distinct but with low divergence values (0,6-2,0%), probably due to their recent origin, and thus it was proposed to use this genetic threshold for recognizing species. The 10-fold rule resulted valuable in several animal taxonomic groups, as North American birds (Hebert *et al.*, 2004a; Hajibabaei *et al.*, 2006a), sardines (Grant and Bowen, 1998), fishes (Ward *et al.*, 2005), moths (Hebert *et al.*, 2003b), springtails (Hogg and Hebert, 2004), crustaceans (Lefebure *et al.*, 2006) and spiders (Paquin and Hedin, 2004), but it resulted poorly resolutive in other taxa as Cnidaria (Shearer *et al.*, 2002), gastropods (Meyer and Paulay, 2005) and butterflies (Wiemers and Fiedler, 2007). The possibility to use a standard cox1 threshold for species diagnosis could be very interesting because could skip the necessity of morphological assayes, but its definition requires to test it also in other geographical regions and taxonomic groups in order to cover all the biodiversity existing for the species under investigation (Hebert *et al.*, 2004a).

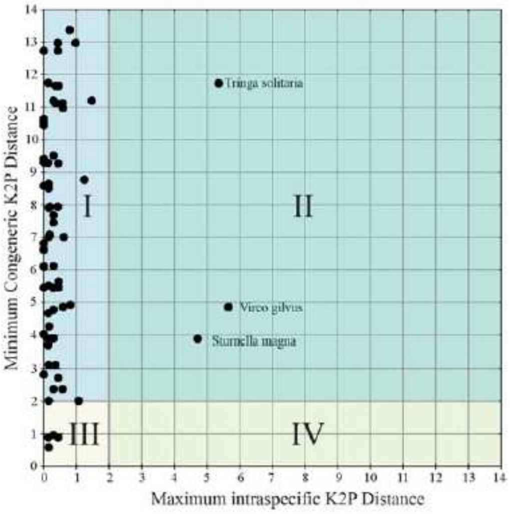


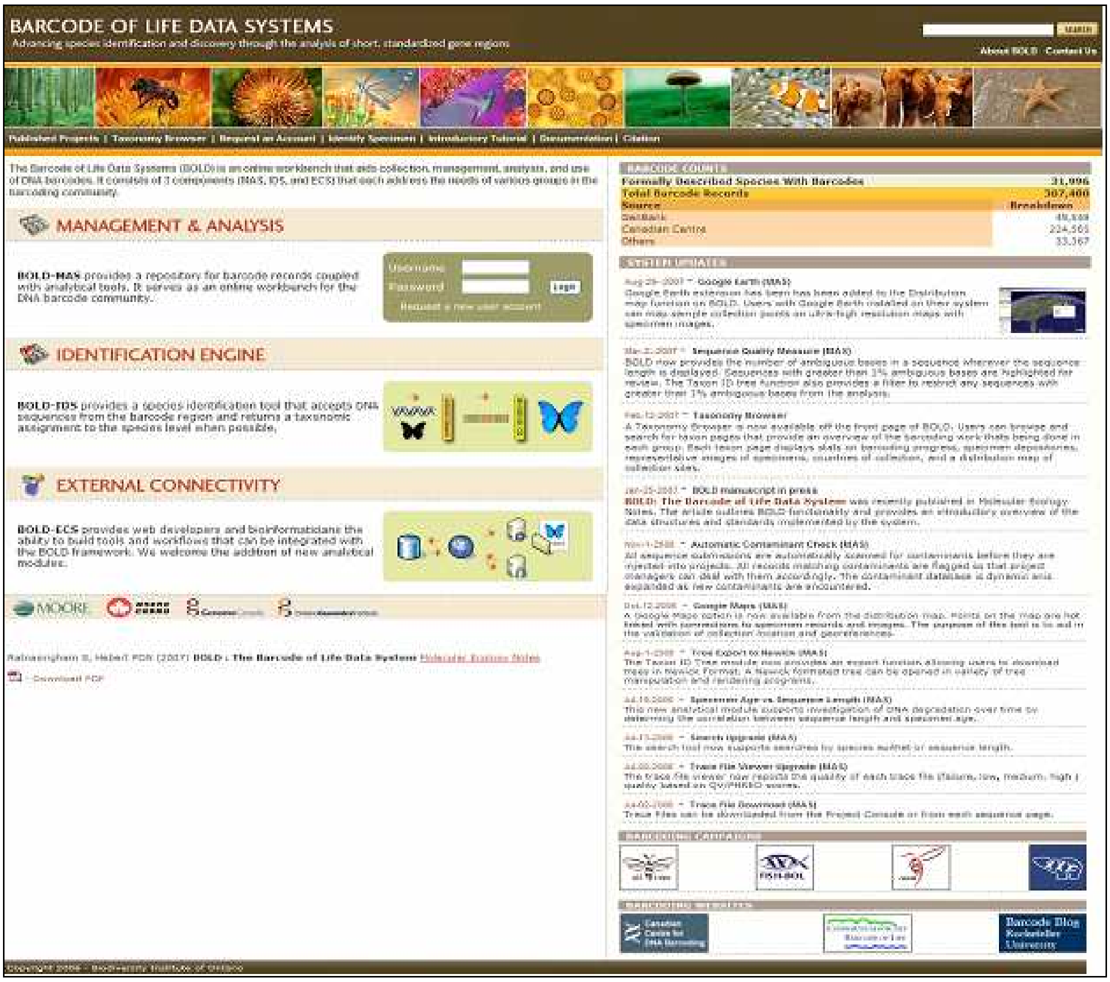
Figure 2. Intraspecific compared to interspecific COI distances (K2P) for individual species in a genetic assay comparing 73 accessions corresponding to as many birds genotypes. For each species, maximum intraspecific variation is compared to minimum interspecific congeneric difference. Only for illustration purposes, an hypothetical cutoff of 2.0% between intra- and interspecific divergence values was chosen. This divides the graph into four quadrants that represent different categories of species: (I) Intraspecific distance < 2% and interspecific distance > 2%: concordant with current taxonomy; (II) Intraspecific distance and interspecific distance > 2%: probable composite species (*i.e.*, candidate for taxonomic split); (III) Intraspecific distance and interspecific distance < 2%: recent divergence, hybridization or possible synonymy; (IV) Intraspecific distance > 2%; interspecific distance < 2%: probable taxonomic misidentification of specimen (modified from Hebert *et al.*, 2004a).

The problem of using the barcoding gap is that it lacks strong biological support and can generate errors, in particular false positive, if populations within one species show high rates of intraspecific divergences, *e.g.* in allopatric populations with interrupted gene flow, and false negatives, when no sequence variation in the barcoding region is found between different species reproductively isolated (species definition in agreement with the Mayr biological species concept) (Mayr, 1963). In these cases the issue becomes distinguishing between populations within the same species and different species and that raises the open question regarding the definition of the species concept. Meyer and Paulay (2005) demonstrated that the barcoding gap existence could be heavly dependent of the sampling of the species. The individuals chosen to represent each taxon in the reference database should cover the major part of the existing diversity otherwise an incomplete sampling could lead to a “barcode gap” that could not correspond to the reality. DNA barcode exclusively promises robust specimen assignment in clades for which the taxonomy is well understood and the representative specimens are widely sampled (DeSalle *et al*., 2005), whereas identification difficulties arise when the unknown specimens come from an under-described taxa (Rubinoff *et al*., 2006a). Therefore it should be proper carrying out an extensive sampling, with specimens from multiple allopatric populations for each species, to assess within species-variability and, mainly, considering species boundaries as a revisable concept (Frezal and Leblois, 2008).

Along with this rule, a second criterium useful to estimate the validity of the assay is the construction of a distance tree (Neighbour-Joining) to give a graphic representation of the genetic distances. The NJ tree does not depend on the barcoding gap, but on the coalescence principle of conspecific populations, *e.g.* individuals belonging to the same species tend to cluster together, but sapearately from different species, and the bootstrapping values give an estimate of the quality of the branching. Anyway, also the NJ tree profile can fail because of incomplete sampling, presence of not reciprocally monophyletic species and when it is applied with closely related species or at intraspecific level, situations that show low divergence values.

Data management on BOLD

Since the advent of DNA barcoding, the construction of a new sequence repository, constituited only by validated nucleotide sequences, is essential for the correct application of this genomic approach. A comprehensive DNA sequence library is essential for correct identification to species, genus, family or even order level (Ekrem *et al*., 2007). Up to now the most common databases freely accessible used as reference systems were the GenBank, EMBL and DDBJ that constitute the International Nucleotide Sequence Database (INSD). The necessity to develop a new reference data set specifically for taxonomic identification was dictated by the fact that these databases, even if they collect sequences of thousands of species, they are not suitable for taxonomic purposes. They are constituited by entries that void of any established taxonomic standards during submission phase, they are often not carefully edited and can suffer from species and population misidentification, missing information and inconsistent terminology (Ross *et al*., 2003). For example, Forster (2003) found that half of all published studies of human mtDNA sequences contain mistakes, not to mention Numts. When GenBank is interrogated by means of BLAST (Basic Local Alignment Search Tool) algorithm, the BIT score (percent identity and E-value) associated with each sequence hit is not a rigorous measure of evolutionary distance or genetic similarity and depends on the size of the database being searched (Karlin and Altschul, 1990). Since these problems could lead the scientists to wrong conclusions in population and evolutionary studies, it is important to develop new affidable sequence databases. In an attempt to catalogue all life forms in DNA terms, the Consortium for the Barcoding of Life (CBOL) was established with the aim of sequencing cox1 gene in all biological species, in a large-scale initiative named the Barcode of Life Initiative (www.barcoding.si.edu) (Savolainen *et al*., 2005; Ratnasingham and Hebert, 2007). Subsequently, the Barcode of Life Data System (BOLD, available on http://www.barcodinglife.com) was born to answer to this necessity and provides support for a large-scale barcode project. BOLD at the beginning was a repository uniquely for cox1 sequences, but currently it is expanding to include also the ITS regions, the official sequences for fungi barcoding, and the combination *matK*/*trnH-psbA* as standard markers for land plants barcoding. In details, BOLD is a collaborative online workbench that includes three different components: the Data Management and Analysis System (BOLD-MAS), the species Identification Engine (BOLD-ID) and the External Connectivity (BOLD-EC).



BARCODE OF LIFE DATA SYSTEMS

Advancing species identification and discovery through the analysis of short, standardized gene regions

SEARCH

About BOLD | Contact Us

Published Projects | Taxonomy Browser | Request an Account | Identify Specimen | Introductory Tutorial | Documentation | Citation

The Barcode of Life Data Systems (BOLD) is an online workbench that aids collection, management, analysis, and use of DNA barcodes. It consists of 3 components (MAS, IDS, and ECS) that each address the needs of various groups in the barcoding community.

MANAGEMENT & ANALYSIS

BOLD-MAS provides a repository for barcode records coupled with analytical tools. It serves as an online workbench for the DNA barcode community.

Username

Password

Login

Request a new user account

IDENTIFICATION ENGINE

BOLD-IDS provides a species identification tool that accepts DNA sequences from the barcode region and returns a taxonomic assignment to the species level when possible.

EXTERNAL CONNECTIVITY

BOLD-ECS provides web developers and bioinformaticians the ability to build tools and workflows that can be integrated with the BOLD framework. We welcome the addition of new analytical modules.

Ratnasingham S, Hebert PDN (2007) **BOLD : The Barcode of Life Data System** Molecular Ecology Notes

Download PDF

BARCODE COUNTS

Formally Described Species With Barcodes	**31,996**
Total Barcode Records	**307,400**
Source	**Breakdown**
GenBank	48,[illegible]
Canadian Centre	224,505
Others	33,367

SYSTEM UPDATES

Aug-29-2007 ~ **Google Earth (MAS)**
Google Earth extension has been has been added to the Distribution map function on BOLD. Users with Google Earth installed on their system can map sample collection points on ultra-high resolution maps with specimen images.

Mar-2-2007 ~ **Sequence Quality Measure (MAS)**
BOLD now provides the number of ambiguous bases in a sequence wherever the sequence length is displayed. Sequences with greater than 1% ambiguous bases are highlighted for review. The Taxon ID tree function also provides a filter to restrict any sequences with greater than 1% ambiguous bases from the analysis.

Feb-12-2007 ~ **Taxonomy Browser**
A Taxonomy Browser is now available off the front page of BOLD. Users can browse and search for taxon pages that provide an overview of the barcoding work thats being done in each group. Each taxon page displays stats on barcoding progress, specimen depositories, representative images of specimens, countries of collection, and a distribution map of collection sites.

Jan-25-2007 ~ **BOLD manuscript in press**
BOLD: The Barcode of Life Data System was recently published in Molecular Ecology Notes. The article outlines BOLD functionality and provides an introductory overview of the data structures and standards implemented by the system.

Nov-1-2006 ~ **Automatic Contaminant Check (MAS)**
All sequence submissions are automatically scanned for contaminants before they are injected into projects. All records matching contaminants are flagged so that project managers can deal with them accordingly. The contaminant database is dynamic and expanded as new contaminants are encountered.

Oct-12-2006 ~ **Google Maps (MAS)**
A Google Maps option is now available from the distribution map. Points on the map are hot linked with connections to specimen records and images. The purpose of this tool is to aid in the validation of collection location and georeferences.

Aug-1-2006 ~ **Tree Export to Newick (MAS)**
The Taxon ID Tree module now provides an export function allowing users to download trees in Newick Format. A Newick formatted tree can be opened in variety of tree manipulation and rendering programs.

Jul-18-2006 ~ **Specimen Age vs. Sequence Length (MAS)**
This new analytical module supports investigation of DNA degradation over time by determining the correlation between sequence length and specimen age.

Jul-13-2006 ~ **Search Upgrade (MAS)**
The search tool now supports searches by species author or sequence length.

Jul-06-2006 ~ **Trace File Viewer Upgrade (MAS)**
The trace file viewer now reports the quality of each trace file (failure, low, medium, high) quality based on QV/PHRED scores.

Jul-02-2006 ~ **Trace File Download (MAS)**
Trace Files can be downloaded from the Project Console or from each sequence page.

BARCODING CAMPAIGNS

BARCODING WEBSITES

Copyright 2006 - Biodiversity Institute of Ontario

MOORE
FISH-BOL
Canadian Centre for DNA Barcoding
Consortium for the Barcode of Life
Barcode Blog Rockefeller University

Figure 2. Home page of *Barcode of Life Data System* (BOLD) web site (Source: www.barcodinglife.org/views/login.php).

Data Management and Analysis System (BOLD-MAS)

DMAS provides a repository for barcode records and it exibhits a simple interface that allows the submission and uploading of new sequences to password-protected projects. It includes information on the place of harvesting and storage for each specimen, photographs and trace files for each sequence record and all these records have to be linked to a voucher specimen. Precisely, BOLD collects currently for each specimens hosted seven data element: (1) species name, (2) voucher data, (3) collection record, (4) identifier of the specimen, (5) cox1 sequence of at least 500 bp, with few ambiguous base-calls, (6) PCR primers used to generate the amplicon and (7) trace files. The core data element in BOLD is a biphasic record consisting of both a ‘‘specimen page’’ and a ‘‘sequence page’’. The former assembles data about source of each specimen including the specimen’s donor and identifier, taxonomy, collection data (including geospatial coordinates and digital images), the repository and catalog number of the voucher specimen. Each specimen page is coupled to a sequence page that records the barcode sequence (FASTA format), PCR primers and trace files, amino acid translation, and ultimately the GenBank accession number. Finally, once the barcode records are submitted in BOLD, then the data are directly uploaded into GenBank because in 2004 GenBank, EMBL and DDBJ databases sealed an accord with CBOL that provides for each barcode standard DNA sequence and relevant supporting data stored in CBOL are automatically moved to GenBank (Savolainen *et al*., 2005). GenBank and the other databases of INSDC expanded the fields for core specimen annotation in their database architecture to more effectively serve barcoding and introduced the keyword ‘BARCODE’’ for those records that meet the appropriate guidelines established by BOLD (Hubert *et al*., 2008).

Identification Engine (BOLD-ID)

The species identification engine is the web tool available for the comparison and matching of sequences from new specimens to the barcode library. The BOLD-ID includes a simple user interface to allow cox1 sequences to be entered into a search field and automatically compared against the existing dataset. BOLD-ID makes use of a combination of BLAST alghorithm and Hidden Markov models based on a global protein alignment for cox1 marker, while for ITS and matK and trnH-psbA it employs only the BLAST algorithm. BOLD provides a probability-based match profile indicating the likely identity of the source species. Additional information is also available, such as links to the species page that provides photographs useful in confirming the identification. Currently, an uploaded version offers the chance to analyse barcode data from other target genes and non-coding regions, more useful in other taxonomic groups, *i.e. matK*/*trnH-psbA* for plants and ITS for fungi.

External Connectivity (BOLD-ECS)

Assembling the sequence information into a comprehensive DNA barcode library requires the development of a data managing system, based on Laboratory Information Management System (LIMS), capable of providing an audit trail for each barcode record. This piece of software, which is under development at the University of Guelph, will be very useful in the handling of data from routine analysis and will extend the capabilities of the current Management and Analysis System (MAS) (Hajibabaei *et al*., 2005).

DNA barcoding technical flowchart

The experimental steps of a DNA barcoding assay are very simple and straightforward:

- *sampling and voucher specimens*: storage in a public repository of all the specimens from which the nucleotide sequences are derived. The sequences have to be retrieved from "holotype" specimens, *i.e.* original individuals stored in public collections (museum, herbarium, zoos, frozen tissue collections and other repositories of biological materials) or newly collected, which are identified by expert taxonomists by means of morphological characters and that provide the basis of the taxonomic system (Dalebout *et al*., 2004). As in most cases it is impossible to obtain the DNA information from these specimens, it is important to select new individuals with certain identities that should be stored as reference specimens. An identification voucher, along with supplemental data such as images, locality information and ecological data, is associated to these specimens that must be conserved as reference for future analyses. For this reason it is important to carry out a long-term storage of the specimens preserving the integrity of the organisms, but for those specimens that have to be completely destroyed to extract DNA, such as for small insects, the only way to conserve some morphological information is to photograph the specimen before destruction (Tautz *et al*., 2003). The need to preserve specimens warrants the transparency of the database because it allows the reviews and re-analyses of a given sample, necessary feature in a discipline, the taxonomy, where the names of organisms are temporary and can be revisionable and the misidentification are common;
- *extraction of genomic DNA*: a tissue sample is taken from the collected individuals and DNA is extracted from them. If the specimen is fresh the DNA isolation should be easy, but in the case of old samples stored in formalin or in the herbarium, the procedure is more complex, requires specific protocol adaptations and sometimes it does not work. Once purified, the genomic DNA must be stored in museum

collections, desiccated or frozen, in way of allowing subsequent amplifications of additional genes (Blaxter, 2004);

- *amplification and sequencing of specific target region*: once extracted, DNA serves as template from which the barcode cox1, ITS, *matK* and *trnH-psbA* markers are amplified by PCR using universal primers (Folmer *et al*., 1994). The development of taxon-specific primers and their combinations are however sometimes necessary to obtain greater intra-generic accuracy (e.g. coral reef, Neigel *et al*., 2007), such as the primers cocktails required for fish species (Ward *et al*., 2005; Ivanova *et al*., 2007) or the primer sets needed to distinguish between primate genera (Lorenz *et al*., 2005). The obtained amplicons are then sequenced bidirectionaly and then manually checked and edited in order to validate sequence quality and detect eventual polymorphic sites, result of co-amplification of nuclear pseudogenes (Bensasson *et al*., 2001);
- *construction of reference database*: sequence information from the voucher samples are deposited in the database accessible from BOLD to allow unambiguous identification of specimens of unknown origin. Only when the barcoding data are validated by the neighbour-joining method and by evaluating genetic distances within and between species, the type specimen and the associated sequence provide a reference record;
- *interrogation of barcode database*: the identification step consists in the submission of the cox1 sequence obtained from an unidentified sample, the ‘query’ sequence, to the BOLD database through the BOLD-IDS in order to find the perfect match. BOLD-IDS accepts the DNA sequence from the barcode region and returns a taxonomic assignment to the species level, when possible, through the same sequence similarity search and the clustering method used for the validation step. In the case of cox1 marker, there are four different sequences subset in function of the validation of the sequences contained: only a subset of BOLD repository is a validated dataset because it includes sequence records with a sequence length of 500 bp, with a species level identification and referred to many species represented by one or two individuals showing less than 2% sequence divergence. BOLD engine delivers a species identification providing the 20 closest matches, with a divergence

value less than 1%, with the reference standard held within the database (Ratnasingham and Hebert, 2007). BOLD also generates a taxonomic identification summary and a NJ tree of species barcode sequences. Then the system can map specimen collection localities on a distribution map with high resolution and allows morphological comparison of voucher specimens when appropriate digital images are loaded. If the match is not obtained, the query sequence is assigned to a genus with a similarity divergence lower than 3%. Above all, if the unknown specimen does not match to any existing records in the barcode library, it should be flagged as a 'problem taxa' that deserves supplemental taxonomic analyses, rather than being discounted as a taxonomic error, suggesting that or the sampling was not complete or we may be in presence of a new species, such as a cryptic species, or a new haplotype or geographical variant.

Overall, there are many technical advantages related to DNA barcoding. The technique is not influenced by subjective assessments, it is reproducible at any time and by any researcher and therefore it represents an universal applicable method, that can be linked to any kind of biological or biodiversity information. The experimental procedure of extracting DNA and amplifying specific markers is technically easy and usually does not require the destruction of the sample, that sometimes is valuable and therefore it should be safeguarded. The technique is fruitfull and effective in terms of cost and time, and enables automated species identification, particularly useful in large sampling campaigns, as of Craig Venter's Global Ocean sampling team (Rusch *et al.*, 2007). The storage of DNA does not need particular attention because the molecule is very stable and any sample can be split into multiple subsamples, which can be sent to many museums as backups. Regarding DNA sequencing step, if the technique was considered expensive in the past, now the technological progress warrants a cheaper and faster way of sequencing (Tauz *et al.*, 2003).

The mitochondrial genome

The mitochondrial genome (mtDNA) is a small circular genome and its size, structure and gene content vary considerably among organisms. It possesses several remarkable characteristics that make it a very useful molecular marker in evolutionary studies. First of all, mtDNA exhibits a non-Mendelian mode of inheritance that determines biased segregation of cytoplasmic genes (Birky, 2001). Generally the inheritance of this genome is maternal, with some exceptions of paternal or biparental mtDNA inheritance (review by Korpelainen, 2004). Second, since it is non-recombining, the entire genome represents a single linkage unit and that, along with its haploid nature, promotes the loss or the fixation of mtDNA haplotypes, reducing the diversity and thus sequence ambiguities from heterozygous genotypes within species(Avise, 1989). Third, although the important cellular functions held by the organellar genes, mtDNA generally evolves faster, about 5-10%, than single-copy nuclear genes at a rate of approximately 2% per million years in bilaterian metazoans (Ballard and Kreitman, 1995), allowing the discrimination of even closely related species (Juan *et al*., 1996; Brown *et al*., 1979). The reason of this high evolutionary rate is due to frequent occurrence of mutations caused by high amount of reactive oxygen radicals (ROS) produced during the respiratory chain, that can chemically alter DNA, coupled with the absence of a compact protein-DNA complex that leaves mtDNA more accessible and, at the same time, more vulnerable to damages caused by ROS (Salgado *et al*., 2008). The evolutionary rate of the mtDNA is not homogeneus, but it displays variation in different regions that are subject to strong functional constraints. Generally, the slowest evolving mitochondrial genes are those encoding the two ribosomal RNAs (rRNAs) and the 22 transfer RNAs (tRNAs), D-loop central domain and nonsynonymous sites in protein-coding genes, while the most rapidly evolving regions are the two peripheral D-loop region domains, called CSB and ETAS, the intergenic sequences and synonimous sites (Pesole *et al*., 1999). Among functional regions in mammals, the highest degree of conservation, with an average pairwise similarity over 75%, was found in the genes coding for the three subunits of the cytochrome c oxidase, the cytochrome b, the 16S rRNA and some tRNAs (Saccone *et al*., 1999). Furthermore, since many mitochondrial genes are highly conserved at the amino acid level, usually the mutations are narrowed at third codon position, with predominance of transitions than transversions, since it is less constrained by selection because of its four-fold degeneracy (Hebert *et al*., 2003a). Therefore, the mutations usually are silent and selective neutral (Brown *et al*., 1979), providing many potentially phylogenetically informative characters. Finally, it was reported that some nucleotides are more susceptible to mutations than other, the frequency of mutation for all four nucleotides is not equal and the direction of mutation is not random. For example, the nucleotide composition at third position site is strongly biased, for instance A-T in arthropods and G-C in chordates, reducing information content (Iannelli *et al.*, 2007; Hebert *et al.*, 2003a). In addition, the mtDNA is present in multiple copies in the cell and that should improve the possibility of amplifying template molecules also in presence of highly degraded DNA, as in processed food, compared to the nuclear encoding single-copy genes. Furthermore, its lack of introns and the low frequency of DNA deletions and insertions simplify sequences alignments of different species because sequence gaps are rare (Saccone *et al.*, 1999). Since its reduced size, it was the first eukariotyc genome to be completely sequenced in human (Anderson *et al.*, 1981) and many other mitochondrial genomes from different organisms were recently sequenced and they are now accessible on the MitBASE Web site, http://www.ebi.ac.uk/htbin/Mitbase/mitbase.pl), an integrated and comprehensive database of mitochondrial DNA. The knowledge of several complete mitochondrial DNA allows not only the design of robust and universal primers enable to routinely recover specific segments of the mitochondrial chromosome in a wide range of eukaryotes (Folmer *et al.*, 1994; Simmons and Weller, 2001), but also specific primers able to amplify in determined species without requiring subsequent sequencing step or other PCR-based techniques (Montiel-Sosa *et al.*, 2000; Lin and Hwang, 2008).

The ideal barcode marker and the cox1 gene

The main difficulty of DNA barcoding was to find the ideal marker that discriminates any species in a given kingdom. In the past, many regions have been tested for species-level biosystematics, but there was not a consensus marker and the choice of the sequence depended on the group under investigation. Selection of an appropriate target market is a critical decision and five criteria must be satisfied to evaluate if the genetic loci are appropriate for DNA barcoding of animals and plants. First of all, an ideal region should be orthologous among taxa, better if amplifyable using universal primers, in order to standardize the procedure across a wide range of taxa (Olmestead and Palmer, 1994; Kress *et al.*, 2005; Taberlet *et al.*, 1996). The use of universal primers is particulary important when environmental DNA, containing a mixture of many species to be identified, is analyzed. Then, it should possess significant species-level genetic variability to allow identification of species, but high conservation rate within species in order to generate the barcode gap (Barrett and Hebert, 2005; Hebert *et al*., 2003a). It should be of appropriate sequence length, about 700 bp, to provide enough phylogenetically informative sites to easily assign species to its taxonomic group (genus, family, etc.), but at the same time to allow PCR amplification and DNA sequencing in one reaction. Shorter regions, even if highly variable, may not provide a sufficient number of variable characters to generate a resolved NJ tree (Shaw *et al*., 2005). Furthermore, the DNA barcode target should be technically simple to sequence, *i.e.* without any long repeat regions, easy to analyze, *i.e.* length-conserved (with more SNPs than In/Dels) to avoid alignments ambiguity and recoverable from degraded DNA samples, such as alcohol-preserved tissues stored in museums, forensic materials or processed food (Telechea *et al*., 2005; Taberlet *et al*., 2007). Finally, identifying hybrids would be desiderable and, in the case of long established natural hybrid species, this should not be problematic (Cowan *et al*., 2006). In cases of recent hybridization or ongoing introgression it is not possible to make a reliable identification using organellar DNA regions, but it requires the use of nuclear regions able to recover different allelic variants from a sample (Chase *et al*., 2005). Nevertheless, in the cases of identification of breeds, geographic origins or individual assignments, markers should possess different features and show consistent intra-specific variability. Therefore, in some cases, a strong haplotypic structure within a species can allow allocation of an individual to a particular geographic population.

Because of its peculiar features, the mitochondrial DNA (mtDNA) has been elected as the molecule of choice for barcoding studies and John Avise (Avise *et al*., 1987) was the first to propose the employment of the mtDNA to recovery the evolutionary history within species. After that a huge mole of phylogenetic studies were published and now the mtDNA represents the first target genome suggested as ideal source of DNA barcoding markers in metazoas. In the past some mitochondrial genes encoding ribosomal DNA (12S, 16S) have been widely exploited, but the presence of frequent insertions and deletions (indels) complicated the sequence alignments (Doyle and Gaut, 2000). Then, the interest was focused on the protein-coding regions that offer the advantage of being arranged into codons. Among the 13 protein-coding genes, cox1 gene was proposed as suitable sequence for DNA barcoding (Hebert *et al*., 2003a). The entire gene is long 1,600 bp, but only the portion of 648 bp located near to the 5' end of the gene proved to be very powerful in discriminating species and phylogeographic groups within species. The cox1 gene was selected as the core of the global bioidentification system for animals because it shares all the criteria above mentioned (Chase *et al.*, 2005). First of all, the universal primer pairs for cox1 allow the routine recovery of the marker from representatives of most animal phyla (Folmer *et al.*, 1994; Zhang and Hewitt, 1997) with no evidence of recovery of the nuclear pseudogenes (Hebert *et al*., 2003a). Second, the alignment of this region is enough easy since the occurrence of insertions and deletions is rare and the evolution source is essentially based on the nucleotide substituions (Hebert *et al*., 2003a). Third, cox1 appears to possess a greater range of phylogenetic signals than any other mitochondrial gene, but its evolutionary rate is not constant among all the metazoan. In common with other protein-coding genes, its third position nucleotides show a high incidence of base substitution, about three times greater than that of 12S or 16S-rDNA regions (De Giorgi *et al*., 1991; Ruttkay *et al*., 1990; Knowlton and Weigt, 1998), but exhibits low nucleotide variation level, for example within Cephalopods (Lindgren *et al*., 2005; Strugnell and Lindgren, 2007) or in plant kingdom (Fazekas *et al*., 2009). Anyway, cox1 evolution showed not only high rates of species discrimination (>95%) in various vertebrate and invertebrate groups (Hebert *et* al., 2003b, 2004b), but also proved to be enough variable to distinguish different phylogeographic groups within a single species (Lynch and Jarrell, 1993; Cox and Hebert, 2001; Wares and Cunningham, 2001). The efficiency of cox1-based barcoding has been documented also for a few groups of fungi (*e.g. Penicillium* spp., Seifert *et al*., 2007; *Aspergillus* spp., Geiser *et al*., 2007), macroalgae (Rodophyta, Saunders, 2005) and protests (Paramecium and Tetrahymenas, Barth *et al*., 2006). Additionally, smaller fragments (*i.e.* 100 bp) of the standard cox1 barcode - 'mini barcodes' - have been shown to be effective for species identifications in specimens whose DNA is degraded or in other situations where obtaining a full-length barcode is not feasible (Hajibabaei *et al*., 2006b).

Land plants: the two-tired approach

As said previously, the rate of genomic evolution in mitochondrion, as well as in nucleo, is not equal for all living species, but can even differ at the ordinal level. Most mitochondrial DNA regions in plants exhibit lower nucleotide substitution rates than plastid or nuclear genomes, unsuitable to distinguish between taxa (Palmer and Herbon, 1988), with some exceptions in specific taxa (Cho *et al*., 2004), and thus land plants, especially angiosperms, seem to be problematic for DNA barcoding. Wolfe *et al*. (1987) showed that rates of synonimous substitution in angiosperm mitochondrial genes are anomalously low, a few-fold lower than in chloroplast genes, from 10 to 20-fold lower than in nuclear genes of both angiosperms and mammals, and from 5- to 100-fold lower than in mammalian mt genes (Cho *et al*., 2004). Furthermore, the mitochondrial genome in plants undergoes rapidly and significant rearrangement (Palmer, 1992) and genome-wide horizontal gene transfer, both at intra and interspecific levels (Wong and Henner, 2003) thereby precluding the existence of universal intergenic spacers useful as identifiers at the species level. As a consequence, all these features exclude species identification based on any mitochondrial regions that resulted inappropriate for discriminating plant species.

Thus for the study of plant barcoding the two primary sources of informations storically are the chloroplast genome (Palmer Herbon, 1988; Clegg and Zurawski, 1992) and nuclear ribosomal DNA repeat region (Baldwin, 1992; Hamby and Zimmer, 1992). The CBOL Plant Working Group (PWG) agrees that the most suitable genome is the chloroplast one (cpDNA) because it may represent the plant counterpart of the animal mtDNA. Chloroplast DNA sequences, both coding and non-coding regions, have been extensively used to infer plant phylogenies at different taxonomic levels (Table 1). The choice of the sequences to adopt depends on the taxonomic group investigated as well as on the phylogenetic level studied in order to select the regions with the more appropriate substitution rate (Shaw *et al*., 2005). Plant studies report a more modest ability of DNA barcoding to discriminate among closely related species compared to animals (Kress and Erickson, 2007). Untill now, the ideal DNA marker for plants that meets all barcode standards was not found yet: “The hope of finding a single, short sequence of DNA from one gene that will reveal the identities of all plants or animals could be akin to a search of Holy Grail” (Rubinoff *et al*., 2006). All the markers, plastid and nuclear, tested singularly to evaluate their ability to discriminate species pairs in plants, exhibited an efficacy lower than the mitochondrial cox1 marker for animals, and less that 85% of the genera examined could be propely identified (Kress and Erickson). This lack of resolution, encountered when only one single DNA region was used for barcoding purposes, has led to develop the idea of an integrated approach based on employing several loci at the same time (Chase *et al*., 2005; Cowan *et al*., 2006; Sass *et al.*, 2007; Fazekas *et al.*, 2008) that was also welcomed by critics of barcoding. Some combinations of DNA regions for a multilocus DNA barcode system have been proposed during the Second International Barcode of Life Conference held in September 2007 in Taipei, Taiwan, but at present no marker combination demonstrated to work universally in all taxonomic groups. In fact, it was demonstrated that not all regions are complementary and universal for all the genera, but certain species are resolved only if differing sets of specific regions are included in the analysis (Fazekas *et al*., 2008). Combining the most variable plastid regions provided only marginally different success rate (Kress *et al*., 2005; Chase *et al*., 2005; Kress and Erickson, 2007; Fazekas *et al*., 2008), suggesting that species discrimination is not always limited by inadequate variability at the chosen locus/loci and raising the issue regarding the discreteness of plant species and the nature of species boundaries on the basis markers from a single genetic linkage group. In fact barcode species resolution, based on monophyly criterion, reaches for equal level of PICs (parsimony-informative characters) values like 90-98% for the animal data sets using only cox1 sequence, while in plants the resolution achieves 46% if using a single plastid gene and a plateu of 71% when several plastid markers are combined (Fazekas *et al*., 2009). Furthermore, when compared the distribution of intraspecific and interspecific genetic distances across animal and plant genera derived from many published projects, it is resulted that the values of interspecific distance are much greater in animals than in plants. In addition the degree of overlap between inter- and intraspecific distance is usually wider in plants than in animals and thus it reduces the ability of the used regions to discriminate species (Fazekas *et al*., 2009).

The most appreciated multi-locus proposal was the “two-tiered approach”, suggested by Newmaster *et al*. (2006), that consists in employing a conservative coding region common across the land plants at a first tier, the “anchor”, that provides resolution at superior ranks (*e.g*., family and genus) and for distantly related plants, and a more variable (coding or noncoding) region as “identifier” to provide resolution for closely related taxa or at lower taxonomic level, below the family level (Gielly and Taberlet, 1994; Olmestead and Palmer, 1994) such as the combination *rbcL* gene - *trnH-psbA* intergenic spacer (Kress and Erickson, 2007). Anyway, the scientific community elected, as standard combination, the plastid gene *matK*, a maturase-encoding gene, with a more rapid substitution rate than *rbcL* useful at the genus and family levels, alone or in combination with *trnH-psbA* (Newmaster *et al*., 2007; Chase *et al*., 2007; Lahaye *et al*., 2008). In addition, Taberlet *et al*. (2007) focused on the feasibility of barcoding plants from highly degraded DNA that is of interest for ancient DNA studies (*e.g*. permafrost samples) and other applied fields (e.g. processed food, customs and medicinal plants). They promoted the chloroplast trnL (UAA) intron or a shorter fragment of this region (the P6 loop, 10-143 bp), which, despite the relative low resolution, can be amplified with highly conserved primers.

The potentials of plastid markers have being tested and several projects have been launched. For example, the "Darwin Initiative for the Survival of Species" funded a project at the Royal Botanical Garden, in Kew, on the barcoding of the orchids of Costa Rica and a project, in collaboration with the University of Johannensburg (South Africa), which aims to barcode the flora of the Kruger National Park in South Africa. Other projects underway are at the Smithsonian Insitute to generate DNA barcodes for all economically plants, especially medicinals and poisonous plants (Cowan *et al*., 2006).

DNA barcoding toward taxonomy, population genetics and phylogeny

The proposal of using DNA barcoding as new identification tool turned on a heated debate about the potential uses of this technique. The advocates of DNA barcoding claim that it will revitalize biological collections and speed up species identification and inventories (Savolainen *et al*. 2005; Gregory, 2005; Schindel and Miller 2005), whereas its opponents argue that it will destroy traditional systematics and turn it into a service industry (Ebach and Holdrege 2005). Mainly the researchers that work with tropical environment are among the most active advocates of DNA barcoding since that habitat is the heart of biodiversity and offers a variability of species, often unknown and thus without any recognized expert taxonomist able to recognize it (Janzen, 2004). DNA barcoding is interested because it involves and complements different scientific fields, in particular taxonomy, molecular phylogenetics and population genetics (**Figure 2**).

The taxonomy's task is to classify all the biodiversity on the Earth employing the Linneum binomial naming system. In the past century specific rules have been introduced by international commissions of scientists in order to standardize this procedure and avoiding cases of synonymies, *i.e.* the same species has two names, and homonymies, *i.e.* the same name is related to different species (http://www.iczn.org). The DNA barcode project does not have the ambition to build the Tree of Life, but rather to produce a simple diagnostic tool based on strong taxonomic knowledge (Schindel and Miller, 2005). DNA barcoding can be just considered an additional instrument complementary to taxonomic surveys for routine species identification and detection of cryptic species in a more standardized way. In this context, DNA barcoding relies on the species concept used previously by taxonomy to define the species. Since DNA barcoding approach is blurred by species-level paraphyly and polyphyly that were proved to be really common, around 21% of cases in animal species, the use of mtDNA barcode may lead to ambiguous or erroneous identification in as many cases (Funk and Omland, 2003)(Funk and Omland, 2003). In addition, in presence of recently diverged species that share alleles for some time after the initial split because of ongoing gene flow, DNA barcoding does not warrant an unequivocal identification. For example, in the case of very recent radiation of cichlid fishes in Lake Victoria, the morphological distinctiveness has built up much faster than has the molecular one determining morphology-based taxonomy more powerful (Meyer *et al.*, 1990). Regarding a second potential use, species discovery, this is not a valid exploitation of the technique because it requires a species concept and a corroboration system and no single source of data can by itself be considered enough to define a species (DeSalle, 2005). As no taxonomist would describe a new species based solely on a single morphological character, so also the barcoding community does not claim that one single gene is enough to characterize all the metazoans. Furthrmore, it would be necessary defining valuable markers and a cut-off value of intraspecific variability in order to discriminate organisms and delimitate species entity that was undergone to interruption of gene flow for a period of time lasting enough to allow the formation of a new species (Savolainen *et al.*, 2005). In particular situations, when crypticism might occur, DNA sequences, like any other molecular markers, from allozymes to DNA markers, can assist in species discovery, by flagging potential candidates for new species units which then need to be confirmed using an integrated taxonomic approach (Witt *et al.*, 2006; Rubinoff, 2006a,b).

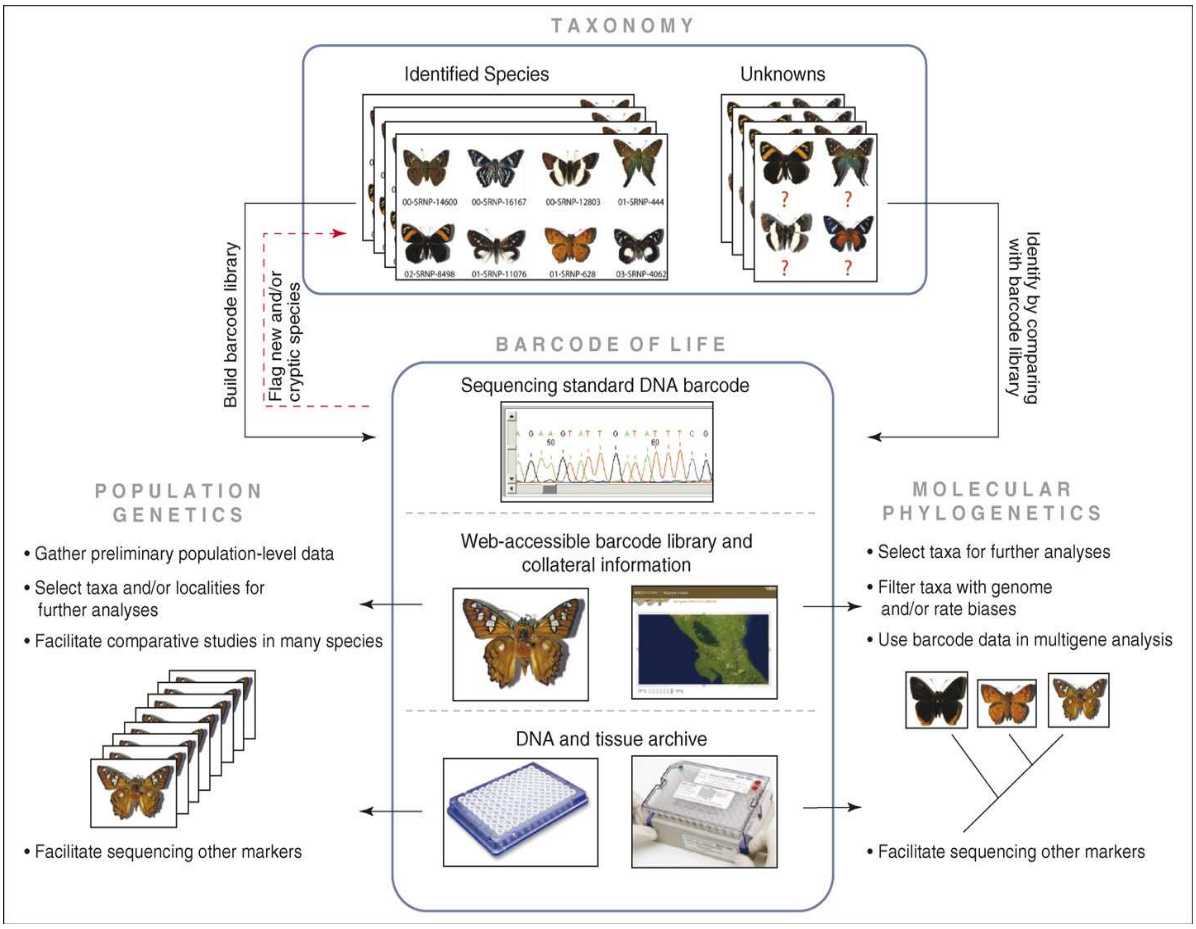

TAXONOMY
Identified Species
Unknowns
00-SRNP-14600
00-SRNP-16167
00-SRNP-12803
01-SRNP-444
02-SRNP-8498
01-SRNP-11076
01-SRNP-628
03-SRNP-4062
?
?
?
?
Build barcode library
Flag new and/or cryptic species
Identify by comparing with barcode library
BARCODE OF LIFE
Sequencing standard DNA barcode
Web-accessible barcode library and collateral information
DNA and tissue archive
POPULATION GENETICS
• Gather preliminary population-level data
• Select taxa and/or localities for further analyses
• Facilitate comparative studies in many species
• Facilitate sequencing other markers
MOLECULAR PHYLOGENETICS
• Select taxa for further analyses
• Filter taxa with genome and/or rate biases
• Use barcode data in multigene analysis
• Facilitate sequencing other markers

Figure 2. Major components of the Barcode of Life projects and their contribution to taxonomy, molecular phylogeny and population genetics. This diagram shows how DNA barcoding libraries can support the conventional taxonomic workflow by high-throughput identification of unknown specimens and by helping to draw attention to new and cryptic species. Barcode sequences and collateral data for each specimen are accessible through a global online data base (e.g. BOLD: http://www.barcodinglife.org). This information can be useful in other contexts, such as phylogenetics (Tree of Life projects) and population-level studies. In addition, archival DNA and tissue specimens collected in barcoding projects provide an excellent resource for other investigations. Butterfly images are taken from the database of Daniel Janzen and Winnie Hallwachs (http://janzen.sas.upenn.edu/) (Hajibabaei *et al.*, 2007).

In poorly studied taxonomic groups, DNA barcoding could be used in the view of “reverse taxonomy”, *i.e.* describing the species first using just the polymorphisms of their mtDNA, rather than analyzing DNA from previously morphologically identified specimens, with the possibility in the future to add morphological information and formal species description (Markmann and Tautz, 2005; Smith, 2005). In this context, it was introduced the concept of MOTU to define taxa, mainly for microbial life where morphological inspection is precluded, without any reference to the correspondence to species concept (Blaxter *et al.*, 2005). Therefore DNA barcoding and metagenomics promise great insights for biodiversity studies of meio- and microfauna, groups frequently underestimated because of their small size (Tringe and Rubin, 2005; Tyson *et al*., 2004).

If the purpose of taxonomy is the identification of organisms, the assignment of the species to higher level taxa is associated with generating phylogenetic hypotheses, which can potentially be inferred directly from DNA sequences. Although the sequences collected within the framework of DNA taxonomy are intended primarily to provide identification, rather than phylogenetic resolution, a DNA taxonomy database will nonetheless constitute an invaluable resource for phylogenetics. In fact, even if the main domain of DNA barcoding is the species identification, it was demonstrated that it can contribute to refining species discovery once that the barcode database is established, flagging candidate exemplar taxa for a comprehensive phylogenetic study (van Velzen *et al*., 2007). Increasing the taxon sampling aids the recovery of the correct phylogeny by reducing branch lengths and homoplasy, both factors that can mislead phylogenies (Zwickl and Hillis, 2002), and Barcode of Life projects can create a perfect taxonomic sampling for conducting phylogenetic studies on different branches of the Tree of Life. It was also advocated that the information contained in the cox1 sequence could have some phylogenetic value because the tree reconstruction above the genus level is often conforming with the classical phylogeny (Ward *et al*., 2005). Actually, the estimation of the species phylogeny through DNA barcoding is not conceptually correct because it derives from the employment of an organellar marker that does not correspond to a gene for the speciation and thus it cannot keep track of the evolutionary history of the taxa (Blaxter *et al*., 2005). Therefore generally the topology of the resultant gene tree is not congruent with the species tree because of several factors (Ekrem, 2007)Ekrem *et al*., 2007). Events, such as interspecific hybridization or repetitive introgression patterns (Bergthorsson *et al.* 2003), polyploidization and horizontal gene transfer (Tautz *et al*., 2003; Dasmahapatra and Mallet, 2006), can create confusion for recovery of taxon affinities. In addition, character convergence and accidental recovery of Nuclear Mitochondrial DNA (NUMT) or Nuclear Plastid DNA (NUPT), nuclear copies of organellar DNA sequences translocated into the nuclear genome of eukaryotic organisms (Zhang and Hewitt, 1996; Williams and Knowlton, 2001), confounds phylogenetic and population genetic analyses since they have different evolutionary patterns and mode of inheritance and they own their particular codon structure, non-synonymous mutations, premature stop codons and insertion-deletions (Strugnell and Lindgren, 2007). Therefore, adopting a multi-locus barcoding system, also called non-cox1 barcode (Bakker, Second International Barcode of Life Conference, Taipei, September 2007), with more than one gene, each representing a distinct linkage group, nucleo and organellar genome, could contribute to “replicate” estimates of the species tree from one or more indipendent gene trees (Moore, 1995).

Finally, it is interesting to evaluate the contribution of DNA barcoding for population genetics. This branch of biology studies genetic variation of populations within a single species to investigate issues, such as migration and geographic drift. The microevolutionary-level assay in the past was investigated by means of allozymes of a particular locus, but their nuclear origin led to concerns regarding allele frequencies and heterozigosities (Avise *et al.*, 1987). Subsequently, the estimation of within-species variability was performed analyzing the mtDNA that provided accessible data for strong genealogical inference and it showed that many species exhibit a deep and geographically structured mtDNA evolutionary history (Tavares and Baker, 2008). Study of the relationship between gene genealogy and population geography constitutes a discipline that can be called intraspecific phylogeography (Avise *et al.*, 1987). The understanding of the evolution of species strongly structured phylogenetically cannot be fully performed without references to the intraspecific phylogeographic structure. DNA barcoding can provide a first signal of the extent and nature of population divergences and can facilitate comparative studies of population diversity in many species. Unfortunately, the genealogy recorded by mtDNA is far from a complete characterization of intraspecific phylogeny, in particular when males and females differ in phylogeographically relevant characteristics. Other difficults arise when intraspecific variation, caused by incomplete sampling or related to a real distinction among specimens (Dasmahapatra and Mallet, 2006), and intragenomic variation, due to heteroplasmy (Terranova *et al.*, 2007), are detected. Therefore, a better approach should be the application of a multi-locus approach because more informative and less sensitive to specific gene genealogies. The availability of high-trhoughput sequencing technology, fine-scale sequence analysis methods, such as SNPs, are contributing to population-level studies (Brumfield *et al.*, 2003).

The character-based approach

Currently, the most common way to use DNA barcode data is based on the phenetic approach, based on genetic distance and clustering method. It has become apparent that this kind of approach has strong limitations, due to the inconstant mtDNA rates of evolution between and within species and between different groups of species resulting in broad overlaps of intra-interspecific distances (Kipling and Rubinoff, 2004). Despite the reported efficiency of the divergence-threshold method in several cases, this approach presents some drawbacks, as said previously.

An alternative to the phenetic approach is the character-based system that focuses on the concept that previously established taxonomic groups can be identified on the basis of a binary signal, presence or absence of a discrete nucleotide substitution, the character state, or combination of characters within a short DNA sequence (Rach *et al*., 2008). Members of a given taxonomic group share sequence polymorphisms, termed "characteristic attributes" (CAs), that are absent from other groups. CAs are diagnostic character states (genes, amino acids, base pairs or even morphological, ecological or behavioural attributes) which are found only in one clade, but not in an alternate group that descends from the same node. CAs are divided in two major groups: i) *pure* CA is shared by all members of the clade and is absent from the other clades, while ii) *private* CA is shared by only some members of a clade, but is absent from the other clades (Rach *et al*., 2008). Both pure and private CA can either be *simple* CA, when confined to a single nucleotide position, or *compound* CAs which are combined states at multiple nucleotide positions (DeDalle *et al*., 2005). These diagnostic characters, in the case of DNA barcoding, are SNPs, an emerging class of molecular markers that include single DNA base mutations and small insertions or deletions that occur at single position in the genome. The challenge of the approach is that the character-based assessment does not convert the sequence polymorphisms in genetic distance, procedure that determines the loss of character-state information. Therofore, for those groups where the genetic distance is small because of scarse number of sequence polymorphisms, such as at the population level, the phenetic approach could be substituted by the character-based system that retains evolutionary information contained in character-state data. Thus a taxa could be distinguished by the presence or absence of a particular CA and, since all classical taxonomic practices are character-based, this makes the DNA characters obtained by DNA barcoding compatible in a diagnostic context with the process of current taxonomic research, allowing the integration of the CAs with the traditional morphological, ecological, behavioural and reproductive traits (DeSalle *et al*., 2005). Furthermore, in contrast with distance-based technique which depends on the degree of "barcoding gap" and thus on the taxon sampling, the character-based system delineate separate groups without reference to the amount of divergence within and among taxa. However, these potential diagnostic entities, called conservation units, CUs, or evolutionary significant units, ESUs (Vogler and DeSalle, 1994), detected by the character system, cannot be considered new species, they require integrated taxonomic to corroborated the species discovery process (Rubinoff, 2006 a,b). The system proved to be a valid tool to discriminate not only different genera and species (Kelly *et al*., 2007), but mainly it has been shown to be applicable at population level (Rach *et al*., 2008). Finally, the application of CAs facilitate the authomatization for the identification of the sequence polymorphisms through the design of a microarray platform or a rapid SNP detection format using PCR technique based on Taq man probe, avoiding the more complex procedure of sequencing and data analyzing.

DNA barcoding potentials in practical fields

Today's society has to resolve many crucial biological issues, among which are i) maintaining biodiversity and thus providing measures of biological diversity, ii) contributing to the conservation and trade surveillance, iii) resolving the Tree of Life, iv) ensuring the bio-security and avoiding pandemics. The achievement of such goals requires accurate taxonomic identification that has traditionally been domain of taxonomists because the classical methods, based on morphology, demanded great skills and time (Frezal and Leblois, 2008). The recent development of faster reliable tools for species identification for both animals and plants, largely based on DNA fingerprinting, is of great support for many aspects of the life, from large-scale biodiversity survayes to forensic science. There are several situations where limited morphological traits are available and, thus, relevant species identification must be molecular-based and DNA barcoding could reveal a powerful resource. DNA barcoding could be of great support to recognize species in all stages of life of an organism, from juvenile to adult forms (Wells and Stevens, 2008; van Velzen *et al.*,

2007) or in presence of small, damaged or incomplete specimens (*e.g.*, stomach extracts) that lack of diagnostic features (Blaxter *et al.*, 2005; Webb *et al.*, 2006). Finally, DNA barcoding is the only tool exploitable for the determination of the taxonomic identity of forensic specimens (Dawnay *et al*., 2007), in food traceability (Wong and Hanner, 2008) or in the protection of the biodiversity against illegal hunting of endangered animals in order to warrant biodiversity conservation and management policies (Palumbi and Cipriano, 1998).

Food traceability

Traceability is defined as "the aptitude to find the history and the usage or localization of an article or activity with the means of a registered identification" (norm ISO 8402). Many aspects of food chain, species origin, geographical region, commercial treatments, food composition and brand name, can be subjected to fraudulent practice and therefore need *a posteriori* verification of the information decleared on the label. The problem of food authentication has emerged recently due to considerable economic impact, health hazards, caused by food containing allergens (Tanabe *et al*., 2007) and food poisoning (Hsieh *et al*., 2002), and ethical and religious issues (Montiel-Sosa *et al*., 2000) associated to the illegal mislabelling trend of food products. In addition, the food concerns caused by the frequent food emergencies (*e.g*., BSE, avian flu, mouth disease, etc.) has reinforced the public awareness regarding the implementation of the traceability and safety of food products sold in the market (Teletchea *et al*., 2005).

In particular, the detection of events of food falsification in seafood products is gaining particular attention because it was demonstrated that mislabeling of these derivatives and the use of vernacular or generic labels for fisheries are known to occur (Marko, 2004). In addition, species identification is necessary in order to prevent the commercialization of species for which a conservation policy exists (Civera, 2003). The extensive and unregulated hunting and trade of whales, though illegal since the 1982, continues and thus pointed out the necessity of developing new systems of monitoring to safeguard protected populations (Palumbi and Cipriano, 1998).

Two important directives regulate the trading exchanges in the European Union (EU) in order to enforce conservation and health-related regulations: i) Reg. CE 853/2004 aims to eliminate toxic products and endangered species from trade and ii) EU Reg.104/2000 establishes that seafood labelling must include clear indications of commercial name, method of production (wild or farmer, organic or intensive) and capture area of the species (Civera, 2003). The DM 14/01/2005 in addition reports the updated list of all the commercial and scientific names for each marine species used for food production.

The task of veterinary inspection consists in the detection of commercial frauds, when there is the substitution of low-quality species for a more valuable one, and sanitary frauds when a hazardous species is sold on the market under a different name. A first analysis is normally realized on the basis of morphological traits, so the skills of the staff are very important. But this kind of species identification for fish products are complicated by many factors, such as: globalization of the seafood industry and consequent introduction in the markets of large numbers of both wild and cultured new species to be examined; ii) sale of processed fish food, as frozen filets, minced meat, fish paste, dried, smoked or canned products, lacking the morphological traits useful for the traditional identification procedure; iii) insufficient trained people employed in species identification (Civera, 2003). The lack of morphological features, lost when the fish is fileted or processed, makes the traditional authenticity tests impossible to carry out. Therefore the species identification demands the development of new analytical methods and the molecular diagnostic techniques have proven to be effective for this aim because they are capable of bypassing the inherent problems of morphology-based identification methods (Wong and Hanner, 2008).

Use of DNA barcoding in crop plants

The adoption of DNA barcoding is not limited to the species level, but there are cases in which it is worth testing the potentials of DNA barcoding also at sub-species level. In the animal kingdom, the application of organellar DNA, in particular cox1 gene, allowed to reconstruct a large number of phylogeopgraphic groups, proving that intraspecific information contained by this marker can be used to improve identification and potentially to identify geographic origins of new species (Teletchea *et al.*, 2005). Instead in the plant kingdom, the application of DNA barcoding to distinguish varieties is complicated by the difficult to find a marker variable so to count enough polymorphisms to distinguish single varieties. The exploitation of the DNA barcoding in crop plants at variety level is relevant in particular cases, such as for potato clones that show different characteristics in relation to their final consumption (Ashkenazi *et al.*, 2001) or the genetically modified organisms

(GMOs) that represent a special case of variety authentication test. Proving the authenticity of crop seeds could be of interest not only for the buyers that seek guaranteed yelds, but also for the plant breeders. In fact plant breeder rights (PBR) on specific plant materials include the exclusive right to breed and to sell a new plant variety in order to guarantee to the breeder the control of the propagation material, the harvest of the variety material and the right to collect royalties on it for a given number of years (Llewelyn and Adcock, 2006).

In Italy, a newly selected variety in order to be registered and commercialized must be distinguishable from all the other varieties, and characterized by uniformity and stability (DUS). DUS testing could be performed on the basis of morphological traits and molecular markers. Traditional systems used nowadays, such as RFLP, SSR or AFLP, are highly discriminating but time-consuming. After the introduction of DNA barcoding, this new technique could represent a valid alternative to traditional ones in order to distinguish one varietal genotype (*i.e.* pure lines, F_1 hybrids, and clones) from another by means of detection of specific SNP markers and/or haplotypes in selected chloroplast regions. The most problematic aspect is that the variety is not a delimited biological entity as the species because a variety is not reproductively isolated and therefore the genetic delimitation is not so marked as at the species level. Although the occurrence of DNA polymorphisms in specific chloroplast regions is less frequent among varieties than species (Newmaster *et al*., 2007), testing the potentials of this technique to distinguish crop varieties could turn out valuable also for the genetic traceability of agri-food products. Therefore DNA barcoding should be further investigated at the sub-species level to ascertain whether it provides essential features to become the new legal standard approach for rapid identification of varieties and authentication of either row materials or their food derivatives.

References

Allander T, Emerson SE, Engle RE, Purcell RH, and Bukh J (2001). A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proc Natl Acad Sci USA* 98(20) :*11609-11614.*

Altschul SF (1990.) Basic local alignment search tool. *J Mol Biol* 215: *403-410.*

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG (1981). Sequence and organization of the human mitochondrial genome. *Nature*. 290(5806) :*457-465.*

Armstrong KF (2005). DNA barcodes for biosecurity: invasive species identification. *Philos T R Soc B-Biol Sci* 360: *1813-1823.*

Arnot DE (1993). Digital codes from hypervariable tandemly repeated DNA sequences in the *Plasmodium falciparum* circumsporozoite gene can genetically barcode isolates. *Mol Biochem Parasitol* 61: *15-24.*

Ashkenazi V, Chani E, Lavi U, Levy D, Hillel J, Veilleux RE (2001). Development of microsatellite markers in potato and their use in phylogenetic and fingerprinting analysis. *Genome* 44: *50–62.*

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, Saunders NC (1987). Intraspecific phylogeography: The mitochondrial bridge between population genetics and systematics. *Annu Rev Ecol Evol S* 18: *489–522.*

Avise J C (1989). Gene trees and organismal histories: a phylogenetic approach to population biology. *Evolution* 43(6) : *1192–1208.*

Ayala FJ (1983). Enzymes as taxonomic characters. In ‘‘Protein Polymorphism: Adaptive and Taxonomic Significance’’ (G. S. Oxford and D. Rollinson, Eds.), Systematics Association Special 24: *3–26*, Academic Press, London.

Baker CS (2000). Predicted decline of protected whales based on molecular genetic monitoring of Japanese and Korean markets. *Philos T R Soc B-Biol Sci* 267: *1191-1199.*

Baker CS (1994). Which whales are hunted – A molecular-genetic approach to monitoring whaling. *Science* 265: *1538-1539.*

Baldwin BG (1992). Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: an example from the Compositae. *Mol Phylogenet Evol* 1: *3–16.*

Ball SL, Hebert PDN, Burian SK, Webb JM (2005). Biological identifications of mayflies (Ephemeroptera) using DNA barcodes. J N Am Benthol. Soc. 24(3): 508-524.

Ballard JWO, and Kreitman M (1995). Is mitochondrial DNA a strictly neutral marker? *Tree* 10: *485–488.*

Balbontin F (2004). Descriptions of larvae of Merluccius australis, Macruronus magellanicus, and observations on a larva of Micromesistius australis from southern Chile (Pisces : Gadiformes). *N Z J Mar Freshwat Res* 38: *609-619.*

Barrett RDH, and Hebert PDN (2005). Identifying spiders through DNA barcodes. *Can J Zool* 83:
481-491.

Barrett RDH (2005). Identifying spiders through DNA barcodes. *Can j zool* 83: *481-491.*

Barth D, Krenek S, Fokin SI, and Berendonk TU (2006). Intraspecific genetic variation in Paramecium revealed by mitochondrial cytochrome c oxidase 1 sequences. *J Eukaryot Microbiol* 53: *20–25.*

Bartlett SE (1992). FINS (forensically informative nucleotide sequencing): A procedure for identifying the animal origin of biological specimens. BioTechniques 13: *518.*

Bensasson D, Zhang DX, Harti DL, and Hewitt GM (2001). Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol Evol* 16(6): *314-321.*

Bergthorsson U, Keith L, Adams KL, Thomason B, Palmer JD (2003). Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature* 424: *197-201.*

Birky CW (2001). The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms and models. *Annu Rev Genet* 35: *125−148.*

Blaxter M (2003). Molecular systematics: counting angels with DNA. *Nature* 421, *122–124.*

Blaxter M (2004). The promise of a DNA taxonomy. *Phil Trans R Soc Lond* B 359: *669-79.*

Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, Abebe E (2005). Defining Operational Taxonomic Units Using DNA Barcode Data. *Phil Trans R Soc Lond* B 360(1462): *1935-1943.*

Brown WM, George M Jr, and Wilson AC (1979). Rapid evolution of animal mitochondrial DNA. *Proc Natl Acad Sci USA* 76(4) :*1967-1971.*

Brumfield RT, Beerli P, Nickerson DA, and Edwards SV (2003).The utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol Evol* 18(5) : *249-256.*

Carr SM (1999). Molecular systematics of gadid fishes: implications for the biogeographic origins of Pacific species. *Can J Zool* 77: *19-26.*

Chapela MJ (2007). Comparison of DNA extraction methods from muscle of canned tuna for species identification. *Food control* 18: *1211-1215*.

Chase MV, Salamin N, Wilkinson M, Dunwell JM, Kesanakurthi RP, Haidar N, Savolainen V (2005). Land plants and DNA barcodes: short-term and long-term goals. *Phil Trans R Soc Lond* B 360: *1889-1895.*

Chase MW, Cowan RS, Hollingsworth PM, van den Berg C, Madriñán S, Petersen G, Seberg O, Jørgsensen T, Cameron KM, Carine M, Pedersen N, Hedderson TAJ, Conrad F, Salazar GA, Richardson JE, Hollingsworth ML, Barraclough TG, Kelly L, Wilkinson M (2007). A proposal for a standardised protocol to barcode all land plants. *Taxon* 56 (2): *295-299.*

Cho Y, Mower JP, Qiu YL, Palmer JD (2004). Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. *PNAS* 101: *17741–17746*.

Civera T (2003). Species identification and safety of fish products. *Vet Res Commun* 27:*481–489.*

Clegg MT, and Zurawski G (1992). Chloroplast DNA and the study of plant phylogeny: present status and future prospects. In: Soltis PS, Soltis DE, Doyle JJ (eds) Molecular systematics of plants. Chapman and Hall, New York, *1-13.*

Costa FO, deWaard JR, Boutillier J, Ratnasingham S, Dooh RT, Hajibabaei M, Hebert PDN (2007). Biological identifications through DNA barcodes: the case of the Crustacea. *Can J Fish Aquat Sci* 64: *272-295.*

Cowan RS, Chase MW, Kress WJ, and Savolainen V (2006). 300,000 species to identify: problems, progress and prospects in DNA barcoding of land plants. *Taxon* 55: *611-616*.

Cox AJ, Hebert PDN (2001). Colonization, extinction, and phylogeographic patterning in a freshwater crustacean. *Mol Ecol* 10 (2): *371-386.*

Dalebout ML, Baker CS, Mead JG, Cockcroft VG, and Yamada TK (2004). A comprehensive and validated molecular taxonomy of beaked whales, family Ziphiidae. *J Hered* 95(6): *459–473.*

Dasmahapatra KK, Mallet J (2006). DNA barcodes: recent successes and future prospects. J *Heredity* 97: *254-255.*

Dawnay N, Ogden R, McEwing R, Carvalho GR, Thorpe RS (2007). Validation of the barcoding gene COI for use in forensic genetic species identification. *Forensic Sci Int* 173: *1-6.*

De Giorgi C, Lanave C, Musci MD, Saccone C (1991). Mitochondrial DNA in the sea urchin Arbacia lixula: evolutionary inferences from nucleotide sequence analysis. *Mol Biol Evol* 8: *515–529*.

De Salle R, Egan MG, Siddall M (2005). The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philos T R Soc B* 360: *1905-1916.*

Doyle JJ, Gaut BS (2000). Evolution of genes and taxa: a primer. *Plant Mol Biol* 42 :*1–23.*

Dunn CP (2003). Keeping taxonomy based in morphology. Tree 18: *270–271.*

Ebach MC, Holdrege C (2005). DNA barcoding is no substitute for taxonomy. *Nature* 434: *697.*

Eddy SR (1998). Profile hidden Markov models. *Bioinformatics* 14: *755-763.*

Ekrem T, Willassen E, Stur E (2007). A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Mol Phylogenet Evol* 43: *530-542.*

Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM, Hajibabaei M, and Barrett SCH (2008). Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS ONE* 3(7): e2802.

Fazekas AJ, Kesanakurti PR, Burgess KS, Percy DM, Graham SW, Barrett SCH, Newmaster SG, Hajibabaei M, and Husband BC (2009). Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? *Mol Ecol Res* 9: *130–139.*

Ferguson JWH (2002). On the use of genetic divergence for identifying species. *Biol J Linn Soc* 75: 509-516.

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

Folmer O, Black M, Hoeth W, Lutz R, and Vrijenhoek R (1994). DNA primers for amplification of mithocondrial cytochrome c oxydase subunit I from diverse metazoan invertebrates. *Mol Mar Biol Biotechnol* 3 (5): *294-299.*

Forster P (2003). To err is human. *Ann Hum Genet* 67*: 2-4.*

Frézal L, Leblois R. (2008). Four years of DNA barcoding: current advances and prospects. *Infect Genet Evol* 8(5): *727-36.*

Funk DJ, Omland KE (2003). Species-Level Paraphyly and Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. *Annu Rev Ecol Evol S 397-423.*

Gielly L, Taberlet P (1994). The Use of Chloroplast DNA to Resolve Plant Phylogenies: Noncoding versus rbcL Sequences. *Mol Biol Evol* 11(5):*769-777.*

Geiser DM, Klich MA, Frisvad JC, Peterson SW, Varga J, and Samson RA (2007). The current status of species recognition and identification in *Aspergillus*. *Stud Mycol* 59(1): *1-10.*

Godfray HCJ (2002). Challenges for taxonomy—the discipline will have to reinvent itself if it is to survive and flourish. *Nature* 417: *17–19.*

Grant WS, Peck R (1998). Shallow population histories in deep evolutionary lineages of marine fishes: insights from sardines and anchovies and lessons for conservation. *J Hered* 89: *415-426.*

Gregory TR (2005). DNA barcoding does not compete with taxonomy. *Nature* 434: *1067.*

Greenstone MH (2005). Barcoding generalist predators by polymerase chain reaction: carabids and spiders. *Mol Ecol* 14: *3247-3266.*

Guggiari M, Bowen BW (2008). The bacterivorous ciliate Cyclidium glaucoma, isolated from a sewage treatment plant: Molecular and cytological descriptions for barcoding. *Eur J Protistol* 44: *168-180.*

Hajibabaei M, deWaard JR, Ivanova NV, Ratnasingham S, Dooh RT, Kirk SL, Mackie PM, Hebert PDN (2005). Critical factors for assembling a high volume of DNA barcodes. *Philos T R Soc B-Biol Sci* 360: *1959-1967.*

Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PDN (2006a). DNA barcodes distinguish species of tropical Lepidoptera. *Proc Natl Acad Sci USA* 103(4): *968-971.*

Hajibabaei M, Singer GAC, Hickey DA (2006b). Benchmarking DNA barcodes: an assessment using available primate sequences. *Genome* 49: *851-854.*

Hajibabaei M, Smith MA, Janzen DH, Rodriguez JJ, Whitfield JB, Hebert PDN (2006c). A minimalist barcode can identify a specimen whose DNA is degraded. *Mol Ecol Notes* 6: *959-964.*

Hamby RK, Zimmer EA (1991). Ribosomal RNA as a phylogenetic tool in plant systematics. In: Soltis P, Soltis D, Doyle J. Chapman & Hall (Eds) *Molecular Systematics of Plants*, New York, *pp. 50–91.*

Hawksworth DL, Kalin-Arroyo MT (1995). Magnitude and distribution of biodiversity. In: Heywood VH and Watson RT (Eds) *Global biodiversity assessment*, *pp. 107-191*. United Nations Environment Programme & Cambridge University Press, Cambridge, UK.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003a). Biological identifications through DNA barcodes. *Proc R Soc Lond* B 270: *313-321.*

Hebert PDN, Ratnesingham S, deWaard JR (2003b). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc Biol Sci* 270: *S96-99.*

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004a). Identification of birds through DNA barcodes. *PLoS Biol* 2: *e312.*

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004b). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *PNAS* 101: *14812-14817.*

Hogg ID, Hebert PDN (2004). Biological identification of springtails (Collembola: Hexapoda) from the Canadian Arctic, using mitochondrial DNA barcodes. *Can J Zool* 82: *749-754.*

Hsieh, Y. H., Shiu, Y. C., Cheng, C. A., Chen, S. K. & Hwang, D. F. (2002). Identification of toxin and fish species in cooked fish liver implicated in food poisoning. *J Food Sci* 67(3)**:** *948-952.*

Iannelli F, Griggio F, Pesole G, Gissi C (2007). The mitochondrial genome of *Phallusia ammillata* and *Phallusia fumigata* (Tunicata, Ascidiacea): high genome plasticity at intra-genus level *BMC Evol Biol* 7:*155*

Ivanova NV, Zemlak TS, Hanner RH, Hebert PDN (2007). Universal primer cocktails for fish DNA barcoding. *Mol Ecol Notes* 7 (4): *544-548.*

Janzen DH (2004). Now is the time. *Philos Trans R Soc Lond B*359 :*731–732.*

Janzen DH, Hajibabaei M, Burns JM, Hallwachs W, Remigio E, Hebert PDN (2005). Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. *Phil Trans R Soc Lond* B 360: *1835–1845.*

Juan C, Ibrahim KM, Oromì P, Hewitt GM (1996). MitochondrialDNA sequence variation and phylogeography of *Pimelia* darkling beetles on the island of Tenerife (Canary Islands). *J Heredity* 77: *589-598.*

Karlin S, Altschul SF (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *PNAS* 87(6): *2264–8.*

Kelly RP, Sarkar IN, Eernisse DJ, DeSalle R (2007). DNA barcoding using chitons (genus *Mopalia*). *Mol Ecol Notes* 7: *177–183.*

Kerr KCR, Stoeckle MY, Dove CJ, Weigt LA, Francis CM, Hebert PDN(2007). Comprehensive DNA barcoding coverage of North American birds. *Mol Ecol Notes* 7: *535-543.*

Kipling WW, Rubinoff D (2004). Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics.* 20*, 47–55.*

Knowlton N (1993). Sibling Species in the Sea. *Annu Rev Ecol Evol S* 24: *189-216.*

Knowlton N, Weigt LA (1998). New dates and new rates for divergence across the Isthmus of Panama. *Phil Trans R Soc Lond* B. 265: *2257-2263.*

Köhler F (2007). From DNA taxonomy to barcoding - how a vague idea evolved into a biosystematic tool. *Zoosyst Evol* 83: *44–51.*

Korpelainen H (2004). The evolutionary processes of mitochondrial and chloroplast genomes differ from those of nuclear genomes. Naturwissenschaften 91: *505–518.*

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005). Use of DNA barcodes to identify flowering plants. *PNAS* 102 : *8369–8374.*

Kress WJ, Erickson DL (2007). A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH- psbA* spacer region. *PloS ONE* 6: *1-10.*

Lahaye R, van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G, Maurin O, Duthoit S, Barraclough TG, Savolainen V (2008). DNA barcoding the floras of biodiversity hotspots. *PNAS* 105: *2923-2928.*

Lefébure T, Douady CJ, Gibert J (2006). Relationship between morphological taxonomy and molecular divergence within Crustacea: proposal of a molecular threshold to help species delimitation. *Mol Phylogenet Evol* 40 : *435–447.*

Llewelyn M, Adcock M (2006). *European Plant Intellectual Property,* Hart Publishing, Oxford and Portland.

Lin W-F, Hwang D-F (2008). A multiplex PCR assay for species identification of raw and cooked bonito. *Food Control* 19 (9): *879-885*.

Lindgren AR, Amezquita E, Katugin ON, Nishiguchi MK (2005). Evolutionary relationships of gonatid squids and implications for reproductive strategies. Mol Phyl Evol 36: 101-111.

Linnaeus C (1756). *Systema Naturae*. Ninth edition. Theodor Haak, Leiden (Lugdunum Batavorum).

Lipscomb D, Platnick N, Wheeler Q(2003). The intellectual content of taxonomy: a comment on DNA in taxonomy. Trends Ecol Evol 18: 65–67.

Lorenz JG, Jackson WE, Beck JC, Hanner R (2005). The problems and promise of DNA barcodes for species diagnosis of primate biomaterials. *Philos Trans R Soc Lond B* 360(1462): *1869-1878.*

Lynch M, Jarrell PE (1993). A method for calibrating molecular clocks and its application to animal mitochondrial DNA. *Genetics* 135: *1197-1208.*

Manwell C, Baker CMA (1963). A sibling species of sea-cucumber discovered by starch-gel electrophoresis. *Comp Biochem Physiol* 10: *39–53.*

Markmann M, Tauz D (2005). Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. *Phil Trans R Soc B* 360 (1462): *1917-1924.*

Marko, P. B., Lee, S. C., Rice, A. M., Gramling, J. M., Fitzhenry, T. M., McAlister, J. S., Harper, G. R., & Moran, A. L. (2004). Mislabelling of a depleted reef fish. *Nature, 430 (6997): 309-310.*

Mayr E. (1963). Animal species and evolution. Cambridge, MA: Harvard University Press.

Meyer A, Kocher TD, Basasibwaki P, Wilson AC (1990). Monophyletic origin of Lake Victoria cichlid fishes suggested by mitochondrial DNA sequences. *Nature* 347: *550–553*.

Meyer CP, Paulay G (2005). DNA barcoding: error rates based on comprehensive sampling. *PloS Biol* 3: *e422.*

Miller SE (2007). DNA barcoding and the renaissance of taxonomy. *PNAS* 104: *4775-4776.*

Montiel-Sosa JF, Ruiz-Pesini E, Montoya J, Roncales P, Lopez-Perez MJ, Perez-Martos A (2000). Direct and highly species-specific detection of pork meat and fat in meat products by PCR amplification of mitochondrial DNA. *J Agr Food Chem* 48 (7): *2829-2832.*

Moore WS (1995). Inferring Phylogenies from mtDNA Variation: Mitochondrial-Gene Trees Versus Nuclear-Gene Trees. *Evolution* 49 (4) :*718-726.*

Newmaster SG, Fazekas AJ, Ragupathy S (2006). DNA barcoding in land plants: evaluation of rbcL in a multigene tiered approach. *Can J Bot* 84: *335-341.*

Newmaster SG, Fazekas AJ, Steeves RAD, Janovec J (2007). Testing candidate plant barcode regions in the Myristicaceae. *Mol Ecol Res* 8: *480-490.*

Olmstead RG, Palmer JD (1994). Chloroplast DNA systematics: A review of methods and data analysis. *Am J Bot* 81(9): *1205-1224.*

Nanney DL (1982). Genes and phenes in *Tetrahymena*. *Bioscience* 32:*783–788.*

Neigel J, Domingo A, Stake J (2007). DNA barcoding as a tool for coral reef conservation. *Coral Reefs* 26(3): *487-499.*

Pace NR (1997). A molecular view of microbial diversity and the biosphere. *Science* 276 (5313): *734-740.*

Padial JM, De la Riva I (2009). Integrative taxonomy reveals cryptic Amazonian species of *Pristimantis* (Anura). *Zool J Linn Soc* 155: *97–122.*

Palmer JD, Herbon LA (1988). Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. *J Mol Evol* 28: *87-97.*

Palmer JD (1992). Mitochondrial DNA in plant systematics: applications and limitations. In: *Molecular Systematics of Plants,* D. Soltis, P. Soltis and J.J. Doyle, eds., Chapman and Hall, pp. 36-49.

Palumbi AR, Cipriano F (1998). Species identification using genetic tools: the value of nuclear and mitochondrial gene sequences in whale conservation. *J Heredity* 89(5): *459-464*.

Paquin P, Hedin M (2004). The power and perils of "molecular taxonomy": a case study of eyeless and endangered *Cicurina* (Araneae: Dictynidae) from Texas caves. *Mol Ecol* 13:*3239-3255.*

Pegg GG, Sinclair B, Briskey L, Aspden WJ (2006). MtDNA barcode identification of fish larvae in the southern Great Barrier Reef, Australia. *Sci Mar* 70(S2): *7–12.*

Pesole G, Gissi C, De Chirico A, and Saccone C (1999). Nucleotide substitution rate of mammalian mitochondrial genomes. *J Mol Evol* 48(4): *427–434.*

Rach J, De Salle R, Sarkar IN, Schierwater B, Hadrys H (2008). Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proc R Soc B* 275: 237-247.

Ratnasingham S, Hebert PDN, (2007). BOLD: The Barcode of Life Data System. *Mol Ecol Notes* 7 (3): *355-364.*

Remigio EA and Hebert PDN (2003). Testing the utility of partial COI sequences for phylogenetic estimates of gastropod relationships. *Mol Phylogenet Evol* 29: *641-647.*

Ross HA, Lento GM, Dalebout ML, Goode M, Ewing G, McLaren P, Rodrigo AG, Lavery S, Baker CS (2003). DNA Surveillance: Web-Based Molecular Identification of Whales, Dolphins, and Porpoises. *J Hered* 94(2): *111–114.*

Rubinoff D (2006a). Utility of mitochondrial DNA barcodes in species conservation. *Conserv Biol* 20(4): *1026-1033*.

Rubinoff D, Cameron S, Will K (2006b). A Genomic Perspective on the Shortcomings of Mitochondrial DNA for "Barcoding" Identification. *J Heredity* 97 (6): *581-594.*

Rusch DB *et al*. (2007). The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific PLoS 5 (3): e77

Ruttkay H, Solignac M, Sperlich D. (1992). Nuclear and mitochondrial ribosomal RNA variability in the obscura group of Drosophila. *Genetica* 85:*131– 138*.

Saccone C, De Giorgi C, Gissi C, Pesole G, Reyes A (1999). Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene* 238(1): *195-209.*

Salgado J, Honorato B, Garcia-Foncillas J (2008). Review: Mitochondrial Defects in Breast Cancer. *Oncology* 2: *199–207.*

Sass C, Little DP, Stevenson DW, Specht CD (2007). DNA barcoding in the Cycadales: testing the potential of proposed barcoding markers for species identification of Cycads. *PLoS ONE* 11: *e1154*.

Saunders GW (2005). Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. *Philos Trans R Soc B* 360 (1462): *1879-1888.*

Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R (2005). Towards writing the encyclopaedia of life: an introduction to DNA barcoding. *Phil Trans R Soc B* 360: *1805-1811*.

Scicluna SM, Tawari B, Clark CG (2006). DNA Barcoding of Blastocystis. *Protist* 157(1): *77-85.*

Schindel DE, Miller SE (2005). DNA barcoding a useful tool for taxonomists. *Nature* 435: *17.*

Seifert KA, Samson RA, deWaard JR, Houbraken J, Lévesque CA, Moncalvo JM, Louis-Seize G, Hebert PDN (2007). Prospects for fungus identification using C01 DNA barcodes, with Penicillium as a test case. *PNAS* 104(10): *3901-3906.*

Shaw J, Lickey EB, Beck JT, Farmer SB, Liu W, Miller J, Siripun KC, Winder CT, Schilling EE, Small RL (2005). The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *Am J Bot* 92: *142–166.*

Shearer TL, Van Oppen MJH, Romano SL, Worheide G (2002). Slow mitochondrial DNA sequence evolution in the Anthozoa (Cnidaria). *Mol Ecol* 11: *2475–2487*

Sheffield CS, Westby SM (2007). The male of *Megachile nivalis* Friese, with an updated key to members of the subgenus *Megachile* s. str. (Hymenoptera: Megachilidae) in North America. *J Hymenopt Res* 16: *178–191.*

Simmons RB, Weller SJ (2001). Utility and evolution of cytochrome *b* in insects. *Mol Phylogenet Evol* 20 (2): *196-210.*

Smith MA, Wood DM, Janzen DH, Hallwachs W, Hebert PDN (2007). DNA barcodes affirm that 16 species of apparently generalist tropical parasitoid flies (Diptera, Tachinidae) are not all generalists. *PNAS* 104 (12): *4967-4972.*

Smith VS (2005). DNA barcoding: Perspectives from a "Partnerships for Enhancing Expertise in Taxonomy" (PEET) debate. *Syst Biol* 54: *841-844.*

Strugnell J, Lindgren A (2007). A barcode of life database for the Cephalopoda? Considerations and concerns. *Rev Fish Biol Fischer* 17: *337-344.*

Taberlet P, Gielly L, Pautou G, Bouvet J (1991). Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Mol Biol* 17: *1105-1110.*

Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermat T, Corthier G, Brochmann C, Willerslev E (2007). Power and limitations of the chloroplast *trn*L (UAA) intron for plant DNA barcoding. *Nucleic Acids Res* 35(3): *e14.*

Tanabe S, Miyauchi E, Muneshige A, Mio K, Sato C, Sato M (2007). PCR method of detecting pork in foods for verifying allergen labelling and for identifying hidden pork ingredients in processed foods. *Biosci Biotechnol Bioch* 71 (7): *1663-1667.*

Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP (2003). A plea for DNA taxonomy. *Trends Ecol Evol* 18(2): *70–74.*

Tavares ES, Baker AJ (2008). Single mitochondrial gene barcodes reliably identify sister-species in diverse clades of birds. *BMC Evol Biol* 8: 81.

Teletchea F, Maudet C, Hanni C (2005). Food and forensic molecular identification: update and challenges. *Trends Biotechnol* 23 (7): *359-366.*

Terranova MS, Brutto SL, Arculeo M, Mitton JB (2007). A mitochondrial phylogeography of *Brachidontes variabilis* (Bivalvia: Mytilidae) reveals three cryptic species. *J Zool Syst Evol Res* 45 (4): *289–298.*

Tringe SG, Rubin EM (2005). Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet* 6 : *805-814.*

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428: *37–43.*

van Velzen R, Bakker FT, Loon JJA (2007). DNA barcoding reveals hidden species diversity in *Cymothoe* (Nymphalidae). *Proc Neth Entomol Meet* 18: *95-103.*

Vences M, Thomas M, Bonett RM, Vieites DR (2005). Deciphering amphibian diversity through DNA barcoding: chances and challenges. *Philos T R Soc B* 360: *1859-1868.*

Viñas J, Tudela S (2009). A validated methodology for genetic identification of tuna species (Genus *Thunnus*). *PLoS ONE* 4 (10): *e7606.*

Vogler AP, DeSalle R (1994). Diagnosing units of conservation management. *Conserv Biol* 8: *354-363.*

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005). DNA barcoding Australia's fish species. *Phil Trans R Soc B* 360 (1462): *1847-1857.*

Wares JP, Cunningham CW (2001). Phylogeography and historical ecology of the North Atlantic intertidal. *Evolution* 55: *2455–2469.*

Webb KE, Barnes DKA, Clark MS, Bowden DA (2006). DNA barcoding: a molecular tool to identify Antarctic marine larvae. *Deep-Sea Res*. Part II 53: *1053–1060.*

Wells JD, Stevens JR (2008). Application of DNA-Based Methods in Forensic Entomology. *Annu Rev Entomol* 53: *103-120.*

Wiemers M, Fiedler K (2007). Does the DNA barcoding gap exist? A case study in blue butterflies (Lepidoptera: Lycanidae). *Front Zool* 4: 8.

Will KW, Mishler BD, Wheeler QD (2005). The Perils of DNA Barcoding and the Need for Integrative Taxonomy. *Syst Biol* 54(5): *844–851.*

Williams ST, Knowlton N (2001). Mitochondrial pseudogenes are pervasive and often insidious in the snapping shrimp genus Alpheus. *Mol Biol Evol* 18(8): *1484-1493.*

Witt JDS, Threloff DL, Hebert PDN (2006). DNA barcoding reveals extraordinary cryptic diversity in an amphipod genus: implications for desert spring conservation. *Mol Ecol* 15: *3073–3082.*

Woese CR, Fox GE (1977). Phylogenetic Structure of the prokaryotic domain: the primary kingdoms. *PNAS* 74(11): *5088-5090.*

Wolfe KH, Li WH, Sharp PM (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *PNAS* 84 (24): *9054-9058.*

Won H, Renner SS (2003). Horizontal gene transfer from flowering plants to Gnetum. *PNAS* 100 (19): *10824-108291.*

Wong EH-K, and Hanner RH (2008). DNA barcoding detects market substitution in North-American seafood. *Food Res Int* 41 (8): *828-837.*

Zhang D-X, Hewitt GM (1996). Nuclear integrations: challenges for mitochondrial DNA markers. *Trends Ecol Evol* 11 (6)*: 247-251.*

Zhang DX, Hewitt GM, (1997). Insect mitochondrial control region: a review of its structure, evolution and usefulness in evolutionary studies. *Biochem Syst Ecol* 25:*99-120.*

Zwickl DJ, Hillis DM (2002). Increased taxon sampling greatly reduces phylogenetic error. *Syst Biol* 51(4): *588-598*.

Web sources

http://www.ebi.ac.uk/htbin/Mitbase/mitbase.pl

http://www.cbd.int

http://www.barcoding.si.edu

http://www.barcodinglife.com/

http://www.barcodinglife.com

http://tolweb.org/tree/

http://www.ebi.ac.uk/htbin/Mitbase/mitbase.pl

Chapter 2

Use of DNA barcoding in seafood products

Use of DNA barcoding for the genetic traceability of commercial seafood products

S. Nicolè†, E. Negrisolo§, G. Eccher†, R. Mantovani‡, T. Patarnello§, D.L. Erickson±, W.J. Kress± and G. Barcaccia†*

†Department of Environmental Agronomy and Crop Science; ‡Department of Animal Science. Faculty of Agriculture. §Department of Public Health, Comparative Pathology and Veterinary Hygiene, Faculty of Veterinary Medicine. University of Padova, Campus of Agripolis – Viale dell'Università 16, 35020 Legnaro, Padova (Italy). ±Department of Botany and Laboratories of Analytical Biology – National Museum of Natural History, Smithsonian Institution, P.O. Box 37012, Washington, DC 20013-7012 (USA).

*Gianni Barcaccia
Phone: +39 049 827 2814 Fax: +39 049 827 2839 E-mail: gianni.barcaccia@unipd.it

Abstract

DNA barcoding is a microgenomic identification system that allows the discrimination of life forms through the analysis of a small portion of the mitochondrial gene cox1 for animals. In this paper we report a practical application of DNA barcoding as a forensic tool to empower genetic traceability of marine organisms, particularly in commercial applications. We adopted a multi-locus approach based not only on cox1, but also on cob and 16S-rDNA genes, using the sequences deposited in BOLD and GenBank databases as reference standards. Our method proved to be a fast and reliable tool to recognize crustaceans, molluscs and fish fillets void of morphological attributes. Five of the 37 analyzed seafood specimens were shown to derive most likely from substitutions, voluntary or by accident, with different species. This approach will clearly be useful in implementing conservation policies, particularly for monitoring the illegal trade of protected and endangered species or to detect mislabeling in commercial processed seafood.

Keywords: DNA barcoding, genetic traceability, BOLD, seafood, mislabeling

Introduction

DNA barcoding is a technique for identifying species by obtaining a short DNA sequence from a known gene and comparing it with databases of orthologous sequences from species of expert-identified voucher specimens (Hebert *et al*., 2003). It aims to obtain a single gene universally amplified across metazoans, so that all species will be delineated by their unique barcode sequence or by a tight cluster of very similar sequences (Ward *et al*., 2005). In fact the core assumption of DNA barcoding is that the variation of the nucleotide sequences within species is much less than the differences among species (Meyer and Paulay, 2005)).

Animal DNA markers suitable for genetic traceability purposes usually belong to the mitochondrial genome. In animals, the mitochondria exhibits a higher rate of nucleotide substitution compared to nuclear DNA, is usually maternally inherited thus minimizing issues of hybridization and its high copy number facilitates PCR and sequence recovery, even from degraded tissues (Saccone *et al*., 1999; Herbert *et al*., 2004b). Furthermore, its simpler composition compared to nuclear DNA due to its lack of introns, pseudogenes or repetitive sequences allows easy global multiple sequence alignments (Lin *et al*., 2005). Finally, the availability of several complete mtDNA genome sequences allowed the design of robust and universal primers, which enable routine recovery from specific regions in a broad range of eukaryotes (Folmer *et al*., 1994; Simmons and Weller, 2001), as well as taxon specific primers, able to amplify only in targeted species without requiring subsequent sequencing step or other PCR-based techniques (Montiel-Sosa *et al*., 2000); (Lin and Hwang, 2008)).

The cox1 gene, encoding for cytochrome oxydase subunit I, was originally proposed as specific mitochondrial marker for DNA barcoding in the animal kingdom. A 648 nucleotide long sequence was selected near to the 5′ end of the gene with two conserved flanking sites where universal primers were designed across a wide taxonomic range of animal groups (Hebert *et al*., 2003; Folmer *et al*., 1994; Rach *et al*., 2008)). The bioidentification system based on cox1 has supplied very reliable results in several animal clades tested so far (Hebert *et al*., 2004a; Hogg and Hebert, 2004; Lin *et al*., 2005; Hajibabaei *et al*., 2006a; Yoo *et al*., 2006; Dawnay *et al*., 2007) and has also provided especially strong resolution at the species level for fish (Hogg and Hebert, 2004; Ivanova *et al.*, 2007; Hubert *et al.*, 2008). Due to these results, the barcoding community has committed itself in an initiative called Fish-BOL (Fish Barcode of Life initiative) that seeks to assemble a comprehensive reference system, based on cox1 marker, for all the 20,000 marine and 15,000 fresh-water fish species estimated on Earth (http://www.fishbol.org/index.php). This project aims to contribute to the management of fish biodiversity and, in conjunction with other web-resources, such as FishBase (http://www.fishbase.org/search.php) or FishTrace (http://www.fishtrace.org/), it will help to develop the Catalogue of Life (http://www.catalogueoflife.org/search.php), an exhaustive database of all known species of organisms on Earth.

Although cox1 is the most popular candidate for DNA barcoding in animal species, other regions have been suggested as barcode markers: the cob gene encoding for Apocytochrome-b (Lin *et al*., 2005; Pepe *et al*., 2007), that represented in the past the most sequenced marker for phylogenetic purposes in several taxa, the cox2 and cox3 genes encoding for mitochondrial cytochrome oxydase subunit II and subunit III, respectively (Park *et al*., 2007), the nad1 gene (encoding for NADH dehydrogenase 1 subunit (Rach *et al*., 2008) and the ribosomal 16S-rDNA (Willows-Munro *et al*., 2005). In contrast the only nuclear DNA region investigated for barcoding potential is that of internal transcribed spacers of the ribosomal RNA genes, ITS1 and ITS2 (Markmann and Tauz, 2005). ITS regions have been officially proposed as the DNA barcode for fungi (Zeng and De Hoog, 2008) and now the identification engine through ITS-based markers is available on the BOLD web site. In addition ITS markers have been successfully used for the identification of plants (Sass *et al*., 2007), protozoa (Guggiari and Peck, 2008) and freshwater sponges (Meixner *et al*., 2007).

DNA barcoding can find application in several fields, from monitoring biodiversity (*e.g.*, taxonomic, ecological and conservation studies) to forensic science for recognizing species in all the circumstances in which distinctive morphological characters, routinely used for the attribution of taxonomic entities, are scanty or absent (Armstrong and Ball, 2005). This potential turns out particularly useful for recognizing organisms in presence of morphological ambiguities, *i.e*. in larval stages (Pegg *et al*., 2006) or because of homoplasy and phenotypic plasticity of a given diagnostic character to environmental factors (Vences *et al*., 2005). In addition, DNA barcoding could contribute to monitor the illegal trade of wildlife products, such as protected and endangered species (Baker and Palumbi, 1994; Baker *et al*., 2000; Shivji *et al*., 2002), or to detect the species origin of commercial processed food items (Dawnay *et al*., 2007; Tanabe *et al*., 2007).

The application of this technique in food authentication is gaining attention because of food safety concerns, caused by the frequent cases of market substitutions (Hsieh, 1996; Marko *et al*., 2004) as well as recent food emergencies (Teletchea *et al*., 2005). Therefore the identification of the origin of feed and food ingredients is of primary importance for the protection of consumers against potential food adulteration and faulty ingredient declaration (Tanabe *et al*., 2007), GMOs (Ronning *et al*., 2005) and food poisoning (Hsieh *et al*., 2002). As reported by U.S. Food and Drug Administration the substitutions of fishes in seafood derivates are getting very common and demand the development of analytical methods to detect voluntary or involuntary mislabeling (http://www.fda.gov/Food/FoodSafety/Product-SpecificInformation/Seafood/RegulatoryFishEncyclopediaRFE/ucm071528.htm).

Several methods are available as identification tools for fish species, from traditional morphological observations to molecular approaches that include genomics and proteomics techniques (Rehbein *et al*., 1995; Martinez and Danielsdottir, 2000; Trotta *et al*., 2005). In this paper we quantify the power of DNA barcoding as a genetic tool implemented in the diagnosis and detection of fish components and/or the identification of species in seafood products. PCR with sequence-specific universal primers was employed to amplify three mitochondrial genes (mt16S-rDNA, cox1 and cob) in raw, frozen and processed commercial seafood. Two approaches were investigated to determine the power of DNA barcode data to correctly identify food products. First sequences were directly compared with existing libraries of DNA sequence using the DNA identification engine at BOLD (Barcode of Life Database, based on the HMM algorithm designed by Eddy (1998), and GenBank using BLAST (Altschul *et al*., 1990). A second tool for product identification was implementation of distance-based approach using NJ trees, which provide a visual inspection of query sequence identity based on tree topology. The aim of the study was to verify the label information of several seafood products in a multi-locus DNA barcoding strategy and also to estimate and compare the reliability of the two most common gene repositories used for phylogenetic and forensic purposes, GenBank and BOLD databases.

Materials and Methods

Collection of seafood samples

A total of 37 seafood samples, including raw, frozen and processed meat, of different commercial brands were collected from markets and groceries of the North-Eastern Italy. Most of them reported on the label a clear indication of both genus and species of the organism, in addition to the common name and capture place, as required law. The others were obtained at the marketplace of Chioggia (Venice) and they showed only the common or vernacular name with, sometimes, the indication of the origin area. In particular, the commercial products included 30 fishes, three crustaceans and four molluscs: some of them were sold as fresh or frozen skinned fillets, while others had undergone different treatments, such as heat treatments and canning processes (**Table 1**). Finally, three seafood products included more than one species and the scientific names of the organisms were indicated on the label.

DNA extraction, amplification and sequencing

Total genomic DNA was extracted and purified from muscle tissue of the 37 samples using GenElute Mammalian Genomic DNA Miniprep Kit (SIGMA) following instructions of the manufacturer with few changes. The specific DNA barcode region of each target gene was amplified in duplicates. All PCR experiments were performed using a GeneAmp PCR System 9700 (Applied Biosystems, Foster City, CA, USA) and the amplification were carried out respecting the instruction supplied in Barcoding Animal Life website (http://www.dnabarcoding.ca/primer/Index.html).

Typical conditions for cox1 amplification include the initial denaturation at 94°C for 1 min, five cycles of 94°C for 30 sec, annealing at 50-55°C for 40 sec, and extension at 72°C for 1 min, followed by 30-35 cycles of 94°C for 30 sec, 55-60°C for 40 sec, and 72°C for 1 min, with a final extension at 72°C for 10 min, followed by indefinite hold at 4°C. We tested only one pair of universal primers for the markers 16S-rDNA and cob, whereas for the cox1 gene we first tested the universal primers from Ward *et al*. (2005) and where the primers failed, a different pair was exploited. The list with the nucleotide sequence for each primer along with annealing temperature and the corresponding reference is reported in **Table 2**. The 25 µl PCR reaction volumes included 1× PCR buffer (100 mM Tris-HCl pH 9.0, 15 mM $MgCl_2$ and 500 mM KCl), 0.2 mM dNTPs, 0.4 µM of each primer, 1 U of *Taq* DNA polymerase and 15 ng of genomic DNA as template. PCR products were purified enzymatically by EXO/SAP (Amersham) and then directly sequenced bi-diretionally according to the original Rhodamine terminator cycle sequencing kit (ABI PRISM Applied Biosystems). Sequences were assembled into contigs, screened for errors in Mega V 4.1 (beta) (Kumar *et al*., 2008) and exported in FASTA format for use in database searches and tree based alignments.

Table 2. List of universal primers used for each mitochondrial marker with their nucleotide sequence.

Marker	**Primer name**	**Primer sequence (5'-3')**	**Ta (°C)**	**References**
cox1	FishF2	TCGACTAATCATAAAGATATCGGCAC	60	Ward *et al*., 2005
	FishR2	ACTTCAGGGTGACCGAAGAATCAGAA	60	Ward *et al*., 2005
	LCO1490	GGTCAACAAATCATAAAGATATTGG	60	Folmer *et al*., 1994
	HCO2198	TAAACTTCAGGGTGACCAAAAAATCA	60	Folmer *et al*., 1994
16S-rDNA	16Sar-5′	CGCCTGTTTATCAAAAACAT	55	Palumbi, 1996
	16Sbr-3′	CCGGTCTGAACTCAGATCACGT	55	Palumbi, 1996
cob	GLUDG-l	TGACTTGAARAACCAYCGTTG	60	Palumbi, 1996
	CB3-H	GGCAAATAGGAARTATCATTC	60	Palumbi *et al*., 1991

Table 1. Commercial samples analyzed by the multi-locus DNA barcoding approach developed (n.a., not available)

No.	Product description	Origin	Species declared in the label	Organism	Family	Processing treatments
16	Blue shark	Pacific Ocean, FAO 71	*Prionace glauca*	Fish	Carcharhinidae	Frozen fillet
15	Atlantic herring	n.a.	*Clupea harengus*	***Fish***	Clupeidae	Smoked, vacuum packaged
33	European anchovy	n.a.	*Engraulis encrasicolus*	***Fish***	Engraulidae	Brine, canned in vegetal oil
34	Atlantic cod	n.a.	*Gadus morhua**	***Fish***	Gadidae	Raw fillet
24	Pacific cod	n.a.	*Gadus macrocephalus*	***Fish***	Gadidae	Dried salted (baccalà)
53	Mako shark	n.a.	*Isurus oxyrhincus*	***Fish***	Lamnidae	Frozen fillet
9	Nile perch	n.a.	*Lates niloticus*	***Fish***	Latidae	Frozen fillet
27	Nile perch	Victoria lake, Africa	*Lates niloticus**	***Fish***	Latidae	Raw fillet
21	Angler	n.a.	*Lophius piscatorius*	***Fish***	Lophiidae	Raw fillet
3	South Pacific hake	South-West Pacific Ocean and Atlantic Ocean	*Merluccius gayi/productus*	***Fish***	Merlucciidae	Frozen, pre-cooked
5	Atlantic hake	South-East Atlantic Ocean	*Merluccius hubbsi*	***Fish***	Merlucciidae	Frozen fillet
8	Scarlet snapper	SouthAfrica Ocean and Indian Ocean	*Merluccius capensis/paradoxus*	***Fish***	Merlucciidae	Frozen, pre-cooked
6	Patagonian grenadier	Pacific Ocean	*Macruronus magellanicus*	***Fish***	Merlucciidae	Frozen fillet
29	Striped catfish	n.a.	*Pangasius hypophthalmus**	***Fish***	Pangasidae	Raw fillet
50	Striped catfish	n.a.	*Pangasius hypophthalmus*	***Fish***	Pangasidae	Raw fillet
13	Turbot	South-East Atlantic Ocean	*Paralichthys isosceles*	***Fish***	Paralichthydae	Frozen fillet
28	European perch	n.a.	*Perca fluviatilis**	***Fish***	Percidae	Raw fillet
4	European plaice	North-East Atlantic Ocean	*Pleuronectes platessa*	***Fish***	Pleuronectidae	Frozen fillet
51	European plaice	n.a.	*Pleuronectes platessa*	***Fish***	Pleuronectidae	Raw fillet
12	Rainbow trout	Farmed in Italy	*Oncorhynchus mykiss*	***Fish***	Salmonidae	Smoked, vacuum packaged
19	Atlantic salmon	n.a.	*Salmo salar*	***Fish***	Salmonidae	Smoked, vacuum packaged
30	Yellow-fin tuna	n.a.	*Thunnus albacares**	***Fish***	Scombridae	Raw fillet
36	Tuna chunks sashimi	n.a.	*Thunnus albacares*	***Fish***	Scombridae	Raw fillet

35	Yellow-fin tuna fillets	n.a.	*Thunnus albacares*	***Fish***	Scombridae	Raw fillet
31	Tuna	n.a.	*Thunnus albacares*	***Fish***	Scombridae	Carpaccio
23	Malabar grouper	n.a.	*Epinephelus malabaricus*	***Fish***	Serranidae	Raw fillet
22	Common sole	n.a.	*Solea solea*	***Fish***	Soleidae	Raw fillet
17	Smoked swordfish	n.a.	*Xiphias gladius*	***Fish***	Xiphiidae	Smoked, vacuum packaged
32	Swordfish carpaccio	n.a.	*Xiphias gladius*	***Fish***	Xiphiidae	Carpaccio
37	Swordfish fillets	n.a.	*Xiphias gladius**	***Fish***	Xiphiidae	Raw fillet
2	Greenshell mussel	Pacific Ocean	*Perna canaliculus*	Mollusc	Mytilidae	Frozen
25	Common octopus	n.a.	*Octopus vulgaris*	Mollusc	Octopodidae	Raw
52	Jumbo squid	n.a.	*Dosidicus gigas*	Mollusc	Ommastrephidae	Raw
18	Great Atlantic scallop	North-East Atlantic Ocean	*Pecten maximus*	Mollusc	Pectinidae	Frozen
11	Northern red shrimp	n.a.	*Pandalus borealis*	Crustacean	Pandalidae	Frozen
7	Pink prawn	Pacific Ocean and Indian Ocean	*Metapenaeus affinis/monoceros*	Crustacean	Penaeidae	Frozen
14	Whiteleg shrimp	n.a.	*Penaeus Vannamei*	Crustacean	Penaeidae	Frozen

*, only the common name is indicated in the label, the scientific name is deducible in agreement with the Italian Ministerial Decree of the 14/01/2005.

BLAST and phenetic analyses

For forensic identification of species identity, both a similarity analysis and a phonetic approach were employed to check the correspondence between sequences of the unique amplicons used as query with the sequences deposited in GenBank and BOLD databases. Homology searches were conducted using the BLAST algorithm against GenBank database and the global alignment through Hidden Markov Model (HMM) against BOLD engine. Therefore two different databases were used as reference system: GenBank, for all the markers, and BOLD, only for cox1 region. In the case of specimen identification through BOLD, there were two tiers of comparison. The first attempt was conducted against a reference subset of the database made up only by validated sequences link to at least three voucher samples. When the BOLD interrogation reported no match, we used the full database that includes every cox1 barcode record, even unvalidated because represented by only one or two specimens.

The phenetic analysis was developed with CLC Sequence Viewer 6.2 for cox1 marker. The genetic distances among sequences were calculated using the K2P parameter and the visual representation was based on the construction of a Neighbour-Joining tree. The phenetic approach consisted in the inference of a NJ tree only for cox1 marker with, when possible, four validated sequences retrieved from BOLD for each species along with the sequences of the samples. In addition, for the species where the cox1 sequence was not available we used sequences from GenBank for those species. The reliability of the clusters formed at the species level in the tree was evaluated by means of a bootstrap test with 1,000 replications. An additional NJ tree was developed using Mega v.4.1 software for the genera that resulted to be polyphyletic, such as *Thunnus*, *Macruronus* and *Gadus*: all the cox1 sequences were retrieved from BOLD, or GenBank when a few entries were available in the former database, to draw a genus-specific tree in order to clarify the relationships among the species within that genus.

Results and Discussion

DNA extraction and PCR-based amplification success

We successfully isolated total genomic DNA from all 37 seafood-derived specimens of different commercial brands, including raw and frozen processed products, and skinned fillets. All of these DNA preparations proved to be accessible to amplification by PCR using universal primers. The PCR conditions as well as the universal primers adopted (see Table 1) were in agreement with the protocol indications supplied on the official barcode website (http://www.dnabarcoding.ca). The primers generated reliable and reproducible single amplification products, with an average length of about 700 bp for cox1, 500 bp for 16S-rDNA and 850 bp for cob gene. All the mtDNA sequences were deposited in NCBI databases on December 12, 2009 (GenBank Accession number: GU324135 - GU324234; **Appendix 1**).

In particular, 16S-rDNA primers worked universally allowing the recovery of the sequences for all commercial products, with one exception. They proved to be highly effective in generating a single amplicon for each target gene in all fish, mollusc and crustacean seafood derivatives, whereas primer pairs specific for cox1 and cob genes were less performing in terms of amplification success and/or specificity. The amplification of the cox1 target region failed in two crustaceans and one mollusc, while the cob-specific primers never worked in molluscs. In the cases of species mixture, sequencing problems were not experienced and double peaks were never detected. This could be due to either the absence of co-listed species or the predominance of one relative to others in the mixture. On the basis of the agarose gel-based electrophoresis analysis and sequence-specific amplification results, it was evident that the DNA was correctly preserved and thus it was possible the direct sequencing of all amplicons. Since the substitution events, fraudulent or by accident, generally involve fresh fillets sold in local marketplaces rather than seafood stuff commercialized by famous brands, we aimed to analyze mainly fresh raw fillets and a few frozen foodstuff, avoiding in this way problems related to the isolation of genomic DNA from processed items (see Table 1). Therefore it is important to test the primer pairs and the PCR conditions used in this assay also in specimens subjected to highly denaturating treatments, such as high temperature and low pH exposures, which often affect the integrity of the DNA hampering the amplification of target regions longer than 200 bp (Chapela *et al*., 2007; Rasmussen and Morrissey, 2008; Espiñeira *et al*., 2009).

Validation of the selected markers

To test and confirm the species declared on the label of each seafood product, we selected as target markers the reference barcode region cox1 along with other two sequences, 16S-rDNA and cob genes. These sequences were chosen because they represent some of the most common regions used for identification and forensic purposes and in fact they showed the widest taxonomic representation in the nucleotide databases of NCBI compared to other very common markers exploited for the same purposes, such as nad1, coding for NADH dehydrogenase subunit 1, or cox2, coding for cytochrome c oxidase subunit 2 (**Figure 1**). The choice of testing more than one target gene, according to a multi-locus DNA barcoding, is mainly due to the possibility of validation of label information contents or attribution of species to a commercial product by using independent replicates. Furthermore, since the BOLD sequence repertory is now far from being complete, using two additional genes improved the chance to find correct matches also for those species for which the cox1 sequence was not available (Dawnay *et al*., 2007). Obviously the central issue is the necessity to develop a reliable database with adequate reference sequence data able to accurately identify species.

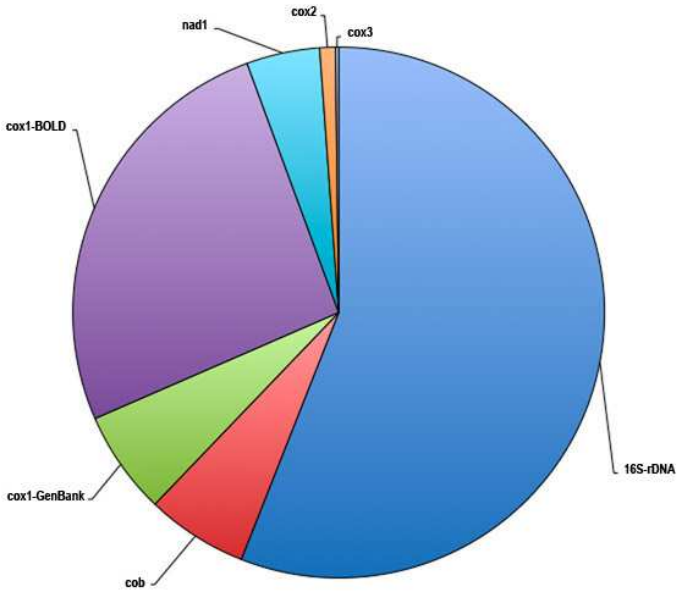


Figure 1. Proportion of sequence accessions related to cox1, nad1, cox2, cox3, 16S-rDNA and cob genes deposited in GenBank and/or BOLD databases.

BLAST and NJ distance-based approaches

A double approach was followed to check the identity of our samples: a similarity search, to establish the correspondence between sequences of the PCR products with that of the gene deposited in the databases, and a distance-based approach, commonly used for barcoding analyses.

To investigate the authenticity of the information reported in the labels, we compared DNA sequences from retail samples with those deposited in two online sequences repositories: GenBank, the gene database developed from NCBI, and BOLD, the new sequence repository born to support the large-scale DNA barcoding projects available, through the dedicated BOLD-ID engine, on the BOLD website (http://www.barcodinglife.org/views/login.php).

BLAST analysis is a suitable technique to find regions of local similarity between sequences, a feature that can turn out useful to identify species in a forensic context. BOLD engine, instead, generates species identification using a quick alignment of a query sequence to the global alignment of all reference sequences followed by a linear search of reference library. This genetic identification system delivers a species identification if the query sequence shows a tight match, less than 1% divergence, to a reference standard

(Ratnasingham and Hebert, 2007). Since the experimental procedure is almost standardized and affordable and fishes are considered an ideal target for cox1 validation in forensic context, the major limitation lies in the saturation of authenticated reference DNA sequences: the richer is the database the more chances there are to recognize an unknown specimen. Since BOLD is being developed using voucher samples, this sequence repository should contain only validated sequences that are promptly to be directly used for identification purposes (Wong and Hanner, 2008). Although this feature, only a subset of BOLD repository is a validated dataset because it includes sequence records referred to species represented by three or more individuals showing less than 2% sequence divergence. Unlike BOLD, it is universally recognized that GenBank contains reliable as well as unverified sequences due to the lack of quality control during the sequence submission phase (Forster, 2003). Thus the recourse to GenBank is motivated by the fact that cox1 sequences for the target species was not always public in BOLD for all the samples and the exploitation of other two genes could improve the chance to find a match with a deposited sequence. This approach allowed us to test the effectiveness of BOLD repository in order to verify if this web resource can be considered a valid tool to identify organisms and eventually to be applied for practical purposes as detecting frauds in seafood trade.

In the BLAST analysis approach, for each query a list of the most similar reference sequences is provided along with the BIT score which incorporates the percent identity (%) estimate and E-value, while in the BOLD search the species level match is valuated by a specimen similarity with divergence value less than 1% and, if the match is not obtained, the query sequence is assigned to a genus with a similarity divergence lower than 3%. On the basis of the mitochondrial DNA barcodes generated in this study, 15 fishes and one mollusc out of 37 selected seafood products could be properly assigned to the species reported in the label by means of all the three marker genes (**Table 3A,B**). Additionally 12 seafood products, of which 11 were fish and one a mollusc, were correctly identified as the species reported on the labels by means of two marker genes, while in other five cases, including two crustaceans, two fishes and one mollusc, the identification was based just on one marker gene. For five commercial products, we did not obtain any match with that declared on the label by means of any marker, even if the standard sequence for that species was available in the databases (see Table 3A,B). This finding led us to conclude that the specimens may have been subjected to substitution events and this idea is also supported by the fact that the match obtained at the species level was the same using all three mtDNA genetic markers.

The region that showed the highest number of positive and unambiguous matches was the cox1 gene (Table 3A). It scored the most frequent matches at the species level, 26 out of 32, with the expected reference sequence contained in GenBank database. When the similarity search was carried out using the BOLD database, the number of matches decreased to 21, mainly because of problems related to the identity of tuna species. In fact, when the ID engine at BOLD was queried, the sequence corresponding to the cox1 region from tuna specimens was always assigned at the level of genus only. Regarding the 16S-rDNA gene, even if the query of GenBank allowed us to assign the origin of the meat to the species level for 28 out of 37 samples, nine of these matches produced equal identity scores with more than one species, so providing no unambiguously reliable identification result (Table 3B). Finally, the cob gene was the most problematic marker. In fact, although it represented the best target for many phylogenetic and forensic studies of animals in the past, now it is becoming replaced by cox1 through the international campaigns, such as Fish-BOL. This sequence scored the worst rate of assignment with only 21 out of 37 products properly identified, five of which produced equal scores with several species (Table 3B). Nevertheless, it is noteworthy that in five situations the missing confirmation of the meat origin by means of the cob marker could be attributable to the unavailability of the reference sequence in the GenBank, events more frequent for this region rather than for the other two sequences (see Table 3). Unlike cob region, the missing standards were only two for 16S-rDNA and for cox1 gene four and two in GenBank and BOLD, respectively.

Table 3A. BLAST results obtained using as query cox1 sequences derived from the commercial seafood products under study.

		cox1					
No.	**Species declared in the label**	**GenBank/Blast**	**E-value**	**Max ID**	**BOLD/HMM**	**Similarity**	**Tree based identification****
2	*Perna canaliculus*	*Perna canaliculus*	0.00	99	*Perna canaliclus*	98.79	*Perna canaliculus*
3	*Merluccius gayi/productus*	*Merluccius hubbsi*	0.00	99	*Merluccius hubbsi*	99.5*	*Merluccius hubbsi*
4	*Pleuronectes platessa*	*Pleuronectes platessa*	0.00	99	*Pleuronectes platessa*	99.67	*Pleuronectes platessa*
5	*Merluccius hubbsi*	*Merluccius hubbsi*	0.00	100	*Merluccius hubbsi*	100*	*Merluccius hubbsi*
6	*Macruronus magellanicus*	*Macruronus magellanicus*	0.00	98	*Macruronus novaezelandiae*	99.54*	*Macruronus* spp.
7	*Metapenaeus affinis* [b]*/monoceros* [a, b]	n.d.			n.d.		
8	*Merluccius capensis/paradoxus*	*Merluccius paradoxus*	0.00	92	*Merluccius paradoxus*	100*	*Merluccius paradoxus*
9	*Lates niloticus*	*Lates niloticus*	0.00	100	*Lates niloticus*	100*	*Lates niloticus*
11	*Pandalus borealis* [a, b]	n.d.			n.d.		
12	*Oncorhynchus mykiss*	*Oncorhynchus mykiss*	0.00	100	*Oncorhynchus mykiss*	100	*Oncorhynchus mykiss*
13	*Paralichthys isosceles*	*Xystreurys rasile*	0.00	99	*Xystreurys rasile*	99.51*	*Xystreurys rasile*
14	*Penaeus vannamei*	*Xystreurys rasile*	0.00	100	*Xystreurys rasile*	100*	*Xystreurys rasile*
15	*Clupea harengus*	*Clupea harengus*	0.00	100	*Clupea harengus*	100*	*Clupea harengus*
16	*Prionace glauca*	*Prionace glauca*	0.00	100	*Prionace glauca*	100*	*Prionace glauca*
17	*Xiphias gladius*	*Xiphias gladius*	0.00	100	*Xiphias gladius*	100*	*Xiphias gladius*
18	*Pecten maximus*[a]	*n.d.*			*n.d.*		
19	*Salmo salar*	*Salmo salar*	0.00	99	*Salmo salar*	100*	*Salmo salar*
21	*Lophius piscatorius*	*Lophius piscatorius*	0.00	100	*Lophius piscatorius*	100*	*Lophius piscatorius*
22	*Solea vulgaris/solea*	*Solea solea*	0.00	99	*Solea solea*	99.84	*Solea solea*
23	*Epinephelus malabaricus*	*Epinephelus areolatus*	0.00	98	*Epinephelus areolatus*	98.71	*Epinephelus areolatus*
24	*Gadus macrocephalus*	*Gadus macrocephalus*	0.00	100	*Gadus ogac*	100*	*Gadus ogac*
25	*Octopus vulgaris* [a]	*Amphioctopus marginatus*	0.00	99	*Amphioctopus marginatus*	100	*Amphioctopus marginatus*
27	*Lates niloticus*	*Lates niloticus*	0.00	100	*Lates niloticus*	100*	*Lates niloticus*
28	*Perca fluviatilis*	*Paralichthys* spp.	0.00	88	*Paralichthys patagonicus*	100*	*Paralichthys patagonicus*
29	*Pangasius hypophthalmus*	*Pangasius hypophthalmus*	0.00	100	*Pangasius hypophthalmus*	100*	*Pangasius hypophthalmus*
30	*Thunnus albacares*	*Thunnus albacares*	0.00	100	*Thunnus obesus*	100*	*Thunnus* spp.
31	*Thunnus albacares*	*Thunnus albacares*	0.00	100	*Thunnus* spp. (1)	100*	*Thunnus* spp.
32	*Xiphias gladius*	*Xiphias gladius*	0.00	99	*Xiphias gladius*	100*	*Xiphias gladius*
33	*Engraulis encrasicolus*	*Thunnus albacares*	0.00	100	*Thunnus* spp. (2)	99.84*	*Thunnus* spp..
34	*Gadus morhua*	*Gadus morhua*	0.00	98	*Gadus morhua*	99.84*	*Gadus morhua*
35	*Thunnus albacares*	*Thunnus albacares*	0.00	100	*Thunnus* spp. (2)	100*	*Thunnus* spp..
36	*Thunnus albacares*	*Thunnus albacares*	0.00	100	*Thunnus* spp. (2)	100*	*Thunnus* spp..
37	*Xiphias gladius*	*Xiphias gladius*	0.00	100	*Xiphias gladius*	100*	*Xiphias gladius*
50	*Pangasius hypophthalmus*	*Pangasius hypophthalmus*	0.00	100	*Pangasius hypophthalmus*	100*	*Pangasius hypophthalmus*
51	*Pleuronectes platessa*	*Pleuronectes platessa*	0.00	100	*Pleuronectes platessa*	100	*Pleuronectes platessa*
52	*Dosidicus gigas*	*Dosidicus gigas*	0.00	99	*Dosidicus gigas*	99.83	*Dosidicus gigas*
53	*Isurus oxyrhincus*	*Isurus oxyrhincus*	0.00	99	*Isurus oxyrhincus*	99.84	*Isurus oxyrhincus*

Table 3B. BLAST results obtained using as query 16S and cob sequences derived from the commercial seafood products under study.

		16S-rDNA			***cob***		
No.	***Species declared in the label***	***GenBank/Blast***	***E-value***	***Max***	***GenBank/Blast***	***E-value***	***Max ID***
2	*Perna canaliculus* [d]	*Perna canaliculus*	8.00E-101	100	n.d.		
3	*Merluccius gayi/productus*	*Merluccius hubbsi*	0.00	100	*Merluccius productus*	0.00	96
4	*Pleuronectes platessa*	*Pleuronectes platessa* (3)	0.00	100	*Pleuronectes platessa*	0.00	99
5	*Merluccius hubbsi*	*Merluccius hubbsi*	0.00	100	*Merluccius hubbsi*	0.00	98
6	*Macruronus magellanicus*	*Macruronus*	0.00	100	*Macruronus magellanicus* (11)	0.00	100
7	*Metapenaeus affinis* [c,d]*/monoceros* [d]	*Litopenaeus vannamei*	0.00	100	*Macruronus spp* (11)	0.00	100
8	*Merluccius capensis/paradoxus*	*Merluccius paradoxus*	0.00	100	*Merluccius paradoxus*	0.00	99
9	*Lates niloticus*	*Lates niloticus*	0.00	99	*Merluccius hubbsi*	0.00	98
11	*Pandalus borealis* [d]	*Pandalus borealis*	0.00	97	*Oncorhynchus mykiss*	0.00	100
12	*Oncorhynchus mykiss*	*Oncorhynchus mykiss*	0.00	99	*Oncorhynchus mykiss*	0.00	100
13	*Paralichthys isosceles*[c, d]	*Xystreurys liolepis*	0.00	96	*Oncorhynchus mykiss*	0.00	100
14	*Penaeus vannamei*	*Penaeus Vannamei*	0.00	100	*Oncorhynchus mykiss*	0.00	100
15	*Clupea harengus*	*Clupea harengus*	0.00	100	*Oncorhynchus mykiss*	0.00	100
16	*Prionace glauca*	*Prionace glauca*	0.00	100	*Prionace glauca*	0.00	100
17	*Xiphias gladius*	*Xiphias gladius*	0.00	99	*Xiphias gladius*	0.00	99
18	*Pecten maximus* [d]	*Pecten maximus*	0.00	99	n.d.		
19	*Salmo salar*	*Salmo salar*	0.00	100	n.d.		
21	*Lophius piscatorius*	*Lophius piscatorius*	0.00	98	*Solea solea*	4.00E-	95
22	*Solea vulgaris/solea*	*Solea solea*	0.00	99	*Solea solea*	0.00	100
23	*Epinephelus malabaricus*	n.d.			n.d.		
24	*Gadus macrocephalus*	*Gadus macrocephalus* (4)	0.00	100	*Gadus macrocephalus* (12)	0.00	99
25	*Octopus vulgaris*	*Octopus spp.* (5)	0.00	99	n.d.		
27	*Lates niloticus*	*Lates niloticus*	3.00E-133	95	*Chelidonichthys lucernus*	0.00	96
28	*Perca fluviatilis*	*Paralichthys patagonicus*	0.00	100	*Paralichthys olivaceus*	0.00	88
29	*Pangasius hypophthalmus*	*Pangasius hypophthalmus*	0.00	99	*Pangasius hypophthalmus* (13)	0.00	99
30	*Thunnus albacares*	*Thunnus albacares* (7)	0.00	99	*Thunnus albacares*	0.00	100
31	*Thunnus albacares*	*Thunnus albacares* (7)	0.00	99	*Thunnus albacares*	0.00	99
32	*Xiphias gladius*	*Xiphias gladius*	0.00	99	*Xiphias gladius*	0.00	99
33	*Engraulis encrasicolus*	*Engraulis encrasicolus* (8)	0.00	99	*Thunnus albacares*	0.00	97
34	*Gadus morhua*	*Gadus morhua*	0.00	99	*Gadus morhua*	0.00	99
35	*Thunnus albacares*	*Thunnus albacares* (7)	0.00	99	*Thunnus albacares*	0.00	100
36	*Thunnus albacares*	*Thunnus albacares* (9)	0.00	99	*Thunnus albacares*	0.00	99
37	*Xiphias gladius*	*Xiphias gladius*	0.00	99	*Xiphias gladius*	0.00	99
50	*Pangasius hypophthalmus*	*Pangasius hypophthalmus*	0.00	100	*Pangasius hypophthalmus* (13)	0.00	99
51	*Pleuronectes platessa*	*Pleuronectes platessa* (10)	0.00	99	*Pangasius spp.* (13)	0.00	98
52	*Dosidicus gigas*	*Dosidicus gigas*	0.00	98	n.d.		
53	*Isurus oxyrhincus*	*Isurus oxyrhincus*	0.00	98	n.d.		

n.d., not determined; [a], [c], [d], no sequence of the labelled species is available in GenBank for cox1, 16S-rDNA and cob, respectively; [b], no sequence of the labelled species is available in BOLD for cox1; **, the threshold divergence value to distinguish different species is 1%, specimens with divergence value minor than 1% cluster together; *, Blast match versus validated sequence BOLD library.

(1), *Thunnus obesus, Thunnus atlanticus*; (2), *Thunnus obesus, Thunnus atlanticus*; (3), *Pleuronectes platessa, Platichthys stellatus*; (4), *Gadus macrocephalus, Gadus ogac*; (5), *Octopus aegina, Octopus marginatus*; (6), *Pangasius hypophthlmus, Pangasius sutchi*; (7), *Thunnus albacares, Thunnus orientalis, Thunnus thynnus thynnus*; (8), *Engraulis encrasicolus, Engraulis eurystole, Engraulis japonicus, Engraulis australis*; (9), *Thunnus albacares, Thunnus orientalis, Thunnus thynnus thynnus, Thunnus alalunga*; (10), *Pleuronectes platessa, Platichthys stellatus, Platichthys flesus, Psettichthys melanostictus, Isopsetta isolepis, Lepidopsetta bilineata, Pseudopleuronectes americanus, Parophrys vetulus*; (11), *Macruronus magellanicus, Macruronus novaezelandiae*; (12), *Gadus macrocephalus, Gadus ogac*; (13), *Pangasius sutchi, Pangasius spp., Pangasius hypophthalmus*.

Furthermore, a phenetic approach based on the construction of a Neighbour-Joining tree, using only the validated cox1 reference sequences, was adopted as additional tool to give a graphic representation of the results obtained using similarity search (**Figure 2**). In this NJ tree, the entries belonging to individuals of a given species were clustered in the same monophyletic group, exception made for the cases of specimens declared as *Thunnus*, *Macruronus* and *Gadus* where the subdivision of the species in distinct clusters was poorly resolved. Regarding the sequences of the collected specimens, most of them grouped with the species declared in the label, allowing their identification, while in few cases the cox1 sequence clustered with a different species, probably because of involuntary substitution or faulty declaration events.

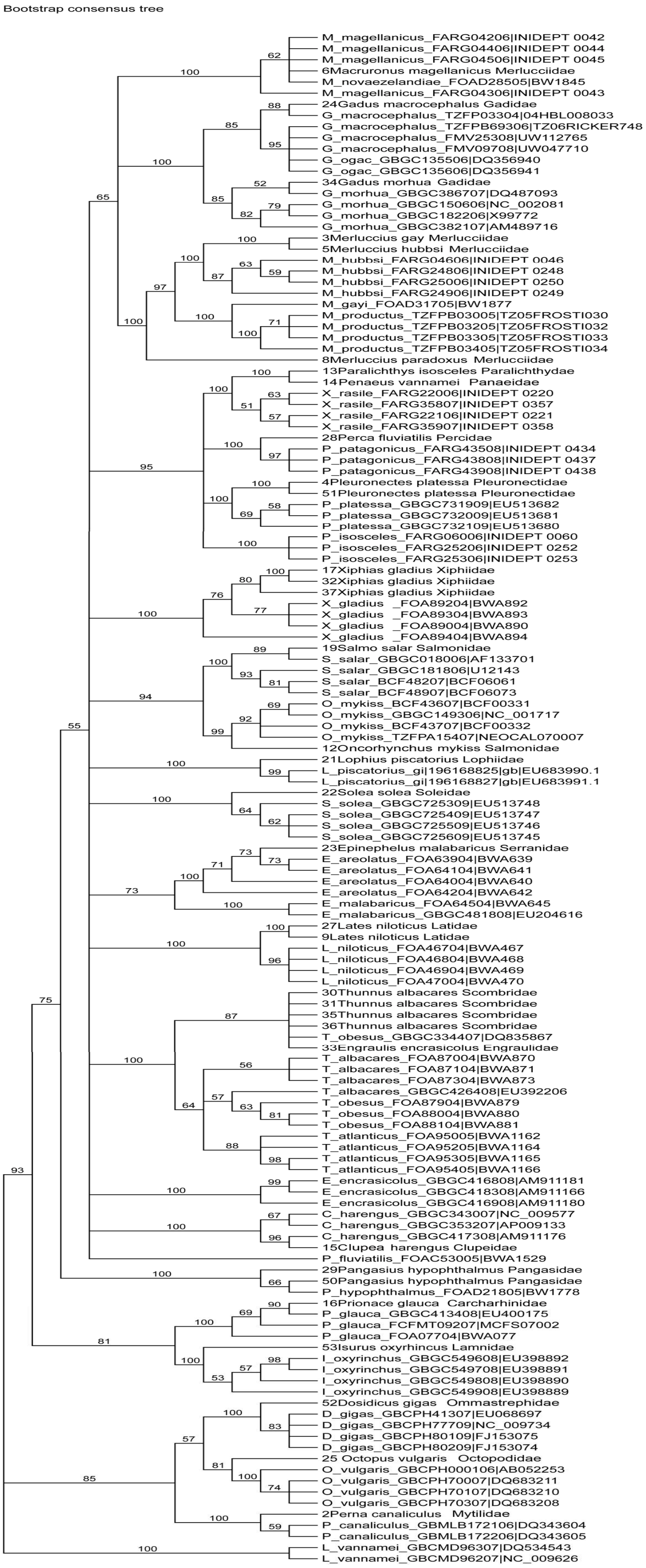

Bootstrap consensus tree
M_magellanicus_FARG04206|INIDEPT 0042
M_magellanicus_FARG04406|INIDEPT 0044
M_magellanicus_FARG04506|INIDEPT 0045
6Macruronus magellanicus Merlucciidae
M_novaezelandiae_FOAD28505|BW1845
M_magellanicus_FARG04306|INIDEPT 0043
24Gadus macrocephalus Gadidae
G_macrocephalus_TZFP03304|04HBL008033
G_macrocephalus_TZFPB69306|TZ06RICKER748
G_macrocephalus_FMV25308|UW112765
G_macrocephalus_FMV09708|UW047710
G_ogac_GBGC135506|DQ356940
G_ogac_GBGC135606|DQ356941
34Gadus morhua Gadidae
G_morhua_GBGC386707|DQ487093
G_morhua_GBGC150606|NC_002081
G_morhua_GBGC182206|X99772
G_morhua_GBGC382107|AM489716
3Merluccius gay Merlucciidae
5Merluccius hubbsi Merlucciidae
M_hubbsi_FARG04606|INIDEPT 0046
M_hubbsi_FARG24806|INIDEPT 0248
M_hubbsi_FARG25006|INIDEPT 0250
M_hubbsi_FARG24906|INIDEPT 0249
M_gayi_FOAD31705|BW1877
M_productus_TZFPB03005|TZ05FROSTI030
M_productus_TZFPB03205|TZ05FROSTI032
M_productus_TZFPB03305|TZ05FROSTI033
M_productus_TZFPB03405|TZ05FROSTI034
8Merluccius paradoxus Merlucciidae
13Paralichthys isosceles Paralichthydae
14Penaeus vannamei Panaeidae
X_rasile_FARG22006|INIDEPT 0220
X_rasile_FARG35807|INIDEPT 0357
X_rasile_FARG22106|INIDEPT 0221
X_rasile_FARG35907|INIDEPT 0358
28Perca fluviatilis Percidae
P_patagonicus_FARG43508|INIDEPT 0434
P_patagonicus_FARG43808|INIDEPT 0437
P_patagonicus_FARG43908|INIDEPT 0438
4Pleuronectes platessa Pleuronectidae
51Pleuronectes platessa Pleuronectidae
P_platessa_GBGC731909|EU513682
P_platessa_GBGC732009|EU513681
P_platessa_GBGC732109|EU513680
P_isosceles_FARG06006|INIDEPT 0060
P_isosceles_FARG25206|INIDEPT 0252
P_isosceles_FARG25306|INIDEPT 0253
17Xiphias gladius Xiphiidae
32Xiphias gladius Xiphiidae
37Xiphias gladius Xiphiidae
X_gladius _FOA89204|BWA892
X_gladius _FOA89304|BWA893
X_gladius _FOA89004|BWA890
X_gladius _FOA89404|BWA894
19Salmo salar Salmonidae
S_salar_GBGC018006|AF133701
S_salar_GBGC181806|U12143
S_salar_BCF48207|BCF06061
S_salar_BCF48907|BCF06073
O_mykiss_BCF43607|BCF00331
O_mykiss_GBGC149306|NC_001717
O_mykiss_BCF43707|BCF00332
O_mykiss_TZFPA15407|NEOCAL070007
12Oncorhynchus mykiss Salmonidae
21Lophius piscatorius Lophiidae
L_piscatorius_gi|196168825|gb|EU683990.1
L_piscatorius_gi|196168827|gb|EU683991.1
22Solea solea Soleidae
S_solea_GBGC725309|EU513748
S_solea_GBGC725409|EU513747
S_solea_GBGC725509|EU513746
S_solea_GBGC725609|EU513745
23Epinephelus malabaricus Serranidae
E_areolatus_FOA63904|BWA639
E_areolatus_FOA64104|BWA641
E_areolatus_FOA64004|BWA640
E_areolatus_FOA64204|BWA642
E_malabaricus_FOA64504|BWA645
E_malabaricus_GBGC481808|EU204616
27Lates niloticus Latidae
9Lates niloticus Latidae
L_niloticus_FOA46704|BWA467
L_niloticus_FOA46804|BWA468
L_niloticus_FOA46904|BWA469
L_niloticus_FOA47004|BWA470
30Thunnus albacares Scombridae
31Thunnus albacares Scombridae
35Thunnus albacares Scombridae
36Thunnus albacares Scombridae
T_obesus_GBGC334407|DQ835867
33Engraulis encrasicolus Engraulidae
T_albacares_FOA87004|BWA870
T_albacares_FOA87104|BWA871
T_albacares_FOA87304|BWA873
T_albacares_GBGC426408|EU392206
T_obesus_FOA87904|BWA879
T_obesus_FOA88004|BWA880
T_obesus_FOA88104|BWA881
T_atlanticus_FOA95005|BWA1162
T_atlanticus_FOA95205|BWA1164
T_atlanticus_FOA95305|BWA1165
T_atlanticus_FOA95405|BWA1166
E_encrasicolus_GBGC416808|AM911181
E_encrasicolus_GBGC418308|AM911166
E_encrasicolus_GBGC416908|AM911180
C_harengus_GBGC343007|NC_009577
C_harengus_GBGC353207|AP009133
C_harengus_GBGC417308|AM911176
15Clupea harengus Clupeidae
P_fluviatilis_FOAC53005|BWA1529
29Pangasius hypophthalmus Pangasidae
50Pangasius hypophthalmus Pangasidae
P_hypophthalmus_FOAD21805|BW1778
16Prionace glauca Carcharhinidae
P_glauca_GBGC413408|EU400175
P_glauca_FCFMT09207|MCFS07002
P_glauca_FOA07704|BWA077
53Isurus oxyrhincus Lamnidae
I_oxyrinchus_GBGC549608|EU398892
I_oxyrinchus_GBGC549708|EU398891
I_oxyrinchus_GBGC549808|EU398890
I_oxyrinchus_GBGC549908|EU398889
52Dosidicus gigas Ommastrephidae
D_gigas_GBCPH41307|EU068697
D_gigas_GBCPH77709|NC_009734
D_gigas_GBCPH80109|FJ153075
D_gigas_GBCPH80209|FJ153074
25 Octopus vulgaris Octopodidae
O_vulgaris_GBCPH000106|AB052253
O_vulgaris_GBCPH70007|DQ683211
O_vulgaris_GBCPH70107|DQ683210
O_vulgaris_GBCPH70307|DQ683208
2Perna canaliculus Mytilidae
P_canaliculus_GBMLB172106|DQ343604
P_canaliculus_GBMLB172206|DQ343605
L_vannamei_GBCMD96307|DQ534543
L_vannamei_GBCMD96207|NC_009626

Figure 2. Neighbour-Joining tree constructed using the 104 mitochondrial cox1 sequences available on BOLD and GenBank databases for each species corresponding to our specimens along with the cox1 sequences obtained experimentally over all specimens. The numbers above the nodes represent bootstrap support after 1,000 replicates.

Three specific cases deserve more attention: in fact in the genera *Thunnus*, *Macruronus* and *Gadus* the genetic distinctiveness of single species was not well delineated. As a consequence, also our samples could not be correctly identified by means of the relative position of branches, affecting in this way the results and so the efficacy of the methodology. To further explore this aspect, a second NJ tree for each problematic genus was constructed with (data not shown) and without (**Figure 3)** the sequences corresponding to the specimens under study. Regarding the genus *Thunnus*, the obtained trees proved to be well resolved, except for the species belonging to the subgenus *Neothunnus* (*i.e*. *T. albacares*, *T. atlanticus* and *T. tonggol*), where the *T. albacares* sequences were polyphyletic, confirming previous findings that showed the cox1 gene as less variable than the mitochondrial DNA control regions (Viñas and Tudela, 2009). Furthermore, *T. alalunga* and *T. orientalis* could not be differentiated because these two species are genetically closely related and thus the chance to distinguish them from each other is influenced by the methodology and the sensibility of the markers used (Alvarado Bremer *et al*., 1997). Since the ability to resolve the species groups by means of a NJ tree is not limited by the number of sequences contained in the database, the lack of discrimination of our samples based on BOLD repository could be attributable to two causes: an initial misidentification of the sequences used as standard references or more likely a more complex phylogenetic history of the genus *Thunnus*, with frequent introgression events which can blur the results. Consequently, for this genus would be essential to select and adopt more than one genetic marker, mitochondrial and nuclear, with an appropriate mutation rate on the basis of previous studies (Viñas and Tudela, 2009).

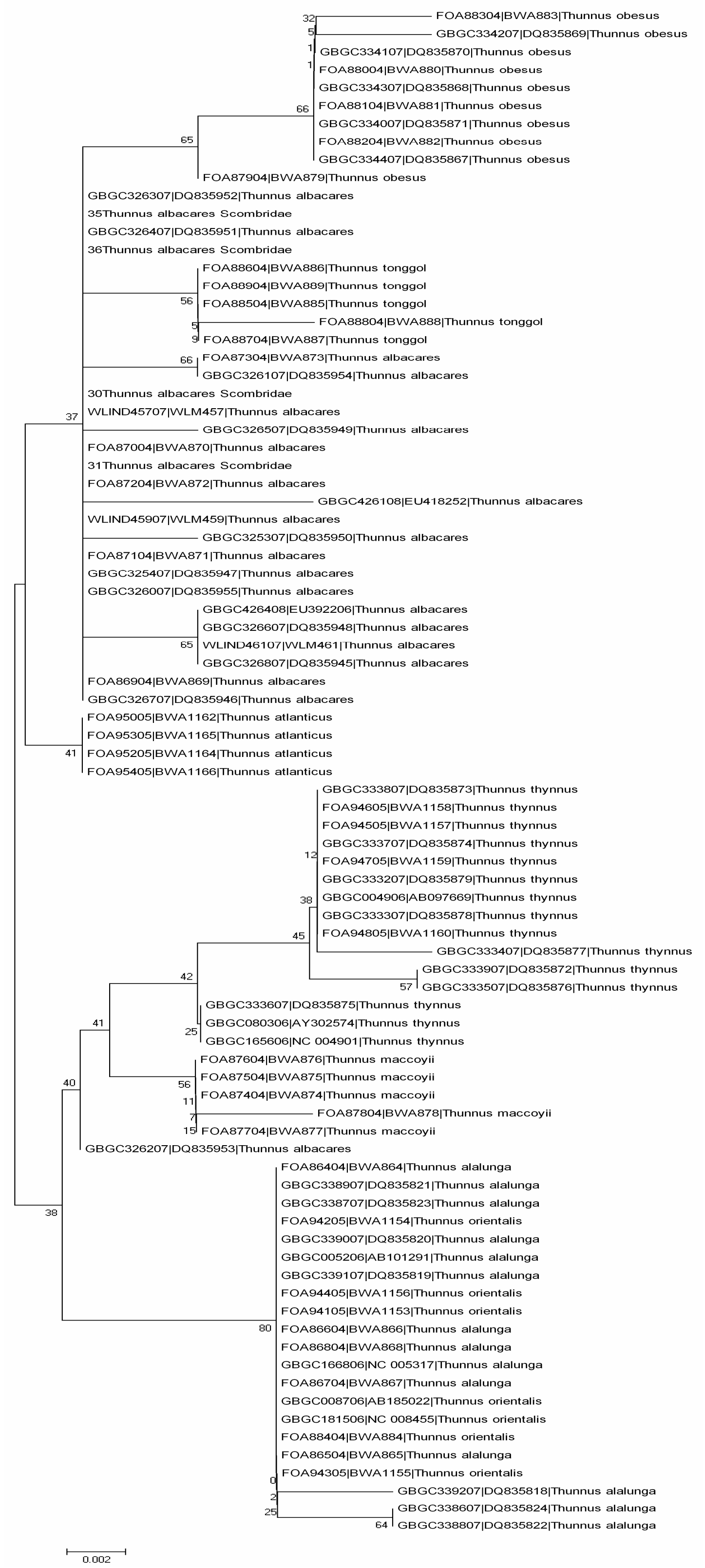
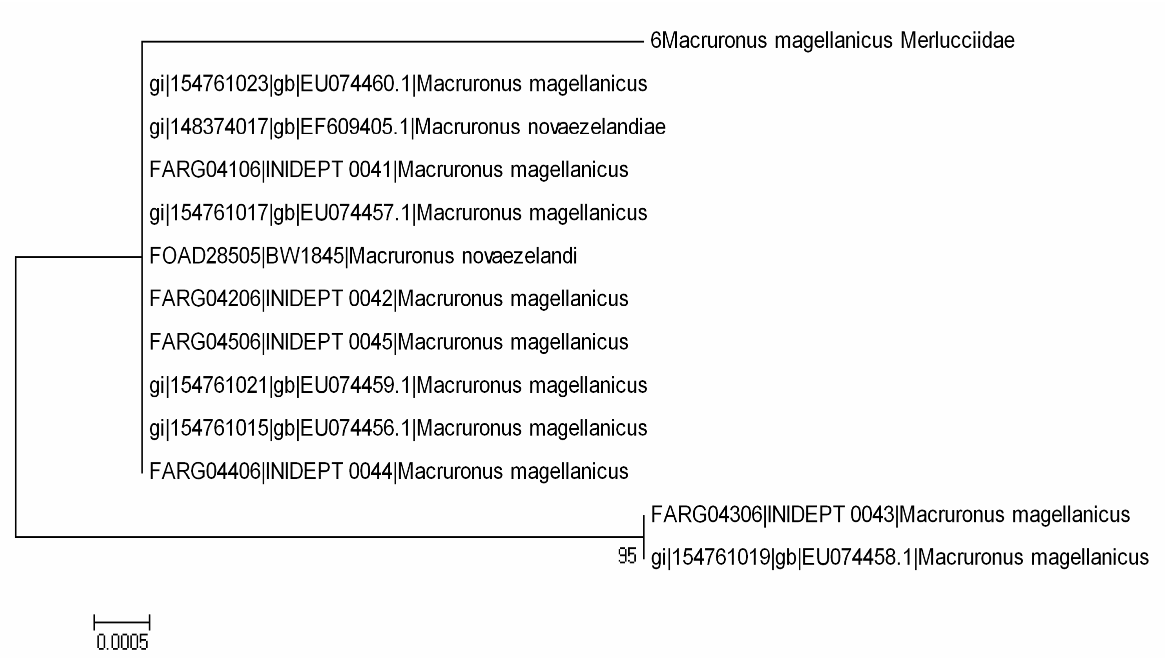

32
FOA88304|BWA883|Thunnus obesus
5
GBGC334207|DQ835869|Thunnus obesus
1
GBGC334107|DQ835870|Thunnus obesus
1
FOA88004|BWA880|Thunnus obesus
GBGC334307|DQ835868|Thunnus obesus
66
FOA88104|BWA881|Thunnus obesus
GBGC334007|DQ835871|Thunnus obesus
65
FOA88204|BWA882|Thunnus obesus
GBGC334407|DQ835867|Thunnus obesus
FOA87904|BWA879|Thunnus obesus
GBGC326307|DQ835952|Thunnus albacares
35Thunnus albacares Scombridae
GBGC326407|DQ835951|Thunnus albacares
36Thunnus albacares Scombridae
FOA88604|BWA886|Thunnus tonggol
FOA88904|BWA889|Thunnus tonggol
56
FOA88504|BWA885|Thunnus tonggol
5
FOA88804|BWA888|Thunnus tonggol
9
FOA88704|BWA887|Thunnus tonggol
66
FOA87304|BWA873|Thunnus albacares
GBGC326107|DQ835954|Thunnus albacares
30Thunnus albacares Scombridae
37
WLIND45707|WLM457|Thunnus albacares
GBGC326507|DQ835949|Thunnus albacares
FOA87004|BWA870|Thunnus albacares
31Thunnus albacares Scombridae
FOA87204|BWA872|Thunnus albacares
GBGC426108|EU418252|Thunnus albacares
WLIND45907|WLM459|Thunnus albacares
GBGC325307|DQ835950|Thunnus albacares
FOA87104|BWA871|Thunnus albacares
GBGC325407|DQ835947|Thunnus albacares
GBGC326007|DQ835955|Thunnus albacares
GBGC426408|EU392206|Thunnus albacares
GBGC326607|DQ835948|Thunnus albacares
65
WLIND46107|WLM461|Thunnus albacares
GBGC326807|DQ835945|Thunnus albacares
FOA86904|BWA869|Thunnus albacares
GBGC326707|DQ835946|Thunnus albacares
FOA95005|BWA1162|Thunnus atlanticus
FOA95305|BWA1165|Thunnus atlanticus
41
FOA95205|BWA1164|Thunnus atlanticus
FOA95405|BWA1166|Thunnus atlanticus
GBGC333807|DQ835873|Thunnus thynnus
FOA94605|BWA1158|Thunnus thynnus
FOA94505|BWA1157|Thunnus thynnus
GBGC333707|DQ835874|Thunnus thynnus
12
FOA94705|BWA1159|Thunnus thynnus
GBGC333207|DQ835879|Thunnus thynnus
38
GBGC004906|AB097669|Thunnus thynnus
GBGC333307|DQ835878|Thunnus thynnus
45
FOA94805|BWA1160|Thunnus thynnus
GBGC333407|DQ835877|Thunnus thynnus
42
GBGC333907|DQ835872|Thunnus thynnus
57
GBGC333507|DQ835876|Thunnus thynnus
GBGC333607|DQ835875|Thunnus thynnus
41
GBGC080306|AY302574|Thunnus thynnus
25
GBGC165606|NC 004901|Thunnus thynnus
FOA87604|BWA876|Thunnus maccoyii
40
56
FOA87504|BWA875|Thunnus maccoyii
11
FOA87404|BWA874|Thunnus maccoyii
7
FOA87804|BWA878|Thunnus maccoyii
15
FOA87704|BWA877|Thunnus maccoyii
GBGC326207|DQ835953|Thunnus albacares
FOA86404|BWA864|Thunnus alalunga
GBGC338907|DQ835821|Thunnus alalunga
38
GBGC338707|DQ835823|Thunnus alalunga
FOA94205|BWA1154|Thunnus orientalis
GBGC339007|DQ835820|Thunnus alalunga
GBGC005206|AB101291|Thunnus alalunga
GBGC339107|DQ835819|Thunnus alalunga
FOA94405|BWA1156|Thunnus orientalis
FOA94105|BWA1153|Thunnus orientalis
80
FOA86604|BWA866|Thunnus alalunga
FOA86804|BWA868|Thunnus alalunga
GBGC166806|NC 005317|Thunnus alalunga
FOA86704|BWA867|Thunnus alalunga
GBGC008706|AB185022|Thunnus orientalis
GBGC181506|NC 008455|Thunnus orientalis
FOA88404|BWA884|Thunnus orientalis
FOA86504|BWA865|Thunnus alalunga
FOA94305|BWA1155|Thunnus orientalis
0
GBGC339207|DQ835818|Thunnus alalunga
2
GBGC338607|DQ835824|Thunnus alalunga
25
64
GBGC338807|DQ835822|Thunnus alalunga
0.002

Figure 3. Neighbour-Joining tree constructed using the 12 mitochondrial cox1 sequences representing the seven species within the *Thunnus* genus retrieved from BOLD and GenBank databases along with the cox1 sequences corresponding to our specimens. The numbers above the nodes represent bootstrap support after 1,000 replicates.

About *Macruronus* species, the NJ tree showed only two clusters grouping entries independently from their species (**Figure 4)**. This tree topology does not surprise because this genus represents another example of taxonomic uncertainty. The division into two species, *M. novaezelandiae* and *M. magellanicus*, corroborated by morphometric analysis and different geographic distributions (Inada, 1990), was recently discounted (Balbontin *et al.*, 2004). The lack of morphological differences in the larval and adult stages, and the genetic divergence in the mitochondrial cob marker would lead to consider these two species as a case of synonymy (Olavarria *et al.*, 2006).



6Macruronus magellanicus Merlucciidae
gi|154761023|gb|EU074460.1|Macruronus magellanicus
gi|148374017|gb|EF609405.1|Macruronus novaezelandiae
FARG04106|INIDEPT 0041|Macruronus magellanicus
gi|154761017|gb|EU074457.1|Macruronus magellanicus
FOAD28505|BW1845|Macruronus novaezelandi
FARG04206|INIDEPT 0042|Macruronus magellanicus
FARG04506|INIDEPT 0045|Macruronus magellanicus
gi|154761021|gb|EU074459.1|Macruronus magellanicus
gi|154761015|gb|EU074456.1|Macruronus magellanicus
FARG04406|INIDEPT 0044|Macruronus magellanicus
FARG04306|INIDEPT 0043|Macruronus magellanicus
95
gi|154761019|gb|EU074458.1|Macruronus magellanicus
0.0005

Figure 4. Neighbour-Joining tree constructed using the 81 mitochondrial cox1 sequences representing the two species within the *Macruronus* genus retrieved from BOLD and GenBank databases along with the cox1 sequences corresponding to our specimens. The numbers above the nodes represent bootstrap support after 1,000 replicates.

Finally, the *Gadus* taxonomy is also problematic and distinct informative characters provided evidence toward different theories: some assert that three species (*G. morhua*, *G. ogac* and *G. macrocephalus*) can be distinguished within the genus *Gadus* on the basis of some morphological traits typical of their larval phase, but others do not agree. In fact, some phenotypical aspects and, most of all, identical mitochondrial cob sequences support the assertion that *G. ogac* and *G. macrocephalus* are synonym (**Figure 5)** (Carr *et al.*, 1999).

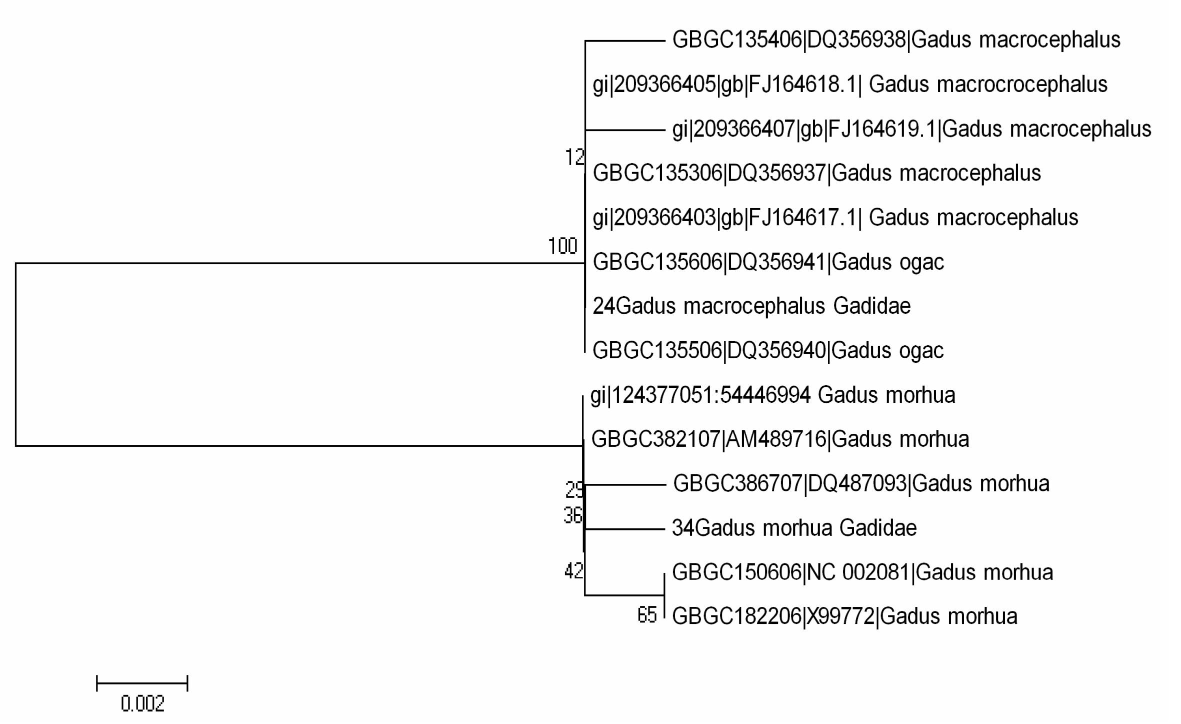

GBGC135406|DQ356938|Gadus macrocephalus
gi|209366405|gb|FJ164618.1| Gadus macrocrocephalus
gi|209366407|gb|FJ164619.1|Gadus macrocephalus
12
GBGC135306|DQ356937|Gadus macrocephalus
gi|209366403|gb|FJ164617.1| Gadus macrocephalus
100
GBGC135606|DQ356941|Gadus ogac
24Gadus macrocephalus Gadidae
GBGC135506|DQ356940|Gadus ogac
gi|124377051:54446994 Gadus morhua
GBGC382107|AM489716|Gadus morhua
29
GBGC386707|DQ487093|Gadus morhua
36
34Gadus morhua Gadidae
42
GBGC150606|NC 002081|Gadus morhua
65
GBGC182206|X99772|Gadus morhua
0.002

Figure 5. Neighbour-Joining tree constructed using the 12 mitochondrial cox1 sequences representing the three species within the *Gadus* genus retrieved from BOLD and GenBank databases along with the cox1 sequences corresponding to our specimens. The numbers above the nodes represent bootstrap support after 1,000 replicates.

This study shows that the molecular approach based on amplification of specific target regions is an efficient tool to ensure the correct detection of food composition and thus to control the label information. The technology of DNA barcoding based on the sequencing of specific mitochondrial DNA markers, is simple, robust and cost-effective, features which make it a valid tool for species authentication. Available data demonstrated that, when misidentification occurs on the basis of one or two genes, the cause may be generally attributable to either absent or erroneous reference sequence entry. This underlines the need to improve the amount of validated cox1 entries in the BOLD repository because a comprehensive DNA sequence library is essential for correct identification to species level (Ekrem *et al*., 2007). Particular cases are represented by the genus *Thunnus*, *Macruronus* and *Gadus.* In the first case, the species identification were reached using 16S and cob genes, but it was narrowed to the genus level on the basis of cox1. The reasons of this failure could be probably related to the use of the cox1 marker that shows inappropriate evolutionary rate for the eight *Thunnus* species and to its inability to detect the frequent introgressive hybridizations among tuna species. In the other two cases, the poor resolution of the tree due to the too low genetic divergence among species could be determined by the genetic identity of the species or by the necessity of a more variable marker. Nevertheless, when all the markers agree on the origin of the seafood product, the misidentification could be proof of species substitution. In this survey among the commercial products, we discovered five events of probable fraudulent substitutions. For instance, the specimen No. 28 was declared to be river perch that, in according to the Italian Ministerial Decree of the 14/01/2005, should be *Perca fluviatilis*. By means of molecular analyses, it was demonstrated that certainly it is not that species, but most likely *Paralichthys* spp., a flounder with lower market value than the perch. In this case, the mislabeling could likely be intentional and in fact the substitution of this pricey species with others less valuable is thought to be very common.

Conclusions

Up to now several different approaches have proved to be feasible for species identification, such as morphological inspection to molecular techniques based on protein analysis, but none of them can be universally applied. In fact during processing, the external features of commercial fish products used by classical assays are removed by slicing and the proteins, exploited by isoelectric focusing, liquid chromatography or immunoassays, undergo heat treatments that denature proteins and thus make them unavailable (Mackie *et al*., 1999). A different source of information is the DNA that, even if partially affected by heating, still represents a more stable molecule not so extensively compromised by high temperature process as occurs for proteins (Unseld *et al*., 1995). Therefore, the development of low cost assays focused on the DNA-based identification approach, that should be able to work independently of the degree of transformation which the food had underwent and without any variability in relation to the fish tissue considered, is getting a basic issue. Outdated gel-based sequencing methods, as PCR-RFLP or PCR-SSCP, the sequence of a target gene can be used to identify an organism even in highly processed foodstuff (Unseld *et al*., 1995). While these techniques required prior knowledge about what may be contained in the product, DNA barcoding does not need this information. Actually this approach, based on amplification, sequencing and interrogation of a sequence database, is not innovative, because more than ten years ago a similar procedure called FINS (Forensically Informative Nucleotide Sequencing) was developed (Bartlett and Davidson, 1992). But only with the introduction of DNA barcoding it became an international resource for molecular identification assays. The main drawback was that FINS exploited different markers for different taxonomic groups, while DNA barcoding offers the possibility to standardize the procedure using a universal region and thus to develop a unique library based on the cox1 sequence for all the metazoans on the Earth. The applications of this analytical method could be the rapid and sensitive monitoring of the meat of commercial interest species in food substrates in order to combat intentional or non-intentional fish substitutions (Logan *et al*., 2008). DNA barcoding in fact revealed feasible to determine the species identity of biological samples including highly processed food. 'Mini barcodes' for the standard cox1 gene were investigated and they proved to be effective for species identifications in specimens whose DNA is fragmented or in other situations where obtaining a full-length barcode is not feasible (Hajibabaei *et al*., 2006b). Our goal was to test the effectiveness of the cox1-based identification system and BOLD repository as a universal and sensitive tool able to recognize the species origin of a food component in frequent commercialized seafood items. The combining data strengthen the key role played by both effective universal primers and good quality DNA. Finally it was highlighted the necessity to develop reliable and comprehensive reference databases for successfully application of DNA barcoding for fish identification of commercial seafood products. So far even if GenBank database still remains the best web tool for forensic purposes, the BOLD ID proved to be enough rich to allow the correct recognition of almost all the specimens.

Acknowledgements

The authors gratefully acknowledge funding for this research from the Italian Ministry of University, Research, Science and Technology (Project: "Development of molecular diagnostic assays for the genetic traceability of agrifood products"). The authors wish to thank Silvio Collani for his help with the editing of nucleotide sequences. We want also to thank the Laboratories of Analytical Biology of the Smithsonian Institution, Washington DC (USA), for their economical support.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* 215*: 403-410.*

Alvarado Bremer JR, Naseri I, Ely B (1997). Orthodox and unorthodox phylogenetic relationships among tunas revealed by the nucleotide sequence analysis of the mitochondrial control region. *J Fish Biol* 50*: 540-554.*

Armstrong KF, Ball SL (2005). DNA barcodes for biosecurity: invasive species identification. *Phil Trans R Soc B* 360 (1462): *1813-1823.*

Baker CS, Palumbi SR (1994). Which whales are hunted? A molecular genetic approach to monitoring whaling. *Science* 265: *1538-1539.*

Baker CS, Lento GM, Cipriano F, Palumbi SR (2000). Predicted decline of protected whales based on molecular genetic monitoring of Japanese and Korean markets. *P Roy Soc B Biol Sci* 267 (1449): *1191-1199.*

Balbontin F, Uribe F, Bernal R, Braun M (2004). Descriptions of larvae of *Merluccius australis*, *Macruronus magellanicus*, and observations on a larva of *Micromesistius australis* from Southern Chile (Pisces: Gadiformes). *New Zeal J Mar Fresh* 38: *609-619.*

Bartlett SE, Davidson WS (1992). FINS (Forensically Informative Nucleotide Sequencing): a procedure for identifying the animal origin of biological specimens. *Biotechniques* 12(3): *408-411.*

Carr SM, Kivlichan DS, Pepin P, Crutcher DC (1999). Molecular systematics of gadid fishes: implications for the biogeographic origins of Pacific species. *Can J Zool* 77: *1-12.*

Chapela MJ, Sotelo CG, Perez-Martin RI, Pardo MA, Perez-Villareal B, Gilardi P, Riese J (2007). Comparison of DNA extraction methods from muscle of canned tuna for species identification. *Food Control* 18 (10): *1211-1215.*

Dawnay N, Ogden R, McEwing R, Carvalho GR, Thorpe RS (2007). Validation of the barcoding gene COI for use in forensic genetic species identification. *Forensic Sci Int* 173 (1): *1-6.*

Eddy SR (1998). Profile hidden Markov models. *Bioinformatics* 14: *755-763.*

Ekrem T, Willassen E, Stur E (2007). A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Mol Phylogenet Evo* 43 (2): *530-542.*

Espiñeira M, Vieites JM, Santaclara, FJ (2009). Development of a genetic method for the identification of salmon, trout, and bream in seafood products by means of PCR-RFLP and FINS methodologies. *Eur Food Res Technol* 229: *785-793.*

Folmer O, Black M, Hoeth W, Lutz R, Vrijenhoek R (1994). DNA primers for amplification of mithocondrial cytochrome c oxydase subunit I from diverse metazoan invertebrates. *Mol Mar Biol Biotech* 3 (5): *294-299.*

Forster P (2003). To err is human. *Ann Hum Genet* 67: *2-4.*

Guggiari M, Peck R (2008). The bacterivorous ciliate Cyclidium glaucoma isolated from a sewage treatment plant: molecular and cytological descriptions for barcoding. *Eur J Protistol* 44 (3): *168-180.*

Hajibabaei M, Janzen DH, Burns J, Hallwachs W, Hebert PDN (2006a). DNA barcodes distinguish species of tropical Lepidoptera. *PNAS 103* (4): *968-971.*

Hajibabaei M, Smith MA, Janzen DH, Rodriguez JJ, Whitfield JB, Hebert PDN (2006b). A minimalist barcode can identify a specimens whose DNA is degraded. *Mol Ecol Notes* 6(4): *959-964.*

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003). Biological identifications through DNA barcodes. *P Roy Soc B-Biol Sci* 270 (1512): *313-321.*

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004a). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *PNAS* 101 (41): *14812-14817.*

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004b). Identification of birds through DNA barcodes. *PloS Biol* 2 (10): *1657-1663*.

Hogg ID, Hebert PDN (2004). Biological identification of springtails (Hexapoda: Collembola) from the Canadian Arctic, using mitochondrial DNA barcodes. *Can J Zool* 82 (5): *749-754.*

Hsieh Y-HP (1996). Species substitution of restaurant fish entrees. *J Food Quality* 21 (1): *1-11.*

Hsieh YH, Shiu YC, Cheng CA, Chen SK, Hwang DF (2002). Identification of toxin and fish species in cooked fish liver implicated in food poisoning. *J Food Sci* 67 (3): *948-952.*

Hubert N, Hanner R, Holm E, Mandrak NE, Taylor E, Burridge M, Watkinson D, Dumont P, Curry A, Bentzen P, Zhang J, April J, Bernatchez L (2008). Identifying Canadian freshwater fishes through DNA barcodes. *PLoS ONE* 3 (6): *e2490.*

Inada T (1990). Family Merluccidae. In: Cohen D. M., Inada T., Iwamoto T., Scialabba N., *An annotated and illustrated catalogue of cods, hakes, grenadiers and other gadiform fishes known to dat*, (ed. Gadiform fishes of the world (Order Gadiformes)) (pp. 319-346). Rome, Italy: FAO Fisheries Synopsis 125.

Ivanova NV, Zemlak TS, Hanner RH, Hebert PDN (2007). Universal primer cocktails for fish DNA barcoding. *Mol Ecol Notes* 7 (4): *544-548.*

Kumar S, Dudley J, Nei M, Tamura K (2008) MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief in Bioinform* 9: 299-306.

Lin WF, Shiau CY, Hwang DF (2005). Identification of four *Thunnus* tuna species using mitochondrial cytochrome b gene sequence and PCR-RFLP analysis. *J Food Drug Anal* 13 (4): *382-387.*

Lin WF, Hwang DF (2008). Application of species-specific PCR for the identification of dried bonito product (Katsuobushi). *Food Chem* 106 (1): *390-396.*

Logan CA, Alter SE, Haupt AJ, Tomalty K, Palumbi SR (2008). An impediment to consumer choice: overfished species are sold as Pacific red snapper. *Biol Conserv* 141 (6): *1591-1599.*

Mackie IM, Pryde SE, Gonzalez-Sotelo C, Medina I, Perez-Martin R, Quinteiro J, Rey-Mendez M, Rehbein H (1999). Challenges in the identification of species of canned fish. *Trends Food Sci Tech* 10 (1): *9-14.*

Markmann M, Tauz D (2005). Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. *Phil Trans R Soc B* 360 (1462): *1917-1924.*

Marko PB, Lee SC, Rice AM, Gramling JM, Fitzhenry TM, McAlister JS, Harper GR, Moran AL (2004). Mislabelling of a depleted reef fish. *Nature* 430 (6997): *309-310.*

Martinez I, Danielsdottir AK (2000). Identification of marine mammal species in food products. *J Sci Food Agr* 80 (4): *527-533.*

Meixner MJ, Luter C, Eckert C, Itskovich V, Janussen D, Rintelen T, Bohne AV, Meixner JM, Hess WR (2007). Phylogenetic analysis of freshwater sponges provide evidence for endemism and radiation in ancient lakes. *Mol Phylogenet Evol* 45 (3): *875-886.*

Meyer CP, Paulay G (2005). DNA barcoding: error rates based on comprehensive sampling. *PLoS Biol* 3 (12): *e422*.

Montiel-Sosa JF, Ruiz-Pesini E, Montoya J, Roncales P, Lopez-Perez MJ, Perez-Martos A (2000). Direct and highly species-specific detection of pork meat and fat in meat products by PCR amplification of mitochondrial DNA. *J Agr Food Chem* 48 (7): *2829-2832*.

Olavarria C, Balbontin F, Bernal R, Baker CS (2006). Lack of divergence in the mitochondrial cytochrome b gene between *Macruronus* species (Pisces: Merlucciidae) in the Southern Hemisphere. *New Zeal J Mar Fresh* 40: *299-304.*

Palumbi SR (1996). Nucleic acids II: the polymerase chain reaction. In: D. M. Hillis, C. Mortiz B. K. Mable, Molecular Systematics, (2nd ed.) (*pp. 205-248*). Sunderland, MA: Sinauer.

Palumbi SR, Martin A, Romano S, McMillan WO, Stice L, Grabowski G (1991). The Simple Fool’s Guide to PCR. V. 2.0. Special Publication University of Hawaii, Department of Zoology Kewalo Marine Laboratory, Honolulu, HI.

Park MH, Sim CJ, Baek J, Min GS (2007). Identification of genes suitable for DNA barcoding of morphologically indistinguishable Korean halichondriidae sponges. *Mol Cells* 23 (2): *220-227.*

Pegg GG, Sinclair B, Briskey L, Aspden WJ (2006). MtDNA barcode identification of fish larvae in the southern Great Barrier Reef, Australia. *Sci Mar* 70 (S2): *7-12.*

Pepe T, Trotta M, Di Marco I, Anastasio A, Bautista JM, Cortesi ML (2007). Fish species identification in surimi-based products. *J Agr Food Chem* 55 (9): *3681-3685.*

Rach J, Desalle R, Sarkar IN, Schierwater B, Hadrys H (2008). Character-based DNA barcoding allows discrimination of genera, species and populations in *Odonata*. *Proc R Soc B* 275 (1632): *237-247.*

Rasmussen RS, Morrissey MT (2008). DNA-based methods for the identification of commercial fish and seafood species. *Compr Rev Food Sci F* 7 (3): *280-295.*

Ratnasingham, S. Hebert PDN (2007). BOLD: The Barcode of Life Data System. *Mol Ecol Notes* 7 (3): *355-364.*

Rehbein H, Etienne M, Jerome M, Hattula T, Knudsen LB, Jessen F, Luten JB, Bouquet W, Mackie I, Ritchie AH, Martin R, Mendes R (1995). Influence of variation in methodology on the reliability of the isoelectric focusing method of fish species identification. *Food Chem* 52 (2): *193-197.*

Ronning SB, Rudi K, Berdal KG, Holst-Jensen A (2005). Differentiation of important and closely related cereal plant species (Poaceae) in food by hybridization to an oligonucleotide array. *J Agr Food Chem* 53: *8874-8880.*

Saccone C, De Carla G, Gissi C, Pesole G, Reynes A (1999). Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene* 238 (1): *195-210.*

Sass C, Little DP, Stevenson DW, Specht CD (2007). DNA barcoding in the Cycadales: testing the potential of proposed barcoding markers for species identification of Cycads. *PLoS ONE* 2 (11): *e1154.*

Shivji M, Clarke S, Pank M, Natanson L, Kohler N, Stanhope M (2002). Genetic identification of pelagic shark body parts for conservation and trade monitoring. *Conserv Biol* 16 (4): *1036-1047.*

Simmons RB, Weller SJ (2001). Utility and evolution of cytochrome *b* in insects. *Mol Phylogenet Evol* 20 (2): *196-210.*

Tanabe S, Miyauchi E, Muneshige A, Mio K, Sato C, Sato M (2007). PCR method of detecting pork in foods for verifying allergen labelling and for identifying hidden pork ingredients in processed foods. *Biosci Biotech Bioch* 71 (7): *1663-1667.*

Teletchea F, Maudet C, Hanni C (2005). Food and forensic molecular identification: update and challenges. *Trends in Biotechnol* 23 (7): *359-366.*

Trotta M, Schonhuth S, Pepe T, Cortesi ML, Puyet A, Bautista JM (2005). Multiplex PCR methods for use in real-time PCR for identification of fish fillets from grouper (*Epinephelus* and *Mycteroperca* species) and common substitute species. *J Agr Food Chem* 53 (6): *2039-2045.*

Unseld M, Beyermann B, Brandt P, Hiesel R (1995). Identification of the species origin of highly processed meat products by mitochondrial DNA sequences. *PCR Methods Appl* 4: *241-243.*

Vences M, Thoma M, Bonett RM, Vieites DR (2005). Deciphering amphibian diversity through DNA barcoding: chances and challenges. *Phil Trans R Soc B* 360 (1462): *1859-1868.*

Viñas J, Tudela S (2009). A validated methodology for genetic identification of tuna species (Genus *Thunnus*). *PLoS ONE* 4 (10): *e7606.*

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005). DNA barcoding Australia's fish species. *Phil Trans R Soc B* 360 (1462): *1847-1857.*

Willows-Munro S, Robinson TJ, Matthee CA (2005). Utility of nuclear DNA intron markers at lower taxonomic levels: phylogenetic resolution among the nine *Tragelaphus* spp. *Mol Phylogenet Evol* 35 (3): *624-636.*

Wong EH-K, Hanner RH (2008). DNA barcoding detects market substitution in North-American seafood. *Food Res Int* 41 (8): *828-837.*

Yoo HS, Eah JY, Kim JS, Min MS, Paek WK, Lee H, Kim CB (2006). DNA barcoding Korean birds. *Mol Cells* 22*: 323-327.*

Zeng J S, De Hoog GS (2008). *Exophiala spinifera* and its allies: diagnostics from morphology to DNA barcoding. *Medical Mycology* 46 (3): *193-208.*

Web sources

http://www.barcodinglife.org/views/login.php

http://www.catalogueoflife.org/search.php

http://www.dnabarcoding.ca

http://www.dnabarcoding.ca/primer/Index.html

http://www.fda.gov/Food/FoodSafety/Product-SpecificInformation/Seafood/RegulatoryFishEncyclopediaRFE/ucm071528.htm

http://www.fishbase.org/search.php

http://www.fishbol.org/index.php

http://www.fishtrace.org/

http://www.ismea.it/flex/AppData/Redational/Normative/20051019000100040.pdf

Appendix 1. List of accession numbers of the sequences retrieved from BOLD and GenBank databases and used for the construction of the Neighbour-Joining tree.

GBGC426408|EU392206;GBGC326507|DQ835949;GBGC326807|DQ835945;GBGC326207|DQ835953;GBGC426108|EU418252;GBGC326707|DQ835946;GBGC326607|DQ835948;GBGC326407|DQ835951;GBGC326307|DQ835952;GBGC326107|DQ835954;GBGC326007|DQ835955;GBGC325407|DQ835947;GBGC325307|DQ835950;WLIND46107|WLM461;FOA87204|BWA872;WLIND45907|WLM459;FOA87104|BWA871;FOA87004|BWA870;WLIND45707|WLM457;FOA86904|BWA869;FOA95405|BWA1166;FOA95305|BWA1165;FOA95205|BWA1164;FOA95005|BWA1162;FOA88104|BWA881;FOA88004|BWA880;FOA87904|BWA879;GBGC334407|DQ835867;FOA88304|BWA883;GBGC334107|DQ835870;GBGC334207|DQ835869;FOA88204|BWA882;GBGC334307|DQ835868;GBGC334007|DQ835871;FOA88804|BWA888;FOA88604|BWA886;FOA88904|BWA889;FOA88504|BWA885;FOA88704|BWA887;GBGC333407|DQ835877;GBGC333907|DQ835872;GBGC333507|DQ835876;GBGC333807|DQ835873;GBGC333607|DQ835875;GBGC080306|AY302574;GBGC165606|NC_004901;GBGC333207|DQ835879;GBGC333307|DQ835878;GBGC333707|DQ835874;FOA94805|BWA1160;FOA94705|BWA1159;FOA94605|BWA1158;FOA94505|BWA1157;GBGC004906|AB097669;GBGC338607|DQ835824;GBGC338807|DQ835822;GBGC339207|DQ835818;GBGC338707|DQ835823;GBGC339107|DQ835819;FOA86404|BWA864;GBGC339007|DQ835820;GBGC166806|NC_005317;GBGC005206|AB101291;GBGC338907|DQ835821;FOA86704|BWA867;FOA86504|BWA865;FOA86804|BWA868;FOA86604|BWA866;FOA88404|BWA884;FOA94405|BWA1156;FOA94205|BWA1154;GBGC181506|NC_008455;GBGC008706|AB185022;FOA94305|BWA1155FOA94105|BWA1153;FOA87804|BWA878;FOA87504|BWA875;FOA87604|BWA876;FOA87404|BWA874;FOA87704|BWA877;FOA95005|BWA1162;FOA95205|BWA1164;FOA95405|BWA1166;FOA95305|BWA1165;FOA89004|BWA890;FOA89404|BWA894;FOA89304|BWA893;FOA89204|BWA892;FOAD31705|BW1877;gi|166898013|gEU271893.1;TZFPB03405|TZ05FROSTI034;TZFPB03305|TZ05FROSTI033;TZFPB03205|TZ05FROSTI032;TZFPB03005|TZ05FROSTI030;FARG04606|INIDEPT0046;FARG25006|INIDEPT0250;FARG24906|INIDEPT0249;FARG24806|INIDEPT0248;gi|154761023|gb|EU074460.1;FOAD28505|BW1845;FARG04506|INIDEPT0045;|FARG04406|INIDEPT0044;|FARG04306|INIDEPT0043;FARG04206|INIDEPT0042;gi|154761019|gb|EU074458.1;gi|154761021|gb|EU074459.1;FARG04106|INIDEPT0041;gi|154761017|gb|EU074457.1;gi|154761015|gb|EU074456.1;gi|148374017|gb|EF609405.1;BCF43707|BCF00332;BCF43607|BCF00331;GBGC149306|NC_001717;TZFPA15407|NEOCAL070007;GBGC018006|AF133701;BCF48207|BCF06061;GBGC181806|U12143;BCF48907|BCF06073;FOA47004|BWA470;FOA46904|BWA469;FOA46804|BWA468;FOA46704|BWA467;GBGC382107|AM489716;GBGC386707|DQ487093;GBGC150606|NC_002081;GBGC182206|X99772;GBGC135406|DQ356938;gi|209366407|gb|FJ164619.1;gi|209366403|gb|FJ164617.1;GBGC135306|DQ356937;gi|209366405|gb|FJ164618.1;GBGC135606|DQ356941;GBGC135506|DQ356940;gi|124377051:54446994;GBGC732109|EU513680;GBGC732009|EU513681;GBGC731909|EU513682;FOAD21805|BW1778 ;FARG25306|INIDEPT0253;FARG25206|INIDEPT0252;FARG06006|INIDEPT0060;GBGC417308|AM911176;GBGC343007|NC_009577;GBGC353207|AP009133;FCFMT09207|MCFS07002;GBGC413408|EU400175;FOA07704|BWA077;GBGC725609|EU513745;GBGC725509|EU513746;GBGC725409|EU513747;GBGC725309|EU513748;GBGC481808|EU204616;FOA64504|BWA645;FOAC53005|BWA1529;GBGC418308|AM91116;GBGC416908|AM911180;GBGC416808|AM911181;GBGC549908|EU398889;GBGC549808|EU398890;GBGC549708|EU398891;GBGC549608|EU398892;FARG35907|INIDEPT0358;FARG35807|INIDEPT0357;FARG22106|INIDEPT0221;FARG22006|INIDEPT0220;FOA64204|BWA642;FOA64104|BWA641;FOA64004|BWA640;FOA63904|BWA639;FARG43508|INIDEPT|0434;FARG43908|INIDEPT0438;FARG43808|INIDEPT0437;gi|196168825|gb|EU683990.1;gi|196168827|gb|EU683991.1;GBCPH77709|NC_009734;GBCPH41307|EU068697;GBCPH80109|FJ153075;GBCPH80209|FJ153074;GBCPH000106|AB052253;GBCPH70007|DQ683211;GBCPH70107|DQ683210;GBCPH70307|DQ683208;GBCMD96307|DQ534543;|GBCMD96207|NC_009626;|GBMLB172106|DQ343604;GBMLB172206|DQ343605.

Chapter 3

Use of DNA barcoding in crop plants: *P. vulgaris* L.

Biodiversity studies in *Phaseolus* spp. by DNA barcoding

Silvia Nicolè*, David L. Erickson†§, Daria Ambrosi*, Elisa Bellucci‡, Margherita Lucchin*, Roberto Papa‡, W. John Kress†§ and Gianni Barcaccia*

*Department of Environmental Agronomy and Crop Science, Università degli Studi di Padova, Viale Università 16 – Campus of Agripolis, 35020 Legnaro, Padova (Italy); †Department of Botany and §Laboratories of Analytical Biology – National Museum of Natural History, Smithsonian Institution, P.O. Box 37012, Washington, DC 20013-7012 (USA); ‡Department of Environmental Sciences and Crop Production, Università Politecnica delle Marche, Ancona, Via Brecce Bianche – Monte d’Ago, 60131 Ancona (Italy).

Keywords: *Phaseolus* spp., cpDNA, ITS, DNA barcoding, varietal groups, SNPs

Abstract

DNA barcoding is a new genomic technique suitable to identify organisms by comparing a sequence of a standardized gene region from an unknown specimen with a comprehensive database of orthologous sequences from species of established identity. Our research aims to test the potential of DNA barcoding as an implemented system for genetic diversity and genetic traceability studies not only of species but also cultivated varieties. The technique was applied to several pure lines of *Phaseolus vulgaris* belonging to wild, domesticated and cultivated common beans, along with several accessions of *P. coccineus*, *P. lunatus* and *Vigna unguiculata*. A multilocus approach was exploited using three chloroplast genic regions (*rbcL, trnL and matK*) and four intergenic spacers (*rpoB-trnC, atpBrbcL, trnT-trnL* and *psbA-trnH*) together with the nuclear ITS1 and ITS2. The main goals were to identify the markers and SNPs that show the best discriminant power at variety level in common bean germplasm, to test two distinct methods (*i.e.* tree-based versus character-based) for biodiversity analysis and traceability assays, and to evaluate the overall utility of plastid DNA barcodes for reconstructing the origin of modern Italian varieties. Our results indicated that the NJ method is a very powerful approach for comparing genetic diversity in plant species, but it is realtive uninformative for the genetic traceability of plant varieties. *Vice versa*, the character-based method was able to identify several distinct haplotypes over all target regions corresponding to Mesoamerican or Andean accessions, with Italian accessions clustered with one or the other gene pool.

Introduction

The genomic advances of the last decade have provided the technological tools for developing a universal DNA-enhanced system of taxonomy suitable to face the current ‘biodiversity crisis’ which requires innovative and informative methods (Tautz *et al.*, 2003). DNA barcoding was proposed as a cost-effective technology able to contribute to the study of biodiversity, which up to now relied predominantly on morphology in the Linnaean classification system (Hebert *et al.*, 2003a). The DNA-based method is fast and not limited by taxonomic impediments, such as missing morphological features of a particular life stage, like eggs and juvenile forms (Wells and Stevens, 2008) or body parts (Wong and Hanner, 2008), or because of homoplasy of some characters (Vences *et al.*, 2005). Although the application of DNA fingerprinting as identification tool is not a new idea, DNA barcoding has earned remarkable success due to the standardization of the procedure by means of the use of a universal barcode sequence across a wide range of organisms (Hebert *et al.*, 2003b). The ambitious idea of using a short piece of DNA to distinguish every species in the world is already a powerful tool in the animal kingdom, but plant biologists have been slower in adapting a universal gene region as a barcode (Hollingsworth *et al.*, 2009). In contrast to the rapid progress in applying barcodes to animals (Ward *et al.*,, 2005), the application of DNA barcoding to the plant kingdom has been constrained by the difficulty of finding an analogous region to animal COI gene. However, recently, the CBOL Plant Working Group (Hollingsworth *et al.*, 2009) has recommended the combination of *rbcL* + *matK* as the plant barcode. This core 2-locus DNA barcoding approach has been proposed as a universal framework for the routine use of DNA sequence data to identify specimens and contribute toward the discovery of overlooked species of land plants. In the same publication a minority position of the Plant Working Group supported the inclusion of the *trnH-psbA* intergeneic spacer as a necessary part of the plant barcode following some earlier publications that outlined some practical difficulties related to the acquisition of *matK* sequences (Kress and Erickson, 2007; Fazekas *et al.*, 2008). The combination of the *rbcL* gene with the *trnH-psbA* intergenic spacer, a more rapidly evolving region than *rbcL* and *matK*, seemed to be a valid alternative to a simple two-locus model: the former distinguishes distantly related plants and the latter to recognize closely related sister species or species groups that have only recently diverged (Kress and Erickson, 2007). Finally, even if the organellar DNA sequences are considered as the main source of information for a barcoding system, it is recognized that in cases of hybridization supplemental analyses with one or more nuclear genes may also be required. Nuclear genes, such as ITS, the ribosomal internal transcribed spacers that is frequently used for phylogenetic analyses, or single-copy nuclear regions have already been considered by some (Cowan *et al.*, 2006) (see also http://www.rbgkew.org.uk/barcoding).

Several DNA fingerprinting and genotyping assays based on molecular markers, such as RFLPs and SNPs, have been developed in the past and are still used in plant genetics and breeding (Mohler and Schwarz, 2008). DNA barcoding could represent an additional system to identify not only species, but also crop varieties and germplasm resources in order to assess the distinctiveness of genotypes as well as the relatedness among genotypes (Pallottini *et al.,* 2004). Testing the potentials of DNA barcoding to distinguish plant varieties of agri-food interest would be extremely valuable for both breeders and farmers. While the ability of DNA barcoding for species identification has been widely investigated, the within-species discrimination of single varietal genotypes, such as clones, pure lines and hybrids, has been poorly investigated and few studies have focused on the use of DNA barcoding as a sufficiently informative technique to be exploited for the genetic identification of closely related crop varieties (Newmaster *et al.*, 2007; Tsai *et al.*, 2008).

Our work focuses on the application of DNA barcoding in cultivated bean germplasm as a new tool for identification and to assess genetic relationships among *Phaseolus* species and varieties *of P. vulgaris*. *Phaseolus* is a genus in the family Fabaceae, the third largest family of flowering plants (Gepts *et al.*, 2005), and is an example of multiple domestications of distinct but related species and multiple populations within the same species, for example as found in *P. vulgaris* and *P. lunatus*. The original natural distribution of this species consists of a fragmented area throughout the Central and Southern American regions, followed by its introduction throughout Europe and Africa after post-Columbian discovery. On the basis of the available data, at least two primary centres of origin have been recognized, one relatively heterogeneous in the Andes (Colombia, Ecuador, Peru, Bolivia, Chile and Argentina) and the other more homogeneous in MesoAmerica (mainly Mexico, Guatemala, Honduras, El Salvador, Nicaragua and Costa Rica), called the Andean and Mesoamerican gene pools, respectively (Gepts *et al*., 1986; Beebe *et al*., 2000; Beebe *et al*., 2001; Papa and Gepts, 2003; Chacon *et al.*, 2005; Papa *et al*., 2006).

In this paper we present our results on the use of DNA barcoding in several pure lines of wild, domesticated, and cultivated common beans, for both coding and non-coding regions from the chloroplast and nuclear genomes. In particular our objectives were: (1) to test how different markers perform as DNA barcodes, mainly below the level of species (*i.e*. Andean and Mesoamerican gene pools); (2) to investigate the genetic differentiation among varieties and how barcode data can be used to reconstruct the origin of modern Italian varieties, and (3) to evaluate the effectiveness of different methods (*i.e*. tree-based versus character-based).

Materials and Methods

Germplasm sampling of *Phaseolus*

A total of 33 varieties of *Phaseolus vulgaris* were selected as representative of Mesoamerican and Andean gene pools on the basis of morphological seed traits, plant descriptors and molecular markers (Rossi *et al.,* 2009). Eight wild and nine domesticated accessions from Central America (Mexico, Costa Rica, Honduras and El Salvador) and ten wild and six domesticated accessions from South America (Argentina, Bolivia, Brazil, Colombia and Peru), were employed. These accessions were obtained from the germplasm banks held at the International Center for Tropical Agriculture (CIAT) and United States Department of Agriculture (USDA) (**Table 1**). Moreover, a total of 22 Italian cultivated accessions of uncertain origin, in terms of progenitor gene pool, were collected from available commercial varieties supplied by CRA, Research unit for Orticulture of Montanaso Lombardo. In addition to these three main sub-groups, two wild accessions from the *P. vulgaris* ancestral gene pool in Peru were included in the analysis. Furthermore, a subsampling of *P. coccineus*, *P. lunatus* and *Vigna unguiculata* accessions were used as reference standards and out-groups. The list of varieties and landraces along with information on their origin is reported in **Table 1**.

Genomic DNA extraction

Genomic DNA was isolated using the Nucleon PhytoPure DNA Extraction (Amersham Biosciences) kit from 0.5-1.0 g of powdered frozen young leaf tissue following instructions of the manufacturer. An additional step of purification with NaOAc was used to remove excess salts and then the DNA pellets were resuspended in 80-100 µl of TE 0.1 Buffer (Tris-HCl 100 mM, EDTA 0.1 mM pH 8). The final concentration of DNA was estimated by electrophoresis on 0.8% agarose/TAE gel and the quantification was conducted by comparison with 1 Kb plus DNA ladder (Invitrogen) of known concentration.

Table 1. List of 63 bean entries with the common name, accession number, origin area and voucher information.

Sample	Species	Accessions	Classification	Origin	Gene pool	Voucher #
PvF8wanc	*P. vulgaris*	G23585	wild-ancestral	South America (Peru)	Ancestral	i.p.
PvG8wanc	*P. vulgaris*	G23587	wild-ancestral	South America (Peru)	Ancestral	i.p.
PvH2mw	*P. vulgaris*	G23652	wild	Central America (Mexico)	Mesoamerican	i.p.
PvA3mw	*P. vulgaris*	G12979	wild	Central America (Mexico)	Mesoamerican	i.p.
PvC3mw	*P. vulgaris*	G23463	wild	Central America (Mexico)	Mesoamerican	i.p.
PvD3mw	*P. vulgaris*	G22837	wild	Central America (Mexico)	Mesoamerican	i.p.
PvB7mw	*P. vulgaris*	G12873	wild	Central America (Mexico)	Mesoamerican	3901-8
PvG7mw	*P. vulgaris*	G12922	wild	Central America (Mexico)	Mesoamerican	i.p.
PvB8mw	*P. vulgaris*	G11050	wild	Central America (Mexico)	Mesoamerican	i.p.
PvC8mw	*P. vulgaris*	G12949	wild	Central America (Mexico)	n.d.	i.p.
PvD8aw	*P. vulgaris*	G21113	wild	South America (Colombia)	Mesoamerican	i.p.
PvE6aw	*P. vulgaris*	G23445	wild	South America (Bolivia)	Andean	i.p.
PvF6aw	*P. vulgaris*	G23444	wild	South America (Bolivia)	Andean	i.p.
PvG6aw	*P. vulgaris*	W618821	wild	South America (Bolivia)	Andean	i.p.
PvH6aw	*P. vulgaris*	G23455	wild	South America (Peru)	Andean	i.p.
PvG3aw	*P. vulgaris*	G23420	wild	South America (Peru)	Andean	i.p.
PvB6aw	*P. vulgaris*	G19893	wild	South America (Argentina)	Andean	i.p.
PvC6aw	*P. vulgaris*	G19898	wild	South America (Argentina)	Andean	i.p.
PvD6aw	*P. vulgaris*	G21198	wild	South America (Argentina)	Andean	i.p.
PvH5aw	*P. vulgaris*	W617499	wild	South America (Argentina)	n.d.	i.p.
PvF7md	*P. vulgaris*	PI201349	domesticated	Central America (Mexico)	Mesoamerican	i.p.
PvG1md	*P. vulgaris*	PI165435	domesticated	Central America (Mexico)	Mesoamerican	3901-10
PvH1md	*P. vulgaris*	PI165440	domesticated	Central America (Mexico)	Mesoamerican	i.p.

PvA2md	*P. vulgaris*	PI309785	domesticated	Central America (Mexico)	Mesoamerican	i.p.
PvH4md	*P. vulgaris*	PI207370	domesticated	Central America (Mexico)	Andean	i.p.
PvE7md	*P. vulgaris*	PI309885	domesticated	Central America (Costa Rica)	Mesoamerican	i.p.
PvD1md	*P. vulgaris*	PI309831	domesticated	Central America (Costa Rica)	Mesoamerican	i.p.
PvF1md	*P. vulgaris*	PI310577	domesticated	Central America (Honduras)	Mesoamerican	i.p.
PvE1md	*P. vulgaris*	PI304110	domesticated	Central America (El Salvador)	n.d.	i.p.
PvC1ad	*P. vulgaris*	BAT93-1	domesticated	South America (Colombia)	Mesoamerican	i.p.
PvC2ad	*P. vulgaris*	BAT93-2	domesticated	South America (Colombia)	Mesoamerican	i.p.
PvH8ad	*P. vulgaris*	BAT881	domesticated	South America (Colombia)	n.d.	3901-11
PvB4ad	*P. vulgaris*	MIDAS	domesticated	South America (Argentina)	Andean	i.p.
PvD5ad	*P. vulgaris*	PI290992	domesticated	South America. (Peru)	Andean	3901-9
PvA7ad	*P. vulgaris*	JALOEEP558	domesticated	South America (Brasile)	Andean	3901-7
Pv1itc	*P. vulgaris*	Cannellino rosso	cultivated	Italy	n.d.	3901-16
Pv3itc	*P. vulgaris*	Montalbano	cultivated	Italy	n.d.	3901-18
Pv6itc	*P. vulgaris*	Munachedda nera	cultivated	Italy	n.d.	3901-19
Pv9itc	*P. vulgaris*	San Michele	cultivated	Italy	n.d.	i.p.
Pv10itc	*P. vulgaris*	Nasieddu viola	cultivated	Italy	n.d.	i.p.
Pv13itc	*P. vulgaris*	Maruchedda	cultivated	Italy	n.d.	i.p.
Pv14itc	*P. vulgaris*	Riso bianco	cultivated	Italy	n.d.	3901-20
Pv16itc	*P. vulgaris*	Cannellino	cultivated	Italy	n.d.	3901-21
Pv19itc	*P. vulgaris*	Verdolino	cultivated	Italy	n.d.	3901-22
Pv22itc	*P. vulgaris*	Blu Lake	cultivated	Italy	n.d.	3901-23
Pv23itc	*P. vulgaris*	Goldrush	cultivated	Italy	n.d.	3901-24
Pv24itc	*P. vulgaris*	Borlotto Clio	cultivated	Italy	n.d.	i.p.
Pv27itc	*P. vulgaris*	Lena	cultivated	Italy	n.d.	3901-25

Pv28itc	*P. vulgaris*	Giulia	cultivated	Italy	n.d.	3901-26
Pv29itc	*P. vulgaris*	Saluggia	cultivated	Italy	n.d.	3901-27
Pv31itc	*P. vulgaris*	Borlotto Lamon	cultivated	Italy	n.d.	3901-28
Pv32itc	*P. vulgaris*	Saluggia	cultivated	Italy	n.d.	3901-29
Pv33itc	*P. vulgaris*	Cannellini	cultivated	Italy	n.d.	3901-30
Pv34itc	*P. vulgaris*	Verdoni	cultivated	Italy	n.d.	3901-34
Pv35itc	*P. vulgaris*	S. Matteo	cultivated	Italy	n.d.	3901-31
Pv36itc	*P. vulgaris*	Zolferini Rovigotti	cultivated	Italy	n.d.	3901-32
Pv37itc	*P. vulgaris*	Neri Messicani	cultivated	Italy	n.d.	3901-33
PcA1mw	*P. coccineus*	PI417608	wild	Central America (Mexico)	n.d.	i.p.
Pc30itc	*P. coccineus*	Venere	cultivated	Italy	n.d.	i.p.
Pc39itc	*P. coccineus*	Spagna	cultivated	Italy	n.d.	i.p.
PlB1md	*P. lunatus*	PI310620	domesticated	Central America (Guatemala)	n.d.	i.p.
Pl38itc	*P. lunatus*	Lima	cultivated	Italy	n.d.	3901-2
Vu40itc	*V. unguiculata*	Fagiolino dall'occhio	cultivated	Italy	n.d.	3905-2

# Plants with flowers and pods are conserved in the herbarium of the Botanical Garden of the University of Padua (Italy).
i.p., Voucher attainment in progress.
n.d., not determined.

DNA barcode markers and PCR assays

Following a multi-locus approach (Chase *et al.*, 2005; Kress and Erickson, 2007; Newmster *et al.*, 2007), several regions were tested using a subset of bean samples in order to detect which markers could be the most informative at the intraspecific level. After this preliminary survey, only seven out of 12 chloroplast gene regions, both coding (*rbcL* and *matK*) and non-coding regions (*trnL* intron, *atpB-rbcL*, *trnH-psbA*, *trnT-trnL* and *rpoB-trnC* intergenic spacers), proved to be variable and informative, while the other regions were observed to be monomorphic and were not adopted for further analysis (*rpl32-trnL*, *ndhF-rpl32*, *trnD-trnT*, *trnS-trnG*, *rpoC1*) (data not shown). Furthermore the two internal transcribed spacers, ITS1 and ITS2, of the rDNA that separate the 5.8S ribosomal gene from 18S and 25S loci, were used to compare the utility of the nuclear genome with the chloroplast genome for resolving relationships at variety level. For three of the selected cpDNA barcode regions, *rbcL*, *trnL* and *atpB-rbcL*, specific primers were designed after the retrieval of the sequences from the NCBI databases for the Fabaceae family. After removal of redundant and unverified entries, serial local multiple sequence alignments were performed by Vector NT software. Specific primer pairs, ranging from 18 to 28-mer in length, were constructed in highly conserved short stretches (300-500 bp) flanking the most variable portions of each region using the PRIMER3 software. In the other cases, universal primers were adopted (**Table 2**). All PCR experiments were performed using a GeneAmp PCR System 9700 (Applied Biosystems). The temperature profile consisted of an initial step of 5 min at 95°C followed by 35 cycles of 30 sec at 95°C, 1.10 min at 56°C for all the markers, except for ITS1 and 2 and *rpoB-trnC* at 54°C, 1.20 min at 72°C, followed in turn by 7 min at 72°C and then held at 4°C. Only for matk marker modified PCR conditions were adopted: 40 cycles of 95°C for 30 sec, 56°C for 1 min and 72°C for 2 min, with initial denaturation 95°C for 5 min and final extension at 72°C for 7 min. The 25 µl PCR reaction volume included 1× PCR buffer (100 mM Tris-HCl pH 9.0, 15 mM $MgCl_2$ and 500 mM KCl), 0.2 mM dNTPs, 0.2 µM of each primer, 0.5 U of *Taq* DNA polymerase, 15 ng of genomic DNA as template and 1× Hi Specific Additive (Bioline) to facilitate the amplification. Sometimes faint double bands were recovered on gel indicating the presence of aspecific products, therefore a second PCR assay was performed using more stringent conditions, higher annealing temperatures and less cycle numbers.

Table 2. List of primers used for each chloroplast and nuclear marker with their nucleotide sequence, amplicon length and reference source.

Marker	**Amplicon length (bp)**				**Primer**	**Primer sequence (5'-3')**	**References**
	P. vulgaris	*P. coccineus*	*P. lunatus*	*V. uguiculata*			
rbcL gene	543	543	543	543	rbcL_F	GCAGCATTYCGAGTAASTCCYCA	This study
					rbcL_R	GAAACGYTCTCTCCAWCGCATAAA	This study
					rbcL 724R*	TCACATGTACCTGCAGTAGC	Lledò *et al*. (1998) mod.
matk gene	695	695	695	695	matK4La	CCTTCGATACTGGGTGAAAGAT	Wojciechowski *et al*. (2004)
					matK1932Ra	CCAGACCGGCTTACTAATGGG	Wojciechowski *et al*. (2004)
trnL intron	350	350	296	357	trnL_F	GGATAGGTGCAGAGACTCRATGGAAG	This study
					trnL_R	TGACATGTAGAATGGGACTCTATCTTTAT	This study
					5'trnLUAAF*	CGAAATCGGTAGACGCTACG	Taberlet *et al*. (1991)
					3'trnLUAAR*	GGGGATAGAGGGACTTGAAC	Taberlet *et al*. (1991)
atpB-rbcL IGS	329	325	326	331	atpB_F	GGTACTATTCAATCAATCCTCTTTAATTGT	This study
					atpB_R	ATGTAAATCCTAGATGTRAAAATAKGCAG	This study
					atpB_R2*	CGCAACCCAATCTTTGTTTC	This study
trnH-psbA IGS	365	365	365	369	psbA3'f	GTTATGCATGAACGTAATGCTC	Sang *et al*. (1997)
					trnHf	CGCATGGTGGATTCACAATCC	Tate and Simpson (2003)
rpoB-trnC IGS	1117	1117	1124	1136	rpoB_F	CKACAAAAYCCYTCRAATTG	Shaw *et al*. (2005)
					trnCGCAR	CACCCRGATTYGAACTGGGG	Shaw *et al*. (2005)
					rpoB_R3*	TTCTTTACAATCCCGAATGG	This study
trnT-trnL IGS	813	837	823	871	trnTUGU2F	CAAATGCGATGCTCTAACCT	Cronn *et al*. (2002)
					5'trnLUAAR	TCTACCGATTTCGCCATATC	Taberlet *et al*. (1991)
ITS1	373	382	355-364	314	ITS5	GGAAGTAAAAGTCGTAACAAGG	White *et al*. (1990)
					ITS2	GCTGCGTTCTTCATCGATGC	White *et al*. (1990)
ITS2	419	418	413	401	ITS3	GCATCGATGAAGAACGCAGC	White *et al*. (1990)
					ITS4	TCCTCCGCTTATTGATATGC	White *et al*. (1990)

*Primers used only for sequencing

The PCR-derived fragments were resolved on 2% agarose/TAE gels and visualized under UV light using ethidium bromide staining. Positive and negative controls were used as references. All amplification products were purified enzymatically by digestion with Exonuclease I and Shrimp alkaline phosphatase (Amersham) and then directly sequenced using forward and reverse primers according to the original Rhodamine terminator cycle sequencing kit (ABI PRISM Applied Biosystems). For some regions a second forward or reverse primer located upstream or downstream that used for all PCR experiments were eventually adopted for replicated sequencing reactions. Finally, in the sequencing mixture of *matK*, DMSO 4% of the total reaction volume was used to overcome some secondary structural problems of the sequence.

Tree-based analysis

The obtained sequences were visualized and manually edited by Sequencer 4.8 for minimizing the possible errors during the sequencing and removing gaps in the coding regions that could cause shifts in the ORF of *rbcL*.

Sequence similarity search was performed using GenBank BLASTn algorithm (http://www.ncbi.nlm.nih.gov/BLAST) against the nucleotide databases of NCBI to check the correspondence between the sequences of the obtained amplicons with the expected sequences. Separate data analyses for each sequence alone, for the combined chloroplast and nuclear data sets individually and together were carried out. Multiple sequence alignments were performed by SeAl v2.0a11 software and the inter- and intraspecific genetic divergences were calculated by means of MEGA 4.1 beta software (Tamura *et al.*,, 2007) according to the Kimura-2-Parameter distance model (Kimura, 1980). Based on the pairwise nucleotide sequence divergences, the Neighbour-Joining (NJ) was estimated and rooted using as outgroup the accessions from different species. A bootstrap statistical analysis (BS) was conducted to measure stability of the obtained branches using 1,000 resampling replicates. All positions containing gaps and missing data were eliminated from the dataset (complete deletion option). To assign each accession to the correct gene pool, the phenetic approach was based on the computation of the genetic distance to see whether the so-called ‘barcode gap’, a discontinuity between intra- and inter-specific variation (Barrett and Hebert, 2005; Hebert *et al.*, 2003a), and the derived “10 x rule” were present in

Phaseolus spp. The polymorphism analysis was performed on the sequence derived by combining the chloroplast DNA regions and the nuclear ITS regions separately.

Character-based analysis

A second approach, the character-based technique, was employed to look for unique sets of diagnostic characters possibly related to single varieties or variety groups of *P. vulgaris*. That is not a hierarchical method and it does not rely on distance trees. It consists in the identification of taxonomic groups through the sharing of specific informative character states, SNPs or In/Dels, narrowed to one nucleotide position or extended to multiple positions (De Salle *et al.*, 2005). Analysis of polymorphism distribution was carried out using DNASP v.4 software (Rozas, *et al.*, 2003) in order to generate a map with information on haplotype data without considering sites with alignment gaps. The program detects positions characterized by the presence of specific character states that could be proper to a particular subgroup within *P. vulgaris* species and shared by all the members of that cluster.

Genetic diversity analysis

Measures of genetic variability were used to estimate the levels of polymorphism within and between different bean accessions. Estimates of nucleotide diversity, such as π (Nei, 1987) and θ (Watterson, 1975) along with Dxy (Nei, 1987), the average number of nucleotide substitutions per site between subgroups of varieties (*i.e.*, Central American, Southern American and Italian accessions), were calculated for the total genotypes of common beans on the basis of the total number of segregating sites and mutations. The π value represents the proportion of nucleotides that differ between two sequences, averaged over all the available pairs of genotypes being compared. For each pairwise comparison of genotypes, $\pi = K/L$, where K is the average number of nucleotide differences per site and L is the gene length in bp (Nei, 1987). The θ estimate indicates the population mutation rate based on the number of segregating sites. For a given population, this parameter is usually computed as $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the specific mutation rate of the population of interest. For chloroplast DNA, $\theta = 2N_e\mu$, where N_e is the effective population size of females (Watterson, 1975). In addition, the haplotype number, H_n, and the haplotype diversity, H_d (Nei, 1987), were calculated. All the genetic diversity statistics for all the accessions and for each of the subgroups were calculated using DNASP software (Rozas *et al.*, 2003).

Differentiation statistics among sub-populations for each SNP and over all SNP markers were also computed using haplotype data information, precisely G_{ST} (Nei, 1973), *i.e.* the fraction of genetic variation within the species that is due to genetic variation between varieties, and from nucleotide sequence information, as F_{ST}, an index of genetic differentiation among populations (Lynch and Crease, 1990). Finally, the gene flow estimate, N_m, was computed for both chloroplast and nuclear markers over all bean accessions. All the genetic differentiation statistics as well as gene flow estimates between subgroups of accessions were calculated using DNASP software (Rozas *et al.*, 2003).

Additional measures of genetic variability were used to estimate the levels of polymorphism within and between different wild and cultivated beans. The average SNP marker frequency (p_i) for each nuclear and chloroplast DNA barcode region was calculated for the accessions from Central America, South America and Italy. The observed number of alleles (n_o) and the effective number of alleles (n_e) per locus were calculated according to Kimura and Crow (1964). The genetic diversity of Nei (1973) were also computed to summarize the data of nuclear and chloroplast SNP markers in *P. vulgaris*. Let p_i denote the frequency of the i^{th} marker allele at a given locus, then the genetic diversity computed as $H_e=1–\sum p_i^2$ is equivalent to the expected heterozygosity. All calculations and analyses were conducted using the software POPGENE version 1.21 (Yeh *et al.,* 1997).

An ordination analysis was performed according to the unweighted pair-group arithmetic average method (UPGMA) clustering algorithm (Sneath and Sokal, 1973), and the centroids of all accessions were constructed from the symmetrical genetic similarity matrix on the basis of Dice’s genetic similarity estimates (Dice, 1945). The principal coordinate analysis technique (Gower, 1996) was applied to compute the first two components out of the qualitative data matrix. The triangular matrix of genetic similarity estimates was double-centered and then bi-dimentionally plotted according to the extracted Eigen-vectors (Rohlf, 1972). The calculations and analyses were conducted using the appropriate routines of the software NTSYS version 1.80 (Rohlf, 1993)

Results

DNA barcoding success and levels of variability

For the selected chloroplast and nuclear markers applied across all 63 accessions of *Phaseolus* spp. PCR amplification success averaged 100% overall, although difficulties due to specific gene regions were sometime experienced giving rise to low quality sequences (**Table 3**). For all doubtful amplicons and sequences, replicated experiments were rerun for either PCR or sequencing. Only *matK* was observed to be a particularly problematic barcode marker in which amplification often failed and when successful, the sequence quality was very low. Similar difficulties have been previously reported in other studies (Kress and Erickson, 2007; Fazekas *et al.*, 2008). Hence it was decided to remove this region from the analysis and focus only on the other easily detectable markers and highly reliable sequences. The primer pairs designed for *trnT-L* and *trnH-psbA* proved to be highly universal with a 100% success for both PCR and sequencing, whereas for the other markers (*i.e.*, *rbcL*, *atpB-rbcL*, *trnL* and *rpoB-trnC*) the primers exhibited a high universality, but the sequence quality was poor for some of the amplificons. In fact, double PCR products were usually not detectable in the gel, but some problems arose during the sequencing likely as a result of multiple co-migrating amplicons of similar size, but different sequence. In a few cases, aspecific amplicons of unexpected length were clearly visible in the gel, as for *rbcL* and *atpB-rbcL*, and therefore a second PCR with more stringent conditions was performed or newly designed primer pairs were eventually adopted for sequencing (see Table 2). Similar problems were experienced and solved also for ITS1 and ITS2 markers (**Table 3**). All the barcode sequences were deposited in NCBI databases on May 5, 2009 and Agoust 31, 2009 (GenBank Accession number: GQ411617-GQ411659 for *rbcL*; GQ411841-GQ411888 for *atpB-rbcL*; GQ411554-GQ411616 for *trnL*; GQ411715-GQ411777 for *trnT-trnL*; FJ951177-FJ951239 for *trnH-psbA*; GQ411660-GQ411714 for *rpoB-trnC* and GQ411778-GQ411840 for ITS1 and ITS2 combined in one sequence).

The sequences were easily aligned for the accessions corresponding to different varieties as the only origin of point mutations was assigned to SNPs, while among sequences corresponding to different species or genera the occurrence of insertions or deletions (*i.e.*, In/Dels) in some portions of the non-coding cpDNA regions required manually editing the alignments. In the case of the ITS regions, heterozygosity was detected at only a few nucleotide positions (Table 3) and the site of nucleotide substitutions was recorded using the conventional code for degenerate bases of the IUB (International Union of Biochemistry).

The single sequences analyzed for cpDNA markers ranged, on average, from 328 bp to 1,124 bp covering a total length of 4,229 bp, whereas for ITS1 and ITS2 markers the amplified sequences were, on average, equal to 358 bp and 413 bp, respectively. In contrast to the presence of several In/Dels and SNPs among *Phaseolus* species, the occurrence of polymorphisms among *P. vulgaris* accessions was limited to single nucleotides. In particular, a total of 17 SNPs were documented across the six investigated chloroplast markers, while 10 SNPs were found for the two ITS regions (Table 3). In common bean accessions, the frequency of SNPs per target chloroplast region varied from zero (for the monomorphic *rbcL* and *atpB-rbcL*) to a maximum of 2.2, with an average value of 0.4 SNPs per 100 bp. The most informative and polymorphic cpDNA barcode regions proved to be *trnH-psbA* and *trnT-trnL* within *P. vulgaris* and among *Phaseolus* species, respectively. The nuclear ITS1 and ITS2 regions scored, respectively, 1.6 and 1 SNP per 100 bp (Table 3).

Tree-based genetic identification method

The distance matrices based on the K2P substitution model for both chloroplast and nuclear regions were recovered and the average values were calculated between *Phaseolus* species and between sub-populations within *P. vulgaris*. Combined DNA barcode sequences showed high interspecific and low intraspecific variation rates (**Table 4**). The genetic distances between *P. vulgaris* and *Vigna unguiculata*, calculated over all barcode regions, were 0.0618 and 0.1651 on the basis of cpDNA and ITS polymorphisms, respectively. Moreover, *P. vulgaris* proved to be more closely related to *P. coccineus* than to *P. lunatus*, according to both chloroplast and nuclear markers. In fact, the average genetic distance with the former was equal to 0.0104 and 0.0231, whereas with the latter it was equal to 0.0173 and 0.0432 on the basis of, respectively, cpDNA and ITS sequence information contents (Table 4). Within *P. vulgaris*, the genetic distance estimated between varietal groups coming from Central America and South America was 0.0022 and 0.0016 according to cpDNA and ITS markers, respectively (**Figure 1**).

Since our interest was mainly focused on the detection of the polymorphisms within *P. vulgaris* accessions useful for discriminating among landraces and varieties within Mesoamerican, Andean and Italian plant materials, a further analysis was done based on the DNA markers scored as polymorphic at the intra-specific level. The degree of nucleotide differentiation between congeneric species was at least 5-fold higher than values estimated within species, whereas no significant sequence divergence rate was scored between the two different gene pools of *P. vulgaris*. Furthermore, as many as 180 comparisons out of 1,600 totally performed at the intraspecific level for the chloroplast and nuclear markers showed no significant differences between varieties.

An approach of genetic distinctiveness based on the "tree method" was also pursued using chloroplast DNA markers. The Neighbor-Joining tree converts the sequence polymorphisms into genetic distances using particular nucleotide substitution models and thus, on the basis of coalescence of conspecific populations, it assembles all the accessions derived from one species, for less than incomplete sampling, in a single group (Wiemers and Fiedler, 2007). Separate analyses for each marker yielded NJ trees that were able to correctly distinguish sister species and different genera, forming separate clusters for *Vigna*, *P. lunatus*, *P. coccineus* and *P. vulgaris* (data not shown). At the same time, the NJ tree profile enabled us to illustrate the lack of discrimination among accessions within the species *P. vulgaris* due to the scarcity or complete lack of informative characters contained in some of the investigated chloroplast regions.

Within *P. vulgaris*, the occurrence of single nucleotide polymorphisms depended on the marker: for *rbcL* and *atpB-rbcL* sequences no SNPs were detected, while for the other regions the absolute number varied from a minimum of two to a maximum of four for *trnH-psbA*. In the NJ tree constructed using the sequence polymorphisms of the four variable chloroplast markers, the members of the species *P. vulgaris*, *P. coccineus* and *P. lunatus* were split into defined clusters, with bootstrap values as high as 99% or 100%, whereas the branching nodes of *P. vulgaris* sub-groups were weakly supported (< 60% in most of the cases) (**Figure 2**).

Table 3. Basic information on the cpDNA and ITS barcode regions, including sequence length of amplicons, inter- and intra-specific number and frequency of SNPs and In/Dels. The percentage of sequence-tagged site PCR and sequencing success is also reported.

	rbcL	***matK***	***trnL***	***atpB-rbcL***	***trnH-psbA***	***trnT-trnL***	***rpoB-trnC***	**ITS1**	**ITS2**
Total No. of *P. vulgaris* entries	57	57	57	57	57	57	57	57	57
No. South American accessions	16	16	16	16	16	16	16	16	16
No. Central American accessions	17	17	17	17	17	17	17	17	17
No. Italian accessions	22	22	22	22	22	22	22	22	22
No. ancestral accessions	2	2	2	2	2	2	2	2	2
Total No. of *Phaseolus* entries	63	63	63	63	63	63	63	63	63
Average amplicon length (bp)	543	695	338	328	366	836	1124	358	413
No. SNPs in *Phaseolus* spp.	8	n.d.	21	14	14	53	48	65	58
Interspecific frequency (SNPs/100 bp)	1.5	n.d.	6.0	4.3	3.8	6.5	4.2	17.4	13.8
No. SNPs in *P. vulgaris*	0	n.d.	4	0	8	3	2	6	4
Intraspecific frequency (SNPs/100 bp)	0	n.d.	1.1	0	2.2	0.4	0.2	1.6	1.0
No. of In/Dels in *Phaseolus* spp.	0	n.d.	1	4	0	5	5	10	5
Average In/Del size (bp)	0	n.d.	58	2	0	7	2	4	5
No. of heterozygous sites	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	3	7
Amplification success (%)	100%	100%	100%	100%	100%	100%	100%	100%	100%
Sequencing success (%)	100%	62%	100%	100%	100%	100%	90%	97%	100%

n.d., not determined; n.a., not applicable.

Table 4. Mean and standard deviation of the inter- and intra-specific genetic divergences calculated using the K2P distance model for the sequence derived from the combination of all chloroplast markers and ITS regions, and overall.

Interspecific K2P distance	***rbcL***	***trnL***	***atpB-rbcL***	***trnH-psbA***	***trnT-trnL***	***rpoB-trnC***	**Overall**	**St. Dev.**	**ITS1**	**ITS2**	**Overall**	**St. Dev.**
P. vulgaris/P. coccineus	0.0037	0.0139	0.0072	0.0107	0.0088	0.0070	0.01035	0.00250	0.0105	0.0169	0.0173	0.0065
P. vulgaris/P. lunatus	0.0074	0.0250	0.0204	0.0226	0.0227	0.0209	0.02314	0.00369	0.0650	0.0438	0.0432	0.0107
P. vulgaris/V. unguiculata	0.0168	0.0459	0.0515	0.0382	0.0852	0.0571	0.06181	0.00718	0.2617	0.1671	0.1651	0.0231
Intraspecific K2P distance												
P. vulgaris	0,0000	0.0041	0.0001	0.0030	0.0008	0.0006	0.00213	0.00066	0.0002	0.0016	0.0006	0.0003
St. Dev.	0,0000	0.0023	0.0001	0.0015	0.0005	0.0002	0.00069		0.0002	0.0005	0.0003	

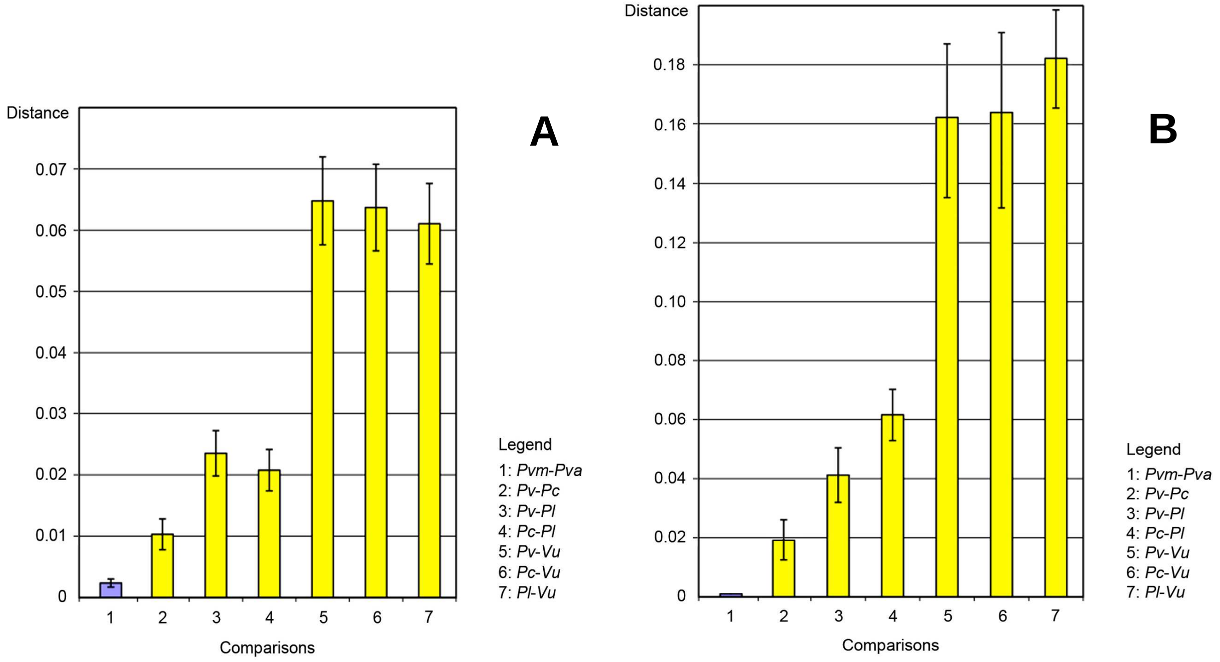


Figure 1. Histograms representing the inter- and intraspecific divergences calculated using chloroplast (A) and nuclear (B) markers. In addition to the mean value, the standard deviation is reported for each comparison within and between species.

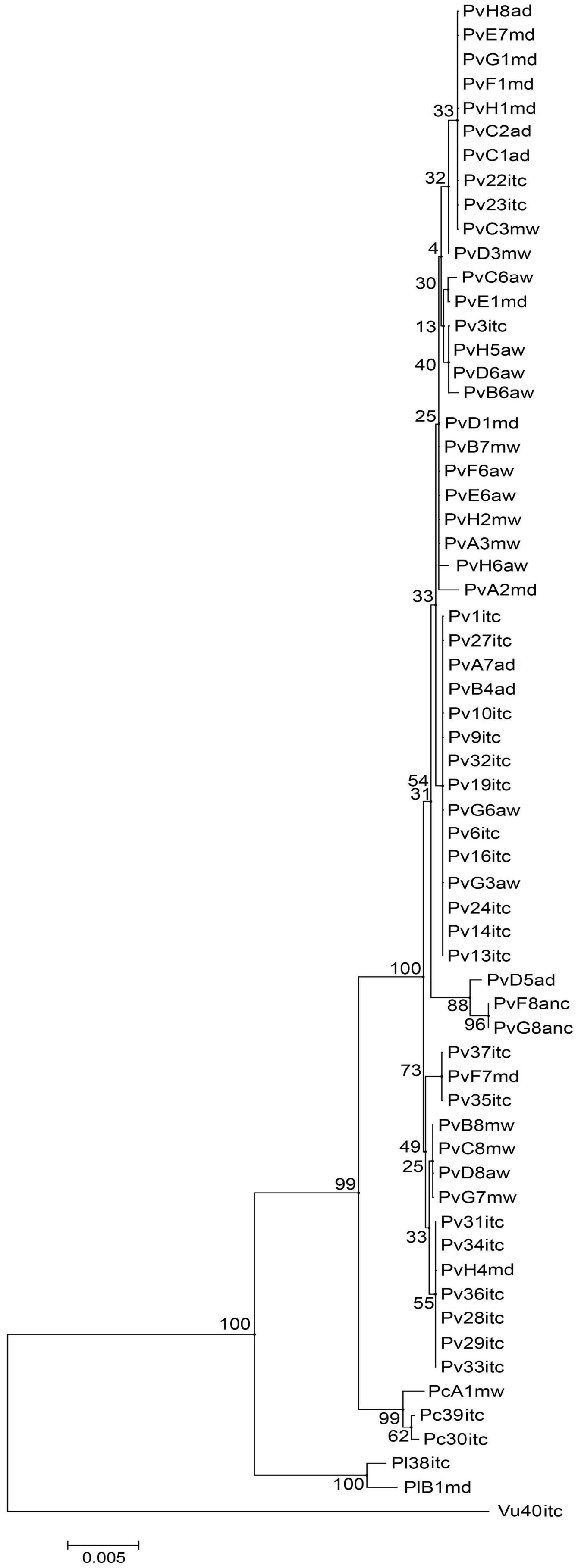


Figure 2. Neighbor-Joining tree based on Kimura 2-parameter for 63 bean entries belonging to *Phaseolus* spp. and rooted using as outgroup the accessions from *Vigna* and *P. coccineus* and *P. lunatus* species. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1,000 replicates) is shown next to the branches. The tree is drawn to scale with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree.

The accessions of *P. vulgaris* derived from either Mesoamerican or Andean gene pools grouped together and formed a few sub-clusters slightly separated from each other with a few exceptions. In four cases the gene pool was in disagreement with the geographic origin. In two of these four cases, *i.e*. PvH4md (from Mexico, but the belonging gene pool, based on the study of Rossi *et al*. (2009), was the Andean one) and PvD8aw (from Colombia, but the belonging gene pool was the Mesoamerican one), the position of these two accessions in the NJ tree was not in conflict with the positions of the other genotypes. In fact PvH4md grouped with Italian cultivars and PvD8aw clustered with two Mesoamerican accessions. In four different cases the indication of the gene pool was absent, but by means of the NJ analysis it was possible to recover this information. Two of these cases were wild accessions and for these genotypes the gene pool coincided with the geographic origin, as it was expected, while the others two were domesticated and their position in the tree suggests that they may have been transferred from one region to another, possibly by human intervention. On the whole, if all bean accessions are classified according to the position in the NJ tree, it is evident that 32 accessions belong to the Andean gene pool, while the remaining 23 to the Mesoamerican gene pool (see Table 1). It is worth mentioning that the ancestral bean accessions were recognized as a separate cluster with a high confidence value and that they grouped with another accession from Peru, the putative primary centre of the ancestral wild gene pool (Debouck *et al.*, 1993) (Figure 2).

The NJ tree constructed using the SNPs recovered from the nuclear ITS regions, based on a lower number of polymorphisms among varieties compared to cpDNA regions, revealed an unstructured distribution of the single nucleotide mutations with no sub-groups for *P. vulgaris* accessions (data not shown).

A drawback of the hierarchical technique applied in this case study was the retrieval of tie trees due to low divergence values among varieties. As a consequence, the NJ tree built for each of the barcode sequence was not unique and this fact compromised the reliability of results.

Character-based genetic characterization method

Owing to the paucity of results using a genetic distance method, a second approach known as "character-based system", was employed to identify shared diagnostic attributes that are common to the members of a given taxonomic group, but are absent from a different clade that descends from the same node (Rach *et al.*, 2009). As for NJ trees, this method does not consider In/Dels, that anyway were not found at intraspecific level, and hence the informative characters employed in the character-based approach was limited only to SNPs. Among the investigated chloroplast DNA markers, *trnH-psbA* and *trnL* showed the highest number of SNPs, proving to be the most suitable regions to discriminate genotypes within a species, along with the nuclear ITS1 and ITS2 markers. Of the other four chloroplast regions, only *trnT-trnL* and *rpoB-trnC* exhibited SNP markers among accessions, although at a lower frequency (see Table 4). On the basis of SNPs as informative characters, the analysis of the entire chloroplast data set revealed the existence of 16 haplotypes out of the 57 accessions of *P. vulgaris* (**Table 5**). It is worth noting that four of them were the most common haplotypes, each being shared by a minimum of six to a maximum of 15 accessions. Unique haplotypes were found for eight of the 57 common bean accessions (Table 5). In particular, the number of haplotypes (H_n) was equal to 9, 9 and 5 for the Central American accessions, the Southern American accessions, and the Italian varieties, respectively. The haplotype diversity (H_d) was 0.875, 0.908 and 0.688, for the three regions, respectively (**Table 6**) with a mean H_d of 0.877 for *P. vulgaris*.

Table 5. Consensus sequence related to the 17 individual SNPs detected in the target cpDNA regions with information on the haplotypes found across all common bean (*P. vulgaris*) entries.

	Marker	***trnL* intron**				***trnH-psbA***								***trnT-trnL***			***rpoB-trnC***	
	SNP position	14	183	264	332	156	219	223	224	225	229	272	283	85	512	673	478	642
Consensus sequence		G	A	T	T	A	T	A	A	A	G	T	C	A	A	T	G	A
Haplotype	No. Entries																	
Hap01	1		C															
Hap02	15					C												
Hap03	10			G								G				G		C
Hap04	3													C			T	
Hap05	6																	
Hap06	7	A			A	C												
Hap07	1												A					n.d.
Hap08	1											G				G		C
Hap09	1					C	C	T	T	T	A							n.d.
Hap10	1													C	G		T	
Hap11	1			G										C			T	
Hap12	1			G								G						C
Hap13	3	A		G	A							G				G		C
Hap14	1	A	C		A													
Hap15	3	A			A													
Hap16	2	A			A		C	T	T	T	A							

n.d., not determined.

Haplotype composition (**Hap01**: PvA2md; **Hap02**: PvA7ad, PvG6aw, PvG3aw, PvB4ad, Pv1itc, Pv6itc, Pv9itc, Pv10itc, Pv13itc, Pv14itc, Pv16itc, Pv19itc, Pv24itc, Pv27itc, Pv32itc; **Hap03**: PvC3mw, PvG1md, PvC1ad, PvH1md, PvC2ad, PvE7md, PvH8ad, PvF1md, Pv22itc, Pv23itc; **Hap04**: PvH5aw,PvD6aw, Pv3itc; **Hap05**: PvH2mw, PvA3mw, PvB7mw, PvE6aw, PvF6aw, PvD1md; **Hap06**: PvH4md, Pv28itc, Pv29itc, Pv31itc, Pv33itc, Pv34itc, Pv36itc; **Hap07**: PvH6aw; **Hap08**: PvD3mw; **Hap09**: PvD5ad; **Hap10**: PvB6aw; **Hap11**: PvC6aw; **Hap12**: PvE1md; **Hap13**: PvF7md, Pv35itc, Pv37itc; **Hap14**: PvG7mw; **Hap15**: PvB8mw, PvC8mw, PvD8aw; **Hap16**: PvF8wanc, PvG8wanc).

Table 6. Summary of genetic diversity, computed separately for chloroplast (A) and nuclear (B) DNA markers for subgroups of geographically distinct accessions and over all accessions of *Phaseolus vulgaris* L. and *Phaseolus* spp. (A,B) and for two different gene pools, and of genetic differentiation indices estimates (C), computed on the basis of cpDNA over all accessions of *Phaseolus vulgaris* L. and among Central-, Southern American and Italian accessions.

A

	Germplasm source		**Geographical origin**			**Gene pool**	
Genetic diversity statistics	***Phaseolus* spp.**	***P. vulgaris***	**Central America**	**South America**	**Italy**	**Mesoamerican[1]**	**Andean[2]**
No. segregating sites (S)	122	17	9	14	7	8	13
Haplotype number (Hn)	21	16	9	9	5	7	9
Haplotype diversity (Hd)	0.898	0.877	0.875	0.908	0.688	0.078	0.74
Average No. differences (K)	8.539	3.358	3.015	3.033	2.364	2.942	1.97
Nucleotide diversity (π/θ)	0.322	0.916	1.176	0.714	1.230	1.285	0.619
π	0.006	0.002	0.002	0.002	0.002	0.002	0,001
θ	0.018	0.002	0.002	0.003	0.001	0.001	0,002

B

	Germplasm source		**Geographical origin**			**Gene pool**	
Genetic diversity statistics	***Phaseolus* spp.**	***P. vulgaris***	**Central America**	**South America**	**Italy**	**Mesoamerican[1]**	**Andean[2]**
No. segregating sites (S)	69	9	5	7	0	6	0
Haplotype number (Hn)	9	5	2	4	1	3	1
Haplotype diversity (Hd)	0.320	0.170	0.120	0.370	0	0.255	0
Average No. differences (K)	3.760	0.620	0.590	0.930	0	0.590	0
Nucleotide diversity (π/θ)	0.240	0.312	0.389	0.434	0	0.532	0
π	0.010	0.0015	0.001	0.002	0	0.002	0
θ	0.042	0.005	0.004	0.005	0	0.004	0

C

	Overall	**Pairwise comparisons**		
Genetic differentiation	***Phaseolus vulgaris***	**M** vs. **A**	**M** vs. **I**	**A** vs. **I**
Average No. substitution	n.a.	0.003	0.002	0.002
Fixation index (G_{ST})	0.087	0.042	0.102	0.036
Differentiation index (F_{ST})	0.190	0.230	0.241	0.094
Differentiation index (N_{ST})	0.190	0.220	0.241	0.106
Gene flow (N_m)	2.26*	n.d.	n.d.	n.d.

n.d., not determined; n.a., not applicable; [1], 23 accessions; [2], 32 accessions; *, on the basis of haplotype data information.

The haplotypes based on chloroplast polymorphisms, corresponding to varietal subgroups within *P. vulgaris* species, were also used for the construction of a NJ tree (**Figure 3**). The majority of haplotypes were nested together in tightly clustered sub-groups supported by low bootstrap values, with the exception of several haplotypes shared by ancestral accessions (*i.e.*, haplotype No. 16) and wild accessions. This latter finding is particular evident for some correlated haplotypes like No. 4, 10 and 11 that are linked to the Andean gene pool, as well as 6, 14, and 15 that are associated with the Mesoamerican gene pool (see Figure 3 and Table 5). Accessions belonging to *P. coccineus*, *P. lunatus* and *Vigna unguiculata* revealed unique haplotypes that were grouped separately for each species.

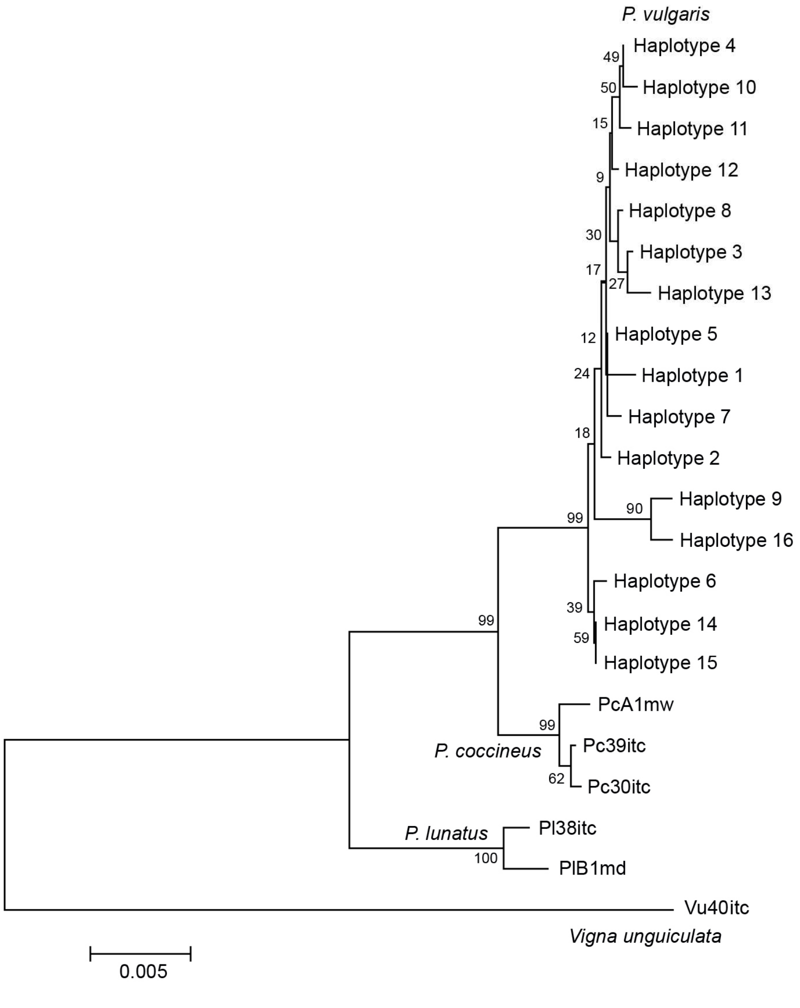


Figure 3. Neighbor-Joining tree based on the 16 haplotypes identified out of the 57 bean accessions of *Phaseolus vulgaris* L. (for details on haplotypes see also Table 5).

The ITS data set of *P. vulgaris* was not informative and all accessions, except the ancestral entries that formed two separate haplotypes, were grouped together in three haplotypes, of which one included most of the accessions (52 samples; data not shown). It is worth noting that the Italian accessions did not show a single polymorphic site, whereas the Southern American accessions were the most variable and scored the highest haplotype diversity (see Table 6).

Genetic diversity and differentiation

In total, in this study an average of 3,642 nucleotides, from both coding and non coding regions, excluding *matK* gene, were analyzed by sequencing six chloroplast markers and nuclear ITS. Among the 27 SNPs detected by comparing the accessions within *P. vulgaris* species, 13 (48%) were transitions, while 14 (52%) were transversions.

Nuclear and chloroplast related polymorphisms were used to estimate the genetic diversity and differentiation for the *P. vulgaris* germplasm. The nucleotide diversity coefficients π and θ, defined per site among chloroplast DNA sequences and considering all the genotypes, were, 2.2 X 10^{-3} and 2.4 X 10^{-3}, respectively, intermediate values between those obtained for accessions within *P. vulgaris* (Gaitan-Solis *et al.*, 2008) and for other legume crops (Zhu *et al.*, 2003; Feltus *et al.*, 2004). These values increased when also *P. coccineus*, *P. lunatus* and *V. unguiculata* were included in the analysis, being π equal to 5.9 X 10^{-3} and θ equal to 18.3 X 10^{-3}. Total data estimates of nucleotide diversity π were as low as 0.002, 0.002 and 0.0016 for the Central American, Southern American and Italian subgroups (Table 6). Regarding the genetic diversity for the ITS regions, π and θ coefficients were equal to 0.0101 and 0.0421, respectively, when considering *Phaseolus* spp. and *Vigna* together, whereas these coefficients considerably decreased when the analysis was based on common bean varieties only, being π and θ equal to 0.0015 and 0.0048, respectively. Within the *P. vulgaris* species, Central American, Southern American and Italian sub-groups scored a π value of 0.0014, 0.0023 and 0, respectively. On the whole, the π differentiation index based on ITS marker scored lower values compared to those computed for chloroplast DNA regions.

Overall summaries of genetic variation statistics for cpDNA and ITS markers, including the frequency of the most common nucleotides and the effective number of nucleotides per SNP site along with Nei's genetic diversity statistics for subgroups of accessions of different geographical origin and over all common bean accessions are reported in **Appendix 1A, 1B**.

On the basis of SNP markers, genetic differentiation statistics and gene flow estimates were also computed. The fixation index was G_{ST}=0.0870, demonstrating that only 9% of the total genetic variation found within the species is due to genetic polymorphisms among Central American, Southern American and Italian accessions. However, it is worth mentioning that on the basis of haplotypes, the fixation index scored the lowest value (0.0363) when comparing Italian accessions with those from South America and the highest one (0.1019) when comparing Italian accessions with those from Central America (see Table 6). These findings were also supported by the genetic differentiation indices F_{ST} and N_{ST} computed for all pairwise comparisons (see Table 6). Moreover, the mean estimate of gene flow (N_m) based on haplotypes was equal to 2.26 (see Table 6).

Taking into account two main sub-groups of accessions, identifying the Mesoamerican and Andean gene pools, the number of segregating sites for chloroplast regions was 8 and 13, respectively. The number of haplotypes (H_n) was equal to 7 for Mesoamerican accessions and to 9 for Andean accessions, while the estimate of haplotype diversity (H_d) resulted almost 10-fold higher in the Andean gene pool (0.7380) compared to that calculated for the Mesoamerican one (0.0775). Estimates of nucleotide diversity were also computed for the two sub-groups, being π equal to 0.0018 and 0.0013 for Mesoamerican and Andean gene pools, respectively, and θ equal to 0.0014 and 0.0021 for Mesoamerican and Andean gene pools, respectively.

These nucleotide and haplotype diversity statistics as well as the re-assignment of undefined accessions to a specific gene pool were also supported by results from ordination analyses based on the genetic similarity estimates computed using the total number of nuclear and chloroplast DNA polymorphisms. The extent of genetic differentiation and the distribution pattern of genetic variation for *P. vulgaris* accessions of Italian, Central American and Southern American geographic origin is clearly observable from the scatter diagram plotted according to the first two coordinates (**Figure 4**). Principal coordinate analysis allowed the definition of centroids for all common bean accessions and confirmed the classification based on haplotypes. In fact, most of the Italian varieties were grouped with accessions belonging to Andean domesticated gene pool, whereas only a few Italian varieties were tightly clustered with accessions of the Mesoamerican domesticated gene pool (Figure 4). Most of the Italian commercial varieties as well as the Andean wild landraces could be discriminated from each other, with a few exceptions, whereas Mesoamerican wild materials and landraces were closely grouped. Several sub-groups of closely related varieties were formed in each quadrant (for details see Figure 4).

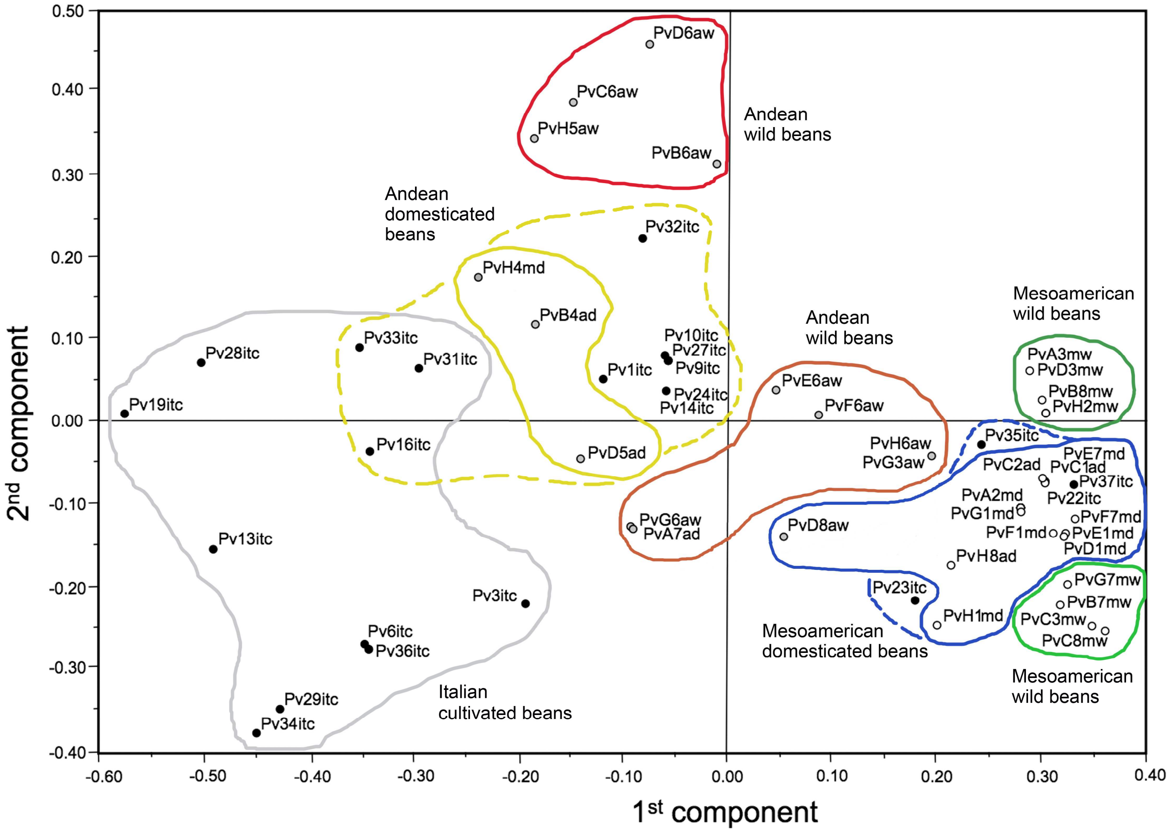


Figure 4. Centroids obtained by the PCA of 54 common bean (*P. vulgaris* L.) accessions, using Dice's genetic similarity estimates based on the whole set of chloroplast and nuclear SNP markers. The first two components were able to explain as much as 68% of the total genetic variation found at the cpDNA and ITS barcoded regions. In particular, the first component explained more than half of the total diversity and it was negatively associated with Italian commercial varieties and positively associated with Mesoamerican wild materials and landraces (Symbols: black bullets, Italian accessions; grey bullets, Andean accessions; white bullets, Mesoamerican accessions. Accession initials: mw, Mesoamerican wild; md, Mesoamerican domesticated; aw, Andean wild; ad, Andean domesticated; itc, Italian cultivated).

The first two principle components were able to explain as much as 68% of the total genetic variation found among the different varieties at the cpDNA and ITS barcoded regions. In particular, the first component, which explains 54.7% of the total diversity, was negatively associated with Italian commercial varieties and positively associated with Central American wild materials and landraces. The second component, which explains 13.2% of the total diversity, was clearly able to discriminate sub-groups of accessions within both Italian commercial varieties and Southern American accessions (Figure 4). The most discriminant nuclear SNP markers between Central and Southern American and Italian accessions proved to be ITS1-141 and ITS1-307 polymorphisms (see **Appendix 1A**, **1B**). These two SNP markers were highly shared in Central American accessions (*i.e.*, T=97% for both nucleotide residues), with intermediate values in Southern American accessions (T=56% and T=50% at positions 141 and 307, respectively) and low frequencies in Italian accessions where the alternative nucleotides were the most common ones (G=63% and C=60% at positions 141 and 307, respectively). The most discriminant chloroplast SNP markers were found in the intergenic spacers *trnH-psbA*, *trnT-trnL* and *rpbO-trnC* at positions 156, 673 and 642, respectively (see Supplementary materials, Table 2S). In particular, the first sequence site showed a fixed nucleotide in Central American accessions (A=100%), with an intermediate value in Southern American accessions (A=57%) and a low proportion in Italian accessions where the alternative nucleotide was the most common one (C=77%).

Discussion

Our results confirm that DNA barcoding is a powerful technique for identification and phylogenetic analyses in *Phaseolus* spp. aimed at reconstructing genetic distances between related species as well as evolutionary patterns. In addition to SNPs, several In/Dels were discovered among *Phaseolus* species. On the whole, the interspecific phylogenetic relationship previously identified by Delgado-Salinas *et al*. (1999) were confirmed in our analysis, with *P. vulgaris* more closely related to *P. coccineus* than to *P. lunatus*.

Since the main goal of this study was to select the markers with the best performance for barcoding at the intra-species level, our attention was focused on the relevance of the nucleotide variability among accessions of *P. vulgaris*. Taking into account the criticisms that were recently raised by the scientific community on the single barcode effectiveness and assuming that shallow variations would have been detected within species, a multi-locus approach was adopted. The criteria used to select the DNA regions suitable for barcoding in order to investigate the genetic distinctiveness of varietal groups and gene pools for common bean were: i) a high number of sequences available in public gene banks to enable the design of primers and to facilitate the identification of species by querying nucleotide databases; ii) an appropriate substitution rate for intraspecific studies on the basis of information available in the literature.

Phenetic tree-building approach versus a character-based system.

To evaluate whether DNA barcoding can be used as an efficient genomic tool for the identification of landraces and cultivars within a given species, two different strategies were adopted and tested: i) a phenetic tree-building approach using genetic distance data and the derived Neighbor-Joining tree to visualize relationships among accessions of *P. vulgaris* as well as among *Phaseolus* species and to determine the gene pool of origin for a set of Italian landraces; ii) a character-based system able to reconstruct haplotypes on the basis of diagnostic characters, fixed and variable among accessions and gene pools, to be exploited for the genetic identification of varietal groups without reference to trees. In addition, a multi-locus SNP marker analysis based on genetic similarities and differentiation statistics was employed to find out the most discriminant polymorphisms among Central American, Southern American and Italian accessions in order to estimate the biodiversity existing within this species.

With respect to the tree-building approach, the use of the divergence values among sequences and the criterion of reciprocal monophyly based on the NJ tree is the standard approach proposed by Hebert *et al*. (2003a) to discriminate among closely related species. One of the basic concepts of a DNA barcode is to employ the distance threshold derived from the barcode gap as a tool for species delimitation. This concept is controversial because a 10X screening threshold of sequence difference is present in some animals groups, such as birds and Lepidoptera (Hebert *et al.*, 2004b; Hajibabaei *et al.*, 2006), but is absent in others, such as cowries (Meyer and Paulay, 2005). This latter observation supports the hyphothesis that the barcoding gap may be an artefact of an incorrect sampling

(Meyer and Paulay, 2005; Wiemers and Fiedler, 2007). An additional tool is the use of the NJ tree profile that allows the assignment of the sequences to the correct species based on the positions of the branches relative to the cluster of the species (Wiemers and Fiedler, 2007). In our study, this kind of system confirmed to be a powerful technique to correctly cluster accessions corresponding to members of the same species by using a standardized genic or intergenic region as a molecular tag. All the sequences, when analyzed separately or together, supported the distinctiveness of different species. In contrast, this approach revealed to be poorly informative for the genetic traceability of cultivars within *P. vulgaris* species. With the exception of *trnH-psbA* and the *trnL* intron, the other chloroplast sequences did not contribute at all or offered only a small contribution to resolve the identify of landraces and varieties. The observed branching pattern of the NJ tree based on this combined data set seemed to be geographically related, with Andean and Mesoamerican bean samples clustering separately. Moreover, most of the 22 Italian varieties were found to cluster with the Andean gene pool with only six classified as Mesoamerican. This result confirms the previous observation about the origin and structure of European (Papa *et al*., 2006; Logozzo *et al.*, 2007), and Italian germplasm of *Phaseolus vulgaris* (Sicard *et al.*, 2005; Angioi *et al.*, 2009).

Unlike the NJ tree based on cpDNA, the distance tree generated by combining the sequences of the nuclear markers did not provide more resolution, but it confirmed previous evidence that discouraged the use of ITS for intraspecific phylogeny because of the occurrence of extensive intragenomic sequence variation (Alvarez and Wendel, 2003). Although the ITS regions scored an average intraspecific frequency of SNPs higher than that found for cpDNA regions (1.3 vs. 0.65 SNPs/100 bp, respectively), the random distribution of their single nucleotide mutations negatively affected the genetic discrimination of accessions and supported the likely occurrence of hybridization among accessions which may favour the occurrence of intragenomic variation. In our case, intragenomic variation is the most likely hypothesis because the inbreeding system of *P. vul*garis would exclude the occurrence of high frequency of heterozygous genotypes.

The discrimination of gene pools and the identification of varieties within *P. vulgaris* through the DNA barcoding standard tree-building approach was not informative because of slow substitution rate. For this reason a character-based system was tested. For the DNA barcoding of multiple individuals within a species, where the genetic distances are very low, it was proposed that the character-based barcode could be a more appropriate approach than the phenetic system (Rach *et al.*, 2008). This method uses DNA sequence information to generate discrete diagnostics for species identification.

To further explore the intra-specific variability, DNASP software was used to discover combinations of character states exclusive to a particular variety as well as polymorphic among varieties. The approach allowed us to detect within the species *P. vulgaris* a total of 16 haplotypes over all cpDNA regions corresponding to as many subgroups, each one made up of Mesoamerican or Andean accessions along with Italian accessions that clustered with one or the other gene pool. The only exception was haplotype No. 5, which was shared by both Mesoamerican and Andean accessions, mostly wild. The fact that the ancestral accessions were recognized as a separate cluster with high bootstrap values (>88%), along with an accession from Peru, agrees with the putative primary centre of the ancestral wild gene pool of common beans hypothesized by Debouck *et al*. (1993).

Differently from chloroplast DNA regions, as expected the nuclear ITS data set of *P. vulgaris* resulted poorly informative and almost all accessions were clustered together in a single group, except for the ancestral entries that clustered apart. In fact, the corresponding NJ tree revealed an unstructured distribution of SNPs with no sub-groups for *P. vulgaris* accessions (data not shown), and without any segregating site among the Italian accessions. These conflicts among molecular data sets (*i.e.*, chloroplast vs. nuclear markers) have been observed in other taxa as well, for example in the *Triticeae* of the grasses (Mason-Gamer and Kellogg, 1996) and the Anacardiaceae (Tingshuang *et al.*, 2004).

The whole set of SNP markers, both from ITS and cpDNA, discovered in *P. vulgaris* was used to compute genetic diversity and differentiation statistics within the ‘core collection’ of *P. vulgaris* to quantify the nucleotide variability of the bean germplasm as well as gene flow among Mesoamerican, Andean and Italian sub-populations. The Southern American accessions were more genetically differentiated than the Central American ones, with a higher number of segregating sites and with slightly higher haplotype diversity values, based on the two sets of regions. However, when the chloroplast data were analyzed alone, genetic variability at the gene pool level proved to be higher in the Andean than

Mesoamerican entries. This result agrees with those recently obtained by Benchimol *et al*. (2007), showing that Andean accessions exhibit greater mean genetic diversity than Mesoamerican accessions. However, with the only exception of SSR markers that have shown similar levels of genetic diversity between the Mesoamerican and Andean gene pools (Kwak and Gepts, 2009), using isozymes and other types of molecular markers, a higher genetic diversity was usually observed in the Mesoamerican gene pool, compared to the Andean one (Koenig and Gepts, 1989; Beebe *et al*., 2000; Beebe *et al*., 2001; Papa and Gepts, 2003; McClean *et al*., 2004; Papa *et al*., 2006). As a matter of fact, in our study the 32 common bean accessions belonging to the Andean gene pool showed estimates of genetic diversity higher than those calculated for the 23 accessions of the Mesoamerican gene pool. This finding could however be affected by the sampling strategy of plant materials, being *P. vulgaris* accessions analyzed in this study arbitrarily selected as representative of Mesoamerican and Andean gene pools on the basis of morphological seed traits and plant descriptors, as well as AFLP markers (Papa and Gepts, 2003; Rossi *et al*., 2009). Most of the Italian commercial varieties as well as the Andean wild materials and landraces, could be discriminated one from another, whereas Mesoamerican wild materials and landraces were closely related. A number of discriminant SNPs was discovered: the most discriminant nuclear SNP markers between Mesoamerican, Andean and Italian accessions were ITS1-141 and ITS1-307 polymorphisms, while the intergenic spacer *trnH-psbA* was the most informative at the chloroplast DNA level.

It is worth emphasizing that the fixation index was equal to about 0.087 for chloroplast markers, demonstrating that less than 9% of the total genetic variation found within the *P. vulgaris* collection is due to sequence polymorphisms among Mesoamerican, Andean and Italian accessions. Thus it supports hybridization and/or introgression between the two major gene pools followed by chloroplast capture, as already reported by Papa and Gepts (2003) and Chacón *et al*. (2005). This is further supported by the mean estimate of gene flow among accessions (N_m=2.26).

The 33 wild and domesticated common bean accessions can be considered a core collection of Mesoamerican and Andean gene pools, as well as the 22 commercial varieties are representative of the Italian cultivated germplasm. Both wild and domesticated accessions within Mesoamerican and Andean gene pools proved to be formed by pure lines that are poorly distinguishable genetically from each other on the basis of the cpDNA haplotypes and ITS polymorphisms. Moreover, our results revealed that genetic variability can be found to some extent within Italian cultivated beans as well as among Italian sub-groups of varieties, underlining the values of improved materials as an irreplaceable bank of diversified genotypes.

To characterize the genetic diversity among common beans different approaches were previously employed, from the analysis of morphological and phaseolin seed protein attributes to the application of several types of molecular markers (for review see Papa *et al*., 2006). By means of these investigative tools, the existence of at least two different major gene pools, *i.e*. Mesoamerican and Andean gene pools, and several racial groups was reported for *P. vulgaris* (reviewed by Chacón *et al*., 2005; see also Rossi *et al*., 2009). With this study a new molecular tool was tested to determine the genetic divergence of the modern common bean cultivars as well as to relate them to wild and domesticated materials from the original bean domestication centres. DNA barcoding combined with the NJ tree-building approach confirmed to be a highly reliable technique for identification purposes at the species-level, while it revealed to be less informative at the variety-level. On one hand, DNA barcoding provided an accurate method for the genetic identification of species of *Phaseolus* by using SNPs and In/Dels of genic or integenic tagged regions; on the other, it can be exploited for the genetic identification of varietal groups within *P. vulgaris* by means of haplotypes.

The incorporation of multiple nuclear regions may be necessary to reliably discriminate and identify single common bean varieties, mainly in groups that exhibit extensive hybridization and repetitive introgression patterns. In addition to ITS, other possible target loci for genetic identification of cultivars within *P. vulgaris* could be single or low-copy nuclear housekeeping genes.

Molecular markers find application in plant science to overcome limitations due to the absence of a standard characterization system and appropriate legal protection of modern varieties and germplasm resources, as already demonstrated in common bean (Pallottini *et al*., 2004) and other major crop species like maize (Barcaccia *et al*., 2003). In such a context, DNA barcoding in plants could be profitably exploited not only for studying biodiversity, but also for assessing genetic identity of crop varieties and foodstuffs.

Acknowledgements

Thanks are due to the A. Gini Fondation (University of Padova) to support Mrs. Silvia Nicolé during her internship at the Smithsonian Institution (Washington DC). We also thank Dr. Bruno Campion, CRA-ORL (Montanaso L., Italy), for supplying the Italian bean varieties. Funding for this project was provided by the Smithsonian Institution and the Ministry of University, Research, Science and Technology.

Conflict of interest

The authors declare no conflict of interest.

References

Alvarez I, Wendel JF (2003). Ribosomal ITS sequences and plant phylogenetic inference. Mol *Phylogenet Evol* 29: 417-434.

Angioi SA, Rau D, Rodriguez M, Lo gozzo G, Desiderio F, Papa R, Attene G (2009). Nuclear and chloroplast micro satellite diversity in *Phaseolus vulgaris* L. from Sardinia (Italy). *Mol Breeding* 23: 413-429.

Barcaccia G, Lucchin M, Parrini P (2003). Characterization of a flint maize (*Zea mays* var. *indurata*) Italian landrace, II. Genetic diversity and relatedness assessed by SSR and Inter-SSR molecular markers. *Genet Res Crop Evol* 50: 253-271.

Barrett RDH, Hebert PDN (2005). Identifying spiders through DNA barcodes. *Can J Zool* 83: 481-491.

Beebe S, Skroch PW, Tohme J, Duque MC, Pedraza F, Nienhuis J (2000). Structure of genetic diversity among common bean landraces of Middle American origin based on correspondence analysis of RAPD. *Crop Sci* 40: 264-273.

Beebe S, Rengifo J, Gaitan E, Duque MC, Tohme J (2001). Diversity and origin of Andean landraces of common bean. *Crop Sci* 41: 854-862.

Benchimol LL, Campos T, Carbonell SAM, Colombo CA (2007). Structure of genetic diversity among common bean (*Phaseolus vulgaris* L.) varieties of Mesoamerican and Andean origins using new developed microsatellite markers. *Genet Res Crop Evol* 54: 1747-1762.

Chacon MI, Pickersgill B, Debouck DG (2005). Domestication patterns in common bean (*Phaseolus vulgaris* L.) and the origin of the Mesoamerican and Andean cultivated races. *Theor Appl Genet* 110: 432-444.

Chase MV, Salamin N, Wilkinson M, Dunwell JM, Kesanakurthi RP, Haidar N, Savolainen V (2005). Land plants and DNA barcodes: short-term and long-term goals. *Phil Trans R Soc B* 360: 1889-1895.

Cowan RS, Chase MW, Kress WJ, Savolainen V (2006). 300,000 species to identify: problems, progress and prospects in DNA barcoding of land plants. *Taxon* 55: 611-616.

Ronn RN, Small RL, Haselkorn T, Wendel JF (2002). Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. *Am J Bot* 89(4): 707-725.

Debouck DG, Torro O, Paredes OM, Johnson WC, Gepts P (1993). Genetic diversity and ecological distribution of *Phaseolus vulgaris* (Fabaceae) in northwestern South America. *Econ Bot* 47: 408-423.

Delgado-Salinas A, Turley T, Richman A, Lavin M (1999) Phylogenetic analysis of the cultivated and wild species of *Phaseolus* (Fabaceae). *Syst Bot* 24: 438–460.

De Salle R, Egan MG, Siddall M (2005). The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Phil Trans R Soc B* 360: 1905-1916.

Dice LR (1945). Measures of the amount of ecological association between species. *Ecology* 26: 297-307.

Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM, Hajibabaei M, Barrett SCH (2008). Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS ONE* 3: e2802.

Feltus F.A., Wan J., Schulze S.R., Estill J.C., Jiang N. and Paterson A.H. (2004). An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. *Genome Res* 14:1812–1819.

Gaitan-Solis E, Choi I-Y, Quigley C, Cregan P, Tohme J (2008). Single nucleotide polymorphisms in common bean: their discovery and genotyping using a multiplex detection system. *Plant Genome* 1: 125-134.

Gepts P, Osborn TC, Rashka K, Bliss FA (1986). Phaseolin-protein variability in wild forms and landraces of the common bean (*Phaseolus vulgaris*): evidence for multiple centers of domestication. *Econ Bot* 40: 451-468.

Gepts P, Beavis WD, Brummer EC, Shoemaker RC, Stalker HT, Weeden NF, Young ND (2005). Legumes as a model plant family. Genomics for food and feed report of the cross-legume advances through genomics conference. *Plant Phys* 137: 1228-1235.

Gower JC (1996). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53: 325-338.

Hajibabaei M, Singer GAC, Hickey DA (2006). Benchmarking DNA barcodes: an assessment using available primate sequences. *Genome* 49: 851-854.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003a). Biological identifications through DNA barcodes. *Proc R Soc Lond* B 270: 313-321.

Hebert PDN, Ratnesingham S, deWaard JR (2003b). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc Biol Sci* 270: S96-99.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004a). Identification of birds through DNA barcodes. *PLoS Biol* 2: e312.

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004b). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proc Nat Acad Sci USA* 101: 14812-14817.

Hollingsworth PM, collaborators of the CBOL Plant Working Group (2009). A DNA barcode for land plants. *Proc Natl Acad Sci USA* 106: 12569-12570.

Kimura M (1980). A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16: 111-120.

Kimura M, Crow JF (1964). The number of alleles that can be maintained in a finite population. *Genetics* 49: 725-738.

Koenig R, Gepts P (1989). Segregation and linkage of genes for seed proteins, isozymes and morphological traits in common bean (*Phaseolus vulgaris* L.). *J Hered* 80: 455-459.

Kress WJ, Erickson DL (2007). A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH- psbA* spacer region. *PloS ONE* 6: 1-10.

Kwak M, Gepts P (2009). Structure of genetic fiversity in the two major gene pools of common bean (*Phaseolus vulgaris* L., Fabaceae). *Theor Appl Genet* 118: 979-992.

Lledó MD, Crespo MB, Cameron KM, Fay MF, Chase MW (1998) Systematics of Plumbaginaceae based upon cladistic analysis of *rbcL* sequence data. *Systematic Botany* 23: *21-29.*

Lynch M, Crease TJ (1990). The analysis of population survey data on DNA sequence variation. *Mol Biol Evol* 7: 377-394.

Logozzo G, Donnoli R, Macaluso L, Papa R, Knupffer H, Zeuli PS (2007). Analysis of the contribution of Mesoamerican and Andean gene pools to European common bean (*Phaseolus vulgaris* L.) germplasm and strategies to establish a core collection. *Genet Res Crop Evol* 54: 1763-1779.

Mason-Gamer RJ, Kellogg EA (1996). Testing for phylogenetic conflict among molecular data sets in the tribe *Triticeae* (Gramineae). *Syst Biol* 45: 522-543.

McClean PE, Lee RK, Miklas PN (2004). Sequence diversity analysis of dihydroflavonol 4-reductase intron 1 in common bean. *Genome* 47: 266-280.

Meyer CP, Paulay G (2005). DNA barcoding: error rates based on comprehensive sampling. *PloS Biol* 3: e422.

Mohler V, Schwarz G (2008). Genotyping tools in plant breeding: from restriction fragment length polymorphisms to single nucleotide polymorphisms. In: Lorz H, Wenzel G (eds.) *Molecular marker systems in plant breeding and crop improvement,* Springer: Berlin. Vol 55, pp. 23-38.

Nei M (1973). Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci USA* 70: 3321-3323.

Nei M (1987). *Molecular evolutionary genetics*. Columbia University Press: New York.

Newmaster SG, Fazekas AJ, Ragupathy S (2006). DNA barcoding in land plants: evaluation of rbcL in a multigene tiered approach. *Can J Bot* 84: 335-341.

Newmaster SG, Fazekas AJ, Steeves RAD, Janovec J (2007). Testing candidate plant barcode regions in the Myristicaceae. *Mol Ecol Res* 8: 480-490.

Pallottini L, Garcia E, Kami J, Barcaccia G, Gepts P (2004). The genetic anatomy of a patented yellow bean. *Crop Sci* 44: 968-977.

Papa R, Gepts P (2003). Asymmetry of gene flow and differential geographical structure of molecular diversity in wild and domesticated common bean (*Phaseolus vulgaris* L.) from Mesoamerica. *Theor Appl Genet* 106: 239-250.

Papa R, Nanni L, Sicard D, Rau D, Attene G (2006). The evolution of genetic diversity in *Phaseolus vulgaris* L.. In: Motley TJ, Zerega N, Cross H (eds) *New approaches to the origins, evolution and conservation of crops*, Darwin's Harvest New York: USA Columbia University Press.

Rach J, De Salle R, Sarkar IN, Schierwater B, Hadrys H (2008). Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proc R Soc B* 275: 237-247.

Rohlf FJ(1972). An empirical comparison of three ordination techniques in numerical taxonomy. *Systematic Zool* 21: 271-280.

Rohlf FJ (1993). *NTSYS-pc Numerical taxonomy and multivariate analysis system*. State University of New York, Stony Brook, NY, Version 1.8.

Rossi M, Bitocchi E, Bellucci E, Nanni L, Rau D, Attene G, Papa R (2009). Linkage disequilibrium and population structure in wild and domesticated populations of *Phaseolus vulgaris* L. *Evol appl* 2: 504-522.

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19: 2496-2497.

Sang T, Crawford DJ, Stuessy TF (1997). Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). *Am J Bot* 84:1120–1136.

Shaw J, Small RL (2005). Chloroplast DNA phylogeny and phylogeography of the North American plums (*Prunus* subgenus *Prunus* section *Prunocerasus*, Rosaceae). *Am J Bot* 92: 2011–2030.

Sicard D, Nanni L, Porfiri O, Bulfon D, Papa R (2005). Genetic diversity of *Phaseolus vulgaris* L. and *P. coccineus* L. landraces in central Italy. *Plant Breed* 124: 464-472.

Sneath PHA, Sokal RR (1973). *Numerical Taxonomy*. WH Freeman and company: San Francisco.

Taberlet P, Gielly L, Pautou G, Bouvet J (1991). Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Mol Biol* **17**: 1105-1110.

Tamura K, Dudley J, Nei M, Kumar S (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24: 1596-1599.

Tate JA, Simpson BB (2003). Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploidy species. Syst Bot 28: *723-737.*

Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP (2003). A plea for DNA taxonomy. *Trends Ecol Evol* 18: 70-74.

Tingshuang Y, Millerb AJ, Wena J (2004). Phylogenetic and biogeographic diversification of *Rhus* (Anacardiaceae) in the Northern Hemisphere. *Mol Phylogenet Evol* 33: 861-879.

Tsai L, Wang J, Hsieh H, Liu K, Linacre A, Lee JC (2008). Bidens identification using the non-coding regions of chloroplast genome and nuclear ribosomal DNA. *Forensic Sci Int Genet 2*: 35-40.

Vences M, Thomas M, Bonett RM, Vieites DR (2005). Deciphering amphibian diversity through DNA barcoding: chances and challenges. *Phil Trans R Soc B* 360: 1859-1868.

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005). DNA barcoding Australia’s fish species. *Phil Trans R Soc B* 360: 1847-1857.

Watterson GA (1975). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* 10: 256-276.

Wells JD, Stevens JR (2008). Application of DNA-Based Methods in Forensic Entomology. *Annu Rev Entomol* 53: 103-120.

White, T. J., T. Bruns, S. Lee, and J. W. Taylor. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. Pp. 315-322 In: *PCR Protocols: A Guide to Methods and Applications*, eds. Innis, M. A., D. H. Gelfand, J. J. Sninsky, and T. J. White. Academic Press, Inc., New York.

Wiemers M, Fiedler K (2007). Does the DNA barcoding gap exist? A case study in blue butterflies (Lepidoptera: Lycanidae). *Front Zool* 4: 8.

Wojciechowski MF, Lavin M, Sanderson MJ (2004). A phylogeny of legumes (Leguminosae) based on analysis of the plastid *matK* gene resolves many well-supported subclades within the family. *American J. Botany* 91: 1846-1862.

Wong EH, Hanner RH (2008). DNA barcoding detects market substitution in North American seafood. *Food Res Int* 41: 828-837.

Yeh FC, Yang RC, Boyle T (1997). POPGENE. CIFOR and Univeristy of Alberta, Canada, Version 1.21.

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2003). Single-nucleotide polymorphisms in soybean. *Genetics* 163: 1123-1134.

Web sources

http://www.barcoding.si.edu

http://www.ncbi.nlm.nih.gov/BLAST

http://www.rbgkew.org.uk/barcoding

Appendix 1A: Summary of genetic variation statistics for cpDNA markers, including the frequency of the most common nucleotides (pi) and the effective number of nucleotides (ne) per SNP site, and genic diversity (h) values referred to Mesoamerican, Andean and Italian accessions, along with the total Nei's expected heterozygosity (H) over all common bean accessions.

		Mesoamerican beans			Andean beans			Italian beans			*Phaseolus vulgaris*		
SNP markers		p_i	n_e	h	p_i	n_e	h	p_i	n_e	h	p_i	n_e	H
tnrL-014	G/A	0,7895	1,4979	0,3324	0,8571	1,3243	0,2449	0,6364	1,8615	0,4628	0,7455	1,6116	0,3795
trnL-183	A/C	0,8947	1,2321	0,1884	1,0000	1,0000	0,0000	1,0000	1,0000	0,0000	0,9636	1,0754	0,0701
trnL-264	T/G	0,4737	1,9945	0,4986	0,9286	1,1529	0,1327	0,8182	1,4235	0,2975	0,7273	1,6575	0,3967
trnL-332	T/A	0,7895	1,4979	0,3324	0,8571	1,3243	0,2449	0,6364	1,8615	0,4628	0,7455	1,6116	0,3795
trnH-psbA-156	A/C	1,0000	1,0000	0,0000	0,5714	1,9600	0,4898	0,2273	1,5414	0,3512	0,5818	1,9478	0,4866
trnH-psbA-219	T/C	1,0000	1,0000	0,0000	0,9286	1,1529	0,1327	1,0000	1,0000	0,0000	0,9818	1,0370	0,0357
trnH-psbA-223	A/T	1,0000	1,0000	0,0000	0,9286	1,1529	0,1327	1,0000	1,0000	0,0000	0,9818	1,0370	0,0357
trnH-psbA-224	A/T	1,0000	1,0000	0,0000	0,9286	1,1529	0,1327	1,0000	1,0000	0,0000	0,9818	1,0370	0,0357
trnH-psbA-225	A/T	1,0000	1,0000	0,0000	0,9286	1,1529	0,1327	1,0000	1,0000	0,0000	0,9818	1,0370	0,0357
trnH-psbA-229	G/A	1,0000	1,0000	0,0000	0,9286	1,1529	0,1327	1,0000	1,0000	0,0000	0,9818	1,0370	0,0357
Mean		0,8526	1,4118	0,1658	0,8908	1,2477	0,1681	0,8636	1,2440	0,1553	0,8671	1,3096	0,1986
St. Dev.		0,2132	0,5073	0,2172	0,1188	0,2769	0,1487	0,2064	0,3063	0,1787	0,1301	0,3089	0,1721

Appendix 1B. Summary of genetic variation statistics for cpDNA markers, including the frequency of the most common nucleotides (pi) and the effective number of nucleotides (ne) per SNP site, and genic diversity (h) values referred to Mesoamerican, Andean and Italian accessions, along with the total Nei's expected heterozygosity (H) over all common bean accessions.

		Mesoamerican beans			Andean beans			Italian beans			*Phaseolus vulgaris*		
SNP markers		p_i	n_e	h	p_i	n_e	h	p_i	n_e	h	p_i	n_e	H
ITS1-080	C/T	0,8889	1,2462	0,2032	0,9688	1,0644	0,0625	1,0000	1,0000	0,0000	0,9519	1,1008	0,0915
ITS1-141	T/G	0,9722	1,0571	0,0556	0,5625	1,9692	0,5081	0,3333	1,8000	0,4571	0,6250	1,8824	0,4688
ITS1-161	C/G	1,0000	1,0000	0,0000	0,9375	1,1327	0,1210	1,0000	1,0000	0,0000	0,9808	1,0392	0,0377
ITS1-168	T/C	1,0000	1,0000	0,0000	0,8750	1,2800	0,2258	0,7500	1,6000	0,3857	0,8750	1,2800	0,2188
ITS1-296	C/T	0,9444	1,1172	0,1079	1,0000	1,0000	0,0000	1,0000	1,0000	0,0000	0,9808	1,0392	0,0377
ITS1-307	T/C	0,9722	1,0571	0,0556	0,5000	2,0000	0,5161	0,3056	1,7373	0,4365	0,5962	1,9287	0,4815
ITS2-102	T/C	1,0000	1,0000	0,0000	0,7353	1,6347	0,4011	0,8333	1,3843	0,2857	0,8611	1,3144	0,2392
ITS2-157	C/T	0,9737	1,0540	0,0526	1,0000	1,0000	0,0000	1,0000	1,0000	0,0000	0,9907	1,0187	0,0183
ITS2-248	A/G	0,9737	1,0540	0,0526	0,7059	1,7101	0,4278	0,6111	1,9059	0,4889	0,7685	1,5523	0,3558
ITS2-357	C/G	1,0000	1,0000	0,0000	0,7941	1,4859	0,3369	0,7778	1,5283	0,3556	0,8611	1,3144	0,2392
Mean		0,9725	1,0586	0,0527	0,8079	1,4280	0,2599	0,7611	1,3956	0,2410	0,8491	1,3470	0,2189
St. Dev.		0,0347	0,0763	0,0640	0,1801	0,3879	0,2043	0,2681	0,3691	0,2147	0,1441	0,3386	0,1742

Chapter 4

Use of DNA barcoding in crop plants: *V. vinifera* L.

DNA barcoding and its potentials for genetic distinctiveness of grapevine cultivars

Silvia Nicolè±*, David L. Erickson†, Gianni Barcaccia±, Marzia Salmaso±, W. John Kress† and Margherita Lucchin±

*Department of Environmental Agronomy and Crop Science, Università degli Studi di Padova, Viale Università 16 – Campus of Agripolis, 35020 Legnaro, Padova (Italy); †Department of Botany and §Laboratories of Analytical Biology – National Museum of Natural History, Smithsonian Institution, P.O. Box 37012, Washington, DC 20013-7012 (USA).

* Silvia Nicolè
Phone: +39 049 827 2867 Fax: +39 049 827 2839 E-mail: silvia.nicole@unipd.it

Keywords: *Vitis* cultivars, DNA barcoding, SNPs, haplotype

Abstract

Vitis vinifera L., with more than 8000 cultivars in existence and a world wide cultivation is one of the most important agricultural crops to society. The difficult to recognize them, by means of morphological features, has prompted the development of new molecular markers able to detect the genetic diversity and discriminate among cultivars. In the present work, we demonstrate how we reconstructed cultivar-specific haplotype performed by means of DNA barcoding and extension into diploid SNP loci, using the character-based system in place of the conventional phenetic approach. Among the 149 *V. vinifera* genotypes studied, on the basis of three nuclear coding regions, GAI gene and two ESTs, it was possible to define 63 haplotypes of which 38 were cultivar-specific, while the other cases were more complex haplotypes grouping several varieties at the same time. Overall, the technique resulted to be successful in inferring haplotypes useful for definition of cultivar genotypes and also allowed us to corroborate some hypotheses, regarding the origin of some local cultivars, that suggested some issues of misidentification (synonymy/homonymy). The obtained data show that a SNP based detection technique will be a suitable tool for grapevine fingerprinting useful for biodiversity and food traceability aims.

Introduction

The Vitaceae family consists of 14 genera and about 900 species, primarily distributed across tropical regions, but with a few genera, such as *Vitis*, present in temperate areas (Soejima and Wen, 2006). The huge economic and agronomic importance of this family derives from the species *Vitis vinifera* L. that is the only species extensively used in the global wine agro-industry. Although a great deal of information is available about the horticultural management of commercial grapes, because grapevine represents one of the major perennial crops in the world, there is a surprising lack of information about the systematic positioning of the family and about the place and period of the two independent domestication events of grape plant (Soejima and Wen, 2006; Jansen *et al*., 2006). A recent work suggested, by means of 15 chloroplast microsatellites, that the probable centre of origin of the species is the Caucasian region since it is the area with the highest degree of biodiversity (Grassi *et al*., 2006). From the primo-domestication site that occurred in the Near-East (Iran, Georgia, Turkey, 7400-7000 BP), the grape moved toward China and gradually spread to Mesopotamia and Egypt until to reach the Mediterranean basin, Greece, Italy, France and Spain, the secondary domestication centre (Grassi *et al*., 2003). After that, the grape cultivations colonized some regions of Northern Europe and then the New World countries where wild species (*i.e. V. ruparia*, *V. rupestris*, *V. berlandieri*, *V. cinerea*) showing natural resistance to some pathogenes (phylloxera, oidium, mildew), were present. These pathogenes in the middle of nineties century were introduced in Europe where became responsible of the spread of pest diseases causing a significant reduction in European wild and cultivated grapevines, sensible to the parasites.

The vast majority of world's grapes are cultivars of *V. vinifera* subsp. *vinifera* (or *sativa*) that is believed to be derived from the wild *V. vinifera* subsp. *silvestris*. During the domestication, the wild ancestor underwent several drastic morphological and physiological changes, such as changes in berry and bunch size, seed and flower morphology, increase of sugar content and greater and more regular productivity (This *et al*., 2006). The cultivated grapevine is a diploid plant, highly heterozygous and nearly all cultivars are hermaphroditic, self-fertile and out-cross easily. Three different processes have had a significance impact on the development of cultivated grapevines: sexual reproduction, vegetative propagation and somatic mutations. New genotypes are produced by sexual reproduction, either by crossing or self-fertilization, and then the adoption of genotypes with desirable traits is realized by vegetative propagation. In fact the marked heterozygosity of grape, the necessity to dispose of genotypes with stable morphological features and the high incidence of inbreeding depression forced the viticulturists to adopt the asexual propagation to ensure conformity to the progeny (see review of Bessis, 2007). Although clonal propagation should warrant that all plants derived from the same mother plant are genetically identical, the occurrence of somatic mutations might eventually lead to the formation of clonal variations and, in the case in which the somatic mutation occur in only one cell layer of the plant, to a genetic chimerism, i.e. the co-existence of cells with different genetic patrimony in the same organism. Thanks to this huge source of mutations, thousands of grape cultivars and even biotypes within a cultivar exist and are generally classified according to their final production, wine and table grapes and raisins. Currently the number of different varieties collected in worldwide germplasm collections is estimated to be around 10000, even if it is also recognized that many cases of synonymy and homonymy exist (Alleweldt and Dettweiler, 1994). Through the use of microsatellite markers, very useful to determine cultivar identity and parentage, it is plausible hypothesizing that a more accurate estimate of the number of cultivars may be around 5000 (This *et al.*, 2006). Italy probably represents the richest country in ampelo-biodiversity due to both the officially native grapevines and the massive presence of regional minor vineyards that together group around 2000 cultivars compared to the only 400 present in France (Schneider, 2005-2006).

Despite this huge biodiversity richness, only a small percent of *Vitis vinifera* varieties are employed for the production of wine (Hidalgo 1993) and therefore this contributes to the genetic erosion and the loss of variability in all those countries were the viticulture practice is really common, as in Italy, Spain, France (Gago *et al.*, 2009). Consequently, the identification and characterization of grape varieties is necessary and must be ensured also for the oldest ones that represent a huge genetic resource for improvement programmes. In addition describing old and local cultivars can turn out useful for the valorisation of wine grapes in the view of food traceability. In fact, varietal authenticity tests of grapes, juices, musts and wines are important to grapegrowers and winemakers since the wine quality depends by the vinification process, the geographical origin of the grapes and the varietal composition of the must (Pinder and Meredith, 2003). In addition, after the introduction of wine labelling laws and trade regulations, now also the marketing of wine requires the development of diagnostic tools able to correctly identify varieties used for the production of wines. For example, the labelling with DOC and IGT marks, conferring an additional value to the product, can arise fraudulent mislabelling events and thus European Legislation (EEC No. 2081/92) was born to protect the geographical indications and designations of origin (Dennis, 1998)

Accurate identification and characterization of grapevine cultivars relies on the choice of appropriate investigative tools. Traditionally, the ampelography, the field of botany concerned with the identification and classification of grapevines, was based on plant morphology. The first complete systematic work assembling several criteria for the identification of 9600 vines dates back to 1952 by Pierre Galet, *Ampélographie Pratique.* Actually the International Organisation of Vine and Wine (OIV) is responsible for the delineation of standards to guarantee the authenticity of grapes and vine products (http://www.oiv.int/uk/accueil/index.php). A list of phenotypic traits employed to distinguish varieties includes for each variety: name and synonyms, morphological aspects, such as descriptions of leaves, growing shoots, shoot tips, petioles, flowers, grape clusters and berries, cultural attitudes, such as disease or insects resistance, and climatic needs. Even with such a wide morphological keys, the task to properly recognize grape cultivars is difficult to achieve and, since the high adaptability of *V. vinifera* species to environmental conditions that can heavily affect its phenotype, the misidentifications are common. Therefore new approaches were developed to guarantee the identification of both grapes and also vine-derived products, such as juice and wine, to which the morphological assays are clearly not applicable (Garcia-Beneytez *et al*., 2002; Siret *et al*., 2002).

Alternatives to ampelography for varietal identification are protein profiling and DNA fingerprinting. The former is a technique based on the detection of macromolecules, such as proteins (Moreno-Arribas *et al*., 1999; Hayasaka *et al*., 2001) or compounds from the secondary metabolism as anthocyanins (Pomar *et al*., 2005). The latter is based on the discovery of nucleotide polymorphisms to characterize a specific genetic entity. Until now, most DNA profiling studies in grapevine have been performed using neutral markers, such as Random Amplified Polymorphic DNA (RAPD; Siles *et al*., 2000), Amplified Fragment Length Polymorphism (AFLP; Ergul *et al*., 2004) and Simple Sequence Repeat (SSR; Salmaso *et al*., 2008). Currently, the SSR markers represent the official diagnostic tool adopted by the international scientific community to define a cultivar and a set of six SSR loci are now considered to be sufficient for genetic identifications of most cultivars (This *et al*., 2004), thus to insert it in the Vitis International Catalogue of Cultivated Varieties (http://www.vivc.bafz.de/index.php; This *et al*., 2004). An other class of markers, more suitable than SSRs, is represented by Single Nucleotide Polymorphisms (SNPs), single base-pair differences in the form of substitutions or Insertion/Deletions (In/Dels), that represent the most frequent source of genetic variability in the human genome (Collin *et al*., 1998). The recent technological advances and the funding of two separate genome sequencing projects (Jaillon *et al*., 2007; Velasco *et al*., 2007) made available the whole sequence of the grapevine nuclear genome, encouraging the analysis of allelic diversity and SNPs characterization. Since the SNP discovery can be easily automated and currently there are several laboratory and computational approaches to detect SNPs within a genome, based on comparative analysis of the same DNA snippet from different individuals, the application of these markers can be useful to characterize and map genes involved in the genetic control of important traits, to detect associations between alleles and phenotypes (Rafalski *et al*., 2002) and for phylogeographic purposes (Brumfield *et al*., 2003). Technically, in order to find the SNPs the nucleotide fragment obtained by PCR amplification, can be analyzed by means of strand conformational polymorphisms (SSCP), melting temperature analysis, heteroduplex analysis (HA), CAPS, or direct sequence analysis, in an approach called DNA barcoding. By means of DNA barcoding, an unknown organism could be identified by matching DNA sequence recovered from the sample to a database of sequences from known organisms, previously described and recognized using morphological keys (Hebert *et al*., 2003a). This technique could be of huge utility for the correlation of the genetic diversity with the phenotypic variability and hence for the definition of cultivars-specific haplotypes exploitable for authentication assays. Anyway, the employment of DNA barcoding at sub-species level is not a conventional application of the method and, as proved by previous results in an other important crop species, such as *Phaseolus vulgaris* (Nicolè *et al*., submitted), it requires the exploitation of a different approach. In fact, since the genetic distance among subgroups within a species is generally too small to allow the definition of a sort of genetic threshold to delimitate different varieties, the employ of the more complex character-based clustering system, founded on the concept of haplotype, could turn out useful for intraspecies study.

The aim of this research is developing a character-state DNA barcoding to unambiguously distinguish varieties within *V. vinifera* species in order to both safeguard the genetic patrimony of the species, for example protecting the local varieties and resolving cases of homonymy and synonymy, and to warrant the authenticity of the grapevine cultivars and their geographical origin.

Materials and methods

Germplasm sampling of *Vitis*

For the molecular analysis, we sampled leaves from 144 different cultivars of *Vitis vinifera*, selected as representatives of the most common cultivars spread in Europe, most of them with final destination for wine production, while a few for table and raisins consumption. Generally only one specimen was collected for each cultivar and, only for a few cases, several individuals, different clones with different origin, were included in the study, for a total of 162 individuals. In details, 135 international certified genotypes within *V. vinifera* species, 85 from Italy, 4 from Rumania, 20 from Spain, 11 from Greece and 16 from Portugal were supplied by certified commercial nurseries. In addition, 24 genotypes of ancient local cultivars, held in two private collections near the Euganean Hills (Padua) plus one cultivar from Breganze (Vicenza), were analyzed as particular study case. Finally, two interspecific hybrids, Bianca and Tintoria, were added and a subsampling of *V. riparia, V. rupestris*, *V. berlandieri*, *V. cinerea* and *V. labrusca* accessions were used as reference standards and out-groups (**Table 1**).

Table 1. List of grapevine accessions with the indication of origin, certification and colour berry.

No.	**Species**	**Cultivar**	**Origin**	**Source**	**Berry**	**Destination**
622	*Vitis vinifera*	Alphonse Lavallez	Italy	certified	red	table
621	*Vitis vinifera*	Cardinal	Italy	certified	red	table
620	*Vitis vinifera*	Moscato d'Amburgo	Italy	certified	red	table
617	*Vitis vinifera*	Palieri	Italy	certified	red	table
705	*Vitis vinifera*	Aledo	Spain	certified	red	table
619	*Vitis vinifera*	Italia	Italy	certified	white	table
738	*Vitis vinifera*	Matilde	Italy	certified	white	table
737	*Vitis vinifera*	Regina	Italy	certified	white	table
736	*Vitis vinifera*	Regina Inzolia	Italy	certified	white	table
728	*Vitis vinifera*	Regina Razaki	Greece	certified	white	table
723	*Vitis vinifera*	Sultanina 919	Greece	certified	white	raisins
724	*Vitis vinifera*	Sultanina 122	Greece	certified	white	raisins
624	*Vitis vinifera*	Aglianico	Italy	certified	red	wine
554	*Vitis vinifera*	Barbera	Italy	certified	red	wine
635	*Vitis vinifera*	Bovale Sardo	Italy	certified	red	wine
559	*Vitis vinifera*	Cabernet Franc	Italy	certified	red	wine
555	*Vitis vinifera*	Cabernet Sauvignon	Italy	certified	red	wine
633	*Vitis vinifera*	Calabrese	Italy	certified	red	wine
632	*Vitis vinifera*	Canaiolo Nero	Italy	certified	red	wine
593	*Vitis vinifera*	Cannonau	Italy	certified	red	wine
610	*Vitis vinifera*	Carignan	Italy	certified	red	wine
594	*Vitis vinifera*	Carmenere ISV	Italy	certified	red	wine
601	*Vitis vinifera*	Carmenere R9	Italy	certified	red	wine
628	*Vitis vinifera*	Ciliegiolo	Italy	certified	red	wine
626	*Vitis vinifera*	Colorino	Italy	certified	red	wine
642	*Vitis vinifera*	Corvina	Italy	certified	red	wine
592	*Vitis vinifera*	Croatina	Italy	certified	red	wine
591	*Vitis vinifera*	Dolcetto	Italy	certified	red	wine
589	*Vitis vinifera*	Franconia	Italy	certified	red	wine
643	*Vitis vinifera*	Freisa	Italy	certified	red	wine
644	*Vitis vinifera*	Grignolino	Italy	certified	red	wine
567	*Vitis vinifera*	Lambrusco Maestri	Italy	certified	red	wine
739	*Vitis vinifera*	Malbech cl.594	Italy	certified	red	wine
740	*Vitis vinifera*	Malbech ISVR6	Italy	certified	red	wine
611	*Vitis vinifera*	Malbo Gentile	Italy	certified	red	wine
602	*Vitis vinifera*	Malvasia Nera	Italy	certified	red	wine
553	*Vitis vinifera*	Merlot	Italy	certified	red	wine
615	*Vitis vinifera*	Montepulciano	Italy	certified	red	wine
564	*Vitis vinifera*	Nebbiolo	Italy	certified	red	wine
636	*Vitis vinifera*	Negroamaro	Italy	certified	red	wine
722	*Vitis vinifera*	Nero d'Avola	Italy	certified	red	wine
609	*Vitis vinifera*	Petit Verdot	Italy	certified	red	wine
629	*Vitis vinifera*	Piedirosso	Italy	certified	red	wine
569	*Vitis vinifera*	Pinot Gris	Italy	certified	red	wine
556	*Vitis vinifera*	Pinot Noir VCR	Italy	certified	red	wine
570	*Vitis vinifera*	Pinot Noir c115	Italy	certified	red	wine
586	*Vitis vinifera*	Primitivo di Gioia	Italy	certified	red	wine
552	*Vitis vinifera*	Raboso Piave	Italy	certified	red	wine
558	*Vitis vinifera*	Raboso Veronese	Italy	certified	red	wine
583	*Vitis vinifera*	Refosco Penduncolo Rosso	Italy	certified	red	wine
639	*Vitis vinifera*	Rondinella	Italy	certified	red	wine
582	*Vitis vinifera*	Sagrantino§	Italy	certified	red	wine
634	*Vitis vinifera*	Sagrantino§	Italy	certified	red	wine
560	*Vitis vinifera*	Sangiovese	Italy	certified	red	wine

641	*Vitis vinifera*	Teroldego	Italy	certified	red	wine
646	*Vitis vinifera*	Tocai Rosso	Italy	certified	red	wine
645	*Vitis vinifera*	Vernaccia Serrapetrona	Italy	certified	red	wine
604	*Vitis vinifera*	Albana	Italy	certified	white	wine
640	*Vitis vinifera*	Arneis	Italy	certified	white	wine
557	*Vitis vinifera*	Chardonnay Blanc	Italy	certified	white	wine
637	*Vitis vinifera*	Falanghina	Italy	certified	white	wine
590	*Vitis vinifera*	Fiano	Italy	certified	white	wine
562	*Vitis vinifera*	Garganega	Italy	certified	white	wine
613	*Vitis vinifera*	Grechetto	Italy	certified	white	wine
631	*Vitis vinifera*	Greco	Italy	certified	white	wine
603	*Vitis vinifera*	Malvasia del Chianti	Italy	certified	white	wine
721	*Vitis vinifera*	Malvasia Istriana	Italy	certified	white	wine
563	*Vitis vinifera*	Manzoni Bianco	Italy	certified	white	wine
588	*Vitis vinifera*	Moscato Bianco	Italy	certified	white	wine
612	*Vitis vinifera*	Moscato Giallo	Italy	certified	white	wine
608	*Vitis vinifera*	Moscato Sardo	Italy	certified	white	wine
587	*Vitis vinifera*	Picolit	Italy	certified	white	wine
568	*Vitis vinifera*	Pinot Blanc	Italy	certified	white	wine
630	*Vitis vinifera*	Prosecco Balbi	Italy	certified	white	wine
623	*Vitis vinifera*	Prosecco Lungo	Italy	certified	white	wine
584	*Vitis vinifera*	Ribolla Gialla	Italy	certified	white	wine
625	*Vitis vinifera*	Riesling Italico	Italy	certified	white	wine
627	*Vitis vinifera*	Riesling Renano	Italy	certified	white	wine
561	*Vitis vinifera*	Sauvignon Blanc	Italy	certified	white	wine
565	*Vitis vinifera*	Tocai Friulano	Italy	certified	white	wine
551	*Vitis vinifera*	Tramier	Italy	certified	white	wine
566	*Vitis vinifera*	Trebbiano Romagnolo	Italy	certified	white	wine
607	*Vitis vinifera*	Trebbiano Toscano	Italy	certified	white	wine
638	*Vitis vinifera*	Verduzzo Friulano	Italy	certified	white	wine
606	*Vitis vinifera*	Vermentino	Italy	certified	white	wine
616	*Vitis vinifera*	Vittoria	Italy	certified	white	wine
605	*Vitis vinifera*	Traminer Aromatico	Italy	certified	pink	wine
726	*Vitis vinifera*	Aghorghitiko	Greece	certified	red	wine
731	*Vitis vinifera*	Moscomavro	Greece	certified	red	wine
725	*Vitis vinifera*	Xinomauro	Greece	certified	red	wine
729	*Vitis vinifera*	Asirtiko	Greece	certified	white	wine
727	*Vitis vinifera*	Korintos	Greece	certified	white	wine
733	*Vitis vinifera*	Moscofilero	Greece	certified	white	wine
730	*Vitis vinifera*	Rhoditis	Greece	certified	white	wine
732	*Vitis vinifera*	Robolla	Greece	certified	white	wine
755	*Vitis vinifera*	Tempranino(Tinta Moriz)	Portugal	certified	red	wine
746	*Vitis vinifera*	Tinta Barroca	Portugal	certified	red	wine
747	*Vitis vinifera*	Tinta Francisca	Portugal	certified	red	wine
745	*Vitis vinifera*	Tinto Cao	Portugal	certified	red	wine
743	*Vitis vinifera*	Touriga Franca	Portugal	certified	red	wine
742	*Vitis vinifera*	Trincadeira	Portugal	certified	red	wine
744	*Vitis vinifera*	Turiga National	Portugal	certified	red	wine
753	*Vitis vinifera*	Alfrocheiro	Portugal	certified	black	wine
756	*Vitis vinifera*	Bastardo	Portugal	certified	black	wine
750	*Vitis vinifera*	Castelao	Portugal	certified	black	wine
748	*Vitis vinifera*	Vinao (Souson)	Portugal	certified	black	wine
751	*Vitis vinifera*	Antao Vaz	Portugal	certified	white	wine
752	*Vitis vinifera*	Arinto Armas	Portugal	certified	white	wine
741	*Vitis vinifera*	Fernao pires	Portugal	certified	white	wine
754	*Vitis vinifera*	Malvasia Fine	Portugal	certified	white	wine
749	*Vitis vinifera*	Rabigato	Portugal	certified	white	wine

649	*Vitis vinifera*	Feteasca Neagra	Rumania	certified	red	wine
647	*Vitis vinifera*	Feteasca Alba	Rumania	certified	white	wine
648	*Vitis vinifera*	Feteasca Regala	Rumania	certified	white	wine
650	*Vitis vinifera*	Mustoasa de Maderat	Rumania	certified	white	wine
712	*Vitis vinifera*	Bobal	Spain	certified	red	wine
715	*Vitis vinifera*	Cannonao	Spain	certified	red	wine
714	*Vitis vinifera*	Cannonao Garnacha	Spain	certified	red	wine
716	*Vitis vinifera*	Cannonao Grenache	Spain	certified	red	wine
719	*Vitis vinifera*	Graciano	Spain	certified	red	wine
708	*Vitis vinifera*	Mencia	Spain	certified	red	wine
720	*Vitis vinifera*	Monastrel	Spain	certified	red	wine
713	*Vitis vinifera*	Prieto Picudo	Spain	certified	red	wine
701	*Vitis vinifera*	Tempranillo	Spain	certified	red	wine
703	*Vitis vinifera*	Tempranillo Tinta Pais	Spain	certified	red	wine
702	*Vitis vinifera*	Tempranillo Tinto de Toro	Spain	certified	red	wine
704	*Vitis vinifera*	Tinta Fina	Spain	certified	red	wine
707	*Vitis vinifera*	Albarino	Spain	certified	white	wine
710	*Vitis vinifera*	Blanca Cayetana	Spain	certified	white	wine
706	*Vitis vinifera*	Macabeo	Spain	certified	white	wine
711	*Vitis vinifera*	Parda	Spain	certified	white	wine
718	*Vitis vinifera*	Parellada	Spain	certified	white	wine
717	*Vitis vinifera*	Pedro Ximenez	Spain	certified	white	wine
709	*Vitis vinifera*	Xarello	Spain	certified	white	wine
528	*Vitis vinifera*	Gruaja*	Breganze, Vicenza	local	red	wine
507	*Vitis vinifera*	Agostana Nera*	Euganean Hills, Padua	local	red	wine
508	*Vitis vinifera*	Cabernet Lispida*	Euganean Hills, Padua	local	red	wine
504	*Vitis vinifera*	Corbinella*	Euganean Hills, Padua	local	red	wine
503	*Vitis vinifera*	Corbinona*	Euganean Hills, Padua	local	red	wine
517	*Vitis vinifera*	Friularo 1*	Euganean Hills, Padua	local	red	wine
518	*Vitis vinifera*	Friularo 2*	Euganean Hills, Padua	local	red	wine
519	*Vitis vinifera*	Friularo 3*	Euganean Hills, Padua	local	red	wine
520	*Vitis vinifera*	Friularo 4*	Euganean Hills, Padua	local	red	wine
521	*Vitis vinifera*	Friularo 7*	Euganean Hills, Padua	local	red	wine
501	*Vitis vinifera*	Gatta*	Euganean Hills, Padua	local	red	wine
510	*Vitis vinifera*	Marzemina Cenerenta*	Euganean Hills, Padua	local	red	wine
511	*Vitis vinifera*	Marzemina Nera	Euganean Hills, Padua	local	red	wine
512	*Vitis vinifera*	Marzemina Nera bastarda*	Euganean Hills, Padua	local	red	wine
506	*Vitis vinifera*	Merlot 181	Euganean Hills, Padua	local	red	wine
505	*Vitis vinifera*	Merlot R3	Euganean Hills, Padua	local	red	wine
513	*Vitis vinifera*	Negrara Veronese*	Euganean Hills, Padua	local	red	wine
514	*Vitis vinifera*	Pattaresca*	Euganean Hills, Padua	local	red	wine
522	*Vitis vinifera*	Raboso Piave 1	Euganean Hills, Padua	local	red	wine
523	*Vitis vinifera*	Raboso Piave 2	Euganean Hills, Padua	local	red	wine
524	*Vitis vinifera*	Raboso Veronese	Euganean Hills, Padua	local	red	wine
509	*Vitis vinifera*	Marzemina Bianca	Euganean Hills, Padua	local	white	wine
502	*Vitis vinifera*	Pignola	Euganean Hills, Padua	local	white	wine
515	*Vitis vinifera*	Schiavetta Doretta*	Euganean Hills, Padua	local	white	wine
618	interspecific hybrid	Perla	Italy	certified	white	table
535	interspecific hybrid	Bianca	Italy	certified	white	wine
516	interspecific hybrid	Tintoria*	Euganean Hills, Padua	local	red	wine
530	*Vitis riparia*	Gloire	CRA ISV collection	local	red	rootstock
531	*Vitis rupestris*	Du Lot	CRA ISV collection	local	red	rootstock
532	*Vitis berlandieri*	wild	CRA ISV collection	local	red	rootstock
533	*Vitis cinerea*	wild	CRA ISV collection	local	red	germplasm
534	*Vitis labrusca*	wild	CRA ISV collection	local	red	germplasm

*, Varieties not registered in the Italian Catalogue of Cultivated Varietes; CRA ISV, Consiglio per la Ricerca e la Sperimentazione in Agricoltura - Istituto Sperimentale per la Viticoltura; §, same clone.

Genomic DNA extraction

Genomic DNA was isolated from frozen young leaf tissue using DNeasy Extraction kit (Qiagen) according to the manufacturer's protocol and the DNA was eluted in 80-100 µl of TE 0.1 Buffer (Tris-HCl 100 mM, EDTA 0.1 mM pH 8). The final concentration of DNA was estimated by electrophoresis on 0.8% agarose/TAE gel and the quantification was conducted by comparison with 1 Kb plus DNA ladder (Invitrogen) of known concentration.

DNA barcode markers and PCR assays

In a preliminary assay, seven different chloroplast markers (*rpoB*, *rps* and *rpl32* genes and *trnH-psbA*, *trnT-trnL*, *atpB-rbcL* and *psbK-psbI* intergenic spacers) were chosen because they proved to be the most polymorphic regions in many taxa (ref). Once verified the inadequacy of the chloroplast genome, we shifted to the nuclear genome and four nuclear cDNA sequences (IF01, IB02, ID04 and IIC08), belonging to an EST (Expressed Sequence Tags) database containing sequences related to four functional classes of genes - sugar metabolism, cell signalling, anthocyanin metabolism and defence related - and the GAI gene, involved in the biosynthetic pathway of the gibberellins (Gas) (Wen *et al*., 2007), were selected and amplified for all the accessions. For each chloroplast and nuclear marker, the PCR reactions were conducted in a volume of 25 µl containing 15 ng of genomic DNA as template, 1× PCR buffer (100 mM Tris-HCl pH 9.0, 15 mM $MgCl_2$ and 500 mM KCl), 0.2 mM dNTPs, 0.2 µM of each primer and 0.5 U of *Taq* DNA polymerase. The primers pairs, along with the relative nucleotide sequences and the reference information, are supplied in **Table 2**. All the PCR amplifications were performed on a GeneAmp PCR System 9700 (Applied Biosystems). The themalcycling conditions for the chloroplast regions were the following: 5 min at 95°C followed by 35 cycles of 30 sec at 95°C, 1.10 min at 50-56°C (in function of the marker), 1.20 min at 72°C, followed in turn by 7 min at 72°C and then held at 4°C. For the nuclear regions and the GAI gene, the temperature conditions used were those recommended by Salmaso *et al*. (2004) and Wen *et al*. (2007), respectively. Positive and negative controls were used as references. The PCR-derived fragments were resolved on 2% agarose/TAE gels and visualized under UV light using ethidium bromide staining.

Table 2. List of primers used for each chloroplast and nuclear marker with their nucleotide sequence, amplicon length and references.

Marker	Length (bp)	Primer name	Primer sequence (5'-3')	Ta (°C)	References
rps16	956	rps_F	GTGGTAGAAAGCAACGTGCGACTT	56	Oxelman *et al*., 1997
		rps_R	TGCGGATCGAACATCAATTGCAAC		Oxelman *et al*., 1997
rpl32 intron	1377	rpl32_F	CTGCTTCCTAAGAGCAGCGT	50	Shaw *et al*., 2007
		rpl32_R	CAGTTCCAAAAAAACGTACTTC		Shaw *et al*., 2007
trnH-psbA IGS	460	psbA3'f	GTTATGCATGAACGTAATGCTC	56	Sang *et al*., 1997
		trnHf	CGCATGGTGGATTCACAATCC		Tate and Simpson, 2003
trnT-trnL IGS	1016	trnTUGU2F	CAAATGCGATGCTCTAACCT	56	Cronn *et al*., 2002
		5'trnLUAAR	TCTACCGATTTCGCCATATC		Taberlet *et al*., 1991
atpB-rbcL IGS	927	atpB-rbcL_F	AACACCAGCTTTRAATCCAA	56	Chiang *et al*., 1998
		atpB-rbcL_R	ACATCKARTACKGGACCAATAA		Chiang *et al*., 1998
trnL-trnF IGS	406	trnL_UNIE	GGTTCAAGTCCCTCTATCCC	50	Taberlet *et al*., 1991
		trnL_UNIF	ATTTGAACTGGTGACACGAG		Taberlet *et al*., 1991
GAI	761	GAI_F	ATGGATGAGCTTCTCGCTGT	50	Wen *et al*., 2007
		GAI_R	TAGAAGTGCATCTGRAGAAT		Wen *et al*., 2007
IF01	607	if01_F	ATGGCTGGCAATCAGGAAGG	60	Salmaso *et al*., 2004
		if01_R	GCCTTGTTGAGCTCCAACAC		Salmaso *et al*., 2004
IB02	481	ib02_F	AAGATTCTTCTGACAACCGGC	60	Salmaso *et al*., 2004
		ib02_R	GCTTGTTGAATACCTCCATCC		Salmaso *et al*., 2004
ID04	419	id04_F	CACCAGTCCCTTACCAGTCT	55	Salmaso *et al*., 2004
		id04_R	CAGTAGAGGAACACAACTGAG		Salmaso *et al*., 2004
IIC08	418	IIc08_F	CAAGGCCTTCTCTTCGTACC	60	Salmaso *et al*., 2004
		IIc08_R	AAGAATTCATATCGCCGACC		Salmaso *et al*., 2004

All amplification products were purified enzymatically by digestion with Exonuclease I and Shrimp alkaline phosphatase (Amersham) and then directly sequenced bidirectionally according to the original Rhodamine terminator cycle sequencing kit (ABI PRISM Applied Biosystems). Only in one case, EST IF01, the sequencing was carried out using only the Reverse primer because of the presence of a long poly-T close to the Forward priming site. In presence of bad quality sequences, a second PCR was conducted. When the sequence quality was poor, the PCR amplification and sequencing steps were repeated.

Character-based analysis

All the obtained nuclear sequences were visualized and manually edited by means of Sequencer 4.8. Nucleotide sites in which only a single nucleotide (=character state, CA, according to the DeSalle's terminology; DeSalle *et al.*, 2005) per site was detected were considered homozygous, whereas when two CAs per site were found the position was considered heterozygous and recorded using the IUB (International Union of Biochemistry) conventional code for degenerate bases. Sequence similarity search was performed using GenBank BLASTn algorithm (http://www.ncbi.nlm.nih.gov/BLAST) against the nucleotide databases of NCBI to check the correspondence between the sequences of the obtained amplicons with the expected sequences. Data analysis for the combined nuclear sequences was carried out for only three out the five markers studied (GAI, ID04 and IIC08). At the moment, the IF01 and IB02 ESTs were not included in the analysis to avoid problems of wrong base calling, made by eye, in correspondence of ambiguous heterozygous sites extremely frequent for these two sequences. Multiple sequence alignments were performed by SeAl version 2.1 software and, since the intrinsic difficult of the DNA barcoding applied at subspecies and population level, the traditional phenetic approach was substituted by the character–based method (Sarkar *et al.*, 2002). Analysis of polymorphisms distribution was performed using Mega version 4.1 to display the aligned combined sequence data and to highlight all the variable sites. To simplify data visualization, all the monomorphic nucleotide positions were excluded from the analysis and kept only those showing a SNP. The information about SNP occurrence were adopted to generate by eye a map with the haplotype reconstruction. to use very short sequences in order to make unlikely the occurrence of recombining events. In addition we defined an haplotype also in presence of heterozygous sites that were dealt as functionally haploid SNPs, *i.e.* without separating the two alleles found for each heterozygous polymorphic position and recording it with the IUB code. The presence of specific character states and combination of character states was evaluated as distinctive of a particular cultivar or, more generally, of a group of cultivars within *V. vinifera* species. The terms *pure*, *simple* and *compound* were employed in agreement to DeSalle's terminology (DeSalle *et al*., 2005): pure to indicate a CA shared among all the individuals belonging to an haplotype and absent form the others, simple to describe a CA narrowed to a single nucleotide position and compound for a combination of particular CAs at determined multiple nucleotide positions.

Results

Nature and frequency of SNPs detected by sequencing

The initial approach was testing the most variable chloroplast regions. The first choice regarded the employment of the *trnH-psbA* intergenic spacer that proved to be the most informative marker within the *Phaseolus* species and in other several taxa (Kress and Erickson, 2007). Once the marker was amplified for all the accessions, it was evident that the sequence was not as much variable as it was hypothesized, but it resulted to be not only monomorphic among different cultivars, but also scarcely variable among *Vitis* spp., with a number of SNPs equal to 0 and 2 when comparing *V. vinifera* cultivars and *Vitis* spp., respectively. The almost complete absence of polymorphism, even among different species of *Vitis* genus, led us to further scavenge the chloroplast genome in order to find other markers with a more appropriate mutation rate for grapevine barcoding. Therefore other six sequences, chosen among the most common markers for angiosperms phylogeny were investigated, the coding region *rps16*, the *rpl32* intron and four intergenic spacers, *trnH-psbA*, *atpB-rbcL*, *trnT-trnL* and *trnL-trnF* (Soejima and Wen, 2006; Shaw *et al*., 2007). The regions were tested only in a subset of accessions, with representatives of every species and with also thirty samples within *V. vinifera*, but an unexpected lack of polymorphisms was found both at the intraspecific and interspecific level (data not shown).

These results led us to move beyond the chloroplast and investigate the nuclear genome, whose analysis in the last decades became really common since it is a recombining and byparentally inherited DNA that allows to shift from the gene trees to multi-locus study of population history (Hare, 2001). Five markers were chosen among 50 gene fragments, considered putative single-copy genes on the basis of a previous study evaluating the degree of polymorphisms of *V. vinifera* by SSCP and sequencing techniques (Salmaso *et al*., 2004). In total 2686 nucleotides from ncDNA were amplified for each accessions (no indels were recovered), but only three regions, GAI, ID04 and IIC08, for a total of 1598 base pairs, were used for the final calculation of the SNP frequency. We encountered some difficulties for scoring the chromatograms of the IB02 and IF01 ESTs because of the presence of several cases of additivity that could not be considered certainly heterozygous. Since the SNP occurrence, both in state of homozygosis and heterozygosis, has to be detected with an high degree of confidence in order to infer the haplotype composition suitable for identification aims, we limited our focus to the regions with no case of ambiguous base calling.

Character-based DNA barcodes specific of cultivars

When comparing all the genotypes, a total of 59 and 53 polymorphic sites in 1598 bp of genomic sequence were counted among *Vitis* spp. and within *Vitis vinifera* species, respectively, with an average frequency of one SNP for every 26.77 bases and 29.3 bases, respectively. Considering the single region individually, the average frequency of CAs occurrence resulted equal to one SNP for every 50.73, 20.95 and 23.22 nucleotides for the region GAI, ID04 and IIC08, respectively, at the intraspecific level and one SNP for every 42.27, 19.95 and 20.9 nucleotides, respectively, at the interspecific level (**Table 3**). On the basis of previous phylogenetic information, the whole sampling was divided in four sub-populations (i) the international cultivars; ii) the local varieties; iii) the interspecific hybrids, Perla, and Bianca, two *V. vinifera* backcross with introgressed genes from non-*vinifera* ancestors, and Tintoria, and iv) the five *Vitis* spp., and the genetic diversity, estimated within each population, was equal to 0.007, 0.0003, 0.0014 and 0.0041, respectively. The genetic distance between the populations was 0.0032 between local and international cultivars, 0.0051 and 0.0021 between putative hybrids and, respectively, international cultivars and local varieties, and 0.0093, 0.0069 and 0.0027 between the outgroups and, respectively, international cultivars, local varieties and hybrids.

Table 3. Information including sequence length of amplicons, number and frequency of SNPs at inter- and intra-specific level and number of haplotypes (Hn) for each nuclear barcodes and for the combined sequence of three regions.

	Lenght (bp)	**No. SNPs**		**Frequency (1SNP/bp)**		**Hn**	
		Vitis spp.	*V. vinifera*	*Vitis* spp.	*V. vinifera*	*Vitis* spp.	*V. vinifera*
GAI	761	18	15	42.27	50.73	23	18
ID04	419	21	20	19.95	20.95	33	28
IIC08	418	20	18	20.9	23.22	14*	11
Combined	1598	59	53	26.77	29.3	67	62

*, missing data.

The number and the composition of haplotypes were derived without the employment of any software because of the difficulty of the programs to work on data file with heterozygous sites and their feature to provide only the most probable haplotypes using statistical algorithms (Table 3). Thanks to the large number of polymorphic sites, it was possible defining a distinct haplotype for unambiguously recognizing each one of the five species of *Vitis*, even if not always the whole combined sequence was available. Considering each single gene individually and excluding the non-*vinifera Vitis* that belong to a specific haplotype on the basis of each marker, the number of haplotypes among grape cultivars and inter-specific hybrids were equal to 18, 28 and 11 for GAI, ID04 and IIC08, respectively, without taking into account the situations were missing data could lead to ambiguous results. When the whole combined sequence was analyzed, all the genotypes, *V. vinifera* cultivars and hybrids, could be divided in at least 63 haplotypes, constituted by one to eight accessions, on the basis of the complete combined nucleotide sequence. **Table 4** shows the character state at all 53 polymorphic nucleotide positions among cultivars, in particular 15, 20 and 18 CAs for GAI, ID04 and IIC08 marker respectively, along with the frequency of each allele per position. Since our accessions are cultivars under strict selection and thus do not represent a random sampling of grapevine populations that follows the equilibrium of Hardy-Weinberg and also for most of the cultivars a single clone was present, all the variable sites were considered, regardless of the restrictive definition that consider a SNP only if the frequency of the most common allele is less than 0.95 (that means it could not be considered informative in a population analysis). In five situations, when multiple individuals were collected for a cultivar, we have never experienced intracultivar variability, but the CAs were shared among all the representatives of the cultivar. This situation happened for Sultanina, Carmenere, Malbech, Cannonao and Sagrantino cultivars, each of them counted two specimens that shared the same polymorphisms. For four of them these CAs allowed to define a cultivar-specific haplotype, whereas for the Carmenere cultivar its haplotype composition was in common with other four different cultivars, Sauvignon Bianco, Schiavetta Doretta, Albana, Piedirosso and Cabernet Lispida.

Table 4. For three nuclear markers, GAI, ID04 and IIC08, information about character state and allele frequency (%) included in parenthesis for each polymorphic position.

Marker	SNP position											
GAI	**156**	**185**	**227**	**232**	**240**	**250**	**284**	**331**	**365**	**464**	**511**	**569**
	G (98.7) R (1.3)	C (98.7) T (0.6) Y (0.6)	C (99.35) Y (0.6)	T (99.35) Y (0.6)	C (98.1) M (1.9)	C (72.4) Y (24.35) T (3.2)	T (99.35) W (0.6)	T (99.35) K (0.6)	C (99.35) Y (0.6)	S (5.8) C (0.9)	C (75.6) Y (21.15) T (3.2)	C (98.7) M (1.3)
	589	**595**	**601**									
	T (97.4) Y (2.6)	G (96.15) A (2.6) R (1.3)	C (99.35) Y (0.6)									
ID04	**28**	**35**	**139**	**140**	**168**	**216**	**227**	**233**	**253**	**263**	**286**	**287**
	G (75.5) K (24.5)	A (73.3) R (22.3)	S (49.7) C (35.55)	A (96.8) R (3.2)	T (99.35) Y (0.6)	A (98.1) M (1.9)	Y (72.9) C (26.45)	A (98.1) M (1.9)	A (98.1) R (1.9)	A (94.2) W (5.2)	G (93.5) S (5.8)	A (99.35) R (0.6)
	316	**327**	**332**	**333**	**345**	**355**	**358**	**376**				
	A (98.7) W (1.3)	A (75.3) R (24) G (0.6)	G (99.35) R (0.6)	K (50) G (40.9) T (9.1)	G (99.3) K (0.7)	G (99.3) R (0.7)	A (99.3) R (0.7)	C (98.6) Y (1.4)				
IIC08	**7**	**13**	**28**	**50**	**53**	**62**	**95**	**125**	**139**	**181**	**193**	**205**
	C (65) Y (20.4) T (14.6)	C (99.4) Y (0.6)	C (99.4) Y (0.6)	G (94.3) A (3.8) R (1.9)	T (99.4) G (0.6)	G (99.4) R (0.6)	C (98.1) Y (1.25) T (0.6)	T (97.5) W (2.5)	C (99.4) Y (0.6)	T (98.7) Y (1.25)	T (97.5) Y (1.9) C (0.6)	C (99.4) S (0.6)
	211	**299**	**301**	**329**	**349**	**376**						
	A (98.7) W (1.3)	A (75.3) R (24) G (0.6)	G (99.35) R (0.6)	K (50) G (40.9) T (9.1)	G (99.3) K (0.7)	G (99.3) R (0.7)						

Table 5. Consensus sequence related to the 53 individual SNPs detected in the three target nuclear regions with information on the haplotypes found across all grapevine (*Vitis* spp.) entries. The number of entries corresponding to local grapes are included in parentheses (see next page).

Genotypes	Hp	GAI	ID04	IIC08
#622_Alphonse	01	GCCCTCYTTCCYCTGCGC	GAGCATACAAAGAAAGGGGAC	CCCGTCGCCTCTTCGGTGTC
#617_Palieri		....	....	....
#730_Roditis		....	....	....
#705_Aledo		....	....	...................?
#746_Tinta Barroca		....	....	....
#751_Antao Vaz		....	....	....
#706_Macabeo		....	..................???	....
#592_Croatina	02	....	....................Y	....
#615_Montepulciano	03	....	..........W..........	....
#724_Sultanina	04	......T....T......	....	....
#723_Sultanina		......T....T......	....	....
#562_Garganega	05	....	...S............K....	Y...............K...
#623_Prosecco Lungo		....	...S............K....	Y...............K...
#611_Malbo Gentile	06	....	...S............K...Y	Y...............K...
#563_Manzoni Bianco	07	......C....C......	KR.S...Y..W...R......	....
#627_Riesling Renano		......C....C......	KR.S...Y..W...R......	....
#589_Franconia	08	......C....C......	KR.S...Y......R......	....
#725_Xinomauro		......C....C......	KR.S...Y......R......	....
#522_Raboso Piave		......C....C......	KR.S...Y......R......	....
#523_Raboso Piave		......C....C......	KR.S...Y......R......	....
#552_Raboso Piave		......C....C......	KR.S...Y......R......	....
#557_Chardonnay		......C....C......	KR.S...Y......R......	....
#519_Friularo		......C....C......	KR.S...Y......R......	....
#520_Friularo		......C....C......	KR.S...Y......R......	....
#518_Friularo		......C....C......	KR.S...Y......R......	....
#517_Friularo		??????????????????	?????????????????????	....
#612_Moscato Giallo	09	......CW...C......	....	....
#588_Moscato Bianco	10	......C....C......	....	....
#608_Moscato Sardo		......C....C......	....	....
#731_Moscomavro		......C....C......	....	....
#733_Moscofilero		......C....C......	....	....
#554_Barbera		......C....C......	....	....
#586_Primitivo Gioia		......C....C......	....	....
#502_Pignola		......C....C......	....	....
#749_Rabigato		......C....C......	....	....
#647_Feteasca Alba		......C....C......	..................???	..................??
#503_Corbinona	11	......C....C......	..........W..........	....
#504_Corbinella		......C....C......	..........W..........	....
#745_Tinto Cao	12	......C...SC......	....	....
#727_Korintos	13	.....RC....C......	....	....
#524_Raboso Veronese	14	......C....C......	...S............K.R..	Y...............K...
#558_Raboso Veronese	15	......C....C......	...S............K....	Y...............K...
#521_Friularo7		......C....C......	...S............K....	Y...............K...
#607_Trebbiano Toscano		......C....C......	...S............K....	Y...............K...
#583_Refosco		......C....C......	...S............K....	Y...............K...
#514_Pattaresca		......C....C......	...S............K....	Y...............K...
#510_Marzem. Cenerenta		......C....C......	...S............K....	Y...............K...
#511_Marzemina Nera		......C....C......	...S............K....	Y...............K...
#509_Marzemina Bianca	16	......C....C......	...G............T....	Y...............K...
#640_Arneis		......C....C......	...G............T....	Y...............K...
#728_Razaki	17	......C....C......	...G............T....	....
#737_Regina		......C....C......	...G............T....	....
#736_Regina Inzolia		......C....C......	...G............T....	....
#633_Calabrese	18	......C....C......	...S............K....	....
#722_Nero Avola		......C....C......	...S............K....	....
#626_Colorino		......C....C......	...S............K....	....
#726_Aghorghitiko		......C....C......	...S............K....	....
#513_Negrara		......C....C......	...S............K....	....
#628_Ciliegiolo		......C....C......	...S............K....	....
#632_Canaiolo Nero		......C....C......	...S............K????	....
#744_Turiga National		......C....C......	...S............K....	....

#505_Merlot	19	......C....C......	...S............K....	...A................
#506_Merlot		......C....C......	...S............K....	...A................
#553_Merlot		......C....C......	...S............K....	...A................
#721_Malvasia Istriana	20	......C....CM.....	...S............K....	....
#593_Cannonau	21	......C...SC......	...S............K....	....
#564_Nebbiolo	22	......C....C......	...S......W.....K....	....
#642_Corvina	23	.....MC....C...Y..	...S......W.....K....	....
#561_Sauvignon Blanc	24	......C....C......	...S............K....	T...............G...
#515_Schiavetta Doretta		......C....C......	...S............K....	T...............G...
#604_Albana		......C....C......	...S............K....	T...............G...
#629_Piedirosso		......C....C......	...S............K....	T...............G...
#508_Cabernet Lispida		......C....C......	...S............K....	T...............G...
#601_Carmenere		......C....C......	...S............K....	T...............G...
#594_Carmenere		......C....C......	...S............K....	T...............G...
#565_Tocai	25	......C....C......	...G............T....	T...............G...
#559_Cabernet Franc	26	......C....C......	....	T...............G...
#649_Feteasca Neagra		......C....C......	....	T...............G...
#603_Malvasia Chianti	27	......C...SC......	....	Y...............G...
#609_Petit Verdot	28	......C....C......	....	Y...............K...
#635_Bovale Sardo		......C....C......	....	Y...............K...
#590_Fiano		......C....C......	....	Y...............K...
#587_Picolit	29	.....RC....CM.....	....	Y...............K...
#631_Greco	30	......C....C......	..........T..........	Y...............K...
#512_Marzem. Nera Bast.	31	......C....C......	KR.S...Y......R......	Y...............K...
#569_Pinot Gris		......C....C......	KR.S...Y......R......	Y...............K...
#570_Pinot Noir		......C....C......	KR.S...Y......R......	Y...............K...
#556_Pinot Noir		......C....C......	KR.S...Y......R......	Y...............K...
#568_Pinot Blanc		......C....C......	KR.S...Y......R......	Y...............K...
#625_Riesling Italico	32	......C....C......	KR.S...Y......R......	T...............G...
#637_Falanghina		......C....C......	KR.S...Y......R......	T...............G...
#738_Matilde		......C....C......	KR.S...Y......R......	T...............G...
#650_Mustoasa		......C....C......	KR.S...Y......R......	T...............G..?
#740_Malbech	33	......C....C......	KR.S...Y......R......	...A................
#739_Malbech		......C....C......	KR.S...Y......R......	...A................
#643_Freisa	34	......C....C......	KR.G...Y......R.K....	Y...............K...
#605_Traminer_aromatico	35	......C....C......	KR.G...Y......R.K....	....
#606_Vermentino		......C....C......	KR.G...Y......R.K....	....
#732_Robolla		......C....C......	KR.G...Y......R.K....	....
#616_Vittoria	36	....	KR.G...Y......R.K....	....
#619_Italia	37	....	KR.G...Y......R.K....	T...............G...
#708_Mencia	38	....	KR.S...Y......R......	Y...............K...
#621_Cardinal	39	....	KR.S...Y......R......	....
#582_Sagrantino	40	....	....	T...............G...
#634_Sagrantino		....	....	T...............G...
#747_Tinta Francisca	41	......C....C..A...	....	...R................
#716_Pedro Ximenez	42	......C...SC..A...	................K....	....
#715_Cannonao	43	......C...SC..A...	...S............K....	....
#714_Cannonao		......C...SC..A...	...S............K....	....
#646_Tocai Rosso	44	......C....C..R...	...S............K....	....
#743_Touriga Franca		......C....C..R...	...S............K....	....
#645_Vernaccia	45	......C....C......	??.G............T....	....
#507_Agostana		......C....C......	...G............T....	....
#756_Bastardo		......C....C.....?	??.G............T....	....
#501_Gatta	46	......C....C......	...G............T....	Y...............K...
#641_Teroldego	47	......C....C......	...S.................	Y...............K...
#710_Blanca Cayetana	48	......T....T......	...S............K....	....
#711_Parda		......T....T......	...S............K....	....
#742_Trincadeira	49	...........C......	....	....
#709_Xarello	50	....	...G............T....	....
#720_Monastrel	51	....	...S............K....	....
#712_Bobal		....	...S............K....	...................?
#630_Prosecco Balbi	52	......C...........	...S............K....	Y...............K...
#584_Ribolla	53	...........C......	...S............K....	....
#719_Graciano		...........C......	...S............K....	....
#718_Parellada	54	..........SC......	...S............K....	...R................
#602_Malvasia Nera	55	..........SC......	...S............K....	T...............G...
#748_Vinao	56	....	...G...Y......R.K....	...A................
#729_Asirtiko	57	....	...S............K....	T...............G...

#555_Cabernet Sauvignon	58	....	....	Y..R............K...
#528_Gruajo	59	......C....C......	KR.G..MYMR...W..K....	Y.......YW.YY..KK..M
#516_Tintoria	60	......C....C......	KR.G..MYMR.S.W..K....	Y.......YW.YY..KK..M
#618_Perla	61	RY.YY.C.KY.C.Y....	TR.G.Y.T...C...R.K.R.	.YY.G.R.TWY.CSR..RCA
#535_bianca	62	.T....C....C......	YR.G..M.MR.SR...K....	........YW..Y......M

#532_*V. berlandieri*	63	RY.YY.C.KY.C.Y..S.	TGRG.Y.T...C..AR.Y.R.	Y...KYR.TWY.CSR..RCA
#534_*V. labrusca*	64	.T....C....C......	...G..MCMR.SR.A.K....	........YW..Y......M
#531_*V. rupestris*	65	A..TC.C.GY.C.C....	TG.G...T...C..A..T...	..T.G...TA..CYA...CA
#530_*V. riparia*	66	A..TC.C.GT.C.C...G	TGRG...T...C..A..T...	T....T..TA..C.A...CA
#533_*V. cinerea*	67	.TS...C....C......	CCCGKCR.TWYTCSG.T.Y.?	????????????????????

#624_Aglianico	na	......C....C......	?????????????????????	T...............G...
#620_Moscato Amburgo	na	....	KR.S...Y......R......	????????????????????
#752_Arinto Armas	na	......C....C......	?G.GR..Y...S..R..????	...................?
#551_Tramier	na	......C....C......	KG.G...Y......G..????	....
#567_Lambrusco Maestri	na	......C....C......	??...................	................K...
#610_Carignan	na	......T....T......	??.G.......S.....????	....
#741_Fernao Pires	na	...........C......	.G.GR..Y...S..R..????	....
#703_Tempranillo	na	....	?G.GR......S....T.???	....
#701_Tempranillo		....	?G.GR......S....T.???	....
#702_Tempranillo		....	?????????????????????	....
#707_Albarino	na	......C....C.....?	.............????????	....
#755_Tempranino	na	......C....C......	?????????????????????	....
#566_Trebbiano Romagnolo	na	??????????????????	...S............K....	....
#717_Pedro Ximenez	na	...........C......	...S............S????	....
#636_Negro Amaro	na	....	??..............K....	....
#639_Rondinella	na	......C....C......	?G.G.......S.....????	....
#613_Grechetto	na	......C....C......	?.........W..........	....
#638_Verduzzo Friulano	na	......C...........	...S............K....	T..................?
#754_Malvasia Fine	na	......C....C......	??.....Y......R......	....
#753_Alfrocheiro	na	......C....C......	???....Y......R......	....
#648_Feteasca Regala	na	......C....C......	....	................G..?
#704_Tinta Fina	na	....	.R.GR......S....K????	...................?
#750_Castelao	na	....	?................????	????................
#591_Dolcetto	na	......C...SCM.....	??..............K....	...................?
#560_Sangiovese	na	??????????????????	?????????????????????	T...............G...

na, haplotype not available because of missing data

Rarely a simple pure CA was identified as peculiar of a cultivar, as in the case of Moscato Giallo that is the only cultivar showing an heterozygous site in position 284 of the GAI gene, indicated by the degenerate nucleotide W. In contrast, frequent compound CAs could be detected, and at least 38 unique cultivar-specific haplotype were discovered, on the basis of the complete combined sequence. All the other haplotype groups, instead, did not identify a single cultivar, but they clustered several modern varieties, with a maximum of 8. Within these complex haplotypes it was difficult to find a correlation among the cultivars because the clusters often grouped very far varieties, with no common history. For example, in the case of haplotype grouping Bovale Sardo, Fiano and Petit Verodt, the three cultivars did not share neither the geographic origin, the first from Sardinia, the second from Campania and the third a French cultivar spread in Veneto and Lazio, or the berry colour (Table 1). In other cases, the haplotypes grouped very close varieties, such as in the case of Regina, where it was impossible to distinguish Razaki, a Regina from Greece, from Italian Regina and Regina Inzolia. Similar results were obtained in the case of Pinot family, where the two accessions of Pinot Noir (570 and 556), Pinot Blanc and Pinot Gris showed the same CAs pattern or for the group of Moscato that included Moscato Bianco, Moscato Sardo and other two closely related cultivars, Moscomavro and Moscofilero. In only one case the DNA typing was able to distinguish two close cultivars: in fact within Prosecco group, a CA in position 250 of the GAI gene allowed to discriminate between Prosecco lungo e Prosecco Balbi that are two different biotypes of Prosecco. In addition, in two cases, Bianca and Perla, a particular genetic haplotype, more similar to non-*vinifera Vitis* species because of the presence of several positions highly heterozygous, was found. The nucleotide composition of these two haplotypes are consistent with the origin of the cultivars that are the result of two separately events of interspecific hybridization. In particular, Perla is an interspecific hybrid between Villard Noir cultivar, a French hybrid grape, x *V. vinifera*, and Bianca, even if can still be considered belonging to *V. vinifera*, owns a more complex pedigree. In fact Bianca is the result of a backcross of the *V. vinifera* cultivar Villard Blanc with the ancestors of Villard Blanc that include accessions of five *Vitis* species, *V. aestivalis, V. berlandieri, V. cinerea, V. lincecumii* and *V. rupestris,* in order to introduce in this cultivar the resistance genes owned by the North America grapes (Csizmazia and Bereznai, 1968 cited by Bellin *et al*., 2009).

Finally, it is worth mentioning that in our reconstruction of diagnostic haplotypes we only employed samples with data at all loci to ensure the set of diagnostic SNP were conserved across accessions. For example, in the case of Tempranillo cultivar, we removed three clones, because none of them had the complete nucleotide sequence and therefore, even if the CAs available were not in disagreement and could suggest an identical haplotype composition shared among the three entries, as happened for Pinot or Regina groups, the missing data affected the results. In total, 21 situations with missing data, attributable to the lack of partial or complete sequence of one or more markers, were recovered. Comparing the sequence of only the GAI region, an other haplotype could be found out because the cultivar Dolcetto showed a typical CAs composition, absent from the other cultivars and characterizing the variety. Similar results could be obtained comparing only the ID04 sequence, and other three new haplotypes could be added, exactly Tramier, Tinta Fina and the Tempranillo clones, while using only the IIC08 any additional haplotype could not be recovered.

Testing the local varieties

Once we established diagnostic haplotypes on the basis of the international references, we tested their utility on some local varieties as study case in order to clarify some genetic relationships among cultivars and resolving eventually situations of synonymy and homonymy.

In the case of Merlot we collected three different individuals, one certified and two local, and all of them shared four CAs specific for that cultivar and absent from all the others. Therefore, comparing the local pattern with the reference standard, we were able to confirm both the CAs pattern unique for the Merlot cultivar and the genetic identity of the local varieties. A second case regards the group of Rabosi, Raboso Piave and Raboso Veronese, and Friularo. In our reference system we had the accessions 552 and 558 corresponding to Raboso Piave and Raboso Veronese, respectively, and they could be distinguished by the belonging to two different haplotypes. When the cultivars from local collection were also added to the analysis their clustering was in accordance with the haplotype composition of the reference standards and in fact the two samples labelled as Raboso Piave, 522 and 523, went to group with 552_Raboso Piave, thus conferming the

SSR results of Salmaso *et al*. (2008), while the local 524_Raboso Veronese was identical to 558_Raboso Veronese, except for one nucleotide site. In addition, the Friularo cultivar was collected and 5 different clones from as many farmers were sampled. From the haplotype reconstruction, it emerged that four out the five clones grouped together in the same hapotype including 552_Raboso_Veronese, while the 521_Friularo7 grouped with 558_Raboso Piave. Other SSR result, confirmed by nuclear DNA barcoding, was the genetic identity among the variety Marzemina Nera and Marzemina Cenerenta, that were different from Marzemina Bianca and Marzemina Nera Bastarda, and among Corbinona and Corbinella that resulted to share the same haplotype. Finally, a last observation regards the two accessions Tintoria and Gruaja. The former exhibited a unique haplotype, highly heterozygous and with many nucleotide sites coomon to *V. labrusca* accession, and the latter revealed a nucleotide composition identical to Tintoria except for one position.

Discussion

Developing a reference system by mean of DNA barcoding

The use of DNA barcoding to test the genetic distinctiveness of grapevine cultivars, and more in general crop varieties, is a novel application of the technique that touches the border-line of its potentials. In fact, DNA barcoding was initially proposed as a diagnostic tool to determine the species identity of an unknown organism. In this paper, it was tested its ability to distinguish modern varieties within *V. vinifera* species, an application that is of huge economic relevance due to the agronomic importance of the crop. A further test was trying to characterize also within the same cultivar different biotypes. The concept of biotype employed in the study is referred to a genotype that differentiated geneticcaly from the original cultivar through occurrence of gemmary mutation, epigenetic effect or their combination, determining the acquisition of a new specific morphological or physiological trait.

The analysis of 144 grapevine cultivars was performed by the character-state method because the application of the conventional phenetic approach is unsuitable for an assay below the species level. Distinguishing genetic entities below the species level requires a more sensitive approach able to conserve all sequence information without converting them in genetic distance. Further, the balance sought for DNA barcode markers is such that within-species genetic diversity is minimized, but in this study it was of principal importance. Thus we combined DNA barcode methods with more intensive DNA fingerpringing using SNP to better define the boundaries among important agronomic cultivars. DNA barcode loci will continue to be important in defining species boundaries, but will be supplemented with SNP data reported here for the purpose of diagnostic traceability of varietal genotypes.

The first attempt of discovering genetic diversity among cultivars was conducted on the chloroplast genome, but it was not sufficiently variable to allow the distinction of crop varieties. The alternative genome for barcoding aims is the nuclear one that shows synonymous substitution rates generally greater than plastid and mitochondrial genes (Wolfe *et al*., 1987). In addition, the nuclear DNA offers the advantage to resolve problems associated with horizontal acquisition of organelles through hybridization events or with introgresson patterns that can be detectable only using byparental markers (Chase *et al.*, 2005). An intrinsic problem of using nuclear sequences is the difficulty of interpreting the frequent occurrence of additivity cases that can often lead to situations of misinterpretation. Since we are working with *V. vinifera* species, that is a diploid species highly heterozygous, frequent cases of intragenomic variation were detected and they could arise because of the presence of more than one allele variant for a particular locus. A second issue is that an haplotype is defined for a non-recombining and haploid genome (Stephens *et al*., 2001). An haplotype is defined as a combination of alleles of closely linked loci on a chromosome or a combination of nucleotide sites linked on the same allele or chromosome that tend to be inherited together. The key issue is that the set of alleles or sites have to be statistically associated on the same chromosome to form a unique linkage group without recombination events and they have to derive from an haploid state, such as the sperm or egg cell or from the cytoplasmic DNA. The employment of haplotype reconstruction to data from nuclear genome only works when the genetic variation is fixed among varieties, including heterozygous states. Generally, in presence of heterozygous sites, it would necessary the separation of the allele variants and the definition of the nucleotide associations for the polymorphic sites. In contrast, in the specific case of *V. vinifera* cultivars, since they are asexually propagated and thus the recombination issue is negligible, the genetic patrimony is fixed allowing the definition of an haplotype independently by the marker distribution on chromosomes. Therefore, the inference of haplotypes from a diploid genome is possible and requires some statistical programs that give a probabilistic definition of haplotypes without the necessity to split the two allelic variants. In the barcoding approach aimed to the variety characterization, since the variety identification requires an unambiguously SNP detection, we decided to carry out a visual inspection on the global sequence alignment to recover the exact haplotype combinations. For this goal, two out of the five nuclear regions amplified were discarded because of too much intragenomic variability.

Out of the 68 haplotypes discovered, five were able to distinguish the *Vitis* spp. and 38 were cultivar-specific, such as for Merlot, Sultanina, Tempranillo, Malbech and Sagrantino, to cite only those with more than one specimen, an interesting result if we think that only 1598 nucleotides were analyzed. Among them, the haplotype composition of the two accessions Perla and Bianca confirms the phylogenetic origin of the two cultivars that are interspecific hybrids with other non-*vinifera Vitis*. An other noteworthy example is the local cultivar Tintoria that was suggested to be an interspecific hybrid with non-*vinifera Vitis*. In fact this cultivar, on the basis of chloroplast SSR markers, showed an haplotype common with American grapevine species (Salmaso *et al*., 2008) and now, the nuclear DNA barocoding seems to further support this hypothesis, even if it is impossible to confirm certainly because too few CAs were available. The other haplotypes, instead, were more complex and they grouped several cultivars that do not seem to have a common history.

Distinguishing among very close varieties, such as Pinot, Moscato and Regina groups, or biotypes, such as Friularo that is considered a biotype of Raboso Piave adapted to Euganean area remains challenging. In the case of Pinot family, it includes the original variety, Pinot Noir, with black berry and the two varieties, Pinot Gris and Pinot Blanc, that are thought to be chimeras, mutant clones derived from the Pinot Noir after the occurrence of a mutation for the berry colour in one cell layer of the berry for the Pinto Gris and in both the cell layers for the Pinot Blanc (white berry). These kinds of somatic mutations are very common in grapevine and contribute to the high incidence of genetic variability. Since the origin of this mutation, probably the only way to resolve the genetic recognition of these three cultivars could be the individuation of a marker mapping on the gene controlling the berry colour and the mutation responsible of the colour change. Thus there are important limits to the resolution we may obtain with genetic markers alone. Even in presence of these multi-varieties haplotypes, some of them allowed to further corroborate some theories suggesting cases of homo- synonymy or parent-offspring relationships. For example, the two cultivars, Nero d'Avola and Calabrese, are known to be synonymous and the haplotype composition put them together even if also with other varieties, while the cultivars Alphonse Lavallèe and Palieri belong to the same haplotype and this is explained by the fact that Palieri is the offspring of Alphonse Lavallèe x Red Malaga (not present in this study).

DNA barcoding and local cultivar

Once specific haplotypes were identified among the international cultivars used as standard references, an additional sampling of ancient local varieties typical of Northeastern Italy were included in the analysis. Characterizing this local germoplasm, that represents an incredible genetic resource for the region, would be the first step of a conservation policy aimed to the preservation and valorization of old native genotypes. The description of this local patrimony represents not only a valuable resource for the territory, since these cultivars still constitute the basis of famous regional wines, such as Raboso Piave or Marzemina, but also would allow to identify potential source of genetic variability exploitable for genetic improvement programmes (breeding program assisted by molecular markers, MAS) providing the information to correlate the genetic variability of grape cultivars with phenotypical differences. The employment of these varieties can be considered an internal test to verify the efficacy of the DNA barcoding approach in order to check the correspondence between the declared origin of the cultivars and the real genetic identity of the sample, resolving eventually cases of synonymy and homonymy, and to compare the results with those obtained previously by nuclear and chloroplast SSR markers (Salmaso *et al*., 2008).

Among the 14 local cultivars employed in this study, six are registered in the Italian Catalogue of Cultivated Varieties, Pignola, Marzemina Bianca, Raboso Piave, Raboso Veronese and Merlot. Merlot is a French variety, grown in the European area and widespread in all Italy since the XIX century, that was included in the analysis as a test case to corroborate the Merlot haplotype obtained by the international accession. Friularo, even if not registered in the Italian Catalogue, is recognized as a biotype of Raboso Piave and, on the basis of both SSR markers and DNA barcoding/fingerprinting technique, they resulted genetically indistinguishable. Also in the Raboso group, the local non-certified genotypes clustered with the correct international reference standards, confirming in this way their genetic identity with these cultivars. The other local varieties, that are not present in the Italian Catalogue are Gatta, Corbinona, Corbinella, Agostana Nera and Tintoria. Corbinona and Corbinella resulted to share the same nuclear haplotype, confirming previous results by nuclear and chloroplast SSRs that showed the synonymy between these two varieties (Salmaso *et al*., 2008). Tintoria and Gruaja are the only two local varieties with a specific haplotype not shared with other cultivars. Tintoria, as said before, is probably an interspecific hybrid, while Gruaja is an old variety whom cultivation is almost disappeared and narrowed to a small area of Vicenza province. The ancient cultivars, such as Gruaja, show an high incidence of mutations and this happens because they cannot be considered unique clones, but they are polyclonal varieties that during the years were adapted to the environment editing their genetic and thus phenotypical traits and originating specific biotypes (Valenti *et al*., 1994). Preserving the ancient varieties is fundamental for genetic improvement programs because, since it is more likely that these varieties accumulate and fix mutations than young cultivars, the high incidence of mutations can be the starting point for the origin of new alleles. The chimeric situation therefore can represent an interesting source of clonal variability from the different cell layers and its recovery might contribute to generate new agronomically useful phenotypes.

In conclusion, even if the results are preliminary, the high number of haplotypes obtained so far demonstrated that the nuclear genome is probably enough variable to function as source of diagnostic markers for traceability studies, allowing the genetic characterization of the main international and local cultivars. Anyway, DNA fingerprinting, based only on only three markers, proved to be unable to distinguish closely related accessions, such as within the Pinot family, or to reflect phylogeographic history of the biotypes, as in the case of Sultanina and Regina groups. Thus the research is still ongoing and it needs additional experimental analyses for increasing the number of sequences assayed to discover more polymorphic sites useful for defining single cultivar identity and ancestry and testing several clones for each cultivar in order to confirm the haplotype composition derived form just one genotypes for variety. Finally it will be necessary performing a more exhaustive assay of the genome and haplotype diversity and comparing DNA barcoding data with previous results regarding nature and frequency of SNPs in grapevine obtained with different molecular markers, such as microsatellites.

References

Alleweldt G, Dettweiler E (1994). The genetic resource of Vitis: world list of Grapevine collections. (2nd edn), Geilweilerhof.

Bellin D, Peressotti E, Merdinoglu D, Wiedemann-Merdinoglu S, Adam-Blondon A-F, Cipriani G, Morgante M, Testolin R, Di Gaspero G (2009). Resistance to Plasmopara viticola in grapevine 'Bianca' is controlled by a major dominant gene causing localised necrosis at the infection site. *Theor Appl Genet* 120:*163–176*.

Bessis R (2007) Evolution of the grapevine (Vitis vinifera L.) imprinted by natural and human factors. Can J Bot 85 (8): 679-690.Brumfield RT, Beerli P, Nickerson DA, and Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol Evol* 18(5): *249-256.*

Chiang TY, Schaal BA, Peng CI (1998). Universal primers for amplification and sequencing a noncoding spacer between the *atp*B and *rbc*L genes of chloroplast DNA. *Bot Bull Acad Sin* 39: *245–250*.

Collins FS, Brooks LD, Chakravarti A (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* 8(12): *1229-1231.*

Cronn RC, Small RL, Haselkorn T, Wendel J F.(2002). Rapid diversification of the cotton genus (*Gossypium*: *Malvaceae*) revealed by analysis of sixteen nuclear and chloroplast genes. *Am J Bot* 89: *707–725*.

Csizmazia J, Bereznai L (1968). A szılı *Plasmopara viticola* és a *Viteus vitifolii* elleni rezisztencia nemesítés eredményei. *Orsz. Szıl. Bor. Kut. Int. Évkönyve*, Budapest. 191-200.

DeSalle R, Egan MG, Siddall M (2005). The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philos T R Soc B* 360: *1905-1916.*

Ergül A, Türkoğlu M, Söylemezoğlu G (2004). Genetic identification of amasya (*Vitis Vinifera L.* Cvs.) genotypes based on AFLP markers. Biotechnol & Biotechnol Eq18: *39-43*

Gago P, Santiago J, Boso S, Alonso-Villaverde V, Grando MS, Martinez MC (2009). Biodiversity and Characterization of Twenty-two *Vitis vinifera* L. Cultivars in the Northwestern Iberian Peninsula. *Am J Enol Viticult* 60(3) :*293-301.*

Galet P (1952). *Précis d'Ampélographie Pratique* Impr. P. Déhan (Montpellier)

Garcia-Beneytez E, Moreno-Arribas MV, Borrego J, Polo MC, Ibanez J (2002). Application of a DNA analysis method for the cultivar identification of grape musts and experimentla and commercial wines of *Vitis vinifera* L. using microsatellite markers. *J Agr Food Chem* 50: *6090-6096.*

Grassi F, Labra M, Imazio S, Spada A, Sgorbati S, Scienza A, Sala F (2003). Evidence of a secondary grapevine domestication centre detected by SSR analysis. *Theor Appl Genet* 107: *1315–1320.*

Hare MP (2001). Prospects for nuclear gene phylogeography. *Trends Ecol Evol* 16: *707-716.*

Hayasaka Y, Adams KS, Pocock KF, Baldock GA, Waters EJ, Hoj PB (2001). Use of electrospray mass spectrometry for mass determination of grape (*Vitis vinifera*) juice pathogenesis-related proteins: a potential tool for varietal differentiation. *J Agr Food Chem* 49: *1830-1839.*

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003a). Biological identifications through DNA barcodes. *Proc R Soc Lond* B 270: *313-321.*

Hidalgo L (1993). La Vitivinicultura Americana, *In Tratado de Viticultura General*, Ediciones Mundi-Prensa, Madrid *765–808.*

Jaillon O, *et al.* (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449: *463-467.*

Jansen RK, Kaittanis C, Saski C, Lee S, Tomkins J, Alverson AJ, Daniell H (2006). Phylogenetic analysis of Vitis (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. *BMC Evol Biol* 6: 32.

Moreno-Arribas MV, Cabello F, Pollo MC, Martin-Alvarez PJ, Pueyo EJ (1999). Assessment of the native electophoretic analysis of total grape must proteins for the characterization of *Vitis vinifera* L. cultivars. *J Agr Food Chem* 47 : *114-120.*

Oxelman B,Lide M, Berglund D (1997). Chloroplast rps16 intron phylogeny of the tribe Sileneae (Caryophyllaceae). *Plant Systemat Evol* 206: *393–410.*

Pinder RM, Meredith CP (2003). Wine – A scientific exploration. Sandler, M.; Pinder, R. (eds). Taylor and Francis, UK, *260 – 273*.

Pomar F, Novo M, Masa A (2005). Varietal differences among the anthocyanin profiles of 50 red table grape cultivars studied by high performance liquid chromatography. *J Chromatogr A* 1094: *34–41.*

Rafalski A (2002a). Applications of single nucleotide polymorphisms in crop genetics. *Curr Opin Plant Biol* 5: *94–100.*

Salmaso M, Faes G, Segala C, Stefanini M, Salakhutdinov I, Zyprian E, Toepfer R, Grando MS, Velasco R (2004). Genome diversity and gene haplotypes in the grapevine (*Vitis vinifera* L.), as revealed by single nucleotide polymorphisms *Mol Breeding* 14: *385–395.*

Salmaso M, Dalla Valle R, Lucchin M (2008). Gene pool variation and phylogenetic relationships of an indigenous northeast Italian grapevine collection revealed by nuclear and chloroplast SSRs. *Genome* 51 (10): *838-855.*

Sang T, Crawford DJ, Stuessy TF (1997). Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). *Am J Bot* 84:1120–1136.

Sarkar IN, Thornton JW, Planet PJ, Figurski DH, Schierwater B, DeSalle R (2002). An automated phylogenetic key for classifying homeoboxes. *Mol Phylogenet Evol* 24(3): *388-399.*

Sefc KM, Pejic I, Maletic E, Thomas MR, Lefort F (2009). Microsatellite Markers for Grapevine: Tools for Cultivar Identification & Pedigree Reconstruction. *Molecular Biology & Biotechnology of Grapevine* Ed. Roubelakis-Angelakis KA.

Shaw J, Lickey EB, Schilling EE, Small RL (2007). Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *Am J Bot* 94: *275-288.*

Siles BA, O'Neil KA, Fox MA, Anderson DE, Kuntz AF, Ranganath SC, Morris AC (2000). Genetic fingerprinting of grape plant (Vitis vinifera) using Random Amplified Length Polymorphic DNA (RAPD) analysis and dynamic size-sieving capillary electrophoresis. *J Agr Food Chem* 48: *5903-5912.*

Siret R, Gigaud O, Rosec JP, This P (2002). Analysis of grape Vitis vinifera L. DNA in must mixtures and experimental mixed wines using microsatellite markers. *J Agr Food Chem* 50: *3822-3827.*

Soejima A, Wen J (2006). Phylogenetic analysis of the grape family (Vitaceae) based on three chloroplast markers. *Am J Bot* 93: *278-287.*

Taberlet P, Gielly L, Pautou G, Bouvet J (1991). Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Mol Biol* 17: *1105-1110.*

Tate JA, Simpson BB (2003). Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploidy species. *Syst Bot* 28: *723-737.*

Tennis MJ (1998). Recent developments in food authentication. *Analyst* 123: *151-156.*

This P, Lacombe T, Thomas RM (2006). Historical origins and genetic diversity of wine grapes. *Trends Genet* 22: *511-519.*

This P, Jung A, Boccacci P, Borrego J, Botta R, Costantini L, Crespan M, Dangl GS, Eisenheld C, Ferreira-Monteiro F, Grando S, Ibáñez J, Lacombe T, Laucou V, Magalhães R, Meredith CP, Milani N, Peterlunger E, Regner F, Zulini L, Maul E (2004). Development of a standard set of microsatellite reference alleles for identification of grape cultivars. *Theor Appl Genet* 109(7): *1448-1458.*

Wen J, Nie Z-L, Soejima A, Meng Y (2007). Phylogeny of Vitaceae based on the nuclear *GAI1*gene sequences. *Can J Bot* 85: *731–745.*

Wolfe KH, Li WH, Sharp PM (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *PNAS* 84 (24): *9054-9058.*

Valenti L, Mastromauro F, Brancadoro L, Bogoni M (1994). **A** methodology for description and evaluation of grapevine germplasm. Atti VIth International Symposium on Grape Breeding, Yalta, Crimea, 4-10 settembre 1994, 29-34.

Velasco R, *et al*. (2007). A high quality draft consensus sequence of the genome heterozygous grapevine variety. *PLoS ONE* 2(12): e1326.

Chapter 5

General conclusions

Species identification and classification have traditionally been domain of taxonomists, but since the classical methods, based on morphology, demand great skills and time and often are difficult to apply in those situations with limited phenotypical traits, recently new molecular-based approaches were developed. DNA barcoding, taxon identification using standardized DNA region, has received much attention in the last decade as a modern genomic tool able to complement the conventional methods in an integrative taxonomy approach. The Consortium for the Barcode of Life has stated: "DNA barcoding will make a huge difference to our knowledge and understanding of the natural world". The rapidity of acquisition of molecular data through PCR amplification and DNA sequencing along with the possibility to set up standard protocols are the most important advantages of the technique. In addition DNA barcoding assays can be applied in all life stages, from juvenile to adult forms and for determination of the taxonomic identity of damaged organisms or fragments (*e.g.*, food stuffs or stomach extracts), important for example in forensic science, in food traceability or in protection of the biodiversity to prevent illegal hunting of endangered animals. Although these unquestionable benefits that confer an invaluable significance to the approach, many criticisms were raised, mainly from taxonomic community that questions the theoretical assumptions on which DNA barcoding is based. The degree of genetic divergence is used as a criterion for species delimitation, *i.e.* to infer if two populations belong or not to the same species, but it can be used only in the framework of Mayr's Biological Species Concept, and thus it does not consider that the species problem is still one of the most discussed biological issues. Therefore several authors belive that DNA barcoding is just ad additional genetic key that can only identify known species and in no way can be considered a replacement of traditional taxonomic practice.

The present research inserted within this debate and intended to provide the first extensive analysis of the possible applications of DNA barcoding in the context of food authentication. In details, the project deals with the study of DNA barcoding applied to the species recognition of fish fillets, often involved in falsification cases, and the genetic distinctiveness of bean and grapevine varieties, two crop species of huge agricultural interest. The necessity of developing new analytical methods able to overpass the taxonomic impediments, *i.e.* the absence of morphological traits as in the case of fish fillets or bean and grape food derivates lost during food-processing, is essential to detect the increasing cases of food falsification.

In fish barcoding, the importance of investigating the application of the technique could be of interest not only for food traceability, detecting mislabeling in commercial processed seafood, but also for conservation policies, monitoring illegal trade of protected and endangered species. Regarding plant barcoding, bean and grapevine were employed as two different study cases, but since the conventional barcoding approach is based on the reproductive isolation, caused by the accumulation of genetic differences, as criterion of distinctiveness of two species, cannot be applied at sub-species level, it was necessary developing a different approach, focused on SNP detection. The results obtained so far confirmed the potentials of DNA barcoding technique as a powerful tool to be exploited for the genetic identification of fish species, confirming to represent a valid alternative to traditional analytical methods to identify the meat origin of seafood derivates. In contrast, the application of the technique for recognizing land plants is known to be more problematic. The technique resulted able to distinguish different species within *Phaseolus* and *Vitis* genera, while at intraspecific level it proved to be less powerful. In the case of bean, SNP markers allowed to recognize some haplo-groups within *P. vulgaris* species related to the geographical origin of the accessions, while within *V. vinifera*, although the research is ongoing, the resolution seems higher and more cultivar-specific haplotypes were discovered.

Future perspectives

The acquisition of these information will allow the development of a microarray technology, able to distinguish hundreds or even thousands of species or varieties simultaneously on the basis of a few specific SNPs, characterizing the genetic entity. Microarray technology is based on the immobilization of thousands of nucleotide sequences on a glass microscope slides. These oligonucleotide probes are complementary to the DNA target sequences to be analyzed. DNA target, which is usually fluorophore-labeled during PCR amplification, hybridises with the oligonucleotide probe on the microarray and can be detected after washing steps by its label. The technology allows the simultaneous screening of several nucleotide sequences of the same gene or different markers, making faster and more powerful the analyses. DNA microarrays, even if extensively used for analysis of gene expression, have been only recently applied for genotyping of organisms thanks to its ability to detect a specific sequence and to recognize genetic variations due to only one single nucleotide polymorphisms (SNP).

General acknowledgements

Finalmente è finitaaaaaaa!!!! Credo che scrivere la tesi di dottorato sia una fatica disumana, fatica che da sola difficilmente sarei riuscita a portare a termine. Ne consegue che è d'obbligo ringraziare tutti coloro che hanno contribuito al completamento di questa tesi!

Naturalmente i miei primi pensieri vanno al gruppo con cui ho lavorato in questi tre anni: Gianni, professore ed amico, che dal primo giorno in cui ho iniziato a lavorare con lui mi ha rassicurato ed incentivato a non lasciarmi mai intimorire e che devo ringraziare dal profondo del mio cuore per avermi spronato alla volta di quel di Washington, sfida che pensavo insormontabile, ma che, superati i primi momenti di sconforto, invece ha saputo arricchirmi come poche esperienze possono fare; i miei colleghi di laboratorio, Giulio, il mio compagno di viaggio per tutta la durata del dottorato, e Daria e Silvio, conosciuti in itinere, sempre disponibili a farsi due risate, ma che nei momenti di bisogno erano sempre lì pronti ad iutarmi (penso alle estrazioni di vite, grazie mille Darietta!! E all'editing delle sequenze di pesci e alla bibliografia, thanks Silvio, great friend! Forse non avrà lo stesso valore di un articolo, ma quello che conta è che ci sia, no?). Ringrazio anche la Prof.ssa Lucchin e tutto il Dipartimento di Produzione Vegetali che mi hanno ospitato nei loro laboratori. Infine è d'obbligo ringraziare tutti gli studenti che hanno fatto le loro comparse come tesisti durante questi tre anni,.penso a Stefano, Giulia, Tommaso e Rodolfo. Grazie amici! Un ringraziamento speciale va anche allo Smithsonian e, in particolar modo a John, Dave, Ida, Vinita e Silvana, persone eccezionali, che stimo sia da un punto di vista professionale che umano, che mi hanno da subito hanno fatto sentire a casa. Thanks a lot dear friends!!Grazie mille per avermi aiutato ad affrontare un argomento ostico come può essere la biologia evoluzionistica di cui non avevo alcun background, ma che ho imparato a conoscere pian pianino..

Per concludere vorrei ringraziare alcune persone che, anche se non direttamente coinvolte nella scrittura della tesi, sento di dover ricordare. Beh innanzitutto i miei genitori che mi hanno sempre sostenuto in tutti i momenti di entusiasmo o sconforto che ho vissuto per i risultati ottenuti. Uno speciale pensiero va a Marina e Virginia, due amiche incredibili che ho avuto l'occasione di conoscere mentre ero a Washington e con cui ho condiviso ogni singolo istante a DC..non sarebbe stato lo stesso senza di voi!

E infine un ultimo ringraziamento al mio amore, pensavi che non ti avrei ringraziato eh? Ma come potrei dopo che hai perso la vista a scrivere la mia bibliografia, anche Linneo ebbene sì!! Che per volare da me, letteralmente fino a DC, ha addirittura mancato la cerimonia in cui veniva proclamato dottore di ricerca, consacrando così tutti i suoi sforzi di dottorando..grazie amore mio!

E per concludere solo un ultima breve osservazione: mi auguro solo che alla fine di questo percorso, il raggiungimento di un traguardo importante, come può essere il conseguimento del titolo di dottore di ricerca, ripaghi delle MILLE ore trascorse al computer ad analizzare migliaia di sequenze (!!!), momenti che hanno determinato una perdita credo di 5 diottrie della mia vista!!!

The end!