Mar Biotechnol
DOI 10.1007/s10126-014-9596-6

ORIGINAL ARTICLE

Utility of DNA Barcoding for Tellinoidea: A Comparison of Distance, Coalescent and Character-based Methods on Multiple Genes

Zhenzhen Yu · Qi Li · Lingfeng Kong · Hong Yu

Received: 16 April 2014 / Accepted: 18 July 2014

Abstract DNA barcoding has become a promising tool for rapid species identification using a short fragment of mitochondrial gene. Currently, an increasing number of analytical methods are available to assign DNA barcodes to taxa. The methods can be broadly divided into three main categories: (i) distance-based methods (the classical approach and the automatic barcode gap discovery (ABGD) approach), (ii) coalescent-based methods (the monophyly-based method and the general mixed Yule coalescent (GMYC) model) and (iii) the character-based method (CAOS). This study is set out to evaluate the availability of each method in barcoding Tellinoidea on the cytomchrome *c* oxidase subunit I (COI) and the 16 small-subunit ribosomal DNA (16S rDNA) genes. As a result, the character-based method was found to be the best in all cases, especially on a genus level. For distance-based methods, the elaborate one gained a success equal or greater than the basic one. The traditional coalescent-based method nicely delimited all of the tellinoideans on a species level. The GMYC model, which is the most radical, clearly inflated the number of species units by 34.6 % for COI gene and by 58.8 % for 16S gene. Thus, we conclude that CAOS better approximates a real barcode, and suggest the use of the ABGD method and the monophyly-based method for primary partitions. Additionally, COI gene may be more suitable as a standard barcode marker than 16S gene, particularly for tree-based methods.

Keywords DNA barcoding · Tellinoidea · Distance-based methods · Coalescent-based methods · The character-based method · Multiple genes

Introduction

Barcode of Life launched in 2003 (Hebert et al. 2003a) and is advertised to make species identification faster and more reliable by employing a short stretch of the mitochondrial cytochrome *c* oxidase I (COI) gene (Waugh 2007; Frézala and Leblois 2008). The initiative goal of this project is to develop a reference sequence library, by which new specimens can be identified simply and automatically via their DNA barcode sequences. Sequence-based species delimitation is becoming invaluable, especially in cases where the traditional identification tools are difficult to apply, such as for larval forms or phenotypically highly plastic species (Neigel et al. 2007; González et al. 2009). The vast majority of barcoding studies, since Hebert et al. (2003a), aim at testing the barcoding methodology, by first sequencing COI gene (or other genes) for numbers of samples, and then by comparing the results obtained with a priori established species mainly based on external morphology (Rach et al. 2008; Zou et al. 2011).

Over the last decade, most methods of DNA barcoding are tree-based and can be broadly divided into two classes. One is distance-based, which firstly converts DNA sequences into genetic distances within and between species, and then bases on the degree of genetic divergences, to establish identification schemes. The classical distance-based method usually sets a raw similarity threshold (e.g. the 3 % cut-off value threshold or the "10× rule" threshold) below which unknown samples are assigned to be described or as new species (Hebert et al. 2003a). Several proponents later coined a notion of "barcoding gap", a distance gap between intra- and inter-

Electronic supplementary material The online version of this article (doi:10.1007/s10126-014-9596-6) contains supplementary material, which is available to authorized users.

Z. Yu · Q. Li (✉) · L. Kong · H. Yu
The Key Laboratory of Mariculture Ministry of Education, Ocean University of China, Qingdao 266003, China
e-mail: qili66@ouc.edu.cn

Published online: 09 August 2014

specific sequence divergences (Meyer and Paulay 2005; Meier et al. 2008). This procedure may be relatively straightforward when the barcode gap is observable; however, the twin distributions of intra- versus inter-specific divergences typically overlap (Hickerson et al. 2006).

Another relies on coalescent theory to delimit species (Hebert et al. 2003b). It is originally developed to construct a gene tree and identify independently evolving clades as species, i.e. the monophyly-based method (Zou et al. 2011). It assumes that a set of lineages—species with orthologous genomic regions in distinct individuals, or other taxa with a tree-like genealogy—is monophyletic (Rosenberg 2007). Posterior probabilities are often used to support barcoding conclusions (Munch et al. 2008). Nevertheless, opponents argued that the gene tree may not completely correspond with the species tree (Kizirian and Donnelly 2004). Moreover, they deemed it somewhat arbitrary to apply a discrete criterion across taxa (Will and Rubinoff 2004).

With the growing interest in species delimitation methods, novel approaches have been put forward, such as the automatic barcode gap discovery (ABGD) approach, the general mixed Yule coalescent (GMYC) model and the character attribute organization system (CAOS). Similar to the classical distance-based method, ABGD will count two DNA sequences as members of distinct groups if their genetic distance is greater than a given threshold (i.e. barcode gap). It can automatically detect where the barcode gap is located by ranking all pairwise genetic distances from smallest to largest, and then partition a DNA sequence dataset into the maximum number of groups (i.e. species) accordingly through an iterative procedure (Puillandre et al. 2012). A main advance over the classical analysis is it can be used even when the two distributions overlap (Paz and Crawford 2012).

The GMYC model estimates species boundary directly from branching rates in a phylogenic tree through a likelihood-based analysis (Pons et al. 2006). Branching patterns of the gene tree within genetic clusters reflect neutral coalescent processes, whereas branching among them reflects the timing of speciating (Monaghan et al. 2009). Thus, conspecific lineages should show a high rate of coalescence relative to a slower rate for heterospecific lineages. This method exploits the switch in the rate and identifies clusters of specimens corresponding to putative species. A threshold (T) is optimized with the GMYC model so that nodes before that are considered as speciation events (Lu et al. 2012). Although GMYC is grounded in a solid likelihood framework, it heavily relies on the correctness of the Yule speciation model (Puillandre et al. 2012).

Without any biological models or assumptions, the character-based identification algorithm (CAOS) has been proposed as an alternative to tree-based barcoding methods. This method bases on the fundamental concept that members of a given taxa share attributes which are absent from sister groups (Sarkar et al. 2008). It characterizes species by identifying a unique combination of diagnostic nucleotides in the target DNA fragment. If the four standard nucleotides (A, T, G and C) are found in fixed states in one species, they are regarded as diagnostics for charactering the species. In other words, species boundaries can be defined by a series of diagnostic characters. To this sense, the approach can increase to any level of resolution by applying multiple genes (Rach et al. 2008), and can identify maternal paraphyletic species regardless the rate of speciation or its phylogenetic history (Yassin et al. 2010). Such a sufficient algorithm has gained remarkable success in several animal taxa so far, for instance, Odonata (Rach et al. 2008), Neogastropoda and turtles (Zou et al. 2011; Reid et al. 2011).

Herein, we focused on assessing the performance of these five analytical methods, which fell into three categories: distance-based methods (the classical approach and the ABGD approach), coalescent-based methods (the monophyly-based method and the GMYC model) and the character-based method (CAOS), by barcoding Tellinoidea across taxonomic hierarchies on multiple genes. The superfamily Tellinoidea is one of the most diverse and representative groups of Veneridae, Heterodonta, Bivalvia (Prezant 1998). It contains approximately 180 living species and has adapted to almost every marine environment (Yonge 1949; Laudien et al. 2003). There are many species with considerable commercial and ecological value in this superfamily, such as *Moerella iridescens*, *Sinonovacula constricta* and *Donax dysoni*. Most tellinoidean species have been readily classified on morphological and ecological characteristics and a morphological taxonomic system has been well built (Bieler et al. 2010; Coan and Valentich-Scott 2012). Therefore, Tellinoidea provides an ideal case for testing the performance of various DNA barcoding methods. By exploring the potential of various types of barcodes, we could get a clear idea of how the information of DNA sequence can be used in taxonomy.

Material and Methods

Sampling Procedure

A total of 83 individuals were newly sequenced consisting of 68 from Tellinoidea (belonging to 16 morphospecies, 8 genera and 5 families), and 15 from Cardiacea (as outgroups) (Table S1). All the samples were collected from 26 widespread localities along the coast of China from 2002–2011 and stored in 95 % ethanol in Laboratory of Shellfish Genetics and Breeding (Fig. S1).

DNA was extracted from small pieces of adductor muscle tissue following a phenol–chloroform procedure modified by

Li et al. (2002). Isolated DNA was resuspended in 1 % TE buffer and stored at −30 °C for use. Partial region of mitochondrial genes was amplified by polymerase chain reaction (PCR). Three pairs of primers were used to amplify COI and 16S recombinant DNA (rDNA) (Table S2). PCR was implemented in a 50-μL mix containing 2 U *Taq* DNA polymerase (Takara), about 100-ng template DNA, 1-μM forward and reverse primers, 200 μM of each dNTP, 1× PCR buffer and 2 mM $MgCl_2$. All PCRs were carried out by the following thermocycler programme: 94 °C for 3 min, 35 cycles of 94 °C for 45 s, 44–56 °C for 1 min and 72 °C for 1 min, then 72 °C for 10 min for extension.

PCR products were firstly visualized on 1.5 % agarose gels with ethidium bromide and then purified by EZ Spin Column DNA Gel Extraction kit (Sangon Biotech). The purified products were used as the template DNA for cycle sequencing reactions performed using BigDye Terminator Cycle Sequencing Kit (Applied Biosystems), and sequencing was conducted on an ABI PRISM 3730 (Applied Biosystems) automatic sequencer. Both DNA strands were sequenced to ensure accuracy. All newly generated sequences were deposited in GenBank (Table S1).

In addition, we mined all the tellinoidean barcode sequences of the two genes from NCBI database and added them to our dataset (Table S3). They were named by their GenBank accession numbers in subsequent analyses.

Processing of DNA Sequences

All the newly generated sequences were edited manually by comparing both strands, filtering the primer sequences and trimming ambiguous based calls using SeqMan software (DNAStar 7.2.1). Alignments were obtained with ClustalW (Thompson et al. 1994) in BioEdit 7.0.9 (Hall 1999). DnaSP 5.00.04 (Rozas et al. 2003) was used to calculate the number of haplotypes.

Distance-based Barcode Analysis

For the classical similarity analysis, pairwise sequence distances were calculated using Kimura's two-parameter (K2P) distance model and analyzed at species, genus and family level in MEGA 4.0 (Tamura et al. 2007) for COI and 16S rDNA genes individually. We run the ABGD program using the web interface at http://www.abi.snv.jussieu.fr/public/abgd/abgdweb.html. A prior for the maximum value of intraspecific divergence (Pmax) ranging from 0.001 to 0.1 was set. Twenty recursive steps within the primary partitions were defined. The default for the minimum relative gap width was limited to 1. K2P was selected as the substitution model to calculate pairwise distances.

Coalescent-based Barcode Analysis

Bayesian tree of COI and 16S rDNA were generated with MrBayes 3.1.2, respectively (Ronquist and Huelsenbeck 2003). Based on the Akaike information criterion, we finally determined the optimal evolution model with jModeltest 0.11: GTR + I + G model for COI and GTR + G model for 16S rDNA (Posada and Buckley 2004). The Bayesian inference analyses started from two different, random trees and ran for 40 million generations with a sample frequency of 1/1,000. The first 2,500 trees of each run were discarded as a burn-in to ensure the stability of final analysis. Posterior probabilities for each clade were shown. Maximum likelihood (ML) trees were inferred severally from unique haplotypes (100 for COI gene and 42 for 16S gene) using PhyML 3.0 (Guindon et al. 2010). The branch lengths on the ML phylograms were clock constrained using r8s 1.71 (Sanderson 2003). The root node was fixed at an arbitrary value of 1.0, then ultrametric trees formed by penalized likelihood (PL). Finally, the putative species on the ultrametric trees were determined using the GMYC method in the SPLITS package for *R* (available at http://r-forge.r-project.org/projects/splits) on a single threshold model (Pons et al. 2006).

Character-based Barcode Analysis

The CAOS algorithm identified pure unique diagnostics for a priori described groups, here termed "characteristic attributes" (CAs). CAs herein was defined as single-nucleotide states which only existed across all numbers of one clade but never in an alternative clade. The phylogenic trees were first produced using the K2P model in PAUP v4.0b10 (Swofford 2002) from the given dataset. Then, the trees were incorporated into NEXUS files with DNA data matrix of Tellinoidea in MacClade v4.06 (Maddison and Maddison 2009), respectively, as guide trees, and were modified manually to ensure that every node is collapsed to single polytomy and all individuals belonging to the same genus were integrated into one group. After that the datasets were executed in P-Gnome to identify CAs (Sarkar et al. 2008). The most variable sites that distinguished all the taxa were chosen and the character states at the nucleotide positions were exhibited. Finally, unique combinations of character diagnostics were identified.

Results

In total, 128 individuals were analyzed for COI gene, consisting of 63 newly generated ones and 65 downloaded ones greater than 500 bp. The data matrices contained 100 unique haplotypes, and harboured 327 variable and 297 parsimony informative sites. The overall nucleotide frequencies were 21.7 % for A; 16.5 % for C; 21.0 % for G and 40.9 % for T. 16S rDNA gene was examined to provide a comparison with species resolution in COI data. The data matrices of 16S rDNA gene, consisting of 53 newly generated sequences and 6 downloaded ones greater than 390 bp, contained 42 unique haplotypes. When sequencing failure occurred in some samples, only one sequence (either COI or 16S rDNA) from those was used in subsequent analysis.

Distance-based Delimitation

(a) Classical distance-based barcode. Relative frequency distributions of genetic distances of COI sequences according to different taxonomic levels within Tellinoidea were compared (Fig. 1). As expected, the degree of genetic divergence increased with higher taxonomic rank. Intraspecific pairwise genetic distances ranged from 0 to 8.6 % with a mean of 0.9 %. Mean pairwise divergence between individuals of congeneric species was 21.1 % (range 9.6–32.8 %). Pairwise genetic distances between specimens of different genera that belong to the same family was 37.2 % on average (range 18.9–59.4 %). The 3 % divergence threshold resulted in splitting 11.5 % (three species) of Tellinoidea. The 10× rule threshold (9.0 % in this study) could correctly distinguish all of the 26 morphospecies. A "distance gap" was detected between intra- and interspecific genetic divergences of COI sequences, and the gap width was 1 %.

Genetic divergences of 16S rDNA for different taxonomic levels within Tellinacea were shown in Fig. S2. Pairwise genetic divergences of conspecific individuals ranged from 0 to 10.3 % with an average of 0.4 %. Mean pairwise divergence between specimens of congeneric species was 19.8 % (range 4.9–35.7 %). Pairwise genetic distances between specimens of different genera that belong to the same family was 32.7 % on average (range 5.3–46.1 %). Both the 3 % divergence threshold and the 10× rule threshold (4.0 % in this study) resulted in splitting of 5.8 % of the 17 tellinaceans (barring some species unidentified by COI gene). Obvious overlap between intraspecific and interspecific genetic distances of 16S rDNA was found.

(b) Automatic barcode gap discovery approach. The ABGD analysis identified an evident "barcode gap" centred around 3 % of divergences of the COI sequences, and revealed 26 genetic clusters as candidate species (Fig. 2). This result was consistent in all recursive partitions with priori genetic distance thresholds between 1.83 and 3.79 %, and we considered it more likely than the other alternatives (such as clustering 53 candidates with intraspecific divergence values below 0.16 %). All of the groups of this 26 species hypothesis corresponded extremely well to the taxa recognized on morphological criteria.

In the ABGD analysis for 16S gene, a major barcode gap was detected at priori genetic distance thresholds ranging from 0.26 to 3.79 %, strongly supporting the presence of 18 clades potentially representing species (Fig. S3). Sixteen of the clades of this 18 species hypothesis were congruent with the currently recognized species. The remaining samples belonging to *Solecurtus divaricatus*, which showed high intraspecific variations, were splitted into two groups. A distinctive barcode gap defining 17 candidate species, with the same number as the currently recognized species, was identified at a priori genetic distance thresholds of 4.83 %. These 17 putative species comprised 14 known species and three taxa whose identification remained to be finalized. Members of *Solecurtus divaricatus* were still splitted into two groups, and, additionally, individuals of *S. constricta* were grouped together with samples of *Sinonovacula rivularis*. Therefore, we considered the 18 species hypothesis was more likely than the alternative.

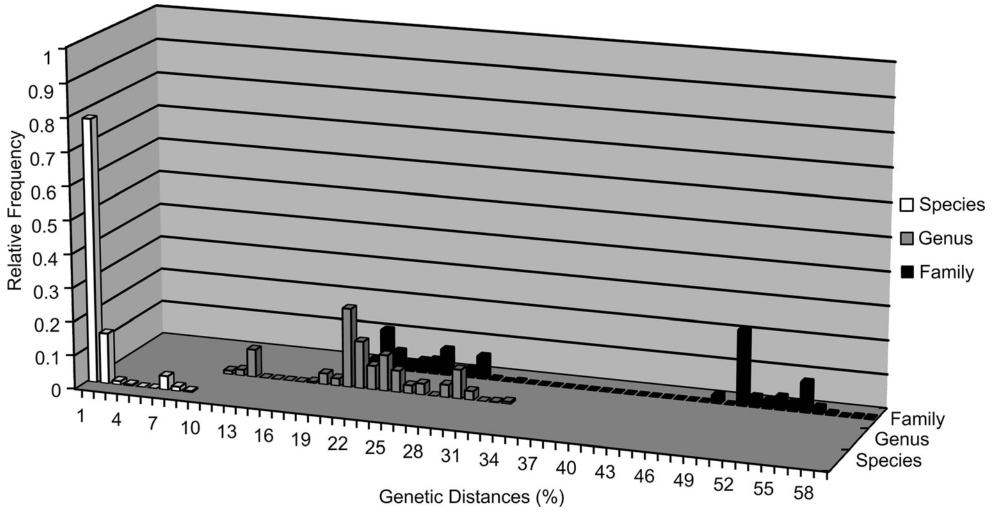
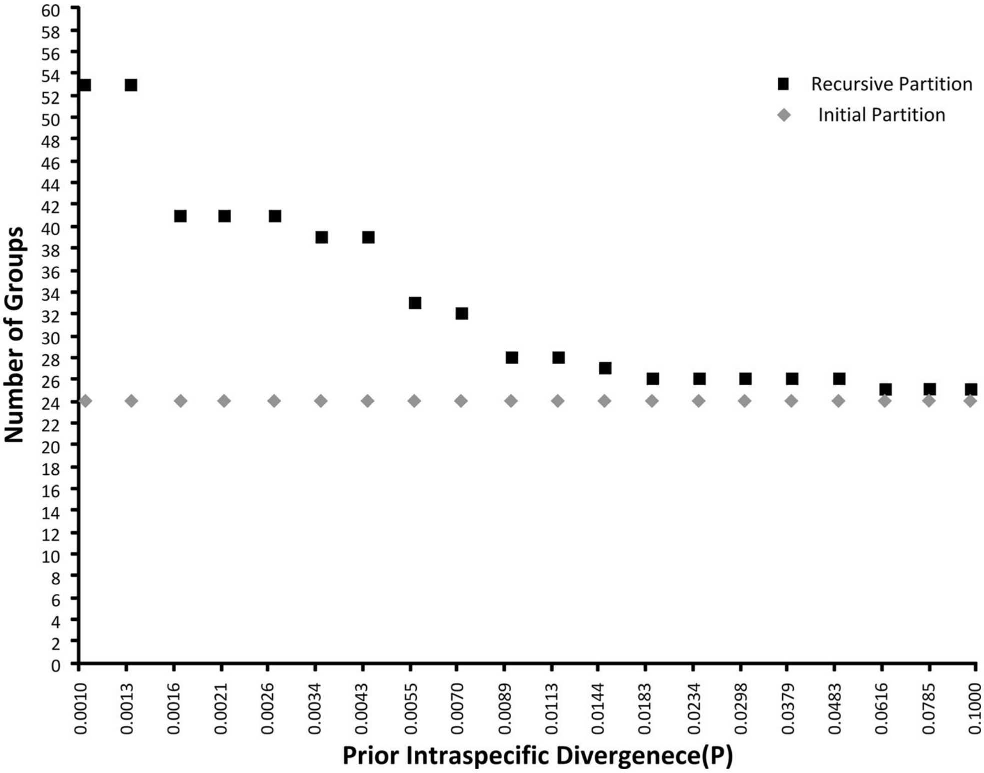


Fig. 1 Relative frequency distributions of intraspecific and interspecific distances according to different taxonomic levels for COI gene

Fig. 2 Automatic partition of tellinaceans based on COI gene. The number of groups inside the partition (initial and recursive) of each given prior intraspecific divergence value were reported



Coalescent-based Assignment

(a) Monophyly-based barcode. Both the COI and the 16S Bayesian trees depicted all the morphology identified species where more than one individual were obtained as monophyletic lineages with 98–100 % supports (Fig. 3 and S4). The seven sequences (five for COI and two for 16S) corresponded to singletons were also flagged as potentially unique in the phylogenetic trees, but no posterior probabilities could be computed. However, only two genera (*Macoma* and *Sinonovacula*) where more than one species were sampled were demonstrated as independently evolving clade with high posterior probabilities for COI gene, and only one (*Sinonovacula*) for 16S gene. Either splitting or lumping occurred among other genera (Fig. 3 and S4).

(b) General mixed Yule coalescent model. The optimal threshold points obtained by the GMYC model for both genes were shown in red line in Fig. 4 and S4, respectively. A total number of 35 lineages, including 24 clusters of more than one individual, were detected as significant GMYC entities based on the COI gene (Fig. 4). Six of the 26 named species were congruently oversplitted as two or more putative species. Twenty-seven GMYC entities, 12 of which included more than one individual, were found in the 16S gene dataset (Fig. S4). Similar to observation of COI gene, six morphospecies were improperly separated.

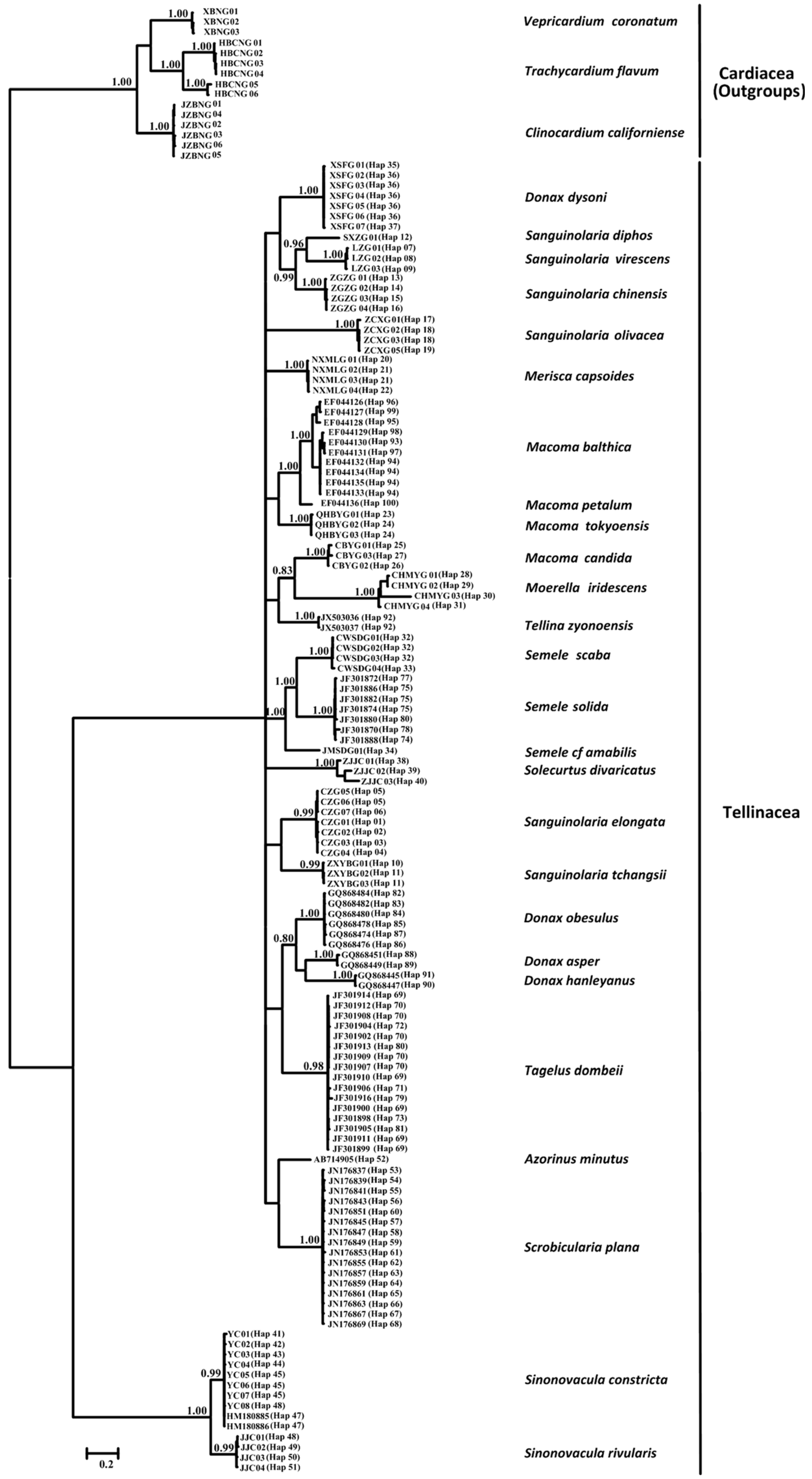
Character-based Identification

(a) Identification on species level. In the COI gene region of Tellinoidea for 26 species, character states at 35 nucleotide positions were found (Table 1). The particular nucleotide positions were selected due to high number of CAs at the key nodes or because of the presence of CAs for groups with highly similar barcoding sequences. All of the 26 tellinoideans revealed a unique combination of character states at 35 nucleotides with at least three different CAs for each species.

The character states at 32 nucleotide positions of 16S rDNA for 17 species of Tellinoidea were shown (Table S4). All species demonstrated a unique combination of character states at 32 nucleotide positions with at least three different CAs for each species.

(b) Identification on genus level. The character states for 12 Tellinoidea genera at 30 nucleotide positions of COI gene region were shown (Table 2). Dashed cells indicated non-significant positions at which at least three different nucleotides occurred

Fig. 3 The Bayesian tree of COI sequences of Tellinoidea with Cardioidea as outgroups using CTR + I + G model. Node support was indicated by posterior probabilities, and were given when ≥0.80



within a genus. All of the 12 genera immediately revealed a unique combination of character states at 30 nucleotides with at least three different CAs for each genus.

The character states for 10 genera of Tellinoidea at 26 nucleotide positions of 16S rDNA gene region were identified (Table S5). All of the 10 genera revealed a unique combination of

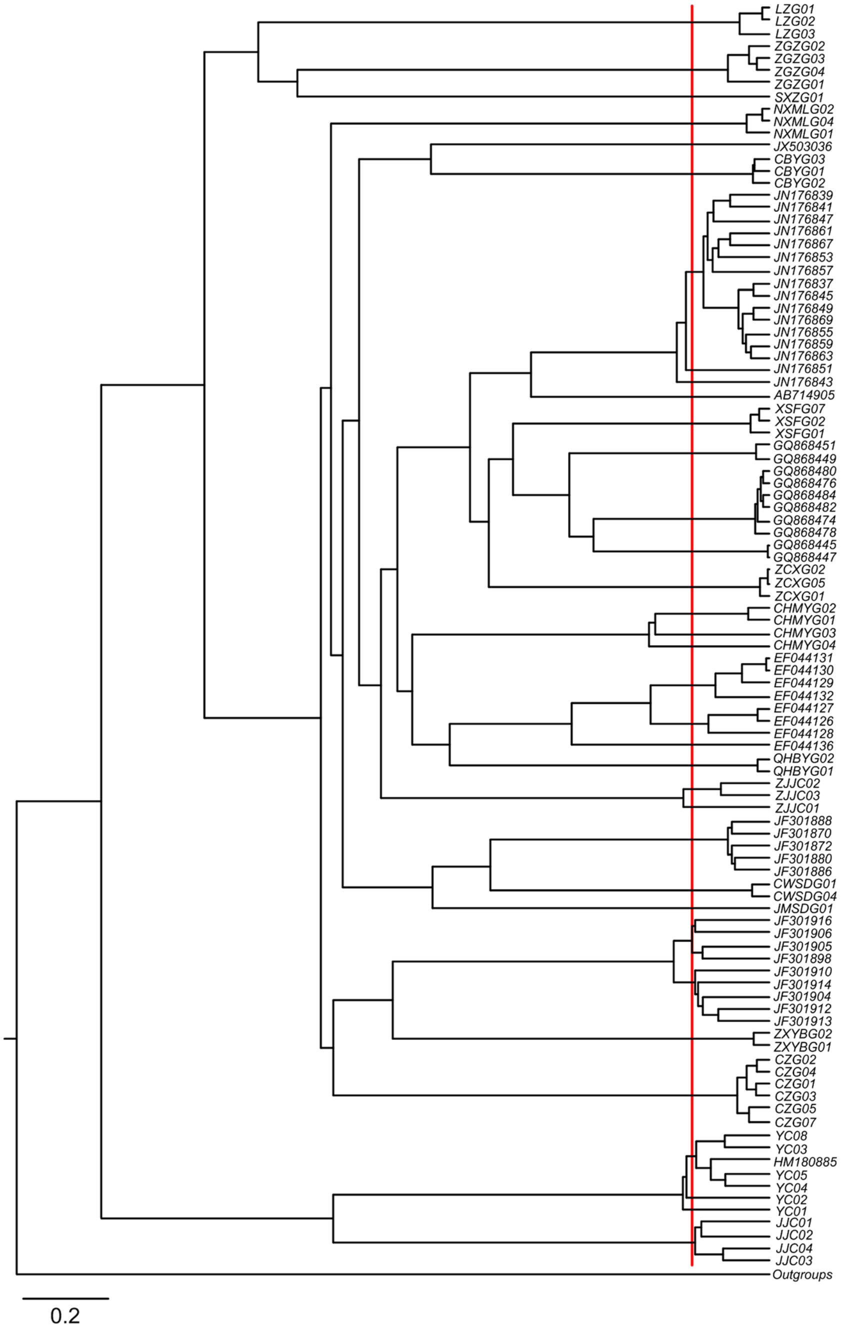


Fig. 4 Ultrametric NJ tree of tellinacean species on based on COI gene, generated from 100 unique haplotypes. The red vertical line in the tree was the threshold point obtained from the GMYC model

diagnostic characters at the selected positions with at least three CAs for each genus.

Discussion

Although still controversial (Meyer and Paulay 2005; Hickerson et al. 2006), the distance-based technique advanced by Herbert et al. (2003a) has become and will probably remain as the standard, workhorse approach in DNA barcoding (Reid et al. 2011). It reduces the information content of all nucleotides into a single distance vector, and usually uses a cut-off value to define categories, i.e. the classical distance method. In this study, 88.5–100 % examined tellinoidean species could be successfully identified by the 3 % criterion or the 10× rule threshold on multiple genes. Additionally, a 1 % width barcoding gap was detected in the COI data. However, this is probably an overestimation caused by undersampling, as (i) poor geographic sampling may leave an open access for high intraspecific divergences and (ii) exclusion of sister taxa

Table 1 Character-based DNA barcodes for 26 tellinacean species; Character states (nucleotides) at 35 selected positions of the COI gene region (ranging from 96–570); Taxa = abbreviations according to Table S1, S3; numbers of individuals analyzed per species were given in brackets

	Positions																																		
Taxa(n)	96	126	132	135	144	168	213	219	243	261	270	273	285	294	297	324	327	339	342	354	369	378	387	390	405	420	423	441	462	477	489	513	516	546	570
S. elongata (7)	T	T	T	G/A	A	T	G	G	T	A	A	A	T	G	A	A	A	T	A	T	A	G	G	A	T	T	A	A	G	T	T	G	G	T/C	G
S. virescens (3)	T	T	A	T	T	T	T	G	T	T	G	A	T	T	A	A	G	T	G	T	T	T	A	A	T	T	G	G	T	T	T	A	T	A	T
S. tchangsii (3)	T	A	T	G	T	T	A	G	G	G	C	G	T	G	A	T	T	T	A	T	A	G	G	A	G	G	T	T	A	T	A	T	T	A	T
S. diphos (1)	G	T	T	T	T	G	T	A	T	A	G	G	G	G	A	T	C	T	A	G	T	T	G	G	T	T	G	G	G	T	T	T	A	T	T
S. chinensis (4)	G	G	A	T	C	T	C	G	T	T	G	G	G	A	A	C	A	T	C	T	C	T	A	G	T	T	A	T	G	T	T	G	G	C	T
S. olivacea (4)	T	C	G	T	G	A	T	G	A	A	A	A	T	A	A	A	G	A	A	T	G	T	A	G	C	T	A	T	T	A	G	G	T	T	T
M. capsoides (4)	G	T	A	T	T	T	T	T	T	G	A	G	G	A	A	A	T	T	A	A	T	T	A	G	T	G	G	A	T	T	A	G	G	G	T
M. tokyoensis (3)	T	G	A	C	G	G	T	G	T	G	A	A	T	A	A	A	T	C	T	A	G	C	A	G	C	G	G	T	T	T	T	A	C	T	T
M. candida (3)	G/A	A	C	A	C	T	T	A	A	G	G	G	T	G	G	T	A	A	A	G	G	G	G	G	T	A	A	A	G	G	G	A	G	A	T
M. balthica (10)	T	T	G	T	T	A	T	A	T	C	G/A	T	T	G/A	A	T	G	A	T/C	A	T	T/C	G	G	T/G	A	A	A	T	T	T	G	T/C	G	T/C
M. petalum (1)	T	T	A	C	T	T	T	G	T	T	A	A	T	A	T	T	G	A	T	G	C	C	G	A	T	T	G	G	T	T	C	G	T	A	C
M. iridescens (4)	T	T/A	A	T	T	G	A	A	G	T	T	A	T	A	T	T	A	T	T	G	G	T	G	A	T	T	T	T	T	A	T	A	T	G	T
T. zyonoensis (2)	A	G	C	A	C	C	G	T	T	G	G	G	G	G	A	G	G	C	C	T	A	C	A	G	G	G	A	G	G	T	T	T	G	A	T
S. scaba (4)	T	T	T	T	T	A	G	G	G	G	G	T	A	A	A	T	T	T	G	A	G	T	G	G	T	G	A	T	T	T	G	A	G	G	T
S. cf amabilis (1)	G	G	T	A	G	T	G	G	T	T	A	A	G	G	A	T	A	G	G	G	G	C	G	G	G	C	A	A	G	T	C	T	A	G	T
S. solida (7)	A	G	G	T/C	G	T	A	A	G	G	A	C	T	A	G	G	G/A	A	A	A	A	T	G	A	T	G	G	G	G	C	C	C	A	T	C
D. dysoni (7)	T	T	T	T	T	A	T	A	A	A	A	A	T	A	A	A	A	T	G	G	T	T	A	G	T	A	A	T	A	C/A	A	A	T	T	T
D. hanleyanus (2)	T	T	T	T	T	T	T	T	G	T	T	A	T	T	A	A	T	C	A	T	C	T	G/A	G	A	G	A	A	G	A	T	G	A	A	G
D. obesulus (6)	T	G/A	A	C	G	A	C	A	G	G	T	A	G	T	T	G	G	T	G/A	A	A	C	A	G	C	T/C	A	A	G	G	C	A	A	A	C
D. asper (2)	T	T	T	T	T	A	A	G	T	A	G	C	T	A	A	A	G	G	A	T	A	T	A	A	T	T	A	A	G	A	A	T	C	A	T
S. divaricatus (3)	G/A	G/A	A	T	T/G	A	G	C	T	G	G	G	C	G/A	A	A	T	T	G	G	G	C	G	G/A	A	T	G	A	G	G	T/G	A	T/A	T/A	A
S. constricta (10)	T	T	A	T	T	T	G	A	T	A	G	T	A	G	A	T	A	A	T	A	T	A	G	T	T	T	T	T	A	A	A	T	G	G	T
S. rivularis (4)	T	T	A	T	T	T	A	T	T	A	A	T	G	G	G	T	G	A	T	G/A	T	A	A	C	T	T	T	T	A	G	G	G	G	T	T
T. dombeii (2)	T	T	A	T	G	T	T	A	A	G	A	A	T	T	A	T	T	C	A	G	G	C	A	A	G	A	G	A	T	A	C	A	T	A	G
A. minutus (1)	T	G	A	A	T	T	A	G	A	A	T	A	T	T	A	A	T	T	C	A	G	A	A	G	G	T	A	A	C	T	T	A	T	T	T
S. plana (17)	C	T	T	T/C	T	T	A	A	T	C	A	A	T	T/A	T	T	T	T	T	T	T	A	T	A	A	T	A	G	T	T	A	A	T	A	T/A

would exclude low interspecific distances at the same time. Given that gene variation represents a product of evolution, an arbitrary cut-off value could not entirely reflect how evolutionary processes impact on it (Zou et al. 2011). Moreover, owing to the diver mechanisms and the various mutation rates of mitochondrion DNA in distinct species (Yassin et al. 2010; Will and Rubinoff 2004), broad overlap of intra- and interspecific distances usually occurs and a universal set of criterion has not been reached (DeSalle et al. 2005; Rubinoff et al. 2006; Vences et al. 2005). Thus, we should be cautious about the classical distance-based method to discriminate species, even though our results confirmed its usefulness in identifying species.

The new proposed distance-based approach, ABGD, is meant to be used as a tool to automatically and rapidly formulate species hypotheses. It statistically infers the barcode gap from the data instead of an arbitrary empirical value and works with multiple thresholds throughout taxa (Puillandre et al. 2012). Nonetheless, ABGD is not an independent tool, and it still suffers limitations from genetic distance and barcoding gap concepts (Jörger et al. 2012). On one hand, the approximate maximum prior intraspecific distance (Pmax) has to be set. Importantly enough, this value needs not be defined precisely as the partitions are stable over a wide range of prior values (Puillandre et al. 2011). On the other hand, the users should decide which grouping option or options to be

Table 2 Character-based DNA barcodes at the genus level: character states (nucleotides) at 30 selected positions of the 16S COI gene region (ranging from 95–566); dashed cells indicate the occurrence of three or all four bases at this particular nucleotide position within a genus, numbers of analyzed species and individuals are shown in brackets

	Positions																													
Genus(Species/n)	96	102	126	135	180	201	207	216	222	225	228	234	246	270	273	279	339	342	378	387	390	402	421	423	459	462	465	507	558	567
Macoma (4/17)	▨	T/G	▨	▨	A/G	G	▨	A/G	T/G	A/T	T	A/G	A/T	A/G	▨	▨	A/C	A/T	▨	A/G	G	A	A/G	A/G	A/T	T/G	A/T	▨	▨	A/T
Merisca (1/4)	G	T	T	T	A/G	A	T	G	T	T	G	G	T	A	A/G	A	T	A	T	A	G	A/G	A	G	A	T	G	G	T	A
Moerella (1/4)	T	T	A	T	G	A	T	A	G	T	T	A	T	T	A	A	T	T	T	G	A	A	G	T	A	T	C	G	T	A
Semele (3/12)	▨	T	T/G	▨	A/G	A/G	T/G	A/G	T	▨	A/T	A/G	T/C	A/G	▨	T/C	▨	A/G	T/C	G	A/G	A	A	A/G	T	T/G	▨	A/G	▨	▨
Donax (5/16)	T	T	▨	T/C	A/G	A/G	▨	A/C	A/G	A/T	▨	A	▨	▨	A/C	A/T	▨	A/G	T/C	A	A/G	A	T/C	A	A/G	A/T	A/T	A	A/T	A/T
Solecurtus (1/3)	A/G	T	A/G	T	G	G	T	A	T	T	A	A	T	G	T	A	T	G	C	G	A/G	A	A	G	T	G	T	A	T	T
Sinonovacula (2/14	T	A/G	T	T	G	A	T	A/G	T	T	A	G	T	A/G	T	T	A	T	A	A/G	T/C	T	G	T	T	A	T	T	A	A/G
Sanguinolaria (6/22)	T/G	T	▨	▨	A/G	A/G	▨	A/G	▨	▨	T/G	A/G	▨	▨	A/G	T	T/G	▨	T/G	A/G	A	A	A/G	▨	▨	▨	T	▨	A/T	T
Tellina (1/2)	A	C	A	A	G	G	C	T	T	G	C	G	G	A	G	T	C	C	C	A	G	A	G	A	T	G	T	A	A	T
Scrobicularia (1/17)	C	T	T	T	A/G	A/G	T	G	T	T	T	A	G	A	A	A	T	T	A	T	A	T	A	A	A	T	A	A	T	T
Azorinus (1/1)	T	T	A	A	G	G	C	T	G	C	A	A	G	T	A	A	T	C	A	A	G	G	T	A	A	C	A	A	G	C
Tagelus(1/16)	T	T	T	T	G	G	T	G	A	C	G	G	G	A	A	T	C	A	C	G	A	A	G	G	C	T	T	G	T	A

used from a number of different ones on their prior information about divergence levels for a particular group (Paz and Crawford 2012). In this case study, ABGD correctly defined all the groups from COI sequences as putative species matching perfectly with known species. For 16S gene, ABGD immediately clustered 16 of 17 morphospecies. Our results highlighted that ABGD may be more objective and may have higher efficiency than the classical distance-based method. We thus recommend it to be used instead of any visual barcode gap definition.

Building of phylogenetic trees for delineating species as independently evolving clades could minimize the failure of identification (Kerr et al. 2009). Herein, the coalescent-based approach on monophyly criterion increased identification success by nearly 10 % over the classical distance-based method. Both COI and 16S rDNA sequences produced similar topologies at the terminal nodes in our study. They revealed that all of the species of interest formed a monophyletic cluster with well supports, although the sample size was low for some. Despite the high efficiency of monophyly-based method for species discrimination, critics have complicated the use of this approach in two cases. Firstly, the long recognized problem of flawed taxonomy will yield gene genealogies that may differ in topologies (Nielsen and Matz 2006). Secondly, the recently divergent taxa may fail to constitute reciprocally monophyletic groups due to lack of time needed to coalesce (Knowles and Carstens 2007). Indeed, several studies have already shown the limitations of monophyly-based methods to identify species (e.g., Trewick 2008; Robinson et al. 2009; Lukhtanov et al. 2009; Yassin et al. 2010). However, in groups with well-established taxonomy, such as Tellinoidea, species identification success has been strong.

Little genera where more than one species were sequenced recovered as monophyly in both trees. The monophyly-based method may be too prescriptive to high taxonomic levels, since it only recognized monophyletic taxa. For instance, applying this approach on *Sanguinolaria* would split the genus into three genera on COI gene, in spite of lacking any other morphological, ecological or reproductive isolation supports. In addition, the genetic information content of such trees is limited (Lowenstein et al. 2009), and the posterior probabilities for monophyly seems to be too conservative and misleads to reject monophyly in some cases (Will and Rubinoff 2004; Little and Stevenson 2007). Given these disadvantages, it seems best to avoid using monophyly-based method. Nevertheless, species resolution of monophyly-based method is generally in agreement with morphological taxonomy (e.g. Dai et al. 2012; Sun et al. 2012; Zou et al. 2011). And due to its powerful computational strengths, it could be still applied to flag species, especially for primary species identification.

The GMYC model tested for the presence of a shift from Yule (between species) to coalescent (within species) branch lengths in an ultrametric tree, but was found not significant relatively to the monophyly-based method in the current paper. Some nodes representing speciation events fell well outside the expected threshold for individuals of the same species at the barcode loci (Fig. 4 and S4). It clearly inflated the number of species units by 34.6 % for COI gene, and by 58.8 % for 16S gene. We thus consider this method as the most radical regarding species assignment, at least for Tellinoidea. As it heavily relies on the Yule speciation model, sampling scheme may be a confounding factor for this test (Lohse 2009). Sampling only a small number of populations is likely to lead to artificial clustering within species under the GMYC model, and there are several lines of evidence to suggest this had some effect on our findings. For one thing, six different cases of morphological species were randomly moved lineages, thereby generated additional GMYC entities. For another, COI gene, including relatively mass samples, gained a higher success ratio than 16S gene. But complete sampling is hardly ever achieved in practice, particularly for most barcoding data (Pons et al. 2006; Papadopoulou et al. 2008). Given the worrisome truth, it seems best to avoid using the GMYC model.

Contrary to phenetic barcodes, the use of diagnostic characters better approximates a real barcode owing to its core benefit of being visually meaningful (Lowenstein et al. 2009). The results of our research depended on multimarkers and implied that character-based barcoding with CAOS could be an effective and reliable technique to discriminate genetic entities at different taxonomic levels. On species level, all the 26 species in COI dataset and 17 species in 16S rDNA dataset revealed a unique combination of character states at least three out of the selected nucleotide positions, respectively. On the genus level, we found a unique combination of character states at 32 nucleotide positions and 26 of COI and 16S rDNA genes with more than three CAs for each of genera, separately. Even though CAs found in one single species may not be representative or less reliable for all others of this genus, such as *Merisca* and *Solecurtus*, they can still be useful in the overall process of genera identification (Rach et al. 2008). The reason is that barcodes of a single species not only increase the overall reliability of barcodes for the whole group, but also provide an important benchmark for the genus. Comparing to tree-based methods focusing on species identification, character-based method is much more suitable for genera delimitation. DNA barcodes in genera will be a powerful expansion for taxonomy and for facilitating biodiversity assessment on our planet.

Another advantage of character-based barcoding is that it is compatible with classical approaches, which is essential to “integrative taxonomy”. Integrative taxonomy previously presented by Dayrat (2005) called to use different sources of evidence in taxonomic practice rather than only relying on morphology. Several cases have already well-resolved problematical identification by means of "integrative taxonomy" based on the combinations of molecular and traditional information (Hebert et al. 2004; Burns et al. 2007).

Finally, we noted a general rise in success ratio of barcoding tellinoideans depending on COI sequences over that of 16S sequences in tree-based methods. It has revealed that the properties of COI made it amenable to be a barcoding marker, other than slowly evolving 16S rDNA gene. Whereas both COI and 16S genes were sufficiently sensitive and well suited as character-based barcode markers for differentiating Tellinoidea on species and genus level. A powerful evidence illustrated that the character-based DNA barcoding could employ more sequence resource for species discrimination, even the relatively conserved genes.

Conclusion

This research effectively demonstrates the potential of DNA barcoding technique in taxonomy of Tellinacea via five different algorithms. The character-based barcoding method performed well in species identification on different taxonomic levels, especially in barcoding the genera. With the great advantage of being compatible with tradition taxonomy, it could offer a powerful and reliable tool for accurate species identification and facilitative biodiversity assessment. Nevertheless, the ABGD approach and the monophyly performed as well as CAOS in barcoding tellinoideans on a species level, and they may be still used to flag species.

Acknowledgments We are extremely grateful to Dr. Jun Chen from Ocean University of China, who collected all the samples used here. The study was supported by research grants from National Natural Science Foundation of China (41276138, 31372524) and Fundamental Research Funds for the Central Universities.

References

Bieler R, Carter JG, Coan EV (2010) Classification of bivalve families. Malacologia 52:113–133

Burns JM, Janzen DH, Hajibabaei M, Hallwachs W, Hebert PDN (2007) DNA barcodes of closely related (but morphologically and ecologically distinct) species of butterflies (Hesperiidae) can differ by only one to three nucleotides. J Lepidopt Soc 61:138–153

Coan EV, Valentich-Scott P (2012) Bivalve seashells of tropical West America. Marine bivalve mollusks from Baja California to northern Peru. Stanford University Press, Barbara, pp 209–258

Dai L, Zheng X, Kong L, Li Q (2012) DNA barcoding analysis of Coleoidea (Mollusca: Cephalopoda) from Chinese waters. Mol Ecol Res 12:437–447

Dayrat B (2005) Towards integrative taxonomy. Biol J Linn Soc 85:407–415

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

Frézala L, Leblois R (2008) Four years of DNA barcoding: current advances and prospects. Infect Genet Evol 8:727–736

González MA, Baraloto C, Engel J et al (2009) Identification of Amazonian trees with DNA barcodes. PLoS ONE 4:e7483

Guindon S, Dufayard JF, Lefort V et al (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98

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

Hebert PDN, Ratnasingham S, deWaard RJ (2003b) Barcoding animal life: cytochrome oxidase subunit 1 divergences among closely related species. Proc R Soc Lond Ser B 270:S96–S99

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

Hickerson MJ, Meyer CP, Moritz C (2006) DNA barcoding will often fail to discover new animal species over broad parameter space. Syst Biol 55:729–739

Jörger KM, Norenburg JL, Wilson NG, Schrödl M (2012) Barcoding against a paradox? Combined molecular species delineations reveal multiple cryptic lineages in elusive meiofaunal sea slugs. BMC Evol Biol 12:245

Kerr KR, Birks SM, Kalyakin MV et al (2009) Filling the gap-COI barcode resolution in eastern Palearctic birds. Front Zool 6:29

Kizirian D, Donnelly MA (2004) The criterion of reciprocal monophyly and classification of nested diversity at the species level. Mol Phylogenet Evol 32:1072–1076

Knowles LL, Carstens BC (2007) Delimiting species without monophyletic gene trees. Syst Biol 56:887–895

Laudien J, Flint NS, van der Bank FH, Brey T (2003) Genetic and morphological variation in four populations of the surf clam *Donax serra* (Roding) from southern African sandy beaches. Biochem Syst Ecol 31:751–772

Li Q, Park C, Kijima A (2002) Isolation and characterization of microsatellite loci in the Pacific abalone, *Haliotis discus hannai*. J Shellfish Res 21:811–815

Little DP, Stevenson DW (2007) A comparison of algorithms for the identification of specimens using DNA barcodes: examples from gymnosperms. Cladistics 23:1–21

Lohse K (2009) Can mtDNA barcodes be used to delimit species? A response to Pons et al. (2006). Syst Biol 58:439–442

Lowenstein JH, Amato G, Kolokotronis SO (2009) The real maccoyii: identifying tuna sushi with DNA barcodes—contrasting characteristic attributes and genetic distances. PLoS One 4:e7866

Lu L, Chesters D, Zhang W et al (2012) Small mammal investigation in spotted fever focus with DNA-barcoding and taxonomic implications on rodents species from Hainan of China. PLoS ONE 7: e43479

Lukhtanov VA, Sourakov A, Zakharov EV, Hebert PDN (2009) DNA barcoding Central Asian butterflies: increasing geographical dimension does not successfully reduce the success of species identification. Mol Ecol Res 9:1302–1310

Maddison WP, Maddison DR (2009) Mesquite: a modular system for evolutionary analysis. Version 2.71. http://mesquiteproject.org. Accessed 23 Mar 2010

Meier R, Zhang G, Ali F (2008) The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. Syst Biol 57:809–813

Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. PLoS Biol 3:2229–2238

Monaghan MT, Wild R, Elliot M, Fujisawa T, Balke M, Inward DJG, Lees DC, Ranaivosolo R, Eggleton P, Barraclough TG, Vogler AP (2009) Accelerated species inventory on Madagascar using coalescent-based models of species delineation. Syst Biol 58:298–311

Munch K, Boomsma W, Willerslev E, Nielsen R (2008) Fast phylogenetic DNA barcoding. Phil Trans Roy Soc B 363:3997–4002

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

Nielsen R, Matz M (2006) Statistical approaches for DNA barcoding. Syst Biol 55:162–169

Papadopoulou A, Bergsten J, Fujisawa T et al (2008) Speciation and DNA barcodes: testing the effects of dispersal on the formation of discrete sequence clusters. Phil Trans Roy Soc B 363:2987–2996

Paz A, Crawford AJ (2012) Molecular-based rapid inventories of sympatric diversity: a comparison of DNA barcode clustering methods applied to geography-based vs clade-based sampling of amphibians. J Biosci 37:887–896

Pons J, Barraclough T, Gomez-Zurita J et al (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst Biol 55:595–609

Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst Biol 53:793–808

Prezant RS (1998) Heterodonta: introduction. In: Beesley PL, Ross GJB, Wells A (eds) Mollusca: the southern synthesis. CSIRO Publishing, Melbourne, pp 289–294

Puillandre N, Lambert A, Brouillet S, Achaz G (2012) ABGD, automatic barcode gap discovery for primary species delimitation. Mol Ecol 21:1864–1877

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 Lond B 275:237–247

Reid BN, Le M, McCord WP, Iverson JB, Georges A, Bergmann T, Amato G, Desalle R, Naro-Maciel E (2011) Comparing and combining distance-based and character-based approaches for barcoding turtles. Mol Ecol Res 11:956–967

Robinson EA, Blagoev GA, Hebert PDN, Adamowicz SJ (2009) Prospects for using DNA barcoding to identify spiders in species-rich genera. Zookeys 16:27–46

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574

Rosenberg NA (2007) Statistical tests for taxonomic distinctiveness from observations of monophyly. Evolution 61:317–323

Rozas J, Sanchez DJC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497

Rubinoff D, Cameron S, Will K (2006) A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. J Hered 97:581–594

Sanderson MJ (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19:301–302

Sarkar IN, Planet PJ, Desalle R (2008) CAOS software for use in character-based DNA barcoding. Mol Ecol Res 8:1256–1259

Sun Y, Li Q, Kong L, Zhen X (2012) DNA barcoding of Caenogastropoda along coast of China based on the COI gene. Mol Ecol Res 12:209–218

Swofford DL (2002) PAUP: phylogenetic analysis using parsimony (and other methods). Version 4.0. Sinauer Associates, Massachusetts

Tamura K, Dudley J, Nei M, Kumar S (2007) Mega 4: molecular evolutionary genetics analyses (mega) software version 4.0. Mol Biol Evol 24:1596–1599

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Trewick SA (2008) DNA barcoding is not enough: mismatch of taxonomy and genealogy in New Zealand grasshoppers (Orthoptera: Acrididae). Cladistics 24:240–254

Vences M, Thomas M, van der Meijden A, Chiari Y, Vieites DR (2005) Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. Front Zool 2:5

Waugh J (2007) DNA barcoding in animal species: progress, potential and pitfalls. BioEssays 29:188–197

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

Yassin A, Markow TA, Narechania AO, Grady PM, DeSalle R (2010) The genus *Drosophila* as a model for testing tree- and character-based methods of species identification using DNA barcoding. Mol Phylogenet Evol 57:509–517

Yonge CM (1949) On the structure and adaptations of the Tellinoidea, deposit-feeding Eulamellibranchia. Phil Trans Roy Soc B 234:29–76

Zou S, Li Q, Kong L, Yu H, Zheng X (2011) Comparing the usefulness of distance, monophyly and character-based DNA barcoding methods in species identification: a case study of Neogastropoda. PLoS One 6:e26619