DNA Barcoding of Animal Life – A Quantitative Assessment

Dept. of Biological Sciences
National University of Singapore

Zhang, Guanyang

Supervisor: Meier, Rudolf

INTRODUCTION

Recently, new methods for species identification and species discovery based on DNA sequence were proposed. First, "DNA barcoding" compares a segment of the *COI* gene of a unidentified specimen against a DNA barcode database for identification (Hebert *et al.* 2003. *Proc. R. Soc. Lond. B.* **270**: 313-321). Second, DNA taxonomy tries to delimit species boundaries exclusively based on DNA sequences for both described and undescribed species (Tautz *et al*. 2003 *TREE* **18**: 70-74). Taking a quantitative approach, this project investigates five issues/questions relating to DNA barcoding/taxonomy.

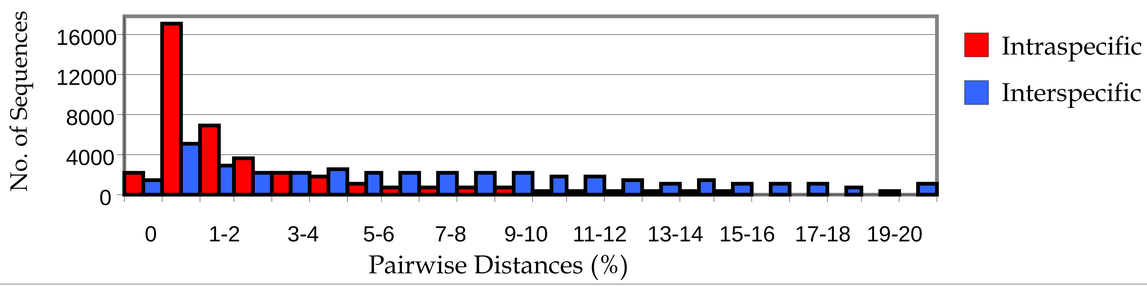
1. **COI variability** – Do intraspecific and interspecific distances overlap?
2. **DNA barcoding identification success** – How accurate is identification based on DNA barcodes?
3. **Assessing DNA taxonomy** - Can we describe species based on genetic distance?
4. **Which gene to use for DNA barcoding and how long should a DNA barcode be?**
5. **Application of DNA barcoding/taxonomy to a Southeast Asian taxon** – *Gastromyzon,* the Borneo suckers

49,841 *COI* sequences representing 14,058 species were downloaded from Genbank and aligned based on amino acid translations. Analyses were performed using *TaxonDNA* (Meier *et al.* 2006. *Syst. Biol.* **55**: 715 – 728). DNA was extracted from 35 species of gastromyzons and outgroups, and *COI, CYTB & 12s w*ere sequenced.

RESULTS & DISCUSSION

1 - COI variability – Distance overlap is extensive

All Metazoans - Average Intraspecific & Minimum Interspecific Distance Overlap



Intraspecific and interspecific genetic distance overlap (excl. 5% margins)

	Taxon	Min. Inter.		Max. Intra.	Perc. of all distances
Arthropods	Coleoptera	0.2%	–	7.2%	40%
	Diptera	0.1%	–	7.6%	25%
	Hymenoptera	0.3%	–	6.1%	29%
	Lepidoptera	0.1%	–	3.2%	34%
	Orthopteroidea	0.2%	–	6.2%	55%
	Paraneoptera	0.9%	–	17.8%	75%
	Crustacea	1.1%	–	12.1%	40%
	Chelicerata	0.2%	–	19.5%	94%
Other invertebrates	Cnidaria	0.0%	–	5.4%	87%
	Acoelomata	1.5%	–	20.0%	72%
	Pseudocoelomata	2.9%	–	19.6%	49%
	Annelida	0.0%	–	15.8%	91%
	Gastropoda	0.2%	–	13.6%	78%
	Bivalvia	0.0%	–	23.6%	92%
	Echinodermata	0.0%	–	2.6%	79%
Vertebrates	Mammalia	0.3%	–	5.0%	49%
	Aves	0.0%	–	1.8%	35%
	"Reptiles"	0.0%	–	18.3%	95%
	Amphibia	0.6%	–	13.5%	35%
	"Fish"	0.3%	–	8.1%	39%

Methods: Average Intraspecific distances and minimum interspecific distances are compiled.

Results: Intraspecific distances and interspecific distances overlap widely. Overlap is still extensive after removing the extreme 5% margins of all intraspecific and interspecific distances.

Discussion:

A good diagnostic character should have low variability within species and high variability between species – **Discontinuity.** However, for *COI*, intraspecific and interspecific distances are not well separated, implying that *COI* is not optimal for species identification or delimitation.

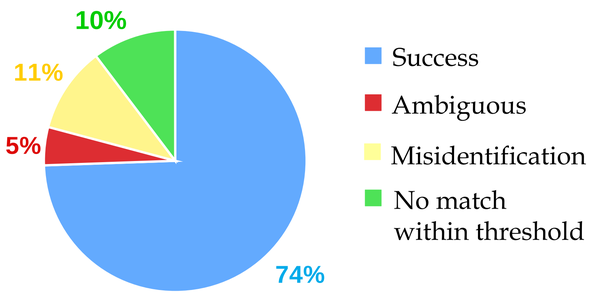
2 – DNA barcoding identification success

Methods: A sequence is taken as a query. We pretend that the query was from an unidentified species and identify it via comparison with the remaining DNA barcodes.

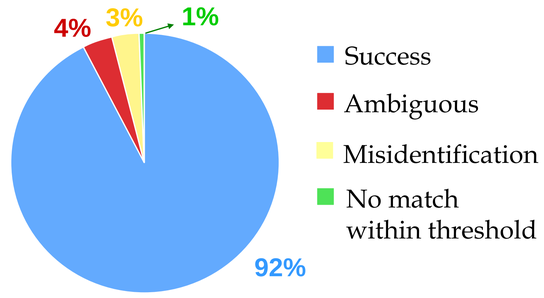
Threshold: A value that 95% of all intraspecific distances fall below

Success: The query sequence is matched with a conspecific sequence within threshold.
Ambiguous: The query sequence is matched with sequences from multiple species within threshold
Misidentification: The query sequence is matched with an allospecific sequence within threshold
No match within threshold: The query is not matched with a sequence within threshold

Best-Close Match based on all sequences



Best-Close Match based on conspecifics



2 – DNA barcoding identification success (Cont.)

Best-Close Match based on all sequences

	Taxon	Success	Ambiguous	Mis-identification	No match within threshold
Arthropods	Coleoptera	68%	3%	12%	17%
	Diptera	74%	6%	13%	7%
	Hymenoptera	65%	3%	9%	23%
	Lepidoptera	75%	4%	6%	15%
	Orthopteroidea	65%	3%	17%	14%
	Paraneoptera	76%	1%	21%	2%
	Crustacea	84%	1%	6%	9%
	Chelicerata	73%	5%	21%	1%
Other invertebrates	Cnidaria	57%	21%	12%	10%
	Acoelomata	90%	1%	8%	2%
	Pseudocoelomata	82%	3%	14%	2%
	Annelida	79%	2%	11%	7%
	Gastropoda	76%	3%	16%	5%
	Bivalvia	76%	6%	15%	3%
	Echinodermata	82%	9%	3%	6%
Vertebrates	Mammalia	83%	1%	4%	13%
	Aves	81%	4%	3%	11%
	"Reptiles"	69%	6%	22%	2%
	Amphibia	79%	0%	10%	11%
	"Fishes"	62%	1%	13%	25%

Best-Close Match based on conspecifics

	Taxon	Success	Ambiguous	Mis-identification	No match within threshold
Arthropods	Coleoptera	92%	3%	3%	1%
	Diptera	89%	7%	4%	0%
	Hymenoptera	92%	3%	3%	2%
	Lepidoptera	92%	4%	3%	1%
	Orthopteroidea	87%	4%	9%	1%
	Paraneoptera	97%	0%	2%	0%
	Crustacea	97%	1%	2%	1%
	Chelicerata	91%	5%	4%	0%
Other invertebrates	Cnidaria	75%	17%	4%	3%
	Acoelomata	98%	0%	1%	0%
	Pseudocoelomata	96%	2%	2%	0%
	Annelida	94%	2%	4%	0%
	Gastropoda	92%	3%	5%	0%
	Bivalvia	89%	7%	4%	0%
	Echinodermata	88%	9%	2%	1%
Vertebrates	Mammalia	98%	1%	1%	0%
	Aves	92%	4%	2%	1%
	"Reptiles"	90%	6%	4%	0%
	Amphibia	96%	0%	3%	1%
	"Fishes"	97%	1%	2%	0%

Results & Discussion:

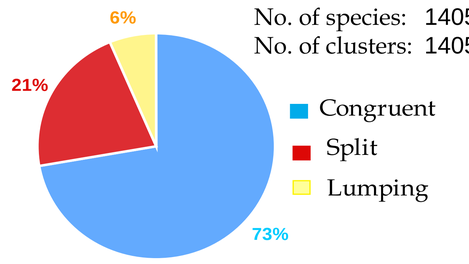
1. Success rates range from 57% – 90% based on all sequences.
   →Identification success will be low with an incomplete database.
2. Success rates reach 75% - 98% when only conspecifics are considered.
   →A complete database is required for high identification success. However, it is difficult to obtain for mega-diverse groups

3 - DNA Taxonomy – Do we need to redescribe 27% of all species?

Methods: Threshold-based species

Sequences with at least one matching sequence within threshold value were clustered.
Congruent: Clusters of all the seq. of a species
Split: Clusters containing some seq. of a species
Lumping: Clusters containing multiple species

All Metazoans – Cluster Result (3% threshold)



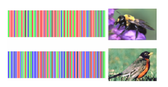
Results & Discussion

1. Numbers of species and clusters are very similar, but cluster content is in conflict for 27% of animal species.

3. Similar results are found for other threshold (1-5%)

	Taxon	Congruent	Split	Lumping
Arthropods	Coleoptera	77%	15%	9%
	Diptera	71%	17%	12%
	Hymenoptera	81%	14%	6%
	Lepidoptera	85%	7%	9%
	Orthopteroidea	51%	40%	9%
	Paraneoptera	74%	22%	4%
	Crustacea	62%	33%	5%
	Chelicerata	57%	43%	0%
Other invertebrates	Cnidaria	57%	27%	16%
	Acoelomata	53%	47%	0%
	Pseudocoelomata	63%	33%	4%
	Annelida	66%	29%	5%
	Gastropoda	68%	32%	0%
	Bivalvia	67%	23%	9%
	Echinodermata	71%	15%	14%
Vertebrates	Mammalia	83%	11%	6%
	Aves	75%	12%	13%
	Reptiles	65%	31%	4%
	Amphibia	56%	40%	4%
	Fishes	85%	8%	7%

4 – Sequence length and comparison between genes

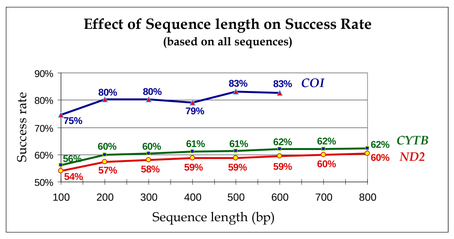


About 5000 *CYTB*, 4000 *ND2* & 3000 COI sequences representing 4000 bird species were used

Sequence length & ID success

Methods:
To assess effect of sequence length on identification success, blocks of complete sequences were obtained ranging from100bp to 800bp in length
Results: Success rate stabilizes after 200bp.
DNA barcodes can be short and Museum samples can be used

Effect of Sequence length on Success Rate (based on all sequences)



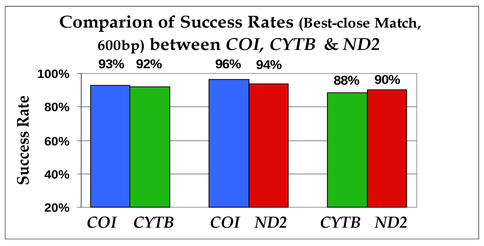
Compare different genes

Methods: Species with data for multiple genes are compared between pairs of genes

Results: No significant difference between the performance of *COI, CYTB & ND2*

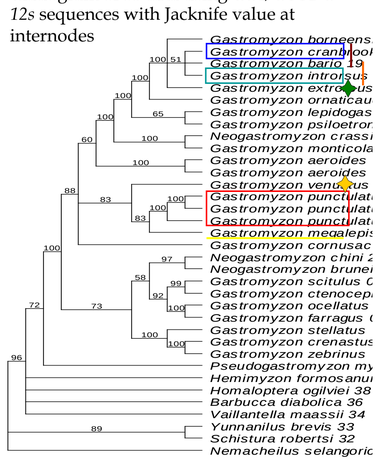
***COI* is applicable to most species. *CYTB* & *ND2* can also be used.**

Comparion of Success Rates (Best-close Match, 600bp) between *COI, CYTB* & *ND2*



5 – Testing DNA barcoding in a SEA tropical taxon

Cladogram of 35 taxa using COI, *CYTB & 12s* sequences with Jacknife value at internodes



Methods:

COI, CYTB & 12s were sequenced. The most parsimonious tree was constructed. (Jacknife, 36% deletion, 250 replicates)

Taxa	Species with Closest COI	Distance	Sister species	COI Distance
G. cranbrooki 03	*G. extrorsus 22*	0.48%	*G. bario 19/ G. introrsus 18*	0.63%/1.93%
G. introrsus 18	*G. extrorsus 22*	1.27%	*G. cranbrooki 03/ G. bario 19*	1.93%/2.21%
G. punctulatus 06	*G. venustus 24*	6.01%	*G. megalepis 07*	7.81%
G. punctulatus 16	*G. venustus 24*	6.32%	*G. megalepis 07*	8.43%
G. punctulatus 17	*G. venustus 24*	5.92%	*G. megalepis 07*	7.53%

Results

Closely-related species do not always have the most similar *COI* sequences as suggested in literature. This is demonstrated in 3 out of 22 *Gastromyzon* species (incl. *Neogastromyzon*) -14%

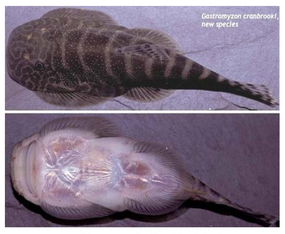
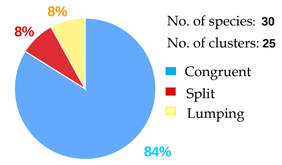


Photo credit: Tan Heok Hui

DNA Taxonomy :

COI cluster at 3%

16% incongruence



U0301960@NUS.EDU.SG