

# Identifying Canadian Freshwater Fishes through DNA Barcodes

Nicolas Hubert<sup>1</sup>, Robert Hanner<sup>2</sup>, Erling Holm<sup>3</sup>, Nicholas E. Mandrak<sup>4</sup>, Eric Taylor<sup>5</sup>, Mary Burridge<sup>3</sup>, Douglas Watkinson<sup>6</sup>, Pierre Dumont<sup>7</sup>, Allen Curry<sup>8</sup>, Paul Bentzen<sup>9</sup>, Junbin Zhang<sup>2</sup>, Julien April<sup>1</sup>, Louis Bernatchez<sup>1\*</sup>

**1** Département de biologie, Pavillon Charles-Eugène-Marchand, Université Laval, Sainte-Foy, Québec, Canada, **2** Canadian Barcode of Life Network, Biodiversity Institute of Ontario, University of Guelph, Guelph, Ontario, Canada, **3** Department of Natural History, Royal Ontario Museum, Toronto, Ontario, Canada, **4** Great Lakes Laboratory for Fisheries and Aquatic Sciences, Fisheries and Oceans Canada, Burlington, Ontario, Canada, **5** Department of Zoology, Vancouver, British Columbia, Canada, **6** Fisheries and Oceans Canada, Central & Arctic Region, Freshwater Institute, Winnipeg, Manitoba, Canada, **7** Ministère des Ressources naturelles et de la faune du Québec, Direction de l'aménagement de la faune de Montréal, de Laval et de la Montérégie, Longueuil, Québec, Canada, **8** Fish and Wildlife Research Unit, University of New Brunswick, Fredericton, New Brunswick, Canada, **9** Department of Biology, Dalhousie University, Halifax, Nova Scotia, Canada

## Abstract

**Background:** DNA barcoding aims to provide an efficient method for species-level identifications using an array of species specific molecular tags derived from the 5' region of the mitochondrial cytochrome c oxidase I (COI) gene. The efficiency of the method hinges on the degree of sequence divergence among species and species-level identifications are relatively straightforward when the average genetic distance among individuals within a species does not exceed the average genetic distance between sister species. Fishes constitute a highly diverse group of vertebrates that exhibit deep phenotypic changes during development. In this context, the identification of fish species is challenging and DNA barcoding provide new perspectives in ecology and systematics of fishes. Here we examined the degree to which DNA barcoding discriminate freshwater fish species from the well-known Canadian fauna, which currently encompasses nearly 200 species, some which are of high economic value like salmon and sturgeons.

**Methodology/Principal Findings:** We bi-directionally sequenced the standard 652 bp "barcode" region of COI for 1360 individuals belonging to 190 of the 203 Canadian freshwater fish species (95%). Most species were represented by multiple individuals (7.6 on average), the majority of which were retained as voucher specimens. The average genetic distance was 27 fold higher between species than within species, as K2P distance estimates averaged 8.3% among congeners and only 0.3% among conspecifics. However, shared polymorphism between sister-species was detected in 15 species (8% of the cases). The distribution of K2P distance between individuals and species overlapped and identifications were only possible to species group using DNA barcodes in these cases. Conversely, deep hidden genetic divergence was revealed within two species, suggesting the presence of cryptic species.

**Conclusions/Significance:** The present study evidenced that freshwater fish species can be efficiently identified through the use of DNA barcoding, especially the species complex of small-sized species, and that the present COI library can be used for subsequent applications in ecology and systematics.

**Citation:** Hubert N, Hanner R, Holm E, Mandrak NE, Taylor E, et al. (2008) Identifying Canadian Freshwater Fishes through DNA Barcodes. PLoS ONE 3(6): e2490. doi:10.1371/journal.pone.0002490

**Editor:** Hans Ellegren, University of Uppsala, Sweden

**Received:** December 5, 2007; **Accepted:** May 19, 2008; **Published:** June 18, 2008

**Copyright:** © 2008 Hubert et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This research was supported through funding to the Canadian Barcode of Life Network from NSERC, Genome Canada (through the Ontario Genomics Institute). Other sponsors listed at www.BOLNET.ca.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: Louis.Bernatchez@bio.ulaval.ca

## Introduction

DNA barcoding is designed to provide accurate, and automated species identifications through the use of molecular species tags based on short, standardised gene regions [1,2]. While humanity is facing increasing evidence of the erosion of Earth's biodiversity, this approach is proving its effectiveness in characterising the complexity of the biodiversity realm at a pace unequalled by other characters [3]. The primary goals of DNA barcoding focus on the assembly of reference libraries of barcode sequences for known species in order to develop reliable, molecular tools for species identification in nature. Current results suggest that, in a large

array of organisms, species are generally well delineated by a particular sequence or by a tight cluster of very similar sequences that allow unambiguous identifications [4,5,6,7,8,9,2,10,11,12].

Despite the great promise of DNA barcoding, it has been controversial in some scientific circles [13,14]. Yet, recent results illustrated some straightforward benefits from the use of a standardised molecular approach for identification [1,2]. First, intraspecific phenotypic variation often overlaps that of sister taxa in nature, which can lead to incorrect identifications if based on phenotype only [e.g. 15]. Second, DNA barcodes are effective whatever the life stages under scrutiny [e.g. 16, 17]. Third, cryptic variation and often spectacular levels of undetected taxonomic

diversity have been frequently reported [e.g. 18, 19, 20]. Finally, DNA barcode libraries are fully available as they are deposited in a major sequence database, and attached to a voucher specimen whose origin and current location are recorded [2,3]. Once libraries are available, recent studies illustrate the vast array of applications that can be applied to them such as forensic engineering [21,22], ecology of cryptic communities [23], the tracking of invasive species [24,25] and identification of prey from predator stomach samples [e.g. 26].

With the aim of assigning specimens to known species based on molecular tags, a 648-bp segment of the 5' region of mitochondrial cytochrome c oxidase I (COI) gene forms the library of primary barcodes for the animal kingdom [1]. Mitochondrial DNA (mtDNA) presents several advantages that make it well suited for large scale molecular tagging. First, this genome is present in a large number of copies yielding substantial amounts of genomic DNA from a variety of extraction methods. Second, the high mutation rate and small effective population size make it often an informative genome about evolutionary patterns and processes [27,28]. For a barcoding approach to species identification to succeed, however, within-species DNA sequences need to be more similar to one another than to sequences in different species. Several processes such as pseudogenes ontogenesis, introgressive hybridisation, and retention of ancestral polymorphism pose potential difficulties in capturing species boundaries using mtDNA sequences [29,30,31,32]. The detection of mixed genealogy between closely related species has been previously estimated to occur in nearly 20 percent of the cases in the wild [30]. Recent barcoding studies emphasised that this percent can vary widely among phyla, yet species assignment failures typically do not exceed 5 to 10 percent in a large array of organisms [2].

The economic importance and identification challenges associated with fishes prompted the launch of an international Fish Barcoding of Life (FISH-BOL) initiative (<http://www.fishbol.org/>) with the aim of barcoding all fishes. In the context of FISH-BOL and for the first time, we examine whether barcoding captures species boundaries and allows species identification among some of the major orders of primary freshwater fishes. Although COI divergence and species identification success has been previously assessed for some marine fishes [7], the average divergence found among freshwater fish species is unknown. The Canadian freshwater fish fauna has been subject to intensive taxonomic analysis for decades [33,34,35,36,37]. Thus, this fauna provides an excellent opportunity to test the efficacy of barcoded-based species delimitation and identification of freshwater fishes over a broad geographic range. Moreover, a large number of species from highly endangered and economically important groups such as salmon and sturgeon are found in Canada. Given their high diversity and dramatic phenotypic changes during development, fish species identification is no easy task. Hence, the development of reliable and universal molecular tags constitutes a major requirement for forensic engineering and conservation strategies involving such emblematic species.

## Materials and Methods

### BARCODE data standard and data management on BOLD

DNA Barcoding has greatly influenced the pace of sequence data acquisition. This approach prompted the development of new protocols and databases to manage the constitution of COI libraries for molecular identification. The Barcode of Life Data System (BOLD; see <http://www.barcodinglife.org>) was developed as a collaborative online workbench that has evolved into a

resource for the DNA barcoding community [3]. The BOLD database currently host specimens records for which essentially, seven data elements are listed:

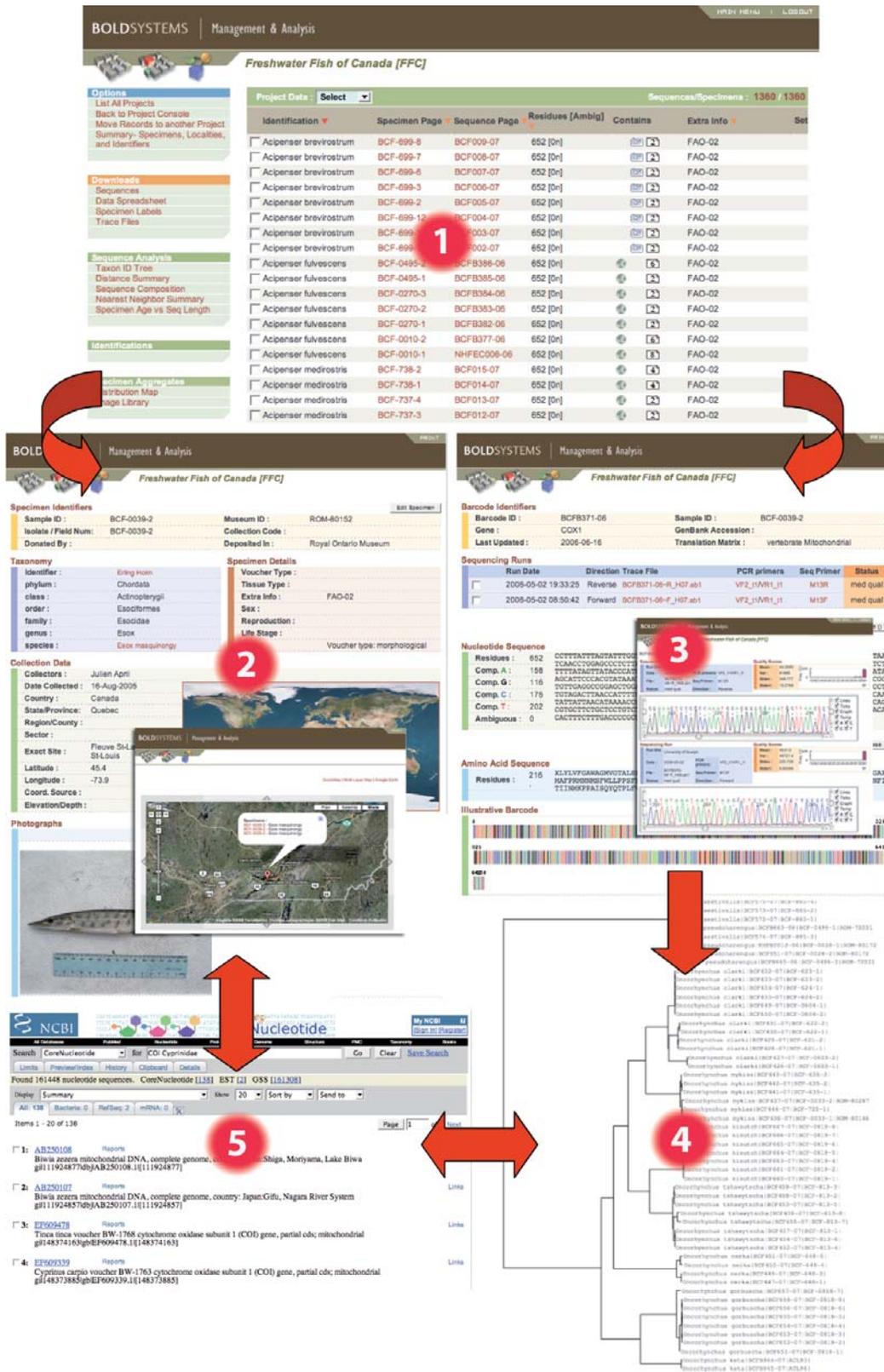
1. Species name
2. Voucher data
3. Collection record
4. Identifier of the specimen
5. COI sequence of at least 500 bp
6. PCR primers used to generate the amplicon
7. Trace files

The core data element in BOLD is a biphasic record consisting of both a “specimen page” and a “sequence page” (Figure 1). Access to these pages is possible through direct link in the project console (1 in Figure 1) that includes a comprehensive list of all specimens included in the project. The specimen page (2 in Figure 1) assembles varied 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 (3 in Figure 1) that records the barcode sequence (FASTA format), PCR primers and trace files, amino acid translation, and ultimately the GenBank accession number as well. Information from both the specimen and sequence pages can be incorporated into taxon ID trees that can be used in the identification system, while onboard mapping functions support investigations into spatial molecular ecology.

After preparing the barcode records in BOLD, data were uploaded into GenBank. Appendix S1 provides the voucher specimen ID, BOLD specimen record number, and GenBank accession number for each record. The Consortium for the Barcode of Life, in cooperation with GenBank and the other members of the International Nucleotide Sequence Database Collaboration (INSDC), have created and implemented the BARCODE data standard. “BARCODE” is a reserved keyword for those records in an INSDC database that meet a higher quality standard that makes them more reliable links between a gene sequence and a species name. All of the GenBank records created by this project and listed in Appendix S1 carry the BARCODE keyword because they include the following data:

1. Bi-directional sequences of at least 500 base-pairs from the approved barcode region of COI, containing no ambiguous sites
2. Links to electropherogram trace files available in the NCBI Trace Archive
3. Sequences for the forward and reverse PCR amplification primers
4. Species names that refer to documented names in a taxonomic publication or other documentation of the species concept used
5. Links to voucher specimens using the approved format of institutional acronym:collection code:catalog ID number

Taken together, the data required under the BARCODE data standard give researchers and other users with unprecedented access to data and metadata associated with the DNA sequence in GenBank. In addition, all of the information related to the present project is publicly available in the ‘Freshwater Fish of Canada’ projects (BCF and BCFB) on the Barcode of Life database (see <http://www.barcodinglife.org>) [3].



**Figure 1. Structure of the Freshwater Fish of Canada (FFC) database in BOLD, functionalities and connections with others public databases.** 1, Project page with the list of the specimens analysed including a link to the specimen and sequence page; 2, Specimen page for an individual of *Esox masquinongy* including voucher information, taxonomy, collection location, collection site maps and specimen image; 3, Sequence page for the same individual of *Esox masquinongy* including sequencing details including links to trace files, amino acid translation of sequence and trace viewer; 4, Taxon ID tree for the Canadian members of the Salmonid genus *Oncorhynchus*; 5, Connections with the GenBank. doi:10.1371/journal.pone.0002490.g001

## Data acquisition and analytical tools

DNA sources for this study included either frozen or ethanol-fixed tissue samples (muscle, liver, blood, or fin). Samples for barcoding originated from expert-identified specimens based on morphological criteria (meristic, morphometric and colouration) currently recognized in recent monographs [33,34,35,36,37]. For each specimens, detailed geographic information and where possible, reference specimens were deposited as vouchers in publicly accessible collections. However, some tissues collected before the beginning of FISH-BOL were obtained through the support of fish taxonomists, particularly for species exhibiting remote geographic distribution. In that case, sequences were generated from tissues lacking proper morphological vouchers. In order to allow the repeatability of the sequences generated, the tissues used for extraction and amplifications were given the status of 'tissue' voucher and distinguished from traditional 'morphological' vouchers. Of the 1360 specimens analysed (190 species), 861 (127 species) sequences were obtained from specimens with vouchers housed in the collection of the Royal Ontario Museum, Toronto (Appendix S1). Hence, samples with specimens housed in museum collections represented 65% of the sequences and 70% of the species analysed in this study.

Previous comparative genetic surveys suggested that freshwater fishes generally exhibit higher levels of inter-population genetic diversity than marine fishes [38]. Hence, we aimed, where possible, to sample three to five individuals per site for at least two sites from different watersheds for widely distributed species to capture a representative part of the molecular diversity. Numbers of specimens per species ranged from one to 17 with a mean of 7.6; nearly twice the number of individuals per species previously analysed for marine fishes [7]. According to the General Status of Wildlife in Canada [39], the Canadian fauna currently includes 203 species of which 194 (96%) have been sampled during the present survey (Appendix S1).

DNA extractions were performed with the NucleoSpin96 (Machery-Nagel) kit according to the specification of the manufacturer under automation with a Biomek NX liquid-handling station (Beckman-Coulter) equipped with a filtration manifold as previously described [40,41]. A 652-bp segment was amplified from the 5' region of the mitochondrial COI gene using either the following primers FishF1-5'TCAACCAACCACAAAGACATTGGCAC3' [7] and FishR1-5'TAGACTTCTGGGTGGCCAAAGAATCA3' [7] or the primer cocktails (including M13 tails to facilitate sequencing) [42] when amplifications failed using the first set of primers. PCR amplifications were performed in 12.5 µl volume including 6.25 µl of 10% trehalose, 2 µl of ultra pure water, 1.25 µl of 10× PCR buffer (10mM KCl, 10mM (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, 20mM Tris-HCl (pH8.8), 2mM Mg SO<sub>4</sub>, 0.1% Triton X-100), 0.625 µl of MgCl<sub>2</sub> (50mM), 0.125 µl of each primer (0.01mM), 0.0625 µl of each dNTP (10mM), 0.0625 µl of *Taq* DNA polymerase (New England Biolabs), and 2 µl of template DNA. The PCR conditions consisted of 94°C for 2 min, 35 cycles of 94°C for 30 s, 52°C 40 s, and 72°C for 1 min, with a final extension at 72°C for 10 min.

All the sequences have been deposited in GenBank and accession numbers for the barcodes, specimen and collection data, sequences, trace files and primers details are available within the BCF and BCFB project files in BOLD (<http://www.barcodinglife.org>). Sequence divergence was calculated using the Kimura 2-parameter (K2P) model [43] and the mid-point rooted Neighbour-joining (NJ) tree of K2P distances was created to provide a graphic representation of the species divergence [44] as implemented in the 'Sequence Analysis' module of BOLD. We checked for a potential sampling bias in the distribution of genetic diversity by plotting the mean intraspecific genetic distance

between haplotypes against the number of individual analysed and tested the significance of the relationship using a covariance analysis as implemented in Statgraphics [45].

## Results

A total of 194 species have been sampled during the present survey and the primers used amplified the target region of all, but four species: *Ctenopharyngodon idella* ( $n = 2$ ), *Lampetra richardsoni* ( $n = 5$ ), *Lampetra camtschaticum* ( $n = 5$ ) and *Catostomus columbianus* ( $n = 5$ ). Thus, a total of 1360 COI barcodes of 652-bp have been obtained for 190 species distributed among 85 genera and 28 families (Appendix S1; BCF abd BCFB projects in BOLD). No insertions/deletions or codon stops were found, supporting the view that all of the amplified sequences constitute functional mitochondrial COI sequences. Moreover, all the amplified sequences were larger than 600-bp, the limit typically observed for nuclear DNA sequences originating from mtDNA (NUMTs) [31]. The entire K2P/NJ tree derived from this study is available in Appendix S2 (or can be generated using BOLD).

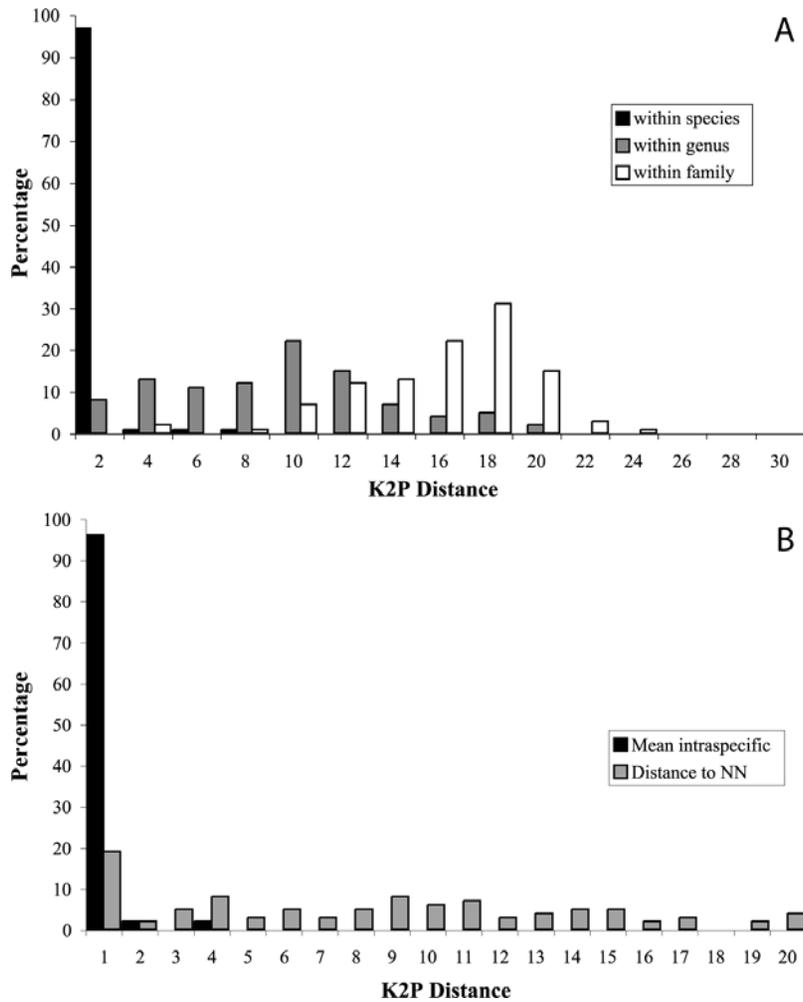
Average intraspecific variation was unrelated to the number of individuals analysed (average intraspecific K2P distance =  $0.015N + 0.135$ ; Covariance Analysis;  $F = 2.22$ ;  $P = 0.138$ ), suggesting representative sampling for the different species. The mean K2P distance of individual within species was 0.302 compared with 8.286 for species within genera (Table 1). Hence, overall, there was a 27-fold more pronounced difference among congeneric species than among conspecific individuals. Distributions of mean K2P distances among conspecific individuals and among congeneric species, however, partially overlapped as K2P distances ranged from 0 to 7.416 among conspecifics and 0 to 19.326 among congeneric species (Table 1).

A steady increase of genetic variation through increasing taxonomic levels was observed, supporting a marked change of genetic divergence at the species boundaries (Figure 2A). The analysis of the distribution of the nearest-neighbour distance (NND), namely the minimum genetic distance between a species and its closest congeneric relative revealed that only 20% of the NND was lower than 1% (Figure 2B) and only 7% of the NND (14 cases) were lower than 0.1% (Table 2). By contrast, the divergence between conspecific individuals was lower than 1% in 96% of cases. NND averaged 7.5%, which was 30-fold higher than the mean within species distance of around 0.3% and 13-fold higher than the mean maximum intraspecific distance of around 0.6%. Overlap in the distribution of the genetic distances between conspecifics individuals and congeneric species may originate from deep intraspecific divergences and low sister-species divergence.

**Table 1.** Summary of genetic divergences (K2P model used for computing distances) for increasing taxonomic levels. Data are from 1360 sequences from 190 species and 85 genera.

Comparisons within	Taxa	Number of comparisons	Min	Mean	Max	SE
Species	190	5865	0	0.27	7.42	0.01
Genus, among Species	85	18933	0	8.37	19.33	0.03
Family, among Genus	28	96992	2.67	15.38	23.22	0.01
Order, among Families	20	76571	14.25	20.06	29.44	0.01
Class, among Orders	2	681968	17.49	24.57	31.20	0.002

doi:10.1371/journal.pone.0002490.t001



**Figure 2. Summary of the distribution of the genetic variability (K2P distances) at COI sequences for the 1360 individuals and 190 species analysed.** A. Distribution of the genetic distance within different taxonomic categories. B. Distribution of the genetic distances to the nearest-neighbour and mean intra-specific distance. doi:10.1371/journal.pone.0002490.g002

In a few cases, we detected deep divergences among individuals that had been assigned to single species. Two lineages, one in the Laurentian Great Lakes area and another one in the St Lawrence River and diverging from 1% to 2% from each other were observed in five species including the common shiner (*Luxilus cornutus*), fathead minnow (*Pimephales promelas*), finescale dace (*Phoxinus neogaeus*), golden shiner (*Notemigonus crysoleucas*) and fantail darter, *Etheostoma flabellare* (Appendix S2). The same pattern was found among samples from the brook stickleback *Culaea inconstans* and the redfin pickerel, *Esox americanus*, where the divergence was even greater as it reached 7% and 3%, respectively. This result supports a genetic differentiation of the two *Esox americanus* subspecies *E. americanus americanus* from the St Lawrence River and *E. americanus vermiculatus* from the Laurentian Great Lakes area to the west. Although a single haplotype was found for each subspecies, more genetic divergence was observed between these two subspecies than with *Esox niger* since *E. americanus* was paraphyletic with its genealogy encompassing that of *Esox niger*. Likewise, a lineage found in the Pacific coast and diverging by 1.5% from the eastern samples was observed in the mottled sculpin, *Cottus bairdii*. Moreover, the Pacific lineage of *C. bairdii* was more closely related to the slimy sculpin, *Cottus cognatus*, than other conspecific samples. This suggests that a careful reappraisal of the current taxonomy for these groups could prove informative.

Cases of shared barcode haplotypes were detected in 13 (7%) of the species analysed including the following pairs: between the lampreys *Ichthyomyzon fossor* and *I. unicuspis*, between the shiners *Notropis volucellus* and *N. buchanaui*, between the shad *Alosa aestivalis* and *A. pseudoharengus*, between the putative species in the cisco species flock, *Coregonus artedii*, *C. hoyi*, *C. kiyi*, *C. nigripinnis* and *C. zenithicus*; and, between the darters *Etheostoma nigrum* and *E. olmstedii*. Nevertheless, we only found evidence of introgressive hybridisation between two diverging species in the case of the darters *Etheostoma nigrum* and *E. olmstedii* with two clades diverging by nearly 6%, each one more closely associated with one of the two species. In all the other cases, COI sequences of the mixed species were tightly clustered and differed by less than 0.1% divergence (Table 2).

## Discussion

This study has shown the efficacy of COI barcodes for diagnosing North American freshwater fishes since most species examined here corresponded to a single, cohesive array of barcode sequences that are distinct from those of any other species. The success of the barcoding approach depends on the distribution of genetic distances between conspecific individuals and heterospecific individuals given that failures in barcode clustering are

**Table 2.** Summary of the Canadian freshwater fish diversity and distribution of the genetic distance of each of the 190 species analysed to the nearest-neighbour at COI (K2P model used for computing distances).

Order	Family	Number of species					
		recorded	barcoded	<0.1	0.1–1.0	1.0–2.7	>2.7
Pleuronectiformes	Pleuronectidae	1	1	0	0	0	1
Cypriniformes	Cyprinidae	54	50	3	6	1	40
	Catostomidae	18	17	0	2	3	12
Scorpaeniformes	Cottidae	9	8	0	2	3	3
Salmoniformes	Salmonidae	29	29	7	6	3	13
Esociformes	Umbridae	2	2	0	0	0	2
	Esocidae	4	4	0	2	0	2
Clupeiformes	Clupeidae	4	4	0	2	0	2
Cyprinodontiformes	Fundulidae	3	3	0	0	0	3
Perciformes	Percidae	16	16	2	0	0	14
	Centrarchidae	13	12	0	0	3	9
	Percichthyidae	3	3	0	0	0	3
	Gobiidae	2	2	0	0	0	2
	Sciaenidae	1	1	0	0	0	1
Gasterosteiformes	Gasterosteidae	5	5	0	0	0	5
Siluriformes	Ictaluridae	10	10	0	0	2	8
Osmeriformes	Osmeridae	4	3	0	0	0	3
Semionotiformes	Lepisosteidae	2	2	0	0	0	2
Acipenseriformes	Acipenseridae	5	5	0	0	2	3
Osteoglossiformes	Hiodontidae	2	2	0	0	0	2
Petromyzontiformes	Petromyzontidae	10	5	2	0	0	3
Percopsiformes	Percopsidae	1	1	0	0	0	1
Gadiformes	Lotidae	1	1	0	0	0	1
	Gadidae	1	1	0	0	0	1
Atheriniformes	Atherinopsidae	1	1	0	0	0	1
Anguilliformes	Anguillidae	1	1	0	0	0	1
Amiiformes	Amiidae	1	1	0	0	0	1
	Total	203	190	14	20	17	139

doi:10.1371/journal.pone.0002490.t002

proportional to the overlap between both distributions [46]. It has been shown that lineages diversify more quickly within species than between species [47]. This is due to the fact that diversification within species is driven by mutation at a rate higher than speciation within lineages. Hence, the branch length between species tends to be much deeper than between conspecific individuals leading to a gap in the distribution of the pairwise distance between conspecific individuals and between species that has been referred to the barcoding gap [46]. The COI locus harbours a high mutational rate even for mtDNA [48]. The present study confirms that, in the vast majority of the taxa examined here (93%), the barcoding gap was observed and the mean genetic distance between conspecifics was generally much smaller than the average distance between individual from distinct species, even if only the sister species were considered.

Although barcode analyses primarily seek to delineate species boundaries at the COI locus for the assignment of unknown individuals to known species, unsuspected diversity and overlooked species are often detected through barcodes analyses, sometimes spectacularly [10,18,47]. The average distance between conspecific individuals was around 0.3% while average NND and

average distance between congeneric species were 7.5% and 8.3%, respectively. When screening for species splits using a threshold of 1% (3 fold higher than the average intraspecific variability), nine species exhibited lineages falling out of the average divergence between conspecific individuals.

Among the set of 190 species, however, 13 species (7%) exhibited barcode sequences that were shared or overlapped with those of other species. Regarding these cases, at least three factors may be involved [30,46]. First, the establishment of reciprocal monophyly between two sister species is a function of time given that fixation of a new coalescent follow the line of descent framework from the coalescent theory [49,50]. Second, the taxa may share polymorphism due to introgressive hybridisation. If hybridisation is due to secondary contact after a stage of isolation and genetic drift, introgressive hybridisation may be detected due to the presence of two divergent clusters, each one being found predominantly in one species or the other. Finally, the barcoding approach first examines species delineation through COI barcodes for species established generally through a traditional approach of taxonomy using phenotypes. Some of the pairs with overlapping barcodes, however, may be a single species. Alternatively, the use

of uniform threshold may be a source of error leading to erroneous assignment of individuals to species [51,52]. In the present case, 34 species would have been undetected by using a 1% threshold. Providing that seven species share polymorphism or harbour mixed genealogy, 24 species with monophyletic COI lineages would have been overlooked with a 1% threshold. Yet, the development of assignment tools based on more realistic probabilistic models under a coalescent framework will likely solve this problem and enhanced the statistical power of individual assignment through the use of a single gene [53,54].

The present study is the first to assess the resolution of barcoding for freshwater fish species from a variety of primary freshwater groups. It is widely appreciated that the fragmentation of the rivers and lakes from continental freshwater network leads to more pronounced genetic structure among populations and deeper divergence among haplotypes than in the marine realm [38]. In the largest barcoding study conducted so far on marine fishes to date [7], the average observed distance between conspecifics was 0.4% while the average divergence reached 9.9% between congeneric species. However, the average distance between conspecifics and congeneric species reached 0.3% and 8.3%, respectively, for freshwater fishes in this study, a pattern strikingly similar to that of marine fishes. Although geographic structure was often detected here among populations, the present survey suggests that the higher geographic structure of freshwater fishes is not necessarily reflected in deeper intraspecific and interspecific divergence than marine species. Although, we failed to capture a substantial amount of population diversity through the present sampling, it remains unlikely that sampling artefacts alone can account for similar intraspecific divergences found among freshwater and marine species. Admittedly however, the Canadian freshwater fish fauna may not be representative of old established population diversity since most of the rivers and lakes of the country have been colonised after the glacial retreat at the end of the Pleistocene [55].

In summary, most of the North American freshwater fish species analysed here exhibit a similar pattern of genetic diversity at COI, each being a single cluster of tightly related mtDNA sequences distinct from all other species. Therefore, the present survey supports the view that the use of COI barcodes is a powerful tool for species identification. Using this method would clearly allow the identification of individually isolated freshwater fish eggs, larvae, fillets and fins, hence providing many new tools useful for the practice of conservation and forensic genetic in these freshwater fishes. From a systematic perspective, COI barcodes provide a new and fast approach for screening the real number of species characterised by private sets of diagnostic characters. The identification of several cases of polyphyletic or paraphyletic COI species genealogy further supports the view that an iterative process of DNA barcoding followed by taxonomic analyses using other characters will be a productive way to catalogue biodiversity [10,56]. The present data set coupled with the functionality in BOLD provides a tool that is already operational for molecular assisted identification of the Canadian species. The entire cataloguing of the North American freshwater fish fauna, which is currently being undertaken by FISH-

BOL, will result in a significant improvement of our knowledge concerning the systematic of the freshwater fishes of the region and also facilitate monitoring changes in the geographic distribution of species that will probably occur in the future.

## Supporting Information

**Appendix S1** Details of species and specimens. Barcode of Life Database (BOLD) specimen numbers given, along with GenBank accession numbers, geographic locality and voucher details.

Found at: doi:10.1371/journal.pone.0002490.s001 (1.28 MB DOC)

**Appendix S2** Neighbour-joining tree of 1360 COI sequences from the 190 freshwater fish species sampled as obtained in BOLD, using K2P distances.

Found at: doi:10.1371/journal.pone.0002490.s002 (0.95 MB DOC)

## Acknowledgments

We wish to thank the following people for their valuable help during the present survey: Adam Betty, Thomas Pratt, Fraser Neave, Robert Devlin, Ruth Withler, John Radford, William Franzin from Fisheries and Oceans, Canada; Patrick Nelson from North/South Consultants Inc.; Lynda Corkum from the University of Windsor; Fred Johnson from Natural Resources, Canada; Marc Mingelbier, Yves Chagnon, Bruno Paré, Francois Girard, Jean Leclerc, Huguette Massé, Bernard Dumas, Chantal Côté, Francis Bouchard, Michel Plante, Philippe Brodeur, Richard Parizeau, Suzanne Lepage, Nathalie La Violette from Natural Resources, Québec; Guy Mayer, commercial fisherman, Québec; Bryan Neff, Nicholas Sweet from the University of Western Ontario; Bernard Côté from the Aquarium du Québec; Bernard Angers from the University of Montréal, Marco Rodriguez from the University of Trois-Rivières, Michael Nelson from the Essex Region Conservation Authority; Chris Wilson from the Trent University; Mark Gautreau from the University of New Brunswick; Julian Dodson, Julie Turgeon from the Laval University; Margaret Docker from the University of Manitoba; Yves Gauthier, Serge Pépin from the Biodôme, Montréal; Andrea Drauch, Josh Israel, John Velasquez, Bernie May from the University of California; John Wenburg, Randy Brown from U.S. Fish & Wildlife Service; Kevin Meyer, Tony Lamansky from the Idaho Department of Fish and Game; Andrew Whiteley from the University of Alaska, Steeve Leadbeater from the University of New Brunswick, Andrew Bentley and Ed Wiley from Kansas University. We acknowledge the support of Richard Winterbottom from the Royal Ontario Museum for providing facilities at the Ichthyology collection and Miranda Elliot for specimen cataloguing there; Tyler Zemlak and Christa Maitland University of Guelph, and Vicky Albert and Lucie Papillon in the Université Laval for the help during the laboratory work. We also thank Paul Hebert for his support and interest throughout the present study. Finally, we thank David Schindel, the academic editor and the anonymous referees for their constructive insights and comments.

## Author Contributions

Conceived and designed the experiments: LB RH NH. Performed the experiments: LB NH JA. Analyzed the data: ET LB RH NH EH NM DW JA. Contributed reagents/materials/analysis tools: ET LB PB NH EH NM MB DW AC JZ JA PD. Wrote the paper: ET LB RH NH NM.

## References

1. Hebert PDN, Cywinka A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London, Series B* 270: 313–321.
2. Hebert PDN, Gregory TR (2005) The promise of DNA barcoding for taxonomy. *Systematic Biology* 54: 852–859.
3. Ratnasingham S, Hebert PDN Bold: The barcode of life data system ([www.barcodinglife.org](http://www.barcodinglife.org)). *Molecular Ecology Notes* 7: 355–364.
4. Hebert PDN, Stoekle M, Zemlak T, Francis CM (2004) Identification of birds through DNA barcodes. *Plos Biology* 2: 1657–1668.
5. Ball SL, Hebert PDN, Burian SK, Webb JM (2005) Biological identification of mayflies (Ephemeroptera) using DNA barcodes. *Journal of the North American Benthological Society* 24: 508–524.
6. Saunders G (2005) Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. *Philosophical Transactions of the Royal Society, Series B* 360: 1879–1888.
7. Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005) DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society, Series B* 360: 1847–1857.

8. Clare EB, Lim BK, Engstrom MD, Eger JL, Hebert PDN (2006) DNA barcoding of Neotropical bats: species identification and discovery within Guyana. *Molecular Ecology Notes* 7: 184–190.
9. Cywinska A, Hunter FF, Hebert PDN (2006) Identifying Canadian mosquito species through DNA barcodes. *Medical and Veterinary Entomology* 20: 413–424.
10. Kerr KC, Stoeckle MY, Dove CJ, Weigt LA, Francis CM, et al. (2007) Comprehensive DNA barcode coverage of North American birds. *Molecular Ecology Notes* doi: 10.1111/j.1471-8286.2007.01670.x.
11. Robins JH, Hingston M, Matisoo-Smith E, Ross HA (2007) Identifying *Rattus* species using mitochondrial DNA. *Molecular Ecology Notes* doi: 10.1111/j.1471-8286.2007.01752.x.
12. Siefert KA, Samson RA, deWaard JR, Houbraken J, Lévesque A, et al. (2007) Prospects for fungus identification using COI DNA barcodes with penicillium as a test case. *Proceedings of the National Academy of Sciences, USA* 104: 3901–3906.
13. Will KW, Rubinoff D (2004) Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics* 20: 47–55.
14. Ebach MC, Holdrege C (2005) DNA barcoding is no substitute for taxonomy. *Nature* 434: 697.
15. Pfenninger M, Cordellier M, Streit B (2006) Comparing the efficacy of morphologic and DNA-based taxonomy in the freshwater gastropod genus *Radix* (Basommatophora, Pulmonata). *BMC Evolutionary Biology* 6: 100.
16. Caterino MS, Tishechkin AK (2006) DNA identification and morphological description of the first confirmed larvae of Heteriinae (Coleoptera: Histeridae). *Systematic Entomology* 31: 405–418.
17. Pegg CG, Sinclair B, Briskey L, Aspden WJ (2006) MtDNA barcode identification of fish larvae in the southern great barrier reef, Australia. *Scientia Marina* 70: 7–12.
18. 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*. *Proceedings of the National Academy of Sciences, USA* 101: 14812–14817.
19. Witt J, Threlloff DS, Doug L, Hebert PDN (2006) DNA barcoding reveals extraordinary cryptic diversity in an amphipod genus: implications for desert spring conservation. *Molecular Ecology* 15: 3073–3082.
20. 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. *Proceedings of the National Academy of Sciences, USA* 104: 4967–4972.
21. Dawney NR, Ogden R, McEwing R, Carvalho GR, Thorpe RS (2007) Validation of the barcoding gene COI for use in forensic genetic species identification. *Forensic Science International* in press.
22. Nelson LA, Wallman JF, Dowton M (2007) Using COI barcodes to identify forensically and medically important blowflies. *Medical and Veterinary Entomology* 21: 44–52.
23. Pfenninger MC, Nowak C, Kley D, Steinke D, Streit B (2007) Utility of DNA taxonomy and barcoding for the inference of larval community structure in morphologically cryptic *Chironomus* (Diptera) species. *Molecular Ecology* 16: 1957–1968.
24. Scheffer SJ, Lewis ML (2006) Mitochondrial phylogeography of the vegetable pest *Liriomyza trifolii* (Diptera: Agromyzidae): diverged clades and invasive populations. *Annals of Entomological Society of America* 99: 991–998.
25. Corin SE, Lester PJ, Abbott KL, Ritchie PA (2007) Inferring historical introduction pathways with mitochondrial DNA: the case of introduced Argentine ants (*Linepithema humile*) into New Zealand. *Diversity and Distributions* doi: 10.1111/j.1472-4642.2007.00355.x.
26. Pons J (2006) DNA-based identification of preys from non-destructive, total DNA extractions of predators using arthropod universal primers. *Molecular Ecology Notes* 6: 623–626.
27. Birky CW, Fuerst P, Maruyama T (1989) Organelle gene diversity under migration and drift: equilibrium expectations, approach to equilibrium, effects of heteroplasmic cells and comparison to nuclear genes. *Genetics* 121: 613–627.
28. Brown WM, George M Jr, Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences, USA* 76: 1967–1971.
29. Pamilo P, Nei M (1988) Relationships between gene trees and species trees. *Molecular Biology and Evolution* 5: 568–581.
30. Funk DJ, Omland KE (2003) Species-level paraphyly and polyphyly: frequency, causes and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology, Evolution and Systematics* 34: 397–423.
31. Zhang DX, Hewitt GM (1996) Nuclear integrations: challenge for mitochondria DNA markers. *Trends in Ecology and Evolution* 11: 247–251.
32. Hubert N, Torricco JP, Bonhomme F, Renno JF (2007) Species polyphyly and mtDNA introgression among three *Serrasalmus* sister-species. *Molecular Phylogenetics and Evolution* 46: 375–381.
33. McPhail JD, Lindsey CC (1970) Freshwater fishes of northwestern Canada and Alaska. *Bulletin of the Fisheries Research Board Canada*, 173.
34. Mandrak NE, Crossman EJ (1992) A checklist of Ontario freshwater fishes. Toronto, Canada: Royal Ontario Museum Publications.
35. Nelson JS, Paetz MJ (1992) The fishes of Alberta. Edmonton, Canada: University of Alberta Press.
36. Scott WB, Crossman EJ (1998) Freshwater fishes of Canada. Oakville, Canada: Galt House Publications.
37. Stewart KW, Watkinson DA (2004) The freshwater fishes of Manitoba. Winnipeg, Canada: The University of Manitoba Press.
38. Ward RD, Woodwark M, Skininski DOF (1994) A comparison of genetic diversity levels in marine, freshwater and anadromous fish. *Journal of Fish Biology* 44: 213–232.
39. Canadian Endangered Species Conservation Council (CESCC) (2001) Wild Species 2000: the general status of species in Canada. Ottawa, Ontario: Minister of Public Works and Government Services Canada.
40. Hajibabaei M, deWaard JR, Ivanova NV, Ratnasingham S, Dooh RT, et al. (2005) Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society, Series B* 360: 1959–1967.
41. Ivanova NV, deWaard J, Hebert PDN (2006) An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Molecular Ecology Notes* 6: 998–1002.
42. Ivanova NV, Zemlak TS, Hanner RH, Hebert PDN (2007) Universal primers cocktails for fish DNA barcoding. *Molecular Ecology Notes* doi: 10.1111/j.1471-8286.2007.01748.x.
43. Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 15: 111–120.
44. Saitou N, Nei M (1987) The neighbour-joining method: a new method for reconstructing evolutionary trees. *Molecular Biology and Evolution* 4: 406–425.
45. Statgraphic (1999) Statistical package. Los Angeles CA: Statistical Graphics Corporation.
46. Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. *PLOS Biology* 3: 2229–2238.
47. Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, et al. (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology* 55: 595–606.
48. Saccone C, De Giorgi C, Cissi C, Pesole G, Reyes A (1999) Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene* 238: 195–209.
49. Kingman JFC (1982) The coalescent. *Stochastic Process and their Applications* 13: 245–248.
50. Tajima F (1983) Evolutionary relationships of DNA sequences in finite populations. *Genetics* 105: 437–460.
51. Hickerson MJ, Meyer CP, Moritz C (2006) DNA barcoding will often fail to discover new animal species over broad parameter space. *Systematic Biology* 55: 729–739.
52. Meier R, Shiyang K, Vaidya G, Ng PKL (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology* 55: 715–728.
53. Abdo A, Golding GB (2007) A step toward barcoding life: a model-based, decision-theoretic method to assign genes to preexisting species groups. *Systematic Biology* 56: 44–56.
54. Nielsen R, Matz M (2006) Statistical approaches for DNA barcoding. *Systematic Biology* 55: 162–169.
55. Bernatchez L, Wilson CC (1998) Comparative phylogeography of Nearctic and Palearctic fishes. *Molecular Ecology* 7: 431–452.
56. Barber P, Boyce SL (2006) Estimating diversity of Indo-Pacific coral reef stomatopods through DNA barcoding of stomatopod larvae. *Proceedings of the Royal Society of London, Series B, Biological Sciences* 273: 2053–2061.