Abstract

Background: DNA sequences have become a primary source of information in biodiversity analysis. For example, short standardized species-specific genomic regions, DNA barcodes, are being used as a global standard for species identification and biodiversity studies. Most DNA barcodes are being generated by laboratories that have an expertise in DNA sequencing but not in bioinformatics data analysis. Therefore, we have developed a web-based suite of tools to help the DNA barcode researchers analyze their vast datasets.

Results: Our web-based tools, available at http://www.ibarcode.org, allow the user to manage their barcode datasets, cull out non-unique sequences, identify haplotypes within a species, and examine the within- to between-species divergences. In addition, we provide a number of phylogenetics tools that will allow the user to manipulate phylogenetic trees generated by other popular programs.

Conclusion: The use of a web-based portal for barcode analysis is convenient, especially since the WWW is inherently platform-neutral. Indeed, we have even taken care to ensure that our website is usable from handheld devices such as PDAs and smartphones. Although the current set of tools available at iBarcode.org were developed to meet our own analytic needs, we hope that feedback from users will spark the development of future tools. We also welcome user-built modules that can be incorporated into the iBarcode framework.

Background

Advancements in DNA sequencing technologies in recent years have resulted in an explosive use of comparative DNA sequence analysis in biological sciences. DNA sequence information has been used in a wide range of applications and for addressing different biological questions from development to evolution and biodiversity. In the early days of molecular biology a handful of sequence analysis software applications existed, several of them have been developed by researchers to address their needs. In last decade or so, development of more robust sequencing platforms, mainly as a result of human and other genome projects, resulted in the introduction of more powerful data analysis packages. Additionally, advancements in computer technologies and applications have been essential for a boom in bioinformatics. With
Here we introduce iBarcode.org, a web-based application server that provides various visualization and analysis tools for DNA barcoding data in a user-friendly environment. These tools have mainly been designed to enable the analysis of large barcode-style data sets, although the features can be used for the analysis of other sequence data. iBarcode.org is free and does not require registration.

Results
The current implementation of iBarcode.org (July 2008) includes a sequence upload and management suite and nine analysis and visualization tools. The sequence upload and management suite enables input, selection, verification, concatenation, and visualization of sequences. The web server provides tools that are divided into three categories. Here we introduce key features of iBarcode.org and provide exemplar cases from barcode data for each analysis and visualization module.

Sequence analysis

a. Haplotype variation
This tool identifies unique haplotypes for each species and provides statistical information on haplotype frequency and nucleotide variation in a user-friendly table format. A simple measure of number of nucleotide difference between sequences is used to calculate haplotype variation across the sequences. Figure 1 demonstrates the screen capture from output of the haplotype variation tool for a set of primate species (partial data set from Hajibabaie et al. [10]). In addition to this table, a reduced dataset containing unique haplotypes is produced in FASTA format. This dataset is stored for further use in other tools (see below) or for download by the submitter.

b. Haplotype map (Barcode-HAPMAP)
This data visualization module provides a graphical view of the nucleotide character variation in a barcode data set. It allows the user to quickly pinpoint nucleotide positions within the barcode sequence that account for barcode variation in a set of species. The tool takes a FASTA alignment of barcode sequences (or the alignment of unique haplotypes created in the Haplotype Analysis tool from a given barcode dataset) as input and highlights variable positions across the barcode sequence in an easy-to-read format. It also shows the nucleotide position for each variable site (counting from 5' to 3') as well as the codon positions they belong to. It is therefore important that the FASTA file of the barcode sequences is in the correct reading frame. This tool works best for focused character-based analysis of a limited number of taxa (i.e. in a species complex or when dealing with cryptic species) as a complement to distance-based methods such as Neighbour-joining analysis [11]. The HTML output format generated by this tool allows robust data transfer to other software packages such as MS-Excel. Figure 2 is an exemplar Bar-
code-HAPMAP of the unique haplotypes in a set of 4 species of skipper butterflies (Lepidoptera:Hesperiidae) [12].

c. Tests of selection at different taxonomic levels
This module uses the popular ratio of non-synonymous to synonymous substitutions (ω) [13] at various taxonomic levels. This ratio has been used for estimating the
degrees of selective pressure in molecular biosystematics. The module uses the program yn00 from the PAML package [14,15] to calculate the ratio of non-synonymous to synonymous substitutions (ω) for all pairs within a set of aligned sequences. It then calculates the average and standard deviation of ω for all sequences pairs that belong to the same species, belong to the same genus, or belong to different genera. A final bar graph depicting these various values is then displayed (Figure 3).

d. DNA barcode cloud visualization
This module takes the popular "word cloud" concept and applies it to number of individuals of each species within a given barcode dataset, producing a visually-appealing means of seeing the relative abundance of species within a dataset. These relative abundances are linearly scaled between font sizes of 50 and 200 points. This feature also provides cloud visualization for sequence divergence within species and haplotype diversity in each species. Each species represented in the cloud visualization output can be selected to create a new subset dataset for further analysis using other tools. Figure 4 provides an example of a barcode cloud for a set of species of primates.

Genetic distance analysis
a. Between- vs. within-species variation graph
DNA barcoding is based on a simple premise: genetic variation between species exceeds that of within species. This tool allows the user to visualize this principle in a given barcode dataset. Specifically, for each species with 3 or more individuals, this tool plots maximum Within Species Divergence (Max-WSD) against minimum Between Species Divergence (Min-BSD) [7]. The input for this tool is a genetic distance matrix (text format) produced either internally (by calculating number of nucleotide differences between and within species) or by common sequence analysis programs such as Mega [16]. Several barcoding studies have used graphs of between- vs. within-species variation. These graphs are considered as one of the standard methods of visualizing barcode data [i.e. [7]], as they allow the user to quickly see outliers that may represent misannotated specimens or sequencing errors.

Tree analysis
a. Organic trees
In Hajibabaei et al. [7], we pioneered a new visually-appealing technique for drawing organic-looking phylogenetic trees. This method maximizes resolution for tips of the tree (i.e. species), which are most important in barcode analysis. The process of building organic trees takes several hours and therefore we have been offering the creation of such trees as an e-mail service.

b. Tree collapse
This tool uses bootstrap values in a phylogenetic tree as a benchmark for visualizing statistical support of a given barcode dataset [10]. This is done by collapsing all the branches that are unsupported by a bootstrap cut-off value that is specified by the user. Although short barcode sequences are not strong phylogenetic markers at deep levels, they are excellent for species-level divergences. A high bootstrap cut-off (i.e. 100%) leads to collapsing most of the branches deeper than species-level, but the majority of the species-level branches are kept intact.
However, exceptionally closely related species may require longer sequences to gain a very high bootstrap support.

c. Tree tip colourization
This visualization tool uses a standard Newick format tree and colourizes the branches leading to individuals of each species (within-species distances) in red and the branches leading to each unique species in blue. It provides a robust method to visually compare different parts of a tree and therefore helps pinpointing exceptional divergence levels or regions of the tree that lack monophyly.

Server details
iBarcode.org is built on the Python-based web.py application framework [17]. Although most analyses are performed using Python itself, visualization and analysis are accomplished via calls to the statistical language R [18], the graphing package GraphViz [19], and the phylogenetic analysis package PAML [14]. We have intentionally kept the interface light and clean so that it loads quickly over low-bandwidth connections, and so that it is viewable and functional from text-based browsers (such as Lynx) or from small handheld devices (cell phones or PDAs).

In the future, we plan to have an application programming interface (API) for our tools, allowing other developers to integrate our analyses into their own tools.

Conclusion
Similarly to several other branches of biology, biodiversity science has increasingly been relying on DNA sequence information. DNA barcoding, as a new global initiative for biodiversity analysis, demands specialized bioinformatics tools and applications. iBarcode.org is a web-based application server developed for visualization and analysis of DNA barcode data. The suite of simple but highly customized tools in iBarcode.org allows the analysis and visualization of barcode data at sequence, genetic distance, and phylogenetic tree levels. Several of these applications have already contributed to barcode publications. iBarcode.org provides a web2.0 environment for developing and sharing tools for barcode data and sets the stage for a new wave of community driven bioinformatics applications.

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
GAS designed the server and developed various tools and applications and edited the manuscript. MH conceived the idea, designed several tools and applications and wrote the manuscript.

Acknowledgements
We acknowledge feedback and support from DNA barcode community especially during the 2nd International Barcode of Life Conference in Taipei (September 2007). We acknowledge the support from the Canadian Centre for DNA Barcoding (CCDB) and an award from the Consortium for Barcode of Life (CBOL).

This article has been published as part of BMC Bioinformatics Volume 10 Supplement 6, 2009: European Molecular Biology Network (EMBnet) Conference 2008: 20th Anniversary Celebration. Leading applications and technologies in bioinformatics. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/10?issue=S6.

References
1. Murphy WJ, Pevzner PA, O’Brien SJ: Mammalian phylogenomics comes of age. Trends Genet 2004, 20:631-639.
2. Marshall E: Taxonomy. Will DNA bar codes breathe life into classification? Science 2005, 307:1037.
3. Hebert PDN, Cywinska A, Ball SL, deWaard JR: Biological identifications through DNA barcodes. Proc Biol Sci 2003, 270:313-321.
4. Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN: DNA barcoding Australia’s fish species. Philos Trans R Soc Lond B Biol Sci 2005, 360:1847-1857.
5. Hajibaei M, Singer GA, Clare EL, Hebert PDN: DNA barcodes distinguish species of tropical Lepidoptera. Proc Natl Acad Sci USA 2006, 103:968-971.
8. Smith MA, Woodley NE, Janzen DH, Hallwachs W, Hebert PDN: DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). Proc Natl Acad Sci USA 2006, 103:3657-3662.

9. Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA: DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. Trends Genet 2007, 23:167-172.

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

11. Saitou N, Nei M: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987, 4:406-425.

12. Burns JM, Janzen DH, Hajibabaei M, Hallwachs W, Hebert PDN: DNA barcodes and cryptic species of skipper butterflies in the genus Perichares in Area de Conservacion Guanacaste, Costa Rica. Proc Natl Acad Sci USA 2008, 105:6350-6355.

13. McDonald JH, Kreitman M: Adaptive protein evolution at the Adh locus in Drosophila. Nature 1991, 351:652-654.

14. Yang Z: PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 1997, 13:555-556.

15. Yang Z, Nielsen R: Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 2000, 17:32-43.

16. Kumar S, Tamura K, Nei M: MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 2004, 5:150-163.

17. web.py [http://webpy.org]

18. R [http://www.r-project.org]

19. GraphViz [http://www.graphviz.org]

20. Clare EL, Lim BK, Engstrom MD, Eger JL, Hebert PDN: DNA barcoding of Neotropical bats: species identification and discovery within Guyana. Mol Ecol Notes 2007, 7:184-190.