

EvolMarkers: a database for mining exon and intron markers for evolution, ecology and conservation studies

CHENHONG LI,* JEAN-JACK M. RIETHOVEN† and GAVIN J. P. NAYLOR‡

*College of Fisheries and Life Science, Shanghai Ocean University, Shanghai 201306, China, †Bioinformatics Core Research Facility, University of Nebraska – Lincoln, Lincoln, NE 68588, USA, ‡Grice Marine Laboratory, College of Charleston, 205 Fort Johnson, Charleston, SC 29412, USA

Abstract

Recent innovations in next-generation sequencing have lowered the cost of genome projects. Nevertheless, sequencing entire genomes for all representatives in a study remains expensive and unnecessary for most studies in ecology, evolution and conservation. It is still more cost-effective and efficient to target and sequence single-copy nuclear gene markers for such studies. Many tools have been developed for identifying nuclear markers, but most of these have focused on particular taxonomic groups. We have built a searchable database, EvolMarkers, for developing single-copy coding sequence (CDS) and exon-primed-intron-crossing (EPIC) markers that is designed to work across a broad range of phylogenetic divergences. The database is made up of single-copy CDS derived from BLAST searches of a variety of metazoan genomes. Users can search the database for different types of markers (CDS or EPIC) that are common to different sets of input species with different divergence characteristics. EvolMarkers can be applied to any taxonomic group for which genome data are available for two or more species. We included 82 genomes in the first version of EvolMarkers and have found the methods to be effective across Placozoa, Cnidaria, Arthropod, Nematoda, Annelida, Mollusca, Echinodermata, Hemichordata, Chordata and plants. We demonstrate the effectiveness of searching for CDS markers within annelids and show how to find potentially useful intronic markers within the lizard *Anolis*.

Keywords: coding sequence, comparative genomics, exon-primed-intron-crossing, nuclear markers, phylogenetics, single-copy gene

Received 20 June 2011; revision received 23 May 2012; accepted 29 May 2012

Introduction

In the last decade, over a thousand genomes have been sequenced. Many more are projected to be completed soon (<http://www.genomesonline.org>). Next-generation sequencing is propelling us into a future in which there will be complete genomes for hundreds of thousands of interesting species.

Nevertheless, it is currently still too expensive to sequence the genome of every species and individual for most studies in ecology, evolution and conservation. Orthologous genes are still the markers of choice for most small-scale research projects. The majority are single-copy genes. The coding sequences (CDS) of those genes are often more conserved and suitable for studies across divergent taxa, whereas faster-evolving noncoding sequences are often chosen for within-species studies. Increasingly, researchers are using comparative genomic approaches to mine the growing number of completed

genomes in order to identify novel nuclear markers. New exon/CDS and intron markers have been developed for mammals (Ranwez *et al.* 2007), birds (Backström *et al.* 2008), turtles (Thomson *et al.* 2008), squamate reptiles (Townsend *et al.* 2008), fishes (Li *et al.* 2007, 2010), among many others.

Li *et al.* (2010) described a pipeline for mining exon-primed-intron-crossing (EPIC) markers in which CDS information is extracted from a well-annotated genome. The CDS are used as query sequences and BLASTed against the genome from which they were derived allowing the user to identify CDS that represent 'single-copy' fragments. The 'single-copy' sequences are then aligned against other genomes to find markers with the appropriate level of divergence for the taxonomic group under investigation. The information about the single-copy CDS can also be used to find introns flanked by those CDS. The pipeline has been designed to identify both CDS and EPIC markers.

Most of the existing tools, however, have been designed with a particular taxonomic group in mind. Applying them to other taxa generally requires

Correspondence: Chenhong Li, Fax: +86 21 61900263;
E-mail: chli@shou.edu.cn

adjustment of the search criteria, modification of the pipeline and repeating the same computation multiple times for different searches. It would be substantially more efficient if current protocols such as those deployed in the Li *et al.* (2010) pipeline could be modified such that BLAST searches were run once and hits stored for subsequent downstream analyses. This would allow CDS and EPIC markers to be extracted from BLAST hit results according to user-supplied criteria such as query type, CDS length and species. Such a framework would facilitate marker identification in any set of taxa for which two or more genome sequences were available. Here, we describe EvolMarkers, a database that implements this framework, and also make suggestions about how to carry out searches to obtain the best results.

Material and methods

Fasta files of genome sequences were downloaded from Ensembl (<http://www.ensembl.org/>), DOE Joint Genome Institute (<http://www.jgi.doe.gov/>), Beijing Genomics Institute (<http://www.genomics.cn/>), Human Genome Sequencing Center (<http://www.hgsc.bcm.tmc.edu/>) and NCBI (<http://www.ncbi.nlm.nih.gov/>). The gene annotation information was extracted from EMBL or from GFF files retrieved from those databases. The genomes used in this study are listed in Table 1 and Supplemental Table S1.

We followed the pipeline described in the study by Li *et al.* (2010) to construct the database. We pre-analysed the data and stored the BLAST results for subsequent searches according to user-supplied criteria. Four steps were used to construct the searchable database: (i) all genome sequences and annotation files are downloaded; (ii) within-species BLAST searches are carried out to identify single-copy CDS larger than 100 bp; (iii) resulting single-copy CDS sequences are BLASTed against a target genome; and (iv) user-provided parameters are taken from the web interface, to parse the BLAST results and post them to the output webpage. Sequence alignments for the selected markers are then made available for downloading.

The database will be updated as new genomes, and their associated annotations are made available by Ensembl. Users are provided with a web-based interface to send request for adding queries and target species of interest. Perl scripts automating these steps are available at <http://bioinformatics.unl.edu/cli/evolmarkers/>. Users can also use the pipeline to carry out tailored searches to reconstruct project-specific databases.

The approach described has been evaluated and applied to Placozoa, Cnidaria, Arthropod, Nematoda, Annelida, Mollusca, Echinodermata, Hemichordata,

Chordata and plants. We have included 82 genomes in the current EvolMarkers database (Table 1, Supplemental Table S1). Of these 82 genomes, 26 were used as queries, while all 82 were used as target species. However, many completed genomes have not been included in the current EvolMarkers database. New species and query–target combinations will be added when requested by the users through the online request form.

Results

The ‘Searching Markers’ page of EvolMarkers provides users with several options (Fig. 1). Users can choose the type of markers (CDS or EPIC) they want to find. If CDS is selected, the minimum length of the CDS sought can be changed. If EPIC is selected, the maximum intron length can be defined. Users must state the number of fasta files they wish to save and the ‘minimum sequence identity in the coding part of EPIC markers’. The program will return EPIC markers with conserved flanking exon regions, if a high identity value is selected. Users also have the option to query target genes across single or multiple species. Usually, the most closely related species to the species of interest is selected as the query species. Below, we present examples using the program to search for CDS and EPIC in a variety of taxonomic groups across different levels of divergence.

CDS markers for studying high-level phylogeny

There are more than 16 000 recognized species of annelids, including fireworms, earthworms, bloodworms, and leeches (Brusca & Brusca 2003). Annelids are important in ecosystems. However, many fundamental questions about annelids remain unresolved due to the lack of knowledge about their diversity and evolutionary relationships. The DOE Joint Genome Institute (<http://www.jgi.doe.gov/>) has sequenced the genomes of a polychaete worm (*Capitella teleta*) and a leech (*Helobdella robusta*). We have explored these two genomes with our EvolMarkers software. We set the minimum CDS length as 500 bp, used the polychaete worm genome to generate queries and used the leech genome as the target and found 16 CDS markers. We then used the leech genome to generate the queries and the polychaete worm as the target species and found 23 CDS markers. There were only four markers shared in the results of the two searches (Supplemental Table S2). The low number of shared markers found is likely due to inconsistencies associated with the draft annotations for the genome data of these two species. Combining the results from the two searches, we obtain 35 CDS markers in total for the annelids. Those markers have an average identity of 55–69% and a length of 404–1630 bp.

Table 1 Species included in the current version of EvolMarkers. The species in bold font are used as the queries. For the full list, see supplemental Table S1

Phylum	Class	Order	Species	Common name
Placozoa			<i>Trichoplax adhaerens</i>	
Cnidaria	Anthozoa	Actiniaria	<i>Nematostella vectensis</i>	Sea anemone
Cnidaria	Hydrozoa	Hydroida	<i>Hydra magnipapillata</i>	Hydra
Arthropoda	Arachnida	Acarina	<i>Ixodes scapularis</i>	Deer tick
Arthropoda	Branchiopoda	Diplosthraca	<i>Daphnia pulex</i>	Water flea
Arthropoda	Insecta	Coleoptera	<i>Tribolium castaneum</i>	Red flour beetle
Arthropoda	Insecta	Diptera	<i>Aedes aegypti</i>	Yellow fever mosquito
Arthropoda	Insecta	Diptera	<i>Anopheles gambiae</i>	African malaria mosquito
Arthropoda	Insecta	Diptera	<i>Drosophila melanogaster</i>	Common fruit fly
Arthropoda	Insecta	Diptera	<i>Culex quinquefasciatus</i>	Southern house mosquito
Arthropoda	Insecta	Hymenoptera	<i>Nasonia vitripennis</i>	Pteromalid parasitoid wasps
Arthropoda	Insecta	Hymenoptera	<i>Apis mellifera</i>	Western honey bee
Arthropoda	Insecta	Hymenoptera	<i>Harpegnathos saltator</i>	Ants
Arthropoda	Insecta	Lepidoptera	<i>Bombyx mori</i>	Silkworm
Arthropoda	Insecta	Phthiraptera	<i>Pediculus humanus</i>	Head louse
Arthropoda	Malacostraca	Amphipoda	<i>Parhyale hawaiiensis</i>	
Nematoda	Chromadorea	Diplogasterida	<i>Pristionchus pacificus</i>	
Nematoda	Chromadorea	Rhabditida	<i>Caenorhabditis briggsae</i>	
Nematoda	Chromadorea	Rhabditida	<i>Caenorhabditis elegans</i>	The elegant worm
Annelida	Clitellata	Rhynchobdellida	<i>Helobdella robusta</i>	Leech
Annelida	Polychaeta		<i>Capitella teleta</i>	Polychaete worm
Mollusca	Gastropoda	Anaspidia	<i>Aplysia californica</i>	California sea hare
Mollusca	Gastropoda	Patellogastropoda	<i>Lottia gigantea</i>	Owl limpet
Echinodermata	Echinoidea	Echinoidea	<i>Strongylocentrotus purpuratus</i>	Sea urchin
Hemichordata	Enteropneusta		<i>Saccoglossus kowalevskii</i>	Acorn worm
Chordata	Ascidacea	Enterogona	<i>Ciona intestinalis</i>	Sea squirt
Chordata	Ascidacea	Enterogona	<i>Ciona savignyi</i>	Sea squirt
Chordata	Cephalochordata	Amphioxiformes	<i>Branchiostoma floridae</i>	Amphioxus
Chordata	Cephalaspidomorphi	Petromyzontiformes	<i>Petromyzon marinus</i>	Lamprey
Chordata	Chondrichthyes	Chimaeriformes	<i>Callorhynchus milii</i>	Elephant shark
Chordata	Actinopterygii	Cypriniformes	<i>Danio rerio</i>	Zebrafish
Chordata	Actinopterygii	Beloniformes	<i>Oryzias latipes</i>	Medaka
Chordata	Actinopterygii	Gasterosteiformes	<i>Gasterosteus aculeatus</i>	Stickleback
Chordata	Actinopterygii	Tetraodontiformes	<i>Tetraodon nigroviridis</i>	Tetraodon
Chordata	Actinopterygii	Tetraodontiformes	<i>Takifugu rubripes</i>	Fugu
Chordata	Actinopterygii	Anguilliformes	<i>Anguilla anguilla</i>	European eel
Chordata	Actinopterygii	Perciformes	<i>Dicentrarchus labrax</i>	European seabass
Chordata	Amphibia	Anura	<i>Xenopus tropicalis</i>	
Chordata	Reptilia	Squamata	<i>Anolis carolinensis</i>	Anole lizard
Chordata	Aves	Galliformes	<i>Gallus gallus</i>	Chicken
Chordata	Aves	Passeriformes	<i>Taeniopygia guttata</i>	Zebra finch
Chordata	Mammalia	Afrosoricida	<i>Echinops telfairi</i>	Lesser hedgehog tenrec
Chordata	Mammalia	Hyracoidea	<i>Procavia capensis</i>	Hyrax

Searching for EPIC markers for ecological and conservation studies

The *Anolis* lizard is widely used as a model species to study adaptive radiation and convergent evolution (Johnson *et al.* 2010; Mahler *et al.* 2010). We used EvolMarkers to identify nuclear EPIC markers flanked by conserved protein-coding sequences in the recently completed genome sequence of *Anolis carolinensis* to facilitate such studies. On the initial search page, 'EPIC markers' was

selected (checked). The maximum intron length was left at the default value, 1000 bp. This dictates that intron length must be <1000 bp in at least one of the species. The identity in flanking exon was set to the default value of 85% identity. *Anolis carolinensis* was used to generate the queries, and *Gallus gallus* and *Taeniopygia guttata* were used as the target genomes. This search produced 1001 EPIC markers (Supplemental Data S1). In each fasta file, the position of the intron is indicated by 'XXXXX'. The intron length is indicated by the number appended at the

1st select the searching parameters

Are you looking for

☒ CDS markers? Minimum length of CDS (no less than 100 bp):

☐ or EPIC markers? Maximum intron length:

Minimum identity in the coding part of the EPIC markers to be saved: ☒ 90 ☐ 85 ☐ 75 ☐ 65Maximum number of markers (fasta files) you want to save: ☒ 100 ☐ 300 ☐ 500 ☐ 1000

2nd select the genomes you want to compare

Query genome (please select only one species)

Subject genome (select one or more species)

Fig. 1 Searching CDS markers using EvolMarkers.

end of each sequence. An example of EPIC sequence obtained is shown in Fig. 2.

Developing markers using expressed sequence tag (EST) or genome survey sequence (GSS) data

We have used our software to compare well-annotated genome sequences to EST or GSS data derived from other species. We used the genome of *Danio rerio* to generate the queries and an EST database of European eel, *Anguilla anguilla* (Coppe *et al.* 2010), as the target to search for EPIC markers. We found 282 EPIC markers (Supplemental Data S2). In a second trial, we used the genome of *Danio rerio* to generate the queries and GSS data of European bass, *Dicentrarchus labrax* (NCBI, <http://www.ncbi.nlm.nih.gov/nucgss>), as the target and found 81 CDS markers (Supplemental Data S3).

Discussion

Advances in genome sequencing and the associated growth of sequence databases have opened up opportunities to improve marker development for evolutionary biologists (Li *et al.* 2007, 2010; Ranwez *et al.* 2007; Backström *et al.* 2008; Thomson *et al.* 2008; Townsend *et al.* 2008). Thomson *et al.* (2010) reviewed the merit of different types of markers and discussed three strategies to develop them. Using EvolMarkers, one can filter the sin-

gle-copy CDS, complete computationally intensive BLAST jobs and save the results for subsequent analysis. Users may search BLAST output for different types of markers (CDS or EPIC) using different criteria, query and target species. EvolMarkers has already been used to develop CDS and EPIC markers in ray-finned fishes, sharks, cnidarians, rodents, ants and plants. Empirical tests of those markers are currently being evaluated in different laboratories. For example, ten CDS markers have been identified and amplified through all major Chondrichthyan lineages (Li *et al.* 2012), and fourteen fish species have been sequenced for 33 EPIC markers (X. Yao, personal communication).

A limitation of EvolMarkers is that at least one query genome with high coverage sequence data and good annotation is required. However, this will become less of a problem as more genomes come online. When using EvolMarkers, users can assign minimum length for the CDS markers. The default is set at 700 bp, based on cost-effectiveness for PCR and Sanger sequencing techniques. However, if more markers are needed, the minimum size of CDS can be lowered. Similarly, the default maximum intron size for EPIC markers is set to ensure that the intron is not too large to be amplified using PCR. Intron length can vary considerably across closely related species. We therefore recommend that users test EPIC markers empirically before embarking on a large EPIC-based study. This is particularly important if EST data are used

1. Gallus gallus	TCAATAACAGATTCCAACCAAAATTCCAAAGXXXXXATCTATGCTATTGGTGATGTAGTTGCTGG376-
2. Anolis carolinensis	TCAACAACAGATTCCAACCAAAAGTGCCAAAGXXXXXATTTATGCTATTGGCGATGTAGTCGCTGG1479
3. Taeniopygia guttata	TCAACAACAGATTCCAACCAAAAGATTCCAAGXXXXXATCTATGCTATTGGTGATGTAGTTGCTGG384-

Fig. 2 Example of an EPIC marker found for *Anolis* lizards.

to find EPIC markers, as EST sequences do not contain information about intron length.

When selecting query and target species, it is important to consider the relatedness of the query to the species of interest. Ideally, the query and target species should be closely related to the taxa of interest for the study, so that PCR primers have a better chance of working. It can also be useful to repeat the search using different queries as this can result in additional markers being found. This is particularly useful when the annotation of the query genomes is incomplete. For example, in our search for annelid markers, we found 16 CDS markers when using the polychaete worm as the query and found 23 CDS markers when using the leech as the query. When combining the results from the two searches, we obtain 35 different CDS markers in total for the annelids.

We anticipate that EvolMarkers will become increasingly useful for gene mapping and population genomic studies (Luikart *et al.* 2003; Thomson *et al.* 2010) as genome sequence databases continue to grow and genome annotation improves. Finally, we believe that the nuclear markers developed using EvolMarkers will prove useful when used in conjunction with next-generation sequencing and target-enrichment technologies (Perry *et al.* 2010; Pool *et al.* 2010).

Acknowledgements

C. Li was supported by National Science Foundation (DEB-0732838). J.-J. M. Riethoven was supported by grant EPSCoR EPS-0701892. G. J.P. Naylor was supported by National Science Foundation (1132229).

References

- Backström N, Fagerberg S, Ellegren H (2008) Genomics of natural bird populations: a gene-based set of reference markers evenly spread across the avian genome. *Molecular Ecology*, **17**, 964–980.
- Brusca RC, Brusca GJ (2003) *Invertebrates*. Sinauer Associates, Sunderland, Massachusetts.
- Coppe A, Pujolar JM, Maes GE *et al.* (2010) Sequencing, de novo annotation and analysis of the first *Anguilla anguilla* transcriptome: EelBase opens new perspectives for the study of the critically endangered European eel. *BMC Genomics*, **11**, 635.
- Johnson MA, Revell LJ, Losos JB (2010) Behavioral convergence and adaptive radiation: effects of habitat use on territorial behavior in *Anolis* lizards. *Evolution*, **64**, 1151–1159.
- Li C, Ortí G, Zhang G, Lu G (2007) A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. *BMC Evolutionary Biology*, **7**, 44.
- Li C, Riethoven J-JM, Ma L (2010) Exon-primed intron-crossing (EPIC) markers for non-model teleost fishes. *BMC Evolutionary Biology*, **10**, 90.
- Li C, Matthes-Rosana KA, Garcia M, Naylor GJP (2012) Phylogenetics of Chondrichthyes and the problem of rooting phylogenies with distant outgroups. *Molecular Phylogenetics and Evolution*, **63**, 365–373.

- Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.
- Mahler DL, Revell LJ, Glor RE, Losos JB (2010) Ecological opportunity and the rate of morphological evolution in the diversification of Greater Antillean anoles. *Evolution*, **64**, 2731–2745.
- Perry GH, Marioni JC, Melsted P, Gilad Y (2010) Genomic-scale capture and sequencing of endogenous DNA from feces. *Molecular Ecology*, **19**, 5332–5344.
- Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. *Genome Research*, **20**, 291–300.
- Ranwez V, Delsuc F, Ranwez S *et al.* (2007) OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evolutionary Biology*, **7**, 241.
- Thomson RC, Shedlock AM, Edwards SV, Shaffer HB (2008) Developing markers for multilocus phylogenetics in non-model organisms: a test case with turtles. *Molecular Phylogenetics and Evolution*, **49**, 514–525.
- Thomson RC, Wang IJ, Johnson JR (2010) Genome-enabled development of DNA markers for ecology, evolution and conservation. *Molecular Ecology*, **19**, 2184–2195.
- Townsend TM, Alegre RE, Kelley ST, Wiens JJ, Reeder TW (2008) Rapid development of multiple nuclear loci for phylogenetic analysis using genomic resources: an example from squamate reptiles. *Molecular Phylogenetics and Evolution*, **47**, 129–142.

C.L. developed the bioinformatics pipeline, wrote the scripts, analysed the data and wrote the manuscript. J.-J.M.R. helped with the Perl scripts and writing. G.J.P.N. wrote the manuscript. All authors have read and approved the final manuscript.

Data Accessibility

EvolMarkers and its documentation are freely accessible at <http://bioinformatics.unl.edu/cli/evolmarkers/>. Supplementary data are available at Molecular Ecology Resources online.

Supporting Information

Additional supporting information may be found in the online version of this article.

Data S1 1001 EPIC markers.

Data S2 282 EPIC markers.

Data S3 81 CDS markers.

Table S1 Species included in the current version of EvolMarkers.

Table S2 CDS markers found for annelids using EvolMarkers.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.