BMC Bioinformatics

BioMed Central

Proceedings

Open Access

EGenBio: A Data Management System for Evolutionary Genomics and Biodiversity

Laila A Nahum[1,2], Matthew T Reynolds[1], Zhengyuan O Wang[1], Jeremiah J Faith[3], Rahul Jonna[4], Zhi J Jiang[1], Thomas J Meyer[1] and David D Pollock*[1,5]

Address: [1]Department of Biological Sciences, Biological Computation and Visualization Center, Louisiana State University, Baton Rouge, LA 70803 USA, [2]Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA 02543, USA, [3]Bioinformatics Program, Boston University, Boston, MA 02215, USA, [4]Division of Developmental Disabilities, Arizona State Department, Phoenix, AZ 85012 USA and [5]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045, USA

Email: Laila A Nahum - lnahum@mbl.edu; Matthew T Reynolds - mreyno5@lsu.edu; Zhengyuan O Wang - zwang3@lsu.edu; Jeremiah J Faith - faith@bu.edu; Rahul Jonna - rjonna@gmail.com; Zhi J Jiang - zjiang1@lsu.edu; Thomas J Meyer - tmeyer5@lsu.edu; David D Pollock* - david.pollock@uchsc.edu

* Corresponding author

from The Third Annual Conference of the MidSouth Computational Biology and Bioinformatics Society
Baton Rouge, Louisiana. 2–4 March, 2006

Published: 26 September 2006

BMC Bioinformatics 2006, **7**(Suppl 2):S7 doi:10.1186/1471-2105-7-S2-S7

Abstract

Background: Evolutionary genomics requires management and filtering of large numbers of diverse genomic sequences for accurate analysis and inference on evolutionary processes of genomic and functional change. We developed *E*volutionary *Gen*omics and *Bio*diversity (*EGenBio*; http://egenbio.lsu.edu) to begin to address this.

Description: *EGenBio* is a system for manipulation and filtering of large numbers of sequences, integrating curated sequence alignments and phylogenetic trees, managing evolutionary analyses, and visualizing their output. *EGenBio* is organized into three conceptual divisions, *Evolution*, *Genomics*, and *Biodiversity*. The *Genomics* division includes tools for selecting pre-aligned sequences from different genes and species, and for modifying and filtering these alignments for further analysis. Species searches are handled through queries that can be modified based on a tree-based navigation system and saved. The *Biodiversity* division contains tools for analyzing individual sequences or sequence alignments, whereas the *Evolution* division contains tools involving phylogenetic trees. Alignments are annotated with analytical results and modification history using our *PRAED* format. A miscellaneous *Tools* section and *Help* framework are also available. *EGenBio* was developed around our comparative genomic research and a prototype database of mtDNA genomes. It utilizes MySQL-relational databases and dynamic page generation, and calls numerous custom programs.

Conclusion: *EGenBio* was designed to serve as a platform for tools and resources to ease combined analysis in evolution, genomics, and biodiversity.

Background

Large-scale genomic technologies have generated an extraordinary amount of data in the past few decades. Consequently, a huge effort has been made toward creating biological databases and systems to organize, analyze, and share information with the world-wide community [1-4]. The application of genomic technologies to molecular evolution has opened new frontiers in the interdisciplinary field of evolutionary genomics, and this has given rise to a great potential to elucidate complex questions in biology [2,5]. Understanding of evolutionary processes is critical, since they determine the sequence, structure, and function of macromolecules, and ultimately shape the higher-level biological complexity of organisms.

Genomic biodiversity has been defined as dense sampling of molecular data from diverse taxonomic groups for large genomic regions or complete genomes [6], and inferences concerning evolutionary processes are greatly improved by adopting combined molecular and computational approaches that include a large amount of genomic biodiversity [6-9]. We have found that the study of evolutionary genomics in the context of a dense sampling of species gives rise to many unique data processing problems, and so have developed the *E*volutionary *Gen*omics and *Bio*diversity (*EGenBio*) project as a web-based system to simplify large-scale evolutionary data management.

The central aim of *EGenBio* is to provide integrated analysis and visualization of raw sequence data, alignments, and phylogenetic trees, to rapidly curate and annotate that data, and to filter that data based on these annotations for further analysis of specific genomic contexts. It is designed to be robust to change and easily extensible to other datasets and other analytical programs. To accomplish this, *EGenBio* has web-based interfaces designed to: (1) access computational tools for phylogenetic and evolutionary analyses; (2) facilitate the construction of large-scale sequence and alignment datasets across diverse taxa; and (3) provide a framework for comparative analysis of diverse genes and genomes. *EGenBio* may also serve to promote the utility of increases in the scale of genomic biodiversity.

Overview

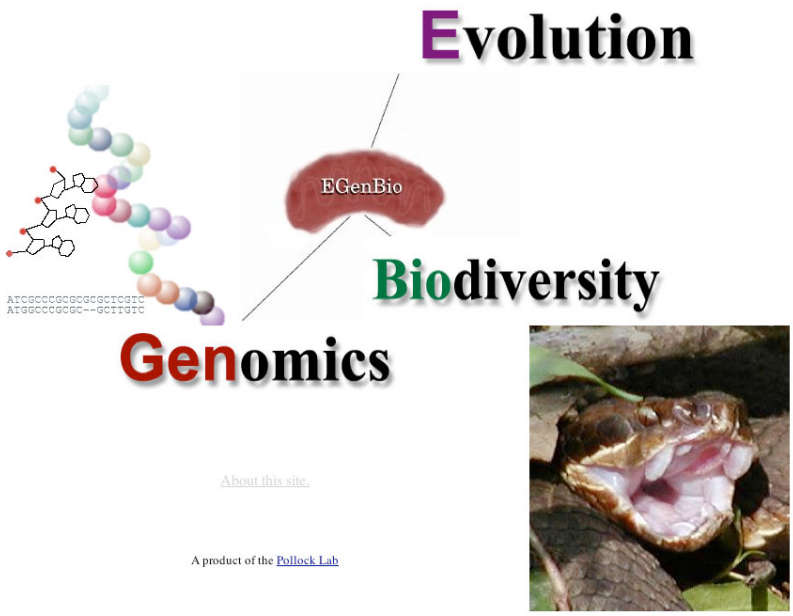
The three conceptual divisions of *EGenBio* (*Evolution*, *Genomics*, and *Biodiversity*) serve to organize web-based access to the primary custom-built tools (Figure 1, Table 1). The *Genomics* division provides access to pre-aligned sequence databases, and serves as a means of flexibly selecting genes and species/genomes of interest and producing a concatenated and annotated alignment for use in further analysis. The data is stored in a MySQL database and accessed "on the fly". Our prototype database contains complete vertebrate mitochondrial genomes that are mostly obtained from NCBI RefSeq annotations [10], but also contains "pre-submission" genomes provided by other investigators. Private access to "pre-submission" genomes allows users to pre-integrate their data with publicly available data for analysis in the primary genome publication, and is available on request. Complex searches can be stored, and registered users can create a personal list of searches that are retained in memory. Alignments for protein-coding genes are based on their amino acid sequences using ClustalW [11], but either amino acid or nucleotide alignments can be selected. The *Biodiversity* division is a collection of tools for analysis of alignments or sets of sequences (Table 1). These tools do not require a phylogenetic tree. Examples include lists of sequences available, database summaries, graphical representation of gene order information, and summary information on selected or uploaded alignments. This section also includes tools to aid experimental analysis of evolutionary genomics predictions, such as generating primers that represent all possible permutations of a set of sequences, or generating degenerate primers that reflect a sample from a posterior probability prediction of a particular ancestral sequence. The *Evolution* division includes tools (Table 1) that incorporate or extract information from phylogenetic trees and that translate from accession numbers to human-readable labels for sequences (e.g., genus species designations, or common names of organisms). This division also includes access to programs for processing and visualization of alignment filters (described below), coevolutionary analysis [12,13], and saturation mutagenesis analysis [14].

In addition to the three main sections, *EGenBio* contains a *Tools* section that serves as a repository for small standalone tools that may also exist as components of other pages, or which serve other simple purposes. The *Help* link leads not to a separate section, but rather to what we will call a separate "framework". The structure of the *Help* framework mirrors the main framework exactly, but instead of linking to actual tools, *Help* pages link to detailed descriptions and documentation for each page. Invisible to most users, a hidden *Design* framework allows for rapid editing and movement of page and site structure information from design to laboratory testing stages, and finally to public access.

Discussion and Conclusion

EGenBio is a web-based system for analysis in evolutionary genomics and biodiversity. It provides tools and resources for quickly creating, modifying, and analyzing large alignment datasets in ways that we have found useful in our own computer-based and experimental evolutionary genomics research. Our prototype database of complete vertebrate mitochondrial genomes represents the densest complete set of genes currently available from

A



B

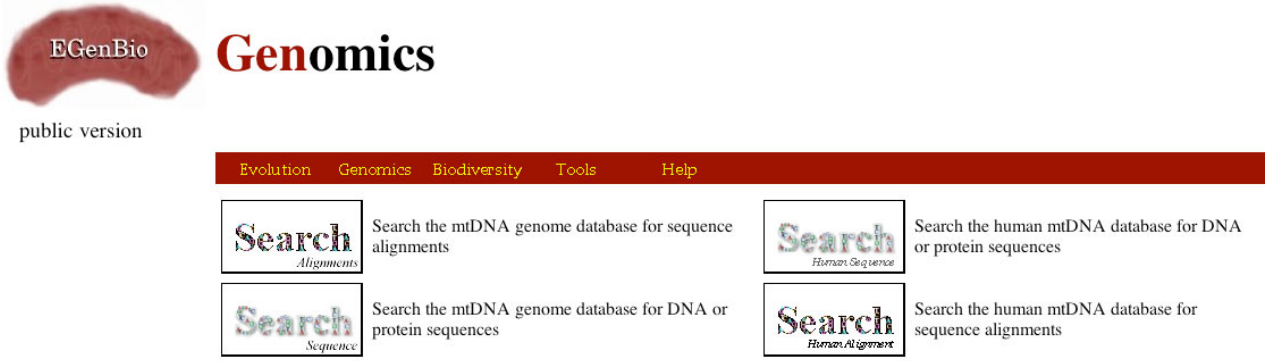


Figure 1
Organization of *EGenBio*. *EGenBio* is accessed through a splash page that links to the three main divisions, *Evolution*, *Genomics*, and *Biodiversity* (A). Each division has its own access page to the tools that are organized in that division. For example, the *Genomics* division page is shown in (B). The division of each page is clearly marked to allow quick movement among divisions and to the *Tools* section and mirrored *Help* pages.

closely related organisms. It can be accessed flexibly according to comma-separated queries or a phylogenetic tree navigation system. *EGenBio* is designed to be easily extensible to use with other protein complexes and other analytical programs. Our goal is to incorporate and utilize as many existing programs as possible, and to develop only "added value" programs. In the current public version, all tools are novel to our system except that alignments are created using ClustalW [11]. The *PRAED* alignment annotation system based on data filters allows alignments to be modified easily according to user interest in annotation features, and allows for the results of anal-

Table 1: Custom tools* currently in the main divisions of *EGenBio*

Division	Tool Name	Tool Description
Genomics	*SearchSequence*	Search the mtDNA genome database for DNA or protein sequences
Genomics	*SearchAlignment*	Search the mtDNA genome database for sequence alignments
Genomics	*SearchHuSequence*	Search the human mtDNA database for DNA or protein sequences
Genomics	*SearchHuAlignment*	Search the human mtDNA database for sequence alignments
Evolution	*TranslateTree*	Translate labels of a tree file
Evolution	*FilterViz*	Visualize filters associated with alignments
Evolution	*TreeReader*	Extract tree clusters along with information on branch lengths
Evolution	*SaturationTool*	Visualize results from saturation mutagenesis MCMC analysis
Evolution	*LnLCorr*	Detect coevolution among residues using LRTs and trees
Biodiversity	*SpeciesList*	List species currently in *EGenBio*
Biodiversity	*SpeciesSearch*	Search species by taxonomic group or NCBI genome identifier
Biodiversity	*LocusOrder*	Display mitochondrial gene order for specified taxa
Biodiversity	*DatabaseSummary*	Provide information about the *EGenBio* databases
Biodiversity	*PrimerPermuter*	Generate permutations for use in primer design
Biodiversity	*PrimerAlternatives*	Produce degenerate primers that reflect amino acid variation

*All tools listed are original to *EGenBio*, except that alignments are based on ClustalW [11].

yses to be returned as further annotations on the alignments. Since it is derived from the *NEXUS* format, it is easy to add batch commands to direct analyses using many common phylogenetic analysis programs. The *PRAED* format and data filters are a unique feature of the *EGenBio* system.

Future modules under development in *EGenBio* include the creation of additional data filters, incorporation of more genes for analysis of functional divergence, development of further visualization tools for statistical analyses of evolutionary dynamics, and automated procedures for analysis using existing programs and tools. We also welcome feedback from the scientific community on areas of general need for integrated evolutionary genomics tools. *EGenBio* is publicly available and can be accessed at http://egenbio.lsu.edu/ via anonymous login. User accounts that allow users to save search parameters and results are provided upon request. Incorporation and private access to pre-publication data can also be accommodated upon request. Replication of the *EGenBio* system would require a Linux-based operating system capable of running Perl, Perl-GD, R, PHP, MySQL, and an Apache web server. It would also require installation of numerous custom scripts in addition to ClustalW.

Abbreviations

LRT: Likelihood ratio test; MCMC: Markov chain Monte Carlo; mtDNA: mitochondrial DNA; NCBI: National Center for Biotechnology Information; *PRAED*: *PR*agmatic *A*nalysis of *E*volutionary *D*ata.

Authors' contributions

LAN is a scientific database curator responsible for the design and documentation of the *Help* and *Design* pages and organism database, and co-wrote this manuscript. MTR developed and managed the system and has created numerous tools. ZOW developed one of the tools and assisted in preparation of the manuscript. JJF began initial development of the system and developed several tools. RJ worked on the *Help* and *Design* pages, the mirror/framework system, and automated generation of pages. ZJJ developed one of the tools and assisted in preparation of the manuscript. TJM assisted in preparation of the manuscript, and review and editing of the web site, and is developing one of the tools. DDP is the principal investi-

gator responsible for the creation, conceptualization, and management of the *EGenBio* system, developed early versions of many of the tools, and co-wrote this manuscript.

Acknowledgements

This work was partly funded by the National Institutes of Health (R22/R33 Innovation and Development grant to David Pollock), the National Science Foundation (CBM²/EPSCOR), and the State of Louisiana (Biological Computation and Visualization Center, Governor's Biotechnology Initiative, and startup funds to David Pollock). We also anonymously thank other current and former members in the Pollock laboratory for assisting in the development and testing of various tools, and thank Chad Jarreau, Jonathan Bonin, Jonny Roberts Jr., Patricia Ledwig, Stephen McCullough, Sujatha Muralidharan, and Yonatan Platt for contributing to the *Biodiversity* image collection.

References

1. Basu S, Bremer E, Zhou C, Bogenhagen DF: **MiGenes: a searchable interspecies database of mitochondrial proteins curated using gene ontology annotation.** *Bioinformatics* 2006, **22(4):**485-492.
2. Crandall KA, Buhay JE: **Evolution. Genomic databases and the tree of life.** *Science* 2004, **306(5699):**1144-1145.
3. Galperin MY: **The Molecular Biology Database Collection: 2006 update.** *Nucleic Acids Res* 2006:D3-D5 [http://].
4. Vasconcelos AT, Guimaraes AC, Castelletti CH, Caruso CS, Ribeiro C, Yokaichiya F, Armoa GR, Pereira Gda S, da Silva IT, Schrago CG, Fernandes AL, da Silveira AR, Carneiro AG, Carvalho BM, Viana CJ, Gramkow D, Lima FJ, Correa LG, Mudado Mde A, Nehab-Hess P, Souza R, Correa RL, Russo CA: **MamMiBase: a mitochondrial genome database for mammalian phylogenetic studies.** *Bioinformatics* 2005, **21(10):**2566-2567.
5. Medina M: **Genomes, phylogeny, and evolutionary systems biology.** *Proc Natl Acad Sci USA* 2005, **102(Suppl 1):**6630-6635.
6. Pollock DD: **Genomic biodiversity, phylogenetics and coevolution in proteins.** *Appl Bioinformatics* 2002, **1(2):**81-92.
7. Faith JJ, Pollock DD: **Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes.** *Genetics* 2003, **165(2):**735-745.
8. Pollock DD, Bruno WJ: **Assessing an unknown evolutionary process: effect of increasing site-specific knowledge through taxon addition.** *Mol Biol Evol* 2000, **17(12):**1854-1858.
9. Pollock DD, Eisen JA, Doggett NA, Cummings MP: **A case for evolutionary genomics and the comprehensive examination of sequence biodiversity.** *Mol Biol Evol* 2000, **17(12):**1776-1788.
10. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Helmberg W, Kapustin Y, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2006:D173-180.
11. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22(22):**4673-4680.
12. Pollock DD, Taylor WR, Goldman N: **Coevolving protein residues: Maximum likelihood identification and relationship to structure.** *J Mol Biol* 1999, **287(1):**187-198.
13. Wang ZO, Pollock DD: **Context dependence and coevolution among amino acid residues in proteins.** *Methods Enzymol* 2005, **395:**779-790.
14. Pollock DD, Larkin JC: **Estimating the degree of saturation in mutant screens.** *Genetics* 2004, **168(1):**489-502.