Published online 13 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D519–D523***
doi:10.1093/nar/gkm973

Enteropathogen Resource Integration Center (ERIC): bioinformatics support for research on biodefense-relevant enterobacteria

Jeremy D. Glasner[1,*], Guy Plunkett III[2], Bradley D. Anderson[1], David J. Baumler[1], Bryan S. Biehl[1], Valerie Burland[1,2], Eric L. Cabot[1], Aaron E. Darling[3], Bob Mau[1], Eric C. Neeno-Eckwall[1], David Pot[4], Yu Qiu[5], Anna I. Rissman[1], Sara Worzella[1], Sam Zaremba[4], Joel Fedorko[4], Tom Hampton[4], Paul Liss[1], Michael Rusch[1], Matthew Shaker[4], Lorie Shaull[4], Panna Shetty[4], Silpa Thotakura[4], Jon Whitmore[4], Frederick R. Blattner[1,2], John M. Greene[4] and Nicole T. Perna[1,2]

[1]Genome Center, University of Wisconsin, 425G Henry Mall, Madison, Madison, WI 53703, [2]Laboratory of Genetics, University of Wisconsin, 425G Henry Mall, Madison, WI 53706, USA, [3]University of Queensland, Institute for Molecular Bioscience, St Lucia Q 4072, Australia, [4]SRA International, Inc., 11300 Rockville Pike, Suite 501, Rockville MD 20852 and [5]University of California, San Diego, Bioengineering, 9500 Gilman Drive, La Jolla, CA 92093, USA

Received September 28, 2007; Revised October 17, 2007; Accepted October 18, 2007

ABSTRACT

ERIC, the Enteropathogen Resource Integration Center (www.ericbrc.org), is a new web portal serving as a rich source of information about enterobacteria on the NIAID established list of Select Agents related to biodefense—diarrheagenic *Escherichia coli*, *Shigella* spp., *Salmonella* spp., *Yersinia enterocolitica* and *Yersinia pestis*. More than 30 genomes have been completely sequenced, many more exist in draft form and additional projects are underway. These organisms are increasingly the focus of studies using high-throughput experimental technologies and computational approaches. This wealth of data provides unprecedented opportunities for understanding the workings of basic biological systems and discovery of novel targets for development of vaccines, diagnostics and therapeutics. ERIC brings information together from disparate sources and supports data comparison across different organisms, analysis of varying data types and visualization of analyses in human and computer-readable formats.

INTRODUCTION

The family *Enterobacteriaceae* includes a variety of pathogens that pose significant threats to human health directly, and indirectly through agricultural crops and livestock. The Enteropathogen Resource Integration Center (ERIC, www.ericbrc.org) is one of the eight Bioinformatics Resource Centers (BRC) for Biodefense and Emerging/Re-Emerging Infectious Diseases (http://www.brc-central.org/). Funded by the National Institute of Allergy and Infectious Diseases (NIAID), ERIC serves as an information resource for enterobacteria on the NIAID established list of select agents related to biodefense—diarrheagenic *Escherichia coli*, *Shigella* spp., *Salmonella* spp., *Yersinia enterocolitica* and *Yersinia pestis*. ERIC seeks to support basic research on pathogenesis and development of novel vaccines, therapeutics and diagnostics for these organisms by:

(i) Adding value to genome data through manual and automated curation with particular focus on biological subsystems relevant to pathogenicity.
(ii) Integrating diverse sources of data ranging from publications on individual genes to large-scale proteomics data sets.
(iii) Developing tools for analyzing and visualizing these data.
(iv) Offering training and specialized analyses to the research community.

THE ERIC–BRC PORTAL OFFERS INTEGRATED ACCESS TO ALL TOOLS AND ANALYSES

The ERIC–BRC is a web portal that provides a single point of access to information about the focus organisms.

*To whom correspondence should be addressed. Tel: +1 608 890 0171; Fax: +1 608 890 0167; Email: jglasner@wisc.edu

Table 1. Genomes (all publicly available complete or draft sequences) contained in ERIC-ASAP as of August 2007

Organism	Complete	Draft	Total
Diarrheagenic *Escherichia coli*	5	7	12
Shigella spp.	8	2	10
Salmonella spp.	6	0	6
Yersinia enterocolitica	1	0	1
Yersinia pestis	6	10	16
Other related genomes	13	8	21
Total	39	27	66

The web portal, implemented with the JBoss Application Server 4.05GA and JBoss Portal Server 2.4.1, provides a single, standardized method of accessing the diverse resources integrated into the system. In addition to the specific resources described in the sections below, the portal provides general information about pathogenic enterobacteria, summaries of the genome database contents, and links to other relevant databases, such as the Immune Epitope Database (1) a curated set of epitopes for the Category A–C select agents. New functionalities and data sets are added to existing sections of the portal when appropriate or incorporated into new portlets within the main ERIC portal. This architecture permits rapid deployment of new components and customizable display of contents.

ERIC–ASAP GENOME ANNOTATIONS

ERIC provides access to continuously updated genome annotations for all ERIC pathogens, as well as information from a variety of other enterobacteria useful for reference and comparison, including *E. coli* K-12 (Table 1). ERIC uses the ASAP genome annotation database system (2) using an Oracle 10 g database for genome annotation and curation. ERIC–ASAP permits database updates continuously, obviating the need for periodic database releases that are a common feature of many genome databases. There are three general types of user accounts available for genome annotation purposes. Administrator accounts permit users the full range of capabilities including the ability to create new genome projects in the system. Curator accounts give users the ability to update ERIC annotations using sophisticated web-based interfaces for manual annotation and curation of information as well as tools for uploads of large sets of annotation data. Annotator accounts provide users with interfaces for manual annotation of individual annotation records. The annotation interfaces are all web-based and can be accessed by any member of the research community that requests an account. The availability of three different types of user accounts is designed to meet the needs of different types of annotators and to encourage training in use of the annotation tools that can be used to update large numbers of annotation records at a time. Genomes in ERIC can be either 'public' or 'private' projects, with users assigned to any of the three types of user accounts. All 'public' genome sequence data and annotations, including any newly added information, are accessible without an account.

Our goal is to provide genome annotations that are accurate, detailed, up-to-date and consistent across genomes. Descriptions of the standard operating procedures (SOPs) used by the ERIC curators are available for download from the portal (http://www.ericbrc.org/portal/eric/aboutasap). Every annotation record includes a description of the evidence supporting the data, and this is the primary way we assess the quality of the annotation information and measure improvements over time. Explanations of the evidence codes and how they are used can be found in the SOP describing gene annotation (http://www.ericbrc.org/portal/eric/sopCdsAnnotation). ERIC–ASAP is open for contribution by the research community to encourage annotation by domain experts. An additional layer of quality control is provided by a 'curation status' tag for each annotation that indicates whether the information has been independently approved by one of a select group of trusted users and dedicated curators.

Sequences and annotations in ERIC can be downloaded in a variety of formats including GenBank flatfile format and GFF3. Files downloaded directly from ERIC reflect continuous updates by the dedicated curatorial staff as well as community-contributed annotations. Snapshots of sequences annotated *de novo* by ERIC are also deposited in GenBank. Examples include the genome of *Y. pestis* strain CA88-4125 (GenBank accession number ABCD00000000) and plasmid pMAR7 from enteropathogenic *Escherichia coli* (3). ERIC is working toward an efficient mechanism for updates of existing GenBank and/or RefSeq records regardless of historical constraints. However, users should be aware that while ERIC provides support for documenting evidence for each individual line of annotation, this is not currently supported by NCBI.

ENTEROFAMS: PROTEIN FAMILIES FOR ENTEROBACTERIA

The first version of the EnteroFams is a collection of 1579 protein families. Each family is represented by a profile-Hidden Markov Model (HMM) similar to Pfam (4) or TIGRfam (5) protein families. EnteroFams differ from these other databases of protein families in that they contain only full-length alignments of proteins from enterobacterial species. The current collection of EnteroFams consist of proteins that are nearly ubiquitous in enterobacteria. Each HMM was constructed from an alignment of putative orthologous proteins from eight genomes (*E. coli* MG1655, *E. coli* EDL933, *Salmonella enterica* Typhimurium LT2, *S. enterica* Typhi CT18, *Y. pestis* CO92, *Yersinia pseudotuberculosis* 32953, *Erwinia chrysanthemi* 3937 and *Erwinia carotovora atroseptica* SCRI1043) and used to scan 11 additional genomes for new members. The threshold for inclusion in a family was defined as the lowest score obtained for a protein from one of the eight seed genomes. The alignment of the seed proteins, the complete alignment of all members and the annotations for each family were manually curated.

All members have a link to the associated EnteroFam page that contains alignments, cutoff thresholds and annotations for each family. EnteroFam HMMs will be made available from each EnteroFam page and for bulk download through the ERIC portal. Annotations were selected to be appropriate for all species so that they can be applied to all members of the family enabling the propagation of high-quality annotations across features in related enterobacteria.

ANNOTATION PROPAGATION: WHEN TO CUT, COPY AND PASTE?

The quality and quantity of annotation data varies within and between genomes. To reduce these inconsistencies, we would like to replace poor annotations with better information. We have developed an 'annotation propagation' tool within ERIC–ASAP to facilitate the comparison, evaluation and replacement of annotations across related genome features. This tool compares the text of annotation data between source and destination features as well as the evidence supporting the annotations. If the source feature annotation is supported by better evidence, the existing annotations for the destination feature are replaced with the annotation from the source feature. The user of the annotation propagation tool chooses the source and destination features and assigns the relative values to different categories of supporting evidence. Using this tool, high-quality annotations from well-curated genomes or protein families can be rapidly applied to other genomes while at the same time preserving any well-supported manual annotations that may already exist. The database retains a record of all annotations, regardless of their approval status, so no information is lost and can be reapplied as necessary.

Propagation of inaccurate or erroneous annotations has potential to do great harm to the quality of genome annotations. Care must be taken to ensure that only high-quality annotations are propagated across appropriate genome features. For example, propagation of annotations to members of EnteroFam families across genomes required that the annotation of the EnteroFam family be 'curated', and that membership in the family was approved by a curator. If a genome already contained an annotation with better supporting evidence, such as a gene product description with an experimental evidence code linked to a publication, the existing annotation was preserved. New annotations added by the propagation procedure all contain an indication that they were added by an automated process and have a link to the SOP describing the procedure.

INSERTION SEQUENCES: ANNOTATING JUMPING GENES

Insertion Sequence (IS) element activity is a significant source of variation between genomes. We have annotated the boundaries of 3412 intact IS elements and 758 IS fragments in a core set of 20 complete genomes based on known IS element sequences collected in the ISfinder Database (6). We have not annotated the IS elements as thoroughly in draft genome sequences since IS elements frequently occur at contig boundaries and are often misassembled in draft sequences. All IS annotations include the identity of the IS element and a link to the related entry in the ISfinder database. IS feature names are assigned to distinguish between multiple copies of the same IS within one genome. The IS annotations can be viewed in ERIC's genome viewers (described below) to examine differences in IS content between genomes.

HIGH-THROUGHPUT EXPERIMENTAL DATA SETS

The ERIC–ASAP database stores and displays results of high-throughput gene expression experiments and proteomics data. For example, ERIC–ASAP contains newly discovered proteomic information about NIAID's Category A–C biodefense organisms obtained through the NIAID-funded Proteomics Research Centers (PRCs). Information about the presence and absence of *S. enterica* Typhimirium proteins under different growth conditions (7) can be obtained from each gene annotation page or downloaded in bulk from ERIC–ASAP. Links are provided from each page to more detailed information about the mass-spectrometry data available at the Administrative Resource Center (http://www.proteomicsresource.org/).

ERIC COMPARATIVE GENOMICS TOOLS

Comparative genomics is a powerful way to identify genes conserved among subsets of related pathogens as well as sequences that differentiate strains and species. ERIC has several components that facilitate genome comparisons and classification of relationships between genes within and across genomes.

Multiple genome alignments

Multiple genome alignments are available for each species of enterobacteria represented in ERIC–BRC (Figure 1). These alignments were constructed using a newly released progressive alignment tool, Mauve 2.0 (8,9), that dramatically improves alignment in regions conserved among subsets of genomes, a particularly important feature for recognition of genomic islands. This new version has significantly improved visualization and navigational tools and provides a powerful mechanism for comparative genomics of bacterial genomes.

Genome alignments are currently available for:

(i) Six complete *Escherichia* genomes
(ii) Ten complete *Escherichia* and *Shigella* genomes
(iii) Five complete *Salmonella* genomes
(iv) Seven complete *Yersinia* genomes

Each alignment includes all available complete published genome sequences as of April 2007 with links directly from the graphical gene display to the annotations in ERIC. As new genomes become available, these alignments will be updated. Older versions will be archived and remain available.

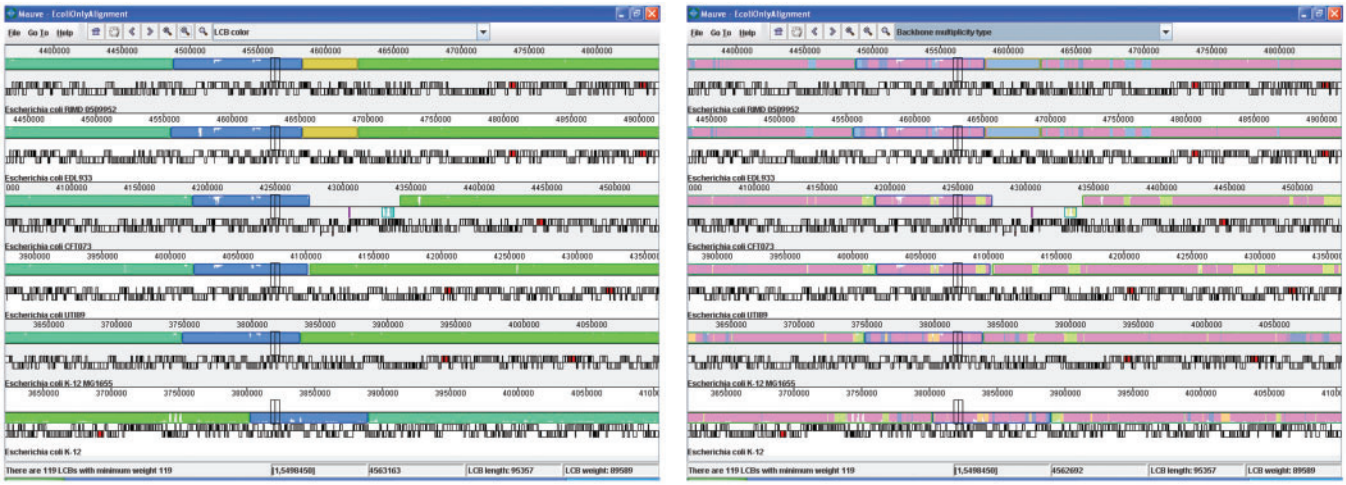

Figure 1. Two views of the same region of the Mauve 2.0 alignment of 6 *E. coli* genomes. The visualization on the left uses the default color scheme based on homologous segments. Each color represents a collinear block that contains regions of homologous sequence. Importantly, islands unique to a single genome or collinear islands common to a subset of genomes are indicated. The visualization of the same aligned region shown on the right is colorized by multiplicity. Here, pink blocks indicate that the region is conserved across all six genomes. Other colors mark regions found in a subset of genomes.

Curated ortholog sets

The ERIC–ASAP database maintains curated sets of proteins predicted to be orthologous. Initial sets of orthologs are constructed by analysis of pairwise BLAST searches between genomes. As described in more detail in the ERIC Ortholog SOP, this process is limited to cases where there is a single unambiguous best reciprocal match, and filtered according to empirically selected comparison-specific thresholds for percent identity and proportion of the proteins aligned. This results in a conservative set of predicted orthologs. These sets are augmented and confirmed by manual review and additional processes such as confirmation of co-linearity in genome alignments.

Generic Genome Browser (Gbrowse)

ERIC provides the GBrowse (10) for querying and viewing genomic data and linking to annotations within the database. Users can search for genes across genomes and zoom in or out on genome maps. There are several tracks of annotation data that can be selected to be displayed. We plan to expand the visualizations available through GBrowse to include high-throughput data sets as well as large-scale bioinformatics predictions such as transcriptional units and regulatory protein-binding sites.

MICROARRAY ANALYSIS SYSTEM

The mAdb microarray database and analysis system (11) is a core component of ERIC. This is a web-based system that supports sharing of data between groups and includes a microarray storage database and a variety of built-in analysis tools. mAdb can import Affymetrix data as well as spotted arrays, and can use the quantitation and composite image files from a number of microarray scanners. The central concept for mAdb is that of creating filtered, reusable data sets for analyzing microarray data. Once the raw data is processed and placed in ERIC's relational database, a user can filter the data for quality, using a variety of filters for spot size, signal/background ratio, excluding those spots marked as 'Bad' or 'Not Found' by the scanner software, as well as a number of other quantitative metrics. Normalization can be done either on the raw data or only on those spots which pass the spot quality criteria set by the user. This creates a parent-filtered data set, which can be further filtered in other ways, such as by expression ratios, by genes (rows) or by arrays (columns), or used directly in the analysis tools. Each data set maintains a history associated with it, so users can see how it was derived.

The analysis tools allow hierarchical, K-means and self-organizing map-based clustering of the data by a number of metrics and linkage methods, as well as other related visualization techniques such as scatter plotting, Principal Components Analysis, and Multidimensional Scaling. Graphics can be exported for publication and protocols for MIAME-formatted data can be stored.

Microarray data is linked to corresponding annotated features in ERIC genomes to provide a way to access up-to-date annotation records while viewing data in mAdb. There are security controls for access to data in mAdb. All users must register for an account. With an account, users have access to publicly available projects and their own private workspace that is only available to themselves and other users that they grant access to the project. Users can collaborate on experiments and analyses in their secure workspace, and if they desire, make data available for analysis by all users.

FUTURE DIRECTIONS

Genome sequencing is ongoing at several institutions for a number of additional strains and isolates of pathogenic

enterobacteria and ERIC will continue to incorporate new sequence data as it becomes available. We plan to continue our efforts of careful manual annotation for these organisms to provide high-quality information that is supported by direct experimentation. The number of genome sequences for these pathogens already available and the large number of new sequences anticipated suggests that manual inspection of every annotation is an impossible task. For this reason, we will focus annotation efforts on a few reference genomes for each group of pathogens as well as continue to carefully annotate protein families from the EnteroFams. Judicious application of the annotation propagation tool will be used to distribute these carefully curated annotations to other genomes.

Use of consistent vocabulary to describe biological entities and functions is critical for comparison of annotations within and between genomes. The Gene Ontology (GO) Consortium is a group dedicated to creating and applying a structured and controlled vocabulary for describing gene products, their functions and locations (12,13). The current annotations of genomes in ERIC contain limited use of the GO, and we plan to expand this in the future.

The use of high-throughput experiments to characterize bacterial genes, proteins and metabolites is increasing and ERIC will continue to integrate these types of data and provide tools for analysis and visualization. The ERIC portal is continually under development to improve data content and usability. The goal of integration of information within ERIC is to provide researchers with a simple-to-use, richly populated database of accurate genome annotation and associated data that will aid in creation of novel diagnostics, therapeutics and vaccines to mitigate the threats posed by pathogenic enterobacteria.

ACKNOWLEDGEMENTS

We thank M. Chandler for providing a copy of the ISfinder database. The ERIC Bioinformatics Resource Center has been wholly funded with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN266200400040C. Funding to pay the Open Access publication charges for this article was provided by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN266200400040C.

Conflict of interest statement. None declared.

REFERENCES

1. Peters,B., Sidney,J., Bourne,P., Bui,H.H., Buus,S., Doh,G., Fleri,W., Kronenberg,M., Kubo,R. *et al.* (2005) The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol.*, **3**, e91.
2. Glasner,J.D., Rusch,M., Liss,P., Plunkett,G.III, Cabot,E.L., Darling,A., Anderson,B.D., Infield-Harm,P., Gilson,M.C. *et al.* (2006) ASAP: a resource for annotating, curating, comparing, and disseminating genomic data. *Nucleic Acids Res.*, **34**, D41–D45.
3. Brinkley,C., Burland,V., Keller,R., Rose,D.J., Boutin,A.T., Klink,S.A., Blattner,F.R. and Kaper,J.B. (2006) Nucleotide sequence analysis of the enteropathogenic Escherichia coli adherence factor plasmid pMAR7. *Infect. Immun.*, **74**, 5408–5413.
4. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
5. Selengut,J.D., Haft,D.H., Davidsen,T., Ganapathy,A., Gwinn-Giglio,M., Nelson,W.C., Richter,A.R. and White,O. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res.*, **35**, D260–D264.
6. Siguier,P., Perochon,J., Lestrade,L., Mahillon,J. and Chandler,M. (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.*, **34**, D32–D36.
7. Adkins,J.N., Mottaz,H.M., Norbeck,A.D., Gustin,J.K., Rue,J., Clauss,T.R., Purvine,S.O., Rodland,K.D., Heffron,F. *et al.* (2006) Analysis of the Salmonella typhimurium proteome through environmental response toward infectious conditions. *Mol. Cell Proteomics*, **5**, 1450–1461.
8. Darling, A.E. (2006) Computational analysis of genome evolution. *Ph.D. Thesis*. University of Wisconsin, Madison 2006.
9. Darling,A.E., Mau,B., Blattner,F.R. and Perna,N.T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.*, **14**, 1394–1403.
10. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
11. Greene,J.M., Asaki,E., Bian,X., Bock,C., Castillo,S., Chandramouli,G., Martell,R., Meyer,K., Ruppert,T. *et al.* (2003) The NCI/CIT microArray database (mAdb) system – bioinformatics for the management and analysis of Affymetrix and spotted gene expression microarrays. In *AMIA Annu. Symp. Proc.,* p. 1066.
12. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
13. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.