D296–D299 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh031

*coli*BASE: an online database for *Escherichia coli*, *Shigella* and *Salmonella* comparative genomics

Roy R. Chaudhuri, Arshad M. Khan and Mark J. Pallen*

Bacterial Pathogenesis and Genomics Unit, Division of Immunity and Infection, Medical School, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

Received August 13, 2003; Revised August 18, 2003; Accepted September 4, 2003

ABSTRACT

We have constructed *coli*BASE, a database for *Escherichia coli*, *Shigella* and *Salmonella* comparative genomics available online at http://colibase.bham.ac.uk. Unlike other *E.coli* databases, which focus on the laboratory model strain K12, *coli*BASE is intended to reflect the full diversity of *E.coli* and its relatives. The database contains comparative data including whole genome alignments and lists of putative orthologous genes, together with numerous analytical tools and links to existing online resources. The data are stored in a relational database, accessible by a number of user-friendly search methods and graphical browsers. The database schema is generic and can easily be applied to other bacterial genomes. Two such databases, *Campy*DB (for the analysis of *Campylobacter* spp.) and *Clostri*DB (for *Clostridium* spp.) are also available at http://campy.bham.ac.uk and http://clostri.bham.ac.uk, respectively. An example of the power of *E.coli* comparative analyses such as those available through *coli*BASE is presented.

INTRODUCTION

Over the past few decades nucleotide sequence data have been accumulating rapidly, driven in recent years by the many complete genome sequencing projects. The comprehensive public databases GenBank (1), EMBL (2) and DDBJ (3) are invaluable in providing access to the raw sequence data and annotation. However, such databases are broad in scope, and there is a gap in the market for smaller curated databases focusing on a particular organism or type of data. As might be expected, the ubiquitous model organism *Escherichia coli* is the focus of many such databases, for example EcoCyc (4), EcoGene (5), Colibri (6), Genobase (7), ECDC (8), RegulonDB (9) and EchoBase (10), to name but a few. The focus of these resources is on the laboratory *E.coli* strain K12, the complete genome of which was sequenced by Blattner *et al*. (11). However, *E.coli* is more than just a model organism. The species *E.coli* incorporates a wide variety of diverse strains and pathotypes, including members of the phylogenetically indistinguishable 'genus' *Shigella* (12). Although most *E.coli* are harmless, many of the strains are pathogenic and cause a variety of diseases in humans and animals (13). Members of the closely related genus *Salmonella* are also important pathogens and the focus of a widespread research effort. Genome sequence data are accumulating for *E.coli*, *Salmonella* and *Shigella*, with 10 complete and annotated genomes currently available (7,11,14–21), together with extensive raw data from 13 genome projects still in progress, several of which are essentially complete (>99% genome coverage). Hitherto there has been no easily accessible tool for comparative browsing and analysis of these genomes.

Here we announce the creation of *coli*BASE, an online resource for *E.coli*, *Shigella* and *Salmonella* comparative genomics, available at http://colibase.bham.ac.uk. Unlike the other *E.coli* databases *coli*BASE is intended to act as a repository for sequence data, annotation and analyses on the full diversity of *E.coli* and its relatives. In the post-genomic era much biological insight can be gained from comparisons between closely related genome sequences, and *coli*BASE is intended to provide the *E.coli* research community with user-friendly access to such analyses.

DATA IN *coli*BASE

The database includes all currently available complete and incomplete *E.coli*, *Shigella* and *Salmonella* genomes. The sequence data are archived using the freely available mySQL relational database management system. Also stored within the database is information concerning sequence features, including annotation from both GenBank (1) and Swiss-Prot (22), references to relevant articles in the literature, and the results of codon usage analyses performed using CodonW. These include information on raw base composition, base composition at synonymous third codon positions, the hydropathicity and aromaticity of the encoded protein, and the Codon Adaptation Index (CAI) value, which has been shown to be associated with expression level (23). Comparative data stored include whole genome alignments generated using MUMmer and PROmer (24). Also available are lists of putative orthologous genes; these genes were identified as 'mutual best hits' during reciprocal BLASTP (25) searches of genome pairs, with the additional requirements that putative orthologues should show >80% identity at the protein level, and that the aligned portion should cover at least 90% of the length of the shorter sequence. For the unfinished

*To whom correspondence should be addressed. Tel: +44 121 414 7163; Fax: +44 121 414 3454; Email: m.pallen@bham.ac.uk

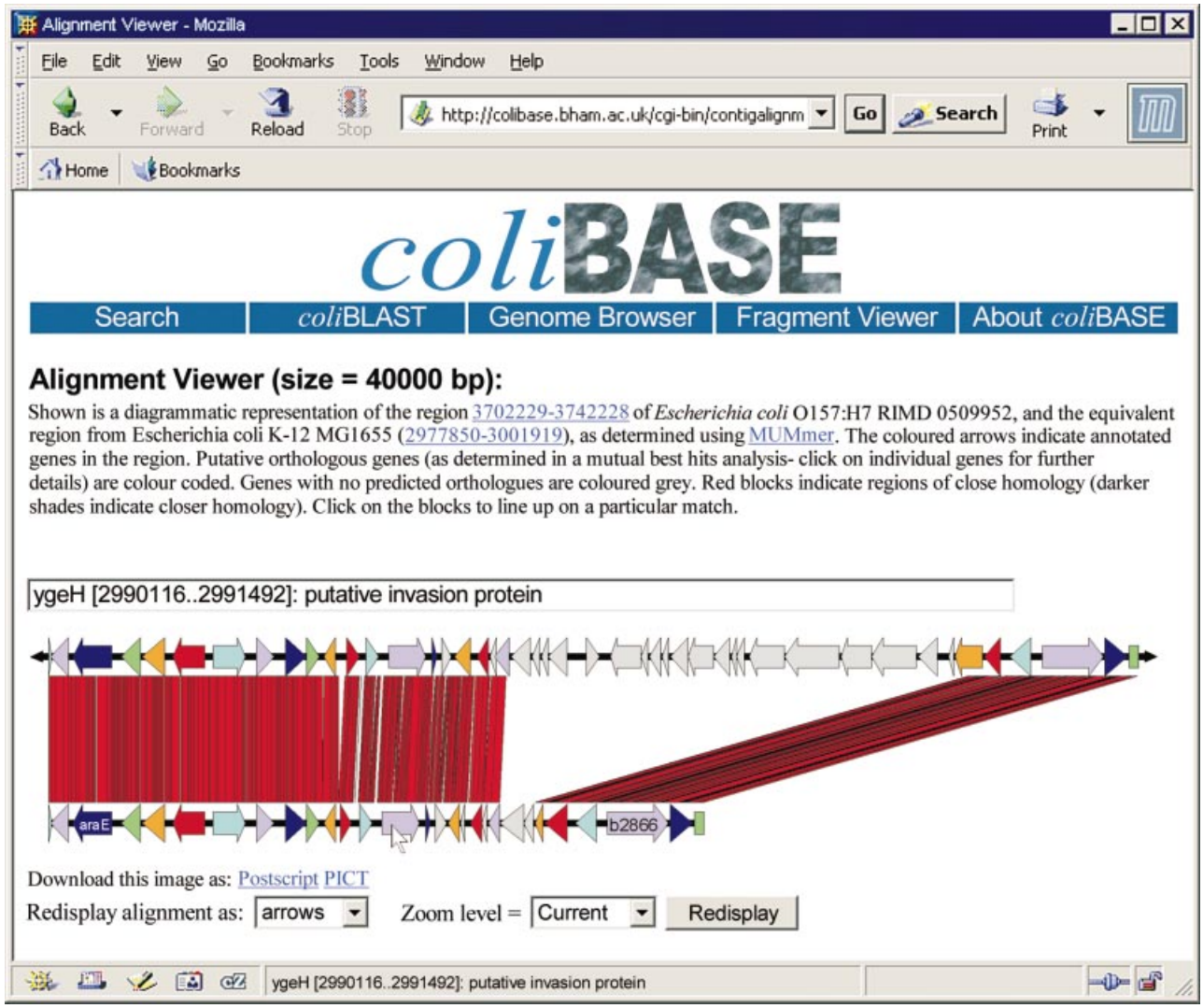

Figure 1. Screenshot of the Alignment Viewer, showing a MUMmer (24) comparison of positions 3702229–3742228 from *E.coli* O157:H7 RIMD 0509952 (Sakai) with the equivalent region from *E.coli* K12 MG1655 (positions 2977850–3001919). This region of the O157:H7 genome contains the type III secretion system ETT2 within O-island 122 (16,19). However, examination of the upstream regions, and comparisons with *E.coli* CFT073 (see Supplementary Material) suggest that a remnant of ETT2 is retained in the K12 genome.

genomes, information about the position of predicted open reading frames, as determined using GLIMMER (26), is also available.

A representation of the database schema is available in the Supplementary Material. The schema is generic and can be easily applied to other bacterial systems. Two such spin-off databases are already in operation: *Campy*DB, for analysis of *Campylobacter* genomes (http://campy.bham.ac.uk), and *Clostri*DB, for comparative analyses of *Clostridium* spp. (http://clostri.bham.ac.uk). Similar databases for other important groups of pathogenic bacteria will be created in the near future. These will include databases for the comparative analysis of pseudomonads, mycobacteria, staphylococci, streptococci, *Chlamydia* and *Rhizobium*/*Sinorhizobium*.

USER INTERFACE

Many of the currently available bioinformatic resources, though powerful, are not readily accessible to bench biologists who may have limited computational experience. Analytical programs are often limited to UNIX platforms and driven by an unfriendly command line interface. To widen accessibility it is desirable to make new resources available via the world wide web, using server-side scripts wherever possible to allow access to those who have not upgraded to the most recent browsers. For this reason we have constructed a web interface to *coli*BASE using Perl/CGI. The main search page consists of a single search box, instantly familiar to users of internet search engines, with searches performed on gene name and annotation. A help page details the use of wildcards to widen searches. The search results page provides links to the individual gene pages for all genes that match the query, together with links for genes (from both complete and unfinished genomes) that did not match the query but are putative orthologues of the genes that did.

Two further search methods are available: an advanced search page, which allows queries to be restricted to particular genomes and database fields, and the *coli*BLAST search page,

which enables databases of the gene, protein and genome sequences in *coli*BASE to be searched using nucleotide or amino acid query sequences. Additionally the Genome Browser and Fragment Viewer pages allow regions of interest to be selected by genomic coordinate.

DATA VISUALIZATION

Novel tools have been developed to visualize the sequence and alignment data stored within *coli*BASE. The Perl module Image::Magick is used to generate a graphical representation of the chromosomal regions surrounding a gene of interest. These are covered by an image map, allowing the user to scan over the image with the mouse pointer to view annotation of the flanking genes. An example of the alignment viewer is shown in Figure 1. All the images in *coli*BASE can be downloaded in PostScript or PICT format, to allow editing in external graphics programs. These formats are generated using the Postscript::Simple Perl module and the qd.pl Perl library, respectively.

ANALYSIS TOOLS

A number of tools are provided to allow the user to further analyse a gene or chromosomal region of interest. These include facilities to obtain the raw sequence data, and to search against the *coli*BLAST databases. A primer design facility, using a customized version of Primer3 (27) allows the rapid design of primers to amplify within a particular region, and this is linked through to *coli*BLAST, to allow the designed primers to be tested for specificity. The applet version of Artemis (28) is integrated into *coli*BASE, allowing rapid access to the powerful visualization and analytical features of this program without the need to download and install additional software. Additionally, *coli*BASE is linked to a number of external resources, allowing it to act as a portal to direct users to relevant online information. These resources include our own ViruloGenome (http://www.vge.ac.uk) for PSI-BLAST searches of incomplete genome sequence data; the NCBI's CDD (29) for the identification of conserved domains within a protein sequence; PubMed, to enable rapid searches for relevant publications in addition to those stored in *coli*BASE; and the *E.coli* specific databases RegulonDB, EcoCyc, EcoGene, Genobase and Colibri.

CASE-STUDY: TYPE-III SECRETION GENES IN *E.coli* K12

The comparative genomics approach that underpins *coli*BASE clearly sheds light on the differences between commensal strains such as K12 and pathogenic strains of *E.coli*. However, using *coli*BASE to align and compare genomes can also provide a functional and evolutionary context for cryptic genes in the model strain, *E.coli* K12. For example, previous authors noted the existence of K12 homologues of genes from the Spi-1 and Spi-3 pathogenicity islands from *Salmonella enterica*, and in particular remarked on the presence in this commensal strain of genes usually associated with type III secretion (30,31). Why these genes occur in K12 remains a mystery until genomic comparisons with several pathogenic strains are performed (see Fig. 1 and Supplementary Material). It then becomes clear that these genes represent 'baggage of history', i.e. a remnant of a much larger pathogenicity island (termed ETT2 by some authors) that potentially encodes a full type-III secretion system in some pathovars.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We would like to acknowledge the contribution of Charles Penn and the other members of the University of Birmingham *E.coli* group (UBEC). *coli*BASE is affiliated to the *E.coli* Index (http://ecoli.bham.ac.uk) through a collaboration with Gavin Thomas (University of York). We thank the British Biotechnology and Biological Sciences Research Council for funding the *coli*BASE, *Campy*DB and ViruloGenome projects (grant references FGT11398, EGA16107 and EGA16174).

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
2. Stoesser,G., Baker,W., van den Broek,A., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V., Lopez,R. *et al.* (2003) The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids Res.*, **31**, 17–22.
3. Miyazaki,S., Sugawara,H., Gojobori,T. and Tateno,Y. (2003) DNA Data Bank of Japan (DDBJ) in XML. *Nucleic Acids Res.*, **31**, 13–16.
4. Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S.M., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002) The EcoCyc Database. *Nucleic Acids Res.*, **30**, 56–58.
5. Rudd,K.E. (2000) EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 60–64.
6. Medigue,C., Viari,A., Henaut,A. and Danchin,A. (1993) Colibri: a functional data base for the *Escherichia coli* genome. *Microbiol. Rev.*, **57**, 623–654.
7. Mori,H., Isono,K., Horiuchi,T. and Miki,T. (2000) Functional genomics of *Escherichia coli* in Japan. *Res. Microbiol.*, **151**, 121–128.
8. Kroger,M. and Wahl,R. (1998) Compilation of DNA sequences of *Escherichia coli* K12: description of the interactive databases ECD and ECDC. *Nucleic Acids Res.*, **26**, 46–49.
9. Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millan-Zarate,D., Diaz-Peredo,E., Sanchez-Solano,F., Perez-Rueda,E., Bonavides-Martinez,C. and Collado-Vides,J. (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.*, **29**, 72–74.
10. Thomas,G.H. (1999) Completing the *E. coli* proteome: a database of gene products characterised since the completion of the genome sequence. *Bioinformatics*, **15**, 860–861.
11. Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
12. Pupo,G.M., Karaolis,D.K., Lan,R. and Reeves,P.R. (1997) Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and *mdh* sequence studies. *Infect. Immun.*, **65**, 2685–2692.
13. Donnenberg,M.S. and Whittam,T.S. (2001) Pathogenesis and evolution of virulence in enteropathogenic and enterohemorrhagic *Escherichia coli*. *J. Clin. Invest.*, **107**, 539–548.
14. Deng,W., Liou,S.R., Plunkett,G.,3rd, Mayhew,G.F., Rose,D.J., Burland,V., Kodoyianni,V., Schwartz,D.C. and Blattner,F.R. (2003) Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. *J. Bacteriol.*, **185**, 2330–2337.
15. Jin,Q., Yuan,Z., Xu,J., Wang,Y., Shen,Y., Lu,W., Wang,J., Liu,H., Yang,J., Yang,F. *et al.* (2002) Genome sequence of *Shigella flexneri* 2a:

insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res.*, **30**, 4432–4441.
16. Hayashi,T., Makino,K., Ohnishi,M., Kurokawa,K., Ishii,K., Yokoyama,K., Han,C.G., Ohtsubo,E., Nakayama,K., Murata,T. *et al.* (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.*, **8**, 11–22.
17. Welch,R.A., Burland,V., Plunkett,G.,3rd, Redford,P., Roesch,P., Rasko,D., Buckles,E.L., Liou,S.R., Boutin,A., Hackett,J. *et al.* (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **99**, 17020–17024.
18. Wei,J., Goldberg,M.B., Burland,V., Venkatesan,M.M., Deng,W., Fournier,G., Mayhew,G.F., Plunkett,G.,3rd, Rose,D.J., Darling,A. *et al.* (2003) Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. *Infect. Immun.*, **71**, 2775–2786.
19. Perna,N.T., Plunkett,G.,3rd, Burland,V., Mau,B., Glasner,J.D., Rose,D.J., Mayhew,G.F., Evans,P.S., Gregor,J., Kirkpatrick,H.A. *et al.* (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**, 529–533.
20. Parkhill,J., Dougan,G., James,K.D., Thomson,N.R., Pickard,D., Wain,J., Churcher,C., Mungall,K.L., Bentley,S.D., Holden,M.T. *et al.* (2001) Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature*, **413**, 848–852.
21. McClelland,M., Sanderson,K.E., Spieth,J., Clifton,S.W., Latreille,P., Courtney,L., Porwollik,S., Ali,J., Dante,M., Du,F. *et al.* (2001) Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature*, **413**, 852–856.
22. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
23. Sharp,P.M. and Li,W.H. (1987) The codon Adaptation Index—a measure of directional synonymous codon usage bias and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.
24. Delcher,A.L., Phillippy,A., Carlton,J. and Salzberg,S.L. (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.*, **30**, 2478–2483.
25. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
26. Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.
27. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
28. Rutherford,K., Parkhill,J., Crook,J., Horsnell,T., Rice,P., Rajandream,M.A. and Barrell,B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.
29. Marchler-Bauer,A., Panchenko,A.R., Shoemaker,B.A., Thiessen,P.A., Geer,L.Y. and Bryant,S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.*, **30**, 281–283.
30. Hueck,C.J. (1998) Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiol. Mol. Biol. Rev.*, **62**, 379–433.
31. Blanc-Potard,A.B., Solomon,F., Kayser,J. and Groisman,E.A. (1999) The SPI-3 pathogenicity island of *Salmonella enterica*. *J. Bacteriol.*, **181**, 998–1004.