Published online 5 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D543–D546***
doi:10.1093/nar/gkm928

*x*BASE2: a comprehensive resource for comparative bacterial genomics

Roy R. Chaudhuri, Nicholas J. Loman, Lori A.S. Snyder, Christopher M. Bailey, Dov J. Stekel and Mark J. Pallen*

Centre for Systems Biology, University of Birmingham, Birmingham B15 2TT, UK

Received September 18, 2007; Accepted October 10, 2007

ABSTRACT

***x*BASE is a genome database aimed at helping laboratory-based bacteriologists make best use of bacterial genome sequence data, with a particular emphasis on comparative genomics. The latest version, *x*BASE 2.0 (http://xbase.bham.ac.uk), now provides comprehensive coverage of all bacterial genomes and features an updated modularized backend and an improved user interface, which includes a taxonomy browser and a powerful full-text search facility.**

INTRODUCTION

*x*BASE grew out of *coli*BASE, *CampyDB* and several similar projects, which were restricted to selected taxonomic groups of bacteria (1,2). In its initial implementation, *x*BASE was simply an umbrella term applied to a set of distinct databases that relied on a similar schema. In creating *x*BASE2, we have developed a new integrated, comprehensive bacterial genomes database, with greatly increased coverage, many new features and an improved user interface and code base.

FEATURES

The implementation of a scalable *x*BASE schema has allowed us to incorporate all bacterial genome sequences available from the NCBI (currently >800), including finished genomes and partially assembled genomes from the WGS division of GenBank. At the time of writing, 516 complete genome sequences and 367 draft genome sequences are included in the database, with the total number expected to reach well over 1000 by the end of 2007. The current design is intended to scale to at least 10 000 genomes. The database is updated monthly via an automated process so that new genome sequence data are incorporated into *x*BASE soon after they are released into the public domain. A numerical scheme for naming updates allows researchers to record and cite the particular version of *x*BASE used in their studies.

The wealth of data within *x*BASE presents new challenges in locating a genome sequence of interest. The user interface provides a number of options to address this problem. The *x*BASE top page contains a complete list of the included genomes, which can be filtered to a more manageable list of 'popular' genomes, where popularity is determined by hit rate of each genome sequence, i.e. is set by the *x*BASE user community. A taxonomy browser is provided, which can be toggled between a full taxonomy view and a list of popular taxa. Although *x*BASE2 maintains the identities of the earlier databases (*coli*BASE, *Campy*DB, etc.) to facilitate the transition to the new version for our current user base, these are now merely flags of convenience for a small subset of all potential taxon-specific views. For example, the *coli*BASE tab takes the user to set of genomes defined by the link from the taxon Enterobacteriaceae. However, users now can select their genome sets of interest from any point on the NCBI bacterial taxonomic hierarchy, from phyla such as Actinobacteria to subspecies like *Zymomonas mobilis* subsp. *mobilis*. Genome sequences can also be located using the full-text search facility, with results ranked on the popularity of the organism.

The number of genomes within *x*BASE2 means that pre-calculating all pairwise genome sequence alignments is now no longer feasible. Instead, MUMmer (3) alignments are now performed on demand, within seconds, and are cached so that they are redisplayed instantly. Nucleotide alignments are now performed using the nucmer component of MUMmer. This has the advantage over mummer in that it allows comparisons between multiple-contig genomes, which is particularly useful when working with unfinished genomes. The alignment display code has been modified to allow multiple pairwise comparisons to be stacked, so that the same region can be viewed in many different genome sequences (Figure 1).

An XY plot allows the plotting of data such as nucleotide composition and principal axes derived from

*To whom correspondence should be addressed. Tel: +44 121 414 7163; Fax: +44 121 414 3454; Email: m.pallen@bham.ac.uk
Present address:
Roy R. Chaudhuri, Cambridge Veterinary School, University of Cambridge, Cambridge CB3 0ES, UK

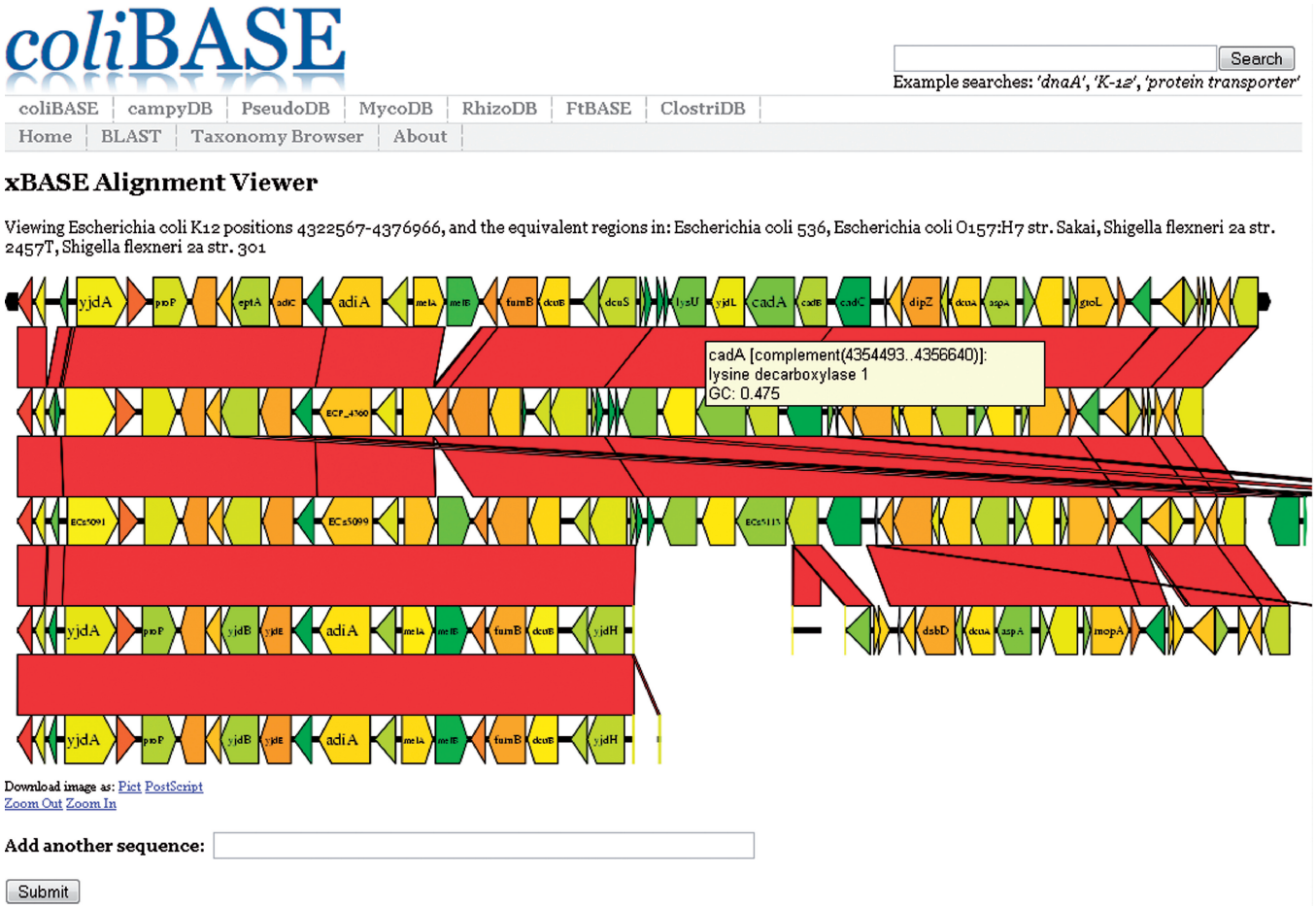

Figure 1. Stacked alignment of regions from multiple *Escherichia coli* and *Shigella* genomes. The region displayed shows the 'black hole' deletion of the *cadA* gene from the *Shigella* genomes. The product of this gene has been shown to attenuate *Shigella* virulence (5).

correspondence analysis of relative synonymous codon usage data (Figure 2). This facilitates identification of genes that show unusual patterns of composition or codon usage that might be indicative of horizontal gene transfer. The graphical display of genomic regions now includes genomic coordinates and at high zoom nucleotide sequences are displayed (Figure 3). Circular chromosomes and plasmids are now handled correctly, so the display can wrap around the origin of the sequence. An *x*BASE blog has been launched, with regular contributions from the development team, including status updates, discussions of new features and requests from users for additional functionality.

*X*BASE2 IMPLEMENTATION

The *x*BASE database backend has been completely rewritten to make use of a modified BioSQL schema, with additional tables to allow storage of genome alignment data and a flattened annotation table to aid rapid full-text searching. The code has also been modularized. The streamlined code brings benefits both to maintainers and end user, allowing new features to be developed rapidly and delivering a significant improvement in response time. Performance has also been improved by investment in new hardware. The full-text indexing facilities offered by the tsearch2 component of Postgresql have proven vital in creating a fast, accurate search facility. Researchers can now rapidly search for text strings such as gene names, phrases and keywords within both genome annotation and the documentation associated with each genome in the *Genomes online database* (GOLD) (4). *x*BASE source code and database is freely available on request.

FUTURE CHALLENGES

*x*BASE is funded until 2012. A key challenge will be maintaining usability in the face of thousands of new genome sequences, which are likely to be delivered within this timeframe by the application of next-generation sequencing to bacteriology. Within *x*BASE, multiple stacked pairwise comparisons will have to be supplemented with a view that collapses down identical regions and only shows representative displays of syntenic regions. Another challenge will be integration of experimental data (e.g. microarray and chromatin immunoprecipitation data) and networks (e.g. of protein–protein interactions) into the *x*BASE facility. Work has already begun on *x*BASE 3.0, which will feature a new 'Web 2.0' user

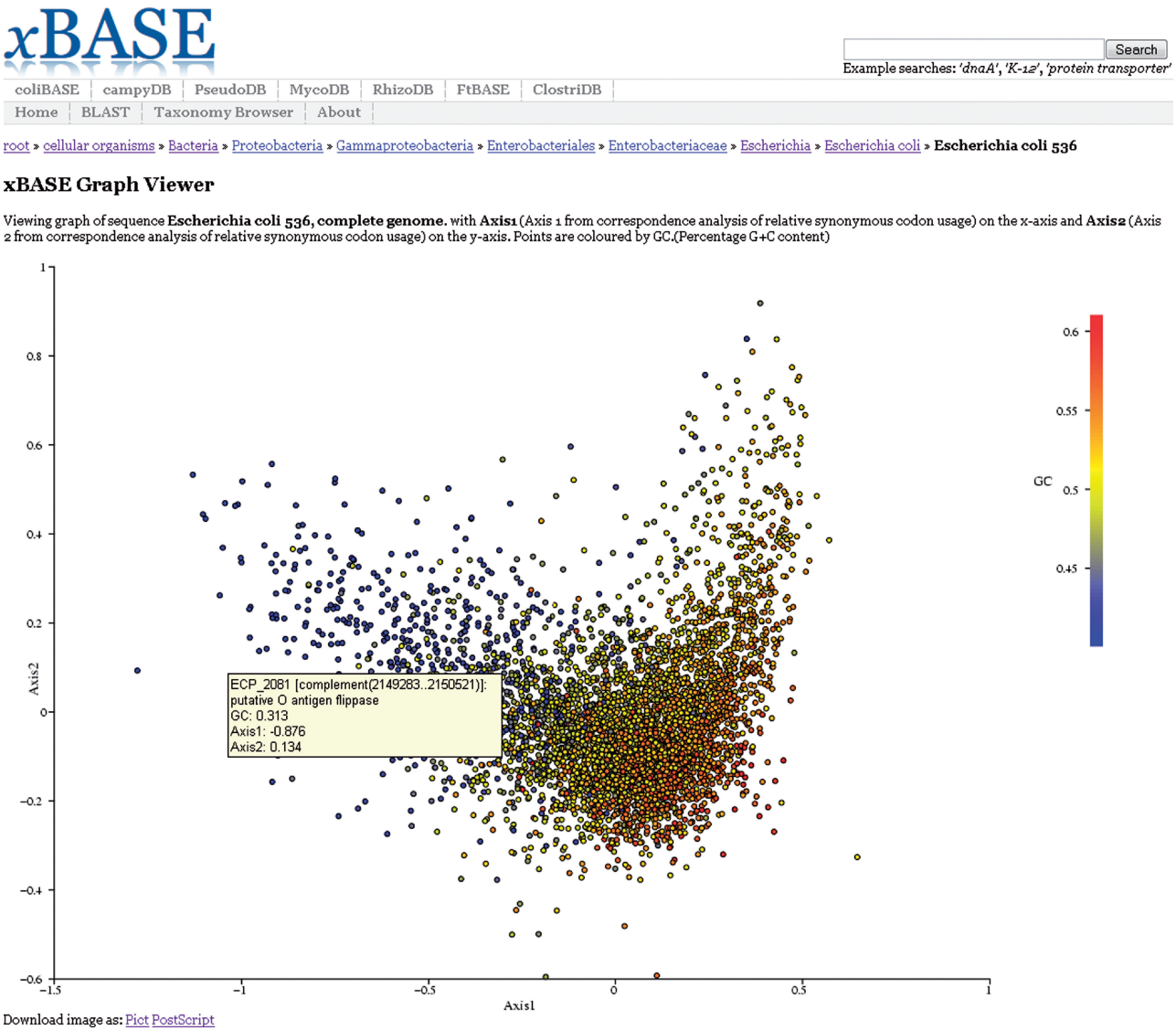

Figure 2. Plot of the two principal axes determined by correspondence analysis of relative synonymous codon usage of genes from *E. coli* strain 536. The plot shows a typical 'rabbit's head' (6), with the right 'ear' corresponding to highly expressed genes with optimal codon usage. The left 'ear' includes genes likely to be of foreign origin, most of which show a low GC content.

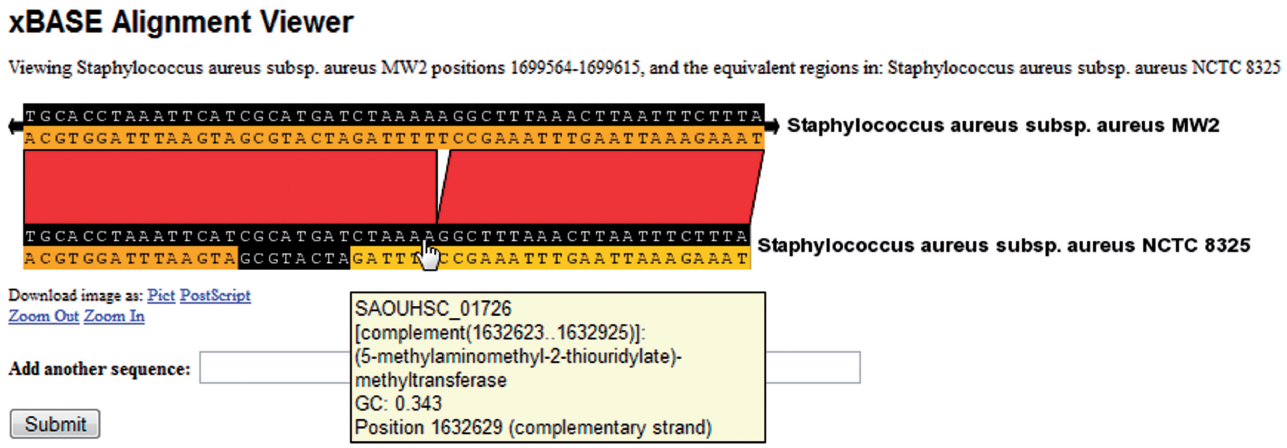

Figure 3. Alignment of two *Staphylococcus aureus* genomes, zoomed in to show the nucleotide sequence. The MW2 gene MW1571, which encodes a tRNA methyltransferase, is split into two in the NCTC 8325 genome. The display indicates that this is due to a single base deletion that occurs in a run of 5 T residues, suggesting that it may be an artefact caused by a sequencing error.

interface exploiting AJAX and advanced browser visual layout techniques, so that different data sets can be overlain and manipulated on the same page as a 'mashup' and full zooming and panning across a genome become possible (as in Google Maps). User tracking will enable us to provide a personalized interface, which will include a community annotation facility. We are thus confident that *x*BASE will continue to serve the bacteriology community well into the second decade of the new millennium. *x*BASE is available at http://xbase.bham.ac.uk.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by BBSRC. This work is supported by the BBSRC, grant number BBE0111791, awarded to MJP and DJS.

Conflict of interest statement. None declared.

REFERENCES

1. Chaudhuri,R.R. and Pallen,M.J. (2006) xBASE, a collection of online databases for bacterial comparative genomics. *Nucleic Acids Res.*, **34**, D335–D337.
2. Chaudhuri,R.R., Khan,A.M. and Pallen,M.J. (2004) coliBASE: an online database for Escherichia coli, Shigella and Salmonella comparative genomics. *Nucleic Acids Res.*, **32**, D296–D299.
3. Kurtz,S., Phillippy,A., Smoot,M., Shumway,M., Antonescu,C. and Salzberg,S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.
4. Liolios,K., Tavernarakis,N., Hugenholtz,P. and Kyrpides,N.C. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res.*, **34**, D332–D334.
5. Maurelli,A.T., Fernandez,R.E., Bloch,C.A., Rode,C.K. and Fasano,A. (1998) 'Black holes' and bacterial pathogenicity: a large genomic deletion that enhances the virulence of Shigella spp. and enteroinvasive Escherichia coli. *Proc. Natl Acad. Sci. USA*, **95**, 3943–3948.
6. Medigue,C., Rouxel,T., Vigier,P., Henaut,A. and Danchin,A. (1991) Evidence for horizontal gene transfer in Escherichia coli speciation. *J. Mol. Biol.*, **222**, 851–856.