BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 23 2008, pages 2803–2804
doi:10.1093/bioinformatics/btn524

Databases and ontologies

The Burkholderia Genome Database: facilitating flexible queries and comparative analyses

Geoffrey L. Winsor, Bhavjinder Khaira, Thea Van Rossum, Raymond Lo, Matthew D. Whiteside and Fiona S. L. Brinkman*

Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada V5A 1S6

Received on September 15, 2008; revised on October 3, 2008; accepted on October 4, 2008

Advance Access publication October 7, 2008

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: As the genome sequences of multiple strains of a given bacterial species are obtained, more generalized bacterial genome databases may be complemented by databases that are focused on providing more information geared for a distinct bacterial phylogenetic group and its associated research community. The Burkholderia Genome Database represents a model for such a database, providing a powerful, user-friendly search and comparative analysis interface that contains features not found in other genome databases. It contains continually updated, curated and tracked information about *Burkholderia cepacia* complex genome annotations, plus other *Burkholderia* species genomes for comparison, providing a high-quality resource for its targeted cystic fibrosis research community.

Availability: http://www.burkholderia.com. Source code: GNU GPL.

Contact: brinkman@sfu.ca.

1 INTRODUCTION

Cystic fibrosis (CF) patients are particularly susceptible to serious infection by *Burkholderia cepacia* complex (BCC) species. BCC are noted for their inherent resistance to antibiotics and the potential for transmission between patients, resulting in an increased risk of mortality (Mahenthiralingam and Vandamme, 2005). As with many pathogenic bacterial species, there is significant interest in comparing BCC genomes with each other, and with other species, to provide insight into species- or strain-specific features that may play an important role in virulence and antimicrobial resistance for these CF-relevant pathogens.

Several high-quality generalized genome databases already exist for searching *Burkholderia* sequence data and annotations, including the Integrated Microbial Genomes (IMG) system (Markowitz *et al*., 2006), the Comprehensive Microbial Resource (CMR) (Peterson *et al*., 2001), Pathema (http://pathema.jcvi.org), National Center for Biotechnology Information (NCBI) Microbial Genomes (Wheeler *et al*., 2007) and Microbes Online (Alm *et al*., 2005). However, these do not focus on the CF *Burkholderia* researchers' needs and do not integrate a flexible annotation search interface with the ability to track curated annotation updates, and methods for navigating species genomes via precise ortholog assignments. While these generalized databases are valuable resources, more phyla-specific databases are also being developed to address the needs of specific research communities, including providing additional resources not available from the more generalized bacterial databases. Phyla-specific databases include the Enteropathogen Resource Integration Center and other NIAID Bioinformatics Resource Centers, as well as the Pseudomonas Genome Database that we previously developed (Glasner *et al*., 2008; Greene *et al*., 2007; Winsor *et al*., 2005). Pathema contains a *Burkholderia*-specific component, but it is focused on *Burkholderia* species of biodefence interest that do not include the BCC of interest to the CF research community.

As the number of bacterial genome sequences increases, including those specific to a single species, there is a need to further improve such resources by providing better comparative genome annotation capabilities and orthologous gene links. We have now built upon the success of the Pseudomonas Genome Database, which offers user-friendly, yet powerful searching of the *P. aeruginosa* PAO1 genome annotation, coupled with continually updated genome annotations obtained from the *Pseudomonas* research community (Winsor *et al*., 2005). We report here the development of the Burkholderia Genome Database that, in addition to having the capability to track continually updated annotations, has a more flexible search interface and integrates multiple genomes with new comparative genome analysis capabilities. The database structure and interface facilitates the easy comparison of multiple genome annotations arising from annotation or sequence-based searches, and couples this with improved methods for ortholog identification and contextual visualization. Resources of specific interest to the BCC research community and CF researchers are also included.

2 IMPLEMENTATION

The front-end web application was developed in a SuSe Linux 10.1 environment using a combination of Java Server Pages 2.0, the Struts 1.2.1 framework, the Java 2 Platform (Standard, v 1.4.2), Perl 5, Apache Web Server 2.0 and Apache Tomcat 5.5. The back-end consists of a MySQL database server (v 14.7). The database can be easily set up to run under most operating systems.

In addition to information about *Burkholderia*-specific bioinformatics analysis and related resources, this database also contains the following features that are of more general utility for a microbial genome database.

*To whom correspondence should be addressed.

2.1 Complex, user-friendly annotation searches

The NCBI, CMR, Microbes Online and IMG databases mentioned above allow searching of genome annotations by entering a keyword and selecting from a list of fields including gene/protein name and references, with the option of filtering by one or more specified strains. The *Burkholderia* Genome Database search page combines these features with a Boolean search interface capable of further refining queries in a user-friendly format. These include options, such as only returning proteins localizing to a specific compartment, returning genes or proteins having an assigned confidence in function or returning genes having no detectable homology to human genes using specific criteria. To help facilitate downstream systems-level analyses, we enable searching of genes/proteins by annually updated functional characterizations, such as TIGRFAM, COG, PFAM or Gene Ontology. Search results link to individual pages ('gene cards') that contain detailed annotation data, a graphical view of the local genomic environment, plus links to a graphical ortholog view (see below) and GBrowse interface (Stein *et al*., 2002) for viewing multiple external annotations, sequence features and orthologs overlaid on a selected genome sequence.

2.2 Tracking changes to genome annotations

Annotation updates can be made based on researcher submissions, updates from other sequence databases or literature review and can be geared towards the interests of the research community. Updates from other databases occur annually, while researcher/literature-based updates occur more frequently. All updates are logged, but most notably a powerful log file search and browse interface is provided. As the history of annotation changes becomes larger and more complex, it is becoming increasingly important to facilitate complex searching of updates. Our updates log can be browsed and ordered by author/date/annotation info parameters, Boolean searched or downloaded. To the best of our knowledge, this level of user-friendly examination of tracked changes to annotations is unique to our database.

2.3 Comparison of gene annotations and context

To enable the comparison of multiple annotations from within or between species or strains, we developed a gene 'clipboard' or cart utility for storing genes selected from a search result. The annotations and sequences associated with genes on this clipboard may be compared. For example, if one searches 'argB', and then selects the genes of interest from the search result, one can view how genomic context changes for this gene in different species. Gene families can be compared in a similar manner. Unlike similar utilities found on the IMG and CMR websites, the orientation of one or more genes being viewed is automatically aligned, and can be selectively reversed for easier comparison of gene order. By clicking on the image map, one can navigate to gene cards for adjacent genes. The sequences for each gene/protein stored on the clipboard can be aligned or used as the basis for further searches.

With the aim of complementing text-based searches for similar genes with sequence-based approaches, an interface for more flexibly viewing BLAST-based search results was also developed. Unlike similar tools from other databases, our BLAST search results can be sorted by multiple fields or linked to the aligned region in GBrowse (for nucleotide sequence) or relevant gene cards (for proteins). Perhaps most importantly, one can add any of these search results to the clipboard or link to individual gene cards where one can view all known orthologs. So, a BLAST search with argB reveals additional genes, not found in an annotation search, which can be further compared using the clipboard and comparison views.

2.4 More comparative analysis: evaluating orthologs

Ortholog prediction is critical for comparative analyses and annotation transfer across species and is typically performed on a genome-scale using a reciprocal best BLAST hits (RBBH) approach. However, RBBH can incorrectly predict a paralog as an ortholog under certain scenarios (Fulton *et al*., 2006). To address this, we incorporated a high-throughput computational method named Ortholuge (Fulton *et al*., 2006) that evaluates predicted orthologs and identifies those that have undergone unusual divergence or were likely falsely predicted by RBBH. In an 'ortholog view' available from a gene card, predicted orthologs for each gene in the database appear along with adjacent genes in a stacked view that makes it easy to compare genomic context as well as the Ortholuge prediction. One can easily determine if a given ortholog is present or absent in a genome. Orthologs can also be viewed in GBrowse, facilitating comparative analysis and contextual navigation between species.

Collectively, this database provides a flexible, biologist-friendly interface for comparative analysis and curated, updated annotation of *Burkholderia* genomes that complements existing resources.

ACKNOWLEDGEMENTS

We thank all *Burkholderia* genome projects. FSLB is a Michael Smith Foundation for Health Research (MSFHR) Scholar and a CIHR New Investigator. MDW received a MSFHR scholarship.

Funding: We thank Cystic Fibrosis Foundation Therapeutics Inc.

Conflict of Interest: none declared.

REFERENCES

Alm,E.J. *et al*. (2005) The MicrobesOnline web site for comparative genomics. *Genome Res.*, **15**, 1015–1022.

Fulton,D.L. *et al*. (2006) Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics*, **7**, 270.

Glasner,J.D. *et al*. (2008) Enteropathogen Resource Integration Center (ERIC): bioinformatics support for research on biodefense-relevant enterobacteria. *Nucleic Acids Res.*, **36**, D519–D523.

Greene,J.M. *et al*. (2007) National Institute of Allergy and Infectious Diseases bioinformatics resource centers: new assets for pathogen informatics. *Infect. Immun.*, **75**, 3212–3219.

Mahenthiralingam,E. and Vandamme,P. (2005) Taxonomy and pathogenesis of the Burkholderia cepacia complex. *Chron. Respir. Dis.*, **2**, 209–217.

Markowitz,V.M. *et al*. (2006) The integrated microbial genomes (IMG) system. *Nucleic Acids Res.*, **34**, D344–D348.

Peterson,J.D. *et al*. (2001) The comprehensive microbial resource. *Nucleic Acids Res.*, **29**, 123–125.

Stein,L.D. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.

Wheeler,D.L. *et al*. (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.

Winsor,G.L. *et al*. (2005) Pseudomonas aeruginosa Genome Database and PseudoCAP: facilitating community-based, continually updated, genome annotation. *Nucleic Acids Res.*, **33**, D338–D343.