D338–D343 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki047

Pseudomonas aeruginosa Genome Database and PseudoCAP: facilitating community-based, continually updated, genome annotation

Geoffrey L. Winsor, Raymond Lo, Shannan J. Ho Sui, Korine S.E. Ung, Shaoshan Huang, Dean Cheng, Wai-Kay Ho Ching, Robert E. W. Hancock[1] and Fiona S. L. Brinkman*

Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, B.C., Canada, V5A 1S6 and [1]Department of Microbiology and Immunology, University of British Columbia, Vancouver, B.C., Canada, V6T 1Z3

Received August 12, 2004; Accepted September 28, 2004

ABSTRACT

Using the *Pseudomonas aeruginosa* Genome Project as a test case, we have developed a database and submission system to facilitate a community-based approach to continually updated genome annotation (http://www.pseudomonas.com). Researchers submit proposed annotation updates through one of three web-based form options which are then subjected to review, and if accepted, entered into both the database and log file of updates with author acknowledgement. In addition, a coordinator continually reviews literature for suitable updates, as we have found such reviews to be the most efficient. Both the annotations database and updates-log database have Boolean search capability with the ability to sort results and download all data or search results as tab-delimited files. To complement this peer-reviewed genome annotation, we also provide a linked GBrowse view which displays alternate annotations. Additional tools and analyses are also integrated, including PseudoCyc, and knockout mutant information. We propose that this database system, with its focus on facilitating flexible queries of the data and providing access to both peer-reviewed annotations as well as alternate annotation information, may be a suitable model for other genome projects wishing to use a continually updated, community-based annotation approach. The source code is freely available under GNU General Public Licence.

BACKGROUND

In 1997, we initiated a community-aided approach for genome annotation that functioned solely through the Internet (1,2) with the goal of critical and conservative genome annotation at reduced cost. This approach was applied to the *Pseudomonas aeruginosa* Genome Project during initial genome annotation efforts, and involved enlisting volunteer researchers from the *Pseudomonas* research community to submit annotations of genes and gene families with which they were familiar. We used this annotation approach because: (i) The *P.aeruginosa* PAO1 genome was the largest bacterial genome sequenced to date; (ii) *P.aeruginosa* is the third most cited bacterium in Medline and has a strong research community studying it. The project, termed the *P.aeruginosa* Community Annotation Project or PseudoCAP, was met with enthusiasm—47 researchers initially expressed interest in the project, and in the end 61 researchers submitted a total of 1741 annotations, a sizeable volunteer contribution for a genome containing 5570 genes. After publication of the complete *P.aeruginosa* PAO1 genome sequence in the year 2000 (1), we have now expanded the methodology of our approach and our core database to facilitate the development of a continually updated genome annotation database for this organism.

A number of community-based approaches to genome annotation have been previously used for other genome projects (3–7) although few of them were exclusively Internet-based. One of the best examples of a successful approach to maintaining a continually updated genome annotation database through an Internet portal has been WormBase (8) for the *Caenorhabditis elegans* genome project (The *C.elegans* Sequencing Consortium 1998). Our approach has been similar to that of WormBase, perhaps with the most notable difference, other than our bacterial focus, being that we have put more emphasis on the development of a user-friendly log file of annotation updates, which is more amenable to searching and includes submission of relevant author information and a detailed description of the updates. Such user-friendly log file search flexibility will become increasingly important as update log files increase in size and complexity. Other excellent community-reviewed annotation systems have been developed that

*To whom correspondence should be addressed. Tel: +1 604 291 5646; Fax: +1 604 291 5583; Email: brinkman@sfu.ca

are suitable for bacterial genomic data including the PeerGAD system (9) and ASAP (10). However, they also appear to lack a Boolean searching facility for the log file or have notable functionalities missing, such as the ability to accept changes in DNA sequence (9) or perform sequence-based searches (10). In addition, alternate annotations made by other research groups/centres are not easily viewed in these systems. We provide a combination of both peer-reviewed centralized, and unreviewed decentralized, data resources that utilize a combination of previously reported approaches, including the 'open annotation' approach (11) and the Distributed Annotation System (12).

DATABASE SCHEMA

An overview of the entire database is shown schematically in Figure 1. The *Pseudomonas* Genome Database is based on three main tables containing (i) the original genome annotation published in the year 2000 (an important reference dataset), (ii) the continually updated annotation and (iii) the log of annotation updates. Visitors entering the site can choose the original or updated annotation and be forwarded to the respective pages in order to browse or search for specific annotations as well as download tab-delimited files of information from these tables. Alternatively, one can log on to the site as a

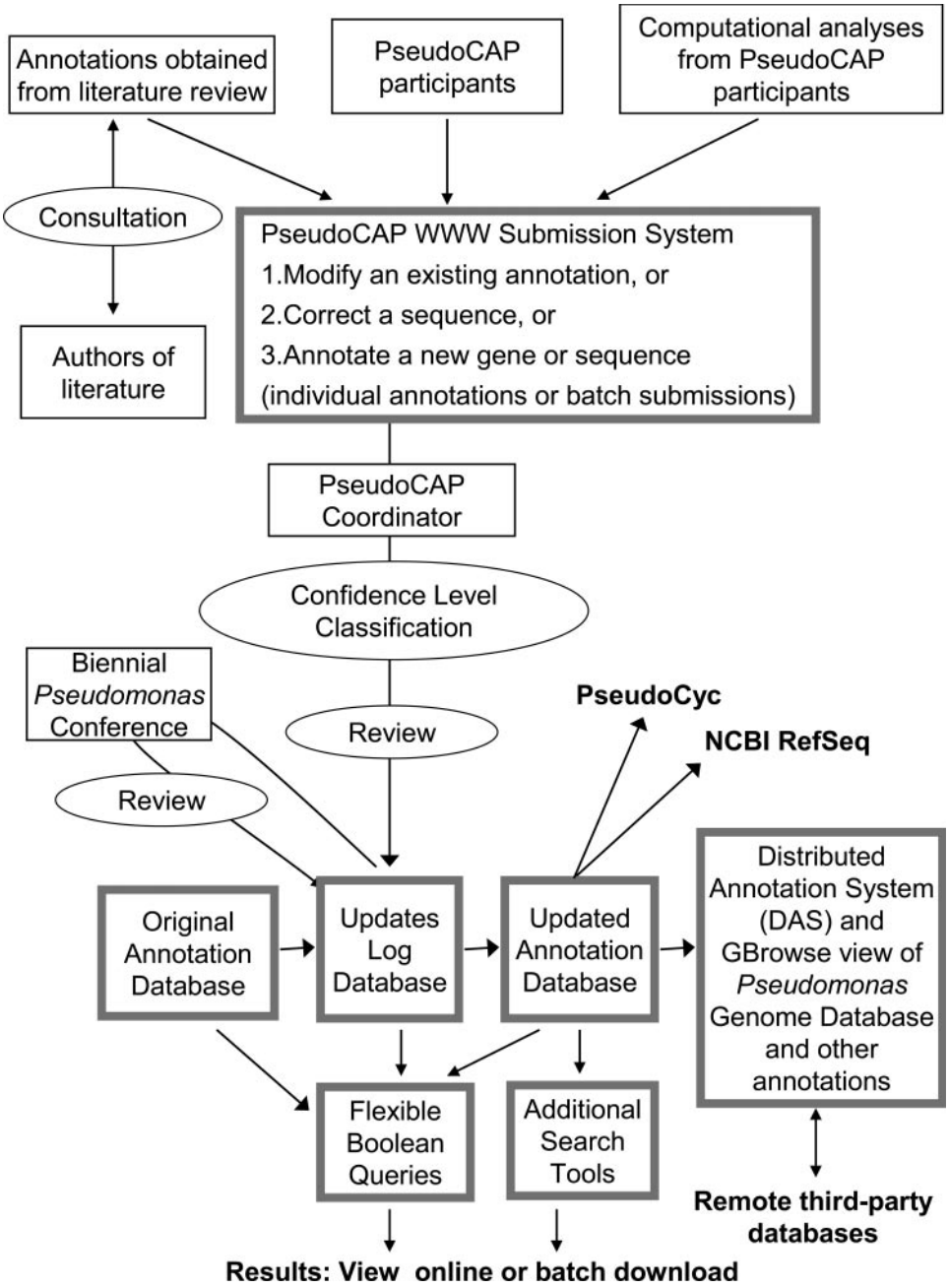

Figure 1. Schematic overview of the *Pseudomonas* Genome Database and PseudoCAP annotation update process. Boxes with thick lines/borders denote database and application components. Ovals reflect processes in the annotation update approach. Elements with thin lines/borders reflect human intervention.

registered PseudoCAP participant and proceed to make an annotation submission through a series of forms.

Records in the original and updated annotation tables are uniquely identified by a locus identifier consisting of 'PA' (for *P.aeruginosa*) followed by a four-digit number representing the order of open reading frames (ORFs) around the chromosome starting at the origin of replication. In consultation with the *Pseudomonas* research community, we have adapted the convention of using a decimal system to account for newly described ORFs (i.e. PA1000.1 would be between PA1000 and PA1001; PA1000.01 would be between PA1000 and PA1000.1). These identifiers can link to external databases such as TIGR's comprehensive microbial resource (CMR) (13), which represents an alternate annotation view, and the NCBI RefSeq (14,15) that we submit updates to. Records in these tables contain information on the primary name associated with the ORF and its product as well as any alternate names that have been used. Furthermore, functional classification, genomic context, structure, predicted localization of the product as well as reactions and predicted pathways the product is involved in, PubMed references and DNA and protein sequences are stored. Fields are searchable using a Boolean search interface with the flexibility to sort the data and then either view the data directly or download the search results in the tab-delimited format. ORFs can also be browsed by their order around the genome or by functional classification of their product. With regard to nucleotide and protein sequences, a BLAST search can be performed against genomic DNA and protein sequences using BLASTALL from the NCBI (16). In addition, subsequences of the PAO1 genome (DNA or translated sequences) can be downloaded by specifying the base pair coordinates of the DNA sequence. Finally, the amino acid sequence of proteins can be obtained by specifying the PA number of the gene encoding it.

FACILITATING CONTINUALLY UPDATED, COMMUNITY-BASED ANNOTATION

The cornerstone of our database, which ensures the quality of its existing annotations, is the ability for PseudoCAP participants to submit new annotation data for review. After entering some personal data related to their affiliation and location, registered participants receive a username and password in order to login and submit modifications to existing annotations, annotations for new genes and sequence corrections as well as view and download his/her submission history. If a participant does not wish to register, the participant may also complete a simple form for one-time direct submissions. Once submitted, the user will receive an email confirming that the annotation submission has been received and is under initial review by the PseudoCAP coordinator. The coordinator examines the submission and then responds to requests for any additional information/clarification, if required. At this point a protein name confidence level is assigned or changed, using our previously developed classification system (1) (see http://www.pseudomonas.com). In addition, a confidence level is assigned/changed for protein subcellular localization, as part of our expanded confidence system. Any suitable GO annotation is also evaluated and associated with the annotation change.

Contrary to a previous review of annotation update systems [(9); this included a review of the PseudoCAP system], our system does facilitate browsing of annotation update histories and the curator/coordinator is not entirely responsible for accepting a new annotation. The entry is reviewed by at least one additional reviewer from the research community and a collection of all annotations made over a two-year period are subjected to additional review at the biennial International *Pseudomonas* conference. We feel that this latter review step is important to provide the community with an efficient mechanism to review annotation updates collectively and examine and discuss any systematic annotation issues. Management of the review stages by a coordinator is also important to ensure consistency in annotation updates and to ensure that additional reviewers chosen from the research community are appropriate for a given annotation update case.

In addition to PseudoCAP participant submissions, the *Pseudomonas* research literature is also reviewed weekly using PubCrawler (17) and the papers that report new gene names, gene functions or other information that may impact on the genome annotation are noted. The corresponding author of the paper is contacted with a proposal for an annotation submission that is based on the paper's work. If agreed to, the submission is directly accepted (because it has already been subjected to peer-review during publication) and the log file notes that this was a submission based on literature review and also provides information on the accepting author and journal citations.

An important consideration for researchers is the access to a history of changes to the genome annotation since its initial release. The *Pseudomonas* Genome Database contains a Boolean-searchable log of all updates that have been made to the genome annotation. Fields include names of the participants who have made the submission along with structured details and the dates of the submission and PA numbers. The log of updates can be browsed and ordered by any of the above parameters or searched using the Boolean search interface that was developed specifically for the log file, or with the search interface for primary annotation information found elsewhere in the database. As with the primary annotation search, results can be viewed and sorted online or downloaded in tab-delimited format. While other web-based genome annotation databases provide access to downloadable update histories, to the best of our knowledge the ability to perform concise Boolean searches and sorting of results is a unique feature of our database. Such functionality will become increasingly important for genome databases, as the log files for updates become larger and more complex.

ALTERNATE ANNOTATIONS AND ADDITIONAL ANALYSES UTILIZING GBROWSE

With the increase in annotations available to microbiologists via the Internet, there is a necessity to visualize genomic annotation information from multiple sources in a single viewer. We also feel that it is important to encourage alternate scientific views by allowing researchers to view any alternate annotations relative to our database's primary, peer-reviewed, annotation information. To facilitate this, we have incorporated a platform-independent web application called

GBrowse developed by Stein *et al*. (18) of the Generic Model Organism System Database Project (GMOD) (http://www.gmod.org). Using checkboxes, the user can select annotation information to view including alternate gene names, protein names, motifs/structures as well as metabolic pathway data and knockout data, and perform a search based on criteria they specify. GBrowse then fetches the region of genome specified by the user's search criteria and presents the specified landmarks to the user in a detailed view containing one or more horizontal tracks representing individual sequence features for that area. The user is free to zoom in and out according to the level of magnification/resolution desired. Landmarks on each track usually contain a link to detailed information on additional websites.

We have incorporated a wide range of locally curated annotations into tracks in our *Pseudomonas* GBrowse including all ORFs, with links to http://www.pseudomonas.com, non-coding RNA genes, intergenic regions (with links to their sequences) and all proteins linked to their respective records at the NCBI. In addition, a variety of third-party annotations are accessible via GBrowse, including the Protein Extraction, Description and ANalysis Tool (PEDANT) online database for protein structure analyses of the *P.aeruginosa* genome that are based on similarity to sequences in the Protein Data Bank (19). A track for the Prokaryotic Database of Gene Regulation and Regulatory Networks (PRODORIC NET) contains information on transcriptional regulators, operons and associated binding sites (20). For pathway analysis, a Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways track (21) and a link to the local PseudoCyc annotation (22), is made available. The PseudoCyc annotation, which is now locally curated, is also subjected to the same annotation updates as are made for the primary, reviewed annotation. A track is also available for computationally predicted rho-independent transcription terminators as described by the Institute for Genomic Research (23) and reported by our group (see section below on additional new tools). As part of a move to incorporate more functional genomics annotation information, tracks are also available for resources such as a transposon mutant library created by the University of Washington Genome Center (UWGC) (24). This library consists of 30 100 mutants derived from the IS50L element of transposon Tn5 which forms either a *phoA* or *lacZ* translational fusion when inserted in the correct orientation into a gene. Clicking on either of the symbols for the *phoA* and *lacZ* fusions in this track links to a more detailed information on the specific transposon including links to details about the library at the UWGC website. We are also in the process of expanding links to genome annotations for other *Pseudomonas* species or *P.aeruginosa* strains. As a first step, links to *Pseudomonas putida* KT2440 and *Pseudomonas syringae* pathovar *tomato* DC3000 putative orthologs (based on reciprocal best-BLAST hits) are made available as GBrowse tracks, with the *P.aeruginosa* PAO1 genome sequence serving as the reference genome. Eventually, we will expand this to include more accurate ortholog identification based on phylogenetic analysis.

GBrowse also enables researchers to upload and privately view their own annotation data as a track in GBrowse by implementing the Distributed Annotation System (DAS) (25). Through DAS, researchers can upload annotation data to an Internet accessible server where others can view these data within a GBrowse track by entering the URL of a reference server. In essence, *Pseudomonas* GBrowse and DAS allow researchers to easily view alternate annotations while promoting open discussion leading to better annotations for the *Pseudomonas* research community.

INTEGRATION WITH OTHER DATA VIEWS–PSEUDOCYC: A PATHWAY/GENOME DATABASE AS AN EXAMPLE

We have recently taken over the curation of PseudoCyc (http://pseudocyc.pseudomonas.com:1555), based on a database schema developed by Karp *et al*. (26) using their Pathway Tools software. PseudoCyc (22) contains information on *P.aeruginosa* PAO1 genes and proteins along with 821 enzymatic reactions and 141 biochemical pathways involving 738 enzymes and 651 compounds. However, one problem with this database is that it uses the gene name as its primary key—a name that can change as annotations are updated. When we took over curation of this database, we implemented the PA number as the new unique identifier for each gene record in PseudoCyc, thus allowing easier integration with records in the *Pseudomonas* Genome Database and other external annotation sources. The updated unique identifiers have facilitated updates to gene and protein names that have changed since the original annotation in the year 2000, while retaining annotation updates that were made during the initial curation of the PseudoCyc pathways (22). Such resources are fully integrated, such that gene records in the *Pseudomonas* genome database can serve as an entry point for viewing their respective entry in PseudoCyc and the context of its product within the various biochemical reactions and pathways described for *P.aeruginosa* PAO1. Another functionality common to PseudoCyc and related databases is the ability to upload gene expression data to PseudoCyc and overlay it on a Metabolic Overview diagram to view results in the context of the various pathways.

ADDITIONAL NEW TOOLS: DNA MOTIF SEARCH AS AN EXAMPLE

In addition to the wide range of locally managed annotations and third-party annotations accessible via the *Pseudomonas* Genome Database, we have created a user-friendly search tool that can be used to find user-specified DNA motifs within the PAO1 genomic DNA sequence. This search tool is powered by a Perl-script capable of accepting as input an IUPAC-formatted variable length DNA sequence and converting it into a regular expression used to search the genome sequence. Upon search completion, an online report or downloadable tab-delimited file is produced containing information on all regions the motif is found in. We used this tool to discover a previously undescribed rho-independent terminator subset containing a common sequence string (Sequence: AAAGC{3,4}SN{5,30}SGGGCTTT; occurrences not previously reported by others are viewable under the Brinkman terminator track under GBrowse). The *P.aeruginosa* PAO1 genome had the highest total number of occurrences of this terminator subset compared to all complete genomes at the time this paper was prepared, with related *Pseudomonas* species genomes containing slightly higher than average

occurrences as a proportion of their genome, perhaps reflecting an evolutionary relationship in the evolution of this terminator sequence. The discovery of this surprisingly terminator-specific and species-specific sequence motif is just one example of how integration of such elementary tools has led to new insights through *Pseudomonas* genome analysis.

CONCLUSIONS

The number of researchers participating in the *Pseudomonas* Genome Database continually updated annotation/PseudoCAP has increased from 61 in the year 2000 to 105 as of July 2004. These individuals, from 11 countries worldwide, have contributed 1019 annotation updates, not including the submissions made prior to the genome sequence publication. These submissions were made by individuals without overt solicitation. However, annotation updates made through review of recent peer-reviewed publications were also useful and are increasingly forming a larger proportion of annotation updates being made. If genome annotations are to remain current, there will be a significant need in the future to improve text-mining approaches for the identification of annotation updates derived from recent research literature.

The database is capable of accommodating changes to annotation data resulting from changes in the genomic DNA sequence, as evident by a recent G insertion at base-pair position 2669175. Such changes reinforce the need for having a clear primary key for all genes annotated in a genome, since the nucleotide coordinates can change for all genes downstream of nucleotide corrections that involve an insertion or deletion. We have utilized the PA number, with its additional decimal system as described above, as a primary key, and discourage the use of the gene name as a primary identifier, due to its potential to change in some cases, as knowledge of a gene's function increases. We encourage all genome projects to have a clear primary key that is not based on any relationship (functional or sequence similarity) with genes in other species, to avoid confusion as orthologous relationships between genes are refined as new sequence data from more diverse organisms are obtained.

The community-driven approach used to annotate the *P.aeruginosa* PAO1 genome represents a successful approach for utilizing the expertise of microbial researchers to aid in microbial genome annotation and analysis. The approach comprises a combination of centralized and decentralized methods by placing annotation data acquired from the community in a centralized, curated database, which is subject to detailed review, with links to additional annotation and alternate interpretations of the data to facilitate communication, differing research views and novel ideas. We believe that both are critical. Continually updated annotation approaches that do not have a reference annotation, which is subjected to review are susceptible to increased confusion, when researchers sort through and deliberate as to which annotation information to trust or report in the context of their laboratory or computational analysis. Continually updated annotation approaches that only provide a single primary reference annotation run the risk of stifling alternate views regarding how gene or cell functions/processes should be described or analyzed. Of course, static annotations that are not subjected to updating may become obsolete and the workload involved in high–quality, re-curation of whole genomes can render occasional re-annotation of a whole genome unfeasible, in contrast to the incremental approach we perform as we keep abreast of the research literature and PseudoCAP submissions.

PseudoCAP has facilitated the collaboration of *Pseudomonas* researchers, capitalized on their experience, and stimulated interaction and collaboration while furthering our understanding of this genome sequence. Through the centralized peer review process, we have been able to attain a high-quality genome annotation for this organism that takes into account levels of confidence for each annotation. This has been achieved with minimal cost by taking advantage of the Internet as an effective means for collecting, distributing and analyzing information, as opposed to a more costly jamboree approach (6) and the payment incentives offered during some projects (3,4). Furthermore, acknowledging researchers for their contributions provides an incentive to participate in such projects in much the same way that publication promotes dissemination of research knowledge. This database and/or our community annotation approach has now been used by other genome project groups, such as the Rhodococcus Genome Project (http://www.rhodococcus.ca/) and *Methanosarcina acetivorans* genome project (SarcinaCAP) (27), either in its existing form or in a slightly modified format. Through its availability as an open source package, with additional utilization of other open source packages such as GBrowse that are already available, this database, and the associated continually updated annotation approach will potentially act as base for additional genome projects in the future, including some already targeted involving the *Pseudomonas* genus. For further information, or to contact us to obtain our source code under a GNU public license, see http://www.pseudomonas.com.

ACKNOWLEDGEMENTS

The authors thank Peter Karp and Randy Gobbel for assistance with implementing the necessary changes to PseudoCyc. Primary funding for this project was provided by the Cystic Fibrosis Foundation. Ancillary funding provided by Genome Prairie, Genome BC, Inimex Pharmaceuticals and the CCFF. R.E.W.H. has a Canada Research Chair and F.S.L.B. is a Michael Smith Foundation for Health Research Scholar.

REFERENCES

1. Stover,C.K., Pham,X.Q., Erwin,A.L., Mizoguchi,S.D., Warrener,P., Hickey,M.J., Brinkman,F.S., Hufnagle,W.O., Kowalik,D.J., Lagrou,M., Garber,R.L., Goltry,L., Tolentino,E., Westbrock-Wadman,S., Yuan,Y., Brody,L.L., Coulter,S.N., Folger,K.R., Kas,A., Larbig,K., Lim,R., Smith,K., Spencer,D., Wong,G.K., Wu,Z., Paulsen,I.T., Reizer,J., Saier,M.H., Hancock,R.E., Lory,S. and Olson,M.V. (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature*, **406**, 959–964.
2. Brinkman,F.S., Hancock,R.E. and Stover,C.K. (2000) Sequencing solution: use volunteer annotators organized via Internet. *Nature*, **406**, 933.
3. Williams,N. (1995) Closing in on the complete yeast genome sequence. *Science*, **268**, 1560–1561.
4. Williams,N. (1996) Yeast genome sequence ferments new research. *Science*, **272**, 481.
5. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 563–567.

6. Pennisi,E. (2000) Ideas fly at gene-finding jamboree. *Science*, **287**, 2182–2184.
7. Buell,C.R., Joardar,V., Lindeberg,M., Selengut,J., Paulsen,I.T., Gwinn,M.L., Dodson,R.J., Deboy,R.T., Durkin,A.S., Kolonay,J.F. *et al*. (2003) The complete genome sequence of the Arabidopsis and tomato pathogen *Pseudomonas syringae* pv. tomato DC3000. *Proc. Natl Acad. Sci. USA*, **100**, 10181–10186.
8. Stein,L., Sternberg,P., Durbin,R., Thierry-Mieg,J. and Spieth,J. (2001) WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res*., **29**, 82–86.
9. D'Ascenzo,M.D., Collmer,A. and Martin,G.B. (2004) PeerGAD: a peer-review-based and community-centric web application for viewing and annotating prokaryotic genome sequences. *Nucleic Acids Res*., **32**, 3124–3135.
10. Glasneri,J.D., Liss,P., Plunkett,G.III, Darling,A., Prasad,T., Rusch,M., Byrnes,A., Gilson,M., Biehl,B., Blattneri,F.R.*et al*. (2003) ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res*., **31**, 147–151.
11. Hubbard,T. and Birney,E. (2000) Open annotation offers a democratic solution to genome sequencing. *Nature*, **403**, 825.
12. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.
13. Peterson,J.D., Umayam,L.A., Dickinson,T., Hickey,E.K. and White,O. (2001) The Comprehensive Microbial Resource. *Nucleic Acids Res*., **29**, 123–125.
14. Maglott,D.R., Katz,K.S., Sicotte,H. and Pruitt,K.D. (2000) NCBI's LocusLink and RefSeq. *Nucleic Acids Res*., **28**, 126–128.
15. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res*., **29**, 137–140.
16. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipmani,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
17. Hokamp,K. and Wolfe,K. (1999) What's new in the library? What's new in GenBank? Let PubCrawler tell you. *Trends Genet*., **11**, 471–472.
18. Stein,L.D., Mungalli,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. and Lewis,S. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*., **10**, 1599–1610.
19. Frishman,D., Mokrejs,M., Kosykh,D., Kastenmuller,G., Kolesovi,G., Zubrzycki,I., Gruber,C., Geieri,B., Kaps,A., Albermann,K. *et al*. (2003) The PEDANT genome database. *Nucleic Acids Res*., **31**, 207–211.
20. Munch,R., Hiller,K., Barg,H., Heldt,D., Linz,S., Wingender,E. and Jahn,D. (2003) PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res*., **31**, 266–269.
21. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
22. Romero,P. and Karp,P. (2003) PseudoCyc, a Pathway-Genome Database for *Pseudomonas aeruginosa*. *J. Mol. Microbiol. Biotechnol*., **5**, 230–239.
23. Ermolaeva,M.D., Khalak,H.G., White,O., Smith,H.O. and Salzberg,S.L. (2000) Prediction of transcription terminators in bacterial genomes. *J. Mol. Biol*., **301**, 27–33.
24. Jacobs,M.A., Alwood,A., Thaipisuttikul,I., Spencer,D., Haugen,E., Ernst,S., Will,O., Kaul,R., Raymond,C., Levy,R., Chun-Rong,L., Guenthner,D., Bovee,D., Olson,M.V. and Manoil,C. (2003) Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proc. Natl Acad. Sci. USA*, **100**, 14339–14344.
25. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R., Stein,L. (2001) The Distributed annotation system. *BMC Bioinformatics*, **2**, 7.
26. Karp,P.K., Paley,S. and Romero,P. (2002) The pathway tools software. *Bioinformatics*, **18** (Suppl. 1), S225–S232.
27. Galagan,J.E., Nusbaum,C., Roy,A., Endrizzi,M.G., Macdonald,P., FitzHugh,W., Calvo,S., Engels,R., Smirnov,Si. and Atnoor,D. (2002) The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. *Genome Res*., **12**, 532–542.