Published online 12 November 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D141–D145***
doi:10.1093/nar/gkn879

The Ribosomal Database Project: improved alignments and new tools for rRNA analysis

J. R. Cole[1,*], Q. Wang[1], E. Cardenas[1], J. Fish[2], B. Chai[1], R. J. Farris[1], A. S. Kulam-Syed-Mohideen[1], D. M. McGarrell[1], T. Marsh[1,2], G. M. Garrity[2] and J. M. Tiedje[1,2]

[1]Center for Microbial Ecology and [2]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA

Received September 15, 2008; Revised October 16, 2008; Accepted October 17, 2008

ABSTRACT

The Ribosomal Database Project (RDP) provides researchers with quality-controlled bacterial and archaeal small subunit rRNA alignments and analysis tools. An improved alignment strategy uses the Infernal secondary structure aware aligner to provide a more consistent higher quality alignment and faster processing of user sequences. Substantial new analysis features include a new Pyrosequencing Pipeline that provides tools to support analysis of ultra high-throughput rRNA sequencing data. This pipeline offers a collection of tools that automate the data processing and simplify the computationally intensive analysis of large sequencing libraries. In addition, a new Taxomatic visualization tool allows rapid visualization of taxonomic inconsistencies and suggests corrections, and a new class Assignment Generator provides instructors with a lesson plan and individualized teaching materials. Details about RDP data and analytical functions can be found at http://rdp.cme.msu.edu/.

DESCRIPTION

As of September 2008 (release 10.3), the Ribosomal Database Project (RDP) maintained 33 082 archaeal and 643 916 bacterial small subunit rRNA sequences. Of these, 142 511 came from cultured organisms while 534 487 were sequences obtained from environmental samples. A total of 5534 sequences are from species type-strains; these sequences help to link taxonomy and phylogeny. As described in our previous update (1), all sequences are tested for anomalies using the Pintail program (2). Slightly fewer than 10% (64 290) of the sequences are marked as being of suspect quality. The RDP sequence collection is updated monthly from the International Nucleotide Sequence Database Collaboration (INSDC: DDBJ, EMBL and GenBank).

In May 2008, RDP introduced the RDP 10 series of releases with completely new bacterial and archaeal alignments based on a major improvement to the RDP alignment strategy (Table 1). These new alignments are created using the Infernal secondary structure based aligner (3), the same aligner used to provide alignments in the Rfam database of untranslated RNA molecules (4). Both Infernal and RNAcad (5) (the aligner used in the RDP 9 series of releases) are stochastic context-free grammar based and provide a high-quality secondary structure aware alignment. The Infernal aligner provides several significant advantages over RNAcad. Infernal is about 25 times faster and can align approximately 44 near full length 16S rRNA sequences per CPU-minute on a 2.66 GHz Xeon processor. It provides a much more intuitive handling of sequencing errors. For example, when a base is missing on one side of a helix, RNAcad disrupts the alignment on the other side of the helix (it does not allow half base pairs), while Infernal will allow the half base pair with a penalty. Infernal also correctly aligns some known problem sequences that have been reported with the RNAcad-based RDP 9 alignment of a small number of short partial sequences (6).

The Infernal aligner, like other probabilistic model-based aligners, is trained on a set of representative sequences. We trained the aligner on a small, hand-curated set of high-quality full-length rRNA sequences derived mainly from genome sequencing projects (508 for the bacterial model and 79 for the archaeal model). In many cases, the annotated rRNA gene start and stop positions, in the genome records, were found to be incorrect and were adjusted using a combination of the RNAmmer web server and hand adjustment [(7); this article also notes the problem in rRNA gene annotation]. Secondary structure information was based on the work

*To whom correspondence should be addressed. Tel: +1 517 353 3842; Fax: +1 517 353 8957; Email: colej@msu.edu

Table 1. Characteristics of the new RDP 10 alignment models

Alignment model	Number of model positions[a]	Number of base pairs modeled[b]	Number of sequences[c]	Percentage bases modeled[d] (%)
Bacterial	1416	431	643 916	92.1
Archaeal	1443	438	33 082	97.3

[a]Number of positions in alignment modeled as columns of homologues.
[b]Number of base pairs in alignment model.
[c]Number of sequences in RDP release 10.3.
[d]Percentage of bases aligned as model positions in RDP release 10.3.

of Gutell and colleagues (8). The relatively small training sets needed for model-based aligners is an advantage over nearest neighbor aligners, which require a much larger seed alignment. It is much easier to maintain reliable hand-adjustment of homology information in the smaller training set.

The second major improvement in RDP 10 over RDP 9 is the former provides an up-to-date, aligned and annotated archaeal data set along with the bacterial data set. As there are significant differences in the conserved secondary structures of bacterial and archaeal 16S rRNA, each data set is provided in a separate alignment to maximize the number of comparable positions. Outgroup sequences are provided in each alignment (*Escherichia coli* for the archaeal and *Methanocaldococcus jannaschii* for the bacterial alignments) for use in phylogenetic reconstruction.

In addition, this release incorporates a new phylogenetically consistent higher order bacterial taxonomy based on published taxonomic opinions, including opinions on environmental groups. Much of the taxonomy is taken from that proposed by Garrity *et al.* (9) with some major modifications from the recent reevaluations to the *Firmicutes* and *Cyanobacteria* proposed by Bergey's Trust (10,11), along with additional published informal taxonomies for the *Acidobacteria* (12), *Verrucomicrobia* (13), OP11 (14) and other less-well-studied areas of microbial diversity. This taxonomy has been used to train the RDP Classifier. The RDP Classifier, which is also available as a standalone program (15), can be easily trained on other alternative taxonomies. In addition, all public RDP sequences can be browsed in the taxonomy offered by GenBank (16).

These improvements have been extended to the *my*RDP user-account system described in our last update (1). All *my*RDP user data have been updated to match the new RDP 10 alignments and taxonomy. As of September 2008, the *my*RDP system had 2652 active users with a total of 1 655 828 private sequences. The Research Buddy feature has proved quite popular, with over 200 researchers participating in the data sharing.

New Pyrosequencing Pipeline

Sequencing SSU rRNA genes from environmental samples is a standard method for determining bacterial community composition. New technologies such as multiplexed pyrosequencing of mixed rRNA amplicons now allow in-depth analysis of bacterial rRNA composition to be carried out rapidly and inexpensively (see e.g. 17).

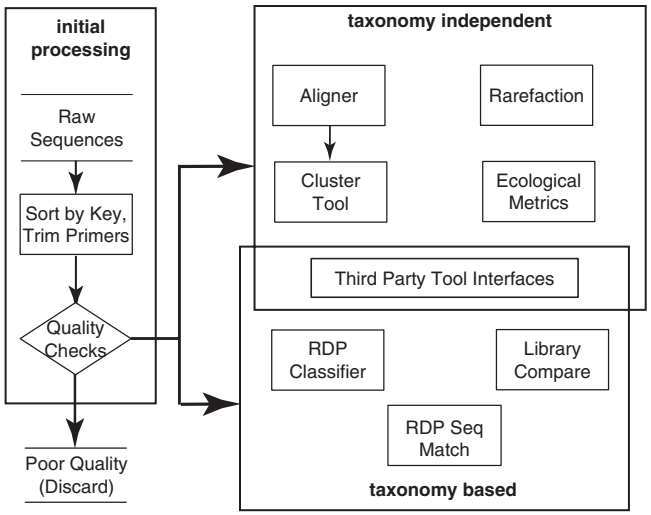

Figure 1. Tools available in the RDP for processing pyrosequencing data.

With these new sequencing methodologies, computational analysis of the large numbers of sequences produced has become a major challenge. The new RDP Pyrosequencing Pipeline offers a collection of tools that automate the data processing and simplify the computationally intensive analysis of large sequencing libraries (Figure 1).

In the initial processing steps, raw sequence reads from multiple samples are sorted using sample-specific key (tag) sequences. The mapping between the tag sequence and sample name is designated in a tag file. Custom designed tags are allowed. Reads with a tag that do not exactly match any tag sequences are put into a separate folder. Four quality filters can be applied in the initial processing step: the maximum edit distance allowed for the forward primer(s), the maximum edit distance allowed for the reverse primer(s), the number of ambiguity codons (N's) and the minimum sequence length after removing primer sequences. The pipeline contains both taxonomy-based and taxonomy-independent methods for community analysis, and each type of method has advantages. Taxonomy-based methods are often faster and provide classification information about the organisms in a sample. However, these methods are not always able to correctly identify novel organisms (organisms not in the preexisting taxonomic framework). For taxonomy-independent analysis, standard species-based comparison methods for both α and β ecological diversity measures are easily applied to clustered data, although divergence-based approaches may yield better resolution (18,19).

For taxonomy-based analysis, the RDP Classifier provides fast and reliable classification of short sequence reads (15,20). The RDP Library Compare program can be used to detect differentially represented taxa between samples. And the RDP Sequence Match tool can be used

to find the closest sequences in the RDP database for each sequence in a sample.

For taxonomy-independent alignment, the trimmed reads are aligned using the fast Infernal aligner mentioned above. Reads are then clustered into Operational Taxonomic Units at multiple pairwise distances using custom code implementing the complete-linkage clustering algorithm. Specialized tools provide common ecological metrics including: Chao1, Shannon Index and rarefaction. In addition, the processed data can be downloaded in formats suitable for common ecological and statistical packages including Spade (http://chao.stat.nthu.edu.tw), EstimateS (http://viceroy.eeb.uconn.edu/estimates) and R (http://www.R-project.org). Other options are available to cluster data from multiple samples, to combine alignments, to extract specific sequences from the dataset, to select representative sequences from clustered sequences and to produce comparative metrics among samples.

Initial tests of the Pyrosequencing Pipeline were carried out with Roche GS FLX data covering a 207-base region of the 16S rRNA gene including the V4 variable region and flanked by conserved targets for amplification primers. Further information about the protocol, including primer and key sequences, can be found at the RDP web site. The Pipeline has been extended to work with GS FLX and GS 20 data covering other regions of the 16S gene and should have no trouble processing the longer pyrosequencing reads produced by the Roche GS FLX Titanium system.

All of the analysis steps are carried out on a small cluster of servers. Most of the steps are relatively fast. A typical pyrosequencing run of 350 000 reads requires about 10 CPU-minutes for the initial processing steps with a 2.66 GHz processor. The RDP Classifier can assign approximately 2900 trimmed 207-base reads per CPU-minute. With an alignment model limited to the 207-base region, reads are aligned by Infernal at a rate of about 2200 reads per CPU-minute. With a complete 16S rRNA model, the short reads are aligned at a rate of around 500 reads per CPU-minute. The RDP complete-linkage Cluster Tool can cluster a maximum of slightly more than 150 000 unique sequences. In practice, this has been sufficient to cluster over 200 000 reads combined from multiple soil samples and a much higher number of reads from less diverse environments. Speed of the Cluster Tool is proportional to the square of the number of sequences. A sample with 10 000 unique sequences requires about 6 min for clustering. Currently, clustering and alignment results are returned to the user through email; all other analyses are returned online, however, an online interface similar to the RDP's *my*RDP space is under development.

New Taxomatic visualization tool

This tool displays a color heatmap representation of a symmetric distance matrix between large sets of sequences chosen by the user (Figure 2). Close and distant relationships are displayed in contrasting colors. This display makes it easy to visualize errors in an underlying taxonomy. Taxa that are phylogenetically incoherent or are misplaced in a hierarchy stand out visually in the representation, as do individual misplaced sequences. Sequences can be selected from the RDP database, from *my*RDP or an aligned fasta file, or a precalculated distance matrix file uploaded by the user. The heatmap displays sequences arranged in a predefined taxonomic order. For RDP and *my*RDP sequences, either the RDP taxonomy or an uploaded user-defined taxonomy can be visualized. Researchers can use pan and zoom features similar to those on Google Maps to examine specific regions of the heatmap. Bounding boxes for each taxon are displayed on mouse-over and the corresponding portions of the taxonomic hierarchy are highlighted. At higher zoom levels, sequence and organism specific information are rendered upon mouse-over in the tooltips display. To accommodate visually impaired users, the colors in the heatmap display can be selected from a drop-down menu. Heatmap images, the underlying distance matrix and the taxonomy are also available for download.

This tool also implements a version of the 'Self-Organizing Self-Correcting Classifier' algorithm (SOSCC) of Garrity and Lilburn (22). SOSCC is an experimental distance matrix optimization algorithm that can be used to automatically detect and reassign misplaced taxa and sequences in taxonomies with

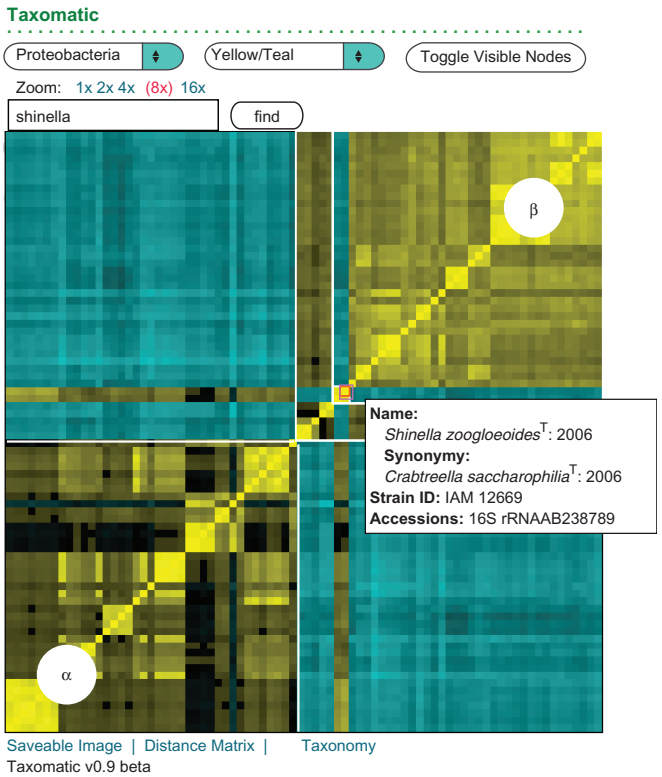

Figure 2. Taxomatic sample screenshot demonstrating a taxonomic anomaly. Yellow indicates more close sequences and teal more distant, with intermediate distances black. The screen is zoomed-in to show a portion of proteobacterial type-strain sequences. α: *Alphaproteobacteria*. β: *Betaproteobacteria*. The mouse points to the genus *Shinella*, originally placed in the *Alphaproteobacteria* (21), but incorrectly moved to the *Betaproteobacteria* in the Taxonomic Outline of the Bacteria and Archaea (9).

minor inconsistencies. Briefly, the SOSCC sorts the matrix to place nearest neighbors next to each other. New taxon boundaries are calculated by expanding each taxon from its centroid sequence to maximize a scoring function based on the numbers of taxon members and nonmembers included in the new boundary. When starting Taxomatic, users are given the option of preprocessing a matrix with the SOSCC, running the algorithm once, or running it in bootstrapping mode (100 iterations). With bootstrapping, SOSCC will generate alignments using a random selection of alignment columns with replacement and perform the SOSCC optimization on each alignment, keeping track of the updated taxonomy for every sequence in each iteration. Reclassifications are accepted that occur in user-predefined minimum number of bootstraps. With bootstrapping, the user is prompted to enter an email address where a link to the results will be sent. The smoothed heatmap and new taxonomic boundaries are displayed at this link. A text file of all machine-generated taxonomic changes is available for download along with the new taxonomy and heatmap images.

New RDP class Assignment Generator

This new educational tool for teaching the basics of rRNA analysis provides instructors a lesson plan along with individualized lesson material. It generates a set of unique sequences for each student that can be easily distributed to a classroom. It provides a set of easy-to-follow instructions for students and an answer key for instructors to help evaluate student performance. The instructor is presented a simple form that asks about the number of students, the level of difficulty (the number of sequences to be assigned for each student) and other information about the class. The tool then produces a set of sequences for each student modified from existing sequences. The sequences are randomly selected from a set of well-known genera from both archaea and bacteria that are programmatically 'evolved' by modifying a small number of bases. The changes preserve rRNA secondary structure and avoid highly conserved positions. Since each student receives a customized set of sequences, there is less chance of students sharing results. Students are asked to analyze the sequences, and the instructor is provided with a key containing the classification for each of the student sequences.

Analysis tools

A new Genome Browser feature added to the RDP Hierarchy Browser, provides a subset of the RDP database that derives from sequenced bacterial and archaeal genomes. Sequences are arranged by organism in the taxonomic hierarchy. Pages for each genome project list the genome size and number of 16S rRNA genes. For each project, links to further resources are provided. These links are automatically updated using the Genomic Rosetta Stone web service from the Genomic Standards Consortium (23). Also, sequences from completed genome projects can be found by searching with the GenBank Project ID from the Hierarchy Browser. Additional tools are described in our previous update (1).

RDP user surveys

Over the last year, the RDP implemented a new user survey system to help obtain user input on directions for the RDP. Each survey asks a single question and is displayed for approximately two weeks as an overlay when users access the RDP web site. Users may either answer the question or decline to answer before continuing to the RDP web site. A browser cookie is used to keep users from seeing the same survey question twice. The results of the RDP surveys are available on the RDP web site Resources page. On average, 810 researchers responded to each RDP user survey.

In addition to extensive Help Files, the RDP hosts a collection of short Video Tutorials demonstrating some of the more complex analysis tasks. These tutorials average 3 min in length. The videos capture the screen as the tasks are performed, while the narrator explains the tasks and the choices available to the researcher. All tutorials are available in Flash, Quicktime and Windows media formats.

RDP ACCESS AND CONTACT

The RDP data and analysis services can be found at http://rdp.cme.msu.edu/. The RDP's mission includes user support. Support questions can be emailed to rdpstaff@msu.edu. Telephone support is available (+1 517 432 4998). The RDP staff may also be contacted via fax (+1 517 353 8957 Attn: RDP) or regular mail.

ACKNOWLEDGEMENTS

We thank Woo Jun Sul for help with development of the RDP Pyrosequencing Pipeline. We thank several individuals for their past contributions: Robin Gutell (and his colleagues), Chuck Parker, Paul Saxman, Bonnie Maidak, Tim Lilburn, Niels Larsen, Tom Macke, Michael J. McCaughey, Ross Overbeek, Sakti Pramanik, Scott Dawson, Mitch L. Sogin, Gary Olsen and Carl Woese.

FUNDING

The Office of Science (BER), U.S. Department of Energy, (DE-FG02-99ER62848, DE-FG02-04ER63933); National Science Foundation (DBI-0328255); National Institute of Environmental Health Sciences Superfund Basic Research Program (P42ES04911). Funding for open access charge: Office of Science (BER), U.S. Department of Energy (DE-FG02-99ER62848).

Conflict of interest statement. None declared.

REFERENCES

1. Cole,J.R., Chai,B., Farris,R.J., Wang,Q., Kulam-Syed-Mohideen,A.S., McGarrell,D.M., Bandela,A.M., Cardenas,E., Garrity,G.M., Tiedje,J.M. *et al*. (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res*., **35**, D169–D172. (doi: 10.1093/nar/gkl889)

2. Ashelford,K.E., Chuzhanova,N.A., Fry,J.C., Jones,A.J. and Weightman,A.J. (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.*, **71**, 7724–7736.
3. Nawrocki,E.P. and Eddy,S.R. (2007) Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput. Biol.*, **3**, e56.
4. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
5. Brown,M.P.S. (2000) Small subunit ribosomal RNA modeling using stochastic context-free grammar. In Bourne,P., Gribskov,M., Altman,R., Jensen,N., Hope,D., Lengauer,T., Mitchell,J., Scheeff,E., Smith,C., Strande,S., *et al.* (eds.) *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB 2000)*, San Diego, CA, pp. 57–66.
6. Ashelford,K.E., Chuzhanova,N.A., Fry,J.C., Jones,A.J. and Weightman,A.J. (2006) New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. *Appl. Environ. Microbiol.*, **72**, 5734–5741.
7. Lagesen,K., Hallin,P., Rødland,E.A., Staerfeldt,H.H., Rognes,T. and Ussery,D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, **35**, 3100–3108.
8. Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D'Souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L.V., Muller,K.M. *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2.
9. Garrity,G.M., Lilburn,T.G., Cole,J.R., Harrison,S.H., Euzeby,J. and Tindall,B.J. (2007) The Taxonomic Outline of Bacteria and Archaea. TOBA Release 7.7. Michigan State University Board of Trustees, MI, USA.
10. Ludwig,W., Schleifer,K.-H. and Whitman,W.B. (2008) *Revised Road Map to the Phylum Firmicutes. Bergey's Manual of Systematic Bacteriology*, Vol. 3, Springer-Verlag, New York.
11. Wilmotte,A. and Herdman,M. (2001) Phylogenetic relationships among the *Cyanobacteria* based on 16S rRNA sequences. In Garrity,G.M., Boone,D.R. and Castenholz,R.W. (eds) *Bergey's Manual of Systematic Bacteriology*, Vol. 1, 2nd edn. Springer-Verlag, New York, pp. 487–493.
12. Barns,S.M., Cain,E.C., Sommerville,L. and Kuske,C.R. (2007) *Acidobacteria* phylum sequences in uranium-contaminated subsurface sediments greatly expand the known diversity within the phylum. *Appl. Environ. Microbiol.*, **73**, 3113–3116.
13. Sangwan,P., Chen,X., Hugenholtz,P. and Janssen,P.H. (2004) *Chthoniobacter flavus* gen. nov., sp. nov., the first pure-culture representative of subdivision two, *Spartobacteria* classis nov., of the phylum *Verrucomicrobia*. *Appl. Environ. Microbiol.*, **70**, 5875–5881.
14. Harris,J.K., Kelley,S.T. and Pace,N.R. (2004) New perspective on uncultured bacterial phylogenetic division OP11. *Appl. Environ. Microbiol.*, **70**, 845–849.
15. Wang,Q., Garrity,G.M., Tiedje,J.M. and Cole,J.R. (2007) Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, **73**, 5261–5267.
16. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2007) Genbank. *Nucleic Acids Res.*, **36**, 25–30.
17. Sogin,M.L., Morrison,H.G., Huber,J.A., Welch,D.M., Huse,S.M., Neal,P.R., Arrieta,J.M. and Herndl,G.J. (2006) Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc. Natl Acad. Sci. USA*, **103**, 12115–12120.
18. Schloss,P.D. (2008) Evaluating different approaches that test whether microbial communities have the same structure. *ISME J.*, **2**, 265–275.
19. Lozupone,C.A. and Knight,R. (2008) Species divergence and the measurement of microbial diversity. *FEMS Microbiol. Rev.*, **32**(4), 557–578.
20. Liu,Z., Desantis,T.Z., Andersen,G.L. and Knight,R. (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res.* [Epub Aug 22, doi:10.1093/nar/gkn491]
21. An,D.S., Im,W.T., Yang,H.C. and Lee,S.T. (2006) *Shinella granuli* gen. nov., sp. nov., and proposal of the reclassification of *Zoogloea ramigera* ATCC 19623 as *Shinella zoogloeoides* sp. nov. *Int. J. Syst. Evol. Microbiol.*, **56**, 443–448.
22. Garrity,G.M. and Lilburn,T.G. (2005) Self-organizing and self-correcting classifications of biological data. *Bioinformatics*, **21**, 2309–2314.
23. Van Brabant,B., Gray,T., Verslyppe,B., Kyrpides,N., Dietrich,K., Glöckner,F.O., Cole,J., Farris,R., Schriml,L.M., De Vos,P. *et al.* (2008) Laying the foundation for a Genomic Rosetta Stone: creating information hubs through the use of consensus identifiers. *OMICS*, **12**(2), 123–127. (doi:10.1089/omi.2008.0020)