Wanchai et al. BMC Bioinformatics 2017, 18(Suppl 14):483
DOI 10.1186/s12859-017-1900-9

BMC Bioinformatics

DATABASE Open Access

dBBQs: dataBase of Bacterial Quality scores

Visanu Wanchai[†], Preecha Patumcharoenpol[†], Intawat Nookaew[†] and David Ussery[*]

From The 14th Annual MCBIOS Conference
Little Rock, AR, USA. 23-25 March 2017

Abstract

Background: It is well-known that genome sequencing technologies are becoming significantly cheaper and faster. As a result of this, the exponential growth in sequencing data in public databases allows us to explore ever growing large collections of genome sequences. However, it is less known that the majority of available sequenced genome sequences in public databases are not complete, drafts of varying qualities. We have calculated quality scores for around 100,000 bacterial genomes from all major genome repositories and put them in a fast and easy-to-use database.

Results: Prokaryotic genomic data from all sources were collected and combined to make a non-redundant set of bacterial genomes. The genome quality score for each was calculated by four different measurements: assembly quality, number of rRNA and tRNA genes, and the occurrence of conserved functional domains. The dataBase of Bacterial Quality scores (dBBQs) was designed to store and retrieve quality scores. It offers fast searching and download features which the result can be used for further analysis. In addition, the search results are shown in interactive JavaScript chart framework using DC.js. The analysis of quality scores across major public genome databases find that around 68% of the genomes are of acceptable quality for many uses.

Conclusions: dBBQs (available at http://arc-gem.uams.edu/dbbqs) provides genome quality scores for all available prokaryotic genome sequences with a user-friendly Web-interface. These scores can be used as cut-offs to get a high-quality set of genomes for testing bioinformatics tools or improving the analysis. Moreover, all data of the four measurements that were combined to make the quality score for each genome, which can potentially be used for further analysis. dBBQs will be updated regularly and is freely use for non-commercial purpose.

Keywords: Genome quality score, Database, Bacteria

Background

It is well known that the current state-of-art of sequencing technologies makes genome sequencing significantly cheaper and quicker. Especially, the third generation sequencing which based on single-molecule sequencing technologies, have gained popularity because of ability of generating the long read [1]. Also, the exponential growth in sequencing data in public databases allow us to explore through large collections of genome sequences [2]. However, it is less known that many genomes in public databases are left as draft genome sequences. A huge number of draft genomes usually comes from difficulty of finishing process of genome sequences generated by second generation sequencing machine. Therefore, many genome projects on major genome repositories were left unfinished [3].

The estimation of errors in draft genome by Denton et al. [4] in 2014 indicated that, by comparing the same genomes with different level of completeness, nearly 40% of all gene families were inferred to have incorrect number of genes in draft genomes. Also, the possible reason of having over predicted genes in unfinished genomes is the fragmentation of genes in many contigs. Hence, these non-finished genome sequences may vary in qualities causing the inconsistent analysis.

Here, we collected both draft and complete genomes for around 100,000 bacterial genomes from major genome repositories: GenBank and GenBank Sequence Read Archive provided by the National Center for Biotechnology Information [5], the Broad Institute [6], the U.S.

* Correspondence: dwussery@uams.edu
[†]Equal contributors
Arkansas Center for Genomic Epidemiology & Medicine and The Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR 72205, USA

Department of Energy Systems Biology Knowledgebase, and the Pathosystems Resource Integration Center [7]. Then all genomes were annotated and assessed for the quality scores with the same method. We designed and implemented database to stores all genomes and their analysis. The website was constructed by the concept of interactive designed which allows users to interact directly with data and get feedback instantly.

Construction and content

Data sources

We retrieved bacterial genomic data from four different sources: GenBank, GenBank - SRA, Broad, Kbase, and PATRIC. These databases are major public genome repositories containing all types of genome completeness ranging from complete gnomes to contigs. The detail of retrieving genome sequences for each database can be described as follows. GenBank genomes were retrieved from the FTP site provided by NCBI [8]. Then each of whole genome sequence in Fasta nucleotide format was download and stored in a directory. The Fasta files of GenBank - SRA, which have already assembled as previously reported by the work of Larsen et al. [9], were downloaded from NCBI SRA FTP site [10] and stored the same way that previously described with GenBank genomes. The genome data from Broad Institute were retrieved from the Broad Olive website [6]. The bundle files of Broad project were extracted and kept only Fasta files. Kbase genomes were obtained through its API which allowed us to easily select any level of completeness and only Fasta format for genomes. Fasta files from PATRIC were searched and downloaded by using its FTP site [11].

Genome quality scores

The genome quality score for each genome was calculated using the method proposed by Land et al. [12]. In order to standardize all genomes in the analysis, all genomes in Fasta format from different sources were predicted for the protein-coding genes using Prodigal [13]. Next, all four individual scores—Sequence Quality score, rRNA score, tRNA score, and Essential gene score—were calculated using RNAmmer [14], tRNAscan-SE [15], and HMMER3 [16] with Pfam-A [17] respectively (Fig. 1). Basically, each of score is the measurement of the completeness of genome sequence: assembly quality, number of rRNA and tRNA genes, and the presence of conserved functional domains. Then all four scores were averaged to estimate the genome quality score.

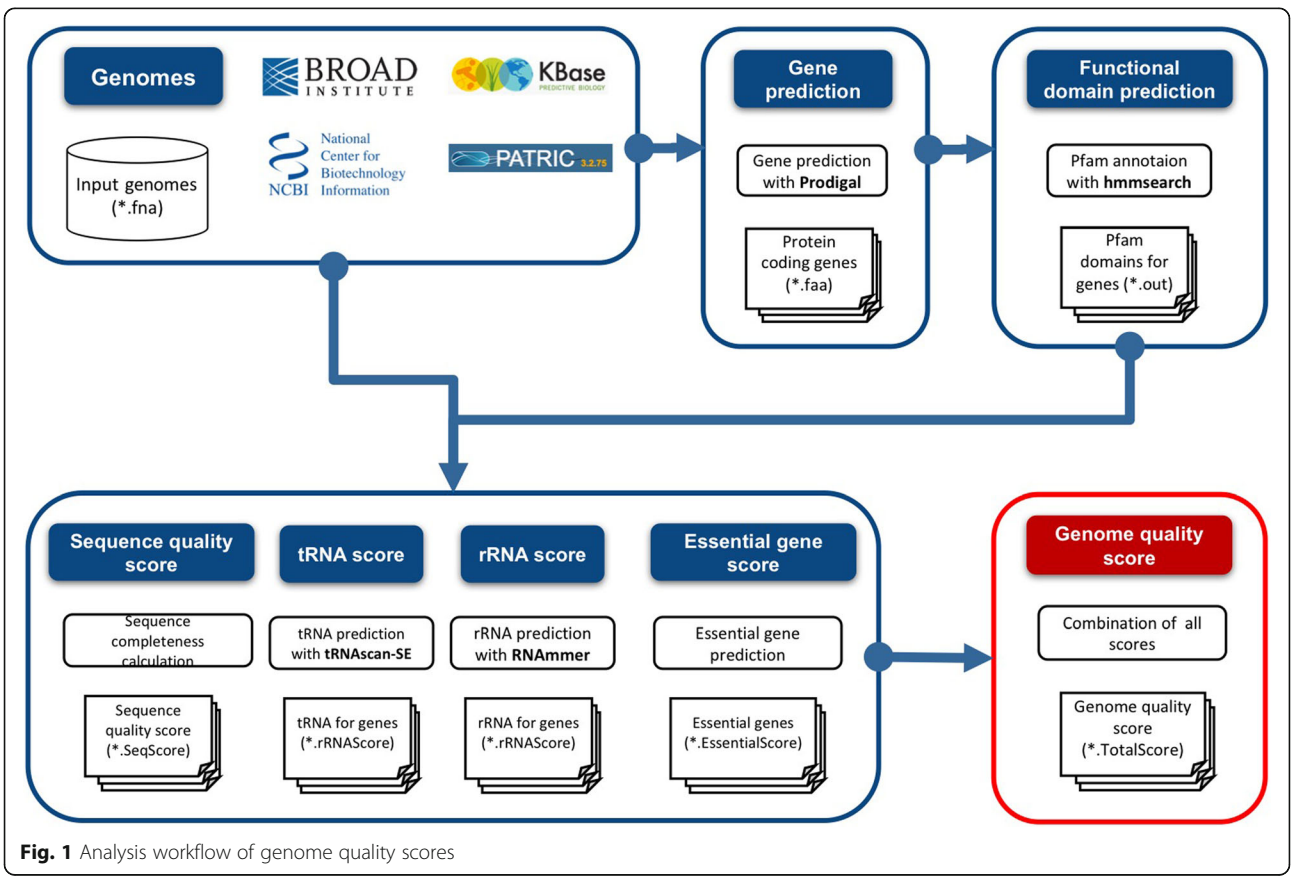

Fig. 1 Analysis workflow of genome quality scores

Database schema and implementation

The database of dBBQs was developed as a relational database using SQlite3. The Apache HTTP 2.4.6 web server was then used to host the website. The API that executed dot commands on SQLite3 and supplied data to webpage was implemented on Python Flask. HTML, JQuery and Bootstrap CSS were used to build the front-end of the website. DC.js and Crossfilter were used to make the dynamic chart features on the website.

The entity relationship diagram (ERD) was designed to store 3 tables representing different kind of information obtained from the analysis: GenomeDetail, QualityScore, Taxonomy. As shown in Fig. 2, each entry in each of the tables demonstrates a field of information contained in the tables. The GenomeDetail table contained name and identifier of all genomes along with basic details such as genome size, number of contigs, GC content. A QualityScore table stored the genome quality scores and other four quality scores. A well curated taxonomy data related to bacterial genomes in GenBank were downloaded from a Namesforlife website. This taxonomy data was reduced to a non-redundant set and then assigned to each genome to make a Taxonomy table.

Utility and discussion

The total number of bacterial genomes stored in the database of dBBQs is 96,167 genomes. These genomes were collected from four different genome repositories: 67,980 genomes from GenBank; 11,768 genomes from GenBank - SRA; 2477 genomes from Broad; 11,944 genomes from Kbase; 1998 genomes from PATRIC. According to the "safe-to-use" genome quality score at 0.8 or better, we found that 65,689 out of 96,167 (~68%) genomes passed this criterion. Table 1 shows the summary of number of bacterial genomes, genome quality scores, and four scores for different sources. As expected, the average of genome quality scores of four different sources met the safety criterion except genomes from GenBank - SRA that have the average score at 0.69. The low average genome quality score was usually because there were too many contiguous pieces for each genome which significantly brought the average sequence quality score down too low and affected the genome quality score.

Comparing the annotations between dBBQs and the original source databases remains a difficult task due to the lack of provided complete annotations for all genomes. However, we still can compare the number of predicted proteins as it is the most complete annotation in the database of bacterial genomes. For purposes of assessing quality of protein prediction, we downloaded the metadata which contains numbers of predicted proteins of all genomes from NCBI [https://www.ncbi.nlm.nih.gov/genome/browse]. As can be seen in Fig. 3, we compared the distribution of predicted proteins between dBBQs and GenBank in four different levels of genome status (Complete Genome, Chromosome, Contig, Scaffold). dBBQs showed very similar at locating proteins in most of genomes in GenBank with a few exceptions even in scaffolds which contain lots of contigs and gaps.

User interface

Interactive chart section

Figure 4 shows the front page of dBBQs which composes of two types of chart (6 bar charts of 'Genome Quality score', 'Sequence Quality Score', 'rRNA Score', 'tRNA Score', 'Essential Gene Score', and 'Taxonomy: Phylum'; 1 donut chart of 'Genome Repositories') and 1 table of genome information. User can select the data category or range of scores from all charts as filters to display on the website. Once any of charts is selected, the other

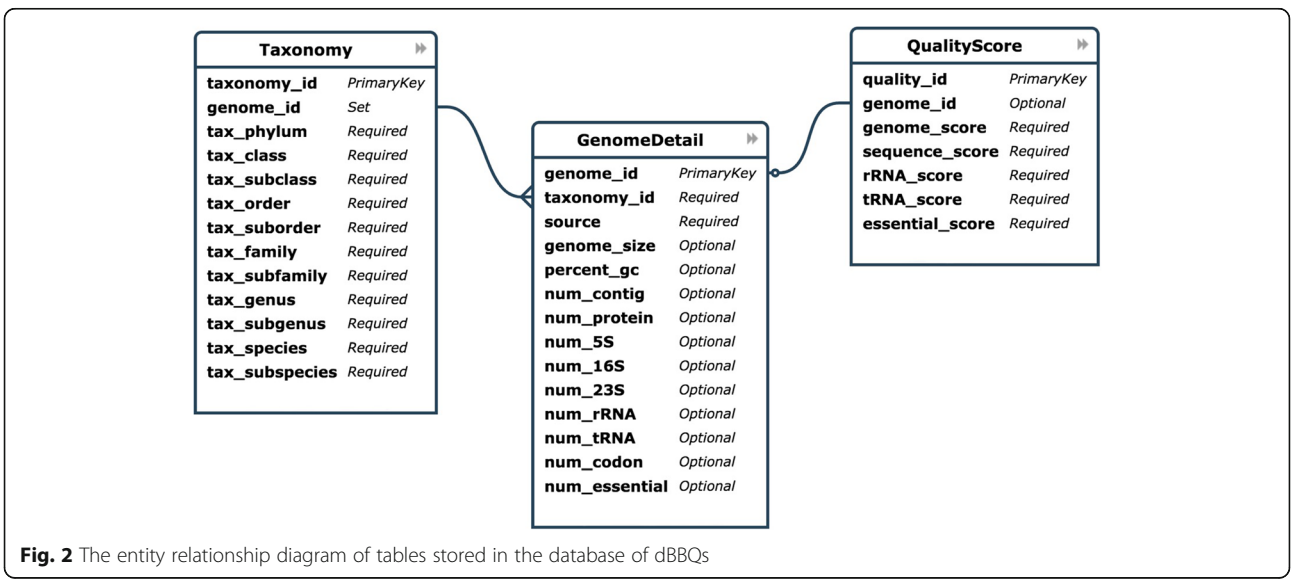

Fig. 2 The entity relationship diagram of tables stored in the database of dBBQs

Table 1 Number of genomes and average scores for each data source used in dBBQs

Data source	Number of genomes	Average of Genome Quality Score	Average of Raw Quality Scores			
			Sequence Quality Score	rRNA Score	tRNA Score	Essential Gene Score
Genbank	67,980	0.85	0.79	0.69	0.91	0.99
Genbank - SRA	11,768	0.69	0.38	0.72	0.74	0.92
Broad	2477	0.94	0.9	0.93	0.95	0.97
Kbase	11,944	0.9	0.82	0.86	0.93	0.99
Patric	1998	0.9	0.81	0.84	0.96	0.99
Total	96,167	0.856	0.74	0.808	0.898	0.972

charts will therefore be updated instantly. For example, when GenBank is selected from the donut chart of Genome Repositories, all bar charts and table will update their information dynamically. To be easy for user to focus at the main score first, we differentiated by picking different colors. The color of bar chart of genome quality score is in red color while other scores is in blue color.

Search section

Through the dBBQs front page, it is integrated with the search function. It can be found at a search box below the Taxonomy bar chart. This search box allows users to scope the group of bacterial genomes by searching for the name. For example, when the '*Escherichia coli*' search term was supplied to the search box all charts and table were filtered and displayed for only genomes containing a word '*Escherichia coli*' in their name. Furthermore, on the right of search box, users can download the search results in the table for the further analysis. A result file will be generated in CSV format, which can be open on many spreadsheet programs such as Microsoft Excel and Number.

Genome quality and statistics section

Any information in detail can be retrieved by clinking at the name of genome on the table. The genome quality and statistics page will start at the new tab on the browser (Fig. 5). This page comprises of five sections represented by five different frames: details of 'all scores and taxonomy' in white frame, details of 'sequence quality score' in green frame, details of 'rRNA score' in blue frame, details of 'tRNA score' in yellow frame, and details of 'essential gene score' in red frame. In addition, users can download data in

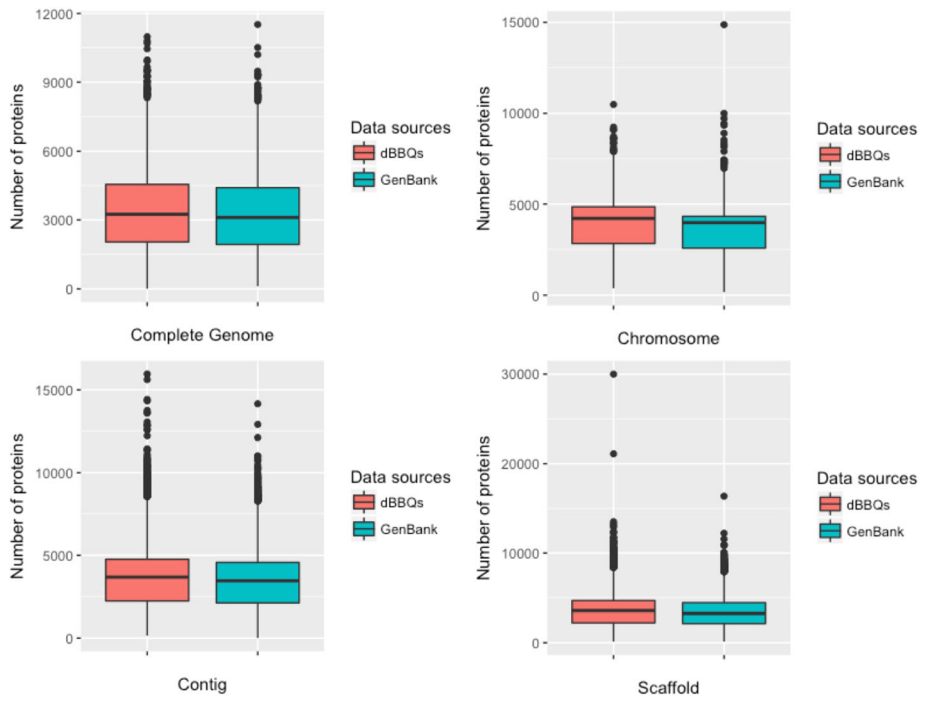

Fig. 3 Comparison of protein prediction between dBBQs and GenBank

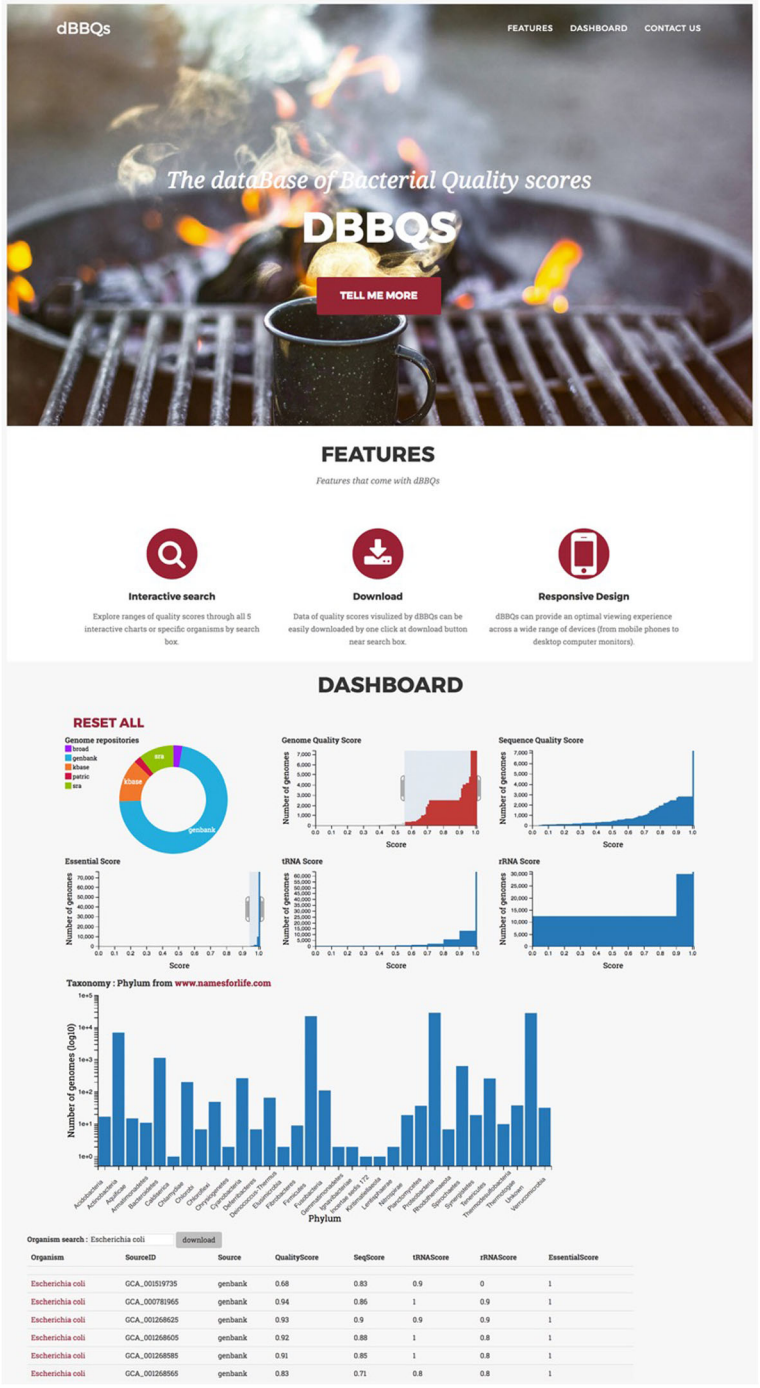

Fig. 4 The front page of dBBQs database and interactive charts of all quality scores

five sections above in CSV format by clicking at the button next to the Genome Quality Score. These attributes will give users details behind a calculation of all quality scores and leverages complex analytics when genomes have very similar scores but different in some details.

Conclusion

dBBQs provides quality scores for all available genome sequences with a user-friendly Web-based user interface. These scores can be used as one of cut-offs to get a high-quality set of genomes for testing bioinformatics tools or improving the analysis. Additionally, all data of

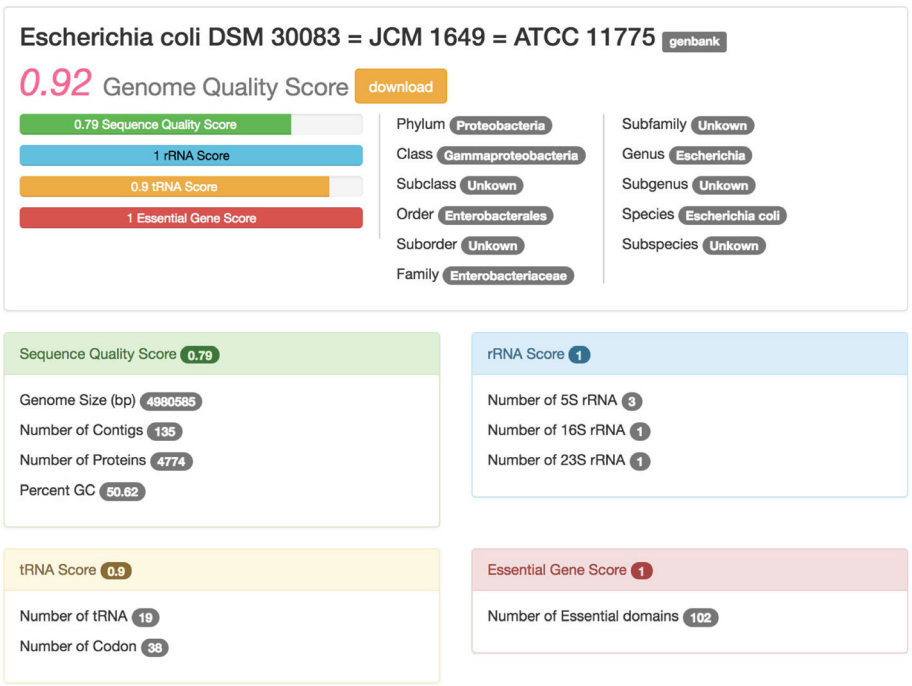

Fig. 5 Genome quality score and statistics card for Acaricomes phytoseiuli DSM 14247

four measurements that were combined to make the quality score for each genome, can be download in CSV format. The data table can be imported to a network and molecular profiling tool like CytoScape. By using CytoScape, the data table can potentially be used as node attributes for further analysis on pathway comparison using KEGG or BioCyc plugins.

Moreover, we plan to release our API to support the connection between other bioinformatics websites and our database. Also, a Web tool for calculating of quality score will be added to the website to allow users to upload genome sequences and get the genome quality scores. The database of dBBQs will be update regularly as number of genomes in public databases growing rapidly and is freely use for non-commercial purpose. These extensions of functionality and long term intention will help contribute largely to the analysis of quality of genomic data in bacterial research community.

Availability and requirements

Project name: dBBQs: dataBase of Bacterial Quality scores Project home page: http://arc-gem.uams.edu/dbbqs Operation system(s): Web based, Platform independent Programming language: HTML, CSS, JavaScript, Python

Abbreviations

API: Application program interface; CSV: Comma-separated Value; dBBQs: dataBase of Bacterial Quality scores; Kbase: The U.S. Department of Energy Systems Biology Knowledgebase; NCBI: National Center for Biotechnology information; PATRIC: Pathosystems Resource Integration Center; SRA: Sequence Read Archive

Acknowledgements

The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin, the Arkansas High Performance Computing Center (AHPCC), multiple National Science Foundation grants, and the Arkansas Economic Development Commission for providing HPC resources that have contributed to the research results reported within this paper. URL: http://www.tacc.utexas.edu and http://hpc.uark.edu.

Funding

The publication cost of this article was funded by NIH/NIGMS grant 1P20GM121293 and from the Helen Adams & Arkansas Research Alliance Endowment in the Department of Biomedical Informatics, College of Medicine at UAMS.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the dBBQs database, http://arc-gem.uams.edu/dbbqs

Authors' contributions

VW designed the project, collected data, performed the analysis, wrote the website, and drafted the manuscript. PP contributed back-end coding and constructed the database. IN helped design the website, discussed the results and interpretation of final data. DU conceived and directed the project. All participated in finalizing and approved the manuscript.

Ethics approval and consent to participate

Not applicable.

About this supplement

This article has been published as part of *BMC Bioinformatics* Volume 18 Supplement 14, 2017: Proceedings of the 14th Annual MCBIOS conference. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-18-supplement-14.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Published: 28 December 2017

References

1. Koren S, Phillippy A. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol. 2015;
2. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLOS Biol Public Library of Science. 2015; 13:e1002195.
3. Mavromatis K, Land ML, Brettin TS, Quest DJ, Copeland A, Clum A, et al. The fast changing landscape of sequencing technologies and their impact on microbial genome assemblies and annotation. Liu Z, editor. PLoS One Public Library of Science. 2012;7:e48837.
4. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. Extensive error in the number of genes inferred from draft genome assemblies. PLoS Comput Biol Public Library of Science. 2014;10:e1003998.
5. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res Oxford University Press. 2016;44:D67–72.
6. Broad Institute. Microbial Genomes Research Areas [Internet]. [cited 2015 Apr 9]. Available from: https://olive.broadinstitute.org/
7. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 2014;42:D581–91.
8. NCBI. Bacterial Genome ftp site. p. ftp://ftp.ncbi.nlm.nih.gov/genomes/. Accessed Jan 2017.
9. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, et al. Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol American Society for Microbiology. 2012;50:1355–61.
10. NCBI. Sequence Read Archive. p. ftp://ftp.ncbi.nlm.nih.gov/sra/.
11. Pathosystems Resource Integration Center (PATRIC) ftp download site. :ftp://ftp.patricbrc.org/patric2/genomes/.
12. Land ML, Hyatt D, Jun S-R, Kora GH, Hauser LJ, Lukjancenko O, et al. Quality scores for 32,000 genomes. Stand Genomic Sci BioMed Central. 2014;9:20.
13. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
14. Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res Oxford University Press. 2007;35:3100–8.
15. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25:955–64.
16. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res Oxford University Press. 2011;39:W29–37.
17. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42:D222–30.