D606–D612 *Nucleic Acids Research, 2020, Vol. 48, Database issue* *Published online 31 October 2019*
doi: 10.1093/nar/gkz943

The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities

James J. Davis[1,2,*], Alice R. Wattam[2,3], Ramy K. Aziz[4,5], Thomas Brettin[1,6], Ralph Butler[2,7], Rory M. Butler[2], Philippe Chlenski[8], Neal Conrad[1,2], Allan Dickerman[3], Emily M. Dietrich[1,6], Joseph L. Gabbard[9], Svetlana Gerdes[8], Andrew Guard[1], Ronald W. Kenyon[3], Dustin Machi[3], Chunhong Mao[3], Dan Murphy-Olson[1,6], Marcus Nguyen[1,2], Eric K. Nordberg[10], Gary J. Olsen[11,12], Robert D. Olson[1,2], Jamie C. Overbeek[1,2], Ross Overbeek[1], Bruce Parrello[1,2], Gordon D. Pusch[8], Maulik Shukla[1,2], Chris Thomas[1], Margo VanOeffelen[8], Veronika Vonstein[8], Andrew S. Warren[3], Fangfang Xia[1,2], Dawen Xie[3], Hyunseung Yoo[1,2] and Rick Stevens[6,13]

[1]University of Chicago Consortium for Advanced Science and Engineering, University of Chicago, Chicago, IL 60637, USA, [2]Division of Data Science and Learning, Argonne National Laboratory, Argonne, IL 60439, USA, [3]Biocomplexity Institute and Initiative, University of Virginia, Charlottesville, VA 22904, USA, [4]Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, 11562 Cairo, Egypt, [5]Center for Genome and Microbiome Research, Cairo University, 11562 Cairo, Egypt, [6]Computing Environment and Life Sciences, Argonne National Laboratory, Argonne, IL 60439, USA, [7]Middle Tennessee State University, Murfreesboro, TN 37132, USA, [8]Fellowship for Interpretation of Genomes, Burr Ridge, IL 60527, USA, [9]Virginia Tech, Blacksburg, VA 24061, USA, [10]Transportation Institute, Virginia Tech University, Blacksburg, VA 24061, USA, [11]Department of Microbiology, University of Illinois, Urbana, IL 61801, USA, [12]Carl R. Woese Institute for Genomic Biology, University of Illinois, Urbana, IL 61801, USA and [13]University of Chicago, Department of Computer Science, Chicago, IL 60637, USA

Received September 12, 2019; Revised October 07, 2019; Editorial Decision October 08, 2019; Accepted October 11, 2019

ABSTRACT

The PathoSystems Resource Integration Center (PATRIC) is the bacterial Bioinformatics Resource Center funded by the National Institute of Allergy and Infectious Diseases (https://www.patricbrc.org). PATRIC supports bioinformatic analyses of all bacteria with a special emphasis on pathogens, offering a rich comparative analysis environment that provides users with access to over 250 000 uniformly annotated and publicly available genomes with curated metadata. PATRIC offers web-based visualization and comparative analysis tools, a private workspace in which users can analyze their own data in the context of the public collections, services that streamline complex bioinformatic workflows and command-line tools for bulk data analysis. Over the past several years, as genomic and other omics-related experiments have become more cost-effective and widespread, we have observed considerable growth in the usage of and demand for easy-to-use, publicly available bioinformatic tools and services. Here we report the recent updates to the PATRIC resource, including new web-based comparative analysis tools, eight new services and the release of a command-line interface to access, query and analyze data.

INTRODUCTION

The Bioinformatics Resource Center (BRC) program was established by the National Institute of Allergy and Infectious Diseases (NIAID) in 2004 with a primary focus on providing access to genome sequence data and analysis tools for studying pathogens. PathoSystems Resource Integration Center (PATRIC) began as one of the original centers tasked with supporting comparative analysis of bacterial pathogens (1–3). In 2009, PATRIC merged with the National Microbial Pathogen Database Resource (NMPDR) BRC (4), which had developed the successful SEED database and RAST (Rapid Annotation using Subsystem Technology) annotation system for uniformly curating and projecting genome annotations across microbial species (5–8). Over the years, the PATRIC resource has expanded

[*]To whom correspondence should be addressed. Tel: +1 630 252 1190; Email: jjdavis@anl.gov
Present address: James J. Davis, Argonne National Laboratory, Computing, Environment and Life Sciences, 9700 S. Cass Avenue, Argonne, IL 60439, USA.

Published by Oxford University Press on behalf of Nucleic Acids Research 2019.

and adapted to keep pace with the growth in bioinformatic datasets and the need for associated analysis tools. As of September 2019, PATRIC includes over 250 000 publicly available microbial genomes and a rich comparative analysis environment.

Since its launch in 2008, RAST (http://rast.nmpdr.org) has performed ∼700 000 genome annotation jobs for private users. By providing access to genome feature identification scripts developed by the academic community and consistent projections of well-curated protein functions from the SEED, RAST serves as a model for a successful bioinformatic service because it alleviates the need for users to build their own custom annotation pipelines, and its consistency enables downstream comparative analyses. Using RAST as a template, in 2014 PATRIC began implementing a variety of bioinformatic services through the website allowing users to assemble and annotate genome sequences, reconstruct metabolic models, analyze SNPs and INDELs, and analyze and compare RNA-seq experiments. The results of these analysis jobs could then be compared with the publicly available genomic and other omic data collections in the resource, while being kept private within the user's workspace environment. By the end of 2016, PATRIC was processing ∼1500 service jobs per month, not including jobs being submitted to the RAST website (3).

Since last described in *Nucleic Acids Research* in 2016 (3), PATRIC has undergone a series of updates and improvements. The data collection has been improved, especially in the area of antimicrobial resistance (AMR) (9); the web browsing environment has been enhanced with new tools and visualizations; and improvements to the workspace have also made it easier to find and share research project data. A command line interface (CLI) for bulk data acquisition and analysis has been built and released for distribution on Mac, Linux and Windows systems. PATRIC has also launched eight new bioinformatic services, with recent emphasis being placed on the ability to analyze data from mixed cultures or metagenomic samples. At last, a rich collection of tutorials has been created to help users with these new tools (https://docs.patricbrc.org/tutorial/). This report describes many of the recent unpublished updates to the PATRIC resource.

WHAT'S NEW IN PATRIC?

Data growth and enhancements

One of the most dramatic changes in supporting bioinformatic work since the beginning of the BRC program has been the exponential growth in publicly available microbial genome sequences (Figure 1). The collection of private user genome sequences that have been annotated and indexed by PATRIC has also grown since the establishment of the workspace environment, and may actually exceed the size of the public genome sequence collection within the next year (Figure 1). Although the private set includes some reanalyzed genome sequences,

we see no indication that microbial genome sequencing and its related bioinformatic analyses is slowing. The increase in publicly available genome sequence data and related structured metadata has also revolutionized the

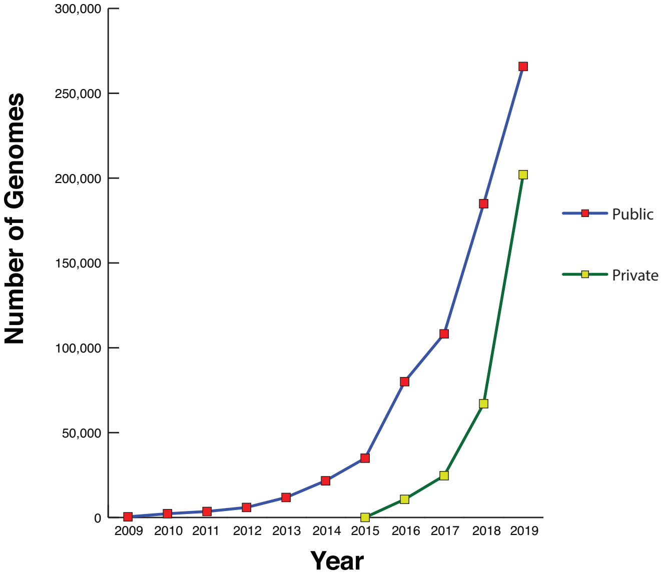


Figure 1. Cumulative growth of public and private genomes in PATRIC.

types of experimental analyses that are possible. For instance, PATRIC provides structured and manually curated metadata associated with each genome, including laboratory-derived AMR phenotypes, host organisms, isolation sources, human body site data and geographical information. These collections of structured metadata provide the foundation for running machine learning and deep learning experiments (10,11), and for providing predictive tools to users (9). We anticipate that the increased use of artificial intelligence techniques in bioinformatics will drive experimental design decisions and ultimately shorten the time required for genetic and other laboratory-based characterization experiments.

Supporting AMR research is a major focus area for data collection and curation at PATRIC. We actively curate both AMR protein annotations and laboratory-derived AMR phenotype data associated with public genomes. The annotation system is able to accurately project over 600 hand-curated AMR protein functions. It also contains a large collection of closely related non-AMR protein functions that have been curated to prevent false predictions of AMR functions. To provide an additional means of comparison, the annotation system also searches for genes with high similarity to those curated by the CARD (12) and NCBI AMR gene database projects (13). The laboratory-derived AMR phenotype collection has been generated by curating data from the literature, NCBI (https://www.ncbi.nlm.nih.gov/pathogens) and other public sources. It has grown to include over 40 000 genome sequences and is being used by researchers worldwide. We have also added over 10 000 plasmid and prophage sequences because of their importance in studying and combatting AMR.

Services

The services provided by PATRIC are designed to enable easy access to complex bioinformatic workflows. They can

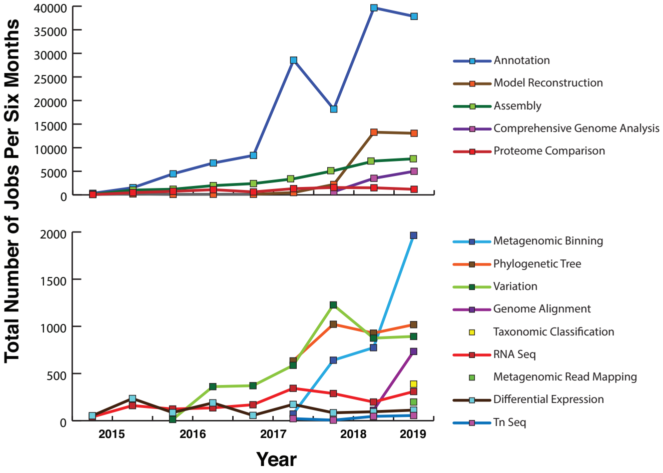


Figure 2. User-initiated analysis jobs completed by the PATRIC bioinformatic services. The top plot shows the use of high-volume services. The bottom plot shows the usage of lower volume and new services. Note the difference in scale between the two plots.

be accessed via the PATRIC web interface and CLI. Most services have the capacity to handle hundreds or even thousands of jobs per day. Jobs are typically run on a series of internal servers, with surge capacity being handled by a large computing cluster. The PATRIC services have grown in popularity since 2014, and as of September of 2019, over 263 000 jobs have been successfully completed (Figure 2).

Noteworthy updates to existing services. Three of our pre=existing services, Genome Assembly, Genome Annotation and RNA-seq analysis, have undergone several noteworthy updates. The Genome Assembly Service has been rebuilt with a new job scheduler that enables a fairer job-queuing process that prevents large jobs from creating bottlenecks (14). In addition to SPAdes (15), we have added Canu (16) for long-read assembly and Unicycler for hybrid long- and short-read assemblies (17). We also provide an image of the assembly graph using Bandage (18), and assemblies can be polished using Racon (19) and Pilon (20) for long- and short-read assemblies respectively. At last, read mapping is performed to generate accurate coverage statistics using Bowtie2 (21) or Minimap2 (22), and SAMtools (23). Two new additions to the Genome Annotation Service include the ability to annotate bacteriophage genome sequences (24) and the computation of genome quality statistics that are based on the CheckM application (25) and an internal RAST model that assesses quality based on the occurrence and completeness of subsystem roles in the genome (26). The RNA-seq analysis Service has also been updated to enable experiments studying host response to microbial infections. To support this, we have added several common eukaryotic host reference genomes including *Caenorhabditis elegans, Danio rerio, Drosophila melanogaster, Gallus gallus, Homo sapiens, Macaca mulatta, Mus musculus, Mustela putorius furo, Rattus norvegicus* and *Sus scrofa.* We have also recently added HISAT2 (hierarchical indexing for spliced alignment of transcripts) (27), a highly-efficient system for aligning reads from RNA-Seq experiments to host genomes and enabled import of datasets from SRA in the RNA-seq interface, further enhancing the capability to perform mixed differential expression analysis of public and private data.

Comprehensive genome analysis. One of the most common use case for analysis of private genomes at PATRIC is for researchers to assemble and then annotate their genome sequences using two separate services. In the Spring of 2018, we launched a streamlined Comprehensive Genome Analysis 'meta-service' that accepts sequencing reads, computes the assembly and annotation, and provides a user-friendly description of the genome. The output includes a genome quality assessment, AMR genes and phenotype predictions, specialty genes, subsystem overview, identification of the closest genome sequences, a phylogenetic tree and a list of features that distinguish the genome from its nearest neighbors. The Comprehensive Genome Analysis Service has quickly risen to be one of the most popular services in PATRIC with over 11 000 jobs being completed since its launch in April 2018.

Phylogenetic trees. The ability to reconstruct and visualize evolutionary relationships lies at the heart of biology. In 2017, PATRIC launched the Phylogenetic Tree Service that enables users to build high-quality phylogenetic trees for public and private genome sequences. The service currently offers two workflows to the user. The first is a protein-based tree-building workflow called 'All Shared Proteins,' which uses the Phylogenomic Estimation with Progressive Refinement (PEPR) pipeline (https://github.com/enordber/pepr). PEPR works by defining shared protein families *de novo* for a genome group using BLAST (28) and HMMER (29) to identify similar proteins and MCL (30) to build clusters. Then alignments are generated using Muscle (31), and trimmed with Gblocks (32). At last, based on the user's preference, PEPR computes the tree using either FastTree (33) or RAxML (34). In 2019, we launched a second, faster, phylogenetic tree building workflow called 'Codon Trees.' It leverages predefined PATRIC global protein families (PGFams) (35), selecting a user-specified number of families (10–1000) that are single-copy (or nearly so) among members of a genome group. Alignments are generated for protein sequences of each family using Muscle (31), and their corresponding nucleotide sequences are aligned to this using the *codonalign* function of BioPython (36). A concatenated alignment of all proteins and nucleotides is written to a PHYLIP-formatted file (37). A partitions file for RaxML (34) is then generated, which describes the alignment in terms of the proteins and nucleotides in the first, second, and third codon positions. Support values are generated from 100 rounds of rapid bootstrapping in RaxML (38).

In addition to the Newick-formatted tree files, the Phylogenetic Tree Service returns a portable document file (PDF), a portable network graphics (PNG) and a scalable vector graphics (SVG) image file of the midpoint rooted tree images generated by FigTree (http://tree.bio.ed.ac.uk/software/figtree/). The phylogenetic tree view on the PATRIC website allows researchers to select nodes and leaves, enabling the user to create groups from specific clades for further analysis. It also generates a genome report that provides a list of the genome sequences and protein families used in the construction of tree and the counts of genes, proteins, amino acids and nucleotides used to compute the tree. At last, problematic genome sequences that could be removed to increase the gene selection and improve the strength of the tree are listed. Since it was built, nearly 5000 jobs have been processed by the Phylogenetic Tree Service.

Fastq utilities. Assessing the quality of sequencing reads is an important first step for ensuring that subsequent analyses, such as assembly, annotation, etc. are accurate. The Fastq Utilities Service, launched in July 2019, enables users to align reads, measure base call quality, and trim low-quality sequences from read files. The service accepts long- or short-read files in single or paired-end format. It can also retrieve read files directly from the NCBI Sequence Read Archive (SRA) using a run identifier as input. The service has three components, 'trim,' 'FastQC,' and 'align,' which can be used independently or in any combination. The trimming component uses Trim Galore (39), which is a Perl wrapper around the Cutadapt (40) and FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc) tools. The FastQC component provides quality control checks on raw sequence data coming from high-throughput sequencing pipelines and enables rapid quality control by indicating problems that could impact downstream analyses. The aligning function aligns reads to a reference genome sequence using Bowtie2 (21,41), saving unmapped reads and generating SamStat (42) reports of the amount and quality of the alignments.

Genome alignment. In November 2018, PATRIC launched the Genome Alignment Service to enable users to compute whole genome sequence alignments. This service uses the progressiveMauve application (43), which constructs positional homology multiple genome sequence alignments in an extension of the original Mauve algorithm (44). The service enables researchers to align up to twenty genome sequences at a time. The output of the service includes a visual display of the genome that allows users to view and explore the entire genome sequence alignment or to zoom in to compare individual regions or genes (Figure 3).

Similar genome finder. When a researcher has a new genome sequence, one of the first things they want to identify is the closest relatives for the organism, but this can be difficult when the public collection is so large. PATRIC provides a service called the Similar Genome Finder to allow researchers to rapidly identify similar genome sequences using Mash (45). Mash works by reducing large sequences to small representative sketches, which can be used to estimate mutation distances based on shared k-mers. PATRIC allows for comparison against all public genome sequences or the NCBI reference genome set. The tool allows researchers to adjust the search sensitivity by selecting the maximum number of k-mers held in common, *P*-value threshold or the distance. The results are returned as a list of the most similar genome sequences with corresponding metadata. As with all PATRIC tables, researchers can select sequences to create groups for later analysis, or download the results.

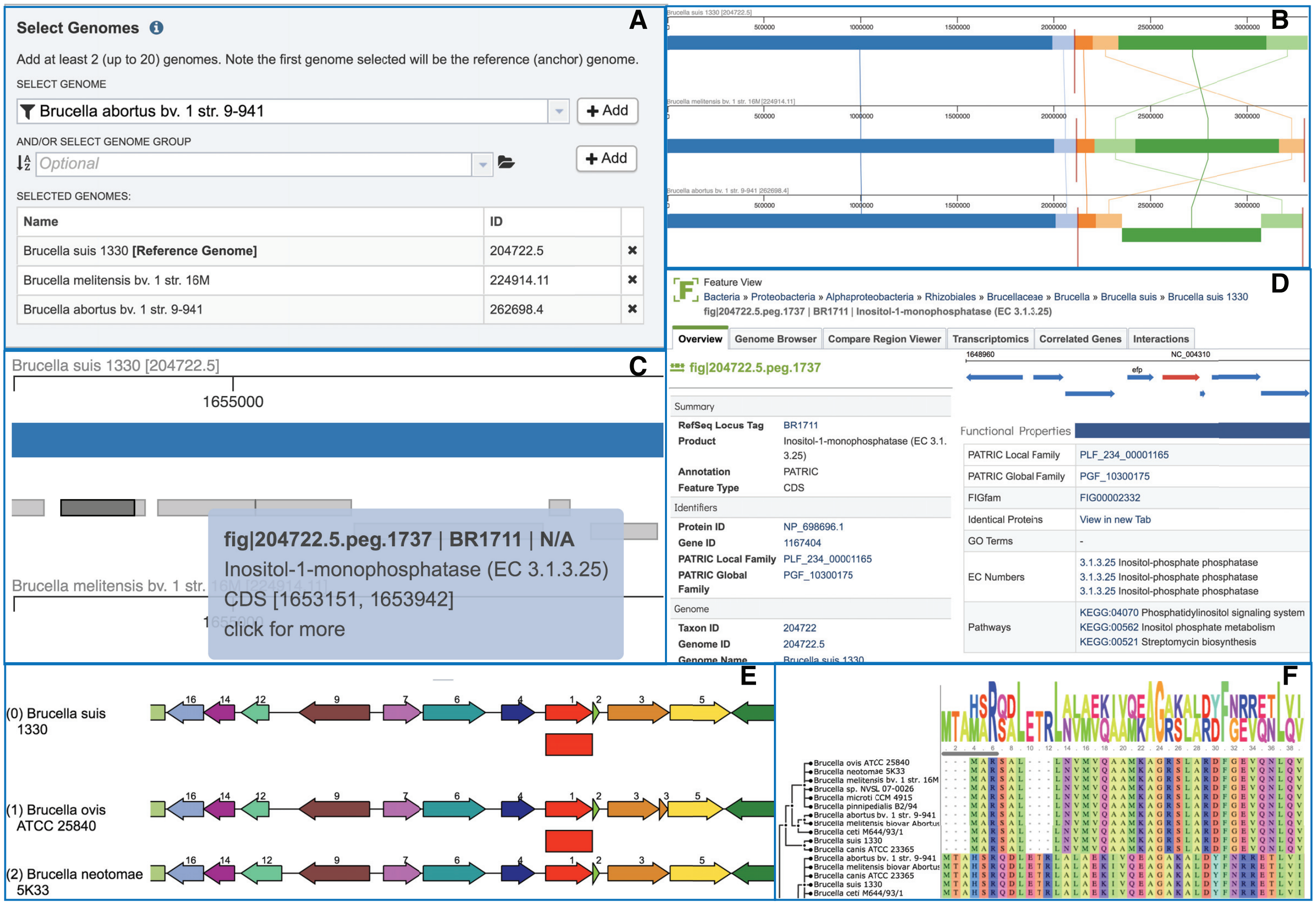
Taxonomic classification. Launched in March of 2019, the Taxonomic Classification Service identifies the taxonomic composition of mixed or metagenomic samples. This service uses the Kraken2 (46) application, which identifies k-mers that are indicative of various taxonomic units. The Kraken database used by the service is a full build that is based on all RefSeq genome sequences (47), the human genome sequence, plasmids and vector sequences. Job output includes the standard Kraken report format, with each bacterial taxon hyperlinked to the matching page in PATRIC. The service also returns a Krona plot (48) that shows the percentage of reads that mapped to each taxon and allows the user to explore selected taxa.

Metagenomic read mapping. Researchers studying AMR or virulence may be interested in analyzing genes in mixed or metagenomic read sets. The Metagenome Read Mapping Service enables researchers to search for these specific genes in a set of reads. It works by aligning reads against a reference gene using KMA, which uses k-mer seeding and the Needleman–Wunsch algorithm to accurately align the reads to the genes of interest (49). Users can currently align against the reference gene sets from the Comprehensive Antibiotic Resistance Database (CARD) (50) and the Virulence Factor Database (VFDB) (51). The service returns html and text versions of the standard KMA report, which shows detailed mapping information, links to genes in PATRIC with high similarity, and a consensus sequence assembled from the aligned reads.

Metagenomic binning. Launched in August 2017, the Metagenomic Binning Service assembles reads from a metagenomic sample into contigs and then attempts to separate these contigs into bins that represent the genomes of individual species. These bins are then fully annotated and detailed quality statistics are computed for each bin. The binning algorithm starts by scanning contigs for specific marker proteins that are almost always singly occurring in the genome. The marker-protein similarity is used to recruit similar genomes from PATRIC, which are then used to recruit additional contigs based on distinguishing protein k-mers. Similar to single isolate genomes, the bins are placed in the user's workspace and indexed within the PATRIC database as private genomes, allowing the full use of the PATRIC comparative analysis and visualization tools for each bin.

Web-based analysis tools

The PATRIC website offers several interactive visual analytic tools that enable users to compare omics datasets. These tools integrate data of various types, perform some computational tasks and render interactive visualizations for the user. PATRIC currently supports many web-based analysis tools, such as the Heat Map Viewer for comparing shared protein content, the Pathway Viewer for exploring metabolic pathways and the Genome Browser for displaying genomic features on the chromosome. We have added



Select Genomes

Add at least 2 (up to 20) genomes. Note the first genome selected will be the reference (anchor) genome.

SELECT GENOME

Brucella abortus bv. 1 str. 9-941 + Add

AND/OR SELECT GENOME GROUP

Optional + Add

SELECTED GENOMES:

Name	ID	
Brucella suis 1330 **[Reference Genome]**	204722.5	✖
Brucella melitensis bv. 1 str. 16M	224914.11	✖
Brucella abortus bv. 1 str. 9-941	262698.4	✖

fig|204722.5.peg.1737 | BR1711 | N/A
Inositol-1-monophosphatase (EC 3.1.3.25)
CDS [1653151, 1653942]
click for more

Feature View

Bacteria » Proteobacteria » Alphaproteobacteria » Rhizobiales » Brucellaceae » Brucella » Brucella suis » Brucella suis 1330

fig|204722.5.peg.1737 | BR1711 | Inositol-1-monophosphatase (EC 3.1.3.25)

Overview | Genome Browser | Compare Region Viewer | Transcriptomics | Correlated Genes | Interactions

fig|204722.5.peg.1737

Summary

RefSeq Locus Tag	BR1711
Product	Inositol-1-monophosphatase (EC 3.1.3.25)
Annotation	PATRIC
Feature Type	CDS

Identifiers

Protein ID	NP_698696.1
Gene ID	1167404
PATRIC Local Family	PLF_234_00001165
PATRIC Global Family	PGF_10300175

Genome

Taxon ID	204722
Genome ID	204722.5
Genome Name	Brucella suis 1330

Functional Properties

PATRIC Local Family	PLF_234_00001165
PATRIC Global Family	PGF_10300175
FIGfam	FIG00002332
Identical Proteins	View in new Tab
GO Terms	-
EC Numbers	3.1.3.25 Inositol-phosphate phosphatase 3.1.3.25 Inositol-phosphate phosphatase 3.1.3.25 Inositol-phosphate phosphatase
Pathways	KEGG:04070 Phosphatidylinositol signaling system KEGG:00562 Inositol phosphate metabolism KEGG:00521 Streptomycin biosynthesis

A
B
Brucella suis 1330 [204722.5]
Brucella melitensis bv. 1 str. 16M [224914.11]
Brucella abortus bv. 1 str. 9-941 [262698.4]
500000
1000000
1500000
2000000
2500000
3000000
C
Brucella suis 1330 [204722.5]
1655000
Brucella melitensis bv. 1 str.
D
1648960
NC_004310
efp
E
(0) Brucella suis 1330
(1) Brucella ovis ATCC 25840
(2) Brucella neotomae 5K33
16
14
12
9
7
6
4
1
2
3
5
F
Brucella ovis ATCC 25840
Brucella neotomae 5K33
Brucella melitensis bv. 1 str. 16M
Brucella sp. NVSL 07-0026
Brucella microti CCM 4915
Brucella pinnipedialis B2/94
Brucella abortus bv. 1 str. 9-941
Brucella melitensis biovar Abortus
Brucella ceti M644/93/1
Brucella suis 1330
Brucella canis ATCC 23365
Brucella abortus bv. 1 str. 9-941
Brucella melitensis biovar Abortus
Brucella canis ATCC 23365
Brucella suis 1330
Brucella ceti M644/93/1

Figure 3. A data analysis workflow in PATRIC using the Genome Alignment Service. (**A**) The website interface allows the selection of genomes; (**B**) visualization of the aligned genomic regions with any deletions, insertions or rearrangements; (**C**) zooming in on the alignment will show the genes on the forward and reverse stands, which can be selected; (**D**) selecting a specific gene from the Genome Alignment viewer opens the PATRIC Feature Page, where all the data available for that gene are shown. (**E**) The Compare Region View tab on the PATRIC Gene Page shows conservation of the selected gene (shown in red), and also the surrounding genes. (**F**) Each gene is assigned to a genus-specific (PLFam) or global (PGFam) protein family that can be selected from the Feature Page, and the family members can be compared using the Multiple Sequence Alignment/Gene Tree tool.

two new visualizations to the PATRIC website that originally existed on the RAST and SEED websites, but required significant reengineering to be functional for use with hundreds of thousands of genomes.

Compare region viewer. The Compare Region Viewer allows researchers to compare gene neighborhoods (genetic loci or chromosomal clusters) across many species. A user selects a gene of interest, the size of the genomic region and the number of genomes for the comparison. The display renders the BLAST similarity of the focus gene, and the similarity of the surrounding genes within the region (Figure 3E).

In RAST, this tool relies on a precomputed database of all-to-all BLAST (28) similarities to determine the set of genomes having a match to the gene of interest, and computes a detailed pairwise comparison of genes in the selected region to color code the data. Due to the number of genomes in the PATRIC database, this method is too slow for real time use. The PATRIC version of this tool bases the focus gene lookup and color coding on either the genus-specific (PLFam) or global (PGFam) protein families (35), which are precomputed for each genome, so the search space is more scoped. However, this visualization is scalable because BLAST is only used to compute protein similarity for the focus genes within the set.

Subsystems. Subsystems are collections of functionally related proteins and are a vital conceptual device for identifying and projecting protein functions across species (7,52). PATRIC now computes and displays subsystem data for each public and privately annotated genome sequence. Subsystems, which result from manual annotation by a team of expert curators, are divided into Superclass (example: Metabolism), Class (example: Stress Response, Defense and Virulence), Subclass (example: Resistance to antibiotics and toxic compounds), Subsystem Name (example: Arsenic resistance) and the functional role of each of the included genes. Clicking on the subsystems tab for any genome provides three different views. The Subsystems Overview shows a pie chart that displays the percent of the genes that are in a particular Superclass. The Subsystems tab includes the number of genes found in a particular Superclass. The Genes tab includes a list of all the genes across all the subsystems, and includes the PATRIC and RefSeq locus tags (47). Subsystem information is not only available for individual genomes, but is also summed for each taxonomic level, all the way up to Superkingdom using the NCBI taxonomy (53). A heatmap view showing presence and absence of specific proteins per selected subsystem across a taxon or a specific genome group can be created by the user.

Command-Line Interface (CLI)

For the past 5 years, the PATRIC data store has been managed using a NoSQL Apache Solr database structure. To accommodate the rapidly growing data collection and to take advantages scalability and resilience, the PATRIC database architecture was converted to an Apache SolrCloud database architecture in the spring of 2019. The SolrCloud database is divided into a series of SolrCores for managing related data types, such as genome features, sequences and transcriptomic data. An underlying application programming interface (API) enables programmatic access to these cores and the data that they contain; however, data acquisition can become complex when navigating and merging fields from the various cores. We have developed a set of command-line scripts that use the API for accessing the data store and performing common analyses. This distribution is available for Mac, Windows and Linux operating systems, including Ubuntu and CentOS 6 and 7, and Fedora 28 and 29 (https://github.com/PATRIC3/PATRIC-distribution/releases). Both the distribution and the PATRIC website contain tutorials on how to use the scripts with examples (https://docs.patricbrc.org/cli_tutorial/). The 482MB distribution contains many of the underlying scripts of the PATIRC environment. Some enable the bulk downloading, merging and manipulation of data and others enable more complex analyses. The distribution also includes useful scripts from earlier SEED (5) and RASTtk (8) projects. A particularly noteworthy functionality offered by the PATRIC CLI distribution is the ability to manage files in the workspace. Users can log into a private workspace, create subdirectories, move files into or out of the workspace and launch annotation and assembly jobs. These scripts provide the means for assembling and annotating hundreds or even thousands of genome sequences. Additionally, we have also made the PATRIC workspace accessible via File Transfer Protocol (FTP), which provides an alternative means of moving large amounts of data into and out of the workspace. Users can access the workspace using the command-line or by using a FTP file manager. We plan to continue developing the command-line tools to allow greater access to services and easier data manipulation.

FUTURE DIRECTIONS

In 2020, the PATRIC team at the University of Chicago, University of Virginia and the Fellowship for Interpretation of Genomes will combine with the viral BRC team that supports the ViPR (Virus Pathogen Database and Analysis Resource) and IRD (Influenza Research Database) resources at the J. Craig Venter Institute (JCVI). The newly formed bacterial and viral BRC team (BV-BRC) will continue to maintain the PATRIC, IRD and ViPR websites while adding new crosscutting functionality. We intend to focus heavily on improving the utility of the new BV-BRC resource for epidemiological analysis, expanding the data store to include other data and metadata types, increasing access to structured data that can be used in artificial intelligence applications, and improving the deployment architecture for the tools and services.

FUNDING

National Institute of Allergy and Infectious Diseases (NIAID) [HHSN272201400027C to R.S.]. Funding for open access charge: NIAID.

Conflict of interest statement. None declared.

REFERENCES

1. Snyder,E., Kampanya,N., Lu,J., Nordberg,E.K., Karur,H., Shukla,M., Soneja,J., Tian,Y., Xue,T. and Yoo,H. (2006) PATRIC: the VBI pathosystems resource integration center. *Nucleic Acids Res.*, **35**, D401–D406.
2. Wattam,A.R., Abraham,D., Dalay,O., Disz,T.L., Driscoll,T., Gabbard,J.L., Gillespie,J.J., Gough,R., Hix,D. and Kenyon,R. (2013) PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.*, **42**, D581–D591.
3. Wattam,A.R., Davis,J.J., Assaf,R., Boisvert,S., Brettin,T., Bun,C., Conrad,N., Dietrich,E.M., Disz,T. and Gabbard,J.L. (2016) Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res.*, **45**, D535–D542.
4. McNeil,L.K., Reich,C., Aziz,R.K., Bartels,D., Cohoon,M., Disz,T., Edwards,R.A., Gerdes,S., Hwang,K. and Kubal,M. (2006) The National Microbial Pathogen Database Resource (NMPDR): a genomics platform based on subsystem annotation. *Nucleic Acids Res.*, **35**, D347–D353.
5. Overbeek,R., Olson,R., Pusch,G.D., Olsen,G.J., Davis,J.J., Disz,T., Edwards,R.A., Gerdes,S., Parrello,B. and Shukla,M. (2013) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.*, **42**, D206–D214.
6. Aziz,R.K., Bartels,D., Best,A.A., DeJongh,M., Disz,T., Edwards,R.A., Formsma,K., Gerdes,S., Glass,E.M. and Kubal,M. (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.
7. Overbeek,R., Begley,T., Butler,R.M., Choudhuri,J.V., Chuang,H.-Y., Cohoon,M., de Crécy-Lagard,V., Diaz,N., Disz,T. and Edwards,R. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, **33**, 5691–5702.
8. Brettin,T., Davis,J.J., Disz,T., Edwards,R.A., Gerdes,S., Olsen,G.J., Olson,R., Overbeek,R., Parrello,B. and Pusch,G.D. (2015) RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep.*, **5**, 8365.
9. Antonopoulos,D.A., Assaf,R., Aziz,R.K., Brettin,T., Bun,C., Conrad,N., Davis,J.J., Dietrich,E.M., Disz,T. and Gerdes,S. (2019) PATRIC as a unique resource for studying antimicrobial resistance. *Brief. Bioinform.*, **20**, 1094–1102.
10. Nguyen,M., Brettin,T., Long,S.W., Musser,J.M., Olsen,R.J., Olson,R., Shukla,M., Stevens,R.L., Xia,F. and Yoo,H. (2018) Developing an in silico minimum inhibitory concentration panel test for *Klebsiella pneumoniae*. *Sci. Rep.*, **8**, 421.
11. Nguyen,M., Long,S.W., McDermott,P.F., Olsen,R.J., Olson,R., Stevens,R.L., Tyson,G.H., Zhao,S. and Davis,J.J. (2019) Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal *Salmonella*. *J. Clin. Microbiol.*, **57**, e01260-18.
12. Jia,B., Raphenya,A.R., Alcock,B., Waglechner,N., Guo,P., Tsang,K.K., Lago,B.A., Dave,B.M., Pereira,S. and Sharma,A.N. (2016) CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, **45**, D566–D573.

13. Feldgarden,M., Brover,V., Haft,D.H., Prasad,A.B., Slotta,D.J., Tolstoy,I., Tyson,G.H., Zhao,S., Hsu,C.-H. and McDermott,P.F. (2019) Validating the NCBI AMRFinder tool and resistance gene database using antimicrobial resistance Genotype-Phenotype correlations in a collection of NARMS isolates. *Antimicrob. Agents Chemother.*, **63**, e00483-19.
14. Yoo,A.B., Jette,M.A. and Grondona,M. (2003) Slurm: Simple linux utility for resource management. *Workshop on Job Scheduling Strategies for Parallel Processing*. Springer, Berlin, Heidelberg, pp. 44–60.
15. Bankevich,A., Nurk,S., Antipov,D., Gurevich,A.A., Dvorkin,M., Kulikov,A.S., Lesin,V.M., Nikolenko,S.I., Pham,S. and Prjibelski,A.D. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.
16. Koren,S., Walenz,B.P., Berlin,K., Miller,J.R., Bergman,N.H. and Phillippy,A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, **27**, 722–736.
17. Wick,R.R., Judd,L.M., Gorrie,C.L. and Holt,K.E. (2017) Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.*, **13**, e1005595.
18. Wick,R.R., Schultz,M.B., Zobel,J. and Holt,K.E. (2015) Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, **31**, 3350–3352.
19. Vaser,R., Sović,I., Nagarajan,N. and Šikić,M. (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.*, **27**, 737–746.
20. Walker,B.J., Abeel,T., Shea,T., Priest,M., Abouelliel,A., Sakthikumar,S., Cuomo,C.A., Zeng,Q., Wortman,J. and Young,S.K. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**, e112963.
21. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
22. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
23. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
24. McNair,K., Aziz,R.K., Pusch,G.D., Overbeek,R., Dutilh,B.E. and Edwards,R. (2018) Phage Genome Annotation Using the RAST Pipeline. In: Clokie,MRJ, Kropinski,AM and Lavigne,R (eds). *Bacteriophages Methods and Protocols*. Humana Press, NY, Vol. **3**, pp. 231–238.
25. Parks,D.H., Imelfort,M., Skennerton,C.T., Hugenholtz,P. and Tyson,G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, **25**, 1043–1055.
26. Parrello,B., Butler,R., Chlenski,P., Olson,R., Overbeek,J., Pusch,G.D., Vonstein,V. and Overbeek,R. (2019) A machine learning-based service for estimating quality of genomes using PATRIC. *BMC Bioinformatics*, **20**, 486.
27. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
28. Boratyn,G.M., Camacho,C., Cooper,P.S., Coulouris,G., Fong,A., Ma,N., Madden,T.L., Matten,W.T., McGinnis,S.D. and Merezhuk,Y. (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.*, **41**, W29–W33.
29. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
30. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
31. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
32. Talavera,G. and Castresana,J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.*, **56**, 564–577.
33. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
34. Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
35. Davis,J.J., Gerdes,S., Olsen,G.J., Olson,R., Pusch,G.D., Shukla,M., Vonstein,V., Wattam,A.R. and Yoo,H. (2016) PATtyFams: Protein families for the microbial genomes in the PATRIC database. *Front. Microbiol.*, **7**, 118.
36. Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F. and Wilczynski,B. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
37. Felsenstein,J. (1993) *PHYLIP (Phylogeny Inference Package), Version 3.5 c*. Joseph Felsenstein, Seattle, Washington.
38. Stamatakis,A., Hoover,P. and Rougemont,J. (2008) A rapid bootstrap algorithm for the RAxML web servers. *Syst. Biol.*, **57**, 758–771.
39. Krueger,F. (2012) Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (28 April 2016, date last accessed).
40. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.*, **17**, 10–12.
41. Langmead,B., Wilks,C., Antonescu,V. and Charles,R. (2018) Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*, **35**, 421–432.
42. Lassmann,T., Hayashizaki,Y. and Daub,C.O. (2010) SAMStat: monitoring biases in next generation sequencing data. *Bioinformatics*, **27**, 130–131.
43. Darling,A.E., Mau,B. and Perna,N.T. (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*, **5**, e11147.
44. Darling,A.C., Mau,B., Blattner,F.R. and Perna,N.T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.*, **14**, 1394–1403.
45. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.
46. Wood,D.E. and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, R46.
47. Haft,D.H., DiCuccio,M., Badretdin,A., Brover,V., Chetvernin,V., O'Neill,K., Li,W., Chitsaz,F., Derbyshire,M.K. and Gonzales,N.R. (2017) RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.*, **46**, D851–D860.
48. Ondov,B.D., Bergman,N.H. and Phillippy,A.M. (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.
49. Clausen,P.T., Aarestrup,F.M. and Lund,O. (2018) Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics*, **19**, 307.
50. McArthur,A.G., Waglechner,N., Nizam,F., Yan,A., Azad,M.A., Baylay,A.J., Bhullar,K., Canova,M.J., De Pascale,G. and Ejim,L. (2013) The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.*, **57**, 3348–3357.
51. Liu,B., Zheng,D., Jin,Q., Chen,L. and Yang,J. (2018) VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.*, **47**, D687–D692.
52. Overbeek,R., Olson,R., Pusch,G.D., Olsen,G.J., Davis,J.J., Disz,T., Edwards,R.A., Gerdes,S., Parrello,B. and Shukla,M. (2013) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.*, **42**, D206–D214.
53. Federhen,S. (2011) The NCBI taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.