Comparative genomic analysis of bacterial data in BV-BRC: an example exploring antimicrobial resistance.

Alice R. Wattam[1], Nicole Bowers[2,3], Thomas Brettin[2,5], Neal Conrad[2,3], Clark Cucinell[1], James J. Davis[2,3], Allan W. Dickerman[1], Emily M. Dietrich[2,5], Ronald W. Kenyon[1], Dustin Machi[1], Chunhong Mao[1], Marcus Nguyen[2,3], Robert D. Olson[2,3], Ross Overbeek[2,4], Bruce Parrello[2,4], Gordon D. Pusch[4], Maulik Shukla[2,3], Rick L. Stevens[5,6], Veronika Vonstein[4], Andrew S. Warren[1]

Affiliations

1. Biocomplexity Institute, University of Virginia, Charlottesville, VA 22904 USA
2. Consortium for Advanced Science and Engineering, University of Chicago, Chicago, IL 60637, USA.
3. Division of Data Science and Learning, Argonne National Laboratory, Argonne, IL 60439, USA.
4. Fellowship for Interpretation of Genomes, Burr Ridge, IL 60527, USA.
5. Computing, Environment and Life Sciences, Argonne National Laboratory, Argonne, IL 60439, USA.
6. Department of Computer Science, University of Chicago, Chicago, IL 60637, USA

Abstract

As genomic and related data continue to expand, research biologists are often hampered by the computational hurdles required to analyze their data. The National Institute of Allergy and Infectious Diseases (NIAID) established the Bioinformatics Resource Centers (BRC) to assist researchers with their analysis of genome sequence and other omics-related data. Recently, the PAThosystems Resource Integration Center (PATRIC), the Influenza Research Database (IRD) and the Virus Pathogen Database and Analysis Resource (ViPR) BRCs merged to form the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) https://www.bv-brc.org/. The combined BV-BRC leverages the functionality of the original resources for bacterial and viral research communities with a unified data model, enhanced web-based visualization and analysis tools, and bioinformatics services.

Key Words

BV-BRC, PATRIC, IRD, ViPR, bacteria, virus, genomics, antimicrobial resistance.

Introduction

The National Institute of Allergy and Infectious Diseases (NIAID) established the Bioinformatic Resource Center (BRC) program in 2004 with the goal of providing data, bioinformatic tools, and workflows to enhance the research experience. The initial eight centers have coalesced over time into two BRCs, one that supports research on bacterial and viral pathogens (BV-BRC)***[1]*** and one that supports eukaryotic pathogens and invertebrate vectors (VEuPathDB)***[2]***. The BV-BRC was formed in 2019 through a merger of three BRC resources: the PAThosystems Resource Integration Center (PATRIC)***[3]***, the Influenza Research Database (IRD)***[4]*** and the Virus Pathogen Database and Analysis Resource (ViPR)***[5]***. The goal in merging these resources was to retain the functionality provided by all three and extend their existing functionality to the other user communities. The BV-BRC provides a single unified website with a suite of analysis tools from the initial resources, supported by back-end database and computing resources.

Here we describe the new comparative genomic functionality that has resulted from the unification of PATRIC, IRD and ViPR, demonstrating how antimicrobial resistance data can be analyzed in the new resource.

Material

The BV-BRC continues to improve the research experience with improvements and additions to the services offered. A key entry point for a user analyzing their own data is the read set produced by next generation sequencing platforms (Illumina, Oxford Nanopore, or Pacific Biosciences) which serve as input to many services. Other services address the diversity of existing genomes and their functional annotations. Each of these services can be accessed by clicking on the 'Tools & Services' tab at the top of any BV-BRC web page (Figure 1). The taxonomic identity of reads or contigs can be determined using the Taxonomic Classification and Similar Genome Finder services. The same reads can be assayed for the presence of antimicrobial resistance genes and/or virulence factors and examined for read quality or mapped to individual genomes. Genome assembly and annotation have been combined into the Comprehensive Genome Analysis service, and the Metagenomic Binning service can isolate high quality genomes from mixed read samples. Once a genome has been annotated within the resource, the Phylogenetic Tree and Gene Tree services can be used to explore the phylogeny of whole genomes or individual proteins. Protein families and their functionality can be examined across hundreds of genomes using the Comparative Systems service. Single nucleotide polymorphisms (SNPs) can be examined using the Variation and MSA and SNP services. An alignment service for whole genomes has also been included.

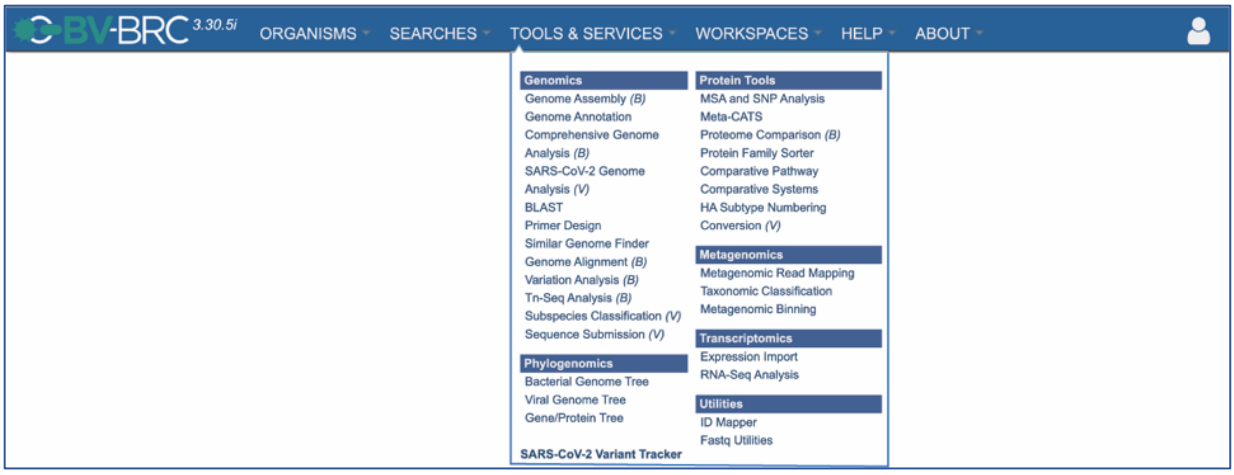

Figure 1. Tools and Services available at BV-BRC.

To demonstrate a potential workflow for research biologists, we will follow a series of publications***[6-8]*** that explore antimicrobial resistance. To demonstrate the type of analysis workflow that can be used in BV-BRC, we will analyze SRA run accession ERR7916262. This is a metagenomic sample collected from a suspected infection and aseptic failure following a joint replacement***[6]***. The steps described below are all available in the publicly available folder used for workshops on antimicrobial resistance (BV-BRC Workshop/ AMR Workshop).

Methods

1. FastQ Utilities – Quality and Trimming

Understanding the quality of fastq reads that come from the sequencer is an essential first step to any of the BV-BRC services that uses them (Assembly, Comprehensive Genome Analysis, Taxonomic classification, Metagenomic read mapping, Metagenomic binning, Variation, RNA-Seq, TN-seq, and Similar Genome Finder). The Fastq Utilities Service provides the capability for aligning, measuring base call quality, and trimming fastq read files to estimate quality. Researchers can submit fastq reads (paired- or single-end, long or short, zipped or not) to the service, as well as Sequence Read Archive run numbers. The four components (trim, paired_filter, fastqc, and align) can be used independently, or in any combination. The pipelines are initiated in the order that they are selected. The Trim component of Fastq Utilities service uses Trim Galore***[9]***, which is Perl wrapper around the Cutadapt***[10]*** and FastQC***[11]*** tools. Paired filter uses Fastq-Pair***[12]***.

Before a significant analysis effort is undertaken, a researcher might want to verify the quality of the reads in the sample, or even massage the data to remove the potential contamination seen in the taxonomic classification analysis. FastQC analysis at BV-BRC produces a fastqc report of the quality of the reads. The FastQC report of ERR7916262 showed good quality except at the end of the reads, which resolved after trimming (Figure 2).

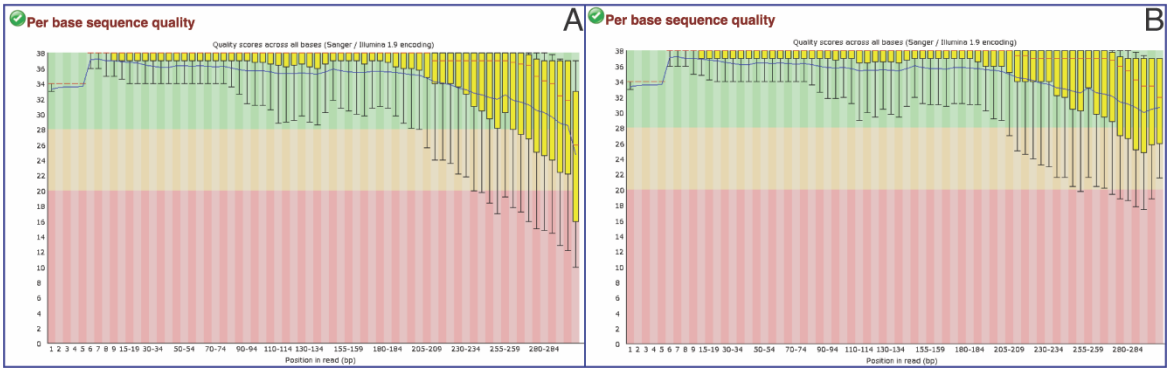

Figure 2. The per sequence quality of the reads from ERR7916262, which is part of the FastQC quality report at BV-BRC. A. FastQC quality report before trimming. B. FastQC quality report after trimming.

2. Taxonomic Classification

The BV-BRC Taxonomic Classification service will identify the microbial composition of metagenomic samples and identify individual isolates. As mentioned for FastQ Utilities above, researchers can submit metagenomic samples as fastq files of reads in a variety of options to the service, using their own data or SRA run numbers; the service will fetch reads from SRA automatically. This service uses Kraken 2***[13]*** that uses exact-match database queries of k-mers. Sequences are classified by querying the database for each k-mer in a sequence, and then using the resulting set of lowest common ancestor (LCA) taxa to determine an appropriate label for the sequence. The service returns a downloadable and viewable taxonomic report that shows the Kraken 2 standard output format, a Krona***[14]*** image, and a text file that contains all taxonomy entries supported by one percent or more of the total hits.

The Taxonomic Classification service returns results in tabular and visual formats, which are also combined into a report that shows the top hits. The visualization of the data is generated by Krona***[14]*** (Figure 3) and different divisions can be visualized by zooming into specific nodes

(Figure 3B). The reads from ERR7916262 mapped predominantly to *Haemophilus influenzae*, which, while an important source of childhood pneumonia and meningitis***[15]***, is also responsible for 2 to 12% of community-acquired pneumoniae in adults***[7]***. It has also been associated with septic arthritis***[16]*** and also with infections in artificial joints***[17-19]***. Evidence that the sample represents a mixed metagenome can be seen in that the reads also mapped to *Streptococcus pneumoniae* and *Homo sapiens* (Taxonomic Classification Report).

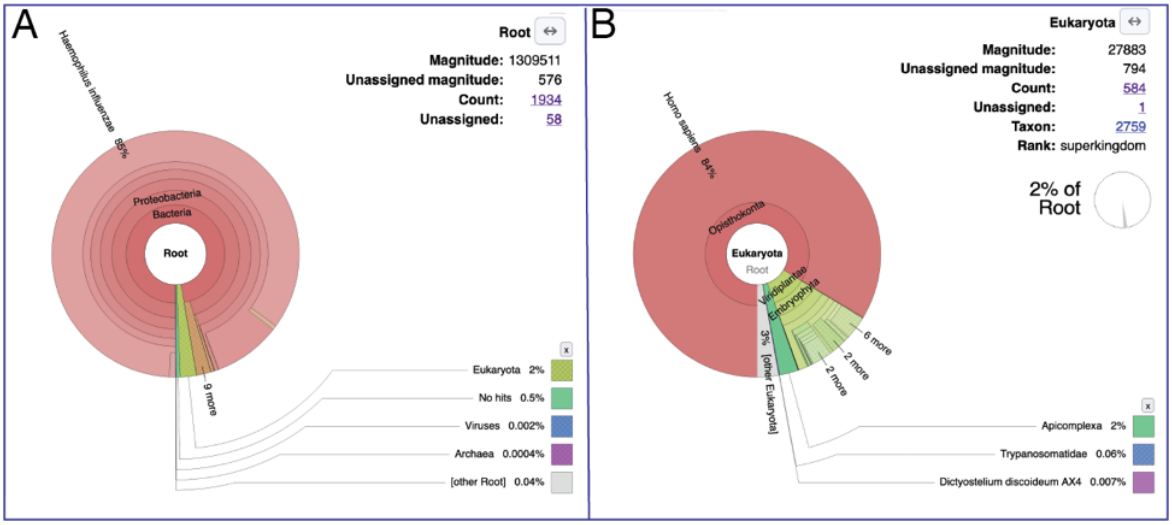

Figure 3. Taxonomic classification analysis, which includes a Krona image of the Kraken 2 results of ERR7916262 reads. A. Unfiltered Krona image. B. Krona image of the reads that map to eukaryotic organisms.

3. Metagenomic Read Mapping

A researcher might want to see if their sample of interest has any sequences that could indicate antimicrobial resistance. The BV-BRC's Metagenomic Read Mapping Service uses KMA***[20]*** to align reads against antibiotic resistance genes or virulence factors. This service accepts reads or a feature group, which is a collection of genes available in the resource that has been selected by the user. It aligns the k-mers to nucleotide sequences from the Comprehensive Antibiotic Resistance Database (CARD)***[21]*** for antimicrobial resistance genes, or the Virulence Factor Database (VFDB)***[22]*** for virulence factors. The service returns a report that includes KMA's standard sample report format, which includes the identifier of the gene from the originating database and the genome in which the gene was described.

Results of the comparison of ERR7916262 reads to the CARD genes are available for viewing at Metagenomic Read Mapping Report. The significant hits include a number of penicillin-binding proteins***[23]***, which contain domains with motifs for the active-site serine penicillin-recognizing enzymes that also includes the class A and C β-lactamases. Among these is NP_439290.1, a *H. influenzae* protein. NP_439290.1 was originally identified as a penicillin-binding protein, but has subsequently been re-annotated at RefSeq (WP_005693446.1) and BV-BRC as a cell division protein FtsI, which plays a part in peptidoglycan biosynthesis***[24]***.

4. FastQ Utilities – Reference Genome Alignment

The Align function of the FastQC Utilities service aligns reads to genomes using Bowtie2***[25]*** to generate BAM files, saving unmapped reads, and generating SAMStat***[26]*** reports of the amount and quality of alignments. The alignment feature of the FastQ Utilities service can be used not

only to access the number of reads that map to a particular genome but can also be used to remove unwanted reads. ERR7916262 showed some human contamination in the Taxonomic Classification analysis. The contaminated reads were removed by aligning the reads to the human genome (genome id 9606.33). This pipeline generates zipped fastq files of reads that mapped, or did not map (designated as unmapped), to the reference genome, as well as a SAMStat report showing the number of the mapped reads (Figure 4).

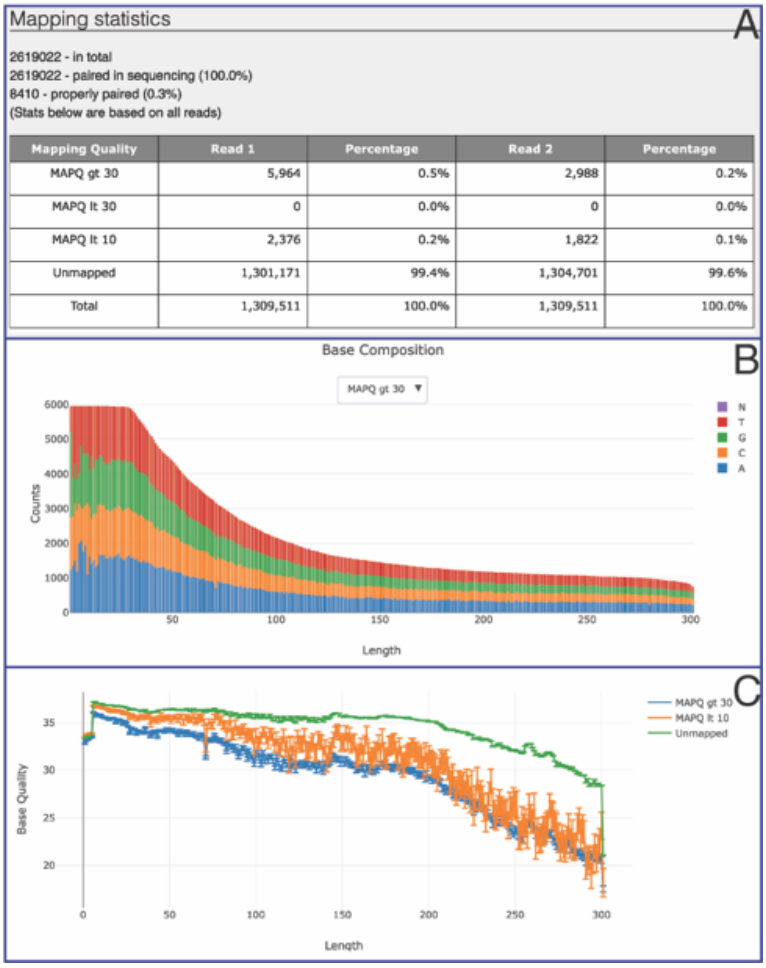

Figure 4. SAMStat report, which is generated by the FastQ Utilities alignment service, showing the number of reads in ERR7916262 that aligned to the human genome, and the number that did not align. A. Mapping statistics show the quality and percentage on a per read basis for a particular set. B. Visualization of the base composition shows the counts and lengths of the reads of the highest quality reads that were aligned to the human genome. C. Visualization showing the quality over length of all reads, including the reads that were unmapped to the human genome.

5. Metagenomic Binning

Once a sample is determined to contain a mixture of bacteria, a researcher might want to see if a genome of good quality could be extracted from the reads in the sample. The Metagenomic Binning service can be used to extract and annotate high-quality, near-complete genomes from reads or metagenomically derived contigs***[27]***. Researchers can submit their metagenomic samples that are reads or contigs to the service, or SRA run numbers to the service.

Researchers can choose to assemble their reads using either MetaSPAdes***[28]*** or MEGAHIT***[29]***. Each set of binned contigs represents a draft genome that will be annotated by an updated

version of RAST called the RAST toolkit (RASTtk)***[30]*** for bacteria, and with VIGOR4***[31, 32]*** or Mat_Peptide***[33]*** for viruses. A structured-language binning report is provided containing quality measurements and taxonomic information about the bins. Contig files and unbinned and unplaced fasta files are also generated. Links to bins are integrated in the BV-BRC database and can be used in any of the downstream services that use genomes.

Submission of ERR7916262 to the Metagenomic Binning service produced one single bin with a genome ranked as good quality, with 100% completeness and 1.5% contamination (Metagenomic Binning Report). Details on the genome can be found in the individual bin results, which are indicated by a checkered flag. Clicking on the row with the flag gives all of the files associated with a genome annotated in BV-BRC, including a hyperlinks to landing pages for the integrated genome in the BV-BRC database and a detailed report (Genome Report) that includes the annotation statistics, as well as a list of genes that appear to be problematic in that there are more, or less, than expected. Hyperlinks for each of these problematic genes are provided as long as they are present in the annotated genome.

6. Similar Genome Finder

While the metagenomic binning service assigns a taxonomic placement to a binned genome, a researcher may want to verify that assignment and identify its closest relatives in the database. BV-BRC provides a service that allows researchers to do this using Mash/MinHash***[34]***, which reduces large sequences and sequence-sets to small, representative sketches, from which global mutation distances can be rapidly estimated. This service accepts reads, contigs, or genomes available in the BV-BRC database.

Following the example of ERR7916262 analysis, examination of the binned genome is available by typing the name (*Haemophilus influenzae* clonal population) or genome ID (727.3208) in the text box for genomes on the Similar Genome Finder landing page. Searches can be made against the NCBI reference and representative genome dataset***[35]*** or all the public genomes in the resource. The search can also be limited to bacterial or viral genomes, or against both at the same time. Submitting the job will return a table that starts with the best hit based on the Mash distance (Figure 5). Individual genomes can be selected and grouped together for downstream analyses comparing them to the annotated genome.

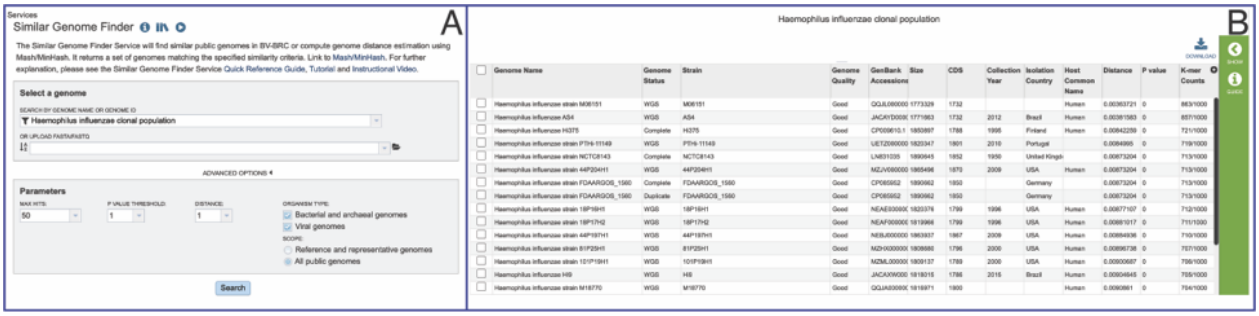

Figure 5. Job submission and result table using the Similar Genome Finder service. A. Job submission interface, where reads or contigs or genome IDs can be uploaded and filtered on the P value or distance. B. Results table showing the genome, its status, quality, GenBank Accession, Size, CDS, collection year, isolation country, host name, Mash distance, P value and k-mer counts.

7. Comprehensive Genome Analysis

One of the most common use cases for analysis of private genomes at BV-BRC is for researchers to assemble and then annotate their genome sequences using two separate services. The streamlined Comprehensive Genome Analysis 'meta-service' addresses this need, computing the assembly and annotation, and providing a user-friendly description of the genome. The output includes a genome quality assessment, AMR genes and phenotype predictions, specialty genes, subsystem overview and a phylogenetic tree. The Comprehensive Genome Analysis Service has quickly risen to be one of the most popular services***[1]***. Researchers can submit their genomic samples that are reads, contigs or SRA run numbers to the service.

Different assembly strategies are provided. Short reads can be assembled using Unicycler***[36]*** or SPAdes***[37]***. Long reads can be assembled with Canu***[38]*** or Flye***[39]***. Hybrid assemblies, which combine short and long reads into a single assembly, can be generated with either Unicycler or Canu. Other parts of the SPAdes package includes assemblies for metagenomic samples (MetaSPAdes***[28]***), plasmids (plasmidSPAdes***[40]***) and reads from single cell sequencing (MDA single cell***[37]***). The assembly service also has options to trim the reads using TrimGalore***[9]***, correct assembly errors (or "polish") using Racon***[41]*** and/or Pilon***[42]***, and also provides the ability to change the minimum length and coverage criteria for saving contigs. Among the files associated with the assembly job are a contigs.fasta file, and an assembly_report.html. This report has details on the assembly and includes a Bandage plot***[43]*** and Quast report***[44]***.

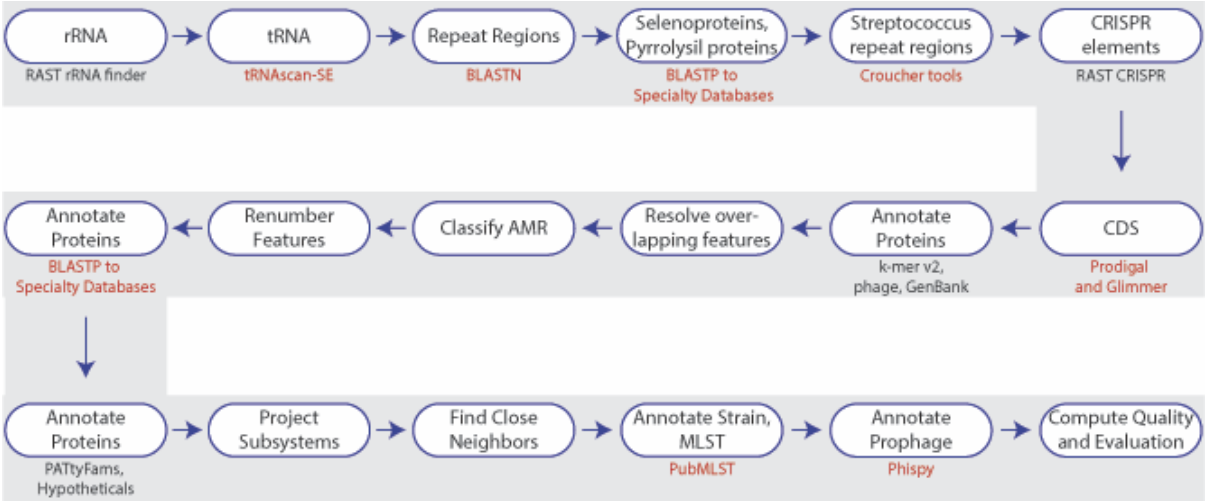

Figure 6. Steps in the RASTtk annotation pipeline which is part of the Comprehensive Genome Analysis service. Red text indicates a component that was not developed by BV-BRC, with black indicating in-house development.

The annotation pipeline uses RASTtk***[30]***. As outlined in Figure 6, an internal tool to RASTtk is first used to call rRNA genes. tRNAscan-SE***[45]*** is used to call the tRNA genes, BLASTN***[46]*** identifies repeat regions within the genome, BLASTP***[47]*** to specialty databases are used to identify seleno- and pyrrolsylproteins, and tools by Croucher***[48]*** are used to identify *Streptococcus* repeat regions. Coding sequences (CDS) are called using Prodigal***[49]***, followed by Glimmer***[50]***. Next, proteins are annotated based on their k-mer signatures, relation to phage and GenBank proteins. Antimicrobial resistance phenotypes are predicted for certain bacterial species using machine learning models***[51]***, followed by an initial protein annotation event that involves taking every protein called in a genome and using BLAT***[52]*** and BLASTP***[47]*** to identify CDSs that have homology to proteins in specialty databases. Genes with homology to

those identified as being involved in antimicrobial resistance are identified by being BLATed against proteins from the Comprehensive Antibiotic Resistance Database (CARD)***[21]***, the National Database of Antibiotic Resistant Organisms (NDARO – https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/), the Antibiotic Resistance Database (ARDB)***[53]*** and a special curation of relevant proteins by BV-BRC curators***[54]***. Possible virulence factors are identified by blasting against a database containing proteins collected from the Virulence Factor Database***[22]***, Violins***[55]***, and a special curation effort by the BV-BRC team***[56].*** Genes with homology to transporters are identified by searching against proteins from the Transporter Classification Database (TCDB)***[57]***, and those similar to genes that have been identified as potential drug targets by comparison to proteins from DrugBank***[58]*** and the Therapeutic Target Database (TTD)***[59]***. Functional annotation and protein families***[60]*** are assigned, and then hypotheticals are identified. All proteins are then mapped to subsystems***[61, 62]*** and PubMLST (www.pubmlst.org) is used to assign sequence types. Among the files associated with the annotation portion of this service are a number of files, including the same GenomeReport.html (described in the Metagenomic Binning service above), and an annotation.contigs.fasta file.

The Comprehensive Genome Analysis service also produces a FullGenomeReport.html, which provides a detailed summary of the genome that begins with a summary of the genome quality scores***[27]***, and then provides information for each step of the service, which includes assembly (if reads were used), annotation, an analysis of specialty genes and functional categories, and a phylogenetic tree of the new genome and its closest high-quality relatives (Figure 7). The newly assembled and annotated genome is integrated into the BV-BRC resource as a private genome where it can not only be viewed by its logged-in owner, but it is also available for any downstream service that uses genomes.

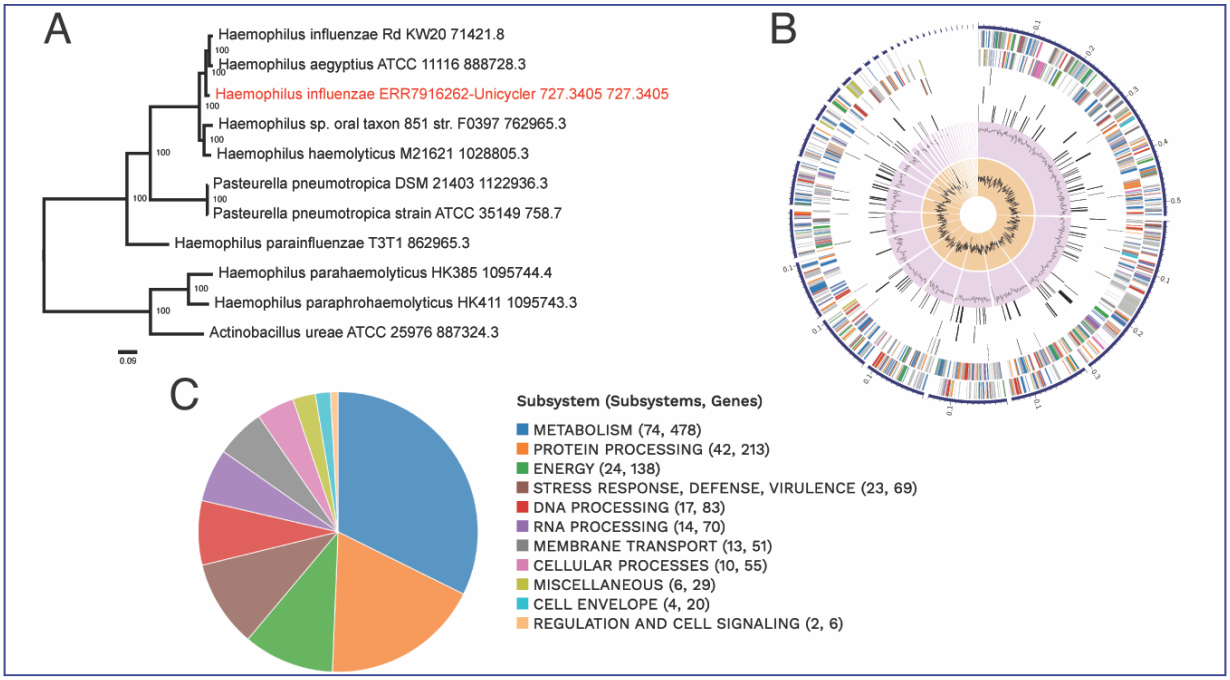

Figure 7. Some visualizations that are part of the Full Genome Report generated by the Comprehensive Genome Analysis service. A. Phylogenetic tree based on the amino acid and nucleotide sequences from five genes, with the private genome shown in red. B. Circular view of the newly annotated genome showing protein-coding genes on the forward and reverse strands, RNA genes,

hits to virulence factors and antimicrobial resistance genes, GC content and GC skew. C. Subsystem analysis of the new genome showing the metabolic profile.

When the reads from ERR7916262 that did not map to the human genome were submitted to the Comprehensive Genome Analysis produced a *H. influenzae* genome ranked as good quality, with 100% completeness and 0% contamination (CGA Genome Report), effectively removing the contamination when the initial reads were submitted to the Metagenomic Binning service described above.

8. Whole Genome Alignment

The Whole Genome Alignment service uses progressiveMauve***[63]*** to align regions conserved in subsets of genomes, which can number from 2 to 20 genomes. This service can be used with any public or private genome in the resource. The service provides a visualization that enables rearrangement of the genome order, zooming into specific areas, links to specific genes and information, and the ability to download an SVG image. It can be used to present overall information, provide suggestions as to contig placement, and identify unique regions.

The exercise using Similar Genome Finder identified the *H. influenzae* M06151 (Genome ID: 727.1771) as the closest public genome to the reads assembled and annotated from the ERR7916262 reads. While any genome can serve as the reference, genomes in fewer contigs produce better images. The M06151 genome has 15 contigs compared to the genome assembled using Unicycler the Comprehensive Genome Analysis service, which is 32 contigs (Figure 8).

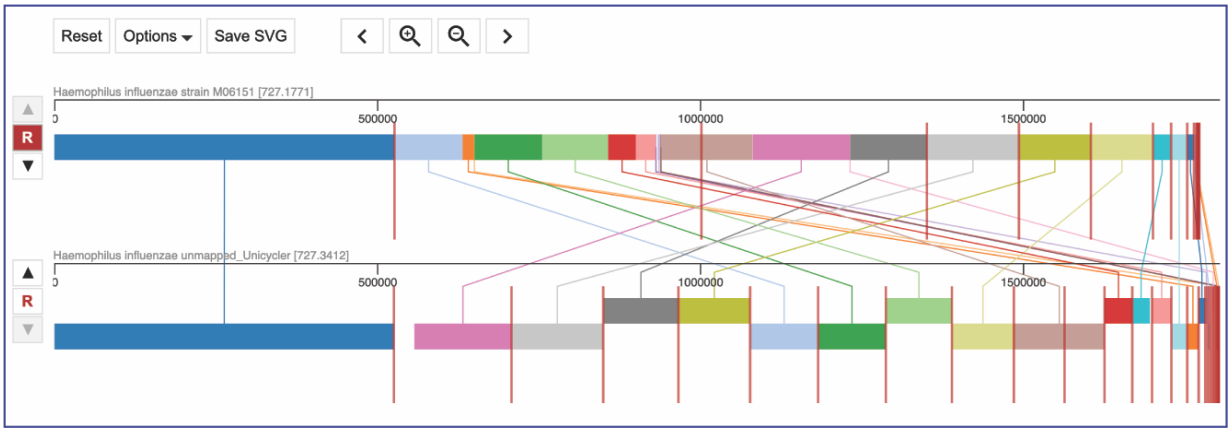

Figure 8. Visualization of the alignment between the H. influenzae M06151 genome and genome assembled from the reads that did not align with the human genome. Red vertical lines indicate contigs and colors indicate shared regions.

9. Bacterial Genome Tree

The Bacterial Genome Tree pipeline generates phylogenies from concatenated alignments of single-copy genes. It selects a user-requested number (10-1000) of the BV-BRC Global Protein Families (PGFams)***[60]*** identified to be single-copy (or within specified tolerances for missing or extra copies). A 50% excess of single-copy families is selected initially, protein sequences aligned using MAFFT***[64]***, and then families are sorted by an alignment quality score favoring lower variability and greater length, and finally the user-requested number is selected from the higher-scoring genes. A sample of each protein alignment (50% by default) is combined into a data file for analysis by RAML***[65]*** to find the optimal protein substitution matrix with the 'PROTCATAUTO' option. Nucleotides are aligned to the aligned amino acids by inserting 3-

base gaps for each gap in the amino-acid alignment. Finally, a combined data matrix is written with both nucleotides and amino acids using a 'partition file' to specify the previously discovered optimal amino acid matrix and the GTRCAT matrix for nucleotides with separate rate parameters for each of the three codon positions. A maximum-likelihood tree is searched for and 100 'rapid' bootstrap support values calculated using the '-f a' option of RAxML. The service returns a reformats the Newick file returned by RAxML into PhyloXML***[66]*** including metadata about the genomes selected by the user upon job submission. Typical metadata fields of interest include taxonomy, host, isolation country and collection year. The tree and associated tip labels can be visualized on the website using the interactive Archaeopteryx Tree Viewer, for which the Tree Viewer Quick Reference Guide***[67]*** provides detailed information about its features and options. This service can be used to explore the phylogenetic relationships of public genomes and the user's private genomes in the BV-BRC resource.

Submitting the private genomes generated by the Metagenomic Binning and Comprehensive Genome Analysis services with the twelve *Haemophilus* reference genomes for the genus to the Phylogenetic Tree service with 100 genes selected produces a tree with strong support values (Figure 9).

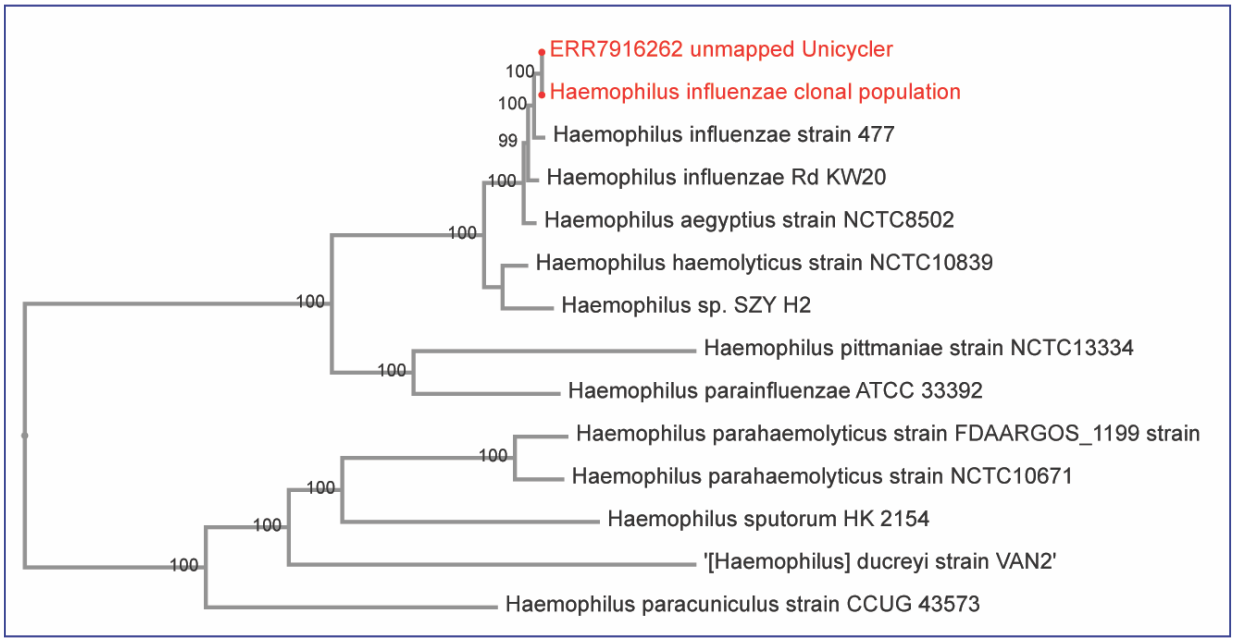

Figure 9. The SVG image that is part of the Bacterial Genome Tree service. Red leaves indicate private genomes.

10. Comparative Systems

The Comparative Systems service at BV-BRC combines two tools, the Protein Family Sorter and the Comparative Pathways Viewer, that were originally part of the PATRIC resource. The new service now includes subsystems***[61, 62]***, a set of functional roles that together implement a specific biological process or structural complex and can also be generalized as pathways. The three independent tools provide summary tables with interactive filters, heatmaps, and other visualizations (Figure 10). Up to 500 genomes can be compared. All three use the BV-BRC protein families (PGFams and PLFams)***[60]*** that are assigned during annotation. The global families (PGFams) can be used for cross-genus comparisons. The local families (PLFams) are for intra-genus comparisons.

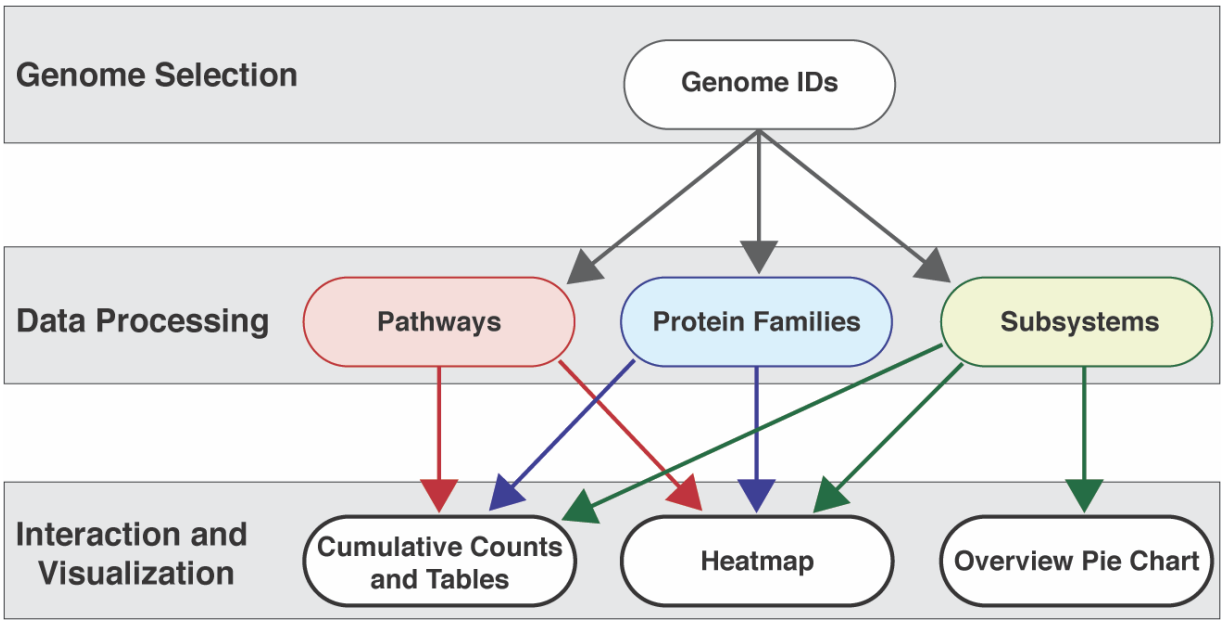

Figure 10.Steps included in the Comparative Systems service.

The Protein Family Sorter allows users to examine the distribution of these protein families across a set of genomes of interest. It initially returns what is commonly referred to as the "pan genome," which in this case refers to the superset of proteins found in all selected genomes. This tool provides various filtering options to quickly locate protein families that are conserved across all the genomes ("core genome"), conserved only in a subset of the selected genomes ("accessory genome"), or that match a specified function. A tabular view shows protein families matching filtering criteria and an interactive heatmap viewer provides a bird's-eye reflecting the filter that shows the distribution of the protein families across multiple genomes, with clustering and anchoring functions to show relative conservation of synteny and identify areas of possible horizontal transfer.

The Comparative Pathway tool analyzes the annotations of a set of genomes to identify metabolic pathways based on taxonomy, EC number, pathway ID, pathway name, and/or specific annotation type. The data are mapped to and summarized on pathway maps from the Kyoto Encyclopedia of Genes and Genomes, commonly known as KEGG***[68]***. This tool also provides a table of unique pathways that match the search criteria (i.e., the genomes or proteins chosen by the researcher, or at any taxonomic level) from which researchers can select specific pathways of interest and view a KEGG Map, or on a heatmap view that summarizes the data, including presence/absence of individual EC numbers within the selected genomes.

Subsystems***[61, 62]***, which was part of the original RAST server***[69]***, are also included in the Comparative Systems service. It includes a distribution pie chart that summarizes the functionality and genes across all genomes selected and also includes a table showing the individual subsystems available. When a subsystem is selected, researchers can see a heatmap showing the genes that are included in it and the presence and absence of those genes across all the genomes in the selection.

Looking at the read set from the orthotic device (ERR7916262) in the Taxonomic Classification and Metagenomic Read Mapping services suggested the presence of a *H. influenzae* genome that had strong BLAT hits to a penicillin-binding protein in the *ftsI* gene. In a summary of genes important in antimicrobial resistance in *H. influenzae**[7]***, Tristam et al. characterize ampicillin-resistant genomes to either produce β lactamase or to lack this enzyme but have variant penicillin-binding proteins, which includes FtsI. There are two types of β lactamase (*bla*$_{TEM}$ and *bla*$_{ROB}$)***[7]***. Some *H. influenzae* lace any β lactamase variants. They are described being β lactamase negative ampicillin-resistant (BLNAR) strains***[8]***.

Searching for *Haemophilus influenzae* in the BV-BRC Global Search, finding Taxa section, and then clicking on the top hit with the most genomes will open the taxa landing page. This shows the combined public and the user's private *H. influenzae* genomes available in the resource. Going to the Features tab for *H. influenzae* will show all the genes across all the genomes in this taxon. The text filter at the top can be used to filter on genes of interest and the protein family assignments can be seen by clicking the small '+' icon in the upper right in the table, and then scrolling down to add the 'PATRIC Local Family' field to the table. Using the text filter to search for the three genes of interest will reveal their PLFam identifiers: 'bla$_{TEM}$' is PLF_724_00004335; *bla*$_{ROB}$ is PLF_724_00012694, and *ftsI* is PLF_724_00000641.

The Comparative Systems service can be used to find the genomes that are BLNAR, and those that have either of the β lactamase genes. Clicking on the AMR Phenotypes tab will open a table and dynamic filter where the genomes can be easily selected based on their AMR metadata. The ampicillin-resistant genomes can be isolated and united into a genome group. A similar group can be selected for the genomes from the ampicillin-susceptible strains.

Submitting the two genome groups (ampicillin-resistant and -susceptible) separately to the Comparative Systems service will summarize the information across all the genomes in protein families, pathways, or subsystems. Clicking on the Protein Families icon at the top of the job results page opens the Protein Family Sorter where the pangenome of the cross-genus or within-genus protein families can be viewed. The within genus families (PLFams) are more applicable as *H. influenzae* is the only taxon examined. The protein family IDs can be used to filter on to show the genes of interest (Figure 11A), either in the table (Figure 11B) or the heatmap view (Figure 11C). In the tabular view, a group for each of the genes in the protein family can be created for exploration in downstream services, and these genes can be viewed by selection from the heatmap. Moreover, a feature group can be created for those *ftsI* genes in the BLNAR genomes can be created to search for specific mutations associated with ampicillin-resistance in downstream services by rearranging the heatmap view and selecting the features of interest.

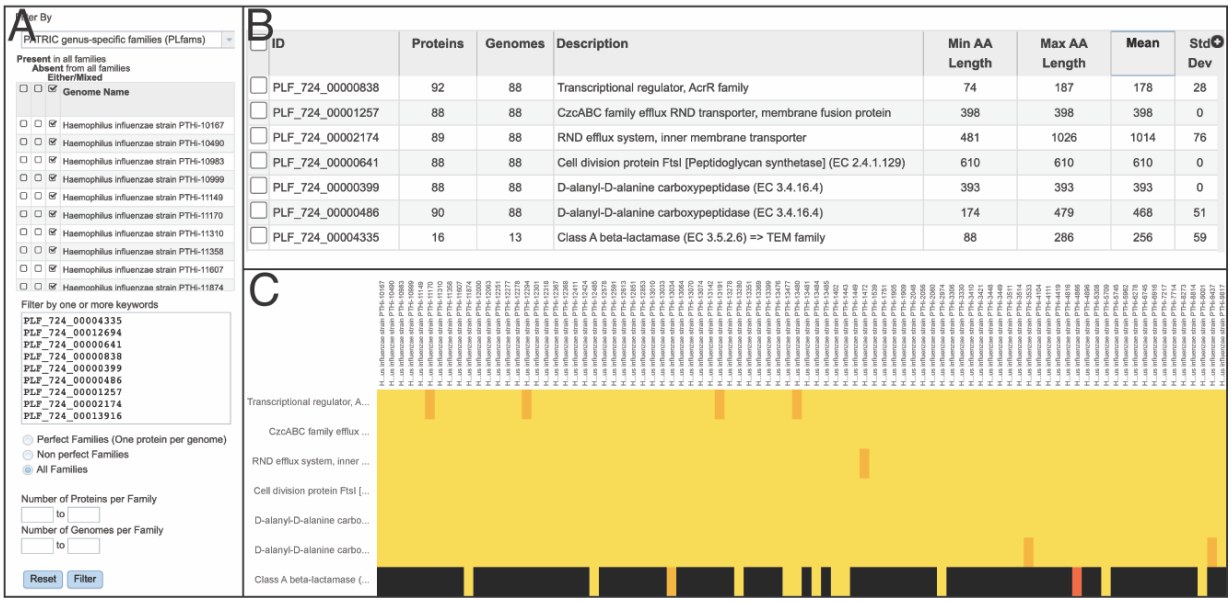

Figure 11. Protein family analysis interface, which is part of the Comparative Systems service. A. The dynamic filter provides the ability to isolate the pan, core or accessory genome, or filter on individual genomes or protein families. B. Tabular view linked to the dynamic filter shows the protein family type and unique identifier, the number of proteins and genomes in that particular family, the minimum and maximum protein length, and the mean and standard deviation based on those lengths. C. The interactive heatmap view of the protein families, linked to the dynamic filter, which shows the proteins in the genomes (y axis) and protein families (x axis).

11. Proteome Comparison

The Proteome Comparison service can be used to readily identify insertions and deletions in up to nine target genomes that are compared with one reference, which can be a genome in the resource (including a researcher's private genome), a genome that has been annotated outside the resource, or a set of proteins that have been saved as a Feature Group in BV-BRC or a fasta file. It is based on the original sequence-based comparison tool that was part of RAST***[62]***. This tool colors each gene based on protein similarity using BLASTP***[47]*** and marks each gene as either unique, a unidirectional best hit or a bidirectional best hit when compared to the reference genome. The output includes a whole-genome schematic that is colored based on BLAST. A table that details all the results can be downloaded for further analysis including coloring that matches the visualization, and hyperlinks in the column with the BV-BRC gene identifiers. A publication-quality Scalable Vector Graphic (SVG) diagram of the results that is publication quality is available for viewing and download.

While the Comparative Systems service shows the presence and absence of protein families, the Proteome Comparison service can be used to examine and even visualize the level of homology between individual proteins. When the binned metagenome was examined in the Similar Genome Finder service, the closest reference genome was *H. influenzae* Rd KW20 (Genome ID 71421.8). When the binned genome was compared to all public genomes, the M06151 strain (Genome ID 727.1771) shared the most k-mers. The Rd KW20 genome was used as the reference and compared to the genome that was generated from the reads that did not map to the human genome (Fastq Utilities) and were then assembled with Unicycler (Comprehensive Genome Analysis), the binned metagenome (Metagenomic Binning , and the *H. influenzae* M06151 genome. The service returns a circular diagram (Figure 12) where a range of BLASTP hits are shown, from the strongest blast hits (blue) to the weakest (red). A downloadable Excel

file also reflects these colors and shows the query coverage and sequence identity for each individual protein compared to the reference.

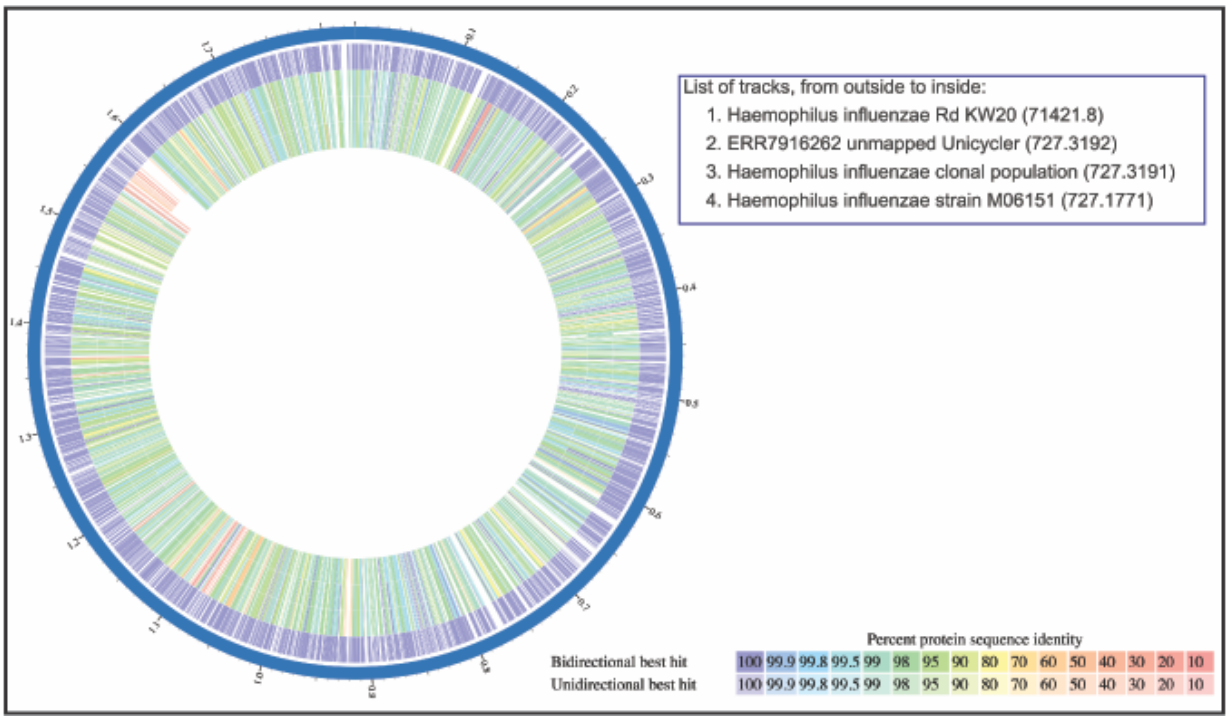

Figure 12. Visualization generated as part of the Proteome Comparison service includes a circular diagram showing the strength of the bidirectional BLASTP hits, the list of the tracks included, and a heatmap legend for interpreting the colors.

12. Compare Region Viewer

The Compare Region Viewer shows all genes in a genomic neighborhood. All genes are shown, but the reference gene (indicated by the red arrow) must be protein-coding. The viewer shows the neighborhood of a gene based on one of several factors: local/global family membership, or a gene neighborhood based on a selected feature group or genome group that was chosen by the user. This viewer also enables selecting the number of regions, and size. The Compare Region Viewer can create a high resolution SVG image that can be exported.

The Proteome Comparison service showed regions of the comparison genomes that had weak BLAST hits (colored red in both the visualization and downloadable Excel file). Clicking on the *H. influenzae* Rd KW20 gene fig|71421.8.peg.1052 in the Excel file, which corresponds to the RefSeq locus tag HI1008 and has the assigned function of DNA uptake protein and related DNA-binding proteins, opens the feature table in BV-BRC. Clicking on the Compare Region Viewer tab will show the gene neighborhood surrounding that gene. Filtering on the genus-specific protein family (PLFam) and selecting a previously created genome group that includes the M06151 public genome that was assembled from the reads that did not map to the human genome and binned metagenome shows the neighborhood across this select group (Figure 13). The 5' region above the pin (red arrow corresponding to fig|71421.8.peg.1052) is conserved, with the arrows having a similar color and number indicating that they belong to the same protein family. The 3' region of the Rd KW20 genome is unique, with different proteins. This corresponds to what was seen in the Proteome Comparison. The same 3' region is conserved across the three comparison genomes.

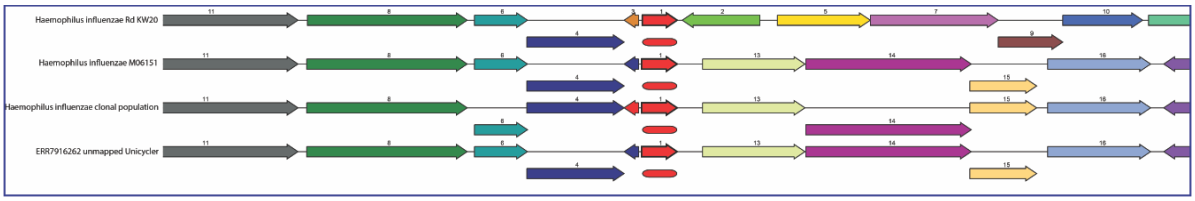

Figure 13. A SVG image of the Compare Region Viewer display of a gene within the H. influenzae Rd KW20 genome compared to two private genomes and the M06151, showing the conserved upstream and diverse downstream regions.

13. Gene/Protein Tree

The Gene/Protein Tree Service enables construction of custom phylogenetic trees built from user-selected genes or proteins. Trees can be built based on either the nucleotide or protein sequences of genomic features from bacteria or viruses. A choice of three tree-inference programs is provided: RaxML***[65]***, PhyML***[70]*** and FastTree***[71]***. The latter choice is best for datasets containing more than 100 very long sequences or over 1,000 small or medium length sequences. The service returns trees in Newick and PhyloXML formats, the latter embedding various metadata fields, which can be rendered in the interactive Archaeopteryx.js Tree Viewer or downloaded and viewed in other software.

The Metagenomic Read Mapping service showed that the reads from ERR7916262 had significant BLAT hits to a penicillin-binding protein that is known in BV-BRC and GenBank as FtsI. This peptidoglycan synthetase that plays a role in resistance to β-lactam antibiotics like ampicillin.***[7]*** The FtsI protein consists of a cytoplasmic domain, a membrane spanning segment, and a periplasmic domain that has transpeptidase activity and is involved in the synthesis of peptidoglycan in the septum of a dividing cell (septal peptidoglycan synthesis)***[8]***. The *ftsI* genes were selected using the text filter on the *Haemophilus* Feature tab and were grouped together. This feature group was submitted to the Gene/Protein Tree service (Figure 14A).

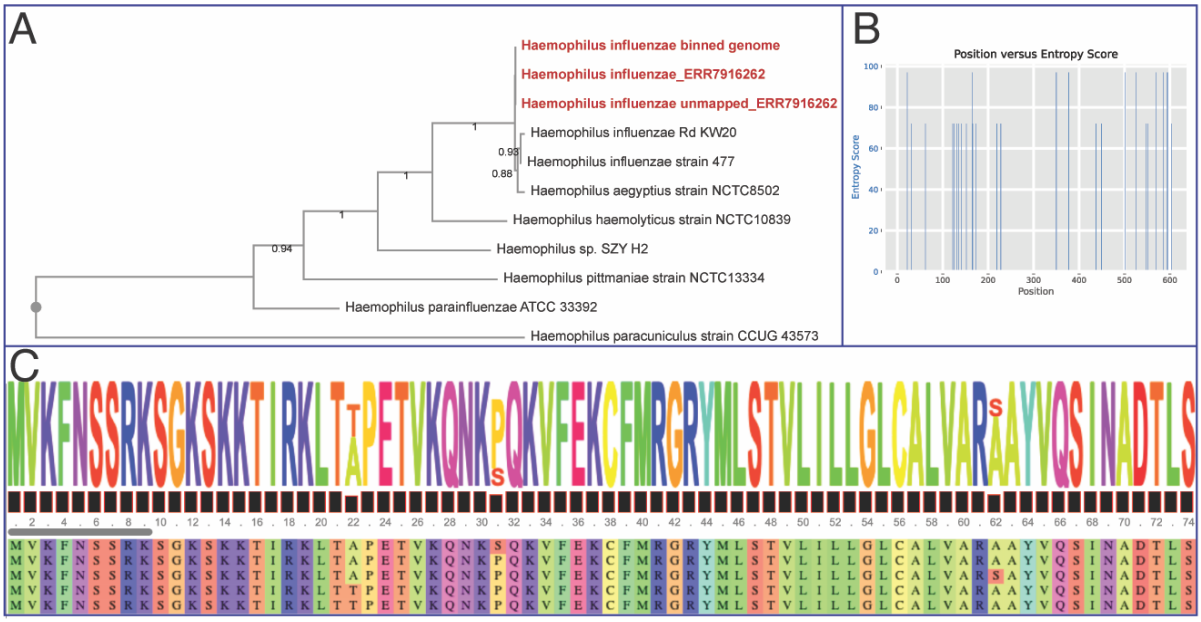

Figure 14. Visualizations that are part of the Gene/Protein Tree and MSA and SNP Variation services. A. SVG image of the FtsI protein from the private genomes (red text) compared to the same protein in selected public genomes (black text). B. The

entropy score, which is generated by the MSA and SNP service, shows the entropy score per position along the FtsI protein. C. Visualization showing SNP mutations in the FtsI protein.

14. Multiple Sequence Alignment and SNP / Variation Analysis

The Multiple Sequence Alignment (MSA) and Single Nucleotide Polymorphism (SNP) Variation Analysis service allows users to choose an alignment algorithm to align sequences selected from a search result, a FASTA file saved to the workspace, or through simply by cutting and pasting. The service can also be used for variation and SNP analysis with feature groups, FASTA files, aligned FASTA files, and user input FASTA records. If a single alignment file is given, then only the variation analysis is run. If multiple inputs are given, the program concatenates all sequence records and aligns them. If a mixture of protein and nucleotides are given, then nucleotides are converted to proteins. Three aligners are provided that include MAFFT***[64]*** and MUSCLE***[72]***. A gene tree is generated using FastTree***[71]***. The service returns a consensus fasta, PNG, and SVG files showing the entropy score plotted against the sequence position, and a SNP tab-separated values (TSV) file. All of the returned files can be downloaded or viewed within the resource.

BLNAR genomes have been shown to have specific SNP mutations***[8]***, including a change from a lysine (K) to asparagine (N) at position 526. A feature group of these specific BLNAR FtsI proteins generated during the Comparative Services exercises shows which of the genomes have that specific mutation (Figure 14 BC).

15. Variation

The BV-BRC Variation Analysis Service can be used to identify and annotate sequence variations in read files compared to a reference genome, including SNPs, SNVs, and indels. The service offers several types of aligners (BWAmem***[73]***, Bowtie2***[25]***, Minimap2***[74]*** and LAST***[75]***) and two types of SNP callers (FreeBayes***[76]*** and BCFTools***[77]***). SnpEff***[78]*** is used to categorizes the effects of variants in genome sequences, annotating variants based on their genomic locations and predicts coding effects. Coding effects such as synonymous or non-synonymous amino acid replacement, start codon gains or losses, stop codon gains or losses, or frame shifts are also predicted.

The service provides a Binary Alignment Map (BAM) and snpEFF.vcf files that can be downloaded or viewed in the BV-BRC Genome Browser***[79]*** (Figure 15). A TSV file for each of the loaded read libraries is also provided, showing the summaries of the locations of the variants, the nucleotide change, and indication of a synonymous or nonsynonymous substitution, or an indel. It can be downloaded or viewed directly in BV-BRC.

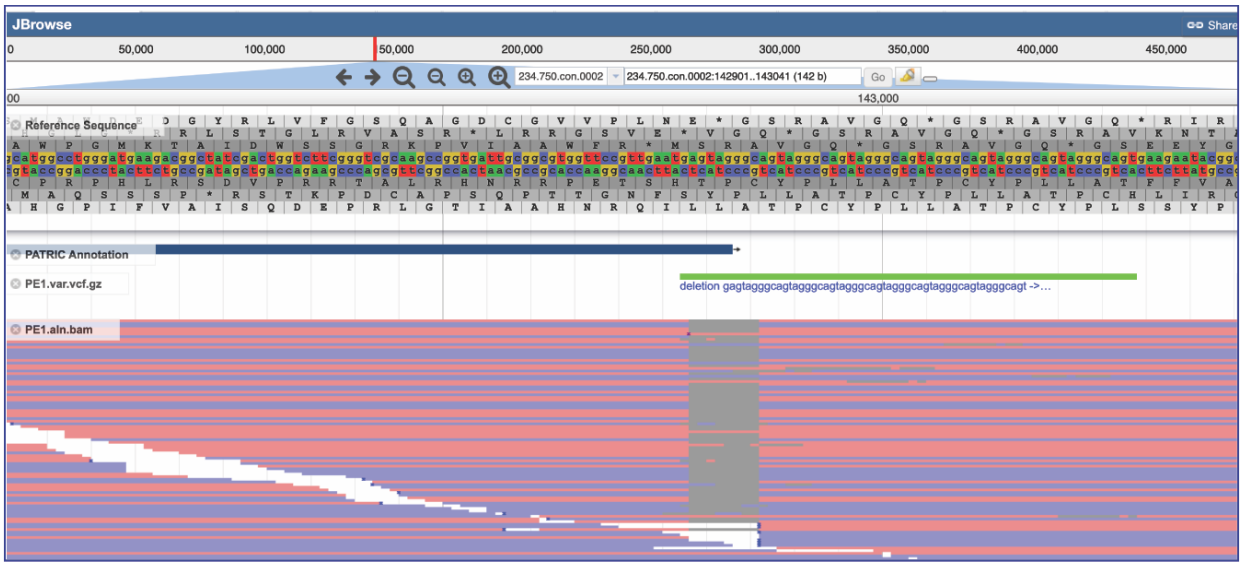

Figure 15. Genome browser view, which is linked to results from the Variation service, where individual SNPs and read files can be seen and compared to the annotated genes and the six-frame sequence translation.

In the Comparative Systems exercise, the protein family IDs were used to filter on to show the AMR genes of interest, and a feature group was created for the *ftsI* genes in the BLNAR genomes to look for specific mutations associated with ampicillin resistance in this gene. From the feature group, users can navigate to the genomes that have those genes, and also see if there are reads available for selected genomes in BV-BRC by clicking on the "+" icon at the top right of the genome table and selecting SRA Accession from the selection window. Those SRA reads can be used in the Variation service to compare to the genome that was assembled from the ERR7916262 file in the Comprehensive Genome Analysis service. There were six SRA read files available for the six BLNAR genomes. This is another way to confirm the presence of the mutations in the *ftsI* gene associated with ampicillin resistance and examine how many reads support that change. These results are provided in the TSV file and are also available for viewing in the Genome Browser, which can be seen by clicking on the browser icon at the upper right of the Results landing page for the job. Clicking on an individual SNP in the browser opens the SNV report that shows the number of reads that have the change when compared to the reference genome (Figure 16).

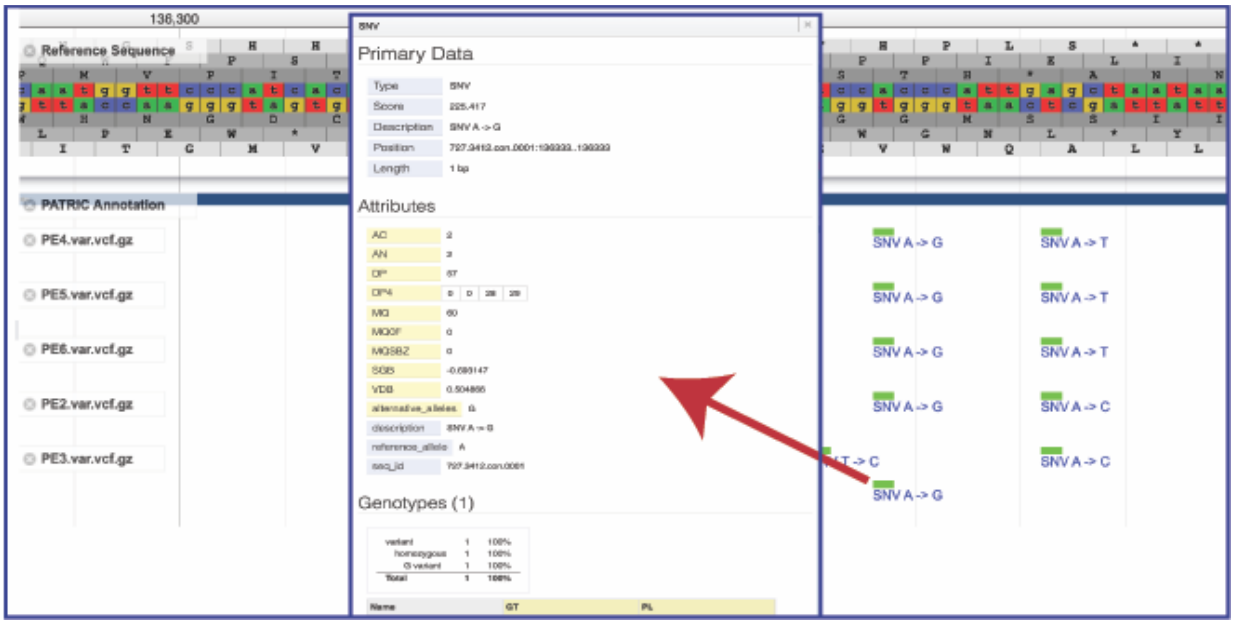

Figure 16. The SNV pop-up that is shown when an individual SNP is clicked on from the browser view. The pop up shows the data associated with the selected SNP.

Conclusion

BV-BRC provides an integrated analysis resource where private data (single, multiple read sets or contigs) can be assembled and annotated, and then compared with any public genome in the resource. An analysis workflow to look for specific mutations associated with antimicrobial resistance demonstrates how the use, interactivity and ease that a researcher is provided with in the resource. The particular example shown above steps through the analysis of a single set of reads, taking them through assembly, annotation, and comparison to other genomes where specific genes are identified and examined for SNPs related to antimicrobial resistance (Figure 17).

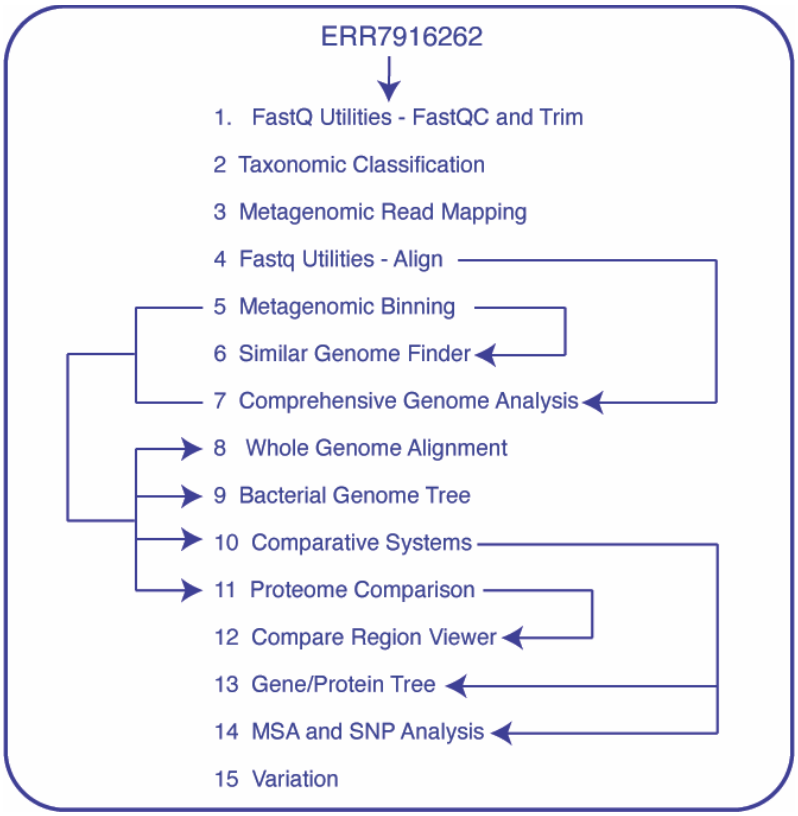

Figure 17. Workflow showing the services used in the analysis of the ERR7916262 read pair. The numbers correspond to sections of this manuscript.

References

1. Olson, R.D., et al., *Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR.* Nucleic acids research, 2023. **51**(D1): p. D678-D689.
2. Amos, B., et al., *VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center.* Nucleic Acids Research, 2022. **50**(D1): p. D898-D911.
3. Davis, J.J., et al., *The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities.* Nucleic acids research, 2020. **48**(D1): p. D606-D612.
4. Zhang, Y., et al., *Influenza Research Database: An integrated bioinformatics resource for influenza virus research.* Nucleic acids research, 2017. **45**(D1): p. D466-D474.
5. Pickett, B.E., et al., *ViPR: an open bioinformatics database and analysis resource for virology research.* Nucleic acids research, 2012. **40**(D1): p. D593-D598.
6. Street, T.L., et al., *Clinical metagenomic sequencing for species identification and antimicrobial resistance prediction in orthopedic device infection.* Journal of Clinical Microbiology, 2022. **60**(4): p. e02156-21.
7. Tristram, S., M.R. Jacobs, and P.C. Appelbaum, *Antimicrobial resistance in Haemophilus influenzae.* Clinical microbiology reviews, 2007. **20**(2): p. 368-389.
8. Ubukata, K., et al., *Association of amino acid substitutions in penicillin-binding protein 3 with β-lactam resistance in β-lactamase-negative ampicillin-resistant Haemophilus influenzae.* Antimicrobial agents and chemotherapy, 2001. **45**(6): p. 1693-1699.

9. Krueger, F., *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.* URL http://www. bioinformatics. babraham. ac. uk/projects/trim_galore/.(Date of access: 28/04/2016), 2012.
10. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet. journal, 2011. **17**(1): p. 10-12.
11. Andrews, S., *FastQC: a quality control tool for high throughput sequence data*. 2010.
12. Edwards, J.A. and R.A. Edwards, *Fastq-pair: efficient synchronization of paired-end fastq files.* bioRxiv, 2019: p. 552885.
13. Wood, D.E., J. Lu, and B. Langmead, *Improved metagenomic analysis with Kraken 2.* Genome biology, 2019. **20**(1): p. 257.
14. Ondov, B.D., N.H. Bergman, and A.M. Phillippy, *Interactive metagenomic visualization in a Web browser.* BMC bioinformatics, 2011. **12**(1): p. 385.
15. Watt, J.P., et al., *Burden of disease caused by Haemophilus influenzae type b in children younger than 5 years: global estimates.* The Lancet, 2009. **374**(9693): p. 903-911.
16. Khattak, Z.E. and F. Anjum, *Haemophilus influenzae*, in *StatPearls [Internet]*. 2022, StatPearls Publishing.
17. Khan, S. and S. Reddy, *Haemophilus influenzae infection of a prosthetic knee joint in a patient with CLL: a vaccine preventable disease.* Case Reports, 2013. **2013**: p. bcr2013010307.
18. Bezwada, H.P., D.G. Nazarian, and R.E. Booth Jr, *Haemophilus influenza infection complicating a total knee arthroplasty.* Clinical Orthopaedics and Related Research (1976-2007), 2002. **402**: p. 202-205.
19. Cichos, K.H., et al., *Efficacy of intraoperative antiseptic techniques in the prevention of periprosthetic joint infection: superiority of betadine.* The Journal of Arthroplasty, 2019. **34**(7): p. S312-S318.
20. Clausen, P.T., F.M. Aarestrup, and O. Lund, *Rapid and precise alignment of raw reads against redundant databases with KMA.* BMC bioinformatics, 2018. **19**(1): p. 307.
21. Alcock, B.P., et al., *CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database.* Nucleic acids research, 2020. **48**(D1): p. D517-D525.
22. Liu, B., et al., *VFDB 2019: a comparative pathogenomic platform with an interactive web interface.* Nucleic acids research, 2019. **47**(D1): p. D687-D692.
23. Zapun, A., C. Contreras-Martel, and T. Vernet, *Penicillin-binding proteins and β-lactam resistance.* FEMS microbiology reviews, 2008. **32**(2): p. 361-385.
24. Sauvage, E., et al., *The penicillin-binding proteins: structure and role in peptidoglycan biosynthesis.* FEMS Microbiology Reviews, 2008. **32**(2): p. 234-258.
25. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature methods, 2012. **9**(4): p. 357.
26. Lassmann, T., Y. Hayashizaki, and C.O. Daub, *SAMStat: monitoring biases in next generation sequencing data.* Bioinformatics, 2010. **27**(1): p. 130-131.
27. Parrello, B., et al., *A machine learning-based service for estimating quality of genomes using PATRIC.* BMC bioinformatics, 2019. **20**(1): p. 1-9.
28. Nurk, S., et al., *metaSPAdes: a new versatile metagenomic assembler.* Genome research, 2017. **27**(5): p. 824-834.

29. Li, D., et al., *MEGAHIT v1. 0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices.* Methods, 2016. **102**: p. 3-11.
30. Brettin, T., et al., *RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes.* Scientific reports, 2015. **5**: p. 8365.
31. Wang, S., J.P. Sundaram, and D. Spiro, *VIGOR, an annotation program for small viral genomes.* BMC bioinformatics, 2010. **11**(1): p. 1-10.
32. Wang, S., J.P. Sundaram, and T.B. Stockwell, *VIGOR extended to annotate genomes for additional 12 different viruses.* Nucleic acids research, 2012. **40**(W1): p. W186-W192.
33. Larsen, C.N., et al., *Mat_peptide: comprehensive annotation of mature peptides from polyproteins in five virus families.* Bioinformatics, 2020. **36**(5): p. 1627-1628.
34. Ondov, B.D., et al., *Mash: fast genome and metagenome distance estimation using MinHash.* Genome biology, 2016. **17**(1): p. 132.
35. RefSeq. *Prokaryotic RefSeq Genomes*. Available from: https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/.
36. Wick, R.R., et al., *Unicycler: resolving bacterial genome assemblies from short and long sequencing reads.* PLoS computational biology, 2017. **13**(6): p. e1005595.
37. Bankevich, A., et al., *SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.* Journal of computational biology, 2012. **19**(5): p. 455-477.
38. Koren, S., et al., *Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.* Genome research, 2017. **27**(5): p. 722-736.
39. Kolmogorov, M., et al., *Assembly of long, error-prone reads using repeat graphs.* Nature biotechnology, 2019. **37**(5): p. 540-546.
40. Antipov, D., et al., *plasmidSPAdes: assembling plasmids from whole genome sequencing data.* bioRxiv, 2016: p. 048942.
41. Vaser, R., et al., *Fast and accurate de novo genome assembly from long uncorrected reads.* Genome research, 2017. **27**(5): p. 737-746.
42. Walker, B.J., et al., *Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement.* PloS one, 2014. **9**(11): p. e112963.
43. Wick, R.R., et al., *Bandage: interactive visualization of de novo genome assemblies.* Bioinformatics, 2015. **31**(20): p. 3350-3352.
44. Gurevich, A., et al., *QUAST: quality assessment tool for genome assemblies.* Bioinformatics, 2013. **29**(8): p. 1072-1075.
45. Lowe, T.M. and S.R. Eddy, *tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.* Nucleic acids research, 1997. **25**(5): p. 955-964.
46. Ye, J., S. McGinnis, and T.L. Madden, *BLAST: improvements for better sequence analysis.* Nucleic acids research, 2006. **34**(suppl_2): p. W6-W9.
47. Johnson, M., et al., *NCBI BLAST: a better web interface.* Nucleic acids research, 2008. **36**(suppl_2): p. W5-W9.
48. Croucher, N.J., et al., *Identification, variation and transcription of pneumococcal repeat sequences.* BMC genomics, 2011. **12**(1): p. 1-13.
49. Hyatt, D., et al., *Prodigal: prokaryotic gene recognition and translation initiation site identification.* BMC bioinformatics, 2010. **11**(1): p. 1-11.
50. Delcher, A.L., et al., *Identifying bacterial genes and endosymbiont DNA with Glimmer.* Bioinformatics, 2007. **23**(6): p. 673-679.

51. Davis, J.J., et al., *Antimicrobial resistance prediction in PATRIC and RAST.* Scientific reports, 2016. **6**: p. 27930.
52. Kent, W.J., *BLAT—the BLAST-like alignment tool.* Genome research, 2002. **12**(4): p. 656-664.
53. Liu, B. and M. Pop, *ARDB—antibiotic resistance genes database.* Nucleic acids research, 2009. **37**(suppl_1): p. D443-D447.
54. Antonopoulos, D.A., et al., *PATRIC as a unique resource for studying antimicrobial resistance.* Briefings in bioinformatics, 2017.
55. Xiang, Z., et al., *VIOLIN: vaccine investigation and online information network.* Nucleic acids research, 2007. **36**(suppl_1): p. D923-D928.
56. Mao, C., et al., *Curation, integration and visualization of bacterial virulence factors in PATRIC.* Bioinformatics, 2015. **31**(2): p. 252-258.
57. Saier Jr, M.H., et al., *The transporter classification database (TCDB): recent advances.* Nucleic acids research, 2016. **44**(D1): p. D372-D379.
58. Wishart, D.S., et al., *DrugBank 5.0: a major update to the DrugBank database for 2018.* Nucleic acids research, 2018. **46**(D1): p. D1074-D1082.
59. Chen, X., Z.L. Ji, and Y.Z. Chen, *TTD: therapeutic target database.* Nucleic acids research, 2002. **30**(1): p. 412-415.
60. Davis, J.J., et al., *PATtyFams: Protein families for the microbial genomes in the PATRIC database.* 2016. **7**: p. 118.
61. Overbeek, R., et al., *The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.* Nucleic acids research, 2005. **33**(17): p. 5691-5702.
62. Overbeek, R., et al., *The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).* 2013. **42**(D1): p. D206-D214.
63. Darling, A.E., B. Mau, and N.T. Perna, *progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement.* PloS one, 2010. **5**(6): p. e11147.
64. Katoh, K. and D.M. Standley, *MAFFT multiple sequence alignment software version 7: improvements in performance and usability.* Molecular biology and evolution, 2013. **30**(4): p. 772-780.
65. Stamatakis, A., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics, 2014. **30**(9): p. 1312-1313.
66. Han, M.V. and C.M. Zmasek, *phyloXML: XML for evolutionary biology and comparative genomics.* BMC bioinformatics, 2009. **10**: p. 1-6.
67. Guide, B.-B.U. *Archaeopteryx Tree Viewer*. 2022; Available from: https://www.bv-brc.org/docs/quick_references/services/archaeopteryx.html.
68. Kanehisa, M., et al., *KEGG for taxonomy-based analysis of pathways and genomes.* Nucleic Acids Research, 2022.
69. Aziz, R.K., et al., *The RAST Server: rapid annotations using subsystems technology.* BMC genomics, 2008. **9**(1): p. 75.
70. Guindon, S. and O. Gascuel, *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.* Systematic biology, 2003. **52**(5): p. 696-704.
71. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree 2–approximately maximum-likelihood trees for large alignments.* PloS one, 2010. **5**(3): p. e9490.
72. Edgar, R.C.J.N.a.r., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* 2004. **32**(5): p. 1792-1797.

73. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows–Wheeler transform.* bioinformatics, 2009. **25**(14): p. 1754-1760.
74. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018. **34**(18): p. 3094-3100.
75. Frith, M.C., M. Hamada, and P. Horton, *Parameters for accurate genome alignment.* BMC bioinformatics, 2010. **11**(1): p. 1-14.
76. Marth, G.T., et al., *A general approach to single-nucleotide polymorphism discovery.* Nature genetics, 1999. **23**(4): p. 452-456.
77. Danecek, P., et al., *Twelve years of SAMtools and BCFtools.* Gigascience, 2021. **10**(2): p. giab008.
78. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* Fly, 2012. **6**(2): p. 80-92.
79. Buels, R., et al., *JBrowse: a dynamic web platform for genome visualization and analysis.* Genome biology, 2016. **17**(1): p. 1-12.

Figure Legends

Figure 1. Tools and Services available at BV-BRC.

Figure 2. The per sequence quality of the reads from ERR7916262, which is part of the FastQC quality report at BV-BRC. A. FastQC quality report before trimming. B. FastQC quality report after trimming.

Figure 3. Taxonomic classification analysis, which includes a Krona image of the Kraken 2 results of ERR7916262 reads. A. Unfiltered Krona image. B. Krona image of the reads that map to eukaryotic organisms.

Figure 4. SAMStat report, which is generated by the FastQ Utilities alignment service, showing the number of reads in ERR7916262 that aligned to the human genome, and the number that did not align. A. Mapping statistics show the quality and percentage on a per read basis for a particular set. B. Visualization of the base composition shows the counts and lengths of the reads of the highest quality reads that were aligned to the human genome. C. Visualization showing the quality over length of all reads, including the reads that were unmapped to the human genome.

Figure 5. Job submission and result table using the Similar Genome Finder service. A. Job submission interface, where reads or contigs or genome IDs can be uploaded and filtered on the P value or distance. B. Results table showing the genome, its status, quality, GenBank Accession, Size, CDS, collection year, isolation country, host name, Mash distance, P value and k-mer counts.

Figure 6. Steps in the RASTtk annotation pipeline which is part of the Comprehensive Genome Analysis service. Red text indicates a component that was not developed by BV-BRC, with black indicating in-house development.

Figure 7. Some visualizations that are part of the Full Genome Report generated by the Comprehensive Genome Analysis service. A. Phylogenetic tree based on the amino acid and nucleotide sequences from five genes, with the private genome shown in red. B. Circular view of the newly annotated genome showing protein-coding genes on the forward and reverse strands, RNA genes, hits to virulence factors and antimicrobial resistance genes, GC content and GC skew. C. Subsystem analysis of the new genome showing the metabolic profile.

Figure 8. Visualization of the alignment between the H. influenzae M06151 genome and genome assembled from the reads that did not align with the human genome. Red vertical lines indicate contigs and colors indicate shared regions.

Figure 9. The SVG image that is part of the Bacterial Genome Tree service. Red leaves indicate private genomes.

Figure 10. Steps included in the Comparative Systems service.

Figure 11. Protein family analysis interface, which is part of the Comparative Systems service. A. The dynamic filter provides the ability to isolate the pan, core or accessory genome, or filter on individual genomes or protein families. B. Tabular view linked to the dynamic filter shows the protein family type and unique identifier, the number of proteins and genomes in that particular family, the minimum and maximum protein length, and the mean and standard deviation based on those lengths. C. The interactive heatmap view of the protein families, linked to the dynamic filter, which shows the proteins in the genomes (y axis) and protein families (x axis).

Figure 12. Visualization generated as part of the Proteome Comparison service includes a circular diagram showing the strength of the bidirectional BLASTP hits, the list of the tracks included, and a heatmap legend for interpreting the colors.

Figure 13. A SVG image of the Compare Region Viewer display of a gene within the *H. influenzae* Rd KW20 genome compared to two private genomes and the M06151, showing the conserved upstream and diverse downstream regions.

Figure 14. Visualizations that are part of the Gene/Protein Tree and MSA and SNP Variation services. A. SVG image of the FtsI protein from the private genomes (red text) compared to the same protein in selected public genomes (black text). B. The entropy score, which is generated by the MSA and SNP service, shows the entropy score per position along the FtsI protein. C. Visualization showing SNP mutations in the FtsI protein.

Figure 15. Genome browser view, which is linked to results from the Variation service, where individual SNPs and read files can be seen and compared to the annotated genes and the six-frame sequence translation.

Figure 16. The SNV pop-up that is shown when an individual SNP is clicked on from the browser view. The pop up shows the data associated with the selected SNP.

Figure 17. Workflow showing the services used in the analysis of the ERR7916262 read pair. The numbers correspond to sections of this manuscript.