AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere

| Computational Biology | Research Article

Beav: a bacterial genome and mobile element annotation pipeline

Jewell M. Jung,[1] **Arafat Rahman,**[1] **Andrea M. Schiffer,**[1] **Alexandra J. Weisberg**[1]

AUTHOR AFFILIATION See affiliation list on p. 13.

ABSTRACT Comprehensive and accurate genome annotation is crucial for inferring the predicted functions of an organism. Numerous tools exist to annotate genes, gene clusters, mobile genetic elements, and other diverse features. However, these tools and pipelines can be difficult to install and run, be specialized for a particular element or feature, or lack annotations for larger elements that provide important genomic context. Integrating results across analyses is also important for understanding gene function. To address these challenges, we present the Beav annotation pipeline. Beav is a command-line tool that automates the annotation of bacterial genome sequences, mobile genetic elements, molecular systems and gene clusters, key regulatory features, and other elements. Beav uses existing tools in addition to custom models, scripts, and databases to annotate diverse elements, systems, and sequence features. Custom databases for plant-associated microbes are incorporated to improve annotation of key virulence and symbiosis genes in agriculturally important pathogens and mutualists. Beav includes an optional *Agrobacterium*-specific pipeline that identifies and classifies oncogenic plasmids and annotates plasmid-specific features. Following the completion of all analyses, annotations are consolidated to produce a single comprehensive output. Finally, Beav generates publication-quality genome and plasmid maps. Beav is on Bioconda and is available for download at https://github.com/weisberglab/beav.

IMPORTANCE Annotation of genome features, such as the presence of genes and their predicted function, or larger loci encoding secretion systems or biosynthetic gene clusters, is necessary for understanding the functions encoded by an organism. Genomes can also host diverse mobile genetic elements, such as integrative and conjugative elements and/or phages, that are often not annotated by existing pipelines. These elements can horizontally mobilize genes encoding for virulence, antimicrobial resistance, or other adaptive functions and alter the phenotype of an organism. We developed a software pipeline, called Beav, that combines new and existing tools for the comprehensive annotation of these and other major features. Existing pipelines often misannotate loci important for virulence or mutualism in plant-associated bacteria. Beav includes custom databases and optional workflows for the improved annotation of plant-associated bacteria. Beav is designed to be easy to install and run, making comprehensive genome annotation broadly available to the research community.

KEYWORDS genomics, annotation, plant-microbe interactions, *Agrobacterium tumefaciens*, mobile genetic elements

Correct and comprehensive genome annotation is critical for characterizing and understanding microbial function and evolution. However, *de novo* annotation of under-studied microorganisms can be challenging. Gene names and proteins may be poorly annotated, and representative sequences of these taxa are often not integrated into databases of existing annotation tools (1). Bacteria with clinical relevance are

Editor Xiyang Dong, Third Institute of Oceanography Ministry of Natural Resources, Xiamen, China

Address correspondence to Alexandra J. Weisberg, Alexandra.Weisberg@oregonstate.edu.

Jewell M. Jung and Arafat Rahman contributed equally to this article. Author order was determined by time contributed to the project.

The authors declare no conflict of interest.

Received 12 March 2024
Accepted 28 June 2024
Published 22 July 2024

overrepresented in gene and genome databases, and important virulence genes from these organisms can be confidently identified and annotated (2). However, annotation of many phytopathogens and plant-associated microbes often fails to identify and name key genes important for plant-microbe interactions, symbiosis, and virulence. For instance, effector proteins secreted by phytopathogens have a fundamental role in the plant disease process, yet they are underrepresented in annotation tools and databases (3). While these key virulence loci may be characterized and classified in species-specific databases and publications, these annotations are often not incorporated into annotation databases or current tools.

Additionally, current whole-genome annotation pipelines typically focus on the annotation and function of individual genes, but often do not report information about their genomic context. Genes may be part of biosynthetic gene clusters or metabolic pathways, be organized into operons, represent loci encoding larger macromolecular structures, or be carried on mobile genetic elements (MGEs) and/or prophage elements. Gene regulatory elements, such as promoters or transcription factor binding sites, are important for understanding how genes are expressed, but these elements are also typically not annotated. Understanding the genomic context and regulation of genes is crucial for understanding their function.

Mobile genetic elements, such as plasmids, integrative and conjugative/mobilizable elements (ICEs/IMEs), integrons, prophages, and other elements can be mobilized from cell to cell and play a major role in the horizontal transfer of genes between bacteria (4). MGEs may be integrated into the chromosome, such as in the case of ICEs and transposons, or they may replicate independently, such as plasmids. MGEs can be incredibly diverse and vary in structure and function. This can make their identification and annotation difficult, especially in draft genome assemblies. The horizontal transfer of MGEs, either directly via conjugation, or indirectly on other elements, enables bacteria to rapidly respond to changes in their environment and acquire new traits that may be selectively advantageous (4–6). For example, plasmids are associated with the movement and transfer of antimicrobial resistance (AMR) genes and genes associated with pathogenicity or mutualist symbioses (7–9). ICEs and IMEs are also recognized as drivers of HGT and have been attributed to the spread of genes involved in AMR, pathogenesis, and symbiosis in diverse microbial taxa (10–12). Integrons are genetic elements that can acquire and shuffle the order of gene "cassettes" relative to a single promoter (13). The order of these genes can be rearranged in response to stress conditions, altering their expression (14). Many integron gene cassettes have been found to encode for traits associated with virulence, resistance, and host-microbe interactions (15).

Bacterial genomes often encode for one or more secretion systems. These secretion systems are diverse multi-protein complexes that have a range of functions, including those that play a fundamental role in the conjugative transfer of MGEs, interbacterial communication and competition, and/or host-microbe interactions (16–18). Conjugation of plasmids and ICEs is typically facilitated by a type IV secretion system (T4SS) (19, 20). Type IV secretion systems sometimes provide other functions beyond conjugation. One of the most well-studied T4SS is in the phytopathogen *Agrobacterium tumefaciens*, where it facilitates the inter-kingdom transfer of DNA (21). Microbial defense systems are also encoded on and transferred via MGEs (22). These diverse systems provide defense against invading DNA, either from phage or plasmids. With an abundance of MGEs in bacteria and genome data, characterizing MGEs and the systems they interact with is essential to answer questions regarding microbial evolution, ecology, and resistance. Despite the importance of MGEs, few tools exist to comprehensively characterize MGEs within bacterial genomes (23).

Comprehensive characterization of specific genetic regions of interest requires multiple genome annotation tools. Many independent tools for the annotation of specific genetic systems have recently been published (24–28). However, the results of these separate analyses must be combined to get a full picture of gene function and context. Additionally, the installation and use of these tools present a challenge, even

for those with computational proficiency. Installation of individual annotation software tools can be laborious and complicated by numerous or conflicting dependencies. This, coupled with unclear or minimal documentation, can limit the accessibility of powerful tools. Moreover, genome analysis with multiple annotation tools also requires manual parsing and cross-correlating numerous output files, which requires experience with the command line.

In response, we present Beav, a command-line tool that streamlines and automates bacterial genome and mobile genetic element annotation. The Beav pipeline incorporates multiple annotation tools, automating the process of running, parsing, and combining results into a single easy-to-read output. The Beav pipeline also includes several tools and databases that enhance the annotation of plant-associated microbes, including genes and regulatory elements important to phytopathogens and mutualist symbionts. Additionally, an optional *Agrobacterium*-specific pipeline identifies the presence of oncogenic Ti and Ri plasmids and classifies them under a published scheme (29, 30). This pipeline also annotates Ti/Ri plasmid-specific regulatory elements and reports the taxonomic classification of the input strain under the *Agrobacterium* biovar/genomospecies scheme (31). Finally, Beav generates a visualization of the position of annotated gene clusters and mobile elements in the genome. Additionally, Beav can generate a separate plot to visualize oncogenic Ti/Ri plasmids, if present.

Beav is a comprehensive genome annotation pipeline for bacteria and associated mobile genetic elements. Beav and its dependencies are available for installation via conda and requires minimal user input for installation and usage. The pipeline uses pre-existing annotation tools in combination with custom scripts and databases to automate annotation and combines results in a single easy-to-read GenBank and/or GFF3 format output. Beav databases and source code are also freely available for download on GitHub at http://github.com/weisberglab/beav.

MATERIALS AND METHODS

Usage of the Beav pipeline

Beav is designed to be user-friendly and includes multiple checks that verify the correct installation of dependencies and valid input arguments before running the pipeline. Beav is a command-line tool written in Python and shell script that runs on Unix-based operating systems such as Linux. The Beav pipeline will clearly indicate if prerequisites are installed correctly, skipped, or causing an error. Each annotation step of the pipeline is listed as it runs and labeled as "Done" when complete. Tables and log files summarizing the output of each annotation program are also produced during the run. The Beav pipeline workflow includes multiple databases and annotation tools (Fig. 1). Beav is packaged in Bioconda for ease of installation (32). Installing Beav via conda will also install all required dependencies in a single environment. A separate program included with Beav will also automatically download and format all databases needed by Beav and its dependencies. Most steps in the workflow are optional and can be skipped. Users input a single file with the nucleotide sequence of their genome assembly in fasta format, along with any other optional parameters. Beav then annotates the genome and proceeds to run other annotation programs. Users can alternatively input a GenBank format annotated genome. In this case, Beav will skip the initial gene annotation step and use these annotations as input to other steps in the pipeline. Finally, the results of each program are parsed, and the initial annotations for each gene are supplemented with information from each of the annotation tools and reported in GenBank and GFF3 format output files. Regions representing mobile genetic elements and gene regulatory elements are also annotated in the final output files. If a Beav run is interrupted or the user wishes to re-run the analysis with additional tools, Beav can also be restarted with the "--continue" option to finish any incomplete analyses.

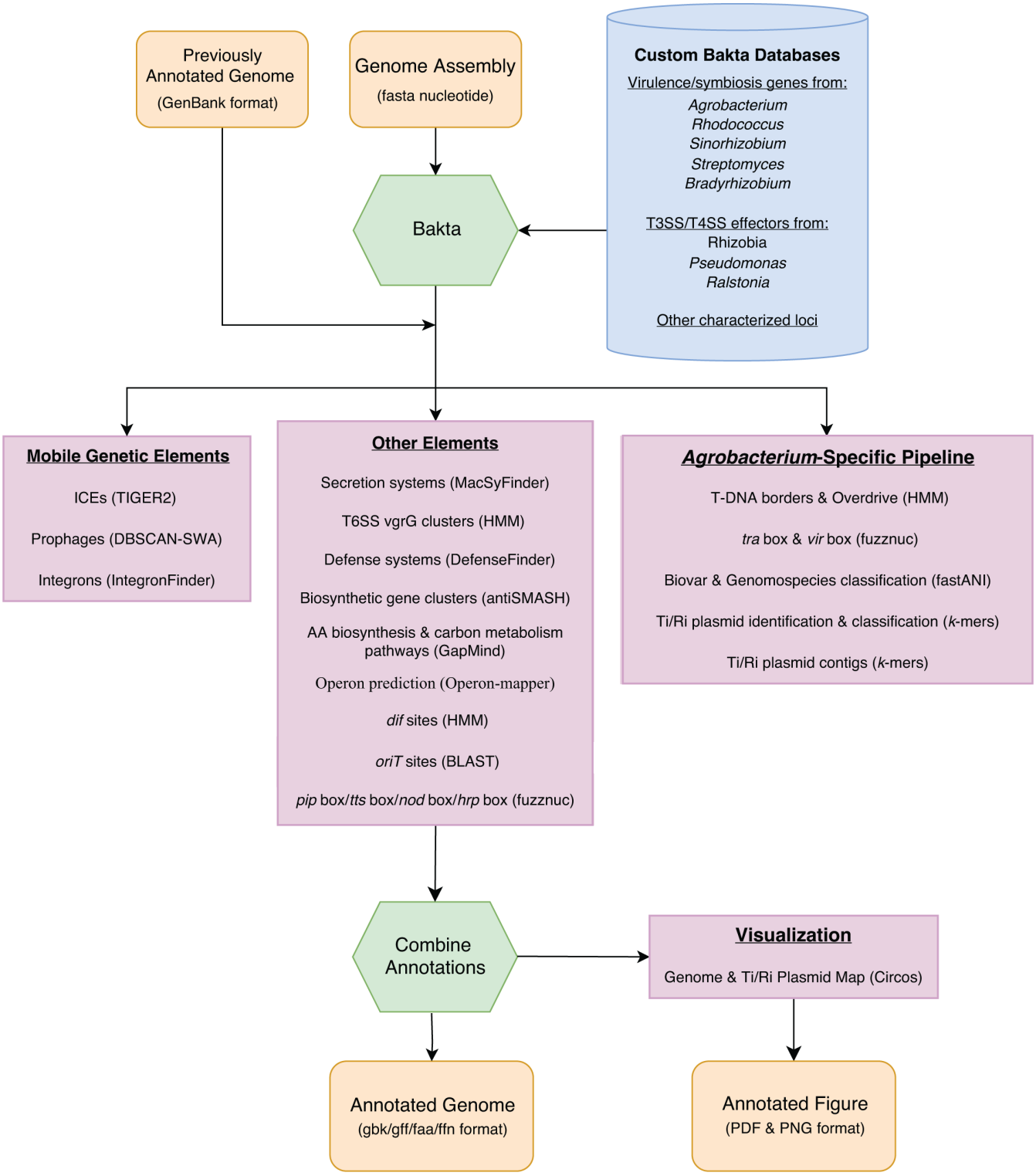

FIG 1 Summary of the Beav pipeline workflow. Beav takes a fasta nucleotide file as input and uses Bakta with a custom gene database to generate preliminary annotations. Alternatively, a previously annotated GenBank file can be used as input. Following this, the pipeline runs several suites of annotation programs to annotate mobile genetic elements and other systems and gene clusters in the genome. An optional *Agrobacterium*-specific pipeline can be run that identifies and classifies Ti/Ri plasmids and annotates features specific to these elements. Finally, annotations are combined and output into standard bioinformatic file formats. Visualizations of annotations across the whole genome as well as Ti/Ri plasmids are generated in raster and vector image formats.

Initial annotation and custom gene databases

Beav takes as input an assembled genome in fasta nucleotide format or an initial annotated genome in GenBank format. If a fasta file is provided as input, Bakta is used for the initial annotation of genes and other loci (33). The Bakta run is supplemented with custom Bakta-formatted gene databases that enrich the annotation of plant-associated microbes. Beav incorporates several custom databases into Bakta to ensure the correct names of virulence and symbiosis genes for numerous phytopathogens and mutualists (Table 1). These databases were compiled from published data sets and online resources (10, 29, 30, 34–38). Databases include known virulence genes and genes encoding secreted effectors from phytopathogenic *A. tumefaciens*, *Pseudomonas syringae*, *Ralstonia solanacearum*, *Rhodococcus fascians*, *Streptomyces scabiei*, and key loci from mutualist *Bradyrhizobium*, *Rhizobium*, and *Ensifer*/*Sinorhizobium*.

Following the Bakta run, the Beav pipeline annotates promoters and binding sites associated with plant-associated microbes, including *pip* box (plant-induced promoter), *tts* box and *hrp* box (regulating genes associated with type III secretion systems), *nod* box (regulating genes associated with nodulation in nitrogen fixation symbiosis), *tra* box (regulation of conjugation loci of Ti/Ri plasmids), and *vir* box (regulation of *Agrobacterium* virulence genes) elements. For several of these elements, the EMBOSS program fuzznuc is used to identify elements based on conserved sequence patterns (40). The *pip* box promoter is identified with the pattern "TTCGBN(15)TTCGB" (41). The *hrp* box promoter is identified with the consensus pattern "GGAAC[CT]N(15,17)CCACNNA" (42–44). The HMMER3 suite program nhmmer is used to search with custom HMM profiles that we have built for other elements from known sequences, including *nod* box and *tts* box regulatory elements, and chromosome *dif* sites (10, 45–52). Bedtools is used to ensure that predicted regulatory sequences do not overlap with coding sequences (53).

Annotation of mobile genetic elements

The Beav pipeline includes several tools to annotate diverse mobile genetic elements in genomes, including ICEs, integrons, and prophage elements. TIGER2 is used to identify and annotate monopartite ICEs integrated into the genome (28). The boundary regions of the identified ICEs, as well as the integrase and predicted target genes and sequences (*attB*, *attP*), are reported in a "mobile_element" feature in the final annotation. Plasmid/ICE origin of transfer (*oriT*) sites are annotated using a database of known *oriT* sequences subset to remove duplicate and very short sequences (54). Blastn with the options "-task blastn-short -outfmt '6 std qlen slen qseq sseq' -dust no -qcov_hsp_perc 20" is used to identify putative *oriT* sequences in the input assembly (55). Blast hits are further filtered to those with an *e*-value of 0.1 or less and a minimum alignment length of 20 bp. DBSCAN-SWA is used to annotate prophages integrated into the genome (27). A mobile_element feature reporting the entire prophage region and classification is also included in the final annotation. IntegronFinder is used to identify and annotate integron

TABLE 1 Custom Bakta annotation databases for plant-associated microbes

Organism	Genes and loci	Reference
Agrobacterium	*vir*, *GALLS*, *acc*, agrocin84 biosynthesis, opine synthases, T-DNA oncogenes, *trb*/*tra*, *upp* attachment cluster	(29, 30, 39)
Bradyrhizobium	*nod*, *nol*/*nop*, *nif*/*fix* symbiosis genes, Type III secreted effectors	(10)
Rhodococcus	*fas*/*att* virulence genes	(34, 35)
Sinorhizobium	*nod*, *nol*/*nop*, *nif*/*fix* symbiosis genes	*Sinorhizobium fredii* HH103 (NCBI: GCA_000283895.1)
Streptomyces	*txt*, *nec1*, *tomA*, *fas* virulence genes	(36)
Ralstonia	Type III secreted effectors	(37)
Pseudomonas	Type III secreted effectors	(38)

loci and cassettes (26). Predicted integron regions are annotated, including the integrase *intI* gene, promoter, and the location of *attC* sequences bordering cassettes.

Annotation of other systems and functional loci

Beav incorporates additional annotation tools that identify other conserved gene clusters or provide further context for gene function. MacSyFinder with the TXSScan models is used to identify genes and gene clusters encoding for diverse secretion systems (24, 56). DefenseFinder is used to characterize the presence of microbial defense systems (57). These systems provide defense against invasion by foreign DNA, including phage and plasmids (58). AntiSMASH is used to identify and annotate biosynthetic gene clusters (25). GapMind is used to associate genes with amino acid biosynthesis pathways as well as genes involved in the catabolism of small carbon metabolites (59, 60). Genes in bacterial genomes are often encoded in and expressed as single transcriptional units called operons (61). Beav can optionally submit jobs to the operon-mapper web server, download completed results, and parse the resulting output (62). Operon tags associating genes to specific predicted operons are added to gene features in the final annotation. If the operon pipeline is run, Beav will also annotate type VI secretion system *vgrG* clusters. These gene clusters encode for the vgrG spike protein as well as adapters, toxins, and cognate immunity genes (63). The HMMER3 program nhmmer is used to search with HMM profiles for vgrG (TIGRFAMs TIGR01646.1 and TIGR03361.1). Operons containing *vgrG* genes are then reported as putative *vgrG* clusters.

Agrobacterium-specific pipeline

In addition to generic annotation tools applicable to all bacteria, Beav includes an optional pipeline for annotating features specific to the phytopathogen and genetic engineering tool *A. tumefaciens*. Using the optional *Agrobacterium* pipeline, *Agrobacterium* genomes can be further annotated with Ti/Ri plasmid-specific loci, including T-DNA borders, overdrive, *vir* box, and *tra* box elements. Fuzznuc with the pattern "RTTDCAWWTGHAAY" is used to annotate the *vir* box virulence gene promoter (64). The Ti/Ri plasmid conjugation loci promoter *tra* box is identified by the consensus pattern "WNGTGMARAWYTGCACDW" (65–67). The HMMER3 suite program nhmmer is used to search with custom HMMs that we developed based on the sequence of known T-DNA border and overdrive sequences (29, 52, 68–70). Beav also automates the identification of oncogenic Ti and Ri plasmid sequences in the input genome assembly and classifies the plasmid type based on a classification scheme (29, 30). The BBTools program compare-sketch.sh is used to identify Ti/Ri plasmid contigs based on a custom database of known oncogenic plasmids (71). Output is reported as the presence of a Ti/Ri plasmid, its classification, and a list of contigs associated with that plasmid. FastANI and a database of representative *Agrobacterium* taxa are used to classify the input *Agrobacterium* strain into biovar and genomospecies-level designations (31, 72). Beav also includes extra scripts to run the *Agrobacterium* analyses independent of the full annotation pipeline.

Parsing and combining annotations

While running multiple annotation tools can be informative, cross-correlating the results of multiple analyses and annotation tools can be challenging. Results are often present in multiple files, and in different formats. Beav solves this issue by automatically parsing the results of each step of the pipeline and incorporating those results into a single output file. A custom Python script and the BioPython library are used to parse the Bakta GenBank output and the results of each step of the pipeline, and add information as new loci or to associated gene features (73). In the GenBank output, mobile genetic elements and prophage are added as "mobile_element" features, and regulatory and other elements are labeled as "misc_feature" elements. Additional information about genes and coding sequences is added as "notes" in the feature qualifiers field of existing features. The annotation software used to make each inference is also listed. Final

annotations are output in GenBank (gbk) and GFF3 formats, and coding sequences are output in fasta nucleotide (ffn) and amino acid (faa) formats. Output files of each tool are organized into directories for easy navigation of results and annotations. This alleviates tedious navigation of multiple output files and keeps results organized. Logs and table format output of each tool are also produced and stored in relevant folders. At the end of a Beav run, all software tools used in the analysis, along with their versions and suggested citations, are listed for ease of reporting results in publications.

Visualization of genome and plasmid annotations

Data visualization is important for understanding large-scale genome structure and organization. However, converting data into the correct format for existing tools and generating publication-quality visualizations can be difficult. Beav automates this process and generates figures of major annotated features and gene clusters across the genome (Fig. 2). The Python package PyCirclize is used to generate Circos plot visualizations (74). Beav generates Circos plots that visualize whole-genome structure, including contigs, the position of secretion system-encoding loci along with their types (i.e., T3SS, T4SS, and T6SS), ICEs, prophage elements, specialized metabolite gene clusters, and tRNA/rRNA position. If Beav is run with the optional *Agrobacterium*-specific pipeline and a Ti or Ri plasmid is detected, a second Circos plot is generated that visualizes Ti/Ri plasmid contigs and associated features (Fig. 3). This plot shows the structure and content of the Ti/Ri plasmid and key loci, including virulence genes, T-DNA regions and borders, plasmid replication (*repABC*) and conjugation (*tra*/*trb*) loci, specialized nutrient (opine) synthase and transport genes, and regulatory elements such as the *tra* box and *vir* box. Beav also includes a supplemental stand-alone command-line tool, beav_circos, to generate these visualizations for other GenBank files.

RESULTS

Comparison of annotations across pipelines

To test the performance and annotation improvements of Beav relative to Bakta alone and the RefSeq PGAP pipelines, we annotated a diverse set of representative bacterial genomes, including *Agrobacterium fabrum* C58 (NCBI: GCA_000092025.1), *Escherichia coli* 131 (NCBI: GCA _005221985.1), *P. syringae* DC3000 (NCBI: GCF_000007805.1), *R. fascians* D188 (NCBI: GCF_001620305.1), and *Aeromonas caviae* WP2-W18-ESBL-01 (NCBI: GCF_014168635.1). We then compared the number of genes annotated as "hypothetical protein" and "uncharacterized protein" by the three pipelines (Table 2). For the Beav pipeline, annotation of the *A. fabrum* C58 genome was run with the optional *Agrobacterium* pipeline, while all other genomes were run with default options. For this test, a gene is considered annotated by Beav if a "note" was added to loci annotated as a hypothetical or uncharacterized protein. Overall, Beav and Bakta predicted a function for more genes than RefSeq PGAP. Beav and Bakta produced comparable numbers of annotated hypothetical proteins for all the genomes, which can be attributed to the Beav pipeline using Bakta for preliminary gene annotations. However, Beav consistently produced additional annotations for genes that Bakta annotated as hypothetical. These results demonstrate that the combined analyses of the Beav pipeline can improve genome annotation over a single tool.

To further assess the usefulness of annotations produced by the Beav pipeline, we summarized the total number of systems, mobile elements, gene clusters, and new features in annotations of diverse bacteria (Table 3). For each test genome, Beav successfully identified the genes of multiple complete biosynthetic gene clusters and defense systems. Further, at least one putative secretion system was identified in each genome. MGE annotations varied based on the analyzed strain. For instance, several ICEs were found in the *A. caviae*, *A. fabrum*, *E. coli*, and *P. syringae* genomes, while zero were found in *R. fascians*. However, these elements might not be present or common in certain bacterial lineages or strains. These results indicate that the Beav pipeline can annotate complete systems and genetic elements applicable to a broad range of bacterial taxa.

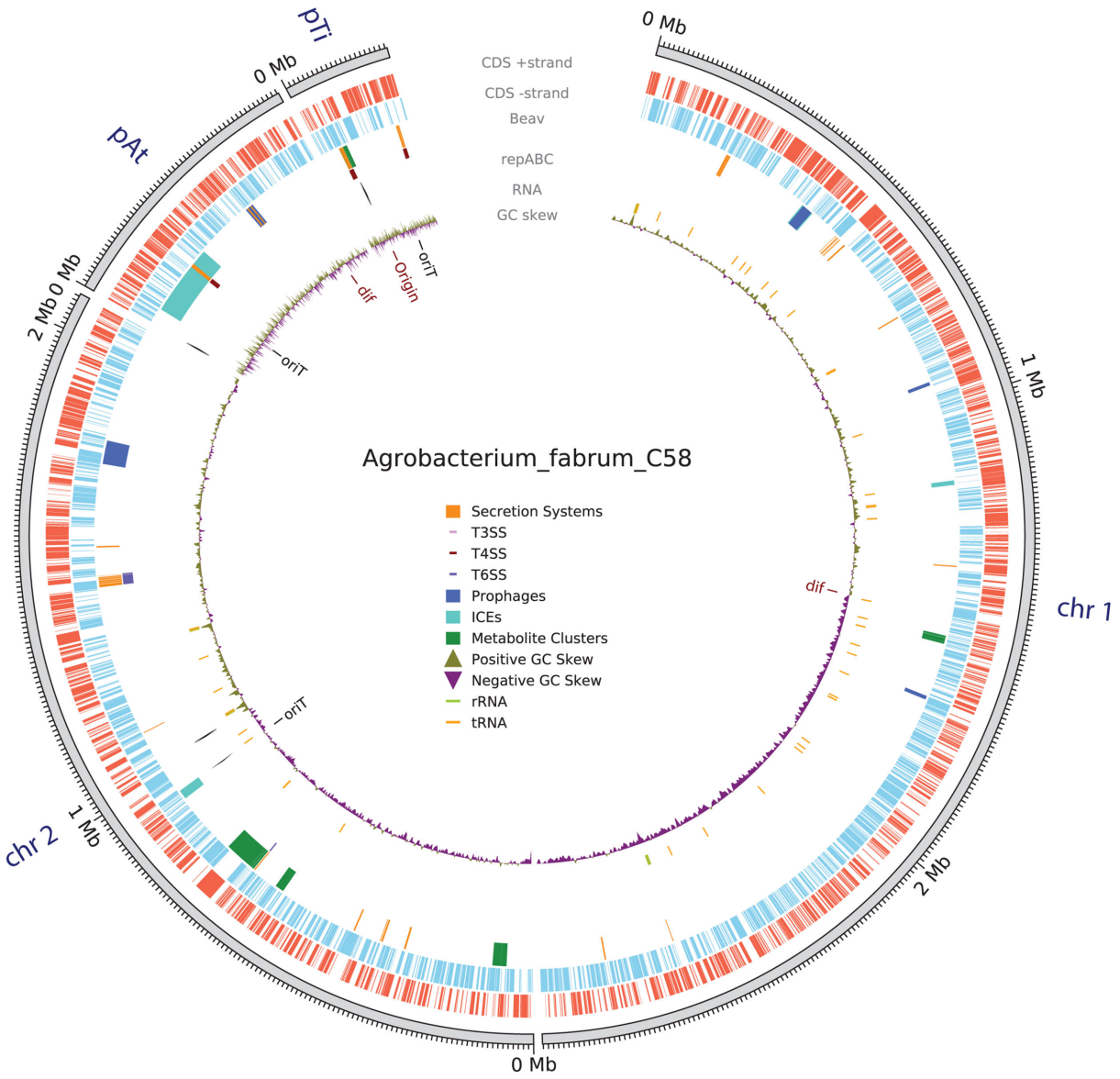

FIG 2 Example Circos plot of whole-genome annotations automatically generated by Beav. There are six inner tracks showing the position of different elements and features. The outermost track shows the length of each assembly contig. The next two tracks indicate CDS position on the ±strand. The third track shows elements annotated by the Beav pipeline, including secretion systems, secondary metabolite gene clusters, and mobile genetic elements. The specific type of secretion system, such as T3SS, T4SS, or T6SS, is plotted as a small colored rectangle along the main feature. Regions encompassing MGEs, such as ICEs and prophages, are indicated as colored blocks. The next track indicates the position of plasmid and secondary chromosome replication loci. The following track indicates the position of ribosomal RNA and other RNA elements. The next track indicates GC Skew across a contig. The innermost track includes markings indicating the position of origin of replication (*oriC*) and MGE origin of transfer (*oriT*) sites.

Analysis of Beav pipeline runtime

To evaluate the runtime of the Beav pipeline, we summarized the amount of time it took Beav and the component steps to run to completion for a diverse set of genomes. Three replicates were run on a cluster computer using eight cores of an AMD EPYC 7601 processor running at 2.2 GHz per core, with 512 GB of available RAM. The average run time of the overall pipeline and each subcomponent was calculated (Table 4). In almost all cases, Beav took less than one hour to run the entire pipeline. The longest steps included the TIGER2 ICE analysis and Bakta which both took 18 min on average to run. TIGER2 runtime increased with the number of ICEs encoded in the genome. For example,

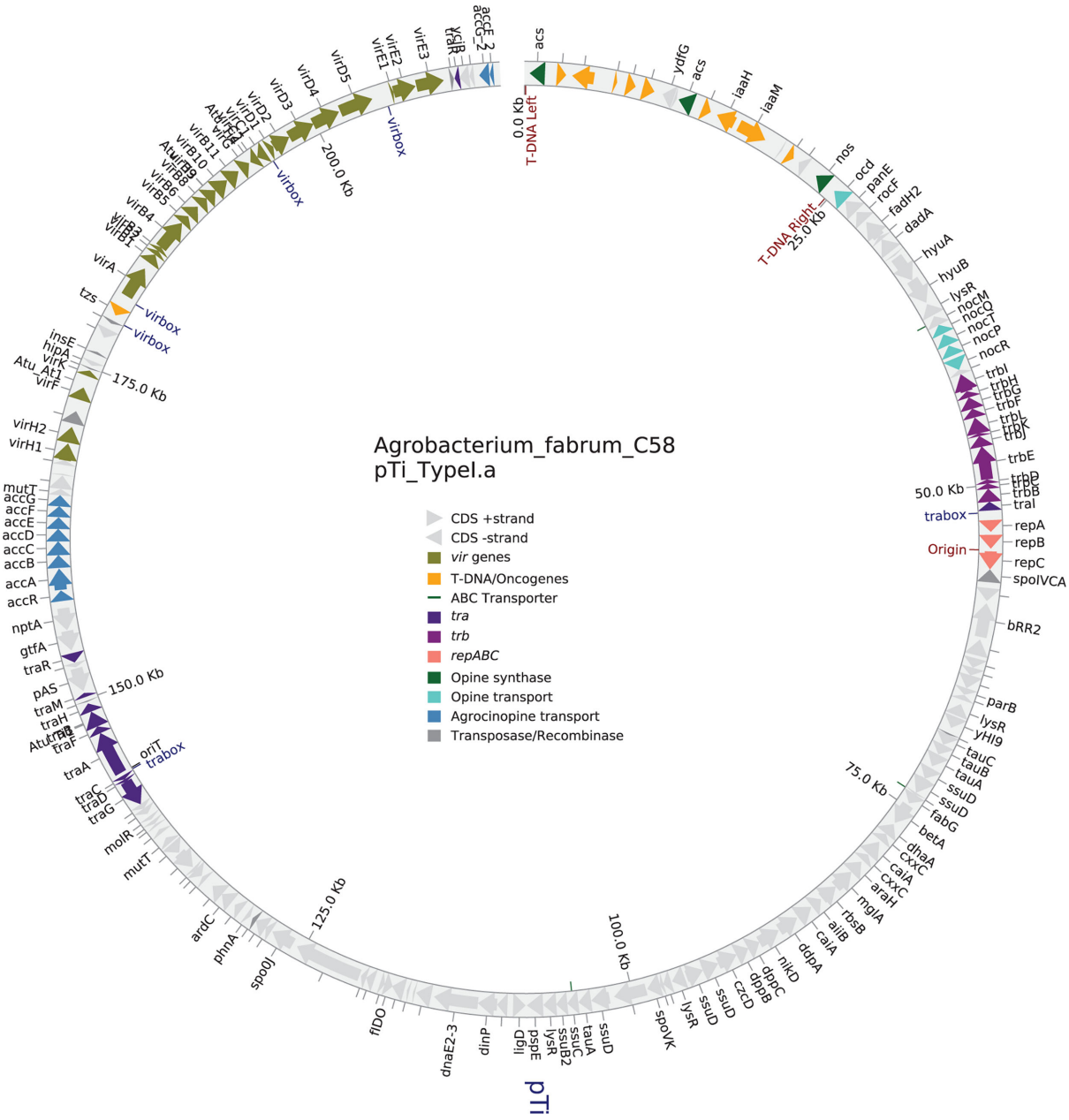

FIG 3 Example Circos plot visualizing oncogenic Ti/Ri plasmids as generated by the optional *Agrobacterium*-specific pipeline. Beav includes an optional *Agrobacterium*-specific pipeline. If this pipeline is run, an additional figure is generated showing contigs identified as representing Ti or Ri plasmids and their annotations. Genes and their strand are visualized as arrows. The presence of key plasmid loci, including T-DNA genes, virulence (*vir*) genes, specialized nutrient (opine) synthesis and catabolism loci, and plasmid replication (*repABC*) and conjugation (*trb*/*tra*) are indicated by arrow color. Other plasmid and regulatory elements, such as T-DNA borders, *vir* box, *tra* box, and the position of ABC transporter genes are indicated. Transposases, integrases, and IS element-associated loci are colored dark gray. All other genes are colored light gray. Gene names, where present, are visualized outside of the ring. If run on draft assemblies with multiple Ti/Ri plasmid contigs, all will be included in the visualization. Several text labels have been adjusted for visibility; otherwise, this figure appears as generated by Beav.

R. fascians D188 encodes 0 ICEs and TIGER2 took about 3 min to run, while *A. fabrum* C58 encodes four ICEs and TIGER2 took 36 min to run. This indicates that the TIGER2 analysis will not greatly increase the runtime if no ICEs are present. However, for genomes that contained multiple ICEs, TIGER2 analysis is a major proportion of Beav runtime. Other steps completed very quickly and did not add much to the overall runtime. The other program's run times vary based on the features in the genome, but their analysis did not

TABLE 2 Counts of hypothetical and uncharacterized proteins in annotations produced by Beav, Bakta, and RefSeq PGAP

Strain	Beav	Bakta	RefSeq (PGAP)
Agrobacterium fabrum C58 (GCF_000092025.1)	126	137	1,848
Escherichia coli 131 (GCF_005221985.1)	368	386	880
Pseudomonas syringae DC3000 (GCF_000007805.1)	137	147	761
Rhodococcus fascians D188 (GCF_001620305.1)	23	24	2,026
Aeromomas caviae WP2-W18-ESBL-01 (GCF_014168635.1)	36	37	1,129

individually exceed 5 min. For example, DefenseFinder, MacSyFinder, and GapMind took around one and a half minutes or less to run for each test genome. These results show that while the Beav pipeline does take longer to run than Bakta, it is not unreasonably longer for the additional annotations Beav provides. Individual programs or steps in the Beav pipeline can be skipped to reduce computational needs or shorten runtime if necessary.

DISCUSSION

Comprehensive genome annotation can be challenging since it requires the use of multiple tools and analyses, some of which are difficult to install or run. It also requires parsing and interpreting the results of these tools, and cross-correlating results for individual genes and loci. Tools can be broadly separated into general genome annotation pipelines, which identify open reading frames and annotate gene and RNA function, and more specialized programs that identify and characterize specific kinds of loci or elements, such as secretion systems or ICEs. For example, tools such as PathoFact and MobileElementFinder can detect MGEs, AMR genes, and metal resistance genes in genome assemblies (75, 76). The mobileOG-db web server and software can annotate various types of mobile elements based on similarity to a database of known MGEs (77). The VRprofile2 web server identifies plasmids, ICEs, and integrons in assembled genomes, though most included tools are built around so-called "ESKAPEE" clinical pathogens (78). Most published annotation pipelines focus on gene identification and annotation and leave the characterization of other kinds of loci to other tools. Few tools combine both kinds of analysis into a single pipeline. Beav is designed to fill this gap and merge results from both general and specialized annotation tools, while adding additional annotations available in no other program. To our knowledge, no other pipeline implements as many diverse annotation tools and analyses as Beav.

Several current genome annotation pipeline options include web-based tools, such as BacAnt, DFAST, Galaxy/Apollo, NCBI PGAP, and RAST (79–83). Web-based services such as Galaxy, Apollo, and Proksee address many of the challenges of annotation by providing a web-based platform that integrates multiple popular annotation tools into a dashboard, making complex analysis and generating figures easier (80, 81, 84). These

TABLE 3 Summary of annotated features and systems predicted by Beav for diverse bacteria

Organism	Biosynthetic gene clusters (antiSMASH)	Secretion systems (MacSyFinder)	Defense systems (DefenseFinder)	ICEs (TIGER2)	Prophages (DBSCAN-SWA)	Integrons (IntegronFinder)	Total new features
Agrobacterium fabrum C58	5	8	8	4	5	0	30
Escherichia coli 131	4	9	9	7	1	0	30
Pseudomonas syringae DC3000	10	9	8	3	0	0	30
Rhodococcus fascians D188	21	1	8	0	0	0	30
Aeromonas caviae WP2-W18-ESBL-01	3	5	9	7	2	2	28

TABLE 4 Runtime for Beav annotation pipeline components

Organism	Beav total run time [hh:mm:ss]	Bakta	MacSyFinder	DefenseFinder	antiSMASH	DBSCAN-SWA	GapMind	TIGER2	IntegronFinder
Agrobacterium fabrum C58 (GCF_000092025.1)	1:04:29	00:16:20	00:00:54	00:01:16	00:02:05	00:02:07	00:01:25	00:36:53	00:00:32
Escherichia coli 131 (GCF_005221985.1)	00:53:50	00:18:25	00:00:50	00:00:41	00:01:42	00:01:28	00:01:11	00:26:23	00:00:31
Pseudomonas syringae DC3000 (GCF_000007805.1)	00:53:24	00:19:26	00:00:57	00:00:42	00:03:06	00:01:02	00:01:14	00:23:35	00:00:48
Rhodococcus fascians D188 (GCF_001620305.1)	00:34:24	00:18:18	00:1:03	00:01:55	00:04:22	00:00:36	00:01:33	00:03:25	00:01:54
Aeromonas caviae WP2-W18-ESBL-01 (GCF_014168635.1)	00:51:34	00:17:20	00:01:10	00:01:08	00:01:21	00:01:11	00:01:06	00:22:36	00:03:43

web services make genome annotation available to users with little to no command-line experience. However, the analysis of whole-genome sequences can be computationally intensive, precluding web servers from high throughput analysis or providing more comprehensive analyses beyond gene identification. Users are often limited to a small number of concurrent jobs, and job wait times can be extensive. This can restrict web server use to analyses of a relatively small number of genomes.

Command-line pipelines are targeted for high-throughput projects and users with some computational experience. There are several established and published command-line annotation pipelines available, including Bakta, Prokka, and MicrobeAnnotator (33, 85, 86). These tools tend to specialize for identification of open reading frames encoding for genes or RNA loci and annotate their predicted function using databases of known genes. Tools such as Bakta work very well for this process and provide high quality annotations for individual genes. Rather than reimplementing this step, we relied on Bakta for initial gene annotations in Beav. We then complement the Bakta annotations with other tools and scripts. However, complex analyses that involve more than one software tool quickly become complicated. Each program produces several outputs and results, and it can be difficult to navigate through every output file. Manually parsing these outputs for valuable information is a time-consuming process. We designed Beav to alleviate the challenges of complex genome annotation and provide a comprehensive tool that users with a basic level of command-line experience can use.

The Beav pipeline makes existing annotation software more accessible. Installation of the Beav pipeline and all its dependencies is made easy using Conda. Conda manages installation so that each program does not need to be installed individually and users do not have to manage dependencies. Likewise, Beav includes a tool that downloads and installs databases and updates for each of the prerequisite programs. Beav was developed with a highly customizable workflow that allows for programs to be run sequentially, independently, or skipped depending on the user's needs. This makes using existing programs easier since running the program is automated and does not require intensive knowledge of each program's usage. Additionally, the pipeline parses the results of multiple annotation tools and combines them into one output. This makes interpretation of annotations easier as results from several programs are merged.

Automation provides convenience to users and simplifies running a pipeline, but that convenience necessitates making decisions on which software to include. We selected tools for each step of the pipeline based on several criteria, including function, run time, open-source code, command line execution, and maximizing information provided to the end user. We selected TIGER2 for the annotation of ICEs over tools such as IslandViewer and ICEberg/ICEfinder because it performs *de novo* prediction rather than prediction by similarity to known islands (87, 88). TIGER2 also predicts exact ICE boundaries and can identify ICEs integrated into sites other than tRNA genes. For annotation of prophage regions, DBSCAN-SWA was selected because

it provides information about the phage components that are present and phage taxonomic classification. Other potential prophage annotation tools are only available as web servers, require docker containers, or use machine learning, which requires large amounts of computing power and/or GPU-based analysis (89). As alternative tools are released in the future, we will consider them as replacements for various steps of the Beav pipeline.

While Beav was designed for annotation of draft or complete genome assemblies, it is also applicable to the annotation of bacterial metagenome-assembled genomes (MAGs). MAGs are sets of assembled contigs representing a single microbial strain extracted from metagenome data (90). Beav supports MAG fasta and GenBank files as input and will produce results similar to those for whole-genome assemblies. However, full metagenome data sets in fasta format are not currently supported due to limitations in gene-calling approaches. For this step, Beav relies on Bakta, which does not currently support complex metagenome samples due to the potential for multiple genetic codes. However, if an annotated bacterial metagenome is provided in GenBank format, Beav will attempt to run the rest of the pipeline on this data. Metagenome data sets containing archaeal or eukaryotic sequences may not be correctly annotated. Runtimes may be extensive given the large size of these data sets.

Having access to these powerful tools makes Beav applicable across many fields of bacterial research. Beav also includes specific databases and tools for improving the quality of annotations for plant-associated microbes, particularly agriculturally relevant phytopathogens, and symbiotic mutualists. Consistent annotation of virulence genes and their names, such as those encoding for secreted effector proteins, is important for effective communication and understanding of pathogens. This makes Beav a valuable tool for plant-microbe and phytopathogen-related studies as the pipeline has gene databases and novel tools that provide reliable naming and annotation conventions. There is currently no single database listing names and representative sequences for virulence genes or effectors for all plant pathogens, and research communities of different pathogens have different standards and naming conventions. Future updates to the Beav database could include genes from other plant-associated microbes, such as additional phytopathogens and mutualists. Unlike other annotation tools, Beav detects promoter and regulatory features unique to plant-associated microbiota. Methods for annotating known promoter and regulatory regions exist and require manual input of patterns or custom models for each region. With Beav, annotation of these elements is automated and greatly reduces the workload that would come with characterizing these regions individually. Accurate and comprehensive whole-genome annotation of phytopathogens and maintaining a reliable repository of genes important for plant-microbe interactions are crucial for pathogen management. We developed Beav to address the need for bioinformatics tools that assist in the analysis of plant-associated microbes and minimize the naming errors commonly associated with these taxa. However, the Beav pipeline and its incorporated tools are broadly applicable to diverse taxa of bacteria.

Beav is the only tool that features an *Agrobacterium*-specific pipeline developed to offer a standardized method for *Agrobacterium* genome annotation. *Agrobacterium* is both an economically important plant pathogen and a biotechnology tool for genetically modifying plants. Thus, this sub-pipeline can aid in *Agrobacterium* pathology and genomics studies, as well as the development of new strains for use in plant transformation and engineering. The *Agrobacterium* oncogenic plasmids, the Ti and Ri plasmids, are diverse in both sequence and content (29, 30). Identifying the presence of an oncogenic plasmid in a genome assembly, and correctly classifying its type, can be difficult, especially in draft assemblies. Consistent classification of plasmids can improve communication on differences between *Agrobacterium* strains. Beav fully automates this process and provides additional annotations for other key virulence elements and regulatory features. The custom gene database also provides for consistent naming of key oncogenic plasmid genes, some of which are frequently misannotated in other tools.

To extend the function of the current Beav pipeline, further work improving annotations and adding other kinds of genome features is still needed. Knowing operon structure is important for understanding gene expression and function, yet associating genes to operons is surprisingly difficult. In our experience, the most accurate *de novo* operon prediction tool that does not require RNA-Seq data is the operon-mapper web server (62). The current version of Beav can submit jobs remotely to this tool. However, using web servers requires an internet connection, limits concurrent jobs, affects version controlling, and web servers may go down or be no longer supported. We hope to adjust the current method of web-based operon mapping towards a local command-line tool, reducing wait times and the potential for network connection and server errors.

Bacterial secreted proteins play a major role in host-microbe interactions for phytopathogenic and clinically relevant bacteria (91). Predicting and identifying genes encoding for secreted proteins is essential for understanding virulence. While Beav annotates genes with similarity to known secreted effectors, it currently does not identify novel effectors *de novo*. There are many published tools for the *de novo* identification of type III and type IV secreted proteins (3). However, in our testing, none of these tools were consistent in identifying known effectors or were only available as web servers, and several tools identified many false positives that are unlikely to encode for secreted proteins. Therefore, we did not include the identification of T3SS and T4SS-secreted proteins in the current iteration of Beav. Tools that can correctly identify secreted proteins with few false positives are needed for accurate annotation.

Other features to investigate for future Beav versions include annotation of other MGEs, such as transposons and plasmids, especially in draft genome assemblies. We did not include tools for the annotation of plasmids, as we found these programs often mis-identify fragmented contigs belonging to genomic islands or ICEs in draft genomes. Comprehensive annotation of regulatory elements, such as promoters and transcription factor binding sites, is also a future development goal. Several programs exist for the prediction of bacterial promoters, such as the sigma70 promoter (92–94). However, most promoter prediction tools were developed based on specific bacterial taxa and do not work well for analyses outside of closely related lineages. There is a need for *de novo* promoter annotation that captures the full breadth of promoters and regulatory elements across diverse bacteria.

ACKNOWLEDGMENTS

We thank Danielle Stevens for testing the Beav pipeline and providing helpful feedback. We also thank Oliver Schwengers for assisting in maintaining dependency compatibility with Bakta, and Björn Grüning for helpful advice on Bioconda integration.

This work was funded by startup funds from the Department of Botany and Plant Pathology at Oregon State University.

AUTHOR AFFILIATION

[1]Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon, USA

AUTHOR ORCIDs

Alexandra J. Weisberg http://orcid.org/0000-0002-0045-1368

AUTHOR CONTRIBUTIONS

Jewell M. Jung, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review and editing | Arafat Rahman, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review and editing | Andrea M. Schiffer, Formal analysis, Investigation, Software, Validation | Alexandra J. Weisberg, Conceptualization, Fund-

ing acquisition, Methodology, Project administration, Software, Supervision, Validation, Writing – review and editing

DATA AVAILABILITY

Beav databases and source code are freely available for download at https://github.com/weisberglab/beav.

REFERENCES

1. Lobb B, Tremblay B-M, Moreno-Hagelsieb G, Doxey AC. 2020. An assessment of genome annotation coverage across the bacterial tree of life. Microb Genom 6:e000341. https://doi.org/10.1099/mgen.0.000341
2. Tatusova T, Ciufo S, Federhen S, Fedorov B, McVeigh R, O'Neill K, Tolstoy I, Zaslavsky L. 2015. Update on RefSeq microbial genomes resources. Nucleic Acids Res 43:D599–D605. https://doi.org/10.1093/nar/gku1062
3. Lovelace AH, Dorhmi S, Hulin MT, Li Y, Mansfield JW, Ma W. 2023. Effector identification in plant pathogens. Phytopathology 113:637–650. https://doi.org/10.1094/PHYTO-09-22-0337-KD
4. Haudiquet M, de Sousa JM, Touchon M, Rocha EPC. 2022. Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations. Philos Trans R Soc Lond B Biol Sci 377:20210234. https://doi.org/10.1098/rstb.2021.0234
5. Weisberg AJ, Chang JH. 2023. Mobile genetic element flexibility as an underlying principle to bacterial evolution. Annu Rev Microbiol 77:603–624. https://doi.org/10.1146/annurev-micro-032521-022006
6. Baltrus DA. 2013. Exploring the costs of horizontal gene transfer. Trends Ecol Evol 28:489–495. https://doi.org/10.1016/j.tree.2013.04.002
7. Baker KS, Dallman TJ, Field N, Childs T, Mitchell H, Day M, Weill F-X, Lefèvre S, Tourdjman M, Hughes G, Jenkins C, Thomson N. 2018. Horizontal antimicrobial resistance transfer drives epidemics of multiple *Shigella* species. Nat Commun 9:1462. https://doi.org/10.1038/s41467-018-03949-8
8. Sullivan JT, Patrick HN, Lowther WL, Scott DB, Ronson CW. 1995. Nodulating strains of *Rhizobium loti* arise through chromosomal symbiotic gene transfer in the environment. Proc Natl Acad Sci U S A 92:8985–8989. https://doi.org/10.1073/pnas.92.19.8985
9. Baltrus DA, Feng Q, Kvitko BH. 2022. Genome context influences evolutionary flexibility of nearly identical type III effectors in two phytopathogenic *Pseudomonads*. Front Microbiol 13:826365. https://doi.org/10.3389/fmicb.2022.826365
10. Weisberg AJ, Rahman A, Backus D, Tyavanagimatt P, Chang JH, Sachs JL. 2022. Pangenome evolution reconciles robustness and instability of rhizobial symbiosis. mBio 13:e0007422. https://doi.org/10.1128/mbio.00074-22
11. Colombi E, Perry BJ, Sullivan JT, Bekuma AA, Terpolilli JJ, Ronson CW, Ramsay JP. 2021. Comparative analysis of integrative and conjugative mobile genetic elements in the genus *Mesorhizobium*. Microb Genom 7:000657. https://doi.org/10.1099/mgen.0.000657
12. Botelho J, Schulenburg H. 2021. The role of integrative and conjugative elements in antibiotic resistance evolution. Trends Microbiol 29:8–18. https://doi.org/10.1016/j.tim.2020.05.011
13. Hall RM, Collis CM. 1995. Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. Mol Microbiol 15:593–600. https://doi.org/10.1111/j.1365-2958.1995.tb02368.x
14. Souque C, Escudero JA, MacLean RC. 2021. Integron activity accelerates the evolution of antibiotic resistance. Elife 10:e62474. https://doi.org/10.7554/eLife.62474
15. Gillings MR. 2014. Integrons: past, present, and future. Microbiol Mol Biol Rev 78:257–277. https://doi.org/10.1128/MMBR.00056-13
16. Juhas M, Crook DW, Hood DW. 2008. Type IV secretion systems: tools of bacterial horizontal gene transfer and virulence. Cell Microbiol 10:2377–2386. https://doi.org/10.1111/j.1462-5822.2008.01187.x
17. Pena RT, Blasco L, Ambroa A, González-Pedrajo B, Fernández-García L, López M, Bleriot I, Bou G, García-Contreras R, Wood TK, Tomás M. 2019. Relationship between quorum sensing and secretion systems. Front Microbiol 10:1100. https://doi.org/10.3389/fmicb.2019.01100
18. Green ER, Mecsas J. 2016. Bacterial secretion systems: an overview. Microbiol Spectr 4. https://doi.org/10.1128/microbiolspec.VMBF-0012-2015
19. Johnson CM, Grossman AD. 2015. Integrative and conjugative elements (ICEs): what they do and how they work. Annu Rev Genet 49:577–601. https://doi.org/10.1146/annurev-genet-112414-055018
20. Costa TRD, Patkowski JB, Macé K, Christie PJ, Waksman G. 2024. Structural and functional diversity of type IV secretion systems. Nat Rev Microbiol 22:170–185. https://doi.org/10.1038/s41579-023-00974-3
21. Nester EW. 2014. *Agrobacterium*: nature's genetic engineer. Front Plant Sci 5:730. https://doi.org/10.3389/fpls.2014.00730
22. Rocha EPC, Bikard D. 2022. Microbial defenses against mobile genetic elements and viruses: who defends whom from what? PLoS Biol 20:e3001514. https://doi.org/10.1371/journal.pbio.3001514
23. Tonkin-Hill G, Corander J, Parkhill J. 2023. Challenges in prokaryote pangenomics. Microb Genom 9:001021. https://doi.org/10.1099/mgen.0.001021
24. Néron B, Denise R, Coluzzi C, Touchon M, Rocha EPC, Abby SS. 2023. MacSyFinder v2: improved modelling and search engine to identify molecular systems in genomes. Peer Community J 3:e28. https://doi.org/10.24072/pcjournal.250
25. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, Weber T. 2021. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res 49:W29–W35. https://doi.org/10.1093/nar/gkab335
26. Néron B, Littner E, Haudiquet M, Perrin A, Cury J, Rocha EPC. 2022. IntegronFinder 2.0: identification and analysis of integrons across bacteria, with a focus on antibiotic resistance in *Klebsiella*. Microorganisms 10:700. https://doi.org/10.3390/microorganisms10040700
27. Gan R, Zhou F, Si Y, Yang H, Chen C, Ren C, Wu J, Zhang F. 2022. DBSCAN-SWA: an integrated tool for rapid prophage detection and annotation. Front Genet 13:885048. https://doi.org/10.3389/fgene.2022.885048
28. Mageeney CM, Lau BY, Wagner JM, Hudson CM, Schoeniger JS, Krishnakumar R, Williams KP. 2020. New candidates for regulated gene integrity revealed through precise mapping of integrative genetic elements. Nucleic Acids Res 48:4052–4065. https://doi.org/10.1093/nar/gkaa156
29. Weisberg AJ, Davis EW, Tabima J, Belcher MS, Miller M, Kuo C-H, Loper JE, Grünwald NJ, Putnam ML, Chang JH. 2020. Unexpected conservation and global transmission of agrobacterial virulence plasmids. Science 368:eaba5256. https://doi.org/10.1126/science.aba5256
30. Weisberg AJ, Miller M, Ream W, Grünwald NJ, Chang JH. 2022. Diversification of plasmids in a genus of pathogenic and nitrogen-fixing bacteria. Philos Trans R Soc Lond B Biol Sci 377:20200466. https://doi.org/10.1098/rstb.2020.0466
31. Keane PJ, Kerr A, New PB. 1970. Crown gall of stone fruit II. Identification and nomenclature of *Agrobacterium* isolates. Aust J Biol Sci 23:585. https://doi.org/10.1071/BI9700585
32. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Köster J, Bioconda Team. 2018. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods 15:475–476. https://doi.org/10.1038/s41592-018-0046-7
33. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. 2021. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microb Genom 7:000685. https://doi.org/10.1099/mgen.0.000685
34. Savory EA, Fuller SL, Weisberg AJ, Thomas WJ, Gordon MI, Stevens DM, Creason AL, Belcher MS, Serdani M, Wiseman MS, Grünwald NJ, Putnam

ML, Chang JH. 2017. Evolutionary transitions between beneficial and phytopathogenic *Rhodococcus* challenge disease management. Elife 6:e30925. https://doi.org/10.7554/eLife.30925
35. Savory EA, Weisberg AJ, Stevens DM, Creason AL, Fuller SL, Pearce EM, Chang JH. 2020. Phytopathogenic *Rhodococcus* have diverse plasmids with few conserved virulence functions. Front Microbiol 11:1022. https://doi.org/10.3389/fmicb.2020.01022
36. Huguet-Tapia JC, Badger JH, Loria R, Pettis GS. 2011. *Streptomyces turgidiscabies* Car8 contains a modular pathogenicity island that shares virulence genes with other actinobacterial plant pathogens. Plasmid 65:118–124. https://doi.org/10.1016/j.plasmid.2010.11.002
37. Peeters N, Carrère S, Anisimova M, Plener L, Cazalé A-C, Genin S. 2013. Repertoire, unified nomenclature and evolution of the type III effector gene set in the *Ralstonia solanacearum* species complex. BMC Genomics 14:859. https://doi.org/10.1186/1471-2164-14-859
38. Lindeberg M, Stavrinides J, Chang JH, Alfano JR, Collmer A, Dangl JL, Greenberg JT, Mansfield JW, Guttman DS. 2005. Proposed guidelines for a unified nomenclature and phylogenetic analysis of type III Hop effector proteins in the plant pathogen *Pseudomonas syringae*. Mol Plant Microbe Interact 18:275–282. https://doi.org/10.1094/MPMI-18-0275
39. Onyeziri MC, Hardy GG, Natarajan R, Xu J, Reynolds IP, Kim J, Merritt PM, Danhorn T, Hibbing ME, Weisberg AJ, Chang JH, Fuqua C. 2022. Dual adhesive unipolar polysaccharides synthesized by overlapping biosynthetic pathways in *Agrobacterium tumefaciens*. Mol Microbiol 117:1023–1047. https://doi.org/10.1111/mmi.14887
40. Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet 16:276–277. https://doi.org/10.1016/s0168-9525(00)02024-2
41. Koebnik R, Krüger A, Thieme F, Urban A, Bonas U. 2006. Specific binding of the *Xanthomonas campestris* pv. vesicatoria AraC-type transcriptional activator HrpX to plant-inducible promoter boxes. J Bacteriol 188:7652–7660. https://doi.org/10.1128/JB.00795-06
42. Zwiesler-Vollick J, Plovanich-Jones AE, Nomura K, Bandyopadhyay S, Joardar V, Kunkel BN, He SY. 2002. Identification of novel *hrp*-regulated genes through functional genomic analysis of the *Pseudomonas syringae* pv. *tomato* DC3000 genome. Mol Microbiol 45:1207–1218. https://doi.org/10.1046/j.1365-2958.2002.02964.x
43. Yang S, Peng Q, Zhang Q, Zou L, Li Y, Robert C, Pritchard L, Liu H, Hovey R, Wang Q, Birch P, Toth IK, Yang C-H. 2010. Genome-wide identification of HrpL-regulated genes in the necrotrophic phytopathogen *Dickeya dadantii* 3937. PLoS One 5:e13472. https://doi.org/10.1371/journal.pone.0013472
44. Fouts DE, Abramovitch RB, Alfano JR, Baldo AM, Buell CR, Cartinhour S, Chatterjee AK, D'Ascenzo M, Gwinn ML, Lazarowitz SG, Lin N-C, Martin GB, Rehm AH, Schneider DJ, van Dijk K, Tang X, Collmer A. 2002. Genomewide identification of *Pseudomonas syringae* pv. *tomato* DC3000 promoters controlled by the HrpL alternative sigma factor. Proc Natl Acad Sci U S A 99:2275–2280. https://doi.org/10.1073/pnas.032514099
45. Passaglia LMP. 2017. *Bradyrhizobium elkanii nod* regulon: insights through genomic analysis. Genet Mol Biol 40:703–716. https://doi.org/10.1590/1678-4685-GMB-2016-0228
46. Perry BJ, Sullivan JT, Colombi E, Murphy RJT, Ramsay JP, Ronson CW. 2020. Symbiosis islands of Loteae-nodulating *Mesorhizobium* comprise three radiating lineages with concordant *nod* gene complements and nodulation host-range groupings. Microb Genom 6:mgen000426. https://doi.org/10.1099/mgen.0.000426
47. Suominen L, Paulin L, Saano A, Saren A-M, Tas E, Lindström K. 1999. Identification of nodulation promoter (*nod*-box) regions of *Rhizobium galegae*. FEMS Microbiol Lett 177:217–223. https://doi.org/10.1111/j.1574-6968.1999.tb13735.x
48. Zehner S, Schober G, Wenzel M, Lang K, Göttfert M. 2008. Expression of the *Bradyrhizobium japonicum* type III secretion system in legume nodules and analysis of the associated *tts* box promoter. Mol Plant Microbe Interact 21:1087–1093. https://doi.org/10.1094/MPMI-21-8-1087
49. Krause A, Doerfel A, Göttfert M. 2002. Mutational and transcriptional analysis of the type III secretion system of *Bradyrhizobium japonicum*. Mol Plant Microbe Interact 15:1228–1235. https://doi.org/10.1094/MPMI.2002.15.12.1228
50. Okazaki S, Okabe S, Zehner S, Göttfert M, Saeki K. 2008. Symbiotic roles and transcriptional analysis of the type III secretion system in *Mesorhizobium loti*, p 235–235. In Dakora FD, Chimphango SBM, Valentine AJ, Elmerich C, Newton WE (ed), Biological nitrogen fixation: towards poverty alleviation through sustainable agriculture. Springer Netherlands.
51. Carnoy C, Roten C-A. 2009. The *dif*/Xer recombination systems in proteobacteria. PLoS One 4:e6531. https://doi.org/10.1371/journal.pone.0006531
52. Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195
53. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033
54. Ares-Arroyo M, Coluzzi C, Rocha EPC. 2023. Origins of transfer establish networks of functional dependencies for plasmid transfer by conjugation. Nucleic Acids Res 51:3001–3016. https://doi.org/10.1093/nar/gkac1079
55. Camargo AP, Call L, Roux S, Nayfach S, Huntemann M, Palaniappan K, Ratner A, Chu K, Mukherjeep S, Reddy TBK, Chen I-MA, Ivanova NN, Eloe-Fadrosh EA, Woyke T, Baltrus DA, Castañeda-Barba S, de la Cruz F, Funnell BE, Hall JPJ, Mukhopadhyay A, Rocha EPC, Stalder T, Top E, Kyrpides NC. 2024. IMG/PR: a database of plasmids from genomes and metagenomes with rich annotations and metadata. Nucleic Acids Res 52:D164–D173. https://doi.org/10.1093/nar/gkad964
56. Abby SS, Rocha EPC. 2017. Identification of protein secretion systems in bacterial genomes using MacSyFinder. Methods Mol Biol 1615:1–21. https://doi.org/10.1007/978-1-4939-7033-9_1
57. Tesson F, Hervé A, Mordret E, Touchon M, d'Humières C, Cury J, Bernheim A. 2022. Systematic and quantitative view of the antiviral arsenal of prokaryotes. Nat Commun 13:2561. https://doi.org/10.1038/s41467-022-30269-9
58. Georjon H, Bernheim A. 2023. The highly diverse antiphage defence systems of bacteria. Nat Rev Microbiol 21:686–700. https://doi.org/10.1038/s41579-023-00934-x
59. Price MN, Deutschbauer AM, Arkin AP. 2020. GapMind: automated annotation of amino acid biosynthesis. mSystems 5:e00291-20. https://doi.org/10.1128/mSystems.00291-20
60. Price MN, Deutschbauer AM, Arkin AP. 2022. Filling gaps in bacterial catabolic pathways with computation and high-throughput genetics. PLoS Genet 18:e1010156. https://doi.org/10.1371/journal.pgen.1010156
61. Jacob F, Perrin D, Sanchez C, Monod J. 1960. [Operon: a group of genes with the expression coordinated by an operator]. C R Hebd Seances Acad Sci 250:1727–1729.
62. Taboada B, Estrada K, Ciria R, Merino E. 2018. Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes. Bioinformatics 34:4118–4120. https://doi.org/10.1093/bioinformatics/bty496
63. Thomas J, Watve SS, Ratcliff WC, Hammer BK. 2017. Horizontal gene transfer of functional type VI killing genes by natural transformation. mBio 8:e00654-17. https://doi.org/10.1128/mBio.00654-17
64. Cho H, Winans SC. 2005. VirA and VirG activate the Ti plasmid *repABC* operon, elevating plasmid copy number in response to wound-released chemical signals. Proc Natl Acad Sci U S A 102:14843–14848. https://doi.org/10.1073/pnas.0503458102
65. Swiderska A, Berndtson AK, Cha M-R, Li L, Beaudoin GMJ, Zhu J, Fuqua C. 2001. Inhibition of the *Agrobacterium tumefaciens* TraR quorum-sensing regulator: INTERACTIONS WITH THE TraM ANTI-ACTIVATOR. J Biol Chem 276:49449–49458. https://doi.org/10.1074/jbc.M107881200
66. Fuqua C, Winans SC. 1996. Conserved *cis*-acting promoter elements are required for density-dependent transcription of *Agrobacterium tumefaciens* conjugal transfer genes. J Bacteriol 178:435–440. https://doi.org/10.1128/jb.178.2.435-440.1996
67. Li P-L, Farrand SK. 2000. The replicator of the nopaline-type Ti plasmid pTiC58 is a member of the *repABC* family and is influenced by the TraR-dependent quorum-sensing regulatory system. J Bacteriol 182:179–188. https://doi.org/10.1128/JB.182.1.179-188.2000
68. Otten L, Schmidt J. 1998. A T-DNA from the *Agrobacterium tumefaciens* limited-host-range strain AB2/73 contains a single oncogene. Mol Plant Microbe Interact 11:335–342. https://doi.org/10.1094/MPMI.1998.11.5.335
69. Shurvinton CE, Ream W. 1991. Stimulation of *Agrobacterium tumefaciens* T-DNA transfer by overdrive depends on a flanking sequence but not on

helical position with respect to the border repeat. J Bacteriol 173:5558–5563. https://doi.org/10.1128/jb.173.17.5558-5563.1991
70. Lu J, den Dulk-Ras A, Hooykaas PJJ, Glover JNM. 2009. *Agrobacterium tumefaciens* VirC2 enhances T-DNA transfer and virulence through its C-terminal ribbon–helix–helix DNA-binding fold. Proc Natl Acad Sci U S A 106:9643–9648. https://doi.org/10.1073/pnas.0812199106
71. Bushnell B. BBMap. Available from: https://sourceforge.net/projects/bbmap/
72. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. https://doi.org/10.1038/s41467-018-07641-9
73. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423. https://doi.org/10.1093/bioinformatics/btp163
74. Shimoyama Y. 2022. pyCirclize: circular visualization in Python [computer software]. https://github.com/moshi4/pyCirclize.
75. de Nies L, Lopes S, Busi SB, Galata V, Heintz-Buschart A, Laczny CC, May P, Wilmes P. 2021. PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. Microbiome 9:49. https://doi.org/10.1186/s40168-020-00993-9
76. Johansson MHK, Bortolaia V, Tansirichaiya S, Aarestrup FM, Roberts AP, Petersen TN. 2021. Detection of mobile genetic elements associated with antibiotic resistance in *Salmonella enterica* using a newly developed web tool: MobileElementFinder. J Antimicrob Chemother 76:101–109. https://doi.org/10.1093/jac/dkaa390
77. Brown CL, Mullet J, Hindi F, Stoll JE, Gupta S, Choi M, Keenum I, Vikesland P, Pruden A, Zhang L. 2022. mobileOG-db: a manually curated database of protein families mediating the life cycle of bacterial mobile Genetic elements. Appl Environ Microbiol 88:e0099122. https://doi.org/10.1128/aem.00991-22
78. Wang M, Goh Y-X, Tai C, Wang H, Deng Z, Ou H-Y. 2022. VRprofile2: detection of antibiotic resistance-associated mobilome in bacterial pathogens. Nucleic Acids Res 50:W768–W773. https://doi.org/10.1093/nar/gkac321
79. Hua X, Liang Q, Deng M, He J, Wang M, Hong W, Wu J, Lu B, Leptihn S, Yu Y, Chen H. 2021. BacAnt: a combination annotation server for bacterial DNA sequences to identify antibiotic resistance genes, integrons, and transposable elements. Front Microbiol 12:649969. https://doi.org/10.3389/fmicb.2021.649969
80. Ramsey J, Rasche H, Maughmer C, Criscione A, Mijalis E, Liu M, Hu JC, Young R, Gill JJ. 2020. Galaxy and Apollo as a biologist-friendly interface for high-quality cooperative phage genome annotation. PLOS Comput Biol 16:e1008214. https://doi.org/10.1371/journal.pcbi.1008214
81. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, et al. 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75. https://doi.org/10.1186/1471-2164-9-75
82. Tanizawa Y, Fujisawa T, Nakamura Y. 2018. DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. Bioinformatics 34:1037–1039. https://doi.org/10.1093/bioinformatics/btx713
83. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 44:6614–6624. https://doi.org/10.1093/nar/gkw569
84. Grant JR, Enns E, Marinier E, Mandal A, Herman EK, Chen C-Y, Graham M, Van Domselaar G, Stothard P. 2023. Proksee: in-depth characterization and visualization of bacterial genomes. Nucleic Acids Res 51:W484–W492. https://doi.org/10.1093/nar/gkad326
85. Ruiz-Perez CA, Conrad RE, Konstantinidis KT. 2021. MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes. BMC Bioinformatics 22:11. https://doi.org/10.1186/s12859-020-03940-5
86. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153
87. Wang M, Liu G, Liu M, Tai C, Deng Z, Song J, Ou H-Y. 2024. ICEberg 3.0: functional categorization and analysis of the integrative and conjugative elements in bacteria. Nucleic Acids Res 52:D732–D737. https://doi.org/10.1093/nar/gkad935
88. Bertelli C, Laird MR, Williams KP, Lau BY, Hoad G, Winsor GL, Brinkman FSL, SimonFraserUniversityResearchComputingGroup. 2017. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. Nucleic Acids Res 45:W30–W35. https://doi.org/10.1093/nar/gkx343
89. Wishart DS, Han S, Saha S, Oler E, Peters H, Grant JR, Stothard P, Gautam V. 2023. PHASTEST: faster than PHASTER, better than PHAST. Nucleic Acids Res 51:W443–W450. https://doi.org/10.1093/nar/gkad382
90. Sangwan N, Xia F, Gilbert JA. 2016. Recovering complete and draft population genomes from metagenome datasets. Microbiome 4:8. https://doi.org/10.1186/s40168-016-0154-5
91. Hogenhout SA, Van der Hoorn RAL, Terauchi R, Kamoun S. 2009. Emerging concepts in effector biology of plant-associated organisms. Mol Plant Microbe Interact 22:115–122. https://doi.org/10.1094/MPMI-22-2-0115
92. Coppens L, Wicke L, Lavigne R. 2022. SAPPHIRE.CNN: implementation of dRNA-seq-driven, species-specific promoter prediction using convolutional neural networks. Comput Struct Biotechnol J 20:4969–4974. https://doi.org/10.1016/j.csbj.2022.09.006
93. Shahmuradov IA, Mohamad Razali R, Bougouffa S, Radovanovic A, Bajic VB. 2017. bTSSfinder: a novel tool for the prediction of promoters in cyanobacteria and *Escherichia coli*. Bioinformatics 33:334–340. https://doi.org/10.1093/bioinformatics/btw629
94. Liu B, Yang F, Huang D-S, Chou K-C. 2018. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics 34:33–40. https://doi.org/10.1093/bioinformatics/btx579