Appendix 1. Bacterial Comparative Genomics Analytic Workflow

This analytic workflow consists of three major parts:

1) Draft genome assembly of novel isolate
2) Genome annotation (structure, function, specialty gene)
3) Genomic and proteomic comparison between genomes of novel isolate and reference

The following script consists of all command lines (grey-shaded), instructions for each software's installation and operation, and example output files and figures.

Please keep in mind that all software listed in this workflow is either appropriate for smaller bacteria genomes, or has been demonstrated to perform better than other comparable software, or is recommended by my advisors Dr. Emma Allen-Vercoe and Dr. Andrew Kropinski. If new bioinformatics tools are shown to be more suitable for certain steps, one can easily do the substitution.

Table of Contents

Before proceeding with the primary content of this analytic workflow, one is advised to install Bioconda, which will enable easy installation of biomedical-related software using the Conda package manager.

Install Bioconda

1. **Install Conda via Miniconda**
   https://bioconda.github.io/user/install.html
   curl -O https://repo.anaconda.com/miniconda/Miniconda3-latest-MacOSX-x86_64.sh
   sh Miniconda3-latest-MacOSX-x86_64.sh
   If Miniconda is successfully installed, one will see "Thank you for installing Miniconda3!"

2. **Set up channels**
   Add the bioconda channel and other channels. It is important to add following command line one line at a time and in order
   conda config --add channels defaults
   conda config --add channels bioconda
   conda config --add channels conda-forge
   Bioconda is now enabled.

   All the package that can be install via Bioconda can be find on this page: https://anaconda.org/bioconda/repo.

Draft genome assembly

FastQC 0.11.9

Before attempting to assemble Illumina paired-end short reads of novel isolate, examining the quality of reads is recommended.

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

1. Install FastQC via Bioconda

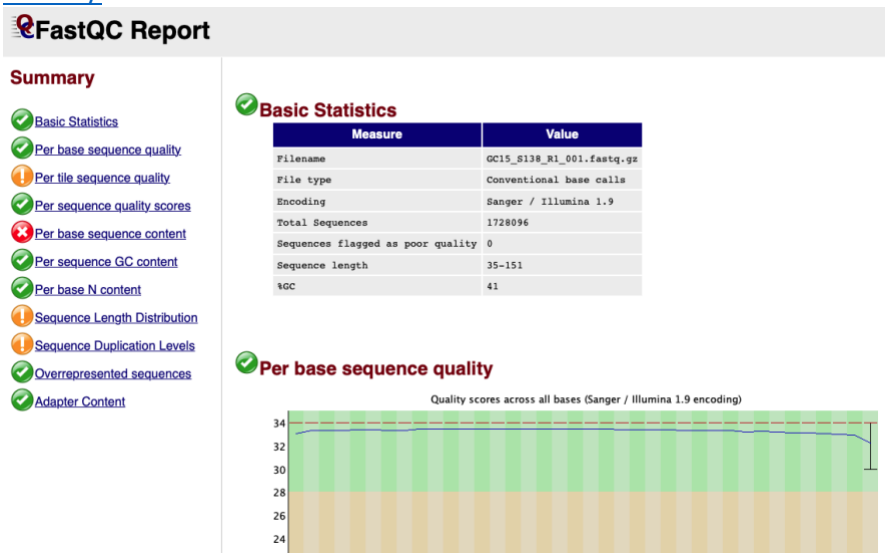
https://anaconda.org/bioconda/fastqc
conda install -c bioconda fastqc
With command: fastqc -v, the version of FastQC can be checked. 0.11.9 would be current latest version.

2. Check quality of reads file

Open FastQC graphical user interface (GUI) with command fastqc. Input file sequences by selecting "File > Open" in the FastQC menu and navigating to the right folder, input FASTQ files can be compressed or decompressed. By hitting hit the "Open" button, FastQC will commence the analysis.

3. Quality report

A series of reports on the sequences will be showed when the analysis has finished. All graph and reports are available for export, use "File > Save report…" before closing. Detailed report interpretations are available on https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/



Trimmomatic 0.39

Trimmomatic is a read trimming tool. Based on the quality results presented by FastQC, any low-quality reads or adaptor sequences will need to be trimmed.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

1. Install Trimmomatic via Bioconda

https://anaconda.org/bioconda/trimmomatic
`conda install -c bioconda trimmomatic`
After trimmomatic is installed, the version can be check by command: `trimmomatic -version`. 0.39 will be current latest version.

2. Quality trim and remove adapter on paired end reads

https://github.com/usadellab/Trimmomatic has detailed instructions of how to select trimming steps and their associated parameters on command line.
After going to directory of paired end sequence files that are pending for process, following conman line can be used to run Trimmomatic:

```
trimmomatic PE -phred33 input_forward.fq.gz input_reverse.fq.gz
output_forward_paired.fq.gz output_forward_unpaired.fq.gz
output_reverse_paired.fq.gz output_reverse_unpaired.fq.gz ILLUMINACLIP:TruSeq3-
PE.fa:2:30:10:2:True SLIDINGWINDOW:4:20 LEADING:3 TRAILING:3 MINLEN:36
```

`PE` is specifying the paired-end Illumina reads are used as input data.
`-phred33` stands for the PHRED quality score system with an ASCII offset of 33.
The next two inputs are the forward and reverse reads in FASTQ files (files can be in compressed format), and following four gun zipped FASTQ files represent output files for paired and unpaired sequences in the forward and reverse format.
`ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:True` will cut adaptor and Illumina-specific sequences.
`SLIDINGWINDOW:4:20` will perform a 4-base wide sliding window trimming which continually scanned through each sequence read, and if the average quality score of each 4 bases is less than 20, the read will be removed.
`LEADING:3 TRAILING:3` will remove base that have a quality score less than 3 at the beginning and end of reads.
`MINLEN:36` will drop reads below the 36 bases long.
The different processing steps will be complete in the order as they are specified by the command line.

SPAdes 3.15.2

SPAdes has been recommended as 'go-to' *de novo* assembler for bacteria with small genome.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology, 19(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

1. **Install SPAdes via Bioconda** https://anaconda.org/bioconda/spades
   conda install -c bioconda spades

2. ***De novo* genome assembly**
   Detailed manual is available on http://cab.spbu.ru/files/release3.15.4/manual.html.
   The input files for SPAdes will be paired-end and high-quality reads that survived the Trimmomatic filter.
   To run SPAdes from the command line, use spades.py [options] -o <output_dir>. Specify the input data: -1 <file_name> for file with forward reads, -2 <file_name> for file with reverse reads. Parameter --careful, will be used to minimizes the number of mismatches and short indels for small genomes during assembly.
   SPAdes requires full path to current working directory, which can be obtained by pwd.
   Example command used in this analytic pipeline is:
   spades.py -1 /Users/scx/Desktop/6999/GC15/GC15_R1_PE.fastq.gz -2 /Users/scx/Desktop/6999/GC15/GC15_R2_PE.fastq.gz -o /Users/scx/Desktop/6999/GC15/SPAdes --careful

3. **Output**
   All output files are stored in <output_dir>.
   - contigs.fasta – resulting contigs
   - assembly_graph.fastg – assembly graph

SeqMan NGen 17.2

This assembler is recommended by Dr. Kropinski and kindly provided for use in this project by DNASTAR. It is a genomic sequence assembly application included in the DNASTAR Lasergene software suite, which automates various tasks: organizing replicates, incorporating BED and VCF files, and detecting variants.

Detailed tutorial is available on: https://www.dnastar.com/manuals/seqman-ultra/17.3.1/en/topic/tutorial-2-whole-genome-de-novo-workflow-with-mate-pair-data

SeqMan NGen®. Version 17.2. DNASTAR. Madison, WI.

QUAST 5.0.2

QUAST is a quality assessment tool for evaluating and comparing genome assemblies with or without reference. It produces summary tables with important metrics including: number of contigs, total length, largest contig, GC (%), N50, L50, number of N's, and plots.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

1. **Install QUAST**
   http://quast.sourceforge.net/install.html
   Via Bioconda, conda install -c bioconda quast.

2. **Evaluate statistical quality of genome assembly**
   Perform with command line python quast.py [options] <contig_file(s)> -o <output_dir>, or with QUAST web interface http://cab.cc.spbu.ru/quast/
   Example command line: quast.py -o QUAST_500 --glimmer GC15_contigs.fasta C_GC15_contigs.fasta D_GC15_contigs.fas

BUSCO 5.3.2

BUSCO is a tool to assess completeness of genome assembly by comparing single-copy orthologs to OrthoDB database.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

1. **Install BUSCO in conda**
   conda install -c conda-forge -c bioconda busco=5.3.2
   Detailed user guide can be found on
   https://busco.ezlab.org/busco_userguide.html#conda-package

3. **Download the lineages database**
   Database for bacteria "bacteria_odb10" can be downloaded from https://busco-data.ezlab.org/v5/data/lineages/ (Creation date: 2020-03-06)

4. **Run BUSCO on assembly contig file**
   busco -i [SEQUENCE_FILE] -l [LINEAGE] -o [OUTPUT_NAME] -m [MODE] [OTHER OPTIONS]
   Example command used: busco -i C_GC15_contigs.fasta -l bacteria_odb10 -o BUSCO_C -m genome *# genome mode: assessing a genome assembly*

CheckM 1.0.18

CheckM offers a set of tools for evaluating the quality of the assembled genome, especially estimating the completeness and contamination. CheckM searched for marker genes (both ubiquitous and single-copy genes) that should be found at a specific number in a particular genome. The less these genes are detected, the lower the completeness. The more they are detected above their theoretical level, the higher the contamination level.
If the completeness of assembly is good, you can proceed with following analysis. If significant potential contamination is found, further steps are required to remove the contamination.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research, 25(7), 1043–1055. https://doi.org/10.1101/gr.186072.114

1. **Run online tool directly on KBase-CheckM App**
   https://kbase.us/applist/apps/kb_Msuite/run_checkM_lineage_wf/release

2. **Input**
   Input file for assessing the genome quality can be assembly object and a binned contigs file. All parameters are set as default.

3. **Or install to Computer**
   Detailed instruction is available on https://github.com/Ecogenomics/CheckM/wiki

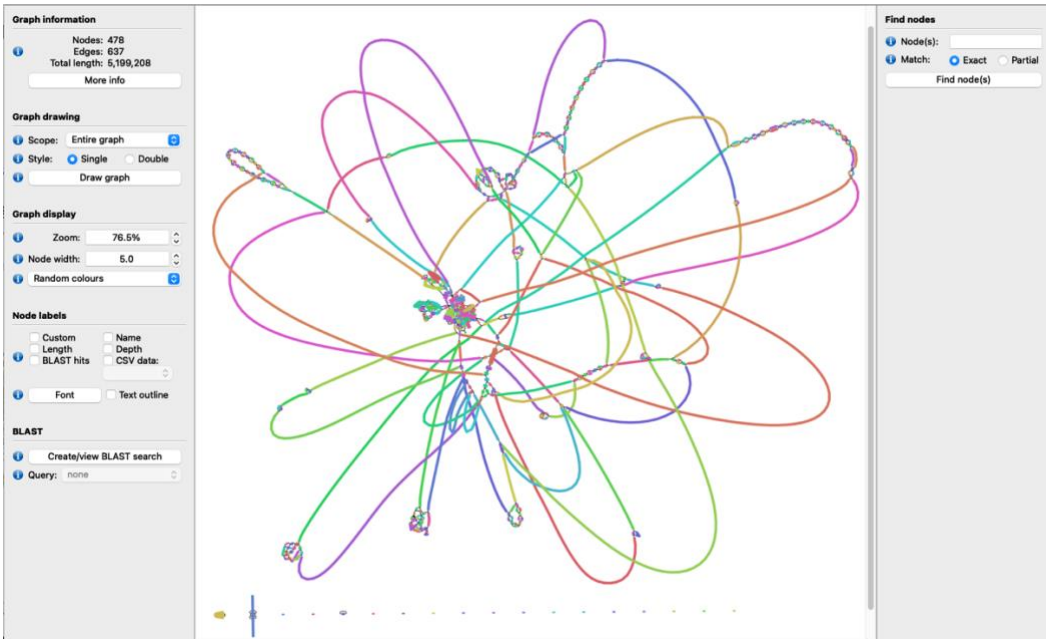
Bandage 0.9.0

Bandage represents a Bioinformatics Application for Navigating De novo Assembly Graphs Easily. The *de novo* assembly program SPAdes produces a graph while the assembly is completed. The graph can be used by Bandage to connect contigs and visualize the de novo assembly. From the visualized graph, you can get some idea of the quality of assembly.

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: Interactive visualization of de novo genome assemblies: Fig. 1. Bioinformatics, 31(20), 3350–3352. https://doi.org/10.1093/bioinformatics/btv383

1. **Installation**
   Install via Bioconda https://anaconda.org/bioconda/bandage
   conda install -c bioconda bandage
   Or download and install from https://rrwick.github.io/Bandage/, Windows and Mac binaries come packaged with all necessary libraries.

2. **Visualization**
   Load the assembly graph from SPAdes in FASTG format (assembly_graph.fastg) as input. After the graph is loaded, by clicking “Draw graph” the assembly graph will be drawn to the screen.



GTDB-Tk 1.6.0

GTDB-Tk (The Genome Taxonomy Database Toolkit) is a software toolkit for taxonomic assignments of bacterial genome based on the Genome Taxonomy Database. It is used as the initial taxon identification in this workflow. The result will provide valued information for following taxonomy assignment.

Chaumeil *et al.* (2020). GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics, 36(6), 1925–1927. https://doi.org/10.1093/bioinformatics/btz848
Parks *et al.* (2021). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Research, 50: D785–D794.

1. **Run online GTDB-Tk Classify directly on KBase**
   https://kbase.us/applist/apps/kb_gtdbtk/run_kb_gtdbtk/release

2. **Input**
   GTDB-Tk requires input file as genome assembly (SPAdes contigs). All parameters are set as default.

FastANI 1.33

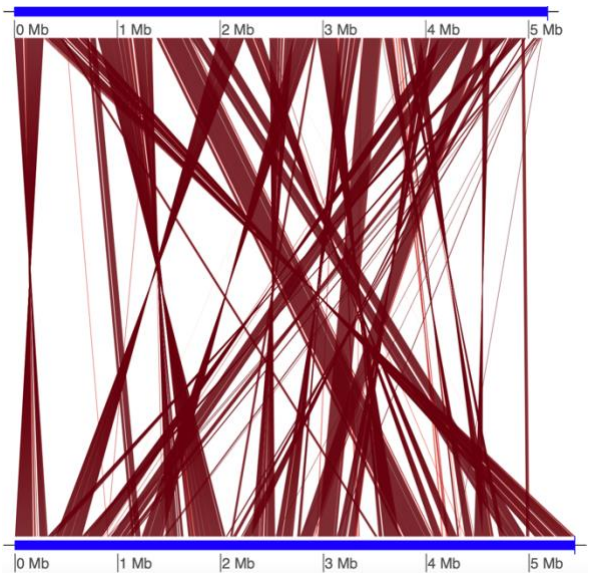
FastANI conducts a series of pairwise comparison of orthologous gene pairs shared between two microbial genomes and calculates the Average Nucleotide Identity.

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications, 9(1), 5114. https://doi.org/10.1038/s41467-018-07641-9

1. **Install package via conda**
   conda install -c bioconda fastani
   If cannot install due to environment issue, try with following command:
   conda create -n foo -c conda-forge -c bioconda fastani
   conda activate foo
   Installation instruction is available on https://github.com/ParBLiSS/FastANI

2. **Compute ANI between single query genome and multiple reference genomes**
   fastANI -q [QUERY_GENOME] --rl [REFERENCE_LIST] -o [OUTPUT_FILE]
   Because GTDB-Tk is able to identify novel isolate as *Phocaeicola* genus. The reference sequences of 13 species of the *Phocaeicola* genus are obtained from the NCBI- RefSeq database. REFERENCE_LIST should be a file (.txt) containing directory paths to all reference genomes, one per line.
   Example command:
   fastANI -q C_GC15_contigs.fasta --rl genome_list.txt -o ANI.txt

3. **Visualization**
   FastANI supports pairwise mapping of matched sequence fragments between two genomes; however, the step will require a one to one comparison with an additional parameter " --visualize".
   fastANI -q [QUERY_GENOME] -r [REFERENCE_GENOME] --visualize -o [OUTPUT_FILE]
   Then, a mapping file with .visual extension will be generated, and can be visualized by a R script "visualize.R" using genoPlotR package. (The R script will be available as a supplementary file.)
   Rscript scripts/visualize.R [QUERY_GENOME] [REFERENCE_GENOME] [MAPPING_FILE]



BLAST 2.13.0

BLAST (basic local alignment search tool) measures the nucleotide sequence similarity between a given query sequence and a database or library of sequences, then returns a list of hits that match the query sequence above a certain threshold. Therefore, can be used for precise taxonomy identification.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

1. **Install BLAST**
   For MaxOSX, download installer archive "ncbi-blast-2.13.0+.dmg" from https://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/, double click the download file and follow the instructions.

2. **Create the path for BLAST**
   `export PATH=$PATH:/usr/local/ncbi/blast/bin`
3. **Set up BLAST database**
   Get NCBI BLAST nucleotide databases with `update_blastdb.pl --decompress nt`.
   Or set up the custom BLAST database with local sequences, more information can be viewed on https://www.ncbi.nlm.nih.gov/books/NBK569841/.
   Genome sequences of 13 *Phocaeicola* species are used to create custom database.
   First to concatenate all 13 reference genomes into one fasta file:
   `cat *.fna > GC15_Database.fasta` *# "*" represents all file that end with. Fna*

   Then use the multi-FASTA file to create the custom BLAST database:
   `makeblastdb -dbtype nucl -in GC15_Database.fasta -parse_seqids -out GC15.blastdb`
   Use `makeblastdb -help` can get more detailed description of command line arguments.

4. **BLAST search against the database**
   Because the nucleotide sequence of draft assembly will be compared with the nucleotide sequences of reference genomes, the blastn function will be used. The blastn command requires at least a `-query` and `-db` option.
   Example command used for this project:
   `blastn -db GC15.blastdb -query /Users/scx/Desktop/6999/GC15/C_GC15_contigs.fasta -out C_BLAST_results.txt -outfmt 7 -max_hsps 1 -max_target_seqs 1`
   # -outfmt 7 will write the output in table form with comments
   # -max_hsps 1and -max_target_seqs 1 will only return the best alignment for each query with each subject

Genome annotation

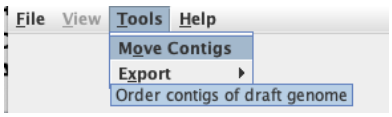
Mauve 2.4.0

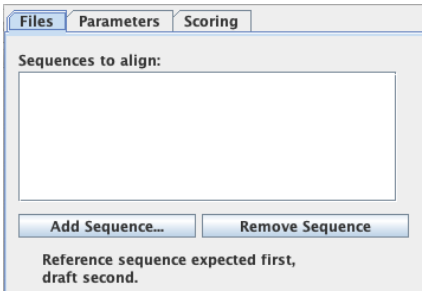
Mauve is a software package including a set of tools for reordering contigs, aligning, and comparing two or more genome sequences by searching for homologous regions.

Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Genome Research, 14(7), 1394–1403. https://doi.org/10.1101/gr.2289704

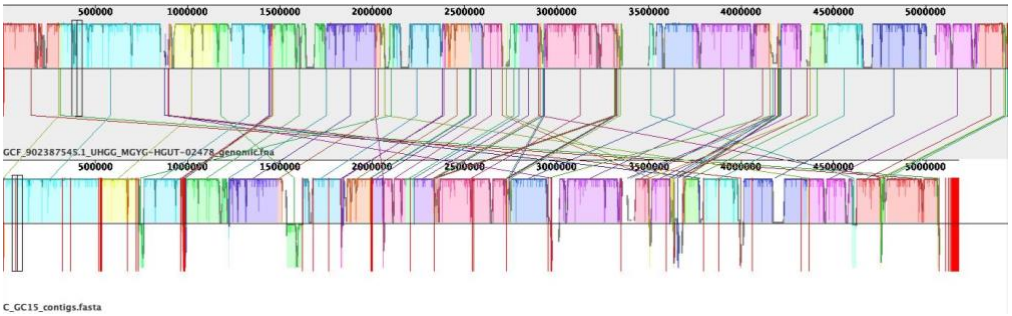
1. **Installation**
   Download the right version of Mauve development snapshots from https://darlinglab.org/mauve/download.html, and install to application.

2. **Reordering contigs and pairwise genomes alignment**
   Detailed user guide is available on https://darlinglab.org/mauve/mauve.html
   Launch the Mauve application from desktop snapshots.
   Choose "Tools" from tab and select "Move Contigs". This function will reorder the contigs position and direction using Mauve Contig Mover (MCM), as well as align the contigs against the reference genome using progressiveMauve.
   In the dialogue box, set the output files to desired directory. Add reference sequence with "Add Sequence" button first, then add draft genome (SPAdes contigs file). Click "Start" to run the reordering with default parameters.





3. **Output**
   The reordering will take several iterations. For each iteration, a new visualization window will appear. The final outputs of reordered contigs from the last iterated alignment will be saved in the subdirectory "alignmentX" with the highest X.
   Pairwise genomes alignment result is available to view. In the alignment map, upper is the reference genome, lower is the draft genome of novel isolate, blocks in the same color and connected with lines are homologous regions.



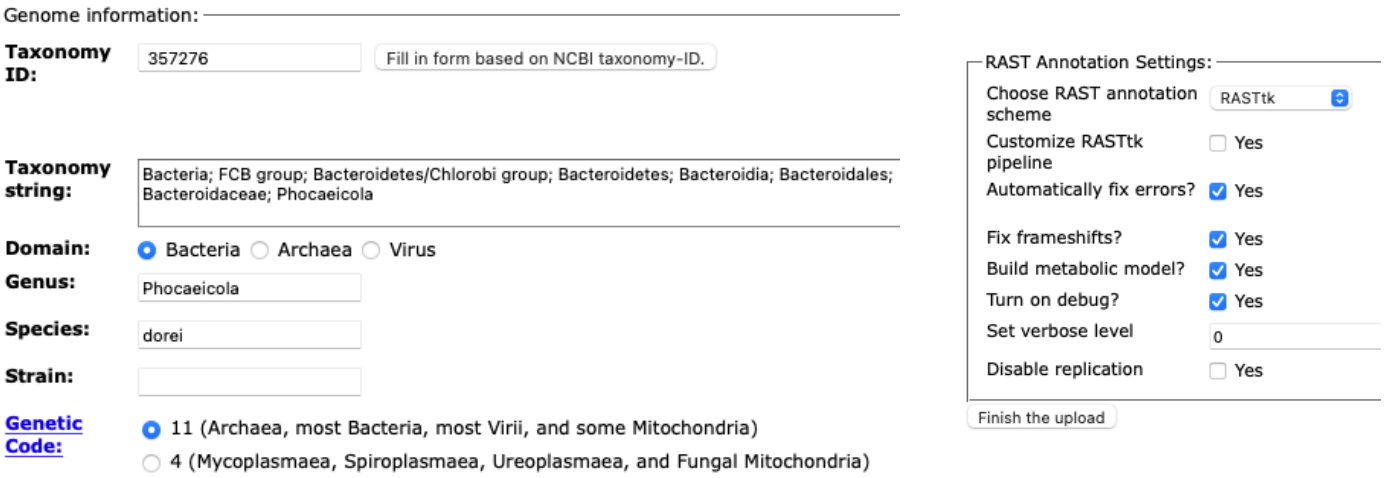
RAST Server and SEED Viewer 2.0

RAST (Rapid Annotation using Subsystem Technology) can perform the fully automated annotation for complete or nearly complete bacterial genome.

Aziz *et al*. (2008). The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics, 9(1), 75. https://doi.org/10.1186/1471-2164-9-75
Brettin *et al*. (2015). RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Scientific Reports, 5, 8365. https://doi.org/10.1038/srep08365
Overbeek *et al*. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Research, 42(Database issue), D206–D214. https://doi.org/10.1093/nar/gkt1226

1. **RAST server**
   Go to http://rast.nmpdr.org/, and register a RAST user account for the first time, then log into the account every time for annotation.

2. **Annotation**
   Select "Upload New Job" under "Your Jobs" tab, then add the draft genome in the form of a set of contigs in FASTA format (SPAdes contigs or Mauve reordered contigs) as a new job.
   On the next page, you can review the contig statistics before moving on next step. It is recommended to provide details about your bacterium in "Genome information" for a more accurate annotation. (eg. I entered the taxonomy ID 357276 for *Phocaeicola dorei* and hit "Look up taxonomy ID at NCBI" button, which would populate the rest of the fields automatically.) You can search for the right taxonomy ID on https://www.ncbi.nlm.nih.gov/taxonomy.
   On the next page, you can choose more options to finalize the RAST annotation pipeline. (For my annotation, I selected RASTtk modular, errors and frameshifts were set to fix automatically, the metabolic model was built, and the debug statements were printed.)



3. Job completion

After the job has been uploaded successfully, it may take several hours or even a day for the job to be completed. An email will be sent when the annotation is complete. Detailed annotation result contains protein-encoding genes (CDSs), RNA-encoding genes (tRNAs and rRNAs), functions for genes. Annotated result can be downloaded in any desired format (GenBank, FASTA, GFF3).

4. SEED Viewer

The finished annotation and identified subsystems can be viewed on SEED Viewer.

It also possesses other functions for genome comparison. Under “Compare” tab, you can run function-based comparison by compare the metabolic reconstruction of genes, and sequence-based comparison which will compare the protein sequence identity between two genomes.

Presence	Category	Subcategory	Subsystem	Role	Organism A	SS active A	Organism B	SS active B
A and B	Amino Acids and Derivatives	Alanine, serine, and glycine	Alanine biosynthesis	Alanine racemase (EC 5.1.1.1)	fig|357276.1422.peg.4013	yes	fig|357276.1424.peg.1881	yes
A and B	Amino Acids and Derivatives	Alanine, serine, and glycine	Alanine biosynthesis	Branched-chain amino acid aminotransferase (EC 2.6.1.42)	fig|357276.1422.peg.3543	yes	fig|357276.1424.peg.298	yes
A and B	Amino Acids and Derivatives	Alanine, serine, and glycine	Glycine Biosynthesis	2-amino-3-ketobutyrate coenzyme A ligase (EC 2.3.1.29)	fig|357276.1422.peg.2724	yes	fig|357276.1424.peg.1513	yes
A and B	Amino Acids and Derivatives	Alanine, serine, and glycine	Glycine Biosynthesis	L-threonine 3-dehydrogenase (EC 1.1.1.103)	fig|357276.1422.peg.2723	yes	fig|357276.1424.peg.1512	yes

Percent protein sequence identity

	100	99.9	99.8	99.5	99	98	95	90	80	70	60	50	40	30	20	10
Bidirectional best hit	100	99.9	99.8	99.5	99	98	95	90	80	70	60	50	40	30	20	10
Unidirectional best hit	100	99.9	99.8	99.5	99	98	95	90	80	70	60	50	40	30	20	10

export table clear all filters

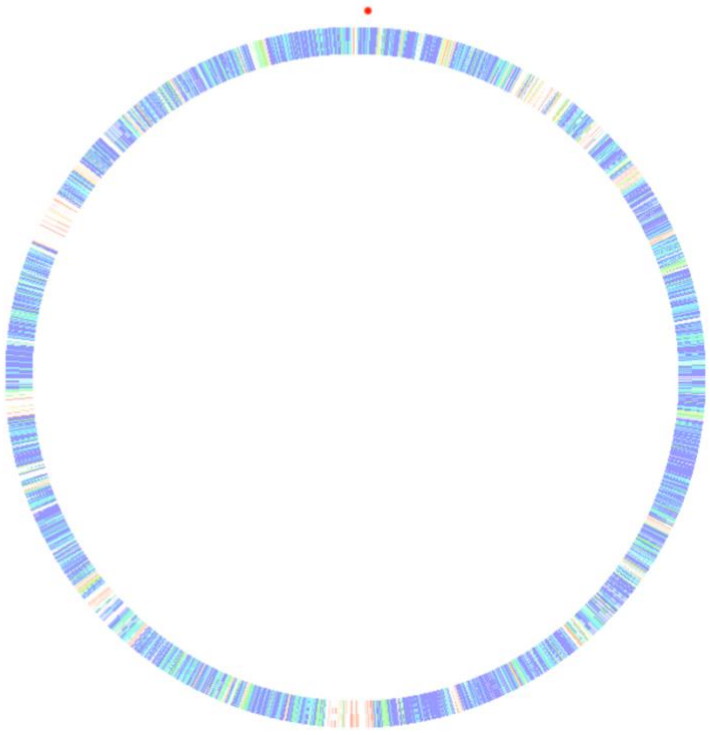
display 30 items per page

«first «prev displaying 2 - 31 of 4560 next» last»

percent identity 357276.1424 <

357276.1422			357276.1424		
Contig	Gene	Length	Hit	Contig	Gene
2	2	51	-		
3	3	51	-		
4	4	51	uni	1	658
5	5	51	uni	1	1989
6	6	991	bi	1	1991
6	7	529	bi	1	1992
6	8	418	bi	1	1993
6	9	662	bi	1	1994
6	10	636	bi	1	1995
6	11	548	bi	1	1996
6	12	718	bi	1	1997
6	13	553	bi	1	1998
6	14	440	bi	1	1999
6	15	1012	bi	1	2000
6	16	359	bi	1	2001
6	17	341	bi	1	2002
6	18	311	bi	1	2003
6	19	30	bi	1	2004
6	20	187	bi	1	2005
6	21	173	bi	1	2006
6	22	1086	bi	1	2007
6	23	703	bi	1	2008
6	24	413	bi	1	2009
6	25	237	bi	1	2010
6	26	442	bi	1	2011
6	27	480	bi	1	2012
6	28	256	bi	1	2013
6	29	119	bi	1	2014
6	30	122	bi	1	2015
6	31	92	bi	1	2016

«first «prev displaying 2 - 31 of 4560 next» last»



Prokka 1.14.6

Prokka (rapid prokaryotic genome annotation) can be used as an alternative genome annotation tool, which is available as a command-line tool on local computer.

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. Bioinformatics (Oxford, England), 30(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

1. **Install PROKKA**

https://anaconda.org/bioconda/prokka
Via Bioconda `conda install -c bioconda prokka`
If encounter installation issue, try to reinstall BioPerl 1.7.2 with `conda install -c conda-forge -c bioconda perl-bioperl=1.7.2`

2. **Annotate the assembly using prokka**

Detailed user manual is available on https://github.com/tseemann/prokka#invoking-prokka
First you will need to move to the directory of assembly contigs or reordered contigs file (FASTA format), and perform annotation with command `prokka [options] <contigs.fasta>`.
Example commands used:
`prokka Reordered_C_GC15_contigs.fasta --outdir C_prokka_reordered --prefix C_GC15 --locustag C_GC15 --kingdom Bacteria --genus Phocaeicola`

3. **Output results**

Some important output files that can be used for downstream analysis:

- .faa: The amino acid of predicted protein coding sequence (CDS)
- .ffn: The nucleotide sequences of all the prediction transcripts (CDS, rRNA, tRNA, tmRNA, misc_RNA)
- .gff: The master annotation in GFF3 format, containing both sequences and annotations. It can be viewed directly in Artemis or IGV
- .txt: The statistics relating to the annotated features found
- .tsv: A summary table of all features

4. **Some counting**

To count the total number of CDS sequences in .faa file, use
`grep '>' name.faa | wc -l`
To count hypothetical protein that been annotated, use
`grep 'hypothetical' spades_c_FAA15AN.faa | wc -l`.
Pay attention that bash is case sensitive or use `grep -i` to ignore all the cases.

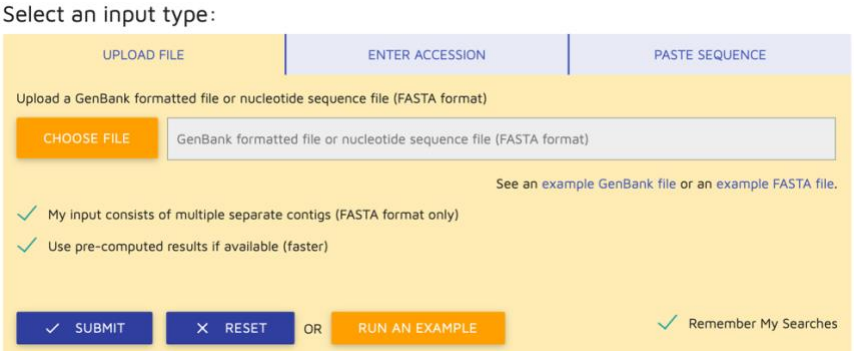
PHASTER

PASTER (PHAge Search Tool Enhanced Release) is web search tool for identification and annotation of prophage sequences within bacterial genomes.

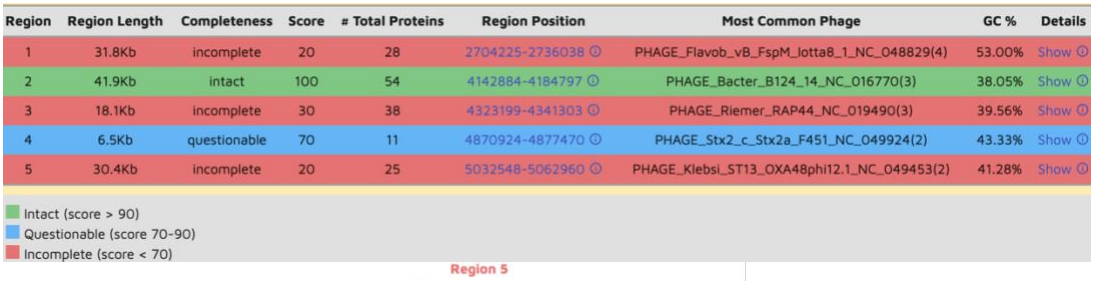
Arndt *et al.* (2016). PHASTER: A better, faster version of the PHAST phage search tool. *Nucleic Acids Research*, 44(W1), W16–W21. https://doi.org/10.1093/nar/gkw387
Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., & Wishart, D. S. (2011). PHAST: A fast phage search tool. *Nucleic Acids Research*, 39, W347-352. https://doi.org/10.1093/nar/gkr485

1. **Web server**
   Prophage annotation can be easily done on web server https://phaster.ca

2. **Create a new job**
   Input file can be nucleotide sequence (assembled contig file, reordered contigs file), or pre-annotated GenBank file (RAST server annotation result) of draft genome and reference genome downloaded from NCBI.
   Detailed instructions are listed on https://phaster.ca/instructions.

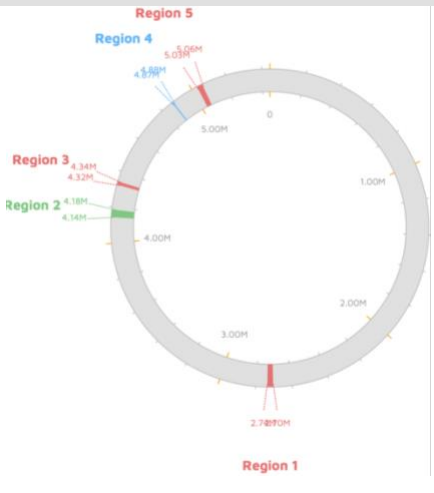


3. **Output**
   Results includes summary, details, and a genome map of identified prophage regions.



Region	Region Length	Completeness	Score	# Total Proteins	Region Position	Most Common Phage	GC %	Details
1	31.8Kb	incomplete	20	28	2704225-2736038	PHAGE_Flavob_vB_FspM_lotta8_1_NC_048829(4)	53.00%	Show
2	41.9Kb	intact	100	54	4142884-4184797	PHAGE_Bacter_B124_14_NC_016770(3)	38.05%	Show
3	18.1Kb	incomplete	30	38	4323199-4341303	PHAGE_Riemer_RAP44_NC_019490(3)	39.56%	Show
4	6.5Kb	questionable	70	11	4870924-4877470	PHAGE_Stx2_c_Stx2a_F451_NC_049924(2)	43.33%	Show
5	30.4Kb	incomplete	20	25	5032548-5062960	PHAGE_Klebsi_ST13_OXA48phi12.1_NC_049453(2)	41.28%	Show

Intact (score > 90)
Questionable (score 70-90)
Incomplete (score < 70)



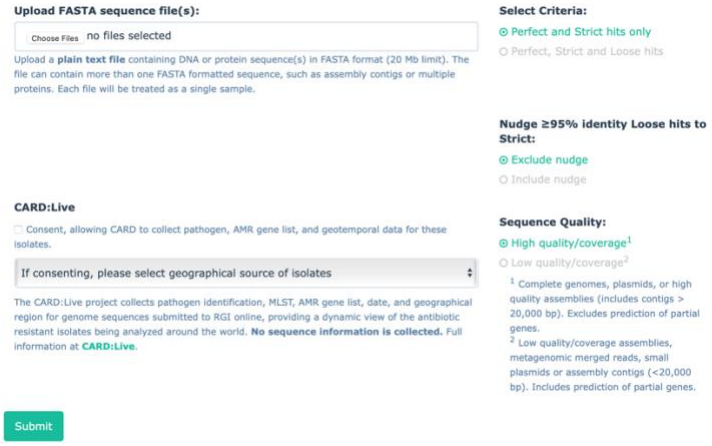
RGI 5.2.1, CARD 3.2.3

The comprehensive antibiotic resistance database (CARD) is a database that contains DNA, protein and SNPs reference sequences for bacterial antimicrobial resistance (AMR) genes. Resistance Gene Identifier (RGI) is a software that analyses and predicts resistomes using CARD's homology and SNP detection methods.

Alcock *et al*. (2019). CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Research, gkz935. https://doi.org/10.1093/nar/gkz935

1. **Web server**
   Resistance genes annotation (for file ≤ 20 Mb) can be performed on web portal https://card.mcmaster.ca/analyze/rgi

2. **Input**
   Create new job by uploading FASTA sequence file (SPAdes contigs, or Mauve reordered contigs, or nucleotide sequence of reference genome downloaded from NCBI) and selecting the right arguments.



3. **Output**
   The summary of annotated antibiotic resistance genes is available for downloading.

RGI Criteria	ARO Term	SNP	Detection Criteria	AMR Gene Family	Drug Class	Resistance Mechanism	% Identity of Matching Region	% Length of Reference Sequence
Strict	tet(X)		protein homolog model	tetracycline inactivation enzyme	glycylcycline, tetracycline antibiotic	antibiotic inactivation	99.74	100.00
Strict	ErmF		protein homolog model	Erm 23S ribosomal RNA methyltransferase	macrolide antibiotic, lincosamide antibiotic, streptogramin antibiotic, streptogramin A antibiotic, streptogramin B antibiotic	antibiotic target alteration	99.62	100.00

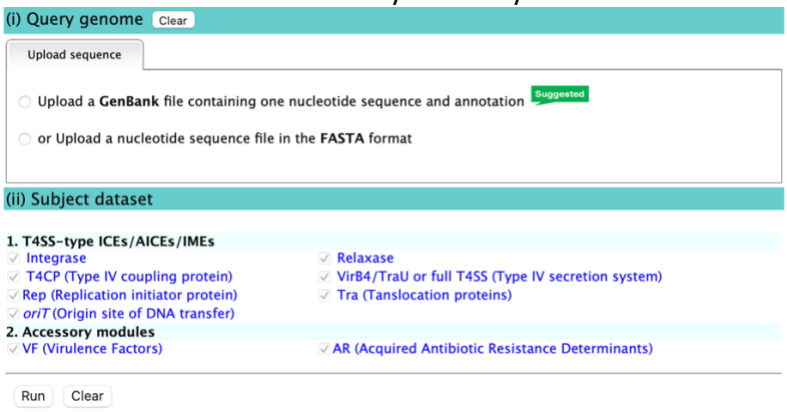
ICEfinder, ICEberg 2.0

ICEFinder is a web-based tool for rapid detection of integrative and conjugative element (T4SS-type ICEs, AICEs) and integrative and mobilizable element in bacterial genomes.
ICEberg is a database of bacterial integrative and conjugative elements.

Liu, M., Li, X., Xie, Y., Bi, D., Sun, J., Li, J., Tai, C., Deng, Z., & Ou, H.-Y. (2019). ICEberg 2.0: An updated database of bacterial integrative and conjugative elements. Nucleic Acids Research, 47(D1), D660–D665. https://doi.org/10.1093/nar/gky1123

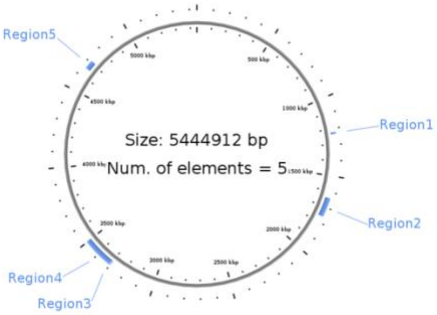
1. **Web server**
   ICEFinder web server https://bioinfo-mml.sjtu.edu.cn/ICEfinder/ICEfinder.html
   Tutorial is available on https://bioinfo-mml.sjtu.edu.cn/ICEfinder/tutorial.html

2. **Create a new job**
   You can upload the sequence and genome annotation in the GenBank format (RAST annotation, or GenBank file download from NCBI) or the nucleotide of bacterial genome in FASTA format (SPAdes contigs, or Mauve reordered contigs file) as input.
   Then select the desired dataset that fits your analysis.



3. **Output**
   Job report contains a summary tab of predicted ICE/IME regions and a genome map, by clicking the region name, you can view the detailed information of each region. The DNA sequence and proteins sequences of identified region can be download for downstream analysis.
   By recording the Job ID, you can retrieve the results any time you want.

#	Name	Location	Length/bp	Type
1	Region1	1222455..1234299	11845	Putative IME
2	Region2	1638565..1759733	121169	Putative ICE with T4SS
3	Region3	3298570..3302521	3952	Putative IME without identified DR
4	Region4	3306832..3502661	195830	Putative ICE with T4SS
5	Region5	4655711..4710700	54990	Putative IME



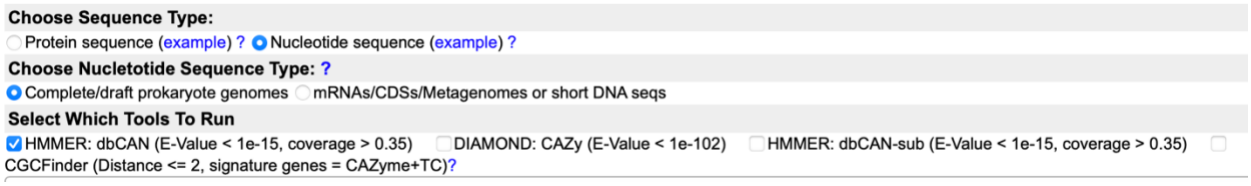
dbCAN2 meta server

It is a web server for automated Carbohydrate-active enzyme Annotation.
One of three automated CAZymes annotation tools on dbCAN2 meta server will be used is this workflow: HMMER for annotated CAZyme domain boundaries according to the dbCAN CAZyme domain HMM database.

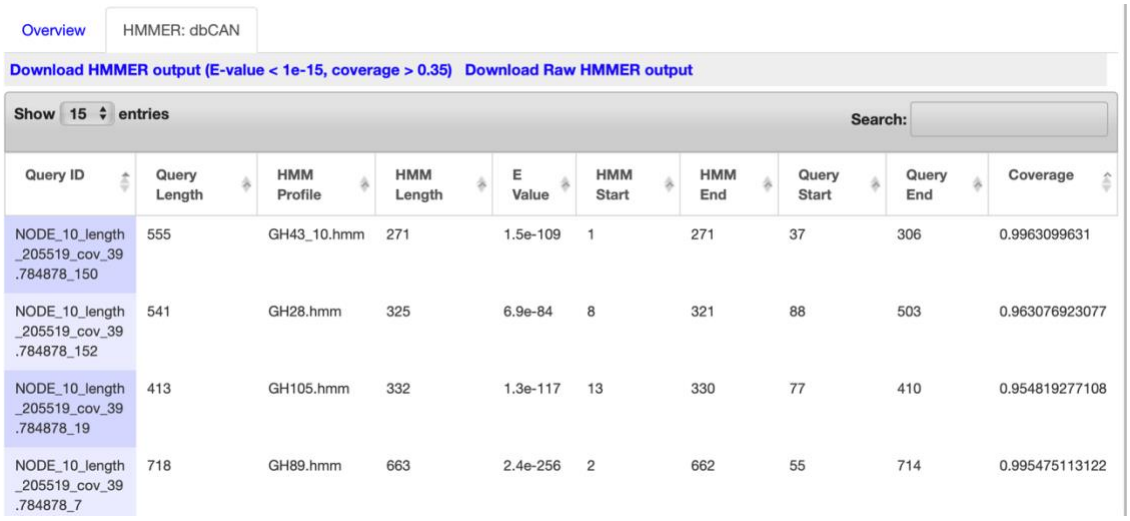
Zhang *et al.* (2018). dbCAN2: A meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Research, 46(W1), W95–W101. https://doi.org/10.1093/nar/gky418

1. **Meta server**
   https://bcb.unl.edu/dbCAN2/index.php

2. **Create a new job**
   Input sequence type can be nucleotide sequence of draft genome (SPAdes contigs, or Mauve recorded contigs) or protein sequence (generated by RAST server annotation). (For my analysis, I upload the reordered contigs nucleotide sequence as a new job.) Then select the right tool to run annotation. In this workflow, I will only use HMMER: dbCAN (E-Value < 1e-15, coverage > 0.35) for CAZymes annotation. After job is completed, annotated CAZymes can be downloaded for further analysis.
   If you wish to use all 3 tools available on dbcan2 meta server, more detailed instructions are available on https://bcb.unl.edu/dbCAN2/help.php



3. **Results**
   The "HMMER:dbCAN" tab displays the results of the HMMER run against the dbCAN database, and is available for download.



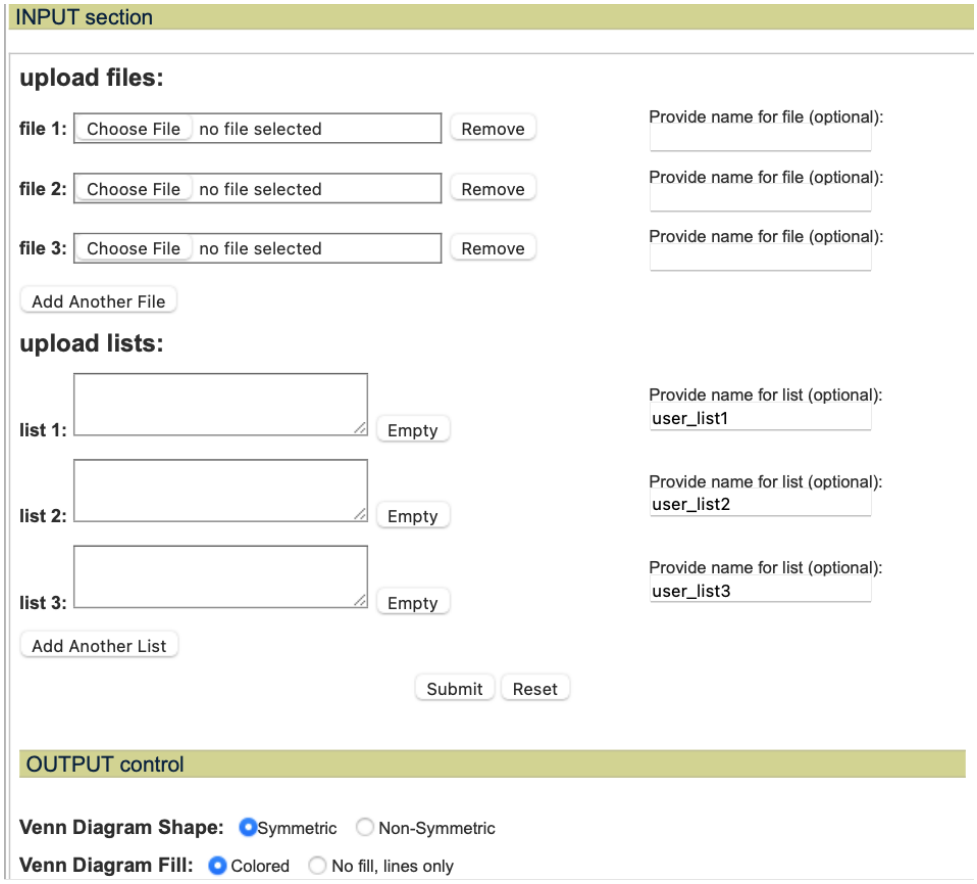
Query ID	Query Length	HMM Profile	HMM Length	E Value	HMM Start	HMM End	Query Start	Query End	Coverage
NODE_10_length_205519_cov_39.784878_150	555	GH43_10.hmm	271	1.5e-109	1	271	37	306	0.9963099631
NODE_10_length_205519_cov_39.784878_152	541	GH28.hmm	325	6.9e-84	8	321	88	503	0.963076923077
NODE_10_length_205519_cov_39.784878_19	413	GH105.hmm	332	1.3e-117	13	330	77	410	0.954819277108
NODE_10_length_205519_cov_39.784878_7	718	GH89.hmm	663	2.4e-256	2	662	55	714	0.995475113122

Comparative Genomics analysis

Venn Diagram

1. **Web server**
   https://bioinformatics.psb.ugent.be/webtools/Venn/

2. **Input**
   The input files for Venn diagram can be the specialty gene annotated in the previous part of workflow (must in plain text format). Such as the CAZymes identified from novel isolate and reference *P.dorei* strain.



3. **Output**
   The report page will generate a Venn diagram and a textual output indicating which genes are shared by two bacteria or are unique to one.

4. **R package**
   R also has package "VennDiagram" to produce high-resolution Venn and Euler plots. Reference manual is available on https://cran.r-project.org/web/packages/VennDiagram/VennDiagram.pdf

Additional software recommended for comparative analysis

OrthoVenn2 (The additional software that is recommended for comparative analysis.)
OrthoVenn is a platform for comparison and analysis of orthologous clusters from two or more genomes (up to 12).

Xu, L., Dong, Z., Fang, L., Luo, Y., Wei, Z., Guo, H., Zhang, G., Gu, Y. Q., Coleman-Derr, D., Xia, Q., & Wang, Y. (2019). OrthoVenn2: A web server for whole-genome comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Research, 47(W1), W52–W58. https://doi.org/10.1093/nar/gkz333

1. **Web server**
   OrthoVenn2 is available as a web service at https://orthovenn2. bioinfotoolkits.net
   Choose "Bacteria" for analysis.

2. Input
   Upload protein sequence file in FASTA format as input (RAST annotated protein for novel isolate, and protein sequence of reference *P.dorei* downloaded from NCBI).
   Set the e-value cut-off (1e-5) for pair-wise protein sequences similarity comparison. Set the inflation value (1.5) for generating orthologous clustering using the Markov Cluster Algorithm. Then select "Annotation", "Protein similarity", and "Cluster relationship" function.
   Presse the submit button to start the analysis. When the job has finished, the results page will be automatically loaded.

3. Output
   The output will generate a table will be an occurrence cluster table showing the results of the overlapping orthologous gene clusters and will display a Venn diagram showing the distribution of shared genes.
   The clusters are functional annotated with GO terms: biological process, molecular function, cellular component, and GO Enrichment by Uniprot and Swiss-Prot.