Assembling and Annotating Prokaryotic Genomes in KBase

Benjamin Allen*[3] (allenbh@ornl.gov), **Chris Henry**[2], **Adam P. Arkin**[1], **Bob Cottingham**[3] and the KBase Team at the following institutions

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Argonne National Laboratory, Argonne, IL; [3]Oak Ridge National Laboratory, Oak Ridge, TN; [4] Brookhaven National Laboratory, Upton, NY; [5] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

http://kbase.us

Project Goals: The DOE Systems Biology Knowledgebase (KBase) is a free, open-source software and data platform that enables researchers to collaboratively generate, test, compare, and share hypotheses about biological functions; analyze their own data along with public and collaborator data; and combine experimental evidence and conclusions to model plant and microbial physiology and community dynamics. KBase's analytical capabilities currently include (meta)genome assembly, annotation, comparative genomics, transcriptomics, and metabolic modeling. Its web-based user interface supports building, sharing, and publishing reproducible, annotated analysis workflows with integrated data. Additionally, KBase has a software development kit that enables the community to add functionality to the system.

For microbiologists, assembling prokaryotic Next-Generation Sequencing reads and constructing an annotated genome with functional information about the coding DNA sequences are necessary first steps for comparative genomics, phylogenetic analysis, or metabolic modeling. In KBase, users can quickly and easily perform de novo assembly on prokaryotic DNA reads using multiple assemblers, and then run two separate annotation tools on the assemblies, calling genes and other genomic features and assigning biological functions, to generate an annotated genome. This genome can be used in downstream analysis within KBase or downloaded for use in other software platforms.

KBase has tools for performing sequencing quality assessment and control to ensure that high quality data is fed into the assemblers. For instance, a user can use the FastQC app on Illumina sequencing reads to determine if any trimming or adapter removal is necessary for raw sequence data, and then use the Trimmomatic app to trim the low quality reads on the ends of the reads and remove any adapters.

After performing quality assessment and control, users can assemble their reads with one of the many open source *de novo* assembly apps available in KBase. Two popular assembly apps in KBase are Velvet and SPAdes. Both Velvet and SPAdes are capable assemblers for prokaryotic reads, although SPAdes has been optimized for single cell data. Velvet is a de Bruijn graph based assembler that combines multiple algorithms for the construction, simplification, and error correction of assembly graphs. SPAdes is also a de Bruijn graph based assembler unique in its utilization of multisized de Bruijn graphs to generate a final assembly from multiple *k*-mers combined with error and mismatch correction tools. Running these assembly apps on a set of

reads generates an Assembly object, which contains the assembled contigs and can be used in downstream analyses or downloaded as a FASTA file. After performing multiple assemblies on a set of sequencing reads, it can be useful to compare the performance of each assembler by analyzing the quality of each assembly. QUAST is an open source assembly quality assessment tool that allow users to compare summary statistics of multiple assemblies. These statistics can be used to determine which assembly is optimal for feeding into downstream annotation pipelines.

Once the assembly has been created, users can annotate structural and functional features on the sequence to generate a genome. Annotating the structural and functional features of an organism's genome allows for comparative analysis against other organisms' genomes and provides a reference for understanding the flow of information within a biological system. Prokaryotic genome annotation in KBase is provided by two apps: Annotate Microbial Assembly and Annotate Assembly with Prokka. Annotate Microbial Assembly takes an assembly generated by one of the assembly apps or imported by the user and runs the sequence through a multi-step pipeline to predict gene locations within the assembly and perform functional annotation using the RAST (Rapid Annotations using Subsystems Technology) toolkit. This annotation pipeline assigns functions from the SEED Subsystems Ontology to genes using a fast *k*-mer based approach. Annotate Assembly with Prokka is a KBase app that calls the popular prokaryotic annotator Prokka. Prokka combines multiple open-source annotation tools in a quick and thorough annotation pipeline for prokaryotic sequences that calls gene annotations from UniProt and RefSeq, then queries the TIGRFAM and Pfam hidden Markov model databases for domain-specific annotations.

The final output of both annotation apps is a genome that can be analyzed for specific biological features called by annotation, such as proteins or regulatory elements. The contents can be explored in a tabular genome viewer that shows summary information about the Genome as well as a list of contigs and the genes that were annotated on each contig. The genome object can be downloaded as a GenBank file or used as input to KBase apps for comparative genomics, metabolic modeling, and more.

To learn more about KBase's tools for assembly and annotation, visit http://kbase.us/assembly-and-annotation/ or try the interactive Narrative tutorial, https://narrative.kbase.us/narrative/notebooks/ws.18188.obj.6.

KBase is funded by the Genomic Science program within the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under award numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725, and DE-AC02-98CH10886.