

Annotation Pipelines

There are a few robust and well designed microbial genome annotation pipelines that you can use to analyze your genome sequences. Each has its own benefits and drawbacks, and these may dictate which pipeline you end up using.

- RAST
- PROKKA
- PATRIC
- NCBI PGAP

Creating an assembled genome to annotate

The same approach that we have talked about in other modules was used to generate a test dataset, namely, downloading fastq data from SRA and then assembling the data with spades. To summarize, these are the commands that were used.

I downloaded the reads from ERS012013, which is part of Kat Holt's *Klebsiella* dataset (Project ID PRJEB2111), and assembled them using spades.

```
fastq-dump --outdir fastq --gzip --skip-technical --readids --read-filter pass --dumpbase -  
spades.py -o assembly -1 fastq/ERS012013_pass_1.fastq.gz -2 fastq/ERS012013_pass_2.fastq.gz
```

Statistic	Value
Number of sequences	4271
Total length	8,716,579
Shortest contig	56
Longest contig	86,652
N50	11,152
N75	28,825

In all the cases below we use the `scaffolds.fasta` output from spades for subsequent analysis.

Example annotation using RAST

Start at the RAST website and from **Your Jobs** choose **Upload a new Job**. This opens up the file chooser page, and at the file chooser

select your `scaffolds.fasta` file. After that file is uploaded, you are presented with a summary of the contigs. Note that RAST may split some of the scaffolds that spades generated, and thus you may have slightly more contigs and slightly shorter sequence size, as shown here. The split happens on runs of `N` bases that

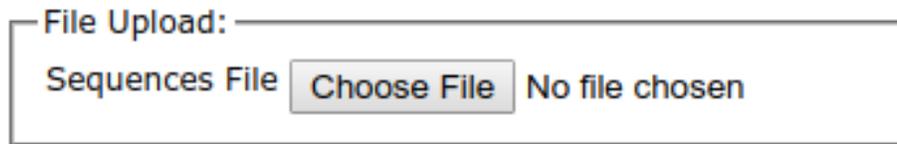


Figure 1: file chooser

spades inserts where it can estimate gaps between contigs based on sequence overlap.

Contig statistics

Statistic	As uploaded	After splitting into scaffolds
Sequence size	8716579	8654096
Number of contigs	4271	6027
GC content (%)	57.1	57.1
Shortest contig size	56	6
Median sequence size	571	374
Mean sequence size	2040.9	1435.9
Longest contig size	86652	86652
N50 value	11152	10463
L50 value	176	185

Figure 2: RAST summary statistics

The bottom of this page asks for information about the organism you have sequenced. If you enter the taxid, as shown here, the form should populate with information from NCBI.

There a series of questions about the annotation pipeline. Two recommended options are to build metabolic models and fix frameshifts, especially if you have a draft genome. Fixing frameshifts is controversial because some genomes (notably *Salmonella enterica serovar Typhi*) have a large number of frameshifts that are an evolutionary trait!

Note: at this stage you can also choose to customize some of the options for the RAST pipeline.

Genome information:

Taxonomy ID: Fill in form based on NCBI taxonomy-ID.

Taxonomy string:

Domain: Bacteria Archaea Virus

Genus:

Species:

Strain:

Genetic Code: 11 (Archaea, most Bacteria, most Virii, and some Mitochondria)
 4 (Mycoplasmata, Spiroplasmata, Ureoplasmata, and Fungal Mitochondria)

Use this data and go to step 3

Figure 3: RAST Job Information

Please consider the following options for the RAST annotation pipeline:

RAST Annotation Settings:

<p>Choose RAST annotation scheme <input type="text" value="RASTtk"/></p> <p>Customize RASTtk pipeline <input type="checkbox"/> Yes</p> <p>Automatically fix errors? <input checked="" type="checkbox"/> Yes</p> <p>Fix frameshifts? <input checked="" type="checkbox"/> Yes</p> <p>Build metabolic model? <input checked="" type="checkbox"/> Yes</p> <p>Turn on debug? <input type="checkbox"/> Yes</p> <p>Set verbose level <input type="text" value="0"/></p> <p>Disable replication <input type="checkbox"/> Yes</p>	<p><i>Choose "RASTtk" for the current modular customizable production RAST pipeline, or "Classic RAST" for the old pipeline.</i></p> <p><i>Customize the RASTtk pipeline</i></p> <p><i>The automatic annotation process may run into problems, such as gene candidates overlapping RNAs, or genes embedded inside other genes. To automatically resolve these problems (even if that requires deleting some gene candidates), please check this box.</i></p> <p><i>If you wish for the pipeline to fix frameshifts, check this option. Otherwise frameshifts will not be corrected.</i></p> <p><i>If you wish RAST to build a metabolic model for this genome, check this option.</i></p> <p><i>If you wish debug statements to be printed for this job, check this box.</i></p> <p><i>Set this to the verbosity level of choice for error messages.</i></p> <p><i>Even if this job is identical to a previous job, run it from scratch.</i></p>
--	--

Finish the upload

Figure 4: RAST Pipeline

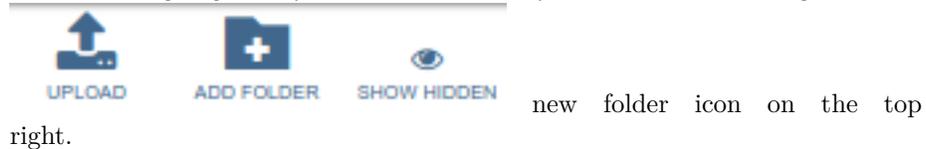
Example annotation using PROKKA

Note: The PROKKA GitHub Site contains many other recipes and advances options for annotating the `scaffolds.fasta` file using PROKKA.

Example submission using PATRIC

To annotate the contigs using PATRIC, I first go to the PATRIC website and log in. If you don't have an account you will need to create one.

Create a new workspace called `Klebsiella` by clicking on the `Workspaces` menu and going to your `home` directory, and then clicking on the



Then use the `p3` commands to submit the `scaffolds.fasta` file for annotation as a genome. You will need to follow these installation instructions to install the `p3` commands, and at the moment they do not provide a CentOS version so it they are not included on the AWS instance.

Once you have installed `p3`, you will need to login:

```
p3-login
```

and provide the same credentials that you use for the website.

For the command, we need to provide several variables:

Variable	Definition
<code>-c</code>	the contigs source file of the contigs (probably scaffolds.fasta from spades output)

Variable	Definition
----------	------------

-n	the name we want to use for our genome
----	--

Variable Definition

-t the
NCBI
Tax-
on-
omy
ID.
For
*Kleb-
siella
pneu-
mo-
niae*
this
is
573.
This
is
used
to
en-
sure
that
the
cor-
rect
pa-
ram-
e-
ters
are
used
for
the
an-
no-
ta-
tion
processes.

Variable	Definition
-d	the domain (Bacteria, Archaea, Eukarya, or Virus)

Then we provide the workspace and the file name to call it in the workspace.

```
p3-submit-genome-annotation --contigs-file scaffolds.fasta -n "Klebsiella pneumoniae NT2114"
```