Bioinformatics Lab Guide 2:

File Naming/Genome Assembly & Quality

James B. Herrick, Ph.D.
James Madison University

Background

- NCBI primer on genome assembly

Assembly using SPAdes (GalaxyTrakr)

1. Make sure your history that contains the read files for **HJ1_SRR5886281** is visible.

2. In the (left) Tools pane, click on “NGS Assembly” and then on the SPAdes tool.

3. Parameters and file selection:

    - Make sure “Run only assembly? (without read error correction)” is clicked ‘No’ (this should be the default)

    - **Turn off ‘Careful Correction’** (click ‘No’)

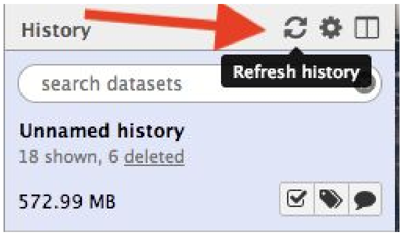
    - Under "K-mers to use, separated by commas” **make sure the following numbers are entered: 21,33,55,77,99,127**

    - Skip down to “Files”. Make sure “separate input files” is selected under ‘Select file format’ (Should be default). Select your forward and reverse fastq files under ‘Forward reads’ and ‘Reverse reads’. The order makes no difference but make sure one is the forward and the other the reverse file (in this example, **SRR5886281_1.fastq** and **SRR5886281_2.fastq**).

    - Near the bottom, click ‘Yes’ on “Output final assembly graph (contigs)?”

4. Hit “Execute”. Assembly can typically take 1 or more hours.

    - Hit the ‘refresh’ icon to monitor progress. The color of output files will change from gray (preparing to upload) to yellow (uploading) to green (uploading complete).



○

How to Name Your Files

We're still (and perhaps always will be) in the process of figuring out the best way to name the myriad kinds of files we generate. First and foremost, **be very careful when re-naming any data files**. For example, for fastq sequence files, keep the VA-WGS or CFSAN or SRR unique numbers. If they're our data, add the short lab strain designation (e.g. "HJ28") in front of the VA or CFSAN (etc.) numbers. Keep other output designations. For example:

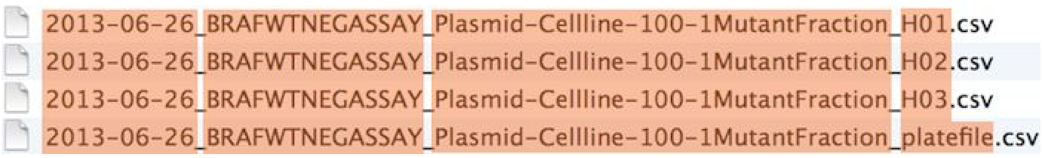
- HJ28_VA-WGS-17106_S8_L001_R1_001.fastq.gz

Here are some general conventions for good file naming. These come from this presentation by Dr. Stephen Turner, our collaborator from the University of Virginia bioinformatics core. Remember, too, that the repository for our data and analysis files is our Open Science Framework page(s).. OSF also has a page with some good 'best practices' in file naming that mirror Dr. Turner's below.

Your file names need to:

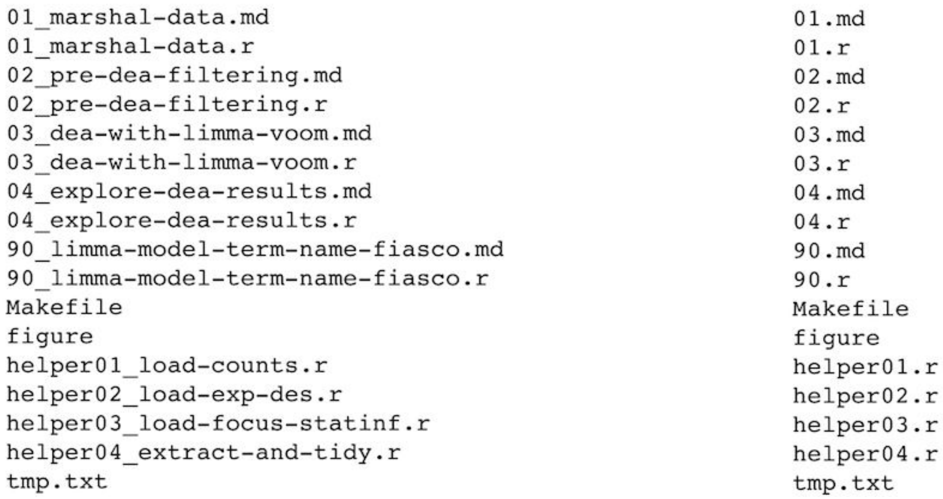
1. Be machine readable
2. Be human readable
3. 'Play well' with default ordering (sorting) schemes

1) For machine readability:
   - ○ Avoid spaces, punctuation, accented characters, case sensitivity
   - ○ Use "-" and "_" to allow recovery of meta-data from the filenames:
     - i) "_" underscore used to delimit units of meta-data I want later
     - ii) "-" hyphen used to delimit words so "my eyes don't bleed" (Turner)



```
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_platefile.csv
```

2) For human readability:
   - ○ **Make sure the name contains understandable info on the content of the file**
   - ○ See above for use of "-" and "_"
     As Dr. Turner says, which filenames do you want at 3 am before a deadline, the ones on the left or the ones on the right?



```
01_marshal-data.md                          01.md
01_marshal-data.r                           01.r
02_pre-dea-filtering.md                     02.md
02_pre-dea-filtering.r                      02.r
03_dea-with-limma-voom.md                   03.md
03_dea-with-limma-voom.r                    03.r
04_explore-dea-results.md                   04.md
04_explore-dea-results.r                    04.r
90_limma-model-term-name-fiasco.md          90.md
90_limma-model-term-name-fiasco.r           90.r
Makefile                                    Makefile
figure                                      figure
helper01_load-counts.r                      helper01.r
helper02_load-exp-des.r                     helper02.r
helper03_load-focus-statinf.r               helper03.r
helper04_extract-and-tidy.r                 helper04.r
tmp.txt                                     tmp.txt
```

3) For ordering schemes:
   - If possible, put something numeric first
   - Use the ISO 8601 standard for dates, e.g. 2018-02-28 (yyyy-mm-dd)
   - Left pad other numbers with zeros, e.g. 01, 02, not 1, 2.

4) **General model for our <u>analysis</u> files:**

   yyyy-mm-dd_isolate-name_analysis-or-other-identifiers.suffix

   <u>Example</u>: 2018-03-19_HJ23_spades-assembly1.fasta

Checking your assembly quality using QUAST (GalaxyTrakr)

Note: can also be done directly at the QUAST website.

1. Click on ‘NGS: Assembly’ and ‘QUAST’ (you don’t have to wait for your assembly to complete to execute this tool).
2. Select your “Contigs/scaffolds output file” by choosing the ‘contigs’ file. A typical name will be “SPAdes on data 2 and data 1: contigs (fasta)”
3. Leave all other parameters as ‘default’ and hit ‘Execute’.

Click on either the .html or .tsv file and compare the outputs to the quality guidelines listed below under ‘Sequencing and Assembly Quality’ and particularly the CDC assembly thresholds for Salmonella. For Salmonella, the “GC(%)” should be 52% and the “total length” around 4.7 mb.

Sequencing and Assembly Quality Standards

General definitions and guidelines for sequencing and assembly quality from *ENGAGE*, the EU Whole Genome Sequencing training consortium:

QC parameters	**Description**
Number of reads	The number of reads refers to the sequence yield, how much was sequenced.
Average read length	The average length of all the reads and is measured in bp.
Depth of coverage, total DNA sequence	Number of bps sequenced divided by the total size (both chromosome and plasmids) of the closed genome (same strain). This number can be rounded to the nearest integer. In essence this number describes the number of times the sequenced bps covers the reference DNA and is often ended with an “x” (e.g. 30x). Coverage greater than 30X indicates good quality.
Size of assembled genome	The total size of all the contigs in bp. The total bp should correspond to the size of the sequenced genome [ca. 4.8 kbp for *Salmonella*].
Total number of contigs greater than 500 bp	The total number of contigs greater than 500 bp. A number of contigs less than 500 contigs normally recommend as good quality.
N50	the N50 length is defined as the length for which the collection of all contigs of that length or longer contains at least half of the sum of the lengths of all contigs, and for which the collection of all contigs of that length or shorter also contains at least half of the sum of the lengths of all contigs. A N50 more than 30 Kb normally indicate good quality.

CDC Assembly thresholds for *Salmonella* (most important in bold):

- Mean Q (Phred score): >30
- Mean depth of coverage: >30x
- N_{50}: >200,000
- Number contigs: **<200**
- Sequence length: **~4.4-5Mbp**

Visualizing your assembly using Bandage

Bandage is a program for visualising *de novo* assembly graphs. It is run locally on your computer and can be downloaded for Mac or Windows (or Linux) here.

1. The input file for Bandage is the assembly graph generated by SPAdes (for example, "**SPAdes on data 2 and data 1: assembly graph**"). Download this file from GalaxyTrakr by clicking on the (underlined) name of the file in the History pane and then clicking on the disk icon [💾] to download it to your computer. Follow the lab file naming conventions above for naming your file.

2. Open up Bandage on your computer. Click on 'File', then 'Load Graph'. Load your assembly graph file. You will see something like this:

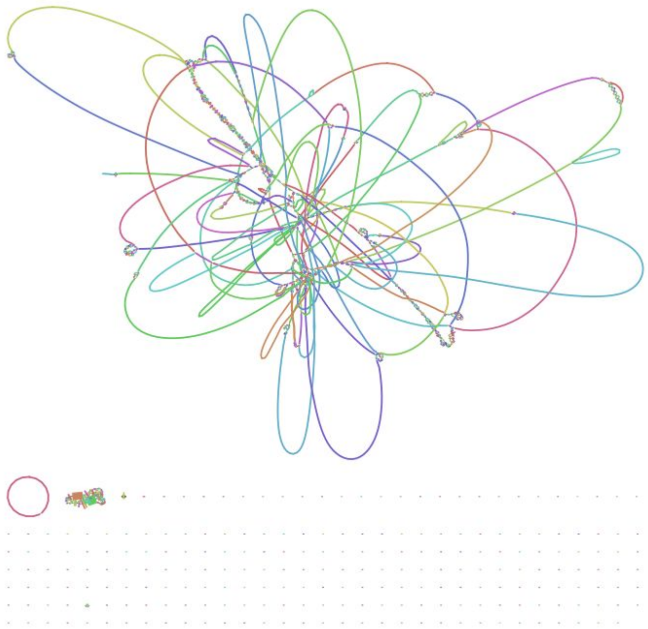


Fig. 1. Bandage output of a SPAdes assembly of HJ27 without trimming or filtering.

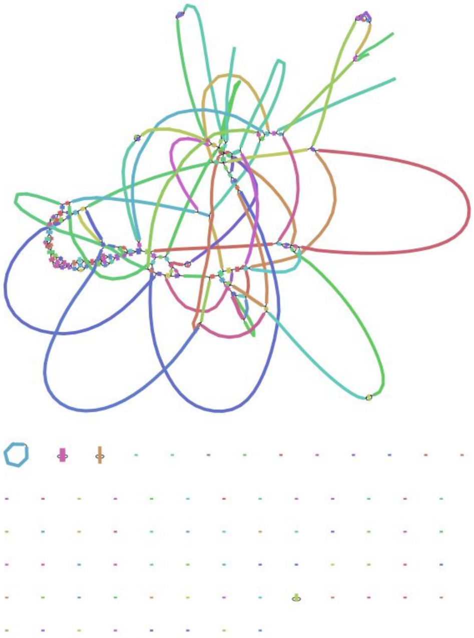


Fig. 2. Bandage output of a SPAdes assembly of HJ27 with SLIDINGWINDOW 4:20 filtering (see below)

Each of the colored lines or bands is a “node” or contig, a region of the chromosome assembled by SPAdes from the reads. The nodes (the colored lines) connect at repeat sequences that are longer than the read length of the sequencer (in the example for HJ27 above, 250 bp). Ideally this figure would be a circle but that essentially never occurs with an assembly from only short reads. Notice in this example the small circle at the bottom (red in the Fig. 1, blue in Fig. 2), a possible plasmid.

Note, too, that in figure 2 there are fewer contigs and just generally a less complex graph compared to figure 1, indicating a possibly better assembly (due to filtering; see below).

You can click on a node, then on ‘Output’ and ‘Web BLAST selected nodes’ to look for similar BLAST regions. In this case, the possible plasmid above (from HJ27) indeed maps 100% to this Salmonella plasmid which, by the way, was quite surprising.