Gonococci assembly lab journal

Darya Nemirich

March 2020

Preliminary preparation

Conda-environment creation. All necessary tools and libraries will be uploaded here.

```
conda create -n gonococci_assembly python=3.8

conda install -c bioconda fastqc
conda install -c bioconda/label/cf201901 spades
conda install -c bioconda prokka
conda install -c conda-forge biopython
conda install -c biobuilds perl=5.22
conda install -c bioconda bbmap
conda install -c conda-forge perl-text-soundex
conda install -c bioconda -c conda-forge multiqc
conda install -c bioconda/label/cf201901 entrez-direct

# There were problems with tbl2asn. The newest version was downloaded
wget ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/converters/by_program/tbl2asn/linux64.tbl2asn.gz
gunzip linux64.tbl2asn.gz
mv linux64.tbl2asn /home/is2/.conda/envs/gonococci_assembly/bin
chmod 755 /home/is2/.conda/envs/gonococci_assembly/bin/linux64.tbl2asn



conda create -n assembly_qc python=3.6
conda install -c bioconda quast
conda install -c bioconda fastani
conda install -c bioconda -c conda-forge busco=4.0.5
conda install -c bioconda roary
conda install -c bioconda gblocks
conda install -c bioconda snippy


conda create --name rgi rgi=4.2.2
```

Reads quality check and trimming

Scripts: `reads_qc.sh`, `reads_trimming.sh`
QC-reports and multiQC report for raw reads are in "/home/is2/internship/scripts" directory
Parameters for trimming:

- minimal average quality (maq) - 20
- k-mer trimming (ktrim) - r (from the right side)
- k-mer length (k) - 21
- minimal k-mer size (mink) - 11
- allowed number of mismatches (hdist) - 2
- quality trim threshold (trimq) - 20
- quality trim to Q20 from both ends (qtrim) - rl
- force-trimming of 20 leftmost (ftl) - 20
- read trimming to the same length (tpe)
- adapter trimming based on pair overlap detection (tbo)

95% of the library is remained (Before - 33,2 M of reads, After - 31,6 M of reads).
Trimming log-file (/home/is2/internship/log_trimming.txt).

Trimmed reads are in "/home/is2/internship/trimmed" directory. QC reports and multiQC report are in "/home/is2/internship/trimmed_qc" directory.

Assembly step

Assembly was performed with default Spades parameters, which include error-correction step, automatic selection of k-mer length. Also "careful" option was applied. Preliminary assembly QC was performed without reference. Assembly quality check was done with QUAST.
Scripts: `assembly.sh`, `assembly_qc.sh`

In order to perform assembly QC and assembly with better quality, all reference genome assemblies for G.n were gotten from NCBI Assembly database (script: get_references.sh) Then, two lists with references and assemblies paths were made. References were compared with assemblies, in order to find the closest to the assembly reference genome.

```
fastANI --ql fastANI_results/query.txt --rl fastANI_results/references.txt -o fastANI_results/fast
```

Script **fastANI_search.sh** The resulting list has been sorted in descending order of ANI value in RStudio. The best hits for each isolate were found. They were two references: "GCA_000695425.1_NG-i19.05" (for samples 1, 3, 12, 14, 16) and "GCA_003428775.1_ASM342877v1" (for others). Assembly QC procedure was repeated with appropriate references (**assembly_qc_with_reference.sh**). The metrics seemed satisfactory to me. I have executed quality checks of assemblies with different k-mer length, in order to ensure that Spades have chosen the best variant.

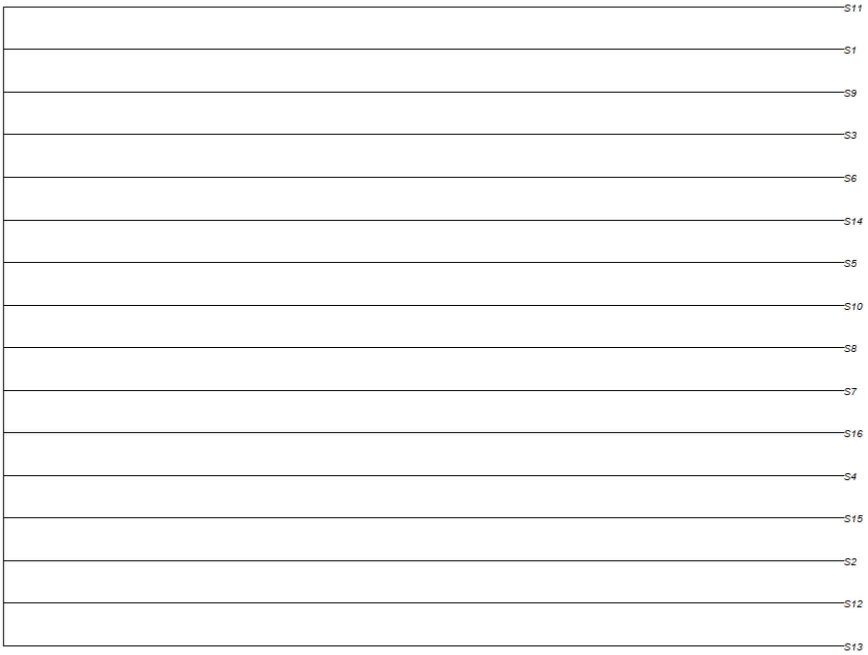
Then another quality check round was performed with BUSCO software. I have downloaded the latest BUSCO database of universal single-copy orthologs for N.g. Scripts: **busco_plots.sh**, **busco_qc.sh**
All in all, there were not that many missing genes. And others were assembled correctly.

Annotation

For assembly annotation step, I have chosen Prokka software. In order to annotate with Prokka more precisely, I've created the own protein database. It consists of identical protein groups proteins, downloaded from here https://www.ncbi.nlm.nih.gov/ipg/?term=neisseria+gonorrhea. Script: **annotation.sh** There are still some unidentified proteins left after the annotation process. Notably, no CRISPR loci were identified by Prokka. I intended to try RepeatModeler to find functional repeats and possible transposons, but it would have taken too much space on the server with all its temporary files.

Variant calling

First, in order to find a reference genome for variant calling procedure, I've performed FastANI comparison of isolates with themselves, to get a distance matrix to build an NJ-tree. However, isolates turned out to be very close to each other, and the obtained results lacked in resolution to build an appropriate tree.



For this reason, I have decided to build a core genomes and perform an alignment. It was done with Roary (**roary_launch.sh**). Gblocks with default parameters was used for blocks extraction. Tree was build using FastTree with GTR model

```
FastTree -nt -gtr < internship/core_genomes_1584498228/core_gene_alignment.aln-gb > my_tree.newick
```

The tree was visualized with FigTree. Variant calling was performed with Snippy (**snp_calling.sh**)

Virulence factors determinants and antibiotics resistance genes search

The latest version of the virulence factors database was downloaded (http://www.mgc.ac.cn/VFs/download.htm). Only N.g virulence factors were extracted from this database (**internship/vfs_search/neisseria_extraction.py**). Blastp search was performed. Best hits were filtrated and than annotated again, in order to determine factors that were found. Scripts: **vfs_search.sh**, **hits_annotation.py**, **hits_annotation_launch.sh**. All isolates are predisposed to have the same virulence factors.

To determine antibiotics resistance determinants, I have tried abricate tool. The results, however, were not satisfying: no significant hits were found. I suppose, it can be explained by a small annotation database. Then I have downloaded the latest version of RGI database (https://card.mcmaster.ca/latest/data) and performed RGI search (**rgi_launch.sh**)

All scripts names, directory structure and programs versions are listed in the report section "Additional information".