# Single Amplified Genome Assembly

**Tyler Alioto**

## Abstract

**Citation:** Tyler Alioto Single Amplified Genome Assembly. **protocols.io**
dx.doi.org/10.17504/protocols.io.nccdasw
**Published:** 23 Mar 2018

## Before start

Software pre-requisites:

Spades

SGA

cutadapt

trim_galore

Augustus

BUSCO

Data sources:
  http://spades.bioinf.spbau.ru/spades_test_datasets/ecoli_sc/

  https://www.nature.com/articles/s41598-017-05436-4
Get from the ENA if possible.
  https://www.ebi.ac.uk/ena/data/view/PRJDB5352
  https://www.ncbi.nlm.nih.gov/bioproject/PRJDB5352

E.coli reference
  https://www.ncbi.nlm.nih.gov/nuccore/NC_000913

## Protocol

### Step 1.

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Download datasets (or generate them if doing a real experiment!)

DATASET

**E. coli reference genome**

**Step 2.**

Run pre-QC (SGA)

DATASET

**SPAdes E. coli (single cell)**

cmd COMMAND

```
cat ecoli_mda.preqc | perl -
e 'while(<>){if (/FragmentSize/){my $line = <STDIN>;while(my $line=<>){if($line=~m/(\d+)/){
print '$1\n';}else{exit;}}}}' > ecoli_mda.fragsizes

module load R
R
#> frag<-read.table('ecoli_mda.fragsizes',header=F)
#> summary(frag)
#       V1
# Min.   :  52.0
# 1st Qu.: 257.0
# Median : 281.0
# Mean   : 259.9
# 3rd Qu.: 301.0
# Max.   :1246.0
#> quit()
```

Check insert size in R

**Step 3.**

Adapter and Quality Trim (trim_galore)

cmd COMMAND

```
trim_galore --gzip -q 10 --paired --retain_unpaired reads.1.fastq.gz reads.2.fastq.gz
```

Trim Illumina adapters and quality trim 3' bases less than Phred Q10

**Step 4.**

Determine composition / Identify genome if known (Kraken) if not already know from 16S RNA sequencing.

cmd COMMAND

```
kraken --db  /path/to/kraken_db/kraken_abfpv_21_08_2016 --threads 4 --preload --fastq-
input --gzip-compressed --paired --quick --only-classified-output --
output ecoli_sdmda.kraken.out $ASSEMBLY/data/ecoli_sdmda.100k.1.fastq.gz $ASSEMBLY/data/eco
li_sdmda.100k.2.fastq.gz
```

**Step 5.**

Run SPAdes (optionally run ccSAG to do co-assembly of SAGs -- includes QC step 2)

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

SOFTWARE PACKAGE (Linux)

**ccSAG**

https://github.com/mstkgw/ccSAG

cmd COMMAND

```
mkdir -p $ASSEMBLY/sags/spades_ecoli_mda
cd $ASSEMBLY/sags/spades_ecoli_mda
# IMPORTANT! make sure to turn on the single-cell mode with --sc
# Optionally add --careful
# Loading the environment with module load in the job script itself ensures the proper opti
mizations are used depending on which partition the job is run.
module purge; module load gcc/4.9.2 SPADES/3.5.0
spades.py --only-assembler --
sc -1 ${ASSEMBLY}/data/ecoli_mda.1.fastq.gz -2 ${ASSEMBLY}/data/ecoli_mda.2.fastq.gz -t 4 -
o ecoli_mda_spades

mkdir -p $ASSEMBLY/sags/spades_ecoli_sdmda
cd $ASSEMBLY/sags/spades_ecoli_sdmda
module purge; module load gcc/4.9.2 SPADES/3.5.0
spades.py --
sc -1 ${ASSEMBLY}/data/ecoli_sdmda.1.fastq.gz -2 ${ASSEMBLY}/data/ecoli_sdmda.2.fastq.gz -
t 24 -o ecoli_sdmda_spades

mkdir -p $ASSEMBLY/sags/spades_bsubtilis_sdmda
cd $ASSEMBLY/sags/spades_bsubtilis_sdmda
module purge; module load gcc/4.9.2 SPADES/3.5.0
spades.py --
sc -1 ${ASSEMBLY}/data/bsubtilis_sdmda.1.fastq.gz -2 ${ASSEMBLY}/data/bsubtilis_sdmda.2.fas
tq.gz -t 24 -o bsubtilis_sdmda_spades

mkdir -p $ASSEMBLY/sags/spades_soil_sdmda
cd $ASSEMBLY/sags/spades_soil_sdmda
module purge; module load gcc/4.9.2 SPADES/3.5.0
spades.py --
sc -1 ${ASSEMBLY}/data/soil_sdmda.1.fastq.gz -2 ${ASSEMBLY}/data/soil_sdmda.2.fastq.gz -
t 24 -o soil_sdmda_spades
```

## Step 6.

Determine completeness (CheckM and/or BUSCO)

cmd COMMAND

```
# Assume you have putative genomes in the directory /home/donovan/bins with fa as the file
extension and want to store the CheckM results in /home/donovan/checkm. To processes these
genomes with 8 threads, simply run:

checkm lineage_wf -t 8 -x fa /home/donovan/bins /home/donovan/checkm
```

CheckM

## Step 7.

(Optional) Align to Reference (MUMMER)

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

SOFTWARE PACKAGE (Linux)

**Mummer**

https://github.com/mummer4/mummer

cmd COMMAND

```
dnadiff -p query.v.ref ecoli_ref_NC_000913.3.fasta ecoli_mda_spades.fasta
mummerplot --png out.delta
```

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited