Assembly Procedure Applied to TARA Oceans Data (Ex. North Pacific)

Benjamin Tully

Abstract

The workflow applied to Tara Oceans raw sequence data for generating assemblies suitable for genomic binning.

Used in:

"290 Metagenome-assembled Genomes from the Mediterranean Sea: Ongoing Effort to Generate Genomes from the Tara Oceans Dataset" - bioRxiv https://doi.org/10.1101/069484

"Undocumented potential for primary productivity in a globally-distributed bacterial photoautotroph" - submitted

Citation: Benjamin Tully Assembly Procedure Applied to TARA Oceans Data (Ex. North Pacific). **protocols.io** dx.doi.org/10.17504/protocols.io.hfqb3mw
Published: 27 Mar 2017

Protocol

Step 1.

cmd COMMAND
```
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR599/ERR599052/*
```

Assemble sequences from the same filter fraction and depth together with Megahit

Step 2.

Ex. Sample TARA138 for the prostistan size fraction (0.8-5.0 µm) collected at the mesopelagic (450-m) which consists of 3 sequence libraries:

ERR1712000 - 190,613,172 reads

ERR1712169 - 48,680,964 reads

ERR868467 - 359,731,894 reads

Total = 599,026,030 PE reads

cmd COMMAND - 1.0.3
```
megahit --presets meta-
sensitive -1 ERR1712000_1.fastq.gz,ERR1712169_1.fastq.gz,ERR868467_1.fastq.gz -2 ERR1712000
_2.fastq.gz,ERR1712169_2.fastq.gz,ERR868467_2.fastq.gz -o tara138_prot_meso.megahit_asm
```

Megahit is available here: https://github.com/voutcn/megahit

EXPECTED RESULTS

Megahit output:

3177295 contigs, total 2172136004 bp, min 200 bp, max 520054 bp, avg 684 bp, N50 746 bp

Repeat step for all samples (station, size fraction, depth)

Step 3.

15 total assemblies for North Pacific stations

Combine all "final.contigs.fa" in to the PRIMARY contig file

Step 4.

Generates tara_northpacific_PRIMARY_contigs.fasta

cmd COMMAND

```
cat */*_asm/final.contigs.fa > tara_northpacific_PRIMARY_contigs.fasta
```

EXPECTED RESULTS

41,167,824 contigs

Size select contigs ≥2kb in length

Step 5.

cmd COMMAND - 0.6.1

```
seqmagick convert --min-
length 2000 tara_northpacific_PRIMARY_contigs.fasta tara_northpacific_PRIMARY_min2000_conti
gs.fasta
```

Seqmagick is available here: http://seqmagick.readthedocs.io/en/latest/

EXPECTED RESULTS

1,231,780 contigs

Longest scaffold 1091714

Number of scaffolds > 10K nt 55757 4.5%

Number of scaffolds > 100K nt 821 0.1%

Mean scaffold size 4099

N50 scaffold length 4081

Run CD-HIT

Step 6.

Remove contigs that have complete overlaps

cmd COMMAND - 4.6

```
cd-hit-est -i tara_northpacific_PRIMARY_min2000_contigs.fasta -
o tara_northpacific_PRIMARY_min2000_contigs.99.fasta -T 90 -M 500000 -c 0.99 -n 10
```

CD-HIT is available here: http://weizhong-lab.ucsd.edu/cd-hit/

Re-number the contigs in the current file

Step 7.

At this point the multiple Megahit assemblie have been sharing the same ID naming scheme.

Re-number the contig IDs. We convert ours to read - "MHASMcontig_###"

New file name = tara_northpacific_PRIMARY_min2000_contigs.renamed.99.fasta

Convert file to AFG format and run minimus2 assembler

Step 8.

EXPECTED RESULTS

END - Elapsed time: 3d 10h 6m 23s

Output

tara_northpacific_PRIMARY_min2000_contigs.renamed.99.fasta = 104,545 contigs

tara_northpacific_PRIMARY_min2000_contigs.renamed.99.singletons.seq = 839,264 contigs

Step 9.

cmd COMMAND

```
cat tara_northpacific_PRIMARY_min2000_contigs.renamed.99.fasta tara_northpacific_PRIMARY_mi
n2000_contigs.renamed.99.singletons.seq > tara_northpacific_SECONDARY_contigs.fasta
```

EXPECTED RESULTS

943,809 contigs

Number of scaffolds > 10K nt 54565 5.8%

Number of scaffolds > 100K nt 927 0.1%

Mean scaffold size 4435

N50 scaffold length 4693

Align raw sequences to SECONDARY contigs using Bowtie2

Step 10.

Build an index file

cmd COMMAND - 2.2.5

```
bowtie2-
build tara_northpacific_SECONDARY_contigs.fasta tara_northpacific_SECONDARY_contigs.bt_inde
x
```

Align raw sequences to SECONDARY contigs using Bowtie2

Step 11.

Perform alignment for each sample (site, fraction size, depth) - as used in the assembly AND repeat for each of the 15 datasets

Future iterations will then convert SAM files to BAM files

cmd COMMAND

```
bowtie2 -
q -1 ERR1712000_1.fastq.gz,ERR1712169_1.fastq.gz,ERR868467_1.fastq.gz -2 ERR1712000_2.fastq
.gz,ERR1712169_2.fastq.gz,ERR868467_2.fastq.gz -
S tara_northpacific_SECONDARY.tara138_prot_meso.sam -
x tara_northpacific_SECONDARY_contigs.bt_index --no-unal -p 50
```

Bowtie2 is available here: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

EXPECTED RESULTS

40.80% overall alignment rate

Use featureCounts to determine number of reads aligned to each contig

Step 12.

Convert FASTA format to SAF format, and run featureCounts

cmd COMMAND - 1.5.0

```
featureCounts -F SAF -a tara_northpacific_SECONDARY_contigs.saf -
o tara_northpacific_SECONDARY.tara138_prot_meso.readcount tara_northpacific_SECONDARY.tara1
38_prot_meso.sam
```

featureCounts is available here: http://bioinf.wehi.edu.au/featureCounts/

Prep data for BinSanity binning tool

Step 13.

cmd COMMAND

```
seqmagick convert --min-
length 7000 tara_northpacific_SECONDARY_contigs.fasta tara_northpacific_SECONDARY_contigs.m
in7000.fasta
```

Convert featurCounts output to coverage values for binning

Step 14.

Using the script in the BinSanity package: Binsanity-profile