***Idotea baltica* RNA-seq *de novo* Assembly**

Trinity 2.1.0

**PREPROCESSING**

REPTILE renamed the reads when separating the FASTQ file into a FASTA and Quality file. Read headers must be changed so that each pair has the same name and ends with /1 or /2, using PYTHON script ‘rename\_fasta.py’

scp rename\_fasta.py ketaya@taito.csc.fi:’$WRKDIR’

mv rename\_fasta.py trinity

python rename\_fasta.py out\_fw\_corrected.fa /1 left.fa

python rename\_fasta.py out\_rev\_corrected.fa /2 right.fa

\*\*\* ~10mins each

Batch jobs file: (based on example)

#SBATCH --cpus-per-task=24

#SBATCH -t 36:00:00

#SBATCH --mem=300000

#SBATCH -p hugemem

module load biokit

Trinity --seqType fa \

--max\_memory 300G \

--left left.fa \

--right right.fa \

--CPU $SLURM\_CPUS\_PER\_TASK \

--normalize\_reads \

--grid\_conf $TRINITY\_HOME/hpc\_conf\_taito.slurm

\*\*\* NOTE! memory should be at least 10x cpus

**RUNNING**

Run

sbatch trinity\_hugemem.sh

\*\*\* ~7hrs (24 cores, 300GB RAM)

Output commands and parameters in file ‘out\_7276491.txt’

***IN SILICO* READ NORMALIZATION**

Removes excess reads beyond 50x coverage

CMD: insilico\_read\_normalization.pl \

--seqType fa \

--JM 300G \

--max\_cov 50 \

--CPU 24 \

--output trinity\_out\_dir/insilico\_read\_normalization \

--left left.fa \

--right right.fa \

--pairs\_together \

--PARALLEL\_STATS

\*\*\* ~ 5.5hrs

9172366 / 235818494 = 3.89% reads selected during normalization.

533883 = 0.23% reads discarded as likely aberrant based on coverage profiles.

0.00% reads missing kmer coverage (N chars included?).

Normalized reads located in files:

left.fa.normalized\_K25\_C50\_pctSD200.fa

right.fa.normalized\_K25\_C50\_pctSD200.fa

*See if changing from 25mers to 20mers makes a difference*

perl /homeappl/appl\_taito/bio/trinity/trinity-2.1.0/utilinsilico\_read\_normalization.pl \

--seqType fa \

--JM 250G \

--max\_cov 50 \

--CPU 24 \

--output trinity\_norm\_20 \

--left left.fa \

--right right.fa \

--pairs\_together \

--PARALLEL\_STATS \

--KMER\_SIZE 20

\*\*\* ~ hrs

**JELLYFISH**

Building a k-mer catalog from reads

* Originally INCHWORM did this on it’s own until a faster method was created: JELLYFISH.

**INCHWORM**

Linear contig construction from k-mers

* Can be up to 32-mers (default is 25-mers)
  + Increase to improve specificity but lose sensitivity
* Inchworm-assembled contigs are output in the file ‘inchworm.K25.L25.DS.fa’
  + K-mer of 25, length at least 25, Double-Stranded as in not strand specific
* This is the most memory (RAM) intensive step
  + Memory needed can be reduced by increasing the minimum k-mer threshold abundance, but will lose some transcript reconstruction sensitivity

Trinity --min\_kmer\_cov 2 (default is 1)

**CHRYSALIS**

Clustering of contigs

**BUTTERFLY**

**OUPUT**

Output Statistics found in file ‘Trinity.timimg’

Statistics:

===========

Trinity Version: v2.1.0

Compiler: GCC

Trinity Parameters: --seqType fa --max\_memory 300G --left left.fa --right right.fa --CPU 24 --normalize\_reads --grid\_conf /appl/bio/trinity/trinityrnaseq-2.1.0/hpc\_conf/taito.slurm

Paired mode

Input data

Left.fasta 1206 MByte

Right.fasta 1195 MByte

Number of unique KMERs: 102470716

Number of reads: 0 Output data

Trinity.fasta 104 MByte

Runtime

=======

Start: Mon Nov 2 09:38:57 EET 2015

End: Mon Nov 2 16:45:30 EET 2015

Trinity 25593 seconds (7.1hrs)

Inchworm 1297 seconds (22mins)

Chrysalis 4857 seconds (81mins)

Butterfly 0 seconds

Rest 19439 seconds (5.4hrs)

Checked Google forums for Trinity and it seems that the timing is incorrect at least for Butterfly and it is a known issue, but as long as the output FASTA file seems alright then there is no need to worry.

Calculate Nx Stats with TrinityStats.pl

/appl/bio/trinity/trinityrnaseq-2.1.0/util/TrinityStats.pl Trinity.fasta

################################

## Counts of transcripts, etc.

################################

Total trinity 'genes': 90,675

Total trinity transcripts: 115,851

Percent GC: 39.29

########################################

Stats based on ALL transcript contigs:

########################################

Contig N10: 3435

Contig N20: 2532

Contig N30: 1963

Contig N40: 1505

Contig N50: 1136

Median contig length: 357

Average contig: 676.70

Total assembled bases: 78,396,052

#####################################################

## Stats based on ONLY LONGEST ISOFORM per 'GENE':

#####################################################

Contig N10: 3075

Contig N20: 2168

Contig N30: 1577

Contig N40: 1145

Contig N50: 810

Median contig length: 313

Average contig: 560.20

Total assembled bases: 50,796,185

**TRY AGAIN WITH** --KMER\_SIZE 20

**Assemble with Rcorrector reads**

#SBATCH --cpus-per-task=24

#SBATCH -t 24:00:00

#SBATCH --mem=256000

module load biokit

Trinity --seqType fq \

--max\_memory 255G \

--left /wrk/ketaya/rcorrector\_data/25mers/out\_fw\_paired.cor.fq \

--right /wrk/ketaya/rcorrector\_data/25mers/out\_rev\_paired.cor.fq \

--CPU $SLURM\_CPUS\_PER\_TASK \

--normalize\_reads \

--grid\_conf $TRINITY\_HOME/hpc\_conf\_taito.slurm

\*\*\* ~4hrs

Job 7592192 failed to allocated ~143GB or memory but it finished normalization. Strangely it only used 130GB out of 250GB allocated.

Normalization:

10,099,252 / 261,953,276 = 3.86% reads selected

529,077 (0.20%) reads discarded as likely aberrant based on coverage

0 reads missing k-mer coverage

Use normalized reads to assemble (trinity\_rcor.sh)

#SBATCH --cpus-per-task=8

#SBATCH -t 24:00:00

#SBATCH --mem=88000

module load biokit

Trinity --seqType fq \

--max\_memory 80G \

--left /wrk/ketaya/trinity/trinity\_rcor\_norm/insilico\_read\_normalization/out\_fw\_paired.cor.fq.normalized\_K25\_C50\_pctSD200.fq \

--right /wrk/ketaya/trinity/trinity\_rcor\_norm/insilico\_read\_normalization/out\_rev\_paired.cor.fq.normalized\_K25\_C50\_pctSD200.fq \

--CPU $SLURM\_CPUS\_PER\_TASK \

--grid\_conf $TRINITY\_HOME/hpc\_conf\_taito.slurm

\*\*\* ~2.5hrs (8 CPUs, 80GB)

Calculate Nx Stats

$TRINITY\_HOME/util/TrinityStats.pl Trinity.fasta

################################

## Counts of transcripts, etc.

################################

Total trinity 'genes': 90663

Total trinity transcripts: 115931

Percent GC: 39.36

########################################

Stats based on ALL transcript contigs:

########################################

Contig N10: 3513

Contig N20: 2580

Contig N30: 1992

Contig N40: 1529

Contig N50: 1154

Median contig length: 358

Average contig: 683.11

Total assembled bases: 79193347

#####################################################

## Stats based on ONLY LONGEST ISOFORM per 'GENE':

#####################################################

Contig N10: 3099

Contig N20: 2190

Contig N30: 1597

Contig N40: 1153

Contig N50: 820

Median contig length: 314

Average contig: 562.60

Total assembled bases: 51006702