***Idotea baltica* Assembly Annotation**

BLASTX (v2.2.31+)

NR (Dec 1, 2015)

**RUNNING**

Load biokit

module load biokit

Create batch job

#!/bin/bash -l

# created: Dec 7, 2015 11:21 AM

# author: ketaya

#SBATCH -J blastx

#SBATCH -n 1

#SBATCH -o out\_%j.txt

#SBATCH -e err\_%j.txt

#SBATCH --cpus-per-task=16

#SBATCH -t 48:00:00

#SBATCH --mem=16000

#SBATCH -p parallel

#SBATCH --mail-type=END

#SBATCH --mail-user=keith.t.yamada@utu.fi

module load biokit

pb blastx -query trinity.K25.raw.fasta -db nr -out idotea\_blastx2nr -outfmt 5 -evalue 0.001

Run

pb blastx -query trinity.K25.raw.fasta -db nr -out result.txt -evalue 0.001

split

Total amount of query sequences: 115931

The job is split into 443 pieces

requesting memory 16000 MB

Batch job ID: 7835793

starting blast\_clusterrun with job ID: 65006

7833953

14:42:18 Jobs: done 0, running 5, waiting 438, failed 0

..........

15:09:17 Jobs: done 0, running 443, waiting 0, failed 0

......

17:11:28 Jobs: done 1, running 442, waiting 0, failed 0

\*\*\* ~20 hrs

**RESULTS**

idotea\_blastx2nr

\*\* ~30GB

Send scripts to Taito

scp parse\_blast.py [ketaya@taito.csc.fi](mailto:ketaya@taito.csc.fi):

scp ReParseBlastbycutoffs.py ketaya@taito.csc.fi:

Load modules

module load openblas

module load python

\*\* Ran for an hour, then killed... too big, need batch job

Parse BlastX output using batch job (parse\_blast.sh)

#SBATCH --cpus-per-task=16

#SBATCH -t 24:00:00

#SBATCH --mem=16000

module load biokit

module load openblas

module load python

python parse\_blast.py idotea\_PARSEDblastx2nr.txt idotea\_blastx2nr

\*\* ~2.5 hrs (file size ~20GB)

Reparse given evalue cutoff (10-5)

#SBATCH --cpus-per-task=16

#SBATCH -t 24:00:00

#SBATCH --mem=16000

module load biokit

module load openblas

module load python

python ReParseBlastbycutoffs.py idotea\_tophitreParsedBlastx2nr\_e5.txt 12 0.00001 idotea\_PARSEDblastx2nr.txt

\*\* ~2 mins

good hits: 38,872

unique matches: 4,270

evalue cutoff 10-10... hopefully number of annotated contigs doesn’t change too much

#SBATCH --cpus-per-task=16

#SBATCH -t 24:00:00

#SBATCH --mem=16000

module load biokit

module load openblas

module load python

python ReParseBlastbycutoffs.py idotea\_PARSEDblastx2nr.txt 12 0.0000000001 idotea\_tophitreParsedBlastx2nr\_e10.txt

\*\* ~2 mins

good hits: 32,930

unique matches: 4,246

ReParse Blast hits to simplify (only query name, accession, subject name, taxonomy, and hsp expect) (ReParseBlast.py)

python ReParseBlast.py idotea\_tophitreParsedBlastx2nr\_e5.txt idotea\_blastx2nr\_e5\_simple.txt

\*\* confirmed same number of annotated contigs (38,872) (54,465 hits)

Use script to ignore “predicted, hypothetical, putative, unknown” (ReParseBlastByTerm.py)

python ReParseBlastByTerm.py idotea\_blastx2nr\_e5\_simple.txt idotea\_blastx2nr\_e5\_simple\_good.txt idotea\_blastx2nr\_e5\_simple\_bad.txt predicted,hypothetical,putative,unknown

\*\* good: 12,941 (21,171 hits) bad: 25,931 (33,294 hits)

Use script to id non-arthropod annotations (GroupTaxon.py)

python GroupTaxon.py idotea\_blastx2nr\_e5\_simple\_good.txt idotea\_blastx2nr\_e5\_simple\_good\_taxon.txt

Use script to filter by taxon “E.coli, Tetrahymena thermophila

...” (FilterByTaxon.py)

python FilterByTaxon.py idotea\_blastx2nr\_e5\_simple\_good.txt idotea\_blastx2nr\_e5\_simple\_good\_NONbacproto.txt idotea\_blastx2nr\_e5\_simple\_good\_bacproto.txt idotea\_filterlist.txt

\*\* good: 10,930 (13,879) bad: 2,011 (7,292)