Starling-May18

Projects/Katarina Stuart/KStuart.Starling-Aug18/Sv3_Genome/Assembly/2020-02-18.L_RNA_scaffolder

PDF Version generated by

Katarina Stuart (z5188231@ad.unsw.edu.au)

on

Jun 23, 2022 @03:53 PM NZST

Table of Contents

2020-02-18.L_RNA_scaffolder

Katarina Stuart (z5188231@ad.unsw.edu.au) - Mar 05, 2020, 6:54 PM NZDT

L_RNA_scaffolder: Scaffolding with Isoseq

DESCRIPTION

L_RNA_scaffolder is a genome scaffolding tool with long trancriptome reads. The long transcriptome reads could be generated by 454/Sanger/Ion_Torrent sequencing, or de novo assembled with pair-end Illumina sequencing. The long reads are aligned to genome fragments using BLAT and alignment file (PSL format with no heading) is used as the input file of L_RNA_scaffolder. L_RNA_scaffolder searches "guider" reads, fragment of which were mapped to different genome fragments. Then the "guider" reads orientated and ordered the genome fragments into longer scaffolds.

https://github.com/CAFS-bioinformatics/L_RNA_scaffolder

Using:
Scaffolds=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/SSPACE_GapfinisherV2/scaffolds_gapfilled_FINAL.fasta
RNA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.1_StarlingIsoseq/analysis/Isoseq3.3_pipeline/polya_8/clustered.hq.fasta

Producing PSL file with BLAT:

https://genome.ucsc.edu/goldenpath/help/blatSpec.html

```
#!/bin/bash

#PBS -N 2020-03-03.IsoseqBLAT.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=24gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/L_RNA_scaffolder

module load blat/35

GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/SSPACE_GapfinisherV2/scaffolds_gapfilled_FINAL.fasta
RNA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.1_StarlingIsoseq/analysis/Isoseq3.3_pipeline/polya_8/clustered.hq.fasta

blat $GENOME $RNA -fine -q=rna -noHead output.psl
```

```
blat $GENOME $RNA -ooc=11.ooc -fine -noHead -q=rna output.psl
```

"-ooc=11.ooc" I don't know what this is or how to make it?

"-ooc=11.ooc tells the program to load over-occurring 11-mers from an external file. This will increase the speed by a factor of 40 in many cases, but is not required."

I was able to run the below:

```
blat $GENOME $RNA -fine -q=rna output.psl
blat $GENOME $RNA -fine -q=rna -noHead output.psl
```

L_RNA_scaffolder

INPUT FILES

PSL file and genome fragment fasta file are necessary for scaffolding. The psl file was generated using BLAT program with "-noHead" option. The genome fragment file should be fasta format, consistent with the subject sequences when using BLAT program. Another file, named overlapping file, contain two columns. This file is not necessary but will avoid some interesting genome fragments not being scaffolding.

COMMANDS AND OPTIONS

L_RNA_scaffolder is run via the shell script: L_RNA_scaffolder.sh found in the base installation directory.

Usage info is as follows:

Required:

-d : the directory where the programs are in.
-i : the output of transcripts alignment with BLAT.
-j : the genome fragments fasta file which will be scaffolded and was used as the database when aligning transcript reads.

Optional:

-r : some fragments which you might be interesting and will not be scaffolded. The file has two columns per row. One row stand for that two fragments might be connected and should not be scaffolded.
-l : the threshold of alignment length coverage (default:0.95). If one read has a hit of which length coverage was over the threshold, then this read would be filtered out.
-p : the threshold of alignment identity (default: 0.9). If one alignment has an identity over the threshold, then the alignment is kept for the further analysis.
-o : the directory where the output file is stored. The default output directory is equal to the program directory. -e : The maximal intron length between two exons (default: 100kb).
-f : the minimal number of supporting reads (default: 1). If the number of the supporting reads for the connection is over the frequency, then this connection is reliable.

Note: a typical L_RNA_scaffolder command might be:

sh L_RNA_scaffolder.sh -d ./ -i input.psl -j genome.fasta

OUTPUT FILES

When L_RNA_scaffolder completes, it will create an L_RNA_scaffolder.fasta output file in the output_dir/ output directory.

echo "Usage: sh `basename $0` -d Program_DIR -i inputfile.psl -j contig.fasta";

Need to give the program write permissions first: https://askubuntu.com/questions/409025/permission-denied-when-running-sh-scripts

```
cd /home/z5188231/programs/L_RNA_scaffolder-src

chmod u+r+x *
```

Run the scaffolding:

```
#!/bin/bash

#PBS -N 2020-03-03.L_RNA_Scaff.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=24gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae
```

```
module load perl/5.28.0

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/L_RNA_scaffolder/

GENOME=/srv/scratch/z5188231/KStuart.Starling-
Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/SSPACE_GapfinisherV2/scaffolds_gapfilled_FINAL.fasta

cp /home/z5188231/programs/L_RNA_scaffolder-src ./

sh L_RNA_scaffolder-src/L_RNA_scaffolder.sh -d L_RNA_scaffolder-src/ -i output.psl -j $GENOME
```

```
mkdir slimsuite && cd $_
module load python/2.7.15
python ~/SLiMSuite/tools/seqsuite.py summarise batchrun="../*.fasta" basefile=scaffolds dna newlog
```

```
#~~#    00:00:04       # ~~~~~~~~ Sequence Summary for L_RNA_scaffolder ~~~~~~~~~ #
#SUM    00:00:30       Total number of sequences: 7,776
#SUM    00:00:30       Total length of sequences: 1,067,321,776
#SUM    00:00:30       Min. length of sequences: 977
#SUM    00:00:30       Max. length of sequences: 31,181,295
#SUM    00:00:30       Mean length of sequences: 137,258.46
#SUM    00:00:30       Median length of sequences: 2,395
#SUM    00:00:30       N50 length of sequences: 7,118,366
#SUM    00:00:30       L50 count of sequences: 38
#SUM    00:00:30       GC content: 41.82%
#SUM    00:00:30       Gap (N) length: 11,598,876 (1.09%)

#~~#    00:00:31       # ~~~~~~~~~~~~ Sequence Summary for scaffold ~~~~~~~~~~~~~ #
#SUM    00:00:34       Total number of sequences: 60 I think this means that 60 RNA reads were used for scaffolding?
#SUM    00:00:34       Total length of sequences: 123,100,963
#SUM    00:00:34       Min. length of sequences: 2,583
#SUM    00:00:34       Max. length of sequences: 31,181,295
#SUM    00:00:34       Mean length of sequences: 2,051,682.72
#SUM    00:00:34       Median length of sequences: 272,362
#SUM    00:00:34       N50 length of sequences: 12,606,761
#SUM    00:00:34       L50 count of sequences: 4
#SUM    00:00:34       GC content: 41.86%
#SUM    00:00:34       Gap (N) length: 1,219,184 (0.99%)

#~~#    00:03:29       # ~~~~ Sequence Summary for scaffolds_gapfilled_FINAL ~~~~ #
#SUM    00:04:17       Total number of sequences: 7,856
#SUM    00:04:17       Total length of sequences: 1,067,313,776
#SUM    00:04:17       Min. length of sequences: 977
#SUM    00:04:17       Max. length of sequences: 30,547,435
#SUM    00:04:17       Mean length of sequences: 135,859.70
#SUM    00:04:17       Median length of sequences: 2,433
#SUM    00:04:17       N50 length of sequences: 6,652,296
#SUM    00:04:17       L50 count of sequences: 40
#SUM    00:04:17       GC content: 41.82%
#SUM    00:04:17       Gap (N) length: 11,590,876 (1.09%)
```

```
#!/bin/bash

#PBS -N 2020-03-04.BUSCO_L_RNA.pbs
#PBS -V
#PBS -l nodes=1:ppn=16
#PBS -l mem=16gb
```

```
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/L_RNA_scaffolder/busco

module load python/3.7.3 blast+/2.2.31 hmmer/3.2.1 augustus/3.3.2 emboss/6.6.0 busco/3.0.2b

export AUGUSTUS_CONFIG_PATH=/srv/scratch/z5188231/programs/augustus
export BUSCO_CONFIG_FILE=/home/z5188231/busco/3.0.2b/config/config.ini

BUSCOSET=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data/BUSCO.2018-08-21

python3 /apps/busco/3.0.2b/scripts/run_BUSCO.py -i ../L_RNA_scaffolder.fasta -o L_RNA_scaffolder -m genome -
l ${BUSCOSET}/aves_odb9/ -c 32 -f
```

Previous results for post gapfilled version:

INFO Results:
INFO C:94.4%[S:92.3%,D:2.1%],F:3.5%,M:2.1%,n:4915
INFO 4638 Complete BUSCOs (C)
INFO 4537 Complete and single-copy BUSCOs (S)
INFO 101 Complete and duplicated BUSCOs (D)
INFO 173 Fragmented BUSCOs (F)
INFO 104 Missing BUSCOs (M)
INFO 4915 Total BUSCO groups searched
INFO BUSCO analysis done. Total running time: 7594.473603963852 seconds

After RNA scaffolding version:

INFO Results:
INFO C:94.3%[S:92.3%,D:2.0%],F:3.5%,M:2.2%,n:4915 for some reason this does down despite the fact that the numbers dont add up this way. Odd, but will proceed anyway.
INFO 4638 Complete BUSCOs (C)
INFO 4539 Complete and single-copy BUSCOs (S)
INFO 99 Complete and duplicated BUSCOs (D)
INFO 170 Fragmented BUSCOs (F)
INFO 107 Missing BUSCOs (M)
INFO 4915 Total BUSCO groups searched