Processing of Pacbio Iso-seq sequences

Bing Cheng; Agnelo Furtado; Robert Henry

Abstract

This protocol is to further process the sequences generated from ICE and Quiver.

Citation: Bing Cheng; Agnelo Furtado; Robert Henry Processing of Pacbio Iso-seq sequences. **protocols.io**
dx.doi.org/10.17504/protocols.io.ja3cign
Published: 10 Aug 2017

Guidelines

Remove of the detected contaminant sequences

If X > Y, then the sequences were removed. This is done subsequently after each detecting steps, therefore, the dataset number is decreased and its name has changed from step 4 onwards.

Before start

Raw data from Pacbio Iso-Seq needs to be processed with RS IsoSeq (version 2.3) pipeline.

Protocol

Remove Primer IIA sequence motifs

Step 1.

To remove the Primer IIA sequence motifs used in library preparations.

Combine the HQ and LQ sequences

Step 2.

LQ output or non-full length coverage sequences may from rare transcripts or lower coverage sequences.

NOTES

GigaScience Database 08 Aug 2017

HQ: high quality sequences, LQ, low quality sequences generated from RS IsoSeq pipeline

Combined sequences were processed with CD-HIT-EST

Step 3.

To further remove the redundant sequences.

NOTES

GigaScience Database 08 Aug 2017

The output dataset was hereafter called **dataset A**

Detecting of chloroplast sequences

Step 4.

BLASTn (1e-10) the **dataset A** against the complete C.arabica chloroplast genome.

NOTES

GigaScience Database 08 Aug 2017

Accession number: EF044213.1 (processed with CLC genomic workbench)

Detecting of mitochondrial sequences

Step 5.

BLASTn (1e-10) the **dataset B** against the N.tabacum and V. vinifera complete mitochondrial genomes (relate species)

NOTES

GigaScience Database 08 Aug 2017

Accession number: BA000042.1 and FM179380.1 (processed with CLC genomic workbench)

Detecting of ribosomal sequences

Step 6.

BLASTn (1e-10) the **dataset C** against the public available ribosomal genes from C. arabica, C.canephora and C.eugenioides

NOTES

GigaScience Database 08 Aug 2017

Accession number: AJ224846, EU650386, DQ153609, AF416459, EU650384, EU650385, AF542981, AF542990, JX459583, JX459584, JX459585, JX459586, JX459587, DQ153593, AF542982, DQ423064, DQ153588, DQ153621, AF542986 (processed with CLC genomic workbench)

Detecting of virus and viroid sequences

Step 7.

BLASTn (1e-10) the **dataset D** against the reference genomes of virus and viroid

NOTES

GigaScience Database 08 Aug 2017

Download from NCBI (processed with CLC genomic workbench)

Detecting of prokaryotic sequences

Step 8.

BLASTn (1e-10) the **dataset E** against the reference genome of prokaryotes

NOTES

GigaScience Database 08 Aug 2017

Download from NCBI (processed with CLC genomic workbench)

Detecting of fungal sequences

Step 9.

BLASTx (1e-10) the **dataset F** against the fungal proteins

NOTES

GigaScience Database 08 Aug 2017

Download from NCBI (processed with CLC genomic workbench)

Find significant hits

Step 10.

Significant matches are filtered with bit score (X) ≥ 300 and identity ≥ 80%

NOTES

GigaScience Database 08 Aug 2017

Processed with CLC genomic workbench

Validation with Cloud BLAST

Step 11.

All the significant matches were confirmed with cloud BLASTn (bit score (Y))

NOTES

GigaScience Database 08 Aug 2017

Processed with CLC genomic workbench

Remove of the detected contaminant sequences

Step 12.

If X > Y, then the sequences should be removed. This is done subsequently after each detecting steps, therefore, the dataset number is decreased and its name has changed from step 4 onwards.

NOTES

GigaScience Database 08 Aug 2017

Processed with CLC genomic workbench

Quality check

Step 13.

Sequence quality was then accessed with the Fasta Statistics through Galaxy/GVL 4.0

NOTES

GigaScience Database 08 Aug 2017

Hereafter the dataset was called **dataset G**