Long-read sequencing of the coffee bean transcriptome reveals the diversity of full length transcripts

Bing Cheng; Agnelo Furtado; Robert Henry

Abstract

Polyploidization contributes to the complexity of gene expression resulting in numerous related but different transcripts. This study explored the transcriptome diversity and complexity of tetraploid Arabica coffee (Coffea arabica) bean. Long-read sequencing (LRS) by Pacbio Isoform sequencing (Iso-seq) was used to obtain full-length transcripts without the difficulty and uncertainty of assembly required for reads from short read technologies. The tetraploid transcriptome was annotated and compared with data from the sub-genome progenitors. Caffeine and sucrose genes were targeted for case analysis._x000D_
 An isoform-level tetraploid coffee bean reference transcriptome with 95,995 distinct transcripts (average 3,236 bp) was obtained. A total of 88,715 sequences (92.42%) were annotated with BLASTx against NCBI non-redundant plant proteins, including 34,719 high quality annotations. Further BLASTn to NCBI non-redundant nucleotide sequences, C. canephora coding sequences with UTR, C.arabica ESTs and Rfam resulted in 1,213 sequences without hits, were potential novel genes in coffee. Longer UTRs were captured, especially in the 5'UTRs, facilitating the identification of upstream ORFs (uORFs). The LRS also revealed more and longer transcript variants in key caffeine and sucrose metabolism genes from this polyploid genome. Long sequences (>10kb) were poorly annotated._x000D_

Citation: Bing Cheng; Agnelo Furtado; Robert Henry Long-read sequencing of the coffee bean transcriptome reveals the diversity of full length transcripts. **protocols.io**

dx.doi.org/10.17504/protocols.io.ja2cige

Published: 10 Aug 2017

Collection

PROTOCOLS

1. Processing of Pacbio Iso-seq sequences

CONTACT: Robert Henry

2. Transcriptome annotation

CONTACT: Robert Henry