

CreateSequenceDictionary Documentation

Description: Reads FASTA or FASTA.GZ files containing reference

sequences, and writes them as a SAM file containing a sequence

dictionary.

Author: Picard team

Contact: Marc-Danie Nazaire, gp-help@broadinstitute.org

Summary

This module uses a FASTA file or a GZIP of FASTA files to create a sequence dictionary in SAM format. However, the output file extension must be DICT to reflect the fact that it is a sequence dictionary.

A sequence dictionary contains the sequence name, sequence length, genome assembly identifier, and other information about sequences.

FASTA files contain sequence data. For more information on the FASTA format, see the NIH description at http://www.ncbi.nlm.nih.gov/BLAST/fasta.shtml. SAM is a standard short read alignment that allows aligners to attach custom tags to individual alignments. For more information about sequence dictionaries, see the SAM specification at http://samtools.sourceforge.net/.

This module uses the CreateSequenceDictionary function from Picard. For more information, see the Picard Web site at http://picard.sourceforge.net/.

Parameters

Name	Description
reference. sequence.file (required)	Input reference FASTA or FASTA.GZ file. For more information on the FASTA format, see the NIH description at http://www.ncbi.nlm.nih.gov/BLAST/fasta.shtml .
genome. assembly (optional)	The genome assembly to put into the AS field of the sequence dictionary entry.
uri (optional)	The URI to put into the UR field of the sequence dictionary entry. The UR is the path to the sequence. If a URI is not specified it defaults to the absolute path of the reference FASTA input file as a URI.

GenePattern

truncate. names.at.white. space (optional)	Make the sequence name the first word from the > line in the reference file. Default: no
num.sequences (optional)	Stop after writing this many sequences.
output.file (required)	The name of the output file. Default: <pre><reference.sequence.file_basename>.dict</reference.sequence.file_basename></pre>

Output Files

1. DICT file

This file is in SAM format, containing header and information for the sequence dictionary. For more information on the sequence dictionary format, see the SAM specification at http://samtools.sourceforge.net/.

Platform Dependencies

Module type: Preprocess & Utilities

CPU type: any **OS:** any

Language: Java (minimum version 1.6)