Introduction to Transcriptomics Analysis

Class 05 - Manipulations of Sequence Files in Linux I.

INSTRUCTOR:
Aureliano Bombarely
Department of Bioscience
Universita degli Studi di Milano
aureliano.bombarely@unimi.it

Outline of Topics

- Exercise 1: Counting sequences in fasta files.
- Exercise 2: Counting nucleotides in fasta files.
- Exercise 3: Counting annotations in fasta files.
- Exercise 4: Counting elements in GFF files.
- Exercise 5: Getting stats for fasta files.

Outline of Topics

- Exercise 1: Counting sequences in fasta files.
- Exercise 2: Counting nucleotides in fasta files.
- Exercise 3: Counting annotations in fasta files.
- Exercise 4: Counting elements in GFF files.
- Exercise 5: Getting stats for fasta files.

- Exercise 1: Counting sequences in fasta files.

  Goal: Using simple Linux commands count the number of FASTA sequences into a file.

  Input: File “Artha_TAIR10_genome.fasta”

  Recommended command: **grep -c “>” filename**

Outline of Topics

- Exercise 1: Counting sequences in fasta files.
- Exercise 2: Counting nucleotides in fasta files.
- Exercise 3: Counting annotations in fasta files.
- Exercise 4: Counting elements in GFF files.
- Exercise 5: Getting stats for fasta files.

- Exercise 2: Counting nucleotides in fasta files.

  Goal: Using simple Linux commands count the number of nucleotides into a file.

  Input: File “Artha_TAIR10_genome.fasta”

  Recommended command: **grep -v “>” filename l wc -m**

Outline of Topics

- Exercise 1: Counting sequences in fasta files.
- Exercise 2: Counting nucleotides in fasta files.
- Exercise 3: Counting annotations in fasta files.
- Exercise 4: Counting elements in GFF files.
- Exercise 5: Getting stats for fasta files.

- Exercise 3: Counting annotations in fasta files.

Goal: Using simple Linux commands count the number of kinase protein sequences for a FASTA file.

Input: File “Araport11_genes.201606.pep.fasta.gz”. You can download it using wget as wget https://www.arabidopsis.org/download_files/Sequences/Araport11_blastsets/Araport11_genes.201606.pep.fasta.gz

Recommended command: **grep “>” filename l grep kinase l wc -l**

Outline of Topics

- Exercise 1: Counting sequences in fasta files.
- Exercise 2: Counting nucleotides in fasta files.
- Exercise 3: Counting annotations in fasta files.
- Exercise 4: Counting elements in GFF files.
- Exercise 5: Getting stats for fasta files.

- Exercise 4: Counting elements in GFF files.

  Goal: Using simple Linux commands count the number of sequence ontology items in a GFF file.

  Input: File “Araport11_GFF3_genes_transposons.201606.gff”.

  Recommended command: **grep -v “#” filename l grep kinase l cut -f3 l sort l uniq -c**

Outline of Topics

- Exercise 1: Counting sequences in fasta files.
- Exercise 2: Counting nucleotides in fasta files.
- Exercise 3: Counting annotations in fasta files.
- Exercise 4: Counting elements in GFF files.
- Exercise 5: Getting stats for fasta files.

- Exercise 5: Getting stats for fasta files.

  Goal: Using fastq-stats command retrieve the stats from a FASTQ file

  Input: All the Artha_*.fastq.gz files

  Recommended command: **fastq-stats filename > filename.stats.txt**