Assignment 1 -EANBiT 2018 Training

Dina Machuve

13 June 2018

Questions

Using the sequence file, *nrf1_seq.fa*, answer the following questions:

1. How many organisms are represented in the sequence file provided? Which command did you use?
2. How many of these sequences are mRNA?
3. How many nucleotides are in the 18th sequence? Which commands did you use.
4. Create a file *(seq_ids.txt)* containing all the sequence IDs, are there duplicates?

Solution 1

There are **36** organisms represented in the *nrf1_seq.fa* sequence file. The commands used are:

- `grep '>' nrf1\_seq.fa >headersnrf.fa`

  to get all the sequence headers and store on temporary file called 'headersnrf.fa'
- `cut -d' ' -f2,3 headersnrf.fa |sort |grep -Ev 'PREDICTED:' >final.fa`

  to get all genus and species (field 2 and 3 respectively) and removed all organisms with foreword 'PREDICTED:' which is in field 2 and saved in temporary file called 'final.fa'
- `grep 'PREDICTED:' headersnrf.fa |cut -d' ' -f3,4 |sort >>final.fa`

  to find all genus and species (field 3 and 4) for organisms with foreword PREDICTED:, sorted them and appended to the file 'final.fa'
- `sort final.fa |uniq |wc -l`

  to get unique count of organisms in the 'final.fa' file

Solution 2

There are **90** mRNA sequences. This was attained using the command

```
grep 'mRNA' nrf1_seq.fa |wc -l
```

Solution 3

There are **214,583** nucleotides are in the 18th sequence. This was attained using the command:

```
grep '>' nrf1_seq.fa|sed -n '18,19p'
```

```
grep -A10000 ">AC161538.7 Mus musculus 6 BAC RP23-1D15 (Roswell Park Cancer Institute
(C57BL/6J Female) Mouse BAC Library) complete sequence" nrf1_seq.fa |grep -B10000
">AC153632.2 Mus musculus 6 BAC RP23-450O1 (Roswell Park Cancer Institute (C57BL/6J
Female) Mouse BAC Library) complete sequence" >check.fa
```

```
tail -n +2 check.fa | head -n -1 |wc -c
```

Solution 4

There are 100 sequence ids on the file *seq_ids.txt*. There are no duplicates.

```
grep '>' nrf1_seq.fa |cut -f1 -d' '|cut -f2 -d'>'>seq_ids.txt
```

```
sort seq_ids.txt |uniq |wc -l
```