Uniwersytet Przyrodniczy we Wrocławiu
Wydział Biologii i Hodowli Zwierząt

Pracownia Informatyczna

AWK

LISTA 5

Szymon Armata
117360

Bioinformatyka, rok III, grupa II

1. SNP

Na theta.edu.pl odnajdź pliki Neomys fodiens_A.txt oraz Neomys fodiens_B.txt. Pliki te występują w formacie vcf (w tym przypadku bez nagłówka), który służy do przechowywania informacji na temat polimorfizmów. Napisz skrypt, który policzy:

a. ile SNP znajduje się w każdym z plików

```
echo -e "\nNeomys_fodiens_A.txt"
awk 'END{ print NR }' Neomys_fodiens_A.txt

echo -e "\nNeomys_fodiens_B.txt"
awk 'END{ print NR }' Neomys_fodiens_B.txt
```

b. ile razy nukleotyd A z genomu referencyjnego został wymieniony na inny nukleotyd

```
awk 'BEGIN{ count=0 }{ if($4~/A/) count++}
END{ print count }' Neomys_fodiens_A.txt
```

c. wypisze do nowego pliku SNP, które są dokładnie takie same u obu zwierząt

```
comm -12 <(cut -f1,2,5 Neomys_fodiens_A.txt) <
(cut -f1,2,5 Neomys_fodiens_B.txt)
```

d. różnice w liczbie SNP pomiędzy obydwoma plikami

```
a=$(awk 'END{ print NR }' Neomys_fodiens_A.txt)
b=$(awk 'END{ print NR }' Neomys_fodiens_B.txt)

echo $((a-b))
```

2. Duplikacje

Plik duplikacje.Btaurus.txt zawiera wybrane fragmenty genomu, które uległy duplikacji w genomie Bos taurus. Ustal jaką długość mają zduplikowane fragmenty i zrób to na podstawie:

a. nagłówków

```
awk -F [:-] '$1~/^>/{print␣$3-$2}'
duplikacje.Btaurus.txt
```

b. sekwencji DNA

```
awk '$1!~/^>/{print␣(length␣($0)+1)}'
duplikacje.Btaurus.txt
```

3. Polimorfizmy

Dla pliku R_norvegicus000020085003.vcf wyświetl na ekran wszystkie polimorfizmy zlokalizowane:

a. na chromosomie 11

```
awk '$1=="Chr11"␣{␣print␣$5␣}'
R_norvegicus000020085003.vcf_.txt
```

b. w pozycji „103508890”

```
awk '$2=="103508890"␣{␣print␣$5␣}'
R_norvegicus000020085003.vcf_.txt
```

c. w pozycji, która zaczyna się od liczb „103”

```
awk '$2~/^103/␣{␣print␣$5␣}'
R_norvegicus000020085003.vcf_.txt
```

d. w pozycji, która zaczyna się od liczb „103” na chromosomie 7

```
awk '$1=="Chr7"&&$2~/^103/␣{␣print␣$5␣}'
R_norvegicus000020085003.vcf_.txt
```

4. Skrypt - operacje

Napisz skrypt, który zastosuje wszystkie powyższe operacje dla plików:
- R_norvegicus000020085003.vcf
- R_norvegicus20029952202.vcf
- R_norvegicus20033325665.vcf

Skrypt powinien wypisywać na ekran nazwę pliku, dla którego wykonywane są poszczególne operacje, poprzedzoną ciągiem znaków „####".

```
!/bin/bash

for file in R_norvegicus*
do
        echo -e  "\n####" $file

        echo -e "\nPolimorfizmy na chr. 11"
        awk '$1=="Chr11" { print $5 }' $file

        echo -e "\nPolimorfizmy w poz. 103508890"
        awk '$2=="103508890" { print $5}' $file

        echo -e "\nPolimorfizmy w poz. od 103"
        awk '$2~/^103/ { print $5}' $file

        echo -e "\nPolimorfizmy w poz. od 103 na chr. 12"
        awk '$1=="Chr12"&&$2~/^103/ { print $5}' $file

        echo -e "\nPolimorfizmy w poz. od 103 na chr. 7"
        awk '$1=="Chr7"&&$2~/^103/ {print $5}' $file
done
```

5. Skrypt - proporcje

Napisz skrypt, który policzy proporcje wszystkich genotypów dla każdego z plików:
- R_norvegicus000020085003.vcf
- R_norvegicus20029952202.vcf
- R_norvegicus20033325665.vcf

Pamiętaj, że możliwe są 3 genotypy, homozygota referencyjna (0/0) oraz alternatywna (1/1), jak również heterozygota (0/1).

```
#!/bin/bash

FILES ="home/upwr/SzAr/z4/*"

for file in $FILES
do
  cat $file | awk '{print␣$10}' | awk -F: '{print␣$1}'
  > z4.mod.txt
  het=$(grep -c "0/1" z4.mod.txt)
  homref=$(grep -c "0/0" z4.mod.txt)
  homalt=$(grep -c "1/1" z4.mod.txt)
echo "###" $file "stosunek␣genotyp w(het:homref:homalt):
"$het : $homref : $homalt
done
```