Uniwersytet Przyrodniczy we Wrocławiu
Wydział Biologii i Hodowli Zwierząt

Pracownia Informatyczna

biologiczne bazy danych

LISTA 3

Szymon Armata
117360

Bioinformatyka, rok III, grupa II

1. Genomy referencyjne

W swoim katalogu domowym utwórz katalog, w którym będą przechowywane genomy referencyjne. Z bazy danych NCBI ściągnij pełny genom Apis mellifera. Znajdziesz go po linkiem:
https://ftp.ncbi.nlm.nih.gov/genomes/refseq/invertebrate/Apis_mellifera/latest_assembly_versions/GCF_003254395.2_Amel_HAv3.1/GCF_003254395.2_Amel_HAv3.1_genomic.fna.gz

```
cd SzAr
mkdir ref_genomes
cd ref_genomes
wget https://ftp.ncbi.nlm.nih.gov/genomes/refseq/
invertebrate/Apis_mellifera/latest_assembly_ver
sions/GCF_003254395.2_Amel_HAv3.1/
GCF_003254395.2_Amel_HAv3.1_genomic.fna.gz
```

1.1. Czy wszystkie sekwencje dotyczą chromosomów?

```
gunzip GCF_003254395.2_Amel_HAv3.1_genomic.fna.gz
less GCF_003254395.2_Amel_HAv3.1_genomic.fna.gz
q
grep '>' GCF_003254395.2_Amel_HAv3.1_genomic.fna
```

nie

1.2. Ile chromosomów zawiera ten plik?

```
grep '>NC_03' GCF_003254395.2_Amel_HAv3.1_genomic.fna
| wc -l
```

16 chromosomów

2. Identyfikatory

Zmień identyfikatory chromosomów w genomie referencyjnym pszczoły miodnej, tak aby zamiast identyfikatora sekwencji (po znaku większości) znajdował się numer chromosomu wraz z przedrostkiem „LG". Kod źródłowy zawrzyj w skrypcie change_names.sh.

```
```

3. Polecenie rsync

Jeśli potrzebujesz dużej ilości danych lub masz problemy z połączeniem sieciowym, wygodniejsze może być użycie rsync do pobierania.

Musisz odwiedzić witrynę Ensembl FTP w przeglądarce internetowej, aby zlokalizować potrzebne pliki, a następnie zmienić adres URL FTP w następujący sposób:

1. Zmień protokół z http: na rsync:
2. Wstaw ensembl w ścieżkę po nazwie domeny zanim wkleisz go do wiersza poleceń.

Na przykład następujące polecenie pobierze wszystkie ludzkie pliki EMBL z http://ftp.ensembl.org/pub/current_embl/homo_sapiens/ do bieżącego katalogu:

```
rsync -av rsync://ftp.ensembl.org/ensembl/pub/
current_embl/homo_sapiens
```

4. rsync w praktyce

Z bazy ENSEMBL (www.ensembl.org) pobierz (rsync) genom jądrowy S. scrofa ("unmasked genomic DNA sequences"). Dopilnuj, aby nie zostały pobrane tzw. contigi (fragmenty genomu nieprzypisane do żadnego chromosomu). Pobrane pliki z chromosomami połącz w odpowiedniej kolejności (autosomy w kolejności rosnącej oraz chromosomy płci).

```
sudo rsync -av rsync://ftp.ensembl.org/ensembl/pub/
release-105/fasta/sus_scrofa/dna/
Sus_scrofa.Sscrofa11.1.dna.chromosome.{[1-9XY],1?}.fa.gz
```

4.1. W jakim formacie przechowywane są te dane?

```
.fa.gz
```

4.2. Jakiego/jakich narzędzia/narzędzi możesz użyć w celu połączenia plików w zadanejkolejności?

```
for i in {1..18} X Y;
do zcat Sus_scrofa.Sscrofa11.1.dna.chromosome.$i.fa;
done > multi_chr_Sus_scrofa.fa
```

4.3. Za pomocą jakiego polecenia sprawdzisz czy kolejność chromosomów w nowym pliku jest prawidłowa?

```
grep '>' multi_chr_Sus_scrofa.fa
```

4.4. Jaki jest rozmiar finalnego pliku, w którym przechowujesz genom referencyjny?

```
wc -c multi_chr_Sus_scrofa.fa
```

2475851005

5. Liczenie plików

Pozostań w katalogu, w którym zostały umieszczone genomy referencyjne. Używając narzędzi linii poleceń, policz ile plików w formacie fasta znajduje się we wspomnianym katalogu.

```
find -name '*.fa' | wc -l
find -name '*.fa.gz' | wc -l
```