Dostęp do serwera

Na dzisiejszych zajęciach będziemy wykorzystywali oprogramowanie serwera linum. Do komunikacji z serwerem wykorzystamy protokół `ssh`. Każdy z Was ma utworzone konto. Login to nazwisko+pierwsza listera imienia (bez polskich znaków).

```
ssh <login>@biotech.uni.wroc.pl
```

Po połączeniu z serwerem należy uzupełnić zmienną systemową `$PATH`, aby móc korzystać z oprogramowania podanego w przykładach. W katalogu domowym znajduje się plik `copy_bashrc`. Należy podmienić plik `~/.bashrc` na plik `copy_bashrc`.

SRA - Sequence Read Archive

Sequence Read Archive (SRA) udostępnia dane dotyczące sekwencji społeczności naukowców, aby zwiększyć powtarzalność i umożliwić nowe odkrycia poprzez porównywanie zestawów danych. SRA przechowuje surowe dane (raw data) z sekwencjonowania z wysokoprzepustowych platform sekwencjonowania, w tym Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics® i Pacific Biosciences SMRT®.

SRA

Dane w SRA zorganizowane są w sposób hierarchiczny, gdzie każda kategoria może zawierać jedną lub więcej pod kategorii. I tak mamy:

- NCBIBioProject: PRJN**** (np. PRJNA257197) zawierajacą całościowy opis pojedyńczej inicjatywy badawczej; projekt odnosi się zazwyczaj do do wielu zbiorów danych.
- NCBI BioSample: SAMN**** i/lub SRS**** (np. SAMN03254300) opisuje materiał biologiczny; każda fizycznie unikatowy obiekt(genotyp) być zarejestrowany jako pojedyncza BioSample z unikalnym zestawem atrybutów.
- SRA Experiment: SRX**** unikalna biblioteka dla określonej próbki (BioSample).
- SRA Run: SRR**** lub ERR**** (np. SRR1553610) jest to link do pliku(-ów) danych połączonego z surowymi danymi sekwencjonowania

Do pobierania danych z archiwum SRA służy narzędzie SRA Toolkit, dostępne dla Windows, macOS oraz Linux. Podstawowe polecenie z tego narzędzia to fastq-dump

```
fastq-dump -h
```

Możliwe jest także pobranie z serwera pliku .sra i użycie programu fasta-dump bezpośrednio na pliku (lokalnie na komputerze)

Polecenie sra-stat generuje nam raport na temat danych w postaci pliku XML.

```
sra-stat --xml --quick SRR1553610 > raport.xml
```

Pobieranie danych z SRA

Do pobrania komunikacji z bazą danych Genbank oraz do pobierania informacji z tej bazy wykorzystamy narzędzie **Entrez Direct**. W pierwszej kolejności pobierzemy numery SRA Run dla Bioprojektu PRJNA515812

```
esearch -db sra -query PRJNA515812

esearch -db sra -query PRJNA515812 | efetch -format runinfo -mode xml | xtract -pattern SraRunInfo -element Run > runinfo.txt
```

Zadanie 1:
Proszę utworzyć plik SRR_ACC_List.txt zawierający numery SRA Run dla Bioprojektu PRJNA515812 jako listę z numerami w jednej kolumnie (na podstawie pliku runinfo.txt)

Stworzymy teraz prosty skrypt bash, który pozwoli nam na automatyzację procesu pozyskiwania danych z SRA. Wykorzystamy program fastq-dump z opcją --split-files, która pozwala na podział odczytów "pair-end" na dwa osobne pliki. Skrypt wczytuje linijka po linijce nazwy archiwów SRA z pliku tekstowego (listy) podanego w wierszu poleceń, a następnie zapisuje odczyty z bazy SRA jako pliki .fastq.

```
#!/bin/bash

while IFS='' read -r line || [[ -n "$line" ]]; do
    AR=${line};
    echo ${AR};
    echo "*** Pobieranie: ${AR}";
    fastq-dump -X 10 ${AR} --split-files
done < $1
```

Zadanie 2:
Zmodyfikować powyższy kod, tak aby pobierał pierwsze 1 000 odczytów z bazy SRA. Możemy teraz uruchomić nasz skrypt dla z plikiem SRR_Acc_List.txt

```
chmod +x SRA_data_col.sh
./SRA_data_col.sh SRR_Acc_List.txt
```

Kontrola jakości plików fastq QC

Czym jest kontrola jakości?

Kontrola jakości (w skrócie: QC) jest procesem poprawy danych poprzez usunięcie możliwych do zidentyfikowania błędów. Zazwyczaj jest to pierwszy krok wykonywany po pobraniu (otrzymaniu) danych. Ponieważ jest to proces, który zmienia dane, musimy być bardzo ostrożni, aby nie wprowadzić nieumyślnie nowych informacji do naszego zbioru danych. Optymalnie chcemy, aby te same dane (informacje) z danych surowych były w danych przefiltrowanych, ale jednocześnie żeby dane były dokładniejsze i łatwiejsze w dalszej analizie (mniejsza ilość błędów).

Ważna uwaga:

Nie należy mieć zbyt wygórowanych oczekiwań względem QC. Pamiętajmy, że kontrola jakości nie zmieni złych danych w dobre dane i nigdy nie uda nam się uratować tego, co jest źle zrobione (np. złe przygotowanie biblioteki, niskiej jakości RNA lub cDNA).

Kiedy dane sekwencjonowania od początku wyglądają naprawdę źle, najlepiej "ruszyć dalej" i zebrać nowe dane. Przeprowadzenie QC nie uratuje takich danych, będzie to tylko zmarnowany nakład pracy.

Im słabiej poznany jest genom nad którym prowadzimy badania, tym ważniejsze jest poprawianie wszelkich błędów.

Jeżeli aligment będzie wykonywany do dobrze poznanego genomu możliwe jest rozpoznanie i zidentyfikowanie błędów w czasie tego procesu. Podczas składania genomu de-novo, błędy mogą skutkować niepowodzeniem tego procesu dlatego ważne, aby w przypadku takich analiz zastosować
bardziej rygorystyczne filtrowanie.

Prawdopodobnie największą trudnością QC jest nasza niezdolność do przewidywania wszystkich możliwych czynników które wpływać na nasze dane.

- Pre-alignment: “raw data” - protokół postępowania identyczny dla każdego przypadku niezależnie jakie analizy będą wykonywane
- Post-alignment: “data filtering” - różne protokoły dla różnych typów analiz

Jak wykonujemy QC

1. Ocena (wizualna) jakość danych.
2. Jeśli ocena QC wydaje się zadowalająca kończymy proces.
3. Jeśli ocena QC jest nie zadowalająca, wykonujemy jeden lub więcej kroków modyfikujących jakość danych, a następnie przechodzimy do kroku 1.

Wiele narzędzi kontroli jakości to oprogramowanie niskiej jakość.
Dynamiczny rozwój technik NGS spowodował, że pisanie programów umożliwiających analizę danych NGS (szczególnie narzędzia QC) stało się "modne", ponieważ takie narzędzia uważano za "łatwe" do opublikowania. Wykorzystanie wyszukiwarki internetowej pozwala znaleźć dziesiątki takich narzędzi. Nie mniej większość z tych narzędzi wykorzystuje "naiwne rozwiązanie", które działają zaskakująco słabo w analizie rzeczywistych danych. Co więcej zaskakująca jest ilość niespójności w
działaniu tych narzędzi. Bardzo często wykonanie tych samych czynności, przy założeniu tych samych parametrów w różnych programach daje odmienne wyniki. Co gorsze istnieje niewiele obiektywnych mierników, standardów i zaleceń, które pozwalają określić dane jako "dobre"

Rekomendowane (subiektywna ocena!) narzędzia do QC to Trimmomatic, BBDuk, flexbar i cutadapt.

Najłatwiej (wstępnie) ocenić jakość odczytów sekwencjonowania poprzez analizę wykresu ilustrującego prawdopodobieństwo błędu, w każdej pozycji odczytu (uśrednione dla wszystkich odczytów). Pozioma oś ilustruje wartości (jakości odczytu 4 linijka pliku fastq) z pliku fastq, ilustrujące prawdopodobieństwo błędnego odczytania zasady.

Jak "jakość" jest kodowana w plikach fastq?
Phred quality score (Q) - system opracowany dla projektu "HGP" w celu ułatwienia automatyzacji procesu sekwencjonowania DNA, wartość Q oznaczana dla każdego nukleotydu. Wartość Phred quality score są
powszechnie akceptowane do oceny jakości odczytu sekwencjonowania DNA. Więcej informacji można znaleźć pod linkiem.

$$Q = log_{10}P$$

- wartość 10 odpowiada 10% błędów (1/10)

- wartość 20 odpowiada 1% błędów (1/100)
- wartość 30 odpowiada 0,1% błędów (1/1000)
- wartosć 40 odpowiada 0.01% błędów (1/10000)

Dwie skale:

- Phred33 (Sanger format); wartości 0 - 93 kodowane za pomocą symboli ASCII 33 to 126 (wartości w odczytach nie przekraczają 60) - Illumina ver 1.8+
- Phred64; wartości 0 - 62 kodowane za pomocą symboli ASCII (w surowych odczytach oczekujemy wartości 0 do 40) - Illumina ver 1.3 - 1.8

Więcej informacji pod linkiem.

Do wizualizacji jakości odczytów wykorzystamy narzędzie FastQC. Program uruchamiany z wiersza poleceń, generuje raporty w postaci pliku html.
Uwaga wstępna:
Jeżeli w raporcie widzimy czerwone znaki nie musimy się nimi (zbytnio) przejmować. Zazwyczaj nie mają one większego znaczenia. Jednak jeżeli stanowią one większość (lub wszystkie) punkty naszego raportu oznacza to że mamy problem z danymi. Do wygenerowania raportu używamy polecenia fastqc:

```
fastqc -h
fastqc  illumina.fq
```

Możemy użyć opcji --nogroup, która spowoduje zminę wyglądu raportu, wyłącza łączenie kolumn.

```
fastqc  illumina.fq --nogroup
```

Ocena jakości w zakładkach "Per base sequence quality" oraz "Per sequence quality scores"

Czym są adaptery sekwencyjne?

Podczas przygotowywania biblioteki unikalnie zaprojektowane adaptery DNA o długościach typowo ponad 30 zasad dodane są do końców 5 'i 3' każdej sekwencji (fragmentu DNA). Po dodaniu adapterów, jednoniciowe sekwencje (fragmenty DNA), które trafią do instrumentu, mają format:

XXXXAAAATTTTGGGGCCCCYYYY

gdzie XXXX i YYYY są adapterami. Aparaty do sekwencjonowania rozpoznają początkowe adaptery, ale zwykle nie wykrywają adapterów końcowych, jeśli zostaną one wprowadzone do odczytu sekwencji. Ma to miejsce kiedy długość odczytu przekracza długość fragmentu DNA, wtedy sekwencja adaptera może pojawić się w danych. Fastqc pozwala nam na zlokalizowanie uniwersalnych sekwencji adaptorów (zakładka Adapter Content). Dodatkowo możemy personalizować sekwencję adaptorów rozpoznawaną przez FastQC poprzez edycję plików konfiguracyjnych programu.

```
ls  ~/src/FastQC/Configuration/
```

```
more ~/src/FastQC/Configuration/adapter_list.txt
```

Czy adaptory muszą zostać ręcznie usunięte?

Zasadniczo jeżeli genom z którego pochodzą odczyty jest znany, adaptory nie powinny powodować problemów. Wiele wysokowydajnych alignmenterów może automatycznie przycinać odczyty i automatycznie usuwać adaptery podczas swojej pracy. Z drugiej strony obecność sekwencji adapterów może powodować znaczne problemy podczas składania nowych genomów i transkryptomów związku z
tym należy je usunąć przed rozpoczęciem tego procesu.

Wykorzystanie narzędzia trimmomatic

Pobieramy 10000 odczytów dla archiwum SRR1553607 (PE)

Sprawdzamy jakość odczytów (FastQC) dla obydwu plików

Który plik wydaje się być gorszej jakości?

Uruchamiamy trimmomatic

```
java -jar trimmomatic-0.36.jar SE ~/Genomika/fastq_QC/SRR1553607_2.fastq
~/Genomika/fastq_QC/trimmed_2.fq SLIDINGWINDOW:4:30
```

lub

```
java -jar ~/src/Trimmomatic-0.36/trimmomatic-0.36.jar SE
~/Genomika/fastq_QC/SRR1553607_2.fastq ~/Genomika/fastq_QC/trimmed_2.fq
SLIDINGWINDOW:4:30
```

Usuwanie adapterów

Proszę stworzyć plik (fasta) z sekwencja adapterów urządzenia ilumina (AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC)

Następnie pobrać archiwum SRR519926 (PE)

Wizualizacja jakości (FastQC), adaptery

Teraz możemy wykorzystać trimmomatic do usunięcia adapterów:

```
java -jar ~/src/Trimmomatic-0.36/trimmomatic-0.36.jar
SE SRR519926_1.fastq output.fq
ILLUMINACLIP:adapter.fa:2:30:5
```

Dla trybu PE

```
java -jar ~/src/Trimmomatic-0.36/trimmomatic-0.36.jar PE SRR519926_1.fastq
SRR519926_2.fastq trimmed_1.fq unpaired_1.fq trimmed_2.fq unpaired_2.fq
ILLUMINACLIP:adapter.fa:2:30:5
```

Zadanie 3:
Z archwum SRR 519926 usunąć odczyty niskiej okno 4 nt, Q<30, z końca odczytu usunąć nukleotydy o niskiej wartości Q <30, usunąć adaptory "Ilumina".
Uwaga:
Kolejność ma wpływ na rezultaty !!!, można eksperymentować z różnymi ustawieniami.

Zadanie 4:
Wykonać analizę jakości dla plików w folderze `/Dydaktyka/fastq/raw/`