Short Read Aligner

Sekwencjonowanie nowej generacji (NGS) produkuje (wiele milionów odczytów), krótkich sekwencji DNA. Przed dalszą analizą konieczne jest zlokalizowanie otrzymanych odcinków (odczytów) w obrębie genomu (lub transkryptomu) referencyjnego. Do mapowania używane jest oprogramowanie typu "Short Read Aligner". Istnieje wiele programów typu "Short Read Aligner" jednak w przypadku analizy danych RNAseq musimy posłużyć się oprogramowaniem, które jest wrażliwe na zjawisko splicingu "splice aware". Przykłady programów mapujących wrażliwych na zjawisko samplingu to:

- hisat2
- STAR

Mapowanie/Aligment

Mapowanie

- Określenie rejonu genomu z którego pochodzi odczyt
- Mapowanie jest poprawne kiedy prawidłowo ustalimy położenie odczytu we genomie

Alignement

- Alignement to prawidłowe umiejscowienie każdej zasady w sekwencji referencyjnej
- Alignement jest poprawny kiedy każda z zasad w sekwencji jest prawidłowo przyporządkowana do sekwencji referencyjnej

Do analiz RNAseq chcemy mieć poprawne mapowanie!

Mapowanie za pomocą STAR

Mapowanie odczytów za pomocą alignera `STAR` składa się z dwóch etapów:

1. Utworzenie indeksu dla genomu referencyjnego
2. Mapowaniem odczytów do genomu referencyjnego

Ćwiczenie 1

Po zalogowaniu na serwer proszę utworzyć w swoim katalogu domowym katalog `STAR_ref_genom`.

Do utworzenia indeksu w programie `STAR` należy uruchomić tryb `genomeGenerate`. Parametry które należy uzupełnić w czasie tworzenia indeksu genomu to:

```
--runThreadN Liczba wątków procesora do użycia w obliczeniach
--runMode genomeGenerate
--genomeDir /ścieżka/do/katalogu/indeksu/genomu
--genomeFastaFiles /ścieżka/do/pliku/fasta/genomu
--sjdbGTFfile /pścieżka/do/anotacji/genomu
--sjdbOverhang dł odczytów-1
```

Jeżeli genom jest bardzo mały dodatkowo powinniśmy zmienić wartość parametru --genomeSAindexNbases na mniejszą wartość korzystając z równania $min(14, log2(GenomeLength)/2 - 1)$.

Ćwiczenie 2

Proszę określić wartość parametru `--genomeSAindexNbases`, dla genomu o sekwencji podanej w pliku `/Dydaktyka/sample_data/refs/22.fa`.

Ćwiczenie 3

Proszę utworzyć indeks dla genomu (chromosom 22). Jako katalog genomu proszę wykorzystać katalog `STAR_ref_genom`. Plik fasta z sekwencją genomu oraz plik gtf z adnotacją genomu znajduje się w katalogu `/Dydaktyka/sample_data/refs/`. W obliczeniach proszę użyć 2 wątków procesora oraz wartości dla parametru `--genomeSAindexNbases` obliczonej w poprzednim zadaniu (część całkowita).

Po utworzeniu indeksu dla genomu referencyjnego możemy przejść do mapowania odczytów.

Aby wykonać mapowanie musimy uruchomić program STAR z parametrami:

```
--runThreadN Liczba wątków procesora do użycia w obliczeniach
--genomeDir /ścieżka/do/katalogu/indeksu/genomu
--readFilesIn /ścieżka/do/odczytu1 [/ścieżka/do/odczytu2 ]
```

Ponieważ program `STAR` wykonuje mapowanie bardzo powoli, aby przyspieszyć ten proces utrzowymy pliki z wybranymi (losowo) 100 odczytami. W tym celu wykorzystamy narzędzie `seqtk`

```
mkdir reads
seqtk sample -s100 /Dydaktyka/sample_data/reads/UHR_1_R1.fq 100 >
reads/UHR_1_R1.fq
seqtk sample -s100 /Dydaktyka/sample_data/reads/UHR_1_R2.fq 100 >
reads/UHR_1_R2.fq
```

Ćwiczenie 4

Proszę zmapować odczyty w plikach `../reads/UHR_1_R1.fq` oraz `../reads/UHR_1_R2.fq za pomocą programu` STAR` oraz indeksu genomu utworzonego w poprzednim ćwiczeniu.

Zadanie 1

Proszę stworzyć skrypt, który wykona mapowanie dla wszystkich plików w folderze `reads` za pomocą programu `STAR`.

Zadanie 2

Proszę stworzyć skrypt, który wykona mapowanie dla wszystkich odczytów próbek UP1-UP6 uzyskanych po filtracji danych w poprzednich ćwiczeniach. (Index genomu dla łubinu znajduje się w katalogu /DANE/STAR_index/Lupin).

Mapowanie za pomocą hisat2

Mapowanie za pomocą programu `hisat2` wygląda analogicznie jak mapowanie za pomocą programu `STAR` i składa się z dwóch zasadniczych etapów:

1. Utworzenie indeksu dla genomu referencyjnego
2. Mapowanie odczytów do genomu referencyjnego

Do budowy indeksu w programie `hisat2` służy polecenie `hisat2-build`. Więcej na temat tego programu możemy się dowiedzieć korzystając z pliku pomocy.

Ćwiczenie 5

Proszę utworzyć katalog `hisat_ref_genom` w swoim katalogu domowym, a następnie utworzyć indeks dla genomu referencyjnego za pomocą polecenia `hisat2-build` w tym katalogu.

Po utworzeniu indeksu dla genomu referencyjnego możemy przystąpić do mapowania odczytów do genomu referencyjnego.

Ćwiczenie 6

Proszę odczyty w plikach `../reads/UHR_1_R1.fq` oraz `../reads/UHR_1_R2.fq` z katalogu `/Dydaktyka/sample_data/` za pomocą programu `STAR` oraz indeksu genomu utworzonego w poprzednim ćwiczeniu. Proszę zmienić parametry dla opcji:

```
--dpad
--gbar
--mp
--np
--n-ceil
--no-mixed
--no-discordant
```

Proszę porównać uzyskane wyniki (procent zmapowanych odczytów) z innymi osobami w grupie

Zadanie 3

Proszę stworzyć skrypt, który wykona mapowanie dla wszystkich plików w folderze `reads` za pomocą programu `hisat2`.

Zadanie 4

Proszę stworzyć skrypt, który wykona mapowanie dla wszystkich odczytów próbek UP1-UP6 uzyskanych po filtracji danych w poprzednich ćwiczeniach. (Index genomu dla łubinu znajduje się w katalogu /DANE/STAR_index/Lupin).

Zadanie 5

Proszę zmapować wszystkie odczyty z katalogu `reads` do genomu `ERCC92`, znajdującego się w katalogu `refs`, za pomocą alignera `hisat2`.