Short Read Aligners

NGS

- Sekwencjonowanie krótkich odcinków (50 – 300 pz)
- Jako wynik otrzymujemy olbrzymią liczbę odczytów (fragmentów DNA) liczoną w setkach milionów

Standardowe sekwencjonowanie vs NGS

- Wczesne algorytmy aligmentu (dopasowania) zostały zaprojektowane w celu znalezienia większości (jeśli nie wszystkich) dopasowań (BLAST)
- Rozwój technik NGS spowodował iż zaistniała potrzeba opracowania nowych i szybkich metod, które mogą bardzo szybko wybrać najlepszą lokalizację dla każdego z odczytów (kosztem braku zlokalizowania wszystkich potencjalnych alternatywnych dopasowań)

Short reads aligners

- Uzasadnieniem nowego paradygmatu było traktowanie sekwencji jako krótkich, niezależnych pomiarów, które będą następnie porównywane ze znanym genomem lub względem siebie nawzajem
- Doprowadziło to do rozwoju całej grupy algorytmów dopasowywania, często określane jako "short read aligners" lub „short read
- mappers"

Short reads aligners

- powszechnie stosowane narzędzi programistyczne w bioinformatyce, zaprojektowane tak, aby wyrównać (align) bardzo duże liczba krótkich odczytów
  - Dominacja urządzeń Illumina spowodowała, że większość działa w zakresie, w jakim te instrumenty generują dane (50-300 pz).

Jak działają Short reads aligners

- Ograniczenia
  1. Większość programów znajdzie jedynie dopasowania, które są „wystarczająco” podobne do celu. To znaczy, że algorytm "rezygnuje" z wyszukiwania poza pewien próg (podobieństwa).
  2. Algorytmy mają za zadanie znaleźć obszary o dużym podobieństwie
  3. Większość programów nie analizuje długich odczytów lub staje się nieefektywna w czasie ich analizy
  4. Istnieje również granica tego, jak krótki może być odczyt

Mapping vs Alignment

- Mapowanie to zlokalizowanie obszaru genomu, z którego pochodzi sekwencja odczytu.
- Mapowanie jest uważane za poprawne, jeśli pokrywa się z prawdziwym regionem.

- Alignment to szczegółowe umieszczenie każdej zasady w odczycie.
- Dopasowanie jest uważane za poprawne, jeśli każda baza jest umieszczona poprawnie.

Mapping vs Alignment

- Poprawne mapownie, niepoprawny aligment (dopasowanie)

- Poprawny aligment (dopasowanie) nieprawidłowe mapowanie

Mapping vs Alignment

- Aligment
  - Identyfikacjia SNP
  - Identyfikacjia variantów genomowych – genotypowanie
- Mapping
  - RNAseq (interesuje nas z jakiego genu pochodzi odczyt, nie jesteśmy szczególnie zainteresowani różnicami w sekwencji)

Jak wybrać program do aligmentu danych

- Algorytm: globalny, lokalny lub pół-globalny?
- Czy istnieje potrzeba zastosowania aligmentu nielinearnego?
- W jaki sposób „radzi sobie” z mutacjami INDEL?
- Czy aligner może opuszczać (lub łączyć) duże obszary (alternatywny splicing) ?
- Czy opcje filtrowania mogą być dostosowane do naszych potrzeb?

Jak wybrać program do aligmentu danych

- BWA (algorytm z najnowszej wersji nie został opublikowany w recenzowanym czasopiśmie !)
  - https://gist.github.com/ialbert/3164967c853b7fd8f44e
- Bowite2

- Histat2 – implementacja algorytmu Bowite2

Jak działają programy do aligmentu danych

- Zasadniczo wszystkie programy działają na tych samych zasadach:

  1. Najpierw budowany jest indeks ze znanego źródła (należy to zrobić tylko raz) – genom feferencyjny.

  2. Odczyty w formacie FASTA / FASTQ są następnie przyrównywane względem tego indeksu.

BWA

- Budowanie indexu
  - Polecenie bwa index

bowite2

- Budowanie indexu
  - bowtie2-build

HISAT2

- Budowanie indexu
  - Polecenie hisat2-build
  - Można stosować zarówno do dużych jak i małych (poniżej 4 miliardów) genomów
  - Małe genomy pliki ht2 (liczby 32-bitowe)
  - Duże genomy ht21 (liczby 64-bitowe)

HISAT2

```
hisat2 [options]* -x <hisat2-idx> {-1 <m1> -2 <m2> | -U <r> | --sra-acc <SRA accession number>} [-S <hit>]
```

Main arguments

`-x <hisat2-idx>`	The basename of the index for the reference genome. The basename is the name of any of the index files up to but not including the final `.1.ht2` / etc. `hisat2` looks for the specified index first in the current directory, then in the directory specified in the `HISAT2_INDEXES` environment variable.
`-1 <m1>`	Comma-separated list of files containing mate 1s (filename usually includes `_1`), e.g. `-1 flyA_1.fq,flyB_1.fq`. Sequences specified with this option must correspond file-for-file and read-for-read with those specified in `<m2>`. Reads may be a mix of different lengths. If `-` is specified, `hisat2` will read the mate 1s from the "standard in" or "stdin" filehandle.
`-2 <m2>`	Comma-separated list of files containing mate 2s (filename usually includes `_2`), e.g. `-2 flyA_2.fq,flyB_2.fq`. Sequences specified with this option must correspond file-for-file and read-for-read with those specified in `<m1>`. Reads may be a mix of different lengths. If `-` is specified, `hisat2` will read the mate 2s from the "standard in" or "stdin" filehandle.
`-U <r>`	Comma-separated list of files containing unpaired reads to be aligned, e.g. `lane1.fq,lane2.fq,lane3.fq,lane4.fq`. Reads may be a mix of different lengths. If `-` is specified, `hisat2` gets the reads from the "standard in" or "stdin" filehandle.

SAM - **Sequence Alignment Maps**

- Plik tekstowy
- Rozdzielony znakami tabulacji
- Zawiera dwie sekcje
  - Nagłówek (jest sekcją opcjonalną, zaczyna się znakiem @)
  - Alignment (każda linijka zawiera informacje o aligmencie), 11 pól zawierających informacje o aligmencie
- Specyfikacja http://samtools.github.io/hts-specs/SAMv1.pdf

BAM

- Binarna wersja SAM
- BAM jest skompresowaną (i najczęściej posortowaną) wersją SAM
- Sortowanie na podstawie
  - Lokalizacji (pozycji) - najczęściej
  - Nazwy odczytów
- Wymiana i przechowywanie danych w formacie BAM jest bardziej efektywne

SAM – cel formatu

- Umożliwia przedstawienie setek milionów dopasowani (aligmentów)
- Metody przetwarzania informacji w plikach SAM / BAM
- Informacja zawarta w tym pliku określa sukces każdej analizy
- Tworzenie tego pliku tak, aby faktycznie zawierał potrzebne informacje zawsze powinny być naszym priorytetem

SAM – co zawiera

- Pliki SAM/BAM utworzone przez różne programy mogą zawierać różną ilość informacji
- Różnice **nie** wynikają jedynie z dokładności i efektywności wykorzystanego narzędzia
- Standar wymaga aby plik SAM zawierał 11 podstawowych kolumn z ściśle określonymi informacjami

SAM – co zawiera

- Jak się okazuje plik SAM utworzone przez różne programy zawierają różne ilości informacji, a różnice między nimi można podzielić na dwa główne kategorie:
  - Które aligmenty są (lub nie są) zareportowane w pliku SAM
  - Informacje w polach poza wymaganym minimum przez standard

SAM – co zawiera

W praktyce konwersja „standardowego” pliku SAM generowanego przez jeden program w inny „standardowy” plik SAM generowany przez inny program jest zadaniem bardzo trudnym, a często wręcz niemożliwym do osiągnięcia

SAM – co zawiera

- Format SAM jest bardzo "skąpy„
- Każdy aligner uzupełni tyle ile informacji, ile wie jak uzupełnić

Co oferują pliki SAM

- Szybki dostęp do aligmentów, które nakładają się na współrzędne. Na przykład wybieramy aligmenty, które się pokrywają z współrzędną 323,567,334 na chromosomie 2
- Łatwy wybór i filtrowanie odczytów na podstawie atrybutów. Na przykład chcemy szybko wybrać aligmenty, które są wyrównane do sekwencji o odwróconej orientacji (-)
- Wydajne przechowywanie i dystrybucja danych. Na przykład, jeden skompresowany plik zawiera dane dla wszystkich próbek jednego typu

BAM to FASTQ

- Ponieważ pliki BAM przechowują skompresowane informacje zdarza się (coraz częściej), że informacjie z „surowych” odczytów przechowuje się jako niewrównane („**unaligned”)** pliki BAM
- Pliki BAM w takiej postaci nie mogą być analizowane, aby można było analizować przechowywane w nich informację konieczne jest ponowne ich przekonwertowanie w pliki FASTQ

Tworzenie plików SAM/BAM

- Pobranie/otrzymanie surowych danych odczytów (pliki FASTQ)
- Pobranie sekwencji genomu referencyjnego
- Indexowanie genomu referencyjnego
- Wyrównanie (aligment) pliku(ów) FASTQ (otrzymujemy plik SAM)
- Sortowanie pliku SAM (samtools)
- Zbudowanie indexu dla pliku BAM (samtools)