1. System analizy danych NGS z paneli genów

(programistyczny)

Sekwenator to instrument odczytujący sekwencję DNA w kilku-kilkudziesieciu probkach na raz. Instrument zapisuje na dysku dane w skompresowanych plikach tekstowych (FASTQ). Zadaniem systemu jest automatyczne wykrycie nowych danych, wyslanie maila na zdefiniowane adresy email, i uruchomienie procesu przetwarzania danych (pipeline, skrypt Python). Pipeline wykrywa mutacje w sekwencjach DNA i generuje statystyki dotyczące danych wejściowych, pośrednich i wyników. Sam pipeline nie jest częścią tego zadania (patrz niżej). Po zakończeniu pipelinu system wysyla powiadomienia emailowe z linkami umożliwiajacymi wyświetlenie statystyk w postaci wykresow i tabel, oraz ściągnięcie wyników (plik tekstowy) do dalszej obróbki i/lub analizy. Całość powinna działac w kontenerze dockera.

Lub

(programistyczno - analityczny)

Implementacja pipelinu do analizy danych NGS z paneli genów (patrz wyżej) i porównanie wyników analiz z wykorzystaniem 2-3 narzędzia do mapowania (np. BWA, bowtie) i 2-3 narzedzi do wykrywania wariantów (np. samtools, freebayes). Porównanie pokrycia i wykrytych wariantów dla ok. 100 próbek, w każdej po ok. 300 genów.

Więcej informacji: pawel.sztromwasser@umed.lodz.pl

2. Federacja baz danych częstości wariantów

(programistyczny)

Częstość wariantów (SNP) w populacji jest ważną informacją, używaną między innymi w poszukiwaniu genetycznych przyczyn rzadkich chorób wrodzonych. Informacje nt. częstości wariantów są jednak rozproszone po wielu instytucjach przeprowadzajacych badania genetyczne (m.in jednostkach badawczych, szpitalach, klinikach). Im większa jest grupa zbadanych osób tym dokładniejsza informacja nt. częstości wariantów, warto więc zintegrować takie dane.

Dane genetyczne ludzi/pacjentów są poufne i dzielenie się nimi otwarcie jest zabronione. Udostępnienie dużej bazy danych zawierającej wyłącznie informacje o częstotliwości występowania wariantów jest mniej problematyczne. Aby uniknąć problemów związanych z centralnym przechowywaniem danych, należy stworzyć federację baz danych o wspólnym interfejsie, który umożliwi agregację informacji o częstości występowania wariantów z dowolnej liczby baz. W ten sposób każda z baz będzie niezależna i będzie zawiarała dane z jednego ośrodka, przetworzone w sposób taki jaki dany ośrodek uzna za najlepszy. Interfejs użytkownika umożliwi pobranie informacji nt. częstości pojedynczego wariantu z wybranych baz danych i przeliczenie częstości jego występowania w zbiorczej populacji. (Uprzywilejowani?) użytkownicy będą mieli możliwość pobrania całej bazy danych i stworzenia lokalnej kopii 'meta-bazy'.

Wymagania dot. oprogramowania:

- zaprojektowanie i stworzenie bazy danych czestosci wariantow zawierajacych przynajmniej:
    - CHR - chomosom
    - POS - pozycja nukelotydowa w chromosomie
    - REF - allel referencyjny
    - ALT - allel alternatywny
    - AC - liczba alleli alternatywnych w bazie
    - AN - calkowita liczba przebadanych alleli (w tej pozycji)
- zaprojektowanie i stworzenie REST API do:
    - skladania zapytan o czestosc wariantu dla krotki (CHR, POS, REF, ALT)
    - sciagania calosci bazy danych
- stworzenie interfejsu uzytkownika do przeszukiwanai federacji baz (predefiniowane URL)
- stworzenie interfejsu uzytkownika dodawania nowych danych do bazy z pliku VCF (jedno i wieloprobkowego) oraz tekstowego
- test na danych 1000Genomes i ExAC

- opcjonalnie: wizualizacja pokrycia genów (ilość próbek/exon)

Część analityczna:
- porównanie częstośći w ExAC i 1000Genomes
- czy są warianty o diametralnie różnych częstościach?
- jak rozkladaja sie rzadkie warianty wzgledem genow/chromosomow?

Więcej informacji: pawel.sztromwasser@umed.lodz.pl

3. Analiza ultrarzadkich wariantów cichych w bazie gnomAD

(analityczny)

Warianty ciche to punktowe zmiany w DNA nie wpływających na sekwencję kodowanego białka (https://pl.wikipedia.org/wiki/Mutacja_cicha). Ewolucyjnie są lepiej tolerowane niż warianty kodujące, czyli takie które zmieniają kodowany aminokwas. Ciekawą grupą będą więc ultra rzadkie warianty ciche i odpowiedź na pytanie dlaczego ich częstość występowania w populacji jest ograniczona.

Baza danych gnomAD (http://gnomad.broadinstitute.org/) zawiera dane o częstości występowania wariantów w kodujących fragmentach genów (egzonach) dla ponad 130 000 osób z różnych populacji. Jest to największa taka baza na świecie i stanowi świetne źródło do poszukiwania odpowiedzi na powyższe pytania.

Przykladowe pytania:

1. Gdzie zlokalizowane są bardzo rzadkie warainty ciche względem sekwencji kodujących (początek, środek, koniec egzonu; który egzon)
2. Czy rzadkie warianty ciche są zlokalizowane pomiędzy wariantami częstszymi?
3. Czy ich lokalizacja nakłada się na:
    - miejsca wiązań czynników transkrypcyjnych
    - miejsca wiazan miRNA,/lincRNA
    - inne regiony regulatorowe
4. Czy ich lokalizacja koreluje się z:
    - ewolucyjną konserwacją
    - scorem dot. wpływu na splicing genu (baza Spidex)
    - ...
5. Czy jest związek z ilością rzadkich wariantów cichych w genie (ew. stosunku cichych rzadkich/częstych), a:
    - długością genu
    - funkcją genu (Gene Ontology)
    - związku genu z chorobami
    -

4. Porównanie metod składania genomów bakteryjnych

(analityczny)

Składanie genomu *de-novo* polega na ułożeniu kompletnej sekwencji genomu z krótszych fragmentów. Im dłuższe fragmenty tym zadanie jest prostsze.

Pod uwagę bierzemy dwie platformy sekwencjonowania: Illumina Miseq oraz Pacbio. Sekwenatory Miseq generują tanio dużą ilość (pokrycie genomu bakterii 300x), dobrej jakości (<1% błędu), ale krótkich odczytów (pary 2x250nt). Samymi krótkimi odczytami jesteśmy w stanie złożyć genom *E. coli* co najwyżej w kilkadziesiąt kawałków. Nie zawsze jest to wystarczające. Sekwenator Pacbio generuje podobne pokrycie długimi odczytami (~ 10000nt) o 15% częstości błędu, co pozwala złożyć genom w jedną całość. Natomiast jego użycie jest kilkukrotnie droższe. Alternatywą jest połączenie w składaniu tanich odczytów Illuminy z niewielką ilością Pacbio. W tym projekcie będziesz mogła/mógł sprawdzić w praktyce czy rzeczywiście się to opłaci.

Porównanie składania 6 genomów *E. coli* i 1 *Salmonella enterica* czterema metodami:

1. krótkie dokładne odczyty Illumina (2x250nt; 1% error rate) u użyciem SPAdes
2. długie niedokładne odczyty Pacbio (~3000nt lub ~10000nt; 15% error rate) z użyciem Canu
3. podejście hybrydowe (długie i krótkie odczyty):
    1. SPAdes - używa krótkich odczytów jak w (1) a potem łączy je w dłuższe z użyciem długich
    2. Canu - używa długich odczytów jak w (2) i poprawia jakość z użyciem krótkich

Dla danych Pacbio są różnice w pokryciu pomiędzy próbkami, co jest dodatkowym czynnikiem do przeanalizowania.