Analiza wyników sekwencjonowania genomowego (NGS) część 1. Annotacja i interpretacja wariantów sekwencyjnych.

Dane, z którymi będziemy pracować pochodzą z projektu sekwencjonowania 1000 genomów osób z Polski[1]. Jest to kompletna sekwencja genomu jednego dorosłego mężczyzny, bez diagnozy chorób rzadkich, uzyskana na platformie Illumina (krótkie odczyty) z pokryciem ok. 30x. Przed ćwiczeniami przeprowadzono najbardziej czasochłonne etapy analizy (ich przeprowadzenie trwało 2-3 dni przy wykorzystaniu silnej wieloprocesorowej stacji roboczej):

Mapowanie Odczyty NGS zostały zmapowane do sekwencji referencyjne genomu człowieka GRCh38 (uwaga: we wszystkich analizach genomu człowieka bardzo ważne jest to, która wersja sekwencji referencyjnej została użyta).

Identyfikacja wariantów. Na podstawie zmapowanych odczytów zidentyfikowano warianty różniące tę sekwencję od sekwencji referencyjnej. Wykorzystano pakiet GATK4[2] z filtrowaniem na podstawie jakości odczytów[3]. Zidentyfikowano w ten sposób 3 813 432 warianty SNP i 915 984 indeli.

Pytanie 1 *Porównaj te wyniki z danymi z projektu 1000 genomów[4]*

Wyniki zapisane zostały w pliku w formacie VCF.

Wstępna annotacja wariantów

Annotację wariantów przeprowadzono za pomocą programu Ensembl Variant Effect Predictor (VEP)[5]. W tej analizie warianty lokalizowane są w odpowiednich genach, przewidywany jest także ich wpływ na funkcję genu. Uzyskujemy w ten sposób informację o tym, czy wariant jest zlokalizowany w obszarach kodujących, intronowych, leżących powyżej lub poniżej genu, itp. Na podstawie tych informacji obliczany jest przewidywany wpływ na funkcję genu (ang. *impact*) i wariant zaliczany jest na jego podstawie do jednej z kategorii:

- HIGH - to warianty, które z dużym prawdopodobieństwem powodują utratę funkcji genu, np. mutacje nonsens, frameshift, itp.
- MODERATE - to warianty mające umiarkowany wpływ na funkcję genu, np. mogące wpłynąć na wydajność składania transkryptu
- LOW - to warianty o niewielkim przewidywanym efekcie fenotypowym, np. zmiany synonimiczne albo w obszarach niekodujących
- MODIFIER - to warianty, których efekt jest trudny do przewidzenia, przeważnie w obszarach niekodujących, intronach (poza miejscami ważnymi dla składania), itp.

Przewidywane konsekwencje wariantu opisywane są za pomocą standaryzowanych terminów *Sequence Ontology*, przedstawionych na rysunku poniżej (ze strony ensembl.org)

---

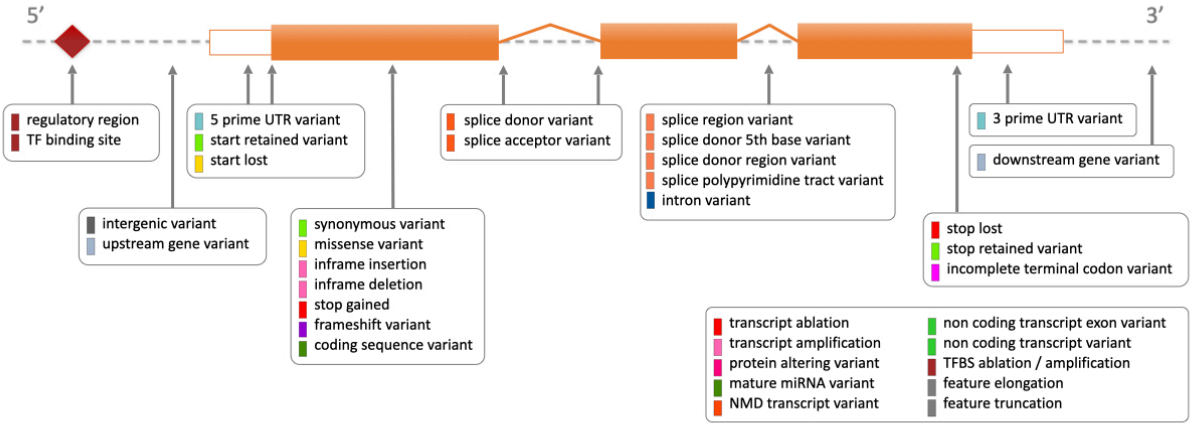
[1] Kaja E, et al. The Thousand Polish Genomes-A Database of Polish Variant Allele Frequencies. Int J Mol Sci. 2022 Apr 20;23(9):4532.

[2] https://gatk.broadinstitute.org

[3] https://gencore.bio.nyu.edu/variant-calling-pipeline-gatk4/

[4] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). https://doi.org/10.1038/nature15393

[5] https://www.ensembl.org/info/docs/tools/vep/index.html



Rysunek 1. Klasyfikacja *Sequence Ontology* efektów wariantów.

Dodatkowo przeszukano bazę danych dbSNP[6] dla stwierdzenia, czy zidentyfikowany wariant znajduje się w bazie danych, a także porównano uzyskane warianty z bazami danych o częstości alleli w populacjach (gnomAD, 1000 genomes). Programem VEP można przeprowadzić jeszcze wiele różnych innych analiz, a stosowana w nim do opisu wariantów terminologia jest standardem w genomice człowieka, dlatego warto zapoznać się z jego opisem.

Annotacje zapisano w pliku VCF zawierającym zidentyfikowane warianty w postaci dodatkowych kolumn (Extra VCF Info Annotations).

Za pomocą narzędzia bcftools[7] podzielono plik VCF na mniejsze zawierające po 2-3 chromosomy. Pliki VCF mogą być kompresowane dla oszczędności miejsca, większość programów wczytuje pliki skompresowane algorytmem gzip.

Dalszą annotację będziemy prowadzić na ćwiczeniach w trybie interaktywnym.

Ćwiczenie: annotacja i analiza wariantów w sekwencji genomu człowieka

Do dalszej analizy wykorzystamy proste interaktywne narzędzie OpenCRAVAT[8] Pozwala ono na zannotowanie wariantów z wczytanego pliku VCF (skompresowanego) przez porównanie z różnymi bazami danych (ponad 150 różnych źródeł), w tym analiz, które już przeprowadzono za pomocą VEP (nie będziemy ich powtarzać).

Program uruchamiamy z terminala wpisując komendę:

```
oc gui
```

W pierwszej kolejności należy kliknąć w zakładkę "Store", która służy do instalacji źródeł danych. Należy upewnić się, że zainstalowane i zaktualizowane są następujące bazy: *Gene Ontology, ClinVar, GWAS Catalog, DGIdb: The Drug Interaction Database* oraz *PharmGKB*. Przeczytaj opis tych baz danych.

---

[6] https://www.ncbi.nlm.nih.gov/snp/

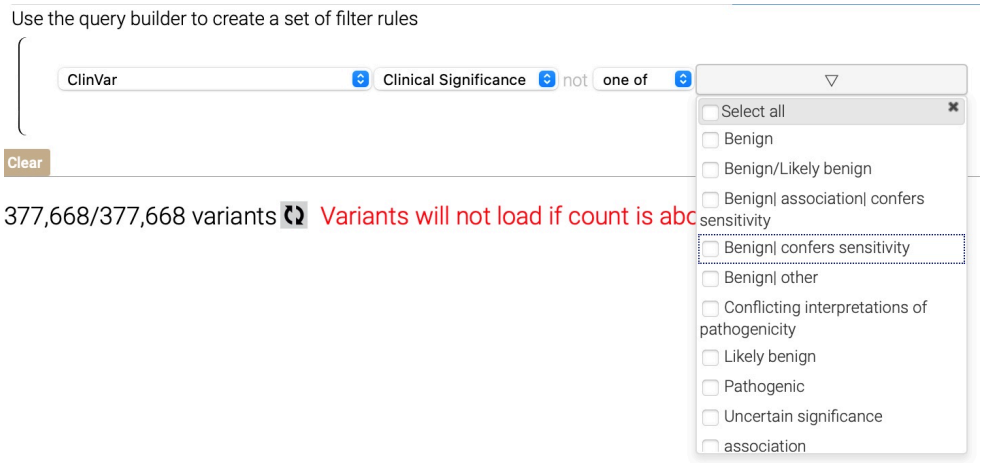
[7] https://samtools.github.io/bcftools/

[8] https://opencravat.org/

Następnie w zakładce "Jobs" klikamy "Add input file(s)" i wczytujemy wskazany plik VCF. Teraz w "Annotations" klikamy "Show all" i wybieramy zainstalowane wcześniej bazy danych. Teraz wystarczy kliknąć "Annotate" i zaczekać obserwując pole "Status". Gdy pojawi się opcja "Open Result Viewer", możemy przystąpić do analizy wyników.

Program nie jest w stanie wyświetlić więcej, niż 100 000 wariantów, trudno zresztą taką liczbę wariantów zinterpretować. Konieczne jest zatem odpowiednie ich filtrowanie. Posłuży do tego zakłada "Filter". Domyślne opcje "Smart filters" są bardzo ograniczone, skorzystajmy więc z zakładki "Query builder". Klikając "+" można dodawać kolejne filtry. Różne filtry można łączyć logicznym "and" lub "or". Możliwości jest wiele, proponuję zacząć od następujących:

1. Sprawdźmy, które warianty mają potencjalnie istotne znaczenie w bazie ClinVar, jak na rysunku poniżej. Z pewnością należy uwzględnić warianty opisane jako "Pathogenic" i "Likely pathogenic", a także "Conflicting interpretations", pozostałe zależnie od tego, jak wiele ich będzie.



Rysunek 2. Konstrukcja przykładowego filtru.

Przy analizie wyników należy zwrócić uwagę na następujące czynniki:

- allel referencyjny ("Ref base") i znaleziony w genomie ("Alt base"). Ważne jest oczywiście, czy wariant jest w postaci homozygotycznej, czy heterozygotycznej - tę informację znajdziemy w bloku "VCF Info" w kolumnie "Zygosity".
- w bloku "Variant annotation" znajdziemy też informacje o tym, w jakim genie jest wariant, czy wariant zmienia sekwencję kodującą i jaki jest jego efekt w klasyfikacji *Sequence ontology* (Rys. 1). W bloku "Gene ontology" będzie więcej informacji o genie i jego funkcji.
- W bloku "Extra VCF Info Annotations" (annotacje wcześniej dodane programem VEP) możemy znaleźć informację o częstości występowania wariantu w populacjach projektu 1000 genomów (kolumna "CSQ MAX AF") i bazy gnomADe ("CSQ GnomADe AF"). Czy częsty wariant może być istotny dla ryzyka rzadkiej choroby? A *vice versa*?
- Zawsze warto wejść do odpowiedniego wpisu w bazie ClinVar - tam znajdziemy informacje o tym, który genotyp może ewentualnie być związany z ryzykiem (nie zawsze ten, który akurat jest w badanym genomie), a także źródła, w tym publikacje.
- Kolumna "CSQ Existing Variation") zawiera odniesienie do bazy dbSNP, jeżeli ten wariant był kiedykolwiek opisywany. Ten sam identyfikator znajdziemy też w bloku "Variant Annotations" w kolumnie "Tags". Sprawdź ten identyfikator w bazie danych dbSNP i na SNPedia.

- W blokach kolumn "DGIDb" i "PharmGKB" znajdziemy informacje o znaczeniu wariantów w farmakogenomice, a w bloku "GWAS Catalog" o ewentualnych asocjacjach (bazy GWAS Catalog i ClinVar nie zawsze się pokrywają).

  Uwaga: bloki kolumn można rozwijać (znak "+") lub zwijać ("-"), można też wyniki sortować klikając na nagłówek kolumny. Kliknięcie na "Layout" pozwoli wybrać, które kolumny mają być widoczne.

2. Podobnie, jak w p. 1, utwórz filtr dla wariantów, które są annotowane w bazie "GWAS Catalog". Czy wyniki GWAS dla pojedynczych SNP zawsze mogą być podstawą predykcji/ diagnozy? Zwracaj uwagę na częstość wariantu w populacjach.
3. Zawsze pozostaje możliwość znalezienia nowego wariantu, który nie jest jeszcze opisany w bazach danych typu ClinVar. Warto przyjrzeć się wszystkim wariantom, dla których przewidywany wpływ na funkcję genu jest duży (filtr: "Extra VCF Info Annotations", "CSQ IMPACT", "equals, "HIGH") lub umiarkowany ("MODERATE"). Kolumna "CSQ Existing Variation") zawiera odniesienie do bazy dbSNP, jeżeli ten wariant był kiedykolwiek opisywany.

Pamiętajmy, że większość annotacji to wyniki działania automatycznych narzędzi. Zanim na podstawie sekwencji genomu wyciągnie się jakiekolwiek wnioski, które można przekazywać pacjentowi, należy bardzo starannie sprawdzić źródła i literaturę dotyczącą zidentyfikowanych wariantów. Niedopuszczalne jest przekazywanie pacjentom surowych wyników annotacji - muszą być wcześniej skontrolowane i zinterpretowane przez odpowiednio wykształconych specjalistów!