
Analiza wyników sekwencjonowania genomowego (NGS) część 2.

Wielogenowe czynniki ryzyka.

Pojedyncze warianty asocjujące z danym fenotypem rzadko mogą być przydatne w przewidywaniu prawdopodobieństwa wystąpienia cechy wieloczynnikowej. Wielogenowe czynniki ryzyka (*Polygenic Risk Scores, PRS*) uwzględniają wkład wielu różnych wariantów w fenotyp, warianty asocjujące z większym lub mniejszym ryzykiem są wliczane z uwzględnieniem siły korelacji jako wagi.

Udostępniona niedawno¹ baza danych² wylicza takie współczynniki dla wszystkich badań GWAS znajdujących się w bazie GWAS Catalog³. Autorzy stworzyli także narzędzie linii komend, które pozwala na obliczenie wielogenowych współczynników ryzyka dla zbadanego genomu. Plikiem wejściowym jest VCF ze wszystkimi wariantami (tu nie można już dzielić na chromosomy), należy także podać, z jakiej populacji pochodzi badana osoba. Wyniki zwracane są w postaci tabeli w pliku tekstowym (format .tsv, kolumny rozdzielone tabulatorami). Plik ten można zaimportować do arkusza kalkulacyjnego i w nim sortować i filtrować. Ten etap został już przeprowadzony, ćwiczenie polegać będzie na interpretacji wyników.

Ćwiczenie.

Otwórz wskazany plik, zapoznaj się z opisami kolumn. Poza oczywistymi (identyfikator badania, nazwa fenotypu, itp.) zwróć uwagę na następujące:

- Score Type: PRS może być podany jako iloraz szans (*Odds ratio, OR*) wskazujący krotność prawdopodobieństwa wystąpienia fenotypu w stosunku do populacji. $OR > 1$ oznacza wzrost ryzyka. OR podaje się dla cech o charakterze dyskretnym (zerojedynkowym). Współczynnik **beta** jest natomiast stosowany dla cech ilościowych i oznacza, o ile jednostek zwiększa się (lub zmniejsza) poziom cechy. Dla jasności lepiej najpierw skupić się na cechach dyskretnych, a później na ilościowych przy sortowaniu i wyborze potencjalnie interesujących wyników - wartości OR i beta nie mogą być porównywane między sobą! Wykorzystaj funkcję filtrowania w arkuszu kalkulacyjnym, żeby analizować tylko OR lub tylko beta. Następnie użyj sortowania dla odnalezienia najwyższych i/lub najniższych wartości.
- Percentile: Ta kolumna informuje, w którym centylu w populacji (dane zaczerpnięte z UK Biobank) mieści się dana wartość OR lub beta. Centyle obliczane są dla populacji odpowiadającej zadeklarowanemu pochodzeniu badanej osoby. Ogólnie - znaczące są wyniki w górnych lub dolnych centylach, wyniki w środkowych centylach oznaczają średnią populacyjną.
- *Risk Variants*: warianty asocjujące z większym ryzykiem, *Protective Variants* - z mniejszym. *Variants without risk allele* - warianty wykrywane w badaniu GWAS, ale nie stwierdzone u

¹ Page, M.L., Vance, E.L., Cloward, M.E. *et al.* The Polygenic Risk Score Knowledge Base offers a centralized online repository for calculating and contextualizing polygenic risk scores. *Commun Biol* **5**, 899 (2022). <https://doi.org/10.1038/s42003-022-03795-x>

² <https://prs.byu.edu/>

³ <https://www.ebi.ac.uk/gwas/>

badanego. *Variants in High LD* - warianty blisko sprzężone i segregujące razem, dla uniknięcia artefaktów uwzględnia się tylko jeden (najsilniejszy) z takiej grupy.

Czy na podstawie uzyskanych wyników można udzielić badanej osobie jakichś konkretnych wskazówek? Dlaczego wysokie wartości OR lub beta nie zawsze odpowiadają wysokim centylom w populacji (wskazówka: sprawdź w bazie *GWAS Catalog* szczegóły badania)?