Analiza asocjacji

Analiza asocjacji obejmuje szereg złożonych zagadnień statystycznych, z którymi na ćwiczeniach zapoznamy się w ograniczonym zakresie. Przed zastosowaniem we własnych projektach polecam rozszerzenie wiedzy, np. przez lekturę artykułu Marees i wsp. (2018)[1] lub innych źródeł.

Podstawowym narzędziem do analizy asocjacji jest program **PLINK**, którego dokumentację można znaleźć na stronie http://zzz.bwh.harvard.edu/plink/ Opisane przykłady są demonstracją podstawowych funkcji programu i wykorzystują uproszczone dane.

Pliki danych. Każda analiza PLINK wymaga conajmniej dwóch tekstowych plików z danymi: rodowodu (z genotypami) oraz mapy *loci* SNP. **Plik rodowodu** ma rozszerzenie .ped, a jego format jest zasadniczo taki sam, jak w pliku przygotowanym do analizy sprzężeń. **Plik mapy** ma rozszerzenie .map, i zawiera cztery kolumny: chromosom, identyfikator SNP, pozycja na mapie genetycznej (cM), pozycja w sekwencji (bp). Obecnie pozycji na mapie genetycznej przeważnie się nie podaje, wtedy w kolumnie tej występują wartości 0. Chromosom oznacza się wartością 1-22, X, Y lub 0 (nieustalony). Oprócz tego można używać osobnych plików do specyfikacji dodatkowych fenotypów, podziału populacji na podpopulacje, itp. (patrz dokumentacja).

Uwaga: dobrze jest używać tej samej nazwy dla pliku z rodowodem i mapą (różne tylko rozszerzenia), np. `przyklad.ped` i `przyklad.map`. Wtedy wystarczy programowi podać wspólną nazwę[2]:

```
plink --file przyklad
```
w przeciwnym razie trzeba podać obie nazwy tak:

```
plink --ped przyklad1.ped --map przyklad2.map
```

Format binarny: pliki rodowodów mogą być bardzo duże. Dla zmniejszenia ich wielkości i przyspieszenia pracy programu można dokonać konwersji na format binarny, który nie jest edytowalny ręcznie. Podstawowa komenda konwersji:

```
plink --file przyklad --make-bed --out nowanazwa
```
utworzy zestaw trzech plików o nazwie nowanazwa i rozszerzeniach .bed, .bim i .fam.

Plik .fam zawiera rodowód (bez genotypów SNP) w standardowym formacie, plik .bim zawiera dane o markerach SNP (pozycja na mapie i allele), zaś plik .bed jest głównym plikiem z genotypami, nieczytelnym dla człowieka (format binarny). Początek plików .fam i .bim możemy obejrzeć komendą head.

Pliki binarne wczytujemy opcją **`plink --bfile nowanazwa.`**

Konwersję formatu można połączyć z **filtrowaniem danych**, które jest niezbędnym krokiem w realnych analizach. Filtrowanie obejmuje usunięcie: SNP brakujących w istotnej części próbek, a następnie osobników, u których brak danych w istotnej liczbie SNP. Dodatkowo można sprawdzić, czy deklarowana płeć zgadza się z heterozygotycznością SNP chromosomu X, a także sprawdzić, czy częstości alleli w grupie kontrolnej odbiegają znacząco od równowagi Hardy'ego-Weinberga. Przykładową kontrolę i filtrowanie danych przeprowadzimy w dalszej części ćwiczeń.

[1] Marees i wsp. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int J Methods Psychiatr Res*. 2018;27:e1608

[2] w niektórych dystrybucjach Linuxa program wywołuje się komendą `plink1`

Podstawowa analiza asocjacji cechy dyskretnej (analiza typu *case/control*). W tej analizie plik rodowodu zawiera osoby, u których występuje badany fenotyp (najczęściej choroba) i osoby kontrolne, u których fenotyp nie występuje. Liczebność obu grup powinna być podobna, aby osiągnąć istotne wyniki wskazane są jak najliczniejsze grupy. Dostępne są różne testy statystyczne.

Wykonanie ćwiczenia:

- przekopiuj pliki `simcasecon.ped` i `simcasecon.map` ze wskazanej lokalizacji do katalogu roboczego. Obejrzyj pliki.
- podstawowa analiza wykorzystująca test chi^2:
  `plink --file simcasecon --assoc --out wynik`
  (oczywiście nazwa pliku wynikowego po --out jest dowolna)
- zlokalizuj plik wynikowy z rozszerzeniem .assoc i obejrzyj w edytorze tekstu. Dla większych zbiorów wynikowych konieczne jest zaimportowanie do arkusza kalkulacyjnego albo programu statystycznego (np. R). W pliku wynikowym A1 oznacza allel rzadki, a A2 allel częsty. F_A i F_U to odpowiednio częstość allelu rzadkiego u chorych i w kontrolach. Podana jest też wartość *P* i OR (*odds ratio*).

Problem istotności. W analizach asocjacji nie można automatycznie przyjmować, że wartość $P<0{,}05$ oznacza istotny wynik! Występuje tu bowiem problem znany w statystyce jako **problem porównań wielokrotnych**, na który trzeba przyjąć poprawkę. W analizach całogenomowych w literaturze przyjęło się stosować próg istotności na poziomie $\boldsymbol{P < 5 \cdot 10^{-8}}$ dla populacji europejskiej (w bardziej różnorodnej populacji afrykańskiej próg obniża się do $P < 1 \cdot 10^{-8}$). Można też oszacować poprawkę Bonferroniego, gdzie próg istotności będzie wynosił 0,05/liczba SNP. Można wreszcie obliczyć różne poprawki w PLINK dodając opcję --adjust:

`plink --file simcasecon --assoc --adjust --out wynik`

Z licznych wyliczonych wartości *P* z poprawkami warto zwrócić uwagę na poprawkę Bonferroniego (BONF) i na kontrolę wyników fałszywie pozytywnych Benjaminiego–Hochberga (FDR_BH). Poprawka Bonferroniego często jest zbyt restrykcyjna (nie uwzględnia sprzężenia blisko położonych SNP i traktuje je jako niezależne porównania).

Całogenomowa analiza cechy ilościowej i prezentacja graficzna. W tym ćwiczeniu analizowany będzie większy zbiór danych z 22 autosomów (z wyników programu HapMap). Fenotypem jest tu cecha ilościowa opisywana wartością liczbową.

Wykonanie ćwiczenia

- przekopiuj pliki `quant.ped` i `quant.map` ze wskazanej lokalizacji do katalogu roboczego. Dokonaj konwersji na format binarny:
  `plink --file quant --make-bed --out quantb`
  Obejrzyj wygenerowany plik z rozszerzeniem .fam, zwróć też uwagę na statystyki prób i SNP podane przez program (są za każdym razem zapisywane do pliku .log)
- przeprowadź podstawową analizę asocjacji na przekonwertowanych plikach (oczywiście nazwa pliku wynikowego po --out jest dowolna):
  `plink --bfile quantb --assoc --out wynik`
  Zwróć uwagę na rozmiar pliku wynikowego! Zauważ też, że ma rozszerzenie qassoc, a program automatycznie wykrył, że prowadzona jest analiza ilościowa. W tej sytuacji stosowana jest statystyka Walda. Jeżeli chcesz podejrzeć kilka pierwszych linijek pliku, użyj komendy `head wynik.qassoc`
- przeglądanie pliku tekstowego tej wielkości (czy nawet jego analiza w arkuszu) jest oczywiście niepraktyczne. Standardowym sposobem wizualizacji wyników analizy na taką skalę jest tzw. ***Manhattan plot.*** Dogodnym

narzędziem jest pakiet qqman w programie R[3]. Aby uzyskać wykres uruchom środowisko R i wprowadź następujące komendy:

```
library("qqman")
wyniki_as <- read.table("wynik.qassoc", head=TRUE)
manhattan(wyniki_as)
```

Wartości na osi Y to -$\log_{10}(P)$. Na których chromosomach lokalizują się punkty (SNP) o najwyższej istotności asocjacji? Można uzyskać przybliżony widok tych chromosomów za pomocą komendy (przykładowo dla chr. 1, podstawiamy właściwy chromosom).

```
manhattan(subset(wyniki_as, CHR == 1))
```

Zobacz, w którym obszarze chromosomu znajdują się interesujące SNP. Można przybliżyć ten obszar komendą (manipulując granicami w opcji xlim możemy przybliżać i oddalać widok na wykresie):

```
manhattan(subset(wyniki_as, CHR == 1), xlim=c(1.0e07,1.1e07))
```

Na wykresie można zaznaczyć nazwy SNP, które spełniają określone kryterium, np, $p < 5 \cdot 10^{-8}$

```
manhattan(wyniki_as, annotatePval = 5e-8, annotateTop = FALSE)
```

Opcje annotacji SNP i przybliżania można oczywiście łączyć, np.

```
manhattan(subset(wyniki_as, CHR == 1),xlim=c(1.11e08,1.12e08), annotatePval = 5e-8, annotateTop = FALSE)
```

Wiele innych opcji (np. kolorowania wykresów) można znaleźć w dokumentacji pakietu[4].

Wykresy można zapisywać do pliku pdf lub jpg. W tym celu **przed** wydaniem komendy tworzącej wykres (`manhattan`) trzeba wpisać (nazwę pliku oczywiście podajemy dowolną)

```
pdf("test2.pdf")
```

a po wydaniu komendy tworzącej wykres wpisujemy

```
dev.off()
```

zamiast pdf możemy użyć jpg (gorsza rozdzielczość, ale plik szybciej otwierany, zwłaszcza przy złożonych wykresach).

- powtórz analizę dodając opcję --adjust. Obejrzyj pierwsze 30 linijek otrzymanego pliku .adjusted (komendą `head -n 30 plik.qassoc.adjusted`).

[3] Instalujemy go w R np. komendą `install.packages("qqman")` na koncie administratora. Wymaga R w wersji 3.5 lub wyższej.

[4] https://cran.r-project.org/web/packages/qqman/vignettes/qqman.html

Dodatek: przydatne komendy Linux/Unix

Przemieszczanie się po katalogach: `cd nazwakatalogu` . Można od razu wchodzić wiele poziomów głębiej, np. `cd katalog1/katalog2` Komenda `cd ..` przejdzie jeden katalog wyżej.

`ls` - lista plików w katalogu, * może zastępować dowolny ciąg znaków, np.

`ls *.bed` pokaże wszystkie pliki z rozszerzeniem .bed, a `ls HapMap_11*` pokaże pliki zaczynające się od HapMap11 Opcja -l (np. `ls -l *.txt`) wyświetli więcej informacji.

`cat` wyświetla plik tekstowy, np. `cat inversion.txt`

Jeżeli plik jest duży, to można wyświetlać z podziałem na strony komendą `less`, np. `less covar_mds.txt`. Spacja daje kolejną stronę, klawisz q wychodzi z przeglądania.

`head` wyświetla początek pliku, a `tail` jego koniec. Można podać, ile linijek ma wyświetlić, np. `head -n 30`

`grep` wyszukuje ciąg znaków w pliku, np. `grep PROBLEM plink.sexcheck`

Umieszczenie po komendzie `>plik` spowoduje, że wynik zamiast na ekran trafi do pliku. Plik zostanie za każdym razem stworzony od nowa, jeżeli chcemy dopisać bez usuwania, musimy zastosować `>>plik`

Usuwanie plików: komenda `rm`. Uwaga: usuwa bez pytania! Np. `rm *.assoc` usunie wszystkie pliki z rozszerzeniem .assoc

`history` przywołuje historię ostatnio wydawanych komend. `!numer` ponownie wywołuje komendę z historii. Możemy znaleźć w historii konkretne komendy tak:

`history|grep plink` znajdzie w historii wszystkie komendy, w których wystąpiło słowo plink.

Szybkie kopiowanie: zaznaczyć myszką tekst i nacisnąć środkowy przycisk (kółko).

Tabulator rozwija nazwę w komendzie.

Strzałki pozwalają poruszać się w historii wydawanych wcześniej komend. Działa też w R, ale tylko do wyjścia z sesji.