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

Ćwiczenie: kontrola i filtrowanie danych i analiza asocjacji cechy dyskretnej). W tym ćwiczeniu poznamy najważniejsze metody kontroli jakości danych całogenomowych, która jest niezbędna w realnych analizach. Jest to nieco uproszczona wersja kursu omówionego w cytowanym artykule Marees i wsp. (2018)[5] z wykorzystaniem towarzyszących mu plików pochodzących z projektu HapMap. Wykorzystamy programy PLINK, R, oraz narzędzia Linuxa takie, jak awk i grep. Na początek stwórz katalog do analizy i skopiuj do niego wskazane przez prowadzących pliki .bed, .bim i .fam. Nadaj tym plikom nazwę, która będzie łatwa do wpisywania (w dalszej części opisu przyjęto nazwę HapMap, ale może być oczywiście dowolna inna, taka sama dla .bed, .bim. i .fam). Skopiuj też pliki o nazwie inversion.txt i heterozygosity_outliers_list.R. Pliki z kolejnych etapów pracy będą wykorzystywane w następnych, uważaj aby niczego nie nadpisać, podając odpowiednią nazwę pliku wynikowego po --out.

1. **Analiza brakujących danych.** W tym kroku usuwamy z danych SNP, dla których u istotnej części osób brak danych, a następnie (kolejność jest istotna) usuwamy osoby, u których brak danych dla istotnej części SNP. Zacznijmy od sprawdzenia, jak wyglądają nasze dane:

   ```
   plink --bfile HapMap --missing
   ```
   W pliku z rozszerzeniem .imiss mamy dla każdej osoby informację o liczbie i odsetku brakujących SNP, zaś w pliku .lmiss dla każdego SNlP informację o tym, u ilu osób nie został oznaczony. Możemy informacje te przedstawić w formie histogramu, w programie R za pomocą komend:
   ```
   indmiss<-read.table(file="plink.imiss", header=TRUE)
   hist(indmiss[,6])
   ```
   (w kolumnie 6 mamy informację o tym, jaki odsetek SNP brakuje u danej osoby), a następnie
   ```
   snpmiss<-read.table(file="plink.lmiss", header=TRUE)
   hist(snpmiss[,5])
   ```
   w kolumnie 5 jest informacja o tym, u jakiego odsetka osób brakuje danego SNP.
   Dla orientacji warto zacząć od stosunkowo łagodnego progu: odrzucamy SNP nieoznaczone u >20% osób.
   ```
   plink --bfile HapMap --geno 0.2 --make-bed --out HapMap_2
   ```
   Oglądamy wyświetloną informację - czy jakieś dane zostały odrzucone? Teraz przeprowadzimy właściwe filtrowanie. Najpierw odrzucamy SNP, których brakuje u więcej, niż 2% osób:
   ```
   plink --bfile HapMap --geno 0.02 --make-bed --out HapMap_2
   ```
   a wreszcie z pozostałego pliku eliminujemy te osoby, u których brakuje więcej, niż 2% SNP:
   ```
   plink --bfile HapMap_2 --mind 0.02 --make-bed --out HapMap_3
   ```
   Zaobserwuj, ile danych odrzuca każda z tych komend, i ile pozostaje. Dlaczego filtrowanie brakujących danych przeprowadza się w tej kolejności?

2. **Kontrola przypisania płci.** Płeć zadeklarowana w rodowodzie nie zawsze odpowiada rzeczywistości. Prostą kontrolą jest sprawdzenie heterozygotyczności markerów znajdujących się na chromosomie X:
   ```
   plink --bfile HapMap_3 --check-sex
   ```
   Obejrzyj wynikowy plik z rozszerzeniem .sexcheck i spróbuj zinterpretować wynik. Dla ułatwienia, można wyszukać wiersze pliku, w których występuje słowo PROBLEM za pomocą komendy
   ```
   grep PROBLEM plink.sexcheck
   ```
   Wynik możemy zapisać w nowym pliku w ten sposób:
   ```
   grep PROBLEM plink.sexcheck > problem.txt
   ```
   Teraz mamy dwie możliwości (wybierz tylko jedną z nich). Albo osoby z problematycznym przypisaniem płci usuwamy z danych (ta komenda usunie z pliku osoby wskazane w pliku problem.txt):
   ```
   plink --bfile HapMap_3 --remove problem.txt --make-bed --out HapMap_4
   ```

[5] 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

albo ustalimy płeć na podstawie SNP chromosomu X, nadpisując dane z rodowodu

```
plink --bfile HapMap_3 --impute-sex --make-bed --out HapMap_4
```

3. **Usunięcie chromosomu X.** Po wykorzystaniu chromosomu X do kontroli płci, do dalszej analizy należy pozostawić wyłącznie autosomy (asocjacje na chromosomie X bada się w osobnej analizie). Chromosom zapisany jest w pierwszej kolumnie pliku .bim. Za pomocą komendy awk wyszukamy w tym pliku SNP na chromosomach od 1 do 22 i zapiszemy w osobnym pliku ich identyfikatory z kolumny 2:

   ```
   awk '{ if ($1 >= 1 && $1 <= 22) print $2 }' HapMap_4.bim>autosomal.txt
   ```
   Teraz wykorzystamy stworzony plik do selekcji danych:

   ```
   plink --bfile HapMap_4  --extract autosomal.txt --make-bed --out HapMap_5
   ```
   Komenda --extract pozostawia markery wskazane w pliku. Jej przeciwieństwem jest komenda --exclude.

4. **Kontrola częstości allelu rzadkiego**. W typowych analizach SNP stosowane są markery polimorficzne. W związku z tym nie spodziewamy się bardzo rzadkich alleli markera. Allele bardzo rzadkie nie dają też istotnych korelacji z częstymi cechami fenotypowymi (zwróć uwagę na to, ile w próbie jest chorych, a ile kontroli). W związku z tym należy wykryć i usunąć SNP, dla których częstość allelu rzadkiego jest poniżej pewnego progu. Wartość tego progu zależy od wielkości próby - im więcej osób w analizie, tym rzadsze allele możemy pozostawić. Przy próbach < 10 000 osób usuwamy SNP o częstości allelu rzadszego (MAF) mniejszej od 5%. Uwaga: istnieją analizy, gdzie w danych uzyskanych z sekwencjonowania całogenomowego poszukuje się właśnie alleli rzadkich. W takiej sytuacji oczywiście ich się nie usuwa. Zobaczmy, jak wyglądają częstości alleli w naszych danych:

   ```
   plink --bfile HapMap_5 --freq --out MAF_check
   ```
   Wyniki zapisane są w pliku MAF_check.frq, który jest zbyt duży, żeby go analizować ręcznie. Można podejrzeć jego format komendą head

   ```
   head MAF_check.frq
   ```
   Jak widać, zdarzają się SNP, w których częstość allelu rzadkiego (kolumna 5) jest bardzo mała. Możemy zobaczyć to na histogramie rozkładu częstości allelu rzadkiego w R.

   ```
   maf_freq <- read.table("MAF_check.frq", header =TRUE)
   hist (maf_freq[,5])
   ```
   Usuńmy wszystkie SNP, dla których MAF (częstość allelu rzadkiego) wynosi mniej niż 5%:

   ```
   plink --bfile HapMap_5 --maf 0.05 --make-bed --out HapMap_6
   ```
   Ile SNP zostało usuniętych, a ile zostało? Powtórz powyższą analizę histogramu częstości MAF dla uzyskanego pliku.

5. **Kontrola równowagi Hardy'ego-Weinberga**. Polimorficzne allele SNP powinny w populacji być niezbyt odległe od równowagi opisywanej prawem Hardy'ego-Weinberga. Należy to sprawdzić za pomocą komendy

   ```
   plink --bfile HapMap_6 --hardy
   ```
   Uzyskany plik plink.hwe jest bardzo duży, obejrzyj jego początek komendą head i zastanów się, jakie są tam informacje. Dla każdego SNP równowaga testowana jest trzykrotnie: w grupie kontrolnej, w grupie chorych, i we wszystkich próbach (w kolumnie 3 odpowiednio UNAFF, AFF, ALL). Wartość P w ostatniej (9) kolumnie mówi nam o istotności odchylenia od równowagi H-W. Zobaczmy histogram tej wartości w programie R:

   ```
   hwe<-read.table (file="plink.hwe", header=TRUE)
   hist(hwe[,9])
   ```
   Aby zobaczyć SNP, dla których P wynosi poniżej $1 \cdot 10^{-5}$ użyjmy programu awk:

   ```
   awk '{ if ($9 <1e-5) print $0 }' plink.hwe
   ```
   Powinno się usuwać SNP dla których P wynosi poniżej $1 \cdot 10^{-6}$ w grupie kontrolnej i poniżej $1 \cdot 10^{-10}$ w grupie chorych. Posłużą do tego komendy:

   ```
   plink --bfile HapMap_6 --hwe 1e-6 --make-bed --out HapMap_7
   ```

```
plink --bfile HapMap_7 --hwe 1e-10 --hwe-all --make-bed --out HapMap_8
```

Czy powyższe komendy usunęły coś z naszych danych? Skonfrontuj odpowiedź z wynikami uzyskanymi wcześniej komendą awk.

6. **Analiza nierównowagi sprzężeń.** Nierównowaga sprzężeń, czyli zależne od siebie dziedziczenie konkretnych alleli różnych SNP może zakłócić niektóre dalsze analizy. Nie zawsze jest konieczne usuwanie takich SNP, ale warto mieć plik zawierający wyłącznie SNP niezależne od siebie. W tym celu wyłączymy obszary genomu, w których wiadomo o występowaniu nierównowagi sprzężeń. Są to obszary niedawnych inwersji i grupy loci podlegających wspólnej selekcji ewolucyjnej (np. obszar MHC). Ich pozycje są podane w pliku inversion.txt. Poniższa komenda wyłączy te obszary, a w reszcie genomu zidentyfikuje obszary, gdzie SNP nie są ze sobą skorelowane.

   ```
   plink --bfile HapMap_8 --exclude inversion.txt --range --indep-pairwise 50 5 0.2 --out indepSNP
   ```

   Plik indepSNP.prune.in będzie zawierał SNP niezależne od siebie. Przyda się w kolejnych etapach.

7. **Analiza heterozygotyczności**. Dla każdej osoby w danych możemy obliczyć współczynnik heterozygotyczności. Osoby o nietypowo niskiej heterozygotyczności w stosunku do reszty populacji są zapewne rezultatem kojarzenia wsobnego, zaś osoby o nietypowo wysokiej heterozygotyczności mogą być artefaktem analizy. Heterozygotyczność testujemy wykorzystując wyłącznie SNP niezależne (nie wykazujące nierównowagi sprzężeń), które zidentyfikowaliśmy w poprzednim kroku.

   ```
   plink --bfile HapMap_8 --extract indepSNP.prune.in --het --out R_check
   ```

   Obejrzyj wygenerowany plik R_check.het. Dla każdej osoby podawana jest tam liczba genotypów SNP homozygotycznych (O(HOM) w kolumnie 3) i całkowita liczba oznaczonych genotypów SNP (N(NM) w kolumnie 5). Heterozygotyczność obliczamy ze wzoru (O(HOM)-N(NM))/N(NM). Aby wygenerować odpowiedni rozkład użyjemy R:

   ```
   het <- read.table("R_check.het", head=TRUE)
   het$HET_RATE = (het$"N.NM." - het$"O.HOM.")/het$"N.NM."
   hist(het$HET_RATE)
   ```

   Czy w histogramie widać osoby o skrajnych wartościach heterozygotyczności? Zanotuj przybliżoną minimalną i maksymalną wartość tego parametru. Zaleca się usunięcie osób, których heterozygotyczność odbiega od średniej o trzykrotność SD. Posłużymy się gotowym już skryptem w R (warto zajrzeć do jego kodu):

   ```
   Rscript --no-save heterozygosity_outliers_list.R
   ```

   Skrypt stworzył plik o nazwie fail-het-qc.txt. Otwórz go edytorem tekstu. Ile osób znalazło się na liście i której części rozkładu z wcześniejszego histogramu odpowiadają? Co w związku z tym możesz o nich powiedzieć?

   Aby usunąć te osoby z danych należy najpierw zmodyfikować plik fail-het-qc.txt. W edytorze tekstu usuń wiersz nagłówka i wszystkie cudzysłowy[6], po czym zapisz plik. Następnie wykorzystaj go do odfiltrowania danych:

   ```
   plink --bfile HapMap_8 --remove fail-het-qc.txt --make-bed --out HapMap_9
   ```

8. **Analiza pokrewieństwa.** W zbiorze danych możemy mieć osoby spokrewnione i zaznaczone jako należące do tej samej rodziny (w pliku rodowodu), a także osoby o ukrytym pokrewieństwie (krewni w pliku .fam nie zaznaczeni jako należący do tej samej rodziny). W typowych analizach populacyjnych GWAS nie powinniśmy mieć w danych osób spokrewnionych (ich wykorzystanie wymaga odrębnych metod). Sprawdźmy, jak wygląda pokrewieństwo szacowane na podstawie alleli SNP dla wszystkich par osób w

---

[6] Koneserzy linii komend mogą tu też skorzystać z komendy sed:

```
sed 's/"// g' fail-het-qc.txt | awk '{print$1, $2}'> fail-het-qc_2.txt
```

zbiorze danych za pomocą komendy:

```
plink --bfile HapMap_9 --genome --min 0.2 --out pihat_min0.2
```

Komenda ta posługuje się współczynnikiem pi-hat. Jego wartość waha się od 1 dla osób o identycznych genotypach (np. bliźniąt monozygotycznych) do 0. Każdy stopień pokrewieństwa oddzielający osoby zmniejszy pi-hat o ½, czyli krewni 1. stopnia (np. rodzeństwo) będą mieli pi-hat=0.5, krewni 2. stopnia - 0.25, i tak dalej. Powyższa komenda zapisała w pliku pihat_min0.2.genome wszystkie pary o pi-hat większym od 0.2. Obejrzyj ten plik w edytorze tekstu. Kolumna RT (5) zawiera opis relacji wynikający z rodowodu. PO to relacja rodzic-dziecko (*parent-offspring*), UN to pary, dla których nie ma informacji o pokrewieństwie (*unrelated*).

Na początek możemy przefiltrować dane zostawiając wyłącznie założycieli (*founders*, czyli osoby niemające rodziców w rodowodzie):

```
plink --bfile HapMap_9 --filter-founders --make-bed --out HapMap_10
```

Powtórzmy teraz analizę pi-hat na pliku zawierającym samych założycieli:

```
plink --bfile HapMap_10 --genome --min 0.2 --out pihat_min0.2
```

I znowu obejrzyjmy plik wynikowy pihat_min0.2.genome. Ile par zostało, jaka jest wartość pi-hat i jakiemu pokrewieństwu może odpowiadać. Znajdź identyfikatory tych osób. Jedną z pary trzeba usunąć z danych. Najlepiej usunąć tę, która ma mniej brakujących oznaczonych SNP. Informację tę znajdziemy w pliku .imiss, który uzyskaliśmy już w etapie 1. Obejrzyjmy jeszcze raz początek tego pliku:

```
head plink.imiss
```

Kolumna N_MISS (4) to liczba brakujących SNP. Odnajdźmy wiersze odpowiadające osobom zidentyfikowanym w pliku pihat_min0.2.genome za pomocą komendy grep:

```
grep identyfikator plink.imiss
```

gdzie *identyfikator* to wartość IID1 lub IID2 z pliku .genome. Która osoba ma większą wartość N_MISS? Stwórz plik tekstowy, do którego skopiujesz identyfikator rodziny i osoby (jako dwie kolumny), zapisz go pod nazwą np. usun.txt. Skopiować możesz z wyniku grep w terminalu. Teraz usuń tę osobę:

```
plink --bfile HapMap_10 --remove usun.txt --make-bed --out HapMap_11
```

To już koniec filtrowania. Ile zostało nam osób w próbie?

9. **Właściwa analiza.** W plikach HapMap_11 (.bed,.bim i .fam) mamy dane, które już mogą nam posłużyć do analizy asocjacji. Zacznijmy od prostej analizy asocjacji wykorzystującej test chi^2, jak na stronie 2 (komendę już sam wpisz, pamiętaj że analizujesz pliki binarne, czyli używasz opcji --bfile a nie --file!). Uzyskany plik .assoc jest bardzo duży, i nie obejrzysz go w edytorze tekstu. Przygotuj na nim wykres typu Manhattan Plot w programie R, tak jak opisano to na stronie 3, najlepiej zapisując też jpeg. Poeksperymentuj z wyborem konkretnych chromosomów, przybliżaniem i zaznaczaniem wybranych SNP (zastosuj próg $p< 1 \cdot 10^{-5}$) tak, jak na stronie 3.

   Następnie, podobnie jak na stronie 2 przeprowadź analizę z poprawkami (--adjust) i obejrzyj pierwsze 25 wyników (zastosuj komendę head).

10. **Analiza stratyfikacji (opcjonalnie)**. Częstym źródłem błędów w analizach korelacyjnych jest istnienie ukrytej zmiennej dzielącej badaną populację i wpływającej na oznaczaną cechę. Zapoznaj się z pojęciem paradoksu Simpsona! Częstym błędem jest branie do analizy populacji niejednorodnej etnicznie. Z badaniem etniczności w genetyce człowieka zapoznamy się na kolejnych ćwiczeniach, tu przyjmijmy, że wszyscy w naszym zbiorze pochodzą z populacji europejskiej. Subtelniejsze podziały wewnątrz badanej populacji można wykryć stosując metody takie, jak analiza głównych składowych (PCA) lub skalowanie wielowymiarowe (MDS). Uwzględnijmy MDS w naszej analizie. Najpierw przeprowadźmy analizę korelacji dla SNP niezależnych, podobną do tej z punktu 7:

    ```
    plink --bfile HapMap_11 --genome --extract indepSNP.prune.in --out HapMap_11
    ```

    Uzyskany plik .genome wykorzystamy w analizie MDS dla 10 wymiarów (zalecana wartość):

```
plink --bfile HapMap_11 --read-genome HapMap_11.genome --cluster --mds-plot 10 --out HapMap_mds
```

Uzyskany plik .mds musimy przekształcić usuwając zbędną kolumnę 3:

```
awk '{print$1, $2, $4, $5, $6, $7, $8, $9, $10, $11, $12, $13}' HapMap_mds.mds>covar_mds.txt
```

Uzyskane kowarianty zastosujemy w analizie korelacji z regresją logistyczną:

```
plink --bfile HapMap_11 -covar covar_mds.txt --logistic --hide-covar --out wyniki_mds_covar
```

Wyniki będą w pliku z rozszerzeniem .logistic

Zanim będzie można na jego podstawie przygotować wykres, musimy usunąć z niego wiersze z wartościami NA komendą:

```
awk '!/'NA'/' wyniki_mds_covar.assoc.logistic> wyniki_2.logistic
```

Uzyskany w ten sposób plik możemy już zwizualizować w R tak, jak wcześniej. Czy wartości P się zmieniły i jak?

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.