# Podejścia Systemowe w Badaniach Biomedycznych
# zajęcia laboratoryjne 3

## Badania asocjacyjne całego genomu

## 1 Badania asocjacyjne całego genomu (GWAS) i kontrola jakości

Badania asocjacyjne całego genomu opierają się na poszukiwaniu związków pomiędzy wariantami genetycznymi a cechami/fenotypami (będącymi obiektem zainteresowania) poprzez ogólnogenomowe genotypowanie. Innymi słowy są to badania oparte na poszukiwaniu korelacji pomiędzy genotypem a fenotypem. Jeżeli jest istotna korelacja pomiędzy genotypem w konkretnym miejscu w genomie a fenotypem badanych organizmów, można wnioskować o istnieniu związku pomiędzy allelem a stanem badanej cechy. Istotne zagadnienia badań asocjacyjnych całego genomu w tym korzyści i ograniczenia GWAS zostały omówione na wykładzie.

## 2 Przygotowanie środowiska pracy

Pobierz stabilną wersję programu *PLINK 1.9*, dostępny na stronie https://www.cog-genomics.org/plink/1.9/. Wypakuj paczkę i następnie przez terminal/wiersz poleceń wykonaj pierwsze testowe uruchomienie (pamiętaj o ustawieniu odpowiednich ścieżek):

```
plink --file toy --freq --out toy_analysis
```

Na podstawie testowych plików wygenerowano raport o częstotliwościach alleli (aby zapoznać się z wykorzystywanymi flagami odwiedź stronę https://www.cog-genomics.org/plink/1.9/filter). Istotne informacje dotyczące narzędzia *PLINK* są także dostępne na stronie http://zzz.bwh.harvard.edu/plink/.

## 3 Analiza GWAS

Do analizy GWAS wykorzystaj dane dostępne na stronie:
https://www.cs.put.poznan.pl/kgutowska/PSwBB/dane/GWAS_lab3.zip, następnie wykonaj poszczególne etapy analizy. Wykorzystane dane powiązane są z badaniami opisanymi w artykule [1].

Wczytywanie odpowiednich danych:

```
plink --covar samples.covar --file adgwas --out input_test --recode
```

Pliki wejściowe *adgwas* zawierają dwa rozszerzenia .ped i .map. Format .ped przechowuje informacje o genotypach, pochodzeniu, o cechach, podczas gdy format .map zawiera mapę markerów, która zawiera informacje o pozycji każdego SNP na chromosomach. Plik o rozszerzeniu .covar zawiera dodatkowe informacje o grupie, nazwach odpowiadających poszczególnym identyfikatorom, ale także wiele innych zmiennych towarzyszących (diagnoza, wiek, lokalizacja itd.).

Test równowagi Hardy'ego-Weinberga:

```
plink --file input_test --hwe .05 --out hwe_rmvd --recode
```

Flaga *hwe* odfiltrowuje wszystkie warianty dla których test równowagi Hardy'ego-Weinberga ma p-value poniżej podanego progu. Należy mieć na uwadze, że rzeczywiste skojarzenia cecha-SNP odbiegają nieco od równowagi Hardy'ego-Weinberga. W związku z tym należy rozważnie wybrać próg filtrowania, aby nie odfiltrować zbyt wielu wariantów.

Częstość allelu rzadszego (ang. minor allele frequency (MAF)):

```
plink --file hwe_rmvd --maf 0.01 --out maf_rmvd --recode
```

Flaga *maf* odfiltrowuje wszystkie warianty z częstotliwością allelu rzadszego poniżej podanego progu (domyślnie 0.01). Można wykorzystać inne flagi, aby określić górną granicę miary MAF lub zarówno górną jak i dolną jednocześnie.

Wskaźniki brakującego genotypu (ang. missing genotype rates):

```
plink --file maf_rmvd --geno .10 --out geno_rmvd --recode
```

Flaga *geno* odfiltrowuje wszystkie warianty na podstawie wskaźnika brakującego genotypu (domyślnie 0.1). Wyklucza warianty, które mają wysoki współczynnik niepowodzeń genotypowania. Natomiast flaga *mind* działa podobnie jak flaga *geno* tylko odfiltrowuje próbki o wysokim odsetku słabych genotypów.

```
plink --file geno_rmvd --mind .10 --out mind_rmvd --recode
```

Należy zapamiętać, że flaga *geno* dotyczy wariantów (ang. per-variant), a flaga *mind* dotyczy próbek (ang. per-sample).

Estymacja IBS/IBD (ang. identity by state/identity by descent):

```
plink --file mind_rmvd --genome --out estimation_results --recode
```

Flaga *genome* generuje plik z raportem w oparciu o obliczenia IBS/IBD, czyli oszacowanie identyczności przez stan lub pochodzenie dla wszystkich par obserwacji. Oszacowanie oparte na IBS wykorzystuje się do wykrywania par obserwacji, które różnią się od siebie bardziej niż można by się spodziewać w losowej próbie. Natomiast oszacowanie par oparte na IBD ma w celu znalezienie par obserwacji, które są do siebie bardziej podobne niż można oczekiwać w przypadku losowej próby.

Redukcja wymiarowości metodą PCA (analiza składowych głównych):

```
plink --file estimation_results --out pca --pca --recode
```

Flaga *pca* domyślnie zwraca 20 składowych głównych. Wektory własne (ang. eigenvec) znajdują się w oddzielnym pliku i są posortowane od najwyższych wartości własnych.

Testy asocjacyjne:

```
plink --adjust --ci .95 --covar pca.eigenvec --covar-number 1 --file pca --logistic
--out results_for_PCA1 --recode
```

Flaga *adjust* generuje plik z raportem podsumowującym testy asocjacyjne, który zawiera kilka podstawowych wielokrotnych poprawek testowych dla surowych wartości p-value. Flaga *ci* pozwala ustawić przedział ufności, np. wartość .95 oznacza, że ustawiono 95% przedział ufności. Flaga *logistic* lub flaga *linear* wykonuje regresję (modele regresji). Jest to regresja logistyczna (wykorzystywana w przypadku asocjacji cecha-wariant (ang. disease traits)) lub liniowa (wykorzystywana w przypadku cech ilościowych (ang. quantitive traits)). Przy okazji wykonywania regresji można wykorzystać pliki zmiennych towarzyszących (ang. covariates), które mogą zostać wykorzystane w modelu regresji (flaga *covar* używana do wybrania odpowiedniego pliku i flaga *covar-number* używana do wybrania odpowiedniego podzbioru).

W poniższym przykładzie (w odróżnieniu od instrukcji powyżej) wyselekcjonowano dwie składowe główne:

```
plink --adjust --ci .95 --covar pca.eigenvec --covar-number 1-2 --file pca --logistic
--out results_for_PCA1-2 --recode
```

# Zadania

Sprawozdanie i kod R prześlij na maila (Kaja.Gutowska@cs.put.poznan.pl). Proszę nadać tytuł maila zgodnie z opisem: PSwBB_Imię_Nazwisko_lab3.

## Zad. 1

Wykonaj polecenia przygotowane w protokole w sekcji *Analiza GWAS* i na ich podstawie przygotuj wstęp do sprawozdania, napisz jakie mamy dane wejściowe i je krótko scharakteryzuj (np. napisz ile mamy wariantów przed filtrowaniem danych i po zastosowaniu różnych filtrów). Zapoznaj się z artykułem [1] i kontekstem biologicznym kryjącym się za przeprowadzoną powyżej analizą, o czym traktują te badania?

## Zad. 2

Po wykonaniu redukcji wymiarowości w *PLINK* wykorzystaj pliki wynikowe, wybrane z nich wczytaj do środowiska R i zwizualizuj. Następnie na ich podstawie oceń ile składowych głównych należy uwzględnić przy wykonywaniu testów asocjacyjnych? Czy należy rozważyć pierwszą składową główną czy może więcej składowych głównych?

## Zad. 3

Na podstawie decyzji podjętej w poprzednim zadaniu, wykonaj odpowiednie testy asocjacyjne tzn. zdecyduj ile składowych głównych należy uwzględnić w testach. Następnie plik wynikowy .logistic wczytaj do środowiska R i znajdź istotne warianty. Posortuj plik wynikowy .logistic na podstawie kolumny $P$, czyli tej zawierającej p-value. Ile wariantów możemy uznać za istotnie powiązanych z cechą (p-value $< 5 \times 10^{-8}$)? Jakie to są warianty? Czy istnieje jakiś filtr/flaga która umożliwia posortowanie wyników w programie *PLINK*?

## Zad 4.

Dla wyselekcjonowanych wariantów (używając identyfikatorów rs):

- Skorzystaj z bazy SNPedia i GWAS catalog, aby scharakteryzować krótko otrzymane warianty. Jakie jest ich znaczenie biologiczne?
- Skorzystaj z bazy gnomAD, aby określić częstotliwość występowania danego wariantu w całej populacji.
- Czy znasz inną bazę danych/narzędzie, która dostarcza dodatkowych informacji o znalezionych wariantach?

# Literatura

[1] Webster, Jennifer A., et al. Genetic control of human brain transcript expression in Alzheimer disease. The American Journal of Human Genetics, 84.4, 2009, 445–458.