

---

## Wstęp do analizy ewolucyjnej populacji człowieka w oparciu o dane genomowe

Dostępność danych o zmienności całego genomu dla reprezentatywnych próbek różnorodnych populacji człowieka, w tym genomów kopalnych, umożliwia odtworzenie złożonych historii wędrówek i mieszania się różnych grup ludzi po opuszczeniu Afryki. Tradycyjne techniki filogenetyczne, odtwarzające historię pojedynczych sekwencji, są tu niewystarczające. Głównym wyzwaniem jest odtworzenie historii na podstawie zmienności sekwencji autosomalnych, które w każdym pokoleniu ulegają tasowaniu przez rekombinację. Wśród wielu różnych metod do najczęściej stosowanych należą analiza struktury/admiksji oraz metody oparte na grupowaniu na podstawie podobieństwa: analiza głównych składowych (PCA) i skalowanie wielowymiarowe (MDS).

**Pliki danych.** Niektóre programy, np. ADMIXTURE wykorzystują format binarny stosowany przez znany już program PLINK (plik rodowodu .bed, plik mapy .map i plik opisujący osobniki .fam). Często zachodzi potrzeba wybrania z dużego pliku danych podzbioru obejmującego wybrane populacje. Przykładową procedurę opisano w dodatku na końcu tego skryptu. Innymi często spotykanymi formatami są pliki ANCESTRYMAP/EIGENSTRAT (pliki .geno, .snp i .ind). Do zamiany formatów służy program CONVERTF<sup>1</sup>. Dane z analiz SNP/NGS są też często udostępniane w formacie VCF, który może być wczytany i zmieniony na inne przez PLINK (w wersji 1.9 lub nowszej).

Liczba niezbędnych SNP zależy od zróżnicowania analizowanych populacji - im bliższe sobie są populacje, tym więcej danych potrzeba. Dla analizy na skalę międzykontynentalną wystarczy ok. 10 000 SNP, do analizy wewnątrz kontynentu trzeba mieć co najmniej 100 000 SNP.

Zalecane jest usunięcie ze zbioru danych SNP pozostających w korelacji (nierównowadze sprzężeń, *LD-pruning*). Procedurę taką opisano w skrypcie omawiającym analizę asocjacji.

### 1. Podstawowa analiza admiksji wybranych populacji z Europy i części Azji

W ćwiczeniu zastosujemy metodę analizy admiksji za pomocą programu ADMIXTURE<sup>2</sup>, stosującego podejście maksymalizacji wiarygodności dla ustalenia pochodzenia każdego osobnika w zbiorze danych.

Analiza admiksji polega na tym, że dla zadanej liczby  $K$  populacji wyjściowych (przodków) program oblicza wkład każdej z tych populacji w genotyp każdego z analizowanych osobników. Analizę najlepiej przeprowadzić dla wielu różnych wartości  $K$  (w analizach obejmujących dane populacji z całej Ziemi stosowano wartości  $K$  od 1 do 20). W celu wskazania najwłaściwszej dla danego zbioru danych wartości  $K$  stosuje się walidację krzyżową (*cross validation*) i wybiera  $K$ , dla którego błąd jest najmniejszy. Zastosowanie walidacji krzyżowej (opcja --cv programu) znacznie wydłuża czas obliczeń, dlatego na ćwiczeniach zastosujemy wartość  $K$  ustaloną wcześniej. Realne analizy tego typu wymagają dużej mocy obliczeniowej - dla nietrywialnych zbiorów danych prowadzi się je zdalnie na silnych stacjach roboczych albo serwerach obliczeniowych. Aby uniknąć przerwania obliczeń przy przerwaniu zdalnego połączenia warto poznać i stosować polecenie `screen` lub `tmux`.

Wyniki obrazuje się za pomocą wykresów słupkowych (*bar graph*), na których każdy słupek odpowiada jednemu osobnikowi, a jego kolory - udziałowi populacji wyjściowych. Do wizualizacji wyników wykorzystuje się zwykle pakiet R.

W przykładowej analizie spróbujemy odpowiedzieć na pytanie, ile populacji źródłowych miało wkład w genomy dzisiejszych mieszkańców Europy. W ich identyfikacji pomogą genomy ze szczątków kopalnych znalezionych w

---

<sup>1</sup> <https://github.com/DReichLab/EIG/tree/master/CONVERTF>

<sup>2</sup> Alexander i wsp. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19:1655–1664. <http://software.genetics.ucla.edu/admixture/index.html>

Luksemburgu (Loschbour), Hiszpanii (LaBrana), Austrii (Iceman) a także znacznie starszego szkieletu z Syberii (MA1). Dane pochodzą ze znacznie większej analizy<sup>3</sup>, obejmującej ponad 2000 próbek.

Dane zawarte są w plikach o nazwie EuPop.

Analizę przeprowadzamy komendą:

```
admixture -j4 EuPop.bed K
```

gdzie  $K$  jest wskazaną przez prowadzącego wartością liczby populacji wyjściowych<sup>4</sup>. Parametr `-j4` oznacza, że należy wykorzystać 4 wątki procesora, dobieramy go tak, aby jak najlepiej wykorzystać możliwości dostępnego sprzętu (w systemie Linux komenda `lscpu` poda liczbę dostępnych rdzeni i wątków).

Wyniki znajdują się w plikach z rozszerzeniem `.P` i `.Q`. W pliku `.P` każdy wiersz odpowiada jednemu allelowi SNP, a wartości w kolumnach częstościom tego allelu w każdej z  $K$  populacji wyjściowych. Najważniejszy jest plik `.Q`, w którym każdy wiersz odpowiada jednemu osobnikowi, a wartości w kolumnach to udziały każdej z  $K$  populacji wyjściowych w jego genotypie. Ten plik będziemy dalej wizualizować.

### Wizualizacja wyników

Podstawowym sposobem wizualizacji wyników takiej analizy jest wykres słupkowy (*barplot*), w którym każdy słupek odpowiada pojedynczej próbce (osobnikowi), a różne kolory oznaczają udział poszczególnych populacji źródłowych zaczerpnięty z pliku `.Q`. Głównym problemem w przedstawianiu wyników jest duża liczba próbek, przez co trudno zmieścić ich nazwy. W ćwiczeniu zastosujemy stosunkowo proste podejście, istnieje wiele pakietów R służących do wizualizacji admiksji w bardziej atrakcyjnej i czytelnej formie (np. z podziałem na populacje)<sup>5</sup>.

W naszych danych nazwa populacji zapisana jest w pierwszej kolumnie pliku `EuAs.fam`. W niektórych zbiorach z danymi informacje te znajdują się w osobnych plikach, wtedy należy odpowiednio zmodyfikować postępowanie.

Kod do wizualizacji zawarty jest w pliku `visualize.R`.

Wykresy sporządzane są dla wyników posortowanych wg. populacji, a także pod względem udziału każdej z  $K$  (tu 3) populacji źródłowych (te wykresy ułatwią identyfikację, który kolor odpowiada której z hipotetycznych populacji źródłowych).

Zadanie dla chętnych: udoskonalić kod z pliku, np. tak aby przygotowywał wykresy dla dowolnej wartości  $K$  podanej jako zmienna.

### Interpretacja wyników

Interpretacja wyników musi być prowadzona bardzo ostrożnie. Szczególnie unikać należy utożsamiania populacji wyjściowych z którąkolwiek z analizowanych populacji współczesnych, chyba że mamy niezależne dane, które to uzasadniają, np. liczne próbki kopalne. Istnieje literatura dyskutująca interpretację takich wyników i możliwe błędy<sup>6</sup>.

Przeanalizuj wyniki w oparciu o informacje z wykładu. Wyszukaj w sieci informację o pochodzeniu tych grup etnicznych, których nazwy nie są Ci znane. W zbiorze danych są cztery próbki kopalne: szkielety z Loschbour i La Braña należały do przedstawicieli łowców-zbieraczy zamieszkujących Europę paleolitu, Iceman to sławny Ötzi, znaleziony w alpejskim lodowcu mieszkaniac Europy sprzed około 5000 lat (epoka miedzi, przełom neolitu i

---

<sup>3</sup> Lazaridis i wsp. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513:409-413

<sup>4</sup> Jeżeli nie znamy wartości  $K$ , dodajemy parametr `--cv` i prowadzimy po kolei analizę dla różnych  $K$  (od 2 do 10-20). Wybieramy  $K$  z najniższym błędem CV. Taka analiza wymaga jednak znacznie więcej czasu, niż mamy na ćwiczeniach. Tu użyjemy  $K = 3$ .

<sup>5</sup> Np. <http://www.royfrancis.com/pophelper/>

<sup>6</sup> Lawson i wsp. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Comm.* 9:3258.

epoki brązu), zaś MA1 to przedstawiciel kultury Mal'ta-Buret' z paleolitu (ok. 24 tys. lat temu) z terenu Syberii. Zwróć uwagę na składową związaną z genomem MA1. Które współczesne grupy etniczne zawierają jej znaczny udział? Skąd wzięła się u mieszkańców Europy i jak wygląda jej udział w różnych częściach kontynentu? Na podstawie genomu z Loschbour ustal, z jakimi pradawnymi populacjami związane są pozostałe składowe, i zastanów się nad ich rozmieszczeniem we współczesnych populacjach.

## 2. Analiza metodą skalowania wielowymiarowego (MDS)

Analizę przeprowadzimy za pomocą PLINK dwuetapowo. Najpierw przygotujemy zbiór z obliczonymi korelacjami SNP (odpowiadającymi podobieństwu genetycznemu wyrażanemu przez IBD):

```
plink --bfile EuPop --genome --out EuPopIBD
```

Następnie przeprowadzimy analizę MDS w dwóch wymiarach (co pozwoli nam na wizualizację na dwuwymiarowym wykresie):

```
plink --bfile EuPop --read-genome EuPopIBD.genome --cluster --mds-plot 2 --out EuPopMDS
```

Uzyskujemy plik EuPopMDS.mds, który zobrazujemy na prostym wykresie w R. Podobnie jak poprzednio, ilość danych wymusza zastosowanie PDF o dużych rozmiarach:

```
mds<-read.table("EuPopMDS.mds", head=T)
pdf(file="mds.pdf", 10, 10)
plot(mds$C1, mds$C2, cex=0.5)
text(mds$C1, mds$C2, mds$FID, cex=0.25, pos=4, offset=0.2)
dev.off()
```

Obejrzyj wynik i przedyskutuj możliwe interpretacje. Jak odległości między populacjami korelują z geografią i z wynikami analizy admiksji? Do bardziej zaawansowanych analiz tego typu warto wykorzystać wyspecjalizowane narzędzia, np. EIGENSOFT<sup>7</sup>.

### Dodatek 1 - przykładowe źródła danych

Zbiory danych do analiz można znaleźć np. na stronach:

<http://evolbio.ut.ee>

<https://reich.hms.harvard.edu/datasets>

### Dodatek 2 - sortowanie i selekcja populacji

Jak wybrać z dużego zbioru danych podzbiór populacji do analizy. Zakładamy, że opisy osobników są w pliku dane.fam, a nazwa populacji jest pierwszą kolumną tego pliku. Najpierw sporządzamy alfabetyczną listę wszystkich populacji.

```
sort -u -t" " -k1,1 dane.fam | awk -F " " '{print $1}' > populations.txt
```

Następnie na podstawie pliku populations.txt tworzymy plik (niech nazywa się wybrane.txt) zawierający tylko te populacje, które chcemy zachować. Teraz tworzymy plik, który będzie zawierał wszystkie osobniki z wybranych populacji:

```
grep -F -f wybrane.txt dane.fam > keep.txt
```

Następnie wykorzystujemy PLINK by stworzyć zbiór, który będziemy analizować:

```
plink --bfile dane --keep keep.txt --make-bed --out wybranedane
```

Uzyskamy pliki wybranedane.bed wybranedane.bim wybranedane.fam ,które możemy już analizować.

---

<sup>7</sup> <https://github.com/DReichLab/EIG/>