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[1]. 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[2], 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

[1] https://github.com/DReichLab/EIG/tree/master/CONVERTF

[2] 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[3], obejmującej ponad 2000 próbek.

Dane zawarte są w plikach o nazwie EuAs.

Analizę przeprowadzamy komendą:

```
admixture -j4 EuAs.bed K
```

gdzie *K* jest wskazaną przez prowadzącego wartością liczby populacji wyjściowych[4]. 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 znajdą 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)[5].

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.

Pierwszym krokiem w R jest wczytanie przygotowanej wcześniej funkcji:

```
source ("barNaming.R")
```

ta komenda wczyta zawartość pliku barNaming.R - prostą pomocniczą funkcję zwiększającą czytelność opisu wykresu. Obejrzyj kod w tym pliku i spróbuj zrozumieć, co robi ta funkcja.

Następnie wczytujemy wyniki do tabeli w R (podano przykład dla *K*=3, dla innych wartości odpowiednio zmieniamy liczbę).

```
tbl<-read.table("EuAs.3.Q")
```

Obejrzyj początek tej tabeli komendą:

```
head(tbl)
```

Zauważ strukturę danych i nagłówek.

W kolejnym kroku wczytamy tabelę z danymi osób, dodatkowo nadając kolumnom nazwy (nagłówki):

```
indTable<-read.table("EuAs.fam", col.names=c("Pop","Sample","c1","c2","c3","c4"))
```

Następnie łączymy obie tabele:

```
merged=cbind(tbl, indTable)
```

Obejrzyj strukturę tej tabeli tak, jak powyżej, zwróć uwagę na nagłówek. W kolejnym kroku przygotujemy tabele posortowane według różnych kolumn. Na początek posortujemy według nazw populacji, co pozwoli nam przeanalizować ich strukturę:

```
ordered=merged[order(merged$Pop),]
```

---

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

[4] 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.

[5] Np. http://www.royfrancis.com/pophelper/

Następnie przygotujemy tabele posortowane według udziału każdej z populacji przodków (w kolumnach V1, V2 - V_K), przykładowo:

```
ordered1=merged[order(merged$V1),]
```

I tak samo dla kolejnych składowych, aż do *K*. Pamiętaj, że strzałka w górę oszczędza wpisywania!

Teraz przygotujemy wykresy. Do eksperymentowania można wyświetlać wykresy na ekranie, jednak ze względu na dużą ilość danych, lepiej jest przygotować pliki PDF o odpowiednio dużej powierzchni. Parametry dobiera się metodą prób i błędów, podane poniżej powinny zadziałać dla analizowanych danych.

```
pdf(file="plot.pdf", 20,4)
mp <- barplot(t(as.matrix(ordered[, 1:3])), col=rainbow(3),border=0, space=0, axes=F, axisname=F)
text(mp, par("usr")[3], labels = barNaming(ordered$Pop), srt = 45, adj = c(1.1,1.1), xpd = TRUE, cex=.2)
dev.off()
```

Kilka słów wyjaśnienia: pierwsza linijka otwiera plik PDF i podaje jego wymiary (w calach). Druga linijka tworzy właściwy wykres, ale bez opisów. Parametr col nadaje inny kolor każdej składowej, w nawiasie podajemy liczbę odpowiadającą wartości *K* (tu 3). Opisy dodawane są do wykresu w trzeciej linijce, tu każdy słupek podpisywany jest nazwą populacji z tabeli (ordered$Pop), ale po "przepuszczeniu" przez funkcję barNaming. Tekst umieszczany jest pod każdym słupkiem, obrócony o 45 stopni i zmniejszony do 20% domyślnego rozmiaru (parametr cex). Parametr adj umieszcza tekst w odpowiedniej odległości od słupka. Czwarta linijka dodaje oś Y. Ostatnia linijka zamyka plik (nie wolno o tym zapomnieć!).

Podobnie sporządzamy wykresy dla każdej z pozostałych posortowanych tabeli, pamiętając o zmianie nazwy pliku i zmianie nazwy tabeli w drugiej i trzeciej linijce.

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[6].

Przeanalizuj wyniki w oparciu o informacje z prezentacji. 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 mieszkaniec Europy sprzed około 5000 lat (epoka miedzi, przełom neolitu i epoki brązu), zaś MA1 to przedstawiciel kultury Mal'ta–Buret' z paleolitu (ok. 24 tys. lat temu) z terenu Syberii. Więcej informacji o nich znajdziesz w sieci.

2. Analiza metodą skalowania wielowymiarowego (MDS)

Z metodą tą spotkaliśmy się już przy analizie asocjacji, gdzie służyła do sprawdzenia, czy badana populacja nie zawiera ukrytych zmiennych, dzielących ją na odrębne subpopulacje. Tu mamy zbiór osób o różnej etniczności, co oczywiście powinno dać zauważalną strukturę. Analizę przeprowadzimy za pomocą PLINK dwuetapowo. Najpierw przygotujemy zbiór z obliczonymi korelacjami SNP (odpowiadającymi podobieństwu genetycznemu wyrażanemu przez IBD):

```
plink --bfile EuAs --genome --out EuAsIBD
```

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

```
plink --bfile EuAs --read-genome EuAsIBD.genome --cluster --mds-plot 2 --out EuAsMDS
```

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

```
mds<-read.table("EuAsMDS.mds", head=T)
pdf(file="mds.pdf",10,10)
```

---

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

```
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[7].

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.

[7] https://github.com/DReichLab/EIG/