# Genomika Ewolucyjna i Populacyjna

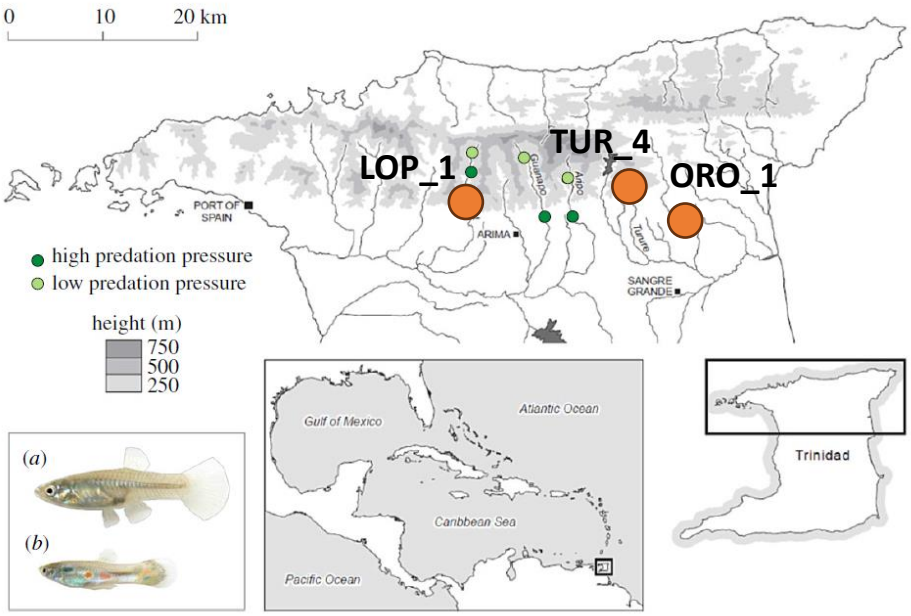
## Ogólnogenomowe resekwencjonowanie dla genomiki populacji

Ćwiczenia 3 Resekwencjonowanie i wywoływanie SNiPów oraz analizy populacyjne

Mateusz Konczal

Na poprzednich ćwiczeniach dowiedzieliśmy się w jaki sposób składa się genomy referencyjne, oraz w jaki sposób mapuje się odczyty do sekwencji referencyjnej, oraz jakie ograniczenia są z tym związanie. Czas przejść do praktycznego zastosowania tej wiedzy. Na tych ćwiczeniach przyjrzymy się w jaki sposób wywołać miejsca polimorficzne oraz przeanalizować wyniki w kontekście populacyjnym.

Spytaj prowadzącego zajęcia gdzie znajdują się pliki udostępnione na dzisiejsze ćwiczenia. Powinieneś móc skopiować je do folderu, który udostępniony jest na maszynie wirtualnej. Dane przekazane Ci do analiz powinny zawierać kilkanaście plików w formacie bam, sekwencji genomu referencyjnej w postaci pliku fasta oraz pliku tekstowego z informacjami o próbkach. Mapowanie odbyło się w podobny sposób do tego o czym uczyliśmy się na poprzednich zajęciach. Sekwencje pochodzą z sekcjonowania genomów gupików - słodkowodnych ryb występujących w rzekach Trynidadu. Na potrzeby ćwiczeń, analizować będziemy tylko chromosom 1. Dwie analizowane populacje pochodzą ze strumieni znajdujących się w zlewni Orropouche (ORO_1, TUR_4), a jedna ze zlewni Caroni (LOP_1).




Ryc. 1. Miejsca próbkowania gupików do resekwencjonowania genomów i do analiz populacyjnych (źródło: Brask i in. 2019, zmodyfikowane).

**1. ZADANIE:**

Prowadzący udostępnił Ci pliki bam. Sprawdź ich wielkość, średnie pokrycie oraz liczbę poprawnie zmapowanych odczytów (możesz wykorzystać metody wykorzystywane na poprzednich zajęciach, lub inne programy, np. `qualimap`). Posortuj wszystkie pliki oraz je zindeksuj. Poniżej podaj polecenia, które użyłeś w tym celu.

W celu prowadzenia analiz, musimy również przeprowadzić indeksowanie pliku fasta:

```
samtools faidx chr1.fasta
```

Informacje o próbkach i wygenerowanych plikach bam opisane są w pliku `Files4Lab3.txt` Pierwsza kolumna zawiera informację o próbce (ta informacja powinna znajdywać się również w polu RQ plików bam), natomiast druga informację o populacji, z której dana próbka pochodzi. Wytnij te kolumny i zapisz w nowym pliku, który potrzebny będzie w dalszym kroku:

```
cut -f 1,2 Files4Lab.txt > Pops.txt
```

Jak dowiedzieliśmy się wcześniej, podczas wykrywania polimorfizmów wykorzystuje się przewidywania prawa Hardy'ego-Weinberga. Aby nasz algorytm działał dobrze, warto uwzględnić informację o populacjach w dalszych analizach, tak że przewidywania HW liczone są wewnątrz każdej ze zdefiniowanych grup.

Wywoływanie miejsc polimorficznych jest zadaniem wymagającym sporo czasu, gdyż wiarygodności liczone są dla każdego miejsca w genomie. Często referencje dzieli się na mniejsze kawałki i obliczenia te wykonuje równolegle. Przykładowo poniżej analizy wykonywane są dla pierwszego miliona miejsc na chromosomie 1, gdzie plik tekstowy zawiera informacje o wszystkich plikach bam, które będą analizowane.

```
ls *.bam > BAM.txt

BAMLIST=BAM.txt

POPS=Pops.txt

time bcftools mpileup -a AD -Ou -f chr1.fasta -r chr1:1:1000000 -b \
${BAMLIST} | bcftools call -G ${POPS} -mv -Ob -o tmp.bcf
```

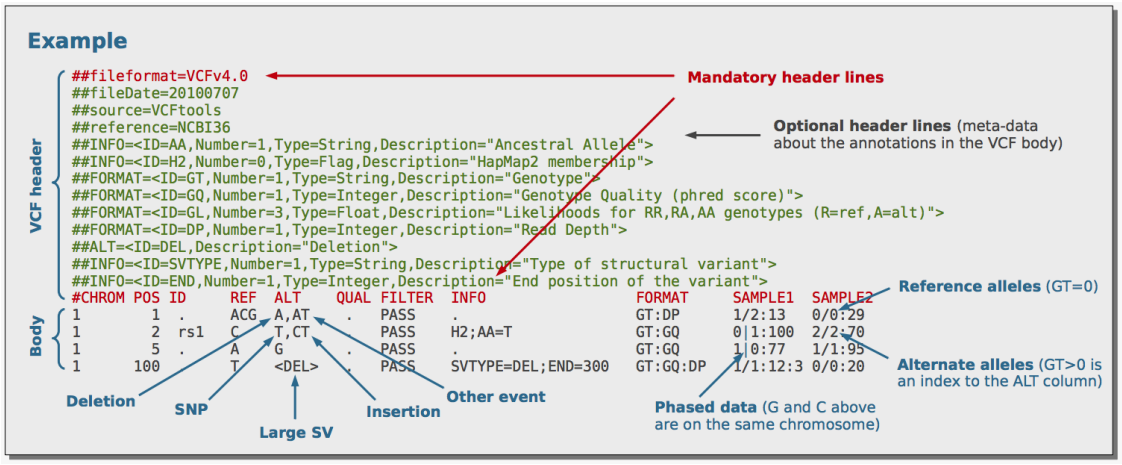
## 2. ZADANIE:

Wykonaj powyższe obliczenia dla fragmentu genomu uzgodnionego z prowadzącym. Podaj polecenia, które w tym celu wykonałaś/eś. Sprawdź i opisz opcje, które wykorzystałeś.

**Format VCF/BCF**

Format bcf jest binarną formą tekstowego pliku vcf. Struktura pliku vcf przedstawiona jest na Ryc. 1. Ogólnie, pomijając wiersze nagłówka (zaczynające się od #), informacje zakodowane są w tabeli, gdzie kolejne wiersze reprezentują kolejne pozycje w sekwencji referencyjnej, natomiast wiersze informują o osobnikach i ich genotypach. Przeformatuj plik bcf do vcf zgodnie z poniższym poleceniem, a następnie podejrzyj jego zawartość oraz przedyskutuj jej znaczenie z prowadzącym ćwiczenia

```
bcftools covert -O v -o tmp.vcf tmp.bcf
```




Ryc. 1 Podstawowe informacje zawarte w pliku vcf (źródło: https://davetang.github.io/learning_vcf_file/)

## 3. ZADANIE:

Policz ile miejsc polimorficznych wykryłeś w Twoim fragmencie genomu. Poniżej podaj polecenie oraz wynik.

**Filtrowanie plików vcf**

W wygenerowanym pliku bcf/vcf znajdują się pozycje, co do których istnieje podejrzenie, że mogą być polimorficzne w analizowanej przez nas próbce. Wykorzystany przez nas program generuje również informację z jakim prawdopodobieństwem dane miejsce jest polimorficzne. Informacja ta zawarta jest w kolumnie QUAL, a prawdopodobieństwo wyrażone jest w fredach (*ang. phreds*), obliczona zgodnie z równianiem 20 w pracy Li (2011).

---

**4. ZADANIE:**

W fredach (*phreds*) kodowane są informacje o jakości nukleotydów, mapowania czy w końcu polimorfizmów. Poniżej opisz jak interpretować te wartości, oraz jak interpretować warości w kolumnie QUAL pliku vcf:

---

Zanim przejdziemy do filtrowania polimorfizmów, dobrze jest dowiedzieć się jak kształtują się rozkłady niektórych statystyk. Użyj poniższych poleceń, żeby wygenerować tabelę, z informacjami o prawdopodobieństwie, że dane miejsce jest polimorficzne (QUAL), średniej jakości mapowania (MQ) oraz sumarycznym pokryciu (DP) w każdym z miejsc.

```
BCF=tmp.bcf

bcftools query -f '%QUAL\t%MQ\t%DP\n' ${BCF} > Stats_QualMQDP.txt

bcftools stats ${BCF) > Stats.stat.txt
```

**5. ZADANIE/DYSKUSJA:**

Użyj wygenerowanego powyżej pliku w celu stworzenia wykresów pokazujących rozkład poszczególnych statystyk. Zinterpretuj wyniki i określ jakie wartości użyłbyś to filtrowania danych. Wykresy oraz interpretację przedstaw poniżej. Następnie przedyskutuj te odpowiedzi z innymi studentami i z prowadzącym.

Gdy zdefiniujemy sobie wartości, których chcemy użyć, możemy przejść do filtrowania danych. W tym celu możemy użyć na przykład programu `bcftools view`, gdzie za pomocą opcji `-e` możemy wykluczyć miejsca o wartościach zdefiniowanych przez określone wyrażenie. W ten sam sposób możemy zdefiniować miejsca, które chcemy zachować (`-i`)

Przykładowo, używając poniższego polecenia, wykluczymy wszystkie miejsca które będą miały QUAL mniejszy niż 60, MQ mniejsze niż 30, a DP będzie się mieścić w zakresie od103 do 414

```
bcftools view -v snps -e 'QUAL < 60 || MQ < 30 || DP < 103 || DP > 414' \
tmp.bcf > tmp.filtered.vcf
```

**6. ZADANIE:**

Przefiltruj swoje dane, używając zdefiniowanych wcześniej wartości. Policz ile miejsc zostało usuniętych a ile zostało do dalszych analiz. Poniżej podaj wykonane polecenia oraz ich wyniki.

**7. ZADANIE:**

Zapoznaj się z wyrażeniami, które wykorzystane mogą być do filtrowania danych:
https://samtools.github.io/bcftools/bcftools.html#expressions
W jaki sposób skonstruowałbyś polecenie, którego celem byłoby:

1) Otrzymanie indeli, których QUAL jest większy niż 100

2) Nie ma żadnych brakujących genotypów

3) średnia jakoś genotypu jest większa niż 30

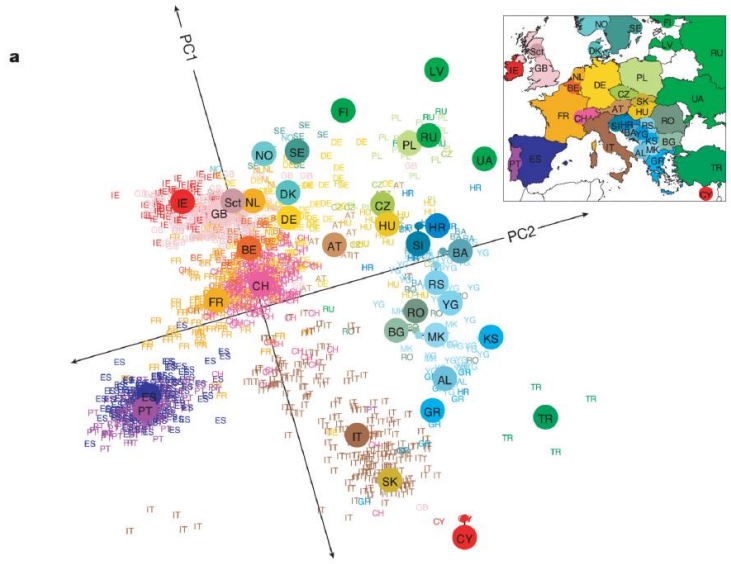
**Analiza głównych składowych – wnioskowanie o strukturze populacji**

Jeżeli mamy już przefiltrowane dane, może przejść do podstawowych analiz z zakresu genetyki populacyjnej. W pierwszej kolejności zobaczmy jak podobne są nasze próbki do siebie. W tym celu możemy użyć programu plink, który policzy nam analizę głównych składowych (*ang. Pricipal Component Analyses, PCA*). Program traktować będzie każdy polimorfizm jako jeden wymiar, po czym przekształci nam dane w taki sposób, żeby pierwszych kilka wymiarów wyjaśniało jak najwięcej zmienności pomiędzy próbkami. Aby przeprowadzić takie analizy wykonaj następujące polecenie:

```
plink --vcf tmp.filtered.vcf --pca
```

## 8. DYSKUSJA/ZADANIE:

Przedyskutuj z prowadzącym wyniki znajdujące się w plikach plink.eigenval i plink.eigenvec. Zastanówcie się w jaki sposób zwizualizować te dane a następnie wykonaj taką wizualizacje. Możecie zainspirować się przykładową ryciną podaną poniżej.




Ryc. 1 Wyniki PCA dla Europejczyków dobrze odzwierciedlają strukturę populacji i podział geograficzny (źródło: Novembre et al. 2008, Nature).

**8*. UWAGA:**

Przeprowadzone przez nas analizy opierają się tylko na kilkunastu osobnikach. Nie pozwala to dobrze oszacować frekwencji alleli potrzebnych dla tego typu analiz. Dlatego nowsza wersja programu plink nie pozwoli na wygenerowanie wyników:

```
plink2 --vcf tmp.filtered.vcf --pca
```

**9. DYSKUSJA:**

Do jakich celów można wykorzystać tego typu analizy. Pomyśl o potencjalnym zastosowaniu w ochronie przyrody, badaniach historycznych czy kryminalistyce.

**Zmienność genetyczna wzdłuż genomu:**

Często interesować nas będzie jak kształtuje się zmienność genetyczna w obrębie populacji. Czy wzdłuż genomu obserwować będziemy zróżnicowanie pomiędzy różnymi regionami chromosomu? Jeżeli tak, to o czym to może świadczyć?

Podstawową miarą, która możemy obliczyć w tym zakresie jest różnorodności nukleotydowa ($\pi$). Miara ta opisuje procent różnic pomiędzy dwoma holotypami wylosowanymi z populacji. Możemy ją obliczyć dla różnych populacji i miejsc na chromosomie. Przykładowo, aby policzyć różnorodność nukleotydową dla populacji ORO_1 w oknach o wielkości 10,000 par zasad możemy wykorzystać następujące polecenia:

```
grep "ORO_1" Pops.txt | cut -f 1 > ORO.txt
```

```
vcftools --vcf tmp.filtered.vcf --keep ORO.txt --window-pi 10000 --out ORO
```

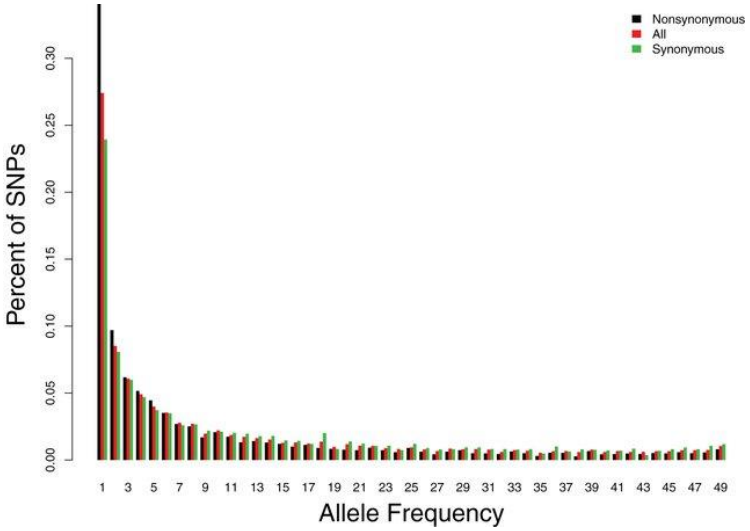
Program vcftools pracuje na wywołanych genotypach zapisanych w formacie vcf. Zakłada on, że każde niewywołane w pliku miejsce jest identyczne z genomem referencyjnym i nie różni się pomiędzy osobnikami, które analizujemy.

**9. ZADANIE:**

Oblicz różnorodność nukleotydową w okach wielkości 10,000 par zasad dla wszystkich 3 populacji dla których masz dane. Porównaj populacje ze sobą i określ, która charakteryzuje się największym, a która najmniejszą zmiennością. Z czego może to wynikać? Zwizualizuj wyniki przedstaw je poniżej.

Inna miarą zmienności genetycznej, opartą na spektrach frekwencji alleli jest statystyka Tajimy D. Statystyka ta przyjmuje wartości bliskie zeru, przy ewolucji neutralnej oraz stałej wielkości populacji. Wartości mniejsze niż zero obserwowane są w przypadku nadmiaru rzadkich wariantów, co związane może być z niedawnym wymiataniem selekcyjnym (*ang. selective sweep*), doborem oczyszczającym, albo/i ekspansją populacji. Wartości większe niż zero obserwujemy, gdy jest brak (lub niedobór) rzadkich wariantów, a więc w przypadku gwałtownego zmniejszenia się wielkości populacji i/lub doboru równoważącego.

Przykładowo, miejsca niesynonimowe znajdują się głównie pod wpływem działania doboru oczyszczającego, dlatego ich spektrum frekwencji alleli (a więc i statystyka Tajimy D) będzie różnić się od polimorfizmów synonimowych tak jak przedstawione jest to na poniższym wykresie.




Ryc. 2 Przykładowe spectrum frekwencji alleli dla populacji ludzkiej (Dania, 25 osób, źródło: Nielsen et al. 2012)

**10. ZADANIE:**

Oblicz wartości Tajimy D, w oknach wielkości 50,000 par zasad, dla wszystkich trzech populacji i porównaj je ze sobą. Poniżej znajdziesz przykładowe polecenie dla jednej z tych populacji. Podaj polecenia oraz zwizualizuj i zinterpretuj wyniki.

```
vcftools --vcf tmp.filtered.vcf --keep ORO.txt --TajimaD 50000 \
--out ORO_TajimaD
```

**Zróżnicowanie pomiędzy populacjami:**

Sprawdźmy teraz jak zróżnicowane pomiędzy sobą są populacje. W tym celu możemy obliczyć na przykład $F_{ST}$ – miara służąca do opisania proporcji całej zmienności genetycznej zawartej w subpopulacjach w stosunku do całej zmienności genetycznej. Miara ta waha się od 1 (jeżeli alternatywne warianty utrwalone są w dwóch populacjach) do 0 (jeżeli nie ma zróżnicowania pomiędzy populacjami) i może być przybliżona następującym wzorem:

$$F_{ST} = \frac{\pi_{between} - \pi_{within}}{\pi_{within}}$$

gdzie $\pi_{between}$ i $\pi_{within}$ reprezentują średnią różnic pomiędzy parą haplotypów wylosowanych z dwóch różnych ($\pi_{between}$) lub z tej samej ($\pi_{within}$) populacji. Fst możemy policzyć dla poszczególnych fragmentów (oknien) genomu o określonej wielkości, identyfikując najbardziej zróżnicowane pomiędzy populacjami regiony genomu. Aby to zrobić możemy wykorzystać na przykład następujące polecenie:

```
vcftools --vcf tmp.filtered.vcf --weir-fst-pop ORO.txt --weir-fst-pop
TUR.txt --fst-window-size 10000 --out oro_tur
```

**11. ZADANIE:**

Obliczy Fst pomiędzy wszystkimi parami w Twojej próbki. Zidentyfikuj najbardziej zróżnicowanie regiony genomu. Sprawdź czy miara $F_{ST}$ koreluje z różnorodnością nukleotydową liczoną wewnątrz populacji. Zwizualizuj i zinterpretuj wyniki.

***12. ZADANIE:**

Powyższe obliczenia prowadzone były na wywołanych genotypach. Na poprzednich zajęciach dowiedzieliśmy się, że obliczenia te mogą być przeprowadzone również na wiarygodnościach. Spróbuj wykonać obliczenia poniżej i porównaj je z tym co uprzednio otrzymanymi wynikami, używając tym razem programu angsd, który pracuje na spektrach frekwencji alleli obliczonych z rozkładów wiarygodności. Poniżej znajdziesz przykładowy sekwencje poleceń:

```
conda install -c bioconda angsd

echo -e 'chr1\t1\t1000000' > PosChr1.bed

angsd sites index PosChr1.bed

awk '$2 == "ORO_1"' Files4Lab3.txt | cut -f 1 | \
sed 's/$/.rmDup.chr1.bam.sorted.bam/g' > O1.txt

angsd -b O1.txt -anc chr1.fasta -out ORO1 -dosaf 1 -gl 1 -sites \
PosChr1.bed

awk '$2 == "TUR_4"' Files4Lab3.txt | cut -f 1 | \
sed 's/$/.rmDup.chr1.bam.sorted.bam/g' > T4.txt

angsd -b T4.txt -anc chr1.fasta -out TUR4 -dosaf 1 -gl 1 -sites \
PosChr1.bed

realSFS ORO1.saf.idx TUR4.saf.idx >ORO1.TUR4.ml

realSFS fst index ORO1.saf.idx TUR4.saf.idx -sfs ORO1.TUR4.ml \
-fstout here

realSFS fst stats2 here.fst.idx -win 10000 -step 10000 \
> slidingwindow
```