# 2024

**dr Bartosz Łabiszak, Zakład Ekologii Roślin i Ochrony Środowiska**

**Genomika ewolucyjna i populacyjna**

**LAB 1**

## Cel ćwiczen:

Na kolejnych trzech spotkaniach będziemy pracować nad zadaniem problemowym. Nadrzędnym celem tych spotkań będzie: ocena **zmienności genetycznej (LAB\_1)** sosny błotnej w porównaniu **do taksonów referencyjnych**, określenie **struktury genetycznej i poziomu zróżnicowania (LAB\_2)** populacji tego gatunku w kontekście ochrony jego zasobów genetycznych. Dowiemy się również o metodach wnioskowania o historii demograficznej taksonów **(LAB\_3).**

## Opis danych do wykonania ćwiczeń

Plik do ćwiczeń: **Genotypy.csv** - zawierają genotypy sekwencji mikrosatelitarnych (**SSR**) loci jądrowego DNA badanych osobników oraz **współrzędne geograficzne badanych populacji**. Analizowane dane pochodzą z 4 izolowanych populacji rzadkiej sosny błotnej (*Pinus uliginosa*) oraz populacji referencyjnych, reprezentujących trzy spokrewnione gatunki sosen (*Pinus sylvestris*, *Pinus mugo*, *Pinus uncinata)*.

Obraz zawierający mapa

Opis wygenerowany automatycznie   
Pinus uliginosa - sosna błotna ▲

Pinus mugo - kosodrzewina ■

Pinus uncinata - sosna hakowata◆

Pinus sylvestris - sosna zwyczajna ●

Z uwagi na charakter zajęć i większą wagę poświęconą interpretacji wyników, **kod w R będzie częściowo podany i wymagający jedynie niewielkich modyfikacji lub krótkiego dopisania własnego kodu**. Funkcje należące do danych pakietów będą oznaczone w kodzie z użyciem operatora „**::**”, np. **poppr::read.genalex()**. Każde zadanie kończy się **pytaniami**, które należy zaadresować w oparciu o wyniki w danej części. Podsumowanie wyników (**ostatnie zadanie w protokole LAB\_3**) jest najistotniejszą częścią trzech zajęć – w której na podstawie wyników analiz odniosą się Państwo do postawionego problemu w formie **sprawozdania**.

## Zadanie I - Wczytanie danych do R i ich pierwszy ogląd

Załaduj wymagane paczki w R (jeśli istnieje konieczność zainstaluj je)

packages <- c(“poppr”, “ggplot2”, “pegas”, “corrplot”, “hierfstat”, “factoextra”, “FactoMineR”, “rnaturalearth”,”rnaturalearthdata”, “sf”, “vegan”,”LEA”, “graph4lg”)

lapply(packages, require, character.only = TRUE)

Wczytaj dane **Genotypy.csv** do R używając funkcji **read.genalex()** tak by otrzymać dane w formacie **genind.** Genind to obiekt wykorzystywany jako podstawowy sposób zapisu danych w pakiecie R **adegenet**. Zawiera informację o genotypach dla każdego osobnika. Więcej o formacie i pakiecie adegenet przeczytasz tu:

<https://cran.r-project.org/web/packages/adegenet/adegenet.pdf>

# wczytanie danych do obiekt klasy genind z danych w formacie genalex

## XXX <- poppr::read.genalex(readalex = … , sep = … ,ploidy = … ,geo = …, genclone = …)

# pamiętaj o tym, że chcemy otrzymać dane klasy genind

# nasze dane zawierają informację o lokalizacji osobników

**Zapoznaj się z informacjami zapisanymi w obiekcie genind (genind$...) i oraz użyj funkcji z pakietu adegenet: summary(), popNames(), pop(), IndNames().**

## Pytania:

1. Ilu loci SSR użyto w tym zestawie danych? **9**

2. Ilu badanych osobników dotyczą te dane? **300**

3. Ile populacji sosen przebadano? **length(unique(pines\_data@pop)) = 12**

4. Jakimi akronimami oznaczono cztery badane gatunki sosen?

* **PUG - *Pinus uliginosa***
* **PM – *Pinus mugo***
* **PUN – *Pinus uncinata***
* **PS – *Pinus sylvestris***

5. W jaki sposób zakodowana była informacja o allelach SSR w formacie genalex, jak zakodowana jest w genind? **W genalex mieliśmy dla każdego osobnika długość loci w parze jego chromosomów, w genid mamy macierz liczności wystąpień każdego locus**

**(możliwe wartości 0, 1, 2)**

6. Które loci odznacza się największą, a które najmniejszą liczbą alleli?

**loci8 największą, a loci2 najmniejszą**

7. Podaj motyw (tj. powtórzenie 2, 3, 4 nukleotydowe) dla każdego z loci:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Loci1 | Loci2 | Loci3 | Loci4 | Loci5 | Loci6 | Loci7 | Loci8 | Loci9 |
| 3 | 3 | 3 | 2 | 3 | 3 | 3 | 2 | 2 |

8. Jaki jest ogólny poziom missing data?

**loci1 loci2 loci3 loci4 loci5 loci6 loci7 loci8 loci9 Mean**

**0.0100 . 0.0033 0.0033 0.0033 0.0033 . . 0.0033 0.0030**

## Zadanie II- Sprawdzenie kompletności danych

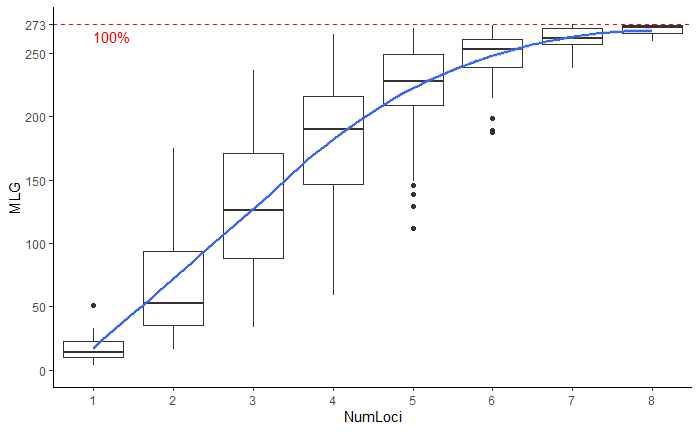
1. Sprawdź przy pomocy krzywej akumulacji genotypów, czy dysponujemy odpowiednią rozdzielczością danych SSR .**Wklej do protokołu powstały wykres**.

poppr::genotype\_curve(gen = …,sample = …, quiet = …)

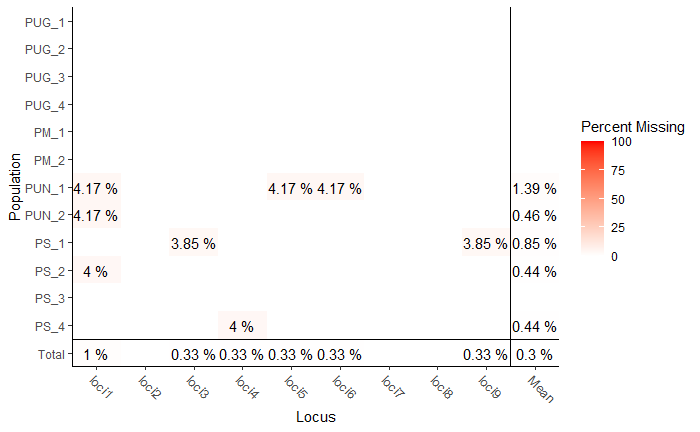
p <- last\_plot()

# last\_plot() używamy by dodać krzywą trendu

p +geom\_smooth(method = “loess”)+theme\_classic()\



Sprawdź w których loci i w których populacjach występują brakujące allele **wklej do protokołu wykres** dla missing data podanych jako **wartość procentowa** i jako **liczba brakujących alleli**.

**Obraz zawierający tekst, diagram, zrzut ekranu, linia

Opis wygenerowany automatycznie**

**poppr::info\_table(gen = …,type = …, plot = …,low = …, scaled = …,percent = …)**

### # wykonaj tę analizę dla liczby i procent brakujących alleli

Wykorzystaj kod podany poniżej do wyłączenia loci i/lub osobników z wysokim poziomem missing data (przyjmując dopuszczalny poziom missing data = 10%)

#### XXX <- missingno(pop = …, type = …, cutoff = …)

## Pytania:

1. Czy w badaniu użyto wystarczającą ilość loci SSR?

**Fajnie byłoby więcej, ale jest jak jest… w sumie, to czy nie było by też więcej missing data, gdybyśmy użyli większej liczby loci?**

2. Czy poziom brakujących danych przemawia za **wyłączeniem loci lub osobników z analiz**? Jeśli tak, to jakie to loci / osobniki? **NIE – bo poziom brakujących danych to po jednym osobniku w populacji któremu brakuje któregoś z 9 alleli**

3. Czy istnieje inny sposób na braki w danych? **Tak – można imputować z występowania w danej populacji**

## Zadanie III- Oszacowanie zmienności wewnątrz populacji i na poziomie gatunków

1. Przeprowadź test równowagi HW **globalnie** (wszystkie osobniki z wszystkich populacji) oraz w **każdej z populacji z osobna**. **Wklej wykres do protokołu**

# test globalny HWE

# test wykonaj na obiekcie genind

**round(pegas::hw.test(XXX, B = 0), digits = 5)**

# test po populacjach

**HWE.test <- (sapply(adegenet::seppop(XXX), function(x) hw.test(x, B=0)[,3]))**

# funkcja seppop() dzieli oryginalny obiekt genind na osobne genind dla

# każdej populacji

# wynikiem jest lista więc stosujemy sapply by wykonać funkcję hw.test

# dla każdego elementu listy. Interesuje nas tylko p value, więc [,3]

# wyciąga nam z powstałej tabeli jedynie 3 kolumnę

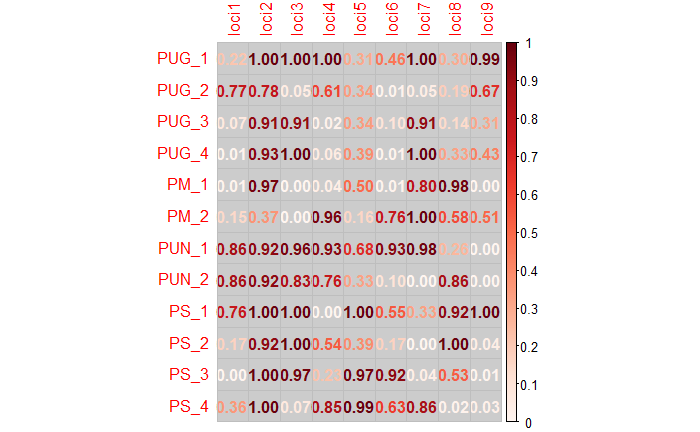
# Używamy transpozycji by dane były w formacie Populacje X Loci

**HWE.test.chisq <- t((HWE.test))**

**round(HWE.test.chisq,5)**

# wizualzacja wartości p jako heatmap

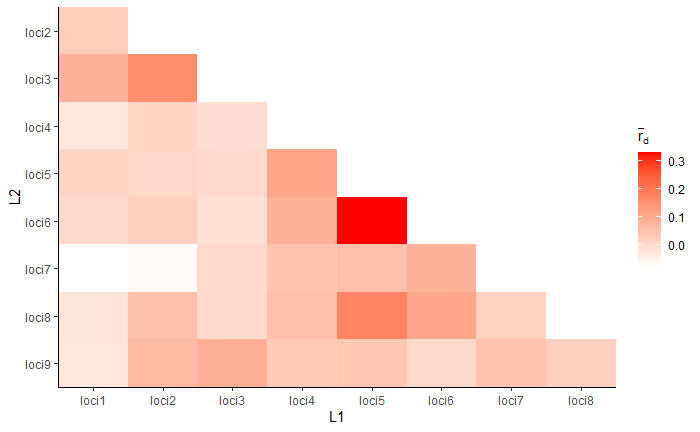
**corrplot::corrplot(corr = HWE.test.chisq,is.corr = F,method = “number”,col = COL1(“Reds”),bg = “grey80”)**



1. Sprawdź, czy nasze loci wykazują nierównowagę sprzężeń (LD). **Wklej wykres do protokołu**.

LD.pair <- pair.ia(XXX,high = “red”, low= “white”)

last\_plot()+ theme\_classic()

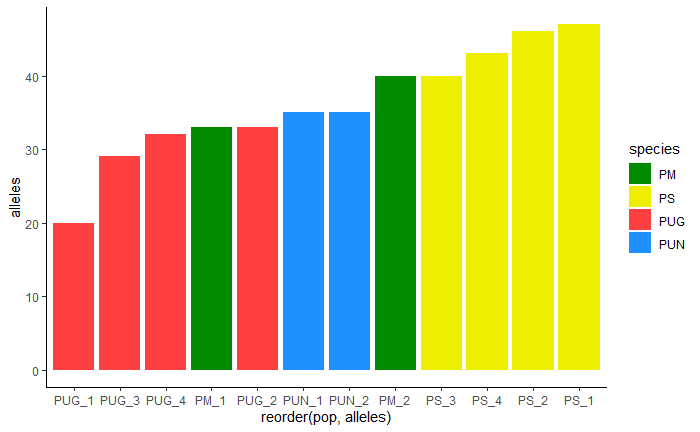
**Wygląda i to na to, że mamy tę nierównowagę sprzężeń – locus5 jest silnie sprzężone z locus6, prawda?**

**W samouczku zastosowanej funkcji jest napisane:**

**Obraz zawierający zrzut ekranu, tekst, Czcionka, linia

Opis wygenerowany automatycznie**

**Skoro indeks sprzężenia jest dla tej pary loci znacznie większy od zera, to znaczy, że istotnie są sprzężone.**

* 1. Wykonaj porównanie liczby alleli w populacjach i **przedstaw wyniki na wykresie słupkowym**. **Wklej wykres do protokołu**. ****

**# Base R**

# stworzenie obiektu zawierającego podsumowanie genind

### XXX <- summary(genind)

# wykorzystaj obiekt z podsumowaniem, nie oryginalny genind

**barplot(XXX$pop.n.all, las=3,xlab = “”, ylab = “Number of alleles”)**

# **kod w ggplot**

# w ggplot mamy większą kontrolę nad wykresem – możemy prosto zmienić

# kolejność barplotów, nadać im kolor zgodny z przynależnością gatunkową

# Stwórz ramkę danych zawierającą 3 kolumny – **liczba alleli**, **populacja**

# oraz **gatunek**

**XXX\_df <- data.frame(alleles = XXX$pop.n.all,**

**pop = …,**

**species = c(rep(XXX,N),rep(XXX,N), rep(XXX,N),rep(XXX, N)))**

**barplot\_XXX <-ggplot(XXX\_df,aes(x= reorder(X,-Y), y= Y, fill = Z))+geom\_bar(stat =”identity”)**

**barplot\_XXX + scale\_fill\_manual(values = c(“green4”, “yellow2”, “brown1”, “dodgerblue”)) + theme\_classic()**

Dla zachowania spójności, będziemy używać następujących kolorów dla naszych gatunków w kolejnych wykresach:

***“green4” – P. mugo, “yellow2” – P. sylvestris, “brown1” – P. uliginosa, “dodgerblue” – P. uncinata***

1. Oblicz **różnorodność alleliczną** (**mean allelic richness, boxplot over species**) **Wklej wykresy do protokołu**.

# allelic richness obliczymy funkcją allelic.richness() z pakietu hierfstat

# musimy wykonać przejście z genind do hierfstat w pierwszym kroku

# wynik prezentowany jest dla każdego loci w każdej populacji – interesuje # nas wyłączenie średnia dla populacji

#### XXX <- genind2hierfstat(XXX)

**allelic.richness(XXX)$Ar**

**mean\_richness <- apply(allelic.richness(XXX)$Ar, MARGIN = 2, FUN = mean) %>% round(digits = 3)**

# do wizualizacji barplotu użyj analogicznego kodu jak w przypadku

# częstości alleli

Obraz zawierający tekst, zrzut ekranu, Wielobarwność, diagram

Opis wygenerowany automatycznie

**# dodatkowo, korzystając z danych allelic.richness(XXX)$Ar**

# możemy zrobić boxplot pokazujący AR na poziomie gatunków

**df <- as.data.frame(t(allelic.richness(XXX)$Ar))**

**df$pops <- rownames(df)**

**df$species <- c(rep(“PUG”,4),rep(“PM”,2), rep(“PUN”,2),rep(“PS”, 4))**

# Zamień na ramkę na długi format do wizualizacji w ggplot

**df\_long <- tidyr::pivot\_longer(data = df,cols = starts\_with(“loci”), names\_to = “loci”, values\_to = “AR”)**

## wizualizacja

**ggplot(df\_long, aes(x = reorder(X,Y), y = Y, fill = Z)) +**

geom\_boxplot() +

labs(x = “Loci”, y = “Allelic richness”) +

**scale\_fill\_manual(values = c(“green4”, “yellow2”, “brown1”, “dodgerblue”)) + theme\_classic()**

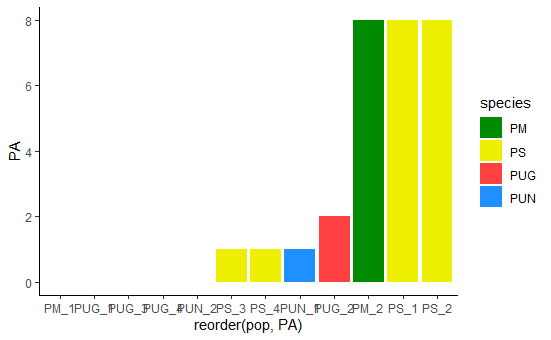
Obraz zawierający tekst, diagram, zrzut ekranu

Opis wygenerowany automatycznie

1. Oblicz **liczbę alleli prywatnych dla populacji i przedstaw ją na wykresie słupkowym**

#### XXX <- poppr::private\_alleles(XXX) %>% apply(MARGIN = 1, FUN = sum)

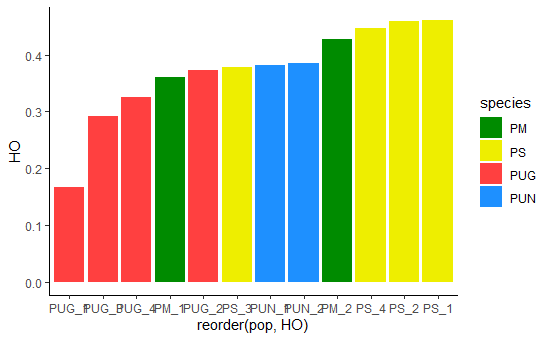
# do wizualizacji barplotu użyj analogicznego kodu jak w przypadku częstości alleli i AR



1. Oblicz heterozygotyczność obserwowaną **(Ho)** i spodziewaną **(He)** oraz współczynnik wsobności **(F)** i **przedstaw je na wykresie słupkowym**

#### XXX\_stats = hierfstat::basic.stats(genind\_object, diploid = TRUE)

## Ho <- apply(XXX\_stats$Ho, MARGIN = 2, FUN = mean, na.rm = TRUE) %>% round(digits = 3)

# analogiczny kod użyj do obliczenia heterozygotyczności spodziewanej i współczynnika wsobności

1. Wykonaj **wizualizację miar zmienności genetycznej na mapie**. Wykorzystaj podany niżej kod:

# zapisanie informacji o lokalizacji osobników jako ramka danych

**loc <- as.data.frame(genind\_object$other)**

# modyfikcja powyższej ramki by zawierała lokalizacje populacji

(hint – każdy osobnik w populacji ma tę samą lokalizację)

**loc\_pops <- XXX(loc)**

# stworzenie ramki danych zawierającej lokalizacje populacji, miary zmienności i oznaczenie gatunkowe

**map\_df <- data.frame( lat =…, long = …,species =… , Ho =…, He =…, Fis =…)**

# przygotowanie mapy

**world <- ne\_countries(scale = “medium”, returnclass = “sf”)**

# wizualizacja pojedynczej miary zmienności, gatunek za pomocą kształtu, miara kolorem

**ggplot(data = world) +**

**geom\_sf() +**

**geom\_point(data = map\_df, aes(x = long, y = lat, shape = species, color = Ho), size = 4) +**

**coord\_sf(xlim = c(-10, 40), ylim = c(40, 70), expand = FALSE)+**

**scale\_shape\_manual(values=c( 15, 16, 17,18))+**

**scale\_color\_continuous(type = “viridis”)+**

**theme\_minimal()**

# analogiczną wizualizację wykonaj dla pozostałych statystyk

##### Pytania:

1. Czy nasze loci wykazują odstępstwa od równowagi Hardego- Weinberga?

2. Uwzględniając wyniki **testu HW** i przyjmując próg istotności p = 0.05 odpowiedz na pytanie, czy któreś z loci należy wykluczyć **z analiz na podstawie testu HW** (czy wykazuje systematyczny wzór odstępstwa od HW dla wszystkich populacji)?

3. Czy **poziom korelacji** pomiędzy loci w teście LD jest **wysoki**?

4. Wymień populacje, które charakteryzują się najmniejszą oraz największą obserwowaną i spodziewaną heterozygotycznością.

5. Który gatunek odznacza się najmniejszą różnorodnością alleli?

6. Czy populacje sosny błotnej **odznaczają się wsobnością**?

7. Czy nasz dane wskazują na obniżoną zmienność u sosny błotnej?