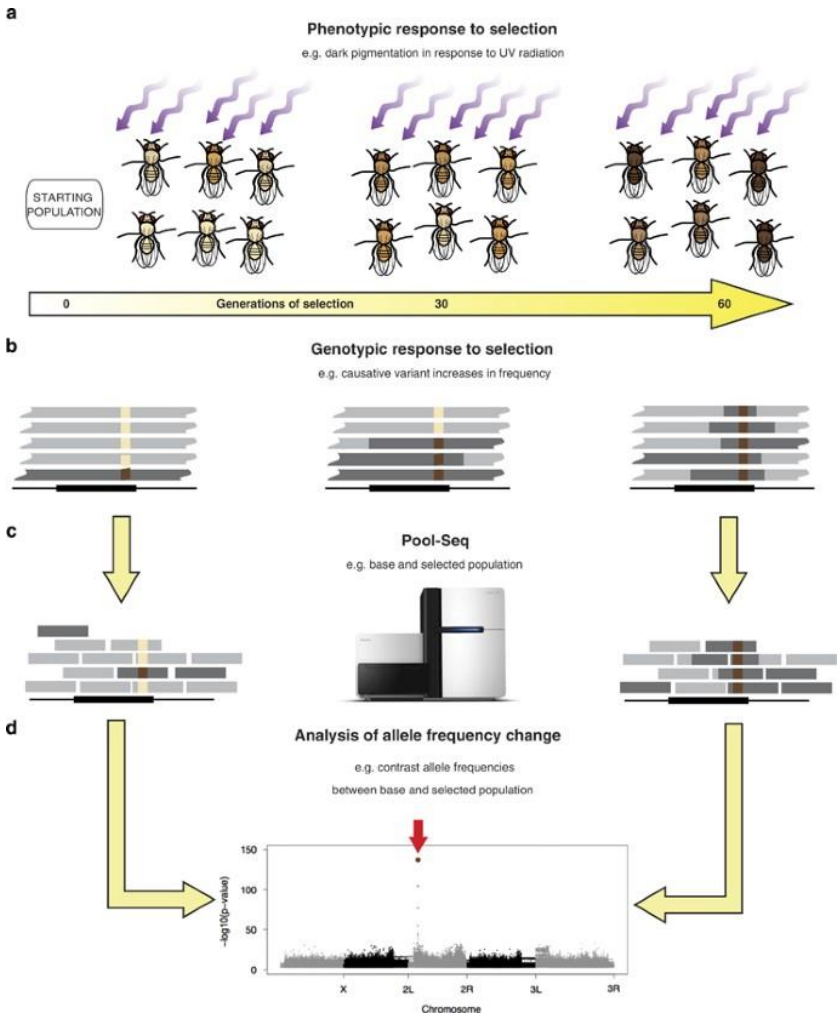
Genomika Ewolucyjna i Populacyjna

Ogólnogenomowe resekwencjonowanie dla genomiki populacji

Ćwiczenia 4 Wyewoluuj i zsekwencjonuj

Mateusz Konczal

Na dzisiejszych zajęciach dowiemy się jak przenalizować eksperyment w ramach eksperymentu typu wyewoluuj i zsekwencjonuj (*ang. Evovle and Resequence*). Eksperymenty tego typu historycznie prowadzone były na muszkach owocowych, obecnie wiele innych gatunków jest poddawanych ewolucji w laboratorium W podobny sposób analizować można eksperymenty na mikroorganizmach, gdzie bakterie selekcjonowane są np. w kontekście obecności antybiotyków, lub eksperymenty na liniach komórkowych. W każdym z tych przypadków optymalne jest sekwencjonowanie pul osobników/ komórek i identyfikowanie genów, w których zaszły największe zmiany frekwencji alleli. Przykładowy schemat analiz przedstawiony jest na poniższej Rycinie.

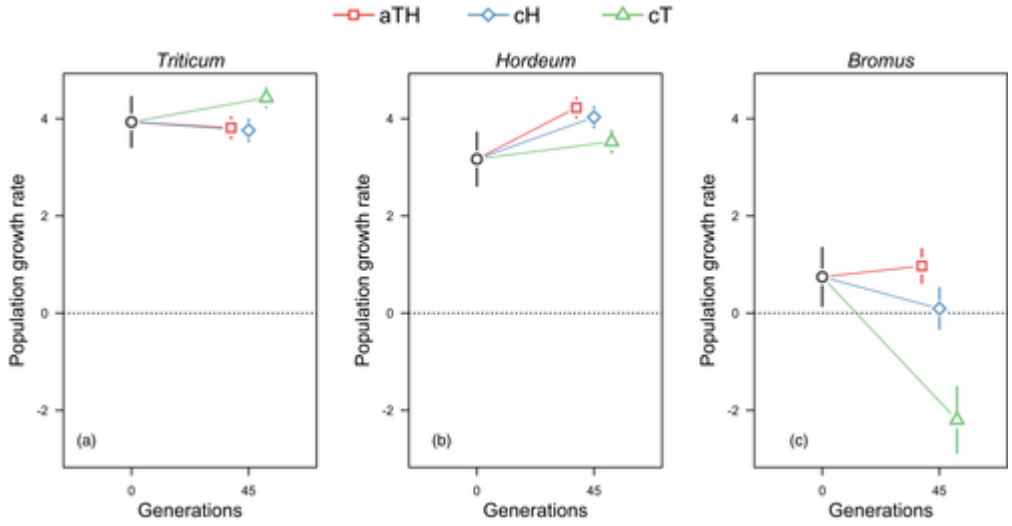


Ryc. 1. Przykładowy schemat analiz w ramach eksperymentu wyewoluuj i zsekwencjonuj (źródło: Schlötterer et al. 2015)

Podczas dzisiejszych zajęć będziemy analizować eksperyment selekcyjny, którego wyniki zostały opublikowane w pracy:

Skoracka A. et al. (2022) Effective specialist or jack of all trades? Experimental evolution of a crop pest in fluctuating and stable environments. Evolutionary Applications https://doi.org/10.1111/eva.13360

Podczas eksperymentu szpeciele z gatunku *Aceria tosichella* selekcjonowane były na różnych roślinach żywicielskich. Szpeciele są pasożytami zbóż uprawnych i występują głównie na pszenicy. W kontrolowanych warunkach laboratoryjnych populacje szpecieli hodowane były na pszenicy (cT) lub jęczmieniu (cH). Trzeci rodzaj populacji to takie, w których roślina żywicielska zmieniał się co trzy pokolenia (3 pokolenia na pszenicy, 3 na jęczmieniu, 3 na pszenicy, 3 na jęczmieniu itd.). Eksperyment przeprowadzony został na 10 niezależnych populacjach w obrębie każdego rodzaju rodziny żywicielskiej. Wszystkie 30 powtórzeń wyprowadzonych zostało z jednej bazalnej populacji i ewoluowało niezależnie przez 45 pokoleń. Po tym czasie populacje zmieniły swoje dostosowania na różnych roślinach żywicielskich zgodnie z tym co zaprezentowane jest na Ryc. 2. Takie fenotypowe zmiany najpewniej były wynikiem działania doboru, działającego różnie w różnych warunkach środowiska. Na dzisiejszych zajęciach będziemy szukać śladów działania tego doboru na poziome genetycznym.



Ryc. 2. Dostosowanie populacji szpecieli *A. tosichella* poddanych ewolucji w oscylujących warunkach środowiska (linie aTH) i w stałych warunkach środowiska, z

Podczas dzisiejszych zajęć każdy dostanie do przeanalizowania sekwencje pochodzące z jednej z wyewoluowanych populacji oraz sekwencje pochodzące z populacji bazalnej z pokolenia zero. Każdy/a student/ka analizować będzie inną linię ewolucyjną, z tym że niektóre linie pochodzą z tego samego reżimu eksperymentalnego. Dlatego też w drugim kroku, pracować będziecie w grupach, żeby zidentyfikować te geny, które powtarzalnie odpowiadają na selekcje w tych samych liniach. Ostatecznie, poproszeni zostaniecie, żeby porównać wyniki uzyskane dla różnych grup.

W przeciwieństwie do poprzednich spotkań zajęcia te mają charakter projektowy. Poniżej przedstawione są przykładowe programy i polecenia, ale ostateczny sposób analizowania danych zostaje do decyzji studentów (po konsultacji z prowadzącym).

Pliki, które będziecie mieć do dyspozycji, to cały genom referencyjny w formacie fasta oraz pliki bam ze zmapowanymi do tego genomu odczytami pochodzącymi z ze zmieszanych z 500 osobników próbek. Pliki bam zawierają tylko te odczyty, które zmapowały się z jakością większą niż 20. Dodatkowo udostępniony Wam zostanie również plik z annotacją strukturalną.

Ogranicz swoje analizy do sekwencji o nazwie contig_37.

```
##Program popoolation i popoolation2, służące do analiz danych typu Evolve and Resequence

mkdir Lab4

cd Lab4

wget \
https://sourceforge.net/projects/popoolation/files/popoolation_1.2.2.zip

unzip popoolation_1.2.2.zip

cd popoolation_1.2.2

P1_DIR=`pwd`

cd ..

wget \
https://sourceforge.net/projects/popoolation2/files/popoolation2_1201.zip

unzip popoolation2_1201.zip

cd popoolation2_1201

P2_DIR=`pwd`

cd ..


##Tworzenie pliku mpielup:

GENOME=Aceto_genome.fasta

BAM1=H_4_15.rmDup.q20.sorted.bam

BAM2=Stock_0.rmDup.q20.sorted.bam

samtools faidx ${GENOME}

MPILEUP1=out1.mpileup

samtools mpileup -r contig_37 -f ${GENOME}  ${BAM1} ${BAM2} > ${MPILEUP1}


#Filtrowanie miejsc wokół indeli
```

```
MPILEUP2=out2.mpileup

perl ${P1_DIR}/basic-pipeline/identify-genomic-indel-regions.pl \
--indel-window 5 --min-count 4 --input ${MPILEUP1} --output indels.gtf

perl ${P1_DIR}/basic-pipeline/filter-pileup-by-gtf.pl --input ${MPILEUP1} \
--gtf indels.gtf --output  ${MPILEUP2}

rm  ${MPILEUP1}


###Filtrowanie po pokryciu pozwala usunąć wątpliwej jakości fragmenty
genomu

##Proste skrypty pythona wykorzystane do filtrowania dostępne są w
repozytorium do ćwiczeń

MPILEUP3=out3

python MeanDPfromMpileup.py ${MPILEUP2} 2

cMin=63

cMax=252

nSamples=2

python FilterMpileup.py ${MPILEUP2} $cMin $cMax $nSamples > ${MPILEUP3}

rm ${MPILEUP2}


##Generowanie pliku SYNC

SYNC=Szp_EEv2.IND.COV.sync

java -Xmx5g -jar ${P2_DIR}/mpileup2sync.jar --input ${MPILEUP3} \
--output ${SYNC} --fastq-type sanger --min-qual 20 --threads 2


#sprawdź wygenerowany plik SYNC

#less -S ${SYNC}


##Fst pomiedzy populacjami  w oknach o wielkości 10kb

perl ${P2_DIR}/fst-sliding.pl --input ${SYNC} --output \
Szp_EEv2.IND.COV_w10k.fst --min-count 6 --min-coverage 40 \
--max-coverage 9999 --window-size 10000 --step-size 10000 --pool-size 500


##Obliczanie różnic we frekwencji alleli pomiędzy próbkami

perl ${P2_DIR}/snp-frequency-diff.pl --input ${SYNC} --output-prefix \
Szp_EEv2.IND.COV_diff --min-count 6 --min-coverage 40 --max-coverage 9999
```

#Test Fishera

cpan Text::NSP::Measures::2D::Fisher::twotailed

```
perl ${P2_DIR}/fisher-test.pl --input ${SYNC} --output Stock_vs_Evolved.fet
\ --min-count 6 --min-coverage 15 --max-coverage 9999 \
--suppress-noninformative
```

#Zidentyfikuj 5% SNPów z najwięszą zmianą frekwencji alleli i/lub z najniższymi wartościami testu Fishera (sprawdź w jakiej formie podawane są te wartości w pliki wynikowym), oraz 5% okien z najwyższym Fst. Sprawdź ile z nich pokrywa się z wynikami dla innych linii będących w tym samym albo w innym reżimie selekcyjnym.

#Jak interpretować zmiany, które są takie same dla większości linii niezależnie od rośliny żywicielskiej.

#Jak interpretować powtarzane wyniki wewnątrz tego samego typu linii, ale nie pomiędzy nimi.

#W jakich genach znajdują się największe zmiany. Z czym związane mogą być te geny.

#Czy obserwowane wzorce mogą być wynikiem działania dryfu genetycznego? Jak testowałbyś tego typu pytanie?

#Zwizualizuj dane, w sposób podobny, do przedstawionego na dole Ryc. 1.

```
######################################################################
```

##Poniższe obliczenia mogą trwać zbyt długo jak na czas standardowych zajęć, ale są ważnym elementem analiz Evolve and Resequence

#Obliczanie różnorodności nukleotydowej dla każdej próbki

```
cut -f 1,2,3,4,5,6 ${MPILEUP3} > S1.mpileup
```

```
cut -f 1,2,3,7,8,9 ${MPILEUP3} > S2.mpileup
```

```
perl ${P1_DIR}/Variance-sliding.pl --measure pi --input S1.mpileup \
--output S1.pi --pool-size 1000 --min-count 3 --min-coverage 15 \
--window-size 20000 --step-size 20000 --fastq-type Sanger
```

```
perl ${P1_DIR}/Variance-sliding.pl --measure pi --input S2.mpileup \
--output S2.pi --pool-size 1000 --min-count 3 --min-coverage 15 \
--window-size 20000 --step-size 20000 --fastq-type Sanger
```

```
#Obliczanie D Tajimy

perl ${P1_DIR}/Variance-sliding.pl --measure D --input S1.mpileup \
--output S1.pi --pool-size 1000 --min-count 3 --min-coverage 15 \
--window-size 20000 --step-size 20000 --fastq-type Sanger

perl ${P1_DIR}/Variance-sliding.pl --measure D --input S2.mpileup \
--output S2.pi --pool-size 1000 --min-count 3 --min-coverage 15 \
--window-size 20000 --step-size 20000 --fastq-type Sanger
```