

# Genomika Ewolucyjna i Populacyjna

## Ogólnogenomowe resekwencjonowanie dla genomiki populacji

### Ćwiczenie 2 Resekwencjonowanie i wywoływanie SNPów – mapowanie do referencji

Mateusz Konczal

## 1. Wstęp

Genetyka populacyjna może być wykorzystywana do identyfikacji wariantów genetycznych segregujących wewnątrz populacji, jak i tych zróżnicowanych pomiędzy populacjami. Z biegiem czasu sekwencjonowanie staje się coraz tańsze, a w związku z tym coraz więcej naukowców (oraz pracowników prywatnych i państwowych firm) decyduje się na resekwencjonowanie całych w celu zrozumienia ogólnogenomowej zmienności. Celem tego ćwiczenia jest zaznajomienie studentów z procesem analizowania surowych odczytów sekwencyjnych i generowania plików do dalszych analiz z zakresu genomiki populacyjnej.

Poradnik przygotowany został na podstawie:

<https://informatics.fas.harvard.edu/whole-genome-ressequencing-for-population-genomics-fastq-to-vcf.html>

oraz artykułu:

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21), 2987-2993.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3198575/>

## 2. Koncept

Istnieje wiele metod służących do izolacji DNA oraz przygotowania bibliotek do sekwencjonowania całych genomów. Te metody są poza zakresem tego kursu. Podczas zajęć zapoznamy się natomiast z dwiema strategiami dotyczącymi projektowania takich badań. Przez pierwsze dwa tygodnie analizować będziemy próbki pochodzące z osobno zsekwencjonowanych diploidalnych osobników, podczas trzecich zajęć natomiast przyjrzymy się analizom próbek przygotowanych z tzw. pul, czyli zmieszanych wielu osobników zsekwencjonowanych jako jedna próbka.

Główną zaletą sekwencjonowania pojedynczych osobników jest możliwość wywołania genotypów, a czasem otrzymania informacji o haplotypach i innych użytecznych statystyk. Przez najbliższe dwa tygodnie skupimy się na zadaniach i metodach pozwalających otrzymanie plików VCF z indywidualnie zsekwencjonowanych osobników diploidalnych. Analizować będziemy sekwencje pochodzące z genomów gupików i zsekwencjonowane za pomocą technologii Illuminy. Zanim jednak do tego przejdziemy, poznamy kilka podstawowych zagadnień teoretycznych i algorytmów wykorzystywanych w tego typu analizach.

---

**DYSKUSJA:**

Przypomnij sobie oraz przedyskutuj z resztą grupy oraz z prowadzącym główne technologie służące do wielkoprzepustowego sekwencjonowania. Czym różni się technologia sekwencjonowania wykorzystywana przez Illuminę od innych technologii. Podaj wady i zalety tych technologii oraz ich potencjalne zastosowanie.

---

### 3. Jak zaprojektować badania

Przy planowaniu sekwencjonowania musimy wziąć pod uwagę kilka czynników, z dużym stopniem wpływających na projektowany budżet przedsięwzięcia. W kolejnych punktach przedyskutujemy w jaki sposób ocenić a) jakość genomu referencyjnego; b) liczbę osobników potrzebnych do sekwencjonowania; c) pokrycie, z jakim chcemy zsekwencjonować osobniki.

a) Czy dostępny jest dla mojego gatunku dobrze złożony i anotowany genom referencyjny?

---

**ZADANIE:**

W przeglądarce internetowej otwórz stronę NCBI, a następnie wyszukaj genomy 3 gatunków oraz odpowiedz na pytania:

- *Homo sapiens*
- *Gallus gallus*
- *Hypsibius exemplaris*

1. Ile złożonych genomów dostępnych jest dla poszczególnych gatunków?

2. Jakiej wielkości mają genomy referencyjne?

3. Ile genów kodujących białka jest anotowane?

3. Czym charakteryzują się statystyki N50 oraz BUSCO? W jaki sposób pomagają one ocenić jakość genomu referencyjnego?

*b) Ile osobników zsekwencjonować?*

Odpowiedź na to pytanie zależy od celu naszych badań. Jeżeli chcemy opisać strukturę populacji i ilość zmienności genetycznej w jej obrębie, to do całogenomowych analiza wystarczy nam dosłownie kilka osobników na populację. Wynika to z faktu, że całe genomy dostarczają dużo informacji, która w dużej mierze jest od siebie niezależna dzięki rekombinacji. Alternatywnie, jeżeli chcemy przeprowadzić dokładną analizę historii demograficznej populacji, mała liczba osobników może być wystarczająca do wykrycia dawnych zdarzeń, ale duże próby są potrzebne, żeby wnioskować o niedawnych zmianach. W końcu identyfikacja zmian frekwencji alleli w konkretnych miejscach czy ogólnogenomowej badania asocjacyjne (tzw. GWAS) wymagają bardzo dużych prób aby osiągnąć odpowiednią moc statystyczną.

---

**DYSKUSJA/ZADANIE:**

Zostałeś poproszony/poproszona o zaprojektowanie badań ogólnogenomowych, których celem ma być porównanie dwóch populacji zagrożonego gatunku ptaka. W szczególności naukowcy chcą dowiedzieć się, która populacja jest bardziej zagrożona (ma mniejszą zmienność genetyczną), oraz jak bardzo różnią się od siebie te populacje. Ile osobników sugerowałbyś zsekwencjonować i dlaczego? Czy wiesz jakie obliczenia zaplanowałbyś, żeby odpowiedzieć na ww. pytania?

---

*c) Jakie pokrycie zaprojektować?*

Pokrycie (lub głębokość sekwencjonowania), to liczba odczytów, które zostały wygenerowane dla określonego miejsca w genomie (Ryc. 1). Złotym standardem w sekwencjonowaniu genomów jest projektowanie eksperymentów tak, aby średnie pokrycie wynosiło 30x na osobnika. Nie oznacza to, jednak, że wszystkie miejsca w genomie będą pokryte 30 odczytami. W wyniku samych tylko losowych procesów część miejsc w genomie będzie zsekwencjonowanych z niższym (lub wyższym pokryciem). Dodatkowo niektóre miejsca w genomie sekwencjonują się gorzej w wyniku czego wariancja jest jeszcze większa.

```
Read 1: CGGATTACGTGGACCATG (read length of 18)
Read 2: ATTACGTGGACCATGAATTGCTGACA
Read 3: ACCATGAATTGCTGACATTCGTCA
Read 4: TGAATTGCTGACATTCGTCA

Depth: 1 1 1 2 2 2 2 2 2 2 3 3 3 3 4 4 3 3 3 3 3 3 3 3 2 2 2 2 2 2 1
```

**Ryc. 1.** Przykład przyrównanych do genomu referencyjnego odczytów pochodzących z sekwencjonowania oraz pokrycia (ang. Depth) w poszczególnych miejscach w genomie.

---

#### ZADANIE:

Otwórz R/Rstudio i wysymuluj oczekiwane pokrycia i jego rozkład używając do tego rozkładu Poissona. Użyj poniższych poleceń, żeby otrzymać wynik dla pokrycia 30x. Następnie zmodyfikuj polecenia tak, żeby zwizualizować oczekiwane wartości dla pokrycia równego 5x, 15x i 50x. Poniżej wklej przykładowe wykresy z odpowiednim opisem oraz odpowiedz na pytanie:

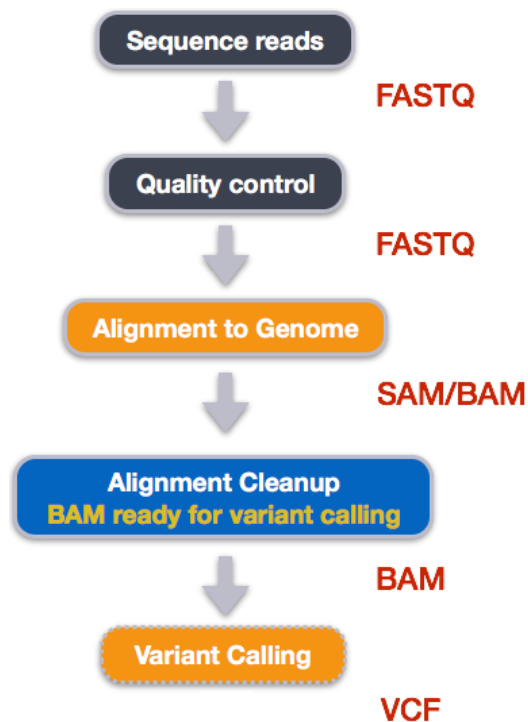
- 1) czy oczekujesz, że niektóre pozycje nie będą w ogóle pokryte odczytami?
- 2) Jak procent ten zmieniać się będzie w zależności od planowanego pokrycia?
- 3) Czy w rzeczywistości spodziewasz się większej wariancji w pokryciu pomiędzy różnymi miejscami w genomie, a jeśli tak to dlaczego?

#### Przykład polecenia w R:

```
> sites=10000
> plot(sites, dpois(sites, lambda=30), type='h')
```

## Mapowanie odczytów do genomu referencyjnego

Po udanym sekwencjonowaniu, odczyty musimy najpierw zmapować do genomu referencyjnego a następnie wywołać miejsca polimorficznych. W tym celu potrzebujemy zazwyczaj wykonać kilka kroków, które przedstawione są na Ryc. 2. Na poprzednich zajęciach przypomnieliśmy sobie w jaki sposób wykonać kontrolę jakości za pomocą programu FastQC oraz jak usunąć odczyty o niskiej jakości za pomocą programu Trimmomatic. Kroki te wyglądają podobnie, niezależnie od tego czy chcemy złożyć genom *de novo*, czy zmapować nasze odczyty do istniejącego już genomu referencyjnego. Poniżej wykonamy kolejne zadania, mające na celu zmapowania wyczyszczonych już odczytów do genomu referencyjnego. Zastanowimy się też nad działaniem algorytmów służących do wywoływania miejsc polimorficznych. Na następnych zajęciach natomiast praktycznie wykorzystamy zdobytą wiedzę, aby wywołać polimorfizmy i wykonać proste analizy populacyjne.



Ryc. 2. Schematyczne przedstawienie kroków potrzebnych do przeanalizowania danych pochodzących z resekwencjonowania genomów.

## Instalacja programu bwa mem2

Mapowanie odczytów do genomu referencyjnego może odbywać się za pomocą kilku powszechnie używanych narzędzi. My podczas zajęć zapoznamy się z jednym z najbardziej popularnych narzędzi, którym jest bwa mem2. Otwórz maszynę wirtualną, stwórz katalog o nazwie Lab2, a następnie przejdź do niego i zainstaluj ww. program, używając do tego dostępnej na maszynie minicondy:

```
conda install -c bioconda bwa-mem2
```

Po udanej instalacji program powinien zostać dodany do ścieżki i być dostępny dla każdego użytkownika.

## Mapowanie do sekwencji referencyjnej

W pierwszej kolejności musimy zindeksować referencje, która powinna być zapisana w formacie fasta:

```
bwa-mem2 index chr1.fasta
```

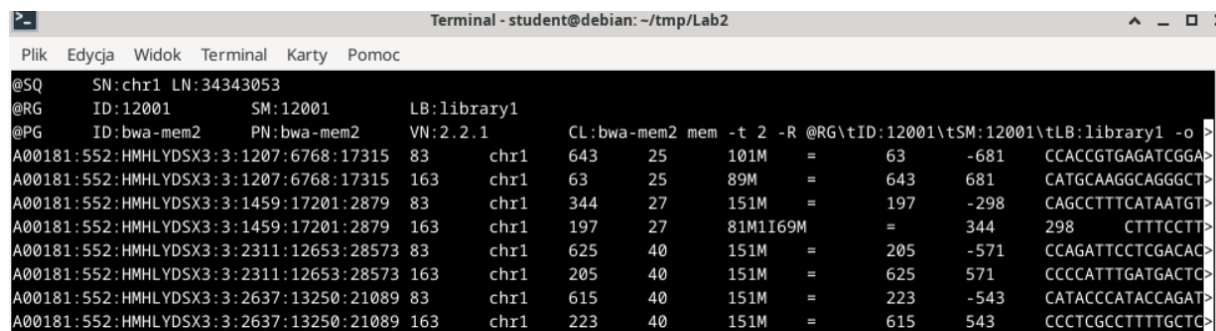
W drugim kroku rozpakujemy pliki fastq oraz zmapujemy je do referencji, używając domyślnych parametrów oraz dodając informację o grupie odczytów:

```
gzip -d 12001.R1.fastq.gz
```

```
gzip -d 12001.R2.fastq.gz
```

```
bwa-mem2 mem -t 2 -R "@RG\tID:12001\tSM:12001\tLB:library1" -o 12001.sam  
chr1.fasta 12001.R1.fastq 12001.R2.fastq
```

## Format plików sam/bam



Read ID	Position	Quality	Mapping Quality	CIGAR	Sequence
A00181:552:HMHLVDSX3:3:1207:6768:17315	83	chr1	643	25	101M
A00181:552:HMHLVDSX3:3:1207:6768:17315	163	chr1	63	25	89M
A00181:552:HMHLVDSX3:3:1459:17201:2879	83	chr1	344	27	151M
A00181:552:HMHLVDSX3:3:1459:17201:2879	163	chr1	197	27	81M1I69M
A00181:552:HMHLVDSX3:3:2311:12653:28573	83	chr1	625	40	151M
A00181:552:HMHLVDSX3:3:2311:12653:28573	163	chr1	205	40	151M
A00181:552:HMHLVDSX3:3:2637:13250:21089	83	chr1	615	40	151M
A00181:552:HMHLVDSX3:3:2637:13250:21089	163	chr1	223	40	151M

Plik sam (bam – binarny format pliku sam) składa się z dwóch części: 1) nagłówki – wiersze zaczynające się od @SQ, @RG, @PG itd., które opisują zawartość pliku i zawierają informacje o próbkach oraz poleceniach generujących i modyfikujących ten plik; 2) przyrównanie – informacja o miejscu oraz podobieństwie z jakim przyrównał się określony odczyt do referencji. Każde przyrównanie reprezentowane jest przez jeden wiersz. Jeden odczyt może mieć więcej niż jedno przyrównanie.

Kolejne kolumny zawierają informacje o: 1) nazwie odczytu, 2) ładzie, 3) nazwie sekwencji referencyjnej, 4) pozycji przyrównania (lewy koniec), 5) jakości mapowania, 6) podobieństwie odczytu do genomu referencyjnego (tzw. CIGAR) itd. W drugiej kolumnie znajduje się flaga, którą zawiera podsumowanie kilku informacji. Zdekodować ją można za pomocą poniższej strony internetowej:

<https://broadinstitute.github.io/picard/explain-flags.html>

---

**ZADANIE:**

Podejrzyj wygenerowany przez siebie plik sam. Wykorzystaj informacje o fladze, oraz podany wyżej link do strony internetowej, żeby zdekodować właściwości zmapowania dla trzech pierwszych odczytów w Twoim pliku. Podaj ich nazwy i uzyskane informacje poniżej.

Dlaczego niektóre odczyty mają takie same nazwy?

Jakiej flagi użyłbyś, żeby znaleźć odczyty, pochodzące z par, w których przynajmniej jeden odczyt nie został zmapowany do referencji?

---

Podsumowanie właściwości zmapowania zakodowanego w flagach możesz wygenerować za pomocą polecenia:

```
samtools flagstats 12001.sorted.bam
```

**Format binarny, sortowanie i indesowanie**

Wiele z dalszych analiz wymaga, żeby przyrównania dostarczone były w formacie binarnym, a odczyty były posortowane zgodnie z miejscem, do którego mapują się w genomie oraz odpowiednio zindeksowane. W tym celu należy wykonać trzy poniższe polecenia.

```
samtools view -b -o 12001.bam 12001.sam
```

```
samtools sort -o 12001.sorted.bam 12001.bam
```

```
samtools index 12001.sorted.bam
```

Teraz przyrównanie można podejrzeć na przykład za pomocą programu `tvview`. Przedyskutuj co widzisz:

```
samtools tvview --reference chr1.fasta 12001.sorted.bam
```

Z powodu ograniczeń finansowych, oraz różnych celów, często konieczne jest sekwencjonowanie do pokrycia niższego niż sugerowane wcześniej 30x. Przeglądając plik za pomocą programu `tvview`, być może zorientowałeś/eś się, że nie zawsze jednoznaczne jest czy w danym miejscu osobnik jest heterozygotą czy homozygotą. Mimo to, dla wielu analiz z zakresu genomiki populacji, bardziej korzystne może być sekwencjonowanie większej liczby osobników do niższego pokrycia. W przypadku tego typu podejścia możliwe jest na przykład oszacowanie spektrum frekwencji alleli, różnorodności nukleotydowej czy struktury populacji w całości na podstawie oszacowań wiarygodności (*ang. likelihood scores*) poszczególnych genotypów. Nie wywołując genotypów, ale szacując ich wiarygodności, można uniknąć błędów i problemów wynikających z sekwencjonowania do niskiego pokrycia. Żeby jednak tak się stało musimy najpierw zrozumieć w jaki sposób tego typu wiarygodności są obliczane. Przyjrzyjmy się w jaki sposób wiarygodności są obliczane dla poszczególnych genotypów diploidalnego osobnika.

---

#### ZADANIE:

Skorzystaj z rozkładu Poissona i oblicz jaka jest prawdopodobieństwo, że heterozygotyczny osobnik będzie miał zsekwencjonowane wszystkie odczyty (*ang. reads*) identyczne, lub tylko jeden odczyt będzie inny od pozostałych. Oblicz takie prawdopodobieństwa dla pokrycia równego 30, 15 oraz 5, zakładając, że prawdopodobieństwo wylosowania (zsekwencjonowania) obu genotypów jest takie same. Możesz posłużyć się poniższym przykładem. Zinterpretuj wynik i przedstaw go poniżej.

#### Przykład polecenia w R:

```
> reads <- 0:30  
> plot(reads, dpois(reads, lambda=max(reads)/2), type='h')
```



Przyjrzyjmy się teraz w jaki sposób obliczane są wiarygodności genotypów:

---

#### PRZYKŁAD DZIAŁANIA ALGORYTMU:

Podążając za oryginalną pracą Henga Li (Li 2011, Bionformatics) postaramy się zrozumieć w jaki sposób obliczane są wiarygodności określonych genotypów dla osobników diploidalnych. Przenalizuj wzór podany niżej oraz przydyskutuj jego interpretację z innymi:

$$\mathcal{L}(g) = \frac{1}{m^k} \prod_{j=1}^l \left[ (m-g)\epsilon_j + g(1-\epsilon_j) \right] \prod_{j=l+1}^k \left[ (m-g)(1-\epsilon_j) + g\epsilon_j \right] \quad (2)$$

gdzie:

- $g_i$  – genotyp (liczba referencyjnych alleli) dla próbki  $i$
- $L(g)$  – wiarygodność (likelihood) określonego genotypu
- $m_i$  – ploidalność określonej próbki (w naszym przykładzie zawsze  $m = 2$ )
- $k$  – liczba odczytów w danym miejscu w genomie (pokrycie)
- $l$  – liczba odczytów z takim samym nukleotydem jak w genomie referencyjnym
- $\epsilon$  – błąd sekwencjonowania

#### ZADANIE:

zaimplementuj go w znanym sobie języku programownia (np. python lub R), lub ewentualnie w Excelu. Przeanalizuj jak zmieniają się wiarygodności oszacowania genotypów ( $g = 0$ ;  $g = 1$ ;  $g = 2$ ), w zależności od pokrycia (np.  $k = 5$ ,  $k = 10$ ,  $k = 30$ ) i błędu sekwencjonowania. Wyniki przedstaw w postaci wykresów i ich interpretacji. Do raportu załącz plik z implementacją.

Powyższy przykład demonstruje w jaki sposób obliczyć można wiarygodność (likelihood) danego genotypu. Znając prawo Hardy’ego-Weinberga, analogicznie obliczyć można prawdopodobieństwo frekwencji alleli w konkretnym miejscu.

---

#### DYSKUSJA:

Przeanalizuj poniższy wzór i zinterpretuj go razem z prowadzącym.

$$\mathcal{L}(\psi) = \sum_{g_1} \cdots \sum_{g_n} \prod_i \Pr\{d_i, g_i | \psi\} = \prod_{i=1}^n \sum_{g=0}^{m_i} \mathcal{L}_i(g) f(g; m_i, \psi)$$

Gdzie:

$$f(g; m, \psi) = \binom{m}{g} \psi^g (1 - \psi)^{m-g}$$

Wyprowadzone jest wprost z prawa HW.

---

Powyżej przedstawiony zostały sposoby obliczenia wiarygodności dla genotypów i frekwencji alleli. Określenie najbardziej prawdopodobnej wartości może odbyć się metodą maksymalnej wiarygodności (*ang. maximum likelihood*), lub wykorzystując twierdzenie Bayesa i tzw. *prior*. Teoria genetyki populacyjnej mówi, że oczekiwana liczba miejsc polimorficznych posiadająca i zmutowanych (w stosunku do stanu ancetralnego) alleli jest odwrotnie proporcjonalna do i. Tak skonstruowany *prior* używany jest do szacowania frekwencji alleli i wywoływania SNPów (dla dociekliwych: równania 20 i 21 w Li 2011).