Protokół z ćwiczeń cz. II: Oszacowanie obciążenia genetycznego

Mikołaj Mieszko Charchuta

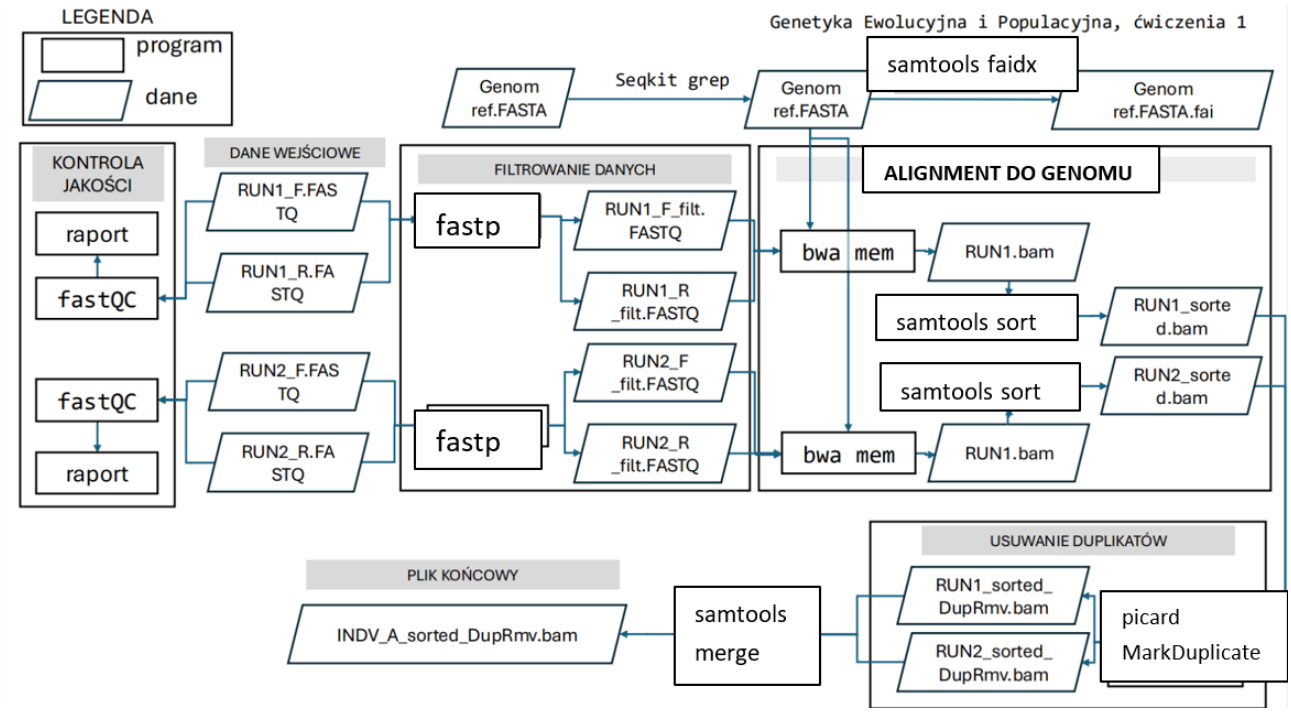
14.02.2025

1 Wprowadzenie

Celem badania było zastosowanie metod ze świeżej publikacji [2] do oszacowania obciążenia genetycznego u dwóch blisko spokrewnionych gatunków ptaków: biegusa łyżkodziobego (*Calidris pygmaea*) oraz biegusa rdzawoszyjego (*Calidris ruficollis*). Biegus łyżkodzioby jest gatunkiem krytycznie zagrożonym wyginięciem, co czyni go idealnym obiektem do badania wpływu małej liczebności populacji na erozję genetyczną. Podchodząc do analiz, spodziewam się, że analizowany osobnik (*C. pygmaea*) będzie miał wyższe obciążenie genetyczne niż osobnik *C. ruficollis*.

2 Materiały i Metody

Schematyczny przebieg podsekcji 2.1-3 przedstawiono na Rysunku 1. Do analizy wykorzystano dane sekwencjonowania genomowego po osobniku z obu gatunków: biegusa łyżkodziobego (C_pyg_26) oraz biegusa rdzawoszyjego (C_ruf_09). Dane ograniczono do analizy scaffoldu 1 genomu referencyjnego biegusa łyżkodziobego.



Rysunek 1: Workflow 1: Przygotowanie danych

2.1 Przygotowanie danych

Sekwencje scaffoldu 1 wyodrębniono z genomu referencyjnego za pomocą narzędzia **seqkit grep**. Plik FASTA scaffoldu 1 zindeksowano przy użyciu **samtools faidx**. Dane sekwencjonowania odczytów zostały pobrane z NCBI SRA (próbki SRS3209979 i SRS3209990, oraz odpowiednio przebiegi: SRR7054135&SRR7054162 i SRR7054133&SRR7054147).

2.2 Analiza jakości odczytów

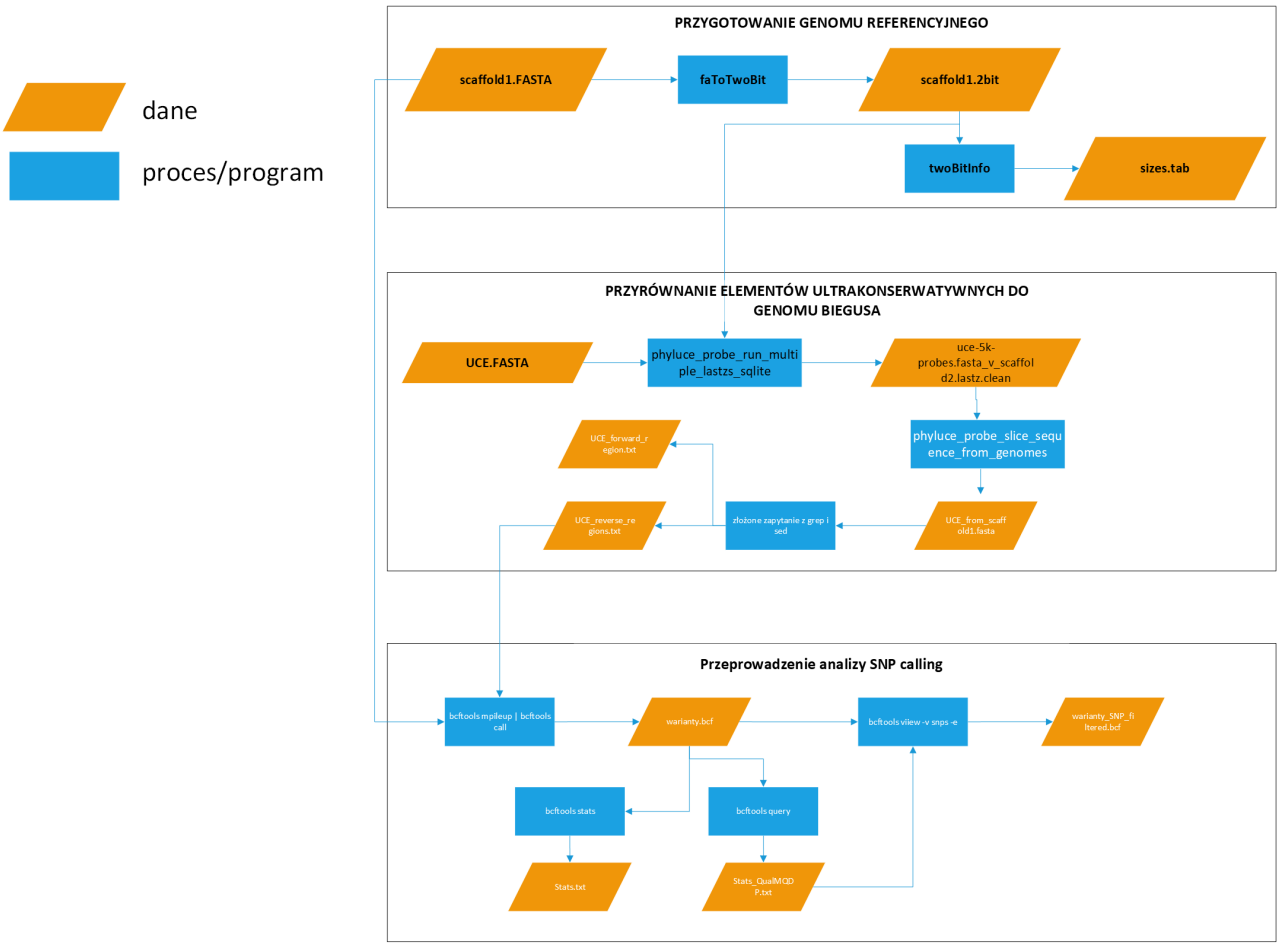
Jakość odczytów przed i po filtrowaniu była wysoka, choć wykryto regiony o niskiej jakości, prawdopodobnie związane z błędami sekwencjonowania. Odczyty sprawdzono oraz przefiltrowano za pomocą narzędzia **fastQC**. Wyniki kontroli i filtrowania przedstawiono w Tabeli 1.

2.3 Mapowanie odczytów

Mapowanie odczytów do scaffoldu 1 wykonano za pomocą Burrows-Wheeler Aligner - **bwa mem**. Pliki pośrednie usunięto, a pliki BAM posortowano programem **samtools sort**. Duplikaty usunięto za pomocą **picard MarkDuplicates**. Końcowy plik .bam uzyskano z obu powtórzeń sekwencjonownia za pomocą **samtools merge**.

Osobnik	Filtrowanie	Run 1	Run 2
C_ruf_09	Przed	Długość odczytów - 125 pz. Wyniki w większości bardzo dobre. Jedna sekwencja jest nadreprezentowana. Blastowana daje różne dziwne wyniki... ale najpewniej to primer Illuminy. Na płytce znajduje się miejsce z sekwencjami o niskiej jakości.	Stała długość odczytów - 125 pz. Wyniki w większości bardzo dobre. Jedna sekwencja jest nadreprezentowana. To samo.
	Po	Odczyty po filtracji mają długość od 42 do 125 pz, głównie $>$ 123 pz. Filtrowanie usunęło tylko małą część nadreprezentowanych sekwencji.	Odczyty po filtracji mają długość od 42 do 125 pz, głównie $>$ 123 pz. Filtrowanie usunęło tylko małą część nadreprezentowanych sekwencji.
C_pyg_26	Przed	Generalnie dobre wyniki. Doszło jednak do problemu na płytce. Te nadreprezentowane sekwencje są spodziewane? Więcej odczytów zawiera GC na poziomie 42% niż przewidziano w teoretycznym rozkładzie.	Wyniki dobre, wykryto nadmiar nadreprezentowanych sekwencji. Więcej odczytów zawiera GC na poziomie 42% niż przewidziano w teoretycznym rozkładzie.
	Po	Usunięto nadreprezentatywne sekwencje. Liczba odczytów z 42% zawartością GC nadal wysoka.	Nadal dużo odczytów z 42% zawartością GC. Usunięto nadreprezentowane sekwencje.

Tabela 1: Analiza jakości odczytów sekwencjonowania



Rysunek 2: Workflow 2: Identyfikacja SNP

2.4 SNPy

Workflow 2 przedstawia proces identyfikacji SNP (Single Nucleotide Polymorphisms).

2.4.1 Konwersja do formatu .2bit

Konieczna dla optymalizacji pamięci. Wykorzystano narzędzie **faToTwoBit**.

2.4.2 UCE

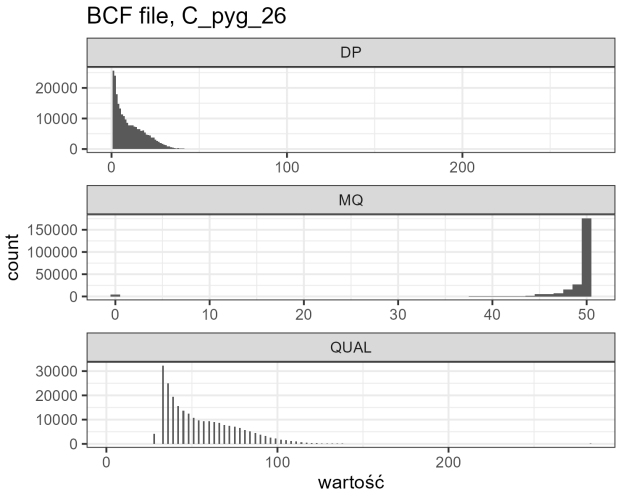
Po przekonwertowaniu danych do formatu .2bit do danych od biegusów przyrównano elementy ultrakonserwatywne programem **LASTZ** w frameworku pakietu **phyluce** wyodrębniając UCE z analizowanych scaffoldów wraz z 1kbp sąsiadujących z UCE. Wszystkie wyodrębnione sekwencje miały orientację REVERSE.

2.4.3 SNP calling

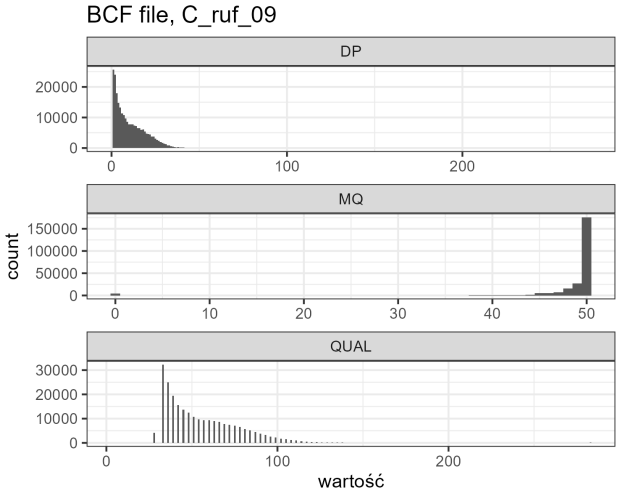
Narzędzie **bcftools mpileup** generuje plik w formacie pileup, który zawiera informacje o odczytach zmapowanych do genomu referencyjnego. Plik pileup zawiera szczegóły na temat pozycji w genomie, alleli referencyjnych i alternatywnych, jakości odczytów oraz liczby odczytów wspierających każdy allel. Narzędzie **bcftools call**.analizuje te informacje o odczytach i decyduje, czy dana pozycja jest polimorficzna (czyli czy występuje wariant).

2.4.4 Filtracja SNPów

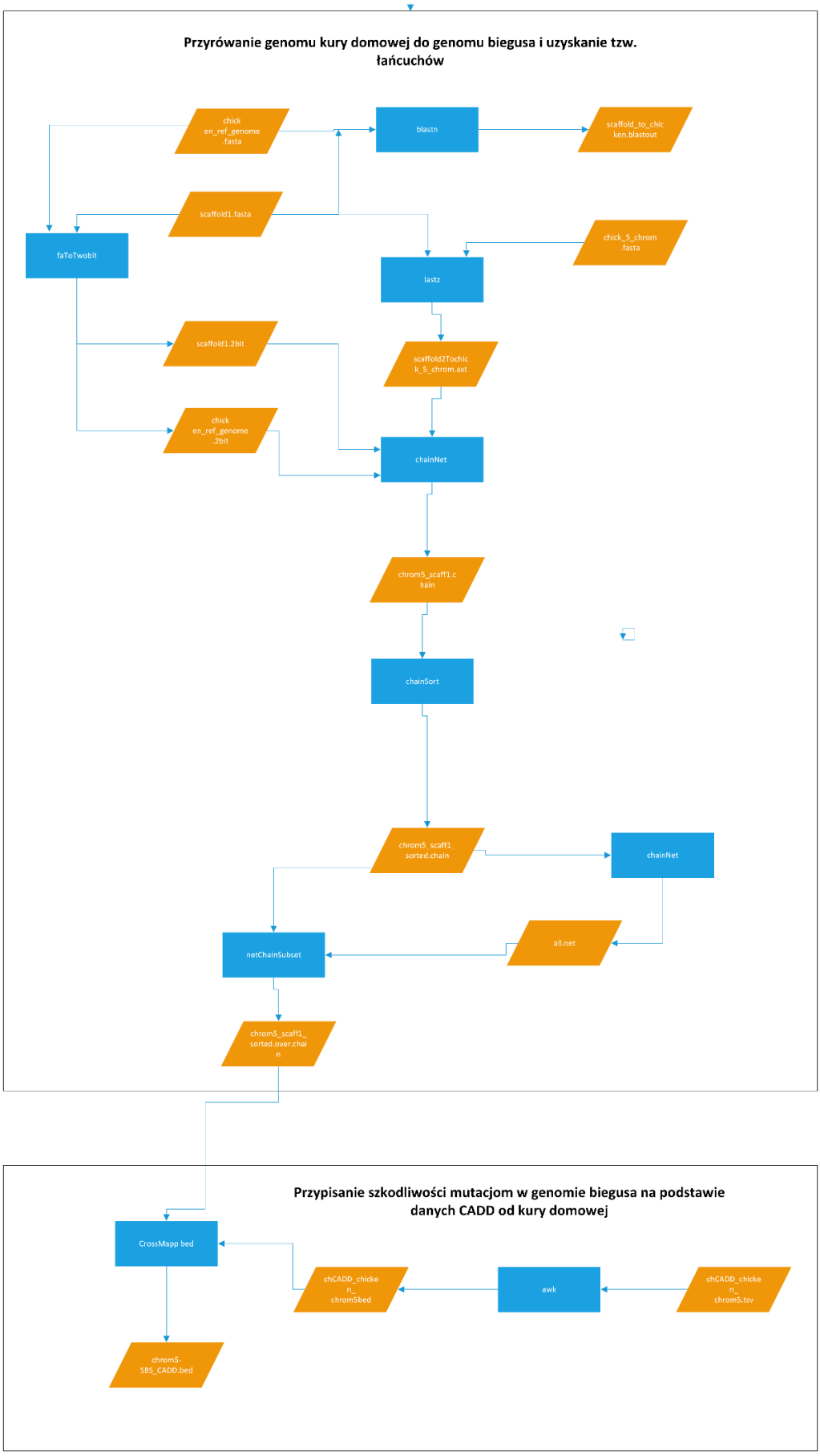
Warianty zostały następnie przefiltrowane przy użyciu **bcftools** i informacji uzyskanych z wykresów 3 i 4, w celu usunięcia niskiej jakości SNP oraz SNP o niskiej częstości alleli. Ostateczne zestawy SNP zostały zapisane w formacie VCF i wykorzystane do dalszych analiz.



Rysunek 3: Statystyki jakości, MQ i DP dla biegusa łyżkodziobego.



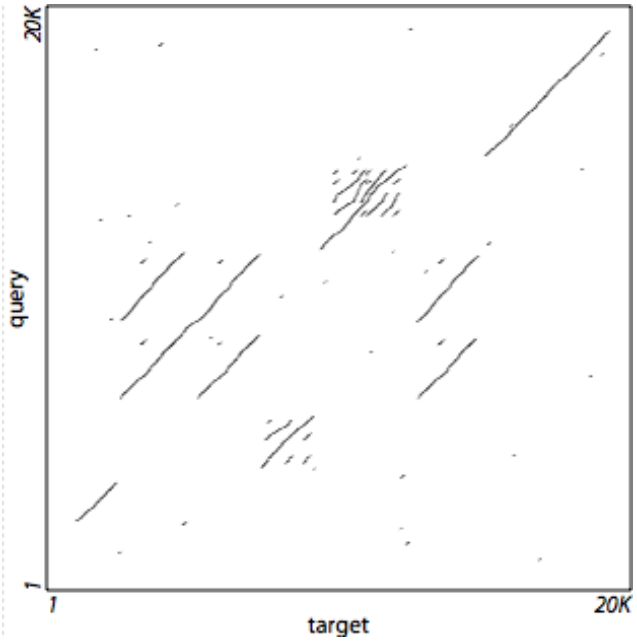
Rysunek 4: Statystyki jakości, MQ i DP dla biegusa rdzawoszyjego.



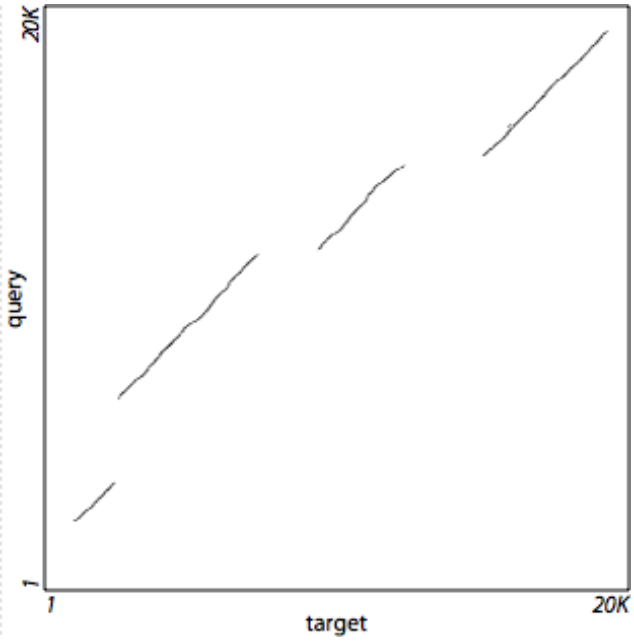
Rysunek 5: Workflow 4:

2.5 Chaining

Wynik działania programu **LASTZ** potraktowano komendą **axtChain** by złączyć je w większe i bardziej spójne bloki (tzw. łańcuchy, chains). Różnica przed i po chainingu przedstawiona jest na Rysunku 6 i 7.



Rysunek 6: Wynik działania programu LASTZ przed chainingiem.



Rysunek 7: Wynik działania programu LASTZ po chainingu.

Utworzone łańcuchy powinny być pofiltrowane. W nowej wersji protokołu w wyniku filtrowania z 1666 łańcuchów pozostało... 256. Wynik filtrowania przedstawiono na Rysunku 8. Zdziwiło mnie to, bo wcześniej filtrowanie usuwało tylko 5 łańcuchów, a teraz usunięto 1410, ale widzę, że dwóm innym studentom pracującym na scaffoldzie 1 również usunęło dużo łańcuchów, więc może to być wynik poprawnego działania programu.

```
(GEiP) st2@bzc1:~/GEiP/Lab3$ grep "chain" galGalChr_5ToSBS_Scaff1.over.chain | wc -l
256
```

Rysunek 8: Wynik filtrowania łańcuchów.

2.6 Przypisanie szkodliwości wariantom

Do przypisania szkodliwości wariantom wykorzystano narzędzie **CADD**, a konkretnie jego model opracowany dla kury (*Gallus gallus*) z publikacji [1]. CADD (ang. Combined Annotation Dependent Depletion) to narzędzie oparte na porównaniu właściwości substytucji (mutacji utrwalonych w linii prowadzącej do człowieka) z właściwościami wymsymulowanych mutacji. Zakłada, że szkodliwe mutacje nie będą pojawiać się wśród substytucji, natomiast będą w danych symulowanych. Właściwości, które porównywane są pomiędzy tymi zbiorami danych, to m.in. informacje o zakonserwowaniu sekwencji (z przyrównania genomów wielu gatunków), poziom ekspresji genów, odległość od granicy egzonu, dane eksperymentalne, informacje o asocjacjach etc. Porównuje się symulowane mutacje z takimi, które obecne są w naturalnych populacjach i w ten sposób trenuje się model. Oszacowane wartości korelują z szacowaną eksperymentalnie patogenicznością mutacji i mogą być obliczone zarówno dla fragmentów kodujących, jak i niekodujących. Wyniki analizy przedstawiono na Rysunkach 9 i 10.

3 Wyniki

3.1 Analiza jakości odczytów

Jakość odczytów przed i po filtrowaniu była wysoka, choć wykryto regiony o niskiej jakości, prawdopodobnie związane z błędami sekwencjonowania.

3.2 Identyfikacja SNP

Dla osobnika C_pyg_26 przed filtrowaniem w pliku bcf było 244504 wariantów, w tym 324 SNP. Po zastosowaniu parametrów filtrowania (QUAL $<$ 20 || MQ $<$ 40 || FORMAT/DP $<$ 3 || FORMAT/DP $>$ 100) pozostały tylko 242 SNP.

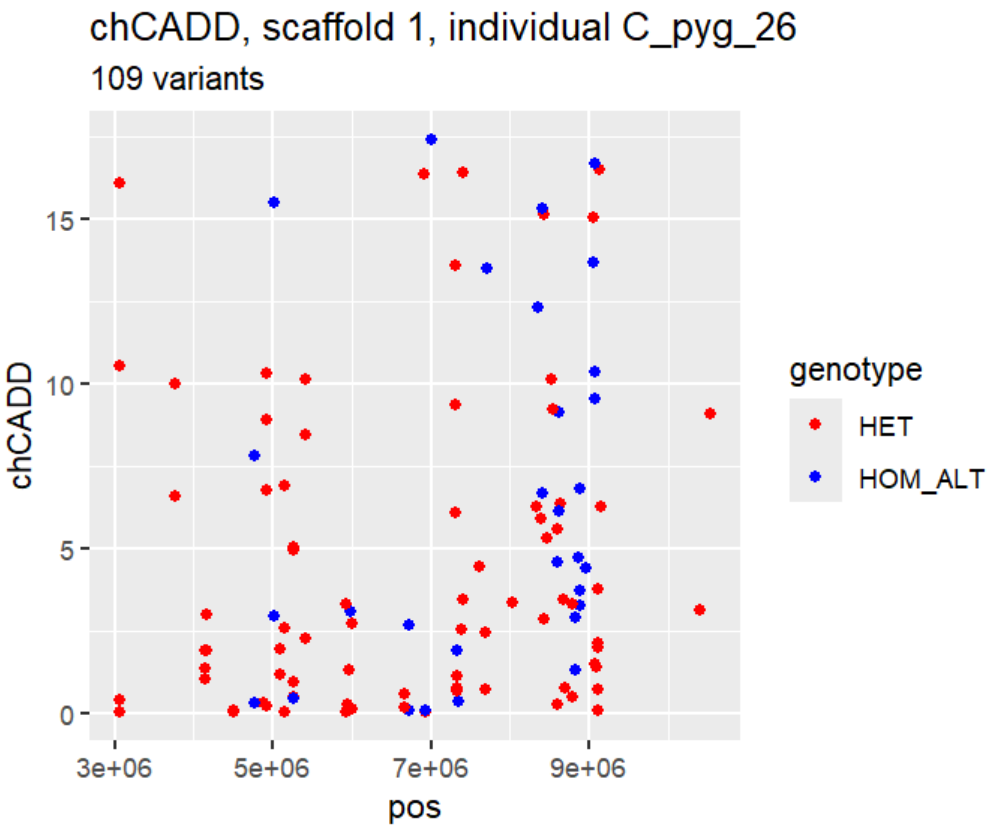
Dla osobnika C_ruf_09 przed filtrowaniem w pliku bcf było 240385 wariantów, w tym 1669 SNP. Po zastosowaniu parametrów filtrowania (QUAL $<$ 25 || MQ $<$ 30 || FORMAT/DP $<$ 2 || FORMAT/DP $>$ 100) pozostało 1314 wariantów, w tym 1314 SNP.

Individual	Mean_CADD	Mean_CADD_HOM	Mean_CADD_HET	VAR	HOM	HET
C_pyg_26	5.029519	6.584459	4.439035	109	30	79
C_ruf_09	4.752241	4.586525	5.121984	501	415	186

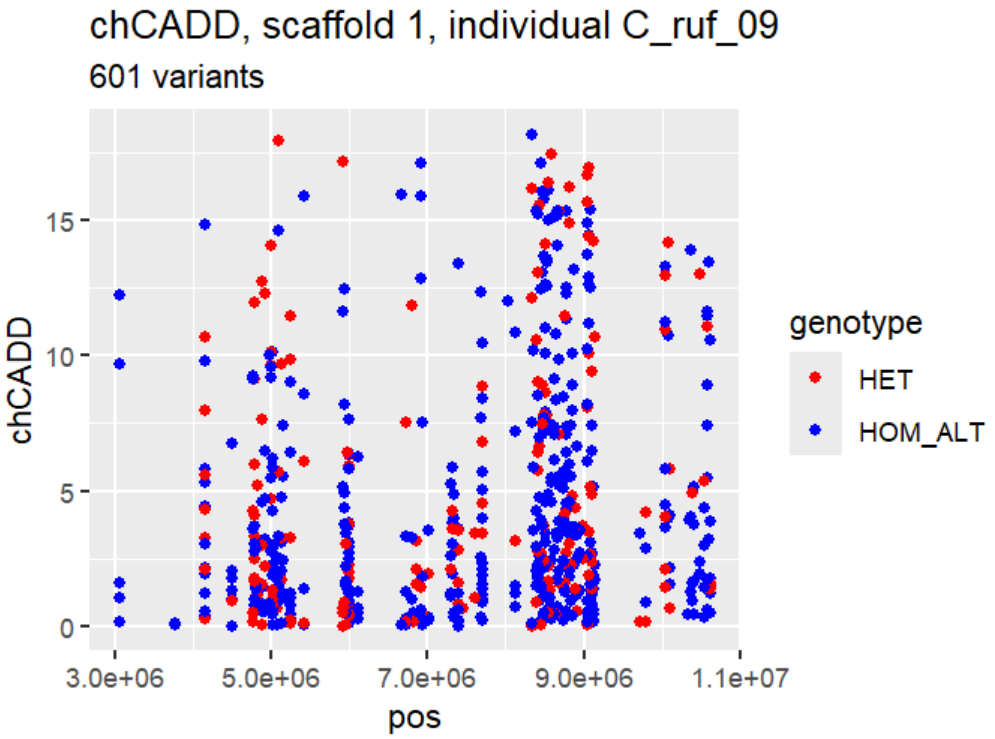
Tabela 2: Podsumowanie punktacji CADD i liczności wariantówdla obu osobników.

3.3 Obciążenie genetyczne

Analizowany biegus rdzawoszyi ma znacznie więcej homozygot w obrębie scaffoldu 1, co może prowadzić do wyższego obciążenia genetycznego. U biegusa łyżkodziobego wykryto więcej heterozygot, co może tłumaczyć niższe obciążenie genetyczne.



Rysunek 9: Rozkład wartości CADD wzdłuż scaffoldu 1 dla biegusa łyżkodziobego.



Rysunek 10: Rozkład wartości CADD wzdłuż scaffoldu 1 dla biegusa rdzawoszyjego.

4 Dyskusja

Wyniki wskazują, że biegus rdzawoszyi ma wyższe obciążenie genetyczne w porównaniu do biegusa łyżkodziobego, co może wynikać z większej liczby homozygotycznych wariantów. Jednak analiza ograniczyła się do jednego scaffoldu, co może nie odzwierciedlać sytuacji w całym genomie. Zastosowanie metod takich jak na zajęciach w hodowli w niewoli może zmniejszyć depresję wsobną i obciążenie genetyczne w populacjach zoo [2].

Literatura

[1] C. Groß, C. Bortoluzzi, D. de Ridder, H. J. Megens, M. A. Groenen, M. Reinders, and M. Bosse. Prioritizing sequence variants in conserved non-coding elements in the chicken genome using chcadd. *PLoS genetics*, 16(9):e1009027, 2020.

[2] S. A. Speak, T. Birley, C. Bortoluzzi, M. D. Clark, L. Percival-Alwyn, H. E. Morales, and C. Van Oosterhout. Genomics-informed captive breeding can reduce inbreeding depression and the genetic load in zoo populations. *Molecular Ecology Resources*, 24(7):e13967, 2024.