Modelowanie genetycznego uwarunkowania cech ilościowych w genomie bydła przy wykorzystaniu narzędzi bioinformatycznych

W pracach wykonanych w ramach projektu porównywano ile i jakie SNP są wybierane jako znaczące przez różne modele. Bazowym modelem był model SNP-BLUP stosowany w ocenie genomowej bydła mlecznego w Polsce. Model ten ma następującą postać:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z_1 q} + \mathbf{e}, \qquad \text{(M1)}$$

gdzie **y** reprezentuje wektor wartości hodowlanych buhajów poddanych deregresji, μ to średnia ogólna, **q** reprezentuje losowy efekt SNP z macierzą wystąpień $\mathbf{Z_1}$, opisany rozkładem $q \sim N(0, \mathbf{I}\sigma_q^2)$ gdzie **I** jest macierzą diagonalną, a σ_q^2 reprezentuje wariancję SNP., **e** jest losowym wektorem efektów błędu opisanym rozkładem $e \sim N(0, \mathbf{R}\sigma_e^2)$ gdzie **R** to macierz diagonalna zawierająca na przekątnej odwrotność efektywnej liczby córek buhaja, a σ_e^2 reprezentuje wariancję błędu.

Powyższy model został porównany z trzema prostszymi modelami zawierającymi stałe (a nie losowe, jak powyżej) efekty SNP, które są znacznie mniej wymagające obliczeniowo lecz rozpatrując każdy SNP osobno nie uwzględniają powiązań pomiędzy SNP wynikających z zaburzenia równowagi Hardyego-Weinberga pomiędzy nimi. Rozpatrywano następujące modele:

- model z efektem pojedynczego SNP: $\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, (M2)

gdzie β reprezentuje stały efekt SNP opisany macierzą wystąpień $\mathbf{X} \in \{-1,0,1\}$, a pozostałe efekty są zdefiniowane jak powyżej;

- model z efektem pojedynczego SNP oraz losowym efektem addytywnie poligenicznym: $\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_2\boldsymbol{\alpha} + \mathbf{e}$, (M3)

gdzie α reprezentuje losowy efekt addytywnie poligeniczny osobnika opisany macierzą wystąpień $\mathbf{X} \in \{-1,0,1\}$, o rozkładzie $N(0, \mathbf{A}\hat{\sigma}_\alpha^2)$ gdzie $\mathbf{A}$ to macierz spokrewnień osobników, a $\hat{\sigma}_\alpha^2$ reprezentuje estymator wariancji addytywnie poligenicznej, pozostałe efekty są zdefiniowane jak powyżej;

- model CAR score zaproponowany przez Zubera i Strimmera (2011) wykorzystujący tzw. CAR scores ω_i, zdefiniowane, jako: $\boldsymbol{\omega} = \mathbf{P}^{-1/2}\mathbf{P}_{\beta y}$, (M4)

gdzie $\mathbf{P}$ reprezentuje macierz korelacji pomiędzy SNP, a $\mathbf{P}_{\beta y}$ wektor korelacji pomiędzy SNP i $\mathbf{y}$.

We wszystkich modelach jako zmienną zależną y rozpatrywano wartości hodowlane buhajów poddane deregresji dla cech: zawartość komórek somatycznych w mleku (SCS) reprezentującą cechę o czysto poligenicznym modelu dziedziczenia, wydajności mleka (MY) i tłuszczu (FY) jako cechy o poligenicznym modelu dziedziczenia, dodatkowo uwarunkowanymi genami o dużych efektach oraz współczynnik niepowtarzalności rui (NRJ) reprezentujący cechę o bardzo silnym wpływie środowiska, z odziedziczalnością jedynie 0.02.

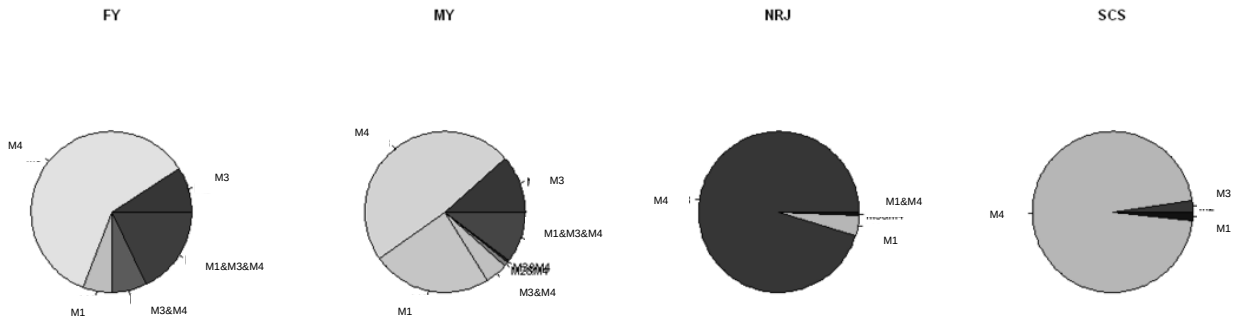
Liczba SNP wybranych, jako znaczące znacznie różni się pomiędzy poszczególnymi modelami (Tabela 1). Dla cech MY, FY oraz SCS największa liczba SNP była zawsze wybierana przez najprostszy model M2. Liczba SNP wybranych na podstawie M2 wahała się pomiędzy 2 242 (SCS) oraz 3 398 (MY) i znacznie przewyższała liczby znaczących polimorfizmów wybranych na podstawie pozostałych modeli, między którymi występowały jedynie nieznaczne różnice w ilości wybranych SNP. Na podstawie M1 wybierano zawsze bardzo zbliżoną liczbę SNP, niezależnie od analizowanej cechy. Wynika to z fakty, że procedurę wyboru SNP w modelu 1 jest oparta na założeniu, że poszczególne estymatory efektów SNP z wektora **q** o rozkładzie N(0,1) – jednakowym dla wszystkich cech.

Tabela 1

Cecha	M1	M2	M3	M4
wydajność tłuszczu	182	2435	72	49
wydajność mleka	153	3398	66	86
liczba komórek somatycznych	163	2242	4	3
współczynnik niepowtarzalności rui	125	0	0	6

Z wyjątkiem modelu M1, dla cech o różnym modelu dziedziczenia uzyskano różne liczby znaczących SNP. Najwięcej znaczących SNP zostało wybranych dla MY oraz FY. Najwięcej z tych polimorfizmów jest zlokalizowanych na chromosomie 14, zawierającym gen DGAT1 o bardzo dużym wpływie na zmienność tych cech. Dla SCS zidentyfikowano znacznie mniejszą liczbę znaczących polimorfizmów. Największe różnice pomiędzy modelami zaobserwowano w kontekście NRJ – na podstawie wyników M2 i M3 obserwowano brak znaczących SNP, sześć polimorfizmów było wybranych przez M4, a model M1 zidentyfikował 125 SNP. Rysunek 1 obrazuje liczbę znaczących polimorfizmów zidentyfikowanych przez modele M1, M3 i M4 dla poszczególnych cech. Niestety stosunkowo nieliczne SNP były wybierane przez wszystkie trzy modele, co wyraźnie sugeruje ostrożność w interpretacji wyników analiz asocjacyjnych dla cech o złożonym modelu dziedziczenia w oparciu o wyniki pojedynczego modelu statystycznego – co jest bardzo częstą praktyką obserwowaną w literaturze.

Rysunek 1



.