Dr hab. prof. IGHZ Mariusz Pierzchała
Instytut Genetyki i Hodowli zwierząt
Polskiej Akademii Nauk
ul. Postepu 36A 05-552 Jastrzębiec

Recenzja Rozprawy doktorskiej mgr Magdy Mielczarek pt.: „The genome-wide distribution of copy number variations in various breeds of domestic cattle (Bos taurus Linnaeus, 1758) based on the next-generation sequencing data ". Wykonanej pod kierunkiem prof. dr hab. Joanny Szydy.

Rozwój nowoczesnej genomiki obejmującej wielkoskalowe analizy zmienności genetycznej jest ściśle związany z badaniami genomu ludzkiego. Intensywne prace prowadzone w celu poznania pełnej sekwencji genomu ludzkiego ukończone w 2003, przyczyniły się do rozszerzenia badań dotyczących nie tylko skali ale i charakteru zmienności genetycznej, w tym analiz wpływu strukturalnego polimorfizmu na różne cechy fenotypowe u ludzi i zwierząt. Wyniki wielu badań wskazują, że główną przyczyną zmienności fenotypowej jest nie tylko polimorfizm pojedynczych nukleotydów (SNP). Analizy asocjacyjne z wykorzystaniem zidentyfikowanych polimorfizmów typu SNP pozwalają wytłumaczyć jedynie część szacowanej zmienności genetycznej. Niejednokrotnie mimo wykrycia asocjacji markerów typu SNP z cechą fenotypową, nie udało się zidentyfikować polimorfizmów funkcjonalnych. Nowe podejście do badań zmienności genetycznej, zapoczątkowane w 2004 roku w badaniach ludzkiego genomu, dowiodło powszechne występowanie licznych dużych zmian strukturalnych. Obserwowane zmiany w porównywanych genomach wykazały zróżnicowaną częstość występowania pewnych segmentów genomu. Zwiększenie lub zmniejszenie liczby kopii obejmowało fragmenty DNA, których wielkość wynosiła od około 500 pz. do nawet 1 mln pz.. Polimorfizm ten, określono mianem zmienności liczby kopii (ang. copy number variation, CNV).

Obecnie uważa się, że zmienność liczby kopii (CNV) jest istotną składową ogólnej zmienności genetycznej u zwierząt. Głównym mechanizmem biologicznym, w świetle obecnej wiedzy, decydującym o zmienności liczby kopii (CNV) jest proces rekombinacji oraz replikacji DNA. W pewnym stopniu odpowiedzialne za CNV są też samorodne zmiany w sekwencji genomu, czyli tzw. mutacje *de novo*. Uważa się, że zmienna liczba kopii CNV wpływając na cechy fenotypowe miała również istotne znaczenie w procesie ewolucji organizmów i ich lepszej adaptacji do środowiska.

Jednakże w pewnych przypadkach zanotowano, że CNV może też być przyczyną zaburzeń w funkcjonowaniu genomu, a w konsekwencji zakłócać przebieg podstawowych procesów komórkowych. Liczne wyniki badań wykazały, że zwiększenie bądź zmniejszenie liczby kopii jest czynnikiem zwiększającym ryzyko, bądź bezpośrednio odpowiedzialnym za występowanie szeregu chorób genetycznych u ludzi m.in. chorób neurodegeneracyjnych takich jak, choroba Huntingtona, Parkinsona oraz Alzheimera, jak i za złożone schorzenia neurologiczne prowadzące do rozwoju autyzmu czy schizofrenii.

Ponadto, wyniki najnowszych badań pozwoliły zdefiniować wiele polimorfizmów typu CNV związanych ze złożonymi chorobami układu sercowo-naczyniowego, chorobami autoimmunologicznymi, metabolicznymi oraz podatnością na zakażenia, i ryzykiem wystąpienia chorób nowotworowych.

Dlatego też, dla pełniejszego zdefiniowania podłoża wielu chorób tak ważne są kompleksowe, wielkoskalowe badania zmienności genetycznej. Jednoczesne badania zmian strukturalnych typu CNV w połączeniu z analizą polimorfizmów typu SNP w całym w genomie, zwiększają szansę na określenie ścisłej korelacji pomiędzy genotypem i fenotypem, a w konsekwencji identyfikacji zmian strukturalnych genomu prowadzących do zaburzeń w funkcjonowaniu organizmu i utraty zdrowia.

W procesie wykrywania polimorfizmu CNV stosowane są głównie technologie mikromacierzowe o wysokiej rozdzielczości. W ostatnich latach jednak, coraz częściej wykorzystuje się technologie sekwencjonowania nowej generacji (NGS), jako dokładniejsze a zarazem o wiele bardziej informatywne. Identyfikacja wielu odkrytych, specyficznych CNV pozwalających określić prawdopodobieństwo wystąpienia określonego schorzenia u ludzi przeszła już do rutynowej diagnostyki.

Coraz więcej jest też wyników badań wskazujących na to, że zmienność liczby kopii (CNV) jest również istotnym źródłem zmienności cech produkcyjnych zwierząt, nie mniej ważnym dla końcowego fenotypu niż wpływ polimorfizmów pojedynczych nukleotydów (SNP) analizowanych w skali genomowej.

W dotychczasowych badaniach dotyczących zmienności genetycznej w genomie bydlęcym stwierdzono, że zmiana liczby kopii (CNV) wpływa na cechy wzrostu, a co za tym idzie może odgrywać kluczową rolę w selekcji i produkcji bydła mięsnego. Dane literaturowe wskazują niejednokrotnie na zbieżne położenie określonych polimorfizmów CNV ze zmapowanymi wcześniej u bydła regionami genomu zawierającymi loci cech ilościowych (QTL), m.in. odpowiedzialnymi za rozwój tkanki mięśniowej np.: CNV100, KCNJ12. Stąd płynie wniosek, że niektóre CNV mogą być kandydatami funkcjonalnych loci warunkujących szybszy rozwoju mięśni u bydła. Szerokie analizy na poziomie genomowym przyczyniły się do ujawnienia również specyficznych wariantów CNV związanych z cechami warunkującymi produkcyjność krów ras mlecznych m.in. rasy holsztyńskiej, czy rasy jersey. Analizy porównawcze genomów rożnych ras bydła dostarczają dowodów, że CNV są istotnie związane z cechami decydującymi o typie użytkowym bydła, zwłaszcza różnicami w przebiegu szlaków metabolicznych u bydła mięsnego i mlecznego. Stały rozwój nowoczesnych technologii wielkoskalowych umożliwia obecnie szeroką identyfikację polimorfizmu typu CNV w skali całego genomu, a co za tym idzie lepszą charakterystykę genetycznego podłoża złożonych cech produkcyjnych bydła i innych gatunków zwierząt gospodarskich.

Pani mgr Magdalena Mielczarek podjęła się w pracy kompleksowej analizy zmienności liczby kopii (CNV) w genomie bydła (*Bos taurus)*.

Przedstawiona do oceny rozprawa doktorska mgr Magdy Mielczarek jest monografią w języku angielskim, w skład której wchodzi 5 głównych rozdziałów (Wprowadzenie, Materiał, Metody, Wyniki i Dyskusja). Praca liczy 65 stron tekstu, w tym jednostronicowe streszczenie w języku polskim, 83 pozycje

literaturowe, 8 stron wprowadzenia obejmującego przegląd literatury, 10 stron opisu materiału i zastosowanych metod badawczych, 10 stron opisu wyników oraz 4,5 strony dyskusji wraz z podsumowaniem.

Od strony formalnej opracowanie spełnia wymagania stawiane rozprawie doktorskiej.

Cele badawcze założone w rozprawie są jasno sformułowane, przedstawiony sposób realizacji badań od strony metodycznej został właściwie udokumentowany odpowiednim odniesieniem do danych literaturowych ze wskazanego obszaru wiedzy.

Ocena Merytoryczna

1. ***Oryginalność tematyki i wartość poznawcza pracy***

Uzasadnienie i omówienie podjętego zagadnienia zostało przedstawione w rozdziale *Introduction*. – wprowadzenie. W rozdziale tym Doktorantka scharakteryzowała zagadnienia dotyczące stosowania technologii sekwencjonowania nowej generacji NGS, wskazując główne zalety stosowania technologii NGS w analizach zmienności genetycznej na poziomie genomu. Szczególną uwagę poświęciła charakterystyce zmienności genetycznej, definiowanej jako zmienna liczba kopii CNV. Autorka szczegółowo omówiła metody detekcji polimorfizmu CNV, wskazując na zalety stosowania sekwencjonowania nowej generacji w identyfikacji nie tylko powszechnych, ale i w odkrywaniu nowych, unikatowych zmian typu CNV. Sekwencjonowanie nowej generacji pociąga za sobą konieczność zastosowania zaawansowanych metod bioinformatycznych, uwzględniających odpowiednie strategie i algorytmy analizowania genomu. Opisując je Doktorantka skupiła się na przedstawieniu pięciu metod bioinformatycznych stosowanych w identyfikacji CNV t.j.: (1)Read pair (2) Split read (3) Read depth, (4) *de novo asembly* oraz (5) łącznej kombinacji wymienionych metod. Autorka klarownie scharakteryzowała zalety i wady każdej ze stosowanych metod w odniesieniu do identyfikacji CNV w zależności od długości fragmentów oraz skuteczności identyfikacji sekwencji unikatowych.

2. Zdefiniowanie celów/ hipotez badawczych

Prezentując uzasadnienie podjętego tematu Autorka wskazuje na wciąż ograniczona liczbę badań obejmujących analizy CNV w odniesieniu do genomu bydła z zastosowaniem technologii sekwencjonowania nowej generacji NGS. Zastosowanie tej technologii w odróżnieniu od metod opartych o analizy mikromacierzowe zwiększa szansę na identyfikację nowych unikatowych polimorfizmów strukturalnych typu CNV. Ponadto, dostarcza obszerniejszej informacji dotyczącej osobniczego zróżnicowania genetycznego na poziomie genomu. Głównym celem podjętym w pracy jest wskazanie jakie są główne różnice miedzy rasami dotyczące polimorfizmu strukturalnego CNV.

3. Zakres i metody badań

Material- Materiał do badań

Materiałem do badań była baza danych sekwencji genomowych 155 osobników należących do 13 ras bydła. Jednakże, poza pięcioma, pozostałe rasy były reprezentowane przez pojedyncze osobniki. O ile w przypadku ras reprezentowanych przez kilkanaście, czy kilkadziesiąt osobników można zdefiniować specyficzną rasowo zmienną liczbę kopii (CNV), to w przypadku pozostałych 8 takowa identyfikacja, specyficznych CNV dla wybranej rasy, była niemożliwa. Dla osobników ras mniej licznie reprezentowanych

można było natomiast określić stopień podobieństwa/dystansu genetycznego, co w jakimś stopniu uzasadniałoby wykorzystywanie tych ras w prezentowanej dysertacji.

Stąd rodzi się pytanie, czy zasadnym jest pisanie, że celem była analiza osobników należących do 13 ras, albowiem analiza statystyczna zmienności CNV objęła tylko 5 ras. Pozostałe pojedyncze osobniki z różnych rasy są tłem, których udział w opracowaniu dysertacji nie został wystarczająco jasno zdefiniowany.

Dodatkowa uwaga dotyczy opisu ras bydła, który jest trochę zbyt ogólnikowy, wskazujący głównie na geograficzną lokalizację danej rasy. Uważam, że warto byłoby się pokusić o dokładniejszą charakterystykę ras, wskazując jaki kierunek selekcji w wybranych rasach był dominujący i jaki jest obecnie ich status użytkowy, wskazując jaka jest dla tych ras średnia wydajność produkcji mleka czy mięsność. Szersza charakterystyka ras, pozwoliłaby lepiej określić, czy chociażby zasygnalizować zależność między zmiennością genetyczną a zmiennością fenotypową analizowanych osobników różnych ras.

Brakuje też w mojej opinii informacji na temat pokrewieństwa miedzy osobnikami, czy takowe było czy osobniki były niespokrewnione. Stopień spokrewnienia wpływa na zmienność genetyczną miedzy osobnikami w rasie. W przypadku braku informacji rodowodowych myślę, że interesującym rozwiązaniem byłoby określenie stopnia podobieństwa genetycznego w obrębie rasy na podstawie polimorfizmów SNPs.

Methods – metody.

W rozdziale dotyczącym metod badawczych Autorka scharakteryzowała narzędzia bioinformatyczne oraz strategie zastosowane w analizach dopasowywania odczytów DNA z sekwencjonowania NGS poszczególnych osobników do genomu referencyjnego. Na podkreślenie zasługuje opracowanie 8 skryptów – „ bash"tzw. plików wsadowych dla powłoki systemu UNIX. Opracowane skrypty świadczą o znajomości narzędzi bioinformaycznych oraz umiejętności reorganizacji zbiorów genomowych baz danych. Opracowane skrypty pozwoliły „zautomatyzować" pracę począwszy od grupowania odczytów sekwencji uzyskiwanych w procesie sekwencjonowania NGS, przez dopasowywanie odczytów do sekwencji referencyjnej, ich porządkowanie, ocenę jakości uzyskiwanych odczytów oraz selekcję odpowiednich osobników, po skrypty "bash" przygotowujące dane sekwencji genomowych do programów wykorzystywanych w detekcji CNV: CNVnator i Pindel.

W dalszej części rozdziału Doktorantka szczegółowo opisuje metodę identyfikacji CNV w zależności stosowanego programu. Autorka wykazała się właściwym krytycznym podejściem do uzyskiwanych wyników przy użyciu wymienionych programów, przeprowadzając pracochłonną edycję zidentyfikowanych CNV.

Odpowiednio zweryfikowane wyniki posłużyły w dalszym etapie do przeprowadzenia analizy funkcjonalnej CNV, oraz klasyfikacji wariantów CNV pod względem lokalizacji w genomie, w regionach kodujących i niekodujących geny oraz stopnia ich potencjalnego funkcjonalnego oddziaływania poprzez wpływ na kodowane transkrypty. Autorka zidentyfikowała 2 duplikacje w regionach niekodujących białek obejmujących transkrypty małego jądrowego RNA (snRNA, SNORD116) oraz rybosomalnego RNA (5s_RNA), które sklasyfikowała jako CNV o potencjalnie dużym wpływie. Jedna z duplikacji zlokalizowana została w regionie kodującym genu (ENSBTAG00000003152), który to gen jak stwierdza Autorka jest prawdopodobnie odpowiedzialny za transport białek z jądra komórkowego. Informacja wymaga powołania się

odpowiednie na dane źródłowe w piśmiennictwie. Kolejne duplikacje zostały zidentyfikowane w intronie genu RHOBTB2 , genu beta defensyny (ENSBTAG00000033545) oraz genu kinazy proteinowej PAK3. Podczas gdy, pozostałe 13 duplikacji zidentyfikowane zostały w intronach genów, w większości pokrywały się z regionami duplikacji CNV bazy danych DGVa. Natomiast w przypadku delecji jedna dotyczyła regionu genu pozostałe 19 w obszarach między genami. Podobnie jak w przypadku duplikacji lokalizacja większości analizowanych 17 z 20 pokrywała się z bazą danych DGV.

W przeprowadzonej analizie statystycznej zastosowane zostały odpowiednie testy nieparametryczne celem określenia rozkładu identyfikowanych CNV, zarówno pod względem liczby jak i wielkości sekwencji DNA. Ponadto oszacowany został stopień indywidualnego zróżnicowania udziału pokrycia genomu przez sekwencje CNV w obrębie rasy.

4 Zakres i wyniki zrealizowanych badań

Opisując wyniki Autorka określiła stopień dopasowania odczytów do sekwencji referencyjnej genomu bydła oraz stopień pokrycia genomu. Uzyskane wyraźne zróżnicowanie pokrycia genomu od 2 do 28, może świadczyć o zróżnicowanej jakości materiału genetycznego wykorzystywanego do sekwencjonowania NGS. Kryterium to zostało wykorzystane w procesie weryfikacji i eliminacji 9 osobników, brakuje jednak informacji na temat ich rasy.

Autorka w pracy przedstawiła ogólna liczbę zidentyfikowanych CNV, liczbę zwalidowanych CNV oraz ogólny udział funkcjonalnych adnotacji, wskazując średnio ok. dwukrotnie większą liczbę CNV zlokalizowanych w regionach niekodujących. Podczas, gdy w przypadku delecji rozkład ten miedzy rasami nie różnił się to w przypadku duplikacji u rasy Fleckvieh i norweskiej czerwonej udział CNV w regionach kodujących był znacząco większy. W tym przypadku wypadałoby skonfrontować uzyskane wyniki cech produkcyjnych tych dwóch ras względem pozostałych ras. Dla lepszego zobrazowania osobniczego zróżnicowania w obrębie rasy wskazanym byłoby uwzględnienie danych rodowodowych, jeśli takowe są dostępne. W stosunku do przeprowadzonych analiz wskazujących najbardziej powszechne identyfikowane CNV brakuje wyjaśnienia, jakie zostało przyjęte kryterium częstości obserwowanych CNV. Wskazanym byłoby zamieszczenie tabeli przedstawiających zestawienie rasowo specyficznych CNV, takie opracowanie byłoby cennym uzupełnieniem i pewnym punktem wyjścia do dalszych badań w celu identyfikacji polimorfizmów o charakterze funkcjonalnym.

Interpretacja wyników

W dyskusji Doktorantka w sposób syntetyczny omówiła uzyskane wyniki, wskazując relatywnie wysoką jakość odczytów DNA uzyskanych w procesie sekwencjonowania NGS w stosunku do danych literaturowych. Główny nacisk w dyskusji doktorantka położyła na proces identyfikacji CNV. Odnosząc się do danych literaturowych, stwierdziła, że kluczowym dla potwierdzenia rzeczywistych zidentyfikowanych wyników jest ich dokładna edycja. Przeprowadzona edycja była kryterium odrzucenia wariantów, które nie mieściły się w zakresie 50 pz. do 5mln pz. Doktorantka stwierdza, że uzyskane wyniki wskazały większą zmienność w przypadku zastosowania programu CNVnator w odniesieniu do duplikacji w porównaniu z programem Pindel podczas gdy, w przypadku delecji sytuacja jest odwrotna. To zestawienie różnic miedzy

programami wymagałoby bardziej szczegółowego omówienia, wskazującego w jakim zakresie identyfikowane CNV są wspólne, a w jakim specyficzne dla danego programu. Wiarygodność wyników uzyskanych przez Doktorantkę potwierdza zestawienie z bazą danych obejmujących strukturalną zmienność w genomie bydła (DGV), uzyskiwane wartości dla duplikacji 44.1% i delecji 44.6% są w dużej mierze zbieżne z danymi zdeponowanymi w wymienionej bazie. Brakuje jednak doprecyzowania, wyjaśnienia czy są to zbiorcze wyniki identyfikacji CNV? Czy zidentyfikowanych jednym z wymienionych programów? Ponadto, jeżeli taka analiza została przeprowadzona, na co wskazuje przedstawione wnioskowanie, to metodyka takich porównań powinna też zostać opisana w materiałach i metodach oraz przedstawiona w wynikach. Interesującym wnioskiem z pracy jest stwierdzenie, że najbardziej wspólne rasowo specyficzne CNV były zlokalizowane w obrębie genów, i są one odpowiedzialne za zmienność zwierząt w obrębie rasy. Wnioskując ze streszczenia pracy zapewne chodzi o rasę Fleckvieh, jednak w głównym opisie dyskusji można mieć wątpliwości czy nie dotyczy to ogółu ras. Omawiana przez Doktorantkę rasowo specyficzna lokalizacja CNV utwierdza mnie w przekonaniu, że załączenie tabeli z najbardziej rasowo specyficznymi polimorfizmami CVN, wskazując ich lokalizację w genomie wraz z odniesieniem do zidentyfikowanych loci genowych, jeśli takowe są znane, byłoby bardzo cennym uzupełnieniem niniejszego opracowania.

W odniesieniu do języka angielskiego praca jest napisana klarownie, niemniej jednak wymaga drobnych korekt językowych, np.:

str 16 was shown on Figure, preferowane jest in Figure.

str 21 należy poprawić nazwę testu Kurskal –Wallis na Kruskal-Wallis,

str 25 The shortest detected duplications was – powinno być "were"

str 35 The non-overlapping criteria of CNV calling in CNVnator involves powinno być involve.

Ponadto, wielu miejscach tekstu należy poprawić formę angielskiego z amerykańskiej na brytyjską np.: characterized, emphasized, revolutionized.

Przedstawione uwagi są natury ogólnej i nie umniejszają wysokiej oceny merytorycznej badań przedstawionych w dysertacji, które uważam za nowatorskie i wartościowe.

Podsumowując recenzję stwierdzam, że niniejsza praca doktorska spełnia wymogi art.13 ust.1 Ustawy z dn. 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. nr 56 poz. 595 z 2003 r. z późn. zm.) i wnoszę o dopuszczeni mgr Magdaleny Mielczarek do dalszych etapów przewodu doktorskiego. Jednocześnie biorąc pod uwagę nowatorski charakter wnoszący wymierne walory poznawcze, pracochłonność oraz czasochłonność przeprowadzonych badań, składam wniosek do wysokiej Rady Wydziału Biologii i Ochrony Środowiska UMK w Toruniu o wyróżnienie niniejszej pracy doktorskiej.

Jastrzębiec 24 Sierpnia 2016r.

dr. hab. Mariusz Pierzchała prof. IGHZ PAN