Poznań, 2.11.2023

prof. dr hab. Paweł Krajewski

Instytut Genetyki Roślin PAN w Poznaniu

Ocena pracy doktorskiej mgr. inż Michała Wojcieszka

„Składanie, adnotacja i porównanie genomów mutantów chemicznych ogórka (*Cucumis sativus* L.)"

Niniejsza ocena została wykonana na podstawie zlecenia Dyrektora Instytutu Nauk Ogrodniczych SGGW zgodnego z uchwałą o powołaniu recenzentów w przewodzie doktorskim mgr. inż. Michała Wojcieszka oraz art. 13 ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. z 2017 r., poz. 1789) w związku z art. 179 Ustawy z dnia 3 lipca 2018 r. Przepisy wprowadzające ustawę – Prawo o szkolnictwie wyższym i nauce (Dz.U. z 2018 r., poz. 1669).

1. Zakres pracy

Praca doktorska mgr. inż. Michała Wojcieszka opisuje wyniki analizy porównawczej genomów trzech mutantów chemicznych ogórka przeprowadzonej za pomocą metod sekwencjonowania nowej generacji. Badane mutanty pochodziły z kolekcji będącej w posiadaniu Katedry Genetyki, Hodowli i Biotechnologii Roślin Instytutu Biologii SGGW, otrzymanej wcześniej w wyniku traktowania etylenoiminą linii wyjściowej B10. Celem badań było znalezienie różnic pomiędzy genomami badanych form i adnotacja tych różnic pod względem strukturalnym i funkcjonalnym. Hipoteza badawcza zakładała nierównomierną podatność poszczególnych regionów genomu na mutacje chemiczne. Badania obejmowały też weryfikację własności fenotypowych mutantów w odniesieniu do ich charakterystyki podanej w opisie kolekcji.

2. Opis formalny pracy

Praca składa się z siedmiu rozdziałów, o układzie typowym dla rozpraw doktorskich. Wstęp zawiera opis terminów podstawowych dla rozprawy takich jak: typy mutacji, rodzaje stosowanych mutagenów i sposoby ich działania, dostępne mutanty ogórka (w tym pochodzące od linii B10) oraz metody sekwencjonowania DNA i sposoby wnioskowania o genomie na podstawie uzyskanych danych. Przedstawiona jest hipoteza badawcza oraz cele szczegółowe pracy. W części „Materiał i metody" opisano sposoby izolacji DNA, przygotowania bibliotek do sekwencjonowania, wstępnego przetwarzania danych zmierzającego do oceny i poprawy ich jakości, składania genomu oraz lokalizacji, weryfikacji i adnotacji wariantów. Po przedstawieniu Wyników następuje Dyskusja, zakończona sformułowaniem ośmiu Wniosków. Cała praca liczy 119 stron.

3. Ocena sposobu realizacji

Rozwijane obecnie metody sekwencjonowania genomów pozwalają na szeroko zakrojone badania form roślin uprawnych znajdujących się w kolekcjach jednostek badawczych, zasobach firm hodowlanych czy też bankach genów. Pozwala to na wyjaśnienie podłoża genetycznego interesujących cech użytkowych. Recenzowana rozprawa opisuje badania wpisujące się w ten nurt.

Do badań wybrano trzy mutanty ogórka: Sh, W19 i WSK charakteryzujące się interesującymi fenotypami w zakresie architektury i rozwoju roślin. W pierwszym etapie prac metodami standardowymi izolowano i sekwencjonowano biblioteki fragmentów DNA pochodzące z roślin uprawianych w dwóch wariantach: polowym i szklarniowym. Uzyskane dane sekwencyjne oceniano pod względem jakości i poddawano usuwaniu adapterów oraz korekcji ze względu na rzadko występujące k-mery.

Otrzymane dane posłużyły do złożenia genomów trzech mutantów narzędziem SOAPdenovo. Przedstawiona jest analiza porównawcza trzech otrzymanych genomów pod względem liczby i struktury kontigów oraz skafoldów. W genomach określono położenie elementów strukturalnych takich jak sekwencje powtarzalne, regiony o niskiej złożoności, geny, transpozony i in. Sekwencje genów zidentyfikowano programem BRAKER i zadnotowano funkcjonalnie przez mapowanie w bazie genów linii referencyjnej. W tej analizie wykorzystano dane sekwencyjne reprezentujące transkryptom ogórka, pobrane z bazy danych. Białka zadnotowano funkcjonalnie narzędziami EggNOG-mapper, InterProScan i GOSlim.

Następnie znaleziono w trzech otrzymanych genomach pozycje różniące się od genomu referencyjnego linii B10, czyli tzw. warianty. Użyto w tym celu mapowania odczytów DNA w genomie referencyjnym oraz narzędzi FreeBayes i DeepVariant. Prawdopodobne efekty wariantów w odniesieniu do kodowanych przez geny białek oceniono programem SnpEff. Losowo wybrane warianty weryfikowano przez sekwencjonowanie odpowiednio zamplifikowanych fragmentów metodą Sangera.

Przeprowadzając powyższe analizy danych sekwencyjnych otrzymano dużą liczbę wyników, które przedstawiono w postaci tabel i wykresów oraz interpretowano statystycznie i biologicznie. Motywem przewodnim interpretacji jest porównanie wyników uzyskanych dla trzech badanych mutantów w zakresie występowania elementów strukturalnych i wariantów. Przedstawione wyniki opisują w zasadzie całość problemów, jakie można poruszyć dysponując danymi takimi, jakie uzyskano przygotowując rozprawę.

Dyskusja wyników dotyczy, w pierwszej części, porównania otrzymanych wyników z wynikami autorów opisujących własności materiałów otrzymanych z linii B10 metodami innymi niż mutageneza chemiczna. Stwierdzono, że mutanty analizowane w pracy charakteryzowały się większą częstością wariantów typu SNP i MNP niż formy otrzymane przez transgenezę i generowanie zmienności somaklonalnej. W drugiej części dyskusji omawiane są poszczególne badane formy w odniesieniu do reprezentowanych fenotypów, również w porównaniu do wiedzy o determinacji tych fenotypów w innych gatunkach roślin.

Należy stwierdzić, że zadanie opisania sekwencji genomowych trzech wybranych mutantów ogórka zostało w pracy zrealizowane. Badania przeprowadzono za pomocą prawidłowo dobranych metod laboratoryjnych i narzędzi obliczeniowych. Jakkolwiek są to metody standardowe, stosowane obecnie w wielu projektach badawczych i hodowlanych, to jednak każde nowe zastosowanie stawia szereg wyzwań wynikających z charakterystyki dostępnego materiału biologicznego, ograniczeń na głębokość sekwencjonowania i dostępności danych wymaganych przez poszczególne narzędzia obliczeniowe. Przedstawiona dyskusja biologicznych aspektów uzyskanych wyników jest oryginalna i dostarcza nowej wiedzy o możliwościach uzyskania zmienności genetycznej za pomocą mutagenezy.

4. Uwagi i pytania

1. W doświadczeniach zastosowano dwa warianty uprawy roślin: polowy i szklarniowy. Informację na temat zgodności fenotypu roślin w tych dwóch wariantach podano tylko dla linii Sh. Ponadto brakuje w pracy wyraźnego wniosku podsumowującego korzyści z użycia dwóch wariantów dla analizy genomów.

2. W części wynikowej wykorzystywane są wyniki sekwencjonowania genomu formy B10. Nie jest jednak jasne, czy sekwencjonowanie i składanie genomu tej formy było przeprowadzone dokładnie takimi samymi metodami jakie zastosowano dla trzech mutantów.

3. W opisie wariantów genomowych stosowane jest słowo „predykcja", które nie jest właściwe, gdyż oznacza proces wnioskowania o zdarzeniach przyszłych (dokładnie, o wartościach zmiennych losowych obserwowanych w przyszłości). Powinno się używać słowa „lokalizacja" lub „identyfikacja". Słowa „predykcja" można użyć w odniesieniu do wpływu wariantów genomowych na kodowane białka lub na fenotypy.

4. Rozkład zidentyfikowanych polimorfizmów w różnych strukturach genomu winien być interpretowany względem udziału ilościowego tych struktur w genomie, zaś liczba wariantów zidentyfikowanych w poszczególnych formach znajdujących się np. w eksonach – względem ogólnej liczby wariantów (np. str. 63: 64% SNV w regionach międzygenowych – lecz jaki jest udział tych regionów w genomie?, liczba polimorfizmów w eksonach największa dla W19 – może tylko dlatego, że ogólnie najwięcej było wariantów w W19).

5. Błędnie jest używane słowo „wpływ" w adnotacji funkcjonalnej wariantów. Np. na str. 65 napisano „Tabela 15. Zestawienie liczby genów będących pod wpływem SNV"; na str. 80 „mutacje o wysokim wpływie na geny". SNV różnicują sekwencję genów i mogą mieć wpływ na wytworzone przez gen transkrypty lub białka.

6. Wniosek 5 dotyczący „przełożenia liczby polimorfizmów na ilość zmian fenotypowych". Jak oszacowano "ilość zmian fenotypowych"?

7. Jedną ze standardowych metod oceny kompletności genomu lub transkryptomu uzyskanego przez sekwencjonowanie jest BUSCO (Benchmarking Universal Single-Copy Orthologs). Czy wykonano taką analizę?

8. Jaki procent danych odrzucono stosując korektę odczytów NGS poprzez filtrowanie rzadkich k-merów (str. 50)? Czy ta korekta miała duży wpływ na jakość uzyskanych sekwencji genomowych?

5. Uwagi stylistyczne i terminologiczne

Poniższe uwagi wynikają w dużej części z licznych usterek stylistycznych i braków w zakresie precyzyjnego stosowania terminologii oraz niedostatecznej weryfikacji tekstu w wersji ostatecznej:

- Zamiast słowa „figura" powinno się używać słowa „rycina".
- str. 21: Genom nie jest „zlokalizowany na siedmiu chromosomach" – składa się z siedmiu chromosomów.
- str. 29: Dane z sekwencjonowania nie zawierają informacji o „lokalizacji na nici DNA" – tę informację uzyskuje się poprzez mapowanie odczytów.
- str. 32: Winno być „... homologii pomiędzy badaną sekwencją ..." – nie można badać „nieznanej sekwencji".
- str. 34: Czułość definiuje się jako „stosunek liczb zdarzeń", a nie jako „stosunek zdarzeń".
- Rycina 4 prezentuje to samo, co Rycina 2 (z dokładnością do długości odczytu i insertu).
- str. 39: Dane w formacie FASTQ zawierają ocenę jakości odczytu każdego nukleotydu, nie „jakości nukleotydu".
- str. 41: Podany odnośnik do bazy danych csgenome.sggw.pl nie działa; strona Konsorcjum nie zawiera żadnych informacji.
- str. 43: Ściśle mówiąc nie ma czegoś takiego jak „teoria prawdopodobieństwa Bayesa" – istnieje twierdzenie Bayesa i wnioskowanie bayesowskie.
- str. 44: Bcftools i Samtools to różne zestawy narzędzi.
- str. 46: Fragment tekstu wyglądający na niedopracowany, pozostawiony z wersji roboczej.
- str. 51: Linia W19 miała 30 077 kontigów, co „wynosi ok. 4,5%" – procent czego?
- str. 52: Winno być „Minimalna długość genu mieści się między 55 a 57 nt" – nie mediana.
- str. 71: Terminologia „6 terminów GO zawierało geny ..." jest niewłaściwa. Można powiedzieć, że „6 terminów GO występowało w adnotacji genów ...". Podobna uwaga dotyczy sformułowania „Najmniej genów otrzymało wspólny termin ...".
- str. 73: Podpis Ryciny 18, winno być „skupisk genów ortologicznych".
- str. 90: W zestawieniu literatury brakuje pozycji Kaufmann i Lower, 1976) cytowanej na str. 90.
- W bardzo wielu miejscach zdania zaczynają się od małej litery.

5. Konkluzja

Recenzowana praca zawiera nowe wyniki naukowe poszerzające wiedzę o genomie ogórka (*Cucumis sativus* L.) i jego reprezentacji w istniejących kolekcjach form o interesujących własnościach fenotypowych. Podaje także nowe informacje o możliwościach generowania zmienności genetycznej za pomocą mutacji chemicznych i porównaniu tej metody z innymi stosowanymi w badaniach

genetycznych. Analizowane mutanty zostały uzyskane metodami, które są dopuszczalne dla materiałów wykorzystywanych do uzyskiwania nowych odmian roślin, tak więc mogą znaleźć zastosowanie praktyczne. W takim przypadku, szeroka wiedza o polimorfiźmie genetycznym warunkującym cechy użytkowe będzie bardzo wartościowa,

Praca nad rozprawą wymagała opanowania szeregu technik eksperymentalnych oraz wielu procedur analizy danych. Doktorant wykazał się umiejętnością prowadzenia badań i interpretacji ich wyników.

W związku z tym uważam, że praca spełnia wymagania stawiane przed rozprawami doktorskimi i wnoszę o dopuszczenie mgr. inż. Michała Wojcieszka do dalszych etapów przewodu doktorskiego.

P. Krzewski