Autoreferat

Opis kariery zawodowej i naukowej, z uwzględnieniem opisu osiągnięcia naukowego, o którym mowa w art. 219 ust. 1 pkt. 2 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (dz. U. Z 2018 r. Poz. 1668 ze zm.)

A. Dane osobowe

imię i nazwisko: Agnieszka Żmieńko

B. Posiadane dyplomy i stopnie naukowe

Tytuł zawodowy magistra

2000: magister biotechnologii; Wydział Biologii Uniwersytetu im. Adama Mickiewicza w Poznaniu; tytuł pracy magisterskiej: „Analiza ekspresji genów hydrolazy S-adenozylohomocysteiny, L-asparaginazy oraz ENOD40 z *Lupinus luteus* we wczesnych etapach rozwoju brodawki korzeniowej"; promotor: prof. dr hab. Andrzej B. Legocki

Stopień doktora

2006: doktor nauk chemicznych w zakresie biochemii; Instytut Chemii Bioorganicznej Polskiej Akademii Nauk w Poznaniu; tytuł rozprawy doktorskiej: „Profilowanie ekspresji genów łubinu wąskolistnego w badaniach nad symbiotycznym wiązaniem azotu"; promotor: prof. dr hab. Andrzej B. Legocki

C. Informacja o dotychczasowym zatrudnieniu w jednostkach naukowych

Podstawowe miejsce zatrudnienia – Instytut Chemii Bioorganicznej PAN w Poznaniu:

07.01.2016 - obecnie; starszy specjalista-biolog; Pracownia Mikromacierzy i Głębokiego Sekwencjonowania / Pracownia Genomiki (po przekształceniu)
01.06.2013 - 06.01.2016; adiunkt; Zakład Biologii Molekularnej i Systemowej
01.04.2007 - 31.05.2013; adiunkt; Centrum Doskonałości CENAT
01.01.2007 - 31.03.2007; asystent; Centrum Doskonałości CENAT
01.05.2001 - 31.12.2006; stypendysta – doktorant; Pracownia Biologii Molekularnej Roślin
01.09.2000 - 30.04.2001; asystent-doktorant; Pracownia Biologii Molekularnej Roślin

Dodatkowe miejsce zatrudnienia – Politechnika Poznańska:

01.03.2014 - 30.09.2020; adiunkt; Instytut Informatyki
03.01.2011 - 31.03.2012; specjalista; Instytut Informatyki

D. Staże w instytucjach naukowych

Staże przed uzyskaniem stopnia doktora

1999-2000; 6-miesięczne stypendium w ramach programu Socrates-Erasmus na Uniwerystecie Arystotelesa w Salonikach (Grecja); praca badawcza w zespole prof. Trianosa Yupsanisa (Zakład Biochemii)

E. Dodatkowe kursy i szkolenia podnoszące kwalifikacje

Umiejętności zawodowe

2015; kurs bioinformatyczny „Tworzenie biologicznych baz danych i stron internetowych", Ideas4Biology, Poznań

2015; kurs bioinformatyczny „Programowanie dla Biologów", Ideas4Biology, Poznań

2011; warsztaty bioinformatyczne „9th Poznan Summer School of Bioinformatics", Wydział Biologii, Uniwersytet im. Adama Mickiewicza, Poznań

2010; warsztaty analizy w R „Metody klasteryzacji hierarchicznej w biologii", Polskie Towarzystwo Genetyczne, Instytut Oceanografii PAN, Sopot

2007; kurs analizy danych mikromacierzowych „EMBO Practical Course on Analysis and Informatics of Microarray Data", European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, Wielka Brytania

2007; kurs analizy danych mikromacierzowych „NGFN Course in Practical DNA Microarray Analysis", Nationalen Genomforschungsnetz, Dortmund-Universtät, Dortmund, Niemcy

2002; kurs bioinformatyczny w Centrum Doskonałości w Biotechnologii Molekularnej (Instytut Biochemii i Biofizyki PAN), Warszawa

2001; kurs bioinformatyczny „Bioinformatics: Computer Methods in Molecular Biology", International Centre for Genetic Engineering and Biotechnology, Triest, Włochy

Umiejętności "miękkie"

2019; szkolenie z komunikacji w środowisku wielokulturowym „Communication in intercultural setting" Centrum Języków i Komunikacji Politechniki Poznańskiej; Poznań

2015; szkolenie z zarządzania zespołem naukowym w ramach projektu SKILLS Fundacji Nauki Polskiej; Poznań

2014; szkolenie z zarządzania projektami badawczymi w ramach projektu SKILLS; Fundacji Nauki Polskiej; Gdańsk

F. Członkostwo w międzynarodowych lub krajowych organizacjach i towarzystwach naukowych

Polskie Towarzystwo Genetyczne (PTGen) – członek od 2017
Polskie Towarzystwo Biologii Eksperymentalnej Roślin (PTBER) – członek od 2013
American Association for the Advancement of Science (AAAS) – członek w latach 2012-2014

G. Nagrody i stypendia

Nagrody i stypendia po uzyskaniu stopnia doktora

2018; nagroda Rektora Politechniki Poznańskiej za osiągnięcia organizacyjne w roku akademickim 2017/2018

2016; nagroda Rektora Politechniki Poznańskiej za osiągnięcia naukowe w roku akademickim 2015/2016

2014-2015; beneficjentka programu SKILLS Fundacji Nauki Polskiej

Nagrody i stypendia przed uzyskaniem stopnia doktora

2000; nagroda Dziekana Wydziału Biologii za wyniki podczas studiów, Uniwersytet im. A. Mickiewicza, Poznań

1999; stypendium Ministra Edukacji Narodowej za wybitne wyniki w nauce i szczególne osiągnięcia naukowe, Warszawa

1998; stypendium Ministra Edukacji Narodowej za wybitne wyniki w nauce i szczególne osiągnięcia naukowe, Warszawa

H. Osiągnięcie naukowe, o którym mowa w art. 219 ust. 1 pkt. 2 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (dz. U. Z 2018 r. Poz. 1668 ze zm.)

Tytuł osiągnięcia naukowego

IDENTYFIKACJA POLIMORFIZMU LICZBY KOPII DNA JAKO ISTOTNEGO SKŁADNIKA KSZTAŁTUJĄCEGO ZMIENNOŚĆ GENETYCZNĄ *ARABIDOPSIS THALIANA*

Publikacje wchodzące w skład osiągnięcia naukowego

wskaźniki naukometryczne:

-liczba cytowań / liczba cytowań bez cytowań własnych habilitantki w czasopismach z kolekcji Web of Science Core Collection, wg bazy Web of Science na dzień 08.09.2020;

-pięcioletni wskaźnik Impact Factor i dane rankingowe czasopism wg wykazu Journal Citation Report obowiązującego w roku opublikowania danego artykułu;

- punktacja MNiSW wg Wykazu czasopism naukowych za rok poprzedzający rok opublikowania (oraz w celach porównawczych – z uwagi na różne systemy punktacji w poszczególnych latach – punktacja MNiSW wg Wykazu czasopism naukowych i recenzowanych materiałów z konferencji międzynarodowych, stanowiącego załącznik do komunikatu Ministra Nauki i Szkolnictwa Wyższego z dnia 18.12.2019 r.)

*autor korespondencyjny

1. Żmieńko A., Samelak A., Kozłowski P., Figlerowicz M.* (2014) **„Copy number polymorphism in plant genomes”.** Theor Appl Genet., 127(1): 1-18.
DOI: https://doi.org /10.1007/s00122-013-2177-7

- 5-letni IF.2013: 3,759
- MNiSW.2013: 45 pkt (MNiSW.2019: 100 pkt)
- Kwartyl: Q1 (PLANT SCIENCES), Q2 (GENETICS & HEREDITY), Q1 (AGRONOMY), Q1 (HORTICULTURE)
- Cytowania: 97/95
- Publikacja wyróżniona jako “Highly cited in field” przez bazę bibliograficzną Web of Science.

2. Zmienko A., Samelak-Czajka A., Kozlowski P., Szymanska M., Figlerowicz M.* (2016) ***„Arabidopsis thaliana* population analysis reveals high plasticity of the genomic region spanning *MSH2, AT3G18530* and *AT3G18535* genes and provides evidence for NAHR-driven recurrent CNV events occurring in this location”.** BMC Genomics, 17:893
DOI: https://doi.org/10.1186/s12864-016-3221-1

- 5-letni IF.2015: 4,276
- MNiSW.2015: 40 pkt (MNiSW.2019: 140 pkt)
- Kwartyl: Q1 (BIOTECHNOLOGY & APPLIED MICROBIOLOGY), Q2 (GENETICS & HEREDITY)
- Cytowania: 11/9

3. Samelak-Czajka A., Marszalek-Zenczak M., Marcinkowska-Swojak M., Kozlowski P., Figlerowicz M., **Zmienko A.*** (2017) **„MLPA-based Analysis of Copy Number Variation in Plant Populations”**. Front. Plant Sci., 8:222
DOI: https://doi.org/10.3389/fpls.2017.00222

- 5-letni IF.2016: 4,672
- MNiSW.2016: 40 pkt (MNiSW.2019: 200 pkt)
- Kwartyl: Q1 (PLANT SCIENCES)
- Cytowania: 9/8

4. Zmienko A.*, Marszalek-Zenczak M., Wojciechowski P., Samelak-Czajka A., Luczak M., Kozlowski P., Karlowski W.M., Figlerowicz M.* (2020) **„AthCNV - a map of DNA copy number variations in the *Arabidopsis thaliana* genome”**. The Plant Cell, 32:1-23.
DOI: https://doi.org/10.1105/tpc.19.00640

- 5-letni IF.2019: 10,144
- MNiSW.2019: 200 pkt
- Kwartyl: Q1 (PLANT SCIENCES), Q1 (BIOCHEMISTRY & MOLECULAR BIOLOGY), Q1 (CELL BIOLOGY)
- Cytowania: 1/1
- Artykuł opatrzony komentarzem edytorskim w sekcji *In Brief*, opublikowanym w tym samym numerze czasopisma: https://doi.org/10.1105/tpc.20.00257
- Artykuł opatrzony komentarzem w sekcji *Plant Science Research Weekly* na portalu Plantae: https://plantae.org/athcnv-a-map-of-dna-copy-number-variations-in-the-arabidopsis-thaliana-genome-plant-cell/?utm_source=TrendMD&utm_medium=cpc&utm_campaign=Plantae_TrendMD_0

Łączne wskaźniki naukometryczne osiągnięcia naukowego:

- Artykuły na liście Web of Science Core Collection: 4
- Sumaryczna wartość wskaźnika IF: 22,851
- Sumaryczna liczba cytowań: 118/113

Omówienie osiągnięcia naukowego

Wprowadzenie

Opracowanie w 1977 r. przez F. Sangera techniki sekwencjonowania DNA przez syntezę (znanej dziś jako sekwencjonowanie pierwszej generacji) otworzyło drogę do poznania sekwencji genomowej wirusów, następnie bakterii, a wreszcie organizmów eukariotycznych. Dzięki wieloletnim wysiłkom międzynarodowych konsorcjów naukowych, w 2000 r. opublikowano pierwszą wersję sekwencji genomu rośliny naczyniowej – rzodkiewnika pospolitego, a w 2001 r. – pierwszą wersję sekwencji genomu człowieka. Początek XXI wieku przyniósł rozwój drugiej generacji technik sekwencjonowania (NGS, ang. *next-generation sequencing*), które umożliwiły masowe równoległe generowanie krótkich odczytów sekwencji z pofragmentowanego DNA genomowego. Takie odczyty można następnie składać metodami bioinformatycznymi w dłuższe ciągi, zwane kontigami, w procesie zwanym asemblacją lub też wykorzystać je bezpośrednio w analizach bazujących na dopasowaniu (mapowaniu) ich do sekwencji referencyjnej. Dzięki technikom NGS do dnia dzisiejszego poznano sekwencje genomowe ponad 6400 gatunków eukariotycznych (źródło: baza GenBank, https://www.ncbi.nlm.nih.gov/genome). Warto przy tym jednak zauważyć, że genomy te mają różny stopień asemblacji odczytów. Złożenie sekwencji na poziomie chromosomów jest dostępne dla mniej niż 900 gatunków, a tylko dla kilkudziesięciu z nich, o genomach nie większych niż 100 Mpz, sekwencję uważa się za kompletną. Niemniej te osiągnięcia stworzyły pole do badań porównawczych, początkowo międzygatunkowych, a w miarę realizacji kolejnych projektów sekwencjonowania – także wewnątrzgatunkowych, skupionych na opisaniu zmienności genetycznej organizmów.

Zmienność genetyczna wynika z istnienia alternatywnych wariantów danej sekwencji genomowej, przy czym w skali populacji liczba i częstość występowania poszczególnych wariantów może być różna. W przypadku genomu człowieka, zmienność genetyczna rozpatrywana jest często z punktu widzenia indywidualnych różnic między osobnikami, z uwagi na możliwe konsekwencje fenotypowe dla nosicieli danego wariantu (np. wystąpienie skojarzonej z nim choroby). Jednak warianty genetyczne wykazują również istotne zróżnicowanie międzypopulacyjne, często będące wynikiem adaptacji do określonych warunków bytowania. U roślin niezwykle istotne jest określenie różnic genetycznych pomiędzy odmianami, liniami hodowlanymi czy też naturalnymi ekotypami. Przyczynia się to do poznania ewolucji roślin, mechanizmów adaptacji do określonych warunków środowiska, a także do poszukiwania i selekcji odmian o cechach pożądanych z punktu widzenia człowieka, takich jak walory hodowlane czy odporność na patogeny. Indywidualne rośliny są traktowane wówczas jako reprezentanci konkretnych linii / odmian / ekotypów, co jest szczególnie dobrze uzasadnione w przypadku roślin samopylnych. Potomstwo takich roślin ma genomy podobne do genomu rodzica, wykazujące wysoki stopień homozygotyczności.

Poszczególne warianty genetyczne mogą obejmować dowolnie długi fragment sekwencji genomowej. Największe są zmiany obserwowane na poziomie kariotypu, polegające na nadmiarze bądź utracie jednego lub większej liczby chromosomów (w całości lub części).

Zmiany kariotypu są badane u człowieka, roślin i zwierząt od wielu lat, z uwagi na łatwość ich obserwacji i analizy różnymi metodami cytogenetycznymi. Ich występowanie zwykle jest związane z wyraźnymi efektami na poziomie fenotypu. Przykładem takiej zmienności genetycznej u człowieka może być trisomia chromosomu 21, powodująca zespół wad wrodzonych, zwany zespołem Downa, której częstość występowania u noworodków wynosi, według różnych szacunków, od 1:600 do 1:1000. Na drugim końcu skali pod względem rozmiaru wariantów genetycznych znajduje się zmienność pojedynczego nukleotydu, polegająca na występowaniu punktowych różnic w sekwencji genomu. Jednonukleotydowe substytucje występujące w populacji z częstością co najmniej 1% noszą nazwę polimorfizmów pojedynczego nukleotydu (SNP, ang. *single nucleotide polymorphism*). Zostały one dość dobrze scharakteryzowane u wielu gatunków roślin i zwierząt. W genomie człowieka SNP obserwuje się średnio co 300 pz, a u kukurydzy – co około 80 pz. Pomimo tej gęstości, zdecydowana większość SNP ma charakter neutralny. Znane są jednak warianty, szczególnie te położone w obrębie sekwencji genów kodujących białka (dla uproszczenia dalej nazywam je genami), które mogą stanowić bezpośrednią przyczynę wyraźnych zmian fenotypowych, takich jak choroby genetyczne u człowieka. Przykładem takiego SNP jest zamiana A → T w jednym z kodonów genu *HBB*, kodującego podjednostkę β hemoglobiny, prowadząca w efekcie do rozwinięcia choroby, zwanej niedokrwistością sierpowatokrwinkową.

Pomiędzy zmianami na poziomie kariotypu, a SNP zawarte jest całe spektrum długości wariantów strukturalnych, które mogą mieć charakter duplikacji, delecji, insercji, inwersji bądź translokacji. Z tego spektrum, ze względu na ograniczenia dostępnych metod, stosunkowo najpóźniej podjęto badania nad wariantami długimi, reprezentującymi zmienność liczby kopii (CNV, ang. *copy number variation*). Terminem CNV obejmuje się zmiany o charakterze niezbalansowanym, a więc modyfikujące liczbę kopii danego regionu w genomie. Warianty CNV mogą mieć zatem charakter insercji, duplikacji lub delecji, choć często współwystępują z innymi typami zmian, tworząc w efekcie złożone rearanżacje strukturalne (Redon i wsp. 2006). Typowo, za CNV uważa się warianty o długości od 1 kpz do wielu Mpz, a warianty krótsze określa się mianem indeli. Należy jednak zaznaczyć, że te granice nie są sztywne (Alkan et al. 2011) i są pokłosiem ograniczonej rozdzielczości pierwszej techniki stosowanej do identyfikacji CNV na skalę całego genomu, jaką była porównawcza hybrydyzacja genomowa do mikromacierzy (CGH lub array CGH, ang. *array comparative genome hybridization*). CGH polega na hybrydyzacji dwóch próbek genomowego DNA do sond DNA unieruchomionych na mikromacierzy, przy czym zwykle jedna z porównywanych próbek to tzw. genom referencyjny, reprezentowany przez te sondy. Względne różnice między próbkami w intensywności sygnału generowanego przez daną sondę odzwierciedlają różnice w liczbie kopii odpowiadającego jej regionu. Rozdzielczość metody CGH zależy od rodzaju zastosowanych sond i gęstości pokrycia sekwencji referencyjnej, jednak zwykle nie przekracza kilku - kilkudziesięciu kpz. Mimo to, właśnie eksperymenty z wykorzystaniem CGH dostarczyły pierwszych informacji o CNV, zarówno u człowieka (Iafrate at al. 2004), jak i u roślin (Springer et al. 2009).

Zaadaptowanie w kolejnych latach technik NGS do badania polimorfizmu liczby kopii umożliwiło zwiększenie rozdzielczości przewidywań regionów CNV, a nawet precyzyjne

określenie granic wielu wariantów, na co nie pozwalała metoda CGH. Nie ulega przy tym wątpliwości, że proces ten był stymulowany przez projekt poznania zmienności genetycznej człowieka (Projekt 1000 Genomów, ang. *The 1000 Genomes Project,* https://www.internationalgenome.org/about#1000G_PROJECT) oraz szereg inicjatyw, zorientowanych na poznanie genetycznych przyczyn chorób u ludzi. Na potrzeby tych badań opracowano wówczas dziesiątki narzędzi bioinformatycznych do identyfikacji CNV (Zhao et al. 2013). Narzędzia te w zdecydowanej większości bazują na analizie wyników mapowania odczytów do sekwencji genomu referencyjnego i wykorzystują jedno lub więcej z trzech podstawowych podejść. W podejściu opartym o analizę głębokości pokrycia (ang. *read depth*) wykorzystuje się informację o rozłożeniu zmapowanych odczytów w genomie i poszukuje obszarów, w których pokrycie odbiega od przeciętnego, wskazując tym samym na obecność duplikacji lub delecji. Wymaga to dodatkowo uwzględnienia różnic wynikających z innych przyczyn, takich jak gorsza wydajność sekwencjonowania regionów o wysokiej zawartości par GC. Aby zminimalizować ryzyko fałszywie pozytywnych wyników, przeszukanie prowadzi się w tzw. oknach (ang. *window* lub *bin*) o ustalonej wielkości. Sąsiadujące ze sobą okna o podobnym pokryciu są następnie łączone i wyznaczane są końcowe koordynaty regionu CNV. Wielkość okien wyznacza dolną granicę długości CNV, które można zidentyfikować tą metodą, przy czym nie może ona być niższa niż długość pojedynczego odczytu. Nie ma natomiast górnej granicy. Z tego powodu podejście to bardzo dobrze sprawdza się w identyfikacji dużych CNV (powyżej 1 kb). Nie pozwala ono jednak na określenie dokładnych koordynat wariantów CNV, ani pozycji genomowej, w której znajdują się dodatkowe kopie, w przypadku duplikacji.

Większą precyzję w określeniu miejsc insercji / delecji wykazuje drugie podejście, oparte na analizie niespójności w mapowaniu dwóch odczytów pochodzących od jednego fragmentu DNA, czyli tzw. odczytów sparowanych (ang. *read pair*). Podejście to sprawdza się jednak dobrze głównie w przypadku delecji i ma bardzo ograniczoną możliwość identyfikacji dużych insercji. Najdokładniejszą (z precyzją do 1 nt) lokalizację miejsc rearanżacji strukturalnych umożliwia ostatnie podejście, tzw. *split read*, w którym poszukuje się niezgodności w mapowaniu, obserwowanych w obrębie indywidualnych odczytów sekwencyjnych. Jednocześnie jednak podejście to generuje nawiększy odsetek wyników fałszywie pozytywnych. Ponadto tylko analizy głębokości pokrycia umożliwiają oszacowanie liczby kopii DNA w genomie dla zidentyfikowanych wariantów. Z tego powodu częstą praktyką jest wykorzystanie narzędzi hybrydowych lub łączenie wielu programów podczas analizy. Takie kombinowane podejście zastosowano w każdej z trzech faz realizacji Projektu 1000 Genomów człowieka, co pozwoliło na identyfikację dziesiątków tysięcy wariantów strukturalnych (Sudmant et al. 2015).

Obecnie szacuje się, że CNV może kształtować zmienność kilkunastu procent genomu ludzkiego, przyczyniając się między innymi do różnic międzypopulacyjnych, z których wiele może mieć znaczenie adaptacyjne. Dobrym przykładem jest tu zróżnicowanie liczby kopii ludzkiego genu *AMY1* kodującego amylazę ślinową – enzym zaangażowany w proces trawienia skrobi. Liczba kopii *AMY1* w diploidalnym genomie może wynosić od dwóch do nawet kilkunastu, przy czym średnio jest wyższa u ludzi w populacjach stosujących dietę bogatą w

skrobię, w porównaniu z populacjami żywiącymi się głównie produktami ubogimi w ten węglowodan (Perry et al. 2007). CNV wydaje się również mieć niebagatelne znaczenie w etiologii takich chorób jak cukrzyca, łuszczyca, choroba Parkinsona czy też różnego rodzaju nowotwory, a nawet w podatności na zarażenie wirusem HIV, aczkolwiek złożoność strukturalna wielu wariantów CNV znacząco utrudnia zasocjowanie ich występowania z konkretnymi fenotypami w badaniach populacyjnych (Usher i McCarrol 2015).

Motywacja i cel badań

Jednym z najlepiej poznanych pod względem genetycznym gatunków roślinnych jest rzodkiewnik pospolity (*Arabidopsis thaliana*), mający mniejszy i prostszy genom w porównaniu z wieloma roślinami użytkowymi (diploidalny, ok. 125 Mpz, 5 par chromosomów). Samopylność oraz łatwość hodowli i transformacji sprawiają, że rzodkiewnik jest często wykorzystywany jako modelowy system, np. w badaniach funkcji genów, mechanizmów i zależności komórkowych, czy też reakcji na stres biotyczny i abiotyczny. Dostępność licznych danych transkryptomicznych, proteomicznych, metabolomicznych, itp. oraz bogate kolekcje mutantów rzodkiewnika ułatwiają prowadzenie złożonych analiz porównawczych i funkcjonalnych, często nieosiągalnych dla innych gatunków roślin.

W 2008 r. podjęto międzynarodową inicjatywę zidentyfikowania wariantów genetycznych w rzodkiewniku, poprzez zbadanie sekwencji genomowej co najmniej tysiąca naturalnych ekotypów tego gatunku, pod hasłem Projektu 1001 Genomów (ang. *The 1001 Genomes Project,* https://1001genomes.org/). W ramach realizacji pilotowej fazy tego projektu, obejmującej analizę 80 ekotypów, skupiono się jednak na opisaniu SNP oraz krótkich indeli (do 20 pz), natomiast analizę CNV potraktowano dość marginalnie (Cao et al. 2011). Zastosowano wówczas analizę głębokości pokrycia i w efekcie zidentyfikowano 1059 regionów CNV o długości od 1 kpz do 13 kpz, które obejmowały łącznie 2% genomu. Tymczasem badania genomu człowieka, ale także np. kukurydzy, sugerowały dużo większy udział CNV w generowaniu zmienności genetycznej. Dlatego, **za podstawowy cel moich badań przyjęłam uzupełnienie istniejącego katalogu wariantów strukturalnych rzodkiewnika o duże indele oraz warianty CNV, poprzez ich identyfikację i charakterystykę. Następnie postanowiłam ustalić, w jakim stopniu polimorfizm liczby kopii wpływa na organizację informacji genetycznej i zmienność genów oraz czy może być powiązany ze zmiennością obserwowaną na poziomie fenotypu. Zamierzałam również ocenić użyteczność wariantów CNV jako markerów w analizach genetycznych i populacyjnych, zwyczajowo bazujących na danych SNP.** Pozwoliłoby to otworzyć pole dla szerokiego zastosowania CNV do analiz fenotypu w badaniach asocjacyjnych całego genomu (GWAS, ang. *genome-wide association study*). Realizowanie tak postawionych celów okazało się tym bardziej zasadne, że w kolejnej fazie Projektu 1001 Genomów (w której badano zmienność genetyczną aż 1135 ekotypów), w ogóle nie podjęto się analizy wariantów strukturalnych większych niż 49 pz (1001 Genomes Consortium 2016).

Omówione poniżej badania, składające się na moje osiągnięcie naukowe, zatytułowane **„IDENTYFIKACJA POLIMORFIZMU LICZBY KOPII DNA JAKO ISTOTNEGO SKŁADNIKA**

KSZTAŁTUJĄCEGO ZMIENNOŚĆ GENETYCZNĄ *ARABIDOPSIS THALIANA*" realizowałam w latach 2012-2020 w Instytucie Chemii Bioorganicznej PAN (ICHB PAN) w Poznaniu, a także w Instytucie Informatyki Politechniki Poznańskiej. Prace prowadziłam w zespole badawczym kierowanym przez prof. dr hab. Marka Figlerowicza. Analiza polimorfizmu liczby kopii u roślin stanowiła całkowicie nowe zagadnienie, nie będące kontynuacją żadnych wcześniejszych badań tego zespołu. **Dla stworzenia warsztatu badawczego niezbędne było zatem również opracowanie adekwatnych podejść eksperymentalnych i bioinformatycznych do analizy CNV, jako cel pośredni w mojej pracy**.

Moje osiągnięcie naukowe stanowi zbiór czterech wieloautorskich publikacji, z których trzy przedstawiają wyniki oryginalnych badań, a jedna jest pracą przeglądową. Prace są opublikowane w modelu otwartego dostępu (ang. *open access*). We wszystkich jestem pierwszą i/lub korespondencyjną autorką. Pliki tych publikacji w formacie PDF oraz towarzyszących im materiałów dodatkowych są dołączone do elektronicznej wersji wniosku. Do publikacji (Zmienko et al. 2020) dołączona jest również kopia raportu z przebiegu recenzji, z uwagami recenzentów i odpowiedziami autorów, dostępnego publicznie na portalu Plant Cell. Oświadczenia habilitantki oraz autora korespondencyjnego (jeśli inny niż habilitantka, zgodnie z wytycznymi w dokumencie pt. „SPOSÓB POSTĘPOWANIA W SPRAWIE NADANIA STOPNIA DOKTORA HABILITOWANEGO w Instytucie Chemii Bioorganicznej Polskiej Akademii Nauk w Poznaniu", stanowiącym Załącznik nr 2 do Uchwały Rady Naukowej ICHB PAN nr 12/RN_117 z dnia 6 listopada 2019 r.), potwierdzające mój wkład w powstanie poszczególnych publikacji, znajdują się w **Załączniku 6**. Ponadto każda z trzech prac zawierających wyniki oryginalne zawiera opublikowaną sekcję „Authors' contributions", która podsumowuje udział poszczególnych autorów w powstaniu pracy.

Polimorfizm liczby kopii DNA w roślinach

W czasie gdy rozpoczynałam badania nad polimorfizmem liczby kopii, doniesienia literaturowe o tym zjawisku w roślinach były stosunkowo nieliczne. Nie istniały również żadne publikacje przeglądowe, prezentujące aktualny stan wiedzy na ten temat. Postanowiłam zatem przede wszystkim zebrać i usystematyzować dostępne informacje o CNV w roślinach. Efektem tego przedsięwzięcia jest publikacja pt. **„Copy number polymorphism in plant genomes" (Żmieńko et al. 2014, Theor Appl Genet)**, której jestem główną autorką i która stanowi część prezentowanego przeze mnie osiągnięcia naukowego. We wprowadzeniu do tej pracy opisałam pokrótce historię badań nad CNV i przedstawiłam molekularne mechanizmy, uważane za główne przyczyny powstawania duplikacji i delecji. Zaprezentowałam również podstawowe sposoby, na jakie CNV może wpływać na ekspresję informacji genetycznej. Duplikacje bądź delecje całych genów mogą zmieniać jego dawkę (ang. *gene doasge*), czyli liczbę kopii genu biorących udział w transkrypcji. Z kolei zmiany liczby kopii regionu promotora czy sekwencji regulatorowych mogą wpływać na poziom ekspresji genu bądź jego poszczególnych kopii, poprzez efekt pozycji (ang. *position effect*). Znane są również częściowe duplikacje bądź delecje, wpływające na zmianę struktury genu, a przez to na sekwencję i funkcjonalność jego produktu.

W dalszej części pracy omówiłam wspomniane wcześniej metody identyfikacji CNV w skali całogenomowej, a następnie przedstawiłam projekty, w których wykorzystano te metody do analizy genomów roślinnych i zidentyfikowano warianty strukturalne o charakterze zmian liczby kopii. W momencie tworzenia tej publikacji, takie dane były dostępne tylko dla siedmiu gatunków roślin: kukurydzy, soi, ryżu, sorgo, rzodkiewnika, pszenicy i pomidora. Pochodziły one z eksperymentów, w których porównywano między sobą od dwóch do ponad stu linii / ekotypów. Zebraliśmy i przeanalizowaliśmy wyniki tych eksperymentów, co podsumowałam w obszernej tabeli, zestawiającej zastosowane metody, liczbę linii / ekotypów wykorzystanych w badaniach, liczbę i typ znalezionych wariantów CNV, a także – o ile było to możliwe – liczbę genów nakładających się na sekwencje zidentyfikowanych regionów. Następnie opisałam najważniejsze inicjatywy i międzynarodowe projekty, mające na celu skatalogowanie zmienności genetycznej roślin, skupiając się na aspektach dotyczących polimorfizmu liczby kopii.

Pierwszym i najlepiej wówczas scharakteryzowanym pod względem CNV gatunkiem roślinnym była kukurydza. Większość informacji na temat polimorfizmu liczby kopii u tego gatunku pochodziła z porównania różnych linii z genomem referencyjnym (linią B73) metodą CGH. Istotnym ograniczeniem takiego podejścia jest preferencja do wykrywania delecji, a nie duplikacji, w badanych liniach. Stąd u kukurydzy zidentyfikowano przede wszystkim warianty, wynikające z braku dużych fragmentów genomu w pewnych liniach, podczas gdy ich obecność stwierdzano w genomie referencyjnym. Dodatkowym minusem CGH jest brak możliwości wykrycia segmentów całkowicie nieobecnych w genomie referencyjnym, z uwagi na brak adekwatnych sond na mikromacierzy. Dopiero zastosowanie metod NGS i badania dużych populacji kukurydzy, jak również ryżu czy soi, rzuciły więcej światła na drugi typ zmian liczby kopii – duplikacje oraz ich powszechne występowanie w genomach roślinnych, co również opisałam w omawianej publikacji przeglądowej.

Pod względem funkcjonalnym szczególnie interesujący wydaje się wpływ CNV na strukturę i aktywność genów, czemu poświęciłam ostatnią część pracy. Omówiłam w niej niektóre spośród zebranych przez nas przykładów powiązań pomiędzy zmianą liczby kopii genów, a istotnymi cechami fenotypowymi roślin, takimi jak walory hodowlane, odporność na patogeny czy metale ciężkie. Zwróciłam również uwagę na ogromny potencjał zjawiska CNV w dynamicznej adaptacji roślin do zmieniających się warunków środowiska. Ponadto wskazałam na możliwość zastosowania wariantów CNV w analizach asocjacyjnych, który to wątek rozwinęliśmy w późniejszych badaniach własnych (por. niżej).

Podsumowując, we wspomnianej pracy przeglądowej po raz pierwszy zostały zgromadzone i przedstawione w sposób kompleksowy informacje na temat badań polimorfizmu liczby kopii u roślin, wskazując na powszechność oraz istotne znaczenie tego zjawiska. Chciałabym podkreślić, że publikacja ta przyciągnęła uwagę środowiska naukowego, o czym świadczy utrzymująca się duża liczba cytowań w kolejnych latach. Już w 2015 roku baza bibliograficzna Web of Science przyznała naszej pracy oznaczenie „Highly cited paper”, z uwagi na wyróżniającą się liczbę cytowań (znalazła się ona – i nadal utrzymuje się – w 1% najlepiej

cytowanych publikacji w obszarze „Agricultural Sciences"). Dotąd zacytowano ją 97 razy, przy czym 25% cytowań pochodzi z lat 2019-2020 (źródło: Web of Science Core Collection).

Charakterystyka liczby kopii genów MSH2, AT3G18530 i AT3G18535, położonych w obrębie złożonego strukturalnie regionu CNV

Jedyne dostępne w owym czasie wyniki przewidywań CNV dla światowej populacji rzodkiewnika, wskazywały m. in. na istnienie dwóch sąsiadujących ze sobą regionów o potencjalnie dużej zmienności liczby kopii, położonych na chromosomie 3 (Cao et al. 2011). Jeden z nich prawie w całości obejmował sekwencje dwóch zachodzących na siebie genów, *AT3G18530* oraz *AT3G18535*. Geny te kodują odpowiednio białko zawierające domenę z powtórzeniami typu armadillo (ARM, ang. *armadillo repeat domain*) oraz ligazę tyrozynową tubuliny, przy czym rola żadnego z nich nie została dotąd poznana. Natomiast drugi region zmienny częściowo obejmował sekwencję konserwatywnego genu *MSH2*. *MSH2* koduje białko zaangażowane w rozpoznawanie niesparowań podwójnej nici DNA (powstałych wskutek błędów replikacji i rekombinacji) oraz uruchomienie szlaku naprawczego, kluczowego dla utrzymania stabilności genomów. Pierwotnie szlak ten odkryto u bakterii, gdzie funkcje rozpoznawania niesparowań pełni białko MutS; MSH2 (czyli MutS homolog 2) jest jego eukariotycznym homologiem. Białko MSH2 wykazuje aktywność po utworzeniu heterodimeru z innymi białkami z rodziny MSH – taki kompleks rozpoznaje różne typy niesparowań, zależne od rodzaju drugiego białka (u rzodkiewnika może to być MSH3, MSH6 lub MSH7), inicjując procesy naprawcze. Uważa się, że geny kodujące białka zaangażowane w interakcje z innymi białkami (a do takich należy MSH2), są bardzo „wrażliwe" na efekty lokalnych duplikacji i delecji w genomie, prowadzących do zaburzenia równowagi pomiędzy składnikami interakcji. Mają one zwykle negatywny wpływ na kondycję organizmu i rzadko utrzymują się w procesie ewolucji – jest to tzw. hipoteza balansu dawki genu (ang. *gene dosage balance hypothesis*) (Birchler i Veitia 2010). Skłoniło mnie to do bardziej szczegółowych badań nad strukturą całego wspomnianego regionu i jego zmiennością. Wymagało to zastosowania skutecznych podejść eksperymentalnych, umożliwiających detekcję zarówno delecji, jak i duplikacji oraz wiarygodną ocenę liczby zduplikowanych kopii genów.

Już podczas przeglądu prac poświęconych CNV w genomach roślinnych moją uwagę zwrócił fakt, że w dotychczasowych badaniach, do eksperymentalnej weryfikacji przewidywań bioinformatycznych wybierano najczęściej warianty typu delecji, podczas gdy obecność duplikacji rzadko była potwierdzana. Miało to swoje oczywiste uzasadnienie w różnym stopniu trudności w zbadaniu delecji (potwierdzanych zwykle przy pomocy PCR bądź sekwencjonowania Sangera) i duplikacji, a więc regionów nieunikalnych, często wielokrotnie powtórzonych w genomie, jednak, w mojej opinii, tym bardziej wymagających pogłębionej analizy. Podobny trend zaobserwowałam zresztą również w przypadku całogenomowych badań CNV u człowieka. Jednym z nielicznych pod tym względem wyjątków była praca, w której identyfikowano geny człowieka występujące w wielu wariantach liczby kopii (Handsaker et al. 2015) – określam je dalej jako geny multialleliczne. W pracy tej skutecznie analizowano geny multialleliczne, stosując emulsyjny PCR (ddPCR, ang. *droplet digital PCR*). Jest to

ilościowa odmiana PCR wykorzystująca znaczniki fluorescencyjne, w której próbkę rozdziela się przed amplifikacją na tysiące małych kropel, stanowiących niezależne mieszaniny reakcyjne. Po zakończeniu reakcji, wyjściową ilość matrycowego DNA określa się poprzez zliczenie w odpowiednim detektorze tych kropel, w których zaszła amplifikacja. Niezbędne do tego jest uzyskanie odpowiednio dużego rozcieńczenia DNA, tak aby w każdej kropli znajdowała się średnio tylko jedna kopia matrycy oraz zastosowanie podczas analiz statystyki Poissona, w celu korekcji przewidywanych odchyleń od tej wartości. W wystandaryzowanych warunkach, różnice między badanymi próbkami (np. ekotypami) w liczbie kropel, w których wykryto sygnał, niezależnie od siły tego sygnału, będą zatem w sposób bezpośredni odzwierciedlać różnicę w liczbie kopii regionów badanych, znajdujących się w genomowym DNA użytym do reakcji. Pomimo niewielkiej możliwości multipleksowania, ddPCR nadaje się do rutynowego genotypowania wybranych regionów w małej i średniej liczbie próbek. We wspomnianej pracy walidacja objęła 22 geny, z których każdy przebadano w 19 - 38 indywidualnie dobranych próbkach, również takich o wysokim stopniu duplikacji badanych genów (do 9 kopii na diploidalny genom). Opierając się na powyższych doniesieniach, postanowiłam wykorzystać emulsyjny PCR do badań wspomnianego regionu CNV u rzodkiewnika.

Równolegle, z myślą o przyszłych zastosowaniach, poszukiwałam podejścia, które byłoby równie skuteczne w badaniu regionów multiallelicznych, a jednocześnie pozwoliłoby na znaczące zwiększenie zarówno liczby badanych genów jak i analizowanych ekotypów. Doskonała do tego celu wydawała się metoda multipleksowej zależnej od ligacji amplifikacji sond (MLPA, ang. *multiplex ligation-dependent probe amplification*). Polega ona na hybrydyzacji do genomowego DNA dwóch półsond DNA, komplementarnych do sekwencji analizowanego regionu, a następnie ich ligacji, co prowadzi do utworzenia pełnej sondy. Taką sondę następnie amplifikuje się metodą PCR (Schouten et al. 2002). Długość i ilość powstałego produktu ocenia się poprzez wysokorozdzielczą elektroforezę kapilarną. Im więcej kopii danego regionu znajduje się w genomie, tym więcej produktu zaobserwujemy w końcowej fazie analizy. Ponieważ zajście ligacji jest warunkowane hybrydyzacją obu półsond w bezpośrednio sąsiadujących pozycjach, gwarantuje to wysoką specyficzność reakcji. Z kolei możliwość zróżnicowania długości sond pozwala na opracowanie zestawów multipleksowych do analizy wielu regionów jednocześnie. Co ważne – z eksperymentalnego punktu widzenia genomowy DNA służy w reakcji wyłącznie do wymuszenia właściwego ułożenia odpowiedniej liczby par półsond względem siebie, zatem całość analizy może być prowadzona w bardzo powtarzalnych i kontrolowanych warunkach, w dużej mierze niezależnych od jakości czy integralności DNA w próbce. Od czasu jej opracowania w 2002 r. metoda MLPA doczekała się licznych adaptacji i modyfikacji. Jedną z nich opracował zespół prof. dr hab. Piotra Kozłowskiego z ICHB PAN, któremu udało się znacząco skrócić długość półsond, dzięki czemu mogą one być całkowicie syntetyczne, podczas gdy wcześniej jedną z nich uzyskiwano przez klonowanie i amplifikację (Marcinkowska et al. 2010). Tą zmodyfikowaną wersję metody MLPA postanowiłam również zaadaptować do badań w rzodkiewniku, co pozwoliłoby mi na efektywne badanie w przyszłości wielu genów w dużej liczbie ekotypów.

W celu zbadania liczby kopii genów *MSH2*, *AT3G18530* i *AT3G18535* zaprojektowałam eksperyment MLPA, w którym wykorzystaliśmy opracowany przez nas multipleksowy zestaw sond do zbadania całego regionu, obejmującego wspomniane geny oraz dodatkowo dwa geny flankujące, *HDA15* i *BRC1*, które – według moich przewidywań, miały nie wykazywać zmienności liczby kopii. Badany obszar miał długość ponad 25 kpz. Eksperyment MLPA objął 189 ekotypów, przy czym ekotyp Col-0 reprezentował genom referencyjny, o niezmiennej liczbie kopii (2 kopie na diploidalny genom) każdego z genów. Ponadto dla każdego z genów zoptymalizowaliśmy warunki reakcji ddPCR, a następnie określiliśmy przy pomocy tej metody ich liczby kopii, w 92 różnych ekotypach.

Dzięki analizie dużej populacji oraz zastosowaniu par sond dla każdego z badanych genów byłam w stanie na podstawie wyników MLPA wyróżnić klastry próbek, którym w większości jednoznacznie przypisałam wartości liczby kopii dla każdego z genów. Rozdzielczość samej metody MLPA nie pozwalała na wydzielenie indywidualnych klastrów próbek z duplikacją o liczbie kopii wyższej niż 6 – pod tym względem zdecydowaną przewagę wykazywał emulsyjny PCR, który pozwolił na ocenę ilościową wysoko zduplikowanych genów, aż do 14 kopii. Poza tym jednak obie metody wykazały niezwykle wysoką zgodność przewidywań. Co istotne, wcześniejsze – bioinformatyczne - przewidywania liczby kopii uzyskane dla badanego przez nas regionu (Cao et al. 2011) okazały się daleko mniej precyzyjne.

Łącznie uzyskane przez nas wyniki wykazały umiarkowany stopień zmienności liczby kopii genu *MSH2*. Nigdy nie obserwowałam jego delecji, co było wynikiem oczekiwanym, biorąc pod uwagę wysoką konserwatywność *MSH2* i jego udział w kluczowych procesach komórkowych. Intrygujące było natomiast zaobserwowanie duplikacji *MSH2* aż w 12 ekotypach, przy połączone wyniki analiz MLPA i ddPCR, obejmujących 4 różne eksony sugerowały, że duplikacje dotyczą dużej części, a może nawet całego genu *MSH2*.

Drugą, nie mniej intrygującą obserwacją był fakt, że liczba kopii *MSH2* była wyraźnie skorelowana z liczbą kopii *AT3G18530* i *AT3G18535* w taki sposób, że duplikacja *MSH2* prawie zawsze towarzyszyła duplikacji pozostałych dwóch genów. Pomimo obserwowanej zbieżności, *AT3G18530* i *AT3G18535* wykazywały wzajemnie dużo większą, bo prawie stuprocentową korelację liczby kopii i były zduplikowane w większej liczbie ekotypów (20) niż *MSH2*. Ponadto niezwykle często ulegały też delecjom (101 ekotypów), a w takich przypadkach liczba kopii *MSH2* zawsze pozostawała na podstawowym, niezmienionym poziomie. Skłoniło mnie to do przeprowadzenia pogłębionych analiz tego regionu i próby zidentyfikowania przyczyn jego wysokiej zmienności.

Udział niealleliczne rekombinacji homologicznej w generowaniu duplikacji oraz delecji genów AT3G18530 i AT3G18535

Poszukując źródła wariantu z delecją obejmującą sekwencje genów *AT3G18530* i *AT3G18535*, przeprowadziłam analizę mapowań odczytów NGS (dostępnych w publicznej bazie danych) dla 1135 ekotypów i ustaliłam, że wariant ten jest obecny u ponad 60% światowej populacji rzodkiewnika. Ekotypy z delecją miały różne pochodzenie geograficzne, przy czym ani analizy

filogenetyczne, ani badania haplotypów czy analizy sprzężeń wokół badanego regionu nie wskazywały na ich wspólne źródło. Sumarycznie, wszystkie dotychczasowe obserwacje sugerowały istnienie powtarzalnego mechanizmu odpowiedzialnego za wielokrotne i niezależne wystąpienie delecji (oraz duplikacji) obu genów jednocześnie. Postawiłam wówczas hipotezę, że mechanizmem odpowiedzialnym za zmienność w badanym regionie może być niealleliczna homologiczna rekombinacja (NAHR, ang. *non-allelic homologous recombination*). NAHR polega na zajściu homologicznej rekombinacji między odcinkami w genomie, które wykazują bardzo wysoki stopień identyczności, ale nie są allelami. Noszą one nazwę segmentalnych duplikacji lub powtórzeń o niskiej liczbie kopii (LCR, ang. *low-copy repeats*). Charakterystyczną cechą wariantów CNV będących wynikiem NAHR jest zatem ich prawie identyczny rozmiar i lokalizacja – ich granice są bowiem wyznaczane położeniem segmentalnych duplikacji w genomie. Według modelu Holliday'a, NAHR zachodząca pomiędzy segmentalnymi duplikacjami w obrębie tej samej podwójnej nici DNA prowadziłaby do powstania wyłącznie delecji, podczas gdy niealleliczna rekombinacja między chromatydami lub homologicznymi chromosomami, prowadziłaby do powstania delecji oraz duplikacji (Kehrer-Sawatzki et al. 2014).

Uważa się, że w NAHR biorą udział segmentalne duplikacje o długości co najmniej 1 kpz i podobieństwie sekwencji większym niż 95%. Takie właśnie dwa powtórzenia zidentyfikowałam w sekwencji genomu referencyjnego w interesującym mnie regionie. Otaczają one geny *AT3G18530* i *AT3G18535*. Każde z nich ma długość 1238 pz i różnią się one między sobą zaledwie w 11 pozycjach. W oparciu o dane literaturowe założyłam, że dwa długie odcinki o całkowitej identyczności, położone w części 5' tych powtórzeń (pozycje 1-230 i 232-768) mogą być miejscami wydajnego inicjowania NAHR. Aby to potwierdzić, wykonaliśmy analizy PCR oraz częściowe sekwencjonowanie spodziewanych miejsc zajścia rekombinacji w 27 ekotypach z delecją i 8 ekotypach z niskim stopniem duplikacji. Uzyskane wyniki były zgodne z zaproponowanym przeze mnie modelem udziału NAHR w generowaniu zmienności liczby kopii *AT3G18530* oraz *AT3G18535*. Dodatkowo, w kilku ekotypach zaobserwowałam przypadki konwersji sekwencji pomiędzy powtórzeniami – zjawisko często towarzyszące rekombinacji.

Opisane wyżej wyniki zebrałam w pracy pt. **„*Arabidopsis thaliana* population analysis reveals high plasticity of the genomic region spanning *MSH2, AT3G18530* and *AT3G18535* genes and provides evidence for NAHR-driven recurrent CNV events occurring in this location" (Zmienko et al. 2016, BMC Genomics)**, której jestem główną autorką i która stanowi część mojego osiągnięcia naukowego. Omówiłam w niej zmienność liczby kopii genów *MSH2*, *AT3G18530* i *AT3G18535* oraz przedstawiłam – po raz pierwszy dla genomu roślinnego rolę NAHR w generowaniu tej zmienności. Zasugerowałam, że scharakteryzowany przez nas region może być dobrym obiektem do badań samego mechanizmu NAHR, jak również jego potencjalnego wykorzystania. Liczne delecje *AT3G18530* i *AT3G18535* w populacji sugerują, że geny te nie są niezbędne dla rośliny, a zatem to właśnie brak presji selekcyjnej spowodował ujawnienie ogromnego potencjału NAHR, generując zmienność strukturalną w tym regionie. W genomie ludzkim wykazano co prawda obecność co najmniej kilkudziesięciu regionów

niestabilnych, narażonych na występowanie NAHR, ale jednocześnie w tych miejscach zajście NAHR zwykle wiąże się z negatywnym wpływem selekcyjnym i występowaniem chorób genetycznych (Liu et al. 2012), przez co nie stanowią one dobrego modelu do badań samego zjawiska i jego częstości. Dodatkowym aspektem naszych badań, który również opisałam w pracy **(Zmienko, et al. 2016)**, jest zastosowanie po raz pierwszy MLPA i ddPCR do analiz liczby kopii DNA w genomie roślinnym oraz bezpośrednie porównanie obu technik.

Przeprowadzone badania nie dały jednoznacznej odpowiedzi w jaki sposób dochodzi do duplikacji genu *MSH2* oraz czy jest ona całkowita. Mimo to, wyniki analiz liczby kopii mocno przemawiają za istnieniem powiazań tego zdarzenia z duplikacją pozostałych dwóch genów, tym bardziej, że w genomie referencyjnym jedno z powtórzeń biorących udział w NAHR obejmuje fragment pierwszego eksonu *MSH2*. Aby zebrać na ten temat dodatkowe informacje, przeprowadziłam dalsze analizy w ramach kierowanego przeze mnie zadania badawczego Miniatura, pt. „Zmienność liczby kopii genu *MSH2* u *Arabidopsis thaliana* - struktura duplikacji, zakres polimorfizmu w naturalnych ekotypach oraz wpływ na ekspresję genów zaangażowanych w naprawę niesparowań DNA”, na realizację którego uzyskałam finansowanie z Narodowego Centrum Nauki w roku 2017. Opracowałam w tym celu panel sond MLPA pokrywających wszystkie 13 eksonów *MSH2*. W każdym z przeanalizowanych przeze mnie ekotypów z duplikacją zaobserwowałam brak zmienności w obrębie eksonów 7-8 oraz 12-13. Wynik ten potwierdziłam przez częściowe sekwencjonowanie obszaru duplikacji w kilku ekotypach. Ponadto liczne zmiany w obrębie sekwencji zduplikowanej kopii genu sugerowały, że nie jest ona funkcjonalna. Co ciekawe jednak – zaobserwowałam podwyższenie ekspresji genu *MSH2* w siewkach tych ekotypów, w których wystąpiła duplikacja *AT3G18530* i *AT3G18535*, niezależnie od stanu liczby kopii *MSH2*. Jak sądzę, może to być związane ze zmianami w obszarze promotora *MSH2*, powstałymi na skutek rekombinacji regionów LCR. Badania tego – najwyraźniej złożonego – regionu strukturalnego kontynuuję obecnie stosując technikę sekwencjonowania długich odczytów.

Zastosowanie techniki MLPA do badania multiallelicznych genów roślinnych

W celu zastosowania techniki MLPA do badania genów *MSH2*, *AT3G18530 i AT3G18535*, niezbędne było zoptymalizowanie wszystkich jej etapów (od projektowania sond po analizę wyników) do badań w rzodkiewniku. Dodatkowym efektem tych działań było określenie przez nas liczby kopii kilkunastu innych, dotąd niescharakteryzowanych pod kątem CNV, genów, w 80 ekotypach. Były to w większości geny multialleliczne, które wytypowałam do badań w oparciu o dane literaturowe i nasze własne, jeszcze wówczas nieopublikowane wyniki bioinformatycznej identyfikacji regionów CNV (por. niżej). Uzyskane wyniki wykorzystałam do zaprezentowania opracowanego przez nas uniwersalnego protokołu MLPA, w postaci publikacji pt. **„MLPA-based Analysis of Copy Number Variation in Plant Populations” (Samelak-Czajka et al. 2017, Front. Plant. Sci.)**, w której jestem autorką korespondencyjną i która stanowi część prezentowanego przeze mnie osiągnięcia naukowego. Opisałam w niej, krok po kroku, tekstowo oraz graficznie, poszczególne etapy eksperymentu MLPA, od zaprojektowania sond aż po analizę wyników, tam gdzie to możliwe wskazując na

ogólnodostępne narzędzia, które mogą okazać się w tym pomocne oraz rozwiązania, które sprawdziliśmy w przypadku analiz CNV u rzodkiewnika. Opracowałam również szablon do półautomatycznego projektowania zestawu sond MLPA, który został dołączony do pracy jako suplement. Szablon ten ułatwia zaprojektowanie reakcji multipleksowej w taki sposób, że do użytkownika należy wyłącznie wybranie sekwencji specyficznej dla badanego regionu oraz decyzja, jaka ma być ostateczna długość sondy. Odpowiednie formuły w szablonie dodają następnie sekwencje adapterowe oraz uniwersalny linker o odpowiedniej długości i wskazują użytkownikowi gotowe sekwencje oligonukleotydów (półsond) do zamówienia.

W omawianej pracy przedyskutowałam również zalety i ograniczenia metody MLPA oraz – wykorzystując wyniki uzyskane dla konkretnych genów – omówiłam różnorodne aspekty analizy, takie jak zastosowanie jednej lub dwóch sond do badania genu, uzyskanie wyniku wskazującego na częściową duplikację bądź delecję, czy też istnienie (w niektórych przypadkach) zależności pomiędzy niespecyficzną hybrydyzacją bądź geograficznym rozmieszczeniem badanych ekotypów, a obserwowaną siłą sygnału MLPA.

W metodzie MLPA niezwykle istotną rzeczą jest zaprojektowanie i zwalidowanie zestawu sond kontrolnych, służących do normalizacji wyników pomiędzy porównywanymi próbkami. Powinny być one nakierowane na taki gen lub jego fragment, który nie wykazuje zmian liczby kopii w badanym zestawie próbek. W przypadku badań populacyjnych na rzodkiewniku, gdzie potencjalnym celem analizy były bardzo różne geny, wymagało to znalezienia sond możliwie uniwersalnych, które będą nadawały się do zastosowania w badaniach obejmujących jak największą liczbę ekotypów. W omawianej pracy wskazałam pięć takich sond kontrolnych, obejmujących geny: *DCL1*, *PS2*, *APG10*, gen oksydoreduktazy (*AT4G21580*) oraz *PFD5*. Sondy te zostały przez nas również z powodzeniem wykorzystane w badaniach regionu obejmującego geny *MSH2*, *AT3G18530* i *AT3G18535* **(Zmienko et al. 2016)**.

Dodatkowym aspektem ujętym w opublikowanym przez nas protokole było dobranie odpowiedniej ilości genomowego DNA do reakcji. Ponieważ MLPA była wcześniej stosowana przede wszystkim do badania genomu człowieka, zalecana zwykle ilość matrycy była stosunkowo duża i wynosiła 50-250 ng DNA na reakcję. Biorąc pod uwagę, że genom rzodkiewnika jest ponad 20 razy mniejszy od genomu człowieka, zasadne wydawało się zmniejszenie tej ilości. Zaplanowałam zatem serię eksperymentów, które pokazały, że możliwe jest znaczące zmniejszenie ilości matrycy w reakcji, nawet do 2 ng **(Samelak-Czajka et al. 2017)**. Biorąc dodatkowo pod uwagę multipleksowość, metoda MLPA okazała się tym samym bardzo wydajnym i skutecznym podejściem do genotypowania liczby kopii genów u rzodkiewnika – i potencjalnie u innych roślin.

Atlas regionów CNV w genomie rzodkiewnika

Równolegle z rozwijaniem warsztatu eksperymentalnego i badaniem wybranych wariantów CNV, prowadziłam działania w kierunku uzyskania kompleksowej mapy regionów CNV w genomie rzodkiewnika metodami bioinformatycznymi. Moim początkowym celem była analiza danych dla 80 ekotypów. W międzyczasie jednak opublikowano dane NGS z

całogenomowego sekwencjonowania 1135 ekotypów (1001 Genomes Consortium 2016), o czym wspominałam wcześniej. Postanowiłam więc wykorzystać ten powiększony zestaw danych, co gwarantowało skuteczniejszą identyfikację wariantów strukturalnych obecnych w populacji. Biorąc pod uwagę, że w takim ujęciu polimorfizm liczby kopii nie był w ICHB PAN przedmiotem wcześniejszych badań, wiązało się to przede wszystkim z koniecznością rozpoznania i opanowania adekwatnych narzędzi. Miałam świadomość, że odmienne podejścia bioinformatyczne, zastosowane do tych samych danych NGS, powodują uzyskanie różnych – wysoce specyficznych – zestawów CNV (Alkan et al. 2011). Wzorując się na osiągnięciach Projektu 1000 Genomów człowieka, zdecydowałam się zatem położyć nacisk na przetestowanie różnorodnych metod dedykowanych identyfikacji CNV, a następnie ich połączenie w celu zbadania genomu rzodkiewnika. Zaznaczam, że etap optymalizacji i testów był tu niezbędny, ponieważ większość tych narzędzi zaprojektowano pod kątem analizy genomu człowieka oraz nie istniały żadne gotowe systemy integrujące ich działanie we wspólny potok analizy. W efekcie, opracowanie wieloetapowej ścieżki analizy i bioinformatyczne przewidywanie regionów CNV było pracą zespołową, przy której pełniłam rolę lidera.

Aby dokonać możliwie pełnej identyfikacji wariantów strukturalnych, zintegrowaliśmy siedem dedykowanych temu programów. Trzy z nich były oparte o analizy głębokości pokrycia (CNVnator, Control-FREEC, Genome STRiP CNV), dwa kolejne wykorzystywały metodę mapowania sparowanych odczytów (BreakDancer i VariationHunter), jeden program bazował na identyfikacji niespójnego mapowania wewnątrz odczytów (Pindel), a ostatni reprezentował podejście hybrydowe (Genome STRiP SV). Opracowaną ścieżkę analizy wykorzystaliśmy do identyfikacji zmian strukturalnych, pominiętych przez konsorcjum Projektu 1001 Genomów. Podzieliliśmy je na dwa rodzaje: duże indele (od 50 pz do 500 pz) oraz regiony CNV (od 500 pz). Jak przedstawiłam we wprowadzeniu, każde z zastosowanych przez nas podejść ma swoje wady i zalety, dlatego uznałam, że właściwym rozwiązaniem będzie nadanie im priorytetów na różnych etapach analizy, tak aby jak najlepiej wykorzystać ich mocne strony. I tak, przy identyfikacji CNV bazowaliśmy głównie na metodach analizy głębokości pokrycia, udokładniając następnie granice regionów przy pomocy pozostałych programów, podczas gdy przy identyfikacji dużych indeli opieraliśmy się przede wszystkim na analizie sparowanych odczytów. W efekcie zidentyfikowaliśmy w genomie rzodkiewnika 70137 dużych indeli oraz 34368 regionów CNV, spośród których 19003 były znalezione przez co najmniej dwa programy. Tym ostatnim nadałam zbiorczą nazwę wariantów AthCNV. Zestawienie naszych wyników z informacją na temat rozkładu SNP i małych indeli w genomie rzodkiewnika pokazało, że odkryte przez nas warianty dopełniają – a nie dublują – listę regionów genomu wykazujących zmienność strukturalną. Proces identyfikacji CNV opisałam w pracy pt. **„AthCNV - a map of DNA copy number variations in the *Arabidopsis thaliana* genome" (Zmienko et al. 2020, Plant Cell)"**, której jestem wiodącą autorką oraz współautorką korespondencyjną i która stanowi część opisywanego przeze mnie osiągnięcia naukowego.

Wpływ genów oraz transpozonów na kształtowanie rozkładu CNV w genomie

Dysponując atlasem wariantów CNV w genomie rzodkiewnika, skupiłam się następnie na ich charakterystyce. Najdłuższy z wariantów AthCNV miał długość blisko 1 Mpz, jednak w większości (92.1%) były one krótsze niż 20 kpz. Warianty wykazywały nierównomierne rozłożenie w genomie, pokrywając ponad 93% obszaru centromerów i tylko 28% genomu poza nimi. Niemniej, nawet pomijając centromery, było to najwyższe pokrycie CNV raportowane dotąd dla genomu rzodkiewnika. Przeprowadziłam zatem szeroko zakrojone porównania naszych wyników z innymi doniesieniami, które pokazały, że - w istocie - atlas, który z założenia miał stanowić katalog wszystkich CNV w rzodkiewniku, obejmuje warianty zidentyfikowane wcześniej w mniejszych projektach populacyjnych, ale także wiarygodnie reprezentuje delecje i – w nieco mniejszym stopniu – duplikacje, wykrywane w asemblowanych *de novo* genomach indywidualnych ekotypów.

Przeanalizowałam następnie rozłożenie AthCNV poza centromerami, pod kątem zawartości pokrytych przez nie obszarów genomu. Najliczniejsze występowanie wariantów zaobserwowałam w obrębie transpozonów (67.5% tych elementów w genomie pokrywało się z co najmniej 1 wariantem AthCNV, zwykle na całej długości). Nakładanie się AthCNV na sekwencje genów było zdecydowanie mniejsze, aczkolwiek aż 18.5% genów było pokryte przez jeden lub więcej wariantów CNV co najmniej w 90%. To nierównomierne rozłożenie można wyjaśnić zarówno poprzez bezpośrednie zaangażowanie transpozonów – mobilnych elementów genetycznych – w tworzenie nowych wariantów strukturalnych, jak i potencjalnie negatywny wpływ CNV na funkcjonalność genów. Wiadomym jest ponadto, że obecność transpozonów może zaburzać ekspresję pobliskich genów, np. poprzez indukowanie wzmożonej metylacji całego regionu, co może stanowić czynnik selekcyjny (Bourque et al. 2018). Rzeczywiście, choć wykazano brak preferencji co do miejsca insercji nowych transpozonów w genomie rzodkiewnika, to ich późniejsza delecja jest już bardzo selektywna i częściej obejmuje transpozony położone w pobliżu genów, kształtując ich wzajemny rozkład (Quadrana et al. 2016). Postanowiłam zatem zbadać, w jakim stopniu ogólne rozłożenie genów i transpozonów w genomie rzodkiewnika zależy od ich stopnia polimorfizmu. Na potrzeby tej analizy każdemu genowi oraz transpozonowi przypisałam status „-CNV", jeżeli pokrywał się on z co najmniej jednym wariantem AthCNV lub „-NONVAR" w przypadku braku takiego pokrycia.

Z moich obserwacji wynikało, że „geny-CNV" leżą średnio bliżej transpozonów niż „geny-NONVAR" **(Zmienko et al. 2020)**. Jednocześnie jednak odległość „transpozonów-CNV" do genów była wyższa niż „transpozonów-NONVAR". Wskazywało to na wypadkowy wpływ różnych – przeciwnie działających – czynników selekcyjnych na zmienność genów i transpozonów, a przez to na ostateczny rozkład tych elementów i regionów CNV w genomie. Aby dokładniej zaobserwować ten wpływ, przeanalizowałam wszystkie pary gen-transpozon, położone względem siebie w odległości nie większej niż 2 kpz. Okazało się, że takie pary mają najczęściej ten sam status (oba „-CNV" lub oba „-NONVAR"), co wydawało się dość naturalne, z uwagi na większą szansę pokrycia obu (lub nie) przez wspólny region AthCNV. Ciekawe jednak było to, że typowa odległość od centromerów sąsiadujących par „gen-CNV" :

„transpozon-CNV" była dużo mniejsza niż par „gen-NONVAR" : „transpozon-NONVAR". Ponadto, w parach o statusie „-NONVAR" zaobserwowałam wzbogacenie w geny zaangażowane w podstawowe procesy komórkowe, takie jak metabolizm kwasów nukleinowych, regulacja procesów rozrodczych czy aktywność czynników transkrypcyjnych, podczas gdy w parach o statusie „-CNV" występowała podwyższona reprezentacja genów białek pozakomórkowych i białek zaangażowanych w odpowiedzi obronne i kataboliczne. Obrazuje to jak liczne współistniejące ze sobą mechanizmy mogą wspólnie kształtować genomy i kierunki ich ewolucji.

Pomimo wyraźnie negatywnej korelacji pomiędzy gęstością wariantów AthCNV, a obecnością genów, w regionach tych zawierała się imponująca liczba ponad 5000 genów. Zbiór ten był wzbogacony w geny młode ewolucyjnie i gatunkowo-specyficzne. Zaobserwowałam też nadreprezentację w regionach CNV genów powstałych w wyniku tandemowych duplikacji oraz genów związanych z odpowiedziami obronnymi rośliny i reakcją na stres.

Zakres liczby kopii genów w 1060 ekotypach

Aby uzupełnić charakterystykę wariantów CNV, postanowiłam wykonać analizę liczby kopii wszystkich wariantów we wszystkich ekotypach. Genotypowanie na tak dużą skalę wymagało podejścia bioinformatycznego. Zastosowaliśmy w tym celu narzędzie Genome STRiP SVGenotyper, które oblicza liczbę kopii wskazanych regionów genomu dla poszczególnych próbek (tu: reprezentujących ekotypy), poprzez jednoczesną analizę statystyczną danych NGS dla wszystkich próbek. Przy takim podejściu, im większa jest badana populacja, tym większa jest poprawność przewidywań dla indywidualnych próbek. Metoda ta bazuje na analizie głębokości pokrycia i dla każdego regionu oblicza jedną wartość liczby kopii w danej próbce, niezależnie od długości tego regionu i równomierności jego pokrycia odczytami sekwencyjnymi. W przypadku złożonych CNV, takie uśrednione wartości mogą być zatem nieprecyzyjne. Przetestowałam i zaprezentowałam ten efekt na przykładzie wariantu AthCNV_7984, obejmującego scharakteryzowane przez nas wcześniej geny *MSH2*, *AT3G18530* i *AT3G18535*. W ekotypach, które, zgodnie z danymi eksperymentalnymi, wykazywały delecję *AT3G18530* i *AT3G18535*, ale miały niezmienioną liczbę kopii *MSH2*, program określał liczbę kopii całego regionu na poziomie pośrednim, tj. około 1. Dużo dokładniejsze przewidywania otrzymaliśmy genotypując mniejsze regiony, wyznaczone przez koordynaty każdego z genów. Uznałam zatem, że najbardziej praktycznym podejściem będzie ocena ilościowa każdego genu oddzielnie, bez względu na to, ile wariantów AthCNV i w jakim stopniu się z nim pokrywa. W konsekwencji wygenerowaliśmy pomiary liczby kopii dla ponad 27 tysięcy genów (polimorficznych i niepolimorficznych), w 1060 ekotypach. Bazując na uzyskanych wynikach porównałam następnie takie parametry jak zakres liczby kopii, wartości maksymalne, minimalne oraz zmienność w populacji, dla genów pokrywających się z regionami AthCNV oraz genów leżących poza tymi obszarami. Z tych porównań wynikało, że regiony AthCNV rzeczywiście obejmują geny o istotnie wyższej zmienności, co potwierdzało prawidłowość stworzonej przez nas mapy CNV w genomie.

Następnie, aby ocenić poprawność przypisanych bioinformatycznie wartości liczby kopii, wykonałam serię analiz MLPA. Objęły one 45 genów, w większości multiallelicznych, zlokalizowanych w różnych regionach AthCNV oraz dodatkowo 4 geny leżące poza regionami zmiennymi. Genotypowanie wykonałam na dużej grupie 314 ekotypów (30% całej badanej populacji), identycznej dla każdego genu, co pozwoliło mi na obiektywną ocenę zgodności analiz bioinformatycznych i eksperymentalnych. Okazała się ona bardzo wysoka, potwierdzając wiarygodność naszych przewidywań. Co ciekawe, dzięki dużej skali analizy udało mi się – jakkolwiek niezamierzenie – potwierdzić raportowane wcześniej błędy w identyfikacji nasion, zdeponowanych w międzynarodowym banku nasion i stanowiących część kolekcji 1001 Genomów. Doszłam przy tym do wniosku, że skład i genotyp regionów CNV może stanowić swoisty identyfikator genetyczny danego ekotypu i tym samym uzupełniać informacje uzyskane dzięki markerom SNP w analizach populacyjnych. Ta obserwacja skłoniła mnie do przetestowania użyteczności CNV jako markerów w badaniach asocjacyjnych, co opisuję poniżej.

Struktura światowej populacji rzodkiewnika odzwierciedlona w rozkładzie wariantów CNV

Z wszystkich znanych analiz genetycznych opartych o dane SNP wynika silne ustrukturyzowanie populacji rzodkiewnika (1001 Genomes Consortium 2016). Rozkład geograficzny ekotypów bardzo mocno pokrywa się z rozkładem podgrup genetycznych, przy czym prawie cała światowa populacja reprezentuje jeden klad. Do najbardziej odmiennych pod względem puli genetycznej ekotypów należą tzw. relikty epoki lodowcowej, tj. indywidualne rośliny, znajdowane przede wszystkim w regionie Morza Śródziemnego, reprezentujące inne klady niż ten, który najbardziej rozprzestrzenił się w Europie i Azji. Niedawno przeanalizowano stopień introgresji puli genów reliktowych w pozostałych genomach, co pozwoliło na zaproponowanie nowego – dwufalowego modelu rozprzestrzeniania się rzodkiewnika w Eurazji (Lee et al. 2017). Wg tego modelu, po pierwszej fali rozprzestrzeniania się polodowcowych reliktów z niewielkiego obszaru południowej Europy na resztę kontynentu, nastąpiła druga fala ekspansji, zainicjowana z obszaru Bałkanów, prawdopodobnie przy udziale człowieka. Ta druga fala spowodowała prawie całkowite zastąpienie innych reliktów młodszymi ewolucyjnie ekotypami, należącymi do wspólnego kladu. Pewne geny tamtych reliktów zachowały się w większym stopniu jedynie w genomach ekotypów zasiedlających północny i południowy kraniec kontynentu, gdzie ostrzejsze warunki klimatyczne wymagały utrzymania w puli dobrze zaadaptowanych wariantów genetycznych.

Postanowiłam sprawdzić, do jakiego stopnia struktura populacji i demograficzna historia rzodkiewnika może być odtworzona na podstawie wariantów AthCNV. W tym celu została wykonana analiza głównych składowych (PCA, ang. *principal component analysis*), w dwóch wersjach. W jednej wersji bazowaliśmy na danych AthCNV, podczas gdy w drugiej wersji – dla porównania – wykorzystaliśmy dane SNP **(Zmienko et al. 2020)**. Okazało się, że markery CNV dużo lepiej odzwierciedlają geograficzny rozkład ekotypów niż SNP. Dodatkowo analizy PCA z zastosowaniem markerów CNV uwydatniły wzajemną „bliskość genetyczną" ekotypów reliktowych z południa Europy z ekotypami z północnej Szwecji (czego nie byliśmy w stanie

wykazać bazując na danych SNP). Potwierdza to zaproponowany model dwufalowej migracji i stanowi dowód na użyteczność markerów CNV w badaniach populacyjnych.

Wpływ polimorfizmu liczby kopii na fenotyp i zastosowanie CNV w badaniu asocjacyjnym całego genomu

W badaniach GWAS, informację o rozłożeniu wariantów genetycznych (najczęściej SNP) w badanej puli genomów zestawia się z informacją na temat fenotypu poszczególnych osobników (np. cech morfologicznych czy wrażliwości na infekcję danym patogenem). Celem GWAS jest wykrycie wariantów, których zmienność jest najlepiej skojarzona ze zmiennością danej cechy. Pomaga to wyznaczyć najbardziej obiecujące miejsca w genomie do dalszych badań, a w efekcie często prowadzi do określenia źródłowej przyczyny zmienności fenotypowej. Od pewnego czasu wskazuje się, że zintegrowane analizy GWAS w oparciu o markery SNP i CNV mogą podnieść istotność statystyczną uzyskiwanych asocjacji lub wskazać nowe asocjacje (McCarroll 2008).

Postanowiłam sprawdzić, czy bezpośrednie zastosowanie markerów CNV (bez danych SNP) będzie wystarczające do wykazania asocjacji z fenotypem w przypadku, gdy to właśnie duplikacja / delecja genu jest przyczyną zmiany fenotypowej. Aby upodobnić sposób analizy do tego stosowanego dla SNP, poszczególnym genom przypisaliśmy odpowiedni status (duplikacja / delecja / brak zmiany) w każdym ekotypie. Następnie przeprowadziliśmy analizy asocjacji tak przygotowanych markerów genetycznych z danymi o rozkładzie 23 cech fenotypowych, pozyskanymi z publicznej bazy danych. Wśród badanych fenotypów znalazły się cechy nadwrażliwości na avrPphB, avrB oraz avrRpm1 – elicitory produkowane przez bakterię *Pseudomonas*. Udało się nam zaobserwować silną asocjację genu *RPS5* z nadwrażliwością na avrPphB oraz genu *RPM1* z nadwrażliwością na avrB i avrRpm1 **(Zmienko et al. 2020)**. Związek pomiędzy utratą genów *RPS5* i *RPM1*, a obniżeniem odporności roślin na infekcję szczepami *Pseudomonas*, produkującymi wymienione wyżej elicitory, jest potwierdzony wcześniejszymi badaniami. Nasza analiza dowiodła zatem, że (przynajmniej w przypadku delecji) zmienność liczby kopii genu potencjalnie można skojarzyć z wynikającymi z niej efektami fenotypowymi, stosując analizy GWAS. Obecnie kontynuuję ten wątek badań, analizując dużą liczbę dostępnych publicznie danych fenotypowych dla rzodkiewnika, pod kątem asocjacji ze zmianami liczby kopii genów.

W przypadku genów multiallelicznych, proste zakodowanie zmiany statusu jako delecja / duplikacja, bez wskazania wartości takiej zmiany, może być jednak niewystarczające dla uzyskania odpowiedniej istotności statystycznej w badaniach asocjacyjnych. Dlatego zainteresowało mnie również zbadanie zależności pomiędzy poziomem liczby kopii genów, a ich ekspresją. Uważa się, że w genomie człowieka multialleliczne CNV mają kilkukrotnie większy wpływ na dawkę genu niż zmiany bialleliczne i że przekłada się to istotnie na zmienność ich ekspresji (Handsaker et al. 2015). W pilotowym badaniu wybrałam jeden z genów multiallelicznych, *SEC10*. W jednej z wcześniejszych prac przedstawiono argumenty wskazujące na to, że dodatkowe kopie tego genu, obecne w niektórych ekotypach, mogą być aktywne funkcjonalnie (Vukašinović et al., 2014). W zgodzie z tymi doniesieniami, wykazałam

istnienie korelacji pomiędzy liczbą kopii, a ekspresją genu *SEC10*, i to zarówno na poziomie RNA, jak i białka **(Zmienko et al. 2020)**. Skłoniło mnie to do kompleksowego zbadania, w jakim stopniu CNV wpływają na zmienność transkryptomu rzodkiewnika. Badania te prowadzę obecnie jako jeden z wątków w ramach kierowanego przeze mnie projektu SONATA pt. „Udział polimorfizmu liczby kopii genów w naturalnym zróżnicowaniu odpowiedzi ekotypów *Arabidopsis thaliana na stres biotyczny*", finansowanego przez Narodowe Centrum Nauki, którego realizację rozpoczęłam w 2018 r.

Przeglądarka statusu liczby kopii genów u rzodkiewnika

Dostępność różnorodnych baz danych, narzędzi do analiz porównawczych oraz przeglądarek genomowych istotnie przyczynia się do rozwoju wiedzy o budowie i funkcjach genomów, a zastawienie danych uzyskanych na różnych poziomach ekspresji informacji genetycznej pozwala na uzyskanie bardziej kompletnego opisu złożonych zjawisk biologicznych. Nasze analizy wykazały, że atlas AthCNV wnosi wartościowe informacje na temat struktury i zmienności genomu modelowej rośliny dwuliściennej, jaką jest rzodkiewnik. Dlatego, aby umożliwić środowisku naukowemu korzystanie z tego zasobu, nasze wyniki udostępniliśmy w formie przeglądarki wariantów AthCNV, wykonanej wg mojego projektu i dostępnej pod adresem: **http://athcnv.ibch.poznan.pl/.** Generowane przez użytkownika interaktywne wykresy pozwalają na szybką ocenę liczby kopii poszczególnych genów i ich zmienności. Ponadto wszystkie zidentyfikowane przez nas warianty strukturalne (AthCNV, pozostałe CNV oraz duże indele) zostały umieszczone w wykazie stanowiącym suplement do publikacji **(Zmienko et al. 2020)**. O użyteczności tych danych może świadczyć fakt, że już opisano ich pierwsze wykorzystanie, w publikacji zdeponowanej w repozytorium bioRxiv (Steidele & Stam, 2020).

Podsumowanie i perspektywy

Podsumowując, za najważniejsze składowe mojego osiągnięcia naukowego uważam:

- zbudowanie od podstaw warsztatu badawczego i wdrożenie wachlarza metod eksperymentalnych i bioinformatycznych do identyfikacji CNV;
- stworzenie katalogu wariantów CNV oraz dużych indeli w genomie rzodkiewnika w oparciu o populację ponad tysiąca ekotypów;
- określenie liczby kopii genów w tych ekotypach i utworzenie publicznie dostępnego zasobu, dostarczającego informacji o ich zmienności w populacji rzodkiewnika;
- wykazanie użyteczności wariantów CNV w badaniach genetycznych i asocjacyjnych;
- opisanie wzajemnego wpływu genów i transpozonów na obserwowany rozkład wariantów strukturalnych w genomie rzodkiewnika;
- szczegółowe scharakteryzowanie regionu CNV wykazującego złożoną strukturę i wysoki stopień zmienności oraz powiązanie mechanizmu NAHR z rozpowszechnieniem się duplikacji i delecji genów *AT3G18530* i *AT3G18535* (a pośrednio prawdopodobnie także ze zmiennością genu *MSH2*) w populacji.

Na swojej stronie internetowej (https://www.1001genomes.org/about.html), Konsorcjum Projektu 1001 Genomów nakreśliło wizję realizacji następnych etapów badań, ujętych jako Projekt 1001G Plus. Wskazano w niej na głęboką potrzebę opisania złożonej zmienności strukturalnej w genomie rzodkiewnika w zakresie, jakiego nie umożliwiały dotąd metody NGS. Nowe możliwości, jakie daje sekwencjonowanie długich odczytów, dają nadzieję na przyspieszenie i udoskonalenie procesu asemblacji genomów, co w jeszcze większym stopniu przybliży nas do wiedzy na temat różnorodności genetycznej organizmów. Zgadzając się w pełni z powyższą wizją, w swoich obecnych badaniach również stosuję sekwencjonowanie długich odczytów, badając wybrane warianty CNV. Jednym z moich obecnych celów badawczych jest określenie związku pomiędzy polimorfizmem liczby kopii wybranych genów rzodkiewnika, a zróżnicowaniem odpowiedzi na stres biotyczny. Badania te prowadzę w ramach wspomnianego wcześniej projektu SONATA pt. „Udział polimorfizmu liczby kopii genów w naturalnym zróżnicowaniu odpowiedzi ekotypów *Arabidopsis thaliana na stres biotyczny*". Jednocześnie uważam, że uzyskane dzięki zastosowaniu danych NGS wyniki, składające się na moje osiągnięcie naukowe, stanowią istotny wkład w poznanie zmienności strukturalnej genomu rzodkiewnika i tworzą użyteczny zasób do dalszych badań w tym temacie.

Literatura uzupełniająca

1001 Genomes Consortium (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. Cell 166: 481-491. doi:10.1016/j.cell.2016.05.063.

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet. 12: 363-376. doi: 10.1038/nrg2958.

Birchler JA, Veitia RA. (2010) The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. 186: 54-62. doi:10.1111/j.1469-8137.2009.03087.x.

Bourque G, Burns KH, Gehring M, Gorbunova, V, Seluanov, A et al. (2018) Ten things you should know about transposable elements. Genome Biol. 19: 199. doi:10.1186/s13059-018-1577-z.

Cao J, Schneeberger K, Ossowski S, Günther, T, Bender, S et al. (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet. 43: 956-963. doi:10.1038/ng.911.

Handsaker RE, Van Doren V, Berman JR, Genovese, G, Kashin, S et al. (2015) Large multiallelic copy number variations in humans. Nat Genet. 47: 296-303. doi:10.1038/ng.3200.

Iafrate AJ, Feuk L, Rivera MN, et al. (2004) Detection of large-scale variation in the human genome. Nat Genet. 36: 949-951. doi:10.1038/ng1416.

Kehrer-Sawatzki H, Bengesser K, Callens T, et al. (2014) Identification of large NF1 duplications reciprocal to NAHR-mediated type-1 NF1 deletions. Hum Mutat. 35: 1469-1475. doi:10.1002/humu.22692.

Lee CR, Svardal H, Farlow A, Exposito-Alonso, M, Ding, W et al. (2017) On the post-glacial spread of human commensal *Arabidopsis thaliana*. Nat Commun. 8: 14458. doi:10.1038/ncomms14458

Liu P, Carvalho CM, Hastings PJ, Lupski JR. (2012) Mechanisms for recurrent and complex human genomic rearrangements. Curr Opin Genet Dev. 22: 211-220. doi:10.1016/j.gde.2012.02.012.

Marcinkowska M, Wong KK, Kwiatkowski DJ, Kozlowski P. (2010) Design and generation of MLPA probe sets for combined copy number and small-mutation analysis of human genes: EGFR as an example. ScientificWorldJournal 10 :2003-2018. doi:10.1100/tsw.2010.195.

McCarroll SA. (2008) Extending genome-wide association studies to copy-number variation. Hum Mol Genet. 17: R135-R142. doi:10.1093/hmg/ddn282.

Perry GH, Dominy NJ, Claw KG, Lee, AS, Fiegler, H et al. (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet. 39: 1256-1260. doi:10.1038/ng2123.

Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc, C, Martienssen, RA et al. (2016) The Arabidopsis thaliana mobilome and its impact at the species level. Elife. 5: e15716. doi:10.7554/eLife.15716.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH et al. (2006) Global variation in copy number in the human genome. Nature 444: 444-454. doi: 10.1038/nature05329.

Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G. (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res. 30: e57. doi:10.1093/nar/gnf056.

Springer NM, Ying K, Fu Y, et al. (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet. 5: e1000734. doi:10.1371/journal.pgen.1000734.

Steidele C, Stam R (2020) Multi-omics approach highlights differences between functional RLP classes in *Arabidopsis thaliana*. bioRxiv. doi:10.1101/2020.08.07.240911.

Sudmant PH, Rausch T, Gardner EJ, Handsaker, RE, Abyzov, A et al. (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526: 75-81. doi:10.1038/nature15394.

Usher CL, McCarroll SA. (2015) Complex and multi-allelic copy number variation in human disease. Brief Funct Genomics. 14: 329-338. doi:10.1093/bfgp/elv028.

Vukašinović N, Cvrčková F, Eliáš M, Cole R, Fowler JM et al. (2014) Dissecting a hidden gene duplication: the *Arabidopsis thaliana SEC10* locus. PLoS One. 9: e94077. doi:10.1371/journal.pone.0094077.

Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 14 Suppl 11: S1. doi:10.1186/1471-2105-14-S11-S1.

I. Przebieg kariery naukowej i pozostałe osiągnięcia naukowo-badawcze

Pogrubioną czcionką oznaczyłam publikacje wchodzące w skład mojego dorobku naukowego; ich lista znajduje się w Załączniku 5.

Wpływ oddziaływania rizobiów na ekspresję informacji genetycznej w roślinach motylkowatych

W pracę naukową zaangażowałam się na 3 roku studiów magisterskich, dołączając do Koła Młodych Przyrodników Zespołu Biologii Molekularnej i Biotechnologii UAM, w ramach którego badaliśmy genom wiroida PSTVd (potato spindle tuber viroid) – pasożytniczej cząsteczki RNA oraz zgłębialiśmy bardzo nowatorskie wówczas zagadnienia związane z potranskrypcyjnym wyciszaniem ekspresji genów w roślinie *Nicotiana benthamiana*, pod opieką dr Krzysztofa Wypijewskiego. Prace eksperymentalne prowadzone w ramach Koła zachęciły mnie do wyjazdu na półroczne stypendium w ramach programu wymiany studentów Socrates-Erasmus (X 1999 - III 2000) na Uniwersytet Arystotelesa w Salonikach (Grecja). Jego głównym (i zrealizowanym) celem było opanowanie techniki badania aktywności nukleolitycznej ekstraktów roślinnych metodą trawienia DNA i RNA w żelu poliakrylamidowym, stosowanej w tamtejszym Zakładzie Biochemii, w zespole prof. Trianosa Yupsanisa. Metodę tą zastosowałam do zbadania własności transgenicznych roślin *N. benthamiana*, które były transformowane wektorem zawierającym cDNA białka kapsydu wirusa ospowatości śliwy, a przez to (poprzez efekt wyciszania genów) odpornych na zakażenie tym wirusem. Konsekwencją tego stażu była propozycja kontynuowania prac w zespole prof. Yupsanisa w ramach studiów doktoranckich, z której jednak nie skorzystałam, ponieważ bardziej pociągała mnie zgłębiana równolegle tematyka symbiozy roślin motylkowatych z bakteriami wiążącymi azot (rizobiami). Badania poświęcone temu zagadnieniu prowadziłam w ramach pracowni magisterskiej w latach 1998-2000, a później w ramach studiów doktoranckich w latach 2000-2006, w ICHB PAN.

Moje ówczesne zainteresowania dotyczyły wpływu oddziaływania rizobiów na ekspresję informacji genetycznej rośliny-gospodarza i wykształcenie brodawek korzeniowych u łubinu wąskolistnego (*Lupinus angustifolius*) oraz łubinu żółtego (*Lupinus luteus*), które wówczas stanowiły modele badawcze w naszej pracowni. W toku realizacji pracy magisterskiej włączyłam się w badania mające na celu określenie sekwencji i charakterystykę funkcjonalną kilku genów z łubinu żółtego: L-asparaginazy (**Borek et al., 2004, Eur. J. Biochem.,** Zał. 5: A.27), genu hydrolazy S-adenozylohomocysteiny oraz genów *ENOD40-1* i *ENOD40-2*. Geny *ENOD40* występują w roślinach motylkowatych, w jednej bądź dwóch kopiach. Badania mutantów pokazywały, że odgrywają one kluczową, choć mało wyjaśnioną rolę w rozwoju brodawek symbiotycznych. Z uwagi na dużą konserwatywność struktury RNA i brak odkrytych produktów białkowych uważano, że *ENOD40* może funkcjonować jako niekodujący RNA i pełnić funkcje regulatorowe. Z drugiej strony, istniały doniesienia pokazujące aktywność translacyjną krótkich otwartych ramek odczytu, również zakonserwowanych w jego sekwencji. Przeprowadziliśmy zatem kompleksową analizę sekwencji genu i promotora oraz aktywności na poziomie RNA obu łubinowych *ENOD40*, demonstrując ich aktywność na wczesnych etapach symbiozy oraz pokazując specyficzność ekspresji *ENOD40-1* dla tego procesu

(**Podkowiński et al., 2009, Acta Biochim. Pol.,** Zał.5: A.21). Dodatkowo zbadaliśmy drugorzędową strukturę RNA *ENOD40-1*, a stosując metodę transformacji i przejściowej ekspresji fuzyjnego genu reporterowego w liściach tytoniu pokazaliśmy aktywność translacyjną dwóch krótkich konserwatywnych ramek odczytu, w szczególności tzw. ORF A o długości 12 aminokwasów. Nasze wyniki pasowały do zaproponowanego wówczas modelu dualnej roli *ENOD40* – jako źródła regulatorowej cząsteczki RNA oraz genu kodującego krótki peptyd. Warto zauważyć, że *ENOD40* (którego homologi występują również w wielu roślinach niemotylkowatych) do dziś uważany jest za klasyczny przykład genu dwufunkcyjnego w roślinach, a jego rola nadal pozostaje enigmatyczna.

Następnie, w ramach realizacji pracy doktorskiej, którą przygotowałam pod kierunkiem prof. dr. hab. Andrzeja B. Legockiego, szukałam genów nowych potencjalnych nodulin, czyli białek zaangażowanych w procesy symbiotyczne, przeszukując w tym celu znormalizowaną bibliotekę około 5000 klonów cDNA łubinu wąskolistnego, metodą hybrydyzacji różnicowej. Wyselekcjonowane klony były poddawane sekwencjonowaniu, bioinformatycznej adnotacji funkcjonalnej oraz badaniom profilu ekspresji w trakcie symbiozy, z zastosowaniem własnoręcznie przeze mnie sporządzonych nylonowych makromacierzy cDNA (**Podkowiński et al. 2001, Zeszyty Naukowe AR we Wrocławiu,** Zał. 5: A.29**; Podkowiński et al. 2006, Journal of Fruit and Ornamental Plant Research,** Zał. 5: A.24). Te badania doprowadziły nas do poznania częściowej sekwencji kilkudziesięciu genów potencjalnych nodulin, w tym specyficznych dla rodzaju *Lupinus* (**Kisiel et al., 2004, w: Understanding the Plant Genome,** Zał. 5: B.2). Łącznie w toku prac nad doktoratem uczestniczyłam w ustaleniu sekwencji prawie 500 fragmentów cDNA genów łubinowych, tzw. sekwencji EST (ang. *Expressed Sequence Tags*). Zostały one zdeponowane w publicznej bazie danych GenBank (https://www.ncbi.nlm.nih.gov/nuccore ID: BG104136-BG104145, BG149099-BG149166, BG153881-BG154166, DV468649-DV468774), gdzie aż do czasu rozpowszechnienia technik NGS, stanowiły znaczący procent sekwencji rodzaju *Lupinus* – taksonu niezwykle interesującego z ewolucyjnego punktu widzenia. Jestem również współautorką pracy przeglądowej poświęconej ewolucji zdolności do symbiotycznego wiązania azotu przez rośliny motylkowate (**Żmieńko et al., 2009, w: Od syntezy chemicznej do biologii syntetycznej,** Zał. 5: B.1).

Wpływ warunków stresowych na ekspresję informacji genetycznej w komórkach

Po doktoracie zamierzałam kontynuować badania ekspresji informacji genetycznej w roślinach, ale już w skali całego transkryptomu. Zbiegło się to w czasie z uruchomieniem w ICHB PAN jednej z pierwszych w Polsce pracowni mikromacierzowych - Centrum Doskonałości w Technologiach Kwasów Nukleinowych (CENAT). W pracowni tej byłam jedną z osób odpowiedzialnych za wdrożenie techniki mikromacierzy, a także zastosowanie jej do badania transkryptomów roślin. Był to dla mnie również czas bardzo intensywnego macierzyństwa (na przestrzeni lat 2006 – 2010 urodziłam kolejno troje dzieci), co czasowo ograniczyło mój potencjał w rozwijaniu samodzielnej kariery naukowej. Moja ówczesna działalność naukowa polegała na współpracy przy licznych projektach badawczych, realizowanych w CENAT, które

pokrótce omawiam poniżej. Mój udział w tych projektach polegał na globalnej analizie transkryptomów roślinnych w różnych warunkach, takich jak odpowiedź na stres czy procesy starzenia i nakreśleniu tła do dalszych badań lub też weryfikacji wcześniejszych hipotez. Moim zadaniem, jako eksperta, było nie tylko uzyskanie i opracowanie wyników mikromacierzowych, ale również ich analiza i interpretacja, w kontekście badanego układu. Pozwoliło mi to uzyskać doświadczenie i rozległą wiedzę na temat odpowiedzi roślin na różnorodne czynniki. Technika mikromacierzy była wówczas popularna na świecie, ale stosunkowo nowa w kraju, co wymagało ode mnie również samokształcenia w zakresie analiz bioinformatycznych. Dlatego regularnie uczestniczyłam w specjalistycznych kursach, zdobywając wiedzę i umiejętności adekwatne do technologii, którą przyszło mi się posługiwać (wymienione w punkcie E Autoreferatu).

Badania w których uczestniczyłam w ramach współpracy z prof. Grażyną Dobrowolską z Instytutu Biochemii i Biofizyki PAN w Warszawie pokazały między innymi, że działanie stresu zasolenia, suszy i metali ciężkich na rzodkiewnik oraz tytoń powoduje globalne zmiany w ekspresji genów, takie jak uruchomienie szlaków sygnalizacji i odpowiedzi obronnych oraz zahamowanie procesów fotosyntezy i modyfikację struktury ścian komórkowych. Zbadaliśmy również rolę genów kodujących białka kinaz SNRK2.4 i SNRK2.10 w odpowiedzi na stres, wywołany podaniem metali ciężkich. Ustaliliśmy, że kinazy te, powszechnie znane jako białka zaangażowane w odpowiedź na stres osmotyczny, są również zaangażowane w sygnalizację indukowaną obecnością jonów kadmu (Cd^{2+}), wpływając na akumulację wolnych rodników oraz przyczyniając się do zahamowania wzrostu korzeni w stresie (**Kulik et al., 2012**, **Plant Physiol.,** Zał. 5: A.18).

Ważnym wątkiem moich ówczesnych badań były analizy transkryptomiczne procesów starzeniowych, zachodzących w jęczmieniu pod wpływem zaciemnienia, prowadzone we współpracy z dr hab. Ewą Sobieszczuk-Nowicką z Instytutu Biologii Eksperymentalnej UAM. Ich celem nadrzędnym było zbadanie roli poliamin (związków zaangażowanych w modyfikację białek) w proces rozpadu starzejącego się liścia oraz regulacji odwracalności tego procesu. W toku prac udało mi się scharakteryzować sekwencję jęczmiennej transglutaminazy – enzymu odpowiedzialnego za potranslacyjne wiązanie poliamin do białek. W oparciu o analizy lokalizacji *in situ*, badania biochemiczne i proteomiczne zaproponowaliśmy udział transglutaminaz w oksydacyjnej deaminacji poliamin w starzejącym się chloroplaście (**Sobieszczuk-Nowicka et al., 2015, Amino Acids,** Zał. 5: A.13). Zbadałam również profile ekspresji genów jęczmienia podczas starzenia i zmiany tych profili w czasie, aż do 10 dnia od momentu zaciemnienia, stosując hybrydyzację do mikromacierzy oligonukleotydowych (**Zmienko et al., 2015, Genomics Data**, Zał. 5: A.12). Uzyskane dane transkryptomiczne pozwoliły nam na kompleksową ocenę równowagi pomiędzy anabolizmem i katabolizmem poliamin w trakcie starzenia, a także wykazanie związku tych procesów z produkcją wolnych rodników, kwasu γ-aminomasłowego i etylenu, co potwierdziliśmy również analizami fizjologicznymi i biochemicznymi. W efekcie pokazaliśmy, że katabolizm poliamin stanowi jeden z czynników regulujących przebieg starzenia (**Sobieszczuk-Nowicka et al., 2016, Front. Plant Sci.,** Zał. 5: A.10).

Kolejnym podejmowanym w moich badaniach tematem była odpowiedź roślin na stres biotyczny. W ramach współpracy z dr hab. Aleksandrą Obrępalską-Stęplowską z Instytutu Ochrony Roślin – Państwowego Instytutu Badawczego w Poznaniu badałam wpływ satelitarnego RNA – patogenicznej cząsteczki subwirusowej, zależnej od obecności wirusa pomocniczego, na przebieg infekcji wirusem karłowatości orzecha ziemnego (PSV), na poziomie transkryptomu. Nasze analizy pokazały, że wpływ ten jest zależny od szczepu wirusa pomocniczego i choć zwykle obecność satelitarnego RNA łagodzi symptomy związane z obecnością wirusa, to w badanym przez nas układzie zaostrzało to przebieg infekcji oraz powodowało intensyfikację odpowiedzi rośliny (**Obrepalska-Steplowska et al., 2018**, **Viruses**, Zał. 5: A.3). W szczególności obserwowaliśmy zmiany w ekspresji genów związanych z procesami fosforylacji, wiązaniem ATP oraz genów białek zlokalizowanych w błonie komórkowej.

Mikromacierze oligonukleotydowe wykorzystałam również do analizy transkryptomu kapusty – rośliny podatnej na infekcję pasożytniczym grzybem *Alternaria brassicicola*, który wywołuje chorobę zwaną czernią. W ramach współpracy z dr Violettą Macioszek z Katedry Biologii i Ekologii Roślin Uniwersytetu w Białymstoku (wcześniej Katedry Genetyki Ogólnej, Biologii Molekularnej i Biotechnologii Roślin Uniwersytetu Łódzkiego) badaliśmy przebieg wczesnych stadiów infekcji (do 48 godzin po inokulacji grzybem). Analizy transkryptomiczne, ultramikroskopowe oraz wyniki pomiarów wydajności aparatu fotosyntetycznego w infekowanych roślinach wyraźnie wskazywały na zahamowanie procesu fotosyntezy, już po 12 godzinach od inokulacji i dalsze stopniowe nasilanie się tego efektu. Obniżenie wydajności fotosyntezy może stanowić element wczesnej odpowiedzi obronnej rośliny, ale możliwy jest również udział nieokreślonego czynnika wirulencji, produkowanego przez grzyb (**Macioszek et al.,** manuskrypt w trakcie recenzji). Wskazuje to na konieczność dalszych badań tego nie analizowanego wcześniej elementu wczesnych interakcji kapusta – *A. brassicicola*.

Dzięki powstaniu Europejskiego Centrum Bioinformatyki i Genomiki (ECBiG) – centrum badawczego ICHB PAN oraz Politechniki Poznańskiej, od 2014 r. umieszczonego na Polskiej Mapie Infrastruktury Badawczej, w którego organizacji aktywnie uczestniczyłam (zob. Zał. 5: O), rozszerzyłam zakres swoich umiejętności praktycznych o technikę NGS, co pozwoliło na mój udział w kolejnych ciekawych projektach badawczych. Jeden z nich, prowadzony we współpracy dr hab. Pauliną Jackowiak z ICHB PAN, był poświęcony zbadaniu transkryptomu w modelu kultury komórek ludzkich, zainfekowanych wirusem zapalenia wątroby typu C. Badania w tym projekcie skupiały się wokół mało dotąd poznanych pofragmentowanych RNA. Pokazaliśmy, że powstają one w wyniku degradacji konkretnych niekodujących RNA, konstytutywnie produkowanych w komórce, takich jak tRNA. W badanym przez nas układzie produkty degradacji RNA stanowiły znaczącą frakcję transkryptomu, a profil ich akumulacji w zainfekowanych wirusem komórkach był różny od komórek kontrolnych, wskazując na potencjalne znaczenie funkcjonalne, np. regulatorowe, tych cząsteczek (**Hojka-Osinska et al., 2016, Acta Biochim Pol.,** Zał. 5: A.7; **Jackowiak et al., 2017, BMC Genomics,** Zał. 5: A.5).

Wkład w rozwijanie technik badania transkryptomów

Specyfika projektów, w które byłam zaangażowana, często rodziła konieczność zastosowania indywidualnych narzędzi badawczych, zwłaszcza że jeszcze na początku XXI w. komercyjne mikromacierze DNA były dostępne zaledwie dla kilku gatunków roślin. Stanowiło to duże wyzwanie w moich badaniach. Testowane przeze mnie rozwiązania obejmowały m. innymi hybrydyzację cDNA do mikromacierzy pokrewnych gatunków, ale także projektowanie własnych mikromacierzy. W tamtym okresie nawiązałam trwałą współpracę z zespołem Instytutu Informatyki Politechniki Poznańskiej (gdzie w latach 2014 -2020 byłam zatrudniona na stanowisku adiunkta). Jednym z naszych osiągnięć było stworzenie od podstaw nowoczesnej oligonukleotydowej mikromacierzy dachówkowatej PlasTiArray, dedykowanej kompleksowej analizie chloroplastowych genomów roślinnych, na potrzeby współpracy z dr hab. Wojciechem Pląderem z Katedry Genetyki Hodowli i Biotechnologii Roślin Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie (**Żmieńko et al. 2011, Plant Methods**, Zał. 5: A.19). Z kolei w przypadku *N. benthamiana* wykorzystałam dostępność opublikowanych w owym czasie sekwencji transkryptomu dla tej rośliny do zaprojektowania mikromacierzy. Początkowo opracowałam testową mikromacierz wysokiej gęstości (244 tysiące sond), która posłużyła do eksperymentalnego ustalenia właściwej orientacji kontigów (jako że opublikowane sekwencje nie miały ustalonej orientacji). Na podstawie tych wyników zaprojektowałam właściwą mikromacierz oligonukleotydową, zawierającą 105 tysięcy sond, którą zastosowaliśmy następnie we wspomnianych wcześniej badaniach infekcji rośliny wirusem karłowatości orzecha ziemnego. Co ciekawe, wyniki uzyskane na testowej mikromacierzy pozwoliły mi na zidentyfikowanie nieprawidłowo złożonych kontigów w opublikowanej wersji v.5. transkryptomu *N. benthamiana*, na podstawie niezgodności profili ekspresji, uzyskanych sondami mapującymi się do różnych końców tych kontigów (**Goralski et al., 2016, Plant Methods,** Zał. 5: A.8).

Z kolei we wspomnianym wcześniej projekcie poświęconym poliaminom, oprócz mikromacierzy DNA, zastosowałam także technikę analizy ekspresji genów metodą emulsyjnego PCR. Jak wspominałam, ddPCR pozwala na bezwzględną ocenę ilości matrycy w badanych próbkach. Dlatego powszechnie uważano, że zastosowanie genów referencyjnych o konstytutywnej ekspresji (ang. *hauskeeping genes*) do normalizacji wyników jest w przypadku tej metody zbędne. Na marginesie badań poliamin wykorzystałam model starzejącego się liścia jęczmienia do pokazania, że normalizacja wyników emulsyjnego PCR z zastosowaniem genów referencyjnych jest często pożądanym zabiegiem i wręcz powinna być przeprowadzona w przypadku eksperymentów obciążonych wysoką zmiennością, np. z powodu obecności inhibitorów w próbce lub częściowej degradacji materiału (**Zmienko et al., 2015, Plos One**, Zał. 5: A.11). Analiza cytowań pracy pokazuje, że ten pogląd spotkał się z akceptacją środowiska użytkowników techniki ddPCR.

Uczestniczyłam również w opracowaniu zestawu genów referencyjnych i metod analizy ekspresji genów w projekcie poświęconym badaniu wpływu osmokondycjonowania na kiełkowanie nasion rzepaku, we współpracy z prof. dr hab. Małgorzatą Garnczarską z Instytutu

Biologii Eksperymentalnej UAM (**Kubala et al., 2015, Plant Science**, Zał. 5: A.14). Ponadto brałam udział w opracowaniu metod analizy transkryptomu kukurydzy. Ta ostatnia praca była częścią projektu poświęconego badaniu odporności odmian kukurydzy na działanie herbicydów i była prowadzona we współpracy z prof. dr hab. Tomaszem Twardowskim z ICHB PAN – jestem współautorką powiązanego z tym działaniem krajowego patentu (**PL 216720 B1**, Zał. 5 K.1).

Z kolei w pracy zespołu Instytutu Informatyki Politechniki Poznańskiej poszukiwaliśmy bioinformatycznych sposobów wydobycia nowych informacji biologicznych poprzez połączenie danych uzyskanych różnymi metodami. Zainteresowaliśmy się między innymi analizami ko-ekspresji genów i ich integracją z wynikami analiz oddziaływań białko-białko oraz rozwojem dedykowanych temu narzędzi. Ten kierunek badań, rozwijany we współpracy z grupą F. Giovaniego z National Research Council of Italy, przyniósł efekt w postaci stworzenia oprogramowania CLAIM, pozwalającego na przewidywanie funkcji nieznanych białek na podstawie ich ko-ekspresji i wzajemnych oddziaływań (**Santoni et al., 2014, OMICS-a Journal of Integrative Biology**, Zał. 5: A.17).

Nabytą wiedzę i umiejętności w zakresie stosowania różnorodnych technik genomiki funkcjonalnej wykorzystywałam także przy tworzeniu licznych prac przeglądowych im poświęconych (**Żmieńko et al., 2001, BioTechnologia**, Zał. 5: A.20, **Kisiel et al., 2003, Biotechnologia**, Zał. 5: A.28; **Kisiel et al., 2004, Kosmos**, Zał. 5: A.26; **Kisiel et al., 2005, Acta Physiol. Plantarum**, Zał. 5: A.25; **Żmieńko et al., 2008, Biotechnologia**, Zał. 5: A.23; **Satyr i Zmienko, 2020, Postepy biochemii,** Zał. 5: A.1). We wszystkich wymienionych pracach jestem pierwszą i/lub korespondencyjną autorką. Ostatnia praca jest poświęcona sekwencjonowaniu nanoporowemu, czyli technice sekwencjonowania długich odczytów, którą obecnie stosuję w swoich badaniach zmienności strukturalnej genomów roślinnych. Jestem również gościnną edytorką tematycznej kolekcji artykułów poświęconej tej tematyce, tworzonej obecnie dla czasopisma Frontiers in Plant Science (https://www.frontiersin.org/research-topics/16585/resolving-the-complexity-of-plant-genomes-and-transcriptomes-with-long-reads).

Zmienność strukturalna genomów

Obserwując efekty działania różnorodnych czynników zewnętrznych na ekspresję informacji genetycznej zaczęłam zadawać sobie pytanie o naturę i udział czynników wewnętrznych, kształtujących genomy eukariotyczne oraz wpływających na zmienność strukturalną. Dzięki współpracy z prof. dr hab. Piotrem Kozłowskim z ICHB PAN odkryłam złożoność tego problemu oraz nauczyłam się techniki MLPA, stosowanej w jego zespole, między innymi do analizy mutacji ludzkiego genu BARD1, uważanego za jeden z potencjalnych czynników genetycznych zwiększających ryzyko zachorowania na nowotwór piersi (**Klonowska et al., 2015, Sci. Rep.,** Zał. 5: A.15, **Klonowska et al., 2016, Oncotarget,** Zał. 5: A.9). Umiejętności te następnie przeniosłam z powodzeniem na grunt badań zmienności strukturalnej genomu rzodkiewnika, a uzyskane w ten sposób wyniki stanowią moje główne osiągnięcie naukowe, które przedstawiłam w sekcji H Autoreferatu.

J. Omówienie osiągnięć dydaktycznych i udział w popularyzacji nauki

W związku z podjęciem pracy na stanowisku adiunkta na Politechnice Poznańskiej, w latach 2014-2020 realizowałam działalność dydaktyczną, skupiającą się głównie na (współ-) prowadzeniu przedmiotów obieralnych o tematyce biologicznej i bioinformatycznej, dla studentów I i II stopnia makrokierunku Bioinformatyka oraz innych kierunków (Zał. 5 sekcja L). Byłam opiekunką naukową ośmiu studentek, realizujących praktyki i krótkoterminowe staże w ECBiG, w latach 2013-2017 (Zał. 5 sekcja N). Byłam również promotorką przy realizacji jednej ukończonej pracy magisterskiej oraz sześciu prac licencjackich (Zał. 5 sekcja M). Warto zaznaczyć, że realizacja dwóch prowadzonych przeze mnie prac licencjackich, zaowocowała współautorstwem studentek w publikacjach naukowych (Goralski et al. 2016, Plant Methods, Zał. 5: A.8 – współautorstwo dyplomantki P. Sobieszczańskiej; Zmienko et al. 2016, BMC Genomics, Zał. 5: A.6 – współautorstwo dyplomantki M. Szymańskiej). Z ramienia ICHB PAN byłam opiekunką naukową / promotorką pomocniczą w dwóch zakończonych przewodach doktorskich oraz jestem promotorką pomocniczą w kolejnym, otwartym przewodzie doktorskim. Ponadto jestem opiekunką naukową doktorantki prowadzącej badania w ramach kierowanego przeze mnie projektu SONATA.

W ramach działalności organizacyjnej i popularyzującej naukę wspomagałam m. innymi organizację Drzwi Otwartych na Politechnice Poznańskiej pod hasłem „Dzień dla Dziewczyn", w latach 2016-2018. Ponadto nawiązałam kilkuletnią (2014-2018) współpracę z kadrą Społecznej Szkoły Podstawowej im. Edwarda hr. Raczyńskiego nr 2 i Społecznego Gimnazjum STO w Poznaniu, polegającą na organizacji i prowadzeniu praktycznych warsztatów w ECBiG dla uczniów najstarszych klas, w czasie których poznawali oni techniki analizy DNA. Udzieliłam również wywiadu na temat życia matki-naukowca, który ukazał się w numerze 4/2011 magazynu PAN „Academia", poświęconym nauce kobiet.

Poznań, 13.10.2020

Agnieszka Żmieńko