Załącznik 2

Autoreferat

1. Imię i Nazwisko: Michał Szcześniak

2. Posiadane dyplomy, stopnie naukowe/ artystyczne – z podaniem nazwy, miejsca i roku ich uzyskania oraz tytułu rozprawy doktorskiej.

Doktor nauk biologicznych w zakresie biotechnologii – 21 czerwca 2013 r; Wydział Biologii, Uniwersytet im. Adama Mickiewicza w Poznaniu. Tytuł rozprawy doktorskiej: *Nowe metody identyfikacji mikroRNA* (praca obroniona z wyróżnieniem).

Promotor: prof. dr hab. Izabela Makałowska, Pracownia Bioinformatyki, Uniwersytet im. Adama Mickiewicza w Poznaniu
Recenzenci: prof. dr hab. Włodzimierz J. Krzyżosiak (Instytut Chemii Bioorganicznej, Polska Akademia Nauk); prof. dr hab. Piotr Zielenkiewicz (Instytut Biochemii i Biofizyki, Polska Akademia Nauk)

Magister biotechnologii: 24 czerwca 2009; Wydział Biologii, Uniwersytet im. Adama Mickiewicza w Poznaniu; promotor: prof. dr hab. Zofia Szweykowska-Kulińska.

Magister bioinformatyki: 24 września 2010; Wydział Biologii, Uniwersytet im. Adama Mickiewicza w Poznaniu; promotor: prof. dr hab. Izabela Makałowska

3. Informacje o dotychczasowym zatrudnieniu w jednostkach naukowych/ artystycznych.

Od 1 października 2013: adiunkt na Wydziale Biologii Uniwersytetu im. Adama Mickiewicza w Poznaniu, w zespole kierowanym przez prof. dr hab. Izabelę Makałowską.

4. Wskazanie osiągnięcia[1] wynikającego z art. 16 ust. 2 ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U.

2017 r. poz. 1789):

Załącznik 2

a) tytuł osiągnięcia naukowego/artystycznego,

Identyfikacja długich niekodujących RNA i badanie ich funkcji pełnionych w kontekście oddziaływań RNA:RNA

b) (autor/autorzy, tytuł/tytuły publikacji, rok wydania, nazwa wydawnictwa, recenzenci wydawniczy),

1. Bryzghalov O*, Szcześniak MW*, Makałowska I. (2016) Retroposition as a source of antisense long non-coding RNAs with possible regulatory functions. Acta Biochim Pol. 2016;63(4):825-833.
   (wydawca: Polskie Towarzystwo Biochemiczne; **IF**: 1.159, **pkt MNiSW**: 15)
2. Szcześniak MW, Bryzghalov O, Ciomborowska-Basheer J, Makałowska I. (2019) CANTATAdb 2.0: Expanding the Collection of Plant Long Noncoding RNAs. Methods Mol Biol. 2019;1933:415-429.
   (wydawnictwo: Springer; IF: brak, pkt MNiSW: brak)
3. Szczesniak MW, Kabza M, Karolak JA, Rydzanicz M, Nowak DM, Ginter-Matuszewska B, Polakowski P, Ploski R, Szaflik JP, Gajecka M. (2017) KTCNlncDB-a first platform to investigate lncRNAs expressed in human keratoconus and non-keratoconus corneas. Database (Oxford). 2017 Jan 10;2017.
   (Wydawnictwo: Oxford University Press; **IF**: 3,978, **pkt MNiSW**: 40)
4. Szcześniak MW, Makałowska I. (2016) lncRNA-RNA Interactions across the Human Transcriptome. PLoS One. 2016 Mar 1;11(3):e0150353.
   (Wydawnictwo: Public Library of Science; **IF**: 2.806, **pkt MNiSW**: 40)
5. Szcześniak MW, Rosikiewicz W, Makałowska I. (2016) CANTATAdb: A Collection of Plant Long Non-Coding RNAs. Plant Cell Physiol. 2016 Jan;57(1):e8.
   (Wydawnictwo: Oxford University Press; **IF**: 4.76, **pkt MNiSW**: 40)
6. Wanowska E, Kubiak MR, Rosikiewicz W, Makałowska I, Szcześniak MW. (2018) Natural antisense transcripts in diseases: From modes of action to targeted therapies. Wiley Interdiscip Rev RNA. 2018 Mar;9(2).
   (Wydawnictwo: Wiley; **IF**: 5,844, **pkt MNiSW**: 35)

* Wspólne pierwsze autorstwo

c) omówienie celu naukowego/artystycznego ww. pracy/prac i osiągniętych wyników wraz z omówieniem ich ewentualnego wykorzystania.

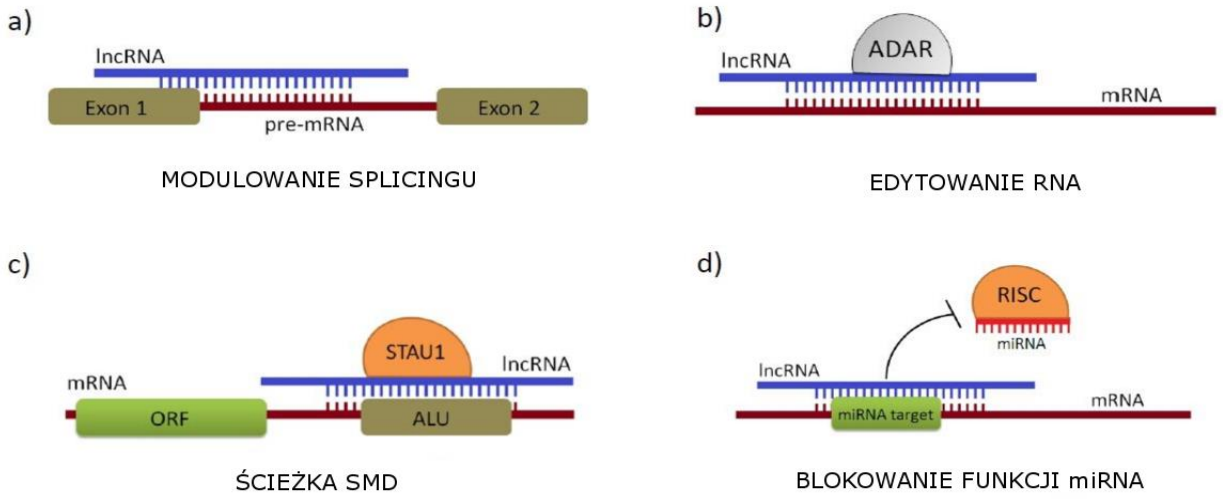
1. Przedstawienie przedmiotu badań oraz celów naukowych

Moje zainteresowanie transkryptomiką i niekodującymi RNA rozpoczęło się w czasie realizacji projektu doktorskiego, którego motywem przewodnim była identyfikacja i adnotacja cząsteczek mikroRNA. Z kolei w ostatnich latach skoncentrowałem się na biologii długich niekodujących RNA. Długie niekodujące RNA (lncRNA, ang. *long non-coding RNA*) stanowią liczną klasę transkryptów, odkrytych u roślin i zwierząt, których wspólną cechą jest brak zdolności kodowania białka oraz długość przekraczająca arbitralnie przyjętą granicę 200 nukleotydów. Baza danych ENSEMBL kolekcjonuje 57,591 lncRNA człowieka, stanowiących aż 28.75% wszystkich transkryptów (Zerbino et al., 2018). Niemniej jednak dotychczas udało się scharakteryzować funkcjonalnie zaledwie 1% tych cząsteczek, istnieje zatem dość pilna potrzeba poszerzania naszej wiedzy na temat pełnionych przez nie funkcji komórkowych. Nie jest to jednak łatwe zadanie, gdyż funkcje przez nie pełnione są różnorodne i są pochodną ogromnego ich zróżnicowania pod względem sekwencji, struktury drugorzędowej, tkankowej specyficzności, lokalizacji komórkowej czy biogenezy. Pośród biologicznych ról tych cząsteczek, można wymienić m.in. modulowanie procesu transkrypcji (Kugel i Goodrich, 2012), działanie jako tzw. gąbki mikroRNA (Tay et al., 2014), modyfikowanie funkcji białek poprzez bezpośrednie oddziaływanie z nimi (Yin et al., 2012), udział w tworzeniu kompleksów makromolekularnych (Kugel i Goodrich, 2012) czy pełnienie roli cząsteczek sygnałowych (Huang et al., 2013). Ze względu na to, że uczestniczą w licznych procesach molekularnych, jak również wielokrotnie powiązano je z chorobami człowieka, stanowią interesujący obiekt badań o potencjalnie dużym znaczeniu dla biologii molekularnej, biotechnologii i medycyny. Znaczenie lncRNA, z naciskiem na te ulegające ekspresji z nici antysensownej względem genów kodujących białko, czyli tzw. NATs (ang. *natural antisense transcripts*), opisałem szczegółowo ze współautorami w pracy przeglądowej (**Wanowska et al., 2018**). Omówiliśmy w niej również molekularne podłoże związku tych cząsteczek z chorobami człowieka, takimi jak nowotwory i choroby neurodegeneracyjne, a na koniec przedstawiliśmy postępy w użyciu nowoczesnych terapii z wykorzystaniem oligonukleotydów nakierowanych na te lncRNA oraz wady i zalety takiego podejścia.

Szczególnie słabo poznanym aspektem biologii lncRNA jest ich udział w dojrzewaniu transkryptów i regulacji ekspresji genów poprzez wchodzenie w oddziaływanie z komplementarnymi cząsteczkami RNA. W takim układzie lncRNA mogłyby uczestniczyć w takich procesach, jak (i) regulacja splicingu na zasadzie maskowania sygnałów splicingowych (Beltran et al., 2008), (ii) edytowanie RNA (Kawahara et al., 2007), (iii) Staufen-mediated decay (SMD), będący szlakiem degradacji RNA u ssaków (Gong i Maquat, 2011), (iv) maskowanie miejsc wiązania mikroRNA (Faghihi et al., 2008). Zostały one schematycznie przedstawione na Rys. 1, a ich krótka charakterystyka przedstawiona została poniżej.

Załącznik 2

Modulowanie splicingu. Oddziaływanie lncRNA z cząsteczką pre-mRNA może prowadzić do maskowania miejsc splicingowych i/lub innych sygnałów splicingowych (Rys. 1a). Takie sparowania są szczególnie prawdopodobne w przypadku antysensownych lncRNA (NATs), które z racji swojego położenia w genomie mogłyby parować się z transkryptami genu znajdującego się na przeciwnej nici DNA, jednakże oddziaływania w *trans* są również możliwe. Na przykład lncRNA o nazwie 5S-OT, silnie zakonserwowany ewolucyjnie u *Eukaryota*, reguluje alternatywny splicing wielu genów człowieka w *trans* poprzez bezpośrednie parowanie z innymi cząsteczkami RNA w regionie elementu powtarzalnego Alu (Hu et al., 2016).



Rys. 1
Schematyczna reprezentacja procesów molekularnych z udziałem lncRNA tworzących sparowania z komplementarną cząsteczką RNA.

Edytowanie cząsteczek RNA. Edytowanie RNA prowadzi do powstania transkryptów o sekwencji nukleotydowej różniącej się od sekwencji kodującego je genu. Najlepiej poznanym mechanizmem edytowania cząsteczek RNA jest hydrolityczna deaminacja adenozyny do inozyny z udziałem enzymów ADAR (ang. *adenosine deaminases acting on RNA*); inozyna paruje się z cytydyną, a więc imituje guaninę, co ma rozmaite konsekwencje funkcjonalne, od zaburzeń splicingu, po deregulację funkcji mikroRNA (Kawahara et al., 2007). Do deaminacji dochodzi w regionach występowania dwuniciowych cząsteczek RNA (ang. *double-stranded RNA*, dsRNA), w związku z czym miejsca oddziaływań pomiędzy lncRNA a pre-mRNA mogą stanowić substrat dla enzymów ADAR (Rys. 1b). Jako że 85% transkryptów człowieka przechodzi proces edytowania (Athanasiadis et al., 2004) oraz znaczna część lncRNA posiada potencjał oddziaływania z innymi cząsteczkami RNA w *cis* i/lub *trans*, udział lncRNA w tym procesie może być znaczący.
Ścieżka SMD. lncRNA posiadające sekwencje powtarzalne Alu mogą kierować cząsteczki mRNA do degradacji w ścieżce SMD (ang. *Staufen-mediated mRNA decay*) (Gong i Maquat, 2011). Kluczową rolę pełni tutaj białko STAU1, które rozpoznaje dsRNA w regionach 3'UTR (region nieulegający translacji, ang. *untranslated region*) mRNA. Fragmenty dsRNA powstają albo na

zasadzie parowań wewnątrzcząsteczkowych, albo z udziałem innych transkryptów, w tym przede wszystkim lncRNA (Rys. 1c). Warto zwrócić uwagę na fakt, że SMD jest ścieżką degradacji cząsteczek kodujących białko, a więc pojedynczy transkrypt lncRNA może decydować o losie wielu cząsteczek mRNA, sam pozostając nietknięty.

Blokowanie funkcji mikroRNA. Formowanie dupleksów lncRNA:mRNA może skutkować podwyższeniem poziomu ekspresji mRNA poprzez zaburzenie funkcji regulatorowych mikroRNA na zasadzie maskowania rozpoznawanych przez nie sekwencji docelowych w regionie 3'UTR (Rys. 1d). Na przykład BACE1AS, transkrypt genu antysensownego względem genu *BACE1*, maskuje sekwencję docelową w mRNA *BACE1* typowo rozpoznawaną przez miR-485-5p, co powiązano z etiologią choroby Alzheimera (Faghihi et al., 2008).

Mając na uwadze takie fakty, jak to że lncRNA stanowią niezwykle liczebną klasę cząsteczek RNA, większość z nich nie została jeszcze scharakteryzowana funkcjonalnie, wiele z tych dobrze opisanych pełni ważną rolę w komórce, jak również ogromny potencjał regulatorowy tych cząsteczek w mechanizmach, w których tworzyłyby bezpośrednie sparowania z innymi RNA, **moim nadrzędnym celem było zidentyfikowanie lncRNA uczestniczących właśnie w takiego rodzaju procesach regulatorowych i bliższe scharakteryzowanie tych funkcji używając podejść bioinformatycznych**. Aby to było możliwe, konieczna była realizacja drugorzędowych celów: i) zbudowanie potoku analitycznego służącego do identyfikacji lncRNA (modyfikacja istniejących już podejść); ii) opracowanie potoku analitycznego do identyfikacji oddziaływań RNA:RNA w skali całego transkryptomu; iii) przygotowanie trzeciego potoku analitycznego, służącego do przewidywania funkcji lncRNA pełnionych w kontekście oddziaływań RNA:RNA; iv) implementacja potoków analitycznych i wykonanie dodatkowych, uzupełniających analiz, celem głębszego scharakteryzowania lncRNA i ich funkcji. Dodatkowym celem było opracowanie dwóch nowych internetowych baz danych, które udostępniałyby w przystępny sposób wyniki uzyskane w trakcie realizacji projektu.

Na realizację przedsięwzięcia otrzymałem finansowanie od Narodowego Centrum Nauki, w postaci grantu SONATA (numer referencyjny 2014/15/D/NZ2/00525) i w przeważającej mierze wykonane było ono w Zakładzie Genomiki Zintegrowanej (Uniwersytet im. Adama Mickiewicza), w grupie badawczej pod kierunkiem prof. dr hab. Izabeli Makałowskiej. Badania były również częściowo finansowane przez Ministerstwo Nauki i Szkolnictwa Wyższego w postaci grantu Mobilność Plus IV (numer referencyjny 1268/MOB/IV/2015/0) na staż w Max Delbrück Center for Molecular Medicine w Berlinie, w grupie prof. Uwe Ohlera. Krótka charakterystyka otrzymanych wyników, ich znaczenie naukowe i potencjał aplikacyjny przedstawione zostały poniżej.

2. Identyfikacja bezpośrednich sparowań typu RNA:RNA w skali transkryptomu (Szcześniak i Makalowska, 2016)

Badanie funkcji lncRNA pełnionych w kontekście oddziaływań RNA:RNA wymaga identyfikacji sparowań międzycząsteczkowych w skali całego transkryptomu. Nie jest to jednak zadanie

trywialne. Po pierwsze, analiza tego rodzaju powinna uwzględniać struktury drugorzędowe tworzone przez cząsteczki RNA, jako że sparowania wewnątrzcząsteczkowe mogą uczynić niektóre fragmenty sekwencji RNA niedostępnymi dla drugiej cząsteczki RNA. Przewidywanie struktur drugorzędowych transkryptów jest jednak zadaniem obciążonym dużym błędem (głównie ze względu na dość dużą długość transkryptów człowieka, ze średnią dużo powyżej 1000 nt), jak również kosztownym obliczeniowo i czasochłonnym, co wykazałem na przykładzie narzędzi RNAplex (Tafer i Hofacker, 2008), RNAduplex (Lorenz et al., 2011) i LncTar (Li et al., 2015). Pomocne mogą tutaj być metody eksperymentalnego badania struktur drugorzędowych, jak SHAPE, PARS i FragSeq, jednak otrzymywane w ten sposób dane nie są dostępne dla pełnych transkryptomów, a szczególnie mało jest ich dla lncRNA, najprawdopodobniej ze względu na stosunkowo niski przeciętny poziom ekspresji tych cząsteczek. Z tych względów opracowałem własne rozwiązanie, polegające na wykorzystaniu narzędzia *lastal* z pakietu LAST (Kiełbasa et al., 2011). *Lastal* służy do przeszukiwania baz danych sekwencji DNA, podobnie jak w przypadku programu BLAST. Jego domyślne parametry oraz macierz substytucji zostały jednak tak przeze mnie zmodyfikowane, aby program szukał fragmentów RNA mogących wchodzić ze sobą w bezpośrednie oddziaływanie. Uwzględniono tutaj kryteria powszechnie wykorzystywane w innych narzędziach, np. służących do identyfikacji sekwencji docelowych dla mikroRNA (Dai i Zhao, 2011). Rozwiązanie obejmuje również test na istotność statystyczną uzyskanych wyników oraz konwersję uzyskanych przyrównań RNA:RNA z domyślnego formatu MAF do innych formatów danych, m.in. BED. W ten sposób problem identyfikacji mogących ze sobą oddziaływać cząsteczek RNA zredukowany został do szukania podobnych fragmentów RNA, co pozwala na **co najmniej tysiąckrotne przyspieszenie analiz w skali transkryptomu w porównaniu z wiodącymi narzędziami**. Następnie użyłem wcześniej zaproponowanej metodologii, by zbadać czułość i specyficzność podejścia (Li et al., 2015). **Obliczyłem, że specyficzność - na poziomie 99.82% - jest wyższa od innych narzędzi, przy porównywalnej czułości, wynoszącej 80%.** Warto wspomnieć, że choć omawiany tutaj algorytm jest stosunkowo wydajny i stanowił podstawę do identyfikowania lncRNA mogących wchodzić w bezpośrednie oddziaływania z innymi cząsteczkami, w dalszych analizach często posiłkowano się dodatkowymi danymi, w celu uwiarygodnienia sparowań. Na przykład wymagane było, aby obie cząsteczki RNA ulegały ekspresji w tej samej tkance lub linii komórkowej.

Algorytm ten wykorzystałem następnie w analizie oddziaływań lncRNA ze wszystkimi transkryptami człowieka z bazy danych ENSEMBL (Zerbino et al., 2018), identyfikując 15,082,791 potencjalnych sparowań z mRNA oraz 56,735,686 z pre-mRNA. Rozróżnienie na oddziaływania z mRNA lub pre-mRNA ma związek z badanymi funkcjami lncRNA, ponieważ niektóre z nich zachodzą w jądrze i dotyczą cząsteczek pre-mRNA (modulacja splicingu, edytowanie RNA), inne zaś zachodzą w cytoplazmie na poziomie cząsteczek mRNA (kierowanie cząsteczek mRNA do degradacji w ścieżce SMD czy deregulacja funkcji mikroRNA). Oddziaływań szukano na poziomie transkryptów, a więc różne formy splicingowe lncRNA mogły wchodzić w identyczne lub bardzo podobne oddziaływania z wieloma izoformami splicingowymi innego genu, przez co podane wyżej liczby są tak duże. W sumie jednak pośród kandydatów

znalazło się tylko 0.098% wszystkich możliwych par gen-gen u człowieka. Ponadto okazało się, że sparowania dotyczą zazwyczaj dwóch lncRNA, zaś te z udziałem mRNA są głównie zlokalizowane w regionie 3' UTR, względem np. 5,19% z nich zlokalizowanych w 5'UTR czy 4,10% w regionie kodującym. Tę obserwację można częściowo wyjaśnić tym, że kumulatywna długość ludzkich 3' UTR jest niemal czterokrotnie większa, niż 5' UTR. Dodatkowo, regiony 3' UTR posiadają więcej regionów powtarzalnych niż np. 5' UTR, co zwiększa szanse na oddziaływanie z lncRNA, które same składają się w ponad 20% z elementów powtarzalnych. Preferencja do parowania z 3'UTR jest ciekawą obserwacją, zważywszy na olbrzymi potencjał regulatorowy zlokalizowany w tych regionach (np. procesy regulatorowe z udziałem mikroRNA, białek HuR czy STAU1). W celu uwiarygodnienia zidentyfikowanych oddziaływań, sprawdzone zostały poziomy ekspresji RNA przewidzianych jako tworzące pary lncRNA:RNA. Obliczenia wykonano dla 63 bibliotek RNA-Seq, obejmujących szereg linii komórkowych i tkanek człowieka. Okazało się, że 46,68% transkryptów ulega ekspresji w przynajmniej jednej próbce, natomiast obie cząsteczki RNA wykazują koekspresję w przypadku 36,33% oddziaływań z udziałem mRNA oraz 28,84% z pre-mRNA. Te liczby pokazują, że w przypadku znaczącego odsetka przewidzianych oddziaływań obie cząsteczki RNA współwystępują, co czyni ich parowanie możliwym, choć na tym etapie nadal nie wiadomo czy obecne są w tym samym kompartmencie komórkowym i czy rzeczywiście wchodzą w bezpośrednie oddziaływanie.

3. Badanie funkcji lncRNA pełnionych w kontekście oddziaływań RNA:RNA

3.1 Badania nad znanymi wcześniej lncRNA człowieka (Szczesniak i Makalowska, 2016)

W celu określenie potencjalnych funkcji pełnionych przez lncRNA w kontekście oddziaływań RNA:RNA, **przygotowano specjalistyczny potok analityczny**, który w oparciu o dane wejściowe, takie jak przewidziane lub wcześniej znane lncRNA i ich bezpośrednie oddziaływania z innymi transkryptami, **sprawdzał czy dana cząsteczka lncRNA może pełnić funkcje w co najmniej jednym z czterech niżej wymienionych procesów molekularnych**.

Modulowanie splicingu poprzez maskowanie miejsc splicingowych. W tym celu przefiltrowano miejsca oddziaływań pomiędzy lncRNA a pre-mRNA, zachowując tylko te, które obejmowały miejsca splicingowe będące przedmiotem alternatywnego splicingu. Zakłada się, że silne oddziaływanie pomiędzy lncRNA a cząsteczką pre-mRNA uniemożliwia rozpoznanie miejsca splicingowego przez spliceosom, przez co dochodzi do alternatywnego zdarzenia splicingowego. Dodatkowo, na koordynaty tych miejsc splicingowych nałożono dane CLIP-Seq dotyczące szeregu czynników splicingowych (U2AF65, PTB, FMRP, QKI, TIAL1, TIA1, HuR, TDP-43 i hnRNPC), aby sprawdzić czy miejsca rozpoznawane przez czynniki splicingowe również ulegają maskowaniu. W toku analiz otrzymałem 3,788,306 interakcji pomiędzy lncRNA a pre-mRNA obejmujących co najmniej jedno miejsce splicingowe, z czego 71% nakładało się z alternatywnymi miejscami splicingowymi. W sumie zidentyfikowano 21,456 unikalnych, alternatywnych miejsc splicingowych, które mogłyby być maskowane przez długie niekodujące RNA. Są one w

większości zlokalizowane w transkryptach niekodujących białka lub w regionach 3'UTR i 5'UTR cząsteczek mRNA. Co ciekawe, 28% tych oddziaływań nakłada się z regionami CLIP-Seq przypisanymi do jednego z dziewięciu badanych czynników splicingowych, w porównaniu z 7% w przypadku oddziaływań obejmujących konstytutywne miejsca splicingowe. Wskazuje to na dość oczywisty związek pomiędzy obecnością alternatywnego miejsca splicingowego i oddziaływaniem z czynnikami splicingowymi, ale stwarza również możliwość, że funkcja regulatorowa lncRNA polega nie tylko na maskowaniu miejsca splicingowego, ale również na uniemożliwieniu wiązania czynników splicingowych z rozpoznawanymi przez nie dodatkowymi sygnałami splicingowymi w cząsteczce pre-mRNA.

Blokowanie funkcji mikroRNA. Koordynaty miejsc oddziaływań pomiędzy mikroRNA a mRNA z eksperymentów CLIP-Seq z bazy danych StarBase 2.0 (Li et al., 2014) nałożono na miejsca docelowe miRNA przewidziane *in silico* programem Miranda (Enright et al., 2003). Część wspólna tych dwóch zestawów miejsc docelowych miRNA oraz miejsc oddziaływań lncRNA z regionami 3'UTR, ustalona narzędziami z pakietu BEDTools (Quinlan i Hall, 2010), stanowi miejsca, w których potencjalnie dochodzi do konkurencji pomiędzy lncRNA a mikroRNA o wiązanie do cząsteczek mRNA - zidentyfikowano 21,204 takich miejsc.

Kierowanie cząsteczek mRNA do degradacji w ścieżce SMD. Użyto narzędzia RepeatMasker (http://www.repeatmasker.org) by zidentyfikować elementy powtarzalne Alu w transkryptach człowieka, po czym nałożono je na koordynaty miejsc oddziaływań między lncRNA i mRNA i zachowano tylko takie przypadki, gdzie element Alu oraz oddziaływanie RNA:RNA występują w regionie 3'UTR transkryptu. lncRNA tworzące tak zdefiniowane pary uznano za potencjalny czynnik modulujący stabilność mRNA, poprzez kierowanie ich do degradacji w ścieżce SMD. W trakcie analizy, elementy Alu zidentyfikowano w regionach 3' UTR 24,886 transkryptów. W przypadku 7,439 z nich, element Alu całkowicie wchodził w region oddziaływania z cząsteczka lncRNA.

Udział w procesie edytowania RNA. Z bazy danych RADAR (Ramaswami and Li, 2014) pobrano genomowe koordynaty miejsc, w których u człowieka dochodzi do edytowania, po czym nałożono je na miejsca oddziaływań z cząsteczkami lncRNA. Zakłada się, że lncRNA oddziałując z innymi cząsteczkami RNA, tym samym współtworząc strukturę dsRNA, dostarczają substratu dla enzymów ADAR, odpowiedzialnych za edytowanie typu adanina --> inozyna. Połączenie zestawu oddziaływań lncRNA z RNA oraz miejsc edytowanie RNA z bazy danych RADAR pozwoliło na zidentyfikowanie 12,853 transkryptów, których edytowanie może zachodzić z udziałem lncRNA.

Podsumowując, w oparciu o adnotacje człowieka, w tym zestaw lncRNA z bazy danych ENSEMBL, wyszczególniłem 57 303 transkrypty człowieka, których dojrzewanie, stabilność i poziom ekspresji mogą być modulowane przez cząsteczki długich niekodujących RNA, na zasadzie bezpośredniego z nimi oddziaływania. Wskazuje to na ogromny potencjał regulatorowy, jakim dysponują cząsteczki lncRNA, jednak konieczne są dalsze badania, w tym eksperymentalne potwierdzenie uzyskanych wyników.

3.2 Badania uwzględniające samodzielnie zidentyfikowane lncRNA (Szczesniak et al., 2016; Szczesniak et al., 2017; Szczesniak et al., 2019)

W dalszej kolejności przystąpiłem do badań nad długimi niekodującymi RNA u roślin oraz w chorobie rogówki oka, zwanej stożkiem rogówki. Liczba lncRNA znanych u roślin jest co najmniej o rząd wielkości mniejsza niż u ssaków, niemniej jednak wciąż stanowią ważny komponent transkryptomu. Na przykład baza danych GreeNC przechowuje 3,008 lncRNA *Arabidopsis thaliana* (Paytuvi Gallart et al., 2016), zaś NONCODE 2016 (Zhao et al., 2016) kolekcjonuje ich 3,763. W przypadku wielu gatunków roślin, wliczając organizmy modelowe, nie opisano jeszcze długich niekodujących RNA, mimo że powiązano je z szeregiem procesów biologicznych, takich jak wernalizacja (Kim et al., 2017) czy organogeneza (Wang et al., 2017). Swoje funkcje pełnią poprzez udział na różnych etapach regulacji ekspresji genów, jak np. remodelowanie chromatyny (Bardou et al., 2014), modulowanie translacji mRNA (Bazin et al., 2017) czy regulacja alternatywnego splicingu. Zważywszy na udowodnione znaczenie funkcjonalne lncRNA u roślin oraz niezwykle skąpą wiedzę na ich temat, wliczając dalece niepełne katalogi tych cząsteczek, postanowiliśmy w pierwszej kolejności przeprowadzić wielkoskalową identyfikację lncRNA, a następnie wykonać analizę ich potencjalnych funkcji komórkowych.

Proces identyfikacji lncRNA składał się z kilku etapów: i) pobieranie i kontrola jakości wyników wysokoprzepustowego sekwencjonowania transkryptomów (RNA-Seq); ii) mapowanie odczytów RNA-Seq do genomu referencyjnego (w pełni lub częściowo złożonego, w zależności od stopnia złożenia genomu); iii) obróbka wyników mapowania i składanie transkryptomu *ab initio*; iv) identyfikacja lncRNA pośród otrzymanego zestawu transkryptów. Samo szukanie lncRNA polegało na zastosowaniu szeregu filtrów, pozwalających na odróżnienie ich od innych transkryptów, jak mRNA, rybosomalne RNA czy snoRNA. W tym celu m.in. porównano złożony *ab initio* transkryptom z zaadnotowanym transkryptomem referencyjnym (o ile był dostępny), wykonano przeszukiwanie bazy danych sekwencji kodujących białko i niekodujących, zbadano potencjał kodujący, jak również usunięto transkrypty o długości poniżej 200 nt. **Wyżej wspomniana metoda identyfikacji lncRNA została zaimplementowana jako wydajny i innowacyjny potok analityczny, który posłużył do analizy danych RNA-Seq 39 gatunków roślin i glonów. W sumie znalazłem 239 631 roślinnych lncRNA i są to absolutnie unikalne wyniki.** Na przykład porównanie z bazą danych ENSEMBL Plants (Kersey et al., 2018) wykazało, że tylko 9% z nich były wcześniej znane, tym bardziej że dla większość badanych roślin i glonów nie opisano wcześniej lncRNA. Jedynie w przypadku *Arabidopsis thaliana,* gatunku o stosunkowo dobrze zaadnotowanym transkryptomie, ponad połowa kandydatów była identyczna z lncRNA zdeponowanymi w ENSEMBL Plants. Warto tutaj wspomnieć, że **wartością dodaną projektu było złożenie transkryptomów kilkudziesięciu gatunków roślin i znaczne poszerzenie repertuaru znanych genów oraz form splicingowych, wraz z uzyskanymi przez nas podstawowymi adnotacjami tych cząsteczek.**

Przystępując do analiz funkcjonalnych lncRNA, ze względu na specyfikę materiału badawczego i dostępność danych, badana była tylko ich rola w modulowaniu splicingu oraz blokowaniu funkcji mikroRNA. Z tych samych powodów konieczna była również modyfikacja naszego potoku

analitycznego służącego do określania funkcji lncRNA pełnionych w kontekście oddziaływań RNA:RNA. **Ostatecznie, uzyskano 11 659 lncRNA powiązanych z modulacją splicingu oraz 440 lncRNA sklasyfikowanych jako czynniki deregulujące funkcje mikroRNA.** Dla uwiarygodnienia wyników sprawdzono, że dwie trzecie oddziaływań dotyczy cząsteczek RNA, które współwystępują w badanych próbkach, co przekłada się na 75% unikalnych lncRNA z przewidzianymi funkcjami.

Analiza funkcjonalna lncRNA była poprzedzona ich identyfikacją również w przypadku badań nad stożkiem rogówki. Stożek rogówki (*keratoconus*, KTCN) to degeneracyjna choroba rogówki oka, polegająca na jej ścieńczeniu i nadmiernym uwypukleniu (Rabinowitz, 1998). Ze względu na zmianę krzywizny rogówki, choroba może prowadzić do znacznego zaburzenia wzroku. Uważa się, że u podłoża choroby stoją czynniki środowiskowe, jak również genetyczne, jednak co do tych drugich dysponujemy jedynie domysłami (Nowak i Gajecka, 2011). Jedną z możliwości jest to, że długie niekodujące RNA przyczyniają się do rozwoju choroby. Aby zbadać to przypuszczenie, wykonano wysokoprzepustowe sekwencjonowanie transkryptomów rogówek osób ze stożkiem rogówki (KTCN) oraz rogówek kontrolnych (tzw. non-KTCN). W oparciu o dane RNA-Seq złożono transkryptom, a następnie przeprowadzono identyfikację lncRNA. **Ostatecznie otrzymałem zestaw 16 331 lncRNA ulegających ekspresji w rogówce, z czego 735 to nieznane wcześniej lncRNA. Dalsza analiza pozwoliła przewidzieć funkcje dla 870 lncRNA, potencjalnie oddziałujących z transkryptami 996 genów**, przy czym były to geny dla których wykazano istotną statystycznie różnicę ekspresji między dwoma zestawami próbek: KTCN i non-KTCN. **Do tego zestawu genów zaliczyć można m.in. *SMAD9, SMAD6, TGFB3* i *TGFBR1* należące do ścieżek sygnalizacyjnych białek TGF-β, Hippo czy Wnt, które już wcześniej wiązano z chorobami narządu wzroku** (Morgan et al., 2013). Co ciekawe, w przypadku 262 genów interakcja z lncRNA ma miejsce tylko w KTCN, ponieważ w pozostałych próbkach lncRNA i druga cząsteczka RNA z pary nie współwystępują, stanowiąc jeszcze jedną przesłankę w kierunku powiązania funkcjonalnego lncRNA z rozwojem stożka rogówki. Geny te wykazują nadreprezentację w ścieżkach sygnalizacyjnych związanych z naskórkowym czynnikiem wzrostu (ang. *epidermal growth factor*, EGF), co można powiązać zaburzeniem procesów regeneracyjnych w rogówce po urazie mechanicznym, które niejednokrotnie mają miejsce w stożku rogówki (Cheung et al., 2014). Dodatkowo, w przypadku około 35% transkryptów, regulacja przez lncRNA może zachodzić na co najmniej dwa sposoby, gdzie większość przypadków dotyczy jednoczesnego powiązania ze ścieżką SMD i edytowaniem RNA. Można to wytłumaczyć w ten sposób, że elementy Alu są preferencyjnie rozpoznawane przez enzymy ADAR i to w nich najczęściej dochodzi do deaminacji adeniny (Daniel et al., 2014). Z drugiej strony, te same elementy Alu są istotnym komponentem struktur dsRNA, nakierowujących transkrypty kodujące białko do rozkładu w ścieżce SMD. Warto również wspomnieć, że około 30% genów wykazujących ekspresję różnicową pomiędzy próbkami KTCN i non-KTCN nakłada się z innymi genami, przy czym prawie zawsze (92,5% par) oba geny wykazują taki sam kierunek zmiany. W większości przypadków jednym z komponentów pary jest lncRNA. Zgodnie z istniejącą literaturą,

taka pozytywna korelacja ekspresji sugeruje, że część lncRNA reguluje ekspresję genów w *cis*, na przykład poprzez udział w modyfikacji histonów czy metylacji DNA (Kugel i Goodrich, 2012).

3.3 Retropozycja jako źródło potencjalnie funkcjonalnych lncRNA (Bryzghalov et al., 2016)

W naszych badaniach przyjrzeliśmy się także funkcjom regulatorowym lncRNA w kontekście ewolucyjnym, zwracając szczególną uwagę na nowe kopie genów powstające w trakcie retropozycji. Retropozycja to proces, w którym cząsteczka mRNA ulega odwrotnej transkrypcji do cDNA, po czym zostaje wbudowana do genomu w lokalizacji innej, niż gen źródłowy (zwany genem rodzicielskim). Powstała w ten sposób nowa kopia genu jest nazywana retrokopią. Retrokopie często są nieaktywne transkrypcyjnie, ponieważ nie posiadają sekwencji promotora, natomiast te, które ulegają ekspresji, zwyczajowo nazywane są retrogenami i stanowią około 7,4% wszystkich genów człowieka. Mogą one wykształcić funkcje inne, niż te pełnione przez ich geny rodzicielskie (tzw. neofunkcjonalizacja) lub pełnić te same funkcje, ale w innym kontekście, na przykład w innej tkance lub na innym etapie rozwoju embrionalnego (subfunkcjonalizacja). Opisane zostały również przypadki, gdzie retrogen zastępuje gen rodzicielski, który to zostaje utracony (Ciomborowska et al., 2013). W naszych badaniach sprawdziliśmy czy retrokopie używają promotorów zlokalizowanych poniżej miejsca insercji, co powinno skutkować ich ekspresją z przeciwnej nici DNA, tzw. nici antysensownej. Ze względu na różnice w sekwencji, takie retrokopie powinny pełnić funkcje inne niż ich geny rodzicielskie. Z drugiej jednak strony, konsekwencją ich pochodzenia jest pełna lub częściowa komplementarność względem genów rodzicielskich, co stwarza możliwość wchodzenia z nimi w bezpośrednie oddziaływania RNA:RNA, mogące skutkować wpływem na obróbkę, stabilność i poziom ekspresji. Mając to na uwadze, **zbadaliśmy 4675 retrokopii człowieka z bazy danych RetrogeneDB (Rosikiewicz et al., 2017) i prawdopodobnie jako pierwsi zidentyfikowaliśmy 58 lncRNA, które są antysensowne wobec retrokopii człowieka**. Taka sama analiza w przypadku szympansa, stanowiącego układ porównawczy w kontekście ewolucyjnym, wykazała brak antysensownych lncRNA, co jednak wynika z faktu, że w bazie danych ENSEMBL zdeponowanych jest bardzo niewiele lncRNA tego gatunku. Z tego powodu wykonaliśmy identyfikację lncRNA u szympansa w oparciu o dane RNA-Seq z bazy danych SRA (Kodama et al., 2012) i wykorzystując wcześniej opisany potok analityczny. **Pozwoliło to na 14-krotne zwiększeniu puli lncRNA szympansa** oraz wytypowanie 23 lncRNA ulegających ekspresji z nici antysensownej retrokopii. Co ciekawe, żaden z przypadków nie jest zakonserwowany ewolucyjnie między człowiekiem i szympansem. Aby znaleźć wsparcie dla hipotetycznego związku funkcjonalnego między antysensownymi lncRNA a genami rodzicielskimi, oszacowaliśmy ekspresję genów w 153 próbkach, korzystając z danych RNA-Seq zdeponowanych w bazie danych ENCODE (ENCODE Project Consortium, 2012). Zauważyliśmy, że w przypadku 27 z 35 lncRNA ulegają one ekspresji w tych samych próbkach, co gen rodzicielski. Następnie zauważyliśmy, że w dwóch przypadkach ekspresja genów z pary jest pozytywnie skorelowana (Spearman rho), a dla innej pary ma miejsce antykorelacja poziomów ekspresji, co dodatkowo wzmacnia hipotezę o powiązaniu funkcjonalnym lncRNA i genów rodzicielskich. Dla pozostałych par nie wykryliśmy korelacji ekspresji, co może wynikać z faktu, że lncRNA i gen rodzicielski nie są powiązane funkcjonalnie,

inne czynniki wpływają na ekspresję obu genów (np. mikroRNA, czynniki transkrypcyjne) albo geny ulegają ekspresji w niewielkiej liczbie próbek, co uniemożliwia uzyskanie istotnych statystycznie wyników w analizie korelacji. Poza tym, niektóre funkcje lncRNA pełnione w kontekście oddziaływań RNA:RNA (np. wywoływanie edytowania cząsteczek RNA) niekoniecznie wiążą się ze wzrostem lub spadkiem poziomu ekspresji genu, co również tłumaczy brak korelacji poziomów ekspresji.

W dalszej kolejności, dla każdej pary lncRNA – gen rodzicielski poszukaliśmy możliwych oddziaływań RNA:RNA i **korzystając z wcześniej opisanego potoku analitycznego udało nam się przypisać funkcje dla 10 antysensownych lncRNA**. Trzy wspomniane wcześniej przypadki dla których wykazano (anty)korelację ekspresji, dotyczące genów rodzicielskich *hnRNPA1*, *CHMP1A* i *RPL23A*, znajdują się wśród tych 10 kandydatów. Przyjrzeliśmy się im dokładniej i dokonaliśmy dość interesujących obserwacji. Na przykład gen rodzicielski *hnRNPA1*, którego białko uczestniczy w obróbce pre-mRNA i innych aspektach metabolizmu RNA (Han et al., 2010), posiada retrokopię *retro_hsap_1933*, której antysensowny transkrypt to lncRNA znany jako AC021224.1-201. Zauważyliśmy, że oddziaływanie lncRNA z pre-mRNA genu rodzicielskiego (hnRNPA1) może prowadzić do maskowania miejsca splicingowego 5’ w szóstym intronie, co skutkuje alternatywnym zdarzeniem splicingowym. Z kolei pod nieobecność lncRNA powstaje inna forma splicingowa, znana jako ENST00000547276, która nie posiada domeny funkcjonalnej potrzebnej do pełnienia podstawowych funkcji przez białko hnRNPA1, wliczając wiązanie cząsteczek RNA i modulowanie splicingu (Mayeda et al., 1994). Z drugiej strony, tylko ta krótsza forma splicingowa uczestniczy w regulacji splicingu i replikacji HIV-1. W związku z tym **zidentyfikowane sprzężenie między lncRNA a jego genem rodzicielskim prowadzi do modyfikacji wzoru splicingu, co skutkuje zmianą funkcjonalności powstającego białka genu *hnRNPA1*.**

4. Opracowanie specjalistycznych baz danych (Szczesniak et al., 2016; Szczesniak et al., 2017; Szczesniak et al., 2019)

W celu upowszechnienia wyników uzyskanych w trakcie realizacji projektu, jak również by ułatwić dostęp do wielkoskalowych danych, ich przeglądanie, przeszukiwanie, wizualizację i pobieranie, utworzone zostały dwie specjalistyczne internetowe bazy danych. Pierwsza z nich, CantataDB, kolekcjonuje roślinne lncRNA, zaś druga, nazwana KTCNlncDB, przechowuje lncRNA zidentyfikowane w rogówce człowieka, wraz informacją na temat ich przewidzianego funkcjonalnego związku z etiologią stożka rogówki.

Baza danych CantataDB 1.0, jak i jej niedawno zaktualizowana wersja określona mianem CantataDB 2.0 (http://rhesus.amu.edu.pl/KTCNlncDB), została zbudowane z wykorzystaniem technik webowych PHP, HTML, CSS, mySQL, JavaScript i Bootstrap, a jej działanie sprawdzono na szeregu przeglądarek internetowych oraz na różnych systemach operacyjnych. Funkcjonalności dostępne w bazie danych można podzielić na cztery kategorie:

- Strona służąca do przeszukiwania danych dostępnych w bazie danych, np. na podstawie nazwy lncRNA, jego poziomu ekspresji czy nazwy gatunku. Po wykonaniu wyszukiwania,

użytkownik otrzymuje informację o użytych kryteriach wyszukiwania, wraz z opcją pobrania uzyskanych wyników. Tabela wynikowa zawiera dane na temat zidentyfikowanych lncRNA, przy czym jeden wiersz w tabeli odpowiada pojedynczej cząsteczce lncRNA. Oprócz kilku podstawowych informacji o lncRNA, jak koordynaty genomowe czy maksymalny poziom ekspresji, znajduje się tutaj również przycisk *Details*, który przekierowuje do strony z dodatkowymi, bardziej szczegółowymi informacjami, jak na przykład sekwencja lncRNA, wykres słupkowy prezentujący profil ekspresji lncRNA w różnych próbkach, zidentyfikowane interakcje lncRNA:RNA oraz ich potencjalne znaczenie funkcjonalne.

- Na osobnej stronie, użytkownik ma możliwość przeszukiwania sekwencji lncRNA wszystkich gatunków uwzględnionych w CantataDB za pomocą programu BLAST.
- Oprócz możliwości pobierania danych związanych z wynikami wyszukiwania według kryteriów użytkownika, możliwe jest również całościowe pobieranie danych, wliczając pliki z sekwencjami lncRNA w formacie FASTA oraz ich adnotacje.
- Na osobnej podstronie wymienione są i krótko scharakteryzowane wszystkie próbki RNA-Seq użyte w analizie danych poszczególnych gatunków roślin i glonów.

Istnieje kilka baz danych, w których zdeponowane są roślinne lncRNA, jednak CantataDB jest najprawdopodobniej najobszerniejszą bazą danych, a dodatkowo, oprócz samych sekwencji lncRNA, przechowuje szereg informacji pomocnych w analizach funkcjonalnych tych cząsteczek. Dla porównania Plant Non-coding RNA Database (PNRD) (Yi et al., 2015) przechowuje sekwencje lncRNA czterech gatunków roślin: *A. thaliana* (2577 lncRNA), *O. sativa* (752 lncRNA), *P. trichocarpa* (538 lncRNA) i *Z. mays* (1704 lncRNA). NONCODE 2016 (Zhao et al., 2016), oprócz lncRNA zwierząt i grzybów, przechowuje 3763 lncRNA zidentyfikowane u *A. thaliana*. LncRNAdb (Quek et al., 2015) przechowuje eksperymentalnie potwierdzone lncRNA dla 9 gatunków roślin: 7 dla *A. thaliana* oraz między 1 a 3 dla pozostałych gatunków. Ponadto, CantataDB 2.0 posiada lncRNA dla 15 gatunków, które nie zostały uwzględnione w GreeNC (Paytuvi Gallart et al., 2016), wliczając organizmy modelowe lub ważne z ekonomicznego punktu widzenia, jak na przykład *Brassica rapa, Brassica napus* czy *Hordeum vulgare*. **Wyróżniającą cechą bazy danych CantataDB jest to, że oprócz samych sekwencji lncRNA, przechowywane są tutaj informacje o oddziaływaniach lncRNA z innymi transkryptami, wraz z ich przypuszczalnymi konsekwencjami funkcjonalnymi** (wersja 1.0 bazy). W sumie, funkcje przewidziano dla 11 896 lncRNA, wliczając 11 659 lncRNA mogących uczestniczyć w modulowaniu splicingu oraz 440 lncRNA potencjalnie powiązanych z blokowaniem funkcji mikroRNA. Dodatkowo, zdeponowane są tutaj informacje na temat potencjału kodującego lncRNA, przewidziane krótkie ramki odczytu i informacja o ich podobieństwie do białek z bazy danych SwissProt (The UniProt Consortium, 2018), jak również poziomy ekspresji lncRNA w szeregu próbek.

Drugą bazą danych utworzoną w trakcie realizacji projektu jest KTCNlncDB (http://rhesus.amu.edu.pl/KTCNlncDB). Została ona zbudowana z wykorzystaniem

nowoczesnych technik webowych, a jej działanie zostało przetestowane na szeregu przeglądarek internetowych oraz na różnych systemach operacyjnych. Dostępne tutaj funkcjonalności i dane można podzielić na cztery kategorie, wykazujące częściowe podobieństwo do bazy danych CantataDB:

- ***Oddziaływania lncRNA z innymi transkryptami***. Jest to główna część bazy danych i składa się z trzech modułów. Pierwszy z nich to wybór kryteriów przeszukiwania; użytkownik może zawęzić wyświetlane wyniki pod kątem przewidzianej funkcji lncRNA, nazwy lncRNA, regulowanego genu, jak również jego bardziej szczegółowego opisu. Drugi moduł służy podsumowaniu wyników przeszukiwania: pojawia się informacja o użytych kryteriach wyszukiwania, o liczbie znalezionych rekordów, wykres kołowy podsumowujący funkcje lncRNA związane z wyszukanymi rekordami. Dostępna jest również opcja pobrania wszystkich wyświetlonych rekordów w postaci pliku tekstowego. Trzecia sekcja zawiera wyniki wyszukiwania, zaprezentowane w formie tabeli, gdzie jeden wiersz odpowiada jednemu oddziaływaniu lncRNA z inną cząsteczką RNA. Pośród dostępnych tutaj danych można wymienić m.in. nazwy oddziałujących cząsteczek RNA czy przewidziane funkcje lncRNA. Dostępny jest również przycisk o nazwie *Details*, który daje dostęp do bardziej szczegółowych danych na temat wybranego rekordu, wliczając poziomy ekspresji obu RNA czy też dodatkowe informacje na temat przewidzianych bioinformatycznie funkcji.
- ***Dane na temat lncRNA***. Ta podstrona pozwala na dostęp do wszystkich lncRNA, wliczając te, dla których nie przewidziano funkcji w kontekście oddziaływań RNA:RNA. Mamy tutaj do czynienia z podobnym podziałem na moduły, jak wyżej. Tabela z wynikami zawiera tutaj m,in. informację o położeniu na genomie czy biotypie transkryptu (np. lincRNA, antysensowny RNA). Po kliknięciu przycisku *Details*, użytkownik otrzymuje dostęp do sekwencji lncRNA w formacie FASTA, danych o poziomie ekspresji, podobieństwie sekwencji do lncRNA z bazy danych NONCODE, ewentualnym podobieństwie do białek zdeponowanych w bazie danych SwissProt oraz krótkich peptydach potencjalnie kodowanych przez lncRNA.
- ***Strona z implementacją programu BLAST***. Użytkownik ma możliwość przeszukiwania wszystkich sekwencji lncRNA zdeponowanych w bazie danych za pomocą programu BLAST.
- ***Pobieranie danych***. Oprócz spersonalizowanego pobierania danych dostępnego dla wyników wyszukiwania uzyskanych przez użytkownika, umożliwiono pobranie wszystkich najważniejszych danych związanych z projektem, jak sekwencje lncRNA czy zidentyfikowane oddziaływania RNA:RNA.

KTCNlncDB jest pierwszym katalogiem lncRNA ulegających ekspresji w stożku rogówki oraz ogólnie w rogówce człowieka; baza danych dostarcza przyjaznej dla użytkownika platformy, która integruje duże zestawy danych i umożliwia efektywne ich przeglądanie, wizualizację, jak również pobieranie.

5. Znaczenie naukowe wyników i ich potencjał aplikacyjny

W trakcie kilkuletnich badań nad długimi niekodującymi RNA udało mi sie zrealizować wyżej wymienione cele i jednocześnie dokonać interesujących odkryć na temat ich biologii. Wymienić tutaj należy:

- Opracowanie specjalistycznych potoków analitycznych służących do identyfikacji lncRNA i badania ich funkcji pełnionych w kontekście oddziaływań RNA:RNA. Wykazano, że są one szybkie i wydajne, co z pewnością stanowi istotny postęp w tej dynamicznie się rozwijającej dziedzinie. Mają one charakter uniwersalny, a więc mogą zostać wykorzystane w innych projektach badawczych, u różnych gatunków i w różnych układach eksperymentalnych (**Szczesniak i Makalowska 2016; Szczesniak et al., 2016; Szczesniak et al., 2019**).
- Opracowanie funkcjonalnej charakterystyki *in silico* dla tysięcy długich niekodujących RNA, u człowieka i roślin, dla których w znakomitej większości wcześniej nie istniały podobne dane. W szczególności, poznane zostały lncRNA, które najprawdpopodobniej powiązane są z etiologią stożka rogówki, choroby o wciąż słabo scharakteryzowanym podłożu molekularnym (**Bryzghalov et al., 2016; Szczesniak i Makalowska 2016; Szczesniak et al., 2016; Szczesniak et al., 2017**).
- Identyfikacja tysięcy nowych lncRNA u 39 gatunków roślin i glonów. Warto zaznaczyć, że dla większości z tych gatunków lncRNA nie były wcześniej w ogóle znane. Co więcej, dokonana została także ich funkcjonalna charakterystyka. Te lncRNA oraz ich adnotacje, wliczając przewidziane funkcje związane z modulowaniem splicingu i blokowaniem funkcji mikroRNA, stanowią bardzo ważny wkład w poznanie biologii tych cząsteczek u roślin. Identyfikacja lncRNA przeprowadzona została takżę w przypadku człowieka (*keratoconus*) oraz u szympansa; w tym drugim przypadku czternastokrotnie zwiększono pulę znanych lncRNA (w porównaniu z baza danych ENSEMBL) (**Bryzghalov et al., 2016; Szczesniak et al., 2016; Szczesniak et al., 2017; Szczesniak et al., 2019**).
- *Złożenie ab initio* transkryptomów 39 gatunków roślin i glonów, co pozwoliło na odkrycie tysięcy nowych genów i form splicingowych (**Szczesniak et al., 2016; Szczesniak et al., 2019**).
- Odkrycie lncRNA powstających w procesie retropozycji i obserwacja, że mogą one pełnić funkcje regulatorowe względem ich tzw. genów rodzicielskich. Te funkcje regulatorowe, pełnione w *trans*, moga zachodzic na różne sposoby i uzyskane zostało dla nich wsparcie bioinformatyczne w postaci sparowań RNA:RNA, analizy ekspresji, w tym korelacji ekspresji, czy konserwacji ewolucyjnej (**Bryzghalov et al., 2016**).
- Opracowanie dwóch nowych internetowych baz danych, nazwanych KTCNlncDB i CantataDB, które udostępniają społeczności naukowe uzyskane wyniki w przystępny sposób (**Szczesniak et al., 2016; Szczesniak et al., 2017; Szczesniak et al., 2019**).
- Praca przeglądowa na temat funkcji długich niekodujących RNA, ich udziale w etiologii chorób człowieka, wraz z omówieniem obiecujących terapii, z których niektóre znajdują się już w różnych fazach testów klinicznych (**Wanowska et al., 2018**).

Opisane tutaj badania stanowiły pierwszą próbę globalnego scharakteryzowania funkcji lncRNA pełnionych poprzez bezpośrednie parowanie z innymi transkryptami. Uzyskane wyniki poszerzają nasze zrozumienie biologii lncRNA, cząsteczek o których wciąż wiadomo stosunkowo niewiele, choć powiązano je już z wieloma ważnymi procesami komórkowymi oraz chorobami człowieka. W szczególności, zidentyfikowane zostały prawdopodobne funkcje tysięcy lncRNA, przez co w sposób istotny poszerzona została wiedza w zakresie procesów regulatorowych na poziomie transkryptomu. Duża część wyników umieszczona została w internetowych bazach danych, przez co są stosunkowo łatwo dostępne, podobnie jak nowo rozwinięte metody i potoki analityczne. Podsumowując, uzyskane wyniki, opracowane metody badawcze, bazy danych i związane z nimi publikacje w czasopismach naukowych posiadają duży potencjał naukowy i cieszą się zainteresowaniem badaczy w dziedzinie biologii molekularnej, medycyny czy biotechnologii, na co wskazują m.in. wskaźniki cytowań oraz odwiedziny na stronach internetowych projektu.

Moje publikacje związane z omawianym osiągnięciem naukowym

1. Bryzghalov O, Szcześniak MW, Makałowska I. (2016) Retroposition as a source of antisense long non-coding RNAs with possible regulatory functions. Acta Biochim Pol. 2016;63(4):825-833.
2. Szcześniak MW, Bryzghalov O, Ciomborowska-Basheer J, Makałowska I. (2019) CANTATAdb 2.0: Expanding the Collection of Plant Long Noncoding RNAs. Methods Mol Biol. 2019;1933:415-429.
3. Szczesniak MW, Kabza M, Karolak JA, Rydzanicz M, Nowak DM, Ginter-Matuszewska B, Polakowski P, Ploski R, Szaflik JP, Gajecka M. (2017) KTCNlncDB-a first platform to investigate lncRNAs expressed in human keratoconus and non-keratoconus corneas. Database (Oxford). 2017 Jan 10;2017.
4. Szcześniak MW, Makałowska I. (2016) lncRNA-RNA Interactions across the Human Transcriptome. PLoS One. 2016 Mar 1;11(3):e0150353.
5. Szcześniak MW, Rosikiewicz W, Makałowska I. (2016) CANTATAdb: A Collection of Plant Long Non-Coding RNAs. Plant Cell Physiol. 2016 Jan;57(1):e8.
6. Wanowska E, Kubiak MR, Rosikiewicz W, Makałowska I, Szcześniak MW. (2018) Natural antisense transcripts in diseases: From modes of action to targeted therapies. Wiley Interdiscip Rev RNA. 2018 Mar;9(2).

Literatura dodatkowa

1. Athanasiadis A, Rich A, Maas S. (2004) Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biol. 2004 Dec;2(12):e391.
2. Bardou F, Ariel F, Simpson CG, Romero-Barrios N, Laporte P, Balzergue S, Brown JW, Crespi M. (2014) Long noncoding RNA modulates alternative splicing regulators in Arabidopsis. Dev Cell. 2014 Jul 28;30(2):166-76. doi:10.1016/j.devcel.2014.06.017.

3. Bazin J, Baerenfaller K, Gosai SJ, Gregory BD, Crespi M, Bailey-Serres J. (2017) Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. Proc Natl Acad Sci U S A. 2017 Nov 14;114(46):E10018-E10027. doi: 10.1073/pnas.1708433114.
4. Beltran M, Puig I, Peña C, García JM, Alvarez AB, Peña R, Bonilla F, de Herreros AG. (2008) A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. Genes Dev. 2008 Mar 15;22(6):756-69. doi: 10.1101/gad.455708.
5. Cheung IM, McGhee CNj, Sherwin T. (2014) Deficient repair regulatory response to injury in keratoconic stromal cells. Clin Exp Optom. 2014 May;97(3):234-9. doi:10.1111/cxo.12118.
6. Ciomborowska J, Rosikiewicz W, Szklarczyk D, Makałowski W, Makałowska I. (2013) "Orphan" retrogenes in the human genome. Mol Biol Evol. 2013 Feb;30(2):384-96. doi: 10.1093/molbev/mss235.
7. Dai X, Zhao PX. (2011) psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res. 2011; 39:W155–9. pmid:21622958
8. Daniel C, Silberberg G, Behm M, Öhman M. (2014) Alu elements shape the primate transcriptome by cis-regulation of RNA editing. Genome Biol. 2014 Feb 3;15(2):R28. doi: 10.1186/gb-2014-15-2-r28.
9. ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247.
10. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. (2003) MicroRNA targets in Drosophila. Genome Biol. 2003; 5(1):R1. pmid:14709173
11. Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, Morgan TE, Finch CE, St Laurent G 3rd, Kenny PJ, Wahlestedt C. (2008) Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nat Med. 2008 Jul;14(7):723-30. doi: 10.1038/nm1784.
12. Gong C, Maquat LE. (2011) lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. Nature. 2011 Feb 10;470(7333):284-8. doi:10.1038/nature09701.
13. Han SP, Tang YH, Smith R. (2010) Functional diversity of the hnRNPs: past, present and perspectives. Biochem J. 2010 Sep 15;430(3):379-92. doi: 10.1042/BJ20100396.
14. Hu S, Wang X, Shan G. (2016) Insertion of an Alu element in a lncRNA leads to primate-specific modulation of alternative splicing. Nat Struct Mol Biol. 2016 Nov;23(11):1011-1019. doi: 10.1038/nsmb.3302.
15. Huang X, Yuan T, Tschannen M, Sun Z, Jacob H, Du M, Liang M, Dittmar RL, Liu Y, Liang M, Kohli M, Thibodeau SN, Boardman L, Wang L. (2013) Characterization of human plasma-derived exosomal RNAs by deep sequencing. BMC Genomics. 2013 May 10;14:319. doi: 10.1186/1471-2164-14-319.
16. Kawahara Y, Zinshteyn B, Sethupathy P, Iizasa H, Hatzigeorgiou AG, Nishikura K. (2007) Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. Science. 2007 Feb 23;315(5815):1137-40.
17. Kersey PJ, Allen JE, Allot A, Barba M, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Grabmueller C et al. (2018) Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Res. 2018 Jan 4;46(D1):D802-D808. doi:10.1093/nar/gkx1011.

18. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. (2011) Adaptive seeds tame genomic sequence comparison. Genome Res. 2011; 21(3):487–93. pmid:21209072
19. Kim DH, Xi Y, Sung S. (2017) Modular function of long noncoding RNA, COLDAIR, in the vernalization response. PLoS Genet. 2017 Jul 31;13(7):e1006939. doi: 10.1371/journal.pgen.1006939.
20. Kodama Y, Shumway M, Leinonen R; International Nucleotide Sequence Database Collaboration. (2012) The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. 2012 Jan;40(Database issue):D54-6. doi: 10.1093/nar/gkr854.
21. Kugel JF, Goodrich JA. (2012) Non-coding RNAs: key regulators of mammalian transcription. Trends Biochem Sci. 2012 Apr;37(4):144-51. doi:10.1016/j.tibs.2011.12.003.
22. Li J, Ma W, Zeng P, Wang J, Geng B, Yang J, Cui Q. (2015) LncTar: a tool for predicting the RNA targets of long noncoding RNAs. Brief Bioinform. 2015;
23. Li JH, Liu S, Zhou H, Qu LH, Yang JH. (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res. 2014 Jan;42(Database issue):D92-7. doi: 10.1093/nar/gkt1248.
24. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF et al. (2011) ViennaRNA Package 2.0. Algorithms Mol Biol. 2011; 6:26. pmid:22115189
25. Mayeda A, Munroe SH, Xu RM, Krainer AR. (1998) Distinct functions of the closely related tandem RNA-recognition motifs of hnRNP A1. RNA. 1998 Sep;4(9):1111-23.
26. Morgan JT, Murphy CJ, Russell P. (2013) What do mechanotransduction, Hippo, Wnt, and TGFβ have in common? YAP and TAZ as key orchestrating molecules in ocular health and disease. Exp Eye Res. 2013 Oct;115:1-12. doi: 10.1016/j.exer.2013.06.012.
27. Nowak DM, Gajecka M. (2011) The genetics of keratoconus. Middle East Afr J Ophthalmol. 2011 Jan;18(1):2-6. doi: 10.4103/0974-9233.75876.
28. Paytuví Gallart A, Hermoso Pulido A, Anzar Martínez de Lagrán I, Sanseverino W, Aiese Cigliano R. (2016) GREENC: a Wiki-based database of plant lncRNAs. Nucleic Acids Res. 2016 Jan 4;44(D1):D1161-6. doi: 10.1093/nar/gkv1215.
29. Quek XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME. (2015) lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Res. 2015 Jan;43(Database issue):D168-73. doi:10.1093/nar/gku988.
30. Quinlan AR, Hall IM. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010; 26(6):841–2. pmid:20110278
31. Rabinowitz YS. (1998) Keratoconus. Surv Ophthalmol. 1998 Jan-Feb;42(4):297-319.
32. Ramaswami G, Li JB. (2014) RADAR: a rigorously annotated database of A-to-I RNA editing. Nucleic Acids Res. 2014 Jan;42(Database issue):D109-13. doi:10.1093/nar/gkt996.
33. Rosikiewicz W, Kabza M, Kosinski JG, Ciomborowska-Basheer J, Kubiak MR, Makalowska I. (2017) RetrogeneDB-a database of plant and animal retrocopies. Database (Oxford). 2017 Jan 1;2017. doi: 10.1093/database/bax038.
34. Tafer H, Hofacker IL. (2008) RNAplex: a fast tool for RNA-RNA interaction search. Bioinformatics 2008; 24(22):2657–63. pmid:18434344
35. Tay Y, Rinn J, Pandolfi PP. (2014) The multilayered complexity of ceRNA crosstalk and competition. Nature. 2014 Jan 16;505(7483):344-52. doi: 10.1038/nature12986.

36. UniProt Consortium T. (2018) UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2018 Mar 16;46(5):2699. doi: 10.1093/nar/gky092.
37. Wang Y, Li J, Deng XW, Zhu D. (2018) Arabidopsis noncoding RNA modulates seedling greening during deetiolation. Sci China Life Sci. 2018 Feb;61(2):199-203. doi:10.1007/s11427-017-9187-9.
38. Yi X, Zhang Z, Ling Y, Xu W, Su Z. (2015) PNRD: a plant non-coding RNA database. Nucleic Acids Res. 2015 Jan;43(Database issue):D982-9. doi: 10.1093/nar/gku1162.
39. Yin QF, Yang L, Zhang Y, Xiang JF, Wu YW, Carmichael GG, Chen LL. (2012) Long noncoding RNAs with snoRNA ends. Mol Cell. 2012 Oct 26;48(2):219-30. doi:10.1016/j.molcel.2012.07.033.
40. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, To JK, Laird MR, Lavidas I, Liu Z, Loveland JE, Maurel T, McLaren W, Moore B, Mudge J, Murphy DN, Newman V, Nuhn M, Ogeh D, Ong CK, Parker A, Patricio M, Riat HS, Schuilenburg H, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Zadissa A, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Cunningham F, Yates A, Flicek P. (2018) Ensembl 2018. Nucleic Acids Res. 2018 Jan 4;46(D1):D754-D761. doi: 10.1093/nar/gkx1098.
41. Zhao Y, Li H, Fang S, Kang Y, Wu W, Hao Y, Li Z, Bu D, Sun N, Zhang MQ, Chen R. (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. Nucleic Acids Res. 2016 Jan 4;44(D1):D203-8. doi: 10.1093/nar/gkv1252.

5. Omówienie pozostałych osiągnięć naukowo - badawczych (artystycznych).

Badanie małych regulatorowych RNA u zwierząt i roślin: rozwijanie specjalistycznego oprogramowania, baz danych oraz analiza danych

Jednym z moich pierwszych obszarów badawczych były mikroRNA, ściśle związane z projektem doktorskim realizowanym w Pracowni Bioinformatyki (Uniwersytet im. Adama Mickiewicza) pod kierunkiem prof. dr hab. Izabeli Makałowskiej. Są one zdefiniowane jako krótkie cząsteczki RNA (typowo 19-24 nt), których funkcje regulatorowe u roślin i zwierząt polegają na hamowaniu translacji bądź degradacji cząsteczek mRNA. Pomimo znacznego postępu w identyfikacji i analizie mikroRNA, wiele z nich pozostaje nieodkrytych, wliczając organizmy modelowe. Zważywszy na niezwykle ważną rolę tych cząsteczek w funkcjonowaniu komórki, wyraźna jest potrzeba rozwijania metod i ich dalszej, wielkoskalowej identyfikacji. Mając to na uwadze, postanowiłem przeprowadzić wielkoskalową analizę sekwencji EST z bazy danych dbEST (Boguski et al., 1993), czyli znaczników sekwencji ulegających ekspresji, celem identyfikacji nowych miRNA. Sekwencje EST były tutaj dobrym punktem startowym, gdyż są dostępne dla setek gatunków roślin i zwierząt, co umożliwiło szukanie miRNA nawet u gatunków, u których wcześniej ich nie analizowano. W tym celu zbudowany został potok analityczny, pozwalający na szukanie nowych miRNA na zasadzie podobieństwa do znanych dojrzałych miRNA. Liczne etapy sprawdzania i filtrowania danych zapewniły bardzo wysoką czułość i specyficzność algorytmu.

Załącznik 2

Potok analityczny wykorzystano w wielkoskalowej analizie sekwencji EST, identyfikując 10 004 miRNA u 221 gatunków zwierząt i 199 gatunków roślin. Te dane wynikowe uzupełniono sekwencjami miRNA z innych źródeł: miRBase, PMRD, microPC oraz z publikacji, jak również szeregiem danych na temat biologii miRNA z różnych źródeł. Dane zdeponowano w nowo utworzonej internetowej bazie danych, nazwanej miRNEST (http://mirnest.amu.edu.pl), **czyniąc ją najprawdopodobniej największa i najbardziej wszechstronną bazą danych mikroRNA. W przypadku dziesiątek gatunków roślin i zwierząt sekwencje mikroRNA można znaleźć wyłącznie w tej bazie (Szczesniak et al., 2012a).** Późniejsza aktualizacja bazy danych miRNEST do wersji 2.0 wiązała się przede wszystkim z dwukrotnym zwiększeniem liczby zdeponowanych tutaj mikroRNA, zaś liczba gatunków roślin i zwierząt oraz wirusów sięgnęła 544. Było to możliwe głównie dzięki wykorzystaniu danych smallRNA-Seq z bazy danych GEO (Clough i Barrett, 2016) i ich analizie nakierowanej na identyfikację mikroRNA, z wykorzystaniem własnego algorytmu. Ponadto, dzięki wykorzystaniu danych smallRNA-Seq, udało się potwierdzić istnienie około 40% miRNA zdeponowanych w wersji 1.0. Dodatkowe cechy nowej wersji bazy danych to: i) wyniki analizy degradomów dziesięciu gatunków roślin, której celem było znalezienie sekwencji docelowych dla miRNA; ii) pojawiła się zakładka o splicingu cząsteczek pri-miRNA; iii) poprawiony został interfejs graficzny, a działanie samej bazy danych zostało przyspieszone; iv) dodano opcję masowego pobierania danych oraz pobierania zawężonego - według kryteriów użytkownika; v) dodano możliwość przysyłania mikroRNA przez użytkowników. **Wszystkie te zmiany sprawiają, że miRNEST pozostaje prawdopodobnie najobszerniejszą bazą danych mikroRNA i posiada szereg unikalnych danych, niedostępnych w innych repozytoriach, jak np. na temat występowania izomirów, sekwencji docelowych miRNA czy miejsc splicingowych (Szczesniak et al., 2014b)**. Warto wspomnieć, że baza miRNEST cieszy się dużą popularnością: ponad 100 tys. odwiedzin w ciągu kilku lat.

Zainteresowało mnie także zagadnienie identyfikacji mikroRNA bez danych transkryptomicznych. Obecnie dostępne narzędzia służące do identyfikacji miRNA *de novo* obarczone są wadami metodologicznymi oraz znacznymi ograniczeniami w ich używaniu. Na przykład często są tworzone wyłącznie z myślą o gatunkach modelowych; do ich testowania wykorzystuje się zbiór treningowy (zbiór testowy i treningowy powinny być rozłączne); rozpatruje się tylko jedną, wybraną metodę nauczania maszynowego; używa się niskiej jakości sekwencji zbiorów pozytywnych i negatywnych; nie uwzględnia się problemu niezbalansowania rozmiaru zestawów pozytywnych i negatywnych, co skutkuje niewłaściwym oszacowaniem wydajności klasyfikatora. Opracowując HuntMi, narzędzie do identyfikacji miRNA *de novo* w oparciu o techniki nauczania maszynowego, podjęliśmy próbę rozwiązania tych problemów, jednocześnie mając na celu maksymalizację czułości i specyficzności. W pierwszej kolejności przygotowaliśmy wysokiej jakości dane wejściowe. Wprowadziliśmy siedem nowych cech do problemu klasyfikacji miRNA i przetestowaliśmy kilka metod nauczania maszynowego, dla każdej z nich przeprowadzając optymalizację parametrów wejściowych. **Podjęliśmy także problem niezbalansowania zbiorów treningowych, tzn. różnicy między wielkością zbioru pozytywnego i negatywnego. W tym celu zaimplementowaliśmy nową technikę, nazwaną**

przez nas ROC-select, która okazała się być lepsza od innych znanych metod rozwiązywania problemu niezbalansowania, przynajmniej w dziedzinie identyfikacji miRNA. Naszą metodę porównaliśmy z wiodącymi narzędziami do identyfikacji miRNA *de novo*: microPred, PlantMiRNAPred (Xuan et al., 2011) i MiRenSVM (Ding et al., 2010). Okazało się, że pod względem wydajności oraz szybkości nasz algorytm je przewyższa. Oprócz wyżej wspomnianych cech, niewątpliwą zaletą HuntMi jest jego elastyczność, gdyż na przykład pozwala użytkownikowi w łatwy sposób tworzyć własne klasyfikatory - w oparciu o dane z dowolnego gatunku - aby je następnie wykorzystać podczas identyfikacji miRNA. Narzędzie można pobrać ze strony http://adaa.polsl.pl/agudys/huntmi/huntmi.htm (**Gudyś et al., 2013**)

W oparciu o doświadczenie, które już na tym etapie posiadałem, wszedłem we współpracę z dr Michaelem J. Axtellem z Pennsylvania State University, USA, w ramach projektu mającego na celu scharakteryzowania transkryptomu małych RNA u jabłoni. U jabłoni (*Malus domestica*) znanych było dotychczas zaledwie 200 mikroRNA, co porównując ze spokrewnionymi gatunkami pozwalało przypuszczać, że jest to jedynie część repertuaru miRNA. Aby lepiej scharakteryzować mikroRNA tego gatunku, wykonaliśmy sekwencjonowanie 12 bibliotek małych RNA - osobno dla drzew odpornych na zarazę ogniową oraz drzew na nią wrażliwych. W toku analizy danych, zidentyfikowaliśmy 116 nowych mikroRNA, jak również potwierdziliśmy istnienie większości dotychczas już scharakteryzowanych mikroRNA. Istnienie wybranych cząsteczek miRNA potwierdziliśmy eksperymentalnie metodami RT-PCR oraz stem-loop qPCR. Zidentyfikowaliśmy również sekwencje docelowe dla mikroRNA w transkryptomie jabłoni. **Analiza porównawcza ekspresji mikroRNA pomiędzy roślinami odpornymi i wrażliwymi na zarazę ogniową pozwoliła wytypować markery wrażliwości na tę chorobę atakującą jabłonie i grusze i przynoszącą ogromne straty ekonomiczne**. W szczególności zidentyfikowaliśmy cztery mikroRNA potencjalnie uczestniczące w budowaniu odporności na zarazę ogniową: mdm-miR169a, mdm-miR160e, mdm-miR167b-g oraz mdm-miR168a,b (**Kaja et al., 2014**). Warto wspomnieć, że te same mikroRNA powiązano już z odpowiedzią na stresy u innych gatunków roślin. Znaczna część badań, wliczając testy laboratoryjne, wykonana została przez mgr Elżbietę Kaja w ramach jej projektu doktorskiego (obronionego 12 kwietnia 2017 r.), **ze mną jako promotorem pomocniczym**. Na zwieńczenie projektu, większość wyników została zdeponowana w nowo utworzonej internetowej bazie danych, dostępnej pod adresem http://lemur.amu.edu.pl/share/apple_miRNAs, zaś nowe mikroRNA dodatkowo umieszczono w bazie danych miRBase.

W ramach badań nad mikroRNA zaangażowałem się w dodatkowe projekty badawcze:

- Uczestniczyłem w tworzeniu bazy danych miREX, kolekcjonującej dane o ekspresji 190 pri-miRNA *Arabidopsis thaliana*, dostępnej pod adresem http://bioinfo.amu.edu.pl/mirex. Poziomy ekspresji określone zostały metodą RT-qPCR. Unikalną cechą tej bazy danych jest to, że pozwala na porównywanie ekspresji miRNA pomiędzy organami oraz stadiami rozwoju rośliny. Ponadto, zdeponowane zostały tutaj starannie dobrane dodatkowe dane, pomagające w interpretacji obserwowanych wyników na temat ekspresji, co wraz z

nowoczesnym interfejsem graficznym czyni miREX-a narzędziem łatwym w obsłudze i przyjaznym dla użytkownika (**Bielewicz et al., 2012**).

- Brałem udział w badaniach niepożądanych efektów obserwowanych w technologii RNAi, we współpracy z grupą prof. dr hab. Włodzimierza Krzyżosiaka z Instytutu Chemii Bioorganicznej, PAN. Mechanizm interferencji RNA (RNAi) wykorzystywany jest w biologii molekularnej do wyciszania ekspresji genów i można tutaj wymienić takie techniki, jak siRNA (ang. *short interfering RNA*), shRNA (ang. *short hairpin RNA*) i sztuczne miRNA (ang. *short hairpin miRNA*, sh-miR). Oprócz wyciszania genów docelowych, często obserwuje się również niepożądane efekty, wliczając stymulację odpowiedzi immunologicznej. Naszym celem było wytypowanie markerów odpowiedzi immunologicznej przy użyciu różnych reagentów RNAi. Wykazaliśmy, że RNAi powoduje szereg niepożądanych efektów, wliczając deregulację endogennych mikroRNA. siRNA nieznacznie indukują odpowiedź immunologiczną z udziałem interferonów, jednak mogą prowadzić do wysycenia ścieżek biogenezy miRNA, co skutkuje spadkiem poziomu ekspresji miRNA, zwłaszcza tych, których komórkowy poziom jest typowo wysoki. Z kolei reagenty oparte na plazmidach nie tylko indukują szereg markerów odpowiedzi immunologicznej, ale również mogą prowadzić do zaburzeń ekspresji miRNA (**Olejniczak et al., 2016**).
- We współpracy z tą samą grupą badawczą uczestniczyłem w badaniach mających na celu analizę czynników wpływających na precyzję obróbki cząsteczek sh-miR. sh-miR (ang. *short hairpin miRNA*) to sztuczne cząsteczki pri-mikroRNA, stosowane do wyciszania ekspresji genów na zasadzie interferencji RNA. Cząsteczki sh-miR są obiecującymi reagentami RNAi, m.in. ze względu na ich niższą toksyczność niż w przypadku innych reagentów, takich jak shRNA. Używając techniki northern blot oraz wysokoprzepustowego sekwencjonowania transkryptomu małych RNA (smallRNA-Seq) zbadaliśmy dokładność, z jaką RNazy Drosha i Dicer tną cząsteczki sh-miR. Zauważyliśmy, że wpływ na nią ma sekwencja funkcjonalnego siRNA, kodowanego przez sh-miR, jak również sekwencja i struktura drugorzędowa całej cząsteczki sh-miR. Wyniki te ułatwią projektowanie bardziej wydajnych reagentów w technologii RNAi (**Galka-Marciniak et al., 2016**).
- Jestem współautorem pracy przeglądowej na temat wysokoprzepustowego sekwencjonowanie transkryptomów w technologii NGS, czyli RNA-Seq (**Conesa et al., 2016**). Jest to technika o bardzo szerokich zastosowaniach, co wiąże się z brakiem jednego, uniwersalnego sposobu analizy danych. Mając to na względzie, przygotowaliśmy opracowanie, w którym udzielamy wskazówek dotyczących różnych aspektów analizy danych RNA-Seq, wliczając kontrolę jakości, mapowanie odczytów do genomu referencyjnego, analizę ekspresji różnicowej, analizę alternatywnego splicingu czy wykrywanie fuzji genowych. Zwracamy przy tym szczególną uwagę na mogące się pojawić trudności i doradzamy jak sobie z nimi poradzić. Poświęciliśmy także dużo miejsca takim zagadnieniom, jak analiza transkryptomów małych RNA (smallRNA-Seq, z

moim decydującym udziałem), integracja danych pochodzących z różnych technik eksperymentalnych z danymi RNA-Seq oraz metody wizualizacji danych. Praca ta cieszy się ogromnym zainteresowaniem społeczności naukowej. Dotychczas była cytowana 337 razy i znalazła się wśród trzech najlepiej cytowanych prac opublikowanych w *Genome Biology* w 2016 roku.

- Współtworzyłem również inną pracę przeglądową, na temat baz danych miRNA oraz aktualnych metod ich identyfikacji, zarówno z wykorzystaniem podejść *in silico,* jak i tych eksperymentalnych, omawiając ich plusy i minusy. Opisaliśmy również sposoby eksperymentalnej walidacji cząsteczek mikroRNA i ich przewidzianych oddziaływań z tzw. sekwencjami docelowymi, a przede wszystkim przedstawiliśmy bazy danych mikroRNA, dostarczając bliższej charakterystyki wybranych przykładów (**Szcześniak et al., 2012b**).

Badanie molekularnych i ewolucyjnych właściwości intronów i miejsc splicingowych

Drugim obszarem moich zainteresowań jest splicing oraz intronowo-egzonowa struktura genów, którym zająłem się już trakcie pisania literaturowej pracy licencjackiej pod kierunkiem prof. dr hab. Zofii Szweykowskiej-Kulińskiej w Zakładzie Ekspresji Genów (Uniwersytet im. Adama Mickiewicza). Po przeredagowaniu, została opublikowana jako praca przeglądowa pt. "Regulacja alternatywnego splicingu" (**Szcześniak i Szweykowska-Kulińska, 2009**). W publikacji tej omówione zostały mechanizmy splicingu i alternatywnego splicingu oraz powiązane z nimi mechanizmy regulatorowe. W ramach pracy magisterskiej na kierunku Biotechnologia, w tej samej grupie badawczej zajmowałem się białkami CBP20 i CBP80 u jęczmienia (*Hordeum vulgare*). Wchodzą one w skład kompleksu białek wiążących się z kapem (ang. *cap-binding complex*, CBC). Białka CBP20 i CBP80 są tutaj niezbędne, aby kap (zwany również czapeczką) prawidłowo spełniał swoje funkcje, m.in. w splicingu pre-mRNA i transporcie RNA z jądra do cytoplazmy. Uzyskane przeze mnie wyniki zostały włączone do niedawno opublikowanej pracy (**Pieczyński et al., 2018**), w której charakteryzowany był intron typu U12 w genie *CBP20*. Struktura genu *CBP20* jest silnie zakonserwowana ewolucyjnie, z ośmioma egzonami i siedmioma intronami, z których czwarty jest tzw. intronem typu U12. W niniejszej pracy zbadano funkcje tego intronu typu U12, poprzez zmianę jego lokalizacji w genie. Okazało się, że relokacja intronu U12 skutkuje znacznym obniżeniem wydajności jego splicingu i akumulacją trranskryptów, których splicing przebiegł w sposób nieprawidłowy. Również zastąpienie intronu U12 intronem typu U2 prowadzi do zaburzeń splicingu. Obserwacje te wskazują, że obecność intronu typu U12 pomiędzy egzonem czwartym a piątym jest niezbędna do prawidłowego splicingu pre-mRNA i produkcji funkcjonalnego białka CBP20. Tematem drugiej pracy magisterskiej (na kierunku Bioinformatyka) było badanie ewolucyjnych aspektów splicingu. Wiedzieliśmy, że geny ortologiczne często różnią się budową egzonowo-intronową, co jest efektem takich procesów, jak fuzja genów czy intronizacja – zjawisko polegające na przekształceniu fragmentu sekwencji egzonu w nowy intron. Zaintrygował nas fakt, że o ile u różnych grup organizmów porównywalnie często obserwuje się zjawisko utraty jak i nabycia nowych intronów, u ssaków nie wykryto ani

jednej intronizacji, wobec wielu przypadków utraty intronów. Aby dokładniej przyjrzeć się temu zagadnieniu, wykonaliśmy analizę genomów ssaków pod kątem intronizacji i udało nam się zidentyfikować dwa geny, które nabyły nowe introny: *RNF113B* oraz *DCAF12L2*. Okazało się, że oba geny są retrogenami, a więc kopiami tzw. genów rodzicielskich powstałymi w procesie retropozycji. Intron w *RNF113B* jest specyficzny dla ssaków naczelnych, natomiast *DCAF12L2* przeszedł dwie niezależne intronizacje: u wspólnego przodka naczelnych i gryzoni oraz u gryzoni. *RNF113B* przeanalizowaliśmy eksperymentalnie i ku naszemu zaskoczeniu okazało się, że intron tego genu jest przedmiotem alternatywnego splicingu: intron jest wycinany z pre-mRNA tylko w jądrach, na dwanaście analizowanych organów. **Jest to nie tylko pierwszy przypadek intronizacji u ssaków, ale również pierwszy opisany przypadek alternatywnego splicingu w retrogenie (Szcześniak et al., 2011).**

W trakcie realizacji projektu doktorskiego pod kierunkiem prof. dr hab. Izabeli Makałowskiej podtrzymałem swoje zainteresowanie zagadnieniem splicingu, zwłaszcza u roślin, gdzie badania nad splicingiem dopiero nabierały rozpędu. Po miesiącach analiz otrzymałem ogromne ilości unikalnych danych, które zostały zdeponowane w nowej internetowej bazie danych, nazwanej ERISdb. ERISdb to baza danych miejsc splicingowych u siedmiu modelowych gatunków roślin: *Arabidopsis thaliana, Chlamydomonas reinhardtii, Glycine max, Oryza sativa, Physcomitrella patens, Selaginella moellendorffii* i *Zea mays*. Tym co odróżnia ERISdb od innych baz danych miejsc splicingowych jest wszechstronność i charakter zdeponowanych tutaj informacji, które zazwyczaj są niedostępne w innych bazach danych lub dostępne tylko częściowo. Należy tutaj wymienić m.in. identyfikację miejsc rozgałęzienia, traktów polipirymidynowych, traktów bogatych w AU, elementów regulatorowych działających w *cis* czy szukanie homologicznych miejsc splicingowych. Wykorzystując sekwencje EST oraz dane RNA-Seq znaleziono eksperymentalne potwierdzenie dla tysięcy miejsc splicingowych. Oprócz tego, zaklasyfikowano wszystkie introny do jednej z dwóch kategorii: U2 lub U12. W tym celu opracowano niezwykle wydajny klasyfikator (metoda lasów losowych) służący do odróżniania intronów typu U2 od U12. Jest to prawdopodobnie pierwsze tego typu narzędzie dla roślinnych miejsc splicingowych. Ostatecznie, przeprowadzono identyfikację miejsc splicingowych w genach mikroRNA, przez co baza ERISdb stała się jedynym repozytorium przechowującym tego rodzaju dane (później podobne dane zamieściliśmy w bazie danych miRNEST). Baza ERISdb jest dostępna pod adresem http://lemur.amu.edu.pl/share/ERISdb/ (**Szczesniak et al., 2013**).

Pozostała aktywność naukowa

W trakcie studiów doktoranckich odbyłem staż na Uniwersytecie Tokijskim w grupie prof. Yutaka Suzuki, gdzie uczestniczyłem w projekcie mającym na celu zsekwencjonowanie, złożenie oraz adnotację genomu nietoperza *Rousettus leschenaulti*. Moim zadaniem było złożenie i staranna adnotacja genomu mitochondrialnego, w oparciu o całogenomowe sekwencjonowanie w technologii Illumina. Zaadnotowany genom został zdeponowany w bazie danych GenBank (ID: KC702803). Dodatkowo, na podstawie analizy filogenetycznej genomów mitochondrialnych 18 gatunków nietoperzy, potwierdzona została słuszność podziału *Chiroptera* na *Yinpterochiroptera* i *Yangochiroptera* (**Szcześniak et al., 2014b**).

Niedawno uczestniczyłem w projekcie kierowanym przez prof. dr hab. n. med. Marzenę Gajęcką z Instytutu Genetyki człowieka PAN i Uniwersytetu Medycznego w Poznaniu, którego celem było scharakteryzowanie molekularnego podłoża stożka rogówki. Współpraca ta zaowocowała wcześniej wspomnianą pracą na temat potencjalnego udziału lncRNA w etiologii tej choroby. W innej fazie projektu po raz pierwszy użyliśmy wysokoprzepustowego sekwencjonowania transkryptomów w technologii NGS (RNA-Seq). Analiza obejmowała 25 próbek pochodzących od pacjentów ze stożkiem rogówki oraz 25 próbek od pacjentów z innymi dolegliwościami rogówki (grupa kontrolna). Analiza ekspresji różnicowej wykazała szereg genów, których ekspresja jest zmieniona u pacjentów ze stożkiem rogówki. Wyniki te wskazują na szeroko zakrojone zaburzenia syntezy kolagenu oraz ścieżek sygnalizacyjnych z udziałem białek TGF-β, Hippo i Wnt, co z kolei powiązano z występującymi w chorobie zmianami organizacji tkanki łącznej rogówki na poziomie molekularnym (**Kabza et al., 2017**).

Moje publikacje w czasopismach naukowych związane z dodatkowymi osiągnięciami naukowymi

1. Bielewicz D, Dolata J, Zielezinski A, Alaba S, Szarzynska B, Szczesniak MW, Jarmolowski A, Szweykowska-Kulinska Z, Karlowski WM. (2012) mirEX: a platform for comparative exploration of plant pri-miRNA expression data. Nucleic Acids Res. 40:D191-D197.
2. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016 Jan 26;17:13. doi:10.1186/s13059-016-0881-8.
3. Galka-Marciniak P, Olejniczak M, Starega-Roslan J, Szczesniak MW, Makalowska I, Krzyzosiak WJ. siRNA release from pri-miRNA scaffolds is controlled by the sequence and structure of RNA. Biochim Biophys Acta. 2016 Apr;1859(4):639-49. doi: 10.1016/j.bbagrm.2016.02.014.
4. Gudyś A, Szcześniak MW, Sikora M, Makalowska I. (2013) HuntMi: an efficient and taxon-specific approach in pre-miRNA identification. BMC Bioinformatics 14:83, doi:10.1186/1471-2105-14-83.
5. Kabza M, Karolak JA, Rydzanicz M, Szcześniak MW, Nowak DM, Ginter-Matuszewska B, Polakowski P, Ploski R, Szaflik JP, Gajecka M. Collagen synthesis disruption and downregulation of core elements of TGF-β, Hippo, and Wnt pathways in keratoconus corneas. Eur J Hum Genet. 2017 May;25(5):582-590. doi: 10.1038/ejhg.2017.4.
6. Kaja E, Szcześniak MW, Jensen PJ, Axtell MJ, McNellis T, Makałowska I (2014) Identification of apple miRNAs and their potential role in fire blight resistance. Tree Genetics & Genomes 11:812.
7. Olejniczak M, Urbanek MO, Jaworska E, Witucki L, Szczesniak MW, Makalowska I, Krzyzosiak WJ. Sequence-non-specific effects generated by various types of RNA

interference triggers. Biochim Biophys Acta. 2016 Feb;1859(2):306-14. doi: 10.1016/j.bbagrm.2015.11.005. Epub 2015 Nov 22.

8. Pieczynski M, Kruszka K, Bielewicz D, Dolata J, Szczesniak M, Karlowski W, Jarmolowski A, Szweykowska-Kulinska Z. A Role of U12 Intron in Proper Pre-mRNA Splicing of Plant Cap Binding Protein 20 Genes. Front Plant Sci. 2018 Apr 16;9:475. doi: 10.3389/fpls.2018.00475.
9. Szcześniak MW, Ciomborowska J, Nowak W, Rogozin IB, Makałowska I. (2011) Primate and rodent specific intron gains and the origin of retrogenes with splice variants. Mol Biol Evol. 28(1):33-7.
10. Szczesniak MW, Deorowicz S, Gapski J, Kaczynski L, Makalowska I. (2012a) miRNEST database: an integrative approach in microRNA search and annotation. Nucleic Acids Res. 40: D198-D204.
11. Szcześniak MW, Kabza M, Pokrzywa R, Gudyś A., Makałowska I. (2013) ERISdb: a Database of Plant Splice Sites and Splicing Signals. Plant Cell and Physiol. 54(2):e10, doi: 10.1093/pcp/pct001.
12. Szczesniak MW, Makalowska I (2014a) miRNEST 2.0: a database of plant and animal microRNAs. Nucleic Acids Res. 42:D74-D77.
13. Szcześniak MW, Owczarkowska E, Gapski J, Makałowska I. (2012b) Bazy danych mikroRNA. Postepy Bioch. 58(1).
14. Szcześniak M, Szweykowska-Kulińska Z (2009) Regulacja alternatywnego splicingu. Postępy Biologii Komórki 36: 3–22.
15. Szcześniak M, Yoneda M, Sato H, Makałowska I, Kyuwa S, Sugano S, Suzuki Y, Makałowski W, Kai C (2014b). Characterization of the mitochondrial genome of Rousettus leschenaulti. Mitochondrial DNA. 2014 Dec;25(6):443-4. doi:10.3109/19401736.2013.809451.

Literatura dodatkowa

1. Boguski MS, Lowe TM, Tolstoshev CM. (1993) dbEST--database for "expressed sequence tags". Nat Genet. 1993 Aug;4(4):332-3.
2. Clough E, Barrett T. (2016) The Gene Expression Omnibus Database. Methods Mol Biol. 2016;1418:93-110. doi: 10.1007/978-1-4939-3578-9_5.
3. Ding J, Zhou S, Guan J. (2010) MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features. BMC Bioinformatics. 2010 Dec 14;11 Suppl 11:S11. doi: 10.1186/1471-2105-11-S11-S11.
4. Xuan P, Guo M, Liu X, Huang Y, Li W, Huang Y. (2011) PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs. Bioinformatics. 2011 May 15;27(10):1368-76. doi: 10.1093/bioinformatics/btr153.

Michał Szczęsny

Poznań, 18 kwietnia 2019 r.