Załącznik 3
Autoreferat w języku polskim
dr Barbara Uszczyńska-Ratajczak

Międzynarodowy Instytut Biologii Molekularnej i Komórkowej w Warszawie

Warszawa, Polska

Załącznik 3

1. IMIĘ I NAZWISKO: Barbara Uszczyńska-Ratajczak

2. DYPLOMY, STOPNIE NAUKOWE:

2019 **Dyplom** (Projektowanie systemów informatycznych z bazami danych)
Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska, Polska

2013 **Doktorat** (Nauki chemiczne z zakresu biochemii)
Instytut Chemii Bioorganicznej, Polska Akademia Nauk, Poznań, Polska
Tytuł: „*Optymalizacja ścieżek analizy danych niestandardowych uzyskiwanych przy użyciu mikromacierzy DNA*"
Promotor: prof. Piotr Kozłowski

2011 **Magister** (Bioinformatyka)
Wydział Biologii
Uniwersytet Adam Mickiewicza, Poznań, Polska
Tytuł: "*Wirtualne laboratorium genomiczne*"
Promotor: prof. Marek Figlerowicz

2008 **Magister** (Biotechnologia)
Wydział Chemiczny
Poliechnika Wrocławska, Polska
Tytuł: "*si-RNA mediated gene silencing of integrin β3 expression inhibits metastatic potential of A459 cells*"
Promotor: dr Anna Nasulewicz-Goldeman

3. HISTORIA ZATRUDNIENIA

2019 – **Starszy Badacz**
Międzynarodowy Instytut Biologii Molekularnej i Komórkowej w Warszawie
Warszawa, Polska

2017 – 2019 **Adiunkt**
Centrum Nowych Technologii, Uniwersytet Warszawski, Polska

2013 – 2016 **Adiunkt**
Centre for Genomic Regulation (CRG), Barcelona, Hiszpania

4. WSKAZANIE OSIĄGNIĘCIA WYNIKAJĄCEGO Z ART. 219 ust. 1 pkt 2 USTAWY:

A) Tytuł:

Genomiczna charakterystyka długich niekodujących RNA w genomach człowieka i myszy

B) Publikacje (liczba cytowań: 174):

* - równy wkład

H1. Julien Lagarde*, Barbara Uszczynska-Ratajczak*, Javier Santoyo-Lopez, Jose Manuel Gonzalez, Electra Tapanari, Jonathan M. Mudge, Charlie Steward, Laurens Wilming, Andrea Tanzer, Cédric Howald, Jacqueline Chrast, Alicia Vela-Boza, Antonio Rueda, Francisco J. Lopez-Domingo, Joaquin Dopazo, Alexandre Reymond, Roderic Guigó & Jennifer Harrow, *Extension of human lncRNA transcripts by RACE coupled with long read high-throughput sequencing (RACE-Seq)* Nat Commun. 17(7), 12339, 2016.

- URL: https://doi.org/10.1038/ncomms12339
- **Impact Factor (2016): 12,454, punkty MNiSW: 45, liczba cytowań: 27**
- **Wkład autora:** Barbara Uszczyńska-Ratajczak wniosła duży wkład w analizę danych oraz proces tworzenia manuskryptu.
- **Procentowy udział:** 30%

H2. Julien Lagarde*, Barbara Uszczynska-Ratajczak*, Silvia Carbonell, Silvia Perez-Lluch, Amaya Abad, Carrie Davis, Thomas Gingeras, Adam Frankish, Jennifer Harrow, Roderic Guigó, Rory Johnson *High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing*, Nat Genet, 49(12),1731-1740, 2017.

- URL: https://doi.org/10.1038/ng.3988
- **Impact Factor (2017): 27,125, punkty MNiSW: 50, liczba cytowań: 64**
- **Wkład autora:** Barbara Uszczyńska-Ratajczak miała duży wkład w zaprojektowanie części eksperymentalnej projektu, jak również w opracowanie części bioinformatycznej. Dodatkowo jej udział obejmował analizę danych oraz przygotowanie manuskryptu. Ponadto Barbara Uszczyńska-Ratajczak jest twórcą orygnialnego kodu stworzonego na potrzeby tego projektu, w tym oprogramowania do demultipleksowania odczytów PacBio.
- **Procentowy udział:** 30%

H3. Barbara Uszczynska-Ratajczak, Julien Lagarde, Adam Frankish, Roderic Guigó and Rory Johnson, *Towards a complete map of the human long non-coding RNA transcriptome*, Nat Rev Genet, 19(9), 535-548, 2018.

- URL: https://doi.org/10.1038/s41576-018-0017-y
- **Impact Factor (2018): 43,704, punkty MNiSW: 50, liczba cytowań: 83**
- **Wkład autora:** Barbara Uszczyńska-Ratajczak miała duży wkład w analizę danych oraz proces tworzenia manuskryptu.
- **Procentowy udział:** 50%

C) CEL NAUKOWY PRACY WRAZ Z OPISEM OSIĄGNIĘTYCH WYNIKÓW

1. Wstęp

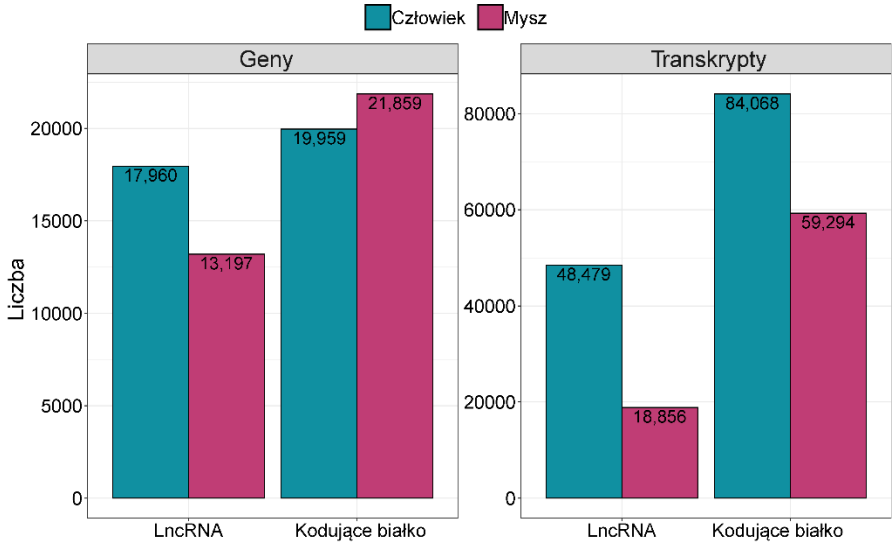
Poznanie sekwencji ludzkiego genomu jest niewątpliwie jednym z największych osiągnięć współczesnej biologii[1,2]. Jednocześnie osiągnięcie to zwróciło naszą uwagę na konieczność poszukiwania elementów funkcjonalnych w ludzkim genomie. Początkowe wysiłki skoncentrowane były głównie na identyfikacji genów kodujących białka i szybko ujawniły, że w ludzkim genomie znajduje się ich tylko ~19,000[3]. Co ciekawe, geny te stanowią tylko 1,1% naszego genomu. Oznacza to, że pozostałe 98,9% nie koduje białek i często jest nazywane niekodującym DNA lub „ciemną materią" genomu. Dane dla innych organizmów wielokomórkowych są mniej oczywiste, ale najprawdopodobniej zbliżone, w szczególności dla myszy[4,5].
Powyższe odkrycie nasuwa dwa bardzo ważne pytania: (1) Dlaczego tak duża część naszego genomu nie koduje białek? oraz (2) Jaką biologiczną funkcję w komórce pełni niekodujące DNA? Naukowcy z całego świata od klinicystów po biologów ewolucyjnych starają się odpowiedzieć na te pytania poprzez projektowanie wysokoprzepustowych, bioinformatycznych przedsięwzięć. Katalogi genów, które są rezultatem tych badań stanowią kluczowe zasoby z punktu widzenia eksploracji genomu. Jednak pomimo technologicznego i obliczeniowego postępu, identyfikacja elementów funkcjonalnych nadal jest bardzo trudna, nawet w przypadku dobrze poznanych genomów ssaków.

1.1. Liczba genów w genomach ssaków

Genom komórkowy zawiera zarówno geny kodujące białka, jak i geny niekodujące. Białka jako końcowy produkt ekspresji genów są łatwe do zidentyfikowania i dalszej charakterystyki. Dlatego też to geny kodujące białka początkowo zdominowały nasze spojrzenie na genomikę. Od dawna zakładano, że złożoność organizmu zależy i wzrasta wraz z ich liczbą. Wczesne szacunki wskazywały, że liczba genów kodujących białka w ludzkim genomie mieści się w przedziale od kilkudziesięciu do kilkuset tysięcy[6]. Zatem dużym zaskoczeniem był fakt, gdy wraz z zakończeniem projektu poznania ludzkiego genomu, liczba ta została zredukowana do ~31,000[2]. Obecnie zarówno RefSeq, jak i GENCODE zgadzają się co do liczby genów kodujących białka w ludzkim genomie na poziomie ~20,000. Natomiast liczba genów niekodujących mieści się w przedziale 15,000 – 18,000[7]. Wartości te dla myszy są dość odmienne w porównaniu z tymi dla człowieka (Rysunek 1). Nie jest jednak jasne czy kwestia ta ma podłoże biologiczne, czy raczej wynika z mniej dojrzałego stanu katalogów genów dla genomu myszy.
Nie do końca wiadomo także, w jaki sposób liczba genów będzie ewoluować w czasie. Geny kodujące białka są relatywnie łatwe do zidentyfikowania z dużą dokładnością, co może wskazywać, iż katalog genów kodujących białka jest niemalże kompletny. Niezależne badania obejmujące użycie metod głębokiego sekwencjonowania, zdają się potwierdzać ten pogląd[8,9]. Chociaż ostatnio byliśmy świadkami próby podważenia tego konsensusu poprzez doniesienie, które wskazywało na odkrycie setki nowych genów

kodujących białka w ludzkim genomie[10]. Wysiłki te jednak okazały się być dość kontrowersyjne, gdyż zidentyfikowane geny miały fałszywie pozytywny charakter[11].

Geny niekodujące białek są znacznie trudniejsze do zidentyfikowania. Dlatego też skład katalogów w kontekście tych genów jest znacznie mniej ostateczny. Wraz z rozwojem metod ukierunkowanych na identyfikację genów niekodujących, oczekuje się, że to właśnie ich liczba będzie znacząco wzrastać w najbliższych latach.



Rysunek 1. Statystyka dla podstawowej wersji katalogu GENCODE (v34) dla genomów człowieka i myszy. *Liczba genów (lewa strona) i transkryptów (prawa strona) dla długich niekodujących RNA (lncRNA, po lewej) oraz genów kodujących białka (po prawej). Pozostałe klasy genów zostały pominięte dla uproszczenia. Biotypy odpowiadają poszczególnym klasom funkcjonalnym genów.*

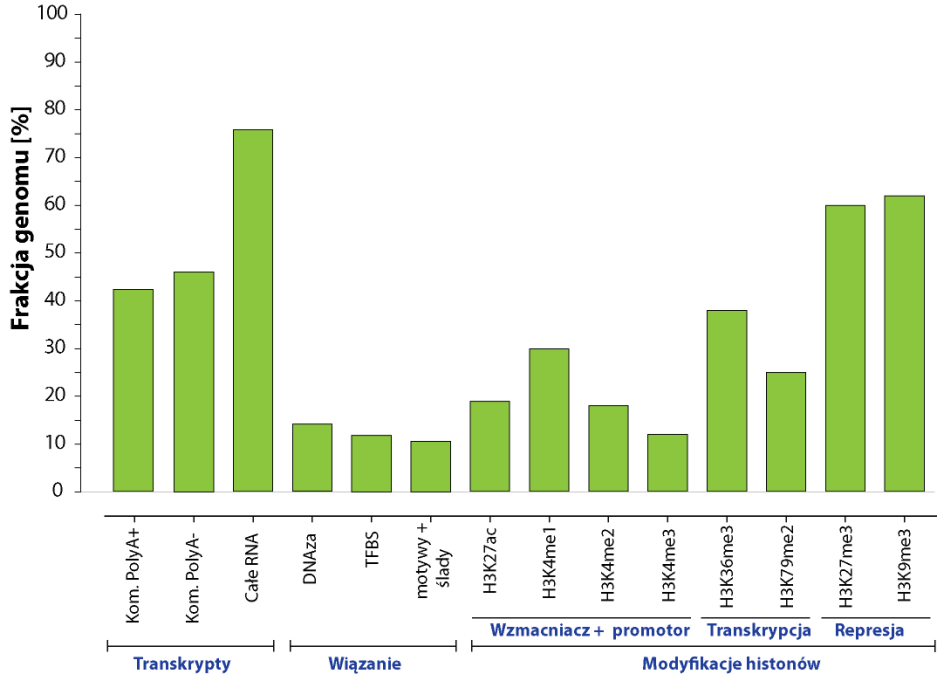
1.2. Jaka część genomów ssaków jest funkcjonalna?

Dzięki postępowi technologii w dziedzinie sekwencjonowania RNA, który pozwolił na dokładną analizę transkryptomów ssaków, zaczęliśmy uzyskiwać bardziej wyrafinowany obraz natury, tożsamości i zakresu elementów funkcjonalnych w genomach człowieka i myszy.

Wielkoskalowe projekty takie jak ENCODE (Encyclopedia of DNA Elements), które przeprowadziły ultra-głębokie analizy przy użyciu sekwencjonowania RNA i ludzkich linii komórkowych, ujawniły, że ~75% ludzkiego genomu ulega transkrypcji, a aż ~50% regionów genomowych wytwarza dojrzałe, poliadenylowane transkrypty[12] (Rysunek 2). Choć wskazane wartości są dość wysokie, powszechnie uważa się, że mogą być one niedoszacowane, ponieważ w projekcie ENCODE użyty został tylko niewielki podzbiór stanów komórkowych. Ponadto dla 11% regionów genomowych, z wyłączeniem 1,1% sekwencji kodujących białka, ENCODE wykazał obecność motywów wiązania czynników transkrypcyjnych oraz śladów DNAzy o wysokiej rozdzielczości (ang. *high-resolution DNase footprints*, DHS) w jednym lub wielu typach komórek (Rysunek 2). Wskazuje to na bezpośredni kontakt tych regionów z elementami regulatorowymi. Wreszcie, modyfikacje histonów związane z regionami promotorowymi lub

wzmacniaczami transkrypcji (ang. *enhancers*) oraz modyfikacje charakterystyczne dla elongacji transkrypcji stanowią odpowiednio ~20% i ~30% ludzkiego genomu (Rysunek 2).

Projekt ENCODE wykazał, że wiele regionów w tym sekwencje niekodujące, regiony niezachowane ewolucyjnie i powtarzające się elementy (ang. *repeat elements*), które wcześniej uważano za niefunkcjonalne, ulegają transkrypcji. Jednakże sama transkrypcja nie wskazuje na jakąkolwiek rolę tych regionów w komórce, a biologiczne znaczenie wszechobecnej transkrypcji jest raczej kontrowersyjne[13]. Wynika to głównie z faktu, że transkrypcja, w szczególności na etapie inicjacji przez polimerazę RNA II (Pol II) jest niespecyficzna i stochastyczna[14]. Dlatego też znaczna część transkryptów międzygenowych może po prostu stanowić produkt uboczny niespecyficznej transkrypcji i zostać przeznaczona do szybkiej degradacji w komórce. Chociaż nie jest jasne, jaki odsetek regionów w genomie pełni biologiczne funkcje w komórce, ostatnie badania sugerują, że 20-25% stanowi górną granicę funkcjonalnej frakcji ludzkiego genomu[15]. Zatem nawet te konserwatywne wyliczenia, dokonane na podstawie analizy zachowawczości ewolucyjnej oraz obciążenia mutacyjnego[16], wskazują, że znaczna część niekodującego genomu najprawdopodobniej jest funkcjonalna.



Rysunek 2. Podsumowanie analizy ludzkiego genomu według danych ENCODE. *Wykres słupkowy przedstawia frakcję ludzkiego genomu dla której wykryto aktywność w ramach konkretnego typu danych ENCODE w co najmniej jednej linii komórkowej lub tkance. Wszystkie wartości procentowe obliczone zostały w odniesieniu do całego genomu.*

1.3. Niekodujące RNA

Choć niekodujące DNA charakteryzuje się brakiem potencjału (lub minimalnym potencjałem) do kodowania białek, wciąż może ono ulegać transkrypcji i produkować niekodujące cząsteczki RNA

o nieznanych dotąd funkcjach. Wiele przykładów niekodujących, ale funkcjonalnych elementów genomowych zostało dobrze poznanych, w tym małe jądrowe RNA (snRNA), małe nuklearne RNA (snoRNA), rybosomalne RNA (rRNA), mikroRNA, transportujące RNA (tRNA) oraz długie niekodujące RNA (lncRNA). Chociaż poczyniono ogromne starania, aby zidentyfikować i dokładnie scharakteryzować niekodujące RNA, jak dotąd opisane przypadki stanowią jedynie niewielką część ludzkiego genomu[5].

Inne funkcjonalnie istotne regiony obejmują *cis*-regulatorowe elementy, kontrolujące ekspresję genów (np. promotory, wzmacniacze, izolatory [ang. *insulators*] oraz wyciszacze transkrypcji), punkty wyjściowe replikacji, telomery i centromery. Co najmniej 8.5% ludzkiego genomu zaangażowane jest w regulację genów w orientacji *cis*, jak wynika z konserwatywnej analizy danych ENCODE typu ChIP-seq oraz śladów DNAzy o wysokiej rozdzielczości w panelu ludzkich linii komórkowych[8].

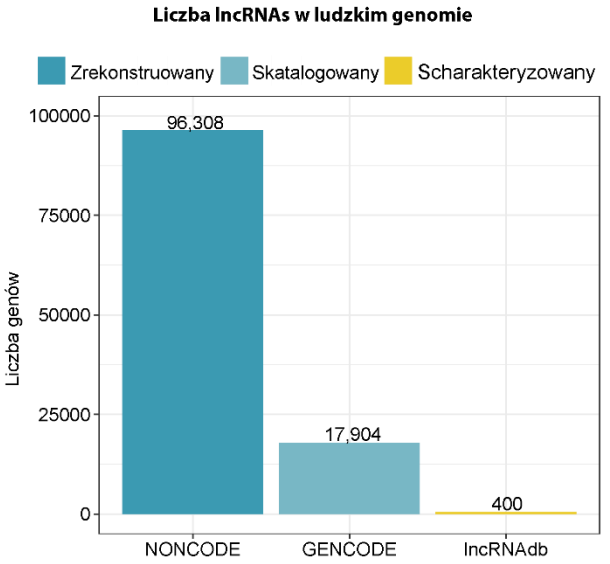
Wzmacniacze transkrypcji są jedną z najbardziej interesujących klas spośród sekwencji regulatorowych. Elementy te działają w układzie *cis* i aktywują transkrypcję genów poprzez szeroki zakres interakcji z promotorami docelowych genów. Wzmacniacze transkrypcji mogą znajdować się zarówno powyżej, jak i poniżej docelowego promotora. W wielu przypadkach odległość ta wynosi od setek do tysięcy nukleotydów[17]. Co więcej, wzmacniacze mogą także same ulegać transkrypcji, wytwarzając tak zwane wzmacniające RNA (ang. *enhancer RNA, eRNA*), które przeważnie nie ulegają procesowi składania transkryptów oraz poliadenylacji[18,19]. Ich cechą charakterystyczną jest także niski poziom ekspresji. Chociaż eRNA pierwotnie uważano za produkty uboczne procesu aktywacji wzmacniaczy transkrypcji, ostatnie doniesienia pokazują, że odgrywają one kluczową rolę w uruchamianiu kaskady molekularnej, prowadzącej do aktywacji wzmacniacza[20].

1.4. Długie niekodujące RNA - „znany nieznany” w biologii i chorobach

Długie niekodujące RNA (ang. *long noncoding RNAs, lncRNAs*) to cząsteczki RNA o długości powyżej 200 nukleotydów, które charakteryzują się ograniczonym lub nawet znikomym potencjałem kodowania białka[21,22]. Transkrypty te przypominają mRNA na kilka sposobów – podlegają one procesowi składania transkryptów (istnieją wyjątki), są poliadenylowane (istnieją wyjątki) i posiadają kap[23]. W przeciwieństwie do mRNA, lncRNA ulegają ekspresji na niewielkim poziomie, charakteryzują się także niższą stabilnością oraz słabszą wydajnością obróbki potranskrypcyjnej[23–25]. Co więcej, lncRNA mają także wysoką specyficzność tkankową. Szacuje się, że liczba lncRNA w ludzkim genomie mieści się w zakresie od 20,000 do 100,000[5,26,27], co znacznie przekracza liczbę genów kodujących białka (~20,000)[3] lub małych niekodujących RNA (7,576 – GENCODE v31)[5]. Dzięki ogólnoświatowym wysiłkom skoncentrowanym na identyfikacji lncRNA, ich detekcja następuje znacznie szybciej, niż ich funkcjonalna charakterystyka. Do tej pory eksperymentalnie określono funkcję dla jedynie niewielkiej liczby (<1%) długich niekodujących RNA[28,29] (Rysunek 3).

1.5. Czy lncRNA są funkcjonalne?

W świecie naukowym od kilku lat ma miejsce ożywiona debata na temat tego, czy lncRNA mają jakąkolwiek funkcję biologiczną[30–32]. W przeciwieństwie do genów kodujących białko (zawierających otwarte ramki odczytu), dla lncRNA związek między sekwencją, a funkcją jest nieznany. Co więcej, lncRNA są słabo zachowane ewolucyjnie pomiędzy gatunkami[33,34]. Niemniej jednak sekwencja lncRNA jest bardziej zachowana ewolucyjnie, niż sekwencje przypadkowe. Rośnie także liczba lncRNA powiązanych z fenotypami komórkowymi[35–37]. Ponadto natura zaprojektowała niektóre lncRNA tak, aby odgrywały istotną rolę w zachowaniu równowagi genetycznej pomiędzy żeńskimi i męskimi organizmami, co sugeruje ich kluczowy charakter w procesie regulacji genów[38–41]. Ostatnie doniesienia dostarczają wielu przykładów wskazujących na regulacyjną rolę kilku lncRNA w procesie różnicowania komórek[42] i organogenezie[43]. Istnieje również coraz więcej dowodów na to, że lncRNA odgrywają także istotną rolę w stanach patologicznych. Wiele lncRNA ma związek z chorobami ludzkimi[29], w tym chorobami sercowo-naczyniowymi[44] i zaburzeniami neurodegeneracyjnymi[45]. Zmiany w profilach ekspresji lncRNA są charakterystyczne także dla różnych typów nowotworów[46]. Co więcej, lncRNA mogą mieć zarówno funkcje onkogenne, jak i supresorowe, które czasem mogą być zintegrowane ze znanymi regulatorami, takimi jak p53[47]. Rosnąca liczba lncRNA jest także łączona ze specyficznymi mechanizmami molekularnymi mających miejsce w procesach rakotwórczych[48]. Wreszcie, istnieją dowody na częste występowanie mutacji w lncRNA w chorobach nowotworowych[49].



Rysunek 3. Szacunkowa liczba genów lncRNA w ludzkim genomie. *„Zrekonstruowany" odnosi się do generowanych obliczeniowo modeli genów opartych na danych RNA-Seq otrzymywanych metodą krótkich odczytów.*

Równoległe, bardzo istotne pytanie brzmi, w jaki sposób lncRNA działają w komórce? Istnieją co najmniej trzy możliwe i nie wykluczające się wzajemnie „modalności funkcjonalne". Po pierwsze, lncRNA mogą być biologicznie aktywne jako dojrzałe cząsteczki RNA, zarówno w orientacji *cis*, jak i *trans*. Po drugie, umiejscowienie lncRNA może pokrywać się z lokalizacją aktywnego DNA regulatorowego (np. LincRNA-P21)[50]. Po trzecie, lncRNA mogą regulować pobliskie geny poprzez sam akt transkrypcji[51]. W dwóch ostatnich przypadkach, cząsteczka lncRNA i jej sekwencja mogą mieć drugorzędne znaczenie. Chociaż proporcje tych klas są nieznane, badania funkcjonalne wskazują, że znaczna populacja lncRNA działa jako dojrzałe cząsteczki RNA[52].

Co ciekawe, istnieją dowody na to, że niektóre geny, początkowo uznane za lncRNA, w rzeczywistości ulegają translacji i produkują białka[53,54]. Zazwyczaj kodują one małe peptydy (<100 aminokwasów), które ze względu na swoją niewielką długość są trudne do wykrycia przy użyciu standardowego oprogramowania stosowanego do budowy katalogów. Zastosowanie wysokoprzepustowego sekwencjonowania fragmentów chronionych przez rybosomy (Ribo-seq) ujawniło nieoczekiwane interakcje pomiędzy lncRNA i rybosomami, co sugeruje, że lncRNA mogą ulegać translacji[55]. Chociaż asocjacja lncRNA z rybosomami niekoniecznie oznacza translację w celu produkcji funkcjonalnego produktu białkowego, opisano kilka przypadków lncRNA faktycznie wytwarzających polipeptydy[56,57]. Pomimo wysiłków podjętych w celu zrozumienia funkcji tej klasy lncRNA, ich rola w komórce wciąż pozostaje nieznana, szczególnie w świetle ostatniego odkrycia wskazującego na różnice w składzie trinukleotydów dla kanonicznych (mRNA) i niekanonicznych (lncRNA) ramek odczytu[58].

Środowisko naukowców badających lncRNA jest zgodne co do tego, że niektóre z tych cząsteczek są funkcjonalne, a inne nie, będąc jedynie produktem ubocznym maszynerii transkrypcyjnej[14]. Jednakże nadal nie osiągnięto konsensusu w kwestii rozmiaru frakcji lncRNA, która ma znaczenie biologiczne w komórce[59]. Najnowsze wielkoskalowe szacunki wskazują na ułamek 2-5% funkcjonalnych lncRNA dla danego fenotypu[60,61]. Biorąc pod uwagę dużą, szacunkwą liczbę lncRNA (Rysunek 3) w genomie człowieka, nawet najbardziej pesymistyczne założenie, w którym znakomita większość z nich – 95% nie ma funkcji biologicznej, pozostawia tysiące ludzkich loci lncRNA o potencjalnym znaczeniu biologicznym lub medycznym. Stąd też lepsze zrozumienie funkcji lncRNA ma kluczowe i bezpośrednie znaczenie z punktu widzenia badań biomedycznych.

1.6. Katalogi lncRNA

W przeciwieństwie do genów kodujących białko, które zostały zidentyfikowane i szczegółowo scharakteryzowane, procesy tworzenia katalogów lncRNA są znacznie mniej zaawansowane. Istnieją trzy główne powody, dla których identyfikacja i charakterystyka lncRNA jest trudna. Po pierwsze, lncRNA ulegają względnie niskiej ekspresji, co utrudnia ich analizę za pomocą standardowych eksperymentów transkryptomicznych, w tym sekwencjonowania RNA i analizy ekspresji genów zależnej od kap (ang. *cap analysis of gene expression*, CAGE). Po drugie, związek sekwencja-struktura-funkcja dla lncRNA jest słabo poznany, co oznacza, że nie można obecnie zastosować żadnych cech sekwencji, ani elementów funkcjonalnych do identyfikacji nowych lncRNA. Wreszcie, lncRNA są słabo zachowane podczas ewolucji, co utrudnia identyfikację ich ortologów lub paralogów na podstawie podobieństwa sekwencji. W konsekwencji detekcja i genomiczna charakterystyka lncRNA prawie całkowicie opiera się na fizycznych dowodach transkryptomicznych.

1.6.1. Katalogi lncRNA tworzone na podstawie rekonstrukcji transkryptomu

Sekwencjonowanie nowej generacji wraz z metodami obliczeniowymi znacznie przyspieszyło identyfikację lncRNA i pomogło znaleźć dziesiątki tysięcy loci kodujących lncRNA u ludzi i myszy[26]. Jednakże większość współczesnych katalogów lncRNA zbudowano na podstawie danych uzyskiwanych za pomocą sekwencjonowania o krótkich odczytach. Sekwencjonowanie drugiej generacji pozwala na głęboką analizę transkryptomów dzięki setkom milionów produkowanych odczytów i choć liczba ta jest imponująca, metody te jednak wciąż nie pozwalają dotrzeć do dolnych granic transkryptomu. Zatem transkrypty o słabej ekspresji, m.in. lncRNA, nie są dobrze reprezentowane podczas standardowych eksperymentów sekwencjonowania RNA metodą krótkich odczytów. Co więcej, takie odczyty są znacznie krótsze niż typowe lncRNA (czy też mRNA) i muszą być połączone razem w celu odtworzenia transkryptów i ich struktur. Chociaż sekwencjonowanie drugiej generacji przy wsparciu metod obliczeniowych, umożliwiających składanie transkryptów pozwoliło na identyfikację transkryptów zarówno dla nowych, jak i znanych genów, podejście to znacznie utrudnia detekcję izoform o pełnej długości. Stąd też większość transkryptów rekonstruowanych z krótkich odczytów posiada niekompletne 5'oraz 3' końce. Powodem tego zjawiska jest niejednolity rozkład zmapowanych odczytów w obrębie transkryptu[62]. Metody składania transkryptów względnie poprawnie identyfikują eksony, ale nie introny[63]. Zasadniczo jakość procesu rekonstrukcji transkryptów zależy od wysokości poziomu ekspresji danego genu oraz jednorodności pokrycia transkryptu mapowanymi, krótkimi odczytami[64], co rzadko kiedy jest spełnione dla lncRNA.

1.6.2.Katalogi lncRNA weryfikowane manualnie

Problem niekompletności transkryptów jest znacząco zredukowany w przypadku katalogów lncRNA weryfikowanych ręcznie[65]. W ramach ręcznego procesu katalogowania, geny kodujące lncRNA oraz modele ich transkryptów są budowane przez ludzi w oparciu o niezrekonstruowane dowody na poziomie genomu i transkryptomu według zdefiniowanych protokołów. Precyzyjna kontrola tworzonych modeli lncRNA skutkuje otrzymaniem katalogów o bardzo wysokiej jakości dla których zminimalizowane jest ryzyko wystąpienia artefaktów obecnych w katalogach budowanych automatycznie, np. pominięcia końcowych eksonów. Jednakże jedną z głównych słabości tego rodzaju katalogów jest czas ich tworzenia, który jest dużo dłuższy w porównaniu do metod katalogowania opartych na rekonstrukcji transkryptów. Ręczne metody budowania transkryptów wymagają także długoterminowego finansowania, co sprawia, że są znacznie droższe niż inne podejścia tworzenia katalogów. Najpopularniejszym katalogiem dla genomów człowieka i myszy, tworzonym ręcznie jest GENCODE[5,23], w którym modele transkryptów są dodatkowo weryfikowane przy użyciu różnych metod eksperymentalnych[63,66].

2. Cele badawcze

Moje badania koncentrują się na budowaniu wysokiej jakości map genowych dla długich niekodujących RNA w genomach człowieka i myszy, jako podstawy do zrozumienia ich fizjologicznych i patologicznych ról w komórce. Główne cele moich badań obejmują:

a. Przyspieszenie procesu katalogowania lncRNA poprzez opracowanie ukierunkowanych, wysokoprzepustowych metod opartych na technologii sekwencjonowania trzeciej generacji, tj. za pomocą długich odczytów. Połączenie przepustowości i dokładności (poprzez ograniczenie do minimum konieczności rekonstrukcji transkryptów) w celu uzyskania katalogu lncRNA o wysokiej jakości przy jednoczesnej redukcji ludzkiej interwencji w ten proces.
b. Poprawa jakości opisanych już modeli transkryptów w katalogu referencyjnym GENCODE oraz identyfikacja brakujących transkryptów w celu utworzenia pełnego katalogu genów lncRNA dla genomów myszy i człowieka.
c. Weryfikacja genomowych cech lncRNA z użyciem ulepszonych katalogów, w tym analiza regionów promotorowych, struktur genów i transkryptów oraz analiza potencjału kodującego białko.
d. Ocena i systematyczna analiza porównawcza istniejących obecnie katalogów lncRNA, w tym ulepszonego katalogu GENCODE, który jest efektem projektu CLS (Capture Long Sequencing).

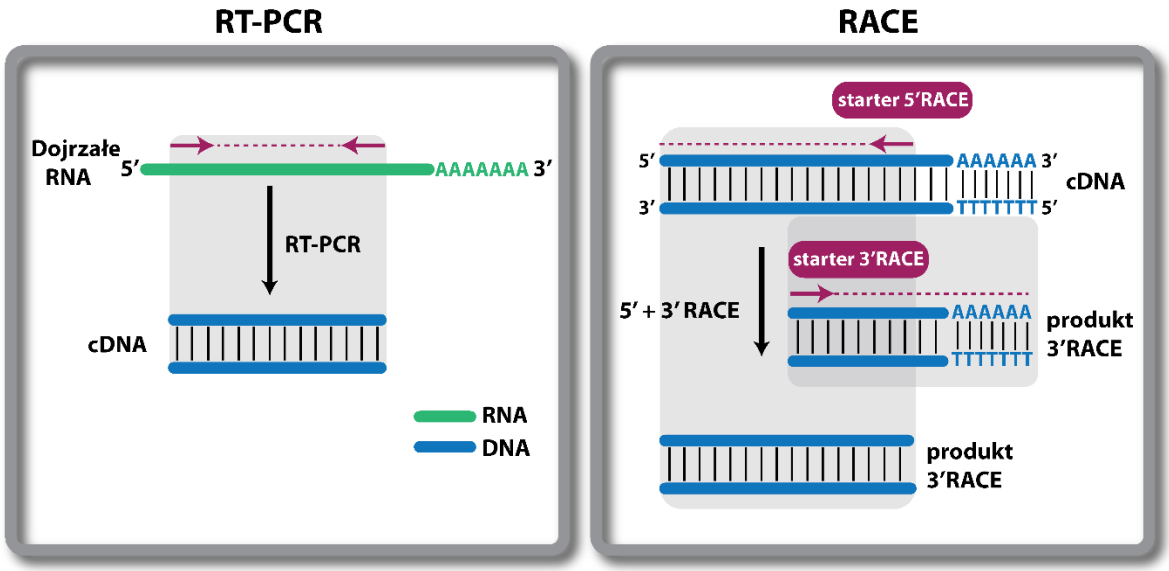
3. Analiza ukrytego transkryptomu za pomocą ukierunkowanych metod sekwencjonowania RNA

Moje badania w znacznym stopniu przyczyniły się do powstania dwóch bardzo czułych metod identyfikacji i katalogowania lncRNA (1) RACE-seq (publikacja H1[67]) oraz (2) CLS - RNA Capture Long Sequencing (publikacja H2[63]). Obie metody polegają na połączeniu procesu wzbogacania wybranych regionów z sekwencjonowaniem metodą odczytów o średniej długości (RACE-seq) oraz długich odczytów (CLS). Wybrane metody sekwencjonowania, tj. Roche GS-FLX 454+ dla RACE-seq oraz PacBio RS II dla CLS były wówczas najnowocześniejszymi technologiami, oferującymi możliwie najdłuższe odczyty dla sekwencjonowania RNA. Projekt RACE-seq stanowi pierwszą próbę zastosowania celowanego sekwencjonowania RNA w procesie katalogowania lncRNA i obejmuje analizę jedynie próbek ludzkich. Po udanej implementacji tej metody do identyfikacji lncRNA, zdecydowaliśmy się na analizę lncRNA w genomie człowieka oraz myszy w ramach projektu CLS.

3.1. RACE-seq (publikacja H1)

Reakcja łańcuchowa polimerazy z odwrotną transkrypcją (ang. *reverse-transcription polymerase chain reaction, RT-PCR*) oraz szybka amplifikacja końców cDNA (ang. *rapid amplification of cDNA ends, RACE)* to dwie dobrze znane metody biologii molekularnej oparte na amplifikacji transkryptów za pomocą

PCR przy użyciu starterów oligonukleotydowych o specyficznej sekwencji (Rysunek 4). Metoda RT-PCR polega na zastosowaniu dwóch zbieżnych starterów, specyficznych dla danego transkryptu. Powstałe produkty RT-PCR są sekwencjami, które znajdują się pomiędzy dwoma oskrzydlającymi starterami. Natomiast RACE polega na zastosowaniu kombinacji starterów uniwersalnych oraz specyficznych dla danego transkryptu. Podczas, gdy starter specyficzny dla danego transkryptu jest komplementarny do wybranego miejsca w obrębie jego sekwencji, uniwersalny starter znajduje się na przeciwległym końcu transkryptu. RACE jest metodą, która pozwala na dokładną analizę końców transkryptów, ponieważ może być wykonywana zarówno w kierunku 5', jak i 3'. Ponadto zastosowanie dodatkowego, zagnieżdżającego startera pozwala zwiększyć specyficzność reakcji RACE[68].



Rysunek 4.Schematyczne przedstawienie technik RT-PCR i RACE. *Regiony amplifikowane przez każdy starter (lub parę starterów - fioletowe strzałki) zostały zaznaczone szarym kolorem.*

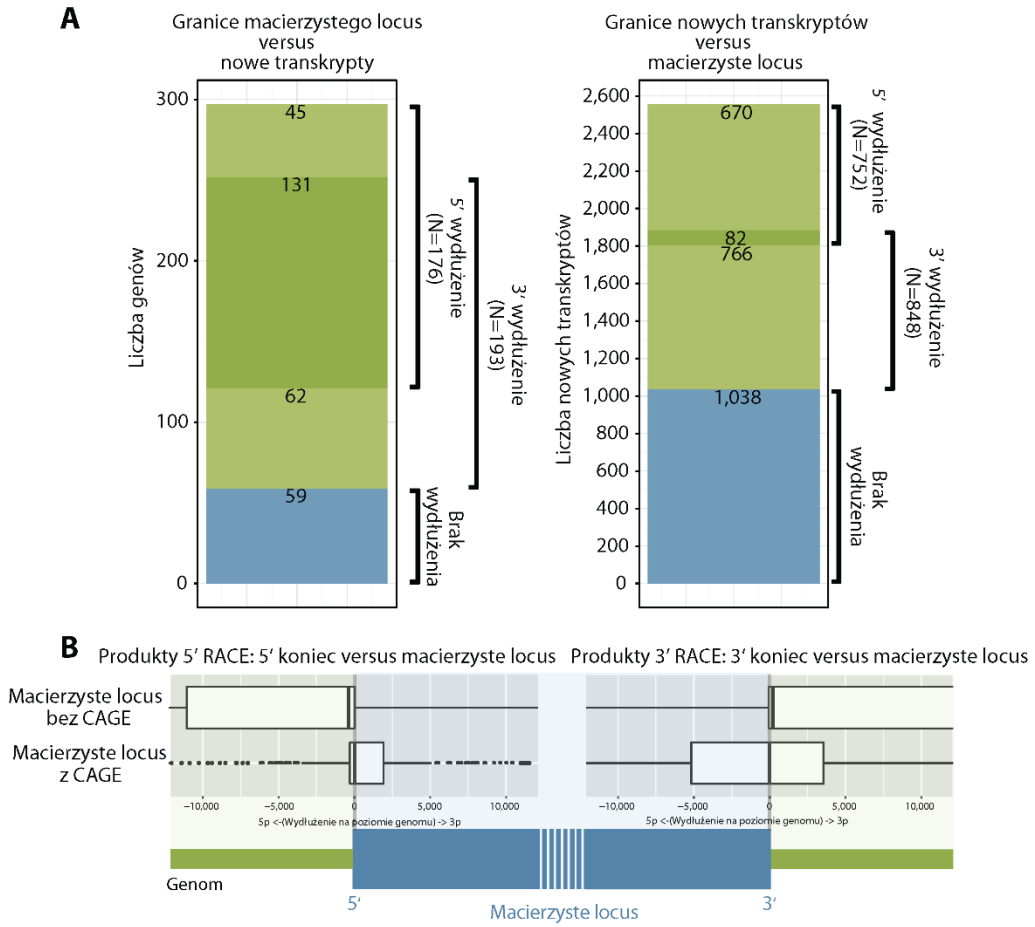
Pomysł na połączenie RACE z wysokoprzepustowym sekwencjonowaniem został wprowadzony przez Olivariousa i wsp.[69]. Choć autorzy określali to połączenie głębokim-RACE, metoda ta w praktyce została zastosowana jedynie dla kilku genów kodujących białka, a uzyskane produkty RACE zostały zsekwencjonowane za pomocą technologii krótkich odczytów.

Nasza praca po raz pierwszy połączyła metodę RACE z sekwencjonowaniem o odczytach średniej długości oraz nadała jej wysokoprzepustowy charakter, dzięki analizie 398 ludzkich lncRNA z katalogu GENCODE w obu kierunkach, tj. 5' oraz 3' z użyciem siedmiu wybranych, ludzkich tkanek. Mieszanina cDNA z reakcji RACE dla każdej z tkanek poddana została sekwencjonowaniu z użyciem platformy 454 (średnia długość odczytu ~600 nukleotydów). Pierwsza partia sekwencji RACE uzyskana została przy użyciu standardowych, niezagnieżdżonych starterów. Następnie używając produktów standardowej wersji RACE, przeprowadziliśmy zagnieżdżoną reakcję RACE, aby poprawić specyficzność oraz czułość detekcji

wzbogaconych sekwencji. Podejście to także miało innowacyjny charakter, ponieważ poprzez zastosowanie zagnieżdżonych starterów, po raz pierwszy byliśmy w stanie ocenić wydajność oraz jakość zagnieżdżonych reakcji RACE. Zastosowanie zagnieżdżonych starterów skutkowało 9,5-krotnym wzrostem detekcji wybranych sekwencji, w porównaniu ze standardową reakcją RACE.

Metoda RACE-seq pozwoliła na detekcję 2,556 nowych transkryptów w obrębie wzbogacanych regionów. Ogółem wydłużenie modeli transkryptów odnotowano dla odpowiednio 752 (670 + 82) i 848 (766 + 82) produktów reakcji RACE w kierunku 5’ oraz 3’ (Rysunek 5A). Osiemdziesiąt dwa nowe transkrypty wydłużyły swoje macierzyste locus zarówno w kierunku 5’ jak i 3’. Z 398 analizowanych genów lncRNA, 238 (60%) wydłużyło swoje granice w kierunku 5’, podczas gdy 131 (30%) genów zostało wydłużonych w obu kierunkach (Rysunek 5A).

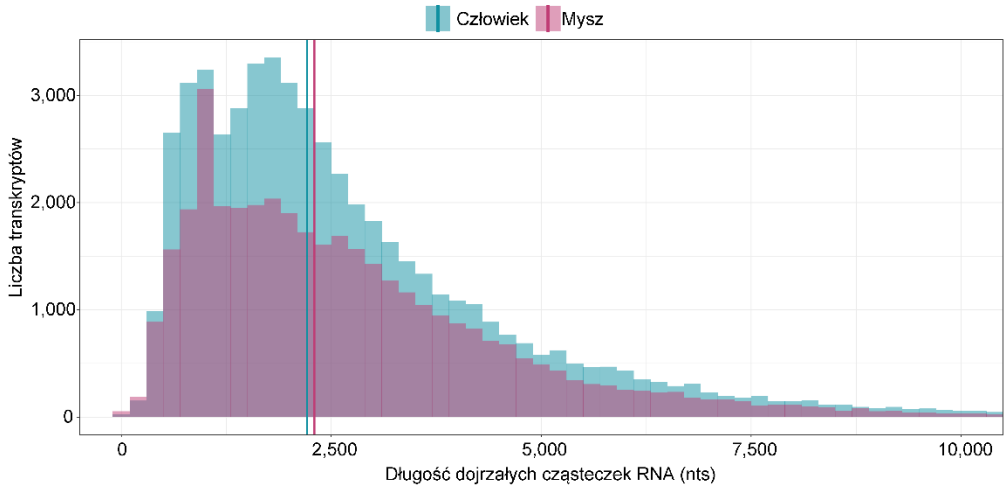
Następnie podzieliliśmy zestaw docelowych lncRNA na kategorie z końcami 5’ zwalidowanymi (N = 180) i niezwalidowanymi (N = 218) przez CAGE (zgodnie z projektem FANTOM[70]) (Rysunek 5B).



Rysunek 5. Wydłużanie locus i nowe granice transkryptu.*(A) Diagramy Venna przedstawiające proporcje loci (lewy panel) i transkryptów (prawy panel) wydłużonych w kierunku 5’ i/lub 3’. (B) Nowe granice dla loci zweryfikowanych (dolny wykres) i niezweryfikowanych przez CAGE (górny wykres). Schematyczne przedstawienie docelowego (wzbogacanego) locus znajduje się poniżej wykresów. Zakres skali wykresów pudełkowych został zmniejszony (-10 000, 10 000 nukleotydów) dla celów przejrzystości.*

Nasze wyniki sugerują, że dla genów zwalidowanych przez CAGE prawdopodobieństwo wydłużenia sekwencji za pomocą RACE jest mniejsze, co potwierdza niezawodność CAGE jako złotego standardu w przewidywaniu miejsc startu transkrypcji (ang. *transcription start sites,* TSS). Co ciekawe, szanse weryfikacji przez CAGE nowych TSS, odkrytych w loci z 5' końcami potwierdzonymi przez CAGE, były znacznie większe, niż w przypadku loci niezweryfikowanych (74% w porównaniu z 56%). Ogólnie rzecz biorąc, podejście RACE-seq pozwoliło na odkrycie wielu niezidentyfikowanych wcześniej elementów, w tym 615 nowych TSS z czego 252 (41%) zweryfikowanych zostało przez dane CAGE. Co więcej, mediana długości dojrzałych transkryptów po RACE-seq wzrosła (nieznacznie) z 623 do 604 nukleotydów.

Podsumowując, wyniki uzyskane za pomocą RACE-seq prowadzą do dwóch głównych wniosków: (1) pomimo znacznych wysiłków czynionych w celu ich dokładnej identyfikacji i analizy, modele zebrane w katalogu GENCODE v7 są niekompletne i nie w pełni odzwierciedlają lncRNA kodowane przez genom człowieka oraz (2) RACE-seq pozwala usprawnić detekcję końców transkryptów. Jednakże pomimo swojej przydatności RACE-seq ma pewne poważne ograniczenia. Jednym z głównych problemów jest identyfikacja transkryptów o pełnej długości za pomocą RACE-seq. Średnia długość dojrzałego transkryptu w genomie człowieka wynosi 2,500 nukleotydów (Rysunek 6), podczas gdy średnia długość odczytu uzyskiwana z użyciem platformy 454 to 600 nukleotydów. W związku z tym zarówno wartości długości, jak i kompletności transkryptów w przypadku RACE-seq mogą być niedoszacowane. Ograniczenie to może być łatwo przezwyciężone poprzez zastosowanie w RACE-seq jednej z aktualnie dostępnych metod sekwencjonowania o długich odczytach, takich jak PacBio lub Oxford Nanopore Technologies (ONT).



Rysunek 6. Rozkład długości transkryptów w genomach człowieka i myszy. *Histogram długości transkryptu w genomie człowieka (zielononiebieski) i myszy (różowy). Oś X: długość dojrzałego RNA (z wyłączeniem intronów) w nukleotydach (nts). Odpowiednie średnie długości (człowiek: 2,213 nts, mysz: 2,298 nts) są przedstawione w postaci pionowych linii dla każdego gatunku przy użyciu tego samego schematu kolorów. Zakres skali osi X został zredukowany do 10,000 nts dla celów przejrzystości.*

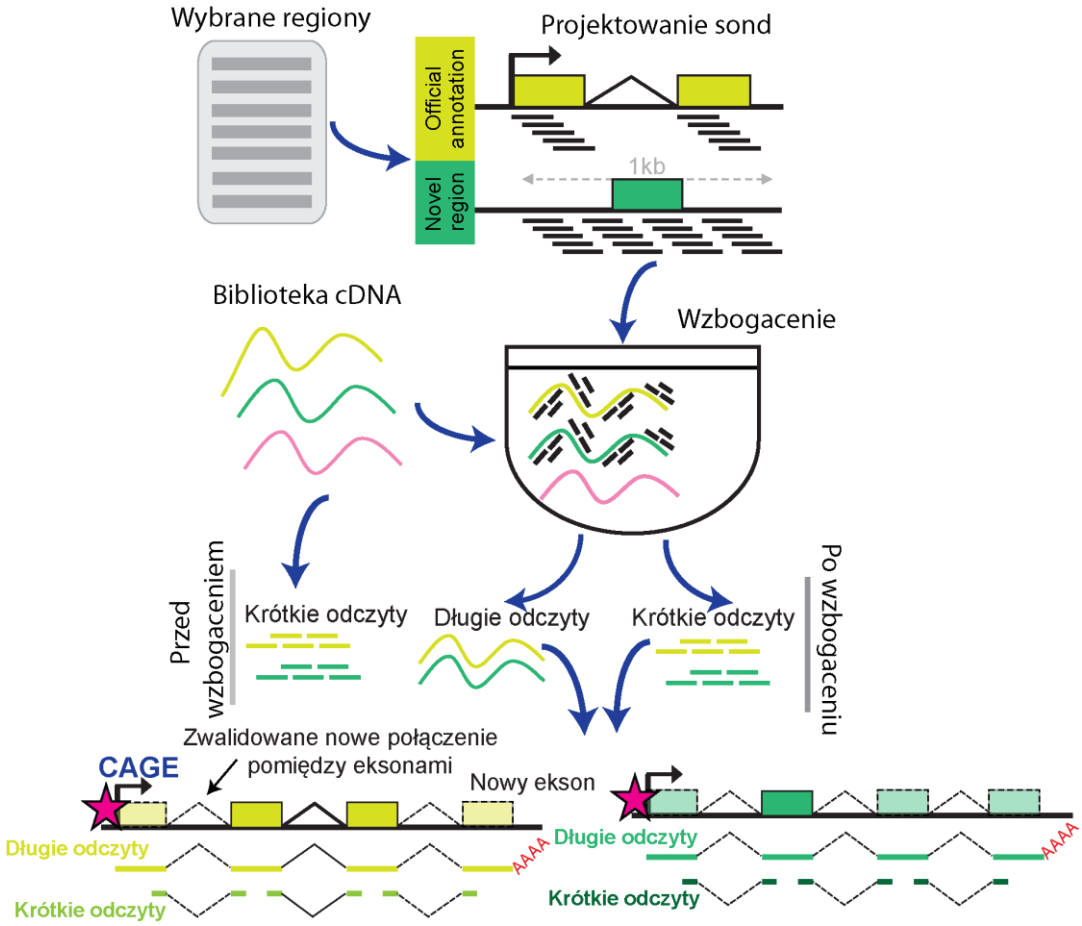
Każda z tych metod generuje znacznie dłuższe odczyty (do 50,000 i 200,000 nukleotydów odpowiednio dla PacBio i ONT), niż metoda 454. Kolejnym ograniczeniem reakcji RACE jest problem wykrywania sekwencji docelowej. Geny wybrane w ramach projektu RACE-seq ulegają relatywnie wysokiej ekspresji z medianą RPKM (ang. *Reads Per Kilobase of exon per Million mapped reads*) na poziomie 8,3. Podczas gdy Derrien i wsp. wykazali, iż ekspresja dla całej populacji lncRNA znajduje się na poziomie <1 RPKM[23]. W związku z tym pojawia się spore ryzyko, iż reakcja RACE nie będzie w stanie skutecznie wzbogacić transkryptów o bardzo niskiej ekspresji. Ponadto RACE obejmuje szereg pracochłonnych kroków, co w znacznym stopniu utrudnia zastosowanie tej metody do analizy większego zestawu transkryptów. Ze względu na te ograniczenia, wybraliśmy zastosowanie metody CLS w naszych kolejnych eksperymentach.

3.2. CLS – wysokoprzepustowa metoda do identyfikacji lncRNA o pełnej długości (H2)

Podejście CLS zostało zaprojektowane w celu przezwyciężenia wspomnianych ograniczeń RACE-seq. Jest to również pierwsze podejście łączące z powodzeniem proces wzbogacania wybranych sekwencji RNA z sekwencjonowaniem metodą długich odczytów – PacBio (Rysunek 7). CLS zaprojektowane zostało w celu ulepszenia istniejących modeli genów i identyfikacji nowych loci w regionach, które potencjalnie mogą zawierać lncRNA. W ramach CLS wybraliśmy niekodujące regiony w ludzkim (9,060) oraz mysim (6,615) genomie, które łącznie stanowią odpowiednio 15,5 oraz 8.3 megazasad. Oligonukleotydowe sondy do wzbogacenia zostały zaprojektowanie z uwzględnieniem zestawu ludzkich (5,953) oraz mysich (1,920) lincRNA (ang. *long intergenic noncodig RNA* – długie międzygenowe niekodujące RNA, które nie pokrywają się z żadnym znanym genem kodującym białko) zlokalizowanych w odległości co najmniej 5,000 nukleotydów od najbliższego genu kodującego białko. Liczby te stanowią odpowiednio 41% oraz 36% wszystkich lincRNA w katalogu GENCODE dla człowieka i myszy. Pozostałe sondy były komplementarne do niekodujących regionów genomowych obejmujących m.in. ultra zachowane elementy (ang. *ultra conserved elements*, UCE), wzmacniacze transkrypcji oraz genów przewidywanych *de novo*, gdzie przewidywaliśmy znaleźć nowe lncRNA. Na etapie projektowania sond uwzględniliśmy także szereg elementów kontrolnych w tym syntetyczne kontrole RNA (ERCC spike-ins), dopasowane poziomem ekspresji do analizowanych lncRNA[71].

Aby zapewnić maksymalną różnorodność lncRNA, wybraliśmy panel dopasowanych, złożonych transkrypcyjnie i istotnych biomedycznie narządów pochodzących od ludzi i myszy, a także dwie szeroko badane ludzkie linie komórkowe (HeLa i K562) oraz dwa mysie embrionalne punkty czasowe (embrionalny dzień 7 [E7] i 15 [E15]). Biblioteki do sekwencjonowania zostały wzbogacone w wybrane regiony za pomocą sond oligonukleotydowych z puli cDNA. Mając świadomość, że platforma PacBio, podobnie jak inne techniki eksperymentalne, faworyzuje krótsze fragmenty, podzieliliśmy wzbogacone biblioteki na trzy

frakcje względem długości fragmentów (1–1,5 kb, 1,5–2,5 kb i> 2,5 kb). Następnie każda z tych frakcji poddana została sekwencjonowaniu za pomocą platformy Pacific Biosciences RS II, uzyskując w sumie ~1 mln odczytów ROI (ang. *reads of insert*) dla każdego gatunku. Zastosowanie sond oligonukleotydowych pozwoliło na znaczne wzbogacenie wybranych sekwencji, tj.19- oraz 11-krotne odpowiednio dla człowieka i myszy w porównaniu ze standardowym, niewzbogaconym sekwencjonowaniem.



Rysunek 7. Metoda CLS. *Schemat pokazujący proces automatycznego budowania modeli transkryptów o wysokiej jakości. Podejście CLS może być stosowane zarówno do ulepszenia istniejących modeli (żółto-zielony), jak i do detekcji nowych (zielony). Sondy oligonukleotydowe są zaprojektowane tak, aby pokrywać w sposób dachówkowy (ang. tiled) wybrane regiony w genomie. Kolejno następuje wzbogacenie bibliotek z użyciem zaprojektowanych sond, a z wzbogaconych próbek przygotowuje się biblioteki do sekwencjonowania metodą krótkich i długich odczytów. Dane uzyskiwane z użyciem metod sekwencjonowania o długich odczytach są wykorzystane do zbudowania modeli transkryptów, natomiast dane uzyskane za pomocą krótkich odczytów służą do walidacji utworzonych modeli transkryptów poprzez weryfikację połączeń pomiędzy eksonami. Kompletność każdego transkryptu jest oceniana na podstawie bliskości klastrów CAGE (różowy) i miejsc terminacji transkrypcji za pomocą sekwencji polyA w długich odczytach, które nie są kodowane przez genom (czerwony). Prostokąty z jaśniejszym cieniowaniem i przerywanymi konturami oznaczają nowe eksony.*

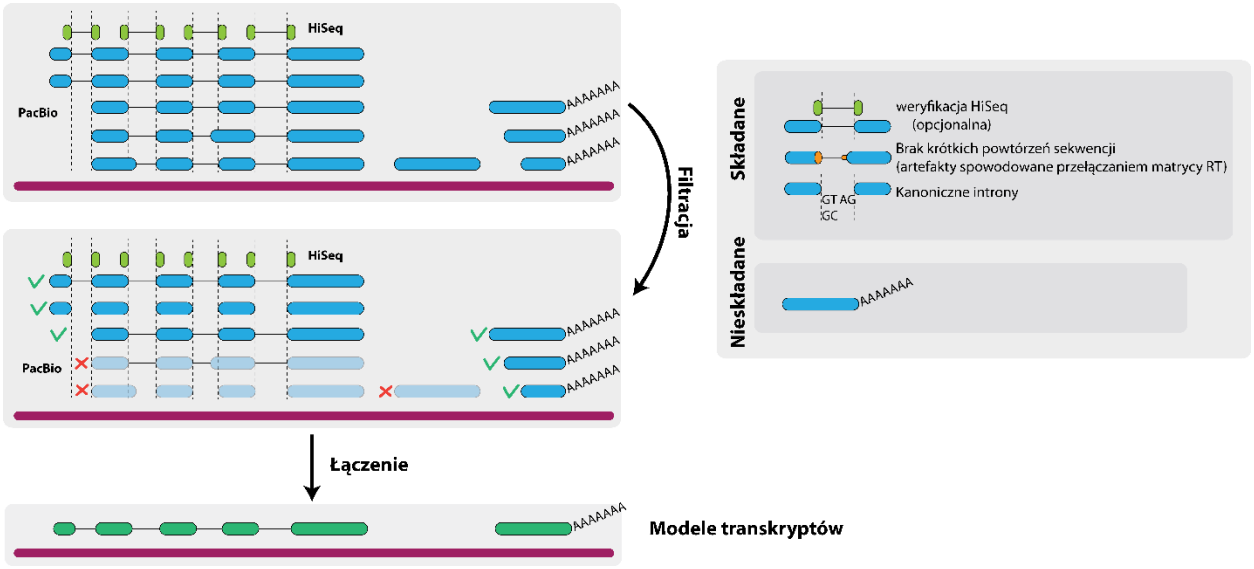
Dzięki zastosowaniu CLS, zidentyfikowaliśmy 179,993/129,556 (ludzkich/mysich) modeli unikalnych transkryptów dla wszystkich biologicznych typów (biotypów) transkryptów. Z 65,736/44,673 modeli o pełnej długości uzyskano 11,429/4,350 struktur dla genów lncRNA z których 8,494/3,168 (74% / 73%)

były nowe. Co więcej, CLS pozwoliło na detekcję nowych transkryptów w przestrzeni międzygenowej, w tym 18,751/20,469 transkryptów w regionach poza eksonami, 637/364 w obrębie wzmacniaczy transkrypcji oraz 433 nowych transkryptów w genomie myszy, które pochodziły z predykcji *de novo*. Co ciekawe, duża liczba (8,616/3,075) wykrytych transkryptów łączyła ze sobą geny różnych biotypów, w tym także geny kodujące białka. Wreszcie, CLS pozwoliło na poprawę jakości istniejących modeli lincRNA oraz na detekcję z dużą dozą dokładności nowych lincRNA w genomach człowieka i myszy. Liczba ludzkich lncRNA dla których zidentyfikowano nowe, zweryfikowane miejsca startu (TSS) oraz miejsca terminacji transkrypcji (ang. *transcription termination sites*, TTS) wzrosła odpowiednio z 1,650 do 2,607 (z 530 do 703 u myszy) oraz z 4,451 do 9,241 (z 1,036 do 1,616 u myszy).

Aby zrozumieć, w jakim stopniu możliwe jest zminimalizowanie ludzkiej interwencji w procesie katalogowania genów, po publikacji artykułu, zespół GENCODE odpowiedzialny za ręczną weryfikację katalogu (Jose Manuel González, Jonathan Mudge i Adam Frankish) ocenił jakość losowo wybranego zestawu 240 modeli transkryptów uzyskanych za pomocą CLS o różnych poziomach ufności (introny zweryfikowane za pomocą krótkich odczytów, kanoniczne introny, kompletne 5’ oraz 3’ końce). W rezultacie > 96% modeli transkryptów zweryfikowanych przez co najmniej jeden z wymienionych wyżej parametrów zostało pomyślnie zatwierdzonych, podczas gdy dla niezweryfikowanych modeli transkryptów poziom walidacji wynosił 62%. Pomimo małej wielkości badanego zestawu, otrzymane wyniki wskazują, że metoda CLS wraz z proponowanym scenariuszem analizy danych pozwala na otrzymanie wysokiej jakości modeli transkryptów i genów w dużej skali. Wynikiem projektu CLS jest testowy zestaw GENCODE+, który powstał poprzez automatyczne połączenie zestawów lncRNA z GENCODE (v20) i CLS odpowiednio dla człowieka i myszy.

3.3. Narzędzie do identyfikacji transkryptów w dużej skali (publikacja H2)

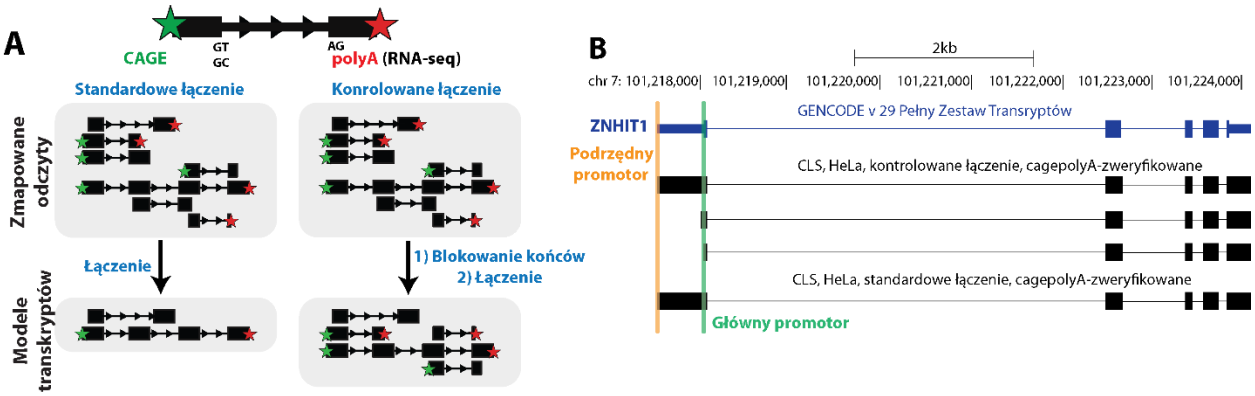
Aby ograniczyć udział człowieka w procesie katalogowania, opracowaliśmy w pełni zautomatyzowane, wysokoprzepustowe narzędzie do tworzenia katalogów genów (Rysunek 8). To bioinformatyczne podejście obejmuje wszystkie etapy analizy danych, w tym mapowanie odczytów, budowanie modeli transkryptów, ocenę jakości oraz etap filtracji, który pozwala utworzyć katalog o jakości zbliżonej do tych weryfikowanych ręcznie (Rysunek 8). Następnie, kompletność 5’ i 3’ końców transkryptów jest oceniana odpowiednio za pomocą danych FANTOM CAGE[70] oraz DHS[72], jak również przy pomocy detekcji TTS w ramach bezpośredniej identyfikacji ogonów polyA w sekwencji odczytów PacBio oraz poprzez ocenę bliskości położenia sygnałów poliadenylacji[73]. Wreszcie, każdy transkrypt jest klasyfikowany zgodnie z jego statusem (znany, nowy), biotypem (kodujący, niekodujący) oraz poziomem kompletności (kompletny, kompletny 5’ albo 3’ koniec, niekompletny).



Rysunek 8. Przetwarzanie odczytów według protokołu CLS. *Odczyty z wzbogaconych bibliotek cDNA, w tym ROI – długie odczyty PacBio (niebieski) i krótkie Illumina HiSeq (żółto-zielony) są mapowane do genomu. Każdy odczyt PacBio przechodzi następnie szereg etapów filtracji w celu zminimalizowania artefaktów wynikających z syntezy cDNA lub niskiej jakości sekwencjonowania. Odczyty PacBio zawierające co najmniej dwa eksony z wątpliwymi intronami, zawierającymi bezpośrednie, krótkie powtórzenia są eliminowane ze względu na ryzyko występowania artefaktów wynikających z przełączania matrycy RT podczas syntezy cDNA. Co więcej, odczyty PacBio o wielu eksonach muszą zawierać tylko kanoniczne introny, podczas gdy weryfikacja indywidualnych połączeń pomiędzy eksonami w odczytach PacBio za pomocą krótkich odczytów HiSeq jest opcjonalna. Natomiast odczyty zawierające pojedyncze eksony muszą zostać poliadenylowane, aby zminimalizować ryzyko wystąpienia zanieczyszczeń genomowym DNA. Odczyty, które przejdą walidację są następnie łączone w unikalny zestaw modeli transkryptów (zielony).*

Kluczowym i oryginalnym podejściem na poziomie analizy danych jest „kontrolowane łączenie" – innowacyjna strategia budowania modeli transkryptów, która zachowuje wewnętrzne miejsca startu transkrypcji (TSS) i miejsca poliadenylacji[63] (Rysunek 9A). Strategia „kontrolowanego łączenia" pozwala na uzyskanie większej liczby modeli transkryptów w porównaniu ze standardowymi metodami łączenia odczytów (kontrolowane łączenie: 179,993/129,556 w porównaniu z 117,258/87,932 dla standardowego łączenia, uzyskanych w wyniku scalenia 771,585/604,199 odczytów ROI we wszystkich biotypach i próbkach CLS). Jedna z niedawno opublikowanych prac naukowych podkreśliła użyteczność tego podejścia. Hirabayashi i wsp. wykryli istnienie dwóch alternatywnych promotorów: głównego i podrzędnego w ludzkim genie *ZNHIT1*, stosując podejście NET-CAGE (Native Elongating Transcript-CAGE)[74] (Rysunek 9B). Główny i podrzędny promotor znajdują się w odległości ~500 nukleotydów od siebie. Według danych NET-CAGE transkrypty wytwarzane przez podrzędny promotor ulegają degradacji znacznie szybciej, niż te uzyskane w ramach głównego promotora. W celu weryfikacji tego odkrycia, tj. wykrycia dwóch różnych form transkryptu, które różnią się jedynie długością regionów nieulegających translacji na 5' końcu (5'UTR), Hirabayashi i wsp. użyli modeli transkryptów o pełnej długości, jakie zidentyfikowane zostały za pomocą CLS w linii komórkowej HeLa. Jak pokazano na Rysunku 9B,

wykrycie transkryptów wytwarzanych przez podrzędny promotor nie byłoby możliwe przy zastosowaniu jedynie standardowej metody łączenia transkryptów.



Rysunek 9. Budowa transkryptów metodą CLS. *(A) W konwencjonalnej („standardowej") metodzie tworzenia transkryptów, wewnętrzne miejsca startu transkrypcji i miejsca poliadenylacji, których lokalizacja pokrywa się z sekwencją innego eksonu są tracone. „Kontrolowane łączenie" zachowuje takie miejsca. (B) Kontrolowane i standardowe łączenie odczytów w ludzkim genie kodującym białko ZNHIT1. Używając NET-CAGE w komórkach HeLa, Hirabayashi i wsp.*[74] *zidentyfikowali dwa odrębne promotory (przedstawione jako pionowe linie) w tym genie: „główny" (zielony) i „podrzędny" (pomarańczowy), produkujące odpowiednio krótkie i długie transkrypty. Katalog GENCODE v29 zawiera jedynie długą formę transkryptu (górna ścieżka, niebieski kolor), gdyż do budowy transkryptów wykorzystywane jest standardowe, bardziej „zachłanne" podejście. Metoda CLS pozwala na uzyskanie transkryptów o pełnej długości w tym (niewzbogaconym) regionie w komórkach HeLa (dwie dolne, czarne ścieżki). Standardowa procedura łączenia skutkuje utratą krótkiej formy transkryptu na skutek włączenia jej w poczet tej dłuższej (ścieżka CLS, „Standardowe łączenie"). Natomiast zestaw transkryptów uzyskanych w ramach „kontrolowanego łączenia" pozwala na zachowanie zarówno krótkich, jak i długich transkryptów (ścieżka CLS, „Kontrolowane łączenie").*

3.4. Dokładna rekonstrukcja transkryptów (publikacja H2)

Jak wspomniano wcześniej, problem rekonstrukcji transkryptów *in silico* z krótkich odczytów uzyskiwanych w ramach sekwencjonowania RNA utrudnia zastosowanie tej techniki do budowania katalogów genów. Aby ocenić przewagę długich odczytów nad krótkimi w tym procesie, porównaliśmy modele transkryptów CLS z modelami, które zostały zrekonstruowane z danych Illumina HiSeq (krótkie odczyty), także otrzymanych z wzbogaconych bibliotek cDNA. Do rekonstrukcji transkryptów użyliśmy StringTie[75]- oprogramowania, które pozwala na uzyskanie z danych sekwencjonowania RNA o krótkich odczytach, modeli transkryptów o najwyższej jakości i powszechnie uznawane jest za referencję w tej dziedzinie[76]. Wyniki wskazują na wyraźną przewagę modeli transkryptów CLS w porównaniu do modeli StringTie. Tylko 116 (~0,8%) modeli StringTie zostało sklasyfikowanych jako transkrypty o pełnej długości, podczas gdy liczba ta wynosiła 4,763 (~20%) dla modeli transkryptów CLS. Otrzymane rezultaty pokazują, że StringTie ma tendencję do nadmiernego wydłużania modeli transkryptów poza rzeczywiste granice ich 5' oraz 3' końców. Nasza analiza po raz pierwszy została wykonana w sposób ilościowy i wykazała wyraźną przewagę długich odczytów nad krótkimi w kontekście tworzenia katalogów lncRNA.

3.5. Nowa definicja genomowych cech lncRNA (publikacja H2)

Rzetelna charakterystyka genomowych i transkryptomicznych właściwości lncRNA ma kluczowe znaczenie dla zrozumienia ich fizjologicznych i patologicznych ról w komórce. Projekt CLS umożliwił weryfikację genomicznych cech lncRNA, które zostały opisane przez wcześniejsze doniesienia naukowe[23,77], poprzez dostarczenie katalogu transkryptów o pełnej długości. W ten sposób CLS przyczynił się do aktualizacji poglądu na temat procesu regulacji transkrypcji i obróbki post-transkrypcyjnej lncRNA. Analiza modeli transkryptów CLS o pełnej długości wykazała, że dojrzałe lncRNA są dłuższe (mediana wynosząca 1,108/1,067 nukleotydów), niż jak wskazano wcześniej (mediana wynosząca 668/715 nukleotydów). Chociaż sformułowanie definitywnych stwierdzeń dotyczących względnych długości mRNA i lncRNA nie było możliwe ze względów technicznych (ograniczona długość odczytów PacBio oraz podział bibliotek cDNA na frakcje względem długości), nie znaleziono dowodów wskazujących na to, że lncRNA są znacznie krótsze od mRNA[23]. Problem analizy długości transkryptów obecnie może zostać łatwo rozwiązany poprzez zastosowanie nowszego sekwenatora PacBio Sequel II, który zapewnia lepszą dokładność oraz znacznie większą przepustowość sekwencjonowania, minimalizując przy tym konieczność podziału bibliotek cDNA względem rozmiaru fragmentów.

Projekt CLS umożliwił także weryfikację doniesienia pokazującego, iż większość transkryptów lncRNA składa się jedynie z dwóch eksonów[23]. Wyniki naszych badań nie potwierdziły tego zjawiska, jednocześnie wskazując, iż rezultat ten może być artefaktem wynikającym z niekompletności modeli lncRNA w katalogach. Według naszych badań, średnia liczba eksonów dla lncRNA o pełnej długości wynosiła 4,27/3,54 w porównaniu z 6,69/6.98 dla mRNA.

Jednym z podstawowych pytań jest to, czy poprawa jakości modeli transkryptów lncRNA ujawni wcześniej niezidentyfikowane otwarte ramki odczytu (ORF). Zaadresowaliśmy ten problem poprzez zastosowanie podejścia *in silico* do badania potencjału kodowania białka w transkryptach CLS o pełnej długości. Do tego celu wybraliśmy dwa najbardziej popularne narzędzia: CPAT[78] (Coding Potential Assessment Tool), który oparty jest na modelu logistycznej regresji bez konieczności przyrównywania sekwencji oraz PhyloCSF[79] – podejście oparte na genomice porównawczej. Obie metody wskazały, że jedynie niewielka frakcja lncRNA w zestawie CLS może kodować białka. Jednocześnie należy zaznaczyć, iż żadna z tych metod nie jest wystarczająco czuła, aby wykryć obecność niekanonicznych ramek odczytu, które czasami identyfikowane są w sekwencji różnych lncRNA[80]. Aby jednoznacznie wyjaśnić problem potencjału kodującego białko w lncRNA, konieczne jest przeprowadzenie dużo bardziej szczegółowej analizy. Przykładowo, niedawno opublikowane wyniki ujawniły różnice w składnie trójnukleotydowym dla kanonicznych i niekanonicznych ramek odczytu, znajdujących się odpowiednio w mRNA i lncRNA[58].

Ogólnie rzecz biorąc, wyniki przeprowadzonej przez nas analizy zachowawczości ewolucyjnej są zgodne z wcześniejszymi doniesieniami i potwierdzają niską zachowawczość eksonów lncRNA. Jednakże

jednocześnie otrzymane przez nas rezultaty jednoznacznie wskazują, iż sekwencje w obrębie startu transkrypcji są zachowane ewolucyjnie. Efekt ten jest widoczny nawet po wyłączeniu z analizy danych dla promotorów dwukierunkowych, które są współdzielone z genami kodującymi białko. Wyniki te mogą wskazywać na *cis*-regulacyjną rolę tych lncRNA w drodze samego „aktu transkrypcji”[51].

Ulepszona definicja 5’ końców w modelach transkryptów CLS pozwoliła nam także porównać regiony promotorowe lncRNA i mRNA. W przeciwieństwie do publikowanych wcześniej wyników, które wskazywały na wyraźnie różnice[25,81], zaobserwowaliśmy serię podobnych i rozbieżnych cech promotorów lncRNA i mRNA. Analiza ta została wykonana przy użyciu zestawu lncRNA i mRNA, dopasowanych pod względem ekspresji oraz danych ChIP-Seq z ENCODE dla linii komórkowych HeLa oraz K562. Co ciekawe, aktywujące modyfikacje histonów w regionie promotorowym, takie jak H3K4me3 (trimetylacja histonu 3 w lizynie 4) i H3K9ac (acetylacja histonu 3 w lizynie 9) były prawie identyczne dla pełnej długości lncRNA i mRNA. Wynik ten pokazuje, że lncRNA nie mają unikalnej architektury regionów promotorowych, a obserwowane wcześniej różnice mogą wynikać z różnic w poziomach ekspresji pomiędzy mRNA i lncRNA lub niekompletności modeli lncRNA w obrębie 5’ końców. Zatem poleganie jedynie na danych zawartych w katalogach genów może skutkować niedokładną analizą regionów promotorowych[25,81].

Nasza analiza wskazała również na obecność cech odróżniających regiony promotorowe dla lncRNA oraz mRNA. Po raz pierwszy wykazaliśmy, że promotory lncRNA ogólnie wykazują podwyższony poziom represyjnych znaczników chromatyny, w tym H3K9me3 (trimetylacja histonu 3 w lizynie 9) i H3K27me3 (trimetylacja histonu 3 w lizynie 27). Wydaje się to być konsekwencją zwiększonej rekrutacji lncRNA do represyjnego kompleksu Polycomb 2 (PCR2), o czym świadczy jego katalityczna podjednostka Ezh2, która deponuje znaki H3K27me3[82,83]. Byliśmy również w stanie rozróżnić promotory lncRNA na podstawie podwyższonego poziomu wiązania CTCF (czynnik wiążący CCCTC) w komórkach HeLa, które jest wielozadaniowym białkiem wiążącym DNA, zaangażowanym w regulację transkrypcji, izolację genów i trójwymiarową organizację genomu[84].

3.6. Analiza ograniczeń CLS – usprawnienia w identyfikacji lncRNA (publikacja H2 and H3)

Podsumowując, publikacja H2 poprzez połączenie wzbogacania RNA z sekwencjonowaniem trzeciej generacji PacBio, metodą długich odczytów znacząco ułatwia proces tworzenia katalogów genów. CLS pozwala na tworzenie transkryptów o bardzo wysokiej jakości przy znacznie niższych kosztach i w wysokoprzepustowej skali. Co ważne, metoda CLS nie jest ograniczona do stosowania wyłącznie platformy PacBio i może być połączona z dowolnie dostępną platformą do sekwencjonowania RNA. Ponadto CLS może znaleźć zastosowanie do wzbogacenia dowolnie wybranych regionów o niskiej ekspresji, w tym także regionów kodujących białko. Dachówkowo ułożone sondy pozwalają na skuteczne

wzbogacenie nawet bardzo dużych regionów genomowych, w przeciwieństwie do wspomnianej wcześniej metody RACE-seq. CLS posiada także imponujący potencjał zwiększenia skali, który został dodatkowo powiększony przez aktualizację wersji zestawów do wzbogacania bibliotek przez firmę Roche, pozwalających obecnie na wzbogacenie do 100 megazasad sekwencji genomowej (w układzie dachówkowym).

Chociaż wykazano, że podejście CLS jest bardzo skuteczne, ma ono kilka ograniczeń, które mogą zostać zaadresowane na etapie przyszłych wdrożeń tej metody. Przykładowo, poziom kompletnych transkryptów może zostać zwiększony przez zastosowanie technik wzbogacania 5'kapu, takich jak CAP Trapper[85,86], które zostały opracowane przez instytut RIKEN. Pilotażowe eksperymenty dla GENCODE pokazują, że metoda CapTrap (Silvia Carbonell Sala) pomaga przezwyciężyć problem niekompletności końców transkryptów (dane nie zostały ujawnione). Rozwój technologii sekwencjonowania trzeciej generacji, który obserwujemy na przestrzeni kilku ostatnich lat, ofiaruje zupełnie nowe możliwości, które nie były dostępne na etapie realizacji projektu CLS. Firma PacBio dokonała znacznych, imponujących ulepszeń swoich platform. Najnowsza platforma Sequel II oferuje znacznie dłuższe odczyty, większą dokładność, która porównywalna jest z dokładnością sekwencjonowania za pomocą platformy HiSeq oraz przepustowość na poziomie > 4M odczytów (~30,000 dla PacBio RS II) w ramach pojedynczego eksperymentu przy znacznie niższym koszcie. Co więcej, proponowana przepustowość eliminuje konieczność podziału bibliotek cDNA na frakcje względem długości. Jednocześnie firma Oxford Nanopore Technologies (ONT) oferuje niedrogą i dostępną alternatywę dla sekwencjonowania PacBio. Chociaż odczyty uzyskiwane za pomocą ONT charakteryzują się niższym poziomem jakości, wyraźną zaletą tej platformy jest możliwość bezpośredniego sekwencjonowania RNA, które pozwala na analizę modyfikacji RNA przy zachowaniu wysokiej przepustowości. Wreszcie analiza frakcji niekodujących RNA, które nie ulegają poliadenylacji, może pomóc zidentyfikować wiele nowych sekwencji regulatorowych[87], w tym wzmacniaczy transkrypcji, które były słabo reprezentowane w projekcie CLS.

3.7. Jaki jest poziom kompletności katalogów lncRNA? (publikacja H3)

Podejście CLS stanowi znaczący krok w kierunku kompletnej mapy genów lncRNA w genomie człowieka i myszy. W ramach tego projektu przeprowadziliśmy także analizę, aby zrozumieć, jak daleko nam do ukończenia katalogów lncRNA. W tym celu stworzyliśmy krzywe nasycenia, które wskazują, iż każdy krok w kierunku głębszej analizy transkryptomu prowadzi do odkrycia większej liczby transkryptów oraz połączeń pomiędzy eksonami we wszystkich badanych tkankach. Co ciekawe, zaobserwowaliśmy ten sam efekt dla zestawu transkryptów CLS o pełnej długości, zarówno w przypadku danych o krótkich, jak i długich odczytach, uzyskanych ze wzbogaconych bibliotek. Brak nasycenia oznacza, że wiele autentycznych transkryptów wciąż pozostaje niezidentyfikowanych. Wynik ten został potwierdzony przez

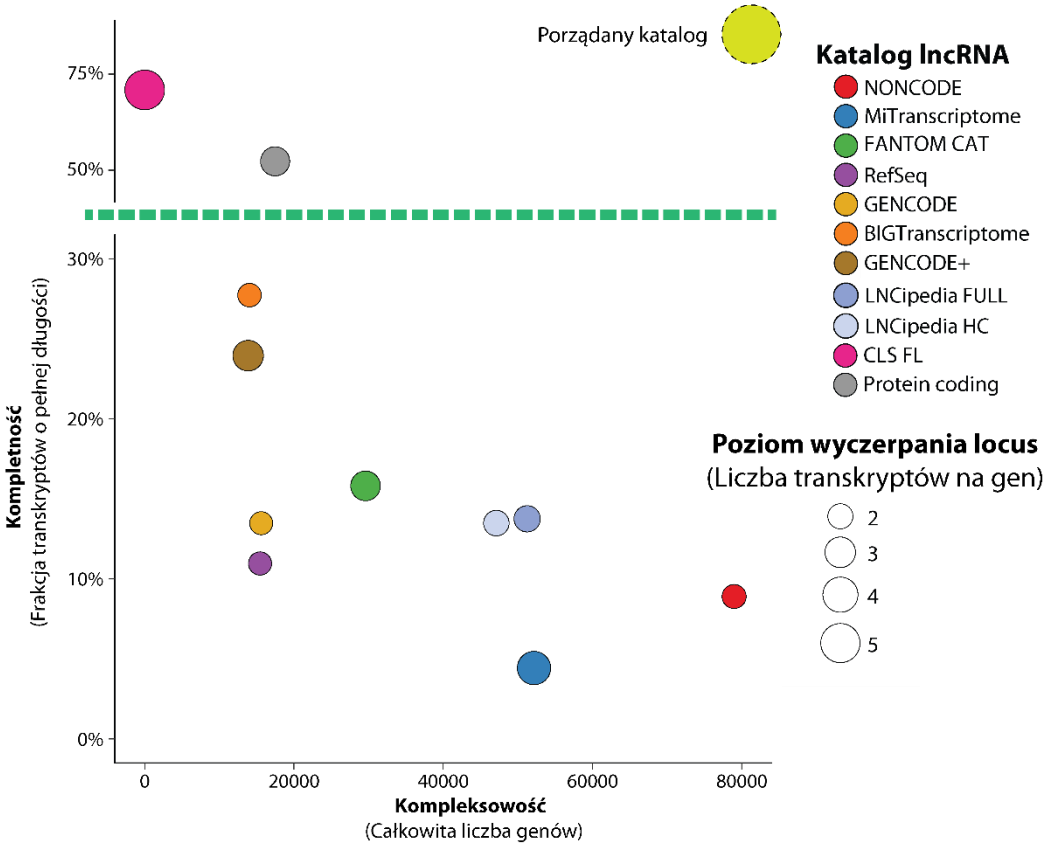
inne, niezależne eksperymenty obejmujące głęboką analizę transkryptomów[88,89]. Stanowi to niezbity dowód na to, że pomimo ogromnych wysiłków włożonych w ich stworzenie, współczesne katalogi lncRNA wciąż pozostają niekompletne, z tysiącami niezidentyfikowanych genów, transkryptów oraz eksonów[65].

W ostatnim czasie Deveson i wsp. zastosowali podejście RNA CaptureSeq – wzbogacenie bibliotek cDNA w sprzężeniu z krótkimi (Illumina HiSeq) oraz długimi odczytami (PacBio RS II), aby w wysokiej rozdzielczości przeanalizować proces transkrypcji w obrębie całego chromosomu 21 w genomie człowieka, w tym w regionach lncRNA oraz genów kodujących białko[88]. Analiza wykazała zasadnicze różnice między genami niekodującymi i tymi kodującymi białka. W rezultacie Deveson i wsp. wprowadzili hipotezę „uniwersalnego alternatywnego składania transkryptów", wskazującą na nieograniczone kombinacje eksonów w strukturach transkryptów lncRNA, ale nie w genach kodujących białka, które wykazały ustaloną liczbę kombinacji eksonów. Model ten wskazuje na istnienie ogromnego rezerwuaru różnych struktur w transkryptach lncRNA. Chociaż biologiczne znaczenie uniwersalnego, alternatywnego składania transkryptów nie zostało jeszcze w pełni ustalone, wyniki te wskazują, iż na ukończenie procesu katalogowania lncRNA przyjdzie nam jeszcze bardzo długo poczekać.

Obecnie dostępnych jest szereg katalogów lncRNA, które różnią się od siebie wieloma aspektami, m.in. sposobem ich tworzenia. Każdy z tych katalogów posiada wady i zalety, które nie są od razu widoczne, ale mogą mieć znaczący wpływ na jakość końcowych wyników. Jednym z głównych celów moich badań było opracowanie metody umożliwiającej porównanie i ocenę istniejących katalogów lncRNA (oraz innych genów) w sposób policzalny. W ramach publikacji H3[65] wprowadzone zostały pojęcia: *kompletności* (ułamek transkryptów o pełnej długości), *kompleksowości* (całkowita liczba genów) oraz *poziomu wyczerpania locus* (liczba transkryptów w danym locus) w celu scharakteryzowania istniejących katalogów lncRNA. Korzystając z tych wskaźników, porównaliśmy różne, publicznie dostępne katalogi lncRNA, w tym NONCODE[26], MiTranscriptome[90], FANTOM CAT[91], RefSeq[92], GENCODE[5] (v27), BIGTranscriptome[93] and GENCODE+, który jest wynikiem połączenia katalogów CLS oraz GENCODE (v20) w sposób automatyczny (Rysunek 10). Całkowita liczba genów oraz transkryptów lncRNA pozostaje nieznana, dlatego też nie można określić docelowych wartości dla kompleksowości oraz wyczerpania locus. Niemniej jednak możliwe jest uzyskanie przybliżonych szacunków ich wartości dla każdego katalogu w celu ich porównania. Aby zapewnić rzetelną analizę katalogów, na nowo zdefiniowaliśmy granice modeli genów za pomocą narzędzia buildLoci, które jest naszym autorskim oprogramowaniem (Julien Lagarde).

Najbardziej uderzającym odkryciem jest brak korelacji między kompleksowością, a kompletnością. Katalogi z największą liczbą genów, takich jak NONCODE lub MiTranscriptome (62,276 i 45,088), wykazały wyjątkowo niski odsetek modeli transkrypcji o pełnej długości (8,9% i 4,4%). Jednocześnie katalogi takie jak BIGTranscriptome lub GENCODE + z największą liczbą kompletnych transkryptów

(27,7% i 24%) uzyskały niski wynik pod względem kompleksowości (12,632 i 13,434). Innymi słowy, istnieje kompromis między jakością, a rozmiarem katalogów lncRNA.



Rysunek 10. Porównanie najpopularniejszych katalogów lncRNA. *Analiza wskaźników jakości dla wybranych katalogów lncRNA. Oś X: Kompleksowość lub całkowita liczba genów; Oś Y: kompletność lub procent struktur transkryptu, których początek został zweryfikowanych za pomocą danych CAGE (cap analysis of gene expression) przygotowanych w ramach projektu FANTOM (Functional Annotation of the Mammalian genome). Koniec 5' uznaje się za kompletny jeśli znajduje się w odległości ± 50 nukleotydów od jednego z silnych klastrów CAGE uzyskanych w ramach fazy 1/2 projektu (n = 201,802). Kompletność 3' końców wyznaczono na podstawie obecności kanonicznych motywów poliadenylacji*[94] *w przedziale 10-50 nukleotydów powyżej tego miejsca. Średnice kół określają średnią liczbę transkryptów na gen. Katalog GENCODE+ powstał poprzez automatyczne połączenie katalogów GENCODE v20 oraz zestawu transkryptów uzyskanych w ramach projektu CLS. Zestaw kodujący białka to zestaw mRNA o wysokiej jakości struktur z katalogu GENCODE, który został wyselekcjonowany zgodnie z opisem przedstawionym w ramach publikacji CLS*[63].

Jeśli chodzi o kompletność, żaden z publicznie dostępnych katalogów lncRNA nie zbliżył się do poziomu genów kodujących białko GENCODE (53,8%). Podejście CLS oraz najnowsze technologiczne ulepszenia metod sekwencjonowania trzeciej generacji dają szansę katalogowi GENCODE na osiągnięcie tego poziomu w najbliższej przyszłości. Ocena różnic między GENCODE, a GENCODE+, wskazuje, że CLS w znacznym stopniu poprawiło jakość transkryptów lncRNA w GENCODE, zwiększając kompletność z 13,5% do 24% oraz poziom wyczerpania locus z 1,9 do 3,3 transkryptów na gen.

Co ciekawe, porównywane katalogi znacząco różniły się od siebie składem genów. Nawet dwa najpopularniejsze katalogi GENCODE i RefSeq zawierają mniej niż 50% wspólnych genów lncRNA. Dokładna analiza połączeń pomiędzy eksonami wykazała, iż niektóre katalogi, np. NONCODE zawierają wysoki odsetek fałszywie pozytywnych struktur.

Ta analiza nie tylko wskazuje mocne i słabe strony współczesnych katalogów, ale także rodzi pytania o potencjalne źródła niekompletności katalogów. Identyfikacja lncRNA odbywa się prawie wyłącznie na podstawie fizycznych dowodów transkryptomicznych. Zatem kwestie techniczne związane z ich uzyskiwaniem mogą wpływać na proces identyfikacji i opisu lncRNA. Na przykład, cząsteczki cDNA często mają tendencję do skracania 5' końców z powodu degradacji RNA i możliwości odłączenia odwrotnej transkryptazy przed osiągnięciem 5' końca matrycy RNA, często w wyniku obecności drugorzędowych struktur RNA[95]. Jak wspomniano wcześniej, problemy techniczne związane z rekonstrukcją transkryptów z danych uzyskiwanych za pomocą sekwencjonowania RNA metodą krótkich odczytów, bardzo poważnie wpływają na jakość tworzonych modeli transkryptów.

Kwestia pozostałych aspektów wpływających na niekompletność katalogów jest znacznie bardziej skomplikowana. Śledzenie zmian ekspresji genów w ciągu życia całego organizmu jest trudne, szczególnie w kwestii wykrywania transkryptów o ekspresji tymczasowej. Ostatnie badania ujawniły eksplozywny charakter transkrypcji u kilku organizmów, w tym ssaków[96,97]. Interakcja pomiędzy wytwarzaniem, a degradacją RNA w komórce może skutkować pojawieniem się tymczasowych, odizolowanych pików ekspresji dla poszczególnych genów[98]. Ze względu na fakt, iż eksperyment RNA pozwala uchwycić skład populacji RNA w komórce, jedynie w danym momencie, niezbędne są ulepszenia usprawniające proces detekcji RNA[99,100]. Pojawienie się metod sekwencjonowania na poziomie pojedynczej komórki (ang. *single cell RNA sequencing*) daje nadzieję na usprawnienie analizy transkryptomu dla różnych typów komórek, w różnych warunkach[101,102]. Jednocześnie lepsze zrozumienie funkcji lncRNA, w tym mechanizmów zachowywania ich funkcji w ramach ewolucji, zdecydowanie usprawniłoby proces identyfikacji lncRNA nie tylko dla genomów człowieka i myszy[33].

4. Wnioski

Przedstawione wyniki pozwalają na wyciągnięcie pięciu głównych wniosków:

1. Połączenie RACE z sekwencjonowaniem o średniej długości odczytów (technologia 454) pozwoliło na identyfikację wielu nowych transkryptów. Jednakże tylko 50% zidentyfikowanych transkryptów posiada kompletne struktury w danym locus, co wskazuje na potrzebę stosowania metod sekwencjonowania o dłuższych odczytach. Ponadto, nasz pilotażowy eksperyment pokazał, iż potencjał skalowalności metody RACE-seq jest ograniczony. Podobnie jak zdolność tej metody do wzbogacania transkryptów o bardzo niskiej ekspresji.

2. Wprowadzenie metody CLS pozwoliło na znacznie usprawnienie identyfikacji transkryptów. CLS jest metodą ukierunkowanego sekwencjonowania, w ramach której po raz pierwszy połączono ze sobą wzbogacanie bibliotek cDNA za pomocą sond oligonukleotydowych oraz sekwencjonowania o długich odczytach. Technika CLS pozwala na uzyskanie modeli transkryptów dla lncRNA o wysokiej jakości, porównywalnej z modelami budowanymi ręcznie, w dużej skali. Uzyskane przez nas wyniki pokazują wyraźną przewagę (także wyrażoną ilościowo) metody CLS nad podejściem rekonstrukcji transkryptów z danych uzyskiwanych za pomocą sekwencjonowania o krótkich odczytach.
3. Zastosowanie CLS oraz metody RACE-seq (w mniejszym stopniu) pozwoliło na poprawę jakości katalogu GENCODE dla genomów człowieka i myszy. Podejście CLS w szczególności umożliwiło identyfikację wielu nowych struktur lncRNA i poprawę definicji granic ich transkryptów. Dlatego też dzięki dobrze zdefiniowanemu zestawowi lncRNA ze strukturami transkryptów o pełnej długości, po raz pierwszy możliwe było przeprowadzenie pełnej charakterystyki genomowych właściwości lncRNA w genomach ssaków.
4. Projekt CLS umożliwił weryfikację ważnych, genomowych cech lncRNA:
    a. Ulepszenie i rozszerzenie struktur lncRNA zawartych w katalogach genów nie skutkowało wzrostem potencjału kodowania białka. Stąd też można uznać, iż lncRNA faktycznie nie kodują białek.
    b. Dojrzałe transkrypty lncRNA są znacznie dłuższe, niż opisano wcześniej, a ich długość nie różni się znacząco od długości transkryptów kodujących białka.
    c. Analiza środowiska chromatyny ujawniła, że lncRNA i mRNA mają podobną architekturę aktywnych promotorów. Jednakże promotory lncRNA są bardziej wzbogacone w represyjne modyfikacje, niż promotory mRNA.
5. Porównanie publicznie dostępnych katalogów wykazało brak korelacji pomiędzy wielkością, a jakością katalogów. Nieoczekiwanie, badane katalogi znacznie różnią się składem genów, nawet te o strukturach weryfikowanych ręcznie (GENCODE, RefSeq). Obserwacje te wskazują na ogromną potrzebę ulepszenia katalogów poprzez zwiększenie ich kompletności, a następnie poprawę ich kompleksowości, dzięki połączeniu ze sobą różnych, dostępnych zasobów. Kluczowym aspektem jest także uzyskanie ujednoliconego zestawu lncRNA.

LITERATURA

1. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science (80-.).* 291, 1304 LP – 1351 (2001).
2. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* 409, 860–921 (2001).
3. Ezkurdia, I. *et al.* Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum. Mol. Genet.* 23, 5866–5878 (2014).

4. Ma, L. *et al.* LncBook: a curated knowledgebase of human long non-coding RNAs. *Nucleic Acids Res.* 1–7 (2018). doi:10.1093/nar/gky960
5. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 1–8 (2018). doi:10.1093/nar/gky955
6. Liang, F. *et al.* Gene Index analysis of the human genome estimates approximately 120 , 000 genes. 25, 239–240 (2000).
7. Willyard, C. New human gene tally reignites debate. *Nature* 558, 354–355 (2018).
8. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* 489, 101–8 (2012).
9. Mele, M. *et al.* The human transcriptome across tissues and individuals. *Science (80-.).* 348, 660–665 (2015).
10. Pertea, M. *et al.* CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* 19, 208 (2018).
11. Jungreis, I. *et al.* Nearly all new protein-coding predictions in the CHESS database are not protein-coding. *bioRxiv* 360602 (2018). doi:10.1101/360602
12. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* 489, 101–8 (2012).
13. Palazzo, A. F. & Lee, E. S. Non-coding RNA: what is functional and what is junk? *Front. Genet.* 6, 2 (2015).
14. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.* 14, 103–105 (2007).
15. Graur, D. An Upper Limit on the Functional Fraction of the Human Genome. *Genome Biol. Evol.* 9, 1880–1885 (2017).
16. Ohno, S. So much 'junk' DNA in our genome. *Brookhaven Symp. Biol.* 23, 366–370 (1972).
17. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nature reviews. Genetics* 14, 288–295 (2013).
18. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465, 182–187 (2010).
19. De Santa, F. *et al.* A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol.* 8, e1000384 (2010).
20. Bose, D. A. *et al.* RNA Binding to CBP Stimulates Histone Acetylation and Transcription. *Cell* 168, 135-149.e22 (2017).
21. Ezkurdia, I. *et al.* The potential clinical impact of the release of two drafts of the human proteome. *Expert Rev. Proteomics* 12, 579–593 (2015).
22. Banfai, B. *et al.* Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* 22, 1646–1657 (2012).
23. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22, 1775–89 (2012).
24. Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* 22, 1616–1625 (2012).
25. Mele, M. *et al.* Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res.* 27, 27–37 (2017).
26. Fang, S. *et al.* NONCODEV5: A comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.* 46, D308–D314 (2018).
27. Volders, P. J. *et al.* An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic acids research* 43, 4363–4364 (2015).
28. Quek, X. C. *et al.* lncRNAdb v2.0: Expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.* 43, D168–D173 (2015).
29. Bao, Z. *et al.* LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res.* 1–4 (2018). doi:10.1093/nar/gky905
30. Palazzo, A. F. & Gregory, T. R. The case for junk DNA. *PLoS Genet.* 10, e1004351 (2014).
31. van Bakel, H., Nislow, C., Blencowe, B. J. & Hughes, T. R. Most 'dark matter' transcripts are

associated with known genes. *PLoS Biol.* 8, e1000371 (2010).
32. Clark, M. B. *et al.* The Reality of Pervasive Transcription. *PLoS Biol.* 9, e1000625 (2011).
33. Ulitsky, I. & Bartel, D. P. lincRNAs: genomics, evolution, and mechanisms. *Cell* 154, 26–46 (2013).
34. Hezroni H *et al.* Priniciples of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* 11, 1110–1122 (2016).
35. Gutschner, T., Baas, M. & Diederichs, S. Noncoding RNA gene silencing through genomic integration of RNA destabilizing elements using zinc finger nucleases. *Genome Res.* 21, 1944–1954 (2011).
36. Stojic, L. *et al.* Specificity of RNAi, LNA and CRISPRi as loss-of-function methods in transcriptional analysis. *Nucleic Acids Res.* 46, 5950–5966 (2018).
37. Liu, Y. *et al.* Genome-wide screening for functional long noncoding RNAs in human cells by Cas9 targeting of splice sites. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4283
38. Chow, J. C., Yen, Z., Ziesche, S. M. & Brown, C. J. Silencing of the Mammalian X Chromosome. *Annu. Rev. Genomics Hum. Genet.* 6, 69–92 (2005).
39. Simon, M. D. *et al.* The genomic binding sites of a noncoding RNA. *Proc. Natl. Acad. Sci. U. S. A.* 108, 20497–20502 (2011).
40. Grant, J. *et al.* Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. *Nature* 487, 254–258 (2012).
41. Brown, C. J. *et al.* The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* 71, 527–542 (1992).
42. Ravasi, T. *et al.* Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.* 16, 11–19 (2006).
43. Grote, P. & Herrmann, B. G. Long noncoding RNAs in organogenesis: Making the difference. *Trends Genet.* 31, 329–335 (2015).
44. Congrains, A. *et al.* Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of ANRIL and CDKN2A/B. *Atherosclerosis* 220, 449–455 (2012).
45. Johnson, R. Long non-coding RNAs in Huntington's disease neurodegeneration. *Neurobiol. Dis.* 46, 245–254 (2012).
46. Huarte, M. The emerging role of lncRNAs in cancer. *Nat. Med.* 21, 1253–1261 (2015).
47. Sanchez, Y. *et al.* Genome-wide analysis of the human p53 transcriptional network unveils a lncRNA tumour suppressor signature. *Nat. Commun.* 5, 5812 (2014).
48. Hosono, Y. *et al.* Oncogenic Role of THOR, a Conserved Cancer/Testis Long Non-coding RNA. *Cell* 171, 1559-1572.e20 (2017).
49. Lanzós, A. *et al.* Discovery of Cancer Driver Long Noncoding RNAs across 1112 Tumour Genomes: New Candidates and Distinguishing Features. *Sci. Rep.* 7, 1–16 (2017).
50. Groff, A. F. *et al.* In Vivo Characterization of Linc-p21 Reveals Functional cis-Regulatory DNA Elements. *Cell Rep.* 16, 2178–2186 (2016).
51. Latos, P. A. *et al.* Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science* 338, 1469–1472 (2012).
52. Carlevaro-Fita, J. *et al.* Unique genomic features and deeply-conserved functions of long non-coding RNAs in the Cancer LncRNA Census (CLC). *bioRxiv* 152769 (2017). doi:10.1101/152769
53. Matsumoto, A. *et al.* mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* 541, 228–232 (2017).
54. Anderson, D. M. *et al.* A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* 160, 595–606 (2015).
55. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324, 218–23 (2009).
56. Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nature reviews. Genetics* 15, 205–213 (2014).
57. Choi, S.-W., Kim, H.-W. & Nam, J.-W. The small peptide world in long noncoding RNAs. *Brief. Bioinform.* 20, 1853–1864 (2019).

58. Brümmer, A., Dreos, R., Marques, A. C. & Bergmann, S. LincRNA sequences are biased to counteract their translation. *bioRxiv* 737890 (2020). doi:10.1101/737890
59. Kowalczyk, M. S., Higgs, D. R. & Gingeras, T. R. Molecular biology: RNA discrimination. *Nature* 482, 310–311 (2012).
60. Liu, S. J. *et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* 355, (2017).
61. Shin, J. Functional Annotation of Human Long Non-Coding RNAs via Molecular Phenotyping. (2019).
62. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* 38, 1–7 (2010).
63. Lagarde, J. *et al.* High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat. Genet.* 49, 1731–1740 (2017).
64. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10, 1177–84 (2013).
65. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nat. Rev. Genet.* 19, 535–548 (2018).
66. Pervouchine, D. D. *et al.* Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. *Nat. Commun.* 6, 5903 (2015).
67. Lagarde, J. *et al.* Extension of human lncRNA transcripts by RACE coupled with long-read high-throughput sequencing (RACE-Seq). *Nat. Commun.* 7, (2016).
68. Yeku, O. & Frohman, M. A. Rapid amplification of cDNA ends (RACE). *Methods Mol. Biol.* 703, 107–22 (2011).
69. Olivarius, S., Plessy, C. & Carninci, P. High-throughput verification of transcriptional starting sites by Deep-RACE. *Biotechniques* 46, 130–2 (2009).
70. Forrest, A. R. R. *et al.* A promoter-level mammalian expression atlas. *Nature* 507, 462–70 (2014).
71. Kralj, J. G. & Salit, M. L. Characterization of in vitro transcription amplification linearity and variability in the low copy number regime using External RNA Control Consortium (ERCC) spike-ins. *Anal. Bioanal. Chem.* 405, 315–320 (2013).
72. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012).
73. Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M. & Gautheret, D. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* 10, 1001–1010 (2000).
74. Hirabayashi, S. *et al.* NET-CAGE characterizes the dynamics and topology of human transcribed cis-regulatory elements. *Nat. Genet.* 51, 1369–1379 (2019).
75. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295 (2015).
76. Shao, M. & Kingsford, C. Accurate assembly of transcripts through phase-preserving graph decomposition. *Nat. Biotechnol.* 35, 1167–1169 (2017).
77. Cabili MN1, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, R. J. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25, 1915–1927 (2011).
78. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* 41, e74 (2013).
79. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* 27, i275-82 (2011).
80. Chen, J. *et al.* Pervasive functional translation of noncanonical human open reading frames. *Science* 367, 1140–1146 (2020).
81. Alam, T. *et al.* Promoter analysis reveals globally differential regulation of human long non-coding RNA and protein-coding genes. *PLoS One* 9, e109443 (2014).
82. Ku, M. *et al.* Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.* 4, e1000242 (2008).

83. Cao, R. & Zhang, Y. The functions of E(Z)/EZH2-mediated methylation of lysine 27 in histone H3. *Curr. Opin. Genet. Dev.* 14, 155–164 (2004).
84. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* 137, 1194–1211 (2009).
85. Carninci, P. & Hayashizaki, Y. High-efficiency full-length cDNA cloning. *Methods Enzymol.* 303, 19–44 (1999).
86. Carninci, P. *et al.* High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* 37, 327–336 (1996).
87. Lorenzi, L. *et al.* The RNA Atlas, a single nucleotide resolution map of the human transcriptome. *bioRxiv* 807529 (2019). doi:10.1101/807529
88. Deveson, I. W. *et al.* Universal Alternative Splicing of Noncoding Exons. *Cell Syst.* 6, 245-255.e5 (2018).
89. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* 32, 903–914 (2014).
90. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* 47, 199–208 (2015).
91. Hon, C. C. *et al.* An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* 543, 199–204 (2017).
92. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745 (2016).
93. You, B. H., Yoon, S. H. & Nam, J. W. High-confidence coding and noncoding transcriptome maps. *Genome Res.* 27, 1050–1062 (2017).
94. Lopez, F., Granjeaud, S., Ara, T., Ghattas, B. & Gautheret, D. The disparate nature of 'intergenic' polyadenylation sites. *Rna* 12, 1794–1801 (2006).
95. Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R. & Siebert, P. D. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* 30, 892–897 (2001).
96. Dar, R. D. *et al.* Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc. Natl. Acad. Sci. U. S. A.* 109, 17454–17459 (2012).
97. Pedraza, J. M. & Paulsson, J. Effects of molecular memory and bursting on fluctuations in gene expression. *Science* 319, 339–343 (2008).
98. Rabani, M. *et al.* Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* 29, 436–442 (2011).
99. Sarropoulos, I., Marin, R., Cardoso-Moreira, M. & Kaessmann, H. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* 571, 510–514 (2019).
100. Field, A. R. *et al.* Structurally Conserved Primate LncRNAs Are Transiently Expressed during Human Cortical Differentiation and Influence Cell-Type-Specific Genes. *Stem cell reports* 12, 245–257 (2019).
101. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382 (2009).
102. Cabili, M. N. *et al.* Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol.* 16, 20 (2015).