Uniwersytet im. Adama Mickiewicza w Poznaniu
Wydział Biologii
Instytut Antropologii

Wojciech Rosikiewicz

Identyfikacja i analiza ekspresji nakładających się genów u człowieka i myszy

Identification and expression analysis of human and murine overlapping genes

Praca doktorska wykonana
w Zakładzie Genomiki Zintegrowanej
Promotor: prof. dr hab. Izabela Makałowska

Poznań 2017

Czasem mam wrażenie, że jestem na dobrej drodze by stać się prawdziwym naukowcem. Nie byłoby tak bez mojej mentorki, prof. dr hab. Izabeli Makałowskiej, której z całego serca dziękuję za prawie dekadę współpracy.

Serdecznie dziękuję dr Joannie Ciomborowskiej-Basheer i dr Michałowi Szcześniakowi, z którymi zawsze i o wszystkim mogłem porozmawiać. Dziękuję również dr Michałowi Kabzie, dr Elżbiecie Kaja, mgr Magdalenie Kubiak, mgr Elżbiecie Wanowskiej i mgr Alexowi Bryzghalov, z którymi tworzyłem zespół i których do końca życia będę nosić w sercu.

Dziękuję także wszystkim koleżankom i kolegom z Wydziału Biologii UAM, bez których nie byłbym tym, kim jestem.

Pracę dedykuję mojej ukochanej żonie Monice i córce Lilii.

Jesteście moją największą motywacją.

Finansowanie

Niniejsza praca doktorska powstała dzięki dofinansowaniu w ramach grantu uzyskanego z Narodowego Centrum Nauki – NCN (NCN 2013/11/N/NZ2/02524), Grantu Dziekana Wydziału Biologii UAM (GDWB-03/2014) oraz Krajowego Naukowego Ośrodka Wiodącego (KNOW) Poznańskiego Konsorcjum RNA (01/KNOW2/2014).

Spis treści

Streszczenie

Od odkrycia u ssaków pierwszych nakładających się genów kodujących białka minęło ponad trzydzieści lat. Przez długi czas fenomen ten uważany był za marginalny, jednakże z biegiem lat przybywało przykładów par genów nakładających się u wielu gatunków. Zjawisku nakładania się genów przypisano wiele funkcji regulatorowych zarówno na poziomie transkrypcji, jak również post-transkrypcyjnie. Niemniej jednak po dziś dzień nie ustalono jednoznacznie wpływu zjawiska nakładania się genów na ich poziom ekspresji.

Wykorzystując koordynaty alternatywnych miejsc startu transkrypcji (TSS) w 73 ludzkich i 10 mysich bibliotekach, zidentyfikowano 582 ludzkie i 113 mysich par genów kodujących białka, nakładających się końcami 5'. Wykazano, że para genów wykorzystująca do ekspresji nakładające się miejsca TSS w jednej bibliotece, może w innej wykorzystywać alternatywne, nienakładające miejsca startu transkrypcji. Tkankowo specyficzne nakładanie się genów może być częściowo związane z rodzajem tkanki w których geny ulegają ekspresji. Wykazano jednak, że również warunki egzogenne, jak np. transfekcja, mogą mieć wpływ na zachodzenie nakładania, co ma związek z alternatywnym wykorzystaniem promotorów. Badania wykazały również, że przynajmniej część genów nakładających się jest regulowana przez promotory dwukierunkowe. Wykazano ponadto, że poziom ekspresji genów nakładających się jest przeciętnie wyższy aniżeli innych genów. Co więcej, w pracy zademonstrowano, że przy wykorzystaniu nakładających się miejsc TSS, geny nakładające się mają wyższy poziom ekspresji aniżeli te same geny, gdy ulegają one ekspresji z nienakładających się miejsc startu transkrypcji.

Nakładanie się genów kodujących białka zostało również przestudiowane w kontekście stanu chromatyny i aktywności polimerazy RNA II. Analiza ta pozwoliła na zidentyfikowanie wzorców aktywności polimerazy oraz wykazanie, że zjawisko interferencji transkrypcji może zachodzić zarówno gdy geny wykorzystują nakładające się jak i nienakładające się promotory. W niektórych przypadkach brak podwyższonej aktywności polimerazy RNA II, a co za tym idzie brak interferencji w rejonie nakładania się genów, połączony został z monoalleliczną ekspresją genów nakładających się.

Kluczowe wyniki zaprezentowane w niniejszej pracy doktorskiej zostały zdeponowane w publicznie dostępnej bazie danych OverGeneDB, która znajduje się pod adresem internetowym http://overgenedb.amu.edu.pl.

Abstract

Over 30 years have passed since the discovery of the first mammalian overlapping protein-coding genes. For a long time, gene overlap was considered to be rather uncommon, but nowadays more and more of different types of overlapping genes, depending on their position and the transcription direction, is reported in diverse species. Gene overlap have been shown to play various regulatory functions on transcriptional and post-transcriptional levels. However, the true influence of the gene overlap phenomenon on the overlapping genes' expression levels is still a matter of debate.

Here, using information of the alternative transcription start sites (TSS) in 73 human and 10 mouse libraries, a total of 582 human and 113 mice 5' end protein-coding overlapping gene pairs were identified. It was shown, that a single gene pair, identified as overlapping in one library, may use a different set of alternative TSS in different libraries, often resulting in the transcription from non-overlapping regions. Tissue-specific gene overlap patterns may partially be related with a tissue type, but it was also shown that gene overlap may be triggered by the environmental changes like the cells transfection, most possibly as a side effect of the alternative promoters' usage. Conducted studies revealed that at least some of the 5' end protein-coding overlapping genes may be regulated by the bidirectional promoters. It was also shown, that overlapping genes on average tend to have a higher expression level then non-overlapping genes. What is more, it was demonstrated that the expression level of overlapping genes is on average higher when they utilize overlapping promoters, than when the same genes are expressed from non-overlapping transcription start sites.

Protein-coding gene overlap was also studied in the context of the chromatin state and RNA polymerase II (RNAPII) activity. Although it was hard to identify transcriptional interference (TI) down-regulatory effect on the expression level based solely on the studies of the overlapping genes' expression levels, it was possible to find conceivable marks of TI within the RNAPII activity patterns. Interestingly these patterns were identified both when genes were overlapping and when they utilized non-overlapping promoters. In some cases, no enhanced RNA polymerase II activity was found within the overlap regions, which was suggested to be possibly connected with the overlapping genes monoallelic expression.

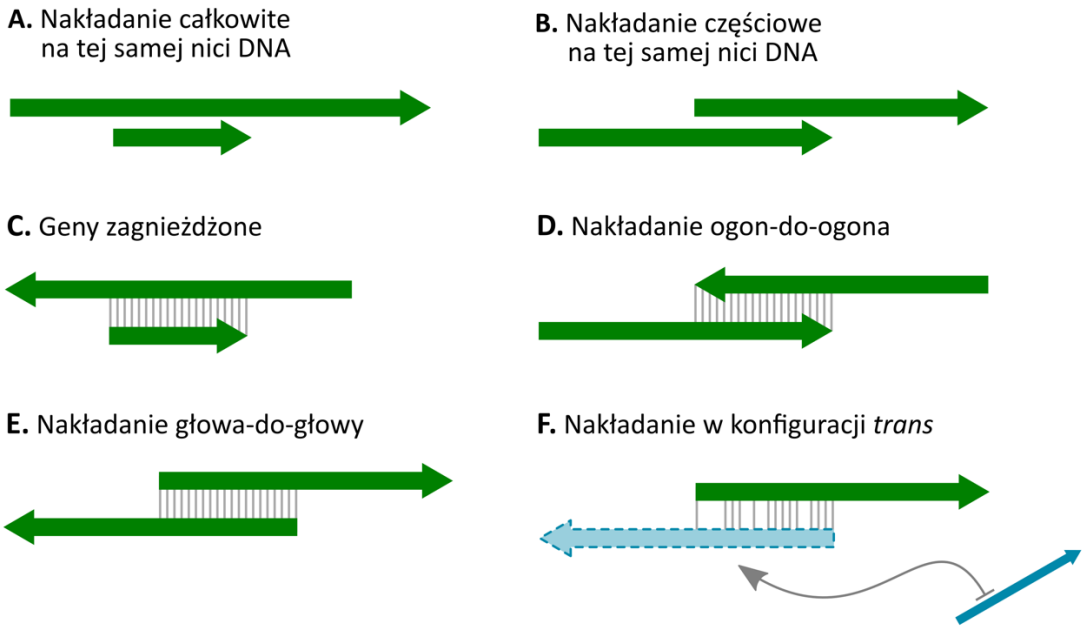
Finally, OverGeneDB database was created so that anyone interested in deeper understanding of the gene overlap phenomenon could explore the key results presented in this project. OverGeneDB is publically accessible under http://overgenedb.amu.edu.pl.

1. Wstęp

Nakładającymi się genami określa się takie geny, które częściowo lub całkowicie współdzielą pewien fragment sekwencji na tej samej lub przeciwnych niciach DNA[1, 2]. Pierwsze wzmianki o nakładających się genach pochodzą sprzed prawie pięćdziesięciu lat, gdzie geny takie odkryte zostały w genomie bakteriofaga λ[3]. Przez wiele lat zjawisko to uważane było za marginalne natomiast wyniki ówczesnych badań sugerowały, że ogranicza się ono głównie do genomów wirusów, które mogą dzięki temu uzyskać znacznie wyższą kompresję genomu[4, 5]. Z biegiem lat zidentyfikowano jednak nakładające się geny w genomie muszki owocowej oraz myszy[6-8], a konsekwentnie również w genomach roślin[9-11], grzybów[12, 13] i innych zwierząt, nie wyłączając genomu człowieka[14-17]. Obecnie szacuje się, że 30% ludzkich i mysich genów może ulegać nakładaniu[18-20], a doniesienia z wyników wielkoskalowych badań prowadzonych przez konsorcjum FANTOM informują, że nawet 72% transkrypcji może być inicjowane na obu niciach DNA tego samego *loci* genomowego[21].

Geny nakładające się można podzielić na wiele typów w zależności od ich wzajemnego umiejscowienia[1]. Pary genów nakładających się na tej samej nici mogą wykazywać nakładanie całkowite, gdy jeden gen w całości znajduje się wewnątrz dłuższego genu i nazywa się go genem zagnieżdżonym (rycina 1 A). Nakładanie genów na tych samych niciach DNA może, choć nie musi, wiązać się z wykorzystaniem innych ramek odczytu, jak zostało to zidentyfikowane przykładowo dla genu *E* bakteriofaga φX174, który zagnieżdżony jest wewnątrz genu *D*[4, 5]. Translacja białek z wykorzystaniem różnych ramek odczytu prowadzi w tym przypadku do kodowania przez oba geny zupełnie innych białek, natomiast mutacja prowadząca do dysfunkcji jednego z nich, niekoniecznie musi się łączyć z dysfunkcją produktów obu genów[4]. Geny na tej samej nici DNA mogą wykazywać również nakładanie częściowe (rycina 1 B), jak zostało to zademonstrowane na przykładzie pary ludzkich genów *BLT1* i *BLT2*, w której rejon 5' UTR genu *BLT1* nakłada się z rejonem kodującym i rejonem 3' UTR genu *BLT2*[22]. Nakładanie się może być zaobserwowane również dla genów znajdujących się na przeciwnych niciach DNA i także tutaj może mieć ono charakter nakładania całkowitego (rycina 1 C) lub częściowego (rycina 1 D-E), które dzieli się na dwa typy. Na przeciwnych niciach geny mogą się bowiem nakładać końcami 3' (rycina 1 D) lub końcami 5' (rycina 1 E), co zwyczajowo nazywa się nakładaniem ogon-do-ogona (ang. *tail-to-tail*) lub głowa-do-głowy (ang. *head-to-head*). Dla każdej z powyżej opisanych kombinacji wzajemnego ułożenia genów w parze można by dodatkowo opisać także pochodne typy w zależności od tego, czy geny nakładają się intronami czy egzonami, oraz czy nakładanie obejmuje rejon kodujący czy niekodujący[1]. Ponadto, geny

kodujące białka mogą tworzyć pary zarówno z innymi genami kodującymi białka, jak również z długimi niekodującymi RNA (lncRNA; z ang. *long non-coding RNA*). Transkrypty genów zlokalizowanych na różnych *loci* genomowych mogą także tworzyć tak zwane pary antysensowne, które jest nakładaniem się w konfiguracji *trans* (rycina 1 F), gdy antysensowny transkrypt znajduje się na innym *loci* genomowym niż gen kodujący białko. Nakładające się w ten sposób transkrypty nie muszą wykazywać pełnej komplementarności w rejonie nakładania, jak ma to miejsce w przypadku par genów nakładających się w konfiguracji *cis* (rycina 1 C-E).



Rycina 1. ***Sześć typów nakładania się genów****. A) nakładanie całkowite pary genów na tej samej nici DNA, przy czym jeden gen jest zagnieżdżony w drugim. B) nakładanie częściowe pary genów na tej samej nici DNA. C) nakładanie się genów na przeciwnych niciach DNA, przy czym jeden z genów jest zagnieżdżony w drugim. D) częściowe nakładanie się genów zlokalizowanych na przeciwnych niciach DNA, przy czym obszar nakładania znajduje się w rejonie 3' końców obu genów. E) częściowe nakładanie się genów zlokalizowanych na przeciwnych niciach DNA, przy czym obszar nakładania znajduje się w rejonie 5' końców obu genów. F) nakładanie się genów w konfiguracji* trans*, gdzie oba geny pochodzą z różnych* loci *genomowych pokolorowanych odpowiednio na zielono lub niebiesko. Pionowe linie reprezentują pełną komplementarność dla par genów nakładających się w konfiguracji cis (C-E), oraz częściową komplementarność dla pary genów nakładającej się w* trans *(F).*

Wraz z odkryciem lncRNA zainteresowanie nakładającymi się parami utworzonymi przez dwa geny kodujące białka bardzo znacząco zmalało. Niemniej jednak, w świetle dynamicznego rozwoju rozmaitych technik biologii molekularnej, ze szczególnym uwzględnieniem sekwencjonowania nowych generacji, pojawiła się doskonała okazja, aby wykorzystać najnowsze osiągnięcia technologiczne do dokładnego zbadania zjawiska

nakładania się genów w tym szczególnym przypadku, gdy jest ono tworzone przez dwa geny kodujące białka. W ramach niniejszej pracy skupiono się na parach genów kodujących białka, które nakładają się końcami 5' oraz funkcjonalnych implikacjach mogących z tego wynikać.

Metody identyfikacji nakładających się par genów

Minione lata obfitowały w opracowania różnych metod identyfikacji genów nakładających się. Każda z tych metod, w zależności od podejścia, prowadziła do identyfikacji innej liczby nakładających się par. Za relatywnie proste podejście można uznać identyfikację nakładających się par z wykorzystaniem programu BLAST, który użyty został do znalezienia komplementarnych sekwencji mRNA w ludzkim genomie [23]. Tym sposobem udało się określić nakładanie 61 par genów końcami 3', 20 par genów końcami 5' oraz cztery pary genów zagnieżdżonych. Podobne podejście oparte zostało dodatkowo, obok sekwencji mRNA, o analizę znaczników sekwencji ulegających ekspresji (EST; z ang. *expressed sequence tag*). Sekwencje mRNA i EST kodowane przez te same geny zostały połączone w jednostki transkrypcyjne między którymi wyszukiwano rejony komplementarne[14]. Podejście to zaowocowało identyfikacją 144 ludzkich i 73 mysich par genów nakładających się, z których część udało się potwierdzić eksperymentalnie z wykorzystaniem metody RT-PCR, co pozwoliło na oszacowanie 84% poprawności obliczeniowej identyfikacji rejonów nakładania. Inna metodologia oparta została o analizę koordynat ludzkich i mysich sekwencji genomowych z bazy GenBank, w efekcie czego zidentyfikowano łącznie 774 ludzkich i 314 mysich par genów nakładających[16]. Ponad połowa z tych par nakładała się końcami 3', podczas gdy najrzadszym typem nakładania było nakładanie całkowite poprzez zagnieżdżenie genów. Pośród tych par Veeramachaneni i współpracownicy zidentyfikowali również 225 par ludzkich genów, które posiadały geny ortologiczne u myszy, jednakże tylko 95 z nich ulegało nakładaniu u obu gatunków. Ponadto znacząca część tych nakładających się par genów ortologicznych posiadała odmienne wzorce nakładania[16]. Nieco wyższe zakonserwowanie zjawiska nakładania wykazano w ramach prac nad wielkoskalowym projektem FANTOM3, gdzie spośród 6141 par transkryptów antysensownych u człowieka, 16% wykazywało nakładanie się także u myszy[24]. Z kolei szczegółowa analiza ludzkich i mysich genów kodujących białka, nakładających się z długimi niekodującymi RNA, która przeprowadzona została przez Wood i współpracowników[25] wykazała, że prawie połowa zidentyfikowanych par u człowieka nie jest zakonserwowana u myszy, oraz że zjawisko nakładania się na końcach 5' jest relatywnie słabiej zachowane niż nakładanie na końcach 3'.

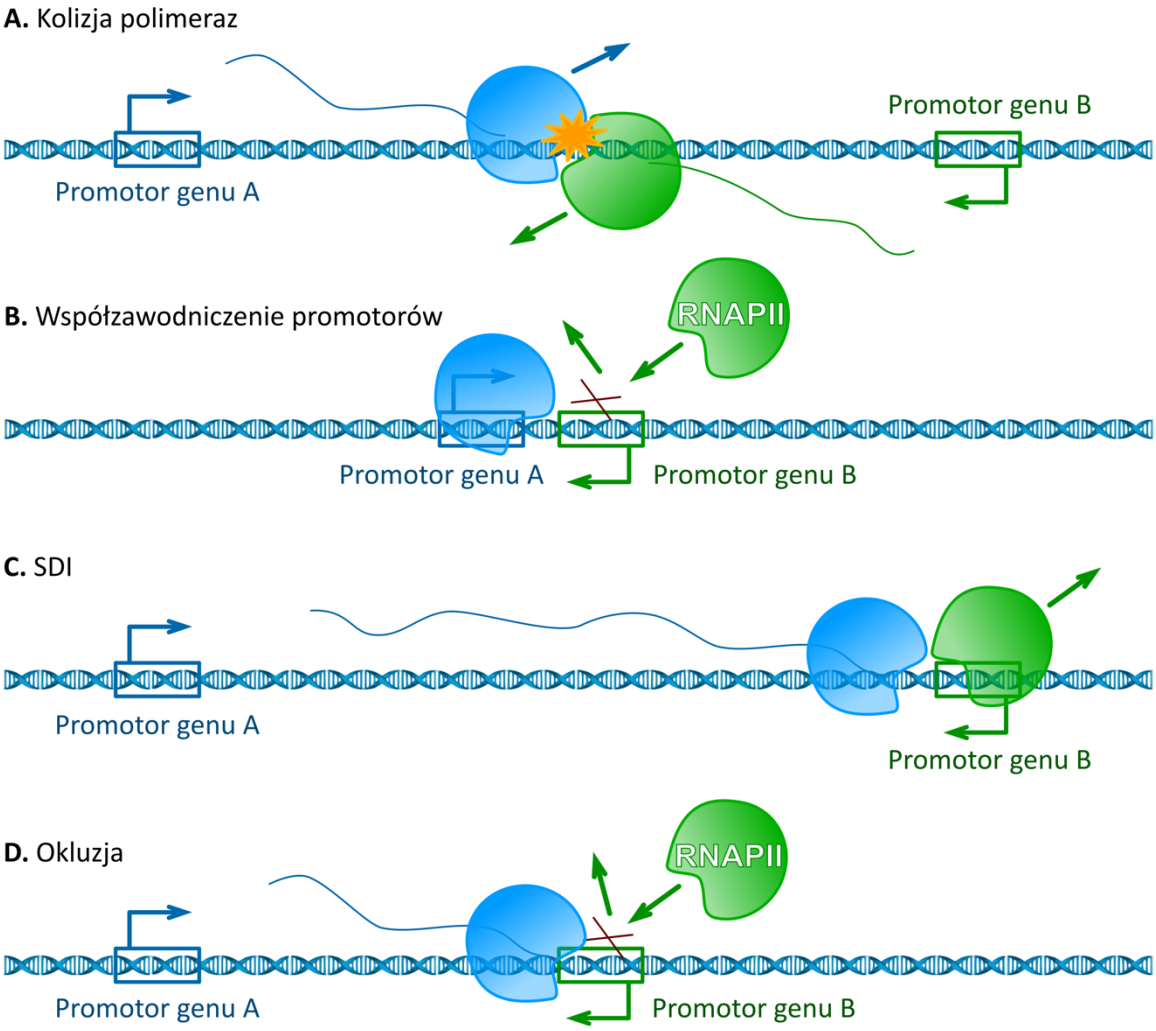
Bardzo dynamiczny rozwój technologiczny, który nastąpił w ostatnich latach, umożliwił identyfikację naturalnych transkryptów antysensownych z wykorzystaniem wielu metod, włączając w to technikę MPSS (z ang. *massively parallel signature sequencing*)[26, 27], ASSAGE (z ang. *asymmetric strand-specific analysis of gene expression*)[28], wysokoprzepustowe sekwencjonowanie całych transkryptomów (RNA-Seq)[29], ze szczególnym uwzględnieniem sekwencjonowania specyficznego dla nici DNA (ssRNA-Seq; z ang. *strand specific RNA sequencing*)[30-32] oraz wiele innych[29, 33, 34]. Conley i Jordan zidentyfikowali tysiące antysensownych miejsc startu trasnkrypcji u człowieka z wykorzystaniem danych z sekwencjonowania 5' końców transkryptów metodą CAGE (z ang. *cap analysis gene expression*)[34]. Ponadto, wykorzystali oni dane ChIP-Seq (z ang. *chromatin immunoprecipitation-sequencing*) z projektu ENCODE, celowane na zbadanie aktywności polimerazy RNA II oraz różnych modyfikacji histonów. Pozwoliło im to na określenie siły promotorów jako słabych, gdy brakowało im aktywności polimerazy a towarzyszyła trimetylacja H3K27 mająca charakter wyciszający, lub jako silnych, gdy znajdowano w nich aktywność polimerazy oraz acetylację dziewiątej lizyny histonu H3, która powiązana jest z aktywnymi promotorami[35]. Wykazali następnie statystycznie istotną nadreprezentację współwystępowania tego samego typu promotorów w antysensownych parach genów, czyli dwóch silnych lub dwóch słabych promotorów. Dodatkowo, analiza sześciu rodzajów tkanek zaowocowała określeniem tkankowo specyficznych wzorców nakładania, które mogą sugerować funkcjonalne znaczenie tego zjawiska[34]. Ling i współpracownicy[36], wykorzystując mikromacierze DNA również zidentyfikowali tysiące naturalnych transkryptów antysensownych wykazując ich tkankowo specyficzne nakładanie oraz międzygatunkowe zachowanie wzorców ekspresji niektórych z nich w dziewięciu tkankach człowieka, myszy i szczura. Wykazano również pozytywną korelację ekspresji transkryptów antysensownych w różnych tkankach, a część z wyników potwierdzona została eksperymentalnie z wykorzystaniem techniki RT-PCR[36].

Potencjalne funkcje nakładania się genów

Większość badań prowadzonych obecnie nad genami nakładającymi skupia się na parach genów kodujących białka, które współdzielą fragment sekwencji DNA z genem długiego niekodującego RNA znajdującym się na nici przeciwnej i tworzącym naturalny transkrypt antysensowny (NAT; z ang. *natural antisense transcript*). Pojawia się coraz więcej doniesień o funkcjonalnym znaczeniu lncRNA. Sugeruje się, że długa niekodująca cząsteczka RNA może odpowiadać za regulację poziomu ekspresji genu na nici przeciwnej[37].

Liczne badania wskazują, że NAT mogą pełnić funkcje zarówno na poziomie transkrypcyjnym, poprzez tak zwane zakłócenia transkrypcyjne, jak również na poziomie post-transkrypcyjnym, poprzez tworzenie dupleksu RNA w rejonie nakładania się transkryptów antysensownych lub interakcję RNA z DNA[2, 37, 38]. Mimo wszystko, znaczenie funkcjonalne zjawiska nakładania w skali globalnej jest wciąż tematem aktywnej dyskusji[34, 37, 39-42]. Naturalne transkrypty antysensowne powiązane zostały z różnego rodzaju chorobami włączając w to chorobę Parkinsona[43], Altzheimera[44], Huntingtona[45], nowotwory[46-49] i wieloma innymi[50-52]. W związku z tym wiele antysensownych lncRNA badanych jest w celach terapeutycznych, ponieważ poprzez regulację poziomu lncRNA można w wielu przypadkach wpływać na poziom ekspresji genu kodującego białko, który znajduje się na nici przeciwnej[53].

Jak wspomniano powyżej, transkrypty antysensowne mogą pełnić funkcje regulatorowe poprzez zakłócenia transkrypcyjne. Zakłócenia transkrypcyjne, zwane również interferencją transkrypcyjną (TI; z ang. *transcriptional interference*), są pośrednim lub bezpośrednim wpływem jednego aktywnie przebiegającego procesu transkrypcji na inny. Zakłócenia podzielić można na cztery podstawowe typy przedstawione na rycinie 2. Pierwszym typem jest kolizja polimeraz, która może wystąpić w przypadku napotkania się dwóch polimeraz RNA przesuwających się w przeciwnych kierunkach (rycina 2 A). Może to skutkować zablokowaniem obu polimeraz, które fizycznie nie mogą kontynuować transkrypcji. Obie z nich pozostają w takim stanie stabilne co wykazane zostało przez Hobsona i współpracowników przy badaniu połowicznego czasu rozpadu kompleksu transkrypcyjnego[54]. Niemniej jednak kolizja polimeraz może w końcu doprowadzić do przedwczesnego zakończenia transkrypcji przez jedną lub obie polimerazy, co zademonstrowano na przykładzie genów *gal7* i *gal10* u drożdży[55]. Innym rodzajem interferencji transkrypcyjnej jest współzawodniczenie nakładających się rejonów promotorowych o inicjację transkrypcji, która w danym momencie może zajść tylko na jednym z takich promotorów (rycina 2 B)[56]. Mechanizm ten może również dotyczyć „nakładających się" rejonów wzmacniających, gdzie jedna sekwencja wzmacniająca odpowiedzialna za regulację dwóch promotorów może wchodzić w interakcję tylko z jednym z nich w danym momencie[57, 58]. Istnieją również doniesienia o odwrotnej sytuacji, w której o jeden rejon promotorowy współzawodniczy wiele sekwencji wzmacniających, co wpływać może na ostateczny poziom aktywności tego promotora[59].



Rycina 2. ***Cztery scenariusze zjawiska interferencji transkrypcyjnej****. A) kolizja dwóch polimeraz RNA II (RNAPII) aktywnie transkrybujących geny w przeciwnych kierunkach tego samego loci. B) Współzawodniczenie blisko siebie usytuowanych promotorów o inicjację transkrypcji. C) Mechanizm SDI (z ang. sitting duck interference), w którym aktywnie transkrybująca polimeraza RNA II przyczynia się do oddysocjowania napotkanego kompleksu preinicjacyjnego genu na nici przeciwnej. D) Okluzja promotora genu przez inną polimerazę, której ekspresja zainicjowana została przez inny promotor. Rycina powstałą w oparciu o pracę przeglądową Shearwin i współpracownicy*[38].

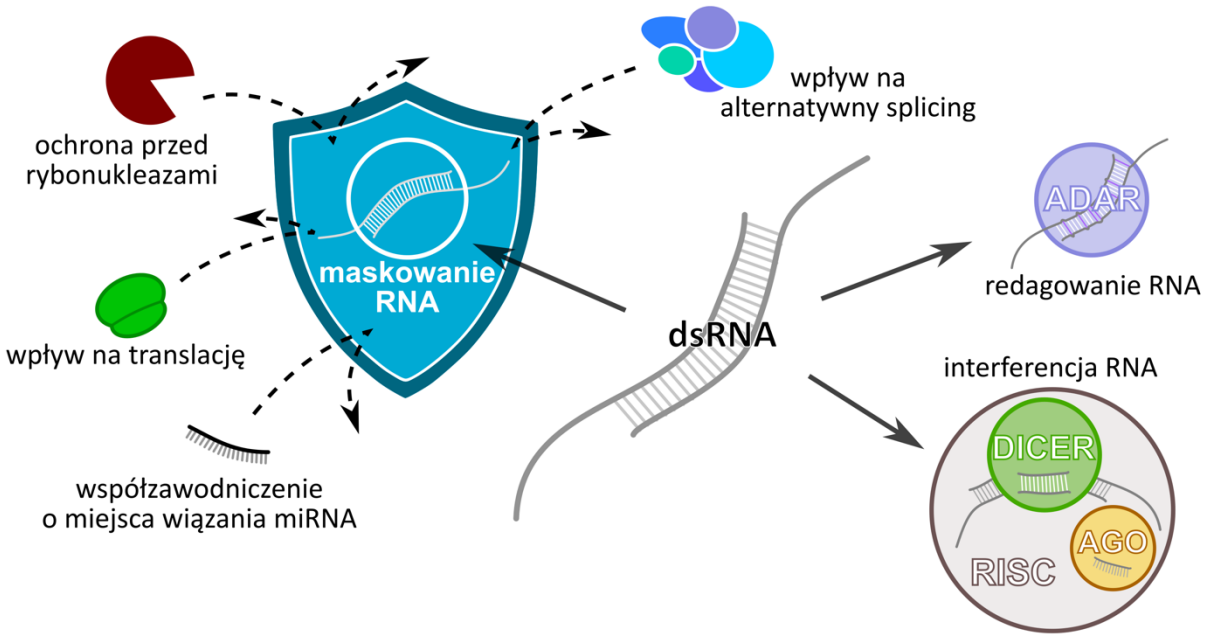
Interferencja transkrypcyjna może również występować w przypadku, gdy przejście polimerazy RNA II z fazy inicjacji transkrypcji do elongacji zabiera relatywnie dużo czasu. Gdy taki kompleks preinicjacyjny napotkany zostaje przez aktywnie transkrybującą polimerazę RNA genu na nici przeciwnej, może on zostać przez nią zdestabilizowany i rozproszony, przerywając przejście polimerazy w fazę elongacji (rycina 2 C). To samo może odnosić się do innych czynników transkrypcyjnych przyłączonych do DNA[60]. Ten mechanizm, zwany SDI (z ang. *sitting duck interference*), zademonstrowany został dla prokariotycznych promotorów których transkrypcja inicjowana w silnym promotorze

destabilizowała polimerazę RNA w słabszym rejonie promotorowym[61]. Podobnym mechanizmem interferencji jest okluzja rejonu promotorowego (rycina 2 D), w której elongacja jednego transkryptu blokuje dostęp do innego promotora, gdy przebiega ona w miejscu jego lokalizacji. Na przykładzie *E.coli* pokazano, że tempo elongacji polimerazy RNA jest na tyle szybkie, że okluzja nie powinna mieć tak negatywnego wpływu na inicjację transkrypcji w blokowanym promotorze jak inne mechanizmy zakłócenia transkrypcyjnego[38, 62]. Niemniej jednak w przypadku mysiego genu *FPGS* wykazano spowolnienie tempa elongacji polimerazy RNA przy jej przechodzeniu przez wyciszony okluzją rejon promotorowy[60, 63]. Również w przypadku retrotranspozonów L1 zagnieżdżonych w intronach innych genów, interferencja transkrypcji zachodząca między ich rejonami promotorowymi połączona została ze spowolnieniem tempa elongacji polimerazy RNA II[64]. Spowolnienie może jako odrębne zjawisko mieć wpływ na proces dojrzewania cząsteczki pre-mRNA poprzez relatywne wydłużenie czasu ekspozycji miejsc alternatywnego splicingu, umożliwiając nawet słabszym z nich być rozpoznanymi[65-69].

Transkrypty nakładających się na przeciwnych niciach DNA genów posiadają w rejonie nakładania całkowicie komplementarną sekwencję. Na poziomie interakcji RNA:RNA może to prowadzić do formowania się dwuniciowego RNA (dsRNA; z ang. *double stranded RNA*). Struktura dsRNA może pełnić rozmaite post-transkrypcyjne funkcje, które podsumowane zostały na rycinie 3. Utworzony dupleks RNA może maskować miejsca alternatywnego splicingu, przyczyniając się w ten sposób do alternatywnego składania cząsteczki pre-mRNA[70]. Proces taki może być niezbędny do translacji białka, jak zademonstrowano to dla ludzkiego genu *Zeb2*, w którym egzon zawierający miejsce wiązania rybosomu znajduje się tylko w jednym z wariantów splicingowych, powstającym przy tworzeniu dsRNA przez transkrypt genu *Zeb2* i jego naturalny transkrypt antysensowny[71]. Z drugiej strony utworzenie dwuniciowego RNA może w niektórych przypadkach być najprawdopodobniej uwikłane w zaburzenie translacji białka. Ebralidze i współpracownicy[72] wykazali przykładowo, że naturalny transkrypt antysensowny zaburza łączenie się czynnika elongacyjnego do mRNA ludzkiego i mysiego genu *PU.1*. Maskowanie RNA może ponadto stabilizować nakładające się transkrypty dzięki blokowaniu dostępu do destabilizujących motywów znajdujących się w transkrypcie kodującym białko[73] lub uniemożliwianiu wiązania się mikro RNA (miRNA; z ang. *micro RNA*) do sekwencji docelowej znajdującej się w rejonie nakładania[74], wydłużając czas połowicznego rozpadu transkryptów antysensownych[75]. Dupleks RNA może również zwiększać stabilność cząsteczki mRNA ochraniając ją przed działaniem rybonukleaz trawiących jednoniciowe

RNA[76]. Liczne badania wykazały również, na przykładach ze świata roślin, zwierząt i grzybów, że dsRNA może stać się źródłem cząsteczek endo-siRNA (z ang. *endogenous small interfereing RNA*) prowadzących do wyciszenia ekspresji genów poprzez zjawisko interferencji RNA[77-85]. Prawdziwą skalę występowania tego zjawiska odkryto dopiero dzięki zastosowaniu technologii sekwencjonowania nowych generacji, które u *Arabidopsis thaliana* pozwoliły na wykazanie, że nawet 4% genów kodujących białka, nakładających się z lncRNA, może być źródłem endo-siRNA[30]. Naturalne transkrypty antysensowne mogą być również źródłem tak zwanych disiRNA (z ang. *dicer independent small interfereing RNA*), które znalezione zostały u *Neurospora crassa*[82] oraz zidentyfikowanych u myszy małych interferujących RNA przypominających piRNA (z ang. *piwi-interacting RNA*)[86], które w odróżnieniu od endo-siRNA nie są zależne od kompleksu RISC (z ang. *RNA-induced silencing complex*). Formowanie się dwuniciowego RNA może również umożliwiać redagowanie sekwencji RNA przez kompleks ADAR (z ang. *adenosine deaminase acting on RNA*), co zademonstrowane zostało przez Petersa i współpracowników dla rejonu nakładania genów *4f-rnp* i *sas-10* u muszki owocowej[87].

Transkrypty antysensowne mogą także post-transkrypcyjnie regulować poziom ekspresji genów kodujących białka poprzez odziaływania RNA – DNA. Oddziaływania te mogą dotyczyć inicjowania i utrzymywania metylacji lub demetylacji rejonów promotorowych i sekwencji wzmacniających genów kodujących białka[49, 88], *loci* genomowych a nawet całych chromosomów[53, 89, 90]. Naturalne transkrypty antysensowne są przykładowo zaangażowane w inaktywację chromosomu X u samic ssaków, gdzie promotor genu *Xist* wyciszony jest przez utrzymywanie w tym rejonie trimetylacji 27 i 36 lizyny histonu H3 przez antysensowny *Tsix*. Aktywacja genu *Xist*, spowodowana wyciszeniem ekspresji *Tsix* na jednym z chromosomów X, prowadzi do konsekwentnej inaktywacji tej kopii chromosomu[91, 92]. Innym przykładem może być regulacja ekspresji ludzkiego genu *BDNF*, którego wyciszenie związanie jest z wprowadzeniem do całego *locus* genomowego w którym się on znajduje trimetylacji 27 lizyny histonu H3 przez kompleks PCR2 (z ang. *polycomb repressive complex 2*)[93]. Aktywność kompleksu PCR2 inicjowana jest z kolei przez naturalny transkrypt antysensowny genu *BDNF*.

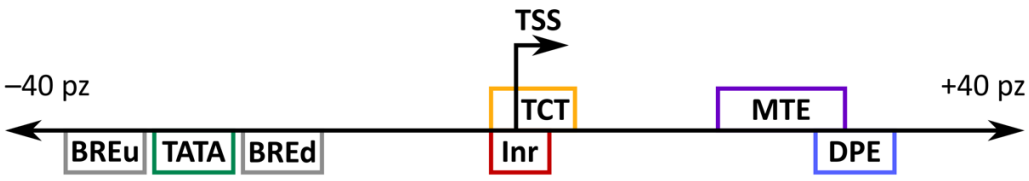


Rycina 3. ***Podsumowanie potencjalnych funkcji pełnionych przez formowanie się dwuniciowego RNA (dsRNA).*** *Obejmują one maskowanie RNA, mogące wpływać na alternatywny splicing i translację, oraz ochraniać transkrypty przed działaniem rybonukleaz trawiących jednoniciowe RNA i współzawodniczyć o miejsca wiązania miRNA znajdujące się w rejonie nakładania. Struktura dsRNA może być ponadto rozpoznana przez kompleks ADAR (z ang. adenosine deaminase acting on RNA) prowadząc do redagowania sekwencji RNA, lub przez kompleks RISC (z ang. RNA-induced silencing complex), prowadząc do interferencji RNA. W skład kompleksu RISC wchodzi białko z rodziny Dicer, przycinające sekwencję dsRNA, oraz białko z rodziny Argonaute (AGO), które pełni kluczową funkcję w wyciszaniu ekspresji genów docelowych przez kompleks RISC.*

Pomimo wielu lat badań poświęconych różnym rodzajom nakładających się genów, wciąż brakuje ostatecznej odpowiedzi na to, czy taki sposób organizacji genomowej wyewoluował jako kolejny poziom regulacji ekspresji genów, czy też jest to produkt uboczny architektury genomowej, który w sposób przypadkowy zyskał funkcjonalne znaczenie dla niektórych genów. Wiele badań wykazało tkankowo specyficzny charakter ekspresji tego typu genów[31, 34, 36], co nie może jednak samo w sobie być ostatecznym dowodem na funkcjonalność zjawiska nakładania[39, 94]. W niniejszej pracy przeanalizowano specyficzną grupę nakładających się genów, w której oba geny z pary są genami kodującymi białka. Skupiono się ponadto na genach zwróconych w stosunku do siebie końcami 5', których ekspresja może potencjalnie być inicjowana przez promotory dwukierunkowe[95]. Z tego względu w dalszej części rozdziału omówiono pokrótce architekturę promotorów, ich rodzaje oraz sposoby regulacji, które mogą przyczyniać się do scharakteryzowania promotora jako jedno lub dwukierunkowy.

Architektura promotorów

Promotorem określa się zestaw wszystkich sekwencji istotnych dla inicjacji transkrypcji genów. Aktywność różnych rodzajów polimerazy RNA może być związana z różnymi rodzajami promotorów, jednakże w kontekście niniejszej pracy doktorskiej skupiono się na promotorach polimerazy RNA II, która odpowiada za transkrypcję genów kodujących białka oraz niektórych długich niekodujących RNA[96]. Większość promotorów złożona jest z przynajmniej kilku elementów promotora podstawowego (ang. *core promoter elements*), które stanowią podstawowy zestaw sekwencji niezbędnych do prawidłowego zainicjowania transkrypcji. Motywy sekwencji promotora podstawowego rozpoznawane są przez ogólne czynniki transkrypcyjne (ang. *general transcription factors*), do których zaliczyć można składające się z wielu podjednostek kompleksy *TFIIA, TFIIB, TFIID, TFIIE, TFIIF* oraz *TFIIH,* które uczestniczą w formowaniu się kompleksu preinicjacyjnego (PIC; z ang. *preinitiation complex*)[97, 98]. Ogólne czynniki transkrypcyjne, wraz z powiązanymi czynnikami transkrypcyjnymi umożliwiają eukariotycznej polimerazie RNA II rozpoczęcie transkrypcji. Wydajność transkrypcji z promotora podstawowego bez oddziaływania z dodatkowymi czynnikami transkrypcyjnymi może być stosunkowo niska. Do promotora podstawowego zaliczyć można kilka elementów, które znajdują się na obszarze obejmującym mniej więcej do 40 nukleotydów powyżej oraz poniżej miejsca startu transkrypcji (rycina 4)[99].



Rycina 4. ***Wybrane elementy promotora podstawowego****. Opis oznaczeń zgodnie z kolejnością alfabetyczną: BREd – z ang. TFIIB recognition element (BRE)-downstream; BREu – z ang. TFIIB recognition element (BRE)-upstream; DPE – z ang. downstream promoter element; Inr – sekwencja inicjatorową; MTE – z ang. motif ten element; TATA – sekwencja TATA; TSS – z ang. Transcription start site. Rycina powstała w oparciu o pracę przeglądową Kadonaga*[100].

W samym centrum promotora podstawowego znajduje się sekwencja inicjatorowa (Inr), której sekwencja konsensusowa, zgodnie z systemem IUPAC (z ang. *international union of pure and applied chemistry*) przyjmuje u człowieka i innych ssaków postać 5'- YYANWYY-3' [101]. Trzeci nukleotyd tej sekwencji najwyższej zgodności, licząc od końca 5', jest miejscem inicjacji startu transkrypcji (TSS; z ang. *transcription start site*) i w stosunku do niego liczona jest odległość innych elementów promotora podstawowego.

Sekwencja inicjatorowa często występuje w promotorze z tak zwaną sekwencją TATA, która odkryta została na przełomie lat siedemdziesiątych i osiemdziesiątych[102, 103]. Nazwa sekwencji TATA wywodzi się od jej sekwencji najwyższej zgodności 5'-TATAWAAR-3', która identyfikowana jest w mniej niż 15% promotorów u ssaków[104, 105]. Sekwencja TATA rozpoznawana jest przez białko *TBP* (z ang. *TATA-binding protein*), która jest podjednostką kompleksu *TFIID*[106]. Po obu stronach sekwencji TATA znajdować się mogą dodatkowe sekwencje regulatorowe, które rozpoznawane są przez podjednostkę *TFIIB* i prawdopodobnie odgrywają one rolę w pozycjonowaniu polimerazy RNA II w odpowiednim miejscu startu transkrypcji[107]. Sekwencje te znane są pod ogólną nazwą BRE (z ang. *TFIIB recognition element*) i w zależności od tego czy znajdują się na 5' czy na 3' końcu w stosunku do sekwencji TATA, noszą one odpowiednio nazwy BREu (z ang. *BRE-upstream*) lub BREd (z ang. *BRE-downstream*)[108-111]. Kolejnym elementem promotora podstawowego z którym często występuje sekwencja inicjatorowa jest sekwencja DPE (z ang. *downstream promoter element*), która po raz pierwszy zidentyfikowana została u muszki owocowej[112]. Znajduje się ona od 28 do 32 nukleotydów poniżej sekwencji inicjatorowej[113] i jest ona przypuszczalnie miejscem wiązania podjednostek *TAF6* oraz *TAF9* kompleksu *TFIID*[97, 114-116]. DPE wchodzi w skład elementów promotora podstawowego prawie wszystkich genów homeotycznych u muszki owocowej[117]. Co ciekawe, sekwencje DPE oraz TATA są bardzo rzadko znajdywane w jednym promotorze, jednakże występowanie ich obu związane jest z sekwencją Inr[112, 113, 118]. Sekwencja DPE nakłada się również z powyżej położonym na sekwencji DNA elementem MTE (z ang. *motif ten element*), który może być regulowany przez te same podjednostki *TAF6* i *TAF9*, które rozpoznają sekwencję DPE[114-116]. Wysoki stopień nakładania się motywów MTE oraz DPE, a co za tym idzie również podobieństwa sekwencyjnego powoduje, że pełnią one najprawdopodobniej podobne funkcje, jednakże oba z tych motywów są od siebie niezależne. Badania wykazały, że element MTE może skompensować inaktywację sekwencji TATA oraz DPE i wraz z sekwencją inicjatorową prowadzić do inicjacji transkrypcji[119]. Ostatni omawiany element promotora podstawowego nosi nazwę TCT, która pochodzi od często występującej w jego sekwencji najwyższej zgodności trójki nukleotydów „TCT". Element ten nakłada się z sekwencją inicjatorową, lecz nie jest on jej substytutem[120]. Badania przeprowadzone u człowieka i muszki owocowej wykazały, że funkcjonalność elementu TCT jest najprawdopodobniej związana z regulacją translacji mRNA i występuje on w prawie wszystkich promotorach białek rybosomalnych oraz innych białkach związanych z translacją[120].

Rodzaje promotorów

Funkcja elementów promotora podstawowego polega na umożliwieniu precyzyjnej regulacji transkrypcji[121, 122]. W poprzednim paragrafie omówiono elementy stanowiące zestaw bloków budulcowych promotorów podstawowych, których kompozycja ma wpływ na charakter promotora. Biorąc pod uwagę budowę, promotory dzieli się na dwie główne kategorie. Pierwszą z nich są promotory skupione (ang. *focused*), których miejsce startu transkrypcji jest precyzyjnie wyznaczone sekwencją inicjatorową i której często towarzyszy sekwencja TATA, DPE lub MTE[123, 124]. Druga kategoria obejmuje promotory rozproszone (ang. *dispersed*), w których wiele miejsc TSS o relatywnie niższej ekspresji znajduje się na obszarze o długości do 100 nukleotydów. Tego typu promotory po raz pierwszy zidentyfikowane zostały wśród genów metabolizmu podstawowego (ang. *housekeeping genes*)[125]. Wśród promotorów tego typu nie znajduje się najczęściej sekwencji TATA, a pozycjonowanie polimerazy RNA II może być regulowane takimi czynnikami transkrypcyjnymi jak *Sp1* czy *NF-Y*[126-128]. Również czynniki epigenetyczne jak pozycjonowanie nukleosomów może mieć wpływ na przynależność promotora do pierwszej lub drugiej kategorii. Promotory rozproszone częściej niż promotory skupione skorelowane są z umiejscowieniem startu transkrypcji w rejonie wolnym od nukleosomów (NFR; z ang. *nucleosome free region*)[129, 130]. Rejon ten z każdej ze stron okalany jest przez silnie pozycjonowane nukleosomy (ang. *well positioned nucleosome*) które zwyczajowo nazywa się nukleosomem +1, dla nukleosomu położonego poniżej rejonu NFR, oraz nukleosomem -1, dla nukleosomu po stronie przeciwnej[131-133]. Obecność rejonu wolnego od nukleosomów może konsekwentnie wpłynąć na rozpoznawanie motywów DNA w nim zawartych przez czynniki transkrypcyjne nie należące do grupy podstawowych czynników inicjujących transkrypcję. Zatem promotor typu rozproszonego, nieposiadający większej części elementów promotora podstawowego, może zostać efektywnie aktywowany za pomocą innych czynników transkrypcyjnych.

Mimo iż promotory podzielić można na opisane powyżej dwie kategorie zależne od precyzji miejsca inicjacji transkrypcji, w praktyce granice te są bardzo płynne i często obserwuje się kilka blisko siebie usytuowanych miejsc inicjacji transkrypcji, pośród których dominuje jedno z tych miejsc[129, 134]. Niezależnie od typu promotora, na jego aktywność mają również wpływ inne czynniki epigenetyczne do których zaliczyć można zróżnicowanie wariantów histonów budujących nukleosom w rejonie startu transkrypcji. Przykładowo wbudowanie w nukleosomy w rejonie promotorowym białek histonowych H2A.Z,

szczególnie w połączeniu z jednoczesnym wbudowaniem wariantu H3.3, prowadzić może do destabilizacji rdzenia nukleosomu stymulując inicjację i elongację transkrypcji[132, 133, 135-139]. Również modyfikacje post-transkrypcyjne *N* końców białek histonowych w nukleosomach znajdujących się w promotorze mogą mieć wpływ na formowanie się kompleksu inicjującego transkrypcję. Modyfikacje takie mogą obejmować acetylację, metylację, fosforylację oraz ubikwitynację, z czego pierwsze dwa rodzaje modyfikacji są na dzień dzisiejszy najdokładniej zbadane[140]. Modyfikacje te mogą być w bardzo dynamiczny sposób dodawane lub usuwane z poszczególnych histonów za pomocą odpowiednich enzymów, np. acetylotransferaz (HAT; z ang. *histone acetylotransferaze*)[141] lub deacetylaz (HDAC; z ang. *histone deacetylase*)[141], które odpowiednio dodają lub usuwają grupy acetylowe z ogonów histonowych. Wspomnianą acetylację histonu łączy się ze zwiększoną aktywnością transkrypcyjną. W kontekście inicjacji transkrypcji acetylacja wariantu histonu H2A.Z w nukleosomie +1 połączona została ze zwiększonym poziomem ekspresji genu[142], podczas gdy acetylacja 27 lizyny histonu H3 (H3K27ac) skorelowana została z aktywnymi rejonami promotorowymi oraz elementami regulatorowymi[143, 144]. Metylacja również może pełnić funkcje aktywujące. Przykładowo trimetylacja 4 lizyny histonu H3 (H3K4me3) powiązana została z aktywnymi rejonami promotorowymi[132, 145]. Badania wykazały, że kompleks *TFIID* może się, poprzez specjalną domenę białka *TAF3*, wiązać bezpośrednio do histonu o takiej właśnie modyfikacji, co może mieć szczególnie ważną funkcję w przypadku inicjacji transkrypcji w promotorach pozbawionych pewnych elementów podstawowych[146]. Z drugiej strony metylacja może prowadzić do ścisłego upakowania chromatyny, prowadząc do wyciszenia danego rejonu w genomie[145, 147, 148].

Ponieważ każdy z nukleosomów składa się z ośmiu białek histonowych, a każde z nich może podlegać różnym post-transkrypcyjnym modyfikacjom, istnieje wiele kombinacji modyfikacji, które mogą wspólnie występować i odzwierciedlać pewien stan chromatyny. Przykładowo wspólne występowanie modyfikacji H3K27me3 oraz H3K4me3 skorelowane zostało z wyciszonymi rejonami promotorowymi[147, 148]. Skorelowanie różnego rodzaju modyfikacji z funkcjonalnymi elementami obserwowanymi w genomach zainspirowało badaczy do stworzenia koncepcji tak zwanego kodu histonowego[149]. Błyskawiczny rozwój rozmaitych technik opartych o sekwencjonowanie nowych generacji umożliwił opracowanie wielu metod pozwalających na globalne analizy rozmaitych modyfikacji histonów w genomach. Obecnie możliwe jest zbadanie występowania wielu rodzajów modyfikacji histonów np. za pomocą techniki ChIP-Seq. Aby ułatwić jednoczesną interpretację wyników z różnych eksperymentów ChiP-Seq, opracowano narzędzia bioinformatyczne oparte o ukryte

modele markowa (HMM; z ang. *hidden marcov models*) oraz nauczanie maszynowe. Przykładami takich narzędzi są ChromHMM[150], oraz jego zmodyfikowana wersja o nazwie Spectacle[151], które dzielą genom na nienakładające się rejony wykazujące taki sam wzorzec post-transkrypcyjnych modyfikacji histonów. Rejony takie nazywane są stanami chromatyny (ang. *chromatin states*) i pozwalają na bardzo efektywną analizę dynamicznie zmieniającego się kodu histonowego[150, 152]. Pamiętać jednak należy, że chociaż wiele badań wykazało wysoki stopień skorelowania między różnymi rodzajami modyfikacji histonów a funkcjonalnymi elementami w genomie, to jednak nie zawsze wszystkie oczekiwane modyfikacje muszą występować. Niedawne badania przeprowadzone dla *D.melanogaster* oraz *C.elegans* wykazały przykładowo, że ekspresja bez znanych „aktywujących" modyfikacji histonów jest możliwa i występuje wśród niektórych genów aktywnych w trakcie rozwoju tych organizmów[153]. Autorzy zasugerowali, że geny o stabilnej ekspresji posiadają silne sygnały modyfikacji histonów, jednakże geny wymagające nagłej aktywacji bądź inaktywacji mogą nie posiadać w chromatynie odpowiednich modyfikacji post-transkrypcyjnych, gdyż ich ekspresja związana jest głównie z czynnikami transkrypcyjnymi przez krótki czas oddziaływującymi z DNA.

Promotory dwukierunkowe

Promotory, w zależności od swojej budowy, czynników transkrypcyjnych oraz modyfikacji epigenetycznych, mogą inicjować transkrypcję w jednym lub dwóch kierunkach i często wykazują pod tym względem tkankową specyficzność[154]. Wiele badań sugeruje, że w przypadku większości promotorów transkrypcja może być inicjowana jednocześnie w dwóch kierunkach, jednakże transkrypty tworzone w jednym z nich stanowią stabilne, kodujące lub niekodujące RNA, podczas gdy transkrypty inicjowane w przeciwnym kierunku tworzą niestabilne i szybko degradowane RNA o nazwie PROMPT (z ang. *promoter-upstream transcript*)[155-159]. Istnieją również promotory dwukierunkowe, które w obu kierunkach inicjują stabilne transkrypty. Promotorom tym poświęcono dalszą część niniejszego rozdziału.

Trinklein i współpracownicy[160] wykazali że 1352 pary ludzkich genów kodujących białka regulowane są przez promotory dwukierunkowe, z których 315 par nakłada się na końcach 5' a pozostałe 1037 ulega ekspresji bez nakładania. Geny te oddalone są od siebie o mniej niż 1000 par zasad, lecz odległość dla większości z nich nie przekracza 300 nukleotydów[160]. Taki dystans między genami w parze może być po części tłumaczony modelem, w którym inicjacja transkrypcji w promotorze dwukierunkowym następuje na obu

końcach rejonu NFR[161-163]. Badania przeprowadzone przez Trinklein i współpracowników wykazały jednak, że dystans pomiędzy genami regulowanymi przez promotory dwukierunkowe może w niektórych przypadkach sięgać więcej niż 2000 nukleotydów[160]. Ogólna charakterystyka promotorów dwukierunkowych u myszy wykazała, że w stosunku do promotorów jednokierunkowych są one regulowane przez więcej czynników transkrypcyjnych, mają dwa osobne kompleksy pre-inicjacyjne oraz geny przez nie regulowane mają średnio wyższy poziom ekspresji[163]. Geny regulowane przez promotory dwukierunkowe mają również ogólną tendencję do ko-ekspresji a ich poziom ekspresji jest przeważnie pozytywnie skorelowany. Reguła ta nie sprawdziła się jednak w przypadku 11% par genów zidentyfikowanych przez Trinkleina i współpracowników[160], które wykazywały negatywną korelację ekspresji. Dwukierunkowe rejony promotorowe znacznie częściej niż promotory jednokierunkowe znajdują się wewnątrz wysp CpG, czyli rejonów DNA o podwyższonym średnim występowaniu par C+G przy jednoczesnym obniżeniu metylacji DNA[164]. Wewnątrz wysp CpG u człowieka znajduje się około 90% promotorów dwukierunkowych oraz jedynie 45% promotorów jednokierunkowych[165]. Promotory dwukierunkowe mogą się do pewnego stopnia różnić od jednokierunkowych również kompozycją elementów promotora podstawowego. Przykładowo sekwencja TATA znaleziona została przez Yang i Elnitski[165] tylko dla 9% genów regulowanych przez promotory dwukierunkowe i 29% genów regulowanych przez promotory jednokierunkowe. Z kolei Park i współpracownicy[166] zasugerowali model regulacji inicjacji transkrypcji, w którym sekwencja TATA stanowi element promotora podstawowego wpływającego na ustalenie kierunku transkrypcji. Model taki tłumaczyłby, dlaczego wewnątrz wysp CpG posiadających mniej sekwencji typu TATA znajduje się więcej promotorów dwukierunkowych.

Na kierunkowy charakter promotora mogą także wpływać czynniki transkrypcyjne[167]. Przykładowo, czynnik transkrypcyjny *GABPA* odpowiada za regulację więcej niż 80% promotorów dwukierunkowych, a jego przyłączenie do promotora jednokierunkowego prowadziło do dwukierunkowej ekspresji aż 67% z testowanych promotorów jednokierunkowych[168]. Badania oparte zarówno o eksperymenty ChIP-Seq jak również predykcje *in silico* wykazały w promotorach dwukierunkowych nadreprezentację miejsc wiązania wielu innych czynników transkrypcyjnych, włączając w to białka *E2F1, E2F4, MYC, NF-Y, NRF-1, SP1, SP3, STAT1* oraz *YY1*[169, 170]. Z drugiej strony wykazano, że obecność miejsc wiązania 73% ludzkich czynników transkrypcyjnych jest niedoreprezentowana wewnątrz dwukierunkowych rejonów promotorowych[170].

Wiązanie niektórych z tych czynników transkrypcyjnych może prowadzić do sytuacji, w której promotor dwukierunkowy inicjuje transkrypcję jednokierunkowo. Przykładami takich czynników są *Creb/ATF, Klf/Sp, NFYA* lub *Zfp161*[171]. Oznacza to, że klasyczne promotory dwukierunkowe posiadają pewien zestaw charakterystycznych dla siebie czynników transkrypcyjnych, które mogą odpowiadać w sposób szczególny za regulację inicjacji ich transkrypcji w dwóch kierunkach.

2. Cel pracy

Większość badań prowadzonych w ostatnich latach skupiona była nad genami kodującymi białka, nakładającymi się z lncRNA. Znacznie mniej wiadomo jednak o parach genów złożonych z dwóch genów kodujących białka. Wyjątkowa architektura genomowa otwiera tutaj możliwość regulacji poziomu ekspresji genów nakładających się np. przez kolizję polimeraz, która może przyczyniać się do obniżenia ekspresji genów nakładających[38]. Z drugiej strony dla par tego typu funkcjonalność zjawiska nakładania mogła mieć drugorzędne znaczenie, gdyż podstawową funkcją obu genów w parze jest kodowanie białka. **Celem niniejszej pracy była weryfikacja skali zjawiska nakładania się genów kodujących białka, skupiając się na genach nakładających się końcami 5' u człowieka i myszy oraz oszacowanie wpływu zjawiska nakładania na poziom ekspresji genów nakładających się, ze szczególnym uwzględnieniem roli jaką odgrywać mogą zakłócenia transkrypcyjne.**

3. Materiały

3.1. Adnotacje referencyjne

Dane adnotacji genomów referencyjnych w wersji GRCh38/hg38 dla człowieka oraz NCBI37/mm9 dla myszy pobrano z wykorzystaniem narzędzia *Table Browser* z bazy danych UCSC[172]. Dla człowieka pobrano dodatkowo adnotacje referencyjne z projektu GENCODE (wersja 24)[173], które wykorzystane zostały podczas analiz wyników pochodzących z wysokoprzepustowego sekwencjonowania transkryptomów.

3.2. Alternatywne miejsca startu transkrypcji

Koordynaty miejsc alternatywnego startu transkrypcji, określone metodą TSS-Seq, pobrane zostały z serwera FTP bazy DBTSS w wersji 9 dla człowieka[174] oraz wersji 8 dla myszy[175]. U człowieka były one dostępne łącznie dla 73 bibliotek (tabela 1), w tym 19 bibliotek pochodzących z organów zdrowego dorosłego człowieka, 5 z organów płodowych, 23 z linii komórkowych powstałych w wyniku hodowli 7 typów komórek w różnych warunkach laboratoryjnych oraz 26 bibliotek TSS-Seq pochodzących z próbek tkanki gruczolakoraka płuc pobranych od różnych pacjentów. Dane pobrane dla myszy obejmowały 4 próbki embrionalne otrzymane kolejno po 7, 11, 15 oraz 17 dniach rozwoju (tabela 1). Dodatkowo, w ramach pracy analizowano dla myszy również koordynaty miejsc TSS w sześciu organach (tabela 1). Koordynaty te określone zostały przez twórców bazy DBTSS z wykorzystaniem tego samego protokołu co pozostałe dane dla człowieka i myszy. Zostały one przygotowane przez zespół profesora Yutaka Suzuki z Uniwersytetu w Tokio na potrzeby niniejszej pracy doktorskiej.

Tabela 1. ***Lista 73 ludzkich i 10 mysich bibliotek TSS-Seq.***

Człowiek *(Homo sapiens)*			
Typ biblioteki	Oznaczenie biblioteki TSS-Seq	Warunki hodowli komórkowej	Odnośnik literaturowy
Zdrowe organy i tkanki	Tkanka tłuszczowa	RNA wykorzystane przez twórców bazy DBTSS do przeprowadzenia eksperymentu TSS-Seq zakupione zostało z firmy Clontech i Ambion.	174, 176
	Nadnercza		
	Mózg 1		
	Mózg 2		
	Mózg 3		
	Pierś		
	Jelito		
	Serce 1		
	Serce 2		

	Nerki 1		
	Nerki 2		
	Wątroba		
	Płuca		
	Limfa		
	Mięśnie		
	Jajniki		
	Prostata		
	Jądra		
	Tarczyca		
	Mózg płodu		
	Serce płodu		
	Nerki płodu		
	Wątroba płodu		
	Grasica płodu		
DLD-1, komórki rakowe jelita grubego	DLD1 Hipoksja HIF1-	Sześć rodzajów warunków uwzględniających niski/normalny poziom tlenu oraz transfekcję celowaną (HIF-1α, Hif2α) i niecelowaną	177
	DLD1 Hipoksja HIF2-		
	DLD1 Hipoksja		
	DLD1 Normoksja HIF1-		
	DLD1 Normoksja HIF2-		
	DLD1 Normoksja		
Beas-2B, nabłonkowe komórki płucne	Beas2B IL4+ STAT6	Osiem rodzajów warunków związanych z ekspresją/nadekspresją czynnika transkrypcyjnego STAT6, dobowej stymulacji interleukinami 4 oraz transfekcji celowanej na wyciszenie czynnika transkrypcyjnego STAT6 lub transfekcji niecelowanej	178
	Beas2B IL4- STAT6		
	Beas2B IL4+ Parent		
	Beas2B IL4- Parent		
	Beas2B IL4- STAT6 siRNA		
	Beas2B IL4- Kontrola siRNA		
	Beas2B IL4+ STAT6 siRNA		
	Beas2B IL4+ Kontrola siRNA		
Ramos, komórki chłoniaka Burkitta	Ramos IL4+	Stymulacja interleukinami 4 w przypadku jednej z linii komórkowych	174
	Ramos IL4-		
MCF-7, nabłonkowe komórki raka piersi	MCF7 Normoksja	Obniżone stężenie tlenu (1%) w przypadku jednej z linii komórkowych	179
	MCF7 Hipoksja		
TIG-3, embrionalne fibroblasty płucne	TIG3 Normoksja	Obniżone stężenie tlenu (1%) w przypadku jednej z linii komórkowych	179
	TIG3 Hipoksja		

HEK-293, embrionalne komórki nabłonkowe nerki	HEK293 Normoksja	Obniżone stężenie tlenu (1%) w przypadku jednej z linii komórkowych	179
	HEK293 Hipoksja		
Komórki HeLa	Hela	Standardowe warunki	174
Komórki gruczolakoraka płuc	A427	Większość linii komórkowych hodowano w standardowych warunkach na pożywce RPMI 1640 firmy Nissui. Linie komórkowe LC2ad, PC3, H1648 i H2347 hodowane były na pożywce kolagenowej firmy IWAKI (ang. collagen Type I-coated dishes)	180
	A549		
	ABC1		
	H1299		
	H1437		
	H1648		
	H1650		
	H1703		
	H1819		
	H1975		
	H2126		
	H2228		
	H2347		
	H322		
	II18		
	LC2ad		
	PC3		
	PC7		
	PC9		
	PC14		
	RERFLCad1		
	RERFLCad2		
	RERFLCKJ		
	RERFLCMS		
	RERFLCOK		
	VMRC-LCD		
Mysz *(Mus musculus)*			
Zdrowe organy	Mózg	RNA wykorzystane przez twórców bazy DBTSS do przeprowadzenia eksperymentu TSS-Seq zakupione zostały z firmy Ambion	dane nieopublikowane
	Serce		
	Nerki		
	Wątroba		
	Śledziona		
	Grasica		
Komórki embrionalne	Zarodek, 7 dni	Standardowe warunki hodowlane	181
	Zarodek, 11 dni		
	Zarodek, 15 dni		
	Zarodek, 17 dni		

3.3. Dane RNA-Seq

Sekwencjonowanie całych transkryptomów przeprowadzone zostało przez twórców bazy DBTSS dla tych samych 26 linii komórkowych gruczolakoraka płuc, które uprzednio przebadane zostały z wykorzystaniem protokołu sekwencjonowania TSS-Seq. Surowe sparowane odczyty powstałe z wykorzystaniem sekwenatora HiSeq 2500 firmy Illumina, zostały pobrane z bazy ENA[182], gdzie zachowane są pod numerem akcesyjnym projektu *PRJDB2256*[180].

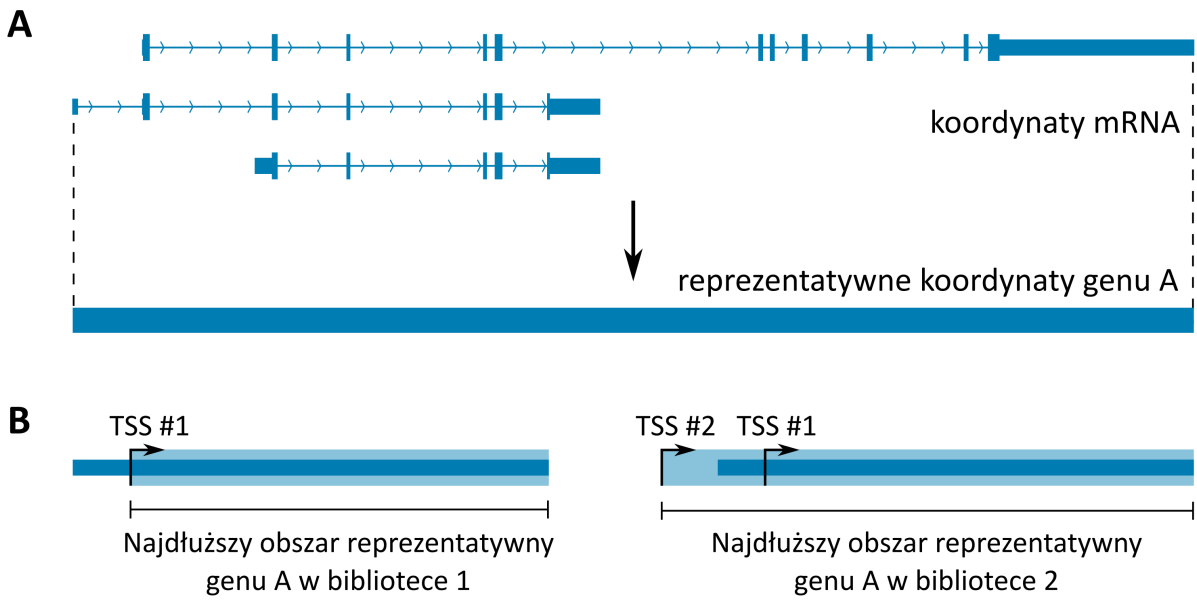
3.4. Dane ChIP-Seq

Dla każdej linii komórkowej gruczolakoraka płuc pobrano z serwera FTP bazy DBTSS zmapowane uprzednio do genomu ludzkiego (w wersji hg38) odczyty w formacie BED3+1, gdzie dodatkowa kolumna (+1) oznacza liczbę odczytów zmapowanych w danym miejscu. Dane te obejmowały wyniki eksperymentów ChIP-Seq celowanych na identyfikację siedmiu rodzajów modyfikacji histonów (H3ac, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3 i H3K9me3) oraz aktywności polimerazy RNA II. Dla każdej linii komórkowej pobrano również dane z odpowiedniego sekwencjonowania kontrolnego.

4. Metody

4.1. Genomowe adnotacje referencyjne i miejsca startu transkrypcji

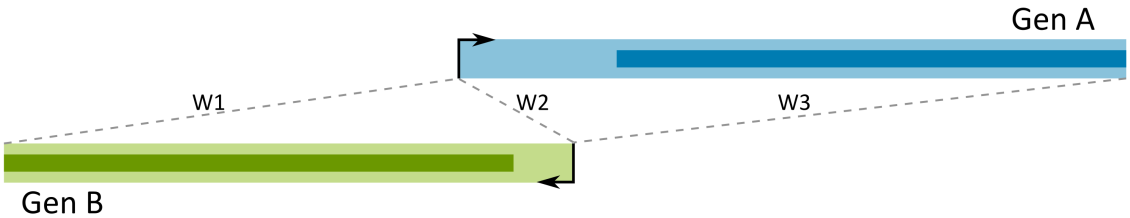
Ludzkie i mysie sekwencje referencyjne z bazy RefSeq przefiltrowane zostały pozostawiając tylko takie, których numer akcesyjny zaczynał się od *NM_**, który to początek w nomenklaturze RefSeq odpowiada sekwencjom mRNA[183]. Użycie w tym kroku adnotacji RefSeq uzasadnione jest wykorzystaniem tych samych adnotacji w bazie DBTSS. Koordynaty transkryptów reprezentujących formy splicingowe danego genu wykorzystane zostały do określenia obszaru zajmowanego przez ten gen, jak zostało to przedstawione na rycinie 5 A. W rezultacie każdy gen reprezentowany był przez jedną parę koordynat, niezależnie od tego jak wiele alternatywnych wariantów splicingowych posiadał. Następnie kolekcje miejsc TSS, wyznaczonych na podstawie poszczególnych bibliotek TSS-Seq, zostały niezależnie przefiltrowane, biorąc pod uwagę tylko takie miejsca TSS, które przez twórców bazy DBTSS oznaczone zostały jako „zaufane" (ang. *confident*) oraz których znormalizowany poziom ekspresji był nie mniejszy niż 5 ppm (liczba odczytów na milion zmapowanych odczytów; ang. *parts per million*). Przyjęcie takiego właśnie minimalnego poziomu ekspresji dla danych typu TSS-Seq umotywowane zostało wynikami badań przedstawionymi przez Yamashita i współpracowników[176]. Kolejnym kryterium przyjętym podczas filtrowania wyników było zmniejszenie dopuszczalnego maksymalnego dystansu pomiędzy miejscem TSS a znanym z adnotacji referencyjnych końcem 5' genu z 50000 nukleotydów, który to dystans przyjęty został w bazie DBTSS[175, 179], do 5000 nukleotydów. W efekcie dla każdej biblioteki utworzono listę genów ulegających ekspresji i przypisanych im miejsc startu transkrypcji. Następnie, na podstawie koordynaty adnotowanego końca 3' oraz koordynaty najdalej wysuniętego miejsca startu transkrypcji, wyznaczony został w każdej bibliotece obszar zajmowany przez dany gen (rycina 5 B). Liczba genów oraz przypisanych im miejsc startu transkrypcji w poszczególnych bibliotekach przedstawiona została w tabeli dodatkowej 1 w aneksie.



Rycina 5. ***Wyznaczanie reprezentatywnych koordynat genów****. Określanie koordynat genu na podstawie lokalizacji jego alternatywnych form splicingowych kodujących białka (A) oraz konsekwentne określanie najdłuższego obszaru reprezentatywnego genu w poszczególnych bibliotekach (B).*

4.2. Identyfikacja par genów nakładających się końcami 5'

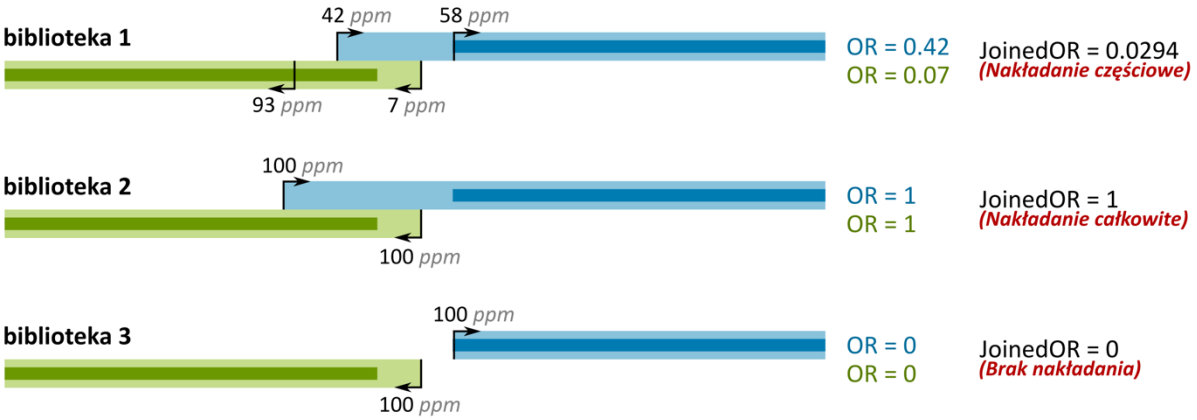
Identyfikacja par genów nakładających się końcami 5' odbyła się niezależnie dla każdej z bibliotek TSS-Seq. Za pary genów nakładających uznawano takie dwa geny, których najdłuższe reprezentatywne koordynaty genów spełniały warunki W1, W2 oraz W3, graficznie zaprezentowane na rycinie 6. Wymagano, aby koniec 5' i 3' reprezentatywnego obszaru genu A, znajdującego się na dodatniej nici DNA, były odpowiednio większe niż koniec 5' i 3' genu B, który zlokalizowany jest na nici ujemnej. Jednocześnie wymagano, aby koordynaty 5' końca genu A znajdowały się poniżej 5' końca genu B. Minimalny przyjęty obszar nakładania wyniósł 1 nukleotyd. Listy wszystkich par genów nakładających się u człowieka i myszy w przynajmniej jednej bibliotece dostępne są w opracowanej na potrzeby niniejszej pracy internetowej bazie danych OverGeneDB, która opisana została w rozdziale 5.9.



Rycina 6. ***Kryteria zastosowane do określenia danej pary genów jako nakładające się końcami 5'.***

4.3. Stopień nakładania się genów

Transkrypcja genów może jednocześnie zostać zainicjowana z wykorzystaniem więcej niż jednego alternatywnego promotora[184, 185]. W wielu przypadkach może to skutkować tym, że jedynie część transkryptów wywodzi się z rejonu nakładania. Sytuacja taka przedstawiona została dla hipotetycznej pary genów na rycinie 7, gdzie oba geny z pary wykorzystują do ekspresji w bibliotece nr. 1 po dwa rejony promotorowe, jednakże tylko bardziej wysunięte z nich nakładają się z genem na nici przeciwnej.



Rycina 7. ***Wartości współczynników OR oraz JoinedOR dla przykładowej pary genów.*** *Para ta wykorzystuje do ekspresji w bibliotekach 1-3 różne rejony promotorowe. W bibliotece 1 dochodzi do nakładania częścią promotorów, które skutkuje przyjęciem przez współczynnik OR obu genów oraz konsekwentnie również JoinedOR, wartości większej od 0 i mniejszej od 1. W bibliotece 2 dochodzi do nakładania się wszystkimi alternatywnymi miejscami startu transkrypcji przez oba geny z pary, co skutkuje przyjęciem przez oba współczynniki wartości równej 1. W bibliotece 3 nie dochodzi do wykorzystania nakładających się miejsc startu transkrypcji, co skutkuje wartościami współczynników OR oraz JoinedOR równej 0.*

Aby określić stopień w jakim gen ulega transkrypcji z wykorzystaniem alternatywnych miejsc startu transkrypcji znajdujących się w rejonie nakładania, opracowano współczynnik nakładania OR (z ang. *Overlap Ratio*). Współczynnik ten wyraża frakcję cząsteczek transkrybowanych z nakładających się promotorów (rycina 7). Aby określić w jakim stopniu geny w danej parze ulegają transkrypcji z wykorzystaniem nakładających się miejsc TSS, opracowano również współczynnik JoinedOR (z ang. *Joined Overlap Ratio*), który jest wynikiem przemnożenia wartości OR obu genów w parze (rycina 7). Maksymalną wartością przyjmowaną przez współczynniki OR oraz JoinedOR jest wartość 1, która oznacza, że wszystkie transkrypty odpowiednio z genu, lub pary genów, zainicjowane były przez nakładające się promotory. Im niższa jest wartość tych współczynników, tym niższa wartość ekspresji przypisana jest do rejonów nakładających się, aż do minimalnej wartości wynoszącej 0, która odzwierciedla brak ekspresji z nakładających się miejsc startu transkrypcji.

4.4. Określenie międzygatunkowego zakonserwowania nakładania się genów

Identyfikacja ortologicznych par genów nakładających się u człowieka i myszy odbyła się w oparciu o adnotacje sekwencji homologicznych zdeponowane w bazie NCBI HomoloGene w wersji 68[186]. Dodatkowym etapem analizy było manualne wyszukanie pośród ortologicznych par genów nakładających się takich par, które ulegają nakładaniu w organach homologicznych między człowiekiem a myszą. Do organów takich należały nerki, serce, wątroba i mózg.

4.5. Szacowanie ekspresji genów nakładających w oparciu o dane RNA-Seq

Surowe odczyty RNA-Seq, pochodzące z 26 linii komórkowych gruczolakoraka płuc poddane zostały kontroli jakości wykonanej przy użyciu programu Trimmomatic (wersja 0.36)[187] z następującymi parametrami: *-phred33; ILLUMINACLIP: adapters/TruSeq3-PE.fa:2:30:10; LEADING: 20; TRAILING: 20; SLIDINGWINDOW:5:20; MINLEN:50*. Jakość poszczególnych bibliotek RNA-Seq zarówno przed jak i po kontroli jakości, została sprawdzona z wykorzystaniem programu FastQC (wersja 0.11.5)[188]. Kontrola jakości na żadnym etapie nie wykazała nadreprezentacji sekwencji rRNA lub tRNA, które w niektórych przypadkach mogą stanowić stosunkowo dużą część wszystkich zsekwencjonowanych odczytów RNA-Seq, negatywnie wpływając na poprawność oszacowanych poziomów ekspresji. Przy użyciu programu HISAT2 (wersja 2.0.5)[189], przefiltrowane odczyty zostały następnie przyrównane do ludzkiego genomu referencyjnego w wersji hg38, który pobrany został z bazy danych UCSC[190]. Użyto domyślnych parametrów za wyjątkiem wykorzystania flagi *--downstream-transcriptome-assembly*, która odpowiada za dostosowanie wyników mapowania do programów składających transkryptomy w trybie *ab initio*. Wykorzystany do mapowania indeks genomu referencyjnego UCSC hg38 został pobrany z oficjalnej strony programu HISAT2 (http://ccb.jhu.edu/software/hisat2). Wynikiem mapowania odczytów były pliki w formacie SAM, które zostały następnie posortowane i przekonwertowane do formatu BAM z wykorzystaniem pakietu programów SAMtools (wersja 1.3.1.)[191].

Całkowity poziom ekspresji genu stanowi sumę poziomów ekspresji wszystkich jego alternatywnych form splicingowych, których ekspresja została obliczona przy użyciu programu StringTie (wersja 1.3.1c)[192]. Program ten uruchomiony został z parametrami *–e* oraz *–B*, które znacznie skracają czas potrzebny do przeprowadzenia analizy biorąc pod uwagę przy szacowaniu poziomu ekspresji transkryptów tylko te odczyty, które zmapowane

zostały wewnątrz koordynat adnotacji referencyjnych. Z tego względu szacowanie poziomu ekspresji genów oparte zostało o adnotacje referencyjne GENCODE (wersja 24)[173], które zawierają więcej alternatywnych form splicingowych niż adnotacje RefSeq. Znormalizowany poziom ekspresji genów oszacowany został w jednostce FPKM (z ang. *fragments per kilobase of exon per million fragments mapped*). Pliki wynikowe w formacie CTAB zostały następnie wykorzystane do analizy ekspresji różnicowej przeprowadzonej za pomocą programu Ballgown (wersja 2.8.0)[193]. Dokładniejszy protokół tej części analizy opisany został w rozdziale 4.6.

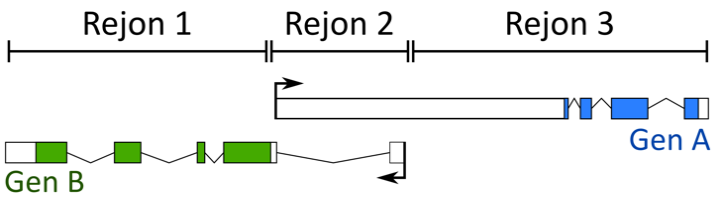
4.6. Badanie wpływu nakładania genów na alternatywny splicing

Analiza potencjalnego wpływu zjawiska nakładania na alternatywny splicing genów przeprowadzona została dla par genów nakładających się przynajmniej w pięciu z 26 bibliotek gruczolakoraka płuc oraz ulegających transkrypcji z wykorzystaniem nienakładających się promotorów w kolejnych pięciu lub więcej bibliotekach. Dla wszystkich przefiltrowanych w ten sposób genów obliczono następnie procentowy udział ich form splicingowych w całkowitej ekspresji. Odbyło się to niezależnie dla każdej z badanych bibliotek gruczolakoraka na podstawie poziomów ekspresji przypisanych poszczególnym wariantom splicingowym, które ustalone zostały zgodnie z protokołem opisanym w rozdziale 4.5. Kolejnym krokiem było obliczenie dla każdego wariantu splicingowego, każdego z genów w parze, średniego udziału w ekspresji w bibliotekach w których geny ulegają nakładaniu, oraz w bibliotekach w których ekspresja obu genów w parze zachodzi z wykorzystaniem nienakładających się miejsc startu transkrypcji. Jeśli dla jakiejś formy splicingowej wartość bezwzględna z różnicy średnich wartości jej udziału w ekspresji genu wynosiła między grupami więcej niż 10%, para genów była raportowana jako kandydat do następnego etapu analizy. Tak przygotowana lista par genów została następnie manualnie przestudiowana z wykorzystaniem przeglądarki genomowej IGV[194, 195]. Do dalszej analizy dopuszczano tylko takie pary genów, w których zmiany procentowego udziału form splicingowych dotyczyły genu, który zawsze wykorzystywał do ekspresji ten sam promotor. Jest to spowodowane tym, że wykorzystanie alternatywnych promotorów samo w sobie mogłoby być powiązane ze zmianami w procentowym udziale poszczególnych form splicingowych genu. Następnie, dla każdej z tak wyłonionych par genów przeprowadzona została zgodnie ze standardowym protokołem[196] analiza ekspresji różnicowej z wykorzystaniem programu Ballgown (wersja 2.8.0)[193]. Analiza ta przeprowadzana była niezależnie dla każdej z par genów dla tych bibliotek gruczolakoraka płuc, w których oba

geny z badanej pary ulegały ekspresji, przy podziale tych bibliotek na te w których ekspresja zachodzi z nakładających się miejsc TSS, oraz te w których ekspresja zachodzi bez nakładania.

4.7. Alleliczna specyficzność ekspresji genów nakładających się

Analiza allelicznej specyficzności ekspresji (ASE, z ang. *allele-specific expression*) przeprowadzona została w oparciu o dane RNA-Seq dla 26 bibliotek gruczolakoraka, które zmapowane zostały do genomu ludzkiego zgodnie z protokołem opisanym w rozdziale 4.5. Pierwszym krokiem analizy było stworzenie dla plików BAM indeksów, co odbyło się z wykorzystaniem narzędzia index z pakietu SAMtools (wersja 1.3.1)[191]. Kolejnym krokiem było zidentyfikowanie polimorfizmów pojedynczych nukleotydów (SNP, z ang. *single nucleotide polymorphism*) oraz krótkich insercji lub delecji, co wykonano niezależnie dla każdej ze zmapowanych bibliotek RNA-Seq za pomocą narzędzia *mpileup* z pakietu SAMtools oraz narzędzia *view* z pakietu bcftools (wersja 0.1.19-96b5f2294a)[197], które uruchomione zostało z parametrami *–bvcg*. Jako wynik otrzymano listę miejsc polimorfizmu pojedynczych nukleotydów oraz indeli, która została następnie przefiltrowana narzędziem *ASEReadCounter* z pakietu GATK (wersja 3.7)[198] z domyślnymi parametrami. W wyniku działania tego narzędzia otrzymano listę koordynat miejsc SNP o potencjalnie biallelicznym charakterze ekspresji wraz z liczbą odczytów RNA-Seq, które potwierdzały w danym miejscu zgodność z referencyjną lub alternatywną sekwencją o pokryciu przynajmniej 10 odczytów. Za sekwencję referencyjną służyła tutaj sekwencja ludzkiego genomu referencyjnego hg38, do której uprzednio mapowane były surowe odczyty RNA-Seq. Dla każdej z 26 bibliotek gruczolakoraka płuc, na podstawie otrzymanych koordynat, miejsca polimorficzne zostały przypisane do genów ulegających ekspresji w danej bibliotece. Wykorzystano w tym celu narzędzie *intersect* z pakietu BEDTools (wersja 2.25.0)[199]. Miejsca SNP przypisywano do genów nakładających się z podziałem na trzy rejony przedstawione na rycinie 8. W ten sposób możliwe było indywidualne zbadanie mono/biallelicznośći każdego z genów nakładających się oraz samego rejonu nakładania.



Rycina 8. ***Podział pary genów na trzy rejony niezależnie badane pod kątem mono/biallelicznośći ekspresji.*** *Rejon pierwszy i trzeci obejmują geny z wyłączeniem rejonu nakładania, podczas gdy ten rozpatrywany jest jako rejon 2. Każdy z trzech rejonów jest niezależnie analizowany pod kątem biallelicznych miejsc SNP, które wykorzystywane są do określenia mono/biallelicznośći ekspresji danego rejonu.*

Dla wszystkich miejsc SNP skojarzonych z tak ustalonymi rejonami obliczono następnie frakcję ekspresji przypisanej do alternatywnych alleli. Jeśli wartość procentowa ekspresji przypisanej do któregoś z alleli wynosiła mniej niż 5%, ekspresja przypisana do danego miejsca SNP uznawana była za potencjalnie monoalleliczną. Poszczególne geny często posiadały więcej niż jedno miejsce SNP. Ekspresję genu uznawano za potencjalnie monoalleliczną tylko w przypadku, gdy wszystkie miejsca SNP skojarzone z tym genem wykazywały ekspresję o charakterze potencjalnie monoallelicznym. W innych wypadkach ekspresja genu uznawana była za bialleliczną.

4.8. Analiza ekspresji różnicowej w kontekście transfekcji

Analiza ekspresji różnicowej w kontekście transfekcji przeprowadzona została na podstawie danych TSS-Seq dla sześciu linii komórkowych Beas2B przedstawionych w tabeli 2. Cztery linie komórkowe poddane celowanej bądź niecelowanej transfekcji potraktowane zostały jako cztery powtórzenia biologiczne linii komórkowych poddanych transfekcji, podczas gdy dwie linie komórkowe oznaczone jako „parent" potraktowane zostały jako kontrola dla tej transfekcji.

Tabela 2. ***Lista sześciu linii komórkowych Beas2B, które podadne zostały analizie ekspresji różnicowej w kontekście transfekcji****. Linie komórkowe oznaczone jako „parent" (pol. Rodzic) stanowią kontrolę do procedury transfekcji, linie oznaczone jako „Kontrola" są poddane zostały niecelowanej transfekcji, natomiast linie komórkowe oznaczone jako STAT6 siRNA+ są liniami poddanymi transfekcji z wykorzystaniem siRNA nacelowanych na wyciszenie ekspresji genu STAT6.*

Linia komórkowa	Kategoryzacja
Beas2B IL4+ parent	Linie komórkowe nie poddane transfekcji
Beas2B IL4- parent	
Beas2B IL4+ Kontrola	Linie komórkowe poddane transfekcji
Beas2B IL4- Kontrola	
Beas2B IL4+ STAT6 siRNA+	
Beas2B IL4- STAT6 siRNA+	

Dla tak pogrupowanych bibliotek, wykorzystując program edgeR[200], przeprowadzono analizę ekspresji różnicowej biorąc pod uwagę łącznie 15634 ludzkie geny, do których przypisano przynajmniej jeden odczyt TSS-Seq w przynajmniej jednej z sześciu bibliotek Beas2B. Potok analityczny oparty został o poradnik użytkownika programu edgeR, z zastosowaniem rekomendowanych parametrów dla przykładowego podejścia opisanego w rozdziale „Quick Start" w wersji z 4 Kwietnia 2016[201]. Jako geny ulegające ekspresji różnicowej uważano takie, dla których prawdopodobieństwo testowe było nie mniejsze niż 0,05. Jako geny ulegające po transfekcji podwyższonej ekspresji uznawano takie, których logarytm o podstawie 2 z krotności zmiany ekspresji genu względem średniego poziomu był większy od 0. Gdy wartość ta była mniejsza, ekspresję genu po transfekcji uznawano za obniżoną względem tej, obserwowanej przed transfekcją.

4.9. Analiza miejsc wiązania czynników transkrypcyjnych

Pierwszym krokiem powiązania rejonów promotorowych z czynnikami transkrypcyjnymi, które mogą je regulować, było zidentyfikowanie potencjalnie tych samych promotorów w różnych bibliotekach TSS-Seq. W tym celu przeprowadzono analizę skupień, łącząc wszystkie miejsca TSS oddalone od siebie w różnych bibliotekach nie więcej niż 300 nt w klastry. Następnie, sekwencje nukleotydowe klastrów TSS, wraz z rejonem flankującym o długości 500 nt w każdą ze stron, zostały zbadane pod katem występowania w nich motywów wiązania czynników transkrypcyjnych pobranych z bazy danych JASPAR[202]. Były to motywy wiązania łącznie dla 367 ludzkich i 141 mysich czynników transkrypcyjnych. Przeszukanie sekwencji nukleotydowych odbyło się z wykorzystaniem narzędzia TFBSTools (wersja 1.14.0)[203], z minimalną wartością relatywnej punktacji (ang. *score*) równą 95%. Wartość relatywnej punktacji wyrażana jest tutaj jako wartość procentowa i określa jakość przyrównania motywu wiązania czynnika transkrypcyjnego w danym miejscu, względem wartości maksymalnej punktacji dla najlepszego możliwego wiązania danego czynnika transkrypcyjnego. Jakość przyrównania obliczana jest w oparciu o matryce PFM (z ang. *position frequency matrix*), które przedstawiają częstotliwość występowania poszczególnych nukleotydów na kolejnych pozycjach motywu wiązania danego czynnika transkrypcyjnego[204]. Tak określona lista czynników transkrypcyjnych, mogących hipotetycznie odpowiadać za regulację poszczególnych promotorów, została następnie dla każdej z bibliotek przefiltrowana, pozostawiając tylko takie czynniki, które w danej linii komórkowej ulegały ekspresji na podstawie danych TSS-Seq.

Nadreprezentacja występowania miejsc wiązania czynników transkrypcyjnych liczona była wykorzystując test na porównanie proporcji dwóch niezależnych grup, oparty o statystykę Z. Porównywano tutaj niezależnie dla każdego czynnika transkrypcyjnego proporcję promotorów posiadających przynajmniej jedno miejsce wiązania danego czynnika wśród rejonów promotorowych genów nakładających się, względem proporcji wśród wszystkich innych genów, które nigdy nie były zidentyfikowane jako nakładające.

4.10. Badania modyfikacji białek histonowych i aktywności polimerazy RNA II

Z bazy DBTSS pobrano koordynaty zmapowanych do genomu ludzkiego odczytów pochodzących z eksperymentów ChIP-Seq, nacelowanych na badanie siedmiu typów modyfikacji histonów oraz aktywności polimerazy RNA II. Dane te dostępne były w formacie BED3+1 dla 26 linii komórkowych gruczolakoraka płuc. Każda z bibliotek ChIP-Seq przekonwertowana została do standardowego formatu BED6. Następnie, posługując się programem MACS2 (wersja 2.1.0)[205], zmapowane odczyty wykorzystane zostały do określenia rejonów wysycenia (ang. *peaks*), tzn. takich, których pokrycie odczytami w stosunku do otoczenia genomowego jest w sposób statystycznie istotny wyższe. Proces ten w języku angielskim zwany jest *peak calling* i przeprowadzony został aktywując w programie MACS2 parametr *--nomodel* dla wszystkich bibliotek ChIP-Seq, oraz dodatkowe parametry *--broad* i *--broad-cutoff 0.1* dla siedmiu modyfikacji histonowych, dla których zaleca się tworzenie dłuższych rejonów wysycenia niż w przypadku analizy czynników transkrypcyjnych[206]. Dane wynikowe z programu MACS2 oraz zmapowane do genomu koordynaty odczytów zostały następnie przygotowane do wizualizacji w przeglądarce genomowej bazy OverGeneDB. W tym celu rejony statystycznie istotnego pokrycia odczytami ChIP-Seq przekonwertowane zostały z formatu BED do bigBed za pomocą narzędzia *bedToBigBed*, które pobrane zostało z bazy UCSC[207]. Wszystkie koordynaty odczytów zmapowanych do genomu zostały natomiast przekonwertowane z formatu BED do bigWig z wykorzystaniem narzędzi *genomecov* z pakietu BEDTools (wersja 2.25.0)[199] oraz *bedGraphToBigWig* z bazy UCSC[207]. Indeksowany format wynikowy bigWig zapewnia tutaj możliwość szybkiego przeglądania pokrycia genomu zmapowanymi odczytami ChIP-Seq.

4.11. Analiza sygnałów aktywności polimerazy RNA II i modyfikacji histonów – modele oparte o dane ChIP-Seq

Zjawisko interferencji transkrypcyjnej przebadane zostało w oparciu o dane ChIP-Seq pochodzące z eksperymentów nacelowanych na badanie aktywności polimerazy RNA II oraz

siedmiu typów modyfikacji histonów. Badano tutaj rejony nakładania genów, w których to zachodzić mogą zakłócenia transkrypcyjne. Celem przeprowadzonych analiz było stworzenie modeli dla rejonu nakładania każdej pary genów nakładającej się w przynajmniej jednej bibliotece gruczolakoraka płuc. Modele rozumiane są tutaj jako uogólnione kombinacje modyfikacji histonów oraz aktywności polimerazy RNA II współwystępujących w rejonie nakładania, charakteryzujące grupy par genów zidentyfikowanych jako nakładające się w przynajmniej jednej bibliotece. Wyodrębnione wzorce epigenetyczne, były następnie przeanalizowane w celu identyfikacji sygnałów, które mogłyby świadczyć o zachodzeniu zjawiska interferencji transkrypcyjnej.

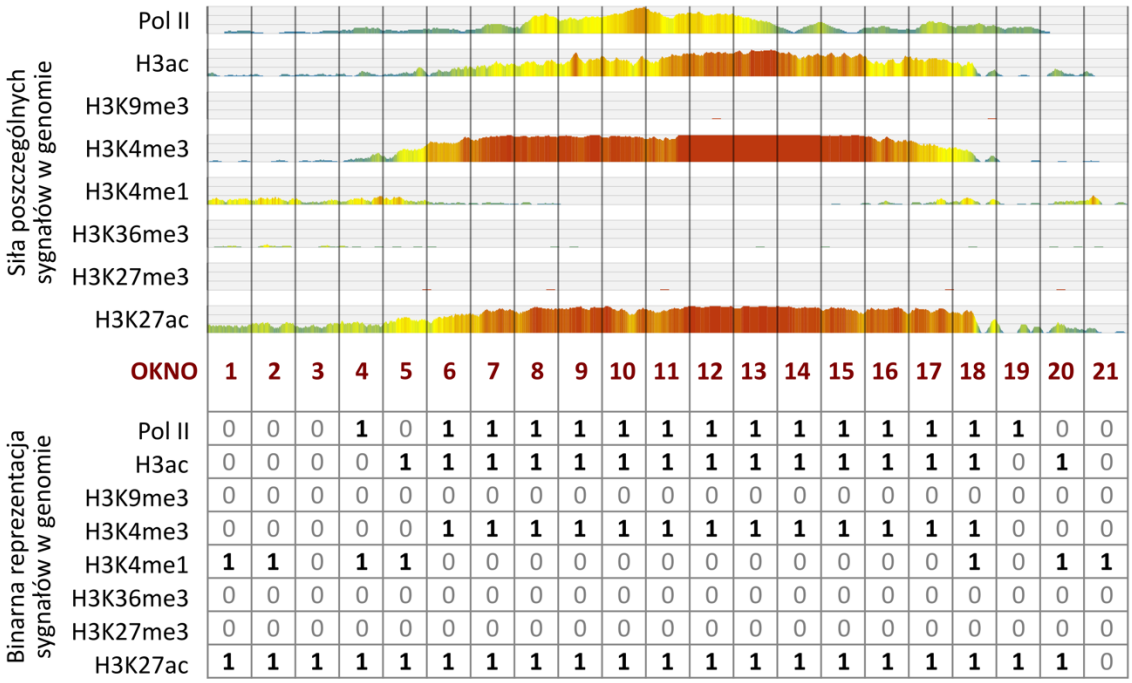
Pierwszym krokiem było określenie rejonu, który włączony zostanie do analizy dla danej pary genów. Zależało to od tego, które geny w parze ulegały w danej bibliotece ekspresji i które promotory, nakładające się lub nie, były wykorzystane. Na tej podstawie pary genów kwalifikowane były do jednej z kategorii:

1. Geny nakładające się – rejon włączony do analizy znajdował się pomiędzy miejscami TSS tworzącymi najdłuższy rejon nakładania pomiędzy genami.
2. Oba geny z pary ulegają ekspresji, ale bez nakładania – podobnie jak w kategorii powyżej, rejon włączany do analizy zlokalizowany był pomiędzy miejscami TSS obu genów. W przypadku, gdy któryś z genów posiadał więcej niż jeden promotor aktywny w danej bibliotece, jako granicę obszaru wybierano TSS najdalej wysunięty w kierunku 5'.
3. Tylko jeden z genów z pary ulega ekspresji – rejon włączany do analizy znajdował się pomiędzy miejscem TSS genu ulegającego ekspresji w danej bibliotece oraz najdalej wysuniętym miejscem TSS genu na nici przeciwnej w bibliotekach, w których gen ten ulegał ekspresji.
4. Żaden z genów z pary nie ulega ekspresji – w przypadku braku ekspresji obu genów w badanej bibliotece, do analizy włączano rejon wyznaczony przez najdalej wysunięte koordynaty miejsc TSS obu genów w innych bibliotekach.

Dla każdego analizowanego rejonu włączano także rejon flankujący, który obejmował obszar o długości 1000 nt w każdą ze stron. Tak przygotowane rejony zapisywane były dla każdej pary genów w formacie BED.

Informacja o aktywności polimerazy RNA II oraz modyfikacjach histonów przygotowana została w sposób niestandardowy z wykorzystaniem polecenia *BinarizeBed* z programu Spectacle[151]. Polecenie to ma za zadanie stworzyć binarną matrycę reprezentującą

występowanie poszczególnych sygnałów w całym genomie. W ramach opracowanego protokołu zmieniono w poleceniu *BinarizeBed* domyślną długość okna (ang. *window*) z 200 do 25 nukleotydów. W programie Spectacle oknem określa się długość obszarów na jakie podzielony zostaje cały genom do którego zmapowane są analizowane odczyty. W każdym z tak wyznaczonych okien program sprawdza występowanie statystycznie istotnego podwyższenia liczby zmapowanych odczytów ChIP-Seq względem kontroli. Wystąpienie takiego obszaru oznaczane jest cyfrą 1, natomiast brak statystycznie istotnych różnic oznaczany jest cyfrą 0. Pozwala to na stworzenie binarnej reprezentacji całego genomu, której przykładowy fragment przedstawiony został na rycinie 9.

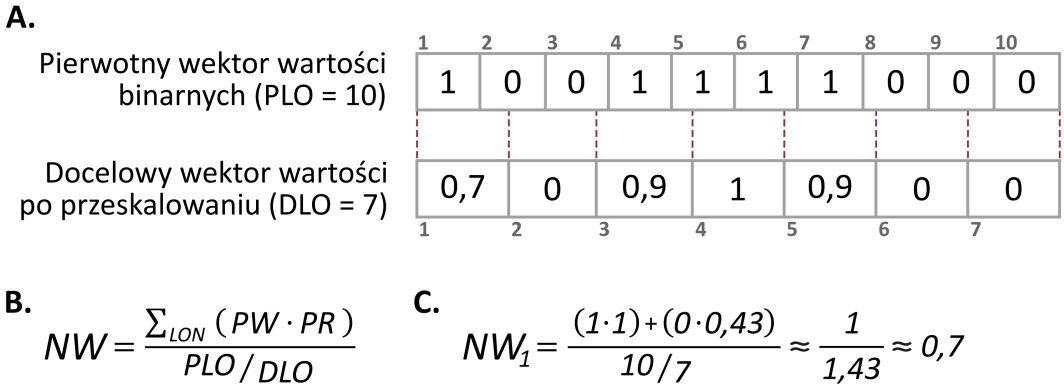


OKNO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Binarna reprezentacja sygnałów w genomie																					
Pol II	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
H3ac	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0
H3K9me3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H3K4me3	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
H3K4me1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1
H3K36me3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H3K27me3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H3K27ac	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0

Rycina 9. ***Binarna reprezentacja sygnałów polimerazy RNA II i siedmiu typów modyfikacji histonów.*** *Utworzona matryca binarna jest jedynie matrycą demonstracyjną, która ma zilustrować ogólną zasadę tworzenia matrycy na podstawie obserwowanych w genomie sygnałów.*

Z tak przygotowanych map binarnych całego genomu wyodrębniono następnie rejony, które zostały uprzednio wyznaczone do analizy dla poszczególnych par genów. W efekcie każda para genów reprezentowana była przez macierz binarną o wymiarach 8 x *N,* gdzie cyfra 8 odpowiada liczbie analizowanych bibliotek ChIP-Seq, natomiast *N* odpowiada długości badanego rejonu, podzielonego przez 25, czyli przez długość okna przyjętą w poleceniu *BinarizeBed*. Kolejnym etapem analizy było przeprowadzenie hierarchicznej analizy skupień. Ponieważ długość analizowanego obszaru zależała od badanej pary genów, przed obliczeniem dystansu pomiędzy poszczególnymi parami binarne mapy reprezentacji badanego rejonu

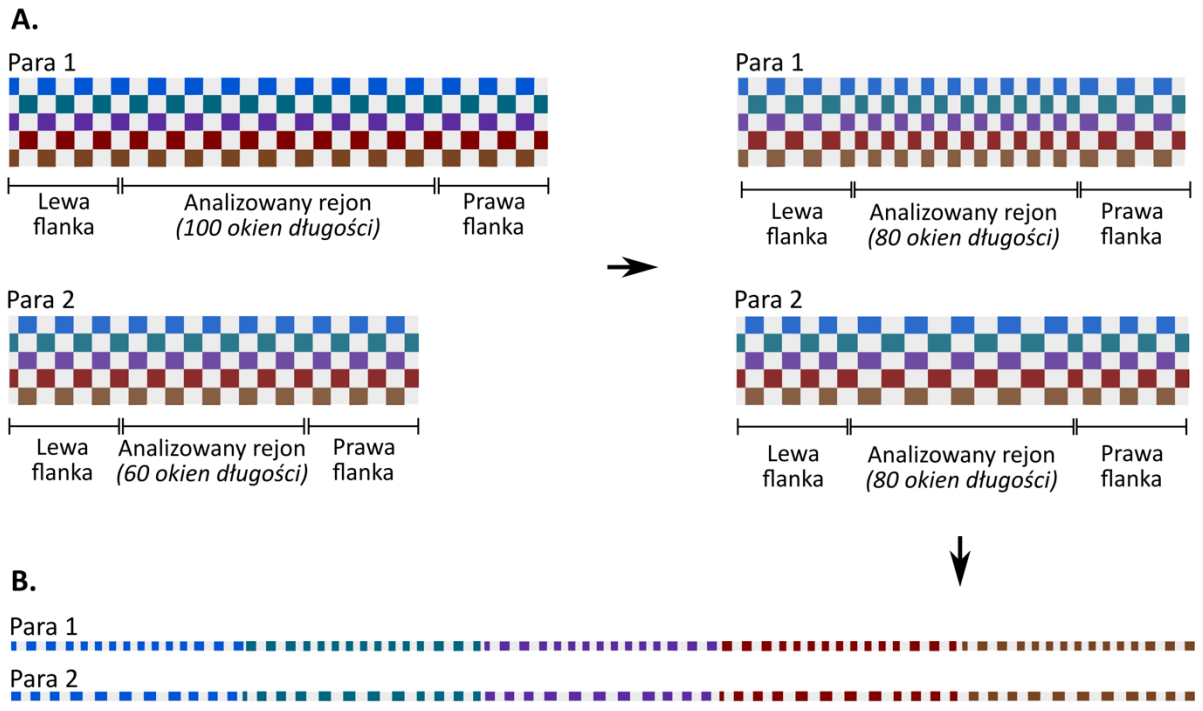
musiały być najpierw przekształcone w wektory o takim samym wymiarze, tzn. o takiej samej liczbie okien. W tym celu informacja zawarta w analizowanym obszarze musiała być przeskalowana, co wiązało się z przeliczeniem wartości z poszczególnych okien. Przeliczenie odbyło się z wykorzystaniem opracowanej w tym celu metody. Zasadę działania omówiono na przykładzie jednego wektora o długości 10 okien, który przeskalowany został do wektora o długości 7 okien, co przedstawione zostało na rycinie 10 A. Obliczenie nowych wartości w poszczególnych oknach odbyło się zgodnie ze wzorem przedstawionym na rycinie 10 B. Dla pierwszego przeskalowanego okna na rycinie 10 A, nowa wartość liczona była na podstawie wartości z dwóch pierwszych okien z pierwotnego wektora. Nowe okno nakładało się w 100% z pierwszym oknem wektora pierwotnego oraz w 43% z oknem drugim. Suma iloczynów wartości w poszczególnych oknach z wektora pierwotnego oraz procentowego pokrycia wyniosła w tym przypadku wartość 1. Wartość ta została następnie podzielona przez iloraz liczby okien w pierwotnej i docelowej matrycy. W efekcie nowa wartość przeskalowanego wektora w oknie pierwszym wyniosła w zaokrągleniu do jednego miejsca po przecinku 0,7 (rycina 10 C).



Rycina 10. ***Przeskalowywanie wektora****. Binarne wartości z pierwotnego wektora o długości 10 okien przeskalowywane są do wektora o docelowej długości siedmiu okien. A) Wektor przed i po przeskalowaniu; B) Wzór zgodnie z którym w obliczane są nowe wartości w poszczególnych oknach docelowego wektora; C) Przykładowe obliczenie nowej wartości okna pierwszego w wektorze docelowym, Skróty: PLO – pierwotna liczna okien; DLO – docelowa liczba okien; NW – nowa wartość okna; NW1 – nowa wartość okna; LON – okna z pierwotnego wektora, które nakładają się z nowym oknem; PW – pierwotna wartość z wektora przed przeskalowaniem; PR – procent w jakim nowe okno nakłada się ze starym oknem.*

Wykorzystując powyższą metodę, badany obszar każdej pary genów przeskalowany został tak, by zajmował 80 okien, podczas gdy rejony flanek, które w każdym przypadku były tej samej długości, nie podlegały przeskalowaniu, zajmując łącznie kolejne 80 okien reprezentacji binarnej jak zostało to graficznie zaprezentowane na rycinie 11 A.

Ponieważ zbyt duże przeskalowanie może wprowadzać pewne zafałszowania sygnału, z dalszej analizy wyłączono pary, których badany rejon przed przeskalowaniem był dłuższy niż 3000 nukleotydów lub krótszy niż 200 nukleotydów. Kryteria takie dobrane zostały na podstawie badań wstępnych przeprowadzonych podczas opracowywania metody. Tak przygotowane matryce przekształcone zostały następnie w wielowymiarowe wektory poprzez kolejne umieszczanie za sobą przeskalowanych wartości z poszczególnych wierszy danych (rycina 11 B). W ten sposób przeskalowano i przekształcono matrycę danych binarnych do wektora o wymiarach 1 x (160*8), gdzie 160 odnosi się do końcowej liczby analizowanych okien danych a cyfra 8 reprezentuje siedem modyfikacji histonów i aktywność polimerazy RNA II. Matryce zostały następnie poddane hierarchicznej analizie skupień. Analiza ta wykonana została z wykorzystaniem skryptów napisanych w języku programowania Python (https://www.python.org/), w których użyto narzędzi *linkage* oraz *dendrogram* z pakietu SciPy (wersja 0.12.0)[208]. Dystans obliczano metodą minimalnej wariancji Warda (ang. *Ward's minimum variance method*).



Rycina 11. ***Przekształcanie binarnej matrycy reprezentującej gen w danej parze, do wektora danych.*** *Wizualizacja zasady przekształcenia binarnej matrycy z informacją o obecności poszczególnych rodzajów modyfikacji histonów lub aktywności polimerazy RNA II do wielowymiarowego wektora danych mogącego być poddanym analizie skupień. Poszczególne kolory na szachownicy reprezentują poszczególne modyfikacje histonów i aktywność polimerazy. Dla uproszczenia przedstawiono tylko 5 wierszy danych. Liczba „okien długości", znajdująca się pod opisem poszczególnych analizowanych rejonów, odnosi się do liczby okien o długości 25 nukleotydów, na które badany obszar genomu podzielony został przez program Spectacle. (A) pary genów o różnej długości badanego rejonu przeskalowane zostały tak, by posiadały one te same wymiary; (B) Utworzenie z przeskalowanych matryc binarnych wektora danych.*

4.12. Maskowanie sekwencji docelowych miRNA przez rejon nakładania

Identyfikacja par genów, których rejon nakładania odpowiadać może za maskowanie sekwencji docelowych cząsteczek miRNA, oparta została o manualną analizę par genów nakładających. Pierwszym krokiem było zidentyfikowanie wszystkich par genów, których rejon nakładania w przynajmniej jednej bibliotece znajduje się w rejonie 3' UTR jednego lub obu genów w parze. Następnie w oparciu o lokalizację sekwencji docelowych miRNA, zdeponowanych w bazach danych mirTarBase[209] oraz Target Scan Human (wersja 7.1)[210] ustalono, które rejony nakładania pokrywają się ze znanymi sekwencjami docelowymi miRNA.

4.13. Implementacja internetowej bazy danych genów nakładających

Baza danych genów nakładających się zaimplementowana została w systemie MySQL (https://www.mysql.com/). Zdeponowane w niej zostały kluczowe informacje dotyczące zidentyfikowanych ludzkich i mysich par genów nakładających się. Publiczny internetowy interfejs bazy danych, dostępny pod adresem http://overgenedb.amu.edu.pl, zaimplementowany został w językach HTML, PHP oraz JavaScript. Szczegółowy podgląd indywidualnej nakładającej się pary genów został dodatkowo zaopatrzony w zagnieżdżoną w interfejs przeglądarkę genomową Dalliance[211].

4.14. Analizy statystyczne i języki programowania

Większość skryptów niezbędnych do przeprowadzenia analiz opisanych w niniejszej pracy doktorskiej zostało zaimplementowane w języku programowania Python w wersji 2.7 (https://www.python.org/). Analizy statystyczne, włączając w to analizę korelacji Pearsona i Spearmana, Test U Manna-Whitneya, oraz różne odmiany testu T studenta, oparte zostały o biblioteki SciPy (wersja 0.12.0; https://www.scipy.org/) oraz NumPy (wersja 1.7.1; http://www.numpy.org/). Wizualizacja wyników w postaci wykresów odbyła się w oparciu o bibliotekę Matplotlib (wersja 1.3.0; http://matplotlib.org/) oraz późniejszą manualną obróbkę graficzną przeprowadzoną w programie InkScape (https://inkscape.org/). Analizy dotyczące działań na sekwencjach nukleotydowych wykonane zostały z wykorzystaniem biblioteki BioPython (wersja 1.64; http://biopython.org/).

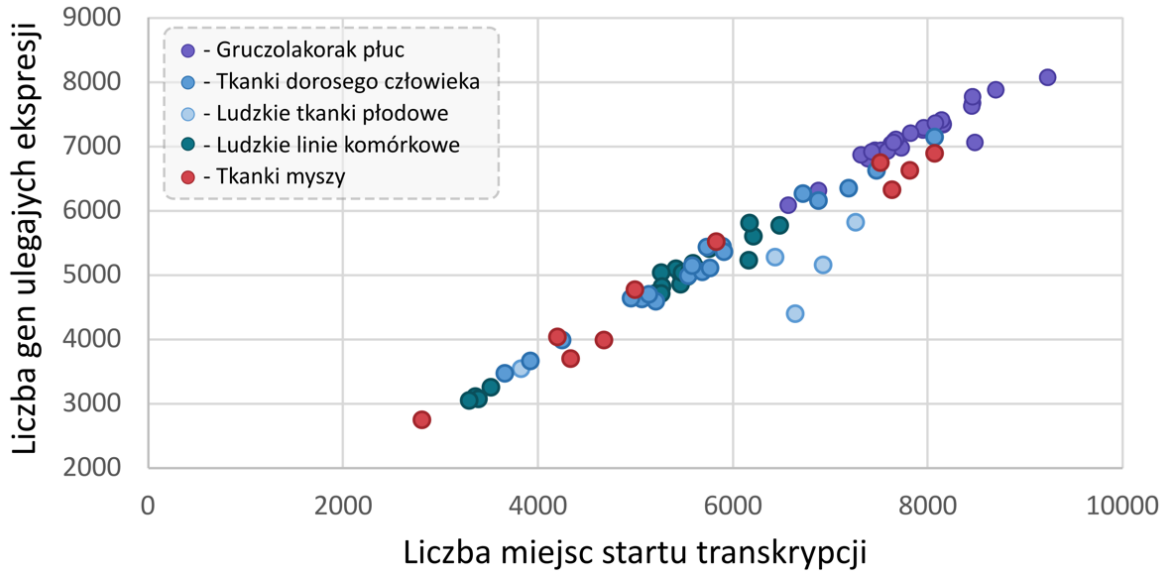
Pozostałe skrypty, głównie związane ze standardowymi analizami wyników sekwencjonowania nowych generacji oraz analizą miejsc wiązania czynników transkrypcyjnych, zaimplementowane zostały w języku programowania R (https://www.r-project.org/). Język R został ponadto wykorzystany do niektórych analiz

statystycznych włączając w to obliczenie jądrowego estymatora gęstości (ang. *kernel density*), oraz wizualizację danych za pomocą wykresów typu *bean plot*[212].

5. Wyniki

5.1. Identyfikacja genów nakładających się u człowieka i myszy

Na podstawie referencyjnych adnotacji z bazy RefSeq ustalono reprezentatywne koordynaty 25 553 ludzkich i 23 899 mysich genów. Spośród nich, w oparciu o dane z bazy DBTSS, zidentyfikowano łącznie 15 778 i 12 508 genów ulegających ekspresji w przynajmniej jednej z 73 i 10 bibliotek odpowiednio u człowieka i myszy. Genom tym przypisano łącznie 46 278 alternatywnych miejsc TSS u człowieka i 21 042 u myszy. Tabela dodatkowa 1, zamieszczona w aneksie, zawiera informację o liczbie aktywnych miejsc TSS w każdej z bibliotek. Liczba ta, zgodnie z oczekiwaniem, jest silnie skorelowana z liczbą wszystkich genów ulegających ekspresji w danej bibliotece (współczynnik korelacji Pearsona = 0.95; p-value < 0,00001), co graficznie przedstawione zostało na rycinie 12.



Rycina 12. ***Wizualizacja zależności między liczbą miejsc TSS a liczbą genów ulegających ekspresji w danej bibliotece****.*

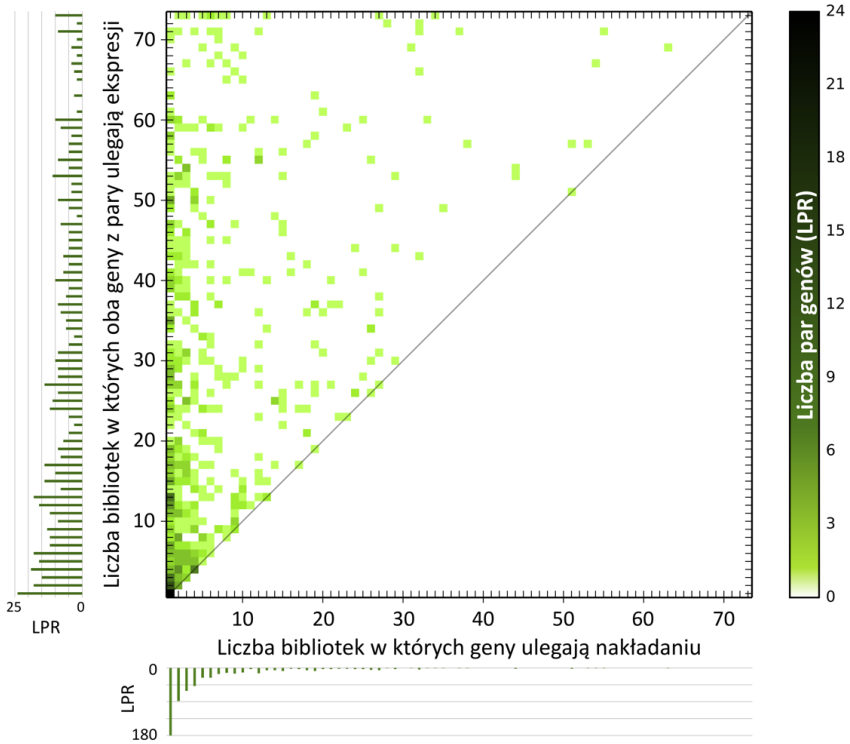
Zarówno ludzkie jak i mysie geny najczęściej wykorzystują jedno, choć nie w każdej bibliotece to samo, miejsce TSS. Taka sytuacja dotyczy odpowiednio 9 065 i 9 349 genów ludzkich i mysich. Geny wykorzystujące do ekspresji zawsze więcej niż jeden promotor należą do rzadkości. U człowieka i myszy jest odpowiednio tylko 106 i 299 tego typu genów. Prawie trzykrotnie wyższa liczba takich genów u myszy w porównaniu z człowiekiem jest najprawdopodobniej związana z tym, że u człowieka analizowana była siedmiokrotnie wyższa liczba bibliotek, co wydaje się zmniejszać prawdopodobieństwo znalezienia genów używających wielu promotorów we wszystkich badanych bibliotekach. Na tle analizowanych

bibliotek nieco wyróżniają się te pochodzące z ludzkich tkanek płodowych, gdzie liczba aktywnych TSS przypadających na liczbę genów ulegających ekspresji w danej tkance jest zawsze wyższa aniżeli w odpowiadającym im tkankach dorosłego człowieka. Jako przykład może tutaj posłużyć mózg, w którym u dorosłego człowieka średnio co dziewiąty gen wykorzystuje do ekspresji dwa miejsca TSS, podczas gdy w płodowym stadium rozwojowym tego organu już średnio co czwarty gen wykorzystuje dwa miejsca startu transkrypcji. Różnica ta jest wyjątkowo duża w przypadku ludzkiego serca, dla którego na jeden gen u osoby dorosłej przypada 1,06 TSS, natomiast w rozwoju płodowym liczba ta sięga 1,51 TSS, co oznacza, że średnio co drugi gen używa do ekspresji dwóch rejonów promotorowych.

Na podstawie koordynat miejsc TSS przypisanych poszczególnym genom zidentyfikowano 582 pary ludzkich i 113 par mysich genów nakładających się na końcach 5' przynajmniej w jednej z 73 i 10 bibliotek TSS-Seq odpowiednio u człowieka i myszy. Na powyższe pary genów składa się odpowiednio 1150 ludzkich i 225 mysich genów, pośród których łącznie 14 genów nakłada się z więcej niż jednym genem na przeciwnej nici DNA. Genom tym przypisanych zostało w sumie 4075 i 518 miejsc startu transkrypcji odpowiednio u człowieka i myszy. Miejsca te wykorzystywane są naprzemiennie w różnych bibliotekach. Średnia liczba miejsc TSS przypadających na gen jest w przypadku genów nakładających się u obu gatunków wyższa niż w przypadku wszystkich innych genów. Dla ludzkich genów nakładających się liczba ta wynosi średnio 3,54 miejsc TSS przypadających na gen, co jest o 23% wyższą wartością niż średnia 2,88 miejsc TSS przypadających na pozostałe ludzkie geny. Z kolei dla mysich genów nakładających i nienakładających się średnia liczba wszystkich miejsc TSS przypadających na gen wynosi odpowiednio 2,3 i 1,71. Średnia długość rejonu nakładania wynosi 1570 pz, minimalna długość wynosi 1 pz i zidentyfikowana została dla dwóch par genów, *GPN1* i *CCDC121* oraz *ARL6IP6* i *PRPF40A*. Maksymalna długość rejonu nakładania wyniosła prawie 50 kpz i odnotowana została w przypadku pary genów *RUNX2* i *SUPT3H.* Zidentyfikowanie w tym przypadku tak długiego rejonu nakładania nie wydaje się być błędne, ponieważ pozycje miejsc TSS znajdują się w okolicy adnotowanych w bazie RefSeq końców 5' obu genów, które same w sobie są wyjątkowo długie, osiągając odpowiednio 222 kpz i 551 kpz długości.

Żadna ze zidentyfikowanych par genów nie ulegała nakładaniu we wszystkich badanych bibliotekach, co dla danych człowieka zostało graficznie zaprezentowane na histogramie prezentującym zależność między liczbą bibliotek, w których dana para ulega ekspresji, a liczbą bibliotek, w których ta para ulega ekspresji z nakładających się TSS ź(rycina 13). Relatywne zagęszczenie znajdujące się w lewym dolnym rogu wykresu

spowodowane jest tym, że większość zidentyfikowanych par genów ulega jednoczesnej ekspresji tylko w stosunkowo małej liczbie bibliotek. Pary genów znajdujące się na zaznaczonej przekątnej wykresu reprezentują takie pary, które zawsze ulegały ekspresji z wykorzystaniem nakładających się promotorów. Pośród zidentyfikowanych par genów nakładających się znaleziono 90 takich właśnie ludzkich i mysich przypadków. Najczęściej wykorzystującą nakładające się miejsca TSS jest para ludzkich genów *ATF5* i *NUP62,* która nakładała się w pięćdziesięciu jeden bibliotekach. Kolejne trzy pary genów ulegające ekspresji wyłącznie z wykorzystaniem nakładających się TSS to pary *FAM120A* i *FAM120OS*, *CIAO1* i *TMEM127* oraz *PPCS* i *ZMYND1*. Każda z tych par ulegała ekspresji w ponad dwudziestu ludzkich bibliotekach. W przypadku myszy, najczęściej wykorzystującą nakładające się miejsca startu jest para genów *Ascc1* i *Anapc16*, która nakłada się w ośmiu bibliotekach. Pozostałe 605 ludzkich i mysich par genów czasem ulegało ekspresji z nakładających się a czasem z nienakładających miejsc alternatywnego startu transkrypcji. Transkrypcja 203 par genów inicjowana była w rejonie nakładania jedynie w jednej bibliotece, podczas gdy oba geny często ulegały ekspresji bez nakładania w wielu innych bibliotekach.



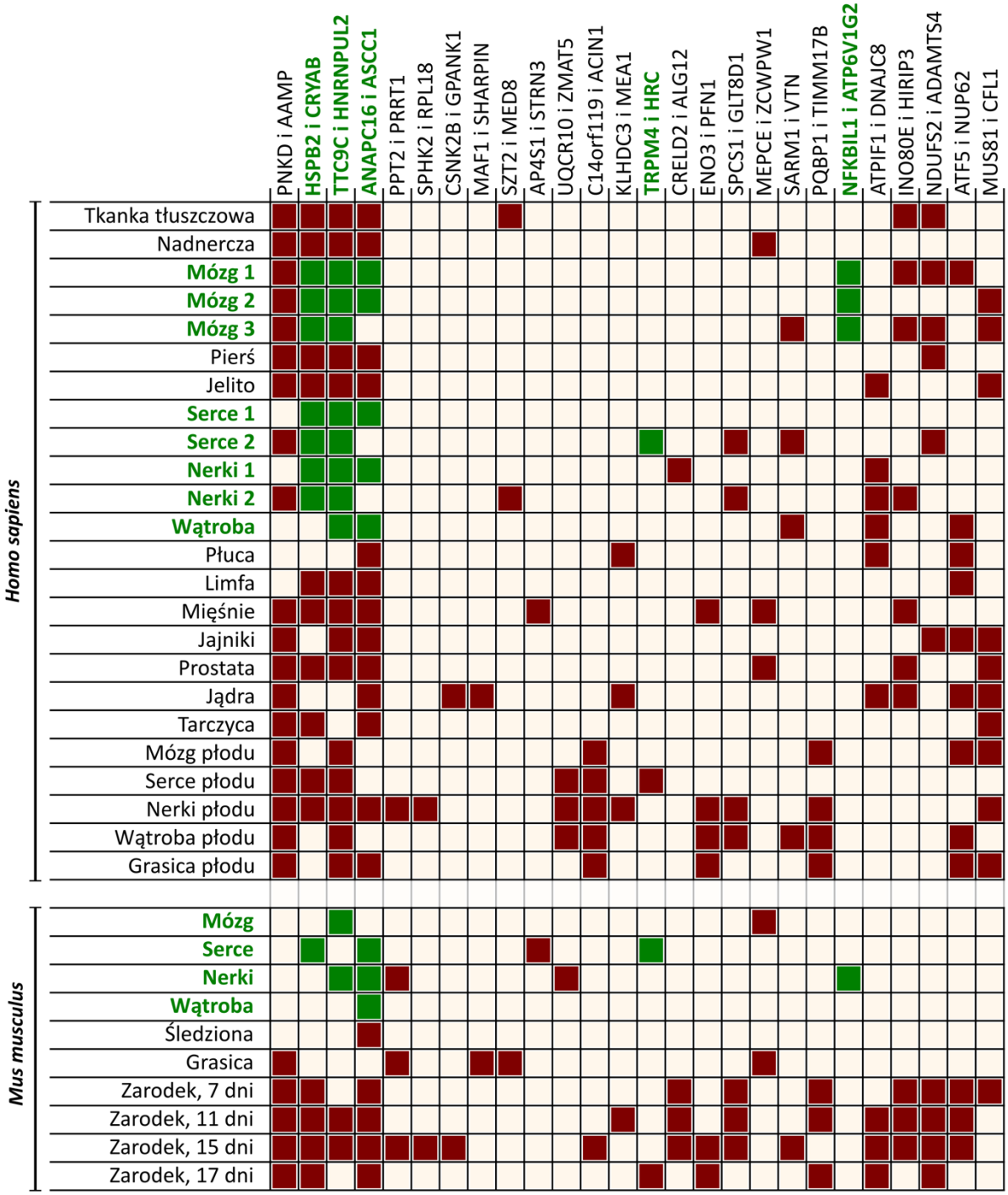
Rycina 13. ***Dwuwymiarowy histogram dla liczby par genów ulegających ekspresji z nakładających lub nienakładających miejsc TSS u człowieka****. Prezentuje on zależność między liczbą bibliotek w których dana para ulega ekspresji z nakładających się miejsc startu transkrypcji (oś X) a liczbą bibliotek w których ta sama para ulega ekspresji niezależnie od zachodzenia zjawiska nakładania (oś Y). Każda z osi jest dodatkowo zaopatrzona w wykres słupkowy prezentujący sumaryczną liczbę par genów dla danej kolumny (wykres dla osi X) lub wiersza (wykres dla osi Y). LPR – liczba par genów.*

5.2. (Nie)Stabilność zachowania nakładania genów

Identyfikacja par genów kodujących białka, które nakładają się końcami 5' przeprowadzona została łącznie dla 73 ludzkich i 10 mysich bibliotek TSS-Seq. Porównanie wyników pomiędzy tymi bibliotekami wykazało znaczne różnice w skali występowania tego zjawiska pomiędzy różnymi organami, tkankami i liniami komórkowymi. Różnice te wynikają z wykorzystania alternatywnego, nienakładającego się rejonu promotorowego, przez co oba geny z pary genów nakładającej się w jednej bibliotece ulegają ekspresji już bez nakładania w innej bibliotece, oraz z braku ekspresji jednego lub obu genów z pary nakładającej się w danej bibliotece. Obserwowane zjawisko może zależeć od czynników gatunkowo lub tkankowo specyficznych lub być zależnym od bardzo ogólnie rozumianych czynników środowiska wewnętrznego i zewnętrznego. W związku z tym przeprowadzono analizę porównawczą ludzkich i mysich genów nakładających się pod kątem zakonserwowania wzorców ich ekspresji i wykorzystania alternatywnych miejsc startu transkrypcji. Analiza została przeprowadzona na czterech poziomach obejmując analizę stopnia zakonserwowania zjawiska nakładania się między gatunkami, różnymi tkankami, warunkami eksperymentalnymi oraz materiałem biologicznym pobranym od różnych dawców.

5.2.1. Zakonserwowanie zjawiska nakładania między człowiekiem i myszą

W oparciu o adnotacje sekwencji homologicznych zdeponowane w bazie HomoloGene[186], spośród 582 ludzkich i 113 mysich par genów nakładających się, zidentyfikowano 26 nakładających się par ortologicznych (rycina 14), co stanowi 4,5 % i 23% wszystkich par genów nakładających się odpowiednio u człowieka i myszy. Zaostrzenie kryteriów do analogicznych międzygatunkowo typów tkanek, przyczyniło się do spadku tej liczby do 5 par genów oznaczonych na rycinie 14 kolorem zielonym.

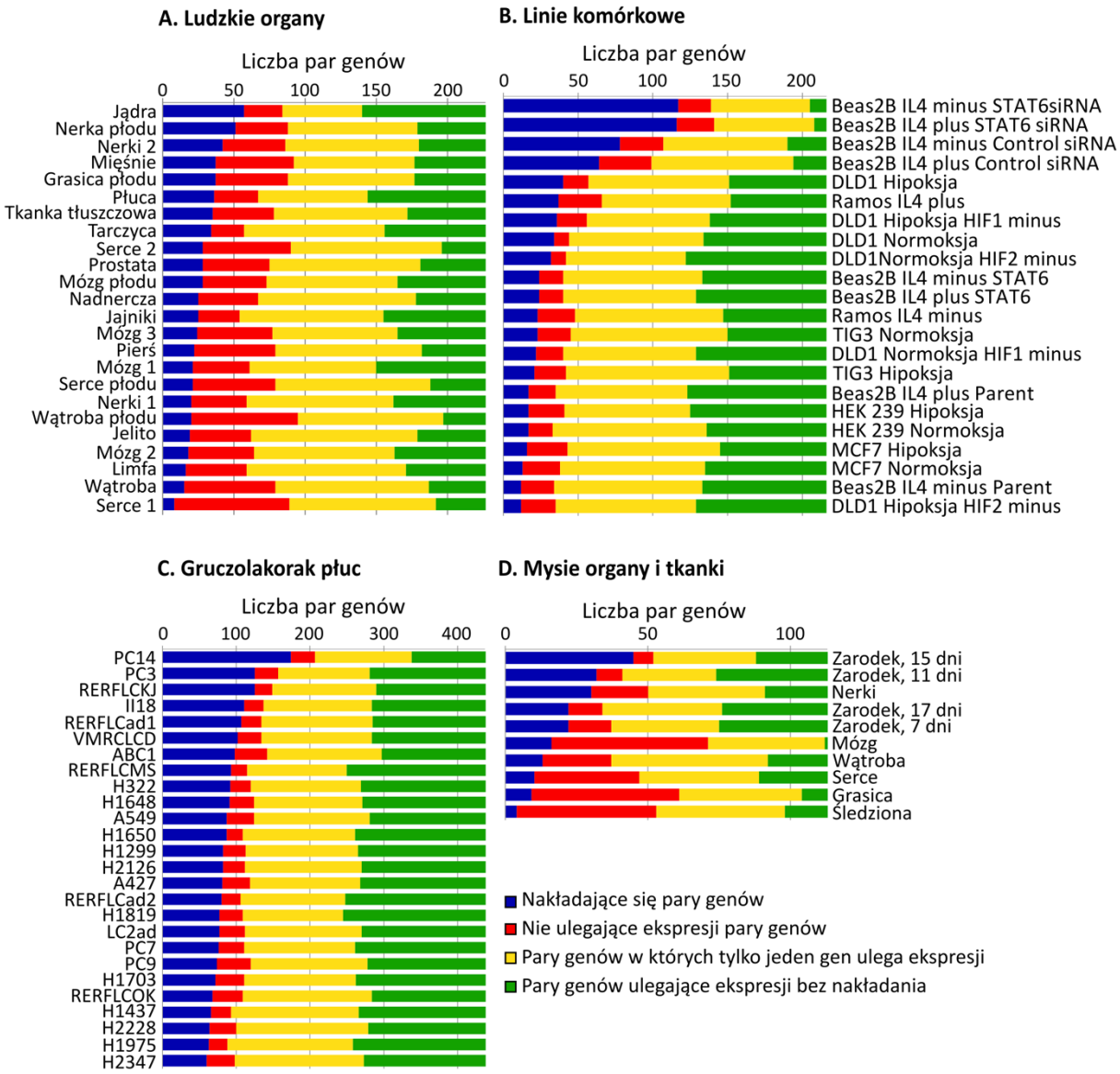


Rycina 14. ***Międzygatunkowe zakonserwowanie zjawiska nakładania się genów****. Kolorem wiśniowym zaznaczono nakładanie się pary genów w danej tkance, podczas gdy kolorem zielonym zaznaczono nakładanie się tych par w tkankach analogicznych między człowiekiem i myszą. Kolorem zielonym zaznaczono również nazwy czterech tkanek analogicznych oraz pięciu par genów nakładających się w nich u obu gatunków.*

5.2.2. Porównanie między różnymi tkankami, warunkami eksperymentalnymi oraz bibliotekami pochodzącymi od różnych dawców

Aby sprawdzić do jakiego stopnia wykorzystanie nakładających się promotorów może być związane z czynnikami tkankowo specyficznymi przeanalizowano 24 biblioteki pochodzące z płodowego i dorosłego stadium rozwojowego organów ludzkich. W tym zestawie zidentyfikowano łącznie 223 pary genów nakładających się przynajmniej w jednej bibliotece. Pośród nich, 53 pary genów ulegały nakładaniu we wszystkich bibliotekach, w których oba geny z pary ulegały ekspresji, jednakże w 20 przypadkach ograniczało to się do jednej biblioteki. Pozostałe 170 par genów ulegało ekspresji z nakładających się miejsc startu transkrypcji jedynie w części bibliotek, w innych natomiast wykorzystywane były nienakładające się promotory. Największa liczba nakładających się par genów została zidentyfikowana w jądrach, gdzie odnotowano 57 takich par, podczas gdy najniższą liczbę, osiem par genów, zidentyfikowano w sercu dorosłego człowieka (rycina 15 A).

Zróżnicowanie liczby zidentyfikowanych par genów nakładających się w poszczególnych tkankach może być związane z ogólną liczbą genów ulegających w nich ekspresji. W przypadku omawianych danych istotnie jądra odznaczają się najwyższą, a serce najniższą liczbą transkrybowanych genów (tabela 1 w aneksie). Jednakże współczynnik korelacji obliczony między liczbą genów ulegających ekspresji w danej bibliotece, a liczbą zidentyfikowanych genów nakładających się, przyjmuje relatywnie niską pozytywną wartość (współczynnik korelacji Pearsona = 0,489; wartość P = 0,015). Zróżnicowanie liczby nakładających się par genów między różnymi bibliotekami zaobserwowano również dla danych myszy. Pośród dziesięciu mysich bibliotek największa liczba 45 par nakładała się w piętnastodniowym embrionie, podczas gdy tylko 4 pary nakładały się w śledzionie (rycina 15 D).



Rycina 15. ***Podsumowanie liczby par genów ulegających ekspresji z nakładających się lub nienakładających miejsc startu transkrypcji.*** *(A) wykres dla 24 ludzkich organów. (B) wykres dla 22 ludzkich linii komórkowych hodowanych w różnych warunkach eksperymentalnych. (C) wykres dla 26 linii komórkowych gruczolakoraka płuc, pobranych od różnych pacjentów. (D) wykres dla 10 ludzkich organów i tkanek.*

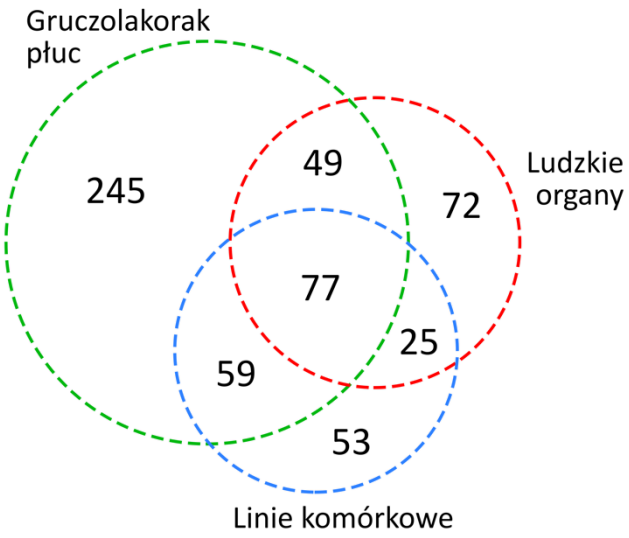
Znaczne zróżnicowanie liczby par genów nakładających się między różnymi tkankami może być spowodowane tkankowo specyficznymi czynnikami. Aby zbadać na ile inne czynniki mogą mieć wpływ na zjawisko nakładania się genów przestudiowano 6 typów ludzkich linii komórkowych, hodowanych w różnych warunkach lub poddanych różnym eksperymentom. Linie te obejmowały próbki raka jelita grubego (DLD1), nabłonka układu oddechowego (Beas2B), chłoniaka Burkitta (Ramos), raka piersi (MCF7), fibroblastów (TIG-3) oraz nerki pochodzenia embrionalnego (HEK 293). W analizie porównawczej nie uwzględniono linii komórkowej Hela, hodowanej tylko w normalnych warunkach tlenowych.

Jak zostało to przedstawione na rycinie 15 B, liczba par genów ulegających ekspresji z nakładających się miejsc startu transkrypcji jest między różnymi liniami komórkowymi bardzo zmienna. Różnice widoczne są również pomiędzy liniami komórkowymi tego samego typu hodowanymi w różnych warunkach lub poddanych różnym eksperymentom. Szczególną uwagę przykuwa 8 linii komórkowych pochodzących z nabłonka układu oddechowego, wśród których zidentyfikowano zarówno największą jak i jedną z najmniejszych liczb nakładających się par genów. Przeprowadzenie transfekcji 4 linii komórkowych Beas2B, niezależnie od tego, czy była ona celowana na wyciszanie ekspresji genu *STAT6*, czy też nie, skutkowała znacznym podwyższeniem liczby genów nakładających się w tych liniach w stosunku do ich kontroli. Jest to doskonale widoczne na rycinie 15 B, gdzie na przykład linia komórkowa *Beas2B STAT6 siRNA+ IL4-,* w której transfekcja nacelowana była na wyciszenie czynnika transkrypcyjnego *STAT6,* odznacza się najwyższą w tym zestawie liczbą 117 par genów nakładających się. W kontroli do transfekcji, tj. bibliotece *Beas2B parent IL4-*, zidentyfikowano natomiast tylko 12 par genów nakładających się. Podobne obserwacje dotyczą wszystkich poddanych transfekcji komórek nabłonkowych i ich kontroli, niezależnie od pozostałych czynników. Transfekcja może być zatem potencjalną przyczyną wykorzystania nakładających się promotorów w miejsce nienakładających. Szersza analiza wpływu transfekcji linii komórkowych Beas2B na wykorzystanie nakładających się promotorów przedstawiona jest w rozdziale 5.6.

Wykazano, że zjawisko nakładania może być związane zarówno z czynnikami tkankowo specyficznymi, jak również egzogennymi. Analizy wykonane na kolejnym etapie skupiły się na poznaniu zróżnicowania zjawiska nakładania się genów pomiędzy liniami komórkowymi tego samego typu, ale pochodzącymi od różnych dawców. Wykorzystano tutaj dane TSS-Seq pochodzące z próbek pobranych od 26 pacjentów chorych na gruczolakoraka płuc. W tym zestawie danych zidentyfikowano łącznie aż 430 par genów nakładających się na końcach 5’ w przynajmniej jednej z bibliotek. Największa liczba 174 nakładających się par genów zidentyfikowana została dla biblioteki PC14 a najmniejsza liczba 60 par znaleziona została w linii H2347. Siedem spośród wszystkich zidentyfikowanych par genów ulegało ekspresji z nakładających się miejsc startu transkrypcji we wszystkich 26 liniach komórkowych.

Wszystkie etapy porównania opisane do tej pory w niniejszym podrozdziale obejmowały łącznie 72 ludzkie biblioteki TSS-Seq, w których nakładało się 580 par genów. Trzysta siedemdziesiąt z tych par genów została zidentyfikowana jedynie w jednym z trzech podzestawów opisanych powyżej, natomiast 77 par genów zostało zidentyfikowanych jako

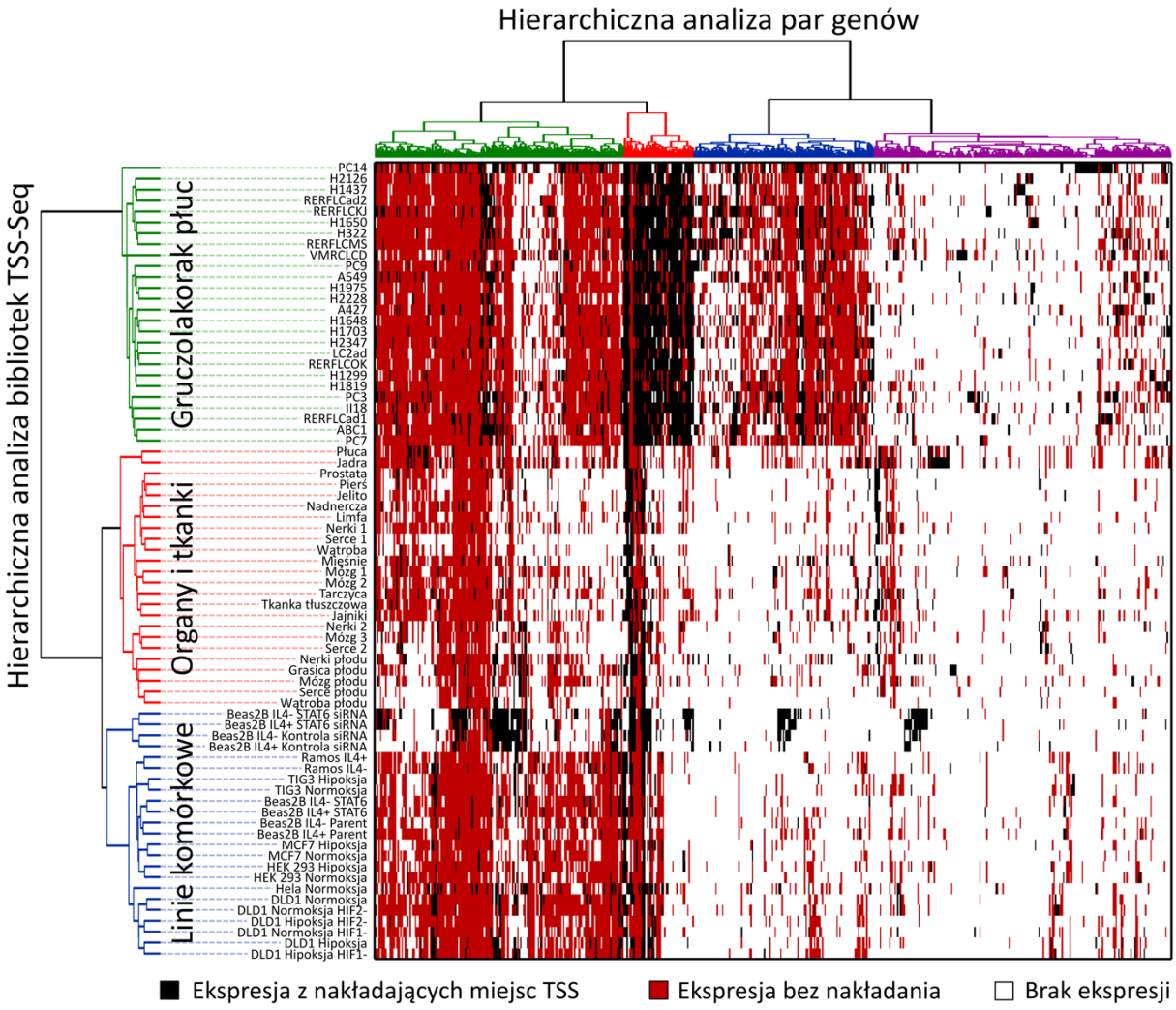
nakładające się co najmniej raz w każdym z trzech zestawów bibliotek (rycina 16). Dziesięć par genów ulegało jednoczesnej ekspresji we wszystkich 72 ludzkich bibliotekach, jednakże żadna z tych par nie ulegała ekspresji z nakładających się TSS we wszystkich z nich.



Rycina 16. ***Diagram Venna przedstawiający zależności pomiędzy analizowanymi zestawami danych****.*

Powyższej opisane wyniki pokazują, że wykorzystanie nakładających się bądź nienakładających promotorów zależy od wielu czynników. Z tego względu zbadano również które tkanki lub linie komórkowe są do siebie podobne pod względem tego jakie geny się w nich nakładają. W tym celu przeprowadzono hierarchiczną analizę skupień. Do analizy tej dodano również wyniki pochodzące z linii komórkowej Hela, która hodowana była tylko w jednych warunkach i nie została uprzednio dołączona do analizy porównawczej przeprowadzonej dla ludzkich linii komórkowych. Wynik dwuwymiarowej analizy skupień dla 582 nakładających się par genów przedstawiony został na rycinie 17.

Na rycinie tej widać, że biblioteki TSS-Seq pogrupowane zostały, w większości przypadków, zgodnie z rodzajem linii komórkowej, tkanki czy organu lub stadium rozwojowym. Zauważa się jednak pewne odstępstwa od tej reguły. Osiem linii komórkowych Beas2B podzielone zostało bowiem na dwa niezależne klastry, z czego jeden klaster zawiera transfekowane a drugi kontrolne linie komórkowe. W odniesieniu do hierarchicznej analizy na poziomie par genów wyodrębniony został klaster, zaznaczony na rycinie 17 kolorem czerwonym, zawierający 52 pary genów nakładających się w wielu bibliotekach gruczolakoraka. Geny te nie ulegają ekspresji w większości innych bibliotek.

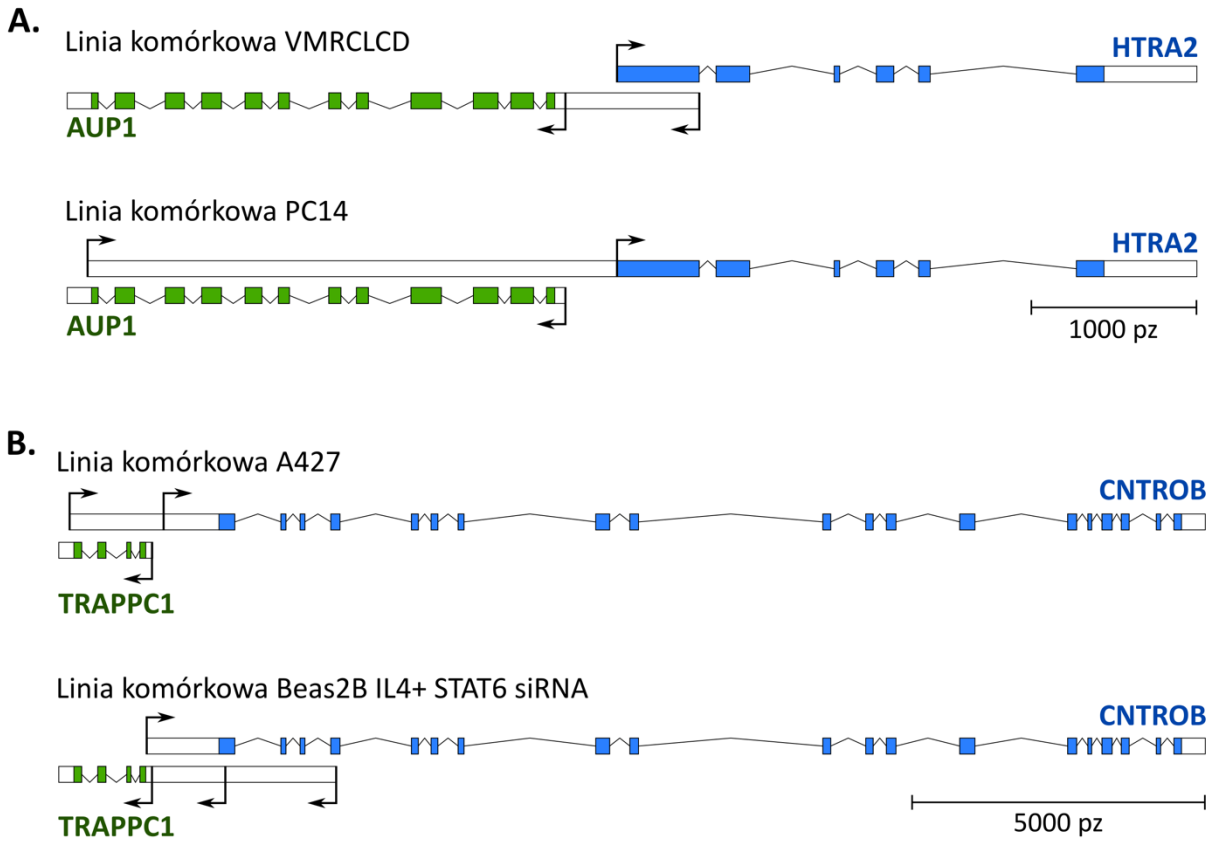


Rycina 17. ***Dwuwymiarowa hierarchiczna analiza skupień par genów nakładających.*** *Kolorem czerwonym zaznaczono ekspresję obu genów bez nakładania, kolorem czarnym ekspresję genów z nakładających się promotorów, podczas gdy kolorem białym zaznaczono brak ekspresji jednego lub obu genów z pary w danej bibliotece.*

5.2.3. Stopień nakładania się genów

Długość rejonu nakładania nie jest dla danej pary genów cechą stałą. W przypadku 300 spośród 695 par genów nakładających się u człowieka i myszy, rejon ten różni się między badanymi bibliotekami o więcej niż 100 pz, natomiast w przypadku 159 par genów ta różnica jest większa niż 1000 pz. Manualna analiza wyników wykazała, że pośród ludzkich par genów nakładających się znajduje się 85 par, których rejon nakładania końcami 5’ nie tylko różni się długością między poszczególnymi bibliotekami, ale jest na tyle przesunięty, że znajduje się w całkowicie innym obszarze. Jako doskonały przykład może tutaj posłużyć para genów *HTRA2* oraz *AUP1*, których nakładanie się zostało zidentyfikowane w pięciu tkankach o charakterze nowotworowym. W linii komórkowej gruczolakoraka płuc VMRCLCD gen *AUP1* wykorzystuje do ekspresji promotor zlokalizowany wewnątrz rejonu

kodującego genu *HTRA2,* tworząc w ten sposób rejon nakładania o długości 504 pz (rycina 18 A). Z kolei w bibliotece PC14 oba geny wykorzystują inne miejsca startu transkrypcji, przy czym promotor genu *HTRA2* zlokalizowany jest rejonie 3' UTR genu *AUP1*, co prowadzi do powstania rejonu nakładania o długości 2930 pz, który obejmuje prawie cały gen *AUP1*.



Rycina 18. ***Lokalizacja rejonu nakładania w zależności od wykorzystanych rejonów promotorowych****. Dla przejrzystości na rycinie przedstawiono tylko jeden wariant splicingowy dla każdego z genów.*

Kolejnym przykładem może być para genów *CNTROB* i *TRAPPC1*, gdzie rejon nakładania w linii komórkowej gruczolakoraka płuc A427 obejmuje całkowicie rejon kodujący genu *TRAPPC1* (rycina 18 B), natomiast w przypadku nabłonkowej linii komórkowej Beas2B IL4+ STAT6 siRNA, rejon nakładania znajduje się głównie wewnątrz genu *CNTROB*.

W przypadku obu opisanych par genów rejon nakładania przesunięty aż do 3'UTR obejmuje miejsce docelowe dla cząsteczki miRNA. Może to mieć bardzo istotne znaczenie regulatorowe przyjmując, iż transkrypty tych genów teoretycznie mogą tworzyć dupleksy RNA:RNA. Konsekwencją wykorzystania tak wysuniętego TSS mogłoby być maskowanie

sekwencji docelowej dla miRNA. Podobny mechanizm regulacji ekspresji może dotyczyć 27 genów, wymienionych w tabeli 3, których rejon 3' UTR, wraz z sekwencjami docelowymi dla miRNA, znajduje się w obszarze nakładania się genów.

Tabela 3. ***Lista 27 genów, których sekwencje docelowe dla miRNA mogą podlegać maskowaniu przez dupleks RNA:RNA****.*

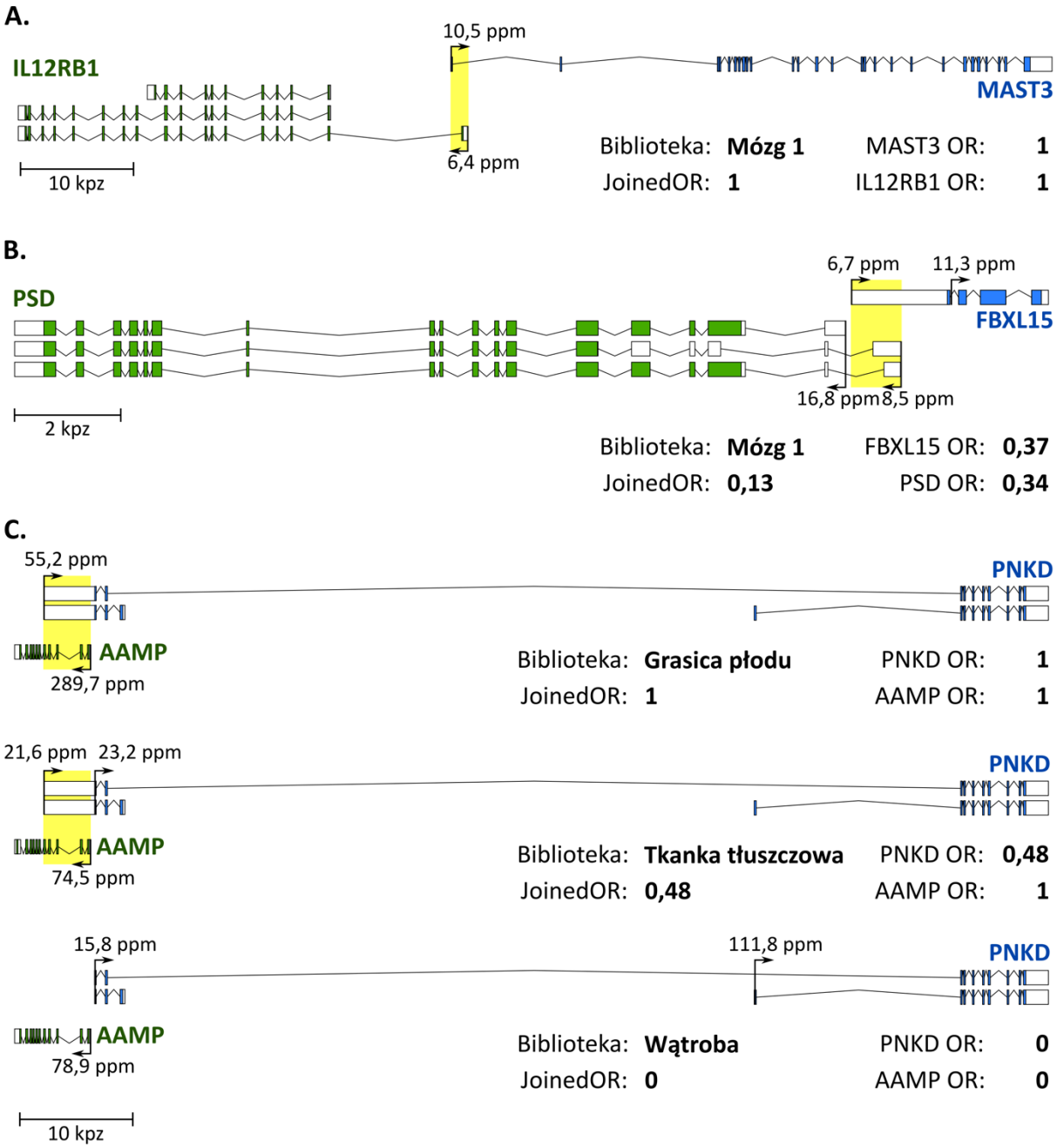
Lp.	**Nazwa genu:**
1.	ALYREF
2.	ATF5
3.	AUP1
4.	B3GALT6
5.	C11orf68
6.	C17orf89
7.	C5orf38
8.	CCDC85B
9.	CPTP

Lp.	**Nazwa genu:**
10.	CRYAB
11.	CYHR1
12.	DPH3
13.	DRAP1
14.	HSPB6
15.	MRPL49
16.	MRPS34
17.	MRPS7
18.	NDUFA2

Lp.	**Nazwa genu:**
19.	NDUFB11
20.	NOP10
21.	PSENEN
22.	RPS23
23.	RPS28
24.	TCTA
25.	TNNC1
26.	TRAPPC1
27.	YRDC

5.2.4. Stopień ekspresji z nakładających się miejsc TSS

Pośród ludzkich i mysich genów odpowiednio 493 i 96 par zawsze wykorzystuje po jednym miejscu startu transkrypcji na gen, gdy ekspresja zachodzi z nakładających się miejsc TSS. Jako przykład może tutaj posłużyć ludzka para genów *MAST3* oraz *IL12RB1*, których nakładanie się w mózgu dorosłego człowieka przedstawione zostało na rycinie 19 A. W przypadku 89 ludzkich i 13 mysich par genów, nakładające się miejsca startu transkrypcji przynajmniej w jednej bibliotece były wykorzystywane obok tych nienakładających się. Równoczesne użycie nakładających się i nienakładających miejsc TSS zostało przedstawione na rycinie 19 B dla przykładowej pary genów *FBXL15* i *PSD*. Ekspresja obu genów zachodzi w mózgu człowieka z wykorzystaniem dwóch miejsc TSS, jednakże tylko dalej wysunięte rejony promotorowe tworzą rejon nakładania.

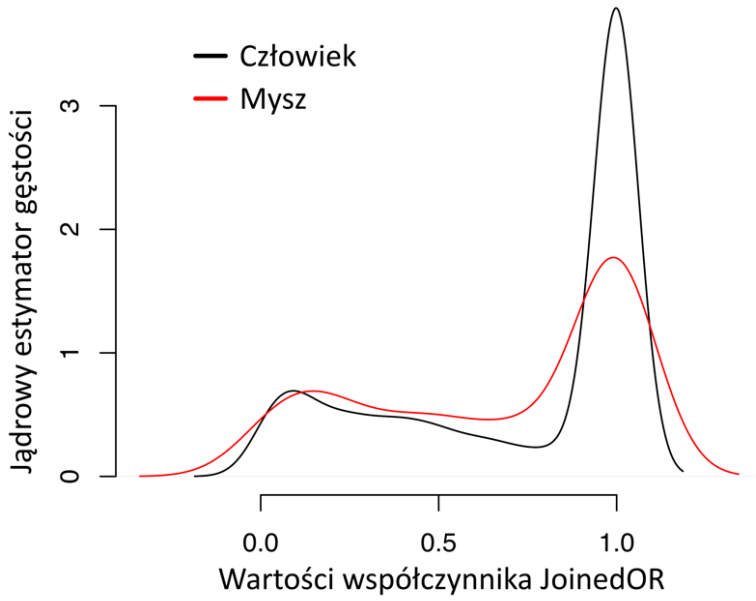


Rycina 19. ***Współczynniki OR oraz JoinedOR w kontekście zmienności wykorzystania rejonów promotorowych dla trzech par genów****. Kolorem żółtym podświetlony został rejon nakładania.*

Aby określić stopień w jakim transkrypcja inicjowana jest w rejonie nakładania, opracowano współczynniki OR oraz JoinedOR. Wyrażają one proporcję między poziomem ekspresji przypisanym do rejonu nakładania a całkowitym poziomem ekspresji odpowiednio genu lub pary genów. Im wartość obu współczynników bliższa jest wartości jeden, tym wyższy stopień ekspresji przypisany został do rejonu nakładania. Wartość 0 oznacza brak ekspresji z rejonu nakładania. W przypadku wspomnianej wcześniej pary genów *FBXL15* oraz *PSD*, przedstawionej na rycinie 19 B, poszczególne geny ulegały ekspresji inicjowanej

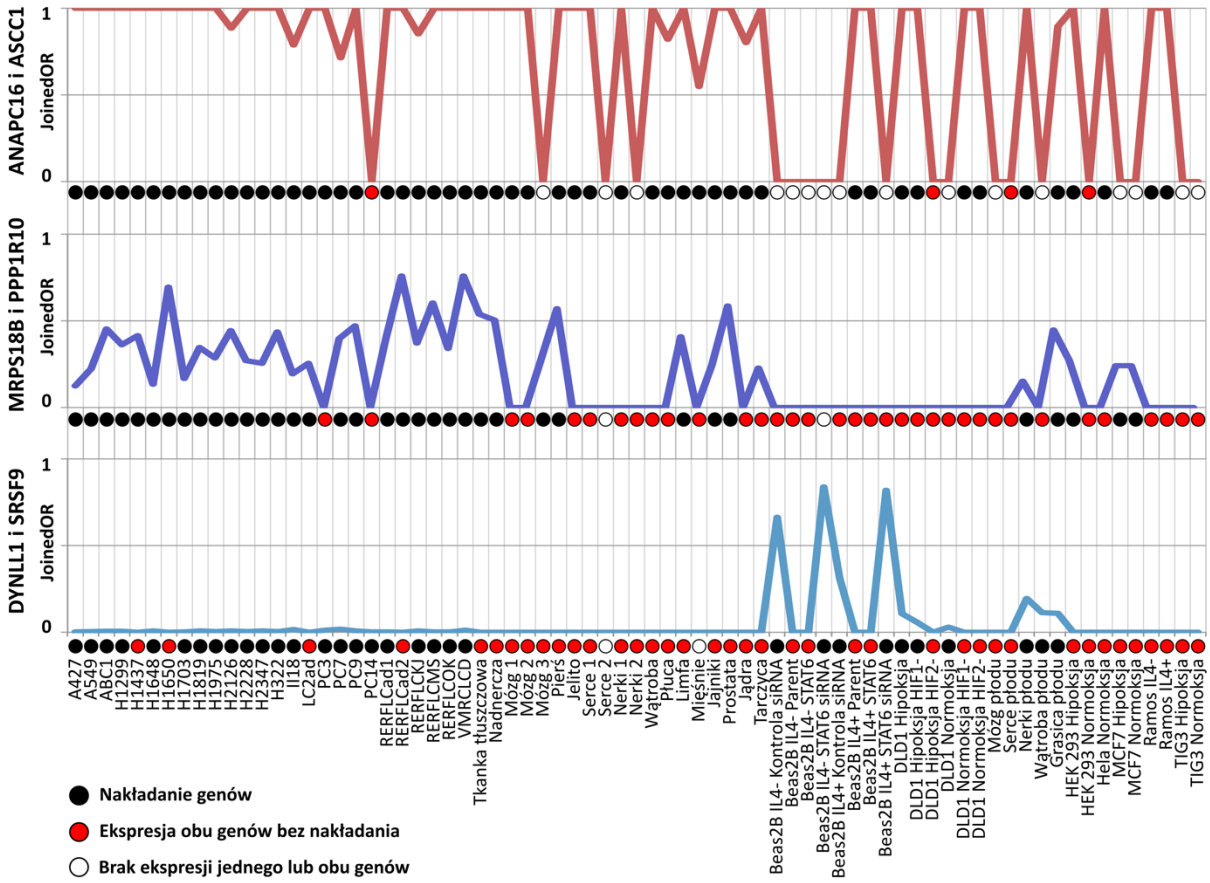
w rejonie nakładania odpowiednio w 37% i 34% (wartości współczynnika OR = 0,37 oraz 0,34). Wartość współczynnika JoinedOR otrzymuje się przez przemnożenie wartości współczynników OR obu genów w parze, przy założeniu jednakowego stopnia ich ekspresji. Założenie takie skutkuje obliczeniem maksymalnej proporcji ekspresji, która może być przypisana do rejonu nakładania a która dla genów *FBXL15* i *PSD* wyniosła 13%. Uwzględniwszy, że poziom ekspresji genów w parze może nie być jednakowy, oznacza to, że nie więcej niż 13% transkryptów w parze mogło być inicjowane wewnątrz nakładającego się rejonu. Kolejnym przykładem może być para genów *PNKD* i *AAMP,* przedstawiona na rycinie 19 C, która ukazuje różnorodne wykorzystanie przez te geny alternatywnych miejsc startu transkrypcji w różnych narządach lub tkankach. W grasicy płodu geny te wykorzystują po jednym, nakładającym się miejscu TSS i wszystkie inicjowane tutaj transkrypty wywodzą się z rejonu nakładania. W tkance tłuszczowej gen *PNKD* wykorzystuje do ekspresji dwa alternatywne rejony promotorowe i mniej niż połowa poziomu jego ekspresji przypisana jest do rejonu nakładania. W wątrobie gen *PNKD* nie wykorzystuje natomiast nakładających się promotorów i ekspresja obu genów w parze zachodzi bez nakładania.

Wdrożenie współczynnika JoinedOR pozwoliło na głębsze zrozumienie zjawiska nakładania się genów na końcach 5’ również w szerszym ujęciu. Spośród analizowanych genów, ekspresja aż 57,3% ludzkich i 44,3% mysich par nakładających się została zainicjowana wyłącznie w rejonie nakładania. W przypadku pozostałych par genów transkrypcja zachodzi zarówno z rejonu nakładania, jak i z poza niego. Rycina 20 pokazuje za pomocą jądrowego estymatora gęstości (ang. *kernel density estimation*) gęstość rozkładu wartości przyjmowanych przez współczynnik JoinedOR. Szczególnie u człowieka widać, że rozkład wartości JoinedOR jest rozkładem bimodalnym, z jednym szczytem usytuowanym dla wartości współczynnika JoinedOR równej jeden oraz drugim w okolicach wartości przybliżonej zeru i asymetrii dodatniej względem wartości 0,5. Oznacza to, że w przypadku gdy geny ulegają ekspresji z więcej niż jednego alternatywnego miejsca startu transkrypcji i nie wszystkie z nich się nakładają, preferowane jest wykorzystywane nienakładających się promotorów.



Rycina 20. ***Rozkład wartości współczynnika JoinedOR u człowieka i myszy****.*

Opracowane współczynniki mogą być użyte do zbadania dynamiki wykorzystania przez geny nakładających się miejsc TSS w różnych bibliotekach. Wzorce ekspresji trzech przykładowych par genów zostały graficznie zaprezentowane na rycinie 21. Pierwszy przykład ukazuje parę genów *ANAPC16* oraz *ASCC1*, które ulegają ekspresji głównie z nakładających się miejsc startu transkrypcji. Wyjątek stanowią tutaj ludzkie mięśnie, gdzie tylko mniej więcej połowa transkrypcji zainicjowana została w rejonie nakładania oraz cztery linie komórkowe, w których jednoczesna ekspresja obu genów nastąpiła z wykorzystaniem nienakładających się miejsc TSS. Druga para genów została zidentyfikowana jako ulegająca ekspresji w większości ludzkich tkanek i linii komórkowych, jednakże nakładająca się tylko w 37 z nich. Oscylacja wartości współczynnika JoinedOR wokół wartości 0,5 oznacza, że mniej więcej połowa transkrypcji inicjowana była w nienakładających się rejonach. Dodatkowo para ta nigdy nie osiąga wartości współczynnika równej 1, co oznacza, że w żadnej z bibliotek ekspresja nie była przypisana wyłącznie do rejonu nakładania. Para genów *DYNLL1* i *SRSF9* stanowi przykład, w którym nakładanie zostało wykryte w 32 bibliotekach, jednakże bardzo znikoma część ekspresji przypisana jest do nakładających się rejonów w większości z tych bibliotek. Jest to szczególnie widoczne dla 22 bibliotek gruczolakoraka płuc, gdzie wartość współczynnika JoinedOR nigdy nie przekracza 0,017.



Rycina 21. ***Wartości współczynnika JoinedOR trzech par genów w różnych bibliotekach****. Kolorem czerwonym oznaczono ekspresję obu genów z pary bez nakładania, kolorem czarnym ekspresję tych genów z nakładających się miejsc TSS, natomiast kolorem białym oznaczono brak ekspresji jednego lub obu genów z pary.*

5.3. Ekspresja genów nakładających się

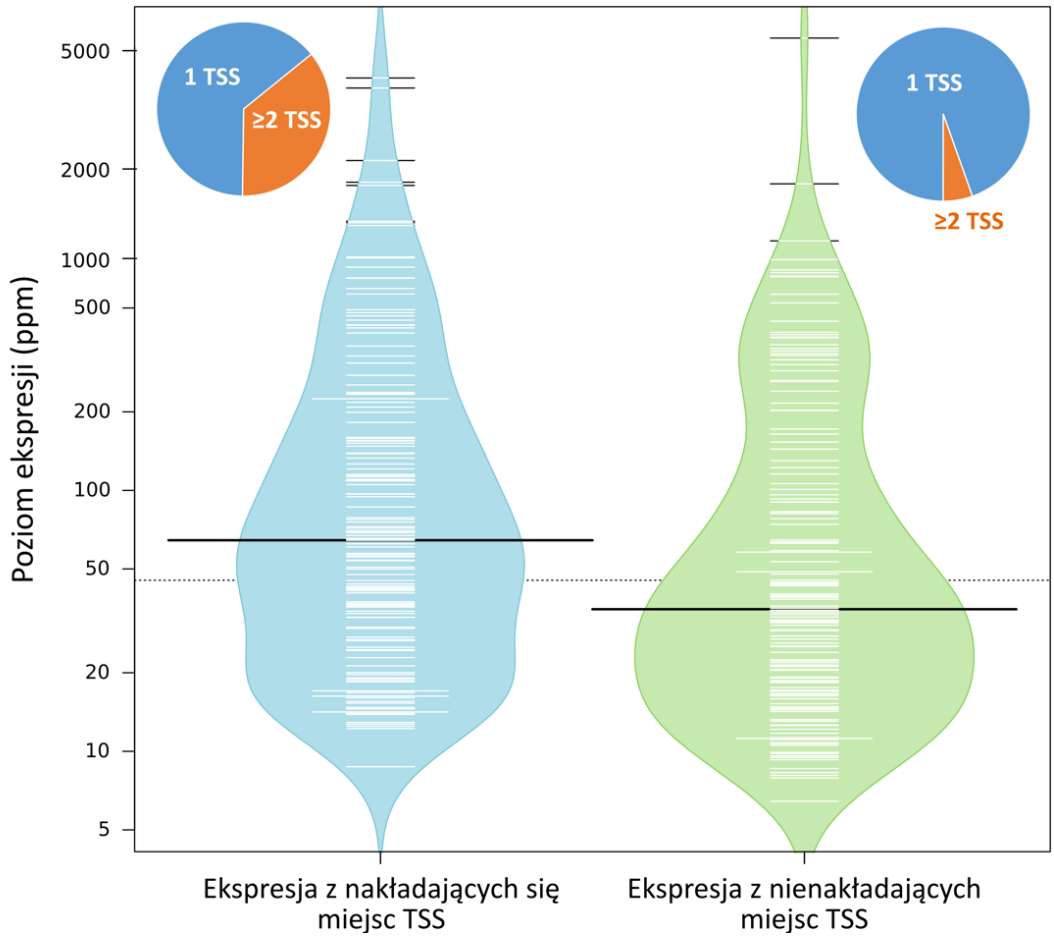
Poziom ekspresji genów wykorzystujących nakładające się miejsca startu transkrypcji osiąga średnią wartość 171 ppm, podczas gdy średnia ekspresja wszystkich innych genów, które w danej bibliotece nie wykorzystywały nakładających się miejsc startu transkrypcji wynosi 117 ppm. Różnica w poziomie ekspresji tych dwóch grup genów jest statystycznie istotna (test T-Studenta dla niezależnych średnich, wartość P = 3e-5). Wynik ten jest nieco zaskakujący biorąc pod uwagę funkcjonującą hipotezę zakłócenia transkrypcyjnego, które skorelowane zostało z obniżonym poziomem ekspresji genów nakładających[38]. Aby sprawdzić jaki dokładnie wpływ zjawisko nakładania ma na ekspresję genów, wyodrębniono 73 pary ludzkich genów, które ulegały ekspresji z wykorzystaniem nakładających się miejsc startu transkrypcji w co najmniej 10 bibliotekach i jednocześnie ulegały ekspresji z wykorzystaniem nienakładających się promotorów w kolejnych dziesięciu lub więcej bibliotekach. Następnie dla każdej z par genów obliczono dla ich ekspresji dwa

współczynniki korelacji Pearsona, pierwszy w bibliotekach, w których ekspresja zainicjowana została w rejonach nakładania, oraz drugi, w których ekspresja obu genów zachodziła z nienakładających się miejsc TSS. Trzydzieści sześć par wykazywało statystycznie istotny wynik dla co najmniej jednej z grup (tabela 4), tzn. współczynnika korelacji liczonego dla ekspresji genów gdy się one nakładają bądź gdy się nie nakładają. Pozytywna korelacja w przypadku nakładania się genów została zaobserwowana dla 29 par genów. W 8 przypadkach te same geny wykazywały także pozytywną korelację ekspresji w sytuacji gdy nie wykorzystywały nakładających się promotorów. Pozostałe 21 par nie wykazywało statystycznie istotnej korelacji w bibliotekach, w których ekspresja zachodziła z nienakładających się miejsc TSS. Cztery pary genów wykazały odwrotne tendencje, pozytywna korelacja zaobserwowana była jedynie w przypadku gdy geny się nie nakładały. Negatywna korelacja zaobserwowana została tylko w przypadku trzech par genów, spośród których para genów *PMPCA* i *SDCCAG3* wykazywała ją w przypadku ekspresji z nakładających promotorów, podczas gdy pary genów *GABPA* i *ATP5J*, oraz *CKS1B* i *SHC1* wykazały negatywną korelację ekspresji przy wykorzystaniu nienakładających miejsc TSS.

Tabela 4. ***Podsumowanie analizy korelacji ekspresji 73 par genów****. Korelacja pozytywna lub negatywna odnosi się do statystycznie istotnych wyników.*

Korelacja poziomu ekspresji w bibliotekach, gdzie transkrypcja zachodzi z:		Liczba par genów
rejonów nakładania	rejonów nienakładających	
pozytywna	brak korelacji	21
brak korelacji	pozytywna	4
pozytywna	pozytywna	8
negatywna	brak korelacji	1
brak korelacji	negatywna	2
brak korelacji	brak korelacji	37

Analiza poziomu ekspresji 73 par genów, gdy zachodzi ona z nakładających się miejsc startu transkrypcji wykazała, że jest ona średnio wyższa, aniżeli ekspresja tych samych par genów z wykorzystaniem nienakładających się miejsc startu transkrypcji (Test T-Studenta dla niezależnych średnich = -2,1; Wartość P = 0,03), co graficznie przedstawiono na rycinie 22. Różnica ta jest bardzo dobrze widoczna w przypadku mediany poziomu ekspresji, która dla genów wykorzystujących nakładające się miejsca startu transkrypcji jest prawie dwukrotnie wyższa, niż obserwuje się to w przypadku ekspresji tych samych genów z wykorzystaniem nienakładających się miejsc TSS.



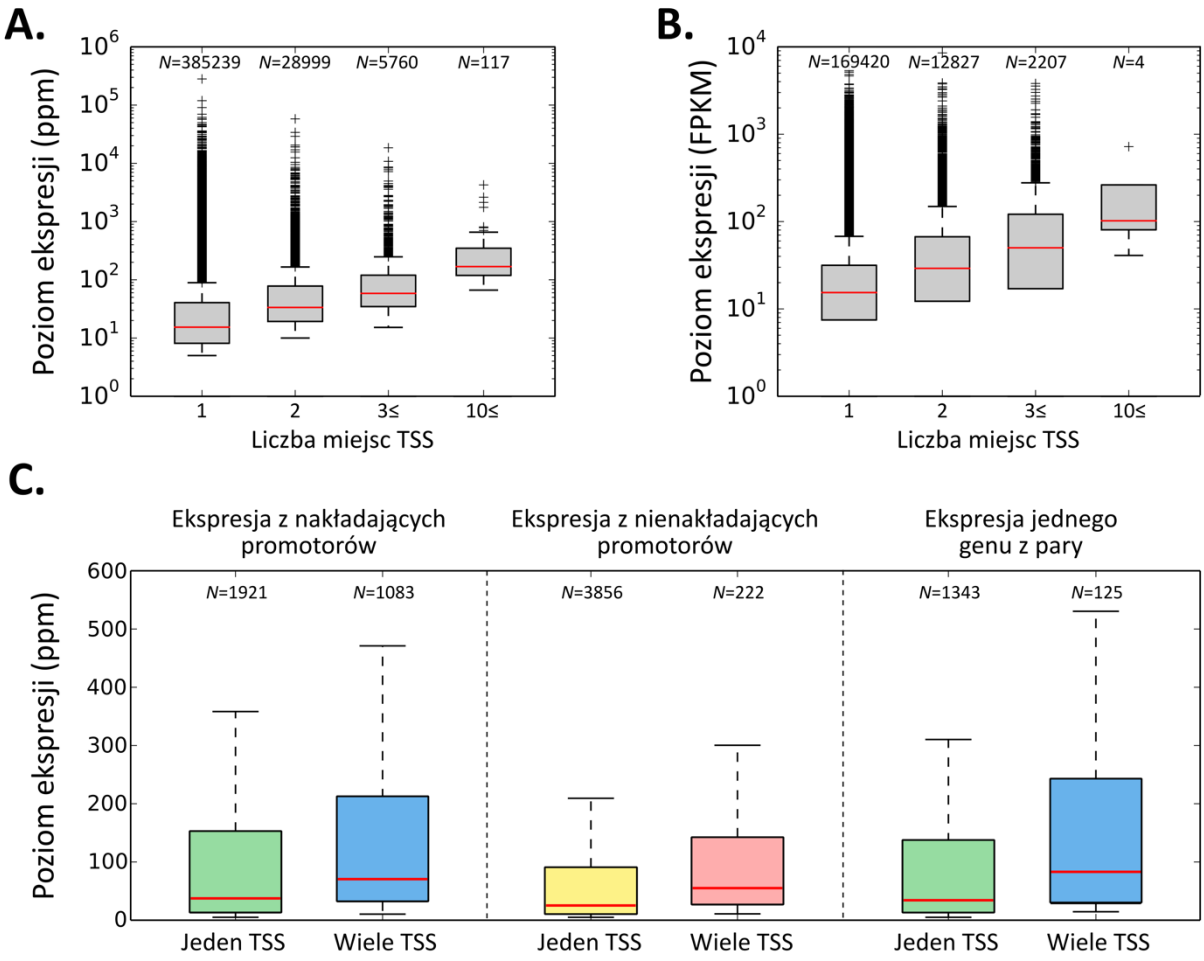
Rycina 22. ***Poziom ekspresji genów przy wykorzystaniu nakładających się lub nienakładających miejsc startu transkrypcji.*** *Pozioma linia koloru czarnego oznacza medianę. Dla każdej z grup przedstawiono również wykres kołowy pokazujący proporcję genów wykorzystujących do ekspresji jeden lub więcej rejonów promotorowych.*

Ponadto analiza wymienionych wyżej 73 par genów wykazała, że wykorzystanie nakładających się promotorów wiąże się nie tylko z przeciętnie wyższym poziomem ekspresji, lecz również z większą liczbą aktywnych rejonów promotorowych, co przedstawione zostało w postaci wykresów kołowych na rycinie 22. W celu sprawdzenia związku pomiędzy liczbą wykorzystywanych TSS, a poziomem ekspresji przeprowadzono analizę wszystkich piętnastu tysięcy ludzkich genów ulegających ekspresji przynajmniej w jednej z 73 bibliotek TSS-Seq. Analiza ta wyraźnie wskazuje na tendencję do wyższej ekspresji genów w bibliotekach, w których wykorzystanych jest więcej alternatywnych miejsc startu transkrypcji co zostało to zobrazowane na rycinie 23 A. Mediana poziomu ekspresji genów wykorzystujących w danej bibliotece jeden TSS wynosi 18 ppm i wzrasta wraz ze zwiększającą się liczbą aktywnych alternatywnych miejsc startu transkrypcji. Poziom ten rośnie do wartości przeciętnej 172 ppm dla genów wykorzystujących jednocześnie 10 i więcej

promotorów. Różnica poziomów ekspresji pomiędzy wszystkimi grupami jest istotna statystycznie przy przyjętej wartości prawdopodobieństwa testowego mniejszej niż 0,05, co zweryfikowane zostało testem Manna–Whitneya.

W celu potwierdzenia tych wyników została niezależnie wykonana analiza ekspresji genów nakładających się wykorzystując dane RNA-Seq. Ze względu na dostępność danych analiza ograniczona była do 26 bibliotek gruczolakoraka płuc. Odczyty poddane zostały kontroli jakości a następnie zostały zmapowane do ludzkiego genomu referencyjnego. Liczba odczytów zmapowanych oraz niezmapowanych dla każdej z bibliotek RNA-Seq została przedstawiona w tabeli dodatkowej 2 w aneksie. Poziomy ekspresji genów wyrażone zostały w znormalizowanych jednostkach FPKM. Jak pokazuje rycina 23 B, mediana poziomu ekspresji genów posiadających jeden aktywny promotor wynosi 15,39 FPKM. Zgodnie z wynikami otrzymanymi na podstawie analizy danych TSS-Seq, mediana także w tym przypadku rośnie wraz z rosnącą liczbą alternatywnych miejsc startu transkrypcji.

Aby ustalić istotność tego czynnika w kontekście poziomu ekspresji wyselekcjonowanych wcześniej 73 par genów nakładających się u człowieka, pary genów podzielono na trzy grupy w zależności od tego czy geny ulegały ekspresji z nakładających się czy nienakładających miejsc TSS lub czy odnotowywana była ekspresja tylko jednego z genów w parze. Dodatkowo każda z trzech grup została podzielona na dwie podgrupy biorąc pod uwagę to, czy dany gen wykorzystał do ekspresji jeden czy więcej miejsc TSS. Poziom ekspresji genów w poszczególnych grupach został porównany z wykorzystaniem testu Manna–Whitneya. Otrzymane wyniki pokazują (rycina 22 C), że niezależnie od tego czy pary genów ulegają ekspresji z nakładających się czy nienakładających miejsc TSS, średni poziom ekspresji genów wykorzystujących więcej alternatywnych promotorów jest zawsze wyższy niż tych używających tylko jeden promotor co potwierdza wyniki uzyskane podczas analizy wszystkich genów. Nie zaobserwowano natomiast różnicy poziomu ekspresji pomiędzy genami, które wykorzystują nakładające się miejsca startu transkrypcji a genami, których ekspresja zachodziła bez ekspresji drugiego genu z pary. Ekspresja przy wykorzystaniu nienakładających się miejsc startu transkrypcji odznacza się statystycznie niższym poziomem w porownaniu do dwóch pozostałych grup zarówno, gdy ta zachodzi z użyciem jednego, jak i wielu miejsc TSS.



Rycina 23. ***Poziom ekspresji genów w zależności od liczby wykorzystanych miejsc TSS****. (A) poziom ekspresji genów wykorzystujących różną liczbę miejsc TSS na podstawie danych TSS-Seq. (B) poziom ekspresji genów obliczony w oparciu o dane RNA-Seq lecz posegregowane w zależności od liczby wykorzystywanych promotorów według danych TSS-Seq. (C) Poziom ekspresji genów pochodzących z analizy 73 par genów ulegających ekspresji z wykorzystaniem nakładających się miejsc TSS w przynajmniej 10 bibliotekach i ekspresji z użyciem nienakładających się miejsc TSS w kolejnych 10 lub więcej bibliotekach. Tym samym kolorem oznaczono wykresy pudełkowe, dla których nie wykazano statystycznie istotnej różnicy (test Manna-Whitneya, wartość P < 0,05).*

Do potwierdzenia tych wyników ponownie wykorzystane zostały dane RNA-Seq. Z uwagi na ograniczoną do 26 liczbę dostępnych bibliotek, jedynie 17 spośród tych par genów spełniło warunek ulegania ekspresji z nakładających się TSS w co najmniej 10 bibliotekach i w kolejnych minimum 10 z nienakładających. Średni poziom ekspresji genów, które według danych TSS-Seq ulegały ekspresji z rejonu nakładania ponownie okazał się istotnie wyższy aniżeli poziom ekspresji tych samych genów z wykorzystaniem nienakładających się promotorów (test t-studenta dla niezależnych średnich = 2,65; wartość P = 0,008).

5.4. Monoalleliczność ekspresji genów nakładających

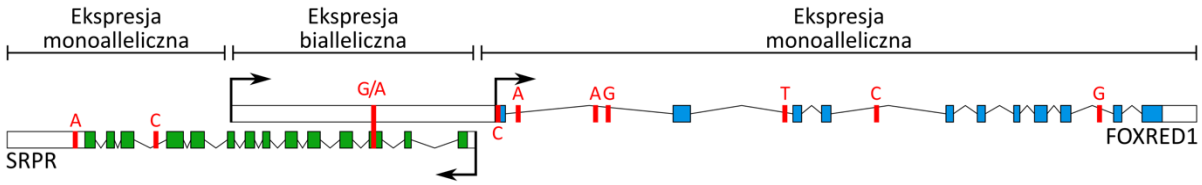
Do niedawna przyjmowało się, że geny na chromosomach autosomalnych ulegają ekspresji z wykorzystaniem obu alleli. Okazuje się jednak, że stabilna ekspresja na obu chromosomach zachodzi najczęściej wśród genów homeotycznych[213, 214], podczas gdy pozostałe geny mogą być na różnych allelach regulowane w nieco odmienny sposób, co prowadzić może do ich monoallelicznej ekspresji. W przypadku pary genów nakładających monoalleliczna ekspresja na różnych chromosomach mogłaby oznaczać swojego rodzaju „ucieczkę" przez zjawiskiem interferencji transkrypcyjnej. Aby przetestować tą hipotezę przeanalizowano dostępne dane z wysokoprzepustowego sekwencjonowania transkryptomów dwudziestu sześciu linii komórkowych gruczolakoraka płuc pod kątem allelicznie specyficznej ekspresji genów nakładających się. Podczas analiz osobno rozpatrywano charakter ekspresji rejonu nakładania i osobno nienakładających się fragmentów genów. W związku z brakiem informacji o sekwencjach genomowych osób, od których pobrany został materiał, w pełni informatywne były jedynie bialleliczne sygnały pochodzące z transkryptów. W przypadku monoallelicznego sygnału niemożliwym jest, w oparciu o same transkrypty, określenie czy wynika to z ekspresji ograniczonej do jednego chromosomu czy też z braku polimorfizmu w badanym rejonie. O potencjalnie możliwej monoallelicznej ekspresji danej pary genów, w której każdy gen transkrybowany jest z innego homologicznego chromosomu, świadczyć może monoalleliczny sygnał z nienakładających się części obu genów i bialleliczny z rejonu nakładania się. Kryteria te co prawda nie wykluczają biallelicznośći, ale pozwalają na wytypowanie potencjalnych kandydatów.

Analiza wzorców ekspresji dla obu genów z pary możliwa była do zbadania jedynie dla 103 par dla których dostępny był sygnał w każdym z wydzielonych rejonów. Spośród tych par w przypadku 93 odnotowany był sygnał świadczący o biallelicznośći w jednym lub obu nienakładających się fragmentach genów. W przypadku pozostałych 10 par genów przedstawionych w tabeli 5, u co najmniej jednego dawcy, miejsca SNP znajdujące się w rejonie nakładania wykazywały charakter bialleliczny, podczas gdy w pozostałych częściach obu genów nie wykryto sygnałów polimorficznych. Jako przykład może tutaj posłużyć para genów *FOXRED1* i *SRPR,* którą zwizualizowano na rycinie 24. W rejonie nakładania zmapowanych zostało 391 odczytów zawierających w badanej pozycji nukleotyd A oraz 116 odczytów zawierających nukleotyd G. Stosunek poszczególnych wariantów wynosi więc około 3:1 i takiego właśnie stosunku można by się spodziewać w sytuacji monoallelicznej ekspresji, gdyż odpowiada on stosunkowi ekspresji tych genów. Dla genu *SRPR* wynosi ona 121 FPKM, a dla genu *FOXRED1* 45 FPKM. Dane te sugerują,

iż gen *SRPR* ulega ekspresji z chromosomu zawierającego allel „A”, natomiast gen *FOXRED1* z chromosomu zawierającego allel „G”.

Tabela 5. ***Lista par genów o monoallelicznej ekspresji obu genów w parze przy jednoczesnym biallelicznym sygnale w rejonie nakładania.***

Para genów	Linie komórkowe gruczolakoraka
ACADVL i *DLG4*	H1648, RERFLCKJ
CCT5 i *FAM173B*	RERFLCad1
FAM120A i *FAM120AOS*	H1703
FOXRED1 i *SRPR*	LC2ad, RERFLCMS
HNRNPH3 i *PBLD*	H2228
IKBKG i *G6PD*	II18, VMRCLCD
NAA15 i *NDUFC1*	H1648
PFKM i *SENP1*	H1819
POLR2A i *ZBTB4*	ABC1
SLC38A6 i *TRMT5*	H2126



Rycina 24. ***Przykładowa para genów FOXRED1 i SRPR o biallelicznym charakterze ekspresji w rejonie nakładania przy jednoczesnej monoallelicznej ekspresji obu genów z pary.*** *Dla uproszczenia na rycinie przedstawiono tylko po jednej, najdłuższej formie splicingowej na gen.*

5.5. Czynniki transkrypcyjne potencjalnie regulujące miejsca TSS nakładających się genów

Analiza danych TSS-Seq pozwoliła na identyfikację 582 ludzkich par genów kodujących białka nakładających się w co najmniej jednej bibliotece. Spośród nich w przypadku aż 98% par w przynajmniej jednej innej bibliotece ekspresji ulegał tylko jeden gen z danej pary. U myszy dotyczyło to 95% spośród 113 par genów nakładających się. Sugeruje to, że geny te mogą być regulowane przez niezależne promotory, nie zaś klasyczne promotory dwukierunkowe, dla których liczne badania wykazywały ko-ekspresję genów przez nie regulowanych[215], a które znane są z możliwości inicjacji nakładających się na końcach 5’ transkryptów[160]. Wykorzystanie przez gen określonych miejsc startu transkrypcji jest bezpośrednim odzwierciedleniem aktywności alternatywnych promotorów tego genu,

która to aktywność jest pozytywnie lub negatywnie regulowana między innymi dzięki czynnikom transkrypcyjnym wiążącym specyficzne motywy sekwencji DNA w rejonie promotorowym[100, 216]. Aby sprawdzić, czy promotory nakładających się genów kodujących białka noszą znamiona promotorów dwukierunkowych, lub czy są regulowane w sposób podobny do jednokierunkowych promotorów nienakładających się genów, przeprowadzono analizę miejsc wiązania czynników transkrypcyjnych w promotorach ludzkich i mysich genów nakładających się.

Identyfikacja miejsc wiązania czynników transkrypcyjnych wykonana została dla promotorów wszystkich genów człowieka i myszy, w tym 4 075 i 518 promotorów genów nakładających się, odpowiednio u człowieka i myszy, oraz 42 202 ludzkich i 21 042 mysich promotorów pozostałych genów. Następnie, na podstawie danych TSS-Seq określono, które z czynników ulegają ekspresji w badanych bibliotekach. W efekcie zidentyfikowano 278 ludzkich i 71 mysich czynników transkrypcyjnych. Różnica pomiędzy człowiekiem i myszą wynika najprawdopodobniej z faktu, że całkowita liczba zdeponowanych w bazie JASPAR motywów wiązania czynników transkrypcyjnych była dużo wyższa dla człowieka (367 czynniki) niż dla myszy (141 czynniki). W przypadku człowieka, analiza wykazała dużo większe zróżnicowanie miejsc wiązania czynników transkrypcyjnych w rejonach promotorów genów nakładających się, przeciętnie dla 46 różnych czynników, w porównaniu z pozostałymi genami, średnio dla 24 różnych czynników. U myszy nie odnotowano podobnej tendencji i średnia liczba czynników transkrypcyjnych potencjalnie regulujących promotory wynosi około 12 zarówno dla genów nakładających jak i wszystkich innych genów.

Najczęściej występującymi czynnikami transkrypcyjnymi są ludzki *GATA2* oraz mysi *Gata1* należące do tej samej rodziny. Miejsca wiązania tych czynników zidentyfikowane zostały w okolicy prawie wszystkich ludzkich i mysich rejonów promotorowych genów nakładających się. Liczebność motywów wiązania tych czynników pośród promotorów genów nakładających porównana została z ich liczebnością w promotorach wszystkich innych genów. W przypadku człowieka wykazano tutaj statystycznie istotną nadreprezentację miejsc wiązania czynnika *GATA2* w okolicach miejsc TSS genów nakładających. Czynnik transkrypcyjny *GATA2* nie jest jedynym, którego nadreprezentację odnotowano wśród czynników potencjalnie regulujących miejsca TSS genów nakładających się. U człowieka nadreprezentowane było łącznie 256 czynników transkrypcyjnych a żaden z czynników nie był niedoreprezentowany (tabela dodatkowa 3 w aneksie). U myszy 12 i 18 czynników transkrypcyjnych odznacza się odpowiednio statystycznie istotnym podwyższeniem lub obniżeniem występowania w stosunku do rejonów promotorowych

genów nienakładających (tabela dodatkowa 4 w aneksie). Wiele nadreprezentowanych czynników transkrypcyjnych może potencjalnie odpowiadać za regulację znacznej części nakładających się promotorów. Wykazano przykładowo, że więcej niż 80% promotorów ludzkich genów nakładających się może być potencjalnie regulowane przez czynniki takie jak *FOXC1, KLF5, MEIS1, MZF1, NFIX* lub *SPI1* (tabela dodatkowa 3 w aneksie). Ponadto miejsca wiązania czynników transkrypcyjnych *E2F3*, *EBF1, ERG*, *GABPA* i *TCF3*, zidentyfikowane zostały jako nadreprezentowane u obu gatunków a dalsze badania wykazały, że czynniki te mogą być zaangażowane w regulację nawet połowy genów nakładających.

Wśród nadreprezentowanych czynników transkrypcyjnych regulujących geny nakładające znalazły się takie czynniki, które w źródłach literaturowych zidentyfikowane zostały jako nadreprezentowane również w promotorach dwukierunkowych[95, 168, 170, 217, 218]. W tabeli 6 przedstawiono listę takich właśnie ludzkich czynników transkrypcyjnych u człowieka.

Tabela 6. ***Lista czynników transkrypcyjnych których miejsca wiązania nadreprezentowane są wśród promotorów genów nakładających się oraz promotorów dwukierunkowych****. Liczba promotorów genów nakładających się wynosi 4 075, natomiast liczba promotorów wszystkich innych genów wynosi 42 202.*

Czynnik transkrypcyjny	% promotorów posiadających motyw wiązania wśród:		Wartość statystyki Z	Krotność różnicy
	genów nakładających	genów nienakładających		
ETS1	98,97%	68,72%	40,80	1,44
YY1	92,29%	50,30%	51,37	1,84
SP1	80,39%	45,13%	43,02	1,78
E2F1	47,12%	19,71%	40,24	2,39
E2F4	39,95%	16,50%	36,77	2,42
NRF1	29,28%	13,74%	26,44	2,13
SP3	22,28%	9,39%	25,62	2,37
STAT1	13,01%	5,27%	19,95	2,47
GABPA	5,25%	1,55%	16,65	3,39
ZNF143	0,86%	0,13%	9,99	6,47

Na szczególną uwagę zasługuje czynnik transkrypcyjny *GABPA*, który według badań przeprowadzonych przez Collinsa i współpracowników[168] reguluje więcej niż 80% dwukierunkowych promotorów. Obecność motywów wiązania tego czynnika odnotowana została w relatywnie małej liczbie ludzkich promotorów genów nakładających się,

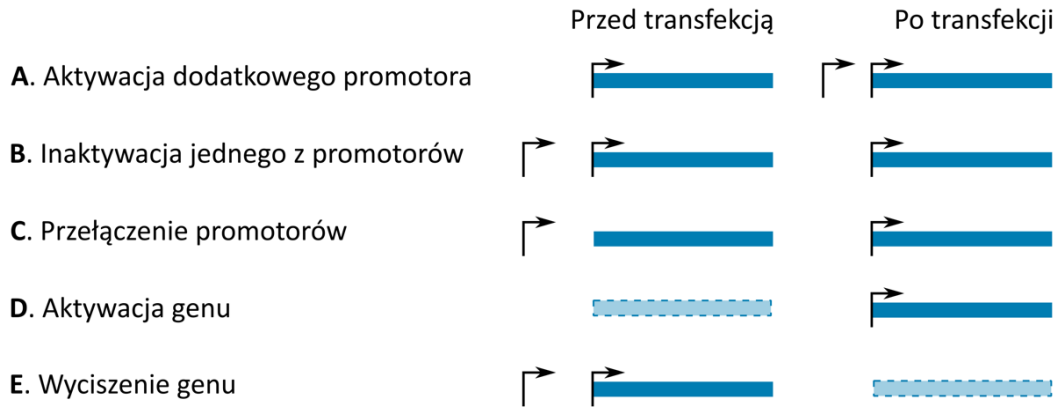
czyli 214 spośród 4075. Promotory te należą jednak do 146 genów i mogą odpowiadać za regulację dwukierunkowej transkrypcji aż 88 ludzkich par genów nakładających się. U myszy miejsca wiązania czynnika transkrypcyjnego *Gabpa* również zidentyfikowane zostały jako nadreprezentowane wśród promotorów genów nakładających się. Czynnik ten może być zaangażowany w regulację 53 par genów, czyli aż 47% par genów nakładających się u myszy. *Gabpa* nie ulega ekspresji w trzech mysich organach: śledzionie, grasicy i mózgu. Dwa pierwsze z tych organów odznaczają się najniższą liczbą zidentyfikowanych genów nakładających się u myszy.

Powyższe wyniki sugerują, że przynajmniej część ze zidentyfikowanych ludzkich i mysich par genów nakładających się może być regulowana przez promotory dwukierunkowe. Niemniej jednak nadreprezentacja znacznej liczby miejsc wiązania czynników transkrypcyjnych, które do tej pory nie były skojarzone z promotorami dwukierunkowymi, może sugerować, że promotory genów nakładających się podlegają dość specyficznej regulacji.

5.6. Wpływ transfekcji na aktywność rejonów promotorowych

5.6.1. Aktywacja i inaktywacja promotorów w wyniku transfekcji

Analiza wpływu transfekcji na aktywność rejonów promotorowych odbyła się w oparciu o sześć linii komórkowych Beas2B, pośród których cztery linie poddane zostały transfekcji z wykorzystaniem celowanego bądź niecelowanego wyciszania, natomiast dwie były kontrolą negatywną, nie będąc liniami poddanymi transfekcji. W oparciu o dane TSS-Seq ustalono, że 6471 genów ulega ekspresji przynajmniej w jednej z tych sześciu bibliotek Beas2B. Pośród nich znaleziono 1335 genów, dla których w odpowiedzi na transfekcję konsekwentnie zachodzi zmiana wzorców wykorzystania promotorów. W zależności od wzorców ekspresji, geny te podzielić można na pięć kategorii przedstawionych na rycinie 25.



Rycina 25. ***Możliwe wzorce wykorzystania promotorów przy porównaniu przed transfekcją i po transfekcji.***

Pierwsze dwa wzorce obejmują geny, które obok promotora aktywnego zarówno przed jak i po transfekcji, aktywowały dodatkowe promotory lub inaktywowały jeden lub więcej z promotorów aktywnych przed transfekcją (rycina 25 A i B). Kolejne wzorce odpowiadają przełączeniu ekspresji genu na alternatywny TSS (rycina 25 C) oraz aktywacji bądź inaktywacji wszystkich rejonów promotorowych, prowadząc odpowiednio do ekspresji genu lub jej wyciszenia (rycina 25 D i E). Największa liczba genów przypisana została do dwóch ostatnich kategorii. W sumie aktywacja lub inaktywacja w wyniku transfekcji dotyczyła 1188 genów (tabela 7). Najrzadszym zjawiskiem jest natomiast przełączenie ekspresji genu na alternatywny promotor, co miało miejsce jedynie w przypadku 18 genów. Ponieważ zmiany opisane przez te scenariusze dotyczą aż 20% genów ulegających ekspresji w bibliotekach Beas2B, pojawia się więc pytanie, czy reakcja taka jest związana z genami, pełniącymi w komórkach jakieś szczególne funkcje oraz jaki wpływ na ich poziom ekspresji miała zaobserwowana zmiana wykorzystania promotorów.

Tabela 7. ***Liczba genów zmieniających w odpowiedzi na transfekcję wzór wykorzystania promotorów.*** *Dla genów ulegających ekspresji przed i po transfekcji przedstawiono dodatkowo podsumowanie wyników analizy ekspresji różnicowej.*

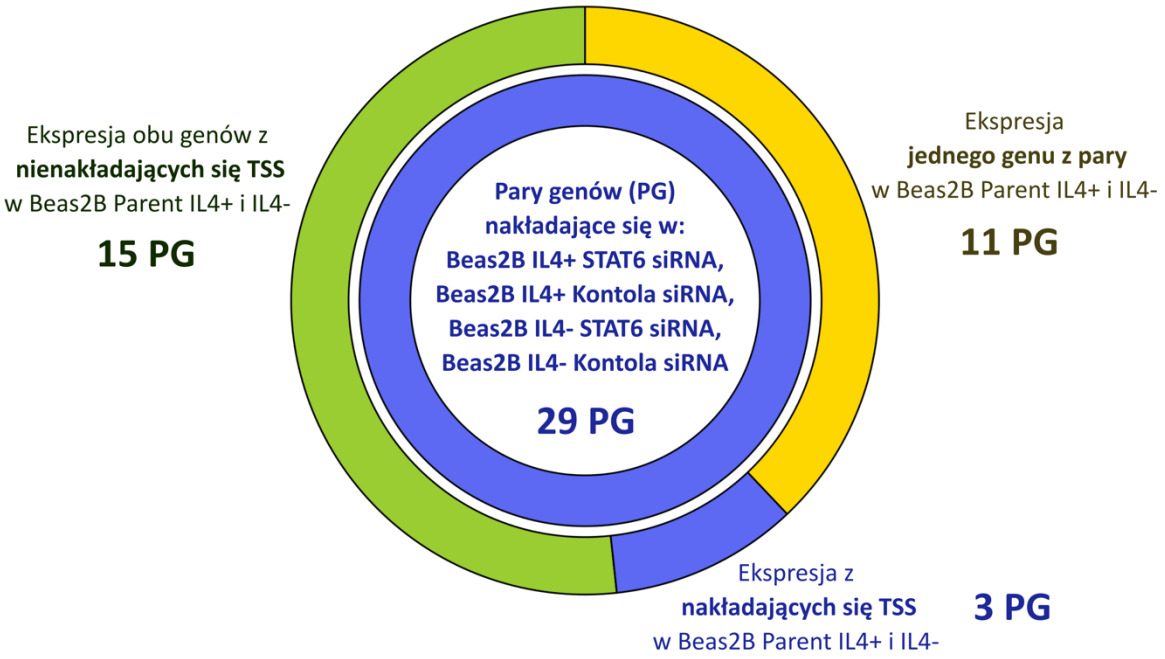
	Całkowita liczba genów		Wpływ zmiany na poziom ekspresji genów		
			Zwiększenie	Zmniejszenie	Brak zmian
Aktywacja dodatkowego promotora	43	w tym	23	5	15
Inaktywacja jednego z promotorów	86		2	36	48
Przełączenie promotorów	18		14	0	4
Aktywacja genu	130				
Wyciszenie genu	1058				

Aby odpowiedzieć na powyżej postawione pytania, w pierwszej kolejności przeprowadzono analizę ekspresji różnicowej dla 147 genów, które w odpowiedzi na transfekcję aktywowały dodatkowy TSS, inaktywowały jeden z promotorów lub przełączyły promotory na alternatywne. Analiza wykazała statystycznie istotną różnicę w poziomie ekspresji 80 z tych genów. Czternaście spośród osiemnastu genów, które w odpowiedzi na transfekcję zmieniły promotor na alternatywny, odznaczyła się podwyższonym poziomem ekspresji (tabela 7). Aktywacja dodatkowych miejsc TSS przyniosła ze sobą zwiększenie jej poziomu u nieco ponad połowy genów, natomiast inaktywacja któregoś z rejonów promotorowych wiązała się w przypadku 36 genów ze statystycznie istotnym obniżeniem poziomu ekspresji.

Analiza funkcjonalna przeprowadzona została z wykorzystaniem bazy Panther (wersja 11)[219] dla poszczególnych grup z ryciny 25. Analiza ta dla grup A-D nie wykazała w tych grupach statystycznie istotnej nad- lub niedoreprezentacji genów pełniących określone funkcje w którejkolwiek z domen ontologii genów (GO; z ang. *Gene Ontology*)[220]. Natomiast w grupie genów, których ekspresja w odpowiedzi na transfekcję została wyciszona (rycina 25 E) zidentyfikowano pewne statystycznie istotne wyniki analizy funkcjonalnej. Najciekawszym wydaje się być to, że pośród inaktywowanych w odpowiedzi na transfekcję genów niedoreprezentowane były geny związane z odpowiedzią immunologiczną (GO:0006955), odpowiedzią na stymulację komórkową (GO:0050896), percepcję sensoryczną bodźców chemicznych (GO:0007606) oraz geny ścieżki sygnałowej związanej z aktywnością receptorów sprzężonych z białkami G (GO:0007186). Pośród nadreprezentowanych znajdowały się natomiast geny odpowiedzialne za organizację transkrypcji zależną od DNA (GO:0006351), procesy metaboliczne RNA (GO:0016070), organizację cytoszkieletu (GO:0007010) i organelli komórkowych (GO:0006996).

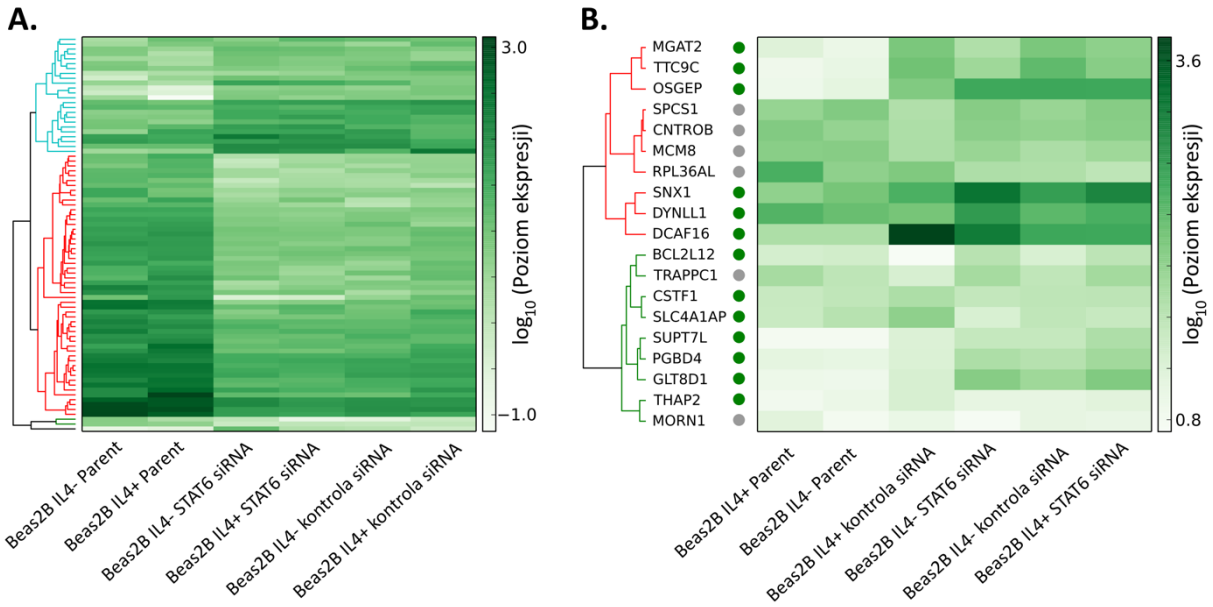
5.6.2. Wpływ transfekcji na wykorzystanie nakładających się promotorów

Problematyka wpływu transfekcji na wykorzystanie nakładających się miejsc startu transkrypcji została już wspomniana w rozdziale 5.2.2, gdzie wśród poddanych transfekcji komórkach nabłonkowych Beas2B zauważono zwiększoną liczbę nakładających się genów. Spośród wszystkich 167 par genów nakładających się przynajmniej w jednej poddanej transfekcji linii komórkowej Beas2B, szczególne zainteresowanie wzbudziło 29 par genów, które we wszystkich transfekowanych liniach wykorzystywały nakładające się miejsca startu transkrypcji (rycina 26).



Rycina 26. ***Pary genów wykazujące konsekwentny wzorzec ekspresji przed i po transkrypcji****. Wewnętrzny krąg wykresu, oznaczony kolorem niebieskim reprezentuje 29 par genów ulegających nakładaniu we wszystkich liniach komórkowych Beas2B poddanych transfekcji (celowanej lub niecelowanej).*

Trzy spośród tych par nakładało się również w liniach Beas2B niepoddanych transfekcji. W przypadku kolejnych 11 par genów nakładających się we wszystkich czterech bibliotekach Beas2B poddanych transfekcji, brak nakładania w dwóch liniach komórkowych będących kontrolą do transfekcji związany był z ekspresją tylko jednego genu z pary. Pozostałe 15 par genów to takie, w których odnotowano konsekwentną zmianę wykorzystania alternatywnych miejsc startu transkrypcji z nienakładających się przed transfekcją, do nakładających się po transfekcji. Związane było to z wykorzystaniem alternatywnych miejsc TSS przez 19 genów. Zmiany tego typu mogły być spowodowane zmianami w poziomie ekspresji czynników transkrypcyjnych w odpowiedzi na stres transfekcji. Aby przetestować tą hipotezę przeprowadzono analizę ekspresji różnicowej czynników transkrypcyjnych. Badanie to wykazało statystycznie istotny wzrost ekspresji 25 oraz spadek ekspresji kolejnych 56 czynników w reakcji na transfekcję (rycina 27 A, tabela dodatkowa 5 w aneksie). Ponadto aktywacja nakładających się promotorów w przypadku 13 spośród 19 genów wiązała się ze statystycznie istotnym wzrostem poziomu ekspresji. Ekspresja pozostałych 6 genów, pozostała na niezmienionym poziomie (rycina 27 B).

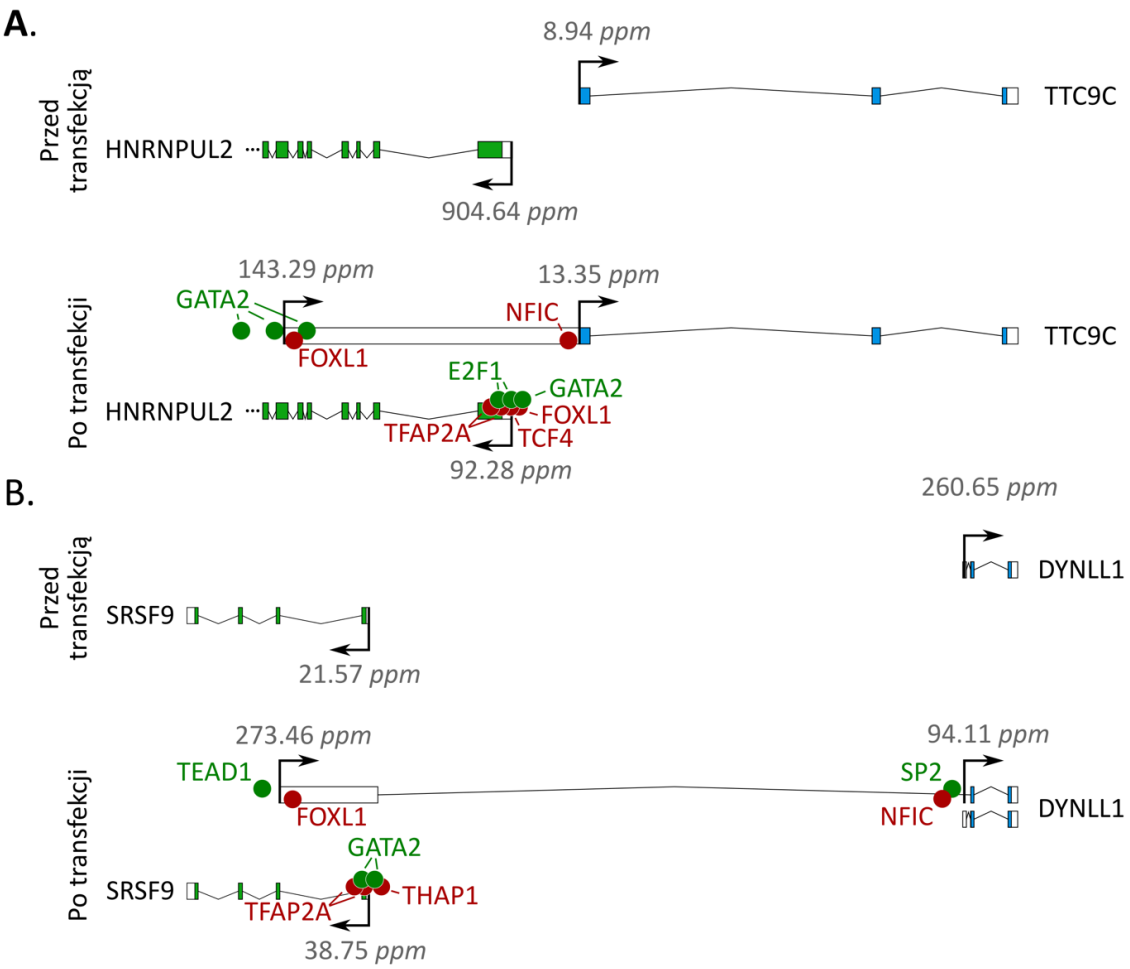


Rycina 27. **Wizualizacja poziomu ekspresji wybranych genów przed i po transfekcji**. A) Wizualizacja ekspresji 81 czynników transkrypcyjnych, dla których wykazano statystycznie istotną zmianę poziomu ekspresji w odpowiedzi na transfekcję. B) Wizualizacja poziomów ekspresji 19 genów, w których w odpowiedzi na transfekcję aktywowane zostały nakładające się promotory. Kółka przy nazwie genu informują o statystycznie istotnej zmianie poziomu ekspresji. Kolor zielony oznacza zwiększenie ekspresji po transfekcji, podczas gdy kolor szary oznacza brak statystycznie istotnej zmiany w poziomie ekspresji.

5.6.3. Studium przypadku: pary genów *TTC9C* i *HNRNPUL2* oraz *DYNLL1* i *SRSF9*

Przykładowa para genów *TTC9C* oraz *HNRNPUL2*, które ulegają nakładaniu po transfekcji, została przedstawiona na rycinie 28 A. Wykorzystanie przez gen *TTC9C* dodatkowego miejsca startu transkrypcji umiejscowionego w obrębie genu *HNRNPUL2* skutkuje utworzeniem rejonu nakładania o długości 3087 par zasad. Aktywacja dodatkowego promotora łączy się ze wzrostem poziomu ekspresji genu *TTC9C* (Log_2 krotności zmiany poziomu ekspresji = 4,58; Wartość P = 5e-9). Jak wcześniej wspomniano, w odpowiedzi na transfekcję aż 81 czynników transkrypcyjnych uległo różnicowej ekspresji. Miejsca wiązania trzech z tych czynników transkrypcyjnych, *GATA2*, *FOXL1* oraz *NFIC* znajdują się w bezpośrednim sąsiedztwie, nie dalej niż 100 nukleotydów, dwóch promotorów genu *TTC9C* (rycina 28 A). Miejsca wiązania czynników *GATA2* i *FOXL1* sąsiadują z nakładającym się promotorem aktywowanym po transfekcji, a czynnika *NFIC* w sąsiedztwie promotora aktywnego także przed transfekcją. W odpowiedzi na transfekcję czynnik transkrypcyjny *GATA2*, który jest znanym aktywatorem[221, 222], uległ podwyższonej ekspresji co może być powiązane z aktywacją nakładającego się promotora. Z drugiej strony

ekspresja czynnika transkrypcyjnego *FOXL1*, który jest znanym represorem[223, 224], uległa statystycznie istotnemu obniżeniu. W najbliższym sąsiedztwie miejsca TSS genu *TTC9C*, które aktywne było przed i po transfekcji, znajduje się miejsce wiązania czynnika *NFIC*, który może pełnić rolę aktywatora[225] lub represora[226]. Po transfekcji ekspresja *NFIC* uległa obniżeniu, ale poziom ekspresji z tego miejsca startu zasadniczo się nie zmienił. Trudno jest więc ocenić czy zmiana poziomu ekspresji tego czynnika miała w tym przypadku znaczenie. Istotne jest natomiast, że ekspresja genu *TTC9C* znacznie wzrosła po transfekcji. Z kolei ekspresja genu *HNRNPUL2,* który znajduje się na przeciwnej nici DNA, spada dziesięciokrotnie mimo wykorzystywania do ekspresji tego samego miejsca TSS. Być może ma to związek z negatywnym wpływem zjawiska nakładania się tych genów, jednakże analiza korelacji ekspresji genów *TTC9C* i *HNRNPUL2* nie wykazała żadnych statystycznie istotnych wyników ani w 34 bibliotekach w których geny te ulegają nakładaniu, ani w 32 bibliotekach w których geny ulegają ekspresji z nienakładających się promotorów. Być może zmiana ta jest związana ze zmianą ekspresji czynników transkrypcyjnych biorących udział w jego regulacji.



Rycina 28. ***Wizualizacja potencjalnego wpływu czynników transkrypcyjnych o różnicowej ekspresji przed i po transfekcji na wykorzystanie nakładających się promotorów po transfekcji.*** *A) Para genów TTC9C i HNRNPUL2. B) Para genów DYNLL1 i SRSF9. Kolorowe kółka oznaczają miejsca wiązania określonych czynników transkrypcyjnych. Ich zabarwienie kolorem czerwonym i zielonym oznacza odpowiednio statystycznie istotne obniżenie i podwyższenie poziomu ekspresji po transfekcji. Skala na rycinie nie została zachowana.*

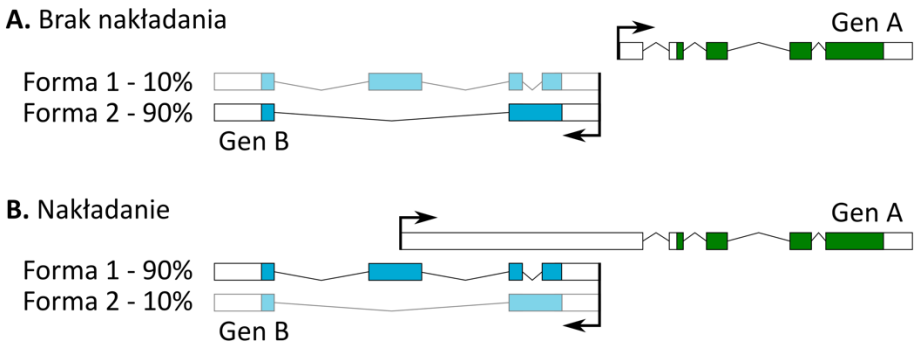
Kolejnym przykładem konsekwentnej zmiany w wykorzystaniu rejonów promotorowych, która skutkuje powstaniem rejonu nakładania jest para genów *DYNLL1* oraz *SRSF9*. Nakładanie się w tym przypadku jest związane ze zmianą promotora przez pierwszy gen z pary (rycina 28 B). Podobnie jak w przypadku genu *TTC9C*, aktywacja nakładającego się promotora genu *DYNLL* po transfekcji może być związana z obniżeniem ekspresji represora *FOXL1*[223]. Spadkowi temu towarzyszy wzrost poziomu ekspresji czynnika transkrypcyjnego *TEAD1* pełniącego funkcję aktywatora[227], co razem mogło przyczynić się do aktywacji nakładającego się promotora. Poziom ekspresji przypisanej do aktywnego przed transfekcją promotora genu *DYNLL1* spada po transfekcji. Może to być spowodowane obniżoną ekspresją wspomnianego wcześniej czynnika transkrypcyjnego *NFIC*, oraz podwyższeniem ekspresji *SP1*, który również może pełnić zarówno funkcję inhibitora

jak i aktywatora[228]. Zmiana ta mogła skutkować zmniejszeniem wydajności tego promotora i przyczynić się do aktywacji dodatkowego, nakładającego się miejsca TSS w celu kompensacji poziomu ekspresji genu. Co ciekawe, ekspresja genu *SRSF9,* znajdującego się na przeciwnej nici, ulega po transfekcji statystycznie istotnemu wzrostowi (Log_2 krotności zmiany poziomu ekspresji = 1,87; wartość P = 0,007), co najprawdopodobniej może mieć związek ze zwiększoną ekspresją czynnika aktywującego *GATA2* oraz spadkiem ekspresji *TFAP2A*, który może pełnić funkcję inhibitora[229, 230].

Do innych par genów nakładających się po transfekcji, których zmianę w wykorzystaniu rejonu promotorowego można tłumaczyć różnicową ekspresją czynników transkrypcyjnych należą pary genów *MGAT2* i *RPL36AL*, *CNTROB* i *TRAPPC1* oraz *SPCS1* i *GLT8D1*. We wszystkich tych parach po transfekcji aktywowany został dodatkowy, nakładający się TSS, który w zależności od pary mógł być regulowany przez podwyższenie ekspresji czynników transkrypcyjnych *GATA2, JUND, E2F1, TEAD1 i TCF7L2,* mogących pełnić funkcje aktywatorów[221, 222, 231-235], oraz obniżenie wyciszającego wpływu *FOXL1*[223].

5.7. Wpływ zjawiska nakładania się genów na alternatywny splicing

Bardzo wiele czynników może wpływać na alternatywny splicing, włączając w to zarówno wykorzystanie alternatywnego rejonu promotorowego[184, 185, 236], jak i spowolnienie elongacji polimerazy RNA II[65-69]. Spowolnienie tempa elongacji może towarzyszyć zjawiskom interferencji transkrypcyjnych wynikającym z inicjacji transkrypcji z nakładających się promotorów[60, 63, 64]. Zjawisko nakładania może też sprzyjać alternatywnemu dojrzewaniu cząsteczki mRNA dzięki oddziaływaniom RNA:RNA[70, 71]. W tym świetle postanowiono sprawdzić, czy nakładanie się genów kodujących białka końcami 5’ może również przyczyniać się do regulacji alternatywnego splicingu. Aby przetestować tą hipotezę należało zidentyfikować takie pary genów, w których zmiana promotora jednego genu z nienakładającego na nakładający się skorelowana była ze zmianą alternatywnych form splicingowych drugiego z genów, przy jednoczesnym wykorzystaniu przez ten gen zawsze tego samego promotora. Taka hipotetyczna para genów przedstawiona została na rycinie 29. Nakładanie się genów w tej parze determinowane jest przez zmianę wykorzystania rejonu promotorowego genu A, podczas gdy ekspresja genu B zachodzi zawsze przy użyciu tego samego miejsca TSS. Ekspresji genu B towarzyszy jednak zależna od zachodzenia nakładania zmiana udziału form splicingowych 1 i 2, która mogłaby być wywołana spowolnieniem polimerazy lub regulacją przez oddziaływania RNA:RNA.

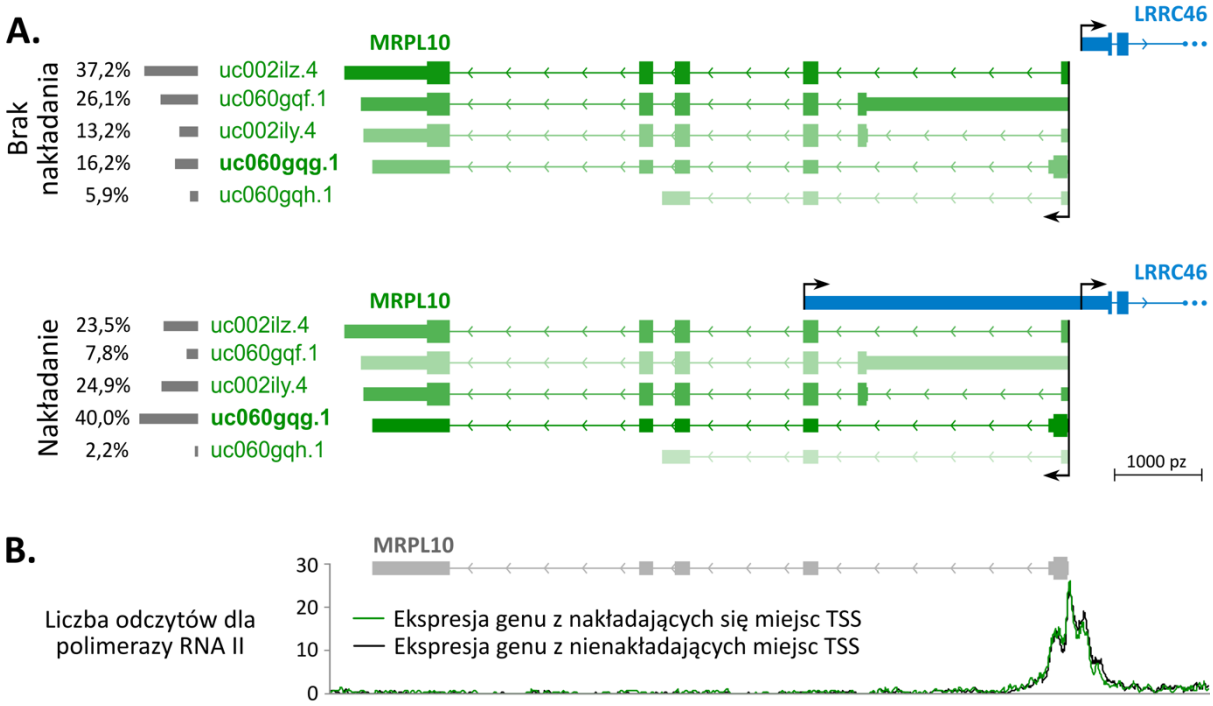


Rycina 29. ***Hipotetyczna para genów na której alternatywny splicing ma wpływ zjawisko nakładania się genów.*** *(A) Przy ekspresji genów w parze z nienakładających się miejsc startu transkrypcji, dominującą formą splicingową genu B jest forma 2, która stanowi 90% wszystkich transkryptów. (B) Przy zachodzeniu nakładania się genów, któremu nie towarzyszy zmiana wykorzystania alternatywnego miejsca startu transkrypcji przez gen B, dominującym wariantem splicingowym tego genu jest forma 1, która stanowi 90% wszystkich transkryptów.*

Analiza powiązania alternatywnego splicingu ze zjawiskiem nakładania się genów przeprowadzona została dla dwudziestu sześciu bibliotek gruczolakoraka płuc, dla których dostępne były dane wysokoprzepustowego sekwencjonowania transkryptomów. Spośród wszystkich par genów zidentyfikowanych jako nakładające w przynajmniej jednej bibliotece gruczolakoraka pozostawiono tylko takie pary, które nakładały się w przynajmniej pięciu bibliotekach oraz w kolejnych pięciu, lub więcej, ulegały ekspresji z wykorzystaniem nienakładających promotorów. W wyniku takiego filtrowania dalszej analizie poddanych zostało 59 par genów. Dla każdej z tych par genów ustalono udział poszczególnych form splicingowych w całkowitej ekspresji genów. Następnie obliczono średni udział poszczególnych form w bibliotekach, w których pary ulegały ekspresji z nakładających się i nienakładających miejsc TSS. Do dalszej analizy pozostawiono 38 par genów, w których różnica średniego udziału przynajmniej jednego wariantu splicingowego wynosiła przynajmniej 10% między bibliotekami w których geny ulegały ekspresji z nakładających i nienakładających promotorów. Większość z tych par genów została następnie odrzucona podczas manualnej analizy, ponieważ zmiany w udziale poszczególnych form splicingowych w całkowitej ekspresji dotyczyły genów, które wykorzystywały alternatywne rejony promotorowe. W takich przypadkach najprawdopodobniej to właśnie zmiana promotora a nie zjawisko nakładania przyczyniło się do zmiany udziału alternatywnych form splicingowych w całkowitej ekspresji genu. Pozytywny wynik filtrowania dla wszystkich etapów analizy otrzymały pary genów *ATPIF1* i *DNAJC8*, *CMC4* i *MTCP1* oraz *LRRC46* i *MRPL10*. Na największą uwagę zasługuje ostatnia z tych par genów. Gen *MRPL10* jest tutaj genem

wykorzystującym zawsze ten sam promotor, podczas gdy nakładanie się genów jest determinowane wykorzystaniem alternatywnego miejsca TSS przez gen *LRRC46*. W sześciu liniach komórkowych, w których *LRRC46* wykorzystuje nakładający się promotor zaobserwowano istotny 23% wzrost udziału wariantu splicingowego *uc060gqg.1*, który należy do genu *MRPL10* (rycina 30 A). Przeprowadzona analiza ekspresji różnicowej wykazała statystycznie istotny wzrost poziomu ekspresji tego wariantu (Log_2 krotności zmiany poziomu ekspresji = 4,95; wartość P = 0,004), przy jednoczesnym braku statystycznie istotnych zmian ekspresji pozostałych form splicingowych. Wariant *uc060gqg.1* posiada odmienny względem innych form rejon kodujący, zlokalizowany całkowicie w pierwszym egzonie. Co więcej, egzon ten jest dłuższy niż w przypadku pozostałych wariantów splicingowych a cały intron znajdujący się między egzonem pierwszym i drugim zlokalizowany jest w rejonie nakładania (rycina 30 A). Warto zauważyć również, że poziom ekspresji genu *MRPL10* nie uległ statystycznie istotnej zmianie. Istotny wzrost udziału jednej z alternatywnych form tego genu odbył się kosztem pozostałych wariantów. Otrzymane wyniki sugerują, że nakładanie się genów mogło mieć w tym przypadku wpływ na proces dojrzewania cząsteczki pre-mRNA.

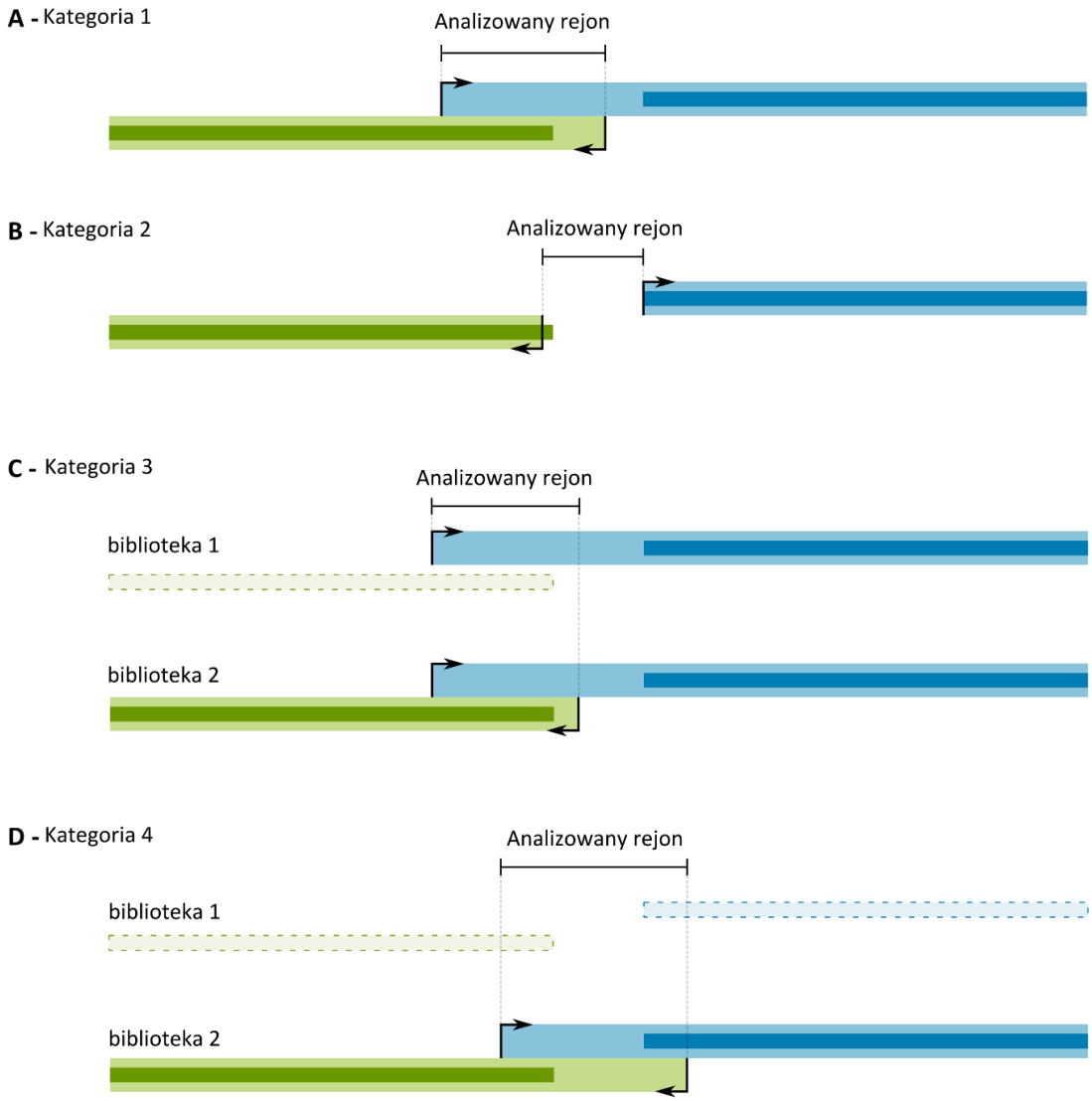
Wyjaśnieniem powyżej opisanej sytuacji mogłoby być spowolnienie w rejonie nakładania tempa elongacji polimerazy RNA II[60, 63, 64]. W konsekwencji takiego spowolnienia wydłużeniu ulega czas dla działania spliceosomu co z kolei umożliwia rozpoznanie alternatywnych i być może słabszych miejsc splicingowych[65-69]. Jeśli istotnie spowolnienie takie następuje, to powinno być to widoczne poprzez podwyższenie w tym rejonie pokrycia odczytami ChIP-Seq dla eksperymentu nacelowanego na badanie aktywności polimerazy RNA II. Niestety, analiza przeprowadzona pod tym kątem dla żadnej z trzech par genów nie wykazała objawów spowolnienia polimerazy objawiającego się istotnym wzrostem pokrycia odczytami w liniach komórkowych, w których transkrypcja zachodzi z rejonów nakładania. Pokazano to na przykładzie pary genów *LRRC46* i *MRPL10* na rycinie 30 B. Sugeruje to, że najprawdopodobniej nie interferencja transkrypcyjna, a jakieś inne mechanizmy, na przykład oddziaływania RNA:RNA, mogą w przypadku genu *MRPL10* wpływać na jego alternatywny splicing.



Rycina 30. ***Potencjalny wpływ zjawiska nakładania się genów na alternatywny splicing.*** *A) Udział wariantów splicingowych pary genów LRRC46 i MRPL10 w bibliotekach w których ulegają one ekspresji z nakładających się i nienakładających miejsc startu transkrypcji. Średni udział każdej z form splicingowych w każdej z grup (nakładanie lub jego brak) podany został jako wartość procentowa oraz zwizualizowany jako uproszczony histogram obok każdego z wariantów. Formy splicingowe których średni udział w ekspresji genu przy nakładaniu lub bez nakładania wynosił mniej niż 1% nie zostały przedstawione na rycinie. B) Wizualizacja aktywności polimerazy RNA II w okolicy genu MRPL10 w bibliotekach w których wykryto nakładanie (kolor zielony), oraz w bibliotekach w których nie odnotowano nakładania (kolor czarny).*

5.8. Analiza sygnałów aktywności polimerazy RNA II i modyfikacji histonów

Zbiorcza analiza sygnałów aktywności polimerazy RNA II oraz modyfikacji histonów miała na celu sprawdzenie na ile stan chromatyny i aktywność polimerazy oddawać mogą obserwowany stan transkrypcyjny genów nakładających się. W tym celu przeprowadzono analizę skupień par genów nakładających się w oparciu o sygnały siedmiu rodzajów modyfikacji histonów (H3ac, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3 i H3K9me3) oraz aktywności polimerazy RNA II w dwudziestu sześciu liniach komórkowych gruczolakoraka płuc. Analiza skupień przeprowadzana była dla obszaru znajdującego się w parach genów pomiędzy aktywnymi miejscami TSS oraz obszaru flankującego o długości 1000 nukleotydów w każdą ze stron. W przypadku, gdy któryś z genów w danej linii komórkowej nie ulegał ekspresji, informacja o położeniu jego miejsc startu transkrypcji brana była z innych bibliotek gruczolakoraka. Pary genów w każdej z dwudziestu sześciu linii komórkowych były dzielone na jedną z czterech kategorii, które graficznie zaprezentowane zostały na rycinie 31 A-D.

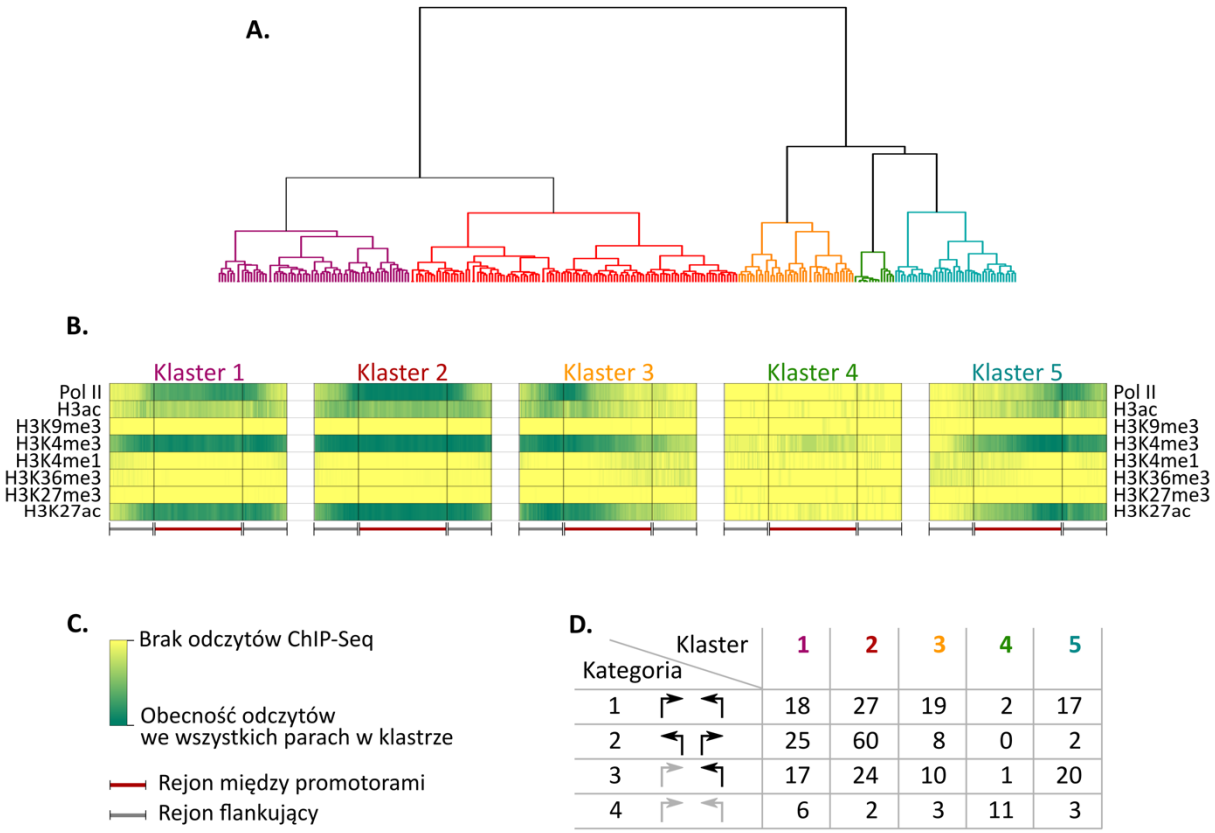


Rycina 31. ***Wizualizacja założeń kategoryzowania par genów w zbiorczej analizie aktywności polimerazy RNA II i modyfikacji histonów****. „Analizowany rejon" jest obszarem, dla którego określa się koordynaty między miejscami TSS analizowanej pary genów. (A) Ekspresja obu genów z nakładających się miejsc TSS; (B) Ekspresja obu genów z nienakładających się miejsc TSS; (C) Ekspresja jednego genu z pary; (D) Brak ekspresji obu genów. Ciemne niebieskie i zielone prostokątne bloki reprezentują ciało genów na nici + i -, podczas gdy jaśniejsze bloki reprezentują najdłuższe jednostki transkrypcyjne. Geny nie ulegające ekspresji w danej bibliotece zaznaczono linią przerywaną.*

Pokrótce, pary genów zaklasyfikowane do pierwszej kategorii w danej bibliotece ulegały nakładaniu, podczas gdy w przypadku par zaklasyfikowanych do drugiej kategorii, oba geny z pary ulegały ekspresji, ale bez nakładania. W przypadku par genów z trzeciej kategorii tylko jeden gen z pary ulegał ekspresji a w przypadku par zaklasyfikowanych do czwartej kategorii żaden gen z pary nie ulegał ekspresji w danej bibliotece według danych TSS-Seq.

Następnie z analizy odrzucono pary genów, których badany rejon, nie wliczając rejonów flankujących, był krótszy niż 200 lub dłuższy niż 3000 nukleotydów. W efekcie analizę przeprowadzono łącznie dla 342 par genów. Hierarchiczna analiza skupień zawsze skutkowała identyfikacją trzech głównych typów modeli modyfikacji histonów oraz aktywności polimerazy RNA II. Modele te odnaleźć można w każdej z 26 analizowanych bibliotek gruczolakoraka płuc. Zostały one omówione poniżej dla przykładowej biblioteki PC3, której dendrogram przedstawiony został na rycinie 32 A.

Pierwszy i drugi klaster reprezentują ten sam model aktywności polimerazy RNA II oraz obecności określonych modyfikacji histonów (rycina 32 B). Na całej długości badane rejony par genów w tych klastrach posiadają trimetylację czwartej lizyny i acetylację 27 lizyny histonu H3, które skorelowane są z aktywnymi promotorami[132, 143, 144, 237]. Relatywnie rzadziej, lecz również na całym obszarze zidentyfikowano obecność modyfikacji H3ac, która szczególnie gdy zlokalizowana jest poniżej miejsca TSS, może korelować ze zwiększoną ekspresją genu[238]. Centralny obszar obu klastrów, znajdujący się pomiędzy miejscami TSS, cechuje się w porównaniu z rejonami flankującymi zwiększoną aktywnością polimerazy RNA II. Mimo iż klastry 1 i 2 zaklasyfikować można jako ten sam model, istnieją między nimi różnice. Główna różnica dotyczy poziomów ekspresji genów należących do pierwszego i drugiego klastra. W pierwszym z nich średnia ekspresja genów zlokalizowanych na nici dodatniej wynosi 103 ppm podczas gdy ekspresja genów na nici ujemnej wynosi 174,4 ppm. Tendencja ta ma odwrotny charakter w klastrze drugim, gdzie dla genów na nici dodatniej ekspresja wynosi średnio 183,8 ppm, natomiast w przypadku genów na nici ujemnej poziom ten osiąga średnio wartość 97 ppm.



Kategoria \ Klaster	1	2	3	4	5
1	18	27	19	2	17
2	25	60	8	0	2
3	17	24	10	1	20
4	6	2	3	11	3

Rycina 32. ***Klasteryzacja sygnałów aktywności polimerazy i modyfikacji histonów w linii komórkowej PC3****. A) Dendrogram hierarchicznej analizy skupień par genów. B) Wizualizacja średniego pokrycia odczytami pochodzącymi z odpowiednich sygnałów modyfikacji histonów oraz aktywności polimerazy RNA II (Pol II). C) Legenda dla podpunktu B. D) Tabela zawierająca liczbę par genów przypadających odpowiednio do klastrów 1-5 oraz kategorii 1-4, które zwizualizowane zostały na rycinie 31 A-D.*

Kolejny model reprezentowany jest przez klastry 3 i 5, stanowiące swoje lustrzane odbicie. W klastrze trzecim obserwuje się tendencje do jednostronnego występowania aktywności polimerazy oraz modyfikacji histonów skorelowanych z aktywnymi promotorami, których największe zagęszczenie znajduje się w okolicy TSS po stronie 5' badanego rejonu (rycina 32 B). Jednocześnie obszar znajdujący się w okolicy końca 3' odznacza się zwiększonym występowaniem metylacji lizyny 4 oraz trimetylacji lizyny 36 histonu H3, które powiązane są odpowiednio z obecnością rejonów wzmacniających[239] oraz aktywnie transkrybowanych genów[132]. W przypadku klastra piątego obserwuje się odwrotne występowanie powyższych cech. Podobnie jak w klastrach 1 i 2, również w przypadku klastra trzeciego i piątego obserwuje się dysproporcję poziomów ekspresji genów znajdujących się na przeciwnych niciach DNA. W przypadku klastra trzeciego poziom ekspresji genów na nici dodatniej wynosi średnio 324 ppm, natomiast na nici przeciwnej jest on ponad dwukrotnie niższy. W klastrze piątym obserwuje się natomiast sytuację odwrotną, w której poziom ekspresji genów na nici dodatniej wynosi 124 ppm i jest trzykrotnie niższy od średniej

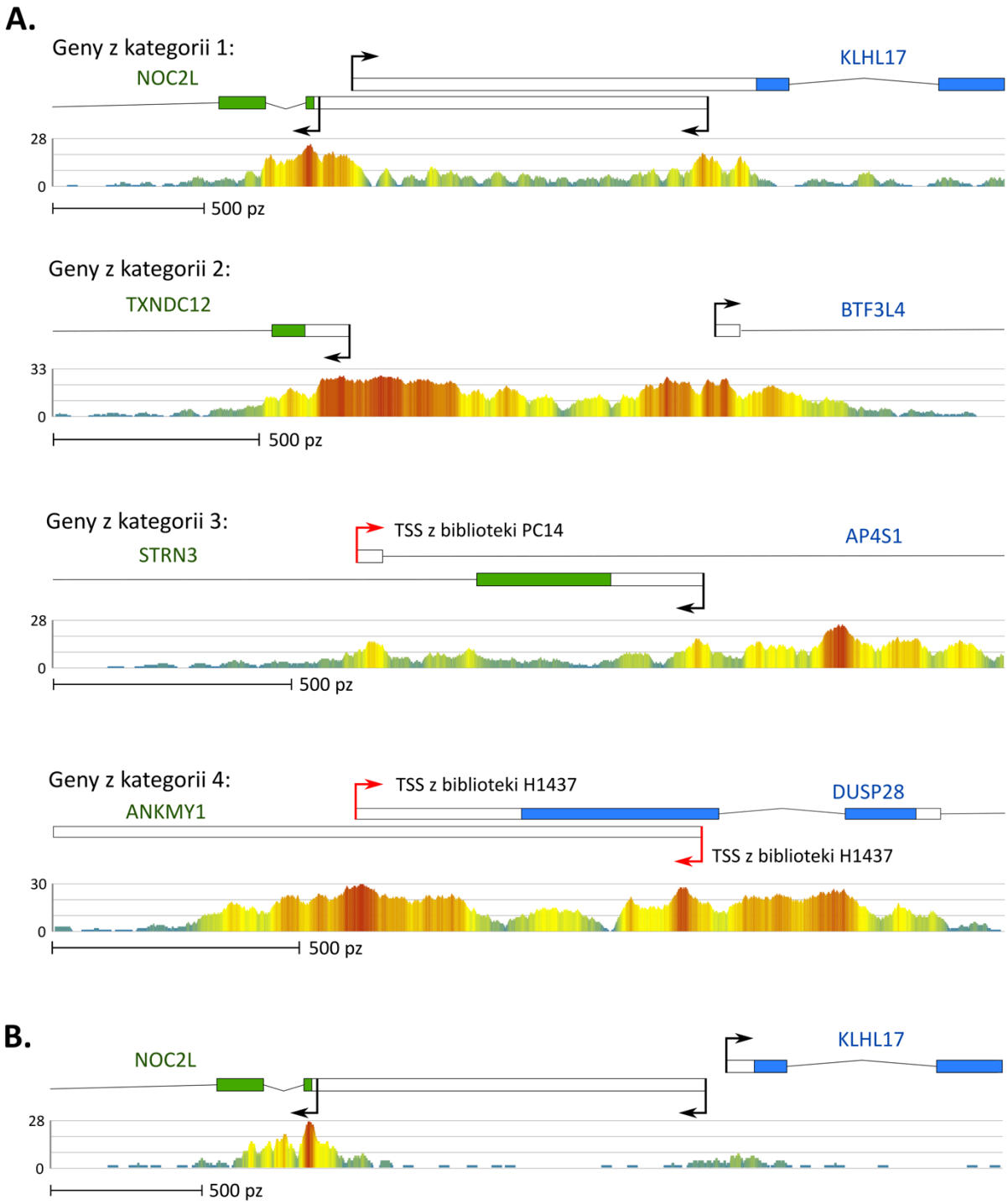
ekspresji genów zakodowanych na nici ujemnej. Zgodnie z tymi tendencjami w grupie par genów, w których tylko jeden gen z pary ulega ekspresji, a które zakwalifikowane zostały do klastra trzeciego, przeważają geny ulegające ekspresji z nici dodatniej, podczas gdy w klastrze piątym przeważają geny ulegające ekspresji z nici ujemnej.

Trzeci model reprezentowany jest przez klaster 4, w którym sygnały wszelkiego typu modyfikacji histonów oraz aktywności polimerazy RNA II są bardzo słabe. Nielicznie występujące tutaj oznaki modyfikacji histonów obejmują również trimetylację dziewiątej i dwudziestej siódmej lizyny histonu H3, które powiązane są z wyciszeniem rejonu chromosomowego[145, 147, 240]. Nie jest zaskakującym, że w klastrze tym zdecydowaną większość stanowią pary, w których żaden z genów nie ulegał ekspresji w linii komórkowej PC3.

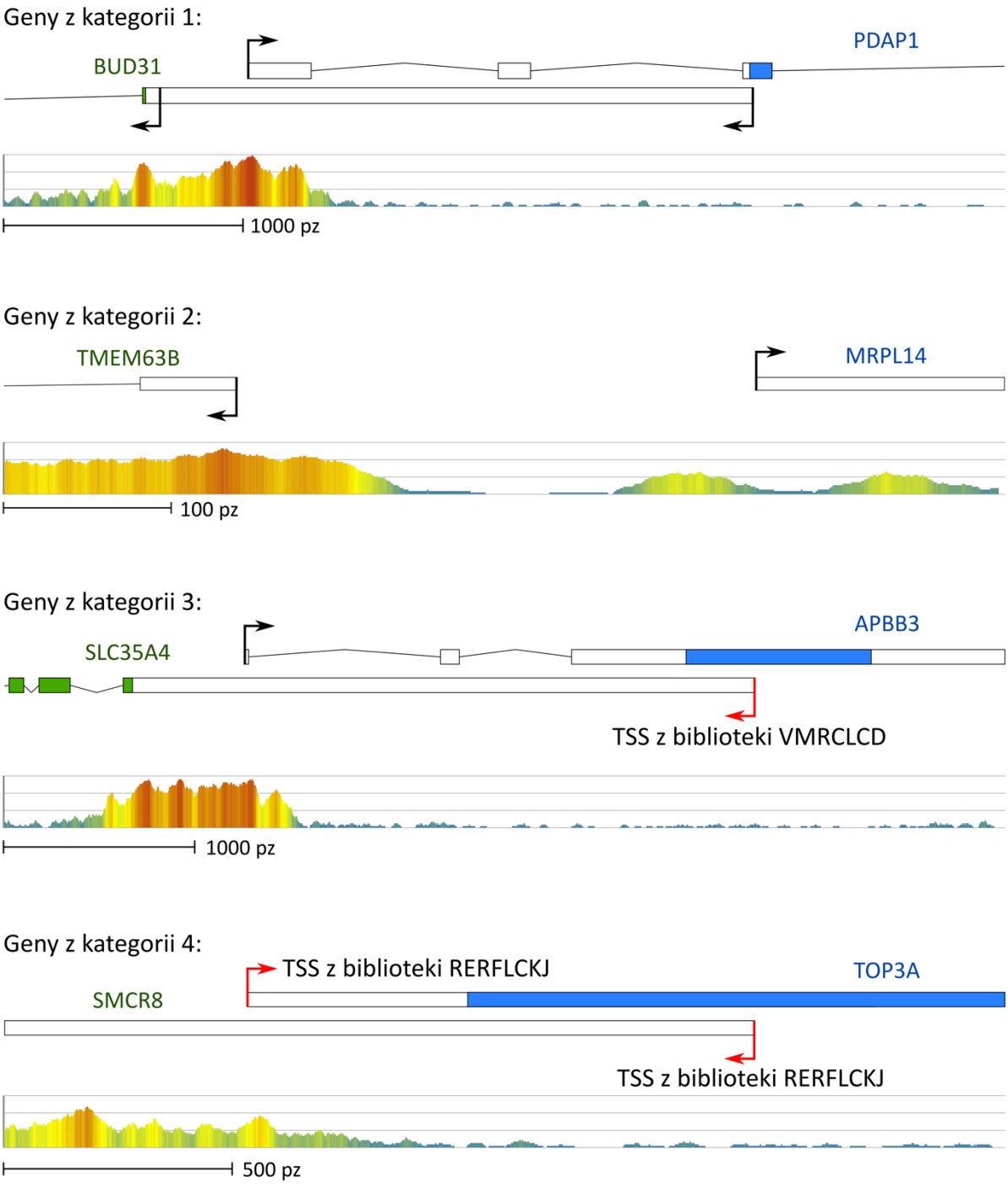
Poszczególne klastry, a co za tym idzie również modele, nie odpowiadają w sposób jednolity poszczególnym kategoriom par genów (rycina 32 D). Przykładowo, do klastra drugiego zaklasyfikowane zostało 63% par genów ulegających ekspresji bez nakładania, lecz również jedna trzecia par genów nakładających się oraz taka sama frakcja par genów w której tylko jeden gen z pary ulegał ekspresji. W klastrze tym znalazły się również dwie pary genów, które w linii komórkowej PC3 nie ulegały ekspresji. Mimo wszystko każda z tych par genów wykazuje w klastrach 1 i 2 podobny wzór kombinacji modyfikacji histonów oraz aktywności polimerazy RNA II. Pośród wszystkich tych sygnałów na szczególną uwagę zasługuje aktywność polimerazy, gdyż to właśnie ona może podlegać zakłóceniom transkrypcyjnym takim jak na przykład kolizja polimeraz. Wzór aktywności polimerazy RNA II dla nienakładających się genów o jednokierunkowych promotorach jest bardzo charakterystyczny i w danych typu ChIP-Seq objawia się statystycznie istotnym wzrostem pokrycia odczytami tylko w okolicy miejsca inicjacji transkrypcji[241, 242]. Na rycinie 33 A przedstawiono przykłady aktywności polimerazy RNA II czterech par genów z różnych kategorii, które zaklasyfikowane zostały do klastrów 1 i 2. Widać tutaj, że wzmożona aktywność polimerazy faktycznie zlokalizowana jest w okolicy aktywnych promotorów genów. Niemniej jednak pomiędzy promotorami tych genów znajduje się obszar o wzmożonej aktywności polimerazy. W przypadku pary genów *KLHL17* i *NOC2L,* która zaklasyfikowana została w linii komórkowej PC3 do kategorii 1, obszar tego wzmocnienia pokrywa się z rejonem nakładania się genów i mógłby być wytłumaczony zachodzeniem zjawiska kolizji polimeraz. Hipoteza ta potwierdza się przy porównaniu do wzorca aktywności polimerazy tej samej pary genów w linii komórkowej II18, w której geny *KLHL17* i *NOC2L* ulegają ekspresji z wykorzystaniem nienakładających się miejsc TSS

(rycina 33 B). Kolizja polimeraz mogłaby też tłumaczyć brak aktywności transkrypcyjnej genu *STRN3*, *DUSP28* oraz *ANKMY1*, należących do 3 i 4 kategorii genów. Najwyższa aktywność polimerazy RNA II w linii komórkowej PC3 odpowiada w przypadku wyżej wymienionych genów lokalizacji miejsc TSS w innych liniach komórkowych (rycina 33 A). Wzór aktywności polimerazy w rejonie między aktywnymi promotorami genów *BTF3L4* i *TXNDC12* pozostaje jednak do pewnego stopnia zagadkowy. Geny te według danych TSS-Seq ulegają ekspresji z wykorzystaniem nienakładających się promotorów, niemniej jednak cały rejon między aktywnymi miejscami TSS wykazuje bardzo wysoką aktywność polimerazy RNA II.

Model drugi, który reprezentowany jest przez klastry 3 i 5 również nie jest jednorodny pod względem kategoryzacji par genów, które zostały do niego włączone. Na rycinie 34 przedstawiono po jednym przykładzie par genów z każdej z czterech kategorii, które włączone zostały do klastra 3. Do klastra tego, tak jak wspomniano już wcześniej, włączane były geny o przeważającej ekspresji genu zakodowanego na nici dodatniej oraz pary genów, w których ekspresji ulegał tylko gen z nici dodatniej. Jest to odzwierciedlone silnym sygnałem aktywności polimerazy RNA II w okolicy 5'. Zaklasyfikowane do tego modelu pary genów nie wykazują też wzmożonej aktywności polimerazy RNA II w centralnej części, która w przypadku modelu pierwszego mogła świadczyć o zachodzeniu interferencji transkrypcji. Nie jest to zaskakujące w przypadku par genów zaklasyfikowanych do kategorii 2-4, w przypadku których zjawisko nakładania nie zachodzi. W przypadku pary genów *PDAP1* i *BUD31* brak obserwacji wzmożonej aktywności polimerazy RNA II w rejonie nakładania może być wytłumaczony bardzo niską frakcją ekspresji genów, która przypisana została do rejonu nakładania. Wartość współczynnika JoinedOR dla tej pary genów wyniosła w bibliotece PC3 jedynie 0,003, co oznacza, że nie więcej niż 3‰ transkrypcji inicjowane jest w rejonie nakładania.



Rycina 33. ***Aktywność polimerazy RNA II wybranych par genów.*** *A) Aktywność polimerazy czterech par genów należących do modelu 1 w linii komórkowej PC3. B) Aktywność polimerazy RNA II dla pary genów KLHL17 i NOC2L w linii komórkowej II18.*

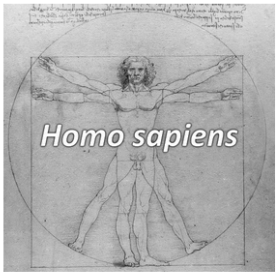


Rycina 34. ***Aktywność polimerazy RNA II wybranych par genów z klastra 3 w linii komórkowej PC3.***

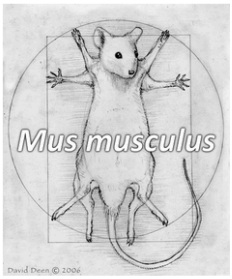
5.9. OverGeneDB – internetowa baza genów nakładających się

OverGeneDB jest internetową bazą danych, publicznie dostępną pod adresem http://overgenedb.amu.edu.pl. Baza ta jest interaktywnym interfejsem bazy danych MySQL, w której skolekcjonowano kluczowe wyniki niniejszej pracy doktorskiej. Interfejs bazy danych przygotowany został w języku Angielskim, co ma zapewnić dostępność prezentowanych wyników szerszej rzeszy odbiorców. Dane skolekcjonowane w bazie mogą

być przeglądane za pomocą zakładki „Browse” (pol. Przeglądaj), gdzie znajduje się tabela z listą wszystkich par genów nakładających się u człowieka i myszy, wraz z podstawową informacją o liczbie bibliotek TSS-Seq, w których dana para ulega ekspresji z nakładających się miejsc startu transkrypcji, oraz w ilu bibliotekach jeden lub oba geny z pary ulegają ekspresji (rycina 35).



species



582	overlapping gene pairs	113
1150	overlapping genes	225
73	libraries	10

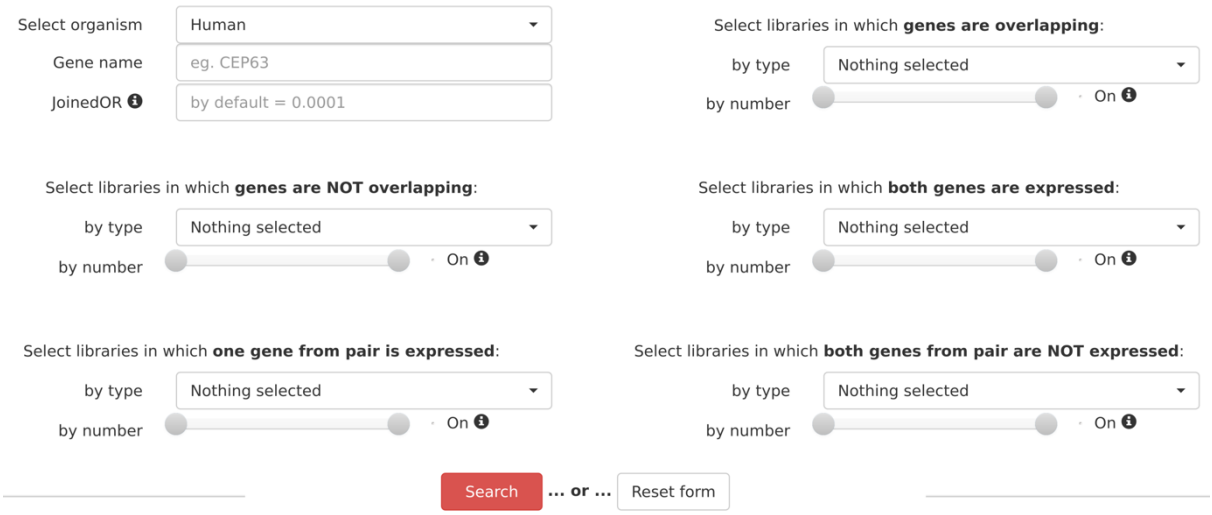
Human overlapping pairs | **Mouse** overlapping pairs

Details	#No.	Gene on positive DNA strand:	Gene on negative DNA strand:	Genes overlap in:	Both genes expressed in:	Gene on positive strand expressed in:	Gene on negative strand expressed in:
» Details »	1.	CEP63	ANAPC13	4 libraries	9 libraries	50 libraries	11 libraries
» Details »	2.	PCNT	C21orf58	1 library	4 libraries	51 libraries	4 libraries
» Details »	3.	TIMM10B	ARFIP2	2 libraries	30 libraries	32 libraries	47 libraries

Rycina 35. ***Zakładka „Browse” w bazie OverGeneDB.***

Bazę danych można swobodnie przeszukiwać również z wykorzystaniem zakładki „Search” (pol. Szukaj) (rycina 36). Użytkownik może tutaj sprecyzować rodzaj lub liczbę bibliotek TSS-Seq, z których interesują go nakładające się lub nienakładające pary genów. Ponadto określić można biblioteki, w których użytkownik jest przykładowo zainteresowany odnalezieniem par genów, w których jeden lub oba geny z pary ulegają ekspresji niezależnie od zjawiska nakładania. Możliwa jest również modyfikacja wartości minimalnej współczynnika JoinedOR, który domyślnie przyjmuje 0,0001. Dzięki temu, definiując wartość JoinedOR jako na przykład 0,7 lub 1, użytkownik z łatwością może wyszukać tylko takie pary genów, które kolejno głównie lub całkowicie ulegają transkrypcji z nakładających się miejsc TSS. Kolejna strategia przeszukiwania zasobów bazy OverGeneDB oparta jest o przeszukiwanie jej za pomocą zadanej przez użytkownika sekwencji nukleotydowej lub białkowej. Sekwencje te zostaną przy użyciu algorytmu BLAST[243, 244] przyrównane

do bazy sekwencji nukleotydowych rejonów nakładania lub sekwencji genów nakładających się.

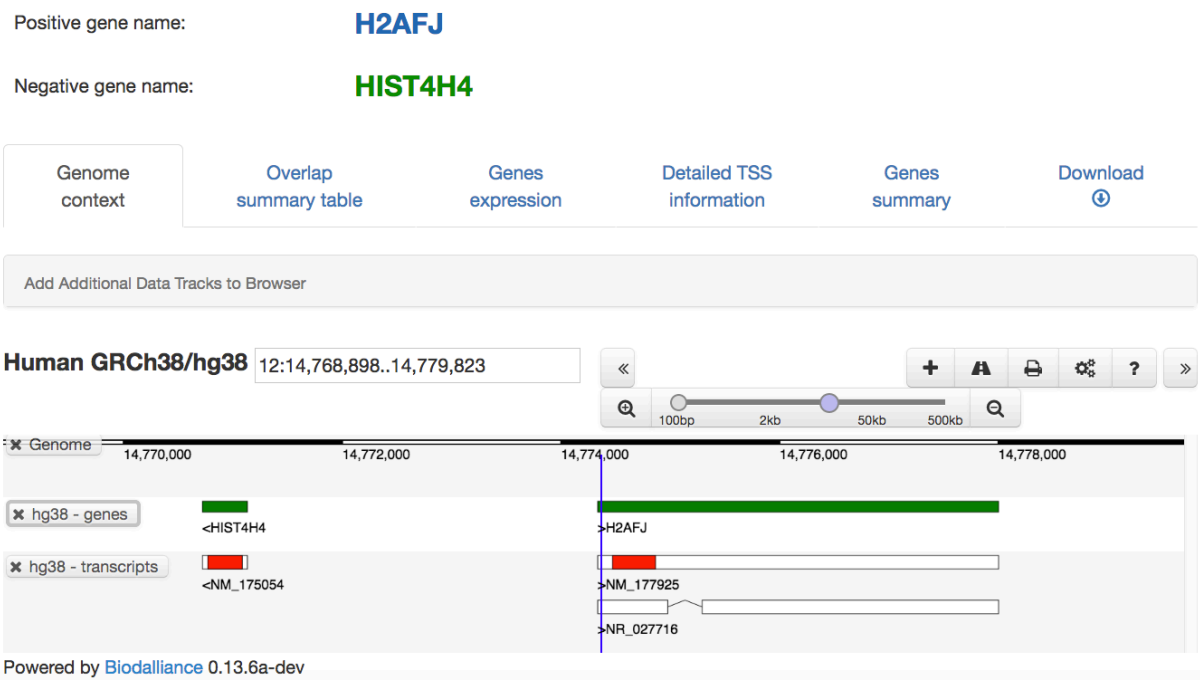


Rycina 36. ***Zakładka „Search" w bazie OverGeneDB.***

Strona podglądu dla pojedynczej pary, do której prowadzą opisane powyżej metody przeszukiwania bazy, wyświetla domyślnie wybraną parę genów w przeglądarce genomowej o nazwie dalliance[211] (rycina 37). Domyślnie wyświetlane są tutaj dane pochodzące z trzech plików, które zwyczajowo po angielsku nazywane są „Tracks" (pol. ścieżka), a które odpowiednio zawierają sekwencję genomu referencyjnego, najdłuższe koordynaty genu obliczone zgodnie z protokołem opisanym w rozdziale 5.1, oraz sekwencje transkryptów z bazy RefSeq, które wykorzystane zostały do oznaczenia koordynat genów. Użytkownik ma możliwość wyświetlenia zawartości innych plików danych klikając na przycisk „Add Additional Data Tracks to Browser", znajdującego się ponad oknem przeglądarki genomowej. Ukazujący się po wyborze tej opcji panel oferuje możliwość włączenia w sumie 691 innych plików danych obejmujących:

- Miejsca alternatywnego startu transkrypcji. Format pliku BED/bigBed.
- Bloki reprezentujące rejony nakładania. Format pliku BED/bigBed.
- Potencjalne miejsca wiązania czynników transkrypcyjnych ulegających ekspresji w wybranej bibliotece. Format pliku BED/bigBed.
- Siedem modyfikacji histonowych i aktywność polimerazy RNA II, gdzie dla każdej biblioteki istnieje możliwość wyświetlenia zarówno rejonów wysycenia (ang. *peaks*) jak również informacji o pokryciu odczytami. Dostępne formaty plików to kolejno BED/bigBed oraz bigWig. Dane dostępne tylko dla człowieka.

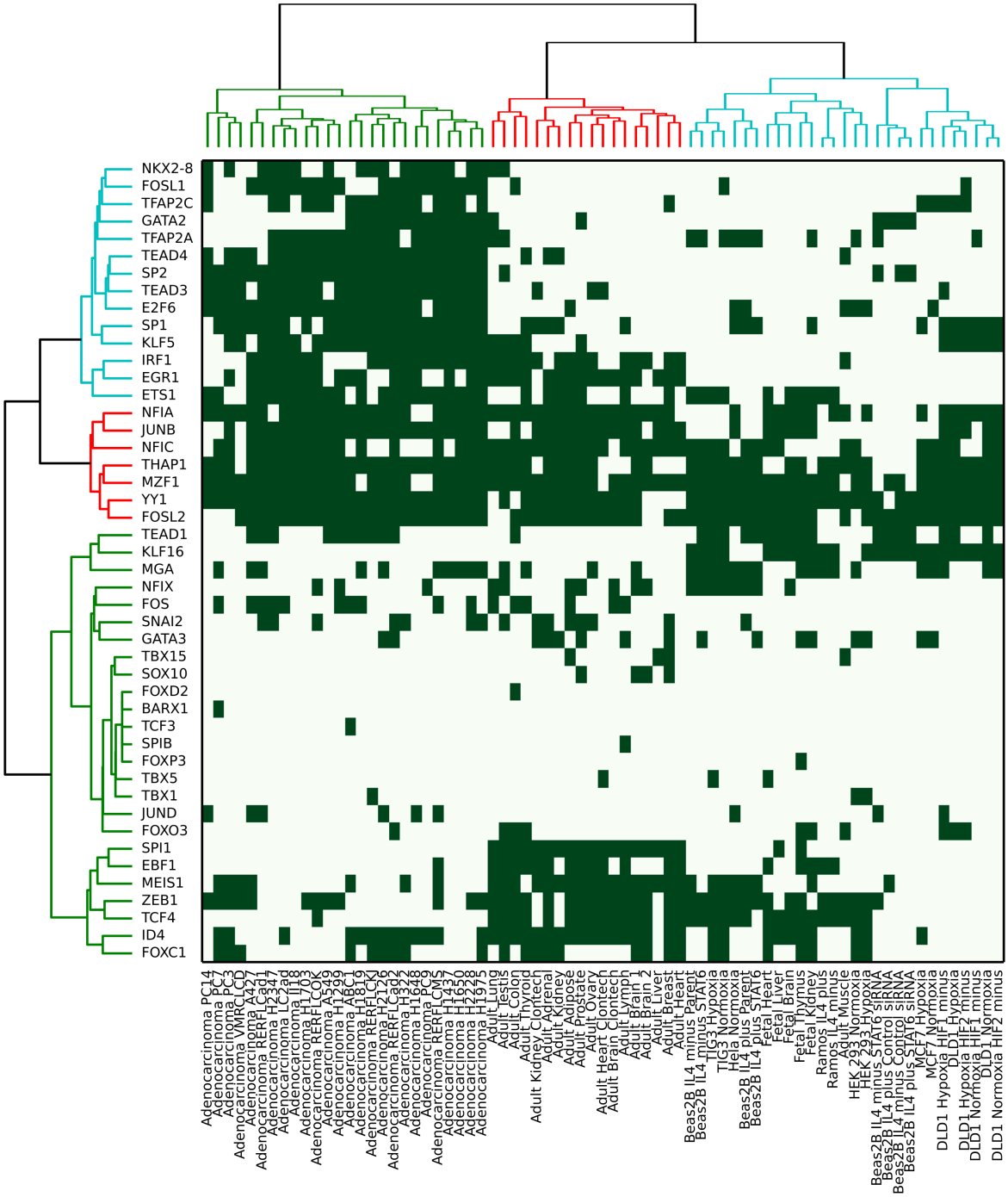
- Zmapowane odczyty RNA-Seq. Format pliku BAM. Dane dostępne tylko dla człowieka.



Rycina 37. ***Szczegółowy podgląd wybranej pary genów w bazie OverGeneDB.***

Szczegółowa informacja odnośnie tego w jaki sposób nakładają się geny w przeglądanej parze, jakie przyjmuje ona wartości OR i JoinedOR, oraz jakie są poziomy ekspresji genów w poszczególnych bibliotekach TSS-Seq / RNA-Seq, znajdują się w zakładkach „Overlap summary table" oraz „Genes expression". Kliknięcie na nazwę biblioteki w zakładce szczegółowych informacji o nakładaniu się genów, uaktywnia dodatkowe okienko, w którym znajduje się wizualizacja nakładania genów w danej bibliotece, wraz ze szczegółową informacją odnośnie poziomów ekspresji przypisanych poszczególnym miejscom TSS. Zakładka szczegółowej informacji o miejscach TSS (ang. *detailed TSS information*) oferuje informacje o wszystkich alternatywnych rejonach promotorowych przypisanych do każdego z genów, wraz z informacją o poziomie ekspresji z danego miejsca TSS w poszczególnych bibliotekach. Zakładka ta zawiera ponadto informacje o czynnikach transkrypcyjnych potencjalnie zaangażowanych w regulację różnych miejsc TSS. Zostały one graficznie podsumowane w postaci dwuwymiarowych map ciepła (ang. *heatmaps*), gdzie kolorem zielonym oznaczono ekspresję a białym brak ekspresji danego czynnika transkrypcyjnego (rycina 38). Każda z map została dodatkowo poddana analizie skupień i dostępna jest do pobrania w trzech formatach: PNG (z ang. *Portable*

Network Graphics), SVG (z ang. *Scalable Vector Graphics*) oraz TSV (z ang. *Tab Separated Value*).



Rycina 38. ***Czynniki transkrypcyjnie potencjalnie zaangażowane w regulację przykładowego promotora „TSS_7000", należącego do ludzkiego genu PARP10****. Kolorem zielonym oznaczono ekspresję a białym brak ekspresji czynnika transkrypcyjnego w danej bibliotece.*

Zakładka „Genes summary" zawiera szczegółowe informacje, które dla każdego z genów w parze obejmują:

- Symbol genu, jego znane synonimy oraz jego oficjalną nazwę, na podstawie danych RefSeq[183].
- Linki do zewnętrznych baz danych takich jak NCBI[245], OMIM[246], HGNC[247], Ensembl[248], HPRD[249] oraz Vega[250].
- Opis genu z bazy RefSeq.
- Opis funkcji pełnionych przez dany gen na podstawie źródeł literaturowych. Dane te pochodzą z bazy NCBI „Gene References Into Functions”[251, 252].
- Adnotacje funkcjonalne z bazy Gene Ontology[253].
- Listę publikacji, które w dowolnym kontekście omawiały wybrany gen, a które pobrane zostały z serwera FTP bazy danych NCBI.

Zakładka „Download” umożliwia ściągnięcie szczegółowych danych powiązanych z przeglądaną parą genów w formatach TSV, BED oraz FASTA.

6. Dyskusja

Inicjacja transkrypcji na obu niciach DNA tego samego *loci* genomowego jest bardzo rozpowszechnionym zjawiskiem zarówno w genomie ludzkim, jak i genomach innych organizmów. Transkrypcja może tutaj skutkować powstaniem rejonu nakładania się genów, czyli fragmentu sekwencji DNA, który przepisany jest do transkryptów powstających na obu niciach DNA. Utworzone w ten sposób pary, tzw. naturalnych transkryptów antysensownych, tworzone są zarówno przez geny kodujące białka, jak również długie niekodujące RNA. Funkcjonalny charakter zjawiska nakładania się genów był wielokrotnie omawiany[37], szczególnie w kontekście par tworzonych przez lncRNA oraz gen kodujący białko. Możliwym jest, że funkcjonalność obserwowana w takich przypadkach wynika z tego, że cząsteczka długiego niekodującego RNA wyewoluowała w taki sposób, aby regulować poziom ekspresji lub alternatywny splicing genu kodującego białko. W przypadku nakładającej się pary genów składającej się z dwóch genów kodujących białka, tego typu funkcja samego zjawiska nakładania nie jest tak oczywista, gdyż główną funkcją obu genów jest kodowanie białek. Pytanie, na które starano się odpowiedzieć w ramach niniejszej pracy doktorskiej dotyczy tego, czy zjawisko nakładania się genów pełni jakąś funkcję regulatorową czy też jest tylko i wyłącznie produktem ubocznym wykorzystania alternatywnych rejonów promotorowych umiejscowionych na przeciwnych niciach DNA w relatywnie bliskim sąsiedztwie. Wiadomo bowiem, że geny mogą wykorzystywać do ekspresji wiele alternatywnych miejsc startu transkrypcji, co skutkować może powstaniem transkryptów o różnej długości na końcu 5'[185, 254]. Przykładowo Tan i współpracownicy[255] wykazali, że 35% genów zaangażowanych w rozwój erytrocytów wykorzystuje w różnych wariantach splicingowych alternatywny pierwszy egzon. Z kolei Kim i współpracownicy[254] zidentyfikowali 1609 genów wykorzystujących do ekspresji w ludzkich fibroblastach jednocześnie wiele rejonów promotorowych. Wykorzystanie przez geny usytuowane relatywnie blisko siebie alternatywnych promotorów może prowadzić do ich nakładania się końcami 5'. Jako że określone promotory mogą być aktywowane i inaktywowane w odpowiedzi na różne warunki wewnątrz i zewnątrzkomórkowe, wykorzystanie alternatywnych promotorów może prowadzić do nakładania się genów w jednych i ekspresji bez nakładania w innych tkankach i liniach komórkowych.

Aby zbadać zjawisko nakładania się genów w tym kontekście, przeanalizowano alternatywne miejsca startu transkrypcji genów kodujących białka w 73 ludzkich i 10 mysich organach, tkankach i liniach komórkowych. Zidentyfikowano łącznie 582 ludzkich i 113 mysich par genów nakładających się częściowo końcami 5' w przynajmniej jednej

ze zbadanych bibliotek TSS-Seq. Liczba zidentyfikowanych par genów jest wyższa niż w przypadku analiz przeprowadzonych przez niektórych innych badaczy. Przykładowo Veeramachaneni i współpracownicy[16] zidentyfikowali 243 pary genów nakładających się końcami 5' u człowieka. Różnica ta może być jednak łatwo wytłumaczona tym, że obecnie mamy znacznie lepszy dostęp do większego zestawu danych. Z drugiej jednak strony liczba par genów zidentyfikowanych w ramach niniejszej pracy doktorskiej jest mniejsza aniżeli ta otrzymana w projekcie FANTOM3, w którym zidentyfikowano aż 1638 par genów nakładających się[24]. Niemniej jednak w projekcie FANTOM w zestawie tym są również pary utworzone przez geny kodujące białko nakładające się z niekodującymi RNA.

Tkankowo specyficzne wzorce zjawiska nakładania się genów kodujących białka

Liczba nakładających się par genów w każdej z analizowanych bibliotek zmieniała się bardzo dynamicznie pomiędzy różnymi rodzajami tkanek i linii komórkowych, co zgodne jest z wynikami badań zaprezentowanymi wcześniej przez Conley oraz Jordana[34]. Również Ling i współpracownicy[36] zidentyfikowali tkankowo specyficzne pary genów nakładających się w dziewięciu organach człowieka, myszy i szczura. Tkankowo specyficzne wzorce nakładania się genów mogą sugerować, że taka konfiguracja genów ma znaczenie funkcjonalne[31, 34, 36]. Niemniej jednak Struhl[39] sugeruje, że sama obserwacja pewnego wzorca ekspresji nie koniecznie odzwierciedla funkcjonalność danego zjawiska, gdyż znaczna część aktywności polimerazy RNA II może być przypadkowa.

Spośród zidentyfikowanych w niniejszej pracy doktorskiej par genów tylko część ulegała nakładaniu we wszystkich bibliotekach, w których oba geny z pary ulegały ekspresji. W ani jednym przypadku geny z danej pary nie ulegały ekspresji we wszystkich badanych bibliotekach TSS-Seq. W większości natomiast para genów ulegała ekspresji z wykorzystaniem nakładających się rejonów promotorowych w jednej lub kilku bibliotekach, przy jednoczesnej ekspresji z wykorzystaniem nienakładających się miejsc startu transkrypcji w innych bibliotekach. Aby uchwycić wzorce ekspresji genów dla 73 ludzkich bibliotek przeprowadzono hierarchiczną analizę skupień. Wykazała ona, że analizowane biblioteki grupują się w trzy duże klastry: linie gruczolakoraka płuc, organy płodowe i dorosłego człowieka oraz laboratoryjne linie komórkowe. Wyłaniający się obraz ukazał klaster genów nakładających się prawie we wszystkich liniach komórkowych gruczolakoraka płuc, przy jednocześnie znikomej ekspresji, najczęściej bez nakładania, w innych bibliotekach. Widoczne są również mniejsze klastry genów ulegających ekspresji w wybranych tkankach, ale wykazujących tendencje do nakładania się w liniach komórkowych nabłonka

oddechowego Beas2B, które poddane zostały transfekcji. Otrzymane wyniki wskazują na to, że przynajmniej niektóre geny mogą w określonych tkankach „preferować" ekspresję z nakładających się promotorów, co ma najprawdopodobniej związek z czynnikami tkankowo specyficznymi.

Przeprowadzone w ramach pracy badania wykazały również, że zjawisko nakładania tej samej pary genów może przyjmować drastycznie odmienny charakter między różnymi bibliotekami. Różnice te mogą dotyczyć zarówno długości jak i lokalizacji rejonu nakładania się genów, który w zależności od biblioteki może znajdować się jedynie w rejonach 5' UTR obu genów, wchodzić w obszar kodujący jednego lub obu genów w parze lub nawet sięgać końców 3'. Jak zostało to zademonstrowane przykładowo dla par genów *HTRA2* i *AUP1* oraz *CNTROB* i *TRAPPC1*, rejon nakładania może być umiejscowiony czasem w jednym a czasem w drugim genie z pary, w zależności od wykorzystania przez te geny alternatywnych miejsc startu transkrypcji. W niektórych przypadkach może mieć to istotne znaczenie regulatorowe na poziomie interakcji RNA:RNA kiedy to rejon nakładania obejmuje na przykład miejsce docelowe dla cząsteczki miRNA. W pracy zidentyfikowano aż 27 przypadków, w których w zależności od lokalizacji rejonu nakładania dane miejsce docelowe może być maskowane lub nie.

Międzygatunkowe zakonserwowanie zjawiska nakładania

Liczne doniesienia literaturowe wykazywały, że zjawisko nakładania się wielu par genów jest gatunkowo specyficzne[16, 24, 25, 256]. Veeramachaneni i współpracownicy[16] pokazali przykładowo, że człowiek posiada 255 par genów nakładających się, które posiadają geny ortologiczne u myszy. Spośród nich jedynie 95 ulegało nakładaniu u obu gatunków, jednakże wzór nakładania się był u obu gatunków różny. Wyniki otrzymane w projekcie FANTOM3[24] zademonstrowały, że nakładanie się mniej niż 20% wszystkich zidentyfikowanych par genów jest zakonserwowane między człowiekiem i myszą. Z kolei badania przeprowadzone przez Wood i współpracowników[25] wykazały, że mniej niż połowa par ludzkich genów kodujących białka nakładających się z długimi niekodującymi RNA, jest zakonserwowana u myszy. Podobnie jak Veeramachaneni, wykazali oni, że zakonserwowanie między oboma gatunkami zjawiska nakładania często nie jest jednoznaczne z zachowaniem wzorców nakładania tych genów. W tych samych badaniach, Wood i współpracownicy wykazali również, że międzygatunkowe zachowanie wzorca nakładania się genów na końcu 5' jest znacznie rzadsze niż na końcach 3'[25].

W takim kontekście zidentyfikowanie w niniejszej pracy doktorskiej jedynie 26 ortologicznych par genów nakładających się zarówno u myszy jak i u człowieka, nie jest w pełni zaskakujące. Warto przypomnieć też, że jedynie 5 par genów ulega u obu gatunków nakładaniu w homologicznych organach. Może być to związane z bardzo wysoką dynamiką wykorzystania alternatywnych miejsc startu transkrypcji, a co za tym idzie nakładania się genów w różnych bibliotekach. Jeśli bowiem między tkankami i organami tego samego gatunku występują tak duże fluktuacje, nie można się spodziewać, że zjawisko to będzie silnie zachowane międzygatunkowo. Nie można jednak wykluczyć, że liczba par genów zakonserwowanych między człowiekiem a myszą znacząco by wzrosła, gdyby przeanalizowano więcej mysich bibliotek TSS-Seq.

Wpływ czynników wewnątrz i zewnątrzkomórkowych na nakładanie się genów

Zmienne warunki wewnątrz i zewnątrzkomórkowe mogą wpływać na to który z promotorów będzie wykorzystany, co z kolei może mieć funkcjonalne implikacje. Przykładowo Zhou i współpracownicy[257] zidentyfikowali 108 genów kodujących białka, które w odpowiedzi na zmianę warunków nakładały się z niekodującymi RNA, a powstały między komplementarnymi transkryptami dupleks RNA stawał się źródłem endogennych siRNA. Warunki egzogenne mogą także wpływać na kompozycję czynników transkrypcyjnych w komórce co przekłada się na zmienne wykorzystanie alternatywnych miejsc startu transkrypcji[178]. W takim świetle nie jest zaskakującym, że aktywacja transkrypcji z nakładających się miejsc TSS może być odpowiedzią na zmienne czynniki tkankowo specyficzne lub też czynniki zewnętrzne. Skala wpływu tych ostatnich na zmianę wykorzystywanych promotorów jest w niektórych przypadkach niespodziewanie duża. Przykładem może być duży wzrost liczby nakładających się genów w transfekowanych liniach komórkowych nabłonka układu oddechowego. W przeprowadzonych analizach zidentyfikowano 15 par, które w reakcji na transfekcję zainicjowały transkrypcję w rejonie nakładania. Analiza miejsc wiązania czynników transkrypcyjnych pokazała, że w przypadku pięciu par aktywacja nakładającego się promotora mogła mieć związek ze zwiększoną po transfekcji ekspresją czynników transkrypcyjnych *GATA2, JUND, E2F1, TEAD1* i *TCF7L2,* które są znanymi aktywtorami[221, 222, 231-235], oraz obniżeniem ekspresji represora *FOXL1*[223]. Niemniej jednak dwie spośród tych 15 par genów, które w liniach Beas2B wykorzystywały w odpowiedzi na transfekcję nakładające się promotory, w liniach komórkowych DLD1 wykazały odwrotne tendencje, nakładając się przed transfekcją a ulegając ekspresji z nienakładających się miejsc TSS po niej. Uwidacznia to, że na to,

który z promotorów będzie wykorzystywany wpływ mają istotnie różne czynniki, w tym przypadku zarówno stres komórkowy wywołany transfekcją jak i typ komórki.

Dodatkowo, przeprowadzona analiza pokazała jak duża może być skala ubocznych efektów przeprowadzanych eksperymentów. W przypadku linii komórkowej Beas2B transfekcja spowodowała nie tylko zmiany poziomu ekspresji wielu genów, ale także jej skutkiem była aktywacja 130 genów i wyciszenie aż 1058. Wśród tych ostatnich niedoreprezentowane były geny pełniące funkcje związane między innymi z odpowiedzią immunologiczną. Oznacza to, że do pewnego stopnia transfekowane komórki ograniczyły wydatkowanie zasobów na inne cele. Liczne źródła literaturowe podają, że efekty uboczne tego typu nie należą do rzadkości[258-262], natomiast transfekcja może poprzez aktywację interferonów stymulować ekspresję genów związanych z odpowiedzią na infekcję komórkową[263, 264].

Ekspresja genów nakładających

Wyniki wielu badań mogą sugerować, że jeśli nakładanie się genów kodujących białka ma funkcjonalne znaczenie, główną rolą jaką będzie ono pełniło jest dostrojenie poziomów ekspresji danych genów. Wniosek taki wynika na przykład z prac Shearwina i współpracowników[38] oraz Yla-Herttuala i Kaikkonen[265]. We wczesnych pracach opisujących zjawisko nakładania wykazano, że naturalne transkrypty antysensowne mają tendencję do ulegania wspólnej i często negatywnie skorelowanej ekspresji[266, 267]. Nowsze badania pokazały natomiast, że ekspresja genów nakładających się może faktycznie być skorelowana, ale pozytywnie[36] lub też korelacji w ogóle nie ma[34]. Wyniki zaprezentowane w niniejszej pracy pokazują, że poziom ekspresji większej części genów nie jest skorelowany, jednakże w tych przypadkach, w których zidentyfikowano jej statystycznie istotną wartość, miała ona charakter pozytywny a nie negatywny. Negatywna korelacja w bibliotekach, w których ekspresja zachodziła z rejonów nakładania potwierdzona została tylko w przypadku jednej pary genów. Jest to zgodne z częścią badań prowadzonych nad naturalnymi transkryptami antysensownymi[34, 36].

Otrzymany wynik zdaje się być w pewnym stopniu zgodny z modelem inicjacji transkrypcji przez promotory dwukierunkowe, które regulują wiele genów u różnych gatunków, w tym u człowieka[160, 170, 268]. Ekspresja genów regulowanych takimi promotorami jest często pozytywnie skorelowana. Można więc przypuszczać, że ekspresja wielu z badanych par genów może być regulowana takimi właśnie promotorami. Co prawda średnia długość rejonu nakładania, która wynosi 1570 pz, jest stosunkowo duża jak na

dwukierunkowe promotory, jednakże istnieją doniesienia mówiące o tym, że mogą one osiągać nawet do 2000 nukleotydów długości[95]. Analiza czynników transkrypcyjnych potencjalnie regulujących rejony promotorowe ludzkich genów nakładających się wykazała statystycznie istotną nadreprezentację kilku czynników transkrypcyjnych które powiązane zostały uprzednio z regulacją promotorów dwukierunkowych[95, 168, 170, 217]. Między innymi był to czynnik transkrypcyjny *ZNF143*, który przez Anno i współpracowników[218] zademonstrowany był jako wiążący 47% promotorów dwukierunkowych. Innym czynnikiem transkrypcyjnym, którego miejsca wiązania nadreprezentowane były wśród promotorów genów nakładających się był *GABPA*, który przez Collinsa i współpracowników[168], zidentyfikowany został jako czynnik wiążący DNA ponad 80% promotorów dwukierunkowych. Collins pokazał ponadto, że w 67% przypadków, związanie się tego czynnika transkrypcyjnego do promotora jednokierunkowego powodowało w nim inicjację transkrypcji w dwóch kierunkach[168]. Niemniej jednak dalsza analiza wykazała, że jedynie 15% par genów nakładających się u człowieka zawiera w promotorach miejsca wiązania *GABPA*. Zupełnie inaczej sytuacja przedstawiła się w przypadku myszy, u której miejsca wiązania czynnika transkrypcyjnego *Gabpa* zidentyfikowane zostały w promotorach aż 47% par genów nakładających się. Wynik taki może sugerować nieco odmienne znaczenie tego czynnika transkrypcyjnego w regulacji nakładania się genów u obu gatunków.

Jeśli zidentyfikowane geny nakładające się regulowane są przez promotory dwukierunkowe lub niezależne promotory współdzielące rejony regulatorowe, to zgodnie z opublikowanymi pracami można by się spodziewać, że będą one ulegały ko-ekspresji[104, 154, 160, 170, 269-271]. Analiza nakładania się genów w 73 ludzkich i 10 mysich organach wykazała jednak, że odpowiednio 98% i 95% par genów ulegających ekspresji z nakładających się miejsc TSS w jednej bibliotece, jest w innych bibliotekach reprezentowane jedynie przez jeden gen. Tak wysoka liczba tego typu przypadków nie potwierdza wspomnianego wcześniej założenia o ko-ekspresji genów nakładających się na końcach 5' i ich kontroli przez promotory dwukierunkowe. Z drugiej jednak strony Rhee i Pugh[272] pokazali, że promotory dwukierunkowe mogą do pewnego stopnia podlegać zróżnicowanej regulacji w każdym z kierunków. Potwierdziły to badania Bagchi i Iyer[215], które wykazały, że promotory dwukierunkowe mogą w niektórych tkankach inicjować transkrypcję w dwóch, a w innych w jednym kierunku, co związane może być np. z działaniem różnych czynników transkrypcyjnych. Być może nie bez znaczenia jest zatem odkrycie w promotorach genów nakładających się u człowieka, miejsc wiązania dla ponad dwukrotnie większej liczby czynników transkrypcyjnych różnego rodzaju w porównaniu z pozostałymi genami.

W przypadku tak bliskiej lokalizacji wymagana może być bowiem bardziej precyzyjna kontrola ich aktywacji, tak aby transkrypcja nie została przypadkowo zainicjowana w niewłaściwym kierunku. Warto zauważyć również, że geny nakładające się mają średnio większą liczbę promotorów przypadających na gen, co również może mieć związek z większym skomplikowaniem regulacji inicjacji ich transkrypcji. Reasumując, otrzymane wyniki sugerują, że przynajmniej część genów nakładających się może podlegać regulacji przez dwukierunkowe rejony promotorowe, lecz z pewnością nie jest to regułą dla wszystkich zidentyfikowanych tutaj genów.

Wiele genów, szczególnie tych, których rejony promotorowe znajdują się w wyspach CpG, wykorzystuje alternatywne miejsc startu transkrypcji[185, 254]. Zmiana profilu ekspresji pary genów z nienakładających się do nakładających jest najczęściej związana ze zmianą promotora na alternatywny bądź aktywacją dodatkowych rejonów promotorowych przez jeden lub oba geny z pary. Analiza ekspresji wykazała, że wykorzystanie większej liczby promotorów wiąże się z podwyższeniem poziomu ekspresji genów. Zostało to pokazane zarówno na przykładzie danych TSS-seq jak i RNA-Seq. Mechanizm ten może być zatem wykorzystywany przez komórkę w momencie, gdy potrzebna jest większa ilość kodowanego przez dany gen białka, niezależnie od tego czy efektem ubocznym może być nakładanie się genów. W pracy pokazano także, że geny ulegające nakładaniu mają średnio wyższy poziom ekspresji niż te same geny, gdy ich ekspresja zachodzi z wykorzystaniem nienakładających się miejsc startu transkrypcji. Jak wykazała analiza porównawcza poziomu ekspresji 73 par genów które ulegały ekspresji z wykorzystaniem nakładających miejsc TSS w części bibliotek i ekspresji tych samych genów bez nakładania w innych bibliotekach, różnica ta nie wynika z faktu, że podczas nakładania wykorzystywanych jest przeciętnie więcej promotorów. Może to sugerować, że potencjalna interferencja transkrypcji w rejonie nakładania, nie ma tak dużego wpływu na ostateczny poziom ekspresji nakładających się genów. Warto tutaj przypomnieć, że w przypadku ekspresji tylko jednego z genów z analizowanych 73 par, jej poziom jest porównywalny do poziomu genów gdy ulegają ekspresji z nakładających się TSS. Nasuwa to więc ciekawe pytanie: dlaczego aktywacja nienakładających się promotorów wiąże się z niższym poziomem ekspresji?

Alternatywny splicing a nakładanie się genów

Zjawisko interferencji transkrypcyjnej, które sugerowane jest w przypadku nakładania się genów, może przyczyniać się do spowolnienia tempa elongacji polimerazy RNA II[60, 63, 64]. To z kolei umożliwia rozpoznanie słabszych sygnałów alternatywnego splicingu[65-69].

Do tej pory przykłady zależności alternatywnego splicingu od nakładania się genów zidentyfikowane zostały zarówno dla par genów utworzonych przez kodujące i niekodujące RNA[71], jak również dla par genów utworzonych przez dwa geny kodujące białka[70, 273]. W przedstawionej pracy przeprowadzono analizę mającą na celu sprawdzenie czy związek alternatywnego splicingu i zjawiska nakładania zachodzi również w przypadku którejś ze zidentyfikowanych nakładających się końcami 5' par genów kodujących białka. Przeprowadzona analiza pozwoliła na zidentyfikowanie trzech par genów, w których alternatywny splicing jednego z genów w parze mógłby być powiązany ze zjawiskiem nakładania się genów. Analiza wzorców polimerazy RNA II zdaje się wykluczać tutaj wpływ interferencji transkrypcji. Wartym dalszego zbadania, szczególnie w przypadku dokładniej opisanej pary genów *LRRC46* i *MRPL10,* pozostaje jednak potencjał do regulacji splicingu poprzez formowanie dupleksów RNA:RNA i maskowanie miejsc splicingowych

Nakładanie się genów a struktura chromatyny i interferencja transkrypcji

Przedstawione powyżej wyniki nie wykluczają jednoznacznie tego, że przynajmniej w przypadku niektórych z par pewne zakłócenia transkrypcji mogą mieć miejsce. Dlatego też sprawdzono na ile dane dotyczące aktywacji TSS i poziomu ekspresji odzwierciedlone są przez stan chromatyny, której stopień upakowania przekłada się na aktywność transkrypcyjną. Przykładowo trimetylacja 4 lizyny i acetylacja 27 lizyny histonu H3 powiązane zostały z aktywnymi promotorami[132, 143-145], podczas gdy już trimetylacja obu tych miejsc skorelowana jest z promotorami wyciszonymi[147, 148]. Analiza współwystępujących na danych nukleosomach kombinacji modyfikacji histonów i funkcjonalnych implikacji tych modyfikacji, czyli tzw. kodu histonowego, była wielokrotnie wykorzystywana w badaniach rejonów transkrypcyjnych[34, 132, 143-145, 147-149]. Również w odniesieniu do nakładających się genów kod histonowy używany był do oszacowania siły promotorów. Przykładowo Conley i Jordan[34] przeanalizowali tysiące nakładających się promotorów u człowieka i określali ich siłę na podstawie obserwacji modyfikacji histonów o aktywującym (np. H3K27ac, H3K4me3 czy H3K4me2) lub wyciszającym charakterze (np. H3K27me3), klasyfikując je odpowiednio jako silne i słabe. Dodatkowo zauważyli oni, że silne promotory były skorelowane z aktywnością polimerazy RNA II, podczas gdy słabe nie odznaczały się taką aktywnością. W swoich badaniach wykazali oni również nadreprezentację nakładających się promotorów o takim samym charakterze, tzn. częściej spotykano pary dwóch promotorów silnych lub dwóch słabych, aniżeli kombinację silnego i słabego promotora. Zasugerowali oni na tej podstawie, że zjawisko nakładania ma charakter raczej aktywujący aniżeli wyciszający[34].

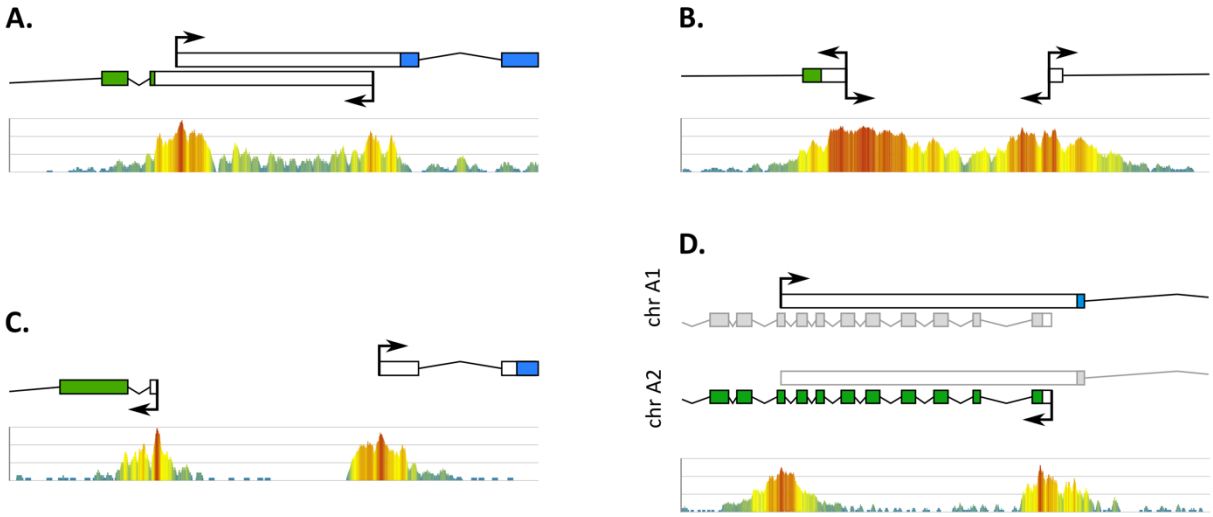
Wniosek ten jest zgodny z wyraźnie wyższym poziomem ekspresji nakładających się genów, jaki zaobserwowano w wyniku przeprowadzonych w trakcie realizacji tej pracy analiz.

Aby określić na ile stan chromatyny odzwierciedla fakt nakładania się genów oraz jaki wpływ nakładanie się genów ma na aktywność polimerazy, we wszystkich bibliotekach gruczolakoraka płuc przeprowadzono klasteryzację genów na podstawie sygnałów epigenetycznych oraz aktywności polimerazy RNA II. W każdej bibliotece uwzględniono rejony wszystkich par genów, które nakładały się przynajmniej w jednej z nich bez względu na to czy w danej bibliotece ulegały ekspresji czy nie. Otrzymane klastry reprezentowały trzy główne modele, a jednym z głównych czynników różnicującym te modele była siła sygnału aktywności polimerazy w badanym rejonie. Bliższa analiza wykazała, co jest szczególnie interesujące, że sygnał, który może świadczyć o zakłóceniach transkrypcji obserwowany był zarówno w przypadku gdy geny się nakładały jak i w takich sytuacjach gdy nakładania nie było. Podobnie brak takiego sygnału cechował zarówno niektóre nakładające się jak i nienakładające się w danych bibliotekach pary.

Obecność sygnału interferencji transkrypcji w przypadku, gdy geny wykorzystywały nakładające się miejsca TSS, nie wymaga większego wyjaśniania gdyż powodem zakłócenia może być kolizja polimeraz aktywnie transkrybujących geny w przeciwnych kierunkach na tym samym odcinku DNA (rycina 39 A). Podniesiony sygnał aktywności polimerazy w obszarze pomiędzy aktywnymi nienakładającymi się promotorami może wiązać się z obecnością promotorów dwukierunkowych typu PROMPT[155-159]. W tym przypadku sygnał może być wynikiem transkrypcji niestabilnych RNA, zachodzącej w kierunku przeciwnym do transkrypcji genu kodującego białko (rycina 39 B). W sytuacji, gdy nienakładające się promotory nie należą do typu PROMPT takiego sygnału możemy nie obserwować, tak jak to przedstawiono na rycinie 39 C.

Najtrudniejszą do wytłumaczenia wydaje się być sytuacja, w której zachodzi nakładanie się genów, ale brak jest sygnału interferencji transkrypcji. Można zadać pytanie w jaki sposób geny te unikają takiego zakłócenia. Hipotetycznie mogłoby to następować w dwóch przypadkach. Po pierwsze, gdy gen wykorzystuje do ekspresji w danym momencie więcej niż jeden alternatywny promotor, z czego nie wszystkie znajdują się w nakładających rejonach. W konsekwencji jedynie część cząsteczek mRNA będzie transkrybowana z nakładających się miejsc. Jak pokazano na przykładzie par genów *ANAPC16* i *ASCC1*, *MRPS18B* i *PPP1R10* oraz *DYNLL1* i *SRSF9*, część poziomu ekspresji przypisanego do rejonu nakładania się genów, którą zbadano z wykorzystaniem współczynników OR oraz JoinedOR, może się skrajnie różnić między badanymi bibliotekami. Doskonałym przykładem

jest tutaj para *PDAP1* i *BUD3,* której współczynnik JoinedOR w bibliotece PC14 wynosił zaledwie 0,003. Oznacza to, że transkrypcja z nakładających się miejsc była bardzo znikoma i w związku z tym interferencja transkrypcji może nie zachodzić. Drugim wyjaśnieniem, zgodnie z wynikami wielu badań[213, 214, 274-278], może być ekspresja monoalleliczna, której mogą ulegać także nakładające się geny. W sytuacji, w której każdy z genów z pary jest transkrybowany na innym z homologicznych chromosomów, fakt ich nakładania się nie ma znaczenia w aspekcie zakłócenia transkrypcji (rycina 39 D). Doskonałym tego przykładem jest para *FOXRED1* oraz *SRPR,* których potencjalnie monoalleliczną ekspresję wykazano w przypadku dwóch próbek gruczolakoraka płuc. Możliwość takiej ekspresji wzmacnia fakt, iż proporcje odczytów z poszczególnymi allelami odpowiadają proporcjom poziomu ekspresji tych genów. Analiza aktywności polimerazy w rejonie nakładania się tych genów, zgodnie z oczekiwaniem, nie wykazała podniesionego poziomu odczytów.



Rycina 39. ***Modele potencjalnie wyjaśniające występowanie lub brak występowania spowolnienia polimerazy RNA II, związanego z zachodzeniem zakłócenia transkrypcyjnego****. A) Para genów nakładających się, widoczna obecność sygnału spowolnienia polimerazy RNA II; B) Para genów wykorzystujących do ekspresji promotory dwukierunkowe typu PROMPT, widoczna obecność sygnału spowolnienia polimerazy RNA II; C) Para nienakładających się genów wykorzystujących promotory jednokierunkowe, brak obecności sygnału spowolnienia polimerazy RNA II; D) Monoalleliczna ekspresja genów nakładających się z dwóch chromosomów homologicznych A1 i A2. Brak obecności sygnału spowolnienia polimerazy RNA II.*

7. Wnioski

W niniejszej pracy doktorskiej zidentyfikowano i konsekwentnie przeanalizowano 695 ludzkich i mysich par genów nakładających się końcami 5', z których oba geny w parze kodowały białka. Wykazano, że zjawisko nakładania nie jest cechą stałą wielu par genów, co prowadzi do tego, że ta sama para genów może ulegać ekspresji z wykorzystaniem nakładających się miejsc TSS w jednej bibliotece i nienakładających się, alternatywnych rejonów promotorowych w innej. Charakterystyka nakładania się w każdej z bibliotek, nawet dla tej samej pary, może być skrajnie różna zarówno pod względem stopnia nakładania się genów, jak również obszaru zajmowanego przez rejon nakładania. Wyniki badań sugerują, że nakładające się pary genów odznaczają się w stosunku do innych genów znacznie bardziej skomplikowanym systemem regulacji poziomu ich ekspresji. W porównaniu do innych genów, rejony promotorowe genów nakładających regulowane są u człowieka przez średnio ponad dwukrotnie większą liczbę różnych czynników transkrypcyjnych, z których część powiązana została z regulacją promotorów dwukierunkowych. W przypadku części par genów wykazano zwiększoną aktywność polimerazy RNA II w rejonie nakładania, co może świadczyć o udziale zjawiska interferencji transkrypcyjnej w regulacji poziomu ekspresji genów.

Bardzo ciekawą obserwacją jest to, że nakładające się geny mają przeciętnie wyższy poziom ekspresji niż te same geny ulegające transkrypcji z nienakładających się TSS. Przy zakłóceniu transkrypcji, wykazanym przynajmniej w niektórych przypadkach, oraz biorąc pod uwagę wcześniejsze doniesienia można było spodziewać się efektu odwrotnego. Jednakże, jak wskazują wyniki przedstawionych badań, zakłócenie transkrypcji może występować także w przypadku gdy geny wykorzystują nienakładające się promotory. W efekcie wpływ interferencji na ekspresję może być niewidoczny przy grupowym porównywaniu poziomu ekspresji genów nakładających i nienakładających się. Istotny wpływ natomiast na wyższy poziom ekspresji genów nakładających się może mieć fakt wykorzystywania bardziej oddalonych promotorów. Promotory te w wielu wypadkach nie są szczególnie silne, ale działając wraz z promotorami nienakładającymi, co jest często obserwowane, mogą wystarczająco efektywnie wzmacniać ekspresję. Wykorzystywanie większej liczby promotorów jest, jak wskazują wyniki, silnie skorelowane z wyższą ekspresją. Natomiast te z nakładających się promotorów, które działają w pojedynkę są najprawdopodobniej przeciętnie silniejszymi promotorami na co może wskazywać średnio wyższa ekspresja nakładających się genów transkrybowanych z jednego TSS w porównaniu z genami, które także wykorzystują jeden TSS, ale nie nakładają się. Wynikający

z przeprowadzonych badań wniosek potwierdza hipotezę o aktywującym charakterze zjawiska nakładania się[34].

Mimo iż przeprowadzone badania nie dostarczyły ostatecznej odpowiedzi odnośnie funkcjonalności zjawiska nakładania się genów w skali globalnej, pokazały one, że ekspresja tych genów może podlegać ściślejszej kontroli niż pozostałych genów, co niewątpliwie powiązane jest z wyjątkową architekturą genomową w jakiej geny te się znalazły.

8. Spis rycin

9. Spis tabel

Tabele dodatkowe:

10. Wykaz najczęściej używanych skrótów

ChIP-Seq – z ang. chromatin immunoprecipitation sequencing
dsRNA – z ang. double stranded RNA
EST – z ang. expressed sequence tag
FPKM – z ang. fragments per kilobase of exon per million fragments mapped
FTP – z ang. file transfer protocol
JoinedOR – z ang. joined overlap ratio
lncRNA – z ang. long non-coding RNA
miRNA – z ang. micro RNA
mRNA – z ang. messenger RNA
NAT – z ang. natural antisense transcript
NFR – z ang. nucleosome free region
OR – z ang. overlap ratio
ppm – z ang. parts per million
PROMPT – z ang. promoter-upstream transcript
RNA-Seq – z ang. RNA sequencing
RNA:RNA - oddziaływania między cząsteczkami RNA
SDI - z ang. sitting duck interference
TSS – z ang. transcription start site
UTR – z ang. untranslated region

11. Literatura

1. Makalowska, I., Lin, C.F. & Makalowski, W. Overlapping genes in vertebrate genomes. *Comput Biol Chem* **29**, 1-12 (2005).
2. Faghihi, M.A. & Wahlestedt, C. Regulatory roles of natural antisense transcripts. *Nat Rev Mol Cell Biol* **10**, 637-643 (2009).
3. Bovre, K. & Szybalski, W. Patterns of convergent and overlapping transcription within the b2 region of coliphage lambda. *Virology* **38**, 614-626 (1969).
4. Barrell, B.G., Air, G.M. & Hutchison, C.A., 3rd Overlapping genes in bacteriophage phiX174. *Nature* **264**, 34-41 (1976).
5. Sanger, F. et al. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687-695 (1977).
6. Henikoff, S., Keene, M.A., Fechtel, K. & Fristrom, J.W. Gene within a gene: nested Drosophila genes encode unrelated proteins on opposite DNA strands. *Cell* **44**, 33-42 (1986).
7. Spencer, C.A., Gietz, R.D. & Hodgetts, R.B. Overlapping transcription units in the dopa decarboxylase region of Drosophila. *Nature* **322**, 279-281 (1986).
8. Williams, T. & Fried, M. A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3' ends. *Nature* **322**, 275-279 (1986).
9. Mol, J.N. et al. Regulation of plant gene expression by antisense RNA. *FEBS Lett* **268**, 427-430 (1990).
10. Quesada, V., Ponce, M.R. & Micol, J.L. OTC and AUL1, two convergent and overlapping genes in the nuclear genome of Arabidopsis thaliana. *FEBS Lett* **461**, 101-106 (1999).
11. Osato, N. et al. Antisense transcripts with rice full-length cDNAs. *Genome Biol* **5**, R5 (2003).
12. Steigele, S. & Nieselt, K. Open reading frames provide a rich pool of potential natural antisense transcripts in fungal genomes. *Nucleic Acids Res* **33**, 5034-5044 (2005).
13. David, L. et al. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A* **103**, 5320-5325 (2006).
14. Shendure, J. & Church, G.M. Computational discovery of sense-antisense transcription in the human and mouse genomes. *Genome Biol* **3**, RESEARCH0044 (2002).
15. Zhou, C. & Blumberg, B. Overlapping gene structure of human VLCAD and DLG4. *Gene* **305**, 161-166 (2003).
16. Veeramachaneni, V., Makalowski, W., Galdzicki, M., Sood, R. & Makalowska, I. Mammalian overlapping genes: the comparative perspective. *Genome Res* **14**, 280-286 (2004).
17. Ge, X., Rubinstein, W.S., Jung, Y.C. & Wu, Q. Genome-wide analysis of antisense transcription with Affymetrix exon array. *BMC Genomics* **9**, 27 (2008).
18. Yelin, R. et al. Widespread occurrence of antisense transcription in the human genome. *Nat Biotechnol* **21**, 379-386 (2003).
19. Chen, J. et al. Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res* **32**, 4812-4820 (2004).
20. Zhang, Y., Liu, X.S., Liu, Q.R. & Wei, L. Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucleic Acids Res* **34**, 3465-3475 (2006).

21. Katayama, S. et al. Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564-1566 (2005).
22. Tager, A.M. & Luster, A.D. BLT1 and BLT2: the leukotriene B(4) receptors. *Prostaglandins Leukot Essent Fatty Acids* **69**, 123-134 (2003).
23. Lehner, B., Williams, G., Campbell, R.D. & Sanderson, C.M. Antisense transcripts in the human genome. *Trends Genet* **18**, 63-65 (2002).
24. Engstrom, P.G. et al. Complex Loci in human and mouse genomes. *PLoS Genet* **2**, e47 (2006).
25. Wood, E.J., Chin-Inmanu, K., Jia, H. & Lipovich, L. Sense-antisense gene pairs: sequence, transcription, and structure are not conserved between human and mouse. *Front Genet* **4**, 183 (2013).
26. Meyers, B.C. et al. Analysis of the transcriptional complexity of Arabidopsis thaliana by massively parallel signature sequencing. *Nat Biotechnol* **22**, 1006-1011 (2004).
27. Nobuta, K. et al. An expression atlas of rice mRNAs and small RNAs. *Nat Biotechnol* **25**, 473-477 (2007).
28. He, Y., Vogelstein, B., Velculescu, V.E., Papadopoulos, N. & Kinzler, K.W. The antisense transcriptomes of human cells. *Science* **322**, 1855-1857 (2008).
29. Arthanari, Y., Heintzen, C., Griffiths-Jones, S. & Crosthwaite, S.K. Natural antisense transcripts and long non-coding RNA in Neurospora crassa. *PLoS One* **9**, e91353 (2014).
30. Li, S., Liberman, L.M., Mukherjee, N., Benfey, P.N. & Ohler, U. Integrated detection of natural antisense transcripts using strand-specific RNA sequencing data. *Genome Res* **23**, 1730-1739 (2013).
31. Lu, T. et al. Strand-specific RNA-seq reveals widespread occurrence of novel cis-natural antisense transcripts in rice. *BMC Genomics* **13**, 721 (2012).
32. Luo, C. et al. Integrative analysis of chromatin states in Arabidopsis identified potential regulatory mechanisms for natural antisense transcript production. *Plant J* **73**, 77-90 (2013).
33. Gan, Q., Li, D., Liu, G. & Zhu, L. Identification of potential antisense transcripts in rice using conventional microarray. *Mol Biotechnol* **51**, 37-43 (2012).
34. Conley, A.B. & Jordan, I.K. Epigenetic regulation of human cis-natural antisense transcripts. *Nucleic Acids Res* **40**, 1438-1445 (2012).
35. Karmodiya, K., Krebs, A.R., Oulad-Abdelghani, M., Kimura, H. & Tora, L. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics* **13**, 424 (2012).
36. Ling, M.H., Ban, Y., Wen, H., Wang, S.M. & Ge, S.X. Conserved expression of natural antisense transcripts in mammals. *BMC Genomics* **14**, 243 (2013).
37. Rosikiewicz, W. & Makalowska, I. Biological functions of natural antisense transcripts. *Acta Biochim Pol* **63**, 665-673 (2016).
38. Shearwin, K.E., Callen, B.P. & Egan, J.B. Transcriptional interference--a crash course. *Trends Genet* **21**, 339-345 (2005).
39. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol* **14**, 103-105 (2007).
40. Werner, A. & Sayer, J.A. Naturally occurring antisense RNA: function and mechanisms of action. *Curr Opin Nephrol Hypertens* **18**, 343-349 (2009).

41. Cui, I. & Cui, H. Antisense RNAs and epigenetic regulation. *Epigenomics* **2**, 139-150 (2010).
42. Nishizawa, M., Okumura, T., Ikeya, Y. & Kimura, T. Regulation of inducible gene expression by natural antisense transcripts. *Front Biosci (Landmark Ed)* **17**, 938-958 (2012).
43. Scheele, C. et al. The human PINK1 locus is regulated in vivo by a non-coding natural antisense RNA during modulation of mitochondrial function. *BMC Genomics* **8**, 74 (2007).
44. Faghihi, M.A. et al. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat Med* **14**, 723-730 (2008).
45. Chung, D.W., Rudnicki, D.D., Yu, L. & Margolis, R.L. A natural antisense transcript at the Huntington's disease repeat locus regulates HTT expression. *Hum Mol Genet* **20**, 3467-3477 (2011).
46. Luo, J.H. et al. Transcriptomic and genomic analysis of human hepatocellular carcinomas and hepatoblastomas. *Hepatology* **44**, 1012-1024 (2006).
47. Pasmant, E., Sabbagh, A., Vidaud, M. & Bieche, I. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J* **25**, 444-448 (2011).
48. Morris, K.V., Santoso, S., Turner, A.M., Pastori, C. & Hawkins, P.G. Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. *PLoS Genet* **4**, e1000258 (2008).
49. Yu, W. et al. Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature* **451**, 202-206 (2008).
50. Khalil, A.M., Faghihi, M.A., Modarresi, F., Brothers, S.P. & Wahlestedt, C. A novel RNA transcript with antiapoptotic function is silenced in fragile X syndrome. *PLoS One* **3**, e1486 (2008).
51. Moseley, M.L. et al. Bidirectional expression of CUG and CAG expansion transcripts and intranuclear polyglutamine inclusions in spinocerebellar ataxia type 8. *Nat Genet* **38**, 758-769 (2006).
52. Michael, D.R. et al. The human hyaluronan synthase 2 (HAS2) gene and its natural antisense RNA exhibit coordinated expression in the renal proximal tubular epithelial cell. *J Biol Chem* **286**, 19523-19532 (2011).
53. Halley, P., Khorkova, O. & Wahlestedt, C. Natural antisense transcripts as therapeutic targets. *Drug Discov Today Ther Strateg* **10**, e119-e125 (2013).
54. Hobson, D.J., Wei, W., Steinmetz, L.M. & Svejstrup, J.Q. RNA polymerase II collision interrupts convergent transcription. *Mol Cell* **48**, 365-374 (2012).
55. Prescott, E.M. & Proudfoot, N.J. Transcriptional collision between convergent genes in budding yeast. *Proc Natl Acad Sci U S A* **99**, 8796-8801 (2002).
56. Wang, P., Yang, J., Ishihama, A. & Pittard, A.J. Demonstration that the TyrR protein and RNA polymerase complex formed at the divergent P3 promoter inhibits binding of RNA polymerase to the major promoter, P1, of the aroP gene of Escherichia coli. *J Bacteriol* **180**, 5466-5472 (1998).
57. Hirschman, J.E., Durbin, K.J. & Winston, F. Genetic evidence for promoter competition in Saccharomyces cerevisiae. *Mol Cell Biol* **8**, 4608-4615 (1988).
58. Conte, C., Dastugue, B. & Vaury, C. Promoter competition as a mechanism of transcriptional interference mediated by retrotransposons. *EMBO J* **21**, 3908-3916 (2002).

59. Lin, Q., Chen, Q., Lin, L., Smith, S. & Zhou, J. Promoter targeting sequence mediates enhancer interference in the Drosophila embryo. *Proc Natl Acad Sci U S A* **104**, 3237-3242 (2007).
60. Palmer, A.C., Egan, J.B. & Shearwin, K.E. Transcriptional interference by RNA polymerase pausing and dislodgement of transcription factors. *Transcription* **2**, 9-14 (2011).
61. Callen, B.P., Shearwin, K.E. & Egan, J.B. Transcriptional interference between convergent promoters caused by elongation over the promoter. *Mol Cell* **14**, 647-656 (2004).
62. Liang, S. et al. Activities of constitutive promoters in Escherichia coli. *J Mol Biol* **292**, 19-37 (1999).
63. Racanelli, A.C., Turner, F.B., Xie, L.Y., Taylor, S.M. & Moran, R.G. A mouse gene that coordinates epigenetic controls and transcriptional interference to achieve tissue-specific expression. *Mol Cell Biol* **28**, 836-848 (2008).
64. Kaer, K. & Speek, M. Intronic retroelements: Not just "speed bumps" for RNA polymerase II. *Mob Genet Elements* **2**, 154-157 (2012).
65. Dujardin, G. et al. How slow RNA polymerase II elongation favors alternative exon skipping. *Mol Cell* **54**, 683-690 (2014).
66. Saint-Andre, V., Batsche, E., Rachez, C. & Muchardt, C. Histone H3 lysine 9 trimethylation and HP1gamma favor inclusion of alternative exons. *Nat Struct Mol Biol* **18**, 337-344 (2011).
67. de la Mata, M. et al. A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell* **12**, 525-532 (2003).
68. Proudfoot, N.J. Dawdling polymerases allow introns time to splice. *Nat Struct Biol* **10**, 876-878 (2003).
69. Howe, K.J., Kane, C.M. & Ares, M., Jr. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae. *RNA* **9**, 993-1006 (2003).
70. Hastings, M.L., Milcarek, C., Martincic, K., Peterson, M.L. & Munroe, S.H. Expression of the thyroid hormone receptor gene, erbAalpha, in B lymphocytes: alternative mRNA processing is independent of differentiation but correlates with antisense RNA levels. *Nucleic Acids Res* **25**, 4296-4300 (1997).
71. Beltran, M. et al. A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes Dev* **22**, 756-769 (2008).
72. Ebralidze, A.K. et al. PU.1 expression is modulated by the balance of functional sense and antisense RNAs regulated by a shared cis-regulatory element. *Genes Dev* **22**, 2085-2092 (2008).
73. Uchida, T. et al. Prolonged hypoxia differentially regulates hypoxia-inducible factor (HIF)-1alpha and HIF-2alpha expression in lung epithelial cells: implication of natural antisense HIF-1alpha. *J Biol Chem* **279**, 14871-14878 (2004).
74. Faghihi, M.A. et al. Evidence for natural antisense transcript-mediated inhibition of microRNA function. *Genome Biol* **11**, R56 (2010).
75. Wang, G.Q. et al. Sirt1 AS lncRNA interacts with its mRNA to inhibit muscle formation by attenuating function of miR-34a. *Sci Rep* **6**, 21865 (2016).

76. Stazic, D., Lindell, D. & Steglich, C. Antisense RNA protects mRNA from RNase E degradation by RNA-RNA duplex formation during phage infection. *Nucleic Acids Res* **39**, 4890-4899 (2011).
77. Werner, A. et al. Contribution of natural antisense transcription to an endogenous siRNA signature in human cells. *BMC Genomics* **15**, 19 (2014).
78. Carlile, M. et al. Strand selective generation of endo-siRNAs from the Na/phosphate transporter gene Slc34a1 in murine tissues. *Nucleic Acids Res* **37**, 2274-2282 (2009).
79. Borsani, O., Zhu, J., Verslues, P.E., Sunkar, R. & Zhu, J.K. Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell* **123**, 1279-1291 (2005).
80. Zhang, X. et al. Genome-wide analysis of plant nat-siRNAs reveals insights into their distribution, biogenesis and function. *Genome Biol* **13**, R20 (2012).
81. Yu, D., Meng, Y., Zuo, Z., Xue, J. & Wang, H. NATpipe: an integrative pipeline for systematical discovery of natural antisense transcripts (NATs) and phase-distributed nat-siRNAs from de novo assembled transcriptomes. *Sci Rep* **6**, 21666 (2016).
82. Lee, H.C. et al. Diverse pathways generate microRNA-like RNAs and Dicer-independent small interfering RNAs in fungi. *Mol Cell* **38**, 803-814 (2010).
83. Tam, O.H. et al. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**, 534-538 (2008).
84. Watanabe, T. et al. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453**, 539-543 (2008).
85. Okamura, K., Robine, N., Liu, Y., Liu, Q. & Lai, E.C. R2D2 organizes small regulatory RNA pathways in Drosophila. *Mol Cell Biol* **31**, 884-896 (2011).
86. Ling, K.H. et al. Derivation of an endogenous small RNA from double-stranded Sox4 sense and natural antisense transcripts in the mouse brain. *Genomics* **107**, 88-99 (2016).
87. Peters, N.T., Rohrbach, J.A., Zalewski, B.A., Byrkett, C.M. & Vaughn, J.C. RNA editing and regulation of Drosophila 4f-rnp expression by sas-10 antisense readthrough mRNA transcripts. *RNA* **9**, 698-710 (2003).
88. Cebrat, M. et al. Mechanism of lymphocyte-specific inactivation of RAG-2 intragenic promoter of NWC: implications for epigenetic control of RAG locus. *Mol Immunol* **45**, 2297-2306 (2008).
89. Li, K. & Ramchandran, R. Natural antisense transcript: a concomitant engagement with protein-coding transcript. *Oncotarget* **1**, 447-452 (2010).
90. Wight, M. & Werner, A. The functions of natural antisense transcripts. *Essays Biochem* **54**, 91-101 (2013).
91. Lee, J.T., Davidow, L.S. & Warshawsky, D. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat Genet* **21**, 400-404 (1999).
92. Ohhata, T. et al. Histone H3 Lysine 36 Trimethylation Is Established over the Xist Promoter by Antisense Tsix Transcription and Contributes to Repressing Xist Expression. *Mol Cell Biol* **35**, 3909-3920 (2015).
93. Modarresi, F. et al. Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation. *Nat Biotechnol* **30**, 453-459 (2012).
94. Wang, J. et al. Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. *Nature* **431**, 1 p following 757; discussion following 757 (2004).

95. Wakano, C., Byun, J.S., Di, L.J. & Gardner, K. The dual lives of bidirectional promoters. *Biochim Biophys Acta* **1819**, 688-693 (2012).
96. Mitra, S.A., Mitra, A.P. & Triche, T.J. A central role for long non-coding RNA in cancer. *Front Genet* **3**, 17 (2012).
97. Thomas, M.C. & Chiang, C.M. The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol* **41**, 105-178 (2006).
98. Gupta, K., Sari-Ak, D., Haffke, M., Trowitzsch, S. & Berger, I. Zooming in on Transcription Preinitiation. *J Mol Biol* **428**, 2581-2591 (2016).
99. Smale, S.T. & Kadonaga, J.T. The RNA polymerase II core promoter. *Annu Rev Biochem* **72**, 449-479 (2003).
100. Kadonaga, J.T. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip Rev Dev Biol* **1**, 40-51 (2012).
101. Javahery, R., Khachi, A., Lo, K., Zenzie-Gregory, B. & Smale, S.T. DNA sequence requirements for transcriptional initiator activity in mammalian cells. *Mol Cell Biol* **14**, 116-127 (1994).
102. Smale, S.T. & Baltimore, D. The "initiator" as a transcription control element. *Cell* **57**, 103-113 (1989).
103. Corden, J. et al. Promoter sequences of eukaryotic protein-coding genes. *Science* **209**, 1406-1414 (1980).
104. Carninci, P. et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**, 626-635 (2006).
105. Cooper, S.J., Trinklein, N.D., Anton, E.D., Nguyen, L. & Myers, R.M. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res* **16**, 1-10 (2006).
106. Burley, S.K. The TATA box binding protein. *Curr Opin Struct Biol* **6**, 69-75 (1996).
107. Deng, W. & Roberts, S.G. TFIIB and the regulation of transcription by RNA polymerase II. *Chromosoma* **116**, 417-429 (2007).
108. Lagrange, T., Kapanidis, A.N., Tang, H., Reinberg, D. & Ebright, R.H. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev* **12**, 34-44 (1998).
109. Deng, W. & Roberts, S.G. A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes Dev* **19**, 2418-2423 (2005).
110. Evans, R., Fairley, J.A. & Roberts, S.G. Activator-mediated disruption of sequence-specific DNA contacts by the general transcription factor TFIIB. *Genes Dev* **15**, 2945-2949 (2001).
111. Deng, W. & Roberts, S.G. Core promoter elements recognized by transcription factor IIB. *Biochem Soc Trans* **34**, 1051-1053 (2006).
112. Burke, T.W. & Kadonaga, J.T. Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev* **10**, 711-724 (1996).
113. Kutach, A.K. & Kadonaga, J.T. The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. *Mol Cell Biol* **20**, 4754-4764 (2000).
114. Burke, T.W. & Kadonaga, J.T. The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. *Genes Dev* **11**, 3020-3031 (1997).

115. Shao, H. et al. Core promoter binding by histone-like TAF complexes. *Mol Cell Biol* **25**, 206-219 (2005).
116. Theisen, J.W., Lim, C.Y. & Kadonaga, J.T. Three key subregions contribute to the function of the downstream RNA polymerase II core promoter. *Mol Cell Biol* **30**, 3471-3479 (2010).
117. Juven-Gershon, T., Hsu, J.Y. & Kadonaga, J.T. Caudal, a key developmental regulator, is a DPE-specific transcriptional factor. *Genes Dev* **22**, 2823-2830 (2008).
118. Gershenzon, N.I. & Ioshikhes, I.P. Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics* **21**, 1295-1300 (2005).
119. Lim, C.Y. et al. The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev* **18**, 1606-1617 (2004).
120. Parry, T.J. et al. The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes Dev* **24**, 2013-2018 (2010).
121. Butler, J.E. & Kadonaga, J.T. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev* **16**, 2583-2592 (2002).
122. Zabidi, M.A. et al. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556-559 (2015).
123. Ni, T. et al. A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat Methods* **7**, 521-527 (2010).
124. Rach, E.A., Yuan, H.Y., Majoros, W.H., Tomancak, P. & Ohler, U. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. *Genome Biol* **10**, R73 (2009).
125. Ohler, U. Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. *Nucleic Acids Res* **34**, 5943-5950 (2006).
126. Kadonaga, J.T. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* **116**, 247-257 (2004).
127. Smale, S.T., Schmidt, M.C., Berk, A.J. & Baltimore, D. Transcriptional activation by Sp1 as directed through TATA or initiator: specific requirement for mammalian transcription factor IID. *Proc Natl Acad Sci U S A* **87**, 4509-4513 (1990).
128. Emami, K.H., Burke, T.W. & Smale, S.T. Sp1 activation of a TATA-less promoter requires a species-specific interaction involving transcription factor IID. *Nucleic Acids Res* **26**, 839-846 (1998).
129. Rach, E.A. et al. Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genet* **7**, e1001274 (2011).
130. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* **13**, 233-245 (2012).
131. Yuan, G.C. et al. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* **309**, 626-630 (2005).
132. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837 (2007).
133. Mavrich, T.N. et al. Nucleosome organization in the Drosophila genome. *Nature* **453**, 358-362 (2008).
134. Hoskins, R.A. et al. Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Res* **21**, 182-192 (2011).

135. Jin, C. et al. H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nat Genet* **41**, 941-945 (2009).
136. Chen, P. et al. H3.3 actively marks enhancers and primes gene transcription via opening higher-ordered chromatin. *Genes Dev* **27**, 2109-2124 (2013).
137. Weber, C.M., Ramachandran, S. & Henikoff, S. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Mol Cell* **53**, 819-830 (2014).
138. Bonisch, C. & Hake, S.B. Histone H2A variants in nucleosomes and chromatin: more or less stable? *Nucleic Acids Res* **40**, 10719-10741 (2012).
139. Jin, C. & Felsenfeld, G. Nucleosome stability mediated by histone variants H3.3 and H2A.Z. *Genes Dev* **21**, 1519-1529 (2007).
140. Haberle, V. & Lenhard, B. Promoter architectures and developmental gene regulation. *Semin Cell Dev Biol* **57**, 11-23 (2016).
141. de Ruijter, A.J., van Gennip, A.H., Caron, H.N., Kemp, S. & van Kuilenburg, A.B. Histone deacetylases (HDACs): characterization of the classical HDAC family. *Biochem J* **370**, 737-749 (2003).
142. Hu, G. et al. H2A.Z facilitates access of active and repressive complexes to chromatin in embryonic stem cell self-renewal and differentiation. *Cell Stem Cell* **12**, 180-192 (2013).
143. Heintzman, N.D. et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-112 (2009).
144. Creyghton, M.P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936 (2010).
145. Mikkelsen, T.S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560 (2007).
146. Vermeulen, M. et al. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* **131**, 58-69 (2007).
147. Bernstein, B.E. et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315-326 (2006).
148. Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R. & Young, R.A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77-88 (2007).
149. Rando, O.J. Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Curr Opin Genet Dev* **22**, 148-155 (2012).
150. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215-216 (2012).
151. Song, J. & Chen, K.C. Spectacle: fast chromatin state annotation using spectral learning. *Genome Biol* **16**, 33 (2015).
152. Hoffman, M.M. et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9**, 473-476 (2012).
153. Perez-Lluch, S. et al. Absence of canonical marks of active chromatin in developmentally regulated genes. *Nat Genet* **47**, 1158-1167 (2015).
154. Balbin, O.A. et al. The landscape of antisense gene expression in human cancers. *Genome Res* **25**, 1068-1079 (2015).
155. Xu, Z. et al. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033-1037 (2009).

156. Preker, P. et al. PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. *Nucleic Acids Res* **39**, 7179-7193 (2011).
157. Core, L.J., Waterfall, J.J. & Lis, J.T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845-1848 (2008).
158. Seila, A.C. et al. Divergent transcription from active promoters. *Science* **322**, 1849-1851 (2008).
159. Preker, P. et al. RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322**, 1851-1854 (2008).
160. Trinklein, N.D. et al. An abundance of bidirectional promoters in the human genome. *Genome Res* **14**, 62-66 (2004).
161. Core, L.J. et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* **46**, 1311-1320 (2014).
162. Duttke, S.H. et al. Human promoters are intrinsically directional. *Mol Cell* **57**, 674-684 (2015).
163. Scruggs, B.S. et al. Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Mol Cell* **58**, 1101-1112 (2015).
164. Antequera, F. Structure, function and evolution of CpG island promoters. *Cell Mol Life Sci* **60**, 1647-1658 (2003).
165. Yang, M.Q. & Elnitski, L.L. Diversity of core promoter elements comprising human bidirectional promoters. *BMC Genomics* **9 Suppl 2**, S3 (2008).
166. Park, D., Morris, A.R., Battenhouse, A. & Iyer, V.R. Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements. *Nucleic Acids Res* **42**, 3736-3749 (2014).
167. Sharon, E. et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30**, 521-530 (2012).
168. Collins, P.J., Kobayashi, Y., Nguyen, L., Trinklein, N.D. & Myers, R.M. The ets-related transcription factor GABP directs bidirectional transcription. *PLoS Genet* **3**, e208 (2007).
169. Xie, X. et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338-345 (2005).
170. Lin, J.M. et al. Transcription factor binding and modified histones in human bidirectional promoters. *Genome Res* **17**, 818-827 (2007).
171. Sherwood, R.I. et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* **32**, 171-178 (2014).
172. Karolchik, D. et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493-496 (2004).
173. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774 (2012).
174. Suzuki, A. et al. DBTSS as an integrative platform for transcriptome, epigenome and genome sequence variation data. *Nucleic Acids Res* **43**, D87-91 (2015).
175. Yamashita, R., Sugano, S., Suzuki, Y. & Nakai, K. DBTSS: DataBase of Transcriptional Start Sites progress report in 2012. *Nucleic Acids Res* **40**, D150-154 (2012).

176. Yamashita, R. et al. Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Res* **21**, 775-789 (2011).
177. Tanimoto, K. et al. Genome-wide identification and annotation of HIF-1alpha binding sites in two cell lines using massively parallel sequencing. *Hugo J* **4**, 35-48 (2010).
178. Kanai, A. et al. Characterization of STAT6 target genes in human B cells and lung epithelial cells. *DNA Res* **18**, 379-392 (2011).
179. Tsuchihara, K. et al. Massive transcriptional start site analysis of human genes in hypoxia cells. *Nucleic Acids Res* **37**, 2249-2263 (2009).
180. Suzuki, A. et al. Aberrant transcriptional regulations in cancers: genome, transcriptome and epigenome analysis of lung adenocarcinoma cell lines. *Nucleic Acids Res* **42**, 13557-13572 (2014).
181. Wijaya, E., Frith, M.C., Suzuki, Y. & Horton, P. Recount: expectation maximization based error correction tool for next generation sequencing data. *Genome Inform* **23**, 189-201 (2009).
182. Toribio, A.L. et al. European Nucleotide Archive in 2016. *Nucleic Acids Res* **45**, D32-D36 (2017).
183. O'Leary, N.A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745 (2016).
184. Ayoubi, T.A. & Van De Ven, W.J. Regulation of gene expression by alternative promoters. *FASEB J* **10**, 453-460 (1996).
185. Davuluri, R.V., Suzuki, Y., Sugano, S., Plass, C. & Huang, T.H. The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet* **24**, 167-177 (2008).
186. Wheeler, D.L. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **29**, 11-16 (2001).
187. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
188. Andrews, S. (2010).
189. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360 (2015).
190. Kent, W.J. et al. The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).
191. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
192. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290-295 (2015).
193. Frazee, A.C. et al. Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat Biotechnol* **33**, 243-246 (2015).
194. Robinson, J.T. et al. Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26 (2011).
195. Thorvaldsdottir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192 (2013).
196. Pertea, M., Kim, D., Pertea, G.M., Leek, J.T. & Salzberg, S.L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* **11**, 1650-1667 (2016).

197. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993 (2011).
198. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
199. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
200. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
201. Yunshun Chen, D.M., Matthew Ritchie, Mark Robinson, Gordon K. Smyth edgeR: differential expression analysis of digital gene expression data. User's Guide. (2016).
202. Mathelier, A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **44**, D110-115 (2016).
203. Tan, G. & Lenhard, B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* **32**, 1555-1556 (2016).
204. Wasserman, W.W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **5**, 276-287 (2004).
205. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
206. Bailey, T. et al. Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput Biol* **9**, e1003326 (2013).
207. Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204-2207 (2010).
208. Eric Jones, T.O., Pearu Peterson and others. (2001).
209. Chou, C.H. et al. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res* **44**, D239-247 (2016).
210. Agarwal, V., Bell, G.W., Nam, J.W. & Bartel, D.P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4** (2015).
211. Down, T.A., Piipari, M. & Hubbard, T.J. Dalliance: interactive genome viewing on the web. *Bioinformatics* **27**, 889-890 (2011).
212. Kampstra, P. Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software* **28** (2008).
213. Deng, Q., Ramskold, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193-196 (2014).
214. Gimelbrant, A., Hutchinson, J.N., Thompson, B.R. & Chess, A. Widespread monoallelic expression on human autosomes. *Science* **318**, 1136-1140 (2007).
215. Bagchi, D.N. & Iyer, V.R. The Determinants of Directionality in Transcriptional Initiation. *Trends Genet* **32**, 322-333 (2016).
216. Juven-Gershon, T., Hsu, J.Y. & Kadonaga, J.T. Perspectives on the RNA polymerase II core promoter. *Biochem Soc Trans* **34**, 1047-1050 (2006).
217. Orekhova, A.S. & Rubtsov, P.M. Bidirectional promoters in the transcription of mammalian genomes. *Biochemistry (Mosc)* **78**, 335-341 (2013).

218. Anno, Y.N. et al. Genome-wide evidence for an essential role of the human Staf/ZNF143 transcription factor in bidirectional transcription. *Nucleic Acids Res* **39**, 3116-3127 (2011).
219. Mi, H. et al. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res* **38**, D204-210 (2010).
220. Gene Ontology, C. The Gene Ontology: enhancements for 2011. *Nucleic Acids Res* **40**, D559-564 (2012).
221. Wu, D. et al. Three-tiered role of the pioneer factor GATA2 in promoting androgen-dependent gene expression in prostate cancer. *Nucleic Acids Res* **42**, 3607-3622 (2014).
222. Matsunaga, H. et al. Essential Role of GATA2 in the Negative Regulation of Type 2 Deiodinase Gene by Liganded Thyroid Hormone Receptor beta2 in Thyrotroph. *PLoS One* **10**, e0142400 (2015).
223. Nakada, C., Satoh, S., Tabata, Y., Arai, K. & Watanabe, S. Transcriptional repressor foxl1 regulates central nervous system development by suppressing shh expression in zebra fish. *Mol Cell Biol* **26**, 7246-7257 (2006).
224. Zhang, G. et al. FOXL1, a novel candidate tumor suppressor, inhibits tumor aggressiveness and predicts outcome in human pancreatic cancer. *Cancer Res* **73**, 5416-5425 (2013).
225. Lee, H.K. et al. Nuclear factor I-C (NFIC) regulates dentin sialophosphoprotein (DSPP) and E-cadherin via control of Kruppel-like factor 4 (KLF4) during dentinogenesis. *J Biol Chem* **289**, 28225-28236 (2014).
226. Nassiri, M. et al. Repressors NFI and NFY participate in organ-specific regulation of von Willebrand factor promoter activity in transgenic mice. *Arterioscler Thromb Vasc Biol* **30**, 1423-1429 (2010).
227. Boam, D.S., Davidson, I. & Chambon, P. A TATA-less promoter containing binding sites for ubiquitous transcription factors mediates cell type-specific regulation of the gene for transcription enhancer factor-1 (TEF-1). *J Biol Chem* **270**, 19487-19494 (1995).
228. Terrados, G. et al. Genome-wide localization and expression profiling establish Sp2 as a sequence-specific transcription factor regulating vitally important genes. *Nucleic Acids Res* **40**, 7844-7857 (2012).
229. Hilger-Eversheim, K., Moser, M., Schorle, H. & Buettner, R. Regulatory roles of AP-2 transcription factors in vertebrate development, apoptosis and cell-cycle control. *Gene* **260**, 1-12 (2000).
230. Bosher, J.M., Totty, N.F., Hsuan, J.J., Williams, T. & Hurst, H.C. A family of AP-2 proteins regulates c-erbB-2 expression in mammary carcinoma. *Oncogene* **13**, 1701-1707 (1996).
231. Jin, T. & Liu, L. The Wnt signaling pathway effector TCF7L2 and type 2 diabetes mellitus. *Mol Endocrinol* **22**, 2383-2392 (2008).
232. Qiu, H., Wang, F., Liu, C., Xu, X. & Liu, B. TEAD1-dependent expression of the FoxO3a gene in mouse skeletal muscle. *BMC Mol Biol* **12**, 1 (2011).
233. Mahanic, C.S., Budhavarapu, V., Graves, J.D., Li, G. & Lin, W.C. Regulation of E2 promoter binding factor 1 (E2F1) transcriptional activity through a deubiquitinating enzyme, UCH37. *J Biol Chem* **290**, 26508-26522 (2015).

234. Agarwal, S.K. et al. Menin interacts with the AP1 transcription factor JunD and represses JunD-activated transcription. *Cell* **96**, 143-152 (1999).
235. Bellizzi, D., Covello, G., Di Cianni, F., Tong, Q. & De Benedictis, G. Identification of GATA2 and AP-1 Activator elements within the enhancer VNTR occurring in intron 5 of the human SIRT3 gene. *Mol Cells* **28**, 87-92 (2009).
236. Vacik, T. & Raska, I. Alternative intronic promoters in development and disease. *Protoplasma* **254**, 1201-1206 (2017).
237. Tie, F. et al. CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development* **136**, 3131-3141 (2009).
238. Rada-Iglesias, A. et al. Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders. *Genome Res* **18**, 380-392 (2008).
239. Heintzman, N.D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-318 (2007).
240. Lee, T.I. et al. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301-313 (2006).
241. Quinodoz, M., Gobet, C., Naef, F. & Gustafson, K.B. Characteristic bimodal profiles of RNA polymerase II at thousands of active mammalian promoters. *Genome Biol* **15**, R85 (2014).
242. Mayer, A., Landry, H.M. & Churchman, L.S. Pause & go: from the discovery of RNA polymerase pausing to its functional implications. *Curr Opin Cell Biol* **46**, 72-80 (2017).
243. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
244. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**, 203-214 (2000).
245. Sayers, E.W. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **37**, D5-15 (2009).
246. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43**, D789-798 (2015).
247. Gray, K.A., Yates, B., Seal, R.L., Wright, M.W. & Bruford, E.A. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res* **43**, D1079-1085 (2015).
248. Flicek, P. et al. Ensembl 2013. *Nucleic Acids Res* **41**, D48-55 (2013).
249. Keshava Prasad, T.S. et al. Human Protein Reference Database--2009 update. *Nucleic Acids Res* **37**, D767-772 (2009).
250. Harrow, J.L. et al. The Vertebrate Genome Annotation browser 10 years on. *Nucleic Acids Res* **42**, D771-779 (2014).
251. Coordinators, N.R. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **45**, D12-D17 (2017).
252. Lu, Z., Cohen, K.B. & Hunter, L. Finding GeneRIFs via gene ontology annotations. *Pac Symp Biocomput*, 52-63 (2006).
253. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).
254. Kim, T.H. et al. A high-resolution map of active promoters in the human genome. *Nature* **436**, 876-880 (2005).

255. Tan, J.S., Mohandas, N. & Conboy, J.G. High frequency of alternative first exons in erythroid genes suggests a critical role in regulating gene function. *Blood* **107**, 2557-2561 (2006).
256. Makalowska, I., Lin, C.F. & Hernandez, K. Birth and death of gene overlaps in vertebrates. *BMC Evol Biol* **7**, 193 (2007).
257. Zhou, X., Sunkar, R., Jin, H., Zhu, J.K. & Zhang, W. Genome-wide identification and analysis of small RNAs originated from natural antisense transcripts in Oryza sativa. *Genome Res* **19**, 70-78 (2009).
258. Fietz, D. et al. Transfection of Sertoli cells with androgen receptor alters gene expression without androgen stimulation. *BMC Mol Biol* **16**, 23 (2015).
259. Kleinman, M.E. et al. Sequence- and target-independent angiogenesis suppression by siRNA via TLR3. *Nature* **452**, 591-597 (2008).
260. Grimm, D. et al. Fatality in mice due to oversaturation of cellular microRNA/short hairpin RNA pathways. *Nature* **441**, 537-541 (2006).
261. Olejniczak, M. et al. Sequence-non-specific effects generated by various types of RNA interference triggers. *Biochim Biophys Acta* **1859**, 306-314 (2016).
262. Jacobsen, L., Calvin, S. & Lobenhofer, E. Transcriptional effects of transfection: the potential for misinterpretation of gene expression data generated from transiently transfected cells. *Biotechniques* **47**, 617-624 (2009).
263. de Veer, M.J. et al. Functional classification of interferon-stimulated genes identified using microarrays. *J Leukoc Biol* **69**, 912-920 (2001).
264. Fensterl, V. & Sen, G.C. Interferons and viral infections. *Biofactors* **35**, 14-20 (2009).
265. Yla-Herttuala, S. & Kaikkonen, M. Does mass balance between sense and antisense transcripts fine-tune the outcome of gene expression? *EMBO Rep* **15**, 125-126 (2014).
266. Chen, J., Sun, M., Hurst, L.D., Carmichael, G.G. & Rowley, J.D. Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. *Trends Genet* **21**, 326-329 (2005).
267. Henz, S.R. et al. Distinct expression patterns of natural antisense transcripts in Arabidopsis. *Plant Physiol* **144**, 1247-1255 (2007).
268. Beck, C.F. & Warren, R.A. Divergent promoters, a common form of gene organization. *Microbiol Rev* **52**, 318-326 (1988).
269. Yamashita, R., Wakaguri, H., Sugano, S., Suzuki, Y. & Nakai, K. DBTSS provides a tissue specific dynamic view of Transcription Start Sites. *Nucleic Acids Res* **38**, D98-104 (2010).
270. Adachi, N. & Lieber, M.R. Bidirectional gene organization: a common architectural feature of the human genome. *Cell* **109**, 807-809 (2002).
271. Li, Y.Y. et al. Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance. *PLoS Comput Biol* **2**, e74 (2006).
272. Rhee, H.S. & Pugh, B.F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483**, 295-301 (2012).
273. Hastings, M.L., Ingle, H.A., Lazar, M.A. & Munroe, S.H. Post-transcriptional regulation of thyroid hormone receptor expression by cis-acting sequences and a naturally occurring antisense RNA. *J Biol Chem* **275**, 11507-11513 (2000).
274. Eckersley-Maslin, M.A. & Spector, D.L. Random monoallelic expression: regulating gene expression one allele at a time. *Trends Genet* **30**, 237-244 (2014).

275. Jiang, Y., Zhang, N.R. & Li, M. SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biol* **18**, 74 (2017).
276. Reinius, B. & Sandberg, R. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nat Rev Genet* **16**, 653-664 (2015).
277. Chess, A. Mechanisms and consequences of widespread random monoallelic expression. *Nat Rev Genet* **13**, 421-428 (2012).
278. Zwemer, L.M. et al. Autosomal monoallelic expression in the mouse. *Genome Biol* **13**, R10 (2012).

12. Aneks

Tabela Dodatkowa 1. ***Lista 73 ludzkich i 10 mysich bibliotek TSS-Seq.*** *Dla każdej biblioteki podano liczbę genów ulegających ekspresji z wykorzystaniem przynajmniej jednego miejsca TSS oraz całkowitą liczbę miejsc TSS aktywnych w danej bibliotece.*

	Nazwa biblioteki	Liczba genów ulegających ekspresji	Liczba aktywnych miejsc TSS
Organy i tkanki, *Homo sapiens*	Tkanka tłuszczowa	5450	5888
	Nadnercza	4988	5538
	Mózg 1	6153	6876
	Mózg 2	5363	5905
	Mózg 3	5050	5683
	Pierś	4594	5207
	Jelito	5112	5767
	Serce 1	3472	3656
	Serce 2	3658	3919
	Nerki 1	5433	5730
	Nerki 2	4627	5065
	Wątroba	3990	4249
	Płuca	6630	7472
	Limfa	5142	5578
	Mięśnie	4640	4956
	Jajniki	6267	6714
	Prostata	4698	5139
	Jądra	7147	8065
	Tarczyca	6351	7185
	Mózg płodu	5824	7261
	Serce płodu	4391	6639
	Nerki płodu	5271	6434
	Wątroba płodu	3539	3825
	Grasica płodu	5158	6927
Linie komórkowe, *Homo sapiens*	DLD1 Hipoksja HIF1-	5769	6479
	DLD1 Hipoksja HIF2-	5451	5831
	DLD1 Hipoksja	5602	6212
	DLD1 Normoksja HIF1-	5412	5755
	DLD1 Normoksja HIF2-	5811	6172
	DLD1 Normoksja	5486	5809
	Beas2B IL4+ STAT6	5112	5623
	Beas2B IL4- STAT6	5014	5460
	Beas2B IL4+ Parent	5230	6159

	Beas2B IL4- Parent	4969	5531
	Beas2B IL4- STAT6 siRNA	3070	3386
	Beas2B IL4- Kontrola siRNA	3108	3360
	Beas2B IL4+ STAT6 siRNA	3047	3293
	Beas2B IL4+ Kontrola siRNA	3255	3519
	Ramos IL4+	4859	5465
	Ramos IL4-	4713	5267
	MCF7 Normoksja	4707	5203
	MCF7 Hipoksja	4821	5270
	TIG3 Normoksja	5063	5558
	TIG3 Hipoksja	5034	5476
	HEK293 Normoksja	5033	5260
	HEK293 Hipoksja	5093	5412
	Hela	5181	5592
Gruczolakorak płuc, Homo sapiens	A427	6971	7725
	A549	6938	7457
	ABC1	6812	7389
	H1299	7087	7656
	H1437	7339	8154
	H1648	7253	7950
	H1650	7675	8459
	H1703	6870	7316
	H1819	7884	8695
	H1975	7288	7961
	H2126	7344	8156
	H2228	6939	7519
	H2347	7204	7826
	H322	7631	8447
	II18	7030	7620
	LC2/ad	7105	7670
	PC3	6307	6877
	PC7	6926	7584
	PC9	6909	7421
	PC14	6088	6570
	RERFLCad1	7414	8141
	RERFLCad2	8067	9231
	RERFLCKJ	7363	8079
	RERFLCMS	7771	8458
	RERFLCOK	7058	7647

	VMRC-LCD	7054	8484
Mus musculus	Mózg	3993	4679
	Serce	4035	4198
	Nerki	5515	5828
	Wątroba	4767	4991
	Śledziona	2747	2810
	Grasica	3701	4330
	Zarodek, 7 dni	6321	7633
	Zarodek, 11 dni	6895	8070
	Zarodek, 15 dni	6744	7510
	Zarodek, 17 dni	6621	7813

Tabela Dodatkowa 2. ***Liczba odczytów zmapowanych i niezmapowanych dla każdej z linii komórkowych gruczolakoraka płuc.***

Biblioteka	Liczba odczytów:	
	zmapowanych	niezmapowanych
A427	96216515	3255881
A549	51416257	3715681
ABC1	91576695	4621729
H1299	116764600	3861524
H1437	115184139	3488127
H1648	92921067	3188863
H1650	64659228	3707002
H1703	215310902	7529570
H1819	180193326	6841956
H1975	83760918	3000700
H2126	109247134	4094648
H2228	126384421	3721751
H2347	112779262	4135662
H322	122142708	4219722
II18	143702793	6846273
LC2ad	101570591	3373997
PC14	123395531	4993911
PC3	113336560	5689702
PC7	117481264	7478394
PC9	78404210	2674108
RERFLCad1	128424012	4176160
RERFLCad2	103709310	2752470
RERFLCKJ	137779846	4552162

RERFLCMS	124618373	3433377
RERFLCOK	76895027	2839895
VMRCLCD	115525002	3662358

Tabela Dodatkowa 3. ***Lista czynników transkrypcyjnych których motywy wiązania są nad lub niedoreprezentowane u człowieka****. W tabeli przedstawiono tylko wyniki istotne statystycznie.*

Czynnik transkrypcyjny	Liczba promotorów regulowanych przez ten czynnik pośród:		Wartość statystyki Z	Krotność zmiany
	genów nakładających	genów nienakładających		
YY1	3761	21226	51,37	1,84
SPIB	3247	16125	51,25	2,09
EBF1	2370	9324	50,59	2,63
NFIX	3533	19176	50,31	1,91
NKX2-8	2838	12916	50,22	2,28
TCF4	2700	12023	49,43	2,33
NFIC	3506	19441	48,73	1,87
FOXC1	3271	17417	47,82	1,94
TCF3	2398	10129	47,81	2,45
ZEB1	1863	6574	47,59	2,93
THAP1	2161	8884	45,73	2,52
MEIS1	3698	22647	45,65	1,69
E2F6	2426	10911	45,33	2,30
SNAI2	2175	9440	43,59	2,39
SP1	3276	19047	43,02	1,78
MZF1	3831	25379	42,80	1,56
KLF5	3402	20523	42,52	1,72
TFAP2A	3156	18132	42,18	1,80
SPI1	3960	27895	40,91	1,47
ETS1	4033	29000	40,80	1,44
GATA3	3152	18657	40,47	1,75
E2F1	1920	8320	40,24	2,39
FLI1	1562	5991	39,81	2,70
OTX1	1897	8324	39,42	2,36
ELK1	1207	3933	39,38	3,18
ID4	1363	5028	38,05	2,81
ETV5	1039	3188	37,97	3,38
E2F4	1628	6962	36,77	2,42
BARX1	2492	13994	35,63	1,84
TEAD3	1874	9003	35,45	2,16
SOX10	2238	11969	35,10	1,94

EHF	1430	5929	35,08	2,50
MGA	1017	3427	34,84	3,07
STAT3	1598	7151	34,67	2,31
TFAP2C	1040	3656	34,03	2,95
FOXD2	2648	15940	33,84	1,72
EGR1	2014	10480	33,77	1,99
FOXP3	2571	15433	33,16	1,73
NR2F1	690	1904	32,92	3,75
NKX2-3	1238	5060	32,70	2,53
MEIS2	1195	4821	32,45	2,57
ETV6	983	3555	32,18	2,86
MAX	1184	4949	31,15	2,48
HOXA5	1814	9538	31,05	1,97
ERG	788	2585	30,98	3,16
HOXB3	2011	11206	30,77	1,86
CEBPB	1271	5628	30,56	2,34
NFIA	1328	6066	30,31	2,27
ETV4	654	1989	29,78	3,41
USF1	959	3812	29,07	2,61
ETV1	602	1792	28,97	3,48
GATA2	4064	34802	28,70	1,21
TBX5	783	2831	28,41	2,86
ELK3	553	1608	28,20	3,56
BARHL2	1540	8053	28,14	1,98
FOXL1	2592	17222	28,09	1,56
ELF1	900	3610	27,81	2,58
KLF16	1131	5145	27,71	2,28
MNT	710	2512	27,48	2,93
NR4A2	724	2598	27,42	2,89
FOXO6	1568	8443	27,35	1,92
USF2	841	3324	27,18	2,62
LHX9	1668	9290	27,13	1,86
SP2	1408	7387	26,49	1,97
TEAD4	1276	6404	26,44	2,06
NRF1	1193	5799	26,44	2,13
FOXI1	1457	7905	25,83	1,91
EN1	1502	8302	25,64	1,87
SP3	908	3961	25,62	2,37
ETV3	469	1389	25,52	3,50
ELF5	767	3102	25,26	2,56
NFATC2	881	3847	25,17	2,37
HOXA2	1170	5876	25,09	2,06

CREB1	727	2893	24,94	2,60
NFKB1	670	2559	24,83	2,71
FOXO4	1263	6677	24,53	1,96
TBX1	575	2039	24,50	2,92
TEAD1	1068	5308	24,11	2,08
PITX3	710	2905	23,94	2,53
ERF	369	1064	22,99	3,59
SP8	619	2450	22,99	2,62
FOXO3	843	3947	22,68	2,21
CEBPA	932	4587	22,58	2,10
NKX6-1	1429	8440	22,43	1,75
HOXB2	1203	6677	22,22	1,87
ISL2	691	3040	21,84	2,35
MSX1	1085	5846	21,82	1,92
SREBF1	426	1445	21,76	3,05
EMX2	1101	5981	21,75	1,91
NRL	474	1730	21,56	2,84
DLX6	1204	6830	21,51	1,83
ISX	1204	6830	21,51	1,83
TFAP2B	470	1731	21,29	2,81
PRRX1	1057	5753	21,18	1,90
TFE3	547	2207	21,11	2,57
NKX3-1	757	3605	20,94	2,17
FOSL2	582	2498	20,45	2,41
STAT1	530	2222	19,95	2,47
MSX2	952	5262	19,48	1,87
NEUROD2	555	2447	19,36	2,35
NKX3-2	517	2219	19,20	2,41
HEY2	264	770	19,20	3,55
RUNX2	497	2125	18,88	2,42
BHLHE40	350	1247	18,82	2,91
HEY1	318	1079	18,69	3,05
ETV2	255	762	18,51	3,47
FOXH1	663	3340	18,12	2,06
MNX1	988	5825	17,97	1,76
JUNB	614	3016	17,96	2,11
FOXG1	655	3350	17,64	2,02
ESX1	569	2746	17,63	2,15
LMX1B	777	4287	17,40	1,88
RAX	747	4074	17,32	1,90
FOXA1	727	3930	17,28	1,92
NFATC3	538	2607	17,02	2,14

SREBF2	255	857	16,83	3,08
TBX15	282	1007	16,80	2,90
LBX2	534	2606	16,80	2,12
GABPA	214	653	16,65	3,39
REL	333	1331	16,43	2,59
ZNF740	259	923	16,11	2,91
TFEC	366	1582	15,89	2,40
MEOX1	579	3059	15,77	1,96
FOS	572	3014	15,72	1,97
LHX6	498	2523	15,41	2,04
LHX2	491	2508	15,12	2,03
FOXD1	393	1833	15,10	2,22
TFEB	318	1357	14,98	2,43
VAX1	560	3032	14,94	1,91
SPDEF	161	471	14,89	3,54
HINFP	227	832	14,67	2,83
JUND	474	2449	14,61	2,00
RFX5	366	1735	14,26	2,18
TCF7L2	256	1039	14,12	2,55
CDX2	602	3477	14,05	1,79
CLOCK	128	355	13,80	3,73
ZBTB18	118	310	13,76	3,94
CDX1	679	4154	13,59	1,69
ESRRA	175	603	13,59	3,01
EMX1	391	1980	13,56	2,05
MLXIPL	208	794	13,50	2,71
HOXC10	402	2129	12,92	1,96
LMX1A	370	1906	12,86	2,01
HMBOX1	263	1179	12,84	2,31
ZBTB7A	76	161	12,67	4,89
ZBTB33	85	200	12,56	4,40
FOSL1	353	1820	12,54	2,01
AR	187	734	12,44	2,64
GMEB2	156	565	12,25	2,86
BHLHE41	145	506	12,21	2,97
HOXA13	345	1835	11,85	1,95
HOXD11	252	1207	11,60	2,16
ZBTB7B	83	219	11,49	3,92
HNF4G	154	610	11,16	2,61
SP4	169	714	10,94	2,45
NR2C2	92	282	10,82	3,38
TFCP2	95	298	10,80	3,30

MAFK	197	904	10,77	2,26
INSM1	130	494	10,67	2,73
MEOX2	326	1826	10,63	1,85
EGR2	125	484	10,27	2,67
RELA	138	573	10,06	2,49
RORA	166	750	10,05	2,29
ZNF143	35	56	9,99	6,47
NFYB	194	957	9,76	2,10
HOXB13	250	1362	9,67	1,90
JUN	285	1676	9,15	1,76
HOXA10	278	1634	9,04	1,76
NFKB2	53	143	9,03	3,84
MEF2C	179	915	8,93	2,03
ELF3	65	208	8,77	3,24
ZBTB7C	61	189	8,72	3,34
MLX	89	342	8,72	2,70
ELF4	89	349	8,54	2,64
POU2F2	209	1163	8,53	1,86
MEF2A	177	941	8,39	1,95
SOX9	174	927	8,29	1,94
CEBPD	141	698	8,25	2,09
MYF6	81	315	8,22	2,66
HNF4A	72	264	8,19	2,82
ZNF263	201	1134	8,18	1,84
CUX1	117	542	8,16	2,24
CEBPE	142	725	7,94	2,03
POU3F4	145	771	7,58	1,95
RUNX1	140	743	7,46	1,95
NFYA	168	955	7,37	1,82
HOXC12	71	289	7,34	2,54
TBX2	99	476	7,16	2,15
ESR2	49	175	6,92	2,90
POU2F1	76	339	6,87	2,32
HES5	28	71	6,85	4,08
HOXD12	79	366	6,69	2,24
POU6F1	83	393	6,68	2,19
HOXC11	105	554	6,50	1,96
MEF2D	97	498	6,50	2,02
FOXB1	154	923	6,44	1,73
NEUROG2	68	308	6,38	2,29
TBX20	54	227	6,18	2,46
OLIG2	83	424	6,04	2,03

TBR1	76	376	6,04	2,09
MSC	51	219	5,86	2,41
JDP2	79	408	5,81	2,01
FOXF2	28	98	5,32	2,96
RFX4	38	157	5,27	2,51
HOXC13	57	282	5,22	2,09
RFX2	43	194	5,09	2,30
BHLHE22	60	314	4,96	1,98
ESR1	5	4	4,95	12,95
POU3F2	61	324	4,89	1,95
HLF	138	923	4,89	1,55
PBX1	24	86	4,82	2,89
RXRB	16	45	4,81	3,68
RFX3	30	122	4,76	2,55
NFE2L2	55	288	4,74	1,98
NFE2	59	318	4,71	1,92
HES7	14	37	4,70	3,92
E2F3	3	1	4,67	31,07
OLIG1	69	404	4,46	1,77
CREB3	15	46	4,35	3,38
POU3F1	64	375	4,29	1,77
PAX5	6	9	4,26	6,90
IRF1	58	335	4,18	1,79
DBP	96	645	4,02	1,54
NFIL3	64	398	3,85	1,67
MEF2B	53	313	3,85	1,75
PAX3	42	232	3,82	1,87
ATF7	32	161	3,82	2,06
GLI2	13	43	3,81	3,13
EGR3	32	162	3,79	2,05
PROX1	9	24	3,74	3,88
FOXP1	88	599	3,73	1,52
XBP1	11	38	3,37	3,00
IRF8	9	29	3,24	3,21
CENPB	8	24	3,23	3,45
CUX2	35	205	3,17	1,77
HSF4	30	171	3,07	1,82
TGIF2	6	16	3,06	3,88
ZIC1	11	44	2,93	2,59
SRF	16	76	2,91	2,18
CEBPG	33	200	2,89	1,71
PRDM1	33	212	2,58	1,61

HSF2	26	158	2,55	1,70
SMAD3	16	84	2,54	1,97
PLAG1	10	44	2,52	2,35
IRF9	4	11	2,44	3,77
ATF4	28	179	2,40	1,62
GLIS3	3	7	2,37	4,44
HNF1B	26	165	2,35	1,63
NR3C1	7	29	2,25	2,50
MAFF	8	36	2,20	2,30
MYBL2	2	4	2,12	5,18
POU3F3	22	141	2,12	1,62
HSF1	21	134	2,09	1,62
PAX7	22	143	2,06	1,59
IRF7	8	42	1,80	1,97
RREB1	5	23	1,69	2,25

Tabela Dodatkowa 4. ***Lista czynników transkrypcyjnych których motywy wiązania są nad lub niedoreprezentowane u myszy****. W tabeli przedstawiono tylko wyniki istotne statystycznie.*

Czynnik transkrypcyjny	Liczba promotorów regulowanych przez ten czynnik pośród:		Wartość statystyki Z	Krotność zmiany
	genów nakładających	genów nienakładających		
Gabpa	115	2991	5,11	1,56
Gata1	518	20344	4,21	1,03
Nr2f6	7	70	3,84	4,06
Tcf12	253	8993	2,77	1,14
Myog	280	10110	2,70	1,13
Myod1	252	9025	2,61	1,13
Zfx	99	3219	2,38	1,25
E2f3	134	4603	2,17	1,18
Erg	262	9667	2,09	1,10
Hoxb5	1	6	2,05	6,77
Tcf3	265	9921	1,81	1,09
Ebf1	73	2430	1,79	1,22
Hoxd9	66	3259	-1,71	0,82
Crem	19	1135	-1,72	0,68
Arid5a	1	202	-1,79	0,20
Twist2	51	2653	-1,88	0,78
Pitx1	102	4981	-2,11	0,83
Gfi1b	51	2738	-2,12	0,76
Foxo1	74	3781	-2,16	0,80
Nr5a2	9	748	-2,22	0,49

Hoxa9	3	407	-2,23	0,30
Arid3a	272	12090	-2,25	0,91
Hoxa5	206	9425	-2,27	0,89
Foxo3	75	4010	-2,63	0,76
Nfatc2	75	4039	-2,70	0,75
Dlx1	46	2968	-3,39	0,63
Prrx2	258	12070	-3,43	0,87
Msx3	115	6170	-3,52	0,76
Dlx3	118	6316	-3,56	0,76
Shox2	110	5973	-3,57	0,75

Tabela Dodatkowa 5. ***Lista czynników transkrypcyjnych, których ekspresja uległa statystycznie istotnej zmianie w odpowiedzi na transfekcję****.*

Czynnik transkrypcyjny	Log_2 krotności zmiany ekspresji
CARM1	6,052720793
MTA1	5,954849377
JUND	5,225556334
NFKBIB	4,824544083
GPS2	4,230687075
SETD3	3,88041796
TRIP13	3,784201277
NOC2L	3,574704019
SUPT7L	3,525479829
MXI1	3,310742828
LPXN	2,9271071
GMEB2	2,858707469
CITED2	2,833219433
ELP2	2,693988758
NAB2	2,39370613
PQBP1	2,333905468
ZNF710	2,087369002
TAF6L	1,983901462
TRIB3	1,979156309
SNW1	1,854653412
RB1	1,828957855
TRRAP	1,77978609
TAF4	1,734574917
SKI	1,675775614
EDF1	1,293987401
CALCOCO1	-1,20257121
SF1	-1,213608801

TAF9	-1,273499304
MED8	-1,296013811
MYSM1	-1,399102427
THRB	-1,423413445
MED21	-1,423903009
NFIB	-1,432929185
BIRC2	-1,43440377
DDX54	-1,451439721
JUN	-1,467877908
TAF1	-1,492838908
URI1	-1,493105175
GABPA	-1,495575223
ZNF136	-1,573978419
ZNF212	-1,584377097
NCOA3	-1,598162822
CTNNB1	-1,619242127
CNOT6	-1,620421183
SAP30	-1,620525481
PSMC3IP	-1,630050184
WWC1	-1,639527267
KDM5A	-1,667080126
NCOA7	-1,686836563
FGF2	-1,711409802
TFAP2A	-1,716852427
ARID5B	-1,725564311
TAF13	-1,787163476
ATF2	-1,787496452
EID1	-1,832583171
TAF2	-1,832712356
SMARCE1	-1,843789686
CDK7	-1,901590087
BRCA1	-1,926217895
CIR1	-1,938316409
PRPF6	-1,941085327
TCF4	-1,948225193
RBFOX2	-1,956083209
PBXIP1	-1,974557669
NFE2L1	-2,030966673
PSMC5	-2,09401249
SS18	-2,172833855
ENY2	-2,282774977
RNF20	-2,287734433

MEIS2	-2,373800976
ING4	-2,398071197
TLE4	-2,437029618
DMAP1	-2,530045016
NFKB2	-2,655757363
E2F8	-2,656939281
NR2F2	-2,659243178
TSG101	-2,969770377
CASP8AP2	-3,201642735
KAT2B	-3,440156973
MINA	-4,027720037
DDX5	-4,261686015