

RNA i transkrypcja

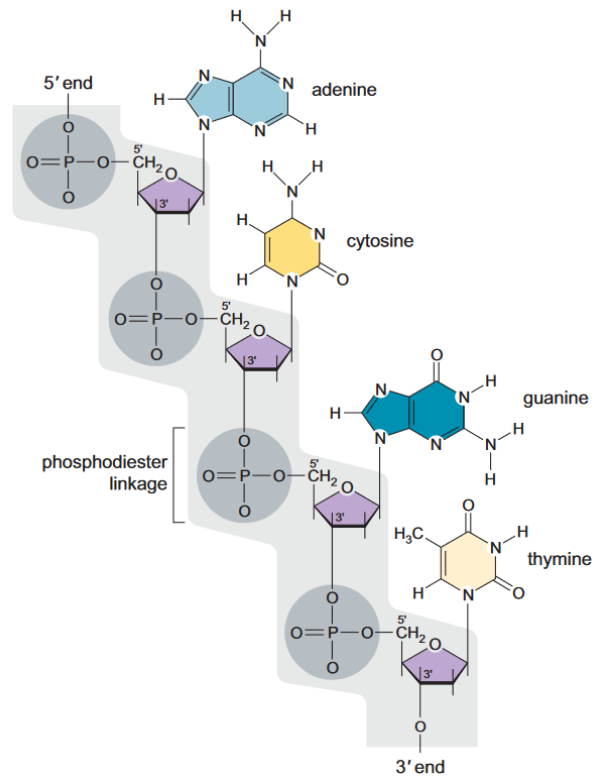
1. Przypomnienie podstawowej wiedzy
2. Metody badań

O mnie

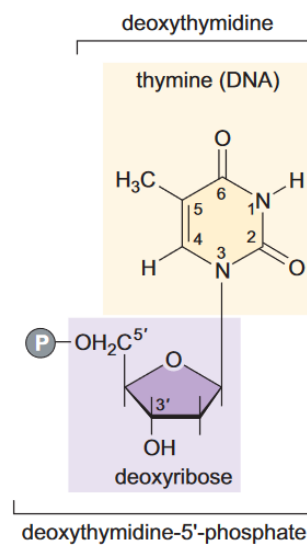
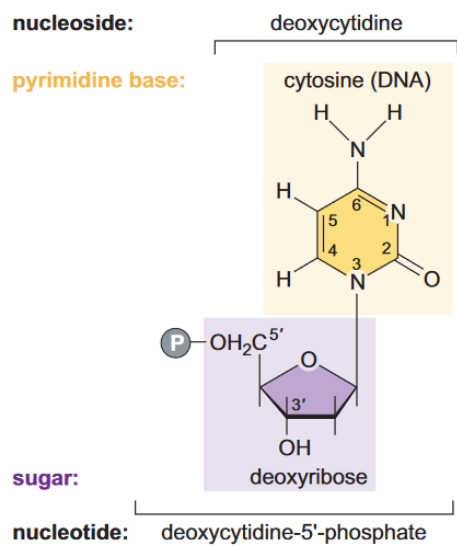
- Michał Szcześniak
- Instytut Biologii i Ewolucji Człowieka
- Pokój 2.43
- miszczyk@amu.edu.pl

DNA

- **Nukleotyd** składa się z ufosforylowanej deokserybozy połączonej z zasadą azotową wiązaniem N-glikozydowym.
- **Nukleozyd** składa się z monocukru deokserybozy, połączonej przy węglu 1' wiązaniem N-glikozydowym z zasadą azotową (guaniną, adeniną, cytozyną lub tyminą).



Nukleotydy



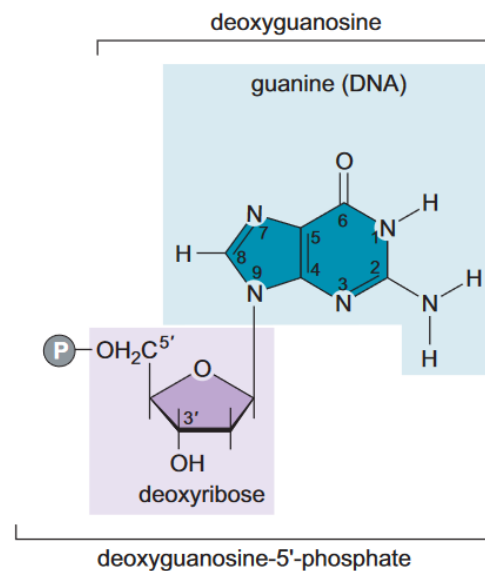
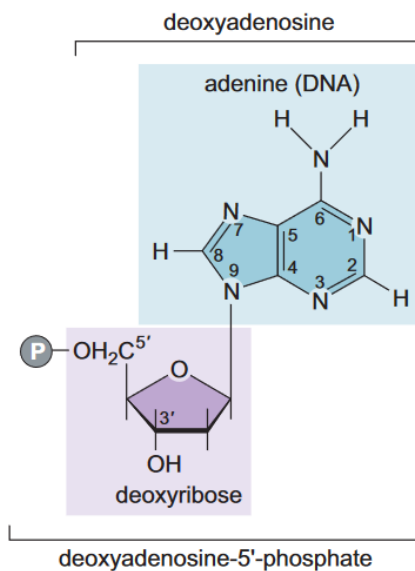
Nukleotydy

nucleoside:

purine base:

sugar:

nucleotide:



Kod IUPAC

Nucleic acid code	Common display	Meaning	Mnemonic
A	A	A	Adenine
C	C	C	Cytosine
G	G	G	Guanine
T (or U)	T	T (or U)	Thymine (or Uracil)
R	A/G	A or G	Purine
Y	C/T	C or T (or U)	Pyrimidines
K	G/T	G or T (or U)	Bases which are ketones
M	A/C	A or C	Bases with amino groups
S	C/G	C or G	Strong interaction
W	A/T	A or T (or U)	Weak interaction
B	C/G/T	Not A (i.e. C, G, T or U)	B comes after A
D	A/G/T	Not C (i.e. A, G, T or U)	D comes after C
H	A/C/T	Not G (i.e. A, C, T or U)	H comes after G
V	A/C/G	Neither T nor U (i.e. A, C or G)	V comes after U
N	A/C/G/T	A or C or G or T (or U)	Nucleic acid
. or -	-	Gap of indeterminate length	

https://www.cottongen.org/help/nomenclature/IUPAC_nt

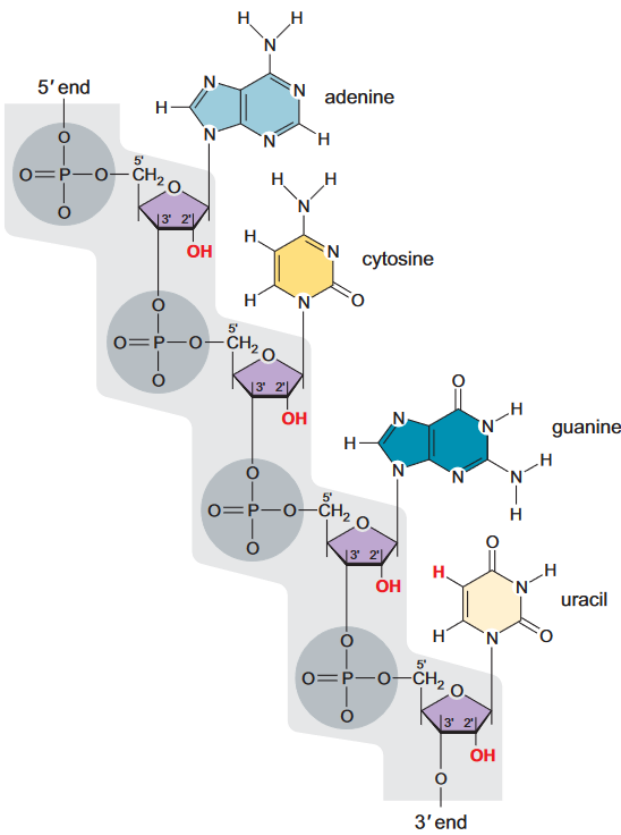
Czym jest RNA?

Czym jest transkrypt?

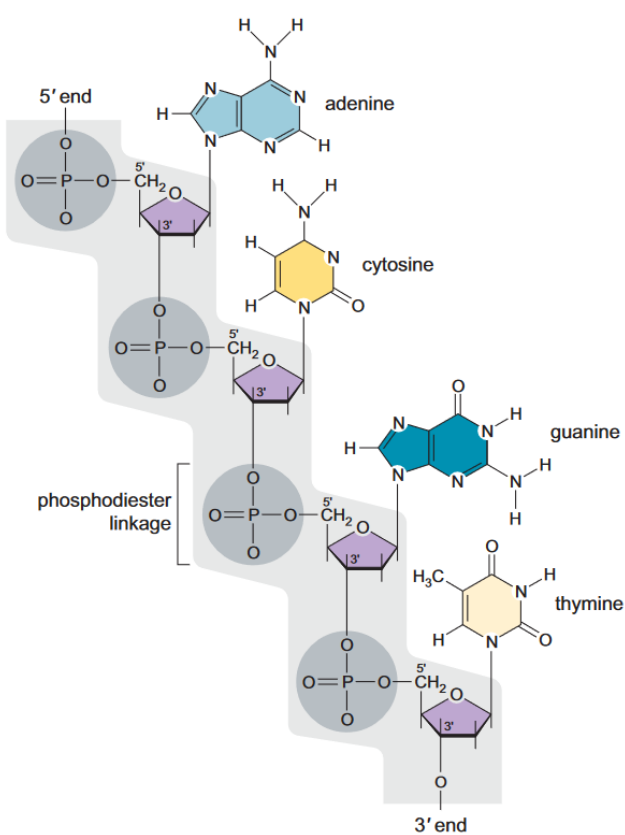
Czym jest transkryptomika?

Czym jest transkryptom?

RNA



DNA



Czym jest RNA?

Czym jest transkrypt?

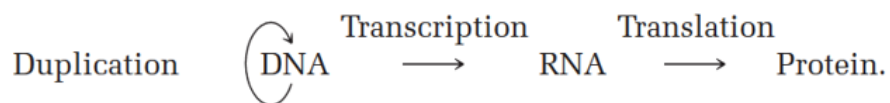
Czym jest transkryptom?

Czym jest transkryptomika?

Centralny dogmat biologii molekularnej

Zaproponowany w 1956 roku przez Francisa Cricka.

Czy nadal się sprawdza? Jak należy go zmodyfikować?



Ekspresja genów

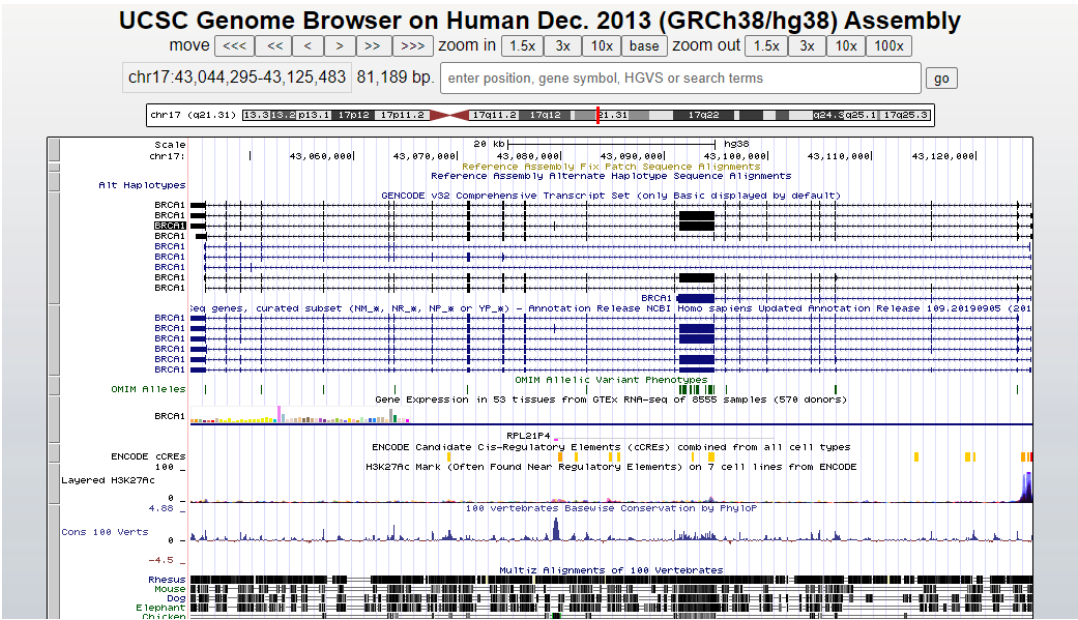
- Ekspresja genów to proces, w którym kod genetyczny genu jest wykorzystywany do utworzenia funkcjonalnego produktu, takiego jak białko czy cząsteczka niekodującego RNA.
- Składa się z dwóch głównych etapów: transkrypcja i translacja
- W przypadku genów, które nie kodują białka, translacja nie ma oczywiście miejsca. Dotyczy to m.in. długich niekodujących RNA, miRNA, tRNA, snRNA, snoRNA i rRNA.

Czym właściwie jest gen?

Dyskusja:

1. Jakie kryteria muszą zostać spełnione by zaklasyfikować „coś” jako „gen”?
2. Czy gen musi ulegać transkrypcji?
3. Czy gen musi pełnić funkcje?
4. Czy istnieje jedna definicja genu?
5. Czym są geny w bazach danych?
6. Gen a transkrypt
7. Gen a *locus*
8. Czy istnieje jedna definicja genu, obowiązująca dla wszystkich organizmów?

Czym właściwie jest gen?



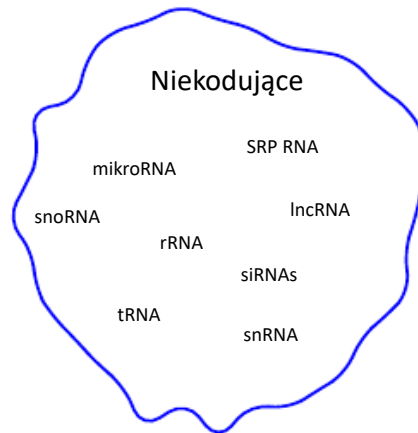
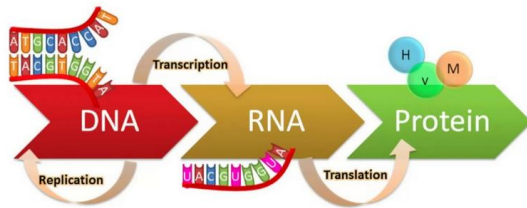
Czym właściwie jest gen?

Statystyki *gene type* z Ensembl (człowiek, wybrane pozycje):

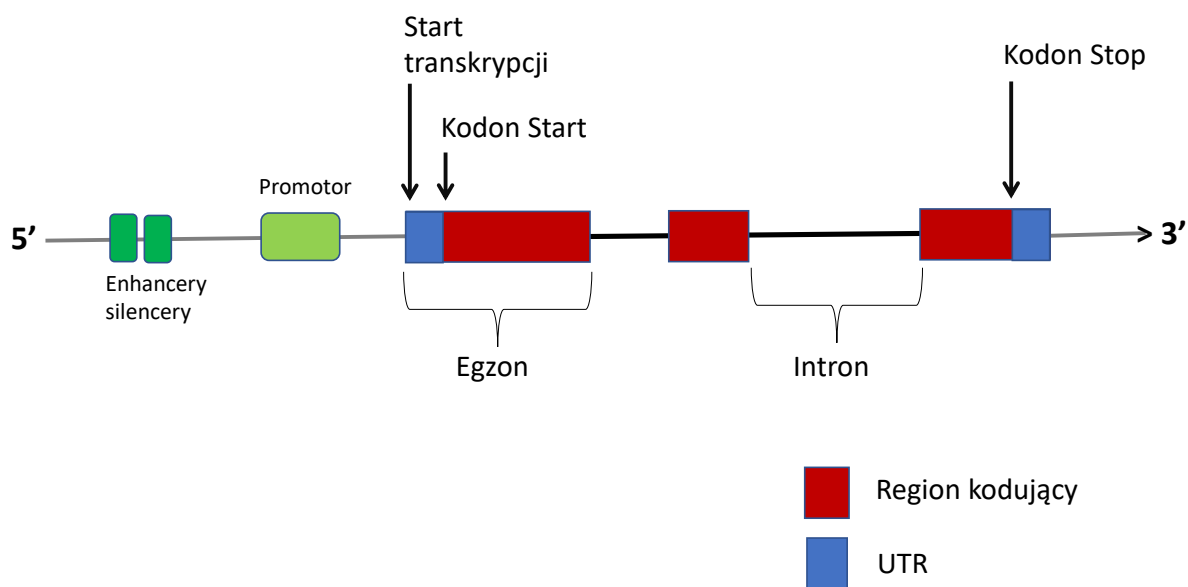
1. rRNA 71
2. rRNA_pseudogene 513
3. scaRNA 51
4. TEC 1115
5. Mt_tRNA 22
6. Mt_rRNA 2
7. protein_coding 22791
8. lncRNA 17959
9. snRNA 2072
10. miRNA 1926
11. processed_pseudogene 10824
12. snoRNA 1010

Wszystkie geny: 67130

Transkrypty kodujące białko i niekodujące



Geny kodujące białko



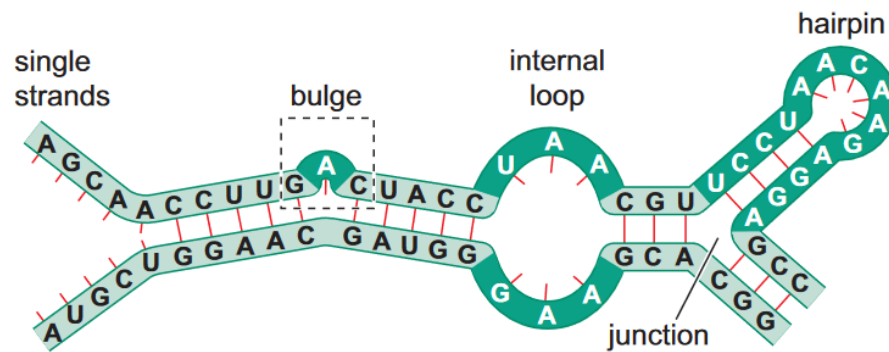
Biologia RNA.
Transkrypcja.

Funkcje RNA (wybrane pozycje)

- mRNA: matryca do syntezy białka
- tRNA: adaptor między kodonami w mRNA a aminokwasami
- rRNA: funkcje strukturalne
- miRNA, lncRNA: funkcje regulatorowe
- Modyfikacje chemiczne innych RNA, np. snoRNA
- Enzymy (rybozomy), np. snRNA

Struktura przestrzenna RNA jest ściśle związana z pełnionymi przez nie funkcjami.

Fałdowanie RNA

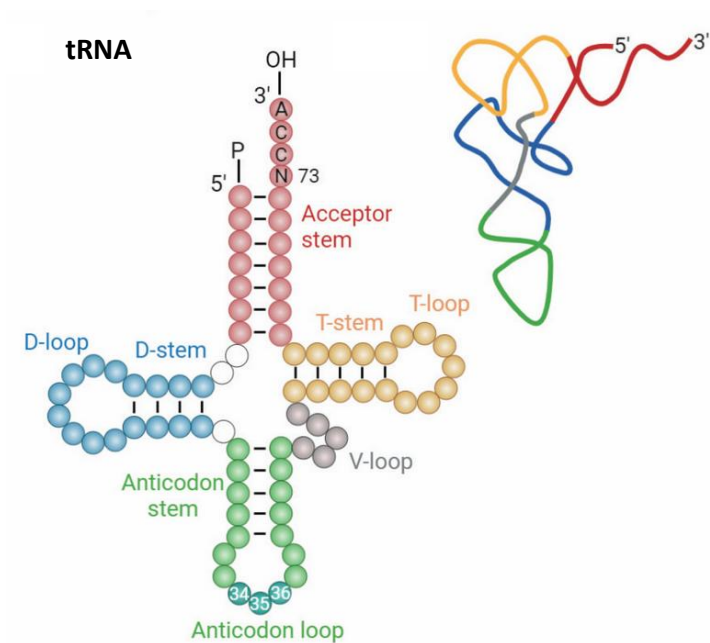


Jakie rodzaje struktur tutaj wyróżniamy?

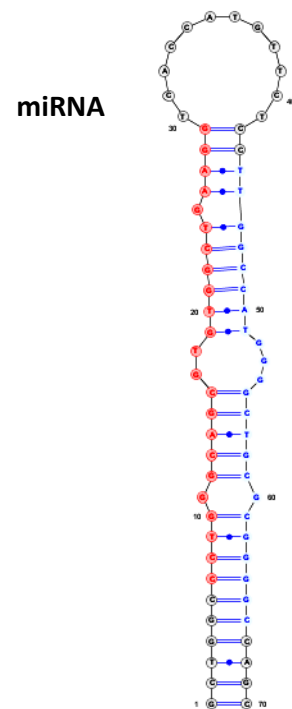
Dlaczego RNA się fałduje?

Przykłady cząsteczek RNA o charakterystycznej strukturze drugorzędowej.

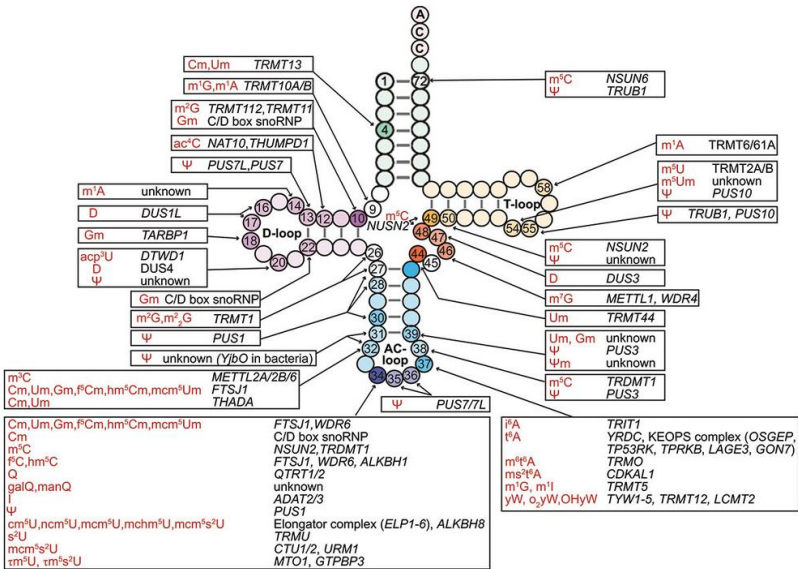
Przykładowe struktura RNA



<https://doi.org/10.3389/fmolb.2020.610617>

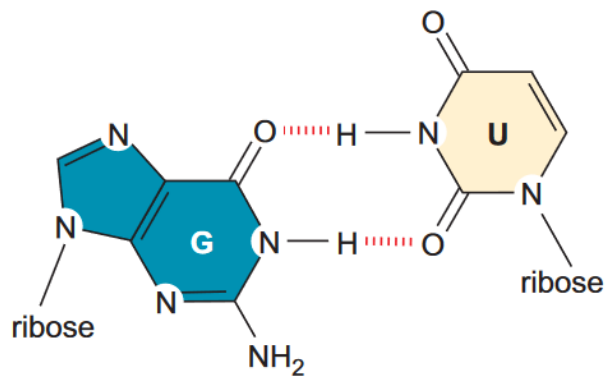


Przykładowe struktura RNA: modyfikacje nukleotydów

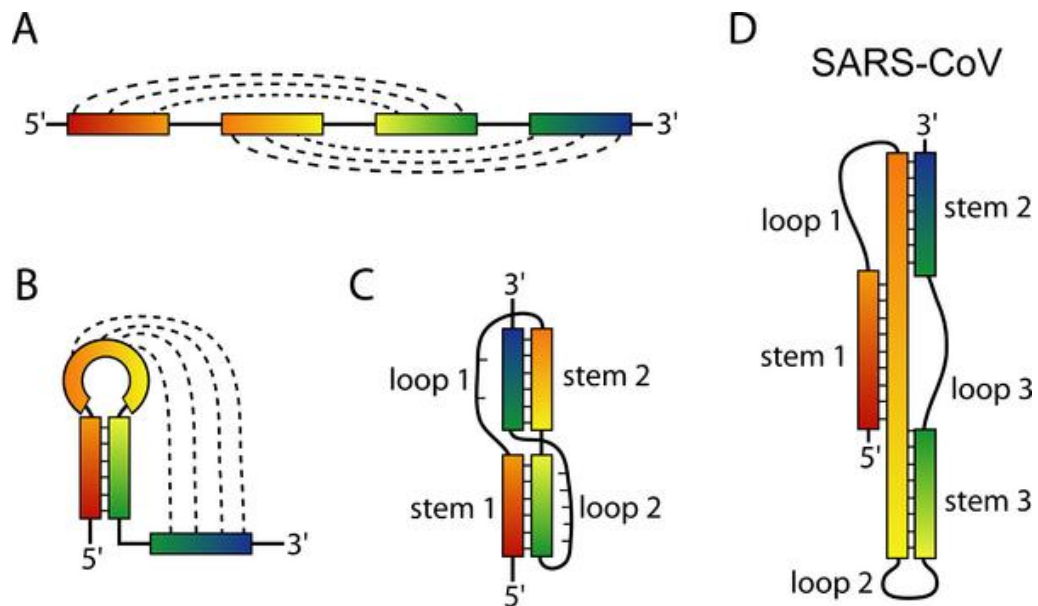


Zhang M, Lu Z. tRNA modifications: greasing the wheels of translation and beyond. *RNA Biol.* 2025 Dec;22(1):1-25. doi: 10.1080/15476286.2024.2442856. Epub 2024 Dec 26. PMID: 39723662. Figure caption: Schematic representation of human cytoplasmic tRNA modifications and their writer enzymes.

Fałdowanie RNA – G:U (wobble)

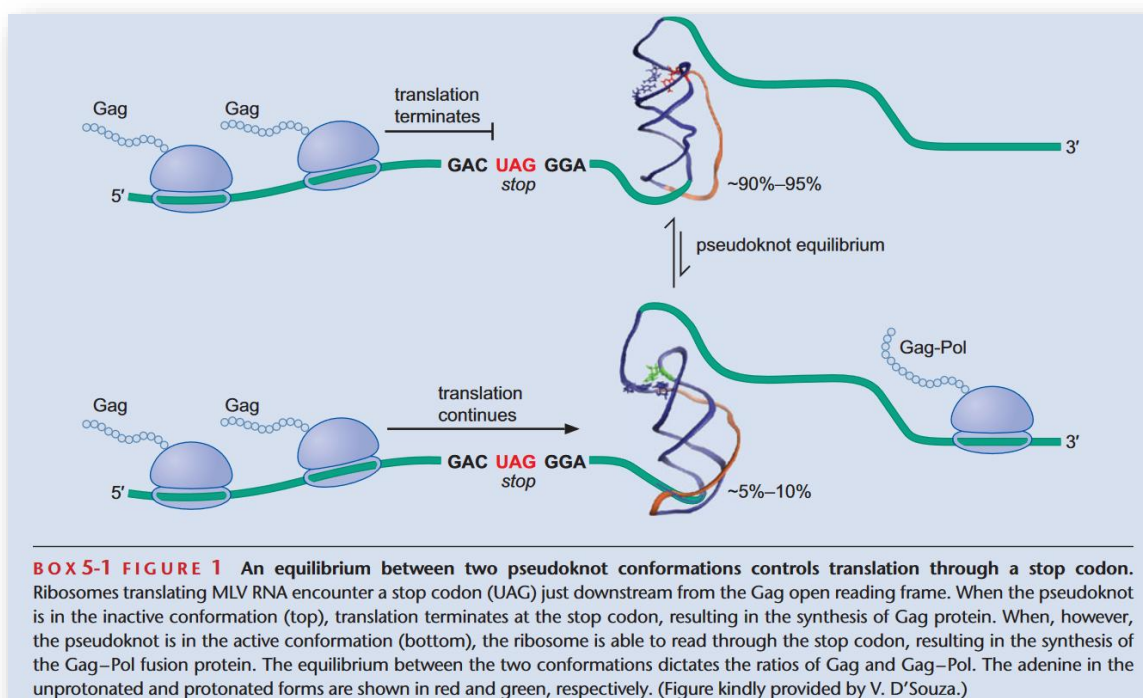


Fałdowanie RNA: pseudowęzły

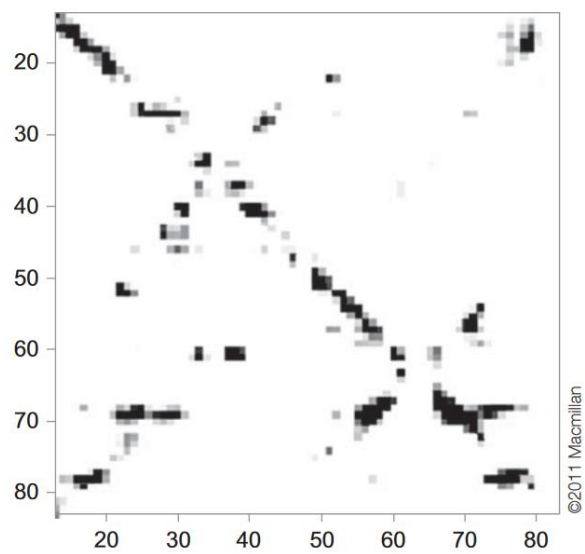


Staple DW, Butcher SE (2005) Pseudoknots: RNA Structures with Diverse Functions. PLOS Biology 3(6): e213.
<https://doi.org/10.1371/journal.pbio.0030213>
<https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0030213>

Poznanwanie struktury RNA

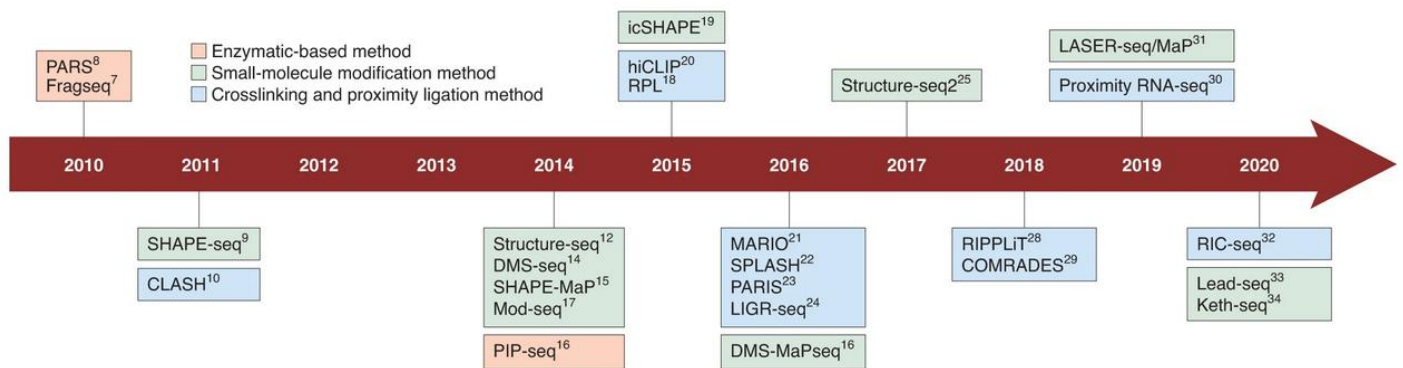


Poznawanie struktury RNA



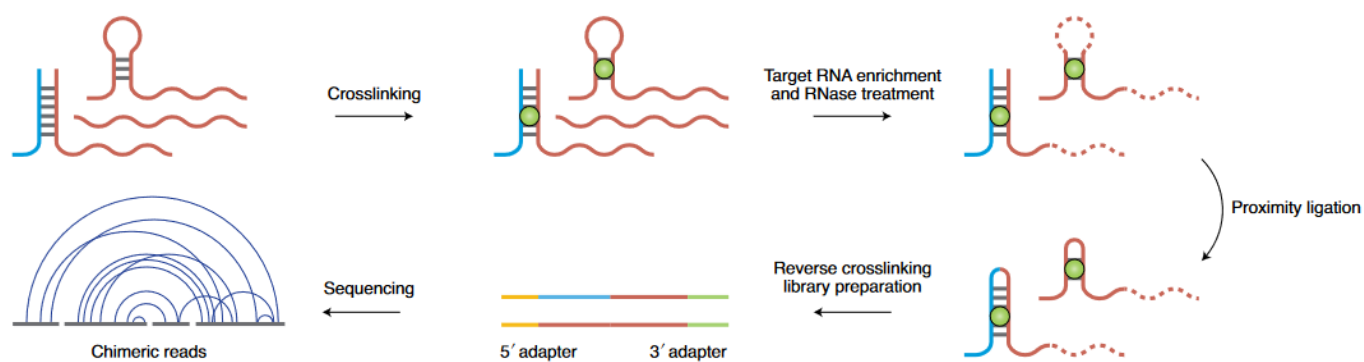
Poznanwanie struktury RNA

From: [RNA structure probing uncovers RNA structure-dependent biological functions](#)

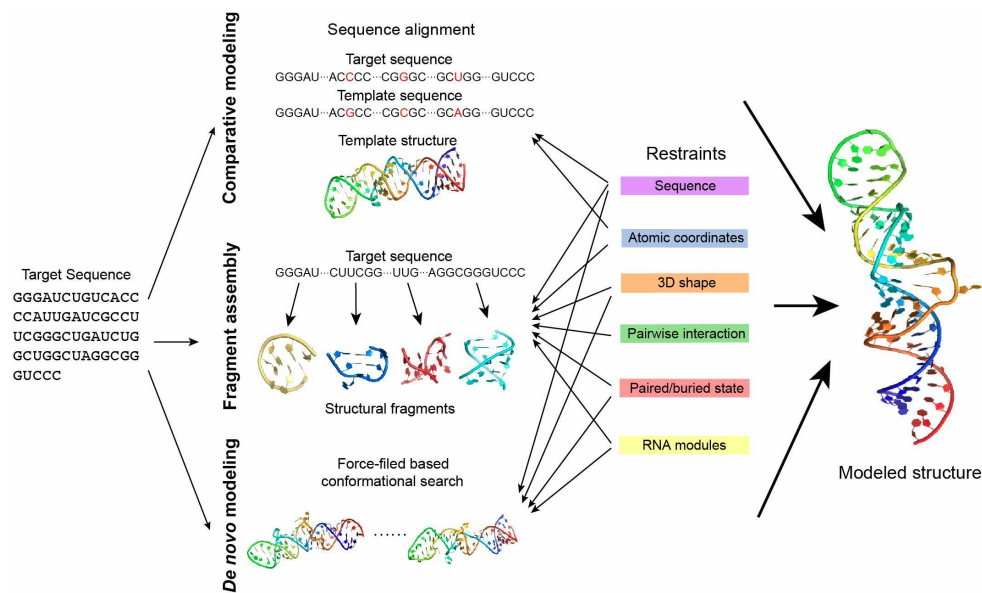


Poznanwanie struktury RNA

Crosslinking and proximity ligation-based method

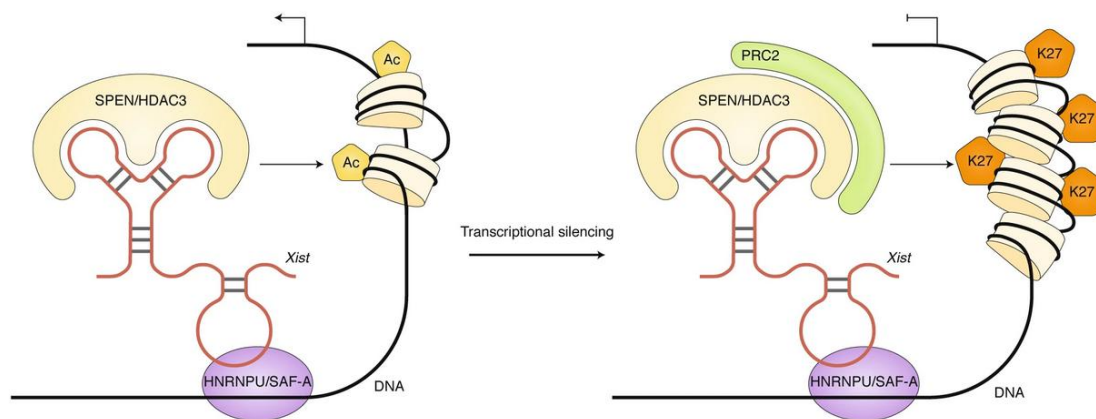


Poznanwanie struktury RNA: metody bioinformatyczne



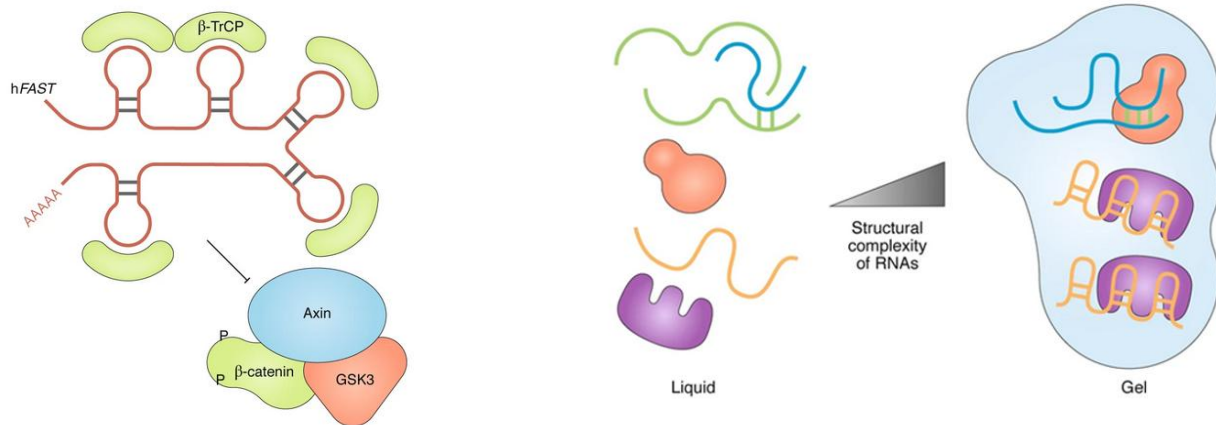
<https://doi.org/10.3389/fgene.2020.574485>

Przykładowe funkcje RNA w kontekście ich struktur przestrzennych



<https://www.nature.com/articles/s41589-021-00805-7>

Przykładowe funkcje RNA w kontekście ich struktur przestrzennych



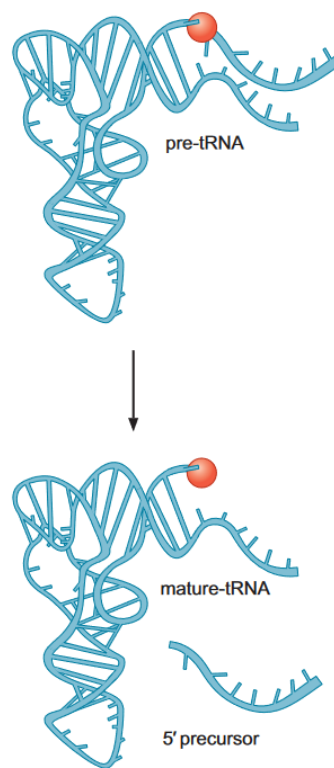
<https://www.nature.com/articles/s41589-021-00805-7>

RNA jako enzymy

- Rybozomy posiadają klasyczne właściwości enzymów, jak miejsce aktywne czy miejsce wiązania substratu
- Jednym z pierwszych poznanych rybozymów jest Rnaza P, endorybonukleaza zaangażowana w dojrzewanie cząsteczek tRNA z cząsteczek prekursorowych
- RNaza P odcina sekwencję liderową (*leader segment*) z końca 5' prekursora
- Rnaza P składa się z RNA i białka, ale tylko komponent RNA posiada właściwości katalityczne

RNA jako enzymy

RNaza P odcina fragment
RNA z końca 5'
prekursora cząsteczki
tRNA



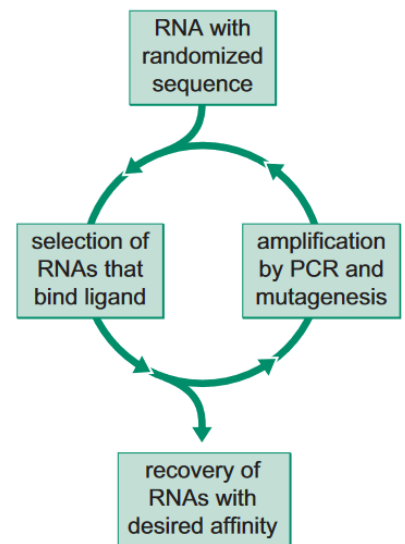
Rybozym hammerhead

- Jest rybonukleazą obecną w wiroidach (czynniki patogenne złożone z RNA, atakujące rośliny)
- Kiedy wiroid się namnaża, powstają jego liczne kopie w postaci pojedynczej nici RNA
- Cięcie tego prekursora wymaga obecności rybozymu hammerhead
- Nie jest to prawdziwy enzym, ponieważ nie sam siebie, zaś klasyczny enzym powinien katalizować reakcję dla wielu cząsteczek. Niemniej jednak da się podzielić cząsteczkę na część substratową oraz część pełniącą funkcje enzymatyczne; ta druga zachowuje się wówczas jak enzym

Inne funkcje RNA omówimy na dalszych wykładach

SELEX: otrzymywanie cząsteczek RNA o żądanych właściwościach

- SELEX: *systematic evolution of ligands by exponential enrichment*
- Z mieszaniny cząsteczek, selekcjonowane są te posiadające określone właściwości, na przykład powinowactwo do wybranych cząsteczek RNA lub białek
- Zróżnicowanie cząsteczek osiągane jest przez mutagenezę pośredniczoną przez polimerazę DNA



Podstawowe techniki molekularne badań nad RNA

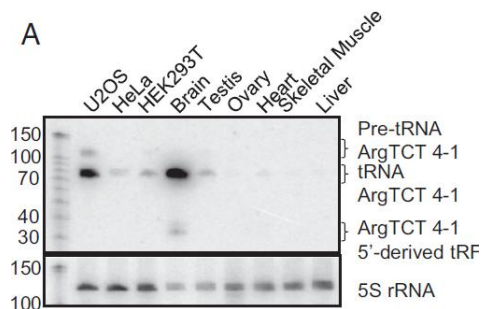
W jaki sposób wykrywamy cząsteczki RNA?

Jak określamy poziom ekspresji transkryptów?

W jaki sposób sekwencjonujemy transkryptomy?

Northern Blot

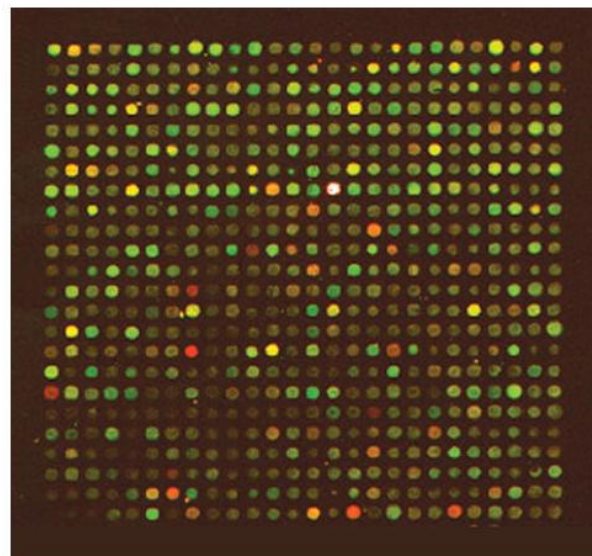
- Badane cząsteczki RNA są wykrywane poprzez hybrydyzację ze specyficzną sondą DNA, z którą tworzą dupleks RNA-DNA.
- W trakcie Northern Blot, cząsteczki RNA są rozdzielane na żelu (elektroforeza) pod względem ich rozmiaru (długości)
- Sondy są znakowane, np. radioaktywnie lub chemoluminescencyjnie
- Northern Blot pozwala określić ilość badanego RNA; może też służyć do porównania poziomów ekspresji pomiędzy badanymi próbkami



<https://doi.org/10.1073/pnas.1821120116>

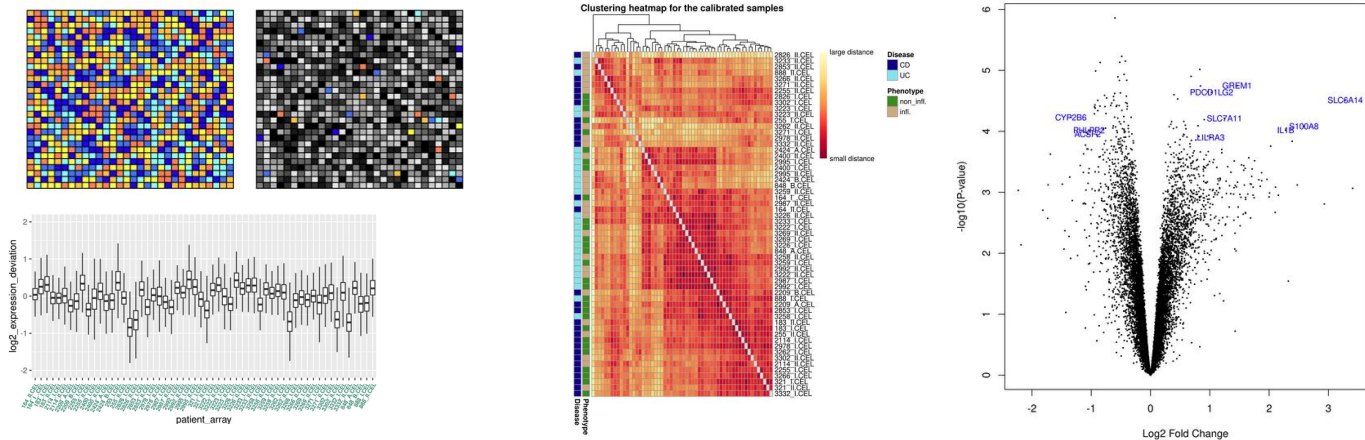
Mikromacierze

- Mikromacierze są koncepcyjnie podobne do Northern/Southern Blot, jednak przeniesione do innej skali
- Odwrócenie znaczenia terminów: tutaj sondy (*probes*) są unieruchomione (np. na szklanej płytce, stanowią znane sekwencje), natomiast badany materiał (cDNA) stanowi tzw. *target*
- Kiedy badane cząsteczki cDNA hybrydują do sond, ich ilość jest proporcjonalna do intensywności świecenia, które jest sczytywane w eksperymencie mikromacierzowym, informując eksperymentatora o poziomie ekspresji genów



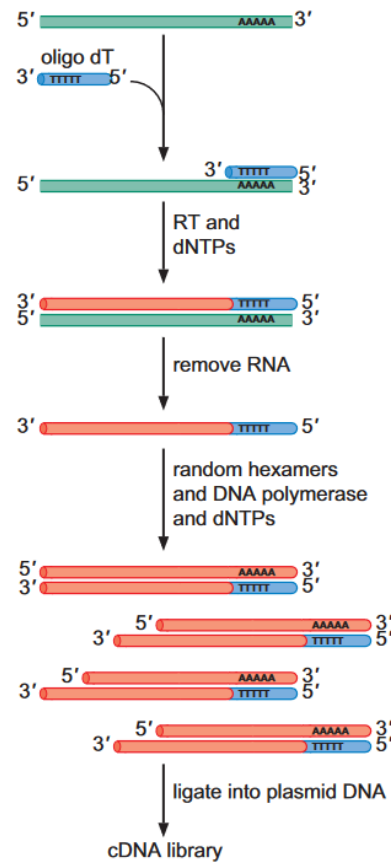
Mikromacierze

An end to end workflow for differential gene expression using Affymetrix microarrays



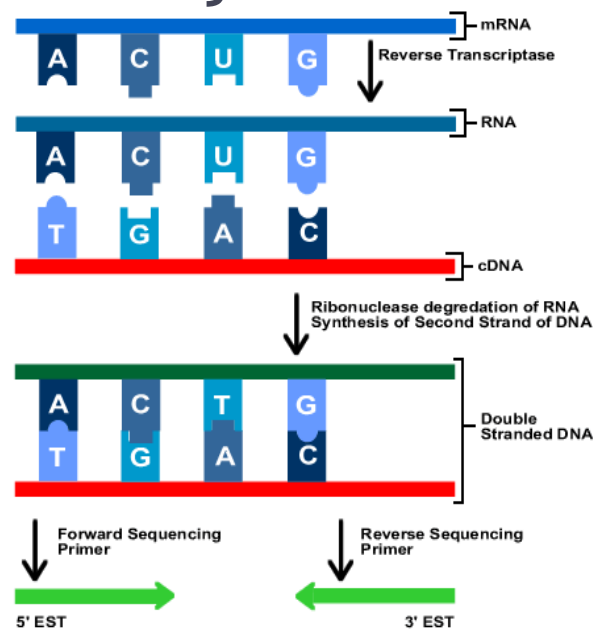
<https://bioconductor.org/packages/release/workflows/vignettes/maEndToEnd/inst/doc/MA-Workflow.html>

Biblioteki cząsteczek cDNA



EST - definicja

- ang. Expressed Sequence Tags – sekwencyjne znaczniki ekspresji
- Krótkie sekwencje nt o długości od 100 do 800 pz
- Otrzymywane przez analizę i sekwencjonowanie komplementarnego DNA (cDNA) wytwarzanego na matrycy mRNA
- 5'EST – zazwyczaj rejon kodujący białko, duży stopień konserwacji między gatunkami
- 3'EST – mogą wchodzić w zakres np. UTR, mniejsza konserwacja



Ryc. Powstawanie ESTów

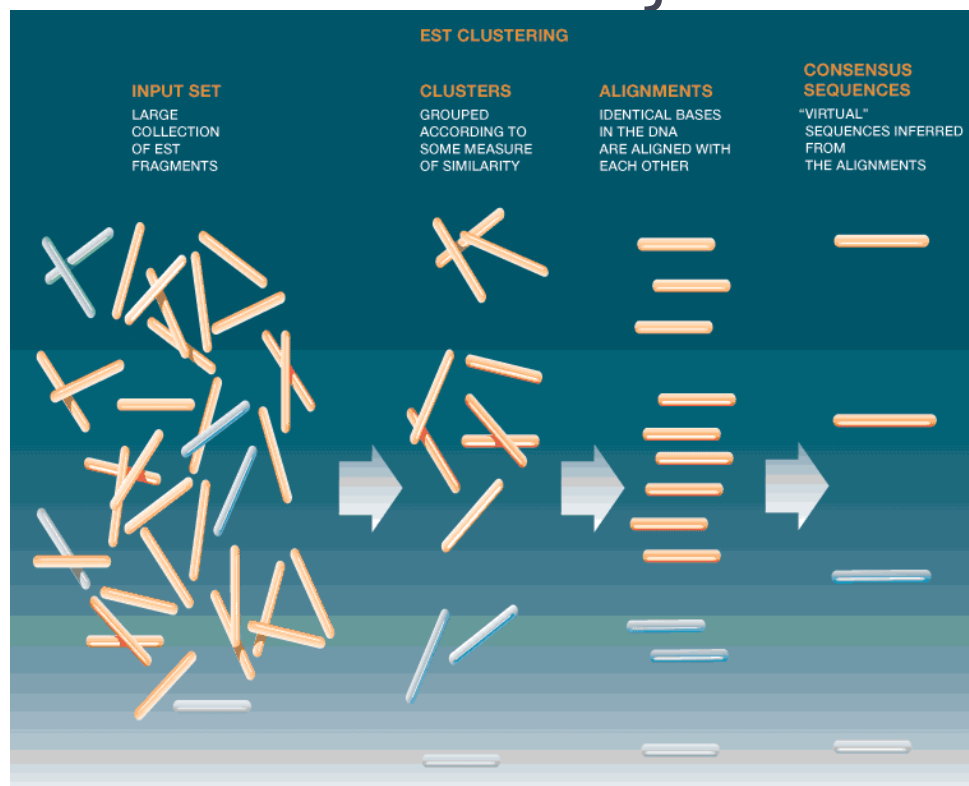
<http://www.ncbi.nlm.nih.gov/About/primer/est.html>

EST - wykorzystanie

- Tworzenie bibliotek (komórka, tkanka, organizm)
- Zestaw EST-ów pozwala zrekonstruować całe cząsteczki cDNA przy użyciu regionów nakładających się (ang. *regions of overlap*) by przewidzieć, które sekwencje powinny do siebie pasować
- Sekwencjonowanie transkryptów
- Bezpośrednia analiza ekspresji genów
- Mapowanie do genomów

Uwaga! mRNA a więc i EST-y nie zawierają intronów, rejonów międzygenowych itp.

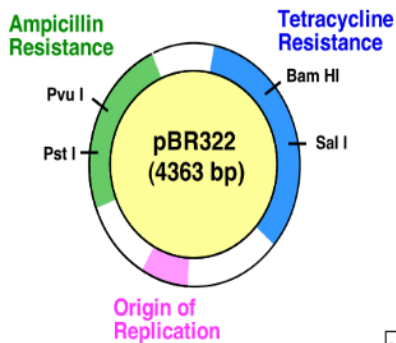
EST - składanie sekwencji



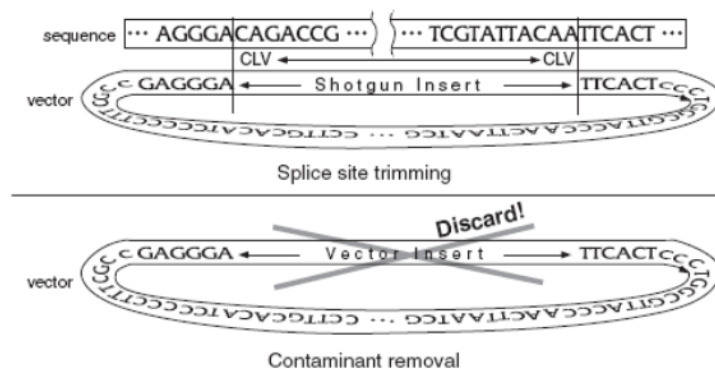
EST - etapy analizy

1. Identyfikacja i usunięcie zanieczyszczeń w sekwencjach (wektory, sekwencje o niskiej jakości)
2. Grupowanie EST-ów na zasadzie podobieństwa (tworzenie klastrów dla danego transkryptu)
3. Składanie sekwencji wewnątrz grup
4. Mapowanie do genomu – analiza transkryptów
5. Adnotacja na podstawie podobieństwa do istniejących białek

Zanieczyszczenia - wektory



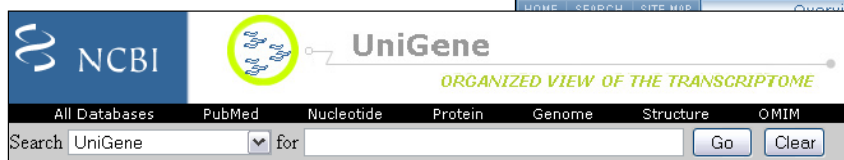
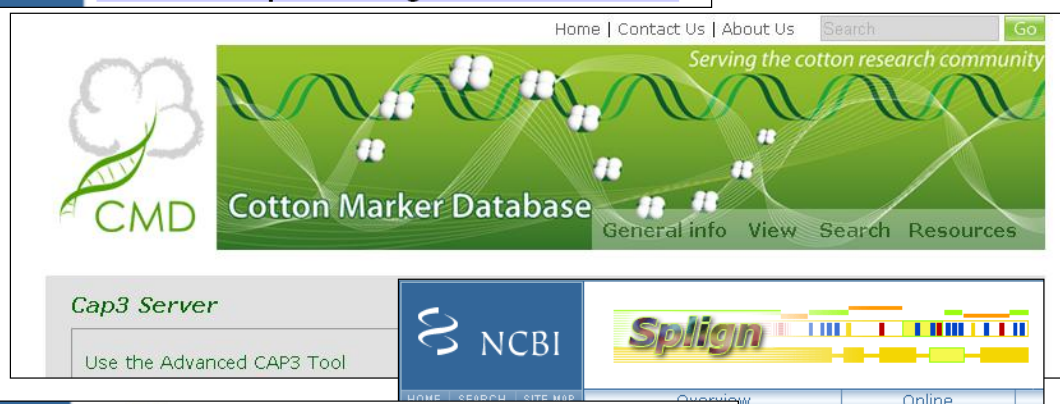
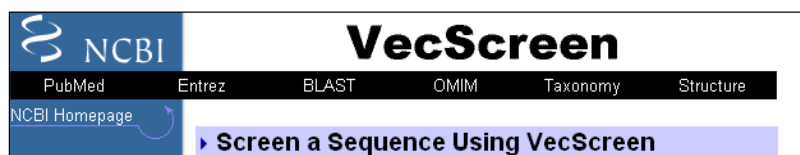
- **WEKTOR** – cząsteczka DNA zdolna do replikacji w komórce danego gospodarza i przez to nadająca się do namnażania innych fragmentów DNA
- klonowanie i namnażanie DNA – także EST
- pozostają fragmenty wektorów → błędy w składaniu



EST - ważne pojęcia

- KLASTER (ang. *cluster*) – grupa podobnych, blisko spokrewnionych sekwencji
- KONTIG (ang. *contig*) - grupa sklonowanych fragmentów DNA, które mają zachodzące na siebie regiony i mogą być złożone w układzie liniowym
- SINGLETONY (ang. *singlets*) – sekwencje, które zostały odrzucone podczas procesu składania (np. z powodu niskiej jakości sekwencji, zbyt małego podobieństwa do pozostałych sekwencji, za małego podobieństwa rejonu nakładającego się, za długiego rejonu niedopasowanego na końcach → kryteria programów!)
- KONSENSUS (ang. *consensus*) – najlepsza sekwencja ustalona w wyniku przyrównania wielu sekwencji

Narzędzia bioinformatyczne

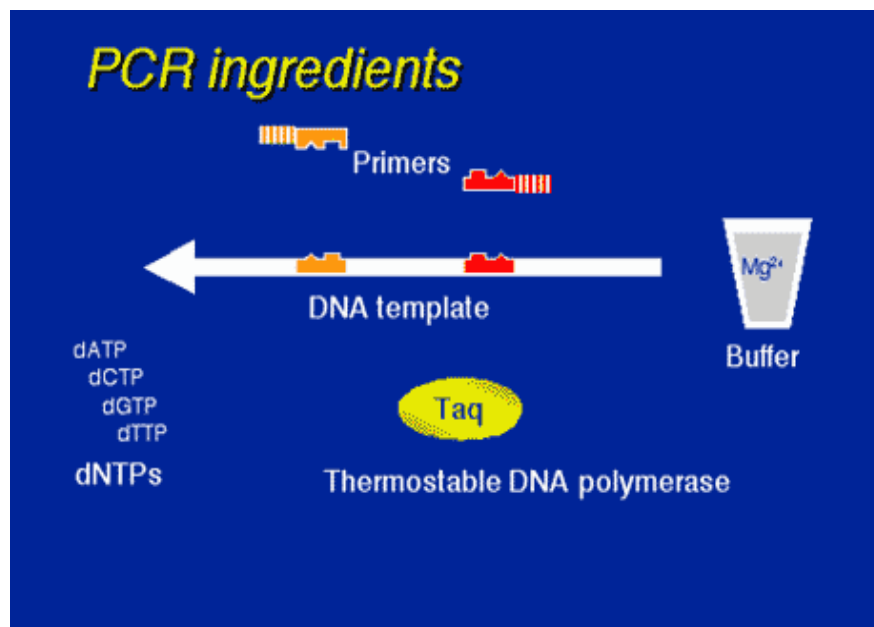


PCR –reakcja łańcuchowa polimerazy (ang. *polymerase chain reaction*)

- Technika umożliwiająca powielenie (amplifikację) wybranego fragmentu cząsteczki DNA
- Polega na wydłużaniu dwóch oligonukleotydowych **starterów** (ang. *primer*), komplementarnych do sekwencji leżących na końcach interesującego nas odcinka DNA
- Wielokrotne powtórzenia **trzech** podstawowych etapów

SKŁADNIKI REAKCJI

1. Matryca DNA
2. Startery
3. Bufor
4. Jony Mg^{2+}
5. Polimeraza
6. dNTP

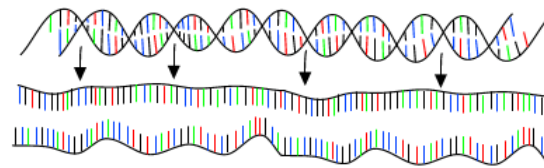


ETAPY REAKCJI

1. DENATURACJA – matryca DNA pod wpływem temperatury ok. 94° C przechodzi w formę jednoniciową
2. HYBRYDYZACJA (przyłączanie, *annealing*) – startery przyłączają się do homologicznych sekwencji DNA matrycy, rozpoczęcie reakcji polimeryzacji w temperaturze ok. 55° C
3. SYNTEZA (*extension, elongation*) – w temperaturze ok. 72° C polimeraza DNA dobudowuje do 3' końca starterów nową nić DNA, komplementarną do matrycy, korzystając z nukleotydów dodanych do reakcji

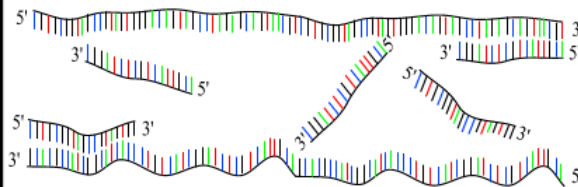
PCR : Polymerase Chain Reaction

30 - 40 cycles of 3 steps :



Step 1 : denaturation

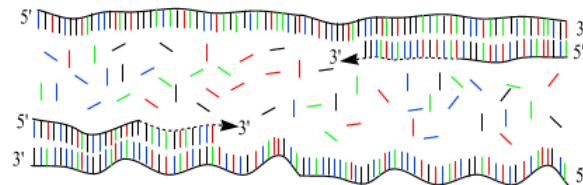
1 minut 94 °C



Step 2 : annealing

45 seconds 54 °C

forward and reverse
primers !!!



Step 3 : extension

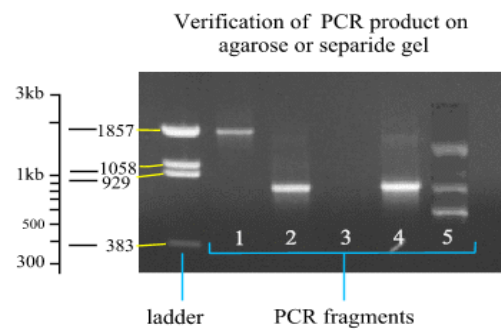
2 minutes 72 °C

only dNTP's

(Andy Vierstraete 1999)

PCR... i co dalej?

- Elektroforeza (PAA lub agarosa)
- Elucja z żelu i dalsze analizy



ZASTOSOWANIA PCR

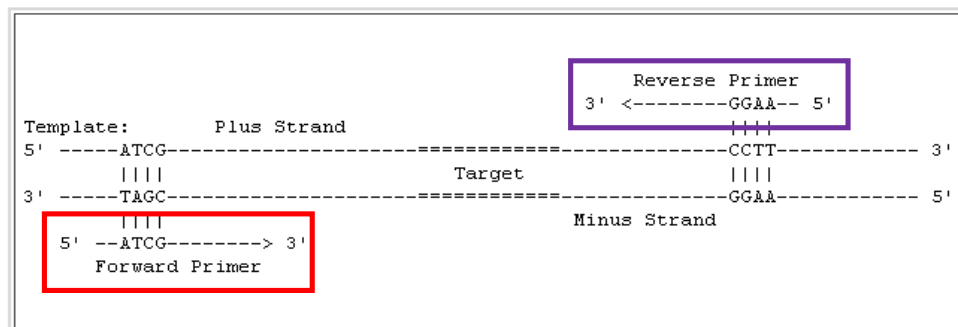
- Identyfikacja określonych sekwencji DNA, np. danego genu
- Analiza transkrypcji – RT-PCR – identyfikacja RNA w materiale genetycznym
- Analiza polimorfizmu – pula produktów charakterystycznych dla genomu
- Mutageneza – wprowadzenie mutacji do amplifikowanego fragmentu
- Identyfikacja mutacji – zmiana wielkości produktów PCR (indel), dalsze testy (mutacje punktowe)
- Mapowanie, sekwencjonowanie DNA, filogenetyka molekularna
- Medycyna sądowa, diagnostyka kliniczna

STARTERY (ang. *primers*)

- Jednoniciowe oligonukleotydy
- Długość: kilkanaście – kilkadziesiąt zasad
- Wiążą się z sekwencją komplementarną (matrycą)
- Miejsce startu syntezy nowej nici DNA

FORWARD (sense)

REVERSE (antisense)



Projektowanie starterów – narzędzia bioinformatyczne

Primer3Plus
pick primers from a DNA sequence

[Primer3Manager](#)
[About](#)

[Help](#)
[Source Code](#)

Task: Detection Select primer pairs to detect the given template sequence. Optionally targets and included/excluded regions can be specified. Pick Primers Reset Form

Main General Settings Advanced Settings Internal Oligo Penalty Weights Sequence Quality

Sequence Id:
[Paste source sequence below](#) Or upload sequence file: Przeglądaj... Upload File

Primer-BLAST
A tool for finding specific primers

NCBI/ Primer-BLAST: Finding primers specific to your PCR template (using Primer3 and BLAST). [more...](#) [Tips for finding specific primers](#)

PCR Template
Enter accession, gi, or FASTA sequence (A refseq record is preferred) Clear

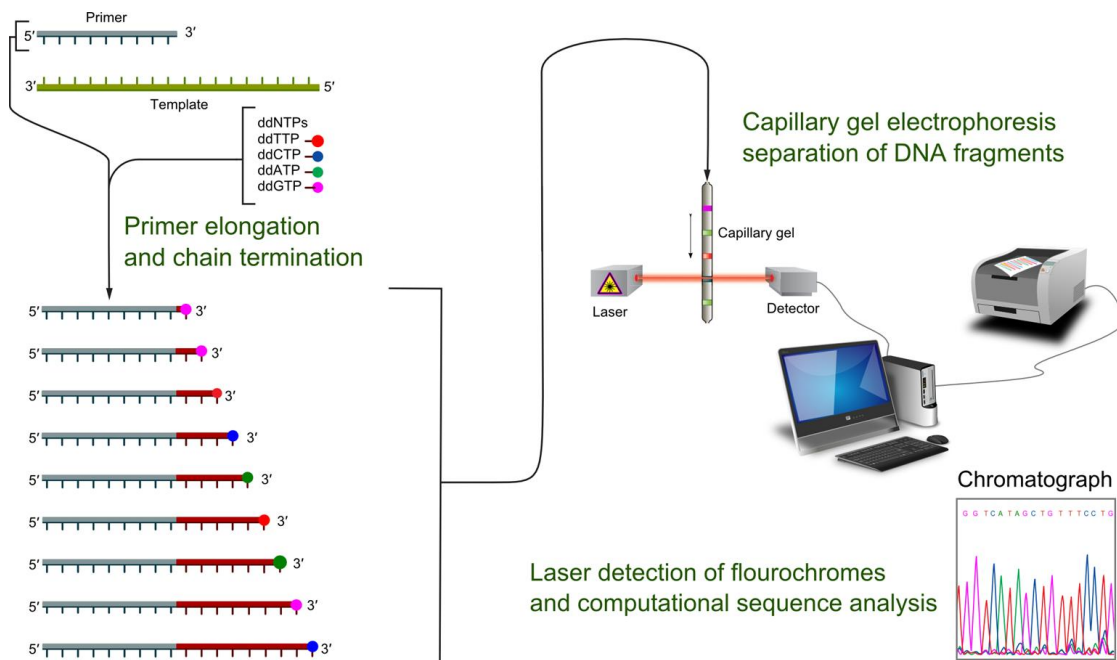
Or, upload FASTA file Przeglądaj...

Range
Forward primer From To Clear
Reverse primer From To

Sekwencjonowanie

I, II, III generacja

Sekwencjonowanie I generacji



Jeśli nie wskazano inaczej, część ilustracji pochodzi z:

Molecular Biology of the Gene, 7th edition

James D. Watson, Tania A. Baker, Stephen P. Bell, Alexander Gann,
Michael Levine, Richard Losick

Pearson 2013

RNA i transkrypcja

1. Przypomnienie podstawowej wiedzy
2. Metody badań

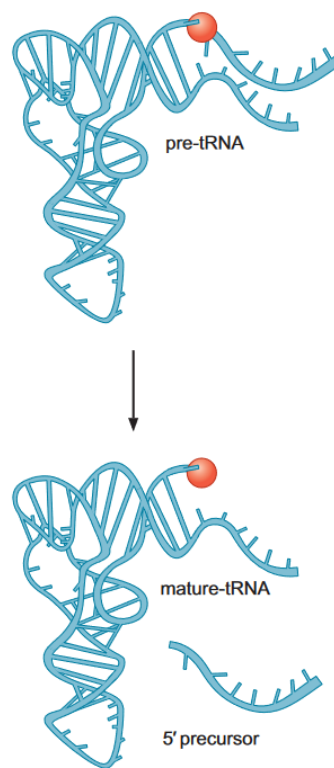
RNA jako enzymy

- Rybozymy posiadają klasyczne właściwości enzymów, jak miejsce aktywne czy miejsce wiązania substratu
- Jednym z pierwszych poznanych rybozymów była Rnaza P, endorybonukleaza zaangażowana w dojrzewanie cząsteczek tRNA z cząsteczek prekursorowych
- RNaza P odcina sekwencję liderową (*leader segment*) z końca 5' prekursora
- Rnaza P składa się z RNA i białka, ale tylko komponent RNA posiada właściwości katalityczne

RNA jako enzymy

RNaza P odcina fragment RNA z końca 5' prekursora cząsteczki tRNA

RNaza P bierze udział w dojrzewaniu końca 5' tRNA u bakterii, archeonów czy Eukariota. In vivo występuje w postaci holoenzymu (katalizę wspomaga kofaktor). U bakterii kofaktor składa się z jednej cząsteczki białka, w archeonów z czterech, zaś u organizmów eukariotycznych z dziewięciu lub dziesięciu. In vitro prezentuje swoje właściwości nawet bez udziału części białkowej. Aktywność enzymatyczną warunkuje RNA M1, rozpoznając właściwe miejsca na nici pierwotnego transkryptu.



Rybozym hammerhead

- Jest rybonukleazą obecną w wiroidach (czynniki patogenne złożone z RNA, atakujące rośliny)
- Kiedy wiroid się namnaża, powstają jego liczne kopie w postaci pojedynczej nici RNA
- Cięcie tego prekursora wymaga obecności rybozymu hammerhead
- Nie jest to prawdziwy enzym, ponieważ nie sam siebie, zaś klasyczny enzym powinien katalizować reakcję dla wielu cząsteczek. Niemniej jednak da się podzielić cząsteczkę na część substratową oraz część pełniącą funkcje enzymatyczne; ta druga zachowuje się wówczas jak enzym

Inne rybozomy - przykłady

- Introny grupy I (np. *Tetrahymena thermophila* rRNA)
- Introny grupy II (np. introny mitochondrialne u drożdży)

Różnią się one rodzajem cząsteczki, która dokonuje ataku na miejsce splicingowe 5'. W intronach grupy I jest nią kofaktor guanozynowy, zaś grupy II – grupa 2'-hydroksylowa reszty adenylanowej intronu.

- Rybosom (częściowo rybozym) – 28S rRNA (eukariota) katalizuje tworzenie wiązań peptydowych podczas translacji
- Rybozomy typu hairpin/spinka są rybozymami małymi i katalizują jeden obrót reakcji. Biorą udział m. in. w replikacji wirusa mozaiki tytoniowej.

Przykładowe zastosowanie rybozymów:

- Terapeutyczne cięcie RNA (gene silencing): rybozomy hammerhead i hairpin mogą być zaprojektowane tak, aby specyficznie przecinały RNA patogenów (np. wirusów HIV, HCV). Wykorzystuje się je w genoterapii chorób wirusowych, blokuje się w ten sposób namnażanie wirusa brodawczaka ludzkiego (*human papilloma virus* – HPV) czy czynnika prowadzącego do transformacji nowotworowej, w wyniku której dochodzi do powstania raka skóry, sromu i ust (to wszystko wciąż w fazie eksperymentalnej).
- Wykorzystywane do precyzyjnego cięcia RNA w eksperymentach *in vitro*.

Rybozomy – podsumowanie

właściwości katalityczne rybozymów wynikają ze zmian konformacyjnymi w ich strukturze trzeciorzędowej. Rybozomy mogą przeprowadzać reakcje transestryfikacji (*hairpin*), transferu nukleotydu (introny), hydrolizy (RNA RNazy P) czy syntezę wiązania peptydowego (rybosom). Są ważnymi pod względem biologicznym cząsteczkami, wciąż prowadzone są badania nad ich aktywnością, ponieważ:

- rybozomy biorą udział w ważnych reakcjach, których zbadanie jest niezbędne do poznania metabolizmu komórki
- mają one prostszą budowę niż enzymy
- selektywność ich działania może być wykorzystywana w terapii celowanej w RNA, choć do tego niezbędne jest poznanie stabilności chemicznej oligonukleotydów w środowisku komórkowym, transportu do wyznaczonych miejsc w komórce czy fałdowania do konformacji aktywnej

Inne funkcje RNA omówimy
na dalszych wykładach

Podstawowe techniki molekularne badań nad RNA

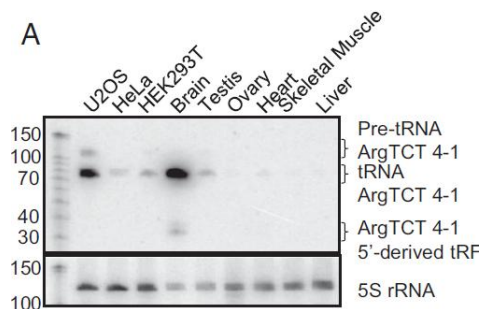
W jaki sposób wykrywamy cząsteczki RNA?

Jak określamy poziom ekspresji transkryptów?

W jaki sposób sekwencjonujemy transkryptomy?

Northern Blot

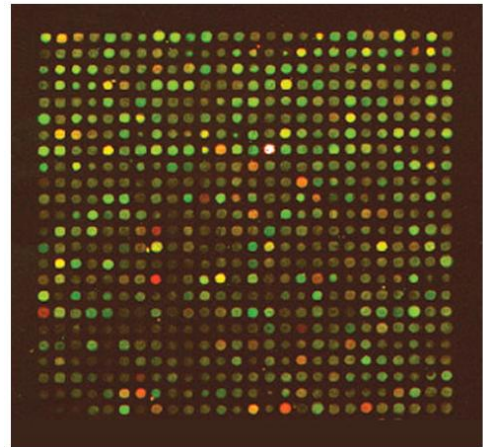
- Badane cząsteczki RNA są wykrywane poprzez hybrydyzację ze specyficzną sondą DNA, z którą tworzą dupleks RNA-DNA.
- W trakcie Northern Blot, cząsteczki RNA są rozdzielane na żelu (elektroforeza) pod względem ich rozmiaru (długości)
- Sondy są znakowane, np. radioaktywnie lub chemoluminescencyjnie
- Northern Blot pozwala określić ilość badanego RNA; może też służyć do porównania poziomów ekspresji pomiędzy badanymi próbkami



<https://doi.org/10.1073/pnas.1821120116>

Mikromacierze

- Mikromacierze są koncepcyjnie podobne do Northern/Southern Blot, jednak przeniesione do innej skali
- Odwrócenie znaczenia terminów: tutaj sondy (*probes*) są unieruchomione (np. na szklanej płytce, stanowią znane sekwencje), natomiast badany materiał (cDNA) stanowi tzw. *target*
- Kiedy badane cząsteczki cDNA hybrydują do sond, ich ilość jest proporcjonalna do intensywności świecenia, które jest sczytywane w eksperymencie mikromacierzowym, informując o poziomie ekspresji genów



Mikromacierz porównująca wzorce ekspresji genów w dwóch typach tkanek (mięśniach i neuronach) u *Caenorhabditis elegans*.

Każde kółko zawiera krótki fragment DNA pochodzący z regionu kodującego pojedynczego genu w genomie *C. elegans*. RNA zostało wyizolowane z mięśni oraz neuronów, a następnie oznaczone fluorescencyjnymi barwnikami (odpowiednio czerwonym i zielonym).

Czerwone kółka wskazują geny ulegające ekspresji w mięśniach, natomiast zielone - w neuronach. Żółte kółka oznaczają geny ulegające ekspresji w obu typach komórek.

Mikromacierze

Etapy eksperymentu mikromacierzowego

- Izolacja RNA z komórek lub tkanek.
- Synteza cDNA lub cRNA w reakcji odwrotnej transkrypcji.
- Znakowanie barwnikami fluorescencyjnymi (np. Cy3 i Cy5).
- Hybrydyzacja znakowanego materiału do sond na mikromacierzy.
- Skanowanie mikromacierzy laserem i detekcja emisji fluorescencji.
- Analiza danych – porównanie poziomów ekspresji genów między próbkami.

Mikromacierze cDNA

Mikromacierze cDNA budowane są z długich sond (500–5000 nukleotydów), które uzyskuje się poprzez amplifikację wybranych fragmentów genów metodą RT-PCR.

- Po oczyszczeniu sondy nanoszone są na płytkę metodą kontaktową (np. za pomocą igieł) lub niekontaktową (np. piezoelektryczną).
- W typowym eksperymencie porównuje się dwie próbki – na przykład tkankę zdrową i nowotworową.
- Jedna próbka jest znakowana barwnikiem zielonym (np. Cy3), druga czerwonym (np. Cy5). Po hybrydyzacji i skanowaniu:
 - zielony sygnał oznacza ekspresję genu w próbce kontrolnej,
 - czerwony – w próbce badanej,
 - żółty – ekspresję w obu próbach.

Mikromacierze oligonukleotydowe

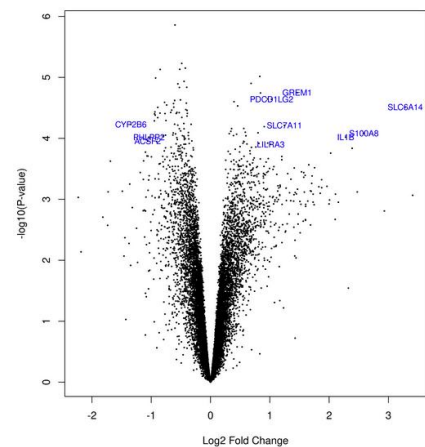
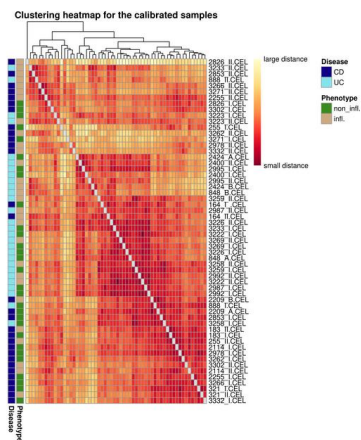
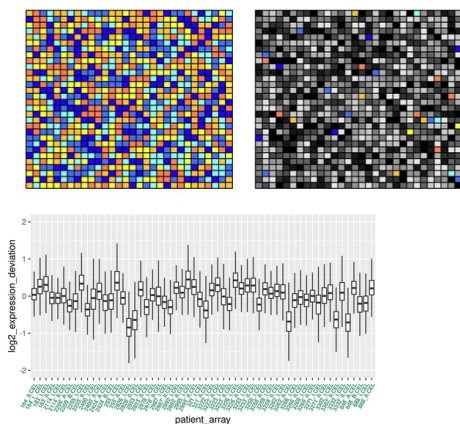
- W mikromacierzach oligonukleotydowych sondy (zwykle 25–70 nukleotydów) są syntetyzowane bezpośrednio na płytce metodą fotolitografii.
- Dzięki temu uzyskuje się bardzo dużą gęstość druku oraz wysoką powtarzalność i specyficzność sond.
- W tych mikromacierzach każda próbka jest analizowana osobno (jedna płytka = jedna próbka).
- Znakowany materiał (zazwyczaj biotynylowany cRNA) po hybrydyzacji wykrywany jest przy użyciu kompleksu streptawidyna–fikoerytryna.

Mikromacierze

Cecha	Mikromacierze cDNA	Mikromacierze oligonukleotydowe
Rodzaj sond	Długie fragmenty cDNA (500–5000 nt)	Krótkie oligonukleotydy (25–70 nt) syntetyzowane in situ
Sposób wytwarzania sond	Amplifikacja RT-PCR i nanoszenie mechaniczne	Synteza fotolitograficzna bezpośrednio na płytce
Liczba analizowanych prób	Dwie jednocześnie (barwniki Cy3 i Cy5)	Jedna próbka na płytkę
Rodzaj znakowanego materiału	cDNA znakowane fluoroforami	cRNA znakowany biotyną, wykrywany przez streptawidynę–fikoerytrynę
Sposób odczytu	Porównawczy (stosunek sygnałów dwóch barwników)	Bezwzględny (intensywność fluorescencji jednej próbki)
Specyficzność hybrydyzacji	Niższa – ryzyko nieswoistych wiązań	Wysoka – krótkie sondy i dokładna lokalizacja
Powtarzalność i standaryzacja	Ograniczona	Bardzo wysoka
Koszt i złożoność	Niska cena, prosta konstrukcja	Wysoka cena, skomplikowana technologia
Zastosowanie	Analiza ekspresji genów w układach porównawczych	Analiza ekspresji, genotypowanie, badanie SNP

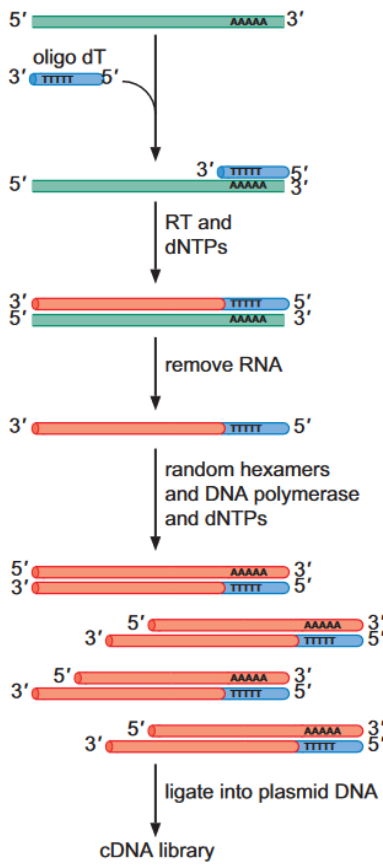
Mikromacierze

An end to end workflow for differential gene expression using Affymetrix microarrays



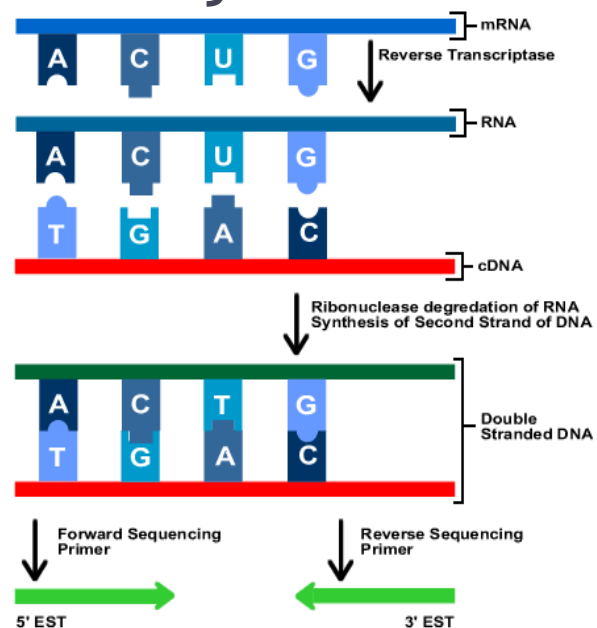
<https://bioconductor.org/packages/release/workflows/vignettes/maEndToEnd/inst/doc/MA-Workflow.html>

Biblioteki cząsteczek cDNA



EST - definicja

- ang. Expressed Sequence Tags – sekwencyjne znaczniki ekspresji
- Krótkie sekwencje nt o długości od 100 do 800 pz
- Otrzymywane przez analizę i sekwencjonowanie komplementarnego DNA (cDNA) wytwarzanego na matrycy mRNA
- 5'EST – zazwyczaj rejon kodujący białko, duży stopień konserwacji między gatunkami
- 3'EST – mogą wchodzić w zakres np. UTR, mniejsza konserwacja



Ryc. Powstawanie ESTów

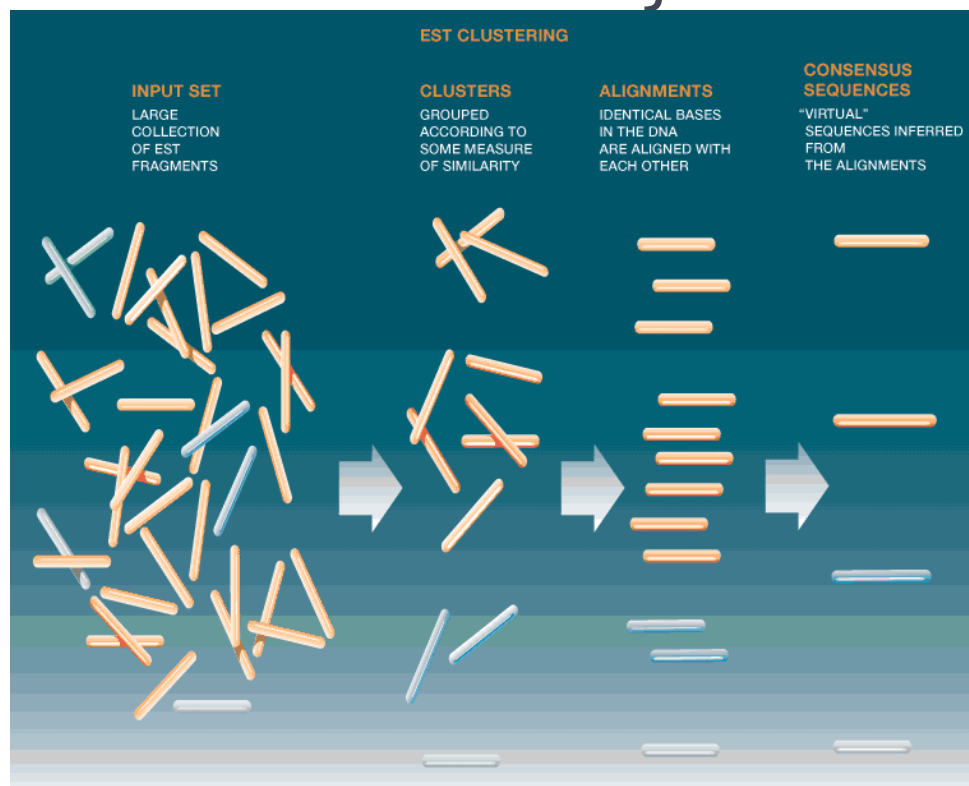
<http://www.ncbi.nlm.nih.gov/About/primer/est.html>

EST - wykorzystanie

- Tworzenie bibliotek (komórka, tkanka, organizm)
- Zestaw EST-ów pozwala zrekonstruować całe cząsteczki cDNA przy użyciu regionów nakładających się (ang. *regions of overlap*) by przewidzieć, które sekwencje powinny do siebie pasować
- Sekwencjonowanie transkryptów
- Bezpośrednia analiza ekspresji genów
- Mapowanie do genomów

Uwaga! mRNA, a więc i EST-y, nie zawierają intronów, rejonów międzygenowych itp.

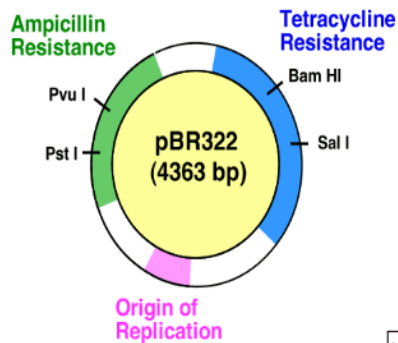
EST - składanie sekwencji



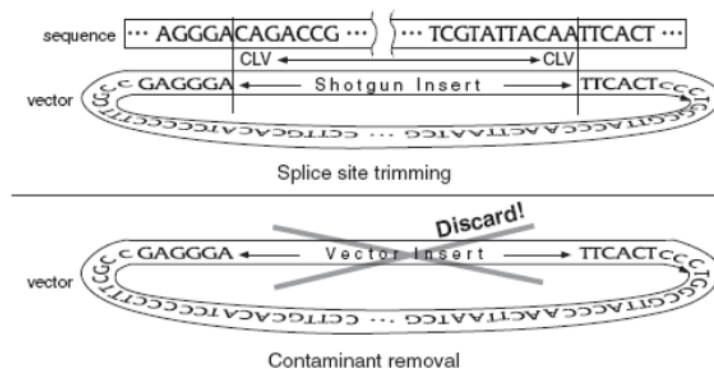
EST - etapy analizy

1. Identyfikacja i usunięcie zanieczyszczeń w sekwencjach (wektory, sekwencje o niskiej jakości)
2. Grupowanie EST-ów na zasadzie podobieństwa (tworzenie klastrów dla danego transkryptu)
3. Składanie sekwencji wewnątrz grup
4. Mapowanie do genomu – analiza transkryptów
5. Adnotacja na podstawie podobieństwa do istniejących białek

Zanieczyszczenia - wektory



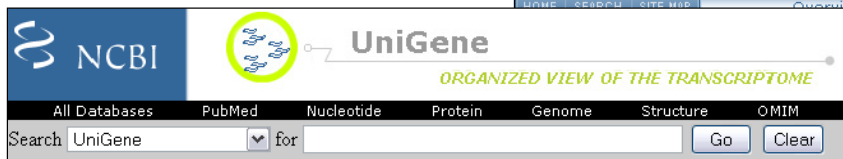
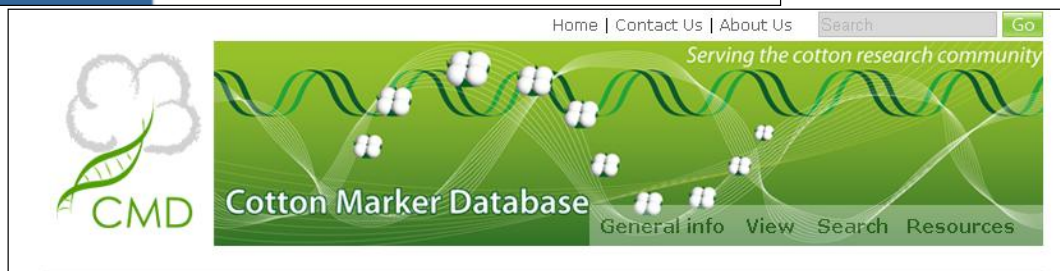
- **WEKTOR** – cząsteczka DNA zdolna do replikacji w komórce danego gospodarza i przez to nadająca się do namnażania innych fragmentów DNA
- klonowanie i namnażanie DNA – także EST
- pozostają fragmenty wektorów → błędy w składaniu



EST - ważne pojęcia

- KLASTER (ang. *cluster*) – grupa podobnych, blisko spokrewnionych sekwencji
- KONTIG (ang. *contig*) - grupa sklonowanych fragmentów DNA, które mają zachodzące na siebie regiony i mogą być złożone w układzie liniowym
- SINGLETONY (ang. *singlets*) – sekwencje, które zostały odrzucone podczas procesu składania (np. z powodu niskiej jakości sekwencji, zbyt małego podobieństwa do pozostałych sekwencji, za małego podobieństwa rejonu nakładającego się, za długiego rejonu niedopasowanego na końcach → kryteria programów!)
- KONSENSUS (ang. *consensus*) – najlepsza sekwencja ustalona w wyniku przyrównania wielu sekwencji

Narzędzia bioinformatyczne



PCR –reakcja łańcuchowa polimerazy (ang. *polymerase chain reaction*)

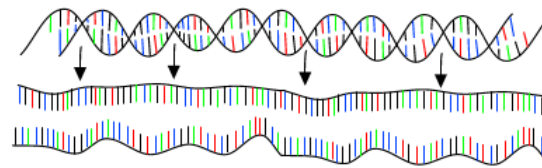
- Technika umożliwiająca powielenie (amplifikację) wybranego fragmentu cząsteczki DNA
- Polega na wydłużaniu dwóch oligonukleotydowych **starterów** (ang. *primer*), komplementarnych do sekwencji leżących na końcach interesującego nas odcinka DNA
- Wielokrotne powtórzenia **trzech** podstawowych etapów

ETAPY REAKCJI

1. DENATURACJA – matryca DNA pod wpływem temperatury ok. 94° C przechodzi w formę jednoniciową
2. HYBRYDYZACJA (przyłączanie, *annealing*) – startery przyłączają się do homologicznych sekwencji DNA matrycy, rozpoczęcie reakcji polimeryzacji w temperaturze ok. 55° C
3. SYNTEZA (*extension, elongation*) – w temperaturze ok. 72° C polimeraza DNA dobudowuje do 3' końca starterów nową nić DNA, komplementarną do matrycy, korzystając z nukleotydów dodanych do reakcji

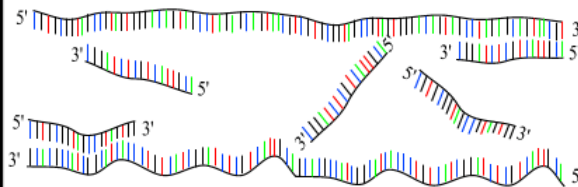
PCR : Polymerase Chain Reaction

30 - 40 cycles of 3 steps :



Step 1 : denaturation

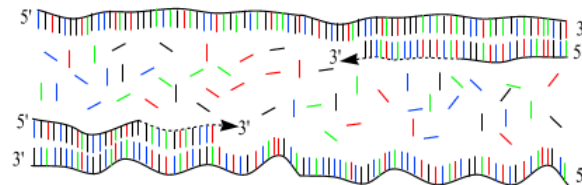
1 minut 94 °C



Step 2 : annealing

45 seconds 54 °C

forward and reverse
primers !!!



Step 3 : extension

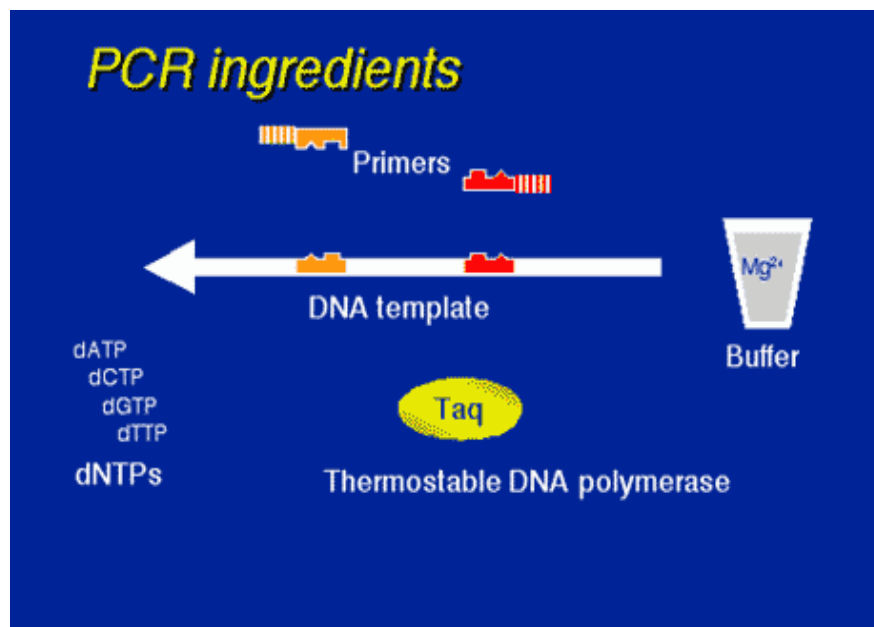
2 minutes 72 °C

only dNTP's

(Andy Vierstraete 1999)

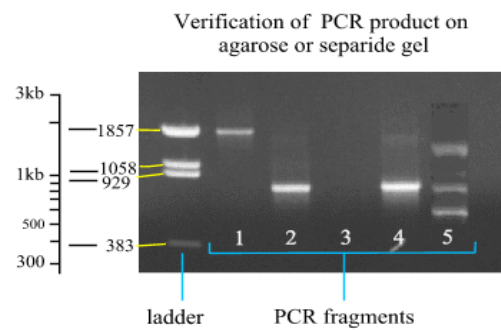
SKŁADNIKI REAKCJI

1. Matryca DNA
2. Startery
3. Bufor
4. Jony Mg^{2+}
5. Polimeraza
6. dNTP



PCR... i co dalej?

- Elektroforeza (PAA lub agarosa)
- Elucja z żelu i dalsze analizy



ZASTOSOWANIA PCR

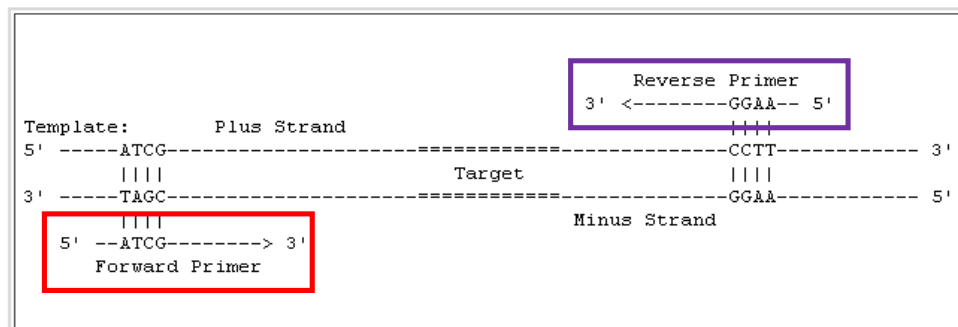
- Identyfikacja określonych sekwencji DNA, np. danego genu
- Analiza transkrypcji – RT-PCR – identyfikacja RNA w materiale genetycznym
- Analiza polimorfizmu – pula produktów charakterystycznych dla genomu
- Mutageneza – wprowadzenie mutacji do amplifikowanego fragmentu
- Identyfikacja mutacji – zmiana wielkości produktów PCR (indel), dalsze testy (mutacje punktowe)
- Mapowanie, sekwencjonowanie DNA, filogenetyka molekularna
- Medycyna sądowa, diagnostyka kliniczna

STARTERY (ang. *primers*)

- Jednoniciowe oligonukleotydy
- Długość: kilkanaście – kilkadziesiąt zasad
- Wiążą się z sekwencją komplementarną (matrycą)
- Miejsce startu syntezy nowej nici DNA

FORWARD (sense)

REVERSE (antisense)



Projektowanie starterów – narzędzia bioinformatyczne

Primer3Plus
pick primers from a DNA sequence

[Primer3Manager](#)
[About](#)

[Help](#)
[Source Code](#)

Task: Detection Select primer pairs to detect the given template sequence. Optionally targets and included/excluded regions can be specified. Pick Primers Reset Form

Main General Settings Advanced Settings Internal Oligo Penalty Weights Sequence Quality

Sequence Id:
[Paste source sequence below](#) Or upload sequence file: Przeglądaj... Upload File

Primer-BLAST
A tool for finding specific primers

NCBI/ Primer-BLAST: Finding primers specific to your PCR template (using Primer3 and BLAST). [more...](#) [Tips for finding specific primers](#)

PCR Template
Enter accession, gi, or FASTA sequence (A refseq record is preferred) Clear

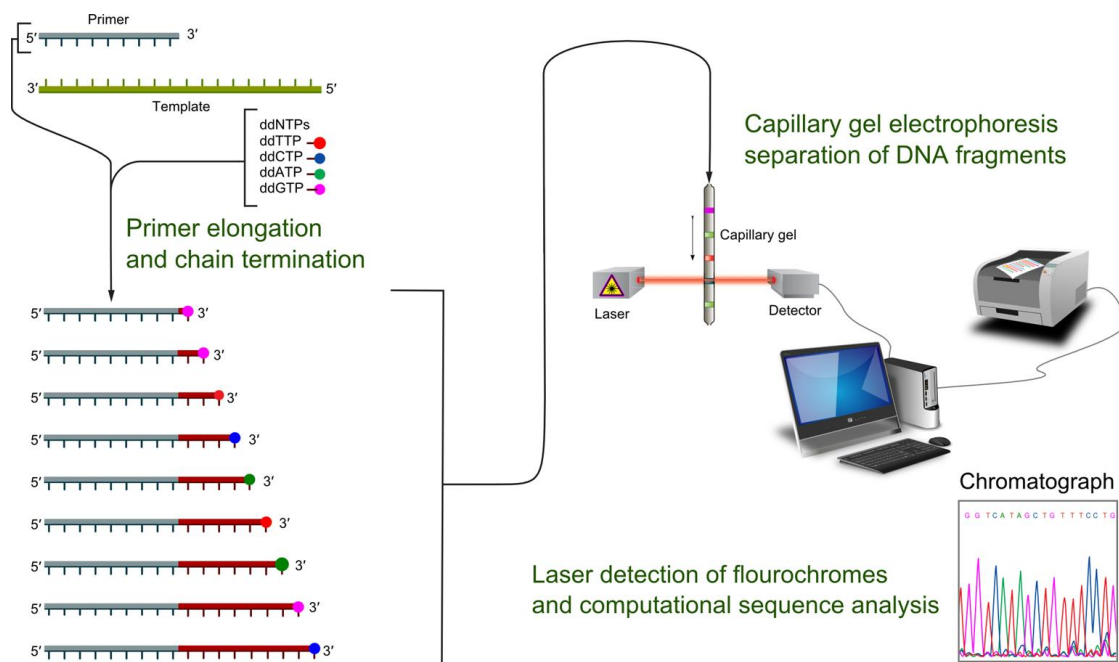
Or, upload FASTA file Przeglądaj...

Range
Forward primer From To Clear
Reverse primer From To

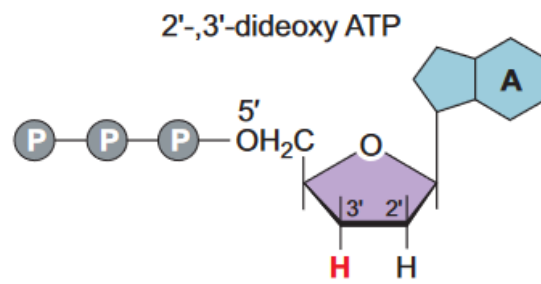
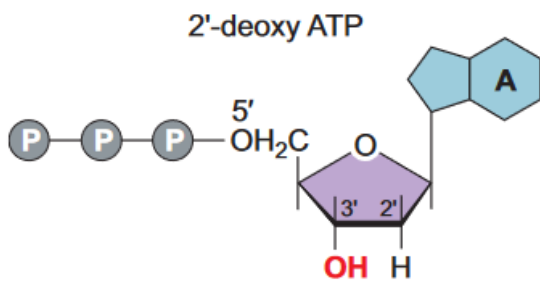
Sekwencjonowanie

I, II, III generacja

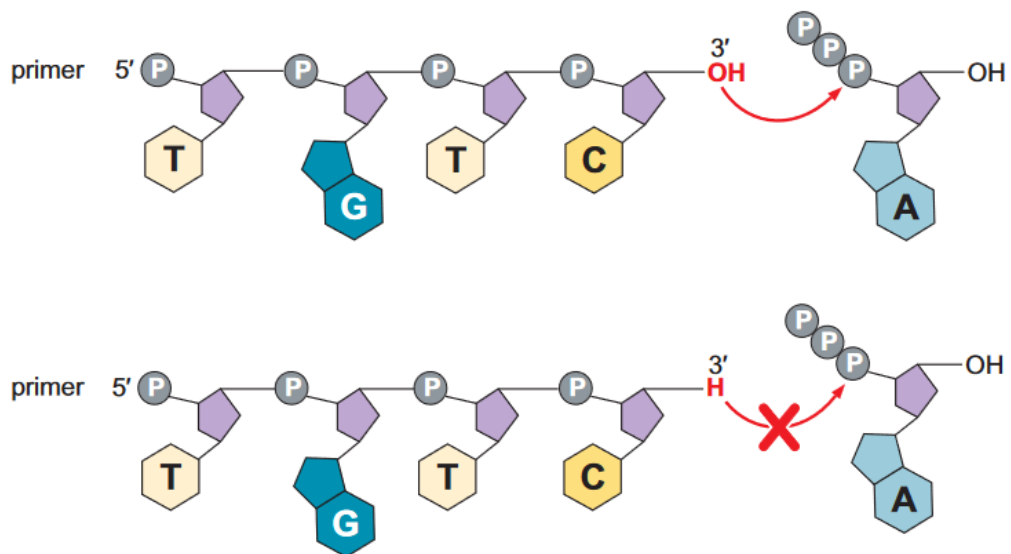
Sekwencjonowanie I generacji



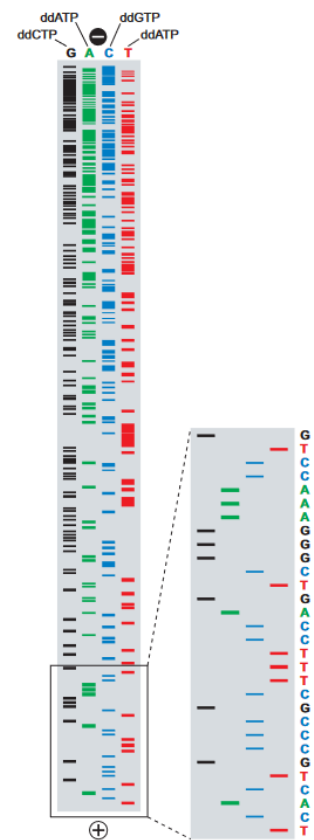
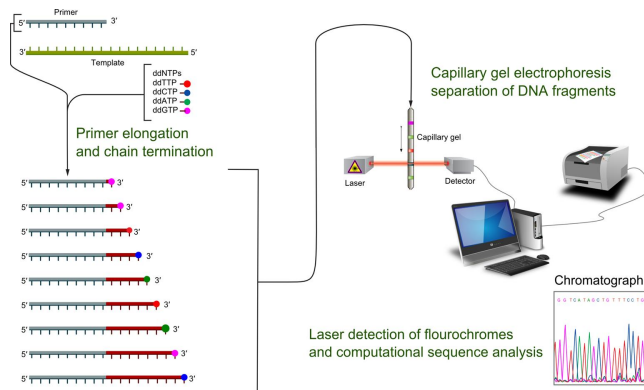
Sekwencjonowanie I generacji



Sekwencjonowanie I generacji



Sekwencjonowanie I generacji



Sekwencjonowanie typu shotgun

- Jako pierwszy genom żywego organizmu zsekwencjonowano bakterię *Haemophilus influenzae* (niewielki genom, ok. 1.8 Mb)
- Genom został pofragmentowany na niewielkie fragmenty, ok. 1 kb
- Te fragmenty wklonowano do plazmidów, tworząc bibliotekę DNA, a następnie zsekwencjonowano metodą Sangera
- Aby mieć pewność, że zsekwencjonowany zostanie cały genom, sekwencjonowaniu poddano ok. 30-40 tys. klonów. Ponieważ zsekwencjonowane fragmenty miały przeciętną długość 600 nt, dało to średnie pokrycie genomu 10x
- Alternatywne podejście, które rozważano, to pofragmentowanie genomu enzymami restrykcyjnymi, a następnie zsekwencjonowanie tak uzyskanych fragmentów DNA. Jednak to podejście się nie sprawdziło, ponieważ uzyskane fragmenty często są o wiele większe niż długość produktu reakcji sekwencjonowania

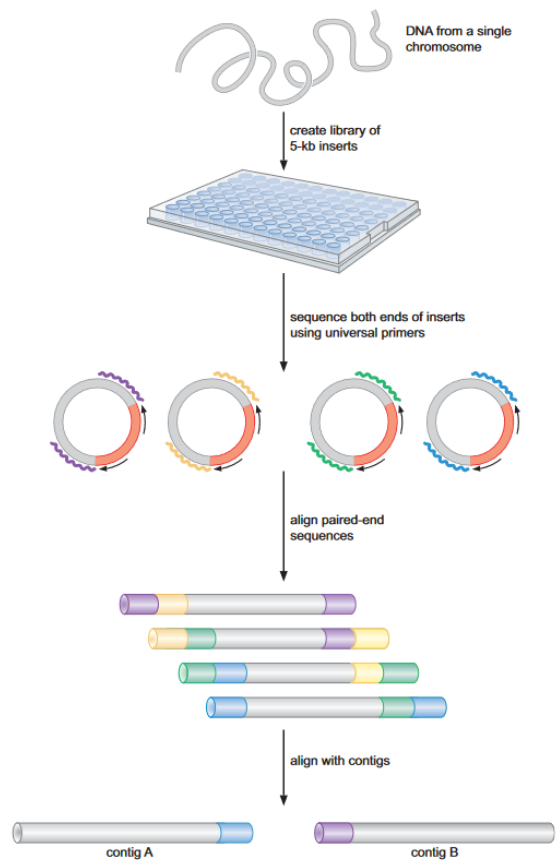
Sekwencjonowanie typu shotgun

- Przeciętny chromosom człowieka ma długość 150 Mb. A zatem wynik sekwencjonowania o długości 600 nt stanowi zaledwie 0.0004% jego długości. W konsekwencji, poznawanie dużych genomów eukariotycznych niesie ze sobą konieczność sekwencjonowania gigantycznej liczby fragmentów.
- Zazwyczaj fragmenty genomu dzieli się na osobne pule, w zależności od ich długości, np. 1, 5 czy 100 kb
- W przypadku człowieka potrzebne jest ok. 2 mln fragmentów, aby zsekwencjonować typowy chromosom z pokryciem 10x
- Krótkie fragmenty DNA są łączone w kontigi, zwykle o długości 50 000 – 200 000 nt. To już dane przydatne biologicznie, bo na przykład u *Drosophila* jeden gen przypada średnio na 10 kb, a więc takie fragmenty mieściłyby już kilka genów. Jednak w przypadku genomów mniej upakowanych w geny, jak genom człowieka, to wciąż zbyt niska wartość (jeden gen na 100 000 nt)
- Z tego powodu kontigi są łączone w dłuższe fragmenty, zwane skafoldami, o długości około 1-2 Mb

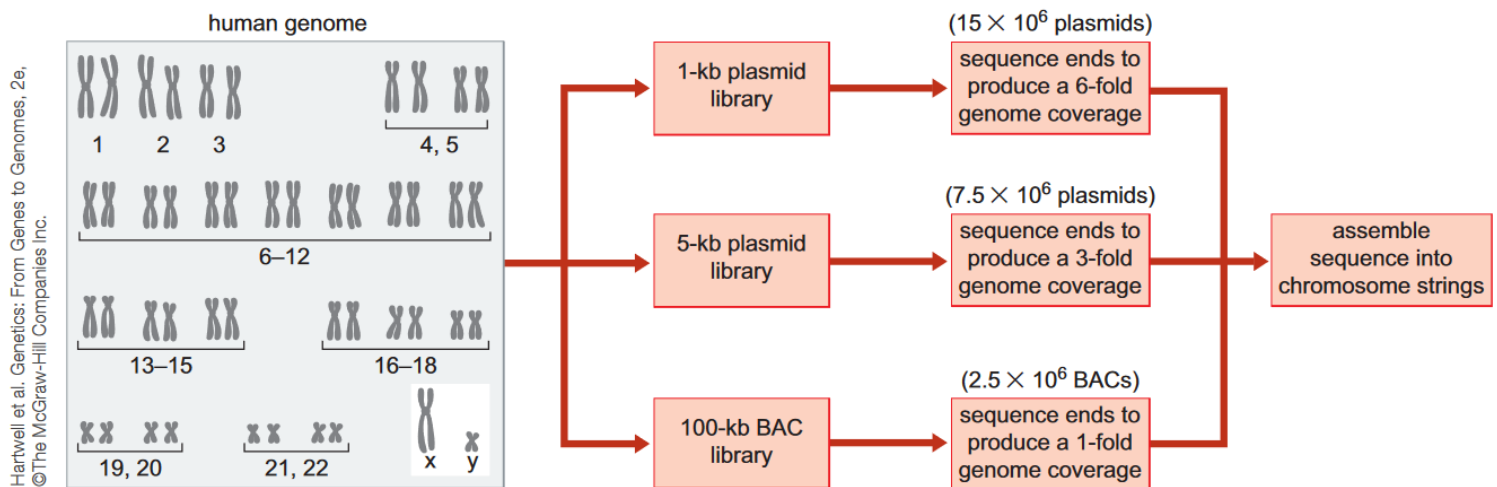
Sekwencjonowanie genomów: sparowane fragmenty

- Głównym ograniczeniem w uzyskiwaniu wystarczająco długich kontigów, skafoldów czy całych chromosomów jest występowanie elementów powtarzalnych: fragmenty, które zawierają nakładające się sekwencje zawierające elementy powtarzalne mogą zostać ze sobą połączone, mimo że nie występują obok siebie w genomie
- Częściowym rozwiązaniem tego problemu jest sekwencjonowanie fragmentów z obu końców, co prowadzi do uzyskiwania sparowanych fragmentów (tzw. *paired-end sequencing*). W tym celu typowo sekwencjonuje się fragmenty o różnej długości, np. 1 kb i 100 kb.
- Ponieważ większość elementów powtarzalnych ma długość poniżej 2-3 kb, sekwencjonowanie fragmentów o długości 5 kb okazuje się bardzo efektywnym podejściem w przypadku genomu człowieka
- Jeśli chcemy połączyć kontigi w skafoldy, to musimy zsekwencjonować końce jeszcze dłuższych fragmentów, o długości około 100 kb; są one wklonowywane np. do tzw. BAC (*bacterial artificial chromosome*). Zsekwencjonowane końce takich fragmentów są następnie mapowane do kontigów, żeby się dowiedzieć, które kontigi można ze sobą połączyć w skafoldy

Sekwencjonowanie genomów: sparowane fragmenty



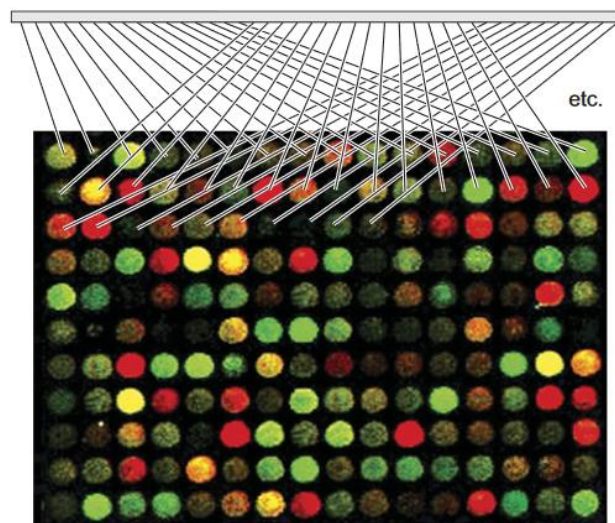
Sekwencjonowanie typu shotgun



Mikromacierze pokrywające genom (*tiling arrays*) jako metoda poznawania transkryptomu

- *Tiling arrays* zawierają immobilizowane genomowe sDNA, typowo o długości około 50 nt
- 1 mln takich 50-merów wystarcza by pokryć cały genom *Drosophila* co 100-150 nt
- sDNA w tak skonstruowanych chipach hybrydyzuje następnie ze znakowanymi fluorescencyjnie cząsteczkami RNA lub cDNA. Ten materiał może pochodzić z konkretnej linii komórkowej, tkanki czy organu
- Sczytawszy sygnał wyzwalany w trakcie hybrydyzacji dowiadujemy się które regiony genomu ulegają ekspresji. Uzyskujemy także informację na temat struktury intronowo-egzonowej genów
- To podejście nadaje się nie tylko do badań nad genami kodującymi białko, ale także niekodującymi
- Dzięki *tiling arrays* dowiedzieliśmy się, że do 30% genomu człowieka ulega transkrypcji, choć tylko 5% z tej puli dotyczy rejonów kodujących białko
- Oczywiście współczesną alternatywą jest wysokoprzepustowe sekwencjowanie cDNA/RNA

Mikromacierze pokrywające genom (*tiling arrays*)
jako metoda poznawania transkryptomu



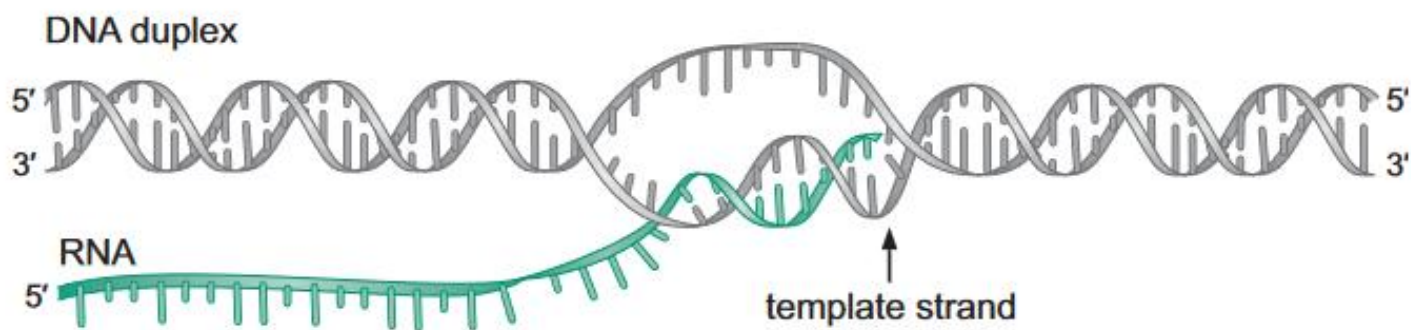
Inne techniki omówimy w dalszej części wykładów



Transkrypcja vs replikacja

- Transkrypcja i replikacja: oba procesy prowadzą do powstania komplementarnej nici kwasu nukleinowego na matrycy DNA
- Podstawowe różnice na poziomie molekularnym to:
 - Transkrypcja prowadzi do powstania RNA, a replikacja: DNA
 - Polimeraza RNA nie wymaga starterów do reakcji syntezy RNA
 - Produkt transkrypcji jest uwalniany (nie pozostaje związany z nicią macierzystą)
 - Możliwa jest jednoczesna transkrypcja tego samego genu przez wiele polimeraz
 - Transkrypcja jest znacznie mniej dokładna niż replikacja (1/10 tys., w porównaniu z 1 / 10 mln) → dlaczego?
 - Transkrypcja działa na wybrane fragmenty genomu, zaś replikacja prowadzi do skopiowania całej informacji genetycznej

Transkrypcja



Polimerazy RNA

- Od bakterii po człowieka, polimerazy RNA mają dość zakonserwowane domeny odpowiedzialne za samą transkrypcję; niemniej jednak, polimerazy RNA składają się z wielu podjednostek, pełniących różne funkcje (wyjątek: niektóre fagowe i organellarne polimerazy RNA, złożone z pojedynczych podjednostek)
- U bakterii wyróżniamy tylko jedną polimerazę RNA
- U *Eukaryota* mamy Polimerazę RNA I, II i III
- Polimeraza RNA I (Pol I) syntetyzuje pre-rRNA 45S, z którego powstają 28S, 18S i 5.8S rRNA.
- Polimeraza RNA II (Pol II) transkrybuje geny kodujące białko, tworząc pre-mRNA. Syntezuje też większość snRNA i małe jąderkowe RNA.
- Polimeraza RNA III (Pol III) syntetyzuje tRNA, 5S rRNA.
- Polimeraza RNA IV i V: obecne tylko u roślin, uczestniczą w transkrypcji małych regulatorowych RNA

Polimerazy RNA

TABLE 13-1 The Subunits of RNA Polymerases

Prokaryotic		Eukaryotic		
Bacterial Core	Archaeal Core	RNAP I (Pol I)	RNAP II (Pol II)	RNAP III (Pol III)
β'	A'/A''	RPA1	RPB1	RPC1
β	B	RPA2	RPB2	RPC2
α^I	D	RPC5	RPB3	RPC5
α^{II}	L	RPC9	RPB11	RPC9
ω	K	RPB6	RPB6	RPB6
	[+six others]	[+nine others]	[+seven others]	[+11 others]

The subunits in each column are listed in order of decreasing molecular weight. (Adapted, with permission, from Ebright R.H. 2000. *J. Mol. Biol.* **304**: 687–698, Fig. 1, p. 688. © Elsevier.)

Transkrypcja: inicjacja, elongacja, terminacja

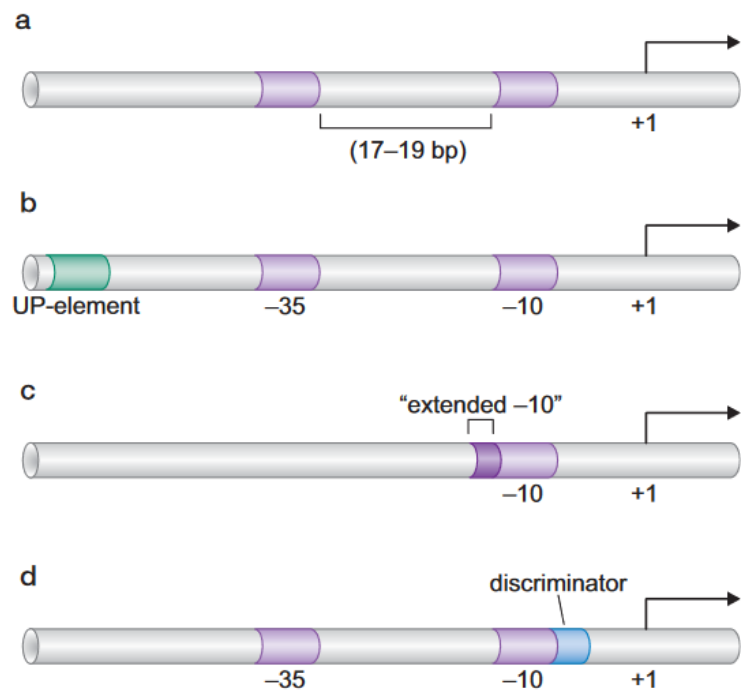
Transkrypcja to proces syntezy różnych rodzajów RNA. Składa się z czterech procesów:

- **Inicjacja:** następuje rozplątanie helisy DNA, przez co polimeraza RNA może przyłączyć się do promotora właściwej nici DNA
- **Elongacja:** polimeraza RNA przesuwa się wzdłuż matrycowej nici DNA a w efekcie tego procesu wydłużana zostaje cząsteczka powstającego RNA.
- **Terminacja:** zakończenie procesu transkrypcji, powiązane z procesem poliadenylacji, czyli dodaniem tzw. ogona poliA na końcu 3' cząsteczki RNA (dotyczy tylko transkryptów kodujących białko)
- **Obróbka RNA:** w trakcie transkrypcji, a częściowo po niej, cząsteczki RNA stają się przedmiotem alternatywnego splicingu, a niekiedy dodatkowej obróbki (np. edytowanie RNA)

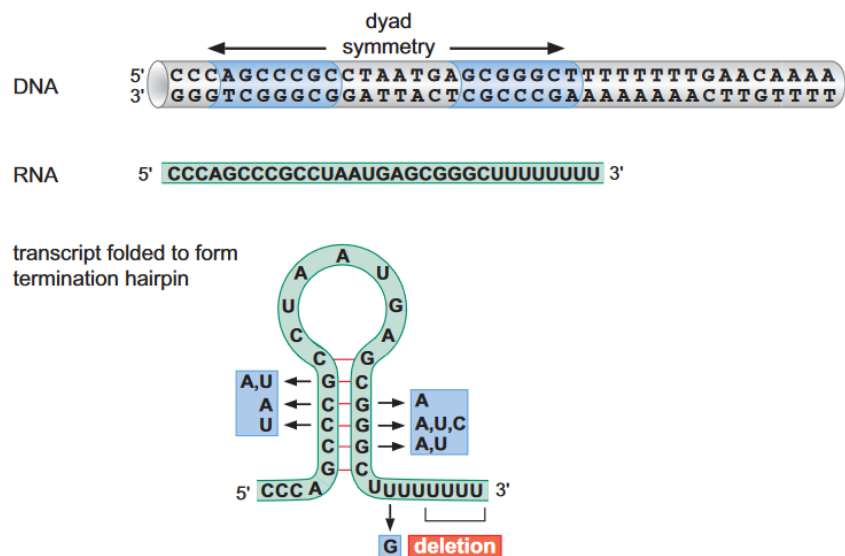
Transkrypcja u bakterii

- Promotor zawiera dwie zakonserwowane sekwencje o długości 6 nt, rozdzielone odcinkiem 19-19 nt
- Te zakonserwowane sekwencje występują 10 oraz 35 nt powyżej miejsca startu transkrypcji, dlatego są nazywane -35 oraz -10.
- Niektóre silne promotory zawierają dodatkowe sekwencje, np. geny rRNA, które zawierają tzw. element UP
- Niektóre silne promotory nie posiadają sekwencji -35, ale za to mają dłuższy fragment -10 w kierunku *upstream* (np. geny gal u *E. coli*, których produkty są odpowiedzialne za metabolizm galaktozy)
- Oprócz tego istnieją promotory posiadające element -10 wydłużony w kierunku *downstream* („dyskryminator”)

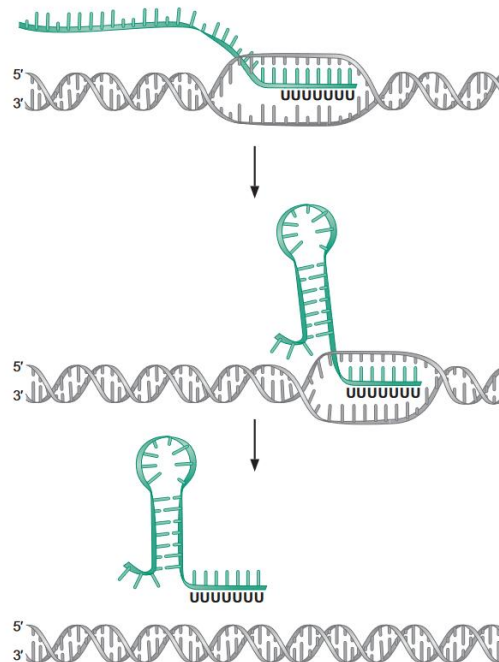
Transkrypcja u bakterii



Transkrypcja u bakterii: terminacja przez sygnały wewnątrz cząsteczki RNA



Transkrypcja u bakterii: terminacja przez sygnały wewnątrz cząsteczki RNA



Jeśli nie wskazano inaczej, część ilustracji pochodzi z:

Molecular Biology of the Gene, 7th edition

James D. Watson, Tania A. Baker, Stephen P. Bell, Alexander Gann,
Michael Levine, Richard Losick

Pearson 2013

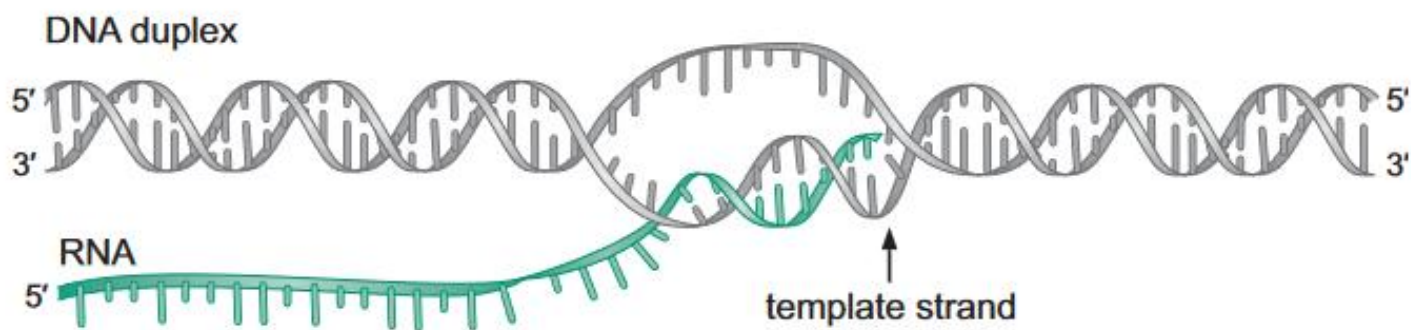
RNA i transkrypcja

1. Przypomnienie podstawowej wiedzy
2. Metody badań

Transkrypcja vs replikacja

- Transkrypcja i replikacja: oba procesy prowadzą do powstania komplementarnej nici kwasu nukleinowego na matrycy DNA
- Podstawowe różnice na poziomie molekularnym to:
 - Transkrypcja prowadzi do powstania RNA, a replikacja: DNA
 - Polimeraza RNA nie wymaga starterów do reakcji syntezy RNA
 - Produkt transkrypcji jest uwalniany (nie pozostaje związany z nicią macierzystą)
 - Możliwa jest jednoczesna transkrypcja tego samego genu przez wiele polimeraz
 - Transkrypcja jest znacznie mniej dokładna niż replikacja (1/10 tys., w porównaniu z 1 / 10 mln) → dlaczego?
 - Transkrypcja działa na wybrane fragmenty genomu, zaś replikacja prowadzi do skopiowania całej informacji genetycznej

Transkrypcja



Polimerazy RNA

- Od bakterii po człowieka, polimerazy RNA mają dość zakonserwowane domeny odpowiedzialne za samą transkrypcję; niemniej jednak, polimerazy RNA składają się z wielu podjednostek, pełniących różne funkcje (wyjątek: niektóre fagowe i organellarne polimerazy RNA, złożone z pojedynczych podjednostek)
- U bakterii wyróżniamy tylko jedną polimerazę RNA
- U *Eukaryota* mamy Polimerazę RNA I, II i III
 - Polimeraza RNA I (Pol I) syntetyzuje pre-rRNA 45S, z którego powstają 28S, 18S i 5.8S rRNA.
 - Polimeraza RNA II (Pol II) transkrybuje geny kodujące białko, tworząc pre-mRNA. Syntezuje też większość snRNA i małe jąderkowe RNA.
 - Polimeraza RNA III (Pol III) syntetyzuje tRNA, 5S rRNA.
 - Polimeraza RNA IV i V: obecne tylko u roślin, uczestniczą w transkrypcji małych regulatorowych RNA

Polimerazy RNA

Polimeraza RNA IV (Pol IV)

- Syntetyzuje jednoniciowe prekursorowe RNA (długości ok. 30-40 nukleotydów)
- Te transkrypty służą jako matryca dla RNA-zależnej polimerazy RNA (RDR2)
- RDR2 przekształca je w dwuniciowe RNA (dsRNA)
- Dicer-like 3 (DCL3) tnie dsRNA na 24-nukleotydowe siRNA
- Transkrybuje głównie sekwencje powtarzalne, transpozony i regiony heterochromatyny

Polimeraza RNA V (Pol V)

- Wytwarza długie niekodujące transkrypty (tzw. *scaffold* RNA); nie produkuje stabilnych transkryptów - jej RNA pełni rolę "rusztowania molekularnego,,
- Te transkrypty rekrutują kompleks wyciszający AGO4-siRNA do określonych miejsc w genomie
- Pozwala zrekrutować metylotransferazę DRM2, co prowadzi do metylacji DNA

Polimerazy RNA

TABLE 13-1 The Subunits of RNA Polymerases

Prokaryotic		Eukaryotic		
Bacterial Core	Archaeal Core	RNAP I (Pol I)	RNAP II (Pol II)	RNAP III (Pol III)
β'	A'/A''	RPA1	RPB1	RPC1
β	B	RPA2	RPB2	RPC2
α^I	D	RPC5	RPB3	RPC5
α^{II}	L	RPC9	RPB11	RPC9
ω	K	RPB6	RPB6	RPB6
	[+six others]	[+nine others]	[+seven others]	[+11 others]

The subunits in each column are listed in order of decreasing molecular weight. (Adapted, with permission, from Ebright R.H. 2000. *J. Mol. Biol.* **304**: 687–698, Fig. 1, p. 688. © Elsevier.)

Transkrypcja: inicjacja, elongacja, terminacja

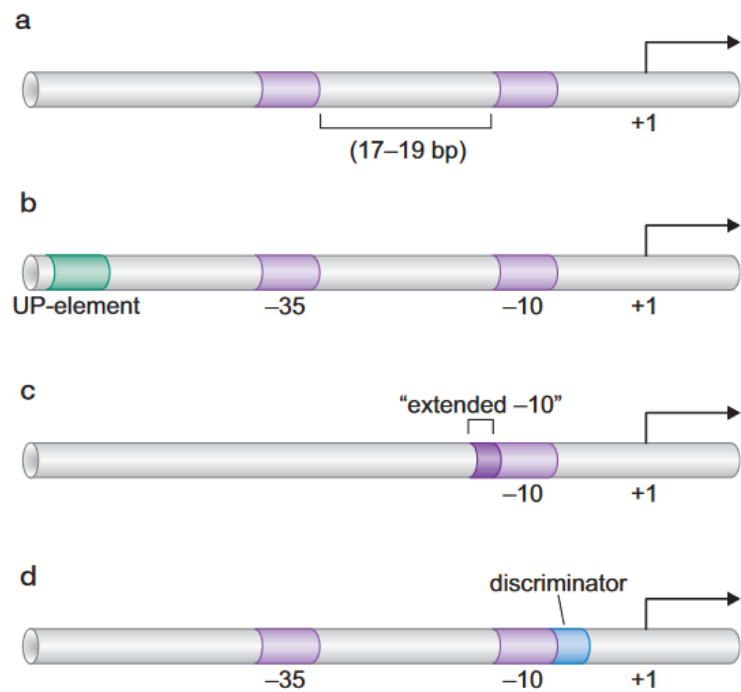
Transkrypcja to proces syntezy różnych rodzajów RNA. Składa się z czterech procesów:

- **Inicjacja:** następuje rozplątanie helisy DNA, przez co polimeraza RNA może przyłączyć się do promotora właściwej nici DNA
- **Elongacja:** polimeraza RNA przesuwa się wzdłuż matrycowej nici DNA a w efekcie tego procesu wydłużana zostaje cząsteczka powstającego RNA.
- **Terminacja:** zakończenie procesu transkrypcji, powiązane z procesem poliadenylacji, czyli dodaniem tzw. ogona poliA na końcu 3' cząsteczki RNA (dotyczy tylko transkryptów kodujących białko)
- **Obróbka RNA:** w trakcie transkrypcji, a częściowo po niej, cząsteczki RNA stają się przedmiotem alternatywnego splicingu, a niekiedy dodatkowej obróbki (np. edytowanie RNA)

Transkrypcja u bakterii

- Promotor zawiera dwie zakonserwowane sekwencje o długości 6 nt, rozdzielone odcinkiem 17-19 nt
- Te zakonserwowane sekwencje występują 10 oraz 35 nt powyżej miejsca startu transkrypcji, dlatego są nazywane -35 oraz -10.
- Niektóre silne promotory zawierają dodatkowe sekwencje, np. geny rRNA, które zawierają tzw. element UP
- Niektóre silne promotory nie posiadają sekwencji -35, ale za to mają dłuższy fragment -10 w kierunku *upstream* (np. geny gal u *E. coli*, których produkty są odpowiedzialne za metabolizm galaktozy)
- Oprócz tego istnieją promotory posiadające element -10 wydłużony w kierunku *downstream* („dyskryminator”)

Transkrypcja u bakterii



Identyfikacja promotorów bakteryjnych - metody bioinformatyczne

1. Wyszukiwanie sekwencji konsensusowych

- Szukamy charakterystycznych motywów:
- Box -10 (Pribnow box): TATAAT (konsensus o70)
- Box -35: TTGACA
- Odległość między nimi: ~17 nukleotydów
- Pozycja: przed (+1) miejscem startu transkrypcji
- Dopuszczenie 1-2 mismatches (promotory naturalne rzadko są idealne)

2. Macierze PWM (Position Weight Matrix)

- Budowanie profilu prawdopodobieństwa na podstawie znanych promotorów
- Uwzględnia częstość występowania każdej zasady w każdej pozycji
- Bardziej elastyczne niż sztywne sekwencje konsensusowe
- Obliczanie wyniku (score) dla kandydujących sekwencji

Algorytmy:

- MEME (Multiple EM for Motif Elicitation); odkrywa nowe
- FIMO (Find Individual Motif Occurrences); identyfikuje znane motywy

3. Analiza regionów upstream

- Przeszukiwanie 150-300 pz przed genem kodującym
- W operonach: tylko przed pierwszym genem

4. Bazy danych i narzędzia

Dedykowane bazy:

- RegulonDB (E. coli)
- cdbprom
- PPD-Prokaryotic Promoter Database (potwierdzone eksperymentalnie)

Programy:

e.g. BacPP (Bacterial Promoter Prediction), 70ProPred

Machine Learning:

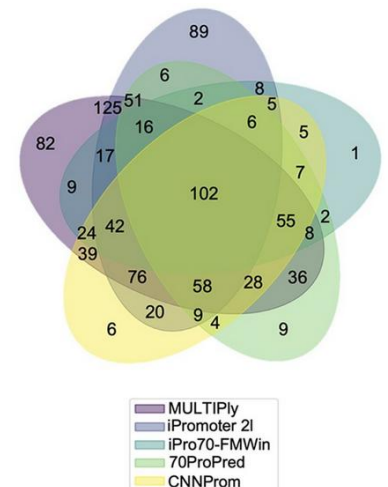
- Modele trenowane na eksperymentalnie zwalidowanych promotorach
- Uwzględniają dodatkowe cechy: struktura DNA, stabilność termodynamiczna
- Random Forest, SVM, sieci neuronowe

Genomika porównawcza:

- Identyfikacja zakonserwowanych motywów w regionach międzygenowych
- Zakładamy, że funkcjonalne promotory są zakonserwowane ewolucyjnie

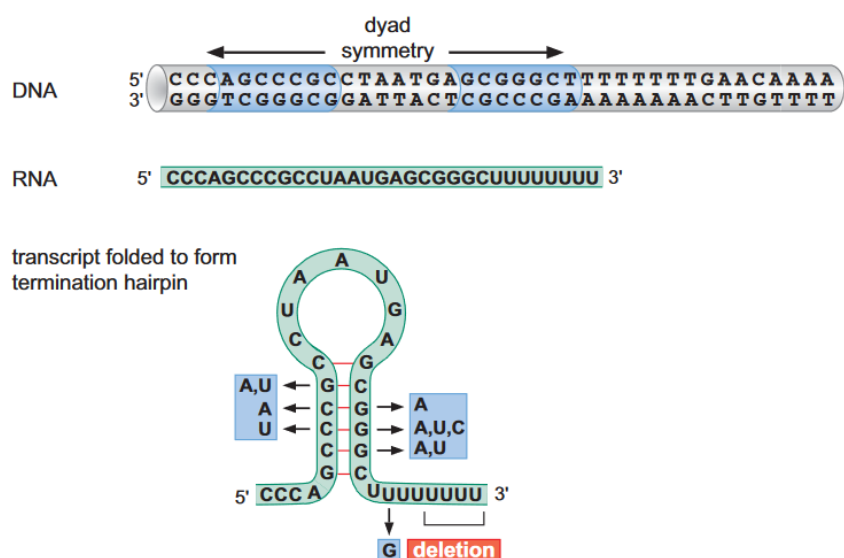
Integracja z danymi eksperymentalnymi

- RNA-seq → określenie poziomu ekspresji i TSS
- ChIP-seq → wiązanie polimerazy RNA lub czynników σ

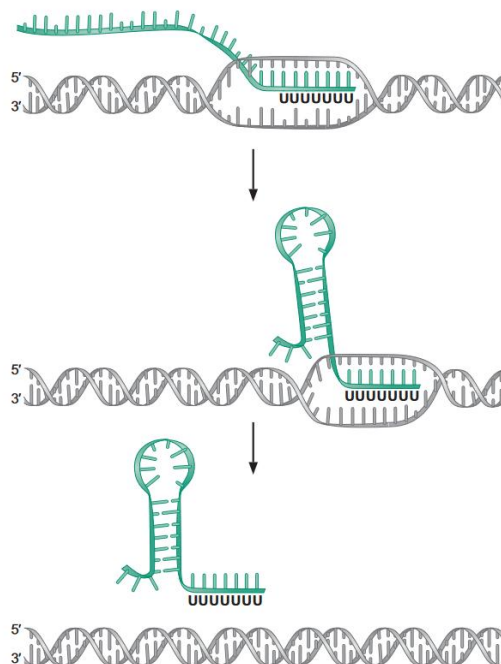


Cassiano MHA, Silva-Rocha R. Benchmarking Bacterial Promoter Prediction Tools: Potentialities and Limitations. mSystems. 2020 Aug 25;5(4):e00439-20. doi: 10.1128/mSystems.00439-20. PMID: 32843538; PMCID: PMC7449607.

Transkrypcja u bakterii: terminacja przez sygnały wewnątrz cząsteczki RNA



Transkrypcja u bakterii: terminacja przez sygnały wewnątrz cząsteczki RNA



Identyfikacja terminatorów transkrypcji u bakterii

Terminatory **rho-niezależne** (intrinsic)

- Struktura hairpin z regionem bogatym w GC
- Po spince: ciąg poli-U (5-8 reszt uracylu)
- Stabilna struktura drugorzędowa ($\Delta G \leq -10$ kcal/mol)

Narzędzia bioinformatyczne:

- TransTermHP (najczęściej używane)
- Przewiduje pozycję, strukturę i stabilność spinki
- Wymaga genomu + adnotacji genów
- Output: pozycja, energia swobodna, pewność (confidence score)

Identyfikacja terminatorów transkrypcji u bakterii

Terminatory **rho-zależne**

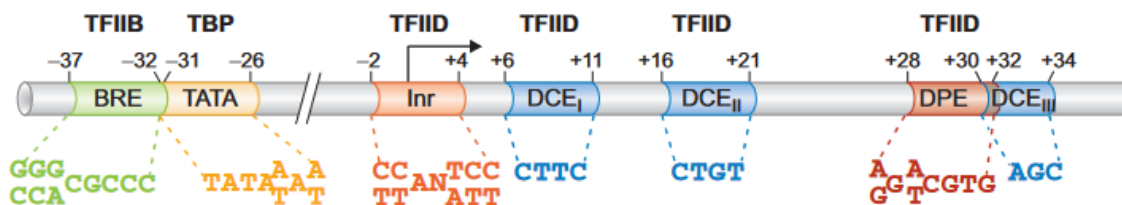
- Brak jednoznacznej sekwencji konsensusowej
- Region bogaty w C i ubogi w G (~70 nt)
- Brak silnych struktur drugorzędowych
- Miejsce wiązania Rho (rut site - rho utilization site)
- Niesparowane regiony bogate w cytozynę

Narzędzia:

- iTerm-PseKNC
- RhoTermPredict

Transkrypcja u Eukaryota

- Promotor ma długość zwykle w zakresie 40-60 nt
- Typowo, promotor posiada kombinację takich elementów, jak BRE, TATA, Inr, DPE, DCE czy MTE.
- Oprócz samego promotora, można wyróżnić dodatkowe sekwencje, które poprzez oddziaływanie z odpowiednimi czynnikami białkowymi wpływają na wydajność transkrypcji, np. wzmacniacze i wyciszacze (*enhancers*, *silencers*)
- Na promotorze, polimeraza II RNA razem z czynnikami białkowymi tworzy tzw. kompleks preinicjacyjny

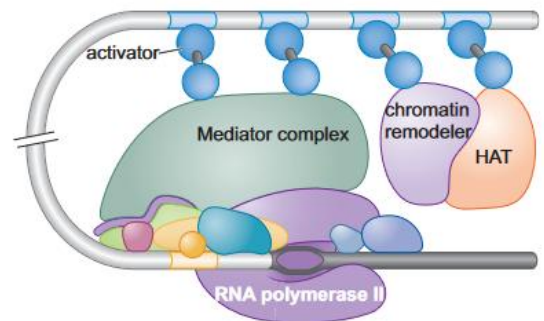


Transkrypcja u Eukaryota

- Tylko ~10-20% genów u ssaków ma klasyczną sekwencję TATA. Większość ma promotory CpG-island (bogate w GC). TATA-less promoters często mają rozmyte TSS (wiele miejsc startu). Implikacja dla nas: Proste wyszukiwanie TATA-box znajdzie tylko niewielką część promotorów!
- Pol II często się zatrzymuje ~20-60 nt za TSS. Potrzebuje P-TEFb (kinaza) aby kontynuować elongację. To główny punkt regulacji (nie sama inicjacja!). Widoczne w ChIP-seq jako "peak" tuż za promotorem. U Drosophila: ~30% genów, u ssaków: ~50-70% genów
- Transkrypcja nie odbywa się równomiernie w jądrze. Pol II tworzy "fabryki" - ogniska transkrypcyjne (kilkaset cząsteczek Pol II). DNA jest "wciągane" do fabryki (nie polimeraza porusza się po DNA!)

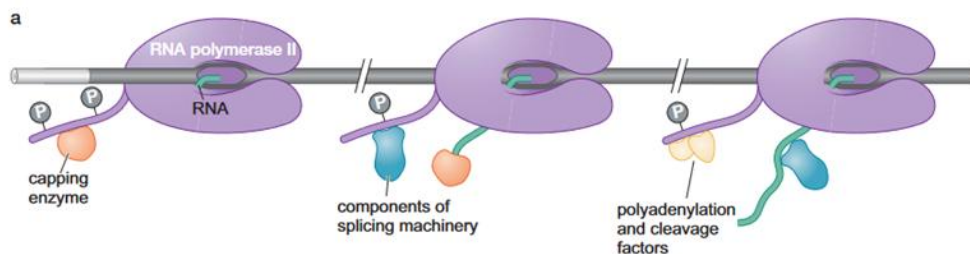
Transkrypcja u Eukaryota

Oprócz samego promotora, polimerazy II RNA oraz podstawowych czynników transkrypcyjnych potrzebne są jeszcze dodatkowe białka, dzięki którym Pol wchodzi w interakcje m.in. z kompleksami modyfikującymi chromatynę. Jednym z takich „łączników” jest białkowy kompleks o nazwie **Mediator**, który jest łącznikiem między polimerazą a białkowymi czynnikami (aktywatorami) przyłączonymi do cząsteczki DNA.



Transkrypcja u Eukaryota

Na etapie elongacji następują przetasowania w kompleksie transkrypcyjnym, m.in. opuszcza go większość czynników związanych typowo z inicjacją transkrypcji, a zastępują je czynniki elongacyjne, jak i enzymy uczestniczące w obróbce RNA.



Transkrypcja u Eukaryota

- Eukariotyczne RNA są przedmiotem wielowątkowej obróbki: dodanie czapeczki na końcu 5', poliadenylacja na końcu 3' czy też splicing; dopiero tak „przetworzona” cząsteczka może zostać eksportowana do cytoplazmy, gdzie staje się matrycą w procesie translacji
- Te procesy są ściśle powiązane z elongacją i terminacją transkrypcji: wiele czynników białkowych jest wspólnych dla tych procesów, prawdopodobnie dla wydajniejszej ich koordynacji
- Pierwszą w kolejności modyfikacją jest dodanie czapeczki (ang. *cap*), stanowiącą modyfikowany (metylowany) nukleotyd guaninowy

Transkrypcja u Eukaryota

- CTD (C-terminal domain) Pol II = kod fosforylacyjny
- CTD zawiera 52 powtórzenia sekwencji YSPTSPS (u człowieka)
- Różne wzorce fosforylacji = różne etapy transkrypcji:
 - Ser5-P → inicjacja, capping
 - Ser2-P → elongacja, splicing
 - Tyr1-P → terminacja

Transkrypcja u Eukaryota

Capping następuje po 20-30 nukleotydach. Cap (7-metylguanozyna) dodawany podczas elongacji. RNA bez cap jest natychmiast degradowane (quality control)

Splicing zaczyna się PRZED zakończeniem transkrypcji. Pierwsze introny mogą być wycinane, gdy ostatnie egzony są jeszcze transkrybowane. Kinetyka transkrypcji wpływa na alternatywny splicing:

- Wolna Pol II → więcej czasu na rozpoznanie słabych splice sites
- Szybka Pol II → pominięcie słabych exonów

To tzw. "kinetic coupling model"

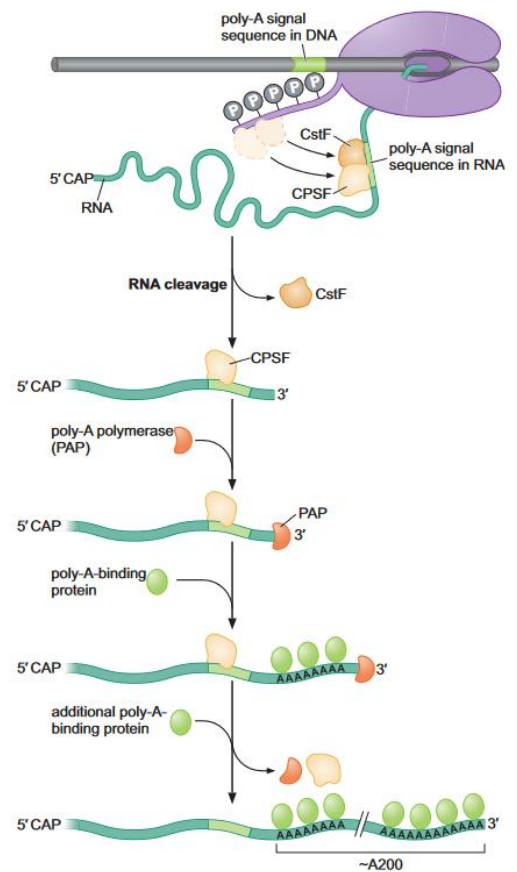
Transkrypcja u Eukaryota

- *Divergent transcription* - od tego samego promotora w obie strony. ~20% promotorów pokazuje dwukierunkową transkrypcję. *Upstream RNA* zwykle krótkie i niestabilne (*cryptic unstable transcripts*)
- *Gene loops*: koniec 3' genu czasami fizycznie kontaktuje się z promotorem (ChIA-PET, Hi-C). Ułatwia to reinicjację transkrypcji. Mediowane przez czynniki terminacji i inicjacji równocześnie. U drożdży bardzo wyraźne, u ssaków subtelniejsze
- Mitochondria używają Pol mt / POLRMT - zupełnie innej polimerazy. Single-subunit polimeraza (jak u fagów T7). Jedna polimeraza dla całego genomu mitochondrialnego. Transkrypty policistronowe, jak u bakterii

Transkrypcja u Eukaryota

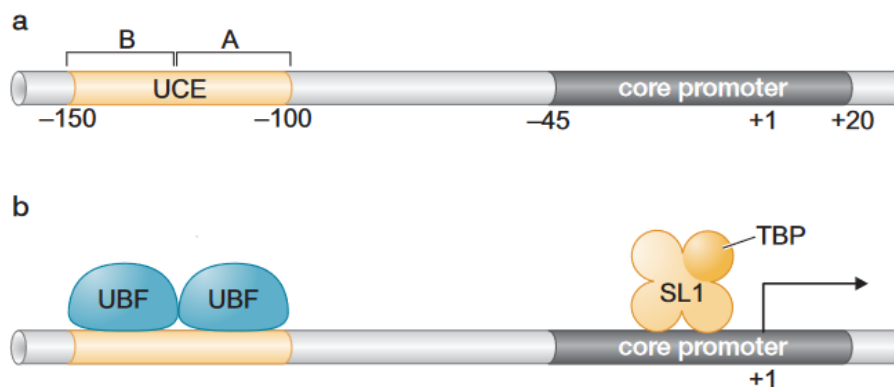
Gdy polimeraza sięga końca genu, napotyka specjalną sekwencję (sygnał poliadenylacji), który będąc przepisany na RNA przyciąga czynniki białkowe uczestniczące w procesie poliadenylacji. Skutkuje to następującymi zdarzeniami:

- cięcie cząsteczki RNA
- dodanie wielu A do końca 3';
- Degradacja fragmentu RNA wciąż związanego z polimerazą
- Terminacja transkrypcji



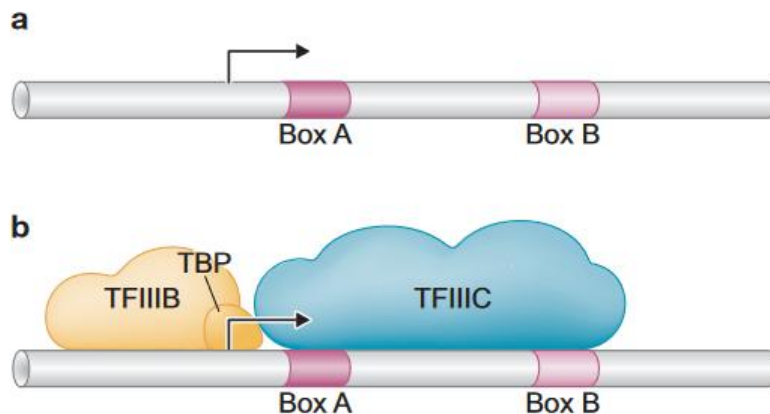
Pol I RNA

- Enzym ten odpowiada za ekspresję zaledwie jednego genu, kodującego prekursor rRNA (wytwarza prekursor 47S (45S) rRNA, który jest następnie cięty do 18S, 5.8S i 28S rRNA rybosomu). rRNA występuje w komórce w wielu kopiach, stąd zapotrzebowanie na tę wyspecjalizowaną polimerazę RNA.
- Promotor Pol I składa się z części głównej oraz elementu UCE (ang. *Upstream control element*)



Pol III RNA

- Zdecydowana większość promotorów tej polimerazy jest zlokalizowana w ciele genu, a więc *downstream*
- Promotor genów tRNA, transkrybowanych przez tę polimerazę, składają się z dwóch części: Box A i Box B. Promotory innych genów (np. 5S rRNA) posiadają natomiast Box A i Box C
- Tak jak w przypadku Pol I i Pol II, wymagane są tutaj czynniki transkrypcyjne, np. TFIIB czy TFIIC.



Bazy danych powiązane tematycznie

Promotory, czynniki transkrypcyjne

EPD: baza danych promotorów u Eukaryota

<https://epd.epfl.ch/>

**EPD**
EUKARYOTIC PROMOTER DATABASE

Computational Cancer Genomics | ExPASy | EPFL

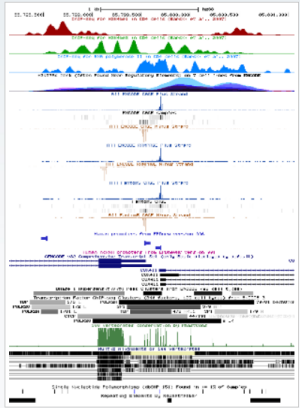
Check out our short survey  Follow @EPD_SIB

Access EPDnew

- H. sapiens
- H. sapiens non-coding
- M. mulatta
- M. musculus
- M. musculus non-coding
- R. norvegicus
- C. familiaris
- G. gallus
- D. melanogaster
- A. mellifera
- D. rerio
- C. elegans
- A. thaliana
- Z. mays
- S. cerevisiae
- S. pombe
- P. falciparum
- Select / Download
- Promoter analysis tools
- FTP site

hsEPDnew, the *Homo sapiens* (human) curated promoter database

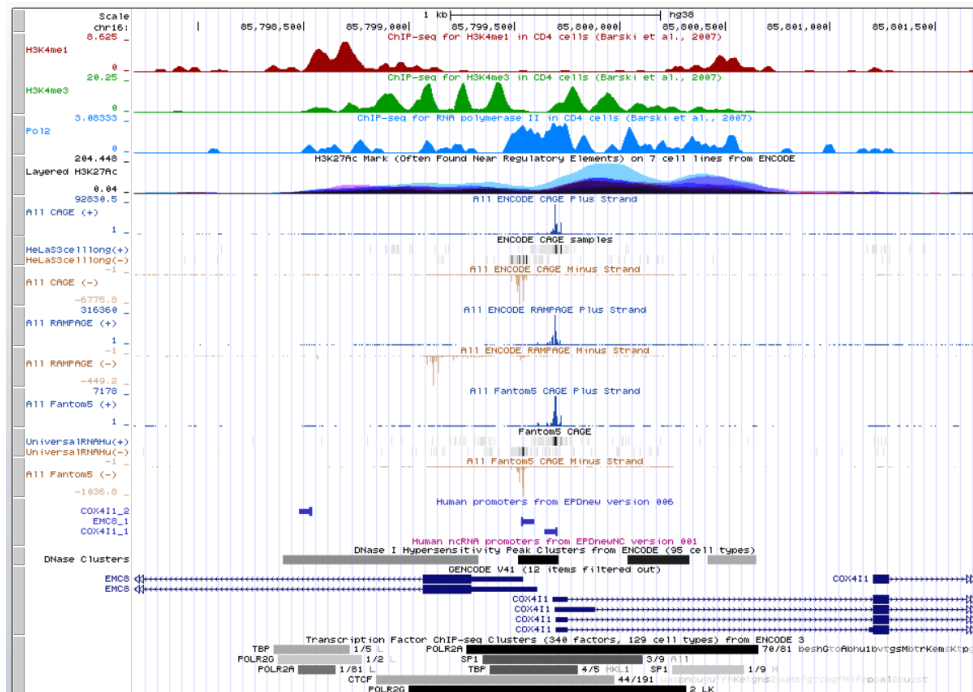
Version	006
Coverage	29598 promoters 16455 genes
Genome assembly	H. sapiens (Dec 2013 GRCh38/hg38)
Gene annotation	Gencode (v28)
Based on data from	Riken/ENCODE CAGE data downloaded from UCSC FANTOM5 data Rampage data EPD (old)
Documentation & Viewer(s)	Promoter assembly pipeline description EPD viewer - hg38 (track content) EPD viewer - hg19 (track content) GM12878 viewer (hg19) FANTOM5/ZENBU (hg38) DBTSS/KERO (hg38) refTSS (hg38) SwissRegulon (hg19)



Promoters shown above: [EMC8](#) and [COX411](#).

Access EPD

<https://epd.epfl.ch/>



EPD: baza danych promotorów u Eukaryota

<https://epd.epfl.ch/>

- wiarygodne pozycje TSS i granice promotorów oparte na danych eksperymentalnych (m.in. CAGE/5'-end) i literaturze.
- Zakres: człowiek, mysz i wiele organizmów modelowych (m.in. Drosophila, ryż, Arabidopsis, drożdże).
- Jakość: zestawy niezduplikowanych promotorów z oceną wsparcia dowodami.
- Funkcje: m.in. wyszukiwanie po genie/transkrypcie, linki do UCSC/Ensembl
- Pobieranie: eksport BED/FASTA/GFF dla pojedynczych promotorów lub całych gatunków.

JASPAR: baza danych czynników transkrypcyjnych

<https://jaspar.genereg.net/>

The screenshot displays the JASPAR 2022 website. The top navigation bar includes the JASPAR 2022 logo, a hamburger menu, a shopping cart icon, and a link to the JASPAR Blog. A left sidebar contains a list of navigation links: Home, About, Search, Browse JASPAR CORE, Unvalidated Profiles, Browse Collections, Tools, RESTful API, Download Data, Matrix Clusters, Genome Tracks, and Enrichment Analysis (marked as 'New'). The main content area features a search bar with the placeholder 'Search JASPAR database...' and a search button. Below the search bar, there are examples of search terms: 'SPI1, P17676, ChIP-seq, Homo sapiens'. A section titled 'Browse JASPAR CORE for 6 different taxonomic groups' displays six categories with representative images: Fungi, Insecta, Nematoda, Plantae, Urochordata, and Vertebrata. To the right of this section is a large banner for 'The high-quality transcription factor binding profile database' with a 'JASPAR Interactive tour' button. At the bottom, there is an 'About JASPAR CORE' section with a brief description of the database and a 'When should it be used?' section. On the far right, there is a 'Citing JASPAR 2022' section with a list of authors and a citation link to the Nucleic Acids Research paper.

JASPAR²⁰²²

Home About Search Browse JASPAR CORE Unvalidated Profiles Browse Collections Tools RESTful API Download Data Matrix Clusters Genome Tracks Enrichment Analysis **New**

Search JASPAR database... Search

Examples: SPI1, P17676, ChIP-seq, Homo sapiens Advanced Options

Browse JASPAR CORE for 6 different taxonomic groups

Fungi Insecta Nematoda Plantae Urochordata Vertebrata

About JASPAR CORE Info about other collections

The JASPAR CORE database contains a curated, non-redundant set of profiles, derived from published collections of experimentally defined transcription factor binding sites for eukaryotes. The prime difference to similar resources (TRANSFAC, etc.) consist of the open data access, non-redundancy and quality.

When should it be used? When seeking models for specific factors or structural classes, or if experimental evidence is paramount

Citing JASPAR 2022 PubMed | NAR

Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N, Fornes O, Leung TY, Aguirre A, Hammal F, Schmelter D, Baranasic D, Ballester B, Sandelin A, Lenhard B, Vandepoele K, Wasserman WW, Parcy F, and Mathelier A
JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles
Nucleic Acids Res. 2022 Jan 7;50(D1):D165-D173.; doi: 10.1093/nar/gkab1113

JASPAR: baza danych czynników transkrypcyjnych

<https://jaspar.genereg.net/>

JASPAR²⁰²²

Home

About

Search

Browse JASPAR CORE

Unvalidated Profiles

Browse Collections

Tools

RESTful API

Download Data

Matrix Clusters

Genome Tracks

Enrichment Analysis new

Detailed information of matrix profile **MA0001.2**

Profile summary Add

Name:

AGL3

Matrix ID:

MA0001.2

Class:

MADS box factors

Family:

MIKC

Collection:

CORE

Taxon:

Plants

Species:

[Arabidopsis thaliana](#)

Data Type:

SELEX

Validation:

8597661

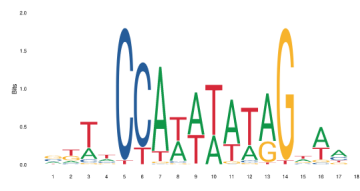
Uniprot ID:

P29383

Source:

Comment:

Sequence logo Download SVG



Frequency matrix JASPAR TRANSFAC MEME RAW PFM Reverse comp.

A	22	16	25	27	0	0	82	40	56	35	65	25	64	0	33	52	45	21
C	29	9	8	16	92	79	1	4	0	0	1	4	0	0	14	5	23	24
G	34	29	4	13	0	0	2	3	1	0	4	3	28	92	15	7	13	26
T	10	41	58	39	3	16	10	48	38	60	25	63	3	3	33	31	14	24

Binding sites information

External links

JASPAR: baza danych czynników transkrypcyjnych

<https://jaspar.genereg.net/>

JASPAR is a regularly maintained open-access database storing manually curated transcription factors (TF) binding profiles as position frequency matrices (PFMs). PFMs summarize occurrences of each nucleotide at each position in a set of observed TF-DNA interactions. PFMs can be transformed to probabilistic or energetic models to construct position weight matrices (PWMs) or position-specific scoring matrices (PSSMs), which can be used to scan any DNA sequence to predict TF binding sites (TFBSs). The JASPAR database provides TFBSs predicted using the profiles in the CORE collection.

The motifs in JASPAR are collected in two ways:

- **Internally:** de novo generated motifs, by analyzing ChIP-seq/-exo sequences using a custom motif discovery pipeline (check the code at our [repository](#)).
- **Externally:** motifs taken directly from other publications and/or resources.

In both cases, the selected motifs are manually curated. Specifically, our curators assess the quality of the motif and search for an orthogonal publication providing support to the motif as the bona fide motif recognized by the TF of interest (e.g., a motif found in ChIP-seq peaks looks similar to one found by SELEX-seq). The Pubmed ID associated with the orthogonal support is provided in the TF profile metadata.

JASPAR is the only database with this scope where the data can be used with no restrictions (open source). For a comprehensive review of models and how they can be used, please see the following reviews

- [Modeling the specificity of protein-DNA interactions Quant Biol. 2013 Jun;1\(2\):115-130](#)
- [Applied bioinformatics for the identification of regulatory elements Nat Rev Genet. 2004 Apr;5\(4\):276-87](#)

JASPAR CORE

The JASPAR CORE database contains a curated, non-redundant set of profiles, derived from published collections of experimentally defined transcription factor binding sites for eukaryotes. The prime difference to similar resources (TRANSFAC, etc.) consist of the open data access, non-redundancy and quality.

When should it be used? When seeking models for specific factors or structural classes, or if experimental evidence is paramount

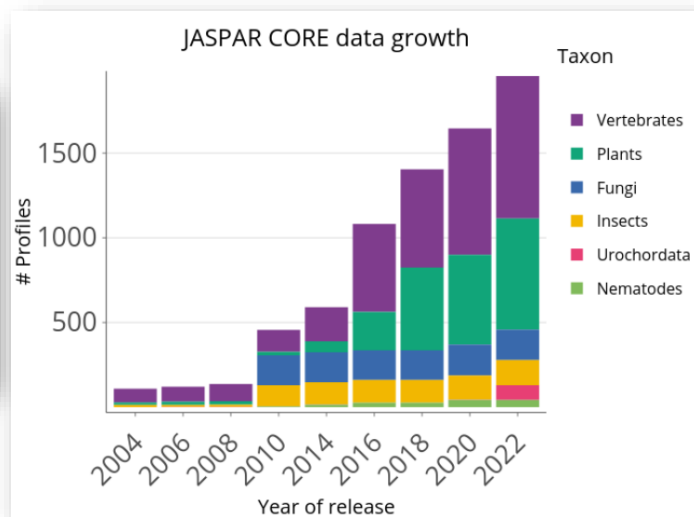
JASPAR: baza danych czynników transkrypcyjnych

<https://jaspar.genereg.net/>

JASPAR CORE

The JASPAR CORE database contains a curated, non-redundant set of profiles, derived from published collections of experimentally defined transcription factor binding sites for eukaryotes. The prime difference to similar resources (TRANSFAC, etc.) consist of the open data access, non-redundancy and quality.

When should it be used? When seeking models for specific factors or structural classes, or if experimental evidence is paramount

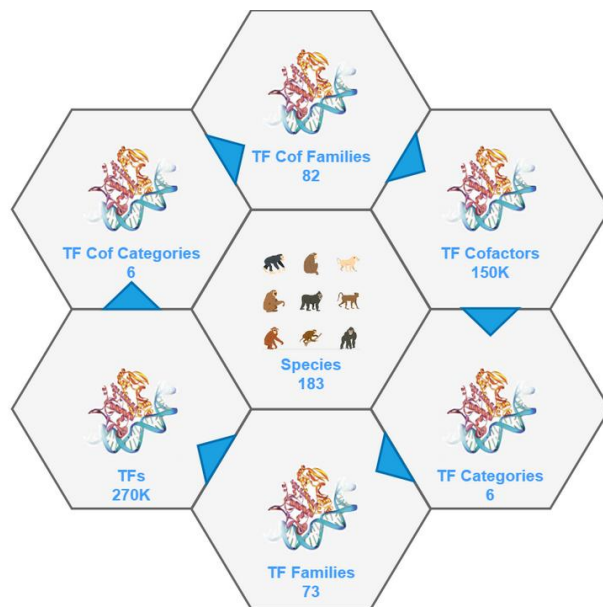


JASPAR: baza danych czynników transkrypcyjnych

<https://jaspar.genereg.net/>

- ogólnodostępna, curated baza profili wiązania TF (motywów).
- Zawartość: profile PFM/PWM/PCM + loga motywów; stabilne ID w formacie MAXxxx.x (wersjonowanie).
- Pochodzenie danych: eksperymenty typu ChIP-seq, SELEX itd. + literatura; zbiór CORE (wysoka jakość, nieredundantne)
- Zakres taksonomiczny: liczne kłady (m.in. kręgowce, rośliny, grzyby, owady, nicienie).
- Funkcje: wyszukiwanie po TF/gene/family, filtrowanie po gatunku/tkance/źródle, skanowanie sekwencji pod TFBS.
- Eksport/API: pobieranie w formatach JASPAR, MEME, TRANSFAC, FASTA; REST API, pakiety TFBSTools (R) i pyjaspar (Python).
- Zastosowania: predykcja miejsc wiązania TF, analiza wzbogacenia motywów, adnotacja promotorów/enhancerów.

AnimalTFDB: baza danych czynników transkrypcyjnych
<http://bioinfo.life.hust.edu.cn/AnimalTFDB4/#/>



AnimalTFDB: baza danych czynników transkrypcyjnych

<http://bioinfo.life.hust.edu.cn/AnimalTFDB4/#/>

Input

The family assignment rules and thresholds determined by established methods ([see details](#)) are used to identify transcription factors from the input protein sequences.

Input the protein sequences in FASTA format (less than 10000 protein sequences)

>Q7Z591

MASSETEIRWAEPLGKGPKRRRWAAEDKRDVDRSSSQSWEEERLFPNATSPELLE
DFR
LAQQHLPPLEWDHPQPDGHQDSGESGETSGEEAEEDVDSPASSHEPLAMLQQGRQ
LDM
TEEEPDGTLGSLVEEEAGESSSRLGYEAGLSLEGHGNTSPMALGHGQARGWVASGEQ
ASG
DKLSEHSEVNPVELSPARSWSSGTVSLDHPSDSLDSWGETDGPQTALAEPLPE
GPS
HHLLSPDGRGTGSVARATPMEFQDSSAPPAQSPQHATDRWRRETTTRFCPQPKHIW
KQT
KTSPKPLPSRFIGSISPLNPQPRPTRQGRPLPRQGATLAGRSSNAPKYGRGQLNYP
LPD
FSKVGPRVRFPKDESYRPPKSRSHNRKPKQAPARPLIFKSPAETVQEVLLSSGEAALA
KDT

Reset

Submit

Result

TF	Family	Full sequence e-value	Best domain e-value	Domain number
P23769	GATA	5.9000000000000005e-33	3.8e-18	2
P15976	GATA	9.800000000000001e-33	5.6e-18	2
P15336	TF_bZIP	1.8e-23	3.3e-23	1
P18846	TF_bZIP	1.2e-17	2.5e-17	1
Q8WY36	HMG_box	1.4e-20	4.9e-20	2

AnimalTFDB: baza danych czynników transkrypcyjnych

<http://bioinfo.life.hust.edu.cn/AnimalTFDB4/#/>

Transcription Factor Binding Site Prediction

Input

Prediction of potential transcription factor binding sites on target genes by inputting the sequence of the target gene (recommended promoter sequences for input genes) ([sell details](#))

Input the sequence in FASTA format

>Example

ATCGATCGACTAGCTAGCTAGCTACGATCGATCGATCAGCTGACTGACTAGCTACGT

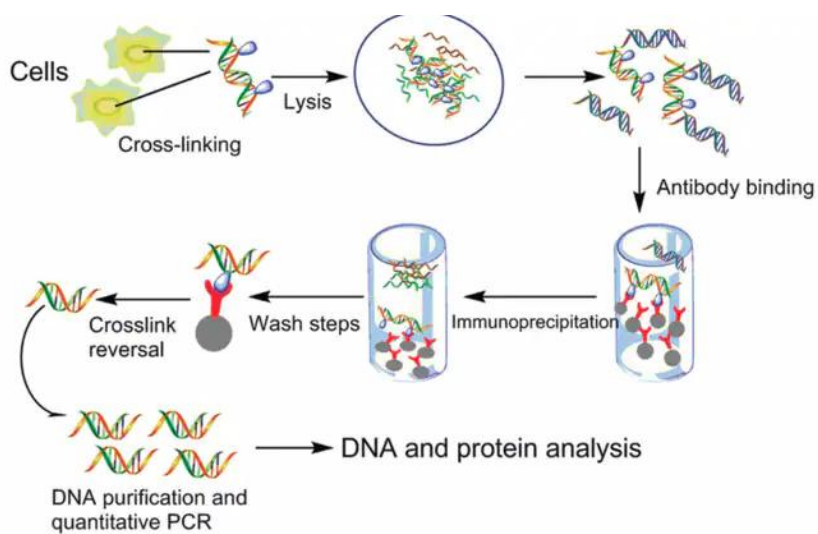
CG

Result

TF	Source	Query	Start	Stop	Strand	Score	P value	Q value	Matched Sequence
CUX2	CISBP	Example	28	39	-	14.8976	2.5e-06	0.00024	CTGATCGATC GA
CUX2	CISBP	Example	28	39	+	15.3853	4.32e-06	0.000207	TCGATCGATC AG
TFAP4	hTFtarget	Example	35	44	-	15.0306	5.33e-06	0.000312	GTCAGCTGA T
TFAP4	hTFtarget	Example	35	44	+	14.6327	5.33e-06	0.000312	ATCAGCTGA C
TFAP4	hTFtarget	Example	35	44	+	14.9796	6.24e-06	0.000312	ATCAGCTGA C

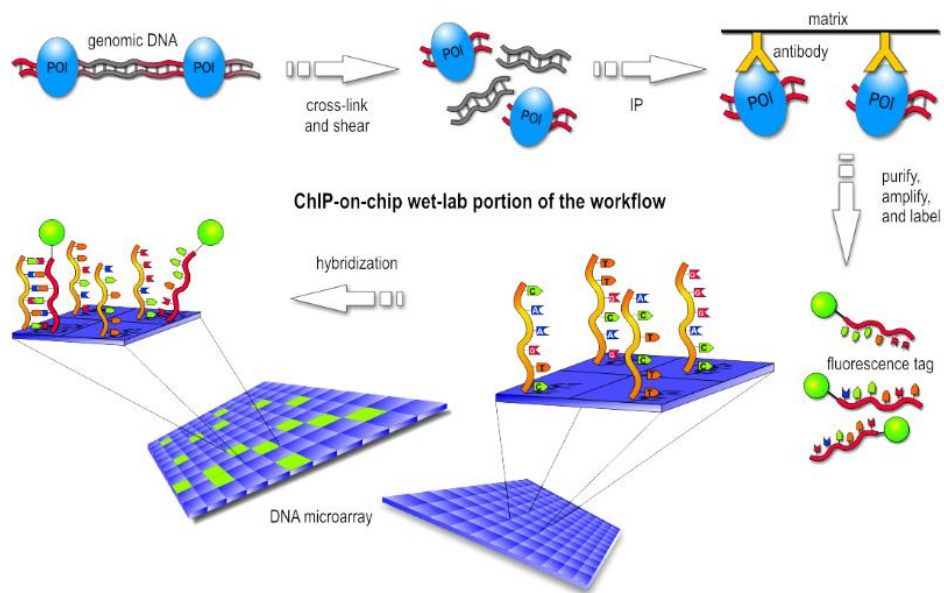
Techniki eksperymentalne

ChIP: badanie oddziaływań białko:DNA



<https://www.sigmaaldrich.com>

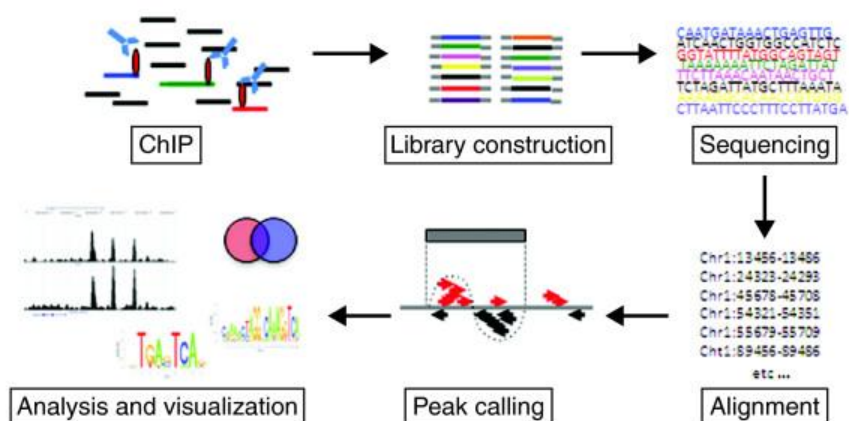
ChIP-on-chip



https://en.wikipedia.org/wiki/ChIP-on-chip#/media/File:ChIP-on-chip_wet-lab.png

ChIP-Seq - chromatin immunoprecipitation sequencing

Sekwencjonowanie fragmentów genomowego DNA, które ulegają precypitacji wraz z określonym, badanym białkiem wiążącym owe fragmenty DNA



ChIP-Seq - chromatin immunoprecipitation sequencing

Podobne metody bazujące na ChIP – Seq:

- **Sono-Seq** - identical to ChIP-Seq but skipping the immunoprecipitation step
- **RIP-Chip** - same goal and first steps, but does not use cross linking methods and uses microarray instead of sequencing
- **Competition-ChIP** - to measure relative replacement dynamics on DNA
- **ChiRP-Seq** - to measure RNA-bound DNA and proteins
- **ChIP-exo** uses exonuclease treatment to achieve up to single base-pair resolution

<http://en.wikipedia.org/wiki/ChIP-sequencing>

ChIP-Seq: co możemy badać?

- ✓ Czynniki transkrypcyjne
- ✓ Polimeraza II RNA
- ✓ Modyfikacje histonów (np. H3K4me3)

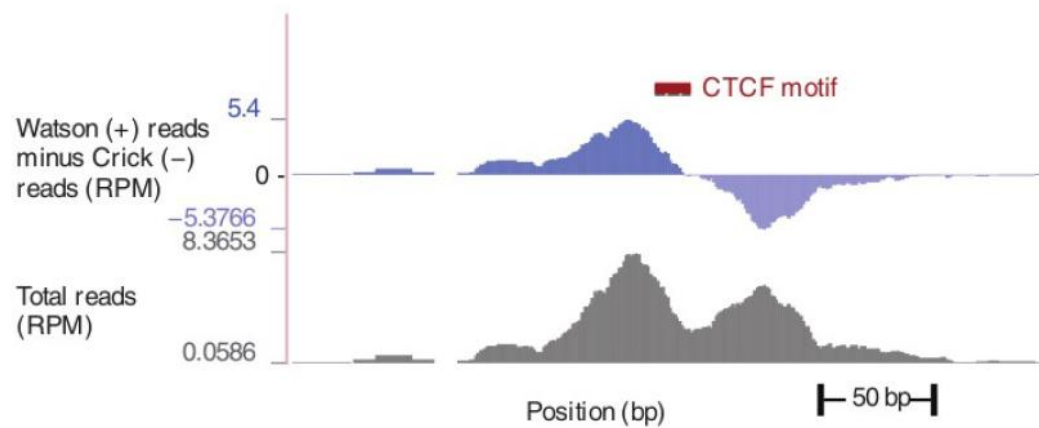
Modyfikacje histonów

H3K4me1: mono-methylation at the 4th lysine residue of the histone H3 protein

Histone modification	Function	Location
H3K4me1	Activation	Enhancers
H3K4me3	Activation	Promoters, bivalent domains
H3K36me3	Activation	Gene bodies
H3K79me2	Activation	Gene bodies
H3K9Ac	Activation	Enhancers, promoters
H3K27Ac	Activation	Enhancers, promoters
H4K16Ac	Activation	Repetitive sequences
H3K27me3	Repression	Promoters in gene-rich regions, developmental regulators, bivalent domains
H3K9me3	Repression	Satellite repeats, telomeres, pericentromeres
Gamma H2A.X	DNA damage	DNA double-strand breaks
H3S10P	DNA replication	Mitotic chromosomes

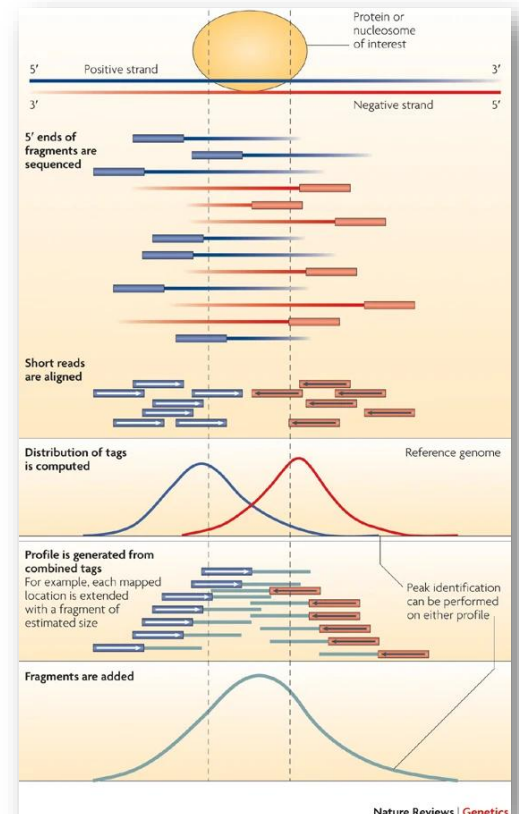
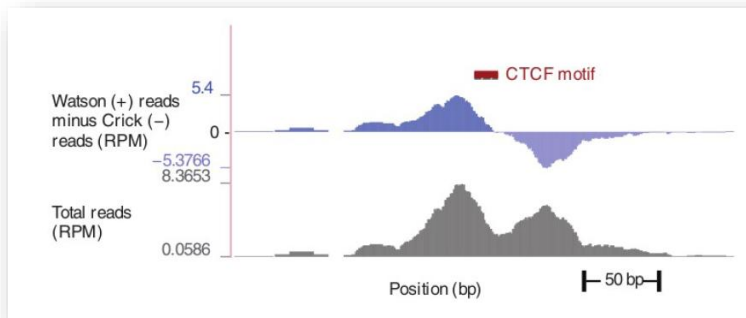
<https://www.abcam.com/epigenetics/histone-modifications>

ChIP-Seq: czynniki transkrypcyjne



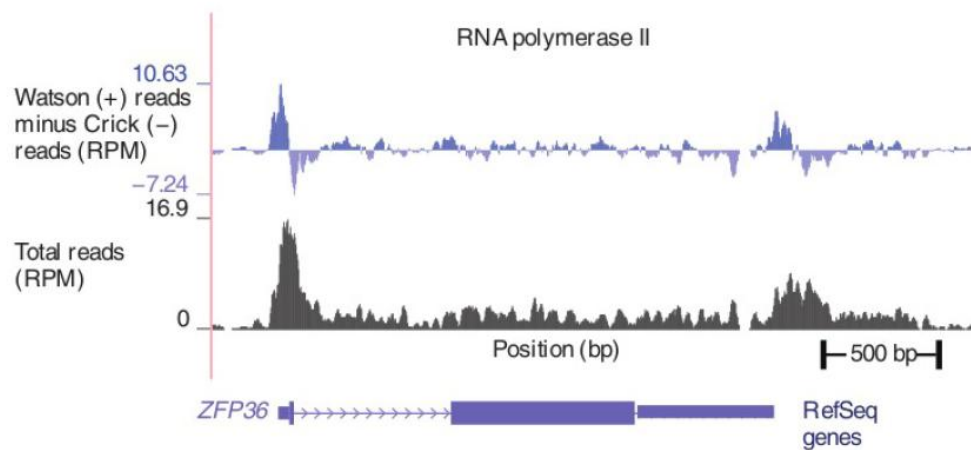
Źródło: Pepke
et al., 2009

ChIP-Seq - chromatin immunoprecipitation sequencing



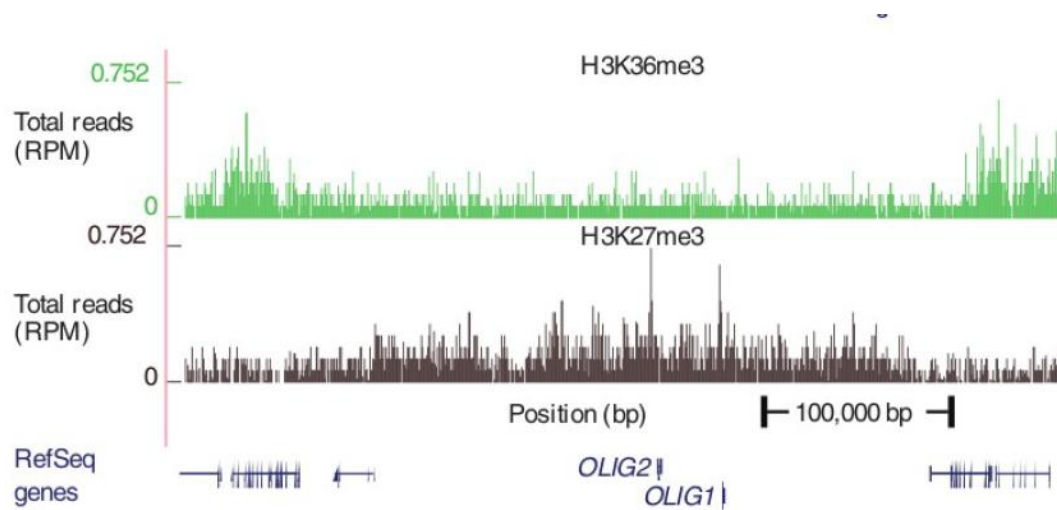
doi.org/10.1038/nrg2641

ChIP-Seq: Polimeraza II RNA



Źródło: Pepke et al., 2009

ChIP-Seq: Modyfikacje histonów



Źródło: Pepke et al., 2009

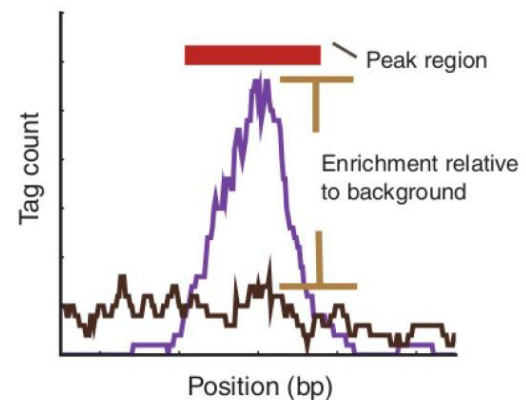
ChIP-Seq: Problem tła

ChIP-Seq opiera się na wzbogaceniu puli DNA we fragmenty wiążące białko, a nie na ich oczyszczaniu

- 60-99% odczytów stanowią sekwencje niespecyficzne (tło)
- Rozkład odczytów tła wzdłuż genomu zależy od wielu czynników – rozmiaru genomu, jego składu oraz artefaktów sekwencjonowania (ang. *sequencing bias*)

Jak poradzić sobie z istnieniem tła?

- Wnioskowanie rozkładu tła z samych danych ChIP-Seq (problematyczne i podatne na błędy)
- Użycie eksperymentalnej kontroli – chromatyny wiązanej przez niespecyficzne przeciwciało (np. anty-IgG) lub nie poddanej immunoprecypitacji (*input DNA*)



Bazy danych: ChIP Atlas

<https://chip-atlas.org/>

 ChIP-Atlas

Peak BrowserTarget GenesColocalizationEnrichment AnalysisPublicationsDocs

Find By ID

 **ChIP-Atlas**

An integrative, comprehensive database to explore public Epigenetic dataset, including ChIP-Seq, DNase-Seq, ATAC-Seq, and Bisulfite-Seq data: ChIP-Atlas covers almost all public data archived in Sequence Read Archive of NCBI, EBI, and DDBJ with over 224,000 experiments.

[Watch movie introduction](#)

What's new

- **New publication** on the NAR web server issue! <https://doi.org/10.1093/nar/gkac199> (2022/03/24)
- Minor bug fix: Peak Browser / Enrichment Analysis UI updates the number of experiments by selecting experiment type (2022/03/01)
- Updated the order of the genome assembly tabs, now GRCh38 is default. And some minor fixes came together! (2022/02/08)
- Added **ATAC-Seq and Bisulfite-Seq**, together with UI improvement including 'peak' icon! (2021/10/04)
- Updated site design: (2021/01/04)
 - [Dataset search](#) is available as a standalone function
 - ID search at the top right corner now accepts GEO Sample ID (e.g. GSM469863) along with SRA Experiment ID (e.g. SRX018625)
 - New information layout for individual experiment list
- Added new genome assemblies: hg38, mm10, dm6, ce11. Happy holidays! (2020/12/01)

Bazy danych: ChIP Atlas

<https://chip-atlas.org/>

ChIP-Atlas

Peak Browser

Target Genes

Colocalization

Enrichment Analysis

Publications

Docs

Find By ID

SRX018625

Go

Advanced

ChIP-Atlas: Peak Browser

Tutorial movies

Visualizes protein binding on given genomic loci with IGV genome browser

H.sapiens (hg38)

H. sapiens (hg19)

M. musculus (mm10)

M. musculus (mm9)

R. norvegicus (m6)

D. melanogaster (dm6)

D. melanogaster (dm3)

C. elegans (ce11)

C. elegans (ce10)

S. cerevisiae (sacCer3)

1. Experiment type

ChIP: Histone (16612)

ChIP: RNA polymerase (2099)

ChIP: TFs and others (15155)

ChIP: Input control (8078)

ATAC-Seq (27318)

DNase-seq (1642)

Bisulfite-Seq (15000)

2. Cell type Class

All cell types (2099)

Adipocyte (6)

Blood (493)

Bone (62)

Breast (255)

Cardiovascular (49)

Digestive tract (239)

Epidermis (71)

3. Threshold for Significance

50

100

200

500

ChIP Antigen (optional)

type to search

All

GTF3C1 (1)

POLR2A (7)

POLR2B (39)

Cell type (optional)

type to search

All

Acute myeloid leukemia (2)

B cells (1)

B-CLL cells (12)

View on IGV

Error connecting to IGV?

Download BED file

Bazy danych: Cistrome

<http://cistrome.org/db/#/>



Cistrome Data Browser

Tips

- Check what factors regulate your gene of interest, what factors bind in your interval or have a significant binding overlap with your peak set. Have a try at [CistromeDB Toolkit](#).
- If you have a Transcription Factor ChIP-seq (and TF perturbed expression) data, [Cistrome-GO](#) help you predict the function of this TF.
- Please help us curate the samples which has incorrect meta-data annotation by clicking the button on the inspector page. Thank you!

Containing word(s):

Species

[All](#)

Homo sapiens

Mus musculus

Biological Sources

[All](#)

1-cell pronuclei

1015c

10326

1064Sk

106A

Factors

[All](#)

AATF

ABCC9

ACSS2

ACTB

ADNP

Results

Batch	Species	Biological Source	Factor	Publication	Quality Control
☐	Homo sapiens	HeLa; Epithelium; Cervix	BTAF1	Johannes F, et al. Bioinformatics 2010	●●●●●
☐	Homo sapiens	HeLa; Epithelium; Cervix	GAPDH	Johannes F, et al. Bioinformatics 2010	●●●●●
☐	Homo sapiens	K562; Erythroblast; Bone Marrow	EGR1	Tang C, et al. Electrophoresis 2010	●●●●●

Bazy danych: NIH Roadmap Epigenomics

<https://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/>



NCBI » GEO » NIH Roadmap Epigenomics

[GEO Publications](#) [FAQ](#) [MIAME](#) [Email GEO](#)

NIH Roadmap Epigenomics Project Data Listings

This page summarizes Roadmap Epigenomics project data that are available through the GEO repository. Additional data will be added as it becomes available.

For data usage terms and conditions, please refer to [Epigenomics Data Access Policies](#)

[Samples](#) [Studies](#) [Matrix](#)

More information and resources are available:

- [Roadmap Epigenomics Homepage](#)
- [Epigenome Browser](#)
- [Epigenome Atlas](#)

[FTP download](#)

[How to use this page](#)

[Search](#) **3256 results** [View tracks](#) [Export](#)

Page of 66 [>](#) [>>](#) Page size

GEO Accession	Sample	Experiment	Track	Center	Download	Embargo end date
GSM409307	H1 cell line	H3K4me1	☐	UCSD	sra bam,bed,wig	2010-05-03
GSM409308	H1 cell line	H3K4me3	☐	UCSD	sra bam,bed,wig	2010-05-03
GSM409312	H1 cell line	H3K36me3	☐	UCSD	sra bam,bed,wig	2010-05-03
GSM410807	H1 cell line	H3K9ac	☐	UCSF-UBC	sra bam,bed,wig	2010-05-03
GSM410808	H1 cell line	H3K4me3	☐	UCSF-UBC	sra bam,bed,wig	2010-05-03
GSM428286	H1 cell line	MRE-Seq	☐	UCSF-UBC	sra bam,bed,wig	2010-10-07
GSM428289	H1 cell line	ChIP-Seq input	☐	UCSF-UBC	sra bam,bed,wig	2010-05-03
GSM428291	H1 cell line	H3K9me3	☐	UCSF-UBC	sra bam,bed,wig	2010-05-03
GSM428295	H1 cell line	H3K27me3	☐	UCSF-UBC	sra bam,bed,wig	2010-05-03
GSM428296	H1 cell line	H3K36me3	☐	UCSF-UBC	sra bam,bed,wig	2010-05-03

<https://www.encodeproject.org/>

[illegible]

ChIP-Seq: narzędzia

- **MACS / MACS2**
- FindPeaks
- PeakSeq
- SISSRS
- QuEST
- GEM
- SICER
- CCAT
- **HOMER**

Jeśli nie wskazano inaczej, część ilustracji pochodzi z:

Molecular Biology of the Gene, 7th edition

James D. Watson, Tania A. Baker, Stephen P. Bell, Alexander Gann,
Michael Levine, Richard Losick

Pearson 2013

Techniki eksperymentalne c.d.
Bazy danych c.d.

Bazy danych: ChIP Atlas

<https://chip-atlas.org/>

The screenshot shows the ChIP-Atlas web application interface. At the top is a dark navigation bar with icons and labels for 'Peak Browser', 'Enrichment Analysis', 'Diff Analysis', 'Target Genes', 'Colo', 'Publications', and 'Docs'. On the right of the bar is a search field with the ID 'SRX018625', a 'Go' button, and a 'Search' icon. Below the navigation bar, the main content area features the 'ChIP-Atlas' logo and a subtitle: 'A data-mining suite for exploring epigenomic landscapes by fully integrating 433,000 ChIP-seq, ATAC-seq and Bisulfite-seq experiments.' The central part of the page contains six interactive tiles arranged in a 2x3 grid. Each tile has an icon, a title, and a list of supported data types in small blue boxes. The tiles are: 'Peak Browser' (icon of a magnifying glass over a DNA strand, supports ChIP, ATAC, Bisulfite), 'Enrichment Analysis' (icon of a heart with a hand, supports ChIP, ATAC, Bisulfite), 'Diff Analysis' (icon of a balance scale, supports ChIP, ATAC, Bisulfite), 'Target Genes' (icon of a target, supports ChIP), 'Colocalization' (icon of four arrows pointing to a center, supports ChIP), and 'Dataset Search' (icon of a magnifying glass, supports ChIP, ATAC, Bisulfite). At the bottom, a 'What's new' section contains a red warning message about server maintenance on November 3rd and a list of recent updates with their dates and links to publications.

ChIP-Atlas
A data-mining suite for exploring epigenomic landscapes by fully integrating 433,000 ChIP-seq, ATAC-seq and Bisulfite-seq experiments.

Peak Browser
ChIP ATAC Bisulfite

Enrichment Analysis
ChIP ATAC Bisulfite

Diff Analysis
ChIP ATAC Bisulfite

Target Genes
ChIP

Colocalization
ChIP

Dataset Search
ChIP ATAC Bisulfite

What's new

- All ChIP-Atlas functions will be temporarily unavailable due to the backend server maintenance: November 3rd 9am - 7th 10pm (GMT+9)
- New publication on the NAR web server issue! <https://doi.org/10.1093/nar/gkac358> (2024/05/16)
- Added Annotation tracks to Peak Browser, together with UI improvement (2023/10/25)
- Launched Diff Analysis tool enabling to detect differential peaks or differentially methylated regions (2023/10/25)
- New publication on the NAR web server issue! <https://doi.org/10.1093/nar/gkac199> (2022/03/24)
- Added ATAC-Seq and Bisulfite-Seq, together with UI improvement including 'peak' icon! (2021/10/04)

Bazy danych: ChIP Atlas

<https://chip-atlas.org/>

ChIP-Atlas

Peak Browser

Target Genes

Colocalization

Enrichment Analysis

Publications

Docs

Find By ID

SRX018625

Go

Advanced

ChIP-Atlas: Peak Browser

Tutorial movies

Visualizes protein binding on given genomic loci with IGV genome browser

H.sapiens (hg38)

H. sapiens (hg19)

M. musculus (mm10)

M. musculus (mm9)

R. norvegicus (m6)

D. melanogaster (dm6)

D. melanogaster (dm3)

C. elegans (ce11)

C. elegans (ce10)

S. cerevisiae (sacCer3)

1. Experiment type

ChIP: Histone (16612)

ChIP: RNA polymerase (2099)

ChIP: TFs and others (15155)

ChIP: Input control (8078)

ATAC-Seq (27318)

DNase-seq (1642)

Bisulfite-Seq (15000)

2. Cell type Class

All cell types (2099)

Adipocyte (6)

Blood (493)

Bone (62)

Breast (255)

Cardiovascular (49)

Digestive tract (239)

Epidermis (71)

3. Threshold for Significance

50

100

200

500

ChIP Antigen (optional)

type to search

All

GTF3C1 (1)

POLR2A (7)

POLR2B (39)

Cell type (optional)

type to search

All

Acute myeloid leukemia (2)

B cells (1)

B-CLL cells (12)

View on IGV

Error connecting to IGV?

Download BED file

Bazy danych: ChIP Atlas

<https://chip-atlas.org/>

- Peak Browser – wizualizacja wiązań czynników transkrypcyjnych i modyfikacji histonów w wybranym regionie genomu
- Enrichment Analysis – analiza wzbogacenia dla wielu genów lub regionów genomowych
- Diff Analysis – różnicowa metylacja DNA
- Target Genes - identyfikacja genów będących pod kontrolą wybranych czynników transkrypcyjnych
- Colocalization – identyfikacja potencjalnych partnerów białkowych dla czynników transkrypcyjnych

Gatunki: człowiek, mysz, szczur, muszka owocowa, *C. elegans*, *S. cerevisiae*

Bazy danych: Cistrome

<http://cistrome.org/db/#/>



Cistrome Data Browser

Tips

- Check what factors regulate your gene of interest, what factors bind in your interval or have a significant binding overlap with your peak set. Have a try at [CistromeDB Toolkit](#).
- If you have a Transcription Factor ChIP-seq (and TF perturbed expression) data, [Cistrome-GO](#) help you predict the function of this TF.
- Please help us curate the samples which has incorrect meta-data annotation by clicking the button on the inspector page. Thank you!

Containing word(s):

Species

[All](#)

Homo sapiens

Mus musculus

Biological Sources

[All](#)

1-cell pronuclei

1015c

10326

1064Sk

106A

Factors

[All](#)

AATF

ABCC9

ACSS2

ACTB

ADNP

Results

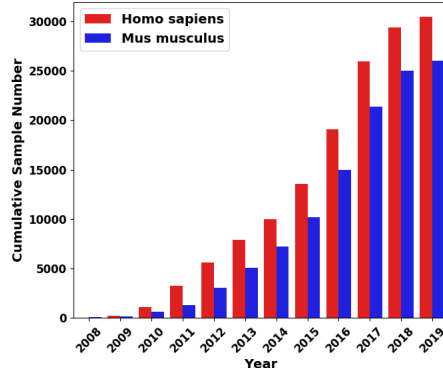
Batch	Species	Biological Source	Factor	Publication	Quality Control
☐	Homo sapiens	HeLa; Epithelium; Cervix	BTAF1	Johannes F, et al. Bioinformatics 2010	●●●●●
☐	Homo sapiens	HeLa; Epithelium; Cervix	GAPDH	Johannes F, et al. Bioinformatics 2010	●●●●●
☐	Homo sapiens	K562; Erythroblast; Bone Marrow	EGR1	Tang C, et al. Electrophoresis 2010	●●●●●

Bazy danych: Cistrome

<http://cistrome.org/db/#/>

Cistrome Data Browser Samples statistics

Species	Homo sapiens	Mus musculus	Total
Factors	13976	10089	24065
Histone Marks and Variants	13006	12394	25400
Chromatin Accessibility	3469	3530	6999
Total	30451	26013	56464



Bazy danych: NIH Roadmap Epigenomics

<https://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/>



NCBI » GEO » NIH Roadmap Epigenomics

[GEO Publications](#) [FAQ](#) [MIAME](#) [Email GEO](#)

NIH Roadmap Epigenomics Project Data Listings

This page summarizes Roadmap Epigenomics project data that are available through the GEO repository. Additional data will be added as it becomes available.

For data usage terms and conditions, please refer to [Epigenomics Data Access Policies](#)

More information and resources are available:

- Roadmap Epigenomics Homepage
- Epigenome Browser
- Epigenome Atlas

[FTP download](#)

[How to use this page](#)

Samples **Studies** **Matrix**

3256 results

Page of 66 >>> Page size

GEO Accession	Sample	Experiment	Track	Center	Download	Embargo end date
GSM409307	H1 cell line	H3K4me1	☐	UCSD	sra bam,bed,wig	2010-05-03
GSM409308	H1 cell line	H3K4me3	☐	UCSD	sra bam,bed,wig	2010-05-03
GSM409312	H1 cell line	H3K36me3	☐	UCSD	sra bam,bed,wig	2010-05-03
GSM410807	H1 cell line	H3K9ac	☐	UCSF-UBC	sra bam,bed,wig	2010-05-03
GSM410808	H1 cell line	H3K4me3	☐	UCSF-UBC	sra bam,bed,wig	2010-05-03
GSM428286	H1 cell line	MRE-Seq	☐	UCSF-UBC	sra bam,bed,wig	2010-10-07
GSM428289	H1 cell line	ChIP-Seq input	☐	UCSF-UBC	sra bam,bed,wig	2010-05-03
GSM428291	H1 cell line	H3K9me3	☐	UCSF-UBC	sra bam,bed,wig	2010-05-03
GSM428295	H1 cell line	H3K27me3	☐	UCSF-UBC	sra bam,bed,wig	2010-05-03
GSM428296	H1 cell line	H3K36me3	☐	UCSF-UBC	sra bam,bed,wig	2010-05-03

Dodatkowy link: https://egg2.wustl.edu/roadmap/web_portal/

Dodatkowy link: https://egg2.wustl.edu/roadmap/web_portal/

[illegible]

<https://www.encodeproject.org/>

ENCODE

Data

Materials & Methods

Help

Search...

REFERENCE

EPIGENOME

Project data from human tissue, cell line, primary cell, and in vitro differentiated cell biosamples organized as reference epigenomes following guidelines set out by IHEC.

All

Roadmap

Non-Roadmap

List

Report

Download

Showing 179 results.

Homo sapiens

Mus musculus

Clear all body map selections

Filter results by body diagram

ASSAY →

← BIOSAMPLE

	polys plus RNA-seq	total RNA-seq	small RNA-seq	microRNA-seq	microRNA counts	DNase-seq	ATAC-seq	WGBS	RBS	TF ChIP-seq	CTCF	POLGA	POLGAphosphoS5	EP300	Histone ChIP-seq	H3K4me3	H3K36me3	H3K4me1	H3K9me3	H3K27me3	H3K27ac	H3K9ac	H3K4me2	H2AFZ	H3K79me2	H4K20me1	H4K8ac	H3K18ac	H3K23ac	H3K4ac	H3K79me1	H2AK5ac	H2BK120ac	H2BK120c	H2BK5ac	H3K14ac	H4K91ac	H2BK15ac	H3K9me1	H3K20ac	H3K36ac	H4K5ac	H3K23me2	
heart left ventricle																																												
heart right ventricle																																												
stomach																																												
sigmoid colon																																												
spleen																																												
adrenal gland																																												
upper lobe of left lung																																												
transverse colon																																												
esophagus muscularis mucosa																																												

?

Help

ChIP-Seq: narzędzia

- **MACS / MACS2**
- FindPeaks
- PeakSeq
- SISSRS
- QuEST
- GEM
- SICER
- CCAT
- **HOMER**

HOMER



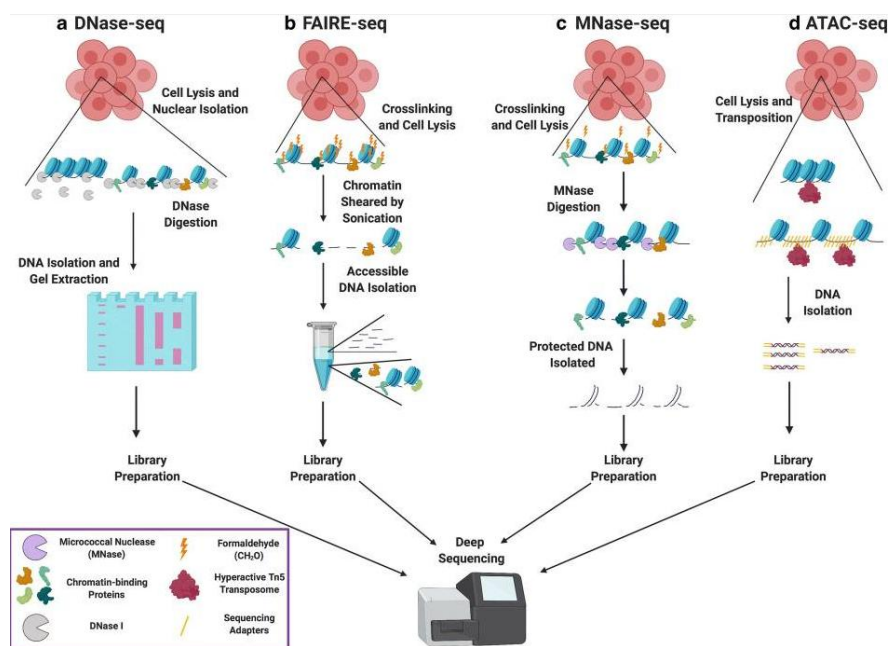
- Zestaw programów służących do:
 - Analizy danych z technologii ChIP-Seq, DNase-Seq, MNase-Seq, GRO-Seq, TSS-Seq i HiC
 - Wyszukiwania i analizy motywów sekwencyjnych
 - Adnotowania rejonów genomowych (np. rejonów wiązania białek lub modyfikacji histonów)
 - Analizy nadreprezentacji terminów Gene Ontology

<http://homer.ucsd.edu/homer/>

HOMER

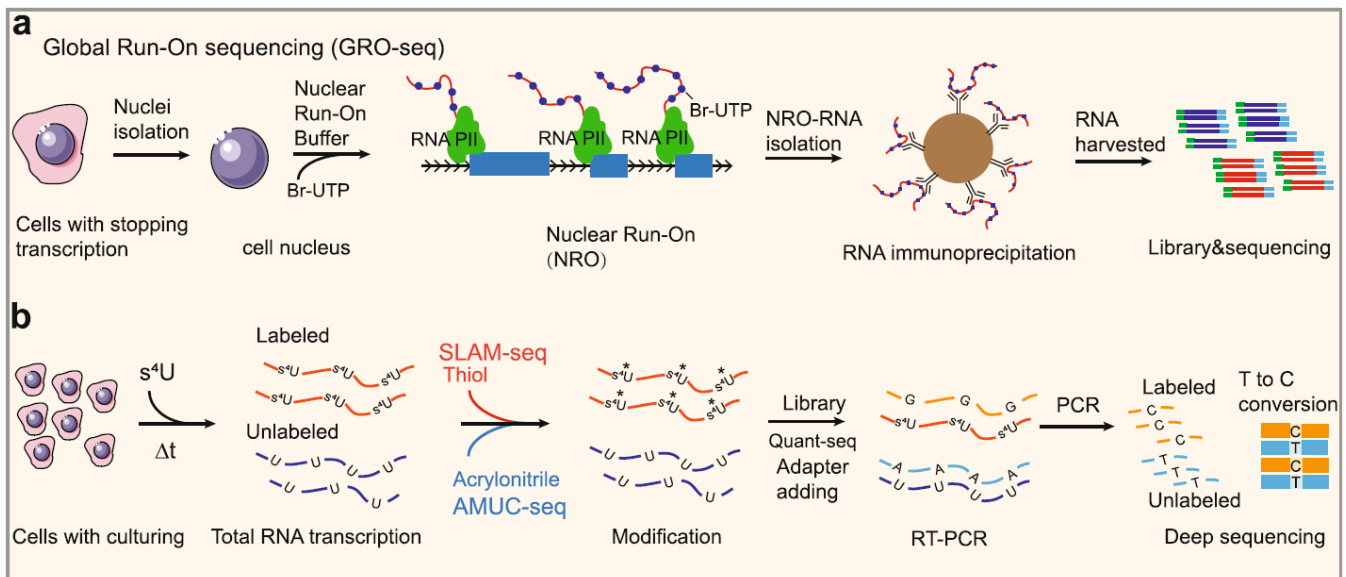
Homer de novo Motif Results									
Known Motif Enrichment Results Gene Ontology Enrichment Results If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into STAMP More information on motif finding results: HOMER Description of Results Tips Total target sequences = 37301 Total background sequences = 35962 <i>* - possible false positive</i>									
Rank	Motif	P-value	log P-value	% of Targets	% of Background	STD(Bp STD)	Best Match/Details	Motif File	
1	TGTTTACATA	1e-12661	-2.915e+04	70.91%	15.19%	40.5bp (65.1bp)	Foxa2(Forkhead)/Liver-Foxa2-ChIP-Seq/Homer More Information Similar Motifs Found	motif file (matrix)	
2	CTTGGCAG	1e-578	-1.332e+03	27.14%	16.52%	54.0bp (65.5bp)	NF1-halfsite(CTF)/LNCaP-NF1-ChIP-Seq/Homer More Information Similar Motifs Found	motif file (matrix)	
3	TTTTATTGGC	1e-384	-8.860e+02	17.77%	10.53%	53.9bp (62.1bp)	Unknown/Homeobox/Limb-p300-ChIP-Seq/Homer More Information Similar Motifs Found	motif file (matrix)	
4	CCTGTAAAT	1e-164	-3.783e+02	3.17%	1.28%	52.2bp (62.9bp)	PH0048.1_Hoxa13 More Information Similar Motifs Found	motif file (matrix)	
5	ATGACTCA	1e-151	-3.485e+02	3.38%	1.47%	50.2bp (65.4bp)	NF-E2(bZIP)/K562-NFE2-ChIP-Seq/Homer More Information Similar Motifs Found	motif file (matrix)	
6	GCCATCTGCTGG	1e-107	-2.485e+02	1.21%	0.35%	56.3bp (69.7bp)	CTCF(Zf)/CD4+-CTCF-ChIP-Seq/Homer More Information Similar Motifs Found	motif file (matrix)	
7	AGATAAGATC	1e-72	-1.671e+02	2.10%	1.02%	55.1bp (58.5bp)	MA0029.1_Evi1 More Information Similar Motifs Found	motif file (matrix)	

Badanie dostępności genomu (= stanu chromatyny)



doi: [10.1007/s10577-019-09619-9](https://doi.org/10.1007/s10577-019-09619-9)

Nascent RNA-Seq



DOI: [10.1186/s13045-020-00945-8](https://doi.org/10.1186/s13045-020-00945-8)

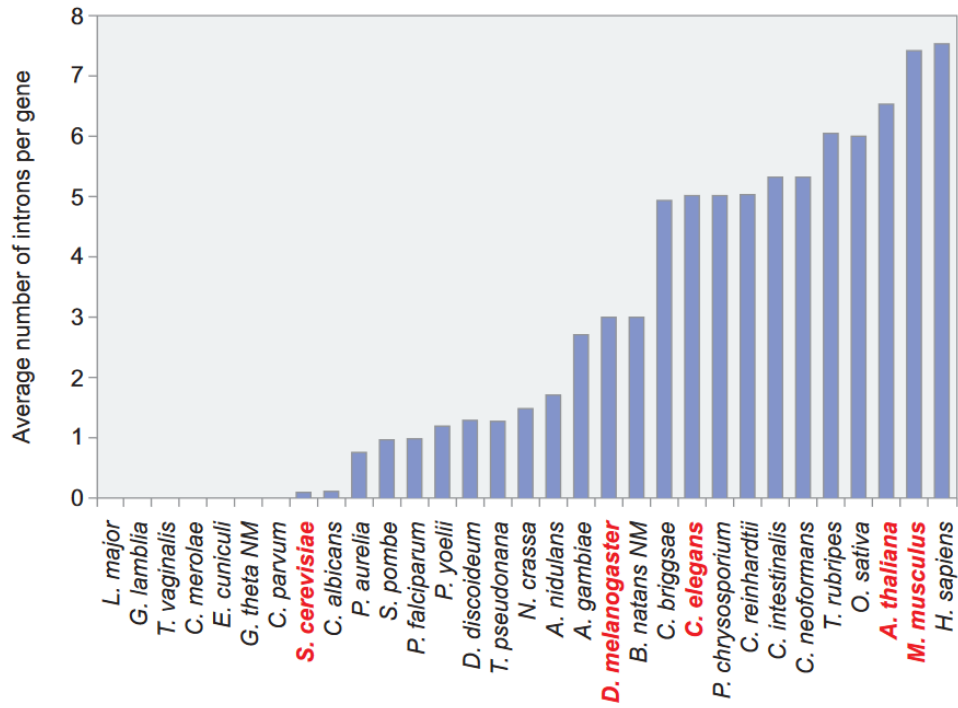
Splicing

Splicing

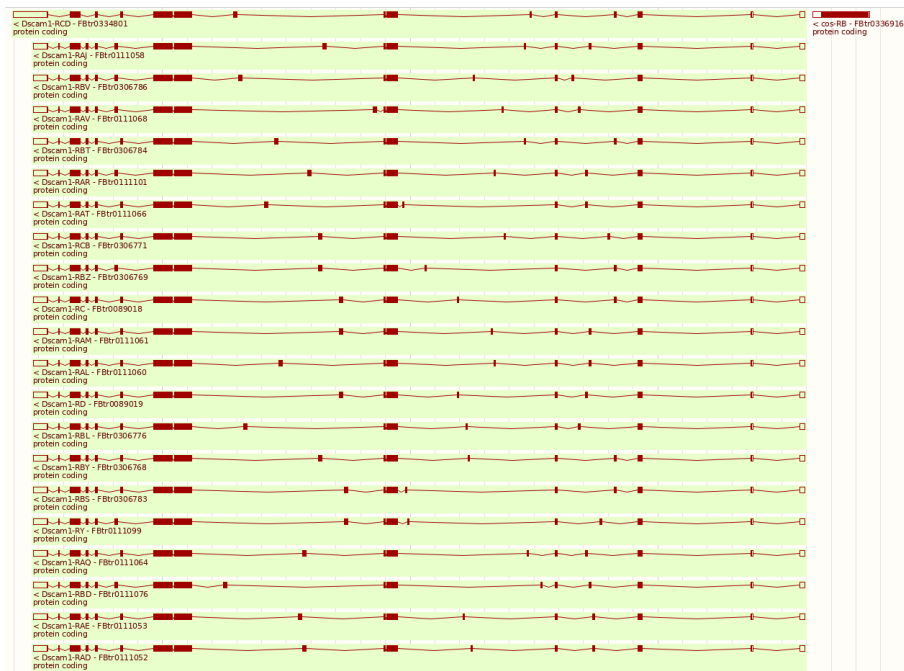
- Geny Eukaryota posiadają od zera do kilkuset intronów. Na przykład *titin* u człowieka: 363 introny
- Ogólnie, liczba intronów rośnie wraz ze wzrostem złożoności organizmu
- Również długością introny bardzo się różnią: od około 20 nt do blisko 1 mln nukleotydów
- Długość pre-mRNA ludzkiej dystrofiny: 2400 kb
- Alternatywny splicing

[illegible]

Splicing

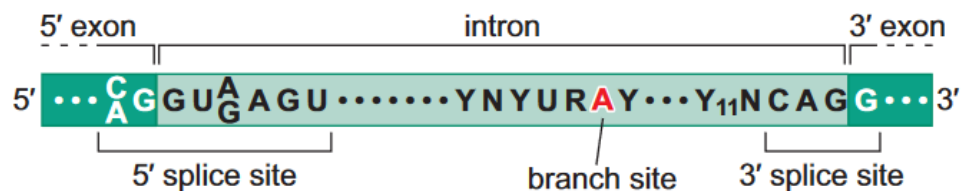


DSCAM (Down syndrome cell adhesion molecule)



Splicing

- Granice pomiędzy egzonami a intronami wyznaczają terminalne dinukleotydy na końcu 5' intronu (GT) oraz na jego końcu 3' (AG); niemniej jednak niektóre introny posiadają inne sekwencje w miejsce GT-AG
- Dodatkowo, introny posiadają tzw. miejsce rozgałęzienia (ang. *branchpoint site*), zazwyczaj zlokalizowane blisko końca 3', po którym występuje trakt polipirymidynowy (ang. *Py tract*)



Introny: przykład

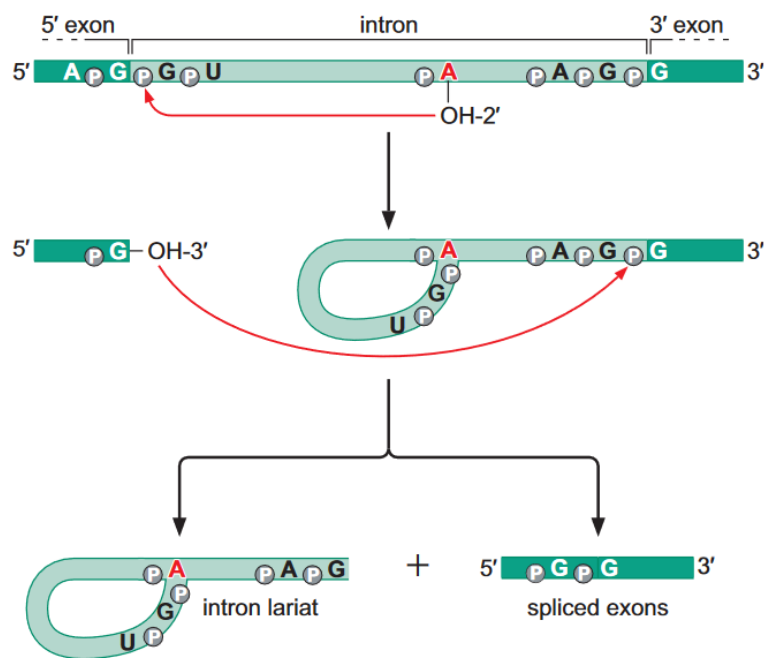
a)

	exon 1 ←		(62 bp)			polypyrimidine tract		exon 2 →
RNF113B Human	GAAG	GT	GGGGGC ...	TT	CTCAC	ATTTAAA--TAAA	TTTTTTTGGTTCTTTT	---AG GA--
RNF113B Chimp	GAAG	GT	GGGGGC ...	TT	CTCAC	ATTTAAA--TAAA	TTTTTTTGGTTCTTTT	---AG GA--
RNF113B Orangutan	GAAG	GT	GGGGGC ...	TT	CTCAC	ACTTAAAAATAAA	TGTTGTTGTTCTTTT	---AG GA--
RNF113B Rhesus	GAAG	GT	GGGAGC ...	TT	CTCAC	ATTTAAAAATAAA	TTGTGGTGGTTCTTTT	---AA GA--
RNF113B Marmoset	GAAA	GT	GGGGAA ...	TT	CTCAC	ACTTAAAAATACA	CTTTGTTGTT-TTTT	---AG GA--
RNF113A Human	GAAG	AG	GGTGGT ...	TT	CTTAA	ATTTAAAAAATAA	ACGTTTTGTTCTTTT	---GG AAAA
RNF113A Mouse	GCAG	AG	GGTGGT ...	TT	CTCAG	-TCTCAAA-ATAA	ATGTTTTGTTGTTTC	---TT AAAA
RNF113A Cattle	GAAG	GA	GGTGGT ...	TT	CTCAA	ATCTCAAA-ATAA	ACGTTTTGTTTGGAAATTTT	AAAA
RNF113A Dog	GAAG	AG	GGTGGT ...	TT	CTCAC	ACCTCAAA-ATA-	AATCTTTGTTTGGAA-ATCTT	AATT

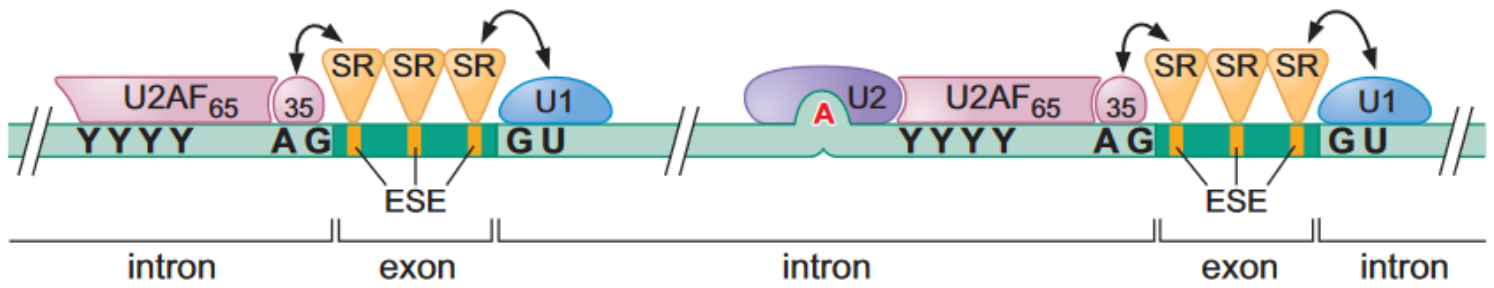
Intron terminal dinucleotides
 branch site
 Polypyrimidine tract

<https://doi.org/10.1093/molbev/msq260>

Splicing

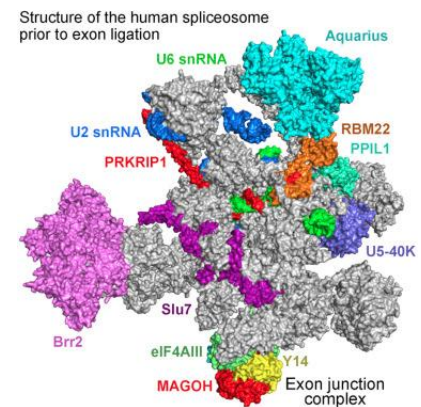


Splicing



Spliceosom

- Reakcje transestryfikacji są pośredniczone przez molekularną „maszynę” nazwaną spliceosomem, składającą się z około 150 białek i pięciu cząsteczek RNA
- Większość funkcji spliceosomu pełnią cząsteczki snRNA (U1, U2, U4, U5, and U6): to one rozpoznają granice egzon-intron i katalizują sam splicing
- snRNA mają długość w zakresie 100-300 nt i wchodzi w interakcje z białkami, tworząc tzw. snRNPs („snurps”, ang. *small nuclear ribonuclear proteins*)
- Jednym z głównych czynników białkowych, niewchodzącym w skład snRNPs, jest U2AF, rozpoznający trakt polipirymidynowy, a także BBP: *branchpoint binding protein*
- W procesie splicingu, spliceosom zużywa energię (ATP)



DOI: <https://doi.org/10.1016/j.cell.2017.04.033>

Spliceosom: U12

atgc highly conserved terminal intronic nucleotides at 5' or 3' end.

atgc branch site.

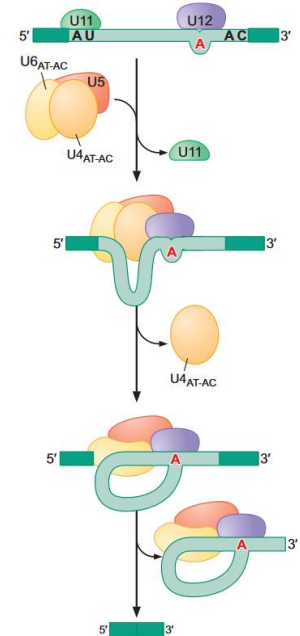
#1
gene: **AT1G02750**, transcript: **AT1G02750.1**
Orthologs: GLYMA13G0150 GLYMA03G37530 LOC_0505g28980 GRMZM5G840013
ATCATAGAGATGCTTGAA**Gatccctt**ttgtttcttttcaaaaccttcaagtc aaattccgtcgtataatgaatgtgatttattgaa**ttccttgac**cgaataatttt**ac**AGTGAGCAGAAGGAAATGTC

#2
gene: **AT1G04130**, transcript: **AT1G04130.1**
Orthologs: GLYMA13G18990 GLYMA10G04660 LOC_0503g53910 PP15301_8V6 Vv07s0005g04590 GRMZM2G146000
TTGCAACTCATCTTGATAT**Gatccctt**ttgccagtttccctttgtttcttattcttgcatctcatttg ... cgtctatagactgttcaatgcgctttatgtcc**ctccttagc**gctctaa**ac**ATGTTTTCAGAGGACAGCCC

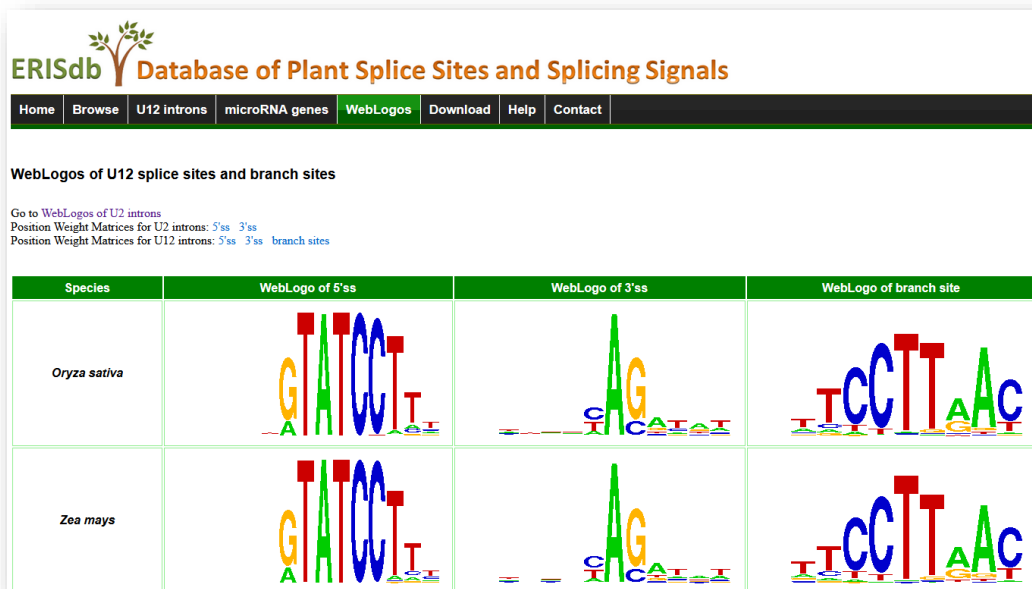
#3
gene: **AT1G10417**, transcript: **AT1G10417.1**
Orthologs: LOC_0510g27180 GRMZM2G390432
ATCAACATCAAGTCGAAG**Gatccctt**atcagaacttgcgtgtcttactgcattttcaagaactagtgaggagatcatttccacgtatatgt**ttacttgac**cagcctt**ac**CTTCAGAAAACTTGCAGGT

#4
gene: **AT1G24050**, transcript: **AT1G24050.1**
Orthologs: LOC_0502g50830 PP1538_297V6 SELMODRAFT_270872 GRMZM2G503738
CCCCAATTCGCTTTT**Gatccctt**aatcttcttaattgtttgtttcatcaaaattgagatttag ... caatgtcttgacctgaattgtgt**atccctga**gagttattgttatt**ac**AAGAGGGTACAAAACCTAGA

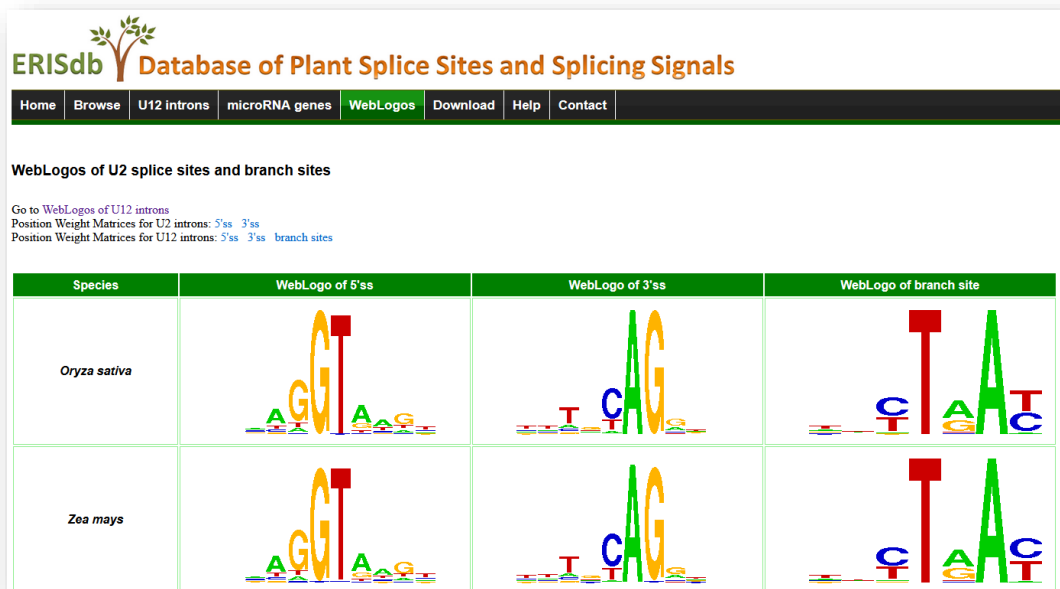
#5
gene: **AT1G26170**, transcript: **AT1G26170.1**
Orthologs: GLYMA02G14280 PP1555_322V6 GRMZM2G457415
TCTCTTTGGGATTGGCTCAA**atatacctt**ttcatgtcattgattgtatccctggagcttgatttaga ... atgagtgtttcccttttttggctgtga**gttcttaac**ctactgtt**ac**TTAGCTGCGGTACTGTAA



Spliceosom: U12



Spliceosom: U2

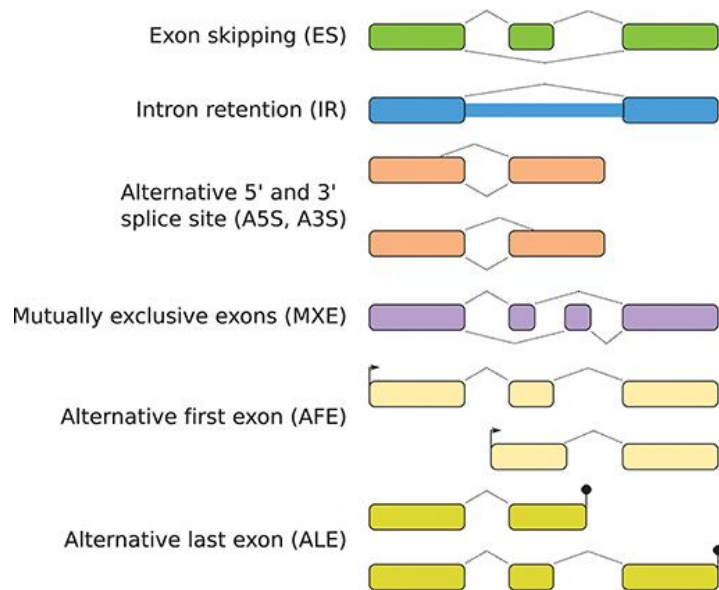


Splicing

TABLE 14-1 Three Classes of RNA Splicing

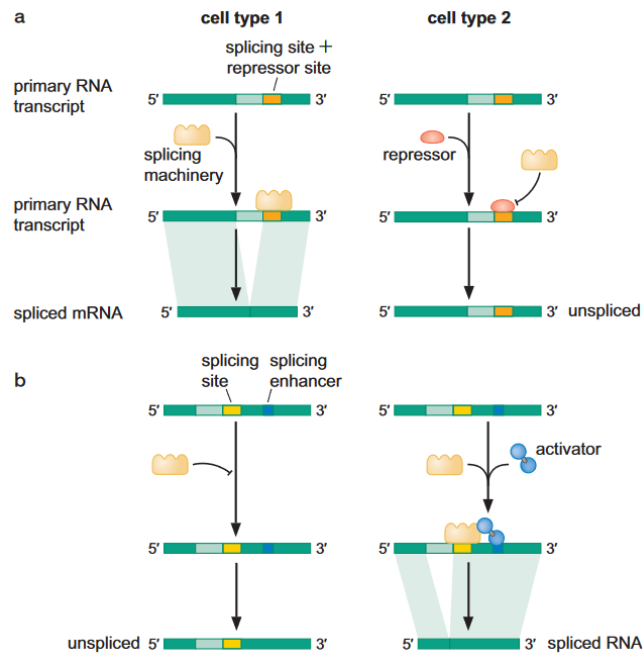
Class	Abundance	Mechanism	Catalytic Machinery
Nuclear pre-mRNA	Very common; used for most eukaryotic genes	Two transesterification reactions; branch site A	Major and minor spliceosomes
Group II introns	Rare; some eukaryotic genes from organelles and prokaryotes	Same as pre-mRNA	RNA enzyme encoded by intron (ribozyme)
Group I introns	Rare; nuclear rRNA in some eukaryotes, organelle genes, and a few prokaryotic genes	Two transesterification reactions; branch site G	Same as group II

Alternatywny splicing



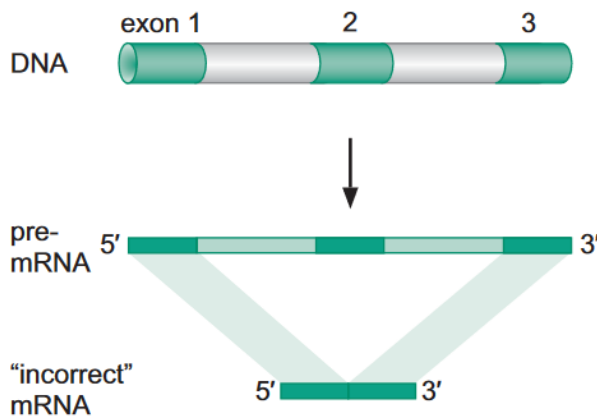
<https://doi.org/10.3389/fmicb.2018.02313>

Alternatywny splicing

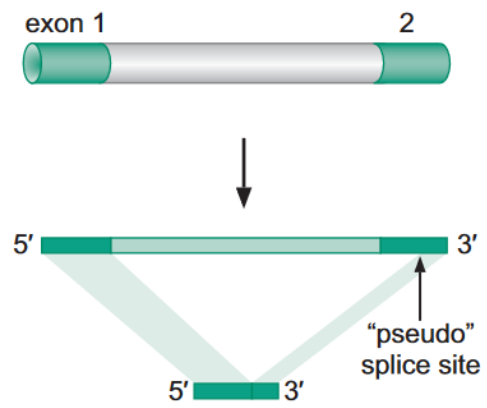


Splicing: błędy

a exon skipping



b pseudo splice-site selection

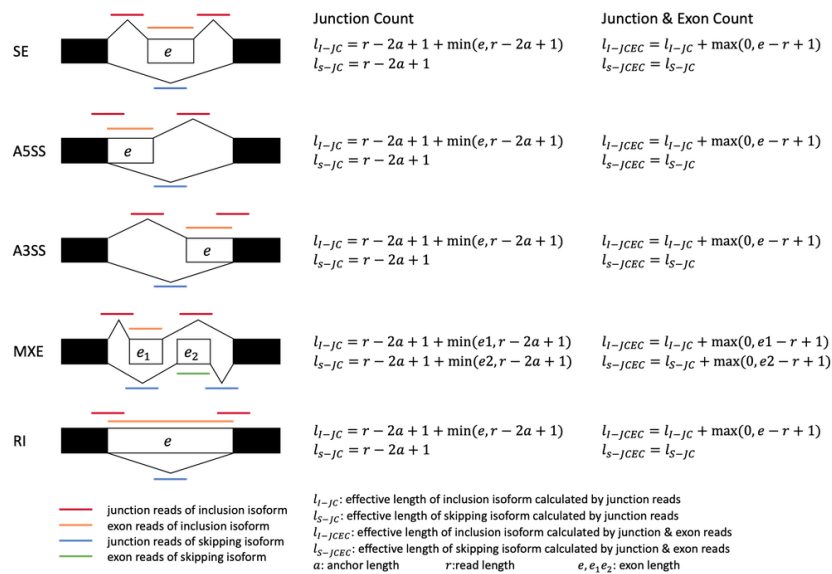


Splicing: błędy

- Ocena się, że 15% mutacji punktowych wywołujących choroby u człowieka tak naprawdę wpływa na splicing. Klasycznym przykładem jest beta-talasemia, charakteryzująca się defektywną produkcją beta-globiny, podjednostki hemoglobiny. Jedną z form tej choroby wywołaną jest przez mutację punktową w pierwszym intronie genu beta-globiny, która zmienia sekwencję TTGGT na TTAGT. Wskutek tego powstaje fałszywe miejsce splicingowe 3'
- Również *familial isolated growth hormone deficiency type II* jest chorobą wywołaną przez defektywny splicing, tym razem w genie kodującym hormon wzrostu
- Frasier syndrome: zaburzony splicing pre-mRNA genu uczestniczącego w rozwoju nerek i gonad.
- Retinis pigmentosa (degeneracja siatkówki, ślepotą): mutacje w genach, których białka wchodzi w skład spliceosomu
- Nowotwory (np. mutacje w genie p73: *squamous cell carcinoma*)

Narzędzia

rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data



<https://www.pnas.org/doi/10.1073/pnas.1419161111>

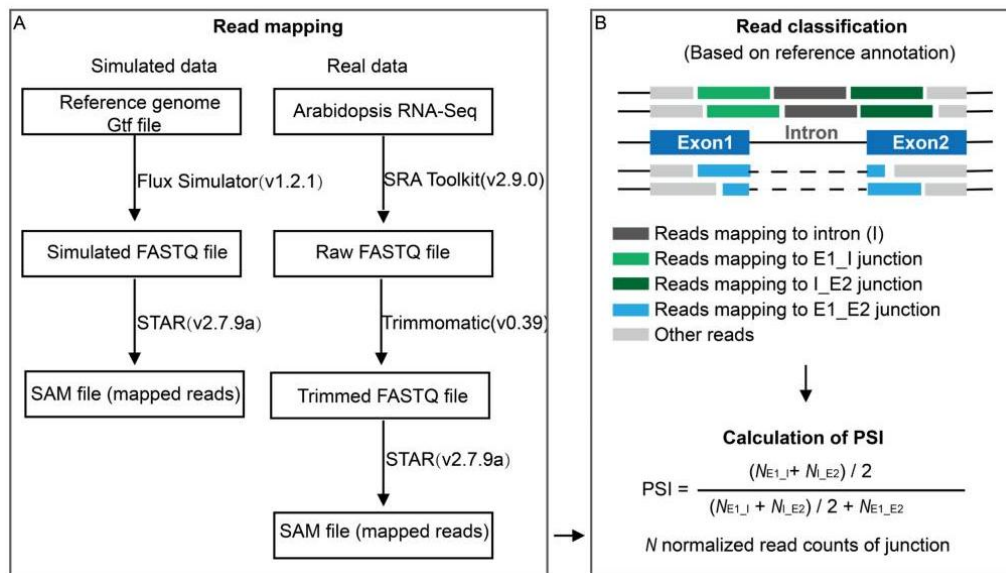
spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data

	Event type	No. of events
	Exon skipping/inclusion (ESI)	1,619
	Mult. exon skipping/inclusion (MESI)	190
	Intron skipping/inclusion ISI	256
	Alternative 5' splice site A5	755
	Alternative 3' splice site A3	733
	Alternative transcription start site ATSS	1,381
	Alternative transcription termination site ATTS	1,125
	Mutually exclusive exons MEE	22
	All events	6,121

Number of individual alternative splicing events identified. A schematic structure of each alternative splicing type, along with the associated name, abbreviation and the number of classified events in Usp49 KD RNA seq data.

<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-81>

ASTool: An Easy-to-Use Tool to Accurately Identify Alternative Splicing Events from Plant RNA-Seq Data



<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9031537/>

Bioconductor: splicing

<https://bioconductor.org/packages/release/bioc/html/ASpli.html>

Integrative pipeline for the analysis of alternative splicing using RNAseq.

<https://bioconductor.org/packages/release/bioc/html/BANDITS.html>

BANDITS is a Bayesian hierarchical model for detecting differential splicing of genes and transcripts, via differential transcript usage (DTU), between two or more conditions. The method uses a Bayesian hierarchical framework, which allows for sample specific proportions in a Dirichlet-Multinomial model, and samples the allocation of fragments to the transcripts. Parameters are inferred via Markov chain Monte Carlo (MCMC) techniques and a DTU test is performed via a multivariate Wald test on the posterior densities for the average relative abundance of transcripts.

Bioconductor: splicing

<https://bioconductor.org/packages/release/bioc/html/GeneStructureTools.html>

GeneStructureTools can be used to create in silico alternative splicing events, and analyse potential effects this has on functional gene products.

<https://bioconductor.org/packages/release/bioc/html/IsoformSwitchAnalyzeR.html>

Analysis of alternative splicing and isoform switches with predicted functional consequences (e.g. gain/loss of protein domains etc.) from quantification of all types of RNASeq by tools such as Kallisto, Salmon, StringTie, Cufflinks/Cuffdiff etc.

Bioconductor: splicing

<https://bioconductor.org/packages/release/bioc/html/regsplICE.html>

Statistical methods for detection of differential splicing (differential exon usage) in RNA-seq and exon microarray data, using L1-regularization (lasso) to improve power.

<https://bioconductor.org/packages/release/bioc/html/SPLINTER.html>

Provides tools to analyze alternative splicing sites, interpret outcomes based on sequence information, select and design primers for site validation and give visual representation of the event to guide downstream experiments.

Jeśli nie wskazano inaczej, część ilustracji pochodzi z:

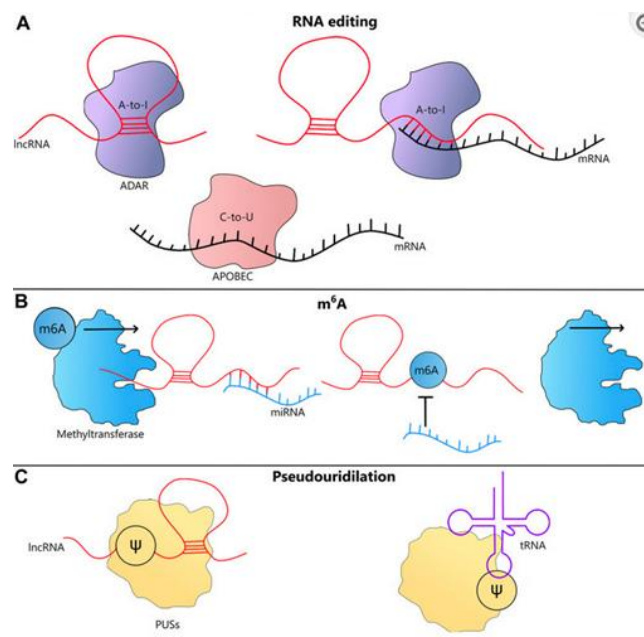
Molecular Biology of the Gene, 7th edition

James D. Watson, Tania A. Baker, Stephen P. Bell, Alexander Gann,
Michael Levine, Richard Losick

Pearson 2013

RNA

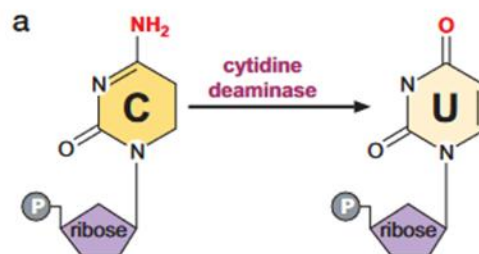
Modyfikacje RNA



doi: [10.3389/fcell.2022.922351](https://doi.org/10.3389/fcell.2022.922351)

Edytowanie RNA

- Dwa rodzaje: deaminacja cytozyn lub adenin oraz insercje/delecje urydyn
- Deaminowana cytozyna jest konwertowana do urydyny – proces ten zachodzi rzadko, w ograniczonej puli tkanek. Przykład: Apolipoprotein B
- Proces przeprowadzany jest przez enzymy APOBEC (C>U)



Edytowanie RNA

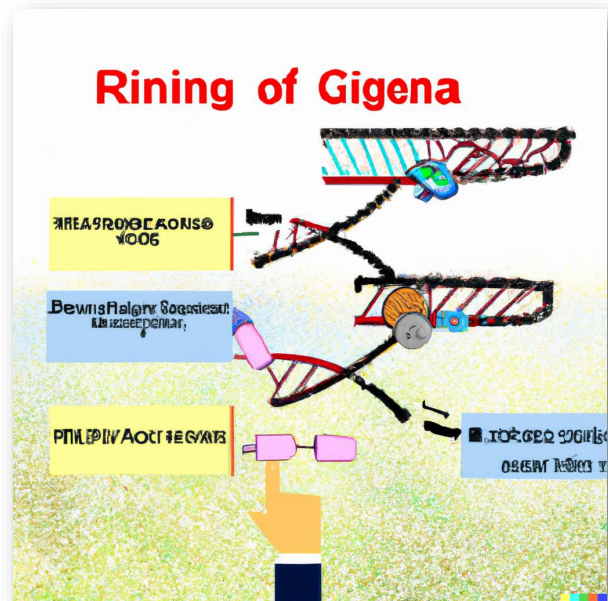
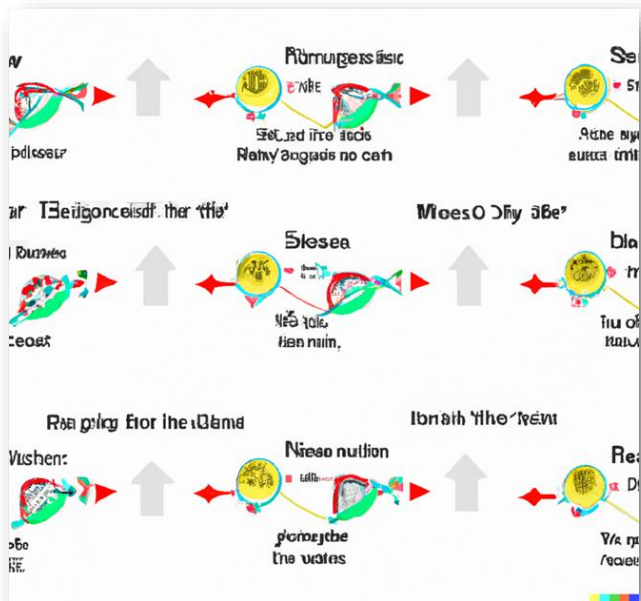


Edytowanie RNA

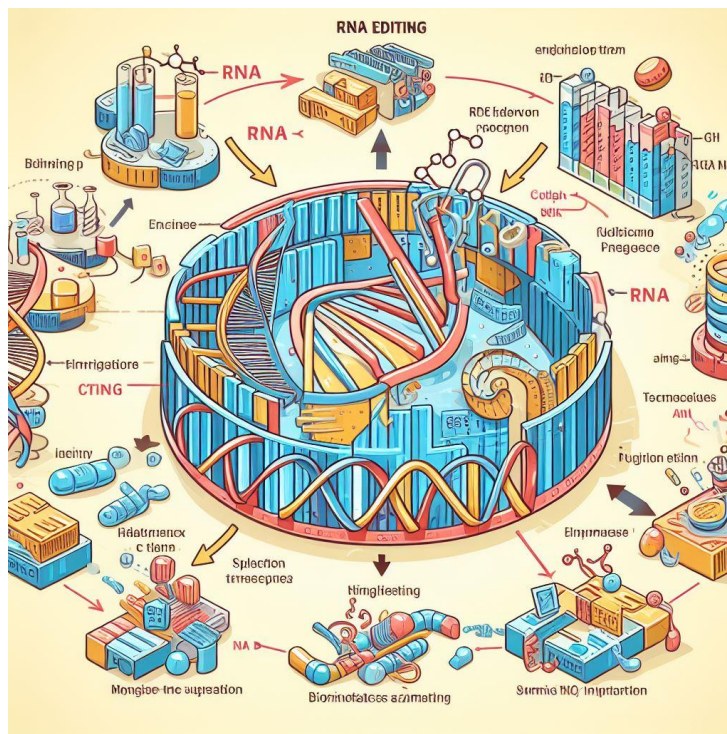
- Drugi rodzaj edytowania to deaminacja adenozy do inozyny (A>I), przeprowadzana przez enzymy ADAR. Inozyna paruje się z cytozyną, a więc inozyna traktowana jest jak guanina
- Ten rodzaj edytowania wykrywany jest w większości pre-mRNA człowieka
- Większość zdarzeń edytowania A>I zdarza się w elementach powtarzalnych Alu, zwłaszcza w regionach 5' UTR, 3'UTR oraz w intronach
- Enzymy ADAR rozpoznają sparowane odcinki RNA (dsRNA)
- Zjawisko powiązane z licznymi chorobami u człowieka; zmiany nie tyle jakościowe, co ilościowe
- Edytowanie a zmiany somatyczne



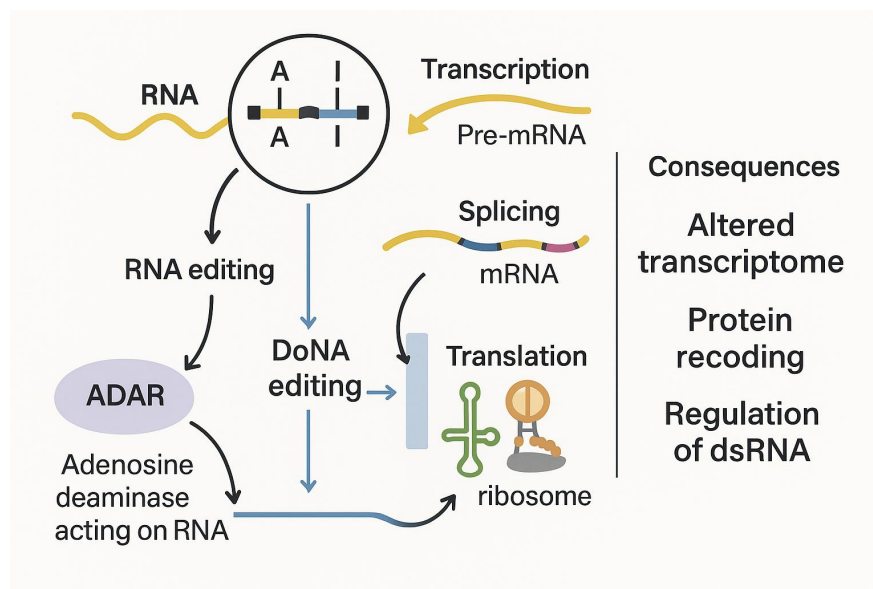
Edytowanie RNA: mechanizmy molekularne



Edytowanie RNA: mechanizmy molekularne



Edytowanie RNA: mechanizmy molekularne



Transport mRNA

- Po pełnej obróbce, mRNA jest transportowane do cytoplazmy, gdzie jest matrycą w procesie translacji
- Jest to proces regulowany: transkrypty poddane błędnej obróbce (np. z intronami) lub nieposiadające ogona poliA pozostają w jądrze
- Udział licznych białek, m.in. SR

RNA

Translacja

- W czasie translacji cząsteczka dojrzałego RNA służy jako matryca do syntezy łańcucha polipeptydowego. Odbywa się w cytoplazmie na rybosomach.
- Rybosomy składają się z dwóch podjednostek (dużej i małej), każda zawierająca rRNA, jak i białka rybosomalne.

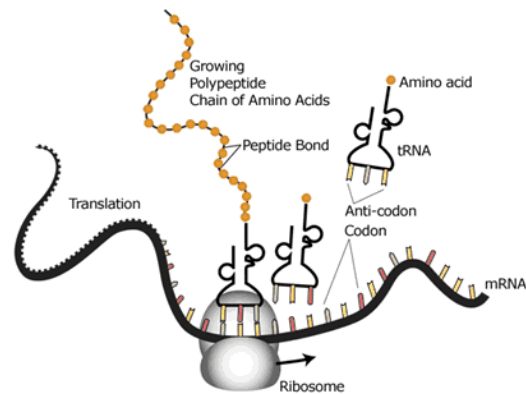
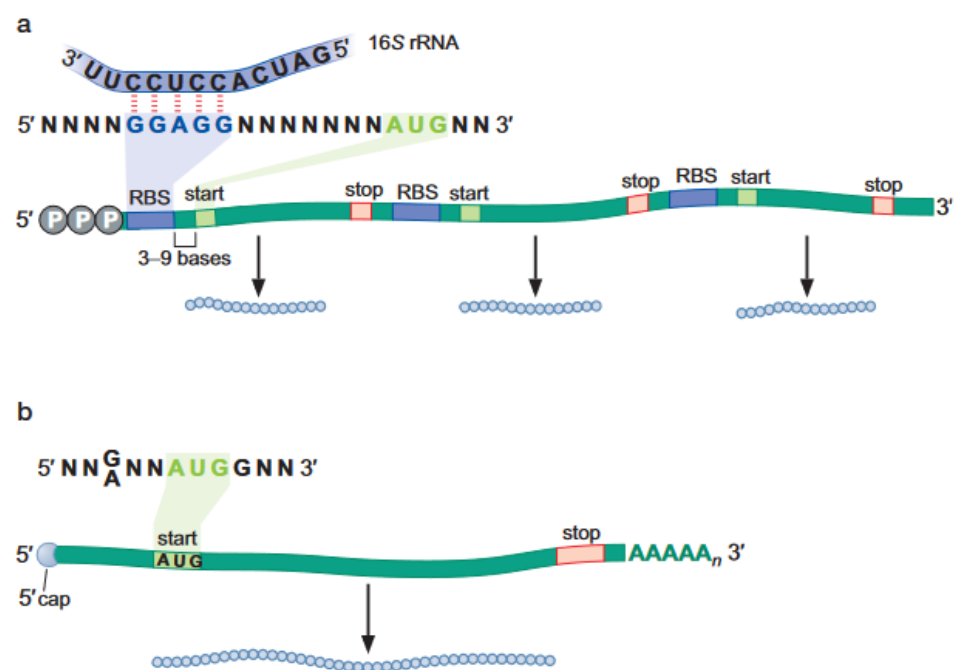


Image adapted from: National Human Genome Research Institute.

Translacja



Translacja

Cztery etapy translacji:

- **Inicjacja:** mała podjednostka rybosomu wiąże się z końcem 5' cząsteczki mRNA i przesuwa w kierunku końca 3', aż napotka kodon start (AUG). Wówczas tworzy kompleks z dużą podjednostką i inicjacyjną cząsteczką tRNA
- **Elongacja:** Poszczególne kodony w cząsteczce mRNA determinują która cząsteczka tRNA przyłączy się do mRNA, a zatem określa kolejność aminokwasów w wydłużanym łańcuchu polipeptydowym.
- **Terminacja:** Translacja jest zakończona gdy kompleks rybosomowy dotrze do kodonu stop (UAA, UAG lub UGA)
- **Potranslacyjna obróbka** białka, dojrzewanie białka

Odczytywanie informacji genetycznej: kod genetyczny

		Second letter					
		U	C	A	G		
First letter	U	UUU Phenyl-alanine UUC UUA Leucine UUG	UCU Serine UCC UCA UCG	UAU Tyrosine UAC UAA Stop codon UAG Stop codon	UGU Cysteine UGC UGA Stop codon UGG Tryptophan	Third letter	U C A G
	C	CUU Leucine CUC CUA CUG	CCU Proline CCC CCA CCG	CAU Histidine CAC CAA Glutamine CAG	CGU Arginine CGC CGA CGG		U C A G
	A	AUU Isoleucine AUC AUA AUG Methionine; start codon	ACU Threonine ACC ACA ACG	AAU Asparagine AAC AAA Lysine AAG	AGU Serine AGC AGA Arginine AGG		U C A G
	G	GUU Valine GUC GUA GUG	GCU Alanine GCC GCA GCG	GAU Aspartic acid GAC GAA Glutamic acid GAG	GGU Glycine GGC GGA GGG		U C A G

<https://readbiology.com/genetic-code/>

Kody genetyczne

<https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?chapter=tgencodes#SG1>

The Genetic Codes

Compiled by Andrzej (Anjay) Elzanowski and Jim Ostell at National Center for Biotechnology Information (NCBI), Bethesda, Maryland, U.S.A.

Last update of the Genetic Codes: Jan. 7, 2019

NCBI takes great care to ensure that the translation for each coding sequence (CDS) present in GenBank records is correct. Central to this effort is careful checking on the taxonomy of each record and assignment of the correct genetic code (shown as a transl_table qualifier on the CDS in the flat files) for each organism and record. This page summarizes and references this work.

The synopsis presented below is based primarily on the reviews by [Osawa et al.](#) (1992) and [Jukes and Osawa](#) (1993). Listed in square brackets [] (under **Systematic Range**) are tentative assignments of a particular code based on sequence homology and/or phylogenetic relationships.

The print-form ASN.1 version of this document, which includes all the genetic codes outlined below, is also available [here](#). Detailed information on codon usage can be found at the [Codon Usage Database](#).

GenBank format by historical convention displays mRNA sequences using the DNA alphabet. Thus, for the convenience of people reading GenBank records, the genetic code tables shown here use T instead of U. The initiator codon - whether it is AUG, CTG, TTG or something else, - is by default translated as methionine (Met, M). The possible initiator codons are marked as 'M' in the second (Starts) row of the translation tables.

Currently, genetic codes can be set independently for nucleus, mitochondria, plastids and hydrogenosomes. The current settings for each of these on the taxonomic tree can be viewed by the four buttons directly underneath the following code list.

The following genetic codes are described here:

- 1. The Standard Code
- 2. The Vertebrate Mitochondrial Code
- 3. The Yeast Mitochondrial Code
- 4. The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code
- 5. The Invertebrate Mitochondrial Code
- 6. The Ciliate, Dinoflagellate, and Hexamita Nuclear Code
- 9. The Echinoderm and Flatworm Mitochondrial Code
- 10. The Euglenid Nuclear Code
- 11. The Bacterial, Archaeal and Plant Plastid Code
- 12. The Alternative Yeast Nuclear Code
- 13. The Ascidian Mitochondrial Code
- 14. The Alternative Flatworm Mitochondrial Code
- 16. Chlorophycean Mitochondrial Code
- 21. Trematode Mitochondrial Code
- 22. *Scenedesmus obliquus* Mitochondrial Code
- 23. *Thraustochytrium* Mitochondrial Code
- 24. Rhabdopleuridae Mitochondrial Code
- 25. Candidate Division SRI and Gracilibacteria Code
- 26. *Pachysolen tansophilus* Nuclear Code
- 27. *Karorelect* Nuclear Code
- 28. *Condylostoma* Nuclear Code
- 29. Mesodinium Nuclear Code
- 30. Peritrich Nuclear Code
- 31. Blastocrithidia Nuclear Code
- 33. Cephalodiscidae Mitochondrial UAA-Tyr Code

Genetic Code in Taxonomy Tree

Mitochondrial Genetic Code in Taxonomy Tree

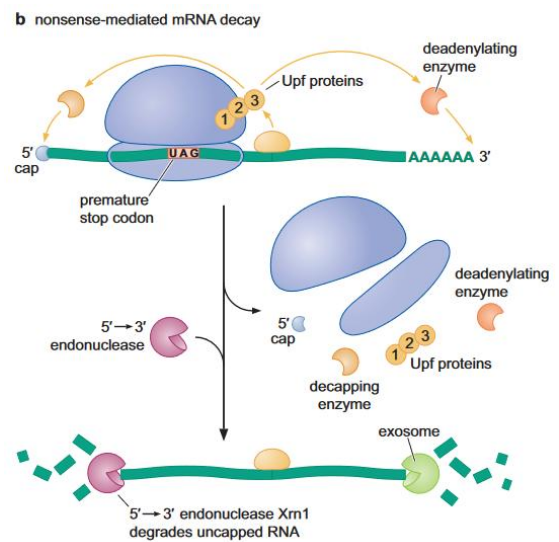
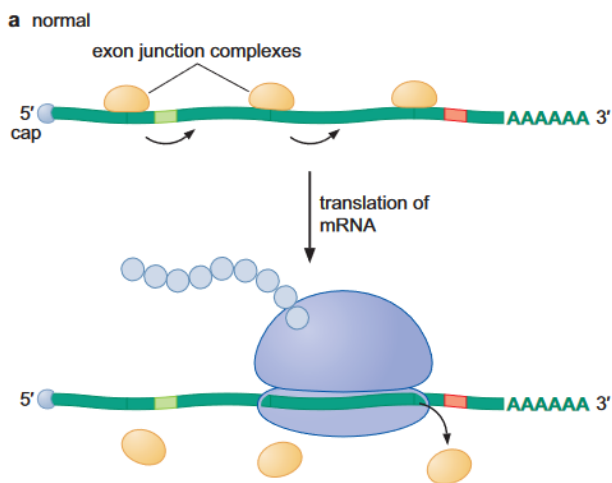
Plastid Genetic Code in Taxonomy Tree

Hydrogenosome Code in Taxonomy Tree

Odczytywanie informacji genetycznej: cechy kodu genetycznego

- **Trójkowy**
Trzy leżące obok siebie nukleotydy tworzą kodon = triplet. Kodon to podstawowa jednostka kodująca aminokwas. Np. kodon UUU oznacza Phe - fenyloalaninę, a kodon GCA oznacza Ala - alaninę.
- **Zdegenerowany**
64 kodony kodują tylko 20 aminokwasów: jeden aminokwas może być kodowany przez kilka kodonów. Nie oznacza to, że dany kodon może kodować więcej niż jeden aminokwas – zobacz poniżej
- **Jednoznaczny**
Kod genetyczny jest jednoznaczny, czyli jedna trójka może kodować tylko jeden aminokwas.
- **Bezprzecinkowy**
Pomiędzy kodonami mRNA nie występują żadne przerwy w zapisie. Rybosom nie pomija żadnego kodonu odczytując mRNA w translacji. Oznacza to również, że nie istnieje kodon oznaczający przerwę w kodzie.
- **Niezachodzący**
Rybosom nie cofa się do poprzedniego nukleotydu odczytując kodony mRNA. Informacja zakodowana jest w kolejnych trójkach nukleotydów i nie występują sytuacją, w której jeden z nukleotydów jest częścią dwóch sąsiednich kodonów.
- **Uniwersalny**
Budowa kodu genetycznego u wszystkich organizmów jest oparta na tych samych zasadach. Kodony w większości oznaczają te same aminokwasy.
- **Kolinearny**
Kolejność ułożenia danych aminokwasów w białku jest wiernym odzwierciedleniem ułożenia odpowiednich kodonów na mRNA (matrycowym RNA).

Nonsense-mediated mRNA Decay (NMD)



Regulacja ekspresji genów

Regulacja ekspresji genów

Regulacja ekspresji genów to ogół procesów które decydują o tym, jaki jest poziom ekspresji poszczególnych transkryptów i białek w komórce. Są to złożone mechanizmy, w których udział biorą wszystkie cząsteczki biologiczne i struktury komórkowe

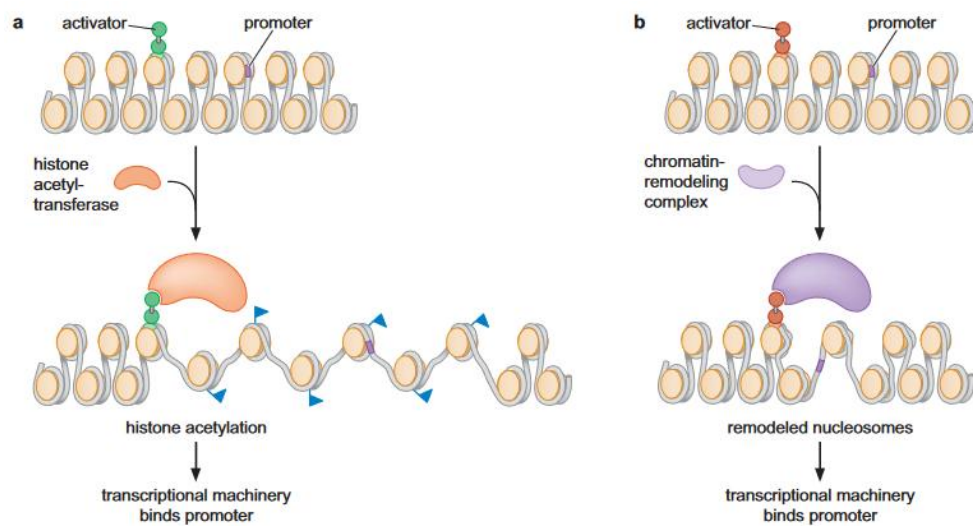
Niektóre geny ulegają ekspresji w sposób ciągły i występują we wszystkich typach komórek – są one związane z podstawowymi procesami, takimi jak oddychanie komórkowe.

Wiele genów ulega ekspresji tylko w niektórych tkankach, organach czy też tylko na pewnych etapach rozwoju osobniczego (różnica może być nie tylko jakościowa, ale też ilościowa, tzn. polegać na zróżnicowanym poziomie ekspresji genu).

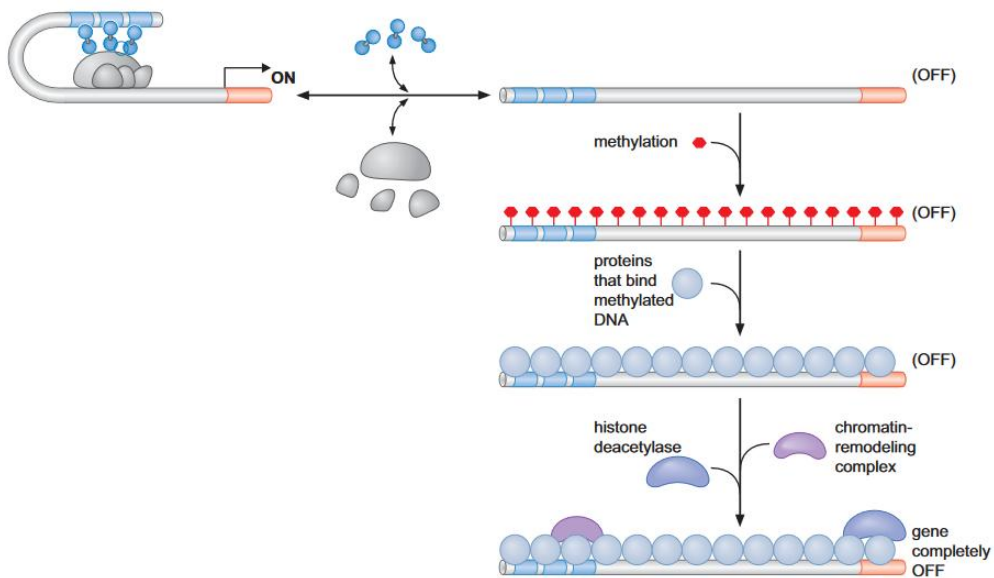
Istnieje wiele mechanizmów regulacji ekspresji genów. Można do nich zaliczyć:

- Regulacja transkrypcji (najbardziej ekonomiczna)
- Regulacja obróbki cząsteczek RNA, np. alternatywny splicing
- Regulacja stabilności cząsteczek RNA (np. miRNA)
- Regulacja translacji
- Obróbka i dojrzewanie białek

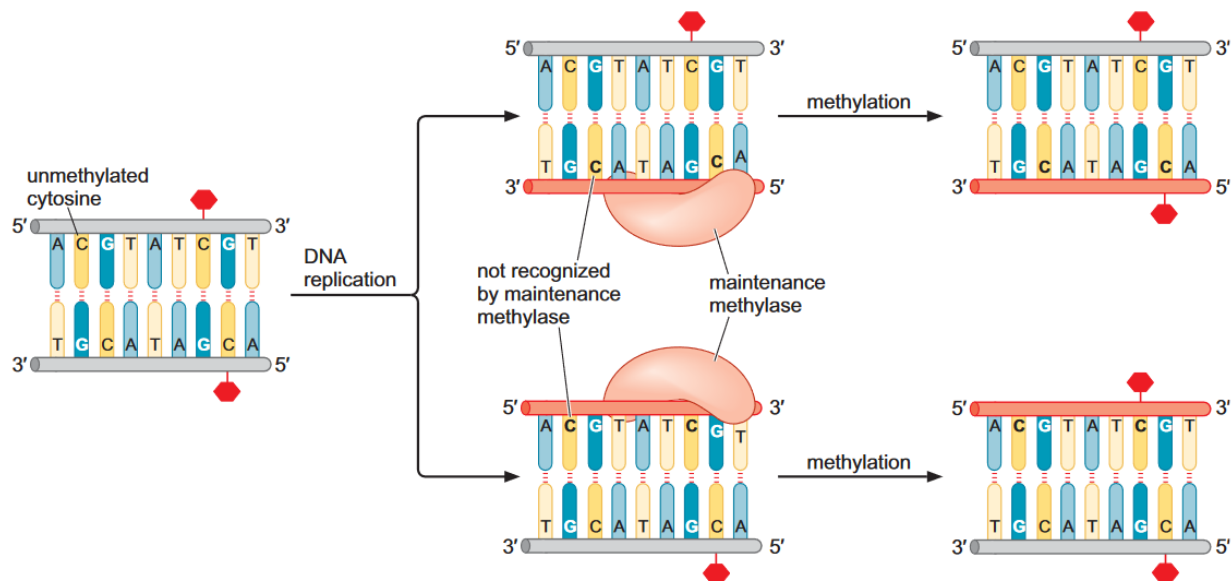
Regulacja transkrypcji



Regulacja transkrypcji: metylacja DNA



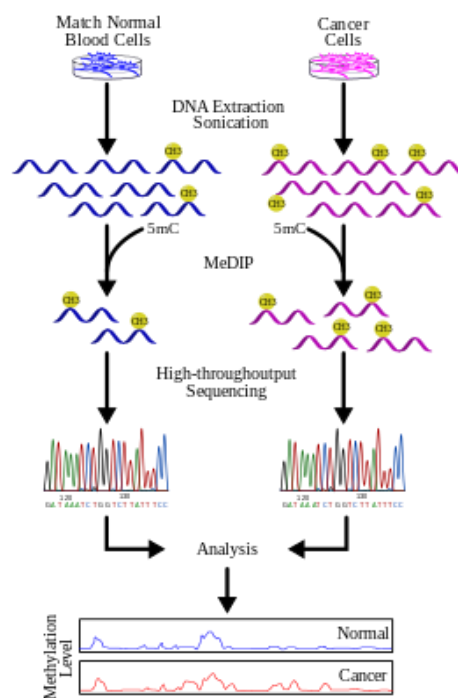
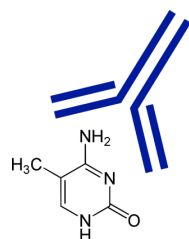
Regulacja transkrypcji: metylacja DNA



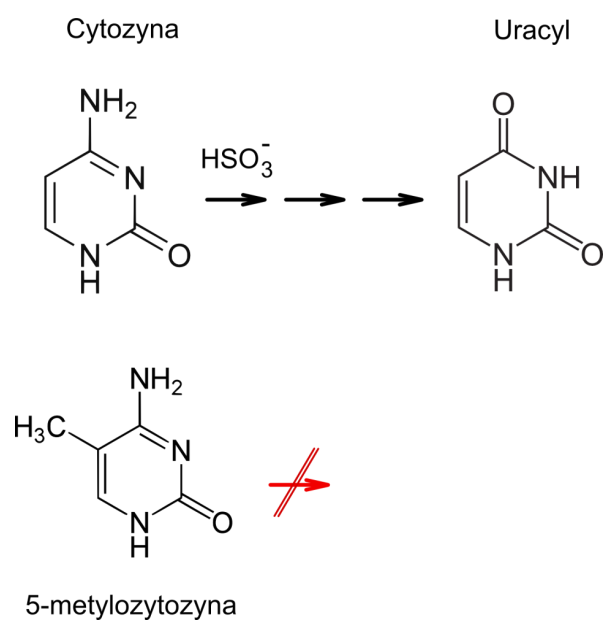
Badanie metylacji DNA

Immunoprecypitacja
metylowanego DNA

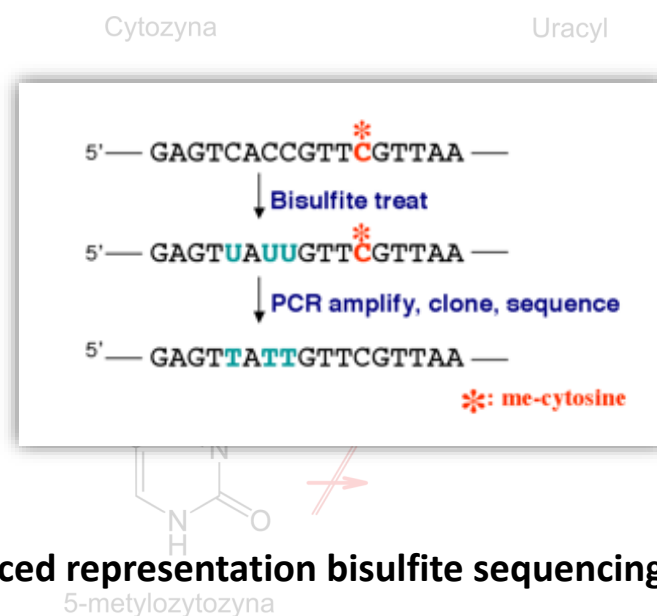
(MeDIP sequencing)



Badanie metylacji DNA BS-seq (bisulfite sequencing)



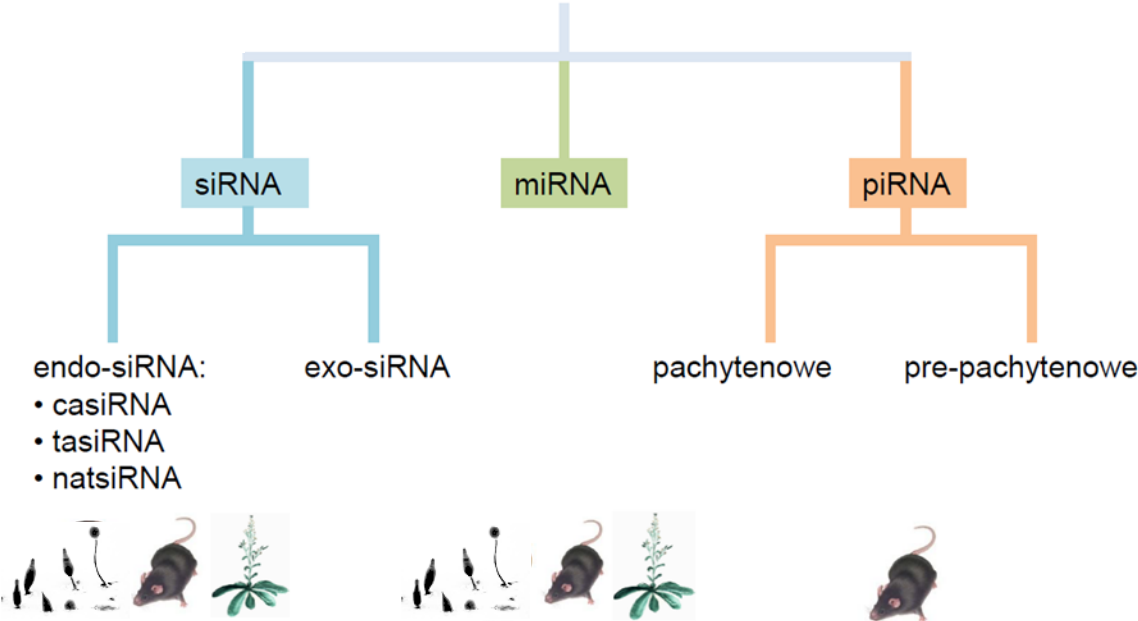
Badanie metylacji DNA BS-seq (bisulfite sequencing)

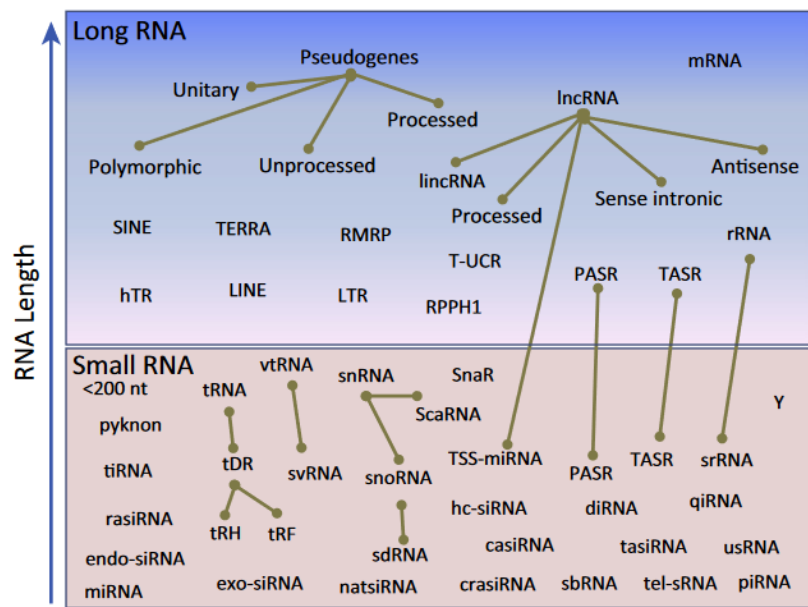


Reduced representation bisulfite sequencing (RRBS)

<http://alphabiolab.com/>

małe regulatorowe RNA





T/BS

Źródło: Vickers KC, Roteta LA, Hucheson-Dilks H, Han L, Guo Y. Mining diverse small RNA species in the deep transcriptome. Trends Biochem Sci. 2015 Jan;40(1):4-7.

sRNA pochodzące z długich, niekodujących prekursorów

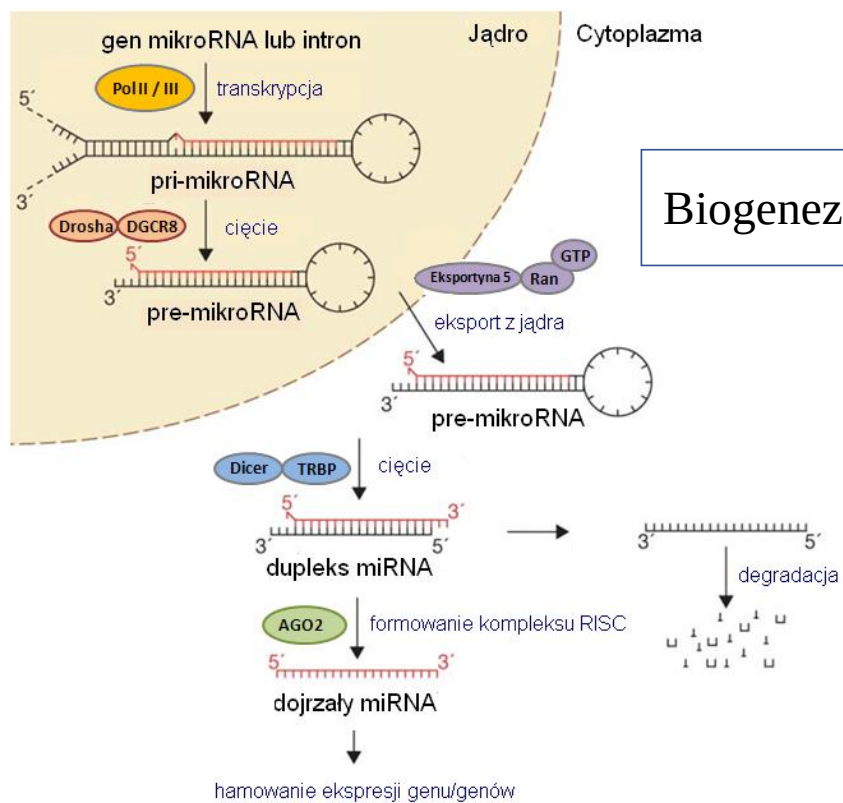
Abbreviation	Name	Class	Size (nt)	Function	Database (website)
miRNA	microRNA	miRNA	19–22	Post-transcriptional gene repression	miRBase.org
TSS-miRNA	Transcriptional start-site-miRNA	miRNA	20–30	Post-transcriptional gene regulation	N/A
moRNA	miRNA-offset RNAs	miRNA	19–22	N/A	N/A
usRNA	Unusually small RNA	miRNA	15–17	Post-transcriptional gene regulation	N/A
endo-siRNA	Endogenous small interfering RNA	siRNA	21	Somatic inhibitor of retrotransposition, post-transcriptional gene repression	N/A
exo-siRNA	Exogenous small interfering RNA	siRNA	21	Gene-targeted silencing, anti-viral	N/A
natsiRNA	Natural antisense siRNA	siRNA	21–24	Pathogen resistance, post-transcriptional gene regulation	bis.zju.edu.cn/pnatdb
casRNA	Cis-acting siRNA	siRNA	24	Transposon methylation, chromatin modification	N/A
tasiRNA	Trans-acting siRNA	siRNA	21	Post-transcriptional gene repression	bioinfo.jit.edu.cn
rasiRNA	Repeat-associated siRNA	siRNA	26–31	Transposon methylation, chromatin modification	deepbase.sysu.edu.cn
hc-siRNA	Heterochromatic small interfering RNA	siRNA	24	Genome maintenance, DNA methylation	N/A
3' U tRF	tRNA-derived fragment	tDR	20	Post-transcriptional gene repression	genome.ucsc.edu
5' tRF	tRNA-derived fragment	tDR	20	Post-transcriptional gene repression, translational repression	genome.ucsc.edu
3'CCA tRF	tRNA-derived fragment	tDR	23	Post-transcriptional gene repression, RNA metabolism	genome.ucsc.edu
tRH	tRNA-derived half	tDR	30–35	Protein synthesis repression, post-transcriptional gene repression	genome.ucsc.edu

Źródło: Vickers KC, Roteta LA, Hucheson-Dilks H, Han L, Guo Y. Mining diverse small RNA species in the deep transcriptome. Trends Biochem Sci. 2015 Jan;40(1):4–7.

sRNA pochodzące z długich, niekodujących prekursorów

Abbreviation	Name	Class	Size (nt)	Function	Database (website)
piRNA	Piwi-interacting RNA	piRNA	25–33	Germline post-transcription gene repression, transposon regulation, chromatin modification	pirnabank.ibab.ac.in
crasiRNA	Centromere repeat-associated small RNA	chRNA	34–42	Cell cycle, centromere formation	N/A
tel-sRNAs	Telomere-specific small RNA	chrRNA	24	Epigenetic regulation	N/A
PASR	Promoter-associated small RNA	CasRNA	20–200	Transcriptional regulation of mRNAs, epigenetic regulation	deepbase.sysu.edu.cn
tiRNA	Transcription initiation smRNA	CasRNA	18	Transcriptional regulation of mRNAs, epigenetic regulation	N/A
TSSaRNA	Transcription start site-associated RNAs	CasRNA	20–90	Transcriptional regulation of mRNAs, epigenetic regulation	N/A
spliRNA	Splice site-associated small RNA	CasRNA	17–18	N/A	N/A
H/ACA sdRNA	H/ACA snoRNA-derived RNA	snRNA	20–24	Post-transcriptional gene repression, alternative splicing	ensembl.org
C/D sdRNA	C/D snoRNA-derived RNA	snRNA	17–19, 30	Post-transcriptional gene repression, alternative splicing	ensembl.org
svRNA	Small vault RNAs	vRNA	22–37	Post-transcriptional gene repression	ensembl.org
srRNA	smRNAs-derived from rRNA	rRNA	24	Post-transcriptional gene regulation	ensembl.org
yDR	Y RNA-derived small RNAs	Y RNA	24–25, 30	N/A	ensembl.org
sbRNA	Stem-bulge RNA	Y RNA	67–133	RNA quality control, chromosomal replication	N/A
diRNA	Double-strand break-induced small RNAs	DSB	21	DNA double-strand break repair	N/A
qiRNA	QDE-2-interacting small RNAs	DSB	20–21	DNA double-strand break repair, protein translation	N/A

Źródło: Vickers KC, Roteta LA, Hucheson-Dilks H, Han L, Guo Y. Mining diverse small RNA species in the deep transcriptome. Trends Biochem Sci. 2015 Jan;40(1):4–7.



Biogeneza mikroRNA

miRNA

pre-miRNA in coding region



pre-miRNA in a noncoding region



pre-miRNA in intron of a protein-coding pre-mRNA



pre-miRNA in intron of noncoding RNA



mikroRNA uczestniczą w:

- morfogenezie
- procesach rozwoju
- odpowiedzi na stres
- szlakach przekazywania sygnałów
- apoptozie
- i wielu innych procesach...



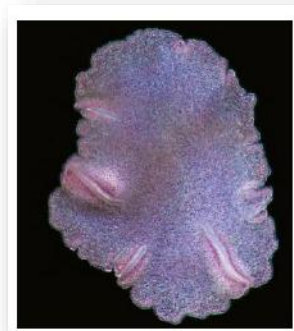
miRNA powiązано z licznymi chorobami u człowieka (miR2Disease):

- 349 miRNA
- 163 choroby



Ewolucja miRNA

- ✓ miRNA niezależnie wykształciły się u roślin i zwierząt
- ✓ wtórna utrata miRNA (*Placozoa*)
- ✓ mała liczba miRNA u prostych zwierząt (*Cnidaria*, *Porifera*)
- ✓ powstawanie genów miRNA *de novo*
- ✓ duplikacja genów miRNA



siRNA u roślin

casiRNA

Wywodzą się z transpozonów, elementów powtarzalnych i powtórzeń tandemowych.
Mają zwykle długość 24 nt.
Są metylowane przez HEN1.
Powstają z udziałem Pol IV.
Funkcja: metylacja DNA i modyfikacja histonów

tasiRNA

W procesie biogenezy wymagane jest cięcie transkryptu przez miRNA lub inne siRNA.
Mają długość 21 nt, wycinane są w fazie.
Są metylowane przez HEN1.

nat-siRNA

Powstają w odpowiedzi na stres.
Mają długość 21 lub 24 nt.
Powstają z udziałem Pol IV.

lsiRNA

Powstają w odpowiedzi na stres
Mają długość 30-40 nt.
Powstają z udziałem Pol IV.



siRNA u zwierząt

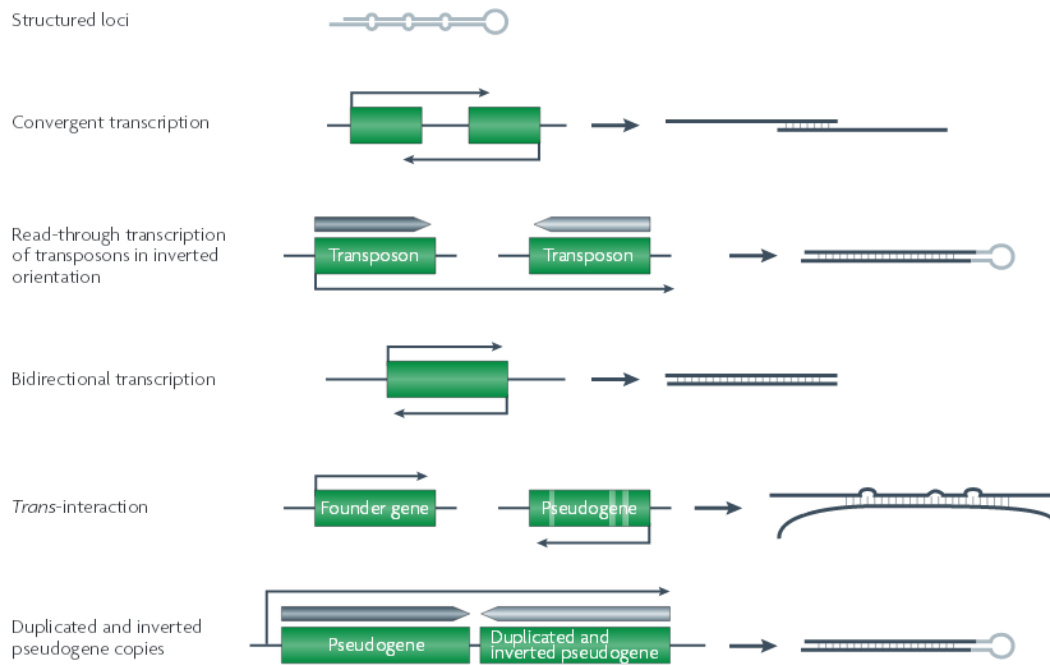
Powstają z transpozonów (np. retrotranspozon L1 u ssaków), regionów międzygenowych, długich cząsteczek RNA, wliczając mRNA.

L1 posiada może ulegać transkrypcji dwukierunkowo, co generuje dwa nakładające się transkrypty, z których mogą powstać siRNA.

Zazwyczaj mają długość 21 nt.

W przeciwieństwie do miRNA i piRNA, nie obserwuje się tendencji, by sekwencja zaczynała się od uracylu.

siRNA u zwierząt: biogeneza



Small silencing RNAs: an expanding universe. Ghildiyal and Zamore, Nature Reviews.

piRNA

Powstają w linii zarodkowej (spermatocyty). Zapewniają stabilność linii zarodkowej przez wyciszanie transpozonów, takich jak L1.

U ssaków wyróżniamy piRNA pachytenowe i pre-pachytenowe.

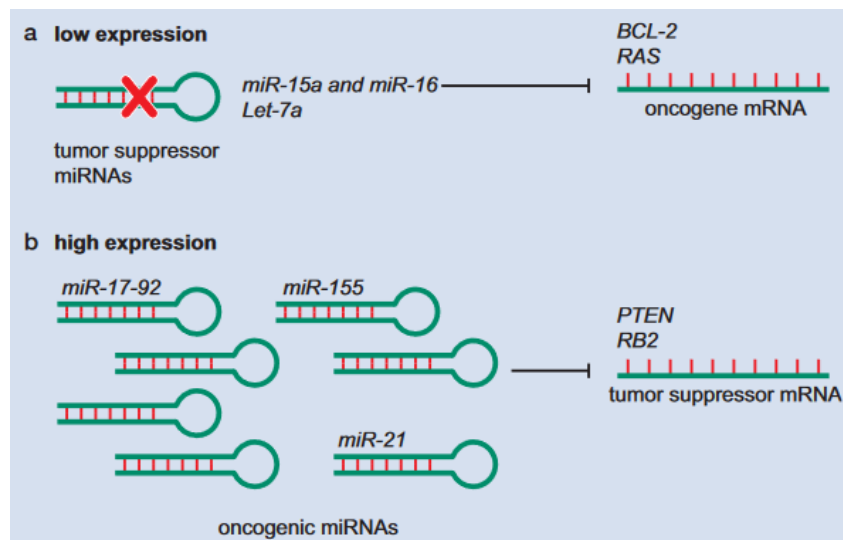
Pre-pachytenowe: wyciszają transpozony poprzez promowanie ich metylacji.

Pachytenowe: funkcja nieznana.

Pachyten to faza mejozy, podczas której chromosomy homologiczne wymieniają materiał genetyczny w procesie *crossing over*.

piRNA są dużo bardziej zróżnicowane niż miRNA, np. u muszki owocowej wykryto ich 1,5 miliona. Po zmapowaniu do genomu tworzą one kilkaset klastrów.

miRNA a nowotwory



Bazy danych małych RNA

Bazy danych sekwencji miRNA

Nazwa bazy danych	Gatunki (liczba)	Rodzaj danych	Metody i źródła danych	PMID
miRBase	zwierzęta, rośliny, wirusy (223)	opublikowane miRNA, referencyjne źródło adnotacji miRNA	literatura, dane od użytkowników, program RNAfold	20205188
PMRD	rośliny (123)	przewidziane in silico miRNA, ich ekspresja i mRNA docelowe	literatura, eksperymenty mikromacierzowe	19808935
microPC	rośliny (125)	przewidziane in silico miRNA	algorytm do identyfikacji miRNA w sekwencjach EST	19660144
miROrtho	zwierzęta (46)	przewidziane in silico miRNA	programy: R-COFFEE, RNAplfold, RNAalifold	18927110
Vir-Mir db	wirusy (1491)	przewidziane in silico miRNA	program Rnaloop, baza danych NCBI	17702763
miRNAmap	zwierzęta (13)	potwierdzone eksperymentalnie miRNA i ich mRNA docelowe	programy: miRanda, RNAhybrid, TargetScan, eksperymenty qPCR	16381831
GrapeMiRNA	winorośl	przewidziane in silico miRNA	program FindMiRNA	19563653
miRNEST	zwierzęta, rośliny, wirusy (544)	miRNA przewidywane in silico i/lub potwierdzone eksperymentalnie, mRNA docelowe, polimorfizm i regulacja ekspresji miRNA	literatura, algorytm do identyfikacji miRNA w sekwencjach EST, 13 baz danych miRNA (patrz: podrozdział miRNEST), GEO, NCBI	22135287

Bazy danych sekwencji docelowych

Nazwa bazy danych	Gatunki (liczba)	Rodzaj danych	Metody i źródła danych	PMID
miRWalk (dawniej: Argonaute)	człowiek, mysz, szczur	przewidziane oraz potwierdzone mRNA docelowe	bazy danych: GenBank, Ensembl, miRBase, programy: DIANA-microT, miRanda, miRDB, PicTar, PITA, RNA22, TargetScan/TargetScanS, miRWalk	21605702
HOCTAR	człowiek	mRNA docelowe	programy: miRanda, TargetScan, PicTar.	21435384
RepTar	człowiek, mysz	przewidziane in silico mRNA docelowe	nowy algorytm oparty na założeniu, że miRNA może posiadać więcej niż jedno miejsce wiązania do pojedynczej sekwencji UTR	21149264
miRTarBase	zwierzęta, rośliny, wirusy (14)	mRNA docelowe	literatura	21071411
miRGator	człowiek, mysz	mRNA docelowe miRNA i ich ekspresja, powiązania miRNA z chorobami	bazy danych: PhenomiR, GEO, ArrayExpress, programy: targetScan, PITA, miRanda, miRbridge	21062822
starBase	człowiek, mysz, C. elegans, rzodkiewnik pospolity, ryż, winorośl	mRNA docelowe	eksperymenty CLIP-Seq i Degradome-Seq	21037263
miRSEL	człowiek, mysz, szczur	mRNA docelowe	bazy danych: HGNC, MGD, Entrez Gene, Swiss-Prot Protein Database, miRGen, miRBase	20233441
miRecords	zwierzęta (9)	mRNA docelowe	literatura, programy: DIANA-microT, MicroInspector, miRanda, miTarget, MirTarget2, Nbmirtar, PicTar, PITA, RNA 22, RNA Hybrid, TargetScan/TargetScanS	18996891
TarBase	zwierzęta (6)	mRNA docelowe (tylko eksperymentalne)	literatura	18957447
miRDB	człowiek, mysz, szczur, pies, kura	mRNA docelowe oraz adnotacja funkcjonalna miRNA	baza danych miRBase, nowy algorytm do szukania mRNA docelowych	18426918
MicroRNA.org	człowiek, mysz, szczur, muszka owocowa, C. elegans	mRNA docelowe i ekspresja miRNA	literatura, program miRanda, bazy danych: miRBase, UCSC	18158296
MiRonTop	człowiek, mysz, szczur	mRNA docelowe	bazy danych: miRBase, NCBI, programy: Targetscan, MicroCosm Targets, Miranda, PicTar	20959382

Bazy danych ekspresji miRNA (1)

Nazwa bazy danych	Gatunki (liczba)	Rodzaj danych	Metody i źródła danych	PMID
CIRCUITSdb	człowiek, mysz	regulacja ekspresji miRNA przez czynniki transkrypcyjne	literatura, bazy danych: TransmiR , TarBase, Myc Target Gene	20731828
mESAdb	człowiek, mysz, danio pregowany	ekspresja miRNA i ich mRNA docelowych	bazy danych: Ensembl, miRBase, microCosm, HUGE, KEGG, GO	21177657
miRNeve	mysz	ekspresja miRNA w oku myszy	Eksperyment: hybrydyzacja RNA in situ z wykorzystaniem modyfikowanych nukleotydów LNA	21171988
dbDEMC	człowiek	ekspresja miRNA w tkankach nowotworowych	literatura	21143814
miReg	człowiek	regulacja ekspresji genów miRNA	literatura	20693604
PuTmiR	człowiek	regulacja ekspresji genów miRNA przez czynniki transkrypcyjne	bazy danych: miRBase, UCSC	20398296
S-MED	człowiek	ekspresja miRNA w sarkomie	eksperymenty z wykorzystaniem systemu BeadArrays	20212452
PhenomiR	człowiek	ekspresja miRNA w chorobach i różnych procesach biologicznych	literatura, bazy danych: OMIM Morbid Map, Gene Ontology, BRENDA Tissue Ontology	20089154
miRGen	zwierzęta (11)	regulacja ekspresji miRNA, polimorfizm, mRNA docelowe	literatura, program mathTM tool (szukanie TFBS), bazy danych: mammalian miRNA expression atlas, UCSC, dbSNP	19850714
TransmiR	zwierzęta	regulacja ekspresji miRNA przez czynniki transkrypcyjne	literatura, baza danych UCbase & miRfunc	19786497

Bazy danych ekspresji miRNA (2)

Nazwa bazy danych	Gatunki (liczba)	Rodzaj danych	Metody i źródła danych	PMID
miR2Disease	człowiek	ekspresja miRNA w chorobach	literatura, baza danych TarBase	18927107
GenomeTraFaC	człowiek, mysz	regulacja ekspresji miRNA przez czynniki transkrypcyjne	bazy danych: Homologene, NCBI, MGI, miRBase	17178752
miSolRNA	pomidor, rzodkiewnik pospolity	ekspresja miRNA oraz ich funkcje w szlakach metabolicznych	literatura	21059227
MirZ (dawniej: mammalian miRNA expression atlas)	człowiek, mysz, szczur	ekspresja miRNA	eksperymenty sekwencjonowania w technologii NGS	17604727
mirEX	rzodkiewnik pospolity	ekspresja miRNA	eksperymenty real-time PCR	22013167
mimiRNA	człowiek	ekspresja miRNA	literatura, programy: TargetScan, RNA22, PicTar, algorytm ExParser, bazy danych: Hypertext cell line database, mammalian miRNA expression atlas, GEO	19933167
mirConnX	człowiek, mysz	regulacja ekspresji miRNA	bazy danych: TarBase, miRBase, DBTSS, UCSC, The Eukaryotic Promoter Database, programy: CoreBoost_HM, PITA, miRANDA, TargetScan, RNAhybrid, Pictar	21558324

Polimorfizm

Nazwa bazy danych	Gatunki (liczba)	Rodzaj danych	Metody i źródła danych	PMID
miRvar	człowiek	polimorfizm i jego funkcjonalne konsekwencje	literatura, bazy danych: SNPdb, UCSC Genome Browser, miRBase, programy: PHDcleav, RISCbinder	21618345
Patrocles	zwierzęta (7)	polimorfizm miRNA i mRNA docelowych	literatura, bazy danych: miRBase, Ensembl, program RNAfold	19906729
PolymiRTS	człowiek, mysz	polimorfizm w mRNA docelowych	bazy danych: dbSNP, miRBase	17099235
dPORE-miRNA	człowiek	polimorfizm i regulacja ekspresji miRNA	bazy danych: UCSC, PhenomiR, Tarbase, KEGG, program BIOBASE MATCH	21326606
dbSMR	człowiek	polimorfizm miRNA	bazy danych: miRBase, Ensembl, programy: miRanda, RNAHybrid, TargetScan	19371411

Funkcje miRNA

Nazwa bazy danych	Gatunki (liczba)	Rodzaj danych	Metody i źródła danych	PMID
UCbase & miRfunc	człowiek, mysz, szczur	funkcje miRNA; konserwacja sekwencji miRNA	bazy danych: miRBase, UCSC, NCBI	18945703
miRNApath	człowiek, mysz, szczur, kura	udział miRNA w ścieżkach metabolicznych	bazy danych: miRBase, miRGen, miRGen, KEGG	18058708
miRò	człowiek	powiązania miRNA-fenotyp	bazy danych: miRBase, mammalian miRNA expression atlas, miRecords, NCBI, GO, Genetic Association Database, programy: TargetScan, PicTar, miRanda	20157481
miREnvironment	zwierzęta, rośliny (17)	powiązania miRNA - fenotyp	literatura	21984757
miTALOS	człowiek, mysz	udział miRNA w szlakach sygnalizacyjnych	programy: TargetScan, TargetScan, PicTar, Pita, RNA22, bazy danych: KEGG, NCBI	21441347

Inne bazy danych

Nazwa bazy danych	Gatunki (liczba)	Rodzaj danych	Metody i źródła danych	PMID
IntmiR	człowiek, mysz	intronowe miRNA, ich mRNA docelowe i deregulacja w chorobach	brak danych	21423893
CoGemiR	zwierzęta (36)	genomika i konserwacja sekwencji miRNA	bazy danych: miRBase, Ensembl, SymAtlas, CoGemiR, program miRNAMiner	18837977
AntagomirBase	człowiek	antagomiry (cząsteczki służące do wyciszania ekspresji genów miRNA)	programy: Sfold, mfold	21904438
HNOCDB	człowiek	miRNA powiązane z nowotworami głowy i szyi oraz nowotworem szczęki	literatura	22024348
EpimiR	człowiek, mysz i 5 innych gatunków	miRNA i regulacja na poziomie epigenetycznym	literatura	24682734
OncomiR	człowiek	onkogenne i supresorowe miRNA	literatura	24651967
OncomiRdbB	człowiek	miRNA w nowotworze piersi	Program: miRanda, zewnętrzne bazy danych	24428888
miR-SynthDB	człowiek	projektowanie syntetycznych miRNA	nowy algorytm	24627222
PASmiR	rośliny(33)	miRNA w stresach abiotycznych	literatura	23448274
miRCancer	człowiek	miRNA w nowotworach	literatura	23325619

miRBase

<http://mirbase.org/>

miRBase

[Home](#)
[Search](#)
[Browse](#)
[Help](#)
[Downloads](#)

Stem-loop

cfa-mir-33a

Accession

MI0007988

Description

Canis familiaris cfa-mir-33a precursor miRNA

Sequence

Copy Sequence

GUGCAUUGUAGUGUCAUUGGcaugucuggcgguaccggugcaauguuuccacagugcauca
(((((((((((((.....((.....))))))))))))..))))).....

Structure

```

--          A UU          uucu  c
GUGCAUUGU G  GCAUUGcaug  gg  g
|||||      |  |||||      ||  g
uacgugaca c  uguaacgugc  cc  u
ac          c  uu        ----  a

```

Annotation confidence

Not enough data

Do you think this miRNA is real?

Yes (+0)

No (-0)

Leave comment

Genome context

chr10: 23456613-23456673 [-]

miRBase



Inne bazy sRNA

piRNABank: <http://pirnabank.ibab.ac.in/index.shtml>

piRNA cluster database
<https://www.smallrnagroup.uni-mainz.de/piRNAclusterDB/>

PlantNATsDB
<http://bis.zju.edu.cn/pnatdb/>

ta-siRNA Database
<http://bioinfo.jit.edu.cn/tasiRNADatabase/>

The MIT/ICBP siRNA Database
<http://web.mit.edu/sirna/>

Analiza sekwencji docelowych

Sekwencje docelowe

- i. U roślin identyfikacja sekwencji docelowych jest łatwiejszym zadaniem niż u zwierząt, gdyż występuje wyższy stopień komplementarności między miRNA a mRNA
- ii. U zwierząt miRNA oddziałują najczęściej z 3'UTR, ale są znane liczne odstępstwa.
- iii. Wyróżniamy podejścia komputerowe (*in silico*) oraz eksperymentalne służące do identyfikacji miejsc docelowych.
- iv. Popularne ostatnio są techniki CLIP-Seq i Degradome-seq, pozwalające identyfikować sekwencje docelowe w sposób wielkoskalowy, a jednocześnie dające dużo bardziej wiarygodne wyniki niż podejście *in silico*.

Pytania

- i. Jakie geny są pod kontrolą danego miRNA?
- ii. Jakie miRNA regulują dany gen i w jakich pozycjach się przyłączają?
- iii. Jaki jest mechanizm działania miRNA?
- iv. Jaka jest ekspresja miRNA?
- v. Czy istnieje korelacja między ekspresją miRNA i mRNA?
- vi. Czy można wykazać oddziaływanie miRNA i mRNA?
- vii. Jaki jest efekt fenotypowy?

Identyfikacja sekwencji docelowych *in silico* oraz bazy danych

Narzędzia:

- psRNATarget
- PAREsnip
- TargetScan
- Miranda
- RNAhybrid
- PITA
- +wiele innych

Bazy danych:

- starBase
- DIANA mirPath
- miRTarBase
- miRNEST

psRNATarget

<https://www.zhaolab.org/psRNATarget/analysis>

psRNATarget: A Plant Small RNA Target Analysis Server (2017 Update)

HomeAnalysisDownloadHelpCitationContact

Analysis [\[Switch to Legacy mode\]](#)

Submit small RNAs

Submit target candidates

Submit small RNAs and targets

Upload sRNA file: [Przełóż...](#) Nie wybrano pliku. [\[Load Demo Data\]](#)

or paste sequences:

- Upload limit: 25MB, FASTA format or sequence only

Choose cDNA library:

Allium cepa (Onion), unigenes, DFCI Gene Index (ONIG), version 2, released on 2008_07_17

- Contact us to add more libraries

Scoring Schema

Schema V1 (2011 release)Schema V2 (2017 release)User-customized Schema

of top targets:

200

Penalty for G-U pair:

0.5

Extra weight in seed region:

1.5

of mismatches allowed in seed region:

2

Expectation:

5

Penalty for other mismatches:

1

Seed region:

2

 -

13

 NT

HSP size:

19

☒ Allow bulge(gap)

psRNATarget

psRNA Target: A Plant Small RNA Target Analysis Server (2017 Update)

[Home](#)
[Analysis](#)
[Download](#)
[Help](#)
[Citation](#)
[Contact](#)

Analysis result

Scoring Schema: Schema V2 (2017 release)

Download: [Download](#)

Filtering by keywords:

expect:

UPE:

Sorting by:

Expectation(E)

mRNA Acc.

Submit

1342 Rows

Page 1/68

mRNA Acc.	Target Acc.	Expect	UPE	Alignment	Target Description	Inhibition	Multiplic
ath-miR398a	AT5G14550.1	0.0	N/A	mRNA 21 UUUCCACUGGACUUGUGU 1 Target 1651 CAGGGUGACUGAGACACA 1671	[Symbols: Core-21-branching beta-1,6-N-acetylglucosaminyltransferase family protein chr5:4691006-4694055 REVERSE LENGTH=1772	Cleavage	1
ath-miR398b	AT5G14550.1	0.0	N/A	mRNA 21 GUUCCACUGGACUUGUGU 1 Target 1651 CAGGGUGACUGAGACACA 1671	[Symbols: Core-21-branching beta-1,6-N-acetylglucosaminyltransferase family protein chr5:4691006-4694055 REVERSE LENGTH=1772	Cleavage	1
ath-miR398c	AT5G14550.1	0.0	N/A	mRNA 21 GUUCCACUGGACUUGUGU 1 Target 1651 CAGGGUGACUGAGACACA 1671	[Symbols: Core-21-branching beta-1,6-N-acetylglucosaminyltransferase family protein chr5:4691006-4694055 REVERSE LENGTH=1772	Cleavage	1
ath-miR156a	AT2G33810.1	0.5	N/A	mRNA 20 CACGAGUGAGAGAAGACAGU 1 Target 787 UUGCUUACUUCUUCUGUCA 806	[Symbols: SPL3 squamosa promoter binding protein-like 3 chr2:14306001-14306072 FORWARD LENGTH=981	Cleavage	1
ath-miR156a	AT3G57920.1	1.0	N/A	mRNA 20 CACGAGUGAGAGAAGACAGU 1 Target 845 GUGCUUCUUCUUCUGUCA 864	[Symbols: SPL15 squamosa promoter binding protein-like 15 chr3:21444321-21446035 REVERSE LENGTH=1377	Cleavage	1
ath-miR156a	AT1G27360.2	1.0	N/A	mRNA 20 CACGAGUGAGAGAAGACAGU 1 Target 1213 GUGCUUCUUCUUCUGUCA 1232	[Symbols: SPL11 squamosa promoter-like 11 chr1:9501077-9503869 FORWARD LENGTH=1464	Cleavage	1

TargetScan

<http://www.targetscan.org/>



Search for predicted microRNA targets in mammals

[\[Go to TargetScanMouse\]](#)

[\[Go to TargetScanWorm\]](#)

[\[Go to TargetScanFly\]](#)

[\[Go to TargetScanFish\]](#)

1. Select a species

AND

2. Enter a human Entrez Gene symbol (e.g. "LIN28A")

AND/OR

3. Do one of the following:

- Select a broadly conserved* microRNA family
- Select a conserved* microRNA family
- Select a poorly conserved microRNA family Note that these families also include small RNAs that
- Enter a microRNA name (e.g. "mmu-miR-1")

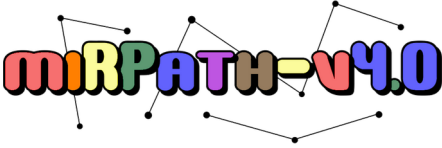
Bazy danych

Sekwencje docelowe

DIANA mirPath

<http://62.217.122.229:3838/app/miRPathv4>

[miRPath v4.0](#) | [Home](#) | [miRNA-centric analysis](#) | [Term-centric analysis](#) | [CRISPR analysis](#) | [Upload dashboard](#) | [Help](#)



Welcome to DIANA-miRPath v4.0, the server for target-based analysis of miRNA functions!

DIANA-miRPath v4.0 webserver offers numerous analysis options for the target-based functional analysis of miRNAs.

Expert miRNA interaction resources lend their power to miRPath, allowing users to perform overrepresentation term analysis in condition-free schemes, using either **experimentally supported targets** – DIANA-TarBase v8.0, DIANA-LncBase v3.0 and miRTarBase v6.0 – or sets of **predicted miRNA targets** – DIANA-microT-CDS and TargetScan. For the first time, miRPath users can also perform tissue-/cell-type-specific investigations in human and mouse, by selecting among 41 high-confidence miRNA-target sets derived from in-house analysis of AGO2-CLIP-Seq datasets with the avant-garde super-learning framework **microCLIP** (Ng et al. 2018).

Each interaction resource may be filtered according to user-selected criteria. Targets supported by direct experimental methods can be retained in TarBase, miRTarBase, TargetScan. For the first time, miRPath users can also perform tissue-/cell-type-specific investigations in human and mouse, by selecting among 41 high-confidence miRNA-target sets derived from in-house analysis of AGO2-CLIP-Seq datasets with the avant-garde super-learning framework **microCLIP** (Ng et al. 2018).

miRPath v4.0 integrates term annotation sources from the [Gene Ontology](#) consortium, [Reactome](#) Pathway Database, KEGG pathways, expertly curated gene sets from [MSigDB](#) and protein domains from [Pfam](#).

Targets Union



Targets Intersection



Pathways Union



ENCORI / starBase

<https://rnasysu.com/encori/>

Gromadzi informacje o sekwencjach docelowych otrzymanych m.in.. z eksperymentów CLIP-Seq i Degradome-Seq.

RNA-binding protein CLIP-seq
PAR-CLIP, CLIP, HITS-CLIP,
eCLIP, HCLIP, FANT-CLIP, etc.

Argonaute CLIP-seq
PAR-CLIP, CLIP, HITS-CLIP,
CLASH, CLEAR-CLIP, eCLIP, etc.

RNA-RNA Interactome
PARIS, LIOR-seq, SPLASH,
and MARDI-seq.

Degradome sequencing
PARE-seq, degradome-seq,
ORFOT-seq, RACE-seq, etc.

StarScan software
RBP Binding Sites and Motifs
GOAUGA

StarScan software
AGO Binding Sites
GOAUGA

StarScan software
RNA-RNA Interactions
StarScan software
Cleavage Sites

Construct regulatory networks

miRNA-target Prediction programs

Annotate Interactions

Annotate Cleavage Sites

miRNA Targetome
>480,000 miRNA-mRNA
>1,200,000 miRNA-mRNA
TMD Potential Prediction

RBP-RNA Interactome
>1,200,000 RBP-mRNA
>1,600,000 RBP-mRNA
Region Binding Preferences

RNA Interactome
>3,700,000 RNA-RNA
Interaction Categories (ncRNA,
ncRNA, mRNA, pseudogenes)

ENCORI Platform

miRNA-RNA

RNA-RNA

Protein-RNA

RBP

Pan-Cancer Analysis
Differential Expression Analysis
Gene Survival Analysis
Gene Coexpression Analysis

Cancer/Disease Genomics Data
>3,280,000 mutations
>100 diseases
>10,000 cancer samples

Diverse Signaling Pathways
Gene Ontology, KEGG,
PANTHER, Reactome, BioCarta,
All Chemical Pathways, etc.

RBP motifs

Differential Expression

Survival Analysis

RNA-RNA Interactome

Expression Networks

RBP-RNA Interactome

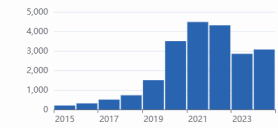
Pathway Enrichment

ENCORI/starBase (Encyclopedia of RNA Interactomes) is an extensive atlas that integrates precise RNA interactions identified by our innovative **StarScan** and **StarScan** algorithms, showcasing the functional and mechanistic insights into the RNA interactomes.

Utilizing our new **StarScan** software, we systematically identified distinct RNA-binding preferences for hundreds of RBPs (RNA-binding proteins), uncovering millions of high-confidence RBP-RNA interactions with diverse cellular functions and significant clinical implications in cancers. Additionally, our novel **StarScan** software revealed the complexity of RNA regulation by discovering millions of reliable RNA-RNA interactions.

Accompany with gene expression data of 32 types of cancers, ENCORI allows researchers performing Pan-Cancer analysis on RNA-RNA and RBP-RNA interactions. ENCORI also provides platforms to perform the survival and differential expression analysis of miRNAs, lncRNAs, pseudogenes and mRNAs.

ENCORI/starBase Citations (2015-2024) Cite Our Paper
>22,000 Citations in Google Scholar (starBase, the pilot version)



Year	Citations
2015	~500
2016	~1000
2017	~1500
2018	~2000
2019	~3000
2020	~4000
2021	~4500
2022	~4500
2023	~3500
2024	~3000

- Average Daily Usage: ~6000 visits from ~500 unique researchers
- Average Annual Usage: >1,000,000 visits from >100 countries

Statistics of ENCORI (last update: 10/18/2024)

- Species: 23 Species
- CLIP-seq data: 2,725 datasets
- Degradome-seq: 100 datasets
- RNA-RNA interactome: 59 datasets
- RNA-seq data: >10,800 samples from 32 cancer types
- miRNA-seq data: >10,500 samples from 32 cancer types
- Disease data: >1,800,000 mutations from 531 disease types
- miRNA-mRNA(CLIP): >460,000 interactions
- miRNA-mRNA(CLIP): >1,200,000 interactions
- RBP-mRNA: >1,290,000 interactions
- RBP-mRNA: >1,600,000 interactions
- RNA-RNA: >3,700,000 interactions
- miRNA-mRNA(degradome): >32,000 interactions
- miRNA-mRNA(degradome): >459,000 interactions
- ceRNA: >2,900,000 pairs
- function annotation: >34,000 functional terms from 21 categories
- Pan-Cancer: Differential Expression, Survival Analysis, CoExpression

How to cite:

- Zhou KR, Huang JH, Liu S, Zheng WJ, Liu SR, et al. An Encyclopedic Regulatory and Functional Atlas of RNA Interactomes. (Manuscript submitted)
- Li JH, et al starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data [2]. Nucleic Acids Res. 2014 Jan;42:D92-7.

Copyright © 2024-2025 Qu Lab & Yang Lab School of Life Science, Sun Yat-sen University, China.

miRTarBase

https://mirtarbase.cuhk.edu.cn/~miRTarBase/miRTarBase_2025/php/browse.php

Baza ta gromadzi dane o potwierdzonych eksperymentalnie sekwencjach docelowych.



miRTarBase

miRBase

HOMESEARCHBROWSEHELPDOWNLOADCONTACT US

↑

Page of 68

◀ Prev

1

2

...

68

Next ▶

Filter for miRNA and target

Search

Example

ID	Species (miRNA)	Species (Target)	miRNA	Target	Validation methods								Sum	# of papers	
					Strong evidence			Less strong evidence							
					Western blot	qPCR	Microarray	NGS	pSILAC	Other	CLIP-Seq				
MIRT000012	Homo sapiens	Homo sapiens	hsa-miR-122-5p	CYP7A1	✓		✓					✓	3	1	
MIRT000364	Homo sapiens	Homo sapiens	hsa-miR-122-5p	IGF1R	✓	✓	✓					✓	4	3	
MIRT000365	Homo sapiens	Homo sapiens	hsa-miR-122-5p	SRF	✓	✓	✓					✓	4	1	
MIRT000663	Homo sapiens	Homo sapiens	hsa-miR-122-5p	RAC1	✓	✓	✓	✓				✓	5	2	
MIRT000717	Homo sapiens	Homo sapiens	hsa-miR-122-5p	RHOA	✓							✓	2	2	
MIRT003006	Homo sapiens	Homo sapiens	hsa-miR-122-5p	CCNG1								✓	7	5	
MIRT003079	Homo sapiens	Homo sapiens	hsa-miR-122-5p	GTF2B			✓					✓	2	1	
MIRT003080	Homo sapiens	Homo sapiens	hsa-miR-122-5p	GYS1		✓	✓	✓				✓	4	2	
MIRT003081	Homo sapiens	Homo sapiens	hsa-miR-122-5p	ANK2	✓		✓					✓	3	1	
MIRT003082	Homo sapiens	Homo sapiens	hsa-miR-122-5p	NFATC2IP	✓		✓					✓	3	1	
MIRT003083	Homo sapiens	Homo sapiens	hsa-miR-122-5p	ENTPD4	✓		✓					✓	✓	4	2

Jeśli nie wskazano inaczej, część ilustracji pochodzi z:

Molecular Biology of the Gene, 7th edition

James D. Watson, Tania A. Baker, Stephen P. Bell, Alexander Gann,
Michael Levine, Richard Losick

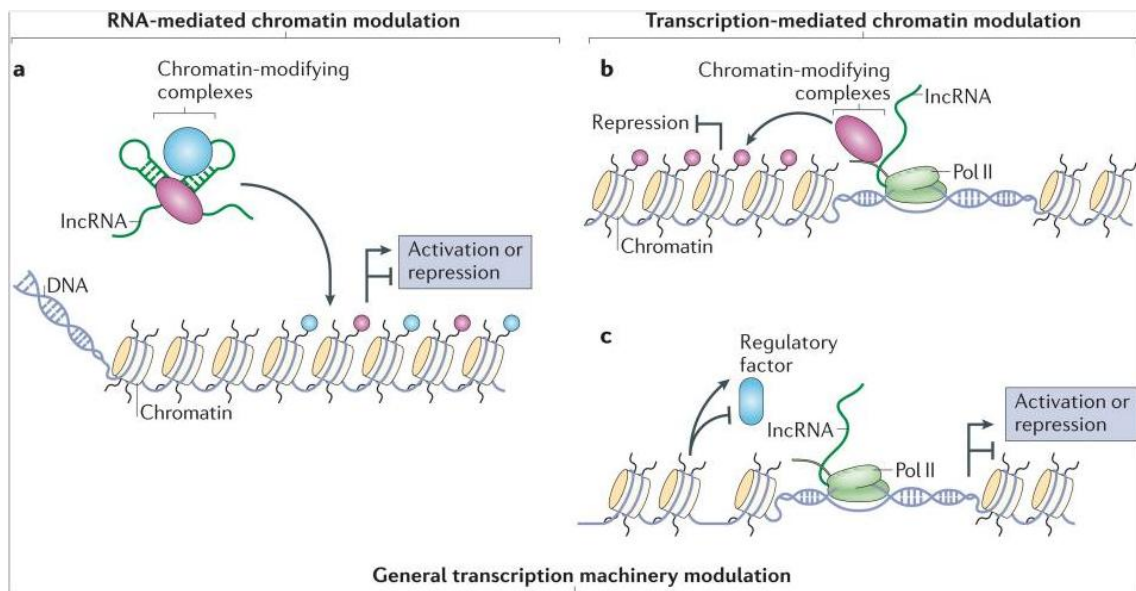
Pearson 2013

Długie niekodujące RNA

Długie niekodujące RNA

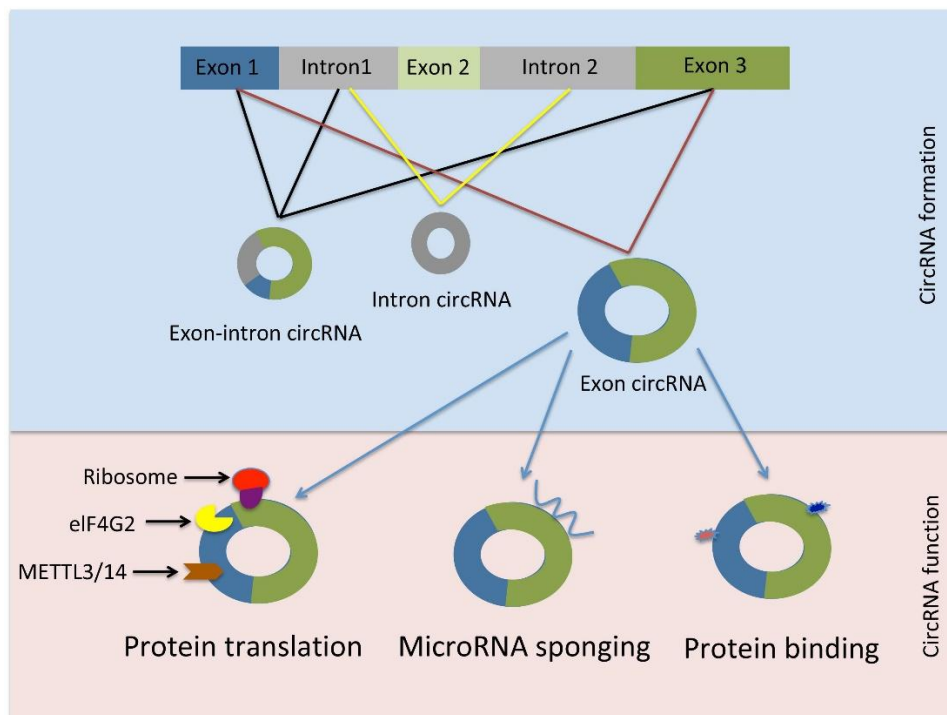
- Długość: co najmniej 200 nt
- Nie kodują białka
- Lokalizacja jądrowa lub cytoplazmatyczna
- Są bardzo liczne (ok. 60 000 w ENSEMBL)
- Pełnią różnego rodzaju funkcje (w *cis* lub w *trans*)

Funkcje lncRNA



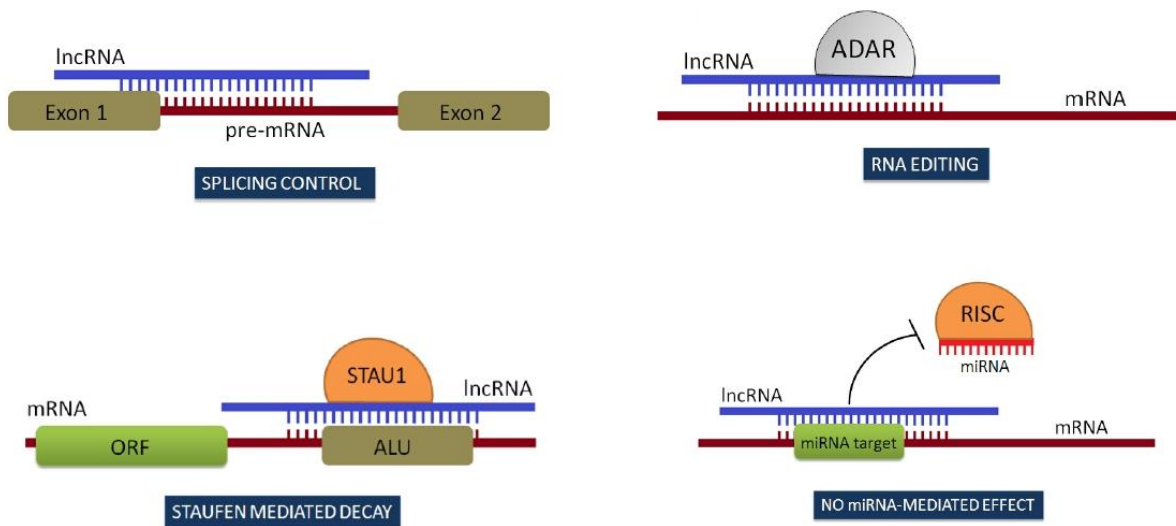
doi: [10.1038/nrm3679](https://doi.org/10.1038/nrm3679)

circRNA



<https://doi.org/10.3389/fphar.2019.00428>

Funkcje lncRNA



Brief Bioinform. 2016 Mar 4; pii: bbw015. [Epub ahead of print]

A comprehensive overview of lncRNA annotation resources.

Xu J, Bai J, Zhang X, Lv Y, Gong Y, Liu L, Zhao H, Yu F, Ping Y, Zhang G, Lan Y, Xiao Y, Li X.

Abstract

Long noncoding RNAs (lncRNAs) are emerging as a class of important regulators participating in various biological functions and disease processes. With the widespread application of next-generation sequencing technologies, large numbers of lncRNAs have been identified, producing plenty of lncRNA annotation resources in different contexts. However, at present, we lack a comprehensive overview of these lncRNA annotation resources. In this study, we reviewed 24 currently available lncRNA annotation resources referring to > 205 000 lncRNAs in over 50 tissues and cell lines. We characterized these annotation resources from different aspects, including exon structure, expression, histone modification and function. We found many distinct properties among these annotation resources. Especially, these resources showed diverse chromatin signatures, remarkable tissue and cell type dependence and functional specificity. Our results suggested the incompleteness and complementarity of current lncRNA annotations and the necessity of integration of multiple resources to comprehensively characterize lncRNAs. Finally, we developed 'LNCat' (lncRNA atlas, freely available at <http://biocc.hrbmu.edu.cn/LNCat/>), a user-friendly database that provides a genome browser of lncRNA structures, visualization of different resources from multiple angles and download of different combinations of lncRNA annotations, and supports rapid exploration, comparison and integration of lncRNA annotation resources. Overall, our study provides a comprehensive comparison of numerous lncRNA annotations, and can facilitate understanding of lncRNAs in human disease.

© The Author 2016. Published by Oxford University Press. For Permissions, please email: journals.permissions@oup.com.

KEYWORDS: annotation resource; database; integration; lncRNA

Selection of different annotation strategies can also contribute to the variations. **CABILI** and **KELLEY**, for example, applied different strategies to the same RNA-seq data (more than three reads, PhyloCSF<100 and without a Pfam hit for **CABILI**; FPKM1,2 exons and PhyloCSF<100 for **KELLEY**), but obtained distinct lncRNA annotations. **CABILI** contained 14 281 lncRNA transcripts, while **KELLEY** contained 21 244 lncRNA transcripts. There were only 4337 common transcripts shared by both resources.

Identyfikacja lncRNA

Table 1. Summary of 24 lncRNA annotation resources reviewed in this study

Source	Data type	Tissue/cell	Samples	lncRNA genes	ResearchMethod scope	Read type	Exon number	Length	Expression	Coding potential	Epigenetic signals	Ref.
CABILI	RNA-seq	24 tissues and cell types	24	8195	lncRNA <i>ab initio</i> assembly	Paired end and single	≥2	>200 bp	≥3 reads per base	PhyloCSF < 100, without hit in Pfam		[2]
KELLEY	RNA-seq	28 tissues and cell lines	70	9164	lncRNA <i>ab initio</i> assembly	Paired end and single	≥2	>200 bp	≥1 FPKM	PhyloCSF < 100		[17]
KRETZ	RNA-seq	keratinocytes	3	654	lncRNA <i>ab initio</i> assembly	Paired end	≥2	>200 bp	>5 RPKM			[15]
DING	RNA-seq	Breast cancer tissues	25	344	lncRNA <i>ab initio</i> assembly	Paired end			>10 read			[29]
KHALIL	ChIP-seq	6 cell lines	12	2510	lncRNA <i>ab initio</i> assembly	Single		>5 Kb			H3K4me3 and H3K36me3	[25]
WHITE	RNA-seq	Lung cancer tissues	728	4067	lncRNA <i>de novo</i> assembly	Paired end		>200 bp		GeneID		[16]
HE	RNA-seq	Prefrontal cortex	38	1888	lncRNA <i>de novo</i> assembly	Single	≥2	>200 bp	>1 read	PhyloCSF < 100, ORF < 100, without hit in Pfam		[20]
HANGAUER	RNA-seq	23 tissues	127	3968	lncRNA <i>de novo</i> assembly	Paired end	≥2	>200 bp	>1 RPKM	ORF < 100		[18]
IYER	RNA-seq	18 organs	7256	52 238	lncRNA <i>ab initio</i> assembly	Paired end and single		>200 bp		Pfam/CPAT		[9]
TRIMARCHI	RNA-seq and ChIP-seq	T-ALL cell lines and primary leukemia samples	14	1984	lncRNA <i>ab initio</i> assembly	Paired end	≥2	>200 bp	≥3 reads	PhyloCSF < 100	H3K4me3, H3K4me1 and H3K27ac	[14]
MORAN	RNA-seq and ChIP-seq	Islets and beta-cells	15	1128	lncRNA <i>ab initio</i> assembly	Paired end		>200 bp	>0.5 RPKM	ORF < 130, without hit in Pfam	H3K4me3	[13]
SIGOVA1	RNA-seq and ChIP-seq	hESC	3	3983	lncRNA <i>ab initio</i> assembly	Paired end		>100 bp	>0.07 FPKM	CPC < 0	H3K4me3	[26]
SIGOVA2	RNA-seq and ChIP-seq	Human endoderm cell	3	3544	lncRNA <i>ab initio</i> assembly	Paired end		>100 bp	>0.07 FPKM	CPC < 0	H3K4me3	[26]
BELL	RNA-seq	Coronary artery smooth muscle cell	3	31	lncRNA <i>ab initio</i> assembly	Single	≥2	>200 bp	>0.7 RPKM	PhyloCSF < 100, without hit in Pfam		[27]
YANG	RNA-seq	Failing LV samples	16	113	lncRNA <i>ab initio</i> assembly	Paired end	≥2		>0.5 RPKM			[21]
NE	RNA-seq	Monocytes	8	2523	lncRNA <i>ab initio</i> assembly	Paired end	≥2	>200 bp				[23]
PARALKAR	RNA-seq	Erythroblasts	15	594	lncRNA <i>ab initio</i> assembly	Paired end	≥2	>200 bp	≥3 read	BlastX, HMMER, PhyloCSF, GetORF		[19]
SOWALSKY	RNA-seq	Castration-resistant prostate cancer (CRPC) tissues	8	2965	lncRNA <i>ab initio</i> assembly	Paired end	≥2	>200 bp				[24]
YAN	RNA-Seq	Preimplantation embryos and hESCs	124	2121	lncRNA <i>de novo</i> assembly	Single	≥2	>1 kb	>1 read	CPC < 0		[28]
NECSULEA1	RNA-Seq	8 organs	185	14 677	lncRNA <i>de novo</i> assembly	Single	≥2	>200 bp	>10 reads			[22]
NECSULEA2	RNA-Seq	2 organs	53	12 667	lncRNA <i>de novo</i> assembly	Single	≥2	>200 bp	>10 reads			[22]
GENCODE	Manually collected			13 869	lncRNA							[1]
LNCCipedia	Intergrative database			17 385	lncRNA							[10]
NONCODE	Intergrative database			54 818	lncRNA							[11]

Xu J, Bai J, Zhang X, Lv Yet al. A comprehensive overview of lncRNA annotation resources. Brief Bioinform. 2016 Mar 4. pii: bbw015.

Brief Bioinform. 2016 Mar 4; pii: bbw015. [Epub ahead of print]

A comprehensive overview of lncRNA annotation resources.

Xu J, Bai J, Zhang X, Lv Y, Gong Y, Liu L, Zhao H, Yu F, Ping Y, Zhang G, Lan Y, Xiao Y, Li X.

Abstract

Long noncoding RNAs (lncRNAs) are emerging as a class of important regulators participating in various biological functions and disease processes. With the widespread application of next-generation sequencing technologies, large numbers of lncRNAs have been identified, producing plenty of lncRNA annotation resources in different contexts. However, at present, we lack a comprehensive overview of these lncRNA annotation resources. In this study, we reviewed 24 currently available lncRNA annotation resources referring to > 205 000 lncRNAs in over 50 tissues and cell lines. We characterized these annotation resources from different aspects, including exon structure, expression, histone modification and function. We found many distinct properties among these annotation resources. Especially, these resources showed diverse chromatin signatures, remarkable tissue and cell type dependence and functional specificity. Our results suggested the incompleteness and complementarity of current lncRNA annotations and the necessity of integration of multiple resources to comprehensively characterize lncRNAs. Finally, we developed 'LNCat' (lncRNA atlas, freely available at <http://biocc.hrbmu.edu.cn/LNCat/>), a user-friendly database that provides a genome browser of lncRNA structures, visualization of different resources from multiple angles and download of different combinations of lncRNA annotations, and supports rapid exploration, comparison and integration of lncRNA annotation resources. Overall, our study provides a comprehensive comparison of numerous lncRNA annotations, and can facilitate understanding of lncRNAs in human disease.

© The Author 2016. Published by Oxford University Press. For Permissions, please email: journals.permissions@oup.com.

KEYWORDS: annotation resource; database; integration; lncRNA

Selection of different annotation strategies can also contribute to the variations. **CABILI** and **KELLEY**, for example, applied different strategies to the same RNA-seq data (more than three reads, PhyloCSF<100 and without a Pfam hit for **CABILI**; FPKM1,2 exons and PhyloCSF<100 for **KELLEY**), but obtained distinct lncRNA annotations. **CABILI** contained 14 281 lncRNA transcripts, while **KELLEY** contained 21 244 lncRNA transcripts. There were only 4337 common transcripts shared by both resources.

Brief Bioinform. 2016 Mar 4; pii: bbw015. [Epub ahead of print]

A comprehensive overview of lncRNA annotation resources.

Xu J, Bai J, Zhang X, Lv Y, Gong Y, Liu L, Zhao H, Yu F, Ping Y, Zhang G, Lan Y, Xiao Y, Li X.

Abstract

Long noncoding RNAs (lncRNAs) are emerging as a class of important regulators participating in various biological functions and disease processes. With the widespread application of next-generation sequencing technologies, large numbers of lncRNAs have been identified, producing plenty of lncRNA annotation resources in different contexts. However, at present, we lack a comprehensive overview of these lncRNA annotation resources. In this study, we reviewed 24 currently available lncRNA annotation resources referring to > 205 000 lncRNAs in over 50 tissues and cell lines. We characterized these annotation resources from different aspects, including exon structure, expression, histone modification and function. We found many distinct properties among these annotation resources. Especially, these resources showed diverse chromatin signatures, remarkable tissue and cell type dependence and functional specificity. Our results suggested the incompleteness and complementarity of current lncRNA annotations and the necessity of integration of multiple resources to comprehensively characterize lncRNAs. Finally, we developed 'LNCat' (lncRNA atlas, freely available at <http://biocc.hrbmu.edu.cn/LNCat/>), a user-friendly database that provides a genome browser of lncRNA structures, visualization of different resources from multiple angles and download of different combinations of lncRNA annotations, and supports rapid exploration, comparison and integration of lncRNA annotation resources. Overall, our study provides a comprehensive comparison of numerous lncRNA annotations, and can facilitate understanding of lncRNAs in human disease.

© The Author 2016. Published by Oxford University Press. For Permissions, please email: journals.permissions@oup.com.

KEYWORDS: annotation resource; database; integration; lncRNA

Selection of different annotation strategies can also contribute to the variations. **CABILI** and **KELLEY**, for example, applied different strategies to the same RNA-seq data (more than 100 million reads) and obtained distinct lncRNA annotations. **CABILI** contained 14 281 lncRNA transcripts, while **KELLEY** contained 21 244 lncRNA transcripts. There were only 4337 common transcripts shared by both resources.

Wybór metody identyfikacji lncRNA ma kluczowe znaczenie

Identyfikacja lncRNA: w praktyce

1. Wysokiej jakości transkryptom (dobrze złożone transkrypty, reprezentatywny zestaw transkryptów)
2. Identyfikacja transkryptów kodujących białko
3. Filtr dotyczący długości transkryptów
4. Filtr dotyczący znanych cząsteczek, które nie są lncRNA
5. Filtr dot. poziomu ekspresji
6. Filtr dot. splicingu
7. Dodatkowe narzędzia sprawdzające np. potencjał kodujący

Identyfikacja lncRNA: wybrane narzędzia

Tool	Model group (type)	Input format	Trained on/Retraining (yes/no)	Installation	Running	Usability score	Reference
COME	RandomForest	gtf	Hg19 / no	3	2=only hg19	5	[24]
CPAT	logReg	bed, fasta	RefSeq, GENCODE / Yes	3	4	7	[21]
CPC2	SVM	fasta	Hg19 / yes	3	4	7	[22]
DeepLNC	DeepLearning (NN)	fasta	RefSeq, Lncipedia / yes	NA/3	0	3	[39]
FEEInc	RandomForest	gtf	Hg38, GENCODE 25 / yes	2= BioPerl installation required; requires much space.	2= Only '.fasta' accepted for genome file name	4	[40]
IRSOM	NN (SOM)	fasta	Ensemble 92, GENCODE / yes	2= Featurer needs specific version of compiler	4	6	[25]
iSeeRNA	SVM	gtf, bed	Hg19 / yes (not recommended)	1= dependent software, preparation of config files	2= only hg19	3	[41]
lncADeep	DeepLearning (DBN)	fasta	RefSeq 75, GENCODE 24 / yes	3	4	7	[26]
LncFinder	SVM	fasta	GENCODE / yes	3	4	7	[23]
LncRNet	DeepLearning (CNN)	fasta	GENCODE 25, Ensembl 87 / yes	3	2= Does not accept 'N' as a nt	5	[42]
lncRScan-SVM	SVM	gtf, fasta	Hg19, GENCODE 19 / yes (not recommended)	3	2= Many preparation steps, hg19	5	[43]
lncScore	logReg	bed, fasta	GENCODE 23 / yes	3	0= Error: 'linear model not converging'	3	[44]
longdist	SVM	fasta	Hg19, GENCODE 19 / yes (not recommended)	3	2= Error in attribute list format needed to be fixed	5	[7]
mRNN	DeepLearning (RNN)	fasta	GENCODE 25 / yes	3	4	7	[11]
RNAse3	DeepLearning (RNN)	fasta	CPC2, FEEInc, mRNN / yes	3	4	7	[27]

<https://doi.org/10.1093/bfpg/elab045>

Bazy danych lncRNA (przykłady)

List of long non-coding RNA databases

From Wikipedia, the free encyclopedia

This is a **list of long noncoding RNA databases**, which provide information about lncRNAs.^{[1][2]}

[Long non-coding RNA databases](#) [\[edit \]](#)

Name ↕	Description	References ↕
deepBase 	Identification, expression, evolution and function of long non-coding RNAs (lncRNAs), small RNAs and circular RNAs from deep-sequencing data	[3]
LNCipedia 	A comprehensive compendium of human long non-coding RNAs.	[4][5]
lncRNAdb 	The Reference Database For Functional Long Noncoding RNAs.	[6]
lncRNAWiki 	A wiki-based, publicly editable and open-content platform for community curation of human long non-coding RNAs (lncRNAs)	[7]
lncBook 	A comprehensive collection of 270,044 human lncRNAs and systematic curation of lncRNAs' annotation by multi-omics data integration, function annotation and disease association	[8]
MONCLdb 	The MOuse NonCode Lung database provides the annotations and expression profiles of mouse long non-coding RNAs (lncRNAs) involved in influenza and SARS-CoV infections.	[9]
NONCODE 	An integrated knowledge database dedicated to ncRNAs, especially lncRNAs	[10]
lncRNome 	A comprehensive searchable biologically oriented knowledgebase for long noncoding RNAs in Humans.	
NRED 	A database of long noncoding RNA expression.	[11]
C-It-Loci 	A tool to explore and to compare the expression profiles of conserved loci among various tissues in three organisms	[12]
MITranscriptome 	A catalog of human long poly-adenylated RNA transcripts derived from computational analysis of high-throughput RNA -Seq data from over 6,500 samples, spanning diverse cancer and tissue types	[13]
sincfy Evolution Browser 	This site contains alignments and evolutionary metrics of conserved lncRNAs.	[14]
Cancer lncRNA Census (CLC) 	Database of long-noncoding RNAs causally implicated in cancer through <i>in vivo</i> , <i>in vitro</i> and other evidence.	[15]
BIGTranscriptome 	High-confidence of coding and noncoding transcriptomes assembled with hundreds of pseudo-stranded and stranded RNA-seq datasets.	[16]
lncRNAKB	A knowledgebase of tissue-specific functional annotation and trait association of long noncoding RNA	[17]

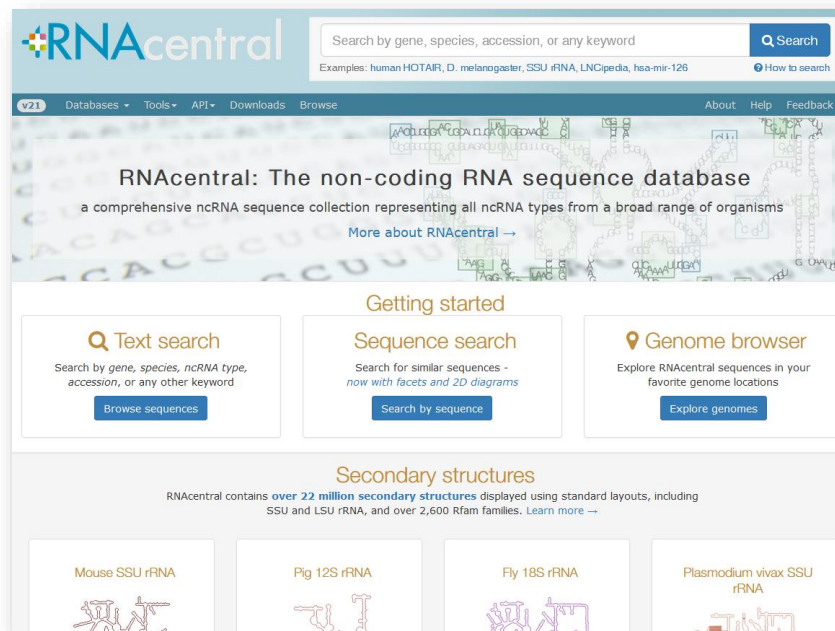
Źródło: Wikipedia

Bazy danych lncRNA (przykłady)

Tool/Database Name	Description	Link	Citation
General lncRNA Databases			
LNCipedia	General lncRNA database, expertly curated	https://lncipedia.org/	[18,19]
LNCBook	General lncRNA database, some community curation	http://bigd.big.ac.cn/lncbook/index	[29]
Expression Databases			
GTEx	Comprehensive public resource to study tissue-specific gene expression and regulation.	https://gtexportal.org/home/	[39,40]
TANRIC	Database of non-coding RNAs in cancer	www.tanric.org	[42]
CANTATAdb	Database of plant lncRNAs	http://cantata.amu.edu.pl/	[48,49]
Protein Coding Potential			
CPC	Protein coding potential prediction	http://cpc.gao-lab.org/	[55]
CPPred	Protein coding potential prediction	http://www.rnabinding.com/CPPred/	[60]
CNIT	Protein coding potential prediction	http://cnit.noncode.org/CNIT/	[62]
Subcellular Localization			
LncSLdb	Database of lncRNA subcellular localization	http://bioinformatics.xidian.edu.cn/lncSLdb/	[72]
LncATLAS	Database of lncRNA subcellular localization	https://lncatlas.crg.eu/	[76]
LncLocator	LncRNA subcellular localization prediction	http://www.csbio.sjtu.edu.cn/bioinf/lncLocator/	[77]
Structural Conformation			
RMDB	Database of RNA structures	https://rmdb.stanford.edu	[94]
RNAfold	RNA structure prediction	http://rna.tbi.univie.ac.at/ http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RFold.cgi	[95]
DMfold	RNA structure prediction with pseudoknots	https://github.com/linyuwangPHD/RNA-Secondary-Structure-Database	[96]

doi: [10.3390/ncrna6040049](https://doi.org/10.3390/ncrna6040049)

Bazy danych lncRNA (przykłady)



The screenshot shows the RNAcentral website homepage. At the top, there is a search bar with the text "Search by gene, species, accession, or any keyword" and a "Search" button. Below the search bar, there is a navigation menu with links for "Databases", "Tools", "API", "Downloads", "Browse", "About", "Help", and "Feedback". The main heading is "RNAcentral: The non-coding RNA sequence database", followed by the subtitle "a comprehensive ncRNA sequence collection representing all ncRNA types from a broad range of organisms". Below this, there is a "More about RNAcentral" link. The page is divided into three main sections under the heading "Getting started": "Text search" (with a "Browse sequences" button), "Sequence search" (with a "Search by sequence" button), and "Genome browser" (with an "Explore genomes" button). At the bottom, there is a section titled "Secondary structures" which states that RNAcentral contains over 22 million secondary structures. Below this, there are four boxes showing examples of secondary structures: "Mouse SSU rRNA", "Pig 12S rRNA", "Fly 18S rRNA", and "Plasmodium vivax SSU rRNA".

Bazy danych lncRNA (przykłady)

The screenshot displays the RNAcentral website interface. At the top, there are two buttons: "Explore secondary structures" and "Visualise your sequence with R2DT". Below this, a section titled "Data from 49 databases" features a carousel of database logos and statistics:

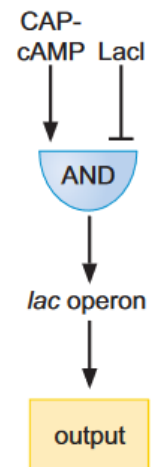
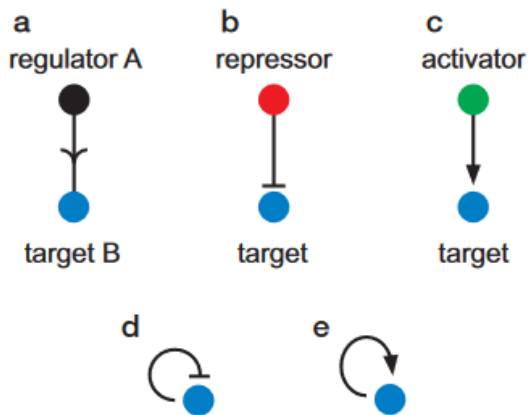
- snoDB**: 1,970 sequences, Example
- LncBase**: 1,151 sequences, Example
- piRBase**: 200,827 sequences, Example
- dictyBase**: 149 sequences, Example
- GENCODE**: 73,117 sequences, Example, Updated

Below the carousel is a "News" section with three columns:

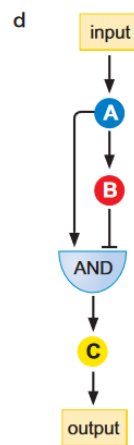
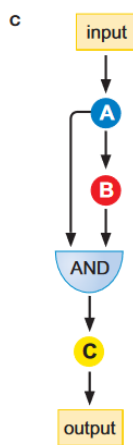
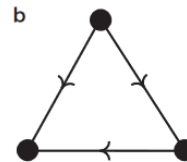
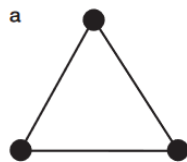
- Tweets**: A tweet from @RNAcentral dated Nov 3, replying to @RNAcentral, discussing fascinating findings and looking forward to future discussions.
- Announcements**: A graphic for "RNAcentral release 21" highlighting 2 new databases (Expression Atlas, snoDB), new data from Ensembl, ENA, RefSeq, and more, and new articles scanned by LitScan. It also includes links to the RNAcentral Blog, RSS feed, and a paper in Current Protocols in Bioinformatics.
- Citing RNAcentral**: A section titled "If you use RNAcentral, please cite our latest paper:" followed by the title "RNAcentral 2021: secondary structure integration, improved sequence search and new member databases" and links to the RNAcentral Consortium, Nucleic Acids Res. 2020, Pubmed, EuropePMC, Full text, and Show abstract.

Regulacja ekspresji: układy

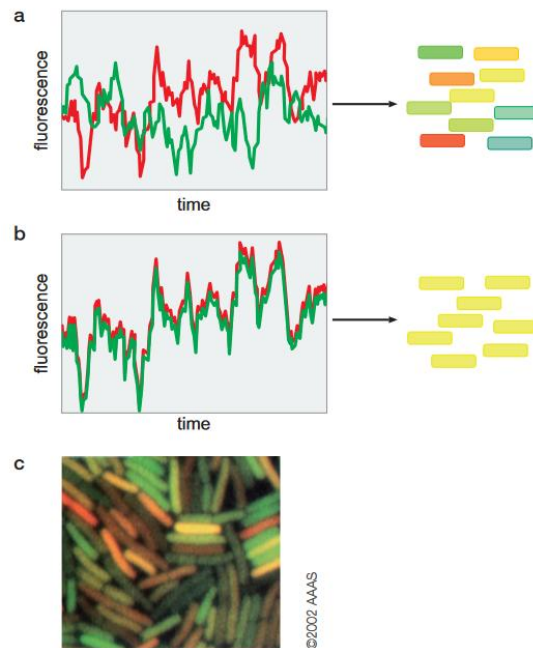
Regulacja ekspresji: układy



Sprzężenie zwrotne

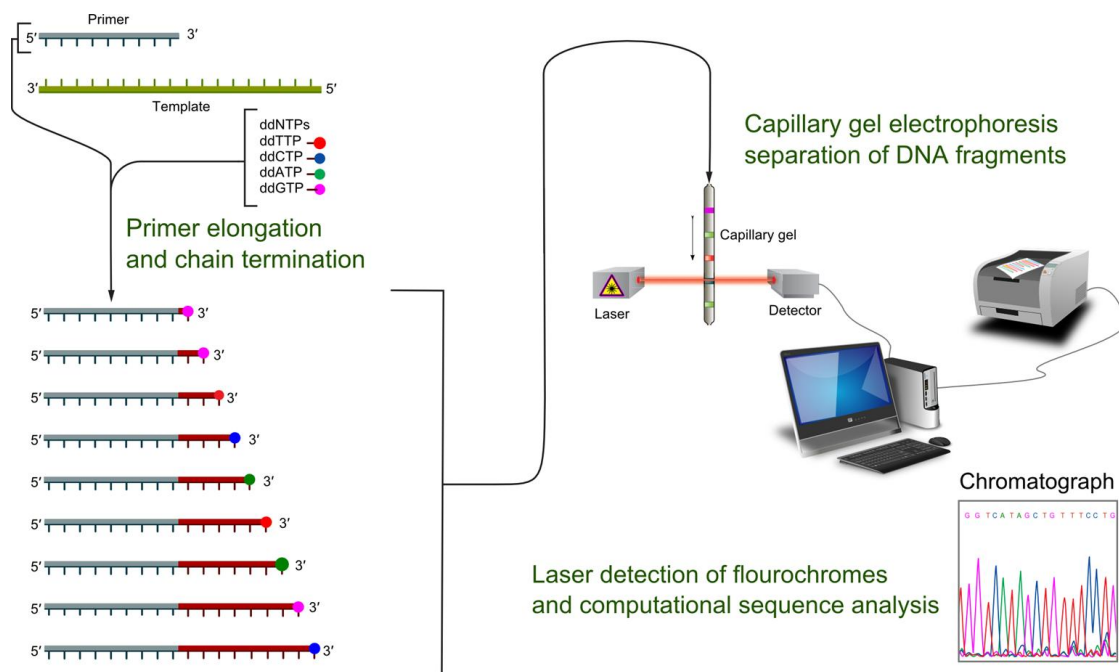


Stochastyczność ekspresji genów



Kilka słów o RNA-Seq

Sekwencjonowanie I generacji



Idea sekwencjonowania nowej generacji

Stworzenie zminiaturyzowanej technologii służącej do sekwencjonowania, która **obniży koszty i zwiększy przepustowość**.

Odbywa się to przez masowe równoległe sekwencjonowanie pofragmentowanej matrycy DNA

Idea sekwencjonowania nowej generacji

Stworzenie zminiaturyzowanej technologii służącej do sekwencjonowania, która **obniży koszty i zwiększy przepustowość**.

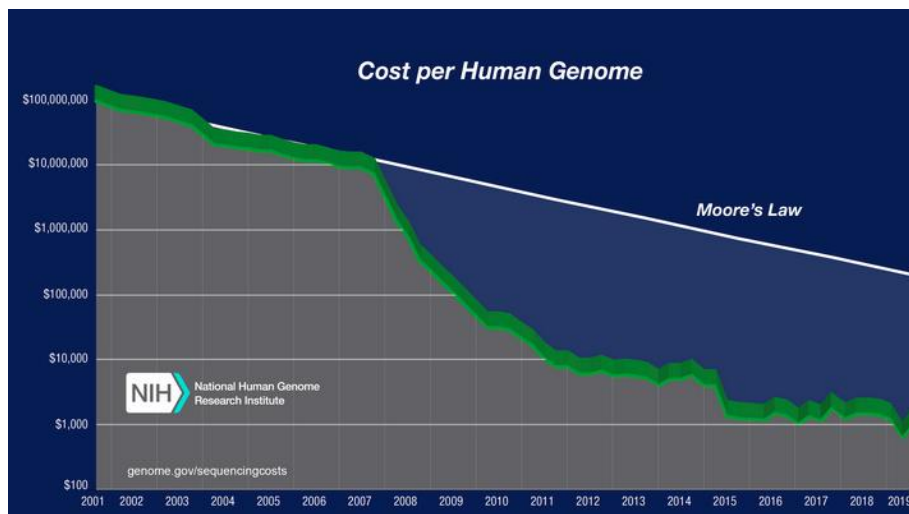
Odbywa się to przez masowe równoległe sekwencjonowanie pofragmentowanej matrycy DNA

Skutek:

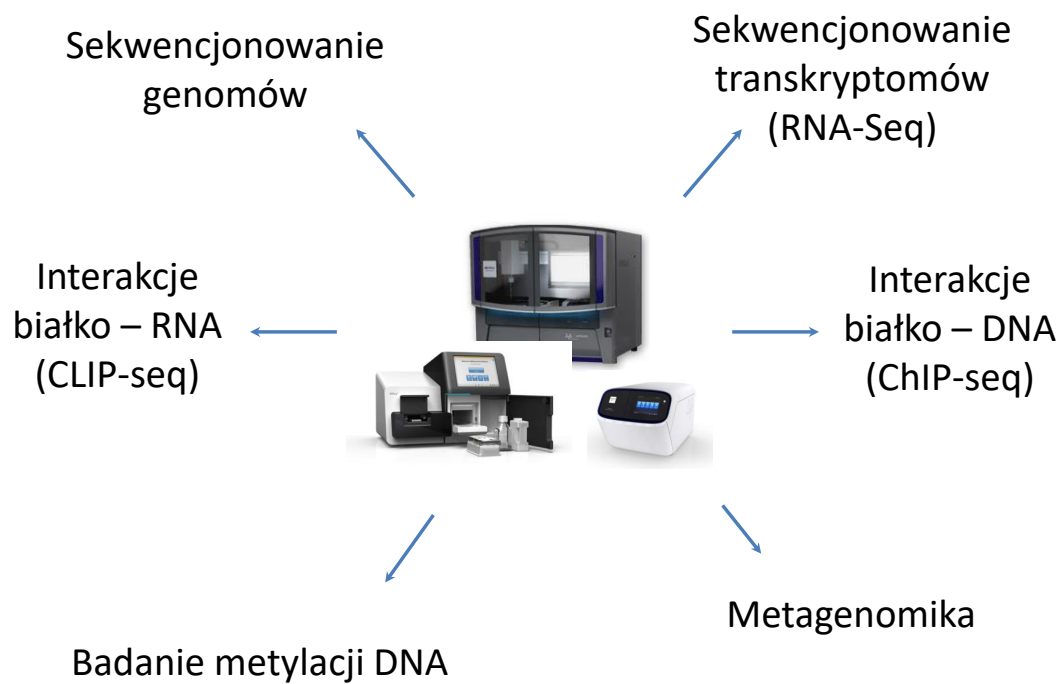
Większość stosowanych obecnie technik, w stosunku do zautomatyzowanej metody Sangera, charakteryzuje się **ogromną liczbą krótkich odczytów**.

Gwałtowny rozwój technologiczny wymusił także rozwój nowych algorytmów oraz formatów plików.

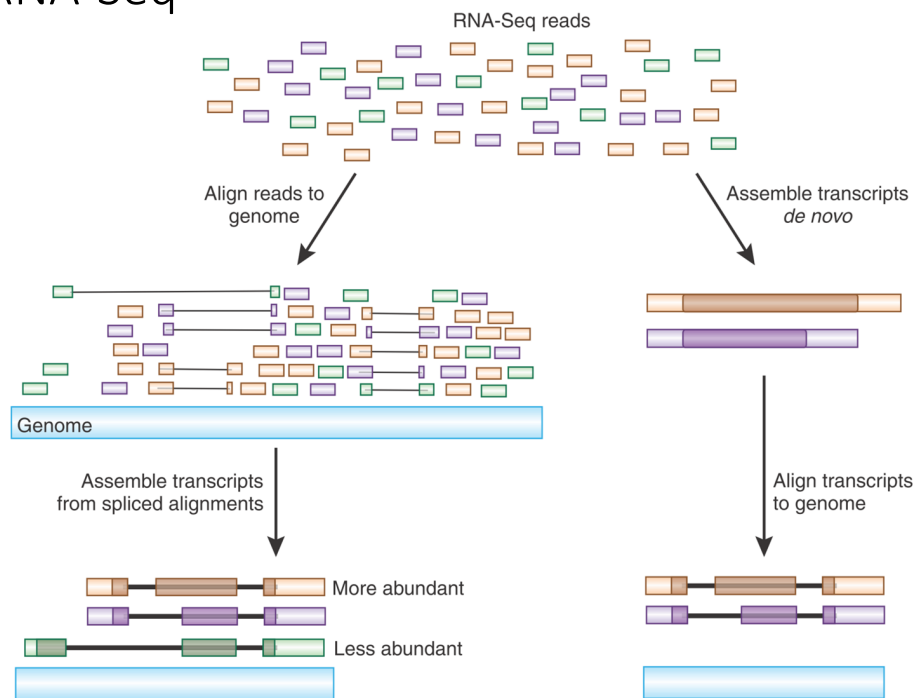
Sekwencjonowanie nowej generacji



<http://www.genome.gov/sequencingcosts/>



Sekwencjonowanie transkryptomów: RNA-Seq



RNA-Seq

Zalety:

- Szeroki wachlarz zastosowań, w tym wykrywanie nowych form splicingowych czy genów
- Lepiej mierzy ekspresję genów niż mikromacierze

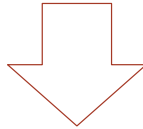
Wady:

- Skrzywienia (np. *primer bias*)
- Odwrotna transkryptaza wprowadza błędy

Cele eksperymentu RNA-seq

STRUKTURA

- Składanie transkryptomów
- Badanie, potwierdzanie struktury genów
- Wykrywanie nowych genów i transkryptów (izoform)



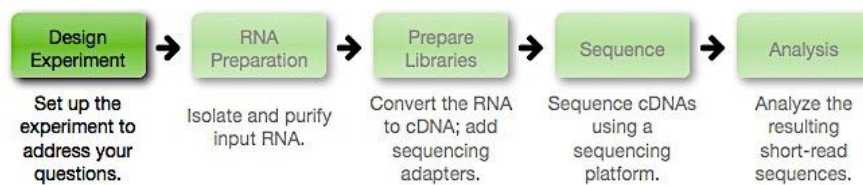
FUNKCJA

- Szacowanie poziomu ekspresji genów i transkryptów
- Ekspresja różnicowa
(np. kontrola vs. stres, kontrola vs. choroba, mutacja)

Sekwencjonowanie transkryptomów

RNA-Seq

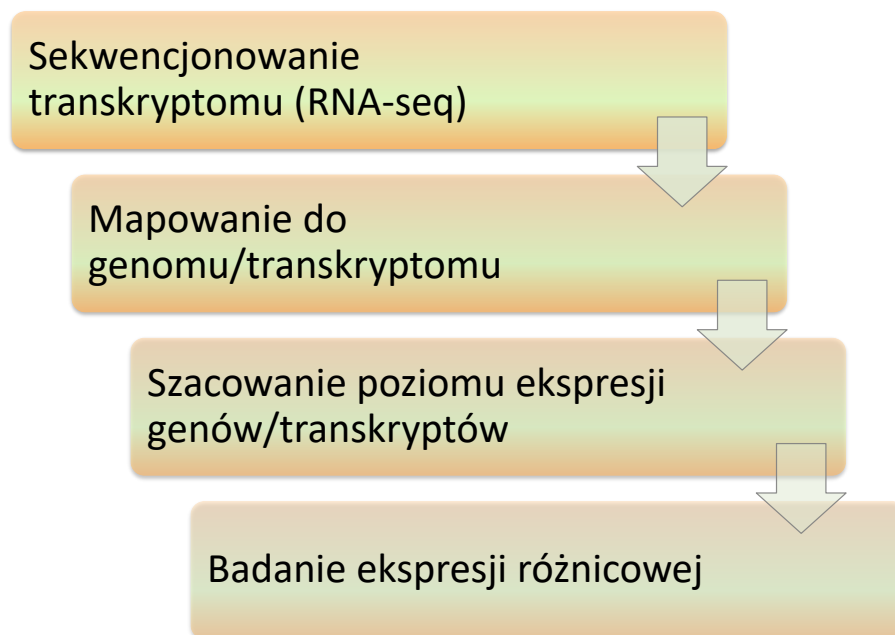
- ✓ Kodujące mRNA
- ✓ Niekodujące RNA → miRNA, małe RNA, długie niekodujące RNA
- ✓ Cały transkryptom



<http://rnaseq.uoregon.edu>

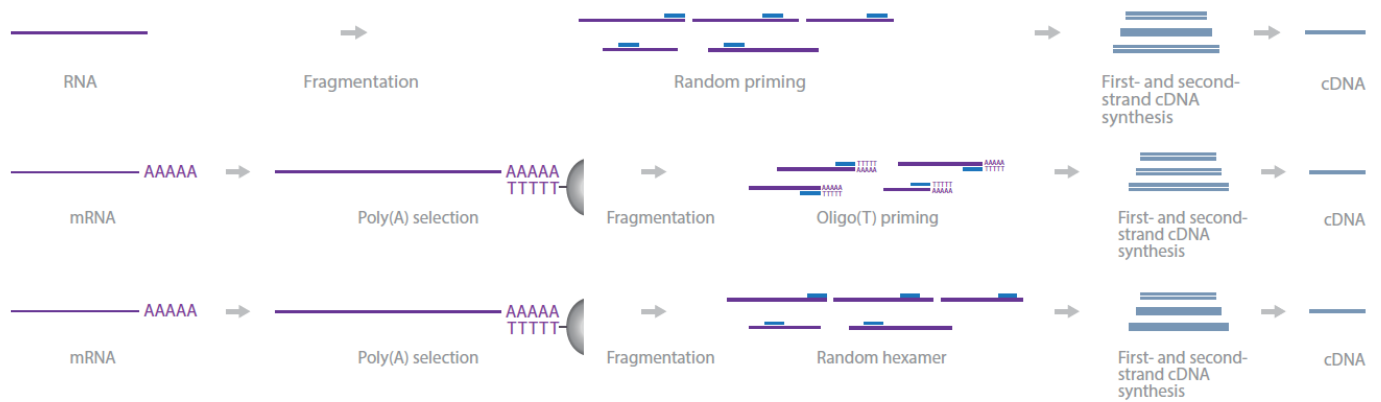
Ogólny schemat działania

Podstawowa ścieżka



Przygotowanie wyników sekwencjonowania

Biblioteki NGS



A schematic overview of RNA-Seq.

Źródło: Illumina

Biblioteki NGS

Główne etapy przygotowywania bibliotek do sekwencjonowania:

- Kontrola jakości RNA
- Fragmentacja i/lub selekcja względem długości
- Selekcja poliA lub usuwanie rRNA
- Odwrotna transkrypcja
- Przyłączenie adapterów do obu końców cDNA
- Kontrola jakości biblioteki

Metody fragmentacji:

- Fizyczna (np. sonikacja)
- Enzymatyczna (np. Dnaza I)
- Chemiczna fragmentacja (kationy metali + ciepło)

RNA-Seq: Illumina

- **Przygotowanie biblioteki**

<https://emea.illumina.com/products/by-type/sequencing-kits/library-prep-kits.html>

- **Generowanie klastrów**

Sekwencjonowany materiał jest immobilizowany na płytce (*flow cell*) → namnożenie poprzez PCR (*solid-phase amplification*), ok. kilkuset kopii

- **Sekwencjonowanie**

SBS: sekwencjonowanie przez syntezę

- **Analiza danych**

Surowe wyniki są w formacie BCL (*binary base call*) → FASTQ. Oprogramowanie Illuminy samo wykonuje tę konwersję, a także demultipleksowanie danych

Oprogramowanie do analizy i kontroli jakości danych NGS

FastQC – wizualna ocena jakości

<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

FASTX-Toolkit – filtrowanie i obróbka pojedynczych odczytów

http://hannonlab.cshl.edu/fastx_toolkit/

Trimmomatic – filtrowanie i obróbka pojedynczych i sparowanych odczytów

<http://www.usadellab.org/cms/?page=trimmomatic>

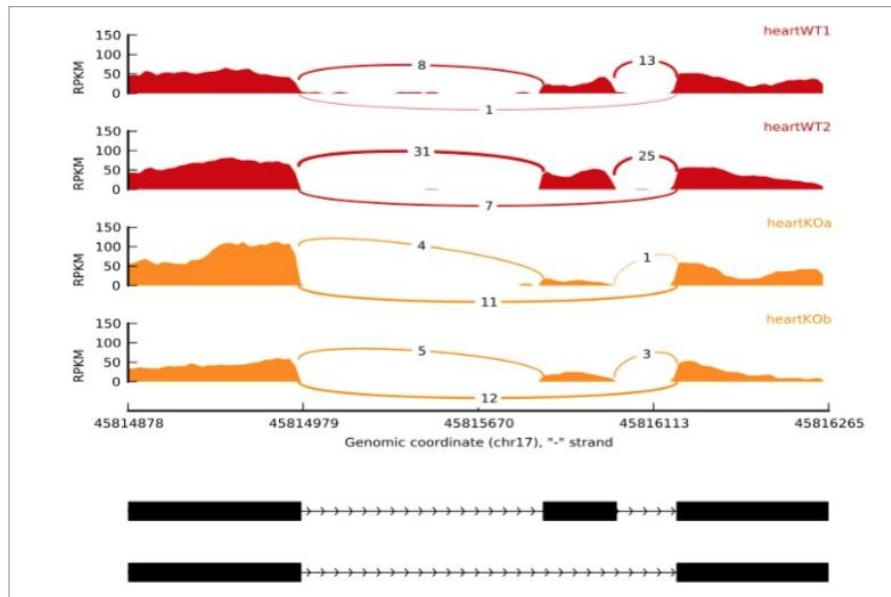
Mapowanie krótkich odczytów

Kroki, które należy wykonać:

- Pobranie sekwencji i adnotacji genomu referencyjnego
- Mapowanie krótkich odczytów do sekwencji referencyjnych (genomu, transkryptomu)
- Dodatkowa obróbka zmapowanych odczytów (plików BAM)

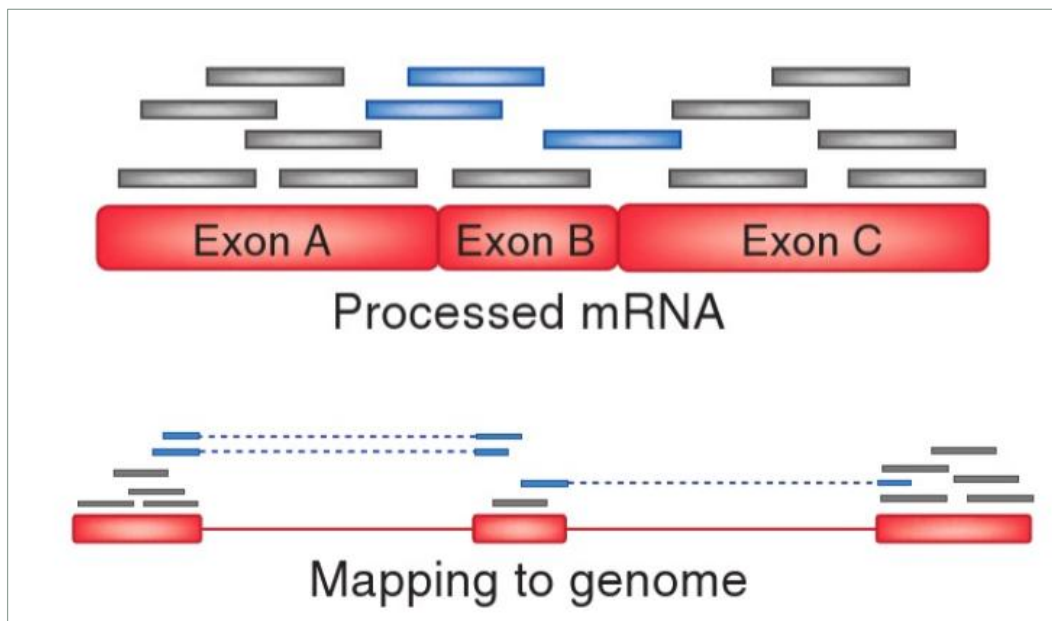
Podstawowe problemy

(1) Jak zmapować krótkie odczyty z uwzględnieniem splicingu?

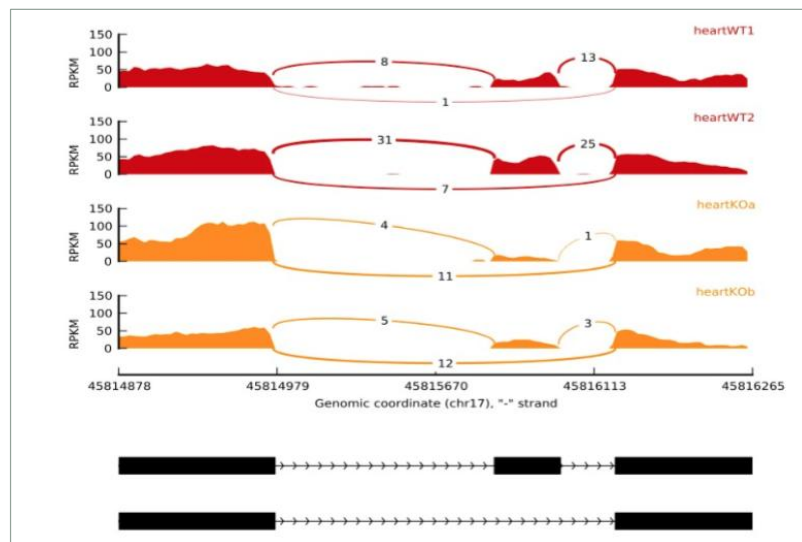


<http://genes.mit.edu/burgelab/miso/>

Mapowanie do transkryptomu



Podstawowe problemy

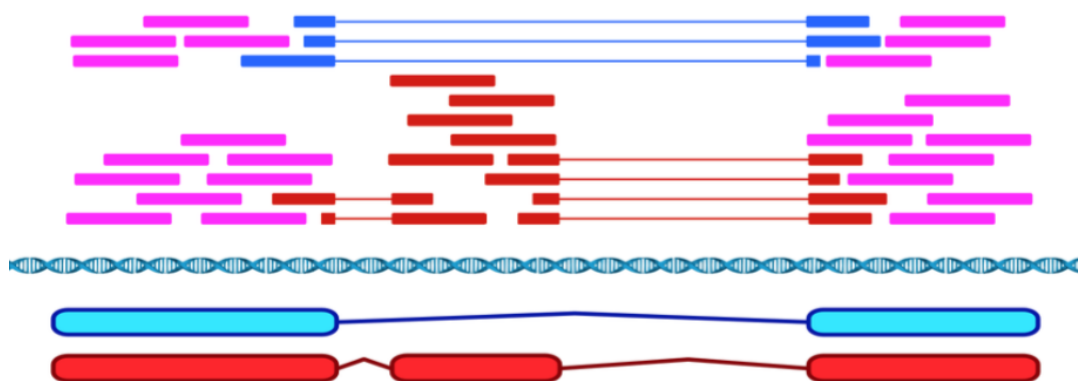


(2) Jak przypisać ekspresję w sytuacji, gdy dany odczyt mapuje się do wielu izoform / bliskich paralogów?

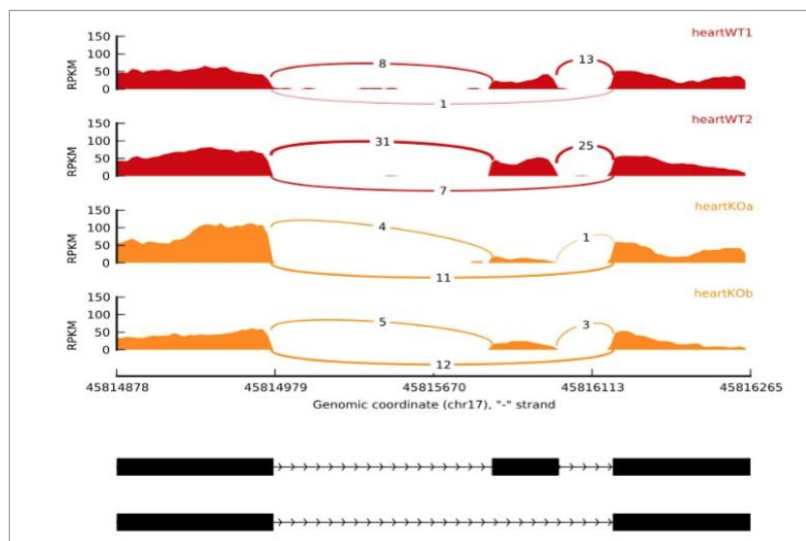
<http://genes.mit.edu/burgelab/miso/>

Wielokrotne zmapowania odczytów

Uniform division – odczyt jest przypisywany równomiernie do wszystkich loci, np. odczyt mapujący się do 10 loci będzie traktowany jako 0.1 odczytu mapującego się do każdego z nich



Podstawowe problemy



(3) Jak znormalizować dane, uwzględniając m.in. liczbę odczytów w danej bibliotece?

<http://genes.mit.edu/burgelab/miso/>

Mapowanie krótkich odczytów

- **Mapowanie bez splicingu (może być wykorzystane do danych RNA-Seq u bakterii)**
 - BWA
 - Bowtie
- **Mapowanie z uwzględnieniem splicingu**
 - STAR
 - HISAT

Przechowywanie zmapowanych odczytów

- Krótkie odczyty zmapowane do sekwencji referencyjnych (genomu lub transkryptomu) przechowywane są w dwóch standardowych formatach: SAM i BAM
- Obydwa formaty plików przechowują dokładnie takie same informacje, ale format BAM jest bardziej wygodny
- Przechowywane dane obejmują lokalizację zmapowanego odczytu oraz dodatkowe informacje, takie jak jakość zmapowania (ang. *mapping quality*, MAPQ) oraz status zmapowania (*concordant* / *discordant*)

Formaty plików

SAM (Sequence Alignment/Map)

- Tabularny plik tekstowy, czytelny dla człowieka
- Duże rozmiary plików (np. dziesiątki GB)
- Zwany czasem TAM (Text SAM format)

BAM (Binary SAM format)

- Dane skompresowane algorytmem BGZF
- Znacznie mniejszy rozmiar pliku
- Oferuje dostęp swobodny (możemy łatwo otrzymać zmapowane odczyty np. z rejonu chr1:1,000,000–2,000,000)

Obróbka plików BAM

Pliki BAM

Operacje na plikach BAM

- Sortowanie pliku BAM
- Tworzenie indeksu pliku BAM
- Oznaczanie zduplikowanych odczytów
- Konwersja BAM → SAM i SAM → BAM

• Oprogramowanie

- Samtools
- <http://www.htslib.org/>
- Bamtools
- <https://github.com/pezmaster31/bamtools>
- Picard
- <http://broadinstitute.github.io/picard/>
- Sambamba
- <http://lomereiter.github.io/sambamba/>

Analiza ekspresji

Normalizacja ekspresji

- Liczba odczytów mapujących się do każdego *locus* jest proporcjonalna do wielkości biblioteki
- Mimo nadziei, że dane RNA-Seq nie będą wymagały intensywnej normalizacji, okazało się, że liczba odczytów zależy m.in. od zawartości par GC i długości transkryptu
- *GC bias* jest uwzględniany w większości programów z użyciem odpowiednich modeli statystycznych
- Powstała potrzeba nowych jednostek ekspresji genów

Jednostki ekspresji

1. RPKM (ang. reads per kilobase per million mapped reads)
2. FPKM (ang. fragments per kilobase per million mapped reads) – alternatywna nazwa, fragment oznacza 2 sparowane odczyty

$$\frac{\text{liczba zmapowanych odczytów}}{\text{długość locus (kbp)} * \text{całkowita liczba zmapowanych odczytów (mln)}}$$

3. TPM (ang. transcript per million) – ilość transkryptów danego rodzaju na milion transkryptów w danej próbce
4. expected_count – ilość odczytów mapująca się do danego transkryptu nie poddana normalizacji – jest to preferowana forma danych dla większości programów do analizy ekspresji różnicowej

RSEM

- Program RSEM mapuje odczyty do transkryptomu lub genomu używając programu **Bowtie** i na podstawie tych zmapowań szacuje, ocenia i normalizuje ekspresję poszczególnych genów i transkryptów.
- Można dostosować algorytm i parametry mapowania w zależności od analizy, wykorzystać wcześniejsze zmapowania.
- Domyślnie program zakłada, że odczyty są niesparowane (single-end) z informacją o jakości (quality score)
- Nie wykrywa nowych izoform splicingowych
- Zawansowane algorytmy statystyczne (m.in. EM), wysoka korelacja uzyskanych wyników z rzeczywistymi wartościami ekspresji, dokładność
- Działa stosunkowo wolno

RSEM

- Bardziej znany
- Wolniejsze działanie
- Większa dokładność (EM)
- Konieczne pełne mapowanie (BAM)
- Geny i transkrypty
- TPM, FPKM, expected count

Salmon

- Nowszy, mniej znany
- Szybsze działanie
- Mniejsza dokładność, inny algorytm
- Quasi-mapowanie (k-mery)
- Geny i transkrypty
- TPM, NumReads

Jeśli chcemy zliczyć liczbę odczytów przypadających na konkretny egzon, transkrypt lub gen – bez zawiłych algorytmów – można użyć BEDTools lub HTSeq.

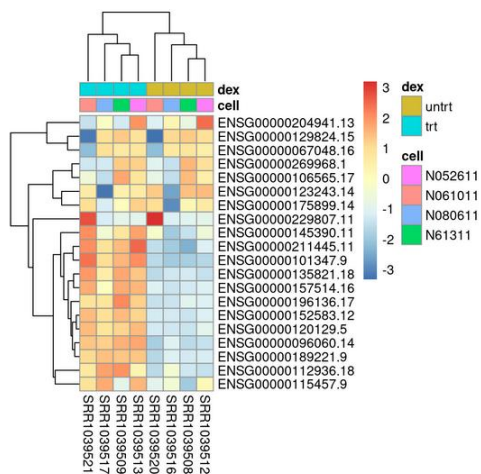
Analiza ekspresji różnicowej

- Polega na identyfikacji genów, których ekspresja zmienia się istotnie statystycznie pomiędzy badanymi warunkami
- Powtórzenia biologiczne
- Normalizacja danych
- Kontrola jakości: wykresy diagnostyczne

Analiza ekspresji różnicowej

Da najpopularniejszych narzędzi można zaliczyć:

- DESeq2
- Limma
- edgeR
- EBSeq



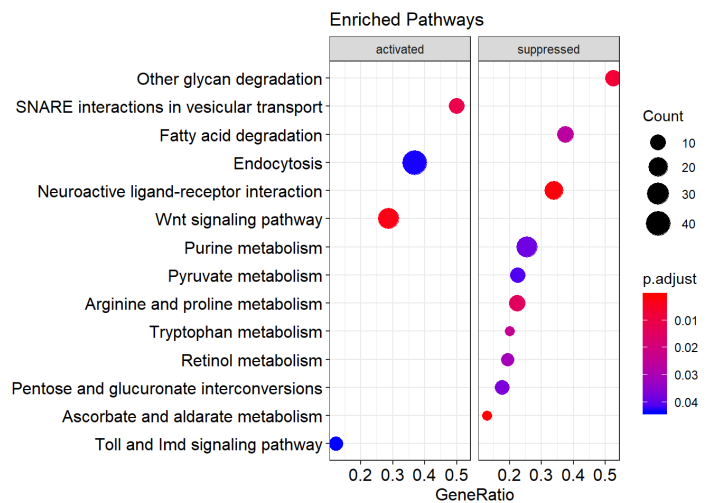
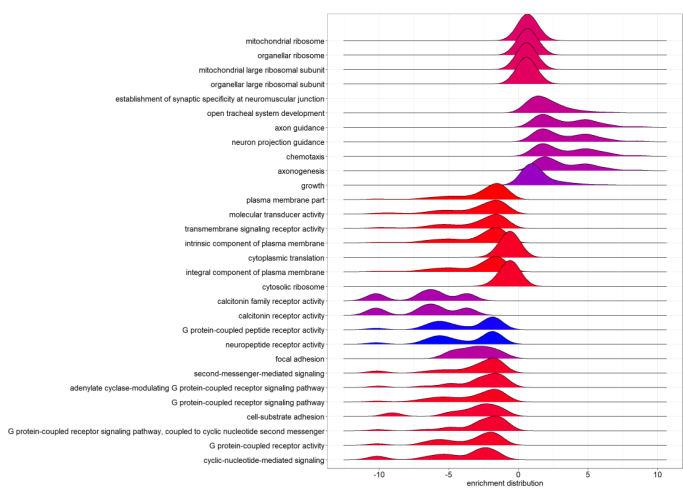
GSEA: gene set enrichment analysis

- Polega na identyfikacji terminów (np. GO) czy ścieżek metabolicznych/sygnalizacyjnych (np. KEGG) pośród genów wykazujących DE
- Ważna jest odpowiednia dobra adnotacja transkryptomów
- Podobnie jak w przypadku DE, tutaj również używa się głównie narzędzi Bioconductor, np. **ClusterProfiler**

GSEA: gene set enrichment analysis

- Polega na identyfikacji terminów (np. GO) czy ścieżek metabolicznych/sygnalizacyjnych (np. KEGG) pośród genów wykazujących DE
- Ważna jest odpowiednia dobra adnotacja transkryptomów
- Podobnie jak w przypadku DE, tutaj również używa się głównie narzędzi Bioconductor, np. **ClusterProfiler**

GSEA: gene set enrichment analysis



GSEA: gene set enrichment analysis

Inne narzędzia:

- GOrilla (<http://cbl-gorilla.cs.technion.ac.il/>); tylko *gene ontology*
- PlantGSEA (<http://structuralbiology.cau.edu.cn/PlantGSEA/>); *gene ontology, pathways*
- DAVID (<https://david.ncifcrf.gov/>); różne analizy
- TopGO (R: <http://www.bioconductor.org/packages/release/bioc/html/topGO.html>)
- GSEAServer (plantgrn.noble.org/GSEAServer); non-model organisms
- KAAS (<http://www.genome.jp/tools/kaas/>); adnotacja transkryptomu

Przykładowe narzędzia online: DAVID



DAVID Bioinformatics Resources
Laboratory of Human Retrovirology and Immunoinformatics (LHRI)



[Home](#) | [Start Analysis](#) | [Shortcut to DAVID Tools](#) | [Technical Center](#) | [Downloads & APIs](#) | [Term of Service](#) | [About DAVID](#) | [About LHRI](#)

Overview

The **Database for Annotation, Visualization and Integrated Discovery (DAVID)** provides a comprehensive set of functional annotation tools for investigators to understand the biological meaning behind large lists of genes. These tools are powered by the comprehensive **DAVID Knowledgebase** built upon the DAVID Gene concept which pulls together multiple sources of functional annotations. For any given gene list, DAVID tools are able to:

- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related many-genes-to-many-terms on 2-D view.
- Search for other functionally related genes not in the list
- List interacting proteins
- Explore gene names in batch
- Link gene-disease associations
- Highlight protein functional domains and motifs
- Redirect to related literatures
- Convert gene identifiers from one type to another.
- And more

Hot Links

Multiple positions available in LHRI

The Laboratory of Human Retrovirology and Immunoinformatics (LHRI) has collaborated with the National Institute of Allergy and Infectious Diseases (NIAID) and supported NIAID clinical trials for patients infected with HIV mutants resisting anti-retroviral therapy. LHRI has isolated the multiple-class drug-resistant (MDR) variants from patients and characterized each variant's drug sensitivity and infectivity. The study aims to define salvage therapy and develop novel therapy (chemotherapy and immunotherapy). During the investigation, LHRI has characterized the emergence of novel mutations on drug susceptibility and viral replication. LHRI is a pioneer in researching the anti-viral cytokine, Interleukin-27, DNA-repair protein (Ku70)-mediated innate immune response against HIV and other virus co-infection, and novel subsets of immune cells. LHRI maintains the Database for Annotation, Visualization and Integrated Discovery (**DAVID**).

(1) **Scientist I - Virology position** available to perform the defective proviral study in our **Basic Research Section**.

(2) **Scientist-Cytokines and HIV** available in our **Basic Research Section**. We are looking for a cytokine immunologist who is interested in virus (HIV/ HSV/KHSV) pathogenesis in myeloid immune cell types (macrophages, dendritic cells and microglia cells).

(3) **Postdoctoral Fellow** available in our **Basic Research Section**. This position is an excellent opportunity for a young Ph.D. who has no experience in virus research and seeks a career in a new research field. You will learn how to handle infectious RNA viruses and investigate the mechanism of virus infection and the interaction of host cell proteins using HIV (lentivirus) variants under an SOP following the NIAID guideline in the BSL2* laboratory. LHRI does not consider virus experience in the past, instead seeks a highly motivated researcher. The knowledge learned in the lab can apply to future COVID or other virus studies.

(4) **Bioinformatics Analyst II** available to perform bioinformatics analyses and develop bioinformatics analysis pipelines in our **Bioinformatics Section**.

DAVID Forum
Forum for DAVID users to ask questions, suggest new functions and help other users by answering their questions.

Przykładowe narzędzia online: ConsensusPathDB



Max Planck Institute
for Molecular Genetics



home
content information
search

- interactions of molecules/pathways
- shortest interaction paths

gene set analysis

- over-representation analysis
- enrichment analysis
- induced network modules

metabolite set analysis

download / data access

documentation



human yeast mouse

Release 35 (05.06.2021)

ConsensusPathDB-human integrates interaction networks in *Homo sapiens* including **binary and complex protein-protein, genetic, metabolic, signaling, gene regulatory and drug-target** interactions, as well as biochemical pathways. Data originate from currently **31** public resources for interactions (listed below) and interactions that we have curated from the literature. The interaction data are integrated in a complementary manner (avoiding redundancies), resulting in a seamless interaction network containing different types of interactions.



Literature:
Herwig, R. *et al.*. Analyzing and interpreting genome data at the network level with ConsensusPathDB. (2016) *Nature Protocols* 11, 1889-1907.
Kamburov, A. *et al.* (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.*
Kamburov, A. *et al.* (2011) ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res.*
Kamburov, A. *et al.* (2009) ConsensusPathDB—a database for integrating human interaction networks. *Nucleic Acids Res.*
Pentchev, K. *et al.* (2010) Evidence mining and novelty assessment of protein-protein interactions with the ConsensusPathDB plugin for Cytoscape. *Bioinformatics*

Current statistics:

unique physical entities:	200,499
unique interactions:	859,848
gene regulations:	18,912
protein interactions:	616,304
genetic interactions:	7,936
biochemical reactions:	25,046
drug-target interactions:	191,650
pathways:	5,578

Przykładowe narzędzia
online: ConsensusPathDB

	name	protein	signaling	metabolic	gene	genetic	drug-target	biochemical
		interactions	reactions	reactions	regulations	interactions	interactions	pathways
	BioCarta	✓	✗	✗	✓	✗	✗	✗
	BioGrid	✓	✗	✗	✗	✓	✗	✗
	MIPS	✓	✗	✗	✗	✗	✗	✗
	ChEMBL	✗	✗	✗	✗	✗	✓	✗
	DIP	✓	✗	✗	✗	✗	✗	✗
	EHMN	✗	✗	✓	✗	✗	✗	✓
	HPRD	✓	✗	✗	✗	✗	✗	✗
	HumanCyc	✗	✗	✓	✗	✗	✗	✓
	INOH	✗	✓	✓	✗	✗	✗	✓
	InnateDB	✓	✓	✗	✓	✗	✗	✗
	IntAct	✓	✗	✗	✗	✗	✗	✗
	KEGG	✗	✓	✓	✗	✗	✓	✓
	MINT	✓	✗	✗	✗	✗	✗	✗
	MIPS-MPPI	✓	✗	✗	✗	✗	✗	✗
	MatrixDB	✓	✗	✗	✗	✗	✗	✗
	NetPath	✓	✓	✗	✗	✗	✗	✓
	PDB	✓	✗	✗	✗	✗	✗	✗
	PDZBase	✓	✗	✗	✗	✗	✗	✗
	PID	✗	✓	✗	✓	✗	✗	✓
	PIG	✓	✗	✗	✗	✗	✗	✗
	PINdb	✓	✗	✗	✗	✗	✗	✗
	PharmGKB	✗	✗	✗	✗	✗	✓	✓
	PhosphoPOINT	✓	✓	✗	✗	✗	✗	✗
	PhosphoSitePlus	✗	✓	✗	✗	✗	✗	✗
	Reactome	✓	✓	✓	✗	✗	✗	✓
	SMPDB	✗	✗	✗	✗	✗	✗	✓
	Signalink	✗	✗	✗	✗	✗	✗	✓
	Spike	✓	✓	✗	✓	✗	✗	✗
	TTD	✗	✗	✗	✗	✗	✓	✗
	Wikipathways	✗	✓	✓	✗	✗	✗	✓
	ConsensusPathDB	✓	✓	✓	✓	✓	✓	✓

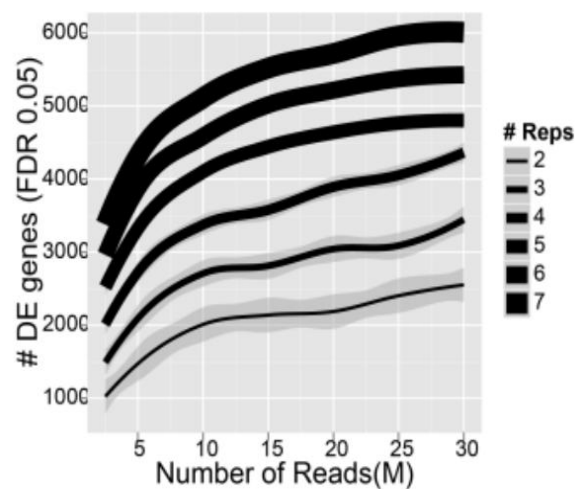
Jakiego pokrycia potrzebujemy?

Category	Detection or Application	Recommended Coverage (x) or Reads (millions)	References
Transcriptome Sequencing	Differential expression profiling	10-25M	Liu Y. et al., 2014; ENCODE 2011 RNA-Seq
	Alternative splicing	50-100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
	Allele specific expression	50-100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
	De novo assembly	>100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
Small RNA (microRNA) Sequencing	Differential Expression	~1-2M	Metpally RPR et al., 2013; Campbell et al., 2015
	Discovery	~5-8M	Metpally RPR et al., 2013; Campbell et al., 2015

Dla porównania...

Category	Detection or Application	Recommended Coverage (x) or Reads (millions)	References
Whole genome sequencing	Homozygous SNVs	15x	Bentley et al., 2008
	Heterozygous SNVs	33x	Bentley et al., 2008
	INDELs	60x	Feng et al., 2014
	Genotype calls	35x	Ajay et al., 2011
	CNV	1-8x	Xie et al., 2009; Medvedev et al., 2010
Whole exome sequencing	Homozygous SNVs	100x (3x local depth)	Clark et al., 2011; Meynert et al., 2013
	Heterozygous SNVs	100x (13x local depth)	Clark et al., 2011; Meynert et al., 2013
	INDELs	not recommended	Feng et al., 2014

Ilu powtórzeń biologicznych potrzebujemy?



Źródło: Liu Y, Zhou J, White KP. *RNA-seq differential expression studies: more sequence or more replication?* Bioinformatics. 2014

Składanie transkryptomów *de novo*

Trinity

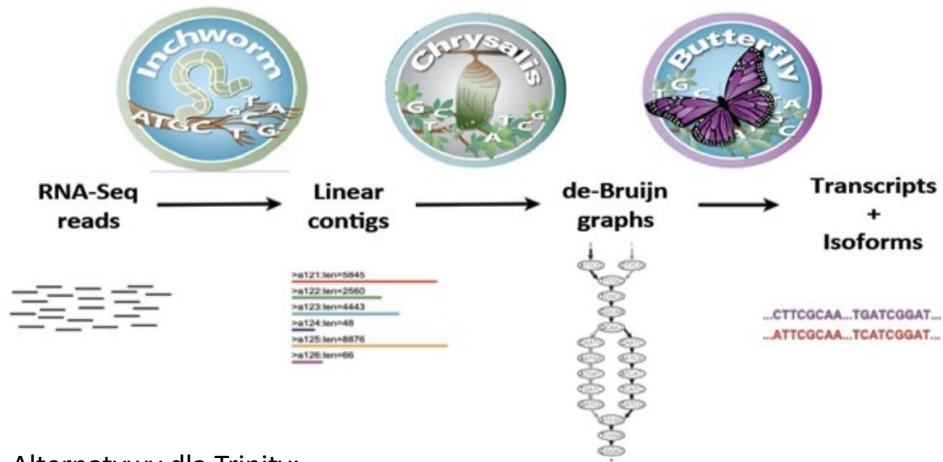
<https://github.com/trinityrnaseq/trinityrnaseq/wiki>

Pakiet trzech odrębnych programów uruchamianych jeden po drugim:

- (1) **Inchworm** – tworzy wstępne kontigi z odczytów RNA-seq
- (2) **Chrysalis** – klasteryzuje powstałe kontigi i tworzy grafy de Bruijn'a dla każdego klastra
- (3) **Butterfly** – analizuje ponownie stworzone klastry biorąc pod uwagę, że odczyty mogą pasować w wiele miejsc (alternatywne formy splicingowe, paralogi itp.)



Trinity



Alternatywy dla Trinity:

- Velvet
- Oases
- Rnnotator
- Trans-ABYSS

<http://slideplayer.com/slide/8071095/>

TransDecoder

<https://transdecoder.github.io/>

- Program, który identyfikuje potencjalne rejony kodujące
- Działa dobrze dla transkryptów składanych *de novo*
- Umożliwia identyfikację ORF,
- Umożliwia poszukiwanie sekwencji homologicznych,

Trinotate

Narzędzie do automatycznej adnotacji transkryptomów (w szczególności złożonych de novo)

Wykorzystuje szereg uznanych w środowisku metod do adnotacji funkcjonalnej transkryptów, takich jak:

- BLAST+/SwissProt do wyszukiwania sekwencji homologicznych
- HMMER/PFAM do identyfikacji domen białkowych
- tmHMM/signalP do identyfikacji białek transbłonowych i peptydów sygnałowych
- GO (Gene Ontology), KEGG

RNA-Seq

Aplikacje

RNA-Seq: zastosowanie

RNA-Seq znajduje szerokie zastosowanie w badaniach biomedycznych i medycynie, m.in. w następujących dziedzinach:

1. Choroby zakaźne
2. Nowotwory
3. Transplantologia
4. Medycyna prenatalna

RNA-Seq: zastosowanie

RNA-Seq pozwala badać szeroki zakres transkryptów, w tym:

- mRNA, w tym izoformy splicingowe
- Niekodujące RNA
- RNA patogenów
- Geny fuzyjne

W szczególności, możemy:

- Wykrywać nieznane wcześniej transkrypty
- Obliczać poziom ekspresji transkryptów
- Wykrywać mutacje
- Badać zjawisko edytowania RNA
- Badać pozakomórkowe RNA

RNA-Seq vs mikromacierze

Przewaga RNA-Seq nad mikromacierzami:

- Wyższa precyzja, zwiększony zakres wykrywanych poziomów ekspresji
- Możliwość mierzenia i detekcji nowych izoform splicingowych, fuzji genowych itd.
- Znajduje zastosowanie dla każdego osobnika, każdego gatunku

Wadą jest stosunkowo większa trudność w przygotowaniu próbki i samym procesie sekwencjonowania.

Ekspresja mRNA

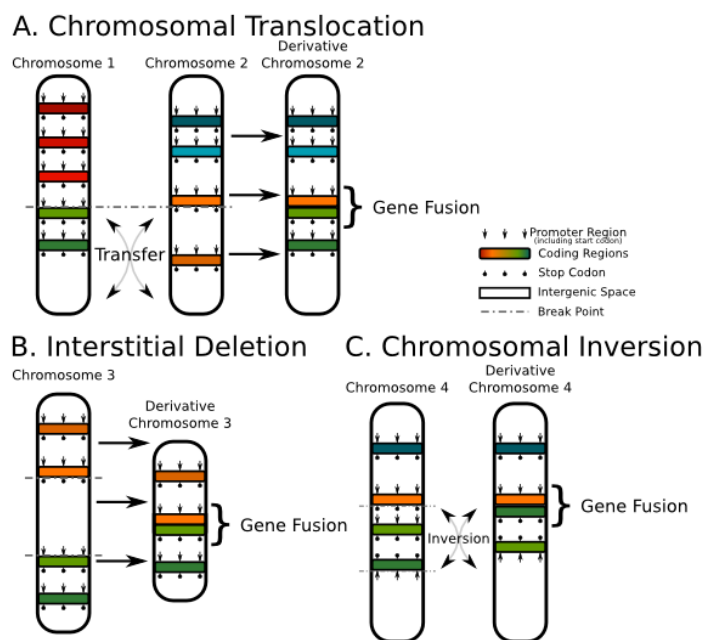
Przykład

W nowotworze piersi zaleca się stosowanie markerów prognostycznych w formie mRNA wybranych genów, które mogą wspomóc decyzję dot. leczenia, obok danych biochemicznych i histopatologicznych. Przykładem jest test OncotypeDx, pozwalający na mierzenie poziomów ekspresji mRNA 21 genów.

Ekspresja mRNA

- **AlloMap** jest nieinwazyjnym testem, polegającym na pomiarze poziomu ekspresji 20 genów, w dużej mierze powiązanych z układem odpornościowym i ścieżkami sygnalizacyjnymi. Stosowany jest u pacjentów po przeszczepie serca, by ocenić ryzyko odrzucenia przeszczepu.
- W opracowaniu jest wiele innych testów, na przykład pomiar ekspresji genów powiązanych z układem odpornościowym i antygenów w połączeniu z Exome-Seq, co ma na celu badanie efektów leczenia w melanomie.

Fuzje genowe



Źródło: <https://en.wikipedia.org>

Fuzje genowe

- Fuzje genowe powiązane z patogenezą szeregu nowotworów, a niejednokrotnie są charakterystyczne dla poszczególnych podtypów nowotworów (np. NSCLC). Na przykład według WHO diagnoza AML (*acute myeloid leukaemia*) może być wykonana w oparciu o aberracje genetyczne, wliczając właśnie fuzję genów *RUNX1–RUNX1T1*.
- Fuzje genowe są także wykorzystywane jako markery prognostyczne i wykorzystywane są jako biomarkery pozwalające na szybki screening w kierunku ryzyka zachorowania na dany rodzaj nowotworu, jak na przykład fuzja genów *TMPRSS2–ERG* w nowotworze prostaty.

Fuzje genowe

Niemniej jednak, poza nielicznymi przypadkami, aktualne wytyczne rzadko rekomendują wykrywanie fuzji genowych z wykorzystaniem RNA-Seq. Na przykład w NSCLC, fuzja *EML4–ALK* jest zazwyczaj wykrywana z wykorzystaniem techniki FISH (fluorescence *in situ* hybridization) i/lub testami immunohistochemicznymi skierowanymi na białko.

Alternatywny splicing

- Alternatywny splicing powiązано z licznymi chorobami, wliczając choroby rozwojowe, neurodegeneracyjne czy nowotwory. Uważa się także, że obecność niektórych form splicingowych może nieść wartość diagnostyczną/terapeutyczną. Na przykład ekspresja jednego z wariantów splicingowych genu AR-V7 (receptor androgenowy) wykryto w komórkach nowotworowych ok. 30% mężczyzn w jednym z podtypów nowotworu prostaty i powiązana jest z obniżoną efektywnością terapii naczelnowanych na receptory androgenowe. Z kolei jedna izoforma genu EGFR (*epidermal growth factor receptor*) jest intensywnie badanym celem terapeutycznym.
- Specyficzne alternatywne formy splicingowe genu BRCA1 odkryto u pacjentów, w których rodzinach występowały wcześniej nowotwory piersi lub jajnika. Co ciekawe, u tych pacjentów nie wykryto klasycznych, znanych mutacji w genach BRCA1 i BRCA2.

Ekspresja allelospecyficzna

Ekspresja allelospecyficzna (z jednego allelu; ASE – *allele-specific expression*) może być efektem działania wielu różnych czynników, wliczając imprinting, inaktywację chromosomu X czy też inne mechanizmy epigenetyczne. Przykładem może być syndrom Angelmana, choroba neurodegeneracyjna, gdzie wyciszona jest ekspresja genu UBE3A z chromosomu macierzystego w mózgu. Z mechanizmem tym powiązane są antysensowne lncRNA.

Pozakomórkowe RNA

- Badania nad exRNA to szybko rozwijająca się gałąź diagnostyki medycznej. exRNA są uwalniane z komórek i są chronione przed degradacją przez to, że występują w pecherzykach, które je transportują, lub są powiązane z RBP i lipoproteinami.
- Mierzenie poziomów ekspresji exRNA jest obiecującą, nieinwazyjną strategią diagnostyczną. Przede wszystkim, biofluidy są łatwiej osiągalne, niż tkanki. Zawierają też mieszanę transkryptów różnych tkanek, na przykład w osoczu można wykrywać transkrypty specyficzne dla mózgu. Konsekwencją jest jednak obniżona specyficzność tkankowa tego podejścia. Niemniej jednak biofluidy różnią się pod tym względem: na przykład w osoczu występują szczególnie duże ilości transkryptów pochodzących z wątroby i serca, z kolei ślina bogata jest w transkrypty ślinianek i przełyku.

Pozakomórkowe RNA

- Wykorzystanie exRNA w diagnostyce nowotworów jest analogiczne do exDNA, z tą różnicą, że geny niosące daną mutację muszą ulegać ekspresji. Dodatkowo, exRNA pozwala na pomiar ekspresji, co daje wgląd w poziom zaawansowania choroby, czego nie można zrobić z wykorzystaniem exDNA.
- exRNA to zarówno szereg mRNA, jak i różne niekodujące RNA.
- Firmy, takie jak Exosome Diagnostics, rozwijają testy diagnostyczne w oparciu o exRNA, które mają pozwolić wykrywać w sposób nieinwazyjny z osocza krwi fuzje genowe (np. *EML4–ALK*) oraz mutacje (np. *EGFR* T790M). Pierwszym takim testem jest **ExoDk**, mierzący poziom ekspresji transkryptów *EML4–ALK* wyizolowanych z egzosomów osocza pacjentów z NSCLC.

Patogenne RNA w osoczu

Małe RNA produkowane przez patogeny i wykrywane w osoczu jako exRNA stanowią obiecujący kierunek badań. Na przykład w przypadku infekcji *Neisseria meningitidis* obserwuje się szereg małych RNA w osoczu, które mogą służyć do diagnostyki tej infekcji.

Niekodujące RNA: circRNA

circRNA są niezwykle interesującym obiektem badań i znamy ich liczne funkcje regulatorowe, jednak jak dotąd nieliczne są doniesienia wiążące je z patogenezą chorób. Niemniej jednak, niektóre circRNA występują w dużych ilościach w płynach ciała i tkankach, a do tego są zwykle bardziej stabilne niż mRNA, co zwiększa zainteresowanie ich zastosowaniem do testów diagnostycznych.

Wirusy

- Wirusy RNA, takie jak HIV, Ebola, dengi, wirusy zapalenia wątroby (A, D, E), SARS, MERS, COVID-19) zazwyczaj wykrywane są z wykorzystaniem qRT-PCR. Ocenia się, że w przyszłości może tutaj szerzej być wykorzystywana technika RNA-Seq.
- Pewien wariant RNA-Seq, nazwany metagenomicznym RNA-Seq, może być wykorzystywany do wszechstronnej diagnostyki pod kątem wirusów.
- RNA-Seq był z powodzeniem wykorzystany by badać początki epidemii i transmisję zakażeń wirusem ebola w zachodniej Afryce w 2014 roku.

Diagnostyka COVID-19



The Illumina COVIDSeq Test is intended for detection of SARS-CoV-2 virus RNA in authorized countries (Canada, Japan, Philippines, South Africa, and U.S.) under each country's respective authorizing agency.* A [Research Use Only \(RUO\) version of this product](#) is available for non-authorized countries.



Illumina
COVIDSeq Test

Illumina COVIDSeq Test

This high-throughput, next-generation sequencing test detects SARS-CoV-2 in patients suspected of COVID-19 and enables virus genome analysis in research use. [Read More...](#)

Key Features and Benefits

- **Accurate:** Detects 98 targets on SARS-CoV-2 for highly accurate detection
- **Comprehensive:** Detects SARS-CoV-2 virus RNA; reports consensus genome calling for virus analysis under research use
- **Quality Controlled:** Built in quality control features in every reaction
- **Flexible:** Seamless end-to-end workflow. Temperature specification allows use of different thermocyclers.
- **Scalable:** Available on NovaSeq 6000, NextSeq 2000, or NextSeq 500/550/550Dx (in RUO mode) systems.

Wyzwania

Istnieją trzy główne trudności od strony analitycznej (bioinformatycznej) na drodze ku rozpowszechnieniu wykorzystania analizy danych RNA-Seq w diagnostyce medycznej:

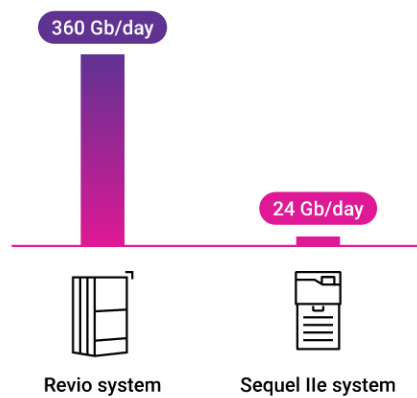
- Brak standardów
- Duża liczba oprogramowania do wyboru; wiele parametrów analizy
- Złożone potoki analityczna składają się z narzędzi, które powstały niezależnie: problemy ze spójnością i licencjami

NGS w medycynie: ograniczenia

- Badania podstawowe nie zawsze idą w parze z realiami klinicznymi, jeśli chodzi o skuteczność biomarkerów
- Trudności z przejściem z badań podstawowych/rozwojowych na poziom komercyjny
- Innym ograniczeniem jest brak regulacji lub niepełne regulacje ze strony odpowiednich organów, jak np. FDA (USA), Ministerstwo Zdrowia itp.
- Szybko zmieniająca się technologia i wiedza
- Źródło materiału (np. bloczki parafinowe – degradacja RNA; lepiej – mrożone w ciekłym azocie) = ograniczenia technologiczne
- Heterogenność próbek
- Przygotowanie biblioteki, sekwencjonowanie i analiza danych (częściowe rozwiązanie: klaster obliczeniowy, GPU, chmura)
- Przechowywanie danych (rozmiar, bezpieczeństwo)
- Problemy natury etycznej (np. zsekwencjonowane komórki Hela – a rodzina; w 2013 r. wycofano wyniki sekw. z ENA)

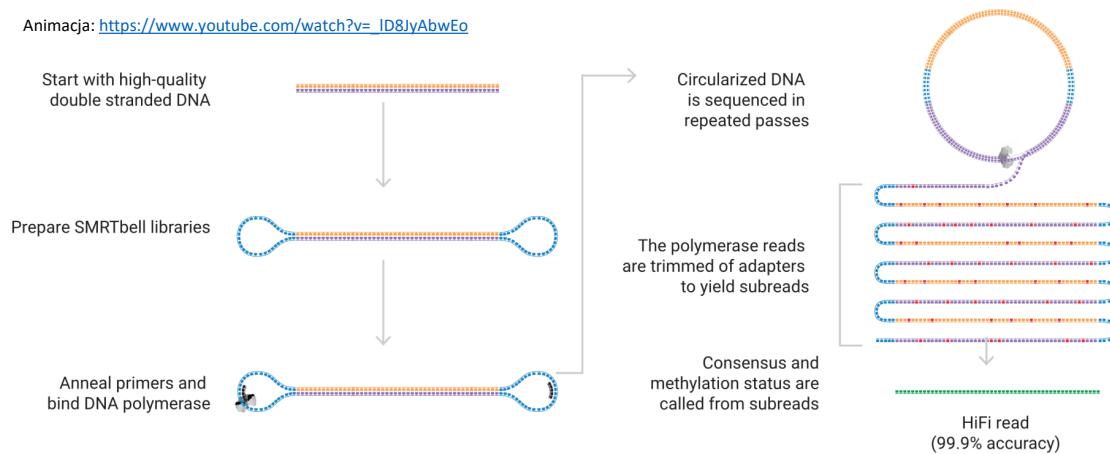
Technologie: c.d.

PacBio



PacBio

Animacja: <https://www.youtube.com/watch?v=ID8JyAbwEo>



PacBio



[PRODUCTS](#) [FOCUS AREAS](#) [ENGAGE](#) [SUPPORT](#) [COMPANY](#)

[CONTACT](#)



RNA SEQUENCING WITH THE ISO-SEQ METHOD

Reveal novel biology that other sequencing methods miss with full-length transcripts using the Iso-Seq[™] method

[Connect with a PacBio scientist](#)



SEQUENCING METHODS

[Whole genome](#)

[Targeted](#)

[RNA](#)

[Epigenetics](#)

[Metagenomics](#)



HUMAN TRANSCRIPTOME STUDIES

Explore how changes in isoform usages contribute to the phenotypic differences between health and disease.

[Learn more](#)



PLANT + ANIMAL TRANSCRIPTOME STUDIES

Advance breeding programs and basic research with full-length cDNA sequencing that does not require a reference genome.

[Learn more](#)



SINGLE-CELL TRANSCRIPTOME STUDIES

Visualize biology at an even higher resolution by capturing a comprehensive and complete transcript sequence from individual cells.

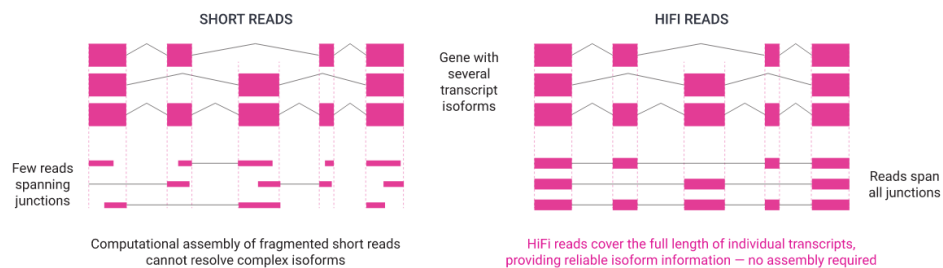
[Learn more](#)

PacBio

INTRODUCTION TO RNA SEQUENCING USING THE ISO-SEQ METHOD

With long and accurate HiFi reads, you can characterize the full diversity of the transcriptome:

- See alternative start and end sites
- Characterize splicing events
- Identify allele-specific isoforms
- Profile expression at isoform resolution
- Detect fusion genes
- Predict full-length open read frames
- ...and more



Oxford Nanopore

The screenshot displays the Oxford Nanopore website's 'RNA and gene expression analysis using direct RNA and cDNA sequencing' page. The page features a dark teal header with the Nanopore logo and navigation links. The main content area has a purple background with a DNA double helix graphic. It includes a breadcrumb trail, a title, a descriptive paragraph, and several key features highlighted with icons. At the bottom, there's a section titled 'How will you use nanopore technology?' with four interactive buttons.

Header: NANOPORE | Products | Applications | Resources | Investors | Careers | News | Search | About | Store | Community | Contact | Login / Register

Breadcrumb: Home > Applications > RNA Sequencing

Buttons: Get started | Talk to us | Subscribe

RNA and gene expression analysis using direct RNA and cDNA sequencing

Unlike traditional RNA-Seq techniques, long nanopore RNA sequencing reads allow for accurate quantification and complete, full-length characterisation of native RNA or cDNA without fragmentation or amplification — streamlining analysis and removing potential sources of bias. Direct RNA sequencing also enables the identification of base modifications alongside nucleotide sequence.

Key Features:

- Characterise and quantify full-length transcripts — up to single-cell resolution
- Get faster access to results with real-time analysis and bespoke tools
- Reduce bias with PCR-free protocols
- Explore epigenetic modifications through direct RNA sequencing
- Scale to your needs using Flongle, MinION, GridION or PromethION

Buy now > **Subscribe for more information >**

How will you use nanopore technology?

- Quantify and study differential gene expression
- Identify fusion transcripts
- Characterise RNA viruses and viral epidemiology
- Detect base modifications

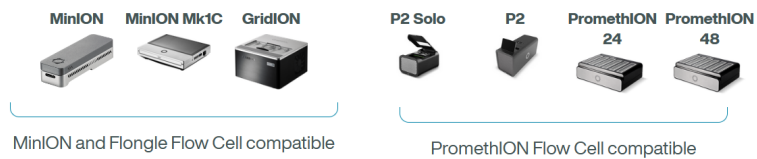
Oxford Nanopore



Configuration	Platform			Techniques		Tech specifications	
Number of flow cells per device	1	1	5	2	2	24	48
Maximum number of channels per flow cell	512	512	512	2,675	2,675	2,675	2,675
Run time	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours
Device TMO†	50 Gb	50 Gb	250 Gb	580 Gb	580 Gb	~7 Tb	~14 Tb
Maximum number of flow cells per year*	104	104	520	208	208	2,596	4,992

<https://nanoporetech.com/products/specifications>

Oxford Nanopore



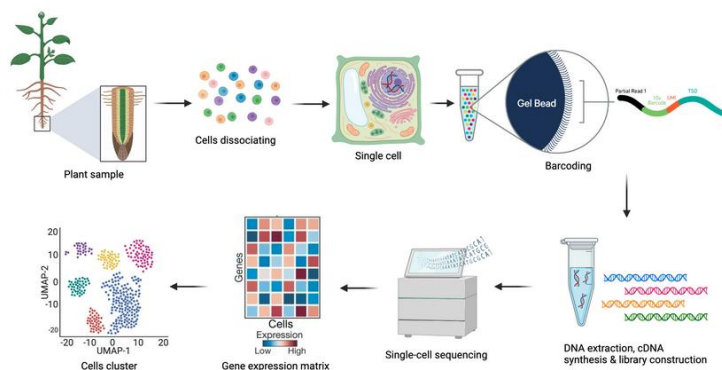
Configuration	Platform			Techniques		Tech specifications	
WGS - small genomes	Yes: Low to medium plex	Yes: Low to medium plex	Yes: Low to medium plex	Yes: Highly multiplexed	Yes: Highly multiplexed	Yes: Highly multiplexed	Yes: Highly multiplexed
WGS - large genomes	Low pass	Low pass	Yes	Yes	Yes	Yes	Yes
Targeted sequencing	Yes: Low to medium plex, and adaptive sampling			Yes: Highly multiplexed			
Metagenomics	Yes: Quantitative species ID			Yes: Quantitative species ID			
RNA sequencing	Yes: Isoform & expression from same experiment			Yes: Isoform & expression from same experiment, including low abundance transcripts			
Epigenetics	Yes - when native DNA used			Yes - when native DNA used			

<https://nanoporetech.com/products/specifications>

scRNA-Seq i transkryptomiki przestrzenna

Transkryptomika pojedynczych komórek (scRNA-Seq)

- pojedyncze komórki są najpierw fizycznie izolowane, a następnie ich RNA jest przepisane na cDNA, amplifikowane i sekwencjonowane – w efekcie dla każdej komórki powstaje odrębna biblioteka cDNA i osobny profil ekspresji genów.

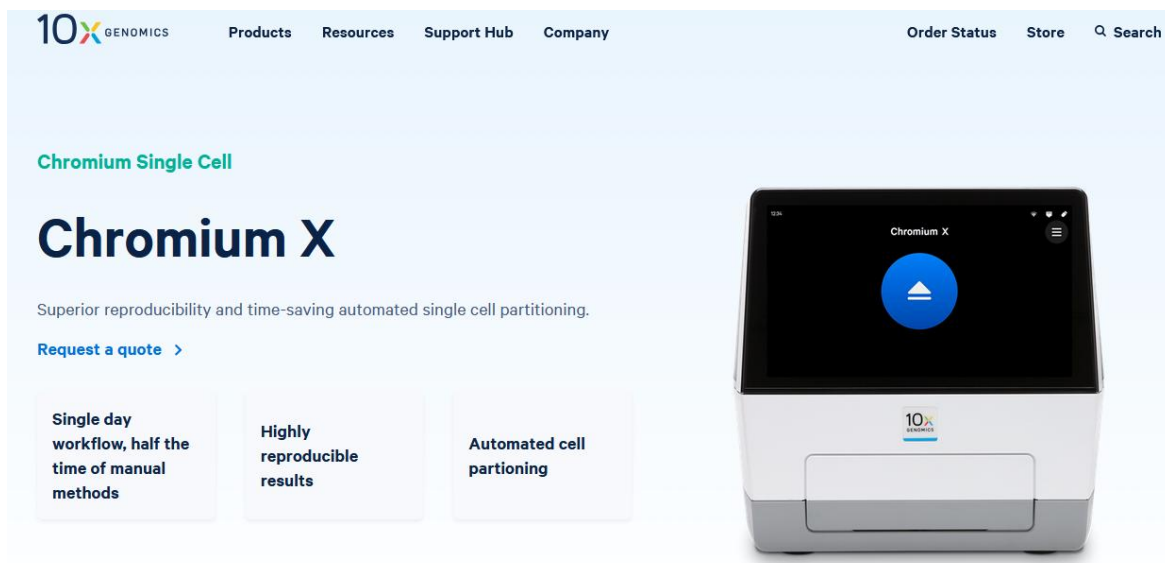


Ali, Muhammad, Tianxia Yang, Hai He, and Yu Zhang. "Plant biotechnology research with single-cell transcriptome: recent advancements and prospects." *Plant Cell Reports* 43, no. 3 (2024): 1-18.

Transkryptomika pojedynczych komórek (scRNA-Seq)

- Pod względem zakresu sekwencjonowanego transkryptu metody te dzieli się na dwie główne grupy: techniki pozyskujące pełnej długości cząsteczki mRNA oraz techniki sekwencjonujące wyłącznie koniec transkryptu (np. fragment przy końcu 3' mRNA).
- Przykładami metod sekw. w pełnym zakresie długości są m.in. Smart-seq2 (opracowana w 2013) czy MATQ-seq umożliwiające one odczyt całych sekwencji mRNA, co pozwala analizować alternatywne formy splicingowe oraz mutacje (polimorfizmy) na poziomie pojedynczej komórki.
- Metody zliczające głównie koniec 3' transkryptu zapewniają zwykle wyższą przepustowość kosztem utraty informacji o pełnej sekwencji. Do tej kategorii należą m.in. platforma Chromium firmy 10x Genomics, system Fluidigm C1, a także Drop-seq i inDrop. Metody te charakteryzują się możliwością równoległego profilowania tysięcy a nawet dziesiątek tysięcy komórek w jednym eksperymencie, przy relatywnie niskim koszcie w przeliczeniu na komórkę. Wadą jest ograniczenie analizy do stosunkowo niewielkich fragmentów transkryptów.

Transkryptomika pojedynczych komórek (scRNA-Seq)



The image is a screenshot of the 10x Genomics website's product page for the Chromium X. The header features the 10x Genomics logo on the left, navigation links for Products, Resources, Support Hub, and Company in the center, and links for Order Status, Store, and a Search icon on the right. The main content area has a light blue background. On the left, it says 'Chromium Single Cell' in green, followed by 'Chromium X' in large bold black text. Below this is the tagline 'Superior reproducibility and time-saving automated single cell partitioning.' and a 'Request a quote >' link. Three white boxes with grey borders list key features: 'Single day workflow, half the time of manual methods', 'Highly reproducible results', and 'Automated cell partitioning'. On the right, there is a photograph of the Chromium X instrument, a white and grey automated cell partitioning machine with a large black touchscreen displaying the 10x logo and a blue upload icon.

10x GENOMICS Products Resources Support Hub Company Order Status Store Search

Chromium Single Cell

Chromium X

Superior reproducibility and time-saving automated single cell partitioning.

[Request a quote >](#)

- Single day workflow, half the time of manual methods
- Highly reproducible results
- Automated cell partitioning



<https://www.10xgenomics.com/instruments/chromium-x>

Transkryptomika pojedynczych komórek (scRNA-Seq)

UCI Genomics Research & Technology Hub

[Home](#) [About Us](#) [Instruments & Services](#) [Seminars & Workshops](#) [Publications](#) [Research Support](#) [Careers](#) [Advisory Board](#) [Search](#)



Fluidigm C1 Single-Cell Auto Prep System

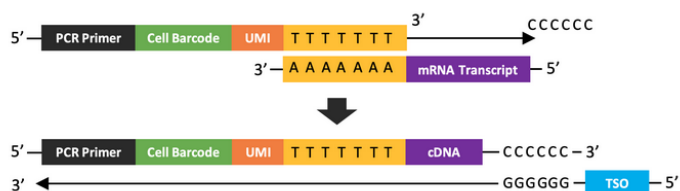
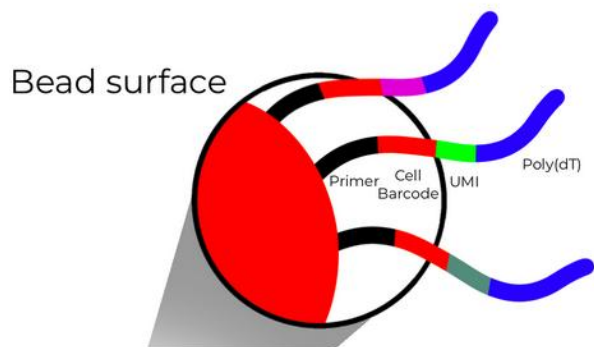
The first automated solution for single-cell genomics research, the C1 Single-Cell Auto Prep system employs innovative microfluidic technology to rapidly and reliably isolate, process and prepare individual cells for genome and transcriptome analysis. The C1 system isolates single cells into individual reaction chambers in Fluidigm's integrated fluidic circuit (IFC). The optically clear IFC allows for automated staining of captured cells and examination by microscopy for viability, surface markers or reporter genes. After staining, cells are automatically lysed and template is quickly prepared for qPCR or sequencing analysis. The applications for this system include, gene expression and single-cell genomics.

[More Information from the Manufacturer](#)

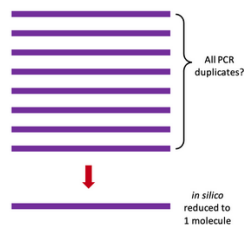
<https://genomics.uci.edu/instrument/fluidigm-c1-single-cell-auto-prep-system/>

Transkryptomika pojedynczych komórek (scRNA-Seq)

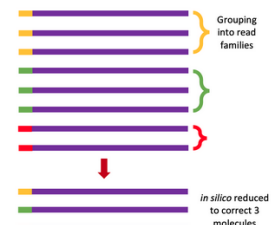
- Ważnym elementem protokołów jest wprowadzanie do cDNA specjalnych sekwencji znaczników: każda komórka otrzymuje unikalny kod kreskowy (*cell barcode*), co pozwala rozdzielić odczyty pochodzące z różnych komórek, oraz dodawane są unikalne identyfikatory molekularne (UMI) do poszczególnych cząsteczek cDNA, co umożliwia odróżnienie unikatowych transkryptów od zduplikowanych kopii powstałych podczas amplifikacji.



PCR duplicate removal **without** UMIs



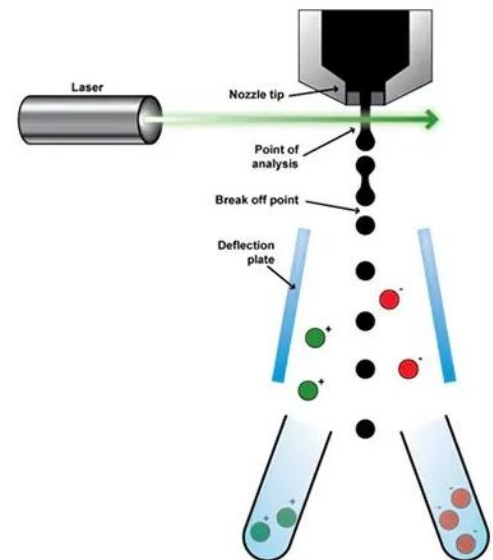
PCR duplicate removal **with** UMIs



<https://help.geneiousbiologics.com/hc/en-us/articles/4781289585300-Understanding-Single-Cell-technologies-Barcodes-and-UMIs>

Transkryptomika pojedynczych komórek (scRNA-Seq)

- Dostępnych jest wiele metod izolacji pojedynczych komórek. Najprostszą techniką jest tzw. limiting dilution, czyli seryjne rozcieńczanie zawiesiny komórek i rozkładanie na płytki w nadziei, że niektóre dołki zawierają tylko jedną komórkę – metoda ta jest jednak mało wydajna i rzadko stosowana w praktyce.
- Bardziej zaawansowane podejście to sortowanie pojedynczych komórek za pomocą cytometrii przepływowej z detekcją fluorescencyjną (FACS). Komórki opatrzone znacznikami fluorescencyjnymi (np. przeciwciałami na markery powierzchniowe) są indywidualnie wyodrębniane na drodze elektrostatycznego sortowania kroplowego. FACS pozwala wyselekcjonować nawet rzadkie typy komórek z populacji, ale wymaga dysponowania znaczną liczbą komórek wejściowych (zwykle >10000) i jest czasochłonny.
- W przypadku tkanek (np. skrawków) unikatową możliwość daje laserowa mikrodyssekcja (LCM) - pod mikroskopem wyposażonym w laser można wyciąć pojedyncze komórki z preparatu tkankowego. LCM pozwala izolować komórki z zachowaniem ich lokalizacji (przed wycięciem można ocenić np. typ komórki w oparciu o morfologię lub barwienie), jednak jest to metoda dość pracochłonna i niskoprzepustowa.
- Najbardziej popularną obecnie klasą metod są rozwiązania mikroprzepływowe, gdzie specjalne urządzenia automatycznie rozdzielają setki lub tysiące komórek. Przykładem jest system 10x Genomics Chromium czy starszy mikrochip Fluidigm C1 – techniki te cechują się wysoką wydajnością izolacji i kontrolą nad przepływem roztworu, a dzięki integracji z automatycznym wykonaniem barcodingu i innych odczynników znacząco redukują błąd ludzki.



<https://www.abcam.com/en-us/knowledge-center/flow-cytometry/fluorescence-activated-cell-sorting-overview>

Transkryptomika pojedynczych komórek (scRNA-Seq)

- Podstawowym wynikiem eksperymentu scRNA-seq jest macierz zliczeń – tabela, w której wierszami są geny, kolumnami poszczególne komórki, a komórki tabeli zawierają liczby odczytów (lub zidentyfikowanych cząsteczek cDNA) dla danego genu w danej komórce. Uzyskanie takiej macierzy wymaga przetworzenia surowych danych sekwencjonowania: odczyty muszą zostać przypisane do odpowiednich komórek (na podstawie dołączonych do sekwencji barcodów) oraz zmapowane do genów.
- Po wstępnym zliczeniu ekspresji rozpoczyna się właściwa analiza – najpierw przeprowadza się kontrolę jakości. Eliminowane są komórki o bardzo małej liczbie odczytów (prawdopodobnie obumarłe lub uszkodzone), a także komórki wykazujące nienaturalne proporcje odczytów, np. zbyt wysoki odsetek transkryptów mitochondrialnych (co wskazuje na ekstrakcję RNA głównie z mitochondriów z powodu pęknięcia komórki). Następnie dane poddawane są normalizacji, aby skorygować różnice w głębokości sekwencjonowania i składzie odczytów między komórkami; jednym z podejść jest sprowadzenie liczby odczytów do wartości na ustalony poziom (np. 10 tysięcy na komórkę) i przeskalowanie, często z dodatkowymi transformacjami (logarytmizacja), tak by porównania między komórkami odzwierciedlały względne różnice w ekspresji genów.

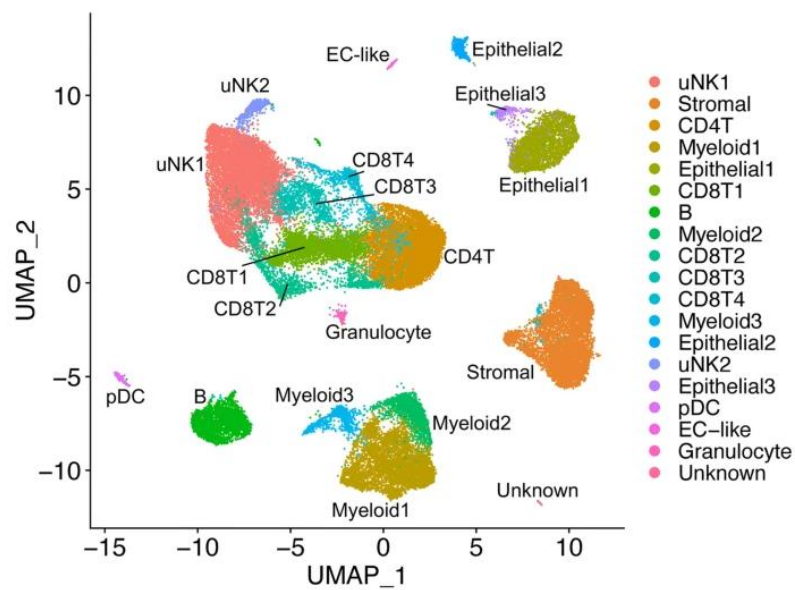
Transkryptomika pojedynczych komórek (scRNA-Seq)

- Kolejnym etapem jest redukcja wymiarowości danych i klasteryzacja komórek. Stosuje się zatem algorytmy takie jak analiza głównych składowych (PCA) do sprowadzenia danych do kilkunastu lub kilkudziesięciu składowych uchwytujących większość zmienności w zbiorze. Następnie często używane są metody nieliniowej redukcji wymiaru i wizualizacji, jak t-SNE czy UMAP, aby przedstawić komórki w dwóch lub trzech wymiarach w sposób odzwierciedlający ich podobieństwo. Równolegle lub po tych krokach wykonuje się grupowanie komórek – np. klasteryzację metodami unsupervised, takimi jak algorytm k-means lub wyspecjalizowane algorytmy dla danych scRNA-seq. Celem jest pogrupowanie komórek o podobnych profilach ekspresji w zbiory odpowiadające potencjalnie odrębnym typom lub stanom komórek. W praktyce popularne narzędzia (np. pakiet Seurat w języku R) udostępniają zautomatyzowane pipeline'y wykonujące powyższe kroki i sugerujące podział na klastry.

Transkryptomika pojedynczych komórek (scRNA-Seq)

- Gdy z danych wyłonią się grupy komórek, następuje faza interpretacji biologicznej, czyli adnotacji klastrow. Na tym etapie badacz stara się przypisać zidentyfikowanym klastrom tożsamość – np. rozpoznać, że klaster 1 to komórki T, klaster 2 to makrofagi, klaster 3 to komórki nowotworowe w stanie proliferacji itd. Standardowym podejściem jest poszukiwanie w każdym klastrze genów markerowych, tzn. genów wykazujących istotnie wyższą ekspresję w danym klastrze niż w pozostałych (analiza różnicowej ekspresji między klastrami). Takie geny można następnie porównać z istniejącą wiedzą (publikacje, bazy danych markerów typu PanglaoDB itp.), aby zidentyfikować typ komórki. Na przykład, jeśli klaster charakteryzuje wysoka ekspresja genów CD3E, CD3D, LCK, można go zinterpretować jako limfocyty T, zaś inny klaster z genami MBP, PLP1 jako oligodendrocyty w mózgu.

UMAP plot



Shih AJ, Adelson RP, Vashistha H, Khalili H, Nayyar A, Puran R, Herrera R, Chatterjee PK, Lee AT, Truskinovsky AM, Elmaliki K, DeFranco M, Metz CN, Gregersen PK. Single-cell analysis of menstrual endometrial tissues defines phenotypes associated with endometriosis. BMC Med. 2022 Sep 15;20(1):315. doi: 10.1186/s12916-022-02500-3. PMID: 36104692; PMCID: PMC9476391.

Ograniczenia i wyzwania scRNA-seq

- Niska efektywność wychwytywania RNA: z pojedynczej komórki udaje się zwykle zaobserwować tylko ułamek faktycznie obecnych w niej cząsteczek mRNA. Wynika to z ograniczeń technicznych - strat podczas preparatyki oraz limitowanej czułości sekwencjonowania. W konsekwencji dane scRNA-seq są *sparse* – zawierają wiele zer (tzw. drop-outs, gdy dany gen nie został wykryty w komórce, mimo że potencjalnie był w niej obecny). Na przykład, szacuje się, że typowa komórka ssaka zawiera $\sim 10^5$ – 10^6 cząsteczek mRNA, ale w pojedynczym eksperymencie scRNA-seq udaje się wykryć naraz najwyżej kilkanaście tysięcy transkryptów. Oznacza to, że geny o bardzo niskiej ekspresji często umykają detekcji, co utrudnia analizę pełnego spektrum różnic między komórkami.
- Zmienność biologiczna niezwiązana z typem komórki: komórki mogą różnić się ilością zebranych transkryptów z powodów niezwiązanych z istotą badanej heterogenności - np. większe komórki lub komórki w fazie G2/M cyklu mają więcej mRNA niż mniejsze lub będące w fazie spoczynku. Te różnice w wielkości/aktywności metabolicznej przekładają się na globalne różnice w liczbie odczytów i mogą zaburzać interpretację wyników, jeśli nie zostaną uwzględnione (np. komórki w mitozie mogą formować osobny klastrowanie głównie z powodu wzmożonej ekspresji genów podziałowych, a nie dlatego, że należą do innego typu).
- Batch effects i reprodukowalność: eksperymenty single-cell wymagają wielu kroków manualnych, co rodzi ryzyko wprowadzenia systematycznych różnic między powtórzeniami technicznymi czy biologicznymi (np. między próbkami przygotowanymi w różne dni lub przez różnych ludzi). Różnice te, jak np. odmienna efektywność odwrotnej transkrypcji czy amplifikacji, mogą powodować klastrowanie komórek według przynależności do partii eksperymentalnej zamiast według biologii. Konieczne jest stosowanie procedur kontrolnych (mieszanie komórek z różnych warunków w jednej partii, dodawanie wewnętrznych standardów kontroli jakości, itp.) oraz odpowiednich metod normalizacji między próbkami, aby zminimalizować te artefakty.
- Artefakty związane z przygotowaniem próbek: sam proces izolacji pojedynczych komórek z tkanki może zmieniać ich transkryptom. Enzymatyczna i mechaniczna dysocjacja tkanki to stres dla komórek; badania wykazały, że procedura ta może indukować ekspresję genów związanych ze stresem i stanem zapalnym lub powodować utratę niektórych transkryptów. Na przykład, trawienie proteolityczne używane do rozluźnienia macierzy pozakomórkowej potrafi aktywować komórki (np. mikroglej w trakcie izolacji zaczyna wykazywać profil prozapalny in vitro, który nie odzwierciedla stanu in vivo). Takie zjawisko nazywane jest dissociation-induced gene expression i stanowi przedmiot badań: opracowano listy genów, których ekspresja gwałtownie zmienia się podczas izolacji (należy je traktować ostrożnie przy interpretacji wyników). Jednym z pomysłów na ograniczenie tego problemu jest stosowanie tzw. zimnej dysocjacji (w niskiej temperaturze) lub bardzo szybkiej izolacji i utrwalenia RNA (np. metodą FIX-seq).
- Złożoność i koszty: scRNA-seq pozostaje technologią wymagającą specjalistycznego sprzętu i wiedzy. Przygotowanie bibliotek z tysięcy komórek wymaga precyzji i często drogiej aparatury. Analiza tych danych wymaga znacznej mocy obliczeniowej oraz umiejętności bioinformatycznych.

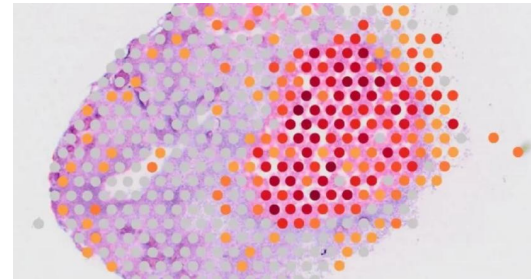
Transkryptomika przestrzenna

- Transkryptomika przestrzenna (spatial transcriptomics) to zbiór technik, które umożliwiają pomiar ekspresji genów bezpośrednio w wycinku tkanki, z zachowaniem informacji o przestrzennym ułożeniu komórek. W przeciwieństwie do scRNA-seq, gdzie komórki są fizycznie rozdzielone i mieszane (co niszczy strukturę tkanki), techniki przestrzenne zachowują in situ relacje między komórkami.
- Obecnie dostępne techniki przestrzenne można podzielić na dwie główne kategorie: metody oparte na sekwencjonowaniu (NGS) oraz metody oparte na obrazowaniu (in situ). Oba podejścia różnią się sposobem detekcji, ale są komplementarne i często używane zamiennie w zależności od potrzeb badania.

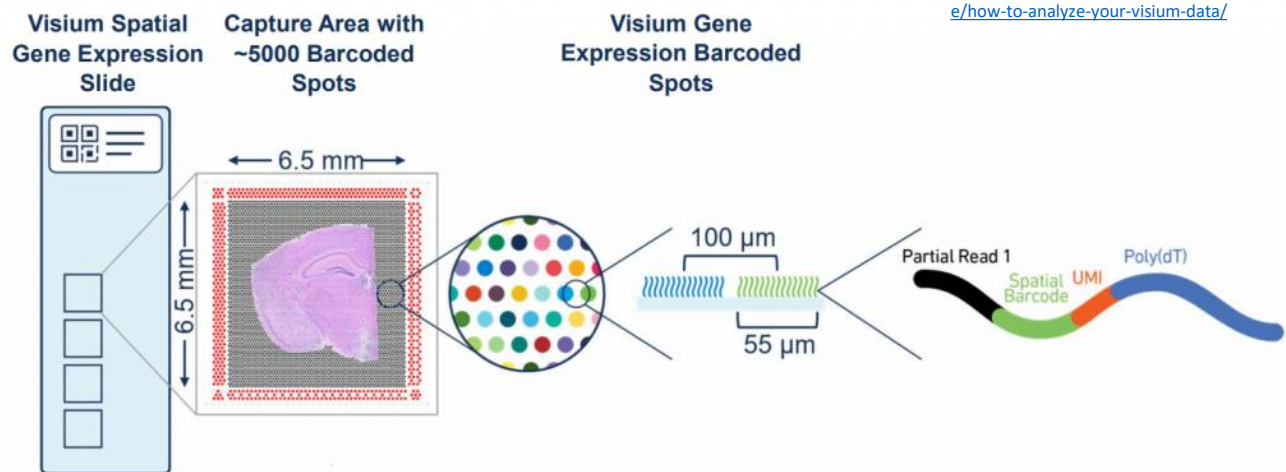
Transkryptomika przestrzenna: metody NGS

- Ich zasadą jest utrwalenie informacji o położeniu transkryptów, a następnie wykorzystanie sekwencjonowania masowego do identyfikacji tych transkryptów. Typowy protokół polega na położeniu skrawka tkanki na szkiełku mikroskopowym, które pokryte jest gęstą siatką sond oligonukleotydowych o znanych współrzędnych. Każda sonda zawiera unikalny barcode oraz sekwencję poli-T do wychwycenia ogona PoliA. Po nałożeniu tkanki, mRNA dyfunduje z komórek i wiąże się z najbliższymi sondami. Następnie wykonuje się odwrotną transkrypcję: powstaje cDNA zawierające sekwencję komplementarną do mRNA oraz dołączony fragment kodu lokalizacji. Z próbki zbierane jest całe powstałe cDNA, które poddaje się standardowej procedurze przygotowania bibliotek i sekwencjonuje na maszynie NGS. Odczyty z sekwenatora zawierają informację o genie (sekwencja cDNA) oraz o położeniu (kod barcode). W rezultacie otrzymujemy zestaw transkryptów przypisanych do konkretnych miejsc na siatce, z których odczytano sygnał.
- Pierwszą tego typu techniką była oryginalna metoda nazwana po prostu Spatial Transcriptomics, opracowana w 2016 roku przez zespół ze Szwecji (Ståhl et al., Science 2016). Obecnie jest ona skomercjalizowana jako platforma Visium firmy 10x Genomics, z ulepszoną rozdzielczością i procedurą kompatybilną zarówno ze świeżymi mrożonymi skrawkami, jak i z utrwalonymi (FFPE). Platforma Visium dysponuje siatką ~5 tys. spots na powierzchni $6,5 \times 6,5$ mm. Rozdzielczość wynosi ok. $55 \mu\text{m}$ – jest to średnica koła, w obrębie którego zbierane są transkrypty.
- Ponieważ typowa komórka ssacza ma średnicę ~10–20 μm , jeden spot Visium obejmuje zwykle kilka do kilkunastu komórek. Otrzymane dane reprezentują zatem połączony transkryptom grupy komórek znajdujących się w danym miejscu tkanki, a nie pojedynczych komórek. Utrata rozdzielczości na poziomie pojedynczej komórki jest istotnym ograniczeniem – sygnał z różnych komórek sumuje się, przez co trudniej wykryć obecność rzadkich typów czy analizować interakcje pojedynczych komórek w gęstych tkankach. Niemniej zaletą Visium jest możliwość badania całego transkryptomu (ponad 15 tysięcy genów) naraz – profil ekspresji jest globalny, a nie ograniczony do wybranych markerów.
- Aby poprawić rozdzielczość przestrzenną, intensywnie rozwijane są nowe NGS-owe metody o wyższej gęstości detekcji. Jedną z nich jest Slide-seq, opracowana w 2019 roku (Rodriques et al., Science 2019). Slide-seq umożliwia mapowanie transkryptomu z prawie jednokomórkową precyzją, choć w praktyce nadal zdarza się, że jedna kulka rejestruje transkrypty z 2–3 małych komórek przylegających do siebie. Udoskonalona wersja Slide-seqV2 poprawiła czułość detekcji RNA o rząd wielkości w porównaniu do oryginału, dzięki optymalizacjom chemicznym. Niezależnie, grupa ze Szwecji (Vickovic et al., 2019) opracowała metodę HDST (High Definition Spatial Transcriptomics), również opartą na kulkach, ale jeszcze mniejszych – ~2 μm średnicy. HDST osiągnęło imponującą rozdzielczość subkomórkową, ale kosztem bardzo skomplikowanej analizy (ogromna liczba małych objętości do odczytu) i wąskiego pola (badano małe kawałki tkanki). Kolejne innowacje, jak Seq-Scope, wykorzystują fotolitografię do osadzenia matrycy sond wprost na szkiełku z gęstością, która daje rozdzielczość rzędu 1–2 μm . Dzięki temu Seq-Scope uzyskuje obraz rozmieszczenia transkryptów niemal na poziomie pojedynczych organelli. Trzeba jednak podkreślić, że choć metody te zbliżają się do obserwacji pojedynczych komórek, to w praktyce pewna agregacja sygnału jest nieunikniona (np. nawet 2 μm kulka może złapać transkrypty z dwóch stykających się jąder komórkowych).

10x Genomics Visium Spatial Gene Expression



<https://www.scdiscoveries.com/blog/knowledge/how-to-analyze-your-visium-data/>

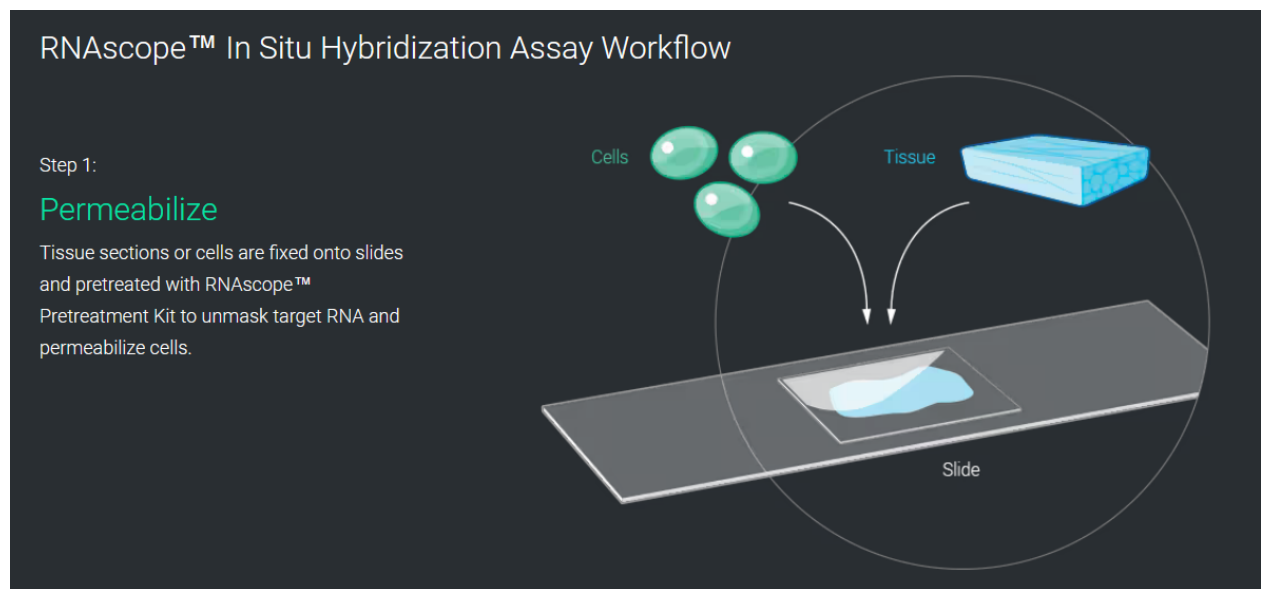


The Visium Spatial Gene Expression Slide (<https://www.10xgenomics.com/>)

Transkryptomika przestrzenna: metody obrazowania *in situ*

- Drugą rodziną technik transkryptomiki przestrzennej są metody, które bezpośrednio wykrywają cząsteczki RNA na preparacie za pomocą sond fluorescencyjnych lub metod sekwencjonowania *in situ*. W odróżnieniu od NGS, tutaj nie izolujemy RNA z tkanki do sekwencjonowania – cała detekcja odbywa się na szkiełku, zwykle przy pomocy mikroskopu fluorescencyjnego o wysokiej rozdzielczości. Najprostszym przykładem jest klasyczna hybrydyzacja *in situ* mRNA przy użyciu sond DNA – współczesna wersja to np. technika RNAscope, w której do wykrycia pojedynczej cząsteczki mRNA używa się pary sond oligonukleotydowych i kaskady reakcji amplifikacji sygnału. RNAscope pozwala wizualizować wybrane transkrypty z dokładnością do pojedynczych cząsteczek na tle obrazu tkanki. Metoda ta jest jednak ograniczona do kilku genów naraz (typowo 1–3 kolory fluorescencji).

RNAscope



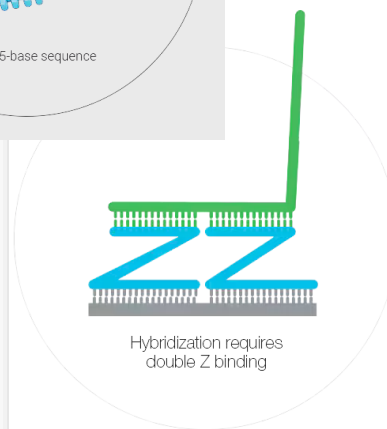
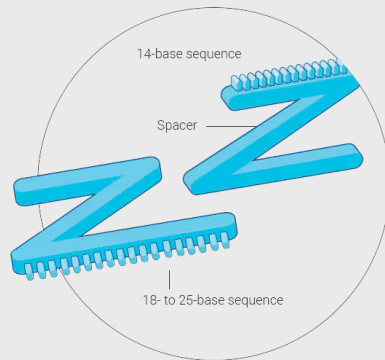
<https://acdbio.com/science/how-it-works>

Each Target Z Probe Contains Three Elements

The lower region of the Z is an 18-to 25-base region that is complementary to the target RNA. This sequence is selected for target specific hybridization and uniform hybridization properties.

A spacer sequence that links the two components of the probe. The upper region of the Z is a 14-base tail sequence.

The two tails from a double Z probe pair forms a 28 base binding site for the pre-amplifier.



RNAscope™ Probe Design and Signal Amplification Strategy

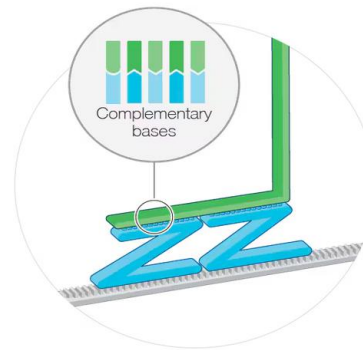
In order to substantially improve the signal-to-noise ratio of RNA ISH, RNAscope employs a probe design strategy much akin to fluorescence resonance energy transfer (FRET), in which two independent probes (double Z probes) have to hybridize to the target sequence in tandem in order for signal amplification to occur. As it is highly unlikely that two independent probes will hybridize to a non-specific target right next to each other, this design concept ensures selective amplification of target-specific signals.

For each target RNA species, ~20 double Z target probe pairs are designed to specifically hybridize to the target molecule, but not to non-targeted molecules.

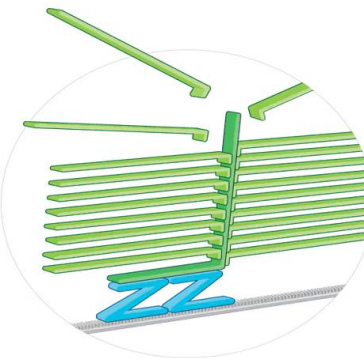
<https://acdbio.com/science/how-it-works>



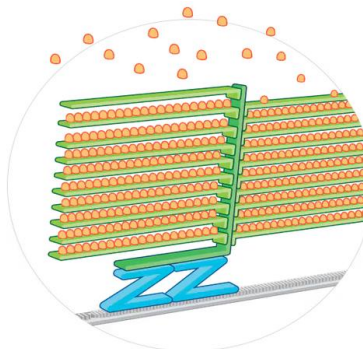
Step 1. Double Z target probes hybridize to the RNA target (~1kb)



Step 2. Pre-amplifiers hybridize to the 28-base binding site formed by each double Z probe



Step 3: Amplifiers are then binding to the multiple binding sites on each preamplifier.



Step 4: Labeled probes, containing a fluorescent molecule or chromogenic enzyme, bind to the numerous binding sites on each amplifier.

<https://acdbio.com/science/how-it-works>

*„Koniec i bomba;
Kto słuchał, ten trąba!”*

