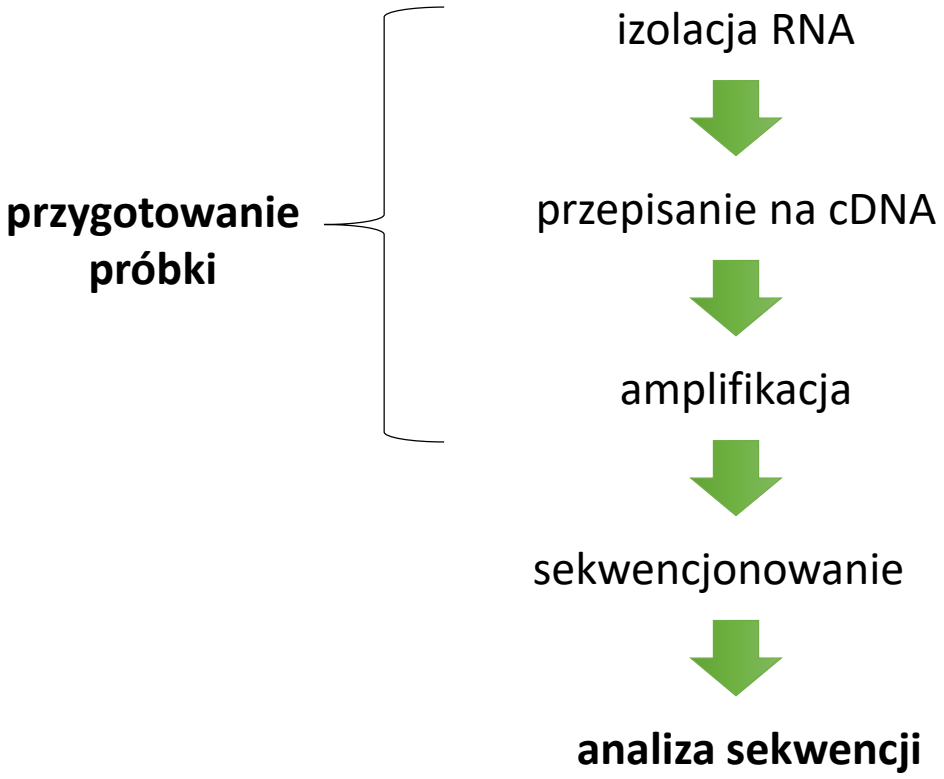
Analiza danych z wysokoprzepustowego sekwencjonowania

W2: Sekwencjonowanie transkryptomu

sekwencjonowanie transkryptomu

Transkryptomika



sekwencjonowanie transkryptomu

Zalety:

- metoda bezpośrednia – widzimy to co jest, a nie to o czym wiemy że jest

- metoda skalowalna – możemy dopasować do swoich potrzeb

- metoda adaptowalna – możliwe badanie różnych zjawisk poprzez odpowiednie przygotowanie biblioteki cDNA

Wady:

- wyższy koszt niż w przypadku mikromacierzy

profilowanie ekspresji genów

A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome

Marc Sultan,[1*] Marcel H. Schulz,[2,3*] Hugues Richard,[2*] Alon Magen,[1] Andreas Klingenhoff,[4] Matthias Scherf,[4] Martin Seifert,[4] Tatjana Borodina,[1] Aleksey Soldatov,[1] Dmitri Parkhomchuk,[1] Dominic Schmidt,[1] Sean O'Keeffe,[2] Stefan Haas,[2] Martin Vingron,[2] Hans Lehrach,[1] Marie-Laure Yaspo[1]†

The functional complexity of the human transcriptome is not yet fully elucidated. We report a high-throughput sequence of the human transcriptome from a human embryonic kidney and a B cell line. We used shotgun sequencing of transcripts to generate randomly distributed reads. Of these, 50% mapped to unique genomic locations, of which 80% corresponded to known exons. We found that 66% of the polyadenylated transcriptome mapped to known genes and 34% to nonannotated genomic regions. On the basis of known transcripts, RNA-Seq can detect 25% more genes than can microarrays. A global survey of messenger RNA splicing events identified 94,241 splice junctions (4096 of which were previously unidentified) and showed that exon skipping is the most prevalent form of alternative splicing.

The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing

Ugrappa Nagalakshmi,[1*] Zhong Wang,[1*] Karl Waern,[1] Chong Shou,[2] Debasish Raha,[1] Mark Gerstein,[2,3] Michael Snyder[1,2,3]†

The identification of untranslated regions, introns, and coding regions within an organism remains challenging. We developed a quantitative sequencing-based method called RNA-Seq for mapping transcribed regions, in which complementary DNA fragments are subjected to high-throughput sequencing and mapped to the genome. We applied RNA-Seq to generate a high-resolution transcriptome map of the yeast genome and demonstrated that most (74.5%) of the nonrepetitive sequence of the yeast genome is transcribed. We confirmed many known and predicted introns and demonstrated that others are not actively used. Alternative initiation codons and upstream open reading frames also were identified for many yeast genes. We also found unexpected 3′-end heterogeneity and the presence of many overlapping genes. These results indicate that the yeast transcriptome is more complex than previously appreciated.

Comparative Functional Genomics of the Fission Yeasts

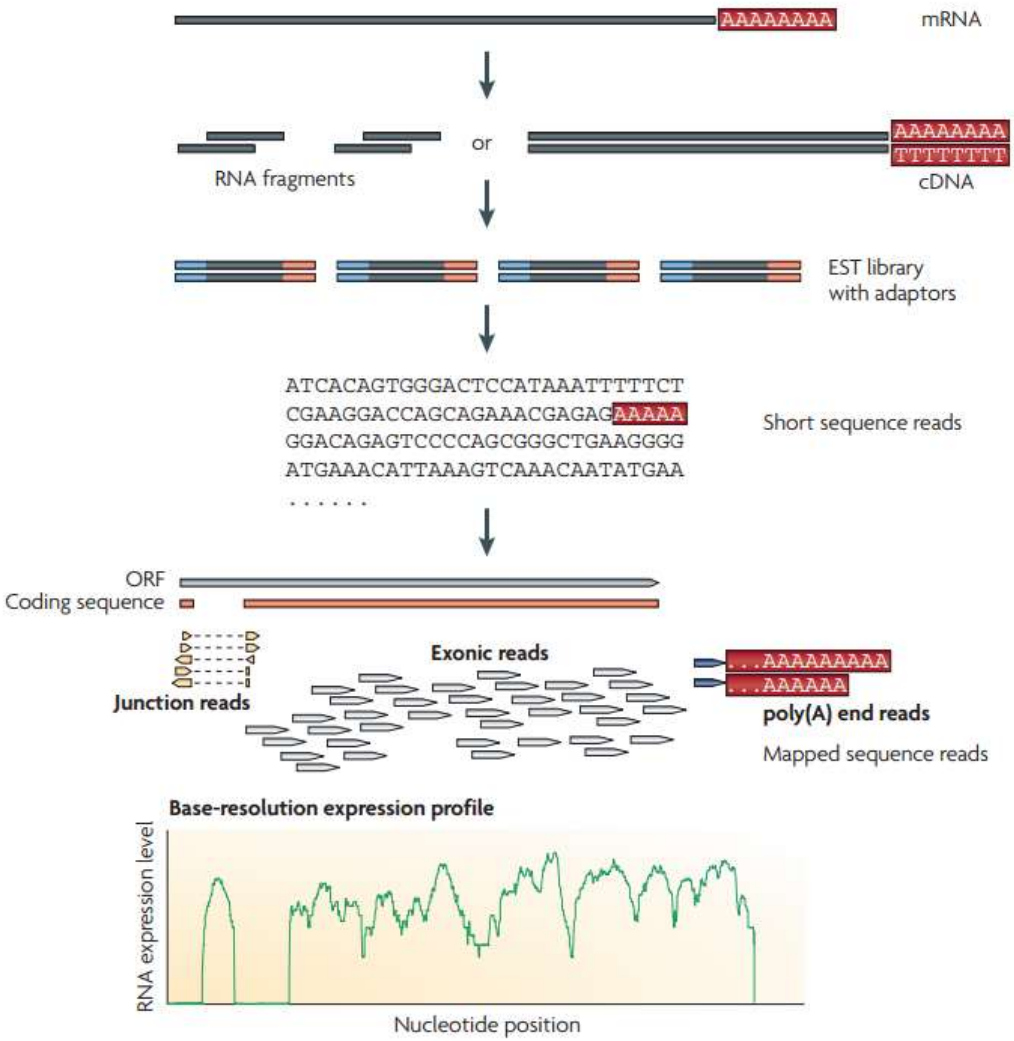
Nicholas Rhind,[1¶] Zehua Chen,[2] Moran Yassour,[3,4,5‖] Dawn A. Thompson,[3‖] Brian J. Haas,[2‖] Naomi Habib,[5,6‖] Ilan Wapinski,[3,7‖] Sushmita Roy,[3,8‖] Michael F. Lin,[8] David I. Heiman,[2] Sarah K. Young,[2] Kanji Furuya,[9] Yabin Guo,[10] Alison Pidoux,[11] Huei Mei Chen,[12] Barbara Robbertse,[13*] Jonathan M. Goldberg,[2] Keita Aoki,[9] Elizabeth H. Bayne,[11]† Aaron M. Berlin,[2] Christopher A. Desjardins,[2] Edward Dobbs,[11] Livio Dukaj,[1] Lin Fan,[2] Michael G. FitzGerald,[2] Courtney French,[6] Sharvari Gujja,[2] Klavs Hansen,[14]‡ Dan Keifenheim,[1] Joshua Z. Levin,[2] Rebecca A. Mosher,[15§] Carolin A. Müller,[16] Jenna Pfiffner,[2] Margaret Priest,[2] Carsten Russ,[2] Agata Smialowska,[17,18] Peter Swoboda,[17] Sean M. Sykes,[2] Matthew Vaughn,[14] Sonya Vengrova,[19] Ryan Yoder,[13] Qiandong Zeng,[2] Robin Allshire,[11] David Baulcombe,[15] Bruce W. Birren,[20] William Brown,[16] Karl Ekwall,[17,18] Manolis Kellis,[8,3] Janet Leatherwood,[12] Henry Levin,[10] Hanah Margalit,[6] Rob Martienssen,[14] Conrad A. Nieduszynski,[16] Joseph W. Spatafora,[13] Nir Friedman,[5,21] Jacob Z. Dalgaard,[19] Peter Baumann,[22,23,24] Hironori Niki,[9] Aviv Regev,[3,4,24¶] Chad Nusbaum[2¶]

Cold-Inducible RNA-Binding Protein Modulates Circadian Gene Expression Posttranscriptionally

Jörg Morf,[1] Guillaume Rey,[2*] Kim Schneider,[1] Markus Stratmann,[1] Jun Fujita,[3] Felix Naef,[2] Ueli Schibler[1]†

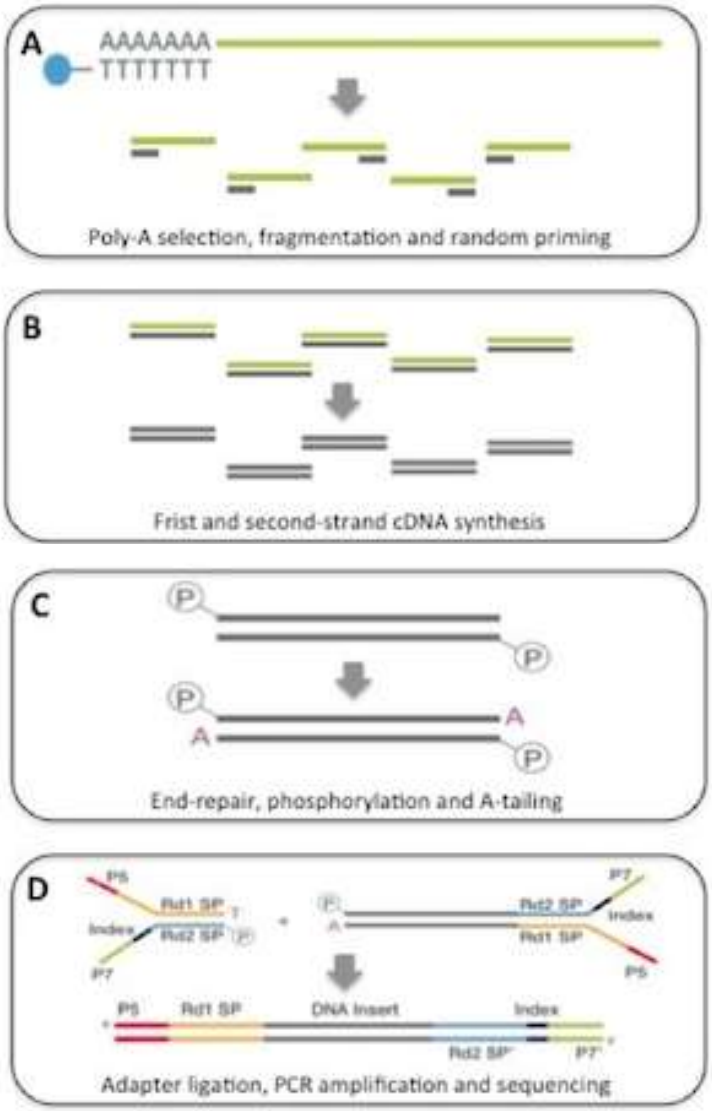
In mammalian tissues, circadian gene expression can be driven by local oscillators or systemic signals controlled by the master pacemaker in the suprachiasmatic nucleus. We show that simulated body temperature cycles, but not peripheral oscillators, controlled the rhythmic expression of cold-inducible RNA-binding protein (CIRP) in cultured fibroblasts. In turn, loss-of-function experiments indicated that CIRP was required for high-amplitude circadian gene expression. The transcriptome-wide identification of CIRP-bound RNAs by a biotin-streptavidin–based cross-linking and immunoprecipitation (CLIP) procedure revealed several transcripts encoding circadian oscillator proteins, including CLOCK. Moreover, CLOCK accumulation was strongly reduced in CIRP-depleted fibroblasts. Because ectopic expression of CLOCK improved circadian gene expression in these cells, we surmise that CIRP confers robustness to circadian oscillators through regulation of CLOCK expression.

profilowanie ekspresji genów - RNA-seq



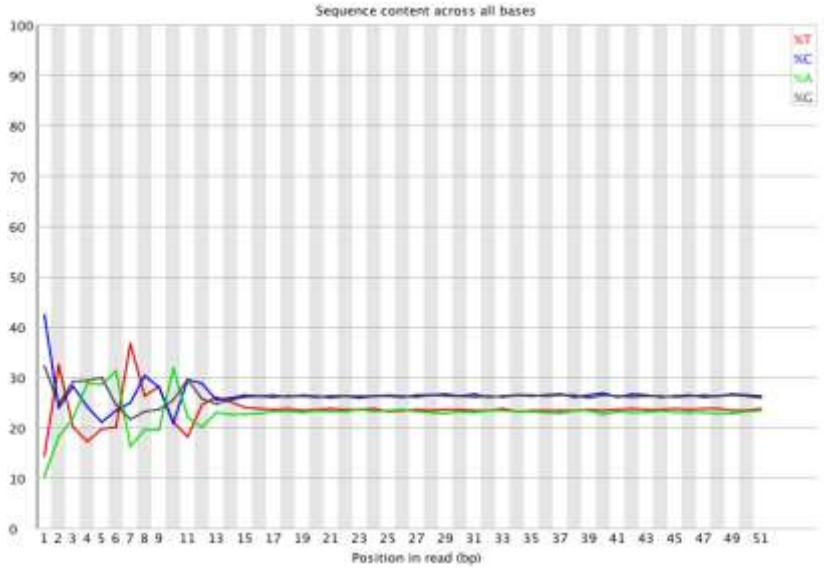
Wang, Z et al. (2009) Nature Reviews Genetics 10:57-63

profilowanie ekspresji genów - RNA-seq illumina Truseq



http://bitesizebio.com

profilowanie ekspresji genów - RNA-seq illumina Truseq



<u>Znaczenie dla analizy:</u>

- możliwość występowania preferencji w rozkładzie nukleotydów na początku sekwencji (związane z losowymi starterami)
- metoda daje dobrą „losowość” fragmentacji - występowanie duplikatów spowodowane jest z dużym prawdopodobieństwem reakcją PCR
- fragmentacja i selekcja odpowiedniej długości insertów powoduje prawie całkowity brak 3’ adaptorów
- utrata informacji o nici z której pochodzi transkrypt – podczas zliczania odczytów należy uwzględnić obie nici

profilowanie ekspresji genów - RNA-seq illumina Truseq

Zalety:

- Minimalna ilość manipulacji próbką przed amplifikacją
- Wysoka dokładność w oznaczaniu poziomu ekspresji genów
- Wysoka jakość biblioteki
- Wysoka dokładność w identyfikacji SNP

Wady:

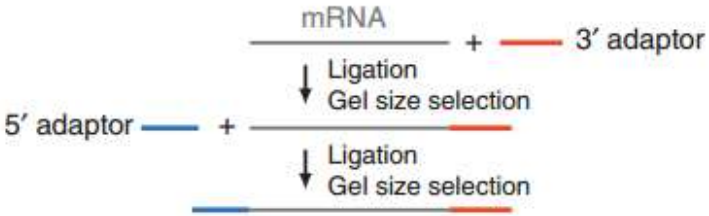
- Utrata informacji o nici z której pochodzą transkrypty

Zastosowanie:

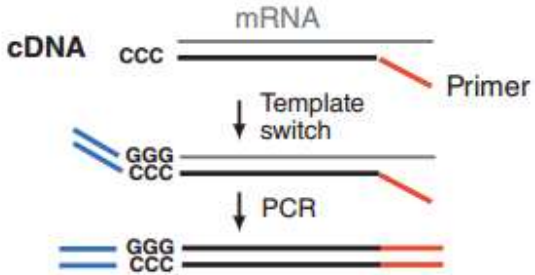
- Badanie poziomu ekspresji GENÓW
- Badanie aktywności transkrypcyjnej genomu
- Badanie polimorfizmu nukleotydowego na poziomie transkryptomu

profilowanie ekspresji genów - RNA-seq
wprowadzenie adaptorów

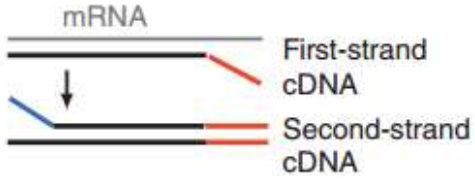
ligacja adaptorów do RNA



SMART



losowe primery z adaptorami



Levin, JZ et al. (2010) Nat. Protcols 9:709

profilowanie ekspresji genów - RNA-seq wprowadzenie adaptorów

Znaczenie dla analizy:

- Utrzymanie informacji o kierunku transkrypcji – przy zliczaniu można rozróżnić geny kodowane na przeciwnych niciach
- Trakty homonukleotydowe dodawane w metodzie SMART bywają trudne do rozróżnienia od naturalnych traktów homopolimerowych
- Możliwe do zaobserwowania preferencje sekwencyjne na obu końcach odczytów związane z preferencjami ligazy

profilowanie ekspresji genów - RNA-seq wprowadzenie adaptorów

Zalety:

- Utrzymanie informacji o kierunku transkrypcji
- Możliwe uzyskanie wariantów splicingowych w skomplikowanych genomach

Wady:

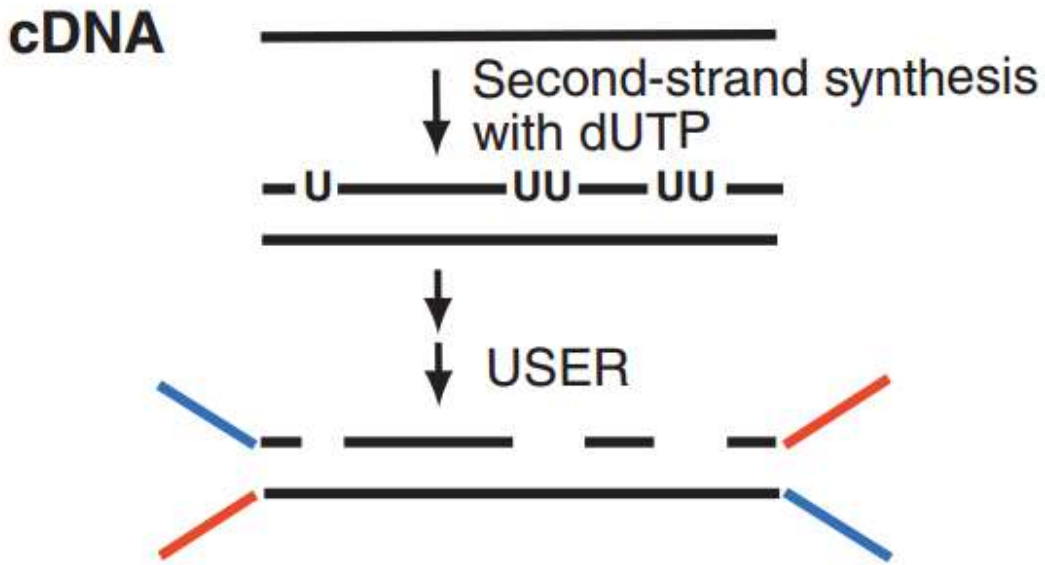
- Zaburzenia odczytywanego poziomu transkryptów poprzez użycie adaptorów
- Stosunkowo niska reproduktywność wyników
- Utrudniona analiza sekwencji

Zastosowanie:

- Badanie poziomu ekspresji GENÓW i TRANSKRYPTÓW
- Badanie aktywności transkrypcyjnej genomu
- Badanie struktury genów
- Identyfikacja izoform splicingowych

profilowanie ekspresji genów - RNA-seq dUTP

usuwanie drugiej nici cDNA z zastosowaniem dUTP



Levin, JZ et al. (2010) Nat. Protcols 9:709

profilowanie ekspresji genów - RNA-seq dUTP

Zalety:

- Utrzymanie informacji o kierunku transkrypcji
- Możliwe uzyskanie wariantów splicingowych w skomplikowanych genomach
- Jakość biblioteki i powtarzalność wyników porównywalna z Truseq

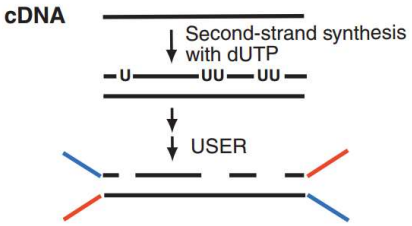
Wady:

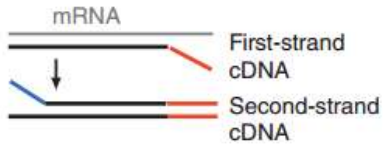
- Stosunkowo skomplikowana procedura

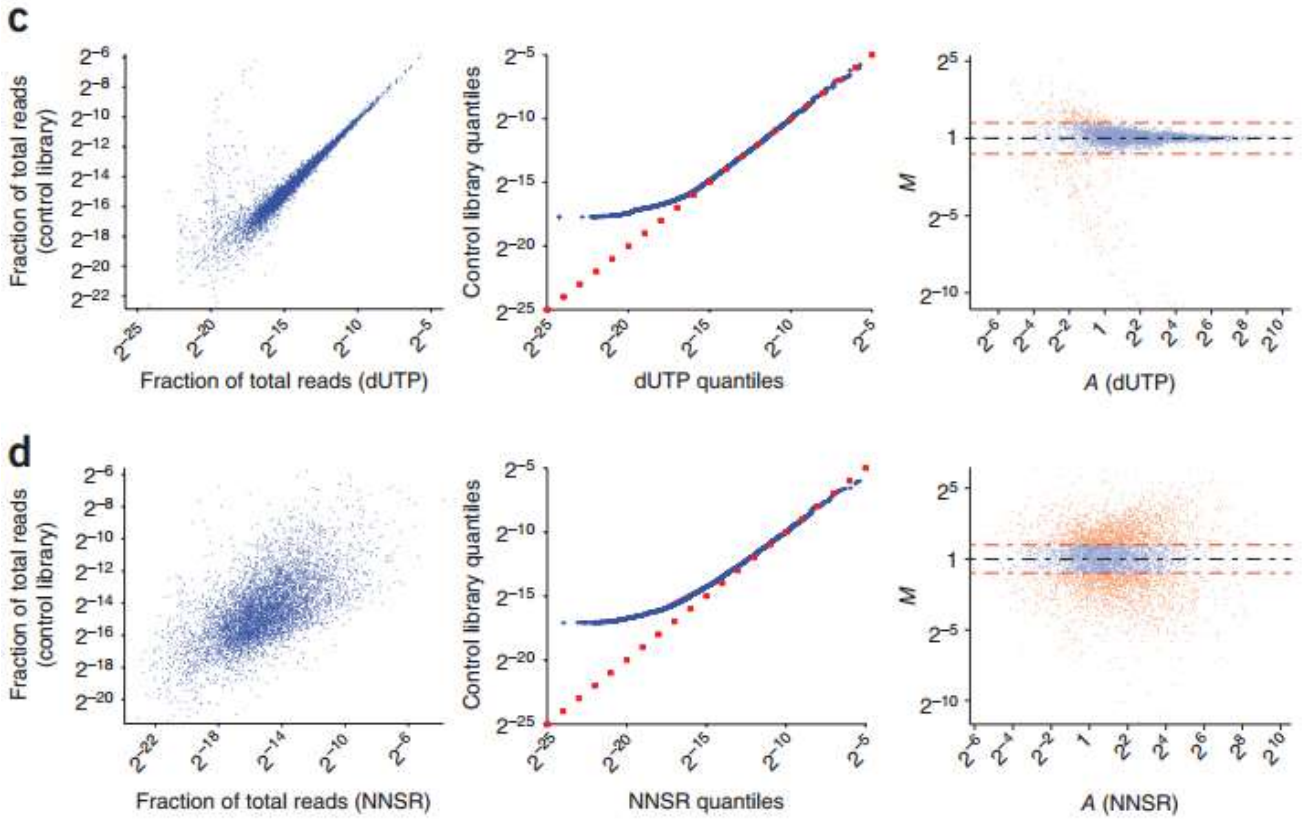
Zastosowanie:

- Badanie poziomu ekspresji GENÓW i TRANSKRYPTÓW
- Badanie aktywności transkrypcyjnej genomu
- Badanie struktury genów
- Identyfikacja izoform splicingowych
- Badanie polimorfizmu nukleotydowego na poziomie transkryptomu

profilowanie ekspresji genów - RNA-seq dUTP

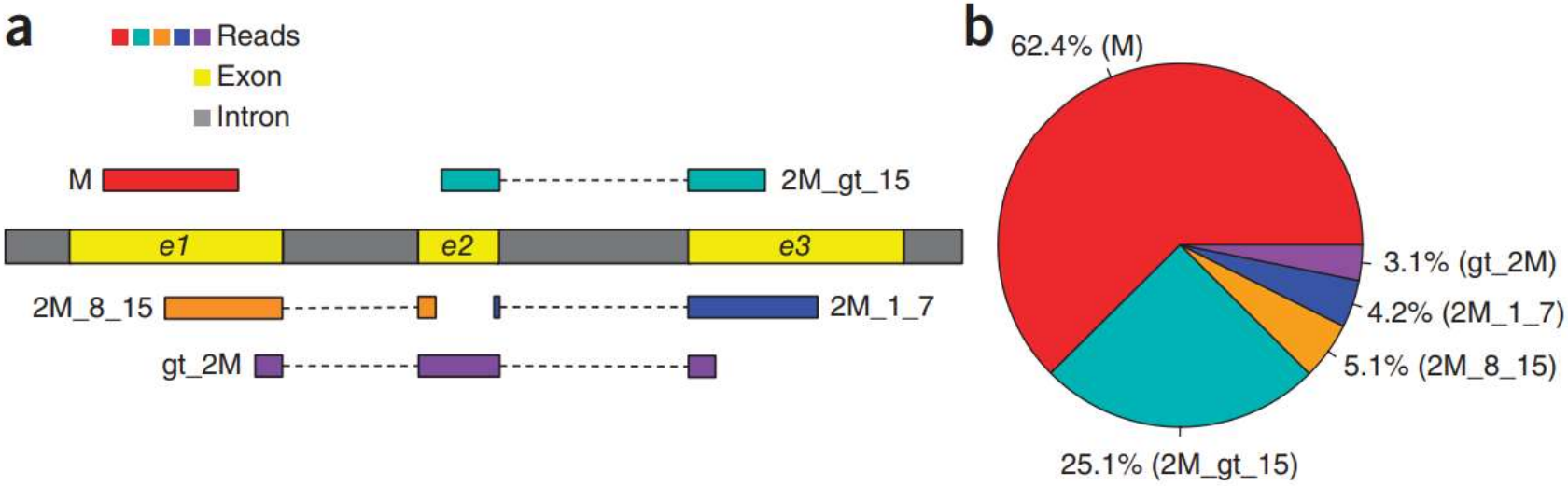






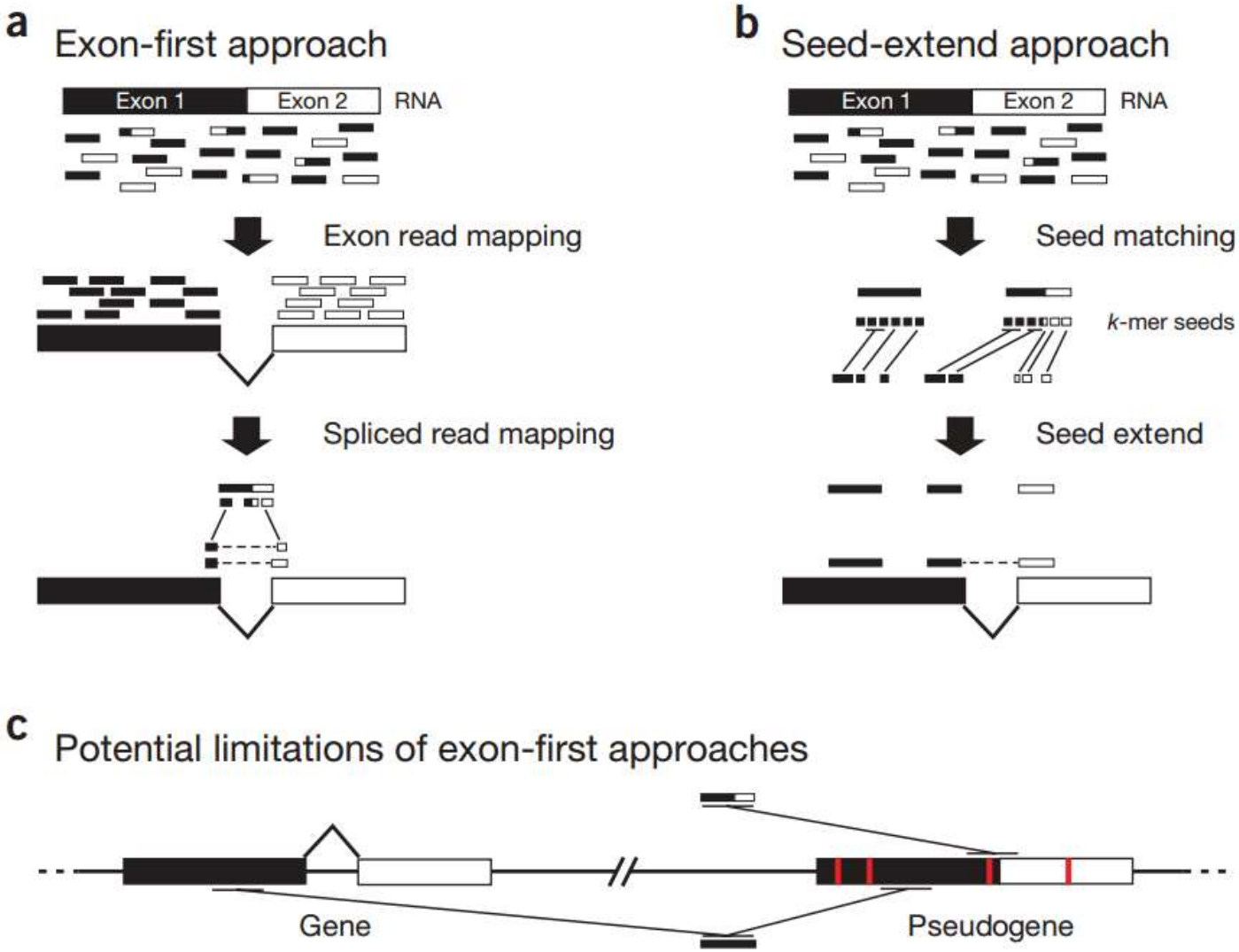
Levin, JZ et al. (2010) Nat. Protocols 9:709

dopasowanie z przerwami?

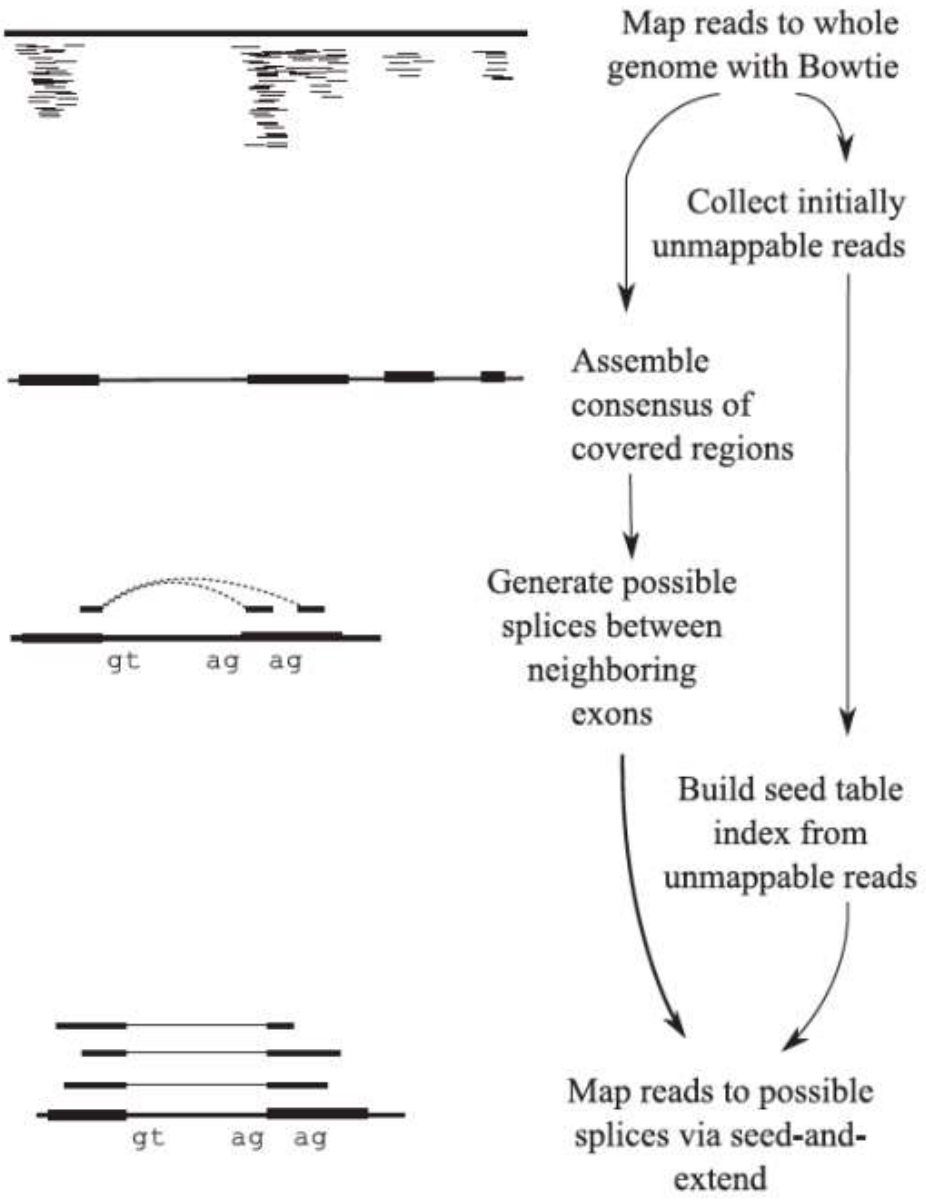


Tak! Dla RNA-Seq!

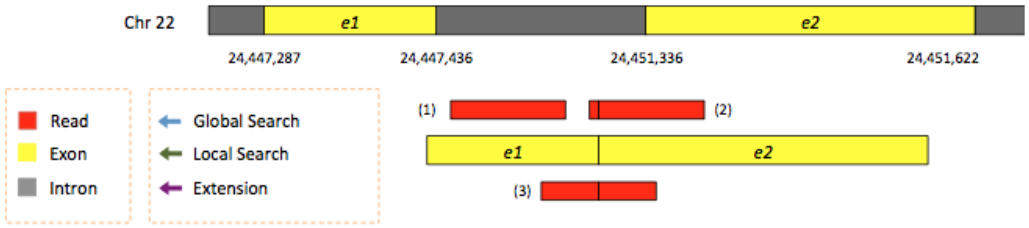
dopasowanie z przerwami

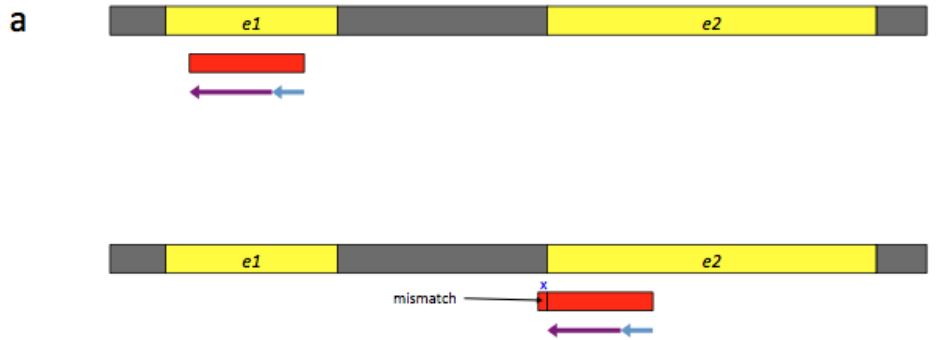


TopHat

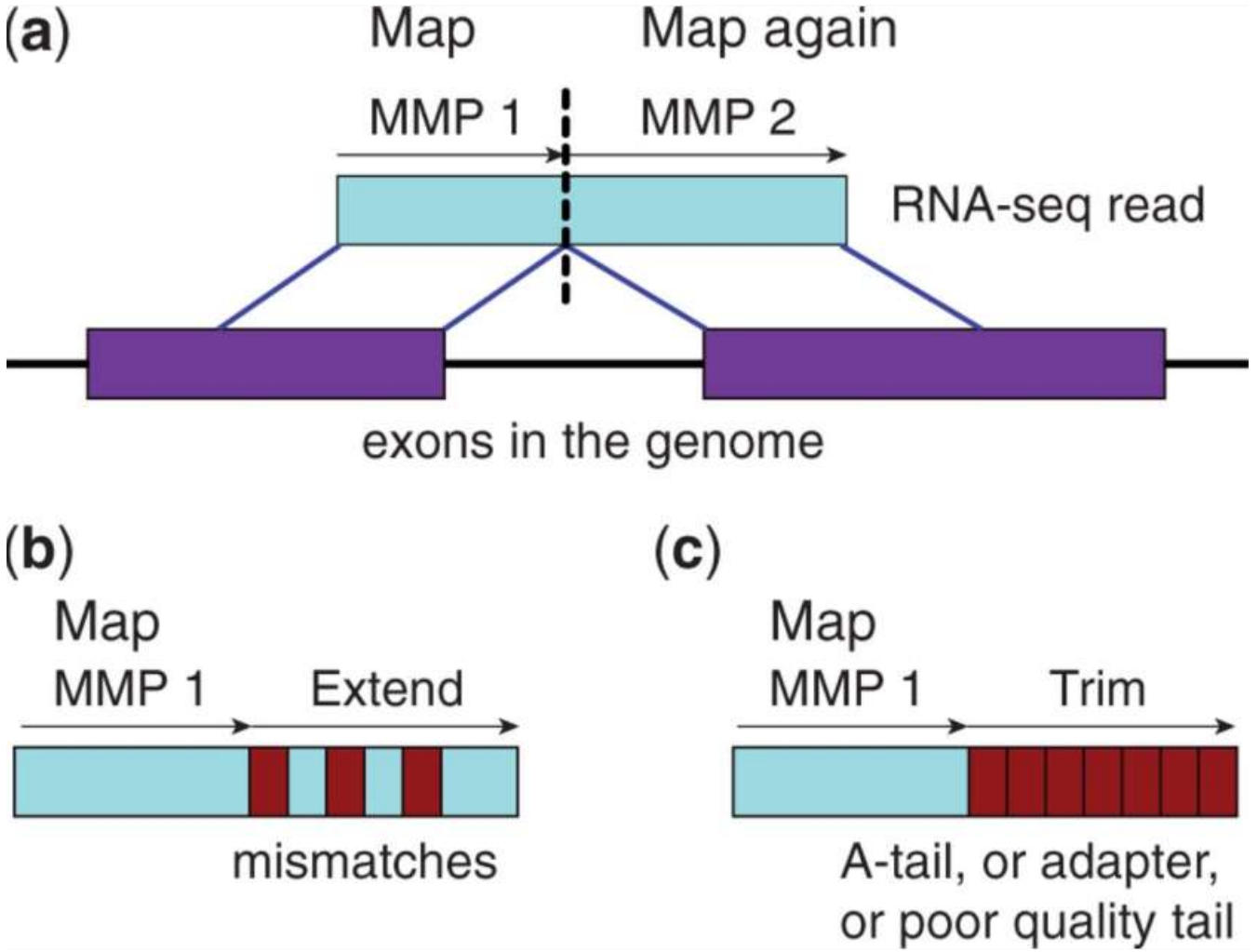

Map reads to whole genome with Bowtie
Collect initially unmappable reads
Assemble consensus of covered regions
Generate possible splices between neighboring exons
Build seed table index from unmappable reads
Map reads to possible splices via seed-and-extend
gt ag ag
gt ag ag

HISAT

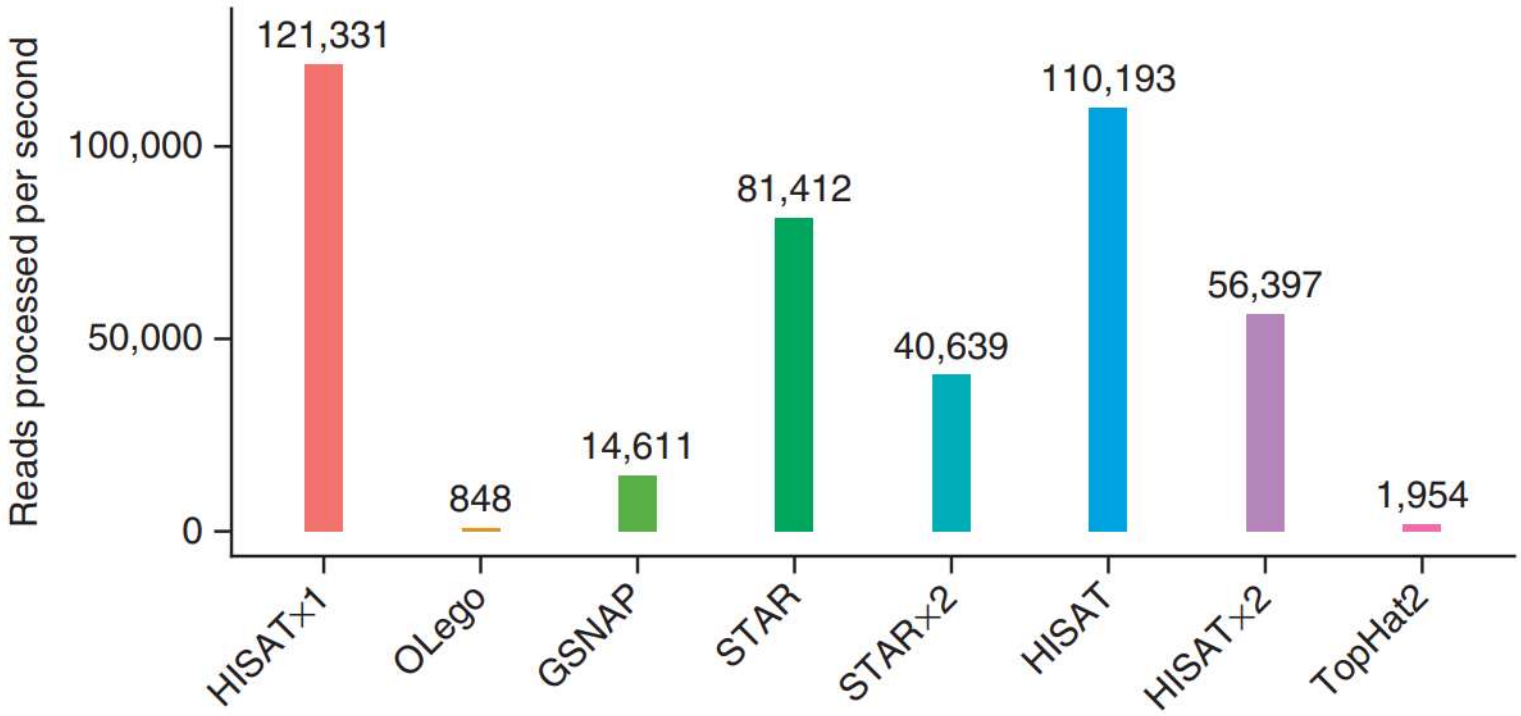

Chr 22
e1
e2
24,447,287
24,447,436
24,451,336
24,451,622
Read
Exon
Intron
Global Search
Local Search
Extension
(1)
(2)
(3)


a
e1
e2
b
mismatch
Local FM index for chr22 from 24,417,280 to 24,482,559
24,447,287
24,451,622
c

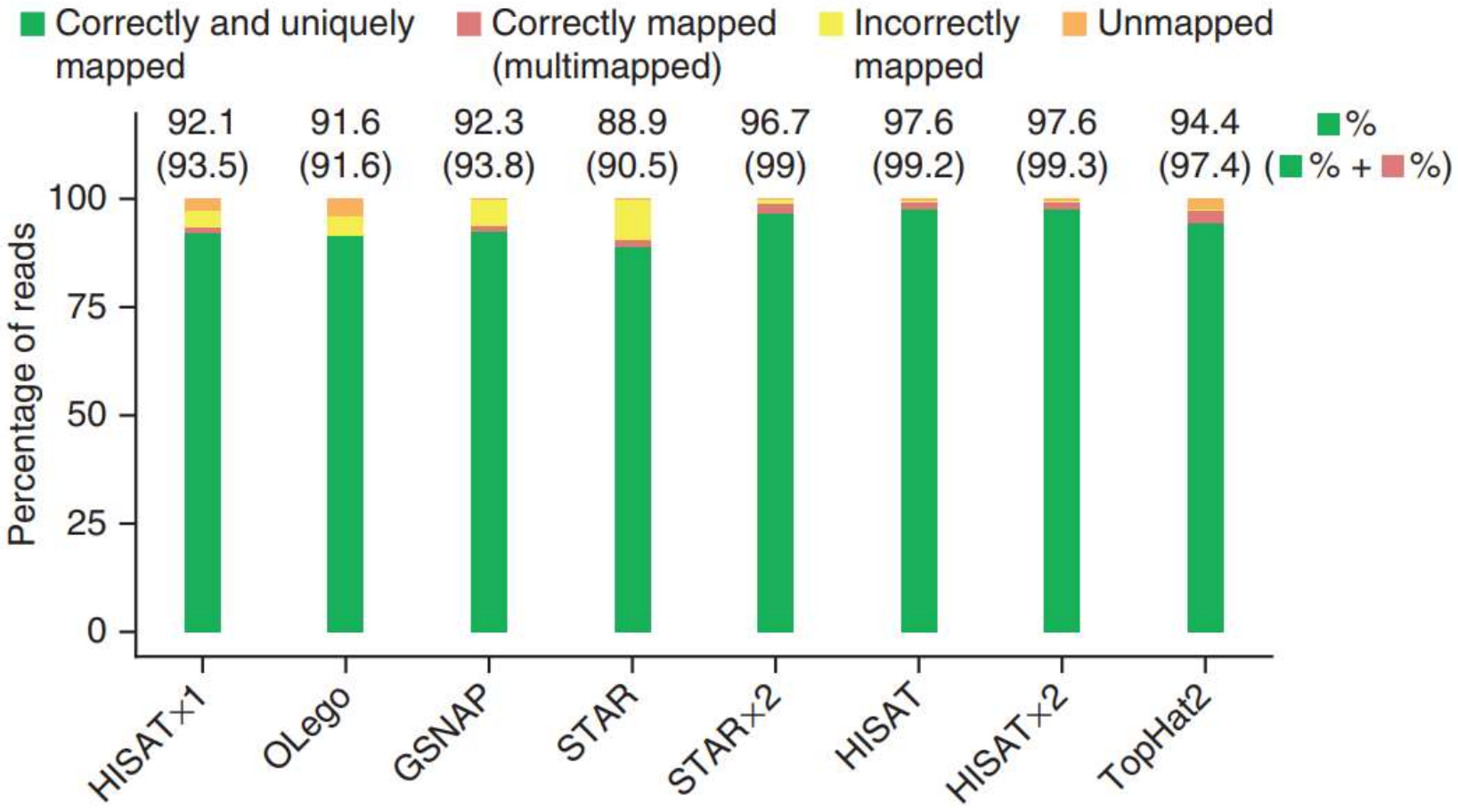
STAR


(a)
Map
Map again
MMP 1
MMP 2
RNA-seq read
exons in the genome
(b)
Map
MMP 1
Extend
mismatches
(c)
Map
MMP 1
Trim
A-tail, or adapter,
or poor quality tail

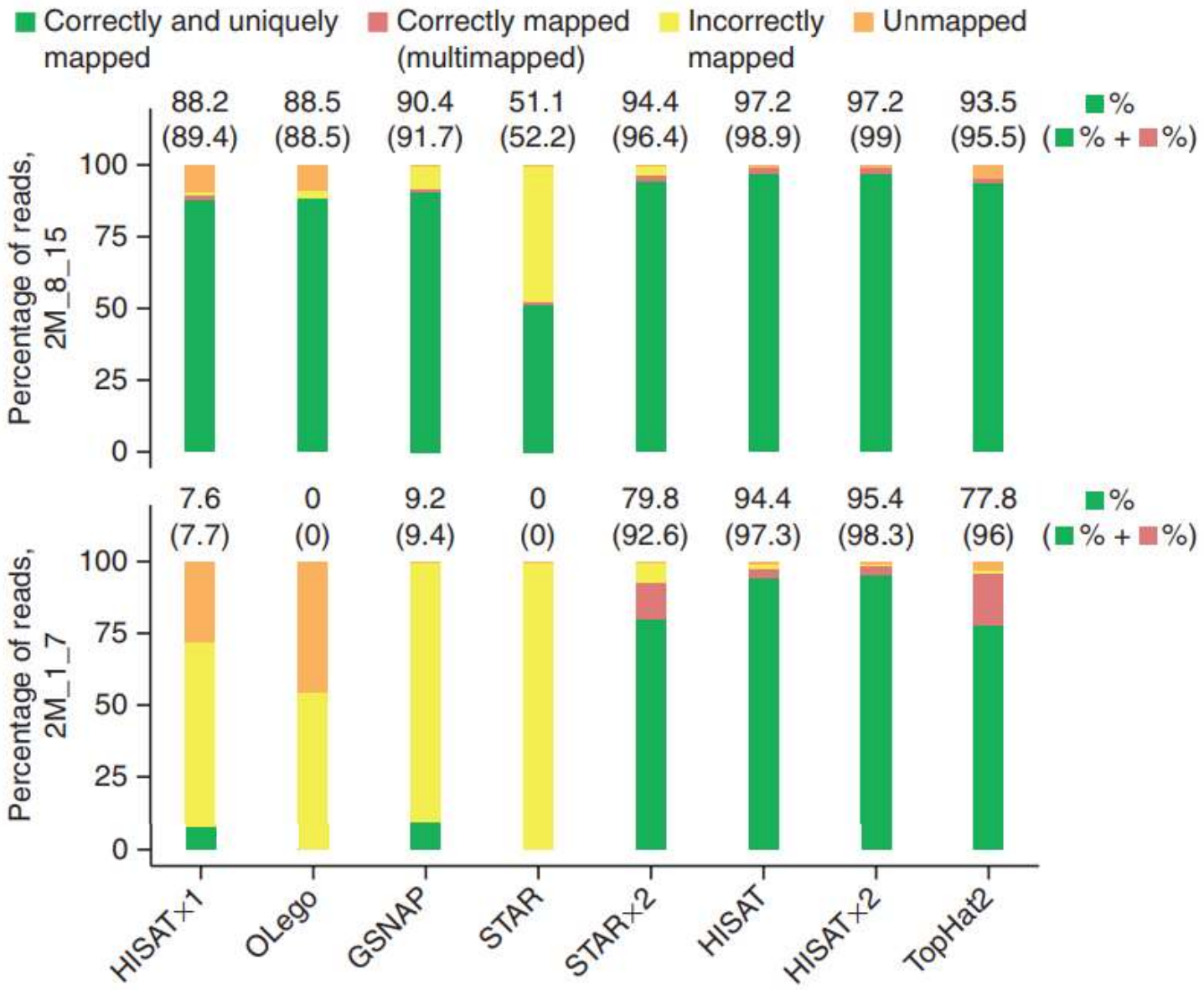
czas działania



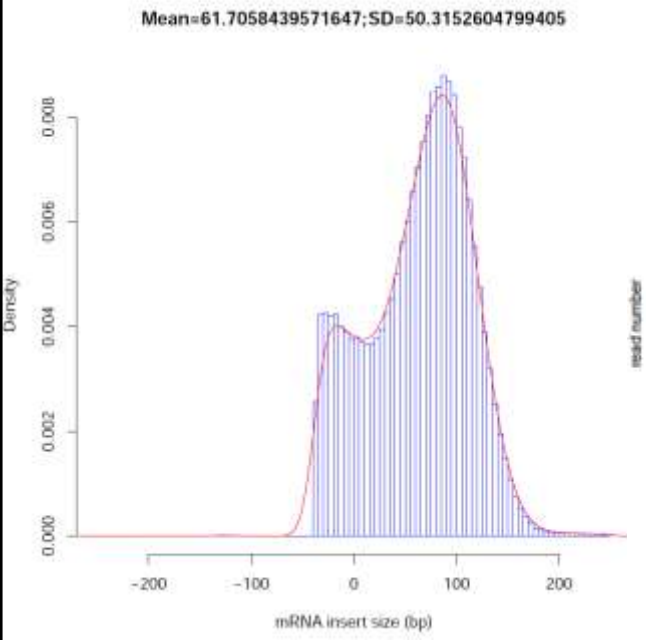
dokładność

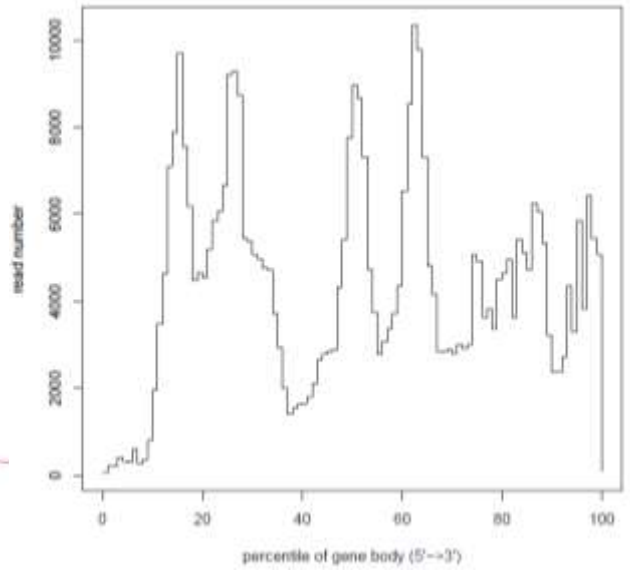

Correctly and uniquely mapped
Correctly mapped (multimapped)
Incorrectly mapped
Unmapped
92.1 (93.5)
91.6 (91.6)
92.3 (93.8)
88.9 (90.5)
96.7 (99)
97.6 (99.2)
97.6 (99.3)
94.4 (97.4)
% (% + %)
Percentage of reads
100
75
50
25
0
HISAT×1
OLego
GSNAP
STAR
STAR×2
HISAT
HISAT×2
TopHat2

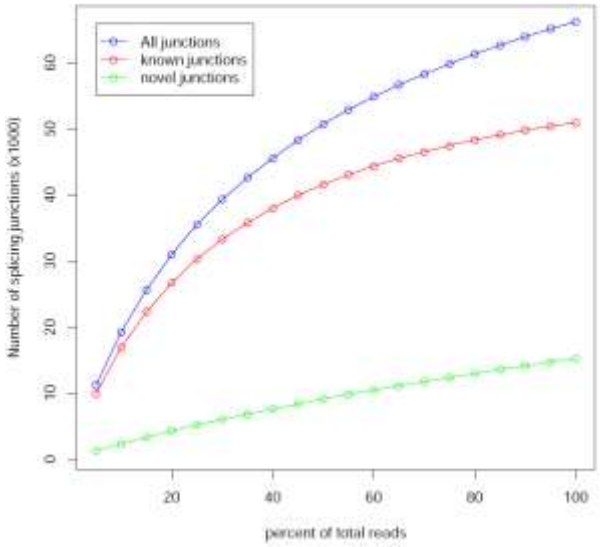
dokładność


Correctly and uniquely mapped
Correctly mapped (multimapped)
Incorrectly mapped
Unmapped
88.2 (89.4)
88.5 (88.5)
90.4 (91.7)
51.1 (52.2)
94.4 (96.4)
97.2 (98.9)
97.2 (99)
93.5 (95.5)
% (% + %)
Percentage of reads, 2M_8_15
100
75
50
25
0
7.6 (7.7)
0 (0)
9.2 (9.4)
0 (0)
79.8 (92.6)
94.4 (97.3)
95.4 (98.3)
77.8 (96)
% (% + %)
Percentage of reads, 2M_1_7
HISAT×1
OLego
GSNAP
STAR
STAR×2
HISAT
HISAT×2
TopHat2

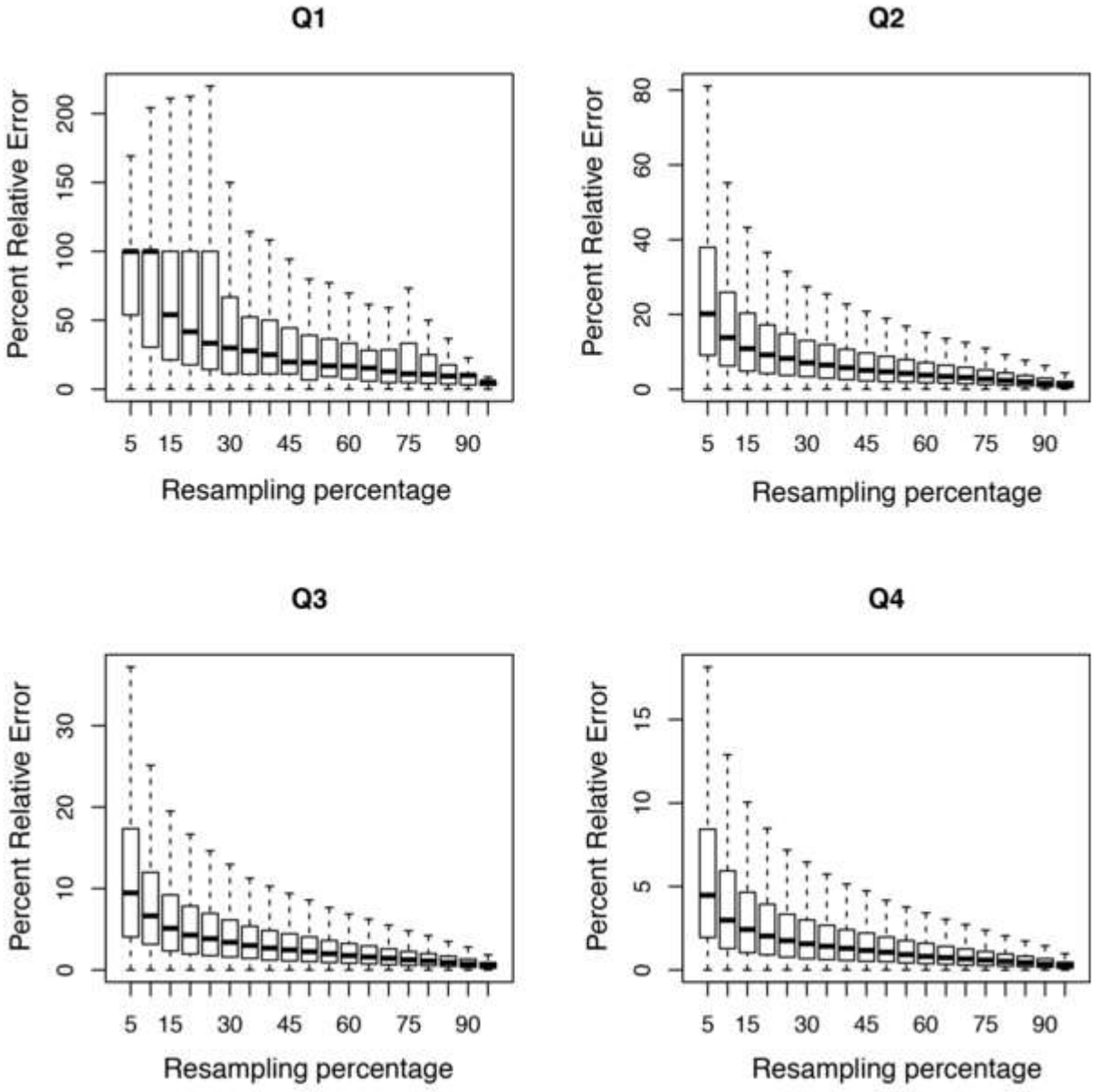
weryfikacja wyników mapowania: RseQC



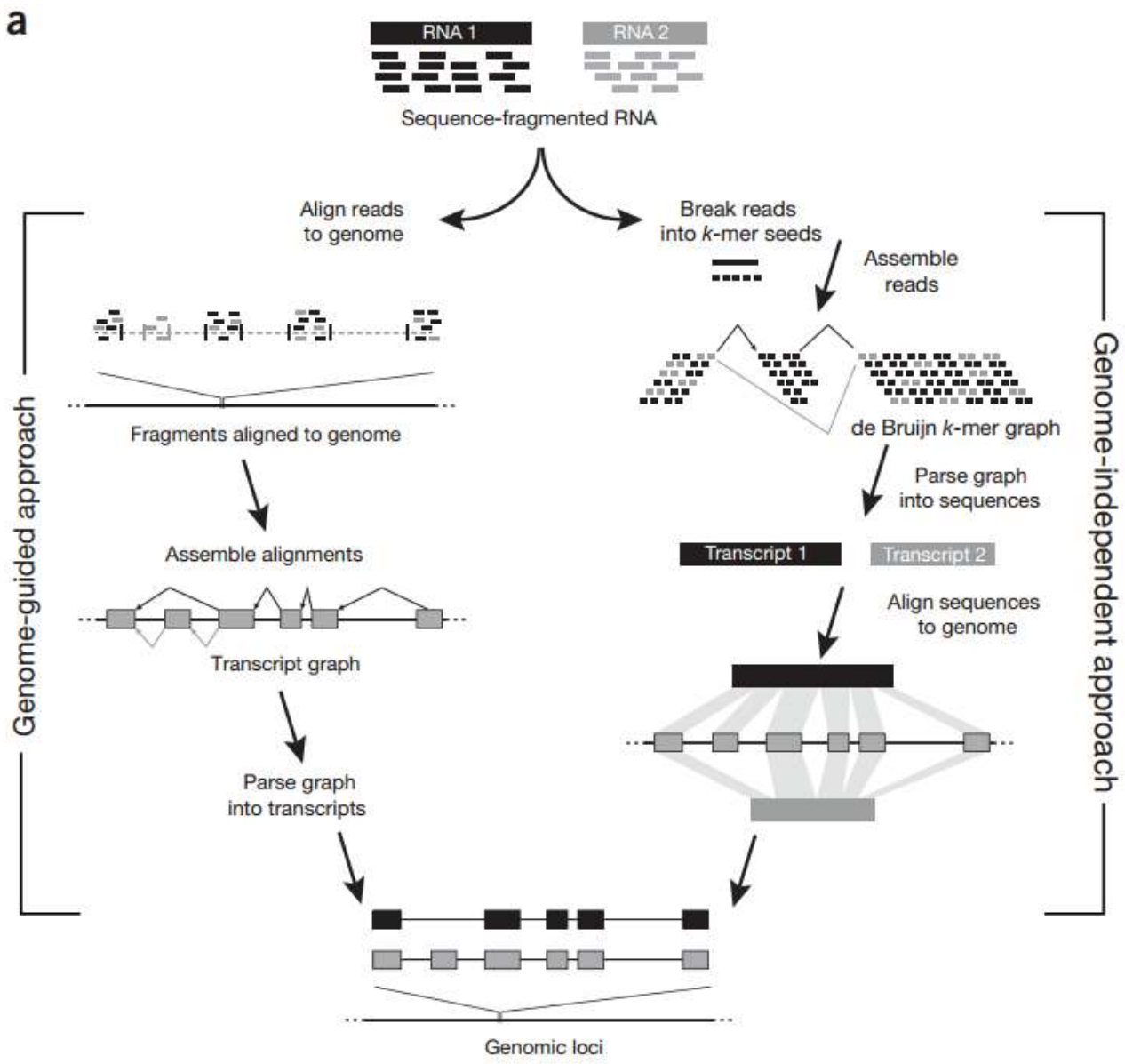




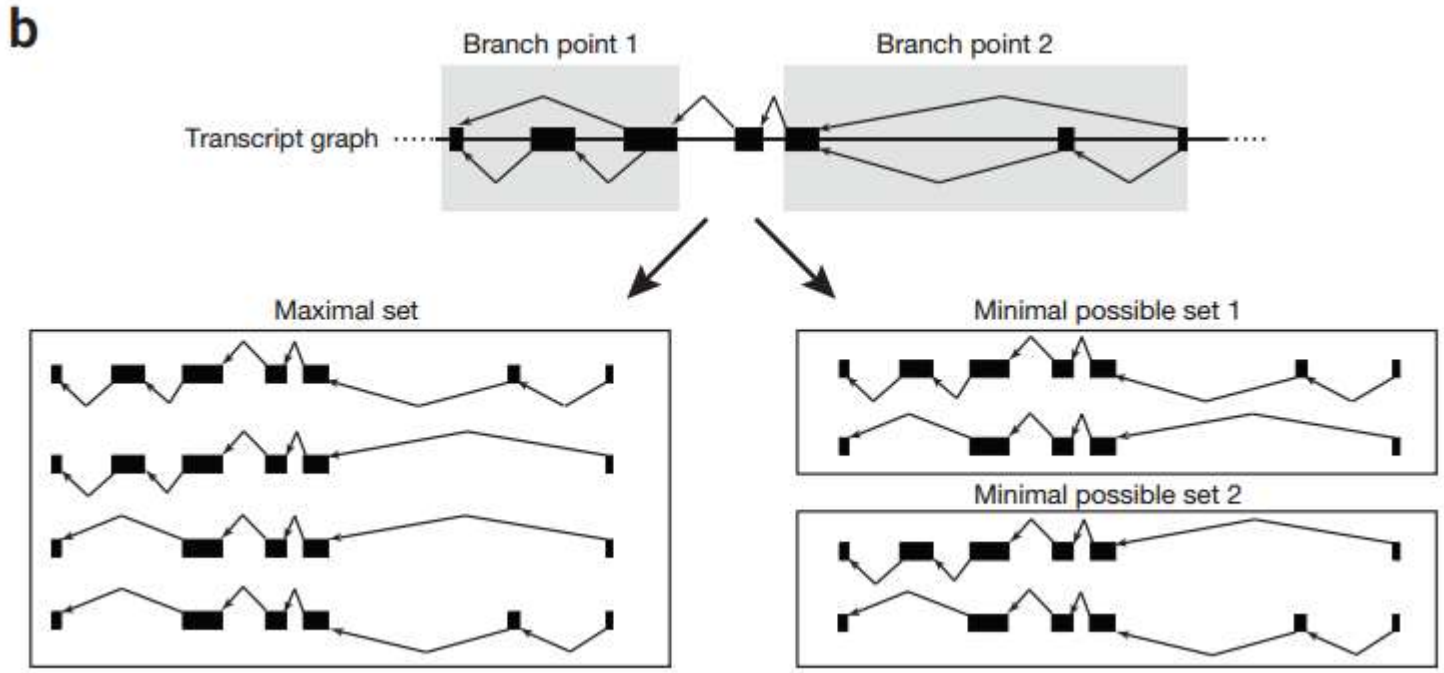
weryfikacja wyników mapowania: RseQC



składanie transkryptów



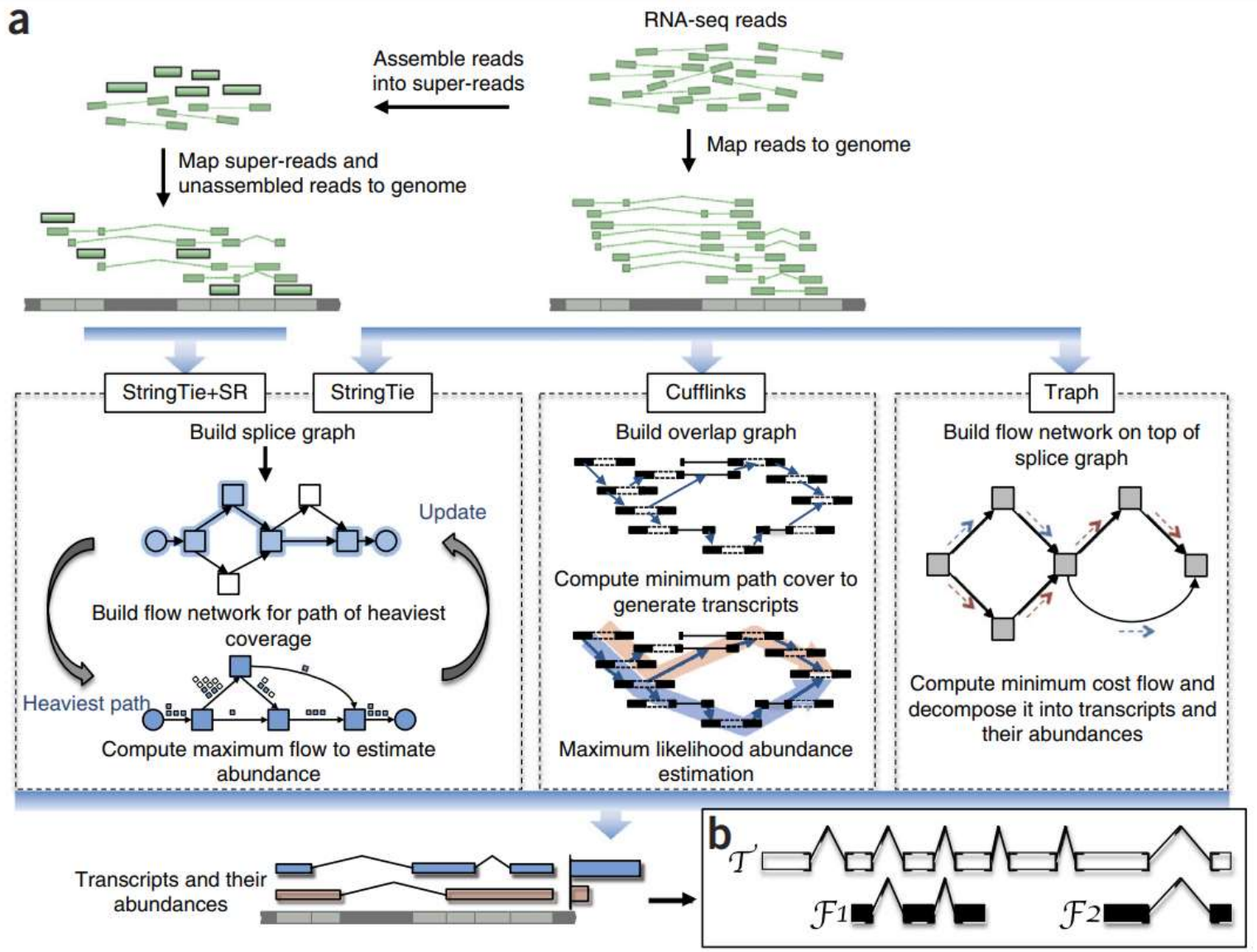
składanie transkryptów



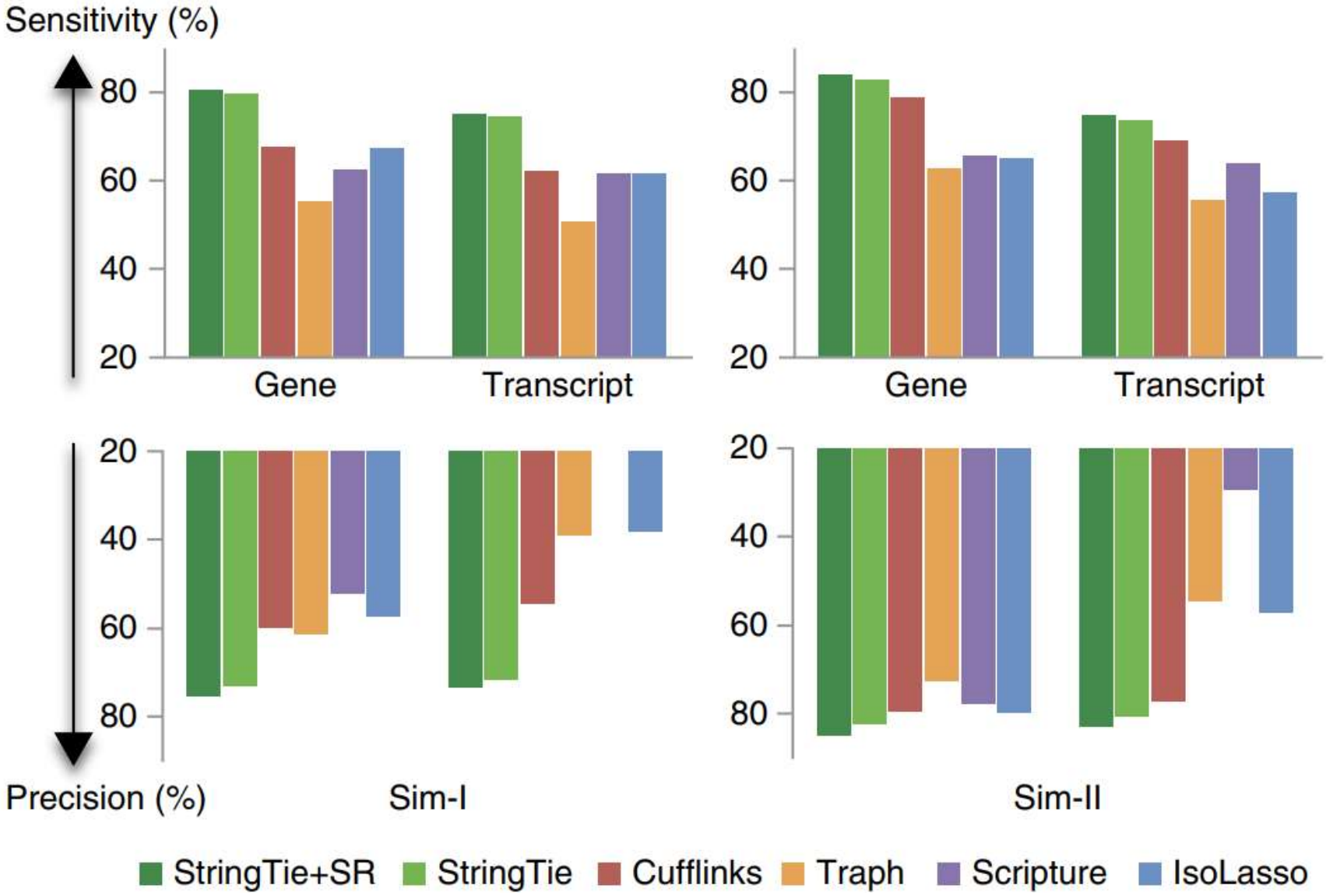
sensitivity
(Scripture)

precision
(Cufflinks)

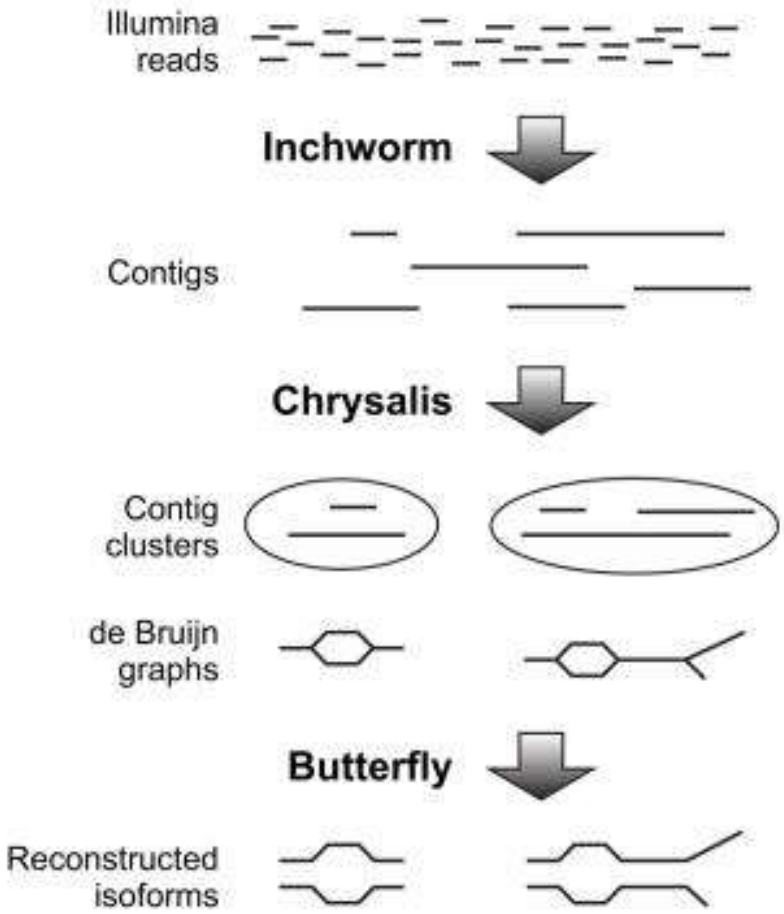
składanie transkryptów

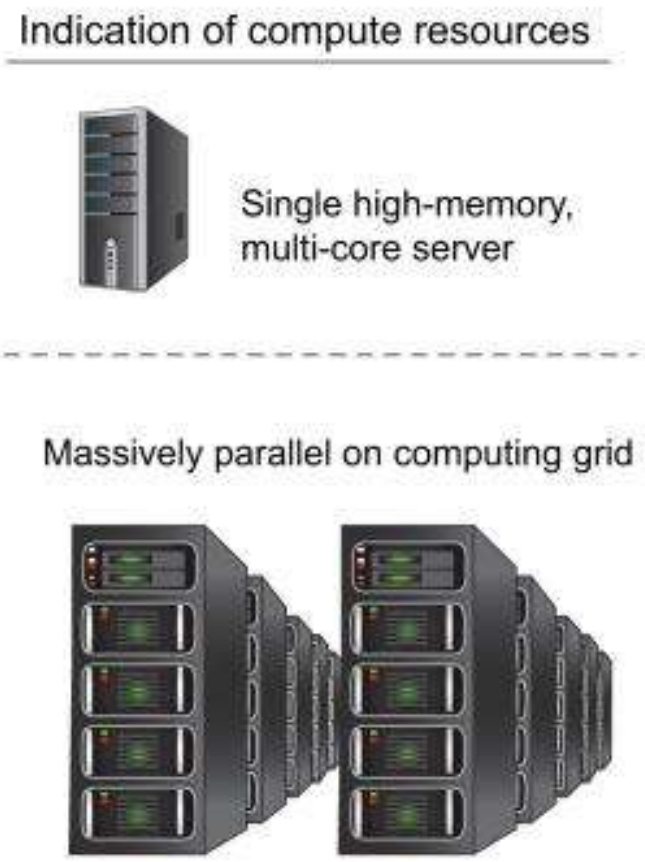


składanie transkryptów

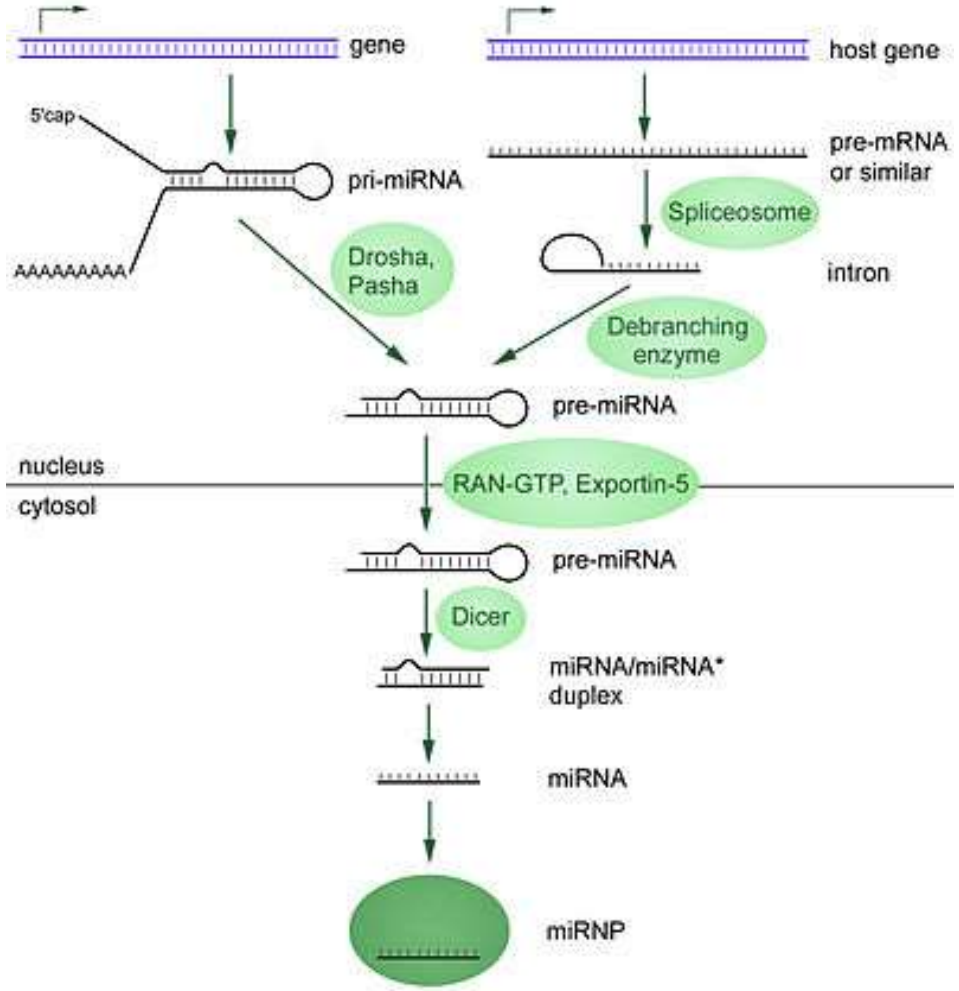


składanie transkryptomu *de novo* Trinity





small RNA-seq

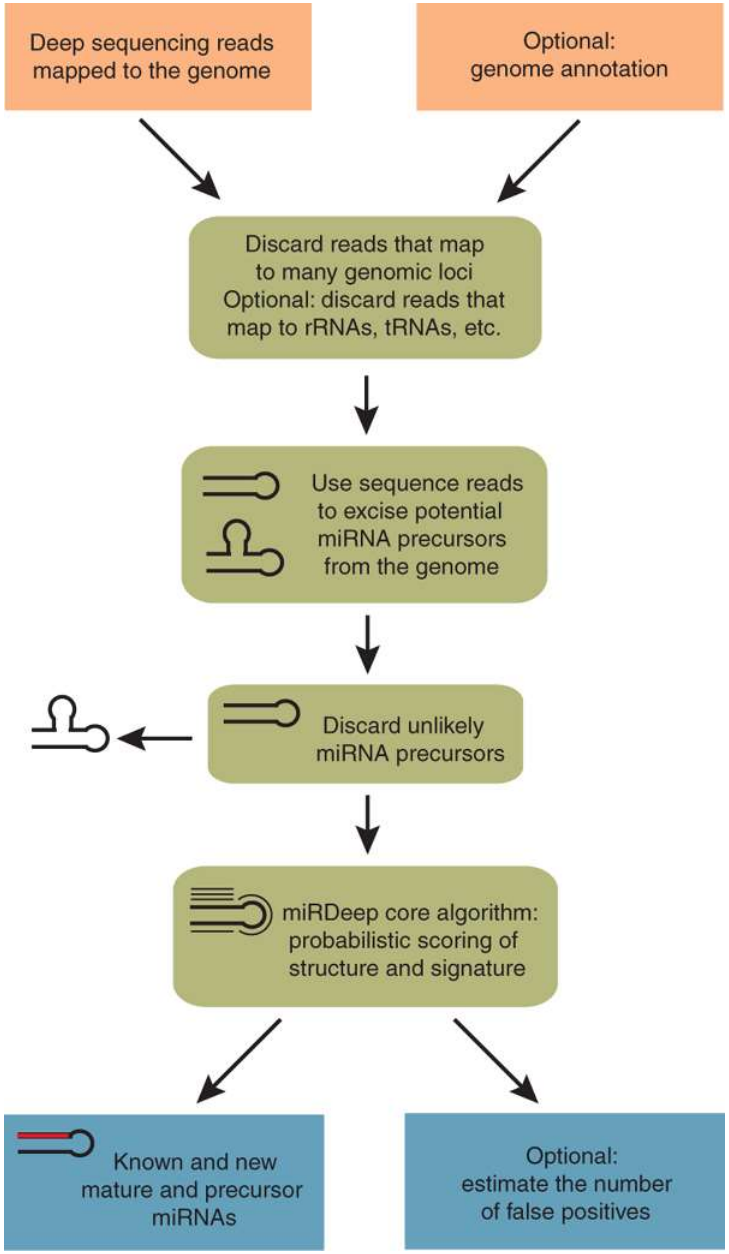


http://wikipedia.org

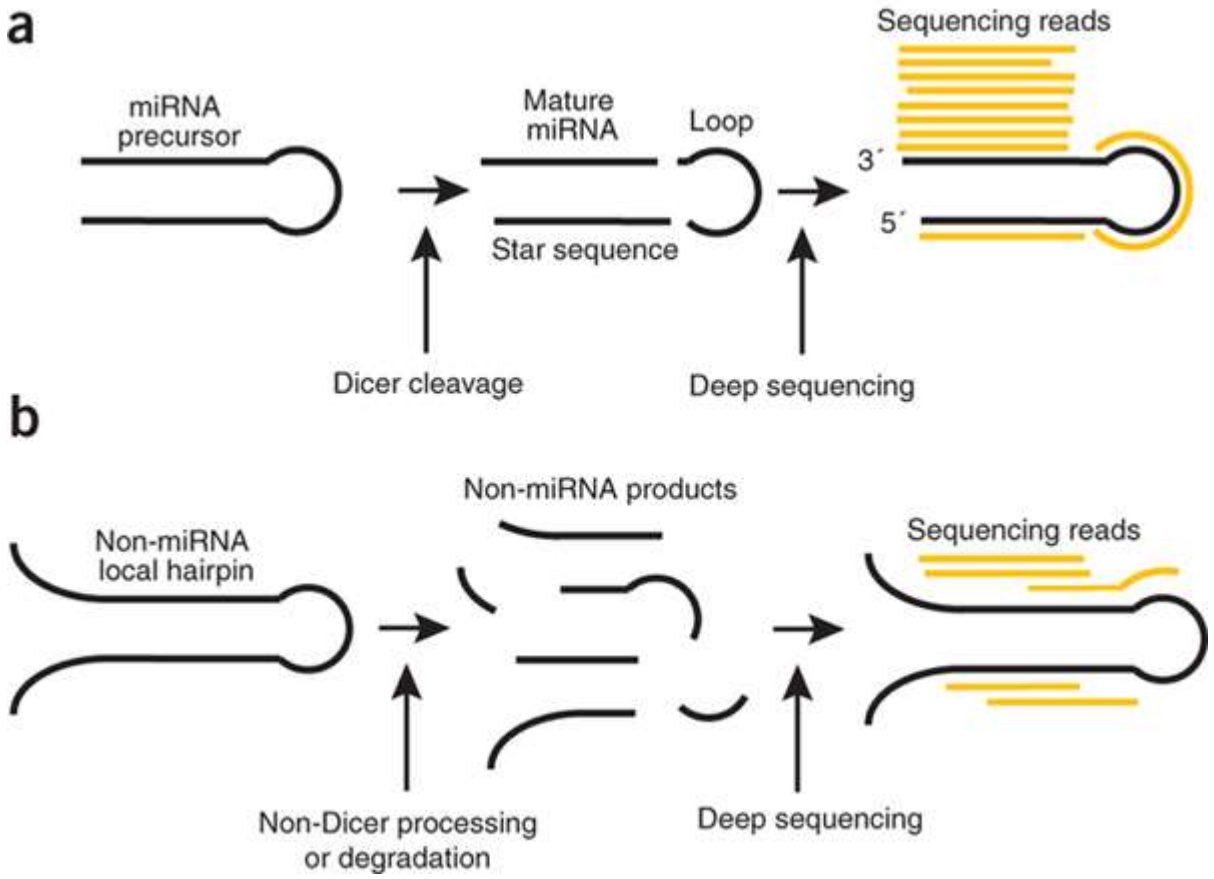
small RNA-seq

1. Izolacja małych RNA (18-30nt)
2. Brak fragmentacji
3. Sekwencjonowanie pojedynczego końca
4. Mapowanie bez uwzględnienia splicingu

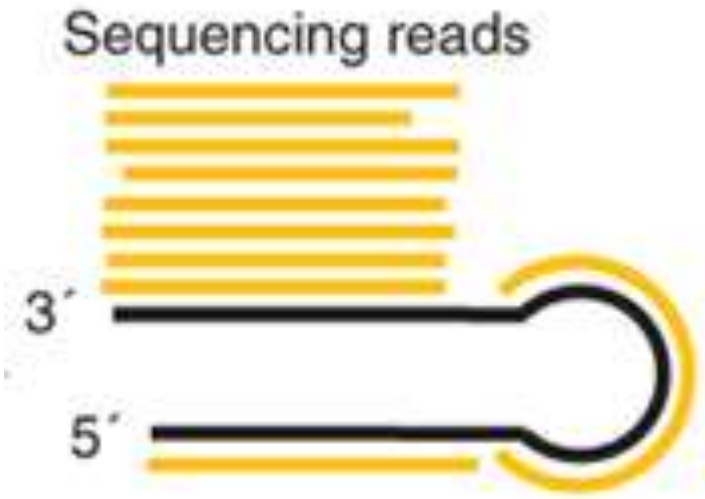
small RNA-seq
mirDeep



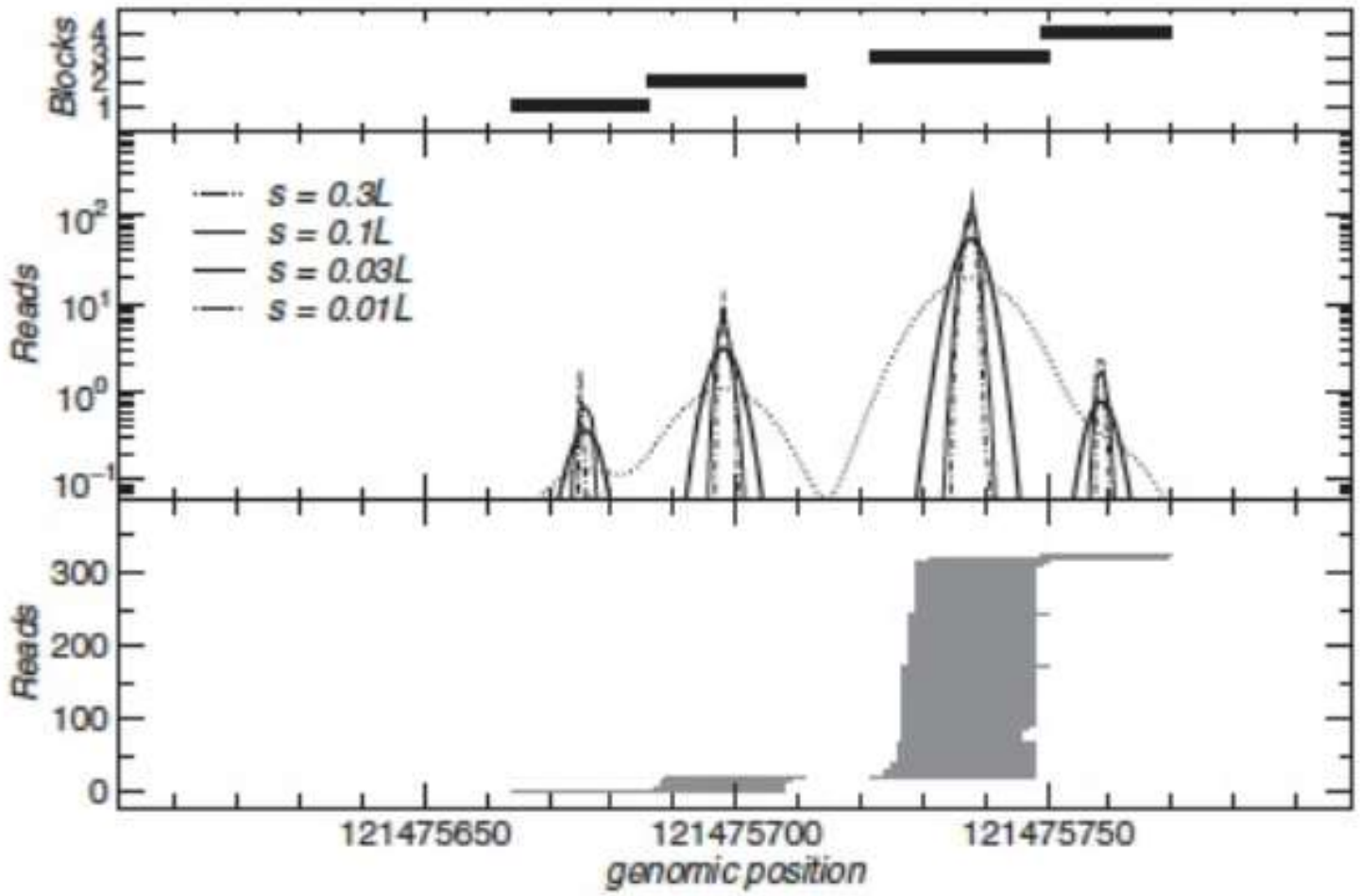
small RNA-seq
mirDeep



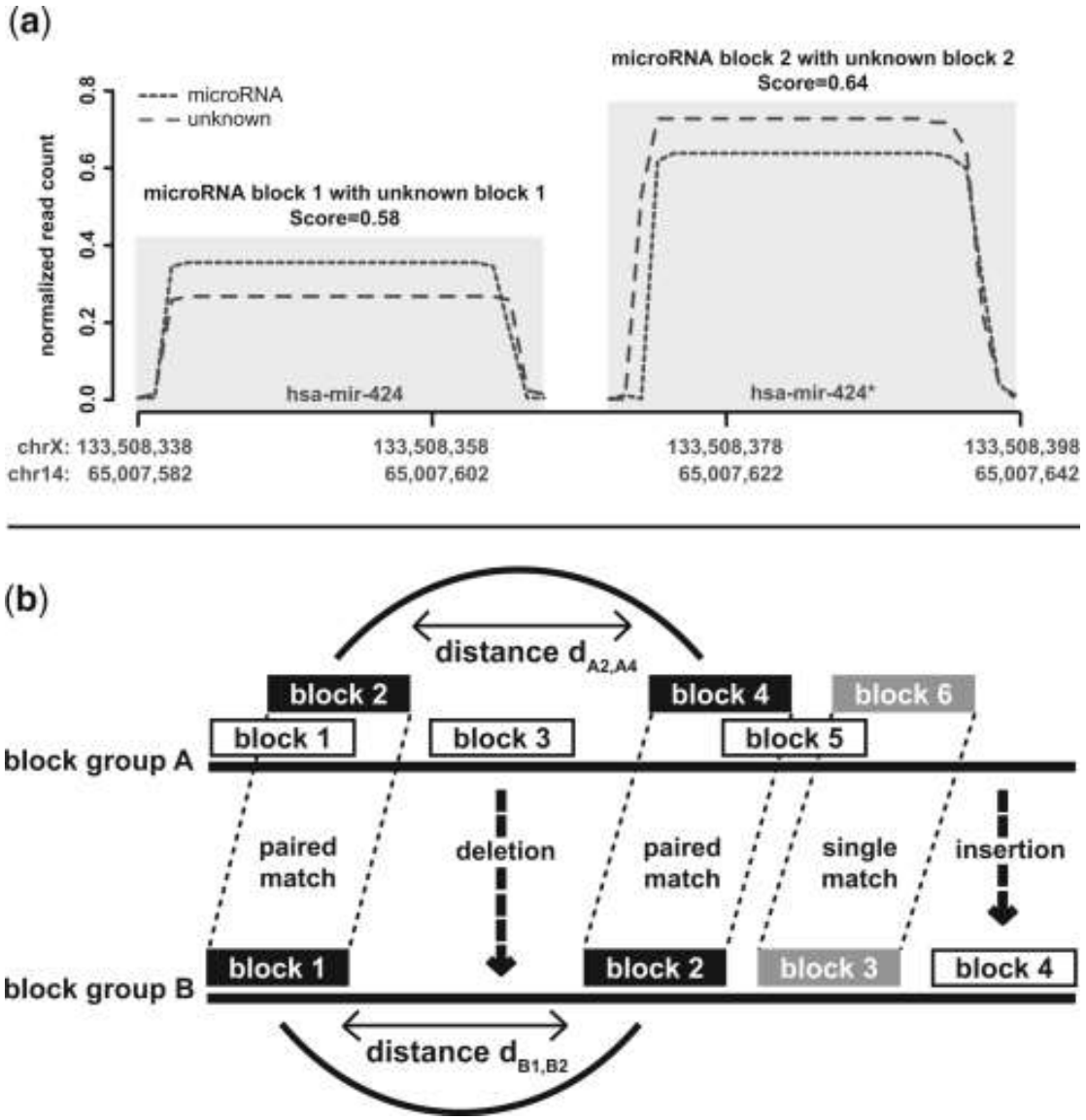
small RNA-seq
poziom ekspresji a duplikaty



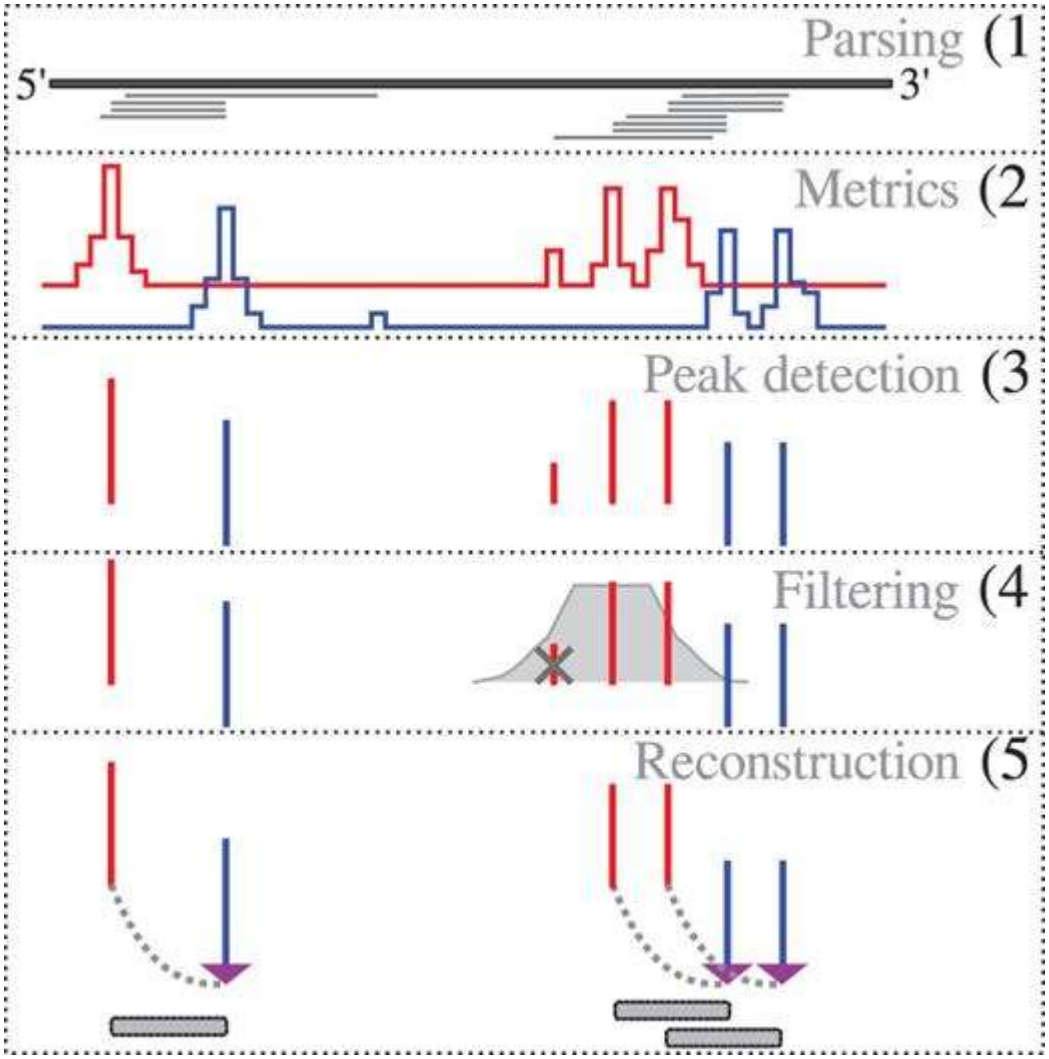
small RNA-seq blockbuster



small RNA-seq deepBlockAlign



small RNA-seq
FlaiMapper



zliczanie poziomu ekspresji

Zliczanie na podstawie adnotacji odczytów przypadających na:
- gen
- transkrypt
- egzon
- …

Programy:

HTSeq-count
- popularny
- bardzo wolny
- zwraca również geny nie ulegające ekspresji

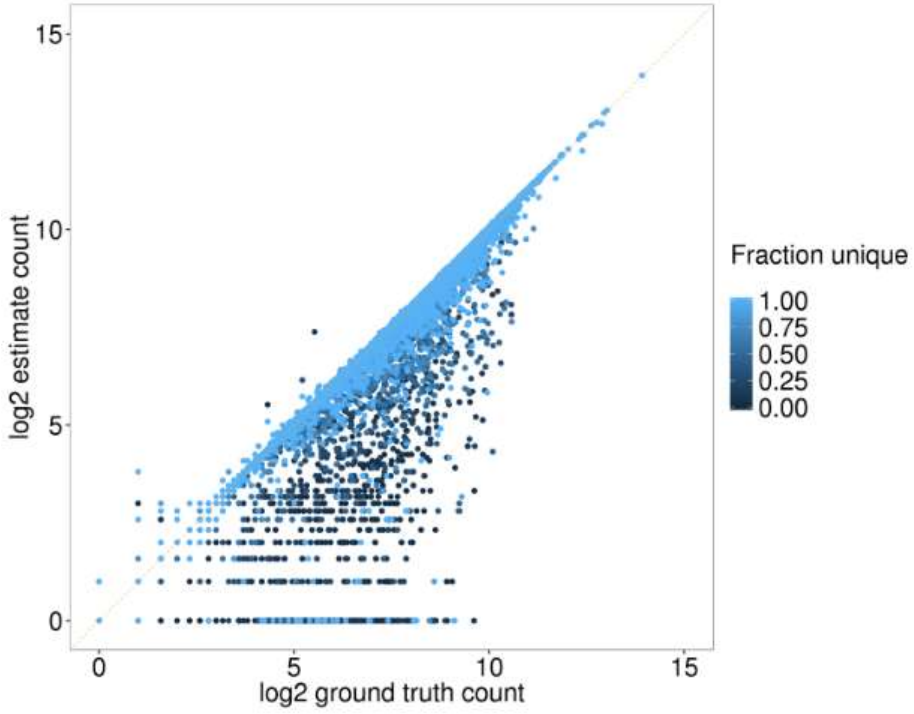
FeatureCounts:
- szybki
- analiza wielu próbek jednocześnie
- duża kontrola nad sposobem zliczania

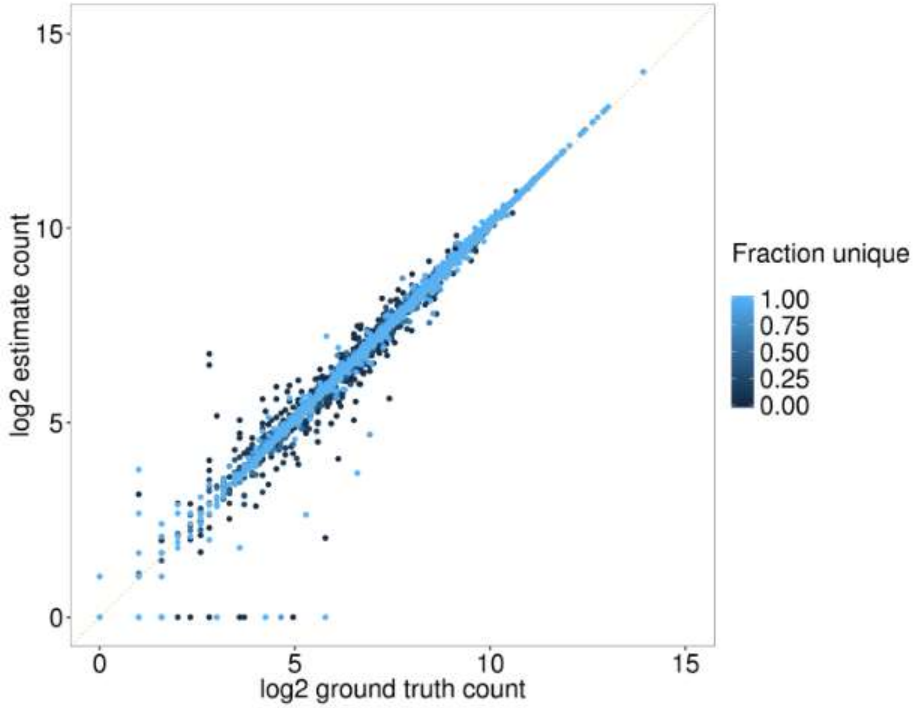
metody alignment-free

- Brak konieczności mapowania odczytów
- Dopasowanie odczytów na podstawie dystansu (np. częstotliwości k-merów)
- Dokładność niezależna od „mapowalności” odczytów
- Bardzo szybkie
- Wyniki problematyczne w interpretacji podczas testowania różnicowej ekspresji

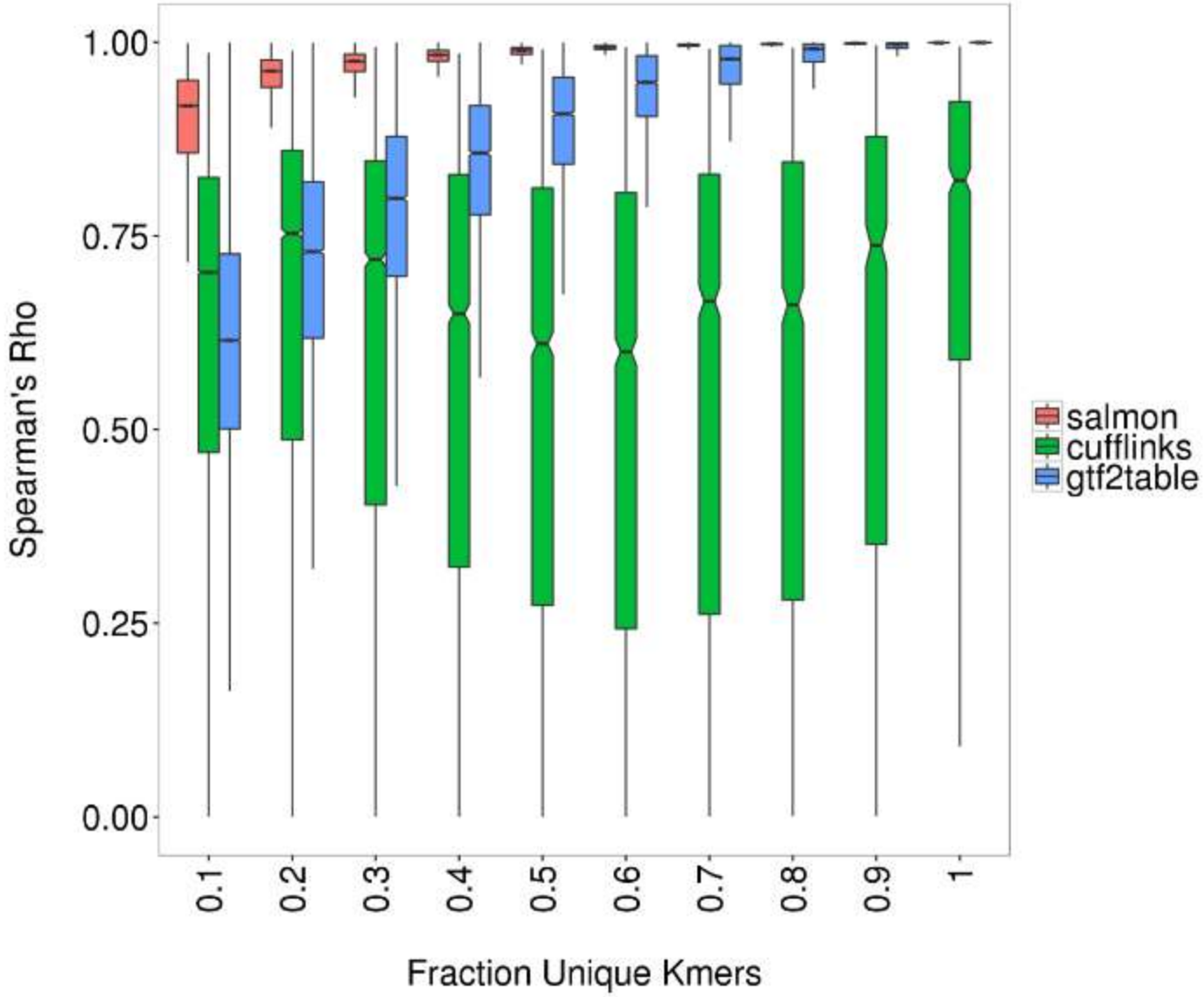
- Kallisto
- Sailfish
- Salomon

metody alignment-free

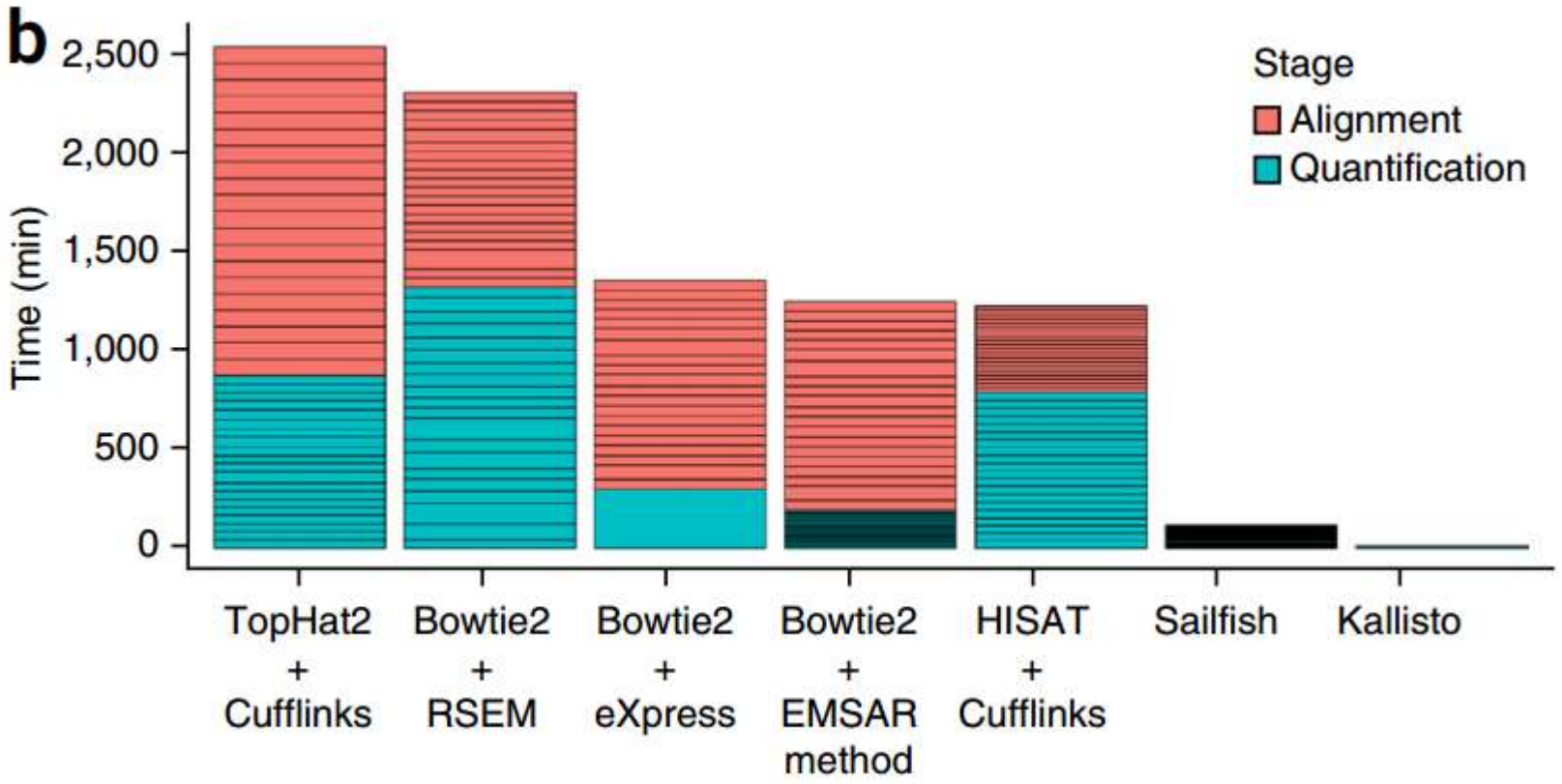




metody alignment-free



metody alignment-free



RNA-seq różnicowa ekspresja

Założenia:

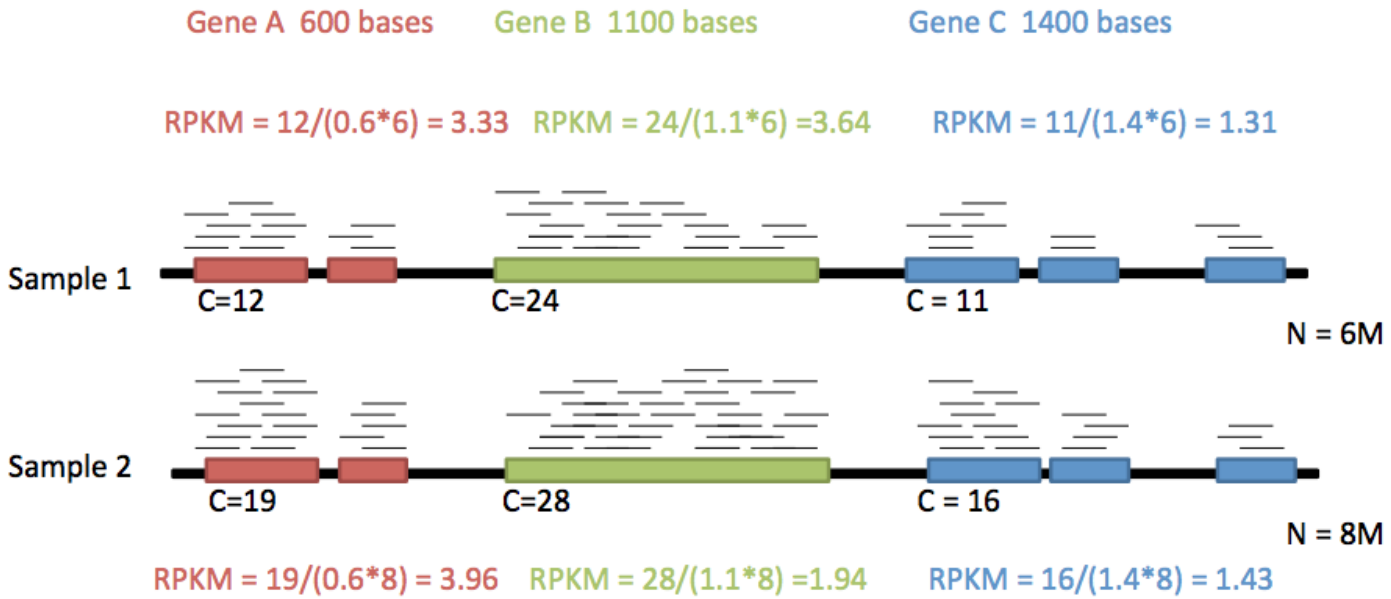
1. Większość genów ulega ekspresji na tym samym poziomie
2. Wariacje w obrębie replikatów technicznych nie występują

Zaburzenia pomiędzy próbkami:

- zróżnicowana ilość odczytów
- zróżnicowana długość genów
- zróżnicowana kompozycja odczytów

normalizacja poziomów ekspresji

RPKM: Reads Per Kilobase of transcript per Milion of mapped reads

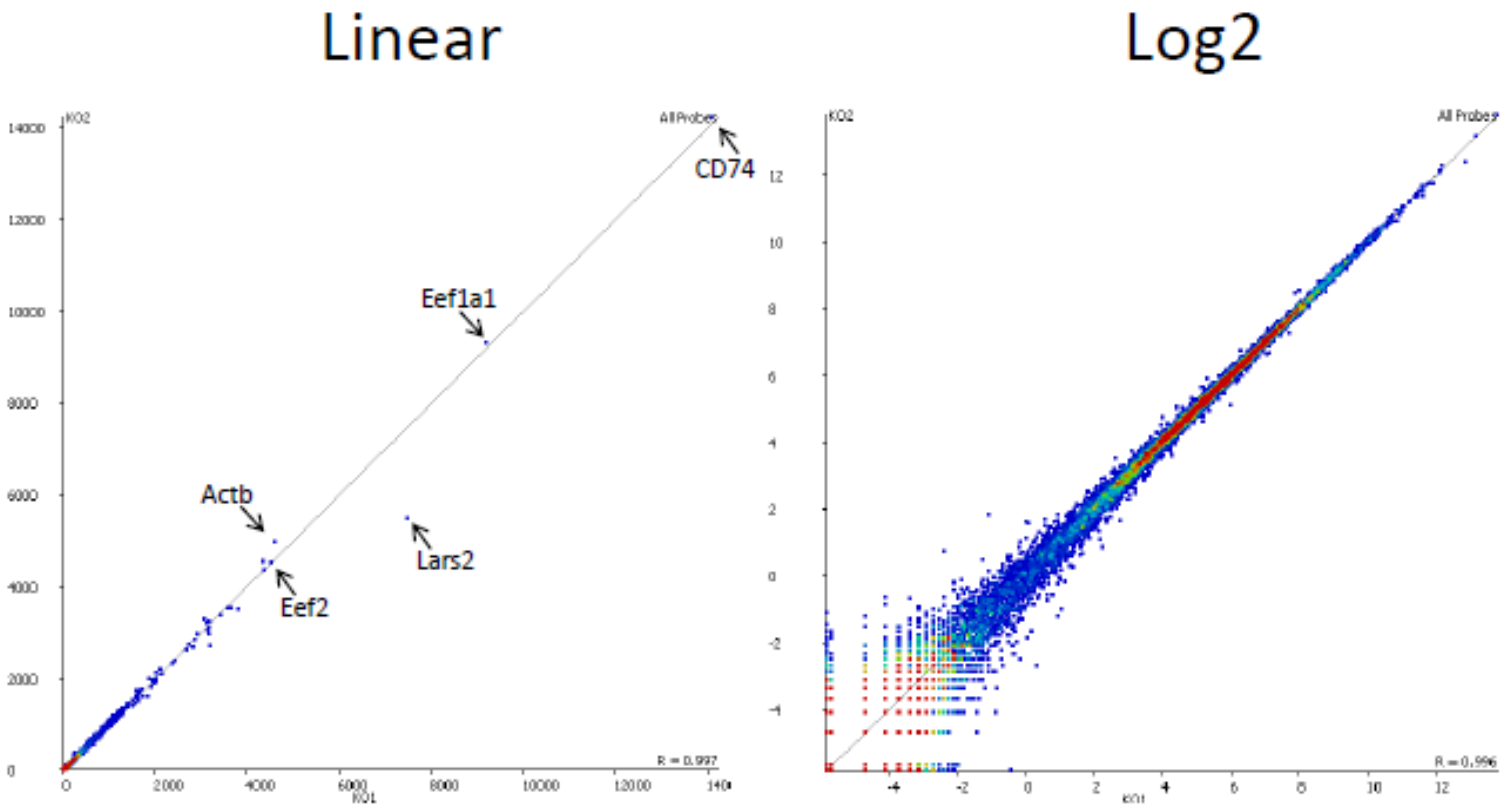


Małe RNA (brak fragmentacji): RPM: Reads Per Milion of mapped reads

wyznaczanie różnic ekspresji

$$\log_2 FC = \log_2(RPKM1/RPKM2)$$

wyznaczanie różnic ekspresji

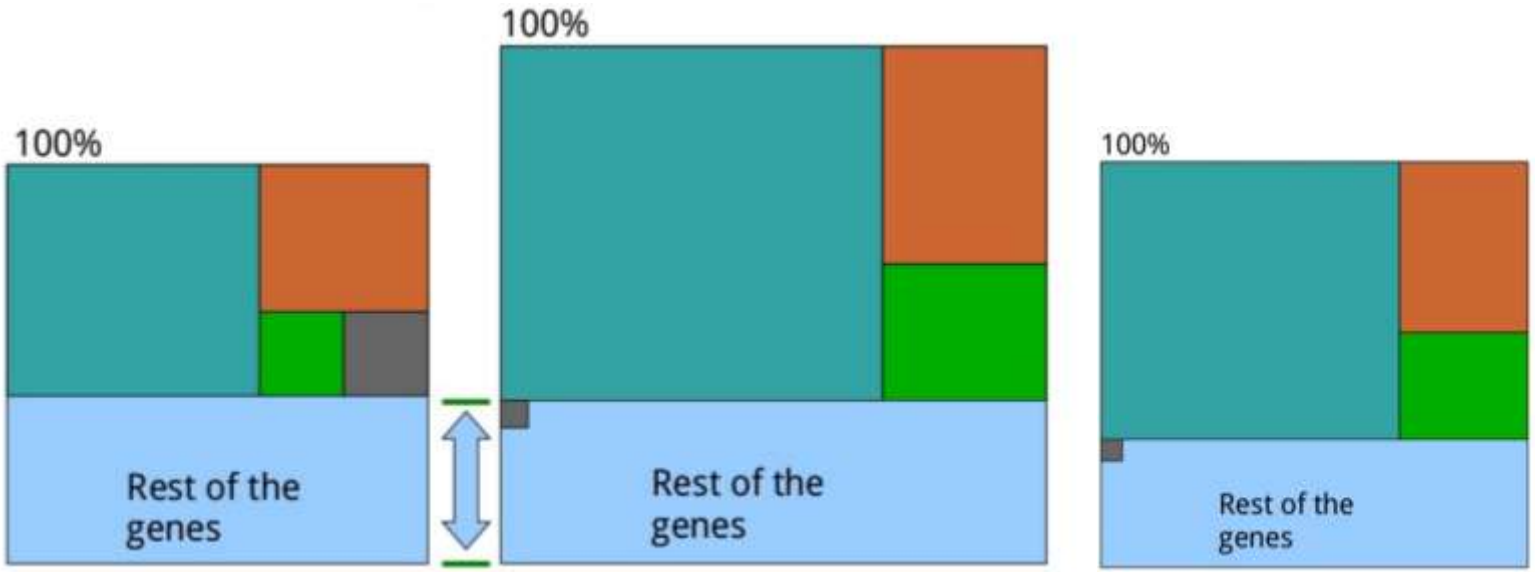


normalizacja poziomów ekspresji

RPKM:

- Bardzo często wykorzystywane
- Zrozumiałe
- Ale problematyczne:
    1. Zmiany w genach ulegających wysokiej ekspresji powodują zmiany pozostałych genów
    2. Odczyty wielokrotnie mapujące zaburzają zliczenia
    3. Ta sama wartość zmiany RPKM pochodzi od:
        - genu o niskim poziomie i dużej wariacji ekspresji
        - genu o wysokim poziomie i małej wariacji eksprsesji

normalizacja poziomów ekspresji

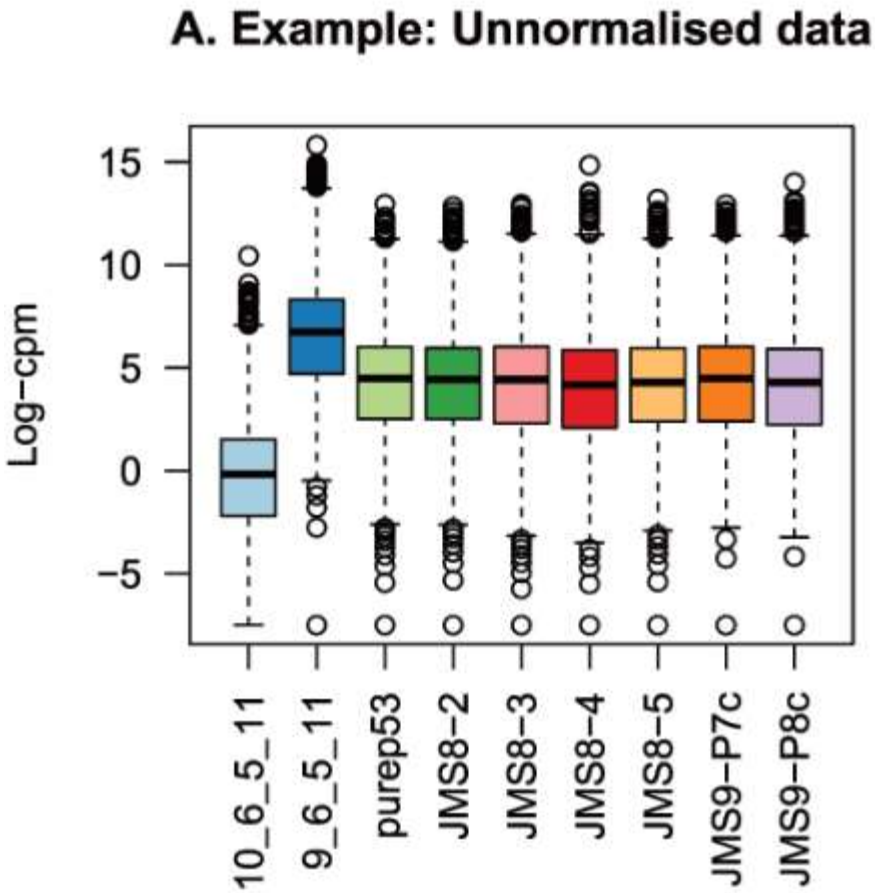


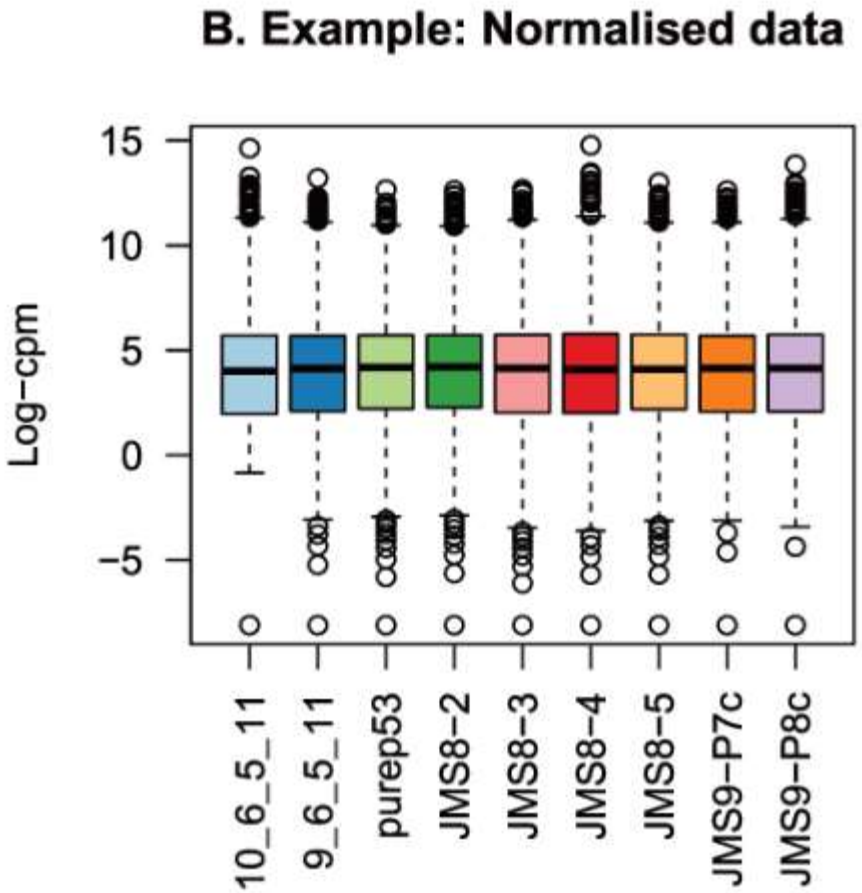
Jakob J. http://slideshare.net 2014

Metody normalizacji

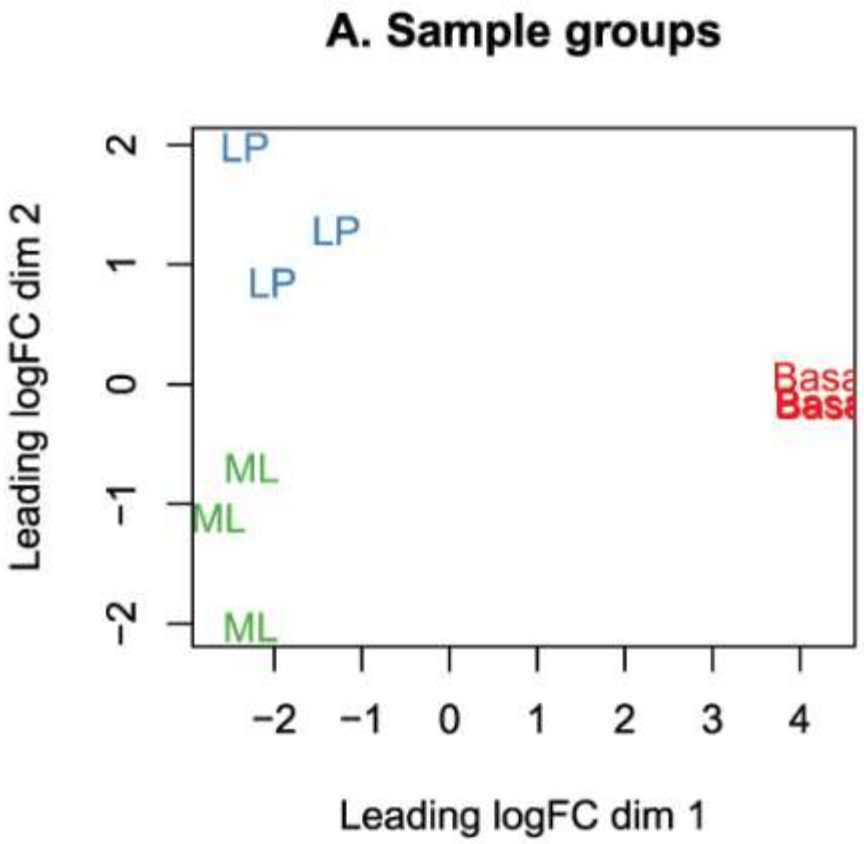
- R/FPKM
  - korekcja różnic w głębokości sekwencjonowania i długości transkryptu
  - cel: porównanie poziomów pomiędzy próbkami i w obrębie próbki
- TMM
  - korekcja różnic w kompozycji próbek spowodowanych pomiarami odstającymi
  - cel: polepszenie precyzji porównań pomiędzy próbkami
- TPM
  - korekcja różnic spowodowanych nieznanymi długościami transkryptów
  - cel: polepszenie precyzji porównań pomiędzy próbkami
- Limma voom (logCPM)
  - stabilizacja wariancji i uniezależnienie jej od średniego poziomu ekspresji

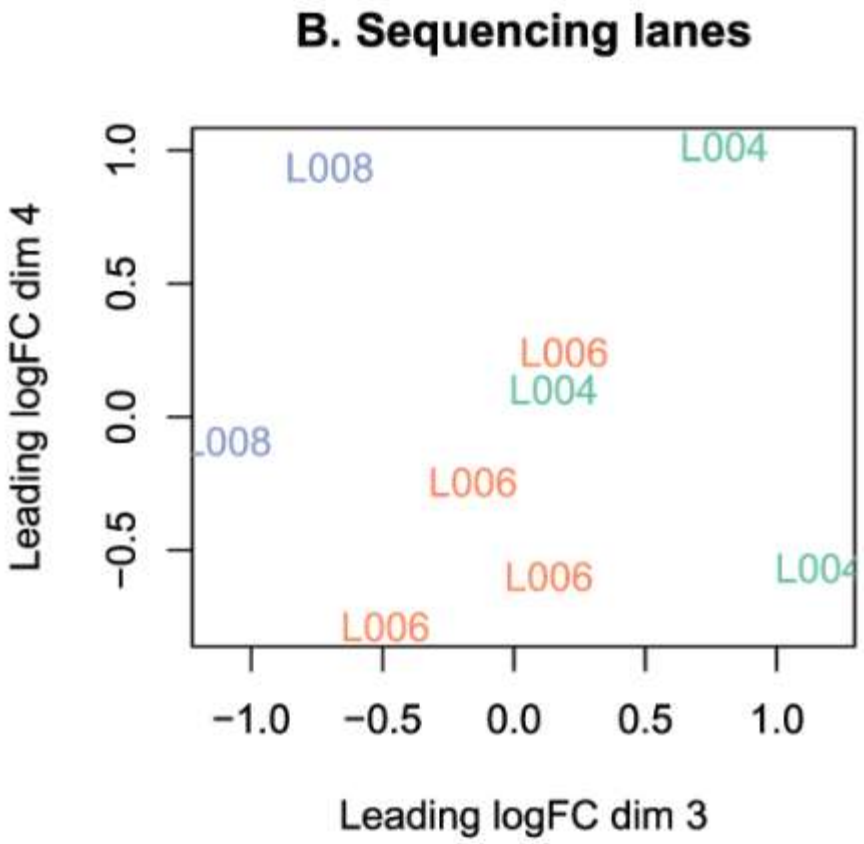
efekt normalizacji



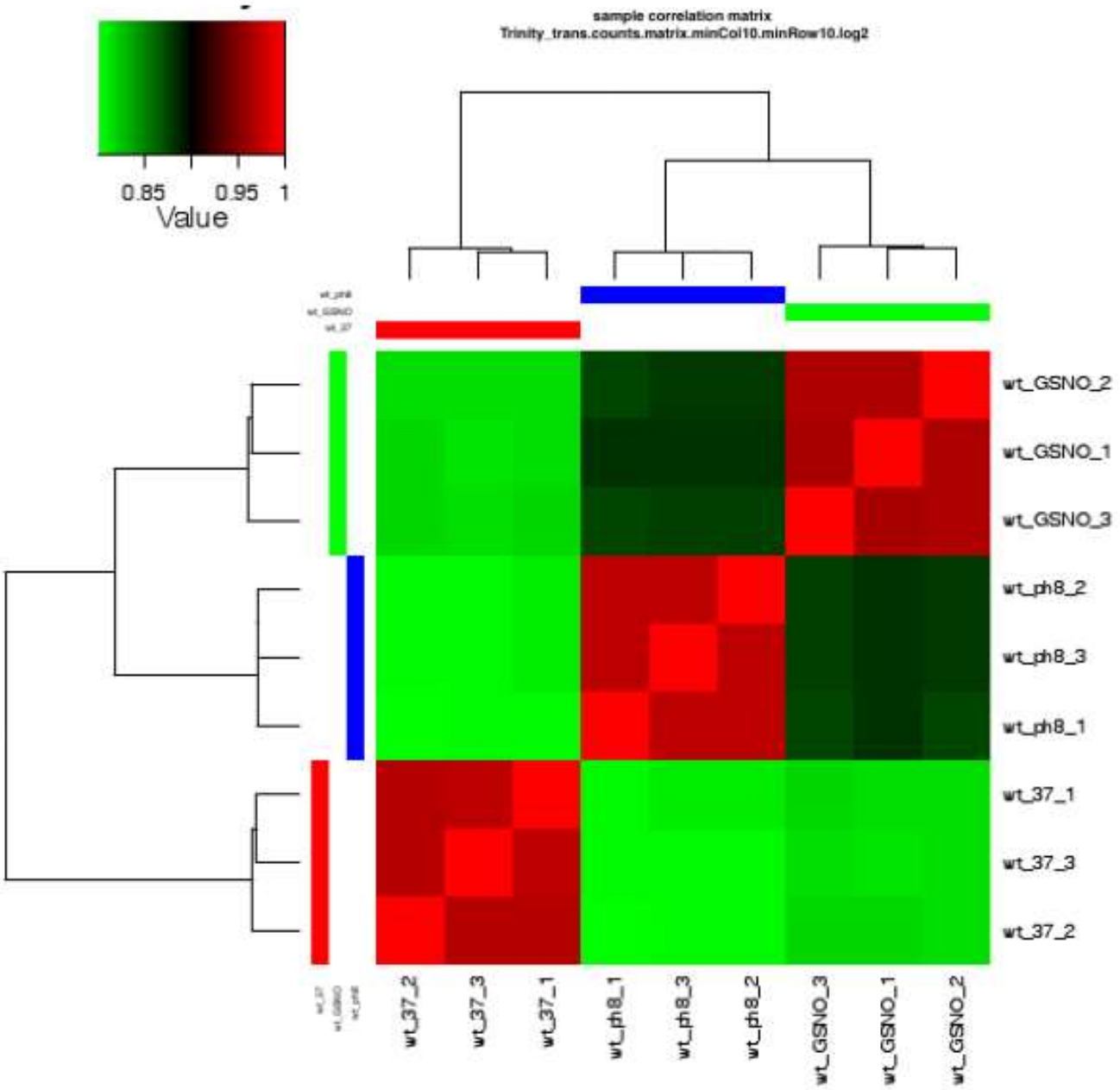


efekt normalizacji

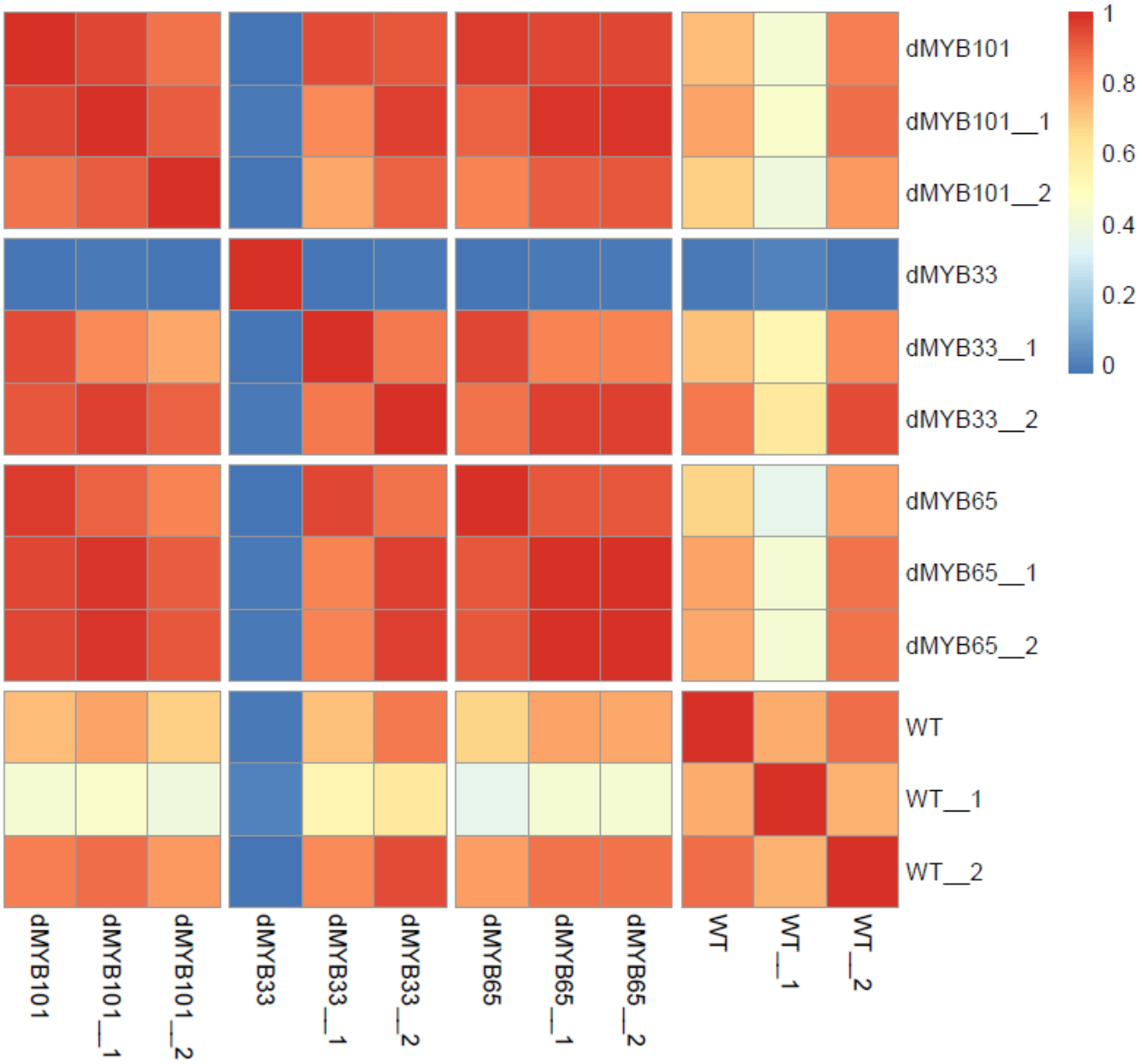




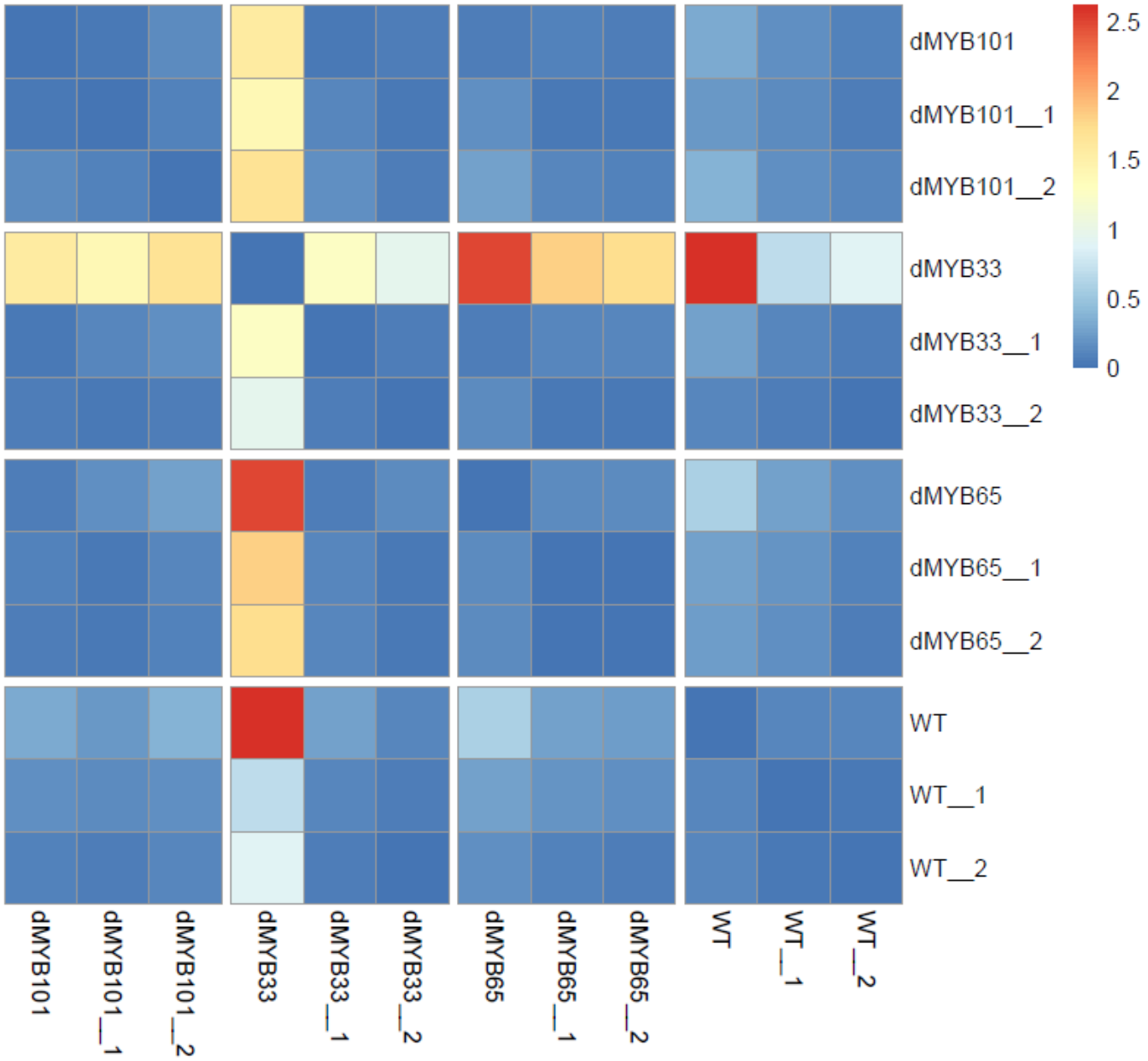
replikaty


sample correlation matrix
Trinity_trans.counts.matrix.minCol10.minRow10.log2
0.85
0.95
1
Value
wt_GSNO_2
wt_GSNO_1
wt_GSNO_3
wt_ph8_2
wt_ph8_3
wt_ph8_1
wt_37_1
wt_37_3
wt_37_2
wt_37_2
wt_37_3
wt_37_1
wt_ph8_1
wt_ph8_3
wt_ph8_2
wt_GSNO_3
wt_GSNO_1
wt_GSNO_2

korelacja próbek - Spearman



korelacja próbek – dystans Poisson



RNA-seq testowanie DE

Metody parametryczne:

- edgeR (negatywna dystrybucja dwumianowa)
- Deseq (negatywna dystrybucja dwumianowa)
- DEGseq (zakłada normalną dystrybucję dla MA)
- baySeq (klasyfikator Bayesa oparty o dystrybujcję NB lub Poisson)

Metody nieparametryczne:

- Test Fisher'a
- Cufdiff
- NOIseq (nie potrzebuje replikatów!)

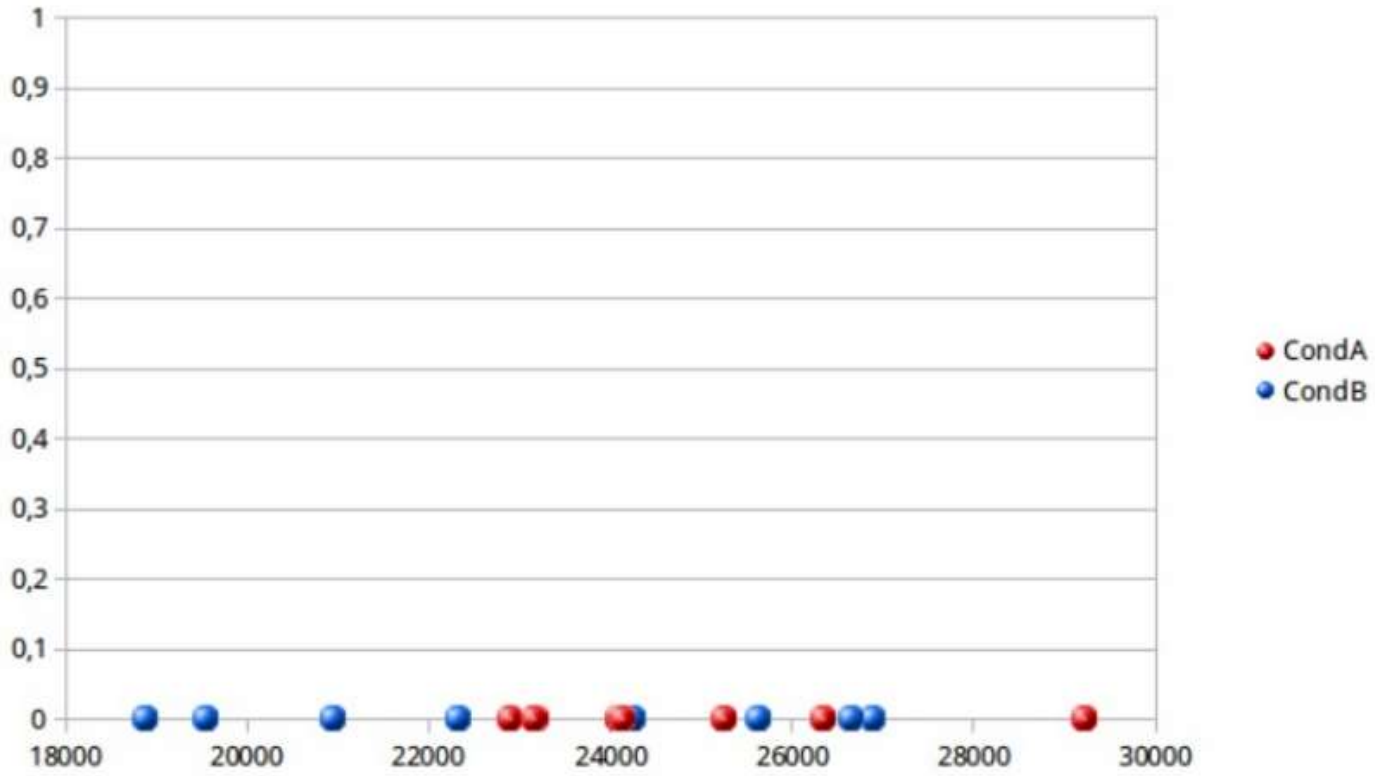
RNA-seq testowanie DE

Parametry:

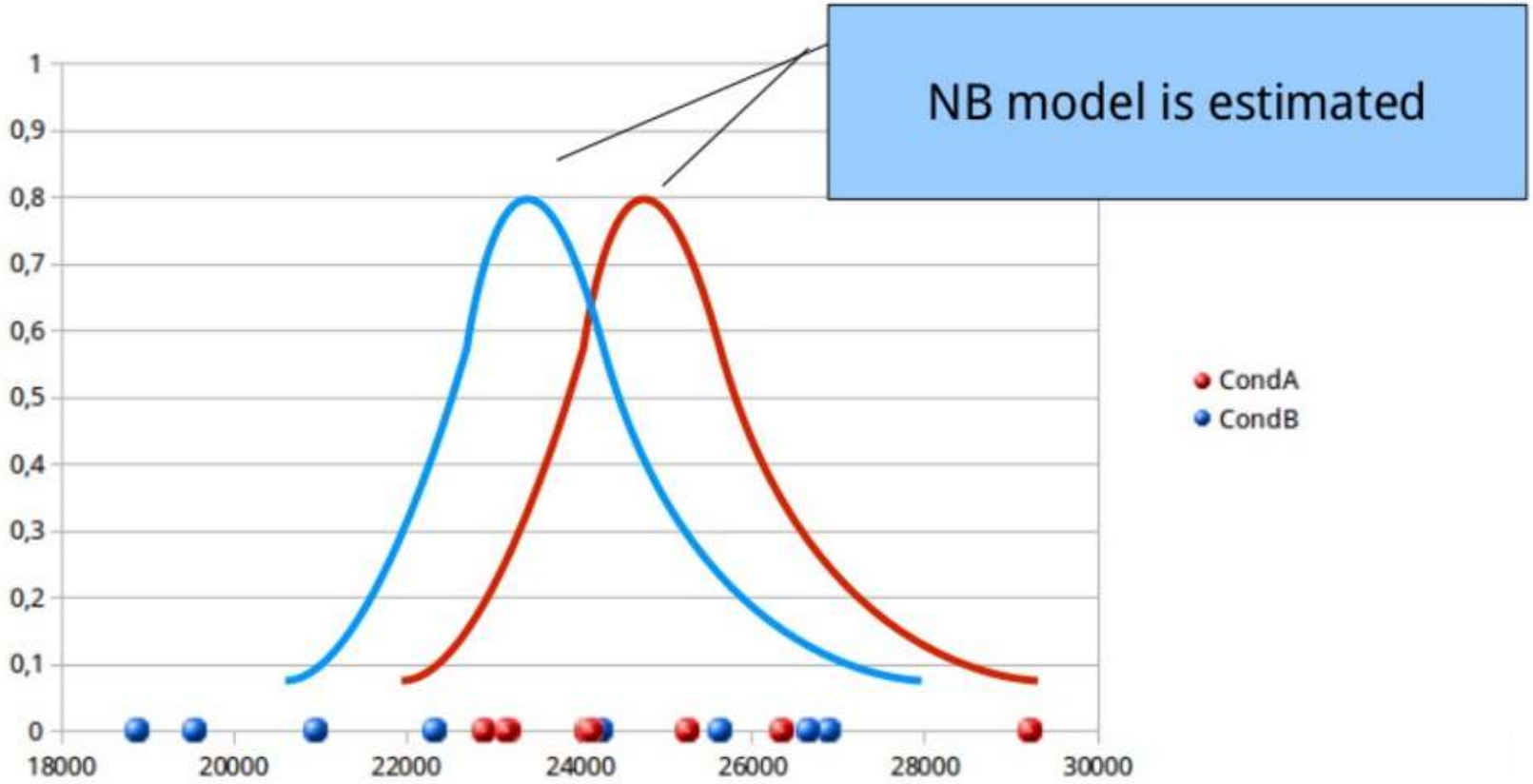
1. Współczynniki wielkości biblioteki
   - Obliczane na podstawie obserwowanych różnic w medianie ekspresji
   - Brak transformacji odczytów

2. Dyspersja w obrębie replikatów
   - Liczona dla każdego genu, bądź w binach i uśredniana
   - zależna od poziomu ekspresji

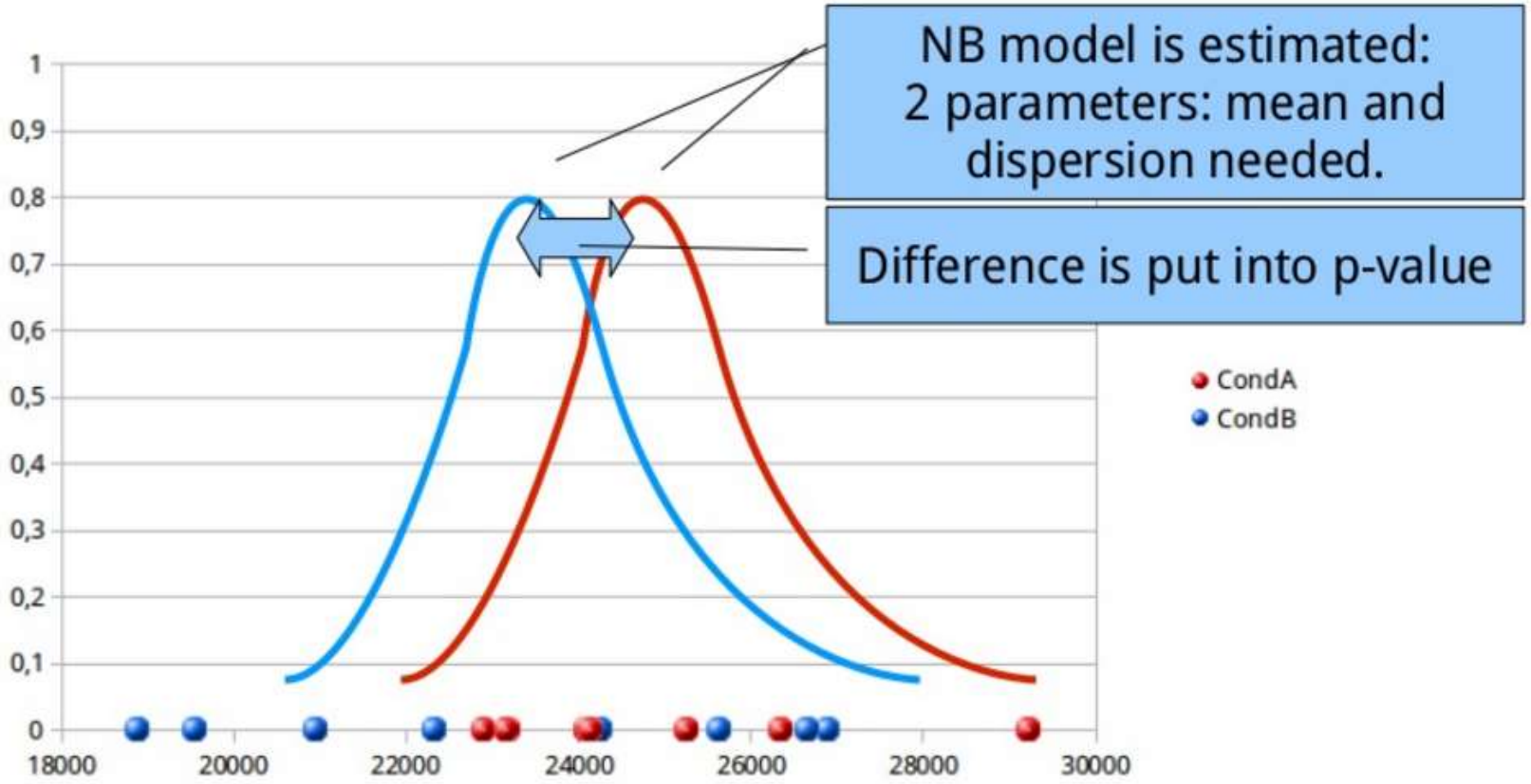
RNA-seq testowanie DE



RNA-seq testowanie DE



RNA-seq testowanie DE



RNA-seq testowanie DE