

Dyskusja biologa ze statystykiem w towarzystwie R - czyli jak znaleźć przydatne informacje w bezmiarze danych biologicznych

A. Szabelska-Beręsewicz, J. Zyprych-Walczak

Katedra Metod Matematycznych i Statystycznych,
Uniwersytet Przyrodniczy w Poznaniu

RBioMeSs - R and Bio-informatics/Medical statistics,
Warszawa 2016

O nas słów kilka...



Projekt 1

Projekt 2



Projekt 3



Projekt 4



O nas słów kilka...

Projekt 1: 'Ostra białaczka szpikowa'

Projekt 2: 'Normalizacja danych RNA-seq'

Projekt 3: 'Wielbłady...'

Projekt 4: 'Oporność na cisplatynę i paklitaksel w komórkach ludzkiego raka jajnika'



Nasz zespół



prof. dr hab. Idzi Siatkowski

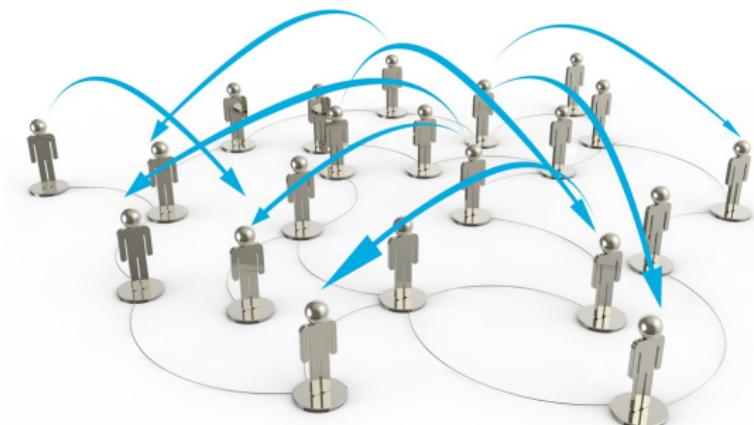


dr Alicja Szabelska-Beręsewicz



dr Joanna Zyprych-Walczak

3+? = nieskończanie wiele możliwości



Źródło: <https://kamiljurowski.files.wordpress.com/>

Współpraca



CENAT ICHB PAN Poznan
dr Luiza Handschuh



ETH Zurich
dr Michał Okoniewski

Współpraca - c.d.



Wojskowy Uniwersytet
Medyczny
dr n. med. Jolanta Szenajch

BOKU Wiedeń
dr Paweł Łabaj

Nasze projekty z przeszłości



Vol. 58, No 4/2011
573–580
on-line at: www.actabp.pl

Regular paper

Impact of DNA microarray data transformation on gene expression analysis — comparison of two normalization methods

Marcin T. Schmidt^{1,2,3}, Luiza Handschuh^{2,4}, Joanna Zyprych⁴, Alicja Szabelska⁴, Agnieszka K. Olejnik-Schmidt¹, Ildzi Siatkowska⁴ and Marek Figlerowicz^{2,3}

¹Department of Biotechnology and Food Microbiology, Poznań University of Life Sciences, Poznań, Poland; ²Institute of Bioorganic Chemistry PAS, Poznań, Poland; ³Poznań University of Medical Sciences, Department of Hematology, Poznań, Poland; ⁴Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Poznań, Poland; Institute of Computing Science, Poznań University of Technology, Poznań, Poland

CENAT ICHB PAN Poznan
dr Luiza Handschuh

Nasze projekty z przeszłości



INTERNATIONAL JOURNAL OF MOLECULAR MEDICINE 32: 688-694 (2013)

Analysis of boutique arrays: A universal method for the selection of the optimal data normalization procedure

BARBARA USZCZYŃSKA¹, JOANNA ZYPRYCH-WALCZAK², LUIZA HANDSCHUCH^{1,3}, ALICJA SZABELSKA^{2,5},
MACIEJ KAZMIERNICKI², WIESŁAWA WORONOWICZ², PIOTR M. M. SIKORSKI¹, MICHAŁ M. SIKORSKI¹,
KACZEKIA A. FOMAŁOWICZ^{1,3}, EDYCJA SŁONECKA^{1,3}, I. MAREK EG. FRONCZAK^{1,3}

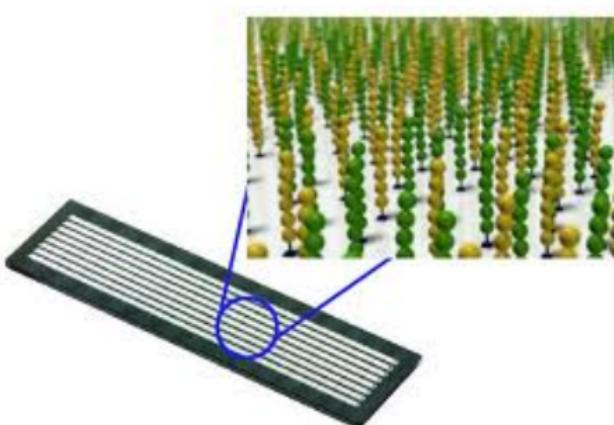
¹Institute of Bioorganic Chemistry, Polish Academy of Sciences, 61-704 Poznań; ²Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, 60-637 Poznań; ³Department of Hematology, Poznań University of Medical Sciences, 60-569 Poznań; ⁴Institute of Computing Science, Poznań University of Technology, 61-138 Poznań, Poland; ⁵Functional Genomics Center Zurich, Swiss Federal Institute of Technology (ETH) Zurich and the University of Zurich, CH-8057 Zurich, Switzerland

Received March 21, 2013; Accepted May 28, 2013

DOI: 10.3892/mmr.2013.144

CENAT ICHB PAN Poznań
dr Luiza Handschuh

Co to jest RNA-Seq?



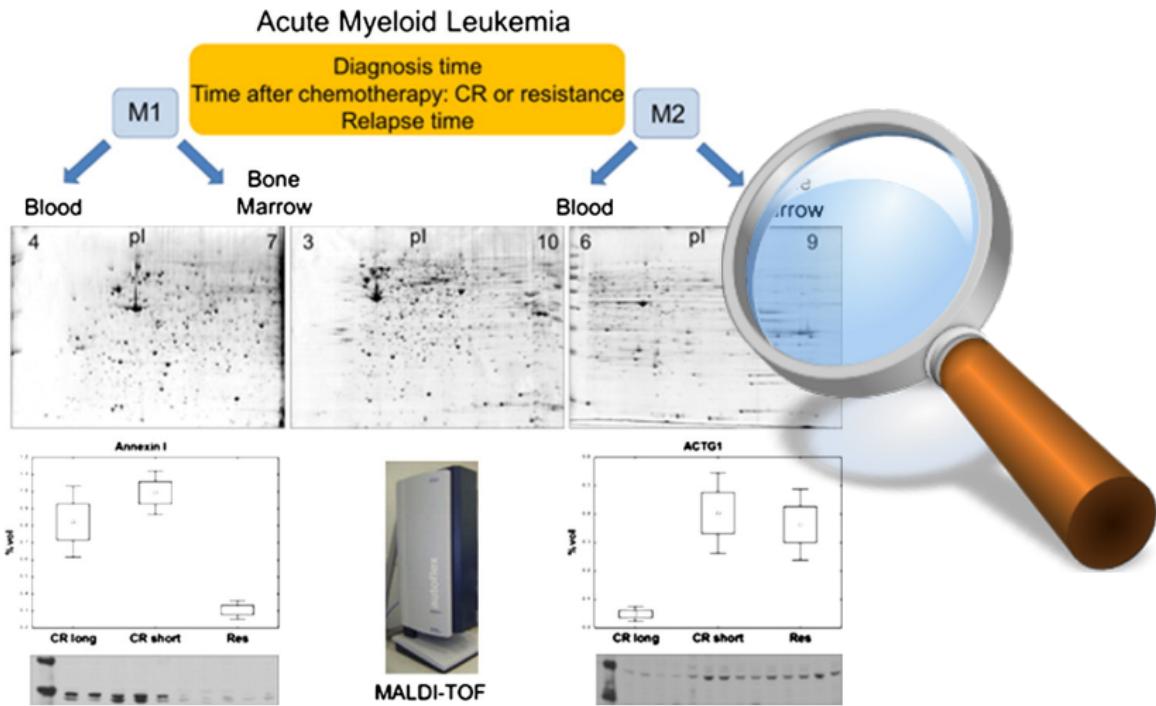
RNA-Seq

Wysoko przepustowa technologia sekwencjonowania, która pozwala na otrzymanie sekwencji badanego materiału genetycznego w celu otrzymania informacji na temat zawartości RNA w próbce.

Przydatność RNA-Seq...

- Identyfikacja ważnych genów związanych z rozważanymi warunkami, np. choroba nowotworowa - analiza różnicowa;
- Odkrywanie nowych genów odpowiedzialnych za rozważany warunek;
- Poszukiwanie mutacji oraz fuzji genów (powiązanych np. z chorobą nowotworową);
- Diagnostyka wielu chorób bakteryjnych i wirusowych;
- Rozwijanie annotacji genomów wielu organizmów.

Projekt 1: 'Ostra białaczka szpikowa - AML'



Rzut oka na dane...

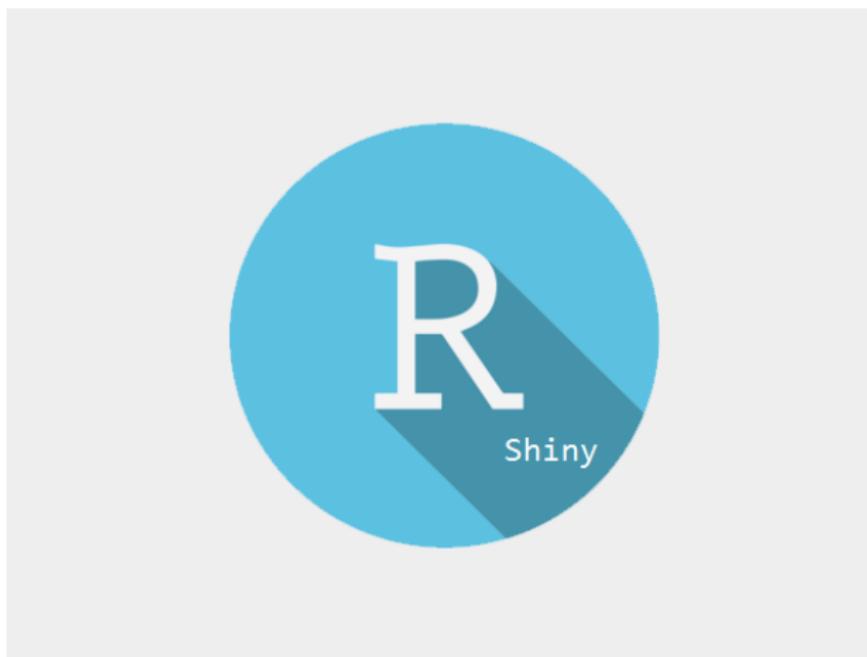
Eksperyment RNA-seq

- 30 próbek krwi (PB) oraz szpiku kostnego (BM)
- 28 pacjentów chorych na ostrą białaczkę szpikową - AML (acute myeloid leukaemia) leczonych w Katedrze i Klinice Hematologii i Transplantacji Szpiku przy Uniwersytecie Medycznym w Poznaniu
- kontrola: próbka szpiku kostnego oraz próbka, która była mieszanką 12 próbek krwi zdrowych ochotników
- odczyty były mapowane korzystając z referencyjnego genomu człowieka UCSC hg19 oraz z wykorzystaniem programu TopHat run (v2.0.6).



Dyskusja biologa ze statystykiem w towarzystwie R

...czyli...





Projekt 2: 'Normalizacja danych RNA-seq'

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 621690, 10 pages
<http://dx.doi.org/10.1155/2015/621690>

Research Article

The Impact of Normalization Methods on RNA-Seq Data Analysis

J. Zyprych-Walczak,¹ A. Szabelska,¹ L. Handschuh,^{2,3} K. Górczak,¹ K. Klamecka,¹
M. Figlerowicz,² and I. Siatkowski¹

¹Department of Mathematical and Statistical Methods, Poznan University of Life Sciences, 60-637 Poznan, Poland

²Institute of Bioorganic Chemistry, Polish Academy of Sciences, 61-704 Poznan, Poland

³Department of Hematology and Bone Marrow Transplantation, Poznan University of Medical Sciences, 60-569 Poznan, Poland

Correspondence should be addressed to J. Zyprych-Walczak; joanna.zyprych@gmail.com

Received 20 March 2015; Revised 17 May 2015; Accepted 18 May 2015

Academic Editor: Ernesto Picardi

Copyright © 2015 J. Zyprych-Walczak et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Problem - mnogość normalizacji...

nazwa	R pakiet	metoda
d ^{TMM}	edgeR	$\log_2(d_j^{TMM}) = \frac{\sum_{g=1}^{G'} w_{gj} M_{gj}}{\sum_{g=1}^{G'} w_{gj}}, M_{gj} = \frac{\log_2(k_{gj}/N_j)}{\log_2(k_{gr}/N_r)}, w_{gj} = \frac{N_j - k_{gj}}{N_j k_{gj}} + \frac{N_r - k_{gr}}{N_r k_{gr}}$
d ^{UQ}	edgeR	$d_j^{UQ} = \frac{\sum_{g \in G} k_{gj}}{UQ_j},$
d ^{DES}	DESeq	$d_j^{DES} = \text{median}_g \frac{k_{gj}}{\left(\prod_{v=1}^m k_{gv}\right)^{1/m}}$
d ^{EBS}	EBSeq	$d_j^{EBS} = 10^{\log Q_j - \frac{1}{m} \sum_{v=1}^m \log Q_v}$
d ^{PS}	PoissonSeq	$d_j^{PS} = \frac{\sum_{g \in G''} k_{gj}}{\sum_{g \in G''} \sum_{j=1}^m k_{gj}},$

m - liczba prób;

k_{gj}, k_{gr} - odczyty dla genu g w j-tej próbie lub w próbie referencyjnej r;

N_j, N_r - ogólna liczba odczytów dla próby j i referencyjnej próby r;

UQ_j - górny kwantyl liczby odczytów różnych od zera w j-tej próbie;

G' - zbiór genów wykluczając 5% najbardziej obfitych genów;

G'' - zbiór genów dla których $GOF_g \in (\epsilon, 1 - \epsilon)$,

gdzie $GOF_g = \sum_{j=1}^m \frac{(k_{gj} - d_j^{tc} k_{g.})^2}{d_j^{tc} k_{g.}}$, $\epsilon \in (0, \frac{1}{2})$,

$d_j^{tc} = \frac{\sum_{g=1}^G k_{gj}}{\sum_{g=1}^G \sum_{j=1}^m k_{gj}}.$

Parę słów o danych

Dane	# prób	# genów	# gaf	mAAG	MAAG	# HG
Cheung	41	52580	12 409	0,024	90180	124
Bodymap	16	52580	13 131	0	934100	131
AML	27	16642	12 749	50,04	482 400	127

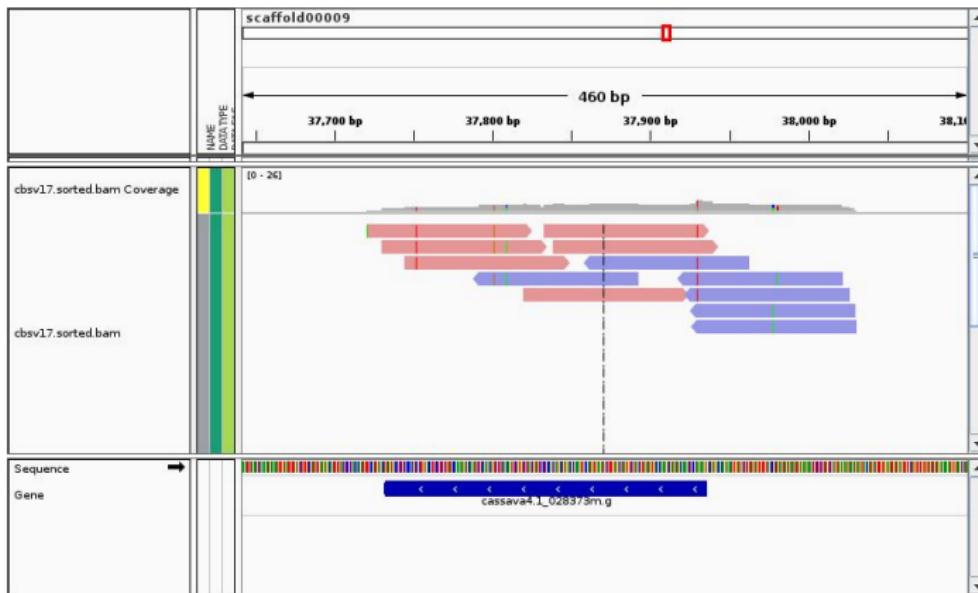
gaf - geny po filtrowaniu

mAAG - minimalna liczba odczytów w genie

MAAG - maksymalna liczba odczytów



Jak wyglądają 'surowe' dane?





Jak wyznaczać tzw. 'read counts'?





Jak wyglądają dane używane w analizach statystycznych?

county

62,722 observations of 14 variables

	row.names	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
1	ENSG00000230021	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	ENSG00000223659	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	ENSG00000225972	11	9	15	17	1	4	2	2	7	4	12	6	7	12
4	ENSG00000225630	1453	1406	1336	1825	1630	1421	1677	1334	2302	1762	1841	1488	2052	2890
5	ENSG00000276171	0	1	0	0	0	0	0	0	2	1	0	0	0	0
6	ENSG00000237973	722	655	497	776	609	447	775	721	1430	1027	970	873	1239	1432
7	ENSG00000278791	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	ENSG00000229344	94	66	51	94	37	25	35	22	110	56	67	43	71	124
9	ENSG00000240489	5	5	1	5	6	2	3	4	15	8	4	4	8	9
10	ENSG00000248527	6348	4782	4146	6672	6495	4122	5990	4735	11877	8163	9281	6440	9671	12583
11	ENSG00000198744	29	22	24	19	22	13	14	11	44	25	24	20	40	47
12	ENSG00000268663	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	ENSG00000273547	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	ENSG00000229376	0	3	0	0	0	0	0	0	0	2	0	1	2	0
15	ENSG00000224956	0	0	0	0	0	1	0	0	1	2	0	0	0	1
16	ENSG00000235373	1	5	3	3	2	3	1	3	5	4	4	4	2	6
... 61,722 omitted ...															

Displayed 1000 rows of 62,722 (61,722 omitted)

Metody w R do analizy różnicowej

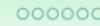
Przegląd metod

- edgeR
- DESeq
- SAMseq
- NOISeq
- EBSeq
- baySeq

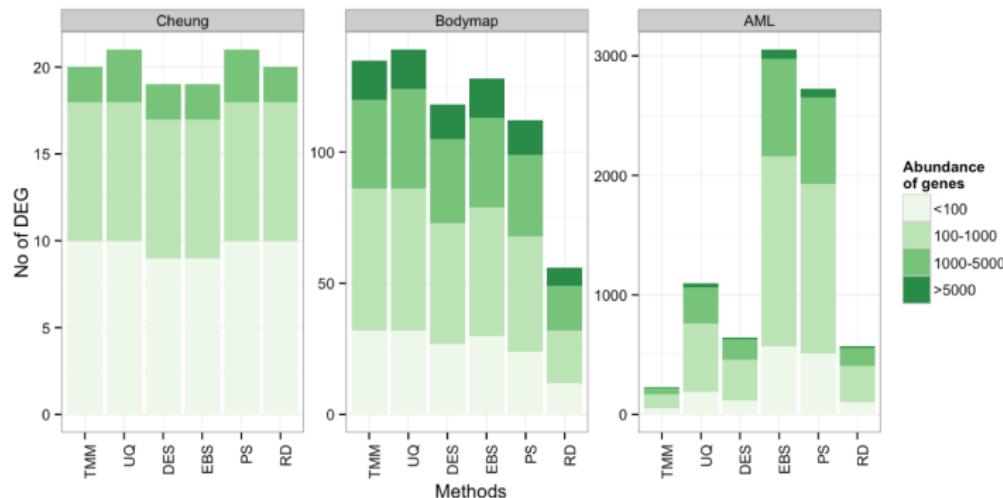
Metody w R do analizy różnicowej

Przegląd metod

- `edgeR`
- `DESeq`
- `SAMseq`
- `NOISeq`
- `EBSeq`
- `baySeq`



Struktura genów różnicujących (DEG)



Analiza dyskryminacyjna

- Weryfikacja przynależności prób do określonych klas bazująca na genach różnicujących;
- 5 klasyfikatorów;
- Walidacja - LOOCV.

Klasyfikatory użyte w analizach:

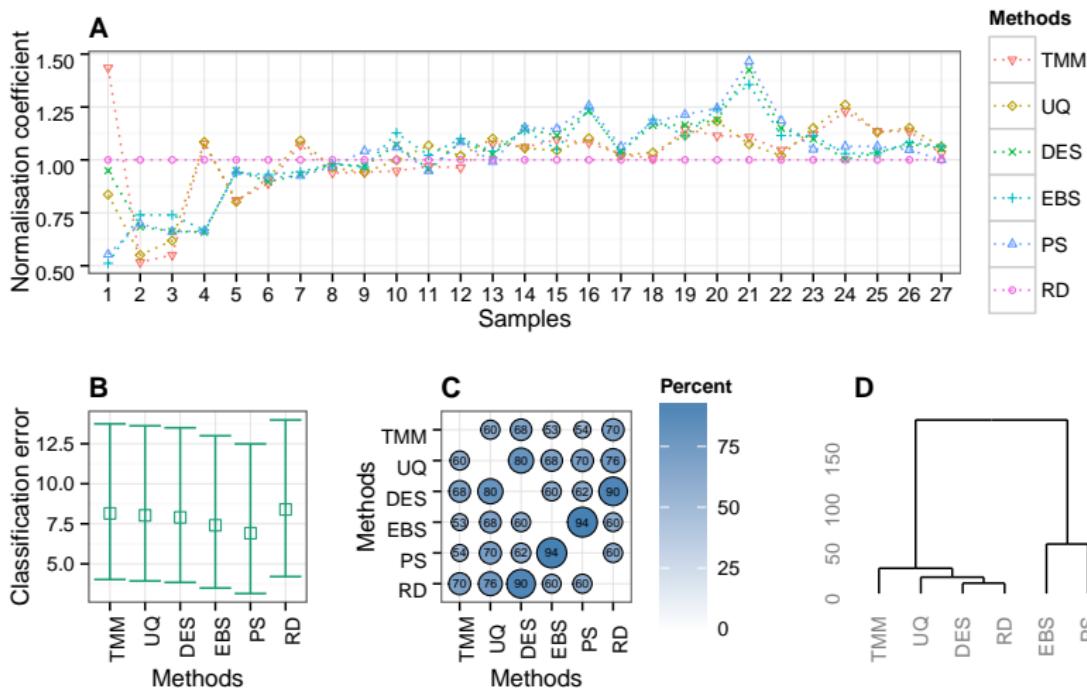
Metoda klasyfikacji	funkcja w R	pakiet
naiwny Bayesowski	NaiveBayesI	MLInterfaces
sieci neuronowe	nnetI	MLInterfaces
k - najbliższych sąsiadów	knnI	MLInterfaces
wektorów wspierających	svmI	MLInterfaces
lasy losowe	randomForestI	MLInterfaces

Podsumowanie wyników - dane AML

kryterium	ranga dla metody				
	TMM	UQ	DES	EBS	PS
ψ	5	4	1	3	2
ϕ	5	4	1	3	2
czułość	5	3	4	1	2
specyficzność	1.5	3	1.5	5	4
błędы predykcji	5	4	3	2	1



Wykresy diagnostyczne - dane AML



R w akcji...





Projekt 3: 'Wielbłądy...'

Journal List > Nucleic Acids Res > v.40(9); May 2012 > PMC3351146

Nucleic Acids Research

Nucleic Acids Res. 2012 May; 40(9): e63.

Published online 2011 December 29. doi: [10.1093/nar/gkr1249](https://doi.org/10.1093/nar/gkr1249)

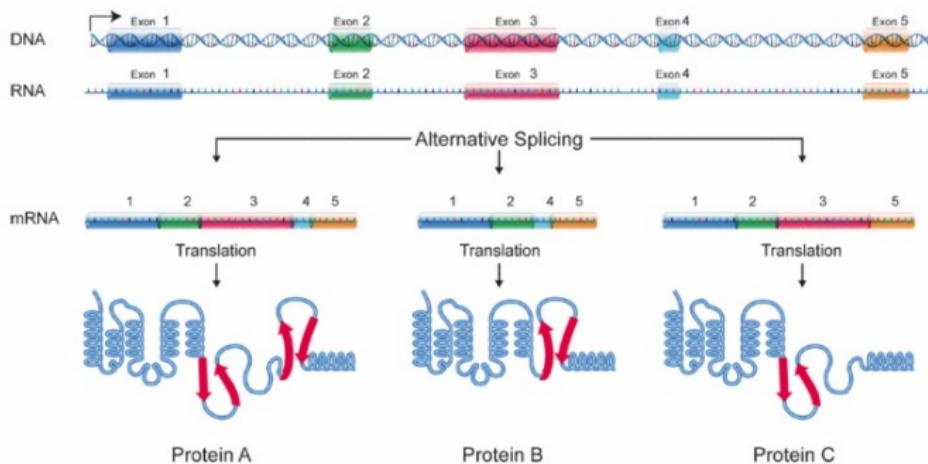
PMCID: PMC3351146

Preferred analysis methods for single genomic regions in RNA sequencing revealed by processing the shape of coverage

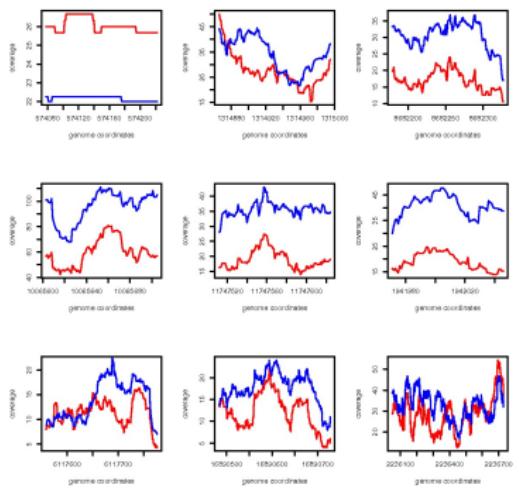
Michał J. Okoniewski,^{1,*} Anna Leśniewska,^{1,2} Alicja Szabelska,³ Joanna Zyprych-Walczak,³ Martin Ryan,¹ Wachtel,⁴ Tadeusz Morzy,² Beat Schäfer,⁴ and Ralph Schlapbach¹

[Author information](#) ► [Article notes](#) ► [Copyright and License information](#) ►

Motywacja biologiczna dla wielbładów



Wprowadzenie do funkcji pokrycia

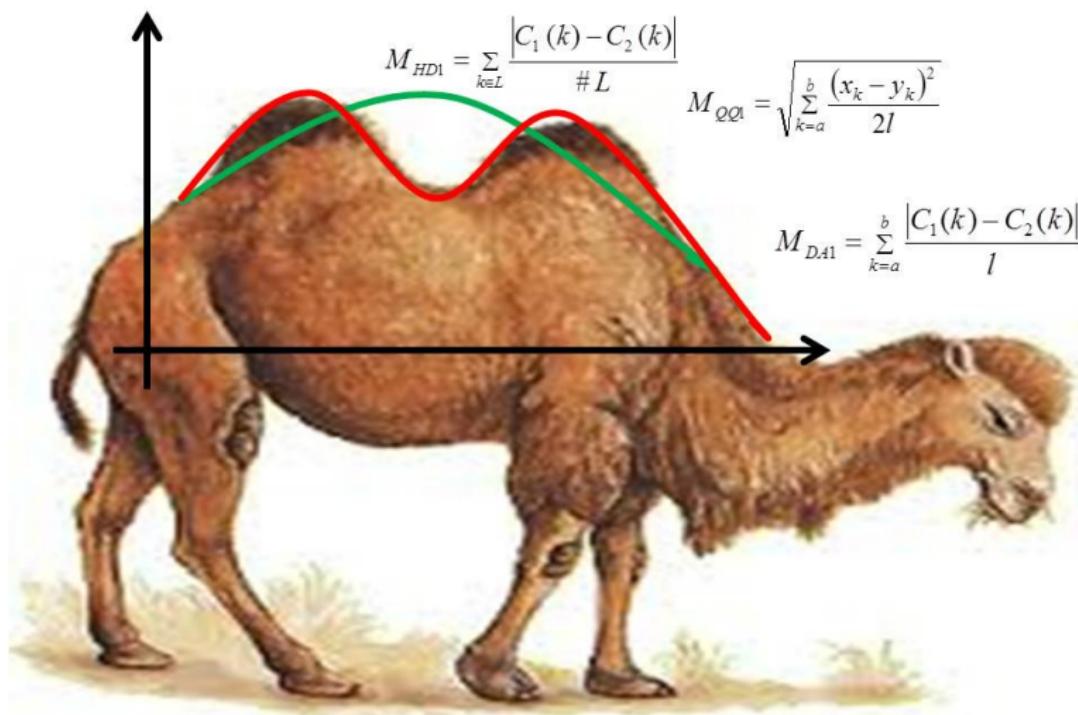


Funkcja pokrycia dla regionu D (Okoniewski et al., 2011)

Dla regionu D, określonego przez parametry $D(\text{chr}, \text{st}, \text{en}, \text{strand})$ o długości $l = \text{en} - \text{st} + 1$, oraz dla zbioru odczytów G_D dopasowanych do regionu D funkcja pokrycia C_D jest zdefiniowana dla każdego nukleotydu $k \in D$ jako ilość odczytów, które są dopasowane do tego nukleotydu. To znaczy:

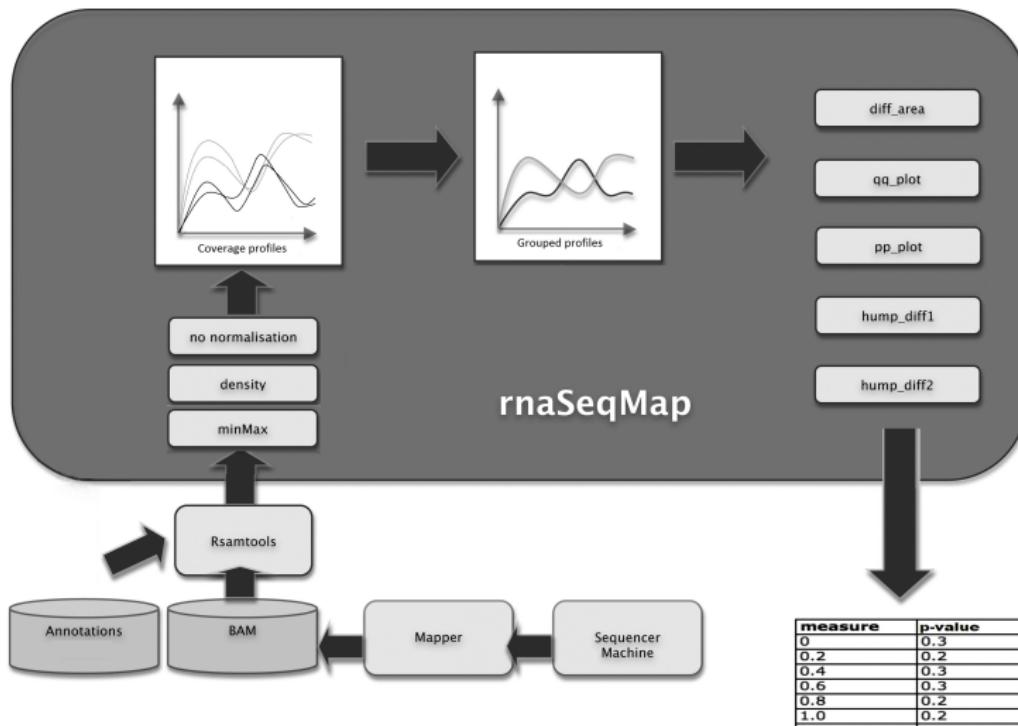
$$C_D(k) = \#\{s \in G_D : k \in s\}, \quad \forall k \in D.$$

Wprowadzenie do funkcji pokrycia..





Procedura testowania krok po kroku





R w akcji...





Projekt 4: 'Oporność na cisplatynę i paklitaksel w komórkach ludzkiego raka jajnika'

ZMIANY W EKSPRESJI SZLAKÓW SYGNAŁOWYCH UCZESTNICZĄCE W WYKSZTAŁCANIU NABYTEJ W ZECNOŚCI ERYTROPOETYNY OPORNOŚCI NA CISPLATYNĘ I **PAKLITAKSEL W KOMÓRKACH LUDZKIEGO RAKA JAJNIKA**

J. Szenajch¹, M. Góralski², A. Świercz², A. Szabelska³, J. Zyprian-Wojciechowski¹, A. Synowiec¹,
I. Siatkowski³, L. Handschuh²

¹ Laboratorium Onkologii Molekularnej, Klinika Onkologii, Wojskowy Instytut Medyczny, Warszawa

² Pracownia Mikromacierzy i Głębokiego Sekwencjonowania oraz Europejskie Centrum Genomiki, Instytut Chemiczny Bioorganicznej PAN, Poznań

³ Katedra Metod Matematycznych i Statystycznych, Uniwersytet Przyrodniczy, Poznań

Uwagi końcowe

- Wizualizacja danych biologicznych jest bardzo istotnym elementem przy współpracy z biologami;
- Potrzeba jest tworzenia grup interdyscyplinarnych, które współpracują nad danym tematem i wzajemnie się wspierają w rozumieniu i interpretacji wyników;
- Spotkania RBioMeSs co kwartał - wymiana doświadczeń grup z różnych miast jako wsparcie dla rozwoju dziedzin: statystyka medyczna i bioinformatyka. Gdzie i kiedy...?

Kontakt

Alicja Szabelska-Beręsewicz: alicja.szabelska@up.poznan.pl

Joanna Zyprych-Walczak: joanna.zyprych@up.poznan.pl

O nas słów kilka...



Projekt 1

Projekt 2



Projekt 3



Projekt 4



Dziękujemy za uwagę!