

Projekt zaliczeniowy nr. 1

Autorzy: Mateusz Kiebała, Sebastian Petryna












Wstępna obróbka danych w celu poprawy ich jakości, wykonana przy użyciu programów Filter FASTQ oraz FASTQ Trimmer.

A1.fastq - przed obróbką	A1.fastq - trimmer 15 od lewej																																
<h2>Summary</h2> <ul style="list-style-type: none">✓ Basic Statistics✓ Per base sequence quality✓ Per tile sequence quality✓ Per sequence quality scores✗ Per base sequence content! Per sequence GC content✓ Per base N content✓ Sequence Length Distribution✗ Sequence Duplication Levels! Overrepresented sequences✓ Adapter Content✗ Kmer Content <h2>✓ Basic Statistics</h2> <table><tr><th>Measure</th><th>Value</th></tr><tr><td>Filename</td><td>A1.fastq</td></tr><tr><td>File type</td><td>Conventional base calls</td></tr><tr><td>Encoding</td><td>Sanger / Illumina 1.9</td></tr><tr><td>Total Sequences</td><td>4830297</td></tr><tr><td>Sequences flagged as poor quality</td><td>0</td></tr><tr><td>Sequence length</td><td>101</td></tr><tr><td>%GC</td><td>42</td></tr></table>	Measure	Value	Filename	A1.fastq	File type	Conventional base calls	Encoding	Sanger / Illumina 1.9	Total Sequences	4830297	Sequences flagged as poor quality	0	Sequence length	101	%GC	42	<h2>Summary</h2> <ul style="list-style-type: none">✓ Basic Statistics✓ Per base sequence quality✓ Per sequence quality scores! Per base sequence content✓ Per base GC content✓ Per sequence GC content✓ Per base N content✓ Sequence Length Distribution✗ Sequence Duplication Levels! Overrepresented sequences! Kmer Content <h2>✓ Basic Statistics</h2> <table><tr><th>Measure</th><th>Value</th></tr><tr><td>Filename</td><td>FASTQ_Trimmer_on_data_17</td></tr><tr><td>File type</td><td>Conventional base calls</td></tr><tr><td>Encoding</td><td>Sanger / Illumina 1.9</td></tr><tr><td>Total Sequences</td><td>4830297</td></tr><tr><td>Filtered Sequences</td><td>0</td></tr><tr><td>Sequence length</td><td>86</td></tr><tr><td>%GC</td><td>42</td></tr></table>	Measure	Value	Filename	FASTQ_Trimmer_on_data_17	File type	Conventional base calls	Encoding	Sanger / Illumina 1.9	Total Sequences	4830297	Filtered Sequences	0	Sequence length	86	%GC	42
Measure	Value																																
Filename	A1.fastq																																
File type	Conventional base calls																																
Encoding	Sanger / Illumina 1.9																																
Total Sequences	4830297																																
Sequences flagged as poor quality	0																																
Sequence length	101																																
%GC	42																																
Measure	Value																																
Filename	FASTQ_Trimmer_on_data_17																																
File type	Conventional base calls																																
Encoding	Sanger / Illumina 1.9																																
Total Sequences	4830297																																
Filtered Sequences	0																																
Sequence length	86																																
%GC	42																																

A2.fastq - przed obróbką

A2.fastq - trimmer 15 od lewej












Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per tile sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)
-  [Kmer Content](#)

Basic Statistics

Measure	Value
Filename	A2.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4830297
Sequences flagged as poor quality	0
Sequence length	101
%GC	42

Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per base GC content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Kmer Content](#)


Basic Statistics

Measure	Value
Filename	FASTQ_Trimmer_on_data_18
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4830297
Filtered Sequences	0
Sequence length	86
%GC	42

B1.fastq - przed obróbką

B1.fastq - quality > 15, trimmer 15 od lewej

Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per base GC content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Kmer Content](#)

Basic Statistics

Measure	Value
Filename	B1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	3765074
Filtered Sequences	0
Sequence length	101
%GC	37

Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per base GC content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Kmer Content](#)





Basic Statistics

Measure	Value
Filename	FASTQ_Trimmer_on_data_14
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1586533
Filtered Sequences	0
Sequence length	86
%GC	41

B2.fastq - przed obróbką

B2.fastq - quality > 15, trimmer 15 od lewej












Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per base GC content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Kmer Content](#)

Basic Statistics

Measure	Value
Filename	B2.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	3765074
Filtered Sequences	0
Sequence length	101
%GC	36

Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per base GC content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Kmer Content](#)












Basic Statistics

Measure	Value
Filename	FASTQ_Trimmer_on_data_20160601_1000000000.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1670083
Filtered Sequences	0
Sequence length	86
%GC	39

C1.fastq - przed obróbką

C1.fastq - quality > 15, trimmer 15 od lewej












Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per base GC content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Kmer Content](#)

Basic Statistics

Measure	Value
Filename	C1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	9037346
Filtered Sequences	0
Sequence length	101
%GC	34

Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per base GC content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Kmer Content](#)


Basic Statistics

Measure	Value
Filename	Filter_FASTQ_on_data_52
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1942679
Filtered Sequences	0
Sequence length	86
%GC	40

C2.fastq - przed obróbką

C2.fastq - quality > 15, trimmer 15 od lewej












Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per base GC content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Kmer Content](#)

Basic Statistics

Measure	Value
Filename	FASTQ_Groomer_on_data_11
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	9037346
Filtered Sequences	0
Sequence length	101
%GC	34

Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per base GC content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Kmer Content](#)

Basic Statistics

Measure	Value
Filename	Filter_FASTQ_on_data_53
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2092187
Filtered Sequences	0
Sequence length	86
%GC	36

input1.fastq - przed obróbką

Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per base GC content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Kmer Content](#)



Basic Statistics

Measure	Value
Filename	input1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4385631
Filtered Sequences	0
Sequence length	101
%GC	39

input1.fastq - quality > 15, trimmer 15 od lewej

Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per base GC content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Kmer Content](#)



Basic Statistics

Measure	Value
Filename	FASTQ_Trimmer_on_data_27
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	564521
Filtered Sequences	0
Sequence length	86
%GC	39

input2.fastq - przed obróbką

Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per base GC content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Kmer Content](#)

Basic Statistics

Measure	Value
Filename	input2.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4385631
Filtered Sequences	0
Sequence length	101
%GC	38

input2.fastq - quality > 15, trimmer 15 od lewej

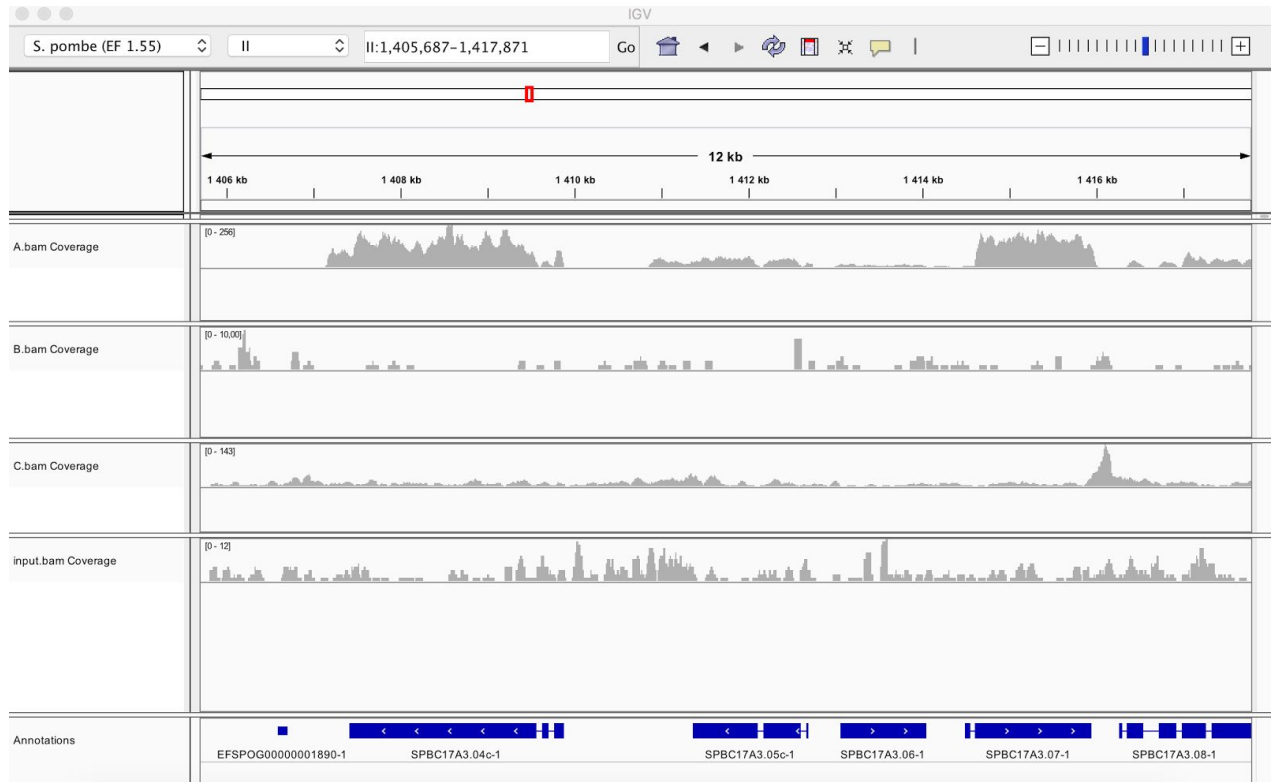
Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per base GC content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Kmer Content](#)

Basic Statistics

Measure	Value
Filename	FASTQ_Trimmer_on_data
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	454296
Filtered Sequences	0
Sequence length	86
%GC	37

Zadanie 1. Odpowiedź: mRNA - próbka A. Wybraliśmy próbkę A, ponieważ jak wynika z obrazu poniżej, jej piki najlepiej pokrywają się z genami.



Zadanie 2. Wiązanie białka - próbki B i C. Wyniki uzyskano za pomocą programu MACS z parametrami:


- effective genome size = 2.70e+09
- tag size = 25
- band width = 300
- no model - ponieważ wszystkie próby obniżenia parametru *fold* się nie udały

a. próbka B

1	2	3	4	5
track name="MACS peaks for MACS_in_Galaxy"				
I	1040	1761	MACS_peak_1	105.10
I	2458	2658	MACS_peak_2	146.76
I	3254	3964	MACS_peak_3	223.62
I	4457	5248	MACS_peak_4	103.90
I	6588	6788	MACS_peak_5	146.76
I	8536	8736	MACS_peak_6	146.76
I	9632	9832	MACS_peak_7	146.76
I	18747	20000	MACS_peak_8	247.31
I	29027	29227	MACS_peak_9	146.76
I	3753433	3755905	MACS_peak_10	1562.69
I	3756410	3756901	MACS_peak_11	186.98
I	3759130	3764251	MACS_peak_12	3100.00
I	3777678	3783950	MACS_peak_13	3100.00
I	3785226	3786487	MACS_peak_14	170.86
I	3787029	3789717	MACS_peak_15	1462.76
I	5570248	5570448	MACS_peak_16	146.76
I	5570977	5572115	MACS_peak_17	135.64
I	5572665	5572865	MACS_peak_18	146.76
I	5573298	5573498	MACS_peak_19	146.76
I	5575175	5575375	MACS_peak_20	146.76

II	1604102	1613990	MACS_peak_21	3100.00
II	1615368	1616275	MACS_peak_22	139.58
II	1616722	1617882	MACS_peak_23	210.83
II	1630825	1631336	MACS_peak_24	107.50
II	1633877	1634495	MACS_peak_25	107.11
II	1635233	1643062	MACS_peak_26	3100.00
II	2113958	2114158	MACS_peak_27	146.76
II	2115539	2116889	MACS_peak_28	283.79
II	2120024	2120224	MACS_peak_29	146.76
II	2135133	2135337	MACS_peak_30	228.00
II	4513710	4514252	MACS_peak_31	107.50
II	4516080	4517736	MACS_peak_32	201.55
II	4518590	4518790	MACS_peak_33	146.76
II	4519444	4519644	MACS_peak_34	146.76
II	4521521	4521721	MACS_peak_35	146.76
II	4522136	4522409	MACS_peak_36	186.98
II	4523249	4523813	MACS_peak_37	107.50
II	4526005	4526205	MACS_peak_38	146.76
II	4527604	4530056	MACS_peak_39	299.73
II	4531077	4531277	MACS_peak_40	146.76
II	4534903	4535103	MACS_peak_41	146.76
III	1071311	1072371	MACS_peak_42	136.87
III	1073032	1090020	MACS_peak_43	3100.00
III	1090718	1091043	MACS_peak_44	311.96
III	1092015	1092215	MACS_peak_45	146.76
III	1108192	1108392	MACS_peak_46	146.76
III	1109188	1138139	MACS_peak_47	3100.00
III	2448143	2448651	MACS_peak_48	107.50

b. próbka C

 1	2	3	4	5
track name="MACS peaks for MACS_in_Galaxy"				
III	2643	3485	MACS_peak_1	418.44
III	4601	5350	MACS_peak_2	256.07
III	5870	6070	MACS_peak_3	205.88
III	10066	10498	MACS_peak_4	151.84
III	11419	11619	MACS_peak_5	205.88
III	13019	16508	MACS_peak_6	520.13
III	23440	23640	MACS_peak_7	205.88
III	989435	989635	MACS_peak_8	205.88
III	1322226	1322426	MACS_peak_9	205.88
III	1592872	1593072	MACS_peak_10	205.88
III	1614170	1614370	MACS_peak_11	205.88
III	2411155	2411355	MACS_peak_12	205.88
III	2439662	2440832	MACS_peak_13	194.28
III	2442954	2443154	MACS_peak_14	205.88
III	2445167	2445367	MACS_peak_15	205.88
III	2447701	2449406	MACS_peak_16	341.41

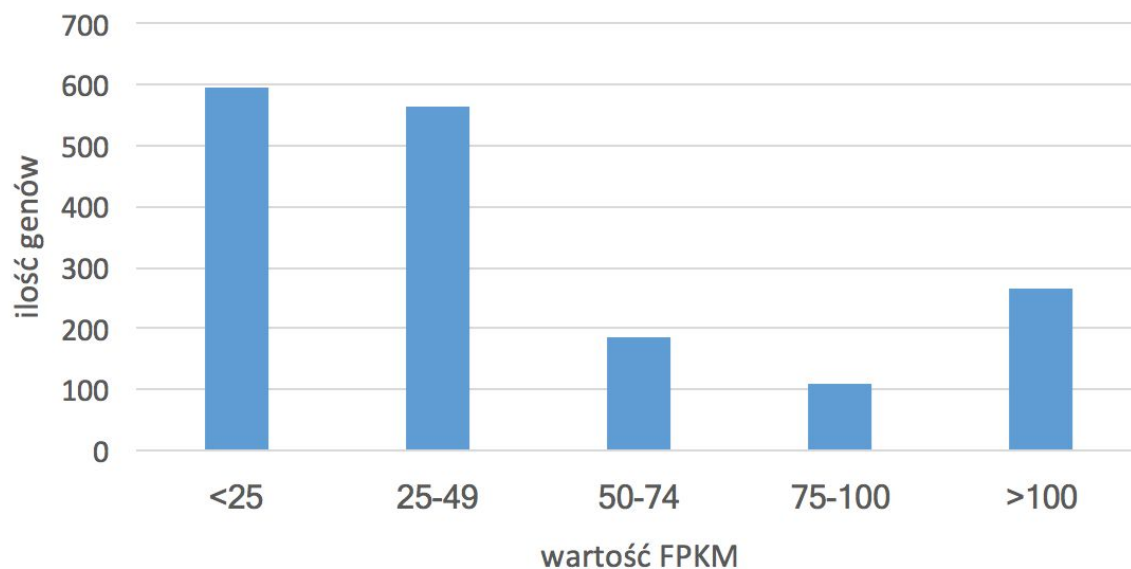
Zadanie 2. ekspresja genów - próbka A. Wyniki uzyskano przy użyciu programu Cufflinks z parametrami:

- max intron length: 300000
- min isoform fraction: 0.1
- pre mRNA fraction: 0.15

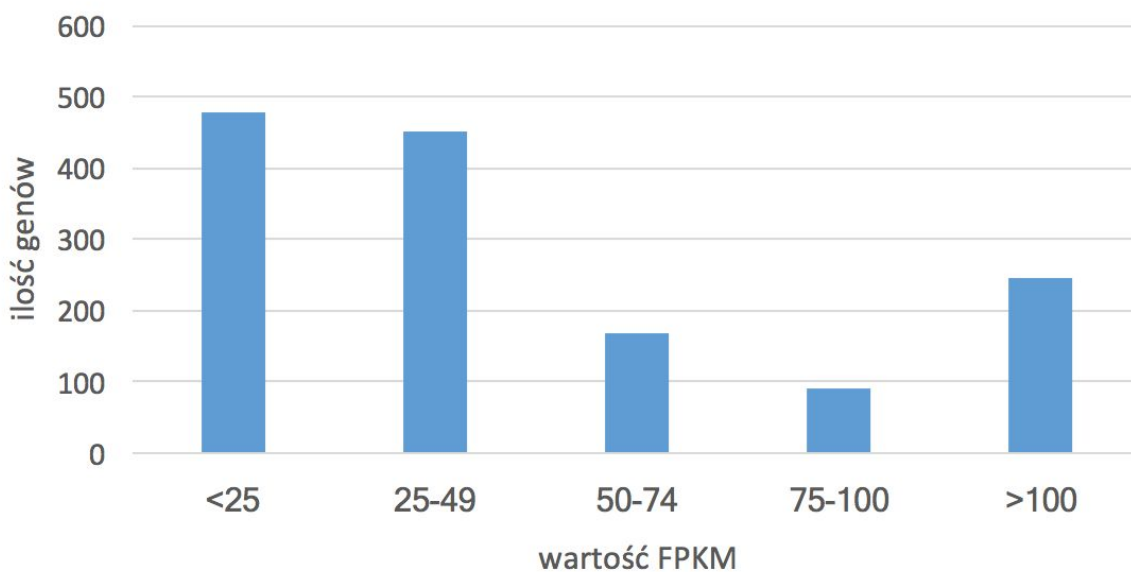
Wycinek wyników:

locus	length	coverage	FPKM
AB325691:3700-4661	-	-	12.3048
AB325691:482-3483	-	-	36.1921
AB325691:14007-15509	-	-	13.8232
AB325691:11284-13643	-	-	45.1893
AB325691:8785-9859	-	-	277.943
AB325691:16082-18204	-	-	16.5527
I:963-2062	-	-	3.8696
I:3710-5329	-	-	3.81843
I:29770-30918	-	-	31.1775
I:21443-22886	-	-	8.55046
I:28669-29628	-	-	8.8945
I:5555702-5557264	-	-	129.561
I:5568716-5569231	-	-	10.5228
I:5565920-5567579	-	-	21.8335
II:5318-6429	-	-	8.11366
II:7526-9340	-	-	30.3281
II:12234-13089	-	-	3.3302
II:21196-21678	-	-	33.7949
II:4513227-4514767	-	-	7.44253
II:4526908-4530423	-	-	4.43193
II:4531943-4532863	-	-	3.63156
III:30590-31669	-	-	7.66946
III:27599-29361	-	-	29.1238
III:32551-34916	-	-	7.39404
III:36950-39291	-	-	25.564
III:2439691-2442803	-	-	901.097
III:2442913-2447551	-	-	5700.61
III:2450602-2452881	-	-	1216.76
MT:0-19431	-	-	4302.92

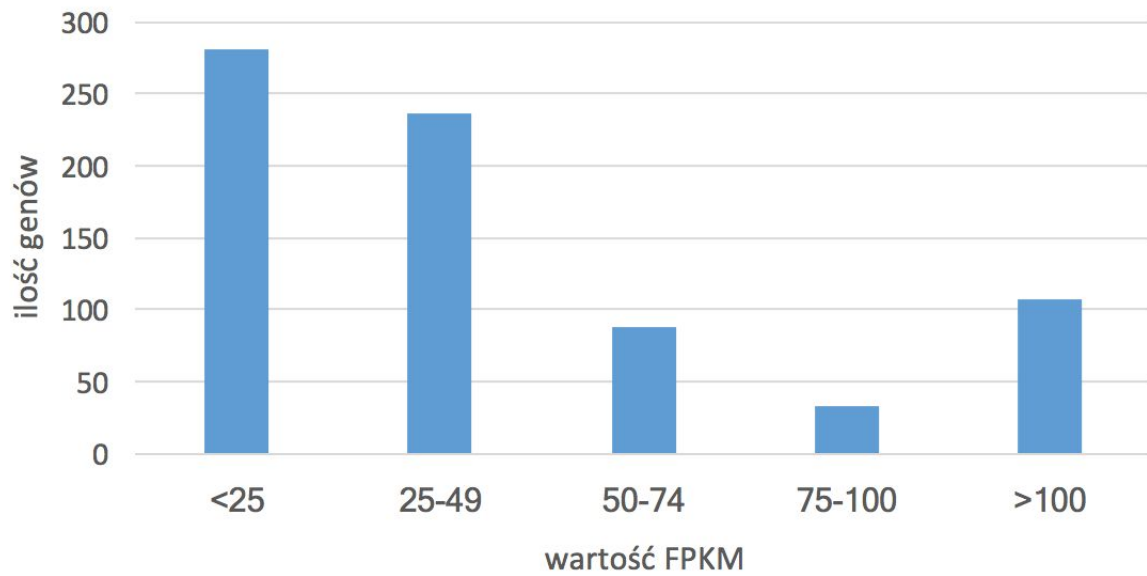
Ekspresja genów - chromosom I



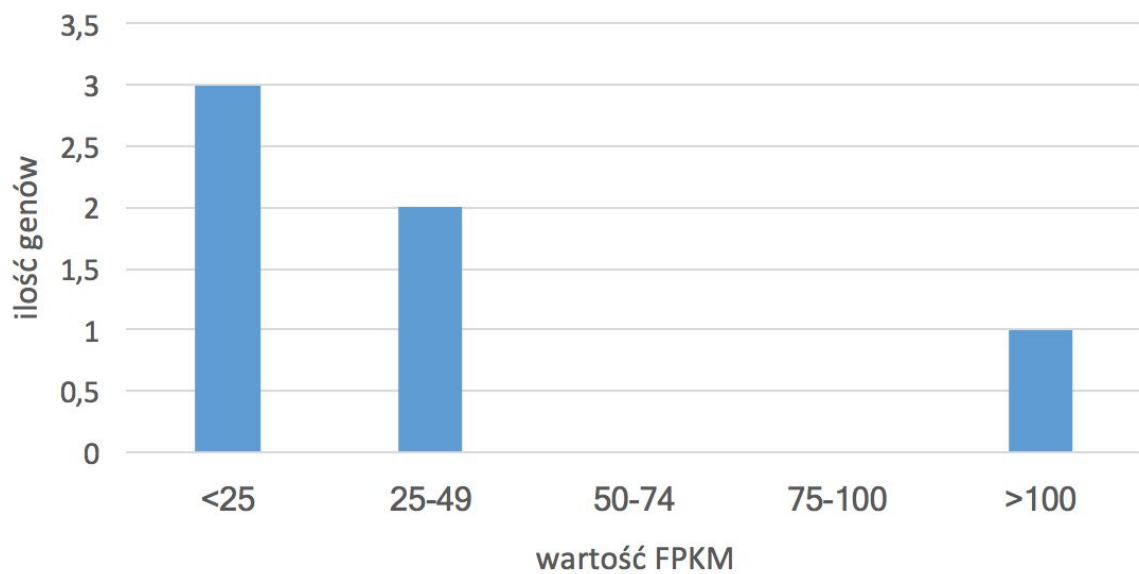
Ekspresja genów - chromosom II



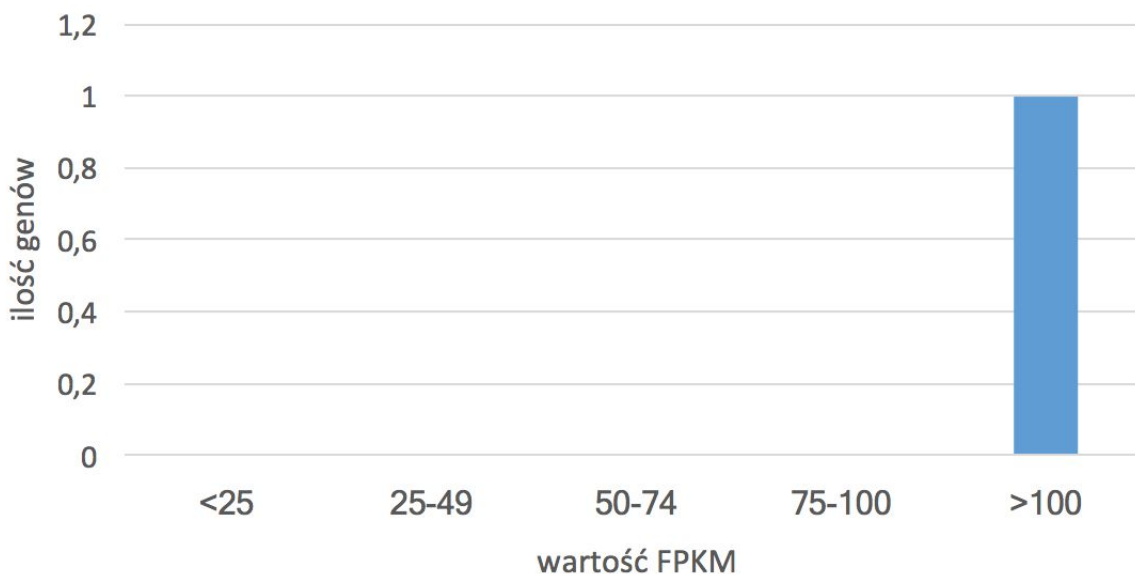
Ekspresja genów - chromosom III



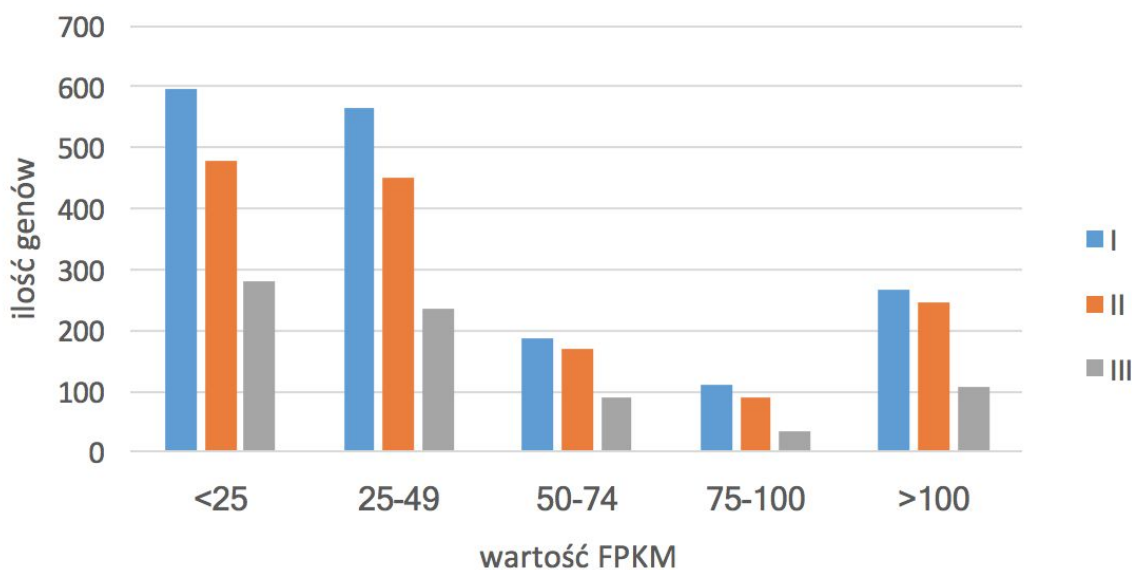
Ekspresja genów - chromosom AB325691



Ekspresja genów - chromosom MT



Porównanie ekspresji genów



Wnioski: Z wykresów widać, że na chromosomach I, II i III większość genów ma wartość FPKM poniżej 50, jednak jest też widoczna część genów z ekspresją powyżej 100 FPKM. Dla chromosomów AB325691 oraz MT mamy za mało genów, aby cokolwiek powiedzieć o zależności poziomów ekspresji.

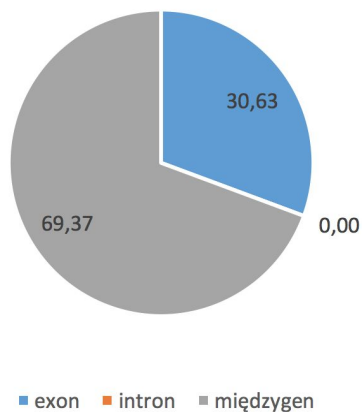
Zadanie 3.

Do charakteryzacji obszarów genomowych napisaliśmy skrypt w języku Python, którym dokonaliśmy charakteryzacji obszarowej (mierzymy pokrycie obszarów pików z obszarami genomu) oraz ilościowej (liczymy ilość przecięć obszarów genomu z obszarami pików).

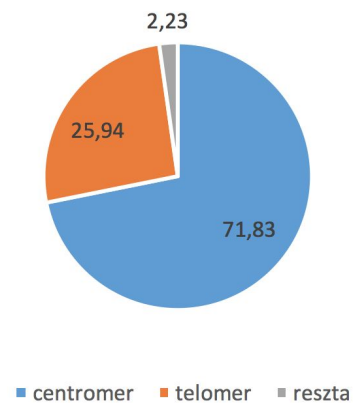
I. Metoda obszarowa

1. Próbką B

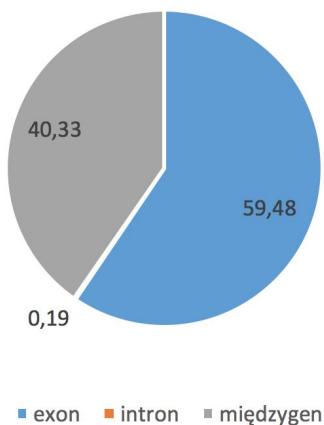
Wiązanie białka - chromosom I



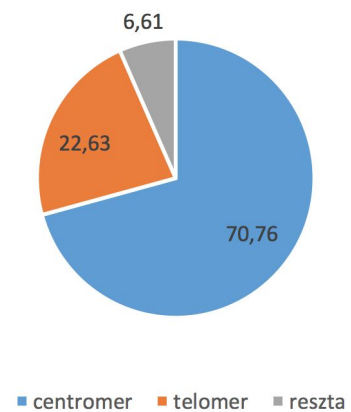
Wiązanie białka - obszar bez genów - chromosom I



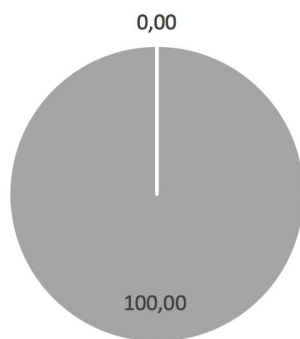
Wiązanie białka - chromosom II



Wiązanie białka - obszar bez genów - chromosom II

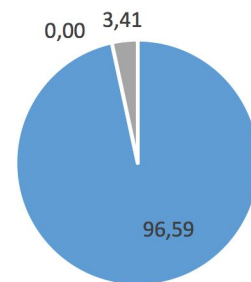


Wiązanie białka - chromosom III



■ exon ■ intron ■ międzygen

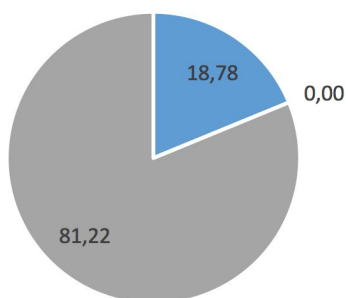
Wiązanie białka - obszar bez genów - chromosom III



■ centromer ■ telomer ■ reszta

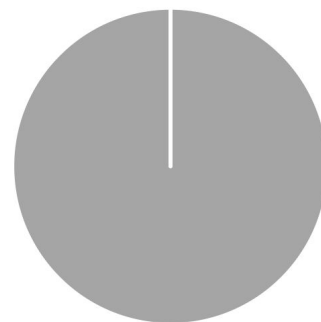
2. próbka C

Wiązanie białka - chromosom III



■ exon ■ intron ■ międzygen

Wiązanie białka - obszar bez genów - chromosom III

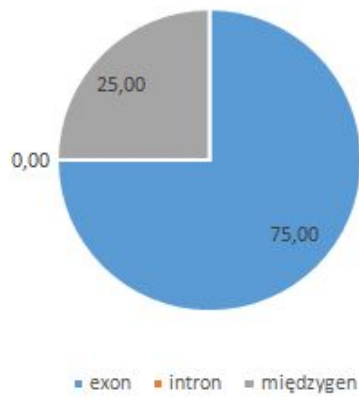


■ centromer ■ telomer ■ reszta

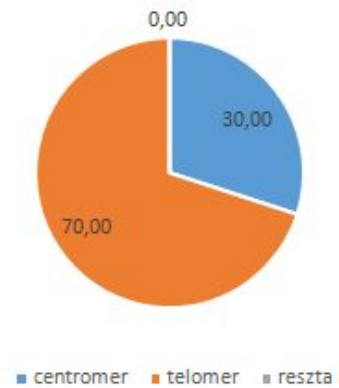
II. Metoda ilościowa

1. Próbką B

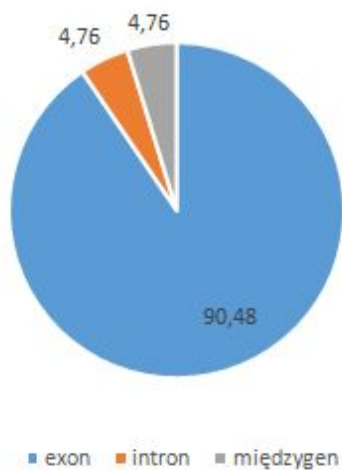
Wiązanie białka - chromosom I



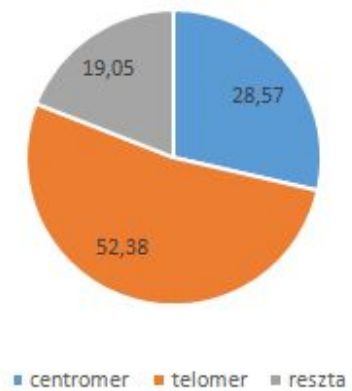
Wiązanie białka - obszar bez genów - chromosom I



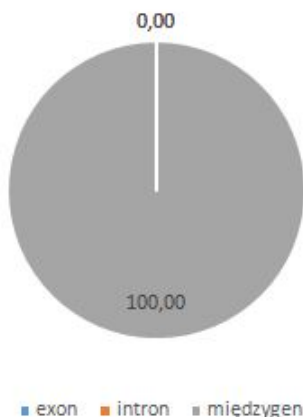
Wiązanie białka - chromosom II



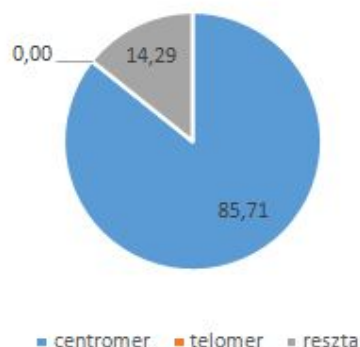
Wiązanie białka - obszar bez genów - chromosom II



Wiązanie białka - chromosom III

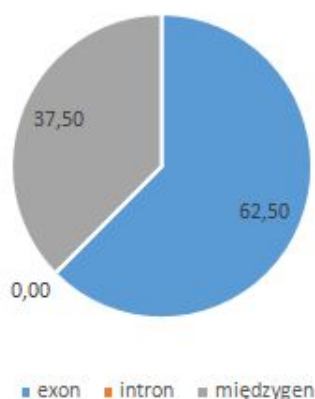


Wiązanie białka - obszar bez genów - chromosom III

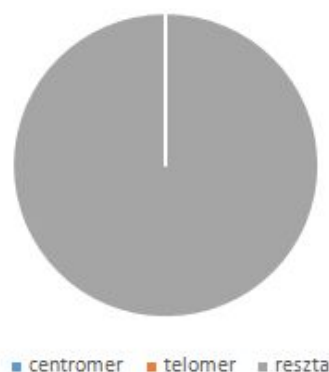


2. Próbką C

Wiązanie białka - chromosom III



Wiązanie białka - obszar bez genów - chromosom III



Wnioski:

Z otrzymanych rezultatów analizy obszarowej wynika, że dla pierwszego i trzeciego chromosomu białko z eksperymentu B wiąże w dużej części na obszarze międzygenowym, zatem być może jest jakoś związane z regulacją procesu transkrypcji, oraz głównie na obszarze centromerowym, co może oznaczać, że odpowiada za strukturę chromosomu. Dla drugiego chromosomu przeważa wiązanie na obszarze eksonów, co wskazuje na to, że wpływa na ekspresję genów. Dla białka z eksperymentu C znalezione zostały piki jedynie na chromosomie III i tam wiążą one przede wszystkim na obszarze międzygenowym.

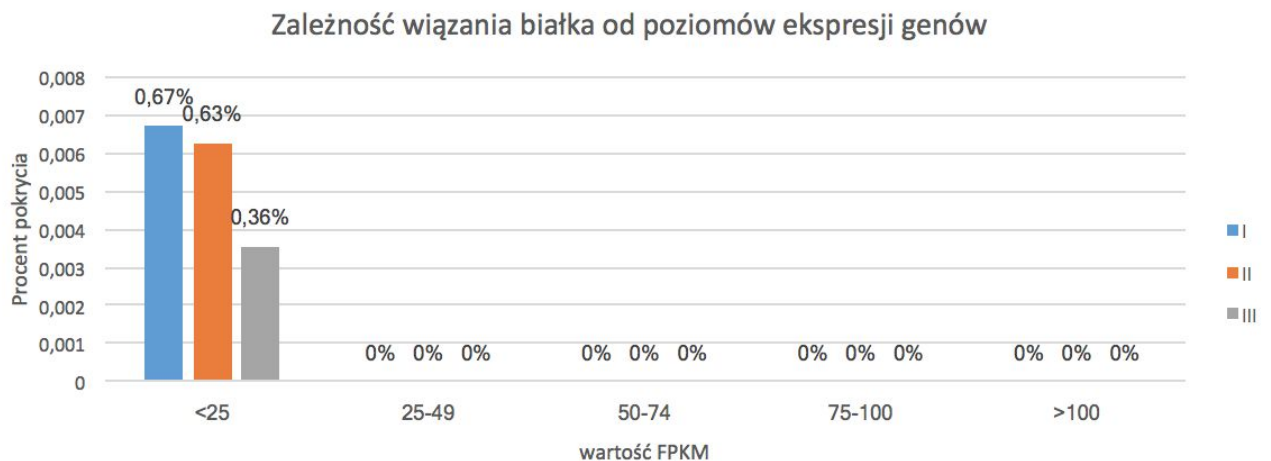
Z analizy ilościowej otrzymane zostały jednak nieco inne wyniki, które wskazywałyby na to, że białko z próbki B na chromosomie II wiąże w 90% na eksonach, a także w przeważającej części na obszarze telomerów. Zastanawiające jest także to, że jedno białko na różnych chromosomach pełni tak różne funkcje. Podejrzany jest też fakt, że dla niektórych wyników pokrycie centromerów przekracza

pokrycie obszaru międzygenowego, co nie powinno zachodzić, jednak trudno określić przyczynę takiego efektu.

Zadanie 4.

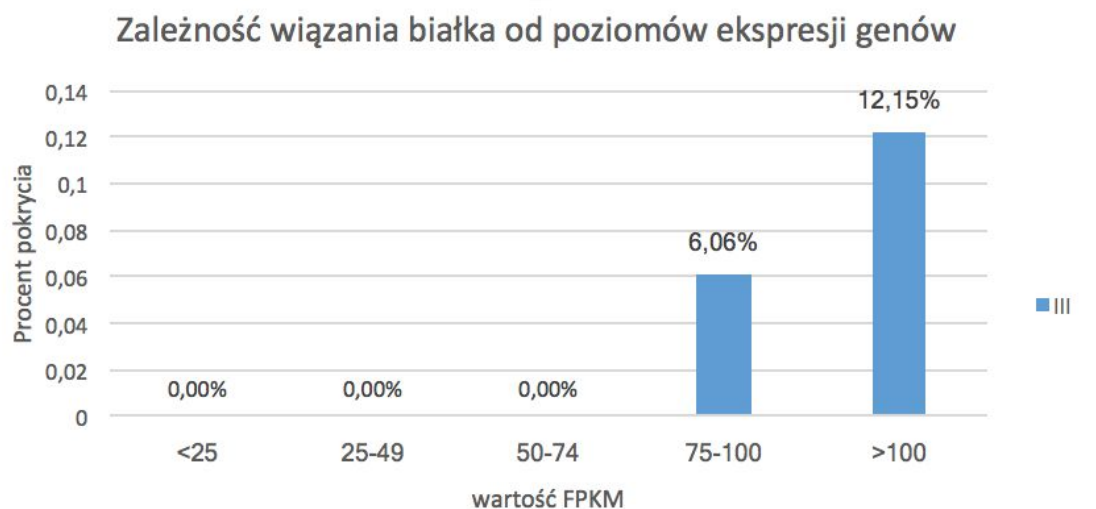
Metoda: Podzieliliśmy geny na 5 grup ze względu na poziom ekspresji. Następnie dla każdej grupy policzyliśmy jaki procent genów w jej obrębie miał pik (obszar genu przecinał się z obszarem piku). Wszystkie obliczenia wykonaliśmy za pomocą skryptu napisanego w języku Python.

1. Próbką B



Wnioski: Próba znalezienia korelacji między ekspresją genów, a wiązaniem do nich białka, skończyła się niepowodzeniem. Jedynie w grupie o wartości ekspresji mniejszej niż 25 FPKM uzyskaliśmy pewny związek, jednak jest on bardzo mały.

2. Próbką C



Wnioski: Otrzymaliśmy pewną korelację między ekspresją genów, a wiązaniem białka. Na wykresie widzimy, że białko wiąże się z genami z ekspresją powyżej 75 FPKM, a szczególnie powyżej 100 FPKM.