



WROCŁAW UNIVERSITY  
OF ENVIRONMENTAL  
AND LIFE SCIENCES

# Analiza danych NGS

Marek Sztuka 117373, Paweł Grygielski  
<https://github.com/paq88/Nextflow-gatk-variant-calling>



# Cel Pracy

- Stworzenie Zautomatyzowanego pipeline'u
- Przeprowadzenie Variant Callingu na danych NGS
- Analiza wzbogacenia GO



# DANE - eksperyment

TALEDs - Transcription Activator-like Effector-linked Deaminases

- Narzędzia do edytowania genomu mitochondrialnego
- Pracują na jednej nici DNA
- A > G
- Problem - mitDNA jest dwuniciowe
- Mismatch > BER > ssDNA > TALED
- Stworzyli bardziej zaawansowaną formę TALED'u
- Wykorzystali ją do wprowadzenia mutacji w mtDNA

# DANE

## Próbki Człowieka

- SRR32281629 Ulepszony TALED6
- SRR32281627 Standardowy TALED

## Parametry:

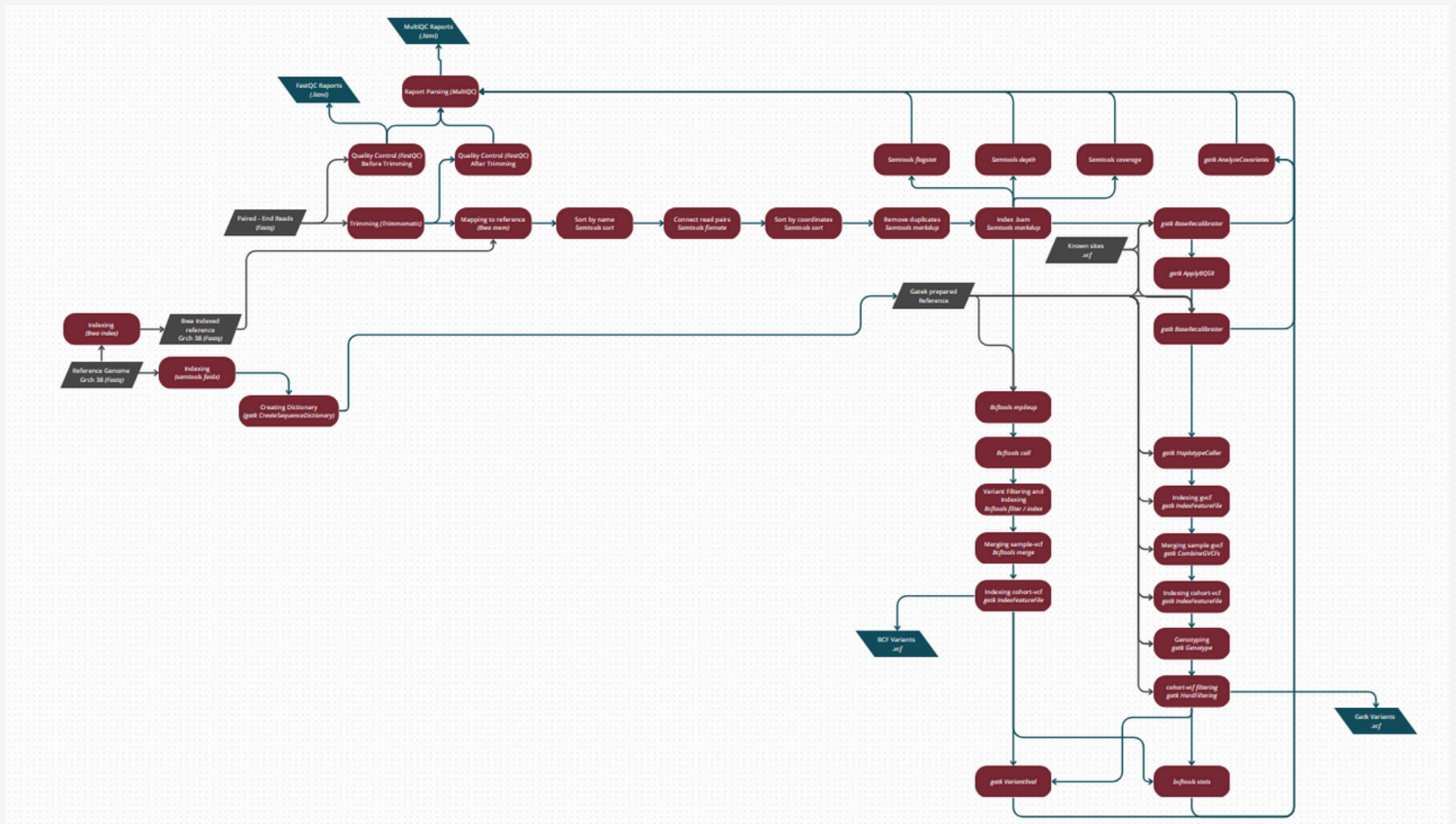
- NovaSeq6000
- WGS
- Sparowane
- Random Selection

# Input

- Próbkki fastq
- Referencja: Homo sapiens grch38
- Known sites: Mills\_and\_1000G\_gold\_standard
- Adaptery
  - PolyA
  - TruSeq

# Pipeline (link)

- Nextflow
- BCFTTools
- GATK
- Zautomatyzowany Setup
- Raporty
  - QC
  - post alignment
  - variant calling
  - base recalibration



# Trimming

ILLUMINACLIP:TruSeq3-PE.fa:2:30:10

- Usuwa adaptory Illumina (z pliku TruSeq3-PE.fa).
- 2 – maks. dozwolona liczba niedopasowań w sekwencji adaptera.
- 30 – wartość palindromeClipThreshold, dotyczy przypadków, gdy adaptory są obecne jako palindromy.
- 10 – simpleClipThreshold, dotyczy klasycznego dopasowania adaptera.
- ILLUMINACLIP:polyA.fa:2:30:10
- Dodatkowe usuwanie sekwencji poli-A (adapterów z pliku polyA.fa), na tych samych zasadach jak powyżej.

## **SLIDINGWINDOW:4:25**

- Ruchome okno o długości 4. Przycina od momentu, gdy średnia jakość w oknie spada poniżej 25.

## **LEADING:25**

- Usuwa bazy o jakości poniżej 25 z początku odczytu.

## **TRAILING:25**

- Usuwa bazy o jakości poniżej 25 z końca odczytu.

## **MINLEN:50**

- Odczyty krótsze niż 50 bp po przycięciu są odrzucane.



# Filtracja wariantów

## Bcftools filter

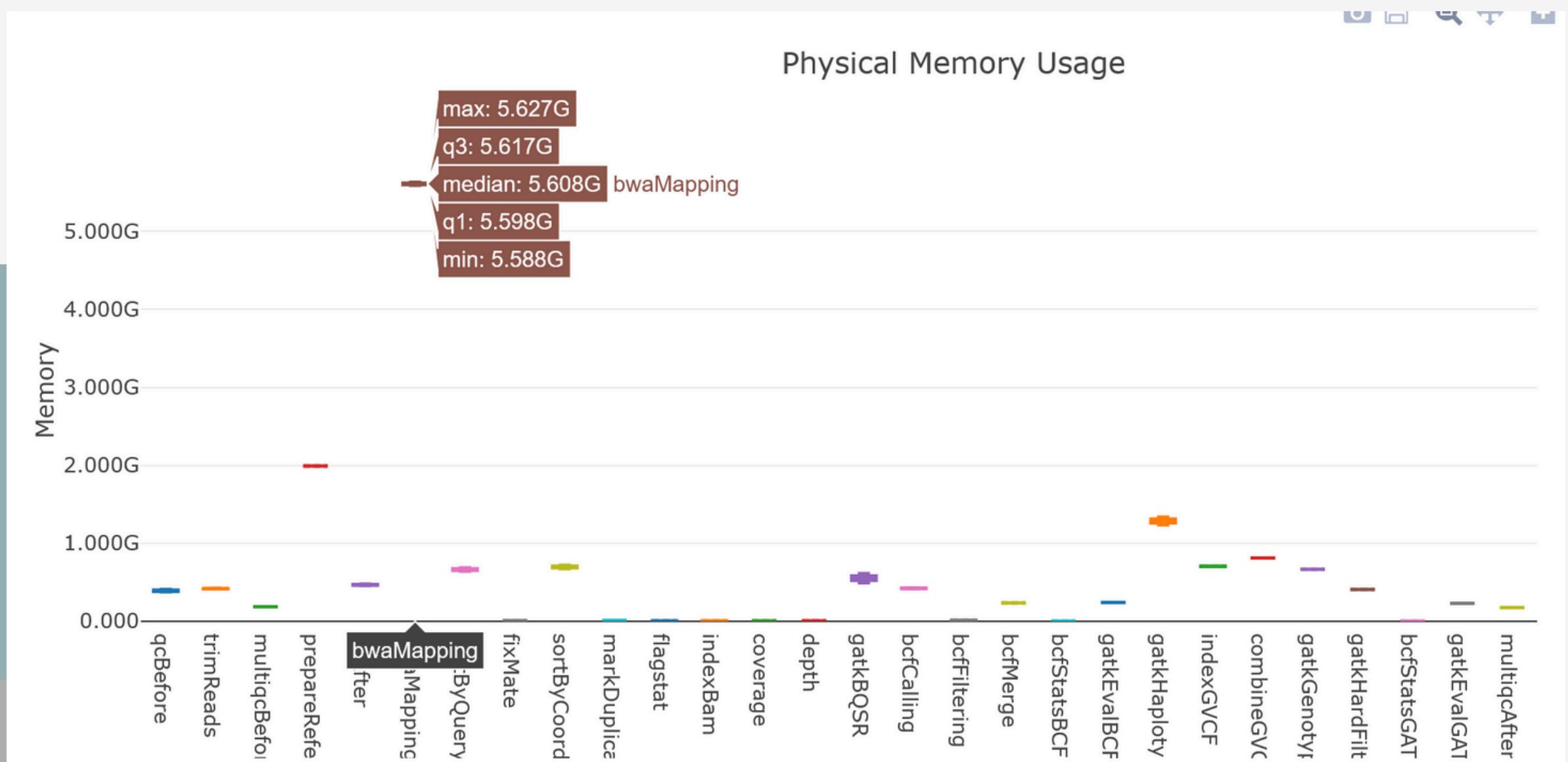
- `QUAL >= 30 && QUAL <= 1000`
- `DP >= 10 && DP <= 1000`
- `MQ >= 40`
- `MQ0F <= 0.1`
- `MQBZ >= -3 && MQBZ <= 3`
- `MQSBZ >= -3 && MQSBZ <= 3`
- `BQBZ >= -3 && BQBZ <= 3`
- `SCBZ >= -3 && SCBZ <= 3`

## Gatk VariantFiltration

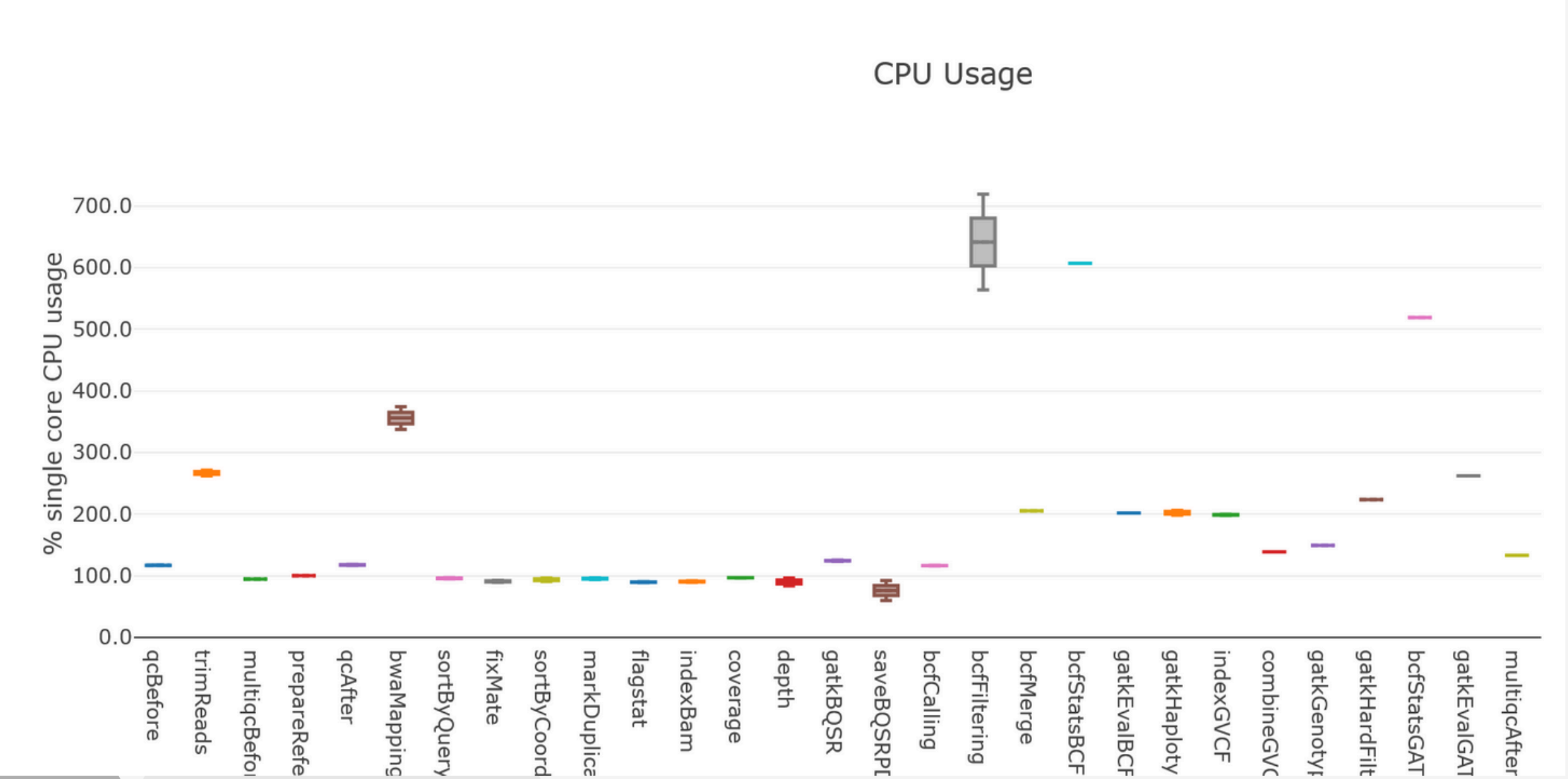
- `QD < 2.0`
- `DP < 10 || DP > 1000`
- `MQ < 40`
- `FS > 60`
- `SOR > 3`
- `MQRankSum < -12.5`
- `ReadPosRankSum < -8`
- `BaseQRankSum < -8`



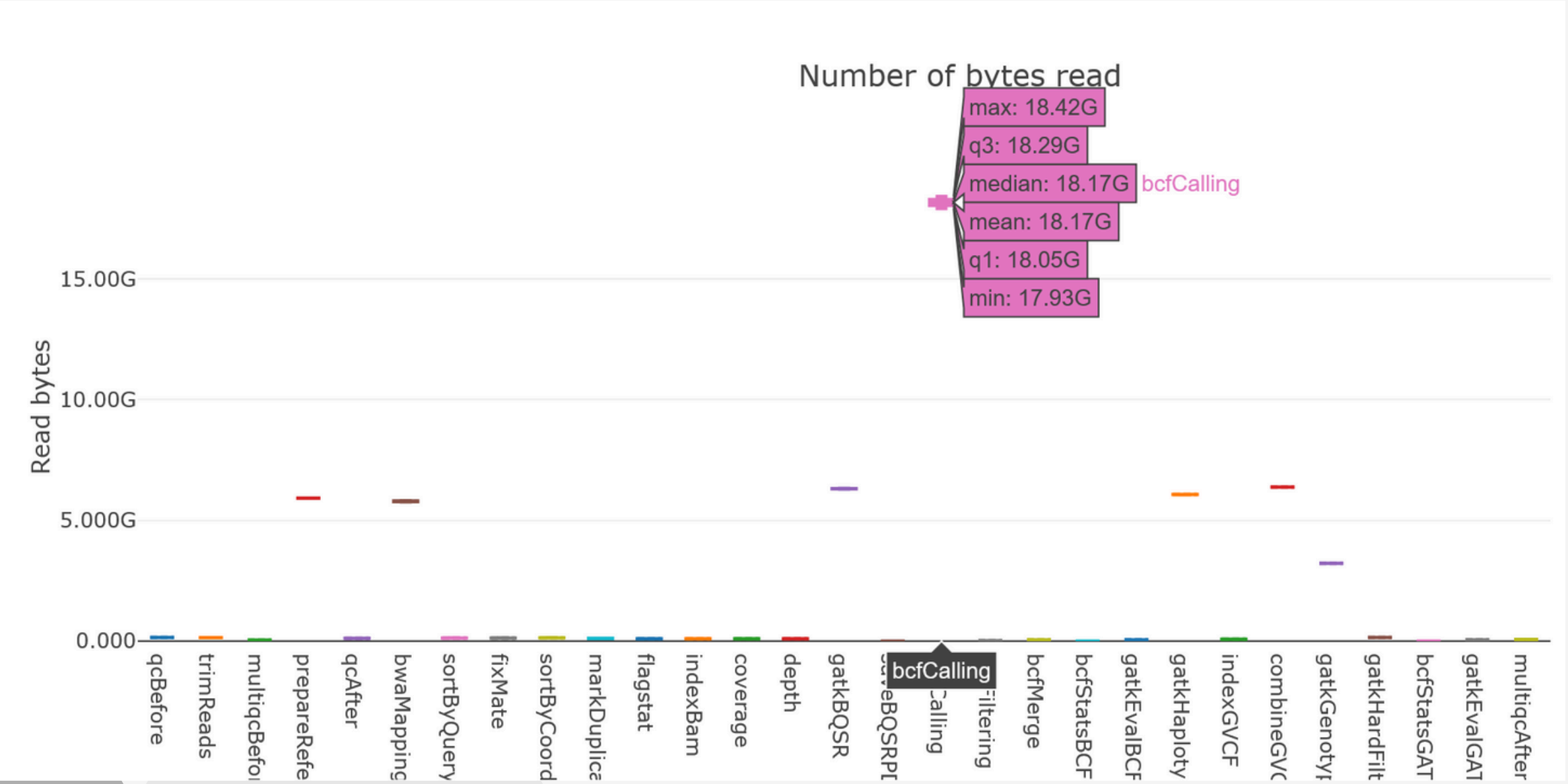
# Zasoby - RAM



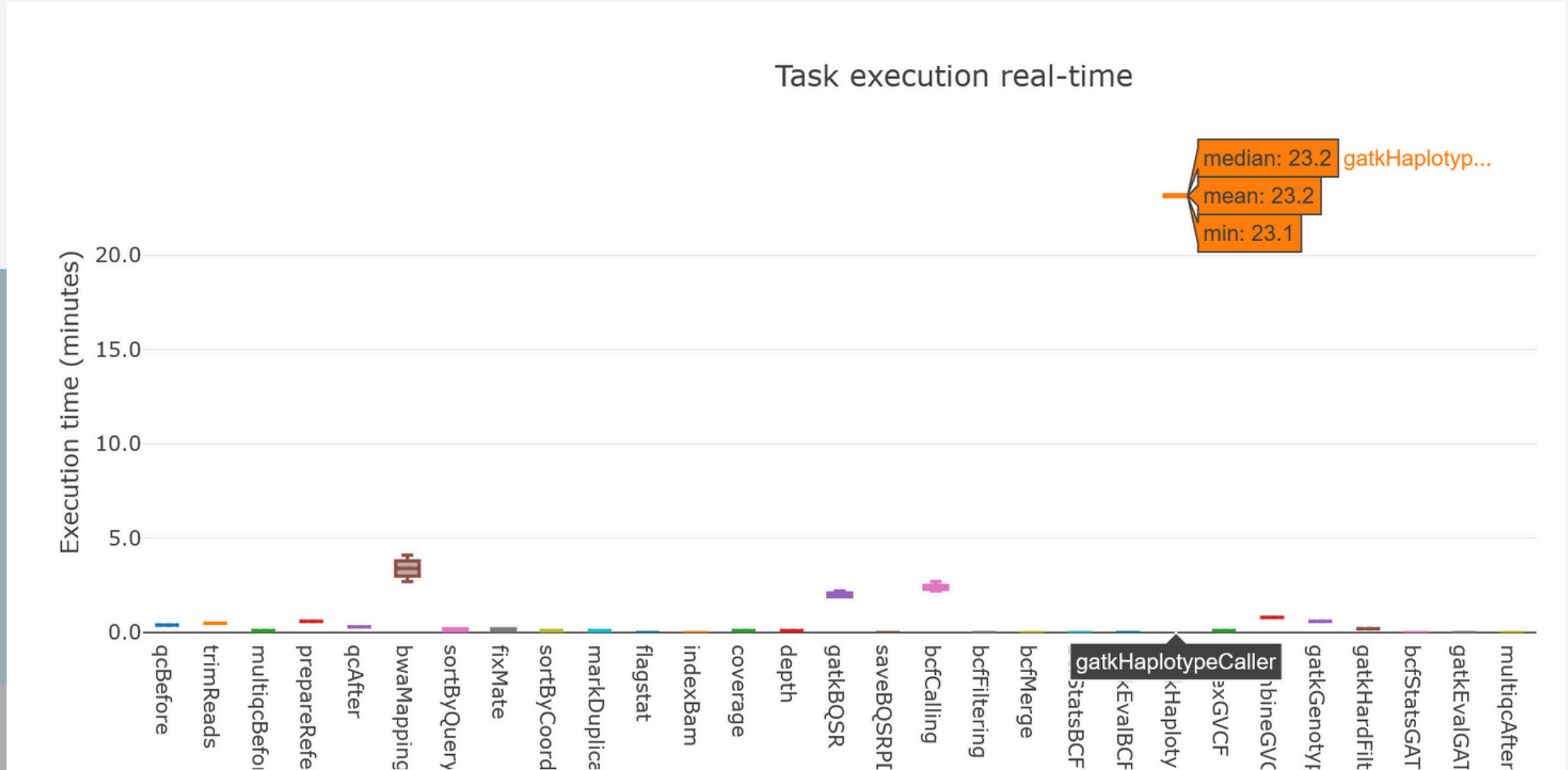
# Zasoby - CPU



# Zasoby - Dysk



# Zasoby - Czas



# WYNIKI



# Kontrola jakości

## General Statistics

Copy tableConfigure columnsScatter plotViolin plotExport as CSV...Showing 4/4 rows and 3/3 columns.Summarize table

Sample Name	Dups	GC	Seqs
SRR32281627_1	18.2%	41.0%	1.0 M
SRR32281627_2	17.8%	41.0%	1.0 M
SRR32281629_1	17.4%	41.0%	1.0 M
SRR32281629_2	17.0%	41.0%	1.0 M

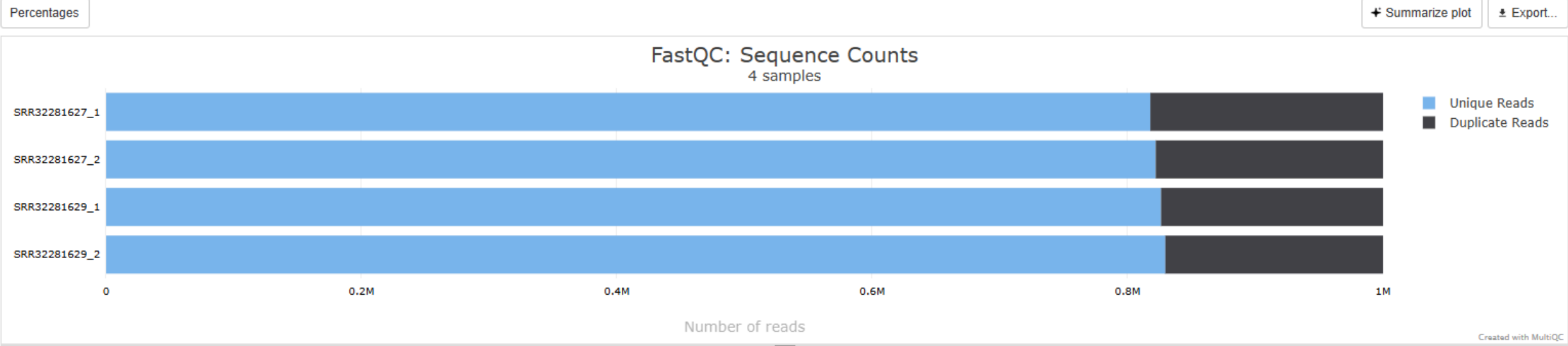
## FastQC

Version: 0.12.1

Quality control tool for high throughput sequencing data. URL: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>

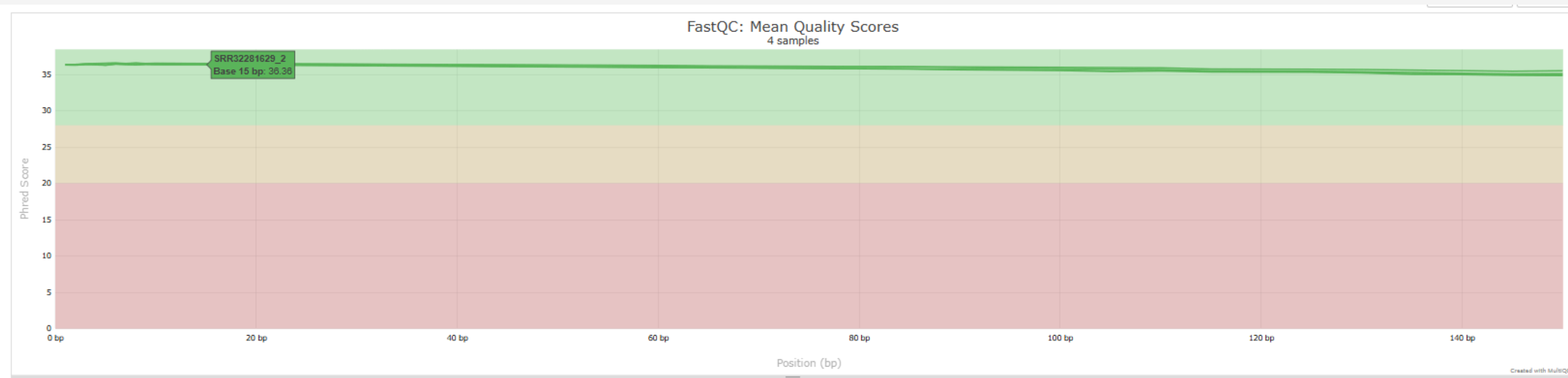
## Sequence Counts

Sequence counts for each sample. Duplicate read counts are an estimate only.



Measure	Value
Filename	SRR32281627_2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1000000
Total Bases	150 Mbp
Sequences flagged as poor quality	0
Sequence length	150
%GC	41

zawartość GC normalna dla człowieka (41%)



### Per Sequence Quality Scores

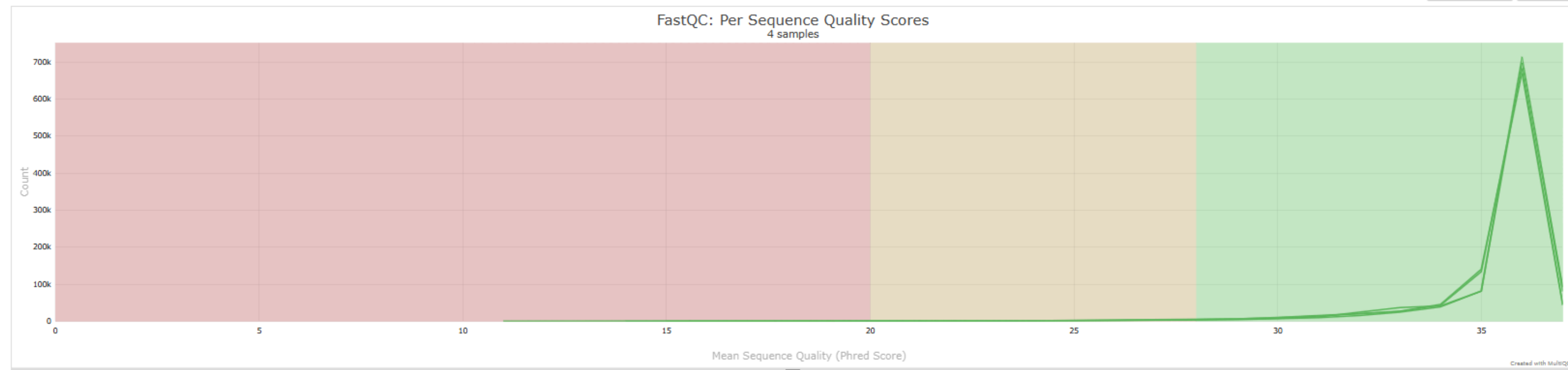
4

The number of reads with average quality scores. Shows if a subset of reads has poor quality.

Help

Summarize plot

Export...



Jakość odczytów jest wysoka i pokrywa się dla obu próbek



## Adapter Content

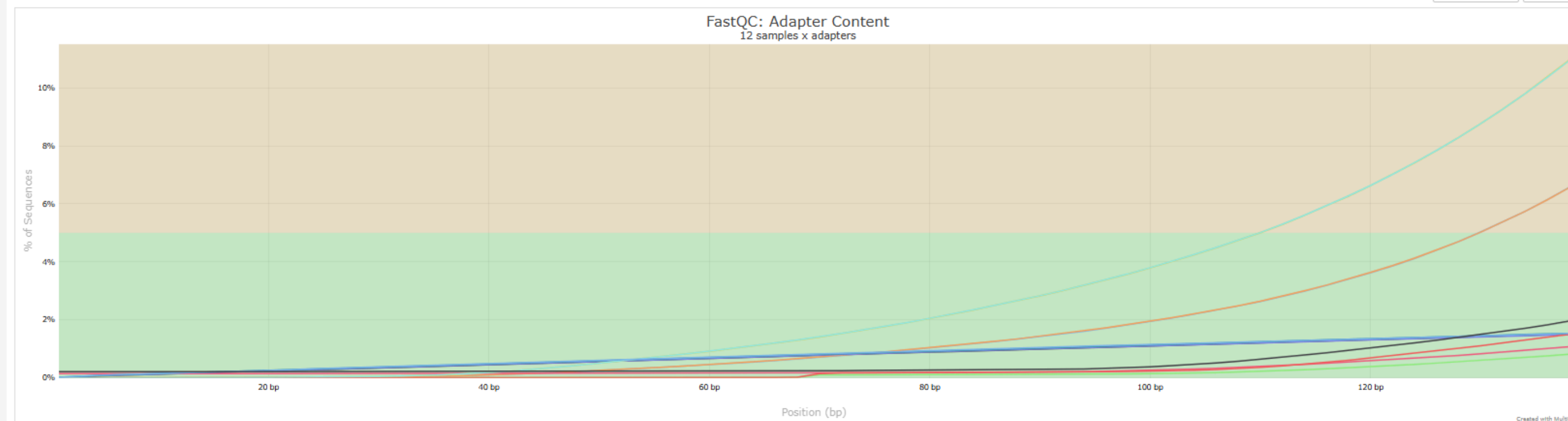
2 2

Help

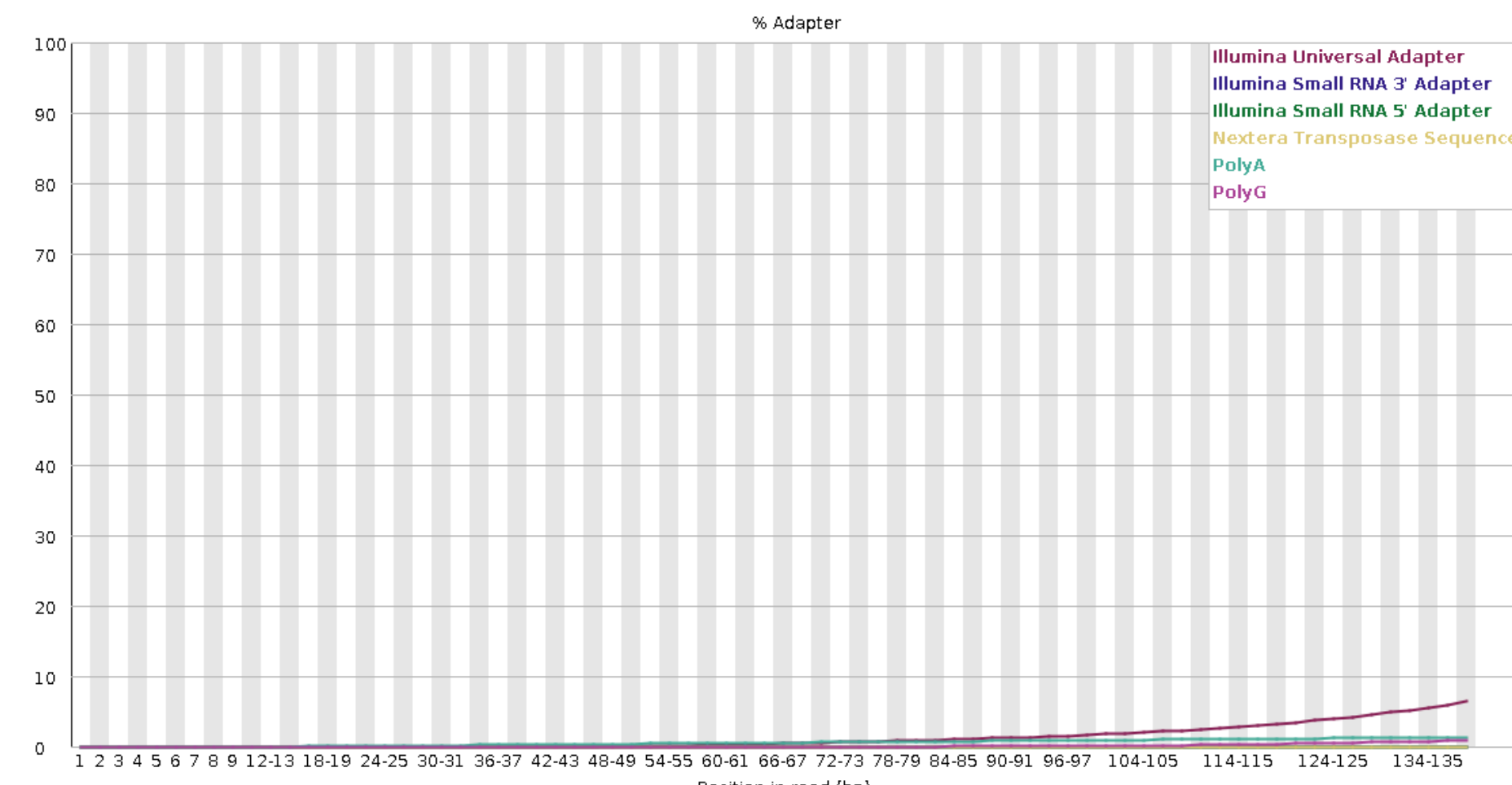
The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.

Summarize plot

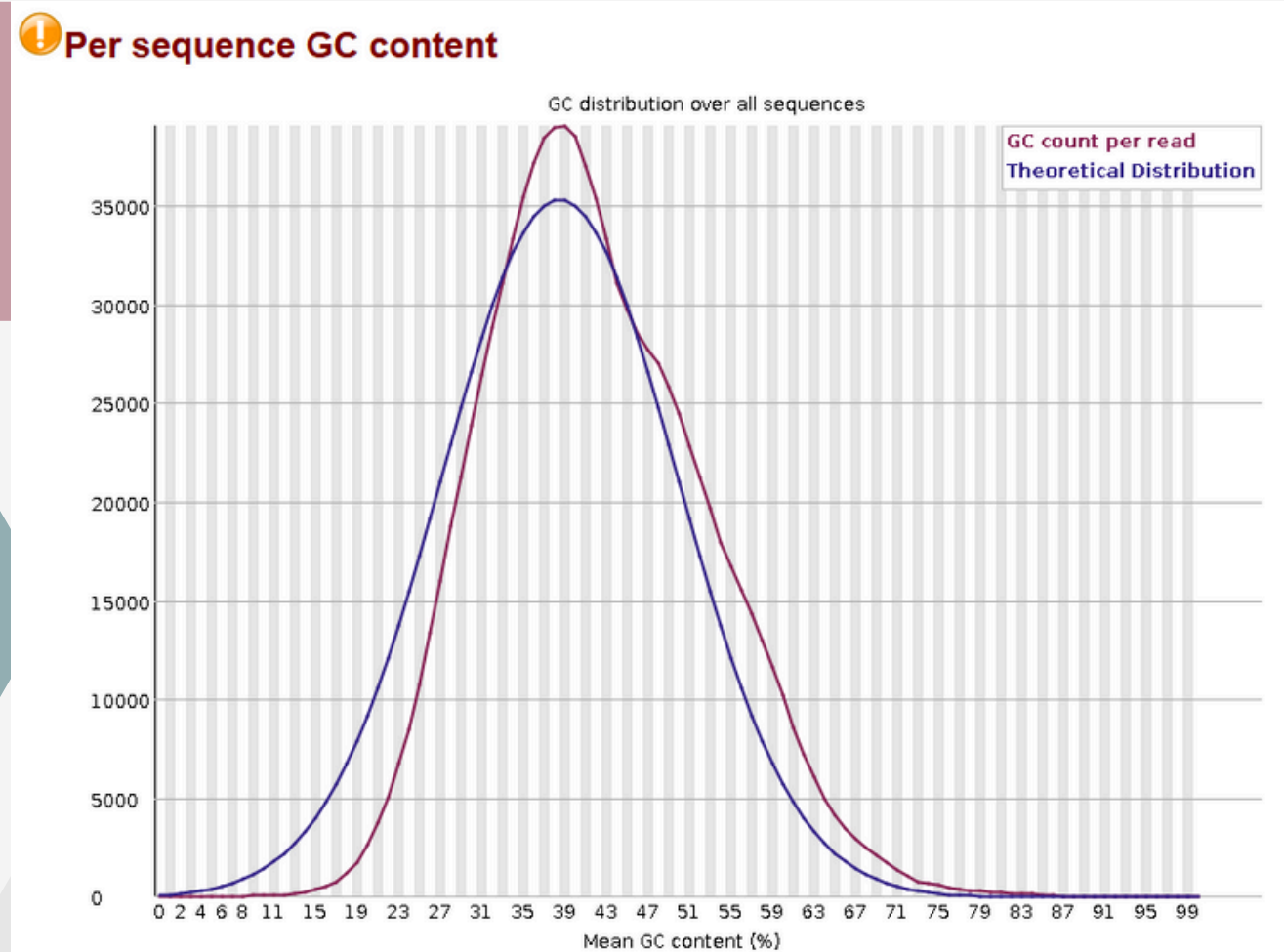
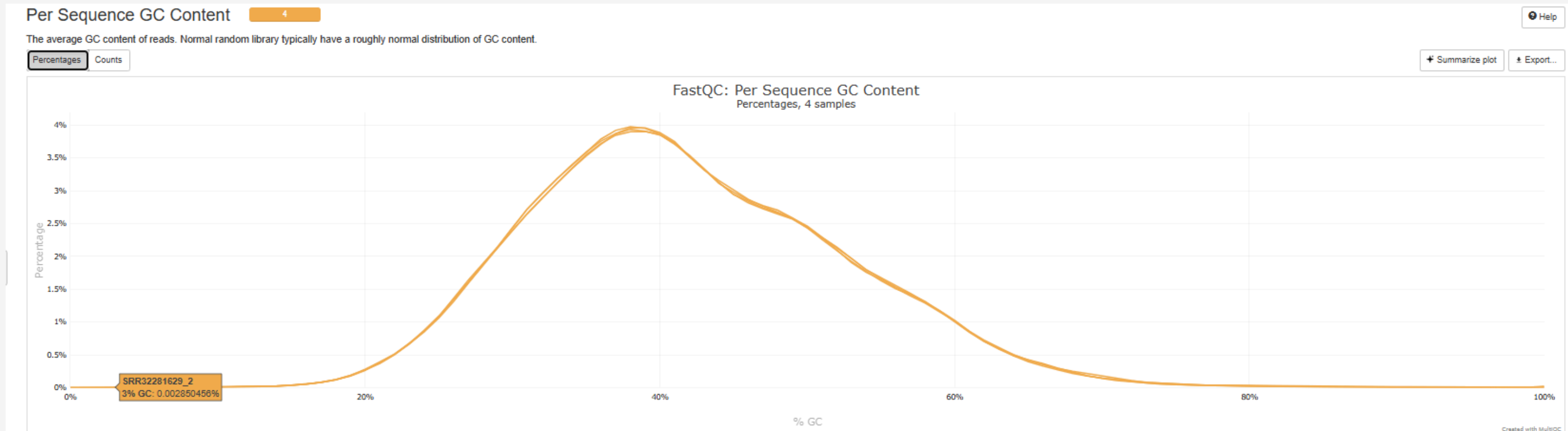
Export...



## Adapter Content



Wzrost obecności sekwencji adapterowych  
pod koniec odczytu  
(dominuje Illumina Universal Adapter)



Rzeczywisty rozkład jest nieco przesunięty w prawo, nie ma jednak poważnych zanieczyszczeń i rozkłady są zbliżone.

# Analiza jakości po trimmingu



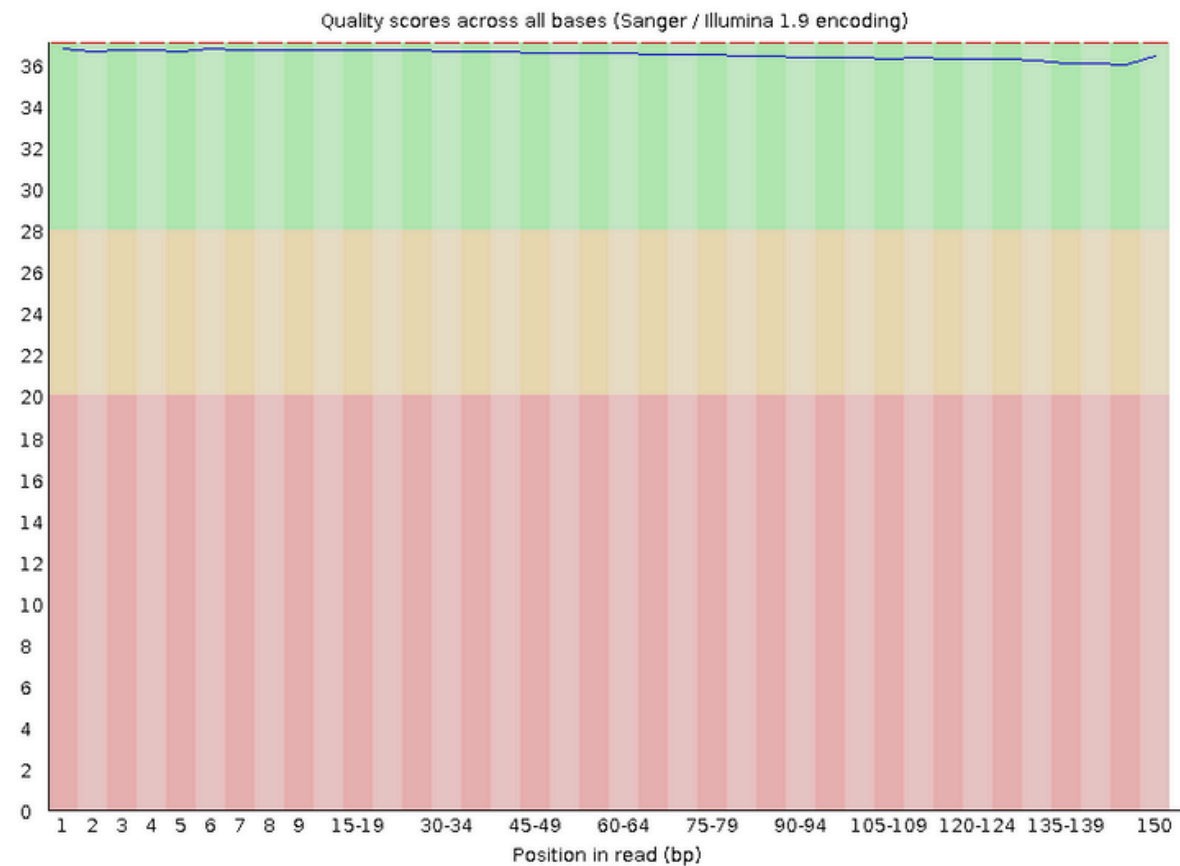
## Basic Statistics

Measure	Value
Filename	SRR32281627_2_trimmed.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	825315
Total Bases	118.8 Mbp
Sequences flagged as poor quality	0
Sequence length	50-150
%GC	41

- Liczba odczytów spadła z miliona do 825 tys.
- Liczba par zasad spadła ze 150 milionów par do 118.8Mbp
- Nie ma odczytów o słabej jakości
- Długości odczytów 50-150 względem poprzedniej długości 150, świadczą o tym, że trimmomatic przyciął te fragmenty, które wymagały przycięcia

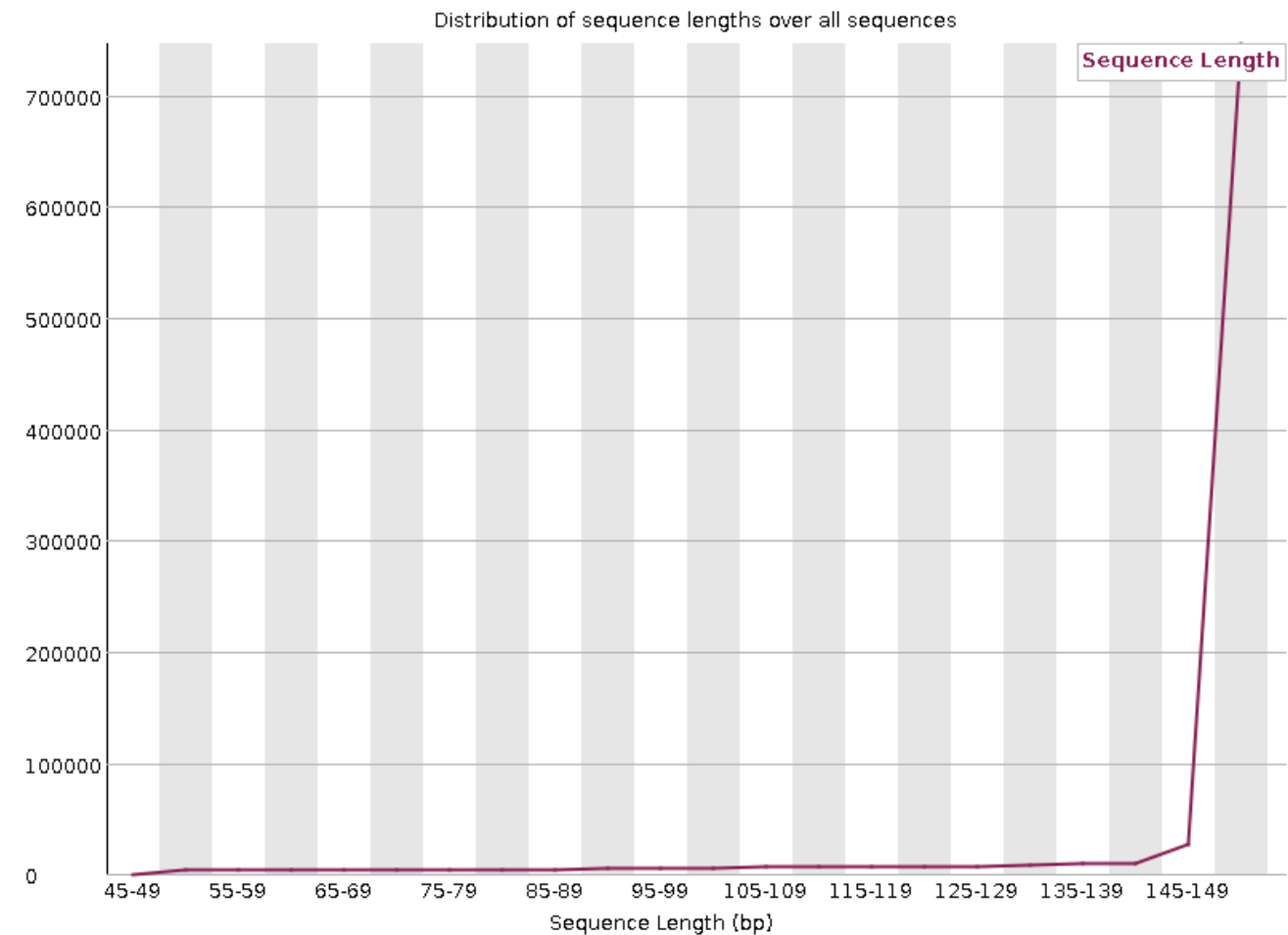
Jakość wciąż jest bardzo dobra.

✔ Per base sequence quality



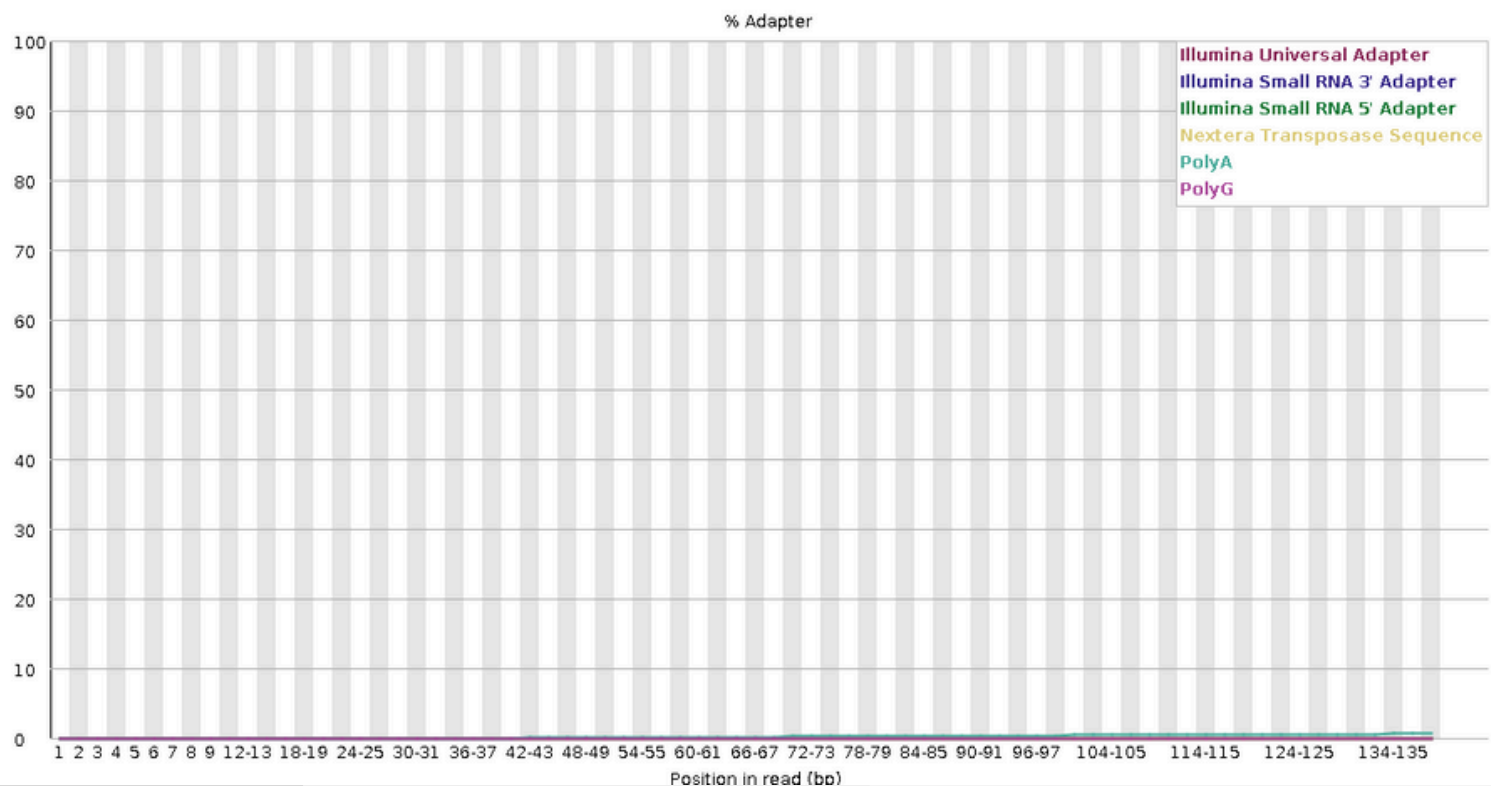
Długość większości odczytów  
jest jednakowa

⚠ Sequence Length Distribution



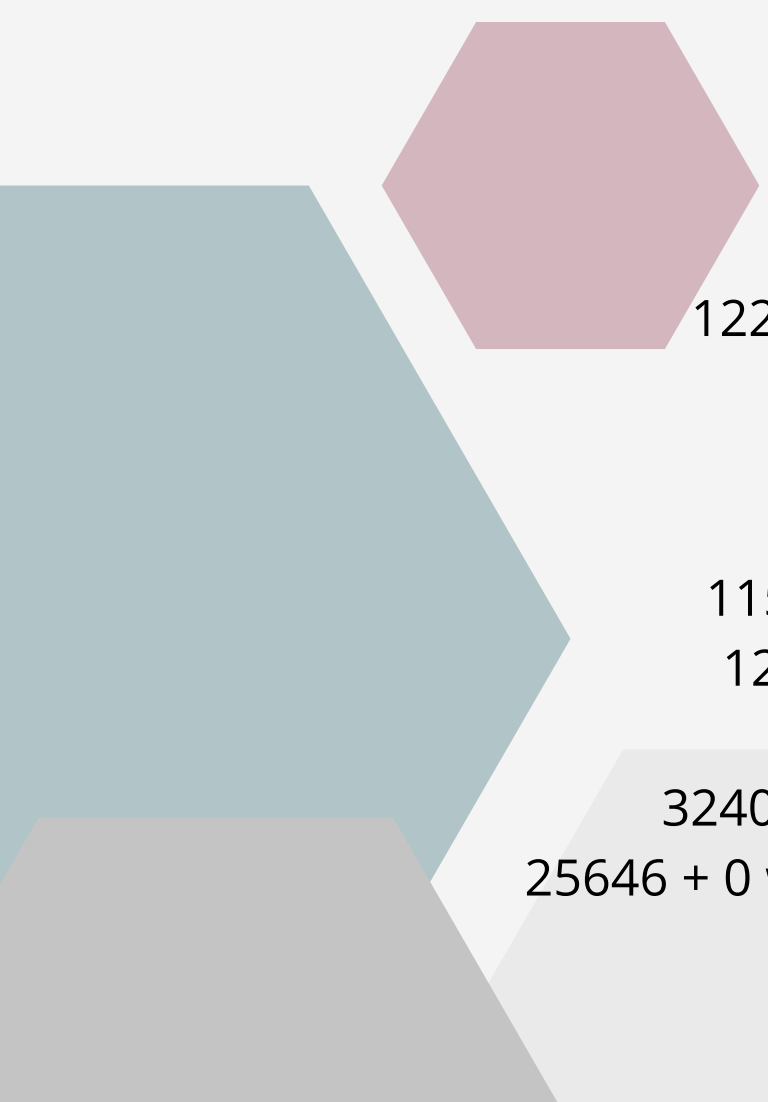
Udało się zmniejszyć ilość  
sekwencji adapterowych prawie do zera

✔ Adapter Content



# Statystyki po mapowaniu

SRR32281627



1269124 + 0 in total (QC-passed reads + QC-failed reads)  
1234244 + 0 primary  
0 + 0 secondary  
34880 + 0 supplementary  
0 + 0 duplicates  
0 + 0 primary duplicates  
1255574 + 0 mapped (98.93% : N/A)  
1220694 + 0 primary mapped (98.90% : N/A)  
1234244 + 0 paired in sequencing  
617103 + 0 read1  
617141 + 0 read2  
1153598 + 0 properly paired (93.47% : N/A)  
1220370 + 0 with itself and mate mapped  
324 + 0 singletons (0.03% : N/A)  
32404 + 0 with mate mapped to a different chr  
25646 + 0 with mate mapped to a different chr (mapQ>=5)

SRR32281629

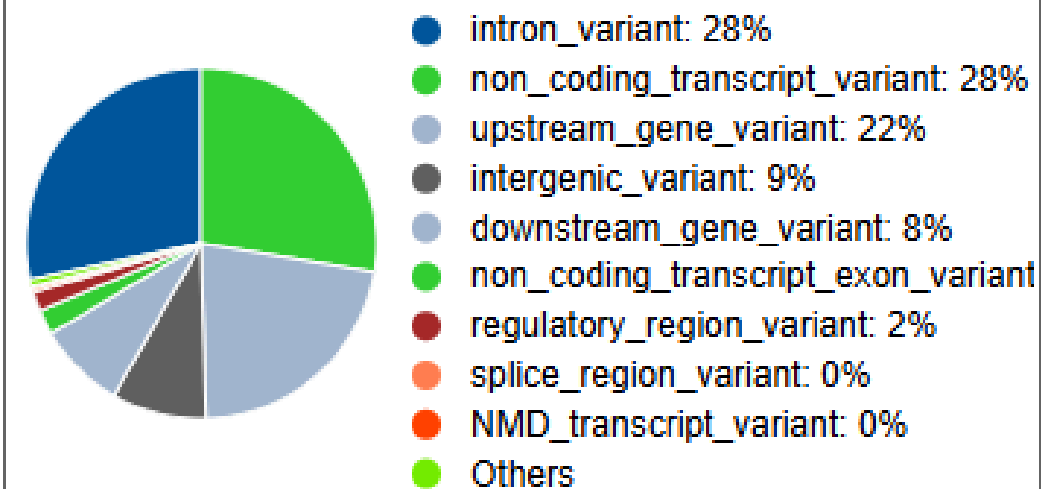
1185881 + 0 in total (QC-passed reads + QC-failed reads)  
1161371 + 0 primary  
0 + 0 secondary  
24510 + 0 supplementary  
0 + 0 duplicates  
0 + 0 primary duplicates  
1179779 + 0 mapped (99.49% : N/A)  
1155269 + 0 primary mapped (99.47% : N/A)  
1161371 + 0 paired in sequencing  
580676 + 0 read1  
580695 + 0 read2  
1110530 + 0 properly paired (95.62% : N/A)  
1155126 + 0 with itself and mate mapped  
143 + 0 singletons (0.01% : N/A)  
21752 + 0 with mate mapped to a different chr  
16682 + 0 with mate mapped to a different chr (mapQ>=5)

# Wyniki GATK (VEP)

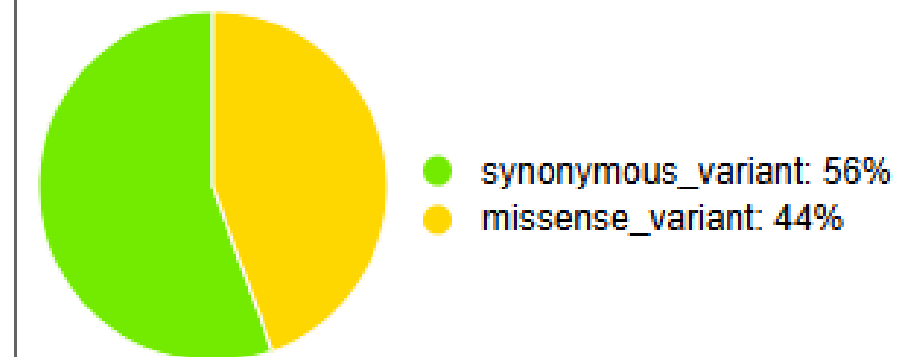
SRR32281627

Category	Count
Variants processed	527
Variants filtered out	0
Novel / existing variants	212 (40.2) / 315 (59.8)
Overlapped genes	57
Overlapped transcripts	191
Overlapped regulatory features	10

Consequences (all)



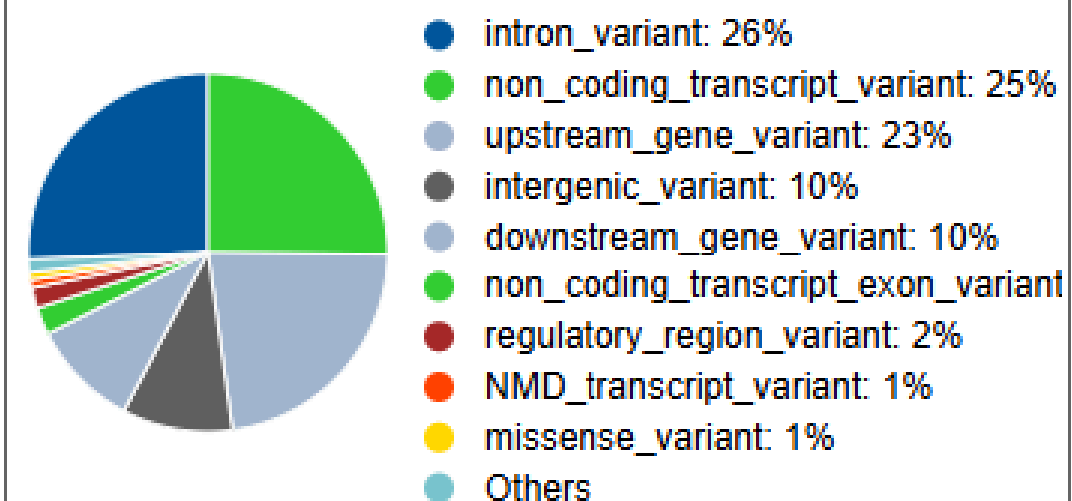
Coding consequences



SRR32281629

Category	Count
Variants processed	593
Variants filtered out	0
Novel / existing variants	257 (43.3) / 336 (56.7)
Overlapped genes	57
Overlapped transcripts	236
Overlapped regulatory features	10

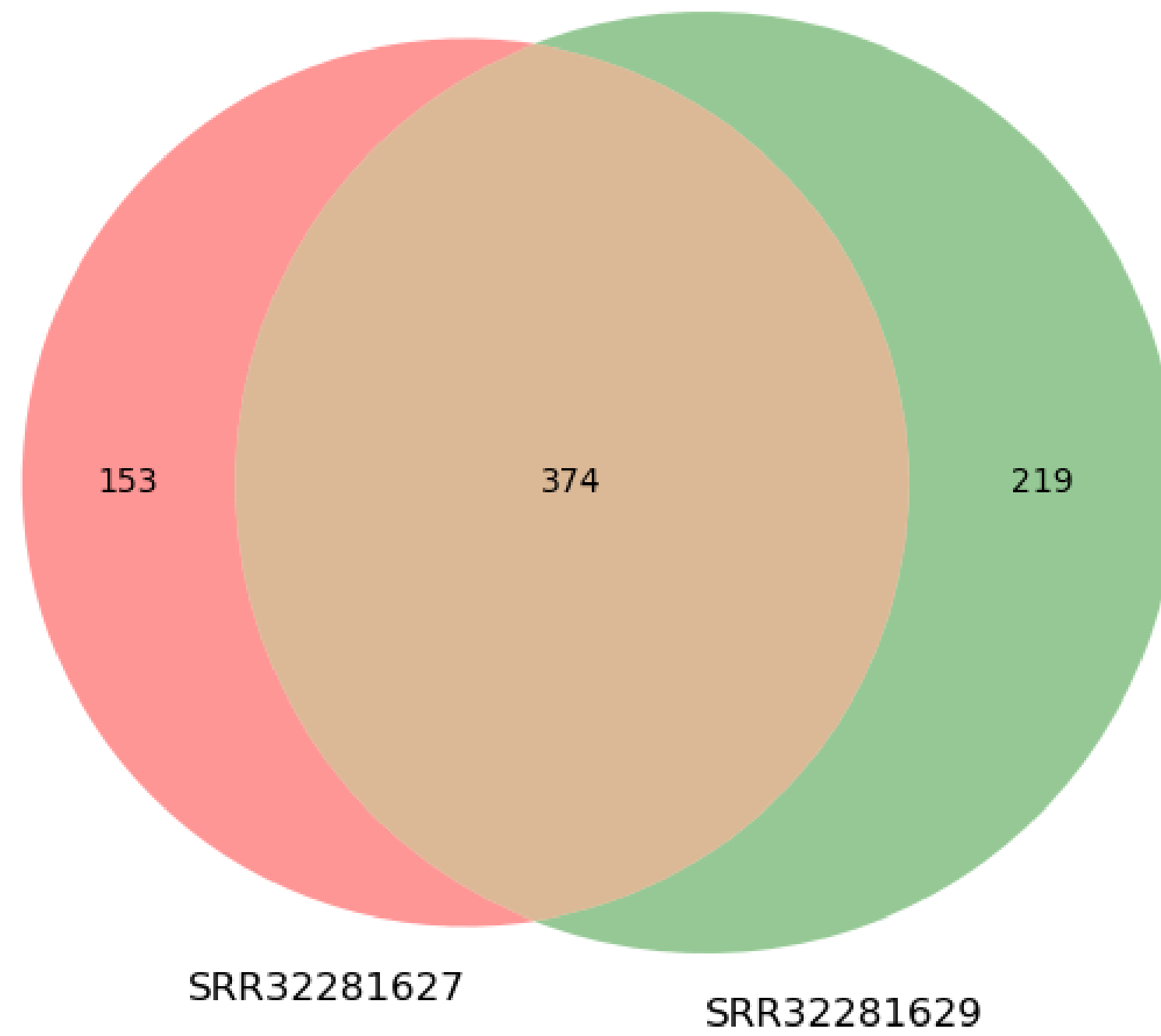
Consequences (all)



Coding consequences



# Porównanie wariantów



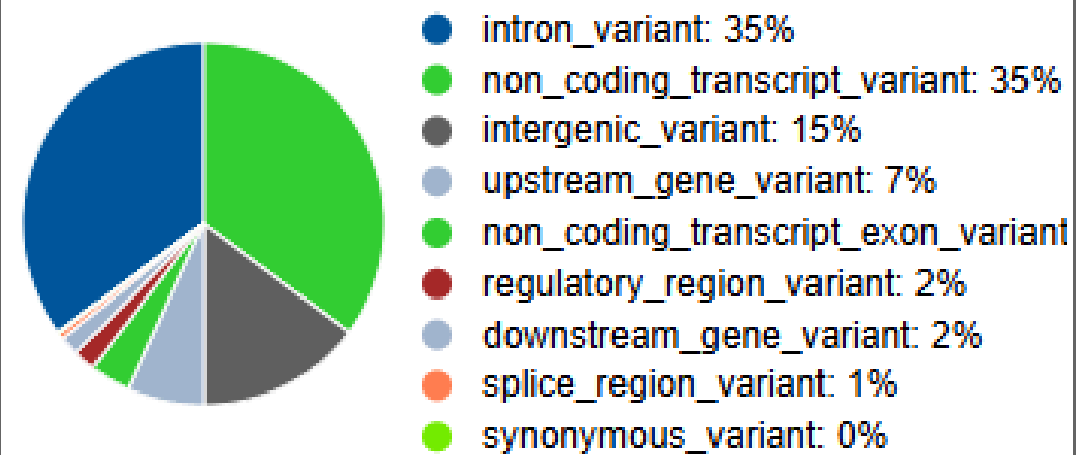


# Wyniki BCF (VEP)

SRR32281627

Category	Count
Variants processed	143
Variants filtered out	0
Novel / existing variants	75 (52.4) / 68 (47.6)
Overlapped genes	21
Overlapped transcripts	71
Overlapped regulatory features	4

Consequences (all)



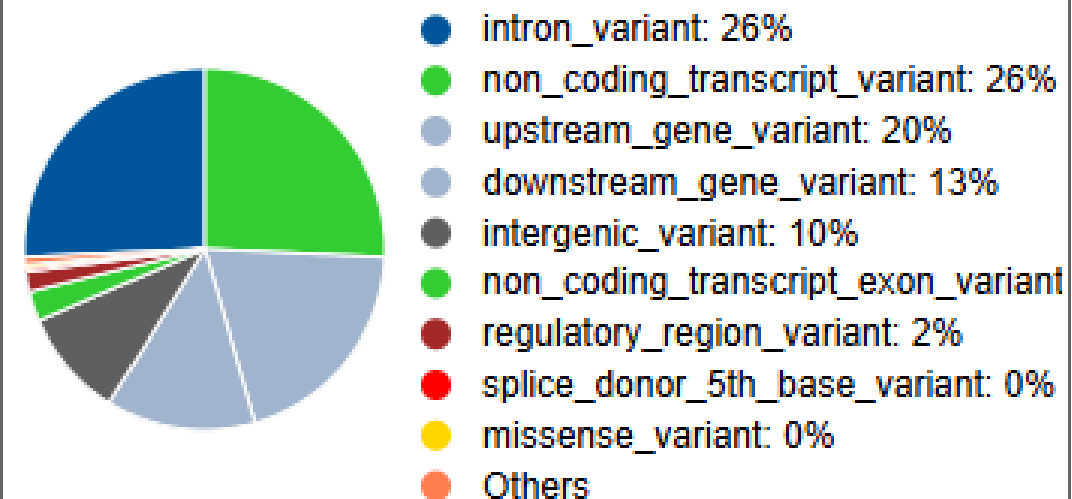
Coding consequences



SRR32281629

Category	Count
Variants processed	183
Variants filtered out	0
Novel / existing variants	91 (49.7) / 92 (50.3)
Overlapped genes	48
Overlapped transcripts	140
Overlapped regulatory features	5

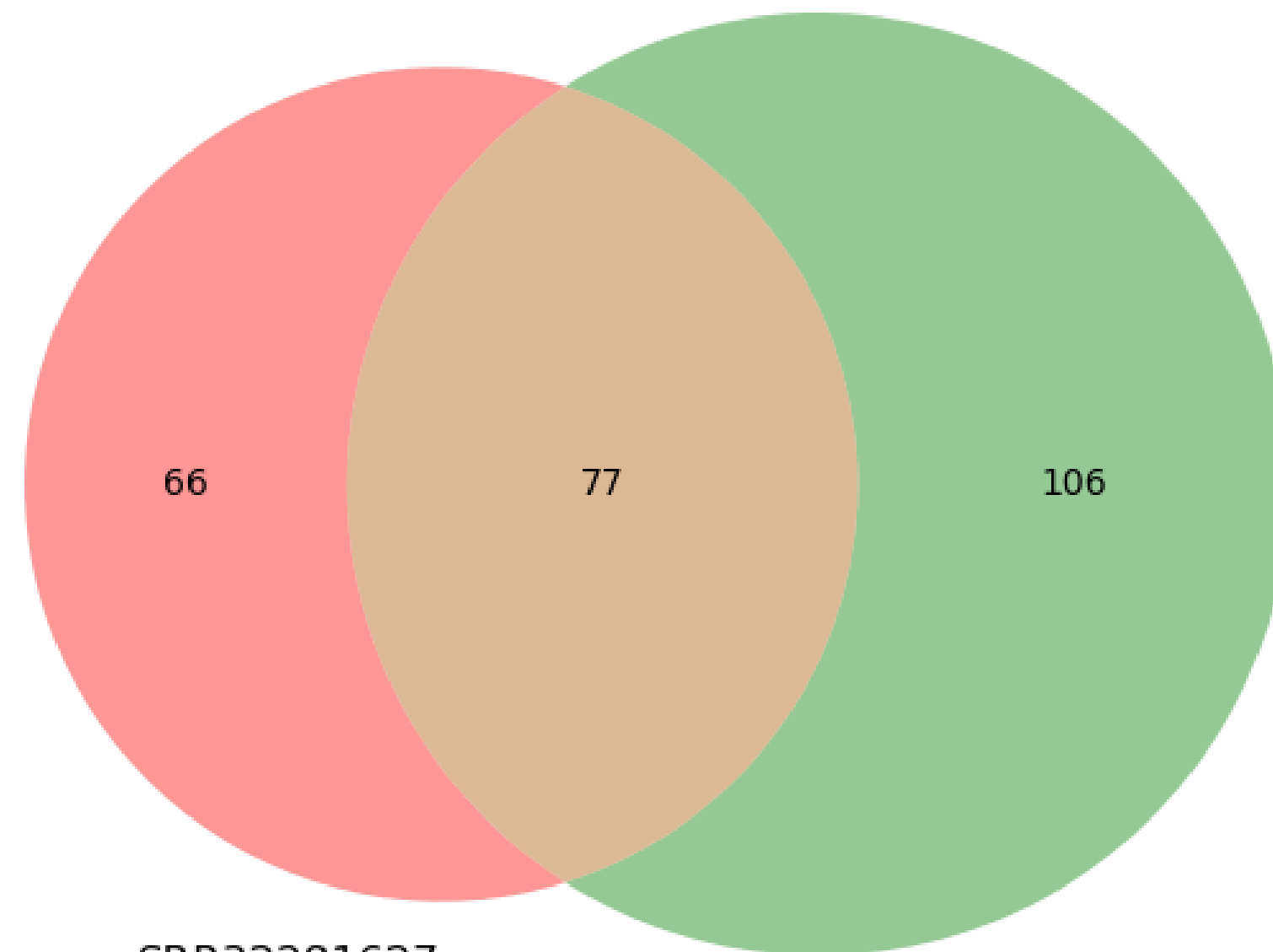
Consequences (all)



Coding consequences



# Porównanie wariantów



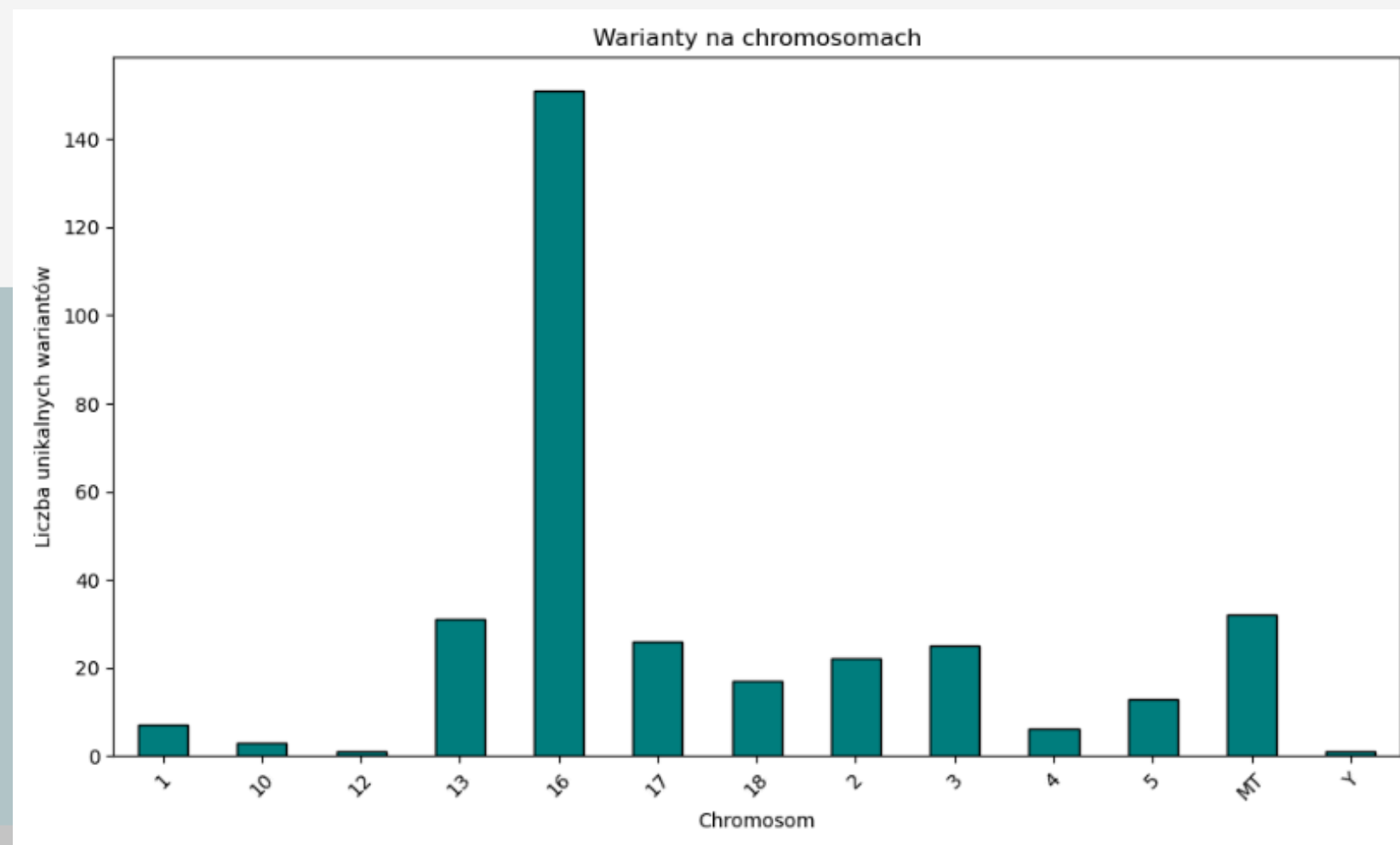
SRR32281627

SRR32281629

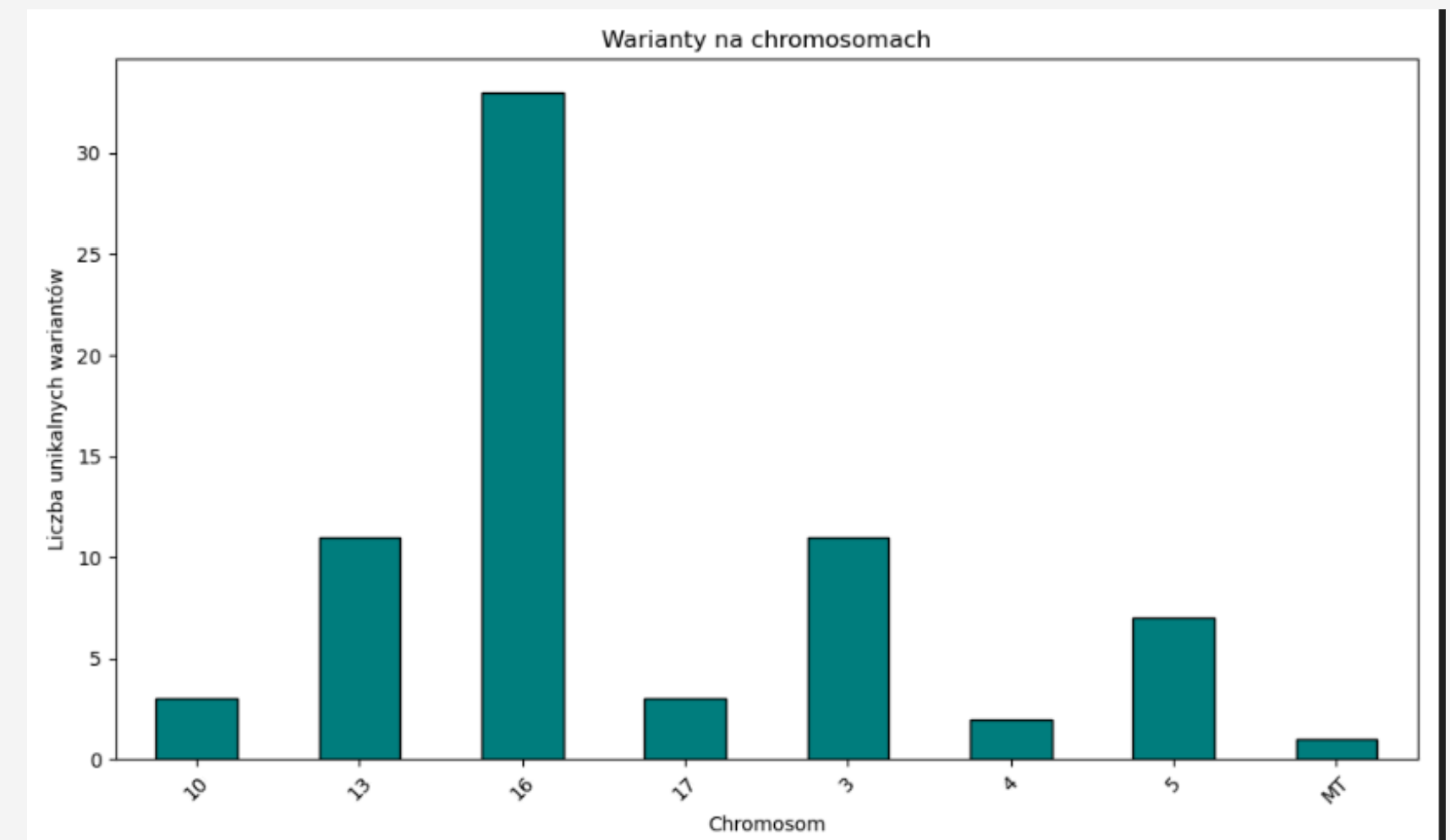
# Liczba wariantów na chromosomach

SRR32281627

GATK



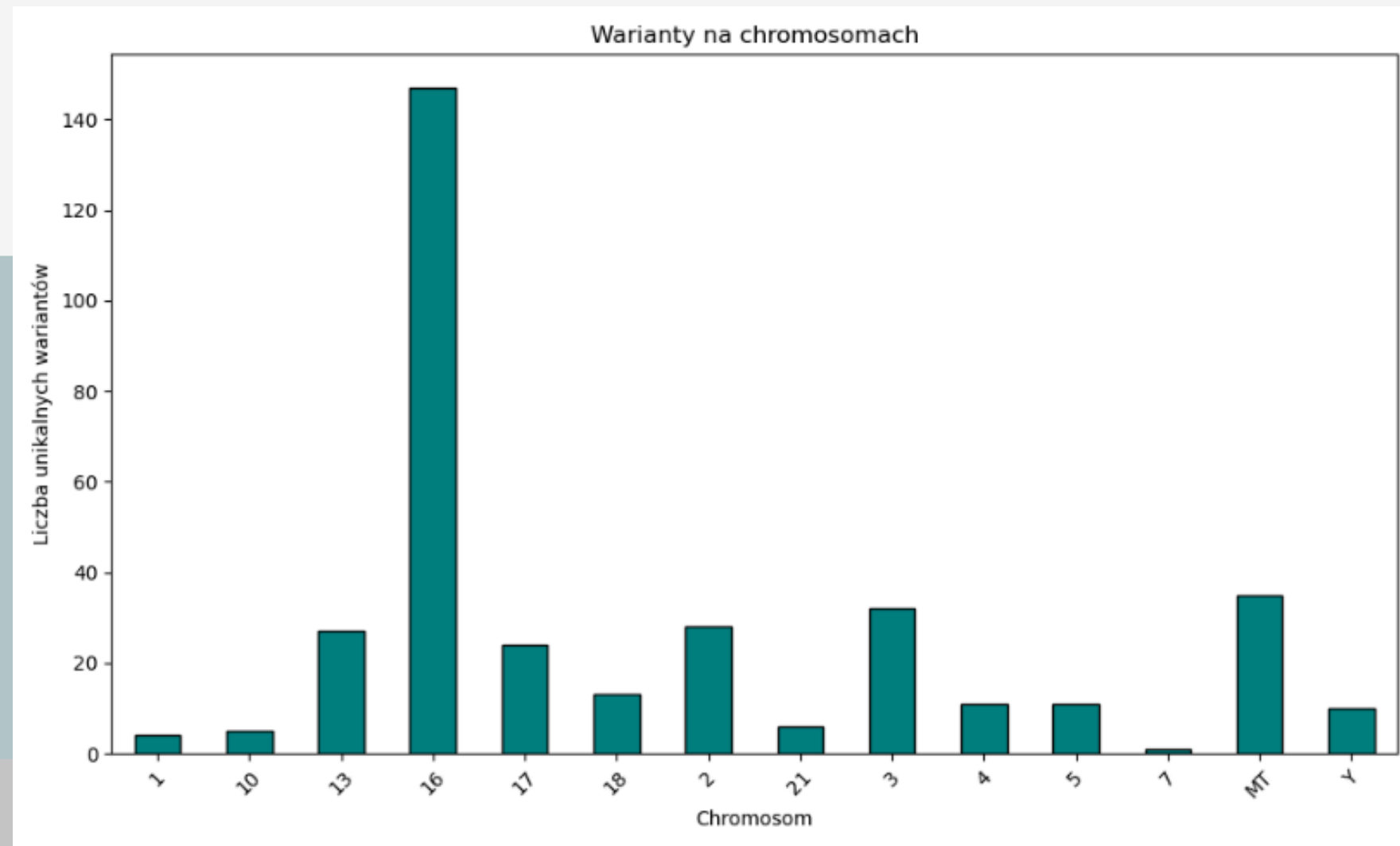
BCF



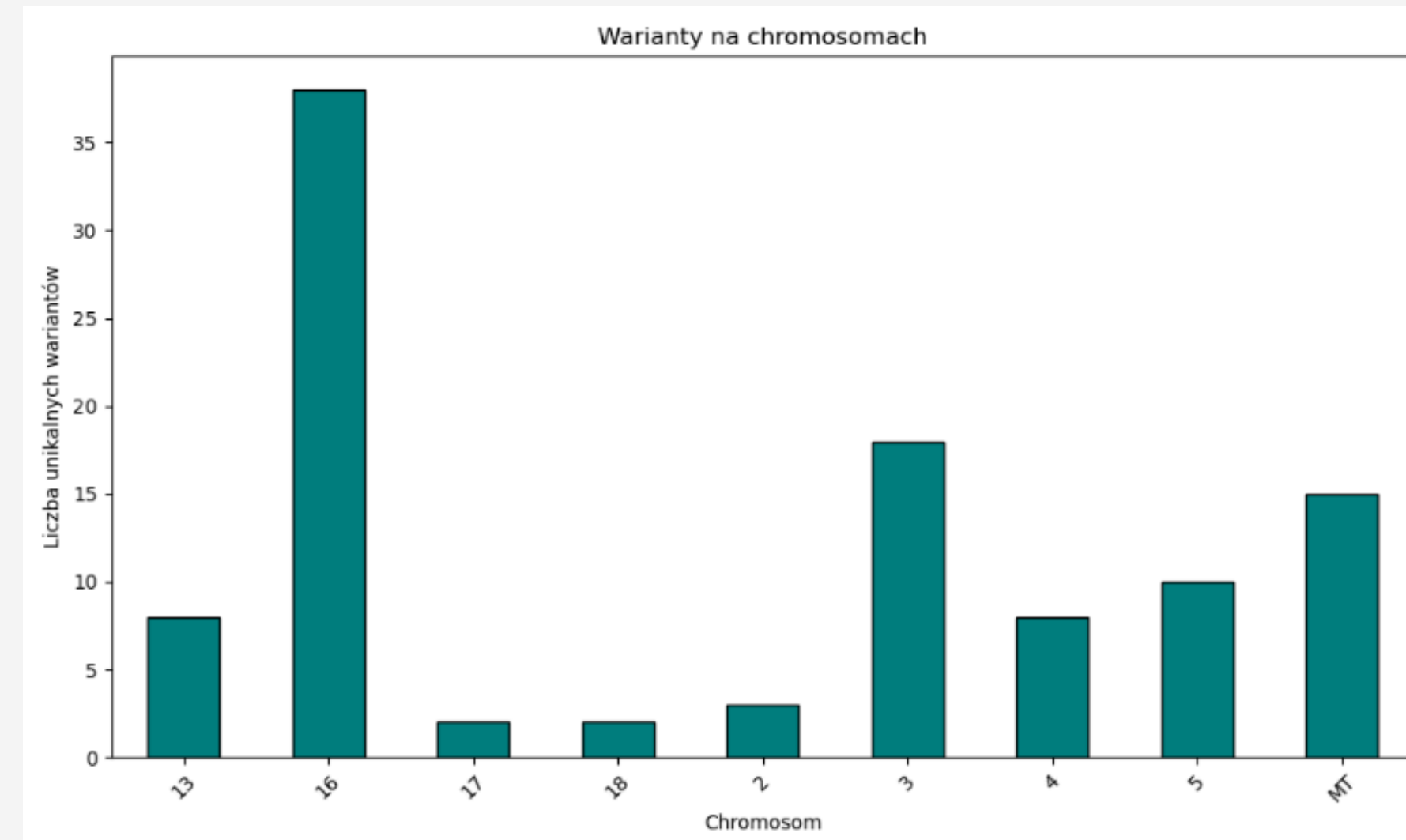
# Liczba wariantów na chromosomach

SRR32281629

GATK



BCF



# Tabela wariantów o wysokim wpływie (GATK)

	Chromosom	Pozycja	ID	Ref	Alt	Influence	Gen
0	13.000000	18211836.000000	.	A	A/T	HIGH	-
1	nan	nan	.	G	G/A	MODERATE	MT-CO3
2	nan	nan	.	A	A/G	MODERATE	MT-ND5
3	nan	nan	.	A	A/G	MODERATE	MT-ND5
4	nan	nan	.	G	G/A	MODERATE	MT-ND5
5	nan	nan	.	C	C/T	MODERATE	MT-CYB
6	nan	nan	.	A	A/G	MODERATE	MT-CYB
7	nan	nan	.	A	A/G	MODERATE	MT-CYB
8	nan	nan	.	A	A/G	MODERATE	MT-CYB

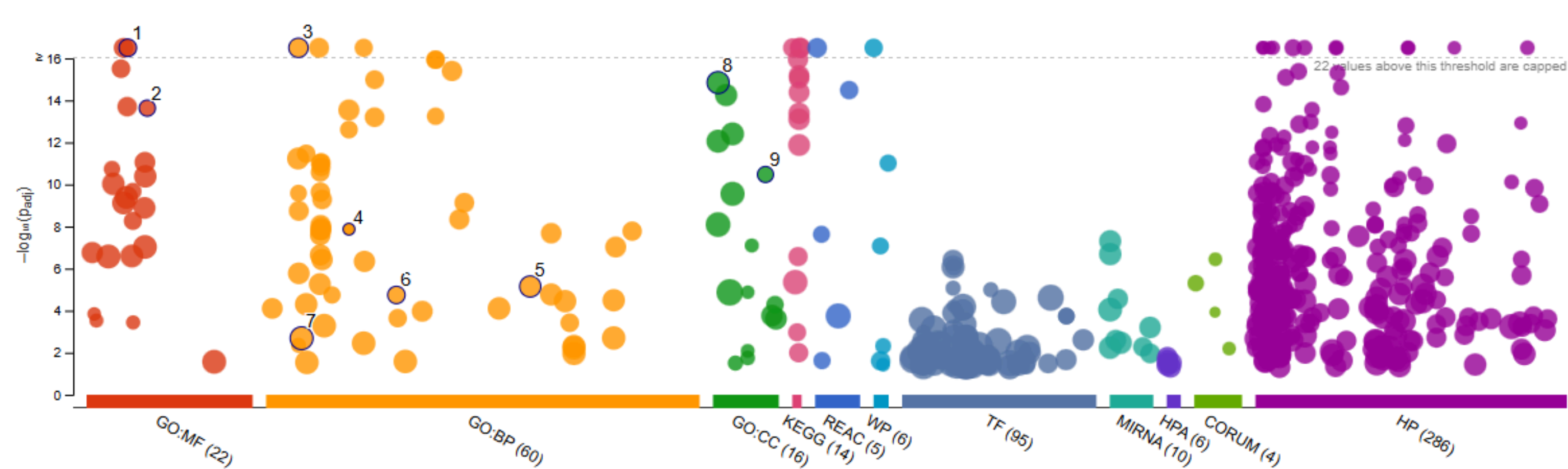
SRR32281627

SRR32281629

	Chromosom	Pozycja	ID	Ref	Alt	Influence	Gen
0	13.000000	18211836.000000	.	A	A/T	HIGH	-
1	nan	nan	.	A	A/G	HIGH	MT-ND5
2	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
3	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
4	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
5	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
6	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
7	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
8	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
9	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
10	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
11	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
12	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
13	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
14	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
15	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
16	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
17	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
18	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
19	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
20	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
21	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
22	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
23	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
24	7.000000	5987357.000000	.	G	G/A	MODERATE	PMS2
25	nan	nan	.	G	G/A	MODERATE	MT-CO3
26	nan	nan	.	A	A/G	MODERATE	MT-ND5
27	nan	nan	.	A	A/G	MODERATE	MT-ND5
28	nan	nan	.	A	A/G	MODERATE	MT-ND5
29	nan	nan	.	G	G/A	MODERATE	MT-ND5
30	nan	nan	.	C	C/T	MODERATE	MT-CYB
31	nan	nan	.	A	A/G	MODERATE	MT-CYB
32	nan	nan	.	A	A/G	MODERATE	MT-CYB
33	nan	nan	.	A	A/G	MODERATE	MT-CYB

# Enrichment analysis

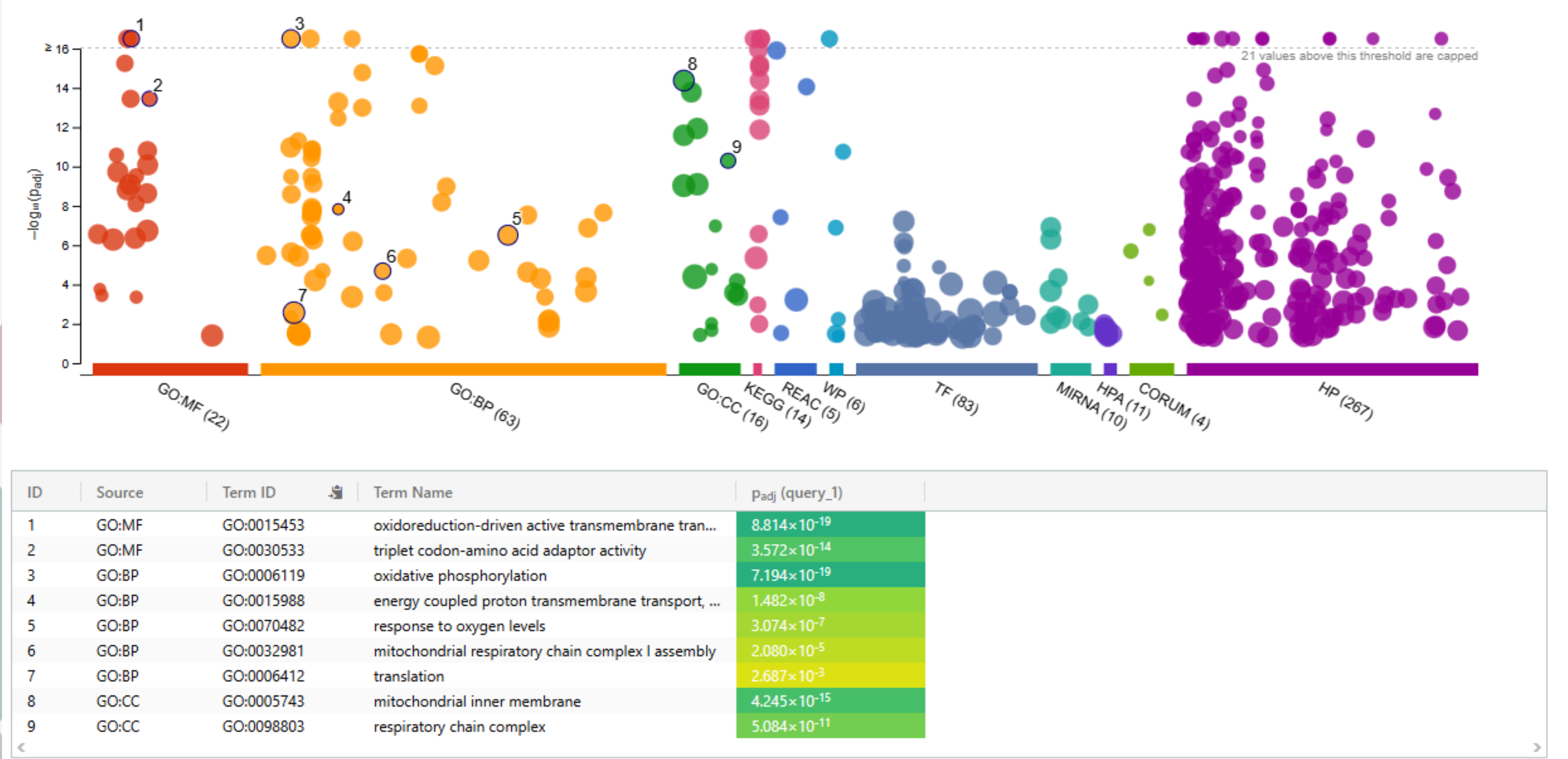
SRR32281629 (GATK)



ID	Source	Term ID	Term Name	padj (query_1)
1	GO:MF	GO:0015453	oxidoreduction-driven active transmembrane tran...	4.872×10 <sup>-19</sup>
2	GO:MF	GO:0030533	triplet codon-amino acid adaptor activity	2.342×10 <sup>-14</sup>
3	GO:BP	GO:0006119	oxidative phosphorylation	3.868×10 <sup>-19</sup>
4	GO:BP	GO:0015988	energy coupled proton transmembrane transport, ...	1.341×10 <sup>-8</sup>
5	GO:BP	GO:0070482	response to oxygen levels	7.109×10 <sup>-6</sup>
6	GO:BP	GO:0032981	mitochondrial respiratory chain complex I assembly	1.800×10 <sup>-5</sup>
7	GO:BP	GO:0006412	translation	2.050×10 <sup>-3</sup>
8	GO:CC	GO:0005743	mitochondrial inner membrane	1.425×10 <sup>-15</sup>
9	GO:CC	GO:0098803	respiratory chain complex	3.347×10 <sup>-11</sup>

# Enrichment analysis

SRR32281627 (GATK)





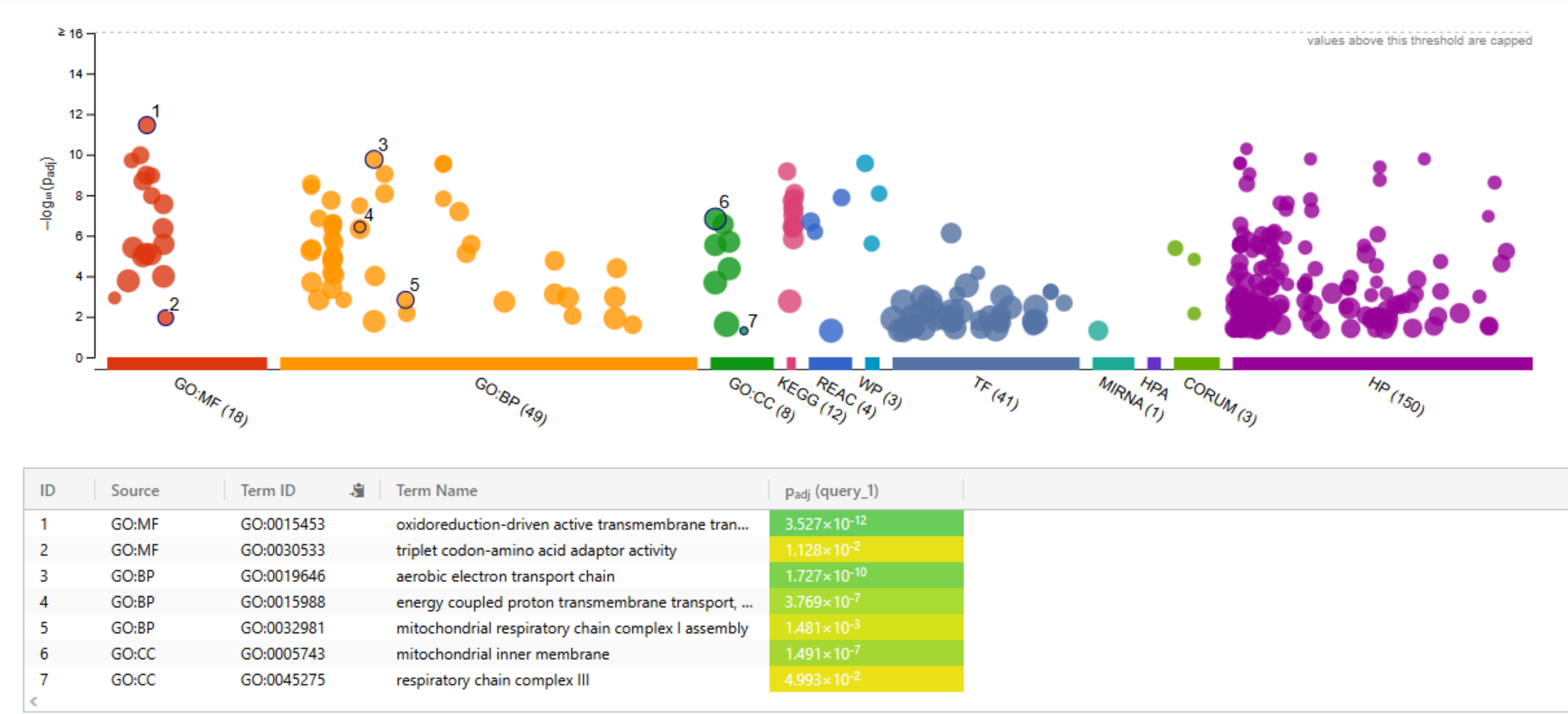
# Enrichment analysis

SRR32281629 (BCF)



# Enrichment analysis

SRR32281627 (BCF)



# Enrichment analysis

- Top 10 najbardziej wzbogaconych pozycji jest niemal identyczne między próbkami w GATK
- (GATK) Jedyna różnica jest na pozycji 4 ale prawdopodobnie dotyczy tego samego procesu biologicznego.
- Wiele z tych pozycji jest związanych z mitochondriami: transport, łańcuch oddechowy, błona mitochondrialna, co może wskazywać na konkretne mutacje, możliwe że któraś z nich jest celowo wprowadzona przez autorów.
- Najbardziej wzbogacony gen dotyczy aktywnego transportu transbłonowego (GO:0015453), pojawia we wszystkich analizach (obydwie próbki, obydwa programy), więc istnieje duże prawdopodobieństwo, że to jest celowa mutacja.
- BCFTools daje większe różnice między próbkami dla próbki \*7 p-wartości są znacznie niższe.
- Większość pozycji pochodzi z kategorii HP (Human Phenotype) – Próbki z człowieka