Marcel Buczkowski 145159

Asemblacja de novo

Dane:

Odczyty sparowane pochodzące z eksperymentu RNA-Seq (Homo sapiens)

- *inf394019/mapped.1.fastq*
- *inf394019/mapped.2.fastq*

Chromosom 22 człowieka jako genom (ze względu na czas działania programów konieczne było ograniczenie danych do 1 chromosomu; pliki mapped.1.fastq oraz mapped.2.fastq zawierają tylko odczyty mapujące się do tego chromosomu)

- *ftp://ftp.ensembl.org/pub/release-88/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.22.fa.gz*

Zadania:

Napisz oraz załącz do sprawozdania skrypt, który w sposób sekwencyjny wykona nastepujące kroki.

1) Przeprowadź kontrolę jakości odczytów przy użyciu FASTQC. Proszę skomentować wyniki kontroli jakości w sprawozdaniu.

```
marcelb@pandora:~/lab7> /home/tools/FastQC/fastqc mapped.1.fastq
```

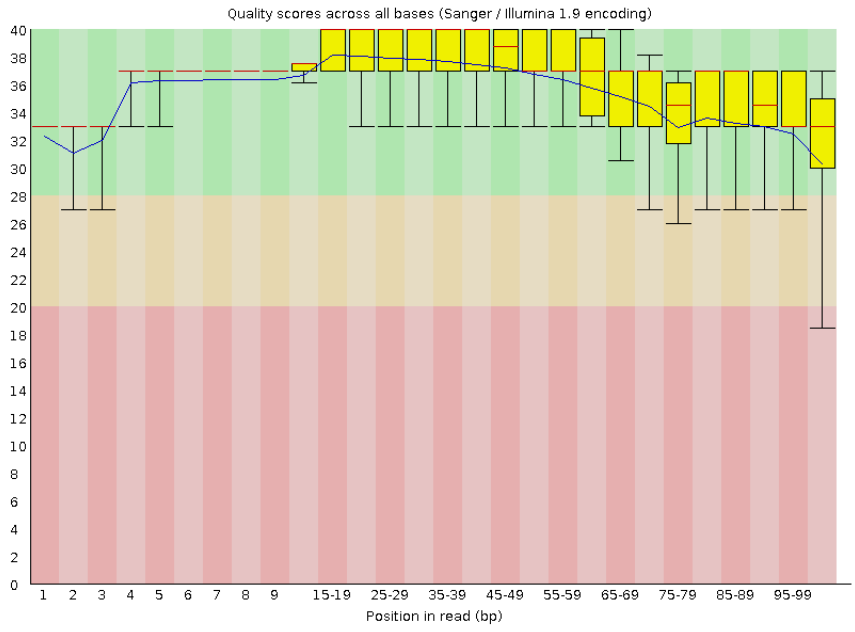
```
marcelb@pandora:~/lab7> /home/tools/FastQC/fastqc mapped.2.fastq
```

*skrypt będzie dołączony na końcu jako zbiór wszystkich komend

Basic Statistics

Measure	Value
Filename	mapped.1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	56006
Filtered Sequences	0
Sequence length	101
%GC	51

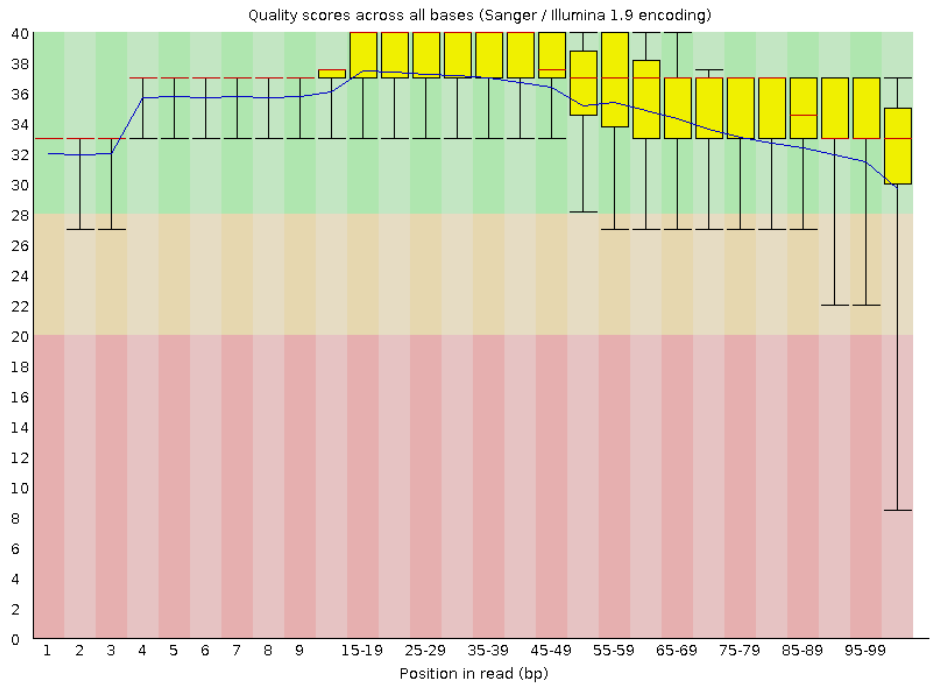
Per base sequence quality



Basic Statistics

Measure	Value
Filename	mapped.2.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	56006
Filtered Sequences	0
Sequence length	101
%GC	51

Per base sequence quality



Jakość odczytów z pierwszego pliku jest na ogół dobra, widać, że końce odczytów są niskiej jakości, ale zostanie to naprawione przy kolejnym kroku – odcinania adaptorów i końców o niskiej jakości.

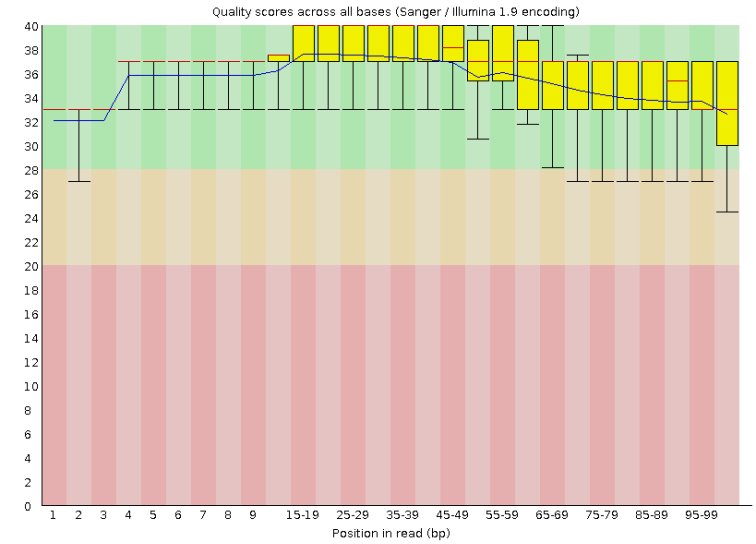
2) Usuń sekwencje adaptorów używanych przez sekwenatory Illumina HiSeq i MiSeq (TruSeq3-PE) , artefakty sekwencjonowania oraz końce o niskiej jakości. Użyj oprogramowania Trimmomatic (w katalogu tools). W sprawozdaniu proszę zamieścić raport programu, opis użytych parametrów oraz uzasadnić ich użycie

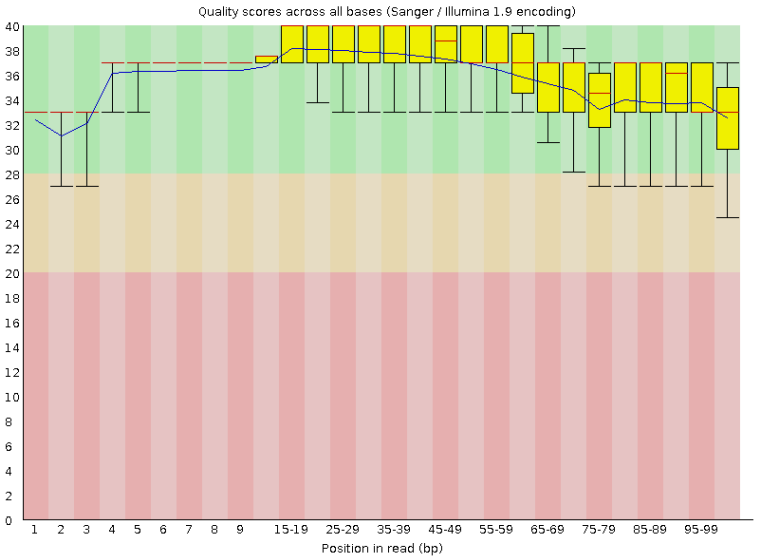
```
marcelb@pandora:~/lab7> java -jar /home/tools/Trimmomatic-0.36/trimmomatic-0.36.jar PE mapped.1.fastq mapped.2.fastq tri
mmed/paired_trimmed_mapped.1.fastq trimmed/unpaired_trimmed_mapped.1.fastq trimmed/paired_trimmed_mapped.2.fastq trimmed
/unpaired_trimmed_mapped.2.fastq ILLUMINACLIP:/home/tools/Trimmomatic-0.36/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAI
LING:3
TrimmomaticPE: Started with arguments:
 mapped.1.fastq mapped.2.fastq trimmed/paired_trimmed_mapped.1.fastq trimmed/unpaired_trimmed_mapped.1.fastq trimmed/pai
red_trimmed_mapped.2.fastq trimmed/unpaired_trimmed_mapped.2.fastq ILLUMINACLIP:/home/tools/Trimmomatic-0.36/adapters/Tr
uSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3
Using PrefixPair: 'TACACTCTTTCCCTACACGACGCTCTTCCGATCT' and 'GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT'
ILLUMINACLIP: Using 1 prefix pairs, 0 forward/reverse sequences, 0 forward only sequences, 0 reverse only sequences
Quality encoding detected as phred33
Input Read Pairs: 56006 Both Surviving: 55832 (99.69%) Forward Only Surviving: 174 (0.31%) Reverse Only Surviving: 0 (0.
00%) Dropped: 0 (0.00%)
TrimmomaticPE: Completed successfully
```

Parametr PE oznacza Paired ends – trzeba go dodać gdyż domyślnie program działa na niesparowanych odczytach, a my pracujemy ze sparowanymi. Parametr ILLUMINACLIP oznacza jakie dokładnie adaptory mamy odcinać. LEADING - usuwanie przednich nukleotydów o niskiej jakości (poniżej 3). TRAILING – usuwanie tylnich nukleotydów o niskiej jakości.

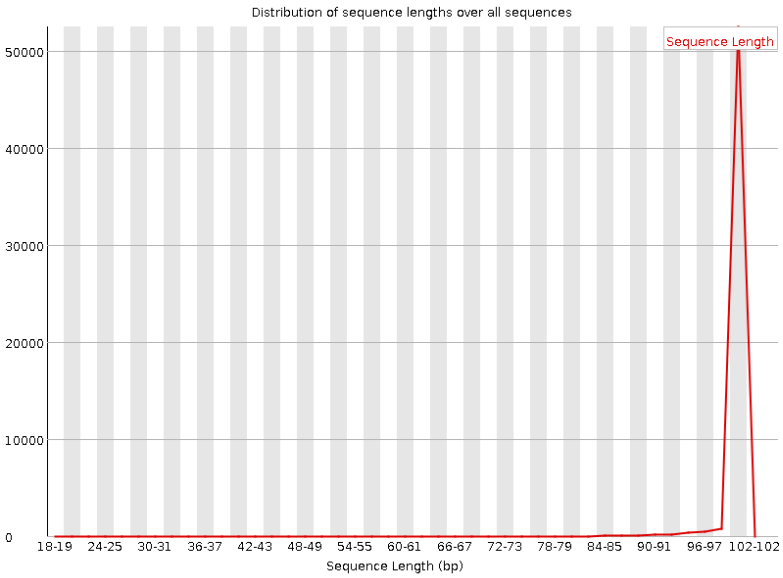
Marcel Buczkowski 145159

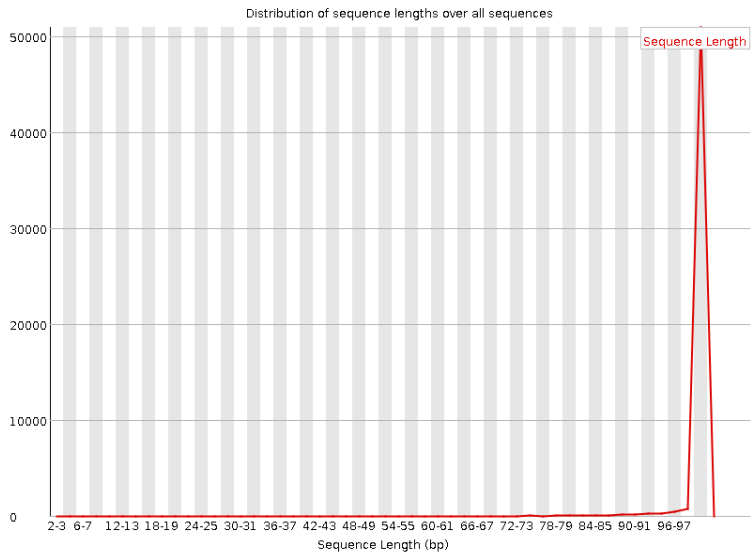
3) Opcjonalnie: ponownie przeprowadź kontrolę jakości odczytów przy użyciu FASTQC. Proszę skomentować wyniki kontroli jakości w sprawozdaniu.





Widać, że na końcach usunięto nukleotydy o słabej jakości. Jednak z jakiegoś powodu adaptery nie zostały odcięte i nie mogę namierzyć tego błędu.
Widać, że długość odczytów nie uległa skróceniu:





4) Uruchom oprogramowanie Trinity w celu złożenia transkryptomu człowieka de novo. Zapisz czas działania programu (polecenie *time*). Przeanalizuj statystyki otrzymanego złożenia przy pomocy skryptu TrinityStats.pl (w katalogu util). Zamieść i skomentuj wyniki w sprawozdaniu.

```
marcelb@pandora:~/lab7> sed -i 's/ //g' trimmed/paired_trimmed_mapped.1.fastq
marcelb@pandora:~/lab7> sed -i 's/ //g' trimmed/paired_trimmed_mapped.2.fastq
marcelb@pandora:~/lab7> time /home/tools/trinityrnaseq-Trinity-v2.4.0/Trinity --seqType fq --max_memory 50G --left trimm
ed/paired_trimmed_mapped.1.fastq --right trimmed/paired_trimmed_mapped.2.fastq --CPU 6
```

```
########################################################################
Butterfly assemblies are written to /home/marcelb/lab7/trinity_out_dir/Trinity.fasta
########################################################################



real    4m44.205s
user    30m47.616s
sys     4m58.788s
marcelb@pandora:~/lab7>
```

```
marcelb@pandora:~/lab7> /home/tools/trinityrnaseq-Trinity-v2.4.0/util/TrinityStats.pl trinity_out_dir/Trinity.fasta > trinityStats_first.log
marcelb@pandora:~/lab7> ls
mapped.1.fastq   mapped.1_fastqc.zip  mapped.2_fastqc      pipeline.sh  trinityStats_first.log
mapped.1_fastqc  mapped.2.fastq       mapped.2_fastqc.zip  trimmed      trinity_out_dir
marcelb@pandora:~/lab7> tri
trimmed/         trinity_out_dir/
marcelb@pandora:~/lab7> cat trinityStats_first.log


################################
## Counts of transcripts, etc.
################################
Total trinity 'genes':  875
Total trinity transcripts:      968
Percent GC: 52.74

########################################
Stats based on ALL transcript contigs:
########################################

        Contig N10: 4092
        Contig N20: 2925
        Contig N30: 2078
        Contig N40: 1602
        Contig N50: 1221

        Median contig length: 394
        Average contig: 725.89
        Total assembled bases: 702666


################################################
## Stats based on ONLY LONGEST ISOFORM per 'GENE':
################################################

        Contig N10: 4102
        Contig N20: 2600
        Contig N30: 1942
        Contig N40: 1494
        Contig N50: 1099

        Median contig length: 387
        Average contig: 690.34
        Total assembled bases: 604047
```

Widać, że po złożeniu transkryptomu człowieka (22 chromosom) za pomocą programu Trinity wykryto 875 genów a 968 transkryptów. Patrząc na to ile jest genów na tym chromosomie wydaje się to być dobra predykcja:

Estimated by	Protein-coding genes	Non-coding RNA genes	Pseudogenes	Source	Release date
CCDS	417	—	—	[1]	2016-09-08
HGNC	424	161	295	[6]	2019-07-08
Ensembl	489	515	325	[7]	2017-03-29
UniProt	496	—	—	[8]	2018-02-28
NCBI	474	392	379	[9][10][11]	2017-05-19

Widzimy też ze średnia długość kontigu to 394 oraz mamy podane statystyki N10-50. Spośród nich najbardziej interesująca jest N50, która oznacza, że 50% asemblowanego genomu jest pokryta przez contigi co najmniej tak duże (długie) jak N50 – czyli w tym wypadku 1221 bp.
Poniżej również jest informacja dotycząca tylko jednego izoformu dla danego genu, czyli są to statystyki bez ‘powtórzonych’ genów tj. różnych transkryptów tego samego genu.

może okazać się usunięcie spacji z plików fastq

5) W celu złożenia transkryptomu przy użyciu genomu referencyjnego należy:
- utworzyć indeks genomu przy użyciu programu STAR
- zmapować odczyty do genomu przy użyciu programu STAR
- przeprowadzić konwersje otrzymanego pliku sam do bam przy użyciu pakietu samtools
- posortować plik bam przy użyciu pakietu samtools
- uruchomić oprogramowanie Trinity w trybie genome-guided. Zapisać czas działania programu.
- Przeanalizować statystyki otrzymanego złożenia przy pomocy TrinityStats.pl. Zamieścić i skomentować wyniki w sprawozdaniu.

a. Utworzenie indeksu genomu:

```
marcelb@pandora:~/lab7> /home/tools/STAR-2.5.3a/bin/Linux_x86_64/STAR --runMode genomeGenerate --genomeDir chr22_index --genomeFastaFiles ./Homo_sapiens.GRCh38.dna.chromosome.22.fa
```

b. zmapować odczyty do genomu przy użyciu programu STAR

*możliwy jest od razu zapis do SAM oraz posortowanie za pomocą narzędzia STAR:

```
marcelb@pandora:~/lab7> /home/tools/STAR-2.5.3a/bin/Linux_x86_64/STAR --genomeDir chr22_index --readFilesIn ./trimmed/paired_trimmed_mapped.1.fastq ./trimmed/paired_trimmed_mapped.2.fastq --outSAMtype BAM SortedByCoordinate
```

```
Jan 23 22:03:44 ..... started STAR run
Jan 23 22:03:44 ..... loading genome
Jan 23 22:03:45 ..... started mapping
Jan 23 22:03:55 ..... started sorting BAM
Jan 23 22:03:56 ..... finished successfully
```

c. Uruchomienie programu Trinity do asemblacji genomu w trybie genome guided

```
marcelb@pandora:~/lab7> time /home/tools/trinityrnaseq-Trinity-v2.4.0/Trinity --genome_guided_bam Aligned.sorted
ByCoord.out.bam --max_memory 50G --CPU 6 --genome_guided_max_intron 10000
```

```
real    5m43.063s
user    24m29.381s
sys     2m23.151s
```

d. analiza wynikow programu Trinity

```
marcelb@pandora:~/lab7> /home/tools/trinityrnaseq-Trinity-v2.4.0/util/TrinityStats.pl trinity_out_dir/Trinity-GG.fasta > trinitySta
ts_second.log
```

```
marcelb@pandora:~/lab7> cat trinityStats_second.log


################################
## Counts of transcripts, etc.
################################
Total trinity 'genes':  966
Total trinity transcripts:      1004
Percent GC: 52.41

########################################
Stats based on ALL transcript contigs:
########################################

        Contig N10: 4080
        Contig N20: 2543
        Contig N30: 1942
        Contig N40: 1471
        Contig N50: 1075

        Median contig length: 374
        Average contig: 669.23
        Total assembled bases: 671910


####################################################
## Stats based on ONLY LONGEST ISOFORM per 'GENE':
####################################################

        Contig N10: 3782
        Contig N20: 2395
        Contig N30: 1892
        Contig N40: 1401
        Contig N50: 1020

        Median contig length: 371.5
        Average contig: 653.27
        Total assembled bases: 631056
```

Otrzymano nieco więcej genów i transkryptów niż w przypadku asemblacji de novo. Contigi tutaj są nieco krótsze na ogół niż w przypadku asemblacji de novo. Statystyka N50 również jest niższa.
**Należy pamiętać o adapterach, które jeśli nie zostały usunięte na pewno miały wpływ na pozostały przebieg analizy.