POST-SEQUENCING ANALYSES

ROTIMI DADA

dadarotimi@hotmail.com

Public Health Alliance for Genomic Epidemiology

Why Quality Control

- Quality control in sequencing is very crucial as it for any laboratory test
- It is the very first step that determines if other downstream analysis can go on or not e.g. AMR, MLST calling, e.t.c.
- The strength of data generated from sequencing is largely dependent on the quality of the entire process

Pre-sequencing quality control

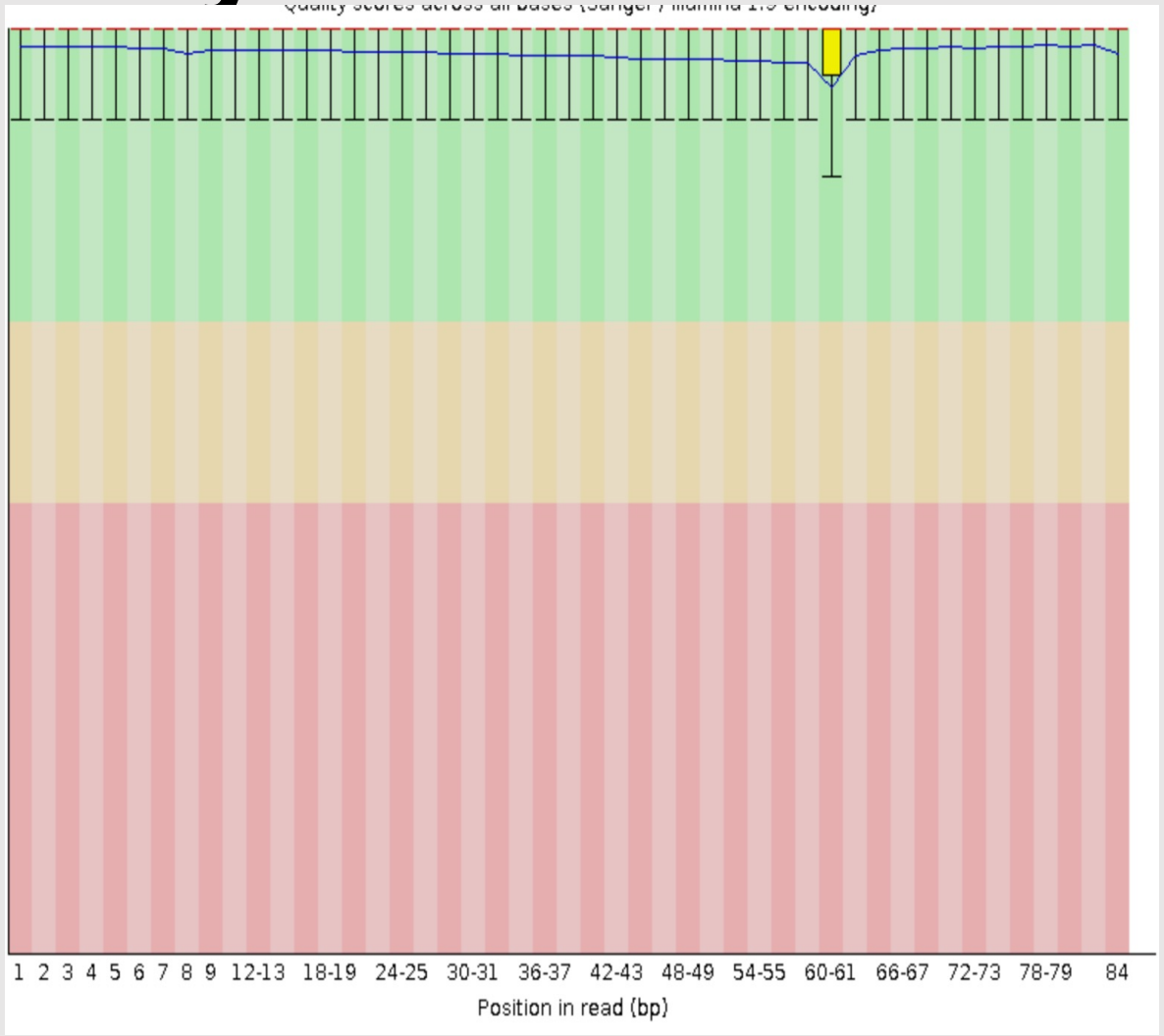
- Sample quality control
  - Isolation and identification of organism(s)
  - Method used for Nucleic Acid Extraction
  - Contamination
  - Elution of nuclear material
  - Quantification of Nucleic Acid

- Library preparation
  - Fragmentation of Nucleic Acid to to optimal fragment length
  - Barcoding
  - Adapter ligation
  - PCR to generate clusters (depending on sequencing technology used
- During sequencing
  - Sequencing Analysis Viewer (SAV)
    - An Illumina tool used to monitor QC during a run
    - Can also be used to check QC after a run

Public Health Alliance for Genomic Epidemiology

Post-Sequencing Quality Control – FastQC

- FastQC is a tool used for providing an overview of basic quality metrics for raw next generation sequencing data (reads)
- Quality metrics are reported in a traffic light manner
  - RED=BAD;
  - ORANGE=WARNING;
  - GREEN=GOOD



Public Health Alliance for Genomic Epidemiology

Quality metrics used in fastqc

FastQC Report

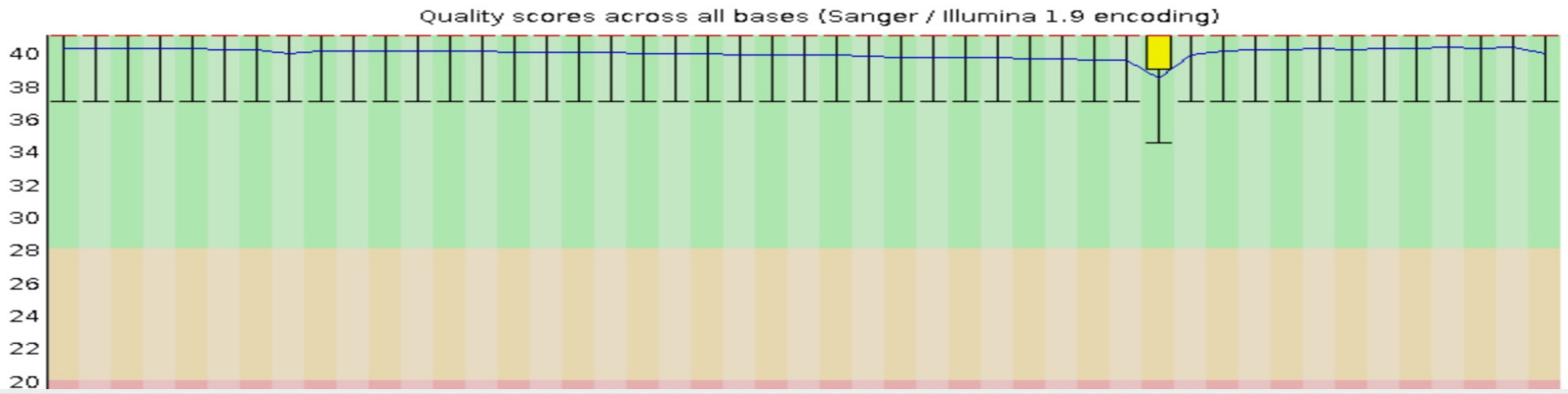
Summary

- Basic Statistics
- Per base 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

Basic Statistics

Measure	Value
Filename	SRR6362403_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2762941
Sequences flagged as poor quality	0
Sequence length	63-84
%GC	51

Per base sequence quality



Public Health Alliance for Genomic Epidemiology

Basic statistics

Name of sequence

Type of quality score encoding

Total number of reads

Sequences flagged as poor quality

Read length and

GC content.

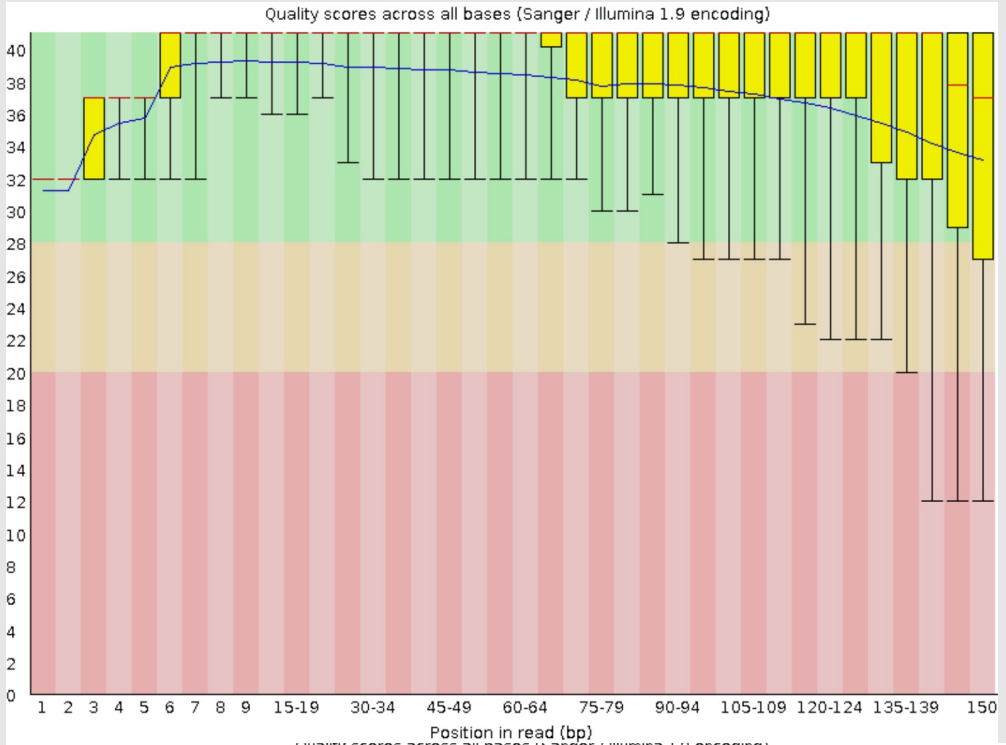
Basic Statistics

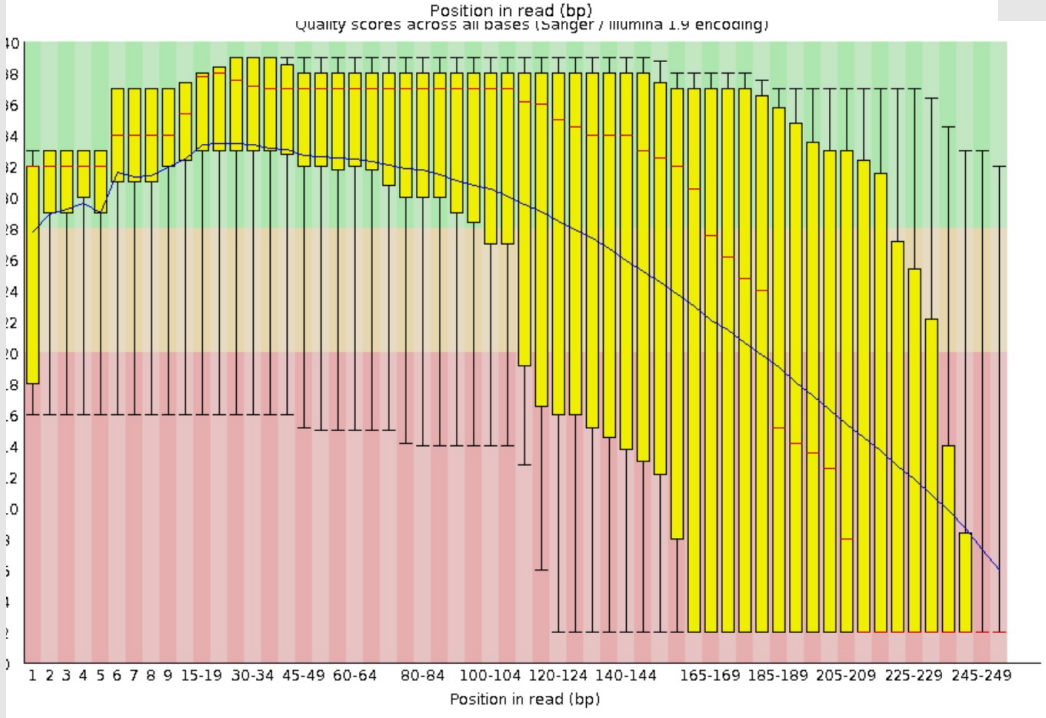
Measure	Value
Filename	SRR6362403_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2762941
Sequences flagged as poor quality	0
Sequence length	63-84
%GC	51

Public Health Alliance for Genomic Epidemiology

Per base sequence quality

- Looks at quality scores at each position along all reads in a fastq file
- Reports quality score in a box-and-whisker plot
- Upper whiskers(green) represent 10th percentile scores
- Medium whiskers (yellow) represent 25th percentile scores
- Lower whiskers (red) represents 90th percentile scores



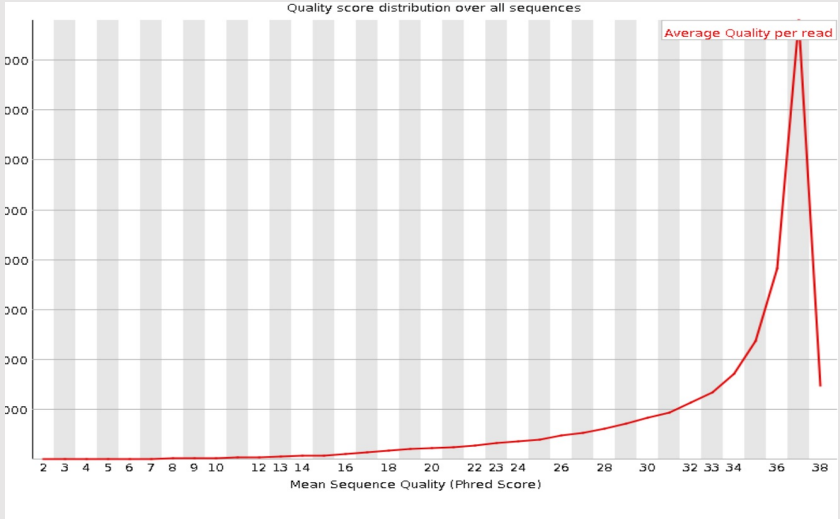


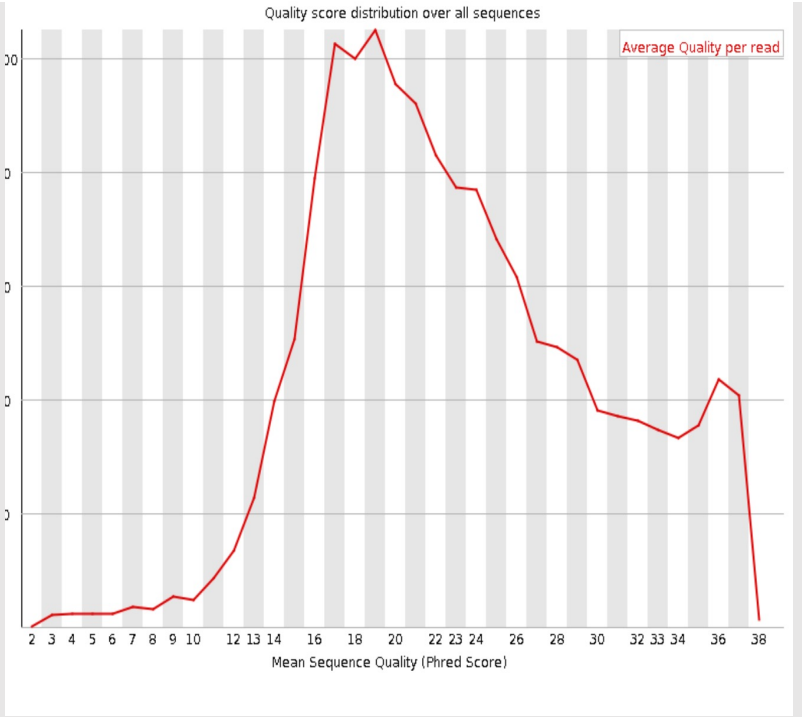
Public Health Alliance for Genomic Epidemiology

per sequence quality scores

- Plots the total number of reads against average quality score (phred score) over the whole length of reads

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

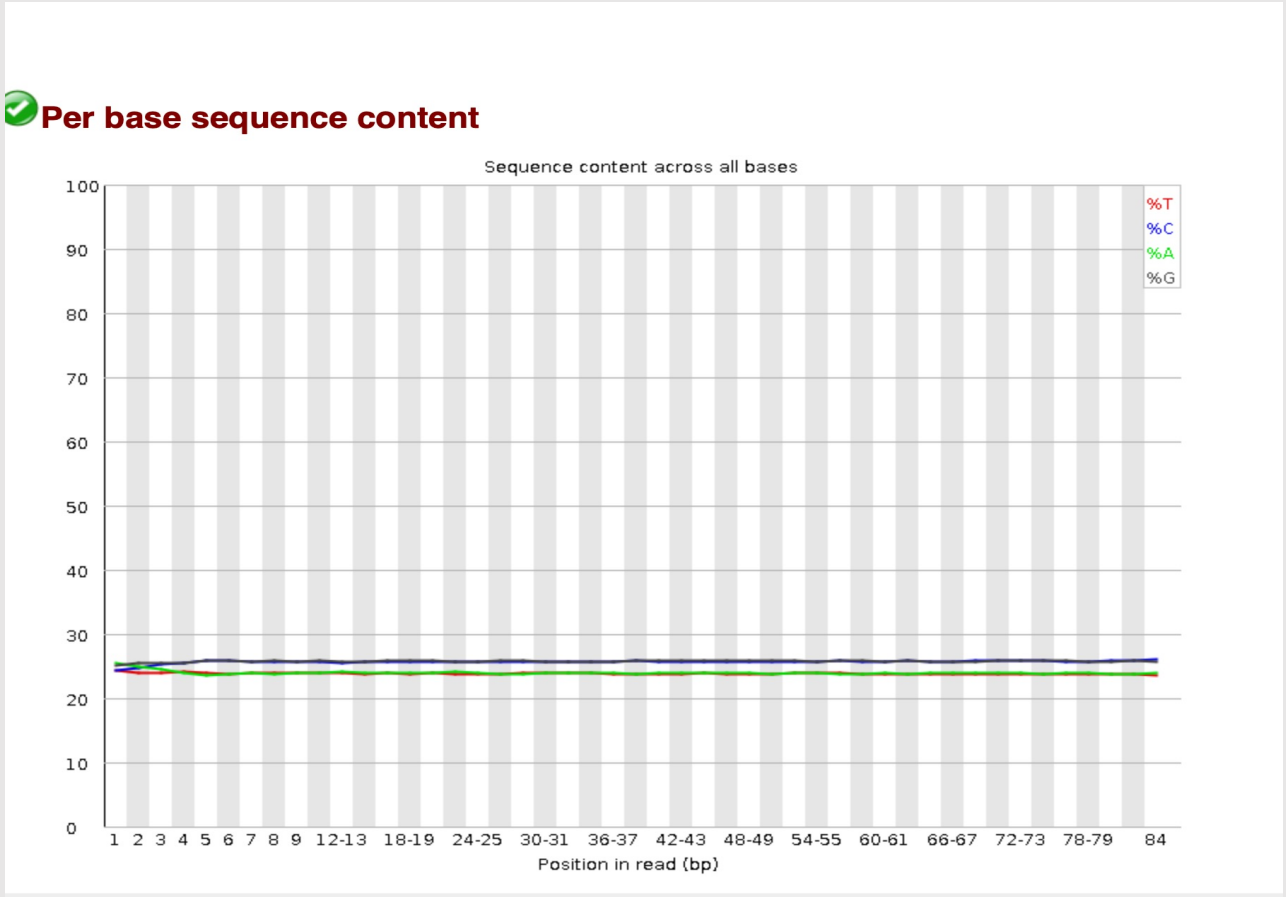




Public Health Alliance for Genomic Epidemiology

Per base sequence content

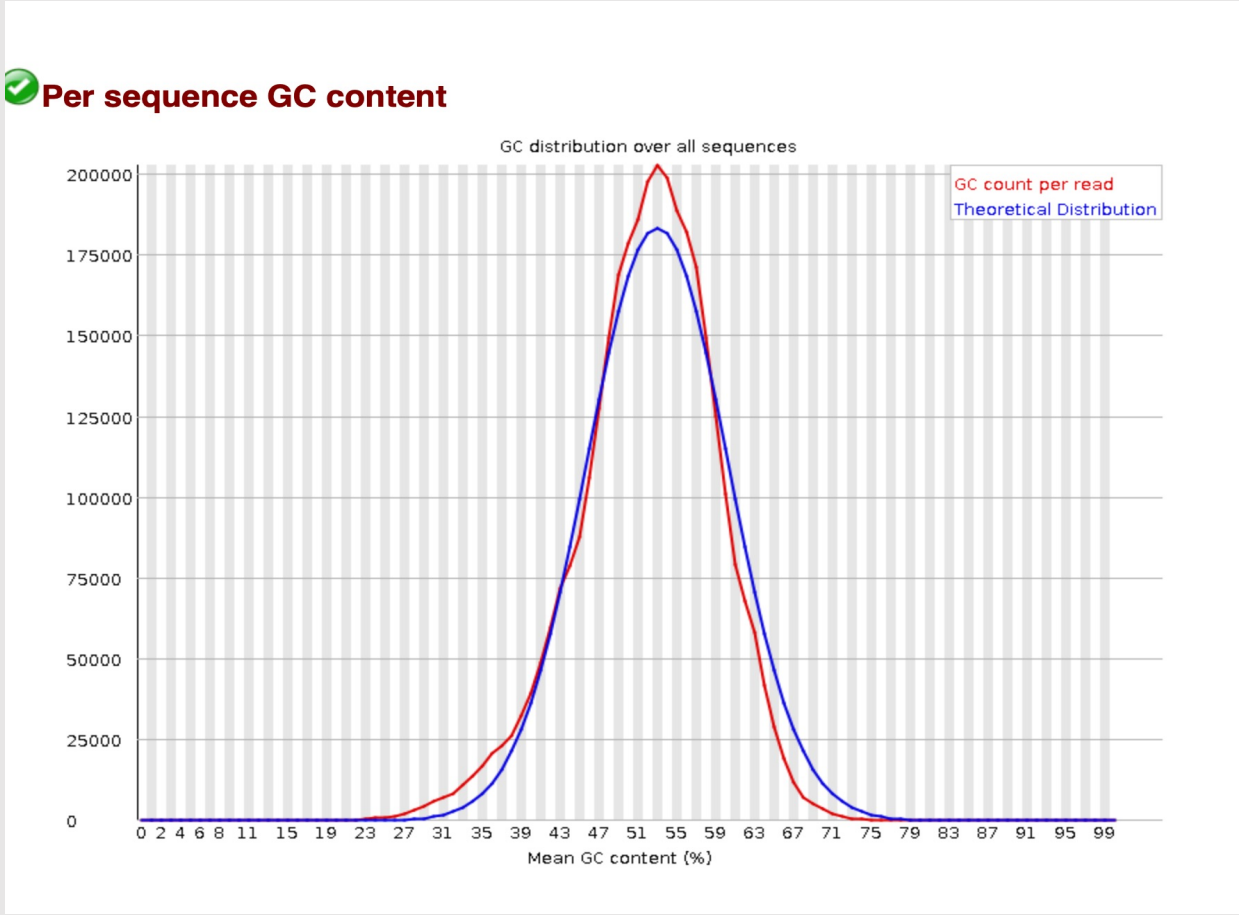
- Here you have reports of the percentage of bases called for the four nucleotides
- The A, C, G, T content is measured across all reads in a file
- The proportion of each of the bases should be fairly constant over the read length



Public Health Alliance for Genomic Epidemiology

Per sequence gc content

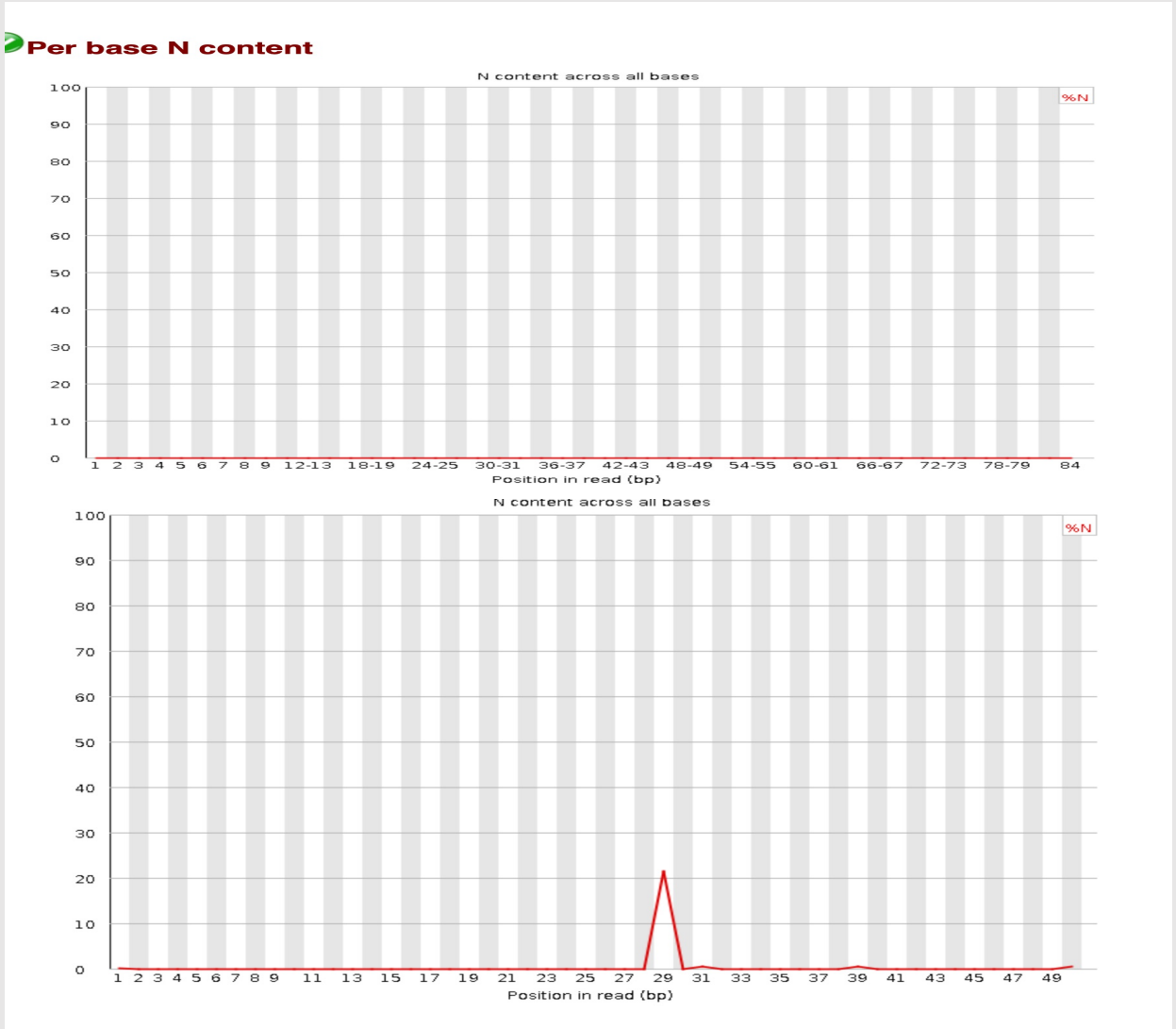
- Number of reads is plotted against GC percentage per read
- Ideally, GC content of all reads should form a normal distribution curve
- Theoretical distribution which assumes a uniform GC content for reads is plotted in blue, while the GC content per read is plotted in red.



Public Health Alliance for Genomic Epidemiology

Per base n content

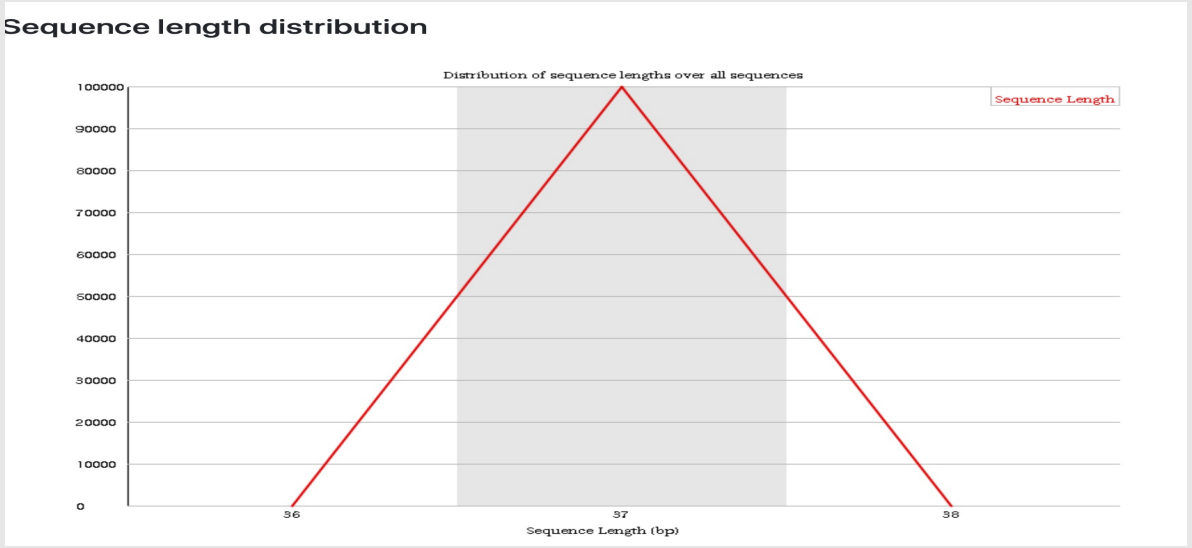
- This describe percentage of bases at each position with no base call.
- Ideally bases should be called at each position across the read length
- When no base is called at any position, a problem must have occurred during the sequencing run.

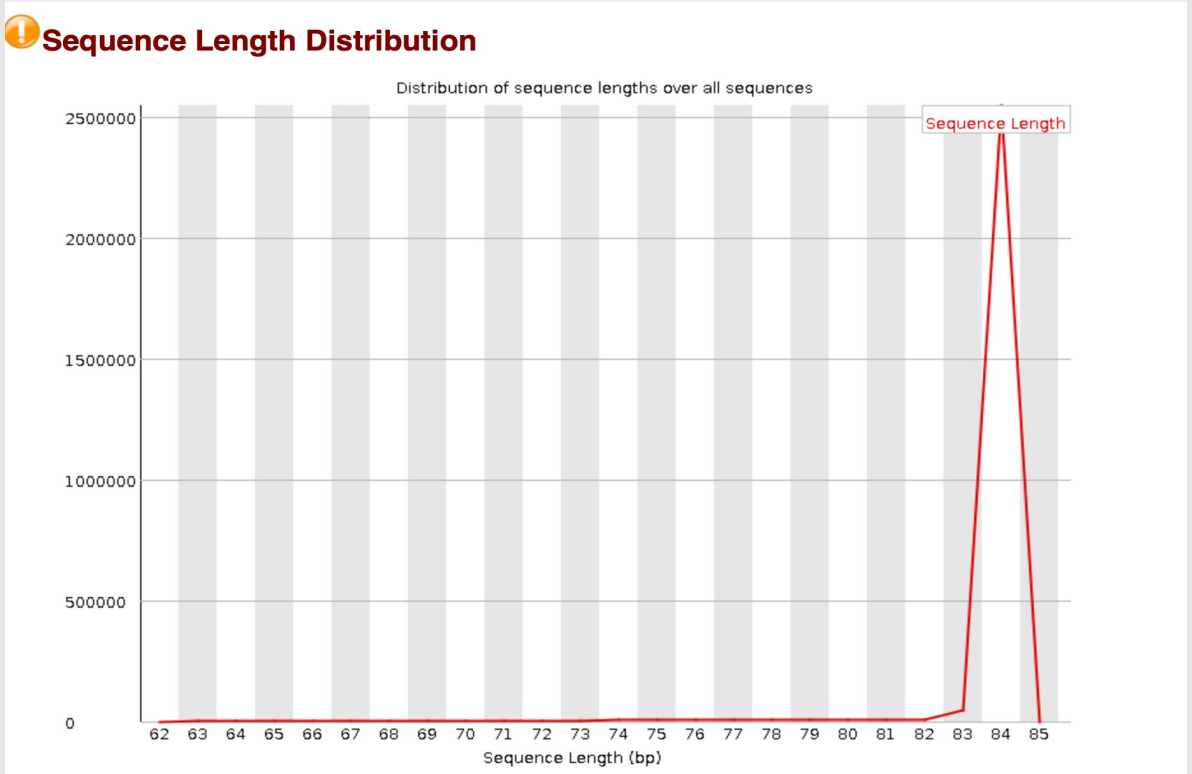


Public Health Alliance for Genomic Epidemiology

Sequence length distribution

- Some sequencing platforms generate sequence fragments of uniform length
- Some others generate reads of widely varying length
- A peak at one side indicates that the reads are of uniform length

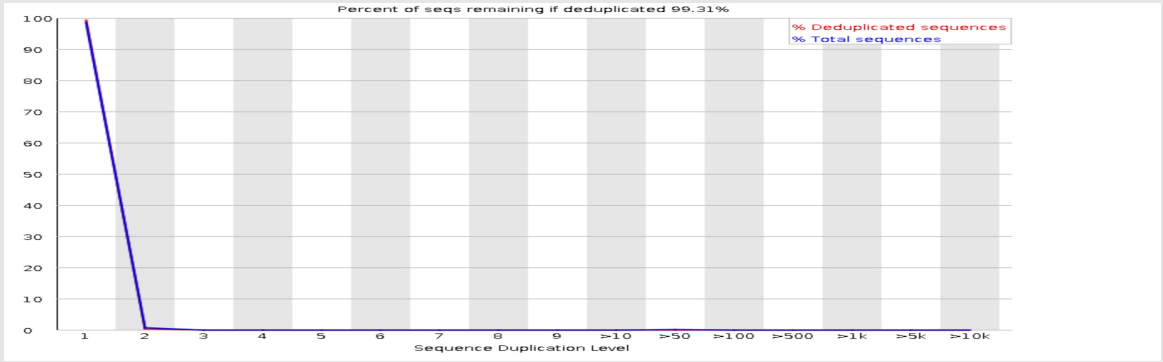


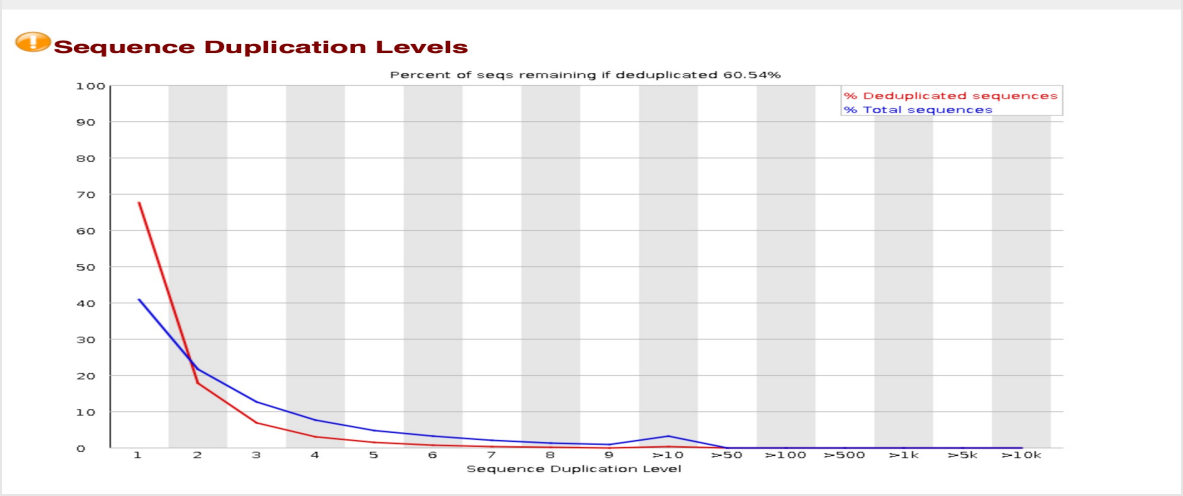


Public Health Alliance for Genomic Epidemiology

Sequence duplication level/over-represented sequences

- There are to main sources of duplicate reads
  - Biased PCR enrichment
  - Truly over-represented reads e.g. transcripts in an RNA-Seq library
- Usually, nearly 100% of your reads should be unique
- There are exceptional situations when sequencing is very deep (> 100x) and some level of duplication is inevitable

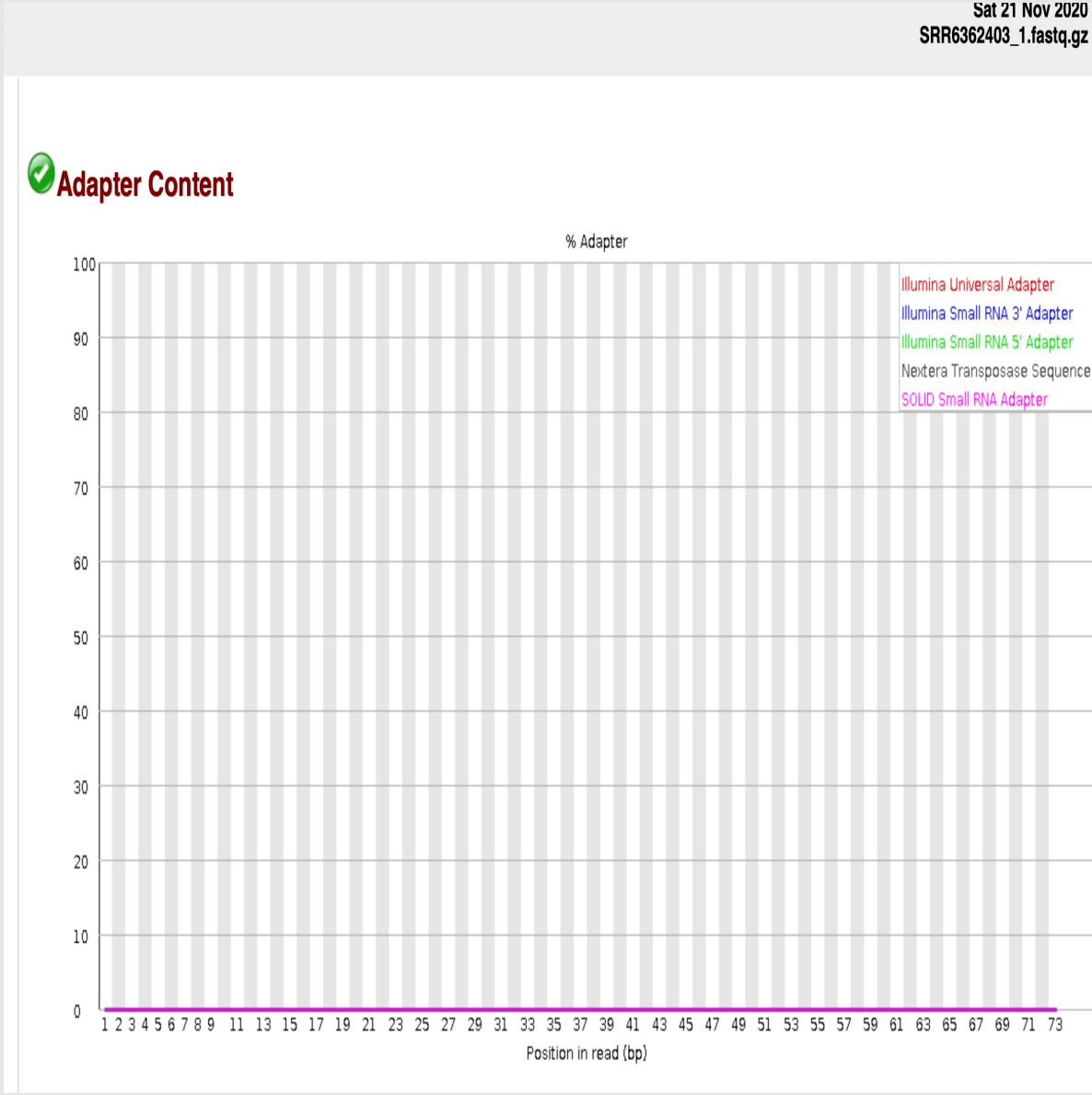




Public Health Alliance for Genomic Epidemiology

Adapter content

- Ideally, your sequence shouldn't have any adapter sequence present
- Once an adapter is detected, it is counted as being present throughout the read length
- Only adapters specific to the library type are searched



Public Health Alliance for Genomic Epidemiology

Multiqc

- Takes input from fastqc output
- Aggregates several reports into one
- Gives a snapshot of different quality metrics for all reads

General Statistics

Copy table | Configure Columns | Plot | Showing 146/146 rows and 4/5 columns.

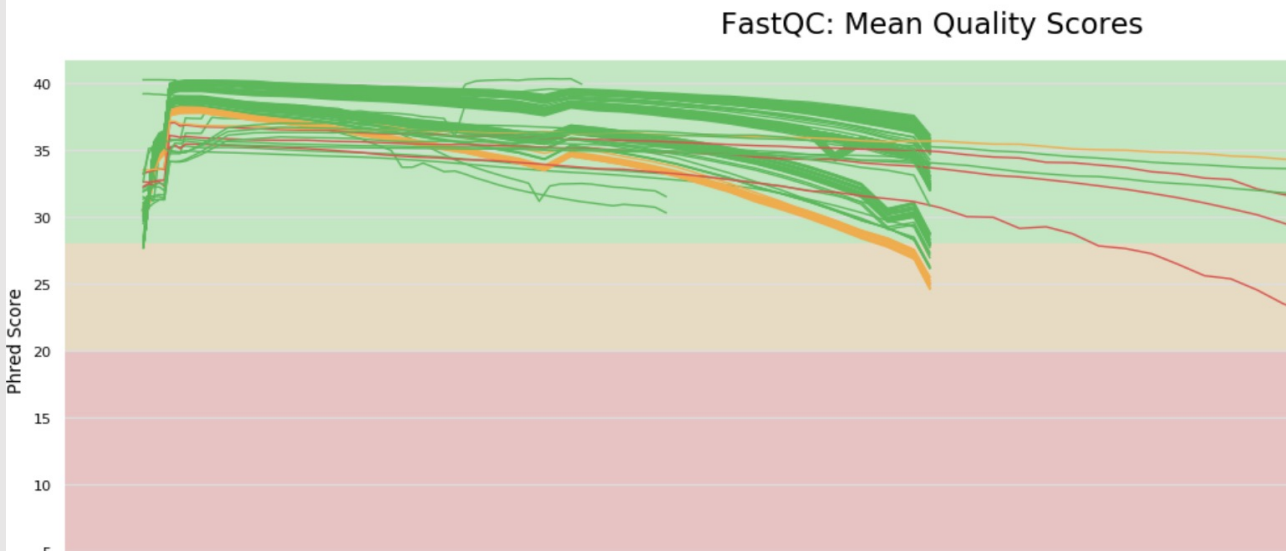
Sample Name	% Dups	% GC	Length	M Seqs
ERR4008003_1	30.3%	58%	149 bp	1.9
ERR4008003_2	29.6%	58%	149 bp	1.9
ERR4046281_1	7.5%	57%	151 bp	1.3
ERR4046281_2	6.2%	57%	151 bp	1.3
ERR4046327_1	11.2%	51%	151 bp	1.4
ERR4046327_2	9.3%	51%	151 bp	1.4
ERR987377_1	5.5%	50%	100 bp	2.0
ERR987377_2	5.0%	50%	100 bp	2.0

Sequence Quality Histograms

128

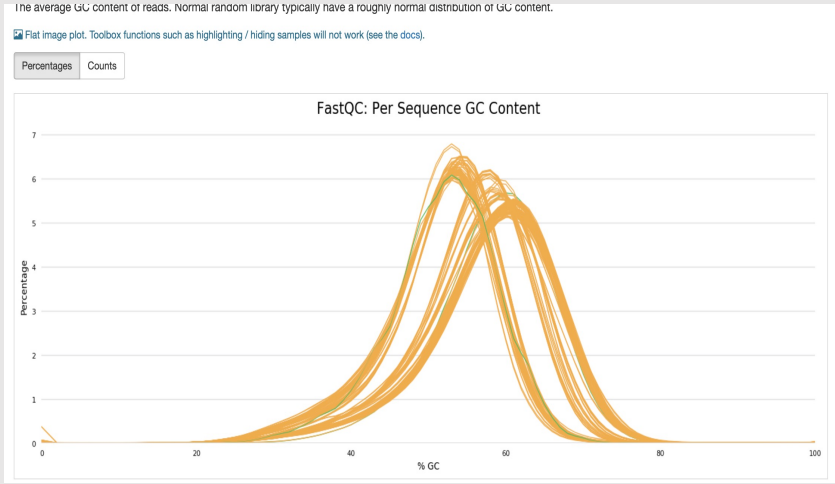
he mean quality value across each base position in the read.

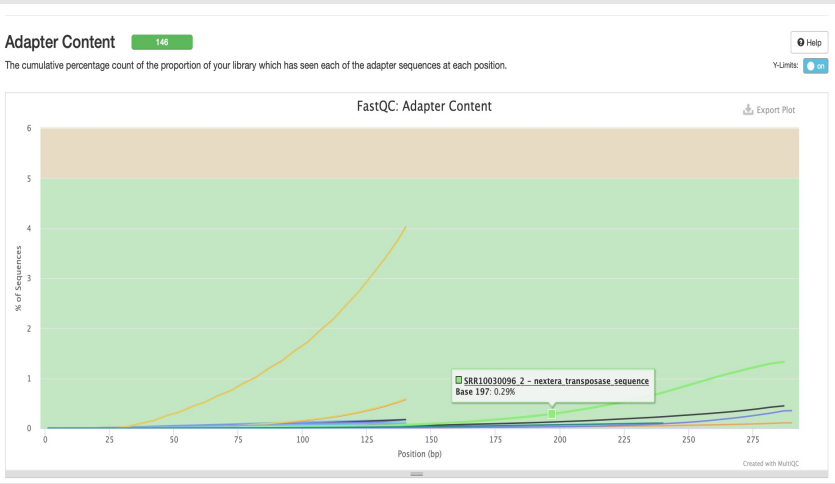
Flat image plot. Toolbox functions such as highlighting / hiding samples will not work (see the docs).

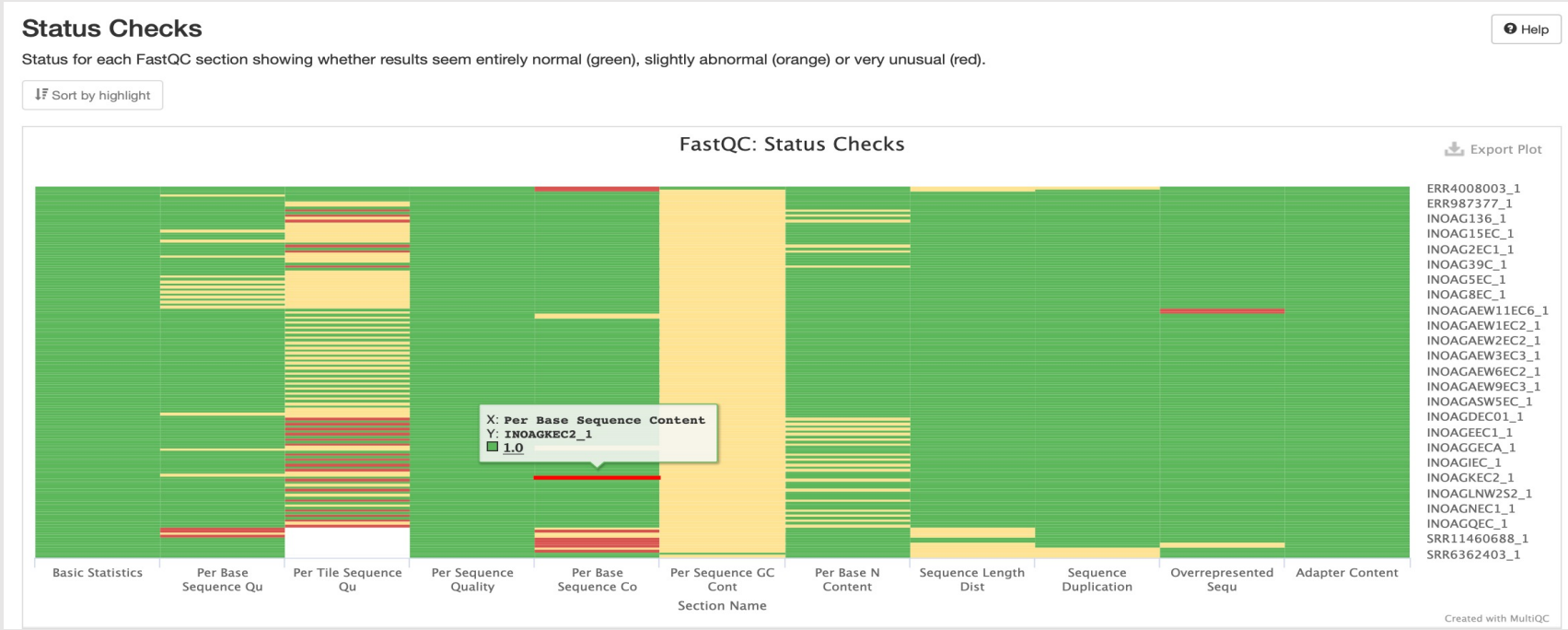


Public Health Alliance for Genomic Epidemiology

- Basically, the quality reports are same as you have in FastQC
- You however have some other visualisation and statistics to help you understand the overall quality statuses of your reads



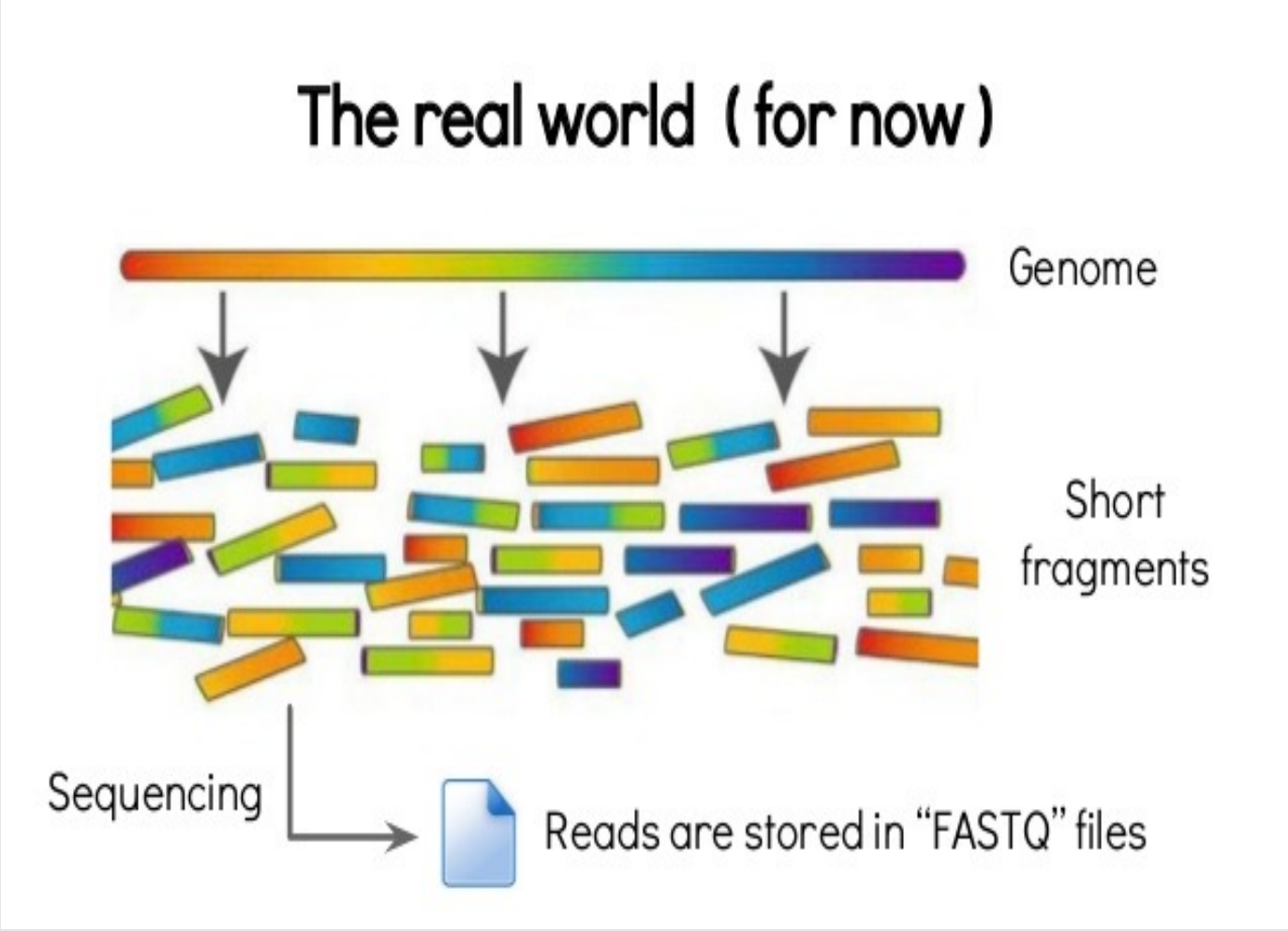




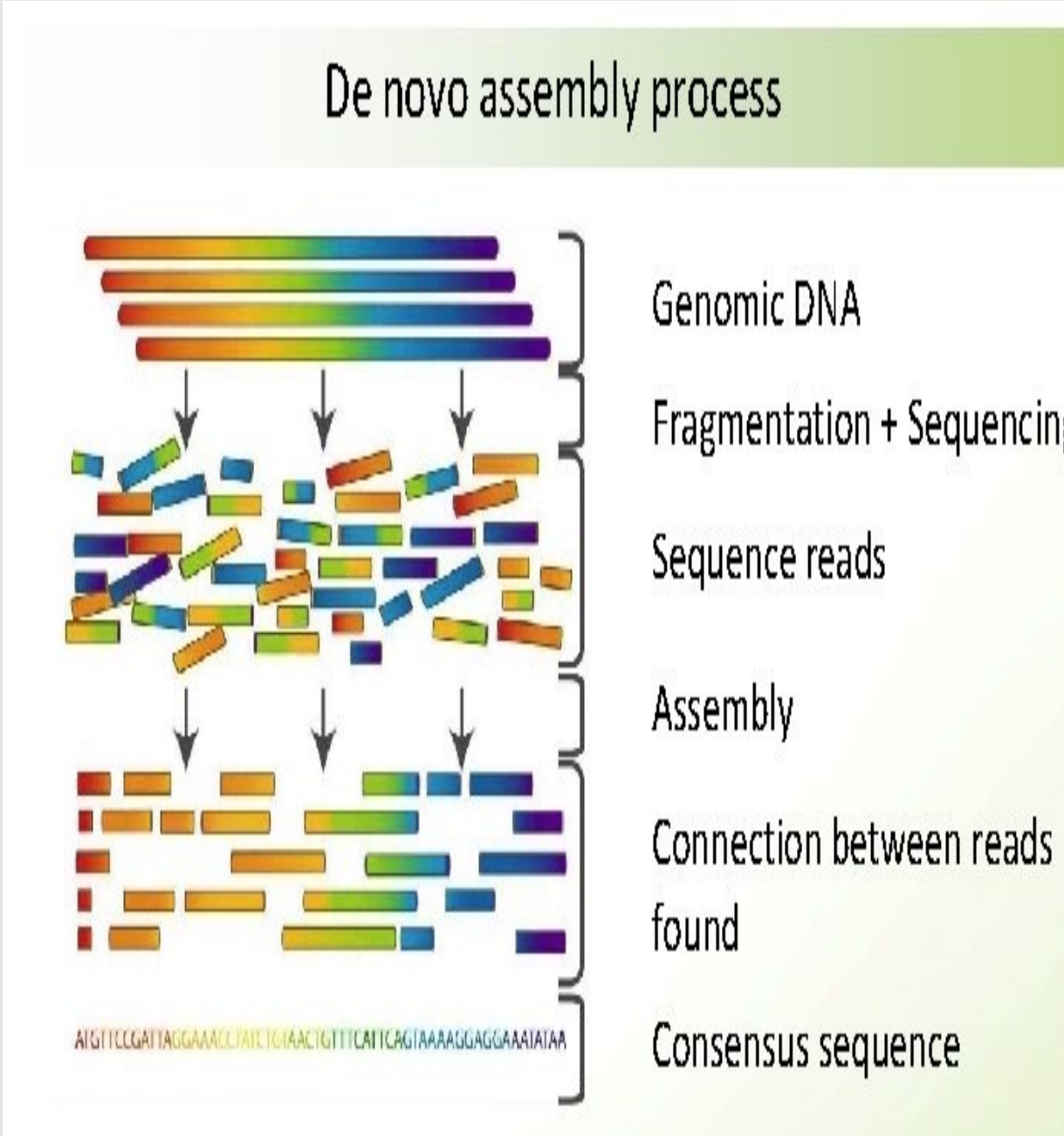
Public Health Alliance for Genomic Epidemiology

Assembling raw reads

- Why assemble?
  - Remember we can't sequence without 'cutting' the DNA
  - Sequencing also took place in 'bits and pieces"
  - Insert size is largely dependent on library preparation, which is also dependent on sequencing platform/technology

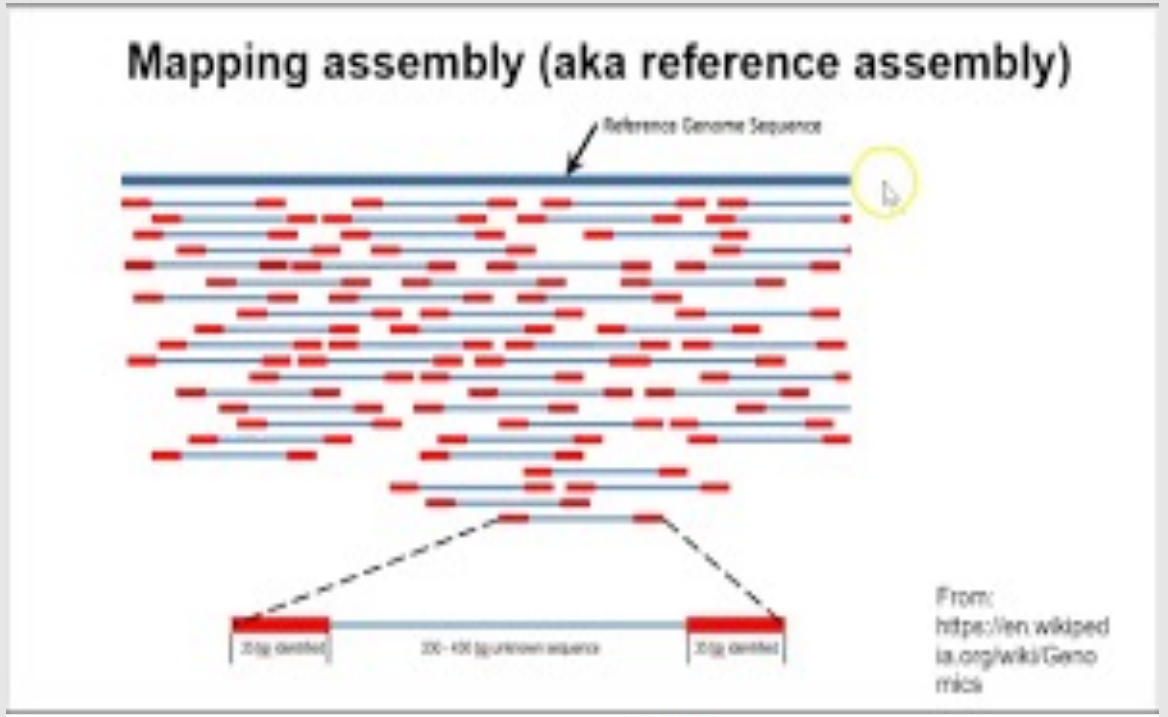


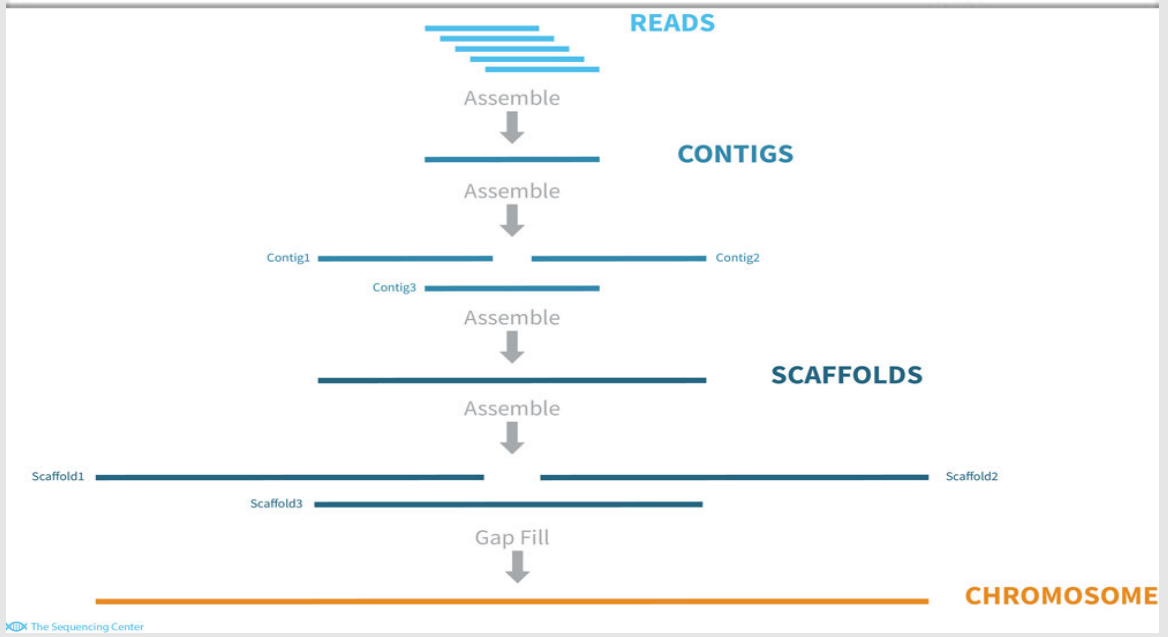
- Arranging reads into chromosomes or plasmids is a huge task
- It is basically the process of determining how reads fit together by looking for overlaps between them
- Data from various platforms yield relatively short stretches



Public Health Alliance for Genomic Epidemiology

- There are two types of sequence assembly protocols
  - Reference-based sequence assembly
  - De-novo sequence assembly
- Quality of reference-based assembly is dependent on the quality of reference used

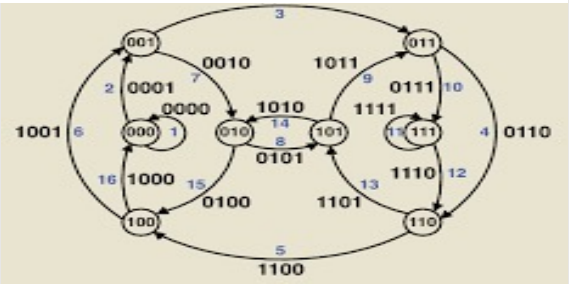


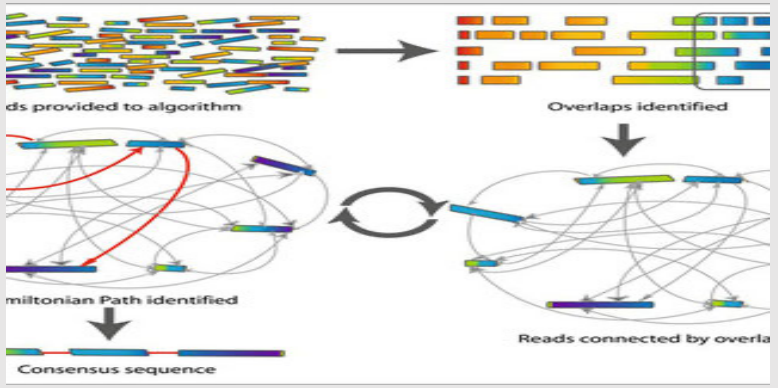


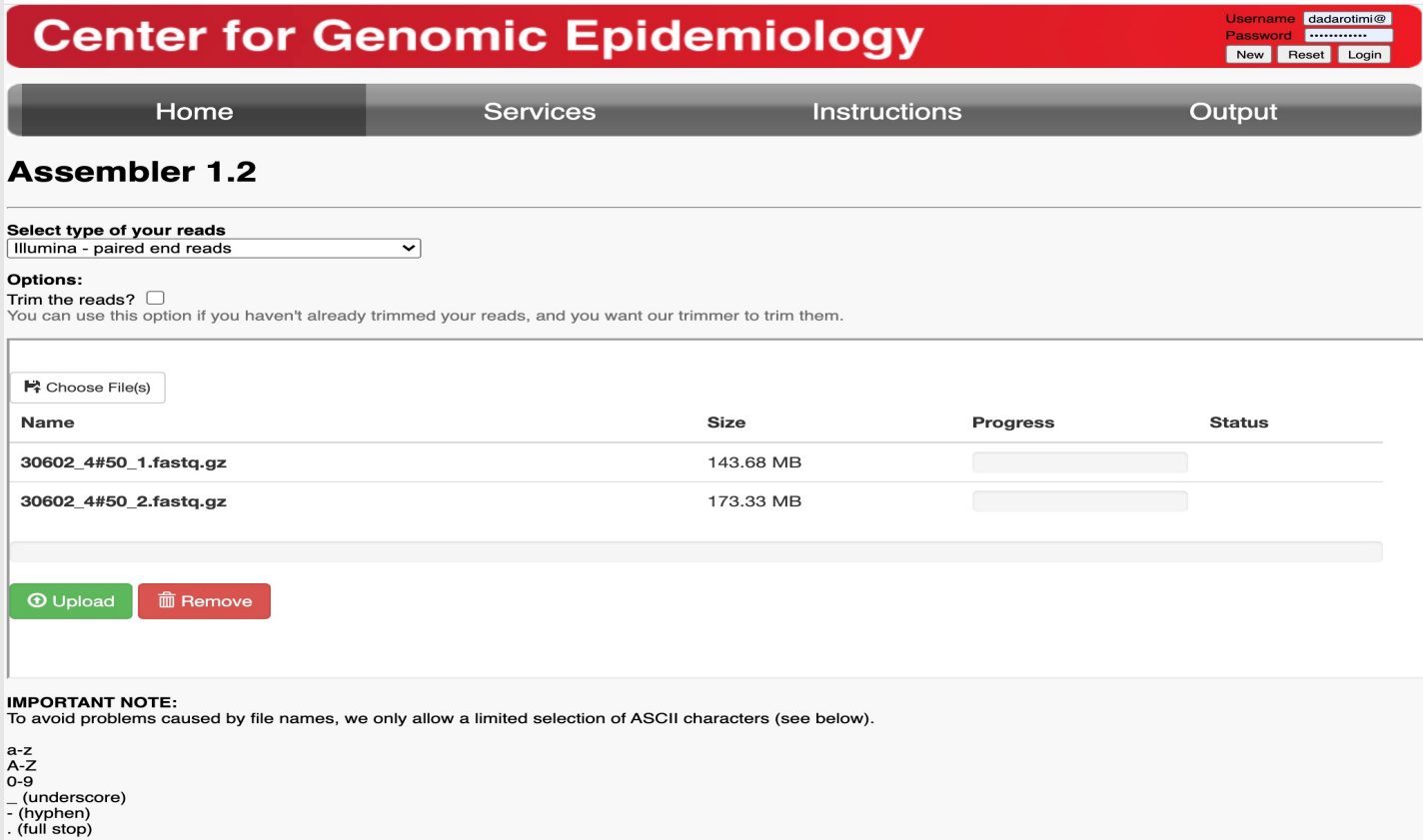
Public Health Alliance for Genomic Epidemiology

- De-novo assemblers used algorithms based on de Bruijn graphs
- Assembly is done using k-mers which are nucleotide patterns of specific length
- Web-based assembler – Center for Genomic Epidemiology - https://cge.cbs.dtu.dk/services/Assembler/
- Some software used for de-novo assembly –Velvet, ABySS, SPAdes

SPAdes







Public Health Alliance for Genomic Epidemiology

Assembly quality metrics

- QC is very essential at every level of genomic study
  - Nucleic Acid extraction
  - Library prep
  - Sequencing
  - Assembly, e.t.c.
- Tools used for measuring assembly quality
  - QUAST
  - SQUAT

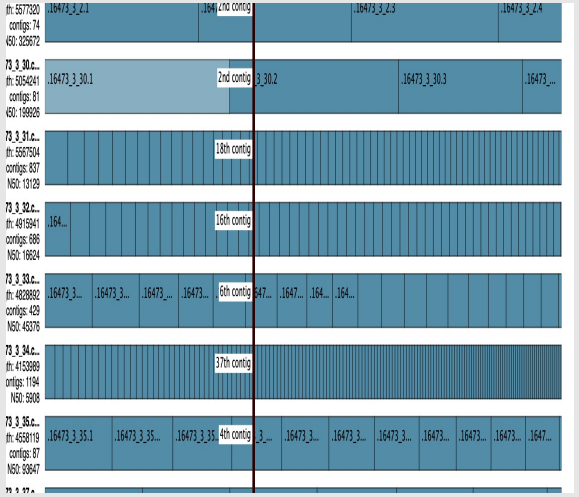
Public Health Alliance for Genomic Epidemiology

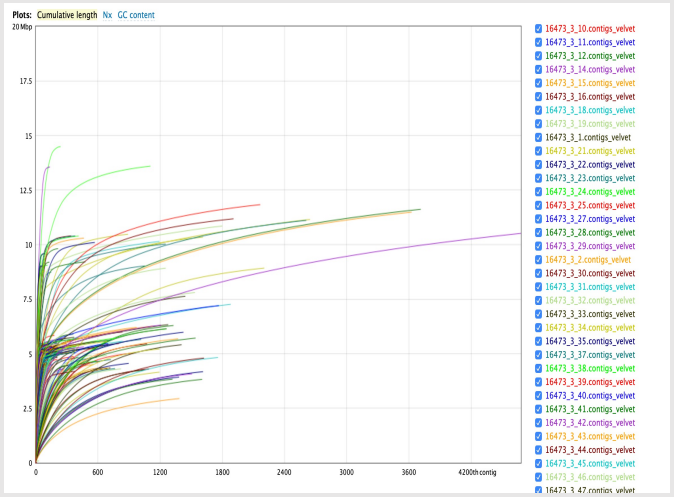
- N50 : length for which the collection of all contigs of that length or longer covers at least 50% of **assembly** length
- Number of contigs
- GC %: total number of G and C nucleotide in the assembly, divided by the total length of the assembly
- L50: minimum number of contigs that produce 50% of the bases of the assembly

Show heatmap

Worst Median Best

Statistics without reference	16473_3_10.contigs_velvet	16473_3_11.contigs_velvet	16473_3_12.contigs_velvet	16473_3_14.contigs_velvet	16473_3_15.contigs_velvet	16473_3_16.contigs_velvet	16473_3_
# contigs	73	94	224	85	96	98	89
# contigs (>= 0 bp)	88	211	588	113	109	191	99
# contigs (>= 1000 bp)	67	69	136	81	92	82	82
# contigs (>= 5000 bp)	50	41	69	62	61	51	62
# contigs (>= 10000 bp)	43	35	58	56	47	44	57
# contigs (>= 25000 bp)	32	30	46	42	37	32	45
# contigs (>= 50000 bp)	24	25	34	36	28	25	32
Largest contig	589 572	359 504	359 891	283 319	509 365	942 935	462 462
Total length	5 003 197	4 947 208	5 270 992	5 280 355	4 871 019	5 372 113	4 597 638
Total length (>= 0 bp)	5 008 775	4 986 779	5 397 899	5 290 537	4 875 882	5 403 968	4 601 050
Total length (>= 1000 bp)	4 999 391	4 929 530	5 210 143	5 277 346	4 868 374	5 360 633	4 593 222
Total length (>= 5000 bp)	4 954 404	4 857 854	5 079 072	5 229 636	4 788 429	5 285 902	4 545 444
Total length (>= 10000 bp)	4 901 729	4 809 771	4 999 952	5 187 415	4 689 872	5 245 617	4 512 963
Total length (>= 25000 bp)	4 735 268	4 721 575	4 809 688	4 937 054	4 545 722	5 036 679	4 307 012
Total length (>= 50000 bp)	4 451 292	4 510 806	4 362 092	4 727 604	4 245 653	4 777 329	3 833 448
N50	163 769	183 803	130 334	134 870	171 332	201 887	104 493
N75	112 102	136 332	63 714	77 702	82 258	111 415	65 699
L50	8	10	13	13	10	8	12
L75	17	17	27	25	19	17	26
GC (%)	50.25	50.33	50.31	50.43	50.74	50.49	50.64
Mismatches							
# N's	44 651	105 330	346 257	75 943	45 646	86 333	8631
# N's per 100 kbp	892.45	2129.08	6569.11	1438.22	937.09	1607.06	187.73





Public Health Alliance for Genomic Epidemiology

THANK YOU FOR YOUR ATTENTION