Quality assessment, read preprocessing and assembly

Sara Monzón Fernández

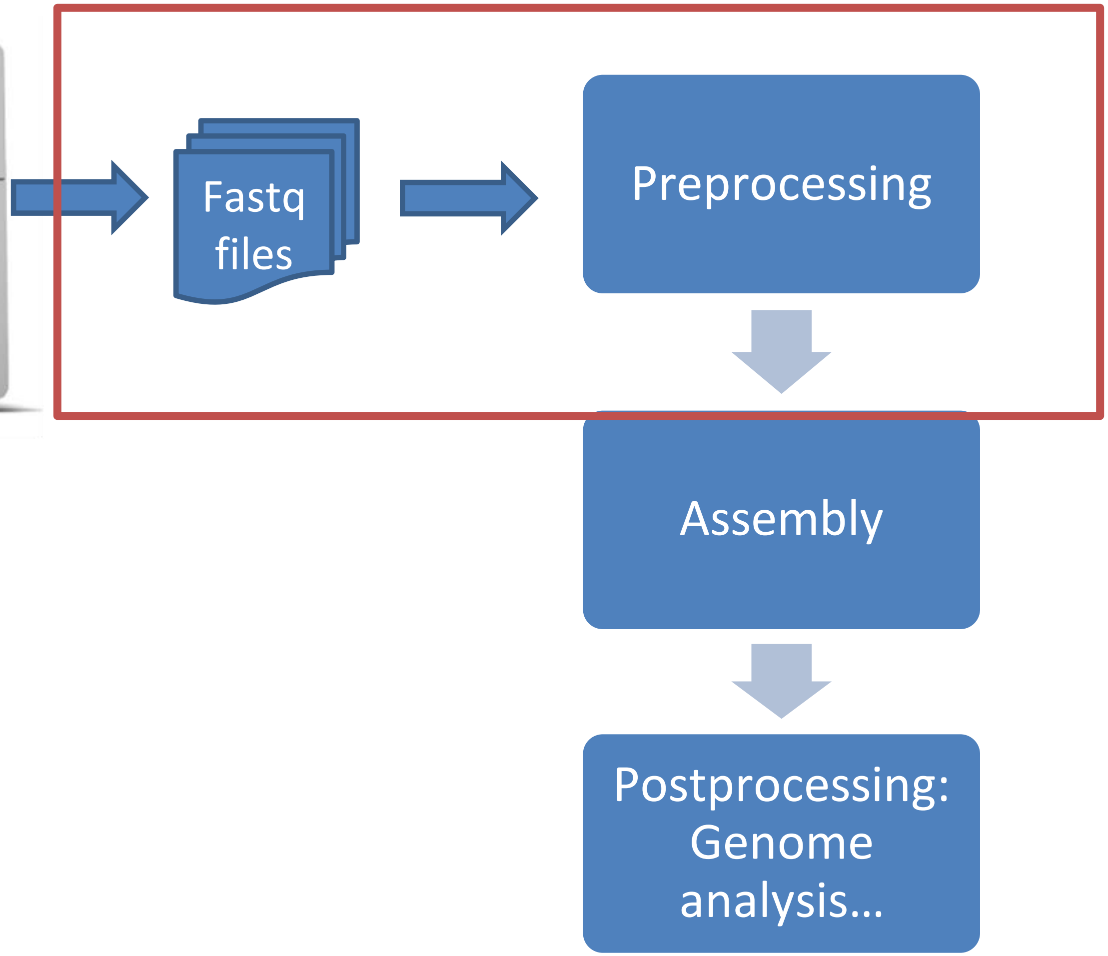
BU-ISCIII

Unidades Comunes Científico Técnicas – SGSAFI-ISCIII

09 Marzo 2021

UAH - ISCIII

Step in the process



Raw output files format

Illumina

.fastq

454 .sff

SOLiD

.fasta
.qual

Nanopore
FAST
5

PacBio RSII
Bax.h
5

FASTQ format

- Is a FASTA file with quality information
- Within HTS, FASTA contain genomes y FASTQ reads

```
>SEQ_ID
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAG
```

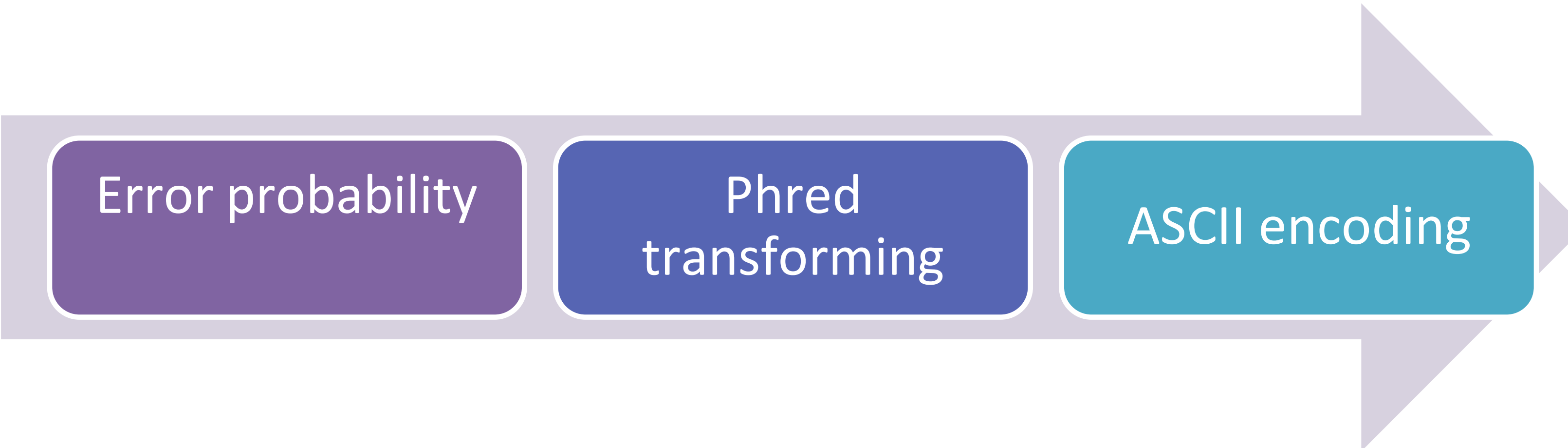
Sequence

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Quality: must be 1 bit

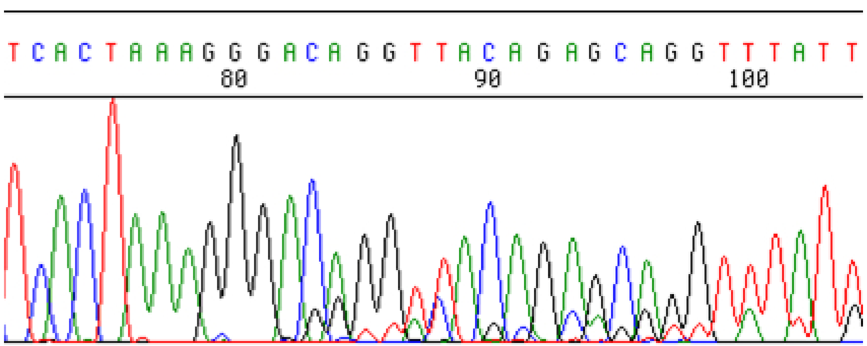
FASTQ format

- Each base has an assigned quality score
  - Sequencing quality scores measure the probability that a base is called incorrectly
- How is it calculated?

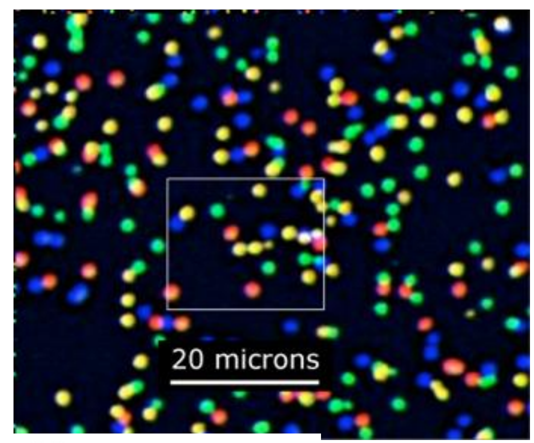


Phred quality and error probability

- **Light intensity** is used to calculate the error probabilities
- Convert error probability into Phred score quality - Ewing B, Green P. (1998)
- Phred originated as an algorithmic approach that considered Sanger sequencing metrics, such as **peak resolution and shape**



wellcome sanger institute



illumina

Phred quality and error probability

- Convert error probability into Phred score quality - in real time on Illumina platforms
- Q scores are defined as a property that is logarithmically related to the base calling error probabilities (P)
- Phred quality range between 0-40 for Sanger and Illumina 1.8+

$$Q = -10 \log_{10} P$$

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%

Phred quality and error probability

- Convert Phred quality score into ASCII, a compact form, which uses only 1 byte per quality value

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

- Phred+33 (Sanger and current Illumina). 0 Phred quality correspond to decimal 33, which is the symbol !

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

- Phred+64 (Solexa and Illumina 1.3-1.5)

Phred quality and error probability

- **Phred 33 example**

```
@HWI-ST731_6:1:1101:1322:1938#1@0/1
NTGACAAAGGGCTAATATCCAGAATCTACAAAGAACTTAAACAAATGTATAAGAATAAAAGTATAGTGCTAACAAT
+
#1:BDDADFDFDD@F>BGFIIIB@CFHIHICAGBC9CBCBGGIGCFF??>GGHFHIGGEGI<FECGDE=FHCHEG=
```

P=0.001 → Q=-10*log10(0.001)=30 → ASCIII 33+30 = 63 → ?

P=0.0001 → Q=-10*log10(0.001)=40 → ASCIII 33+40 = 73 → I

```
Quality encoding: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
                  |         |         |         |         |
   Quality score: 0.........10........20........30........40
```

FASTQ format

Illumina read header

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

```
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:2458:1027 1:N:0:ACAGTG
AGAAAAAACCTTGGANGGAAAAAAATCAGACATTTTCTAGAGGTGGAAGGCAAACTGAACAAAGAAATAATTCACA
+
DGGGEDHHHHGGGFE#CBACBCA<?HHHHBHHHHHHHHHDHHHHEHEFEGGGGGG/GGDDDGHFHGFCHFHHEHEH8
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3082:1029 1:N:0:ACAGTG
GGTAATACAGACTGANATGATCAAAGGCATGCTGGAAACAAACCTATTAAAGATAAGCTTGGATCAAGCTTTCATT
+
B:B:?BB/:=55177#55877<775EDD>E=B?BBBBGGGDDAG@G>GGGGGG@)EEEEBEG>GGGGGGGAAA?<D
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3185:1033 1:N:0:ACAGTG
TCTGGGACATTGCTCNTGGCTGGGAGTCACCTGTCTGGGACATTGCTCAGGGCTGGGAGACACGTGTTGGAGGGAC
+
BC??A66;)74781<#7??;452.27'64(8,851DDG8GB?###################################
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3268:1033 1:N:0:ACAGTG
ATTCAAATTAGAAGANAGTTGATCGTTCTTCATGATGCCCAAAAATTTCACTGAGAAAACCCTTTTTTAAGCCCAC
+
IIIIIIIIIIFFFFE#ABACFEEFFIIGIIIFIHE@BIIIIIIIIIHHIIFIIF>HHIHIFGDIIIIIIGFHIEGH
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3400:1035 1:N:0:ACAGTG
TCCTGCTTTAGGAGANTCCTCATGCTCTGACAGGATGCTCTCTATGTGAGTTGAGCTGGTCTTCTCACTTTTATAG
+
IIIIIHIHIIGGEGG#AACA@?=?BHHIIIIIHHIHIIIIHIHIHHGIHIHGHGIGIHGEGGGGHG@EFGGCEFAB
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3962:1033 1:N:0:ACAGTG
CCACCAACACAGTCTNCACCTTCTGTTGCTGGTGATAGATTTTTGCACCTTTCCATCCTCCAGGTTTCAAAATAGC
+
HHFHHDHDHH>C?CA#EEEE>?A?>HHDGHEGBGBCEEEEGHHF8HEHEEHECH,=>>==EAEE>BEBBAEAACAB
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:4491:1028 1:N:0:ACAGTG
AGAGAGAGAGAGAGANAGAGGACTCTGGAGATGCCGAAGCACAAGCCTGCAAGAGTCCCAGCAAAGAAAATAAAAA
+
GADGGEGGEGBBB?B#@=@@72;64GGGFGB>GGGBDG<DBGB<DA??/?##########################
```

ASCII-coded (0-40):

- “!"#$%” lowest quality
- “FGHI”   highest quality

Sequencing quality assessment

- To asses quality, software uses **Phred per-base quality** score is used
- Is the **first quality control step** after sequencing. There should be one after every step of the analysis
- After quality assessment user can know how **reliable** are their datasets
- QC will determine the next **filtering** step
- Filtering decisions will **impact** directly in **further analysis**
- Many other steps also use this quality as variable in their **algorithms**

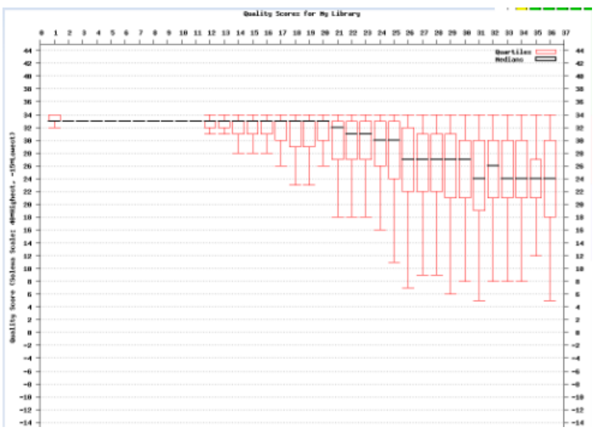
Sequencing quality assessment: Artifacts

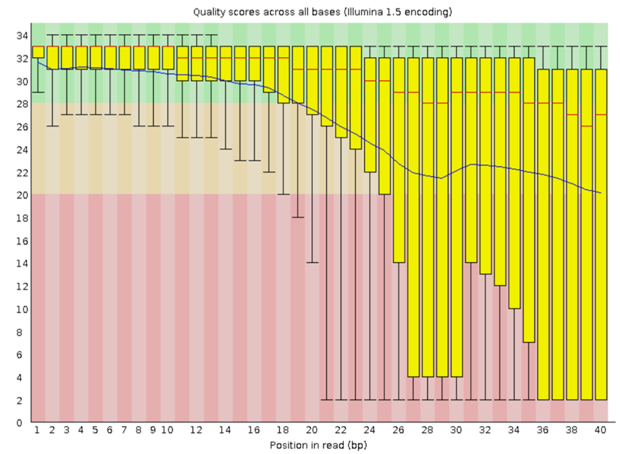
HTS methods are bounded by their technical and theoretical limitations and sequencing errors cannot be completely eliminated (Hadigol M, Khiabanian H. 2018)

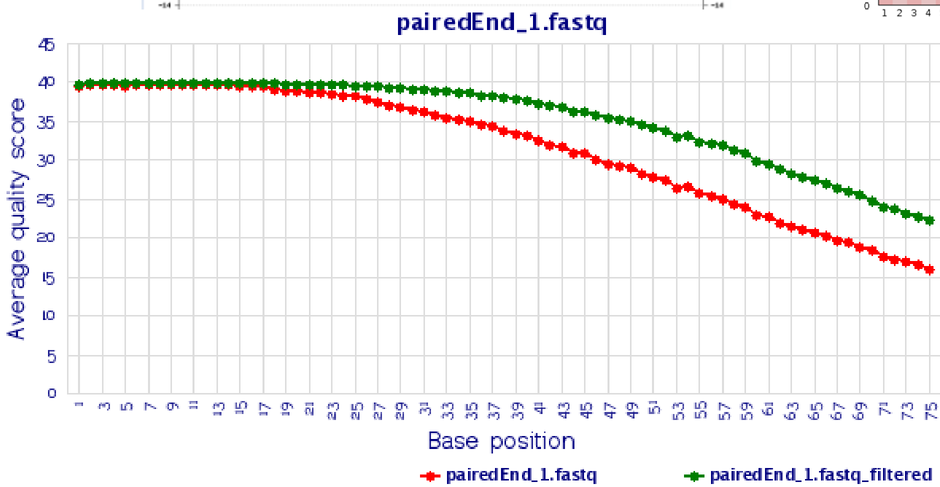
- **Artifacts in library preparation**
  - Remaining adapters
  - High rate of duplicates
  - GC regions bias
  - Polymerase error rate
  - DNA damage during breakdown
- **Artifacts during secuencing**
  - Low quality in sequence ends(Phasing: cluster loose sync)
  - Complication in certain regions:
    - Repetitions
    - Homopolymers
    - High CG content

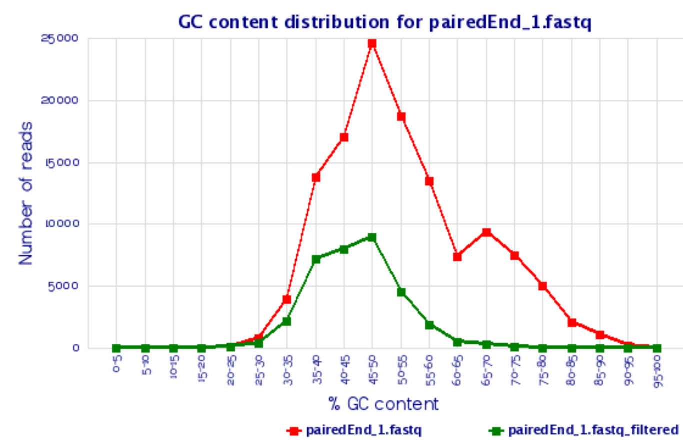
Sequencing quality assessment

- FastQC, fastx-toolkit, sfftools, NGSQCToolkit, etc…

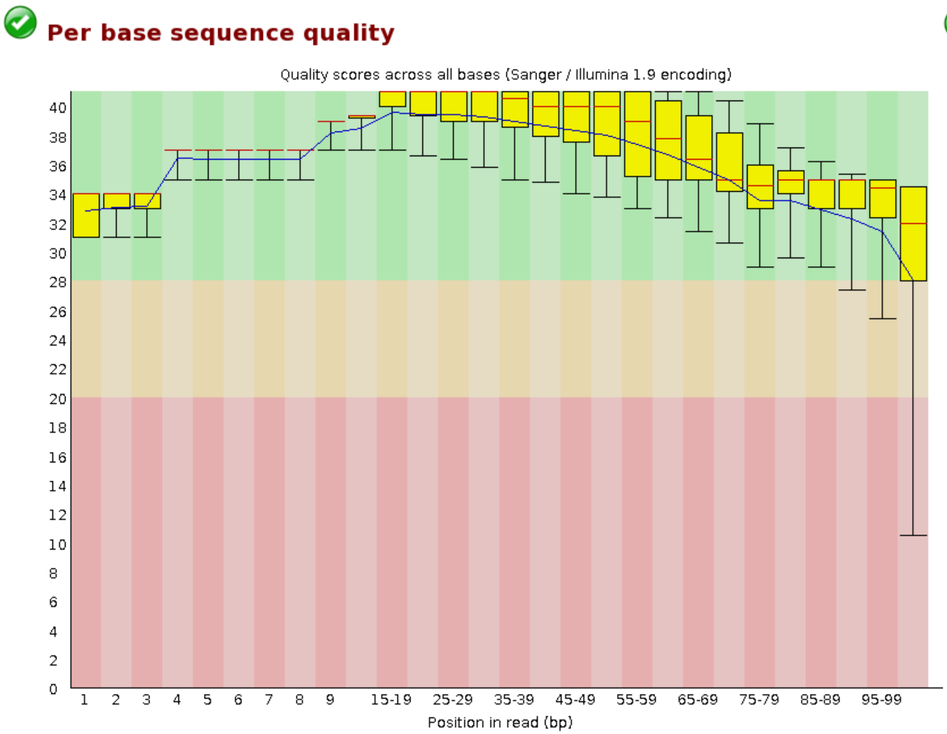


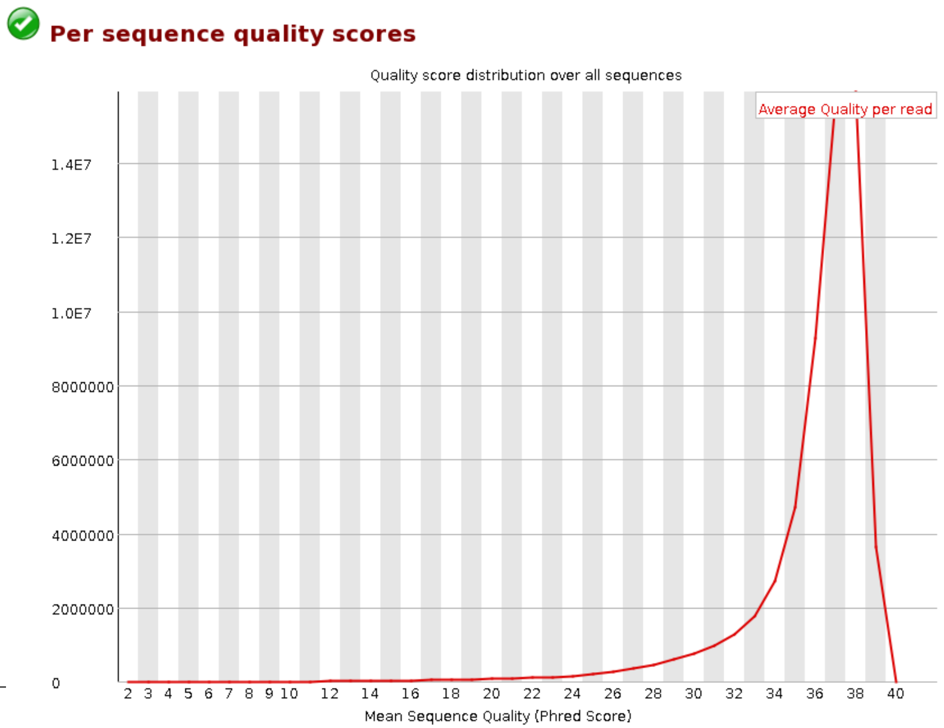






Sequencing quality assessment: FastQC





https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FastQC: Basic Statistics

- Self defined overall stats
  - Encoding: Phred33 or Phred64

Basic Statistics

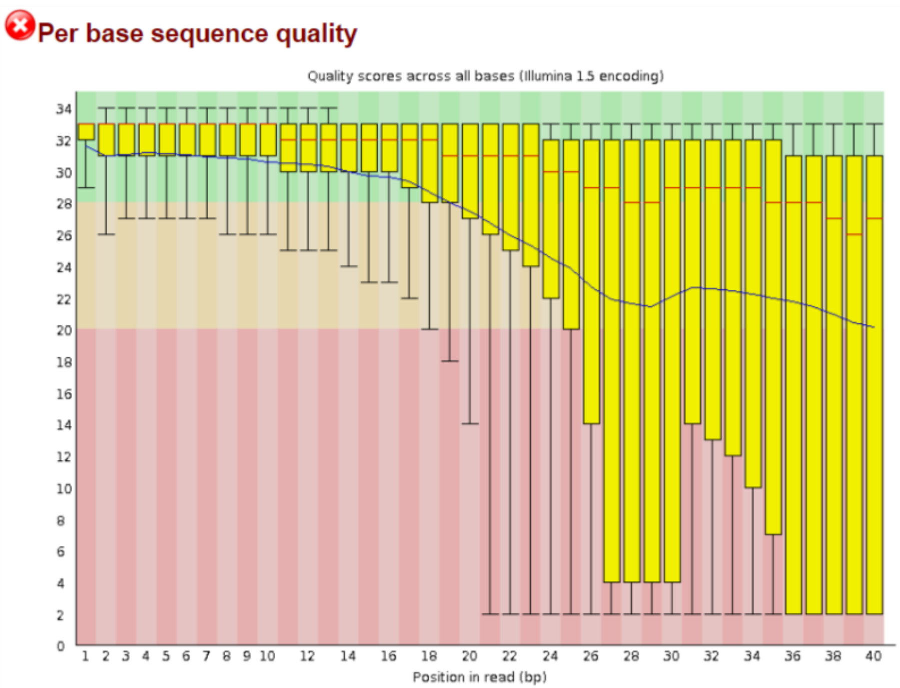
Measure	Value
Filename	bad_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	395288
Sequences flagged as poor quality	0
Sequence length	40
%GC	47

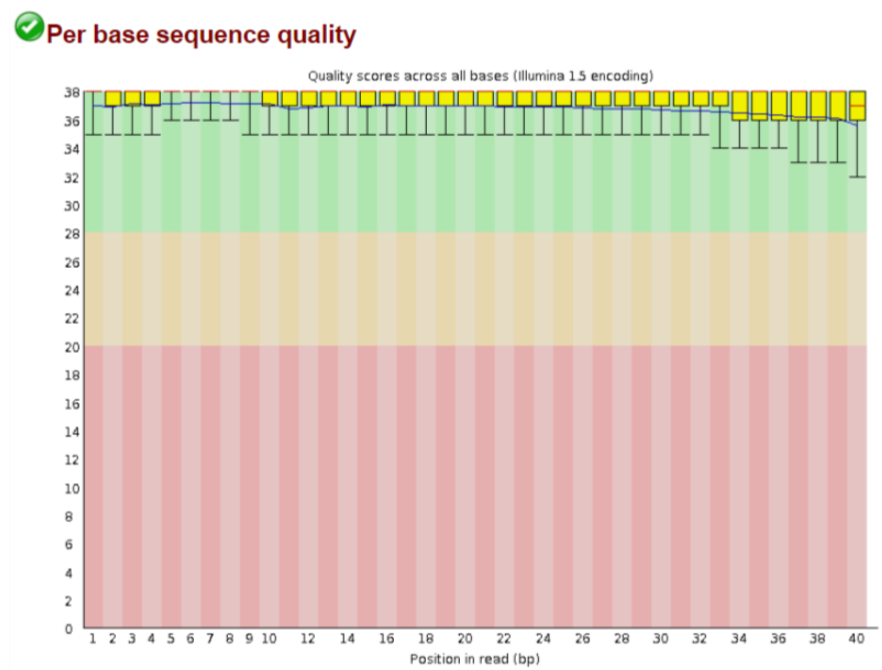
Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

FastQC: Per base sequence quality

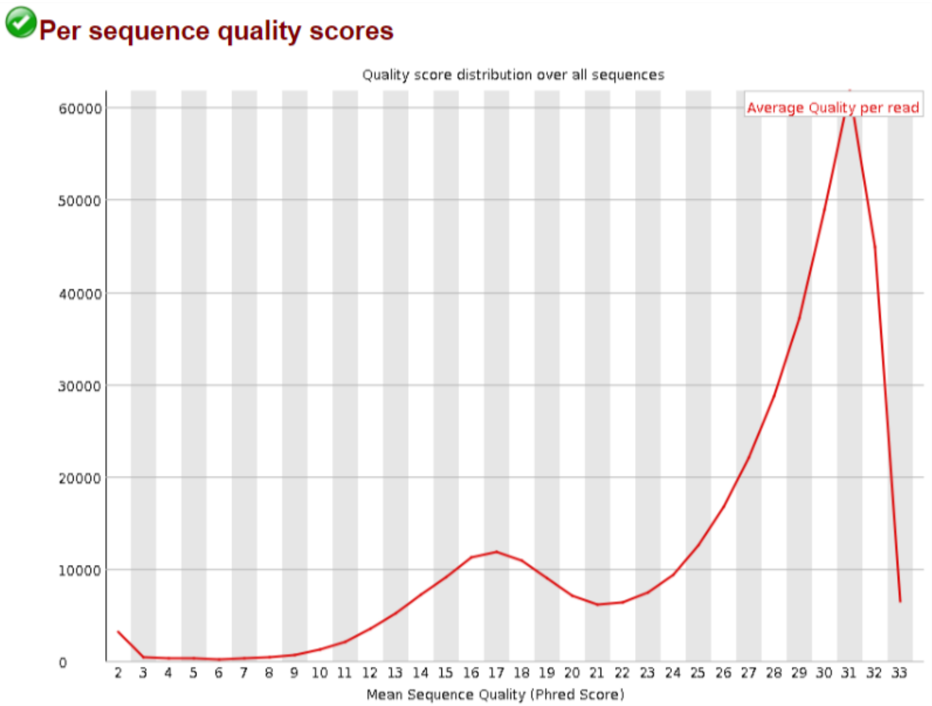
- Overview of the range of quality values across all bases at each position in the FastQ file
- **Median**, **inter-quartile range** **(25-75%), 10-90% points**, **mean quality**

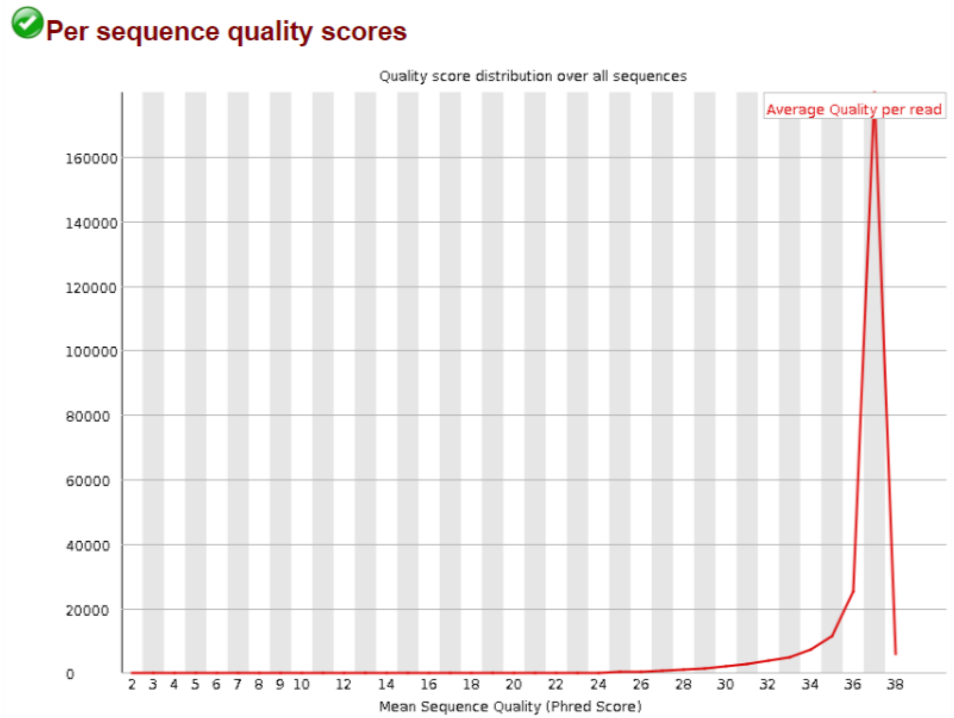




FastQC: Per sequence quality score

- Number of sequences with the same mean quality

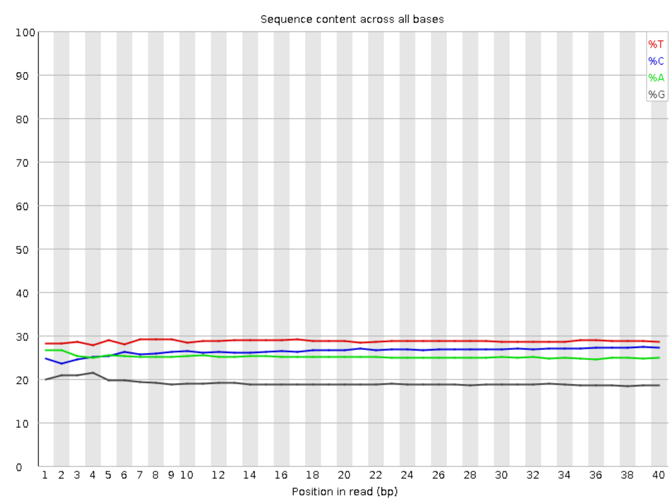




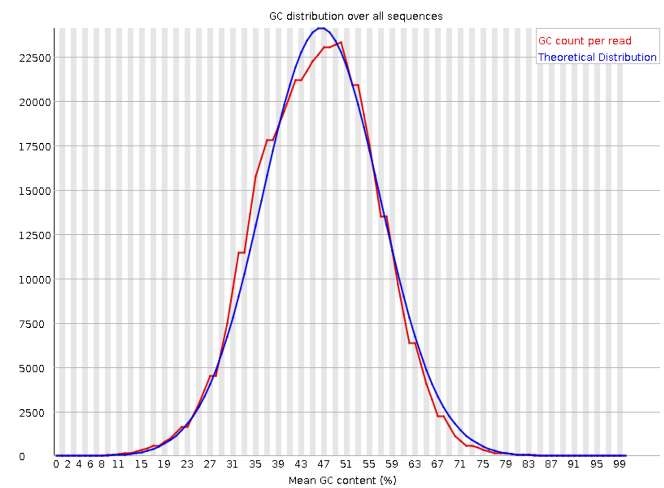
FastQC: Nucleotide related errors

- How expected nucleotide distribution deviates from expected
  - Per base sequence content
  - Per base GC content
  - Per sequence GC content
  - Per base N content

Per base sequence content

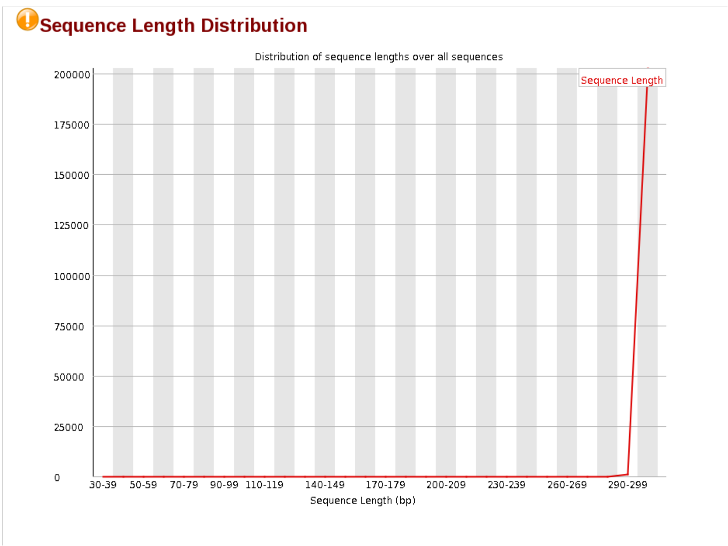


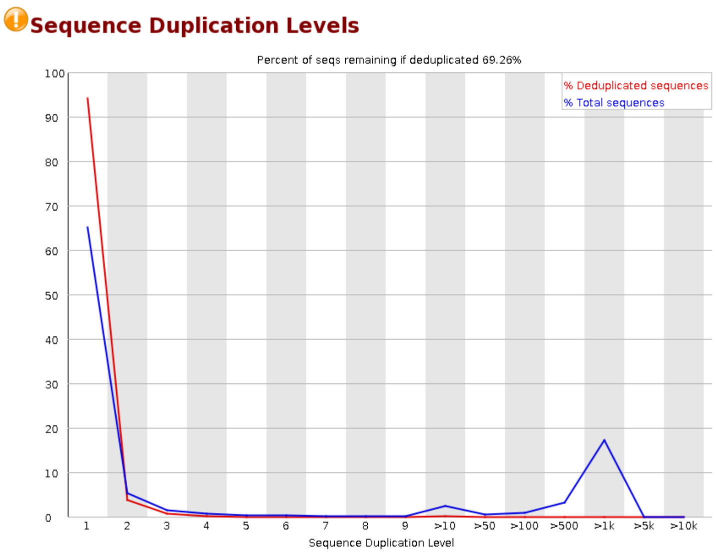
Per sequence GC content



FastQC: Sequence related errors

- How expected nucleotide distribution deviates from expected
  - Sequence Length Distribution - Fragments
  - Sequence Duplication Levels
  - Overrepresented sequences
  - Adapter Content





Sequence filtering

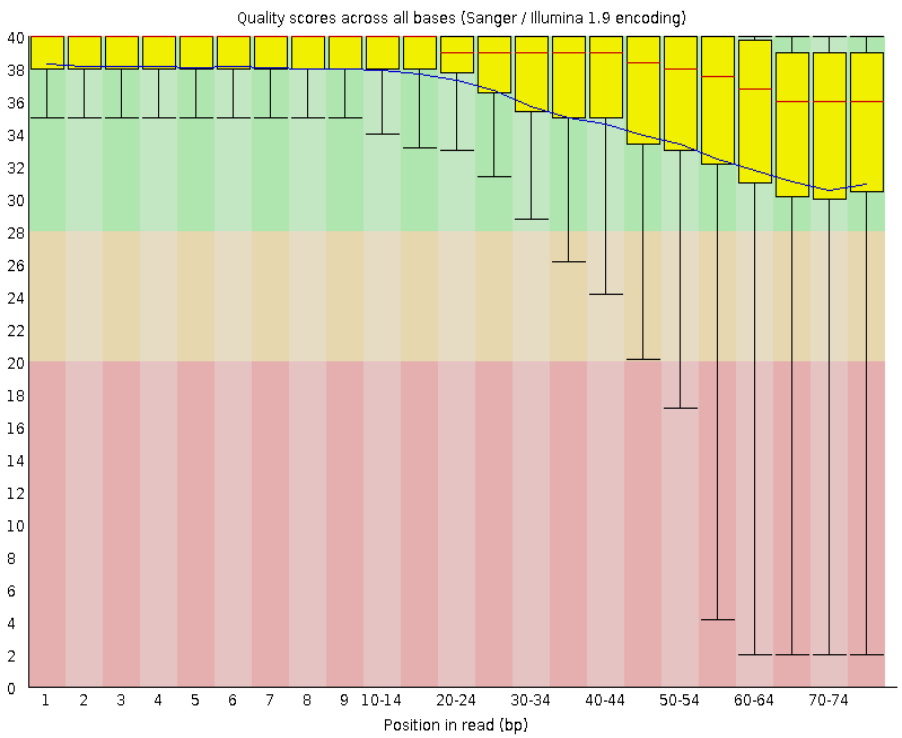
- **Remove residual adapters**
  - Depending on used library

- **Filtering parameters**
  - Quality filtering
    - Overall mean quality
    - Local mean quality
      - Sequence end
      - Sliding window
  - Size filtering
    - Overall sequence size
    - Remaining sequence size after filtering

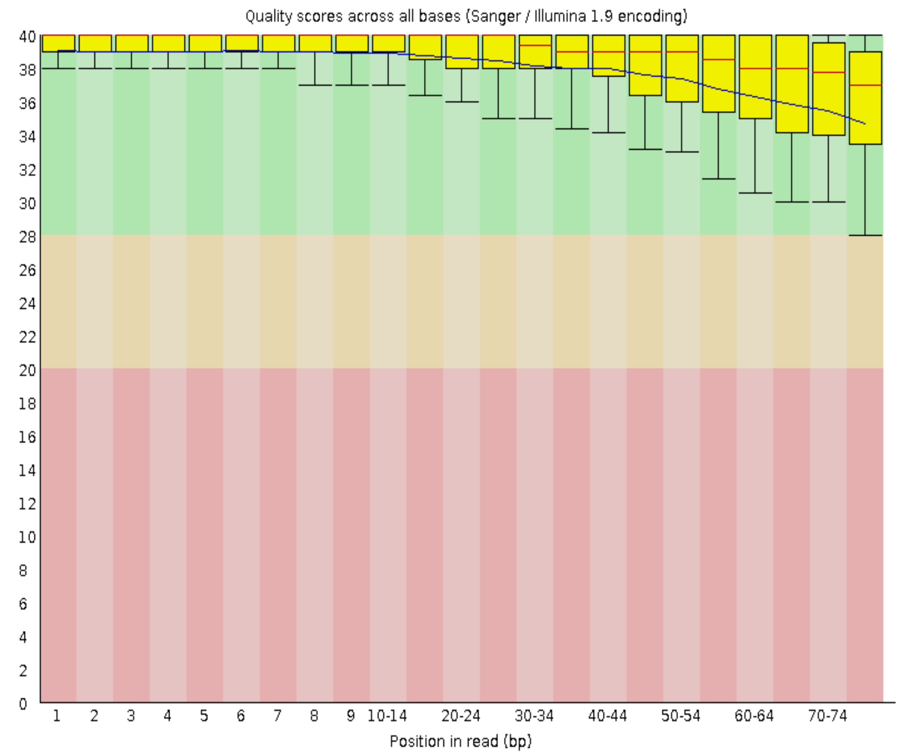
Sequence filtering

- **Example of quality filtering**

Per base sequence quality



Per base sequence quality



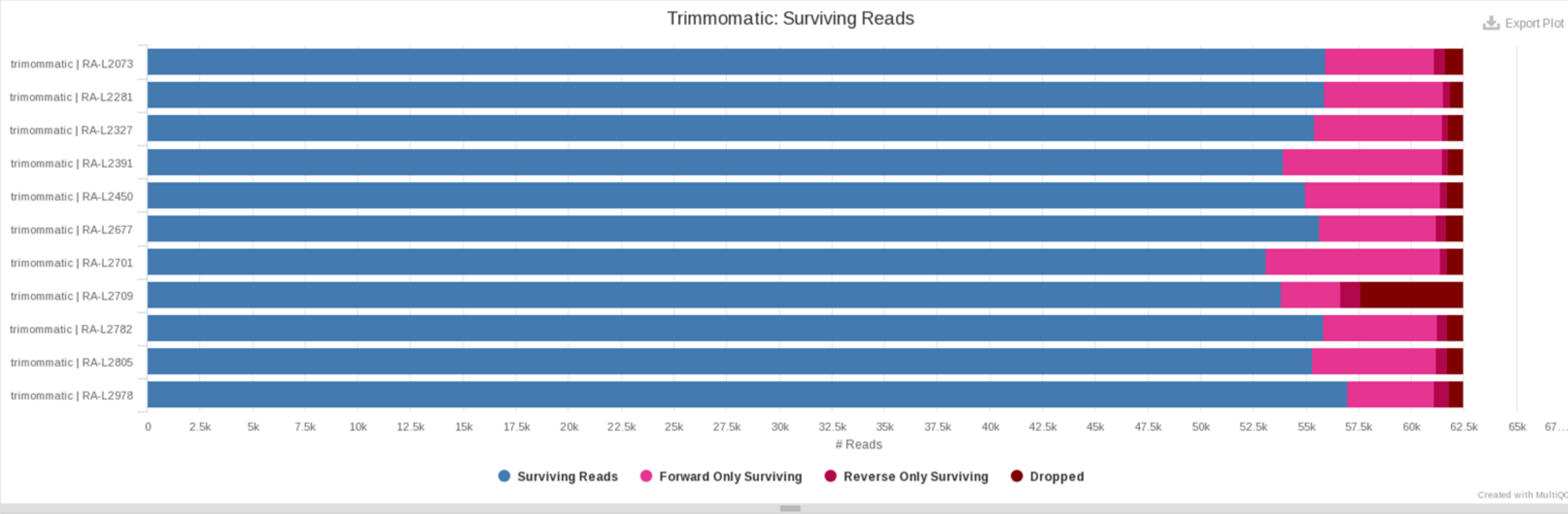
Sequence filtering: stats with MultiQC

Sequence filtering: stats with MultiQC

Trimmomatic

Trimmomatic is a flexible read trimming tool for Illumina NGS data.

Number of Reads | Percentages



Assembly

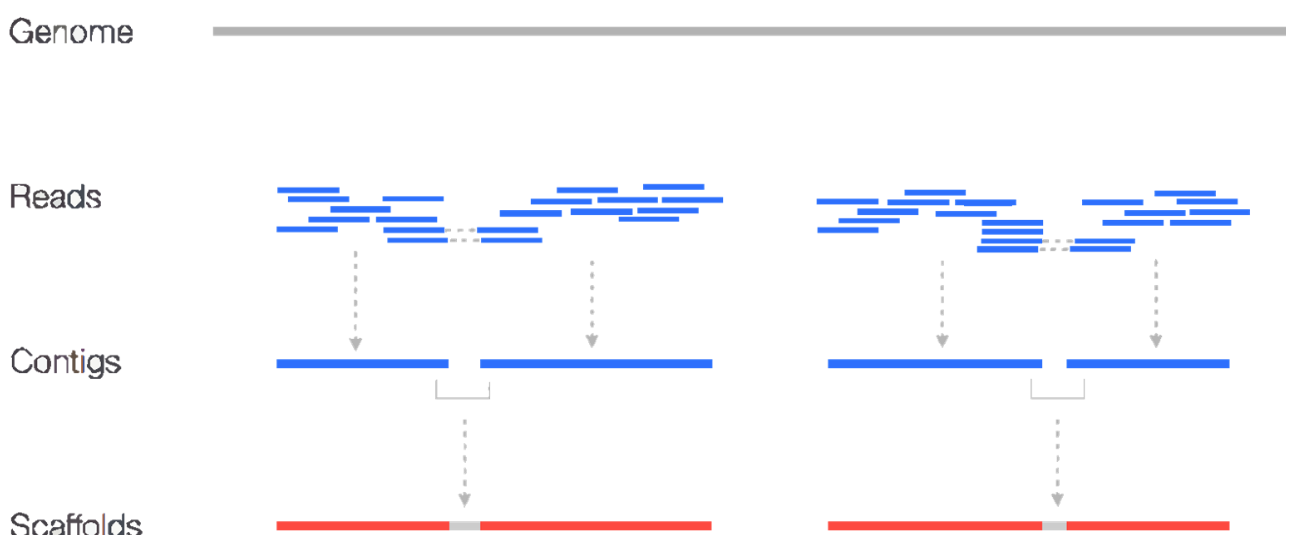
Reconstruct a representation of the original DNA from shorter DNA sequences or small fragments known as reads

- ***De novo:*** with no previous knowledge of the genome to be assembled. It overlap the end of the end of each read in order to créate a longer sequence.

- ***Assembly with reference:*** A similar but not identical genome guides the assembly process. Map reads over supplied genome.

Assembly: contig y scaffold

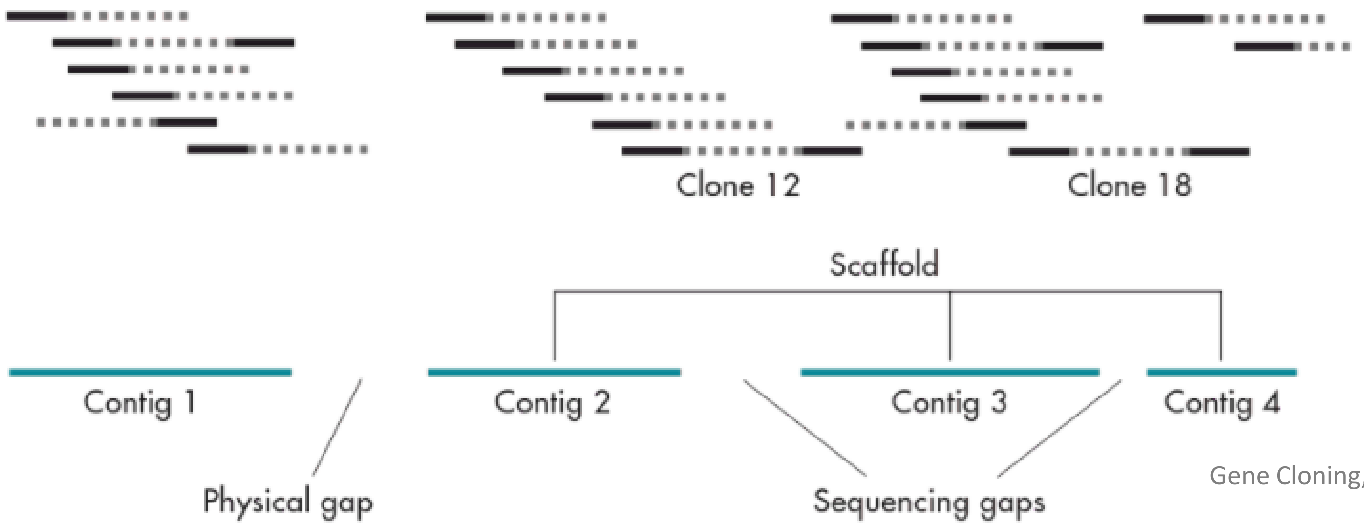
- **Contig:** continuous sequence made up of overlaping shorter sequences
- **Scaffold:** two or more contigs located and rearranged according to spatial information(pair-end, mate pair, reference)



https://www.biostars.org/p/253222/

Assembly: gaps

- **Sequencing gaps:** Position and orientation known by spatial information
- **Physical gaps:** No information about adjacent contigs



Gene Cloning, Lodge *et al.*

Assembly: Algorithms

- **De Brujin Graph (DBG: k-mer graph)**

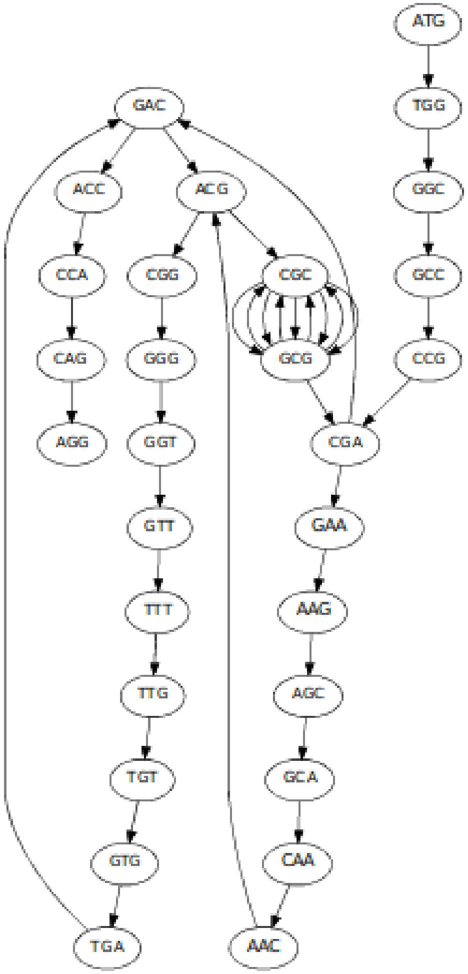
Chopping reads into much shorter k-mers (fi
length fragments) and then using all the
mers to form a DBG and infer the contigs.

- Nodes in the graph are k-mers
- Edges represent consecutive k-me

(which overlap by k-n symbols)

Ex. SPAdes, ABySS, Velvet, AllPaths, Soap….

https://medium.com/@han_chen



Algorithms: DBG

- To be able to use de Bruijn graphs, we need reads of length **L** to overlap by **L-1** bases.
- Not all reads will overlap another read perfectly.
  - Read errors
  - Coverage "holes"
- Not all reads are the same length (depending on technology and quality cleanup)

To help us get around these problems, we use all k-length subsequences of the reads, these are the k-mers.

Algorithms: DBG

TTGACACTTACCGA **Read**

TTGACACTTACC
TGACACTTACCG
GACACTTACCGA

k-mers for k=12

TTGAC
TGACA
GACAC
ACACT
CACTT
ACTTA
CTTAC
TTACC
TACCG
ACCGA

k-mers for k=5

Algorithms: DBG

Example #1:

HAPPI PINE INESS APPIN

All 4-mers:

HAPP PINE INES APPI

APPI NESS PPIN

Unique 4-mers:

HAPP APPI PINE PPIN INES NESS

Algorithms: DBG

Example #1:

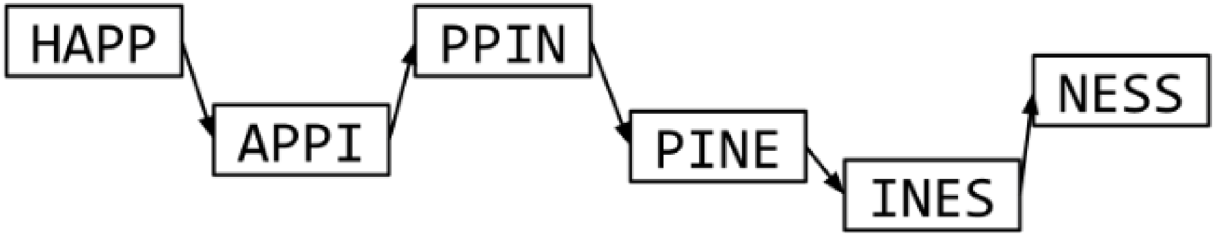
HAPPI PINE INESS APPIN

k = 4 k-mers:

HAPP APPI

PINE PPIN

INES NESS



Algorithms: DBG

Example #1:

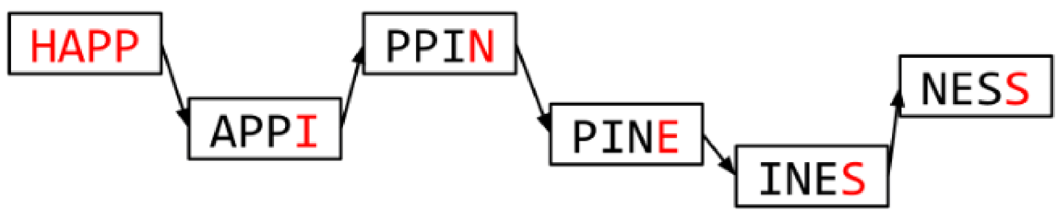
HAPPI PINE INESS APPIN

k = 4 k-mers:

HAPP APPI

PINE PPIN

INES NESS



HAPPINESS

Easy!

Algorithms: DBG

Example #2:

```
MISSIS SSISSI SSIPPI
```

All 4-mers (9):

```
MISS    SSIS    SSIP
 ISSI    SISS    SIPP
  SSIS    ISSI    IPPI
```

Unique 4-mers (7):

```
MISS SSIS SSIP ISSI SISS SIPP IPPI
```

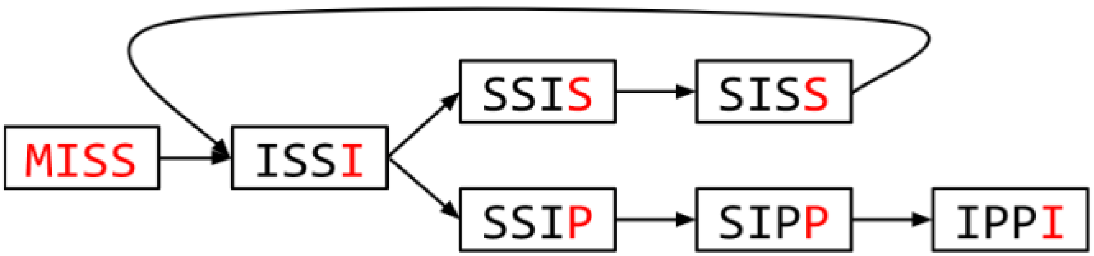
Algorithms: DBG

Example #2:

MISSIS SSISSI SSIPPI

All 4-mers:

MISS ISSI SSIS SISS SSIP SIPP IPPI



MISSISSIPPI or MISSISSISSISSIPPI or ...

Algorithms: DBG

Example #2a:

MISSIS SSISSI SSIPPI

All 5-mers (6):

MISSI SSISS SSIPP

ISSIS SISSI SIPPI

Unique 5-mers (6, no duplicates):

MISSI ISSIS SSISS SISSI SSIPP SIPPI

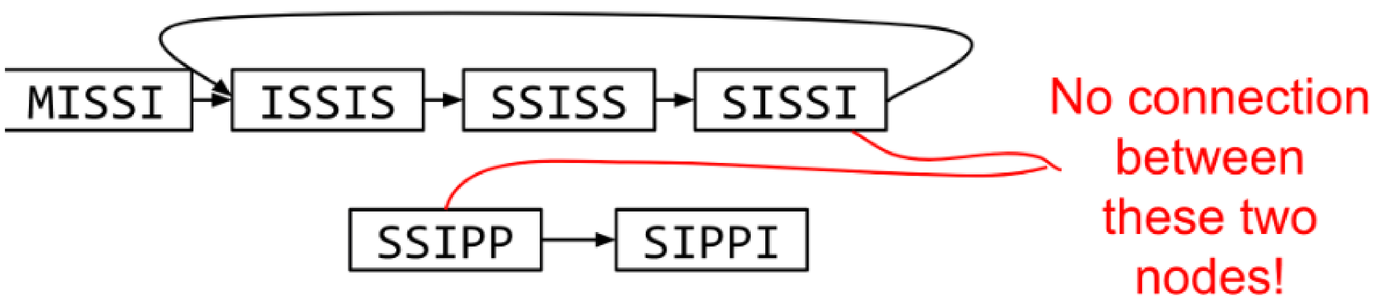
Algorithms: DBG

Example #2a:

MISSIS SSISSI SSIPPI

This time k = 5 k-mers:

MISSI ISSIS SSISS SISSI SSIPP SIPPI



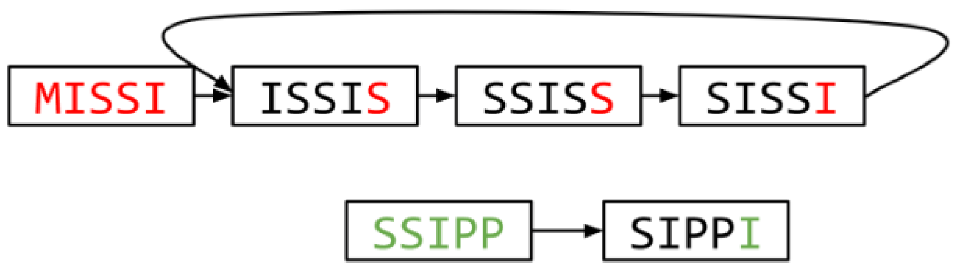
Algorithms: DBG

Example #2a:

MISSIS SSISSI SSIPPI

This time k = 5 k-mers:

MISSI ISSIS SSISS SISSI SSIPP SIPPI



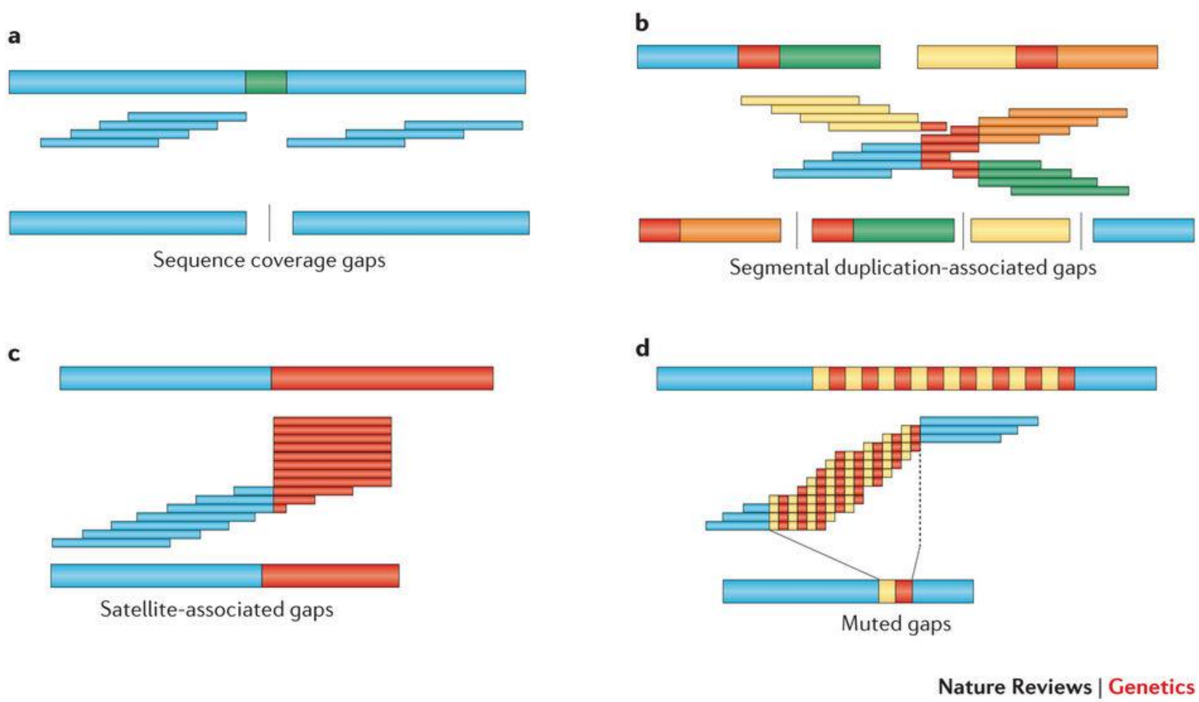
MISSISSIS SSIPPI

Algorithms: DBG

- Lower k
  - More connections
  - Less chance of resolving small repeats
  - Higher k-mer coverage
- Higher k
  - Less connections
  - More chance of resolving small repeats
  - Lower k-mer coverage

Optimum value for k will balance these effects.

Assembly: Errors



- **A. Gaps – non sequenced region**
- **B. Long repeats**
  - Cuimera
- **Collapsed repetitive regions**
  - **C. Terminal**
  - **D. Interstícial**

Genetic variation and the de novo assembly of human genomes
Chaisson *et al.*

Assembly: Scaffolding

- **From draft:**

  Order contigs (Nucmer, if there is reference it can be used to align and guide)

  Fill the GAPs (GapFiller, fill sequencing gap (not physical gap)

  Solve repeated sequence ambiguities (Expander)

  Resequence with different library:

  - Longer fragments and/or distance

- **Tools for assembly improvement**

  SSPACE (Scaffolding) REAPR (evaluate scaffolding, breaking incorrect scaffolds)

- **Assembly visualyzing**

  Artemis, ACT (compare two or more sequences), Icarus (Quast)

Assembly: Evaluation

- Software that evaluate differets algorithms & parameters
  iMetAMOS, *Koren et al., BMCBioinformatics 2014, 15:126*
  GAGE-B, *Magoc et al.,* Bioinformatics 2013,29(14):1718-25

- **Graph evaluation**: Bandage, Wick R.R., Schultz M.B., Zobel J. & Holt K.E. (2015)

- **Assembly evaluation**: Quast, *Gurevich et al., Bioinformatics 2013, 29:8*

- **Metrics for a good assembly:**
  Large N50
  Sum closest to expected
  Low n
  Low L50

Assembly: Evaluation - Quast

- Assembly evaluation: Quast, *Gurevich et al., Bioinformatics 2013, 29:8*

Worst Median Best

Show heatmap

Genome statistics	RA_L2073_paired_assembly	RA_L2391_paired_assembly	RA_L2677_paired_assembly	RA_L2978_paired_assembly	RA_L2281_paired_assembly	RA_L2450_paired_assembly	RA_L2701_paired_assembly
Genome fraction (%)	81.079	88.828	84.92	90.172	85.733	88.172	92.463
Duplication ratio	1	1	1.001	1.001	1.001	1	1
# genomic features	1736 + 824 part	2113 + 600 part	1881 + 768 part	2157 + 611 part	1992 + 637 part	2073 + 643 part	2368 + 412 part
Largest alignment	16 612	33 033	21 336	25 068	29 638	30 305	40 471
Total aligned length	2 405 510	2 635 297	2 519 300	2 675 166	2 543 440	2 615 874	2 743 222
NGA50	3176	6162	4234	5948	5104	5358	9519
LGA50	267	151	219	153	166	166	96
Misassemblies							
# misassemblies	23	1	14	2	17	12	4
Misassembled contigs length	84 193	9611	45 868	6390	111 490	72 879	37 962
Mismatches							
# mismatches per 100 kbp	17	18.78	15	16.71	341.39	15.75	13.49
# indels per 100 kbp	1.21	1.25	1.87	1.94	7.27	1.45	0.87
# N's per 100 kbp	0	0	0	0	0	0	0
Statistics without reference							
# contigs	748	546	684	569	569	584	392
Largest contig	16 612	33 033	21 336	25 068	30 915	30 305	40 471
Total length	2 440 656	2 676 227	2 562 578	2 714 287	2 629 607	2 618 624	2 787 129
Total length (>= 1000 bp)	2 439 127	2 676 227	2 559 569	2 714 287	2 628 029	2 615 105	2 785 415
Total length (>= 10000 bp)	257 236	739 181	320 638	811 392	700 516	658 319	1 419 641
Total length (>= 50000 bp)	0	0	0	0	0	0	0

Extended report

Assembly: Evaluation - Quast

- Assembly evaluation: Quast, *Gurevich et al., Bioinformatics 2013, 29:8*

Assembly: Assemblers

Name	Type	Technologies	Author	Presented /Last updated	Licence*	Homepage
DNASTAR Lasergene Genomics Suite	(large) genomes, exomes, transcriptomes, metagenomes, ESTs	Illumina, ABI SOLiD, Roche 454, Ion Torrent, Solexa, Sanger	DNASTAR	2007 / 2016	C	link
Newbler	genomes, ESTs	454, Sanger	454/Roche	2004/2012	C	link
Canu	Small and large, haploid/diploid genomes	PacBio/Oxford Nanopore reads	Koren et al.[8]	2001 / 2018	OS	link
SPAdes	(small) genomes, single-cell	Illumina, Solexa, Sanger, 454, Ion Torrent, PacBio, Oxford Nanopore	Bankevich, A et al.	2012 / 2017	OS	link
Velvet	(small) genomes	Sanger, 454, Solexa, SOLiD	Zerbino, D. et al.	2007 / 2011	OS	link

***Licences:** OS = Open Source; C = Commercial; C / NC-A = Commercial but free for non-commercial and academics