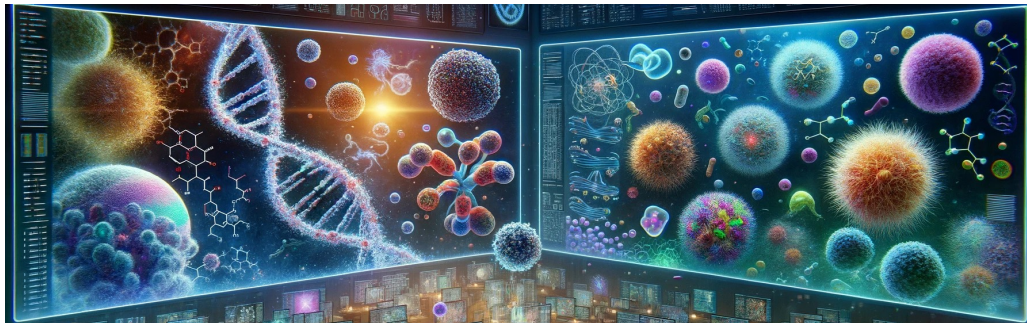


Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

UCD DUBLIN

www.ucd.ie/cfs
@CfsUcd

Advanced Applications of Next Generation Sequencing in Food Safety

Dr Guerrino Macori, BSc, MSc, PhD
Assistant Professor
School of Biology and Environmental Science
University College Dublin, Ireland
UCD Centre for Food Safety
guerrino.macori@ucd.ie
@guerrinomacori

Quality assessment raw data

Fastq

- Text-based format for storing both a DNA sequence and its corresponding quality scores

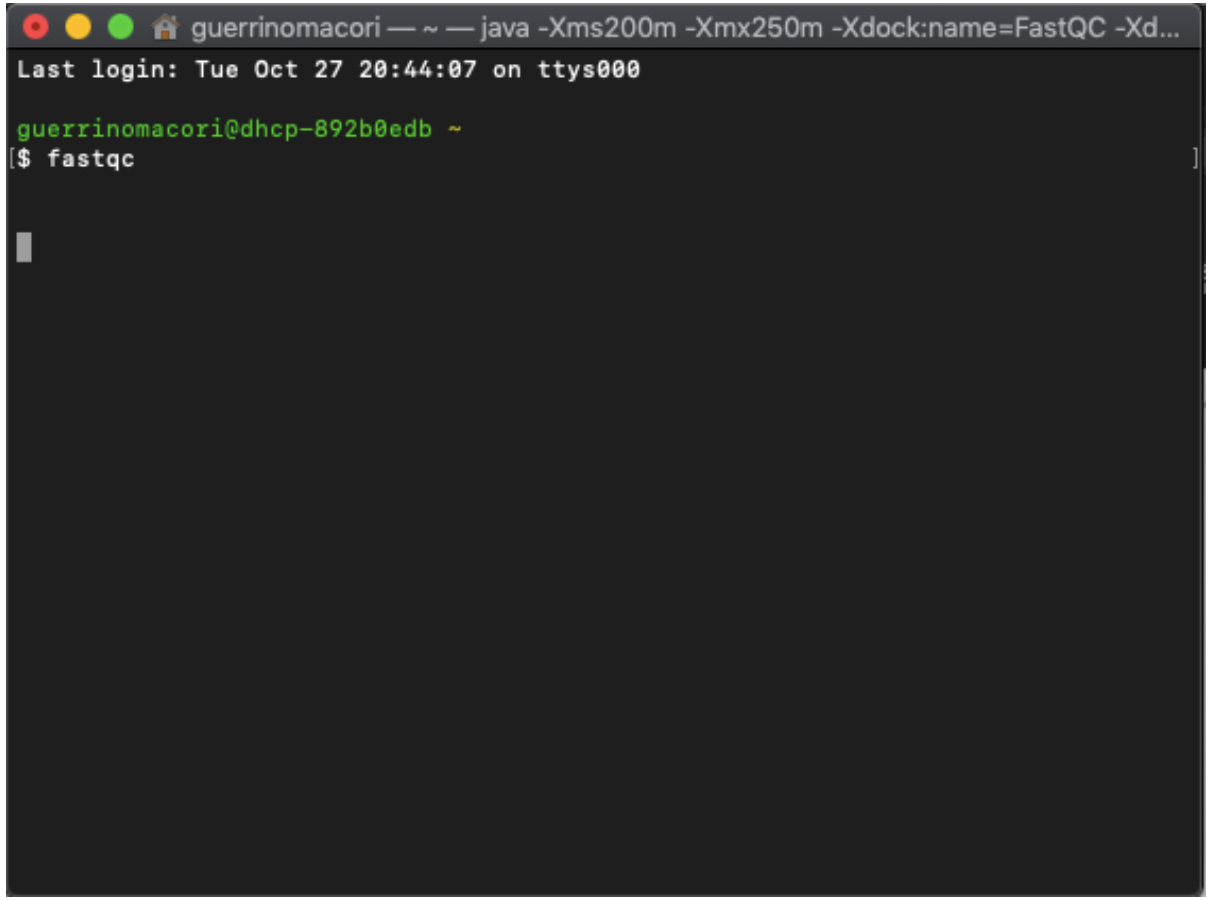
```
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN
CCTGCGCGGTCGCGACTACGACCTGGCGCTCGCCGATCATGGCTGGCACTGGCGCGGGN
+
CCCCCDGGGGGFEG7FGGGGGGGGGGGFFGC@GGGGGGEGGGGFGGCGFFF8FCFGGDF#:A
```

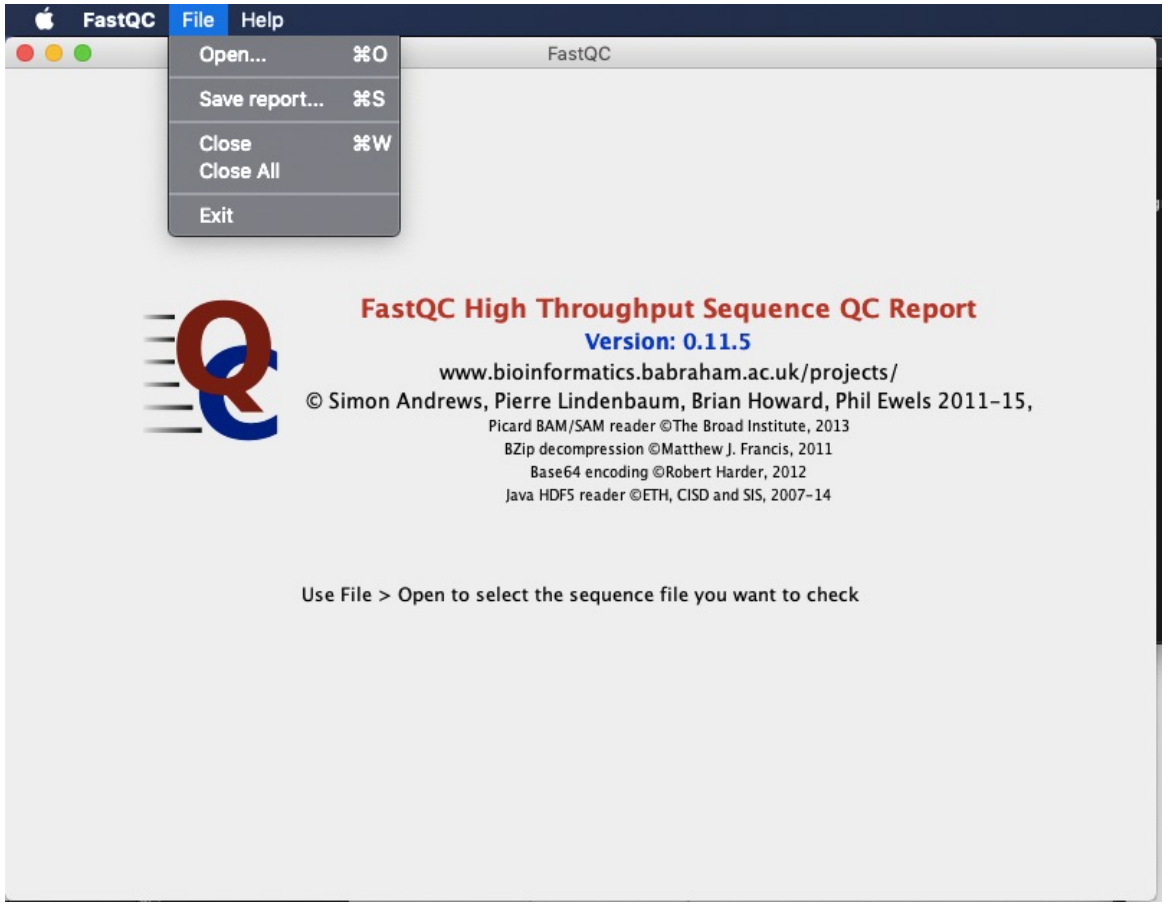
-> Sequence name
-> Sequence bases
-> Quality line break
-> Quality scores

@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN	the unique instrument name
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN	the run id
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN	the flowcell id
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN	flowcell lane
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN	tile number within the flowcell lane
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN	'x'-coordinate of the cluster within the tile
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN	'y'-coordinate of the cluster within the tile
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN	the member of a pair
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN	Yes or Not if the read fails
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN	0 the sequence NOT identified as control (PhiX) >0 sis a control
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN	index sequence

Quality assessment raw data

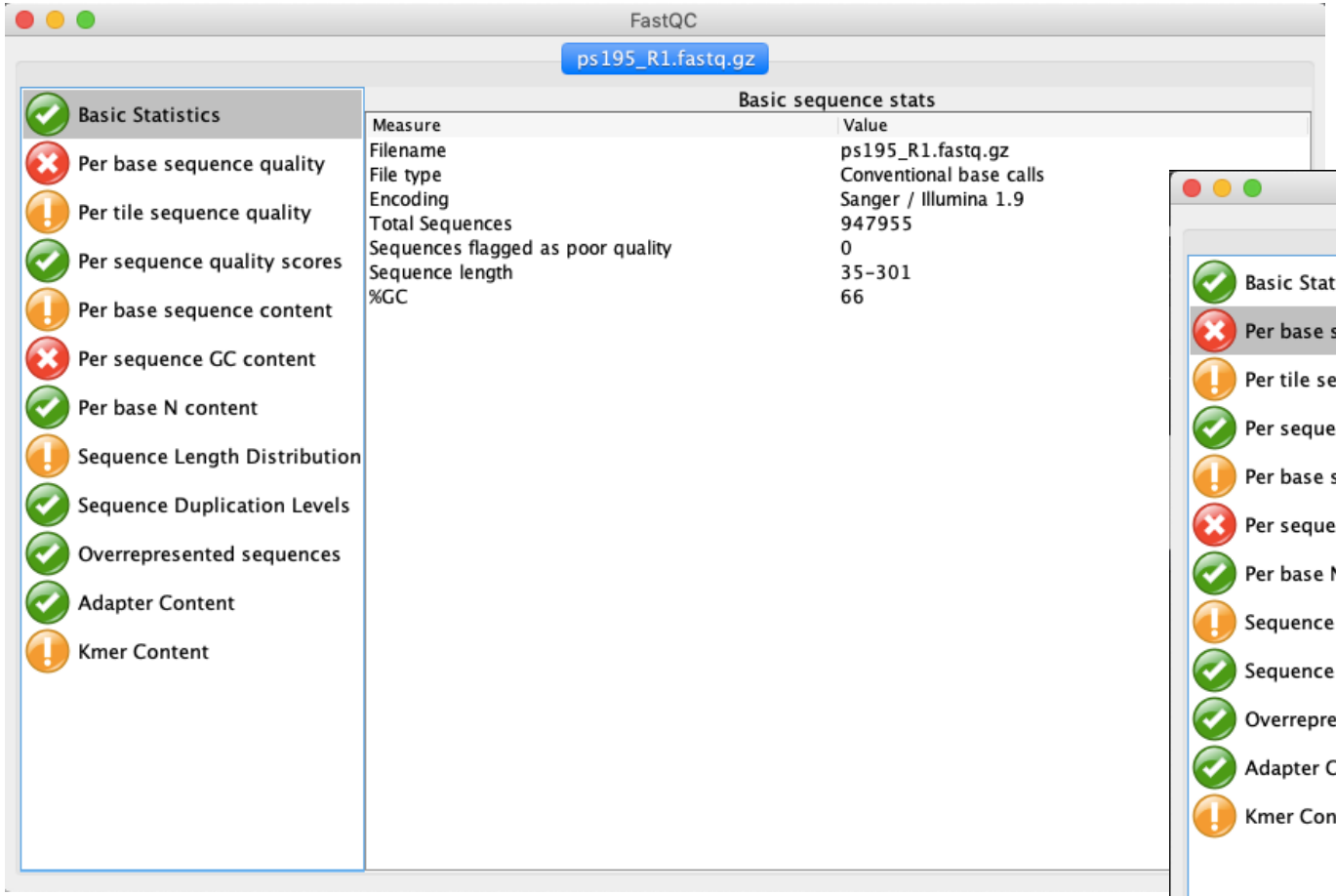
Fastq

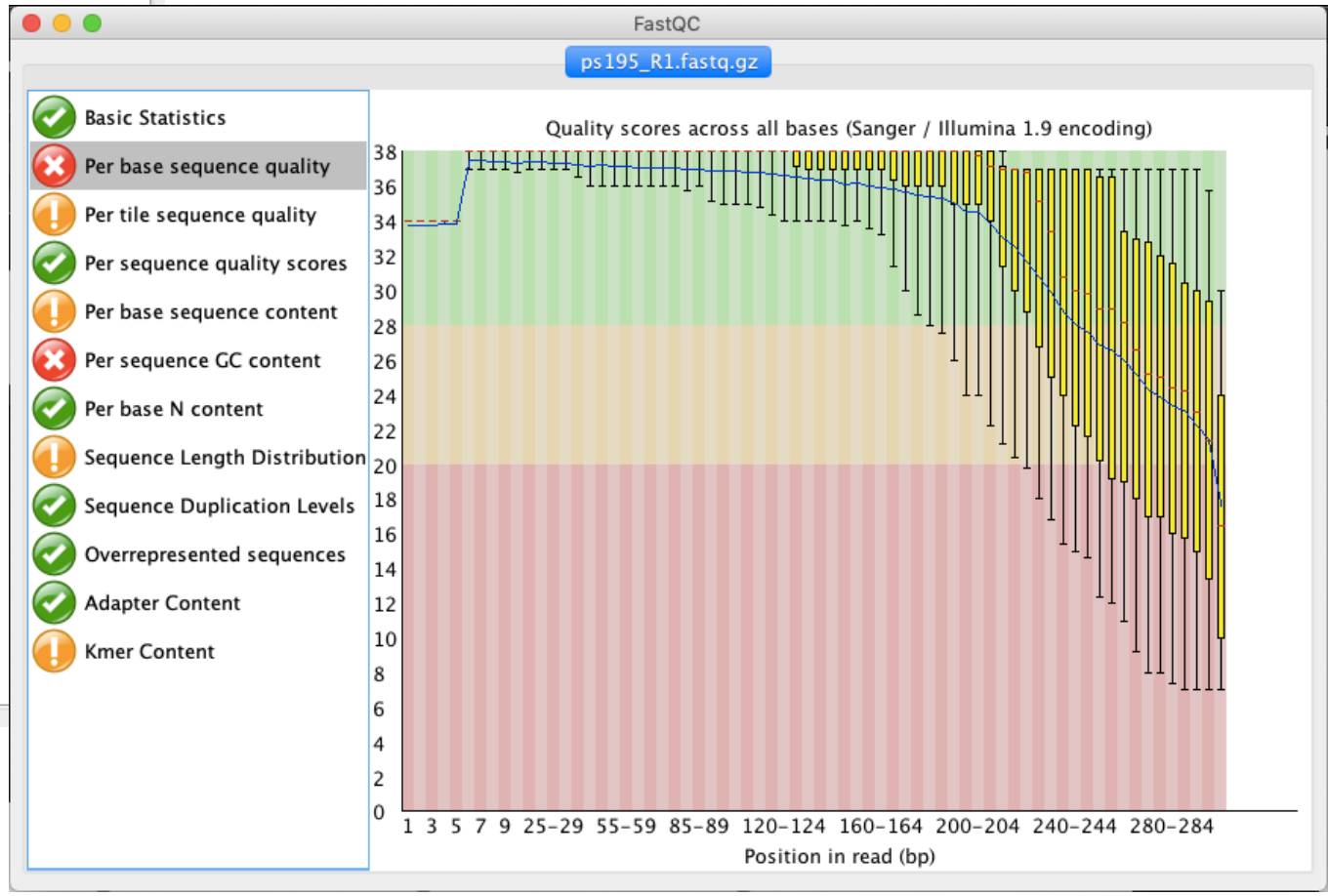




Quality assessment raw data

Fastq





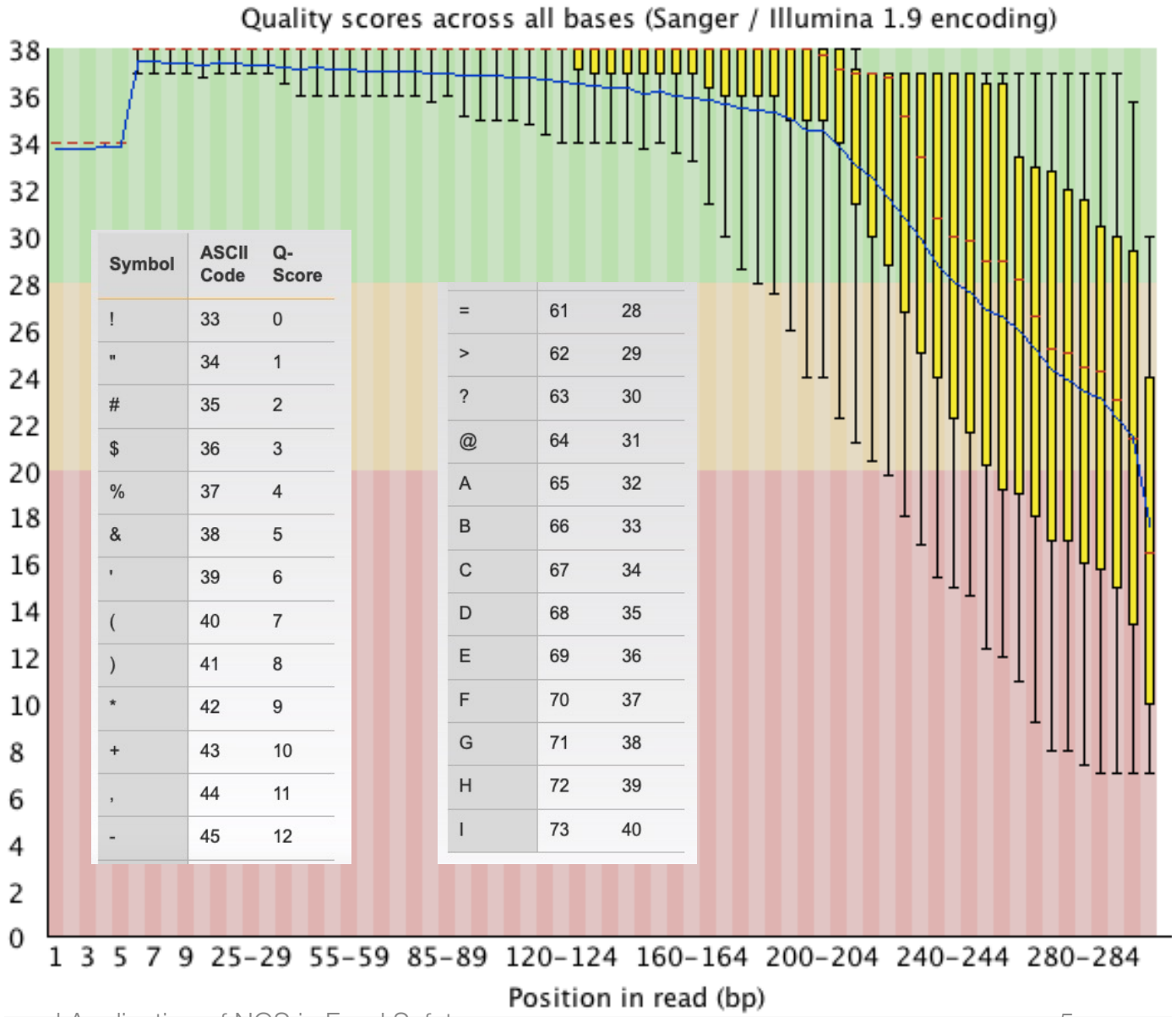
Quality assessment raw data

Fastq

A quality value Q is an integer mapping of p
(i.e., the probability that the corresponding base call is incorrect).

FASTQ quality scores are ASCII characters
– ASCII - American Standard Code for Information Interchange
– letters {A-Z, a-z}, numbers {0-9}, special characters {@ £ # $ % ^ & * ( ) , . ? “ / | ~ ` { } }
– Each character has a numeric whole number associated with it
• When people speak of quality scores, it is typically in the region of 0 to 40
– For example, trimming the ends of reads below Q25
• To translate from ASCII characters to Phred scale quality scores – YOU SUBTRACT 33 (-33)
• But … what does a quality score actually mean???

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|    |    |    |    |    |    |    |    |
0....5...10...15...20...25...30...35...40
|    |    |    |    |    |    |    |    |
worst.................................best
```

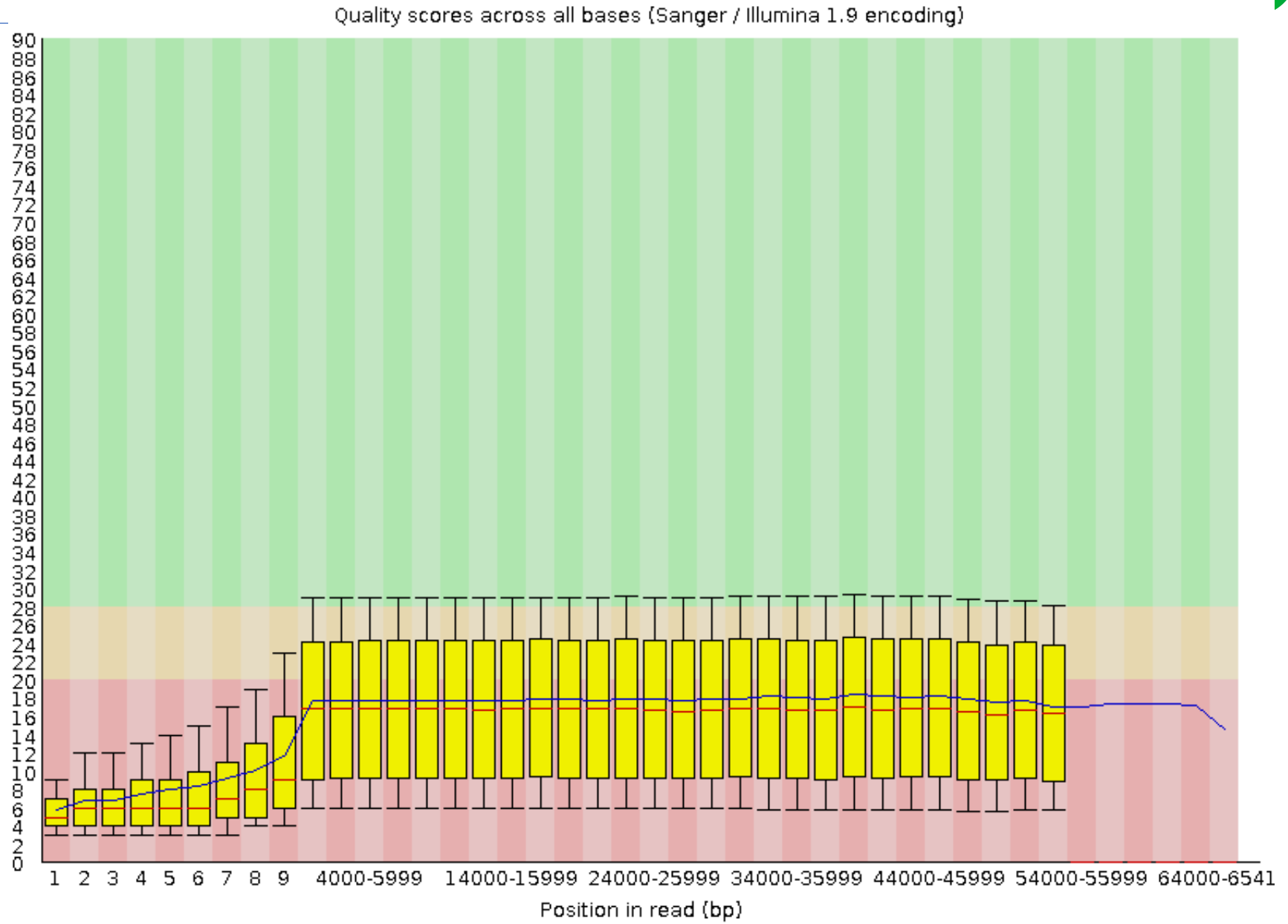


Symbol	ASCII Code	Q-Score
!	33	0
"	34	1
#	35	2
$	36	3
%	37	4
&	38	5
'	39	6
(	40	7
)	41	8
*	42	9
+	43	10
,	44	11
-	45	12

=	61	28
>	62	29
?	63	30
@	64	31
A	65	32
B	66	33
C	67	34
D	68	35
E	69	36
F	70	37
G	71	38
H	72	39
I	73	40

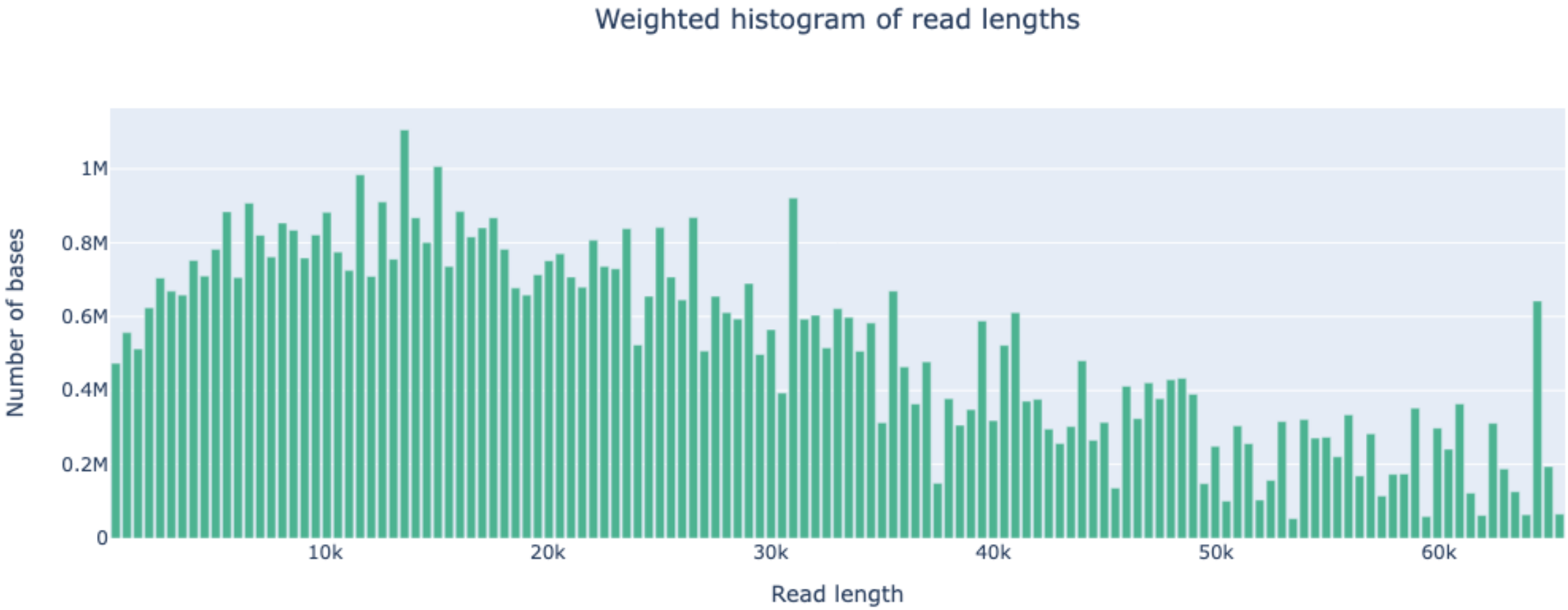
Quality assessment raw data

Fastq



Quality assessment raw data

Fastq quality check tool NanoPlot



This plot shows the distribution of read lengths, weighted by the number of bases.

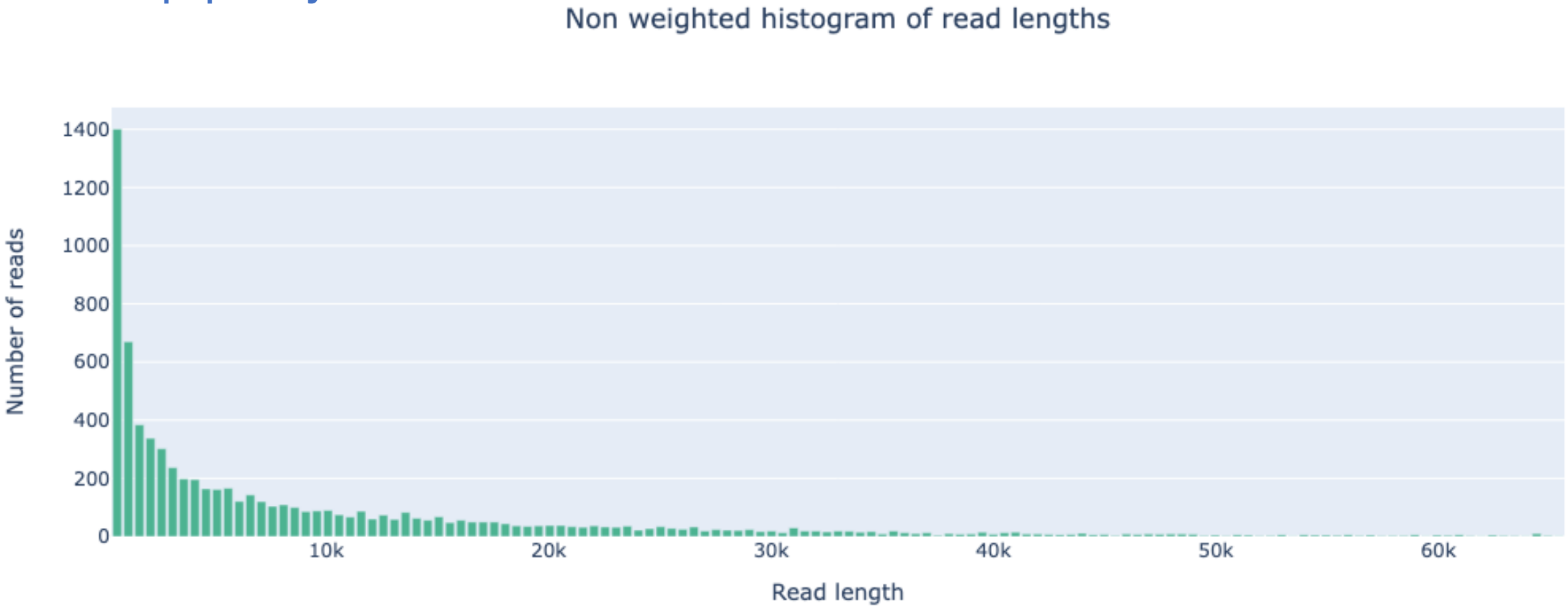
Quality assessment raw data

Fastq quality check tool NanoPlot

This plot is particularly useful because it illustrates the relationship between read length and quality. Check f longer reads tend to have lower quality, which is often the case with Nanopore data. The dot representation gives a clear visualization of each individual read.

Quality assessment raw data

Fastq quality check tool NanoPlot



Unlike the weighted version, this plot shows the distribution of read lengths without considering the number of bases, offering a view of the sheer number of reads of different lengths.

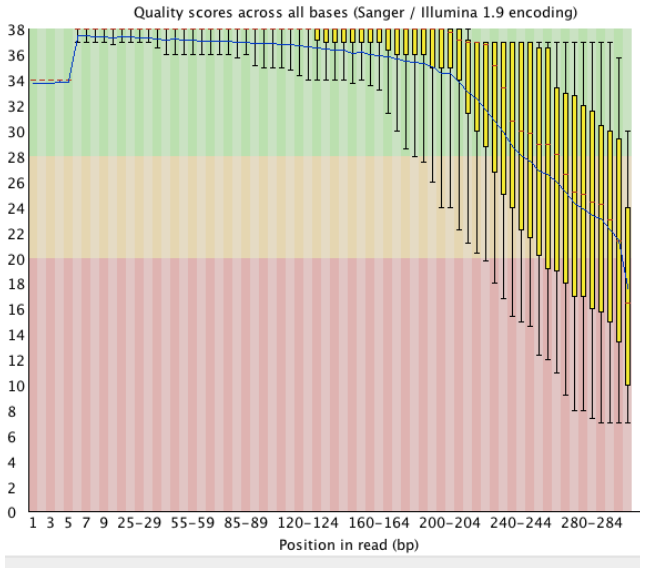
Quality assessment raw data

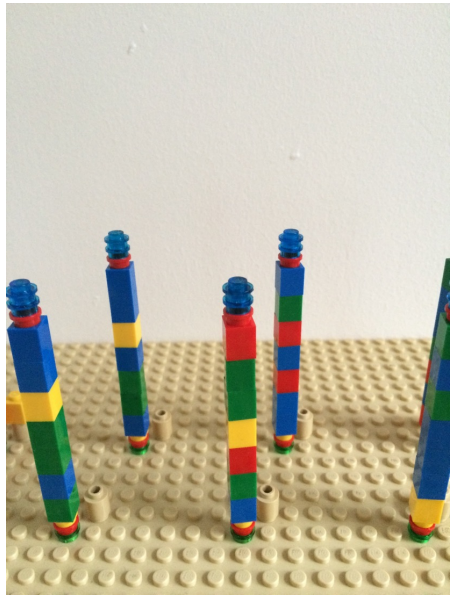
Fastq

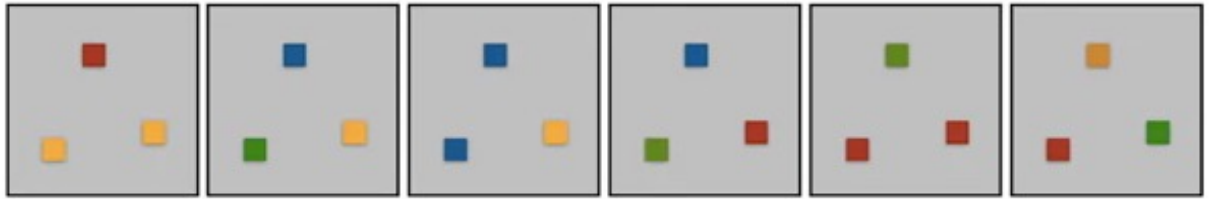
- FASTQ Quality Scores

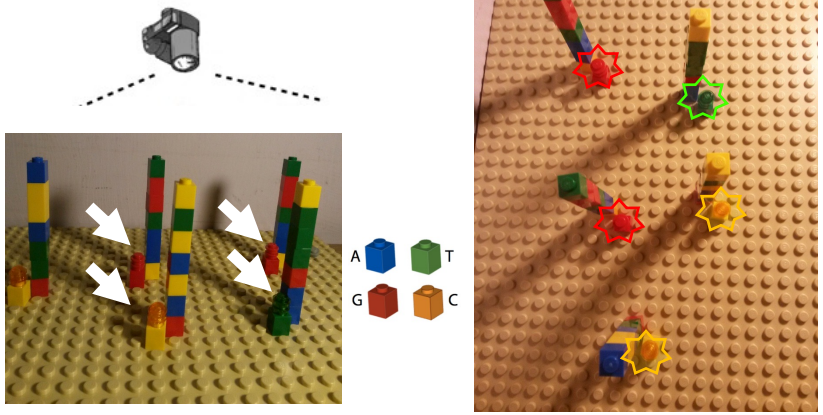
What does the quality mean?

- Quality scores tend to decrease along the read
- The read represents the sequence of a Cluster – Cluster is comprised of ~1,000 DNA molecules
- As the sequencing progresses more and more of the DNA molecules in the cluster get out of sync
  - Phasing: falling behind: missing and incorporation cycle, incomplete removal of the 3' terminators/fluorophores
  - Pre-phasing: jumping ahead: incorporation of multiple bases in a cycle due to NTs without effective 3' blocking



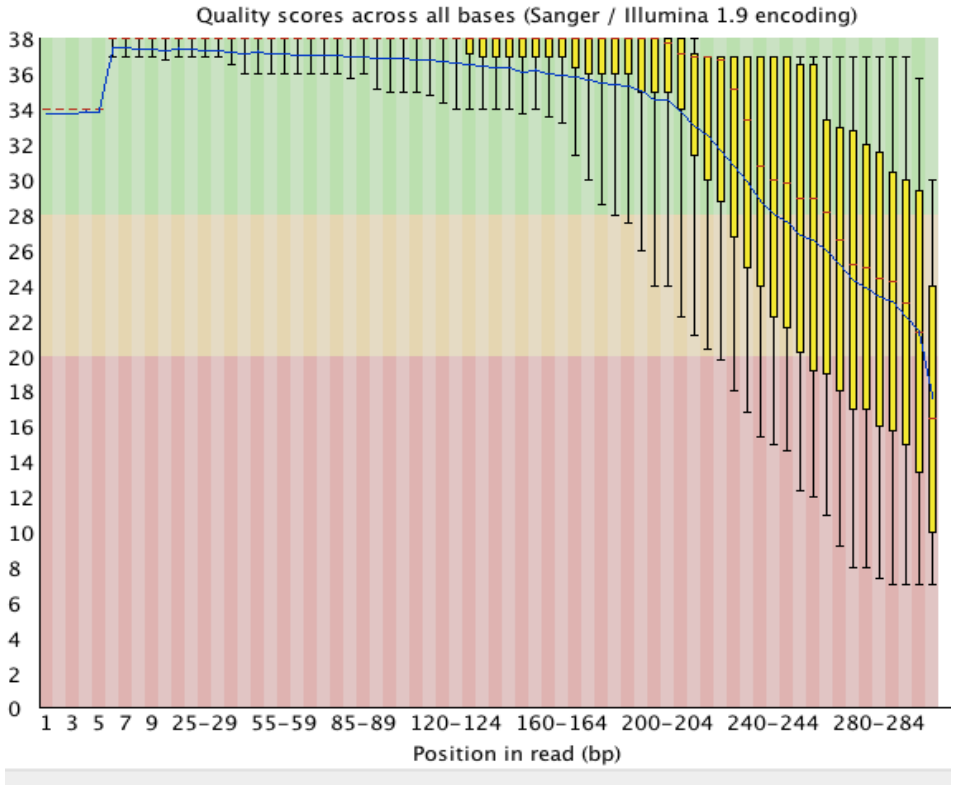


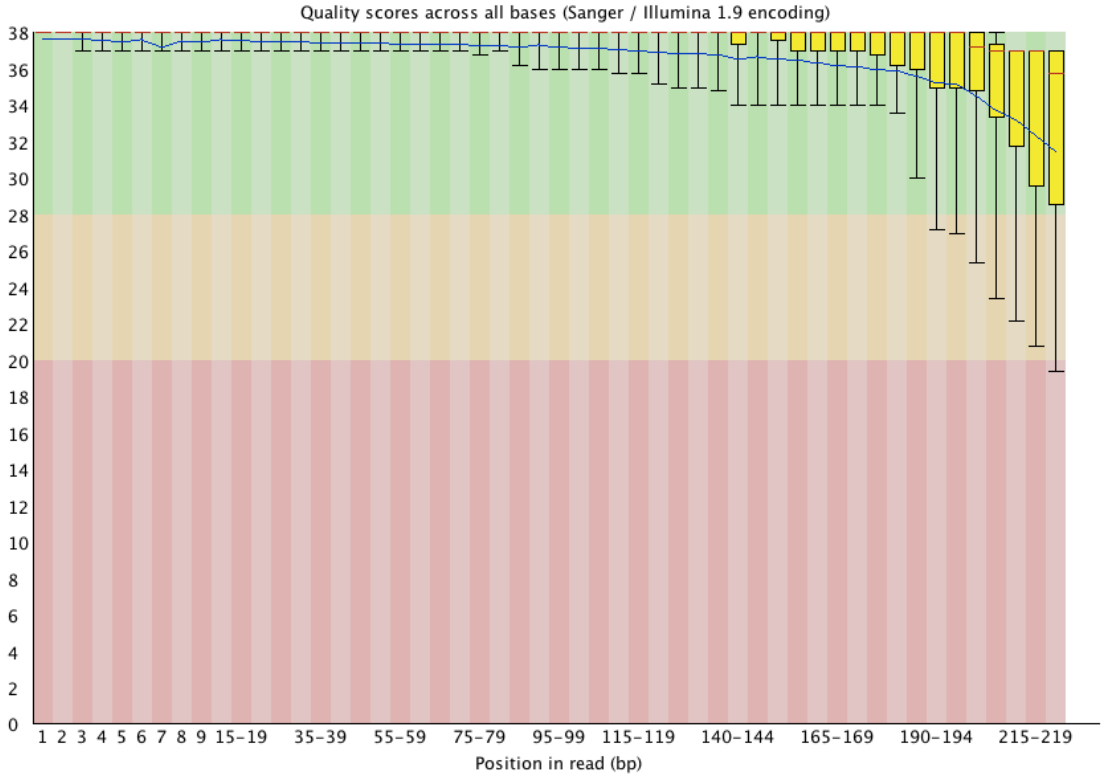




Quality assessment raw data

Fastq





ps195_R1.fastq | ps195t1_R1.fastq.gz

Basic sequence stats

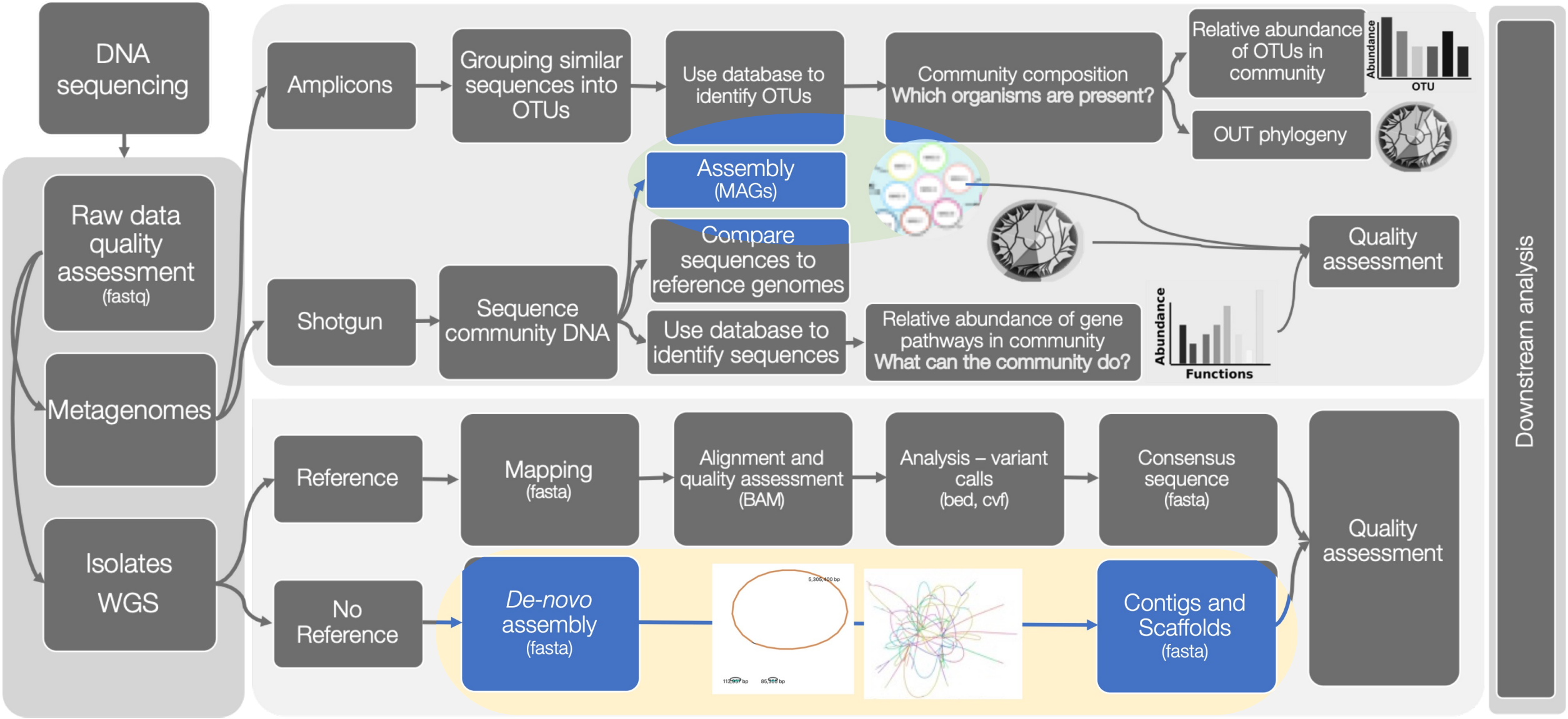
Measure	Value
Filename	ps195_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	947955
Sequences flagged as poor quality	0
Sequence length	35-301
%GC	66

ps195_R1.fastq | ps195t1_R1.fastq.gz

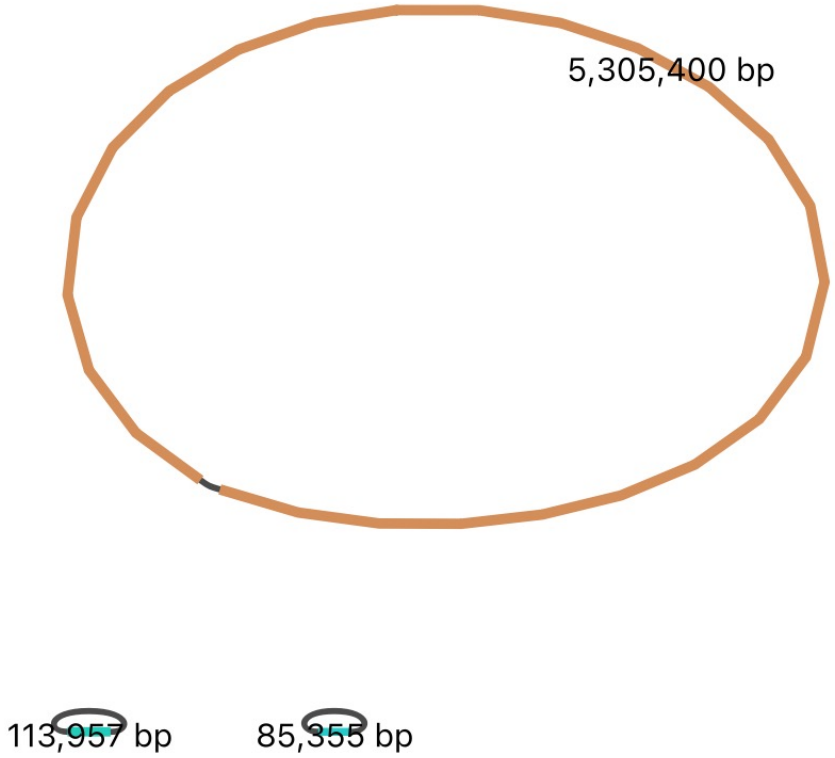
Basic sequence stats

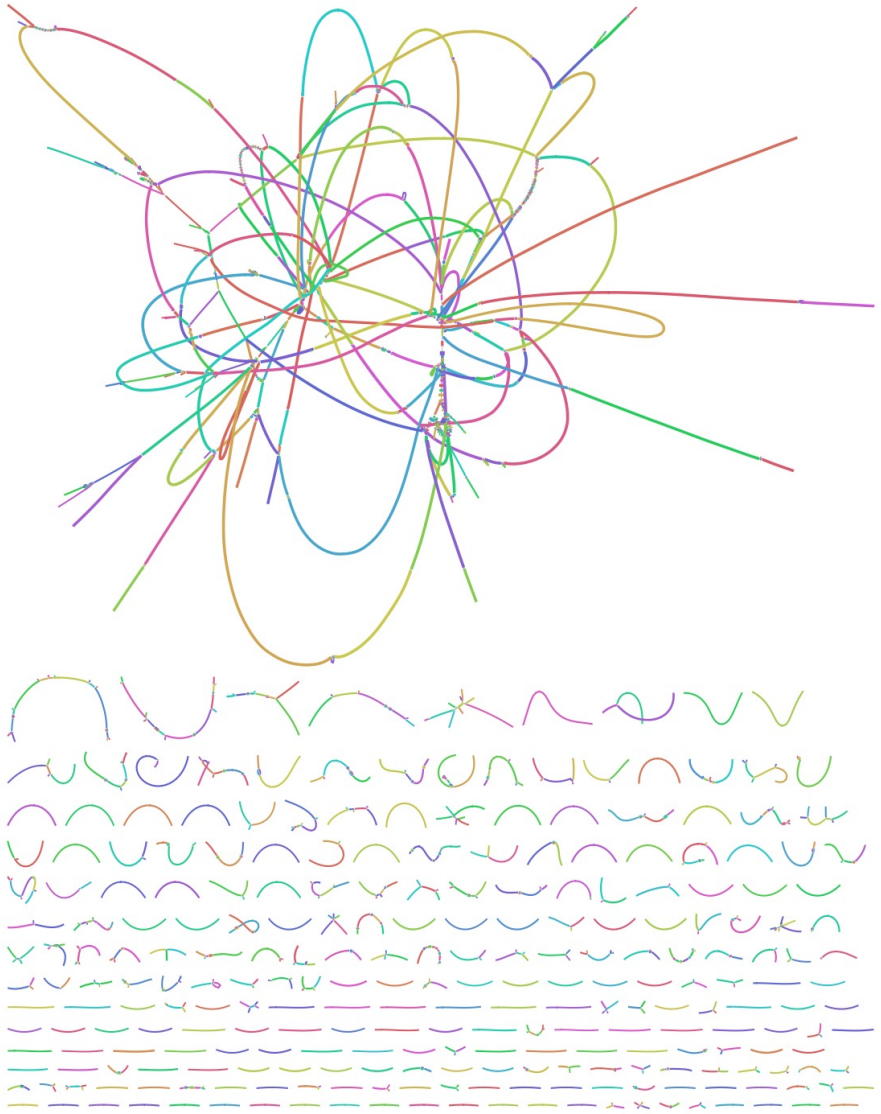
Measure	Value
Filename	ps195t1_R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	878816
Sequences flagged as poor quality	0
Sequence length	37-224
%GC	66

Applications - Food Safety Bioinformatics

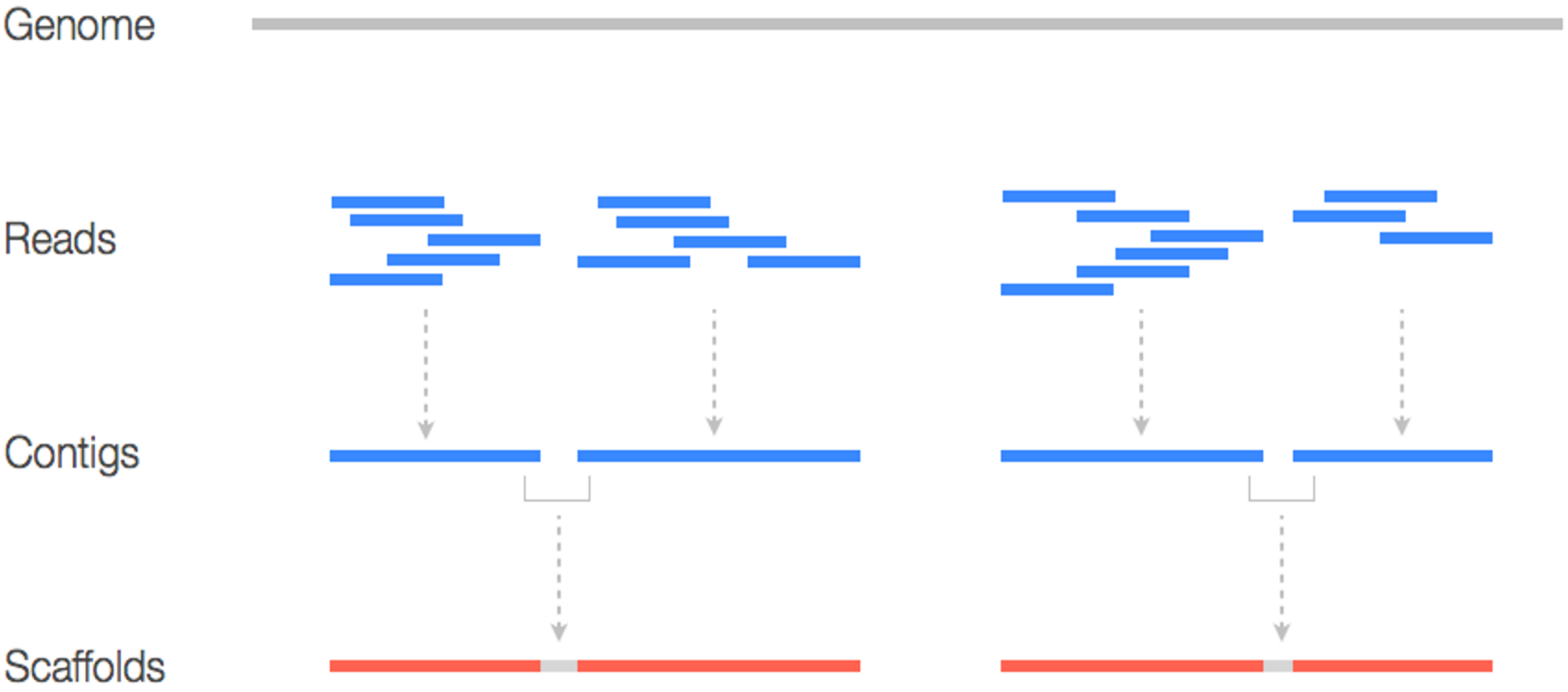


Applications - Food Safety Bioinformatics - Assembly





Applications - Food Safety Bioinformatics - Assembly



Applications - Food Safety Bioinformatics - Assembly

TATTCTTCCACGTAGGGCCTTCCACGCTTCG

Applications - Food Safety Bioinformatics - Assembly

TATTCTTC

CTTCCACG

CACGTAGG

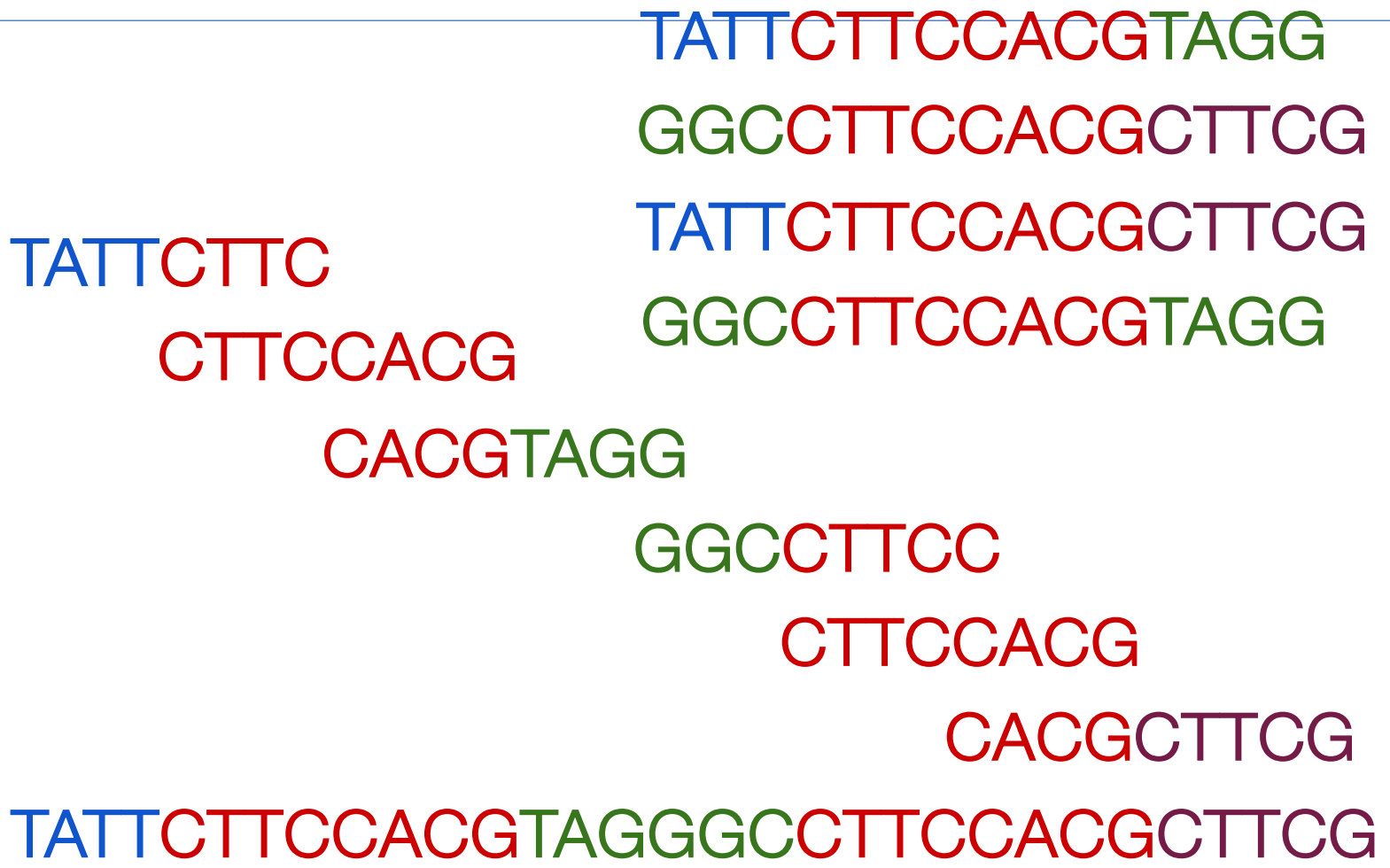
GGCCTTCC

CTTCCACG

CACGCTTCG

TATTCTTCCACGTAGGGCCTTCCACGCTTCG

Applications - Food Safety Bioinformatics - Assembly



Applications - Food Safety Bioinformatics - Assembly

Genomic repeats

TATTCTTCCACGTAGG

GGCCTTCCACGCTTCG

TATTCTTCCACGCTTCG

GGCCTTCCACGTAGG

Applications - Food Safety Bioinformatics - Assembly

Genomic repeats

TATTCTTCCACGTAGG

ACGTAGGGCCTT

GCCTTCCACGCCTTCG

TATTCTTCCACGTAGGGCCTTCCACGCTTCG

Applications - Food Safety Bioinformatics - Assembly

Two different methods

- **Reference guided** – When you have “map” of final product, and try to match your pieces together to look like final product
    - Jigsaw puzzle – you have picture and put pieces together
    - Lyrics of song – you know words to song, and put words in correct order

- *De novo* – Put pieces together by what “makes sense”
    - Jigsaw puzzle – you may not have picture, but can put pieces together by what fits together
    - Lyrics of song – you may not know words of song, but can figure out sentences from the words.

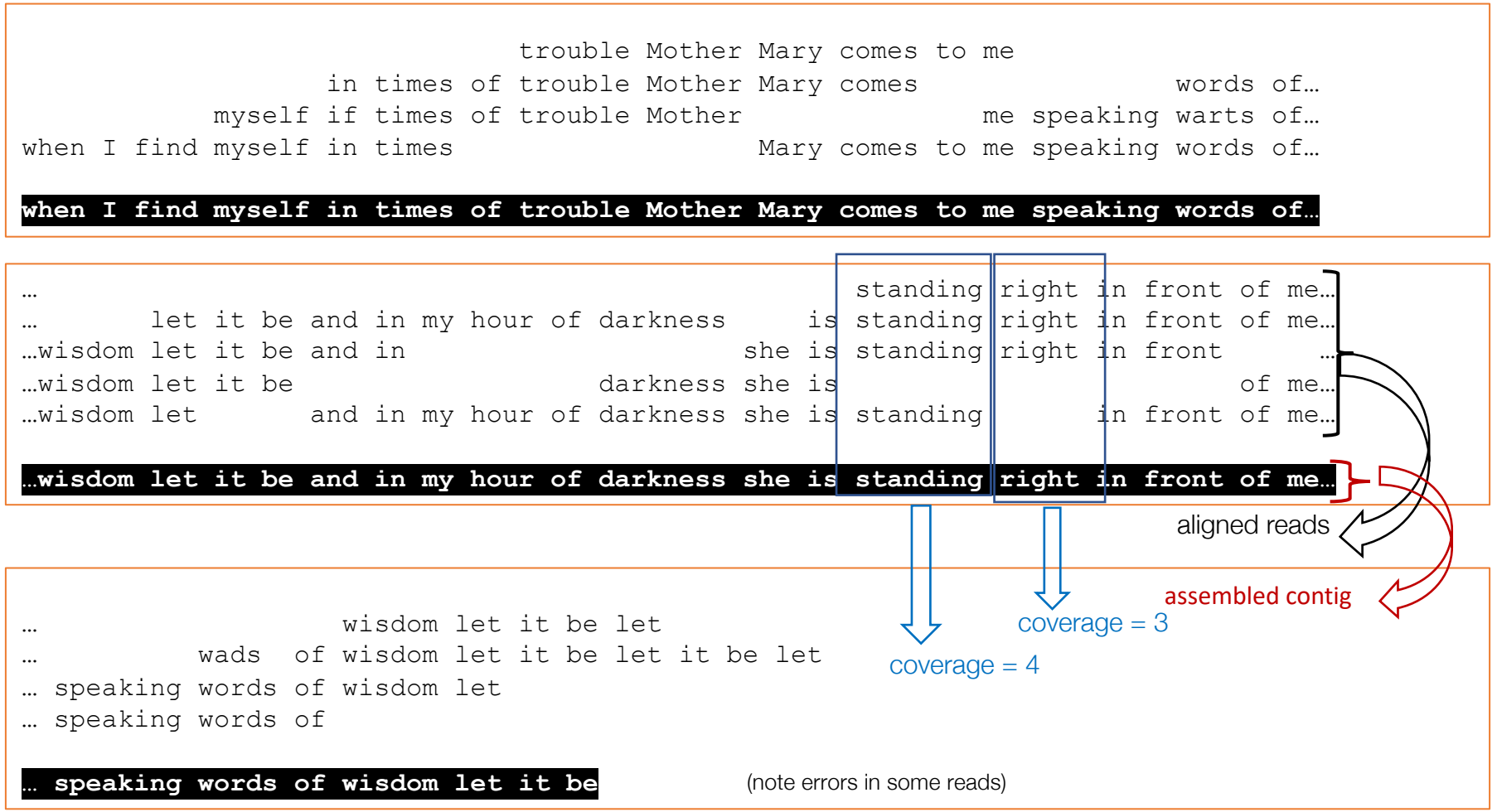
Applications - Food Safety Bioinformatics - Assembly

Example: *De novo* assembly of lyrics, using “reads” of 4 to 6 words (each word is a base pair/amino acid):

1: yeah there will be an
2: tomorrow let it be o will
3: let it be and though they
4: me speaking words of wisdom
5: let it be let
6: the night is cloudy there
7: be let it be let it
8: be whisper words of wisdom
9: on me shine until tomorrow
10: let it be let it be let
11: be and when the broken hearted
12: answer let it be and though
13: it be let it be
14: and though the night is
15: the broken hearted people living
...

Applications - Food Safety Bioinformatics - Assembly

Assembly “pile-up”



Applications - Food Safety Bioinformatics - Assembly

The repeats

```
                    let it be let it
                it be let it be let                 words …
            let it be let it be     it be speaking words …
… of wisdom let it be     it be let it be                    ?

… of wisdom let it be let it be let it be speaking words …
```

```
                    let it be let it        it be speaking words …
                it be let it be let    be let it be                 ?
            let it be let it be     it be let it
… of wisdom let it be     it be let it be                words …

… of wisdom let it be let it be let it be let it be speaking words …
```

```
                                                                 ?
                    let it be let it           be let it be
                it be let it be let    be let it be let it              words …
            let it be let it be     it be let it be     it be speaking words …
… of wisdom let it be     it be let it be     it be let it be

… of wisdom let it be let it be let it be let it be let it be speaking words …
```

Applications - Food Safety Bioinformatics - Assembly

Final assembly

3 contigs:

```
and though the night is cloudy there is still a light that shines on me
shine until tomorrow let it be o will I make up to the sound of music
Mother Mary comes to me speaking words of wisdom
```

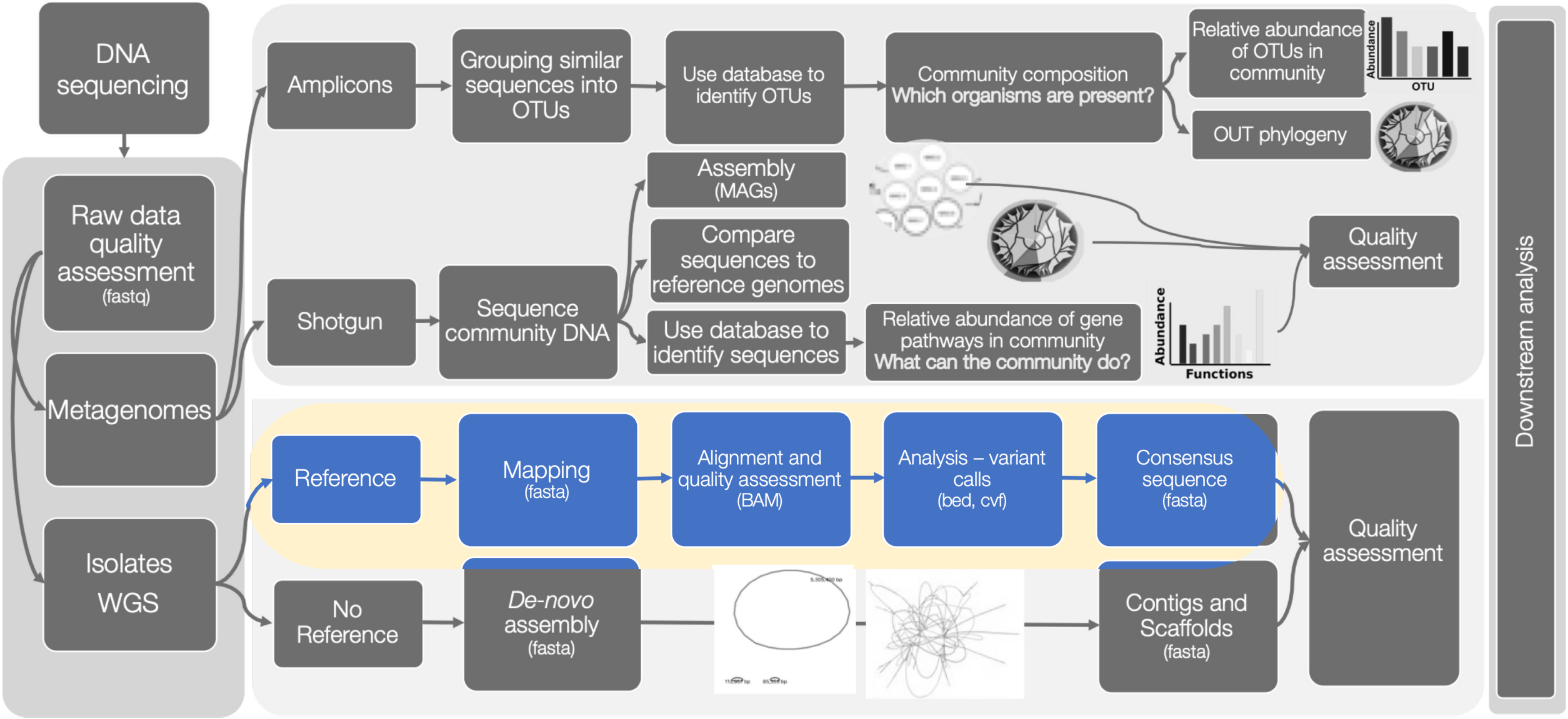
```
when I find myself in times of trouble Mother Mary comes to me speaking
words of wisdom let it be and in my hour of darkness she is standing right
in front of me speaking words of wisdom
```

```
whisper words of wisdom let it be and when the broken hearted people living
in the world agree there will be an answer let it be and though they may be
parted there is still a chance that they will see there will be an answer
```

Many unused reads:

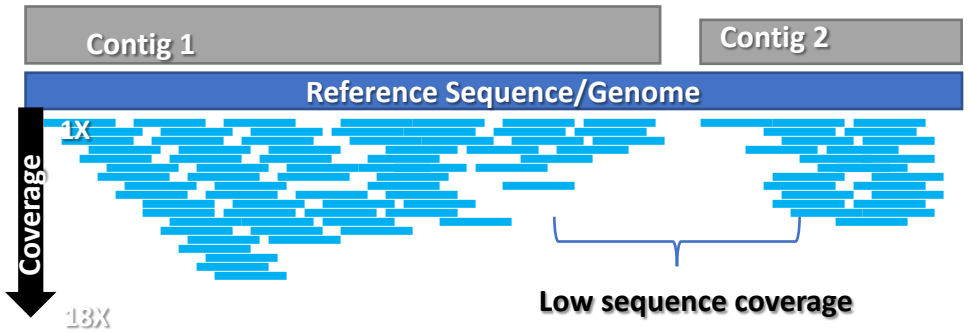
```
be let it be
let it be let it
it be let it be
let it be let it be let
be let it be let
be let it be
it be let it be let
…
```

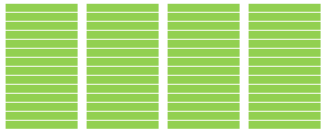
Applications - Food Safety Bioinformatics



Applications - Food Safety Bioinformatics - Assembly

Reference-Guided (Mapped) Assembly





UNMAPPED READS

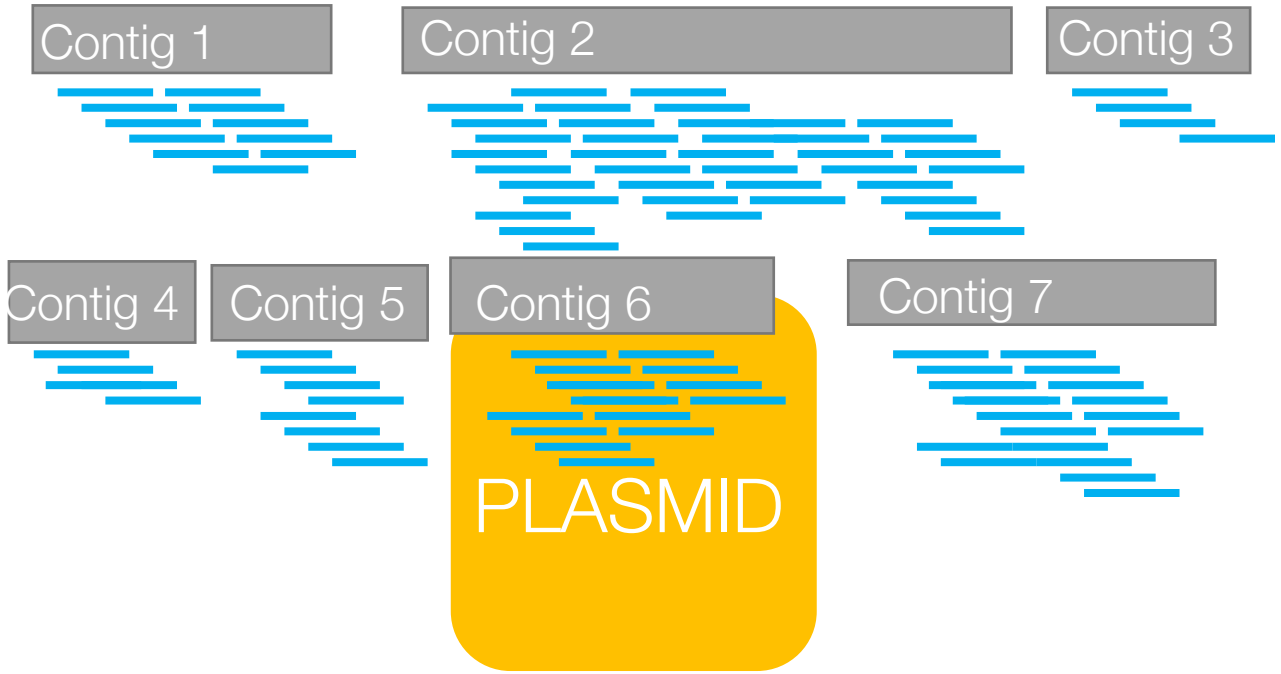
1. **Sequences not present in the reference.**
2. **Plasmids or other extrachromosomal.**
3. **DNA Structural Variation/Rearrangement**

ADVANTAGES: Relatively fast, well-suited to highly-conserved genomes.
DISADVANTAGES: Issues with high diversity, mobile elements

Example software: BWA (https://github.com/lh3/bwa)
breseq (https://github.com/barricklab/breseq)

Applications - Food Safety Bioinformatics - Assembly

De-novo assembly



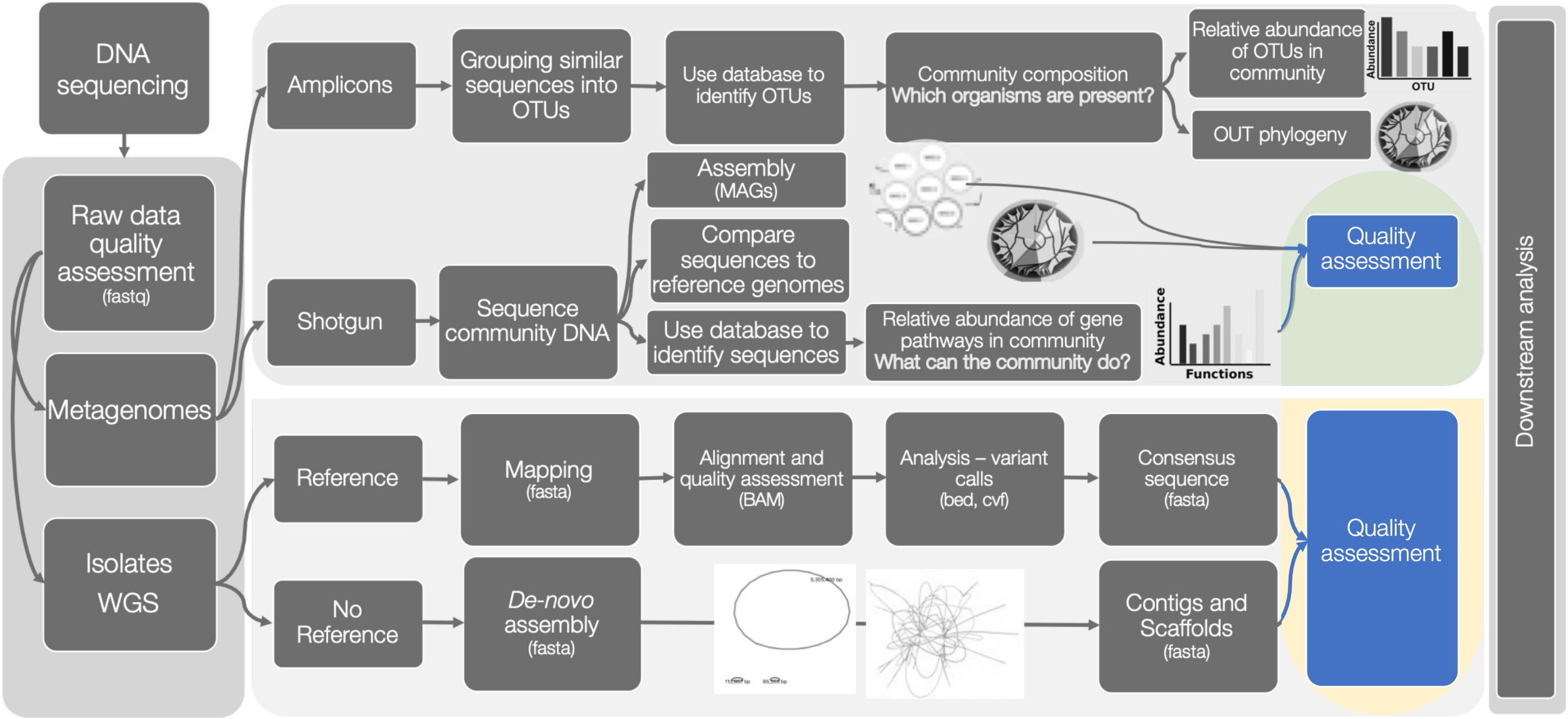
ADVANTAGES: Reference agnostic: assembles all the reads it can. Various algorithms.
DISADVANTAGES: Doesn't always get things right. Particularly with complex repeats.

Example software: SPAdes (http://bioinf.spbau.ru/spades)
List: https://en.wikipedia.org/wiki/Sequence_assembly#Available_assemblers

Longer vs Shorter Reads

- Read length depends on instruments used
- Depending on goal, one may be better than other
- Shorter reads
  - Pro – economic
  - Con – repeats, missing areas
- Long reads
  - Pro - provide better coverage and higher consensus
  - Con - expensive

Applications - Food Safety Bioinformatics



Quality of assemblies Basic evaluation

06 November 2022, Sunday, 11:46:09

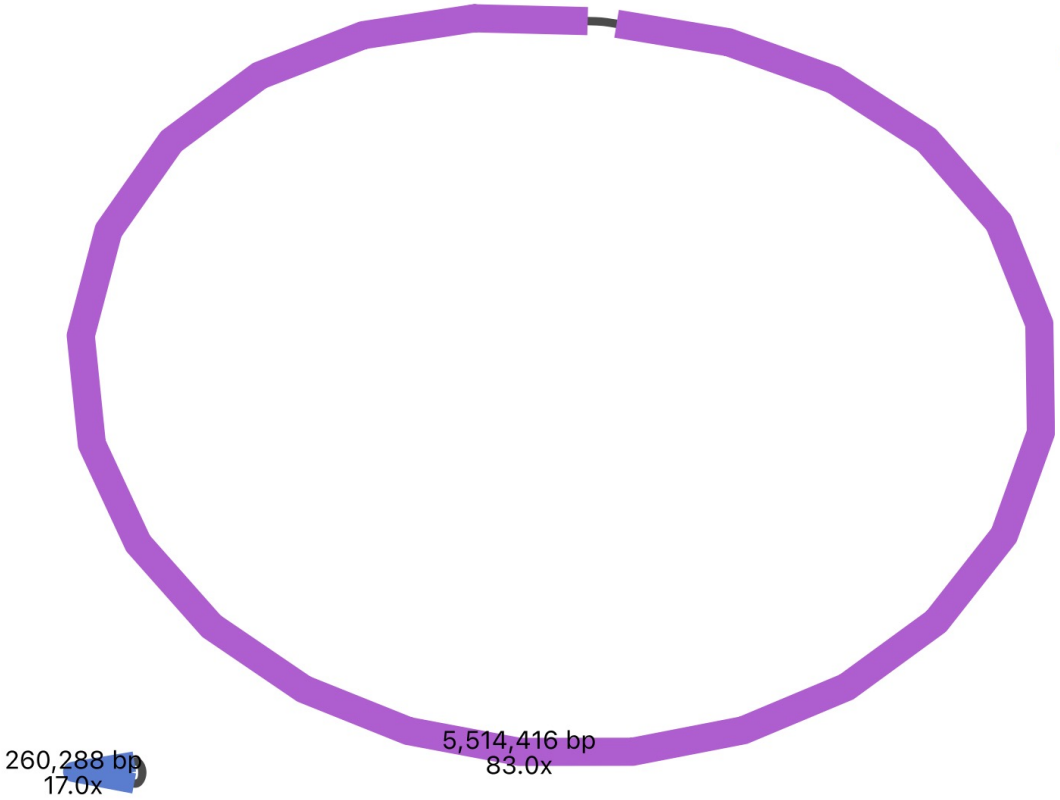
View in Icarus contig browser

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Worst Median Best ☑ Show heatmap

Statistics without reference	CFS1073_ONT	CFS1073_ill	CFS1080_ONT	CFS1080_ill	CFS4321_ONT	CFS4321_ill	CFS4331_ONT	CFS4331_ill
# contigs	2	1628	3	587	2	16	2	17
# contigs (>= 0 bp)	2	1628	3	587	2	16	2	17
# contigs (>= 1000 bp)	2	189	3	481	2	12	2	12
# contigs (>= 5000 bp)	2	29	3	307	2	10	2	10
# contigs (>= 10000 bp)	2	26	3	224	2	10	2	10
# contigs (>= 25000 bp)	2	25	3	110	2	10	2	10
# contigs (>= 50000 bp)	2	19	2	45	2	8	2	8
Largest contig	5 514 416	743 105	5 513 354	328 825	5 420 830	3 157 918	5 417 151	3 157 653
Total length	5 774 704	6 872 116	9 615 475	9 685 624	5 505 270	5 496 177	5 524 377	5 496 378
Total length (>= 0 bp)	5 774 704	6 872 116	9 615 475	9 685 624	5 505 270	5 496 177	5 524 377	5 496 378
Total length (>= 1000 bp)	5 774 704	5 957 976	9 615 475	9 610 642	5 505 270	5 493 381	5 524 377	5 492 464
Total length (>= 5000 bp)	5 774 704	5 728 977	9 615 475	9 172 552	5 505 270	5 486 975	5 524 377	5 486 709
Total length (>= 10000 bp)	5 774 704	5 708 717	9 615 475	8 580 011	5 505 270	5 486 975	5 524 377	5 486 709
Total length (>= 25000 bp)	5 774 704	5 690 225	9 615 475	6 777 244	5 505 270	5 486 975	5 524 377	5 486 709
Total length (>= 50000 bp)	5 774 704	5 476 870	9 579 472	4 536 919	5 505 270	5 425 680	5 524 377	5 425 414
N50	5 514 416	382 077	5 513 354	46 677	5 420 830	3 157 918	5 417 151	3 157 653
N90	5 514 416	745	4 066 118	8704	5 420 830	359 344	5 417 151	359 344
auN	5 277 592	355 297	4 880 849	72 680	5 338 980	2 017 471	5 314 087	2 017 202
L50	1	7	1	52	1	1	1	1
L90	1	459	2	240	1	5	1	5
GC (%)	57.05	56.9	49.5	49.79	57.49	57.5	57.46	57.5
Mismatches								
# N's per 100 kbp	0	0	0	1.01	0	1.76	0	1.8
# N's	0	0	0	98	0	97	0	99

Quality of assemblies Basic evaluation



2 contigs

06 November 2022, Sunday, 11:46:09

View in Icarus contig browser

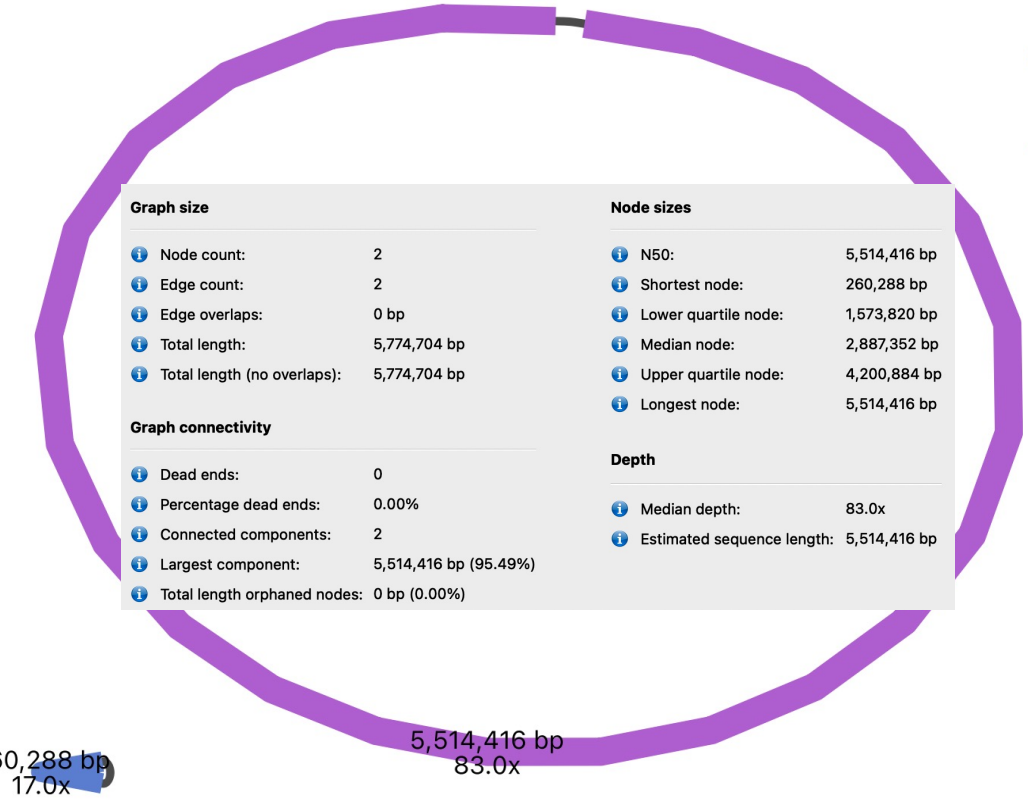
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Show heatmap

Worst Median Best

Statistics without reference	CFS1073_ONT	CFS1073_ill	CFS1080_ONT	CFS1080_ill	CFS4321_ONT	CFS4321_ill	CFS4331_ONT	CFS4331_ill
# contigs	2	1628	3	587	2	16	2	17
# contigs (>= 0 bp)	2	1628	3	587	2	16	2	17
# contigs (>= 1000 bp)	2	189	3	481	2	12	2	12
# contigs (>= 5000 bp)	2	29	3	307	2	10	2	10
# contigs (>= 10000 bp)	2	26	3	224	2	10	2	10
# contigs (>= 25000 bp)	2	25	3	110	2	10	2	10
# contigs (>= 50000 bp)	2	19	2	45	2	8	2	8
Largest contig	5 514 416	743 105	5 513 354	328 825	5 420 830	3 157 918	5 417 151	3 157 653
Total length	5 774 704	6 872 116	9 615 475	9 685 624	5 505 270	5 496 177	5 524 377	5 496 378
Total length (>= 0 bp)	5 774 704	6 872 116	9 615 475	9 685 624	5 505 270	5 496 177	5 524 377	5 496 378
Total length (>= 1000 bp)	5 774 704	5 957 976	9 615 475	9 610 642	5 505 270	5 493 381	5 524 377	5 492 464
Total length (>= 5000 bp)	5 774 704	5 728 977	9 615 475	9 172 552	5 505 270	5 486 975	5 524 377	5 486 709
Total length (>= 10000 bp)	5 774 704	5 708 717	9 615 475	8 580 011	5 505 270	5 486 975	5 524 377	5 486 709
Total length (>= 25000 bp)	5 774 704	5 690 225	9 615 475	6 777 244	5 505 270	5 486 975	5 524 377	5 486 709
Total length (>= 50000 bp)	5 774 704	5 476 870	9 579 472	4 536 919	5 505 270	5 425 680	5 524 377	5 425 414
N50	5 514 416	382 077	5 513 354	46 677	5 420 830	3 157 918	5 417 151	3 157 653
N90	5 514 416	745	4 066 118	8704	5 420 830	359 344	5 417 151	359 344
auN	5 277 592	355 297	4 880 849	72 680	5 338 980	2 017 471	5 314 087	2 017 202
L50	1	7	1	52	1	1	1	1
L90	1	459	2	240	1	5	1	5
GC (%)	57.05	56.9	49.5	49.79	57.49	57.5	57.46	57.5
Mismatches								
# N's per 100 kbp	0	0	0	1.01	0	1.76	0	1.8
# N's	0	0	0	98	0	97	0	99

Quality of assemblies Basic evaluation



06 November 2022, Sunday, 11:46:09

View in Icarus contig browser

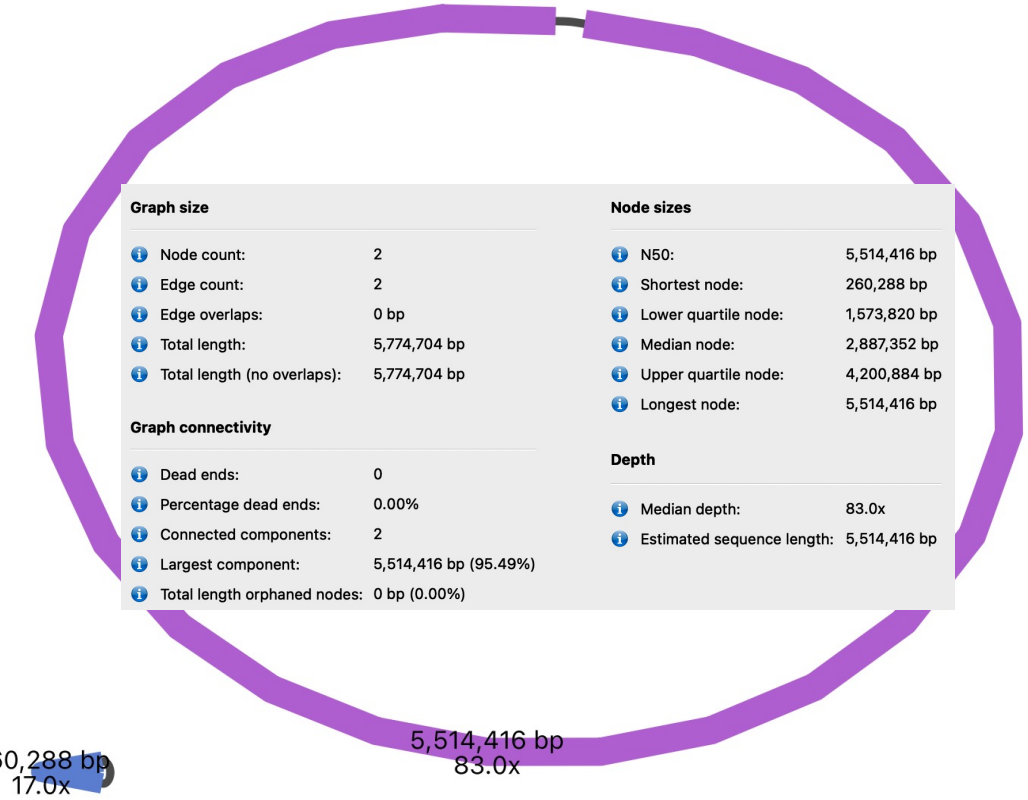
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Worst Median Best Show heatmap

Statistics without reference	CFS1073_ONT	CFS1073_ill	CFS1080_ONT	CFS1080_ill	CFS4321_ONT	CFS4321_ill	CFS4331_ONT	CFS4331_ill
# contigs	2	1628	3	587	2	16	2	17
# contigs (>= 0 bp)	2	1628	3	587	2	16	2	17
# contigs (>= 1000 bp)	2	189	3	481	2	12	2	12
# contigs (>= 5000 bp)	2	29	3	307	2	10	2	10
# contigs (>= 10000 bp)	2	26	3	224	2	10	2	10
# contigs (>= 25000 bp)	2	25	3	110	2	10	2	10
# contigs (>= 50000 bp)	2	19	2	45	2	8	2	8
Largest contig	5 514 416	743 105	5 513 354	328 825	5 420 830	3 157 918	5 417 151	3 157 653
Total length	5 774 704	6 872 116	9 615 475	9 685 624	5 505 270	5 496 177	5 524 377	5 496 378
Total length (>= 0 bp)	5 774 704	6 872 116	9 615 475	9 685 624	5 505 270	5 496 177	5 524 377	5 496 378
Total length (>= 1000 bp)	5 774 704	5 957 976	9 615 475	9 610 642	5 505 270	5 493 381	5 524 377	5 492 464
Total length (>= 5000 bp)	5 774 704	5 728 977	9 615 475	9 172 552	5 505 270	5 486 975	5 524 377	5 486 709
Total length (>= 10000 bp)	5 774 704	5 708 717	9 615 475	8 580 011	5 505 270	5 486 975	5 524 377	5 486 709
Total length (>= 25000 bp)	5 774 704	5 690 225	9 615 475	6 777 244	5 505 270	5 486 975	5 524 377	5 486 709
Total length (>= 50000 bp)	5 774 704	5 476 870	9 579 472	4 536 919	5 505 270	5 425 680	5 524 377	5 425 414
N50	5 514 416	382 077	5 513 354	46 677	5 420 830	3 157 918	5 417 151	3 157 653
N90	5 514 416	745	4 066 118	8704	5 420 830	359 344	5 417 151	359 344
auN	5 277 592	355 297	4 880 849	72 680	5 338 980	2 017 471	5 314 087	2 017 202
L50	1	7	1	52	1	1	1	1
L90	1	459	2	240	1	5	1	5
GC (%)	57.05	56.9	49.5	49.79	57.49	57.5	57.46	57.5
Mismatches								
# N's per 100 kbp	0	0	0	1.01	0	1.76	0	1.8
# N's	0	0	0	98	0	97	0	99

2 contigs

Quality of assemblies Basic evaluation



2 contigs

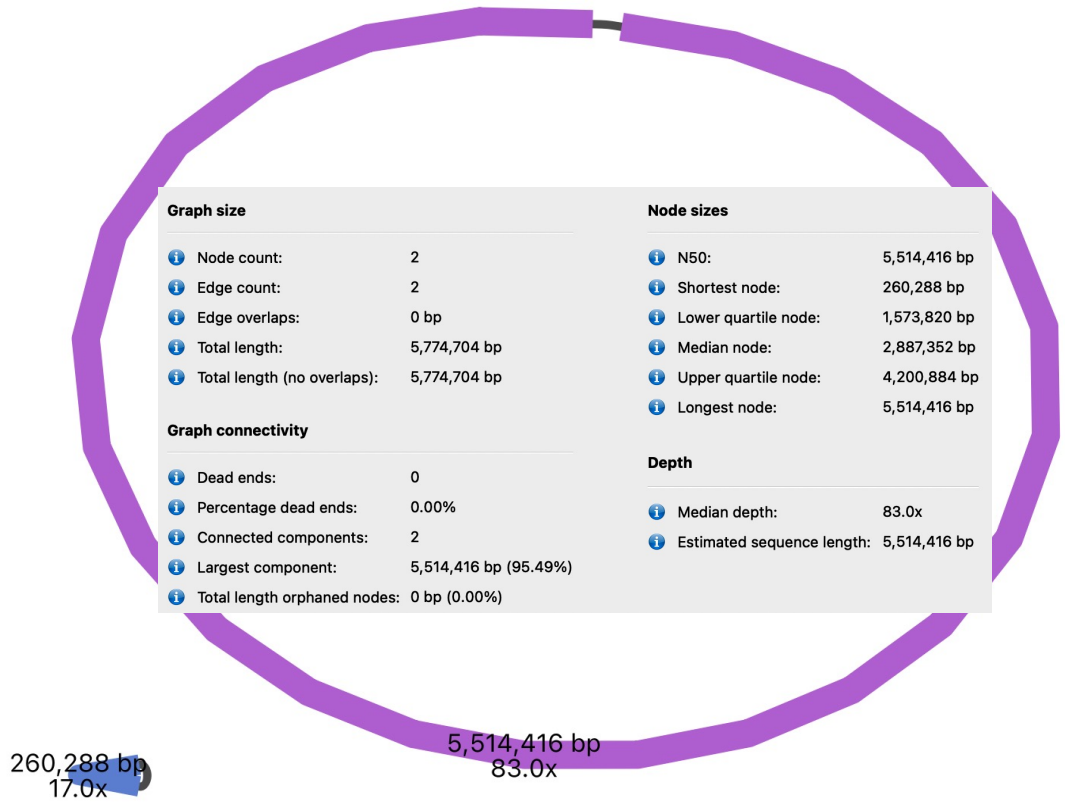
Worst Median Best — Show heatmap

Statistics without reference	CFS1073_ONT
# contigs	2
# contigs (>= 0 bp)	2
# contigs (>= 1000 bp)	2
# contigs (>= 5000 bp)	2
# contigs (>= 10000 bp)	2
# contigs (>= 25000 bp)	2
# contigs (>= 50000 bp)	2
Largest contig	5 514 416
Total length	5 774 704
Total length (>= 0 bp)	5 774 704
Total length (>= 1000 bp)	5 774 704
Total length (>= 5000 bp)	5 774 704
Total length (>= 10000 bp)	5 774 704
Total length (>= 25000 bp)	5 774 704
Total length (>= 50000 bp)	5 774 704
N50	5 514 416
N90	5 514 416
auN	5 277 592
L50	1
L90	1
GC (%)	57.05
Mismatches	
# N's per 100 kbp	0
# N's	0

Parameters

- Number of contigs
- Size
- N50
- N90
- L50
- L90
- GC

Quality of assemblies Basic evaluation



2 contigs

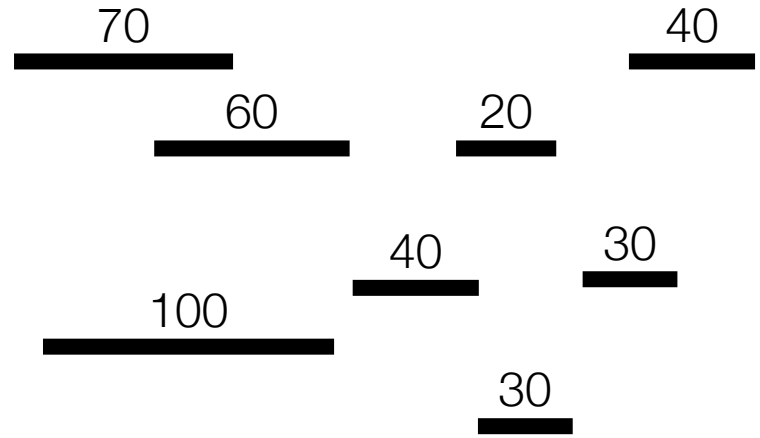
Worst Median Best — Show heatmap

Statistics without reference	CFS1073_ONT
# contigs	2
# contigs (>= 0 bp)	2
# contigs (>= 1000 bp)	2
# contigs (>= 5000 bp)	2
# contigs (>= 10000 bp)	2
# contigs (>= 25000 bp)	2
# contigs (>= 50000 bp)	2
Largest contig	5 514 416
Total length	5 774 704
Total length (>= 0 bp)	5 774 704
Total length (>= 1000 bp)	5 774 704
Total length (>= 5000 bp)	5 774 704
Total length (>= 10000 bp)	5 774 704
Total length (>= 25000 bp)	5 774 704
Total length (>= 50000 bp)	5 774 704
N50	5 514 416
N90	5 514 416
auN	5 277 592
L50	1
L90	1
GC (%)	57.05
Mismatches	
# N's per 100 kbp	0
# N's	0

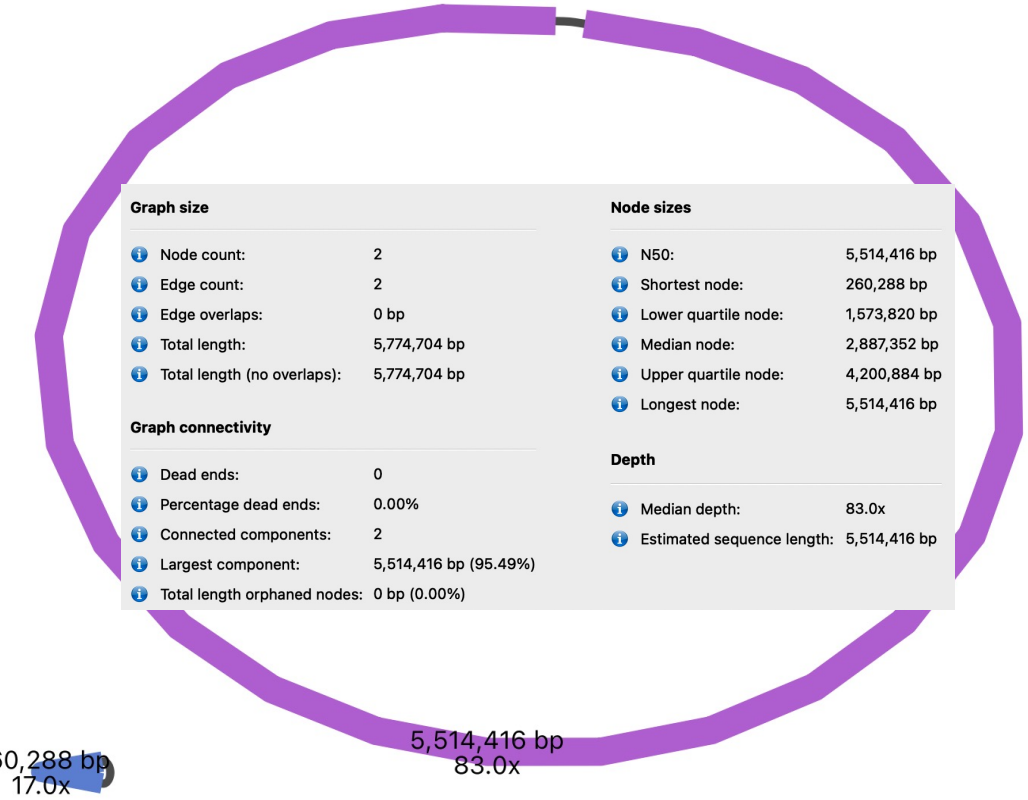
Parameters

- N50
- N90

The maximum length X for which the collection of all contigs of length >= X covers at least 50% of the assembly



Quality of assemblies Basic evaluation



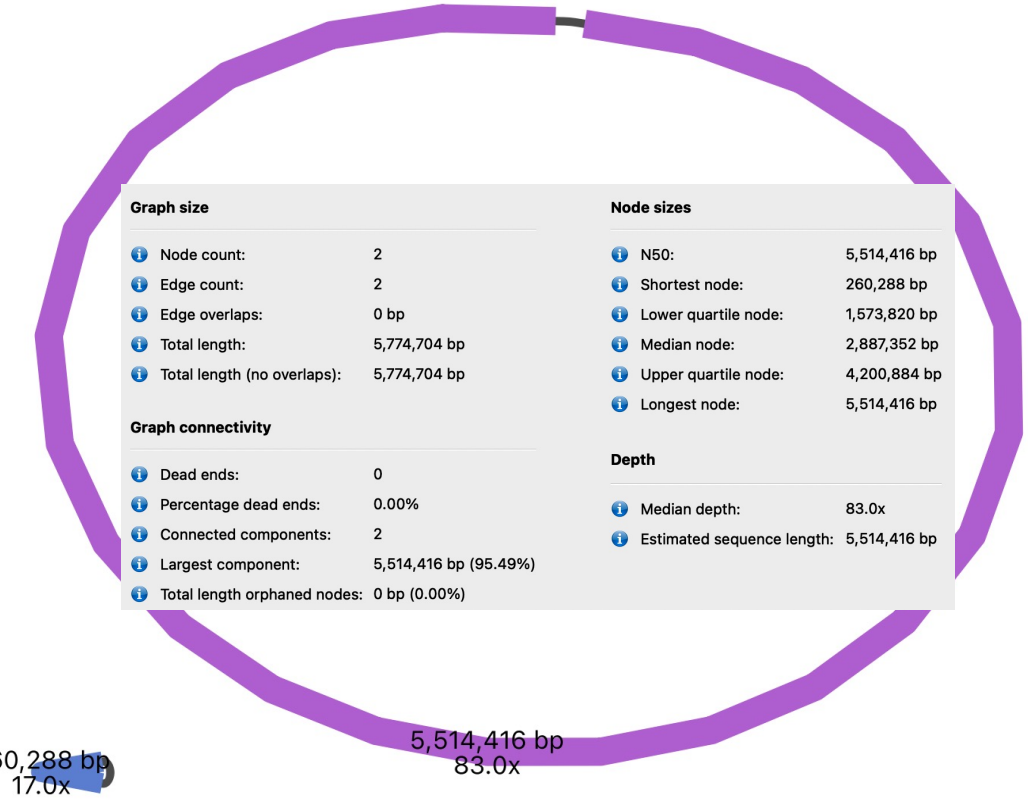
2 contigs

Parameters

- N50
- N90

The maximum length X for which the collection of all contigs of length >= X covers at least 50% of the assembly

Quality of assemblies Basic evaluation

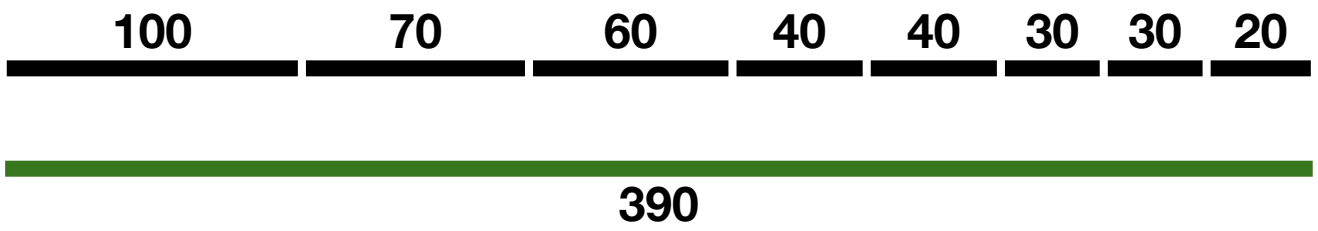


2 contigs

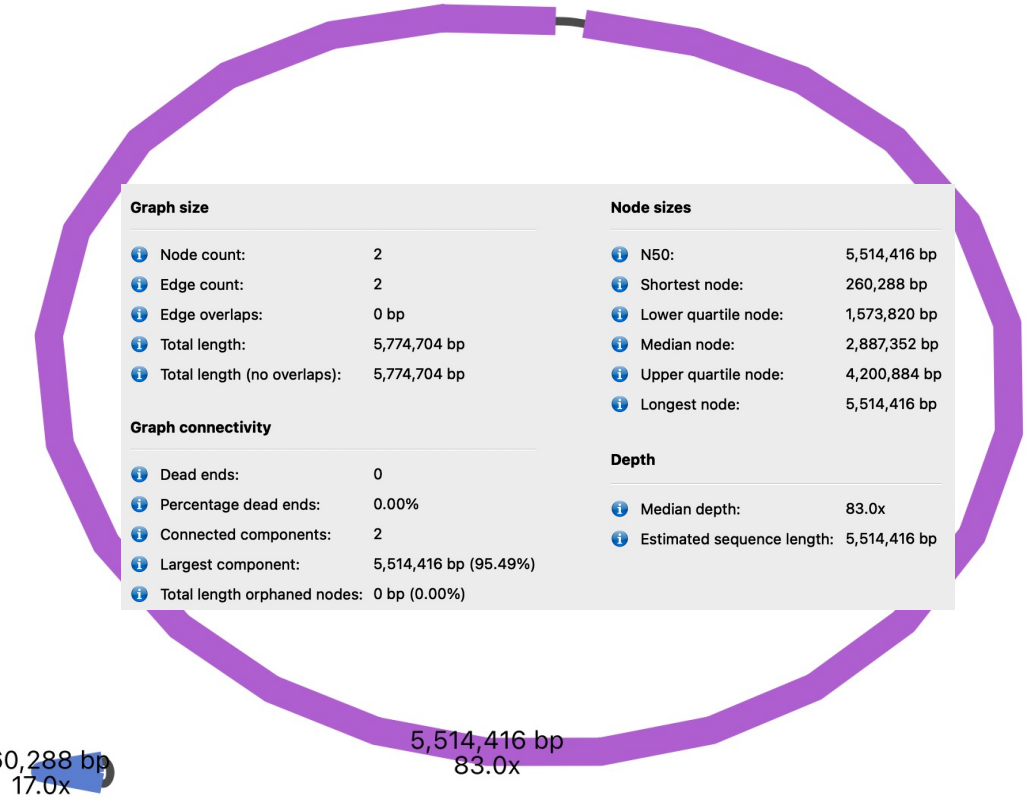
Parameters

- N50
- N90

The maximum length X for which the collection of all contigs of length >= X covers at least 50% of the assembly



Quality of assemblies Basic evaluation

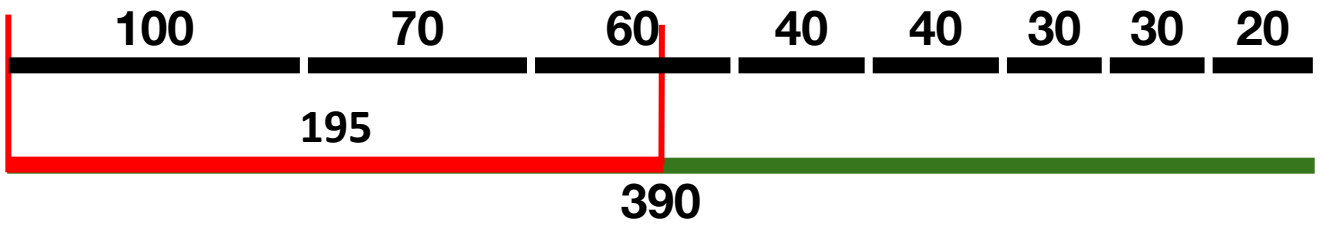


2 contigs

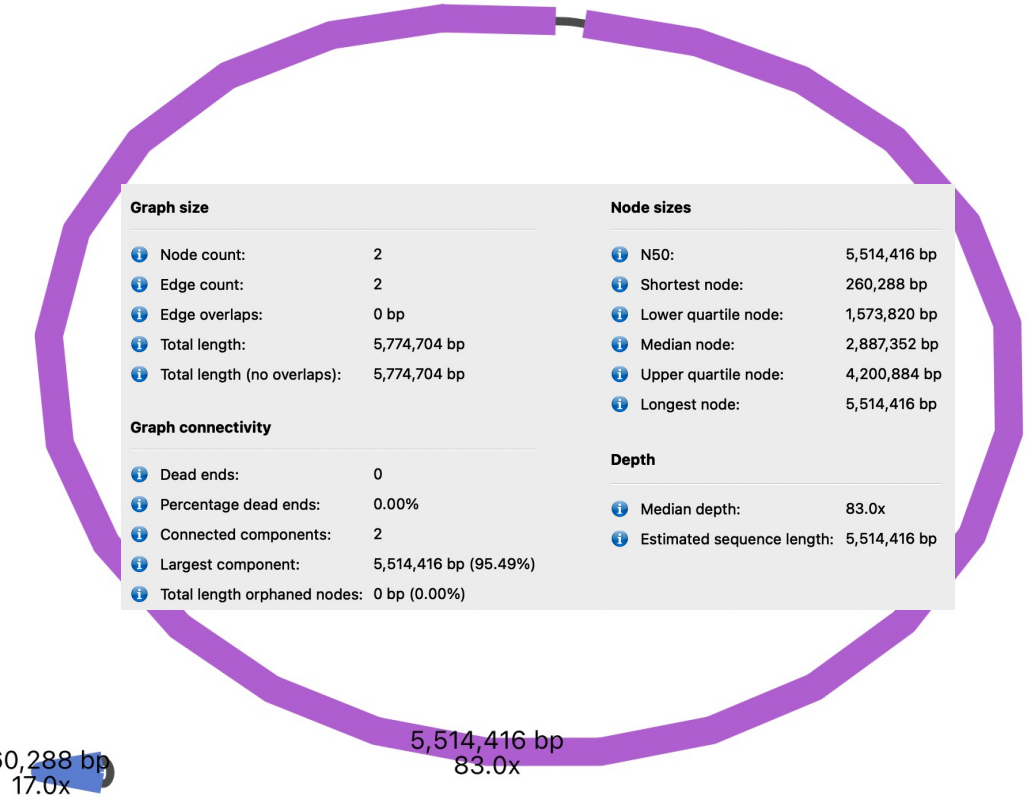
Parameters

- N50
- N90

The maximum length X for which the collection of all contigs of length >= X covers at least 50% of the assembly



Quality of assemblies Basic evaluation

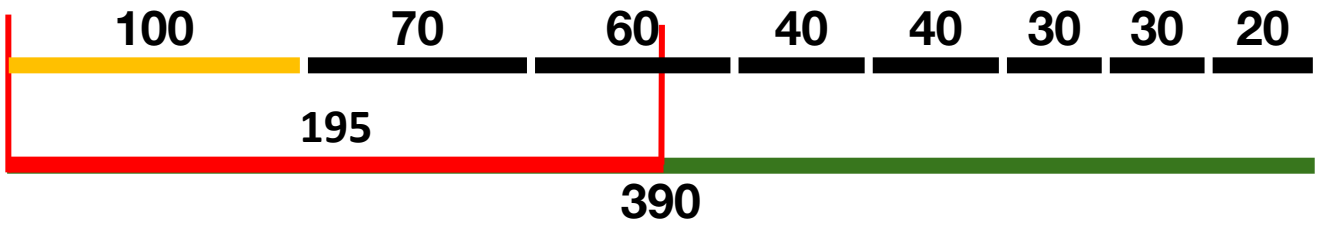


2 contigs

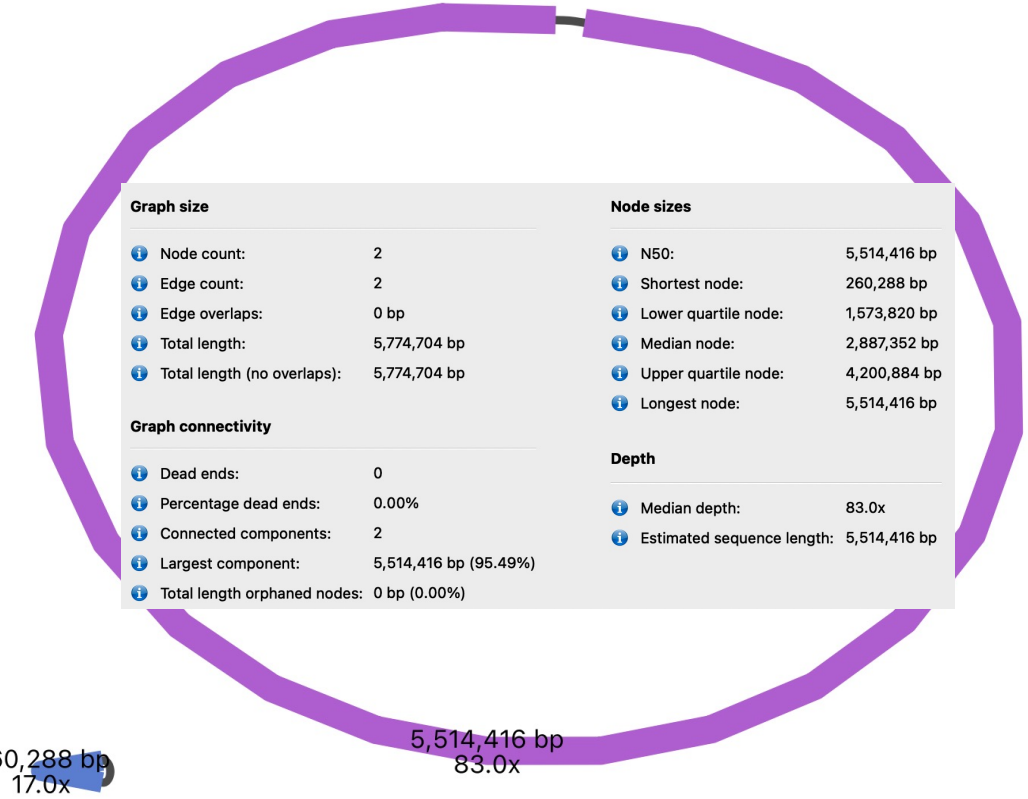
Parameters

- N50
- N90

The maximum length X for which the collection of all contigs of length >= X covers at least 50% of the assembly



Quality of assemblies Basic evaluation

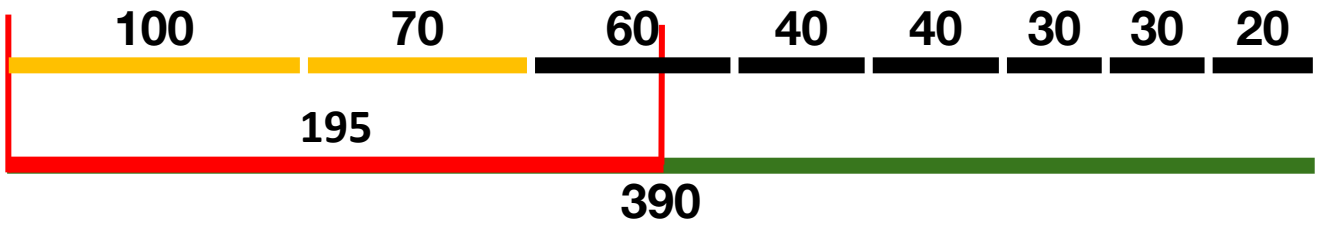


2 contigs

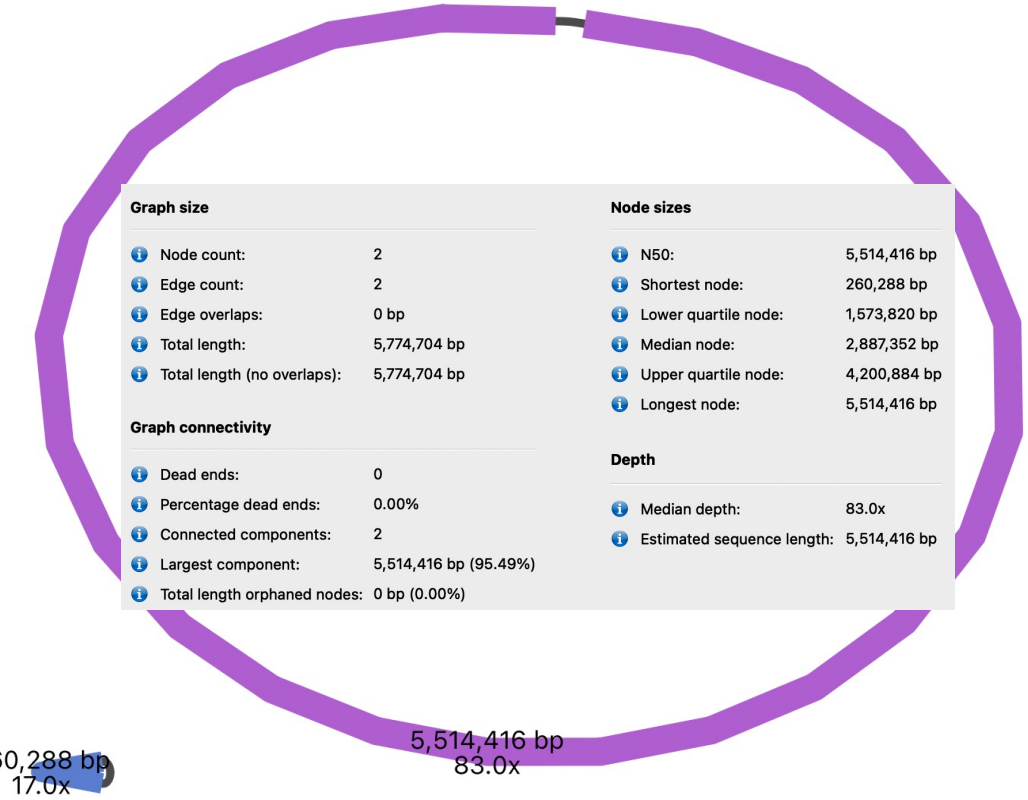
Parameters

- N50
- N90

The maximum length X for which the collection of all contigs of length >= X covers at least 50% of the assembly



Quality of assemblies Basic evaluation

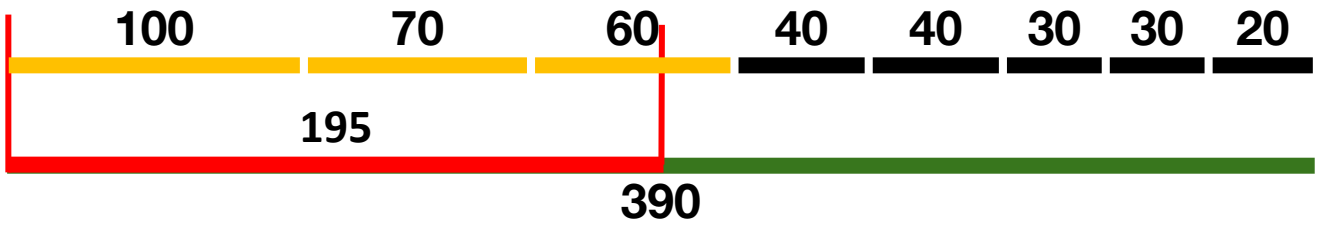


2 contigs

Parameters

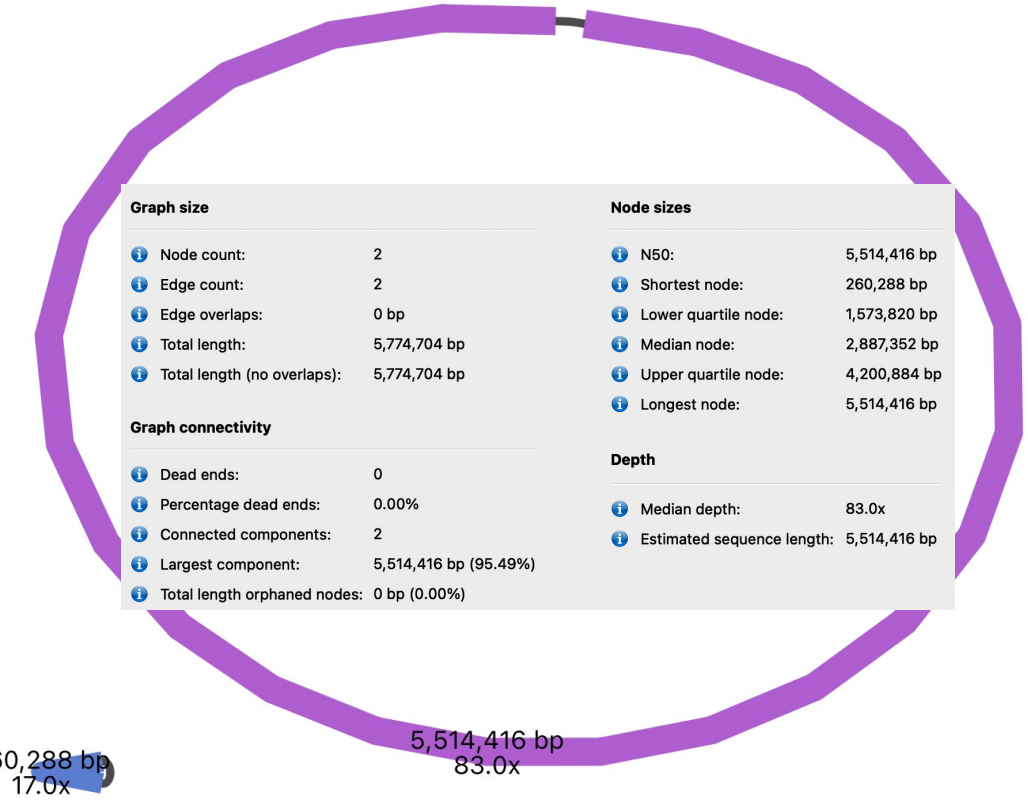
- N50
- N90

The maximum length X for which the collection of all contigs of length >= X covers at least 50% of the assembly



N50=60

Quality of assemblies Basic evaluation

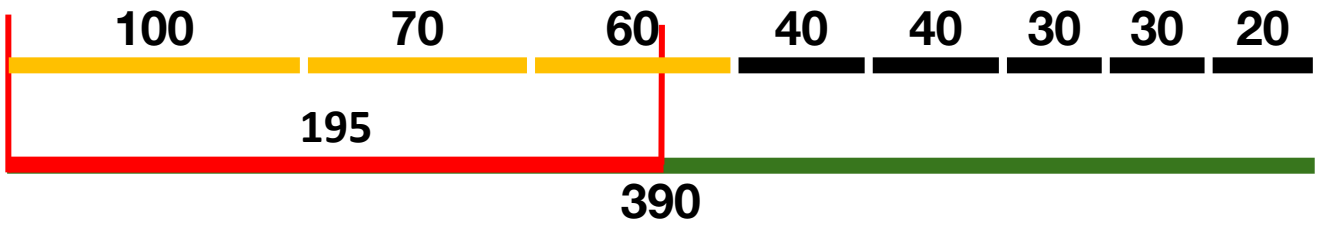


2 contigs

Parameters

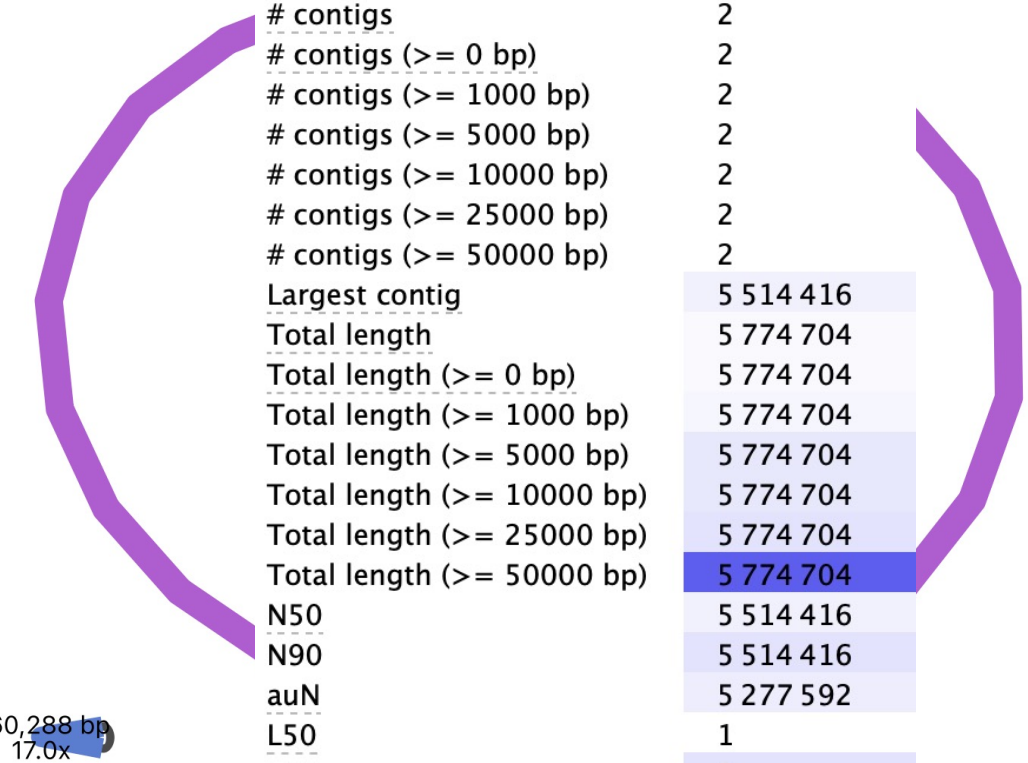
- L50
- L90

The minimum number X such that X longest contigs cover at least 50% of the assembly



L50=3

Quality of assemblies Basic evaluation



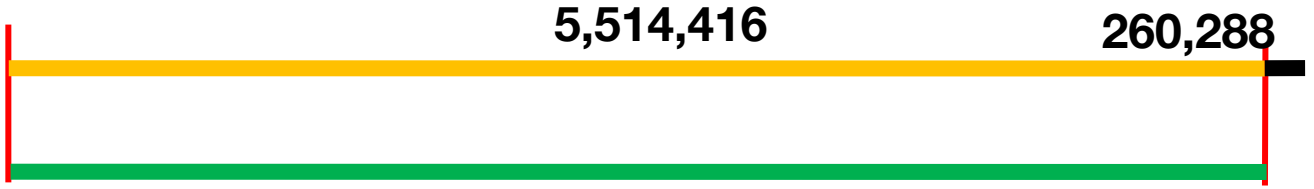
Worst Median Best — Show heatmap

Statistics without reference	CFS1073_ONT
# contigs	2
# contigs (>= 0 bp)	2
# contigs (>= 1000 bp)	2
# contigs (>= 5000 bp)	2
# contigs (>= 10000 bp)	2
# contigs (>= 25000 bp)	2
# contigs (>= 50000 bp)	2
Largest contig	5 514 416
Total length	5 774 704
Total length (>= 0 bp)	5 774 704
Total length (>= 1000 bp)	5 774 704
Total length (>= 5000 bp)	5 774 704
Total length (>= 10000 bp)	5 774 704
Total length (>= 25000 bp)	5 774 704
Total length (>= 50000 bp)	5 774 704
N50	5 514 416
N90	5 514 416
auN	5 277 592
L50	1
L90	1
GC (%)	57.05
Mismatches	
# N's per 100 kbp	0
# N's	0

Parameters

- L50
- L90

The minimum number X such that X longest contigs cover at least 50% of the assembly

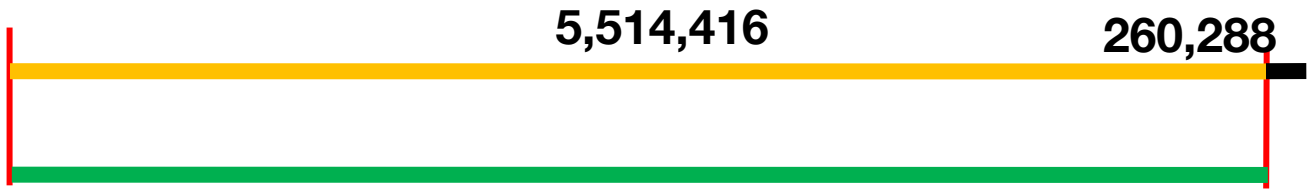


N50=5,514,416 | L50=1

Quality of assemblies Basic evaluation

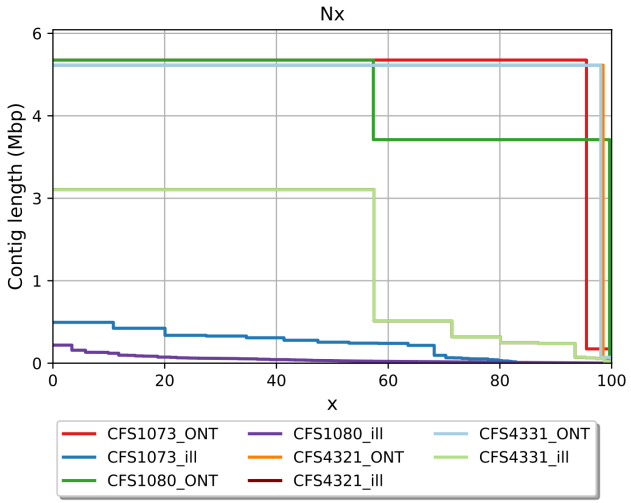
Given a de novo assembly, we often measure the “average” contig length by N50.

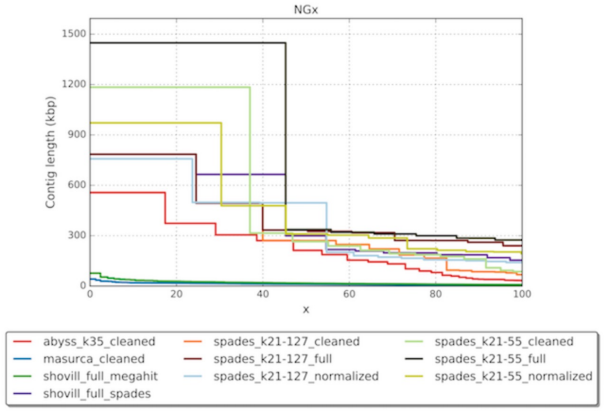
N50 is neither the real average nor median. It is the length of the contig such that this and longer contigs cover at least 50% of the assembly. A longer N50 indicates better contiguity. We can similarly define Nx such that contigs no shorter than Nx covers x% of the assembly. The Nx curve plots Nx as a function of x, where x is ranged from 0 to 10



N50=5,514,416 | L50=1

Quality of assemblies Basic evaluation





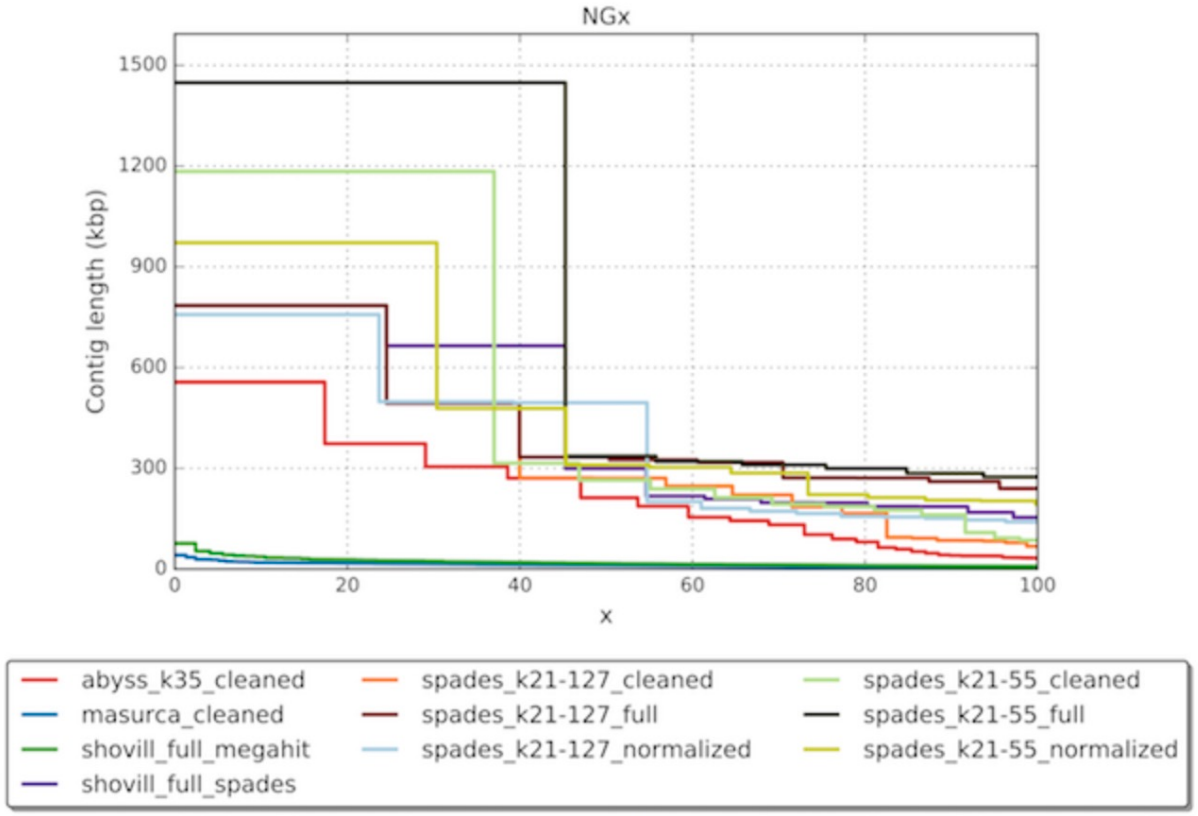
https://lh3.github.io/2020/04/08/a-new-metric-on-assembly-contiguity

NG50 (similar to N50) of several assemblers/settings are about the same around 300kb, but it is clear the black curve achieves better contiguity – a single contig on that curve covers more than 40% of the assembly.

Intuitively, a better Nx curve is “higher”, or has a larger area under the curve. Then we can take the area under the curve, abbreviated as “auN”, as a measurement of contiguity. The formula to calculate the area

$$\text{auN} = \sum_i L_i \cdot \frac{L_i}{\sum_j L_j} = \sum_i L_i^2 \Big/ \sum_j L_j$$

Quality of assemblies Basic evaluation



where Li is the length of contig i.

$$\text{auN} = \sum_i L_i \cdot \frac{L_i}{\sum_j L_j} = \sum_i L_i^2 \Big/ \sum_j L_j$$

It is more stable and less affected by big jumps in contig lengths.
It considers the entire Nx curve.
Connecting two contigs of any lengths will always lead to a longer auN

https://lh3.github.io/2020/04/08/a-new-metric-on-assembly-contiguity

Quality of assemblies Basic evaluation

Worst Median Best

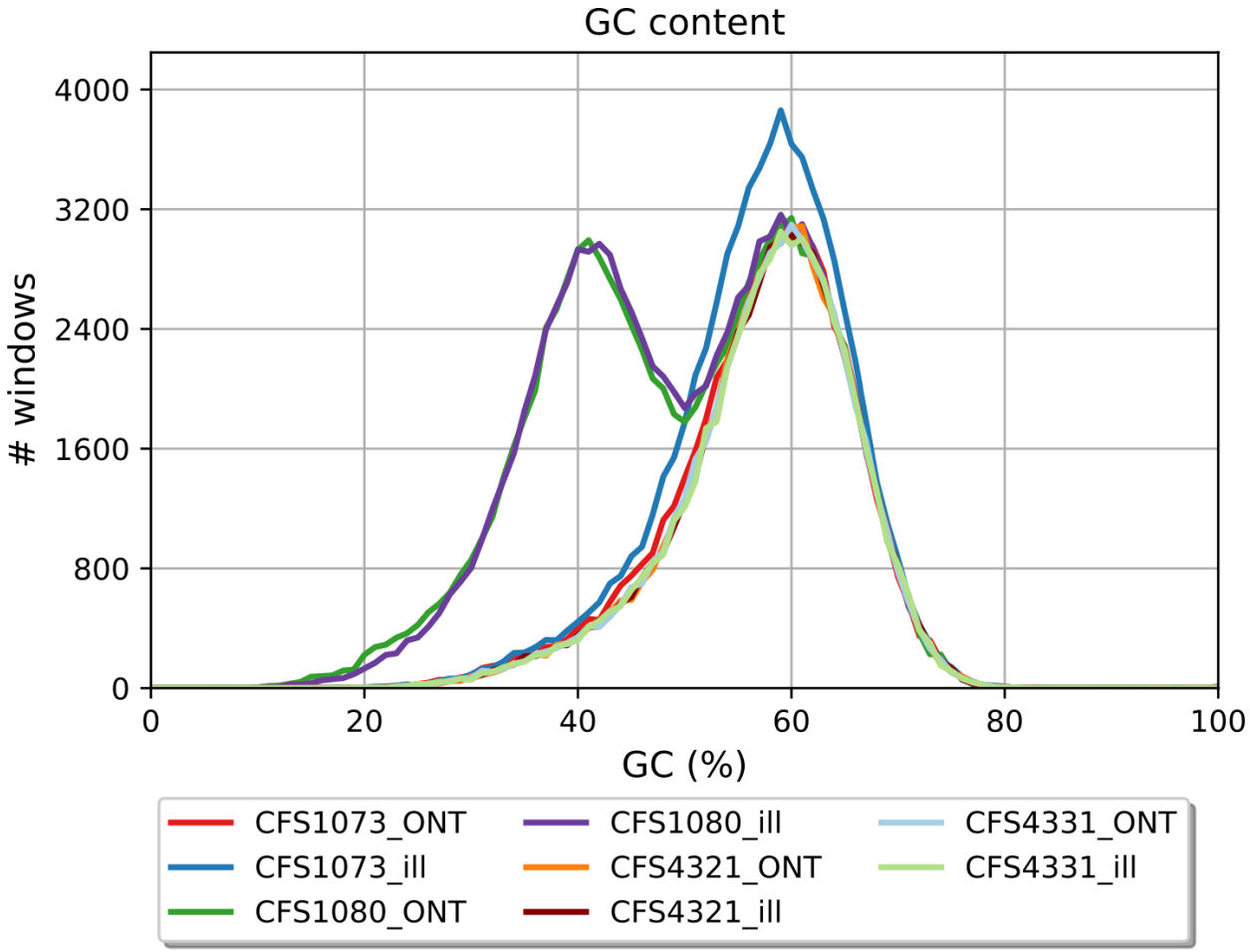
Show heatmap

Statistics without reference	CFS1073_ONT
# contigs	2
# contigs (>= 0 bp)	2
# contigs (>= 1000 bp)	2
# contigs (>= 5000 bp)	2
# contigs (>= 10000 bp)	2
# contigs (>= 25000 bp)	2
# contigs (>= 50000 bp)	2
Largest contig	5 514 416
Total length	5 774 704
Total length (>= 0 bp)	5 774 704
Total length (>= 1000 bp)	5 774 704
Total length (>= 5000 bp)	5 774 704
Total length (>= 10000 bp)	5 774 704
Total length (>= 25000 bp)	5 774 704
Total length (>= 50000 bp)	5 774 704
N50	5 514 416
N90	5 514 416
auN	5 277 592
L50	1
L90	1
GC (%)	57.05
Mismatches	
# N's per 100 kbp	0
# N's	0

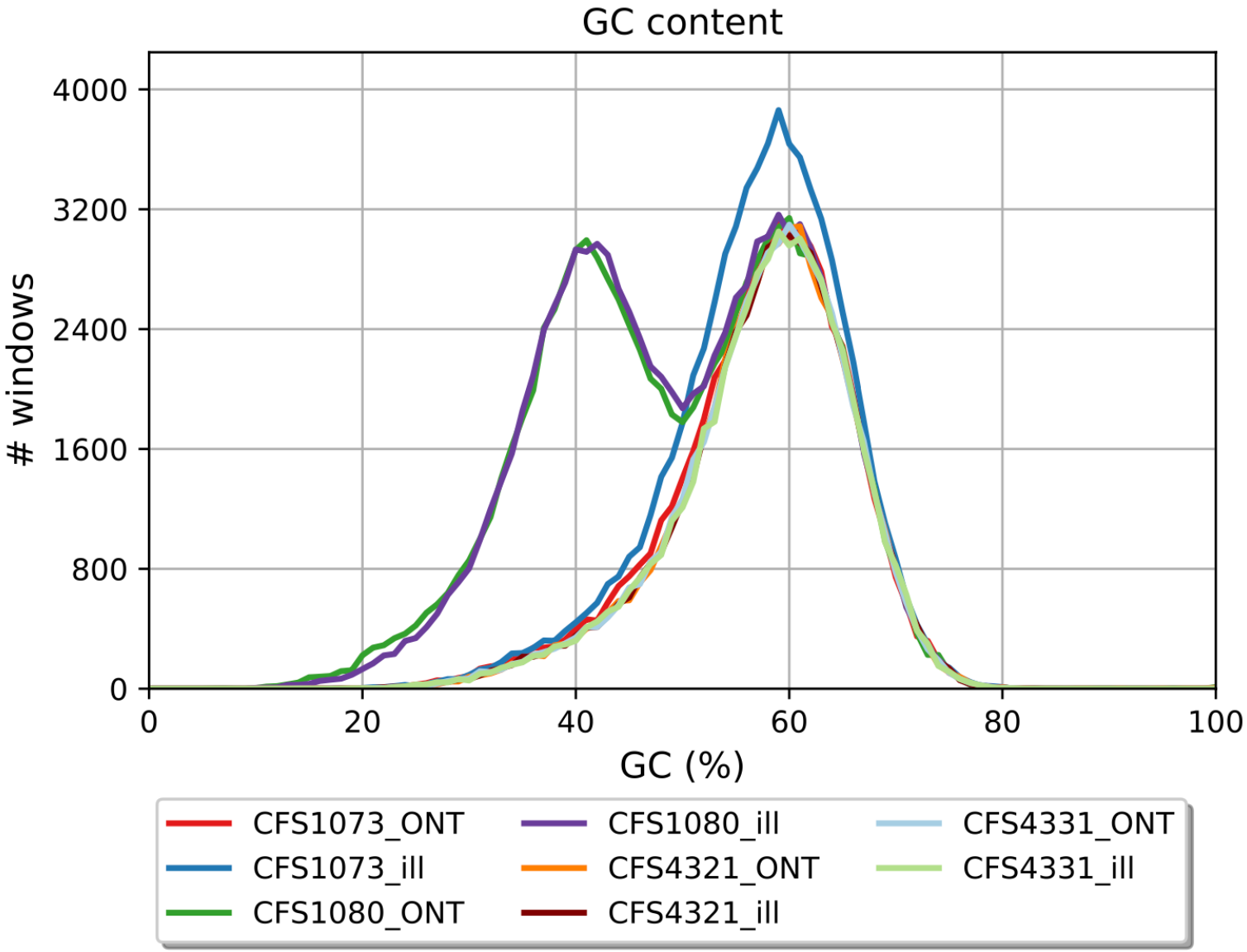
260,288 bp
17.0x

Parameters

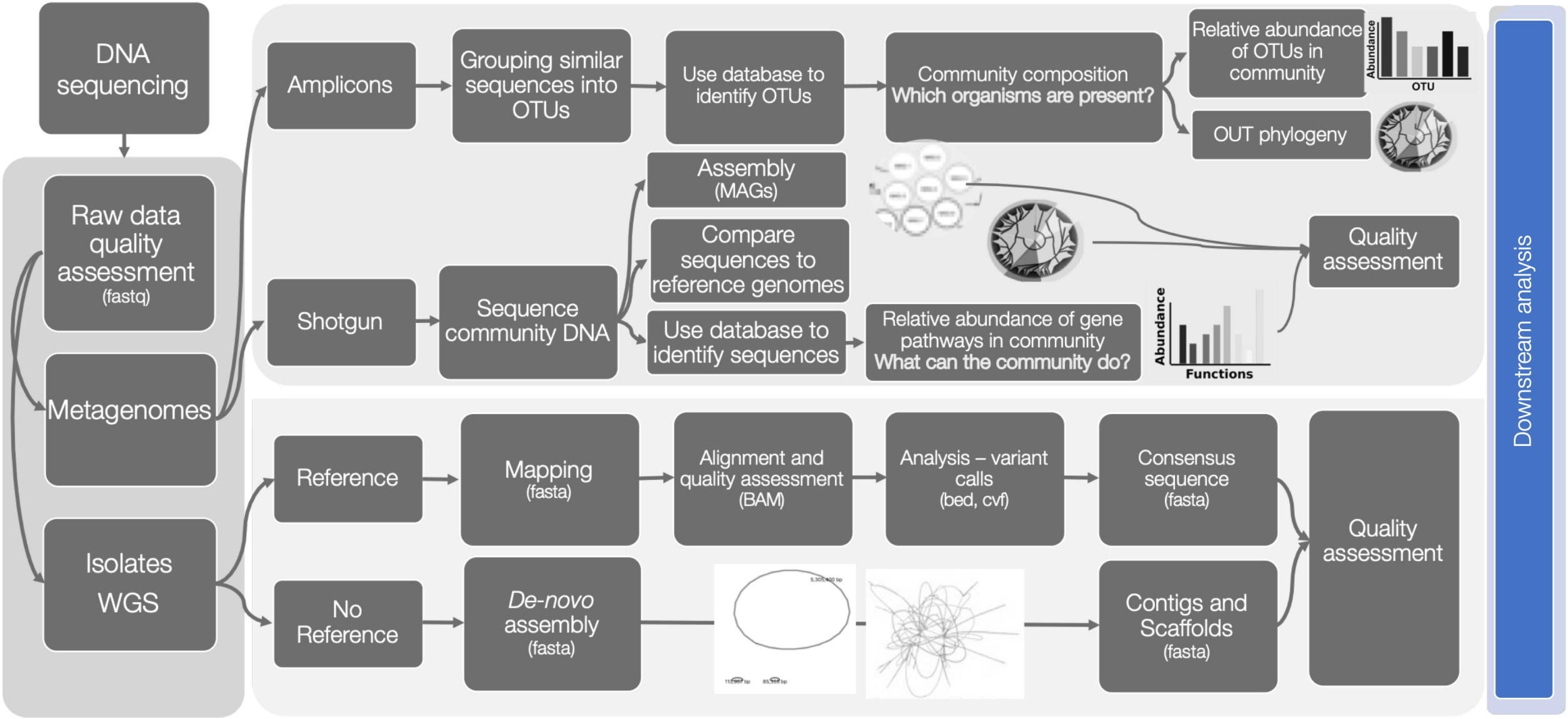
- GC % content



Quality of assemblies Basic evaluation



Food Safety Bioinformatics



Annotation

Downstream analysis

Microbial genome annotation often consists of running an automatic annotation pipeline followed by manual curation of the results

Homology methods to transfer information from a closely related reference genome to the new sequence.

Automatic pipelines with errors and then manual curation step to catch and remove these.

Whole genome annotation

Downstream analysis

Tools for bacterial whole genome annotation

Prokka – Standalone command line tool, takes just a few minutes per genome. This is the best way to get good quality annotation in a flash, which is particularly useful if you have loads of genomes or need to annotate a pangenome or metagenome. Note however that the quality of functional information is not as good as RAST, and you will need several extra steps if you want to do functional profiling and pathway analysis of your genome(s)… which is in-built in RAST.

PGAP: NCBI has developed an automatic prokaryotic genome annotation pipeline that combines ab initio gene prediction algorithms with homology based methods. The first version of NCBI Prokaryotic Genome Automatic Annotation Pipeline (PGAAP; see Pubmed Article) developed in 2005 has been replaced with an upgraded version that is capable of processing a larger data volume. NCBI's annotation pipeline depends on several internal databases and is not currently available for download or use outside of the NCBI environment.

BEACON (automated tool for Bacterial GEnome Annotation ComparisON), a fast tool for an automated and a systematic comparison of different annotations of single genomes. The extended annotation assigns putative functions to many genes with unknown functions. BEACON is available under GNU General Public License version 3.0 and is accessible at: http://www.cbrc.kaust.edu.sa/BEACON/.

BlastKOLA: Assigns K numbers to the user's sequence data by BLAST searches, respectively, against a nonredundant set of KEGG GENES. KOALA (KEGG Orthology And Links Annotation) is KEGG's internal annotation tool for K number assignment of KEGG GENES using SSEARCH computation. Annotate Sequence in KEGG Mapper and Pathogen Checker in KEGG Pathogen are special interfaces to this server and can be executed in an interactive mode. BlastKOALA is suitable for annotating fully sequenced genomes.

PAGIT: Provides a toolkit for improving the quality of genome assemblies created via an assembly software. PAGIT compiled four tools: (i) ABACAS which classifies and orientates contigs and estimates the sizes of gaps between them; (ii) IMAGE uses paired-end reads to extend contigs and close gaps within the scaffolds; (iii) ICORN for identifying and correcting small errors in consensus sequences and; (iv) RATT for help annotation. The software was mainly created to analyze parasite genomes of up to about 300 Mb.

MAKER: A portable and easily configurable genome annotation pipeline. MAKER allows smaller eukaryotic and prokaryotic genome projects to independently annotate their genomes and to create genome databases. It identifies repeats, aligns ESTs and proteins to a genome, produces ab-initio gene predictions and automatically synthesizes these data into gene annotations having evidence-based quality values. MAKER's inputs are minimal and its ouputs can be directly loaded into a Generic Model Organism Database (GMOD). They can also be viewed in the Apollo genome browser; this feature of MAKER provides an easy means to annotate, view and edit individual contigs and BACs without the overhead of a database. MAKER is available for download and can be tested online via the MAKER Web Annotation Service (MWAS).

MyPro is a software pipeline for high-quality prokaryotic genome assembly and annotation. It was validated on 18 oral streptococcal strains to produce submission-ready, annotated draft genomes. MyPro installed as a virtual machine and supported by updated databases will enable biologists to perform quality prokaryotic genome assembly and annotation with ease.

Annotation

Downstream analysis

Prokka is a software tool for the rapid annotation of prokaryotic genomes. A typical 4 Mbp genome can be fully annotated in less than 10 minutes on a quad-core computer, and scales well to 32 core SMP systems. It produces GFF3, GBK and SQN files that are ready for editing in Sequin and ultimately submitted to Genbank/DDJB/ENA.

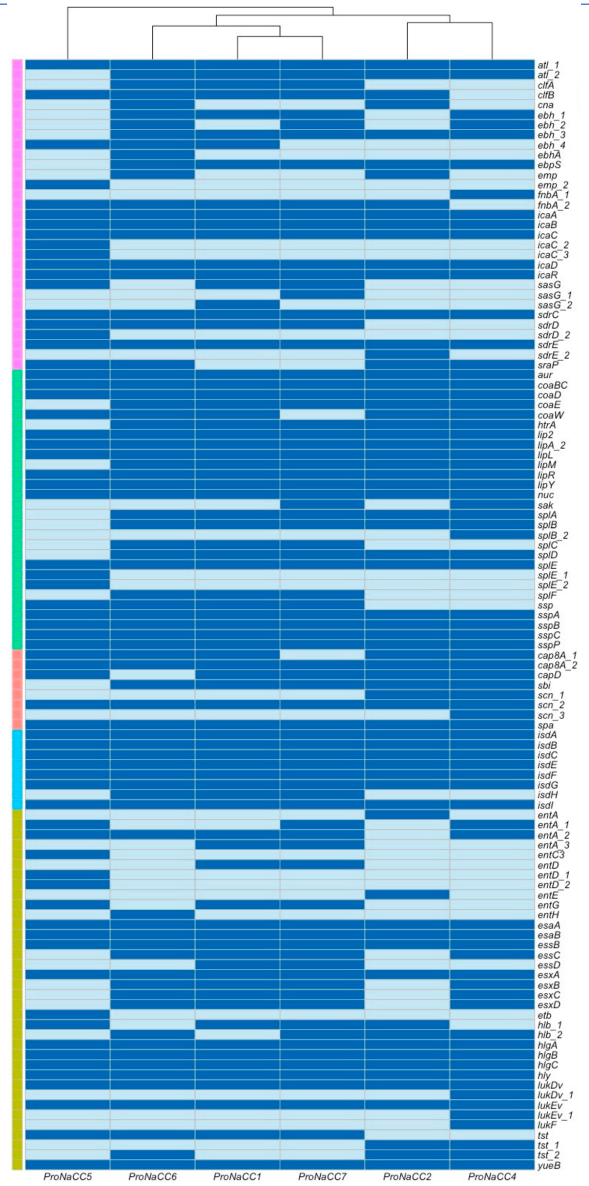
Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068-9. PMID:24642063

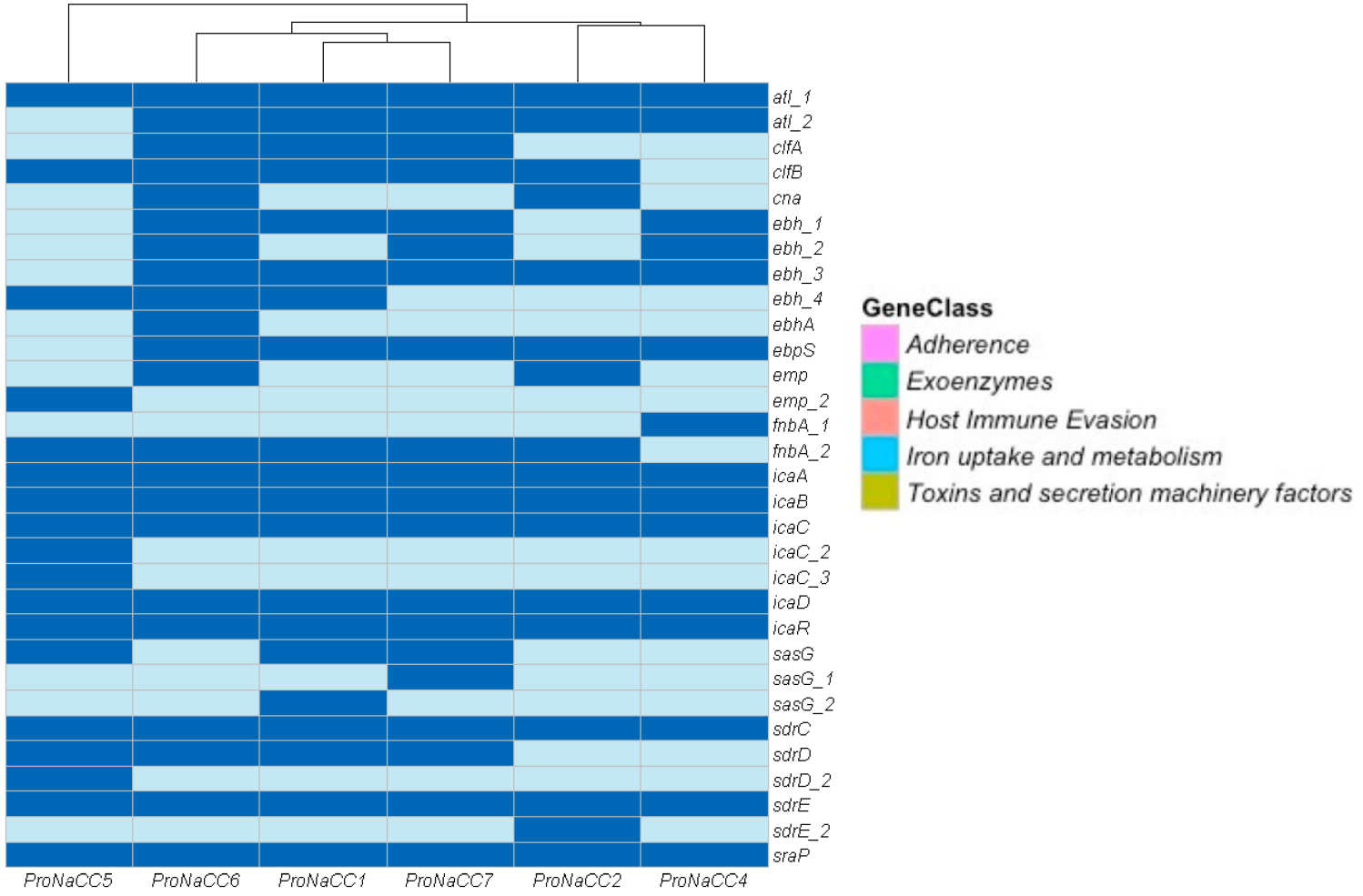
PROKKA_11092021

locus_tag	ftype	length_bp	gene	EC_number	COG	product
JCJMHEKL_00001	CDS	300				hypothetical protein
JCJMHEKL_00002	CDS	1794				hypothetical protein
JCJMHEKL_00003	CDS	300				hypothetical protein
JCJMHEKL_00004	CDS	1239	tyrS_1	6.1.1.1	COG0162	Tyrosine--tRNA ligase
JCJMHEKL_00005	CDS	1257	nicP_1			Porin-like protein NicP
JCJMHEKL_00006	CDS	1209	bcr_1			Bicyclomycin resistance protein
JCJMHEKL_00007	CDS	423	farR			HTH-type transcriptional regulator FarR
JCJMHEKL_00008	CDS	243				hypothetical protein
JCJMHEKL_00009	CDS	1428	ccoN1_1	1.9.3.1	COG3278	Cbb3-type cytochrome c oxidase subunit CcoN1
JCJMHEKL_00010	CDS	1503	norG_1		COG1167	HTH-type transcriptional regulator NorG
JCJMHEKL_00011	CDS	1431				hypothetical protein
JCJMHEKL_00012	CDS	1674	sir_1	1.8.7.1	COG0155	Sulfite reductase [ferredoxin]
JCJMHEKL_00013	CDS	489				hypothetical protein
JCJMHEKL_00014	CDS	807	hpcH	4.1.2.52	COG3836	4-hydroxy-2-oxo-heptane-1,7-dioate aldolase
JCJMHEKL_00015	CDS	804	hpcG	4.2.1.163		2-oxo-hept-4-ene-1,7-dioate hydratase
JCJMHEKL_00016	CDS	1305	nicT			Putative metabolite transport protein NicT
JCJMHEKL_00017	CDS	393	hpcD_1	5.3.3.10		5-carboxymethyl-2-hydroxymuconate Delta-isomerase
JCJMHEKL_00018	CDS	924	hpcB	1.13.11.15	COG3384	3,4-dihydroxyphenylacetate 2,3-dioxygenase
JCJMHEKL_00019	CDS	1461	betB_1	1.2.1.8	COG1012	NAD/NADP-dependent betaine aldehyde dehydrogenase
JCJMHEKL_00020	CDS	780	hpcE_1		COG0179	Homoprotocatechuate catabolism bifunctional isomerase/decarboxylase
JCJMHEKL_00021	CDS	660	hpcE_2		COG0179	Homoprotocatechuate catabolism bifunctional isomerase/decarboxylase
JCJMHEKL_00022	CDS	912	rhaR_1			HTH-type transcriptional activator RhaR
JCJMHEKL_00023	CDS	807	neo	2.7.1.95		Aminoglycoside 3'-phosphotransferase
JCJMHEKL_00024	CDS	570				hypothetical protein
JCJMHEKL_00025	CDS	2187	bphP	2.7.13.3		Bacteriophytochrome
JCJMHEKL_00026	CDS	588				hypothetical protein
JCJMHEKL_00027	CDS	1386	ppnN	3.2.2.-	COG1611	Pyrimidine/purine nucleotide 5'-monophosphate nucleosidase
JCJMHEKL_00028	CDS	510				hypothetical protein
JCJMHEKL_00029	CDS	1191	sotB_1			Sugar efflux transporter B

Annotation

Downstream analysis



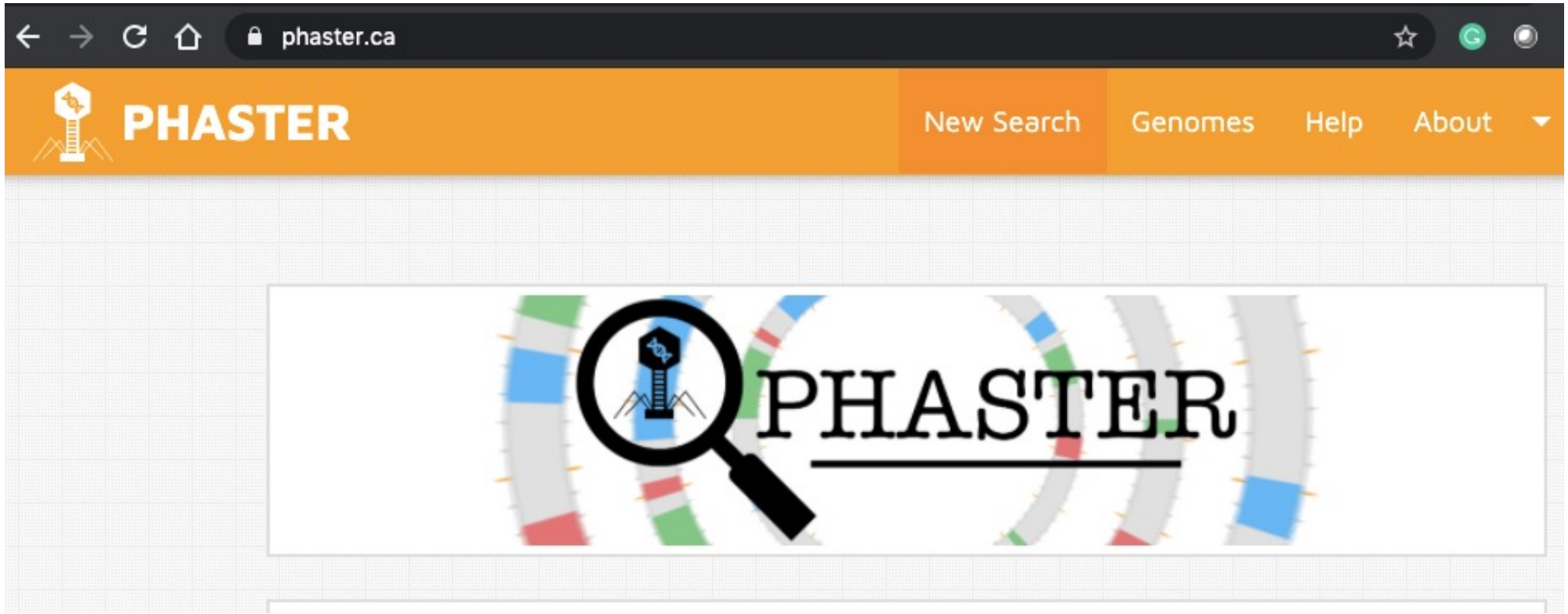


Annotation

Downstream analysis

In silico Phage Hunting

PHASTER (PHAge Search Tool Enhanced Release)



Annotation

Downstream analysis

In silico Phage Hunting

PHASTER (PHAge Search Tool Enhanced Release)

Total: 1 prophage regions have been identified, of which 1 regions are intact, 0 regions are incomplete, and 0 regions are questionable.

Region	Region Length	Completeness	Score	# Total Proteins	Region Position	Most Common Phage	GC %	Details
1	38.2Kb	intact	131	55	156682-194926	PHAGE_Pseudo_JD024_NC_024330(51)	64.35%	Show

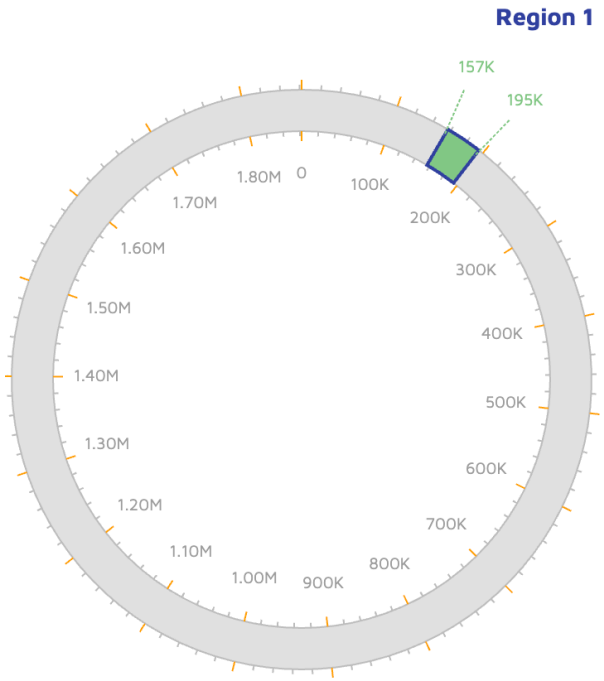
Intact (score > 90)
Questionable (score 70-90)
Incomplete (score < 70)

Annotation

Downstream analysis

In silico Phage Hunting

[">NODE_1_length_1862038_cov_39.509666"]



Prophage Region 1

Start: 156682
End: 194926
 # CDS: 55
Predicted Type: intact
GC%: 64.35

Intact (score > 90)
Questionable (score 70-90)
Incomplete (score < 70)

Viewer Options

Hide Region Labels
Show Label Lines
Hide Markers
Condense Labels
Save Image

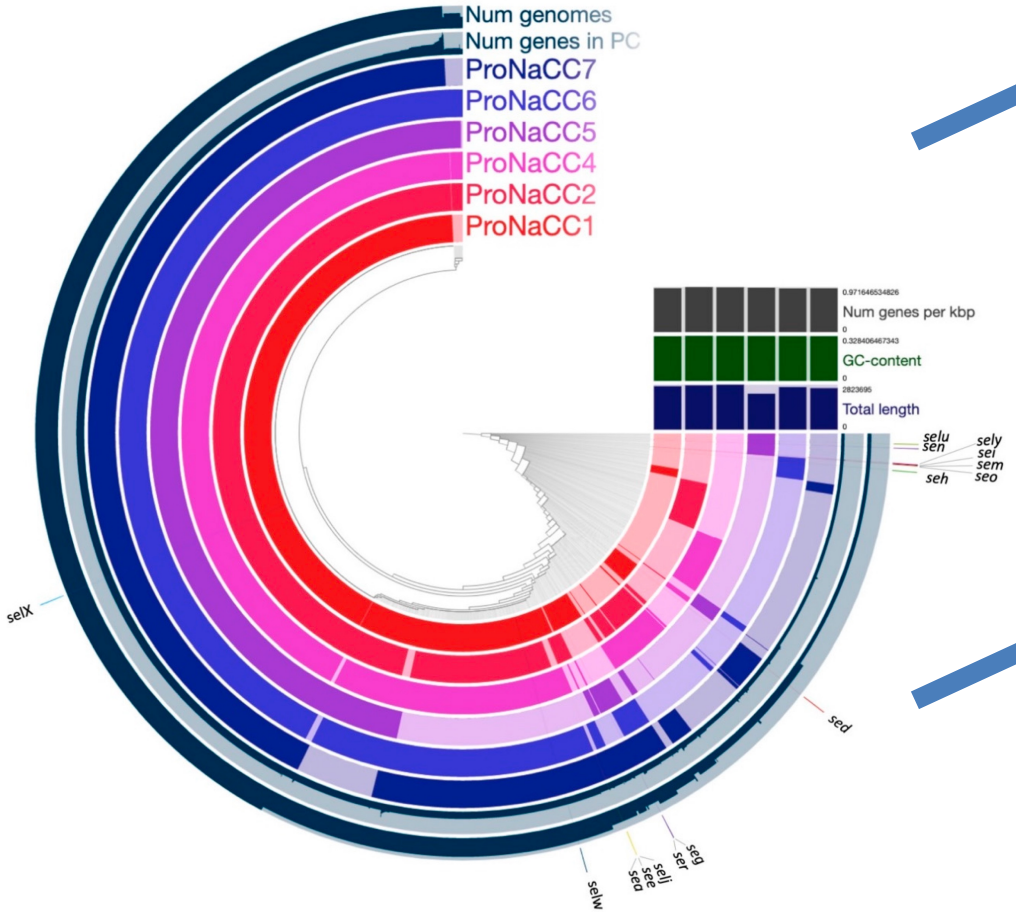
Length: 1862038 bps
Phages: 1

ion	1
ion Length	38.2Kb
npleteness(score)	intact(131)
cific Keyword	transposase,capsid,virion,tail
ion Position	156682-194926
RNA	0
otal Proteins	55
hage Hit Proteins	55
ypothetical Proteins	0
ge + Hypothetical Protein %	100%

Annotation

Downstream analysis

In silico Phage Hunting



GGTCGTATTATCGCACGTCCATATGTTGGTGAACCAGGAAACTTTAC
ACGTACATCTAATCGACATGACTATGCGTTAAAACCTTTTGGTAAAA
CTGTCTTAGATCATTTGAAAGACGGTGGTTATGATGTTATTGCCATC
GGTAAAATTAATGACATTTATGATGGTGAAGGTGTAACAGAAGCGGT
TCGTACGTAGAGTAACATGGACGGTATGGATCAATTGATGAAAATTG
TTAAGAAAGATTTCACAGGTATTAGCTTCTTAAACTTAGTAGACTTT
GATGCATTATACGGTCATCGTCGTGATAAACCAGGTTATGCACAAGC
AATTAAAGATTTCGATGATCGCTTGCCAGAACTGTTTAGCAACTTAA
AAGAAGACGATTTAGTAATTATTACAGCAGACCATGGTAATGACCCG
ACAGCGCCAGGTACGGACCATACGAGAGAATATATCCCAGTAATTAT
GTACAGTCCGAAATTTAAAGGTGGTCATGCACTAGAAAGTGATACTA
CATTCAGTTCTATCGGTGCAACTATAGCAGATAATTTCAACGTAACA
TTACCAGAGTTCGGTAAAAGTTATTTAAAGGAATTGAAATAGAATAA
ATTTAGATATTATAAAAACAGCAGTGAAGTTAACTATAACAATAGTT
TTCTTCACTGCTGTTTTTATTATAATAGAGAAACGTAAGACG

GGTCGTATTATCGCACGTCCATATGTTGGTGAACCAGGAAACTTTAC
ACGTACATCTAATCGACATGACTATGCGTTAAAACCTTTTGGTAAAA
CTGTCTTAGATCATTTGAAAGACGGTGGTTATGATGTTATTGCCATC
GGTAAAATTAATGACATTTATGATGGTGAAGGTGTAACAGAAGCGGT
TCGTACGTAGAGTAACATGGACGGTATGGATCAATTGATGAAAATTG
TTAAGAAAGATTTCACAGGTATTAGCTTCTTAAACTTAGTAGACTTT
GATGCATTATACGGTCATCGTCGTGATAAACCAGGTTATGCACAAGC
AATTAAAGATTTCGATGATCGCTTGCCAGAACTGTTTAGCAACTTAA
AAGAAGACGATTTAGTAATTATTACAGCAGACCATGGTAATGACCCG
ACAGCGCCAGGTACGGACCATACGAGAGAATATATCCCAGTAATTAT
GTACAGTCCGAAATTTAAAGGTGGTCATGCACTAGAAAGTGATACTA
CATTCAGTTCTATCGGTGCAACTATAGCAGATAATTTCAACGTAACA
TTACCAGAGTTCGGTAAAAGTTATTTAAAGGAATTGAAATAGAATAA
ATTTAGATATTATAAAAACAGCAGTGAAGTTAACTATAACAATAGTT
TTCTTCACTGCTGTTTTTATTATAATAGAGAAACGTAAGACG

Bioinformatic tools for characterising bacterial pathogens

Publicly available antibiotic resistance and virulence databases

Antimicrobial resistance databases	Virulence databases
CARD (https://card.mcmaster.ca/) (*23*)	VFDB (http://www.mgc.ac.cn/VFs/) (*24*)
RAC (http://rac.aihi.mq.edu.au/rac/) (*25*)	PATRIC (https://www.patricbrc.org/) (*26*, *27*)
ResFinder (https://cge.cbs.dtu.dk/services/data.php) (*21*)	Victors (http://www.phidias.us/victors/) (*28*)
ARDB (https://ardb.cbcb.umd.edu/) (*29*)	PHI-BASE (http://www.phi-base.org/) (*30-33*)
NDARO https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/	MvirDB (http://mvirdb.llnl.gov/) (*34*)

Options bioinformatics analysis for Food Safety

- Command line
- Stand alone web-resources (for example http://www.genomicepidemiology.org/services/)
- Online pipelines (for example https://galaxyproject.org/eu/)

Bioinformatic tools for characterising bacterial pathogens

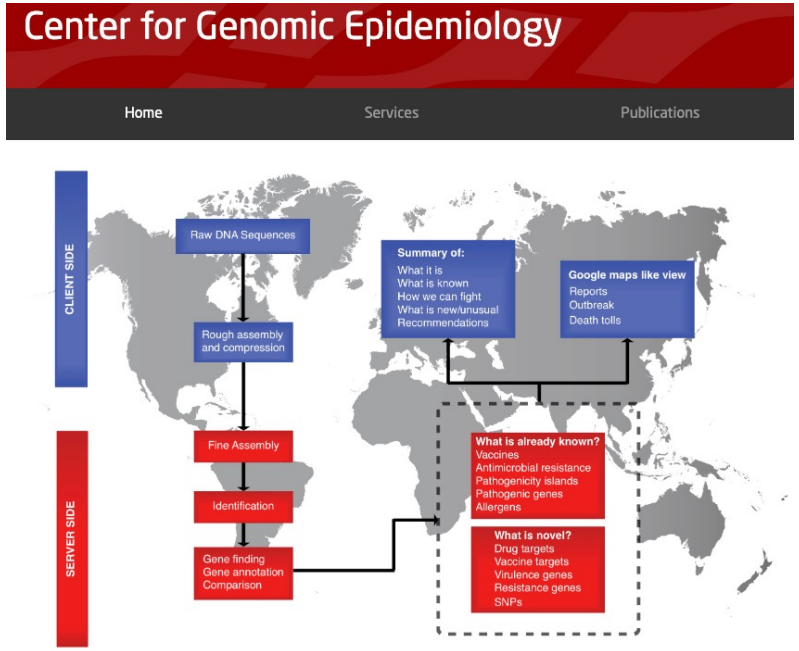
Live exercise

Using command line/terminal

```
spades -1 read1.fastq -2 read2.fastq --careful --only-assembler -t 48 --cov-cutoff auto -o Sample1
```

```
abricate --db resfinder CFS1073_ONT.fasta
```

Bioinformatic tools for characterising bacterial pathogens



Welcome to the Center for Genomic Epidemiology

The use of sequencing technologies is currently transforming almost every aspect of biological science. In relation to infectious diseases, the advances are rapidly changing our scientific discoveries, as well as diagnostic and outbreak investigations. The ability to analyze sequencing data and take advantage of the rapid progress, is however, not equally distributed between institutions and countries.

The aim of the Center for Genomic Epidemiology (CGE) is to provide access to bioinformatics resources also for those with limited experience and thereby allow all countries, institutions and individuals to take advantage of the novel sequencing technologies. Doing so, we hope to facilitate more open data sharing around the world and provide more equal opportunities for all.

CGE is entirely non-commercial and operates a number of free online bioinformatics services. Funding is provided as core funding from the Technical University of Denmark (DTU) and from a range of public and private sources as listed below.

If you want to read more about us and our research activities, please visit the Global Surveillance website.

Overview of Services

Phenotyping

ResFinder
Identfcation of acquired antibiotic resistance genes.

ResFinderFG
Identfcation of functional metagenomic antibiotic resistance determinants.

LRE-finder
Identfcation of genes and mutations leading to linezolid resistance.

KmerResistance
Identfcation of acquired antibiotic resistance genes using Kmers.

PathogenFinder
Prediction of a bacteria's pathogenicity towards human hosts.

VirulenceFinder
Identfcation of acquired virulence genes.

Restriction-ModificationFinder
Determination of Restriction-Modification sites (based on REBASE.)

SPIFinder
SPIFinder identifies Salmonella Pathogenicity Islands.

ToxFinder
ToxFinder identifies genes involved in mycotoxin synthesis.

Typing

MLST
Multi Locus Sequence Typing (MLST) from an assembled genome or from a set of reads.

PlasmidFinder
PlasmidFinder identifies plasmids in total or partial sequenced isolates of bacteria.

pMLST
Multi Locus Sequence Typing (MLST) from an assembled plasmid or

Phylogeny

MINTyper
Identification of SNPs with automatic filtering, masking and site validation together with inferred phylogeny based on both long and short sequencing data.

CSIPhylogeny
CSI Phylogeny calls SNPs, filters the SNPs, does site validation and infers a phylogeny based on the concatenated alignment of the high quality* SNPs.

NDtree
NDtree constructs phylogenetic trees from Single-End or Pair-End FASTQ files.

Evergreen
Evergreen generates a forest of constantly updated phylogenetic trees with publicly available whole-genome sequencing data from foodborne, bacterial isolates that were deposited in the short sequencing read archives (NCBI SRA/ENA).

TreeViewer
Phylogeny Tree Viewer.

Metagenomics

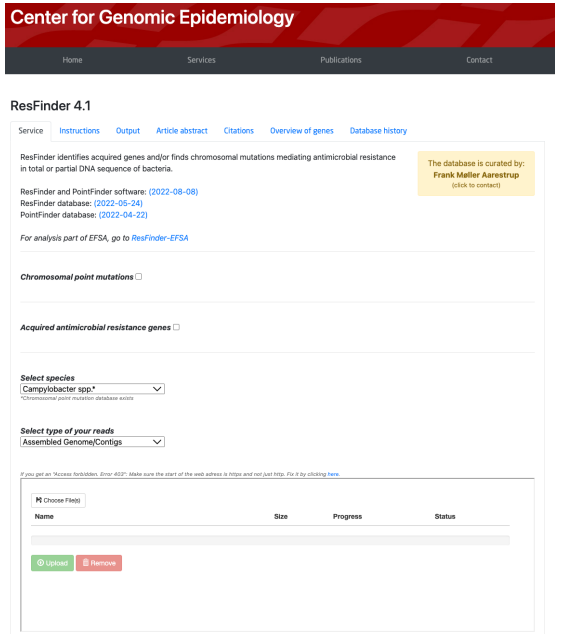
CCMetagen
CCMetagen: Comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data.

PCR-tools

RUCS
RUCS: Rapid Identification of PCR Primers Pairs for Unique Core Sequences.

Other

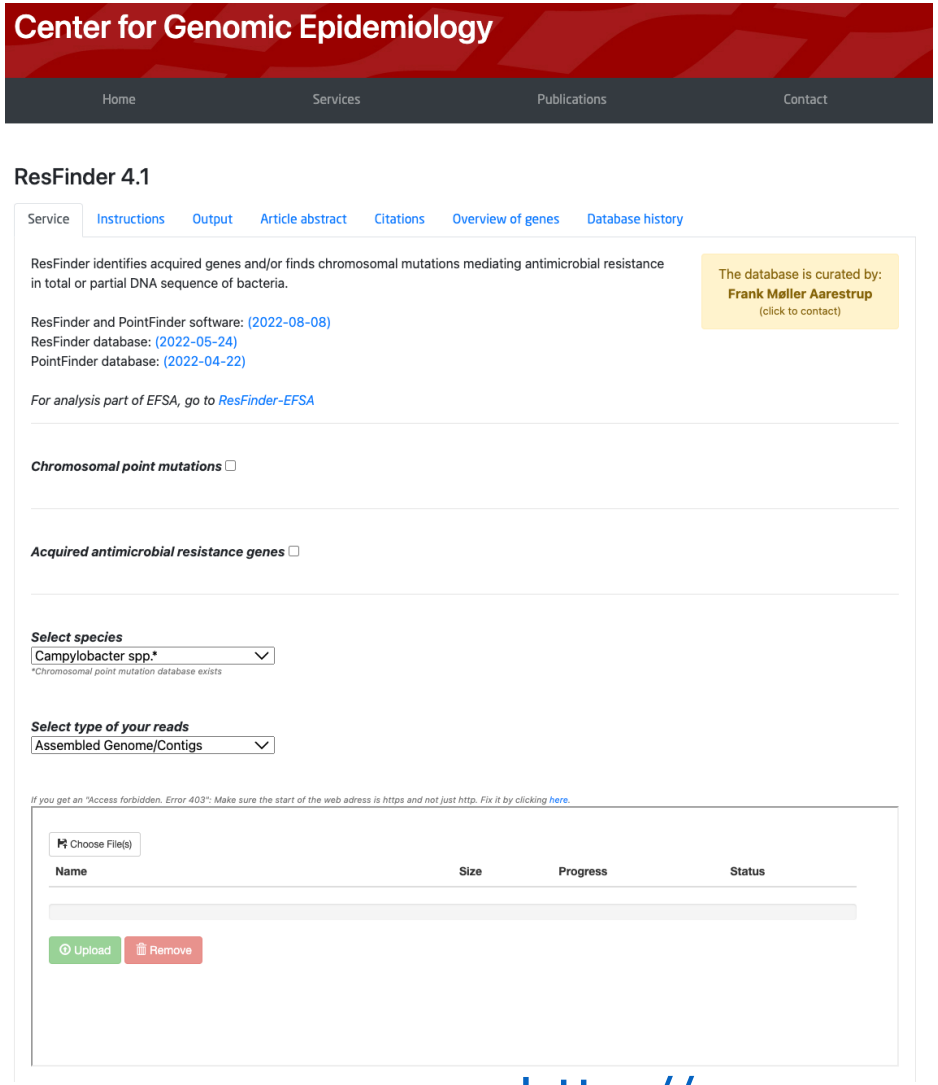
MyKMAfinder



Live exercise
web resources

http://www.genomicepidemiology.org/services/

Bioinformatic tools for characterising bacterial pathogens



Try different services using the sequence (fasta file) included in the lecture material!

Live exercise
web resources
http://www.genomicepidemiology.org/services/

Bioinformatic tools for characterising bacterial pathogens

Galaxy Europe

The homepage of the European Galaxy community

Galaxy is an **open-source** platform for **FAIR data analysis** that enables users to:

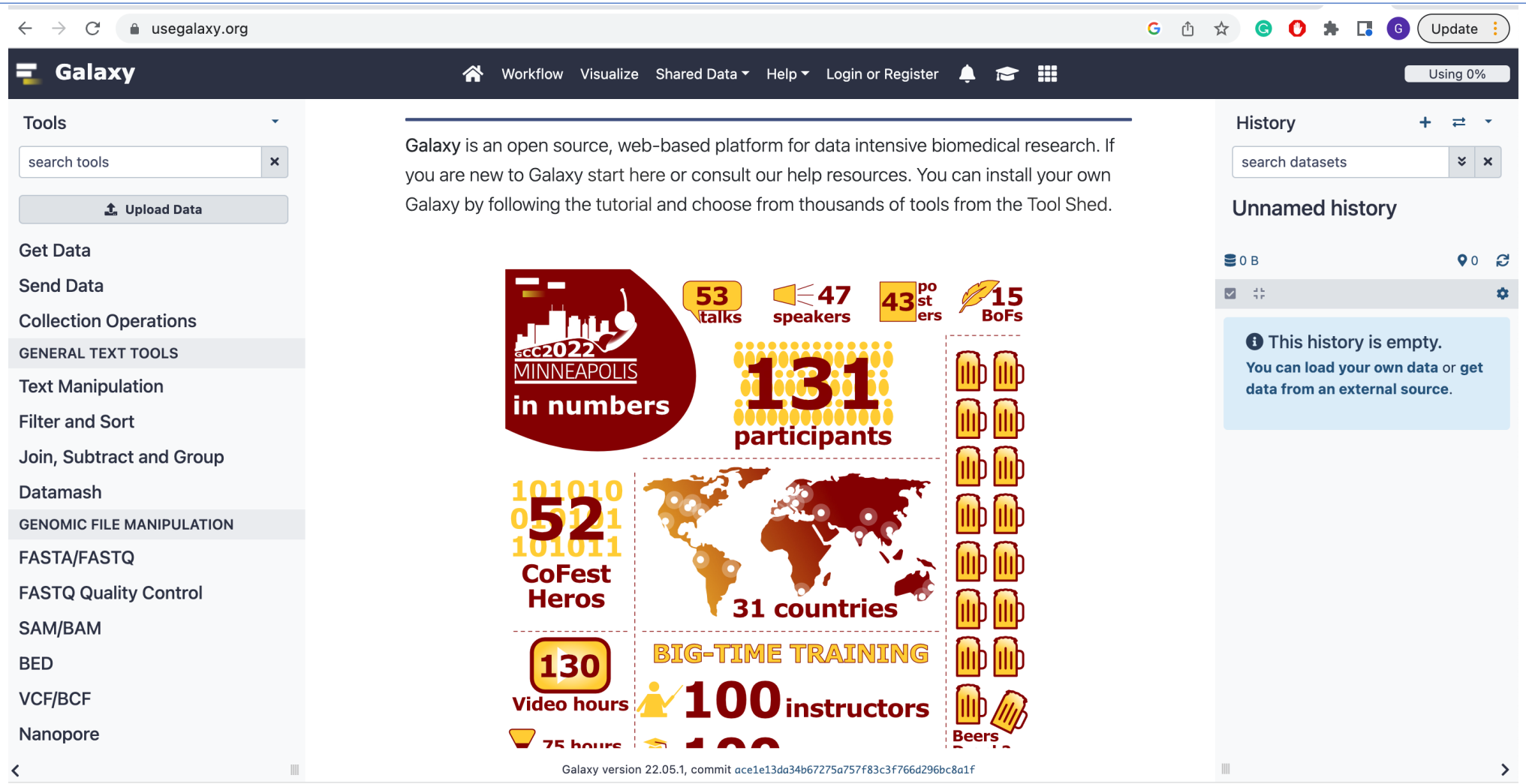
- use **tools** from various domains (that can be plugged into **workflows**) through its graphical web interface.
- run code in **interactive environments** (RStudio, Jupyter...) along with other tools or workflows.
- **manage data** by sharing and publishing results, workflows, and visualizations.
- **ensure reproducibility** by capturing the necessary information to repeat and understand data analyses.

The **Galaxy Community** is actively involved in helping the ecosystem improve and sharing scientific discoveries.

Live exercise

Using https://galaxyproject.org/eu/

Bioinformatic tools for characterising bacterial pathogens



Live exercise

Using https://galaxyproject.org/eu/

Summary and Glossary

- **Big data/grid/cloud** - With the increasing volume and heterogeneity of data sets (often referred to as “Big Data”), high performance computing is needed for analysis of the data. Many bioinformatics methods have been adapted to run on clusters of multiple computers (grid computing) and on large remotely located servers (cloud computing). Galaxy and KNIME are two popular software solutions to integrate and distribute larger data analysis tasks to the grid/cloud. Examples of cloud-based bioinformatics applications: HBLAST (BLAST, the most used bioinformatics sequence alignment tool); TPP, a proteomic analysis tool; HIPPIE: promoter analysis provided as Amazon Machine Image; BG: bacterial genome annotation based on Amazon Web Services .
- **Data mining** - Statistical and machine learning techniques to determine trends in typically large data sets. Unsupervised techniques (sample grouping is not explicitly used in the analysis) include: principal component analysis (PCA) and clustering algorithms (e.g. K-means, hierarchical). Supervised techniques (sample grouping is taken into account) include: ANOVA, Mann–Whitney U test, partial least squares analysis (PLS), machine learning (e.g. by support vector machines (SVM) and random forest): a computational model is trained to use properties derived from samples to predict the status of samples.
- **Virtual machines (VM)** - A large computer file (disk image) that consists of an operating system (e.g. Linux), software tools and data. The image can be run on an actual computer using virtual machine software that emulates an actual computer. In other words, a computer in a computer. The advantage of VMs is that they are portable (can be run on many different types of computer hardware), easy to backup and more straight-forward to maintain. Examples of the use of VMs are the generic bioinformatics tools in the NEBC Bio-Linux distribution and the 16s analysis suite Qiime
- **Databases** - Databases are organized collections of biological data. Bioinformatics is only successful if databases with high-quality data are available, together with structured vocabularies that describe the content of the data sets. An updated overview of relevant biological databases can be found here: http://www.oxfordjournals.org/our_journals/nar/database/cap/.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4793891/

Summary and Glossary

- **Genome sequencing** - Determining the complete genome sequence of a microbial strain of interest. Next-generation sequencing (NGS) techniques allow for high-throughput and high-quality sequencing results. Especially the combination if different techniques (e.g. Illumina and Pacific Biosystems or PacBio) result in high-quality (circular) genomes.
- **Sequencing data (FASTQ)** - Sequencing data are represented in FASTQ format. These files provide, next to the raw sequence data, additional information regarding the quality of the reads. In this manner, quality control and trimming can be applied.
- **Assembly** - Raw sequence reads of different NGS technologies can be assembled into contigs, long stretches of DNA sequence representing part of the genome. Most of the assembly methods are based on alignment of sequence reads with each other (de novo assembly) or against a reference genome (mapping assembly), thereby generating long DNA sequences from the fragments generated by the sequencing. Some examples of assembly tools are SPAdes, Ray, MIRA and IDBA.
- **Scaffolding** - Organizing the contigs from the assembly to larger, gapped, DNA sequences. Some NGS techniques (e.g. Illumina) allow the synthesis of paired end (PE) or mate pair (MP) libraries; libraries with a fixed insert size that are sequenced at both ends. As reads span a larger DNA fragment, the matched reads pairs can be used to order contigs, even if the sequence in between the contigs has not been assembled. In general, most assembly tools allow for scaffolding, but also dedicated tools exist, such as SSPACE.
- **Gap closure strategies** - After scaffolding, genome sequences will most often contain gaps. Common strategies to fill these gaps are generating new sequencing data using, for example, PacBio's long reads, or predicting the most likely order and orientation of the contigs using bioinformatics tools like Projector2 or Mauve. These tools infer contig order by comparing them to one or more reference sequences.
- **Gene function annotation** - Gene function is typically inferred from similarity in amino acid sequence. Gene functions can be predicted by comparing sequences to databases containing genes with known functions with tools like RAST and Prokka.
- **Orthology** - Genes in different organisms are orthologous when they were the same gene in the last common ancestor. Reconstructing the evolutionary history of genes allows the prediction of functional equivalence (i.e. orthologous genes are likely to have similar functions). Tools are OrthoMCL and Orthagogue.

Summary and Glossary

- **Comparative genomics** - All analyses in which genome sequences or genome content of multiple organisms are compared.
- **Metabolic modelling** - Prediction of growth, and recruitment of metabolic pathways, of microbes by using the genome sequence as an inventory of all possible metabolic reactions. Genome-scale metabolic models can be constructed using automated or comparative genomics analyses. Once constructed, the models can be used to simulate growth by, for example, flux balance analysis (FBA) and to determine the boundaries of fluxes by flux variability analysis (FVA). Tools for modelling are Pysces, the SEED and VANTED.
- **Microbiome analysis** - All microbes present in a particular niche are termed a microbiome. Analysis of microbiomes can be done using different next-generation sequencing-based techniques (see below).
- **16s rRNA sequencing** - 16s amplicon sequencing is the generation of sequence reads from conserved regions of the 16s gene. Amplicon sequencing (e.g. by Illumina) is used to identify the bacterial (and sometimes archaeal) component of microbial communities. Examples of software to infer community composition from sequencing data are Qiime and Mothur. 16s sequencing is a relatively cheap and well-established technique and as such an ideal starting point for characterization of complex cultures for which limited prior knowledge is available.
- **Functional prediction** - 16s sequences derived from a particular ecological niche indicate the taxa present and their relative abundance. From these data, presence of gene functions in those taxa can be performed using, e.g., PICRUSt. It infers the presence of gene functions in given taxa using already sequenced genomes part of the same taxa.
- **Shotgun metagenomics and metatranscriptomics** - Random fragments of the DNA or (enriched) mRNA of a given microbiome are sequenced with next-generation sequencing. Metagenomics and metatranscriptomics techniques are powerful, as they allow circumventing growing microbes while still determining their gene content or gene expression. This provides insight into the molecular functions encoded by the DNA, taxonomic assignment of that DNA fragments or inferring similar information for expressed mRNAs. This method can be used in addition to sequencing of individual isolates from complex cultures. Sequencing of individual isolates, however, has the advantage that comparative genomics can be done and metabolic models can be built more straight-forwardly, provided that the isolates under study are representative of the biodiversity present in the complex culture.

Summary and Glossary

- **Assembly** - Using the sequence overlap, the DNA/RNA-derived sequences can be assembled into larger contigs. Functional annotation of these larger fragments is more straight-forward, but the fraction of reads that can be assembled into contigs depends on both the complexity of the microbiome (many different microbes with varying abundances) as well as the presence of microdiversity (many different microbes with similar genome sequences).
- **Annotation** - Similar to the genome of a single bacterium, the sequences of a metagenome can be functionally and taxonomically annotated by comparing (assembled) sequences or predicted gene products against one or more reference databases with sequences with known functions from known taxonomic origin. Gene context such as operons are, however, primarily missing in shotgun metagenomics reads/contigs. A few tools are: PhymmBL (taxonomic classification using sequence-based models), MG-RAST (functional and taxonomic classification using alignment to reference databases) and MetaPhlAn (taxonomy prediction using taxon-specific marker genes).
- **Strain typing and tracking** -Pinpointing the presence of a particular microbe (strain) in a biological sample. Using MLST markers (multi-locus sequence typing), PCR based on unique DNA fragments or strain-specific markers, the abundance of particular strains can be followed during the course of a fermentation. Potential downside of these techniques is that only known biodiversity can be traced. Therefore, the performance must be evaluated on new strains. New biodiversity can be uncovered, provided that a genomic target is well-designed (e.g. targeting a gene that is single copy with sufficient resolution to distinguish between strains).
- **Predicting phenotypes** - Gene–trait matching: machine learning or statistics methods are used to predict the phenotype of a bacterial strain based on the presence/absence of particular genes, (parts of) metabolic pathways or classifications from experts. Transcriptome–trait matching: gene expression data (based on microarray or RNAseq) instead of gene presence are used. Transcriptome data from multiple strains grown under the same condition or the same strain grown under different conditions can be used.
- **Metabolomics** -The simultaneous measurement of multiple metabolites in biological samples. Metabolomics is a technique that can be applied to describe reaction products of microorganisms in defined media and in food samples. Its data are very suitable to be associated to results from sensory measurements.

Summary and Glossary

- consensus sequence - is the calculated order of most frequent residues, either nucleotide or amino acid, found at each position in a sequence alignment. It serves as a simplified representation of the population. It represents the results of multiple sequence alignments in which related sequences are compared to each other and similar sequence motifs are calculated. Such information is important when considering sequence-dependent enzymes such as RNA polymerase.