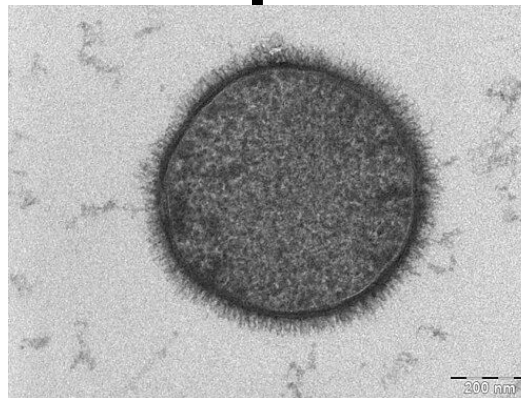


Asemblacja genomu *Bacillus subtilis* z odczytów Nanopore

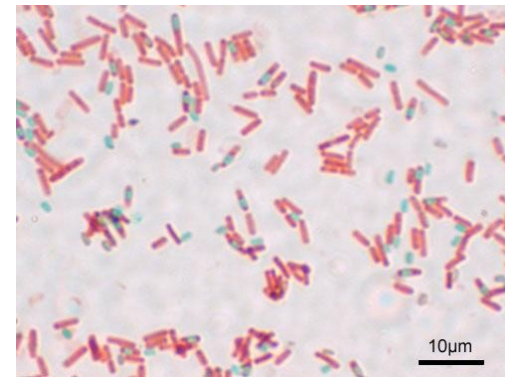


Genomika Funkcjonalna
Gorzelańczyk, Konrad

Bioinformatyka(3 rok),
Wydział Informatyki I Telekomunikacji,,
Politechnika Poznańska



Przekrój poprzeczny *B. subtilis*
(mikrofotografia wykonana przy pomocy
[TEM](#)). Średnica komórki wynosi ok. 0,7
μm.



Endospory laseczki siennej

Najważniejsze fakty o genomie (co wiemy “twardo”)

Referencyjny szczep laboratoryjny: *B. subtilis* subsp. *subtilis* str. 168

- **Chromosom:** kolisty, kompletna sekwencja RefSeq **NC_000964.3** ma **4,215,606 pz.**
- **Zawartość GC:** około **43.5%** (często podawana jako średnia dla genomu tego gatunku i szczepu referencyjnego).
- **Liczba genów** (KEGG dla 168): około **4174 genów białkowych** oraz **178 genów RNA**.
- **Kamień milowy genomiki:** pełna sekwencja *B. subtilis* była jednym z pierwszych wielkich genomów bakterii Gram-dodatnich opisanych w literaturze (Nature, 1997).
- **Dużo duplikacji i transporterów:** w pracy Nature wskazano m.in. rozrost rodzin genów, w tym dużą rodzinę transporterów ABC (ATP-binding transport proteins).

Elementy mobilne i “sprzątanie” genomu

Warianty inżynieryjne pochodne 168 bywają odchudzane przez usuwanie **profagów i wysp AT-bogatych**, żeby zwiększyć stabilność i przewidywalność w biotechnologii (np. szczepy typu delta 6).

Sprawozdanie z asemblacji genomu prokariotycznego: *Bacillus subtilis* (Nanopore)

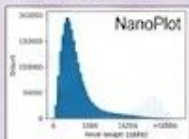


KONTEKST PROJEKTU: Asemblacja *de novo* genomu bakterii *B. subtilis* przy użyciu długich odczytów Nanopore.



1. DANE WEJŚCIOWE I QC

Surowe Odczyty (FASTQ)



Seqkit/NanoStat: ~26k odczytów
Suma: ~178 Mb
N50: ~9.1 kb
Max dl.: ~122 kb
Jakość śr.: ~Q8 (typowo dla raw Nanopore)

2. MAPOWANIE KONTROLNE

Minimap2 (map-ont) -> Samtools -> Mosdepth



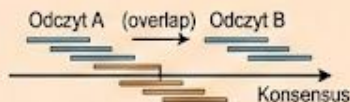
Do referencji *B. subtilis*
Zmapowano: ~99%
Średnie pokrycie: ~41x

WNIOSEK:

Dane wystarczające do asemblacji

3. ASEMBLACJA DE NOVO (FLYE)

Teoria OLC (Overlap-Layout-Consensus)



FLYE ASEMBLER (dane ~44x)

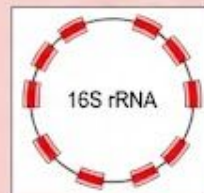
WYNIK GŁÓWNY:
1 KONTIG (Chromosom kolisty)
Długość: ~4.03 Mb

4. WIZUALIZACJA I POWTÓRZENIA

Bandage & Barnap/BLAST



Graf asemblacji (44x):
Rozwiązany



WYZWANIE: Operony rRNA
Barnap wykrył: 10 kopii 16S
(regiony powtarzalne)

5. WALIDACJA JAKOŚCI (QUAST & CheckM)

QUAST (vs. Referencja)



Misassemblies: 0
Genome fraction: >99.9%
Indels/Mismatches:
Wysokie (cecha surowego Nanopore)

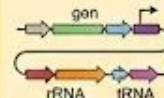
CheckM (Marker genes)



Kompletność: ~81%
Kontaminacja: ~1%

6. ADNOTACJA (PROKKA)

Prokka Pipeline



Znalezione cechy:
CDS (białka): 8189
rRNA: 29 (pełny zestaw)
tRNA: 82

EKSPERYMENT PORÓWNAWCZY: WPŁYW POKRYCIA NA GRAF OLC



PEŁNE DANE (~44x)

Wynik Flye: 1 Kontig

Wniosek: Wysokie pokrycie rozwiązuje powtórzenia



DOWNSAMPLING (~20x)

Wynik Flye: 2 Kontigi (3 disjointigs)
Efektywne pokrycie overlaps: ~15x

Wniosek: Niskie pokrycie zrywa graf w trudnych miejscach

Illumina vs PacBio vs Nanopore (w kontekście asemblacji)

Illumina



krótkie odczyty (50–300 bp),
bardzo wysoka dokładność (Q30+)

PacBio HiFi



długie odczyty (10–25 kb) +
wysoka dokładność

Nanopore

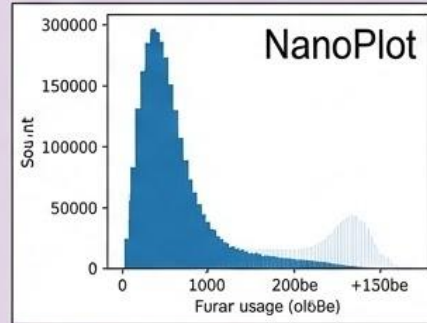
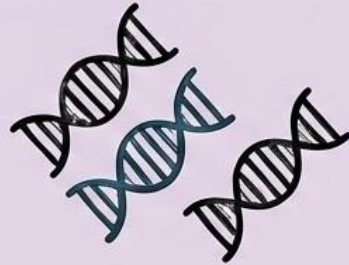


bardzo długie odczyty
(10–100 kb, ultra-long >100 kb),
surowo większy błąd,
ale świetne do repeatów i SV

Q-score: $Q = -10 \log_{10}(P_{\text{błędu}})$, np. Q30 = 0.1% błędu

1. DANE WEJŚCIOWE I QC

Surowe Odczyty (FASTQ)



Seqkit/NanoStat: ~26k odczytów

Suna: ~178 Mb

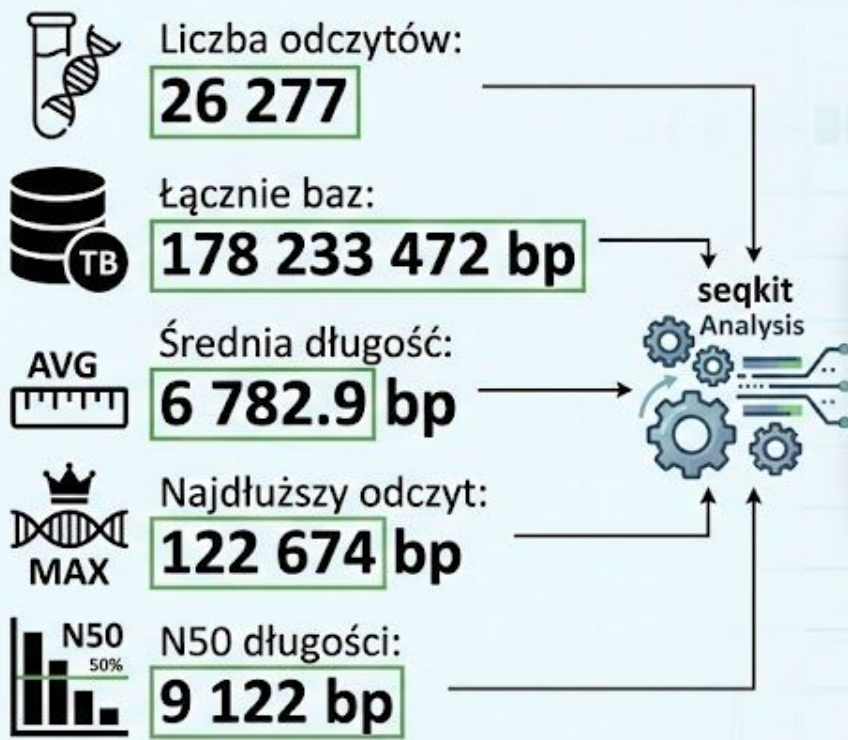
N50: ~9.1 kb

Max dł: ~122 kb

Jakość śr.: ~Q8 (typowo dla raw Nanopore)



FASTQ: ilość danych i długości odczytów (seqkit)



```
(nl) konrad_guest@konrad-hp:~/ExternalDisk/functional_genomics/gorzalanczyk_asenblacja$  
(ml) konrad_guest@konrad-hp:~/ExternalDisk/functional_genomics/gorzalanczyk_asenblacja$  
$ seqkit stats b_subtilis_reads.fastq.gz > seqkit_stats.txt  
-----  
file                format type  num_seqs  sum_len  min_len  avg_len  max_len  
-----  
b_subtilis_reads.fastq.gz  FASTQ  DNA      26,277   178,233,472   140     6,782.9   122,674  
(nl) konrad_guest@konrad-hp:~/ExternalDisk/functional_genomics/gorzalanczyk_asenblacja$  
$
```

Jakość odczytów i rozkłady (NanoStat + NanoPlot)



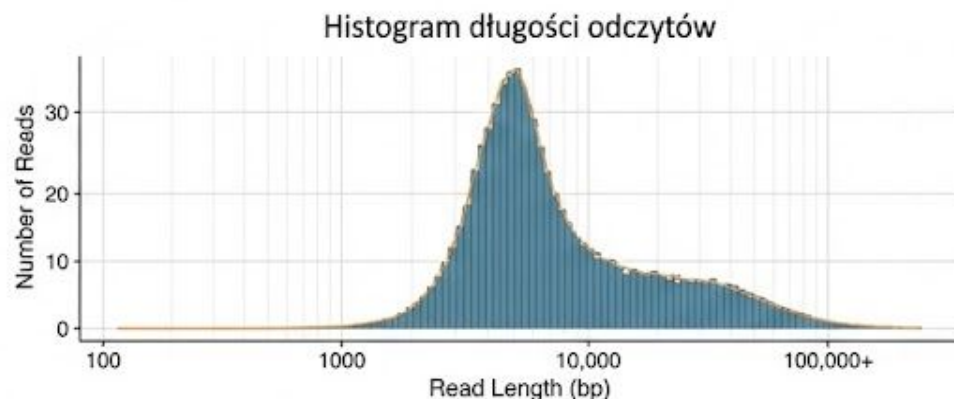
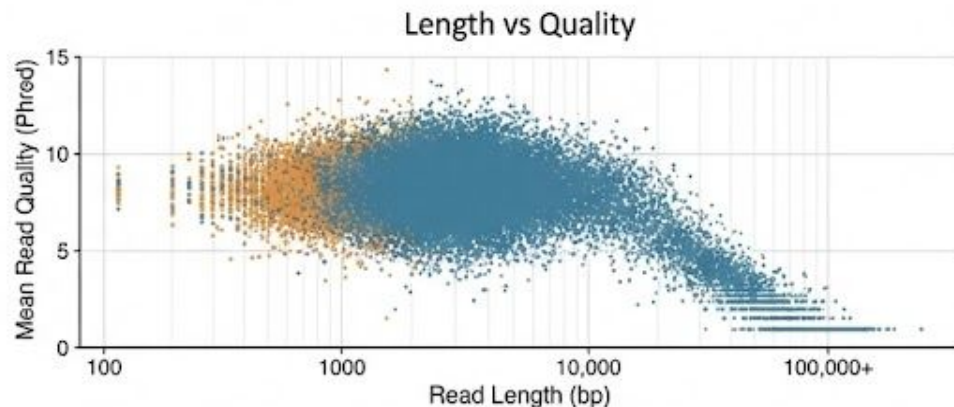
Średnia jakość: $Q \approx 8.1$
(typowe dla surowych ONT)



N50: 9.1 kb

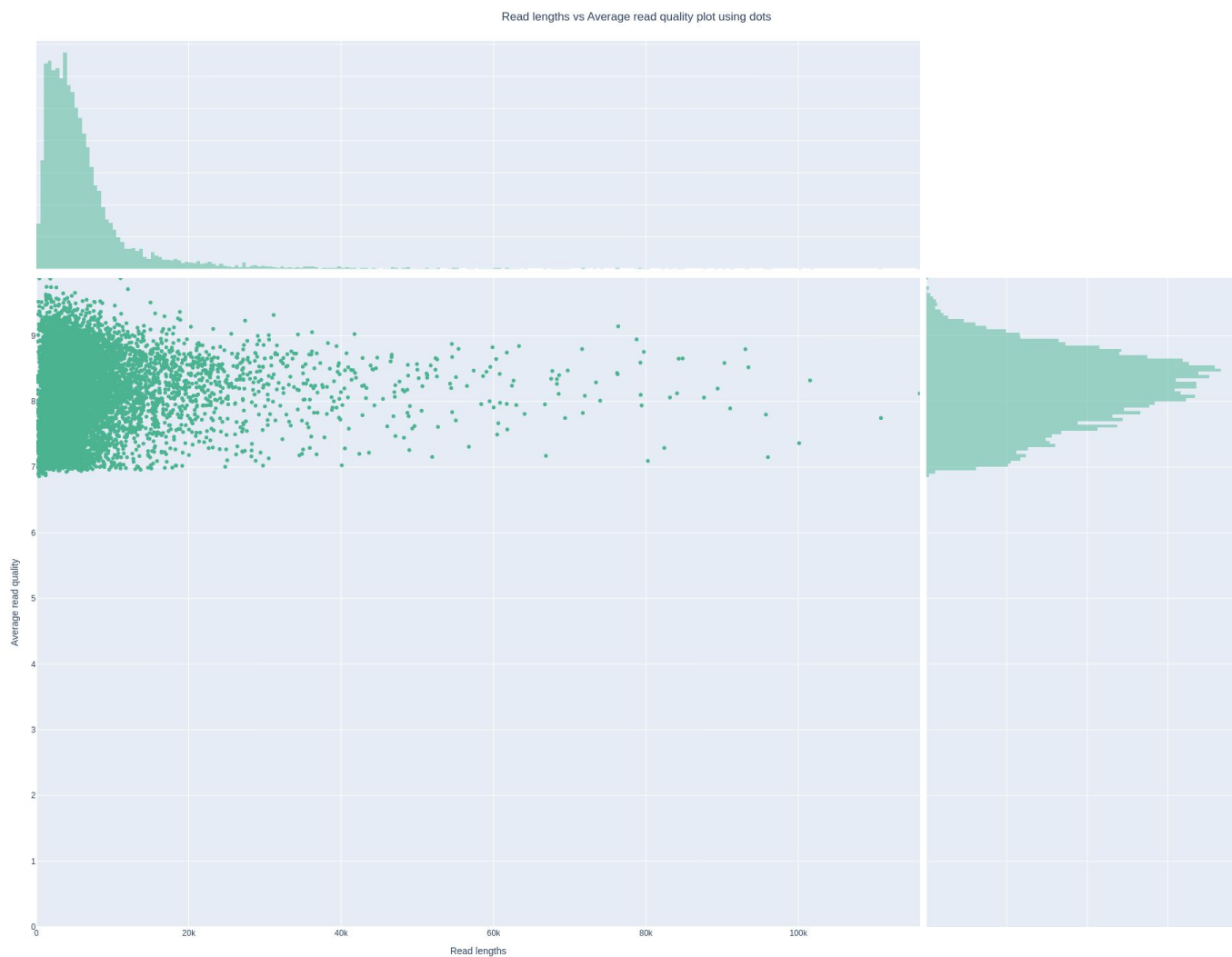


Obecne odczyty >100 kb

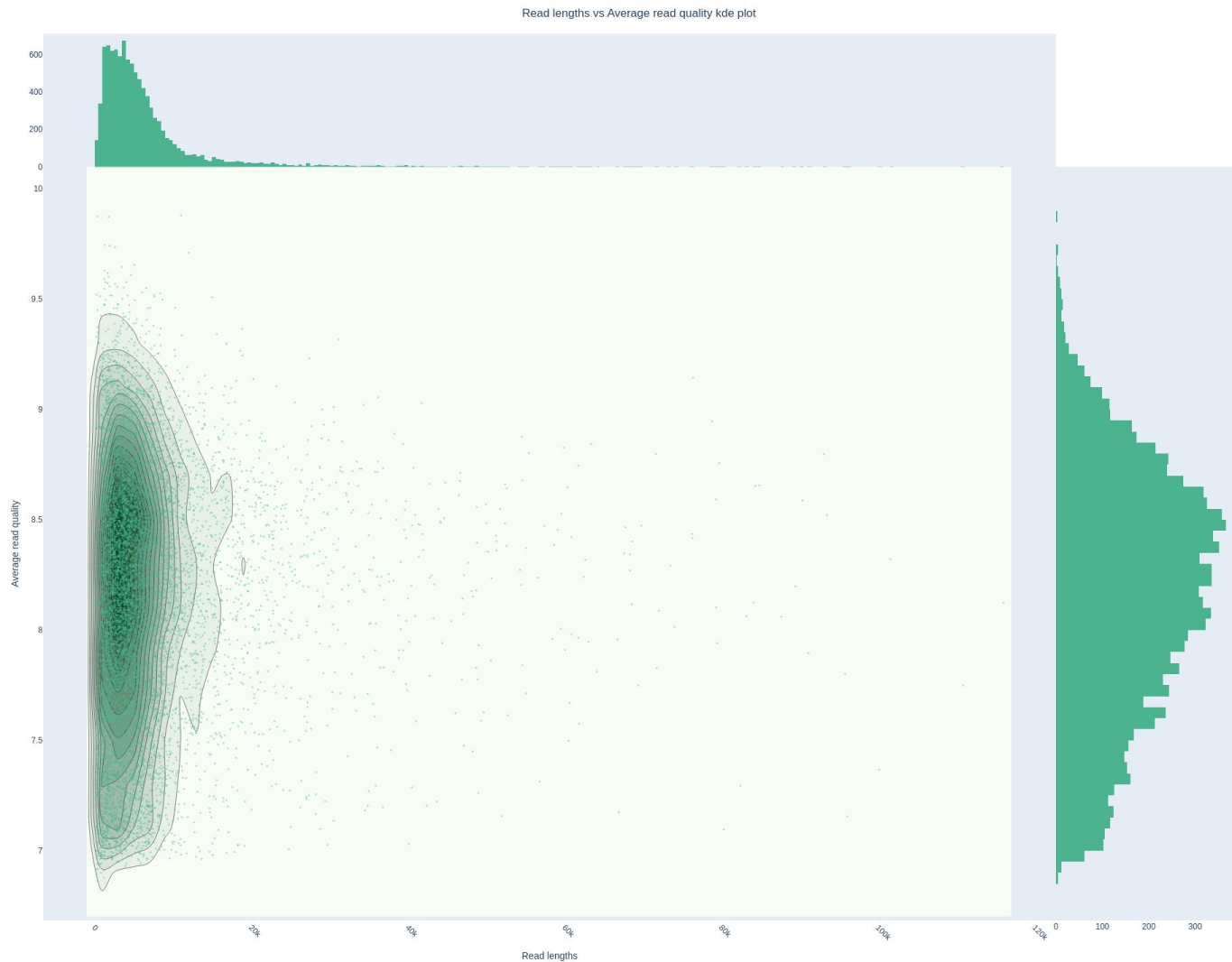


Kontekst: Asemlacja *Bacillus subtilis*, dane Nanopore.

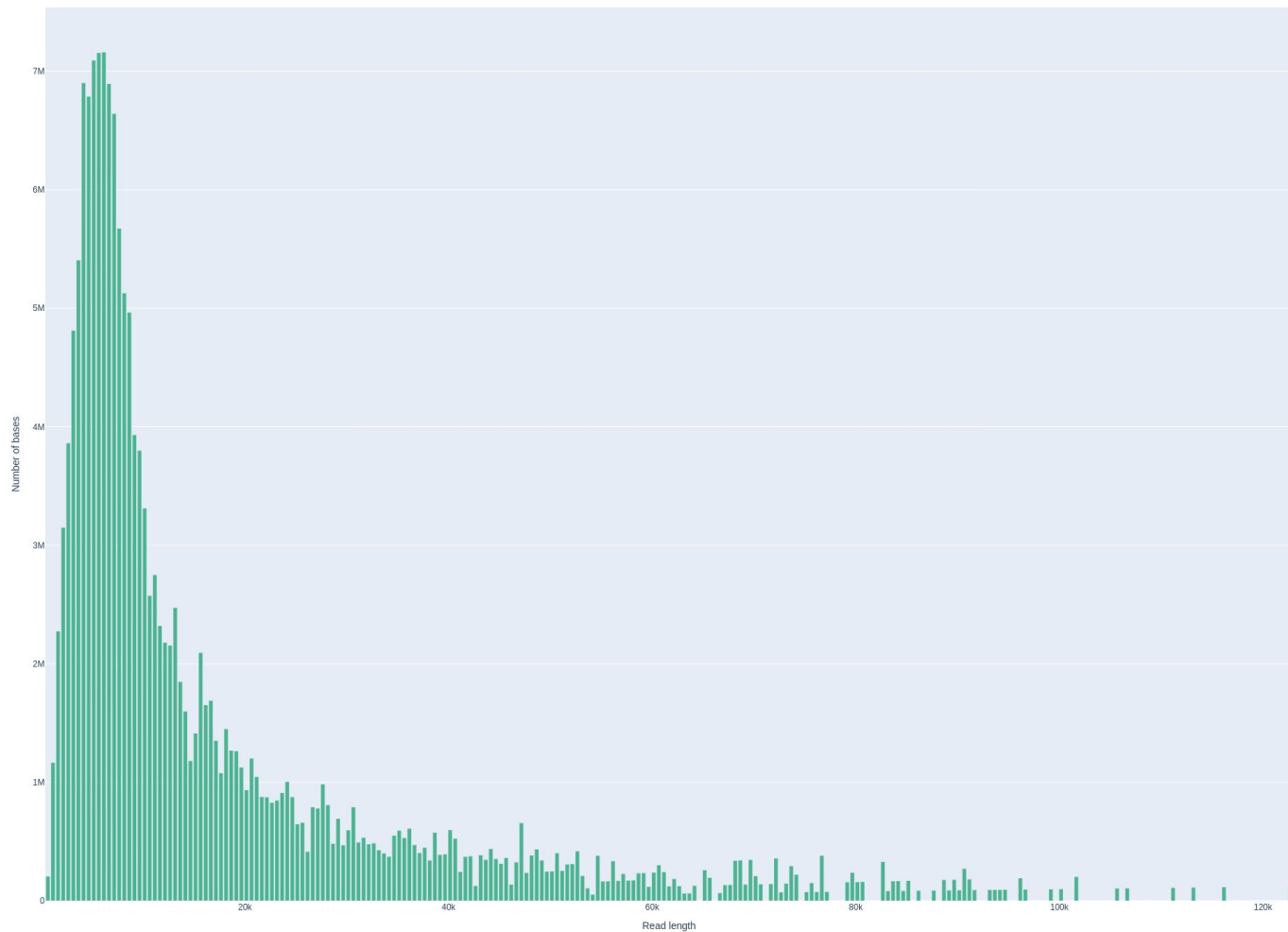
Read lengths vs Average read quality plot using dots



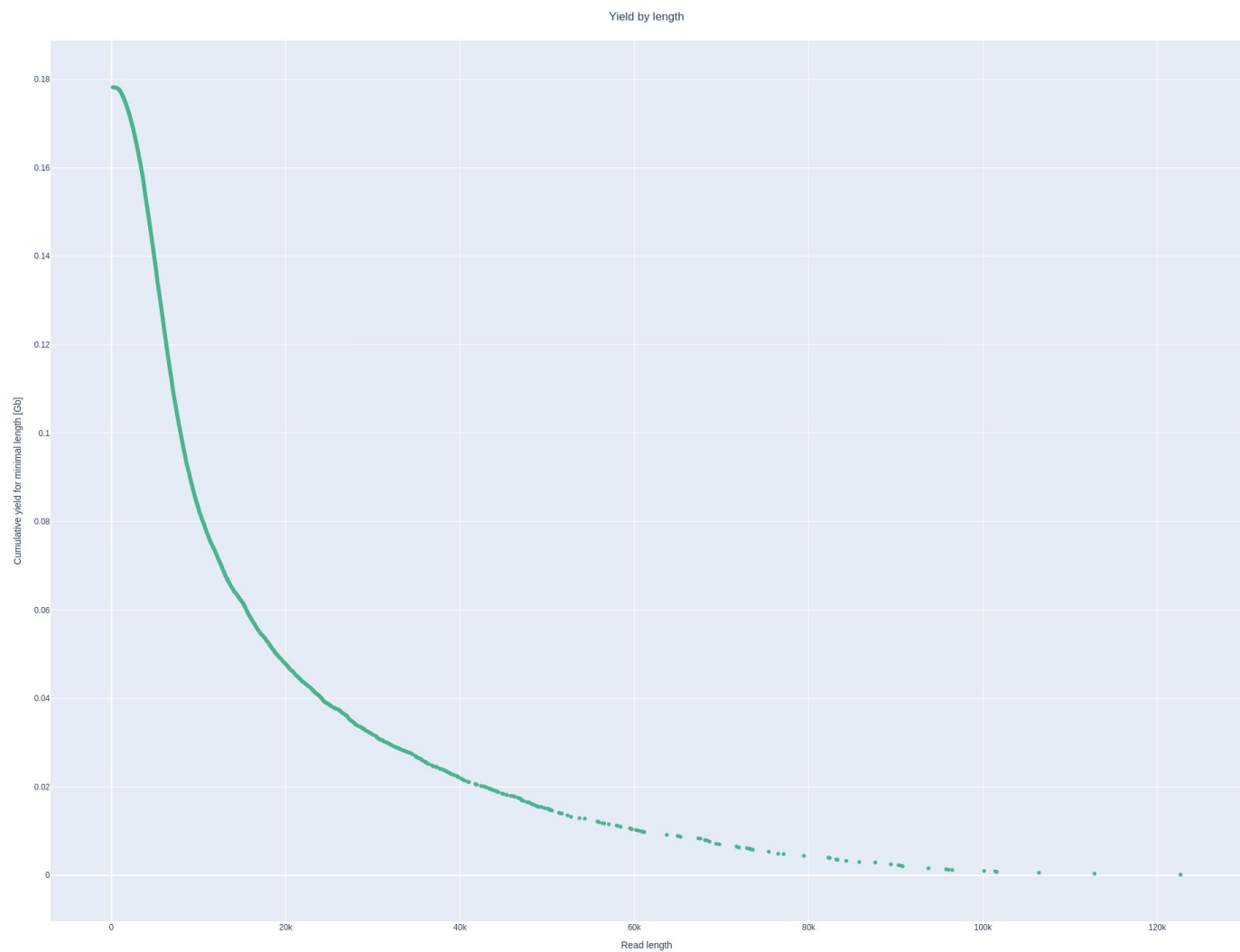
Read lengths vs Average read quality kde plot



Weighted histogram of read lengths

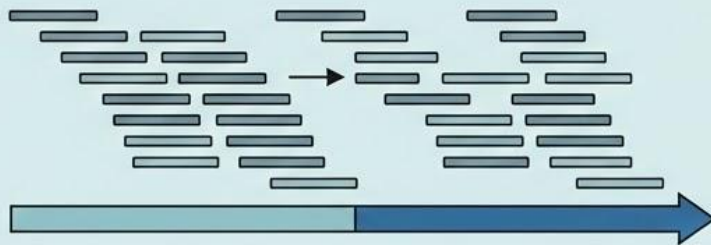


Yield by length



2. MAPOWANIE KONTROLNE

Minimap2 (map-ont) -> Samtools -> Mosdepth



Do referencji *B. subtilis*

Zmapowano: ~99%

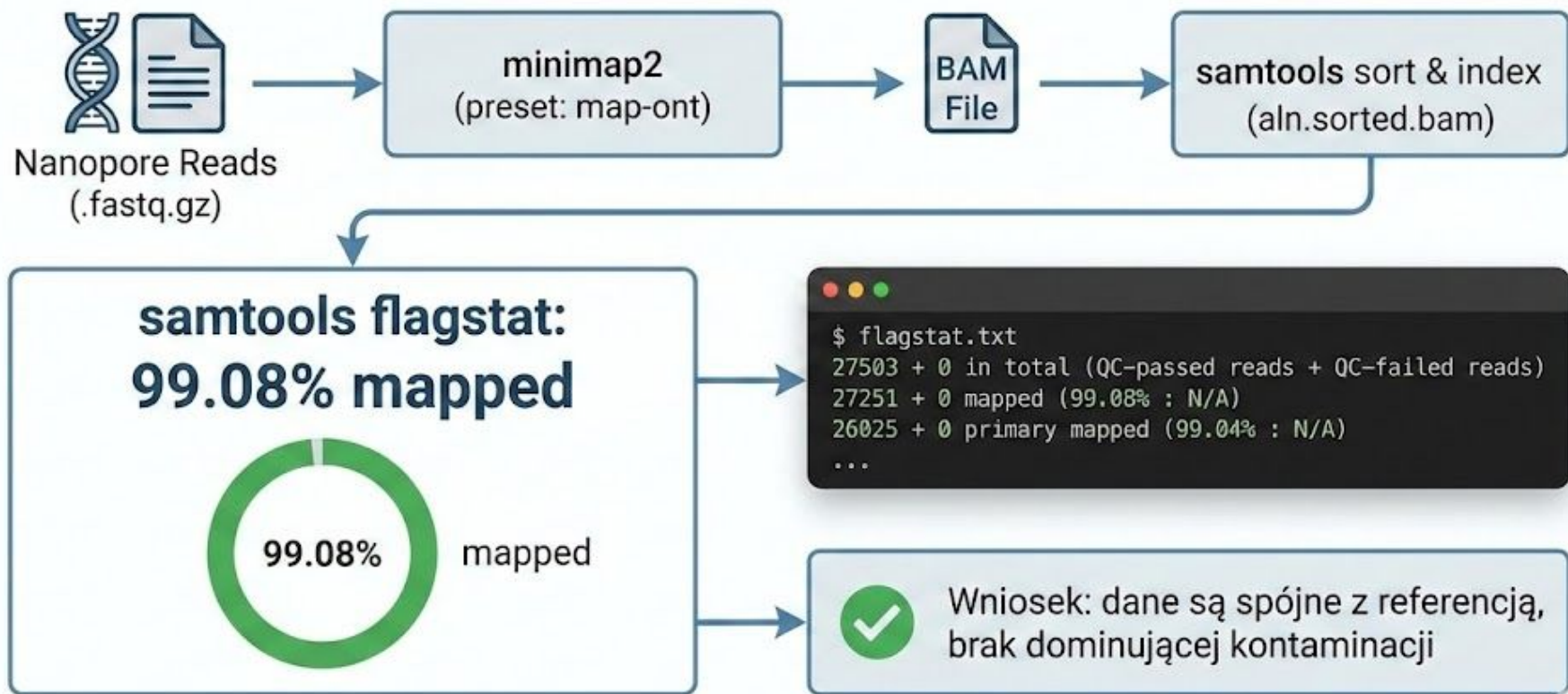
Średnie pokrycie: ~41x



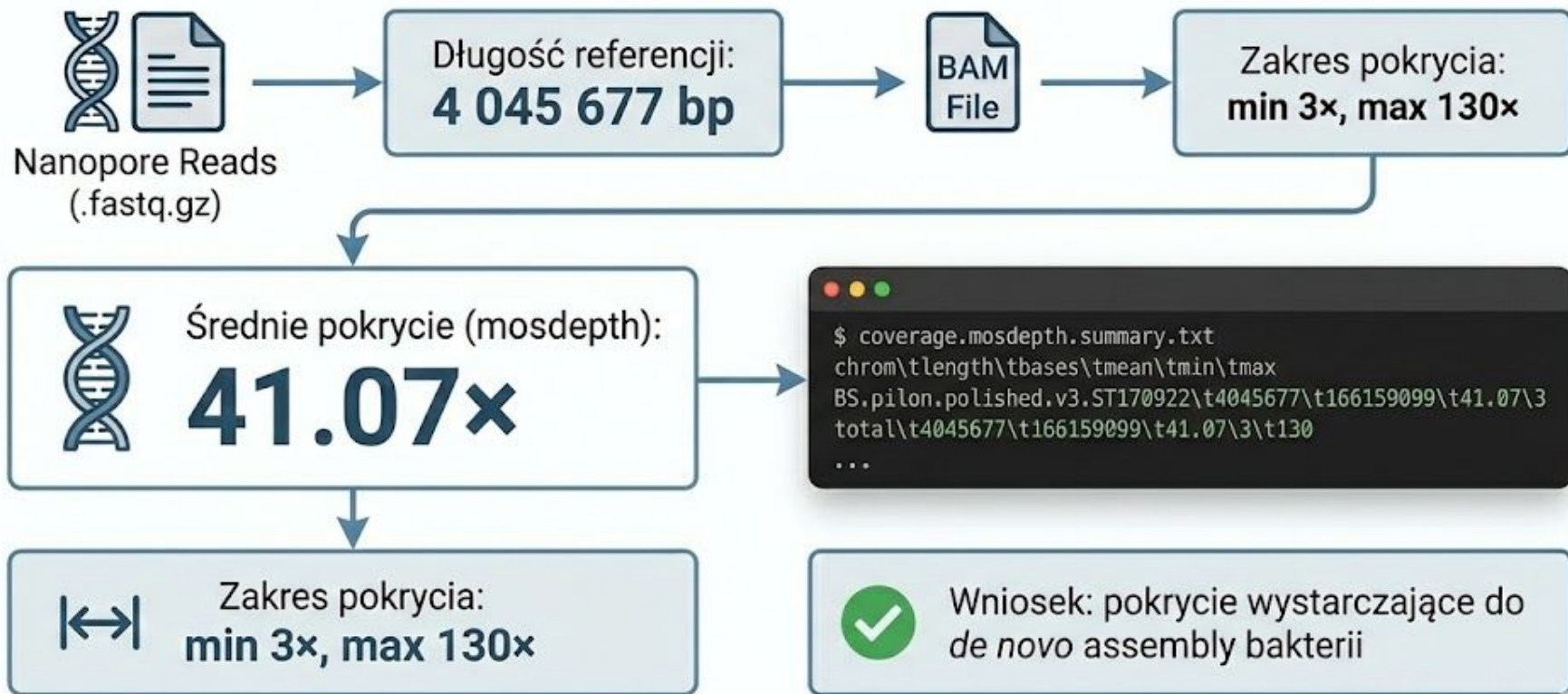
WNIOSEK:

Dane wystarczające do asemblacji

Kontrola danych przez mapowanie (minimap2 + samtools)



Pokrycie (coverage) genomu



Ocena “czy starczy do złożenia genomu?”



Teoretycznie:

$$\frac{178.2 \text{ Mb}}{4.05 \text{ Mb}} \approx 44\times$$



W praktyce (mosdepth + samtools)

~41x
(pokrycie)

+ 99%
mapped

Długie odczyty (Nanopore)



N50: **9.1 kb**

Max: **122 kb**



Odpowiedź:

TAK, dane są wystarczające do *de novo* assembly

Grafiy w asemblacji: OLC vs de Bruijn

OLC (Overlap-Layout-Consensus)

Overlap → Layout → Consensus

- Wierzchołki = odczyty 
- Krawędzie = nakładki (→ )

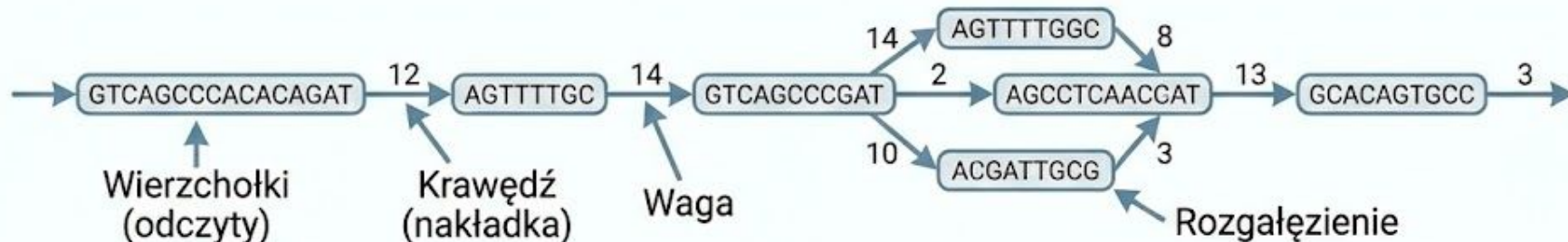
de Bruijn

GTCAG CCCAC k-MER
ATCAG CCCAC k-MER
CAGAT CAGAT GTTTT
ATATT
AGGTT

- Wierzchołki = k-mery (k-mery)
- Dobrze dla krótkich odczytów 



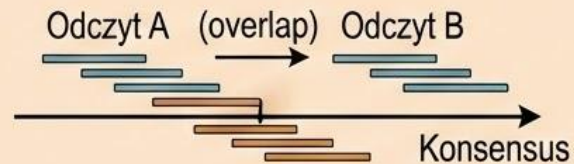
Nanopore często "idzie" w stronę OLC / overlap-based rozwiązań



Przykład grafu nakładania (OLC) z wagami (długość nakładania/wynik dopasowania)

3. ASEMBLACJA DE NOVO (FLYE)

Teoria OLC (Overlap-Layout-Consensus)



FLYE ASSEMBLER (dane ~44x)

WYNIK GŁÓWNY:

1 KONTIG (Chromosom kolisty)
Długość: ~4.03 Mb

Flye: wynik asemblacji dla pełnego zestawu (~44x)

Flye command: `flye --nano-raw b_subtilis_reads.fastq.gz --genome-size 4.1m`

Wynik Asemblacji



1
contig



Długość contigu:
4 032 848 bp



Średnie pokrycie Flye:
~43x



Interpretacja:
"prawie zamknięty"
chromosom

Assembly statistics (Flye output)

```
Total length: 4032848
Fragments: 1
Fragments NS0: 4032848
Largest frg: 4032848
Scaffolds: 0
Mean coverage: 43
```

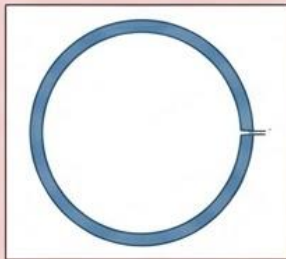
seqkit stats assembly

```
$ seqkit stats "$assembly"
file format type num_seqs sum_len min_len avg_len max_len
flye_out/assembly.fasta FASTA DNA 1 4,032,848 4,032,848 4,032,848 4,032,848
```

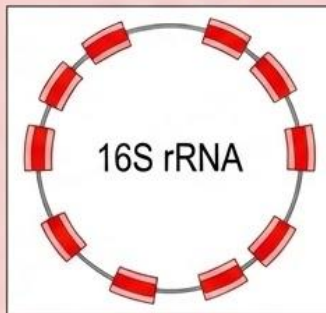
Name	Size	Modified	
 00-assembly	2 items	29 gru 2025	☆
 10-consensus	3 items	29 gru 2025	☆
 20-repeat	7 items	29 gru 2025	☆
 30-contigger	7 items	29 gru 2025	☆
 40-polishing	9 items	29 gru 2025	☆
 assembly.fasta	4,1 MB	29 gru 2025	☆
 assembly.fasta.fai	26 bytes	29 gru 2025	☆
 assembly.fasta.seqkit.fai	26 bytes	29 gru 2025	☆
 assembly_db.ndb	20,5 kB	29 gru 2025	☆
 assembly_db.nhr	71 bytes	29 gru 2025	☆
 assembly_db.nin	128 bytes	29 gru 2025	☆
 assembly_db.njs	499 bytes	29 gru 2025	☆
 assembly_db.not	20 bytes	29 gru 2025	☆
 assembly_db.nsq	1,0 MB	29 gru 2025	☆
 assembly_db.ntf	16,4 kB	29 gru 2025	☆
 assembly_db.nto	8 bytes	29 gru 2025	☆
 assembly_graph.gfa	4,0 MB	29 gru 2025	☆
 assembly_graph.gv	193 bytes	29 gru 2025	☆
 assembly_info.txt	92 bytes	29 gru 2025	☆
 flye.log	36,3 kB	29 gru 2025	☆
 params.json	92 bytes	29 gru 2025	☆

4. WIZUALIZACJA I POWTÓRZENIA

Bandage & Barrnap/BLAST



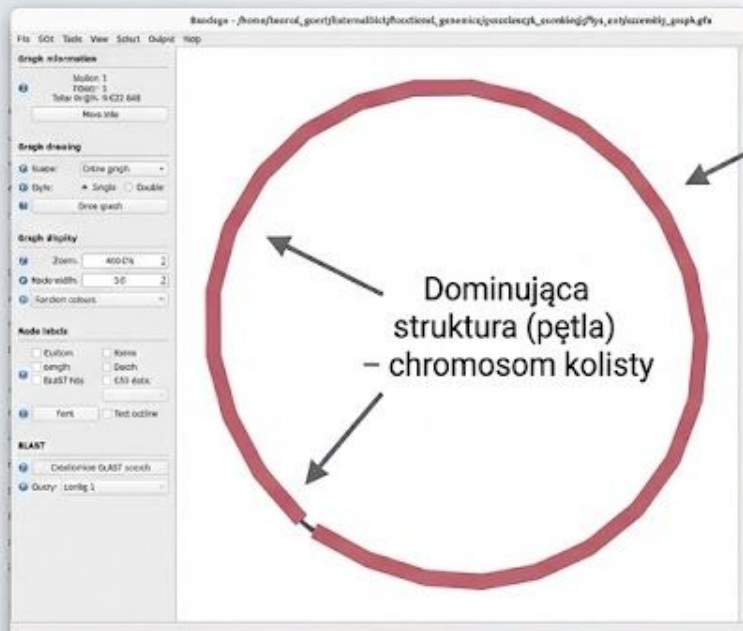
Graf asemblacji (44x):
Rozwiązany



WYZWANIE: Operony rRNA
Barrnap wykrył: 10 kopii 16S
(regiony powtarzalne)

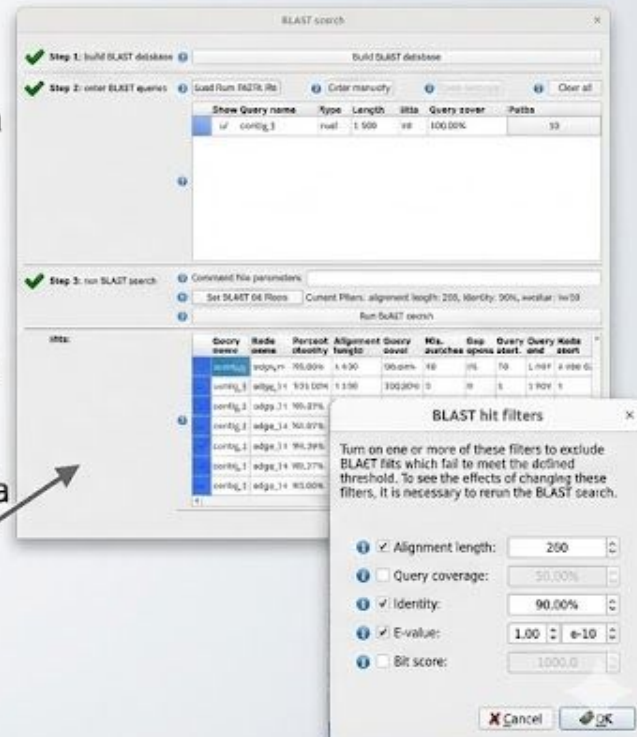
Bandage: – wizualna kontrola grafu + BLAST

Wizualizacja Grafu (GFA)



- **Graf GFA:** jedna dominująca struktura (spójna składowa)
- Dla bakterii często widoczna pętla (chromosom kolisty)
- Panel BLAST pozwala "przypiąć biologię" do grafu (np. 16S, geny oporności)

Panel BLAST: Przypisywanie Biologii



Graph information

Nodes: 1
Edges: 1
Total length: 4 032 848

More info

Graph drawing

Scope: Depth range

Min: 10,0

Max: 100,0

Style: ☒ Single ☐ Double

Draw graph

Graph display

Zoom: 473,5%

Node width: 5,0

Random colours

Node labels

☒ Custom

☒ Name

☒ Length

☒ Depth

☒ BLAST hits

☒ CSV data:

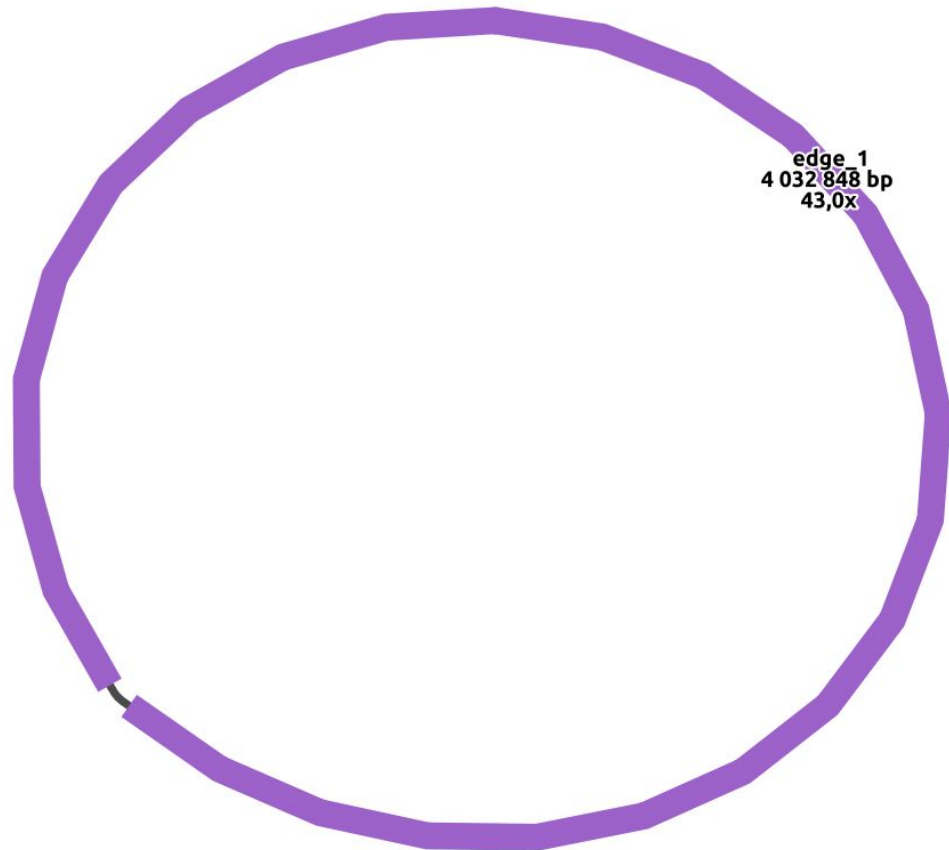
Font

☒ Text outline

BLAST

Create/view BLAST search

Query: none



Find nodes

Node(s):

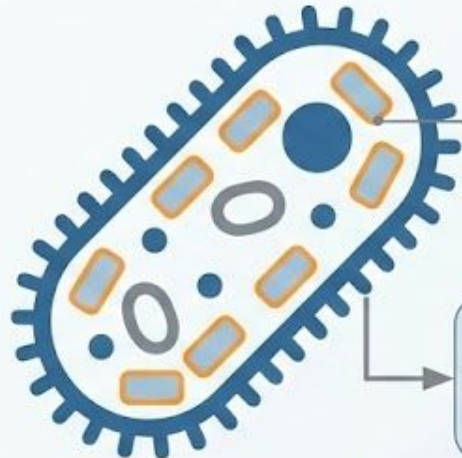
Match: ☒ Exact ☐ Partial

Find node(s)

Operony rRNA i powtórzenia: 16S

barrnap output:

```
[barrnap] Found 29 ribosomal RNA features.  
Found: 16S_rRNA contig_1 ...  
Found: 23S_rRNA contig_1 ...  
Found: 5S_rRNA contig_1 ...
```



Liczba kopii
16S rRNA: 10



Wniosek: 16S jest wielokopiowy,
dlatego jest hotspotem
niejednoznaczności w grafach

BLAST results					
	Name	Node start	Node end	contigu	
		100%	100%		
BLAST	16S_rRNA	11	19	contig_1	
BLAST	16S_rRNA	70	76	contig_1	
BLAST	16S_rRNA	728	828	contig_1	
BLAST	16S_rRNA	1012	1036	contig_1	
BLAST	16S_rRNA	1568	1877	contig_1	
BLAST	16S_rRNA	2136	2180	contig_1	
BLAST	16S_rRNA	2434	2343	contig_1	
BLAST	16S_rRNA	2627	2345	contig_1	
BLAST	16S_rRNA	2354	2797	contig_1	
BLAST	16S_rRNA	3434	3917	contig_1	
BLAST	16S_rRNA	4383	4013	contig_1	
BLAST	16S_rRNA	4600	4308	contig_1	
BLAST	16S_rRNA	4575	5513	contig_1	

BLAST hits w Bandage pokazują
wielokrotne kopie 16S w różnych
miejscach kontigu

Graph information

Graph information

Total length

Graph drawing

Scope:

Min:

Max:

Style:

Graph display

Zoom:

Node width:

BLAST hits

Node labels

☐ Custom☐ Length☒ BLAST hits

Font

BLAST

Create

Query: all

Find nodes

Exact ☐ Partial

node(s)

BLAST search

Step 1: build BLAST database

Build BLAST database

Step 2: enter BLAST queries

Load from FASTA file

Enter manually

Clear selected

Clear all

Show	Query name	Type	Length	Hits	Query cover	Paths
☒	5S_rRNA::contig_1:10495-10594(-)	nucl	99	9	100,00%	9
☒	5S_rRNA::contig_1:1240119-1240228(+)	nucl	109	9	100,00%	9
☒	5S_rRNA::contig_1:161196-161305(-)	nucl	109	9	100,00%	9
☒	5S_rRNA::contig_1:140657-140766(-)	nucl	109	9	100,00%	9
☒	5S_rRNA::contig_1:80741-80850(-)	nucl	109	9	100,00%	9
☒	5S_rRNA::contig_1:3288863-3288973(-)	nucl	110	9	100,00%	9
☒	5S_rRNA::contig_1:4914-5021(-)	nucl	107	9	100,00%	9

Step 3: run BLAST search

Command line parameters:

Set BLAST hit filters

Current filters: none

Run BLAST search

Hits:

	Query name	Node name	Percent identity	Alignment length	Query cover	Mis-matches	Gap opens	Query start	Query end	Node start
#30...	16S_rRNA::contig_1:3583978-3585517(-)	edge_1-	100,00%	1 539	100,00%	0	0	1	1 539	447 332
#30...	16S_rRNA::contig_1:3583978-3585517(-)	edge_1-	98,83%	1 542	100,00%	8	10	1	1 539	3 866 61
#30...	16S_rRNA::contig_1:3583978-3585517(-)	edge_1-	98,70%	1 539	99,94%	13	7	1	1 538	3 887 15
#30...	16S_rRNA::contig_1:3583978-3585517(-)	edge_1-	98,51%	1 543	99,94%	14	9	1	1 538	3 947 32
#30...	16S_rRNA::contig_1:3583978-3585517(-)	edge_1-	98,38%	1 543	100,00%	15	10	1	1 539	739 148
#30...	16S_rRNA::contig_1:3583978-3585517(-)	edge_1-	98,25%	1 542	100,00%	20	7	1	1 539	4 017 58
#30...	16S_rRNA::contig_1:3583978-3585517(-)	edge_1-	97,80%	1 543	100,00%	20	14	1	1 539	3 953 20

Close

File Edit Tools View Select Output Help

Graph information

Nodes: 1
Edges: 1
Total length: 4 032 848

[More info](#)

Graph drawing

Scope: Depth range

Min: 10,0

Max: 100,0

Style: ☒ Single ☐ Double

[Draw graph](#)

Graph display

Zoom: 350,8%

Node width: 5,0

BLAST hits (rainbow)

Node labels

☐ Custom ☐ Name

☐ Length ☐ Depth

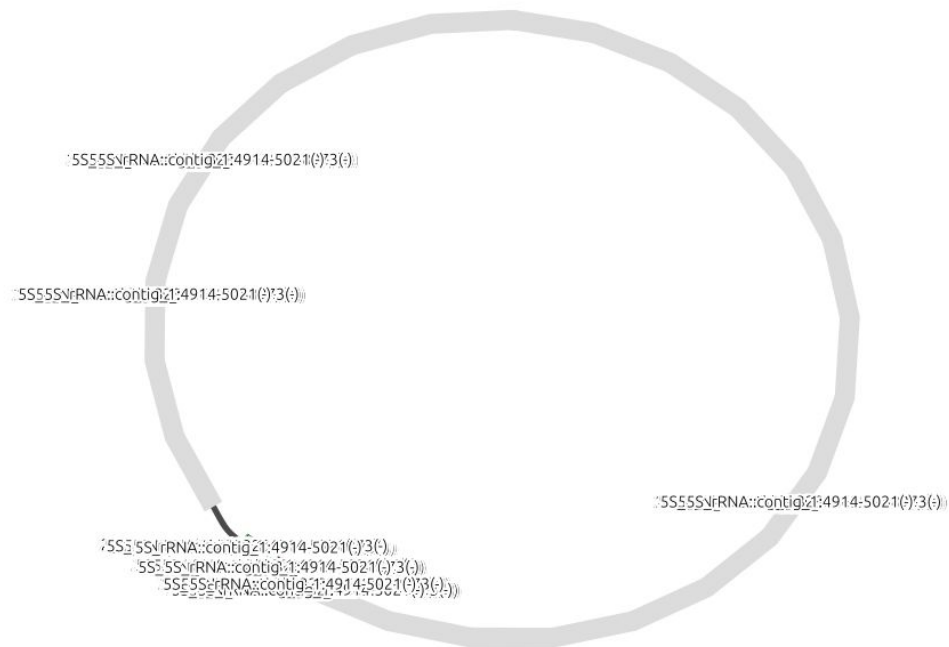
☒ BLAST hits ☐ CSV data:

[Font](#) ☒ Text outline

BLAST

[Create/view BLAST search](#)

Query: all



Find nodes

Node(s):

Match: ☒ Exact ☐ Partial

[Find node\(s\)](#)

Graph information

Nodes: 1
Edges: 1
Total length: 4 032 848

More info

Graph drawing

Scope: Depth range
Min: 10,0
Max: 100,0
Style: ☒ Single ☐ Double
Draw graph

Graph display

Zoom: 2391,9%
Node width: 5,0
BLAST hits (rainbow)

Node labels

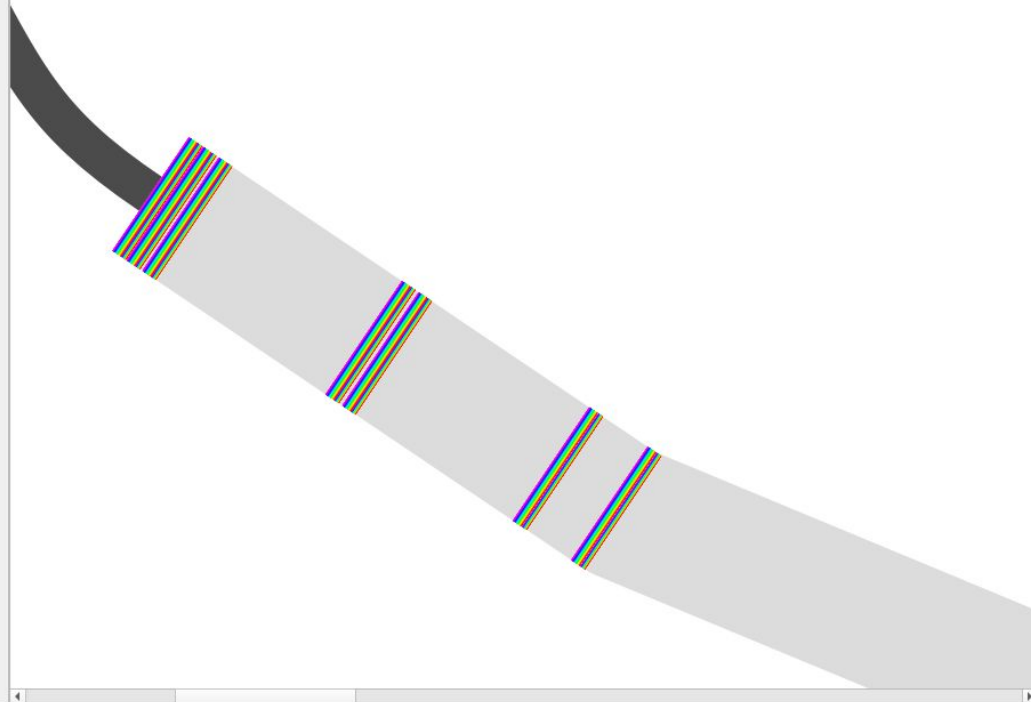
☐ Custom ☐ Name
☐ Length ☐ Depth
☐ BLAST hits ☐ CSV data:
Font ☐ Text outline

BLAST

Create/view BLAST search
Query: all

Find nodes

Node(s):
Match: ☒ Exact ☐ Partial
Find node(s)



5. WALIDACJA JAKOŚCI (QUAST & CheckM)

QUAST

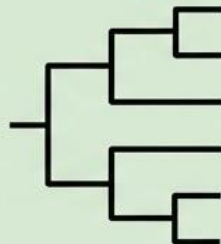
(vs. Referencja)



Misassemblies: 0
Genome fraction: >99.9%
Indels/Mismatches:
Wysokie (cecha
surowego Nanopore)

CheckM

(Marker genes)



Kompletność: ~81%
Kontaminacja: ~1%



Jakość asemblacji (referencja + markery)

QUAST (Referencja)

 1
contig



Genome fraction:
~99.998%



Misassemblies:
0

Błędy bazowe typowe dla ONT:
mismatch ~217/100kbp, indel ~365/100kbp

QUAST Report

The economic lot-275106-m is relevant to the report output.

Show 10 entries

Table	Contig	Edges	Genomes	Fraction	Misassemblies	Measles
contig	1	402 538	~99.998%	1999.90%	0	4 032 129
Misassemblies 1	0	0	-	-	0	N/A
Misassemblies 2	0	0	-	-	0	N/A
Misassemblies 3	0	0	-	-	0	N/A
Misassemblies 4	0	0	-	-	0	N/A
Misassemblies	0	0	-	-	0	N/A
Misassemblies	0	0	-	-	0	N/A
Misassemblies	0	0	-	-	0	N/A
Total assembl	0	0	-	-	0	N/A

CheckM (Markery)



Completeness:
81.18%



Contamination:
0.99%



Problem RAM: pplacer OOM → użyto --reduced_tree

\$ CheckM mode for ONT_RNA/External/Result_assembly_sftp

Marker lineage ID: 1

Marker status: GRF/51A/2k1A0

Marker-SIast completeness: 81.18%

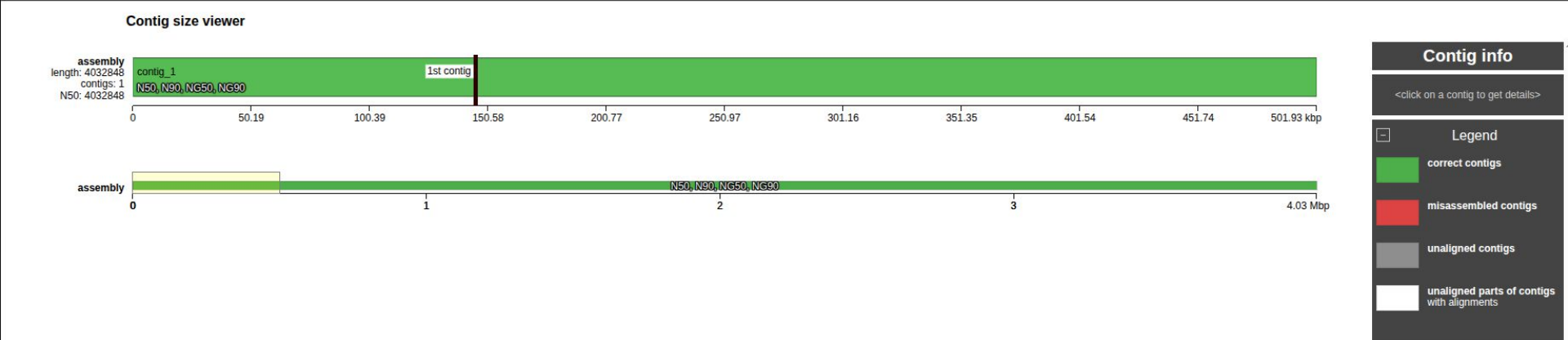
Marker sources: dsaesste_ceria

Marker lineage: deseted_referentz_analysis

Marker lineage espts.klekrr.coult-8

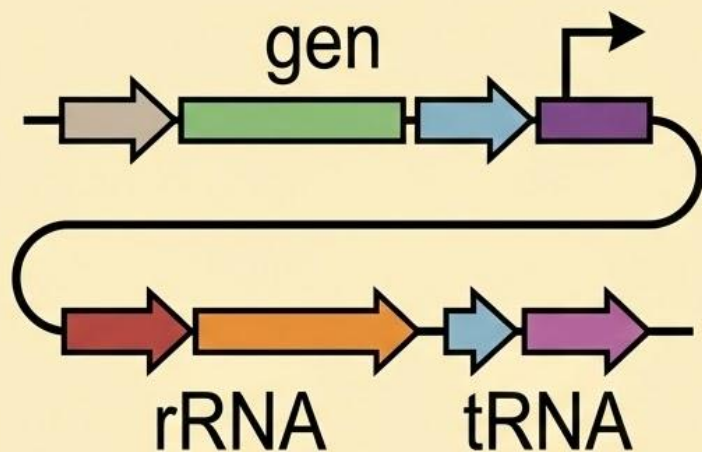
\$ CheckM output table

Bin Id	Marker Lineage	Completeness	Contamination	Presorted	Concgs	Query cove	Total
91	Aglaeirra	81.10%	0.99%	RCF	53	0	35223
92	Beetresateca_pellacet	82.89%	0.99%	RLF	58	0	91258
93	Fewsoyveestas	81.10%	0.99%	RLF	53	0	71826
94	Kareto_copis	81.10%	0.99%	RCF	37	0	106530
95	Bareline_sesiam	81.10%	0.99%	RLF	38	0	185529
96	Gereine_hostajes	89.62%	0.99%	RLF	59	0	22734
97	Bacte_auesose	81.18%	0.99%	RLF	77	0	119687
98	Avpen_enera	87.88%	0.99%	2082	3718	15	1228733
99	Gacop_ardibiss	81.10%	0.99%	2082	2224	15	526226
0110	Atocapi_osta	81.10%	0.99%	RLF	58	1	115185
0111	Accosentebon	81.12%	0.99%	RLF	25	1	110899
0112	Accapacator_justan	81.89%	0.99%	RLF	37	3	168797
0113	Rditecs_avar	73.69%	0.99%	RLF	7	1	5526
0114	Rditecs_apstlast	77.93%	0.99%	905	13	0	52120



6. ADNOTACJA (PROKKA)

Prokka Pipeline



Znalezienie cechy:
CDS (białko): 8189
rRNA: 29 (pełny zestaw)
tRNA: 82

Prokka: predykcja genów i elementów RNA

Podsumowanie (b_subtilis.txt)

```
b_subtilis.txt — 60x24
organism: Bacillus subtilis strain
contigs: 1
bases: 4032848
CDS: 8189
rRNA: 29
repeat_region: 1
tRNA: 82
tRNA: 1
```

Statystyki i Pliki Wyjściowe



Contigs: 1, Bases: 4 032 848



CDS: 8189



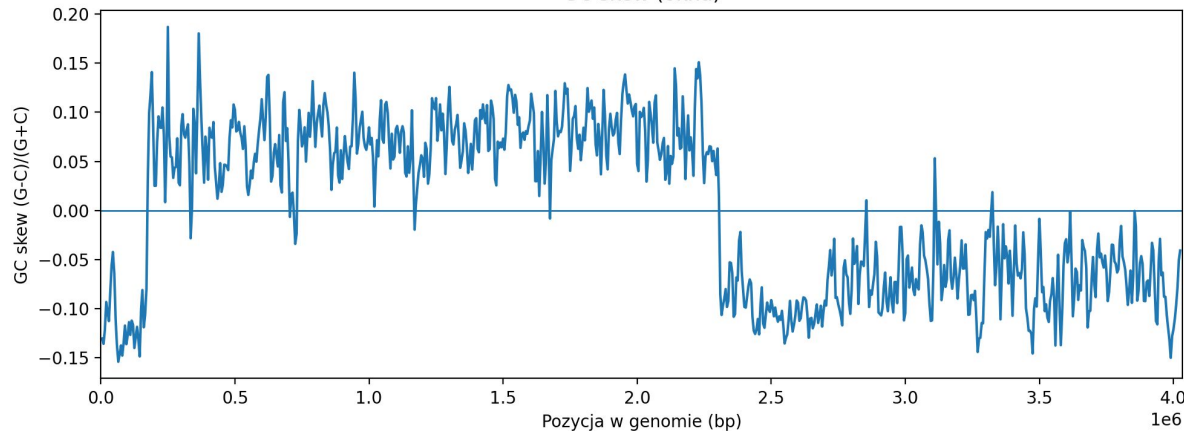
rRNA: 29, tRNA: 82, tmRNA: 1

Output:



Gotowe do dalszych analiz

GC skew (okna)



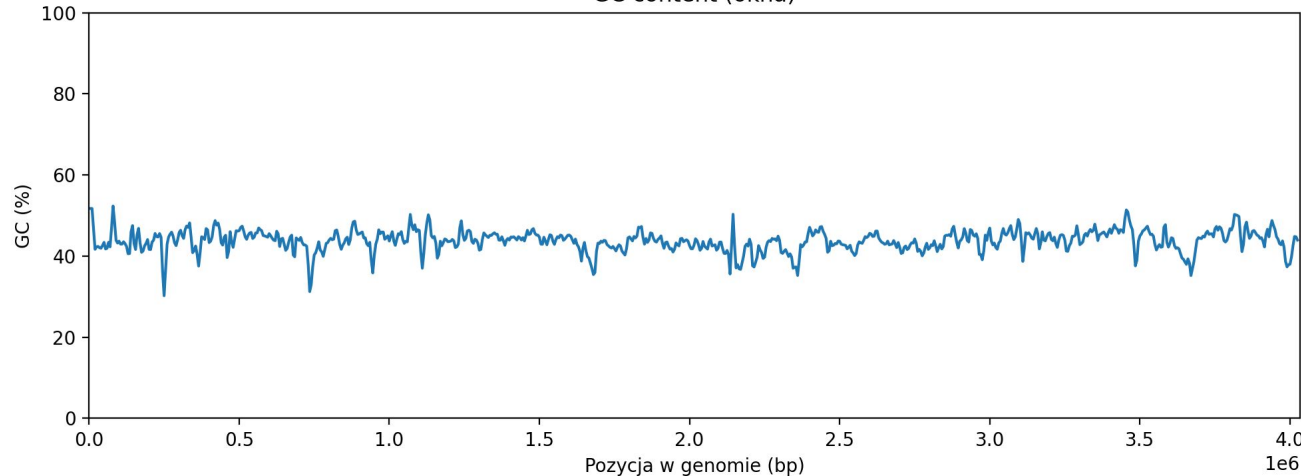
W bakteryjnym, kolistym chromosomie GC skew często rozdziela się na dwie połówki (replichory):

- jedna strona ma przewagę G (skew > 0),
- druga przewagę C (skew < 0).

Punkty przejścia zwykle odpowiadają okolicom:

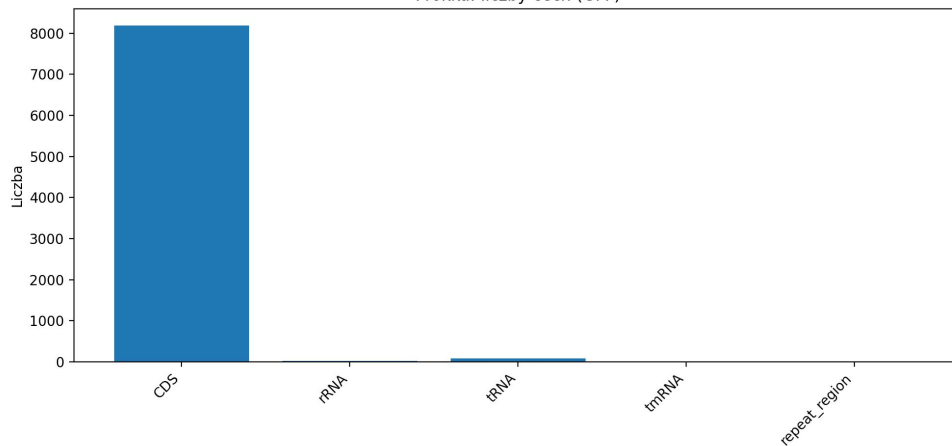
- **oriC** (origin of replication),
- **ter** (terminus).

GC content (okna)



- To pasuje do *Bacillus subtilis* (GC ~43–44%).
- Lokalna zmienność jest normalna. Najczęściej oznacza:
 - klastry rRNA/tRNA,
 - wyspy genomowe, prophage,
 - fragmenty po HGT (transfer poziomy),
 - regiony powtórzeń.

Prokka: liczby cech (GFF)



To co wygląda dobrze:

- **tRNA ~ 80–90**: to pasuje do *B. subtilis* (około 86).
- **tmRNA = 1**: to jest normalne.
- **rRNA kilkadziesiąt**: też pasuje (operony rRNA dają łącznie wiele genów 16S/23S/5S).

To co wygląda podejrzanie:

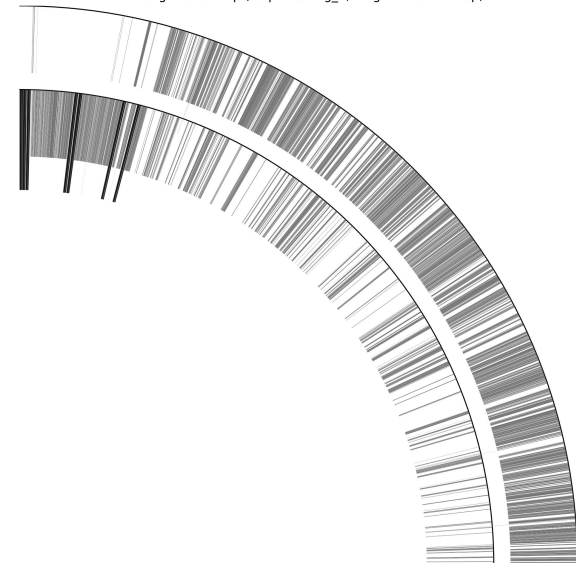
- **CDS ~ 8200** jest zdecydowanie za dużo dla *B. subtilis*.

Najprawdopodobniej:

Jest dobra asemlacja w sensie jeden contig ale **nadal sporo indeli** (QUAST pokazał indels/100 kbp wysokie).

Indele robią **frameshift**, a to rozrzuca geny na kawałki. Prodigal (w Prokka) wtedy przewiduje **więcej krótkich CDS** zamiast jednej pełnej sekwencji genu.

Prokka genome map (seqid=contig_1, length=4032848 bp)

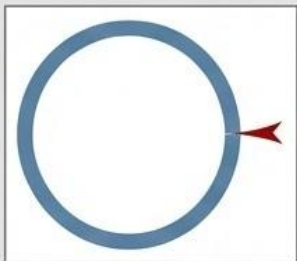


- dużo gęstych kresek na obu pierścieniach, czyli geny na + i - nici
- brak wielkich “pustych” odcinków bez adnotacji

To wygląda jak typowy bakteryjny genom: geny gęsto upakowane. Brak wielkich dziur sugeruje, że nie masz ogromnych niezmapowanych fragmentów albo śmieciowej sekwencji.

EKSPERYMENT PORÓWNAWCZY: WPŁYW POKRYCIA NA GRAF OLC

PEŁNE DANE (~44x)



Wynik Flye: 1 Kontig

Wniosek: Wysokie pokrycie rozwiązuje powtórzenia

DOWNSAMPLING (~20x)



Wynik Flye: 2 Kontigi (3 disjointigs)
Efektywne pokrycie overlaps: ~15x

Wniosek: Niskie pokrycie zrywa graf w trudnych miejscach

Mamy:

- genom: **4,045,677 bp**
- dane: **178,233,472 bp**
- obecnie ~44x, mosdepth ~41x

Proporcja dla 20x:

$$p = \frac{(20 * \text{genome length})}{\text{total bases}} \approx \frac{20 * 4045677}{178233472} \approx 0.454$$

cd

/home/konrad_guest/ExternalDisk/functional_genomics/gorzelanczyk_assemblacja

```
genome_len=$(seqkit fx2tab -l b_subtilis_reference.fasta |  
awk '{s+=$3} END{print s}')
```

```
total_bases=$(seqkit stats -T b_subtilis_reads.fastq.gz |  
awk 'NR==2{print $5}')
```

```
p=$(awk -v g="$genome_len" -v b="$total_bases" 'BEGIN{print  
(20*g)/b}')
```

```
echo "Genome length = $genome_len"  
echo "Total bases = $total_bases"  
echo "Sampling p(20x)= $p"
```

Zrobienie FASTQ o pokryciu ~20x:

Zrobienie FASTQ o pokryciu ~20x:

```
```bash  
(bio) konrad_guest@konrad-
hp:~/ExternalDisk/functional_genomics/gorzelanczyk_assemblacja
$ seqkit sample -p "$p" -s 13 b_subtilis_reads.fastq.gz -o
b_subtilis_reads_20x.fastq.gz
```

```
[INFO] sample by proportion
[INFO] 11976 sequences outputted
(bio) konrad_guest@konrad-
hp:~/ExternalDisk/functional_genomics/gorzelanczyk_assemblacja
$
```
```

Wychodzi prawie idealnie:

```
(bio) konrad_guest@konrad-  
hp:~/ExternalDisk/functional_genomics/gorzelanczyk_assemblacja  
$ echo "Bases in 20x FASTQ = $bases20"
```

Bases in 20x FASTQ = 81749935

```
(bio) konrad_guest@konrad-  
hp:~/ExternalDisk/functional_genomics/gorzelanczyk_assemblacja  
$ echo "Approx coverage = $cov20 x"
```

Approx coverage = 20.2067 x

Downsampling 20x: kiedy “zacząyna się psuć ciągłość”

Cel: sprawdzić stabilność assembly przy mniejszym pokryciu

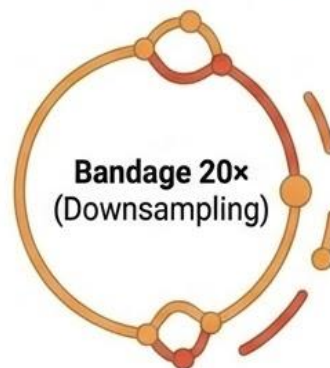
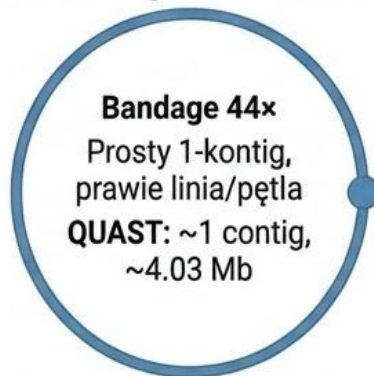
20x: $p=0.454$

$p \approx 0.454$, dane ~81.7 Mb,
~20.2x

Flye 20x: pojawiają się
dodatkowe contigi, graf
robi się bardziej złożony

QUAST 20x: 100%
genome fraction,
0 misassemblies, ale
więcej mismatch/indel
(mniej konsensusu)

Grafika: Bandage 20x vs 44x + QUAST grep (contigs, mismatches, indels)



Dodatkowe contigi,
graf bardziej złożony

QUAST:
~2 contigs (≥ 500 bp),
~4.04 Mb Total,
1 Unaligned

| Metryka | 44x (Pełny) | 20x (Downsampled) | Komentarz |
|----------------------------|---------------|-------------------|----------------------------|
| Genome Fraction | 100% | 100% | Zachowana kompletność |
| Misassemblies | 0 | 0 | Brak błędów strukturalnych |
| # Contigs (≥ 500 bp) | 1 | 2 | Spadek ciągłości |
| Mismatches / 100kbp | Niższa (~216) | Wyższa (~465) | Mniej precyzyjny konsensus |
| Indels / 100kbp | Niższa (~364) | Wyższa (~532) | Mniej precyzyjny konsensus |

Wniosek: Przy 20x genom jest kompletny, ale traci ciągłość i rośnie liczba błędów na poziomie pojedynczych zasad z powodu mniejszej redundancji danych.

File Edit Tools View Select Output Help

Graph information

Nodes: 3
Edges: 1
Total length: 4 083 750

[More info](#)

Graph drawing

Scope: Entire graph

Style: ☒ Single ☐ Double

[Draw graph](#)

Graph display

Zoom: 369,2%

Node width: 2,5

Colour by depth

Node labels

☒ Custom ☒ Name
☒ Length ☒ Depth
☒ BLAST hits ☒ CSV data:

Font ☐ Text outline

BLAST

[Create/view BLAST search](#)

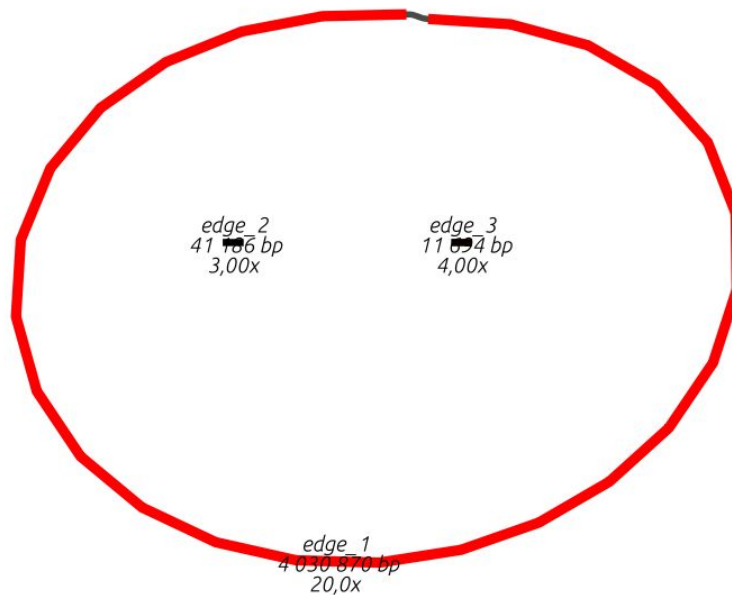
Query: none

Find nodes

Node(s):

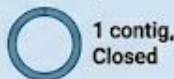
Match: ☒ Exact ☐ Partial

[Find node\(s\)](#)

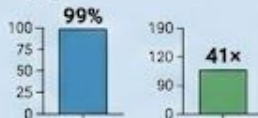




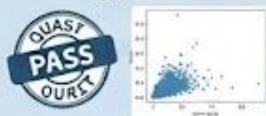
Dane ONT ~44x pozwoliły na praktycznie **zamknięty genom** w 1 **contigu** (~4.03 Mb)



Kontrola jakości: >99% mapped, pokrycie ~41x



QUAST: brak błędów strukturalnych, błędy bazowe typowe dla ONT



Powtórzenia rRNA (16S x10) to naturalne źródło trudności w grafach



Downsampling 20x pokazuje, że pokrycie krytycznie wpływa na rozwiązywanie powtórzeń



Prokka daje kompletny zestaw plików do dalszej analizy funkcjonalnej

