GoSeqIt

Ex. 4 Recap

Q1. How many contigs of size >= 500 bp do the two assemblies each have?

Answer: contigs_trimmed = 29 contigs, contigs_untrimmed = 272 contigs.

Q2. What is the N50 value for each of the two assemblies?

Answer: contigs_trimmed = 189.171, contigs_untrimmed = 19.579

Q3. The N50 value is defined as the length of the shortest contig in the set of longest contigs that together make up at least half the assembly size (half the “Total length”). How is N75 defined?

Answer: The length of the shortest contig in the set of longest contigs that together make up at least 75% of the assembly size. Note that you can find any Nx value by clicking the link “Nx” right above the plot in the bottom of the report.

Q4. Which assembly would you use in your further analysis?

Answer: Contigs_trimmed.fasta

Ex. 4-extra

PURPOSE

To assemble reads generated by PacBio (or Oxford Nanopore) on their own using minimap2/ miniasm for assembly and racon for polishing. Next we will use SPAdes to create a hybrid assembly based on both the PacBio data and trimmed Illumina reads.

DATA

PacBio reads from the isolate *Staphylococcus aureus* ATCC 25923:
SRR2104768.fastq.gz (> 2 GB unzipped, ~900 MB when gzipped)

The trimmed Illuimna reads that we have previous worked with. They are also from *S. aureus* ATCC 25923:
SRR4114395_1_5end_qual_trimmed.fastq.gz
SRR4114395_2_5end_qual_trimmed.fastq.gz

Finally, the reference genome for the same strain:
NZ_CP009361.fasta

COMMENTS

- Many of the tools must be compiled before they can run, which makes the installation a bit tricky…
- **Minimap2** finds overlaps between the reads
- **Miniasm** creates the raw assembly (SRR2104768_minimap2_miniasm.fasta)
- **Racon** polishes the assembly using the PacBio reads (SRR2104768_racon_corrected.fasta)
- **SPAdes** is used for generating hybrid assembly using both PacBio and Illumina reads (SRR2104768_SRR4114395_hybrid_assembly.fasta)

QUAST

Quality Assessment Tool for Genome Assemblies by CAB

16 August 2018, Thursday, 09:36:10

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).

Suggestion: assembly contigs_untrimmed contains continuous fragments of N's longer than or equal to 10 bp. You may consider rerunning QUAST using --scaffolds (-s) option!

Aligned to "NZ_CP009361" | 2 778 854 bp | 1 fragment | 32.88% G+C

Worst Median Best — Show heatmap

Genome statistics	SRR2104768_minimap2_miniasm	SRR2104768_racon_corrected	hybrid_assembly	contigs_untrimmed	contigs_trimmed
Genome fraction (%)	0.013	97.836	99.555	98.569	98.617
Duplication ratio	0.989	1.01	1	1.011	1
Largest alignment	360	562 778	1 727 799	67 407	508 890
Total aligned length	360	2 746 006	2 766 945	2 768 258	2 741 485
NGA50	-	433 044	1 727 799	19 579	189 171
LGA50	-	3	1	44	4
Misassemblies					
# misassemblies	0	10	0	1	0
Misassembled contigs length	0	2 766 259	0	681	0
Mismatches					
# mismatches per 100 kbp	274.73	23.5	0.61	1.13	0.18
# indels per 100 kbp	3296.7	280.09	0.43	0.07	0.07
# N's per 100 kbp	0	0	0	3.57	0
Statistics without reference					
# contigs	7	7	6	272	29
Largest contig	1 509 913	1 446 627	1 727 799	67 407	508 890
Total length	2 976 675	2 845 188	2 792 688	2 798 824	2 767 228
Total length (>= 1000 bp)	2 976 675	2 845 188	2 792 688	2 771 320	2 765 885
Total length (>= 10000 bp)	2 976 675	2 845 188	2 792 688	2 237 142	2 747 270
Total length (>= 50000 bp)	2 948 288	2 817 865	2 766 691	237 532	2 558 562
N50	1 509 913	1 446 627	1 727 799	19 579	189 171
N75	844 729	802 001	434 124	11 242	98 015
L50	1	1	1	45	4
L75	2	2	2	93	10
GC (%)	33.85	32.73	32.81	32.76	32.73

What does it mean for the downstream analysis?

	Reference (NZ_CP009361)	**Illumina, untrimmed reads** (uncontigs_trimmed)	**Illumina, trimmed reads** (contigs_trimmed)	**PacBio w. minimap2/ miniasm** (SRR2104768_minimap2_miniasm)	**PacBio w. minimap2/ miniasm/ racon** (SRR2104768_racon_corrected)	**Hybrid, PacBio + Illumina** (SRR2104768_SRR4114395_hybrid_assembly)
N50	2.778.854	19.070	189.171	1.509.913	1.446.627	1.727.799
MLST	ST243 (all alleles are perfect matches)	ST243 (all alleles are perfect matches)	ST243 (all alleles are perfect matches)	Unknown ST (all alleles present but w. app. 90% ID)	Unknown ST (all alleles present but w. app. 99% ID)	ST243 (all alleles are perfect matches)

MLST, reference, contigs_trimmed, contigs_untrimmed, hybrid_assembly

Multilocus Sequence Type (MLST) result

MLST Profile: saureus

Sequence Type: ST243

Gene	% Identity	Alignment Length	DB allele Length	Gaps	Best match
arcc	100	456	456	0	*arcc_2*
aroe	100	456	456	0	*aroe_2*
glpf	100	465	465	0	*glpf_5*
gmk	100	417	417	0	*gmk_2*
pta	100	474	474	0	*pta_6*
tpi	100	402	402	0	*tpi_3*
yqil	100	516	516	0	*yqil_2*

MLST, PacBio minimap2/miniasm

Results for MLST

Subdatabase: saureus

Sequence Type: Unknown ST

Gene	% Identity	Alignment Length	DB Allele Length	Gaps	Best match
arcc	89.51	486	457	38	arcc_162
aroe	88.06	444	456	38	aroe_52
glpf	91.84	478	465	27	glpf_648
gmk	88.73	417	417	32	gmk_121
pta	86.47	436	474	42	pta_61
tpi	87.88	429	402	36	tpi_121
yqil	85.06	502	516	64	yqil_639

MLST, PacBio minimap2/miniasm/racon

Results for MLST

Subdatabase: saureus

Sequence Type: Unknown ST

Gene	% Identity	Alignment Length	DB Allele Length	Gaps	Best match
arcc	99.34	455	455	1	arcc_551
aroe	99.78	455	456	1	aroe_2
glpf	99.57	467	465	2	glpf_5
gmk	100	413	417	0	gmk_2
pta	98.14	483	474	9	pta_6
tpi	99.75	403	402	1	tpi_3
yqil	99.23	520	516	4	yqil_2