ДЗ11_bioinfa

Сорокина Тамара, Б06-204

Отчёт по улучшению сборки генома вируса гриппа

На семинаре мы работали с данными секвенирования вируса гриппа, которые лежат по пути:

/projects/mipt_dbmp_biotechnology/genome_de_novo/

Для начала провели сборку с разными наборами параметров, в основном через SPAdes, а также попробовали Platanus. Качество сборки сравнивали с помощью Quast.

Что получилось в начале:

Statistics without reference	platanus_k39_contig	platanus_k49_contig	platanus_k63_contig	scaffolds
# contigs	42	42	40	49
# contigs (>= 0 bp)	42	42	40	49
# contigs (>= 1000 bp)	0	1	1	1
# contigs (>= 5000 bp)	0	0	0	0
# contigs (>= 10000 bp)	0	0	0	0
# contigs (>= 25000 bp)	0	0	0	0
# contigs (>= 50000 bp)	0	0	0	0
Largest contig	895	1032	1032	1069
Total length	8094	8078	7606	15 064
Total length (>= 0 bp)	8094	8078	7606	15 064
Total length (>= 1000 bp)	0	1032	1032	1069
Total length (>= 5000 bp)	0	0	0	0
Total length (>= 10000 bp)	0	0	0	0
Total length (>= 25000 bp)	0	0	0	0
Total length (>= 50000 bp)	0	0	0	0
N50	327	298	235	440
N90	71	71	71	148
auN	371.1	409.7	429.8	484.1
L50	9	8	7	12
L90	25	27	29	37
GC (%)	44.22	43.98	44.33	45.47
Mismatches				
# N's per 100 kbp	0	0	0	0
# N's	0	0	0	0

SPAdes показал себя лучше, чем Platanus — сборки получились более полные и с длинными контигами. Поэтому дальше решено было улучшать именно сборку через SPAdes.

Что дальше:

Чтобы получить ещё более качественную сборку, добавим больше значений k-меров (21,33,55,77,99,111). Это нужно для того, чтобы получить длинные контиги, которые лучше "перекрывают" повторы и реже дают разрывы. Минус в том, что длинные k-меры могут «пропустить» участки с редкими вариантами, но при хорошем покрытии это не критично.

Запустим сборку так:

```
python3 /projects/mipt_dbmp_biotechnology/soft/SPAdes-4.1.0-Linux/bin/spades.py \
--careful -k 21,33,55,77,99,111 \
-1 /projects/mipt_dbmp_biotechnology/genome_de_novo/7_S4_L001_R1_001.fastq \
-2 /projects/mipt_dbmp_biotechnology/genome_de_novo/7_S4_L001_R2_001.fastq \
-o ~/homeworks/hw_11/genome_assembly_results/spades
```

Прогоним всё через Quast:

```
~/soft/bin/quast.py -o ~/homeworks/hw_11/assembly_analysis -m 0 --threads 1 \
~/classes/class_11/genome_assembly_results/k39/platanus_k39_contig.fa \
~/classes/class_11/genome_assembly_results/k49/platanus_k49_contig.fa \
~/classes/class_11/genome_assembly_results/k63/platanus_k63_contig.fa \
~/classes/class_11/genome_assembly_results/spades/scaffolds.fasta \
~/homeworks/hw_11/genome_assembly_results/spades/hw_scaffolds.fasta
```

Что изменилось:

Statistics without reference	platanus_k39_contig	platanus_k49_contig	platanus_k63_contig	scaffolds	hw_scaffolds
# contigs	42	42	40	49	18
# contigs (>= 0 bp)	42	42	40	49	18
# contigs (>= 1000 bp)	0	1	1	1	1
# contigs (>= 5000 bp)	0	0	0	0	0
# contigs (>= 10000 bp)	0	0	0	0	0
# contigs (>= 25000 bp)	0	0	0	0	0
# contigs (>= 50000 bp)	0	0	0	0	0
Largest contig	895	1032	1032	1069	1089
Total length	8094	8078	7606	15 064	9207
Total length (>= 0 bp)	8094	8078	7606	15 064	9207
Total length (>= 1000 bp)	0	1032	1032	1069	1089
Total length (>= 5000 bp)	0	0	0	0	0
Total length (>= 10000 bp)	0	0	0	0	0
Total length (>= 25000 bp)	0	0	0	0	0
Total length (>= 50000 bp)	0	0	0	0	0
N50	327	298	235	440	611
N90	71	71	71	148	312
auN	371.1	409.7	429.8	484.1	653.5
L50	9	8	7	12	6
L90	25	27	29	37	14
GC (%)	44.22	43.98	44.33	45.47	45.89
Mismatches					
# N's per 100 kbp	0	0	0	0	0
# N's	0	0	0	0	0

- Общая длина сборки стала ближе к реальной длине генома гриппа (~13 Кб). Раньше было немного больше — возможно, были пересобраны повторы. Теперь могли что-то недособрать, но зато меньше лишнего.
- **N50** стал больше — это значит, что среди самых "вкладных" контигов теперь есть более длинные, а значит, сборка стала менее фрагментированной.
- **L50** стал меньше — то есть для того, чтобы покрыть половину сборки, теперь нужно меньше контигов, что тоже хорошо.
- То же самое с **N90** и **L90** — улучшения видны.
- **auN**, которая обобщает все метрики N1–N100, тоже выросла — значит, в целом сборка получилась лучше.

Вывод:

Добавление длинных k-меров и параметра --careful дало лучший результат — метрики выросли, сборка стала аккуратнее и ближе к настоящему геному. Такую стратегию можно использовать как хороший подход для сборки небольших вирусных геномов.