Отсчет. Сборка генома E.Coli.

Автор: Гавриленко Александр.

Сперва сборка генома прочтениями от PacBio проводилась с помощью Flye для каждого покрытия (N), соотвественно.

```
flye --nano-raw ./pacbio_Nx.fq.gz
--genome-size 4.64m --out-dir ./flye_output_Nx
```

Логи сборки для N = 10:

> Total length: 4 333 634
> Fragments: 93
> Fragments N50: 61 597
> Largest frg: 155 648
> Scaffolds: 0
> Mean coverage: 9

Логи сборки для N = 20:

> Total length: 4 681 275
> Fragments: 24
> Fragments N50: 303 960
> Largest frg: 589 942
> Scaffolds: 0
> Mean coverage: 19

Логи сборки для N = 40:

> Total length: 4 642 048
> Fragments: 1
> Fragments N50: 4 642 048
> Largest frg: 4 642 048
> Scaffolds: 0
> Mean coverage: 39

Логи сборки для N = 80:

> Total length: 4 644 117
> Fragments: 3
> Fragments N50: 4 640 521
> Largest frg: 4 640 521

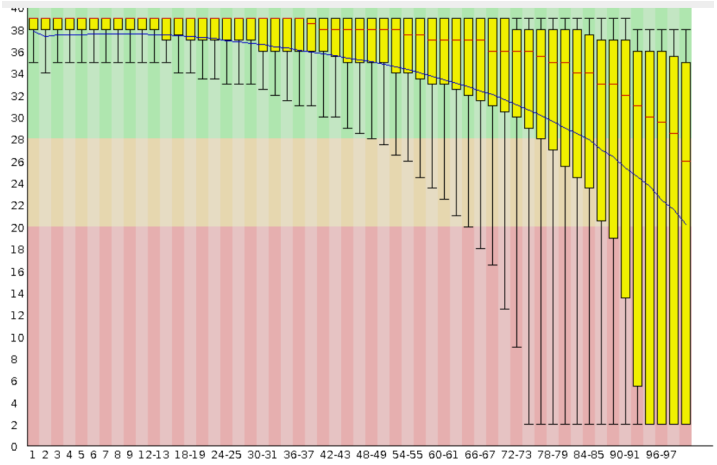
> Scaffolds: 0
> Mean coverage: 78

Увеличение покрытия генома прочтениями значительно улучшают сборку - увеличивается N50, сокращается количество контигов, увеличивается длина наибольшего фрагмента.

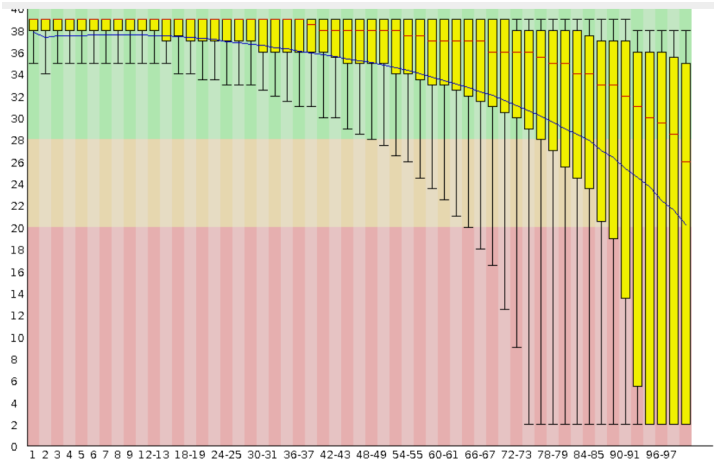
Сборка прочтениями Illumina.

1. Анализ качества парных прочтений fastqc.

illumina.100x.1_fastqc.html



illumina.100x.2_fastqc.html



Видно что к концу прочтений качество падает.

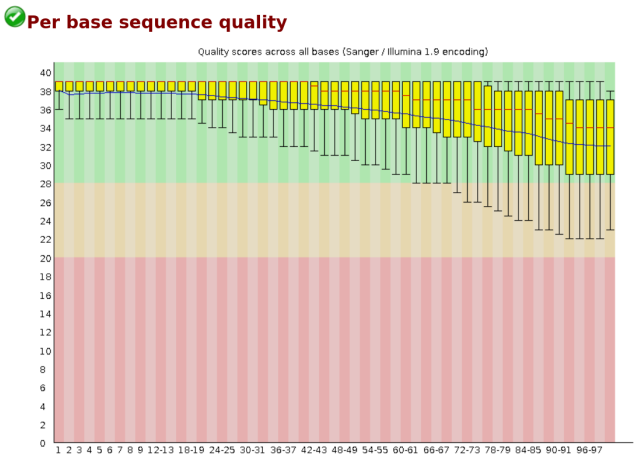
2. Фильтрация ридов с помощью Trimmomatic.

```
java -jar trimmomatic-0.39.jar PE -phred33 ./illumina.100x.1.fq.gz
./illumina.100x.2.fq.gz output_forward_paired.fq.gz
output_forward_unpaired.fq.gz output_reverse_paired.fq.gz
output_reverse_unpaired.fq.gz  LEADING:3 TRAILING:3 MINLEN:36 SLIDINGWINDOW:4:15
TrimmomaticPE: Started with arguments:
-phred33 ./illumina.100x.1.fq.gz ./illumina.100x.2.fq.gz
```

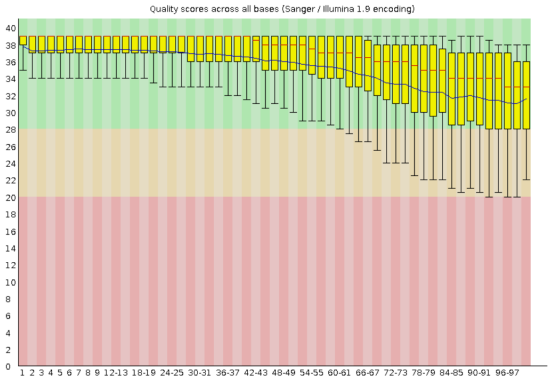
```
output_forward_paired.fq.gz output_forward_unpaired.fq.gz
output_reverse_paired.fq.gz output_reverse_unpaired.fq.gz
LEADING:3 TRAILING:3 MINLEN:36 SLIDINGWINDOW:4:15
Input Read Pairs: 2500000 Both Surviving: 2401660 (96.07%)
Forward Only Surviving: 55222 (2.21%) Reverse Only Surviving: 36855 (1.47%)
Dropped: 6263 (0.25%)
TrimmomaticPE: Completed successfully
```

Качество после фильтрации.

output_forward_paired_fastqc.html



output_reversed_paired_fastqc.html



3. Сборка прочтений с помощью Spades

```
spades.py -1 output_forward_paired.fq.gz -2 output_reverse_paired.fq.gz
-o ./illumina_assembly
```

4. Гибридная сборка каждого вида прочтения (N) Pacbio и Illumina.

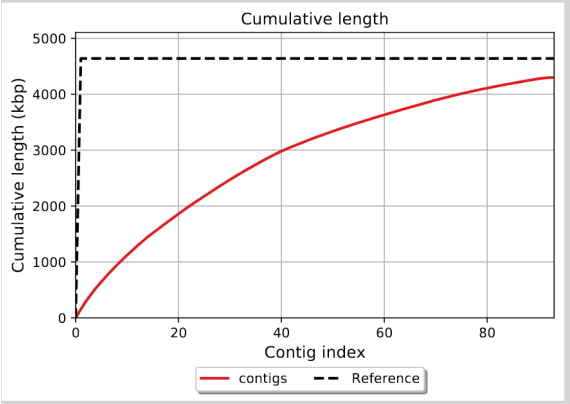
```
spades.py -1 output_forward_paired.fq.gz -2 output_reverse_paired.fq.gz
--pacbio pacbio_Nx.fq.gz -o ./flye_output_Nx
```

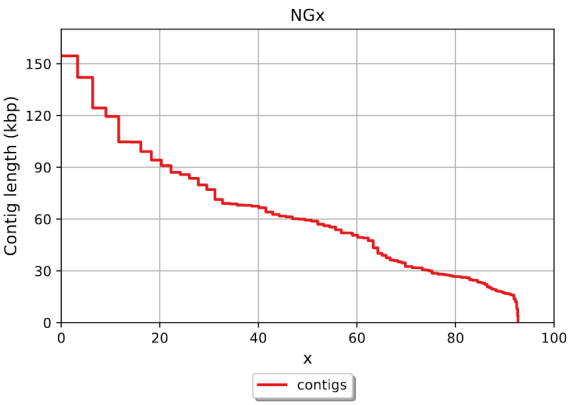
5. Анализ сборок с помощью Quast

```
quast contig -r refence.fasta -o assembly_metrics
```

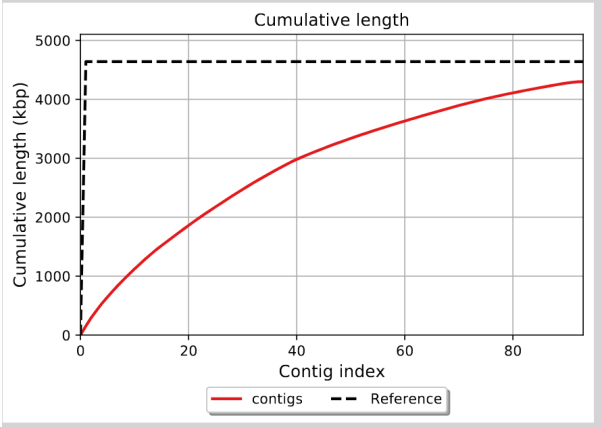
Сравнение базовых статистик

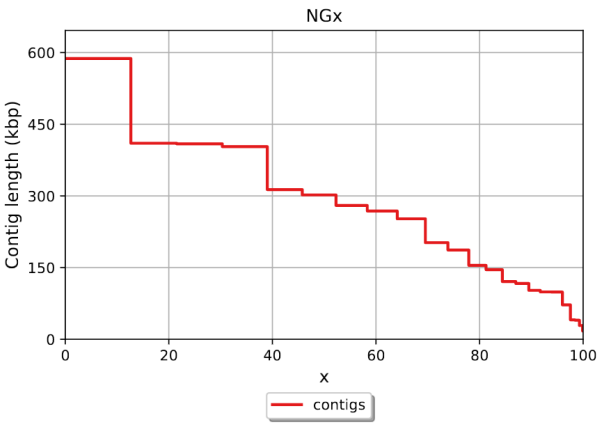
1) 10x Pacbio



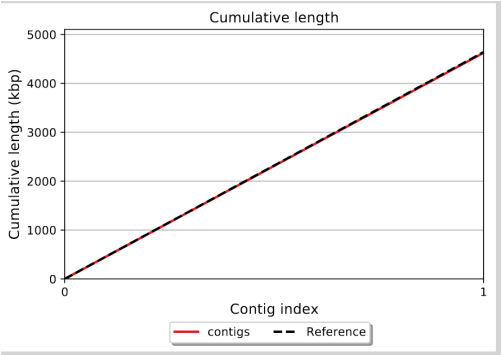


2) 20x Pacbio

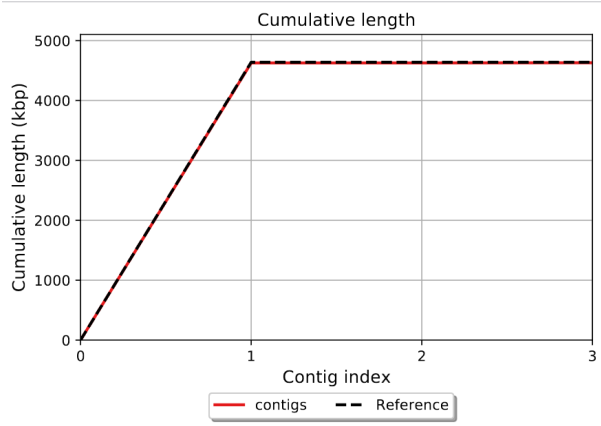


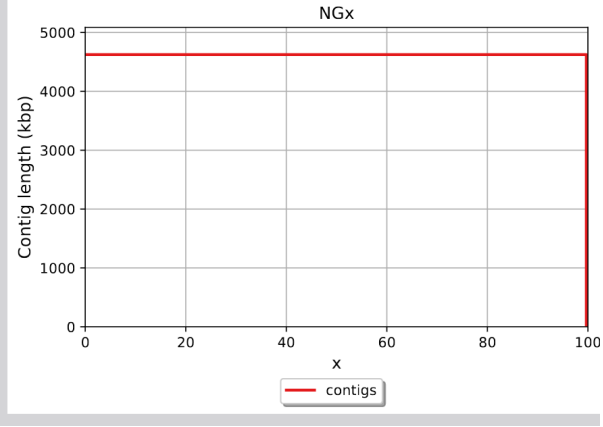


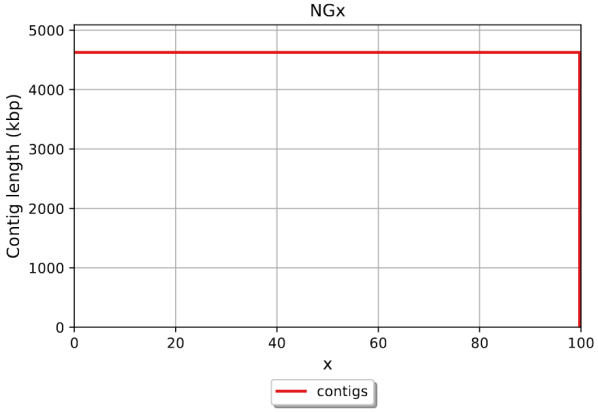
3) 40x Pacbio



4) 80x Pacbio

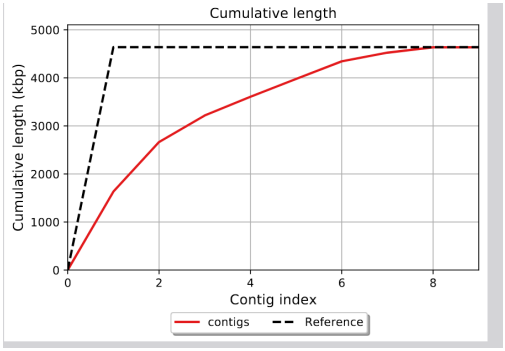


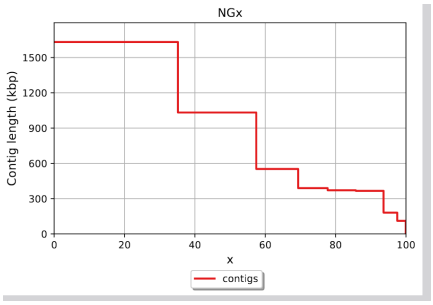




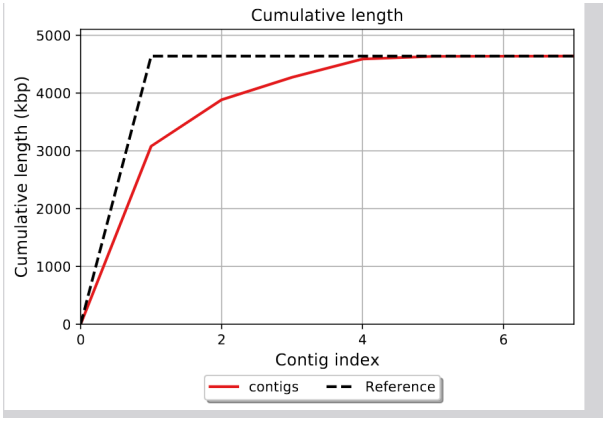
Видно, что с увеличением покрытия растет качество сборки - длина контигов растет. В конечном счетае при покрытии больше чем 40x геном покрывается одним контигом.

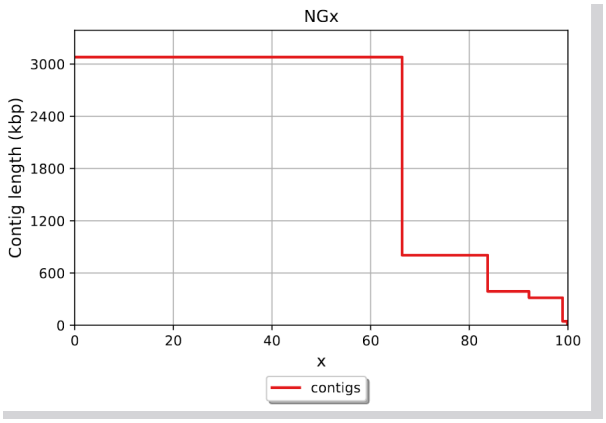
Illumina + 10x Pacbio





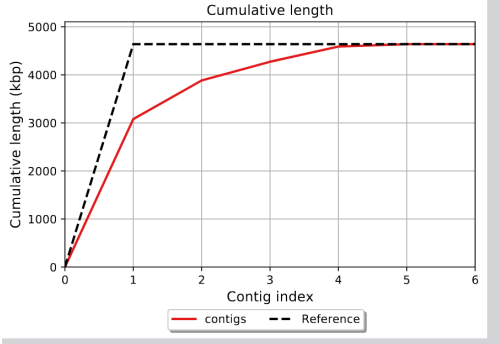
Illumina+ 20x pacbio

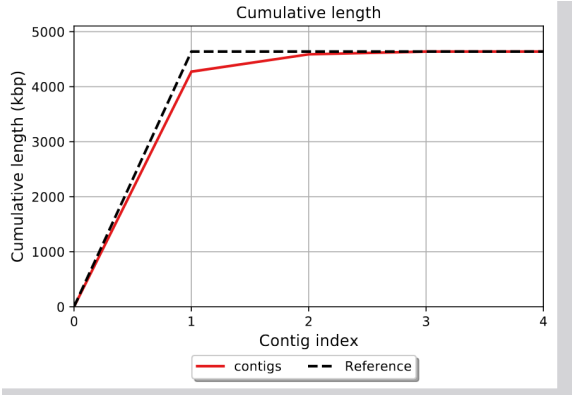


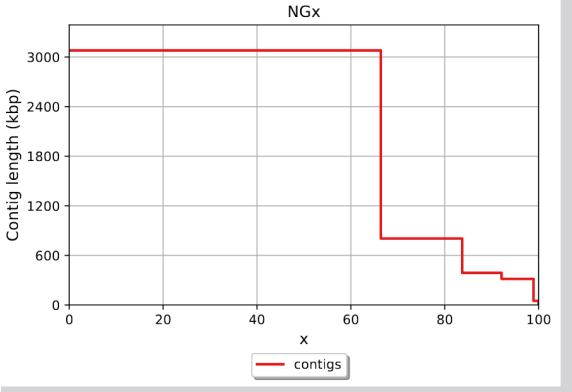


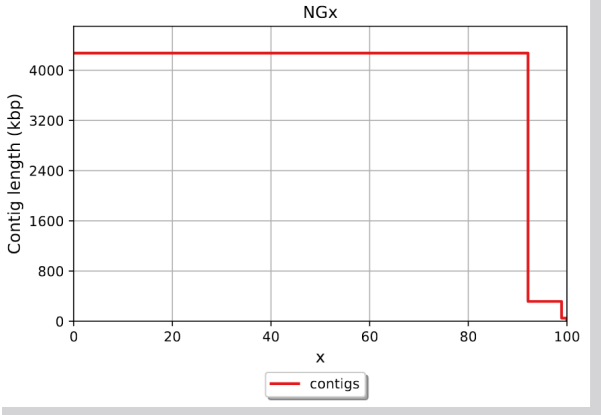
Illumina+ 40x pacbio

Illumina+80x pacbio



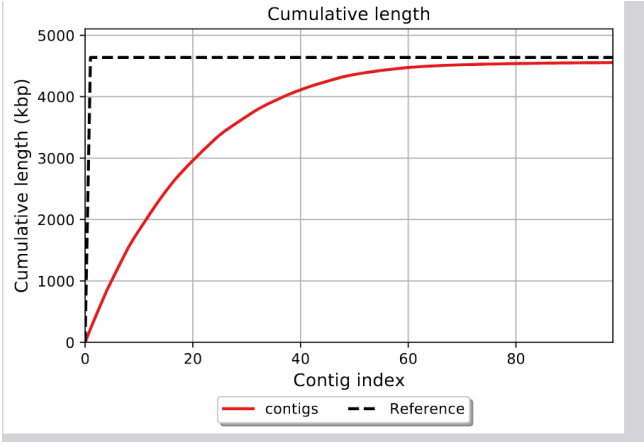


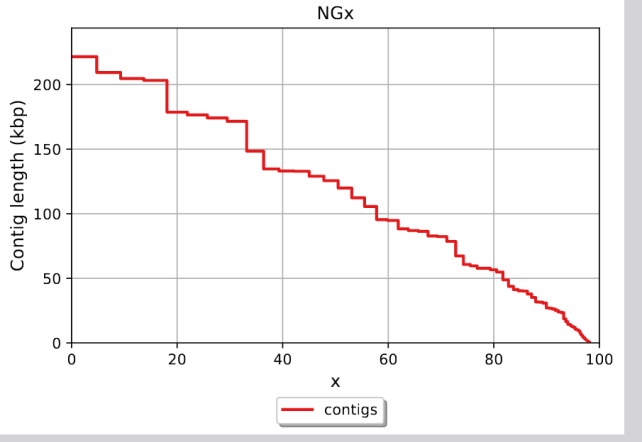




Видно, что с увеличением покрытия pacbio прочтений растет длина контигов, что покрывать все большую часть генома одним контигом.

Illumina only





Данные pacbio с большим количеством прочтений дает лучшее качество NGx. Требуется меньшее количество контигов в сборке для покрытия генома.

Сравнительная таблица reference-based метрик:

	Genome fraction (%)	Duplication ratio	Largest alignment	Total aligned length	NGA50	LGA50
Illumina only	98.138	1	221 546	4 554 328	125 608	14
10x Pacbio	87.794	1.014	153 521	4 129 003	54 753	29
20x Pacbio	99.238	1.004	587 087	4 621 197	300 804	6
40x Pacbio	99.976	0.997	2 159 397	4 622 960	944 840	2
80x Pacbio	99.974	0.998	3 016 666	4 629 254	3 016 666	1
10x Pacbio + Illumina	99.854	1.001	1 032 624	4 635 622	572 345	3
20x Pacbio + Illumina	99.907	1.001	2 301 238	4 635 484	722 395	2
40x Pacbio + Illumina	99.926	1.001	2 301 238	4 636 385	722 395	2
80x Pacbio + Illumina	99.927	1.001	3 023 578	4 636 183	3 023 578	1

Данные Illumina значительно показатели метрик данные Pacbio, это обусловено тем, что технология секвенировния Illumina более точная, короткие прочтения обуславливают большое покрытие, что позволяет алгоритму сборщика точнее детектировать ошибки секвенирования.

Лучшее качество показывает гибиридная сборка - длинные прочтения с большим покрытием Pacbio позволяют разрешать длинные повторы, перед которыми короткие прочтения Illumina уязвимы.