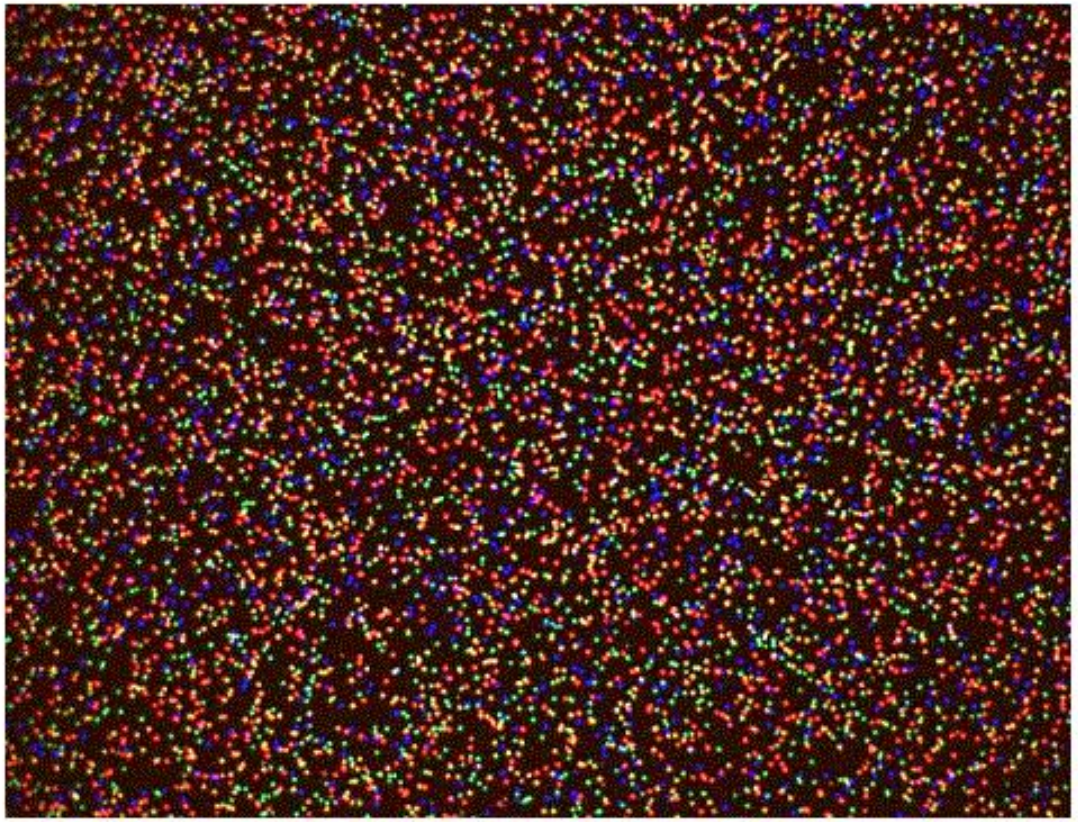
Лекция 7

Анализ данных NGS

Single Tile 4-Color Overlay



Формат данных FastQ

```
@M02435:68:000000000-AUW5P:1:1101:14971:1352 1:N:0:92
GCTGAACCATGGCCAACAGGCTATCACCGGCGAAATGATGATCCGAATCTTCACCCAGGCCGAGGAAATGATC
TCTGCTCCCTCCCACCGGGATTTCTTCATCAAGGATGGGGTAGTTCACTGGAAGGGGGTGAAGGTTGGTCAGA
TGGTAGATGGCCGACTGTCCGCAGATGAACAATTCAAGAACCAGGAGGACTTGCTATGTCAGTTATTGGCACA
CAGATAGGGTTCCGTAAGAACCAGATCATCGC
+
3AABABFBFFFFAFFGGGGGGGFHHHHHGGGGGEEHFFFFFGHHHGGGGHHHHHHHGFGHGC0>@EEHHBGFG
HHHFHHFGFHHEEGGGGGG?AGHHHHGHHHHHHHFHGGHGFGCGFFHHHHGGHHGGGC@@.1FGHHHCAGHHH
HHGGHBGHGGFFC-
;@EGGFFGGB?CBBFBFFCFFF0C9FGGGFGEEDDF/BFFFFFFF9BBFFFBFFFFEFFFBFFFFFEA;.A9.
FFDFF/BEFEFF/BBFF.
@M02435:68:000000000-AUW5P:1:1101:16805:1554 1:N:0:92
ATCAGTTCCTTGGCGAAACCACGGGGGCTGAACCCATTGGATGCGAGGCCGGCGTAATGGTCAGCGGTACCCG
AGGGTCCCATTCCGATGTAGATTTTTGCGGTACAGCAAGAGGGAAAGTCTCTGCGGTGCATTAGTAATCTCCT
TTGATTAATCGTTTGGTTGATAAATCTATTCCGTCAATGCCAGAGCCACAAGCACTGACACAGCCAGGAAGCC
CAGGGGGAGTATGATGGTTAGTAATAGGAAGG
+
>ABBBFFFFFFFGGGCGGGGGGGGGGGGGGHHHHGGHHHHFHHHGGGGGGGGGGGGGGGHHGHHHHCEEGGHHG
GGGGGHHHEHHGHHGGGHGHHHHHHHGHGGGFGFHHHGHHHHGDHEHGHHHHHHGGGGGGHHHHHGHFGGGGG
GGGEFFFGGGGFGGGGCEGGGFGGFFGGGGFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFEFFFFFFEF
FFFFFAC=D;BFFFFFFFFFFF/FFBFFFFD
```

FastQC

FastQC Report

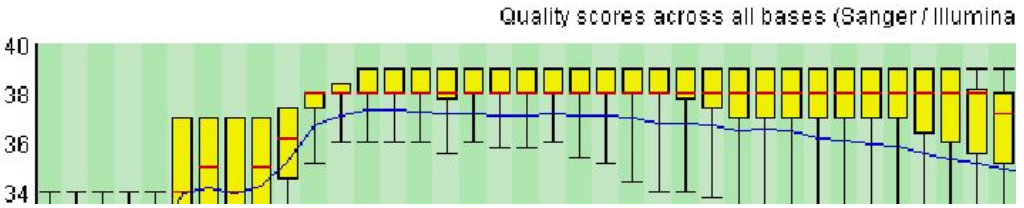
Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Basic Statistics

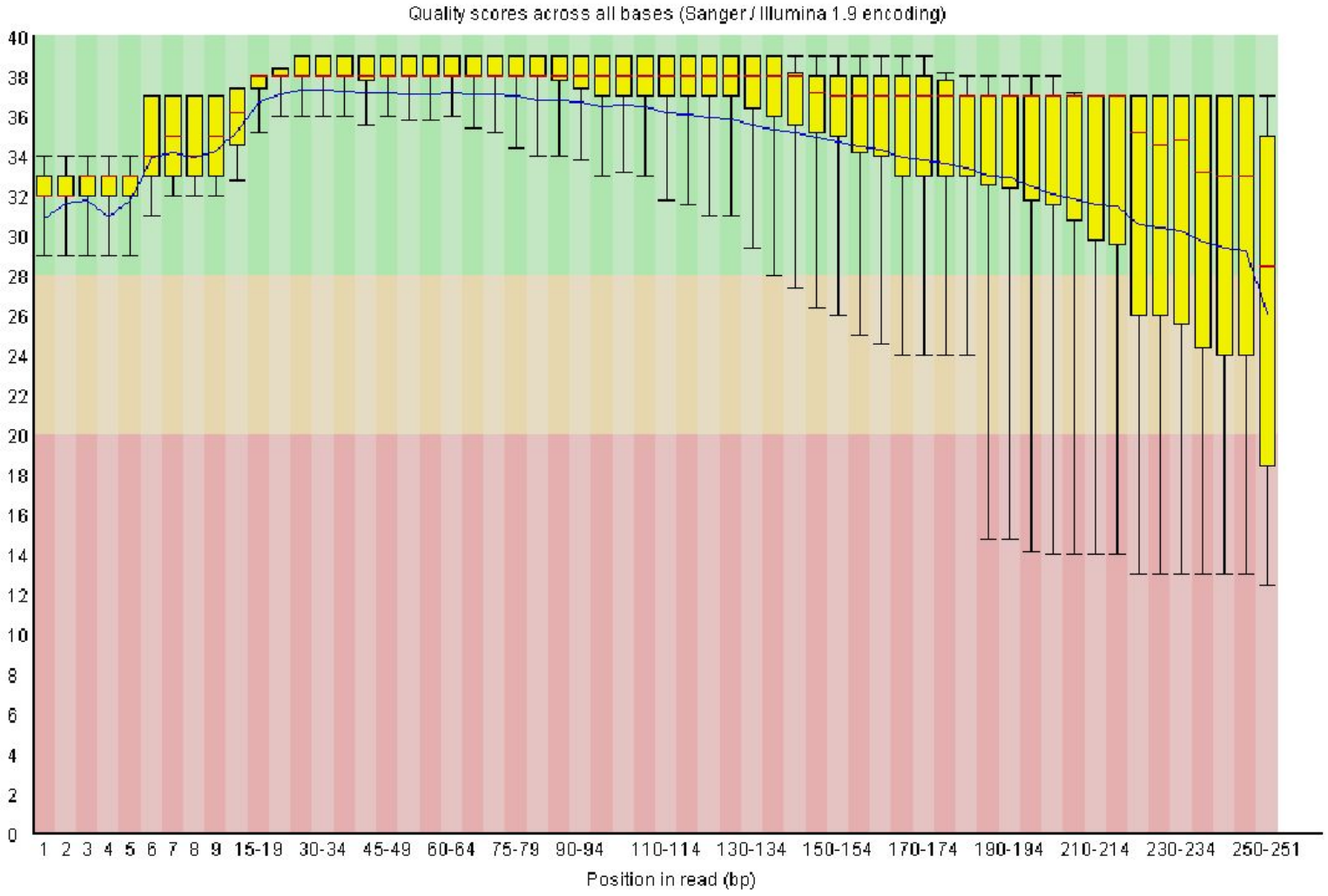
Measure	Value
Filename	92_S92_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	114027
Sequences flagged as poor quality	0
Sequence length	35-251
%GC	60

Per base sequence quality



FastQC

Per base sequence quality



FastQC

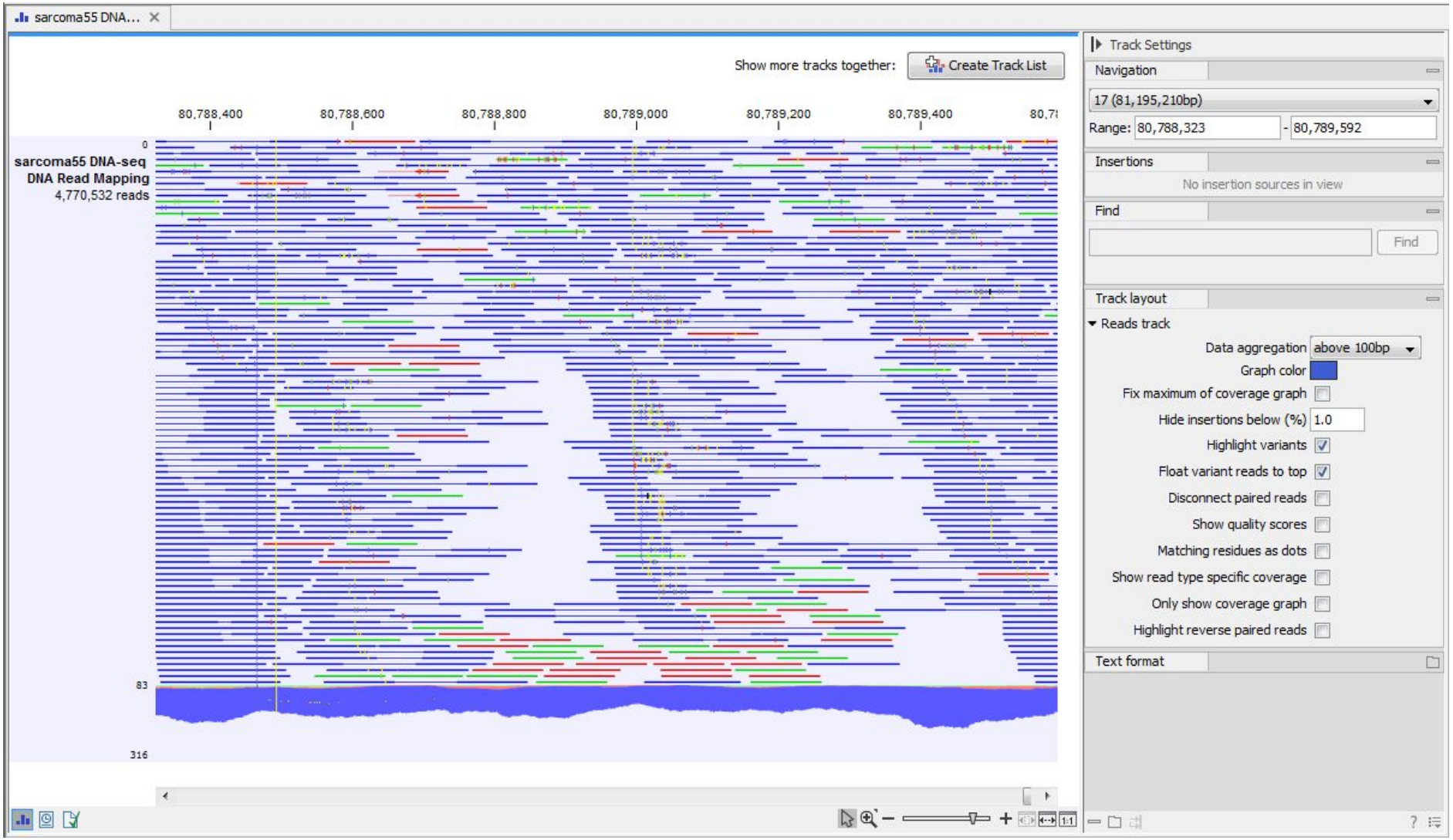
Overrepresented sequences

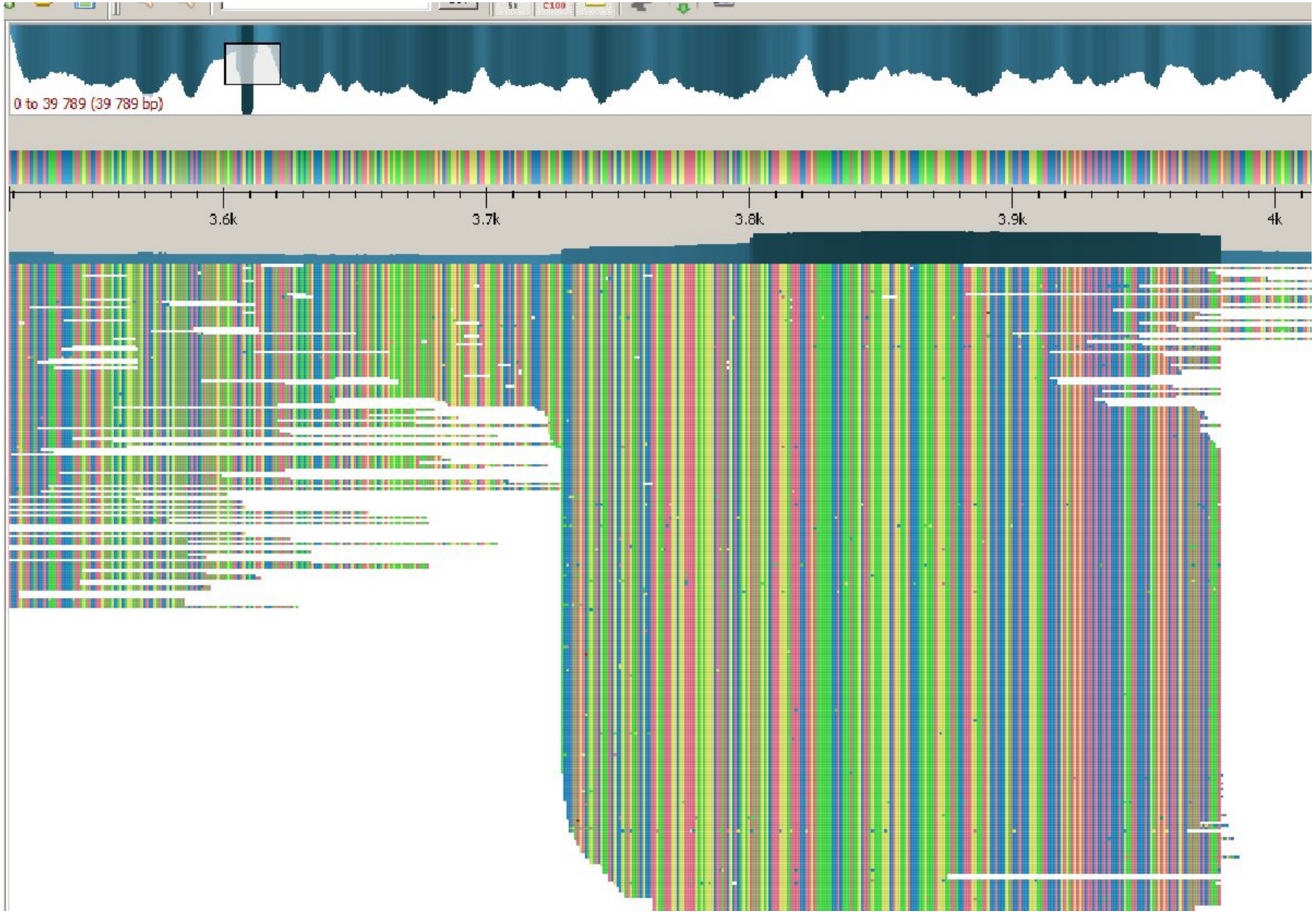
Sequence	Count	Percentage	Possible Source
CCCGACTCTGATCAGCCCACCCTTCCCCCGTAGCCCCTCGTCTGGTGGGC	216	0.18942881949012078	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGCGATAGATCTCGTAT	215	0.18855183421470353	TruSeq Adapter, Index 1 (97% over 36bp)
CTCCGGCCTTAAACCCACATCCAAAAGAGAGAGAATCGCATGAGCTTTCT	122	0.10699220360090154	No Hit

Алгоритмы сборки *de novo*

- OLC – overlap-layout-consensus (Arachne, Celera, CAP3, Phrap)

- Графы де Брёйна (Velvet, SOAPdenovo, SPAdes)


sarcoma55 DNA...
Show more tracks together:
Create Track List
80,788,400
80,788,600
80,788,800
80,789,000
80,789,200
80,789,400
sarcoma55 DNA-seq
DNA Read Mapping
4,770,532 reads
0
83
316
Track Settings
Navigation
17 (81,195,210bp)
Range: 80,788,323 - 80,789,592
Insertions
No insertion sources in view
Find
Find
Track layout
Reads track
Data aggregation above 100bp
Graph color
Fix maximum of coverage graph
Hide insertions below (%) 1.0
Highlight variants
Float variant reads to top
Disconnect paired reads
Show quality scores
Matching residues as dots
Show read type specific coverage
Only show coverage graph
Highlight reverse paired reads
Text format


0 to 39 789 (39 789 bp)
3.6k
3.7k
3.8k
3.9k
4k