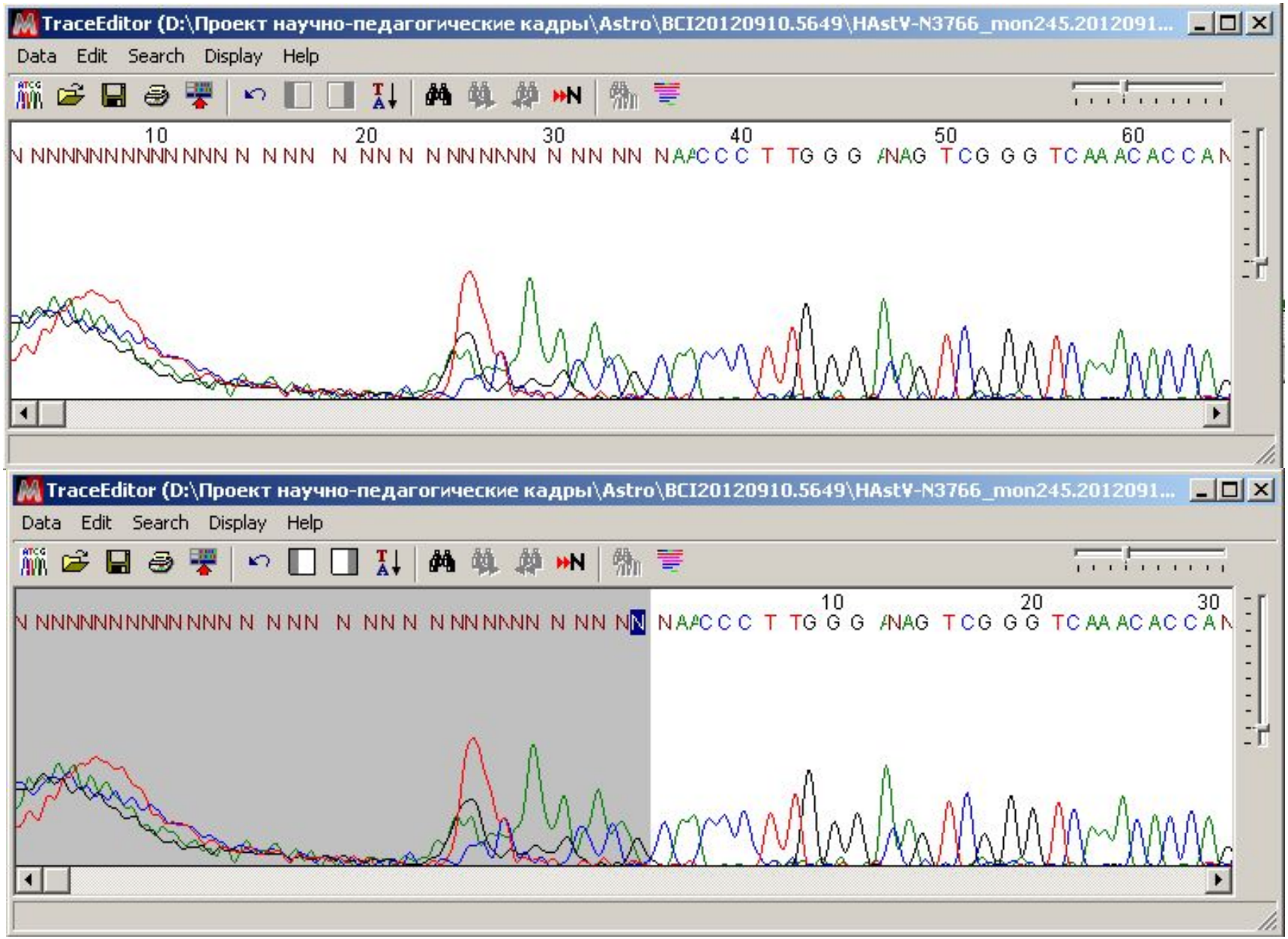
Лекция 5

Анализ данных секвенирования

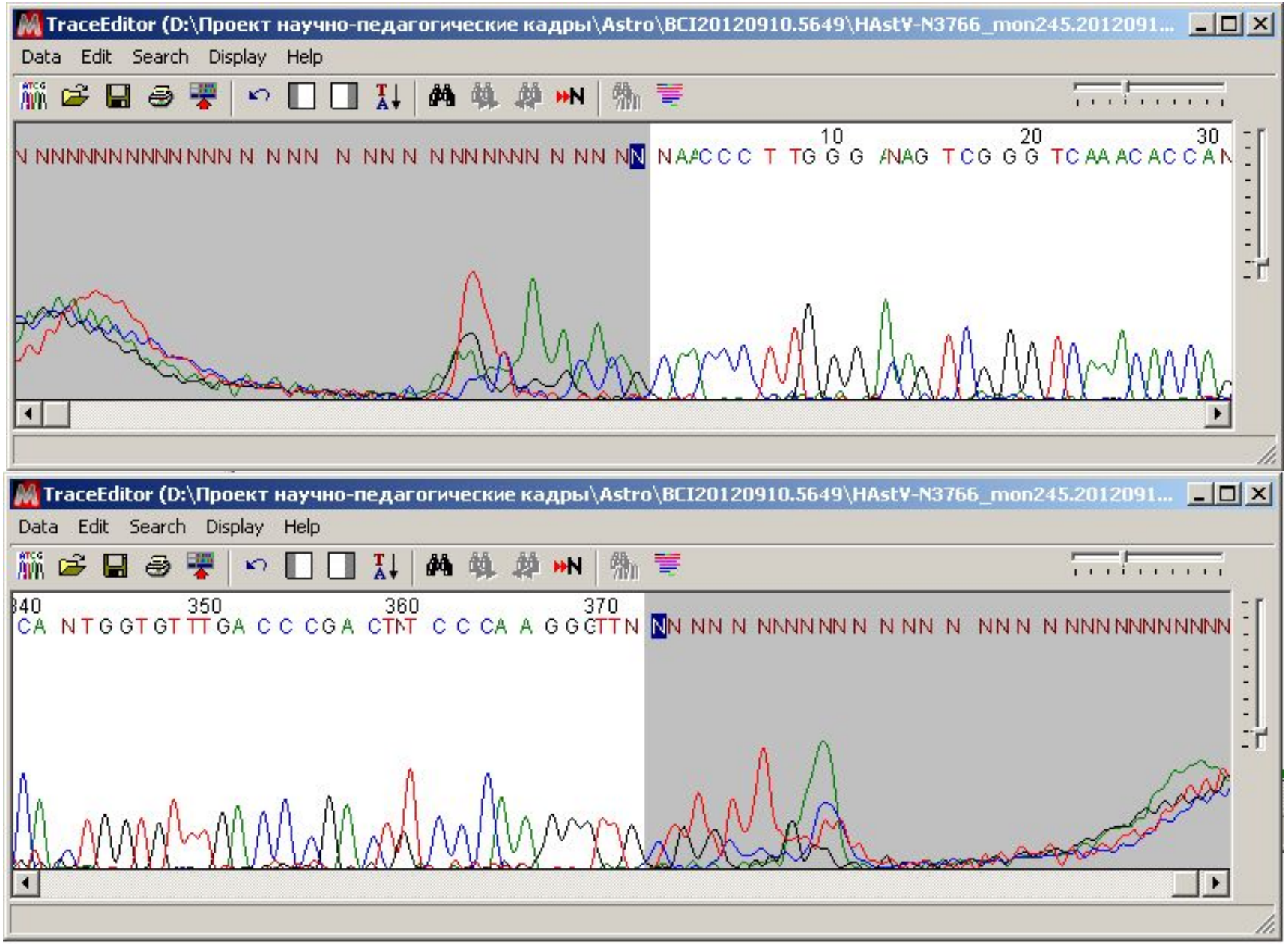
Визуализаторы

- FinchTV (бесплатный)
- Sequencher (платный)
- TraceEditor (входит в пакет ПО MEGA)
- Sequence Scaner (бесплатный)

TraceEditor (1)

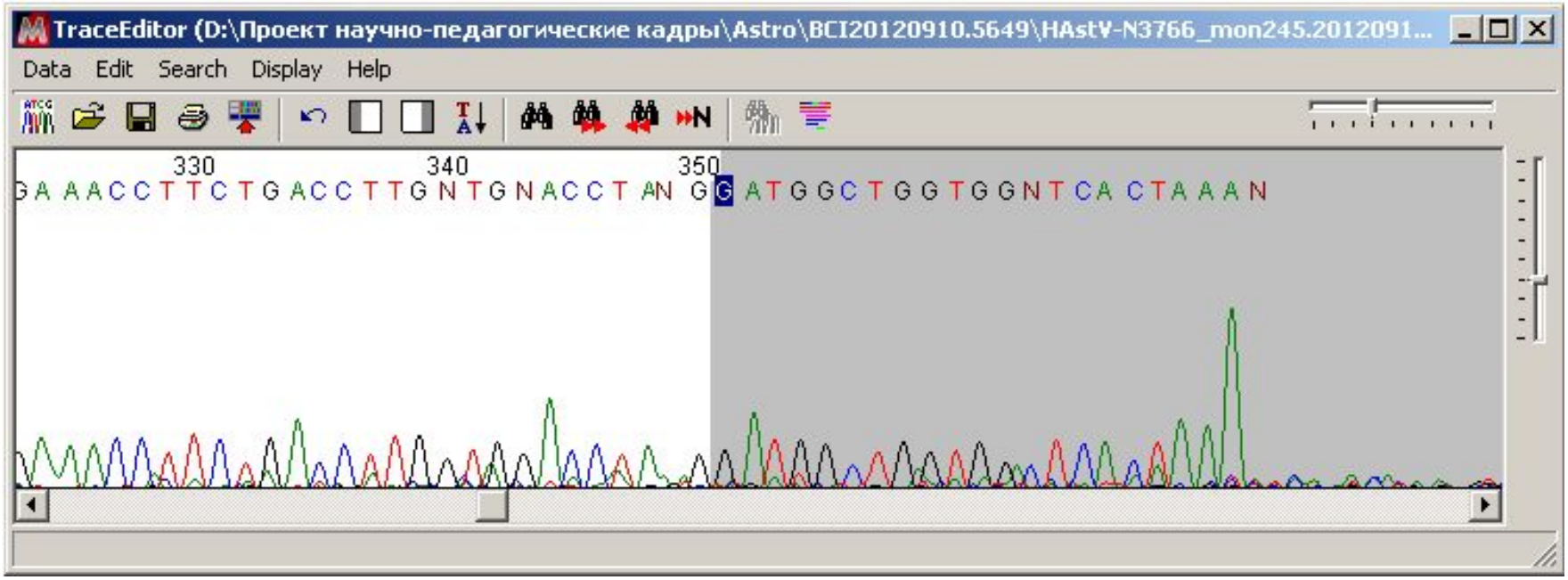


TraceEditor (2)



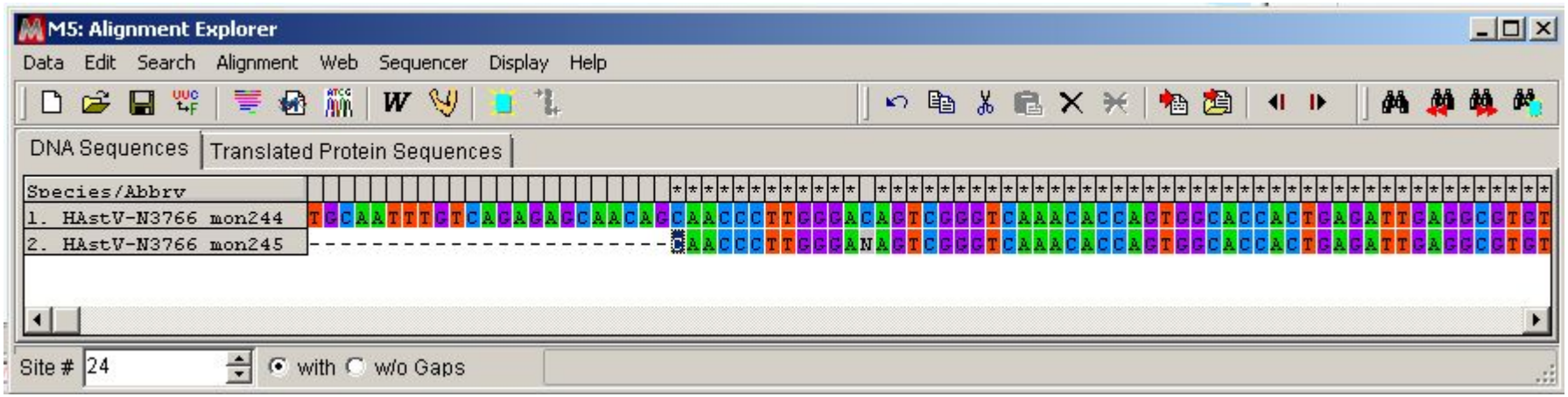
TraceEditor (3)

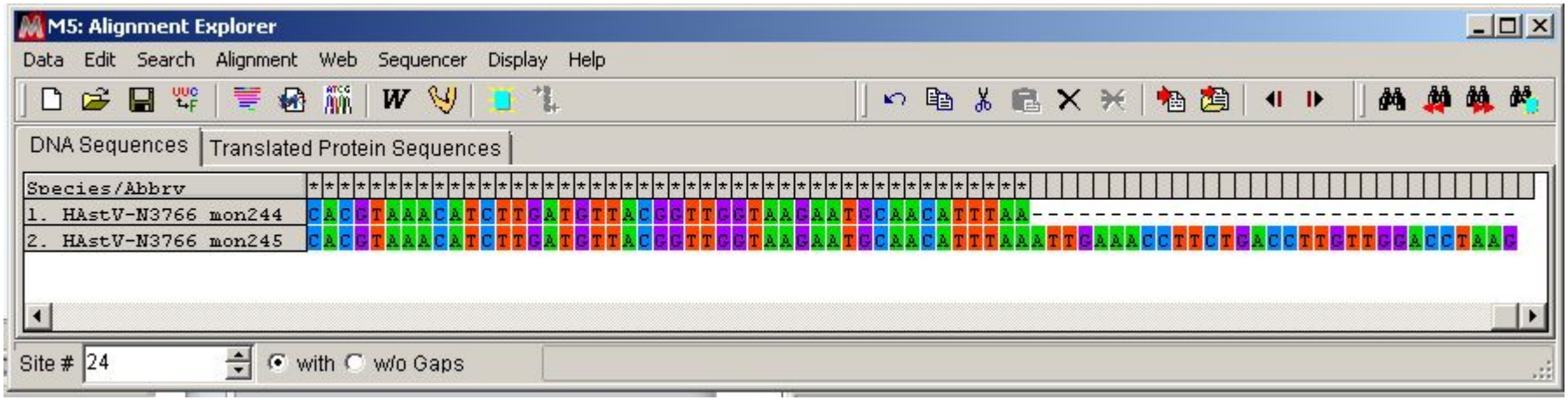
TraceEditor (4)



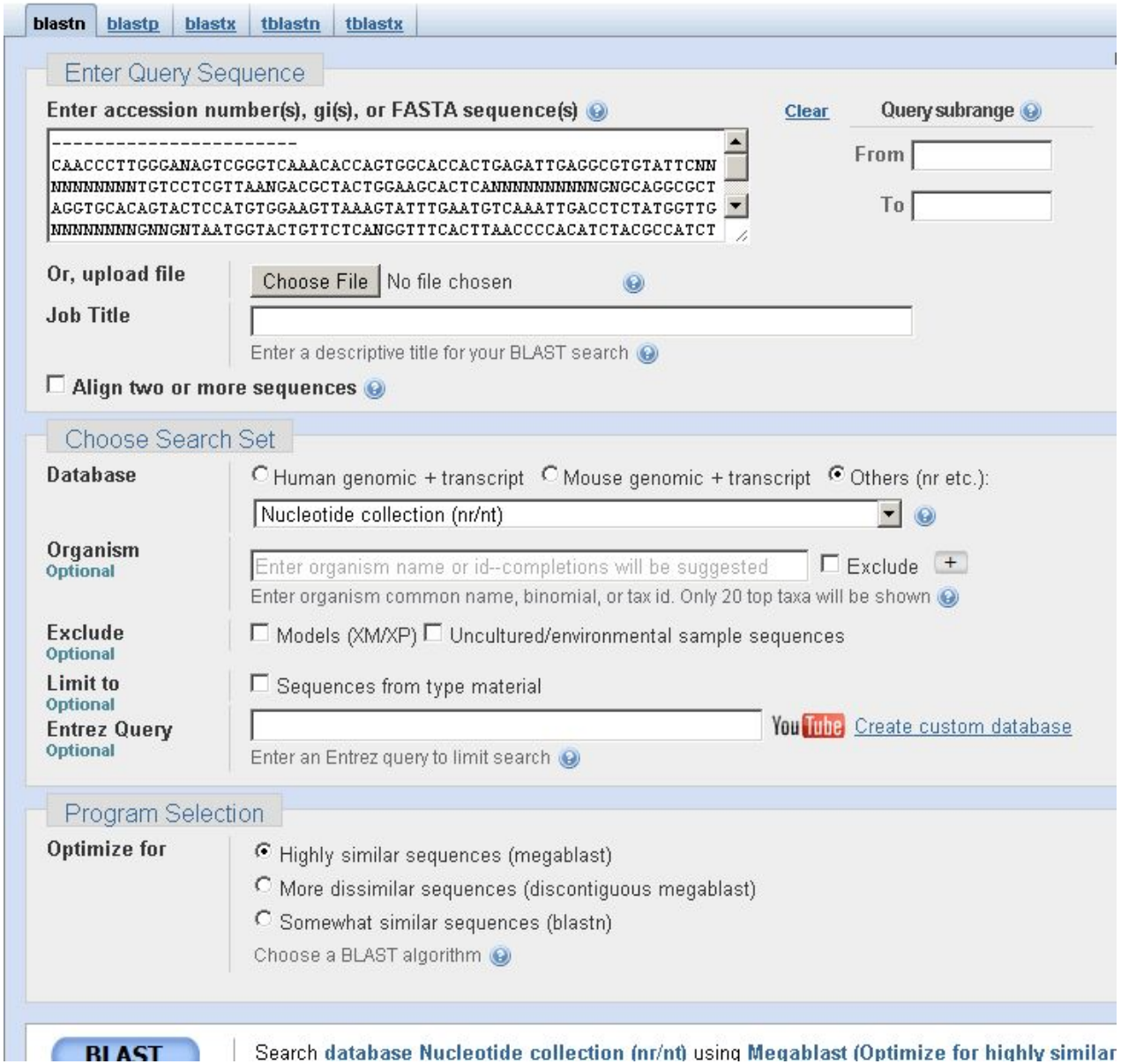
5’-TTAGTGAGCCACCACCAGCCATC-3’

Alignment


M5: Alignment Explorer
Data Edit Search Alignment Web Sequencer Display Help
DNA Sequences
Translated Protein Sequences
Species/Abbrv
1. HAstV-N3766 mon244
2. HAstV-N3766 mon245
Site # 24
with
w/o Gaps

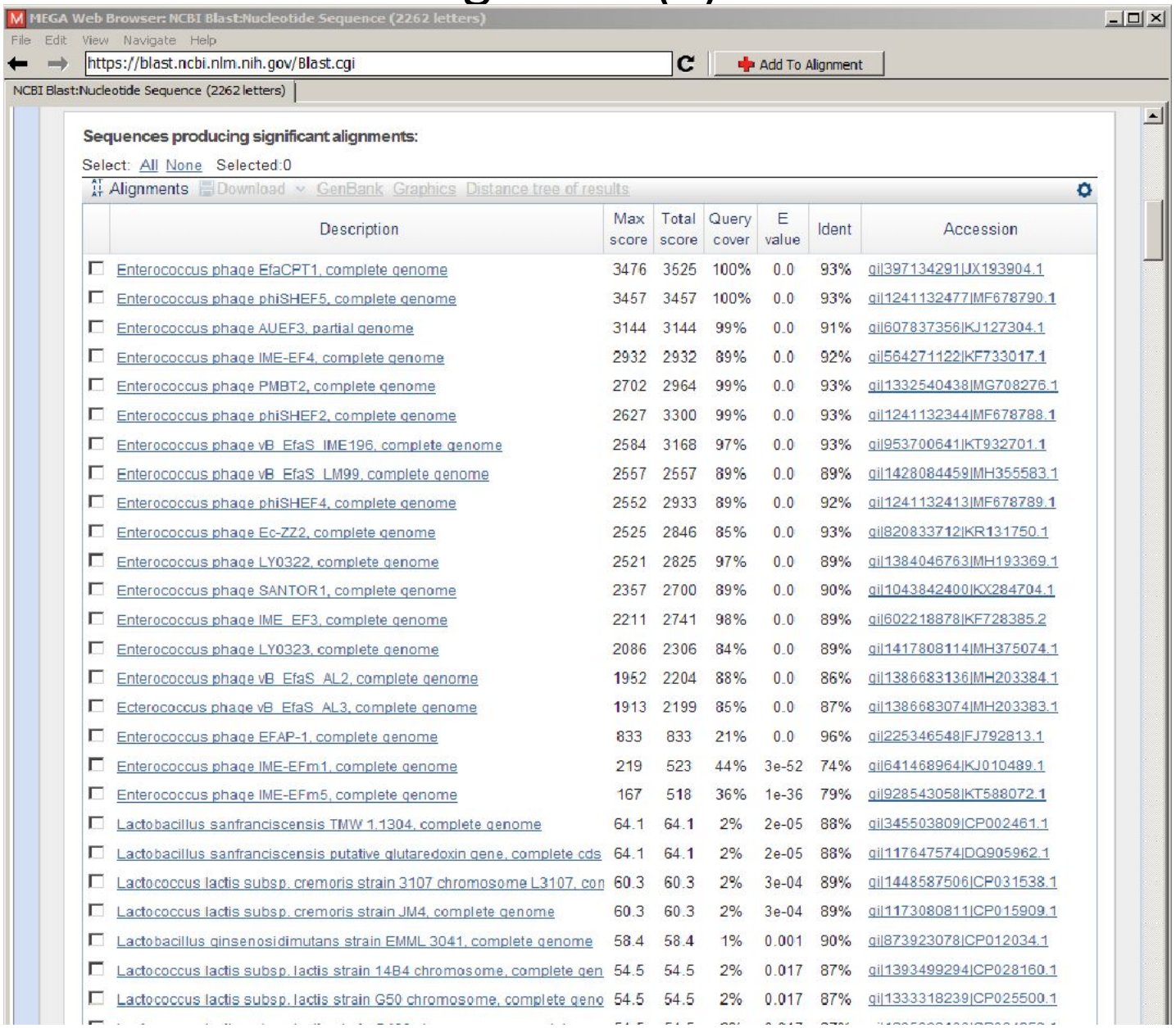

M5: Alignment Explorer
Data Edit Search Alignment Web Sequencer Display Help
DNA Sequences
Translated Protein Sequences
Species/Abbrv
1. HAstV-N3766 mon244
2. HAstV-N3766 mon245
Site # 24
with
w/o Gaps

Megablast (1)


blastn
blastp
blastx
tblastn
tblastx
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To

CAACCCTTGGGANAGTCGGGTCAAACACCAGTGGCACCACTGAGATTGAGGCGTGTATTCNN
NNNNNNNNTGTCCTCGTTAANGACGCTACTGGAAGCACTCANNNNNNNNNNGNGCAGGCGCT
AGGTGCACAGTACTCCATGTGGAAGTTAAAGTATTTGAATGTCAAATTGACCTCTATGGTTG
NNNNNNNNGNNGNTAATGGTACTGTTCTCANGGTTTCACTTAACCCCACATCTACGCCATCT
Or, upload file
Choose File
No file chosen
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Human genomic + transcript
Mouse genomic + transcript
Others (nr etc.):
Nucleotide collection (nr/nt)
Organism
Optional
Enter organism name or id--completions will be suggested
Exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Limit to
Optional
Sequences from type material
Entrez Query
Optional
YouTube
Create custom database
Enter an Entrez query to limit search
Program Selection
Optimize for
Highly similar sequences (megablast)
More dissimilar sequences (discontiguous megablast)
Somewhat similar sequences (blastn)
Choose a BLAST algorithm
BLAST
Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar

Megablast (2)

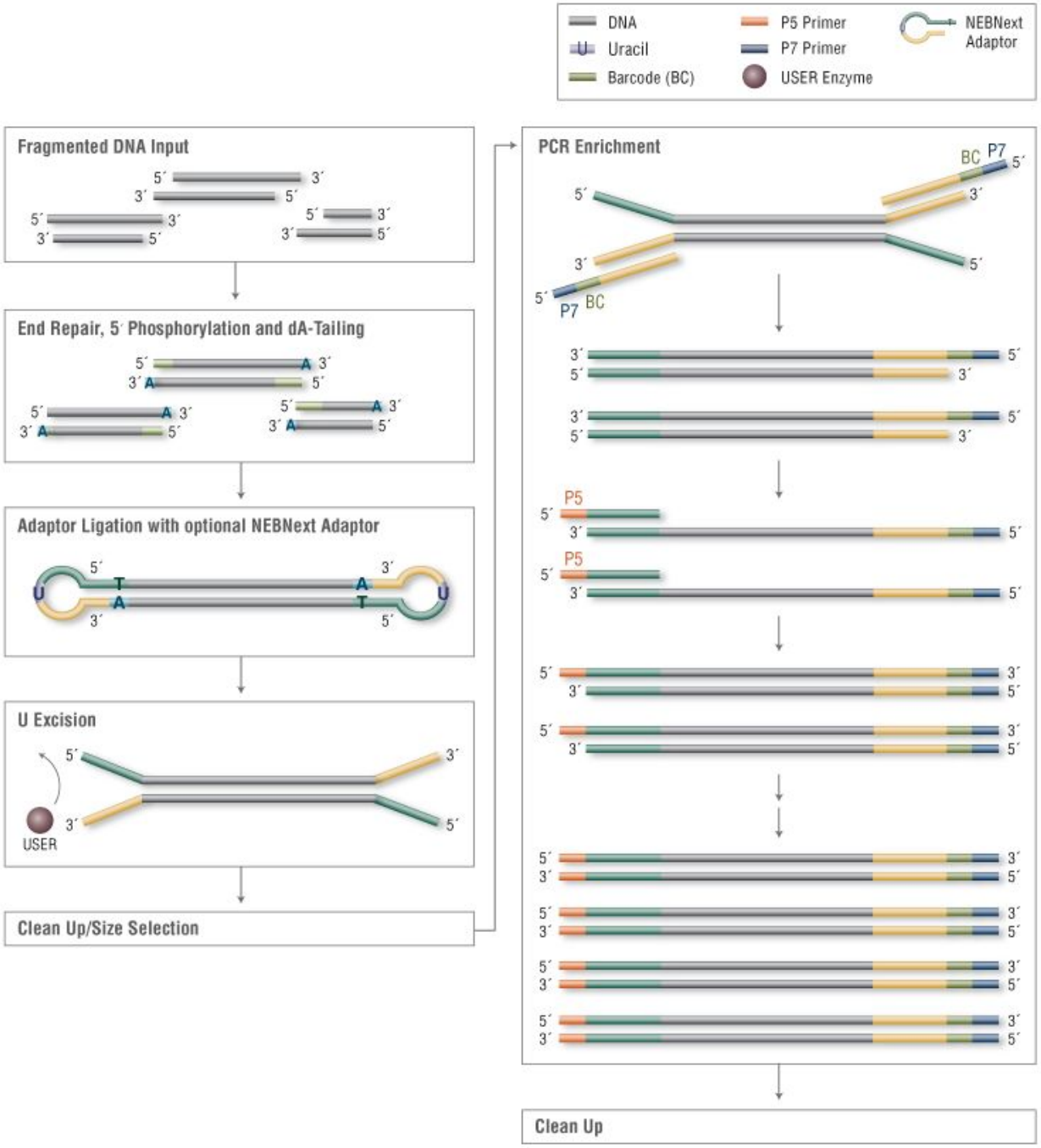


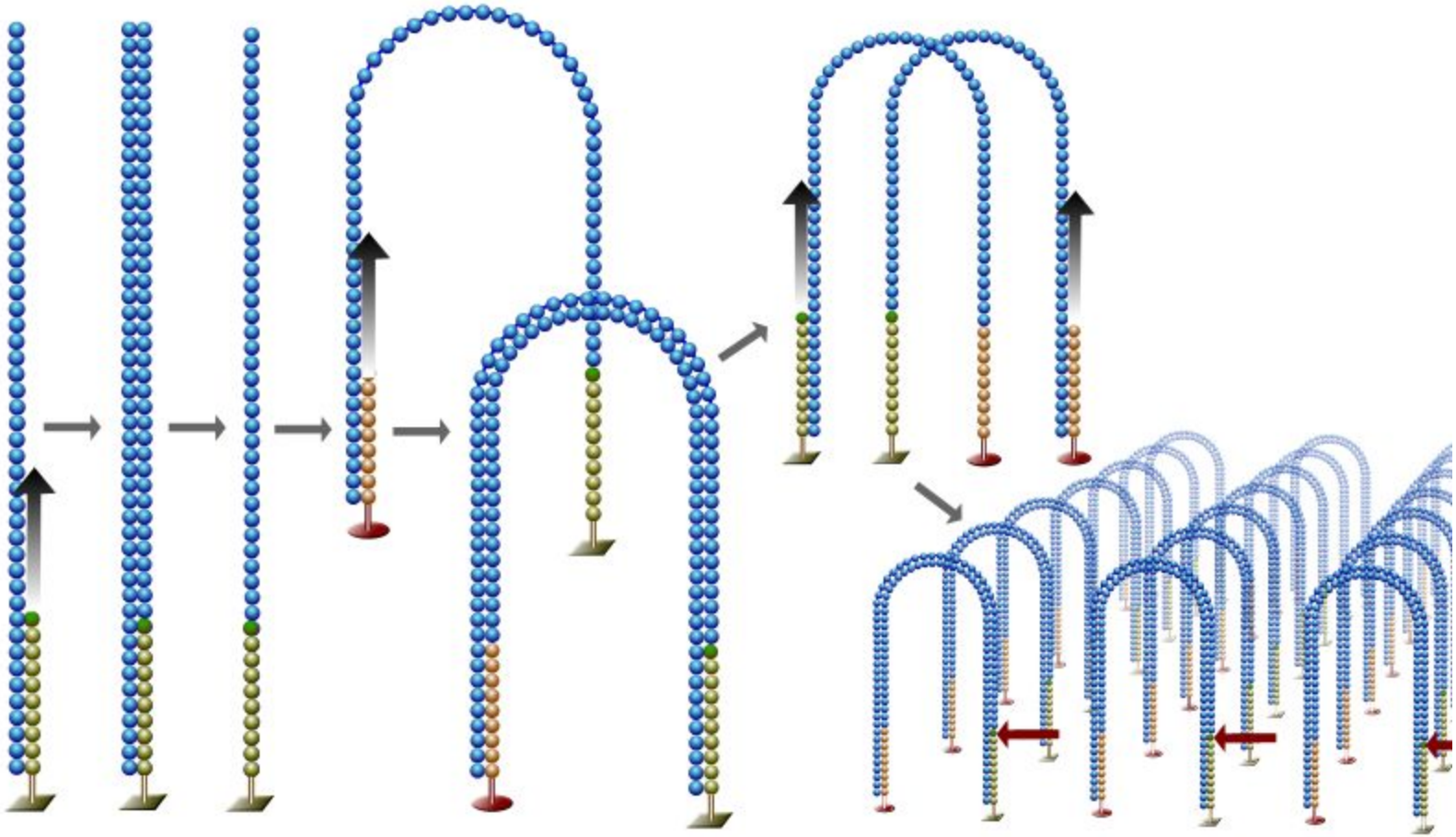
Sequences producing significant alignments:

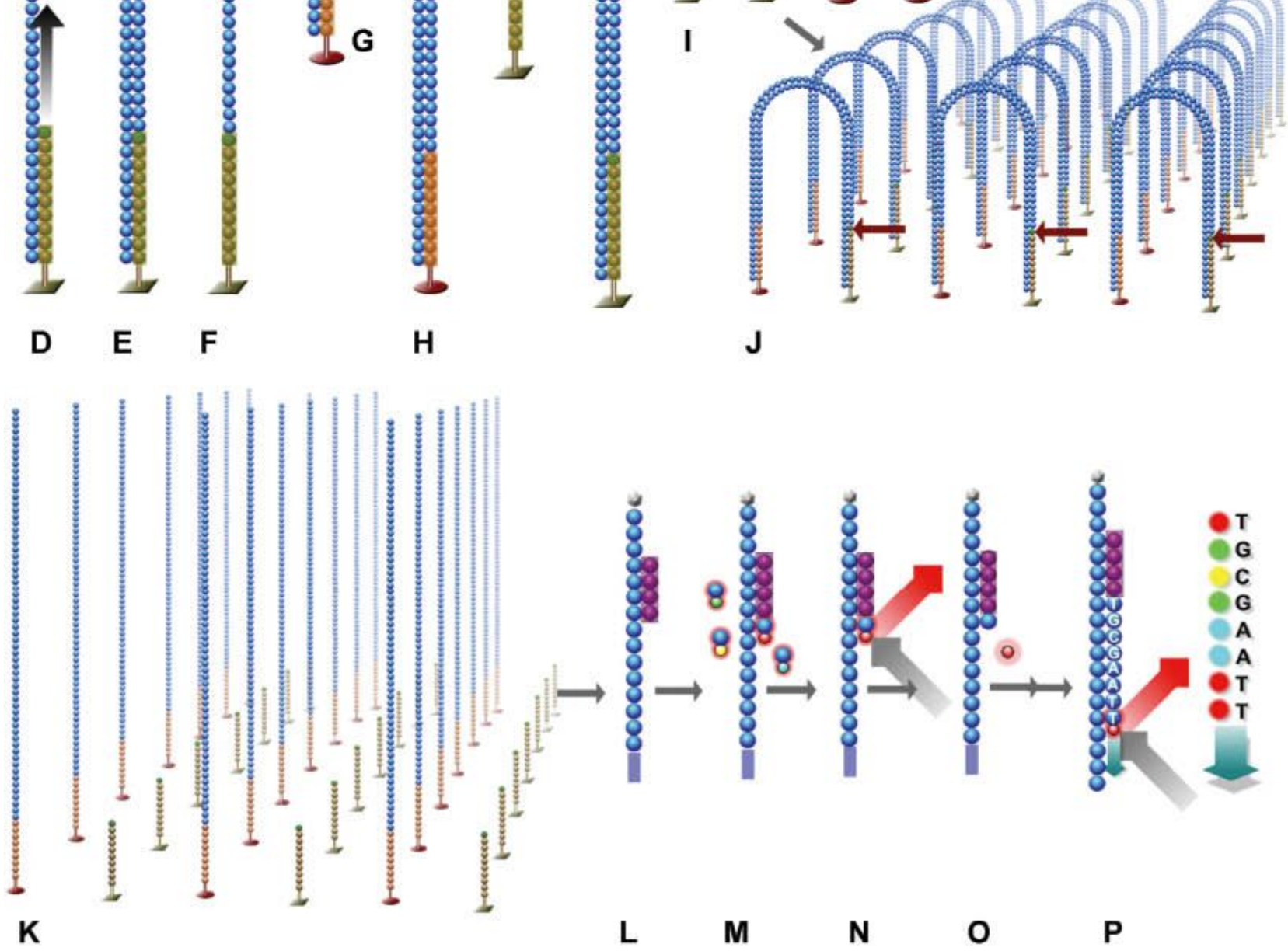
Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

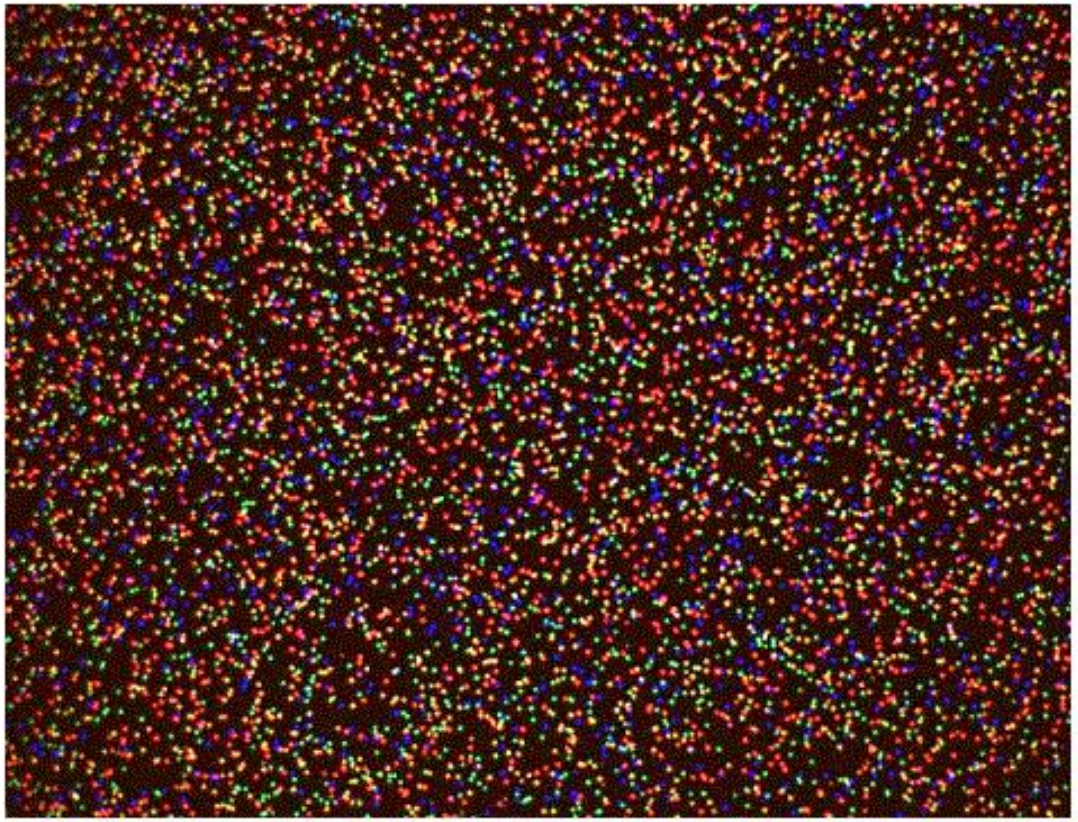
Description	Max score	Total score	Query cover	E value	Ident	Accession
Enterococcus phage EfaCPT1, complete genome	3476	3525	100%	0.0	93%	gi|397134291|JX193904.1
Enterococcus phage phiSHEF5, complete genome	3457	3457	100%	0.0	93%	gi|1241132477|MF678790.1
Enterococcus phage AUEF3, partial genome	3144	3144	99%	0.0	91%	gi|607837356|KJ127304.1
Enterococcus phage IME-EF4, complete genome	2932	2932	89%	0.0	92%	gi|564271122|KF733017.1
Enterococcus phage PMBT2, complete genome	2702	2964	99%	0.0	93%	gi|1332540438|MG708276.1
Enterococcus phage phiSHEF2, complete genome	2627	3300	99%	0.0	93%	gi|1241132344|MF678788.1
Enterococcus phage vB_EfaS_IME196, complete genome	2584	3168	97%	0.0	93%	gi|953700641|KT932701.1
Enterococcus phage vB_EfaS_LM99, complete genome	2557	2557	89%	0.0	89%	gi|1428084459|MH355583.1
Enterococcus phage phiSHEF4, complete genome	2552	2933	89%	0.0	92%	gi|1241132413|MF678789.1
Enterococcus phage Ec-ZZ2, complete genome	2525	2846	85%	0.0	93%	gi|820833712|KR131750.1
Enterococcus phage LY0322, complete genome	2521	2825	97%	0.0	89%	gi|1384046763|MH193369.1
Enterococcus phage SANTOR1, complete genome	2357	2700	89%	0.0	90%	gi|1043842400|KX284704.1
Enterococcus phage IME_EF3, complete genome	2211	2741	98%	0.0	89%	gi|602218878|KF728385.2
Enterococcus phage LY0323, complete genome	2086	2306	84%	0.0	89%	gi|1417808114|MH375074.1
Enterococcus phage vB_EfaS_AL2, complete genome	1952	2204	88%	0.0	86%	gi|1386683136|MH203384.1
Ecterococcus phage vB_EfaS_AL3, complete genome	1913	2199	85%	0.0	87%	gi|1386683074|MH203383.1
Enterococcus phage EFAP-1, complete genome	833	833	21%	0.0	96%	gi|225346548|FJ792813.1
Enterococcus phage IME-EFm1, complete genome	219	523	44%	3e-52	74%	gi|641468964|KJ010489.1
Enterococcus phage IME-EFm5, complete genome	167	518	36%	1e-36	79%	gi|928543058|KT588072.1
Lactobacillus sanfranciscensis TMW 1.1304, complete genome	64.1	64.1	2%	2e-05	88%	gi|345503809|CP002461.1
Lactobacillus sanfranciscensis putative glutaredoxin gene, complete cds	64.1	64.1	2%	2e-05	88%	gi|117647574|DQ905962.1
Lactococcus lactis subsp. cremoris strain 3107 chromosome L3107, con	60.3	60.3	2%	3e-04	89%	gi|1448587506|CP031538.1
Lactococcus lactis subsp. cremoris strain JM4, complete genome	60.3	60.3	2%	3e-04	89%	gi|1173080811|CP015909.1
Lactobacillus ginsenosidimutans strain EMML 3041, complete genome	58.4	58.4	1%	0.001	90%	gi|873923078|CP012034.1
Lactococcus lactis subsp. lactis strain 14B4 chromosome, complete gen	54.5	54.5	2%	0.017	87%	gi|1393499294|CP028160.1
Lactococcus lactis subsp. lactis strain G50 chromosome, complete geno	54.5	54.5	2%	0.017	87%	gi|1333318239|CP025500.1


DNA
Uracil
Barcode (BC)
P5 Primer
P7 Primer
USER Enzyme
NEBNext Adaptor
Fragmented DNA Input
End Repair, 5´ Phosphorylation and dA-Tailing
Adaptor Ligation with optional NEBNext Adaptor
U Excision
USER
Clean Up/Size Selection
PCR Enrichment
BC
P7
P5
Clean Up




G
I
D
E
F
H
J
K
L
M
N
O
P
T
G
C
G
A
A
T
T

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

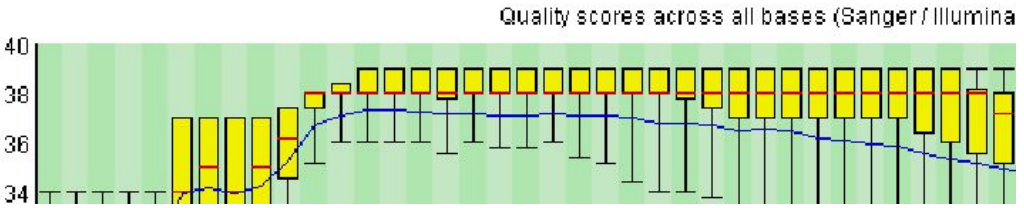
Summary

- 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

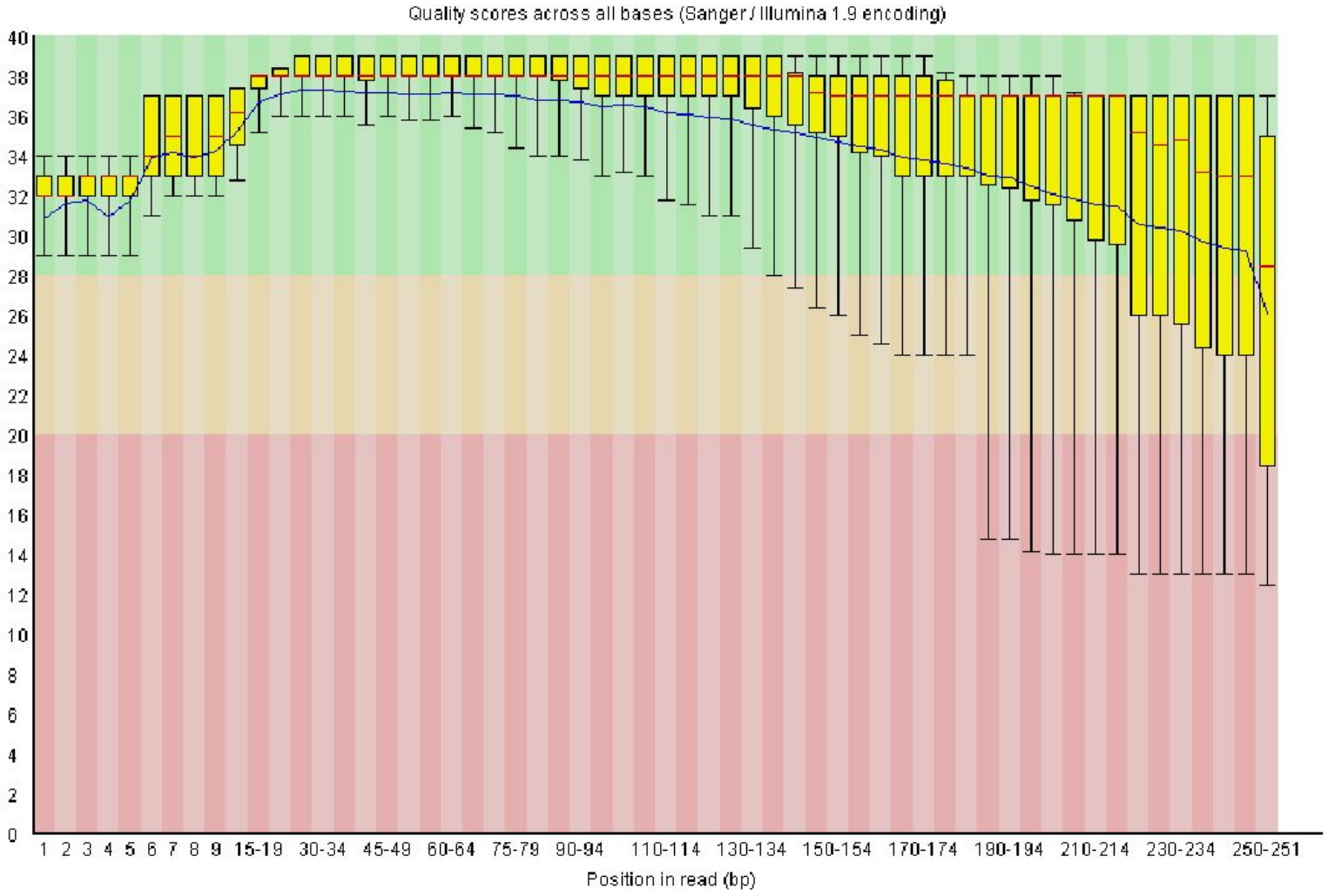
Measure	Value
Filename	92_S92_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	114027
Sequences flagged as poor quality	0
Sequence length	35-251
%GC	60

Per base sequence quality



FastQC

Per base sequence quality



FastQC

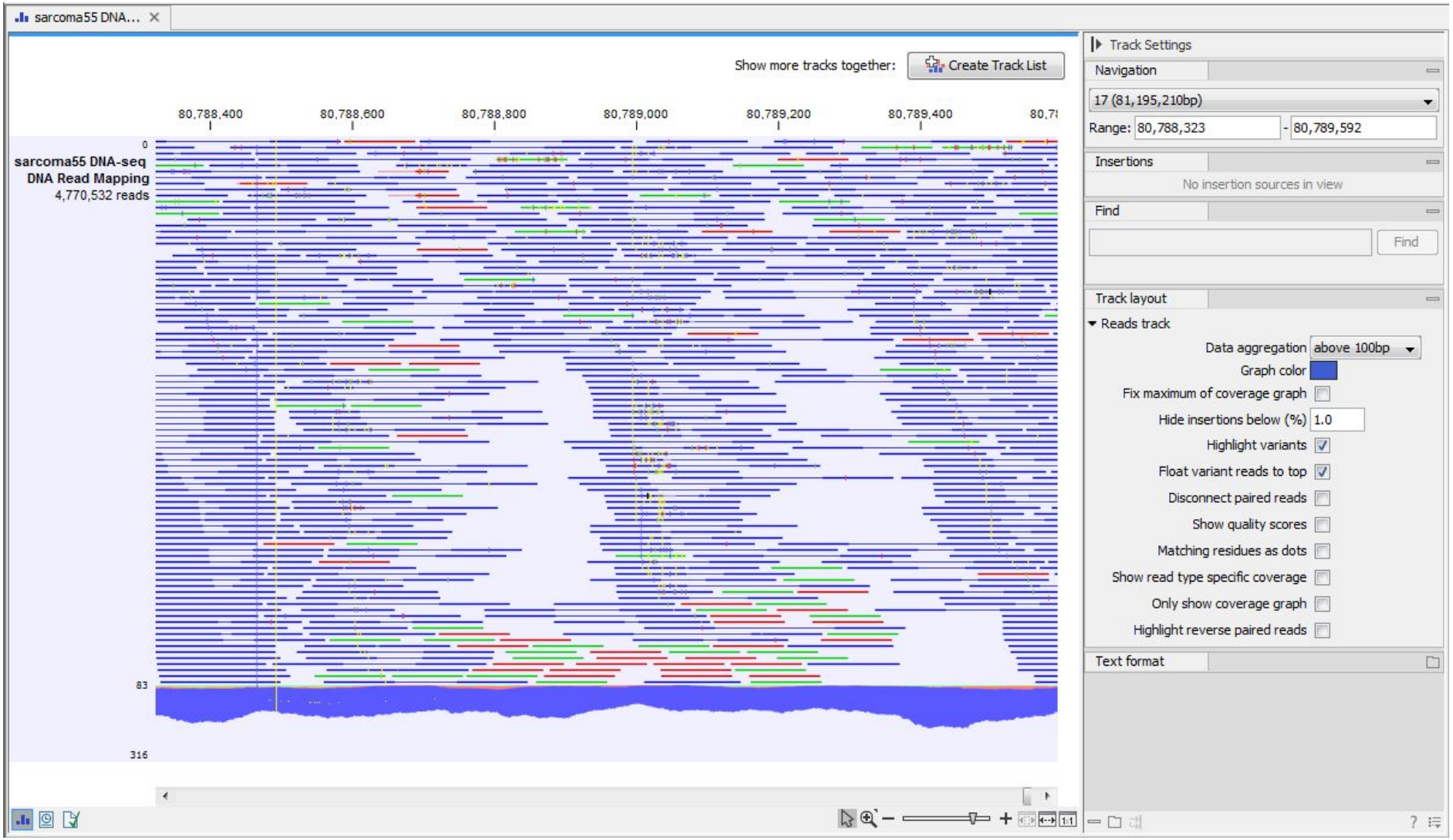
Overrepresented sequences

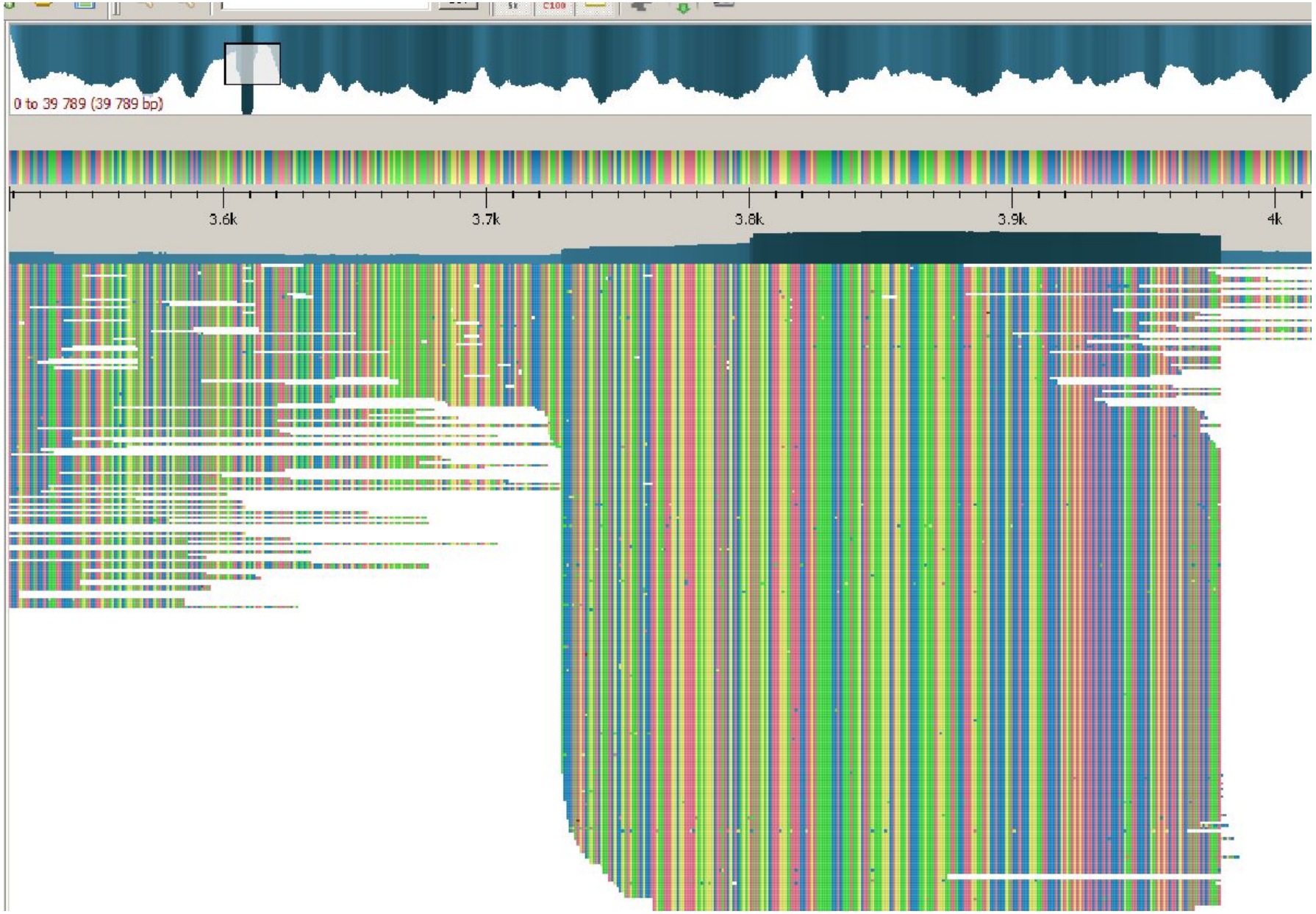
Sequence	Count	Percentage	Possible Source
CCCGACTCTGATCAGCCCACCCTTCCCCCGTAGCCCCTCGTCTGGTGGGC	216	0.18942881949012078	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGCGATAGATCTCGTAT	215	0.18855183421470353	TruSeq Adapter, Index 1 (97% over 36bp)
CTCCGGCCTTAAACCCACATCCAAAAGAGAGAGAATCGCATGAGCTTTCT	122	0.10699220360090154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