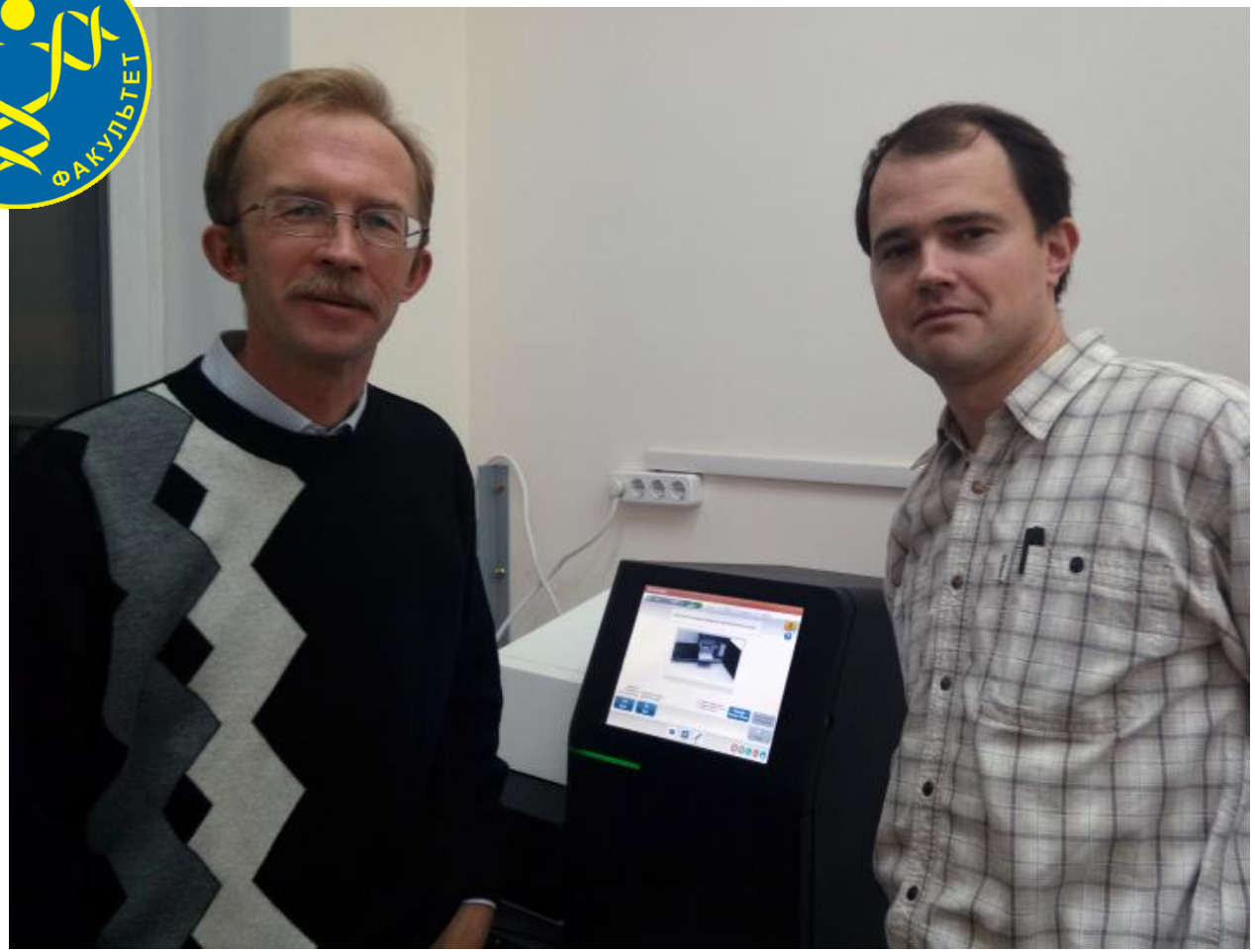
Направления работы Института микробиологии НАН Беларуси в области геномики прокариот

Валентович Л.Н.

Лаборатория «Центр аналитических и генно-инженерных исследований»
Институт микробиологии Национальной академии наук Беларуси

Биологический факультет БГУ

сотрудничает с нами…



Отчёт о первой пятилетке

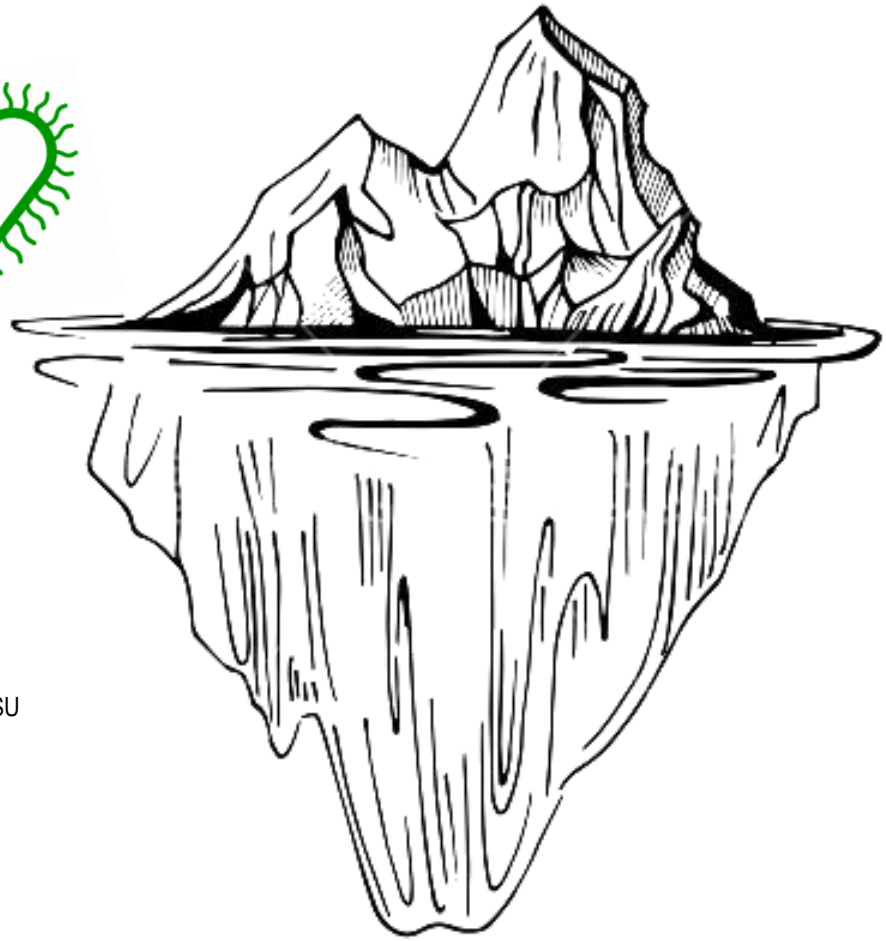
Список проектов по секвенированию (2014-2018 гг.)

- **Дрожжей:**

1) *Glaciozyma antarctica* t3-1a

- **Бактерий:**

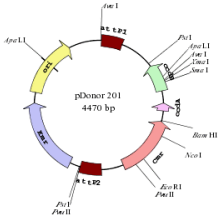
1) *Pectobacterium carotovorum* 3-2 **(Genbank: CP024842)**
2) *Pectobacterium carotovorum* 14A **(Genbank: CP034276-CP034278)**
3) *Pectobacterium atrosepticum* 36A **(Genbank: CP024956)**
4) *Bacillus velezensis* BIM B-439D **(Genbank: CP032144)**
5) *Staphylococcus warneri* 22-1 **(Genbank: CP032158-CP032159)**
6) *Erwinia amylovora* E2 **(Genbank: CP024970-CP024971)**
7) *Erwinia amylovora* 1/79
8) *Pectobacterium carotovorum* 25.1
9) *Bifidobacterium animalis* subsp. *lactis* H1 (BioProject ID: PRJNA263203)
10) *Bifidobacterium animalis* subsp. *lactis* H3 (BioProject ID: PRJNA263288)
11) *Pseudomonas brassicacearum* S-1
12) *Pseudomonas corrugate* 3'
13) *Pseudomonas syringae* pv. *lachrymans* 8
14) *Pseudomonas fluorescens* BIM B-582
15) *Pseudomonas guineae* 7т.4.20.1
16) *Pseudomonas lundensis* 2т.2.5.2
17) *Pseudomonas* sp. 3.C. 1.9
18) *Leifsonia rubra* Хт.6.20.4
19) *Stenotrophomonas maltophilia* 92
20) *Sporosarcina psychrophila* 5t.3.20
21) *Carnobacterium* sp. 11т.7.20.2
22) *Porphyrobacter sanguineus* gip-4
23) *Bacillus pumilus* 63-2-2
24) *Bacillus pumilus* 11-1-1
25) *Bacillus pumilus* F6
26) *Bacillus pumilus* BIM B- 263
27) *Rhodococcus pyridinivorans* L5A-BSU
28) *Clavibacter michiganiensis* Н.П.



Бактериофагов:

1) бактериофаг Pf-10 **(Genbank: KP025626)**
2) бактериофаг φEa2809 **(Genbank: KP037007)**

Мегаплазмид:



1) pBS72 **(Genbank: KX711616)**

- **Метагеномных образцов** (16S рДНК):

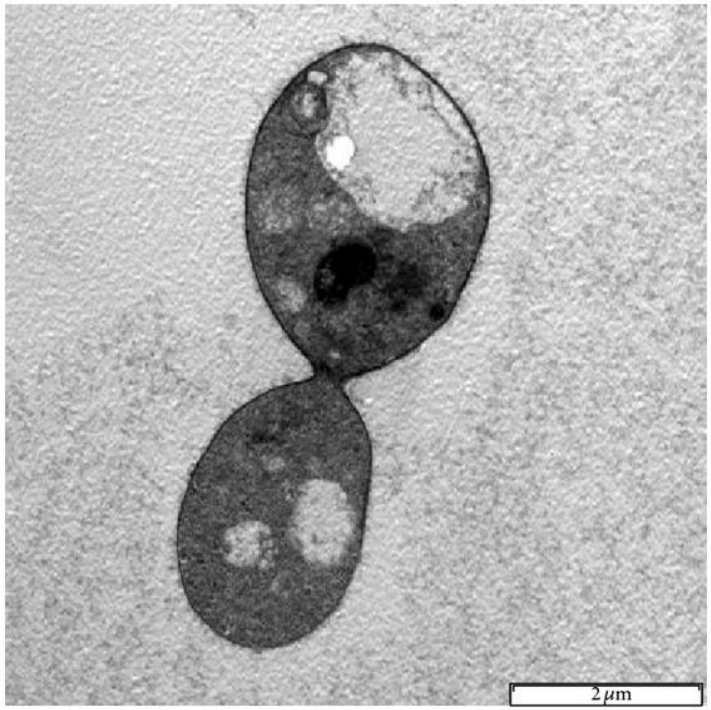
1) Озера Нарочь
2) Различных биотопов Антарктики
3) Микробиома кишечника человека

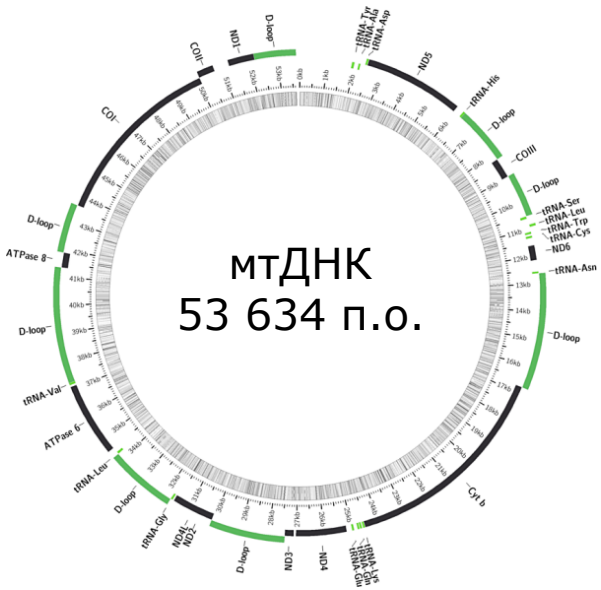
Дрожжи

Glaciozyma antarctica t3-1a

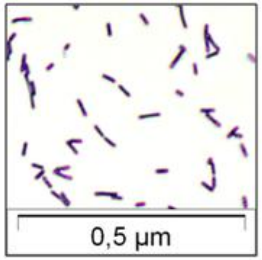
Размер генома: ≈ 20 000 000 п.н.
Прочитано нами: ≈ 19 628 587 п.н.
Количество полученных контигов: ≈ 4000
Среднее покрытие: 18

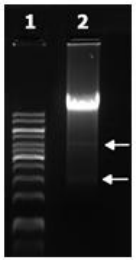
Количество хромосом: ?

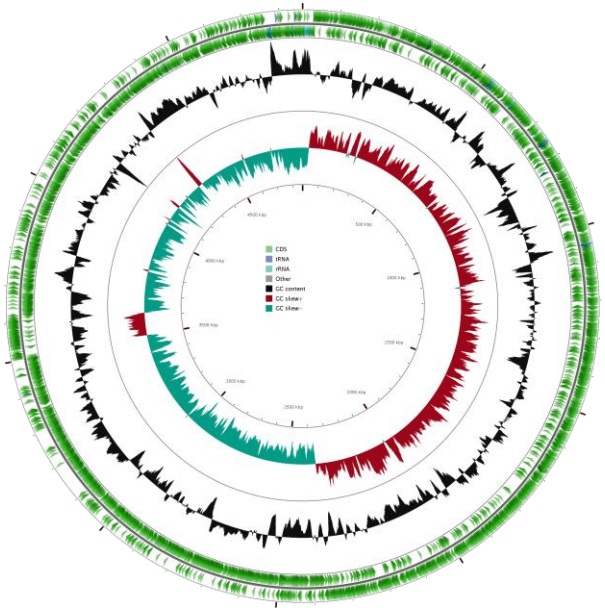


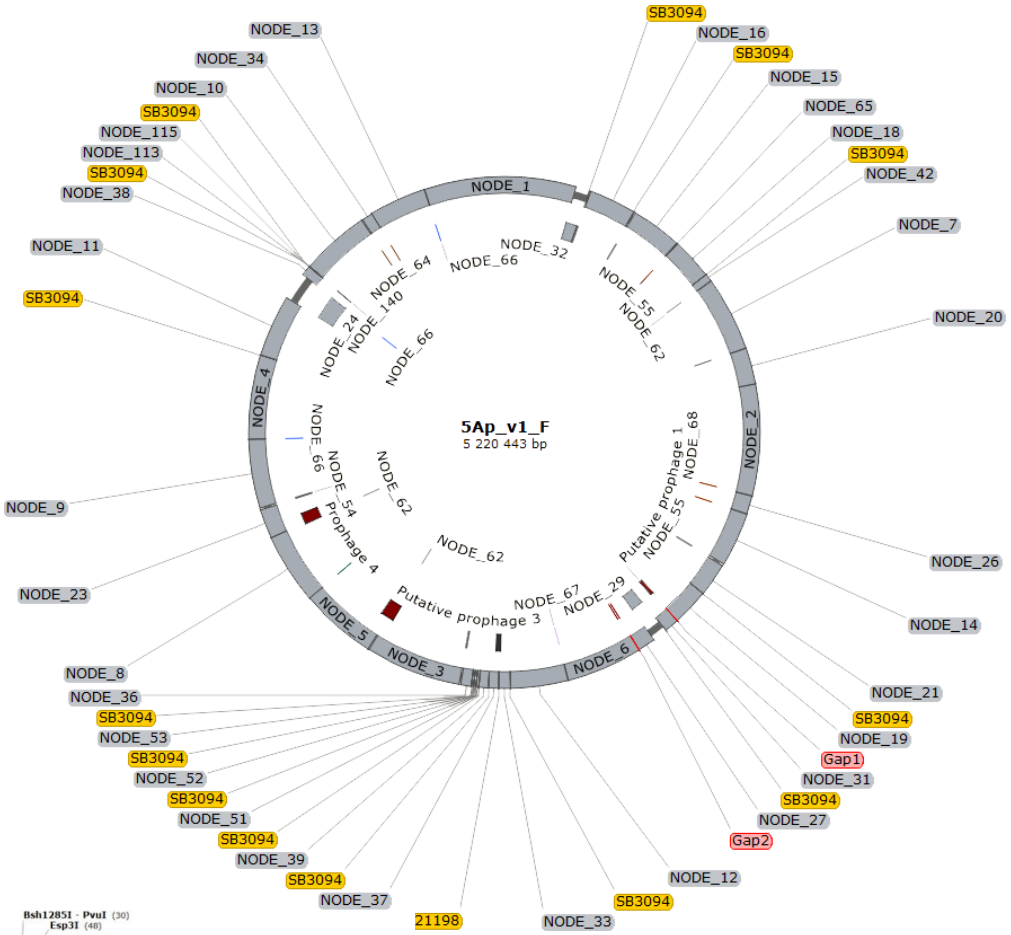


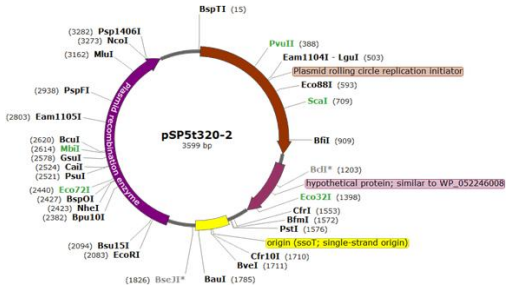
Бактерии

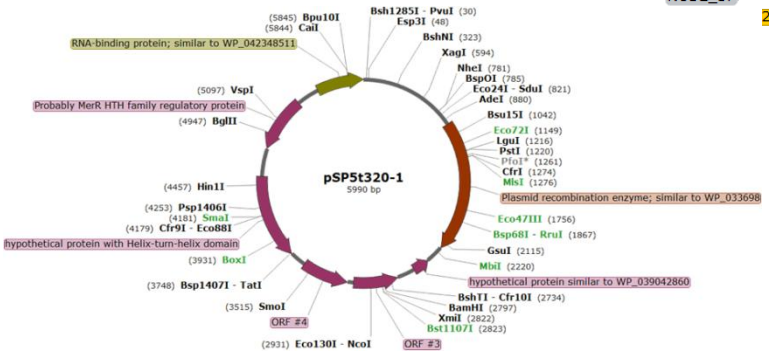

0,5 µm


1
2

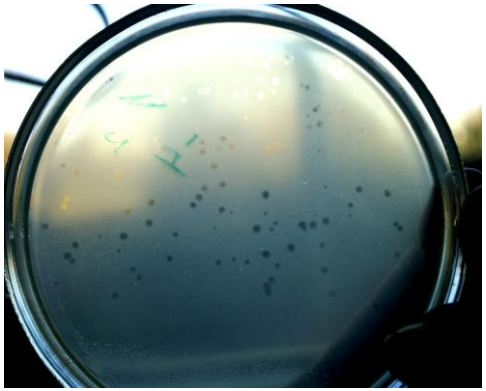


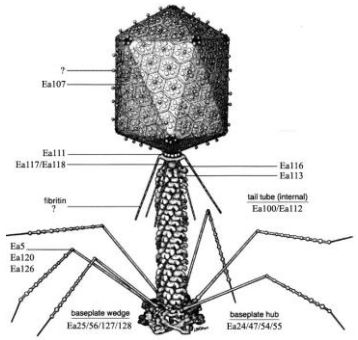

5Ap_v1_F
5 220 443 bp
NODE_1
NODE_2
NODE_3
NODE_4
NODE_5
NODE_6
Putative prophage 1
Putative prophage 3
Prophage 4
SB3094
Gap1
Gap2
21198

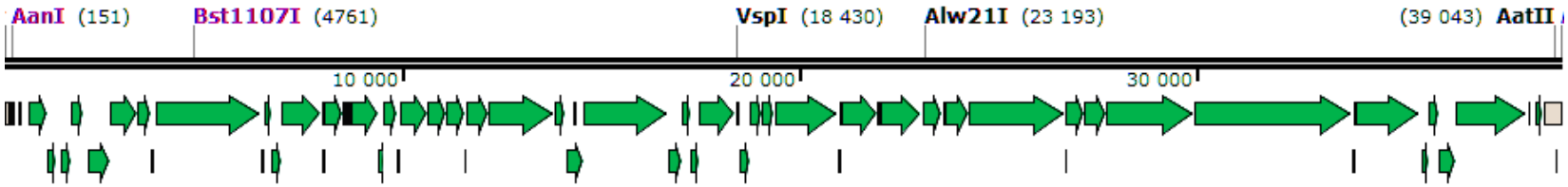

pSP5t320-2
3599 bp
origin of replication
origin (ssoT; single-strand origin)
Plasmid rolling circle replication initiator
hypothetical protein; similar to WP_052246008


pSP5t320-1
5990 bp
RNA-binding protein; similar to WP_042348511
Probably MerR HTH family regulatory protein
hypothetical protein with Helix-turn-helix domain
Plasmid recombination enzyme; similar to WP_033698
hypothetical protein similar to WP_039042860
ORF #4
ORF #3

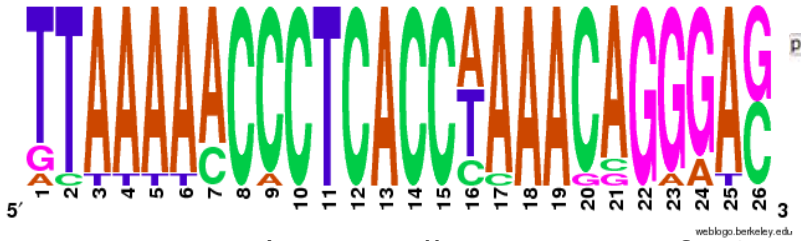
Вирусы и плазмиды



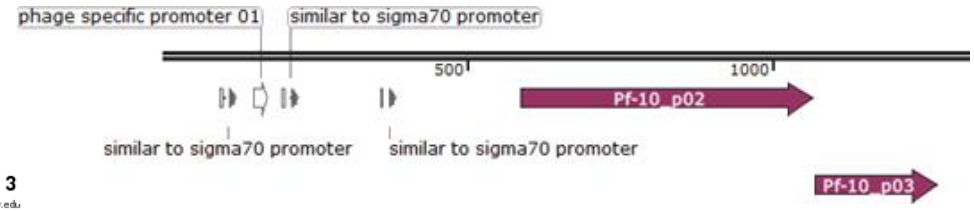




Pf-10
39 167 bp



Фагоспецифический промотор Pf-10



Используемые приборы

Секвенатор **LI-COR 4300**

Определение нуклеотидной последовательности по методу Сэнгера. Размер прочтений – до 1000 пар оснований.

Секвенатор **Illumina MiSeq**

Определение нуклеотидной последовательности с помощью технологии синтеза (SBS) от компании Illumina. Размер прочтения – до 300 пар оснований.

Пробоподготовка:
Nextera XT DNA Library Prep Kit – Illumina

MuSeek Library Preparation Kit, Illumina compatible - Thermo Fisher

Секвенирование: Illumina MiSeq, наборы MiSeq Reagent Kit v3 - 2×300

Исходные данные для работы

Секвенирование по технологии illumina:
на выходе миллионы строк в FASTQ формате

instrument : run#: flowcell_ID :lane: tile : x-pos : y-pos

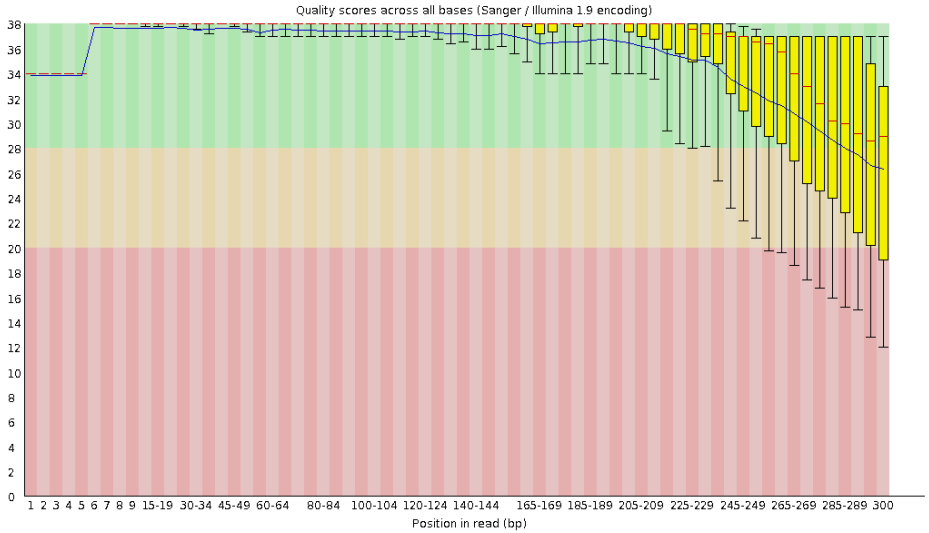
read :
filtered :
control# :
index_sequence

```
@HWI-ST992:147:D22HDACXX:8:2313:8106:8355  1:N:0:ATGTCA
CCTCGTTAAGAATGAACGTAGGGAAGCCGTTAGGCATAGTTACGGAGACCCTTGCATAT
+
CCCFFFFFHHHHHJJJJJHIJJJJJJJJJHIIJJJJJJJIJJJJJGIGIJJJIJJJIJH
@HWI-ST992:147:D22HDACXX:8:2313:8144:8417  1:N:0:ATGTCA
ATAATGAATCTGAAATATAAATTTGGAATTATGATTTACAGATGTTCAAAACTTTGAAC
+
CCCFFFFFHHHGFIGGHIJIIIJJIEGHJGIHGIIGHIIJGJHGIGFIIGHJEHHIJIJ
@HWI-ST992:147:D22HDACXX:8:2313:8391:8324  1:N:0:ATGTCA
AGTCTATTTTTGAATTTAGTGGTATGACCAGGGCTACTACATTTTCCAATATCTTCCAA
+
CC@FFFFFHHHHFIJJJJJHIJFHJJJJJJJJJJJJJJJJJJJJJJJIJJJJJJJJJJJ
@HWI-ST992:147:D22HDACXX:8:2313:8483:8332  1:N:0:ATGTCA
AAATTAATTACAGTTATTGTAATTATTATTATATATTAGGATGATTCACAAATTGGAAT
+
CCCFFFFFHGHHHIIIIJJGGHEIGIIJJJJJJJJIGIFIJIDGGJJIJJJJIGHJJII
```

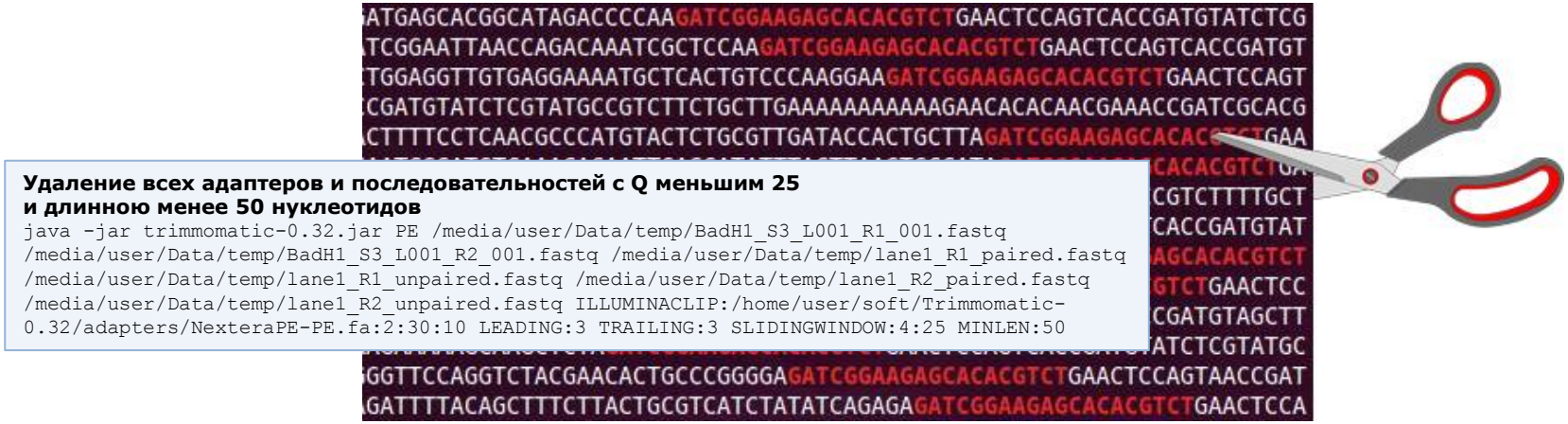
«Стандартный» путь обработки данных

- Анализ и обработка «сырых» данных (прочтений): **FastQC** и **Trimmomatic**
- Сборка прочтений - получение контигов: **SPAdes и A5**
- Фильтрование контигов: **Blast2GO и собственные скрипты**
- Анализ оставшихся контигов и сборка их в скаффолды: **тонкая ручная работа в программе Snapgene**
- Секвенирование по Сэнгеру спорных мест. Проверка с помощью ПЦР.
- Проверка сборки – поиск однонуклеотидных вариантов, небольших вставок и делеций: **Pilon**
- Проверка сборки – картирование прочтений на собранный геном: **BWA и Bowtie 2**, визуализация: **Tablet**
- Автоматическая аннотация генома: **Rast, Prokka, Phast (Phaster)**
- “Ручная” аннотация генома: **Sigmoid, PHIRE, BLAST** и др.

Анализ и обработка «сырых» данных (прочтений): FastQC и Trimmomatic



https://www.bioinformatics.babraham.ac.uk/projects/fastqc/



Удаление всех адаптеров и последовательностей с Q меньшим 25 и длинною менее 50 нуклеотидов

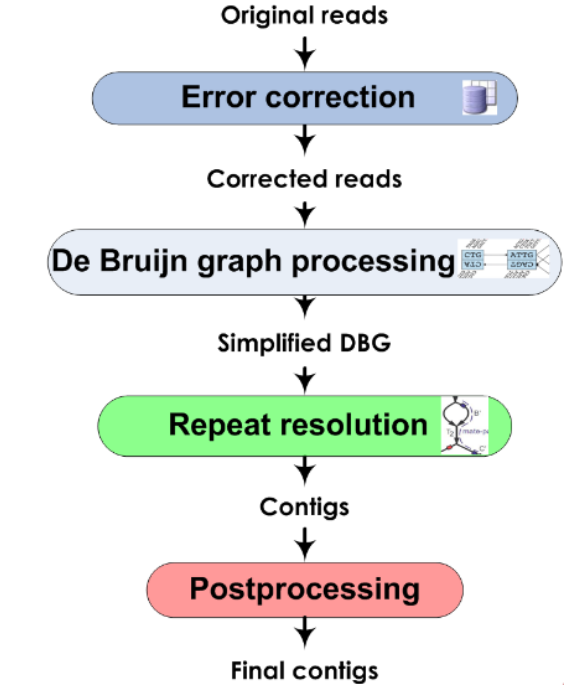
```
java -jar trimmomatic-0.32.jar PE /media/user/Data/temp/BadH1_S3_L001_R1_001.fastq
/media/user/Data/temp/BadH1_S3_L001_R2_001.fastq /media/user/Data/temp/lane1_R1_paired.fastq
/media/user/Data/temp/lane1_R1_unpaired.fastq /media/user/Data/temp/lane1_R2_paired.fastq
/media/user/Data/temp/lane1_R2_unpaired.fastq ILLUMINACLIP:/home/user/soft/Trimmomatic-
0.32/adapters/NexteraPE-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:25 MINLEN:50
```

http://www.usadellab.org/cms/?page=trimmomatic

Сборка прочтений - получение контигов:
SPAdes и A5

SPAdes

http://cab.spbu.ru/software/spades/

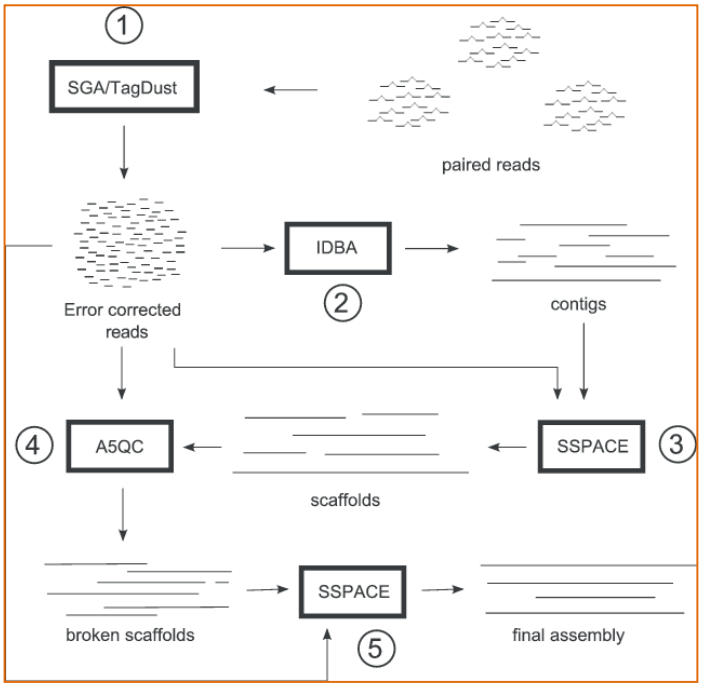


1	2	3
#name	length	coverage
NODE_6	777	14.4252
NODE_4	4666	24.3604
NODE_3	5925	12.647
NODE_2	35102	11.9551
NODE_5	790	62.2327
NODE_1	132734	11.9265

A5-miseq

https://sourceforge.net/projects/ngopt/

Read cleaning: **Trimmomatic**;
Error correction: **SGA**;
Contig assembly: **IDBA-UD**;
Crude scaffolding;
Misassembly correction;
Final scaffolding



Фильтрование контигов: Blast2GO и собственные скрипты

https://www.blast2go.com/

D:\Work\sequences\NGS\H3\blast2go-h3\blast2go_20140811_0859.dat - Blast2GO V.2.8.0

File Blast Mapping Annotation Analysis Statistics Select Tools View Support

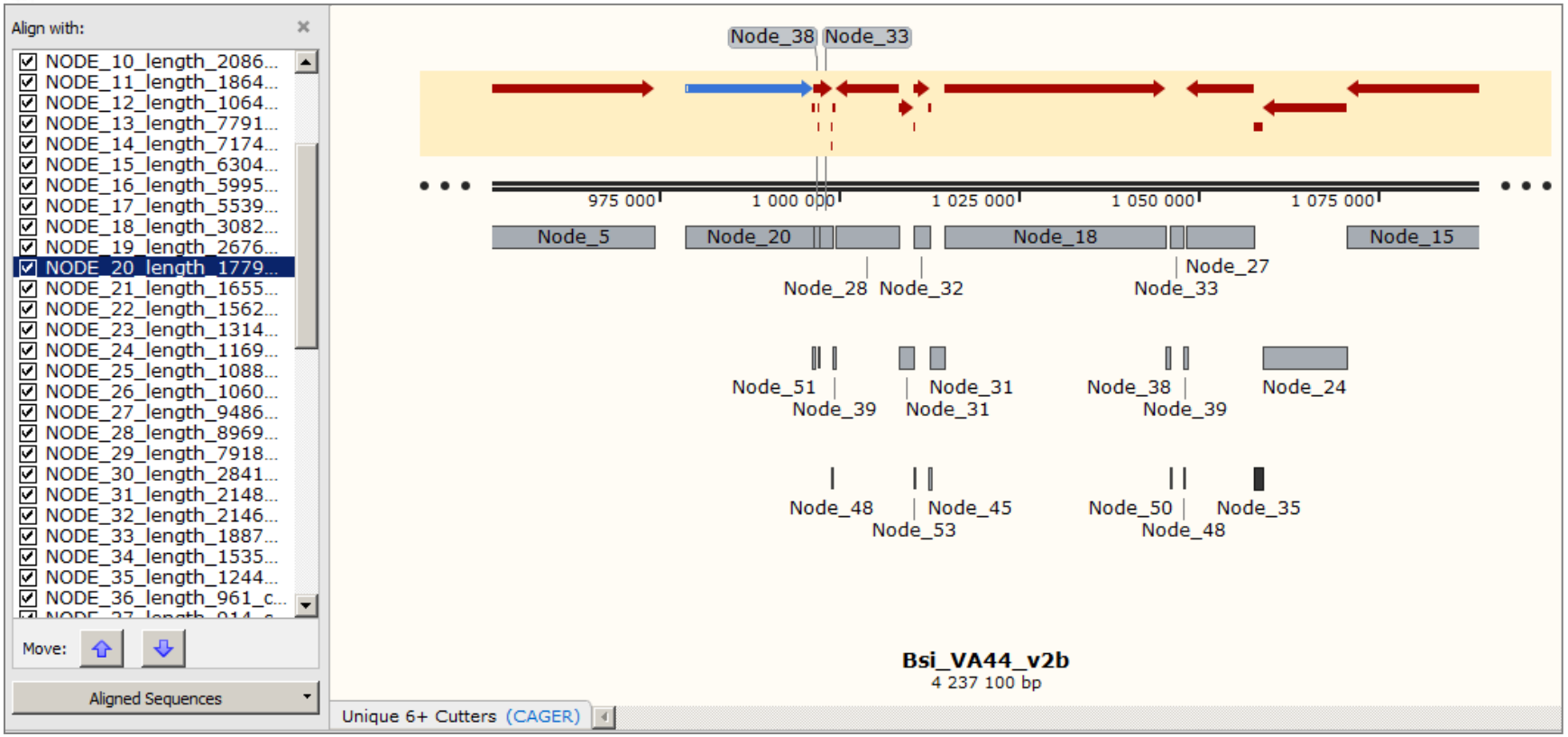
GO:0007067,GO:0016021

nr	sequence name	seq description	length	#hits	min. eValue	sim mean
11	NODE_11_length_5513_cov_7.47791_ID_21	streptococcus vt complete genome	5513	3	0,0E0	99.67%
12	NODE_12_length_3986_cov_51.3921_ID_23	bifidobacterium animalis strain complete g...	3986	3	0,0E0	100.0%
13	NODE_13_length_3069_cov_35.8838_ID_25	bifidobacterium animalis strain complete g...	3069	3	0,0E0	100.0%
14	NODE_14_length_2445_cov_84.5207_ID_27	bifidobacterium animalis lactis bi- complet...	2445	3	0,0E0	100.0%
15	NODE_15_length_1776_cov_58.5992_ID_29	bifidobacterium animalis lactis complete g...	1776	3	0,0E0	99.67%
16	NODE_16_length_1455_cov_3.61295_ID_31	pectobacterium carotovorum carotovorum c...	1455	3	0,0E0	89.33%
17	NODE_17_length_1333_cov_132.844_ID_33	bifidobacterium animalis lactis complete g...	1333	3	0,0E0	100.0%
18	NODE_19_length_1273_cov_0.613438_ID_...	erwinia phage ea9- complete genome	1272	1	0,0E0	98.0%
19	NODE_21_length_1117_cov_0.919192_ID_...	erwinia phage ea9- complete genome	1117	1	0,0E0	99.0%
20	NODE_24_length_1029_cov_5.99335_ID_47	bacillus amyloliquefaciens plantarum com...	1029	3	0,0E0	99.0%
21	NODE_26_length_999_cov_1.79587_ID_51	pseudomonas brassicacearum brassicac...	999	3	0,0E0	94.0%
22	NODE_27_length_993_cov_13.6871_ID_53	pectobacterium carotovorum carotovorum c...	993	3	0,0E0	90.33%
23	NODE_28_length_984_cov_4.61494_ID_55	pseudomonas fluorescens complete geno...	984	3	0,0E0	93.33%
24	NODE_29_length_980_cov_10.4572_ID_57	pectobacterium carotovorum carotovorum c...	980	3	0,0E0	90.67%
25	NODE_31_length_959_cov_15.5264_ID_61	pectobacterium carotovorum carotovorum c...	959	3	0,0E0	87.67%
26	NODE_37_length_925_cov_12.312_ID_73	pectobacte				
27	NODE_39_length_907_cov_6.04231_ID_77	pectobacte				
28	NODE_41_length_906_cov_0.596919_ID_81	erwinia ph				
29	NODE_44_length_893_cov_8.81462_ID_87	bacillus ar				
30	NODE_46_length_886_cov_8.79183_ID_91	---NA---				
31	NODE_49_length_872_cov_4.0953_ID_97	pectobacte				

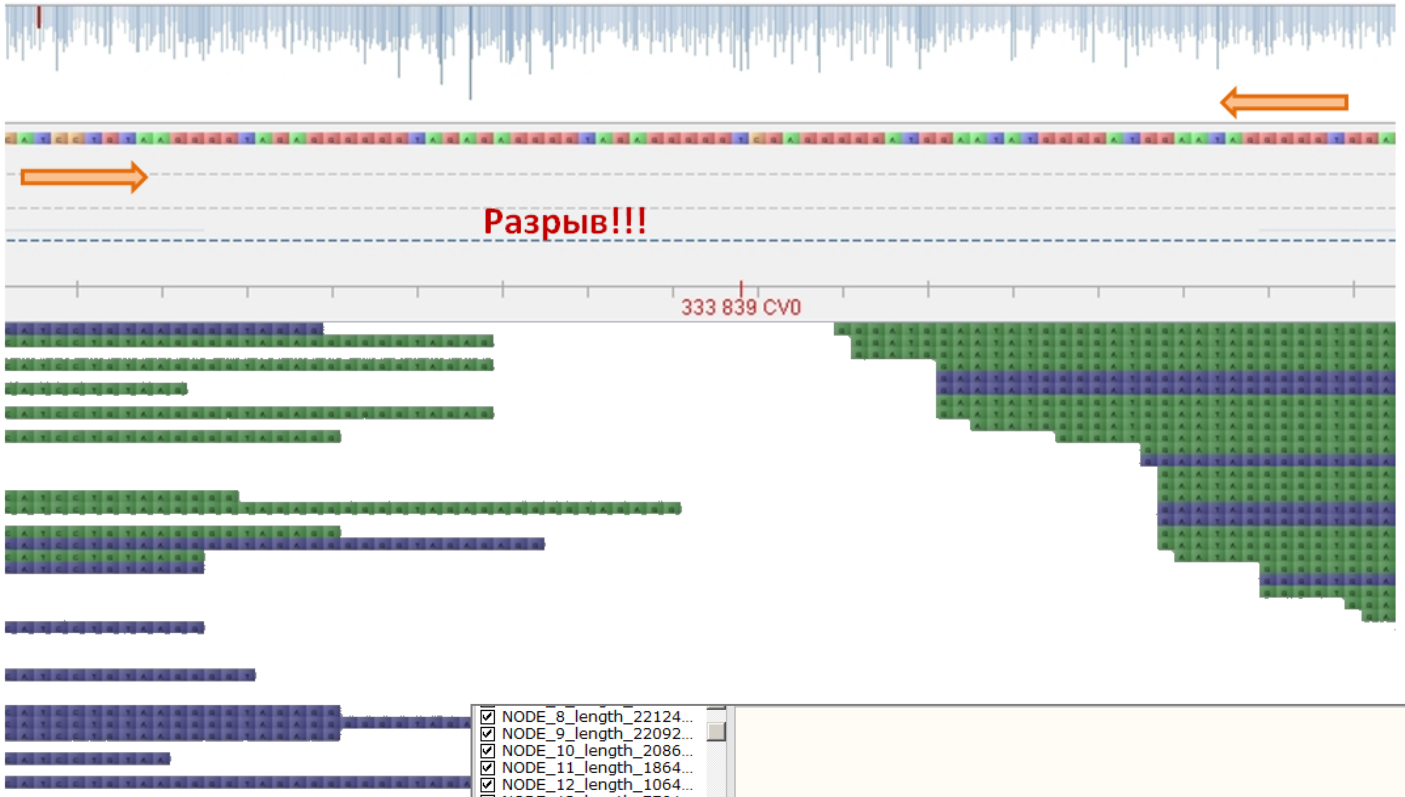
№	Название	Длина	Покрытие	ГЦ	Коэфф.	Аннотация	Начало	Конец
1	NODE_1_length_290866_cov_14.696621	290866	14,696621	61,08	1.0	Pseudomonas corrugata BS3649	-	-
2	NODE_2_length_270694_cov_15.249916	270694	15,249916	59,87	1.0	Pseudomonas corrugata BS3649	-	H_NODE_181 NODE_185_K
3	NODE_3_length_256694_cov_13.726800	256694	13,726800	60,42	0.9	Pseudomonas corrugata BS3649	-	-
4	NODE_4_length_216614_cov_12.766633	216614	12,766633	59,41	0.8	Pseudomonas corrugata RM1-1-4	-	-
5	NODE_5_length_176073_cov_12.056540	176073	12,056540	60,71	0.8	Pseudomonas corrugata RM1-1-4	-	-
6	NODE_6_length_157428_cov_14.060972	157428	14,060972	60,55	0.9	Pseudomonas corrugata RM1-1-4	-	NODE_112_K
7	NODE_7_length_153812_cov_14.955864	153812	14,955864	61,14	1.0	Pseudomonas corrugata BS3649	-	NODE_182_K NODE_184_K

Анализ оставшихся контигов и сборка их в скаффолды: тонкая ручная работа в программе Snapgene

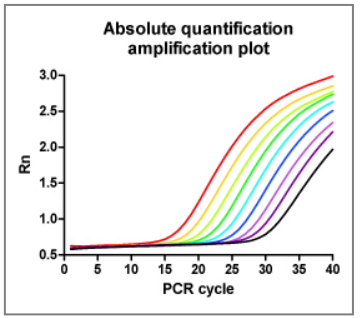
SnapGene®
Software for molecular biology

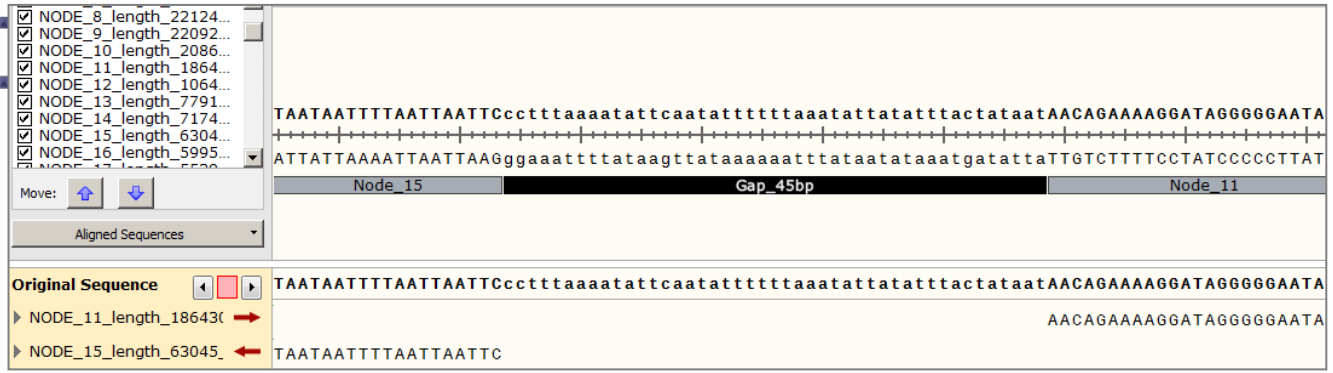


Секвенирование по Сэнгеру спорных мест. Проверка варианта сборки с помощью ПЦР



А также проверка количества повторяющихся элементов с помощью количественной ПЦР



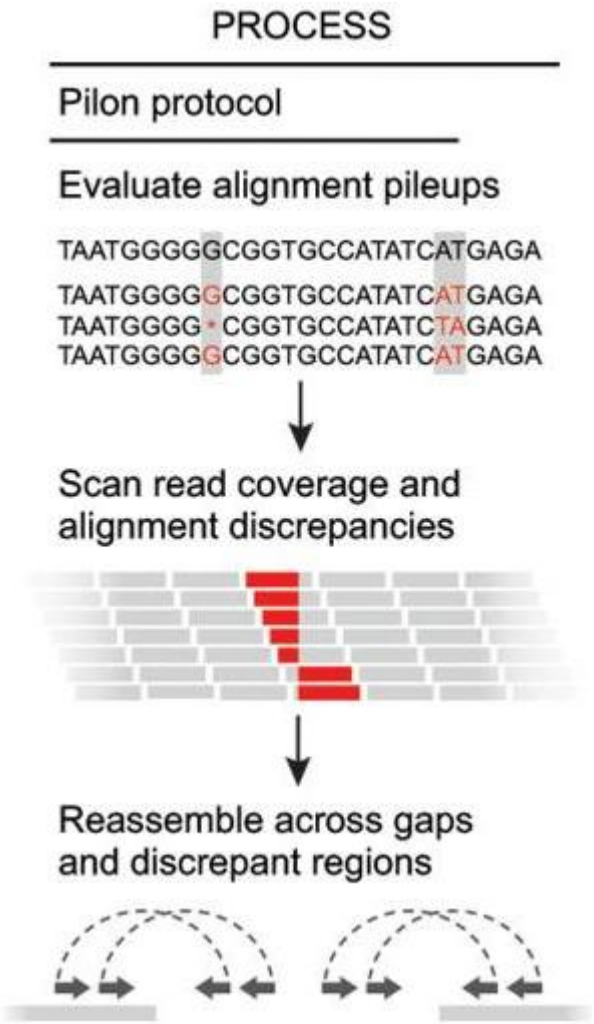


Проверка сборки – поиск однонуклеотидных вариантов, небольших вставок и делеций: Pilon

Pilon

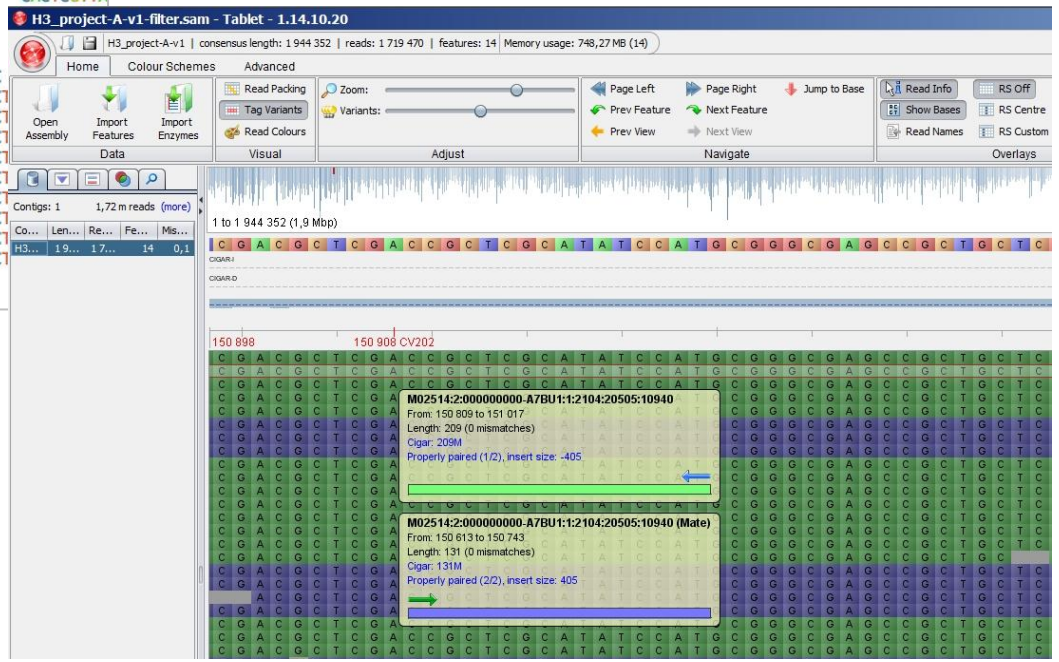
Java

https://github.com/broadinstitute/pilon/



Проверка сборки – картирование прочтений на собранный геном: BWA и Bowtie 2, визуализация: Tablet

https://ics.hutton.ac.uk/tablet/



Burrows-Wheeler Aligner

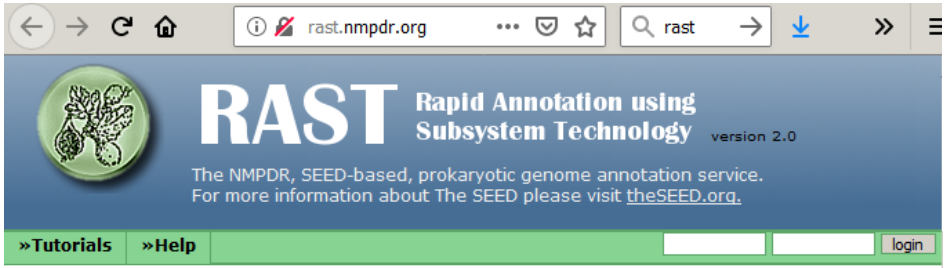
http://bio-bwa.sourceforge.net/

http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

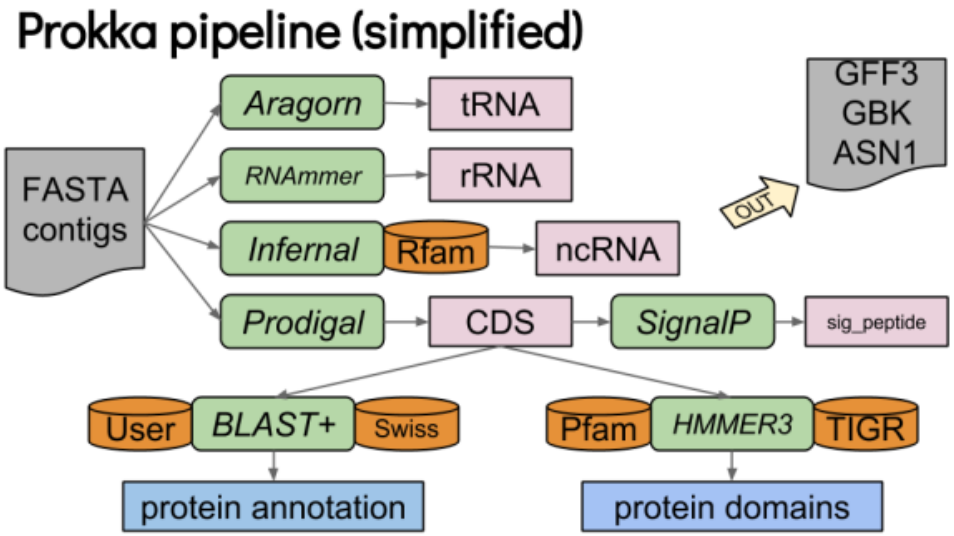
The James Hutton Institute

Java

Автоматическая аннотация генома: Rast, Prokka, Phaster и др.

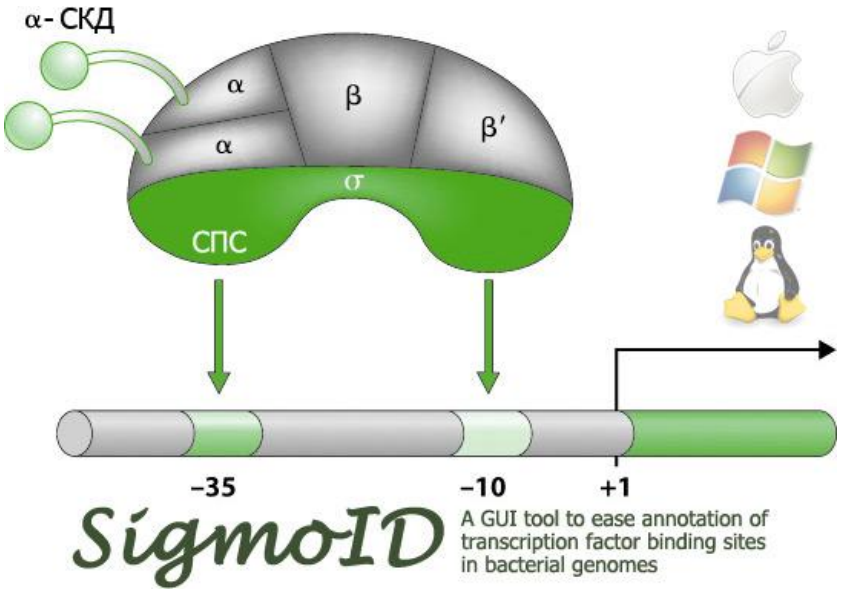


http://phaster.ca/

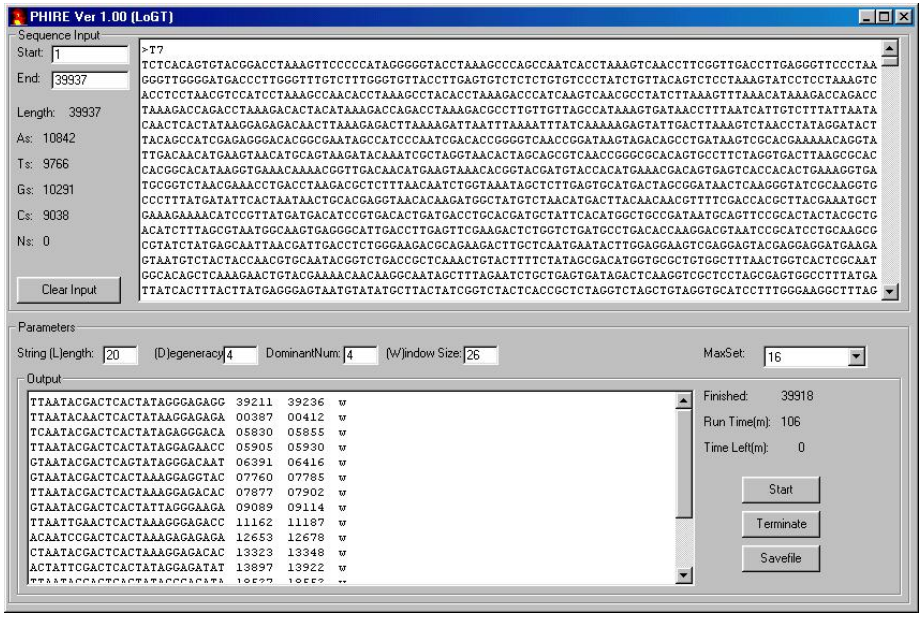


http://www.vicbioinformatics.com/software.prokka.shtml

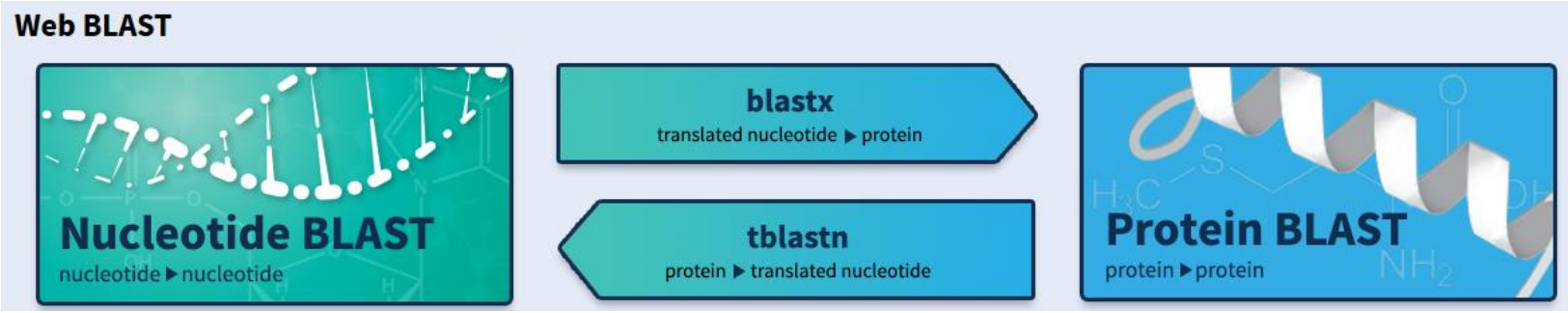
“Ручная” аннотация генома: Sigmoid, PHIRE, BLAST и др.



https://github.com/nikolaichik/SigmoID

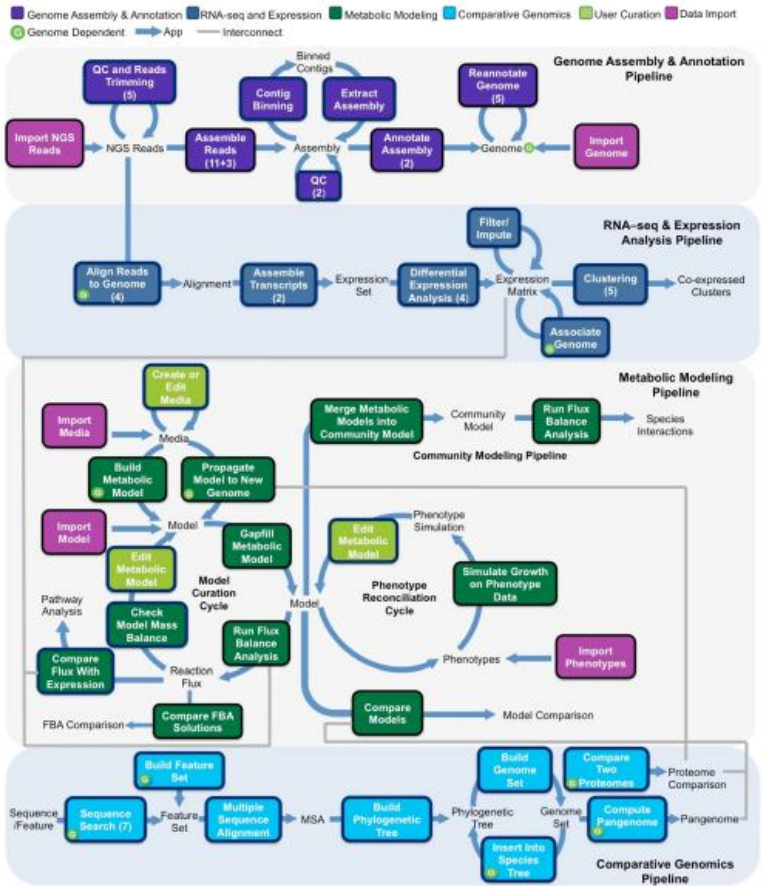


PHIRE (PHage *In silico* Regulatory Elements)
https://www.biw.kuleuven.be/logt/PHIRE.htm



Использование сторонних серверов с предустановленным ПО

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase



ORIONE
dataintensivebiology

--- Common tools ---
--- Tools By Topic ---
Microbiology
Variant calling
Metagenomics
NGS: RNA Analysis
--- Other tools ---
OMERO.biobank

Bioinformatics, Volume 30, Issue 13, 1 July 2014, Pages 1928–1929, https://doi.org/10.1093/bioinformatics/btu135

Работа с фрагментами 16S рДНК от метагеномных образцов

QIIME
- Quality control
- Chimera detection
- OTU clustering
- Pick representative sequences
- Assign taxonomy
- Taxonomy table

mothur
- Quality control
- Align sequences
- Clean alignment
- Pre-cluster sequences
- Chimera detection
- Classify sequences
- Remove non bacterial sequences
- Generate distance matrix
- OTU clustering
- Classify OTU
- Taxonomy table

MG-RAST
- Upload sequences and metadata
- Quality control
- RNA identification
- RNA clustering
- Assign taxonomy
- Taxonomy table

Б/Д для классификации метагеномных данных

NCBI taxonomy classification
NCBI RefSeq bacterial genomes
NCBI RefSeq viral genomes
NCBI RefSeq GRCh38 human genome
Kraken DB: MiniKraken 4Gb database
CLARK-l DB: RefSeq bacterial+viral genomes
CLARK-l DB: RefSeq viral genomes
DIAMOND DB: UniRef50
DIAMOND DB: UniRef90

ONE CODEX

https://www.onecodex.com/

http://taxonomer.iobio.io/

Работа с метагеномными данными

Нарочь
2016

Образец донных отложений

Получено прочтений: более 4 млн. × 2 (по 300 п.о.)

ОЗУ для сборки с помощью SPAdes: 72 Гб
Собрано контигов: 237 861
Общая длина контигов: 161 633 370 п.о.

Как с этим работать???

Бактерии	86%
Грибы	2,10%
Археи	1,80%
Вирусы	0,04%
Остальное	10%


Спасибо за внимание