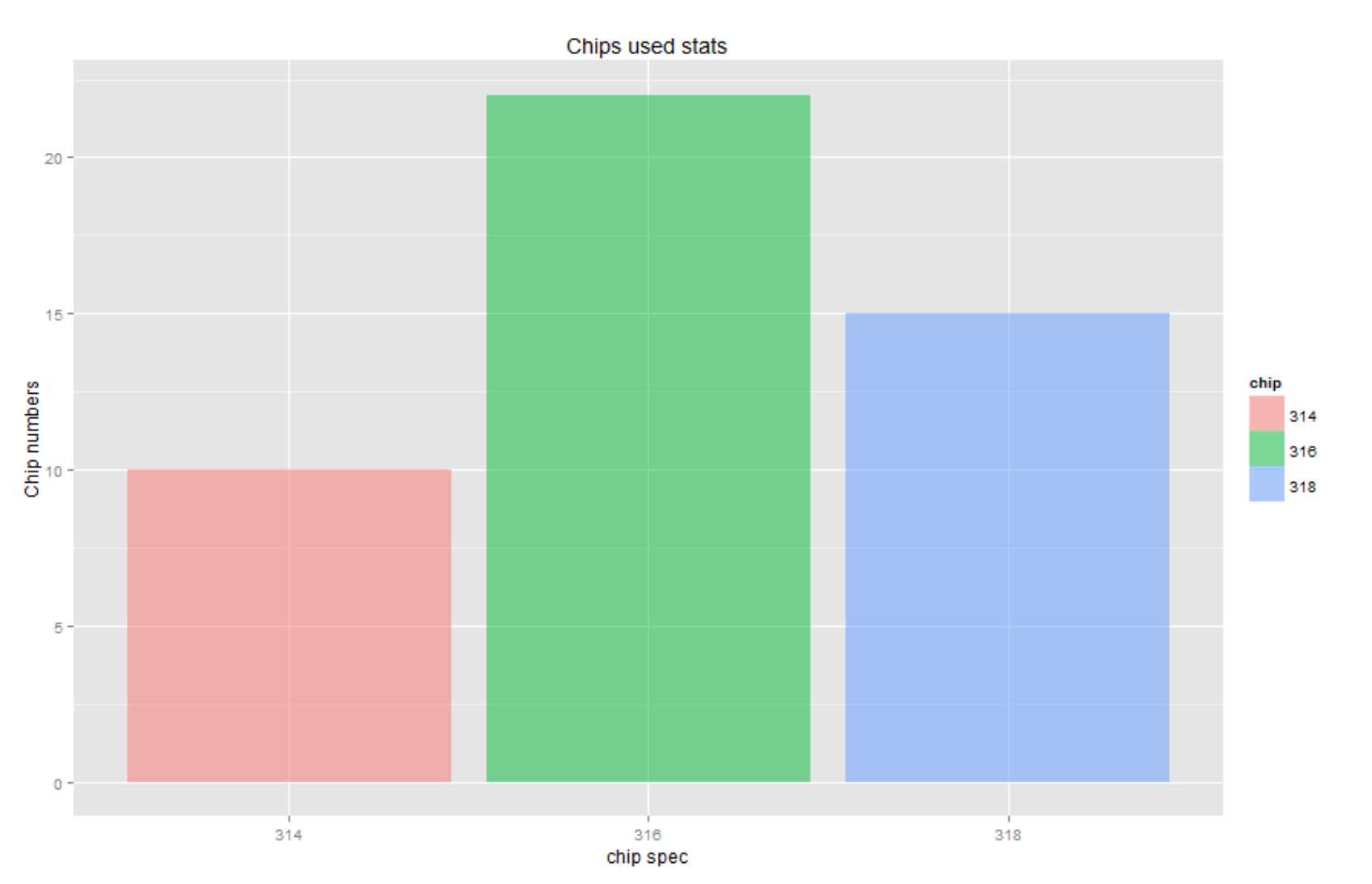
Наши

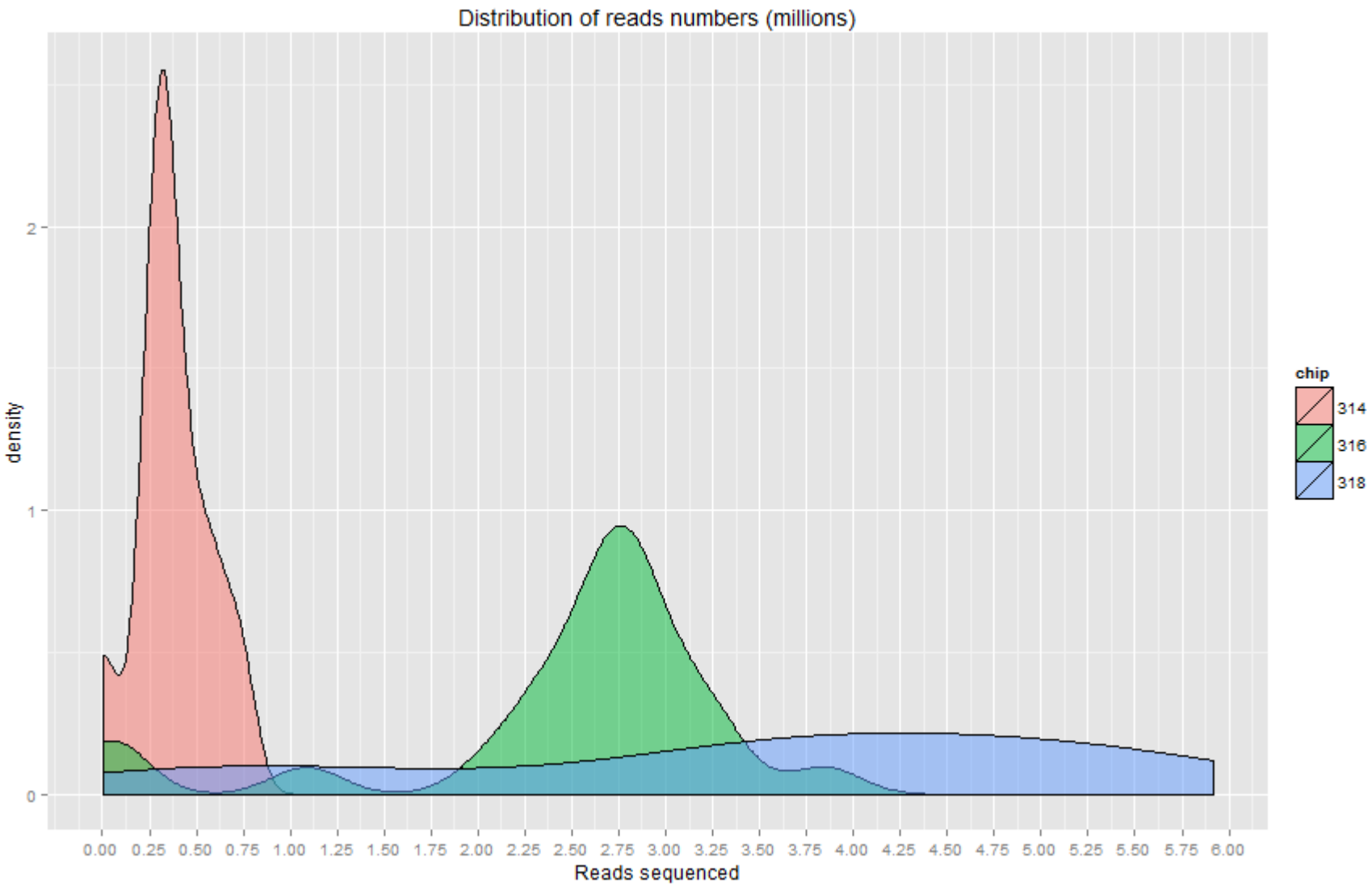
~~Мои~~ первые 100 проектов на IonTorrent - рассказ биоинформатика

Алексеев Дмитрий
зав. лаб. Биоинформатики НИИ ФХМ

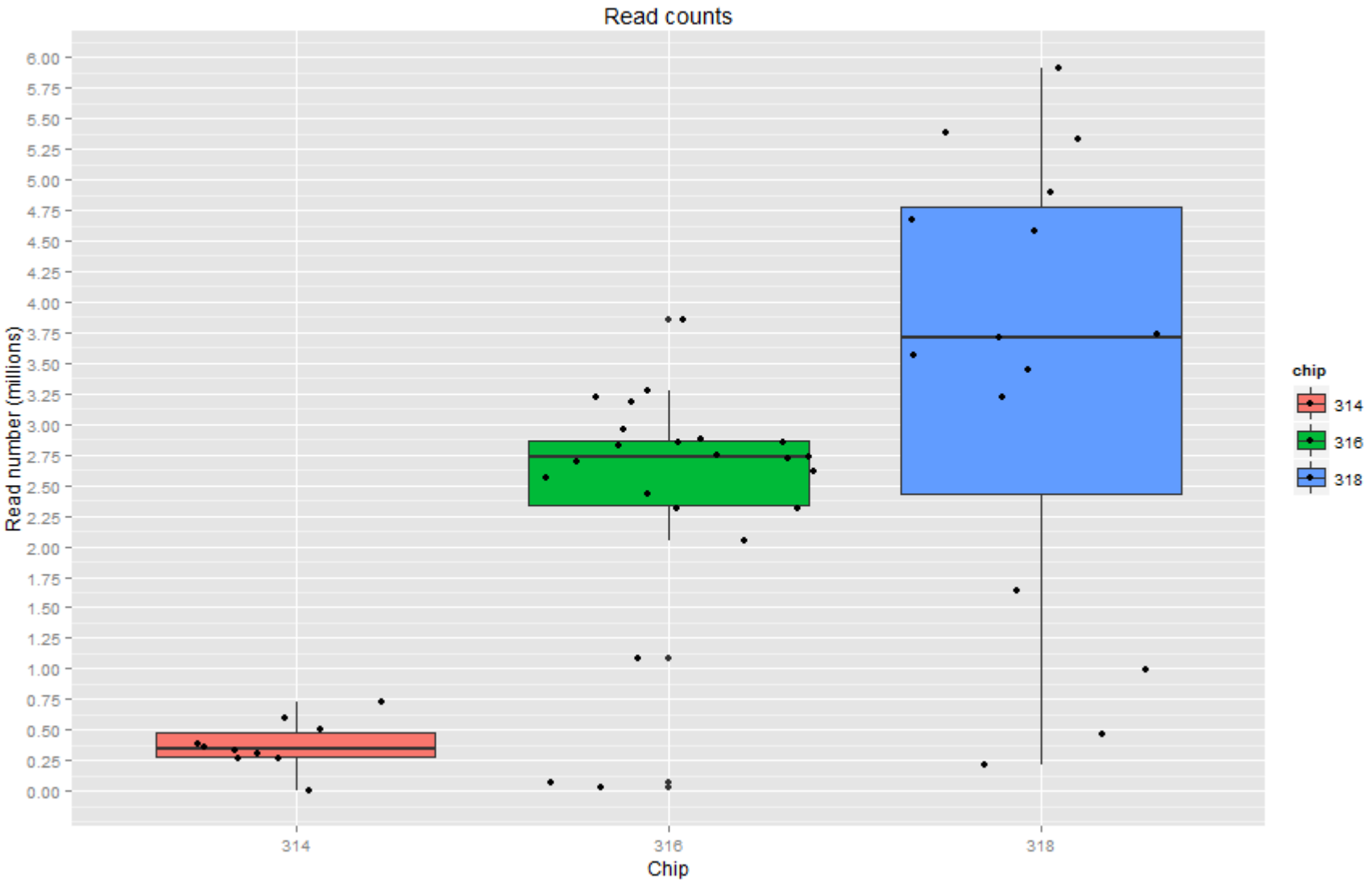
Статистика – 54 чипа



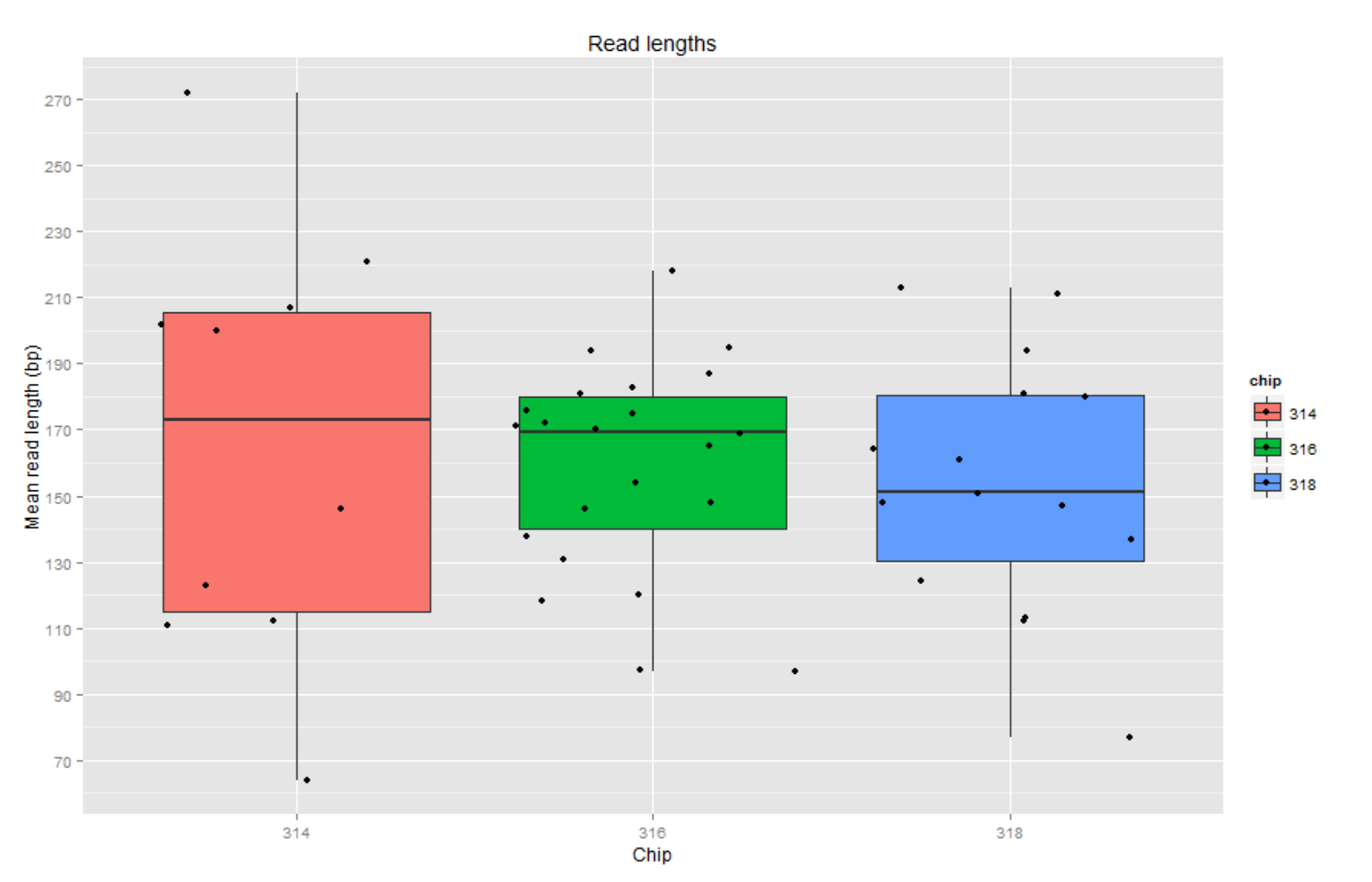
Кол-во ридов


Distribution of reads numbers (millions)
density
0
1
2
0.00 0.25 0.50 0.75 1.00 1.25 1.50 1.75 2.00 2.25 2.50 2.75 3.00 3.25 3.50 3.75 4.00 4.25 4.50 4.75 5.00 5.25 5.50 5.75 6.00
Reads sequenced
chip
314
316
318

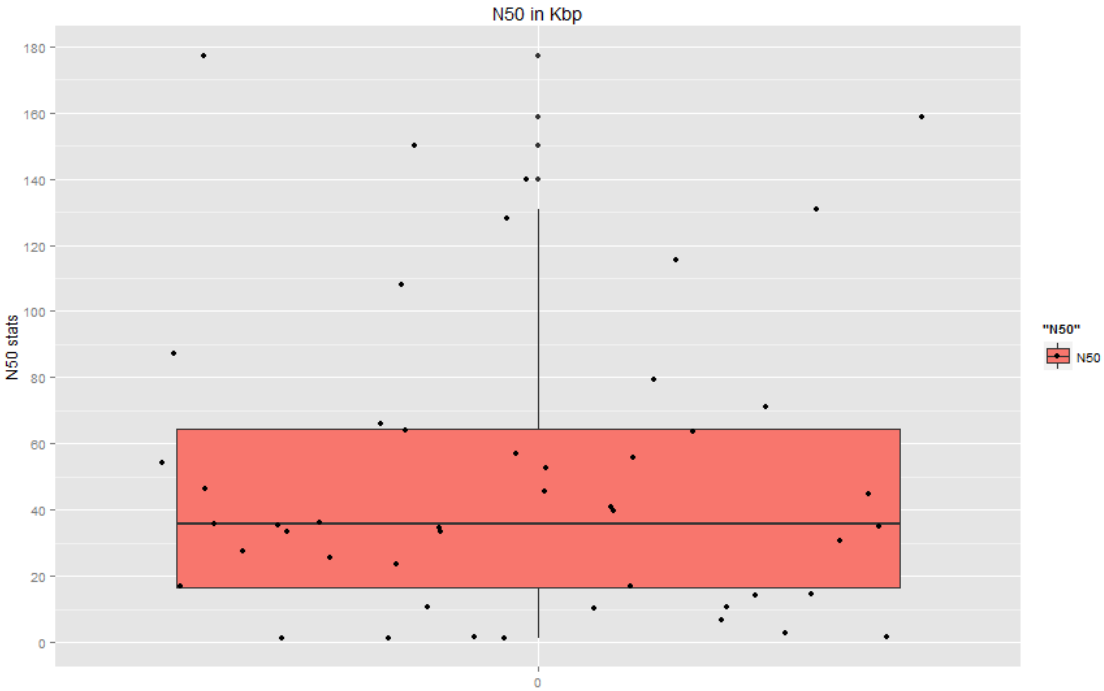
Кол-во ридов

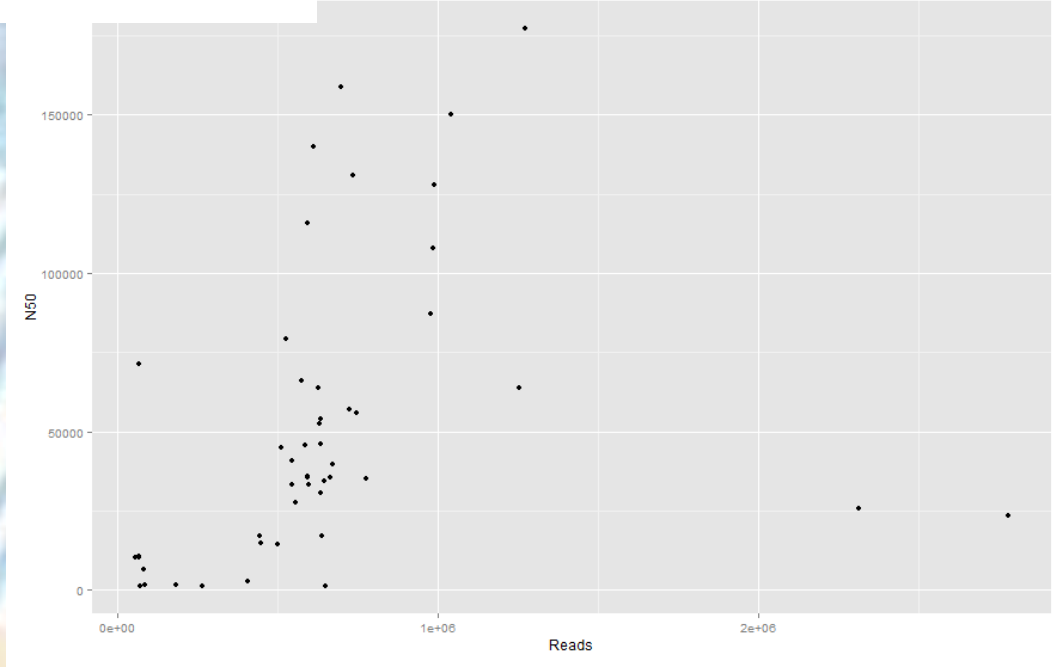

Read counts
Read number (millions)
6.00
5.75
5.50
5.25
5.00
4.75
4.50
4.25
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
314
316
318
Chip
chip
314
316
318

Длины ридов

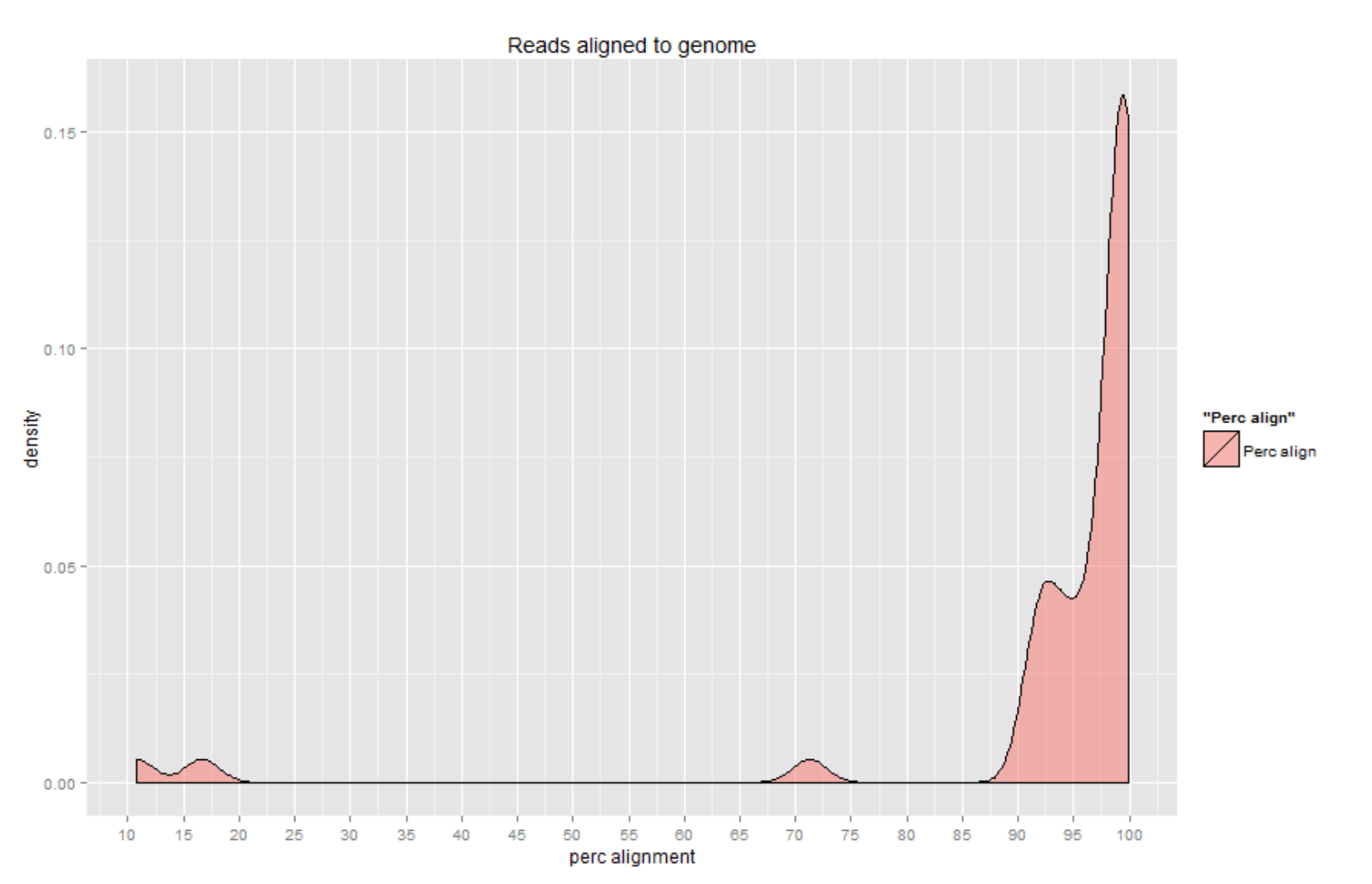

Read lengths
Mean read length (bp)
270
250
230
210
190
170
150
130
110
90
70
314
316
318
Chip
chip
314
316
318

N50


N50 in Kbp
N50 stats
180
160
140
120
100
80
60
40
20
0
0
"N50"
N50


N50
150000
100000
50000
0
0e+00
1e+06
2e+06
Reads

Риды картировавшиеся на сборку (mira)



Ross *et al. Genome Biology* 2013, **14**:R51
http://genomebiology.com/2013/14/5/R51

Genome Biology

RESEARCH Open Access

Characterizing and measuring bias in sequence data

Michael G Ross*, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J Lennon, Ryan Hegarty, Chad Nusbaum and David B Jaffe

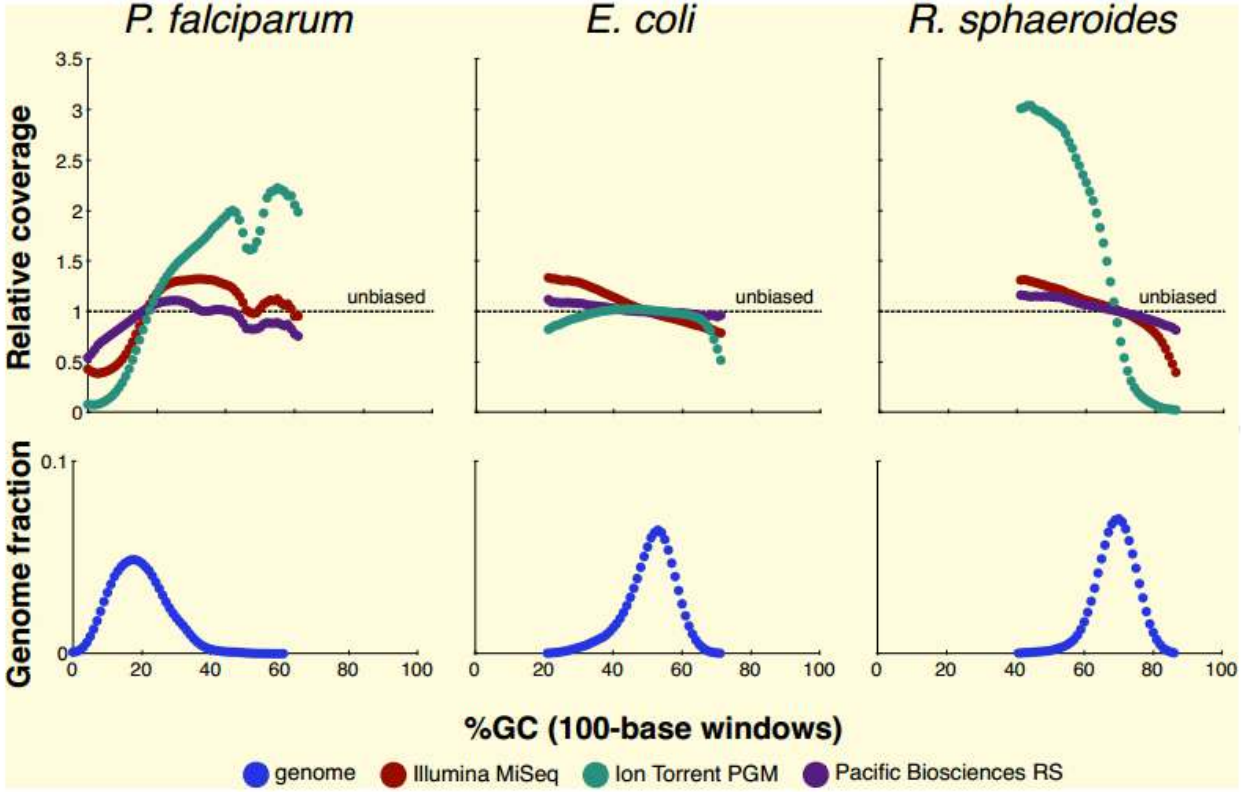


Figure 2 GC-bias plots for three microbial genomes. Top: plots showing the relative coverage GC-bias for Illumina MiSeq, Ion Torrent PGM, and Pacific Biosciences RS on the *P. falciparum* (19% GC), *E. coli* (51%), and *R. sphaeroides* (69%) genomes (Table 2, data sets 1 to 9). Unbiased coverage would be represented by a horizontal line at a relative coverage = 1 (black dashed line). Relative coverage is only plotted for GC percentages for which there are at least 1,000 100-base windows in the genome. Bottom: the GC composition distribution of each genome.

Парные фрагменты - SSPACE

- Позволяет удлинять и соединять (скаффолдинг) контиги на основе одной и более библиотек парных ридов

BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 4 2011, pages 578–579
doi:10.1093/bioinformatics/btq683

Genome analysis Advance Access publication December 12, 2010

Scaffolding pre-assembled contigs using SSPACE

Marten Boetzer[1,2], Christiaan V. Henkel[3], Hans J. Jansen[3], Derek Butler[1] and Walter Pirovano[1,*]

[1]BaseClear B.V., Einsteinweg 5, 2333 CC Leiden, [2]Leiden Institute for Advanced Computer Science, Leiden University, Niels Bohrweg 11, 2333 CA Leiden and [3]ZF-screens B.V., Niels Bohrweg 11, 2333 CA Leiden, The Netherlands

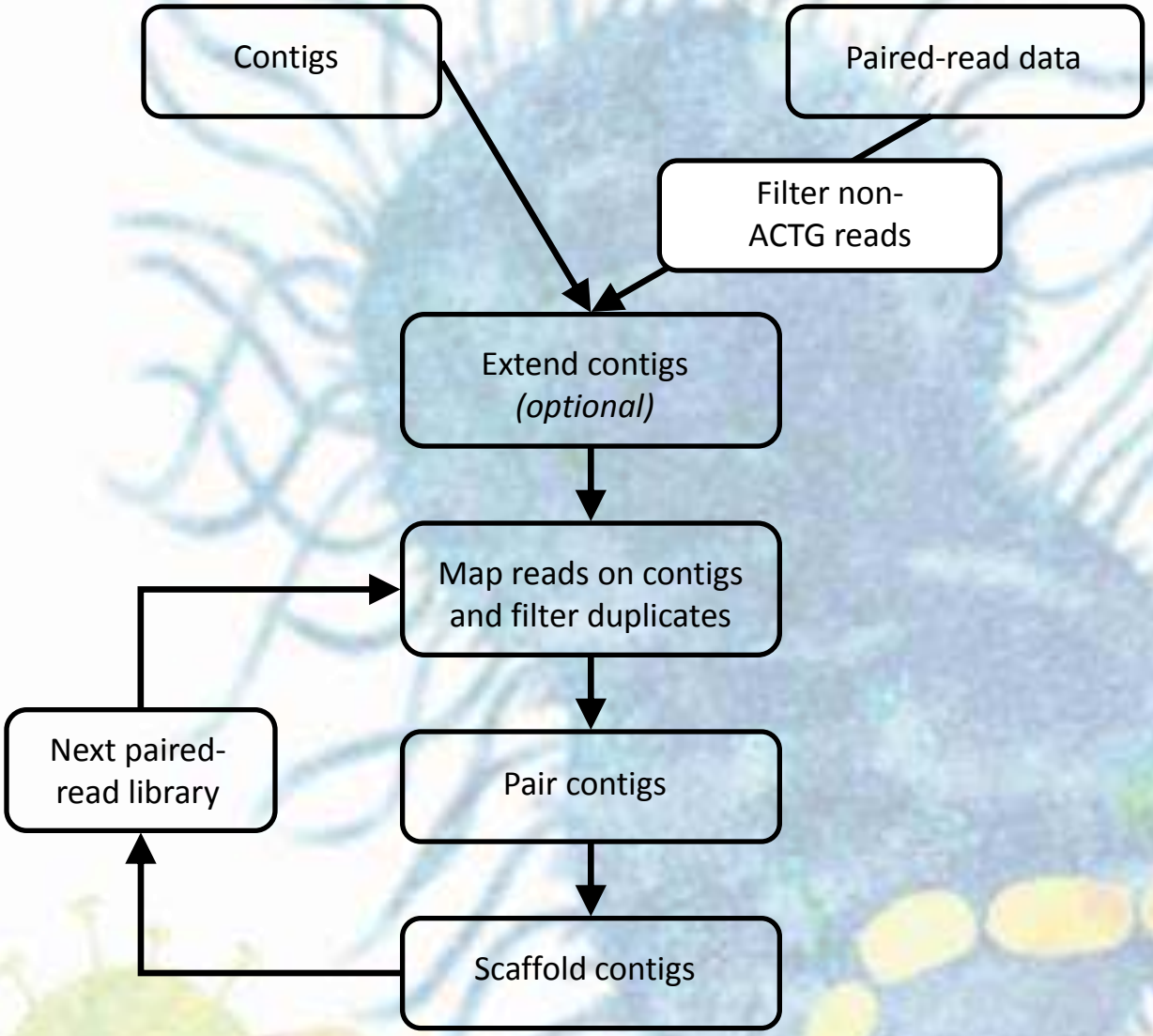
Associate Editor: John Quackenbush

ABSTRACT

Summary: *De novo* assembly tools play a main role in

To date only a few programs are able to scaffold pre-assembled contig sequences. A commonly used tool is Bambus (Pop *et al.*,

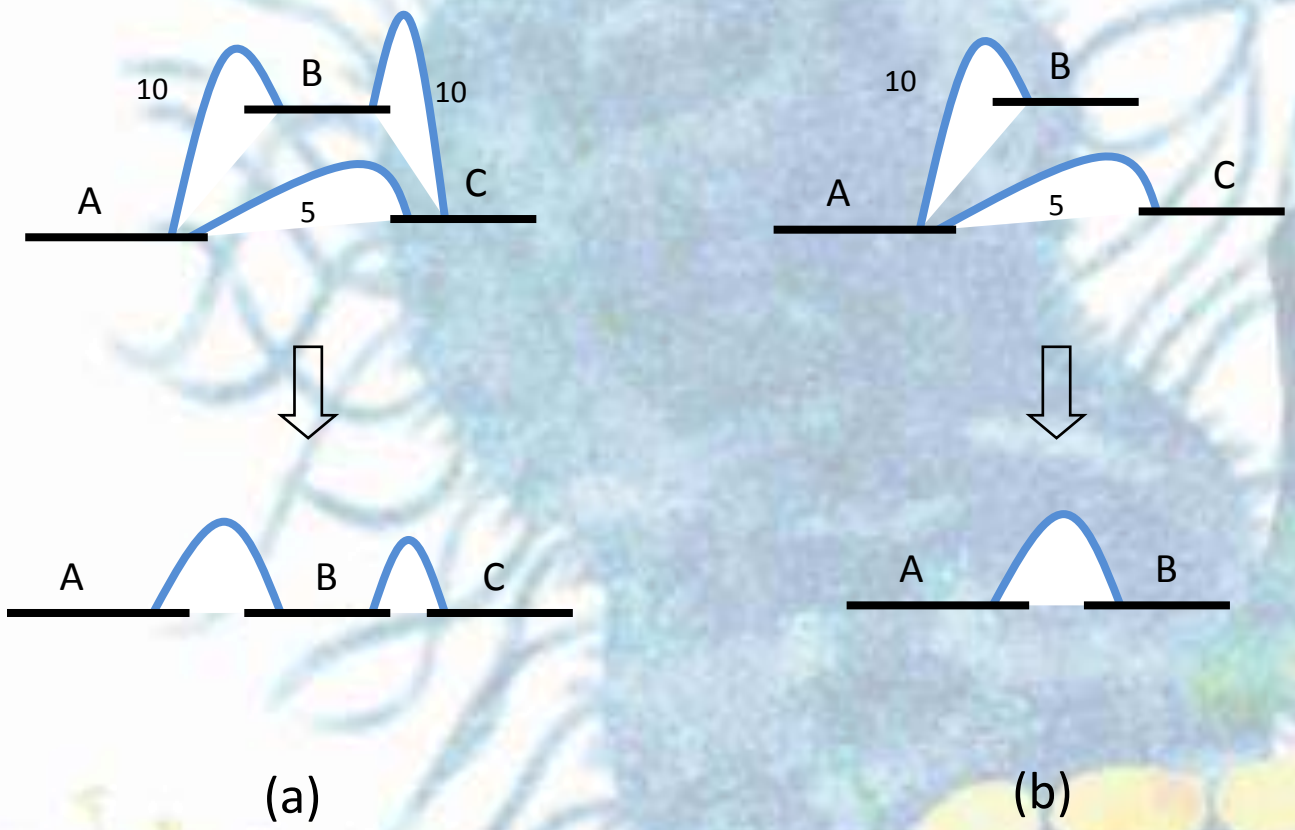
SSPACE algorithm



SSPACE algorithm

- Вероятные пары контигов определяются на основе картирования парных ридов на сборку
- Картирование считается успешным, если расстояние между ридами удовлетворяет указаному при запуске диапозону
- Скаффолдинг начинается с наиболее длинного контига. Если число связей между ним и другим контигом превышает порог, то они объединяются в скаффолд (итерационный процесс)
- В случае наличия альтернативных путей (A <-> B, A <-> C) вычисляется отношение числа связей между ними. Если оно меньше, чем порог (0.7) выбирается лучший путь. Если больше – скаффолдинг прерывается

Альтернативные пути скаффолдинга



(a) (b)

Представление скаффолдов: .dot файл, dotty

Пример статистики скаффолдинга

- READING READS r2b1: Total inserted pairs = 437,615
- READING READS r3b1: Total inserted pairs = 460,399
- Inserted contig file;
  - Total number of contigs = 135
    N50 = 41,390
- After scaffolding r2b1:
  - Total number of scaffolds = 51
  - N50 = 143,152
- After scaffolding r3b1:
  - Total number of scaffolds = 25
  - N50 = 321,140

Пайплайн для сбора статистики

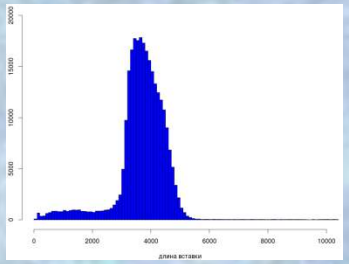
пары

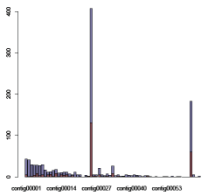
Риды 1 | Риды 2

FASTQ

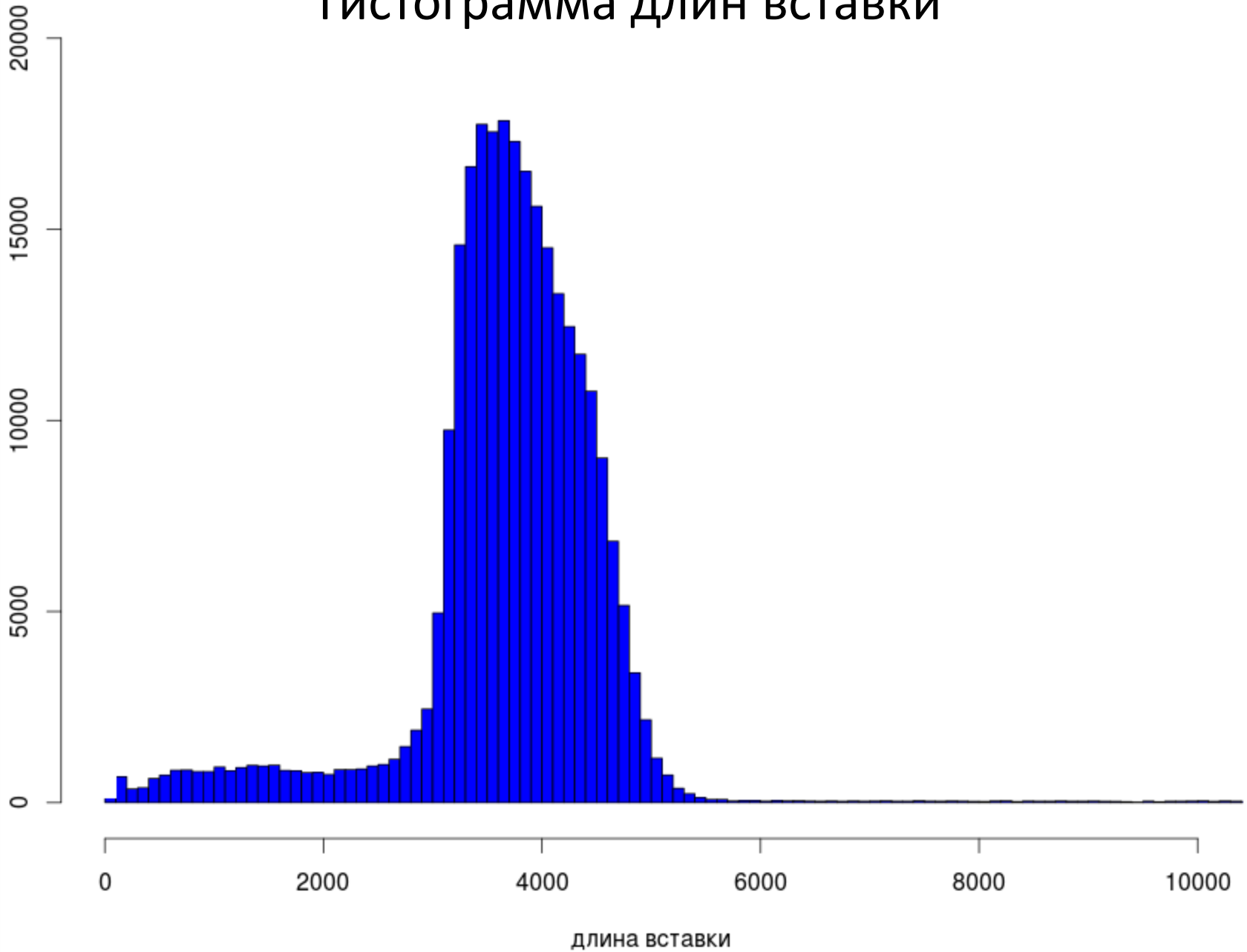
контиги → Bowtie 2, R

- Гистограммы длин вставок
- подробная информация о картировании пар

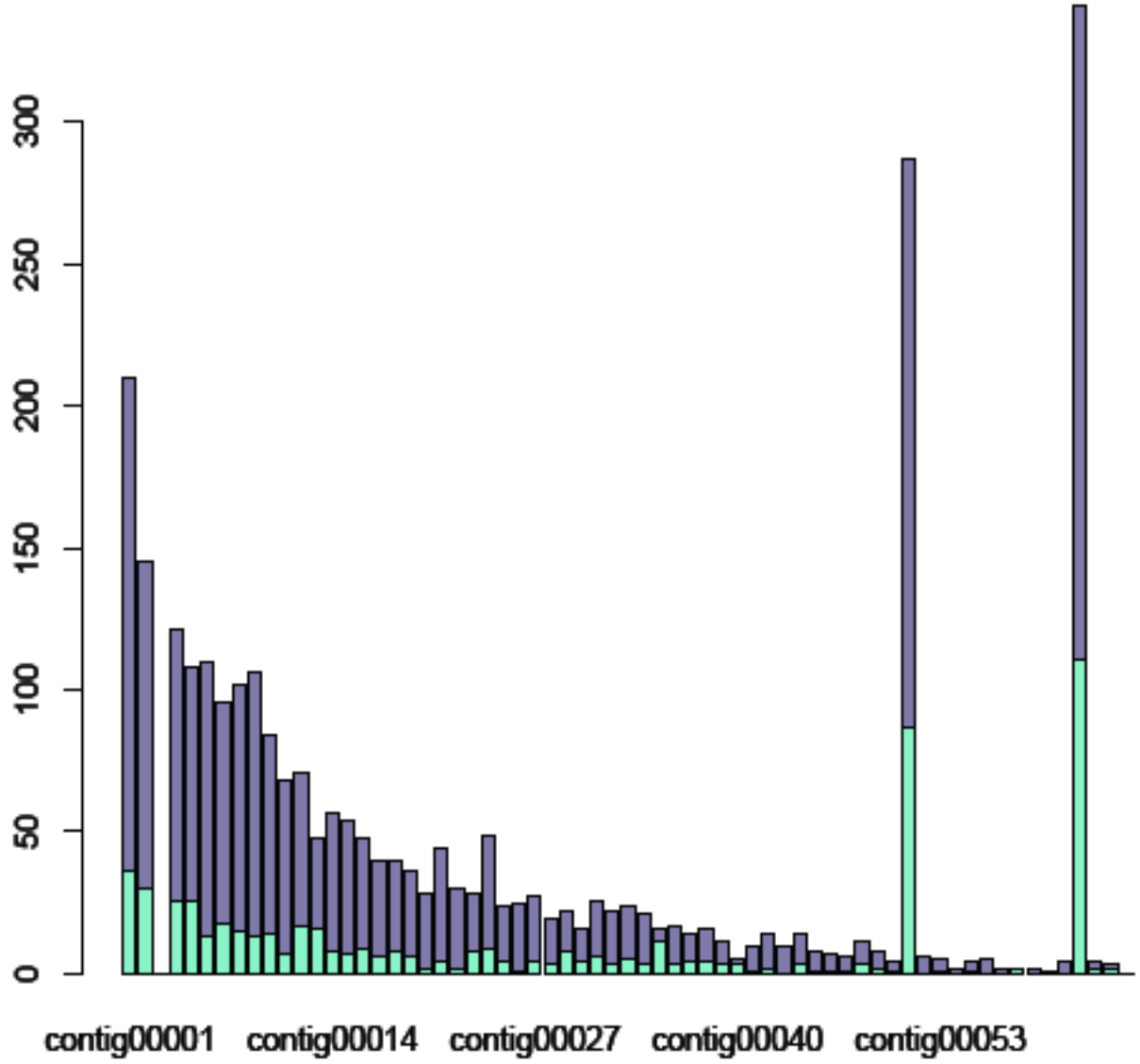




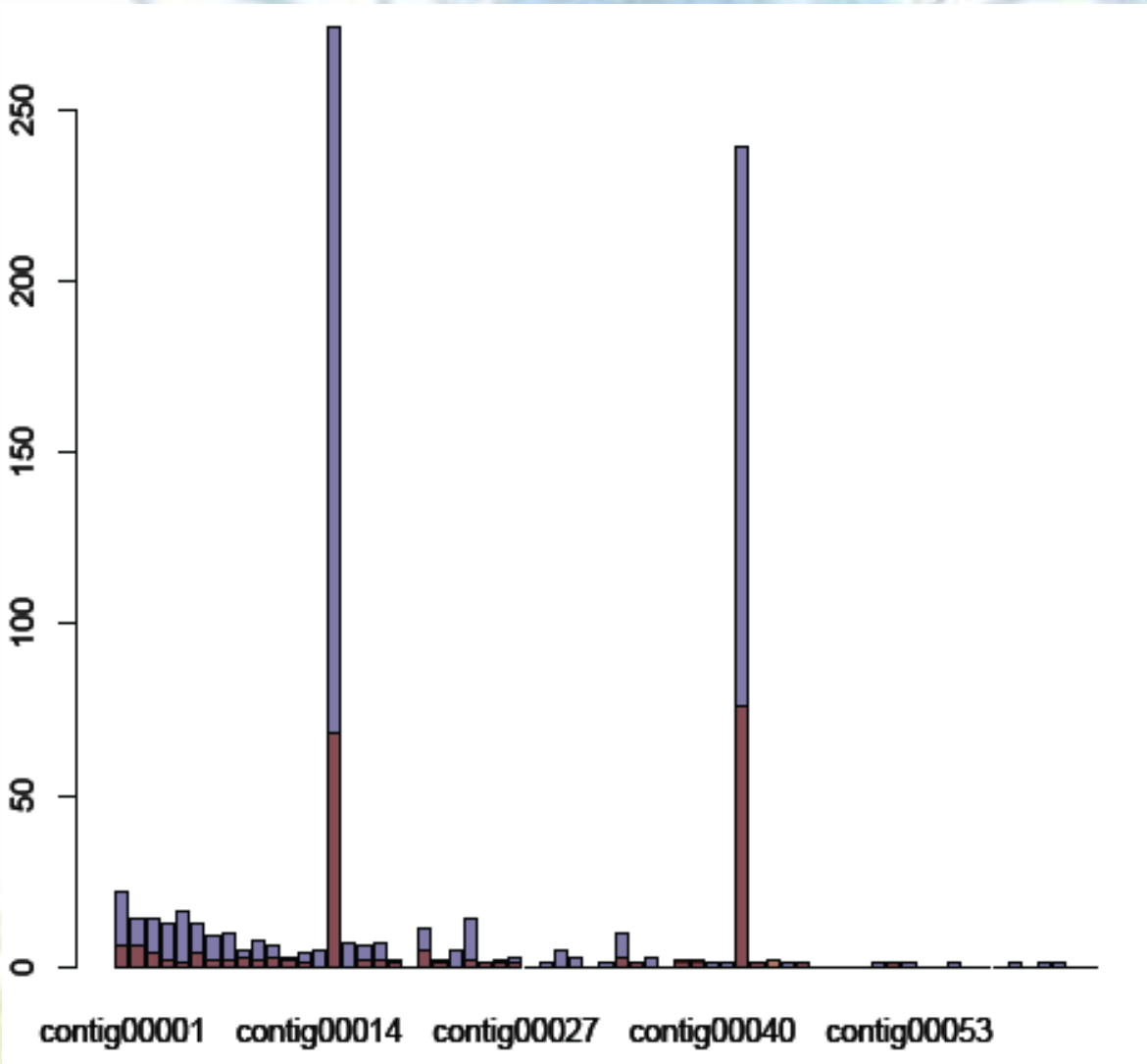
Гистограмма длин вставки



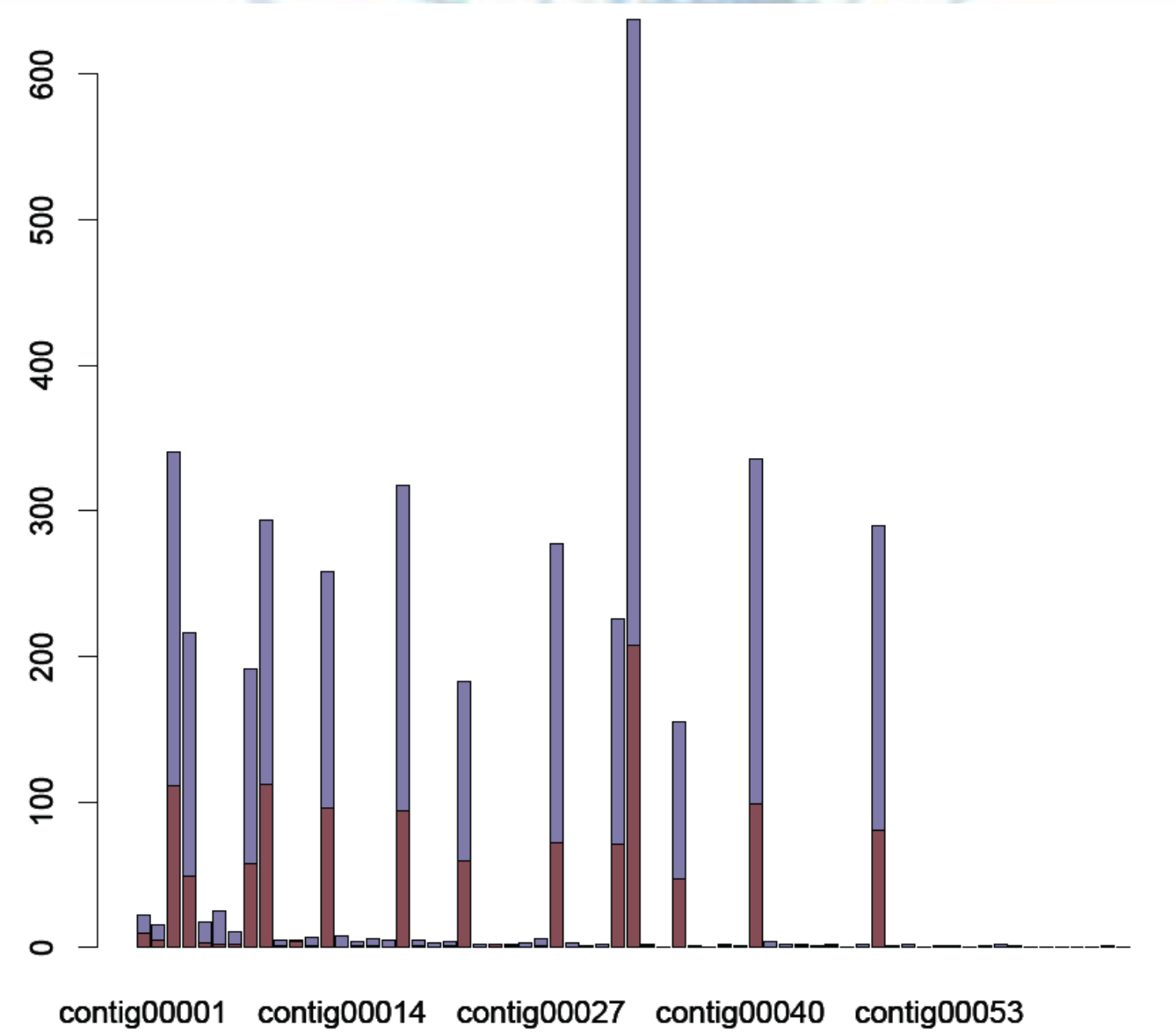
Contig 3: определить пары просто
Биологические повторы - хорошая воспроизводимость



Contig 37: определить пары просто



Contig 62: определить пары сложно



Скаффолдинг сборок 454 и Ion Torrent

- Pseudomonas *stutzeri,* 454+Ion Torrent
  - Total number of contigs = 91
  - N50 = 130,512
  - Total number of scaffolds = 5
  - N50 = 2,709,945
  - Max scaffold size: 2,709,945

- Pseudomonas *stutzeri ,* Ion Torrent
  - Total number of contigs = 369
  - N50 = 47,929
  - Total number of scaffolds =74
  - N50 = 377,361
  - Max scaffold size: 1,618,709

Скаффолдинг сборок 454 и Ion Torrent

- Enterococcus faecium, 454
  - Total number of contigs = 70
  - N50 = 95,885
  - Total number of scaffolds = 15
  - N50 = 441,245
  - Max scaffold size: 1,575,801

- Enterococcus *faecium,* Ion Torrent
  - Total number of contigs = 156
  - N50 = 100,114
  - Total number of scaffolds =50
  - N50 = 1,960,091
  - Max scaffold size: 1,960,091

M.tuberculosis
454 FLX
149 контигов
N50 = 101797

Библ парных фрагментов
(2-3 тыс + 4-5 тыс)
(под разными баркодами)

49 scaffolds
N50 = 494662

M.tuberculosis
454 FLX
165 контигов
N50 = 88489

Библ парных фрагментов
(2-3 тыс + 4-5 тыс)
(под разными баркодами)

35 scaffolds
N50 = 738326

Тема:

Расшифровка механизмов формирования детерминант патогенности стрептококками группы *viridans* на основании сравнения полноразмерных нуклеотидных последовательностей их геномов

Проблема:

Прояснение молекулярных механизмов формирования патогенного потенциала клинически значимыми микроорганизмами рода *Streptococcus*, в частности *S. pneumoniae* (пневмококка), вызывающего у человека пневмонию, сепсис, менингит

Задачи:

- Геномное секвенирование клинических изолятов зеленящих (*viridans*) стрептококков, отличных по видовой принадлежности (*S. mitis* и *S. pneumoniae*) и клиническим проявлениям со стороны пациента (инвазивные и неинвазивные *S. pneumoniae*)
- Инвентаризация и систематизация известных клеточных компонентов и структур (ферментов, протеинов клеточной стенки, капсулы, пилей и др.), обычно рассматриваемых в качестве факторов патогенности и вирулентности, на основании анализа полноразмерных нуклеотидных последовательностей геномов изолятов зеленящих стрептококков подгруппы mitis.
- Выявление новых факторов, которые могут играть роль в адаптации и инвазии микроорганизмов рода *Streptococcus* в макроорганизме, методами сравнительной геномики.
- Выявление структурных и функциональных особенностей генома, определяющих выбор между паразитизмом и комменсализмом на примере сравнения геномов *S. pneumoniae* и *S. mitis*.

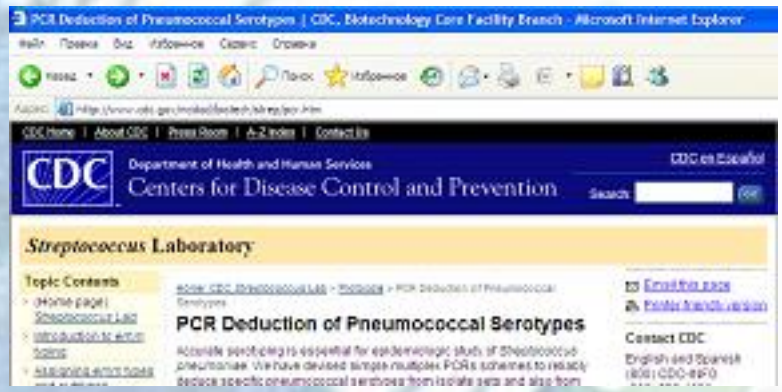
Sequencing Platform	Sequencing Depth	Assembly Method	Large Contigs
Ion Torrent PGM™	35x	GS De Novo Assembler v. 2.8	115
Ion Torrent PGM™	34x	GS De Novo Assembler v. 2.8	103
Ion Torrent PGM™	40x	GS De Novo Assembler v. 2.8	107
Ion Torrent PGM™	31x	GS De Novo Assembler v. 2.8	124
Ion Torrent PGM™	20x	GS De Novo Assembler v. 2.8	226
Ion Torrent PGM™	26x	GS De Novo Assembler v. 2.8	180
Ion Torrent PGM™	19x	GS De Novo Assembler v. 2.8	208
Ion Torrent PGM™	55x	GS De Novo Assembler v. 2.8	88
Ion Torrent PGM™	75x	GS De Novo Assembler v. 2.8	60
Ion Torrent PGM™	48x	GS De Novo Assembler v. 2.8	110
Ion Torrent PGM™	74x	GS De Novo Assembler v. 2.8	37
Ion Torrent PGM™	53x	GS De Novo Assembler v. 2.8	67
Ion Torrent PGM™	77x	GS De Novo Assembler v. 2.8	34

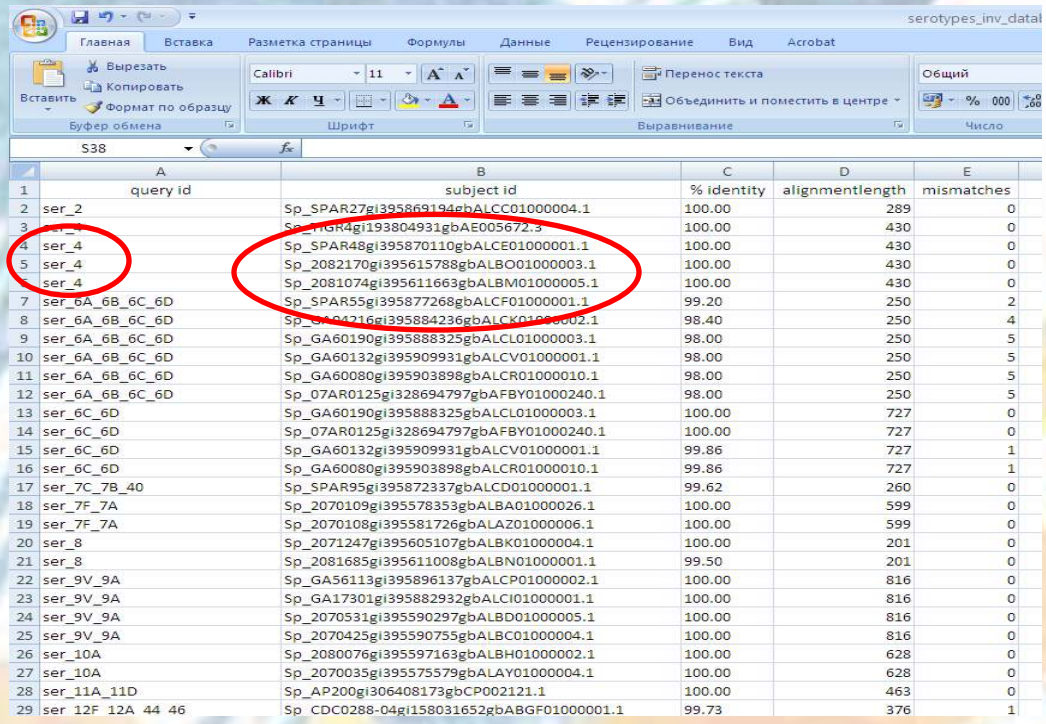
К настоящему моменту прочитаны геномы 8 штаммов пневмококков (4 инвазивных, 4 неинвазивных), 5 штаммов *S. mitis* и одного штамма псевдопневмококка. Все геномы можно найти в базе данных NCBI

Вид	Штамм	Accession в NCBI
Streptococcus pneumoniae	357	PRJNA201317
Streptococcus pneumoniae	2009	PRJNA201318
Streptococcus pneumoniae	801	PRJNA201319
Streptococcus pneumoniae	845	PRJNA201320
Streptococcus pneumoniae	1488	PRJNA201321
Streptococcus pneumoniae	1542	PRJNA201322
Streptococcus pneumoniae	3051	PRJNA201323
Streptococcus pneumoniae	1779	PRJNA206047
Streptococcus mitis	11/5	PRJNA201324
Streptococcus mitis	13/39	PRJNA201325
Streptococcus mitis	17/34	PRJNA206048
Streptococcus mitis	18/56	PRJNA206049
Streptococcus mitis	29/42	PRJNA206050
Streptococcus pseudopneumoniae	G42	в процессе оформления

Первичная характеристика штаммов по геномным данным: «виртуальное серотипирование» *Streptococcus pneumoniae*

Рекомендации CDC по определению 40 серотипов с помощью сероспецифичной мультипраймерной ПЦР

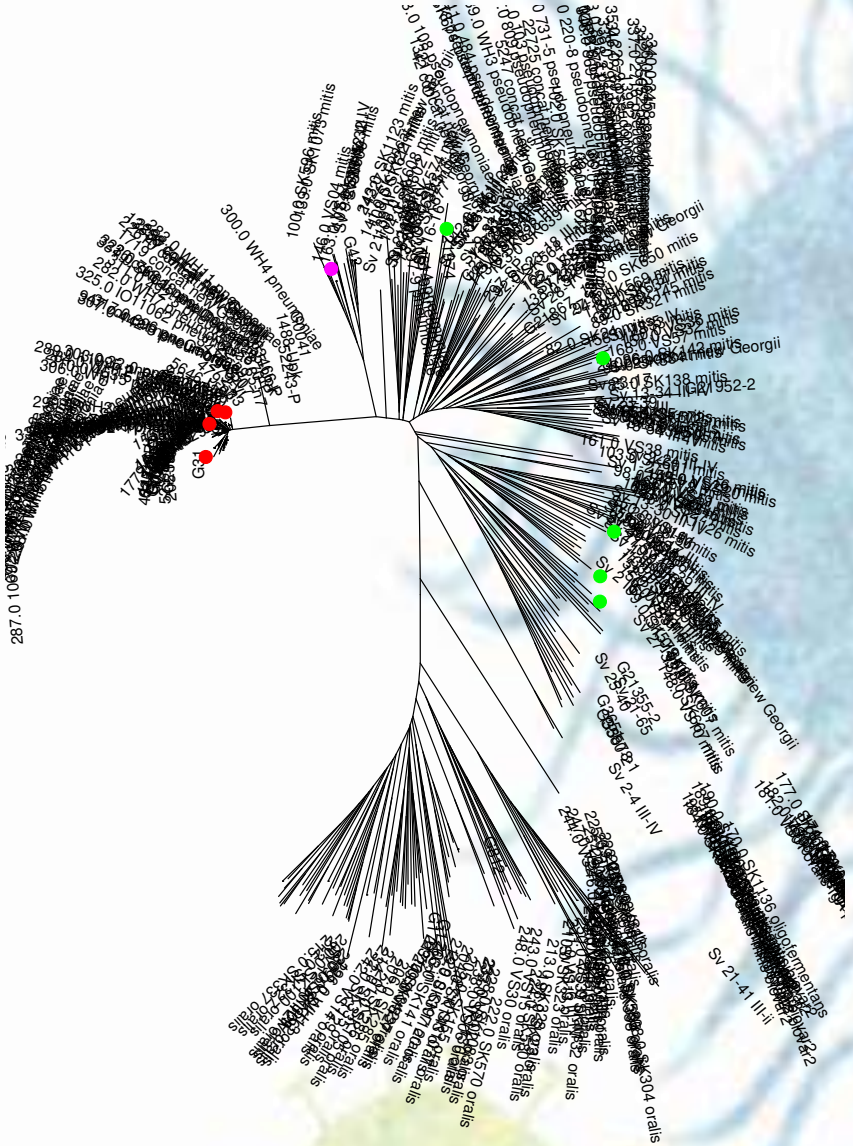




	A	B	C	D	E
1	query id	subject id	% identity	alignmentlength	mismatches
2	ser_2	Sp_SPAR27gi395869194gbALCC01000004.1	100.00	289	0
3	[illegible]	Sp_HGR4gi193804931gbAE005672.3	100.00	430	0
4	ser_4	Sp_SPAR48gi395870110gbALCE01000001.1	100.00	430	0
5	ser_4	Sp_2082170gi395615788gbALBO01000003.1	100.00	430	0
	ser_4	Sp_2081074gi395611663gbALBM01000005.1	100.00	430	0
7	ser_6A_6B_6C_6D	Sp_SPAR55gi395877268gbALCF01000001.1	99.20	250	2
8	ser_6A_6B_6C_6D	Sp_GA94216gi395884236gbALCK01000002.1	98.40	250	4
9	ser_6A_6B_6C_6D	Sp_GA60190gi395888325gbALCL01000003.1	98.00	250	5
10	ser_6A_6B_6C_6D	Sp_GA60132gi395909931gbALCV01000001.1	98.00	250	5
11	ser_6A_6B_6C_6D	Sp_GA60080gi395903898gbALCR01000010.1	98.00	250	5
12	ser_6A_6B_6C_6D	Sp_07AR0125gi328694797gbAFBY01000240.1	98.00	250	5
13	ser_6C_6D	Sp_GA60190gi395888325gbALCL01000003.1	100.00	727	0
14	ser_6C_6D	Sp_07AR0125gi328694797gbAFBY01000240.1	100.00	727	0
15	ser_6C_6D	Sp_GA60132gi395909931gbALCV01000001.1	99.86	727	1
16	ser_6C_6D	Sp_GA60080gi395903898gbALCR01000010.1	99.86	727	1
17	ser_7C_7B_40	Sp_SPAR95gi395872337gbALCD01000001.1	99.62	260	0
18	ser_7F_7A	Sp_2070109gi395578353gbALBA01000026.1	100.00	599	0
19	ser_7F_7A	Sp_2070108gi395581726gbALAZ01000006.1	100.00	599	0
20	ser_8	Sp_2071247gi395605107gbALBK01000004.1	100.00	201	0
21	ser_8	Sp_2081685gi395611008gbALBN01000001.1	99.50	201	0
22	ser_9V_9A	Sp_GA56113gi395896137gbALCP01000002.1	100.00	816	0
23	ser_9V_9A	Sp_GA17301gi395882932gbALCI01000001.1	100.00	816	0
24	ser_9V_9A	Sp_2070531gi395590297gbALBD01000005.1	100.00	816	0
25	ser_9V_9A	Sp_2070425gi395590755gbALBC01000004.1	100.00	816	0
26	ser_10A	Sp_2080076gi395597163gbALBH01000002.1	100.00	628	0
27	ser_10A	Sp_2070035gi395575579gbALAY01000004.1	100.00	628	0
28	ser_11A_11D	Sp_AP200gi306408173gbCP002121.1	100.00	463	0
29	ser_12F_12A_44_46	Sp_CDC0288-04gi158031652gbABGF01000001.1	99.73	376	1

- **8 наших геномов *S. pneumoniae***
- **64 генома из базы данных NCBI**

Первичная характеристика штаммов по геномным данным: **филогенетика**



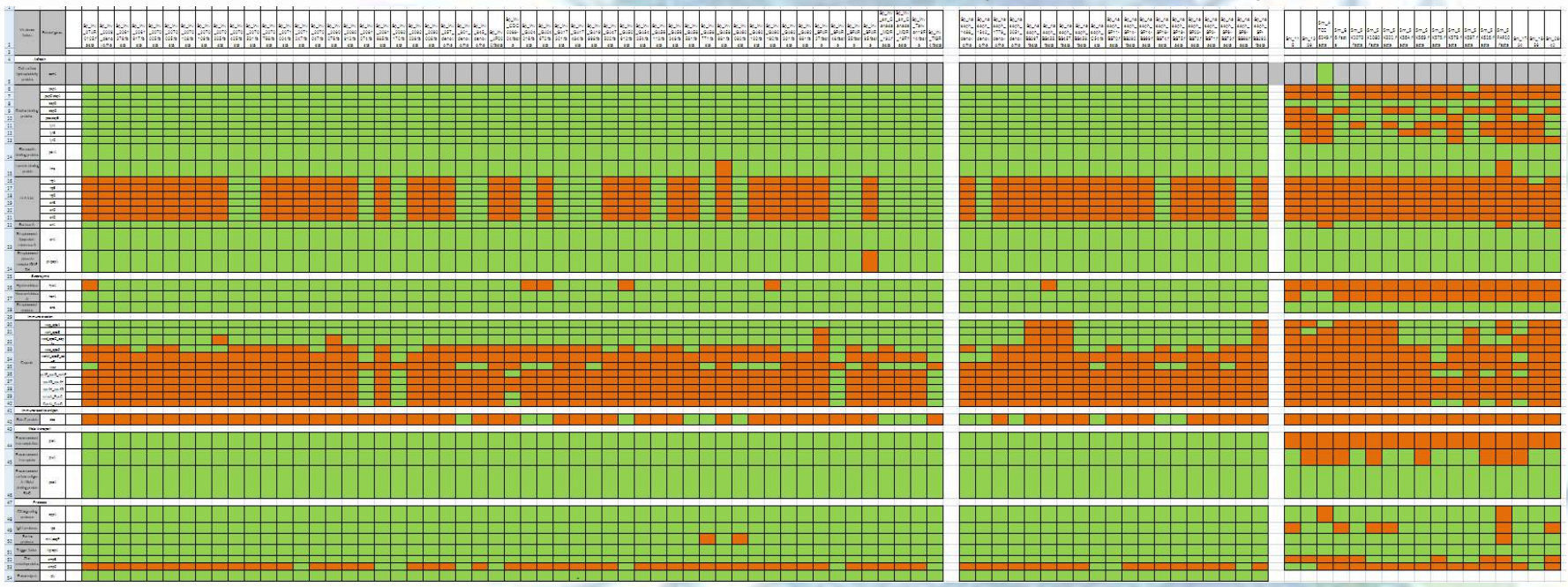
- 427 штаммов из базы данных MLSA S. viridans
- 64 генома из базы данных NCBI
- 14 геномов из нашей коллекции

Сравнительный анализ факторов вирулентности штаммов подгруппы mitis группы зеленящих стрептококков

S. pneumoniae (инвазивные)

S. pneumoniae (неинвазивные)

S. mitis



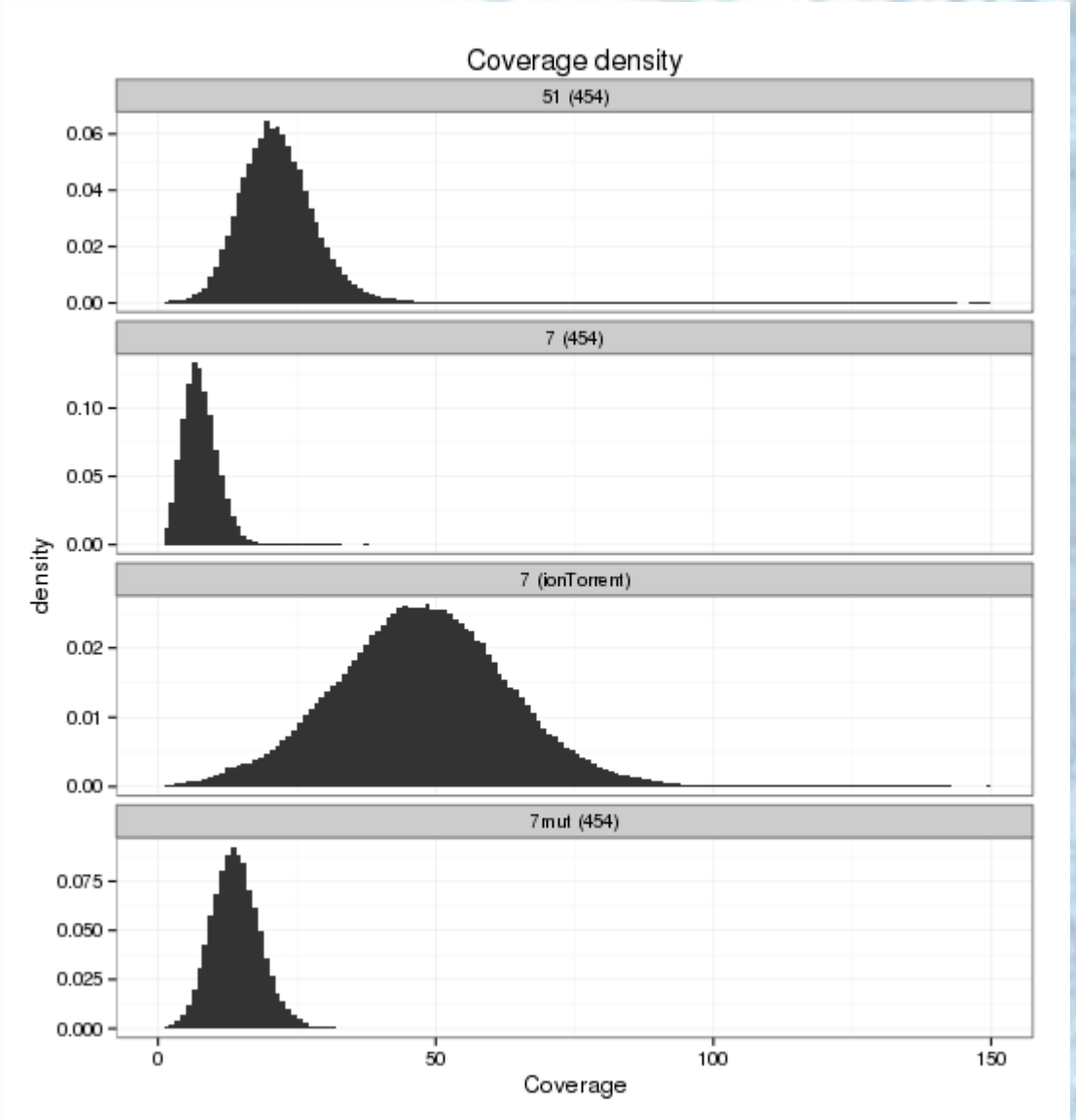
Структура капсульного оперона представителей группы зеленящих стрептококков

Капсула – один из основных факторов вирулентности пневмококка

Neisseria gonorrhoeae
Клинические изоляты

	PEN	TET	CIPR	SPEC	CRO	Azi
NG3						
NG3mut						
NG19						
NG51						

Neisseria gonorrhoeae

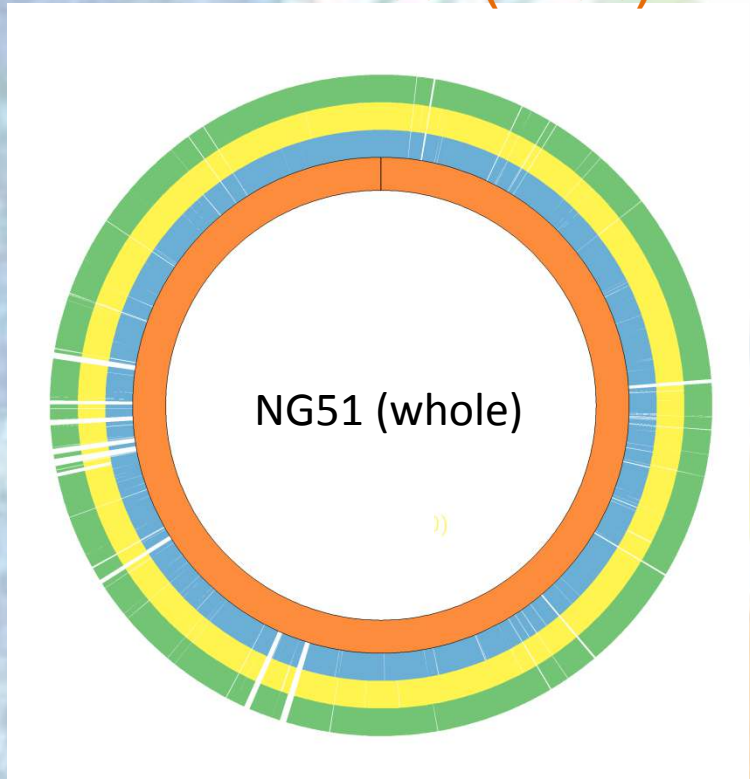


NG7mut (454)

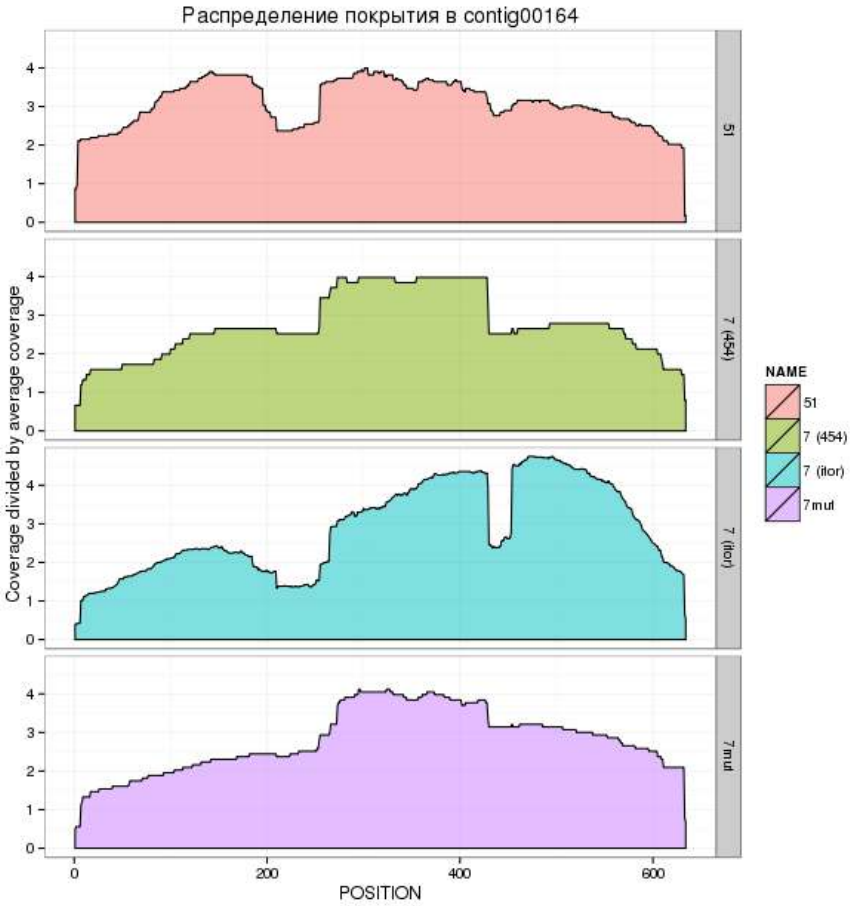
NG7 (ionTorrent)

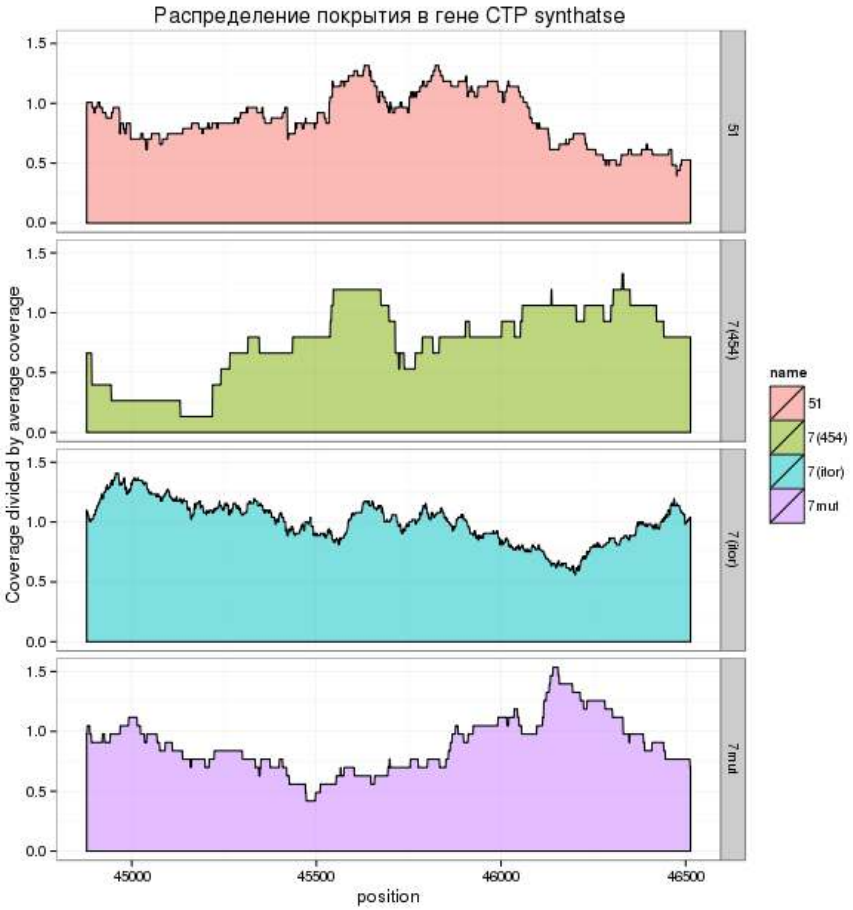
NG7 (454)

NG51 (whole)



Neisseria gonorrhoeae





Neisseria gonorrhoeae Внутригеномные повторы

- DNA uptake sequences 5’-ATGCCGTCTGAA-3’

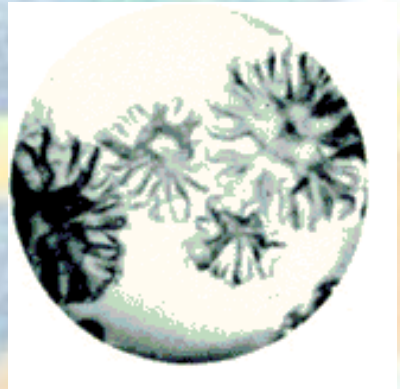
Штамм	Количество повторов	
	DUS10	**DUS12**
N. gonorrhoeae NG3	1849	1419
N. gonorrhoeae NG3mut	1905	1451
N. gonorrhoeae NG19	1930	1479
N. gonorrhoeae NG51	1913	1462
N. gonorrhoeae FA1090	1965	1522
N. gonorrhoeae NCCP11945	1966	1520
N. gonorrhoeae 35/02	1915	1475

Сборка de-novo и анализ генома *Pseudomonas stutzeri KOS6*

Объект исследования: *Pseudomonas stutzeri*

- Грамотрицательные бактерии, которые широко распространены в природе
- Впервые были извлечены из спинномозговой жидкости. Условно-патогенные микроорганизм
- Обитают в очень разнообразных средах: почва, толща морской воды и осадки, сточные воды
- Обладают очень широким спектром метаболических функций
- Модельный организм для изучения денитрификации
- Способности организма усваивать широкий спектр субстратов может сделать его применимым в биоремедиации и очистки сточных вод.
- Высокая частота естественной трансформации
- Использовались итальянским микробиологом Giancarlo Ranalli для реставрации старинных фресок

Pseudomonas stutzeri KOS6



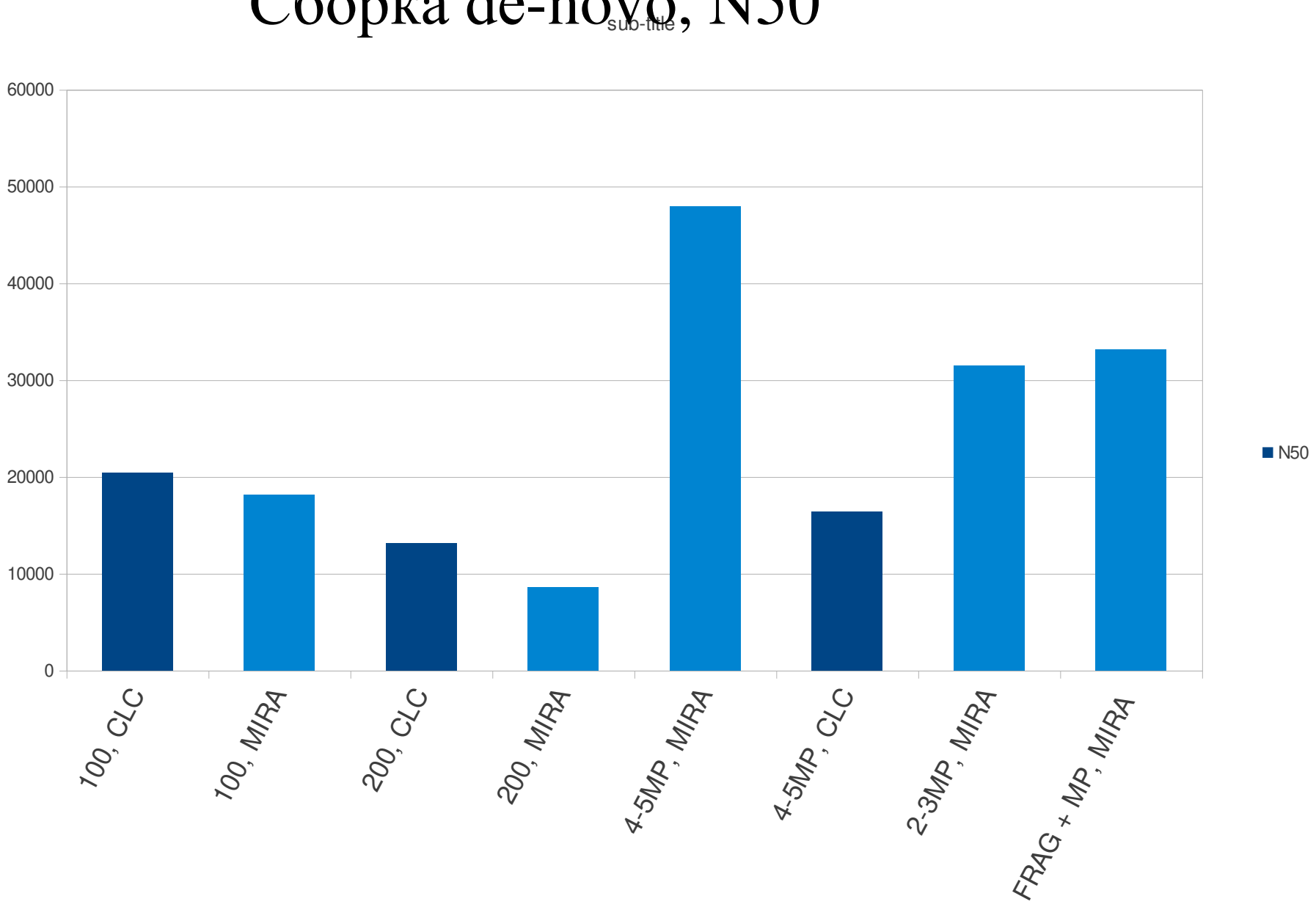
Результаты секвенирования

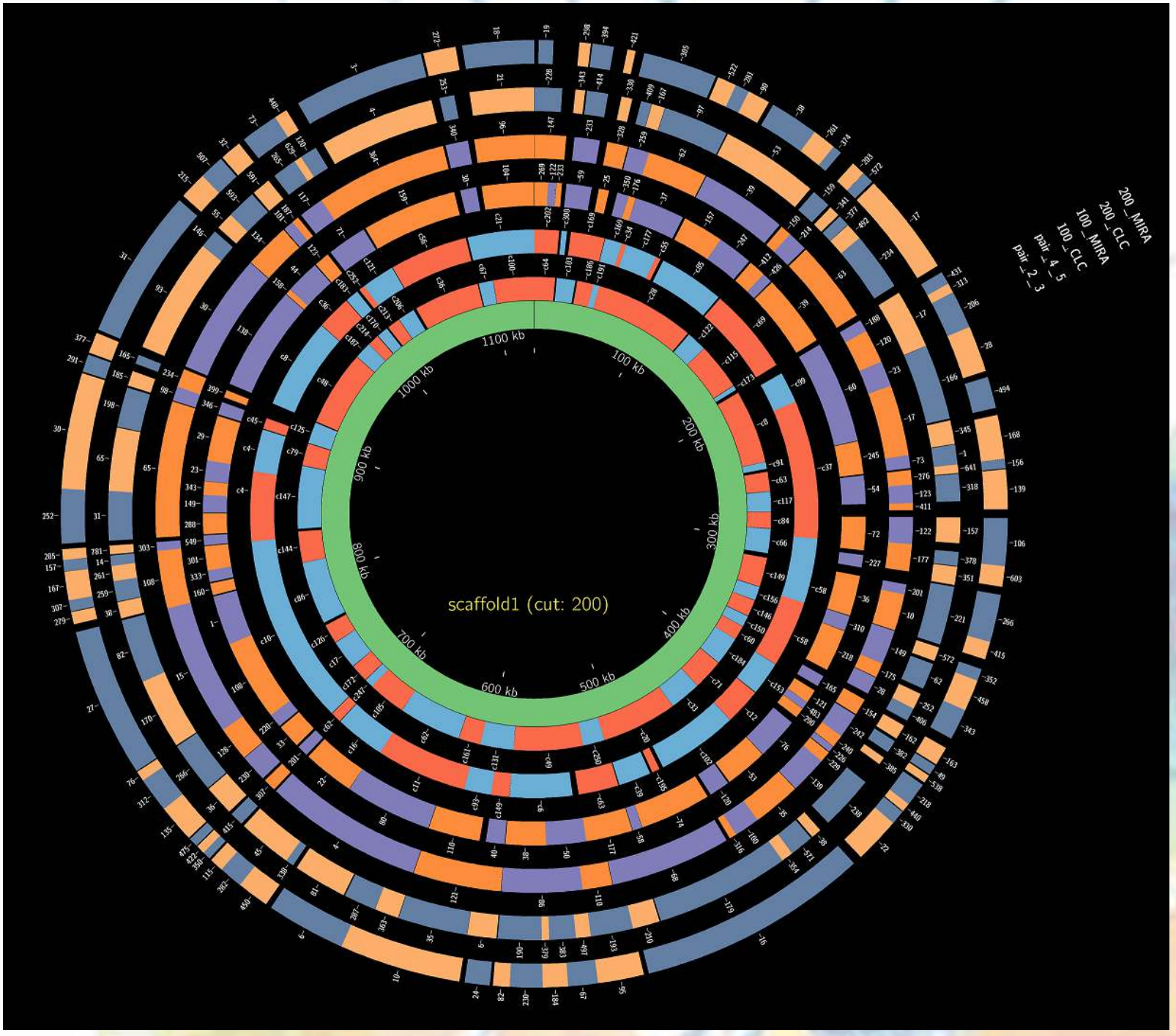
Ion Torrent PGM

- *Две фрагментные библиотеки:*
  - 100 пн (покрытие 57x)
  - 200 пн (покрытие 135x)

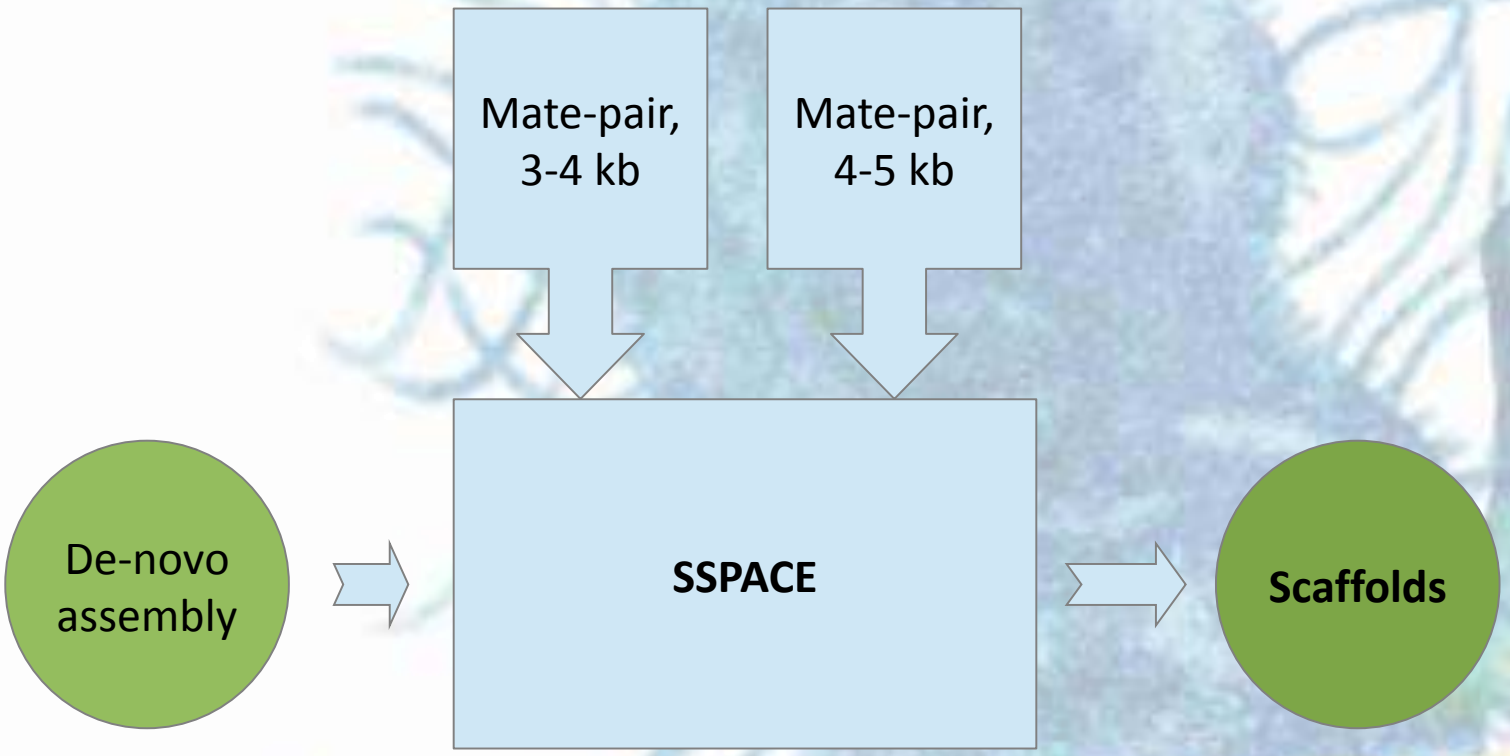
- *Две парные (mate-pair) библиотеки:*
  - вставки 2.5 кб (покрытие 149x, 73% парных)
  - вставки 4.5 кб (покрытие 142x, 87% парных)

Сборка de-novo, N50




scaffold1 (cut: 200)
100 kb
200 kb
300 kb
400 kb
500 kb
600 kb
700 kb
800 kb
900 kb
1000 kb
1100 kb
pair_2_3
pair_4_5
100_CLC
100_MIRA
200_CLC
200_MIRA

Скаффолдинг



Скаффолдинг

- Mate pair 4-5 кб:

  - 74 скаффолдов

- Mate pair 2-3 кб

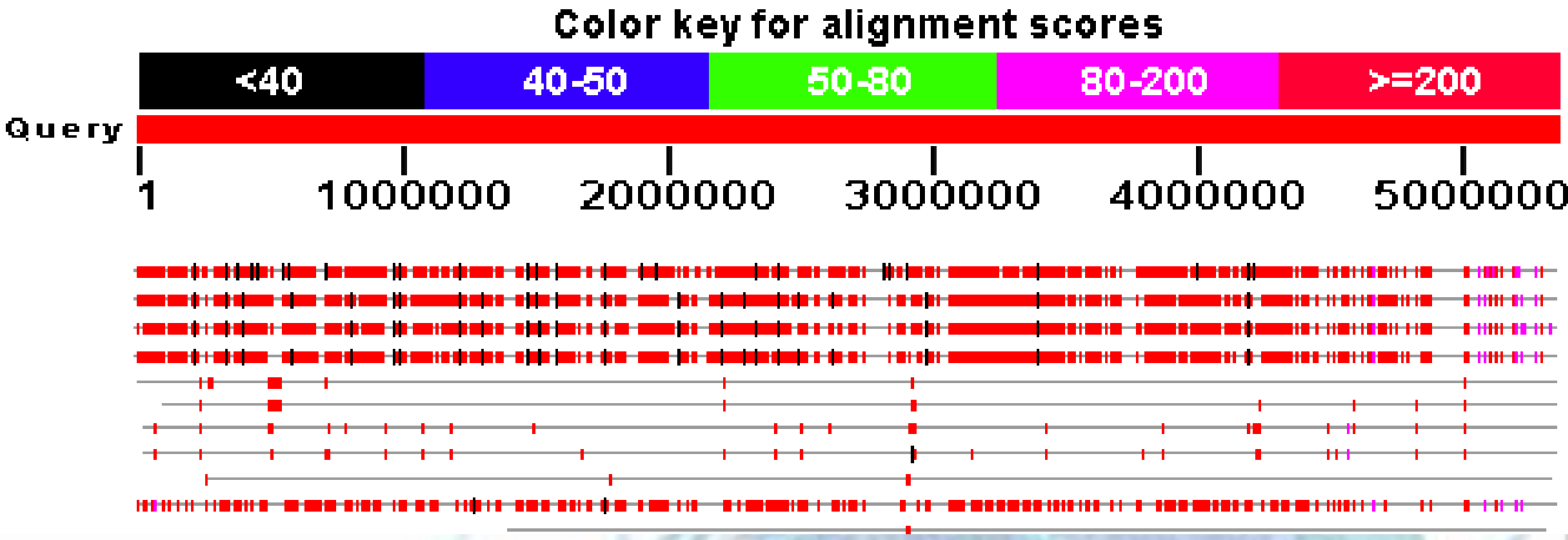
  - 90 скаффолдов

- Fragment + mate pair («ручная», Vector NTI)

  - 26 скаффолдов

Анализ генома

Whole-genome BLAST



Pseudomonas stutzeri CCUG 29243
Pseudomonas stutzeri DSM 4166
Pseudomonas stutzeri ATCC
Pseudomonas stutzeri A1501
Tolumonas auensis DSM 9187
Parvibaculum lavamentivorans DS-1

Горизонтальный перенос

- ***Tolumonas auensis***

Бактерия производящая толуол. Обитает в том числе в бескислородных озерных отложениях

Толуо́л (от исп. Tolu, толуанский бальзам) — метилбензол, бесцветная жидкость с характерным запахом, относится к аренам. Продукт каталитического риформинга бензиновых фракций нефти.

- ***Parvibaculum lavamentivorans DS-1***

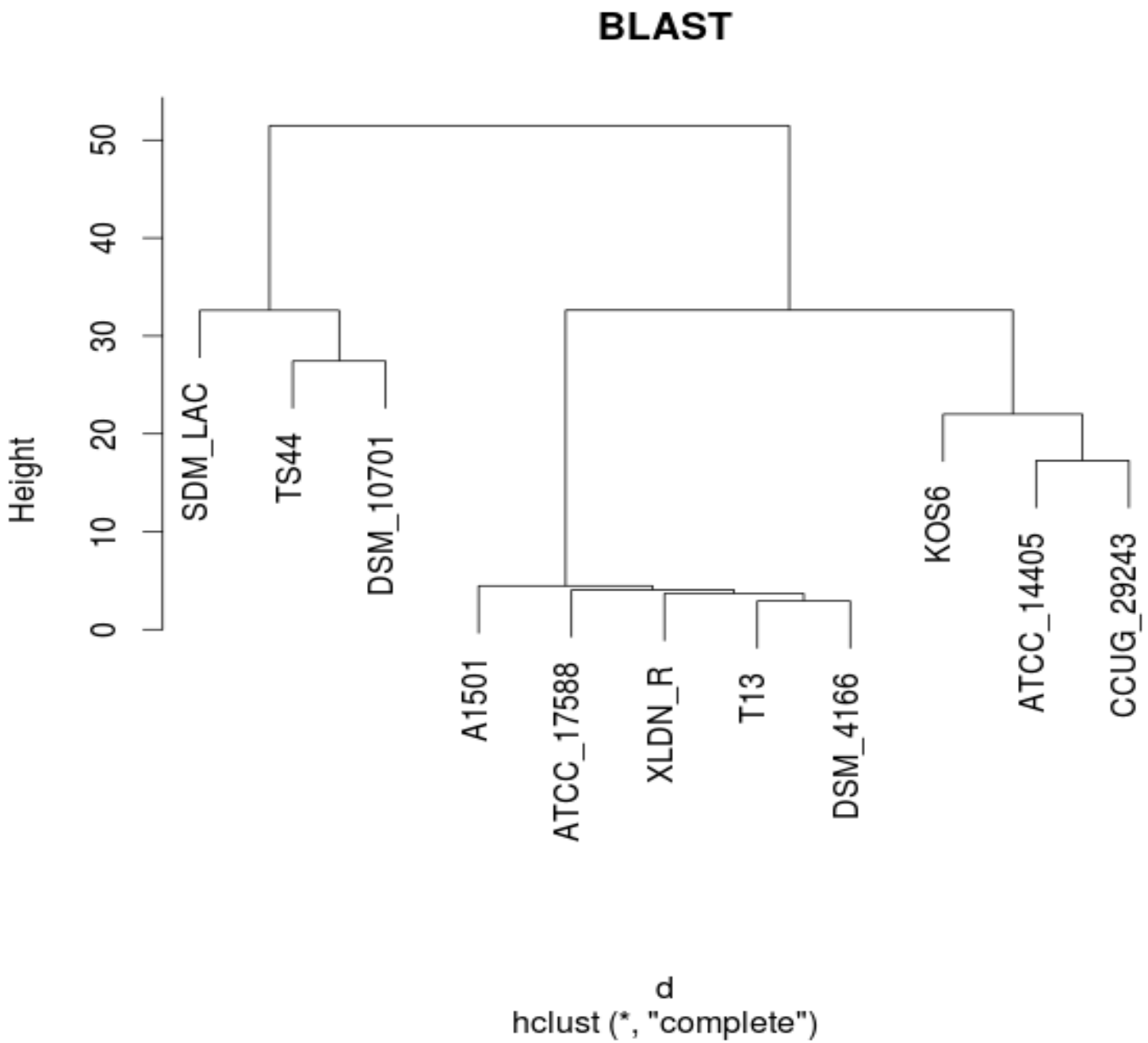
Был извлечен из очистных сооружений (Германия). Участвует в переработке линейных алкилбензолсульфонатов (в сульфофинилкарбоксилаты)

115 генов

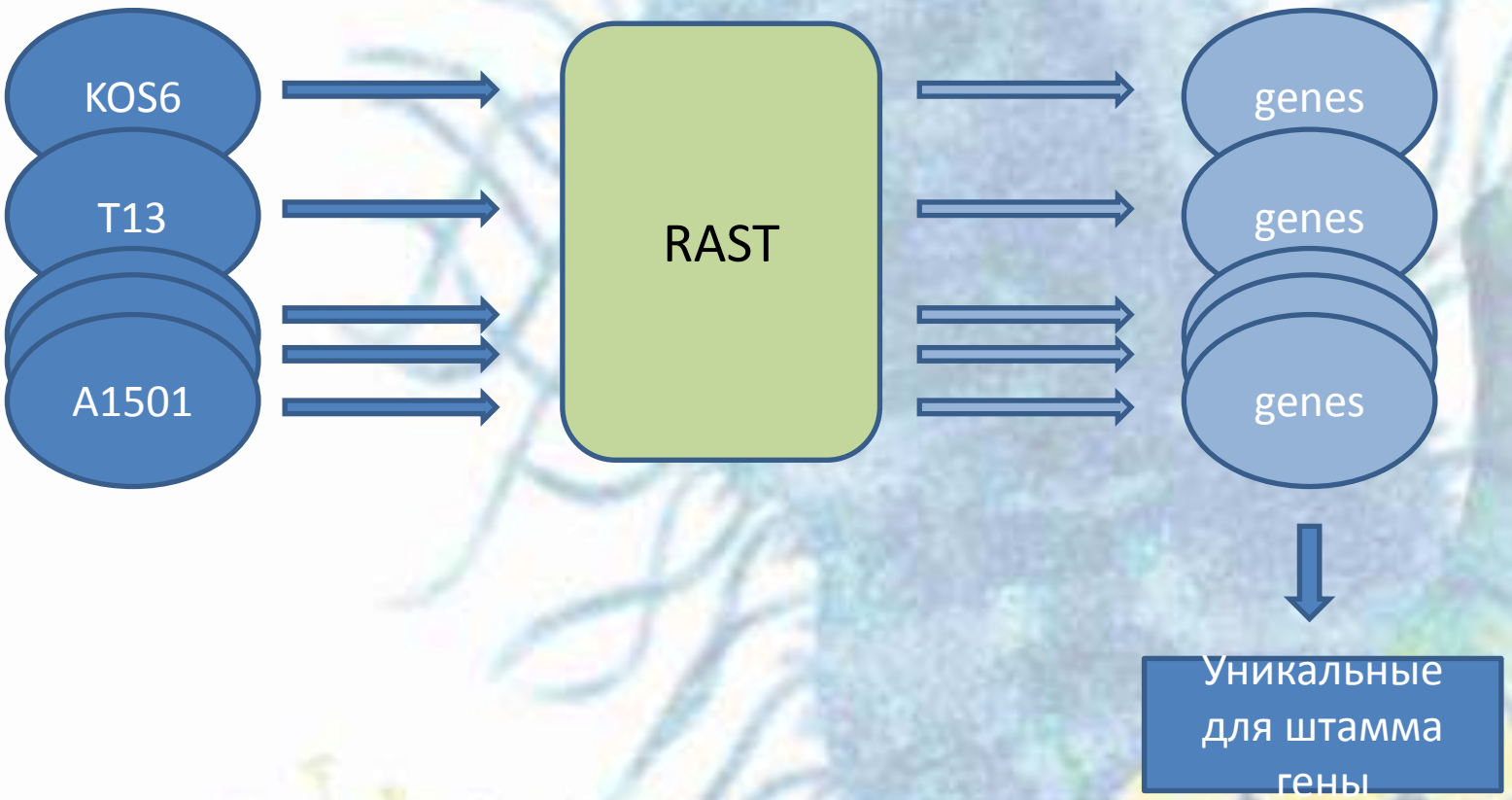
76 – гипотетические (неизвестные)

Известные штаммы *P. stutzeri*

KOS6	**шлам**
CCUG_29243	Нафталин дегр, загр. м. ос.
ATCC_14405	морск
DSM_4166	ризосф
ATCC_17588	клинич
A1501	ризосф
DSM_10701	Почва бтрансформ прир в-в
SDM_LAC	Утилиз лактат
T13	Активный ил
TS44 -	Загрязн почва трансф арсенита
XLDN_R -	Дегр крезола


BLAST
Height
0
10
20
30
40
50
SDM_LAC
TS44
DSM_10701
A1501
ATCC_17588
XLDN_R
T13
DSM_4166
KOS6
ATCC_14405
CCUG_29243
d
hclust (*, "complete")

Как найти «уникальные» гены



- [1] "Enoyl-CoA hydratase [branched-chain amino acid degradation] (EC 4.2.1.17)"
  [2] "Aldehyde dehydrogenase A (EC 1.2.1.22)"
  [3] "Sialic acid transporter (permease) NanT"
  [4] "Acyl dehydratase"
  [5] "FIG016502: iron uptake protein" - не уникальный...
  [6] "RecD-like DNA helicase Atu2026"
  [7] "RecD-like DNA helicase YrrC"
  [8] "VgrG-3 protein" - не уникальный...
  [9] "TRAP-type C4-dicarboxylate transport system, small permease component"
  [10] "TRAP dicarboxylate transporter, DctM subunit, unknown substrate 5"
  [11] "TRAP dicarboxylate transporter, DctQ subunit, unknown substrate 5"
  [12] "TRAP transporter solute receptor, unknown substrate 5"
  [13] "2,4-dihydroxyhept-2-ene-1,7-dioic acid aldolase (EC 4.1.2.-)"
  [14] "2-hydroxyhepta-2,4-diene-1,7-dioate isomerase (EC 5.3.3.-)"
  [15] "2-oxo-hepta-3-ene-1,7-dioic acid hydratase (EC 4.2.-.-)"
  [16] "3,4-dihydroxyphenylacetate 2,3-dioxygenase (EC 1.13.11.15)"
  [17] "5-carboxymethyl-2-hydroxymuconate delta-isomerase (EC 5.3.3.10)"
  [18] "5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase (EC 1.2.1.60)"
  [19] "5-carboxymethyl-2-oxo-hex-3- ene-1,7-dioate decarboxylase (EC 4.1.1.68)"
  [20] "Homoprotocatechuate degradative operon repressor"
  [21] "Transcriptional activator of 4-hydroxyphenylacetate 3-monooxygenase operon, XylS/AraC family"
  [22] "gliding motility protein MglA"
  [23] "Cyn operon transcriptional activator"
  [24] "Phage capsid scaffolding protein"
  [25] "Phage tail fiber protein"
  [26] "Phage major tail tube protein"
  [27] "Phage tail completion protein"
  [28] "Phage tail length tape-measure protein"
  [29] "Phage tail protein"
  [30] "Phage tail sheath monomer"
  [31] "ParD protein (antitoxin to ParE)"
  [32] "Vanillate O-demethylase oxidoreductase (EC 1.14.13.-)"
  [33] "Carbon monoxide dehydrogenase large chain (EC 1.2.99.2)"
  [34] "Carbon monoxide dehydrogenase medium chain (EC 1.2.99.2)"
  [35] "Carbon monoxide dehydrogenase small chain (EC 1.2.99.2)"
  [37] "transcriptional regulator, Crp/Fnr family"

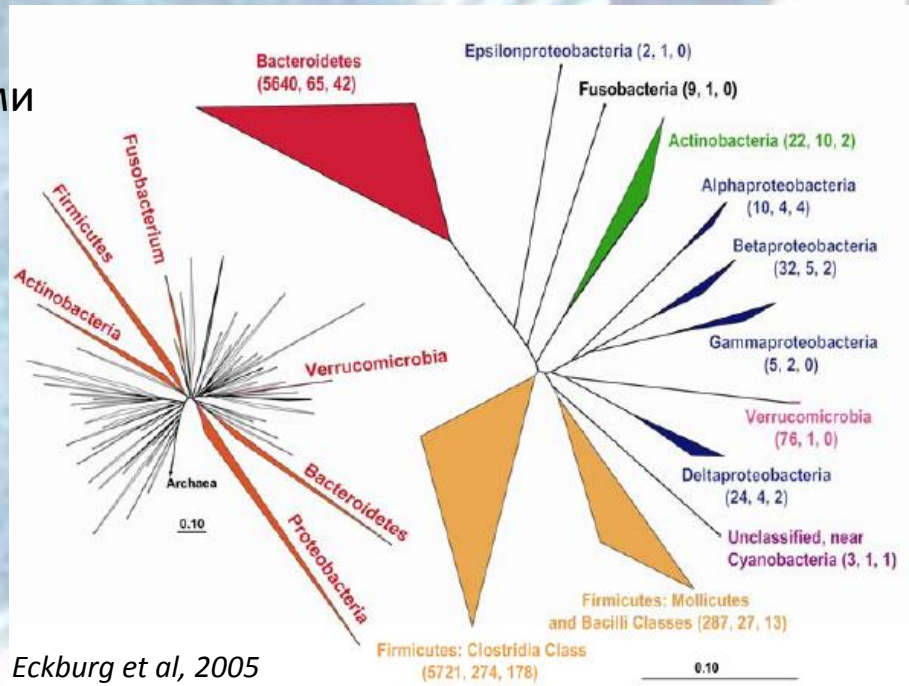
		Денитрификация, nor, nir	**Азотфиксация** nif	**Деградация нафталена, nah**	VI type
KOS6	**шлам**	**+**	**+**	**+**	**+**
CCUG_29243	Нафталин дегр, загр. морск. осадки.	+	-	+	+
ATCC_14405	Морская вода	+	-	-	-
DSM_4166	Ризосферы	+	+	-	-
ATCC_17588	Клинический	+	-	-	-
A1501	Ризосферы	+	+	-	-
DSM_10701	Почва	+	-	-	-
SDM_LAC	Утилизатор лактата	+	-	-	-
T13	Активный ил	+	-	-	-
Tolumonas_auensis	озер. отложения	+	+	-	+
TS44	Загрязненная почва трансф. арсенита	+	-	-	+
XLDN_R	Деград-я крезола	+	-	-	-

Выводы

- Произведена сборка de-novo и скаффолдинг *P. stutzeri KOS6*
  - Лучшие показатели: mate-pair, MIRA
- Присутствует 100 кб вставка, гомологичная участкам геномов бактерий, вероятно, обитающих в сходных условиях
- Комбинация спобностей к денитрификации, азотфиксации, деградации нафталена является уникальной среди других известных штаммов *P. stutzeri*

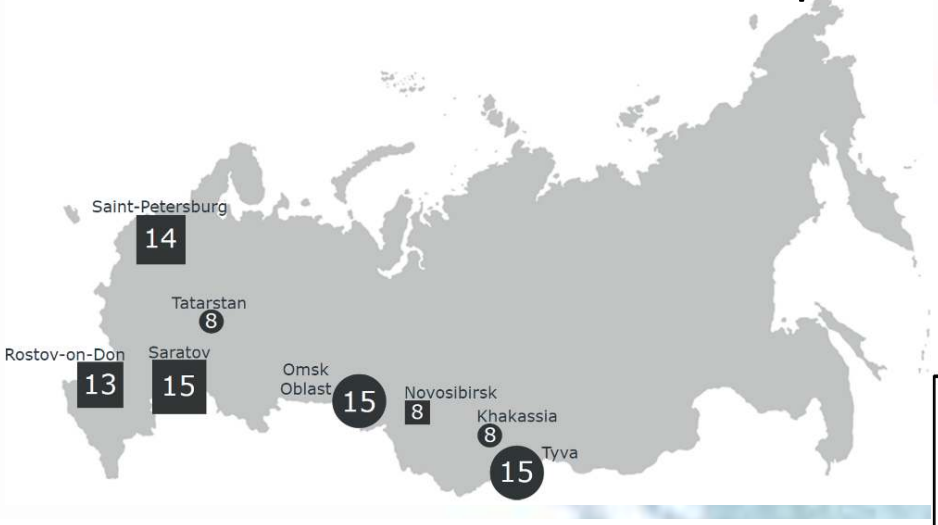
Микробиота кишечника человека

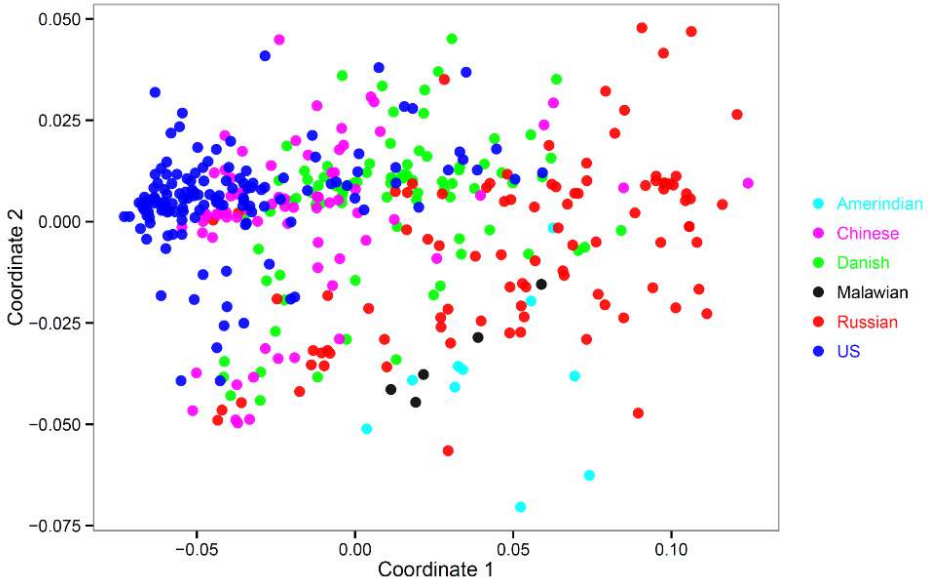
- Метагеном
- 300-1000 видов бактерий
- Число клеток ~ 10^{15}: на 2 порядка раз больше, чем клеток человека
- Число генов в метагеноме ~ 10^{6}: на 2 порядка больше, чем генов человека
- Функции:
  - Метаболизм
  - Предотвращение колонизации патогенами
  - Иммунитет человека
  - Защита от воспалительных заболеваний
  - Связь с высшей нервной деятельностью

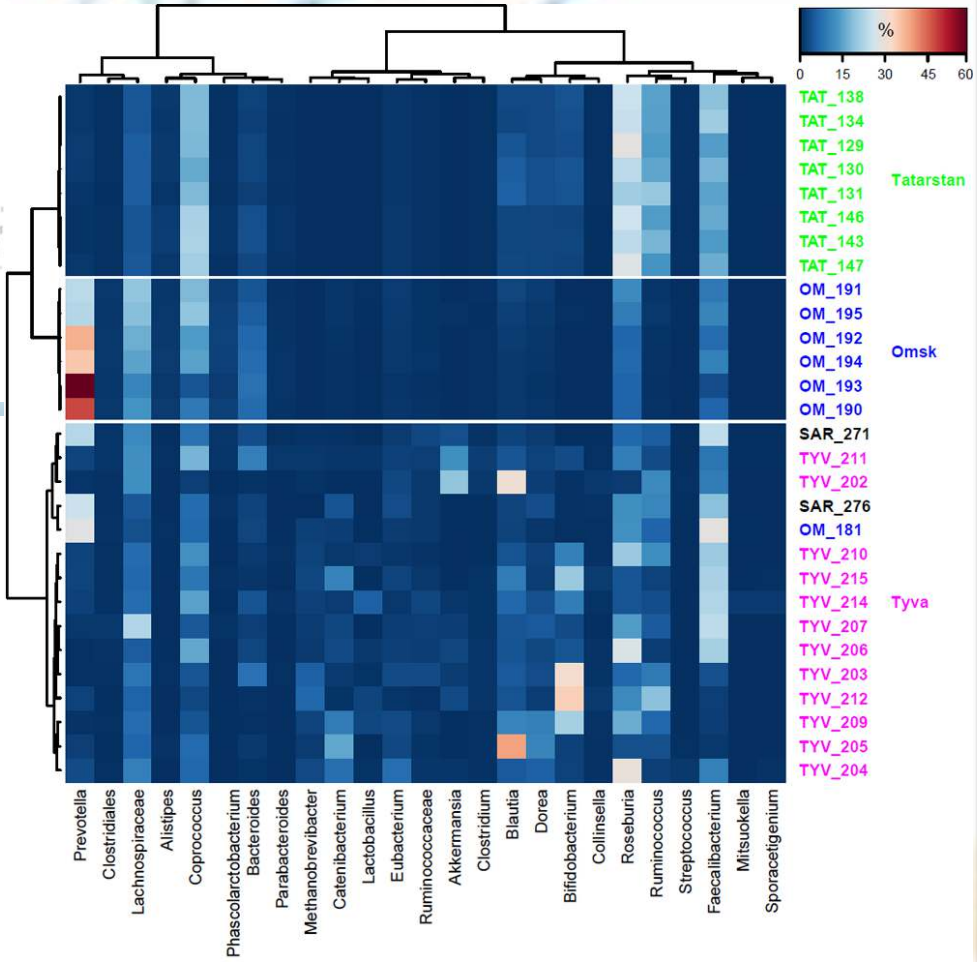


Eckburg et al, 2005

Российский метагеномный проект





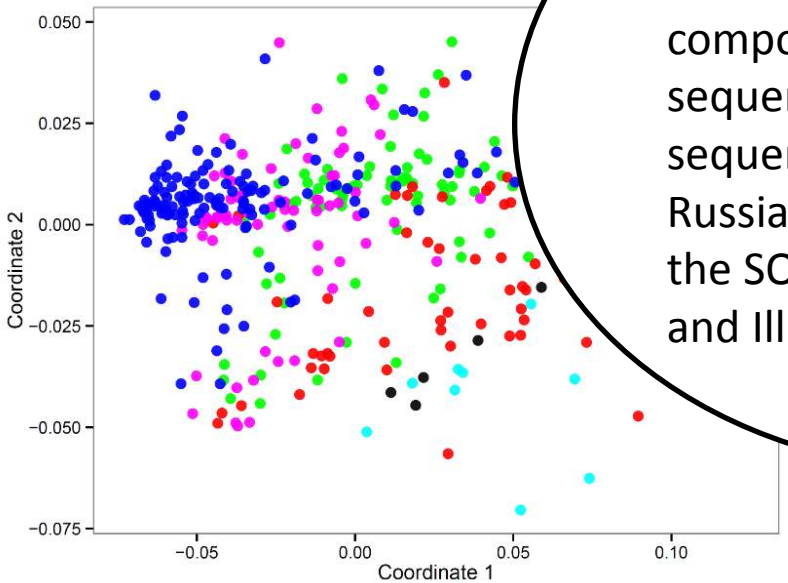


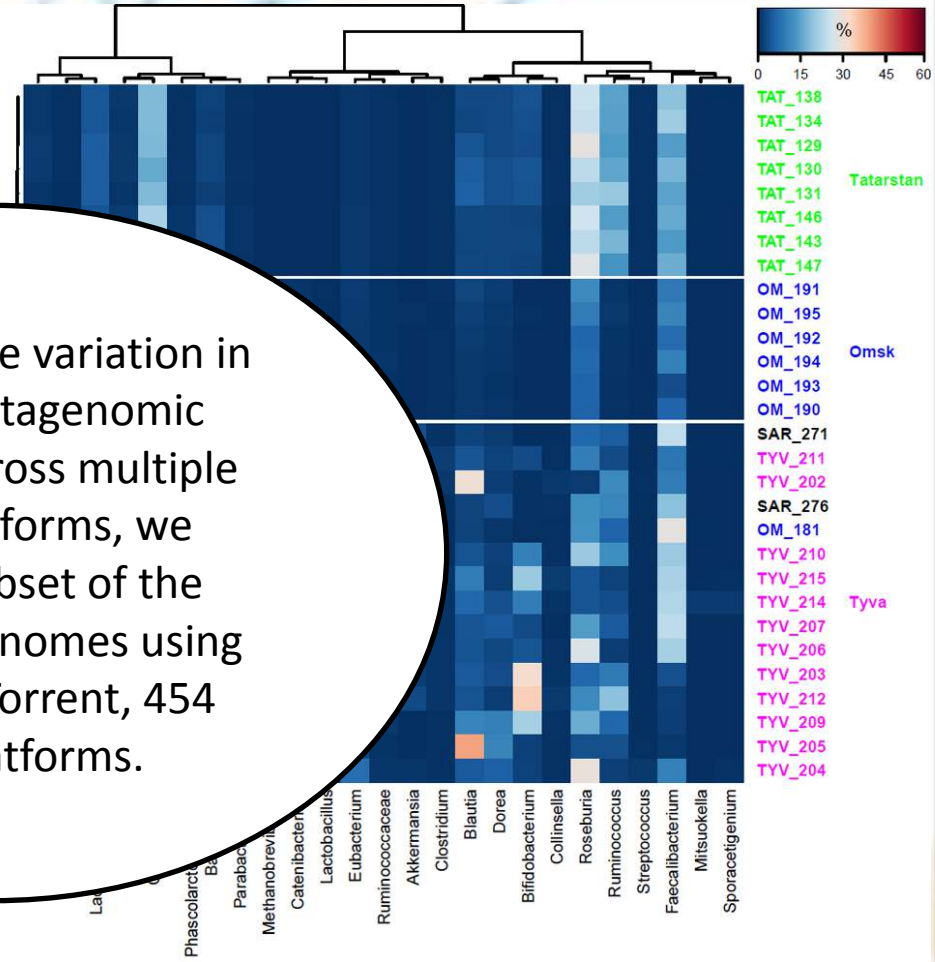
Tyakht, A. *et al*. **Novel human gut microbiota community structures in urban and rural populations in Russia**. *Nature Communications*, 2013 (accepted).

http://www.metagenome.ru

metagenome.ru

Российский метагеномный проект





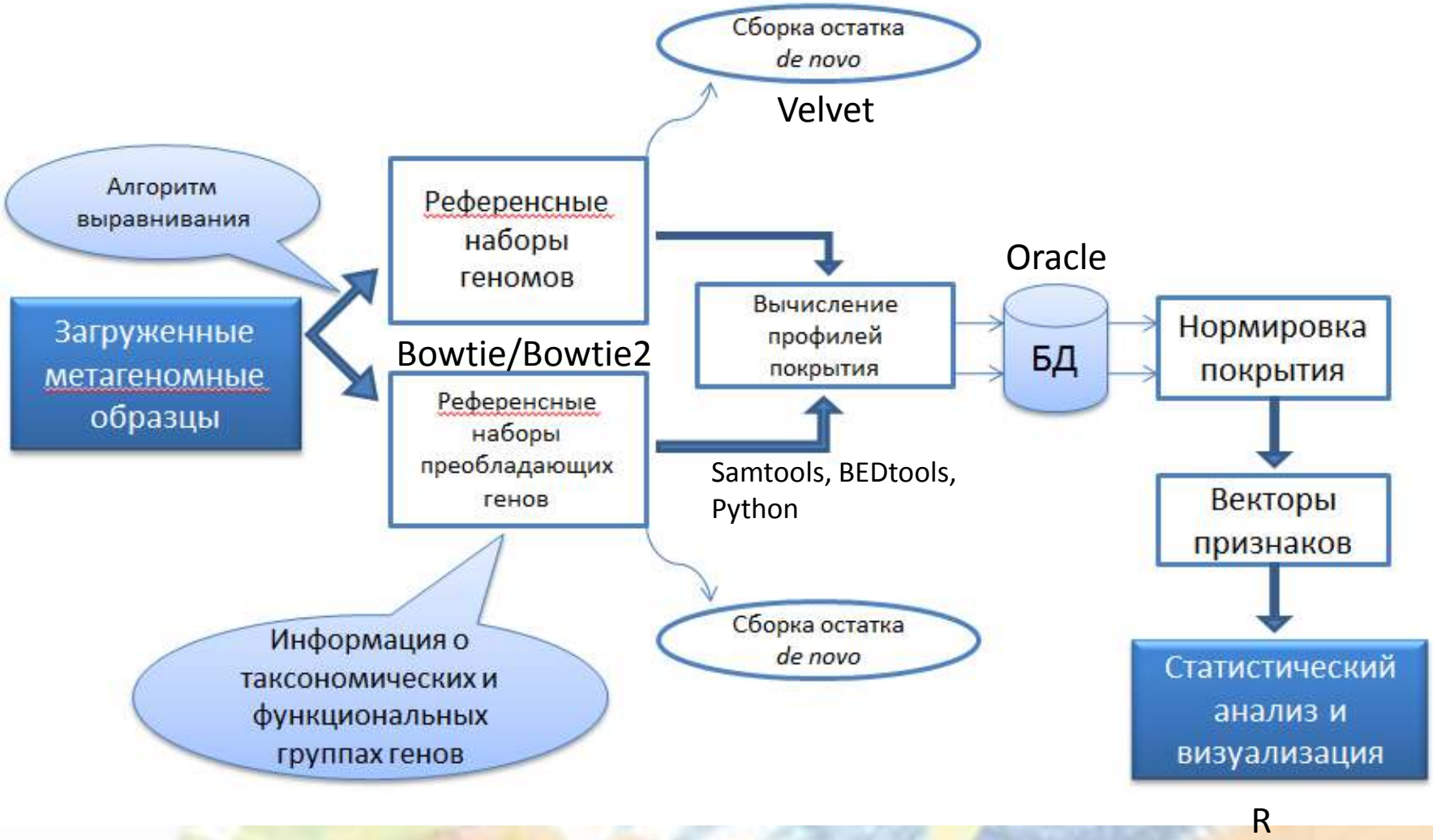
...To evaluate the variation in the shotgun metagenomic composition across multiple sequencing platforms, we sequenced a subset of the Russian metagenomes using the SOLiD, Ion Torrent, 454 and Illumina platforms.

Tyakht, A. *et al*. **Novel human gut microbiota community structures in urban and rural populations in Russia**. *Nature Communications*, 2013 (accepted).

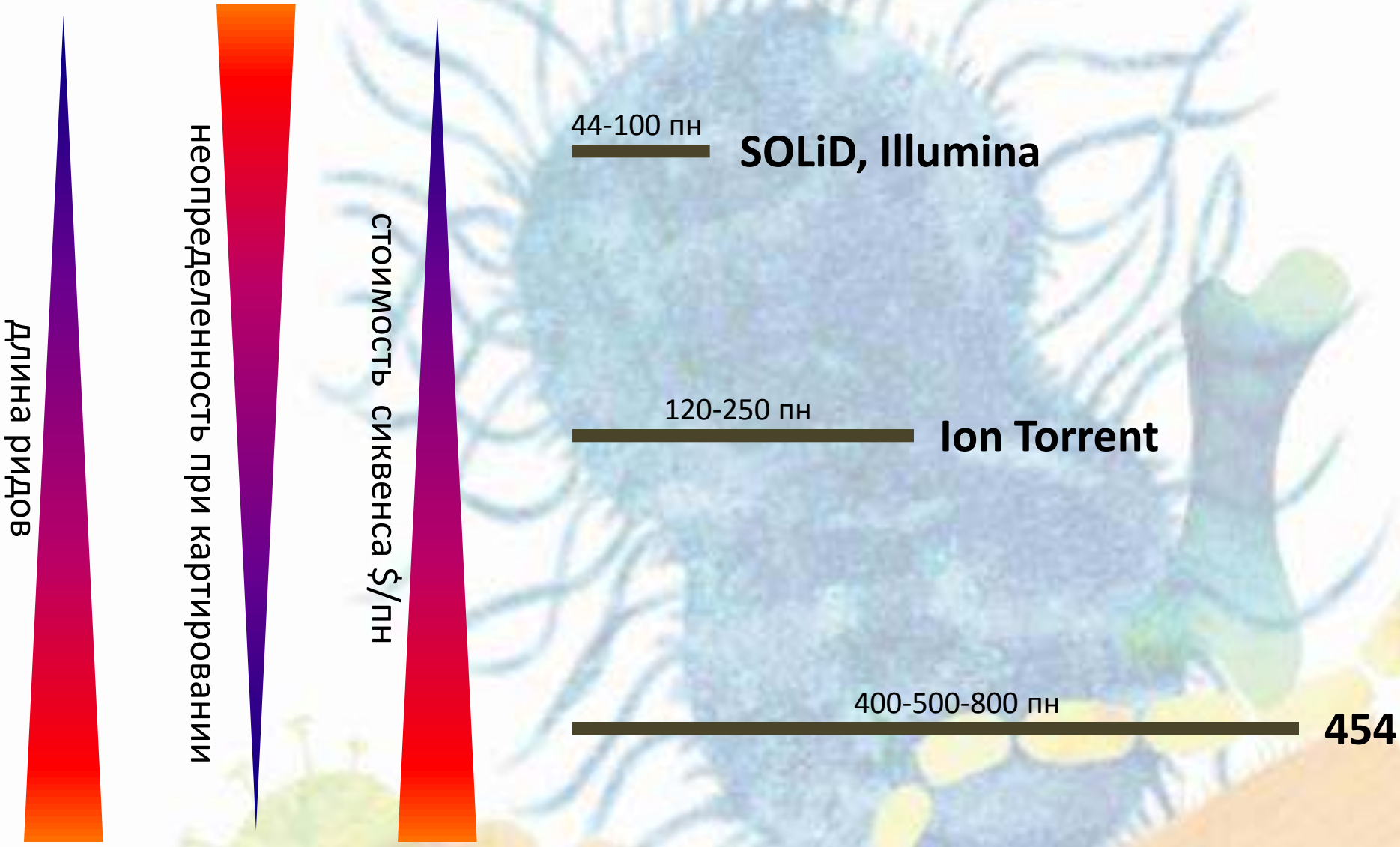
http://www.metagenome.ru

metagenome.ru

Высокопроизводительный конвейер для анализа метагеномных данных

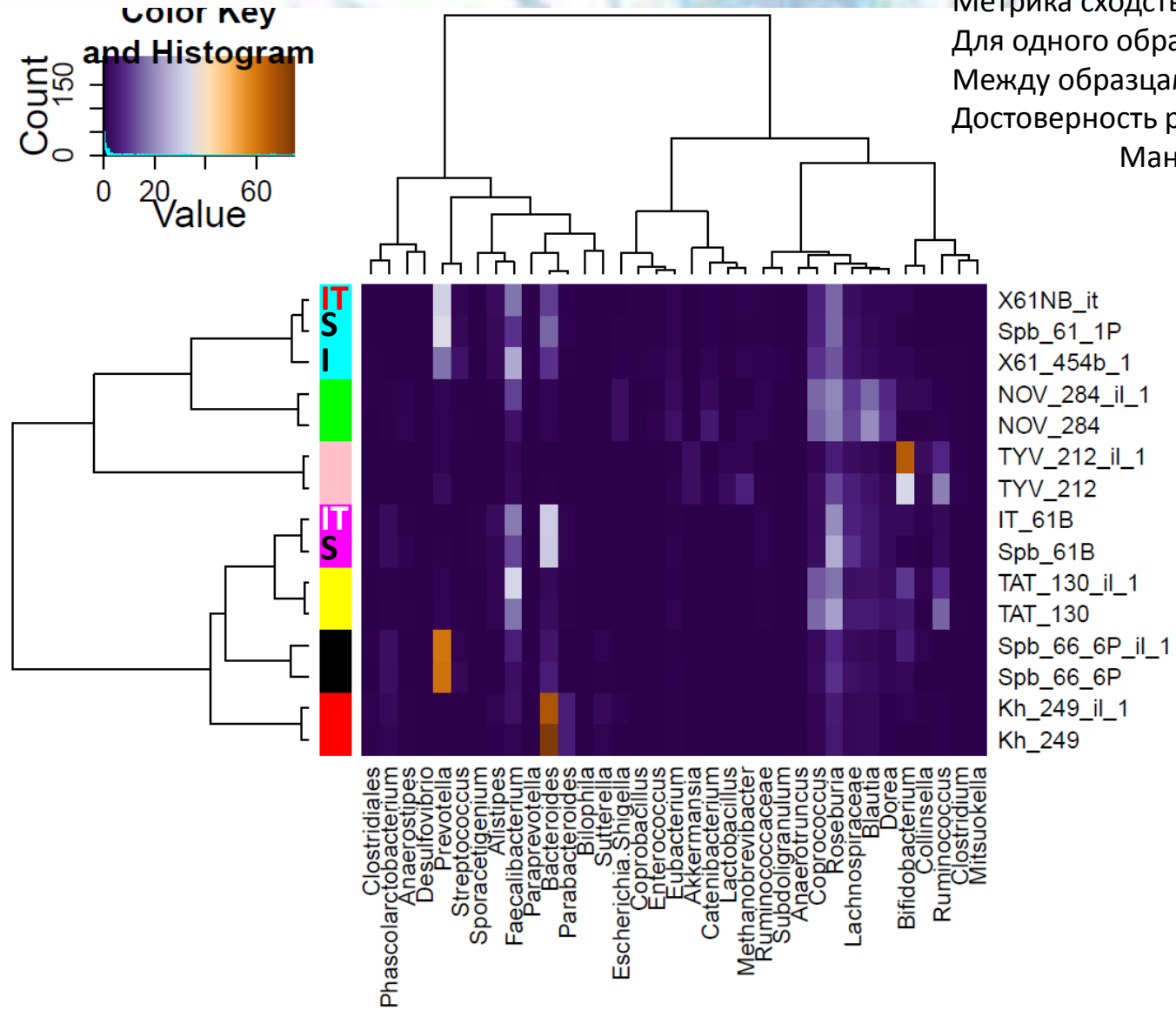


Выбор платформы секвенирования для метагеномного исследования



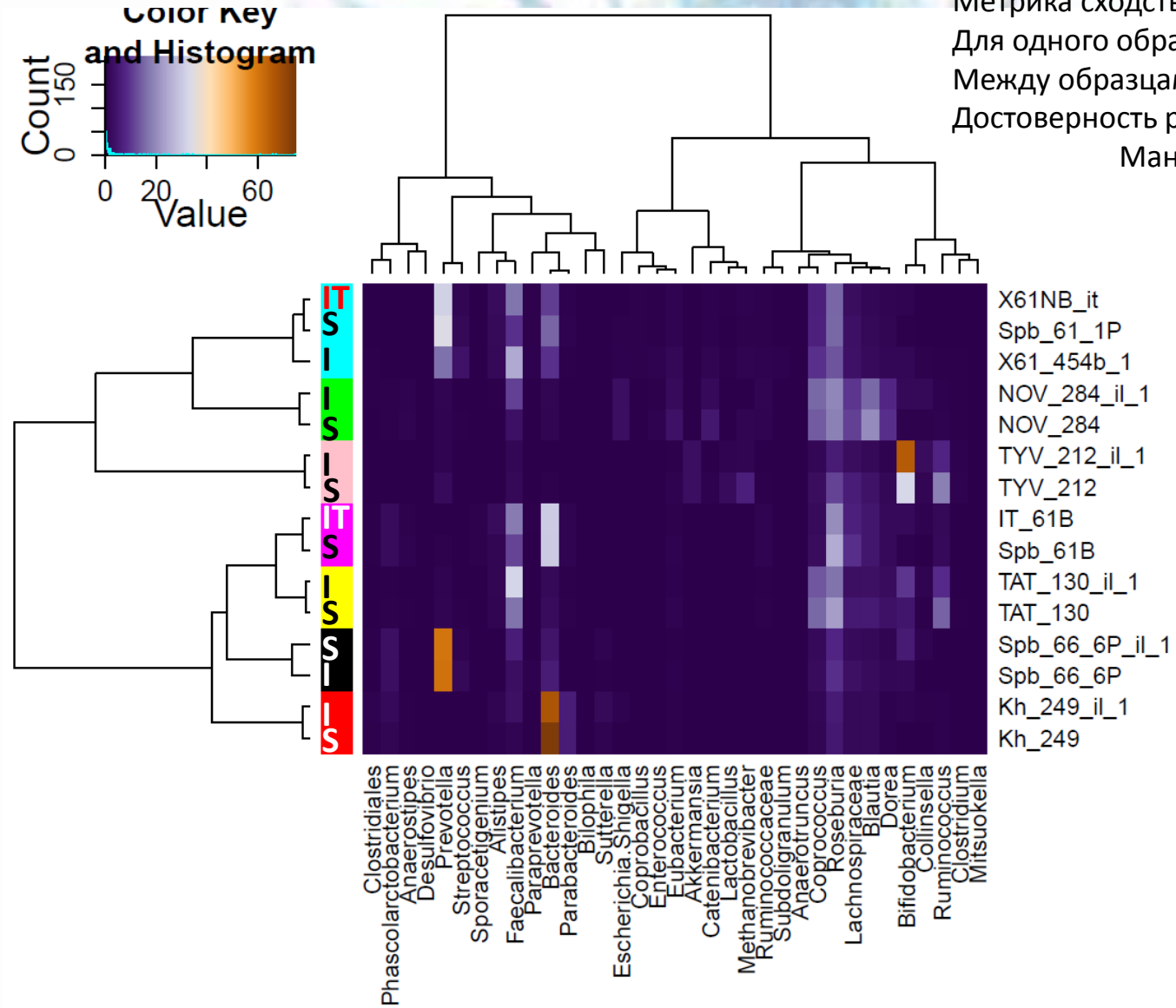
Сравнение состава микробного сообщества, полученного на разных платформах

Признаки: родовой состав
Метрика сходства: корреляция Спирмена
Для одного образца: 0,95 ± 0,03
Между образцами: 0,74 ± 0,10
Достоверность различия: тесты Манна-Уитни, P < 0,01



Сравнение состава микробного сообщества, полученного на разных платформах

Признаки: родовой состав
Метрика сходства: корреляция Спирмена
Для одного образца: 0,95 ± 0,03
Между образцами: 0,74 ± 0,10
Достоверность различия: тесты Манна-Уитни, P < 0,01



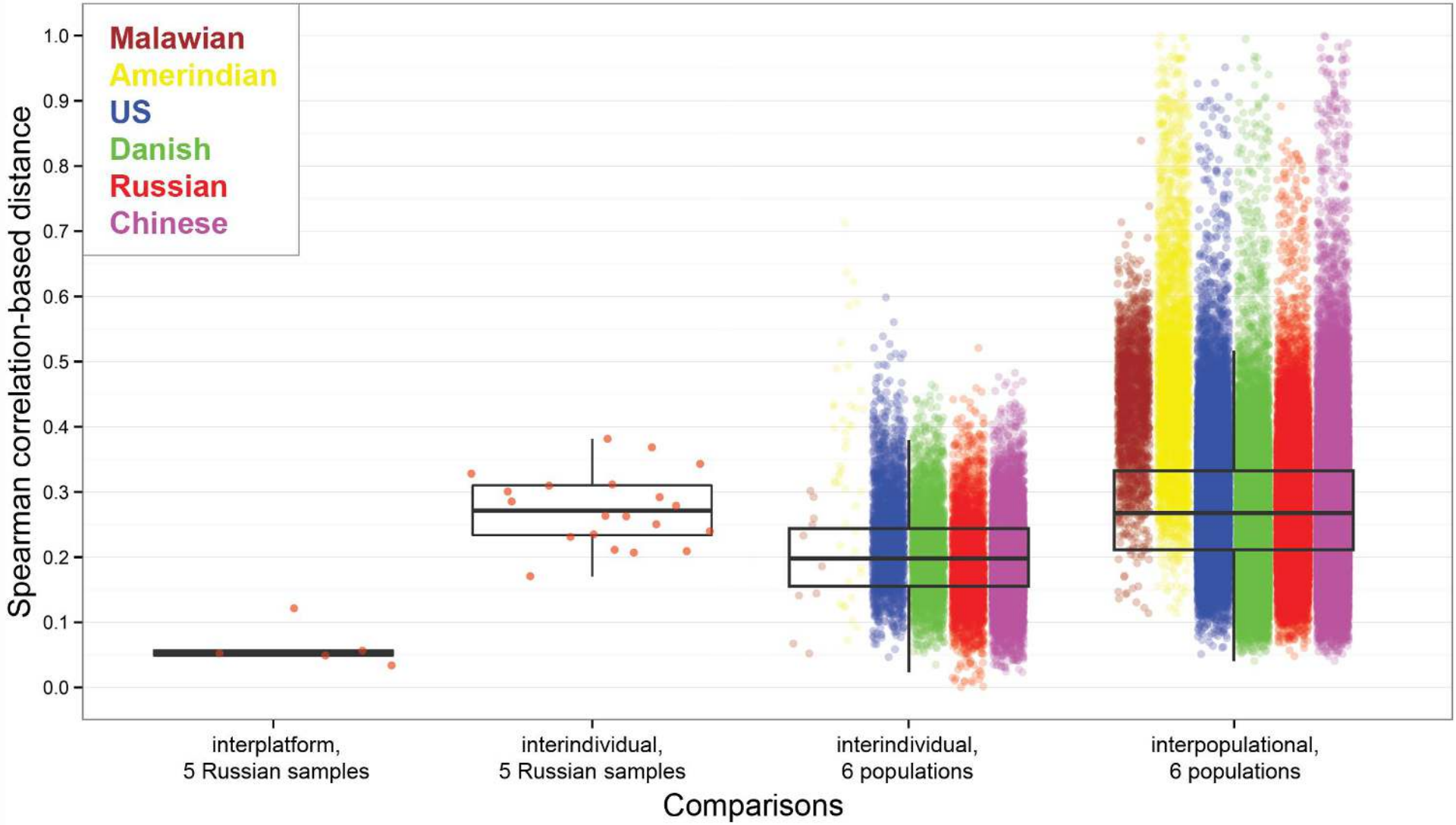
Разные длина ридов на Ion Torrent

	Roseburia	Faecalibacterium	Lachnospiraceae	Coprococcus	Blautia	Prevotella	Eubacterium	Dorea	Bifidobacterium	Ruminococcus	Collinsella	Lactobacillus	Akkermansia	Methanobrevibacter	Catenibacterium	Ruminococcaceae	Subdoligranulum	Phascolarctobacterium	Anaerostipes	Mitsuokella	Bacteroides	Clostridiales	Clostridium	Coprobacillus	Enterococcus	Anaerotruncus	Escherichia.Shigella	Holdemania	Streptococcus	Actinomyces	Acidaminococcus	Gordonibacter	Marvinbryantia	Bilophila	Alistipes	Parabacteroides
Ion Torrent, short reads	23.83	34.56	8.3	7.22	5.44	4.26	2.21	2.17	2.48	1.37	2.04	0.81	1.11	0.58	0.66	0.56	0.61	0.36	0.16	0.38	0.14	0.19	0.1	0.09	0.09	0.07	0.04	0.05	0.02	0.05	0	0.03	0	0.02	0	0
Ion Torrent, long reads	22.66	23.37	10.55	8.21	10.07	5.31	4.07	3.17	1.97	1.31	1.53	1.39	1.32	1.12	1.22	0.51	0.44	0.36	0.19	0.15	0.27	0.16	0.12	0.11	0.12	0.07	0.13	0.04	0.05	0	0	0	0	0	0	0
SOLiD	37.48	13.98	12.17	13.88	6	3.71	2.27	3.08	0.62	1.28	0.22	0.67	0.43	1.08	0.86	0.47	0.22	0.22	0.3	0.04	0.17	0.2	0.17	0.15	0.09	0.08	0.03	0.04	0.01	0	0.03	0	0.02	0	0.01	0.01

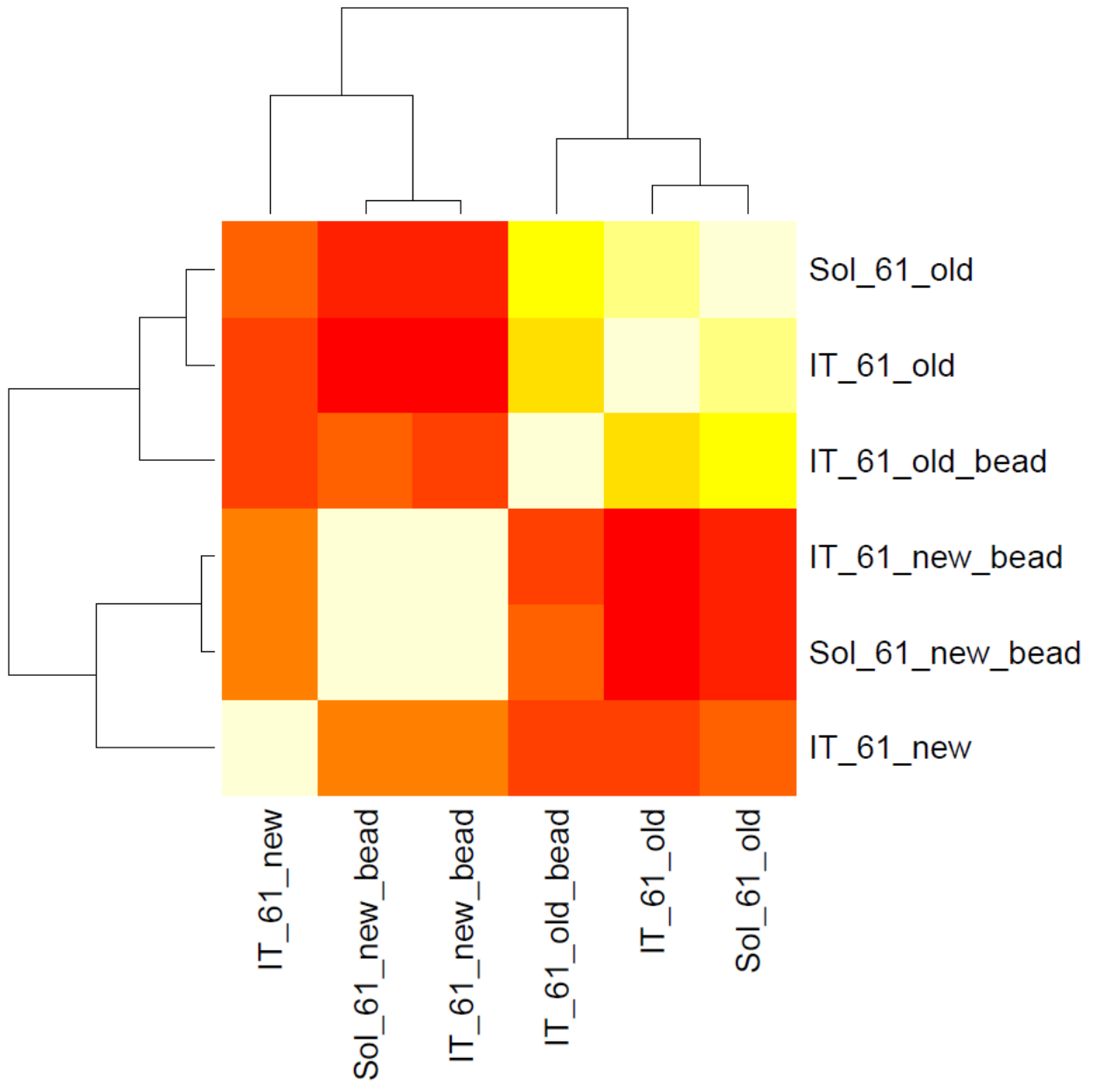
- Ion Torrent
  - 120 пн
  - 250 пн
- Solid – 50 пн

- Корреляция: 0.95 ± 0.02 s.d.

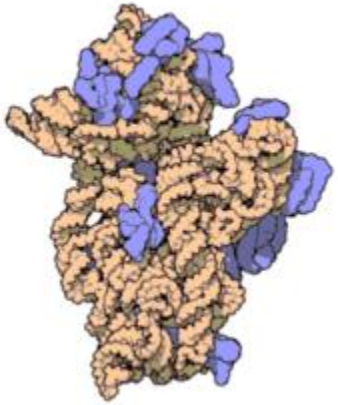
Фоновая корреляция



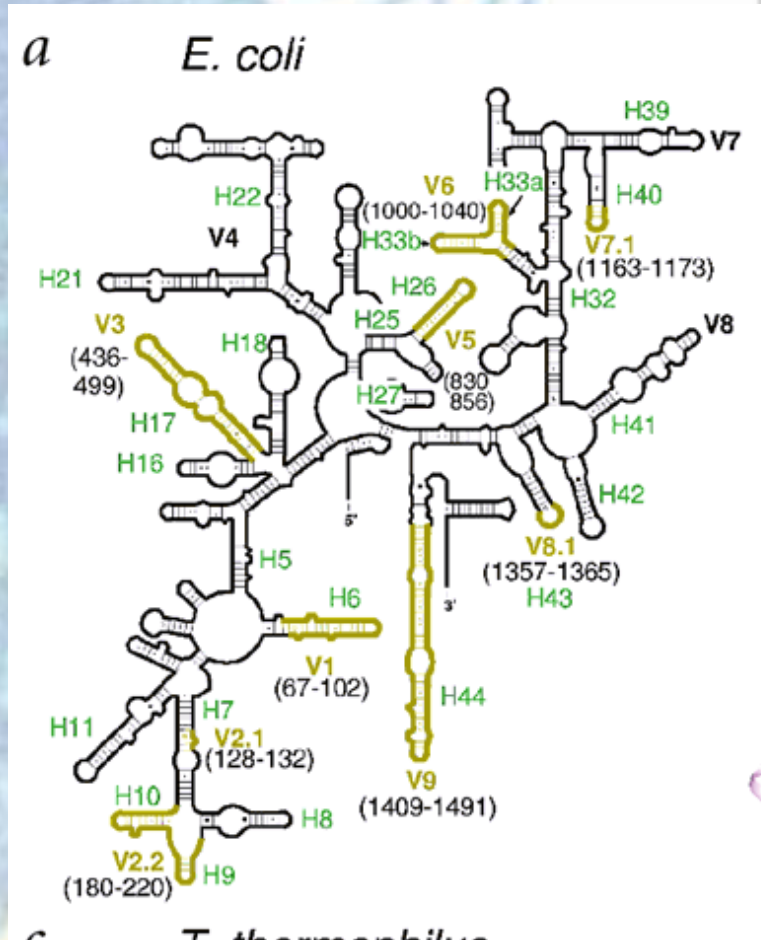
Факторы вариабельности микробиотного профилирования: временная вариация > метод пробоподготовки > платформа



Анализ микробного состава по секвенированию последовательностей 16S рРНК



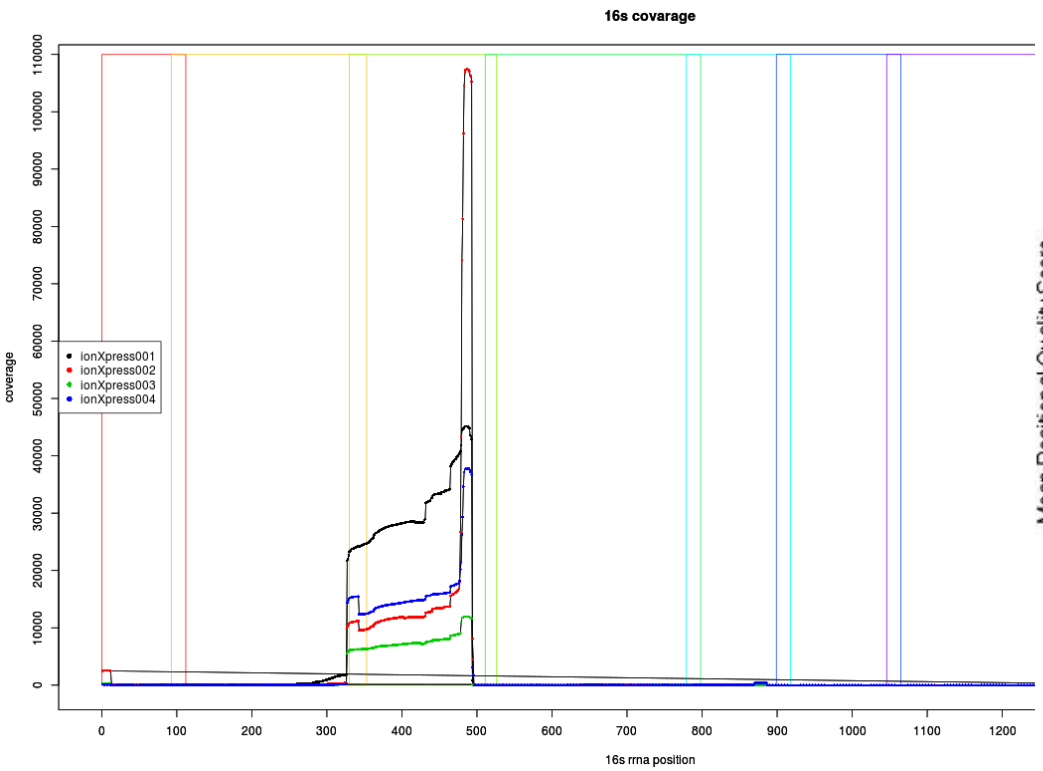
- Внутривидовое сходство сиквенса 98-99%
- Выделяется из тотального ДНК с помощью универсальных праймеров
- Секвенируется целиком либо вариабельные области
  - Для микробиоты кишечника: стандартный протокол Human Microbiome Project (http://www.hmpdacc.org) для V13, V35
- Для эукариот – можно использовать 18S рРНК.

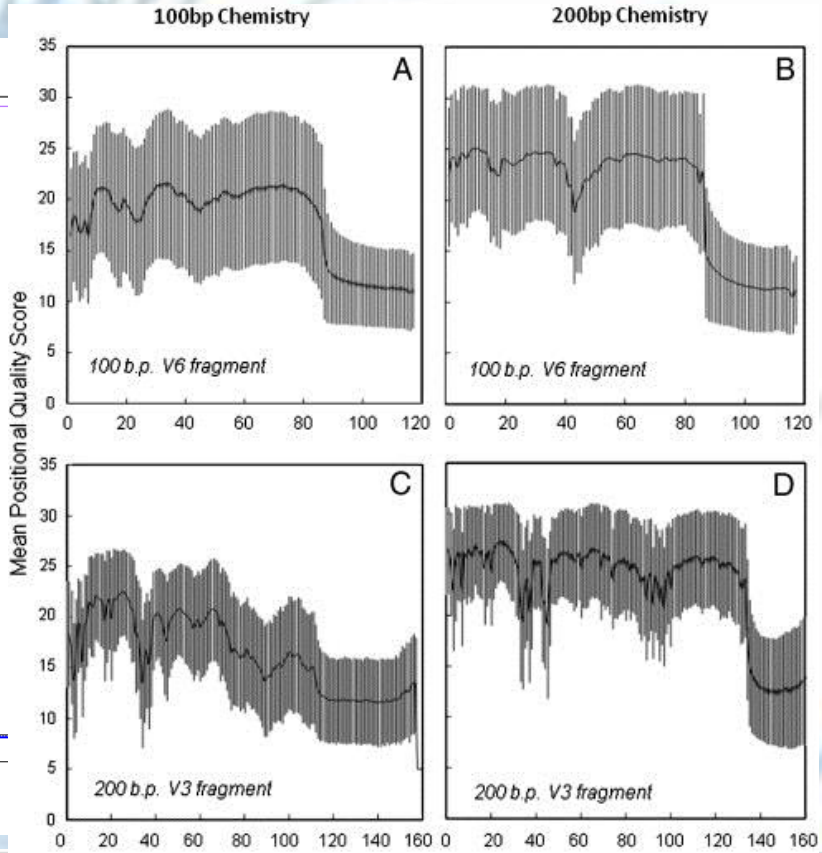


Преимущества 16S формата

- Эффективность, поточный анализ
- Секвенирование из малого количества бактериальной ДНК (биоптат тканей человека, мокрота)
- Детекция бактериального заражения даже при низкой глубине покрытия

Ion torrent 16S





ELSEVIER
Volume 91, Issue 1, October 2012, Pages 80–88

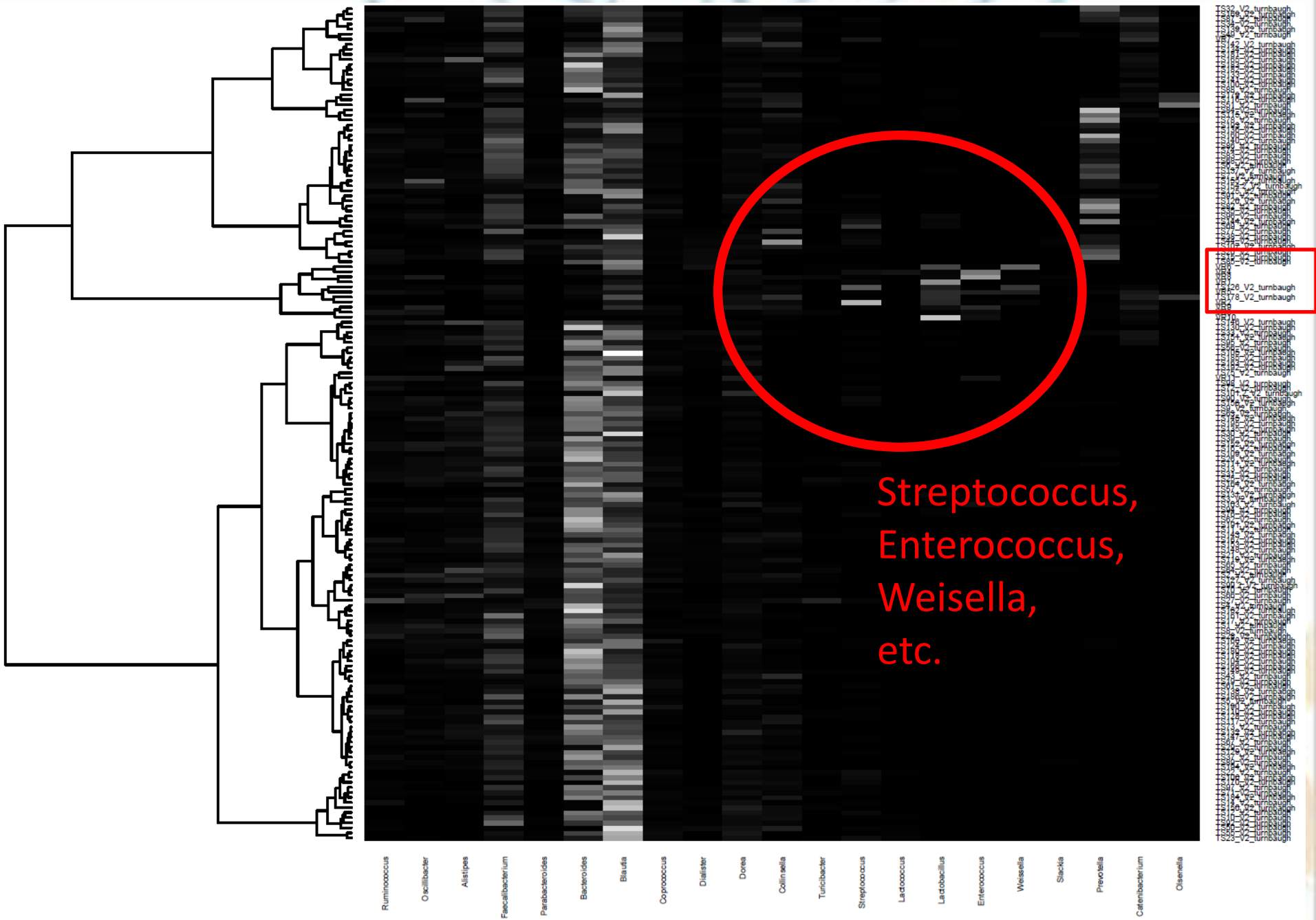
Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) Platform

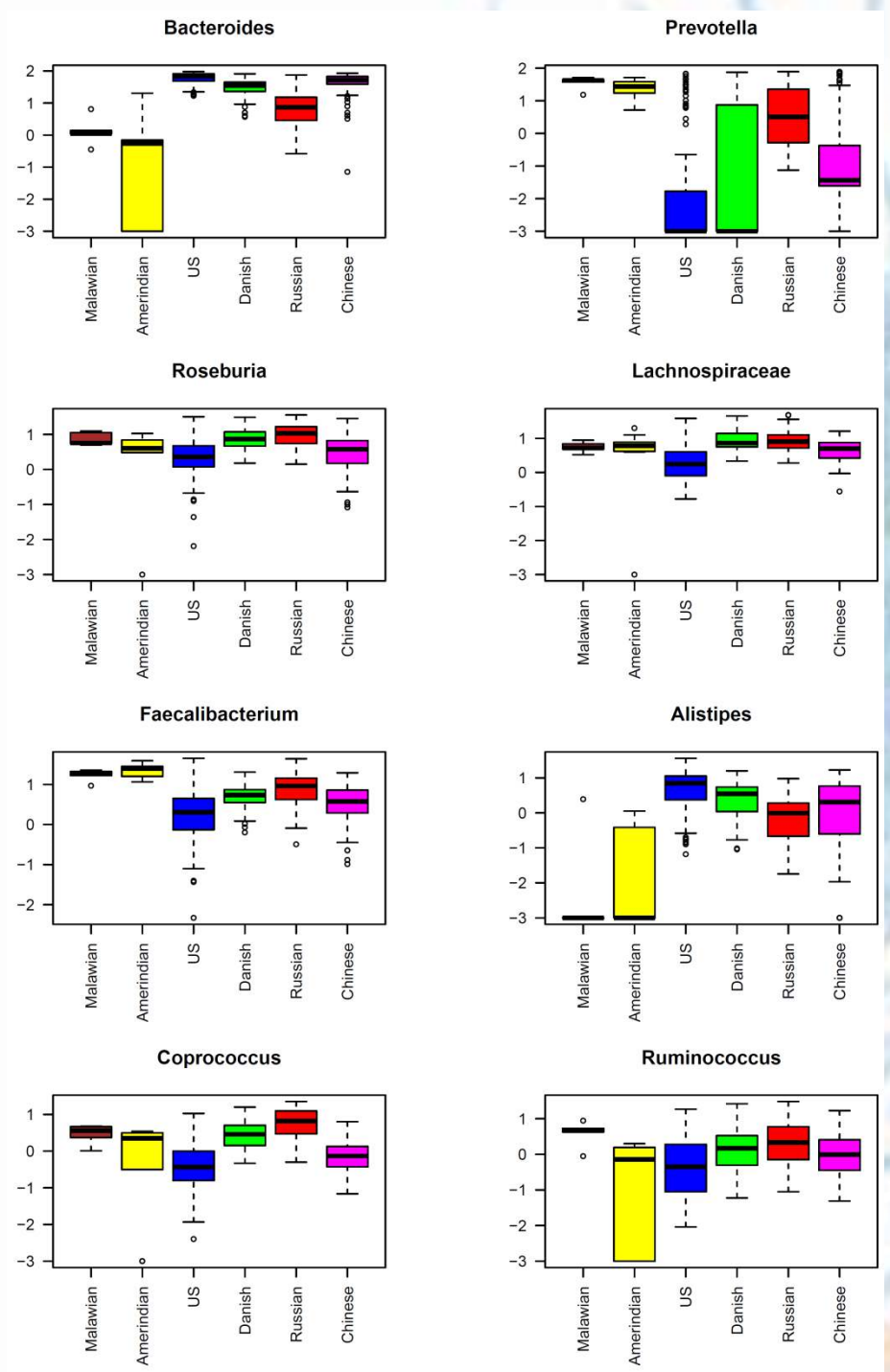
Andrew S. Whiteley[a, b], Sasha Jenkins[b, c], Ian Waite[b, c], Nina Kresoje[d, e], Hugh Payne[f], Bruce Mullan[f], Richard Allcock[d, e], Anthony O'Donnell[b, c]

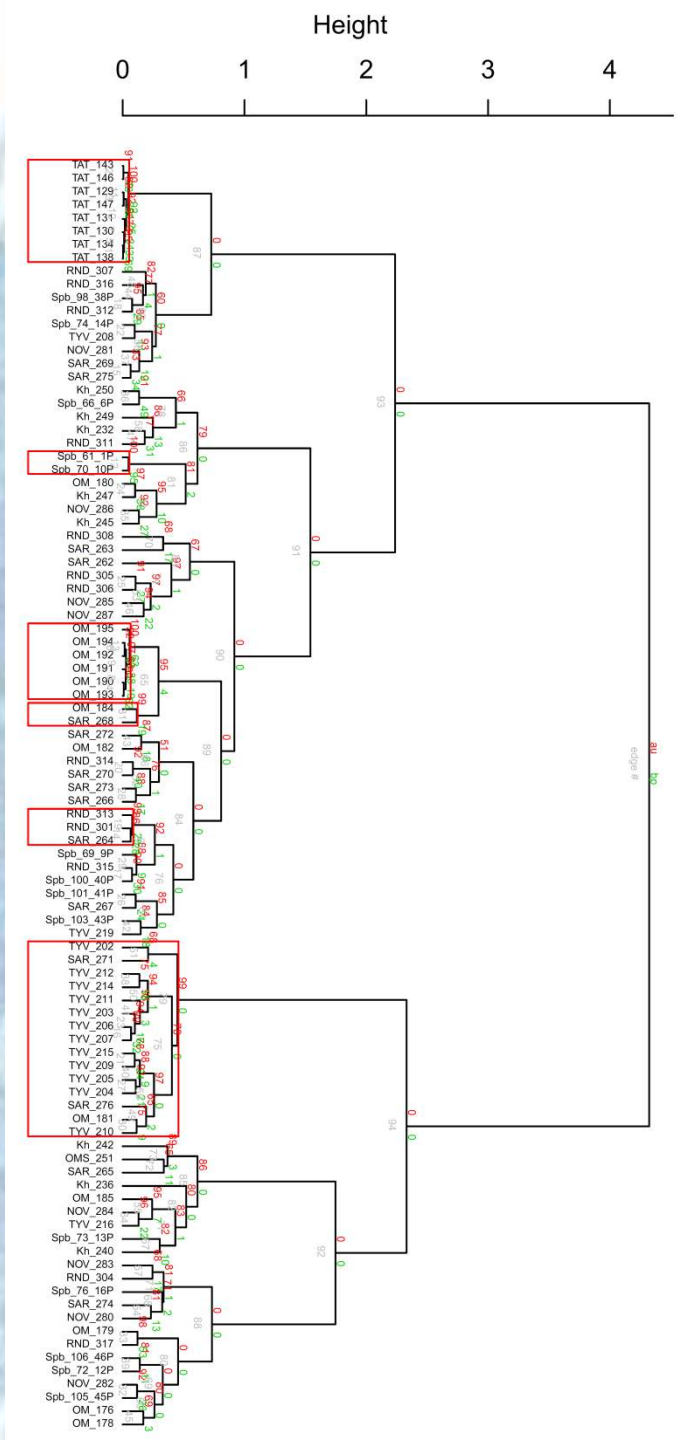
Пример: Влияние среды на микробиоту. Работники вредного производства (n=13), QIIME+R

Consensus Lineage	VR9	VR4	VR3	VR8	VR10	VR11	VR2	VR5	VR7	VR1	VR6	#OTU ID
k__Bacteria;p__Firmicutes;c__Clostridia;o__Coriobacteriales;f__Coriobacteriaceae;g__;s__	173		694	559	15		99	102				65
k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__[Coprobacillaceae];g__;s__	8	258	7	4	4	570	53		60	784		892
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Dorea;s__formicigenerans	271	231	67	13	11	549	13	78	1151	15		1188
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Dialister;s__invisus											2343	1524
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Lactococcus;s__garvieae		1784	1	1		6	2					1689
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__[Ruminococcus];s__	192		33		1	66	19	4	1081	75		1920
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae;g__;s__	725	38	147	359	33	152	242	55	166	294		2349
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Coprococcus;s__	16	334	22	2		35	34	25	1453	62	2	2467
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus;s__	1519	9121	737	12588	654	126	239	448		37	107	2663
k__Bacteria;p__Firmicutes;c__Clostridia;o__Coriobacteriales;f__Coriobacteriaceae;g__Collinsella	954		373	761	54		1171	75	544	208		3181
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae;g__Oscillospira;s__	876	30	953	2319	60	1256		115	3			3250
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus;s__ruminis	1005		275		6		1880	290		4772	65	3363
k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__[Eubacterium];s__biforme	366	1	639	81	9		890	32	1656	6706		3452
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae;g__;s__		1	1854									3632
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__Clostridium		56	4280	78	2461		12	195		147	17	3669
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Blautia;s__	394	725	57	58	4	42	237	106	22	262	238	3858
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus;s__		58		21	125		269	3		3844		3968
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Peptostreptococcaceae;g__;s__	4	41	1117	3643	507		167	39		1529	241	4202
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae		402	19	257	9631		3434	267	4	1	6322	4266
k__Bacteria;p__Firmicutes;c__Clostridia;o__Coriobacteriales;f__Coriobacteriaceae;g__;s__		975				59				327		4270
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Dorea;s__	279			52		89	9	2	2304		9	4910
k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__[Coprobacillaceae];g__Catenibacterium;s__	315		183		276		1268	245	2780			5222
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Blautia;s__obeum	101	1879	204	334	43	86	233	71	1250	408	24	5348
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus;s__	558	3398		20	6	1977	2	2			1	5464
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae	1483		13	70	5	15		4		9		5474
k__Bacteria;p__Tenericutes;c__Mollicutes;o__RF39;f__;g__;s__		1	3374							2		6363
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae	183	7	103	1	4	109	5	6	1216	23	55	6384
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Coprococcus;s__	217	258	141	1	1	2089	5	75	382	19	310	6450
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus;s__brevis		252	2	78	27		14	252			3605	6460
k__Bacteria;p__Firmicutes;c__Clostridia;o__Coriobacteriales;f__Coriobacteriaceae;g__Adlercreutzia;s__	47	751		5		46				521		7153
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__Clostridium	3	551	6917	119	590		250	32	1	236	68	8232
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Peptostreptococcaceae;g__;s__		1139	353	208		1	8	8			87	8542
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__;s__		190		4	1	8	3511	1	196	121	38	8895
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Blautia;s__	752	231	130	120	15	323	1592	85	2156	1367	12152	8900
k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__		81		3		3899		1	75			9055
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__Escherichia;s__	878	39	3	1	2	1049	27	20		46	5	9331
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__[Ruminococcus];s__gnavus	1095	1414	56	836	24	1972	1978	7	2450	1145	32	9979
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Leuconostocaceae;g__Weissella;s__				83	3		843	424	4		10301	10066
k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__[Eubacterium];s__cylindroides	1608				26			482				10493
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__luteciae		2		10	1		4558	40			4	10892
k__Bacteria;p__Firmicutes;c__Clostridia;o__Coriobacteriales;f__Coriobacteriaceae;g__;s__			122	348	106		202	41	1090			11432
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae	153	484	623	267	239	1427	108	28	65	594		11492
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__;g__;s__	1			79	1628	1	656	63	50		2	11702

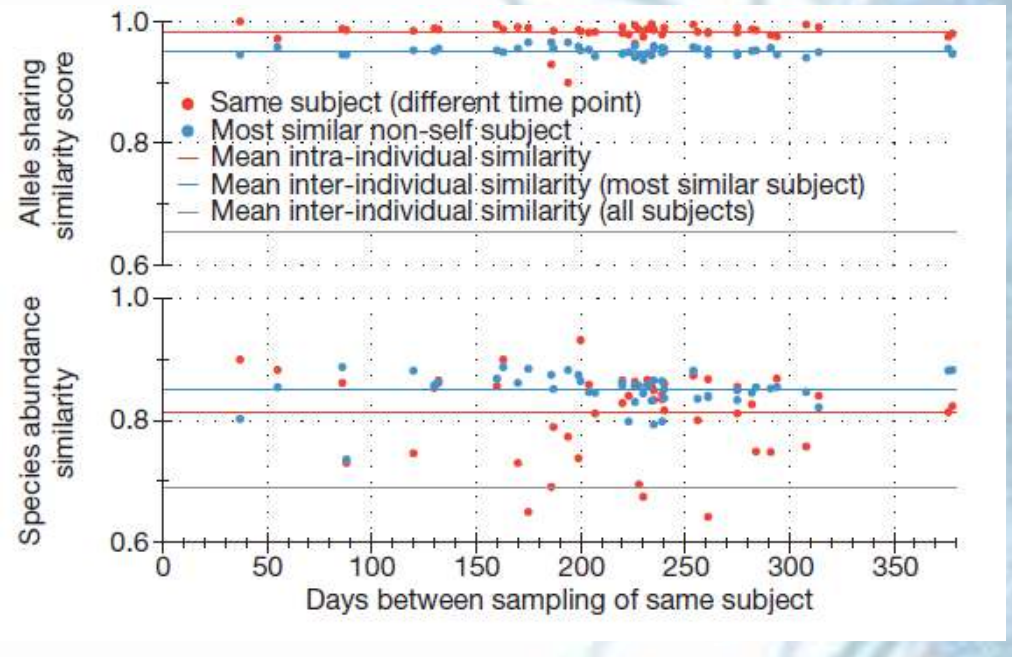
Пакет R: сравнение с внешней контрольной группой (Turnbaugh et al, 2010, 154 образца)




Bacteroides
Prevotella
Roseburia
Lachnospiraceae
Faecalibacterium
Alistipes
Coprococcus
Ruminococcus
Malawian
Amerindian
US
Danish
Russian
Chinese


Height
0
1
2
3
4

SNP variations in the metagenome



Temporal stability of genomic variation patterns.

Data: 139 (USA), 110 (Europe) metagenomic samples

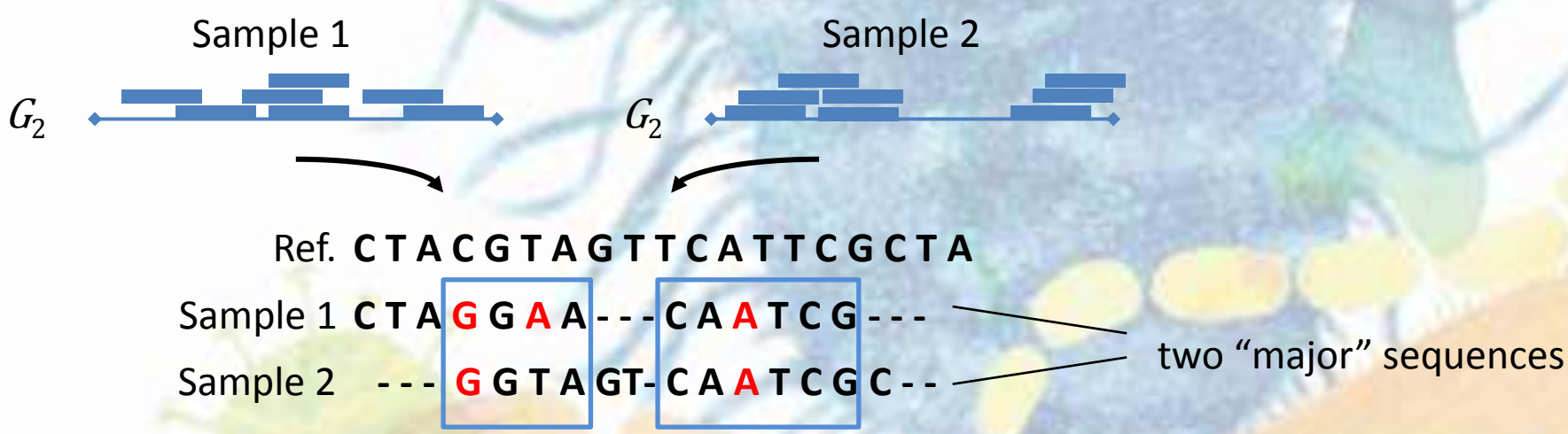
Distance measure: fixation index and nucleotide diversity

Results:

- Individual-specific variation patterns are stable over time
- Absence of clear continental stratification apart from small number of species.

From Schloissnig et al., Nature 2012

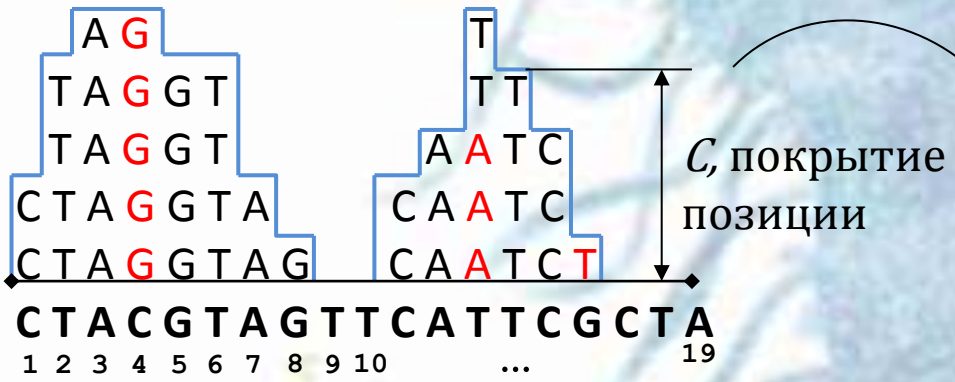
Our method of study of metagenome genetic variety



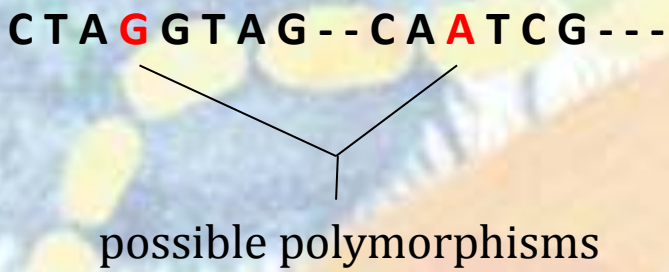
Sources of nucleotide differences in metagenomes

- True SNPs
- Mapped reads from other bacterial genome
- Errors of sequencing

So the main problem was proper choice of coverage, allele sharing and allele support thresholds to generate “major” sequence for each sample. In our opinion, distance between such major sequences is more robust characteristic than fixation index.



$$\begin{array}{c} C_A \\ C_T \\ C_G \\ C_C \end{array} \begin{pmatrix} 0 & 0 & 5 & 0 & 0 & \cdots & 0 \\ 0 & 4 & 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & 5 & 4 & \cdots & 0 \\ 2 & 0 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$



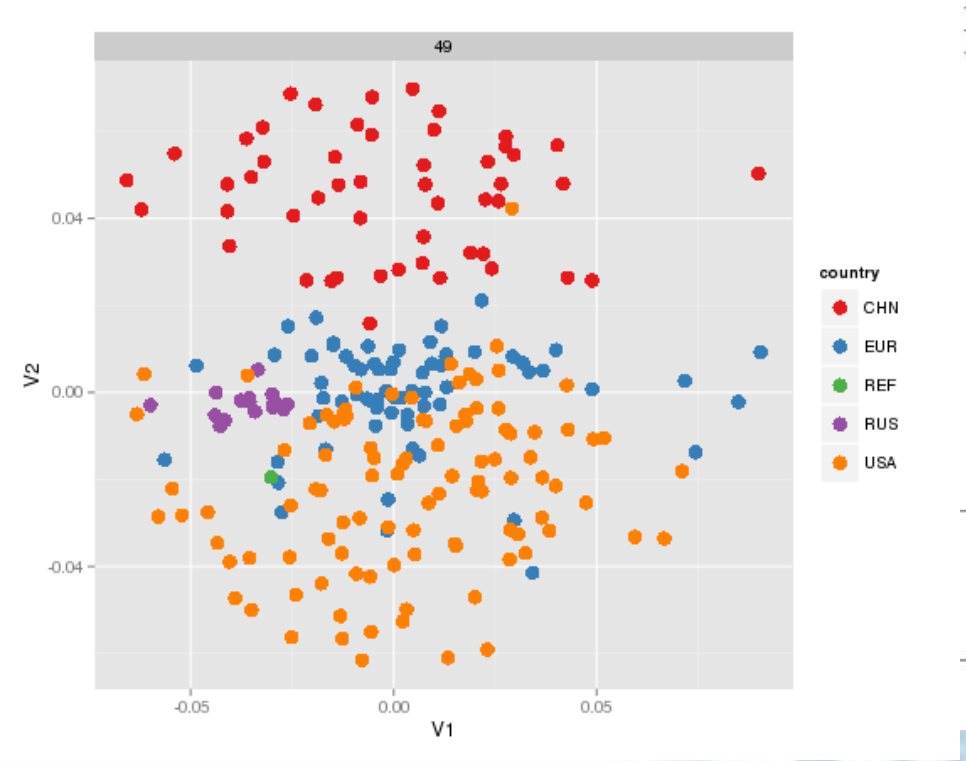
Data

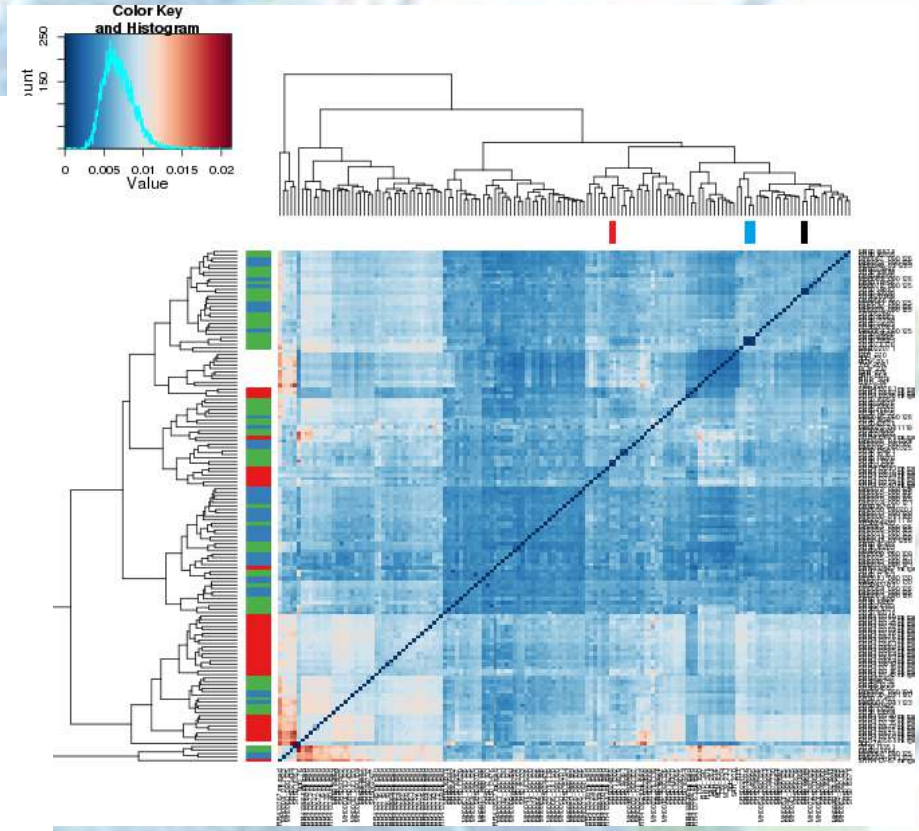
Host country	Source of the data	Number of samles	Number of individuals	Sequencing platform
USA	HMP	138	50 (sampled once) 41 (sampled twice) 2 (sampled tree times)	Illumina
Denmark	MetaHit	84	84	Illumina
China	BGI-Shenzhen	100	50 (type 2 diabet) 44 (healthy)	Illumina
Russia	RIPCM	20	20	SOLiD

Total: 342 gut matagenome samples of individuals from 4 countries.
Metagenome reads have been mapped on the catalog of 444 bacterial genomes

We have chosen 93 bacterial genomes that have abundance level greater than 1% at least in 50 individuals.

Result





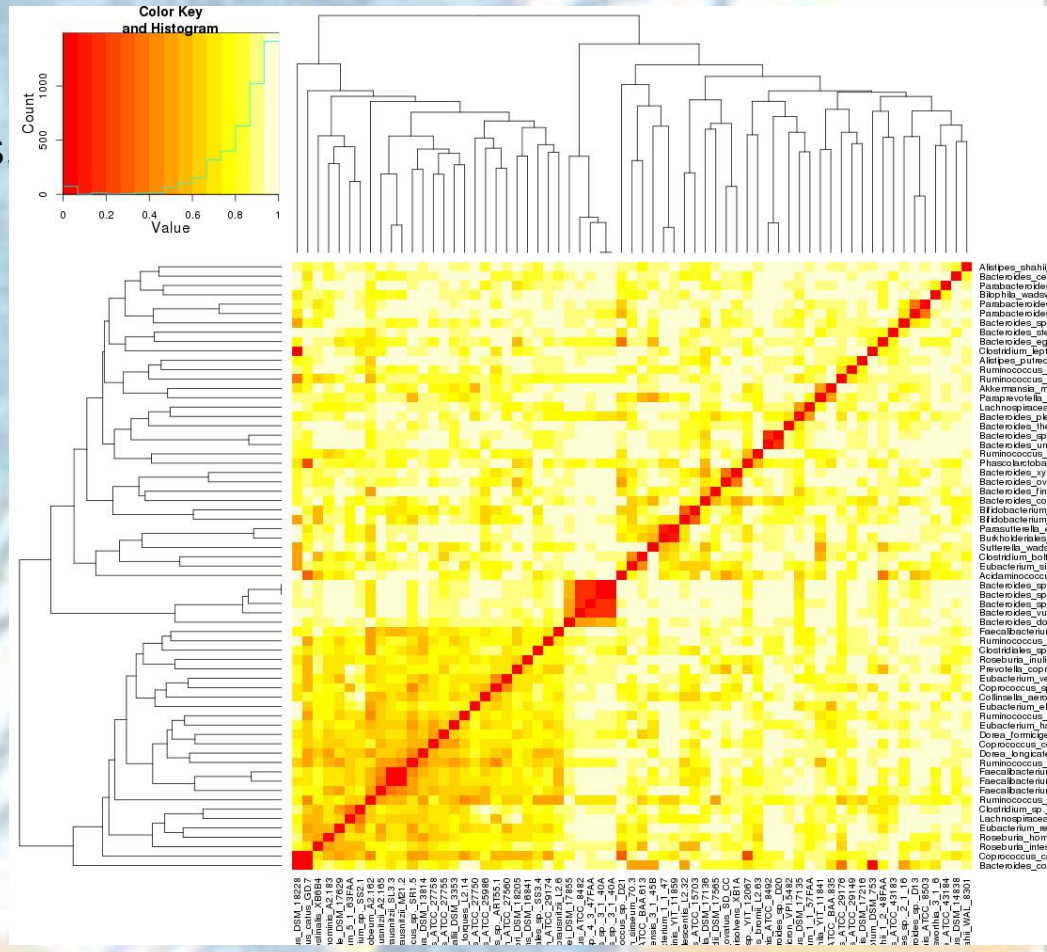
Example: *Faecalibacterium prausnitzii A2-165*

Consistency of results

On the heatmap we can see dense clusters of different strains of the same bacterial species:

- *Faecalibacterium prausnitzii* (4strains)
- *Bifidobacterium adolescentis* (2 strains
- *Bacteroides vulgatus* (2 strains.)

…and some others

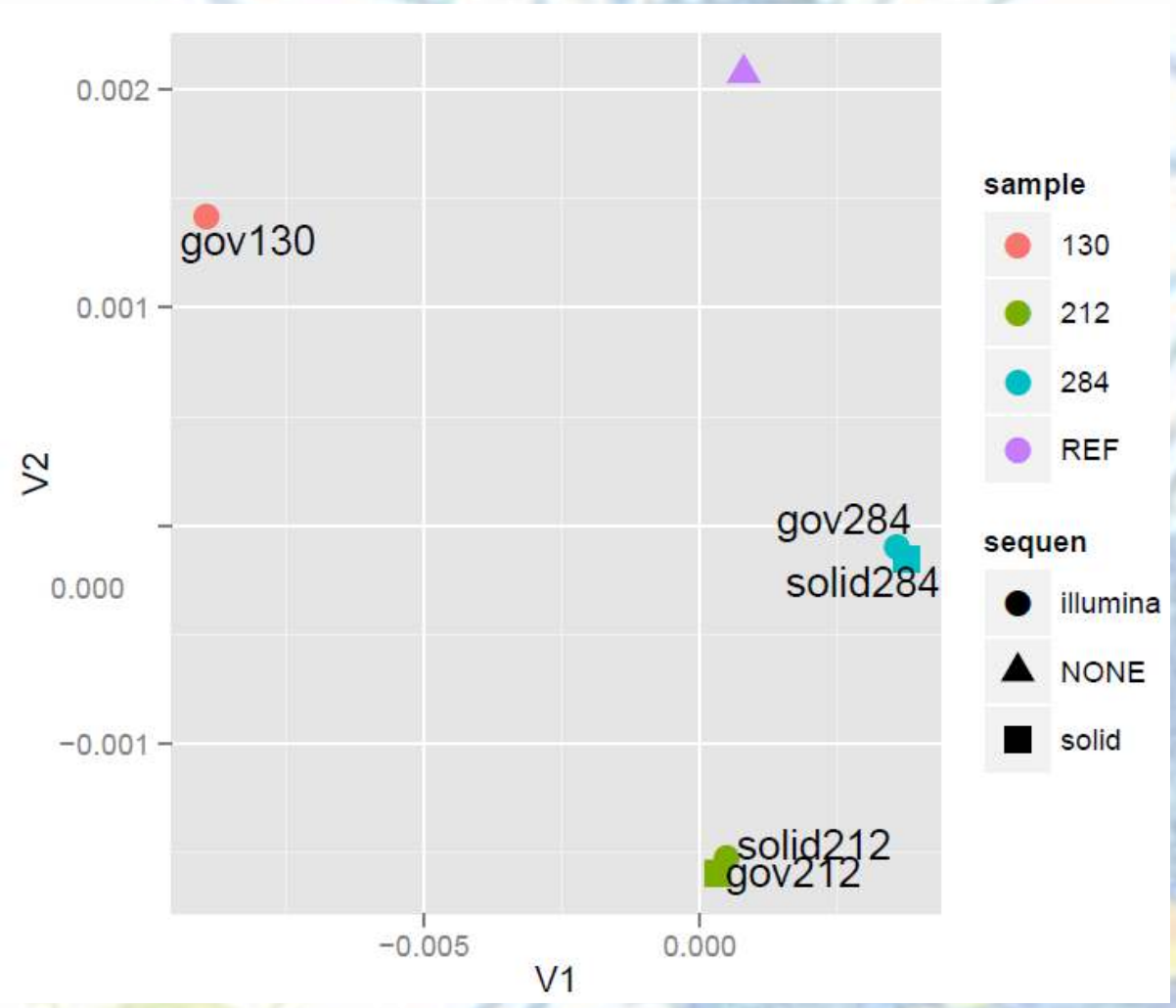


Similarity measure:

$$d(M_i, M_j) = 1 - |cor(M_i, Mj)|$$

$cor(M_i, Mj)$ - Mantel test Spearman correlation between distance matrices

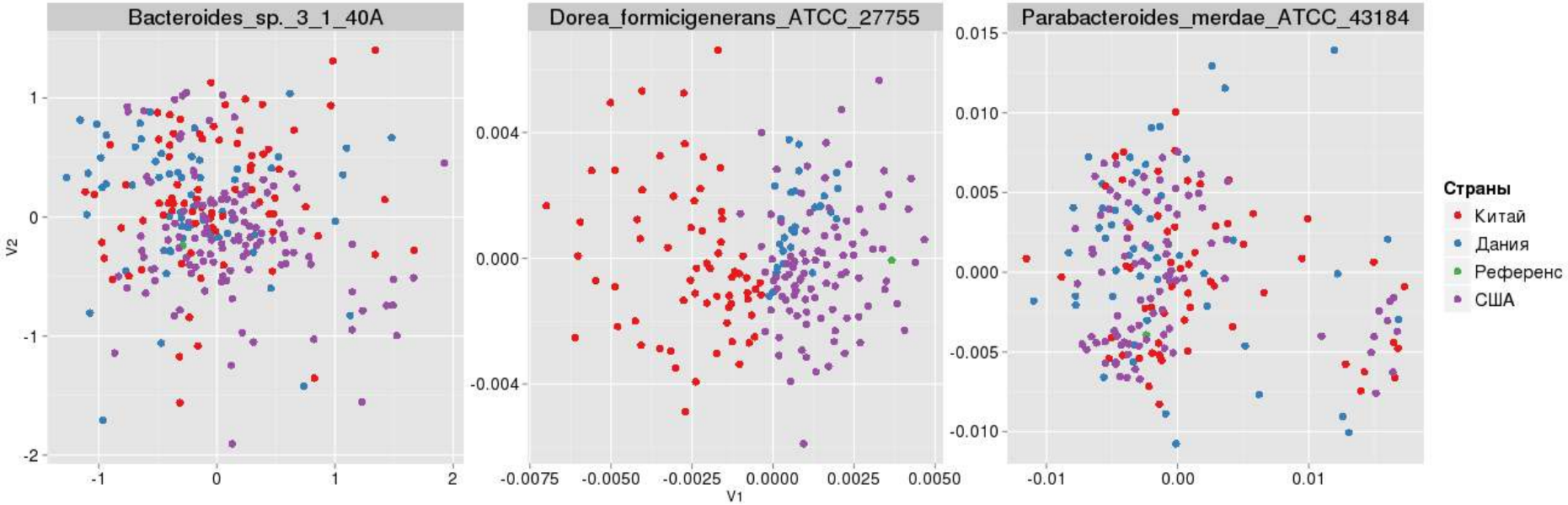
Dependence on sequencing platform



We've sequenced 5 metagenomic samples on different platforms.

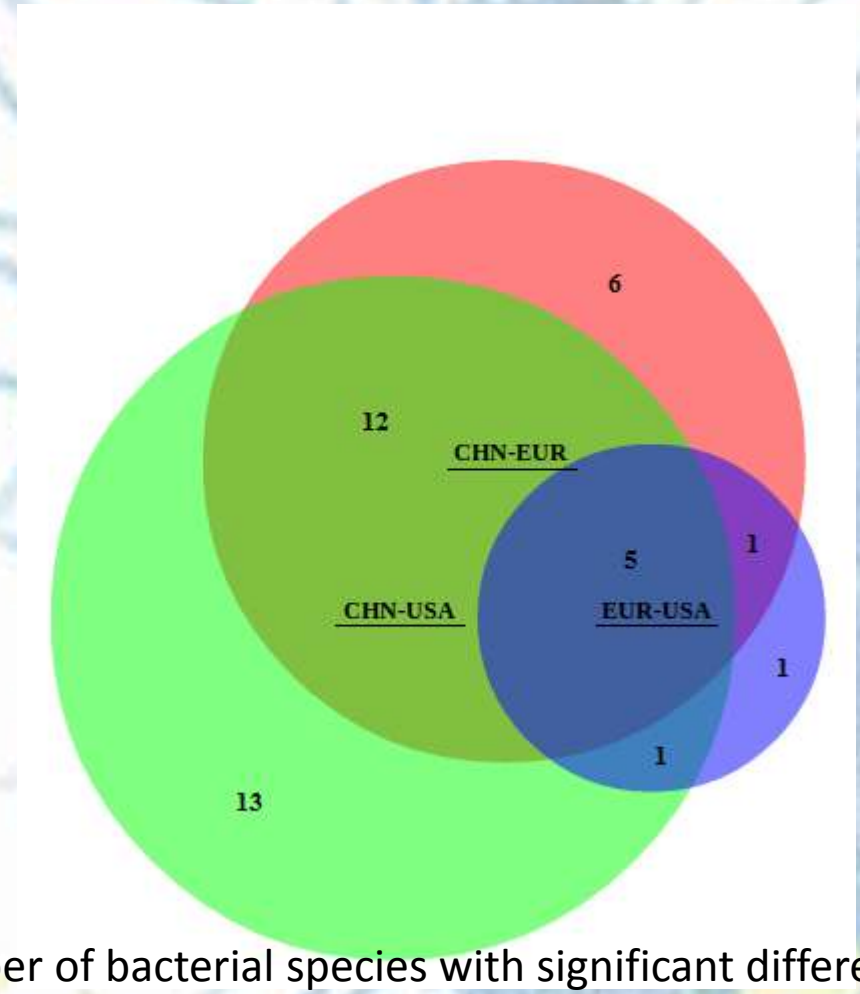
Stratification patterns of different bacterial species

Absence of any stratification Geographic factor Non-geographic factor



MDS-projections of distance matrices of different bacterial species.

Geographic stratification in different “directions”

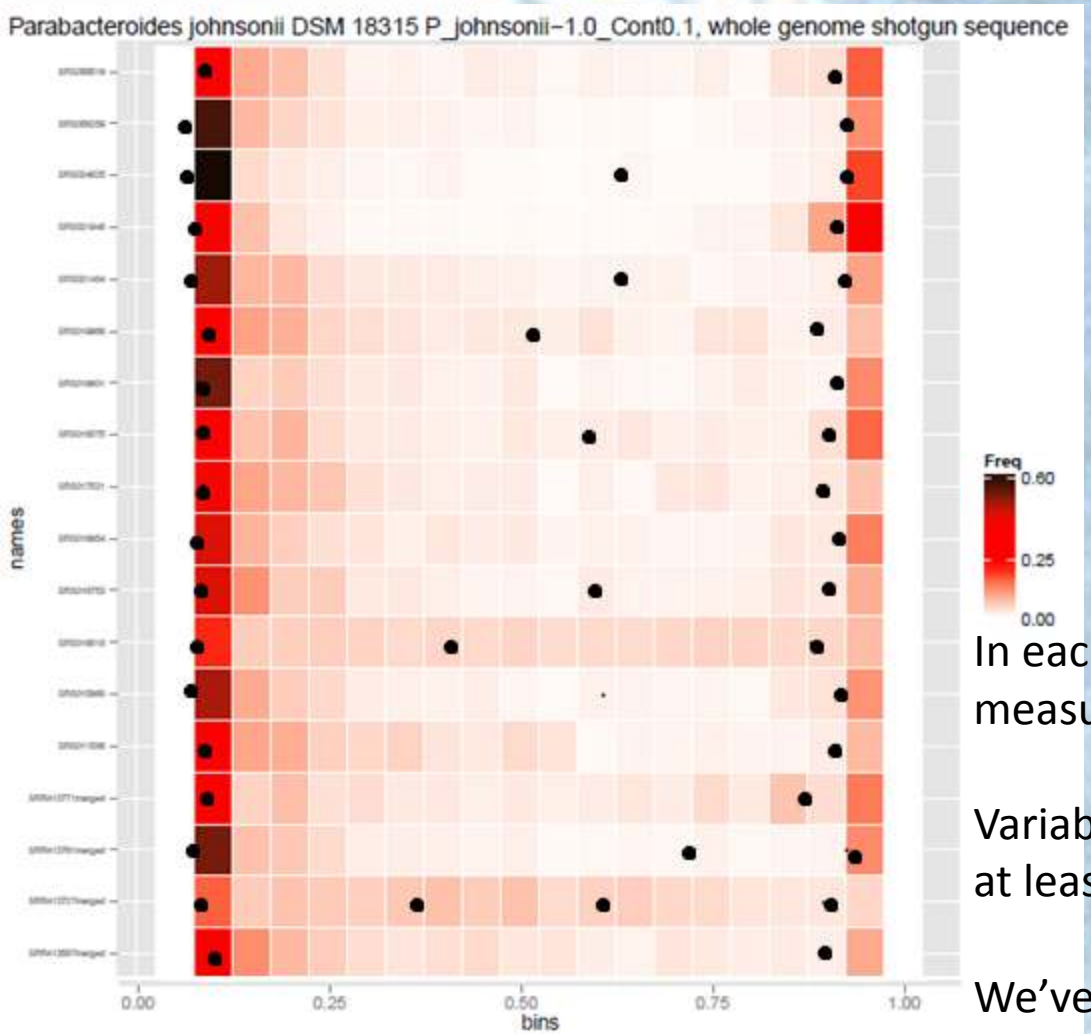


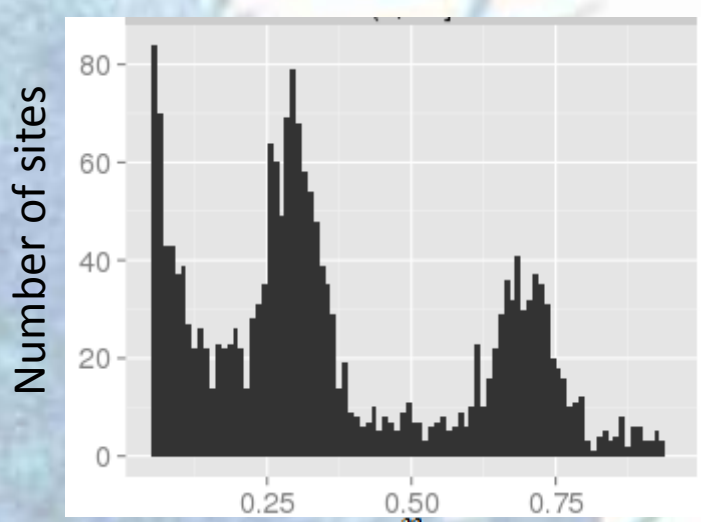
Circle means the number of bacterial species with significant difference in Wilcoxon test in particular “direction” (inter-country)

Total number of tested bacterial species: 65.

“Heterogenity” of variable sites in samples

Is it a manifestation of co-existence subpopulation of same bacterial species in the same organism?





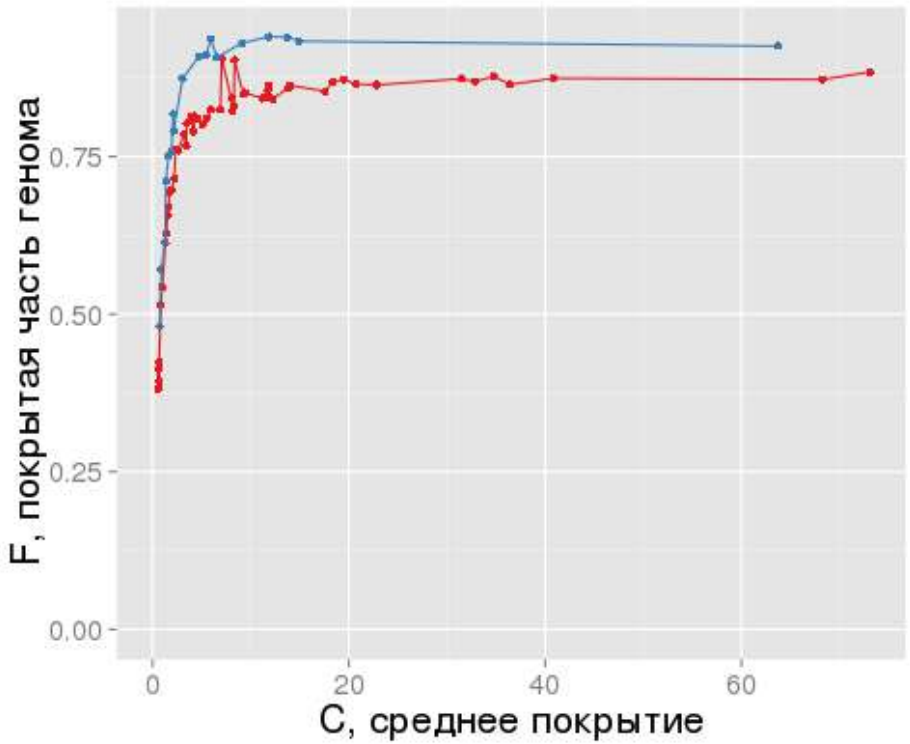
$$\frac{n_r}{n_d + n_r}$$

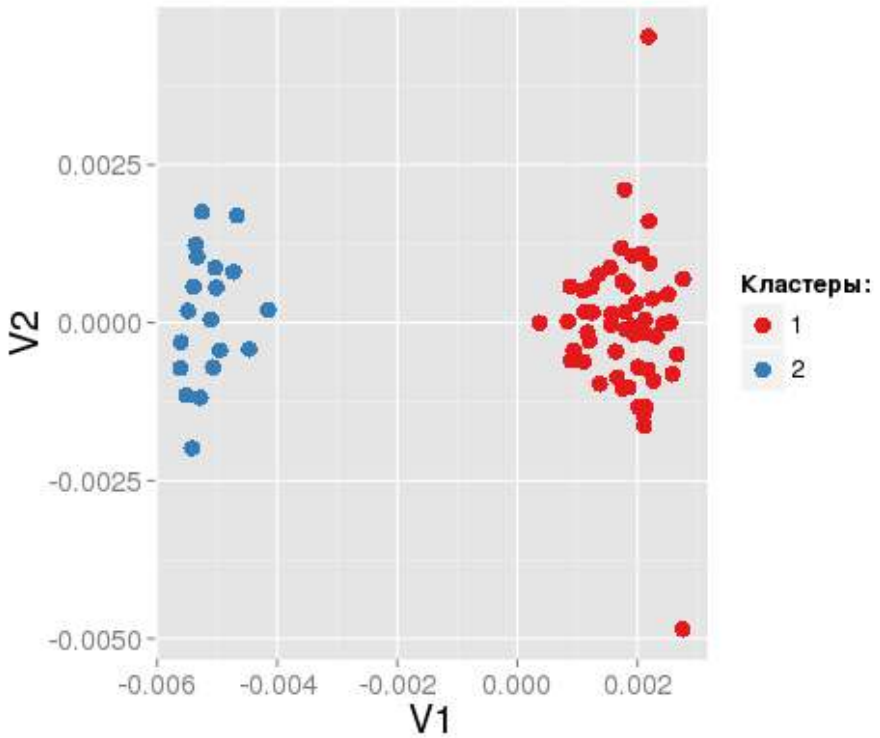
In each variable site of all samples we have measured “recessive” allele fraction.

Variable site means that it differs from reference at least N of M samples.

We’ve used different thresholds: {N, M}

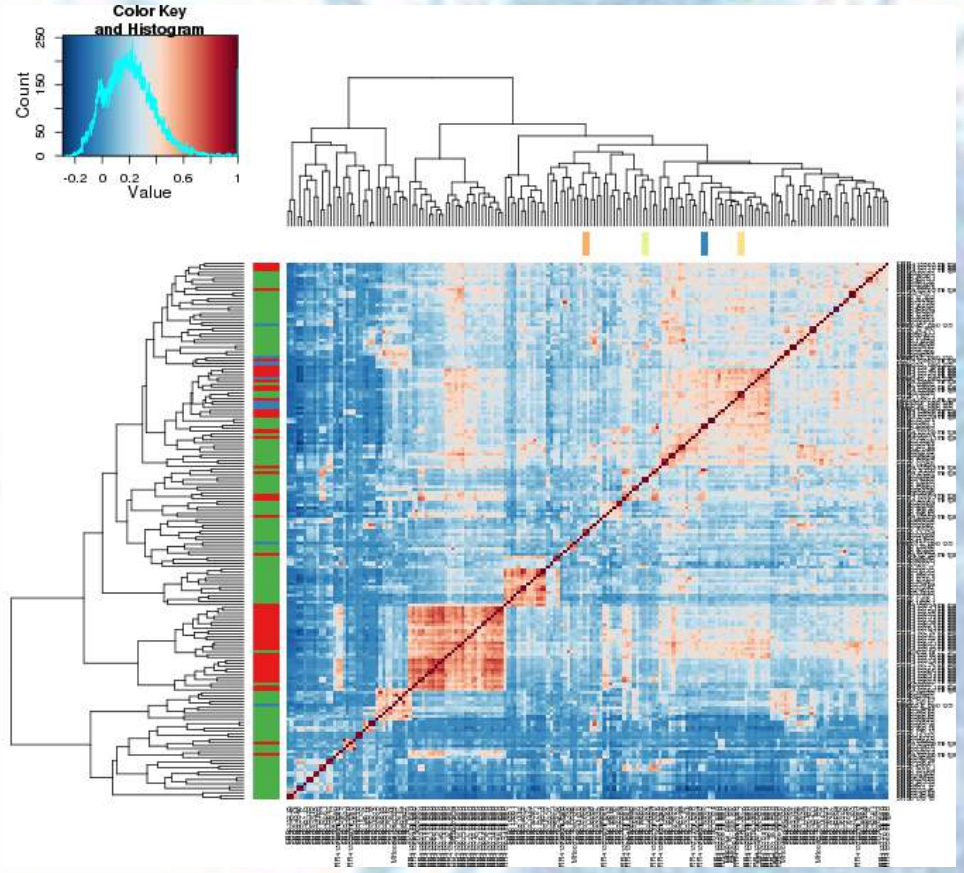
Polymorphism-level difference is connected with gene composition difference between samples





$$F \approx \beta(1 - e^{C}) = \beta\left(1 - \exp\left(\frac{NL}{G}\right)\right)$$

Correlation between coverage profiles of different samples of same species



We have chosen samples with coverage ≥ 5. Then

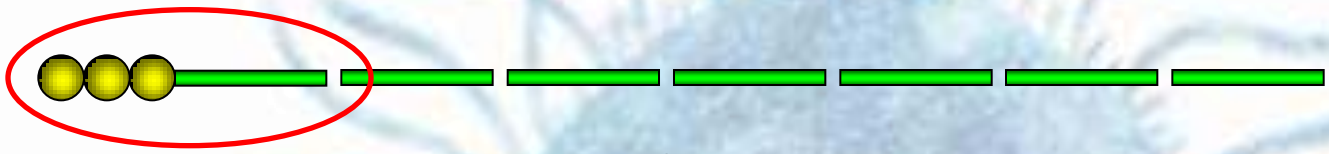
Разработка метода картирования сайтов инициации транскрипции с помощью технологии секвенирования второго поколения на примере *Mycoplasma gallisepticum*

Фисунов Г.Ю., Горбачёв А.Ю., Семашко Т.А., Мазин П.В.

Зачем искать сайты инициации транскрипции

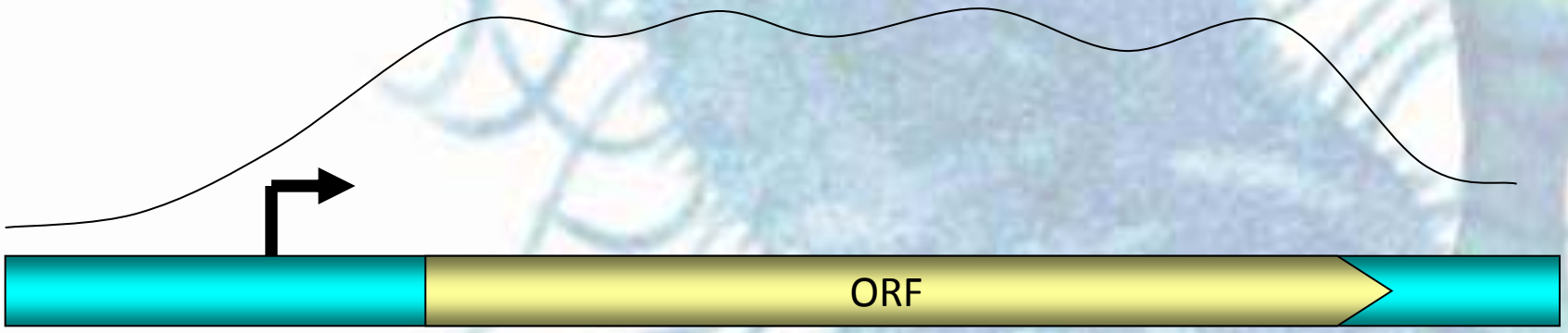
- Поиск сайтов связывания транскрипционных факторов

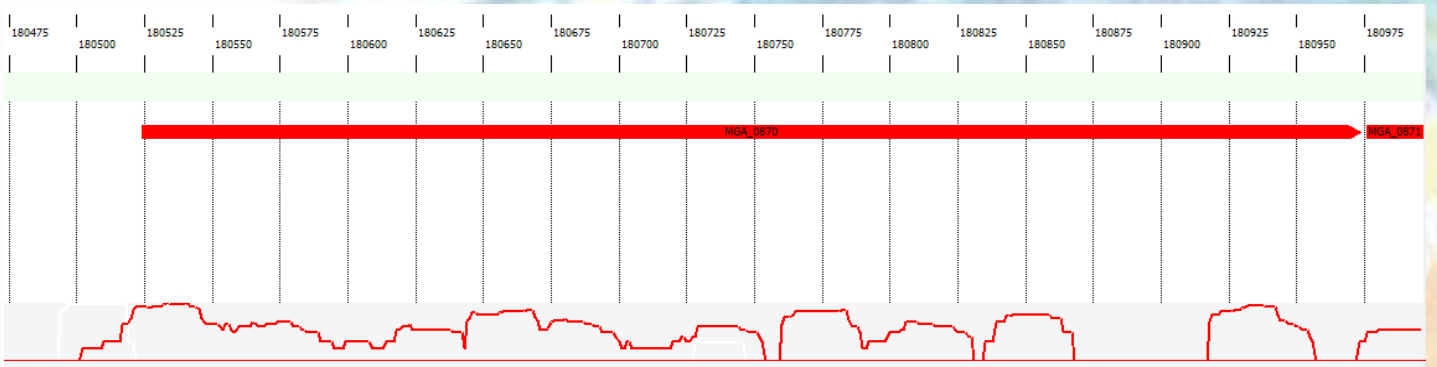
Технологическая проблема



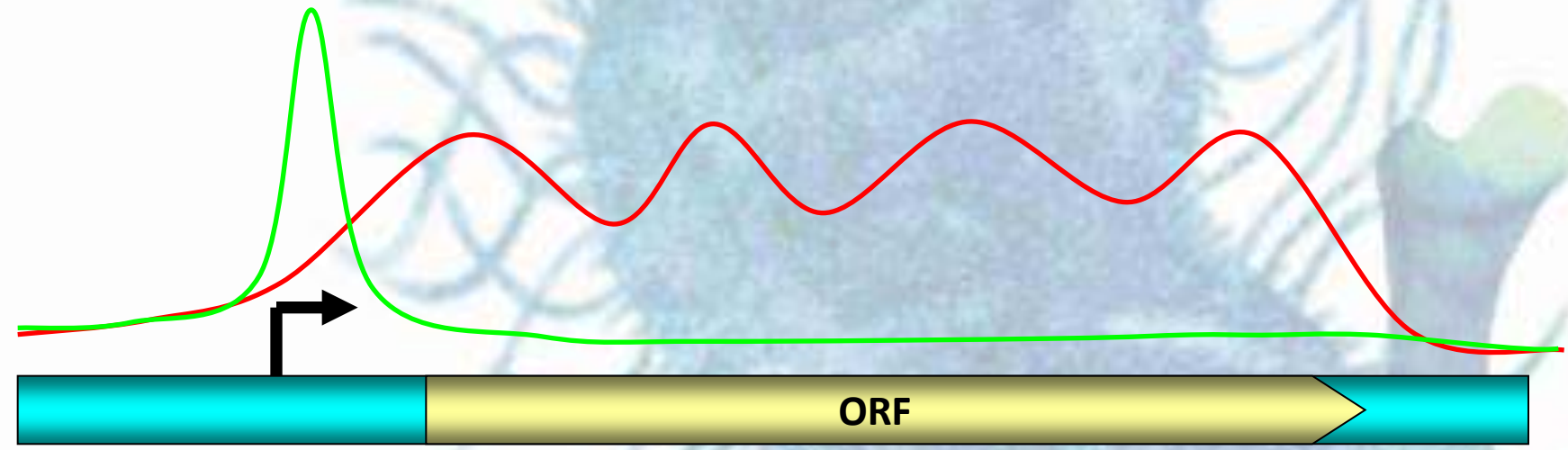
мРНК после фрагментации

Покрытие при стандартном приготовлении библиотеки





Идея метода



Покрытие при стандартном приготовлении библиотеки

Покрытие 5`-обогащённой библиотеки

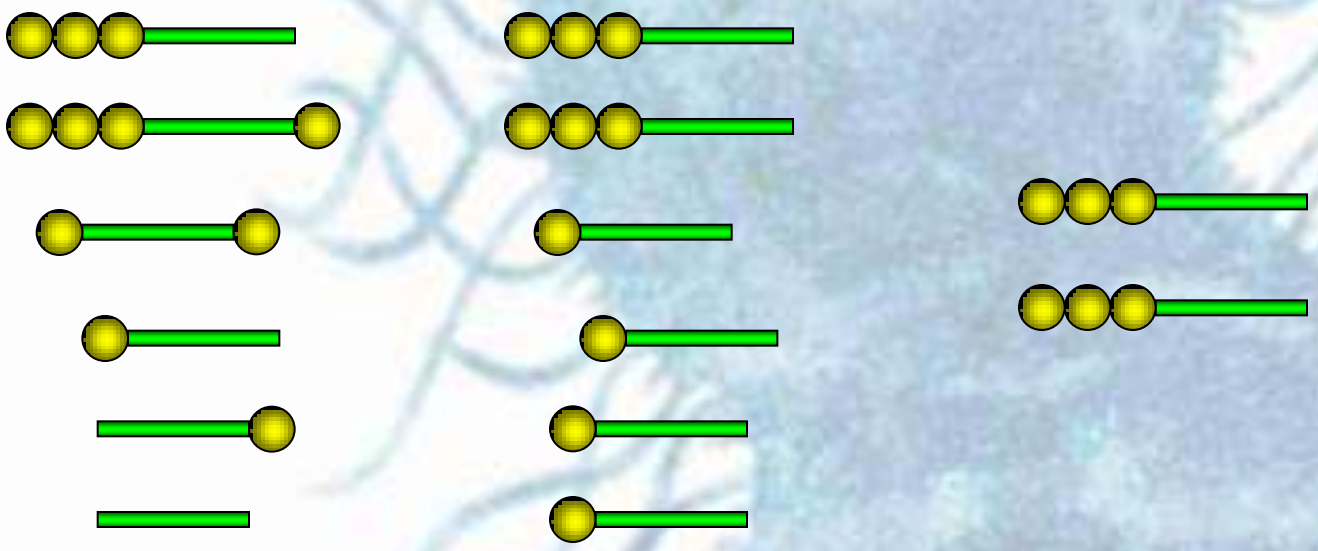
Шаг 1 - фрагментация

= Фосфат

Шаг 2 – обогащение 5`-последовательностями

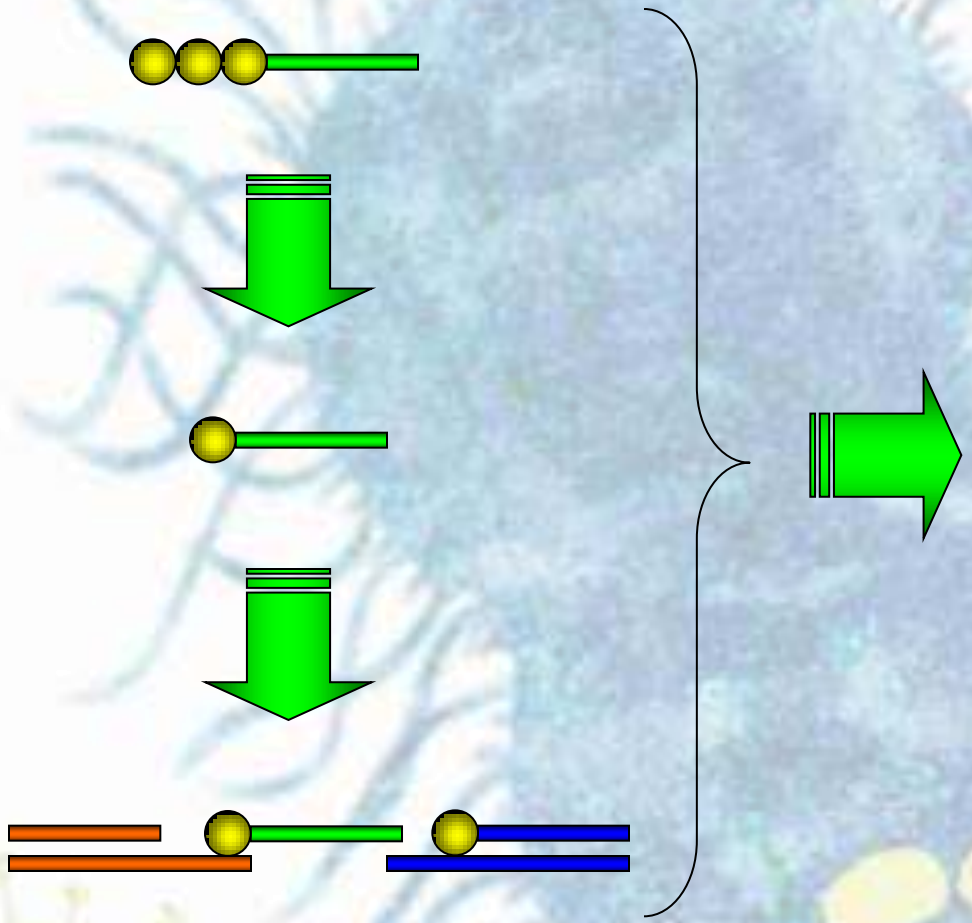
Починка концов

Обработка 5`-фосфат-зависимой нуклеазой



= Фосфат

Шаг 3 – снятие защитной группы и лигирование адаптеров

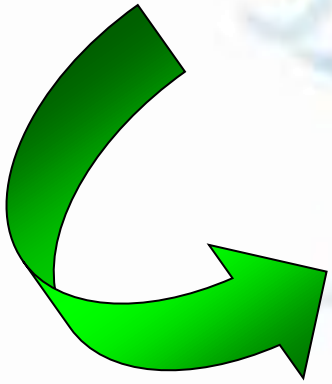


Синтез кДНК и нормализация библиотеки

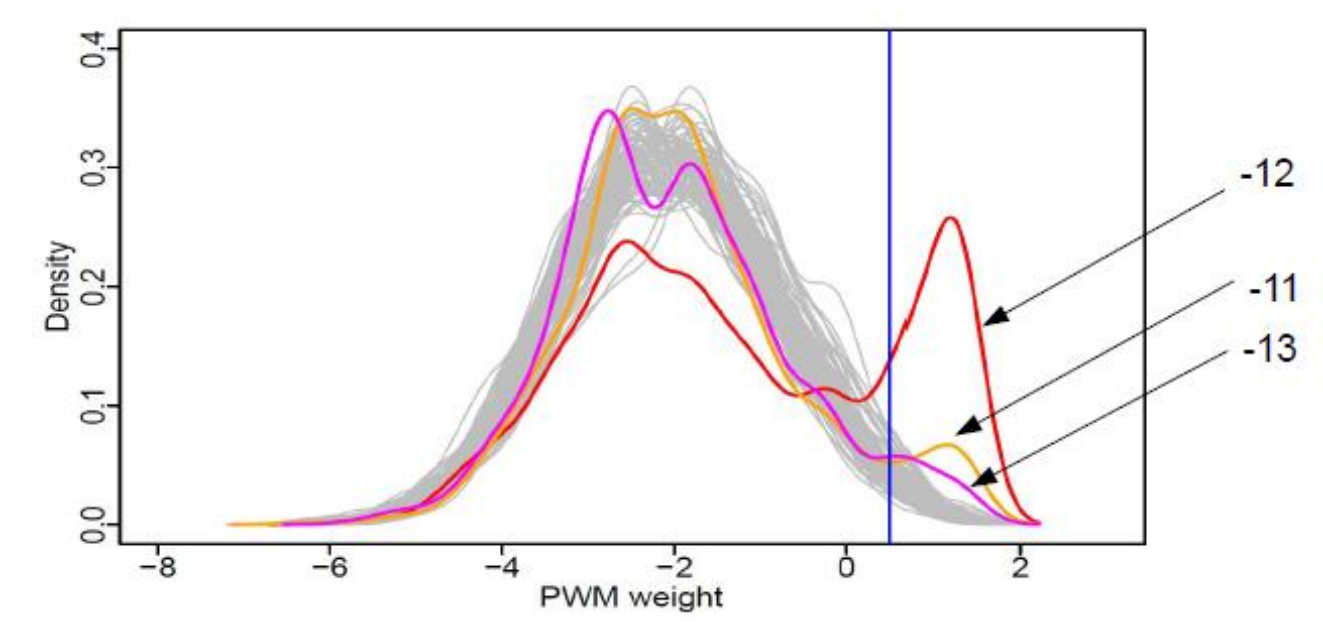
= Фосфат

Шаг 4 - секвенирование

Шаг 5 – анализ результатов

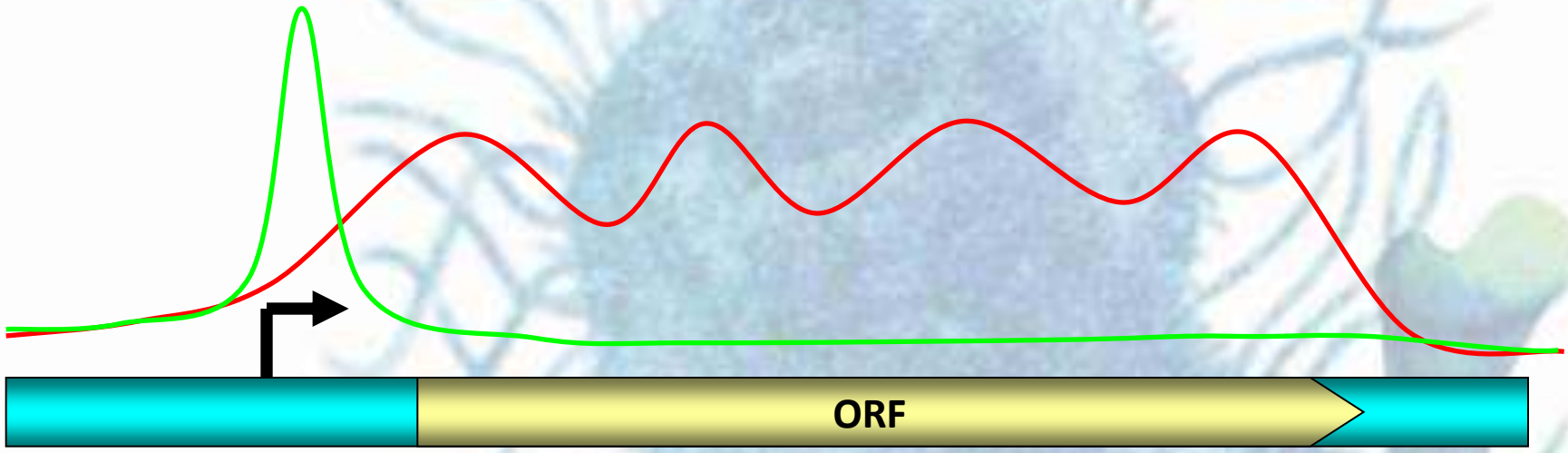


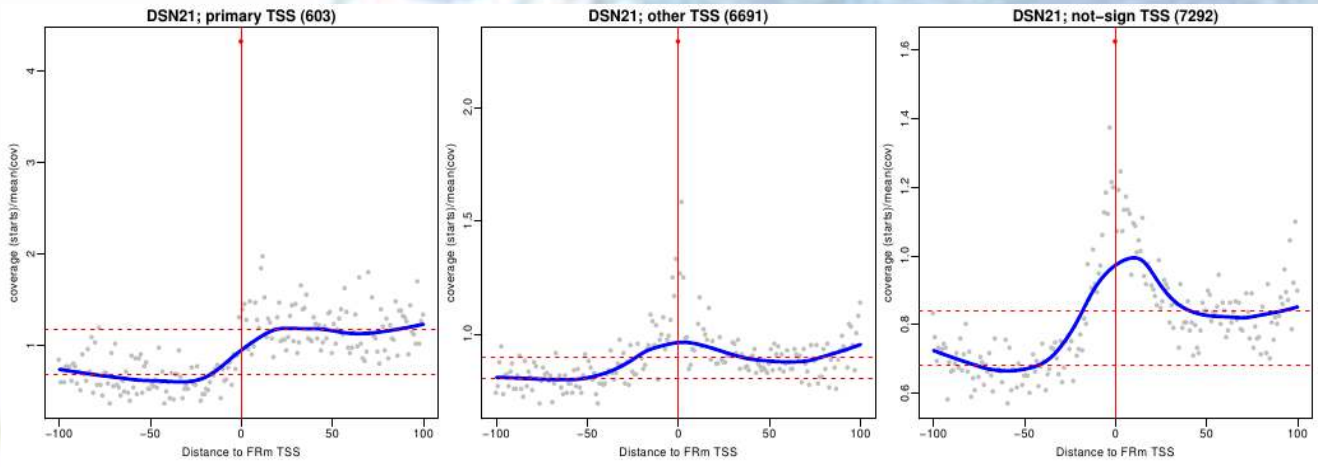
Веса TATA-боксов, расположенных в разных позициях относительно старта



Шаг 5 – анализ результатов

Проверка по наличию ступеньки в покрытии



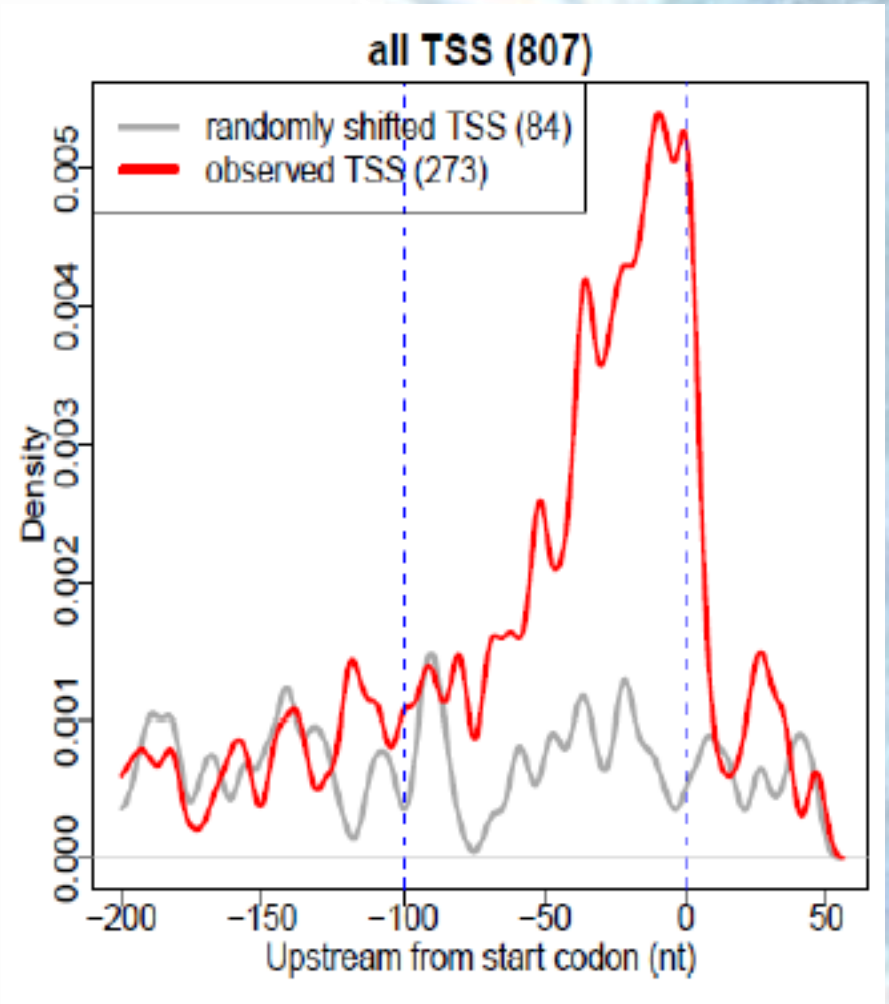


Покрытие по + цепи

Покрытие по - цепи

Покрытие по + цепи с обогащением 5`- концами

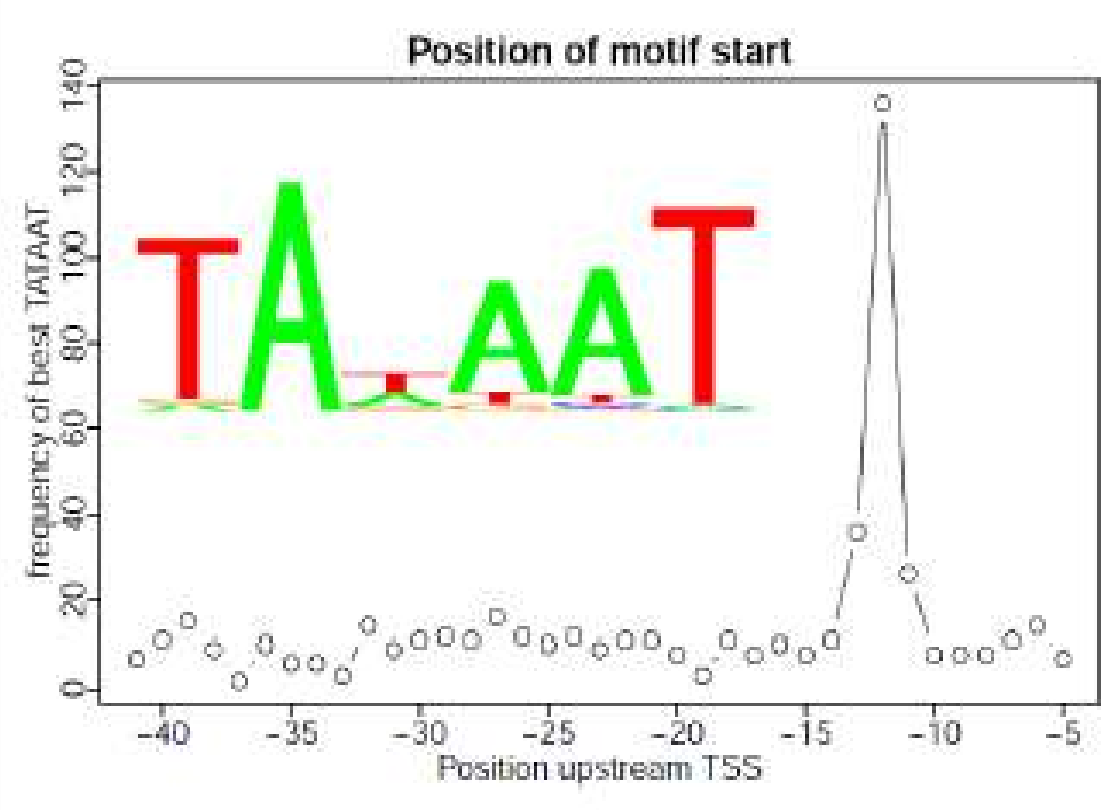
Результаты – как соотносятся старты транскрипции (TSS) и ORF (длина 5`-UTR)



TSS в среднем располагается от 0 до -50 bp от старта рамки

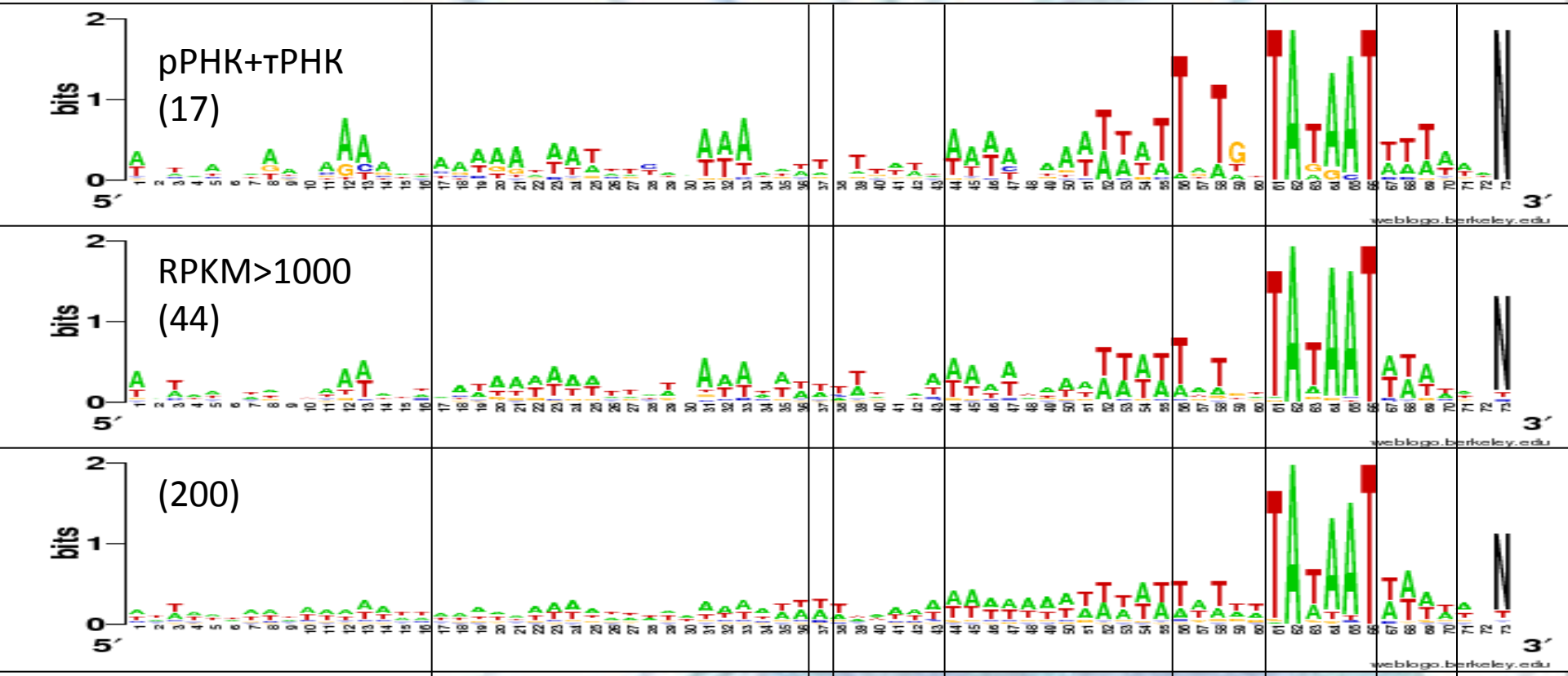
Результаты

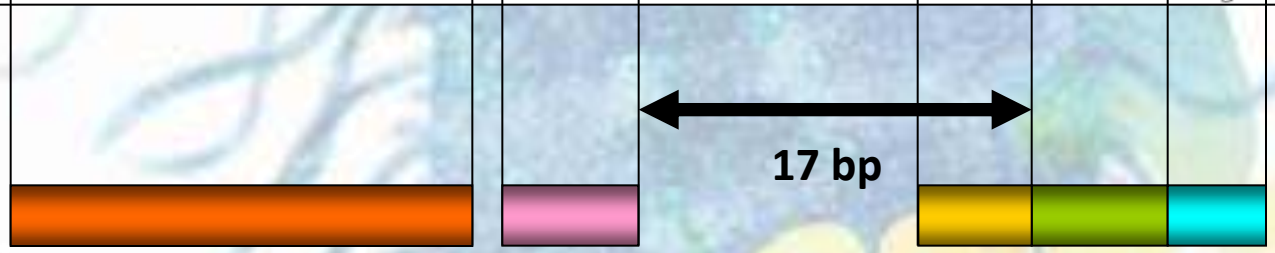
Всего идентифицировано 807 TSS
200 TSS имеют TATA-бокс, стоят вблизи рамки и имеют ступеньку в покрытии



TATA-бокс стоит на позиции -12 от старта транскрипции

Структура микоплазменного промотора





UP-element

-35

TTGACA

Ext

TRTGN

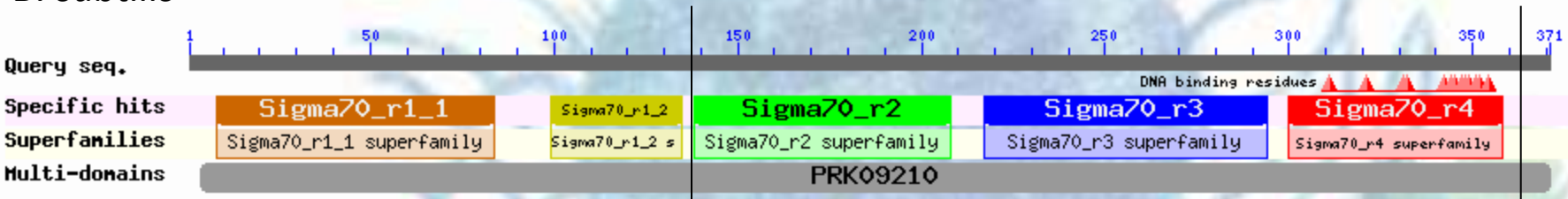
-10

Dis

GCGC

Строение микоплазменного сигма-фактора

B. subtilis



M. gallisepticum

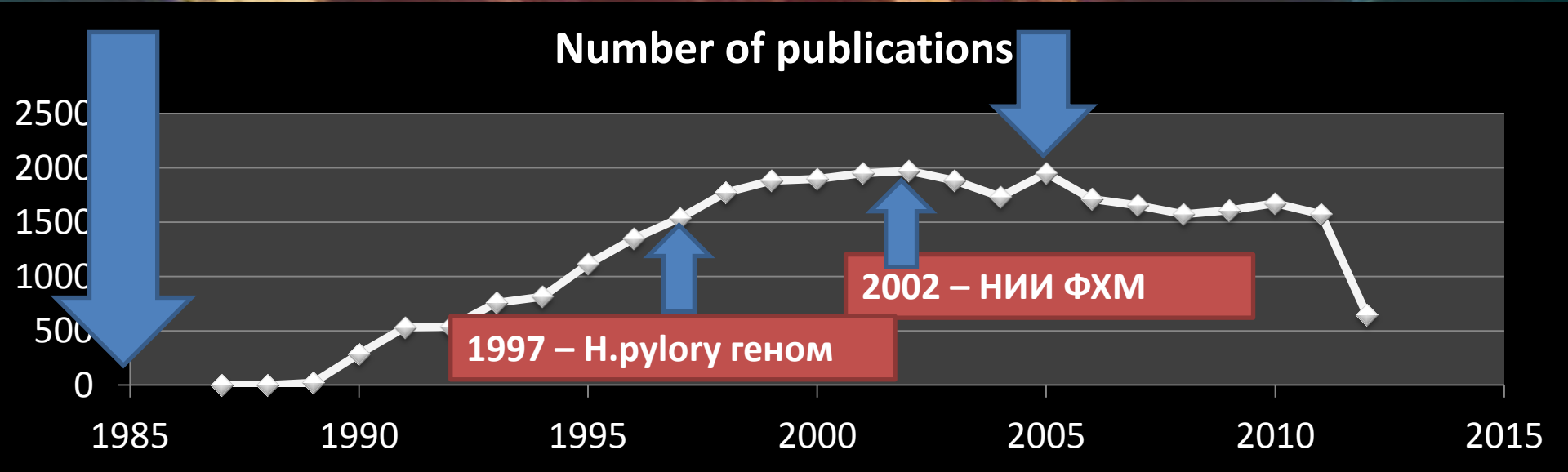
Dis **-10-box** **Ext** **-35-box**

Выводы

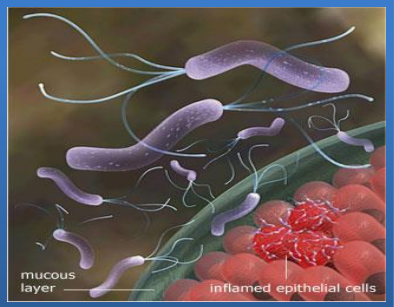
- Метод работоспособен
- Ключевые элементы микоплазменного промотора – TATA-бокс и Ext-элемент
  - Консенсус - WWWTTRTGNTAWAATWWW
- Консервативный -35 элемент отсутствует
- Длина 5`-UTR порядка 50 bp

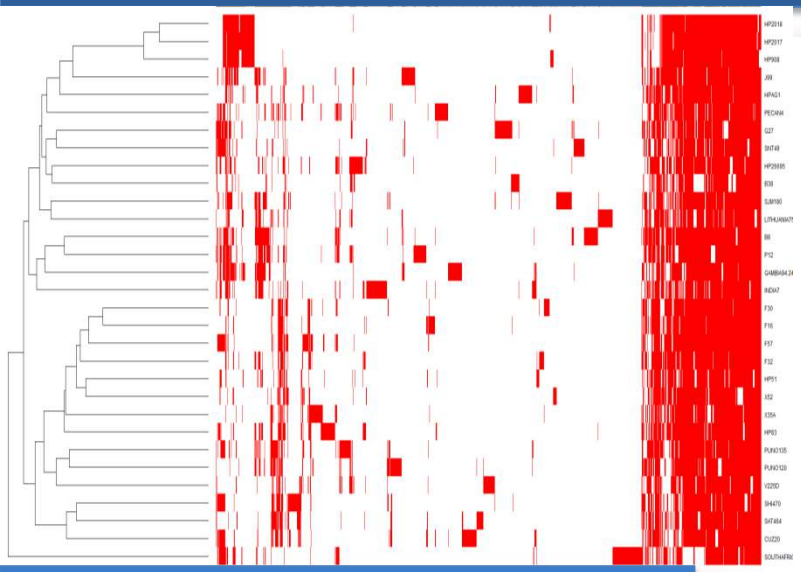
Открытие бактерии
J Robin Warren, Barry Marshall

Нобелевская премия

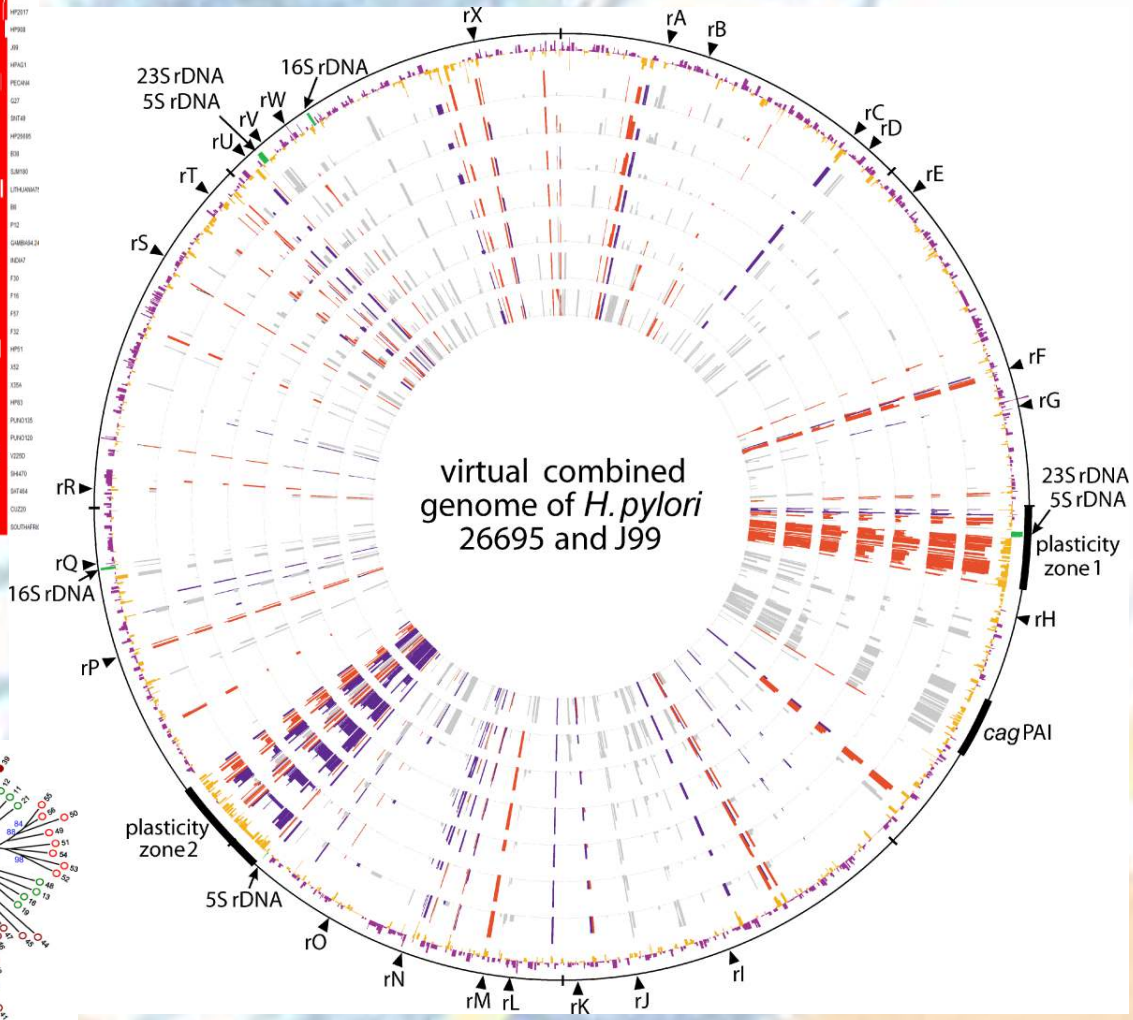


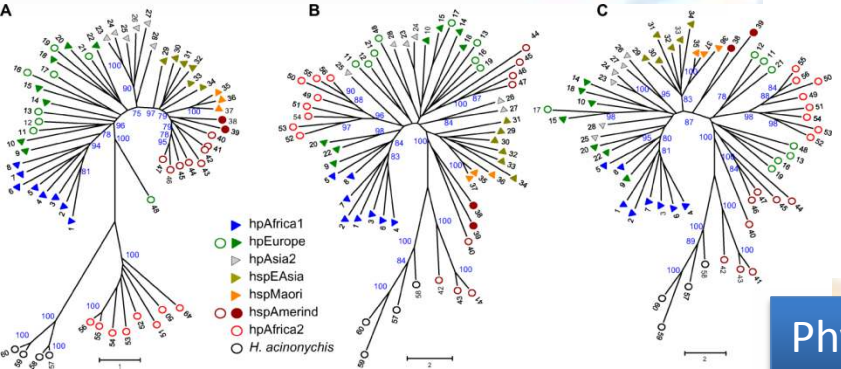
Micro and macro heterogenity
Helicobacter pylori





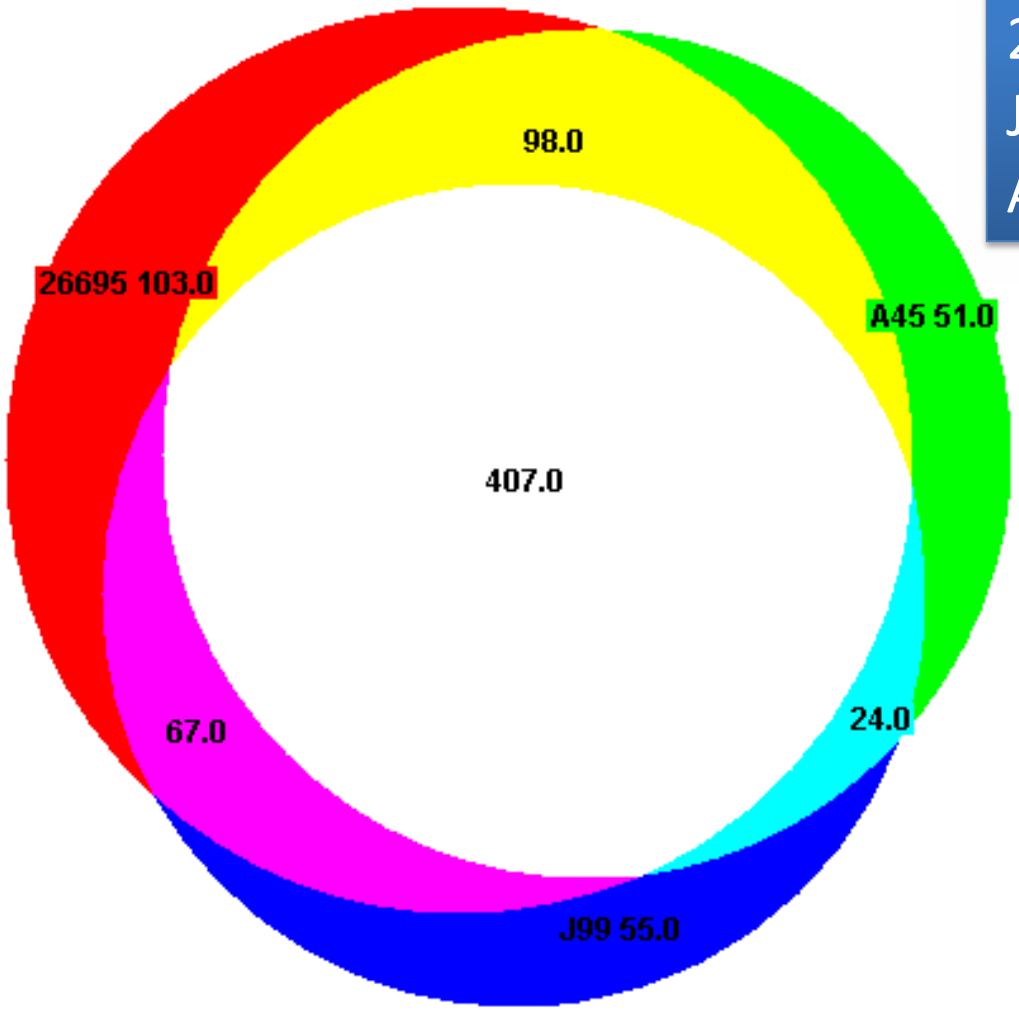
3186 genes in 31 strain having 1051 shared gene



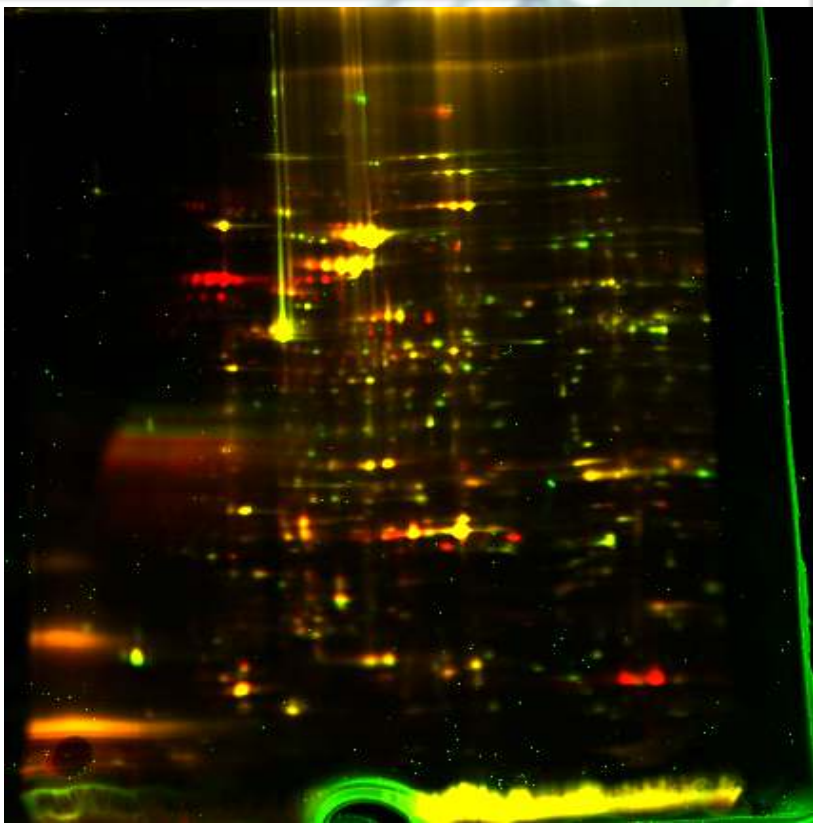


Phylogeny based on different genes

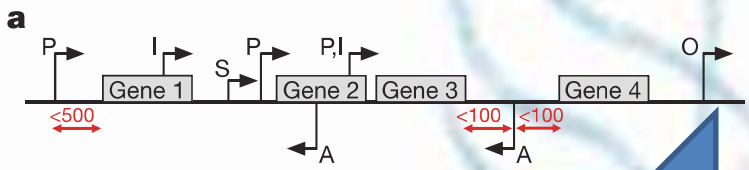
Протеомные карты *Helicobacter pylori* – первый взгляд

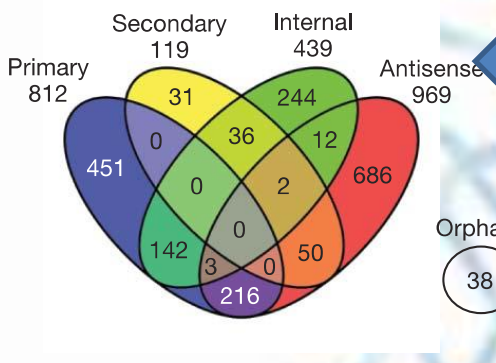


26695 – 675 белков
J99 – 553 белков
A45 – 580 белков



Транскриптомное профилирование бактерий

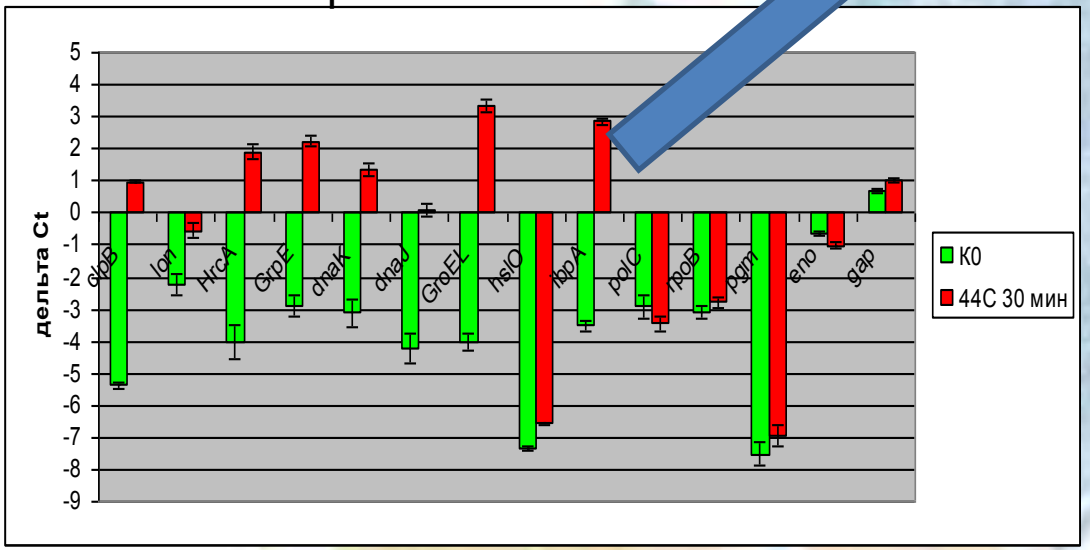




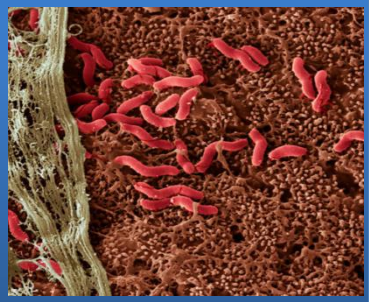
Более половины РНК - Антисмысловое

Полное транскрипционное профилирование

ПЦР



Метилирование
Helicobacter pylori - Phasevariation

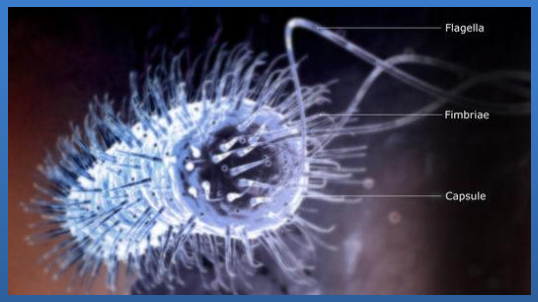


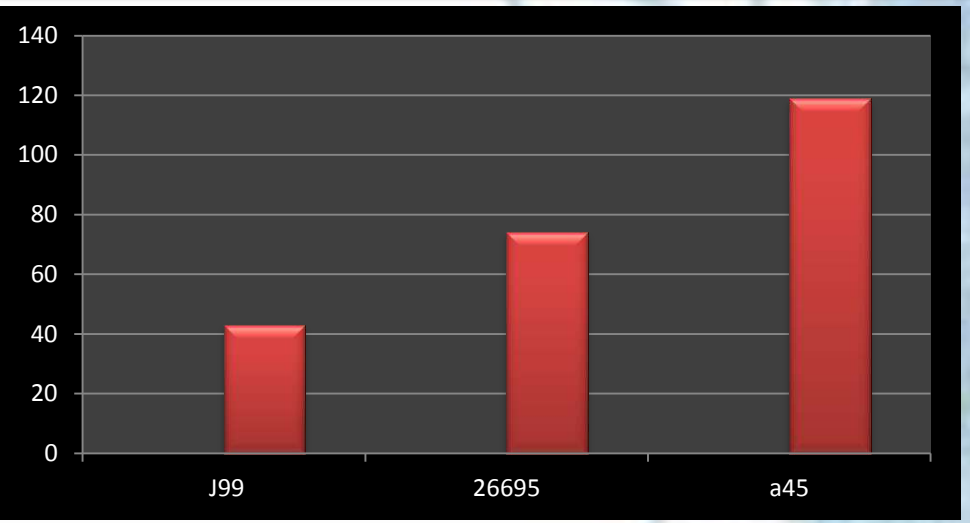
a45	J99	26695
ACGT	ACGT	ACGT
	ACNGT	ACNGT
ATTAAT	ATTAAT	ATTAAT
CATG	CATG	CATG
CCATC		
CCGG	CCGG	CCGG
	CCNNGG	
CCTC	CCTC	CCTC
CCTTC		CCTTC
	CGWCG	
		CTGCAG
		GAAGA
GANTC	GANTC	GANTC
GATC	GATC	GATC
GCGC	GCGC	GCGC
GGCC		
GRRG	GRRG	GRRG
GTAC	GTAC	GTAC
GTNNAC	GTNNAC	GTNNAC
	GTSAC	
TCGA	TCGA	TCGA
TCNGA		
TCNNGA	TCNNGA	TCNNGA
TGCA		TGCA

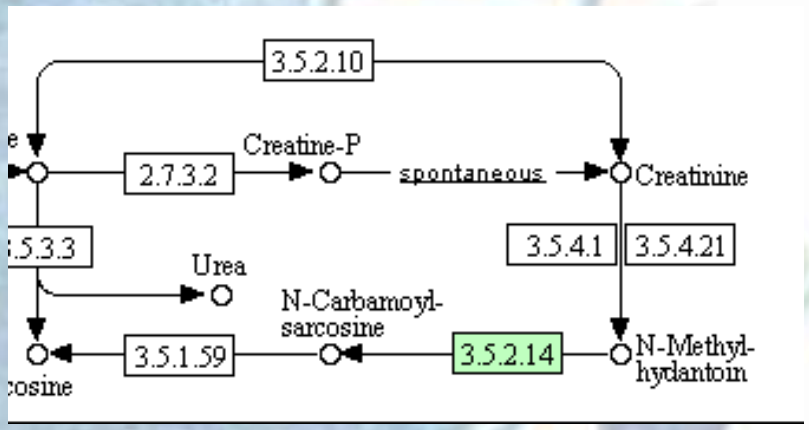
Although we cannot exclude another as yet undescribed
role for these modH phase variation in H. pylori biology, we have
confirmed phasevarion mediated epigenetic mechanism of gene
expression does operate in H. pylori -
Phasevarion Mediated Epigenetic Gene Regulation in
Helicobacter pylori – Plos One – 12.2011

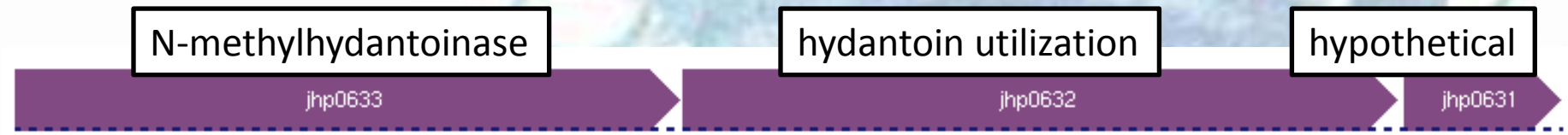
Метилирование
Helicobacter pylori - Phasevariation

Motive TGCA – 158 protein









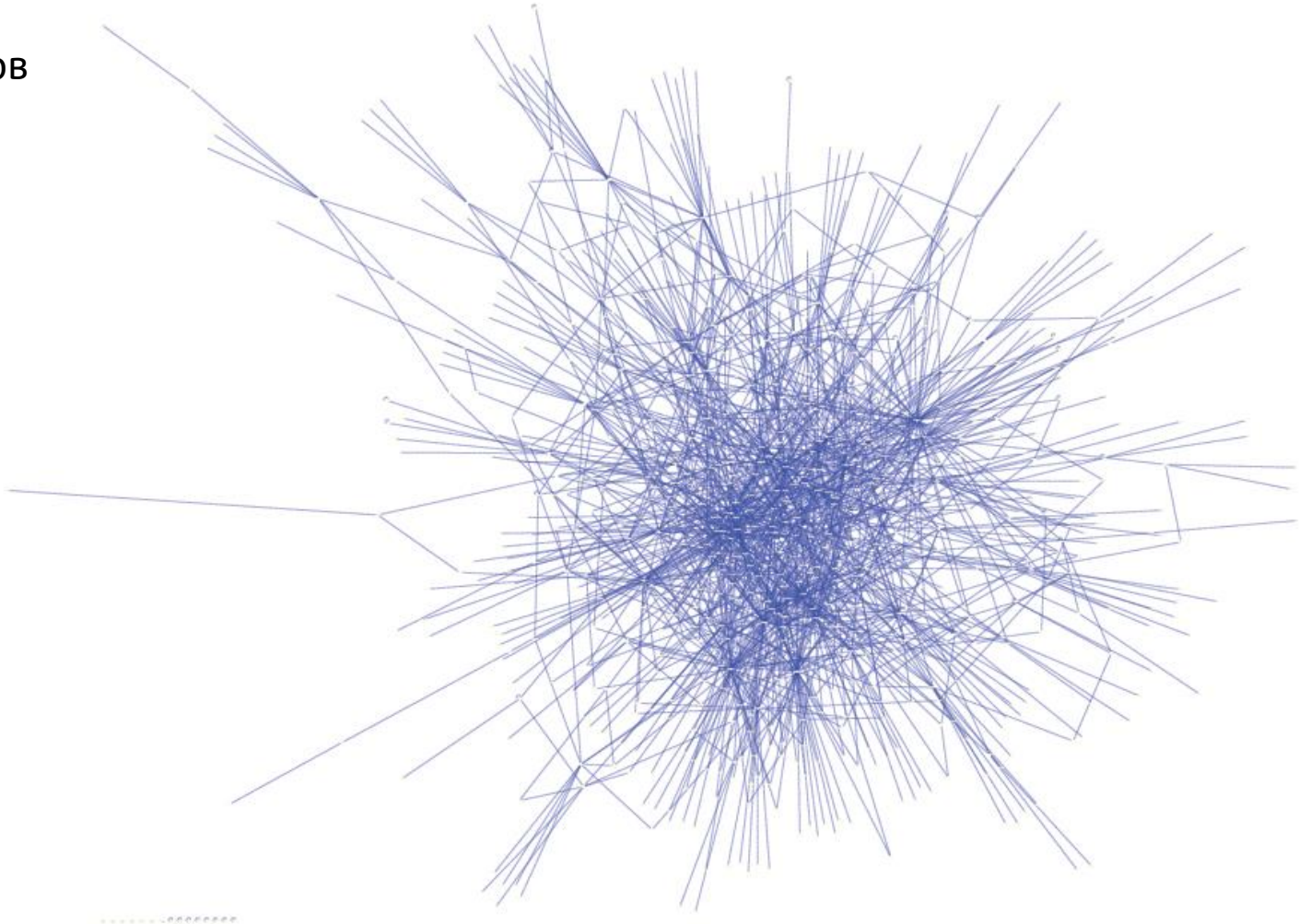
12

8

4

Интерактом *Helicobacter pylori*

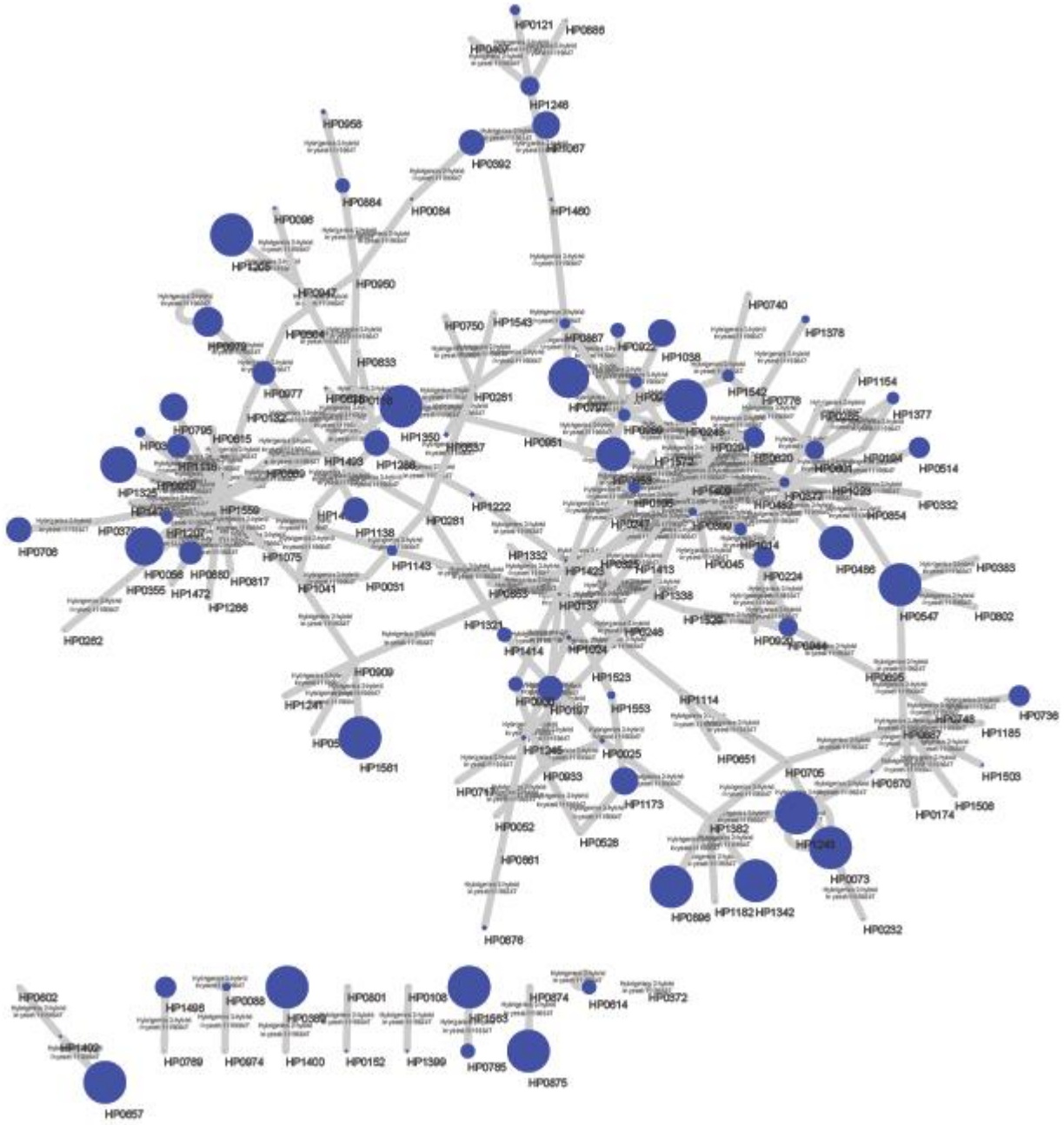
1200 белков



26695

A45

J99



Nonmetric MDS

Отличие мутантов по белковой представленности
почти такое же как и межштаммовое

Coordinate 2

1.5
1.0
0.5
0.0
−0.5

−1.0
−0.5
0.0
0.5

T228

2_fullGelDigest_IDA_1−T198
1_fullGelDigest_IDA_1−T197

T227_mut1_2_20130319_2uL_1
T226_mut1_1_20130319_2uL_1

T195_helicob_J99_1_fullGelDigest_IDA_1−T195
T196_helicob_J99_2_fullGelDigest_IDA_1−T196
T194_helicob_26695_2_fullGelDigest_IDA_1−T194
T193_helicob_26695_1_fullGelDigest_IDA_1−T193

T203_hp_1325−1_IDA_4uLload_1−T203
T206_hp_1325−2_IDA_4uLload_1−T206
T204_hp_hpy−1_IDA_4uLload_1−T204
T207_hp_hpy−2_IDA_4uLload_1−T207

Благодарю за внимание