ВВЕДЕНИЕ В МЕТАГЕНОМИКУ

Школа по биоинформатике NGS 2017

Федеральный Научно-клинический Центр Физико-химической Медицины

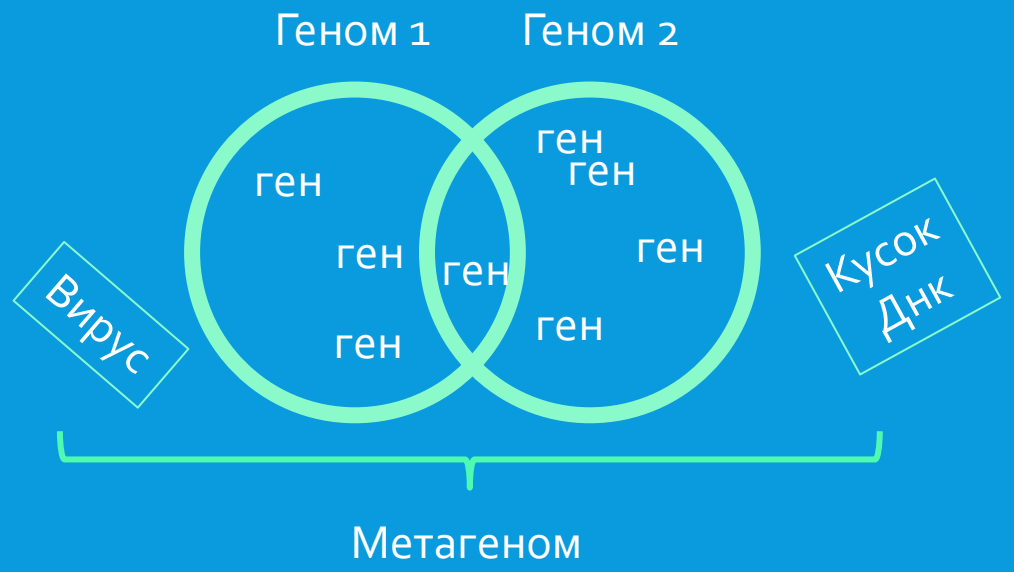
Лаборатория Биоинформатики

Галкин Фёдор

f.a.Galkin@gmail.com

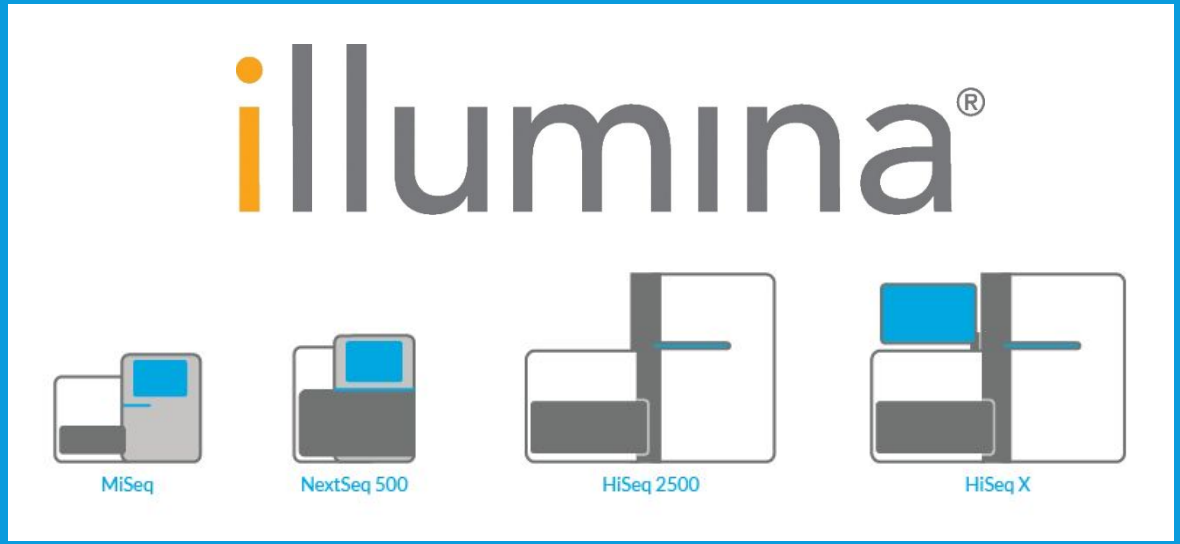
ГЕНОМ И МЕТАГЕНОМ

- Геном — последовательность нуклеотидов, присущая какой-либо биологической единице (виду / организму / клетке).
- Метагеном — генетическая информация, содержащаяся во всех биологических единицах данной среды и в самой среде.

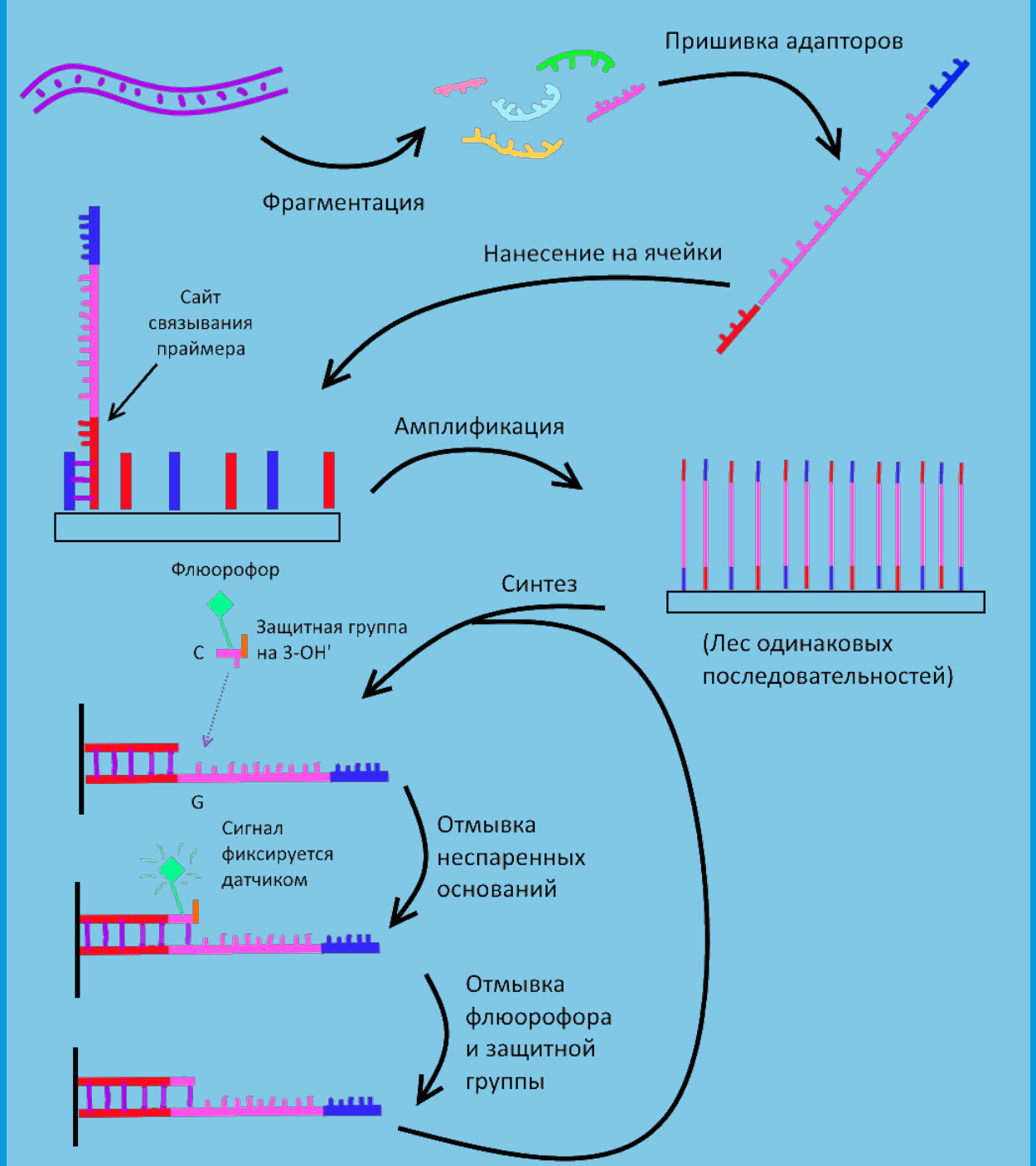


СЕКВЕНИРОВАНИЕ СИНТЕЗОМ (ILLUMINA)

1. Выделение ДНК (из одного организма или сообщества);
2. Дробление ДНК на множество коротких (250-600 nt) последовательностей;
3. Присоединение линкеров к последовательностям;
4. Распределение ДНК по ячейкам;
5. Амплификация матрицы;
6. Добавление к матрице меченых A/T/C/G;
7. Фиксация сигнала от присоединихшихся нт Многократное повторение раундов репликации;
8. Обработка данных (отсечение линкеров, оценка качества, устранение чужеродных последовательностей)



1-4: Пробоподготовка
5-7: Секвенирование
8+: Биоинформатика

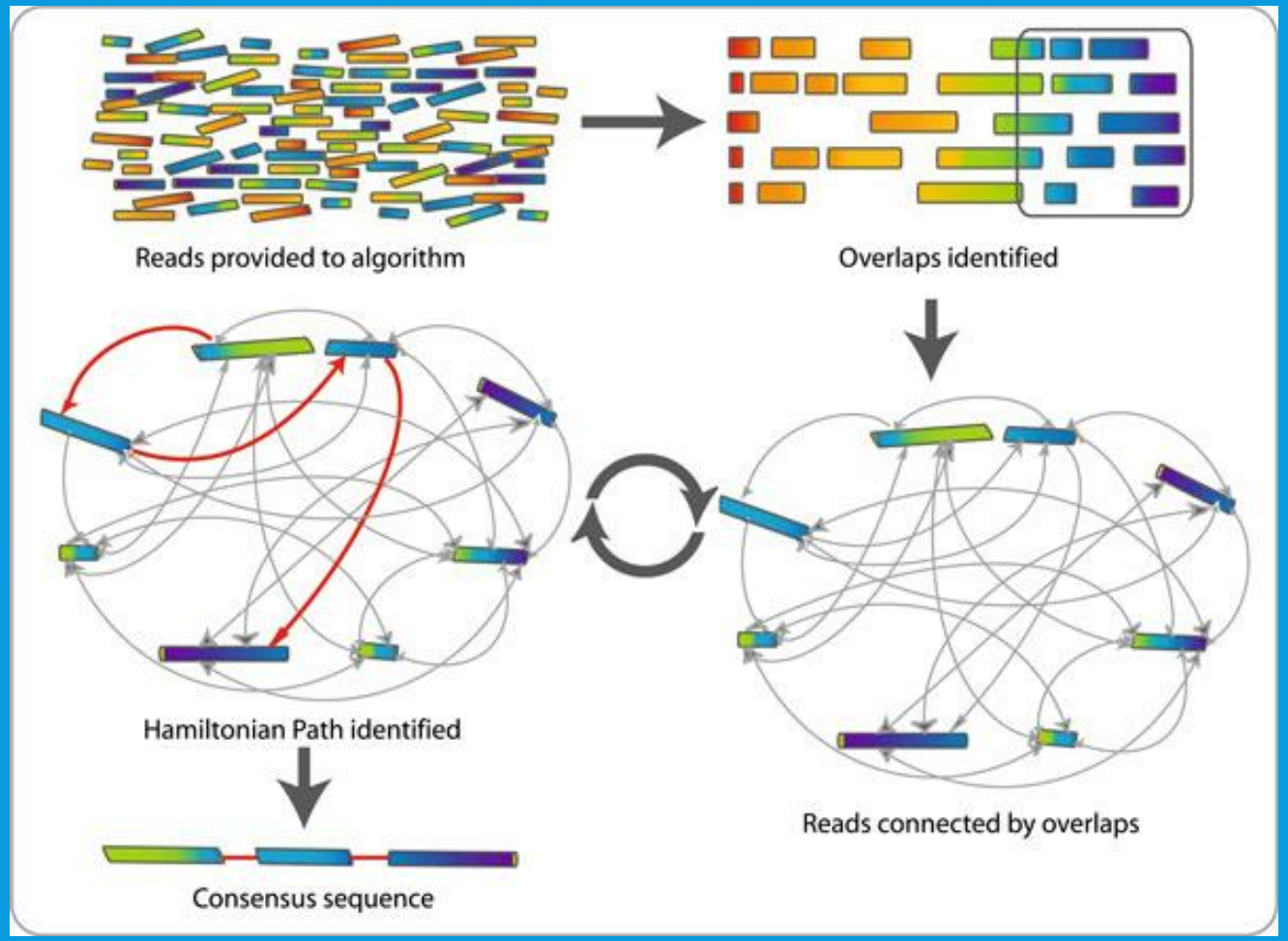



```
@ERR1316078.1 10317.000039927B_0/1
TACGTAGGGTGCAAGCGTTATCCGGAATTATTGGGCGTAAAGCGCGCGCAGGCGGTT
CGTCCCGTCCGGTGTGAAAGCCCATCGCTTAACCCCGGAACTGCATCGGGTACGGGC
ATCCTTGCGTCCGGTCGGGGTGGTCGGAATTCCCGG
+
AAAA>C>A>B>>A1GGGGGAGEHEEE0BG12DDBG0?FECA21B?/AE/>>>E>E//
/>?///>?//E@<@ECFHHF00?<B0??B/?1?1//>>//<1??111..-.>>EC@-
-//<<00/-.--:;----:----.0;-;9ABB9/9-

@ERR1316078.2 10317.000039927B_1/1
TACGTAGGTGGCGAGCGTTATCCGGAATGATTGGGCGTACAGGGCGCGTAGGTGGCG
TACTAAGTCTGTAGTAAAAGGCAATGGCTCAACCATTGTAAGCTATGGAAACTGGTA
TGCTGGAGTGCAGAAGAGGGCGATGGAATTCCATGT
+
>>>3>BCA54>BA2EEEGGEGGHEE22B3B5DDFGAAGE01D3B1AAE>EEG@EFA?
>>/?@@FFBGHHFHH44BB3//?0?FGBFGG3BF/FGHHHGH?B1D111B?GDBGHF
DGD2>100@@@D110110??/--->C0<<>1=0<<=
```

OXFORD NANOPORE

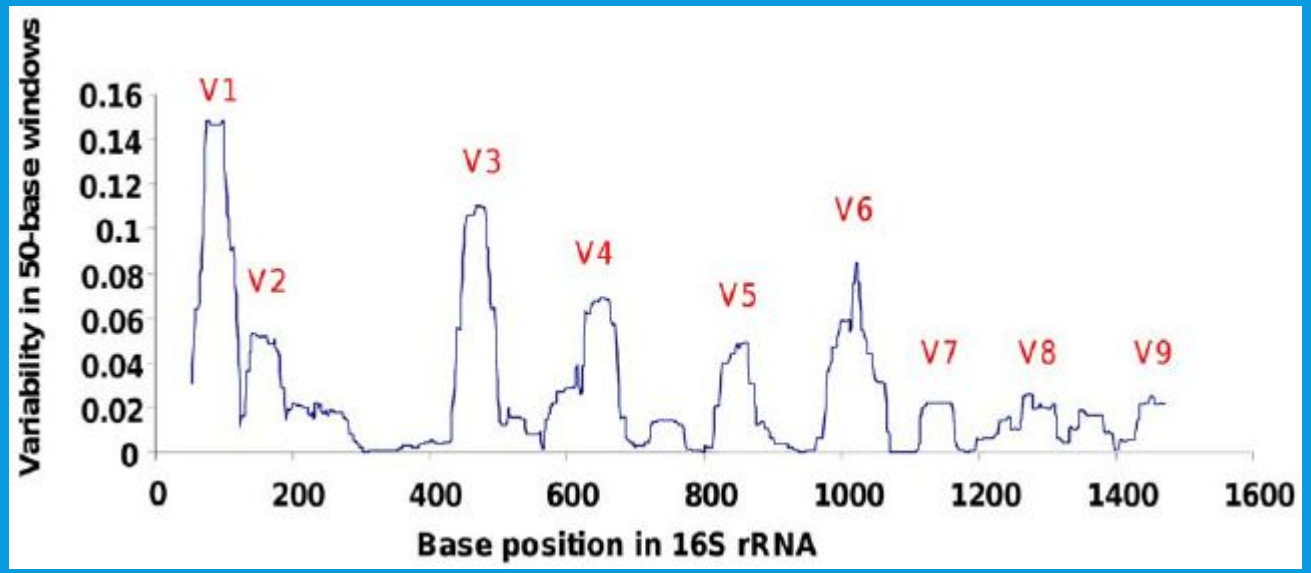
СБОРКА ГЕНОМА



Гамильтонов путь — путь в графе, проходящий через каждую вершину ровно один раз

16S VS WGS

- Рибосома — универсальная биомашина. У бактерий малая единица кодируется 16S-ДНК, у эукариот — 18S.
- 16S ДНК очень консервативна, а профиль мутаций в её гипервариабельных участках видоспецифичен.
- Чтобы определить вид можно амплифицировать его 16S участок и секвенировать только его.



16S VS WGS

Секвенирование ампликона (16S)	Полногеномное секвенирование
Большая глубина	Большее покрытие
Меньше информации (мегабайты, 10^1 Mbp)	Больше информации (гигабайты, 10^2-10^4 Mbp)
Подходит для определения вида, если амплифицированы маркерные гены	Подходит для определения вида
Амплификация может дополнительно искажать информацию	Степень и форма искажения информации зависит от выбора платформы

ГЕНОМ И МЕТАГЕНОМ

- Геном — последовательность нуклеотидов, присущая какой-либо биологической единице (виду / организму / клетке).
- Метагеном — генетическая информация, содержащаяся во всех биологических единицах данной среды и в самой среде.

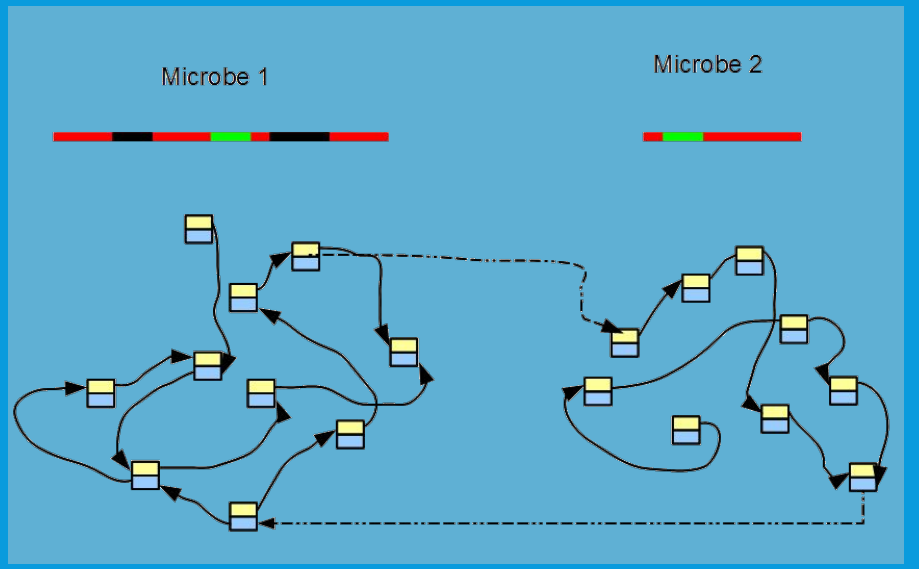
Геном человека

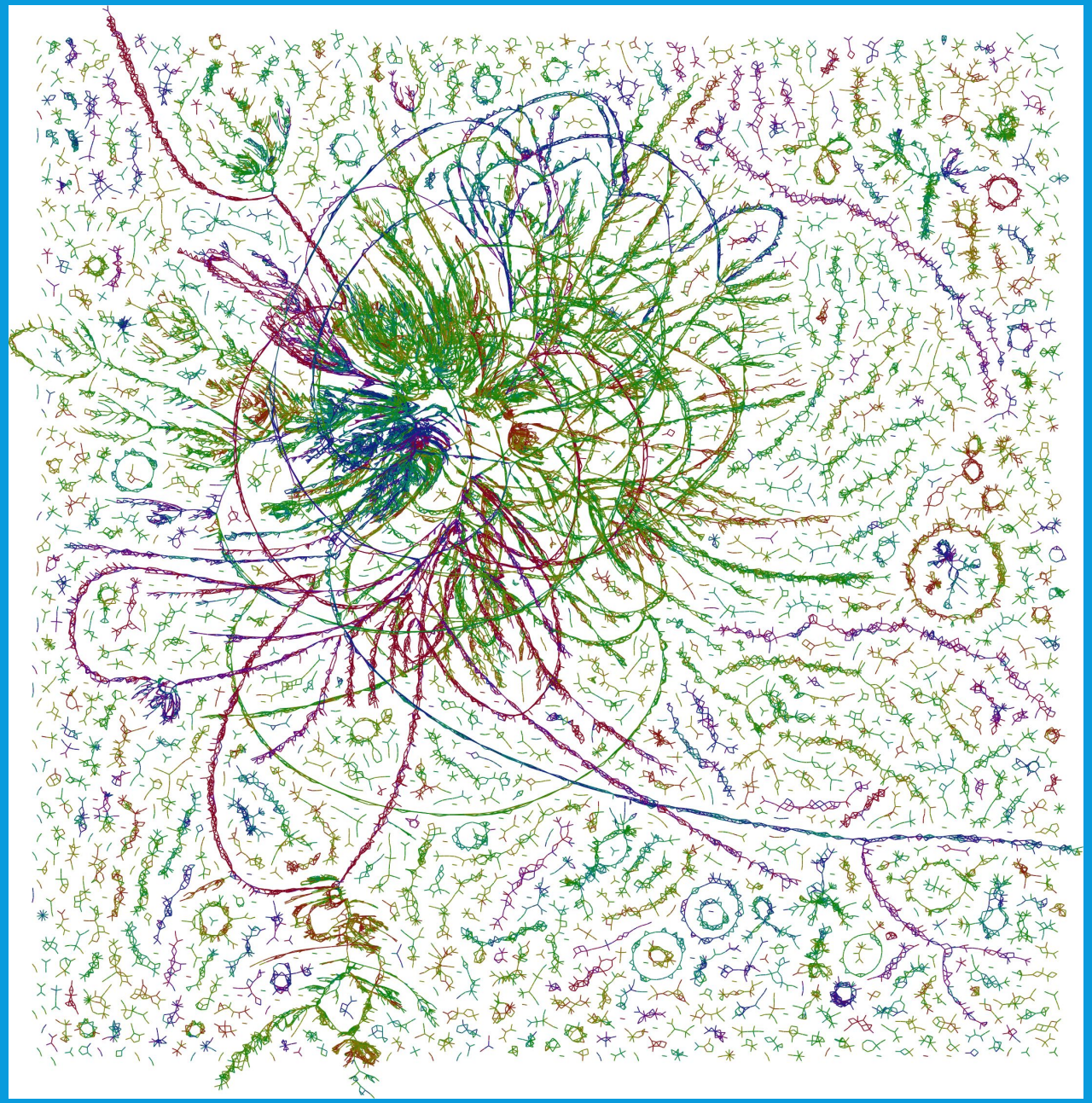
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA

…

GTCAAGACCCGCTGCCAGGCCAGCCTCTGCCTGACCGCCGGCTCACCTC

x25



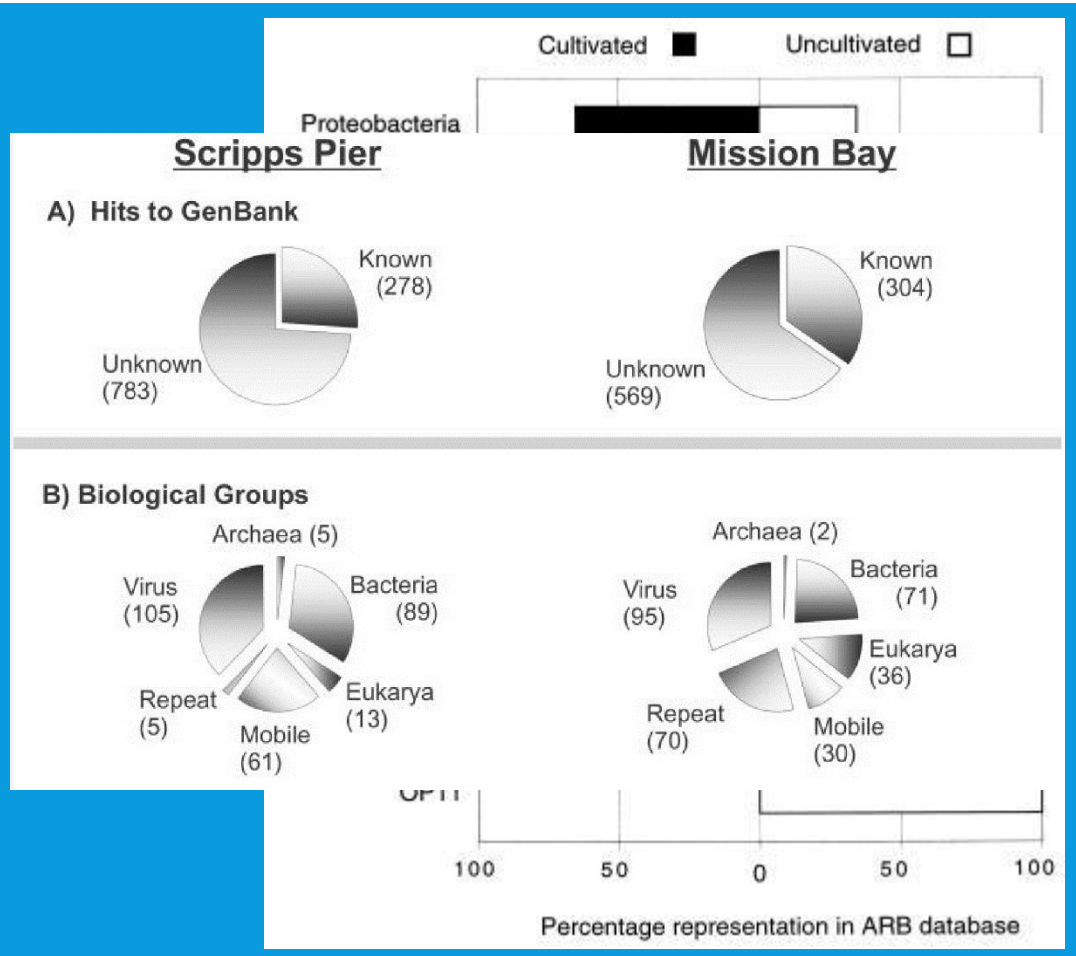


ГЕНОМ VS МЕТАГЕНОМ

Геном	Метагеном
Нужно изолировать организм	Нужно изолировать сообщество
Необходимо культивировать бактерию, чтобы выделить достаточно ДНК	Чтобы выделить больше ДНК достаточно взять больше образца
Несёт информацию о функциях организма	Несёт информацию о функциях сообщества
Цель — создать консенсус, присущий всем биологическим единицам какого-либо объекта (особям в виде)	Цель — показать разнообразие биологических единиц внутри одного объекта (видов в сообществе)
---	Можно разделить на геномы (иногда)
Должен включать максимум подпоследовательностей, присущих объекту (WGS)	Может содержать только маркерные последовательности таксонов (16S или другие)
Термин введён в 1920	Термин введён в 1998

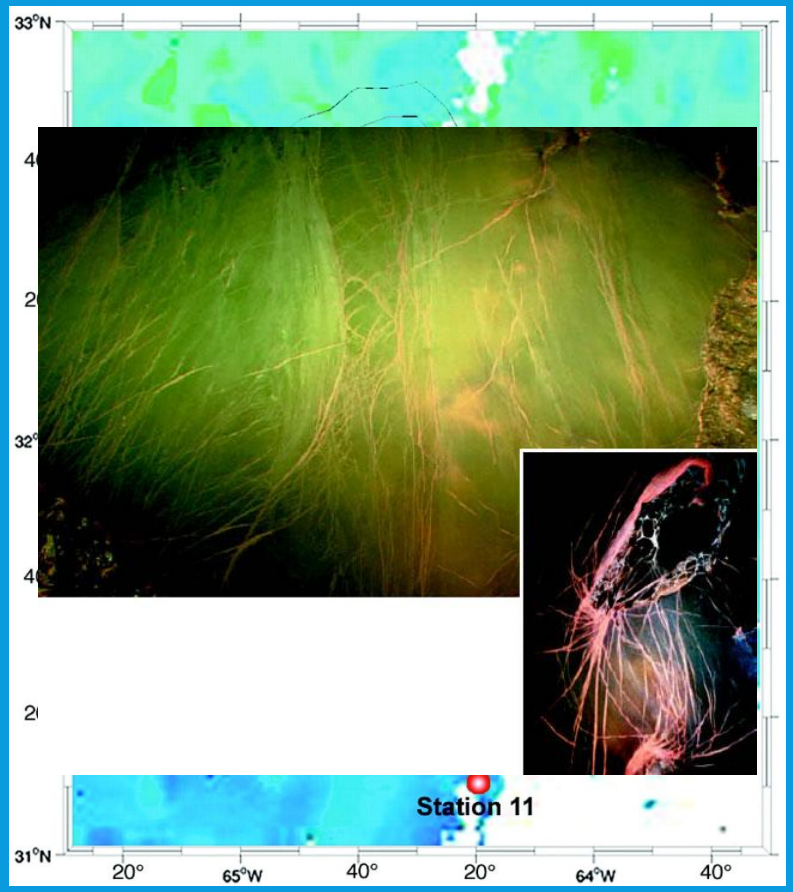
КЛЮЧЕВЫЕ ИССЛЕДОВАНИЯ В МЕТАГЕНОМИКЕ

- 1998 — секвенирование ДНК, выделенной из сообществ, показало, что только 1% микроорганизмов культивируемы (PMC107498)
- 2002 — секвенирование вирусной ДНК из морской воды показало ранее неизвестное разнообразие вирусов. Вирусы становятся самым большим депо генетической информации (PMC137870).



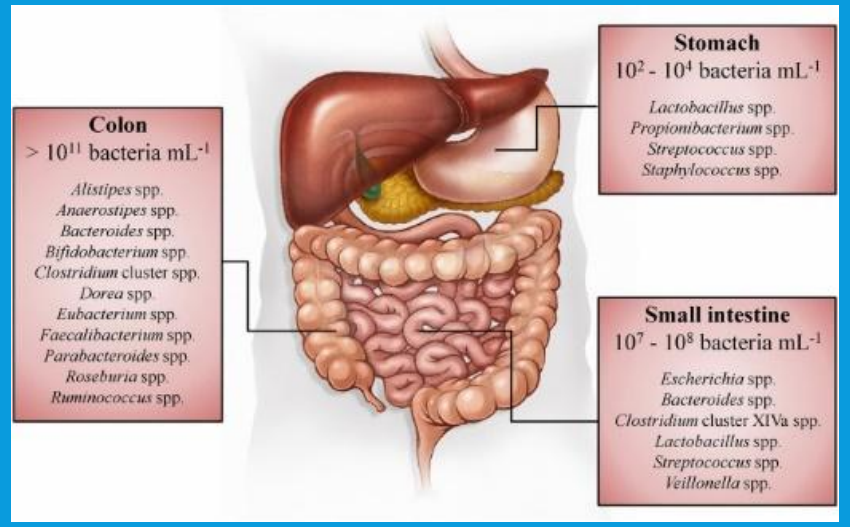
КЛЮЧЕВЫЕ ИССЛЕДОВАНИЯ В МЕТАГЕНОМИКЕ

- 2004 — экспедиция в Саргассовом море секвенировала 1.2кк белок-кодирующих генов (х10 раз больше, чем было тогда известно), найдено 150 ранее неизвестных бактерий (PMID: 15001713)
- 2005 — секвенирование метагенома шахтовых стоков позволило полностью восстановить 2/5 геномов этого сообщества, смоделировать метаболизм сообщества и подобрать условия культивации 1 из бактерий (PMID: 14961025).



МИКРОБИОТА ЧЕЛОВЕКА

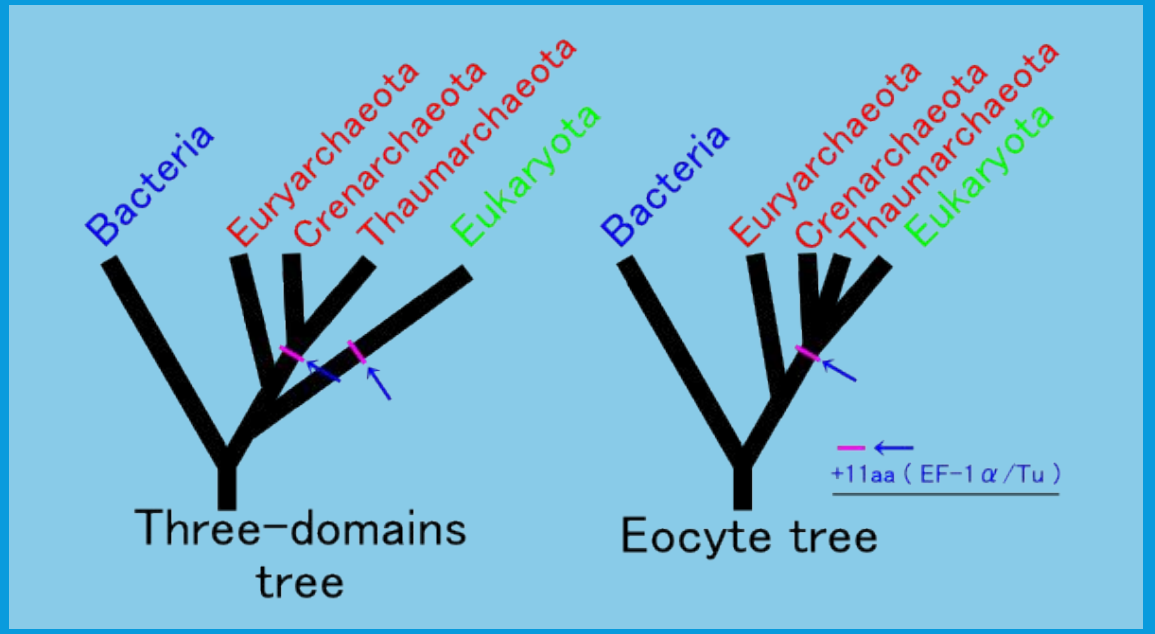
- 2008 — старт NIH Human Microbiome Project, в котором было отсеквенировано >5k образцов из 15-18 микробных сообществ на теле 242 американцев. Доступны 16S-, WGS- и частично обработанные данные
- 2011 — старт коммерческого проекта American Gut Project, объединившего 200k 16S секвенирований со всего света (MiSeq).



(PMC4528021)

ДВА ИЛИ ТРИ ДОМЕНА ЖИЗНИ

- 1985 — Карл Вёзе издал работу о трёхдоменном дереве жизни на основе сравнения рРНК разных организмов
- 2015 — после секвенирования образцов со дна Атлантики учёные собрали геном локиархеи. Её гены содержат 3% эукариотических белков (PMC4444528). В образцах не найдено 18S-эукариотических генов и все эукариотические гены фланкированы бактериальной ДНК

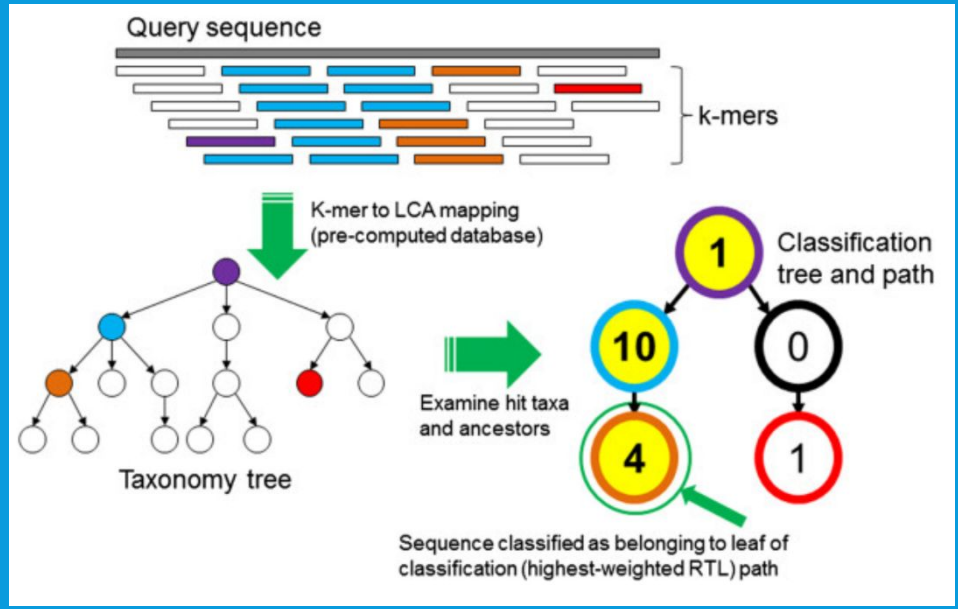


ЧТО МОЖНО ДЕЛАТЬ С МЕТАГЕНОМОМ?

- Описать состав сообщества;
- Искать отличия между сообществами;
- Описать функционал сообщества;
- Собрать метагеном;
- Собрать геном;

BINNING / OTU CALLING

- **Биннинг** — соотнесение каждого рида таксономической единице (Observed Taxonomy Unit).
- **Closed reference binning** — выравнивание ридов против БД характеристических последовательностей.
- **Silva, Greengenes** — БД рибосомальных последовательностей.
- **Kraken, Metaphlan** — самые быстрый классификатор, использующий closed ref. подход (PMC4053813).



Kraken проверяет точные совпадения со своей БД ДНК, характерных для таксонов разных уровней.

DE NOVO BINNING

Определить виды не по последовательностям, а по их статистикам:

- GC%;
- Частота кодонов;
- Ди-/Три-/Тетрануклеотдное распределние.

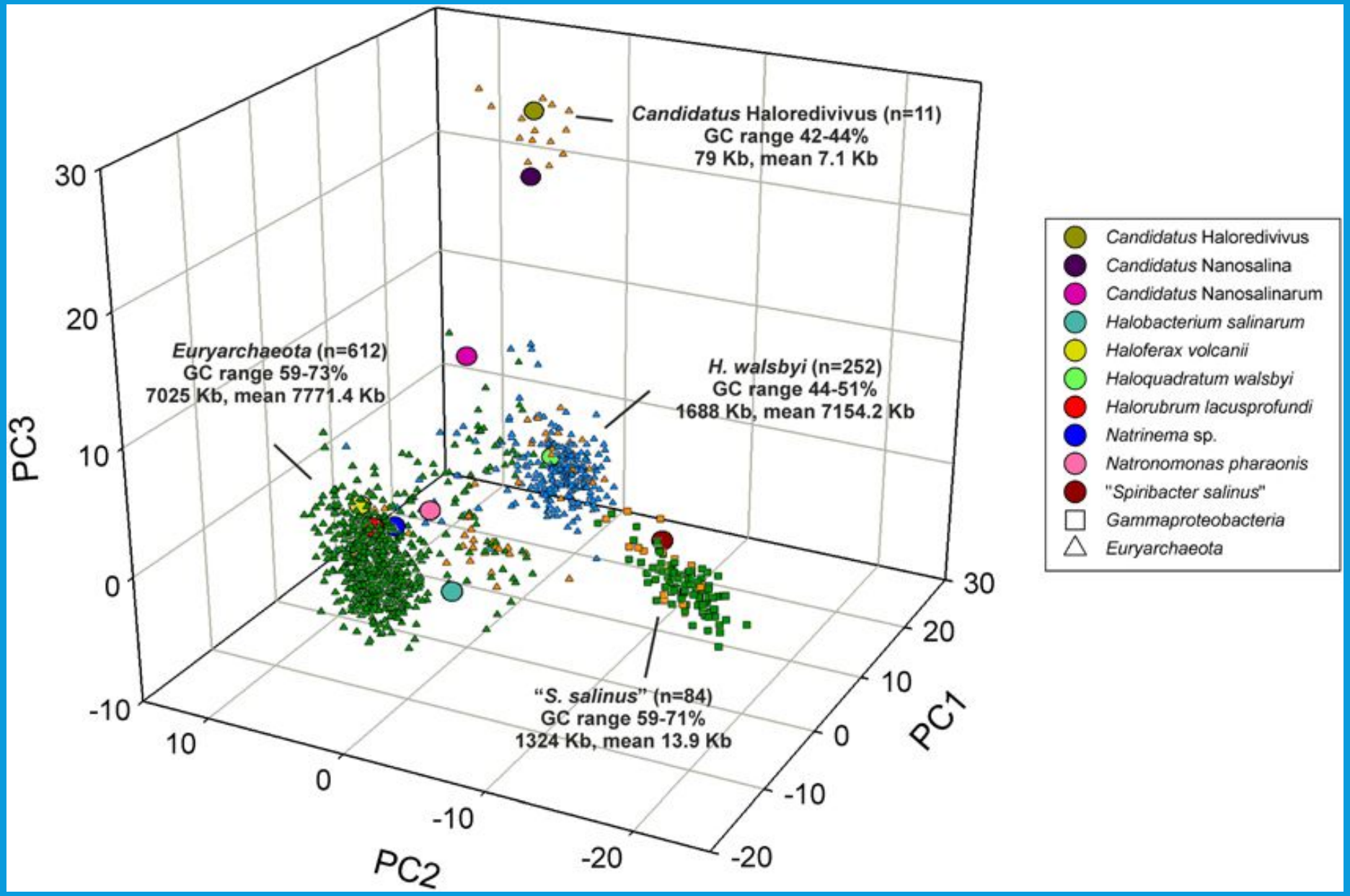
- Метод используется, когда картирование ридов не помогло

…ACCTGGAATCGGAAA…

L = 15
GC% = 47%

Тетрануклеотиды:

```
ACCTGGAATCGGAAA
ACCT  AATC
 CCTG  ATCG
  CTGG  TCGG
   TGGA  CGGA
    GGAA  GGAA
     GAAT  GAAA
```


Candidatus Haloredivivus (n=11)
GC range 42-44%
79 Kb, mean 7.1 Kb
Euryarchaeota (n=612)
GC range 59-73%
7025 Kb, mean 7771.4 Kb
H. walsbyi (n=252)
GC range 44-51%
1688 Kb, mean 7154.2 Kb
"S. salinus" (n=84)
GC range 59-71%
1324 Kb, mean 13.9 Kb
PC3
PC2
PC1
30
20
10
0
-10
-20
Candidatus Haloredivivus
Candidatus Nanosalina
Candidatus Nanosalinarum
Halobacterium salinarum
Haloferax volcanii
Haloquadratum walsbyi
Halorubrum lacusprofundi
Natrinema sp.
Natronomonas pharaonis
"Spiribacter salinus"
Gammaproteobacteria
Euryarchaeota

МЕТОДЫ БИННИНГА

De novo	Closed reference
Работает без каталога	Нужен каталог характерестических последовательностей
Не подходит для сравнения ридов, полученных с разных ампликонов	Можно объединять дата сеты, полученные после секвенирования разных ампликонов
Вычисления не параллелизуются	Вычисления параллелизуются
Филогенетическое дерево строится заново и непредвзято	Филогенетическое дерево задано заранее
Позволяет обнаружить скрытое разнообразие	Определяются только известные таксоны

СРАВНЕНИЕ СОСТАВА СООБЩЕСТВ

Для этого используются многочисленные метрики, в том числе привнесённые из экологии:

- Сравнение численности всех таксонов между выборками;
- Jaccard index;
- Bray–Curtis dissimilarity;
- Jensen–Shannon divergence;
- Unifrac (учитывает филогенетическое расстояние)
- …

$$J(\mathbf{x},\mathbf{y}) = 1 - \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)}$$

Σmin(#ридов в общих видах)

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

Σ (#ридов в уникальных видах)

J = 1 - (10 + 0) / (500 + 200)

	Cont1	Cont2	…
Tax1	10	500	…
Tax2	200	0	…
…	…	…	…

	Exp1	Exp2	…
Tax1	100	300	…
Tax2	200	200	…
…	…	…	…

BC = 1 – 2*(10) / (200 + 0)

- T-test
- Mann-Whitney
- Wilcoxon

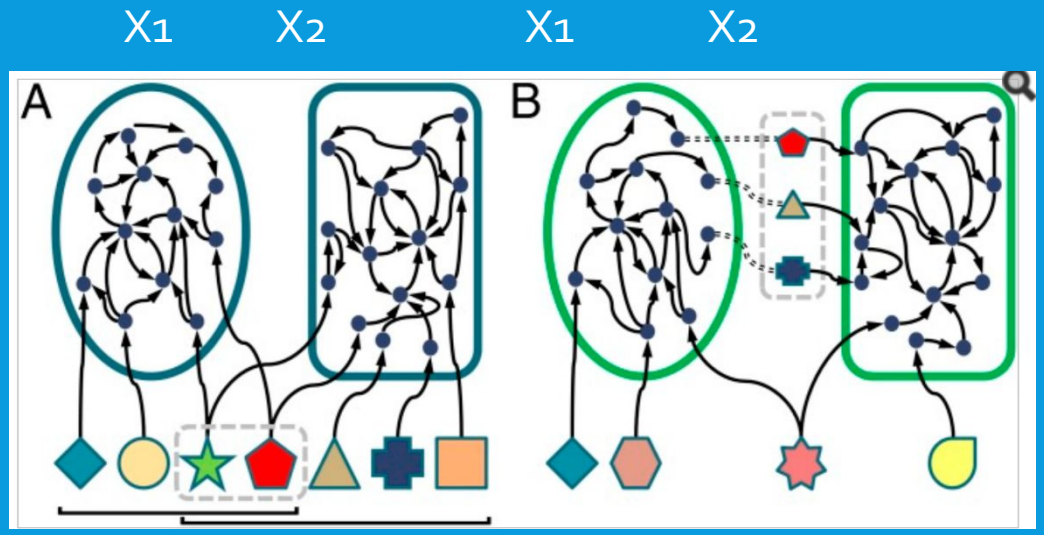
КОРРЕЛЯЦИОННАЯ ТАБЛИЦА

Численность каких таксонов скоррелирована?

Heatmap — способ иллюстрации матрицы корреляций.
Красный — негативная корреляция, зелёный — положительная,
белый — нет корреляции

ФУНКЦИОНАЛЬНЫЙ АНАЛИЗ

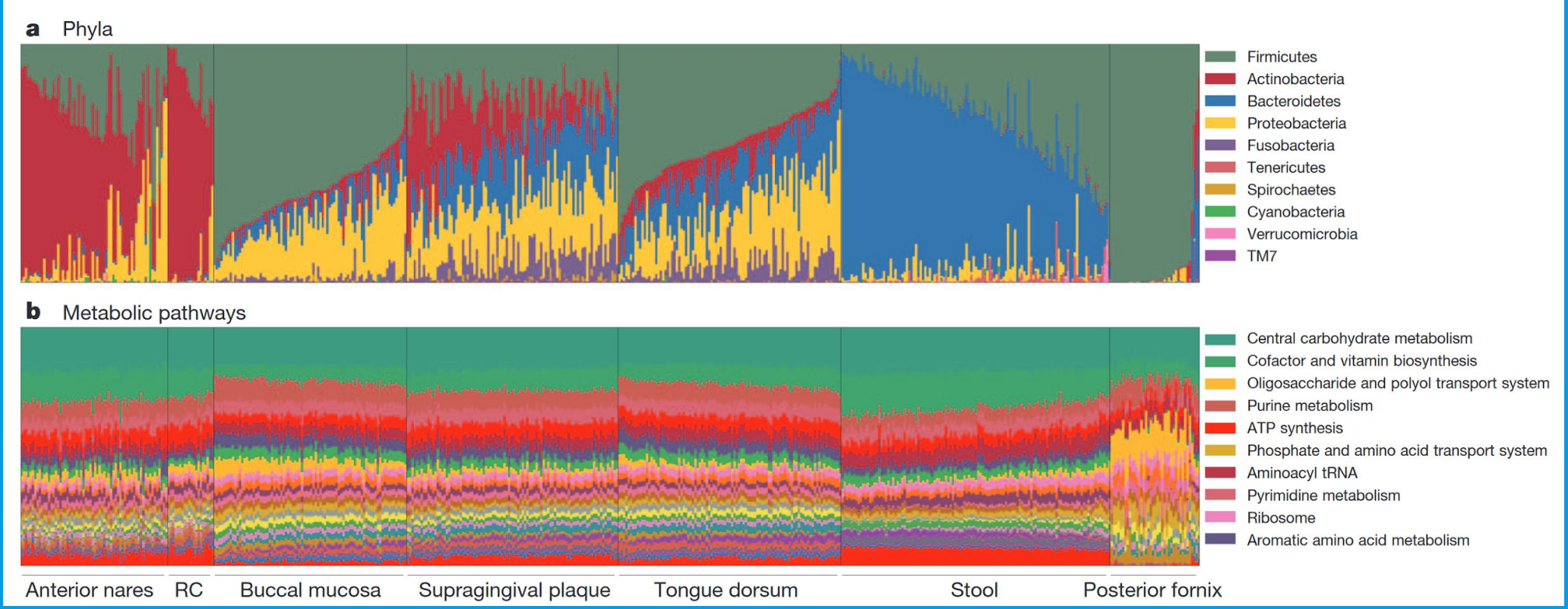
- WGS даёт информацию о генах в сообществе;
- Если есть только 16S:
  16S -> Виды -> Геномы видов -> Функциональное моделирование / Предсказание генов / Анализ литературы и БД
- Существует множество способов количественно выразить сотрудничество / конкуренцию между видами: Metabolic Complementarity / Competition Index, Biosynthetic Support Score… (PMC3732988 — хорошая статься по теме)



Comp.(x1,x2) = 2/4 = 50%
Comp.(x2,x1) = 2/5 = 40%

Coop.(x1,x2) = 0/3 = 0%
Coop.(x2,x1) = 3/5 = 60%

СООТВЕТСТВИЕ ФУНКЦИОНАЛА И ВИДОВОГО СОСТАВА



Вертикальные полоски – образцы от разных людей. Таксономический состав сильно различается, тогда как генетический неизменен (PMC3564958)

TAKE HOME MESSAGE

- Метагеномика позволяет увидеть скрытое разнообразие микромира;
- Метагеномика позволяет оценить, как микроорганизмы взаимодействуют между собой и с окружающей средой (функциональный анализ);
- WGS и 16S-секвенирование предназначены для разных задач;