

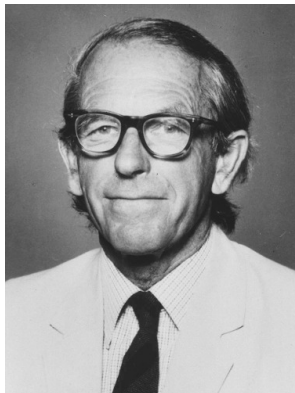
# Нуклеотидные банки данных

Ваня Русинов

# Экскурс в историю секвенирования

# История секвенирования

- ▶ 1951 – Первая последовательность:  
бычий инсулин, цепь В (белок!)



Frederick Sanger

# История секвенирования

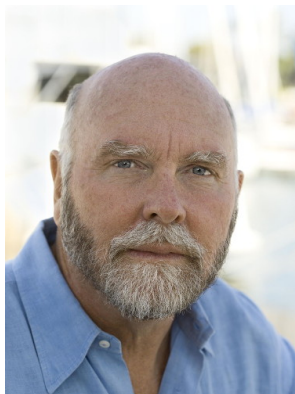
- ▶ 1951 – Первая последовательность: бычий инсулин, цепь В (белок!)
- ▶ 1976 – Первый полный геном: бактериофаг MS2 (РНК)



Walter Fiers

# История секвенирования

- ▶ 1951 – Первая последовательность: бычий инсулин, цепь В (белок!)
- ▶ 1976 – Первый полный геном: бактериофаг MS2 (РНК)
- ▶ 1995 – Первый геном бактерии: *Haemophilus influenza*



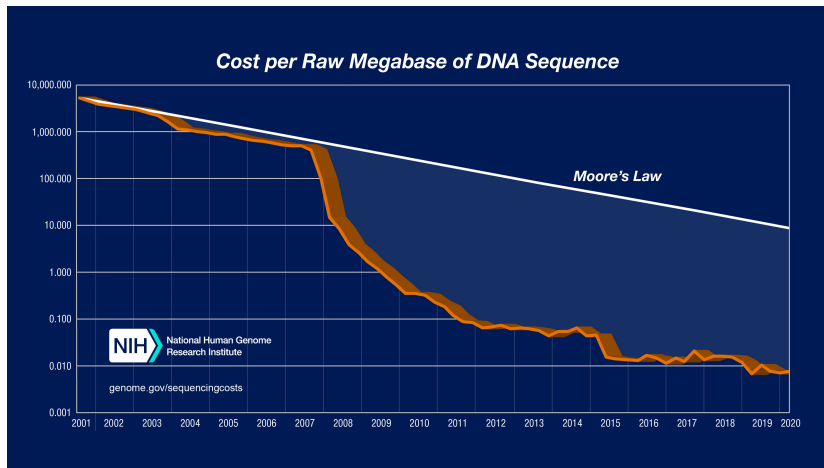
Craig Venter

# История секвенирования

- ▶ 1951 – Первая последовательность:  
бычий инсулин, цепь В (белок!)
- ▶ 1976 – Первый полный геном:  
бактериофаг MS2 (РНК)
- ▶ 1995 – Первый геном бактерии:  
*Haemophilus influenza*
- ▶ 2001 – Первый геном человека:  
международный консорциум vs. Celera  
Genomics

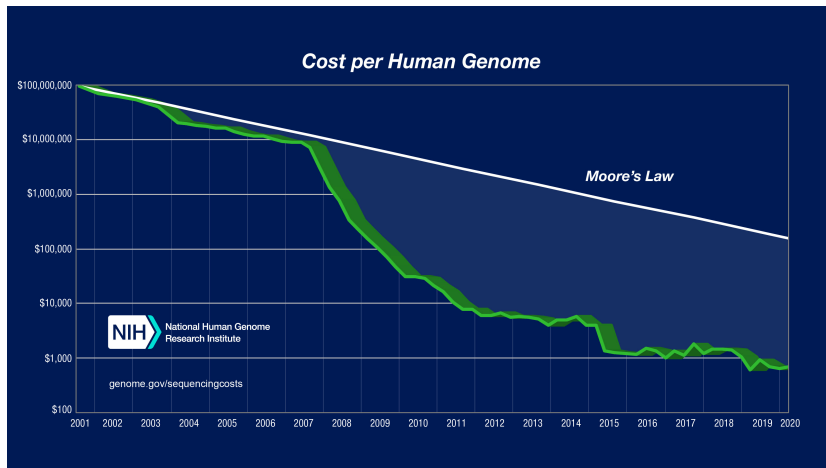
# Next Generation Sequencing

# Стоимость секвенирования





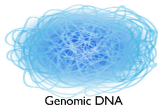
# Стоимость секвенирования



# Sanger vs. NGS

## Human Genome Sequencing

Generating a Reference  
Genome Sequence  
(e.g., Human Genome Project)



Genomic DNA

Break genome into  
large fragments and  
insert into clones



Order clones



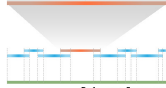
Break individual  
clones into  
small pieces



Generate thousands  
of sequence reads  
and assemble  
sequence of clone

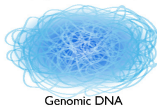


Assemble sequences  
of overlapping clones  
to establish  
reference sequence



Reference Sequence

Generating a Person's  
Genome Sequence  
(e.g., Circa ~2016)



Genomic DNA

Break genome  
into small pieces



... TATGCGATGCGTATTTCTAATA ...

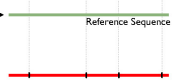
Generate millions  
of sequence reads

Align sequence reads  
to established  
reference sequence



Reference Sequence

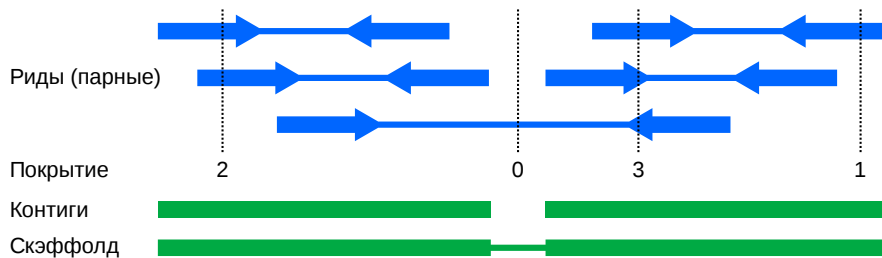
Deduce starting  
sequence and identify  
differences from  
reference sequence



# Термины NGS

- Прочтение (рид)** последовательность, полученная из секвенатора
- Качество прочтения** величина  $Q = -10 \log_{10} p$ , где  $p$  – вероятность ошибочного прочтения нуклеотида;  $Q$  вычисляется для каждого нуклеотида рида
- Контиг** секвенированный без пропусков фрагмент ДНК, собирается в компьютере из прочтений
- Скэффолд** набор контигов, про которые известно взаимное расположение и примерное расстояние; разрывы заполняют соответствующим количеством букв N

# Сборка NGS



# Показатели качества сборки

**(Среднее) покрытие** среднее число ридов, в которые попал каждый нуклеотид

**N50** длина самого длинного контига, такого, что этот и все более длинные контиги покрывают более половины генома

**L50** номер контига (при упорядочивании по убыванию длины), длина которого равна N50

# Технологии NGS

Table 1 | **Main characteristics of current NGS technologies**

Technology	Run type			Maximum read length	Quality scores	Error rates
	Single end	Paired end	Mate pair			
Illumina	Yes	Yes	Yes	300 bp	>30	0.0034–1%
SOLiD	Yes	Yes	Yes	75 bp	>30	0.01–1%
IonTorrent	Yes	Yes	No	400 bp	~20	1.78%
454	Yes	Yes	No	~700 bp (up to 1 kb)	>20	1.07–1.7%
Nanopore	Yes	No	No	5.4–10 kb	NA	10–40%
PacBio	Yes	No	No	~15 kb (up to 40 kb)	<10	5–10%

Escalona, Rocha, Posada. *Nat Rev Genet.* 2016; 17(8): 459–469. PMID: 27320129; PMCID: PMC5224698.

# Масштабы бедствия

По данным NCBI Genome, на данный момент секвенированы (или в процессе секвенирования) геномы:

- ▶ 282 246 прокариот
- ▶ 41 416 вирусов
- ▶ 24 437 плазмид
- ▶ 17 234 органелл
- ▶ 13 199 эукариот

NCBI Genome содержит далеко не все секвенированные геномы.

# Массовые геномные проекты

...

2013 100000 геномов людей

2012 100000 геномов патогенных бактерий

2012 1000 геномов грибов

2011 5000 геномов насекомых

2010 10000 геномов позвоночных

2009 1000 геномов растений

2008 3000 геномов риса

2008 1000 геномов людей

2008 1001 геном Арабидопсиса

...



# Что секвенируют?

- ▶ Геномы
- ▶ Экзомы
- ▶ Транскриптомы
- ▶ Метиломы
- ▶ Метагеномы
- ▶ Сайты связывания ДНК белком (ChIP-seq)
- ▶ Сближенные участки хромосом (HiC)
- ▶ И многое другое

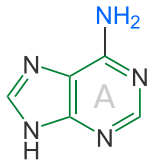
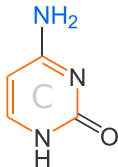
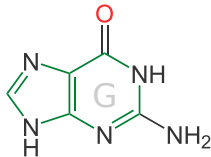
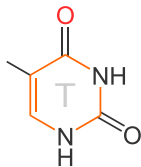
# Нуклеотидные банки данных

# Соглашения о хранении последовательностей

- ▶ Хранят последовательность однобуквенных обозначений нуклеотидов,
- ▶ записанную от 5'-конца к 3'-концу
- ▶ только для одной цепи ДНК.
- ▶ В случае РНК урацилы тоже обозначают буквой Т.
- ▶ Нуклеотиды нумеруют, начиная с 1 (а не 0).
- ▶ Координаты указывают включительно с обоих концов.

# Нуклеотидные коды IUPAC

Код	Значение	Мнемоническое правило
A	A	<b>A</b> dentine
C	C	<b>C</b> ytosine
G	G	<b>G</b> uanine
T	T	<b>T</b> hymine
K	G или T	<b>K</b> eto
M	A или C	<b>aM</b> ino
R	A или G	<b>puR</b> ine
Y	C или T	<b>pY</b> rimidine
S	C или G	<b>S</b> trong interaction
W	A или T	<b>W</b> eak interaction
B	не A	после A по алфавиту
D	не C	после C по алфавиту
H	не G	после G по алфавиту
V	не T	после T и U по алфавиту
N	любой	<b>aN</b> y



Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. *PNAS*. 1986;83:4-8. PMID: 2417239; PMCID: PMC322779.

# Типы баз данных

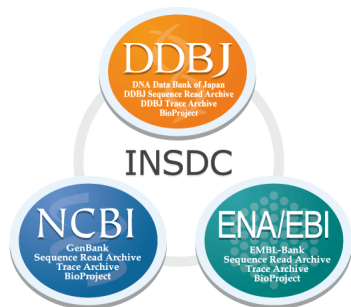
**Архивные** записи создают сами экспериментаторы, они же отвечают за достоверность информации

**Курируемые** за создание и редактирование записей отвечают специальные люди, кураторы

**Автоматические** записи создаются автоматически компьютерными программами

## International Nucleotide Sequence Database Collaboration:

- ▶ Объединяет 3 крупнейших нуклеотидных архива: GenBank, ENA, DDBJ
- ▶ Ежедневный обмен данными
- ▶ Единый формат таблицы локальных особенностей
- ▶ Рекомендации по использованию терминов и ключевых слов в аннотациях
- ▶ И некоторые прочие унификации (например, таблицы генетического кода)



# Структура данных в INSDC

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	<a href="#">Sequence Read Archive</a>	European Nucleotide Archive ( <a href="#">ENA</a> )	<a href="#">Sequence Read Archive</a>
Capillary reads	<a href="#">Trace Archive</a>		<a href="#">Trace Archive</a>
Annotated sequences	<a href="#">DDBJ</a>		<a href="#">GenBank</a>
Samples	<a href="#">BioSample</a>		<a href="#">BioSample</a>
Studies	<a href="#">BioProject</a>		<a href="#">BioProject</a>

# ENA data domains

- Study/Project** информация о проекте по секвенированию
- Sample** информация об образце – источнике биологического материала
- Read** прочтения NGS и описание методики секвенирования
- Assembly** информация о сборке скэффолдов и хромосом из ридов и контигов
- Analysis** производные данные о сборке или аннотации
- Contig set** содержит ссылки на записи с контигами
- Sequence** собранные и аннотированные последовательности
- Taxon** таксономическая информация
- Checklist** описание требований к метаданным на момент создания записи о проекте



# ENA data classes

**STD** собранные и аннотированные последовательности

**CON** скэфолды

**WGS** геномные контиги

**TSA** транскриптомные контиги

**HTG** High Throughput Genomic data

**HTC** High Throughput Transcriptomic data

**EST** Expressed Sequence Tags

**GSS** Genome Survey Sequence

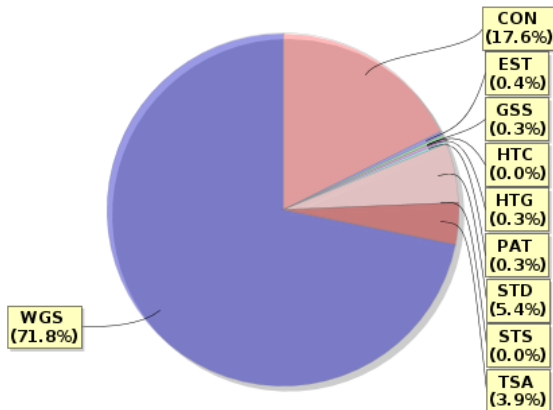
**STS** Sequence Tagged Site

**PAT** последовательности из патентов

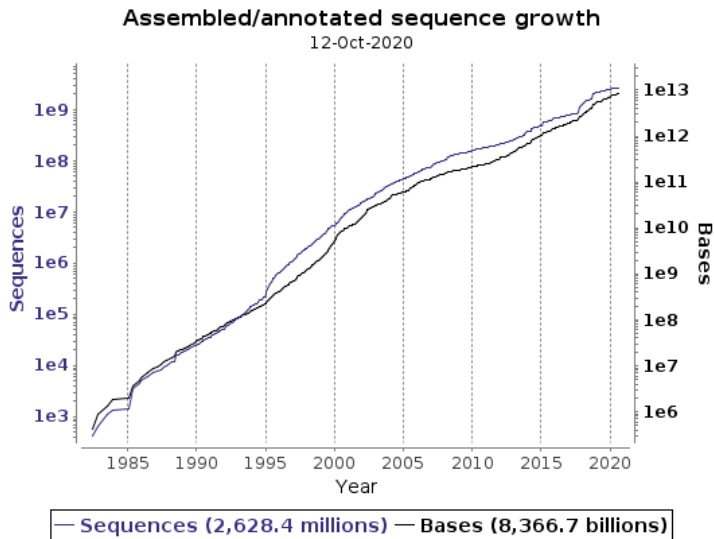
# ENA: распределение записей по классам

## Assembled/annotated sequence bases by dataclass

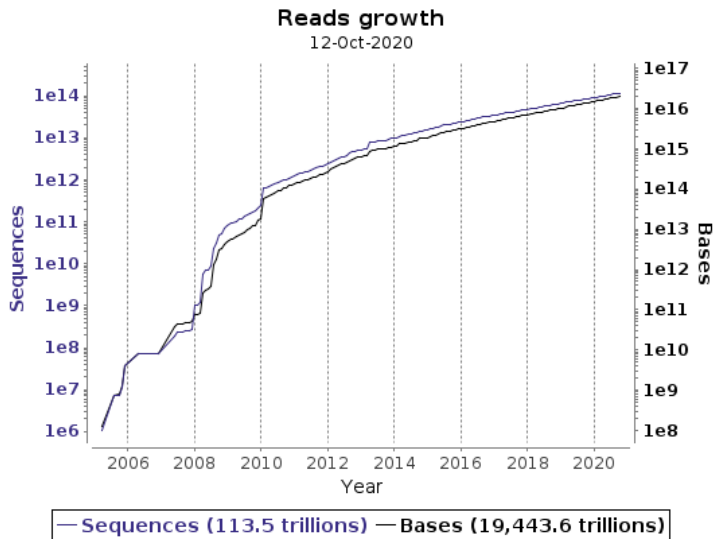
12-Oct-2020



# ENA: рост числа записей



# ENA: увеличение объема SRA



# Формат записи в ENA

```
ID   LR694071; SV 1; linear; genomic DNA; STD; VRT; 335 BP.
XX
AC   LR694071;
XX
PR   Project:PRJEB20083;
XX
DT   01-AUG-2019 (Rel. 141, Created)
DT   12-SEP-2019 (Rel. 142, Last updated, Version 2)
XX
DE   Carcharodon carcharias isolate C_car_delaware_2007 genome assembly,
DE   chromosome: 1
XX
KW   .
XX
OS   Carcharodon carcharias (great white shark)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Chondrichthyes;
OC   Elasmobranchii; Galeomorphii; Galeoidea; Lamniformes; Alopiidae;
OC   Carcharodon.
XX
RN   [1]
RA   Canner M., Saneer T.;
RT   ;
RL   Submitted (31-JUL-2019) to the INSDC.
RL   Saneer-Weeksbooth Bashhouse
XX
DR   MD5; a7e1a34e45308b36003800689dc43934.
DR   BioSample; SAMN04526268.
XX
FH   Key                Location/Qualifiers
FH
FT   source              1..335
FT                       /organism="Carcharodon carcharias"
FT                       /chromosome="1"
FT                       /isolate="C_car_delaware_2007"
FT                       /mol_type="genomic DNA"
FT                       /db_xref="taxon:13397"
XX
SQ   Sequence 335 BP; 323 A; 12 C; 0 G; 0 T; 0 other;
caaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa
...
caaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaac                                     60
//                                                                                   335
```

# Формат записи в GenBank

```
LOCUS          LR694071                335 bp    DNA     linear   VRT 12-SEP-2019
DEFINITION    Carcharodon carcharias isolate C_car_delaware_2007 genome assembly,
              chromosome: 1.
ACCESSION     LR694071
VERSION       LR694071.1
DBLINK        BioProject: PRJEB20083
              BioSample: SAMN04526268
KEYWORDS      .
SOURCE        Carcharodon carcharias (great white shark)
  ORGANISM    Carcharodon carcharias
              Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Chondrichthyes;
              Elasmobranchii; Galeomorphii; Galeoidea; Lammiformes; Alopiidae;
              Carcharodon.
REFERENCE     1
  AUTHORS     Canner,M. and Saneer,T.
  TITLE       Direct Submission
  JOURNAL     Submitted (31-JUL-2019) Saneer-Weeksbooth Bashhouse
FEATURES             Location/Qualifiers
  source          1..335
                  /organism="Carcharodon carcharias"
                  /mol_type="genomic DNA"
                  /isolate="C_car_delaware_2007"
                  /db_xref="taxon:13397"
                  /chromosome="1"
ORIGIN
  1 caaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa
    ...
  301 caaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaac
//
```

# Базы метаданных NCBI

- Taxonomy** Таксономическая база данных
- BioProject** Содержит информацию о проектах секвенирования
- BioSample** Содержит информацию об образце, из которого получена НК
- Assembly** Информация о сборке, ссылки на последовательности контигов, скэффолдов, и репликонов
- Genome** Информация о геномах организмов на основании имеющихся геномных сборок

# RefSeq

Нуклеотидная база данных в NCBI, созданная для снижения избыточности и унификации аннотаций.

- ▶ Автоматическая (по большей части) база данных
- ▶ Свой формат AC (содержат \_)
- ▶ Значительно меньше вырожденность, чем в GenBank (меньше повторяющихся последовательностей)
- ▶ Аннотации валидируются и обновляются
- ▶ Аннотации более унифицированные, чем в GenBank
- ▶ Некоторые записи даже курируются
- ▶ Формат записей практически идентичен GenBank



# Nucleotide

Для упрощения поиска нуклеотидных последовательностей через web-интерфейс и Entrez (nucscore).

- ▶ Содежит нуклеотидные записи из: INSDC (GenBank/ENA/DDBJ), RefSeq, PDB, TPA (Third Party Annotation)
- ▶ Одна и та же последовательность может встречаться много раз (например, идентичные записи INSDC и RefSeq)
- ▶ Содержит все нуклеотидные последовательности в NCBI (после слияния с GSS и EST в июле 2019).

# Nucleotide collection (nr/nt)

- ▶ Содержит только последовательности и их идентификаторы в INSDC, RefSeq, PDB.
- ▶ Идентичные последовательности кластеризуются.
- ▶ Служит для поиска с помощью BLAST.

# Поиск по аннотациям и загрузка данных

- NCBI**
1. Глобальный поиск по базам NCBI (<https://www.ncbi.nlm.nih.gov/search>)
  2. E-Utilities – Web API к системе Entrez (<https://www.ncbi.nlm.nih.gov/books/NBK25501/?term=e-utilities>)
  3. Batch Entrez – по файлу со списком идентификаторов (<https://www.ncbi.nlm.nih.gov/sites/batchentrez>)
  4. FTP (<ftp://ftp.ncbi.nlm.nih.gov/>)
- ENA**
1. Search & Browse (<https://www.ebi.ac.uk/ena/browse>)
  2. Advanced search (<https://www.ebi.ac.uk/ena/browser/advanced-search>)
  3. REST URLs (<https://www.ebi.ac.uk/ena/browse/data-retrieval-rest>)
- DDBJ**
1. Search & Analysis (<https://www.ddbj.nig.ac.jp/services-e.html>)
  2. ARSA (<http://ddbj.nig.ac.jp/arsa>)
  3. GetEntry (<http://getentry.ddbj.nig.ac.jp/top-e.html>)
  4. FTP (<ftp://ftp.ddbj.nig.ac.jp/>)