



# **Нуклеотидные банки данных**

**Дарья Владимировна Диброва**

к.б.н.,

с.н.с. НИИ ФХБ имени А.Н. Белозерского МГУ имени М.В. Ломоносова,  
доцент ФББ МГУ имени М.В. Ломоносова

13 октября 2026 г.

# **1. Нуклеотидные последовательности и где они обитают**

# Разные типы последовательностей

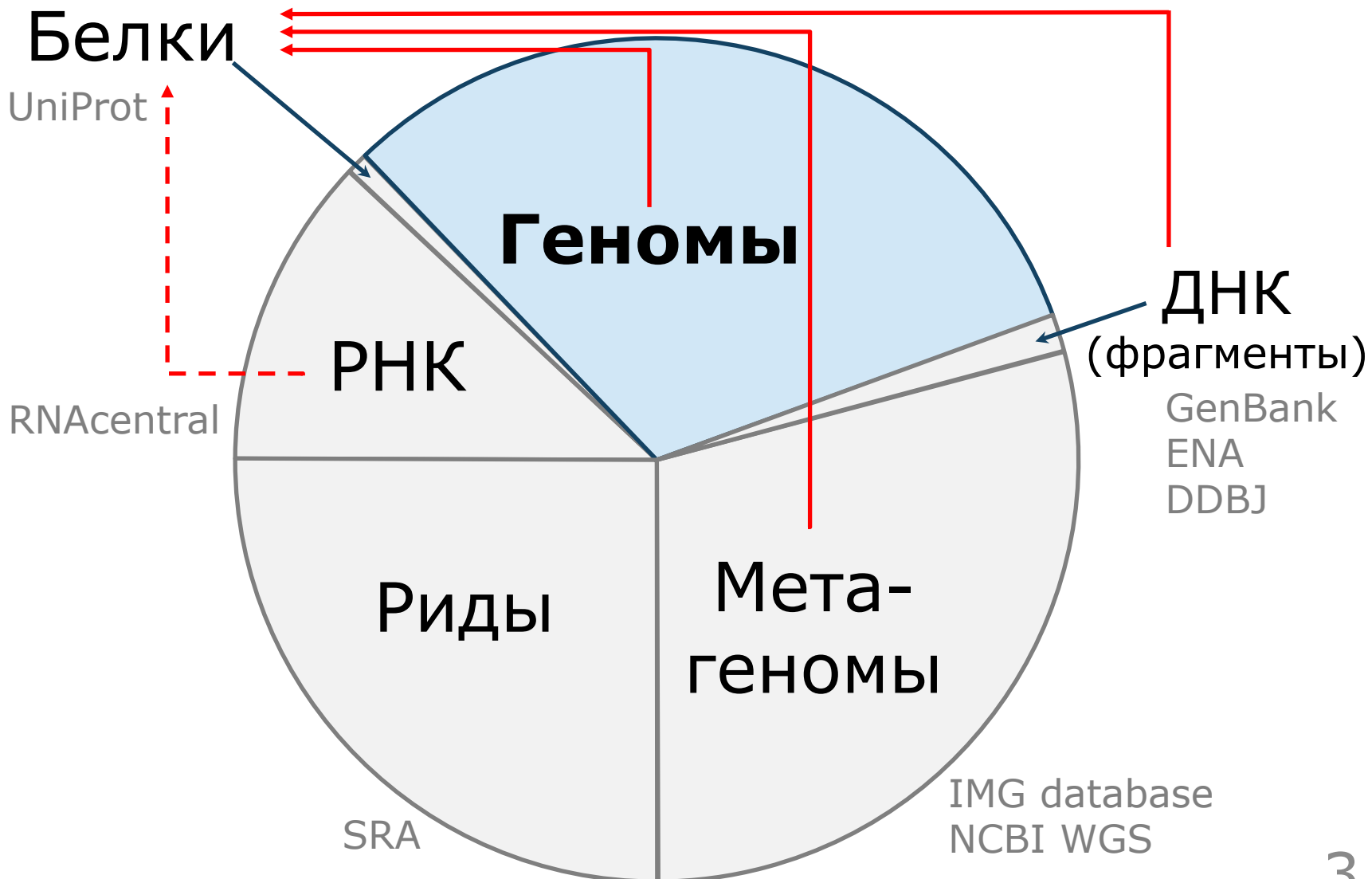
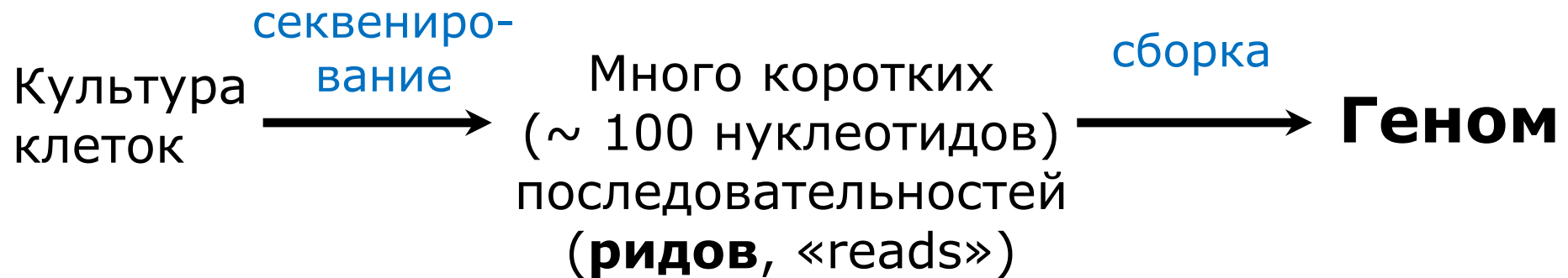


Схема достаточно упрощенная; размеры секторов не отвечают реальным объемам данных

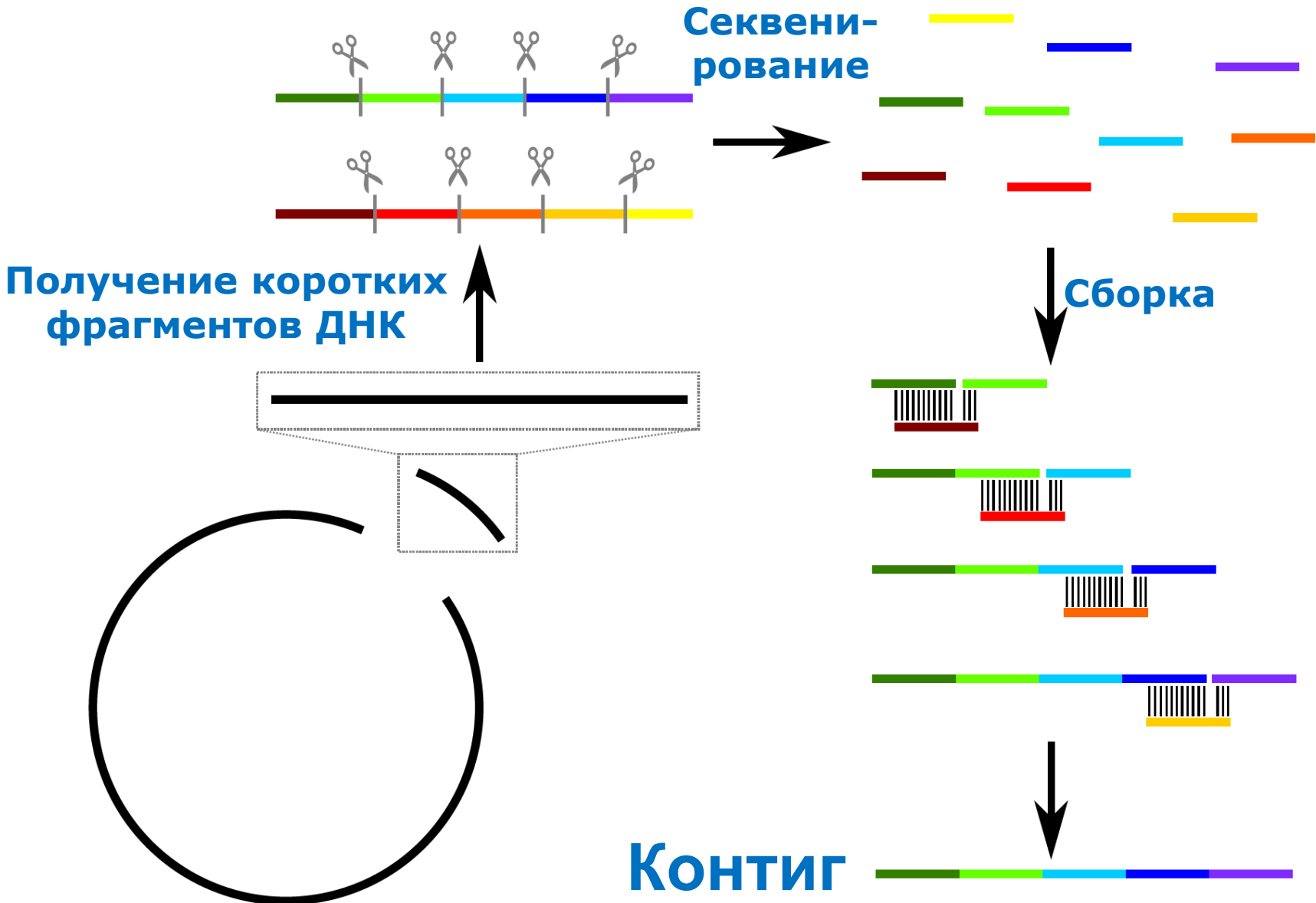
# Откуда берутся последовательности геномов?



**Геном** – совокупность генетической информации организма.

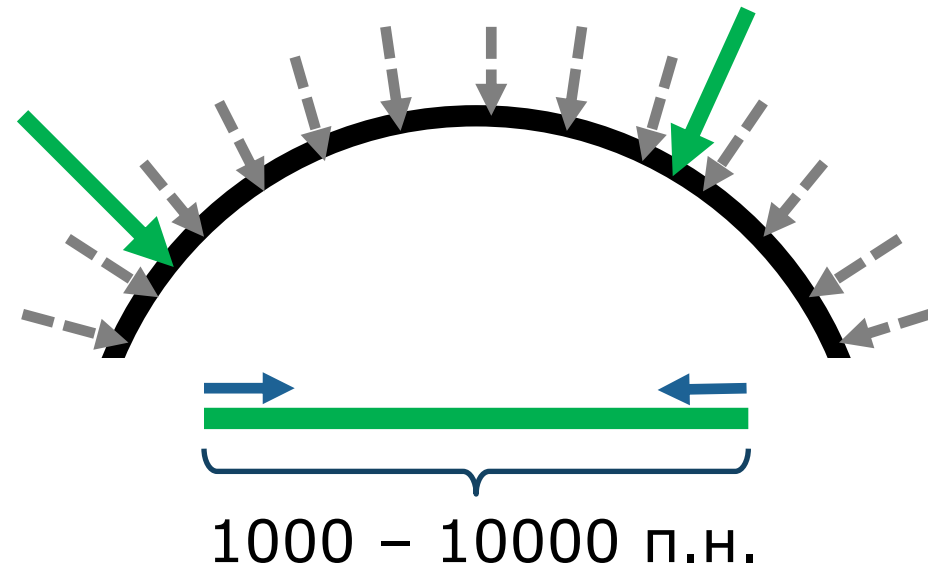
Секвенировать и собирать можно, конечно, не только целые «геномы», но и вообще любую ДНК/РНК

# Сборка генома (очень упрощенно)



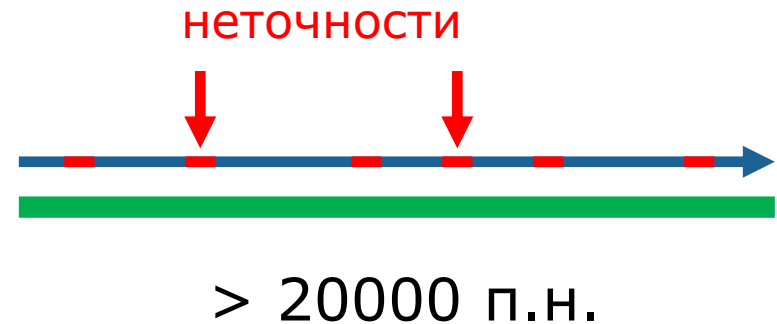
# Scaffolding: от контигов до хромосом

## Mate-pair sequencing



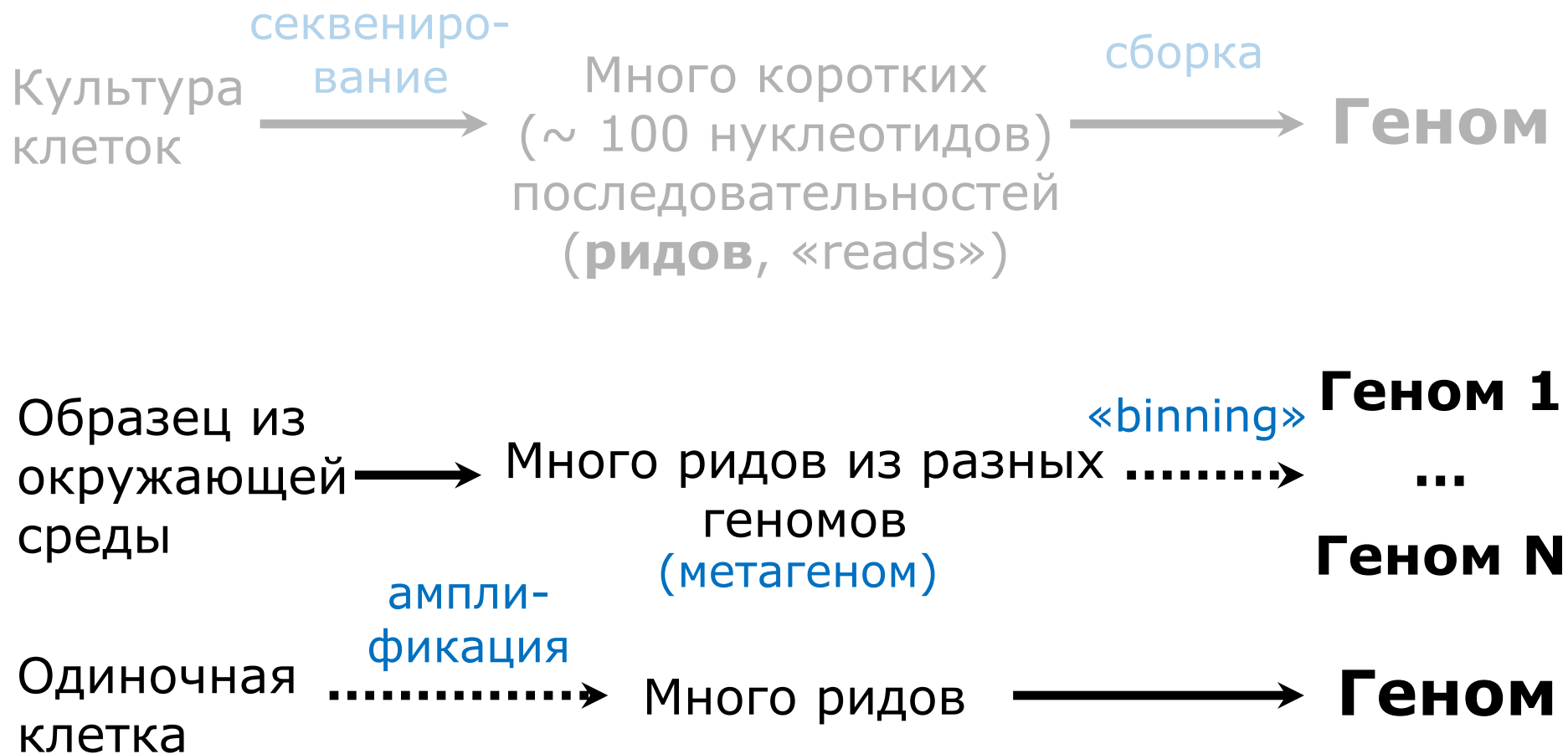
- Разрезание ДНК на более длинные куски
- Секвенирование с концов (получаются парные риды, **mate pair**)
- Учет этой информации при сборке

## Single molecule real time sequencing

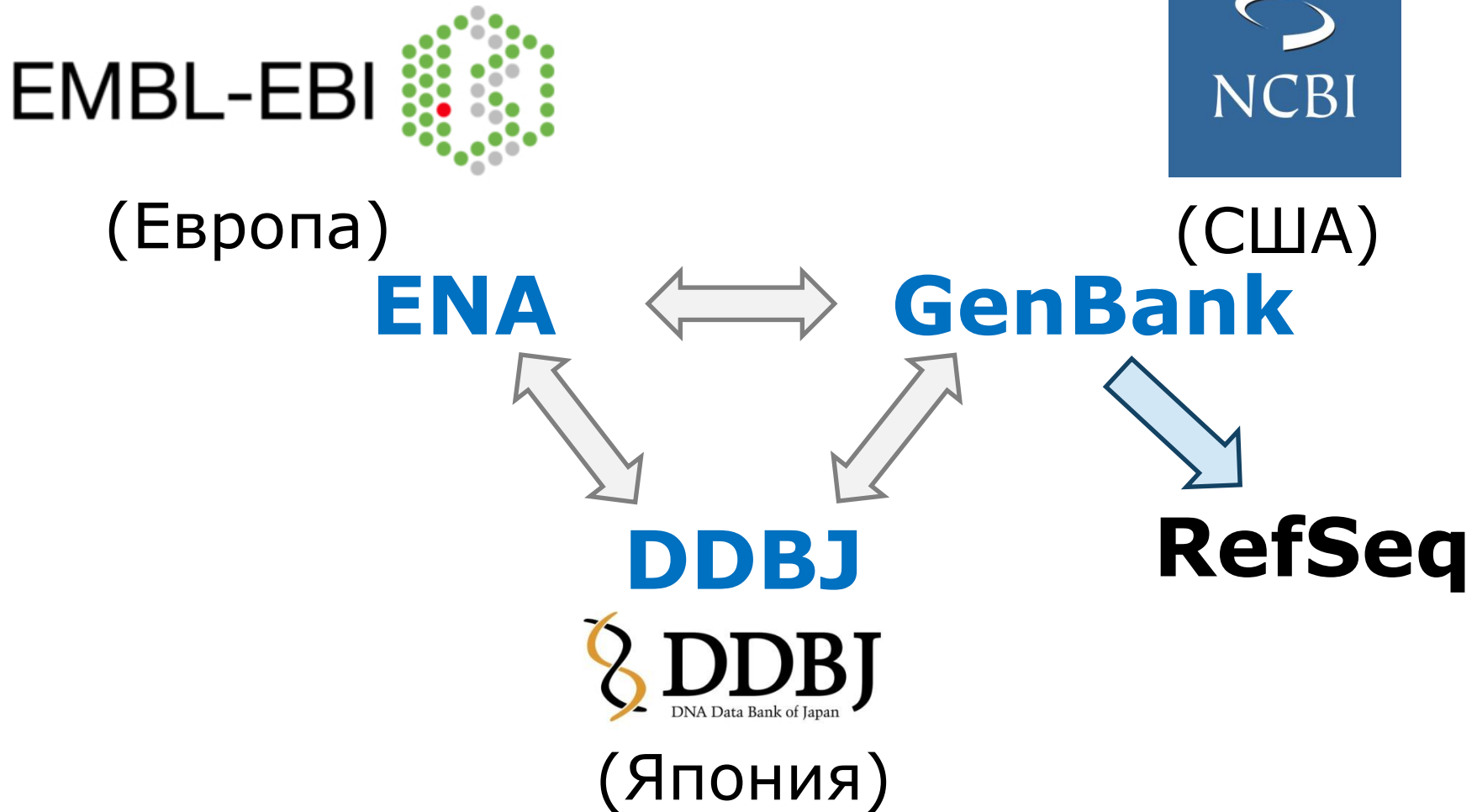


- Такие секвенаторы достаточно дорого стоят
- Точность секвенирования низкая, но для нужд сборки ее может хватать

# Нет культуры клеток — нет генома?



# International Nucleotide Sequence Database Collaboration



# Размер GenBank и WGS

Число последовательностей

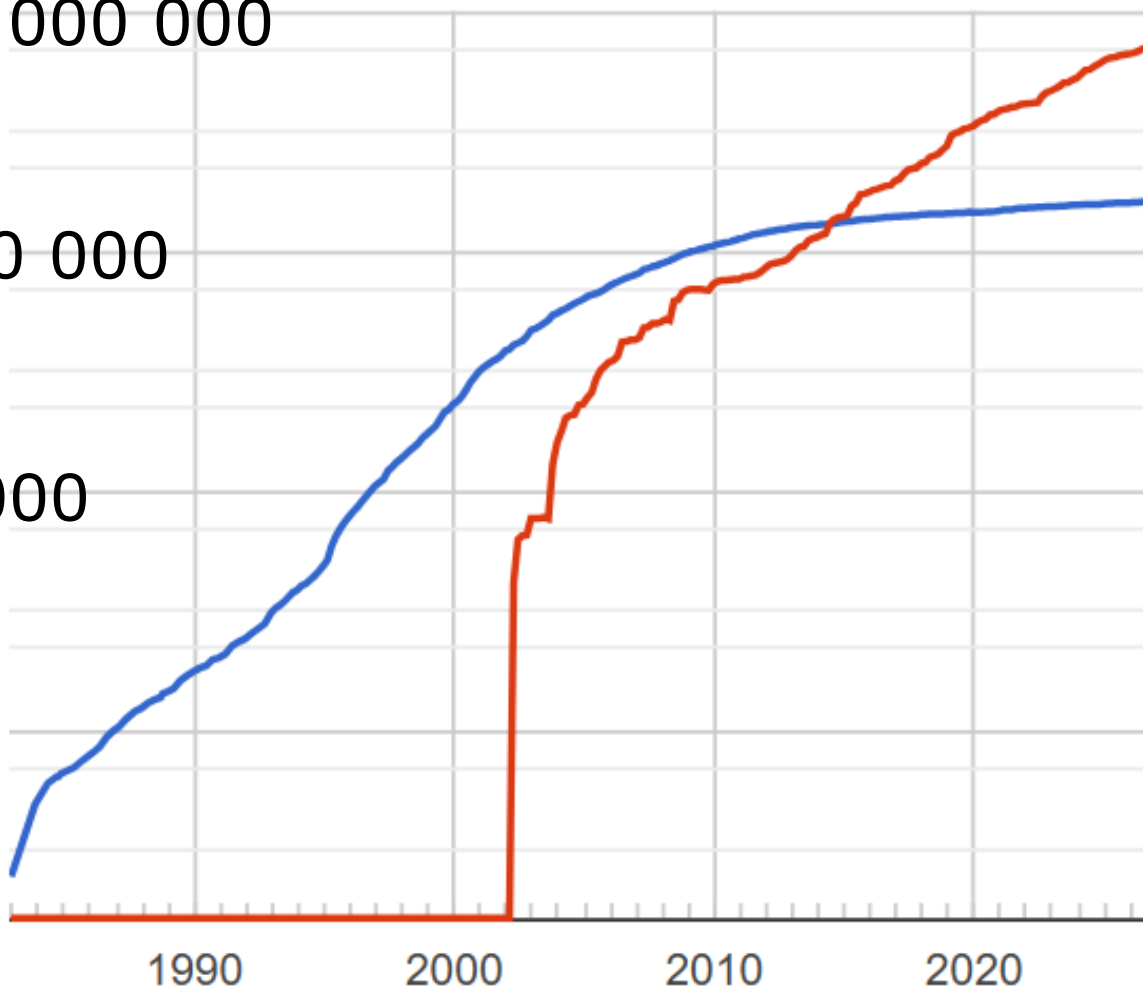
10 000 000 000

100 000 000

1 000 000

10 000

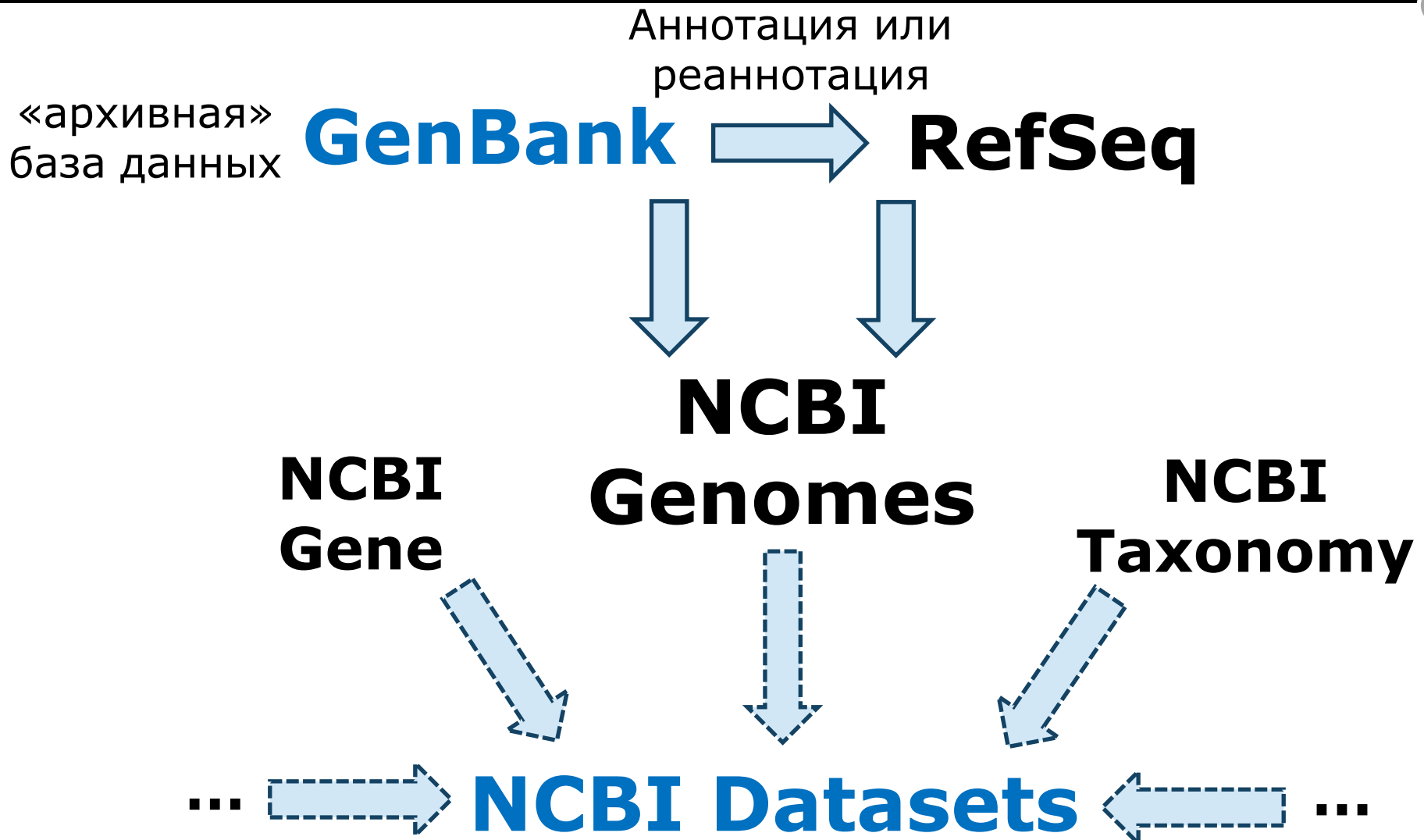
0



**WGS** (Whole  
Genome Shotgun)

**GenBank**

# NCBI Genomes, NCBI Datasets



## **2. Сборки геномов и система для их поиска**

# Важные параметры сборок

Две сборки *Deinococcus radiodurans* ATCC 13939:

| Assembly     | Genome size | Number of chromosomes | Number of scaffolds | Scaffold N50 | Scaffold L50 | Number of contigs | Contig N50 | Contig L50 | Assembly level  |
|--------------|-------------|-----------------------|---------------------|--------------|--------------|-------------------|------------|------------|-----------------|
| ASM2054668v1 | 3.3 Mb      | 4                     | 4                   | 2.6 Mb       | 1            | 4                 | 2.6 Mb     | 1          | Complete Genome |
| ASM68789v1   | 3.2 Mb      | -                     | -                   | -            | -            | 51                | 150.8 Kb   | 7          | Contig          |

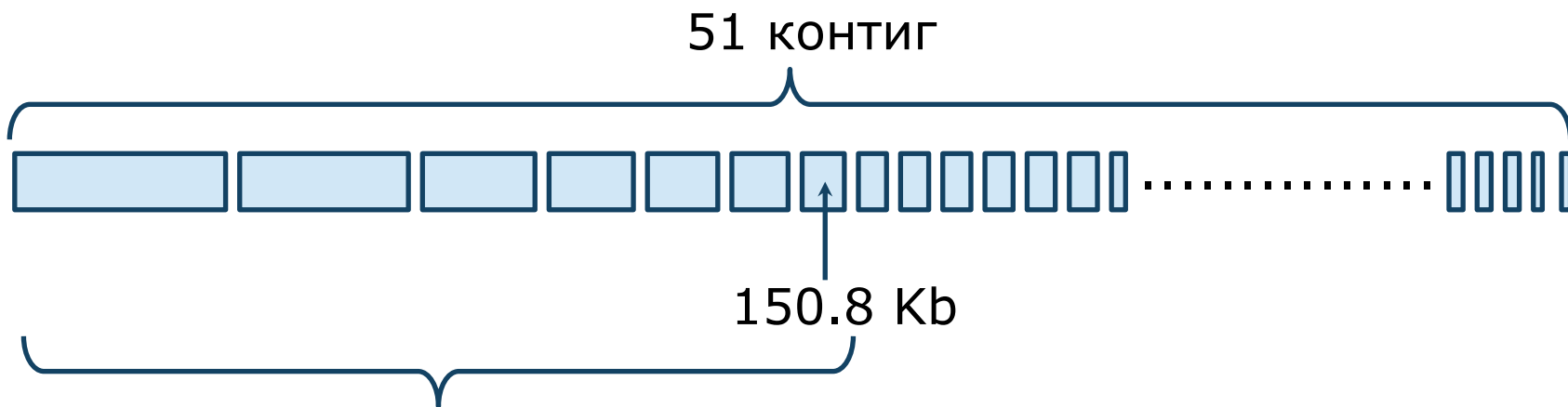
**N50:** Длина контига, для которого половина (**50%**) всех нуклеотидов сборки содержится в контигах такой и большей длины

**L50:** Число контигов (наименьшее), в которых содержится половина (**50%**) всех нуклеотидов сборки

# Параметры N50 и L50

Две сборки *Deinococcus radiodurans* ATCC 13939:

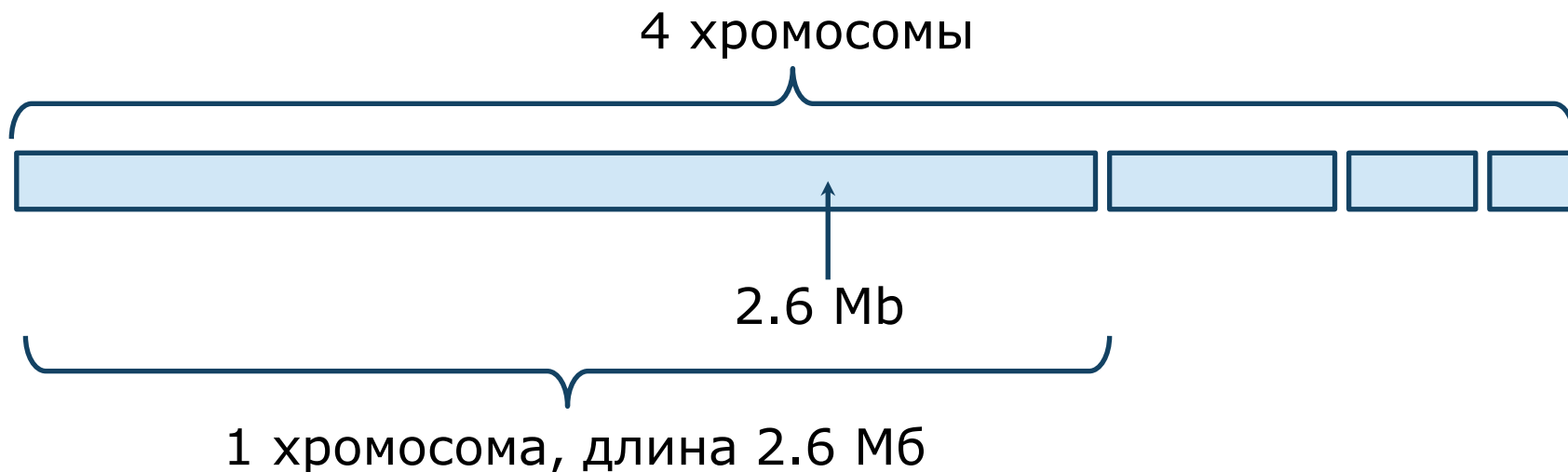
| Assembly     | Genome size | Number of chromosomes | Number of scaffolds | Scaffold N50 | Scaffold L50 | Number of contigs | Contig N50 | Contig L50 | Assembly level  |
|--------------|-------------|-----------------------|---------------------|--------------|--------------|-------------------|------------|------------|-----------------|
| ASM2054668v1 | 3.3 Mb      | 4                     | 4                   | 2.6 Mb       | 1            | 4                 | 2.6 Mb     | 1          | Complete Genome |
| ASM68789v1   | 3.2 Mb      | -                     | -                   | -            | -            | 51                | 150.8 Kb   | 7          | Contig          |



# Параметры N50 и L50

Две сборки *Deinococcus radiodurans* ATCC 13939:

| Assembly     | Genome size | Number of chromosomes | Number of scaffolds | Scaffold N50 | Scaffold L50 | Number of contigs | Contig N50 | Contig L50 | Assembly level  |
|--------------|-------------|-----------------------|---------------------|--------------|--------------|-------------------|------------|------------|-----------------|
| ASM2054668v1 | 3.3 Mb      | 4                     | 4                   | 2.6 Mb       | 1            | 4                 | 2.6 Mb     | 1          | Complete Genome |
| ASM68789v1   | 3.2 Mb      | -                     | -                   | -            | -            | 51                | 150.8 Kb   | 7          | Contig          |



# Значения N50 и L50: scaffold vs. contig

Сборка гриба *Phytophthora agathidicida*:

| Assembly     | Genome size | Number of chromosomes | Number of scaffolds | Scaffold N50 | Scaffold L50 | Number of contigs | Contig N50 | Contig L50 | Assembly level |
|--------------|-------------|-----------------------|---------------------|--------------|--------------|-------------------|------------|------------|----------------|
| ASM2572299v1 | 57 Mb       | 10                    | 10                  | 7.5 Mb       | 4            | 17                | 4.9 Mb     | 5          | Chromosome     |

```
>CP106971.1 Phytophthora agathidicida strain 3770 chromosome 1
```

```
<...>
```

```
AGGGTAAAGGGAACCGCCCTTTACAAACCCAGAAGACACAAACCACCCTCCCTTTAGTGACATGCATAT  
CTTAGCCTACATTTTCAGTTACGGAGAGGTTACTAACTGGTAAATATGACACATANNNNNNNNNNTTCCG  
CGCCAGCCACCGCATTTTCGGGGGTGGCACTGCGCCAACACCGGCTGCGGCACACACAGTGGCAATGCCT
```

```
<...>
```

Контиг 1

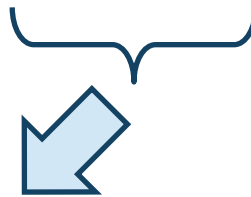
NNN – обозначение соединения КОНТИГОВ

Контиг 2

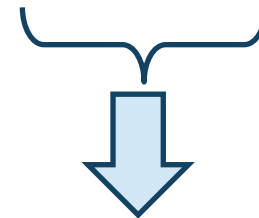
# Значения N50 и L50: scaffold vs. contig

Сборка гриба *Phytophthora agathidicida*:

| Assembly     | Genome size | Number of chromosomes | Number of scaffolds | Scaffold N50 | Scaffold L50 | Number of contigs | Contig N50 | Contig L50 | Assembly level |
|--------------|-------------|-----------------------|---------------------|--------------|--------------|-------------------|------------|------------|----------------|
| ASM2572299v1 | 57 Mb       | 10                    | 10                  | 7.5 Mb       | 4            | 17                | 4.9 Mb     | 5          | Chromosome     |



- 1) В **четырёх** скэффолдах содержится 50% нуклеотидов сборки;
- 2) Длина наименьшего скэффолда из них равна **7.5 Mb**.



- 1) «Разрежем» скэффолды по участкам NNN...;
- 2) Получим 17 контигов, в **пяти** из которых содержится 50% нуклеотидов сборки;
- 3) Длина наименьшего из них **4.9 Mb**.

# Интерфейс поиска в NCBI Genomes через NCBI Datasets

## Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

Selected taxa  
Deinococcus radiodurans R1 = ATCC 13939 = DSM 20539 Enter one or more taxonomic names

Filters

DOWNLOAD

Select columns

11 Genomes

Rows per page

20

1-11 of 11

| <input type="checkbox"/> | Assembly     | GenBank         | RefSeq          | Scientific name                  | Modifier             | Annotation               | Action |
|--------------------------|--------------|-----------------|-----------------|----------------------------------|----------------------|--------------------------|--------|
| <input type="checkbox"/> | ASM2054668v1 | GCA_020546685.1 | GCF_020546685.1 | Deinococcus radiodurans R1 = ... | ATCC 13939 (strain)  | NCBI RefSeq<br>Submitter | ⋮      |
| <input type="checkbox"/> | ASM856v1     | GCA_000008565.1 | GCF_000008565.1 | Deinococcus radiodurans R1 = ... | R1 (strain)          | NCBI RefSeq<br>Submitter | ⋮      |
| <input type="checkbox"/> | ASM2054670v1 | GCA_020546705.1 | GCF_020546705.1 | Deinococcus radiodurans R1 = ... | ATCC 13939 substr... | NCBI RefSeq              | ⋮      |
| <input type="checkbox"/> | ASM163882v1  | GCA_001638825.1 | GCF_001638825.1 | Deinococcus radiodurans R1 = ... | R1 (strain)          | NCBI RefSeq<br>Submitter | ⋮      |
| <input type="checkbox"/> | ASM2054702v1 | GCA_020547025.1 | GCF_020547025.1 | Deinococcus radiodurans R1 = ... | ATCC 13939 substr... | NCBI RefSeq              | ⋮      |
| <input type="checkbox"/> | ASM2054680v1 | GCA_020546805.1 | GCF_020546805.1 | Deinococcus radiodurans R1 = ... | ATCC 13939 substr... | NCBI RefSeq              | ⋮      |
| <input type="checkbox"/> | ASM2054708v1 | GCA_020547085.1 | GCF_020547085.1 | Deinococcus radiodurans R1 = ... | ATCC 13939 substr... | NCBI RefSeq              | ⋮      |
| <input type="checkbox"/> | ASM2054678v1 | GCA_020546785.1 | GCF_020546785.1 | Deinococcus radiodurans R1 = ... | ATCC 13939 substr... | NCBI RefSeq              | ⋮      |
| <input type="checkbox"/> | ASM2137829v1 | GCA_021378295.1 | GCF_021378295.1 | Deinococcus radiodurans R1 = ... | R1 (strain)          | NCBI RefSeq<br>Submitter | ⋮      |
| <input type="checkbox"/> | ASM68789v1   | GCA_000687895.1 | GCF_000687895.1 | Deinococcus radiodurans R1 = ... | ATCC 13939 (strain)  | NCBI RefSeq              | ⋮      |
| <input type="checkbox"/> | ASM68598v1   | GCA_000685985.1 | GCF_000685985.1 | Deinococcus radiodurans R1 = ... | DSM 20539 (strain)   | NCBI RefSeq              | ⋮      |

1 организм  
!= 1 геном

Сбор-  
ки

# Фильтрация результатов

Filters

INCLUDE ONLY

☐ Reference genomes

☐ Annotated genomes

☐ Annotated by NCBI RefSeq

☐ Annotated by GenBank submitter

☐ Metagenome-assembled genomes (MAGs)

☐ Type material i

EXCLUDE

☐ Atypical genomes

☐ Metagenome-assembled genomes (MAGs)

☐ Genomes from large multi-isolate projects

SEARCH WITHIN RESULTS

ASSEMBLY LEVEL

contig

scaffold

chromosome

complete

YEAR RELEASED

1980

2026

## Особенно полезные фильтры:

- Наличие аннотации (т.е. предсказания генов).
- Смотреть только референсные геномы.
- Выбор уровня сборки.
- Год выпуска (проверить, не появилось ли новых сборок в таксоне, который вы уже ранее изучали).

# В таблице по умолчанию не отображаются некоторые колонки

## Select columns

### Basic information

- ☒ Assembly
- ☐ BioProject
- ☐ BioSample
- ☒ GenBank
- ☒ Level
- ☒ RefSeq
- ☒ Release date
- ☒ Sequencing technology
- ☒ Submitter
- ☐ WGS accession

### Taxonomy

- ☐ Modifier
- ☒ Scientific name
- ☒ Tax ID
- ☐ Taxonomy check status
- ☐ Type material
- ☐ Annotation
- ☐ Annotation Date
- ☐ Annotation pipeline
- ☐ BUSCO
- ☐ Genes

### Assembly statistics

- ☒ CheckM completeness (%)
- ☐ CheckM contamination (%)
- ☐ CheckM marker set
- ☐ Chromosomes
- ☐ Contig N50 (kb)
- ☐ Contigs
- ☐ GC percent
- ☒ Scaffold N50 (kb)
- ☐ Scaffolds
- ☒ Size (Mb)

### Annotation

- ☒ Annotation
- ☐ Annotation Date
- ☐ Annotation pipeline
- ☐ BUSCO
- ☐ Genes
- ☒ Protein-coding
- ☐ Pseudogenes

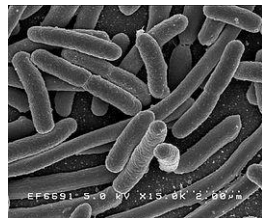
Например, можно посмотреть:

- Показать **параметры полноты генома** посчитанные CheckM (<https://ecogenomics.github.io/CheckM/>);
- N50 для контигов и скэффолдов** (если они очень похожи, вероятно скэффолдов мало и/или они короткие);
- Показать числа **белок-кодирующих генов и псевдогенов**.

# **3. Структурная и функциональная аннотация геномов**

# Рассмотрим на примере «прокариот»

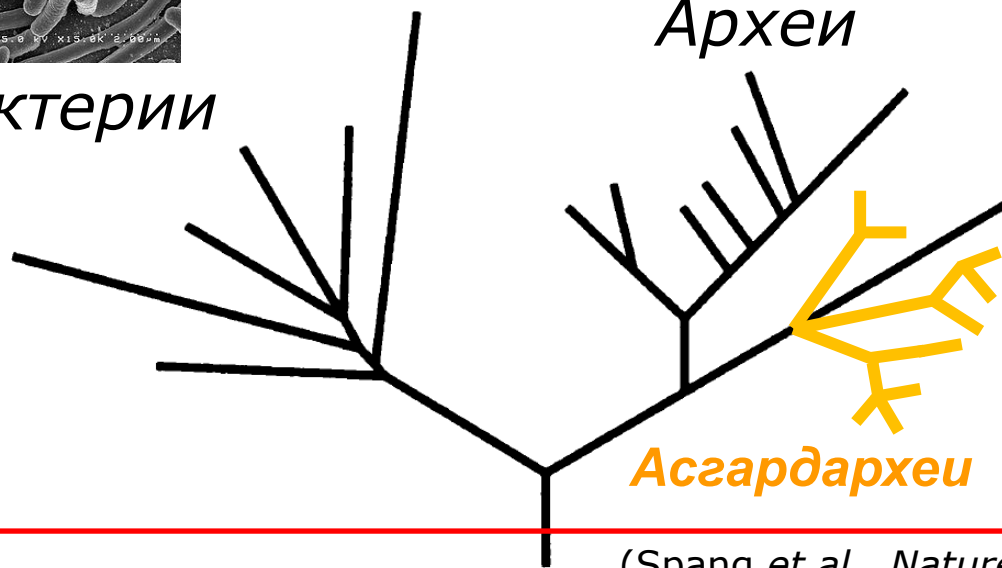
прокариоты



Бактерии



Археи



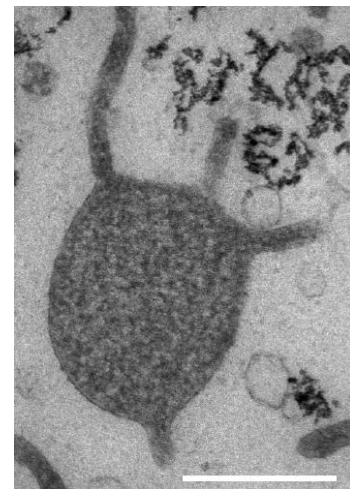
(Spang et al., *Nature*, 2015)

На основании сходства 16S рРНК  
все клеточные формы жизни  
разделяются на **три домена**

(Woese et al., *PNAS*, 1977; Woese et al., *PNAS*, 1990)

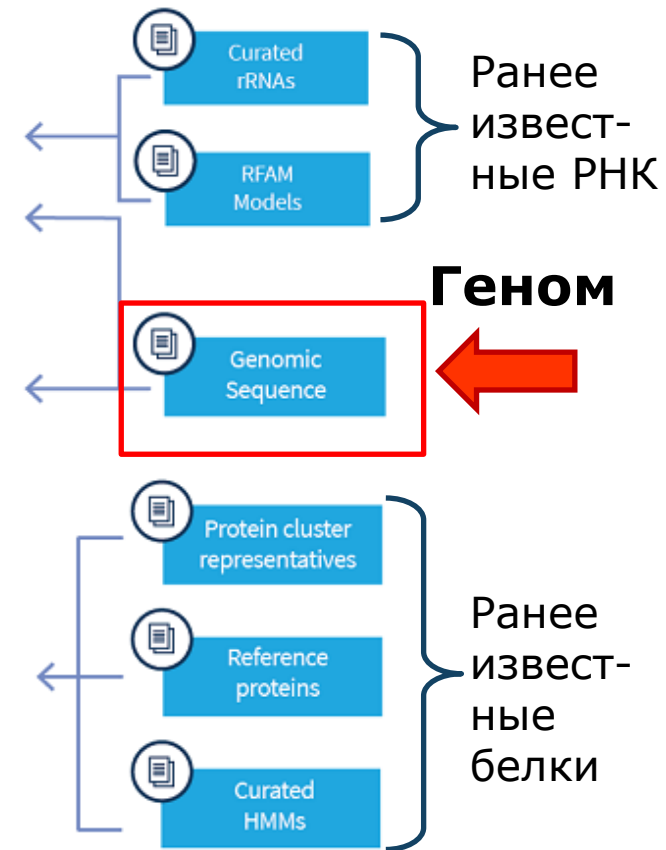
Изображения представителей доменов взяты из Википедии

Эукариоты

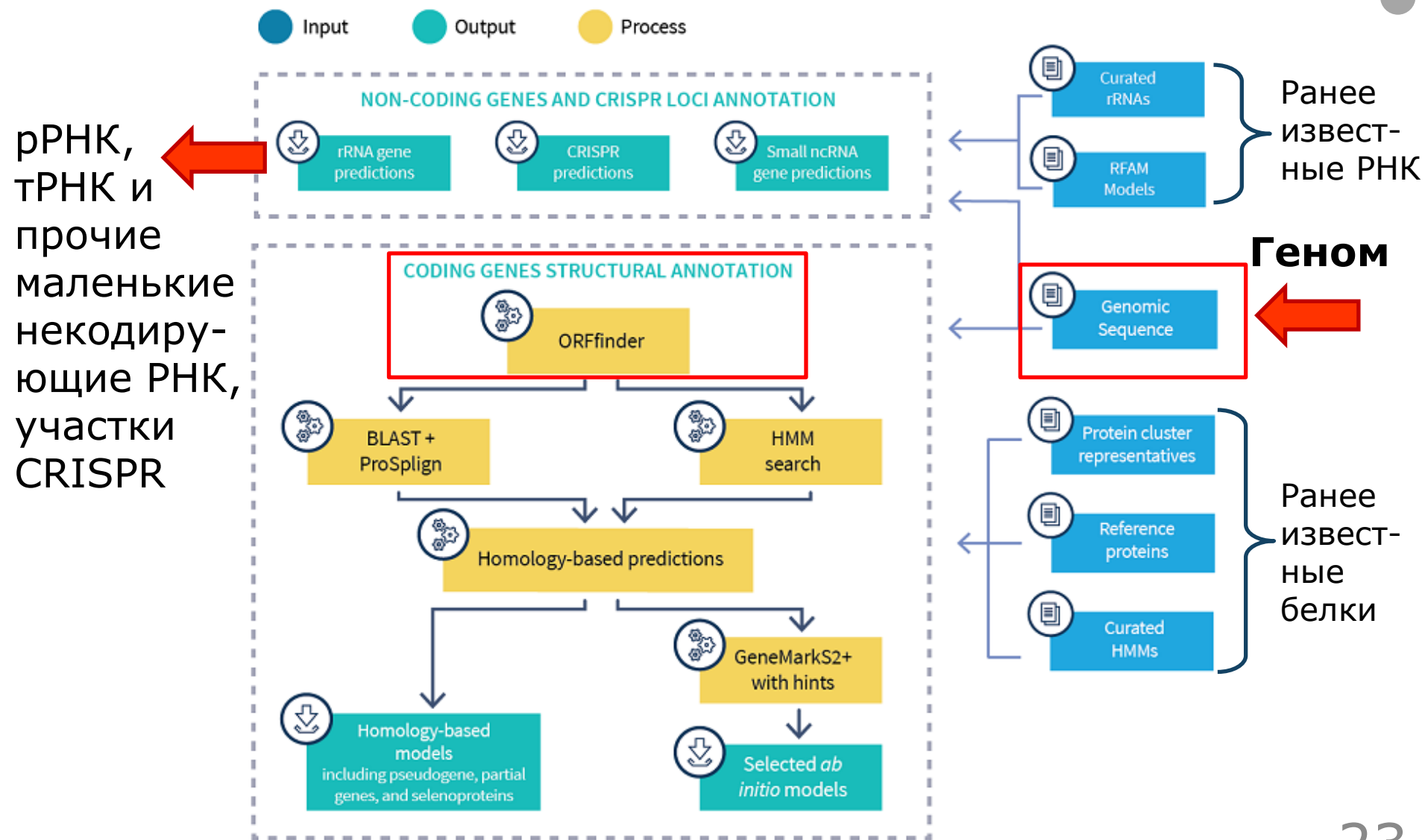


(Imachi et al., *Nature*, 2020)

# Структурная аннотация генома прокариот (поиск генов)



# Структурная аннотация генома прокариот (поиск генов)



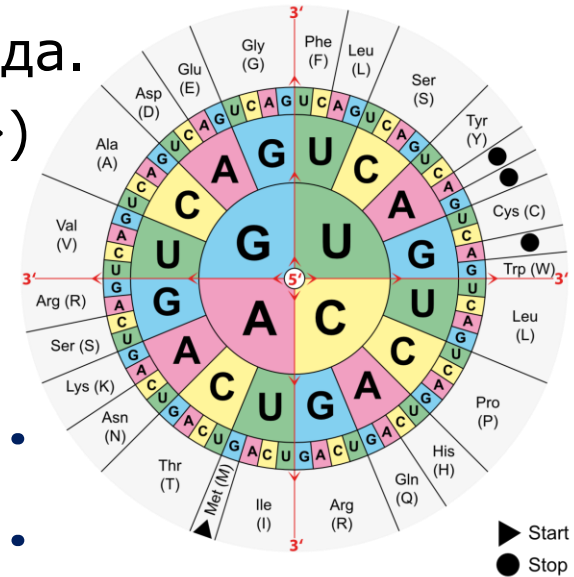
# Рамки считывания (reading frames)

- Начнем «считывать» с первого нуклеотида.
- Каждым трем нуклеотидами («триплету») соответствует одна аминокислота (по «генетическому коду»)

У Т Р И Р К Т . . .

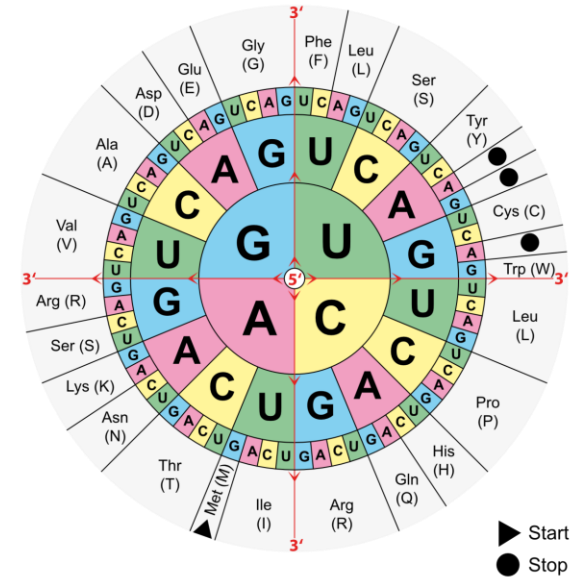
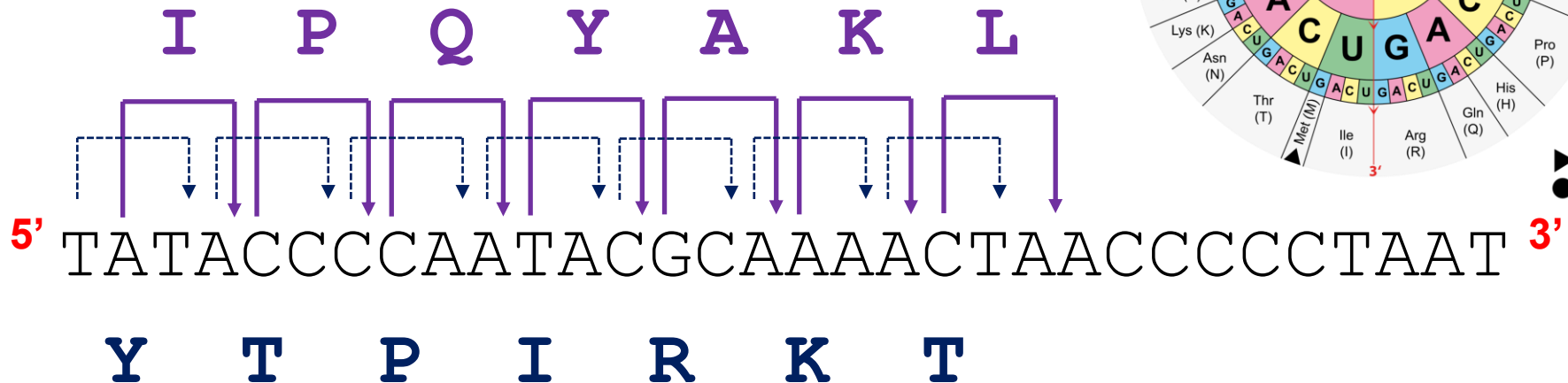


5' TATACCCCAATACGCAAAACCTAACCCCTAAT 3'



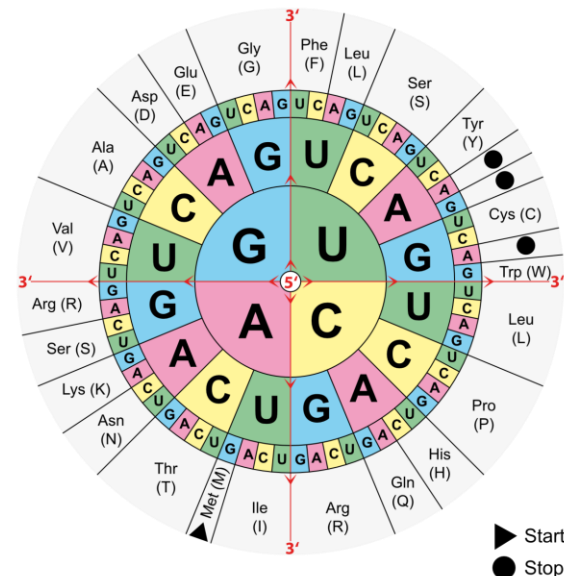
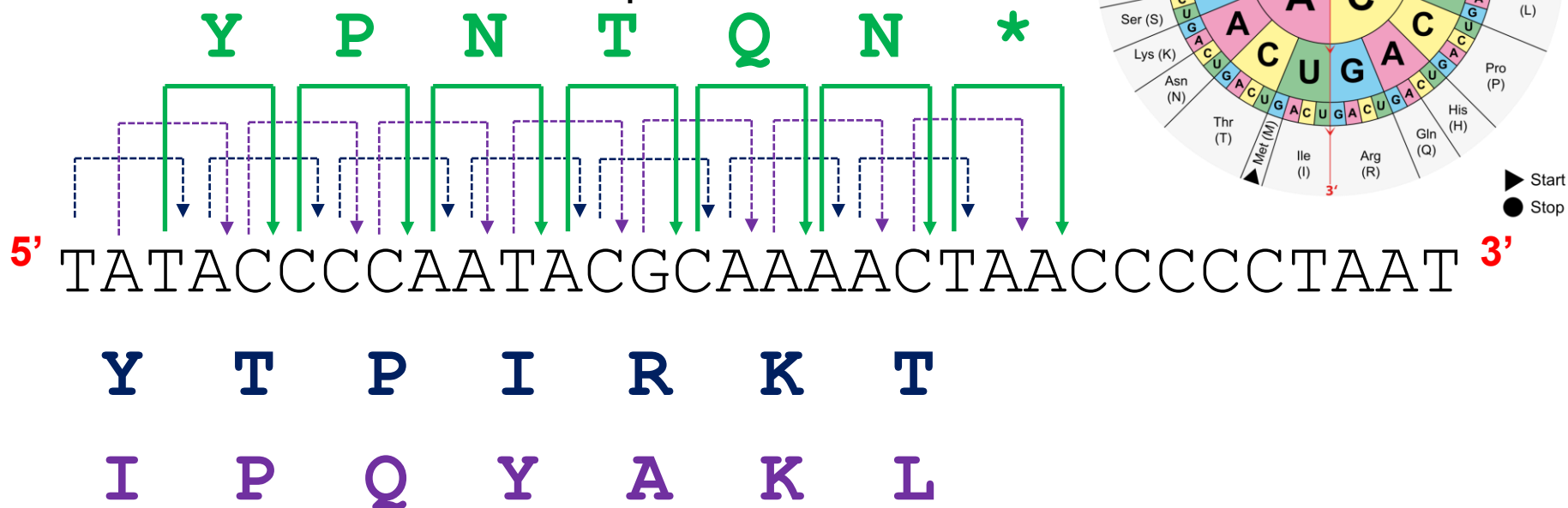
# Рамки считывания (reading frames)

- Начнем «считывать» со второго нуклеотида.
- Получающаяся последовательность белка не имеет ничего общего с первой!



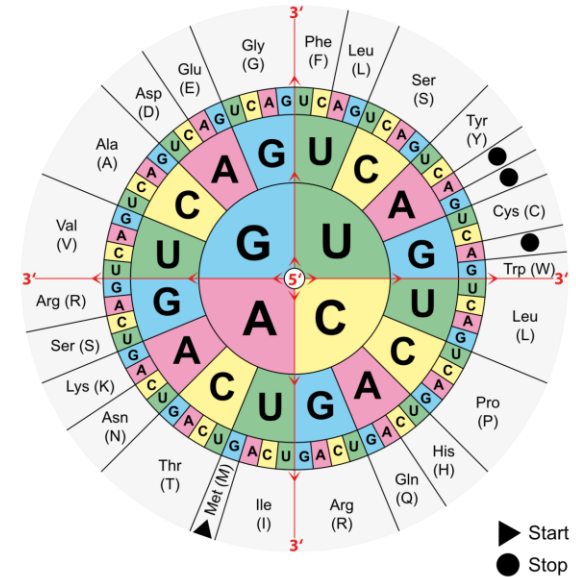
# Рамки считывания (reading frames)

- Начнем «считывать» с третьего нуклеотида.
- Получаем третью, совершенно новую последовательность белка.
- А если начать с четвертого?..



# Рамки считывания (reading frames)

- Но есть еще **комплементарная** цепь.
- И с нее можно точно так же считывать в трех «**рамках считывания**», которые дадут еще 3 последовательности белка.
- Обратите внимание на **порядок считывания и ориентацию** получающихся пептидов!



5' TATACCCSSAATACGCAAAASTAACCCCSSTAAT 3'

||||||||||||||||||||||||||||||||

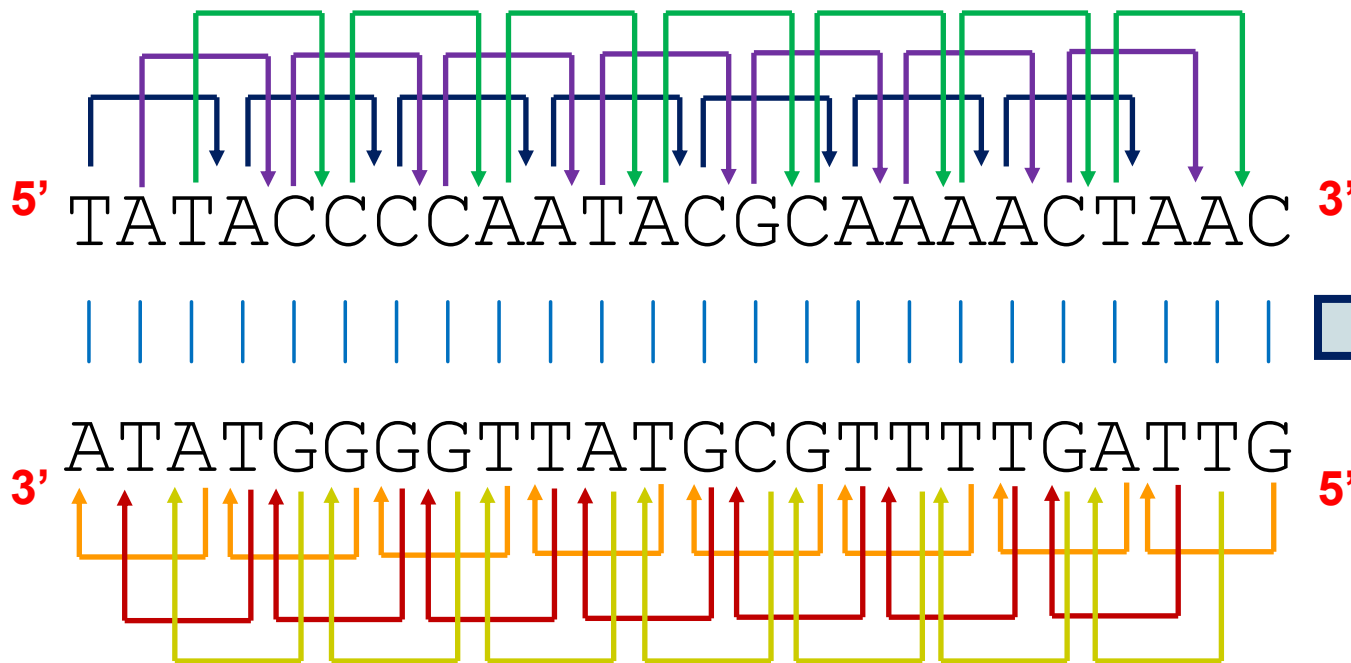
ATATGGGGTTATGCGTTTTTGATTGGGGGGATTA

3' ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ 5'

C- I G W Y A F S V G \* -N

# Рамки считывания (reading frames)

- Всего есть **6** рамок считывания: три **прямые** (forward) и три **обратные** (reverse)
- Самое распространенное обозначение: прямые **+1, +2, +3** и обратные **-1, -2, -3**



>frame +1  
YTPIRKTN

>frame +2  
IPQYAKLT

>frame +3  
YPNTQN\*X

>frame -1  
VSFAYWGI

>frame -2  
\*FCVLGYX

>frame -3  
LVLRIGVX

# Открытые рамки считывания (ORFs)

- **Открытая рамка считывания** – последовательность ДНК/РНК, которая *потенциально* может кодировать белок.

1. Начинается старт-кодоном;
2. Кончается стоп-кодоном;
3. Не содержит стоп-кодонов.

Стандартный старт-кодон: **ATG**

Стандартные стоп-кодоны: **TAA, TAG, TGA**

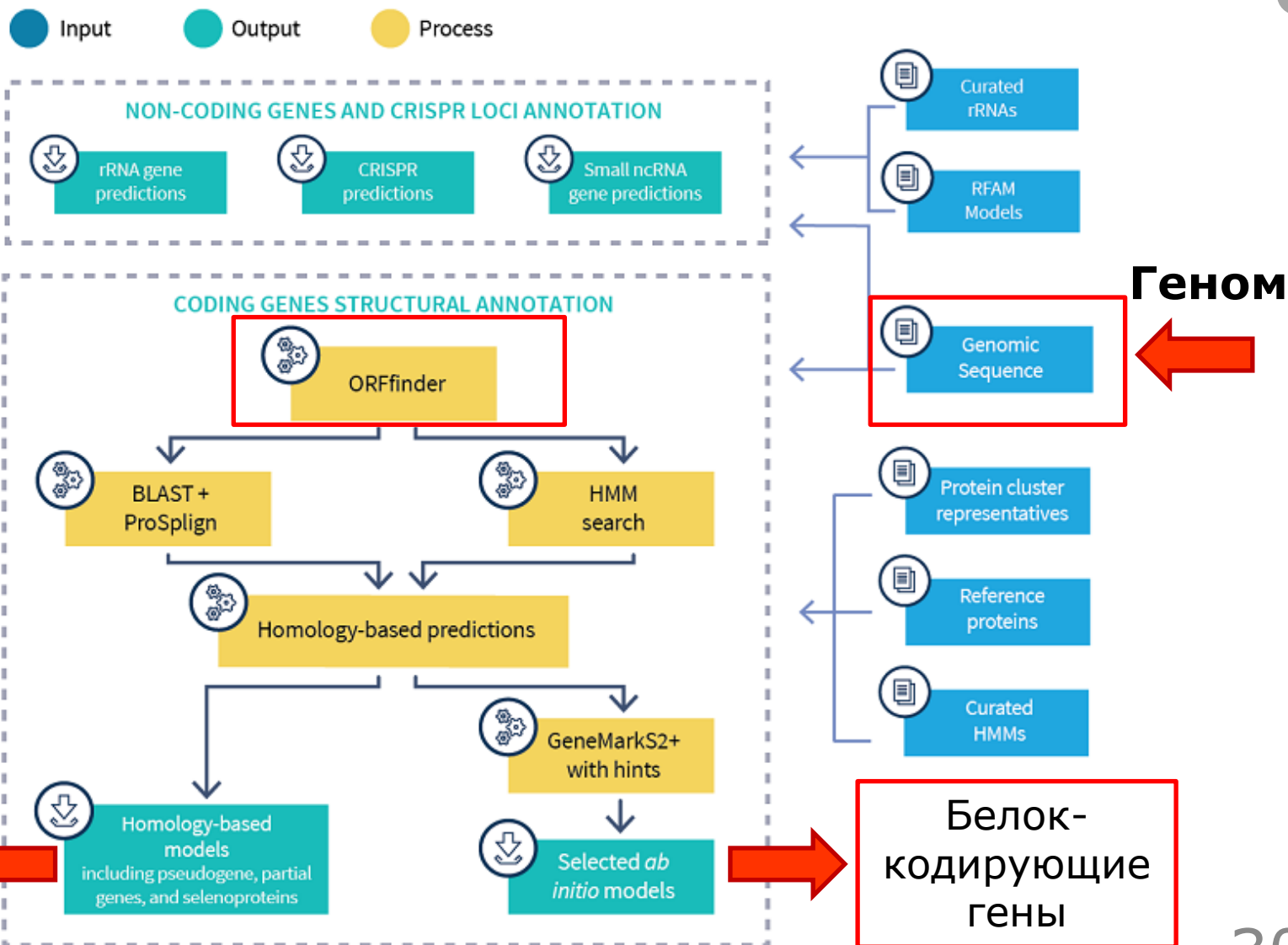
**!** Но даже в референсном геноме *Escherichia coli* (NC\_000913.3) далеко не всегда используется «стандартный» старт. И между видами стоп-кодоны могут варьировать. Как и генетический код (немного)

|            |             |            |          |
|------------|-------------|------------|----------|
| <b>ATG</b> | <b>3890</b> | <b>ATT</b> | <b>4</b> |
| <b>GTG</b> | <b>338</b>  | <b>CTG</b> | <b>2</b> |
| <b>TTG</b> | <b>80</b>   | <b>TTC</b> | <b>1</b> |

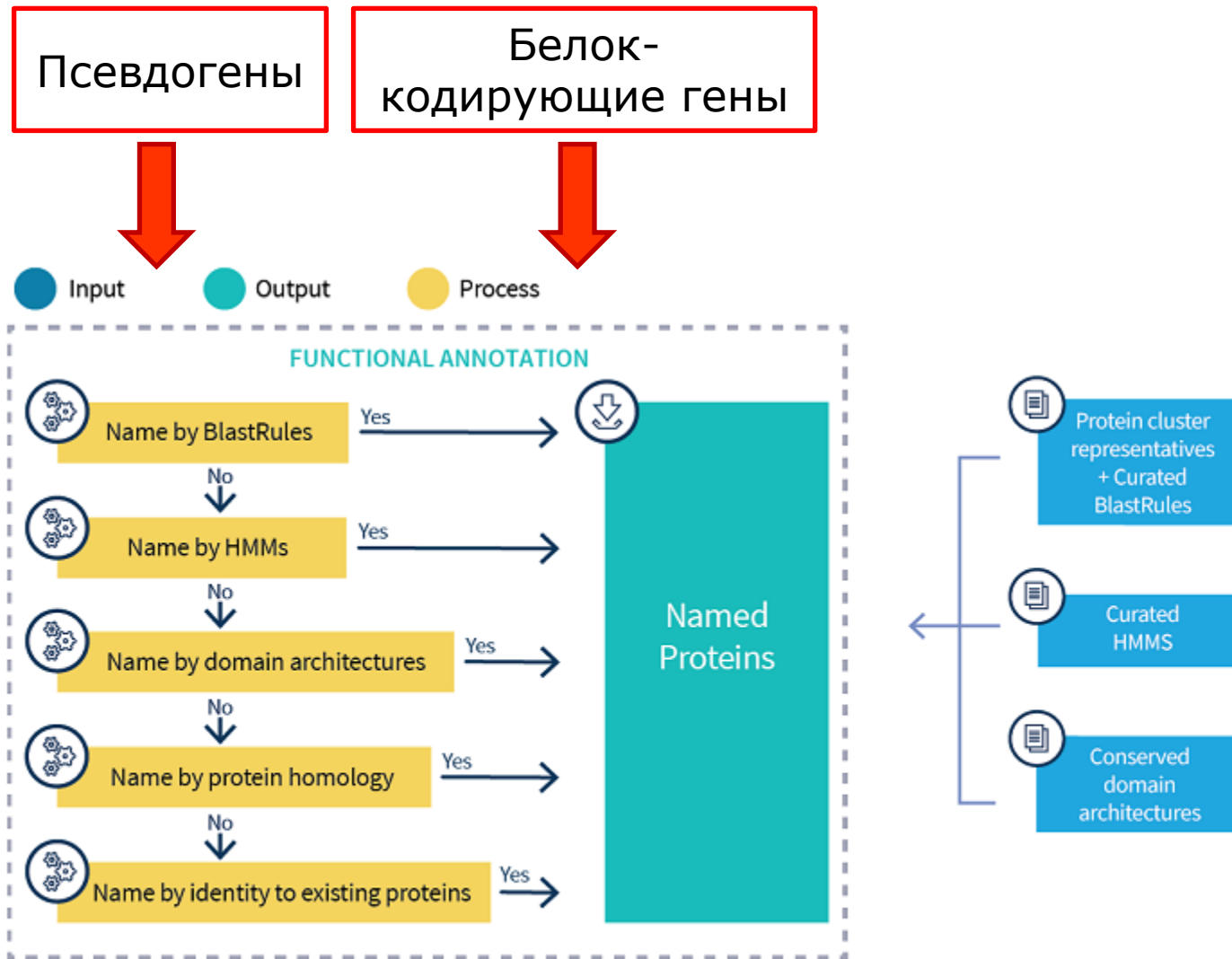
См. также: Belinky F., Babenko V.N., Rogozin I.B., Koonin E.V. Purifying and positive selection in the evolution of stop codons. *Scientific Reports*, 8:9260, **2018**.

Villegas A. & Kropinski A.. An analysis of initiation codon utilization in the Domain Bacteria - concerns about the quality of bacterial genome annotation. *Microbiology (Reading)*, 154(Pt 9):2559-2661, **2008**.

# Не все ORFs – гены белков



# Функциональная аннотация генома в NCBI (прокариоты)



## **4. Основной файл сборки генома (формат GBFF)**

# Доступные данные для сборки генома

**Download Package**

1 genome selected for download

Select file source

☒ All

☐ RefSeq only

☐ GenBank only

Select file types

☒ Genome sequences (FASTA)

☐ Annotation features (GTF)

☐ Annotation features (GFF)

☐ Sequence and annotation (GBFF)

☐ Transcripts (FASTA)

☐ Genomic coding sequences (FASTA)

☐ Protein (FASTA)

☐ Sequence report (JSONL)

☒ Assembly data report (JSONL)

Your selected data will be downloaded as a ZIP archive

Estimated file size is 2 MB

Name your file

ncbi\_dataset.zip

[Cancel](#) [Download](#)

Последовательности **ДНК генома**  
(хромосомы, плазмиды, скэффолды,  
континги)

Аннотация в виде таблицы

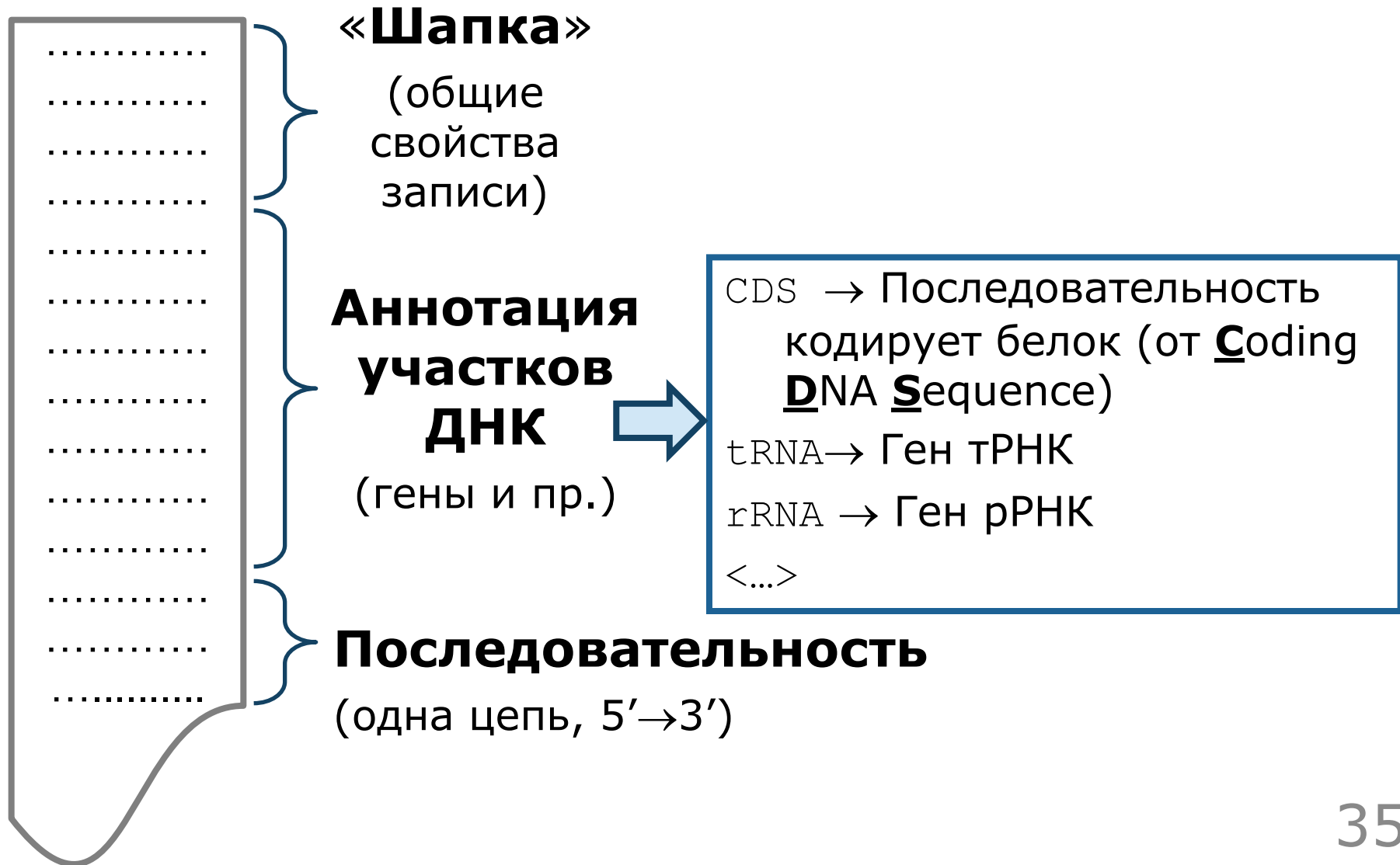
**Аннотированный геном**

Белок-кодирующие  
последовательности

Последовательности **белков**



# Файл генома в формате GenBank (.gbff)



# Фрагмент файла .gbff (аннотация гена тРНК и гена белка)

```
104730 gene { 60485..60561
104731 /locus_tag="E5E91_RS13600"
104732 /old_locus_tag="E5E91_13950"
104733 /db_xref="GeneID:69518947"
104734 tRNA { 60485..60561
104735 /locus_tag="E5E91_RS13600"
104736 /old_locus_tag="E5E91_13950"
104737 /product="tRNA-Met"
104738 /inference="COORDINATES: profile:tRNAscan-SE:2.0.7"
104739 /note="Derived by automated computational analysis using
104740 gene prediction method: tRNAscan-SE."
104741 /anticodon=(pos:60519..60521,aa:Met,seq:cat)
104742 /db_xref="GeneID:69518947"
104743 gene { complement(60635..61123)
104744 /locus_tag="E5E91_RS13605"
104745 /old_locus_tag="E5E91_13955"
104746 /db_xref="GeneID:69518948"
104747 CDS { complement(60635..61123)
104748 /locus_tag="E5E91_RS13605"
104749 /old_locus_tag="E5E91_13955"
104750 /inference="COORDINATES: similar to AA
104751 sequence:RefSeq:WP_017869241.1"
104752 /note="Derived by automated computational analysis using
104753 gene prediction method: Protein Homology."
104754 /codon_start=1
104755 /transl_table=11
104756 /product="hypothetical protein"
104757 /protein_id="WP_010889315.1"
104758 /db_xref="GeneID:69518948"
104759 /translation="MLFACPALFASASAGSIPGTQVSYKLAQDALTHGNASMVFVDAS
104760 EDPTGKTYMALVCDLGKTTFRFYSKESLGSRAGEKVLFFFRADDQGLRNAHGTLREYY
104761 KTGNLVLDFSYDSSELFRTFASTQKKVVVRVNRNGMASMTLTFPVKGFMVAYRAIGS
104762 CG"
```

# Запрограммированный сдвиг рамки при трансляции

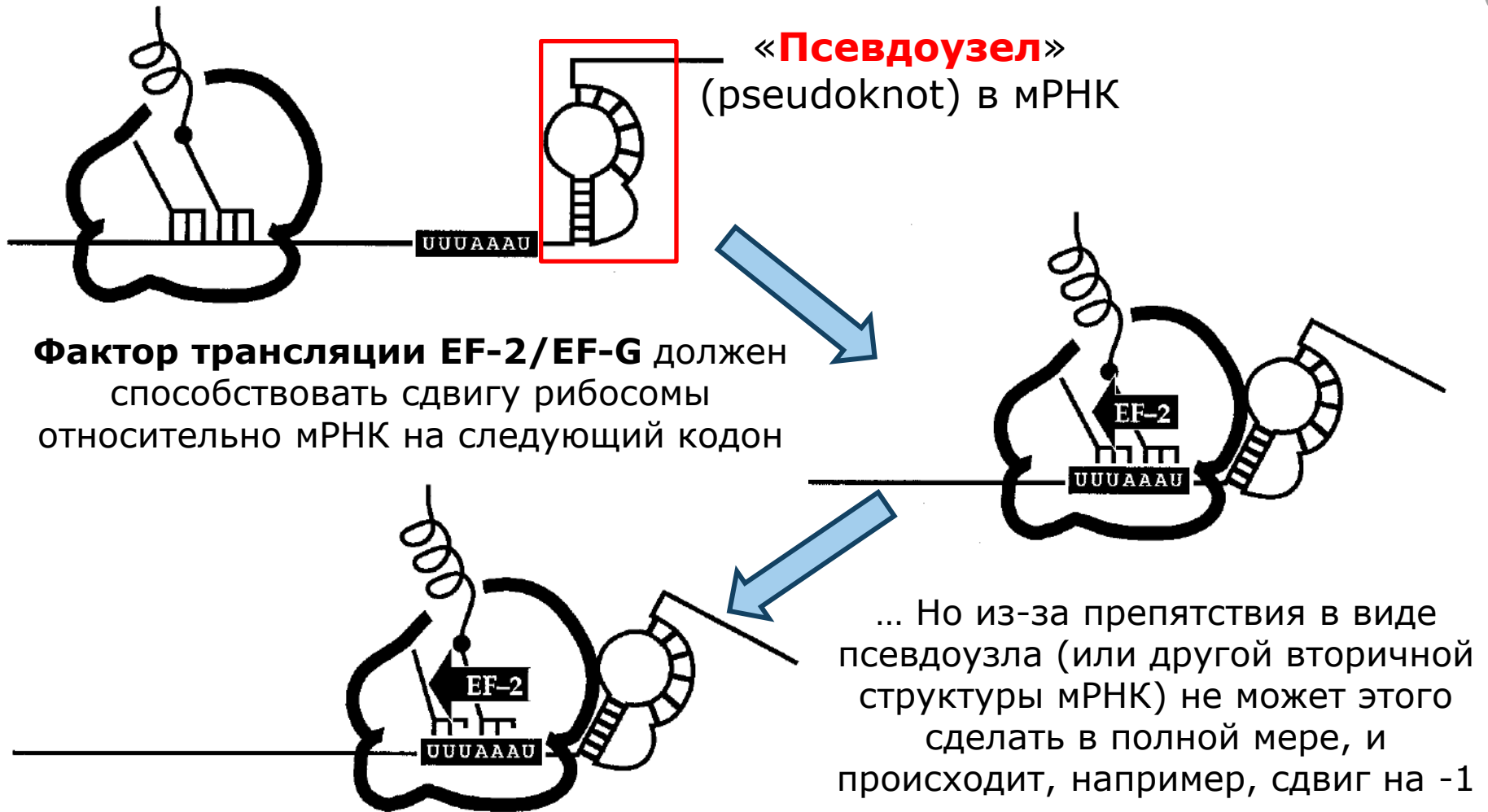


Рисунок: Farabaugh P.J. Programmed Translational Frameshifting. *Microbiological Reviews* 60(1):103–134, **1996**.  
См. также: Ketteler R. On programmed ribosomal frameshifting: the alternative proteomes. *Front Genet.* 3:242, **2012**.

Atkins J.F., Loughran G. Bhatt P.R., Firth A.E., Baranov P.V. Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. *Nucleic Acids Res.* 44(15):7007–7078, **2016**.

# Отражение этого процесса (у прокариот и вирусов) – слово `join`

Пример из генома *E. coli*: NC\_000913.3

|    | CDS |                                                                           |
|----|-----|---------------------------------------------------------------------------|
| 1  |     | <code>complement(join(3035184..3036206,3036208..3036282))</code>          |
| 2  |     | <code>/gene="prfB"</code>                                                 |
| 3  |     | <code>/locus_tag="b2891"</code>                                           |
| 4  |     | <code>/gene_synonym="ECK2886; supK"</code>                                |
| 5  |     | <code>/ribosomal_slippage</code>                                          |
| 6  |     | <code>/codon_start=1</code>                                               |
| 7  |     | <code>/transl_table=11</code>                                             |
| 8  |     | <code>/product="peptide chain release factor RF2"</code>                  |
| 9  |     | <code>/protein_id="NP_417367.1"</code>                                    |
| 10 |     | <code>/db_xref="UniProtKB/Swiss-Prot:P07012"</code>                       |
| 11 |     | <code>/db_xref="ASAP:ABE-0009488"</code>                                  |
| 12 |     | <code>/db_xref="ECOCYC:EG10762"</code>                                    |
| 13 |     | <code>/db_xref="GeneID:947369"</code>                                     |
| 14 |     | <code>/translation="MFEINPVNNRIQDLTERS DVLRGYLDYDAKKERLEE VNAELEQP</code> |
| 15 |     | <code>DVWNEPERA QALGKERS SLEAVVDTLDQMKGLEDVSG LLELAVEADDEETFNEAVAE</code> |
| 16 |     | <code>LDAL EEKLAQLEFR RMFSGEYDSADCYLDIQAGSGGTEAQDWASMLERMYLRWAESRG</code> |
| 17 |     | <code>FKTEII EESEGEVAGIKSVTIKISGDYAYGWLRTETGVHRLVRKSPFDSGGRRHTSFS</code>  |
| 18 |     | <code>SAFVYPEVDD DIDIEINPADLRIDVYRTSGAGGQHVNRTE SAVRITHIPTGIVTQCQN</code> |
| 19 |     | <code>DRSQHK NKDQAMKQMKAKLYELEMQKKNAEKQAMEDNKSDIGWGSQIRSYVLDDSR IK</code> |
| 20 |     | <code>DLRTGVETRNTQAVLDGSLDQFIEASLKAGL"</code>                             |

- **У вирусов** это регулярно происходит для записи альтернативных вариантов белков, считываемых с одного гена
- **У эукариот** через слово `join` объединяют экзоны

# Фрагмент файла .gbff

## (начало последовательности ДНК)

```
111850 ORIGIN
111851      1 tctttgctcg ccatacccaa agtctacacg ctgattttca cgtttccaga cctgcccctc
111852     61 tcgctactca gctctccaag tttgctcgct tgatgaatga tcaaatcttt taaagataaa
111853    121 aacaaaaaaa tgatgatcat catcataggg cggcatgttc tcacgtctag gacggcgttt
111854    181 ttcgattttt ctccacaaaag tgccacaggt aaacggccccg aattccacaa agtgccacag
111855    241 gcaaacggcc tttttctcca caaagtgcc aaggcaaacg gtccaaatgc caciaaagtgc
111856    301 cacaggtaat tccacaaagt gccacaggca ggggcattca aactccacaa agtgccgcag
111857    361 gcgagggcgc tcaaaactcca caaagtgcc aaggtaattc caciaaagtgc cacaggcaaa
111858    421 cggcccaa at tccacaaagt gccacaggta attccacaaa gtgcccacagg caggggtcatt
111859    481 caaactccac aaagtgccac aggtattcgt ctagacaatt acaggtctat taagaaaatt
111860    541 ttttgctgctg aagaatcatc tgtttctacg caatctaatt ttttcggtac taccagttgt
111861    601 ccagcacttc ctgacgttca aggaggaaca gtcgattgca gacctttcgg atatggtgag
111862    661 cgctgtgaaa acactcacgg tgttcaacca tgccggaggc gcaggcaaga ccagtttgac
111863    721 tttgaatggt ggctatgagc tggcacgggg cgggctgcgg gtgctgctgc tcgacctcga
111864    781 cccacaggcg aatttgaccg gctggctggg catttctggg gtcacgcggg agatgacggg
111865    841 ttaccacagt gccgtggacg gtcagccgct gccttcccc a gtcaaggcgt ttggcctgga
111866    901 cgttattcct gcacatgtca gtctggccgt ggcggaaggc cagatgatgg gccgagtcgg
111867    961 tgctcagggg cgtctgcgcc gcgctttagc agaggtatcc ggcgactacg acgtagccct
111868   1021 cattgatagt ccgcctagcc ttggacagct cgccatcttg gccgcgctgg ccgctgacca
111869   1081 gatgatcgtg ccggtaccga cgcgacaaaa aggcctagat gctttgcctg gggttacaggg
111870   1141 cgccctgact gaataccgtg aggtgcggcc tgacttgacg gtggccctgt acgttccgac
111871   1201 cttttatgac gctcggcgct gacacgatca ggaagtctta gccgatctga aagcgcacct
111872   1261 ttctccctg gcgcgcccgg ttccgcagcg tgaagcagtc tggctcgact cgacggccca
111873   1321 gggagccccg gtcagcgaat atgcgccggg aacaccggtt catgccgacg tgcagcgccct
111874   1381 gaccgctgat atagctgccg ctatcggcgt agcgtatccg ggggaaaacg catgaccgag
111875   1441 cgctcgtccag agcgtcggcg cgacctgctt ggtctgttgg gtgaaacgcc agttgatctc
111876   1501 agtcaagcca acgatatccg tgccctgccc gttaatgaac tgaaagtagg aagcaccag
111877   1561 ccccgccgct cttttgatct ggagcggcta agtgagctgg cagaaagtat ccgggcgcat
111878   1621 ggggtattgc agccgctgct ggtgcgttcg gtggatgggc agtacgagat tgttgaggg
111879   1681 gagcgtcgtt ggcgagccgc ccagctggct ggcttggctg aagttccggg cgtcgttcgg
111880   1741 cagctcagca acgagcaggc cagagctgct gccctcattg aaaacttgca gcgcgacaac
111881   1801 ctgaatgtca ttgatgaggt ggacgggaaa cttgagttga tagcgttgac gttggggttg
111882   1861 gaaagggaag aagcgcgcaa gcgcctgatg caactgctgc gagccgtccc aggcgacgag
```

# Фрагмент файла .gbff (конец ДНК, начало новой записи)

```
118705 411241 ttctaagacg ctcacagccc tctggacggc ctccgagcca tatgcctggg ccagagtgag
118706 411301 ggctcgcccc ccgctgaccc cgtagcgggc caggacggcg atggcctccg ggtcgactgc
118707 411361 cgcgcttttg gtgctggcaa aggtgtagtg caccatgcgc gttttgcctc gccccggta
118708 411421 ttccacctca cgcagaaaac ctatggcgat caattcctcg tgcgctgggt tgagcgcccc
118709 411481 catgactgta tcgagacgca tctcgtgcat cccagatggg gtcgcccagg tgttgagcgg
118710 411541 cacagcgaag gccagcaccg gctcctggcc gacctgatcc cgctgaaaag acagcgtgcg
118711 411601 gaacagattg cgggtcagag gctgactcag tcgcgccagc atctcaaaat cgagcggcaa
118712 411661 aatgtgtccg gcccgatat tgctgatcag gctgtcgtca agccggatag cgagaagtgt
118713 411721 ctgggcacgc cactgcccgg cattggtcac atccgccgta ttgtcttcag ccccaggcg
118714 411781 gacaatcaaa ctgaagctta cgctccgcca ccggtggcgg ctgccgtcat accagctgtc
118715 411841 ggtcacgtcg tatgcggtgt gttgcaatcg ccgcaacgaa gcttgcatat tgcgataagc
118716 411901 acgggcaactg ggcgtaatcc cgctgagctg caaaagtctg cgggtcgtca gacgaacggt
118717 411961 gtcatcctgc ggcattccct gcaacaccac agcgttgacg agtccaagga gaacatcggt
118718 412021 gtccacccca tgaggaacca acgcgttacg aggcgcagtg cagcgcacgc tgacgcctcc
118719 412081 aatggcatcg aaggcaatgt gccgggcccc ttcgtagagc tcggtgcggt tgggagcgag
118720 412141 caccaggttc agccgtgcca gatttggctc gtcgtgcctg agtgccagc
```

118721 //

```
118722 LOCUS      NZ_CP038666          45503 bp      DNA      circular CON 15-JAN-2022
118723 DEFINITION  Deinococcus radiodurans ATCC 13939 plasmid pCP1, complete sequence.
118724 ACCESSION   NZ_CP038666
118725 VERSION     NZ_CP038666.1
118726 DBLINK      BioProject: PRJNA224116
118727             BioSample: SAMN11313123
118728             Assembly: GCF_020546685.1
118729 KEYWORDS    RefSeq.
118730 SOURCE      Deinococcus radiodurans ATCC 13939
118731 ORGANISM    Deinococcus radiodurans ATCC 13939
118732             Bacteria; Deinococcus-Thermus; Deinococci; Deinococcales;
118733             Deinococcaceae; Deinococcus.
```

# Для любителей командной строки и программирования

1

Вся информация, которую вы смотрели на сайте, есть в **огромных** сводных таблицах:

[https://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY\\_REPORTS/](https://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/)

Для GenBank: `assembly_summary_genbank.txt`

Для RefSeq: `assembly_summary_refseq.txt`

2

Конкретную сборку можно найти по адресу:  
(на примере `GCA_020546685.1`)

[https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/020/546/685/GCA\\_020546685.1\\_ASM2054668v1/](https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/020/546/685/GCA_020546685.1_ASM2054668v1/)

Кстати, файлов там доступно больше, чем через web-интерфейс

3

Для скачивания конкретных файлов можно пользоваться удобной программой `wget`:

<https://www.gnu.org/software/wget/>