Как сохранить геном человека в базе данных?

Придумаем вместе

Сколько последовательностей ДНК составляют геном «человека»

Референсный геном содержит по одной последовательности каждой молекулы ДНК данного вида (для бактерий – штамма).

В референсном геноме гомологичные ДНК представлены одной последовательностью

Гомологичные ДНК – те, которые в эволюции произошли от одной и той же ДНК и различаются только мутациями.

У человека совпадение последовательностей гомологичных хромосом порядка 99.9% - 99%

У человека 22 аутосомы, X-chr, Y-chr, mito-DNA =
25 последовательностей в референсном геноме.

В геноме одной клетки одного человека ……
молекул ДНК

Сколько последовательностей ДНК составляют геном «человека»

Референсный геном содержит по одной последовательности каждой молекулы ДНК данного вида (для бактерий – штамма).

В референсном геноме гомологичные ДНК представлены одной последовательностью

Гомологичные ДНК – те, которые в эволюции произошли от одной и той же ДНК и различаются только мутациями.

У человека совпадение последовательностей гомологичных хромосом порядка 99.9% - 99%

У человека 22 аутосомы, X-chr, Y-chr, mito-DNA =
25 последовательностей в референсном геноме.

В геноме одной клетки одного человека 46 chr + много mito-DNA молекул ДНК

Геном клетки – совокупность ДНК клетки

Последовательность ДНК содержит информацию, которую умеют извлекать клеточные механизмы.
Поэтому говорят:
Геном — это совокупность всей наследственной информации

ДНК всех клеток организма имеет практически совпадающие геномы. Отличия в геномах разных клеток редки и называются соматическими мутациями.
Исключения: клетки иммунитета и раковые клетки

Геном организма – «референсный» геном его клеток

У некоторых вирусов носителем генома являются молекулы РНК.

Обратимся к более простому геному

Геном бактерии Buchnera aphidicola str. JF99 состоит из одной молекулы ДНК.
Как хранить геном в банке данных?

Банк данных Refseq

- Банк данных состоит из записей
- Одна запись банка содержит информацию об одной референсной последовательности
- Каждая запись имеет уникальный идентификатор
- Форматы информации в записи унифицированы

Сохранить ПОСЛЕДОВАТЕЛЬНОСТЬ ГЕНОМА `Buchnera aphidicola str.JF99` из Refseq в файле

- Нужен уникальный идентификатор ……
  ……
  ……
- Какую минимальную информацию добавить к последовательности для скачивания её в одном файле?
  ……
  ……
- Придумать формат файла удобный и для программ, и для людей.
  ……
  ……
  ……
  ……
  ……
  ……

Сохранить ПОСЛЕДОВАТЕЛЬНОСТЬ ГЕНОМА *Buchnera aphidicola str.JF99* из Refseq в файле

- Нужен уникальный идентификатор NC_017253.1
  NC_017253 называется Accession Code (AC),
  1 – номер версии
- Какую минимальную информацию добавить к последовательности для скачивания её в одном файле?

  ……

  ……
- Придумать формат файла удобный и для программ, и для людей.

  ……

  ……

  ……

  ……

  ……

  ……

Сохранить последовательность genoma `Buchnera aphidicola str.JF99` из записи в базе данных Refseq в файле

- Нужен уникальный идентификатор NC_017253.1
  NC_017253 называется Accession Code (AC),
  1 – номер версии
- Какую минимальную информацию добавить к последовательности для скачивания её в одном файле?
  Buchnera aphidicola str. JF99 (Acyrthosiphon pisum), complete sequence
- Придумать формат файла удобный и для программ, и для людей.

……
……
……
……
……
……

Сохранить последовательность genoma Buchnera aphidicola str.JF99 из записи в базе данных Refseq в файле

- Нужен уникальный идентификатор NC_017253.1
  NC_017253 называется Accession Code (AC),
  1 – номер версии
- Какую минимальную информацию добавить к последовательности для скачивания её в одном файле?
  Buchnera aphidicola str. JF99 (Acyrthosiphon pisum), complete sequence
- Придумать формат файла удобный и для программ, и для людей. Формат fasta

```
>NC_017253.1 Buchnera aphidicola str. JF99 (Acyrthosiphon pisum), complete sequence
ACTACTTATCCACAGATTTGTTCTTTACTAATAATAATAGTAATTATTATTTTTTATTTTTTTATTTTTT
TGAATTTAAATCTTAAAGAAAAGAAAAAGATCTTTTTTTTAAGATATTATGTTTTTAAGATTAACATGTG
TTATCTTGAATAAAATATTAATACTATTTAAATATTTTTAAATTTTTAAAAGGTTTTTATATGTTTAATT
```

……

……

……

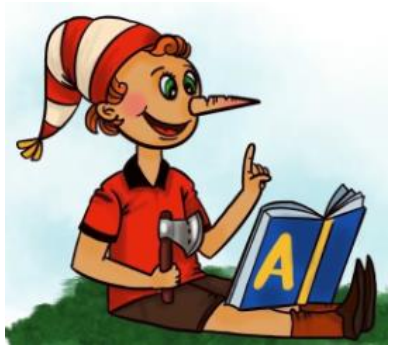
В Refseq хранится последовательность одной цепочки ДНК

- Какой из 2х комплементарных?
  - Решает автор записи (секвенировавший геном)
- ДНК бактерии кольцевая. В каком месте сделать разрыв?
  - Решает автор записи, есть предпочтения **(Mackiewicz et al., Where does bacterial replication start? Rules for predicting the oriC region. Nucleic Acids Res. 2004)**
- В каком направлении записывается посл-ть ДНК?

ОТ 5’ КОНЦА К 3’ КОНЦУ

слева 5’ =======> 3’ направо

Заруби себе на носу!



В Refseq хранится последовательность одной цепочки ДНК

- Как написать последовательность комплементарной цепочки?

  Дано: ACTACTTATCCACAGA
  Комплементарная цепочка

  ……

REVERSE COMPLEMENT!

В Refseq хранится последовательность одной цепочки ДНК

- Как написать последовательность комплементарной цепочки?

  Дано: ACTACTTATCCACAGA
  Комплементарная цепочка
  TCTGTGGATAAGTAGT

REVERSE COMPLEMENT!

Последовательность генома однозначно определяет химическую формулу молекулы ds DNA – носителя генома

При этом в термине ГЕНОМ не учитываются модификации ДНК, необычные конформации ДНК и ошибки секвенирования

Аннотация генома features

Запись последовательности генома из Refseq включает дополнительную информацию

Вот что содержится в АННОТАЦИИ геноме `Buchnera aphidicola str.JF99` и других геномов.
Эту информацию можно посмотреть и скачать из Refseq

Аннотация включает:

- Таксономию бактерии
- Ссылки на литературу
- Информацию о генах
  - Координаты гена в последовательности
  - На какой он цепочке, прямой или комплементарной
  - Название продукта гена (белка)
  - Идентификатор белка в БД белковых последовательностей
  - Последовательность белка

Фрагмент аннотации генома Buchnera aphidicola str.JF99 в базе данных Refseq

```
LOCUS       NC_017253              641716 bp    DNA     circular CON 01-NOV-2022
DEFINITION  Buchnera aphidicola str. JF99 (Acyrthosiphon pisum), complete
            sequence.
ACCESSION   NC_017253
VERSION     NC_017253.1
SOURCE      Buchnera aphidicola str. JF99 (Acyrthosiphon pisum)
  ORGANISM  Buchnera aphidicola str. JF99 (Acyrthosiphon pisum)
            Bacteria; Pseudomonadota; Gammaproteobacteria; Enterobacterales;
            Erwiniaceae; Buchnera.
REFERENCE   1  (bases 1 to 641716)
  AUTHORS   MacDonald,S.J., Thomas,G.H. and Douglas,A.E.
  TITLE     Genetic and metabolic determinants of nutritional phenotype in an

     gene            201..2087
                     /gene="mnmG"
                     /locus_tag="CWS_RS00005"
                     /old_locus_tag="CWS_00005"
     CDS             201..2087
                     /product="tRNA uridine-5-carboxymethylaminomethyl(34)
                     synthesis enzyme MnmG"
                     /protein_id="WP_012619398.1"
                     /translation="MFNLRNFDVIVVGAGHAGTEAAMASSRMGCKTLLLTQKISDLGA
```

Ген на прямой цепи

```
gene            3501..3986
                /locus_tag="CWS_RS00020"
                /old_locus_tag="CWS_00020"
CDS             3501..3986
                /locus_tag="CWS_RS00020"
                /old_locus_tag="CWS_00020"
                /EC_number="7.1.2.2"
                /inference="COORDINATES: similar to AA
                sequence:RefSeq:WP_014498977.1"
                /GO_function="GO:0015078 - proton transmembrane
                transporter activity [Evidence IEA]"
                /GO_process="GO:0015986 - proton motive force-driven ATP
                synthesis [Evidence IEA]"
                /note="Derived by automated computational analysis using
                gene prediction method: Protein Homology."
                /codon_start=1
                /transl_table=11
                /product="F0F1 ATP synthase subunit B"
                /protein_id="WP_009873966.1"
                /translation="MNLNATILGQAISFVLFVWFCMKYIWPPIILAIETRQKEIKESL
                TNAKKAQDELYILEKKIHQNIIDAKQKASNILNSANKQKVSILEDARNQALEESKKII
                LNTQSEINIAITHARKNLHKEVVDLSISMAEKIIKKNISKDDNQELLDELVTSLSQVK
                N"
```

Ген на обратной цепи

```
gene            complement(229363..230160)
                /gene="speD"
                /locus_tag="CWS_RS01055"
                /old_locus_tag="CWS_01095"
CDS             complement(229363..230160)
                /gene="speD"
                /locus_tag="CWS_RS01055"
                /old_locus_tag="CWS_01095"
                /EC_number="4.1.1.50"
                /inference="COORDINATES: similar to AA
                sequence:RefSeq:WP_012619379.1"
                /GO_function="GO:0004014 - adenosylmethionine
                decarboxylase activity [Evidence IEA]"
                /GO_process="GO:0008295 - spermidine biosynthetic process
                [Evidence IEA]"
                /note="Derived by automated computational analysis using
                gene prediction method: Protein Homology."
                /codon_start=1
                /transl_table=11
                /product="adenosylmethionine decarboxylase"
                /protein_id="WP_079172804.1"
                /translation="MIKLQKLKLYGFNNLTKSLSFCIYDICYANTNDSRNSYISYIDE
                QYNAIRLTKILKKTCSIIGANVLNIFHQDYEPQGASVTILVCEEPMSMEKINALNKNI
```

Ген РНК (rRNA)

```
gene            275191..276738
                /locus_tag="CWS_RS01235"
                /old_locus_tag="CWS_r03198"
rRNA            275191..276738
                /locus_tag="CWS_RS01235"
                /old_locus_tag="CWS_r03198"
                /product="16S ribosomal RNA"
                /inference="COORDINATES: nucleotide
                motif:Rfam:14.4:RF00177"
                /inference="COORDINATES: profile:INFERNAL:1.1.1"
                /note="Derived by automated computational analysis using
                gene prediction method: cmsearch."
                /db_xref="RFAM:RF00177"
```

Аннотации генов можно сохранить в формате «Хромосомной таблицы»

Таблица содержит только главную информацию о генах.

Второстепенная информация не включена в таблицу.

НАПРИМЕР такая:

```
/note="Derived by automated computational analysis using
gene prediction method: tRNAscan-SE."
```

# feature	class	genomic_ accession	start	end	strand	name	symbol	product_ accession
gene	protein_coding	NC_017253.1	8915	11326	-		gyrB	
CDS	with_protein	NC_017253.1	8915	11326	-	DNA topoisomerase (ATP-hydrolyzing) subunit B	gyrB	WP_009873972.1
gene	protein_coding	NC_017253.1	11453	12553	-		dnaN	
CDS	with_protein	NC_017253.1	11453	12553	-	DNA polymerase III subunit beta	dnaN	WP_012619401.1
gene	protein_coding	NC_017253.1	12558	13922	-		dnaA	
CDS	with_protein	NC_017253.1	12558	13922	-	chromosomal replication initiator protein DnaA	dnaA	WP_009873974.1
gene	protein_coding	NC_017253.1	14373	14516	+		rpmH	
CDS	with_protein	NC_017253.1	14373	14516	+	50S ribosomal protein L34	rpmH	WP_009873975.1

Запомните

- Референсный геном содержит по одной последовательности каждой молекулы ДНК в данном виде (для бактерий – штамме)
- Каждая последовательность Refseq имеет уникальный идентификатор (AC – код доступа)
- Формат fasta: строка с 1м символом “>” содержит идентификатор и через пробел информацию о последовательности; следующие строки – последовательность
- Последовательность пишется от 5’ к 3’ концу
- Цепочка ДНК выбирается произвольно
- Разрыв кольцевой ДНК для записи последовательности может быть выбран произвольно, есть предпочтения – они не всегда соблюдаются
- Последовательность комплементарной цепочки пишется как REVEVERSE COMPLEMENT
- Последовательность генома однозначно определяет химическую молекулы ДНК (с оговорками)
- Геном — это совокупность всей наследственной информации: носитель *in vivo* совокупность ДНК клетки, носитель *in silico последовательности всех ДНК вида*
- Аннотация последовательности включает таксономию бактерии, Ссылки на литературу, Информацию о генах
- Минимальная информация о гене координаты в последовательности (от..до), на какой цепочке – прямой или комплементарной – расположен, тип гена – белок или РНК, название гена, последовательность белка (для генов белков)

БУДЕТ КОНТРОЛЬНАЯ ПО ВСЕМУ МАТЕРИАЛУ О ГЕНОМЕ!!!!

А кто слушал – молодец!

КОНЕЦ