МФК Биоинформатика

февраль – май 2022

Полезная информация

Сайт МФК: «Биоинформатика 2022»
https://kodomo.fbb.msu.ru/wiki/Main/mf_2022s

С орг-вопросами обращайтесь к
Сергею Александровичу Спирин (организатору МФК и лектору)
sas@belozersky.msu.ru

Кроме лекций для того, чтобы разобраться в теме предлагается выполнять домашние задания. Бывают задания на поиск в интернет, бывают школьного уровня, бывают посложнее

Зачёт автомат можно получить при зачёте оговоренного числа домашних заданий, выдаваемых после каждой лекции

На сайте найдёте

- Список записавшихся на МФК
- Ссылки на презентации
- Домашние задания (после каждой лекции)
- Форма для вопросов преподавателям

Попозже

- Правила получения зачета
- Ведомость с результатами проверок ДЗ
- Новости

ОФИЦИАЛЬНЫЙ список записавшихся в ректорате, - 70 студентов, - уже включен на сайт (проверьте есть ли вы) . Он дает право на получение зачёта в зачётку и в зачётную ведомость МГУ.

Желающие могут попросить включить их в список участников МФК Биоинформатика.

Это не меняет их официальный статус. Дополнительно записанные студенты – вольнослушатели - не получают права на проставление зачёта по МФК.
Узнавайте что можно сделать в своей учебной части.

Преподаватели не обязаны проверять домашние задания вольнослушателей. Но если захочет – может

Обзор тем.
Порядок тем на установлен, дополнения и изменения возможны

- ДНК, РНК, их последовательности.
- Геном – что такое? Как кодируется и для кого?
  - На примере простейших геномов вирусов
- Геном человека: история и практика
- Как расшифровывают последовательности ДНК, РНК, белков?
- Выравнивание последовательностей. BLAST
- Как реконструировать эволюция по последовательностям
- Что можно узнать реконструируя эволюцию по последовательностям современных и сохранившихся архаическим молекул ДНК.
- Разнообразие мира РНК
- Пространственные структуры биологических молекул - белков, ДНК и др.
- Современные способы разработки лекарств против известной мишени – белка.
- Интересные примеры применения науки с использованием биоинформатики. Пример генной терапии. Ребенок трех родителей. Метилирование и возраст, можно ли вернуть клетке молодость? И др.

Л1. ГЕНОМ

Андрей Владимирович Алексеевский

aba@belozersky.msu.ru

1.Особенные молекулы в составе живого на Земле

- Живые организмы (клеточные) и вирусы состоят из молекул Вирусы не клетки))) живые ли…
- У живого (на планете ЗЕМЛЕ! Кто знает про другие миры?) есть особенные молекулы: **ДНК, РНК, белки,** липиды
  Есть и другие молекулы, без которых не проживёшь)))

Наверное, можно так сказать: *если физический объект не включает в свой состав ни ДНК, ни РНК, то он не является живым в земном понимании жизни.*

Биоинформатика – информатика применяемая к массовым молекулярно-биологическим данным. Современные биотехнологии дают данные. Для того чтобы получить из них новые биологические и медицинские знания необходима биоинформатика.

Биоинформатика. Объекты

- Биологические последовательности
  - ДНК
    - Например, есть специальный метод секвенирования определенных коротких последовательностей, который позволил получить важную инфорамцию о **пространственной организации молекул ДНК в хромосомах человека**
  - РНК
  - белки
- Пространственные структуры биологических макромолекул
- Любые массовые биологические данные.

2. Геном это информация*, которая закодирована в молекулах ДНК, содержащихся в клетке организма

- **Физический носитель генома** – совокупность молекул ДНК, содержащихся в живой клетке
- **У вирусов**, хотя вирусы – не клетки, **тоже есть геном**. Геномы вирусов, кроме молекул **ДНК**, бывают закодированы молекулами **РНК**.

Соответственно, бывают ДНК вирусы(пример -аденовирусы) и РНК вирусы (коронавирусы, вирус гриппа)

*) эта информация передаётся по наследству

Вопросы

- Что такое информация? Нашел такое определение:

  ИНФОРМАЦИЯ — сведения независимо от формы их представления))) [1)]

Теория информации основанная Шенноном – математическая теория передачи данных[2)] – используется в биоинформатике, по слишком формализована и проста для объяснения живого:)

- Какая информация закодирована в геноме?
- Сведения: от кого? кому (получатель)?
- Как кодируется информация в геноме?
- Как прочитать её людям и зачем?

[1)] Wiki со ссылкой на Когаловского Р.М. специалиста по информационным систем.

[2)] C.Shannon, “The Mathematical Theory of Communications” , 1948

Что такое информация?

Армянское радио

- «Правда ли, что Иштоян выиграл в лотерею машину?»

  Иштоян известный футболист Арарата в 70-х

- «**Правда**. Но не Иштоян, а Петросян, не машину, а швейную машинку; не в лотерею, а в карты; и не выиграл, а проиграл»

  Петросян чемпион мира по шахматам в 60-х

«Сколько информации в этом сообщении?»
И.М.Гельфанд

Молекулы ДНК или РНК

3. КОДИРОВАНИЕ ИНФОРМАЦИИ

Как кодируется информация в молекуле ДНК

Уровень 1: ГЕНОМИКА
Информация кодируется – **основной химической формулой молекулы ДНК** (или РНК).

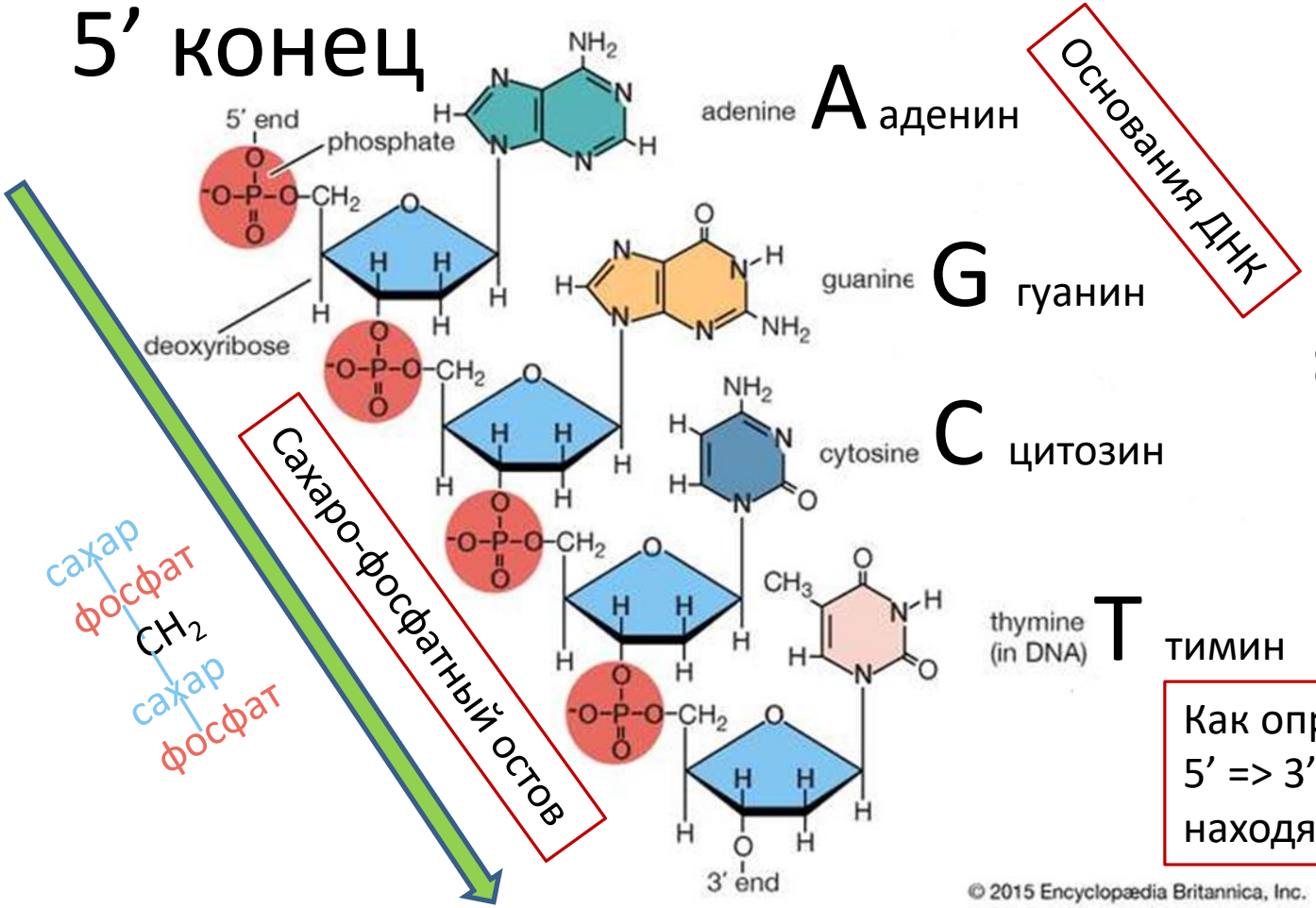
Уровень 2: ЭПИГЕНОМИКА

Кодирование информации сохраняющимися при делении клетки химическими модификациями молекулы ДНК (или РНК) или белков, стабильно связанных с ДНК

Как заметки на полях книги. Содержание не меняют, но смысл для читателя меняют. Заметкам Ленина на полях прочтенных книг посвящен 29 том собрания сочинений:))))

Формула молекулы ДНК определяется последовательностью букв **A, T, G, C**

5’ конец



Основания ДНК

A аденин

G гуанин

C цитозин

T тимин

= **AGCT**

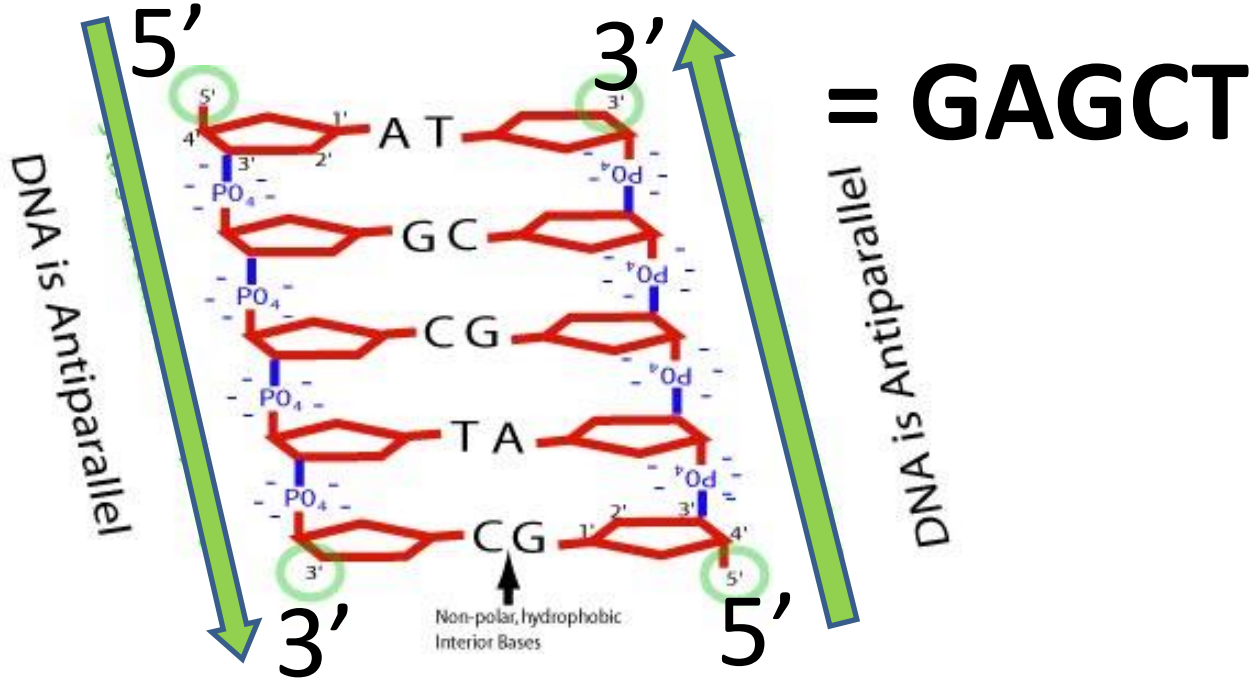
Сахаро-фосфатный остов

сахар
фосфат
CH_2
сахар
фосфат

Как определить направление
5’ => 3’
находясь в середине цепочки?

3’ конец

ДНК обычно состоит из двух антипараллельных комплементарных цепочек



= GAGCT

Можно ли изучая кусочек ДНК (см. предыдущий слайд) определить какая цепочка первая (основная) а какая вторая (дополнительная)? НЕТ в силу полной симметрии остова ДНК.

И направление 5’ => 3’ каждой цепочки? ДА

Это нужно молекулярным машинам и белкам, которые «читают» геном.

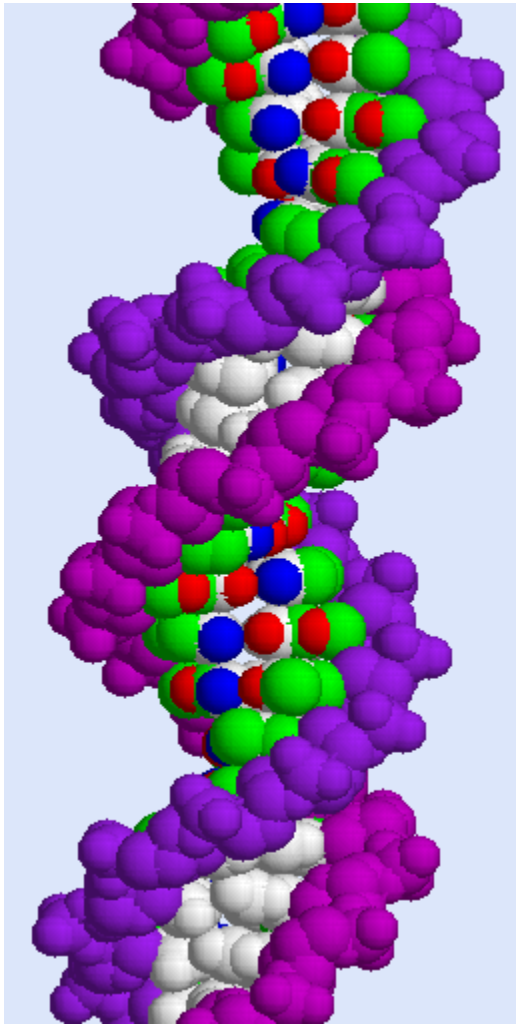
И биоинформатикам:
в банках данных лежат последовательности полных геномов (например, NCBI Genomes).
Какой из цепочек? В каком направлении?

Задание 1. Известна последовательность одной цепочки ДНК:

A T G A C C A A A

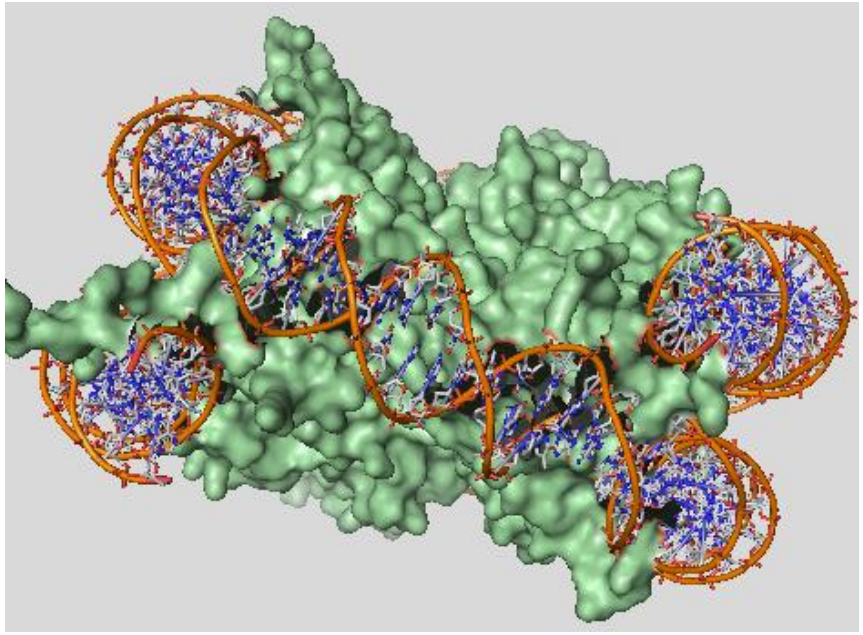
Напишите последовательность второй цепочки.

4. Лучше один раз увидеть .....



Двойная спираль ДНК.

Раскраска моя ААл ☺



Нуклеосома:
ДНК человека на “катушке” из гистонов: вид сбоку (гистоны – такие белки)

Обе структуры расшифрованы с помощью рентгеноструктурного анализа.

Считывать информацию – последовательност ДНК белки и молекулярные машины могут двумя способами

а) **расплетая цепочки** и различая основания ДНК по их свойствам

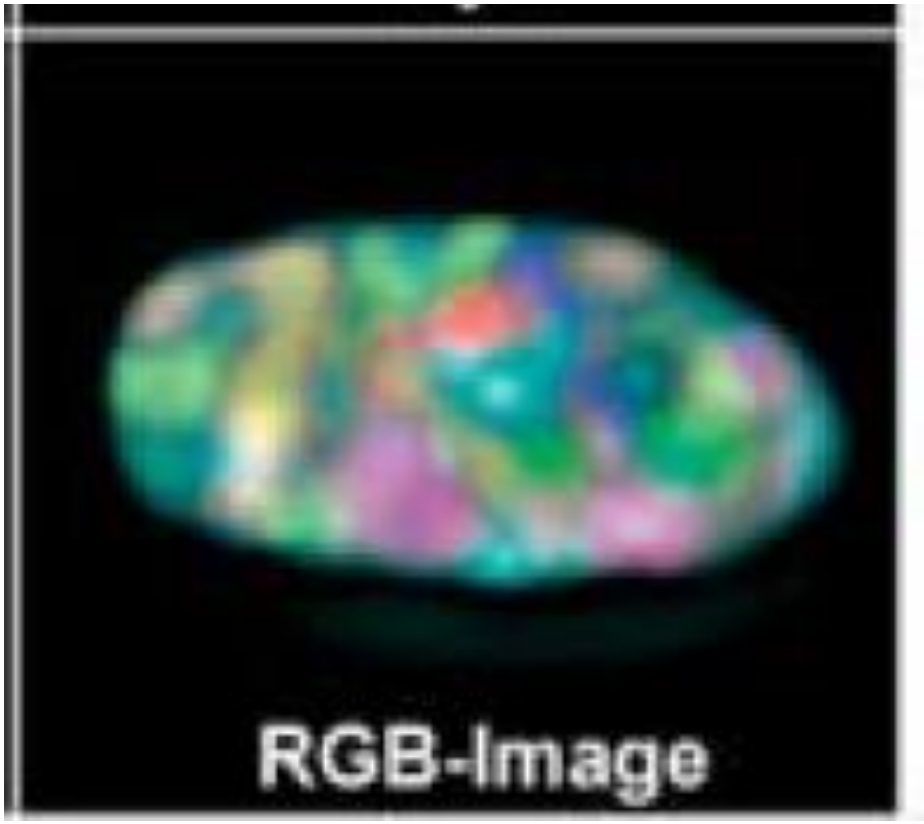
б) **без расплетения** - анализируя атомные группы в большой бороздке ДНК (раскрашены на пред. слайде) и малой бороздке – не раскрашены; также локальная кривизна ДНК немного зависит от последовательности …

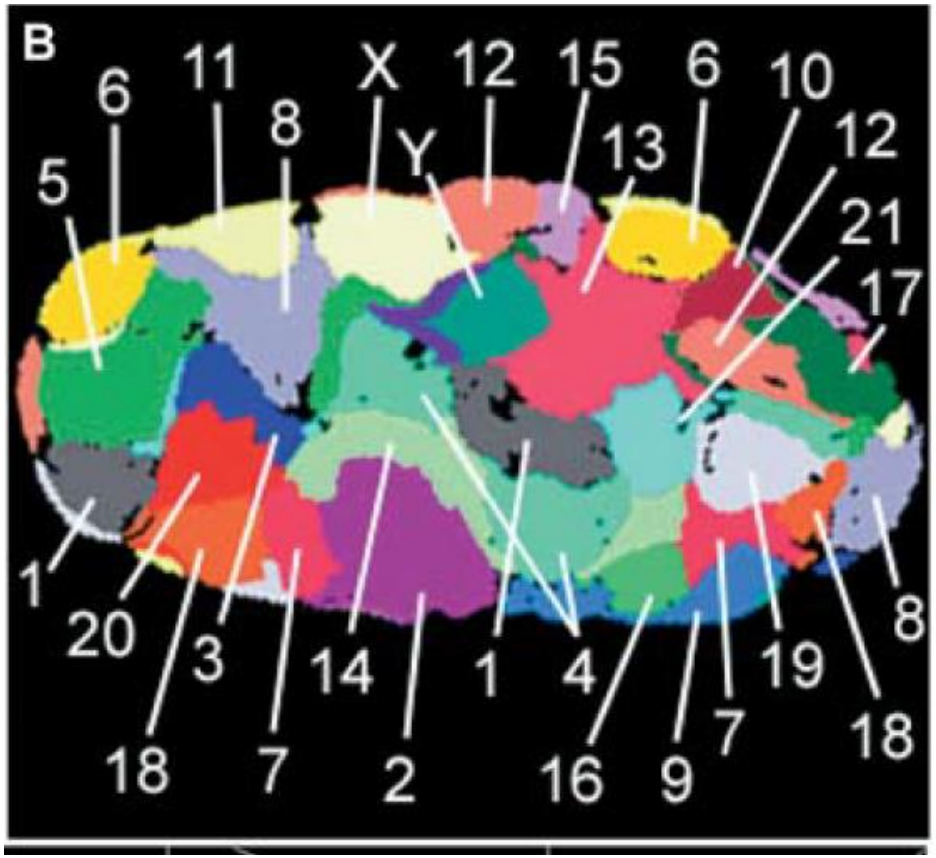
Цепочки ДНК с комплементарными последовательностями образуют 2х цепочечные ДНК (дцДНК) в силу законов физики *(гибридизация)*. Возможна гибридизация и ДНК – РНК

Пары комплементарных оснований A – T и G – C в дцДНК связаны водородными связями (двумя и тремя соответственно). Каждая водородная связь – уменьшение энергии, физическая система стремится к минимуму энергии.

Это свойство лежит в основе устойчивости молекул ДНК (задание 7: для ДНК какой древности удалось прочитать её последовательность) и получения микроскопического изображения на след. слайде

ДНК в ядре клетки фибробласта человека. Разные ДНК покрашены в разные цвета, одинаковые по последовательности (>99%) – в одинаковые цвета





Клетка находится перед стадией деления: каждая хромосома состоит из 2х одинаковых ДНК после удвоения (репликации). При делении клетки он разойдутся в разные дочерние клетки

Bolzer A et al. Three-dimensional maps of all chromosomes in fibroblast nuclei and prometaphase rosettes. PLoS Biol. 2005

Как получена микрофотография

К каждой ДНК подобраны многочисленные пробы – ДНК полностью комплементарные участку данной ДНК.

Probe size can range from few kb to megabases (Mb), depending on the application.

К концам проб присоединены флюорофоры одного из 7 цветов. Пропорции проб к одной ДНК с флюорофорами разных цветов подобраны специально, чтобы различать разные ДНК.

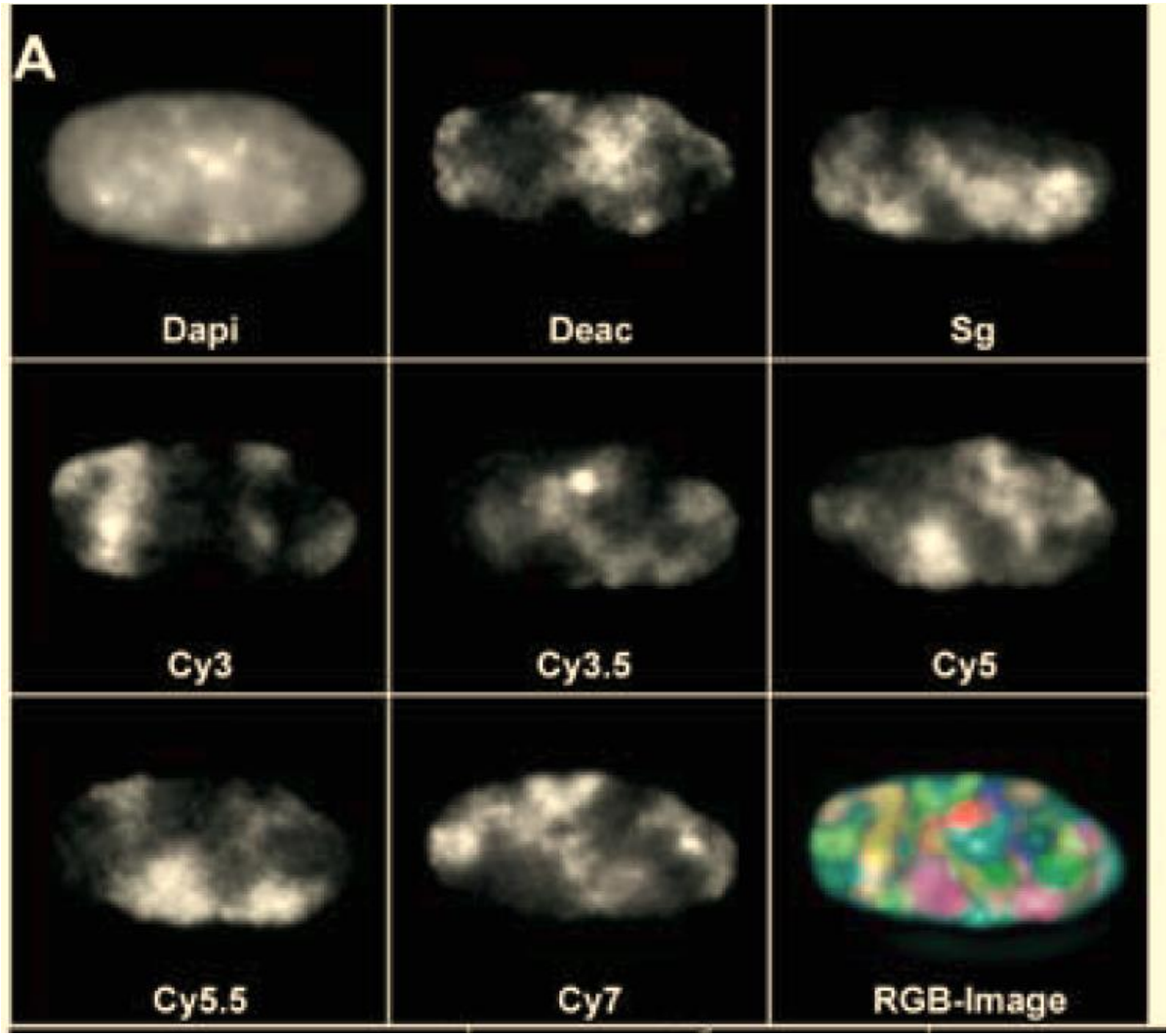
(A) Деконволюция (объединение) флюоресцентных микрофотографий в 7 каналах

(one channel for DAPI (DNA counterstain and seven channels for the following fluorochromes: diethylaminocoumarin (Deac), Spectrum Green (SG), and the cyanine dyes Cy3, Cy3.5, Cy5,Cy5.5, and Cy7))

Окраска изображения разных хромосом (1–22, X, and Y) в 24 цвета получена наложением 7 каналов

(B) Прорисовка с искусственно подобранными цветами изображения (A)

Исходные микрофотографии в восьми каналах и их совмещение

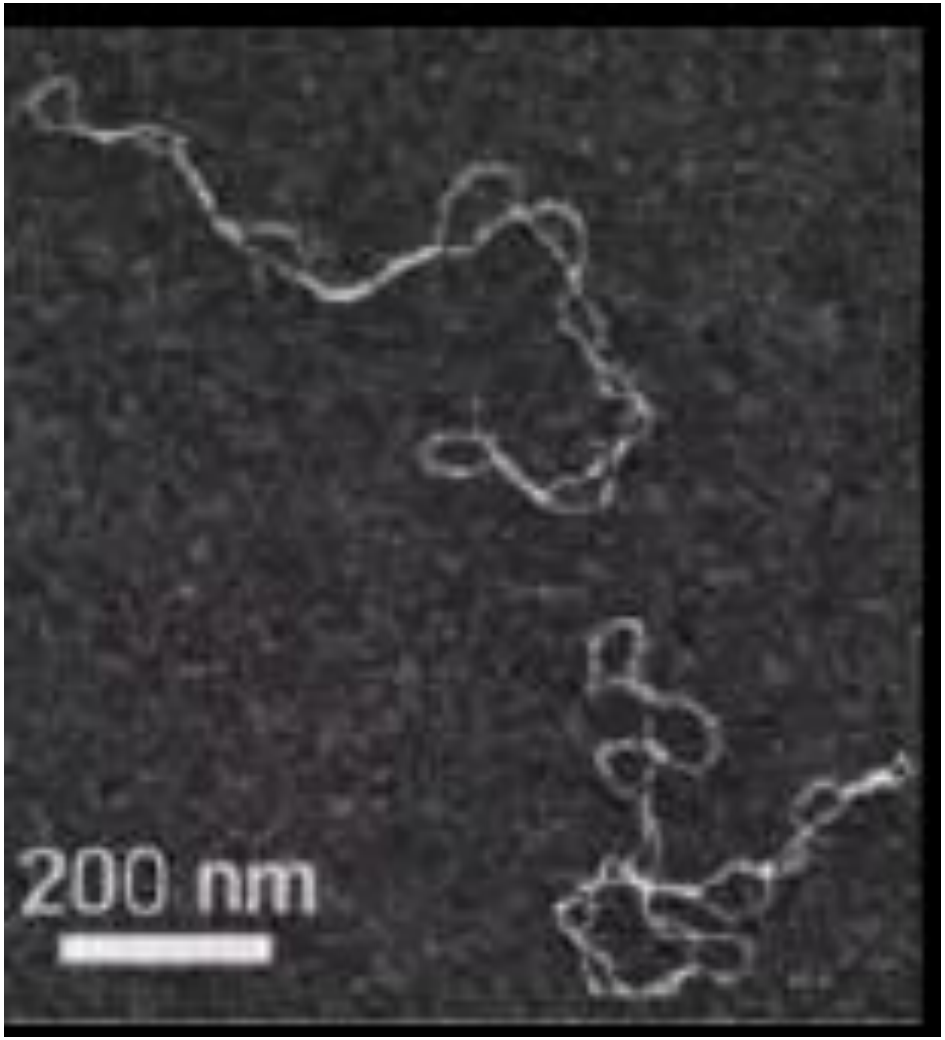


Микрофотография маленькой кольцевой ДНК бактерии - плазмиды

Маленькая плазмида бактерии

Электронная микроскопия

Маленькая кольцевая ДНК бактерии – плазмида[1)] (плазмида – не весь геном бактерии!)



Просвечивающий электронный микроскоп

[1)] про плазмиды
https://www.quora.com/What-is-the-importance-of-plasmids-in-biotechnology

Пример: Геном адено-ассоциированного вируса AAV2

- В вирусных частицах он хранится в виде одноцепочечной ДНК (оцДНК)
- Адено-ассоциированные вирусы человека используются для генной терапии (м.б. расскажу сильно позже в моей второй лекции).
- На следующих двух слайдах – полный геном!

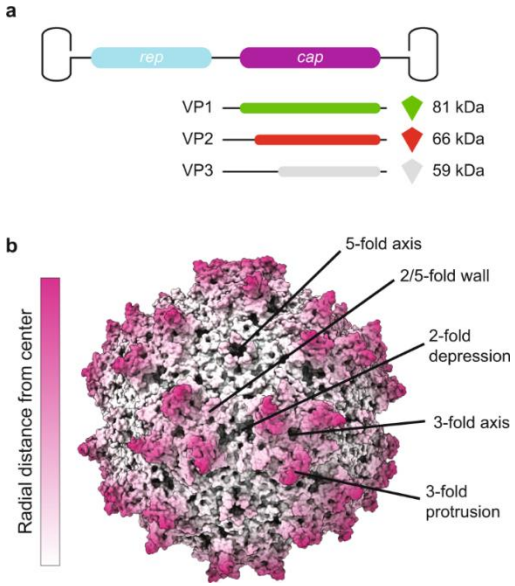


Fig. 1: Adeno-associated virus (AAV) capsid structure.

T.P. Wörner et al., Adeno-associated virus capsid assembly is divergent and stochastic, Nature Communications 12:1642 (2021)

>NC_001401.2 Adeno-associated virus - 2, complete genome 4679 bp
TTGGCCACTCCCTCTCTGCGCGCTCGCTCGCTCACTGAGGCCGGGCGACCAAAGGTCGCCCGACGCCCGG
GCTTTGCCCGGGCGGCCTCAGTGAGCGAGCGAGCGCGCAGAGAGGGAGTGGCCAACTCCATCACTAGGGG
TTCCTGGAGGGGTGGAGTCGTGACGTGAATTACGTCATAGGGTTAGGGAGGTCCTGTATTAGAGGTCACG
TGAGTGTTTTGCGACATTTTGCGACACCATGTGGTCACGCTGGGTATTTAAGCCCGAGTGAGCACGCAGG
GTCTCCATTTTGAAGCGGGAGGTTTGAACGCGCAGCCGCCATGCCGGGGTTTTACGAGATTGTGATTAAG
GTCCCCAGCGACCTTGACGAGCATCTGCCCGGCATTTCTGACAGCTTTGTGAACTGGGTGGCCGAGAAGG
AATGGGAGTTGCCGCCAGATTCTGACATGGATCTGAATCTGATTGAGCAGGCACCCCTGACCGTGGCCGA
GAAGCTGCAGCGCGACTTTCTGACGGAATGGCGCCGTGTGAGTAAGGCCCCGGAGGCCCTTTTCTTTGTG
CAATTTGAGAAGGGAGAGAGCTACTTCCACATGCACGTGCTCGTGGAAACCACCGGGGTGAAATCCATGG
TTTTGGGACGTTTCCTGAGTCAGATTCGCGAAAAACTGATTCAGAGAATTTACCGCGGGATCGAGCCGAC
TTTGCCAAACTGGTTCGCGGTCACAAAGACCAGAAATGGCGCCGGAGGCGGGAACAAGGTGGTGGATGAG
TGCTACATCCCCAATTACTTGCTCCCCAAAACCCAGCCTGAGCTCCAGTGGGCGTGGACTAATATGGAAC
AGTATTTAAGCGCCTGTTTGAATCTCACGGAGCGTAAACGGTTGGTGGCGCAGCATCTGACGCACGTGTC
GCAGACGCAGGAGCAGAACAAAGAGAATCAGAATCCCAATTCTGATGCGCCGGTGATCAGATCAAAAACT
TCAGCCAGGTACATGGAGCTGGTCGGGTGGCTCGTGGACAAGGGGATTACCTCGGAGAAGCAGTGGATCC
AGGAGGACCAGGCCTCATACATCTCCTTCAATGCGGCCTCCAACTCGCGGTCCCAAATCAAGGCTGCCTT
GGACAATGCGGGAAAGATTATGAGCCTGACTAAAACCGCCCCCGACTACCTGGTGGGCCAGCAGCCCGTG
GAGGACATTTCCAGCAATCGGATTTATAAAATTTTGGAACTAAACGGGTACGATCCCCAATATGCGGCTT
CCGTCTTTCTGGGATGGGCCACGAAAAAGTTCGGCAAGAGGAACACCATCTGGCTGTTTGGGCCTGCAAC
TACCGGGAAGACCAACATCGCGGAGGCCATAGCCCACACTGTGCCCTTCTACGGGTGCGTAAACTGGACC
AATGAGAACTTTCCCTTCAACGACTGTGTCGACAAGATGGTGATCTGGTGGGAGGAGGGGAAGATGACCG
CCAAGGTCGTGGAGTCGGCCAAAGCCATTCTCGGAGGAAGCAAGGTGCGCGTGGACCAGAAATGCAAGTC
CTCGGCCCAGATAGACCCGACTCCCGTGATCGTCACCTCCAACACCAACATGTGCGCCGTGATTGACGGG
AACTCAACGACCTTCGAACACCAGCAGCCGTTGCAAGACCGGATGTTCAAATTTGAACTCACCCGCCGTC
TGGATCATGACTTTGGGAAGGTCACCAAGCAGGAAGTCAAAGACTTTTTCCGGTGGGCAAAGGATCACGT
GGTTGAGGTGGAGCATGAATTCTACGTCAAAAAGGGTGGAGCCAAGAAAAGACCCGCCCCCAGTGACGCA
GATATAAGTGAGCCCAAACGGGTGCGCGAGTCAGTTGCGCAGCCATCGACGTCAGACGCGGAAGCTTCGA
TCAACTACGCAGACAGGTACCAAAACAAATGTTCTCGTCACGTGGGCATGAATCTGATGCTGTTTCCCTG
CAGACAATGCGAGAGAATGAATCAGAATTCAAATATCTGCTTCACTCACGGACAGAAAGACTGTTTAGAG
TGCTTTCCCGTGTCAGAATCTCAACCCGTTTCTGTCGTCAAAAAGGCGTATCAGAAACTGTGCTACATTC
ATCATATCATGGGAAAGGTGCCAGACGCTTGCACTGCCTGCGATCTGGTCAATGTGGATTTGGATGACTG
CATCTTTGAACAATAAATGATTTAAATCAGGTATGGCTGCCGATGGTTATCTTCCAGATTGGCTCGAGGA

CACTCTCTCTGAAGGAATAAGACAGTGGTGGAAGCTCAAACCTGGCCCACCACCACCAAAGCCCGCAGAG
CGGCATAAGGACGACAGCAGGGGTCTTGTGCTTCCTGGGTACAAGTACCTCGGACCCTTCAACGGACTCG
ACAAGGGAGAGCCGGTCAACGAGGCAGACGCCGCGGCCCTCGAGCACGACAAAGCCTACGACCGGCAGCT
CGACAGCGGAGACAACCCGTACCTCAAGTACAACCACGCCGACGCGGAGTTTCAGGAGCGCCTTAAAGAA
GATACGTCTTTTGGGGGCAACCTCGGACGAGCAGTCTTCCAGGCGAAAAAGAGGGTTCTTGAACCTCTGG
GCCTGGTTGAGGAACCTGTTAAGACGGCTCCGGGAAAAAAGAGGCCGGTAGAGCACTCTCCTGTGGAGCC
AGACTCCTCCTCGGGAACCGGAAAGGCGGGCCAGCAGCCTGCAAGAAAAAGATTGAATTTTGGTCAGACT
GGAGACGCAGACTCAGTACCTGACCCCCAGCCTCTCGGACAGCCACCAGCAGCCCCCTCTGGTCTGGGAA
CTAATACGATGGCTACAGGCAGTGGCGCACCAATGGCAGACAATAACGAGGGCGCCGACGGAGTGGGTAA
TTCCTCGGGAAATTGGCATTGCGATTCCACATGGATGGGCGACAGAGTCATCACCACCAGCACCCGAACC
TGGGCCCTGCCCACCTACAACAACCACCTCTACAAACAAATTTCCAGCCAATCAGGAGCCTCGAACGACA
ATCACTACTTTGGCTACAGCACCCCTTGGGGGTATTTTGACTTCAACAGATTCCACTGCCACTTTTCACC
ACGTGACTGGCAAAGACTCATCAACAACAACTGGGGATTCCGACCCAAGAGACTCAACTTCAAGCTCTTT
AACATTCAAGTCAAAGAGGTCACGCAGAATGACGGTACGACGACGATTGCCAATAACCTTACCAGCACGG
TTCAGGTGTTTACTGACTCGGAGTACCAGCTCCCGTACGTCCTCGGCTCGGCGCATCAAGGATGCCTCCC
GCCGTTCCCAGCAGACGTCTTCATGGTGCCACAGTATGGATACCTCACCCTGAACAACGGGAGTCAGGCA
GTAGGACGCTCTTCATTTTACTGCCTGGAGTACTTTCCTTCTCAGATGCTGCGTACCGGAAACAACTTTA
CCTTCAGCTACACTTTTGAGGACGTTCCTTTCCACAGCAGCTACGCTCACAGCCAGAGTCTGGACCGTCT
CATGAATCCTCTCATCGACCAGTACCTGTATTACTTGAGCAGAACAAACACTCCAAGTGGAACCACCACG
CAGTCAAGGCTTCAGTTTTCTCAGGCCGGAGCGAGTGACATTCGGGACCAGTCTAGGAACTGGCTTCCTG
GACCCTGTTACCGCCAGCAGCGAGTATCAAAGACATCTGCGGATAACAACAACAGTGAATACTCGTGGAC
TGGAGCTACCAAGTACCACCTCAATGGCAGAGACTCTCTGGTGAATCCGGGCCCGGCCATGGCAAGCCAC
AAGGACGATGAAGAAAAGTTTTTTCCTCAGAGCGGGGTTCTCATCTTTGGGAAGCAAGGCTCAGAGAAAA
CAAATGTGGACATTGAAAAGGTCATGATTACAGACGAAGAGGAAATCAGGACAACCAATCCCGTGGCTAC
GGAGCAGTATGGTTCTGTATCTACCAACCTCCAGAGAGGCAACAGACAAGCAGCTACCGCAGATGTCAAC
ACACAAGGCGTTCTTCCAGGCATGGTCTGGCAGGACAGAGATGTGTACCTTCAGGGGCCCATCTGGGCAA
AGATTCCACACACGGACGGACATTTTCACCCCTCTCCCCTCATGGGTGGATTCGGACTTAAACACCCTCC
TCCACAGATTCTCATCAAGAACACCCCGGTACCTGCGAATCCTTCGACCACCTTCAGTGCGGCAAAGTTT
GCTTCCTTCATCACACAGTACTCCACGGGACAGGTCAGCGTGGAGATCGAGTGGGAGCTGCAGAAGGAAA
ACAGCAAACGCTGGAATCCCGAAATTCAGTACACTTCCAACTACAACAAGTCTGTTAATGTGGACTTTAC
TGTGGACACTAATGGCGTGTATTCAGAGCCTCGCCCCATTGGCACCAGATACCTGACTCGTAATCTGTAA
TTGCTTGTTAATCAATAAACCGTTTAATTCGTTTCAGTTGAACTTTGGTCTCTGCGTATTTCTTTCTTAT
CTAGTTTCCATGGCTACGTAGATAAGTAGCATGGCGGGTTAATCATTAACTACAAGGAACCCCTAGTGAT
GGAGTTGGCCACTCCCTCTCTGCGCGCTCGCTCGCTCACTGAGGCCGGGCGACCAAAGGTCGCCCGACGC
CCGGGCTTTGCCCGGGCGGCCTCAGTGAGCGAGCGAGCGCGCAGAGAGGGAGTGGCCAA

Задания 5, 6, 9 – про гномы экстремальной длины

5. Геном - неизвестный текст, который хочется изучить. С чего начнем?

- В алфавите четыре буквы A, T, G, C (понятно)

- Буквы идут неупорядоченно,
  похоже на случайную последовательность?

Лингвистический анализ текста

- Одинаковы ли частоты букв?
- Часто и редко встречающиеся слова
  (т.е. короткие последовательности)
- Равномерность частоты букв и слов вдоль текста

Все эти вопросы изучаются и имеют биологически смысл!
Примеры наблюдений:

- #C ≈ #G, #T ≈ #A (# = число)
- Слов CG *мало* в определенных геномах

  Слов TA *мало* во всех геномах

(Задание 8* - нужно уметь программировать)

Мало по сравнению с ожидаемым при случайном расположении букв в тексте.
Самое простое предположение:
Ожидаемое #CG = частота(C)*частота(G)* число букв в геноме. Для др. слов аналогично

- В некоторых геномах #C > #G в одной части и

#G > #C в другой части («GC skew»)

Всерьёз думают о живых вакцинах, основанных на вирусах с увеличенным числом CG или TA!

Interestingly, most mammalian RNA viruses have low frequencies of CpGs (45,46). Furthermore, viruses with high CpG frequencies may be more recognizable by pathogen innate immune sensors (47–50).

Attenuation of the classical oral poliovirus vaccine is based on very few point mutations, which can revert to virulence after a few rounds of viral replication (144). These pioneering results obtained with recoded polioviruses suggest that codon-usage in recoded viruses may be much more stable than most RNA virus point mutants, and could possibly enable the development of live attenuated RNA virus vaccines with superior genetic stability.

Martinez et al., 2019, NAR

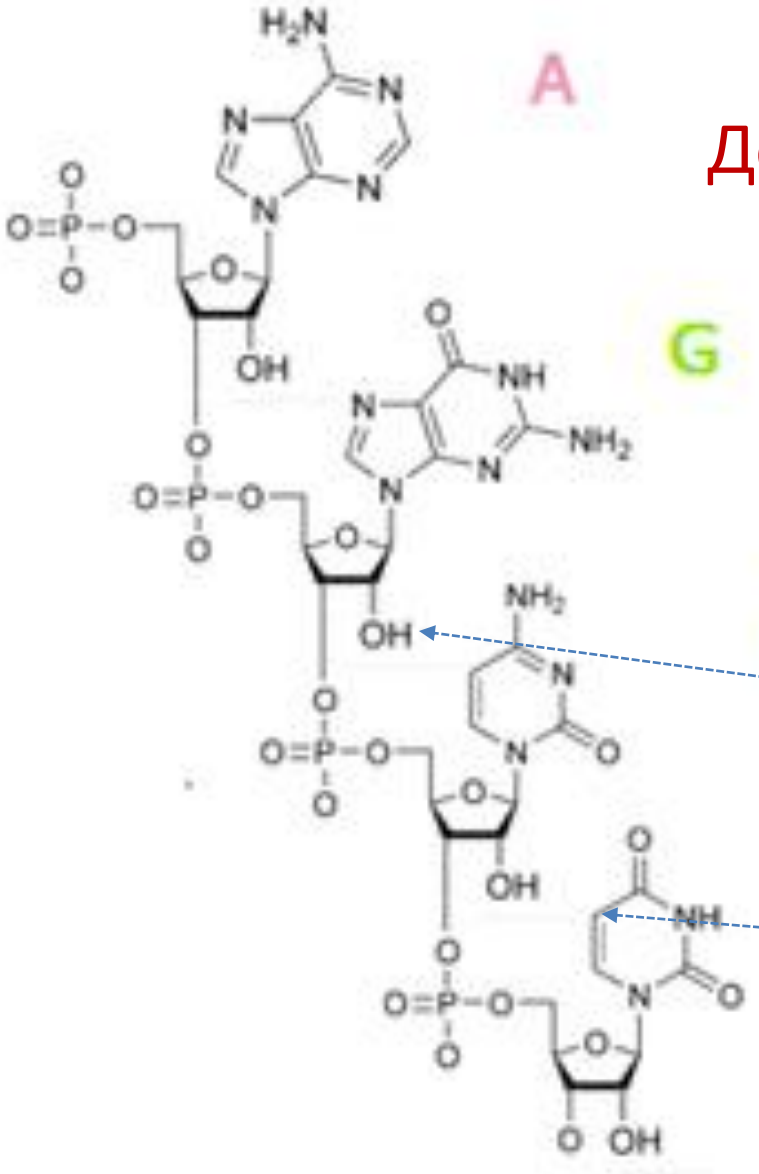
Take home messages

- Геном клетки организма закодирован во всех молекулах ДНК в ней (Задание 2. Число молекул ДНК в клетке человека)
- Вирусы не образуют клетки. Геномы вирусов – ДНК или РНК.
- Последовательность оснований однозначно определяет химическую формулу одной цепочки молекулы ДНК (не считая модификаций, они не сохраняются при репликации ДНК)
- Последовательность одной цепочки 2х цепочечной ДНК однозначно определяется последовательностью другой цепочки (не считая ошибок – неправильно спаренных оснований ДНК) (Задание 1. Обязательное)
- По фрагменту 2х цепочечной ДНК можно определить направление 5’ => 3’ каждой цепочки
- Последовательность оснований читается молекулярными машинами и записывается людьми всегда от 5’-конца к 3’-концу

На примере надоевшего всем генома SARS-CoV-2

6.ЧТО ЗАКОДИРОВАНО В ГЕНОМЕ

Геном коронавируса - молекула РНК

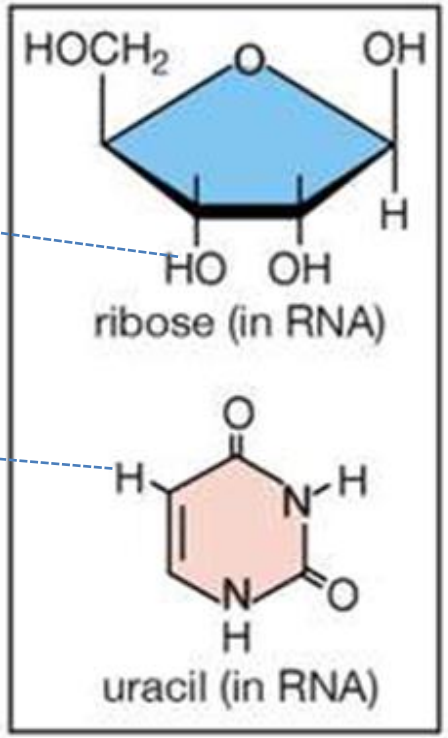


Отличия РНК от ДНК:

ДезоксирибоНуклеиновая Кислота

- **РибоНуклеиновая Кислота**
- **одноцепочечная**

U урацил вместо T



Всего-то отличий, а какая разница в биологии!!!

Небольшие отличия в формуле – большие отличия в пространственной структуре и функциях

Референсный геном SARS-CoV-2

>NC_045512.2 Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1,complete genome

ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA
GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTG
TTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTG
TCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTG
CTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAA
ATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAA
CCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCAC

Геном SARS-CoV-2 - одна молекула РНК

- Файл с геномом лежит в банке данных: https://www.ncbi.nlm.nih.gov/nuccore/NC_045512
  Содержит:
  - аннотацию – описание генома
  - последовательность; хотя геном – РНК, используются буквы A, T, G, C, а не A, U, G, C. Так принято, для единообразия. Что это РНК написано в аннотации
- В аннотации – описаны гены белков - участки последовательности , кодирующие последовательность аминокислот белка, и другая информация, полученная авторами расшифровки последовательности.

Геном содержит инструкцию для клетки организма хозяина (человека) как размножить вирус SARS-CoV-2. При этом клетка умрёт и, если заражено много клеток, хозяин заболеет!!!

Информация: Текст и читатель

Информацию из генома «читают» белки или молекулярные машины, состоящие из белков и других молекул, например, иногда РНК. И выполняют записанную в геноме инструкцию.

脊髄性筋萎縮症遺伝子治療製品 オナセムノゲンアベパルボベク（ゾルゲンスマ®）の薬理学的特性と臨床試験成績

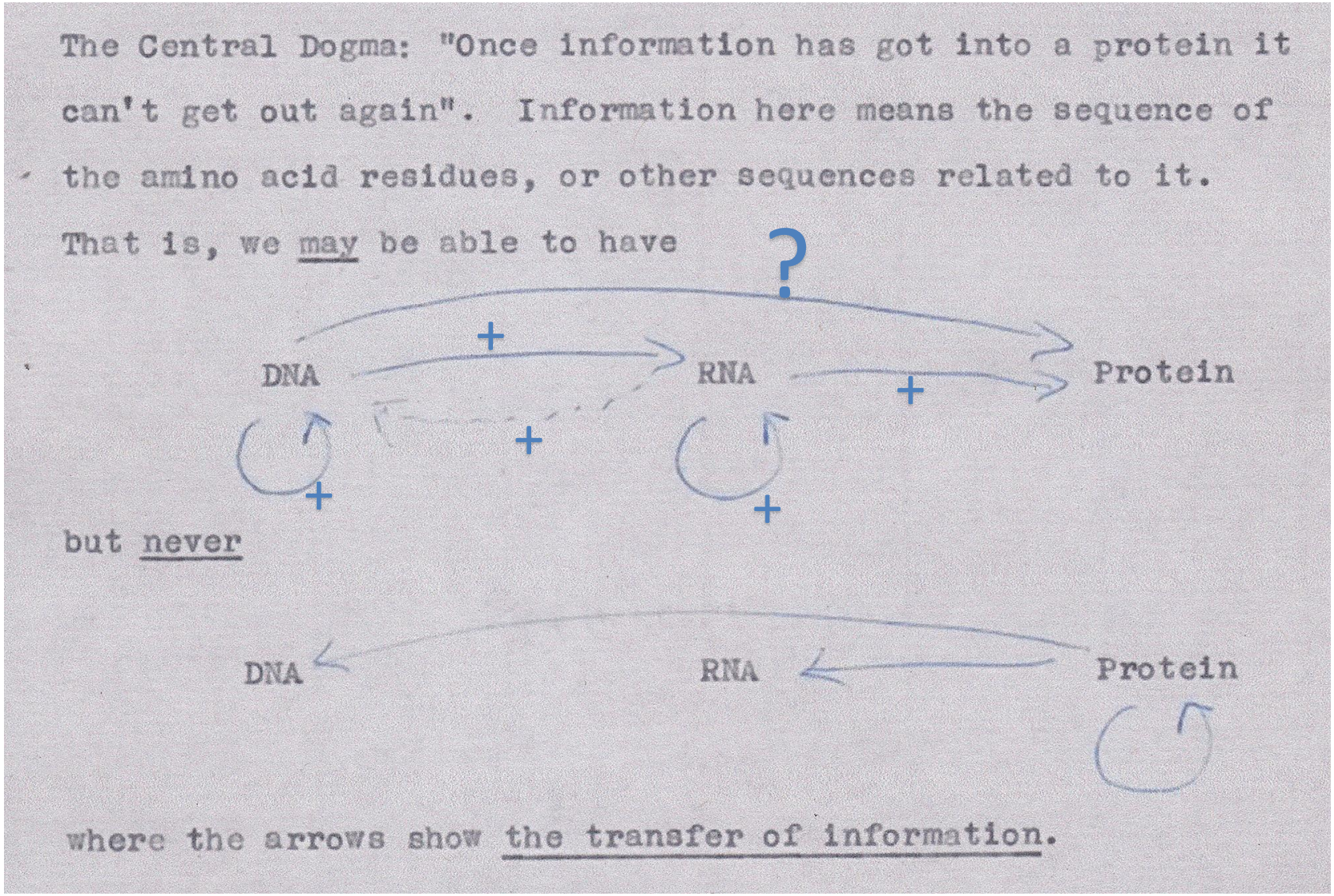
渥美 綾香[1]，米田 智廣[2]，土田 健[2]，香川 雄輔[2]，冨永 俊輔[2]，川瀬 一穂[1]，菊地 信孝[1]

オナセムノゲンアベパルボベク（製品名 ゾルゲンスマ®，開発コード AVXS-101）は，機能的なヒト Survival motor neuron（*SMN*）遺伝子を脊髄性筋萎縮症（SMA）患者の運動ニューロンに届けられるよう設計された，アデノ随伴ウイルス 9 型カプシドを有する非増殖性遺伝子組み換えアデノ随伴ウイルスベクター製品である．2020 年 3 月 19 日に 2 歳未満の「SMA（臨床所見は発現していないが，遺伝子検査により SMA の発症が予測されるものも含む）」を対象に承認された．静脈内に投与された本品は，SMA の根本原因である *SMN1* 遺伝子の機能欠損を補って運動ニューロンの変性・消失を防ぎ，神経及び筋肉の機能を高め，筋萎縮を防ぐことで，SMA 患者の生命予後及び運動機能を改善することが期待される．また，導入された *SMN* 遺伝子は患者のゲノム DNA に組み込まれることなく細胞の核内にエピソームとして留まり，運動ニューロンのような非分裂細胞に長期間安定して存在するように設計されていることから，1 回の静脈注射で治療が完結する．SMA モデルマウスへの本品投与により，SMN タンパク質の持続的発現，体重増加，運動機能の改善，生存期間の延長等が認められた．臨床でのオナセムノゲンアベパルボベクの有効性は，Ⅰ型 SMA 患者（CL-101 試験）及び未発症の SMA 患者（CL-304 試験）を対象とした臨床試験にて確認され，両試験において自然歴に比べ有意に「出生から永続的な呼吸補助が必要となる又は死亡までの期間」が延長されることが示された．また，両試験とも自然歴では見られない運動マイルストーンの達成も確認

Что записано в геноме

- **Гены белков** – участки ДНК, кодирующие аминокислотную последовательность белка.
  - **ДНК-зависимая РНК-полимераза** переписывает участок ДНК содержащий ген белка в матричную РНК с той же последовательностью оснований (с заменой Т на U)
  - **Рибосома** в соответствии с триплетами мРНК синтезирует белок
- **Гены молекул РНК,** отличных от мРНК
- **Сигналы** для белков и молекулярных машин

Crick's first outline of the central dogma, from an unpublished note made in 1956.



Что записано в геноме

- **Гены белков** – участки ДНК, кодирующие аминокислотную последовательность белка.
  - **ДНК-зависимая РНК-полимераза** переписывает участок ДНК содержащий ген белка в матричную РНК с той же последовательностью оснований (с заменой Т на U)
  - **Рибосома** в соответствии с триплетами мРНК синтезирует белок
- **Гены молекул РНК,** отличных от мРНК
- **Сигналы** для белков и молекулярных машин

Что записано в файле с геномом SARS-CoV-2 для людей?

- Последовательность
- Аннотация – формализованное описание того, что известно про последовательность

Из аннотации записи NC_045512 в банке данных Refseq на сайте NCBI

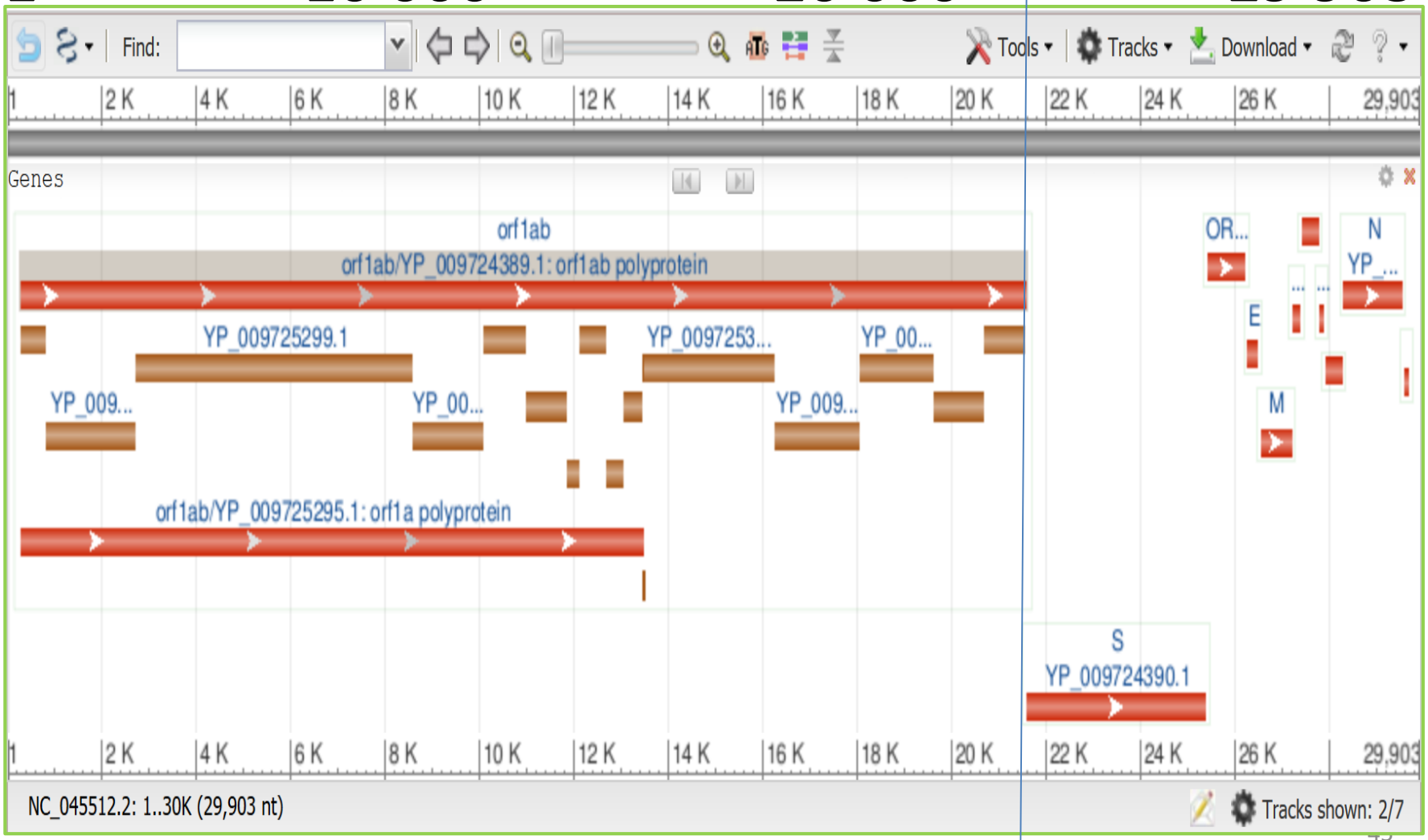
LOCUS NC_045512 29903 bp ss-RNA linear VRL 18-JUL-2020
DEFINITION Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome.
ACCESSION NC_045512

gene 21563..25384 /gene="S" /gene_synonym="spike glycoprotein"
CDS 21563..25384 /gene="S" spike protein" /product="surface glycoprotein" /protein_id="YP_009724390.1"
/translation="MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFR
SSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIR
GWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVY
……
……

ГЕНЫ БЕЛКОВ ранние и поздние гены разделены линией

По оси X нуклеотиды РНК:

1 10 000 20 000 29 903



Гены изображены красными и коричневыми полосками

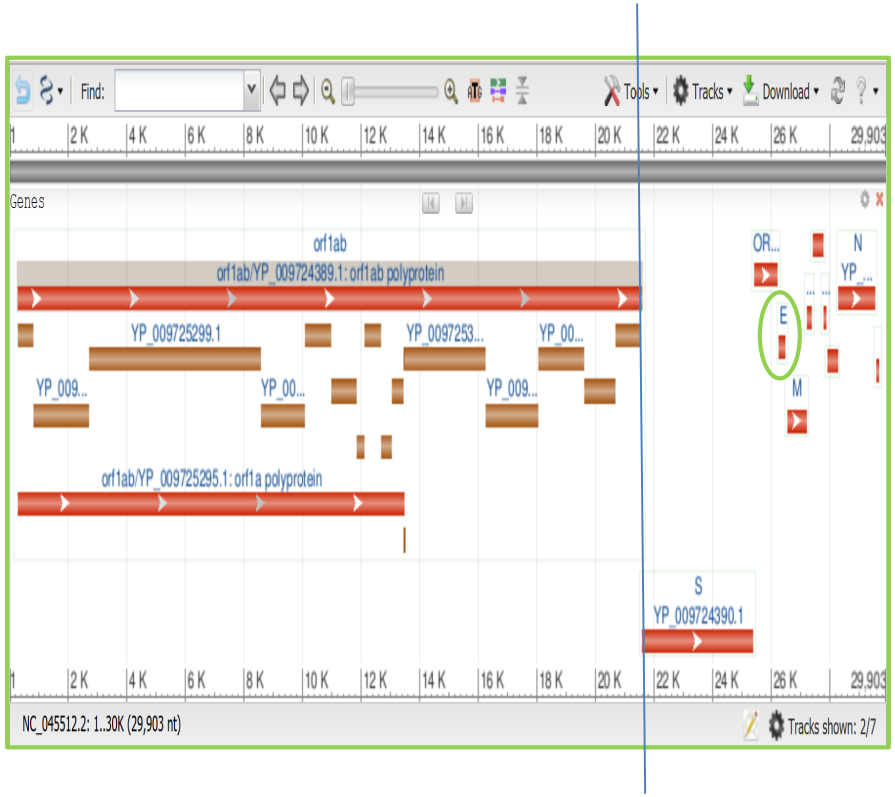
ГЕНЫ ORF1ab и ORF1a

красные полоски до вертикальной линии

А зрелые белки, которые они кодируют, изображены коричневыми полосками.

Как так?

И почему два гена на одном месте, но разной длины?



При заражении covid-19 в клетке хозяина (человека) оказывается РНК вируса.

РИБОСОМА (молекулярная машина для синтеза белков) опознаёт ее как мРНК - матричную РНК гена

Как удаётся коронавирусу выдать свою РНК за мРНК? Нужны соответствующие сигналы…

ТРАНСЛЯЦИЯ: СИНТЕЗ МОЛЕКУЛЫ БЕЛКА

В клетке для синтеза белка нужна матричная РНК

- мРНК это молекула РНК, на которой записана копия гена
  (комплементарная к комплементарной цепочке гена!
  «минус на минус = плюс»)
- Белок синтезирует рибосома используя мРНК
- В клетке человека много РНК разных типов

  (в лекции А.Жариковой будет об этом).
- Рибосома отличает мРНК по двум сигналам
  - Специальная группа атомов cap на 5’ конце.
  - ПолиА на 3’-конце

В клетке хозяина (человека) оказывается РНК коронавируса

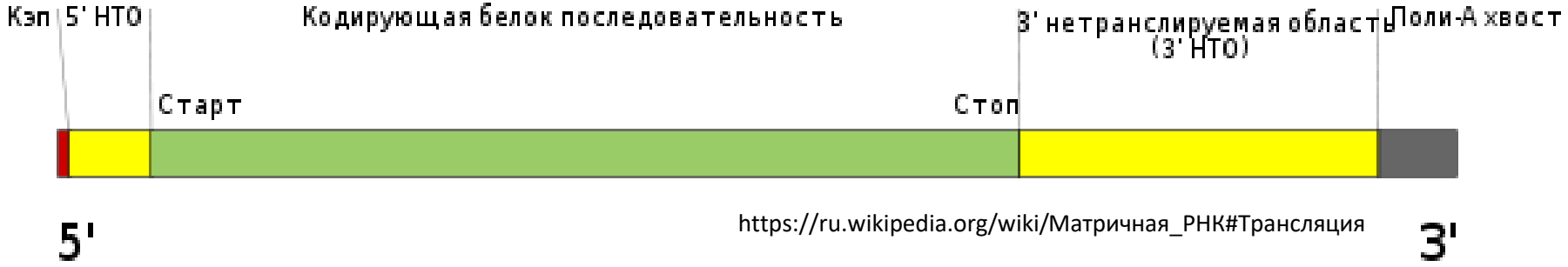
РНК коронавируса несет оба эти сигнала.

- ПолиА на 3’-конце конце РНК видны в файле с РНК (см)
- Кэп тоже есть

……AAAATTAATTTTAGTAGTGCTATCCCC
ATGTGATTTTAATAGCTTCTTAGGAGAAT
GAC**AAAAAAAAAAAAAAAAAAAAAAAAAA**
AAAAAAA

Одна мРНК – один ген и один белок

У человека и других эукариот. Бывают исключения, но они редки!!!



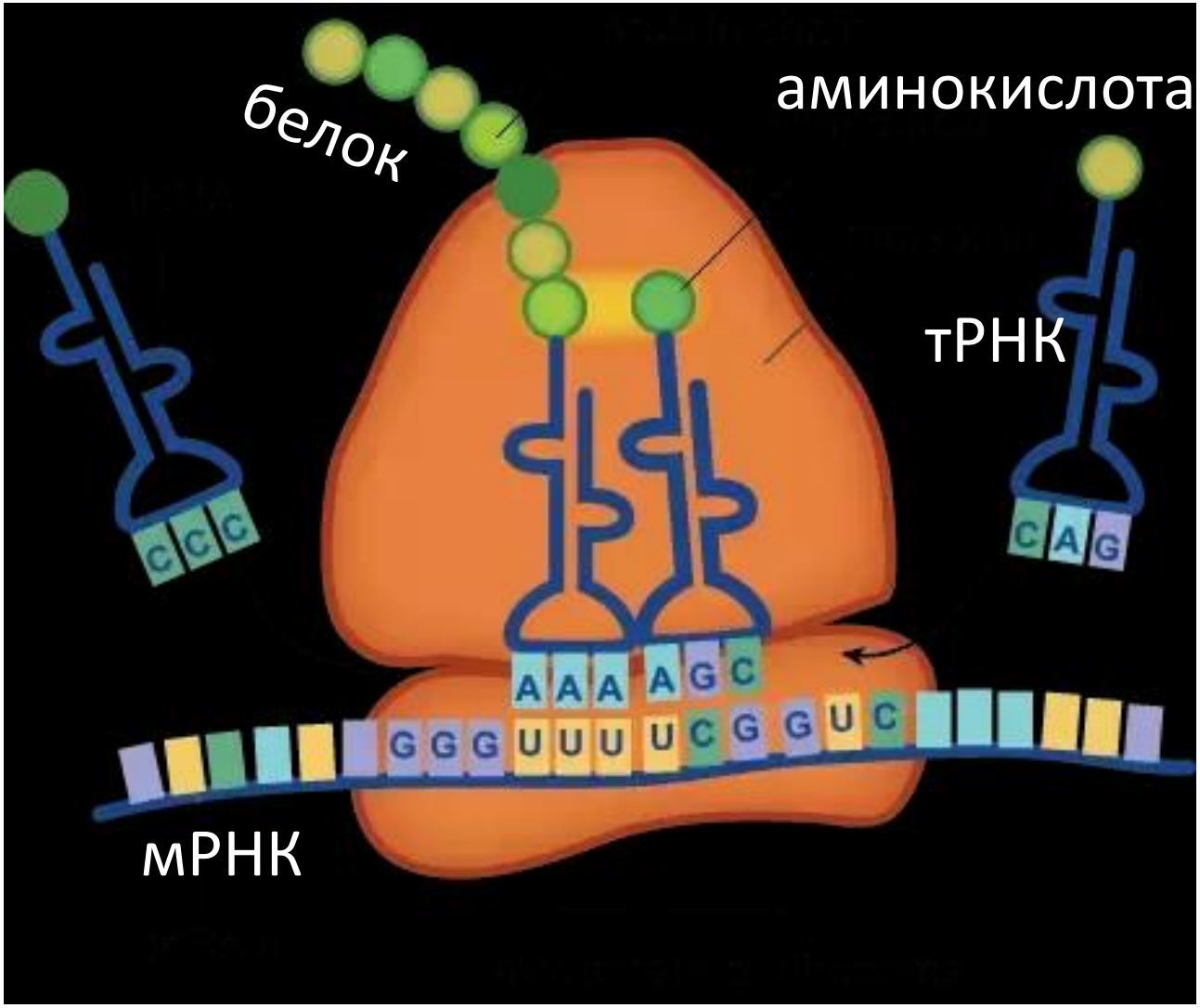
* Малая субъединица рибосомы узнает кэп и сканирует мРНК до появления кодона ATG.
Причем не просто ATG, а ATG в подходящем окружении, так называемая последовательность Ко́зак

* С триплета ATG начинается синтез белка рибосомой, по кодонам. Заканчивается на стоп кодоне.

По таблице генетического кода!

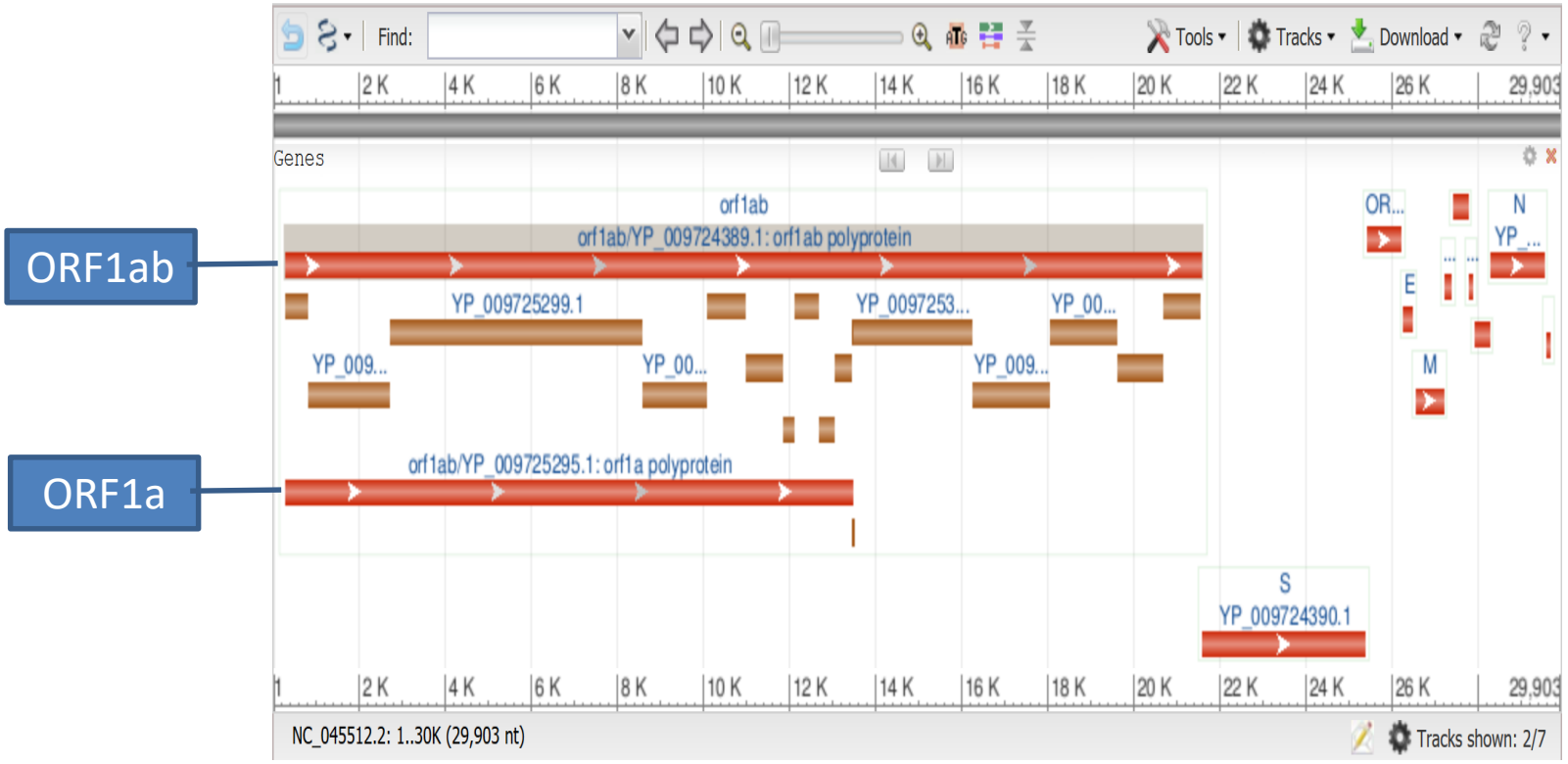
AAA	K	CAA	Q	GAA	E	TAA	**Stop**
AAG	K	CAG	Q	GAG	E	TAG	**Stop**
AAC	N	CAC	H	GAC	D	TAC	Y
AAT	N	CAT	H	GAT	D	TAT	Y
ACA	T	CCA	P	GCA	A	TCA	S
ACG	T	CCG	P	GCG	A	TCG	S
ACC	T	CCC	P	GCC	A	TCC	S
ACT	T	CCT	P	GCT	A	TCT	S
AGA	R	CGA	R	GGA	G	TGA	**Stop**
AGG	R	CGG	R	GGG	G	TGG	W
AGC	S	CGC	R	GGC	G	TGC	C
AGT	S	CGT	R	GGT	G	TGT	C
ATA	I	CTA	L	GTA	V	TTA	L
ATG	M	CTG	L	GTG	V	TTG	L
ATC	I	CTC	L	GTC	V	TTC	F
ATT	I	CTT	L	GTT	V	TTT	F

Так рибосомы делают белки



Ген ORF1a заканчивается стоп-кодоном, как положено.

Так и транслируется рибосомой человека

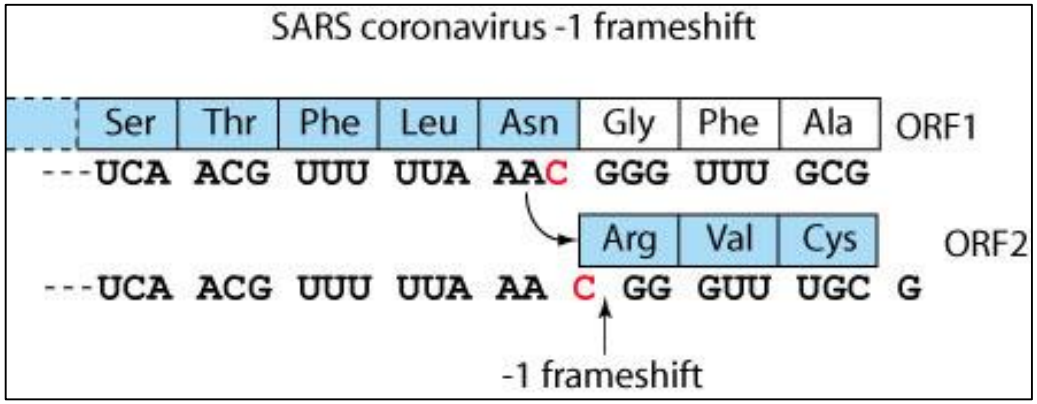


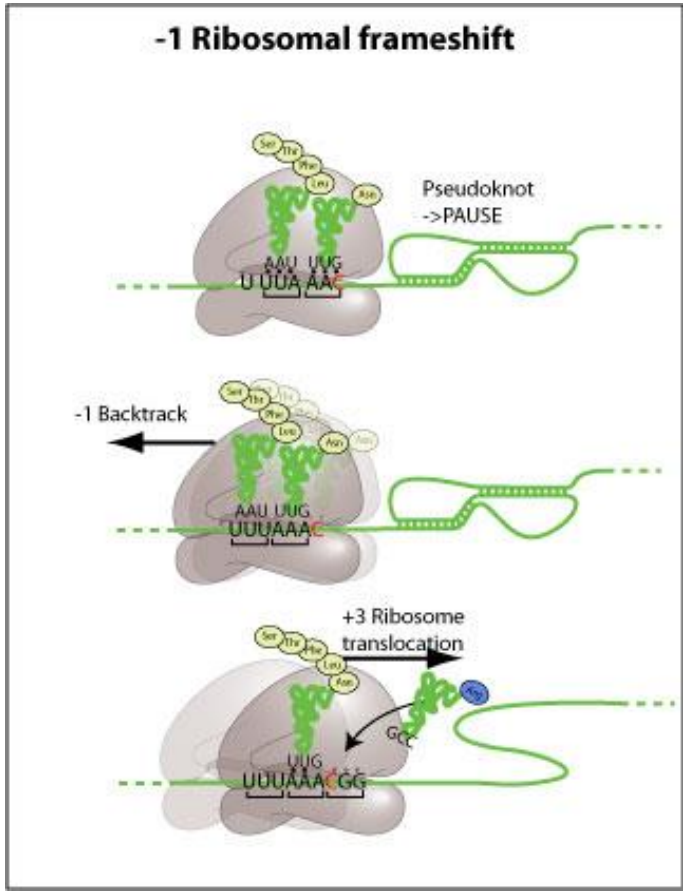
Как же транслируется ген ORF1ab???

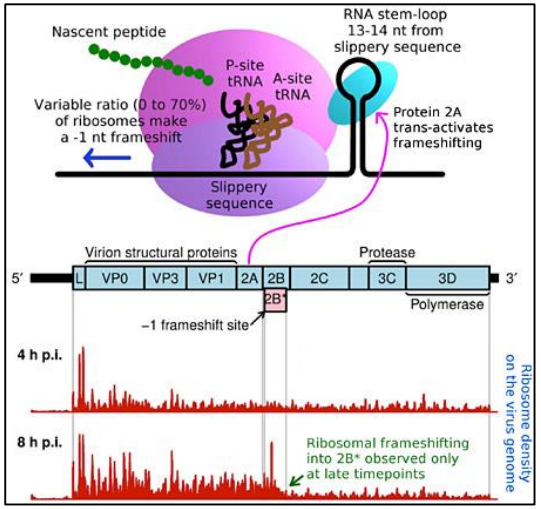
Не бывает правил без исключений!

<u>Программируемый</u> рибосомный сдвиг.

Рибосома останавливается из-за шпильки на РНК и slippery sequence. Отскакивает на ОДИН нуклеотид(букву). И продолжает синтез белка







Продукты генов ORF1ab и ORF1a - Большие белки – полипротеины. См. след слайд

Они сами себя разрезают на отдельные белки

ORF1a = polyprotein 1a (pp1a)

ORF1b = polyprotein 1ab (pp1ab)

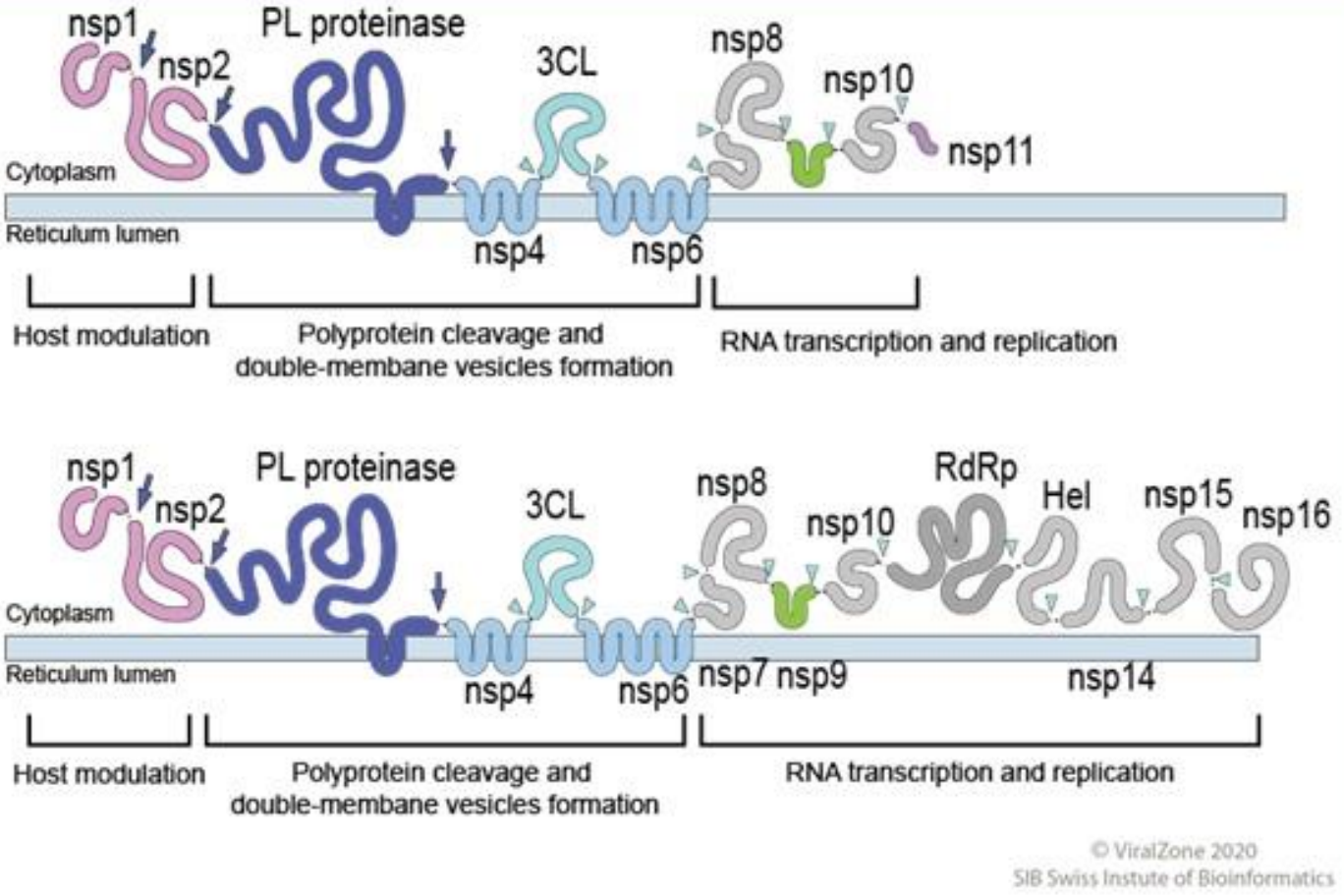


Figure 2: SARS-Cov-2 polyproteins

There are two major, partially overlapping, ORFs in the 5' two-thirds of the viral genome. ORF1a is frameshifted to ORF1b, and these encode the replicase polyproteins pp1a and pp1ab respectively. Polyproteins pp1a and pp1ab are proteolytically cleaved into 16 putative non-structural proteins (nsps).

Функции некоторых ранних белков

Name	Число а/к	зачем нужен
NSP1	180	Деградирует некоторые хозяйские РНК
NSP3a	1945	Протеиназа, отрезает nsp1, nsp2, nsp3
NSP5a	306	Протеиназа, режет полипротеин в 11 местах
NSP8	198	Помогает при репликации РНК
NSP12a	932	Полимераза – по РНК делает комплементарную РНК (RDRp)
NSP13	601	Хеликаза (расплетает двойную спираль РНК)
NSP14	527	Присоединяет сар к РНК
NSP15a	346	Уклонение от защиты хозяйских клеток

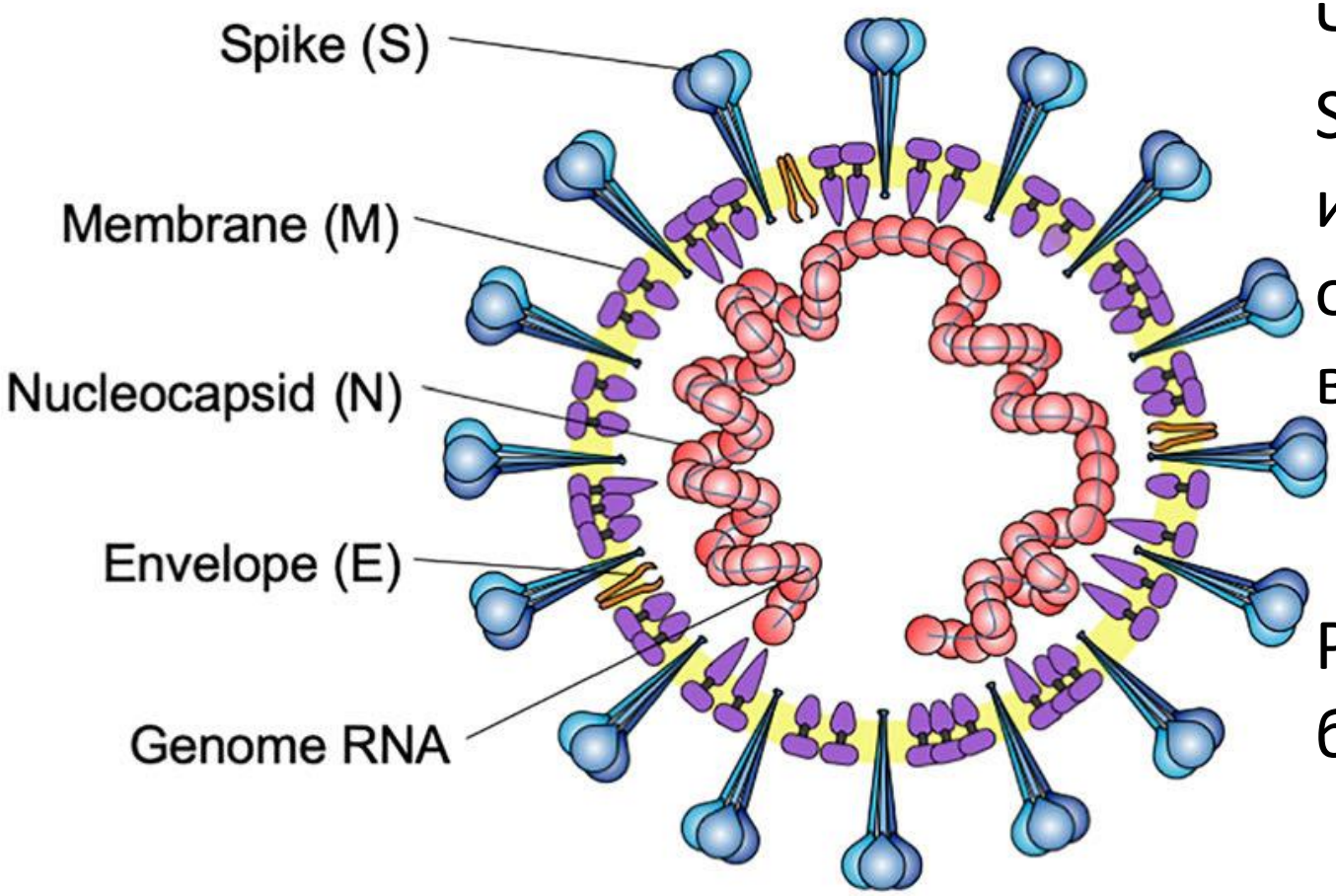
Yoshimoto, *Protein J* **39,** 198–216 (2020).

Как транслируются поздние белки?

Среди поздних генов все белки составляющие вирион, который существует пока вирус вне хозяина.
Оно и понятно – вирион собирается в конце заражения, когда есть много РНК – геномов и пора выходить из клетки.

Коронавирус вирус: белки

Схема вириона, существующего между заражениями



Четыре белка:
S, M, E, N
и РНК
оставляют
вирион

РНК облеплена
белками N

Xu J, et al. Antibodies and vaccines against Middle East respiratory syndrome coronavirus. Emerg Microbes Infect 2019, Review. MERS-CoV

Функции белков

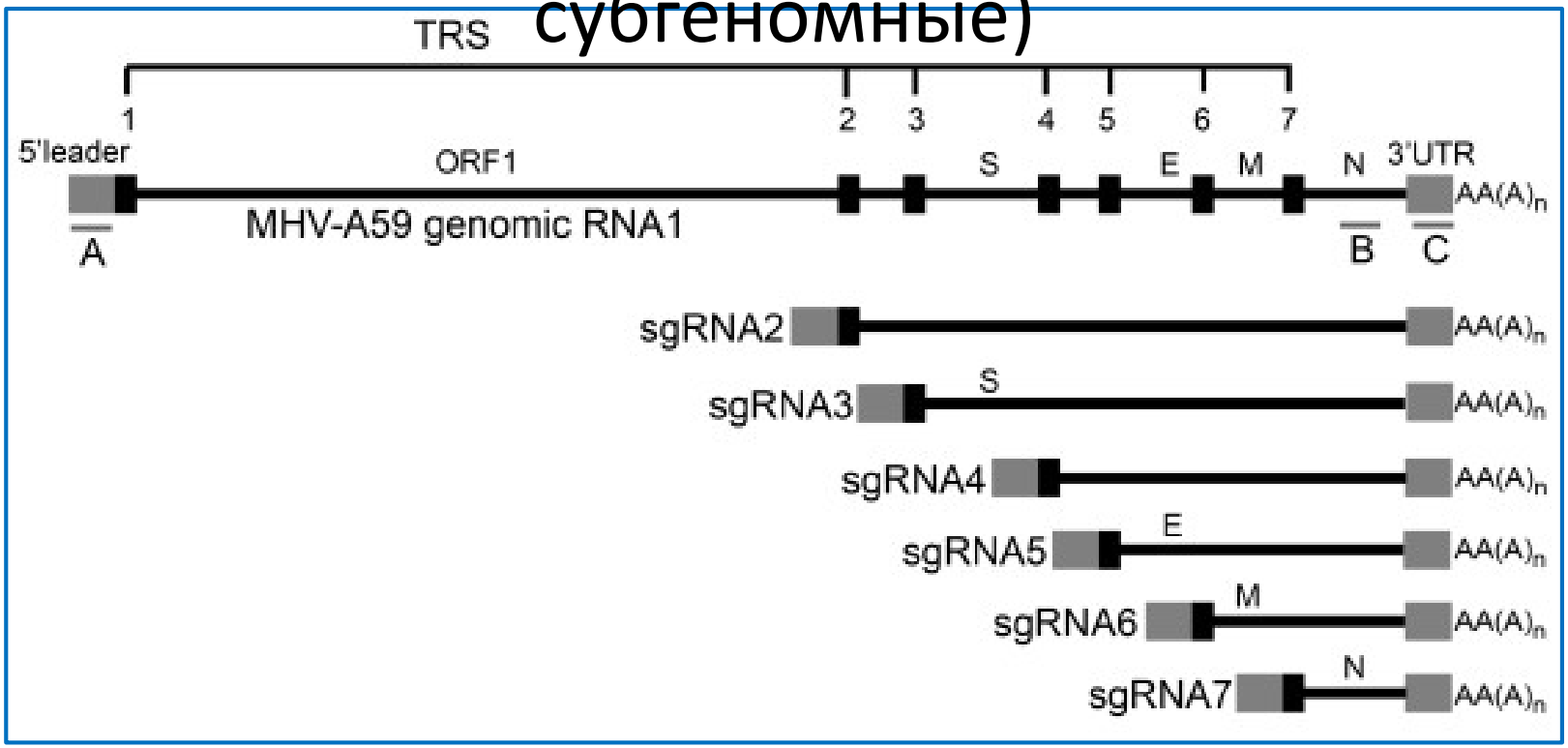
- M белок составляет оболочку (капсид) вириона, вместе с липидной мембраной (желтая)

- E белок нужен для правильной кривизны капсида. 2я функция - в хозяйской клетке. Пентамер E является ионным каналом в мембране органеллы ERGIC[1)] Коронавирусы, лишенные E, могут размножаться, хотя и менее патогенны

- N белок облепляет РНК в конформации **бусы на струне** для сохранности генома. При сборке капсида он обладает удивительной способностью связываться только с РНК коронавируса!

1) Между эндоплазматическим ретикулумом и аппаратом Гольджи)

Коронавирус синтезирует отдельные мРНК для поздних генов (называют сгмРНК, субгеномные)

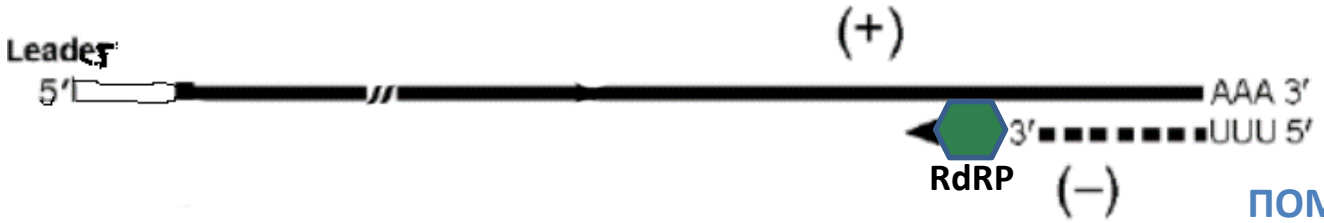


ИДЕЯ коронавируса: лидерную последовательность «склеить» с участком начиная от позднего гена и до конца!

Сохраняются все 5’ концевые и 3’ концевые сигналы (КЭП, полиА и др.)

СДЕЛАВ ЭТО ПРЕДОК КОРОНАВИРУСОВ ЗАКРИЧАЛ ЭВРИКА! И заразил множество хозяев.

РНК зависимая РНК полимераза коронавируса



ПОМНИМ:
В РНК U вместо Т

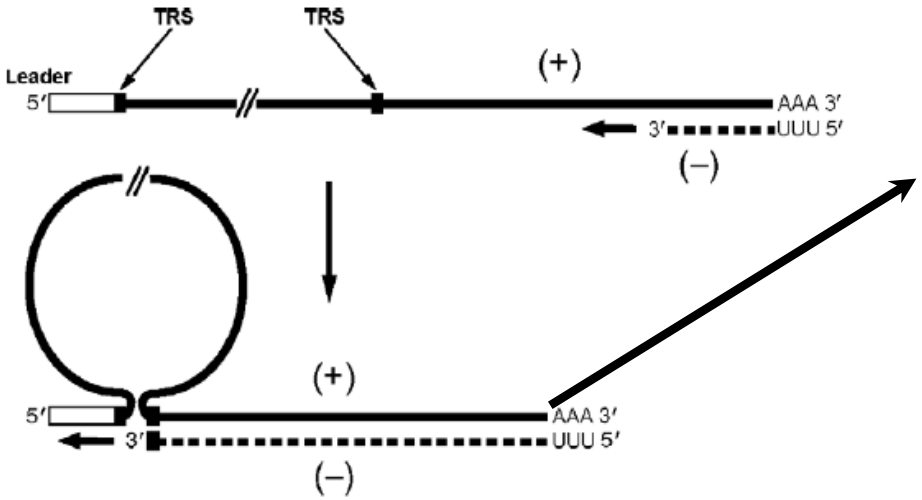
- Для синтеза новых РНК нужных для сборки новых частиц вируса нужен белок «РНК зависимая РНК полимераза, RdRP», его ген 11й среди зрелых ранних белков коронавирусов.

- С РНК коронавируса, которую обозначают +РНК, он делает комплементарную копию, называемую -РНК

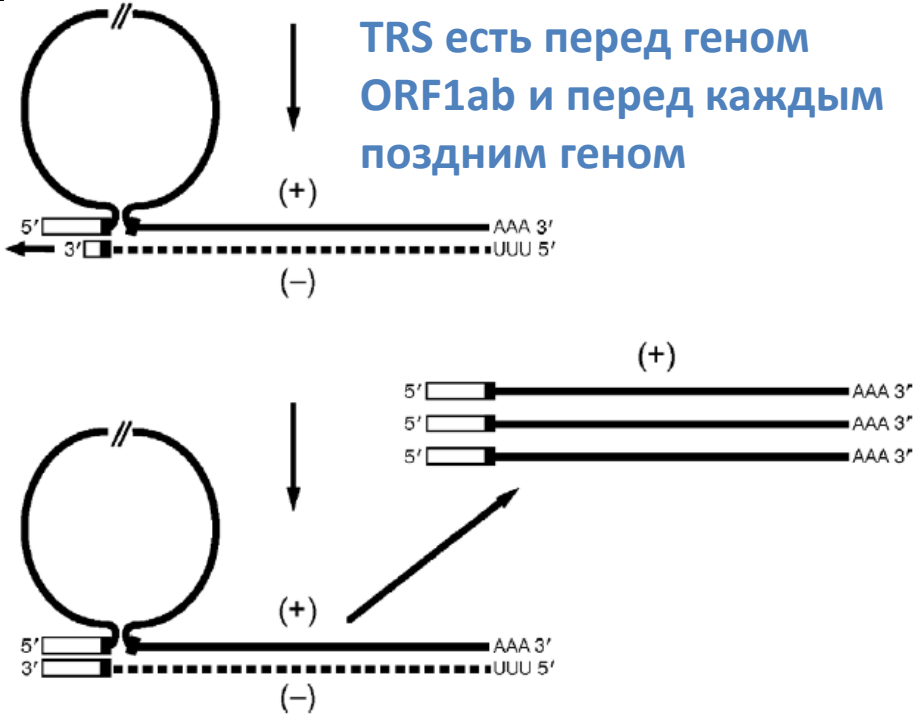
- RDRP может копировать любую РНК, в частности, -РНК

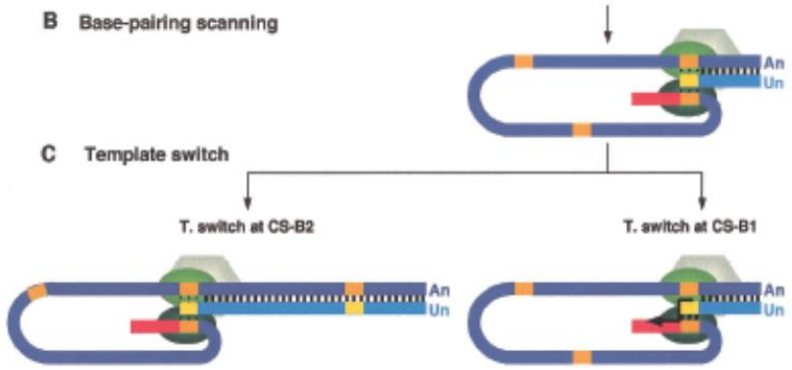
- Минус на минус будет плюс!!!

Коронавирусам (эволюции) пришлось долго ломать голову чтобы придумать такое! Сигналы TRS – последовательности похожие на CTAAAC – обозначены черным прямоугольником, желтым на цветном рисунке.



TRS есть перед геном ORF1ab и перед каждым поздним геном





По –сгРНК полимераза RdRP делает +сгмРНК.
И рибосома транслирует первый ген на ней в белок

ОТКУДА ВСЕ ИЗВЕСТНО?

- Последовательность генома SARS-CoV-2
  - Севенирование нового поколения (NGS)

- Определение родства
  - биоинформатика
- Гены
  - Биоинформатика – сравнение с белками известных коронавирусов
  - Эксперимент+биоинформатика: NGS транскриптом
- Белки
  - Масс-спектроскопия
  - биоинформатика

Сигналы

- Сигнал полиаденилирования мРНК (последовательность полиА не содержится в гене, а присоединяется после транскрипции)
  - Алгоритм поиска
- Программируемый сдвиг рамки считывания (FrameShift)
  - алгоритмы предсказания
  - Рибосомный профайлинг

- Сайты протеолиза (по которым режутся полипротеины)
  - Биоинформатика
  - Эксперимент, масс-спектроскопия

- Сигналы разрывной транскрипции РНК SARS-Cov-2
  - Транскриптомы
  - биоинформатика

SARS-CoV-3. 30K - большой геном или маленький?

Как посмотреть!

		ответ – вид и число букв в геноме	
Суперцарство	Примеры	Маленький (самый)	Большой (самый)
Вирусы	ВИЧ, Коронавирусы, вирус полиомиелита, ….		
Прокариоты	Бактерии и археи (похожи на бактерий, но другое суперцарство)		
Эукариоты	Животные, растения, грибы, простейшие …		
Абсолютная категория	(может выше забыл каких-нибудь?:)		

Задание 9 (на сайте с ДЗ)

При выполнении этого задания опишите вид и источник информации
Наука идет вперед, мои знания могли устареть(((

Толщина мембраны

Размеры разных клеток человека см. в интернете, например, https://www.youtube.com/watch?reload=9&v=qz53ud-i_sY

Толщина мембраны (липидного бислоя) - 5 нм

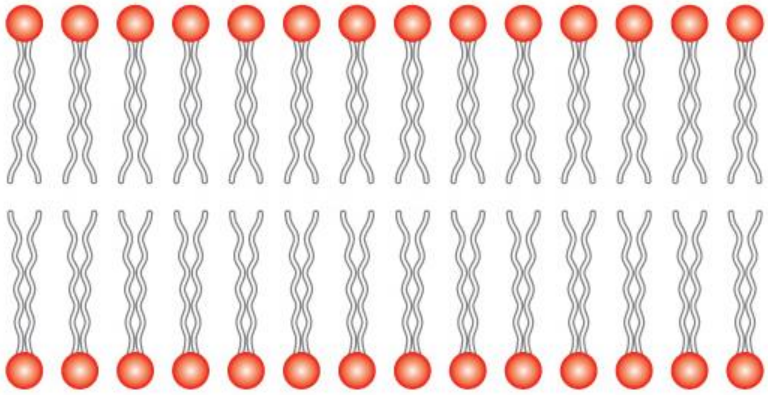
1 метр = 1 000 миллиметров

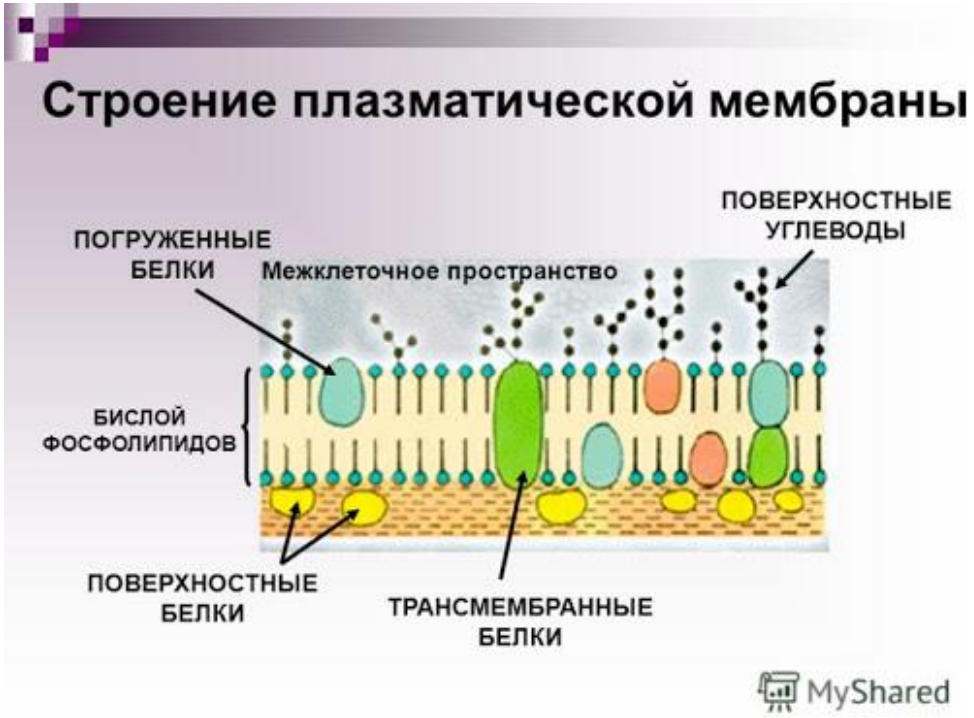
1миллиметр = 1 000 микрометров

1 микрометр = 1 000 нанометров

Задание 4 (классика). Возьмем примерно шарообразную клетку человека. Пропорционально увеличим ее до размера арбуза. Какой толщины будет мембрана?

Липидный бислой





В мембрану включены или к ней присоединены другие молекулы, белки и углеводы

КОНЕЦ ПРЕЗЕНТАЦИИ

S белок главный для проникновения в клетку хозяина.

Шипик короны состоит из трех S.

Клетка защищена мембраной, как крепостная стена, но со всех сторон.

Штурм или хитрость?

S – троянский конь

В крепости есть ворота. Но не каждого пускают.
S-белок мимикрирует под белок человека, с которым знаком рецептор, который называется ACE2 – ангиотензин превращающий фермент 2.

Другой белок протеаза (разрезатель белка) serine protease TMPRSS2 активирует S белок, путем отрезания “шапочки” S-белка

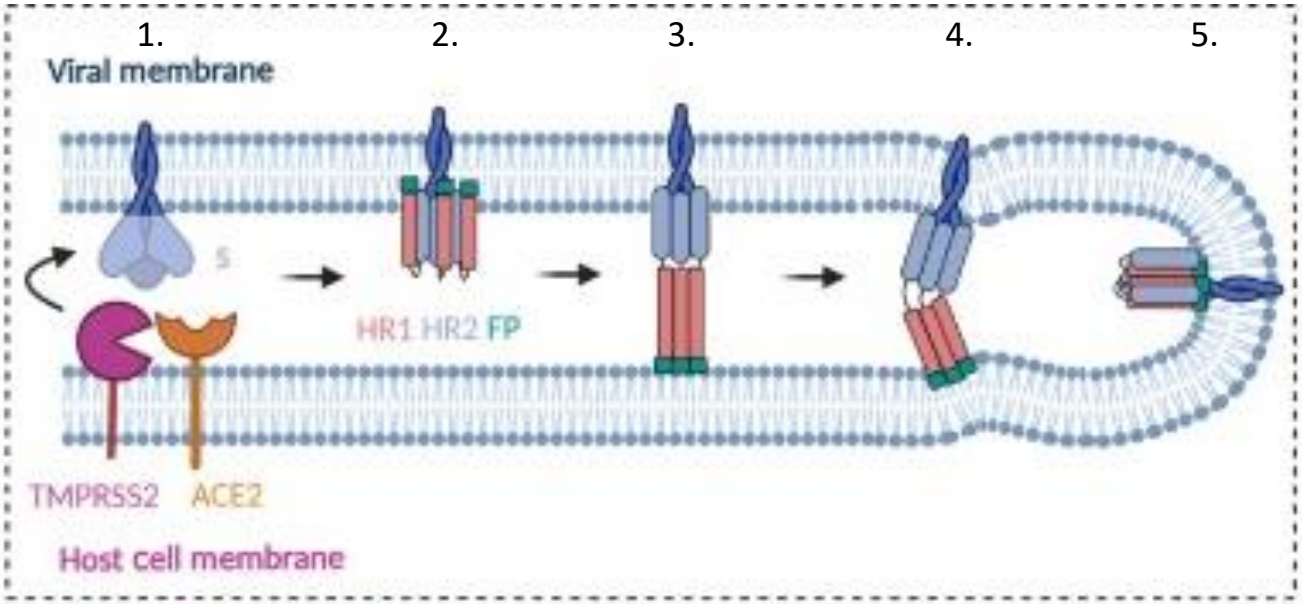
Получив добро, S стягивает мембрану вируса и мембрану хозяйской клетки; мембраны сливаются, РНК оказывается в клетке хозяина.

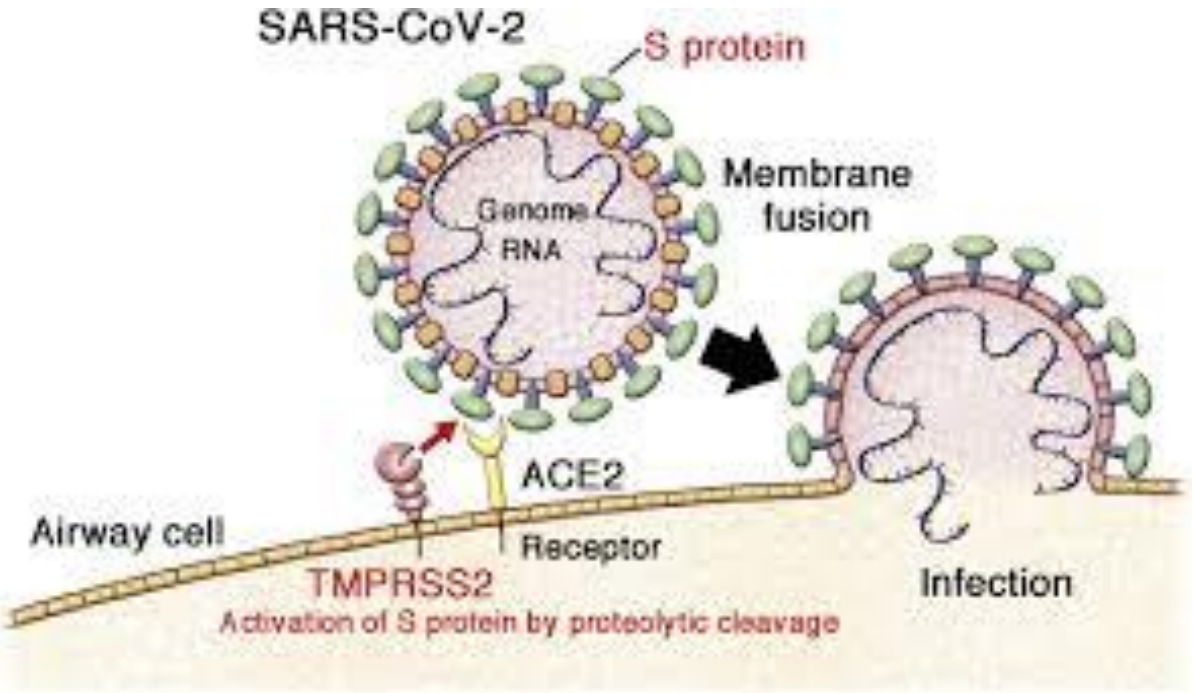
S-белком определяется кого и какую ткань заражает вирус

Как вирус вводит РНК в клетку (в картинках)

1. S-белок связывается с ACE2
2. Протеаза TMPRSS2 отрезает «шапочку» с верхней части S-белка
3. Верхушка S-белка перестраивается и зеленые участки (гидрофобные)
   заякориваются в мембрану клетки
4. S-белок складывается в прежний вид.
5. Мембраны вируса и клетки сливаются
   (как сливаются два мыльных пузыря)

   РНК оказывается внутри клетки хозяина

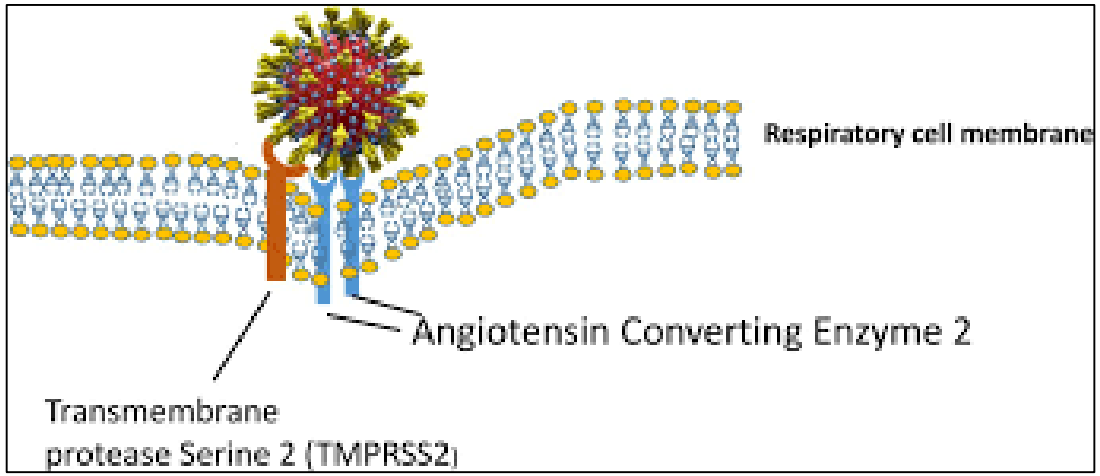




Два белка (ACE2 и TMPRSS2) должны быть рядом в мембране клетки чтобы произошло заражение!!!

Исследуется этим ли определяется выраженность симптомов covid-19

И иммунитетом, конечно.



Задача построения пространственной структуры белка по его последовательности имеет долгую историю

Есть серьёзные успехи alpha-fold