ВОЗМОЖНОСТИ СЕКВЕНИРОВАНИЯ

Анастасия Жарикова – 2022
azharikova89@gmail.com

Что такое секвенирование?

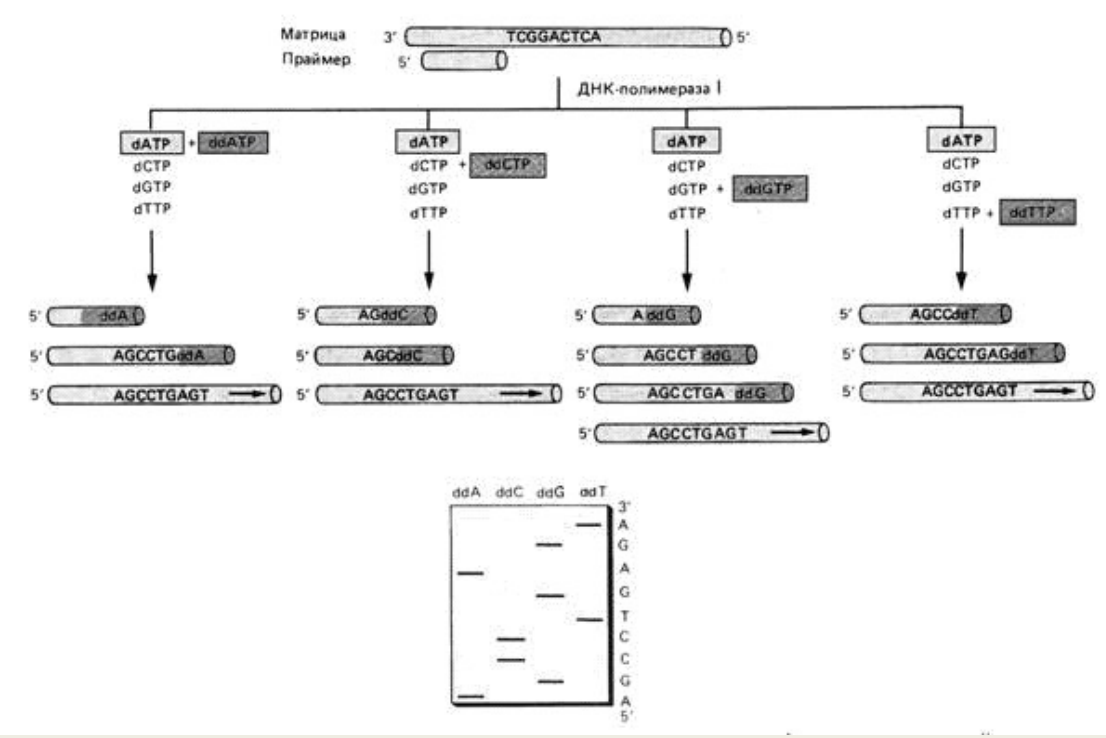
Определение последовательности макромолекулы

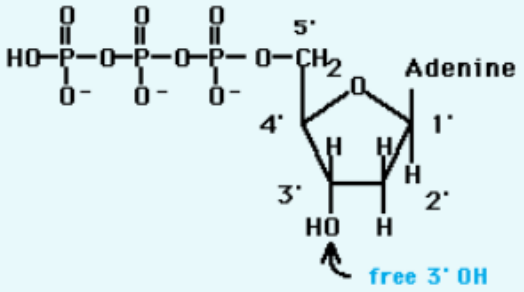
Будем говорить о нуклеиновых кислотах =>

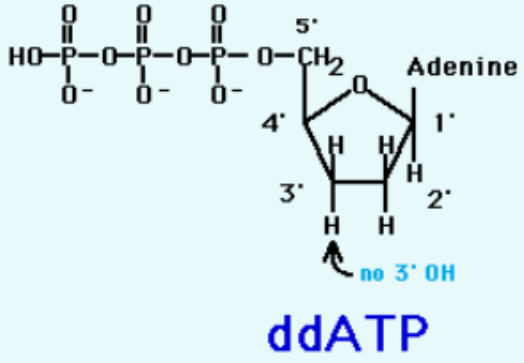
РНК (рибонуклеиновая кислота) или **ДНК** (дезоксирибонуклеиновая кислота) =>

Последовательность нуклеотидов

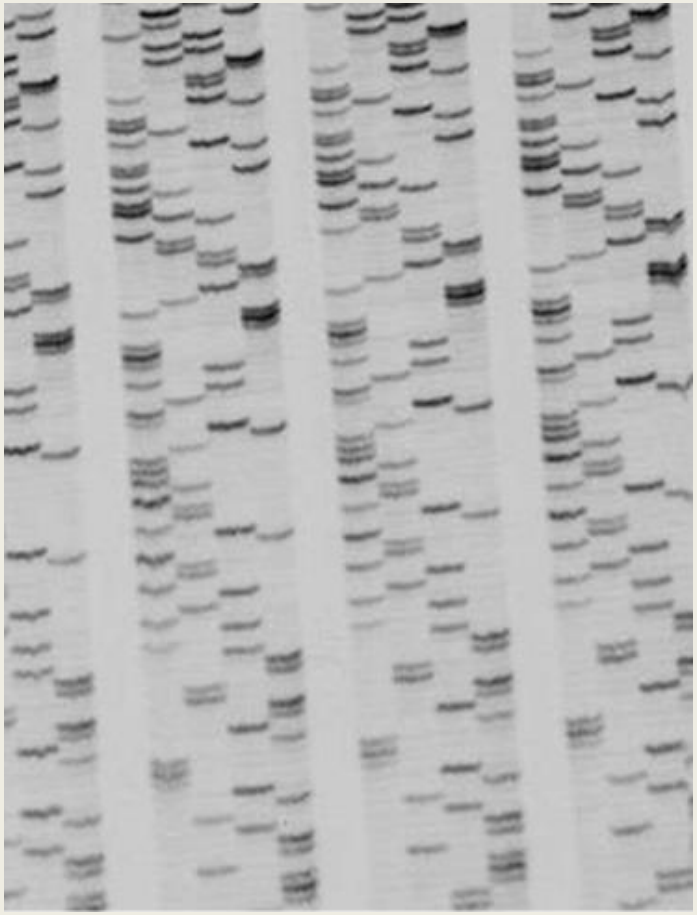
Метод «терминаторов»
секвенирование ДНК по Сэнгеру – 1977г





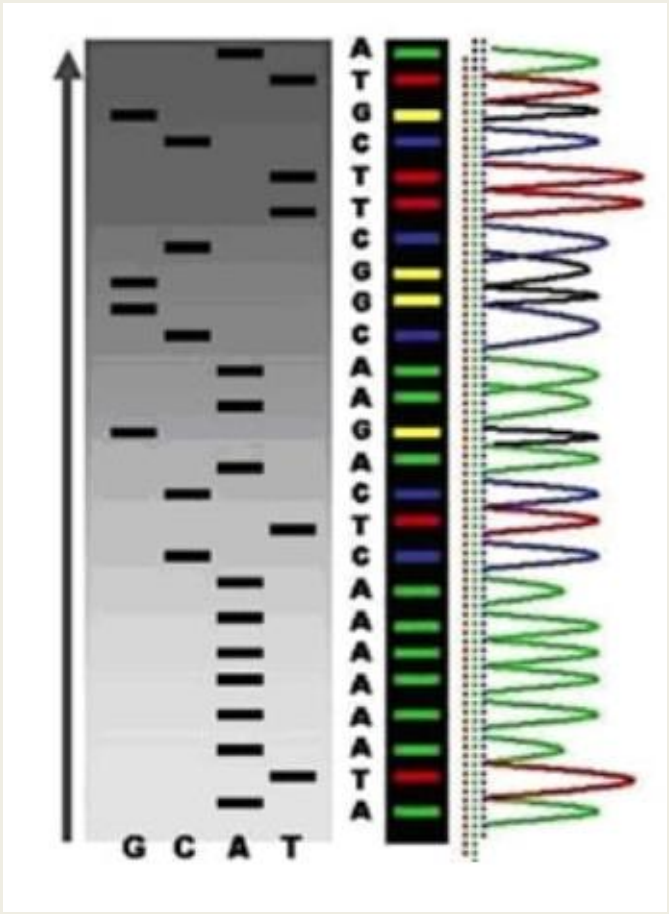


Метод «терминаторов»



Было

Стало

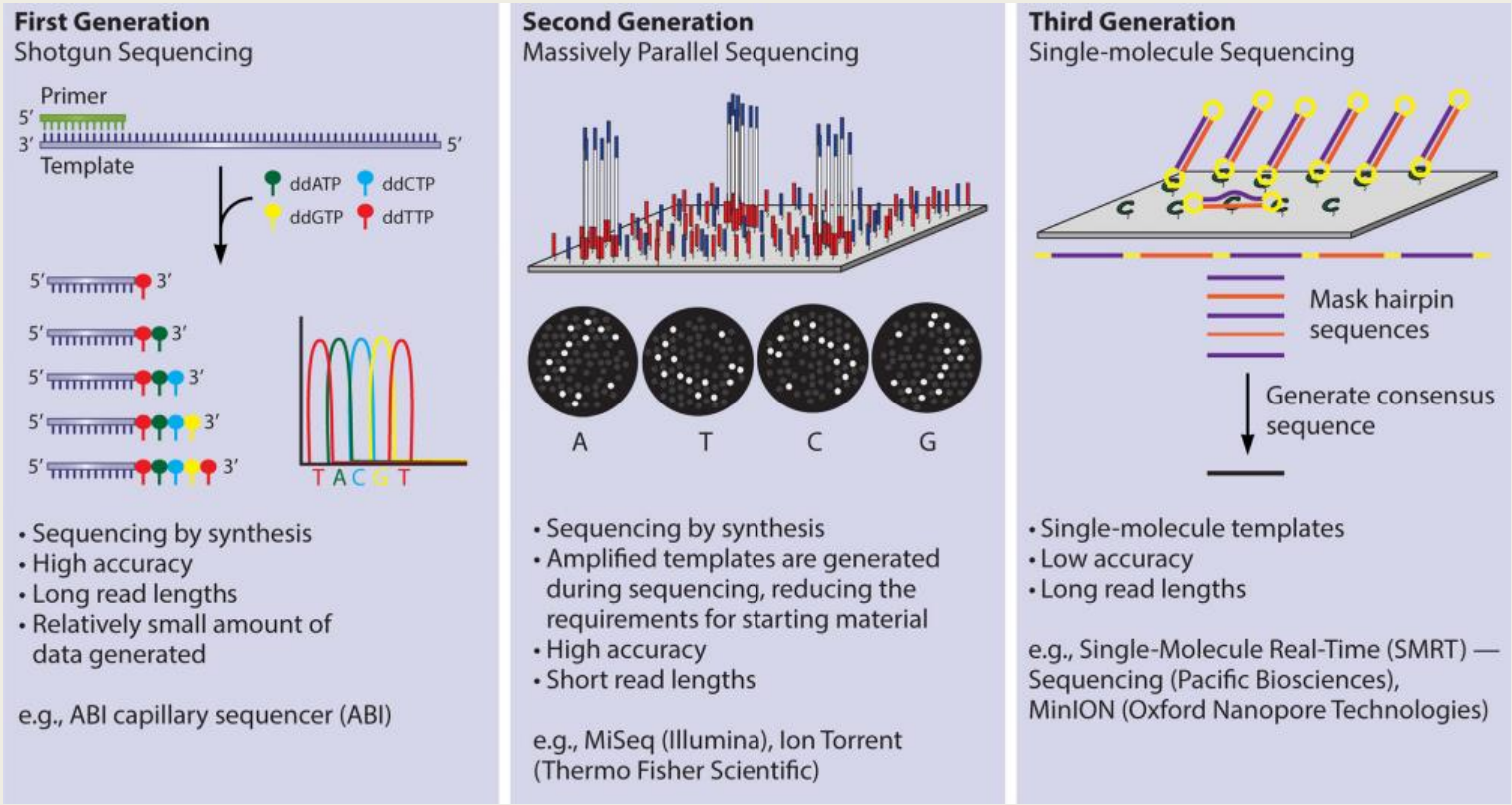


Секвенирование нуклеиновых кислот

«Золотой стандарт»
~ 1000 п.н.

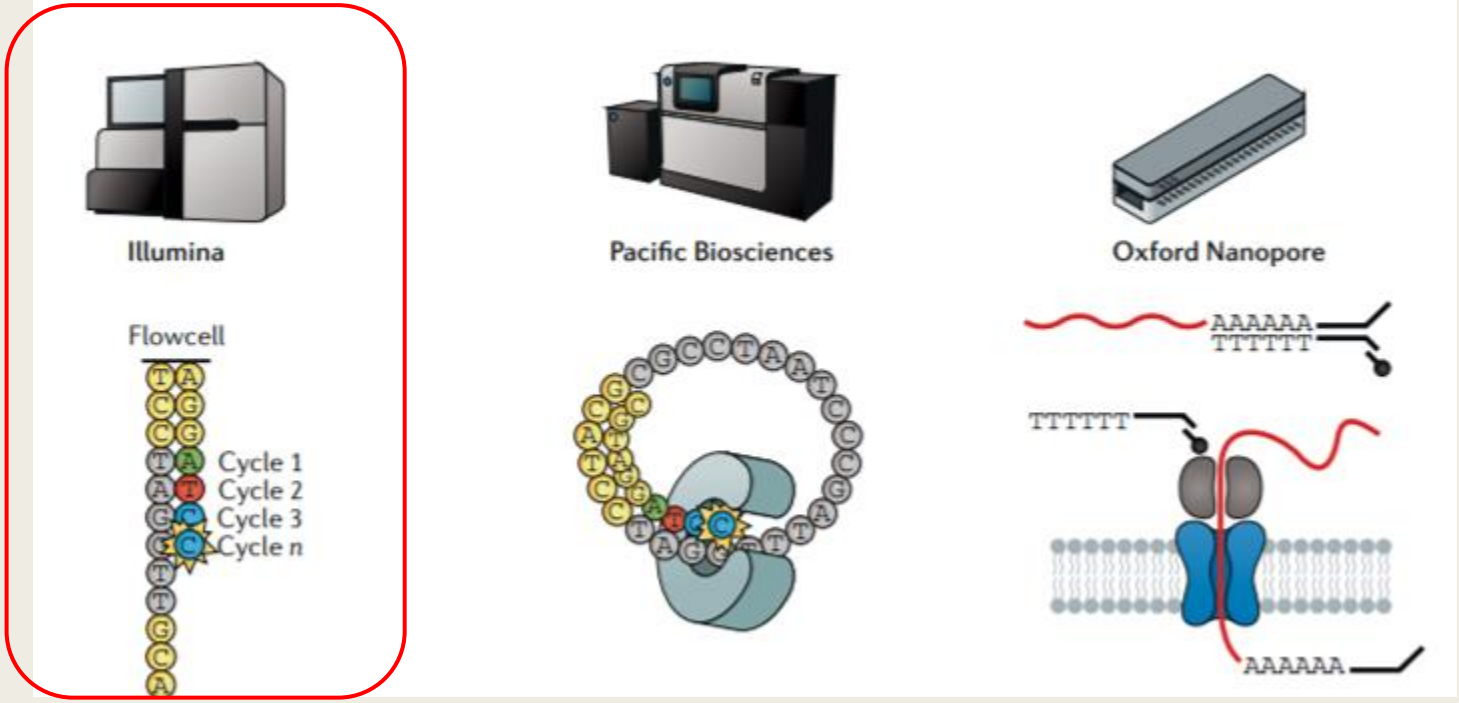
Поколения подходов к секвенированию

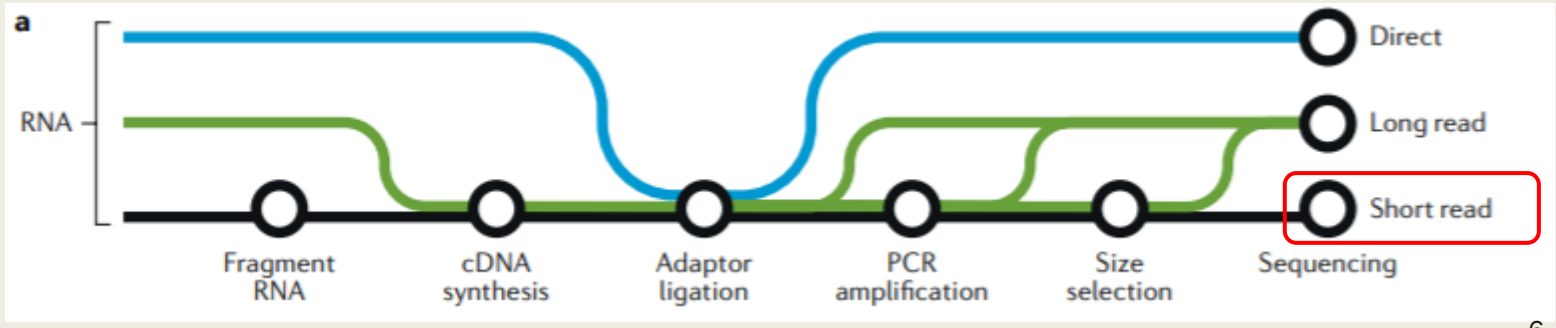
https://pubmed.ncbi.nlm.nih.gov/27559074/



Секвенирование нуклеиновых кислот

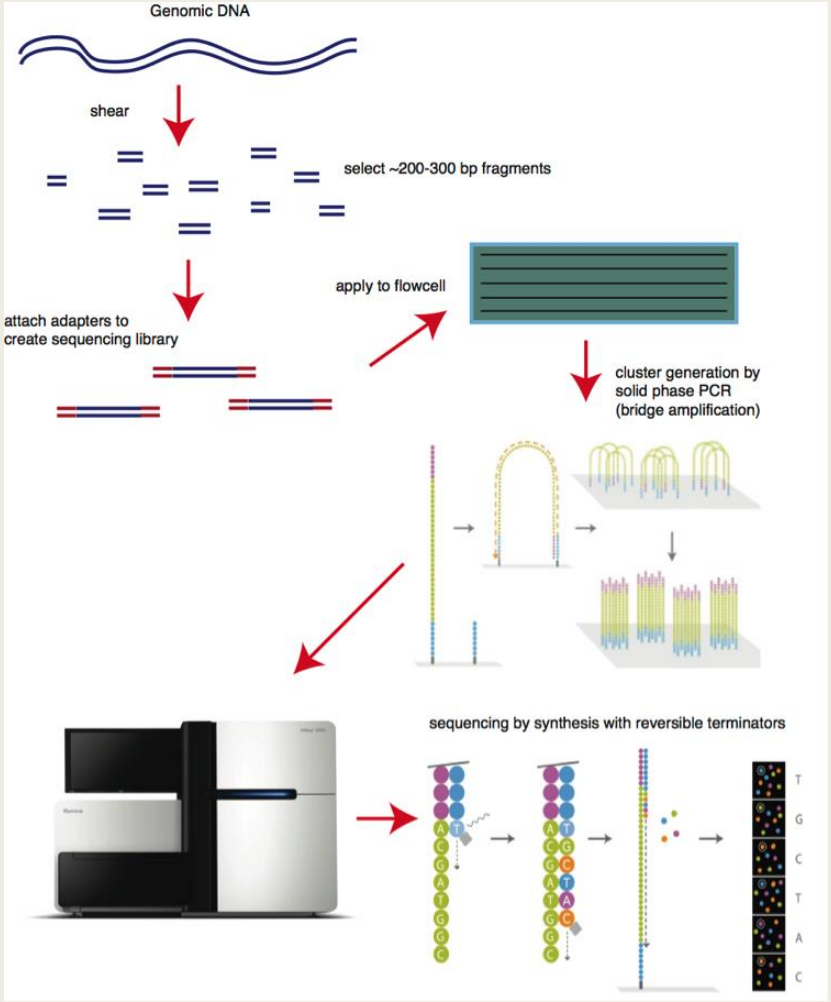
Разнообразие платформ





https://pubmed.ncbi.nlm.nih.gov/31341269/

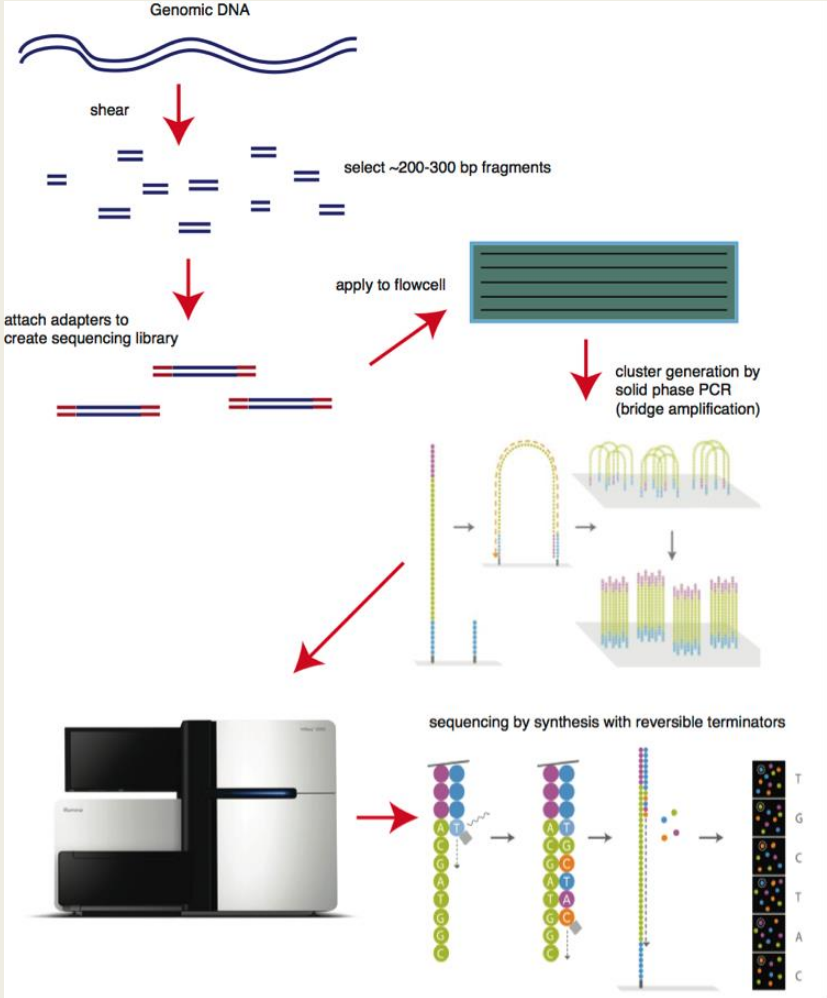
Next generation sequencing – NGS - Illumina



Выделение ДНК
Фрагментация
Подготовка библиотеки
Секвенирование
мультик

https://bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology/

Next generation sequencing – NGS - Illumina



Выделение ДНК
Фрагментация
Подготовка библиотеки
Секвенирование
мультик

Умеем секвенировать только ДНК
А что же с РНК?
РНК => обратная транскрипция => кДНК => секвенирование

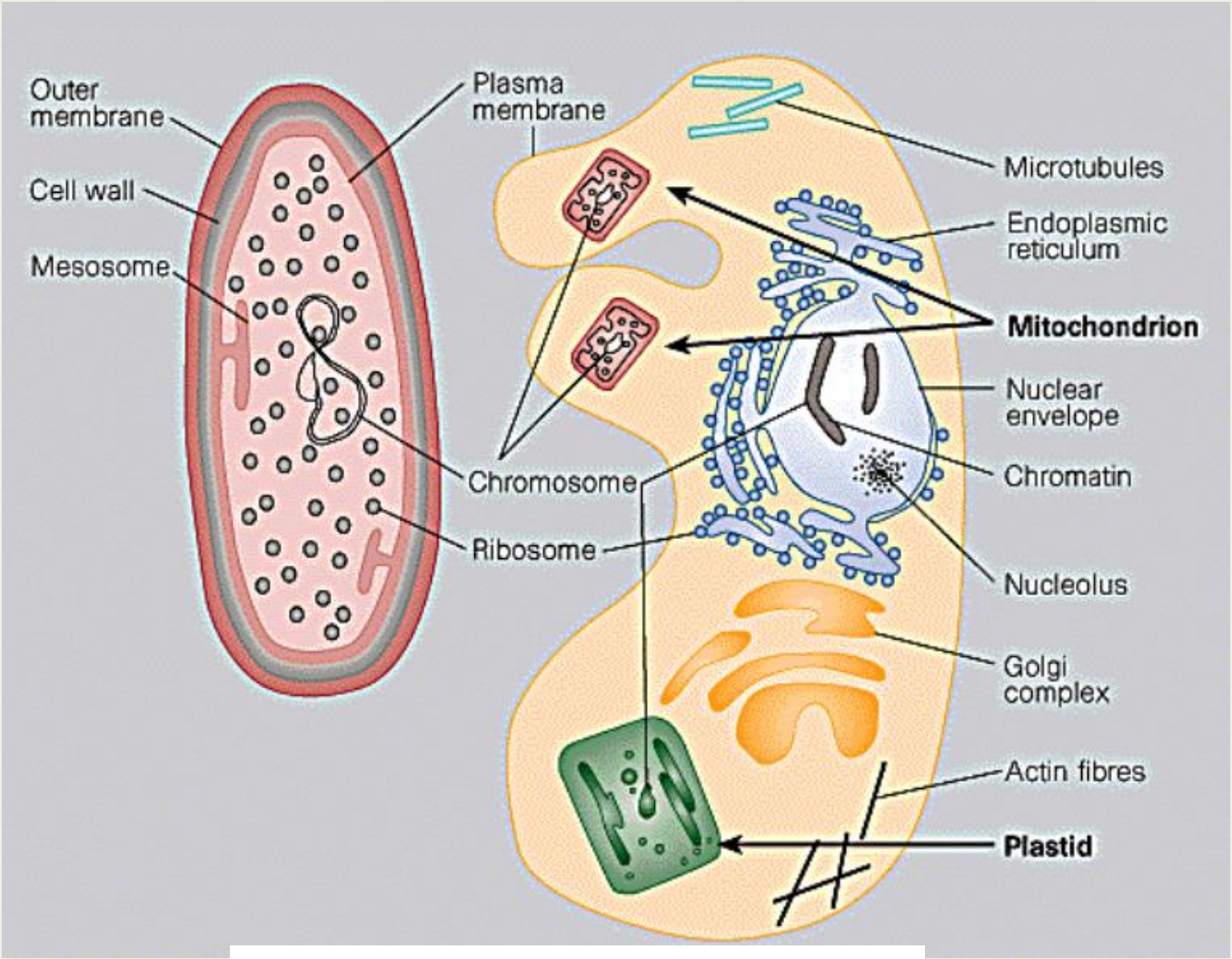
В итоге мы в любом случае секвенируем короткие фрагменты ДНК

DNA-seq & RNA-seq

- Стремимся секвенировать
  - *Всю ДНК => геном*
  - *Конкретные области генома*
  - *Всю РНК => транскриптом*
  - *Отдельные фракции транскриптома*

https://www.nature.com/scitable/topicpage/eukaryotic-cells-14023963/

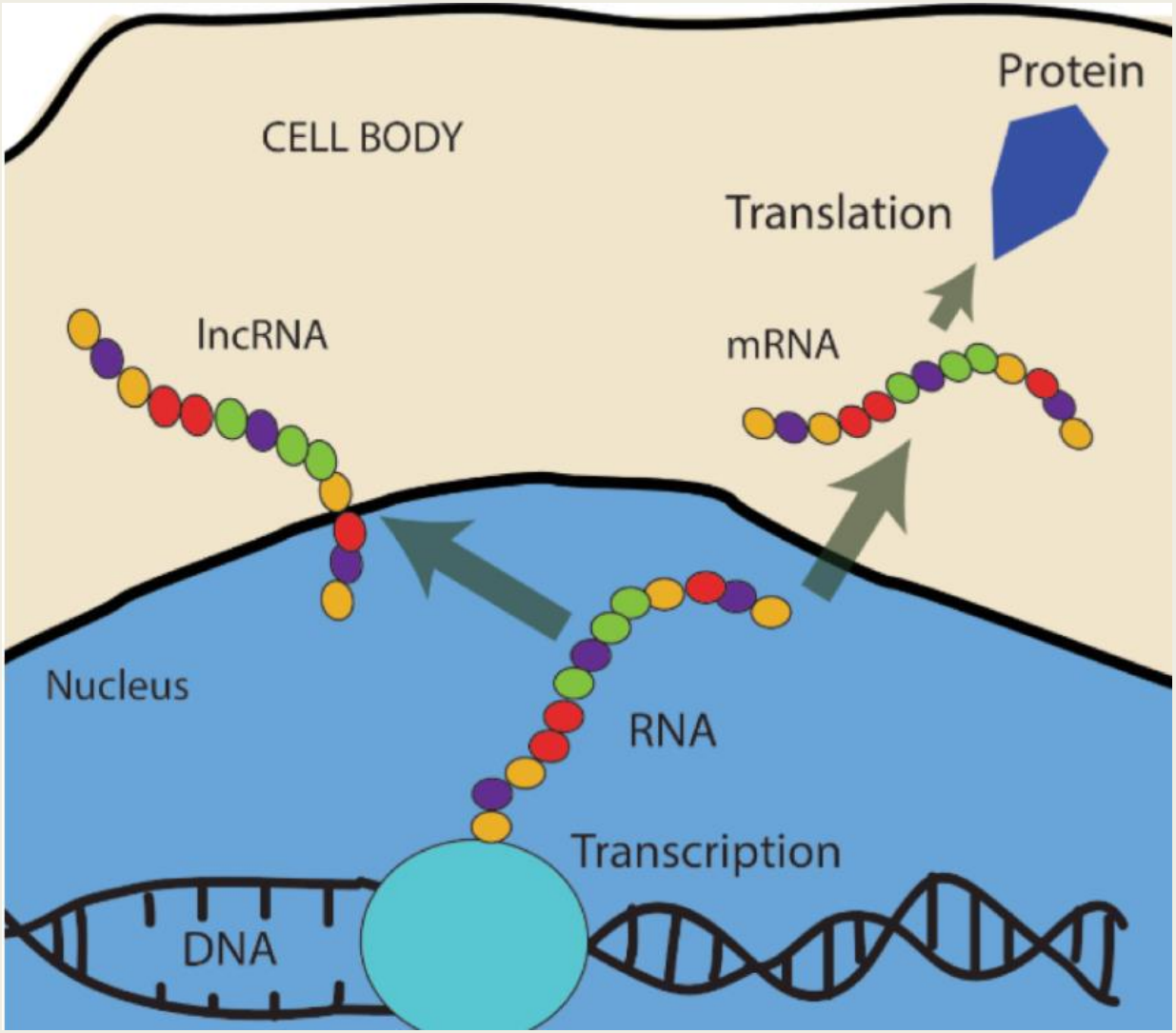
Где находится ДНК?



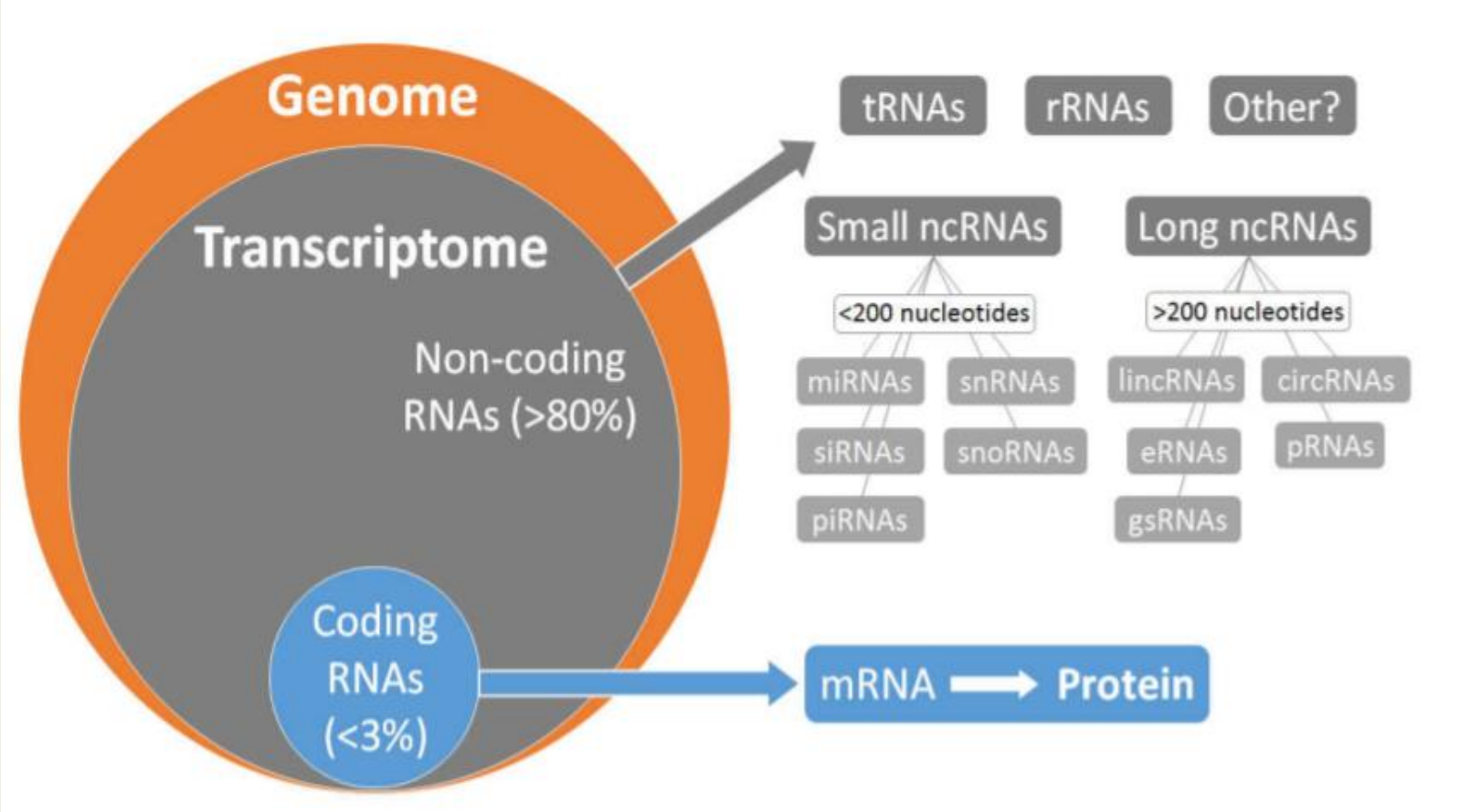
Typical prokaryotic (left) and eukaryotic (right) cells

https://sitn.hms.harvard.edu/flash/2014/long-noncoding-rnas/

Откуда берется РНК?



Какие бывают РНК?



https://pubmed.ncbi.nlm.nih.gov/30934986/

GENCODE
разметка генов человека

https://www.gencodegenes.org/human/

Total No of Genes	61533
Protein-coding genes	19982
Long non-coding RNA genes	18811
Small non-coding RNA genes	7567
Pseudogenes	14763
- processed pseudogenes	10662
- unprocessed pseudogenes	3557
- unitary pseudogenes	243
- polymorphic pseudogenes	50
- pseudogenes	15
Immunoglobulin/T-cell receptor gene segments	
- protein coding segments	409
- pseudogenes	236

Количество белок-кодирующих генов у разных видов

- Картофель – 39 000
- Человек ~ 20 000
- Черви – 14 000
- Мухи – 12 000
- Грибы – 6 000
- Бактерии – 2 000 – 4 000
- Микоплазмы - 500
- Вирус гриппа – 12

Некодирующие РНК

https://pubmed.ncbi.nlm.nih.gov/30635897/

Noncoding RNAs Databases: Current Status and Trends

Vinicius Maracaja-Coutinho, Alexandre Rossi Paschoal, José Carlos Caris-Maldonado, Pedro Vinícius Borges, Almir José Ferreira, and Alan Mitchell Durham

Abstract

One of the most important resources for researchers of noncoding RNAs is the information available in public databases spread over the internet. However, the effective exploration of this data can represent a daunting task, given the large amount of databases available and the variety of stored data. This chapter describes a classification of databases based on information source, type of RNA, source organisms, data formats, and the mechanisms for information retrieval, detailing the relevance of each of these classifications and its usability by researchers. This classification is used to update a 2012 review, indexing now more than 229 public databases. This review will include an assessment of the new trends for ncRNA research based on the information that is being offered by the databases. Additionally, we will expand the previous analysis focusing on the usability and application of these databases in pathogen and disease research. Finally, this chapter will analyze how currently available database schemas can help the development of new and improved web resources.

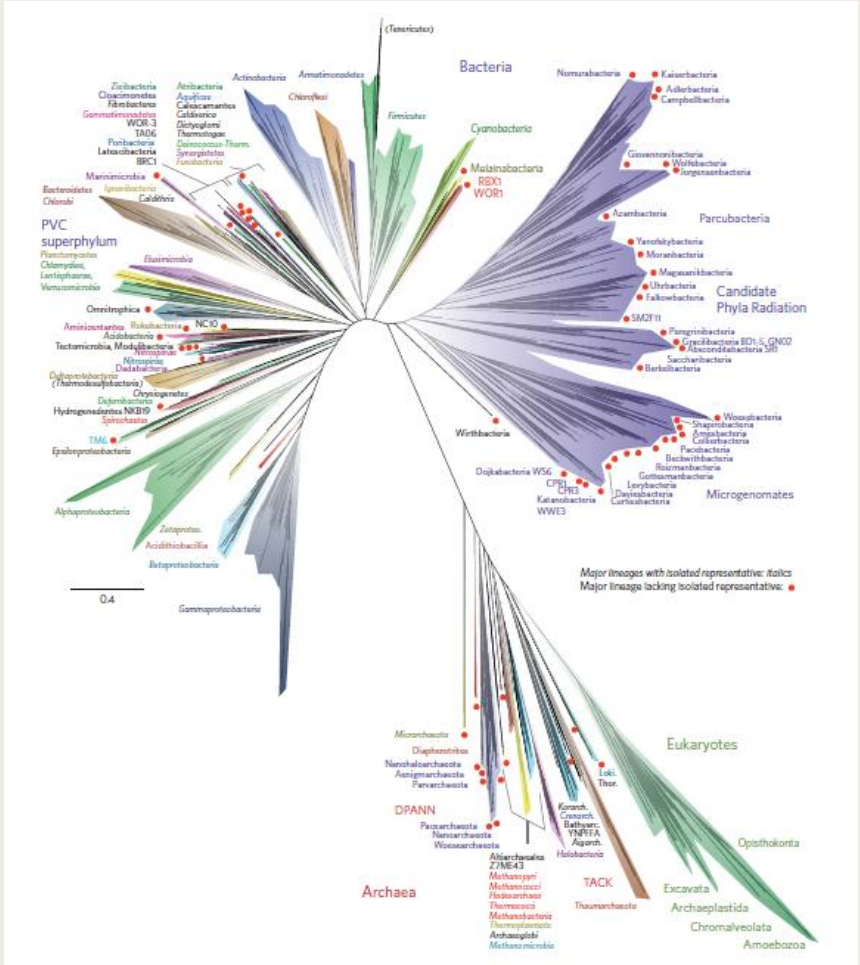
https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/

Геномы организмов в NCBI

- Eukaryotes (21680)
- Prokaryotes (390644)
- Viruses (47571)
- Plasmids (36213)
- Organelles (21972)

Задачи:

- Сборка геномов
- Сравнительная геномика



https://www.nature.com/articles/nmicrobiol201648.pdf

Число хромосом у разных видов

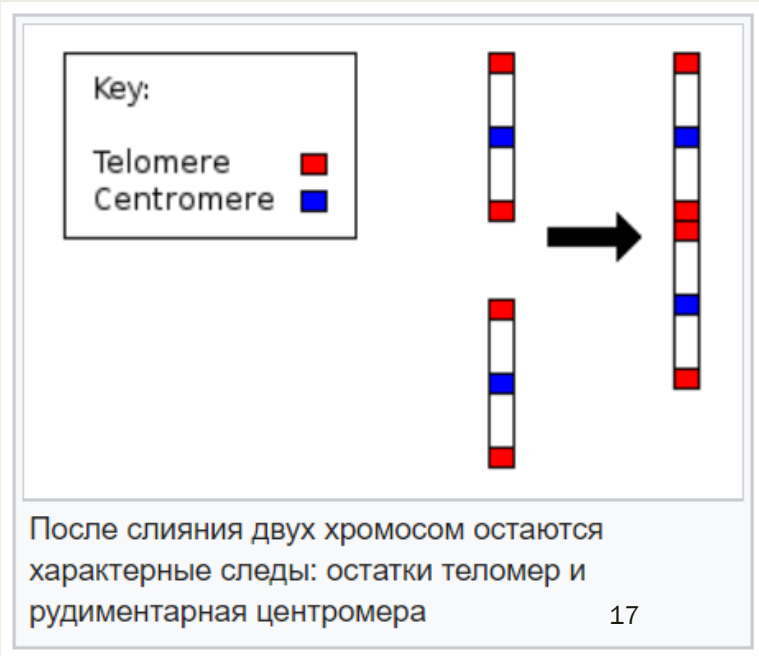
Гиббоны - 44

48

46

Макака - 42

Капуцин - 54

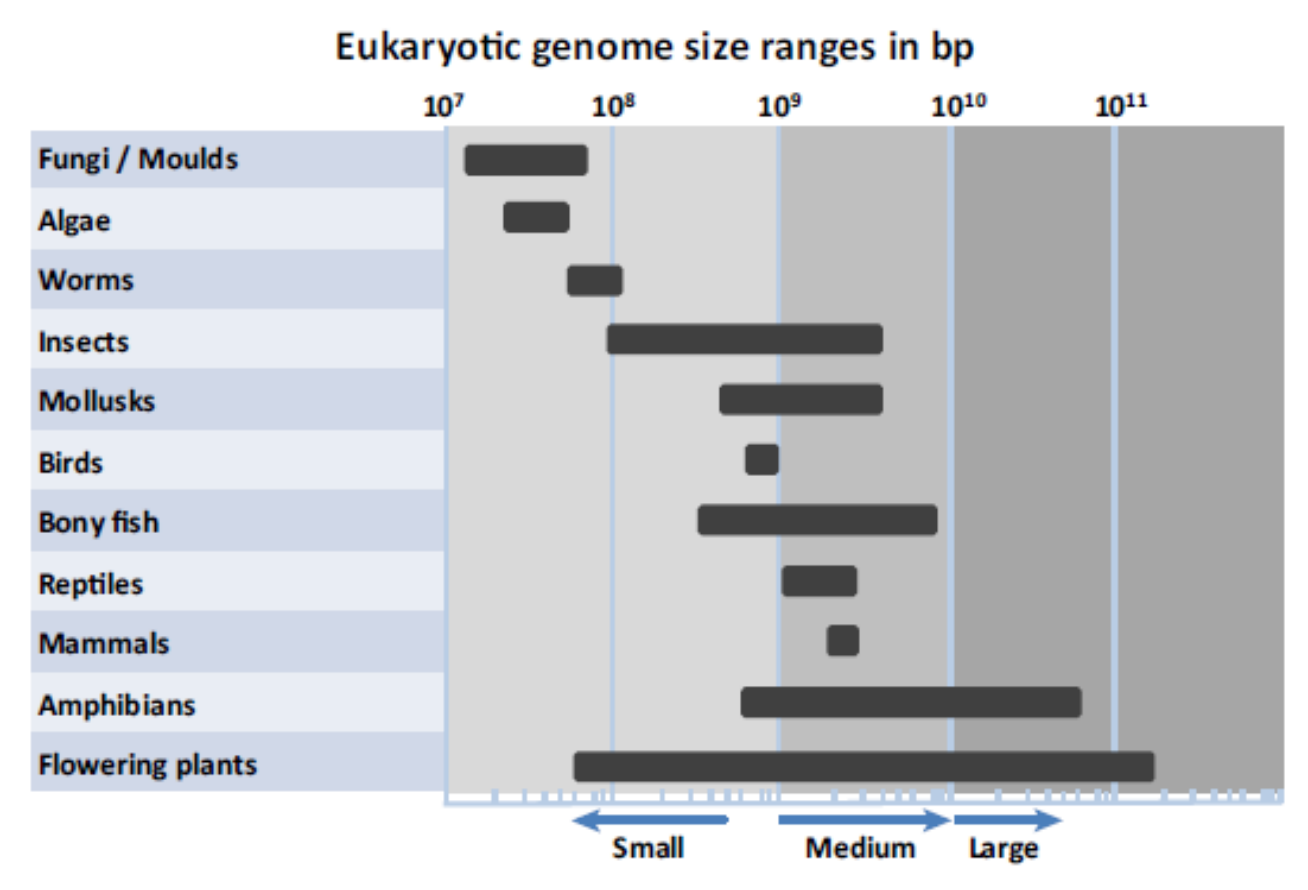


После слияния двух хромосом остаются характерные следы: остатки теломер и рудиментарная центромера

Число хромосом у разных видов

Муравей (Myrmecia pilosula) – 2
Плодовая мушка - 8
Арабидопсис – 10
Голубь – 16
Кошка – 38
Лиса - 34
Мышь - 40
Собака – 78
Утка – 80
Сазан - 104
Корова – 120
Рак (Cambarus clarkii) – 200
Хвощ – 216
Краб - 254
Бабочка – 380

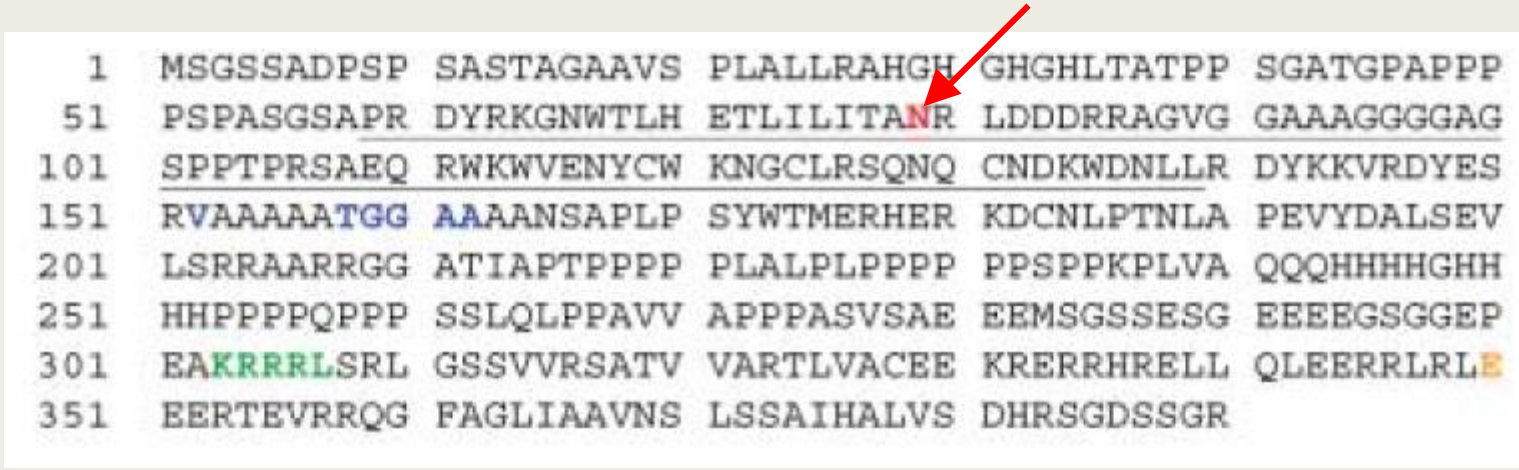
Размер генома



Размер генома человека ~ 3 млрд п.н.

База про размеры геномов: https://www.genomesize.com/statistics.php

Доместикация риса



Дикий рис – AAG – лизин

Культурный рис – AAT – аспарагин

DNA-seq

- Изучение отличий образца от референсного генома
- Можно секвенировать
  - *Геном*
  - *Экзоны мРНК => экзом (< 2%)*
  - *Конкретные области или гены*

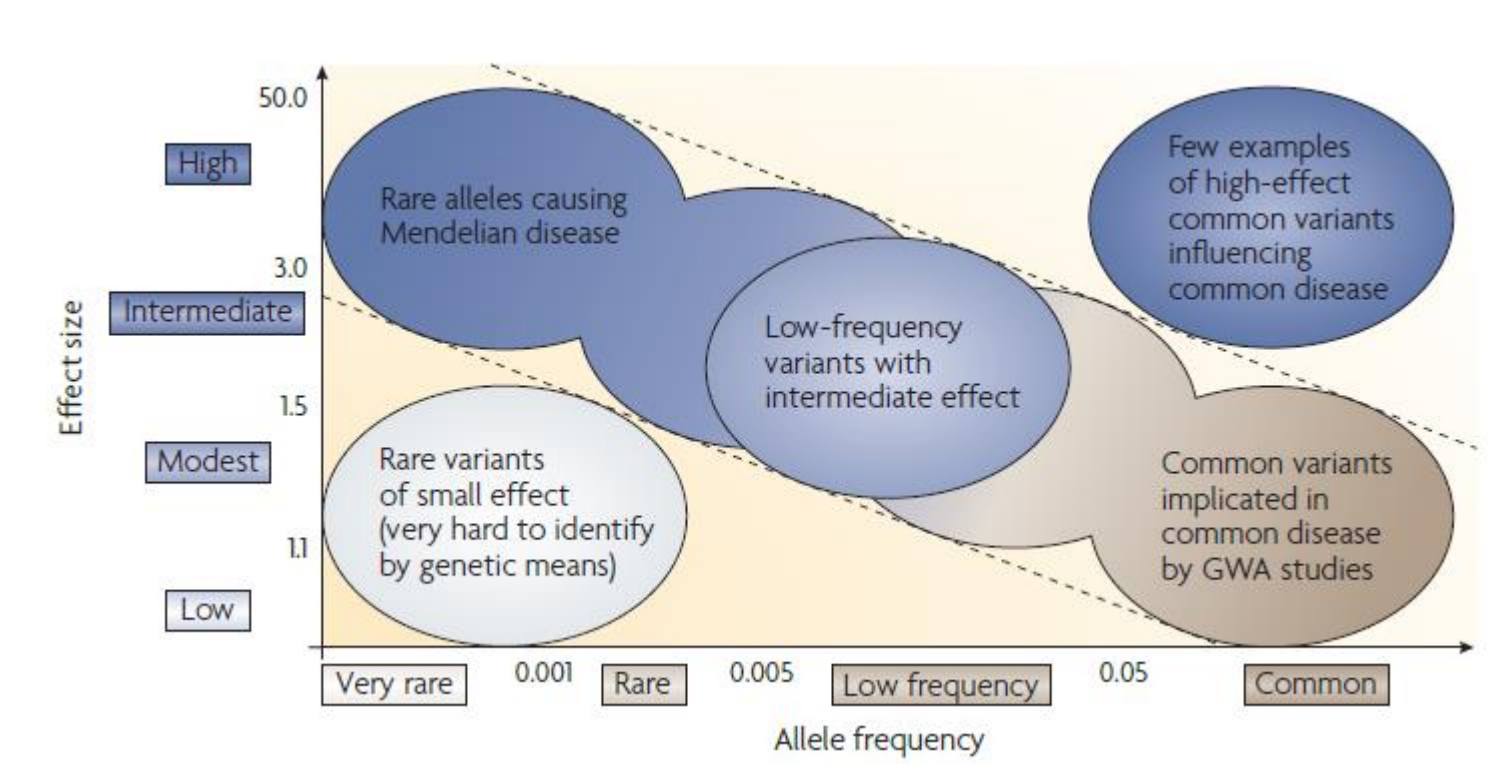
Коллекции

- 1000 genomes project
- The genome aggregation database (gnomAD)
- 100000 genomes
- UK Biobank
- …

Зачем?

- Геногеография – изучает географическое распространение генетических признаков живых организмов
- Медицинская геномика – изучает влияние геномных нарушений на развитие наследственных заболеваний
- Фармакогенетика – изучает генетические особенности человека, влияющие на индивидуальную переносимость лекарственных средств
- …

Генетика и заболевания

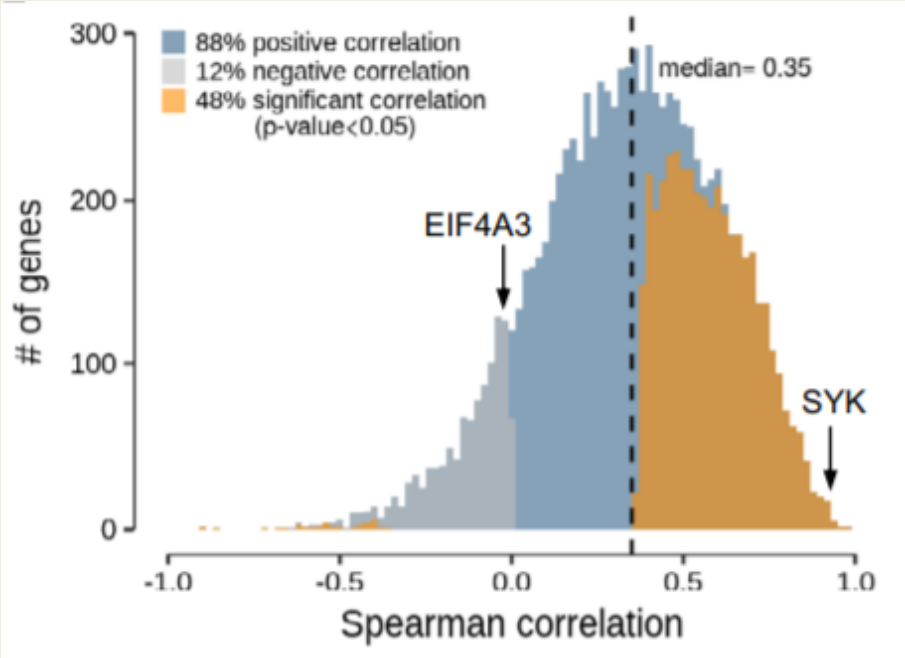


https://www.nature.com/articles/nrg2793

RNA-seq - Задачи

- Экспрессия генов
- Дифференциальная экспрессия генов
- Сплайсинг
- Редактирование РНК
- Сборка транскриптома

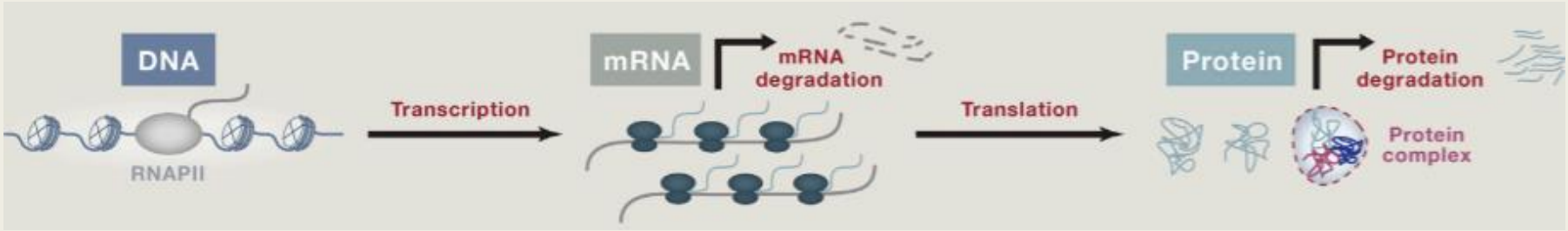
мРНК



Транскрипционный профиль белок-кодирующих генов

Можем косвенно судить о концентрации белков

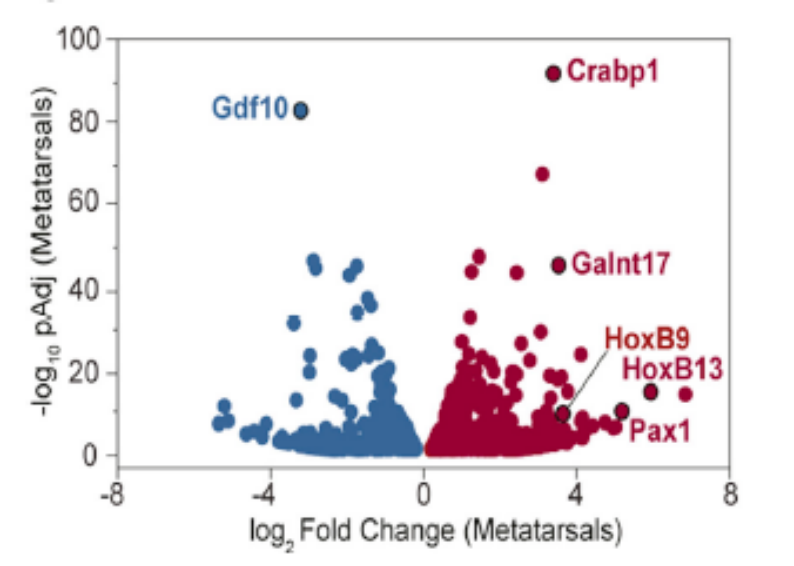
[A deep proteome and transcriptome abundance atlas of 29 healthy human tissues]

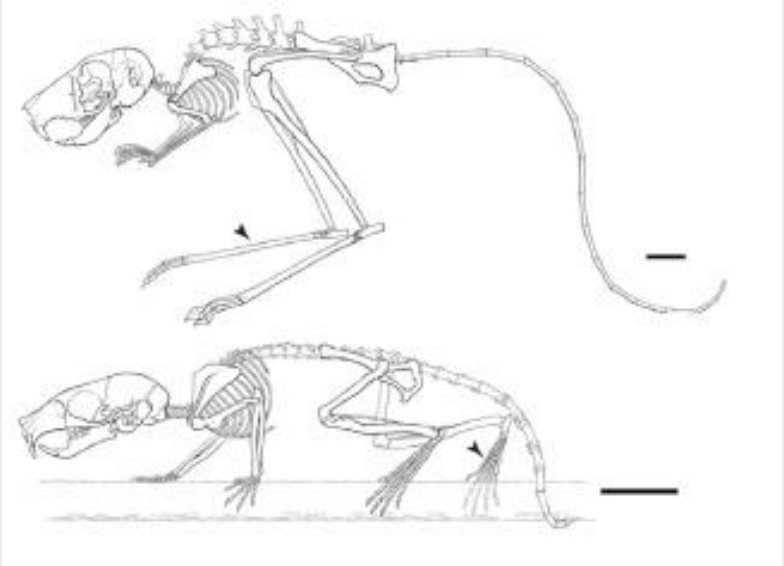


[On the Dependency of Cellular Protein Levels on mRNA Abundance]

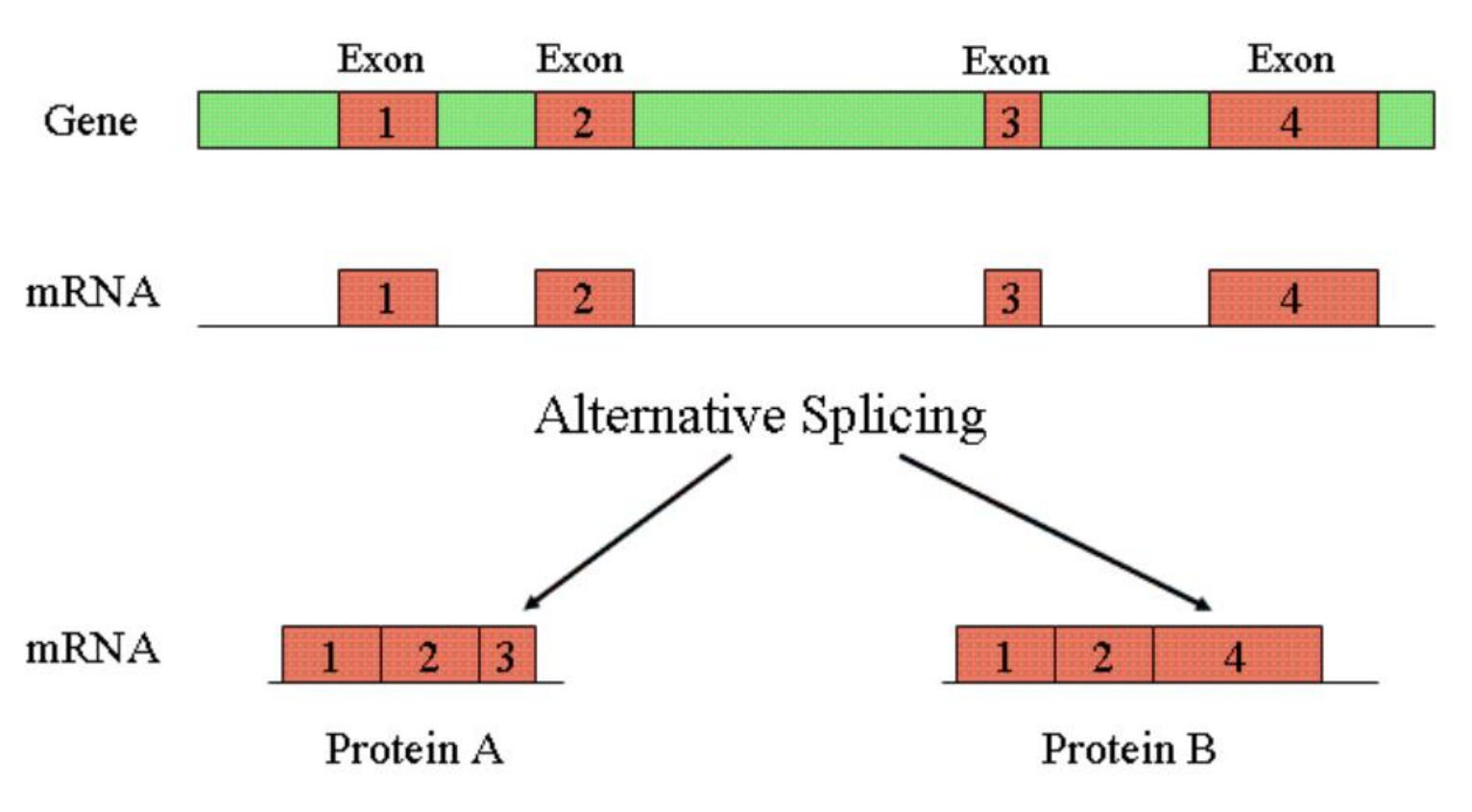
Дифференциальная экспрессия

Interspecies transcriptomics identify genes that underlie disproportionate foot growth in jerboas



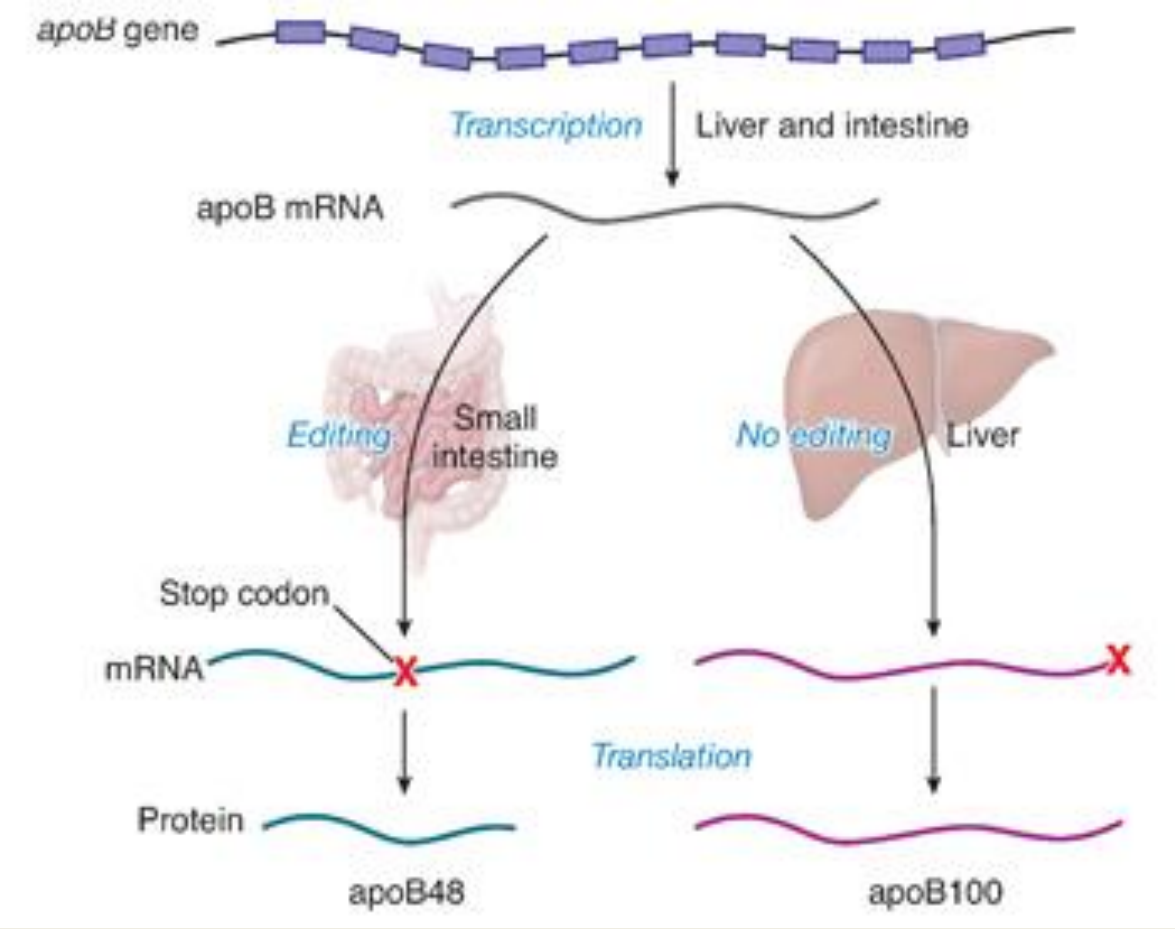


Альтернативный сплайсинг



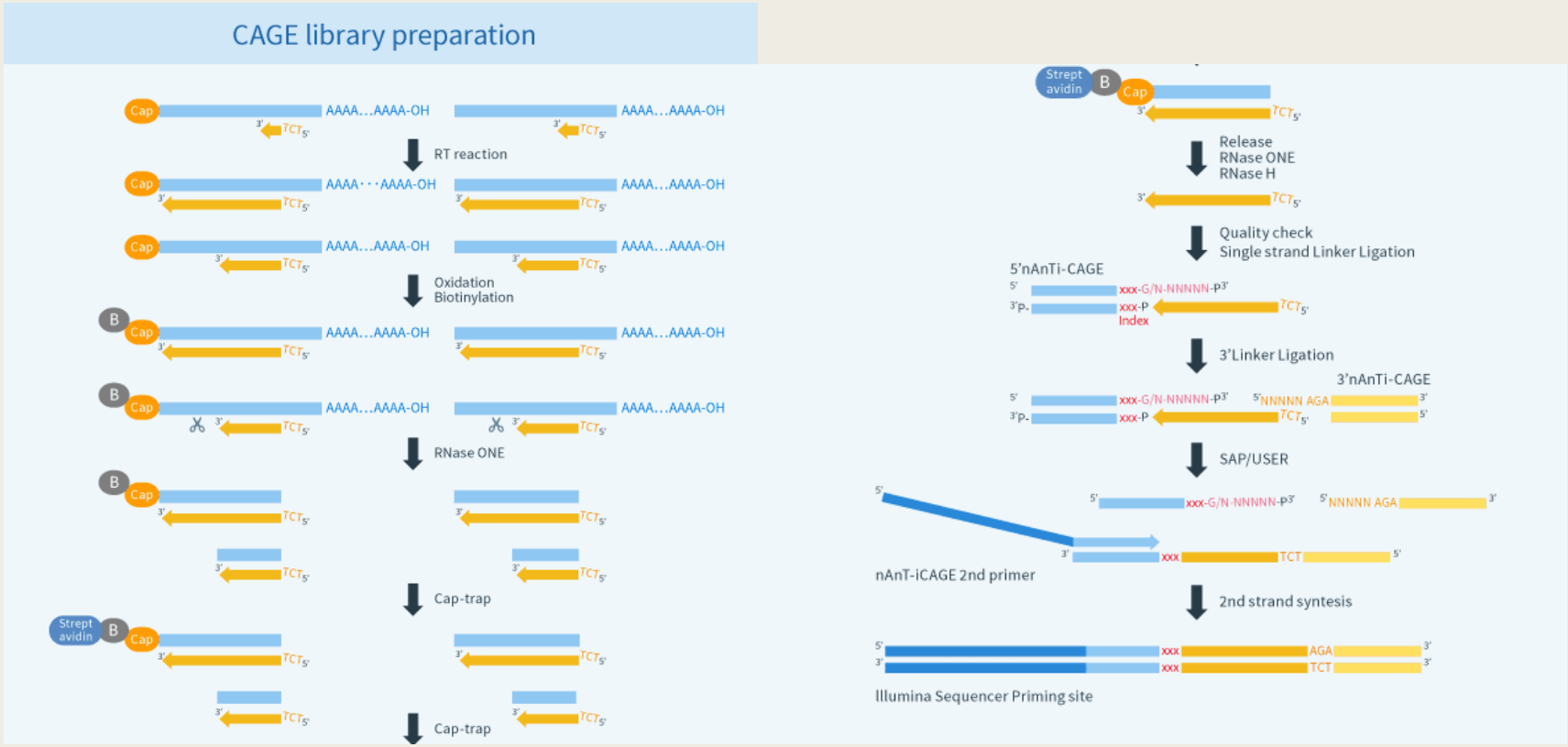
https://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/alternative_splicing.html

Редактирование РНК

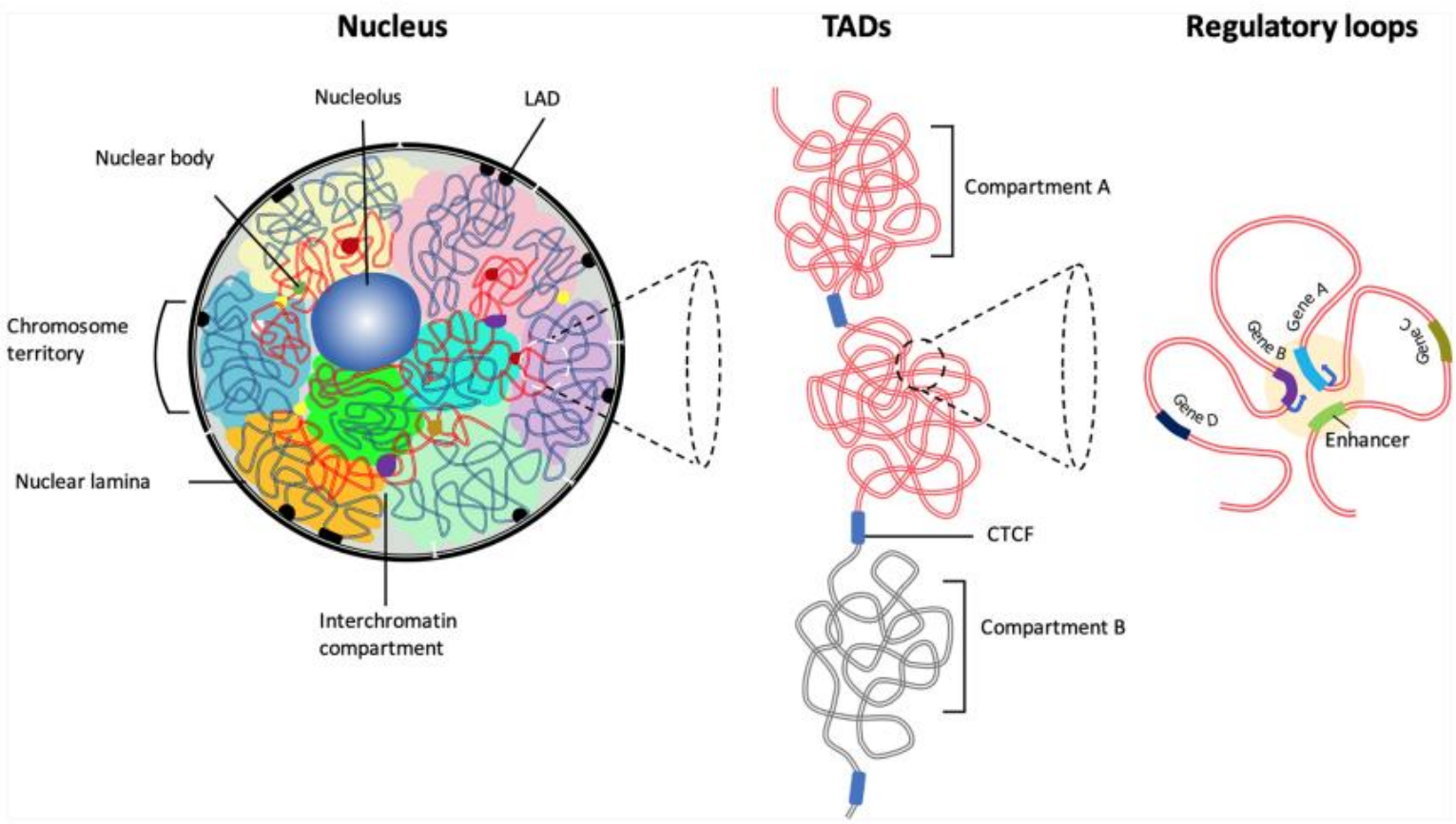


https://basicmedicalkey.com/pharmacology-of-cholesterol-and-lipoprotein-metabolism/

CAGE
уточнение старта транскрипции

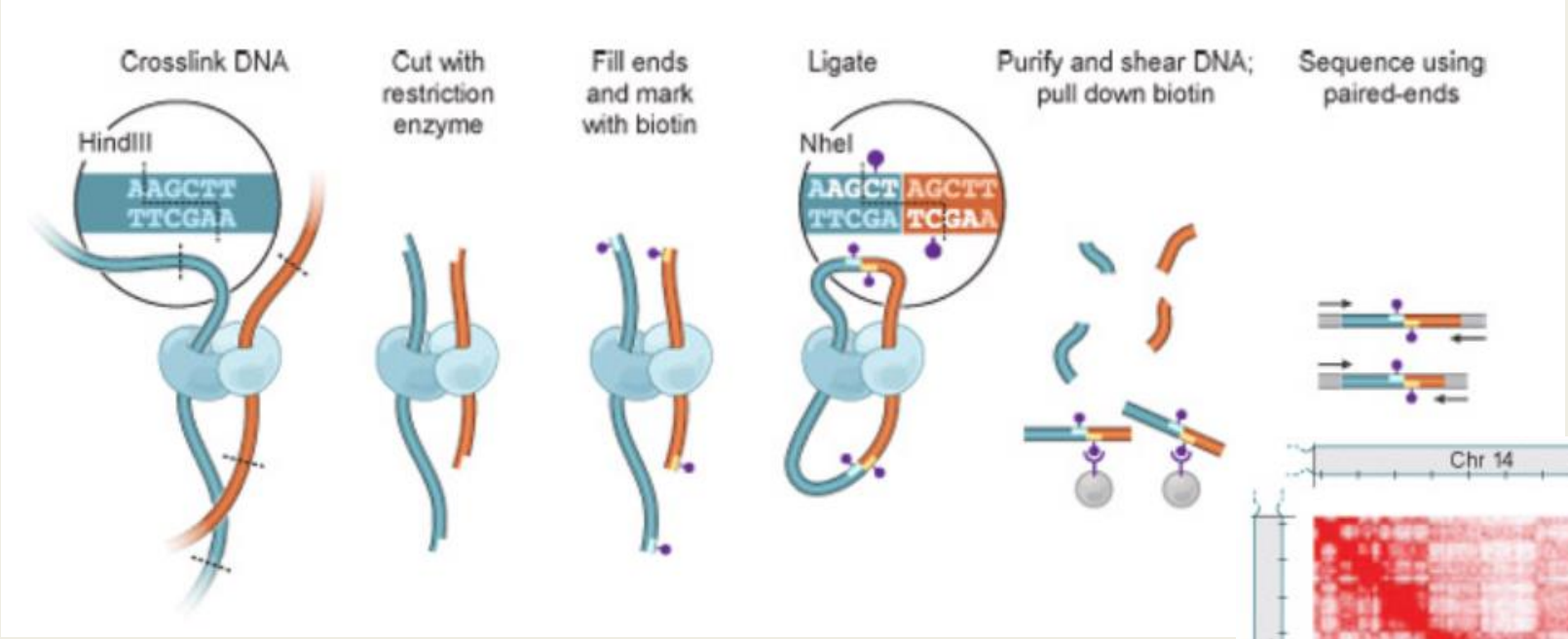


Укладка ДНК в ядре



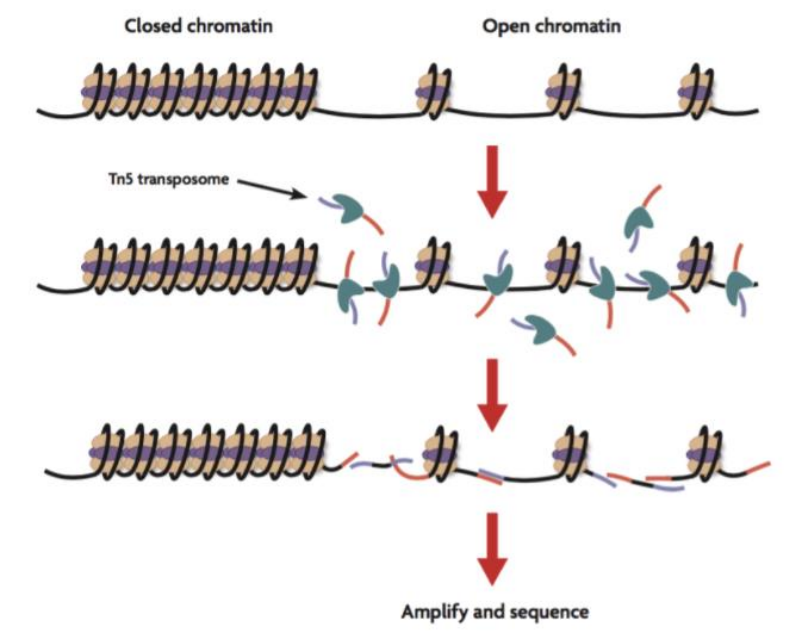
https://pubmed.ncbi.nlm.nih.gov/33854505/

Hi-C - ДНК+ДНК
определение 3D-структуры хроматина



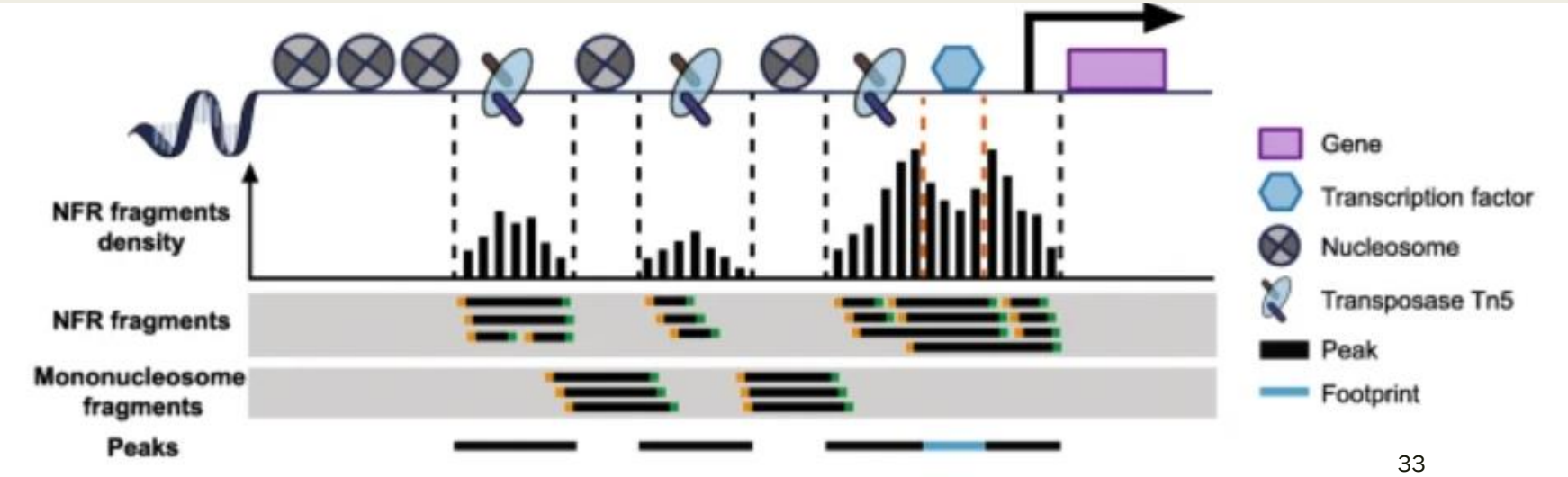
https://pubmed.ncbi.nlm.nih.gov/20461051/

ATAC-seq - доступный хроматин

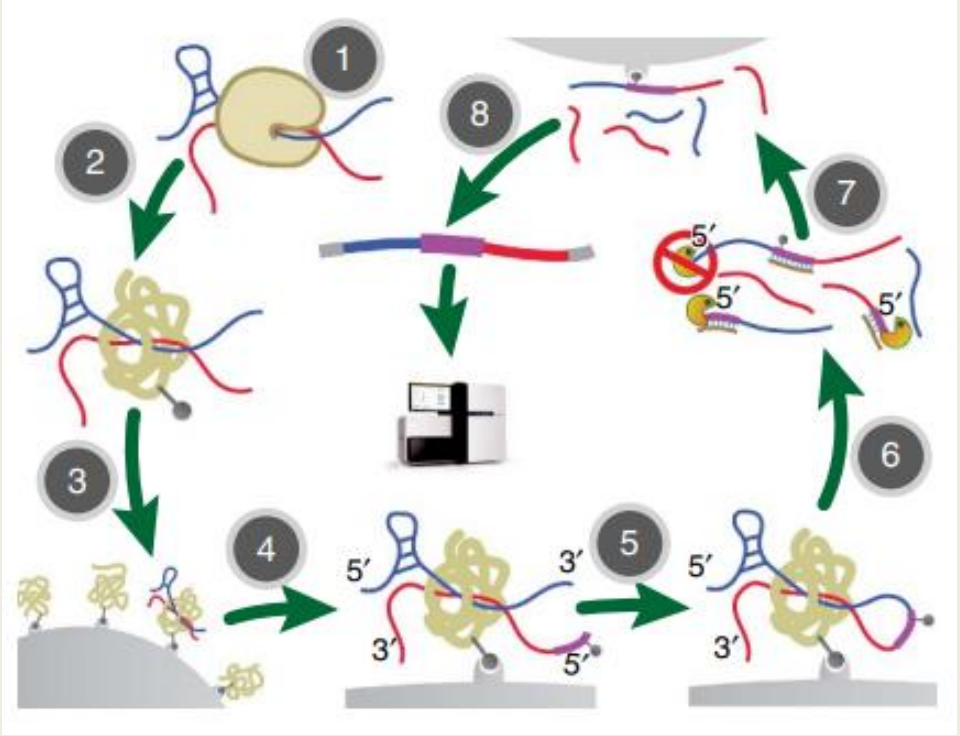


https://www.activemotif.com/blog-atac-seq

Высокоактивная транспозаза Tn5 вносит двуцепочечные разрывы в открытые участки хроматина и вставляет в области разрывов адаптеры для секвенирования



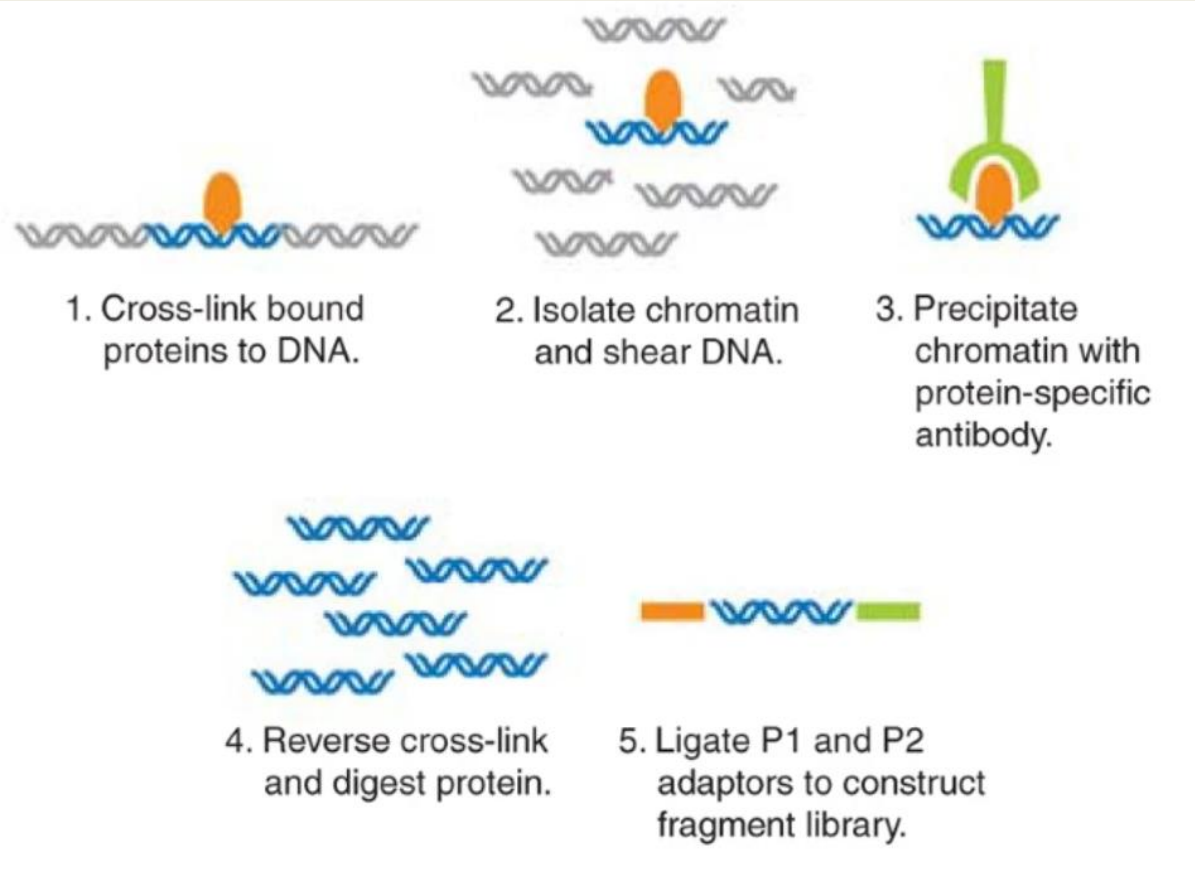
MARIO – PHK+PHK



(1) cross-linking RNAs to proteins;
(2) RNA fragmentation, protein denaturing and biotinylation;
(3) immobilization of RNA-binding proteins at low density;
(4) ligation of a biotinylated RNA linker;
(5) proximity ligation under a dilute condition;
(6) RNA purification and RT;
(7) biotin pull-down;
(8) construction of sequencing library

https://www.nature.com/articles/ncomms12023.pdf

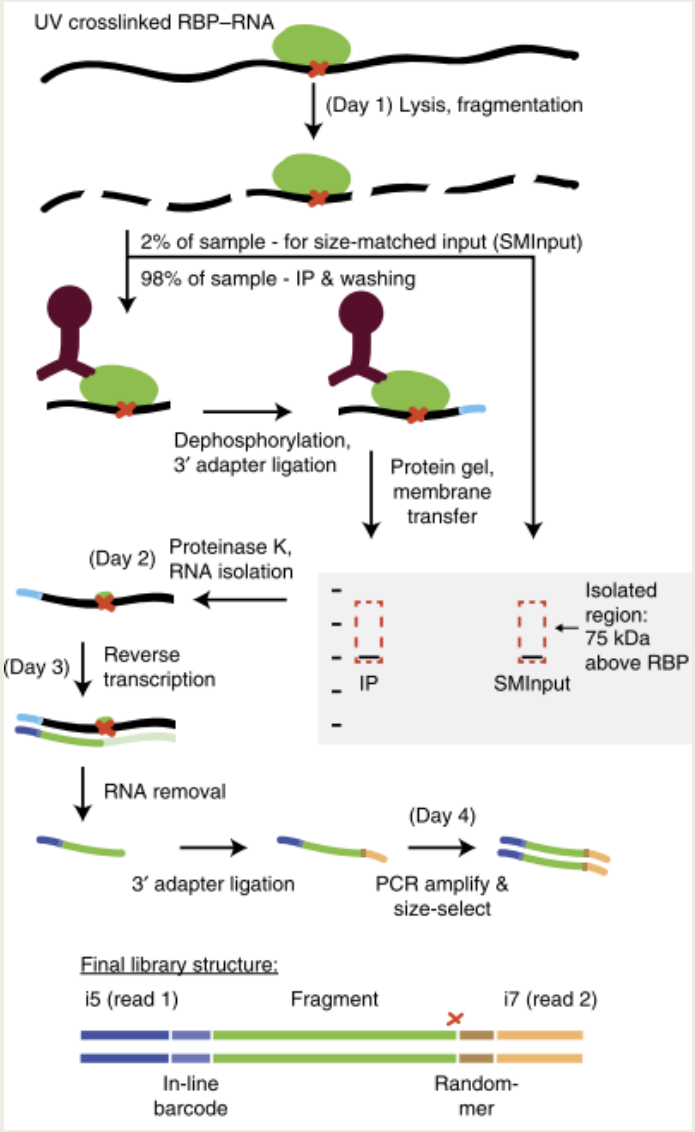
Chip-seq – ДНК+белок



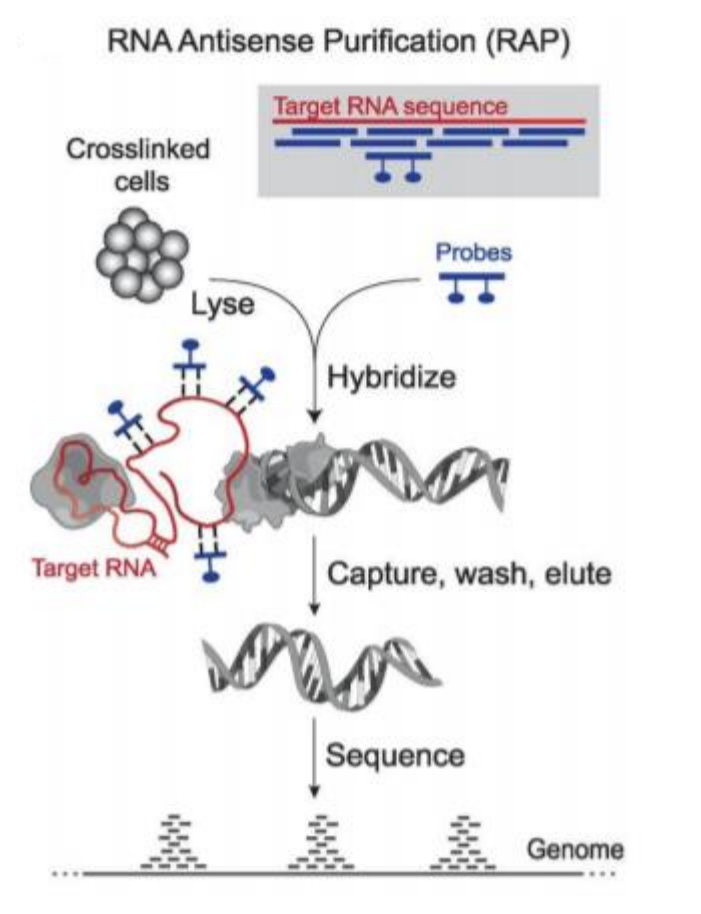
Антитела специфичны к одному белку => за один эксперимент устанавливаем сайты связывания по всему геному, но для одного белка

https://www.nature.com/articles/nmeth.f.247

eClip – РНК+белок

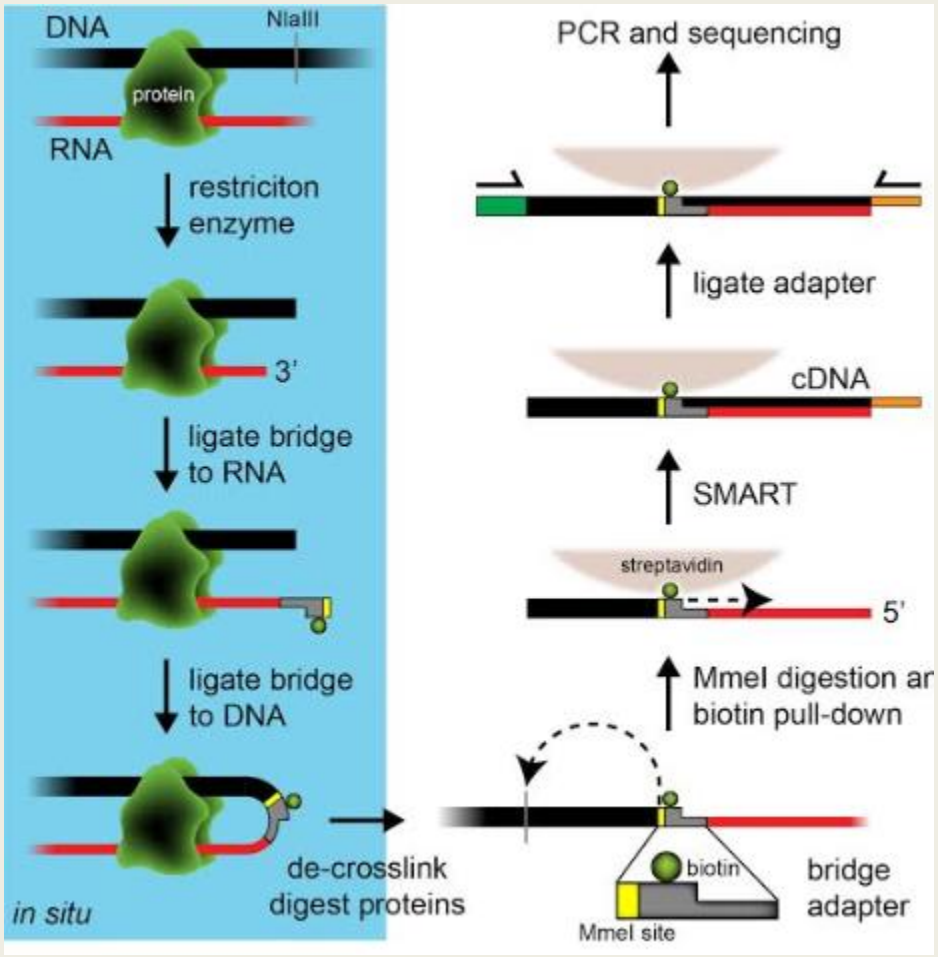


RAP-seq – РНК+ДНК



Для одной заранее (!) известной РНК устанавливаем локусы взаимодействия с хроматином

Red-C – РНК+ДНК

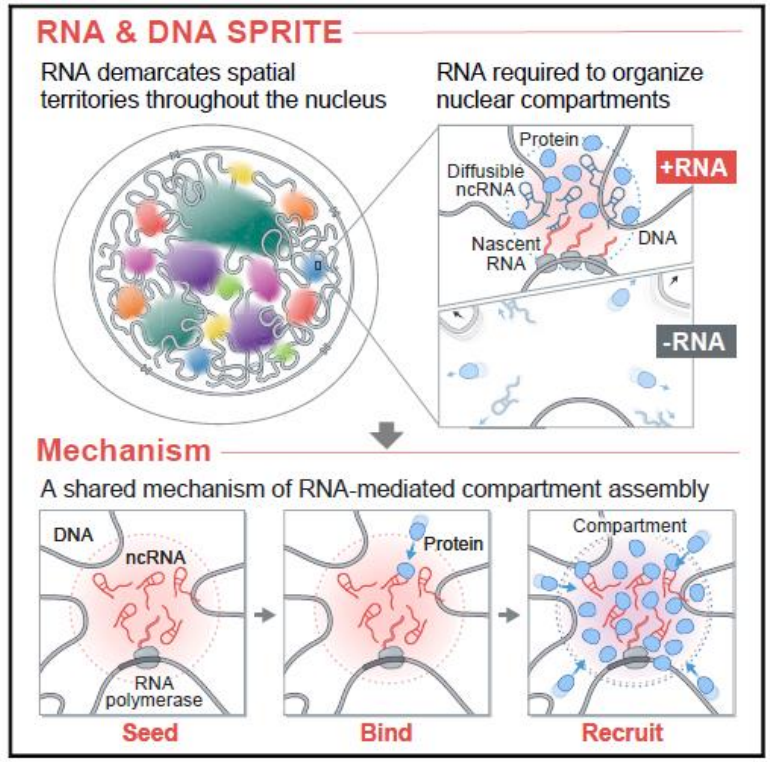


Для всех РНК, независимо от типа и длины, устанавливаем все локусы взаимодействия с хроматином

Есть другие методы: GRID-seq, RADICL-seq, ...

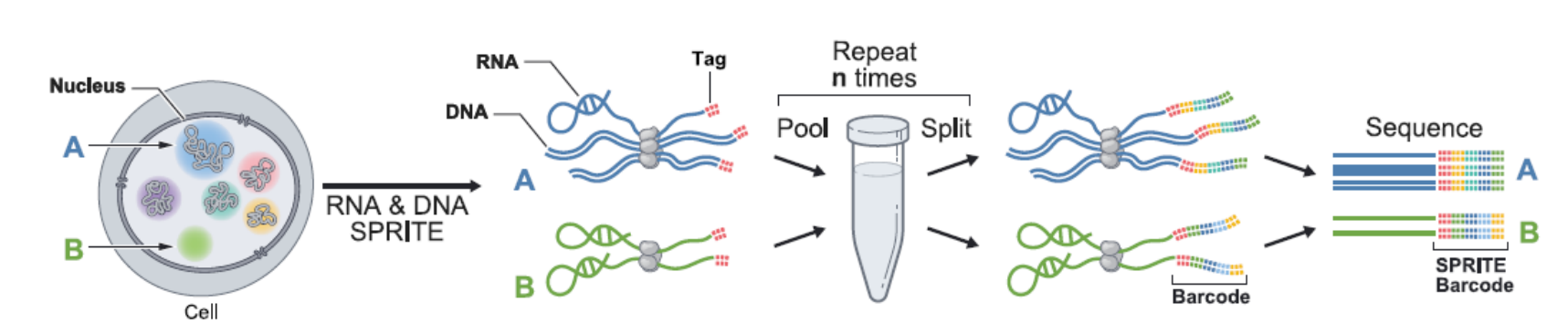
https://pubmed.ncbi.nlm.nih.gov/32479626/

RD-SPRITE – РНК+ДНК



Можно детектировать крупные макромолекулярные комплексы, ядерные структуры

Нет стадии лигирования близко расположенных молекул



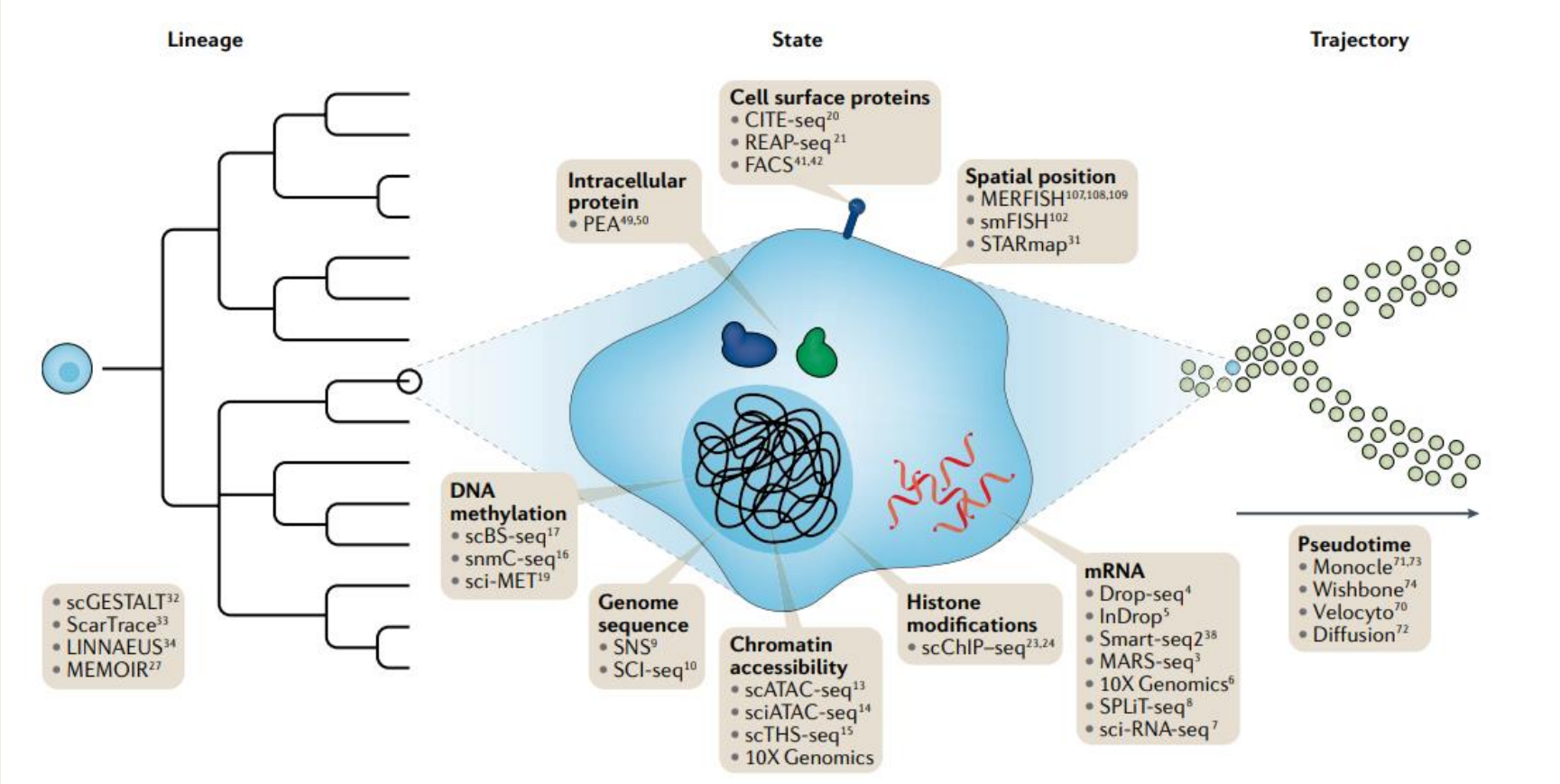
https://pubmed.ncbi.nlm.nih.gov/34739832/

enSEQlopedia

- http://enseqlopedia.com/enseqlopedia/
- Описание большого количества протоколов
- Есть далеко не все

https://www.nature.com/articles/s41576-019-0093-7.pdf

А еще есть single cell методы



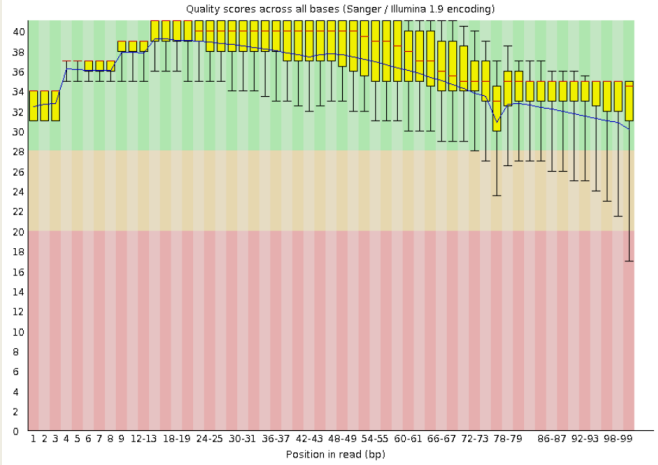
Важно

- У всех методов есть свои ограничения, особенности, «подводные камни», источники «шума», вариации протокола
- Необходимо понимать природу данных (организм, ткань, клеточная линия, процесс выделения и обработки)
- Разобраться во всех этапах пробоподготовки и контролях

Как обрабатывать данные?

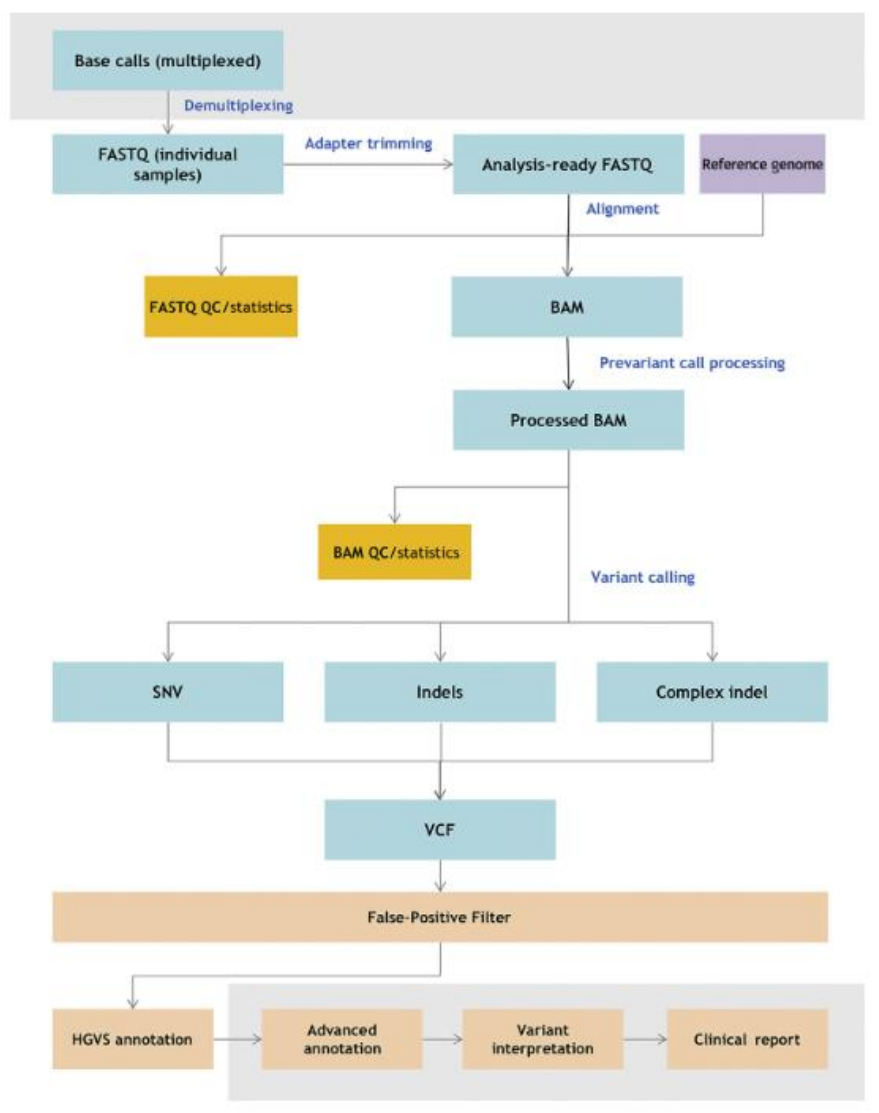
- Говорим только о Illumina
- В любом случае получаем чтения или «риды»
- Для каждого типа секвенирования разрабатывают свой подход биоинформатической обработки
- Есть общие шаги в обработке чтений

Всегда нужно



- Проверить качество чтений
- Проверить наличие адаптеров или других технических последовательностей в чтениях
- *Картировать чтения на референсный геном или собрать референсный геном\транскриптом*
- *Проверить качество сборки или картирования*

Программный конвейер



Каждый блок – шаг обработки, отдельная программа

Не все эксперименты одинаковые

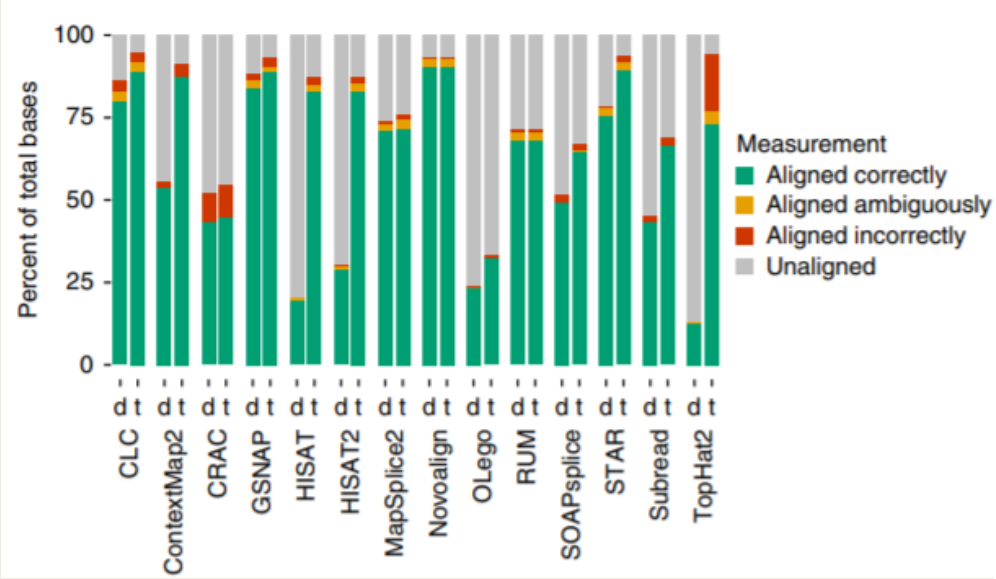
«Стандартный» протокол обработки может не подойти

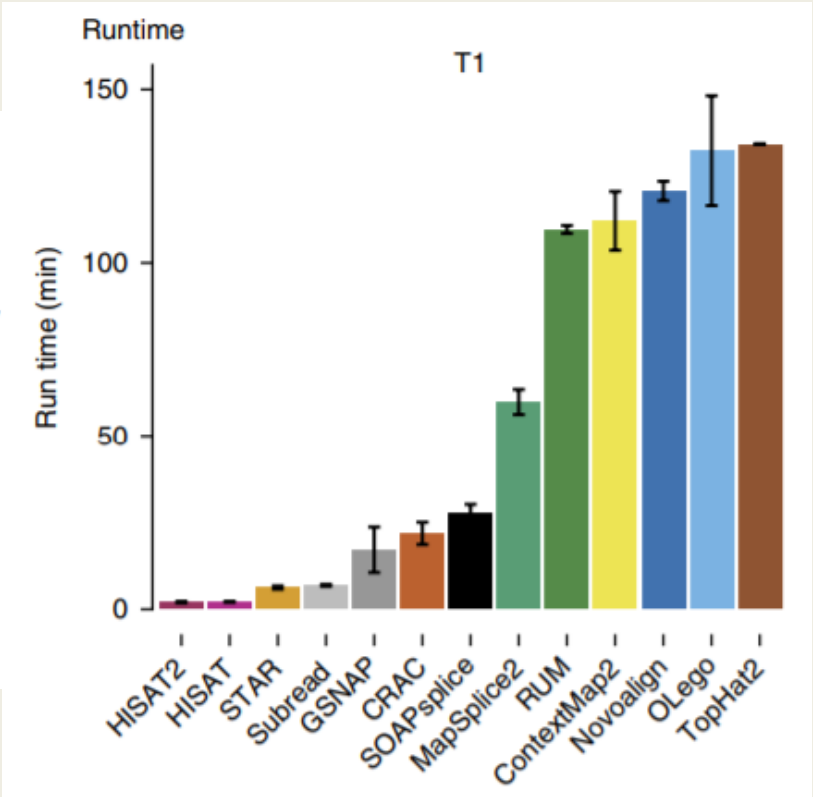
https://pubmed.ncbi.nlm.nih.gov/29154853/

Benchmarks

Практически для каждого шага программного конвейера можно нейти несколько похожих программ

Нужно быть в курсе актуального ПО





https://pubmed.ncbi.nlm.nih.gov/27941783