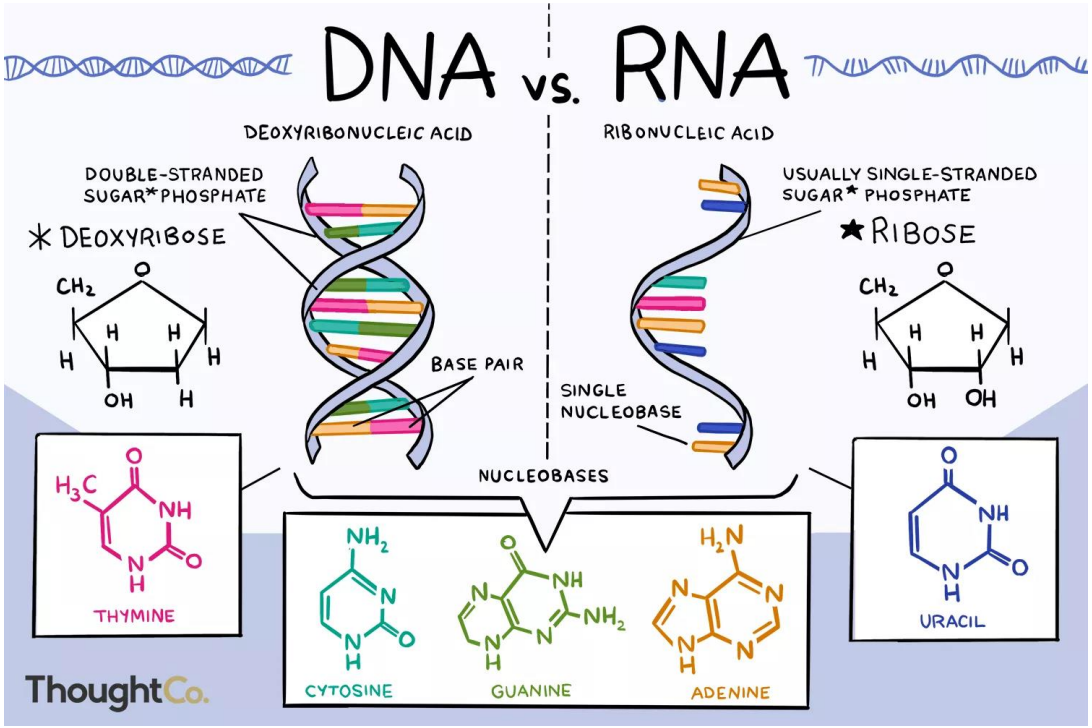
СЕКВЕНИРОВАНИЕ ОТ N ДО Y

Анастасия Жарикова
azharikova89@gmail.com
ФББ МГУ - 5 апреля 2023

СЕКВЕНИРОВАНИЕ

- Установление последовательности нуклеиновых кислот (с белками тоже можно, но иначе)



ref

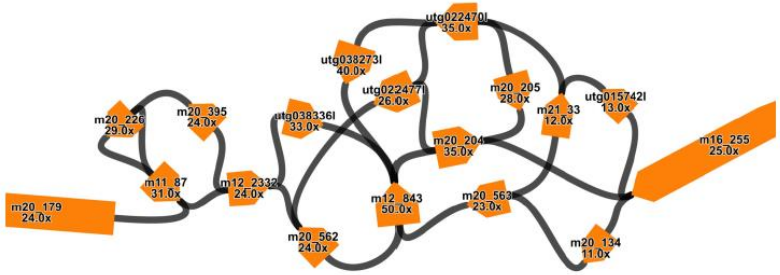
ЧТО КОНКРЕТНО

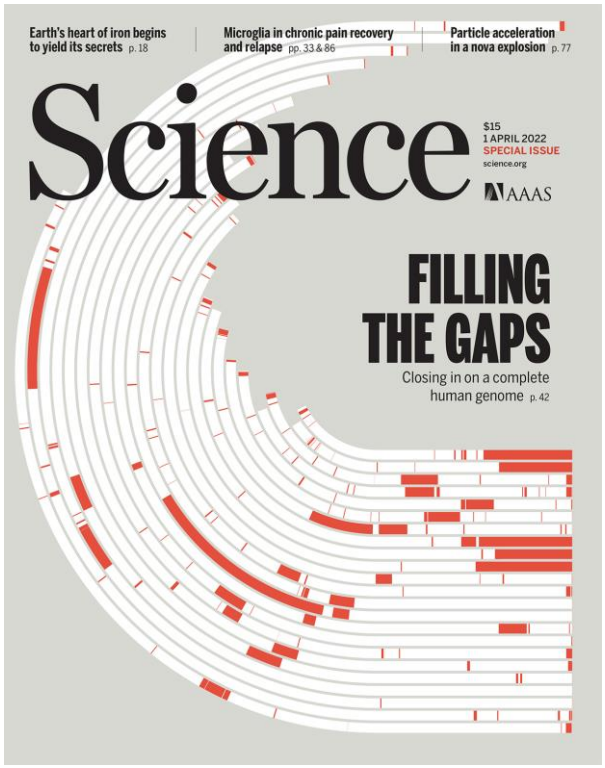
- ДНК: секвенировать весь геном

Откуда взять геном?

Нужно выделить ДНК

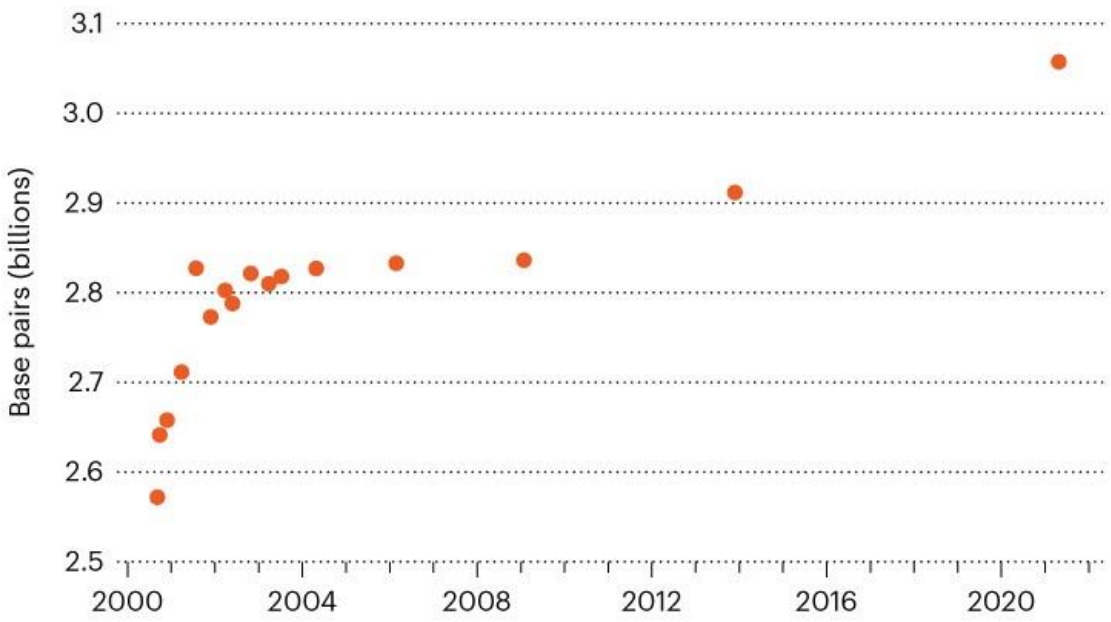
T2T





COMPLETING THE HUMAN GENOME

Researchers have been filling in incompletely sequenced parts of the human reference genome for 20 years, and have now almost finished it, with 3.05 billion DNA base pairs.



0.3% of sequence might still have errors. Includes X but not Y chromosome. Count excludes mitochondrial DNA.

©nature

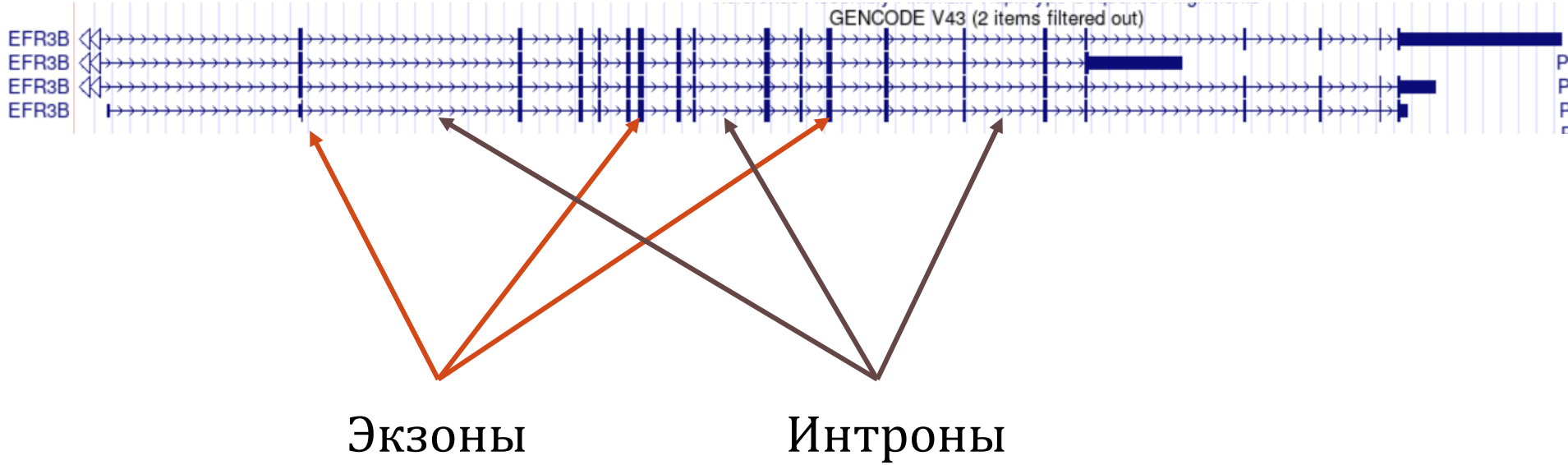
https://www.science.org/toc/science/376/6588

А ЕСЛИ НЕ ВЕСЬ ГЕНОМ?

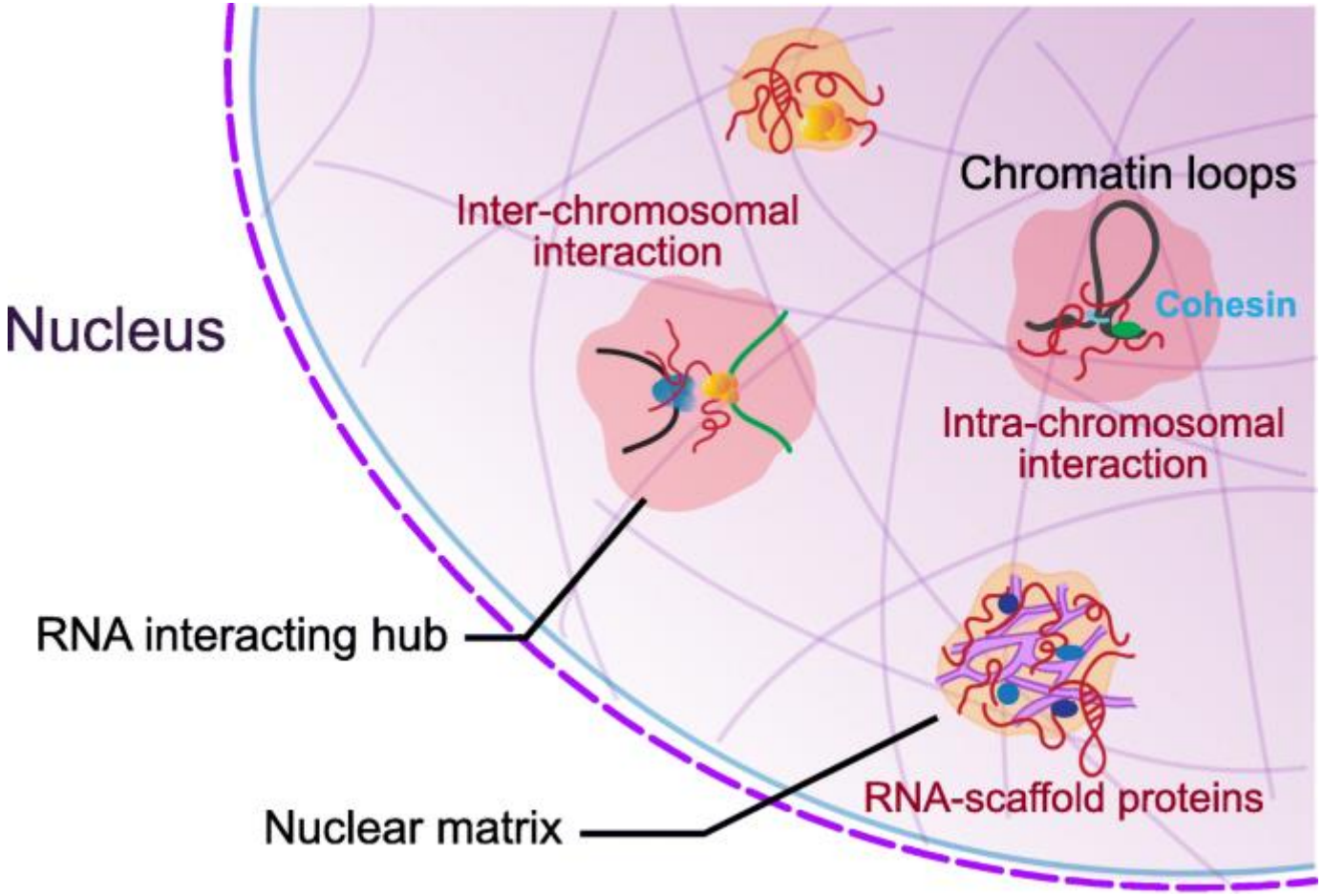
- Какие фрагменты ДНК есть смысл секвенировать?

ЭКЗОМ

- Экзоны белок-кодирующих генов
- ~1-2% от всего генома

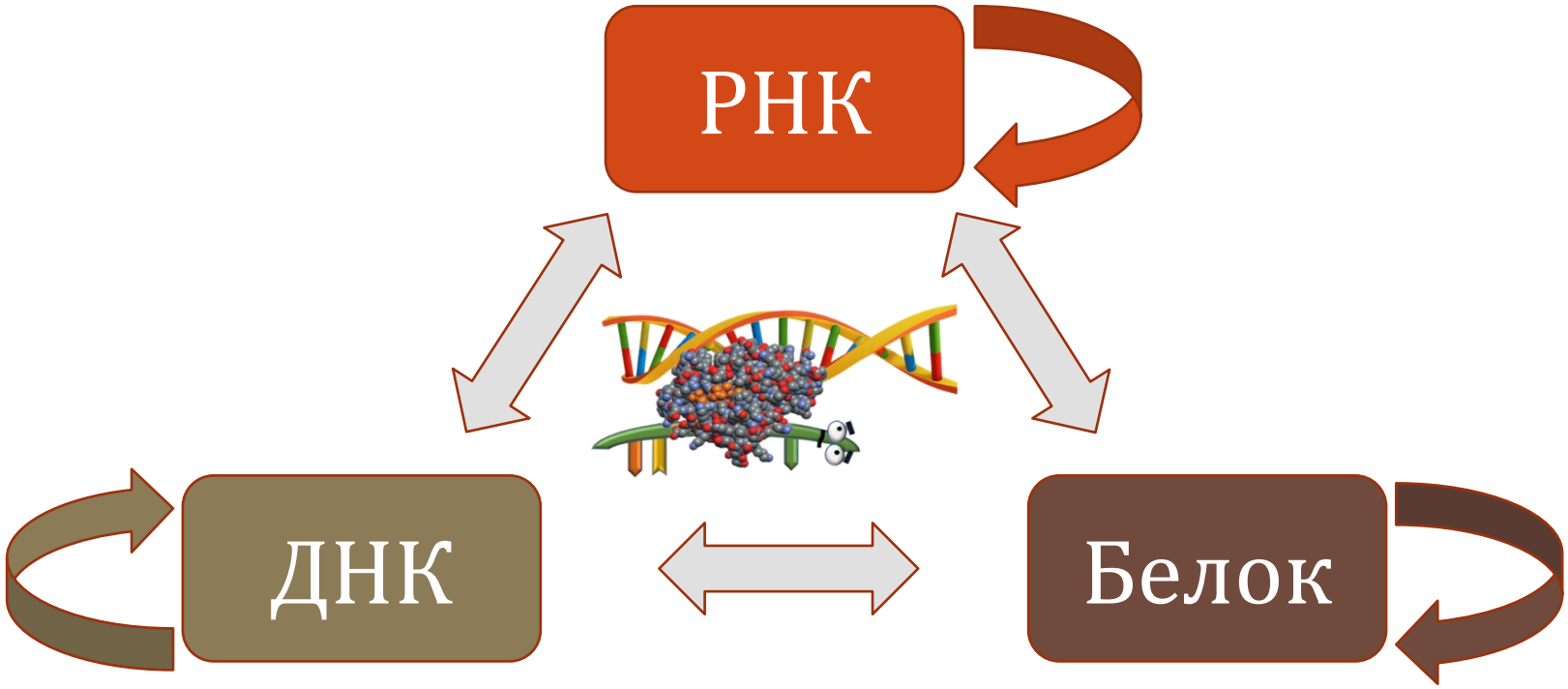


МАКРОМОЛЕКУЛЯРНЫЕ ВЗАИМОДЕЙСТВИЯ



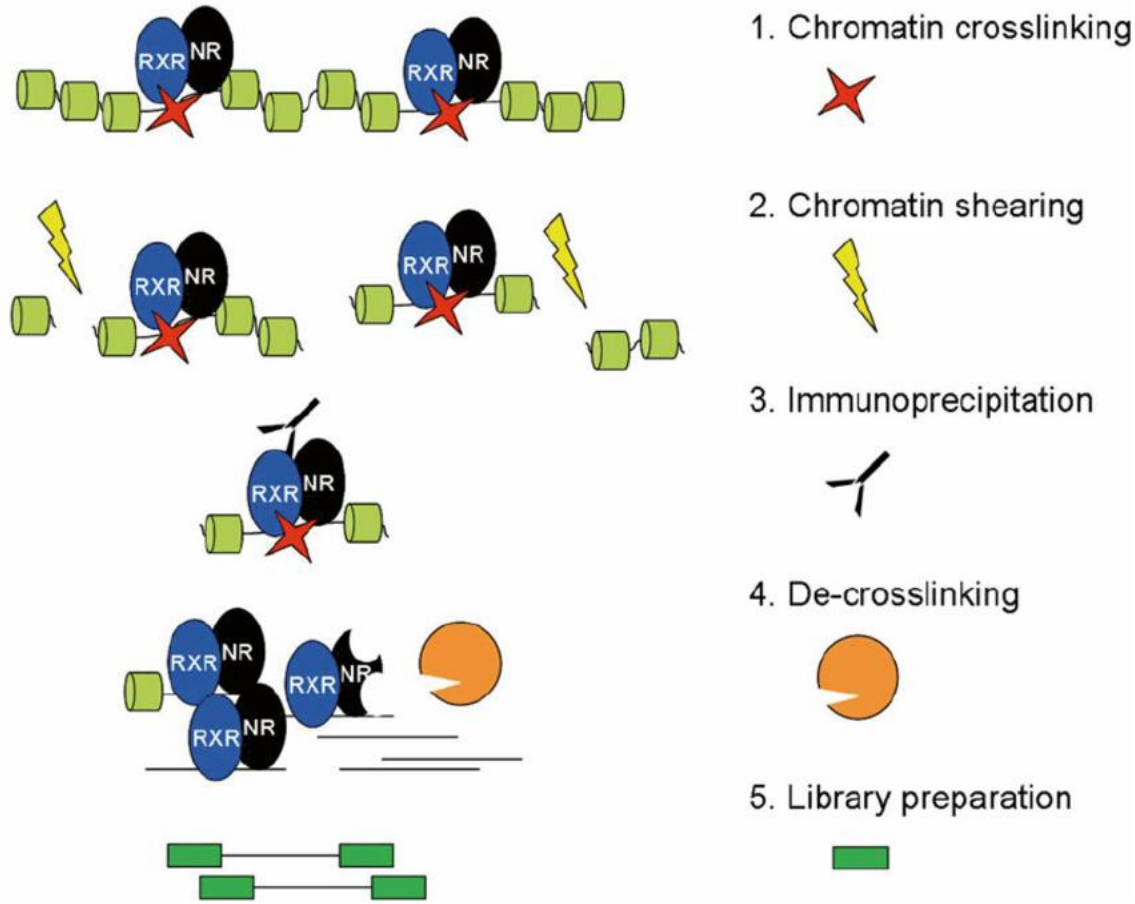
ref

МАКРОМОЛЕКУЛЯРНЫЕ ВЗАИМОДЕЙСТВИЯ



CHIP-SEQ

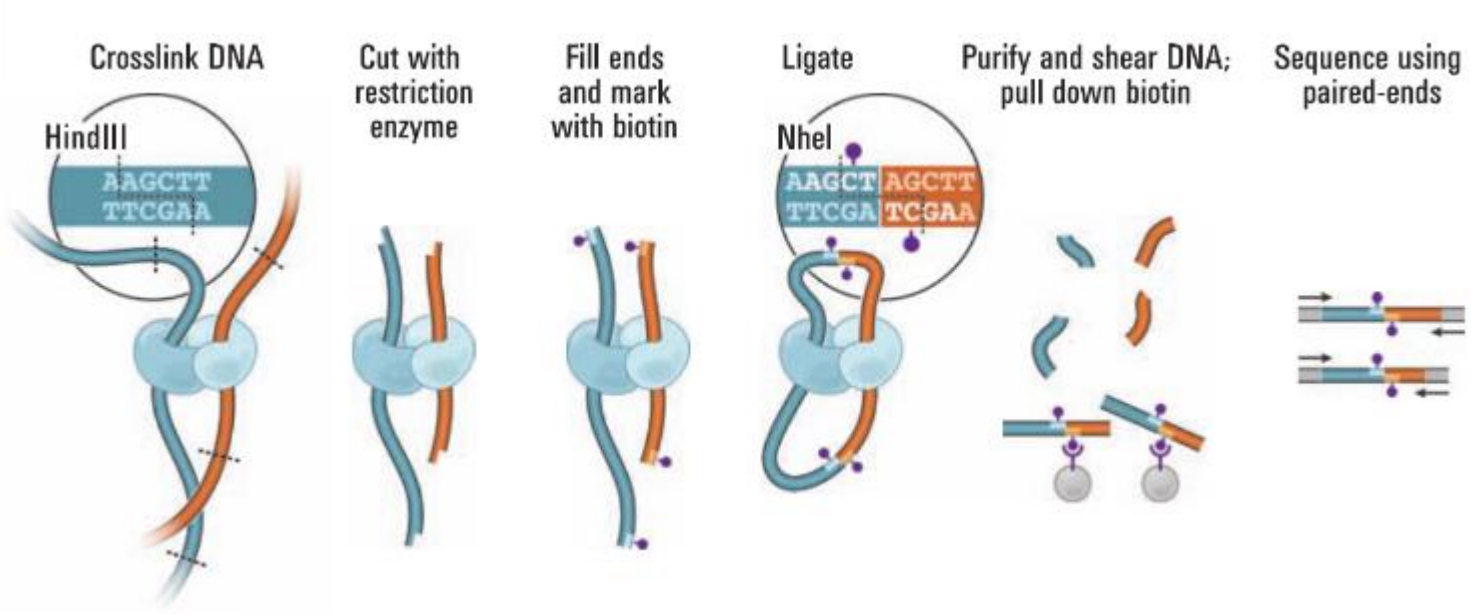
- ДНК-белок



ref

HI-C

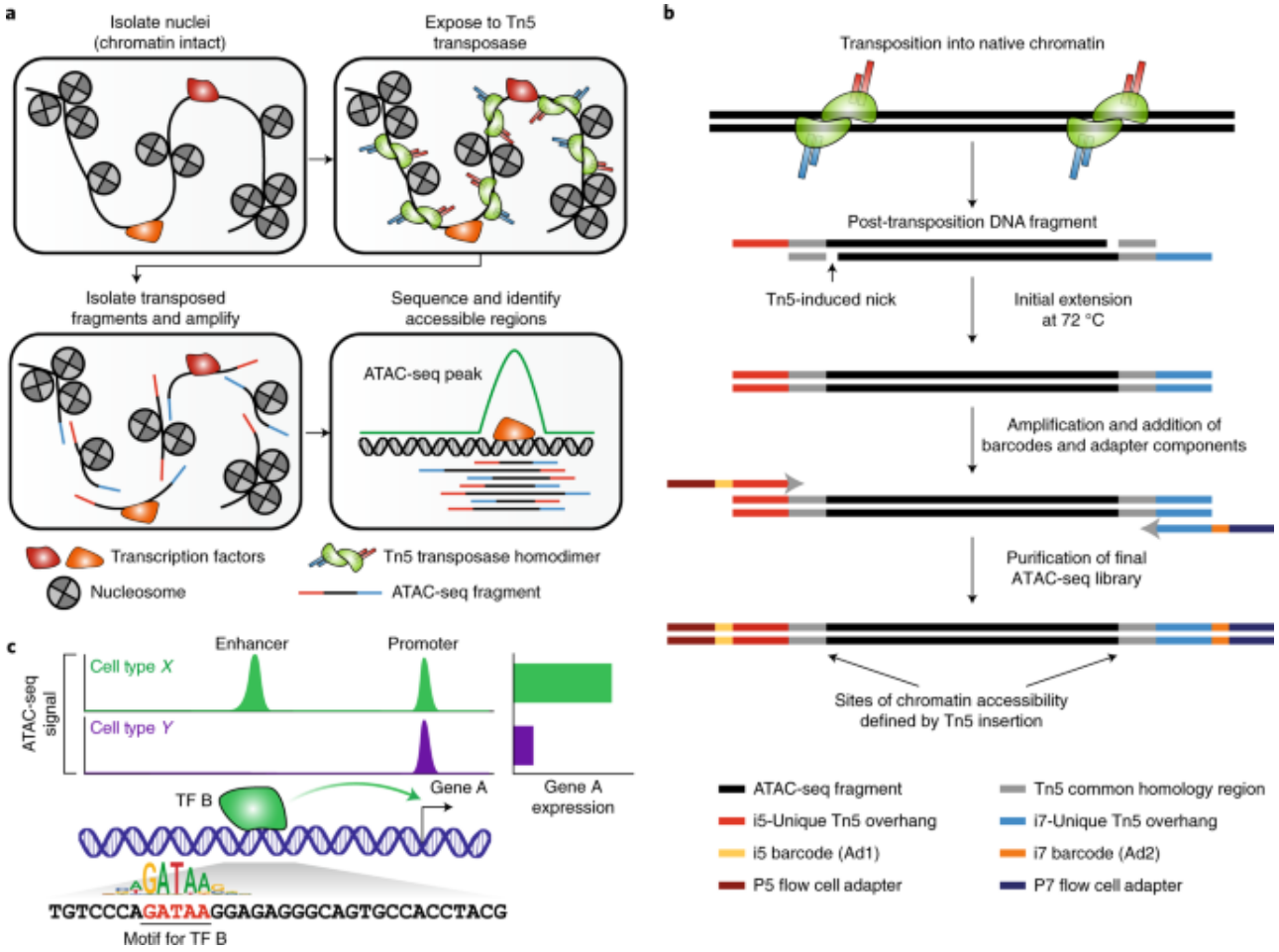
- ДНК-ДНК



Lieberman-aiden et al., 2009

ATAC-SEQ

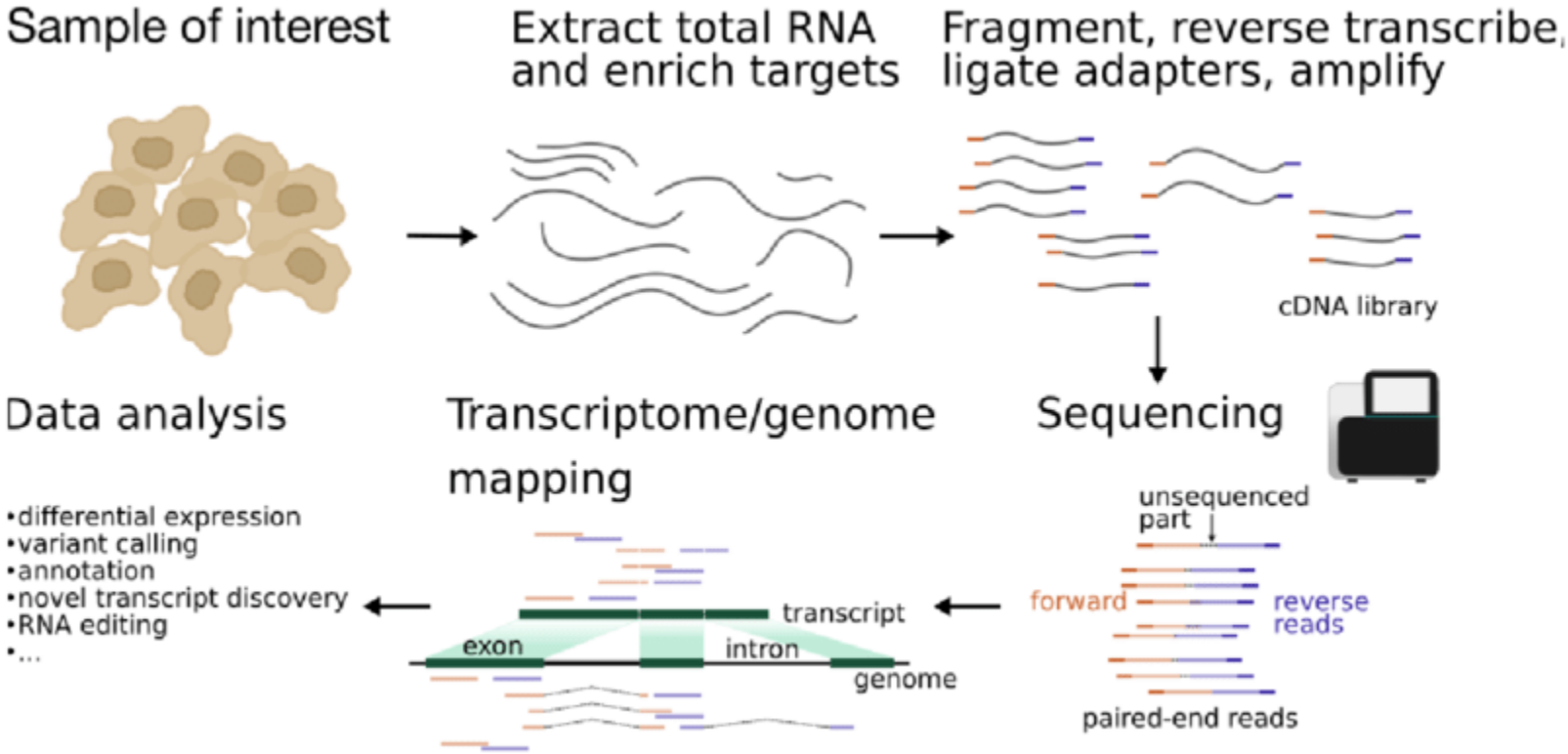
- Доступность хроматина



RNA-SEQ

- Секвенирование РНК
- **Illumina** умеет секвенировать только ДНК
- Что делать?

RNA-SEQ


Sample of interest
Extract total RNA and enrich targets
Fragment, reverse transcribe, ligate adapters, amplify
cDNA library
Sequencing
unsequenced part
forward
reverse reads
paired-end reads
Transcriptome/genome mapping
transcript
exon
intron
genome
Data analysis
•differential expression
•variant calling
•annotation
•novel transcript discovery
•RNA editing
•...

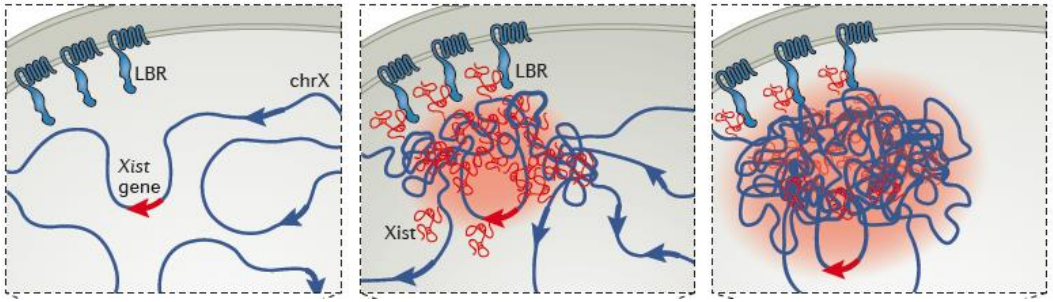
ref

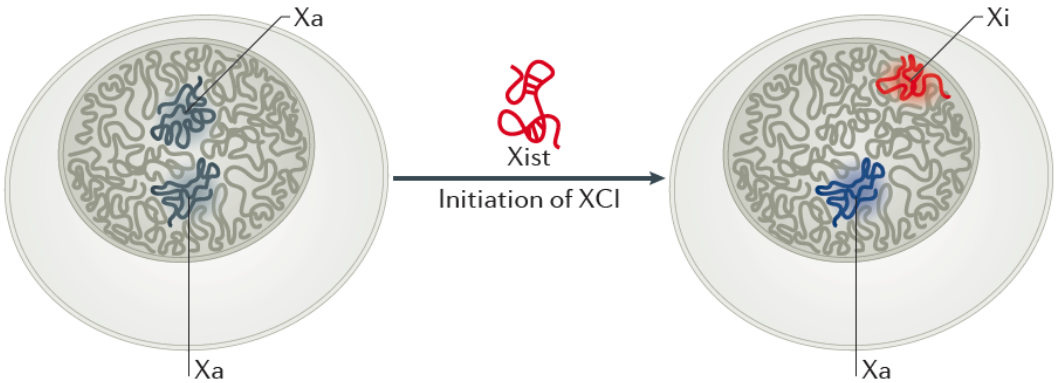
ТИПЫ РНК В КЛЕТКЕ

- Тотальная
- полиА
- Без фракции рРНК
- По размеру:
  - Малые (**snRNA, microRNA, …**)
- По внутриклеточной локализации:
  - Ядерные
  - Цитоплазматические

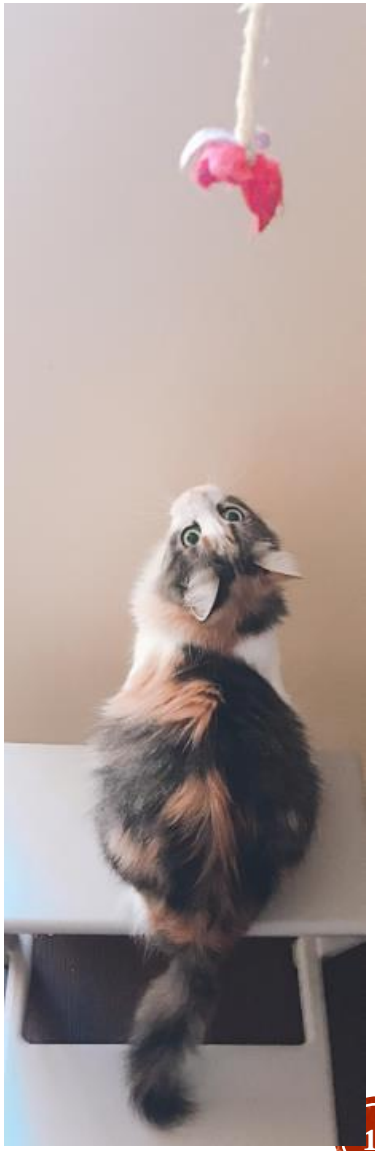
ЗАЧЕМ УБИРАТЬ МРНК?

XIST – инактивация X хромосомы у самок млекопитающих





Engreitz et al., 2016

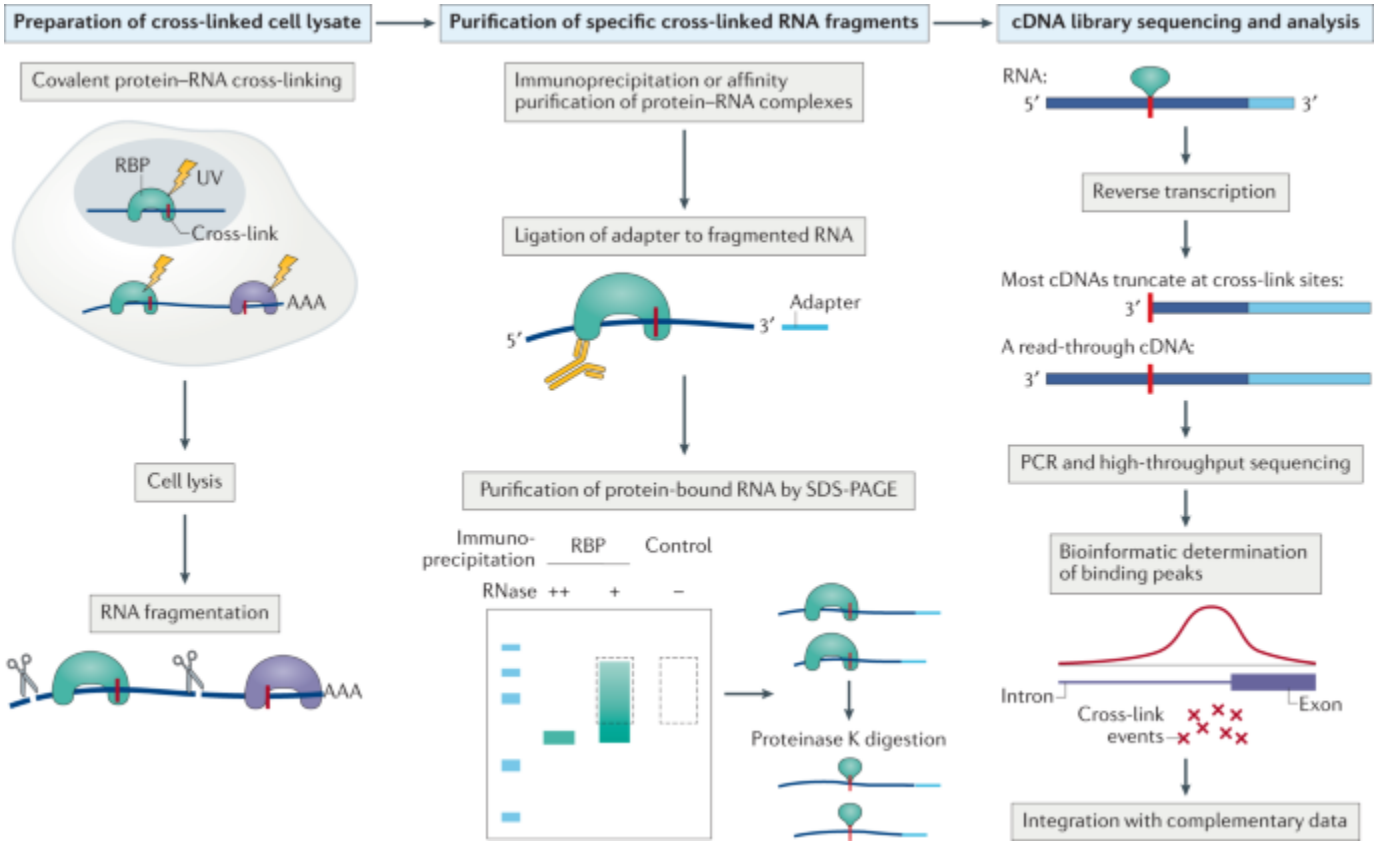


ПРОЦЕСС

- Подготовка нужной фракции РНК
- Проверка качества РНК
- Обратная транскрипция => кДНК
- Фрагментация (~200-300 нукл)
- Секвенирование
- Реплики

ECLIP

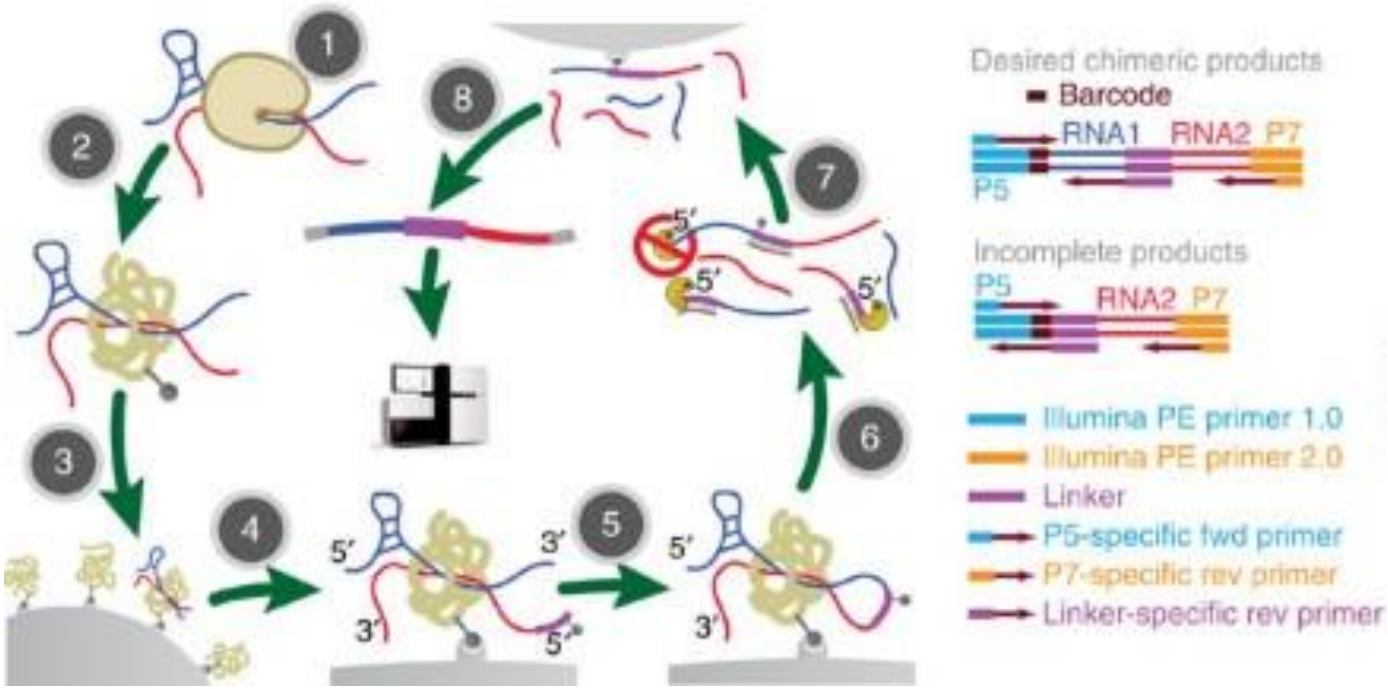
- РНК-белок



ref

MARIO

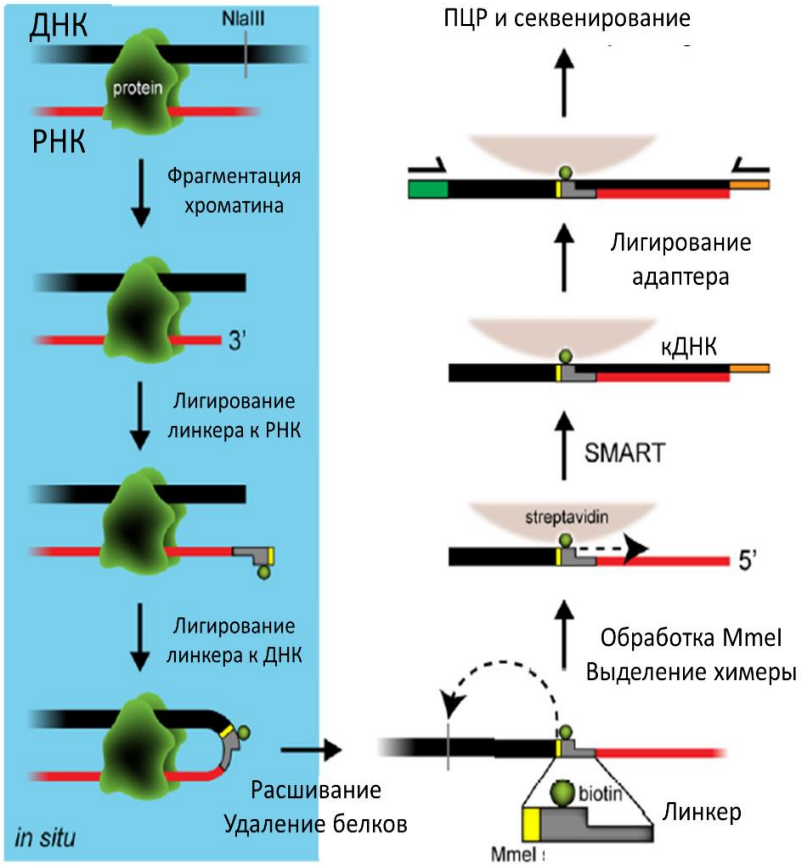
- РНК-РНК

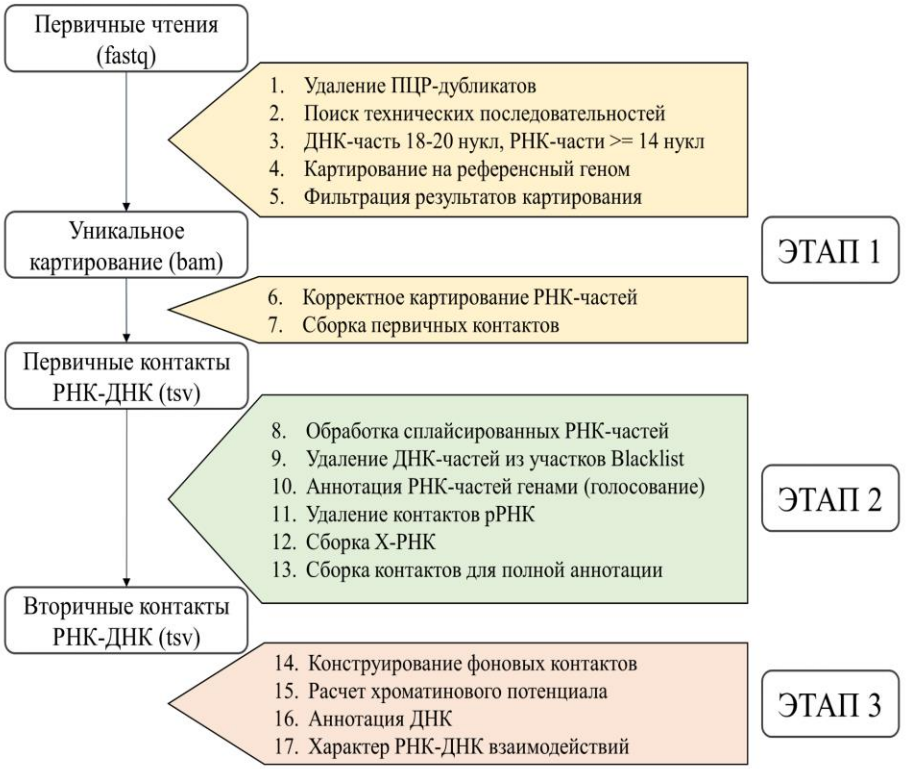


ref

RED-C

- РНК-ДНК

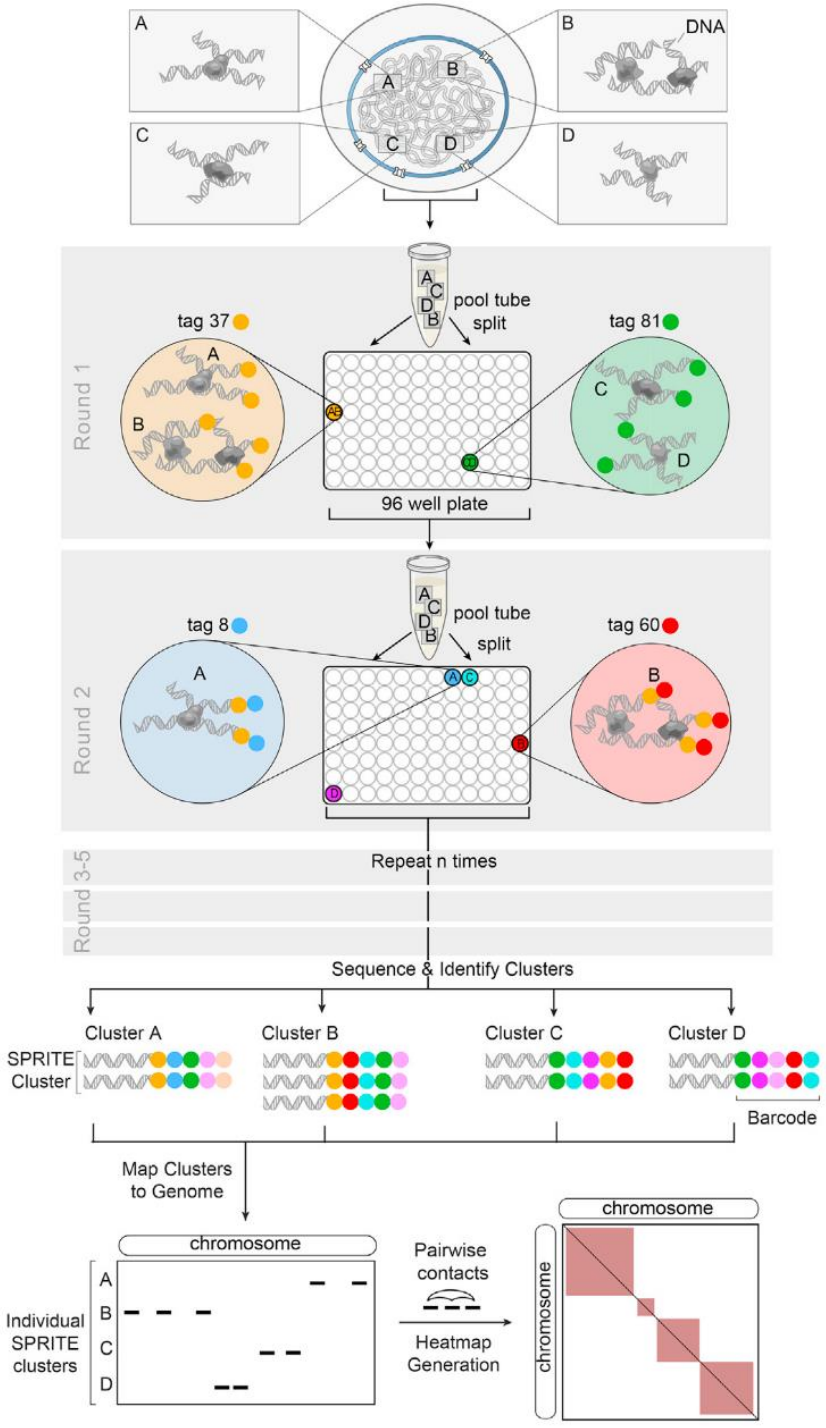




ref

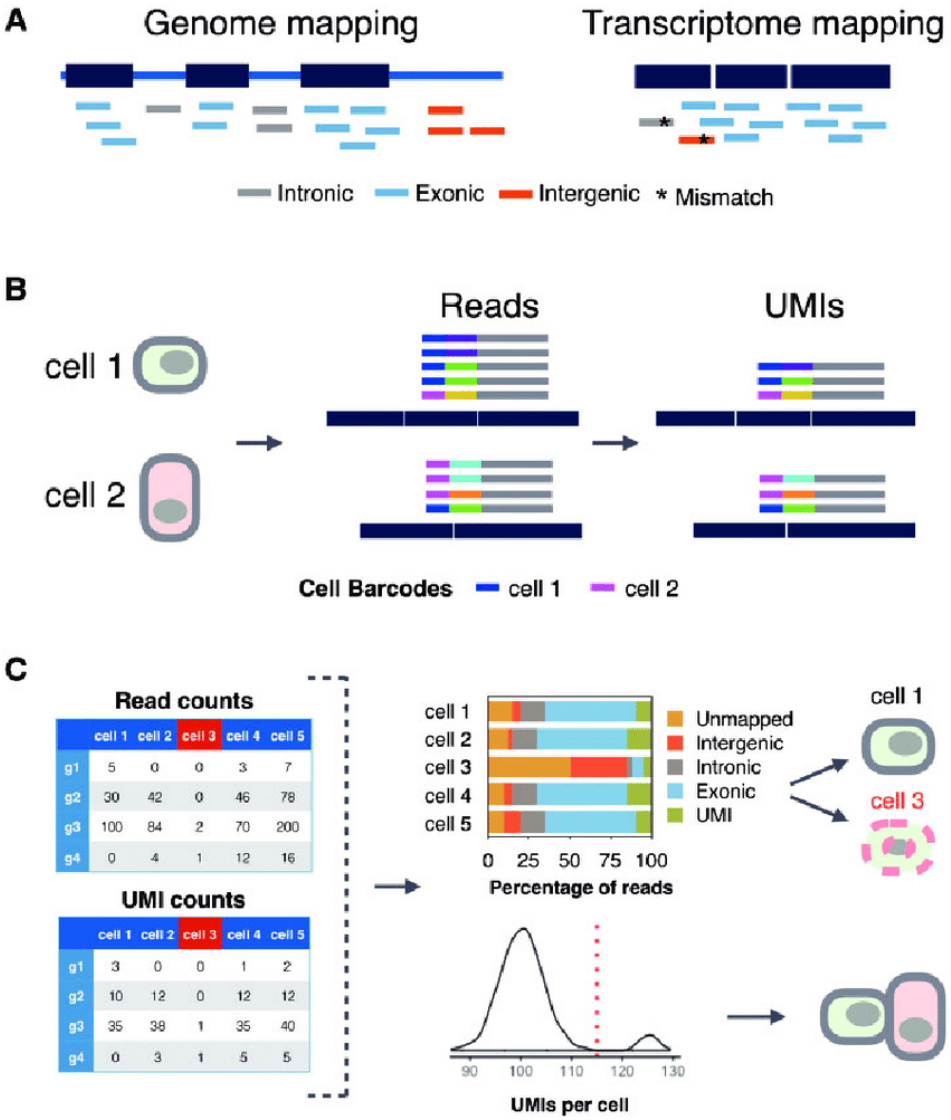
RD-SPRITE

- интеграктомика



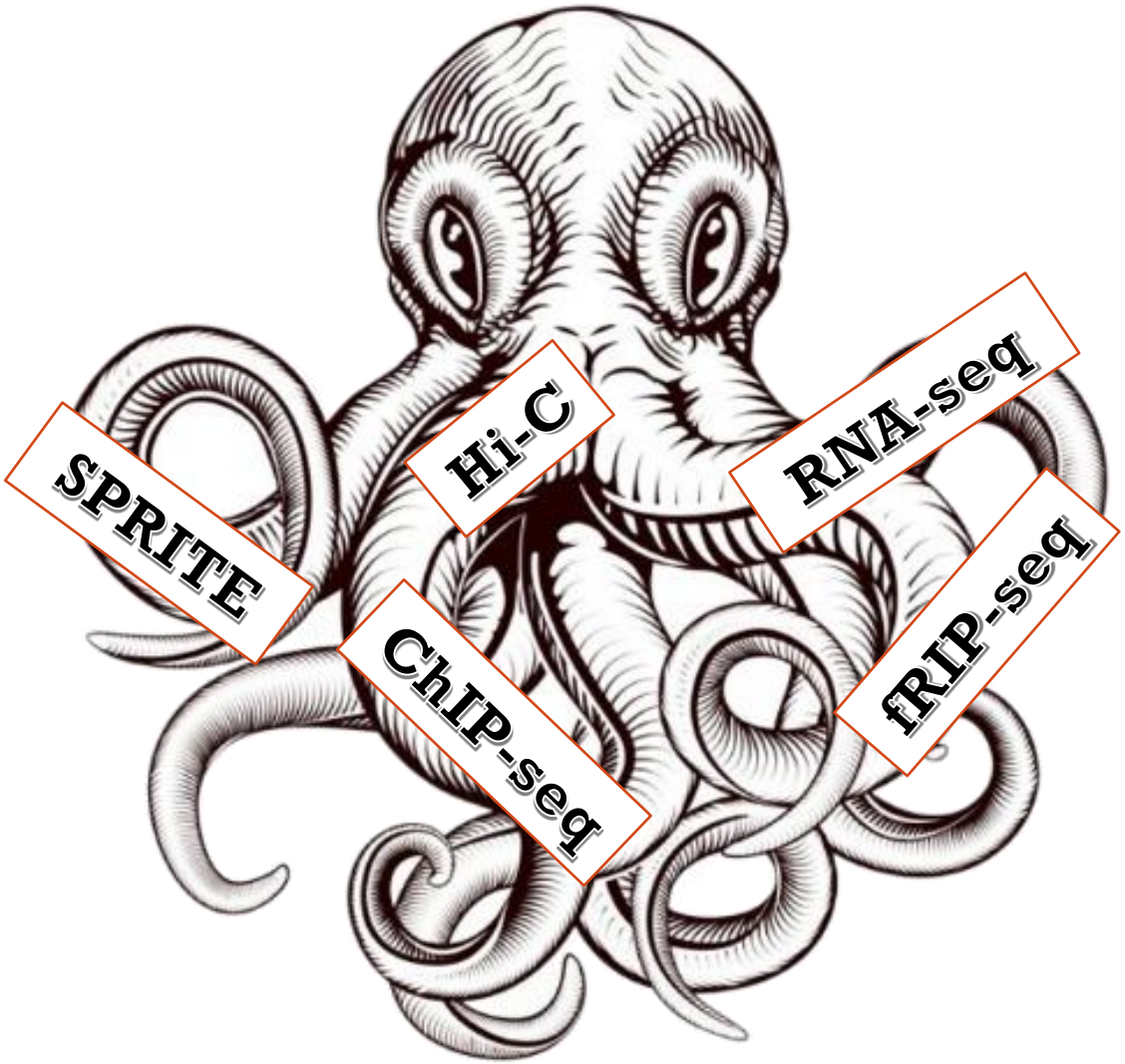
Quinodoz et. al. Cell 2018

SINGLE CELL



ref

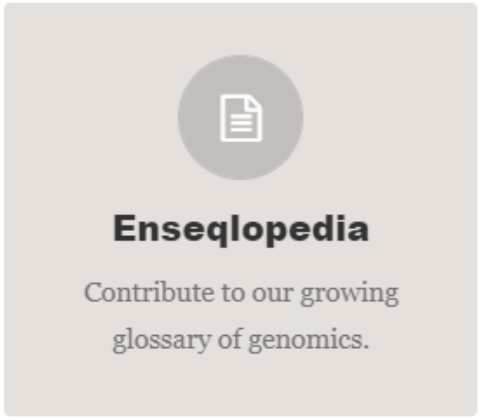
МНОГОСЕК

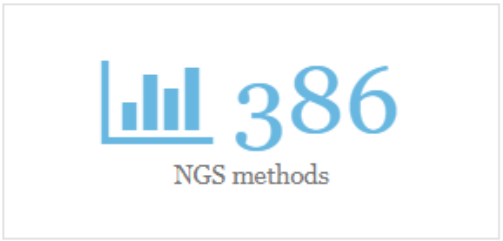


ОДНО И ТО ЖЕ...

ENSEQLOPEDIA

- http://enseqlopedia.com/





- RNA-Protein Interactions
  - AGO-CLIP
  - CLASH
  - CLIP-Seq or HITS-CLIP
  - DLAF
  - eCLIP
  - hiCLIP
  - iCLIP
  - miR-CLIP
  - miTRAP
  - PAR-CLIP
  - PIP-Seq
  - Pol II CLIP
  - RBNS
  - Ribo-Seq or ARTSeq
  - RIP-Seq
  - TRAP-Seq
  - TRIBE
  - BrdU-CLIP
  - HiTS-RAP
  - irCLIP

ЗАЧЕМ

- Зачем биоинформатику знать, что в пробирке?

ОТ ЭТОГО ВСЕ ЗАВИСИТ

- От протокола пробоподготовки зависит биоинформатический анализ
- Какой брать референсный геном (или не брать?)
- Как выстраивать протокол обработки данных?
- Какую задачу вообще решаем?
- Как лучше визуализировать результаты?

ВСЕ СДЕЛАНО ДО НАС
ENCODE

- https://www.encodeproject.org/

ENCODE

Showing 17208 results

List | Report | Download | Visualize | {;}

ASSAY → / ← BIOSAMPLE

	TF ChIP-seq	Histone ChIP-seq	DNase-seq	total RNA-seq	Mint-ChIP-seq	polyA plus RNA-seq	ATAC-seq	microRNA-seq	scRNA-seq	snATAC-seq	eCLIP	DNAme array	small RNA-seq	WGBS	long read RNA-seq	ChIA-PET	RAMPAGE	RNA microarray	genotyping array	CAGE	microRNA counts	RNA Bind-n-Seq	Repli-seq
tissue	439	2175	945	447		399	225	305	393	283	2	121	67	162	112	1	104	2	7	17	101		
dorsolateral prefrontal cortex	59	188	56	120				43							9								
adrenal gland	8	39	33	26		11	8	23	59	15	2	5	2	5	19		4				2		
heart left ventricle	16	84	10	9		2	15	5	54	65		3		4	4		2				2		
heart	6	81	39	21		16	7	25	40	8			1	10	6		2				8		
liver	42	91	24	3		20	8	7	8	8		1	1	9			2				7		
cell line	2836	773	180	136	54	155	124	39	6	32	250	87	110	18	56	104	29	67	73	50	8	92	
K562	750	19	4	17		15	4	2		3	145	3	7	1	4	9	1	9	2	9	1	6	
HepG2	814	15	2	6		11	2	2		3	105	3	3	2	3	4		6	2	6	1	6	
GM12878	190	15	1	5		13	2	2	2	3		3	6	1	4	4	1	7	2	6	1	6	
MCF-7	150	18	4	1		4	1	2		3		2	7		1	4		2	2	3	1	6	
HEK293	204	6										2				2		1	2				
primary cell	74	506	558	241	715	136	71	52	3	29		38	24	7	6	60	16	57	37	30	1	12	
naive thymus-derived CD4-positive, alpha-beta T cell		16	16	5	74	8	2	2		2						2							
T-cell		11	62	7	18	1	6							1		4							
naive thymus-derived CD8-positive, alpha-beta T cell			13	2	66	6	3	1		1						1							
CD14-positive monocyte	2	21	8		41	8				2			1	1						1			

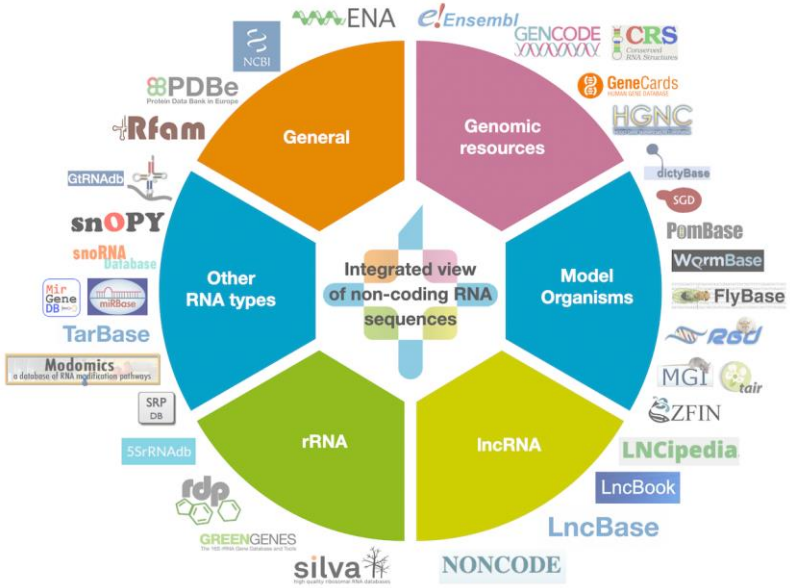
Унификация и единообразие!!!!!

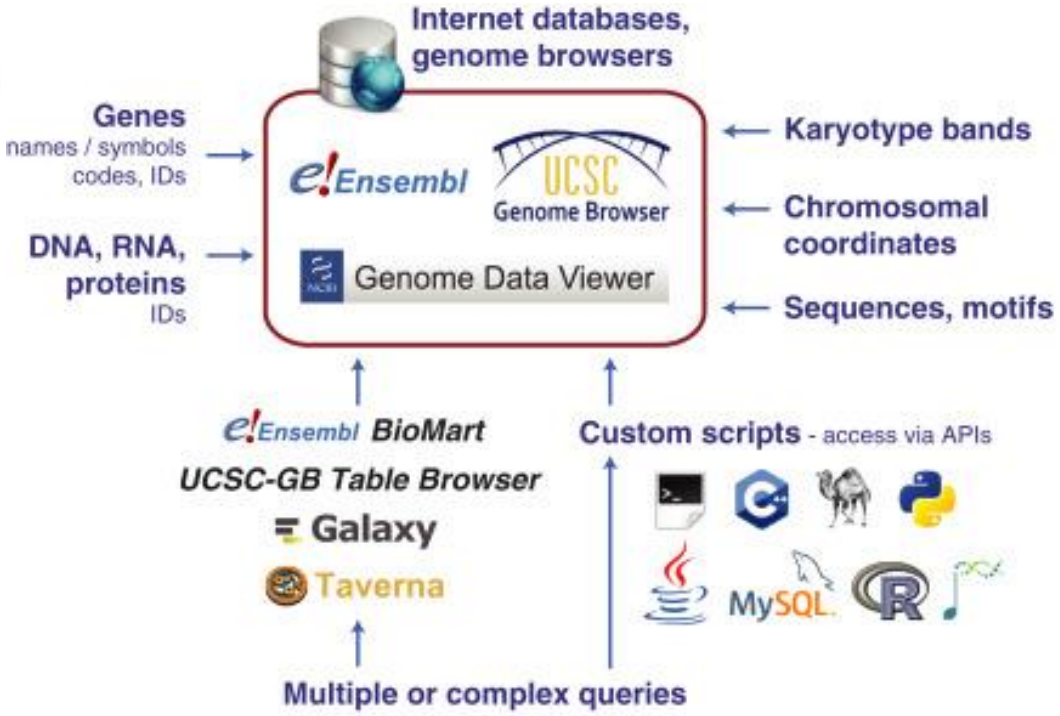
ЧТО ДАЛЬШЕ?

- Результаты анализа:
  - Дифференциально экспрессирующиеся гены
  - Таргеты транскрипционных факторов
  - …

Нужна какая-то аннотация, анализ сопредставленности, просто информация о гене или белке, …

РАЗНООБРАЗИЕ БАЗ ДАННЫХ





Это еще далеко не все

ref

ref

GENECARDS

- https://www.genecards.org/

Jump to section									
	Aliases	Disorders	Domains	Drugs	Expression	Function	Genomics	Localization	Orthologs
	Paralogs	Pathways	Products	Proteins	Publications	Sources	Summaries	Transcripts	Variants
Research Products	Antibodies	Assays	Proteins	Inhib. RNA	CRISPR	Exp. Assays	miRNA	Drugs	Animal Models
	Cell Lines	Clones	Primers	Genotyping					

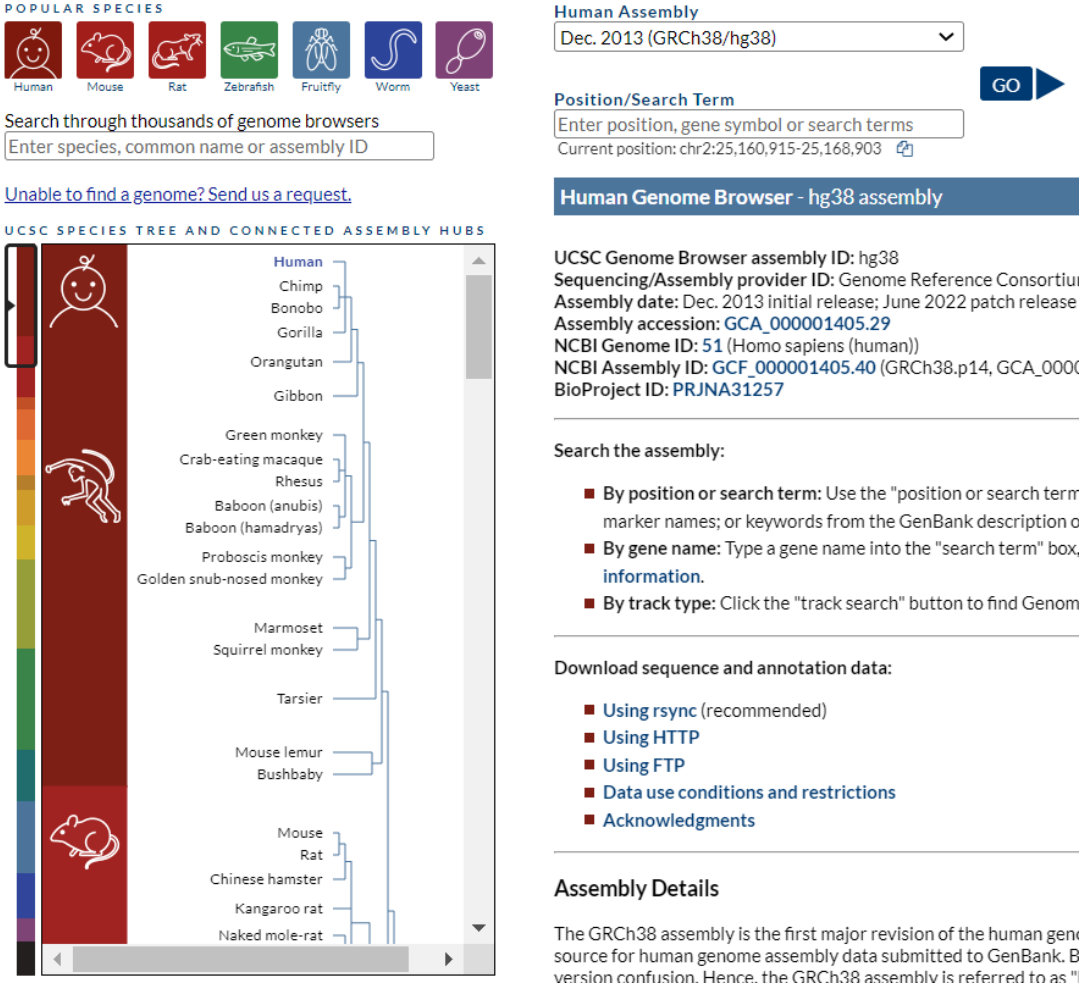
GeneCards Version 5.15 (Updated: Mar 27, 2023)

Total genes	415,866
HGNC approved	43,617
Disease genes	19,871
Hot genes	500

Category	# of Genes
Protein-coding	21,667
ncRNA genes	292,110
lncRNAs	130,578
piRNAs	111,811
miRNAs	6,904
rRNAs	1,272
tRNAs	1,160
snoRNAs	1,905
SRP_RNAs	9,022
circRNAs	120
Other ncRNAs	29,338
Functional elements	76,694
Pseudogenes	22,201
Genetic loci	1,387
Gene clusters	10
Uncategorized	1,797

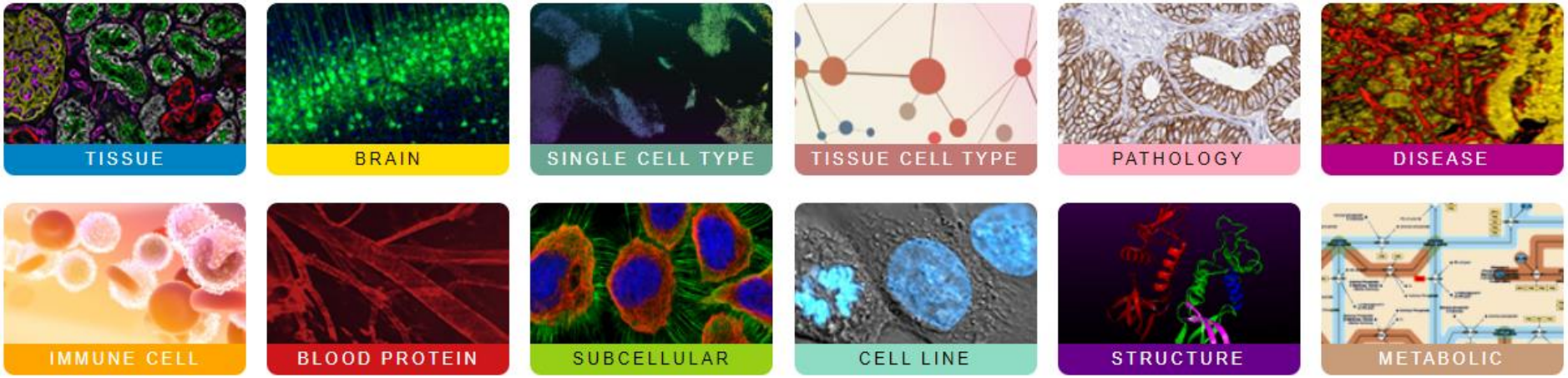
ГЕНОМНЫЙ БРАУЗЕР

- https://genome.ucsc.edu/index.html



HUMAN PROTEIN ATLAS

- https://www.proteinatlas.org/
- Не только protein



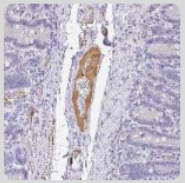
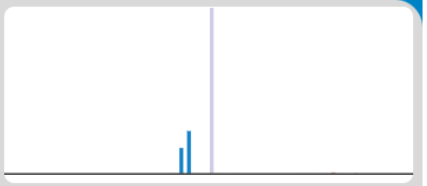
HUMAN PROTEIN ATLAS

APOB

SUMMARY | TISSUE | RNA BRAIN | RNA SINGLE CELL | RNA TISSUE CELL | PATHOLOGY | MS DISEASE | RNA IMMUNE | IA BLOOD | SUBCELL | RNA CELL LINE | N/A STRUCTURE | METABOLIC

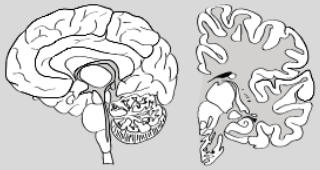
PROTEIN SUMMARY
SECTION OVERVIEW
RNA DATA
ANTIBODY DATA

TISSUE[i]

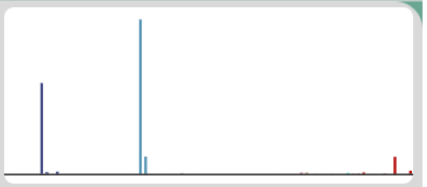
Tissue expression cluster[i]	Liver - Metabolism (mainly)
Tissue specificity[i]	Group enriched (intestine, liver)
Protein expression[i]	Distinct positivity in plasma.

BRAIN[i]



Human regional specificity[i]	Not detected
Pig regional specificity[i]	Not detected
Mouse regional specificity[i]	Not detected

SINGLE CELL TYPE[i]

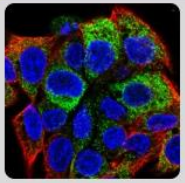
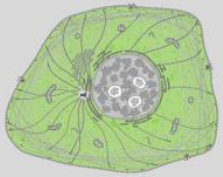



Single cell type expression cluster[i]	Hepatocytes - Metabolism (mainly)
Cell type specificity[i]	Group enriched (Hepatocytes, Proximal enterocytes)

TISSUE CELL TYPE[i]

- Adipocytes (Subcutaneous)
- Adipocytes (Visceral)
- Adipocytes (Breast)
- Hepatocytes

SUBCELLULAR[i]

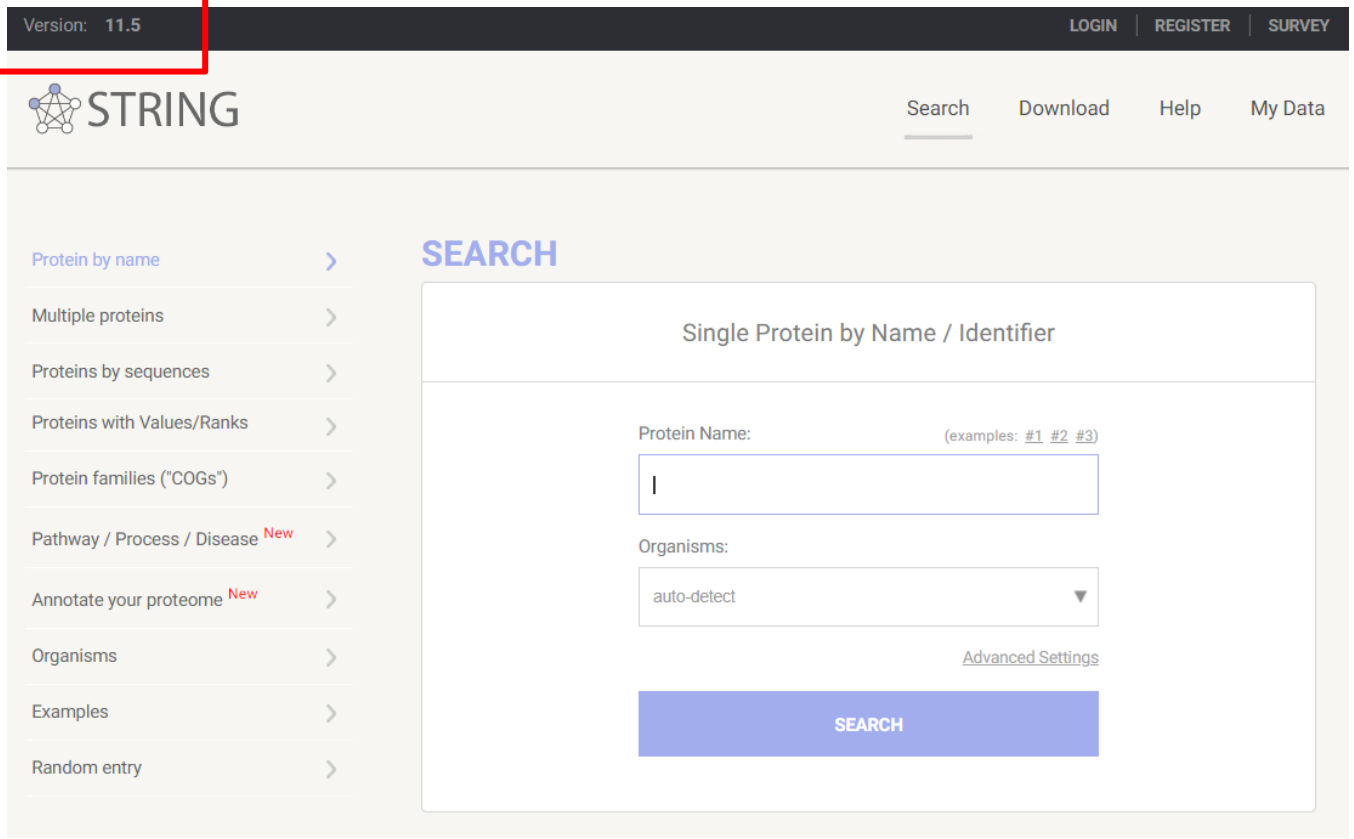
Main location[i]	Localized to the Vesicles, Cytosol

ЧТО ДЕЛАТЬ

- Если сразу МНОГО генов?
- Анализ обогащения наборов
- Гены аннотированы и сгруппированы
- Например: участвуют в определенном метаболическом пути
- Статистический метод, который определяет, что гены из заранее заданного набора встречаются в вашем списке чаще, чем вы бы ожидали по случайным причинами (перепредставлены)

STRING

- https://string-db.org/



STRING

Node Color

colored nodes:
query proteins and first shell of interactors

white nodes:
second shell of interactors

Node Content

empty nodes:
proteins of unknown 3D structure

filled nodes:
some 3D structure is known or predicted

Known Interactions

from curated databases

experimentally determined

Predicted Interactions

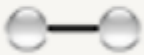
gene neighborhood

gene fusions

gene co-occurrence

Others

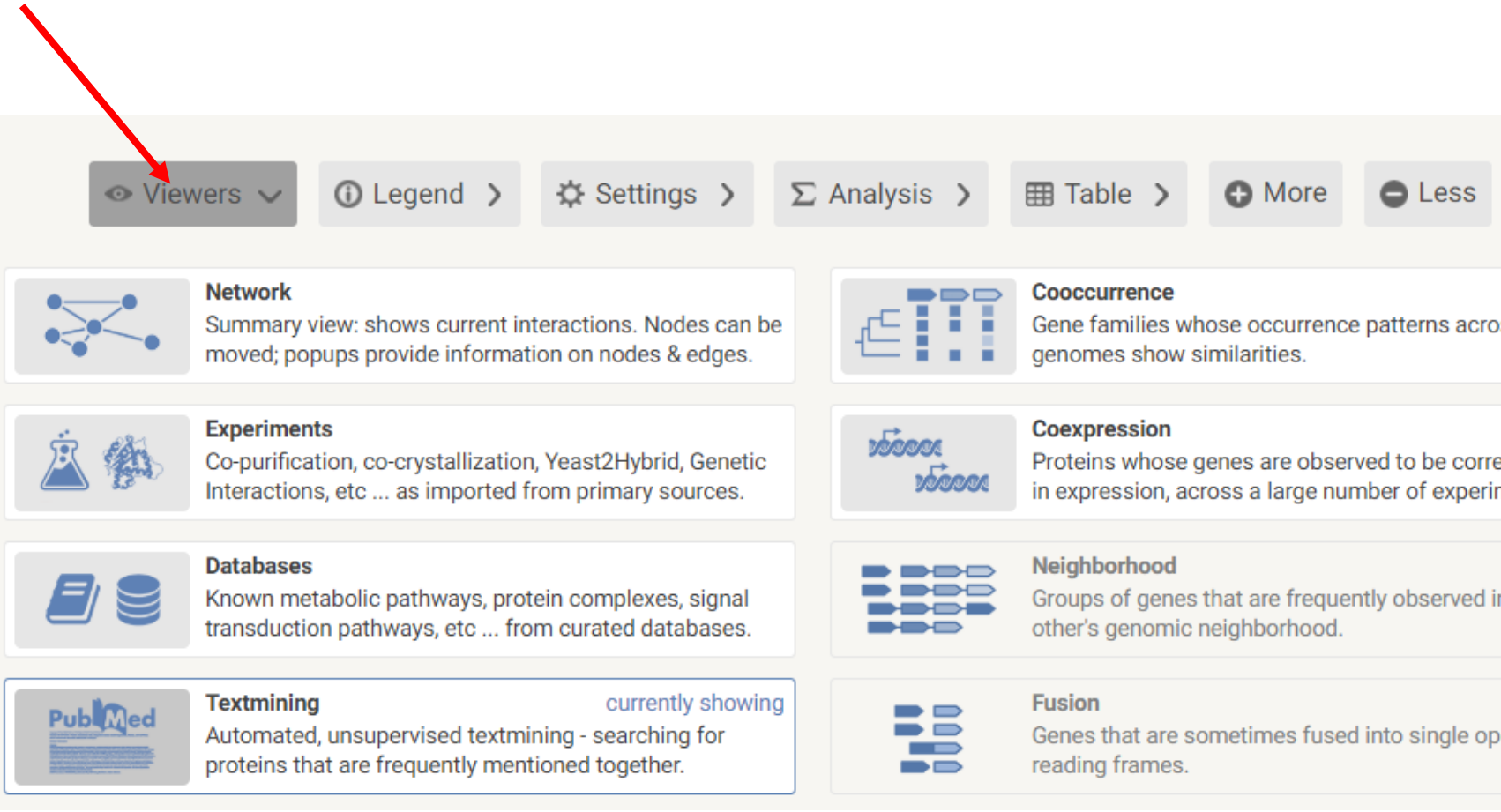
textmining



co-expression

protein homology

STRING


Viewers
Legend
Settings
Analysis
Table
More
Less
Network
Summary view: shows current interactions. Nodes can be moved; popups provide information on nodes & edges.
Cooccurrence
Gene families whose occurrence patterns across genomes show similarities.
Experiments
Co-purification, co-crystallization, Yeast2Hybrid, Genetic Interactions, etc ... as imported from primary sources.
Coexpression
Proteins whose genes are observed to be correlated in expression, across a large number of experiments.
Databases
Known metabolic pathways, protein complexes, signal transduction pathways, etc ... from curated databases.
Neighborhood
Groups of genes that are frequently observed in each other's genomic neighborhood.
PubMed
Textmining
currently showing
Automated, unsupervised textmining - searching for proteins that are frequently mentioned together.
Fusion
Genes that are sometimes fused into single open reading frames.

KEGG

- Метаболические пути
- https://www.genome.jp/kegg/kegg2.html

KEGG - Table of Contents

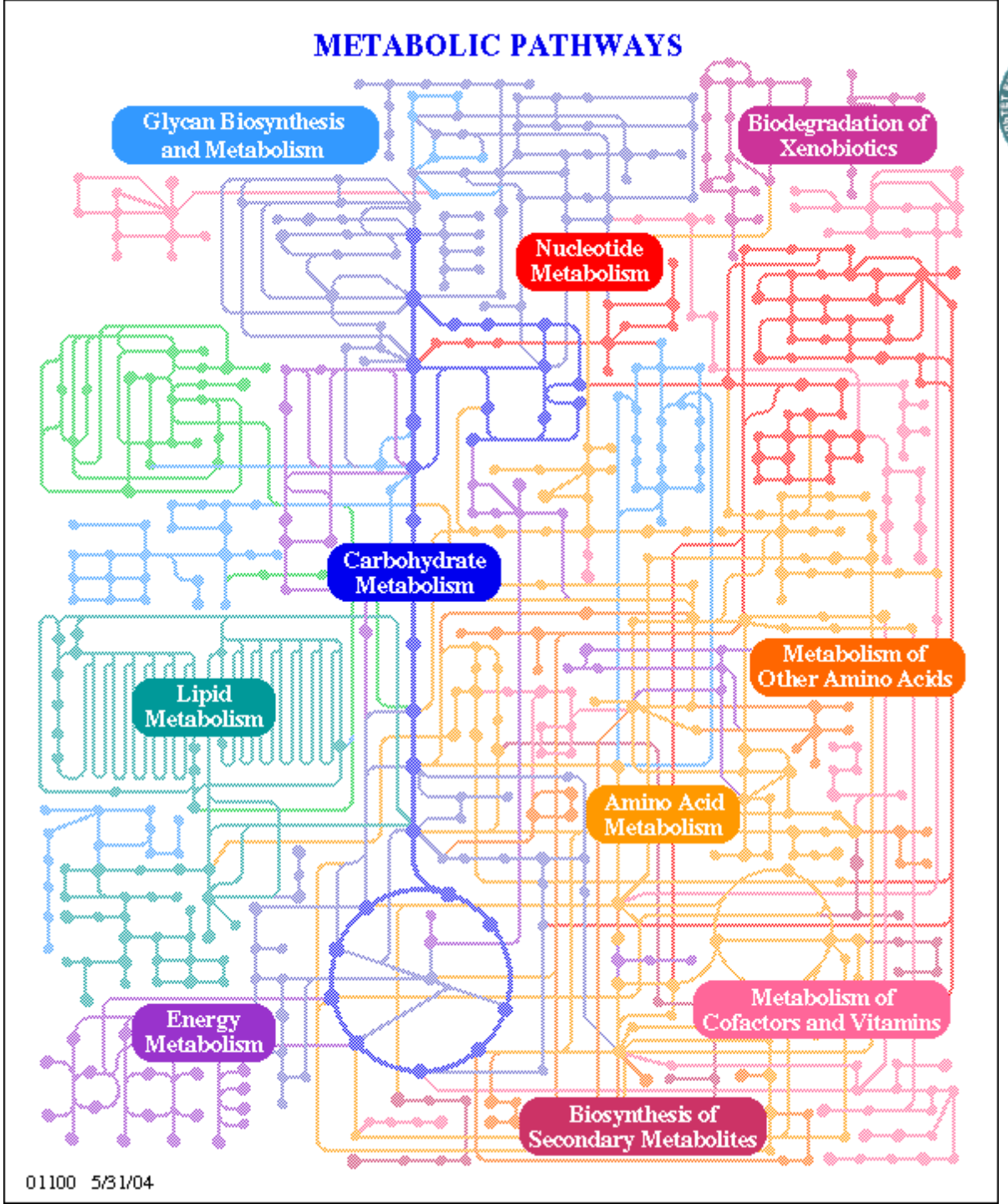
KEGG2 PATHWAY BRITE MODULE KO GENES COMPOUND NETWORK DISEASE DRUG

Search references cited in KEGG (PMID, DOI, author, title, journal) Go

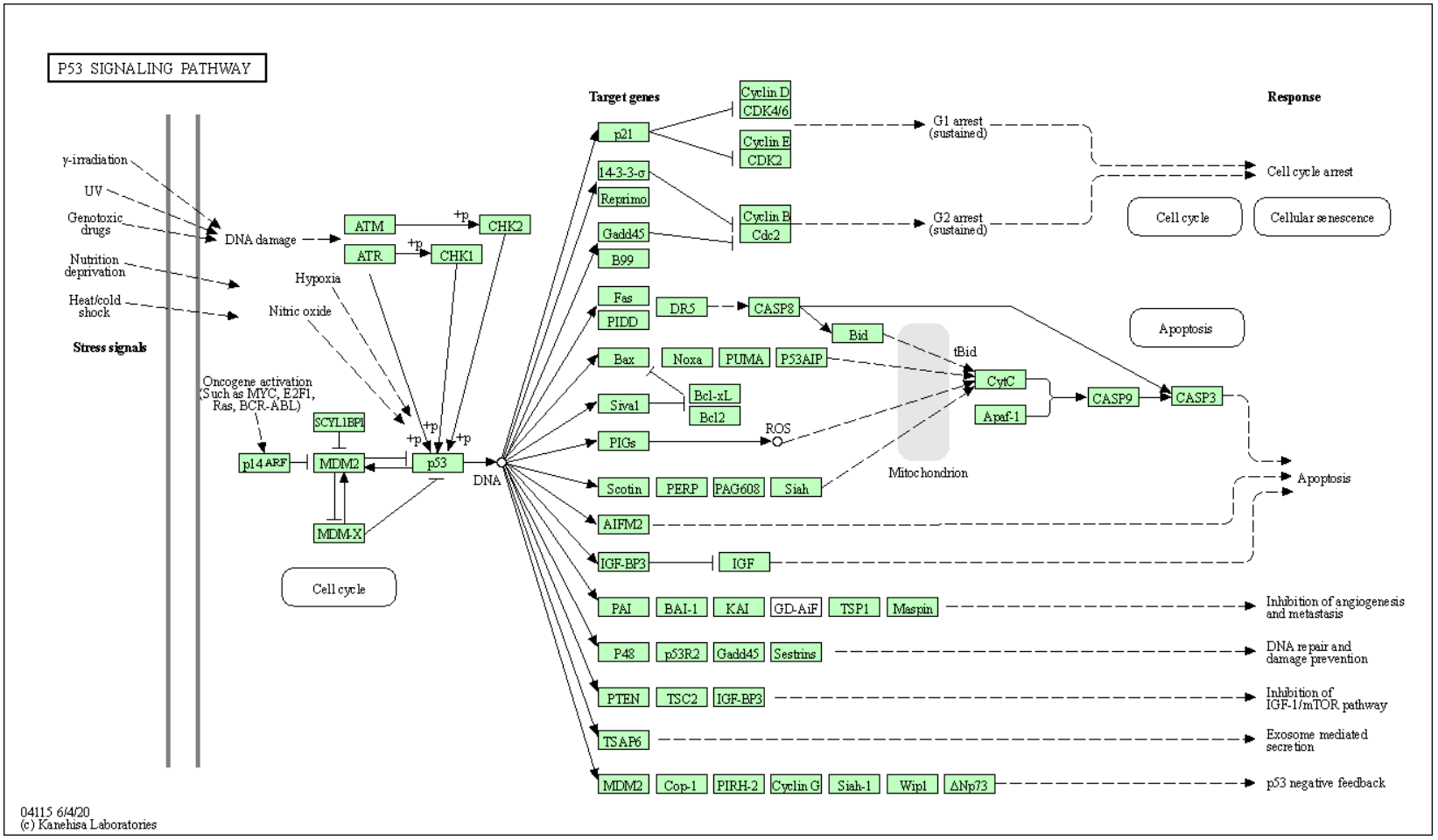
Number of references (2023/4/1)

total	69,123	pathway	6,518	ko	27,728	glycan	919	network	2,393
		brite	444	genome	5,994	reaction	2,020	variant	1,366
		module	1,089	agenes	3,023	enzyme	15,880	disease	10,211

KEGG

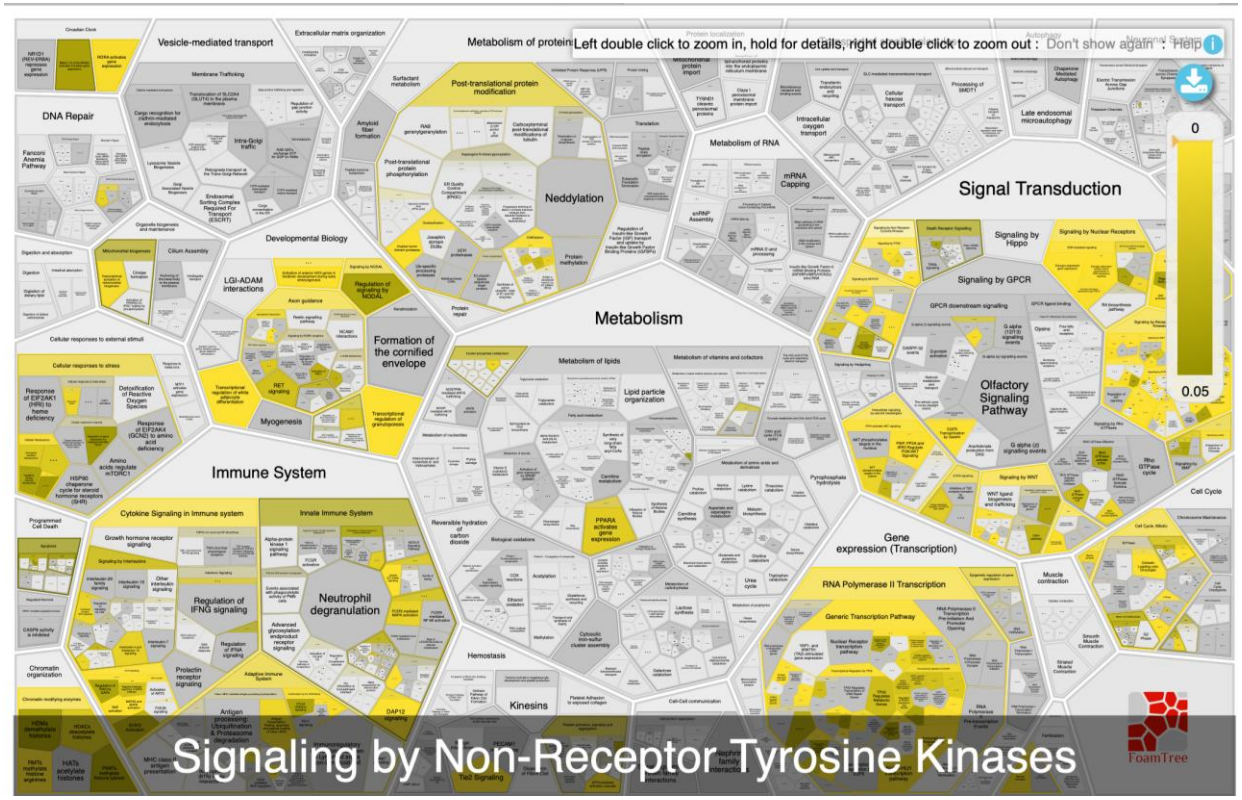

METABOLIC PATHWAYS
Glycan Biosynthesis and Metabolism
Biodegradation of Xenobiotics
Nucleotide Metabolism
Carbohydrate Metabolism
Metabolism of Other Amino Acids
Lipid Metabolism
Amino Acid Metabolism
Metabolism of Cofactors and Vitamins
Energy Metabolism
Biosynthesis of Secondary Metabolites
01100 5/31/04

KEGG


P53 SIGNALING PATHWAY
γ-irradiation
UV
Genotoxic drugs
Nutrition deprivation
Heat/cold shock
Stress signals
DNA damage
ATM
ATR
CHK2
CHK1
Hypoxia
Nitric oxide
Oncogene activation (Such as MYC, E2F1, Ras, BCR-ABL)
SCYL1BP1
p14 ARF
MDM2
MDM-X
p53
DNA
Cell cycle
Target genes
p21
14-3-3-σ
Reprimo
Gadd45
B99
Fas
PIDD
Bax
Noxa
PUMA
P53AIP
Siva1
Bcl-xL
Bcl2
PIGs
Scotin
PERP
PAG608
Siah
AIFM2
IGF-BP3
IGF
PAI
BAI-1
KAI
GD-AiF
TSP1
Maspin
P48
p53R2
Gadd45
Sestrins
PTEN
TSC2
IGF-BP3
TSAP6
MDM2
Cop-1
PIRH-2
Cyclin G
Siah-1
Wip1
ΔNp73
Cyclin D
CDK4/6
Cyclin E
CDK2
Cyclin B
Cdc2
DR5
CASP8
Bid
tBid
CytC
Apaf-1
CASP9
CASP3
ROS
Mitochondrion
G1 arrest (sustained)
G2 arrest (sustained)
Response
Cell cycle arrest
Cell cycle
Cellular senescence
Apoptosis
Apoptosis
Inhibition of angiogenesis and metastasis
DNA repair and damage prevention
Inhibition of IGF-1/mTOR pathway
Exosome mediated secretion
p53 negative feedback
04115 6/4/20
(c) Kanehisa Laboratories

REACTOME

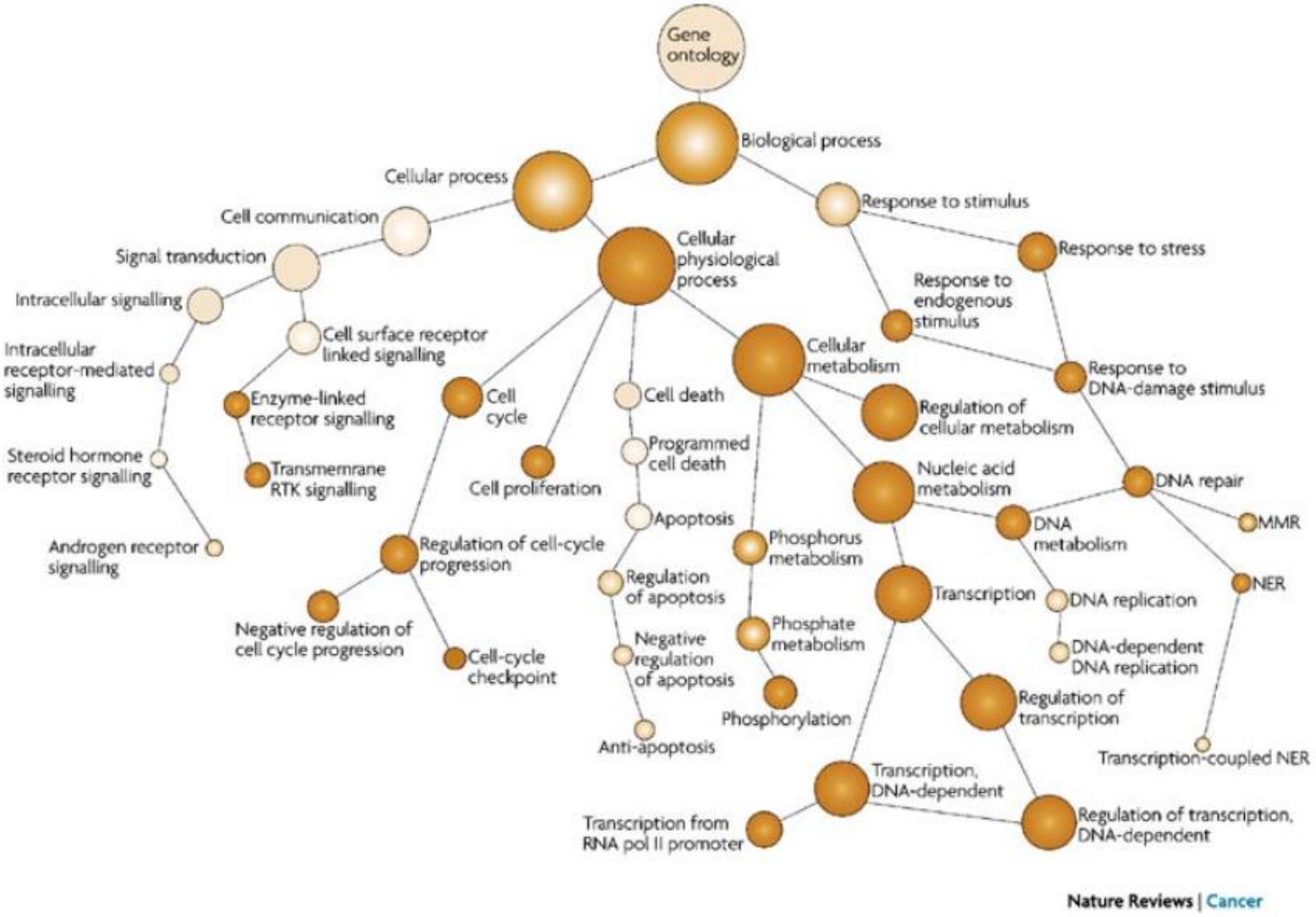
- Метаболические пути
- https://reactome.org/



GENE ONTOLOGY

- http://geneontology.org/

GENE ONTOLOGY



Hu P et al, Nature Reviews Cancer 2007

PUBMED

MEDLINE PubMed Production Statistics

	FY2022	FY2021	FY2020	FY2019	FY2018
MEDLINE Citations Indexed (Annual)	1,369,611	1,291,807	952,919	956,390	904,636
MEDLINE Citations Cumulative Total	29,807,639	28,444,654	27,149,277	26,196,358	25,239,968
MEDLINE Journal Titles	5,282	5,282	5,274	5,243	5,251
PubMed Citations (Annual)	1,714,780	1,733,089	1,514,199	1,366,447	1,329,148
PubMed Citations Cumulative Total	34,693,538	33,136,289	31,563,992	30,178,674	28,934,389
PubMed Searches	2.58 Billion	2.57 Billion	3.3 Billion	3.1 Billion	3.3 Billion
Web/Interactive	1.283 Billion	1.186 Billion	1.076 Billion	896 Million	831 Million
Script/E-Utilities	1.303 Billion	1.391 Billion	2.2 Billion	2.2 Billion	2.5 Billion

ref

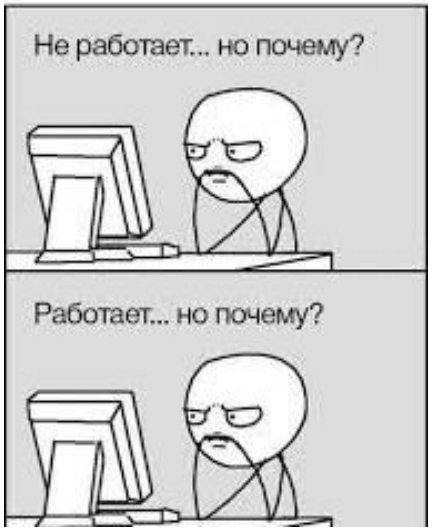
ЧЕМ ЗАНИМАЕТСЯ БИОИНФОРМАТИКА СЕГОДНЯ?

- Анализ медицинских и биологических данных разнообразной специфики:
  - Аминокислотные и нуклеотидные последовательности
  - Эволюция
  - Структуры биологических молекул
  - Разработка лекарств
  - Популяционные исследования
  - Обработка сигналов (МРТ, ЭЭГ, …)
  - Обработка изображений (микроскопия)

А ЕЩЕ

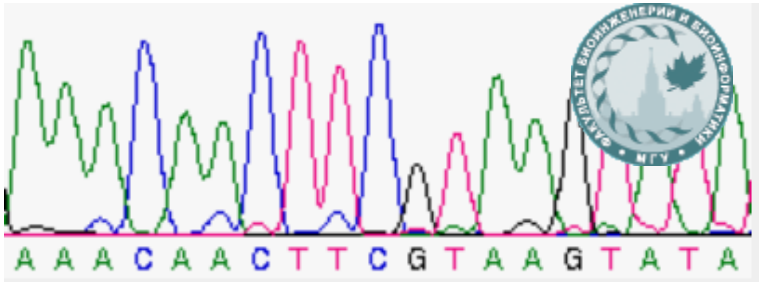
- Проектирование баз данных для хранения информации
- Систематизация данных
- Разработка и совершенствование алгоритмов и программ для анализа биологических данных

- Написание статей, грантов, патентов
  - Рецензирование публикаций
    - Участие в конференциях и научная коммуникация
      - Преподавание
        - Ведение дипломных и курсовых проектов
          - Чтение чужих статей
            - Самообразование
              - Поиск себя…

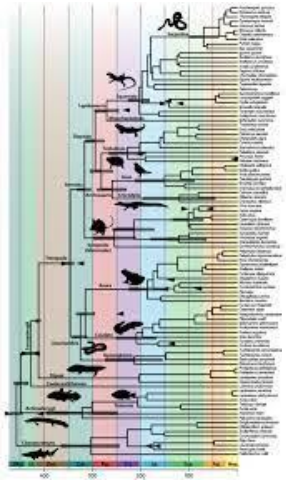


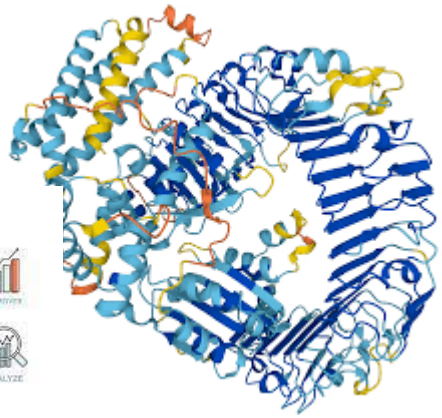
Давайте по порядку

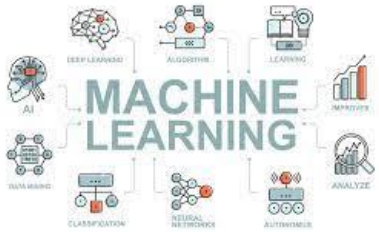
python



ДАННЫЕ

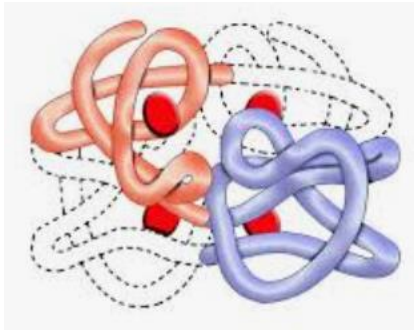




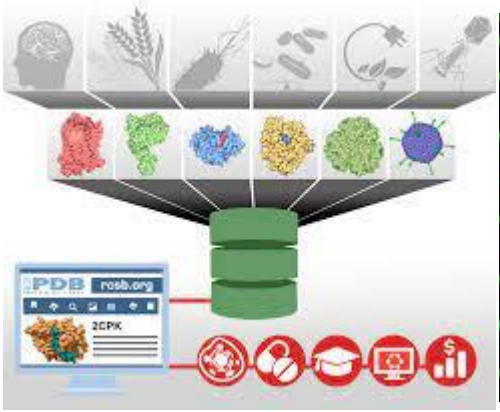


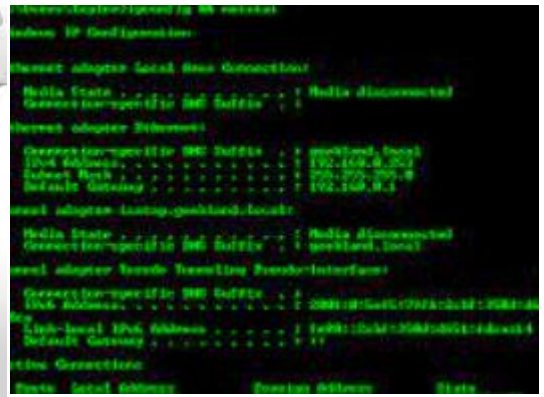
- Откуда берутся данные?
- Какие они бывают?
- Сколько их вообще?
- Где они хранятся?
- Как их доставать, анализировать и что для этого нужно знать и уметь?
- Какой в этом биологический смысл?

BASH
THE BOURNE-AGAIN SHELL



R

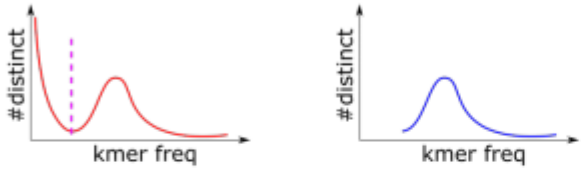




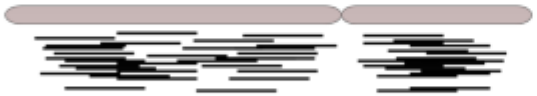
БУДНИ БИОИНФОРМАТИКА

Read clipping

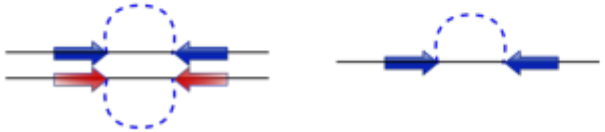
Read error correction



Read mapping

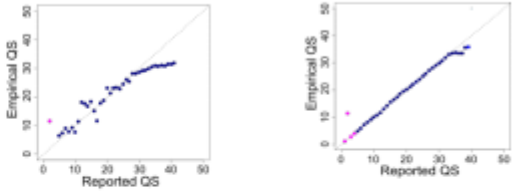


Duplicate removal

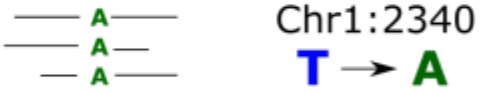


Re-alignment

Base Quality Adjustment

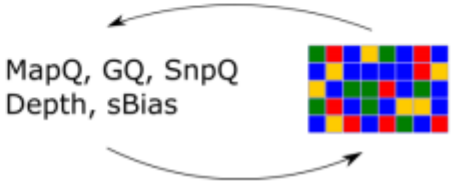


SNP calling



SNP filtering

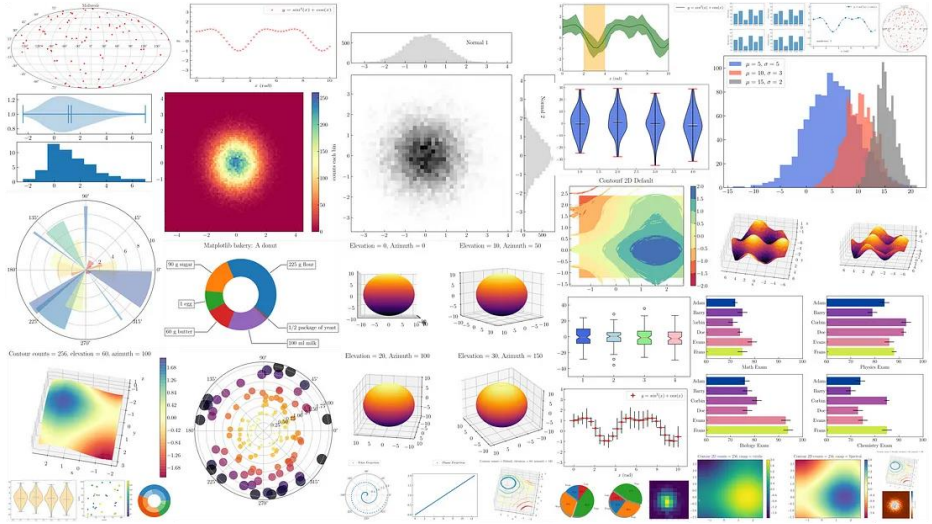
Evaluation



ref

ВИЗУАЛИЗАЦИЯ

- Важный этап анализа – представление своих результатов
- Табличка -> картинка



ref

ref

ВСЕ ТЕЧЕТ, ВСЕ МЕНЯЕТСЯ

- Версии референсов
- Базы данных
- Протоколы обработки данных
- Актуальные программы
- Интерфейсы сервисов
- Пакеты и программы

- ВСЕ МЕНЯЕТСЯ И ОБНОВЛЯЕТСЯ!
- Нужно фиксировать версии и не путать их в рамках одного проекта
- Помните о воспроизводимости

ДОМАШНЕЕ ЗАДАНИЕ

- **Вариант 1**
- Взять любой протокол секвенирования (смотрите Enseqlopedia, статьи, …), но не из лекции
- Описать:
  - Общую задачу (изучение сайтов связывания транскрипционных факторов, пространственной структуры хроматина, …)
  - В общих чертах (без подробностей!!!) пробоподготовка и обработка
  - Как использовали метод, что показали в рамках 1-2 статей (ищите в **PubMed**)

ДОМАШНЕЕ ЗАДАНИЕ

- **Вариант 2**
- Возьмите любой ген (искать: **HPA, GeneCards**, Википедия, …)
- Кратко опишите функцию гена
- Найдите ген в геномном браузере, приведите картинку
- Подайте на вход сервису **STRING**
- Приведите полученный граф (можно увеличить)
- Опишите его (любой из 8 вариантов вкладки **Viewers**)
- Найдите информацию о анализе вашего списка с помощью **KEGG** (вкладка **Analysis**)
- Попробуйте интерпретировать
- Есть ли выбранный ген или генный продукт в **HPA**? Какая у него клеточная локализация?

ДОМАШНЕЕ ЗАДАНИЕ

- При выполнении домашнего задания не стесняйтесь добавлять иллюстрации и рассуждения
- Вся информация, которую Вы взяли из внешних источников, должна содержать соответствующие ссылки
- Оформите домашнее задание в виде документа (**.pdf**)
- Домашнее задание будет проверено, если понятно, что сделано
  - Описана задача: в чем заключалось задание
  - Понятен ход рассуждения: что вы сделали для выполнения задания
  - Присутствует логика изложения: каждое последующее действие вытекает из предыдущего
  - Указаны все ссылки на внешние источники
  - Есть минимальный анализ (ваш собственный!) результатов

НО! Это не курсовая! Без фанатизма! Рассчитывайте на 2-4 страницы

ДОМАШНЕЕ ЗАДАНИЕ

- Назовите документ **FBB_MFK_2023_Zharikova**(заменить)**.pdf**
- Отправьте на почту: **azharikova89@gmail.com**


СПАСИБУШКИ