

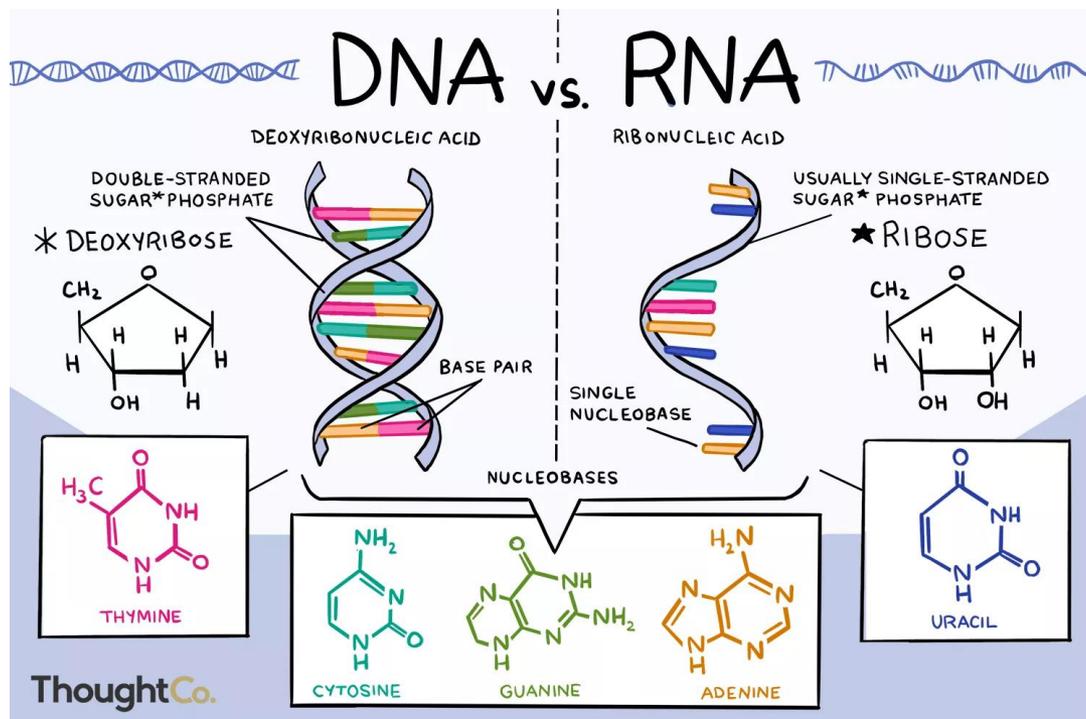
СЕКВЕНЦИРОВАНИЕ ОТ N ДО Y

Анастасия Жарикова
azharikova89@gmail.com
ФББ МГУ - 13 марта 2024



СЕКВЕНИРОВАНИЕ

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





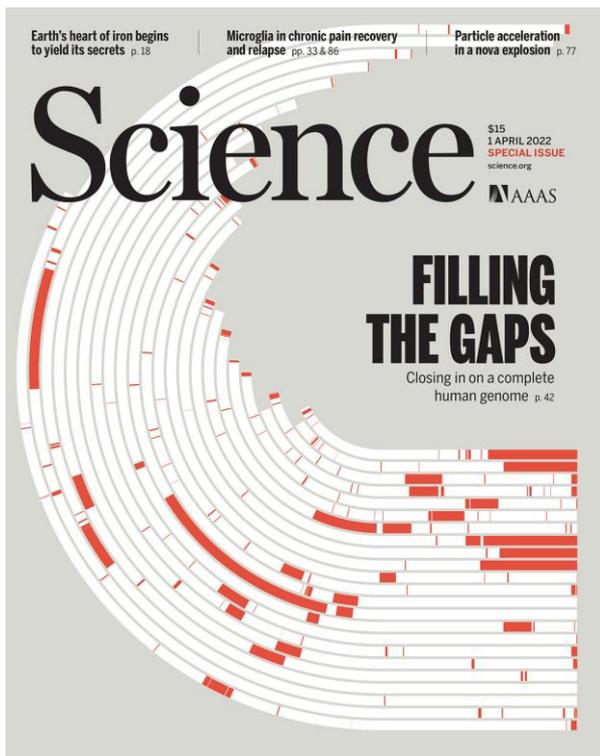
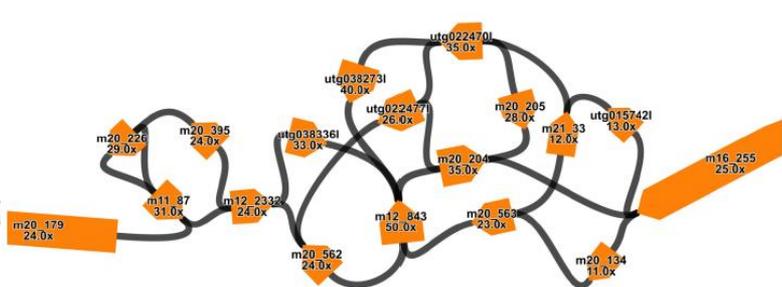
ЧТО КОНКРЕТНО

- ДНК: секвенировать весь геном

Откуда взять геном?

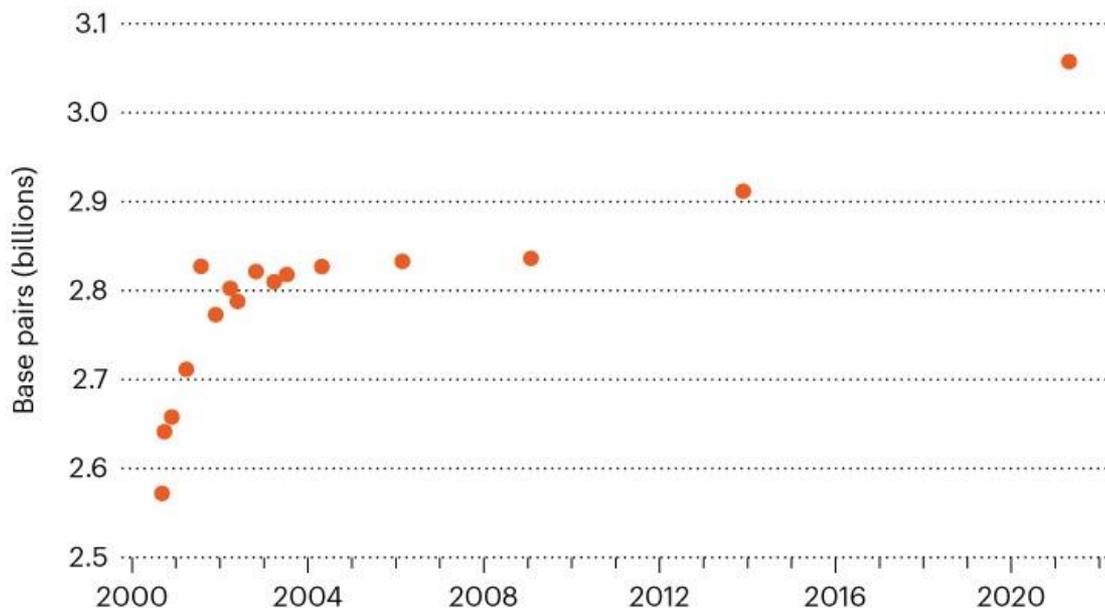
Нужно выделить ДНК

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>

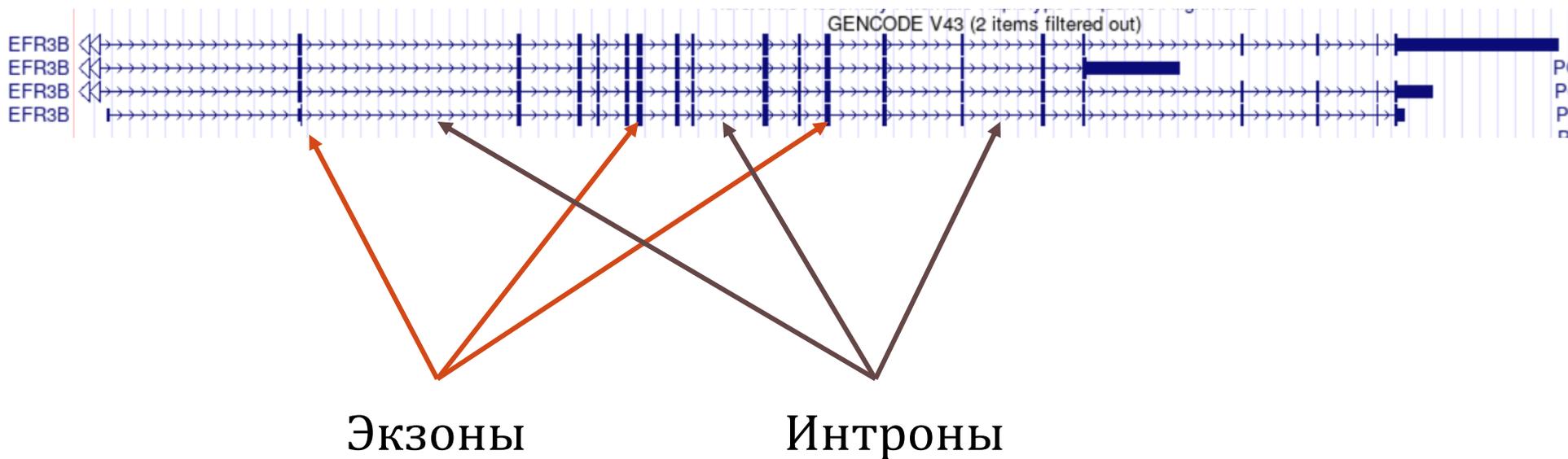


А ЕСЛИ НЕ ВЕСЬ ГЕНОМ?

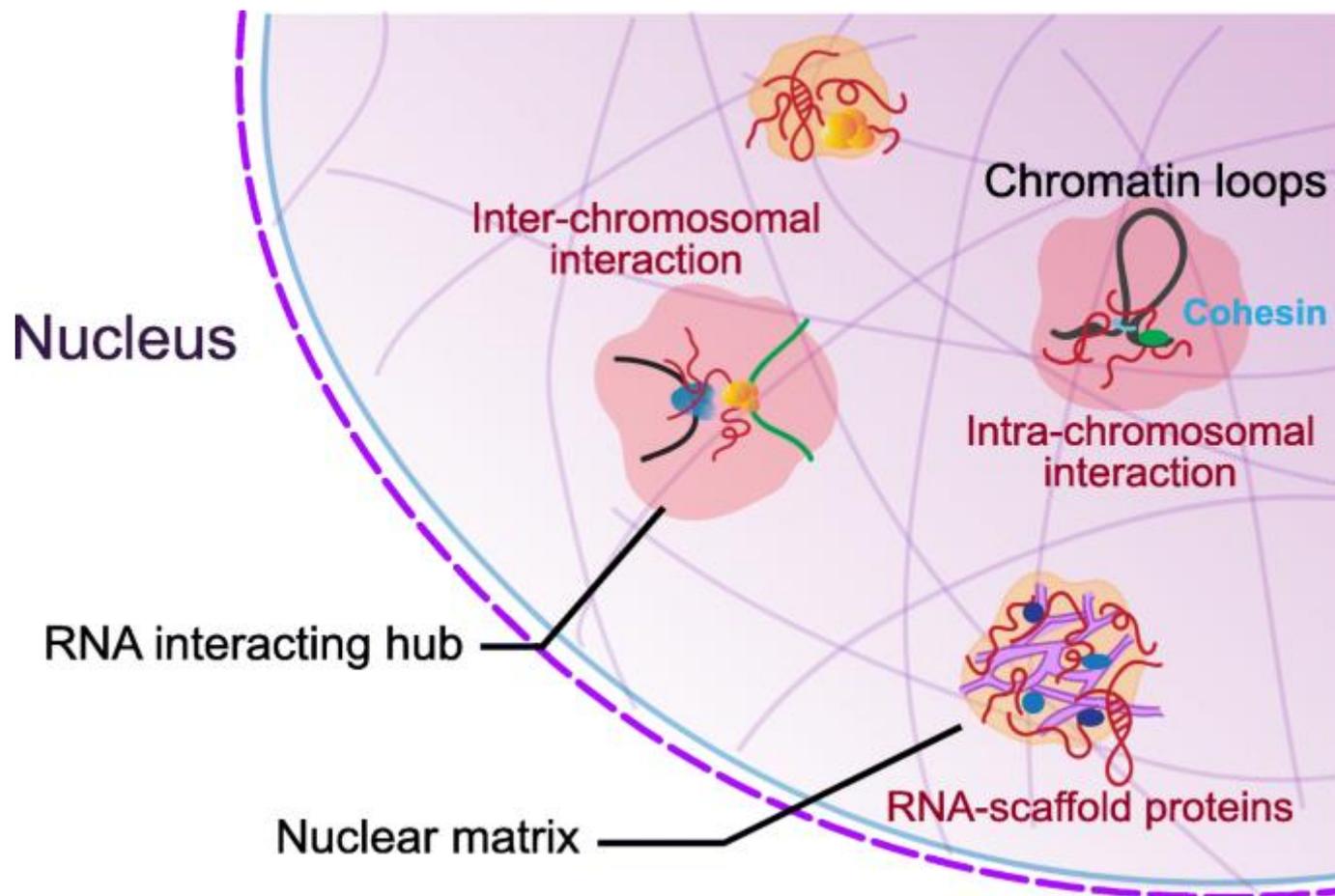
- Какие фрагменты ДНК есть смысл секвенировать?

ЭКЗОМ

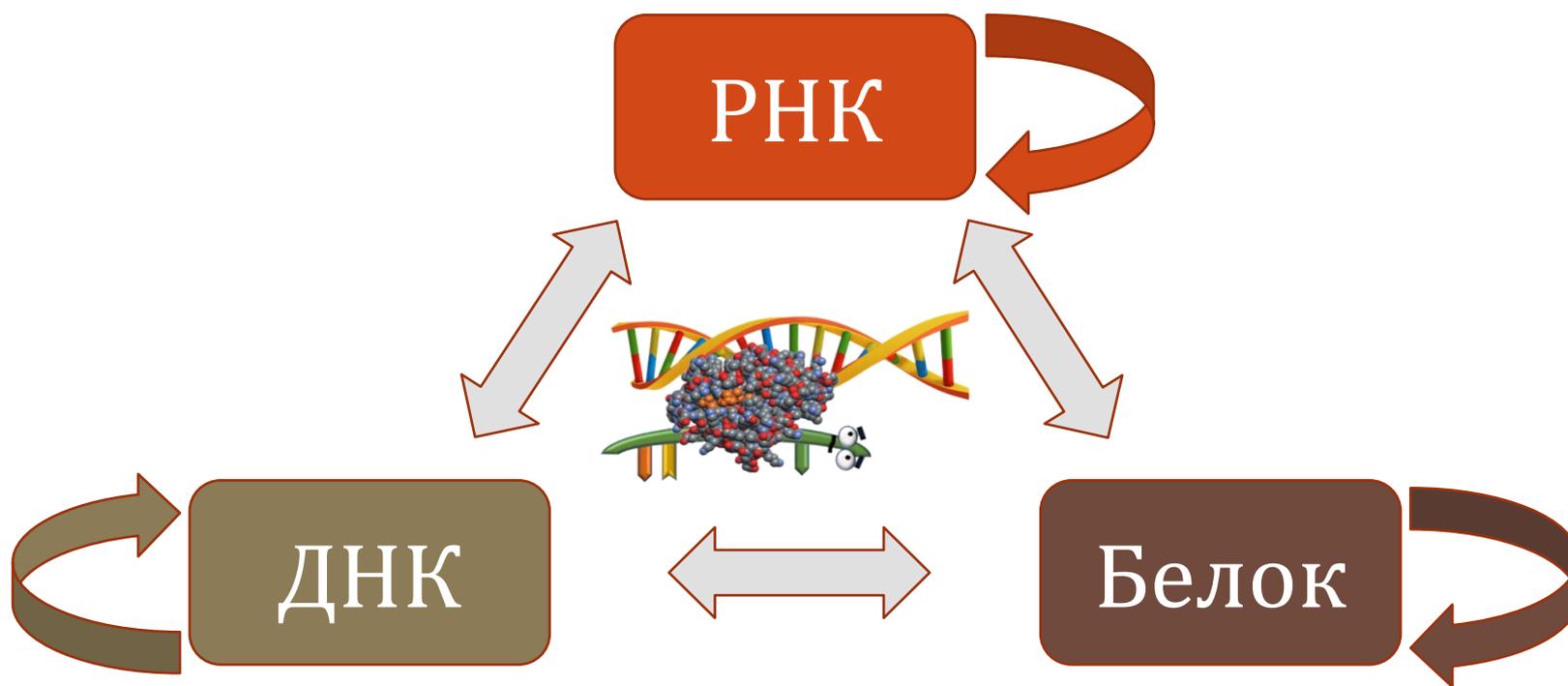
- Экзоны белок-кодирующих генов
- ~1-2% от всего генома



МАКРОМОЛЕКУЛЯРНЫЕ ВЗАИМОДЕЙСТВИЯ

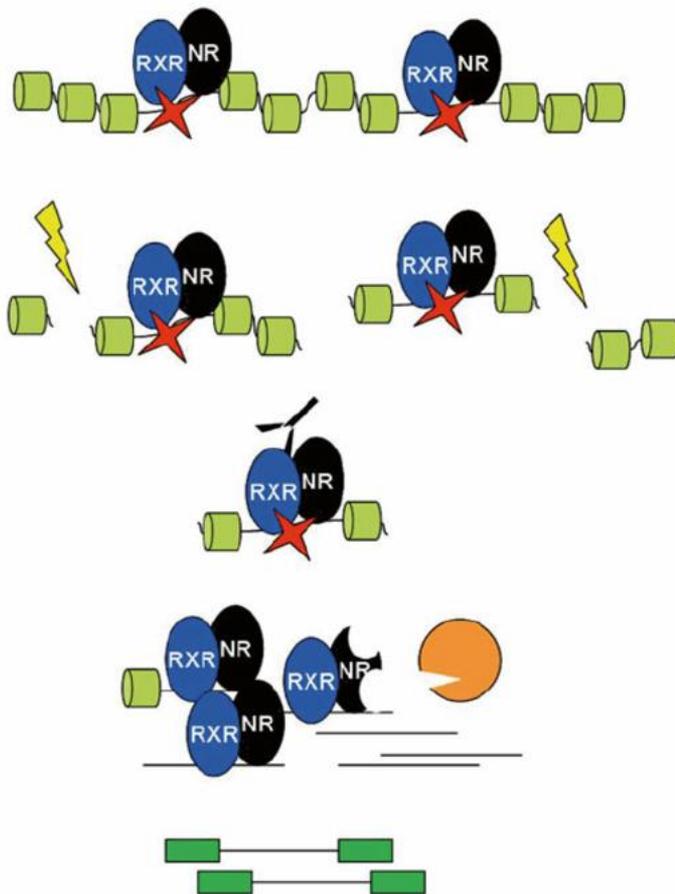


МАКРОМОЛЕКУЛЯРНЫЕ ВЗАИМОДЕЙСТВИЯ



CHIP-SEQ

- ДНК-белок



1. Chromatin crosslinking



2. Chromatin shearing



3. Immunoprecipitation



4. De-crosslinking

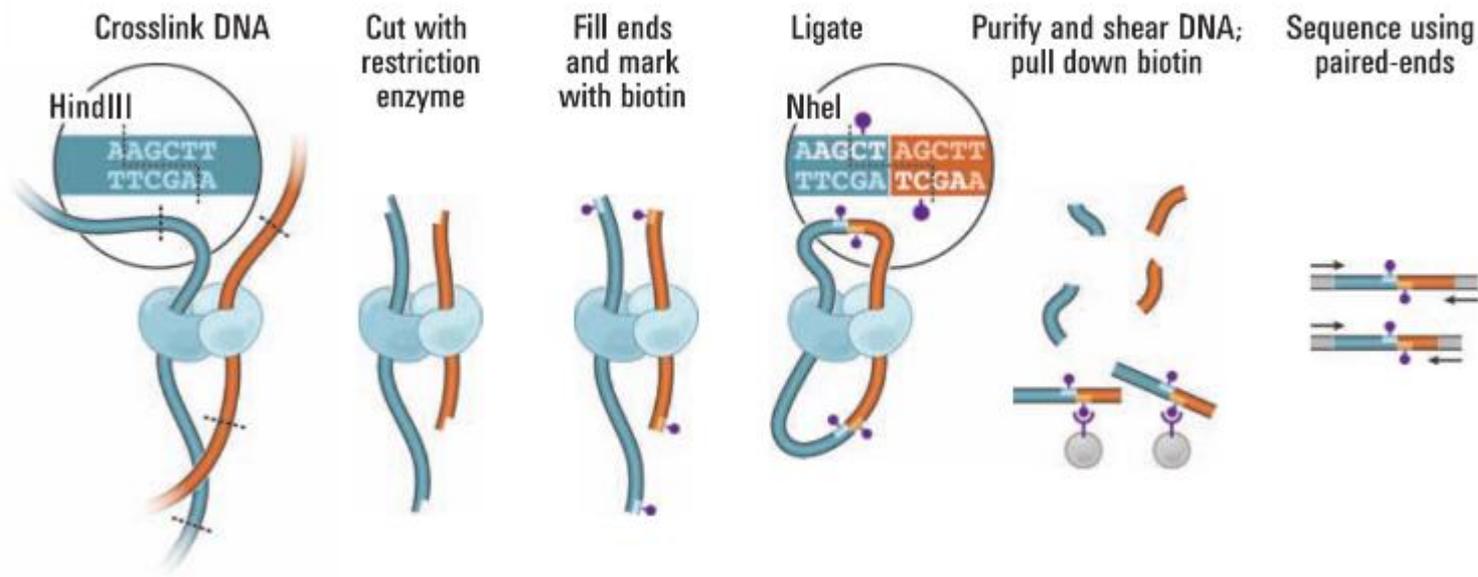


5. Library preparation



HI-C

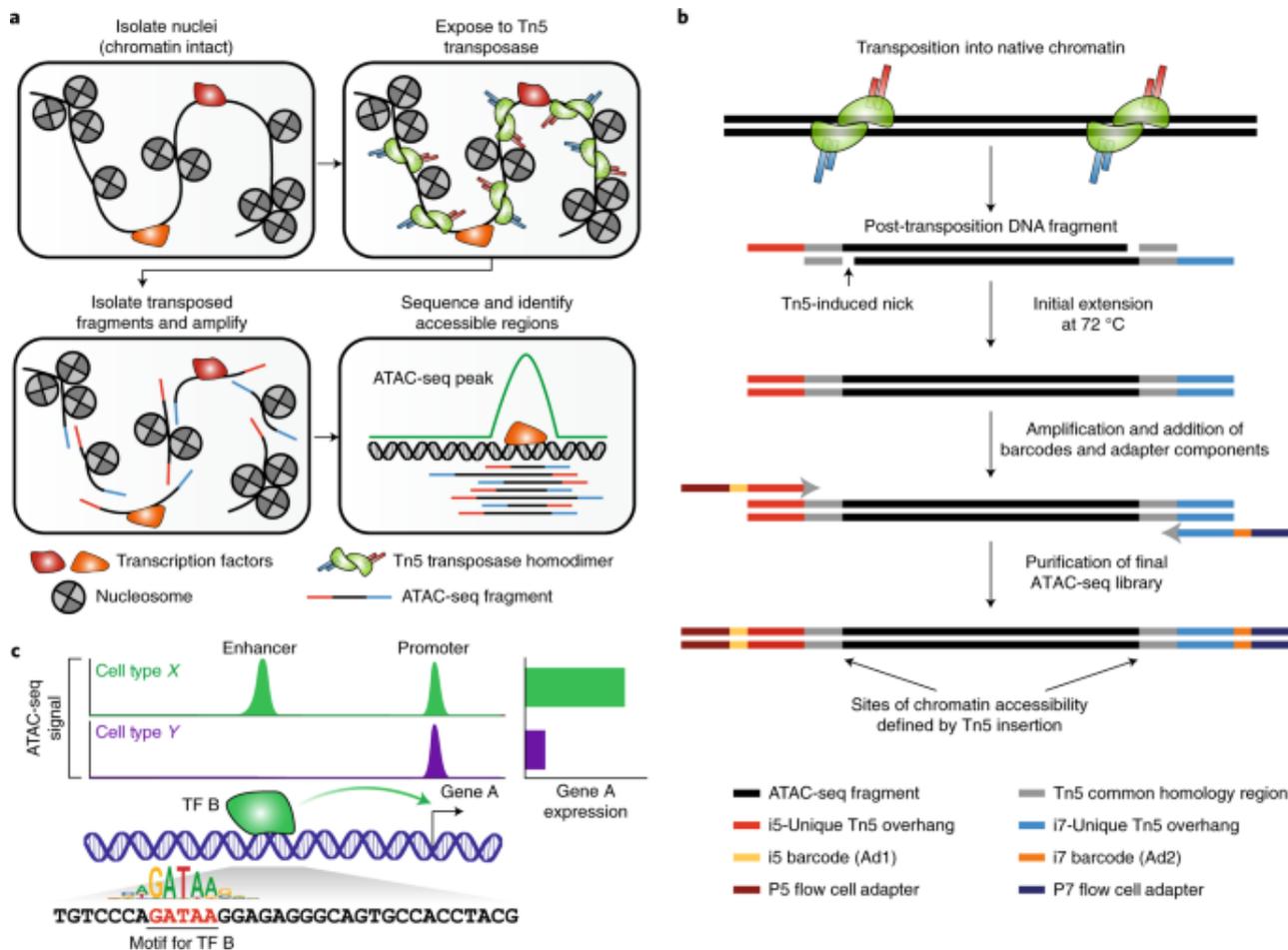
- ДНК-ДНК



Lieberman-aiden et al., 2009

ATAC-SEQ

Доступность хроматина



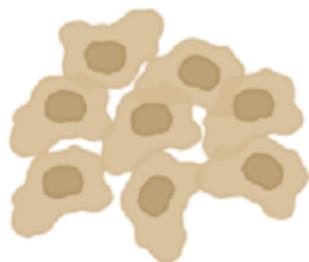


RNA-SEQ

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

RNA-SEQ

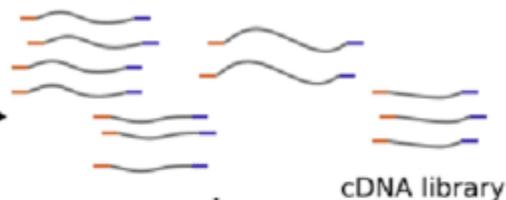
Sample of interest



Extract total RNA and enrich targets



Fragment, reverse transcribe, ligate adapters, amplify

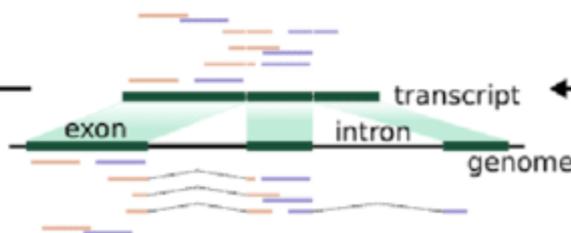


cDNA library

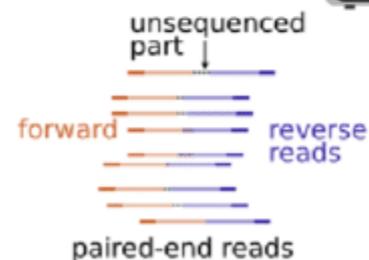
Data analysis

- differential expression
- variant calling
- annotation
- novel transcript discovery
- RNA editing
- ...

Transcriptome/genome mapping



Sequencing



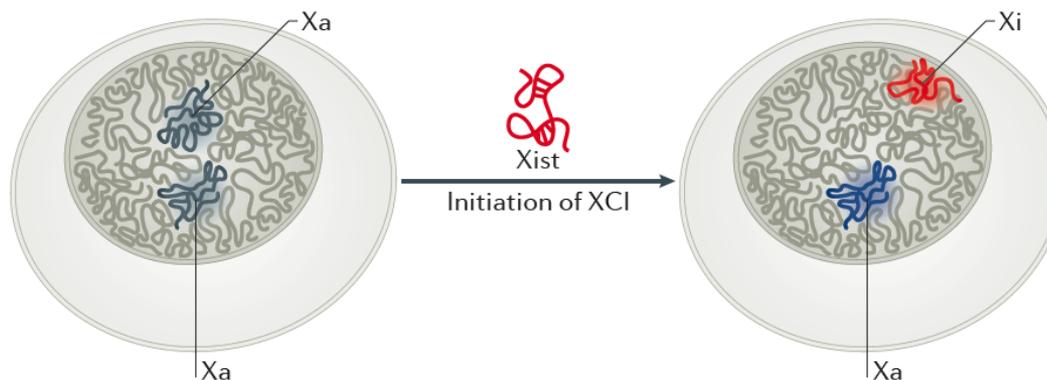
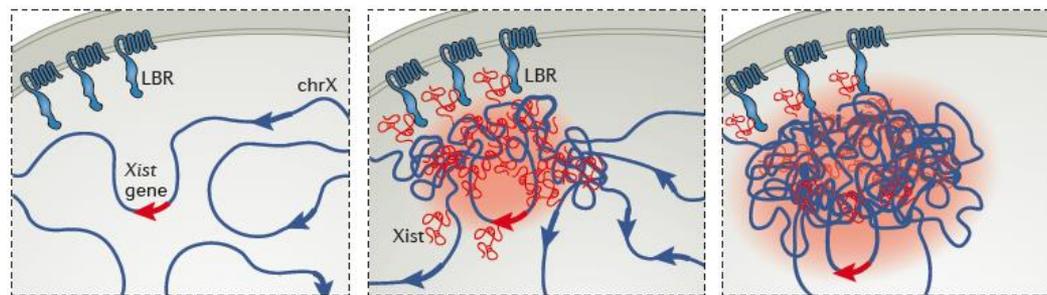


ТИПЫ РНК В КЛЕТКЕ

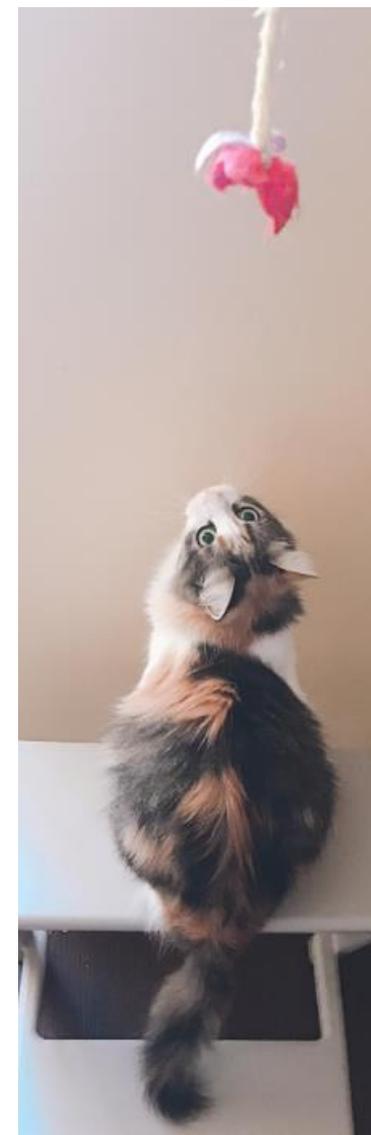
- Тотальная
- полиА
- Без фракции рРНК
- По размеру:
 - Малые (snRNA, microRNA, ...)
- По внутриклеточной локализации:
 - Ядерные
 - Цитоплазматические

ЗАЧЕМ УБИРАТЬ МРНК?

XIST – инактивация X хромосомы у самок млекопитающих



Engreitz et al., 2016



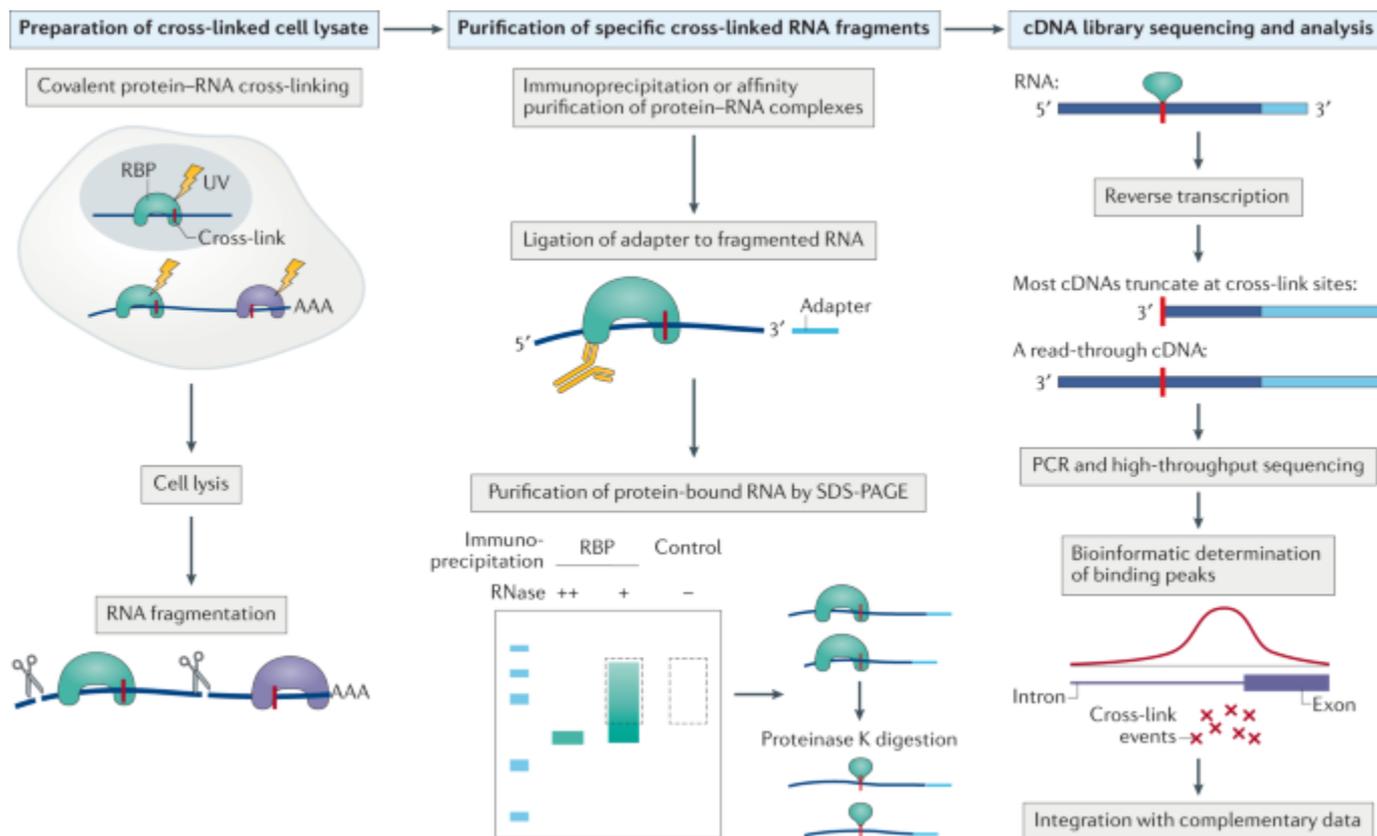


ПРОЦЕСС

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

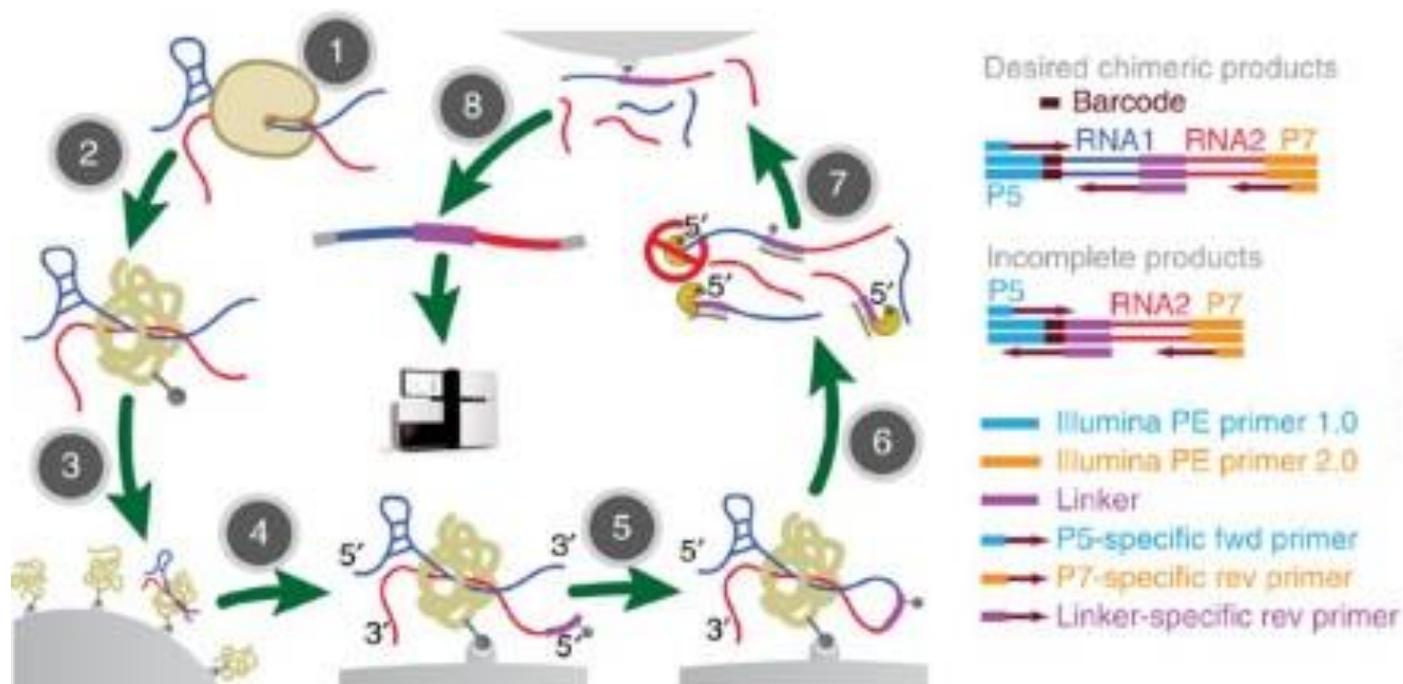
ECLIP

■ РНК-белок



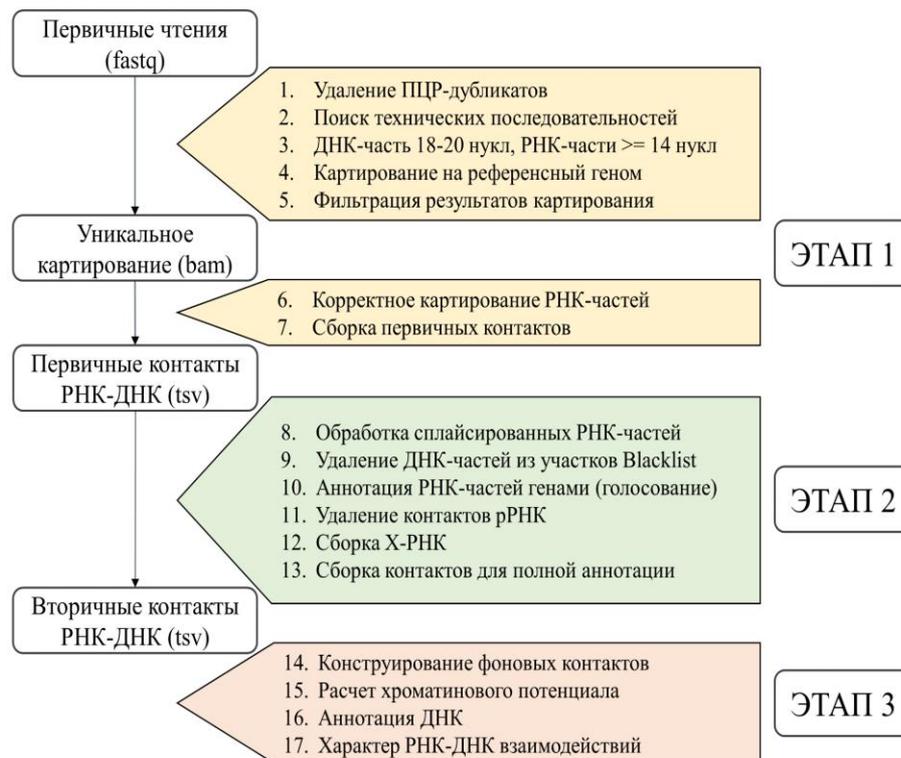
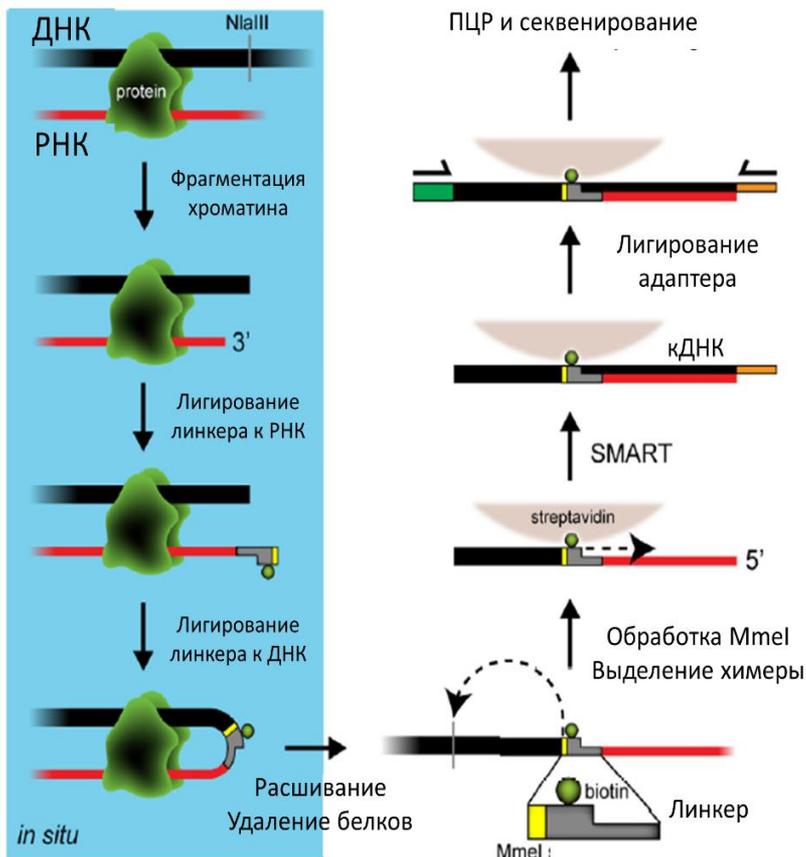
MARIO

- PHK-PHK



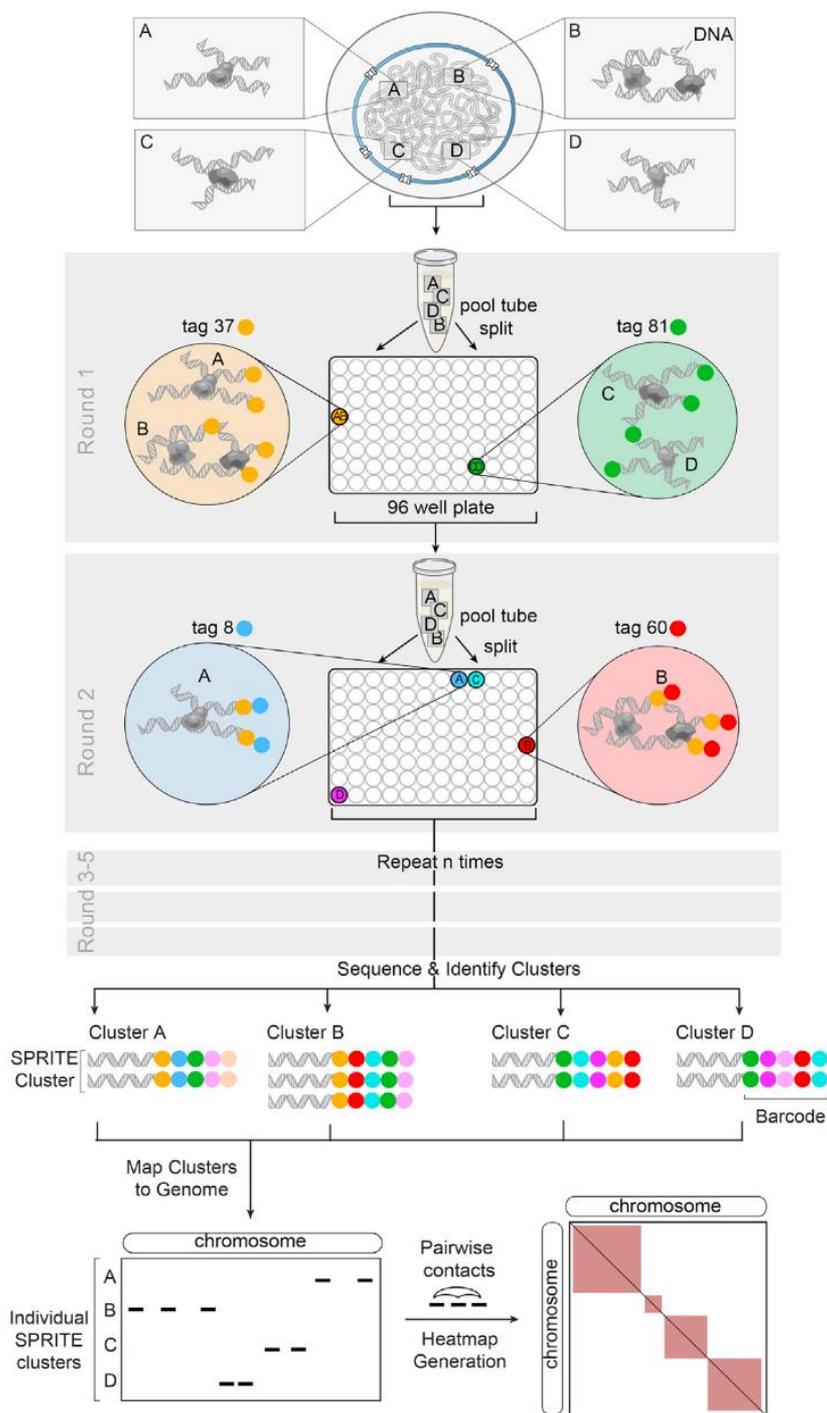
RED-C

■ РНК-ДНК

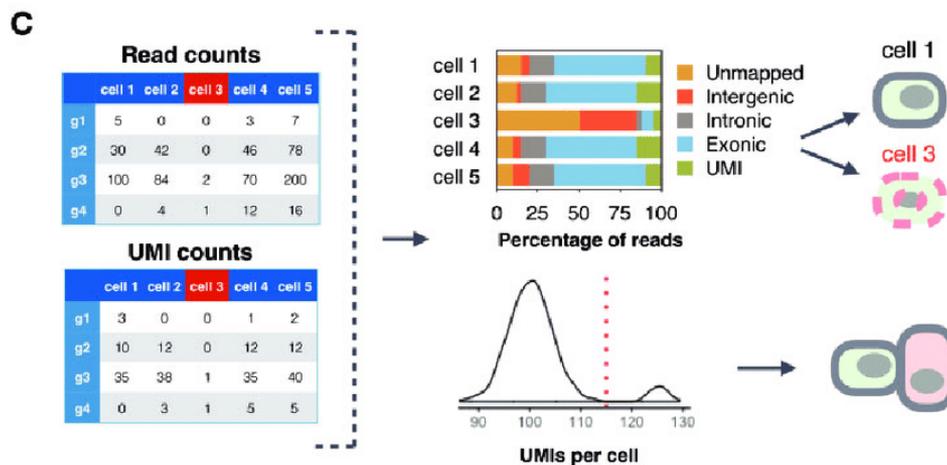
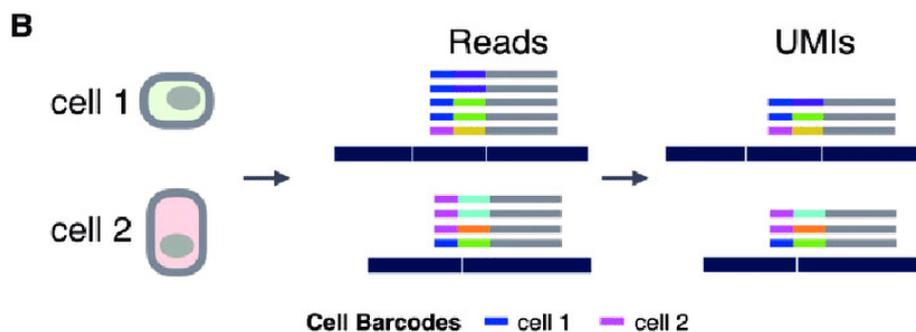
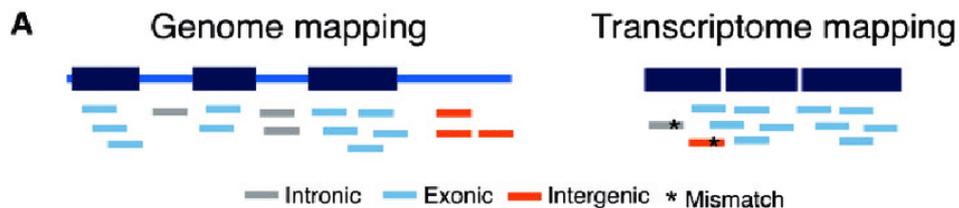


RD-SPRITE

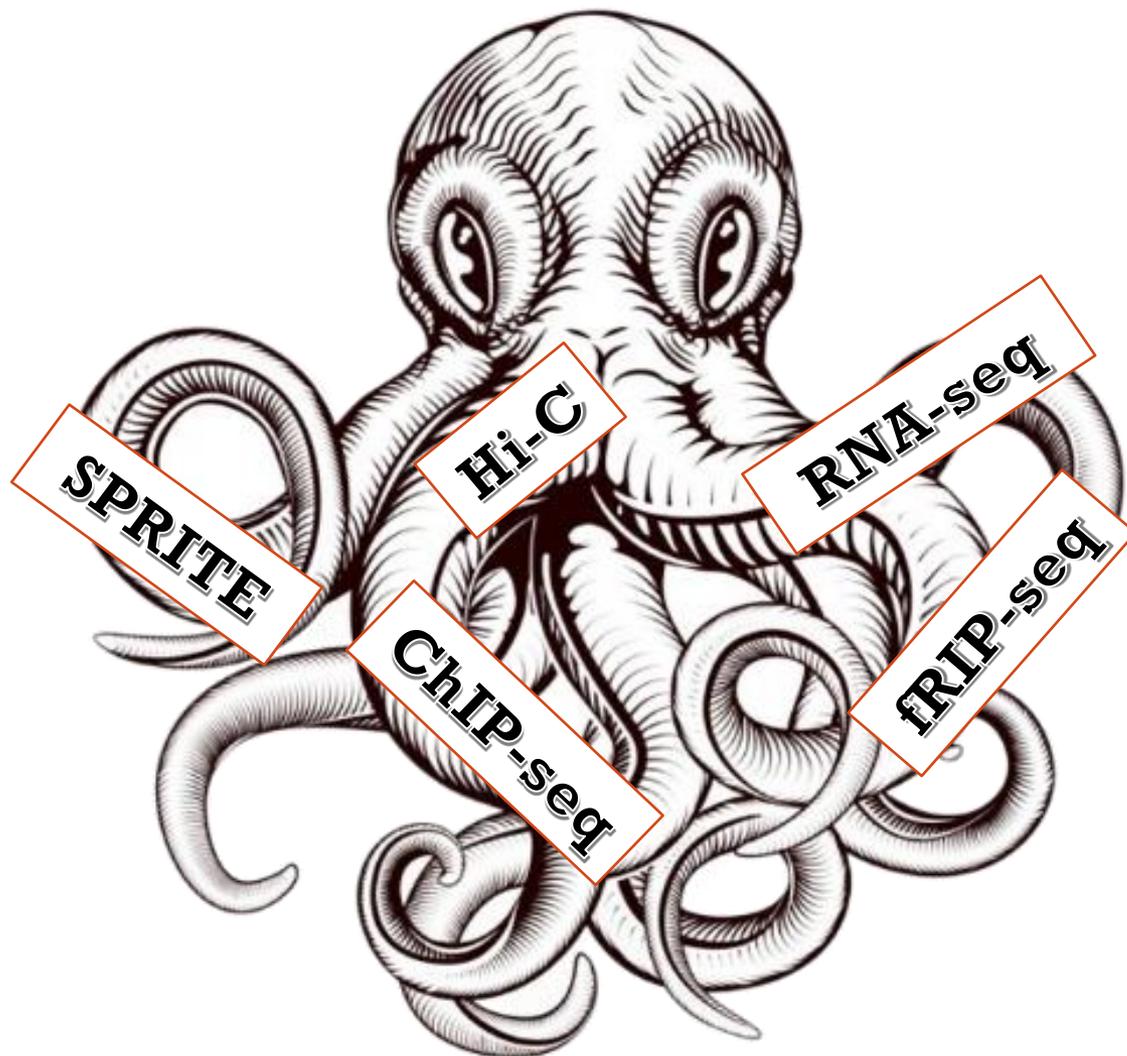
- интегратомика



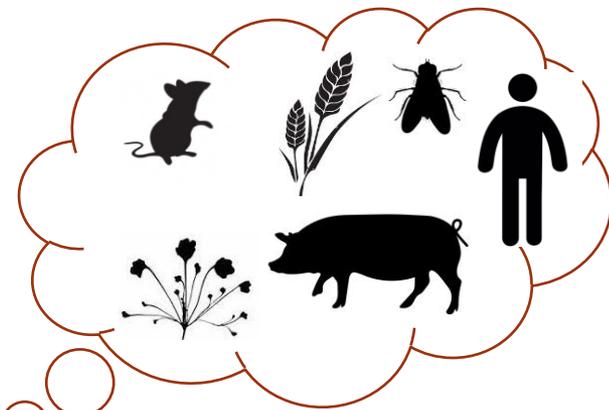
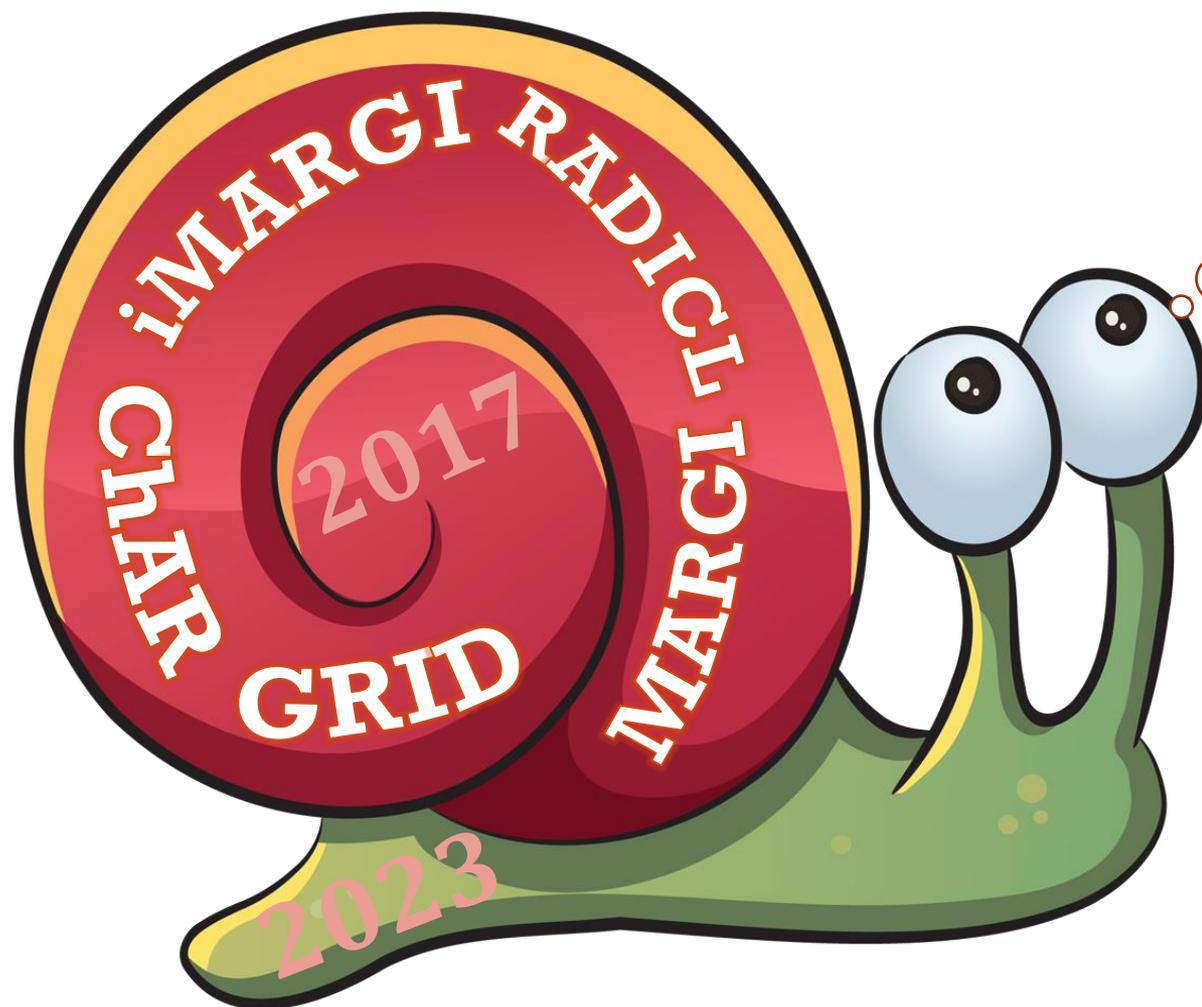
SINGLE CELL



МНОГОСЕК



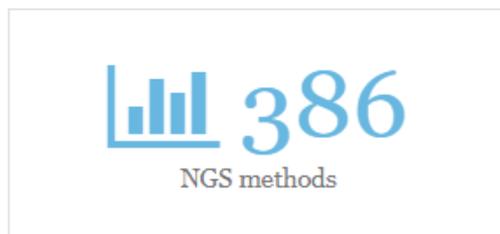
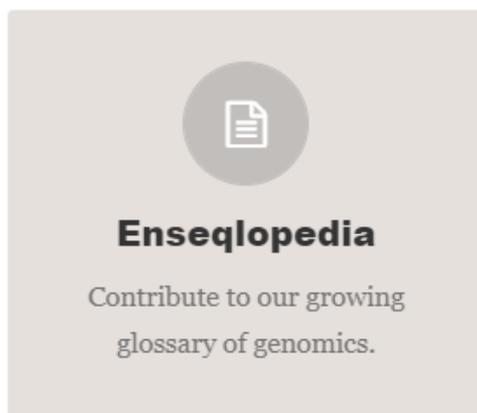
ОДНО И ТО ЖЕ...





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/>

The screenshot displays the ENCODE project website interface, organized into a grid of resource cards. The top navigation bar includes three main categories: Functional genomics (blue), Functional characterization (orange), and Encyclopedia of elements (green). Below this, the resources are categorized into three main color-coded sections: blue, grey, and red.

| Functional genomics | Functional characterization | Encyclopedia of elements |
|----------------------------|--|----------------------------|
| Rush Alzheimer's | EN-TEx | Deeply profiled cell lines |
| Protein knockdown (Degron) | Computational and integrative products | Human donors |
| ENCORE | Stem cells development | Imputed experiments |
| Immune cells | Mouse development | Reference epigenome |
| Functional genomic series | Single-cell experiments | RNA-seq |
| Region search | Encyclopedia browser | ChIP-seq experiments |

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!!!!

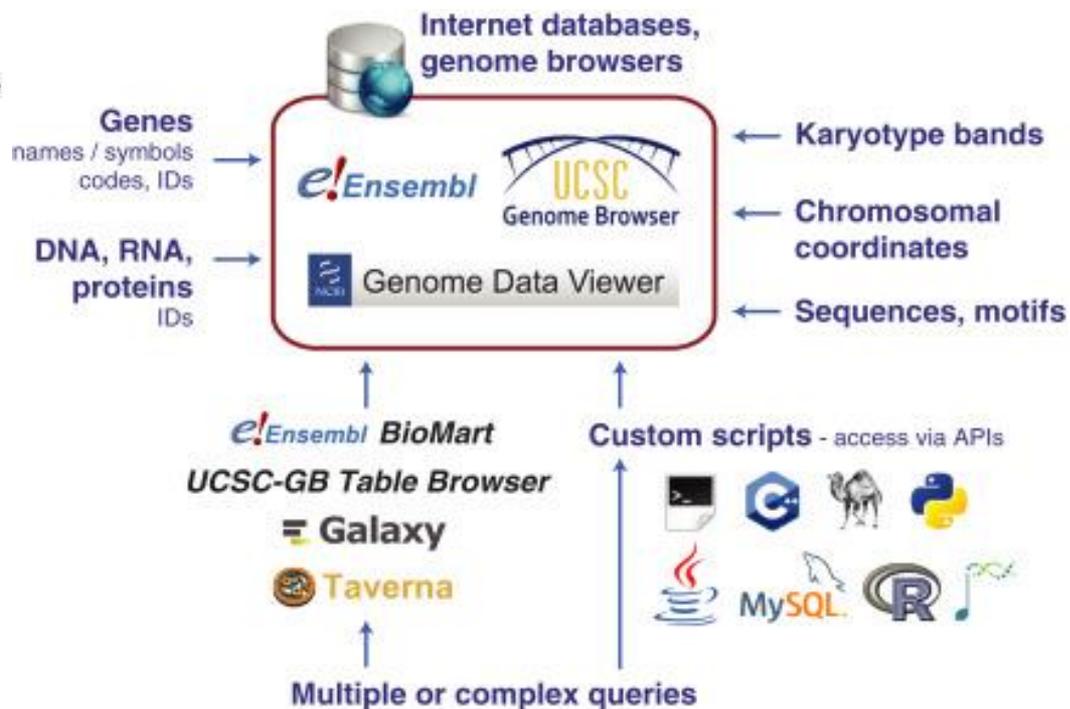
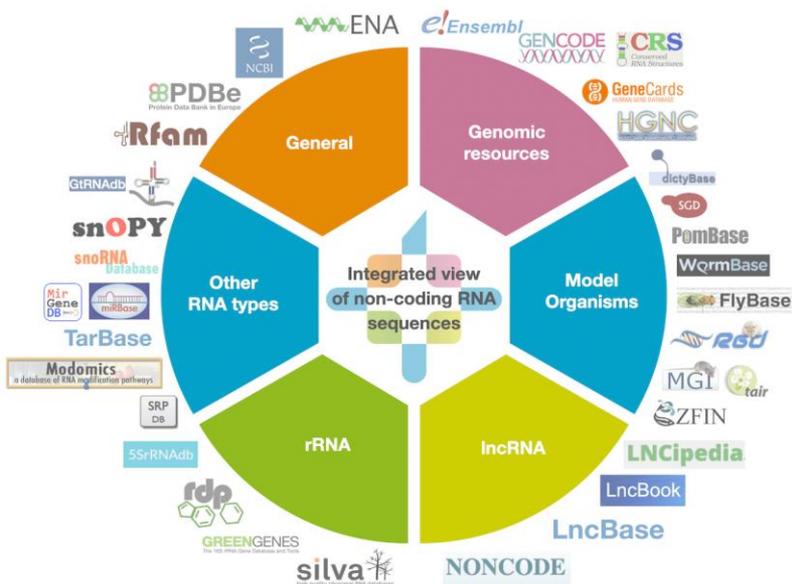


ЧТО ДАЛЬШЕ?

- 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



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 | Category ? | # of Genes |
| HGNC approved | 43,617 | Protein-coding | 21,667 |
| Disease genes | 19,871 | ncRNA genes | 292,110 |
| Hot genes | 500 | 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>

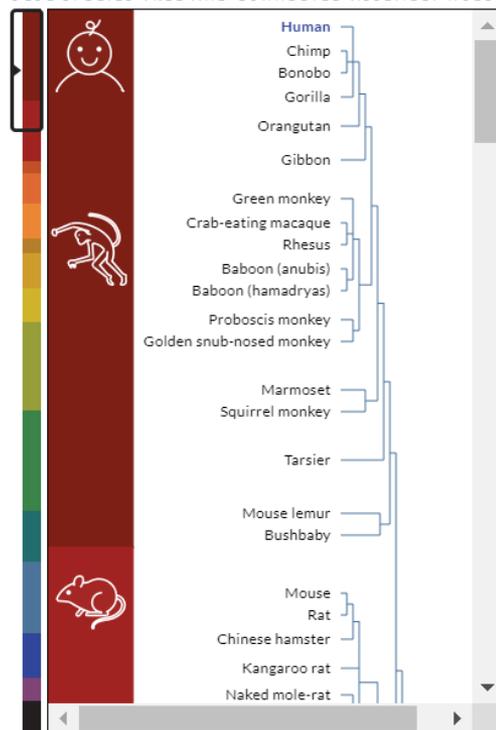


Search through thousands of genome browsers

Enter species, common name or assembly ID

[Unable to find a genome? Send us a request.](#)

UCSC SPECIES TREE AND CONNECTED ASSEMBLY HUBS



Human Assembly

Dec. 2013 (GRCh38/hg38) ▾



Position/Search Term

Enter position, gene symbol or search terms

Current position: chr2:25,160,915-25,168,903 [↗](#)

Human Genome Browser - hg38 assembly

UCSC Genome Browser assembly ID: hg38

Sequencing/Assembly provider ID: Genome Reference Consortium

Assembly date: Dec. 2013 initial release; June 2022 patch release

Assembly accession: [GCA_000001405.29](#)

NCBI Genome ID: 51 (Homo sapiens (human))

NCBI Assembly ID: [GCF_000001405.40](#) (GRCh38.p14, GCA_000001405.29)

BioProject ID: [PRJNA31257](#)

Search the assembly:

- **By position or search term:** Use the "position or search term" marker names; or keywords from the GenBank description or [information](#).
- **By gene name:** Type a gene name into the "search term" box, [information](#).
- **By track type:** Click the "track search" button to find Genom

Download sequence and annotation data:

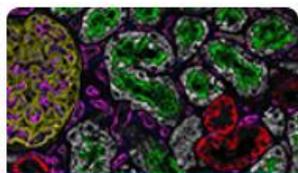
- **Using rsync** (recommended)
- **Using HTTP**
- **Using FTP**
- **Data use conditions and restrictions**
- **Acknowledgments**

Assembly Details

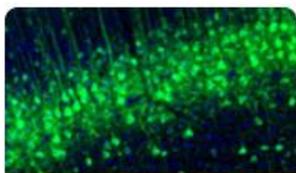
The GRCh38 assembly is the first major revision of the human genome source for human genome assembly data submitted to GenBank. It is a version confusion. Hence, the GRCh38 assembly is referred to as "GRCh38.p14"

HUMAN PROTEIN ATLAS

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



TISSUE



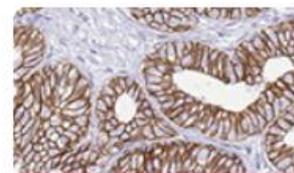
BRAIN



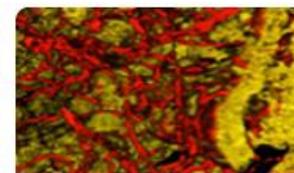
SINGLE CELL TYPE



TISSUE CELL TYPE



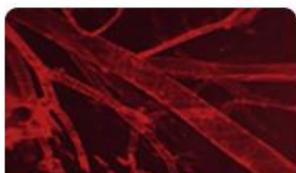
PATHOLOGY



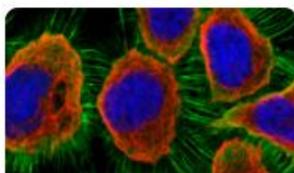
DISEASE



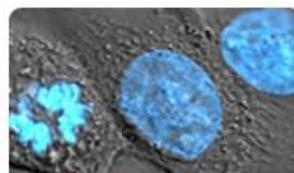
IMMUNE CELL



BLOOD PROTEIN



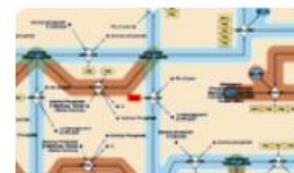
SUBCELLULAR



CELL LINE



STRUCTURE



METABOLIC

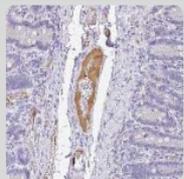
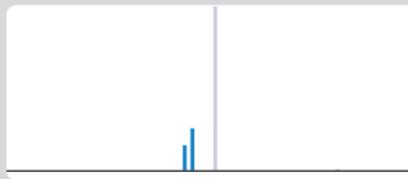
HUMAN PROTEIN ATLAS

APOB



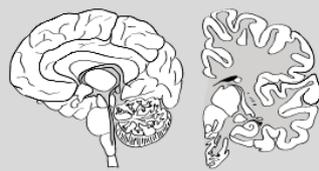
- PROTEIN SUMMARY
- SECTION OVERVIEW
- RNA DATA
- ANTIBODY DATA Y

TISSUE¹

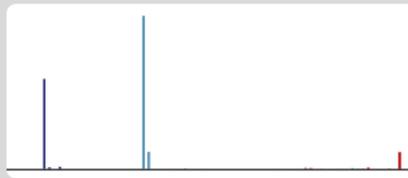
| | |
|--|-----------------------------------|
| Tissue expression cluster ¹ | Liver - Metabolism (mainly) |
| Tissue specificity ¹ | Group enriched (Intestine, liver) |
| Protein expression ¹ | Distinct positivity in plasma. |

BRAIN¹



| | |
|---|--------------|
| Human regional specificity ¹ | Not detected |
| Pig regional specificity ¹ | Not detected |
| Mouse regional specificity ¹ | Not detected |

SINGLE CELL TYPE¹

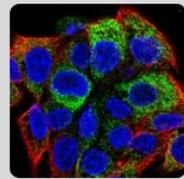
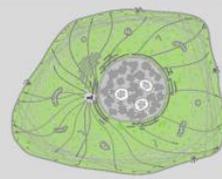



| | |
|--|--|
| Single cell type expression cluster ¹ | Hepatocytes - Metabolism (mainly) |
| Cell type specificity ¹ | Group enriched (Hepatocytes, Proximal enterocytes) |

TISSUE CELL TYPE¹

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

SUBCELLULAR¹

| | |
|----------------------------|------------------------------------|
| Main location ¹ | 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 причинам (перепредставлены)



STRING

- <https://string-db.org/>

Version: 11.5

LOGIN | REGISTER | SURVEY

STRING

Search | Download | Help | My Data

Protein by name >

Multiple proteins >

Proteins by sequences >

Proteins with Values/Ranks >

Protein families ("COGs") >

Pathway / Process / Disease **New** >

Annotate your proteome **New** >

Organisms >

Examples >

Random entry >

SEARCH

Single Protein by Name / Identifier

Protein Name: (examples: #1 #2 #3)

Organisms:

auto-detect ▼

[Advanced Settings](#)

SEARCH

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.

 **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

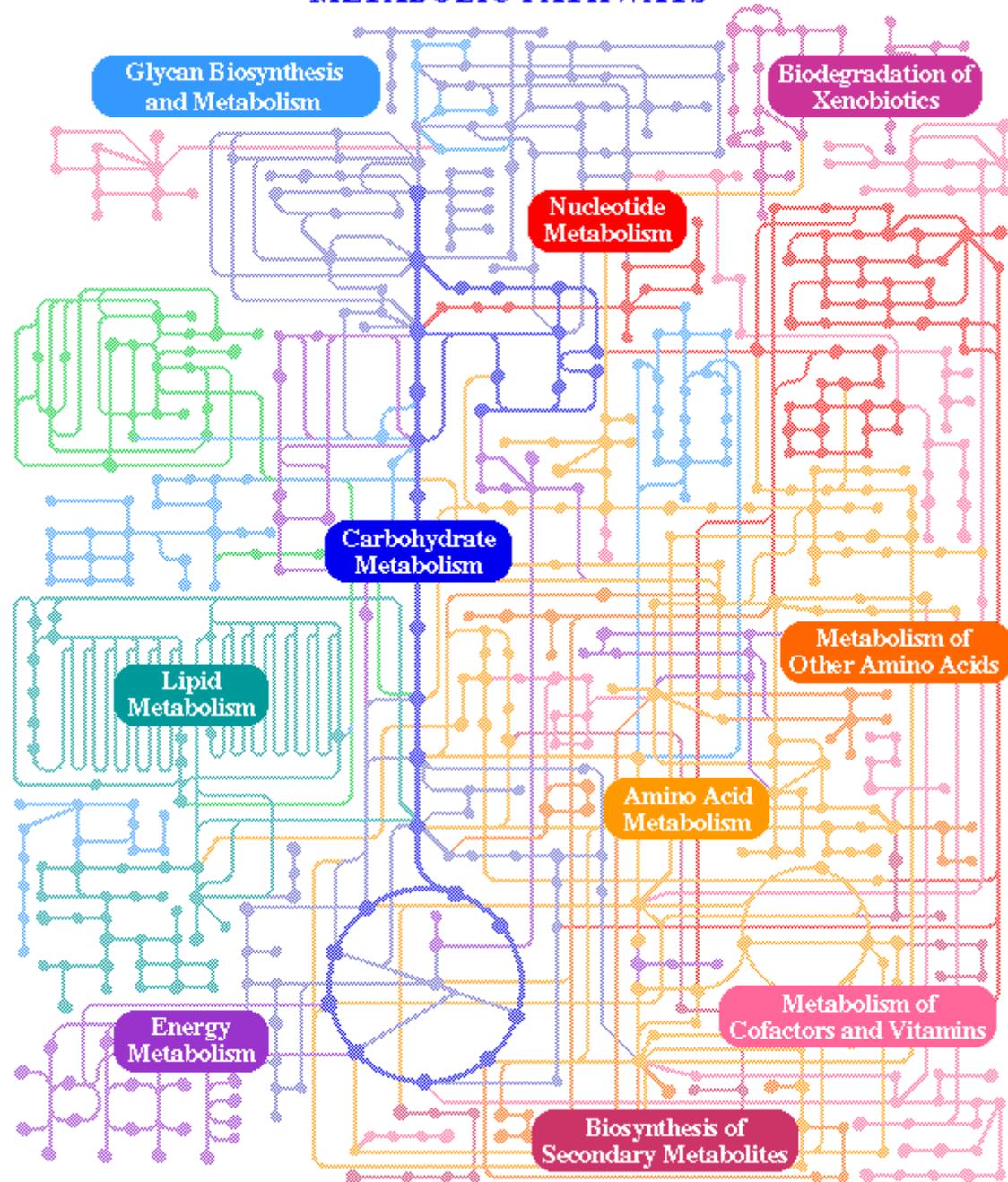
Search references cited in KEGG

Number of references (2023/4/1)

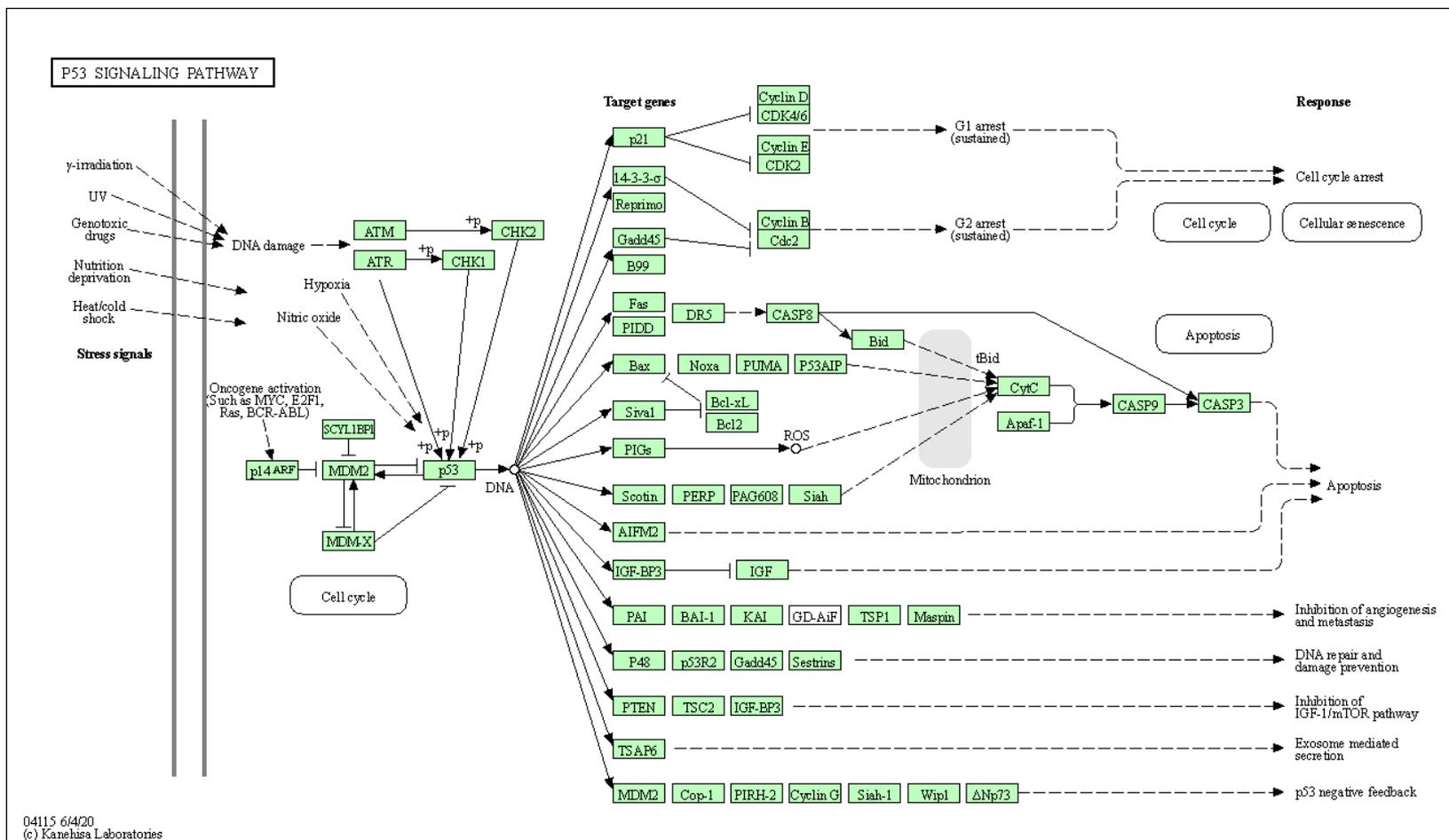
| | | | | | | | | | |
|-------|--------|---------|-------|--------|--------|----------|--------|---------|--------|
| total | 69,123 | pathway | 6,518 | ko | 27,728 | glycan | 919 | network | 2,393 |
| | | brite | 444 | genome | 5,994 | reaction | 2,020 | variant | 1,366 |
| | | module | 1,089 | agenes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

KEGG

METABOLIC PATHWAYS

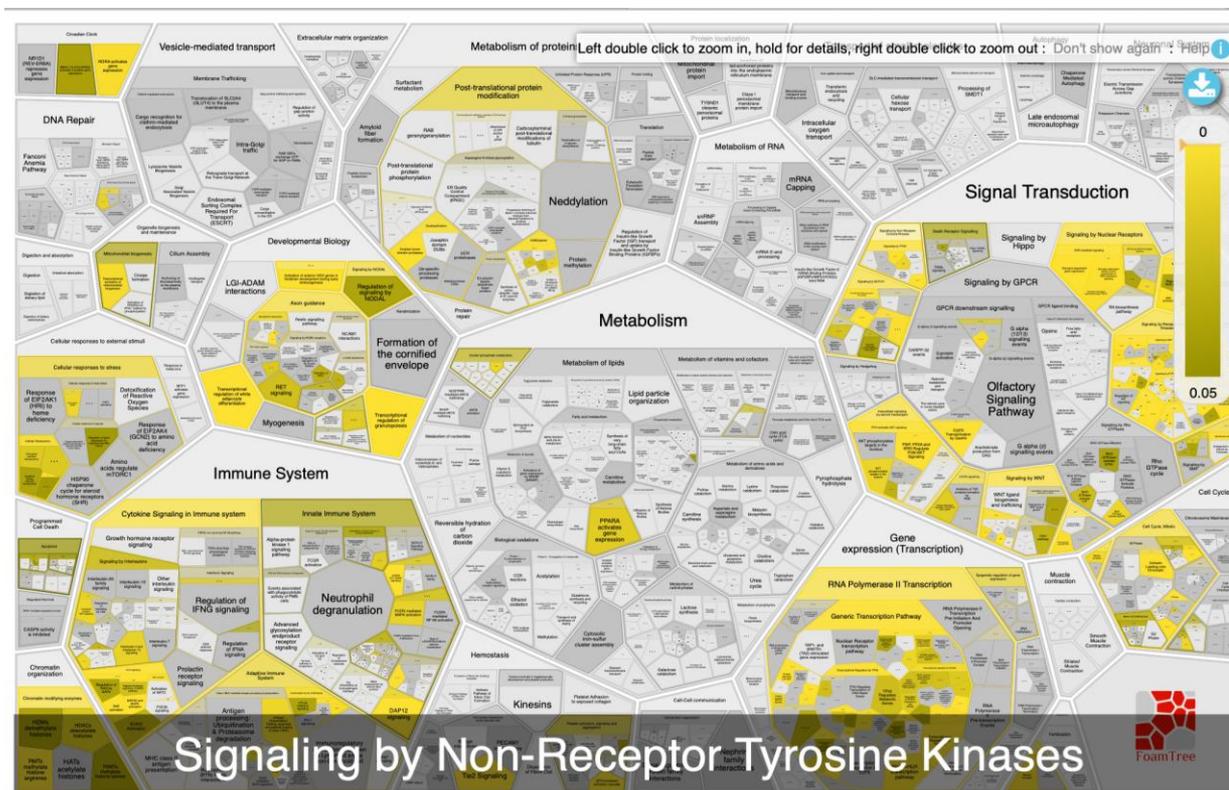


KEGG



REACTOME

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





GENE ONTOLOGY

- <http://geneontology.org/>

THE GENE ONTOLOGY RESOURCE

The mission of the GO Consortium is to develop a comprehensive, **computational model of biological systems**, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life.

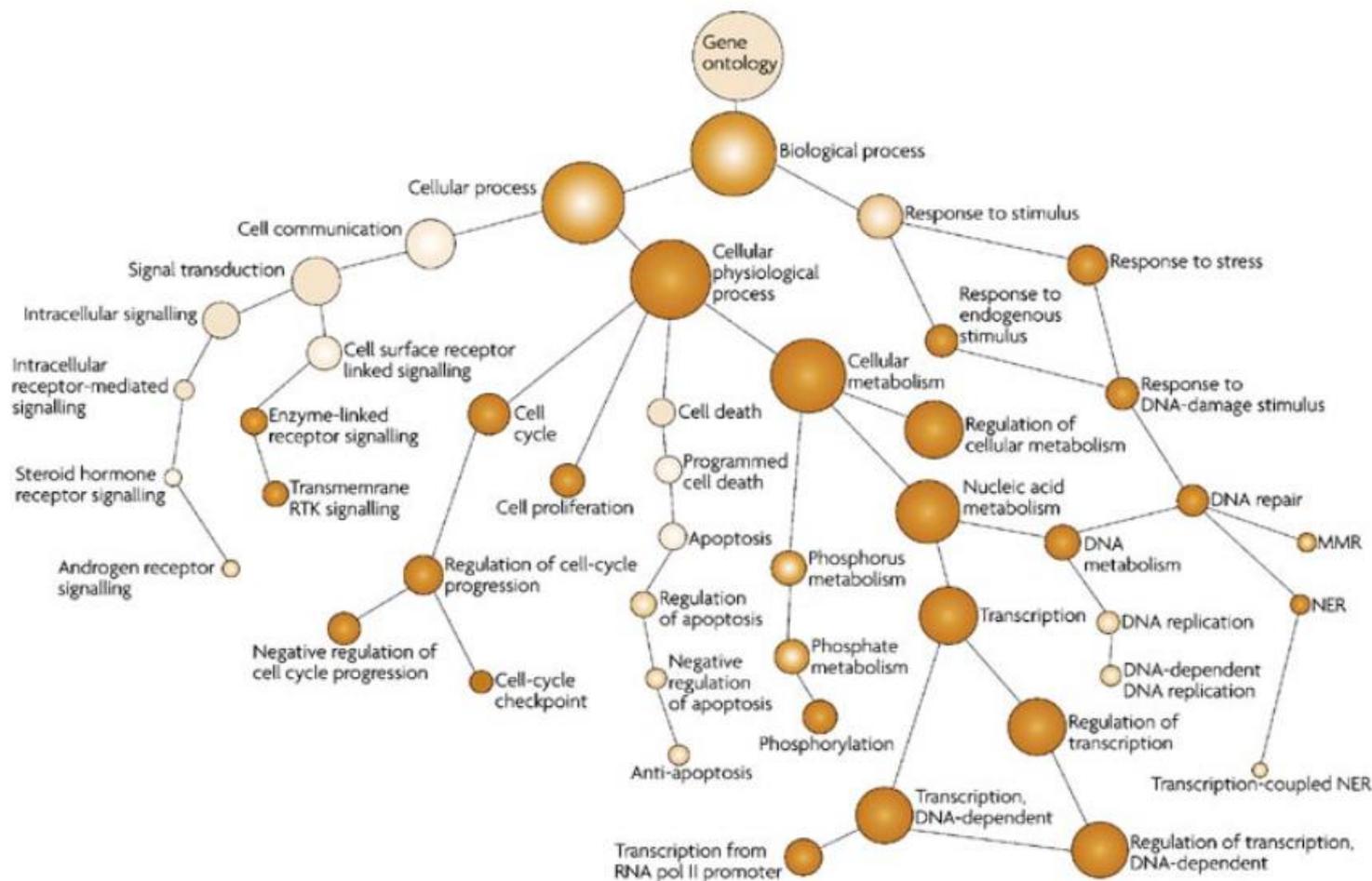
The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.

Search GO term or Gene Product in AmiGO ...



Any Ontology Gene Product

GENE ONTOLOGY





PUBMED

MEDLINE PubMed Production Statistics

| | <u>FY2022</u> | <u>FY2021</u> | <u>FY2020</u> | <u>FY2019</u> | <u>FY2018</u> |
|---|---------------|---------------|---------------|---------------|---------------|
| 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 |

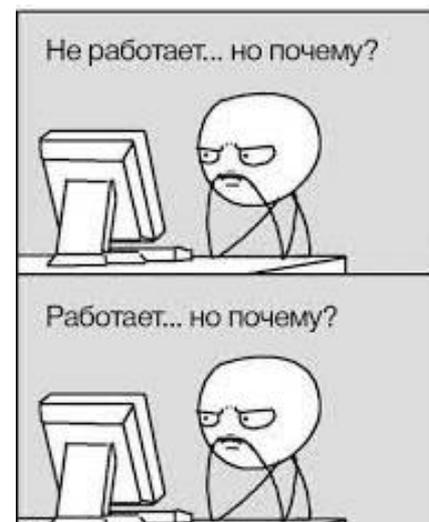


ЧЕМ ЗАНИМАЕТСЯ БИОИНФОРМАТИКА СЕГОДНЯ?

- 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...



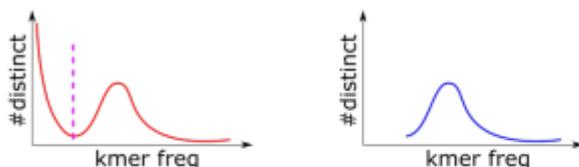
Давайте по порядку

БУДНИ БИОИНФОРМАТИКА

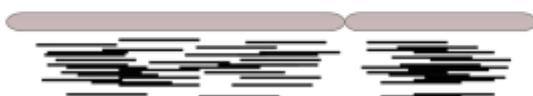
Read clipping



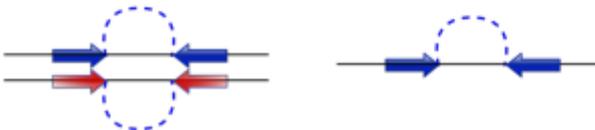
Read error correction



Read mapping



Duplicate removal

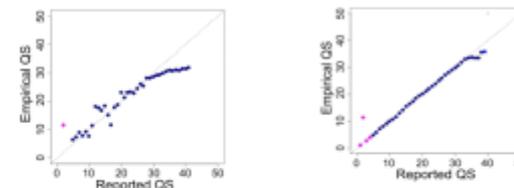


Re-alignment

TTAAAAAACGT
TTA-AAA--CGT
TT--AAAA-CGT

TTAAAAAACGT
TT---AAACGT
TT---AAACGT

Base Quality Adjustment



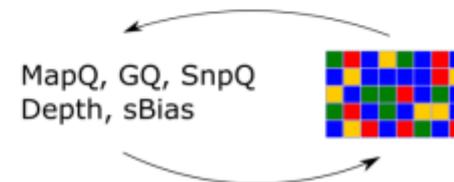
SNP calling

— A —
— A —
— A —

Chr1:2340
T → A

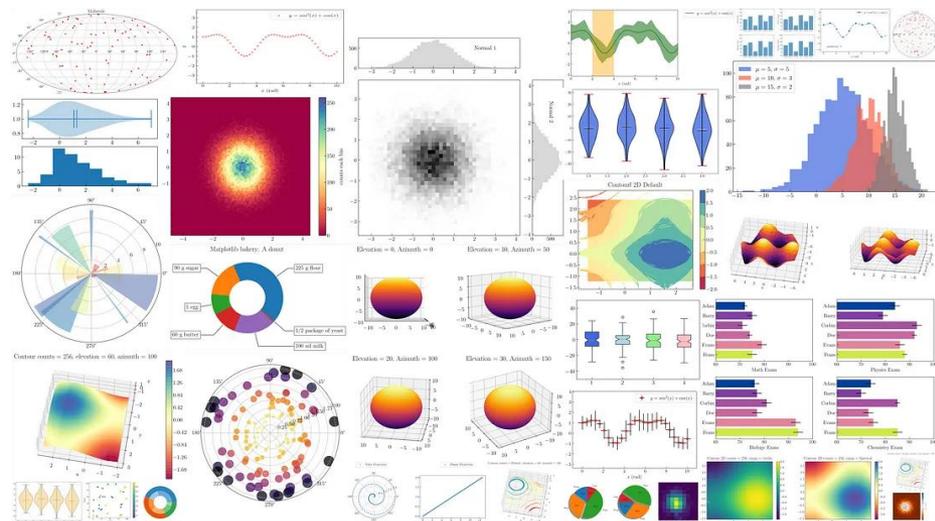
SNP filtering

Evaluation



ВИЗУАЛИЗАЦИЯ

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





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_2024_ZharikovaA(заменить).pdf**
- На первой странице документа укажите полностью свои фамилию, имя, курс и факультет
- Отправьте на почту: azharikova89@gmail.com
- Если есть вопросы – напишите мне!

