Лекция №5

Интеграция разрозненных источников информации по биомедицинской тематике

м.н.с. лаб. гликогеномики
Тийс Евгений Сергеевич,
tiys@bionet.nsc.ru

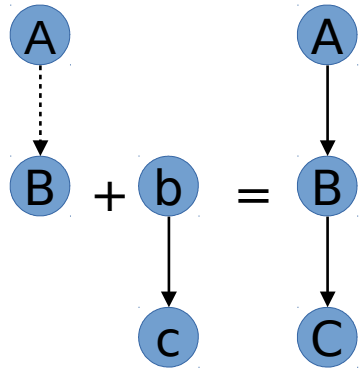
План лекции

1. Введение.
2. Обзор Интернет-ресурсов, интегрирующих информацию по биомедицинской тематике.
3. Ресурсы, интегрирующие биологическую информацию из разнородных источников и представляющие ее в виде генных сетей: ANDSystem, STRING, GeneMania, Pathway Commons.
4. Практическое применение инструментов интеграции.

План лекции

1. Введение.
2. Обзор Интернет-ресурсов, интегрирующих информацию по биомедицинской тематике.
3. Ресурсы, интегрирующие биологическую информацию из разнородных источников и представляющие ее в виде генных сетей: ANDSystem, STRING, GeneMania, Pathway Commons.
4. Практическое применение инструментов интеграции.

Цель интеграции биомедицинских данных – понимание и моделирование мехнизмов биологических процессов



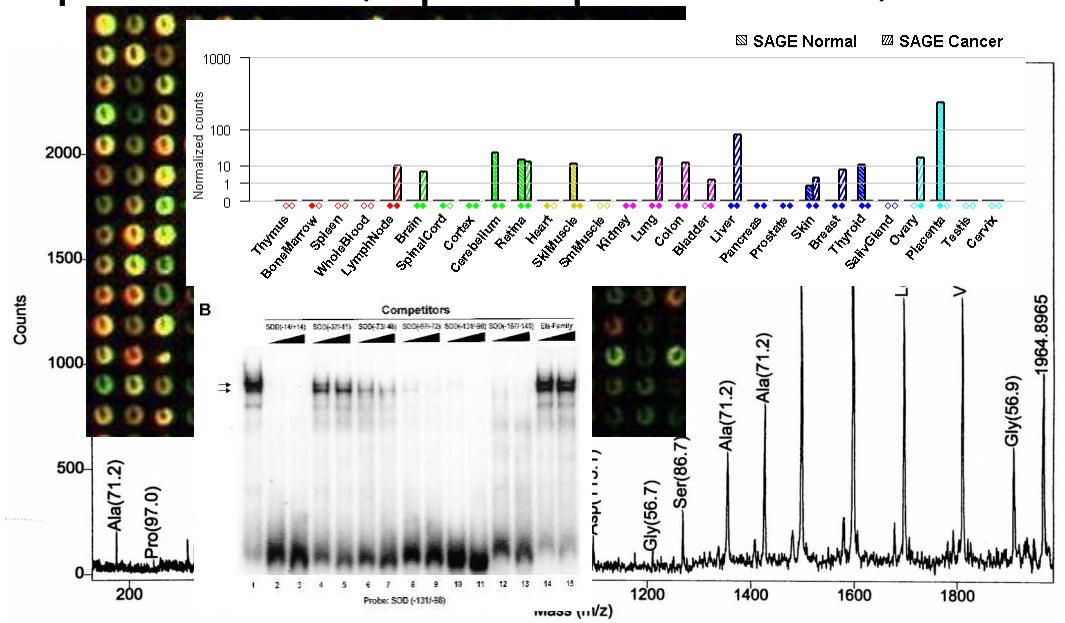
Преимущества интеграции биологических данных:

- Единый формат – возможность автоматизации обработки и сбор статистики.
- Единый способ графического и текстового представления – упрощение осмысления, отсутствие штрафов на переключения внимания.
- Возможность выполнения поискового запроса ко всем данным
- Установление эквивалентности объектов из различных источников с учетом синонимии, что позволяет избежать дублирования информации
- Выявление противоречий, ошибок, пробелов в информации
- Возможность выявления наиболее достоверной информации

По Мигинский Д. С. и др. Технология семантической интеграции баз данных в системной биологии //Вычислительные технологии. – 2008. – Т. 13. – №. 6.

Процесс интеграции

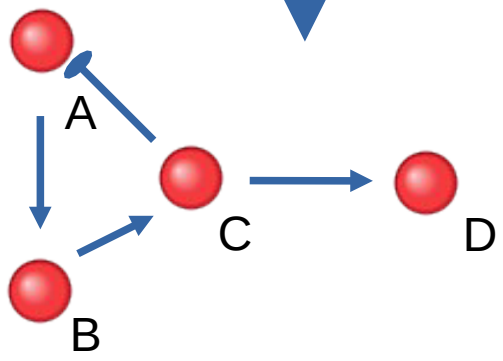
Экспериментальные данные:
Протеомные, транскриптомные, метаболомные



базы данных

PubMed

Скорость появления публикаций **1,4** в минуту
http://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html



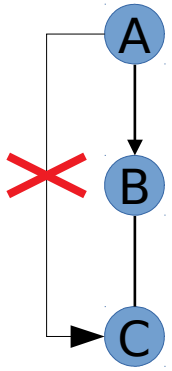
Генная сеть – группа координированно функционирующих генов, взаимодействующих друг с другом как через свои первичные продукты (РНК и белки), так и через разнообразные метаболиты и другие вторичные продукты функционирования генных сетей, которая контролирует какой-либо фенотипический признак организма.

Интегрируемая информация

Компоненты генной сети:

1) группа координированно экспрессирующихся генов, составляющая ядро сети;
2) белки, кодируемые этими генами;
3) низкомолекулярные компоненты (гормоны и другие сигнальные молекулы, энергетические компоненты, метаболиты);
4) связи между участниками сети (в том числе отрицательные и положительные обратные связи).

По Колчанову Н.А. с соавт. ГЕННЫЕ СЕТИ // Вавиловский журнал генетики и селекции, 2013, Т. 17:4/2.

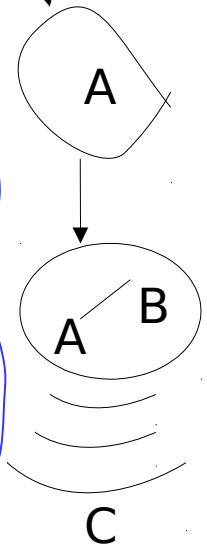


“A gene network is a mixed graph G : =(V,U,D) over a set V of nodes, corresponding to gene activities, with unordered pairs U, the undirected edges, and ordered pairs D, the directed edges.”

Pinna A., Soranzo N., De La Fuente A. From knockouts to networks: establishing direct cause-effect relationships through graph analysis //PloS one. – 2010. – Т. 5. – №. 10. – С. e12912.

“Gene network is a graphical illustration for exploring the functional linkages and the potential coordinate regulations of genes.”

Wang et al.Gene Network Exploration of Crosstalk between Apoptosis and Autophagy in Chronic Myelogenous Leukemia // BioMed Research International, 2014.



“A gene network is a directed labelled graph, where each node represents a gene and each arc represents a relation between the genes.”

Rung J. et al. Building and analysing genome-wide gene disruption networks //Bioinformatics. – 2002. – Т. 18. – №. suppl 2. – С. S202-S210.

“A gene network is a collection of effective interactions, describing the multiple ways through which one gene affects all the others to which it is connected.”

Zhu Y., Pan W., Shen X. Support vector machines with disease-gene-centric network penalty for high dimensional microarray data //Statistics and its interface. – 2009. – Т. 2. – №. 3. – С. 257.

Ассоциативные генные сети являются молекулярно-генетическими сетями, ассоциированными с какими либо биологическими процессами, фенотипическими признаками или заболеваниями.

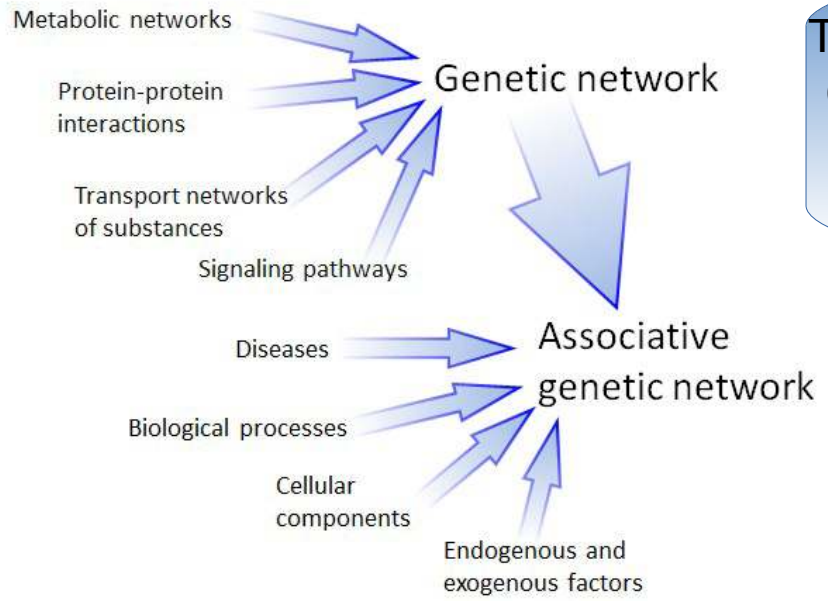
Ассоциативные генные сети в качестве вершин включают следующие типы объектов:

(1) **молекулярно-генетические объекты**. К этим типам объектов относятся гены, РНК, белки, метаболиты, клеточные компоненты;

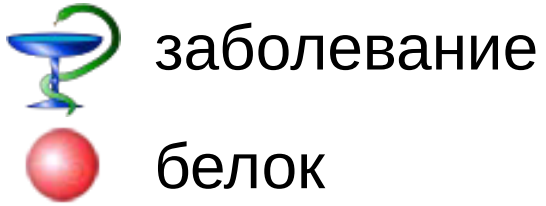
(2) **биологические процессы и системы** – метаболические пути, пути передачи сигналов, транспортные пути и т.п.);

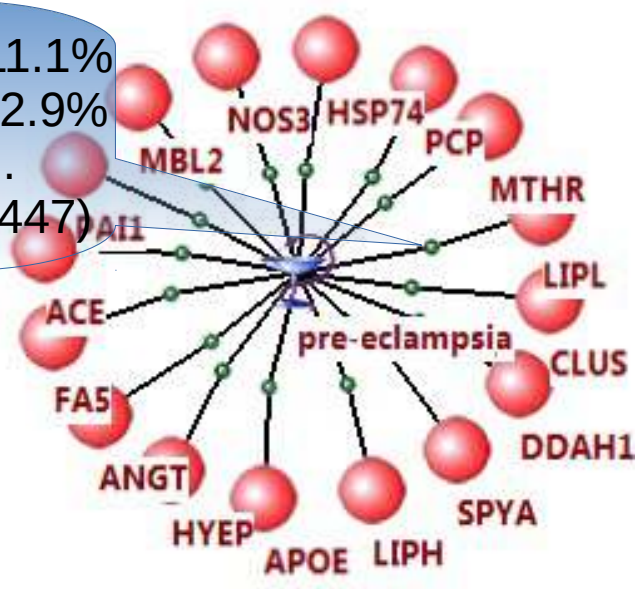
(3) **фенотипические признаки** – поведенческие характеристики или функциональные состояния организма или генетических систем, включая заболевания;

(4) **внутренние и внешние факторы**, воздействующие на систему (мутации, эпигенетический контроль, температура, давление, лекарства и другие химические соединения).



The N291S variant was identified in 11.1% of pre-eclamptics as compared with 2.9% of pregnancy controls (p = 0.008). (Hubel CA et al. 1999, PMID: 10636447)





Примеры систем, интегрирующих биологические данные.

GeneCards - интеграционный ресурс по генам ЧЕЛОВЕКА, включает информацию из более чем 125 интернет-доступных ресурсов.

Ensembl - геномный браузер для исследования геномов позвоночных 87 видов, интегрирует более ста ресурсов.

UCSC Genome Browser - геномный браузер для исследования геномов около 50 видов живых организмов, интегрирует более ста ресурсов.

GEO - репозиторий, в котором хранятся разнородные данные по экспрессии генов, полученные в 2 124 727 экспериментах.

Uniprot - содержит информацию по структуре и функциям белков 457010 бактерий, 168308 вирусов, 12163 архей, 894013 эукариот, интегрирует информацию из 145 ресурсов.

NCBI – ресурс, который обеспечивает общее информационное пространство для 66 баз данных.

Список баз данных, представленных на сайте NCBI (https://www.ncbi.nlm.nih.gov/).

Assembly	GeneReviews	Probe
BioProject (formerly Genome Project)	Genes and Disease	Protein Clusters
BioSample	Genetic Testing Registry (GTR)	**Protein Database**
BioSystems	Genome	**PubChem BioAssay**
Bookshelf	Genome Reference Consortium (GRC)	**PubChem Compound**
ClinicalTrials.gov	HIV-1	**PubChem Substance**
ClinVar	HomoloGene	**PubMed**
CloneDB (formerly Clone Registry)	Influenza Virus	**PubMed Central (PMC)**
Computational Resources from NCBI's Structure Group	Journals in NCBI Databases	PubMed Health
Consensus CDS (CCDS)	MedGen	Reference Sequence (RefSeq)
Conserved Domain Database (CDD)	MeSH Database	**RefSeqGene**
Database of Expressed Sequence Tags (dbEST)	National Library of Medicine (NLM) Catalog	Retrovirus Resources
Database of Genome Survey Sequences (dbGSS)	NCBI C++ Toolkit Manual	SARS CoV
Database of Genomic Structural Variation (dbVar)	NCBI Education Page	Sequence Read Archive (SRA)
Database of Genotypes and Phenotypes (dbGaP)	NCBI Glossary	Structure (Molecular Modeling Database)
Database of Major Histocompatibility Complex (dbMHC)	NCBI Handbook	Taxonomy
Database of Short Genetic Variations (dbSNP)	NCBI Help Manual	Third Party Annotation (TPA) Database
GenBank	NCBI Pathogen Detection Project	Trace Archive
Gene	NCBI Website Search	UniGene
Gene Expression Omnibus (GEO) Database	**Nucleotide Database**	UniGene Library Browser
Gene Expression Omnibus (GEO) Datasets	**Online Mendelian Inheritance in Man (OMIM)**	Viral Genomes
Gene Expression Omnibus (GEO) Profiles	PopSet	Virus Variation

Результат поискового запроса к ресурсу NCBI по ключевому слову YY1

Search NCBI databases Help

yy1 Search

Results found in 33 databases for "yy1"

Literature

Books	54	books and reports
MeSH	14	ontology used for PubMed indexing
NLM Catalog	0	books, journals and more in the NLM Collections
PubMed	1,274	scientific & medical abstracts/citations
PubMed Central	5,868	full-text journal articles

Health

ClinVar	49	human variations of clinical significance
dbGaP	3	genotype/phenotype interaction studies
GTR	4	genetic testing registry
MedGen	0	medical genetics literature and links
OMIM	49	online mendelian inheritance in man
PubMed Health	0	clinical effectiveness, disease and drug reports

Genomes

Assembly	0	genome assembly information
BioProject	64	biological projects providing data to NCBI
BioSample	100	descriptions of biological source materials
Clone	770	genomic and cDNA clones
dbVar	859	genome structural variation studies
Genome	17	genome sequencing projects by organism
GSS	5	genome survey sequences
Nucleotide	3,104	DNA and RNA sequences
Probe	658	sequence-based probes and primers
SNP	6,011	short genetic variations
SRA	147	high-throughput DNA and RNA sequence read archive
Taxonomy	0	taxonomic classification and nomenclature catalog

Genes

EST	730	expressed sequence tag sequences
Gene	1,465	collected information about gene loci
GEO DataSets	235	functional genomics studies
GEO Profiles	103,477	gene expression and molecular abundance profiles
HomoloGene	10	homologous gene sets for selected organisms
PopSet	32	sequence sets from phylogenetic and population studies
UniGene	104	clusters of expressed transcripts

Proteins

Conserved Domains	7	conserved protein domains
Protein	979	protein sequences
Protein Clusters	1	sequence similarity-based protein clusters
Structure	19	experimentally-determined biomolecular structures

Chemicals

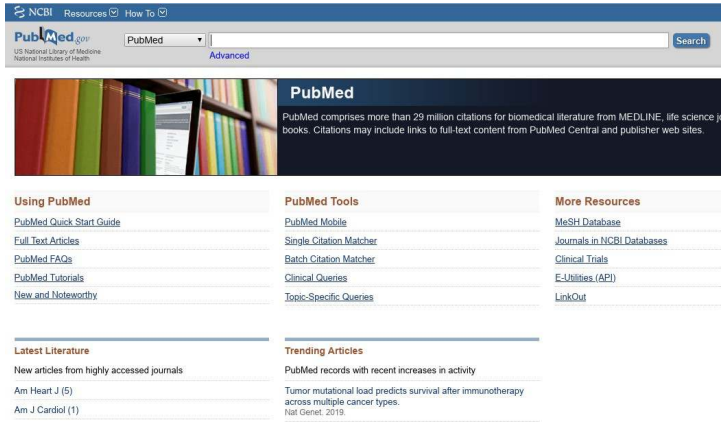
BioSystems	767	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	7	bioactivity screening studies
PubChem Compound	1	chemical information with structures, information and links
PubChem Substance	280	deposited substance and chemical information

План лекции

1. Введение.
2. Обзор Интернет-ресурсов, интегрирующих информацию по биомедицинской тематике.
3. Ресурсы, интегрирующие биологическую информацию из разнородных источников и представляющие ее в виде генных сетей: ANDSystem, STRING, GeneMania, Pathway Commons.
4. Практическое применение инструментов интеграции.

Источники информации

Важнейшим источником информации по биомедицинской тематике являются научные публикации. В системе **PubMed**, которая является крупнейшей базой данных научных статей, доступных через интернет, собрано более 29 миллионов публикаций.



Интерфейс PubMed

Журнал **Nucleic Acids Research (NAR) database issue** (IF=8,65) публикует списки **баз данных**. На январь 2019 года работали 1613 баз:

http://www.oxfordjournals.org/nar/database/c

Категории баз данных

Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
Metabolic and Signaling Pathways
Human and other Vertebrate Genomes
Human Genes and Diseases
Microarray Data and other Gene Expression
Proteomics Resources
Other Molecular Biology Databases
Organelle databases
Plant databases
Immunological databases
Cell biology

Как Вы думаете, по какому запросу к PubMed будет найдено больше статей?

1. Beta vulgaris - сахарная свёкла

2. Panax ginseng – женьшень обыкновенный

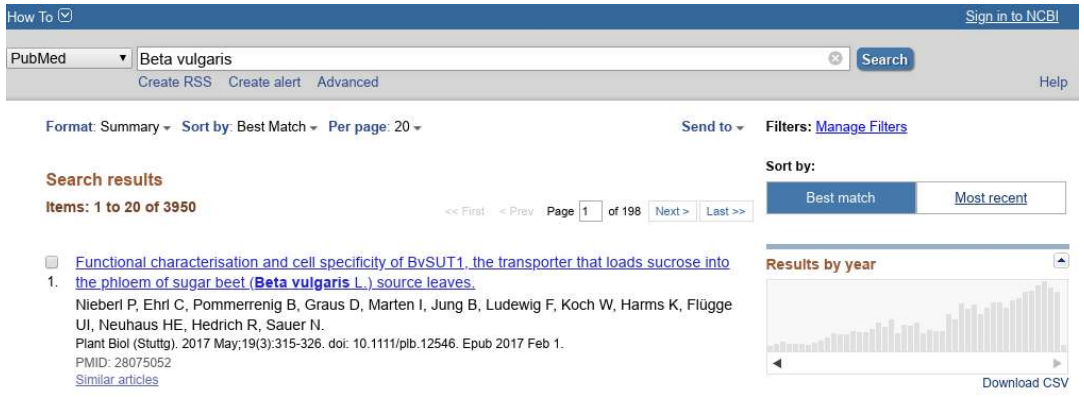
Как Вы думаете, по какому запросу к PubMed будет найдено больше статей?

1. Beta vulgaris - сахарная свёкла:

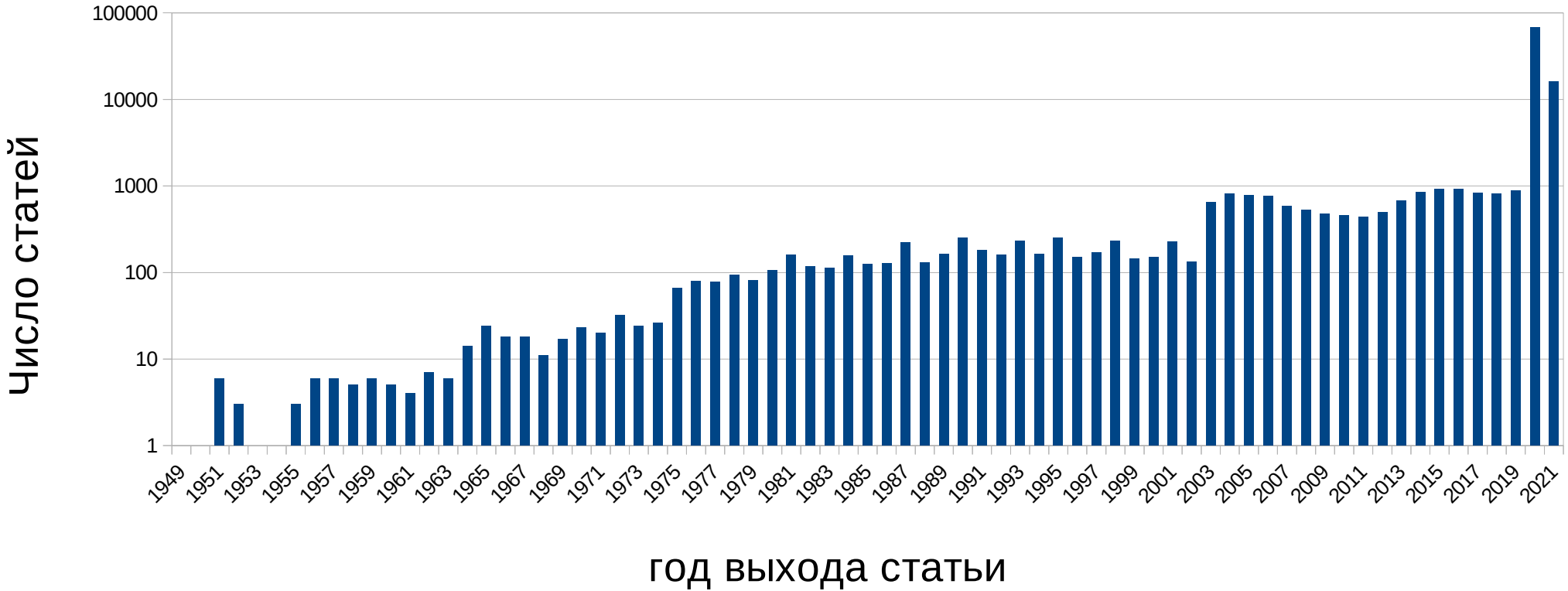
3950 документа

2. Panax ginseng – женьшень обыкновенный:

6911 документов



Число публикаций по запросу “coronavirus”



Компоненты генной сети:

1) группа координированно экспрессирующихся генов, составляющая ядро сети;
2) белки, кодируемые этими генами;
3) низкомолекулярные компоненты (гормоны и другие сигнальные молекулы, энергетические компоненты, метаболиты);
4) связи между участниками сети (в том числе отрицательные и положительные обратные связи).

По Колчанову Н.А. с соавт. ГЕННЫЕ СЕТИ // Вавиловский журнал генетики и селекции, 2013, Т. 17:4/2.

5) И другие биологические объекты

Базы данных, содержащие информацию по генам.

- HGNC
- NCBI Gene (Entrez Gene)
- NCBI Nucleotide (GenBank)
- GeneCards
- Ensembl
- UCSC Genome Browser
- GEO – хранилище данных по экспрессии генов
- другие

HGNC - HUGO Gene Nomenclature Committee

- Доступна через интернет по адресу http://www.genenames.org/
- Содержит верифицированные названия и синонимы человеческих генов.
- Идентификаторами являются числа. Например, ген YY1 имеет идентификатор 12856.
- Содержится информация по

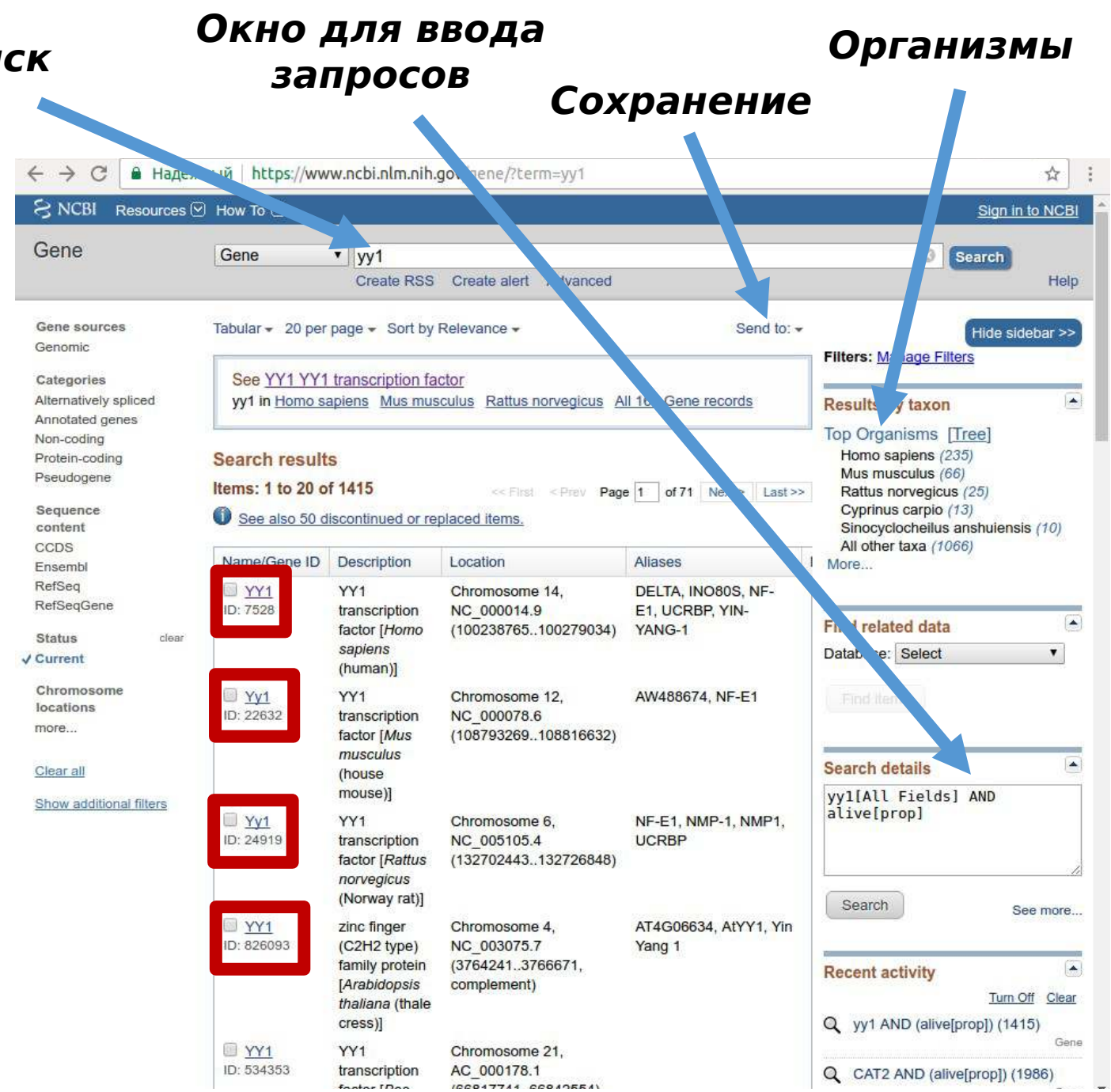
✓Синонимам

✓Хромосомному локусу

✓Ссылки на другие ресурсы

NCBI Gene

- Доступна через интернет по адресу https://www.ncbi.nlm.nih.gov/gene
- Интегрирует информацию по генам 9479 эукариот, 1151 бактерий, 7152 вирусов и др.
- Идентификаторами генов являются числа. Например, ген транскрипционного фактора YY1 человека имеет идентификатор 7528, мыши – 22632, крысы – 24919, Arabidopsis thaliana – 826093.



NCBI Gene - Информационная карточка гена

NCBI Resources How To Sign in to NCBI

Gene Gene Search

Advanced Help

Full Report Send to: Hide sidebar >>

HDAC2 histone deacetylase 2 [*Homo sapiens* (human)]

Gene ID: 3066, updated on 3-Mar-2020

Summary

Official Symbol	HDAC2 provided by HGNC
Official Full Name	histone deacetylase 2 provided by HGNC
Primary source	HGNC:HGNC:4853
See related	Ensembl:ENSG00000196591 MIM:605164
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	HD2; RPD3; YAF1; KDAC2
Summary	This gene product belongs to the histone deacetylase family. Histone deacetylases act via the formation of large multiprotein complexes, and are responsible for the deacetylation of lysine residues at the N-terminal regions of core histones (H2A, H2B, H3 and H4). This protein forms transcriptional repressor complexes by associating with many different proteins, including YY1, a mammalian zinc-finger transcription factor. Thus, it plays an important role in transcriptional regulation, cell cycle progression and developmental events. Alternative splicing results in multiple transcript variants. [provided by RefSeq, Apr 2010]
Expression	Ubiquitous expression in testis (RPKM 14.8), endometrium (RPKM 10.6) and 25 other tissues See more
Orthologs	mouse all

Table of contents

NCBI Nucleotide

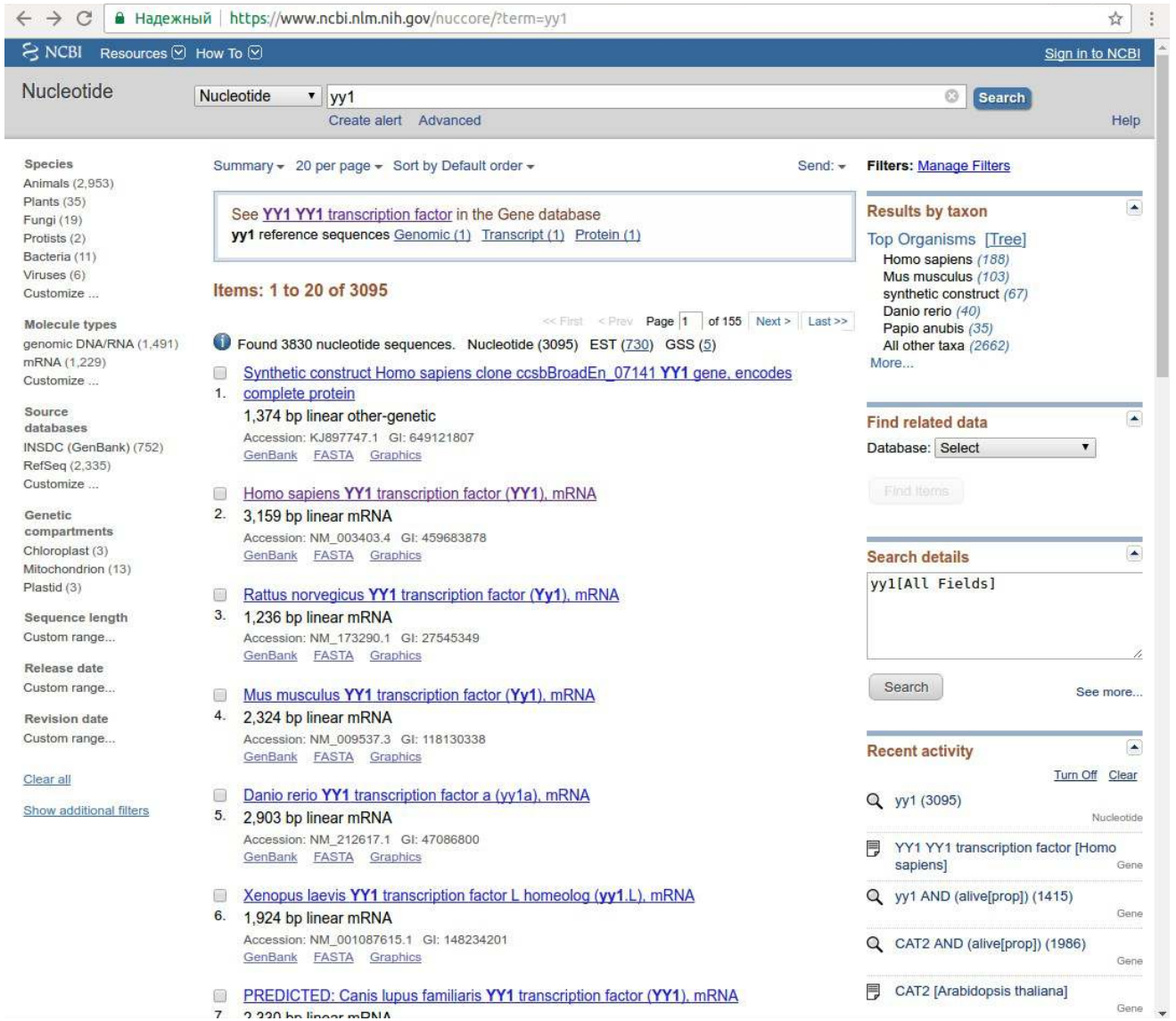
- https://www.ncbi.nlm.nih.gov/nuccore
- Коллекция нуклеотидных последовательностей из нескольких источников GenBank, RefSeq, TPA и PDB.
- Идентификаторы обычно начинаются с двух заглавных букв, например: NM_003403.4

Nucleotide database

G – gene
M – mRNA
P – protein

Версия документа

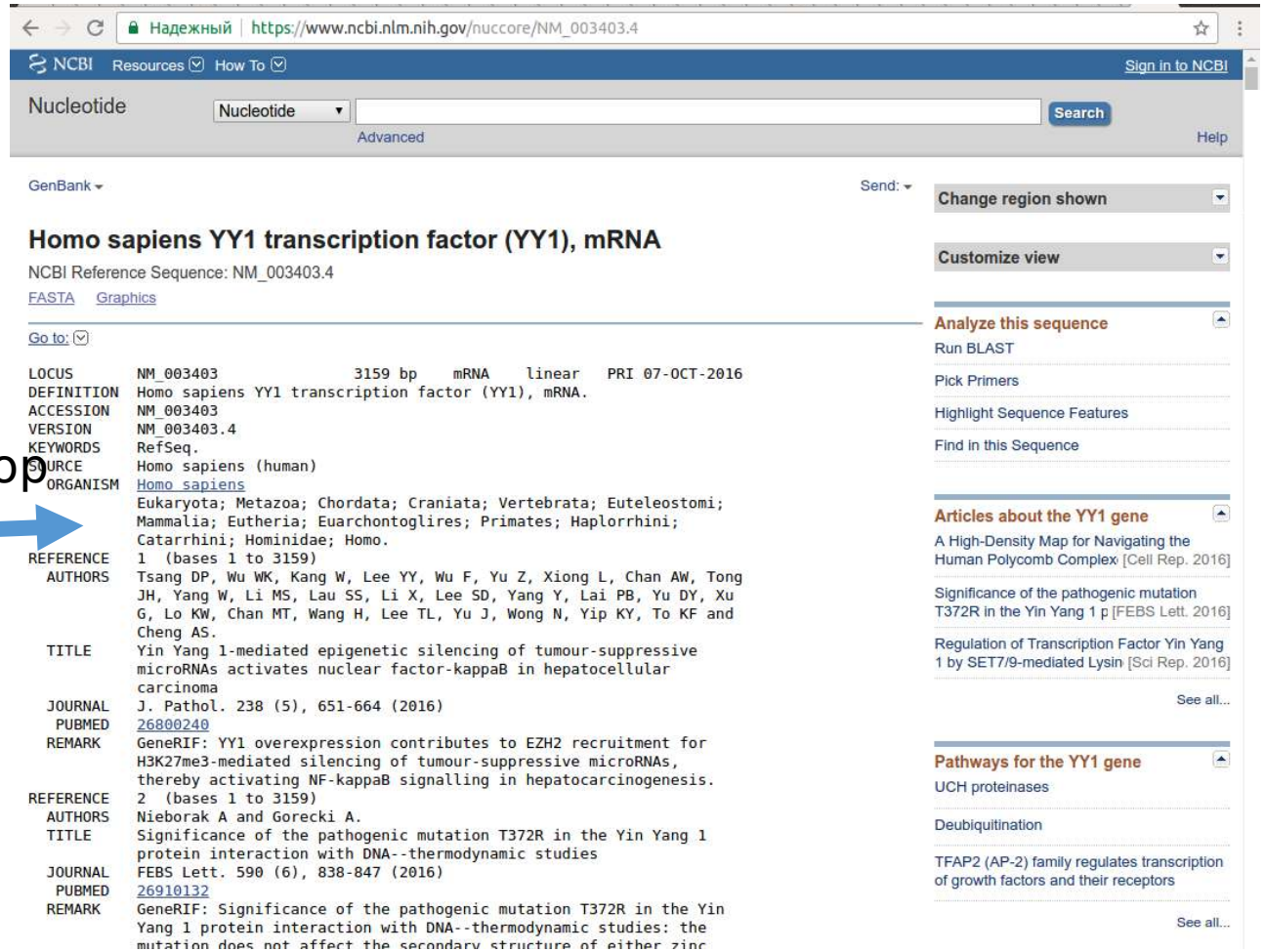
Номер mRNA, не совпадает с номером гена или белка



NCBI Nucleotide. Карточка последовательности гена

Общая информация

Локус
Идентификатор
Организм и таксономия
Ссылки на литературу



Особенности нуклеотидной последовательности

```
exon            1..1542
                /gene="YY1"
                /gene_synonym="DELTA; INO80S; NF-E1; UCRBP; YIN-YANG-1"
                /inference="alignment:Splign:2.0.8"
exon            1543..3159
                /gene="YY1"
                /gene_synonym="DELTA; INO80S; NF-E1; UCRBP; YIN-YANG-1"
                /inference="alignment:Splign:2.0.8"
STS             1729..1852
                /gene="YY1"
                /gene_synonym="DELTA; INO80S; NF-E1; UCRBP; YIN-YANG-1"
                /standard_name="RH1618"
                /db_xref="UniSTS:42115"
STS             1954..2640
                /gene="YY1"
                /gene_synonym="DELTA; INO80S; NF-E1; UCRBP; YIN-YANG-1"
                /standard_name="YY1__3892"
                /db_xref="UniSTS:462950"
STS             2035..2173
                /gene="YY1"
                /gene_synonym="DELTA; INO80S; NF-E1; UCRBP; YIN-YANG-1"
                /standard_name="G06170"
                /db_xref="UniSTS:49520"
regulatory      2690..2695
                /regulatory_class="polyA_signal_sequence"
                /gene="YY1"
                /gene_synonym="DELTA; INO80S; NF-E1; UCRBP; YIN-YANG-1"
polyA_site      2714
                /gene="YY1"
                /gene_synonym="DELTA; INO80S; NF-E1; UCRBP; YIN-YANG-1"
STS             2879..3086
                /gene="YY1"
                /gene_synonym="DELTA; INO80S; NF-E1; UCRBP; YIN-YANG-1"
                /standard_name="RH68412"
                /db_xref="UniSTS:51218"
STS             3001..3100
                /gene="YY1"
                /gene_synonym="DELTA; INO80S; NF-E1; UCRBP; YIN-YANG-1"
                /standard_name="RH44921"
                /db_xref="UniSTS:36888"
regulatory      3106..3111
                /regulatory_class="polyA_signal_sequence"
                /gene="YY1"
                /gene_synonym="DELTA; INO80S; NF-E1; UCRBP; YIN-YANG-1"
polyA_site      3132
                /gene="YY1"
                /gene_synonym="DELTA; INO80S; NF-E1; UCRBP; YIN-YANG-1"
ORIGIN
      1 agggcgaacg ggcgagtggc agcgaggcgg ggcgggctga ggccagcgcg gaagtctcgc
     61 gaggccgggc ccgagcagag tgtggcggcg gcggcgagat ctgggctcgg gttgaggagt
    121 tggtatttgt gtggaaggag gcggaggcgc aggaggaagg gggaagcgga gcgccggccc
    181 ggagggcggg aggaggcgcg gccagggcgg gcggttgcgg cgaggcgagg cgaggcgggg
    241 agccgagacg agcagcggcc gagcgagcgc gggcgcgggc gcaccgaggc gagggaggcg
    301 gggaagcccc gccgccgccg cggcgcccgc cccttccccc gccgcccgcc ccctctcccc
    361 ccgcccgctc gccgccttcc tccctctgcc ttccttcccc acggccggcc gcctcctcgc
    421 ccgcccgccc gcagccgagg agccgaggcc gccgcggccg tggcggcgga gccctcagcc
```

Продолжение информационной карточки

GeneCards

Информационная карточка гена

http://www.genecards.org/

- Интеграционный ресурс по генам ЧЕЛОВЕКА, включает информацию из более чем 125 интернет-доступных ресурсов.
- Содержится информация по

✓ синонимам гена

✓ Идентификаторам

✓ Описание, функция

✓ структуре нуклеотидной последовательности

✓ продуктам гена (соответствующим белкам)

✓ Биологические пути, в которых ген участвует

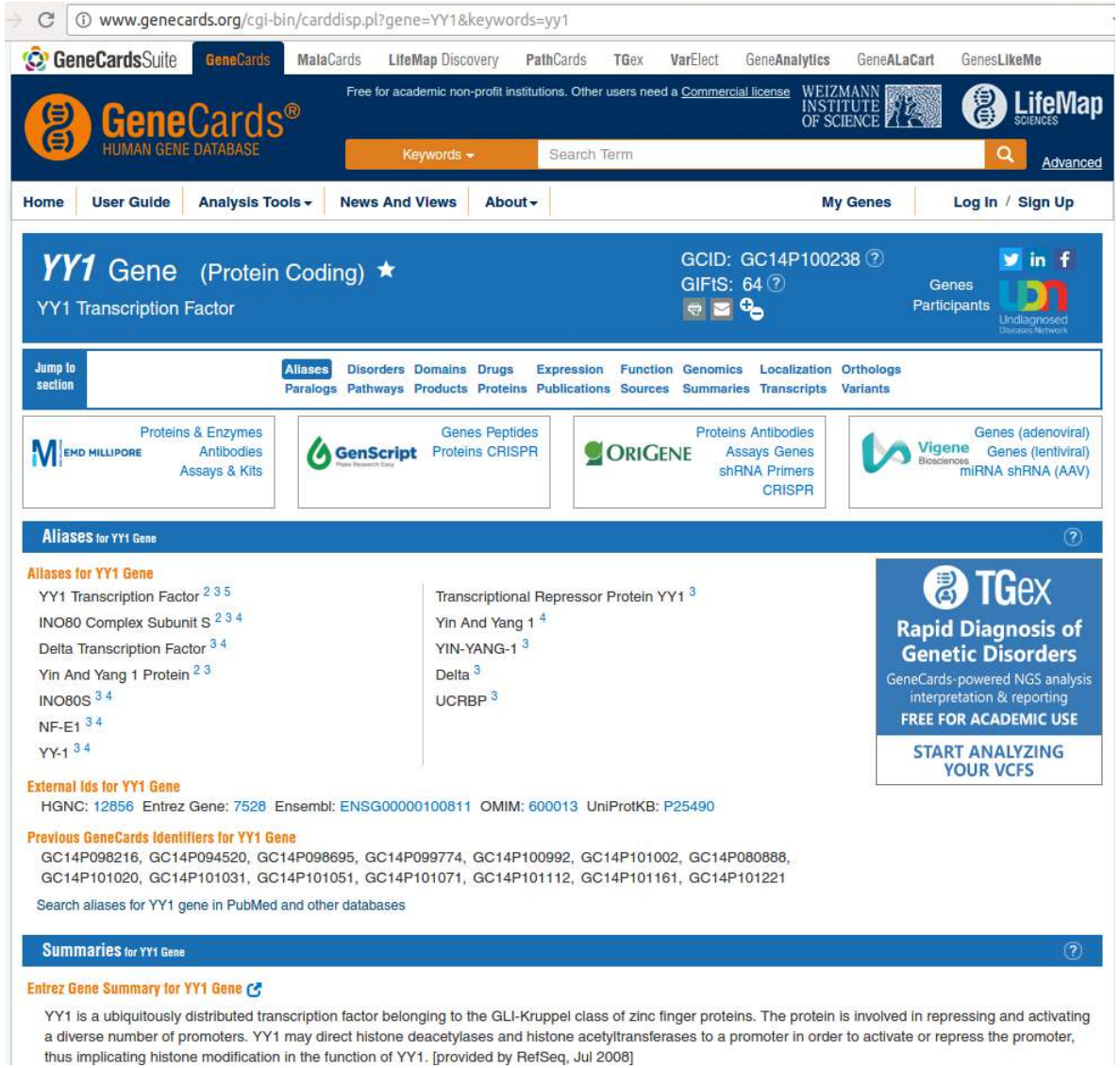
✓ Экспрессии

✓ Ортологам

✓ Полиморфизму

✓ Ассоциированным заболеваниям

✓ Публикациям



Ensembl

http://www.ensembl.org/index.html

- Геномный браузер для исследования геномов позвоночных (87 видов).
- Идентификаторы типа ENS код организма G/T/P число. Например, ген YY1 человека имеет идентификатор ENSG00000100811, один из транскриптов - ENST00000636393, ген YY1 мыши - ENSMUSG00000021264.
- Удобен для проведения сравнительных геномных исследований.
- Содержится информация по:

✓Синонимам, описанию, локализации

✓Последовательности

✓Ортологам и паралогам

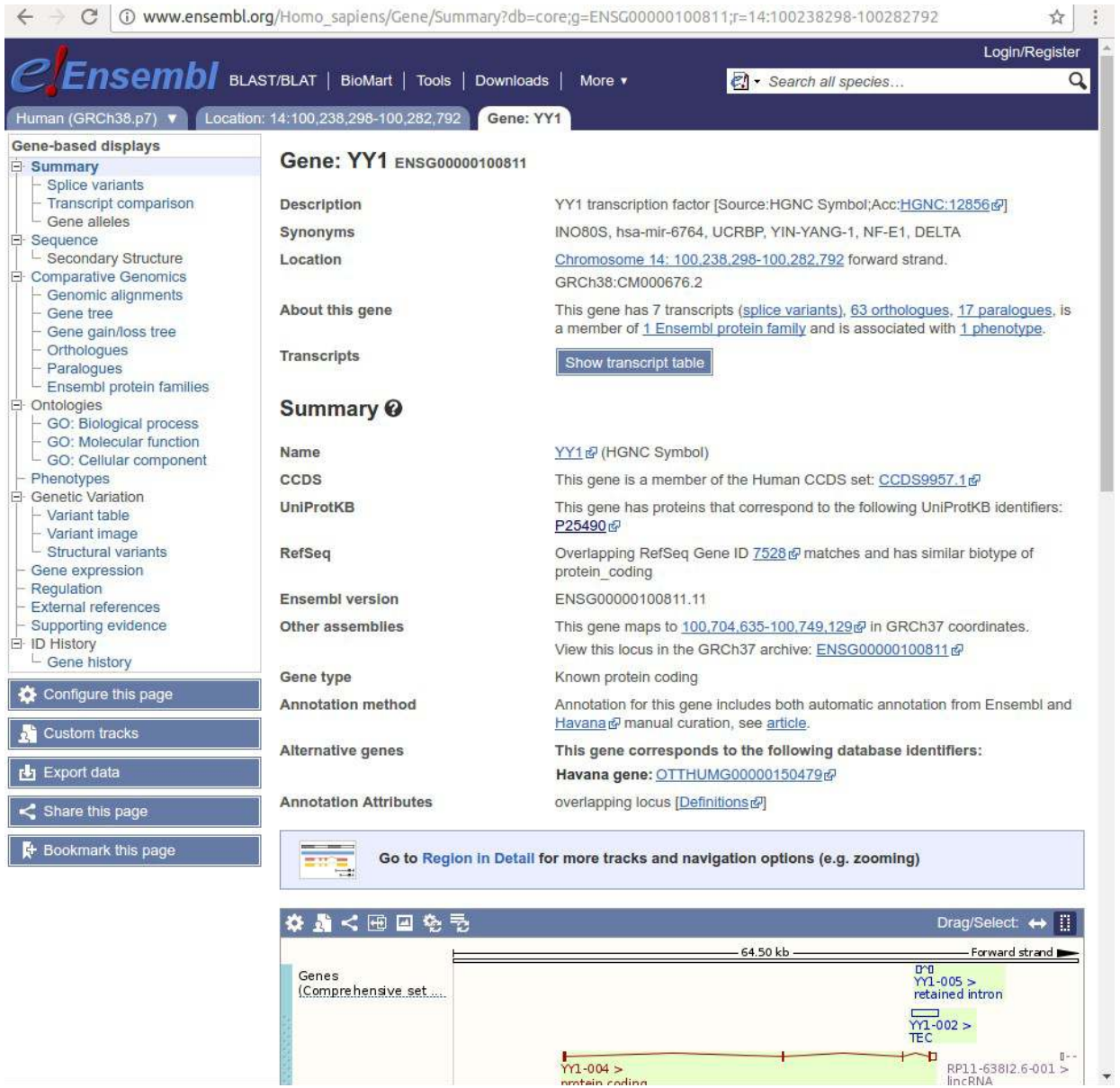
✓Биологическим процессам

✓Полиморфизмам

✓Экспрессии

✓Регуляции и др.

Информационная карточка гена



UCSC (University of California, Santa Cruz) Genome Browser

https://genome.ucsc.edu/

- Геномный браузер для исследования геномов около 50 видов живых организмов.
- Идентификаторы типа uc три цифры три буквы . цифра. Например, один из транскриптов гена YY1 человека имеет идентификатор uc001ygy.3, другой - uc059fch.1
- Содержится информация по:

✓По интрон/экзонной структуре гена

✓Последовательности

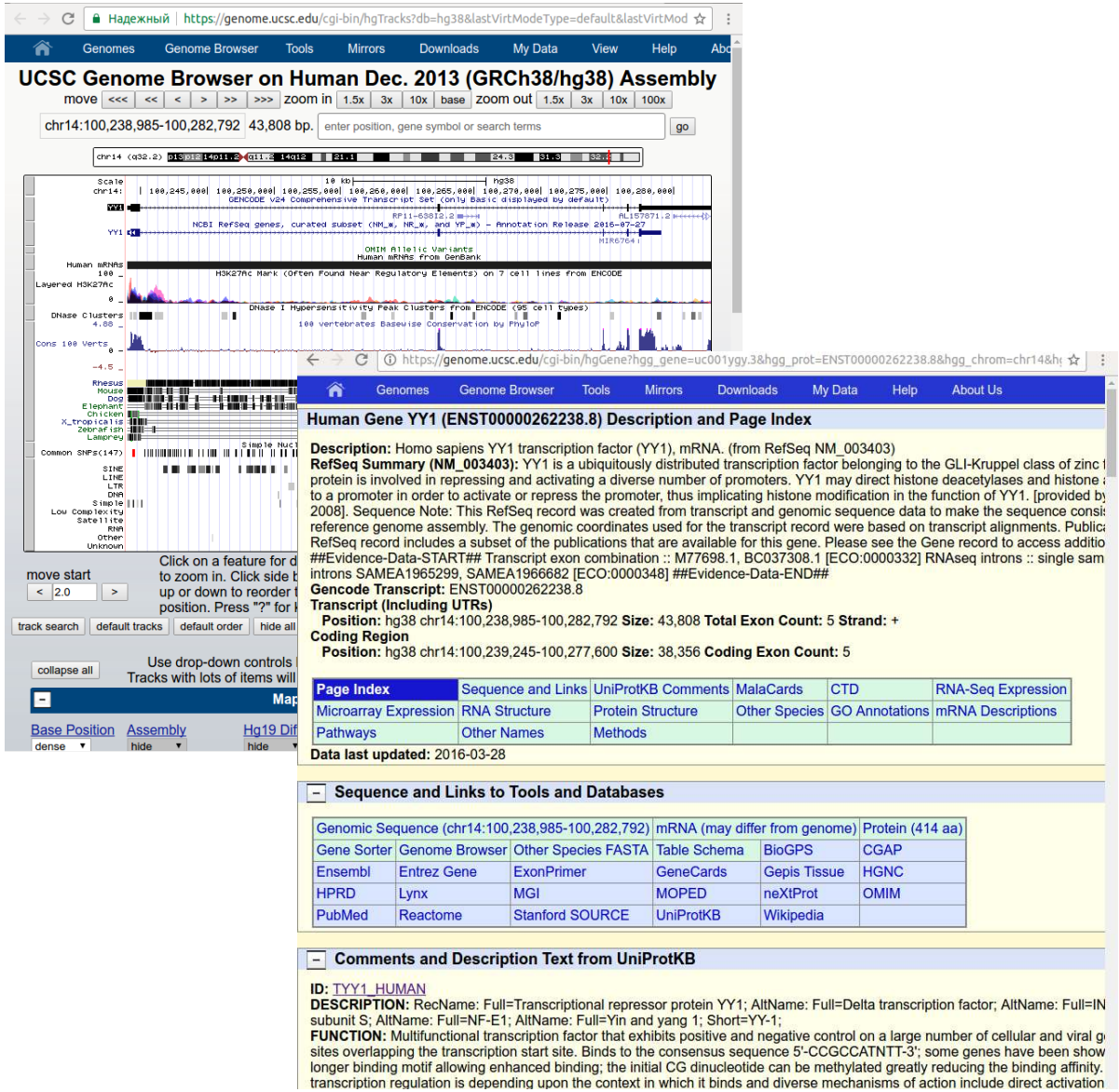
✓Биологическим процессам

✓Полиморфизмам

✓Экспрессии

✓Регуляции и др.

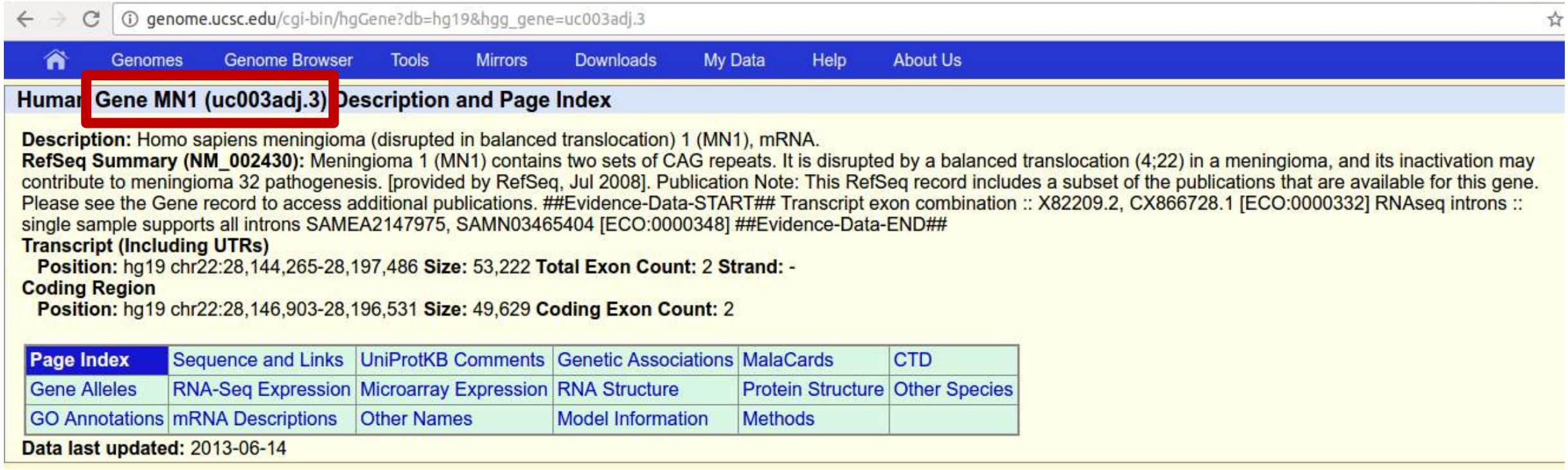
Информационная карточка гена



В базе данных UCSC genome browser для человека представлено 82 960 идентификаторов и информационных карточек. Генов человека по разным оценкам от 20 до 30 тысяч.

Почему же тогда в UCSC genome browser присутствует такая цифра 82 960?

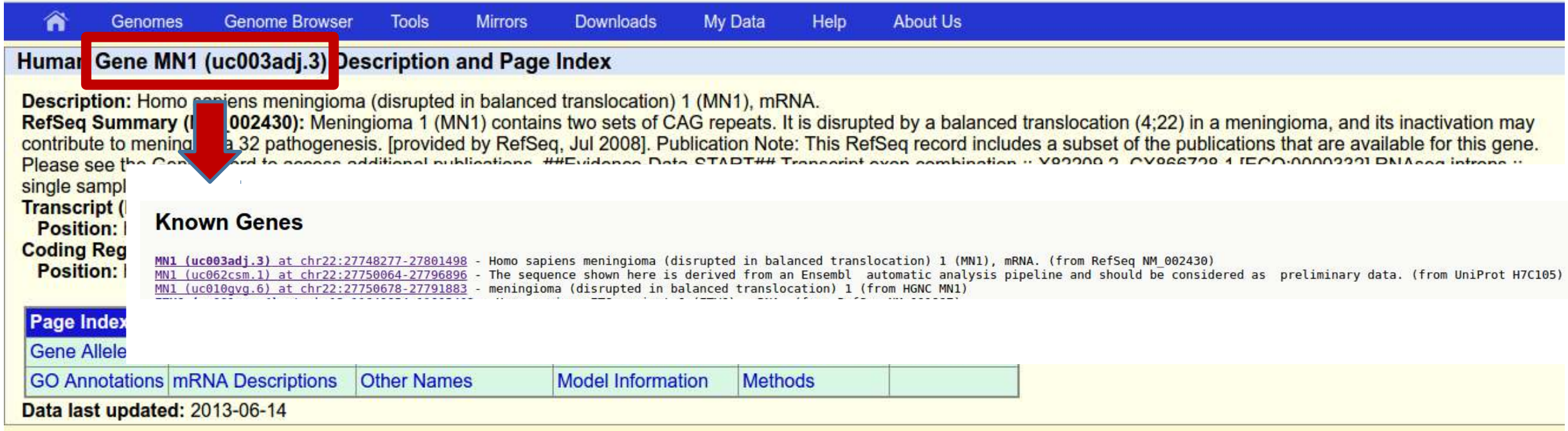
Пример информационной карточки в UCSC genome browser.



В базе данных UCSC genome browser для человека представлено 82 960 идентификаторов и информационных карточек. Генов человека по разным оценкам от 20 до 30 тысяч.

Почему же тогда в UCSC genome browser присутствует такая цифра 82 960?

В UCSC genome browser идентификатором «uc три цифры три буквы . цифра» (UCSC ID) обозначается один из транскриптов гена. В общем случае с гена может считываться несколько функциональных транскриптов.



GEO - Gene Expression Omnibus

- https://www.ncbi.nlm.nih.gov/geo/
- Данные по экспрессии генов. Высокопроизводительные эксперименты по оценке уровней экспрессии генов (методы ДНК-микрочипов и RNA-seq).
- Идентификаторы для DataSet - GDSчисло и для DataSeries - GSEчисло. Например, серия экспериментов, посвященных анализу уровня экспрессии генов в HeLa клетках при нокауте генов YY1 и YY2, имеет идентификатор GSE14964; DataSet - GDS3788, а отдельный эксперимент этой серии - GSM373617.
- Содержится информация по:

✓Дизайну экспериментов,

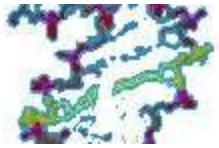
✓методам, участникам, публикациям

✓Сырые и часто нормированные данные по экспрессии

Отношения между платформами, сериями и образцами

Базы данных, содержащие информацию о полиморфизме в генах

dbSNP
Short Genetic Variations



NCBI

SNPedia

GWAS Catalog

- dbSNP
- ClinVar
- SNPedia
- GWAS

dbSNP

- https://www.ncbi.nlm.nih.gov/SNP/
- Содержит информацию по полиморфизму 10 организмов. Для человека представлена информация по более чем 600 млн коротких вариаций.
- Идентификаторы начинаются с двух букв rs далее число. Например, полиморфизм в гене человека YY1 имеет идентификатор rs61992955.
- Содержит информацию по

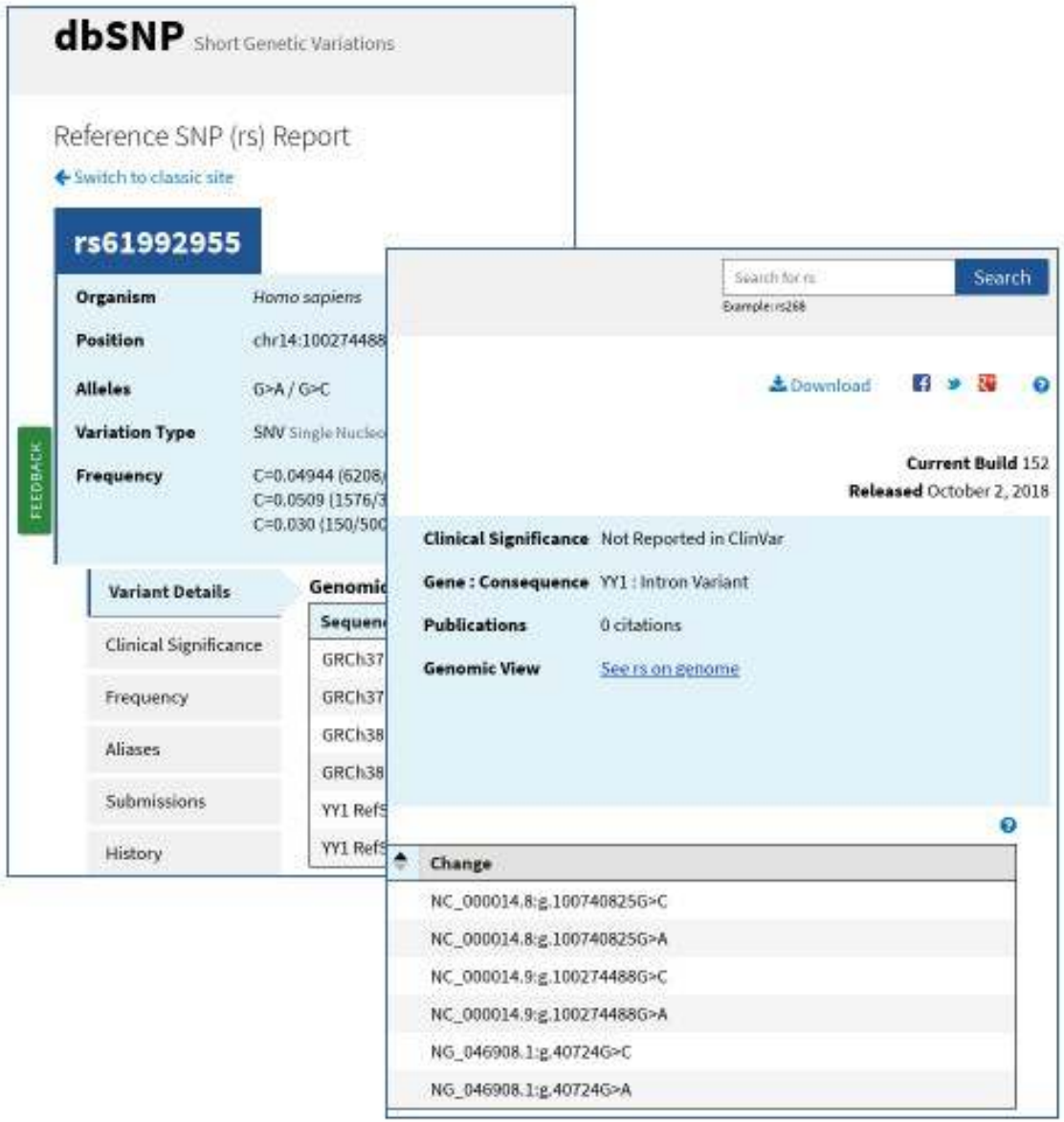
✓Частотам встречаемости генетического варианта в разных популяциях

✓Позиции в геноме

✓Фланкирующим последовательностям

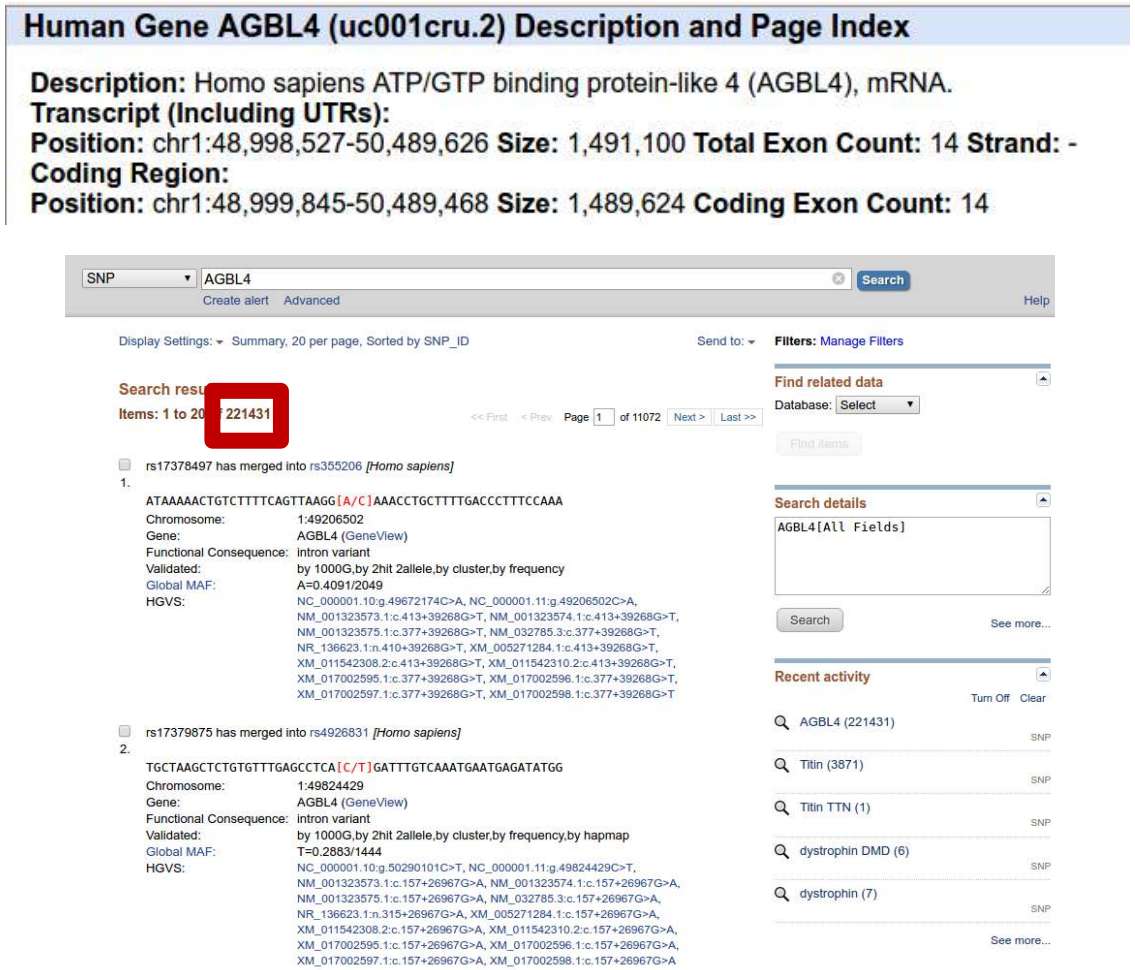
✓Эффекту и др.

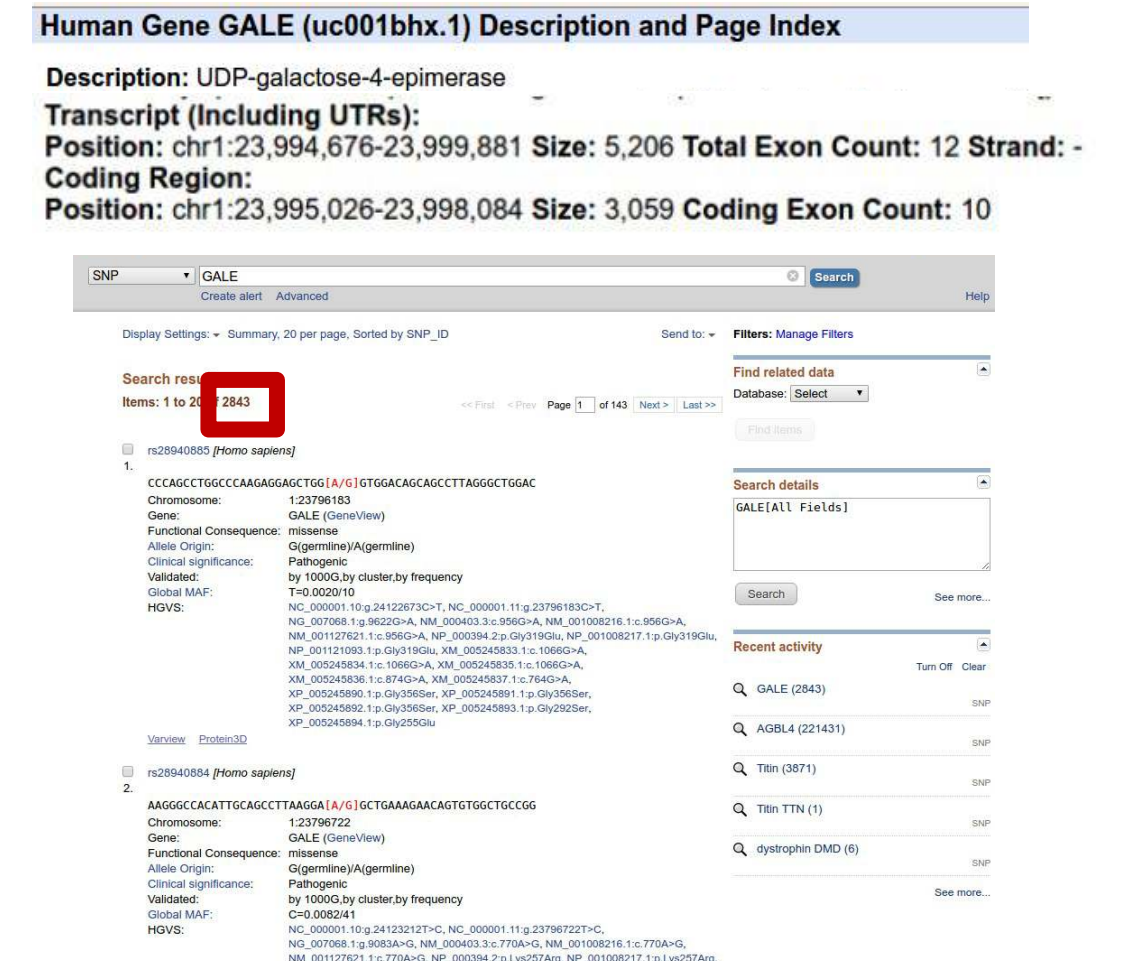
Информационная карточка вариаци



В базе данных dbSNP для гена AGBL4 присутствует информация по 221 431 SNP, а для гена GALE только по 2 843 SNP. Число 221 431 в ~78 раз больше числа 2 843.

Значит ли это что ген AGBL4 в большей степени подвержен мутагенезу?

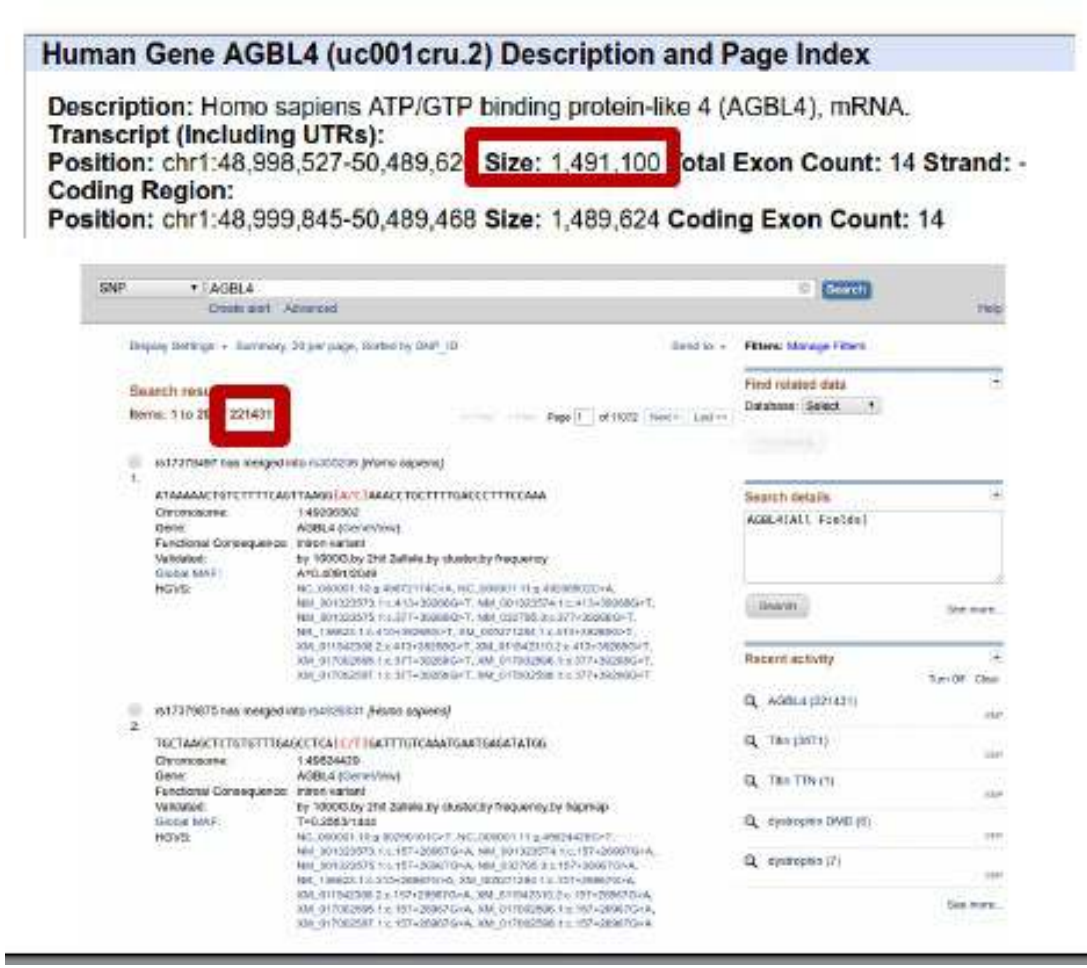


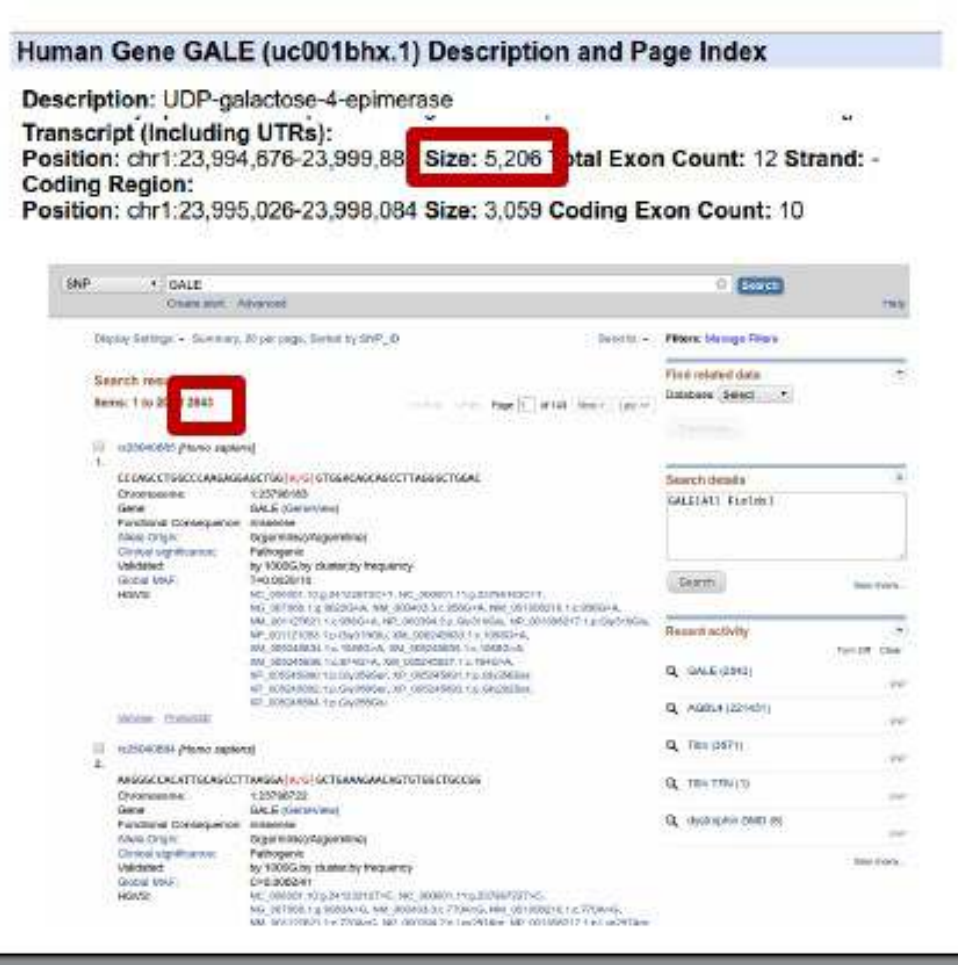


В базе данных dbSNP для гена AGBL4 присутствует информация по 221 431 SNP, а для гена GALE только по 2 843 SNP. Число 221 431 в ~78 раз больше числа 2 843.

Значит ли это что ген AGBL4 в большей степени подвержен мутагенезу?

Длина гена AGBL4 составляет 1 491 100 п.н., а гена GALE – 5 206 п.н., т.е. ген AGBL4 в 286 раз длиннее гена GALE. Таким образом, в гене AGBL4 на 100 п.н. приходится 14,85 SNP, а в гене GALE – 54,61, т.е. число известных мутаций на п.н. в гене GALE значительно превосходит это же число для гена AGBL4.





ClinVar

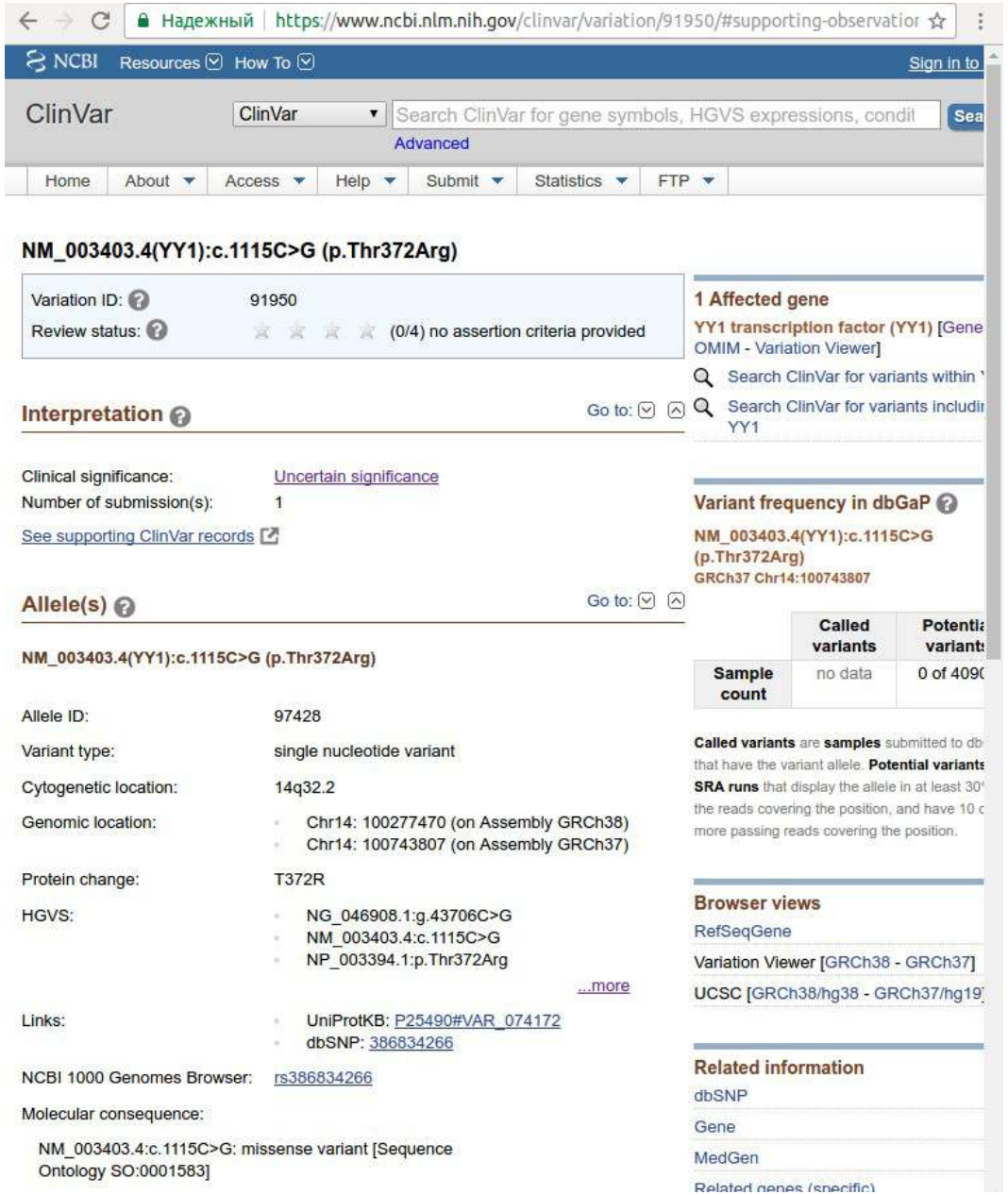
- https://www.ncbi.nlm.nih.gov/clinvar/
- Содержит информацию по генетическим вариантам человека и их связи с заболеваниями.
- Идентификаторы – числа, но также используются более разветнутые названия. Например, вариант гена человека YY1 имеет идентификатор 91950, а также NM_003403.4:c.1115C>G (p.Thr372Arg).
- Содержит информацию по

✓Клинической значимости генетического варианта

✓Типу генетического варианта

✓Позиции в геноме и в белке

Информационная карточка варианта



SNPedia

- Доступна через интернет по адресу https://www.snpedia.com/
- Организована в форме Википедии. Любой желающий может вносить и изменять информацию по полиморфизму человека и его связи с заболеваниями.
- Используются идентификаторы из базы данных dbSNP - начинаются с двух букв rs далее число. Например, полиморфизм в гене человека FCER1A имеет идентификатор rs2494262.
- Содержит информацию по

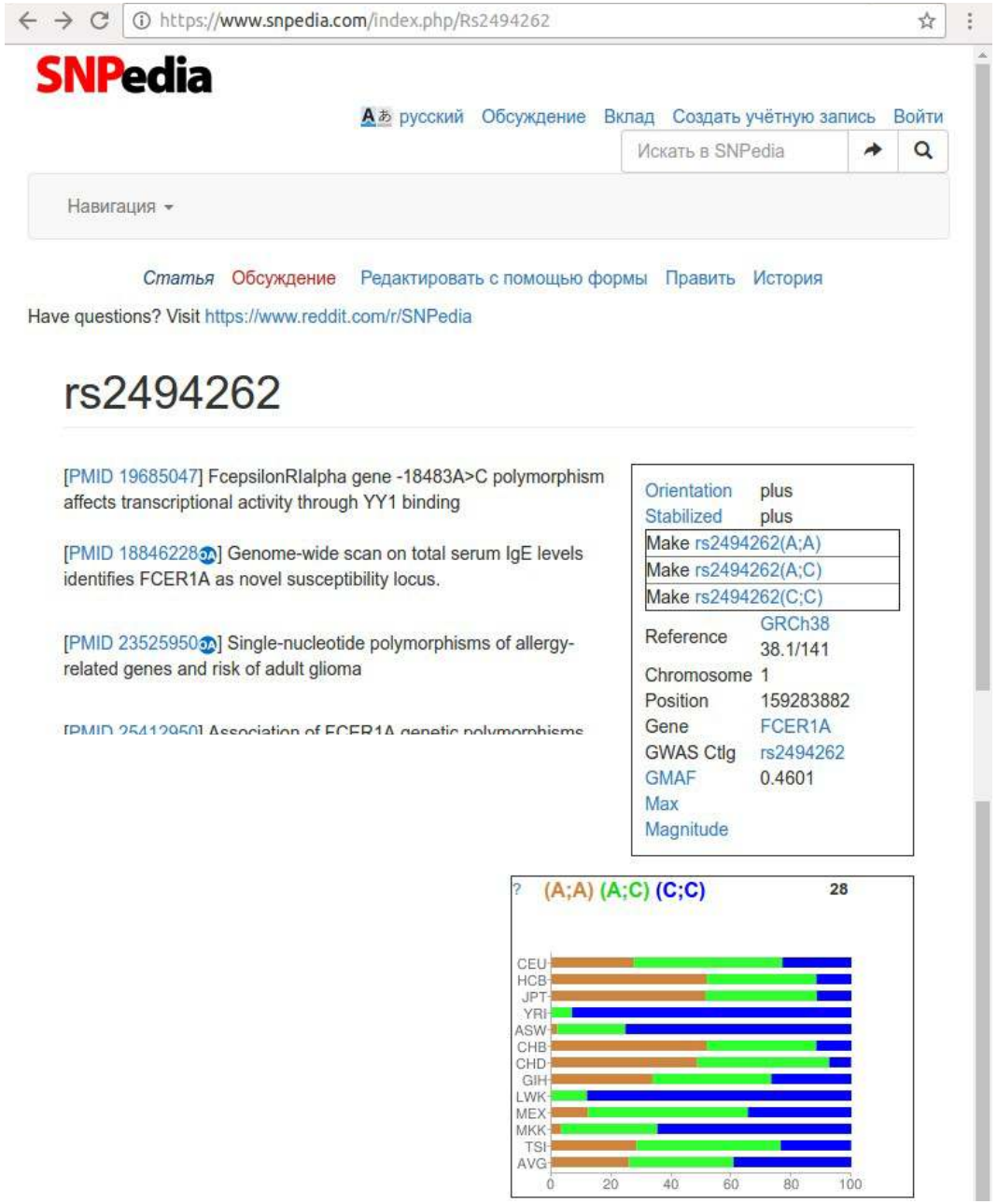
✓ Клинической значимости полиморфизма

✓ Позиции в геноме

✓ Частотам встречаемости аллельных вариантов в разных популяциях

✓ Публикациям и др.

Информационная карточка полиморфизма



GWAS Catalog

- Доступна через интернет по адресу https://www.ebi.ac.uk/gwas/
- Содержит информацию по связи полиморфизмов человека и с заболеваниями и фенотипами.
- Используются идентификаторы из базы данных dbSNP - начинаются с двух букв rs далее число. Например, полиморфизм в гене человека YY1 имеет идентификатор rs2766692.
- Содержит информацию по

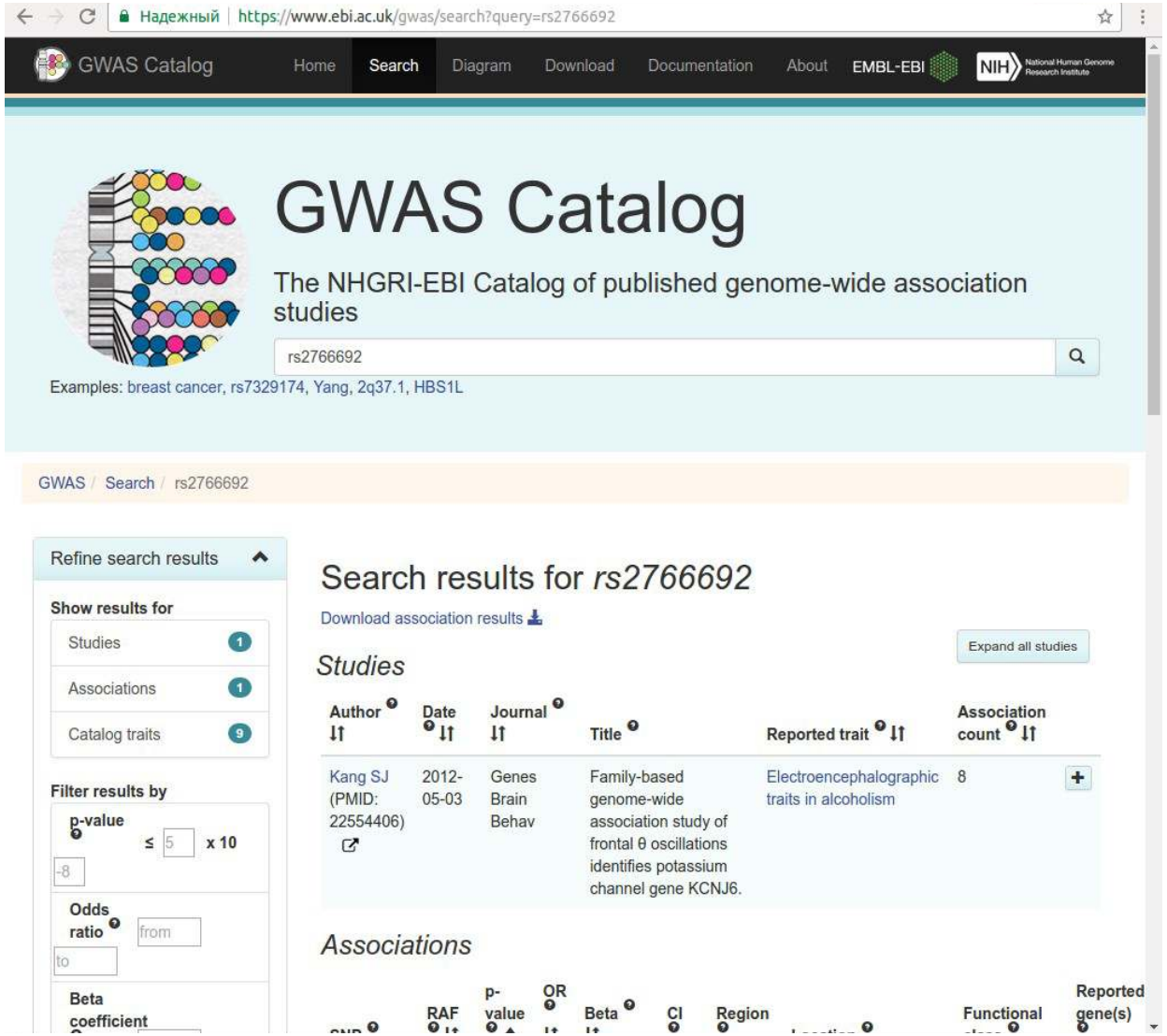
✓Ассоциированным заболеваниям и фенотипам

✓Позиции в геноме

✓Частотам встречаемости полиморфизма в разных популяциях

✓Публикациям и др.

Информационная карточка полиморфизма



Базы данных, содержащие информацию по белкам

Uniprot (Universal Protein Resource)

http://www.uniprot.org/

- Содержит информацию по структуре и функциям белков 457010 бактерий, 168308 вирусов, 12163 архей, 894013 эукариот.
- База содержит два раздела: 1) Swiss-Prot, включающий карточки белков, проаннотированные вручную экспертами;
- 2) TrEMBL, включающий карточки белков, сформированные автоматическими средствами.
- Идентификаторы двух типов 1) шесть знаков: цифры и заглавные буквы 2) короткое название белка_организм. Например, белок YY1 человека P25490 и TYY1_HUMAN, мыши - Q00899 и TYY1_MOUSE.
- Содержит информацию по

✓ синонимам и функциям

✓ Локализации в клетке

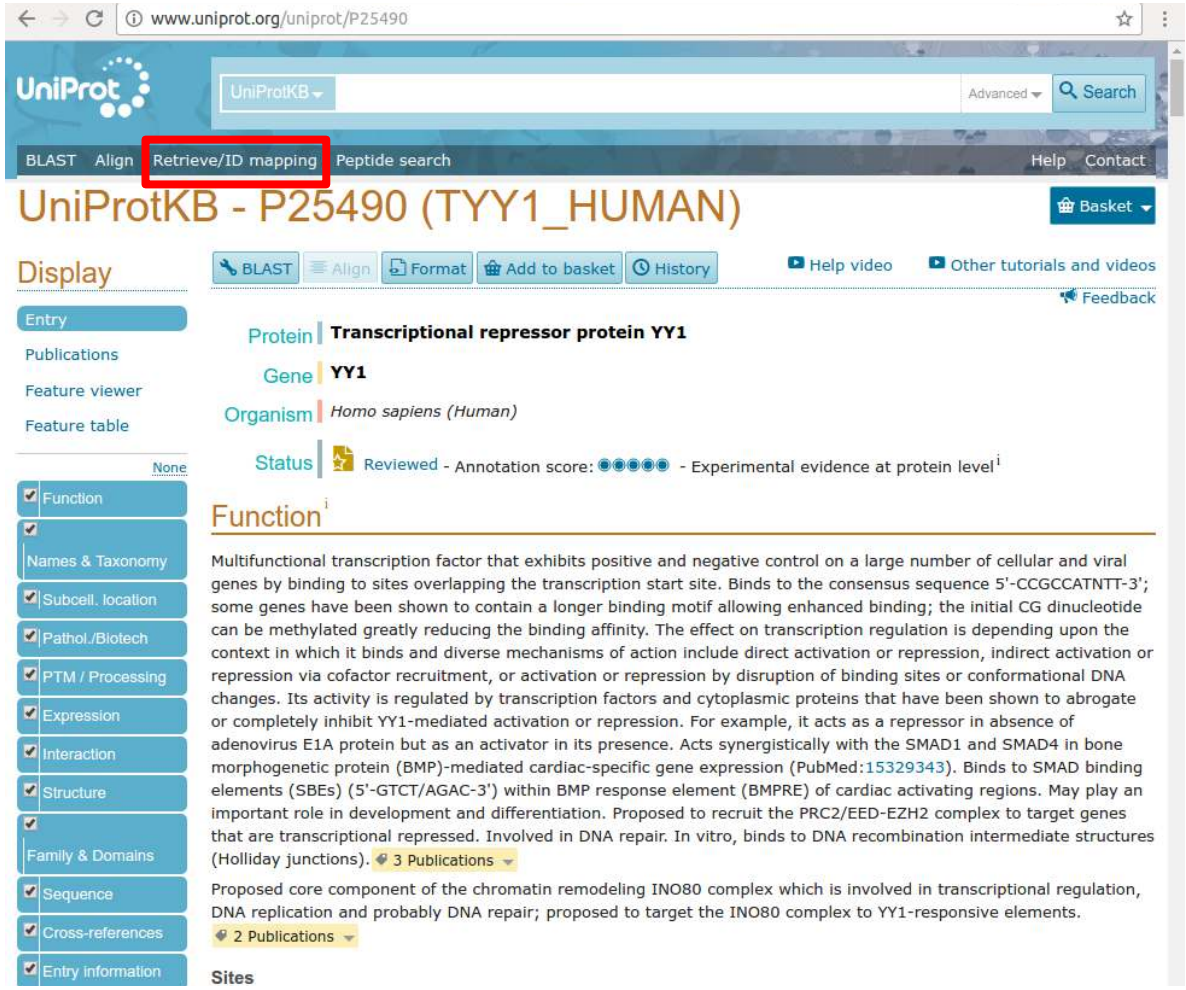
✓ Последовательности и др.

Лайфхак

H0YJV7_HUMAN

Unreviewed (TrEMBL)

Информационная карточка белка



NCBI Protein

- https://www.ncbi.nlm.nih.gov/protein/
- Содержит информацию по аминокислотным последовательностям белков различных организмов.
- Используются идентификаторы различных баз данных, таких как GenBank и RefSeq. Например, белок YY1 человека имеет идентификатор AAH65366.1
- Содержит информацию по

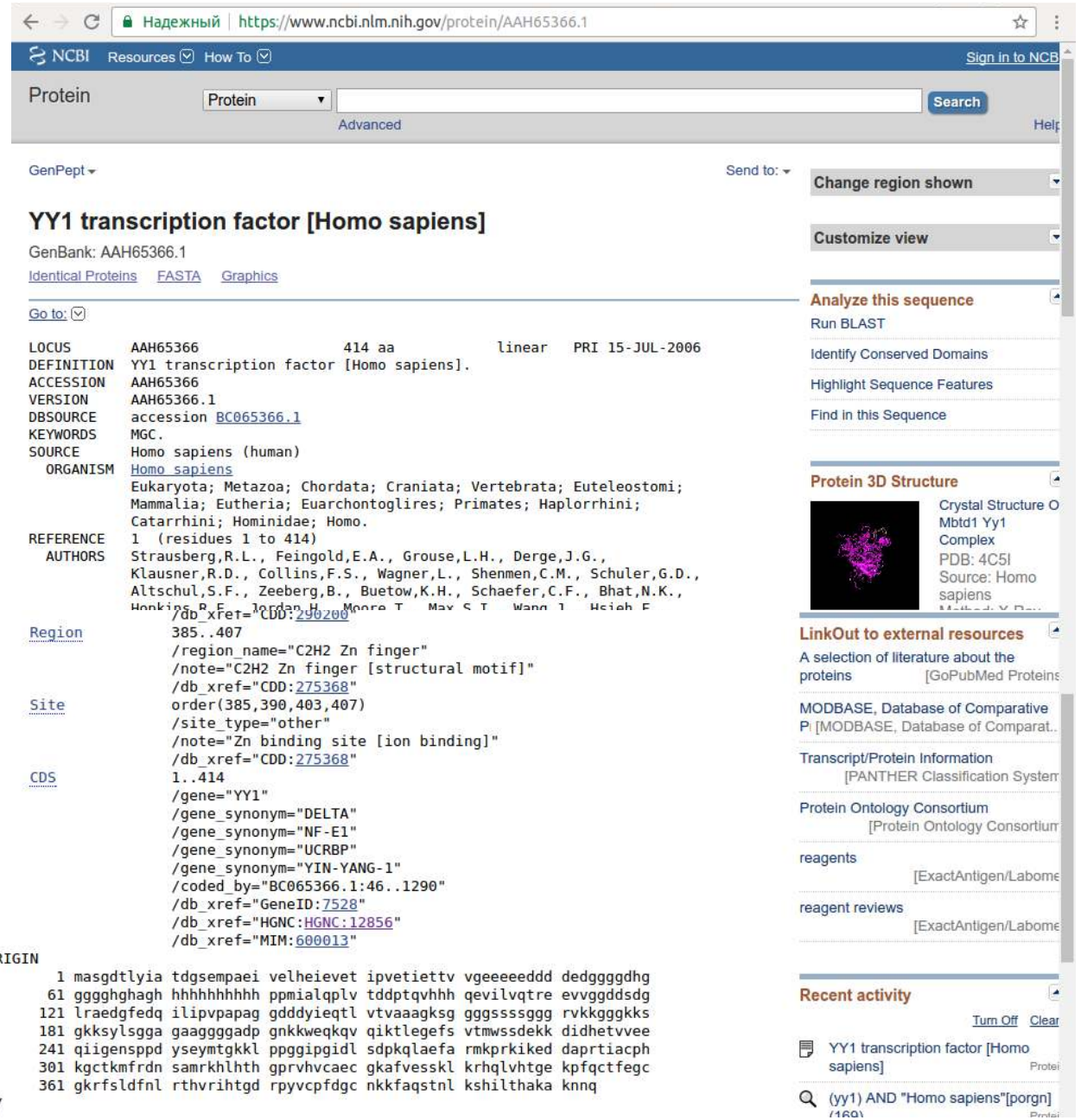
✓ Организмам, таксономическому положению организма

✓ Ссылки на публикации

✓ Разметке функциональных сайтов белка

✓ Аминокислотной последовательности и др.

Информационная карточка белка



PDB - Protein Data Bank

- http://www.rcsb.org/pdb/home/home.do
- Содержит информацию по 3D структуре белков и белковых комплексов различных организмов.
- Идентификаторы четыре символа: цифры и заглавные буквы. Например, белок HUMAN ERYTHROCYTE CATALASE имеет идентификатор 1DGB.
- Содержит информацию по

✓ Функции белка

✓ Ссылки на публикации

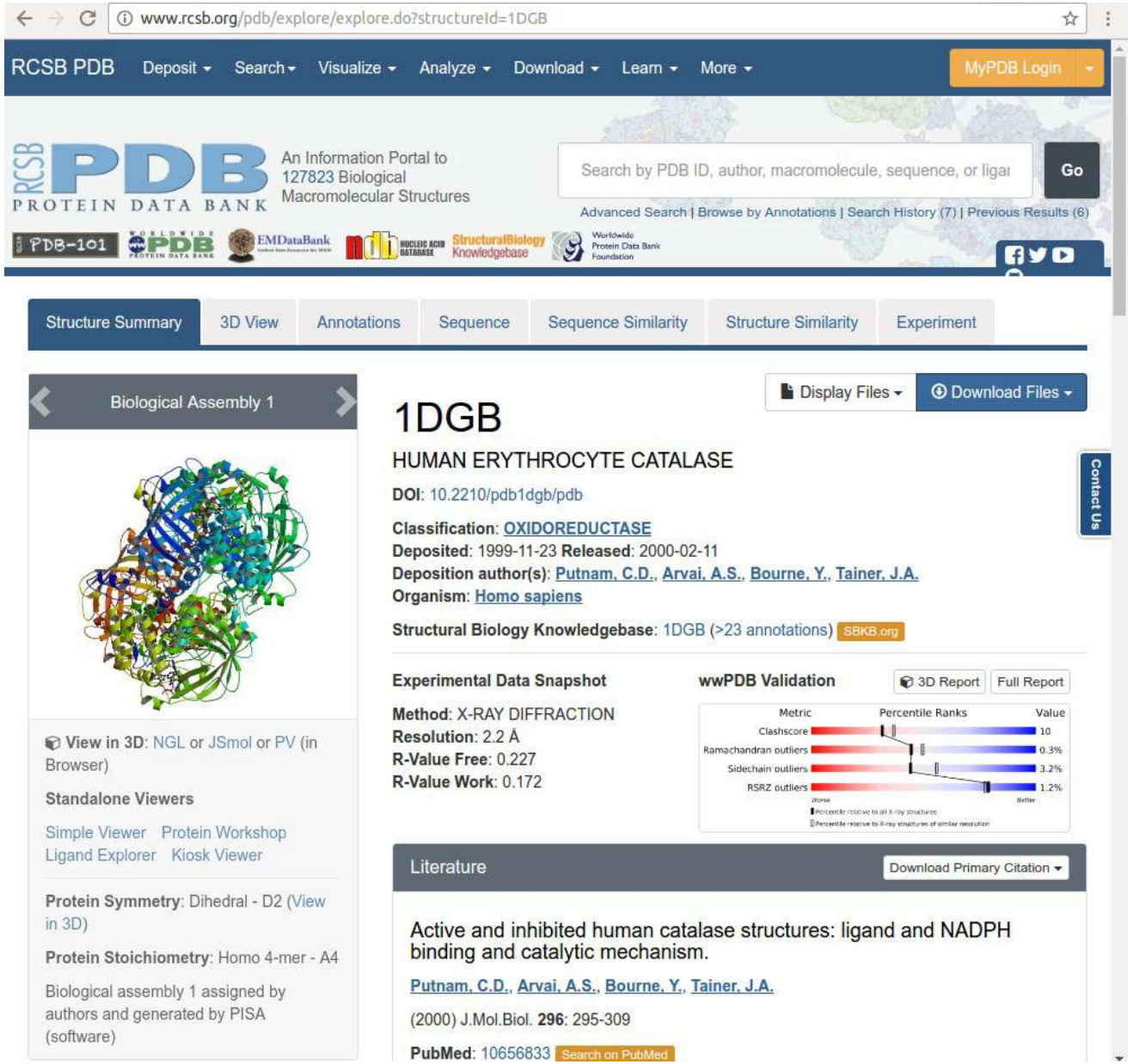
✓ Лигандам

✓ 3D структуре

✓ Аминокислотной последовательности

✓ Вторичной структуре белка и др.

Информационная карточка белка



Базы данных, содержащие информацию по микроРНК

miRBase

Информационная карточка микроРНК

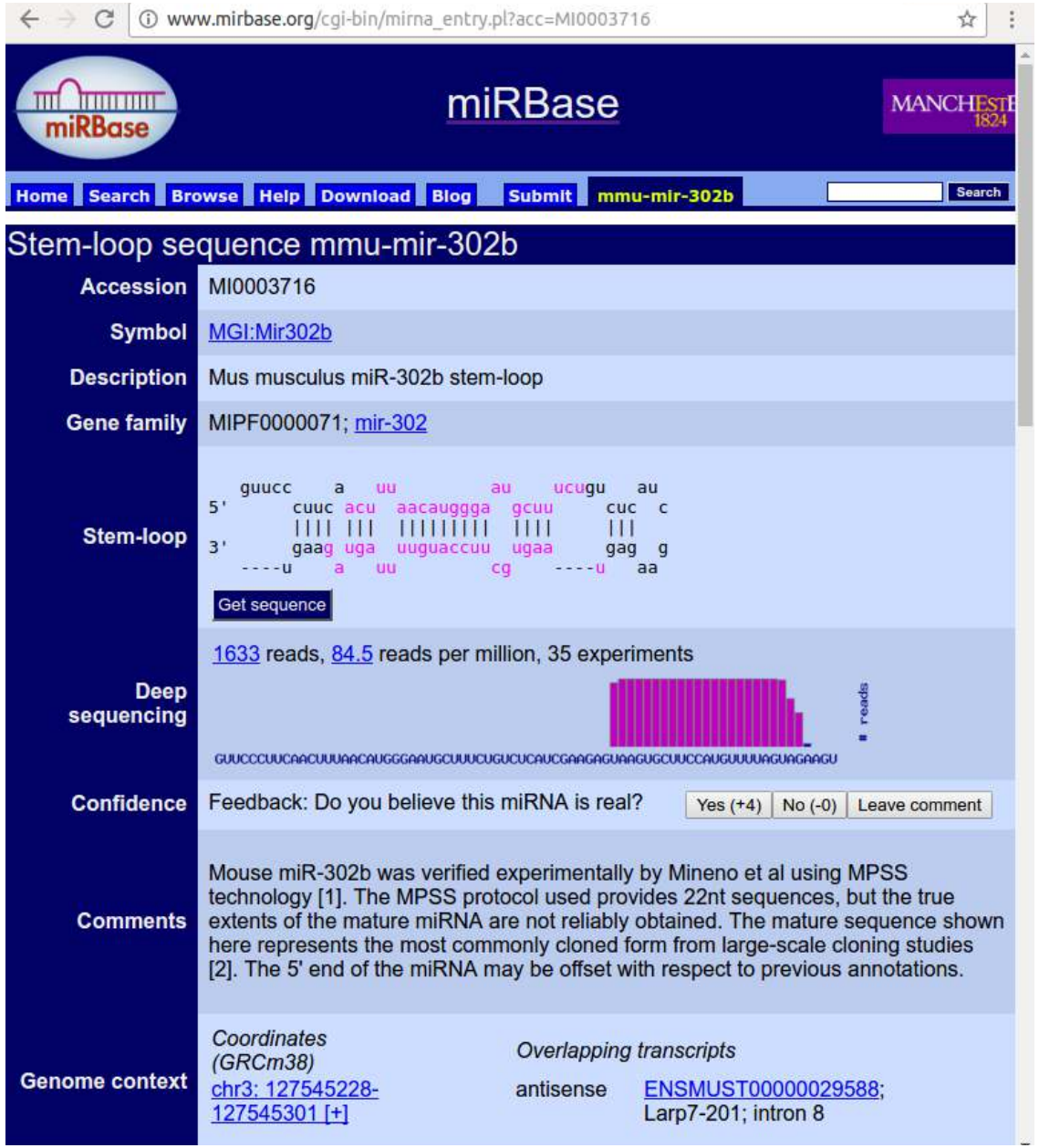
- http://www.mirbase.org/
- Содержит информацию по структуре и последовательности микроРНК 223 организмов.
- Идентификаторы:MI затем семь цифр. Например, микроРНК мыши mmu-mir-302b имеет идентификатор MI0003716.
- Содержит информацию по

✓Последовательности шпильки

✓Ссылки на публикации

✓Расположению в геноме

✓Зрелым микроРНК и др.



Информационная карточка микроРНК.

miRDB http://www.mirdb.org/miRDB/index.html

Содержит информацию о последовательностях микроРНК и их, спрогнозированным с помощью биоинформатики, мишеням для геномов человека, мыши, крысы, собаки и курицы.

Идентификаторы: название микроРНК. Например, hsa-miR-302b-3p.

- ✓ Приставка «mir» отделяется дефисом, вслед за ней следует номер, говорящий о порядке именования (mir-123 была открыта и названа раньше, чем mir-302).
- ✓ «mir-» обозначает пре-микроРНК, «MIR-» ген, кодирующий микроРНК, а «miR-» — для обозначения зрелой формы.
- ✓ К названию микроРНК с последовательностями, отличающимися на один или два нуклеотида, приписывается строчная буква (miR-123a и miR-123b).
- ✓ Пре-микроРНК, дающие начало на 100 % идентичным микроРНК, но локализованные в разных местах генома, имеют в названии цифру, отделенную дефисом (hsa-mir-194-1 и hsa-mir-194-2).
- ✓ Вид, из которого была выделена микроРНК, обозначается в названии трёхбуквенной приставкой (hsa-miR-123 человека).
- ✓ Когда две зрелые микроРНК образуются сразу с 3’ и 5’ концов исходной пре-микроРНК, к ним добавляется суффикс −3p или −5p. А если известен уровень экспрессии для этих микроРНК, тогда микроРНК с меньшей экспрессией помечают звёздочкой (miR-123 и miR-123* имеют общую исходную шпилечную пре-микроРНК, но в клетке обнаруживается больше miR-123).

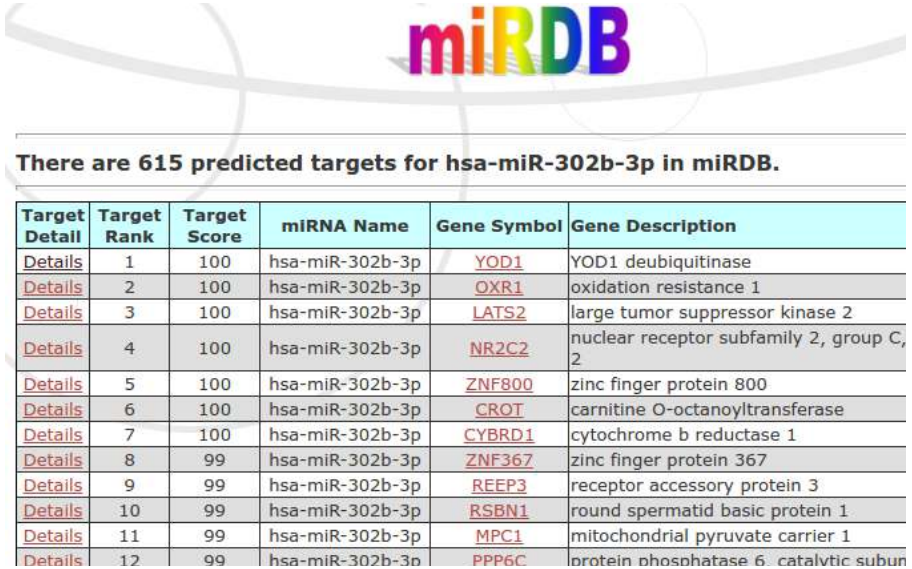
miRDB

MicroRNA and Target Gene Description:

miRNA Name	hsa-miR-302b-3p	**miRNA Sequence**	UAAGUGCUUCCAUGUUUUAGUAG
Previous Name	hsa-miR-302b		
Target Score	100	**Seed Location**	984, 1137, 1538, 1600, 3494
NCBI Gene ID	55432	**GenBank Accession**	NM_018566
Gene Symbol	YOD1	**3' UTR Length**	5171
Gene Description	YOD1 deubiquitinase		

3' UTR Sequence

```
   1 CCTATGCATG AATGAGGGTT GAAGCCTACT ACCTCACACA TCCAGAAGGC TCTGGGTTTT
  61 CCAATAAGCT ATGGTAACCC TAAAGAACAA AGGATACAAT GCTTGAACCA TCCTTTTAAC
 121 TTAAAACCAC TAAGACACTG AAATTCCTTG TTAAGATTAA AATTAGTGTG CAAGTTTACA
 181 GATGTGTGTC TACAGTGGTA AACTGTACAT ACATGCCTCT TTCTGCTGGA GTGACAGAAT
```



miRDB

There are 615 predicted targets for hsa-miR-302b-3p in miRDB.

Target Detail	Target Rank	Target Score	miRNA Name	Gene Symbol	Gene Description
Details	1	100	hsa-miR-302b-3p	YOD1	YOD1 deubiquitinase
Details	2	100	hsa-miR-302b-3p	OXR1	oxidation resistance 1
Details	3	100	hsa-miR-302b-3p	LATS2	large tumor suppressor kinase 2
Details	4	100	hsa-miR-302b-3p	NR2C2	nuclear receptor subfamily 2, group C, 2
Details	5	100	hsa-miR-302b-3p	ZNF800	zinc finger protein 800
Details	6	100	hsa-miR-302b-3p	CROT	carnitine O-octanoyltransferase
Details	7	100	hsa-miR-302b-3p	CYBRD1	cytochrome b reductase 1
Details	8	99	hsa-miR-302b-3p	ZNF367	zinc finger protein 367
Details	9	99	hsa-miR-302b-3p	REEP3	receptor accessory protein 3
Details	10	99	hsa-miR-302b-3p	RSBN1	round spermatid basic protein 1
Details	11	99	hsa-miR-302b-3p	MPC1	mitochondrial pyruvate carrier 1
Details	12	99	hsa-miR-302b-3p	PPP6C	protein phosphatase 6, catalytic subun

miRTarBase

- http://mirtarbase.mbc.nctu.edu.tw/
- Содержит информацию по экспериментально подтвержденным взаимодействиям микроРНК-мишень для 23 организмов.
- Идентификаторы: микроРНК человека hsa-miR-26b-5p имеет идентификатор MIRT029499.
- Содержит информацию по

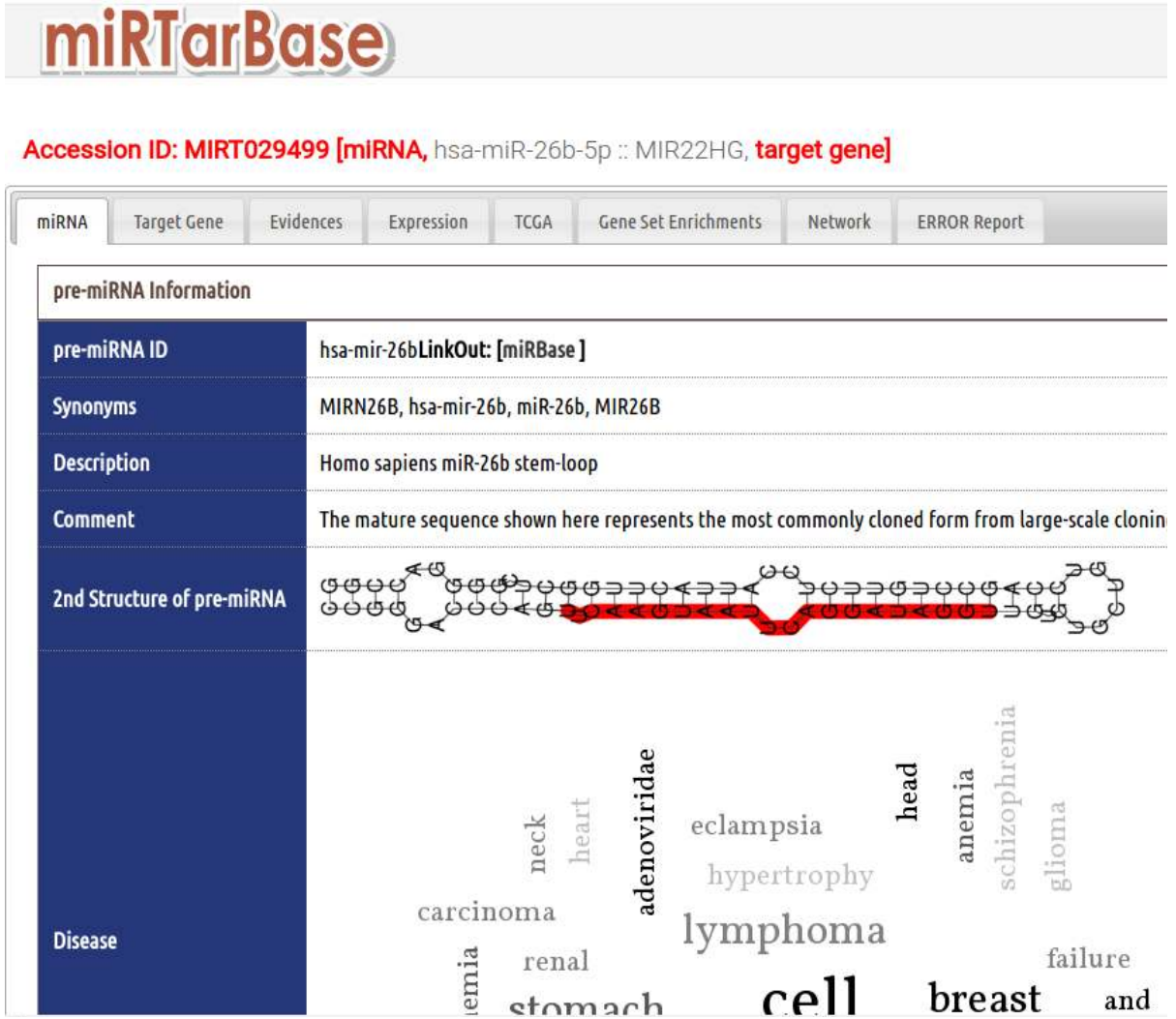
✓Мишеням (23054 генов для 4076 микроРНК)

✓Последовательности

✓Ассоциированным заболеваниям

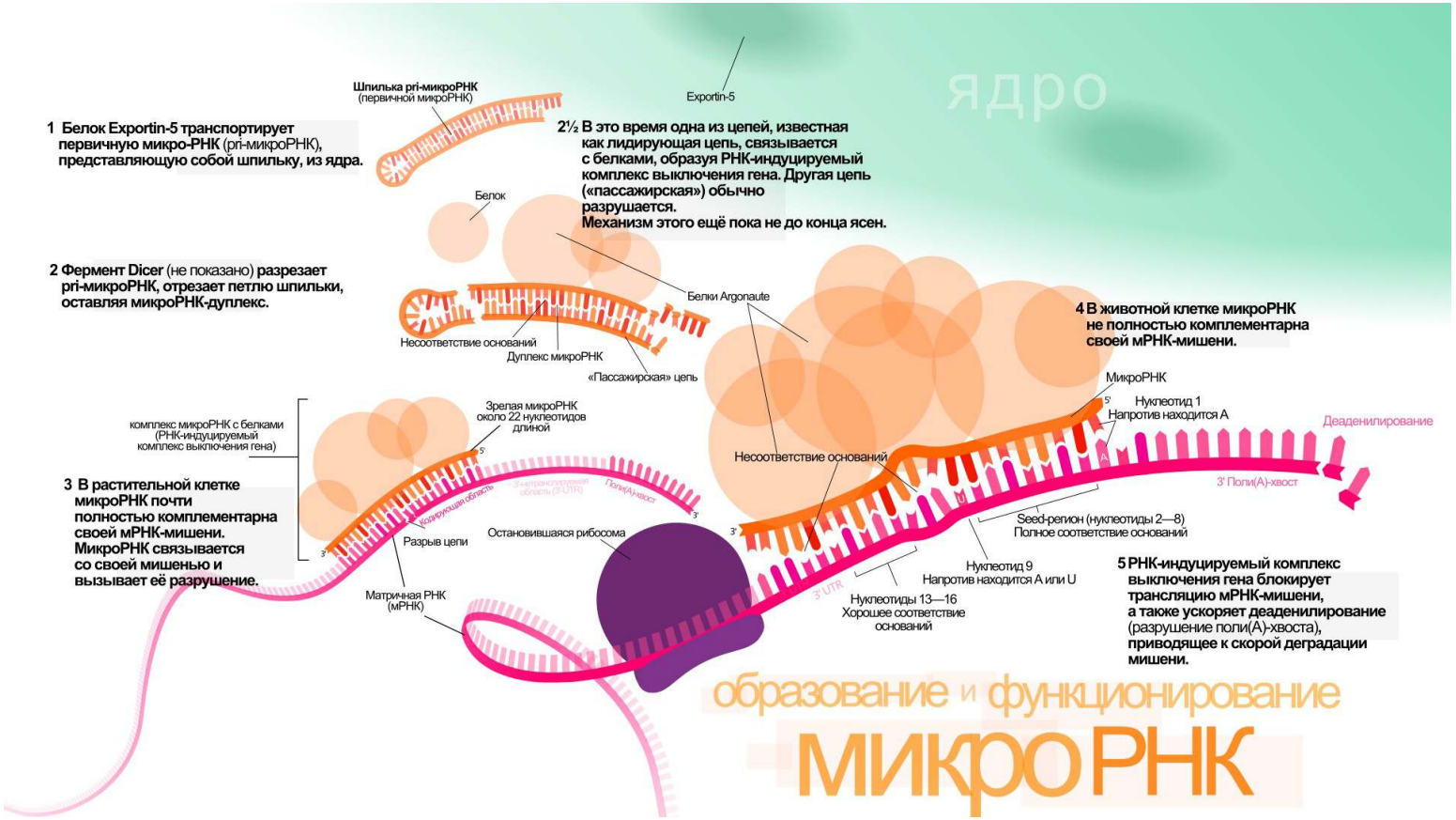
✓Ссылки на публикации и др.

Информационная карточка микроРНК.

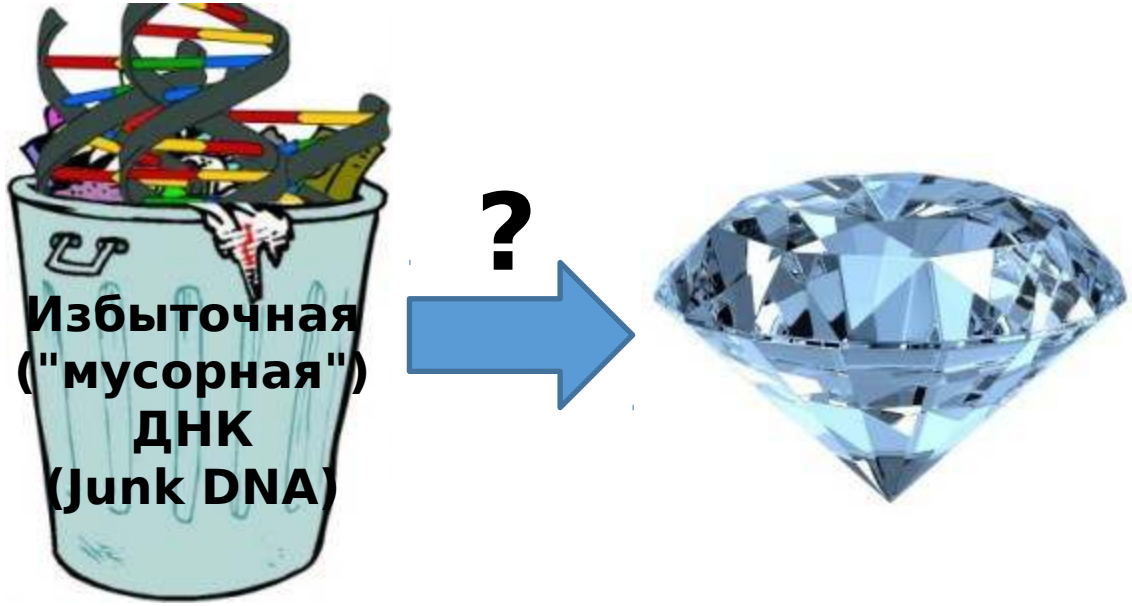


МикроРНК (англ. microRNA, miRNA) — малые некодирующие молекулы РНК длиной 18—25 нуклеотидов (в среднем 22), принимающие участие в транскрипционной и посттранскипционной регуляции экспрессии генов путём РНК-интерференции.

А какие еще классы некодирующих РНК Вы можете назвать?



1. tRNA - transfer RNA
2. siRNA - Small (short) interfering RNA
3. snoRNA - Small nucleolar RNAs
4. piRNA - Piwi-interacting RNA
5. circRNA - Circular RNA
6. shRNA - short (small) hairpin RNA
7. lncRNA - Long non-coding RNA
8. vlincRNA – very long non-coding RNA



Базы данных, содержащие информацию по метаболитам

ChEBI

PubChem

hmdb
The Human Metabolome Database

ChEBI

https://www.ebi.ac.uk/chebi/init.do

- Посвящена малым химическим соединениям, как природного, так и искусственного происхождения.

Содержит информацию по

✓ функциям

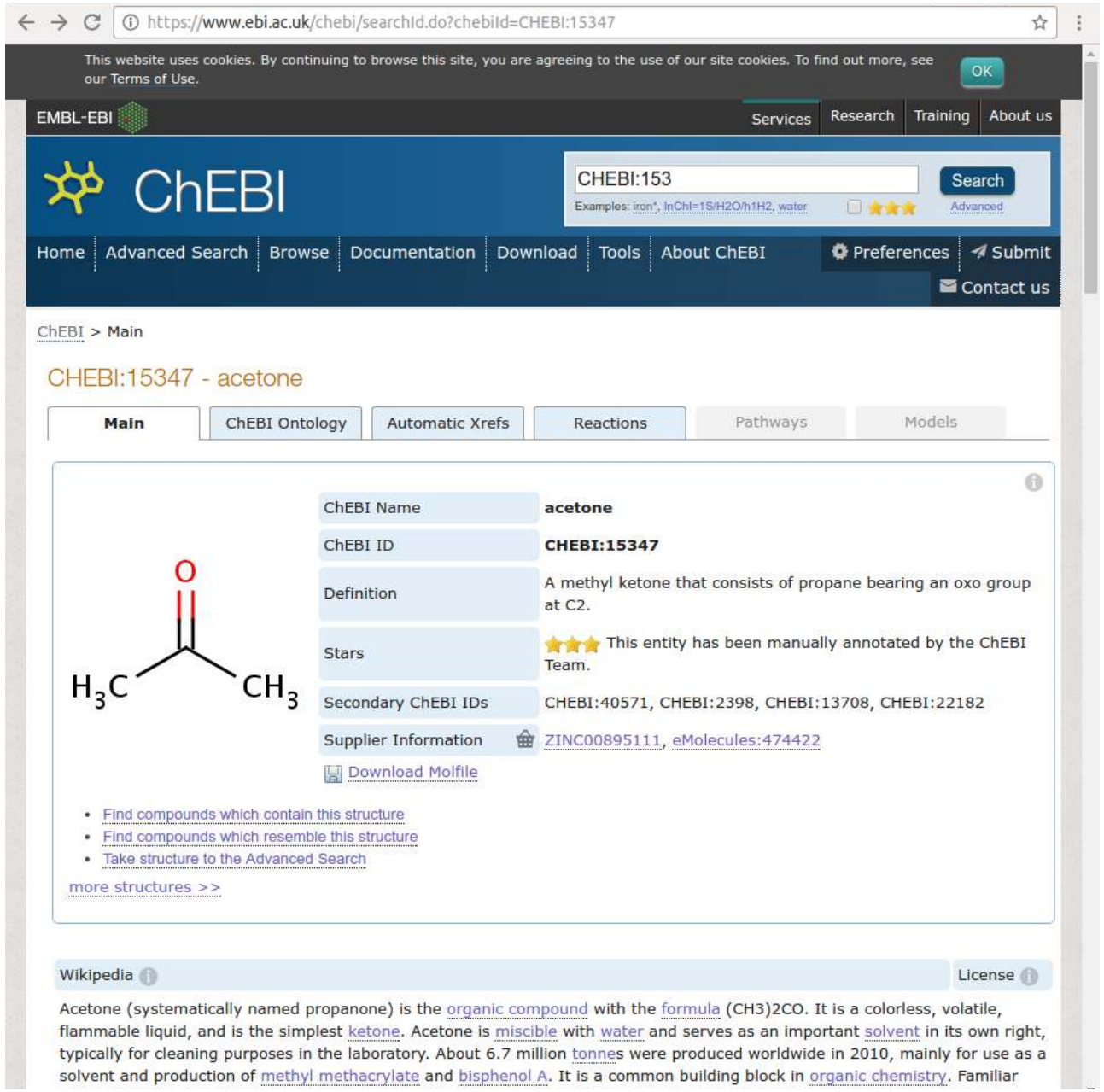
✓химической формуле

✓молекулярному весу

✓биологической роли

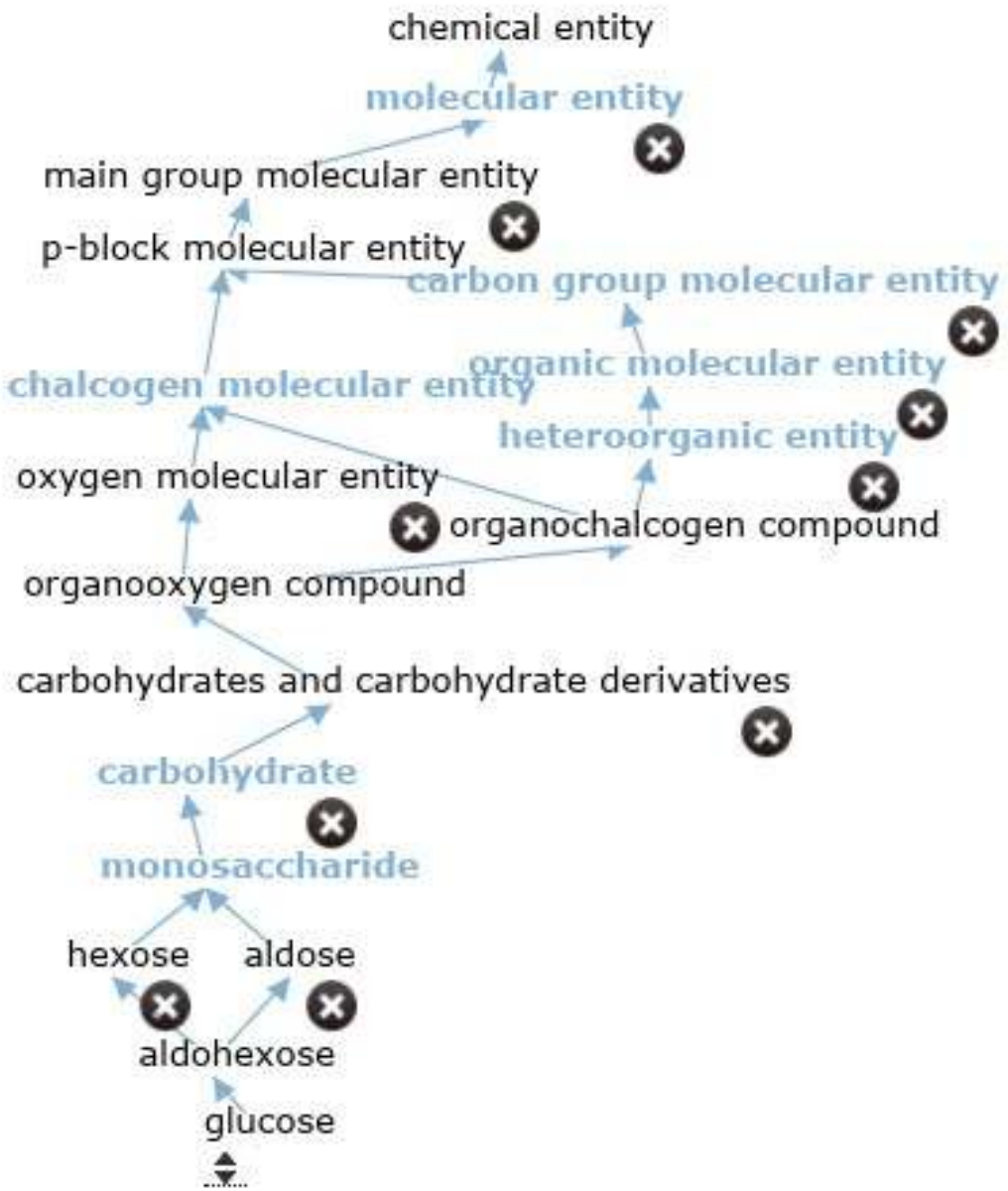
✓онтологии (родство с другими химическими соединениями)

Информационная карточка метаболита



ChEBI

Онтология



CHEBI:24431 chemical entity
- CHEBI:23367 molecular entity
  - CHEBI:33579 main group molecular entity
    - CHEBI:33675 p-block molecular entity
      - CHEBI:33582 carbon group molecular entity
        - CHEBI:50860 organic molecular entity
          - CHEBI:33285 heteroorganic entity
            - CHEBI:36962 organochalcogen compound
              - CHEBI:36963 organooxygen compound
                - CHEBI:36586 carbonyl compound
                  - CHEBI:17087 ketone
                    - CHEBI:51867 methyl ketone
                      - **CHEBI:15347 acetone**
                        - CHEBI:48057 2-oxopropylidene group
                        - CHEBI:48059 2-oxopropylidyne group
                        - CHEBI:47220 chloroacetone
                        - CHEBI:51845 bromoacetone
                        - CHEBI:27957 hydroxyacetone
                        - CHEBI:17906 aminoacetone
                        - CHEBI:37551 1-hydroxy-3-methoxyacetone
                        - CHEBI:48056 acetonyl group
                    - CHEBI:26292 propanones
                      - **CHEBI:15347 acetone**

Relationship Types

- is a
- has part
- is conjugate base of
- is conjugate acid of
- is tautomer of
- is enantiomer of
- has functional parent
- has parent hydride
- is substituent group from
- has role

Status

- Checked
- Unchecked

PubChem

https://pubchem.ncbi.nlm.nih.gov/compound/

- Посвящена малым химическим соединениям, как природного, так и искусственного происхождения.
- Содержит информацию по

✓синонимам

✓ функциям и физико-химическим свойствам

✓химической формуле

✓молекулярному весу

✓биологической роли

✓производителям и коммерческим названиям

✓патентам

Информационная карточка метаболита

HMDB

http://www.hmdb.ca/

Посвящена метаболитам человека.

- Содержит информацию по

✓синонимам

✓ функциям и физико-химическим свойствам

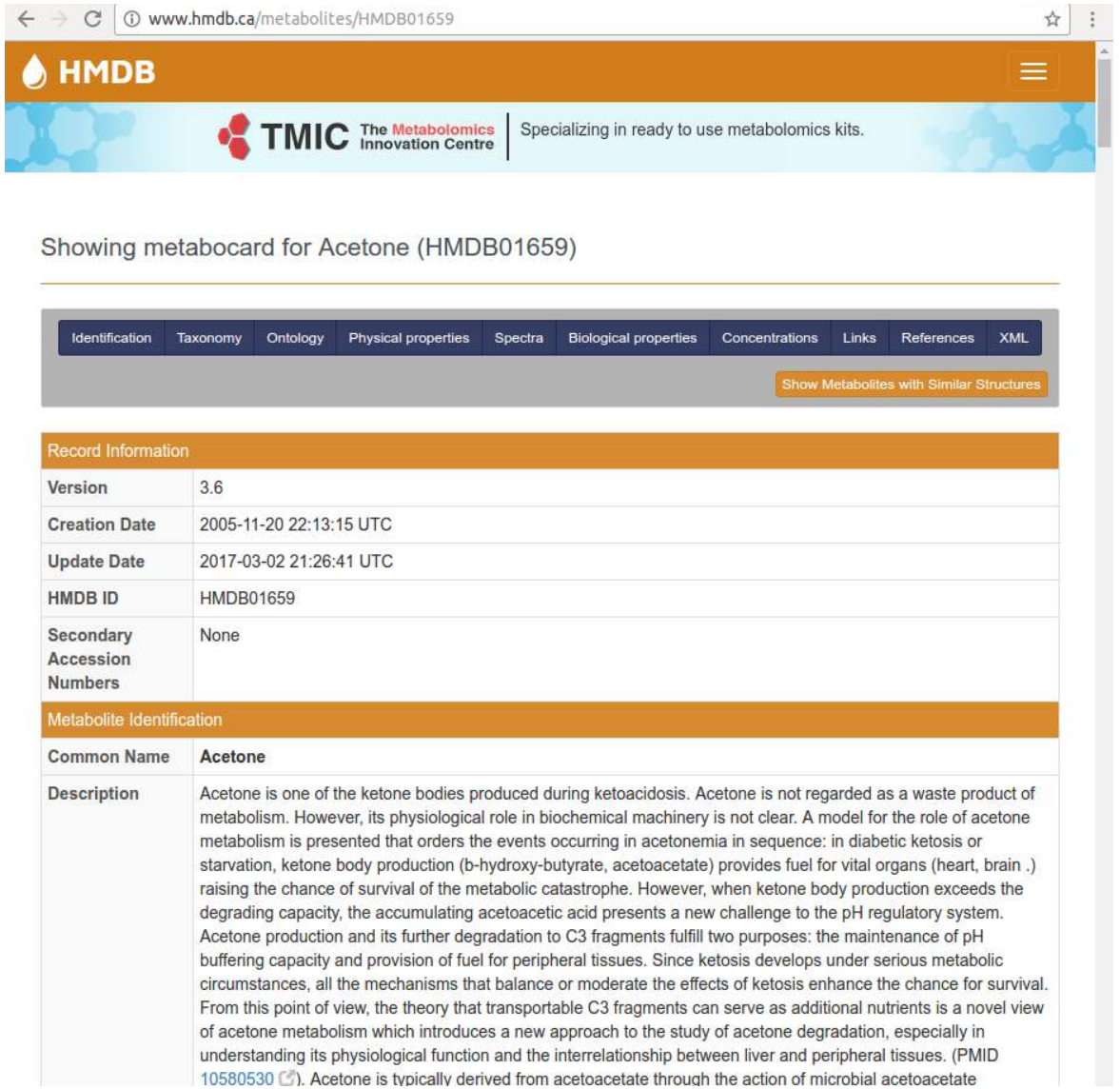
✓химической формуле, молекулярному весу

✓ассоциированным заболеваниям

✓концентрации в норме и при патологии в различных тканях

✓биологической роли и др.

Информационная карточка метаболита



Базы данных, содержащие информацию по заболеваниям.

OMIM®

ICD-11

DISEASE ONTOLOGY

OMIM

Информационная карточка заболевания

- https://www.omim.org/
- Наследуемые заболевания человека
- Содержит информацию по

✓синонимам

✓ ассоциациям с генами

✓наследуемости

✓биохимических особенностях

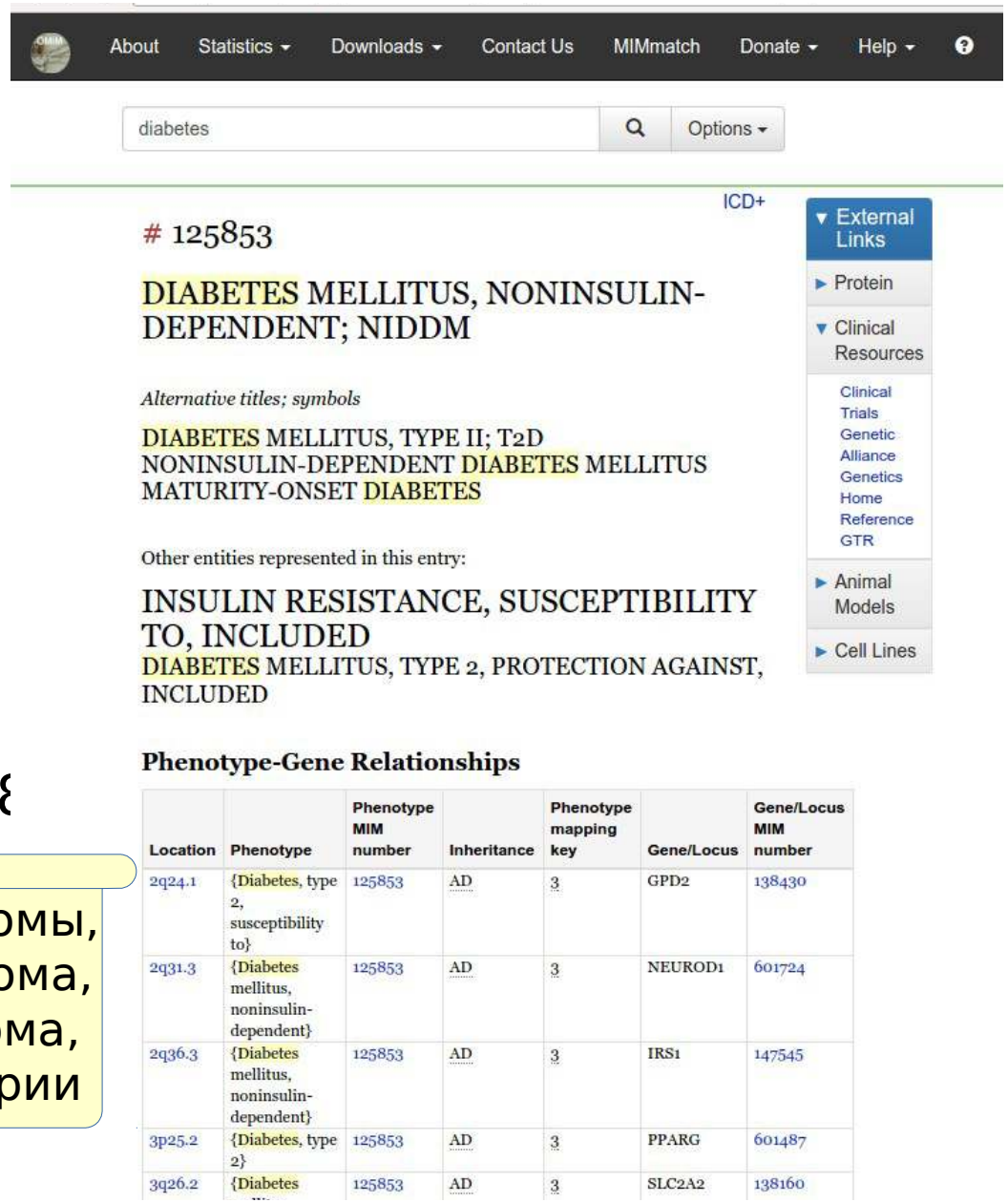
✓патогенезе

✓животным моделям и др.

Идентификаторы: #1258530 (или MIM:1258

* - ген
# - фенотип, не один локус
+ -ген/фенотип (напр. + 159555)
%- фенотип, неизвестны гены
^ - устаревший идентификатор

1,2,6 – аутосомы,
3 – X-хромосома,
4 – Y-хромосома,
5 - митохондрии



Phenotype-Gene Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key	Gene/Locus	Gene/Locus MIM number
2q24.1	{Diabetes, type 2, susceptibility to}	125853	AD	3	GPD2	138430
2q31.3	{Diabetes mellitus, noninsulin-dependent}	125853	AD	3	NEUROD1	601724
2q36.3	{Diabetes mellitus, noninsulin-dependent}	125853	AD	3	IRS1	147545
3p25.2	{Diabetes, type 2}	125853	AD	3	PPARG	601487
3q26.2	{Diabetes	125853	AD	3	SLC2A2	138160

ICD11 - International Statistical Classification of Diseases and Related Health Problems 11th Revision

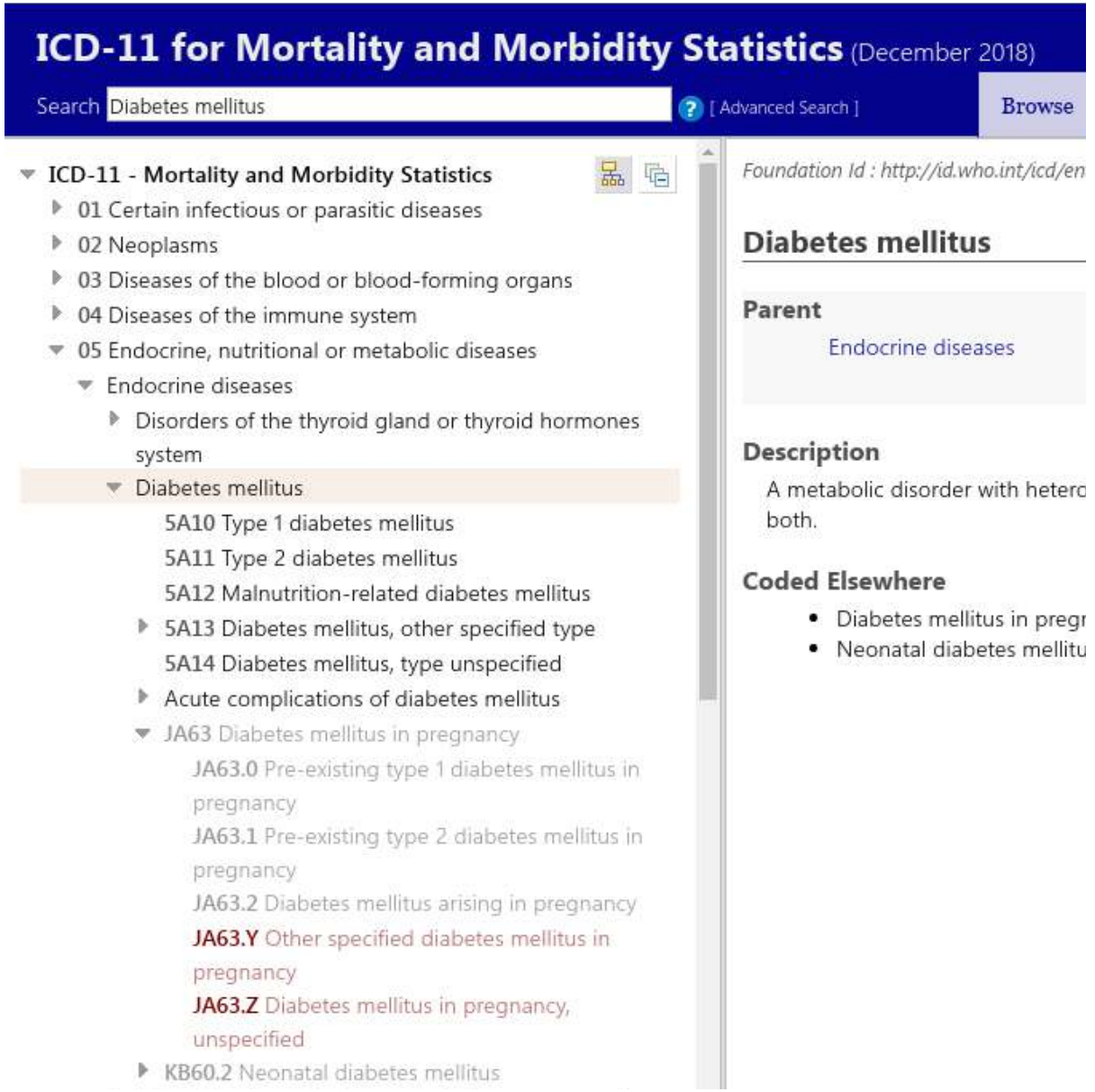
- https://icd.who.int/
- Содержит классификацию заболеваний человека, принятую международным сообществом и используемую практикующими врачами.

Идентификаторы: от 1A00.00 до ZZ9Z.ZZ

Подробнее в Reference guide раздел 2.2.1

- Нелирическое отступление:
- Как найти идентификаторы в текстовых редакторах с помощью ctrl+F?

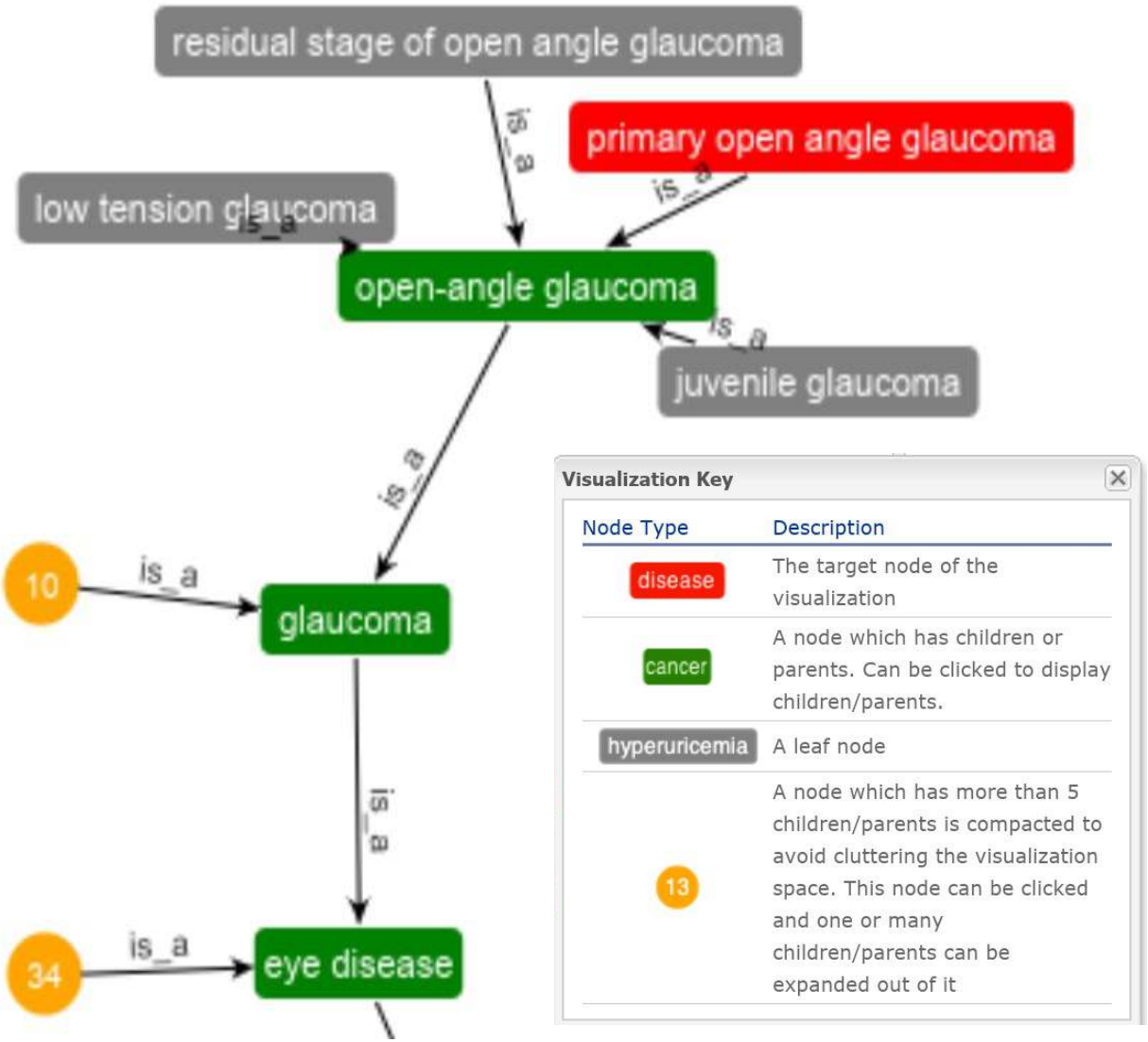
Использовать регулярные выражения (regexp) или подстановочные знаки. Пример : \w[A-Z]\d\w\.\w{2} \w{4}\.\w{2}



Disease Ontology

- Доступна через интернет по адресу http://disease-ontology.org/
- Содержит классификацию заболеваний человека по 8 основным группам.
- Идентификаторы DOID:число. Например, Diabetes mellitus это раздел с идентификаторами от E10 до E14. Заболевание diabetes mellitus имеет идентификатор DOID:9351 и относится к группе glucose metabolism disease -> carbohydrate metabolism disease -> acquired metabolic disease -> disease of metabolism.
- Содержит краткую информацию по патогенезу заболевания и ссылки на другие источники.

Онтология



Оцените, сколько в настоящее время известно заболеваний?

Оцените, сколько в настоящее время известно заболеваний?

В МКБ-10 описано около 12,5 тысяч заболеваний (МКБ-10 одобрена ВАЗ в 1990 году), .

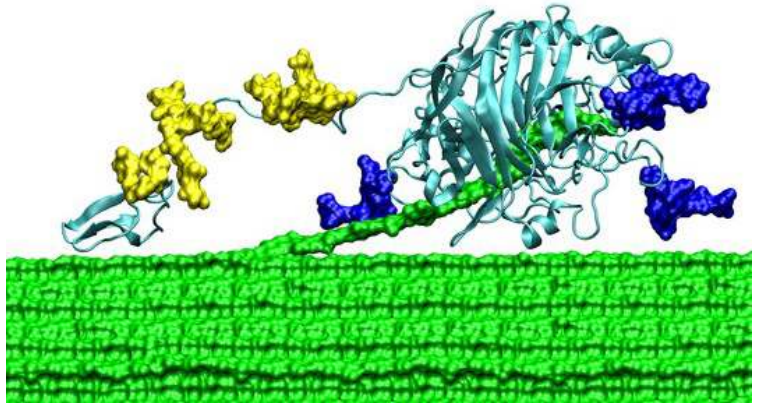
В МКБ-11 – около 16 тысяч (одобрена в мае 2019).

Базы данных, содержащие информацию по биологическим путям/процессам.

Что такое биологический процесс?

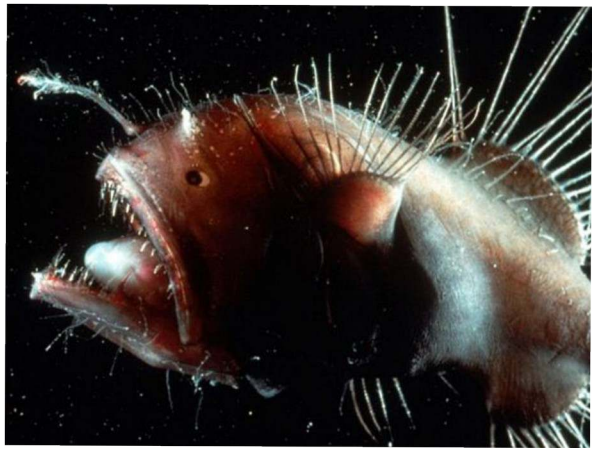
функция

Биохимическая активность



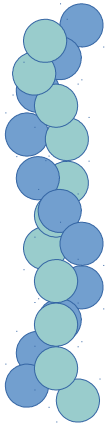
Molecular function (MF)

Биологическая цель



Biological process (BP)

Структура



Cellular component (CC)

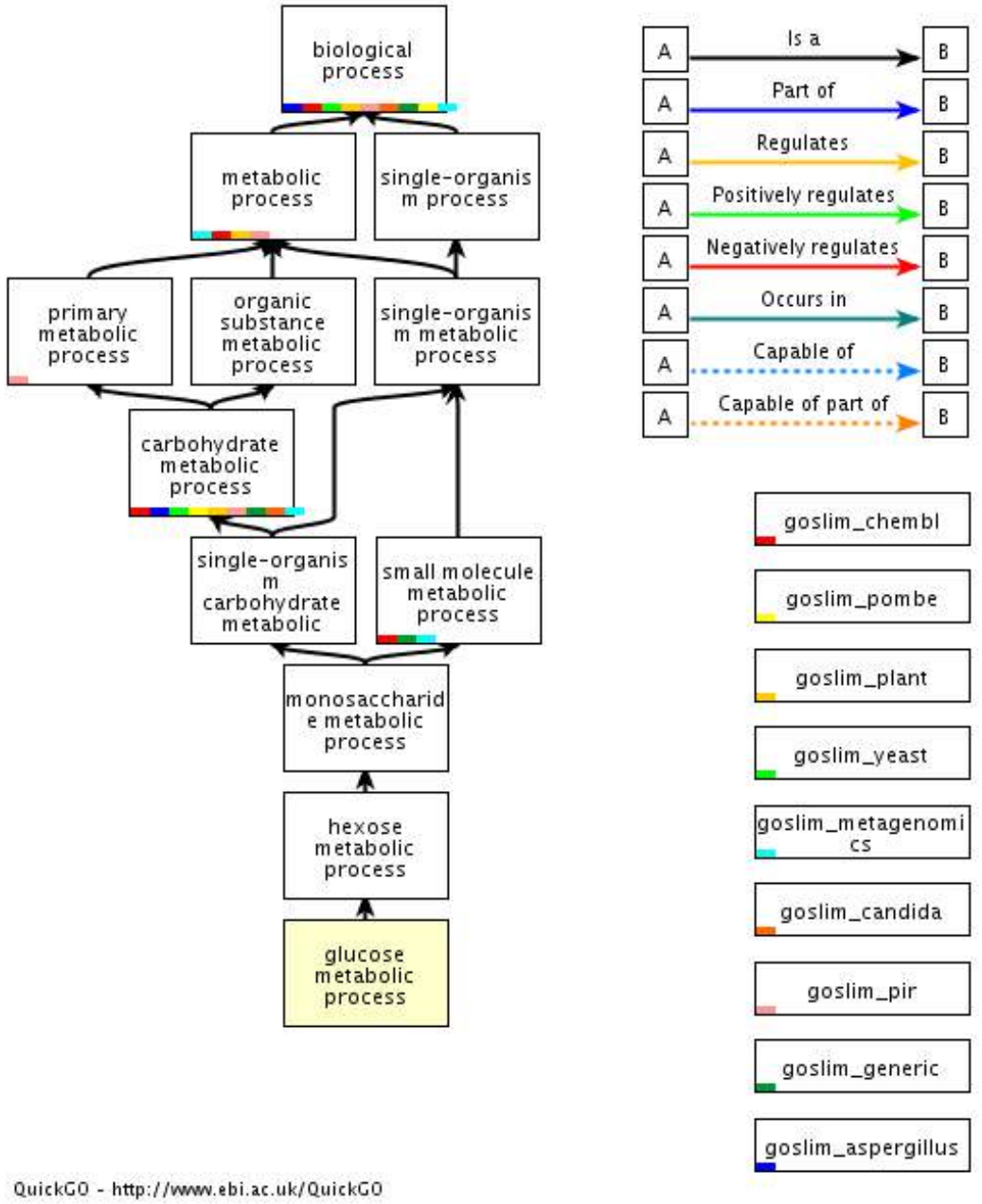
Gene Ontology

- http://www.geneontology.org/
- Является онтологией, описывающей биологические процессы (BP), клеточные компоненты (CC) и молекулярные функциям (MF). Имеется привязка к генам различных организмов).
- Идентификаторы: GO:семизначное число. Например, glucose metabolic process имеет идентификатор GO:0006006.
- Содержит:

✓ Синонимы

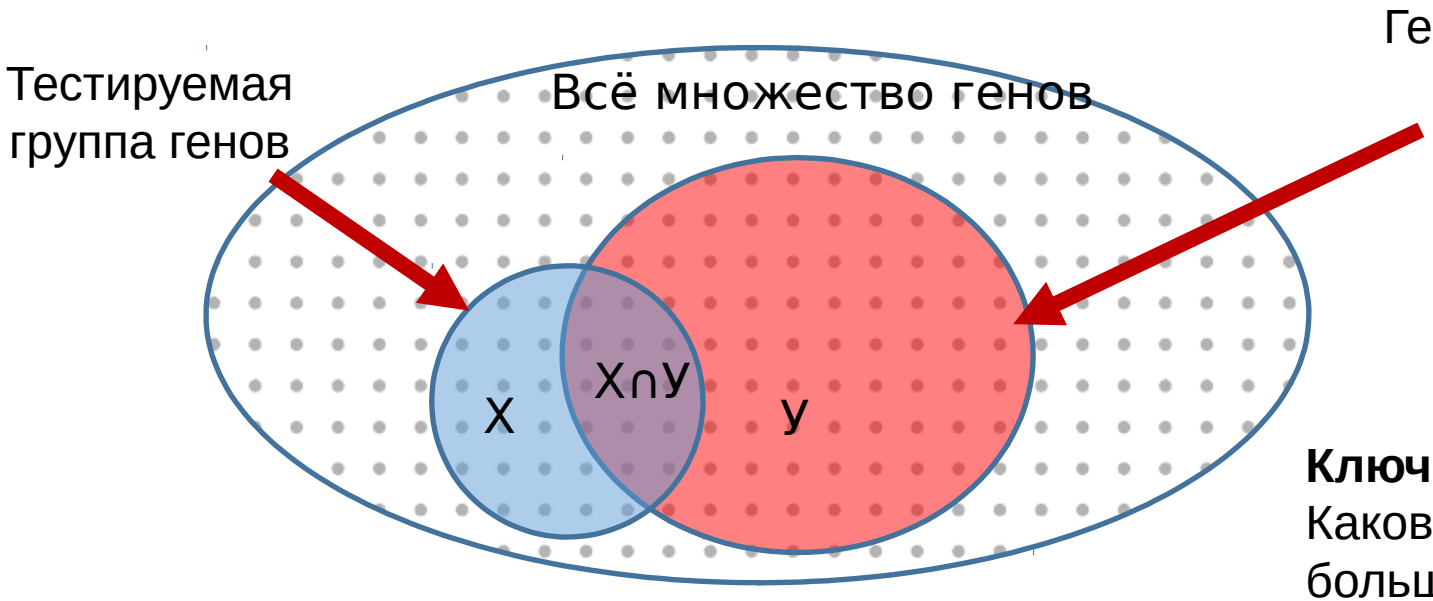
✓ определение

✓ ассоциации с генами (https://www.ebi.ac.uk/GOA)



Поиск сверхпредставленных биологических процессов

Сверхпредставленность процесса (фенотипа, заболевания) – повышенная представленность генов, приписанных данному процессу в исследуемой группе генов. Названия на английском: Over-represented biological processes, Gene Enrichment Analysis, Gene Ontology Enrichment Analysis.

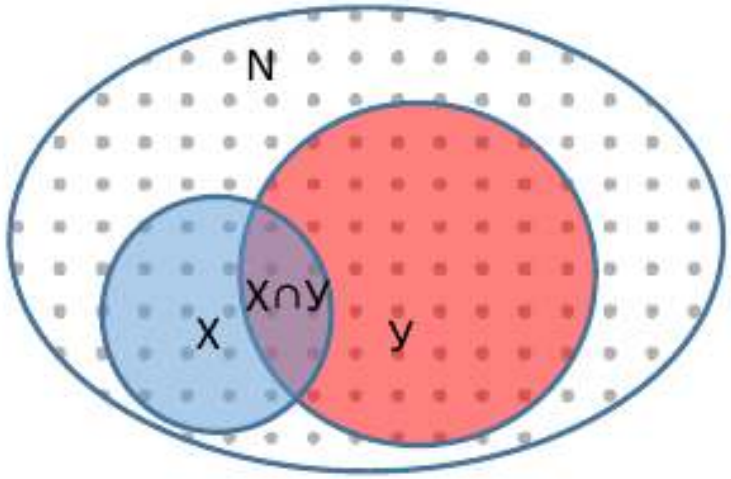


Ключевой вопрос:
Какова вероятность ожидать столь большое
пересечение по случайным причинам?

Алгоритм:

1. Выявляется тестируемая группа генов X (например, с помощью эксперимента типа случай/контроль).
2. Выявляется набор генов Y, относящихся к какому-то термину Gene Ontology.
3. С помощью точного критерия Фишера (гипергеометрического распределения) или других статистических методов оценивается вероятность обнаружить такое же или большее пересечение между множествами по случайным причинам (p-value). При этом учитываются такие величины как X∩Y, X, Y и число всех генов.
4. Процедура из пунктов 2, 3 повторяется для всех Gene Ontology терминов. Так как всего Gene Ontology терминов много (порядка 10 тысяч), необходимо делать поправку на множественное сравнение, например, Бенджамини-Хокберга.

Точный критерий Фишера



Что наблюдается? Таблица сопряжённости (2x2)

	Число генов в наборе X	Число генов **не** в наборе X	Полное число генов
Число генов в наборе Y	|X∩У|=5	|Y|-|X∩У|=1	|Y|=6
Число генов **не** в наборе Y	|X|-|X∩У|=1	N-|XUY|=23500--6-6+5=23493	N-|Y|=23500-6==23494
Полное число генов	|X|=6	N-|X|=23500-6==23494	N=23500

Что ожидается по случайным причинам?

$$N(N-1)(N-2)\ldots(N-|X|+1)=\frac{N!}{(N-|X|)!}$$

Ген А, ген Б,... ген В

...

Ген Б, ген А,... ген В

⇒ |X|!

$$\frac{N!}{(N-|X|)!\,X!}=C_N^{|X|}$$

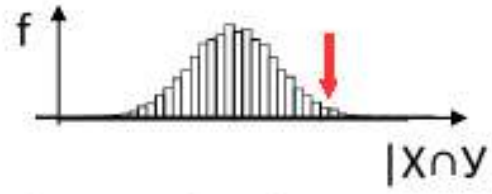
Полное число вариантов $X_{случайный}$

Сколько вариантов генов из $Y_{случайный}$ окажутся в $X_{случайный}$, если известно, что их |X∩У| (как в наблюдении)?

$$C_{|Y|}^{|X\cap Y|}$$

Сколькими способами можно заполнить оставшиеся места в $X_{случайный}$ генами не из $Y_{случайный}$ $C_{N-|Y|}^{|X|-|X\cap Y|}$

Какова ожидаемая частота (вероятность)?



$$\frac{C_{|Y|}^{|X\cap Y|}\cdot C_{N-|Y|}^{|X|-|X\cap Y|}}{C_N^{|X|}}$$

$$p\text{-значение}=\frac{C_6^5\cdot C_{23494}^1}{C_{23500}^6}+\frac{C_6^6\cdot C_{23494}^0}{C_{23500}^6}=6{,}03\cdot10^{-19}$$

Как посчитать в R?

```
x<-matrix(c(5,1,1,23493),2,2)
fisher.test(x, alternative = "greater")
```

Пакеты Bioconductor: GOStats, TopGO

KEGG - Kyoto Encyclopedia of Genes and Genomes

- http://www.genome.jp/kegg/
- Содержит описание биологических процессов на молекулярно-генетическом уровне, составленное вручную экспертами.
- Идентификаторы: трехбуквенный идентификатор организма и пятизначное число. Например, Glycolysis/Gluconeogenesis имеет идентификатор hsa00010.
- Содержит информацию по

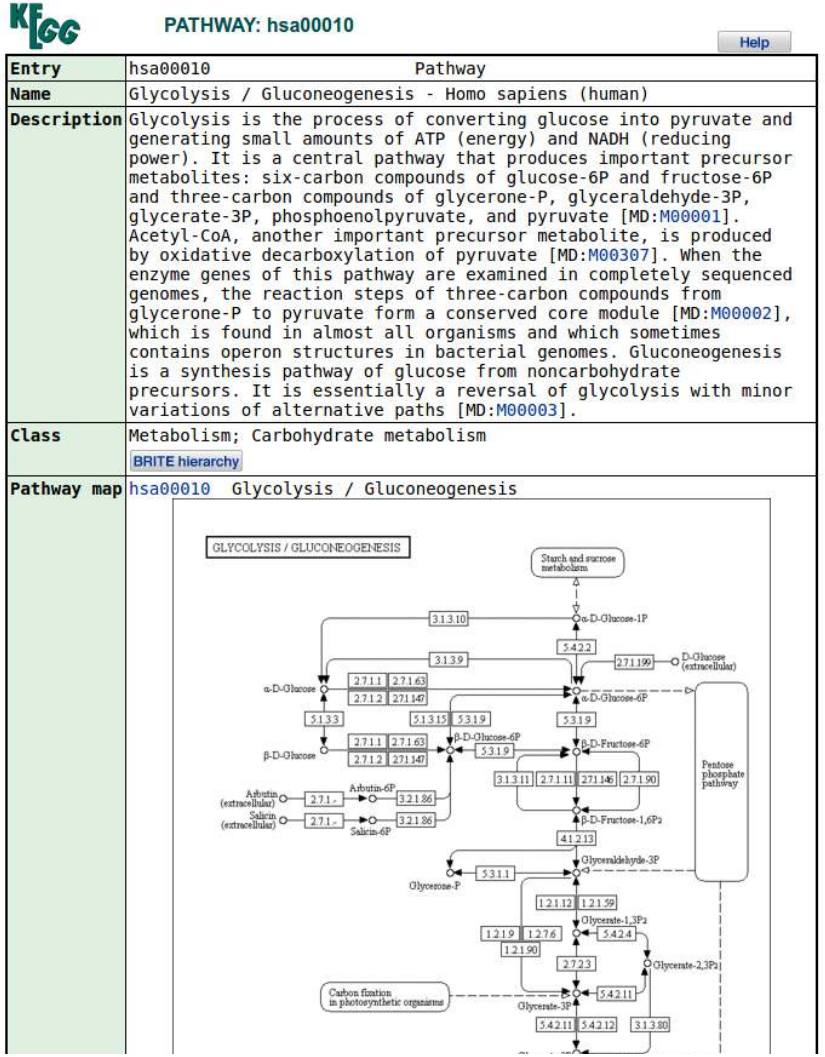
✓Описанию

✓Входящим в процесс генам/белкам, метаболитам и др.

✓Ассоциированным заболеваниям

✓ ссылкам на литературу и др.

Информационная карточка



PATHWAY: hsa00010

Help

Entry	hsa00010 Pathway
Name	Glycolysis / Gluconeogenesis - Homo sapiens (human)
Description	Glycolysis is the process of converting glucose into pyruvate and generating small amounts of ATP (energy) and NADH (reducing power). It is a central pathway that produces important precursor metabolites: six-carbon compounds of glucose-6P and fructose-6P and three-carbon compounds of glycerone-P, glyceraldehyde-3P, glycerate-3P, phosphoenolpyruvate, and pyruvate [MD:M00001]. Acetyl-CoA, another important precursor metabolite, is produced by oxidative decarboxylation of pyruvate [MD:M00307]. When the enzyme genes of this pathway are examined in completely sequenced genomes, the reaction steps of three-carbon compounds from glycerone-P to pyruvate form a conserved core module [MD:M00002], which is found in almost all organisms and which sometimes contains operon structures in bacterial genomes. Gluconeogenesis is a synthesis pathway of glucose from noncarbohydrate precursors. It is essentially a reversal of glycolysis with minor variations of alternative paths [MD:M00003].
Class	Metabolism; Carbohydrate metabolism BRITE hierarchy
Pathway map	hsa00010 Glycolysis / Gluconeogenesis

Reactome

- http://www.reactome.org/
- Содержит описание биологических процессов на молекулярно-генетическом уровне, составленное вручную экспертами.
- Идентификаторы: код группы процессов - трехбуквенный идентификатор организма и пятизначное число. Например, Glucose metabolism имеет идентификатор R-HSA-70326.
- Содержит информацию по

✓Описанию

✓Входящим в процесс генам/белкам, метаболитам и др.

✓Положению в иерархической структуре процессов

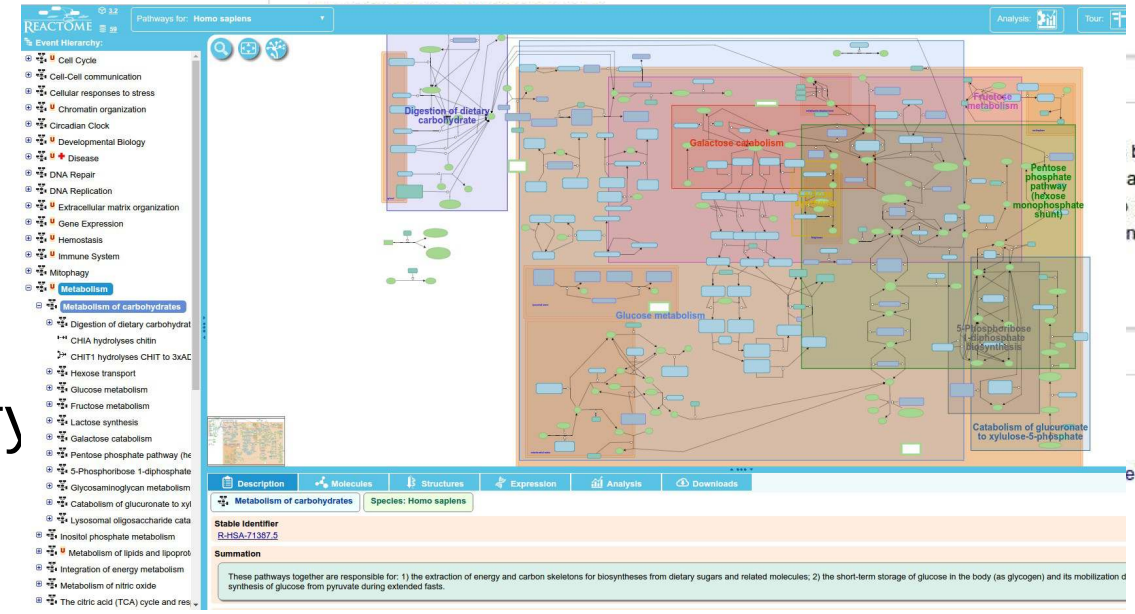
✓ «ортологичным» процессам и др.

Информационная карточка

Glucose metabolism

Stable Identifier	R-HSA-70326
Type	Pathway
Species	Homo sapiens

Locations in the PathwayBrowser



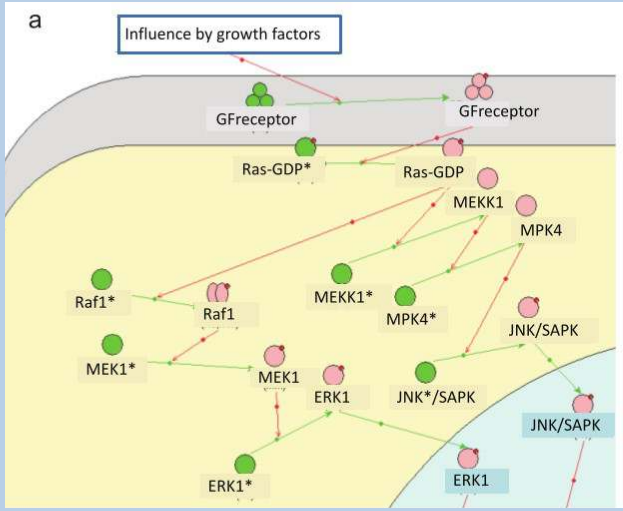
План лекции

1. Введение.
2. Обзор Интернет-ресурсов.
3. Ресурсы, интегрирующие биологическую информацию из разнородных источников и представляющие ее в виде генных сетей: ANDSystem, STRING, GeneMania, Pathway Commons.
4. Практическое применение инструментов интеграции.

Два типа инструментов интеграции

1. Ручная реконструкция генных сетей экспертами с использованием специальных программных средств – редакторов генных сетей.

ИЦиГ СО РАН, 1998, N.A. Kolchanov, E.A. Ananko, N.L. Podkolodny, I.L. Stepanenko, E.V. Ignatieva, O.A. Podkolodnaya et al.



MAP-киназный путь передачи сигнала в ядро клетки, контролирующий процесс клеточного деления, активируемый ростовыми факторами.

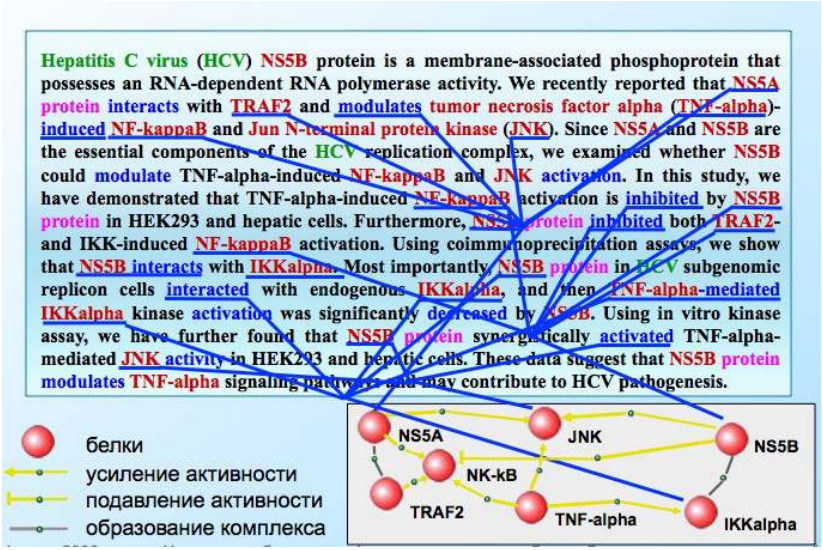
2. Автоматическое извлечение знаний о молекулярно-генетических взаимодействиях различного типа из текстов научных публикаций и баз данных компьютерными методами (методы text-mining). Примерами являются ANDSystem, STRING, GeneMania, Pathway Commons и др.

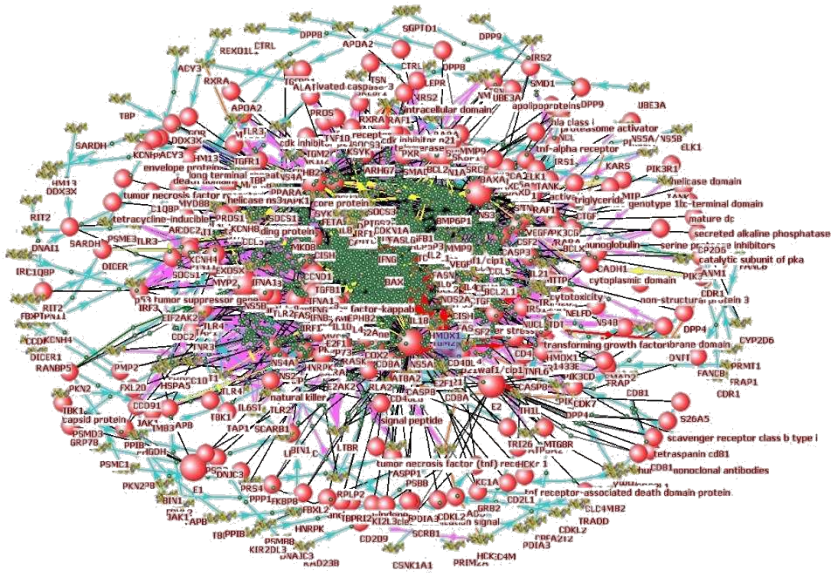
Система ANDSystem для автоматического извлечения знаний из текстов научных публикаций, международных патентов и баз данных в области биомедицины.

В ИЦиГ СО РАН разработана компьютерная технология автоматического извлечения знаний из текстов научных публикаций и международных патентов ANDSystem. Проведен автоматический анализ более 25 млн. научных публикаций и 10 млн. международных патентов.

Система ANDSystem включает
- модуль онтологии предметной области
- модуль текст майнинг
- базу знаний
- интерфейс пользователя.

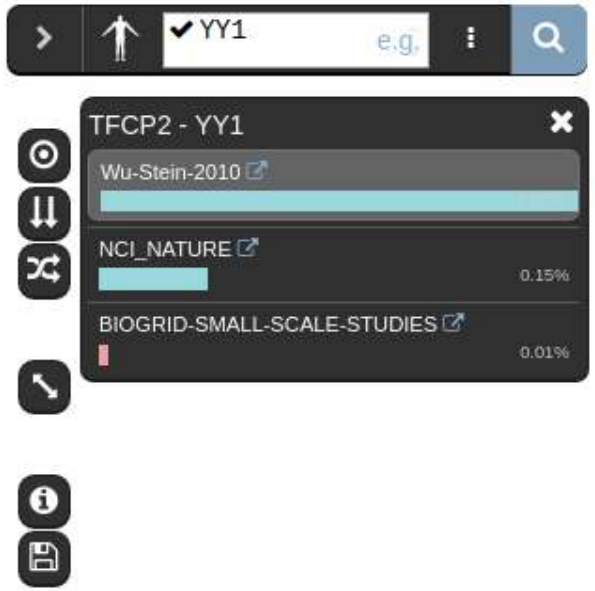
Создана база знаний, содержащая более 15 млн. фактов, значимых для биомедицины, включающих взаимосвязи между
- молекулярно-генетическими системами и процессами
- заболеваниями и фенотипическими признаками
- факторами окружающей среды и др.

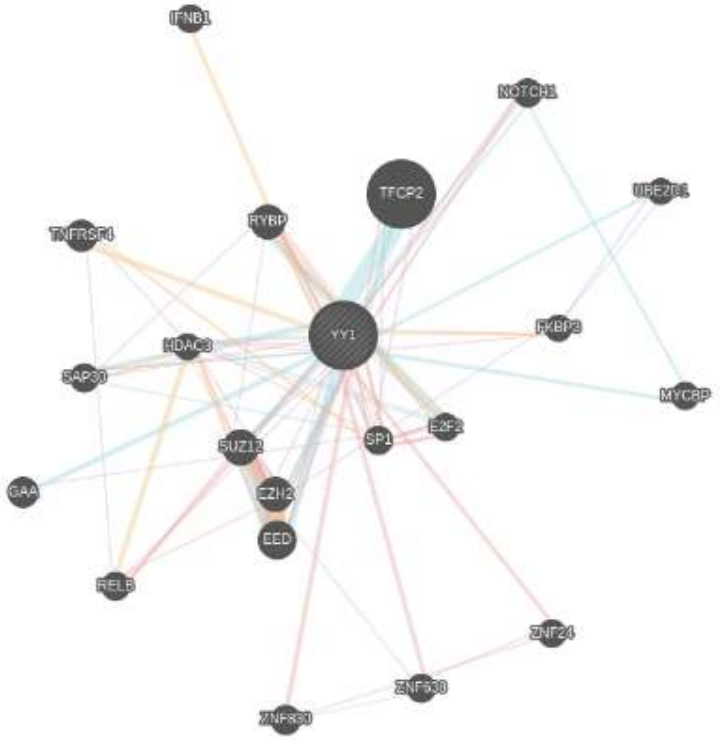


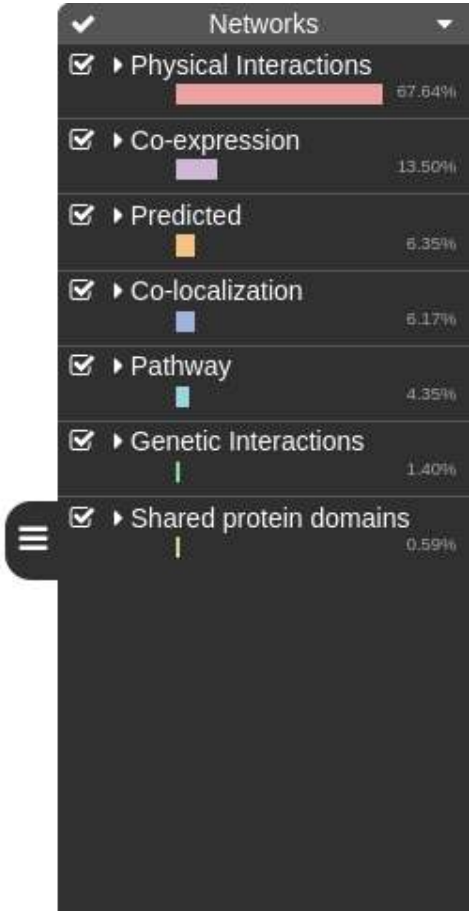


Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. BMC Systems Biology 2015, 9(Suppl 2):S2.

GeneMANIA

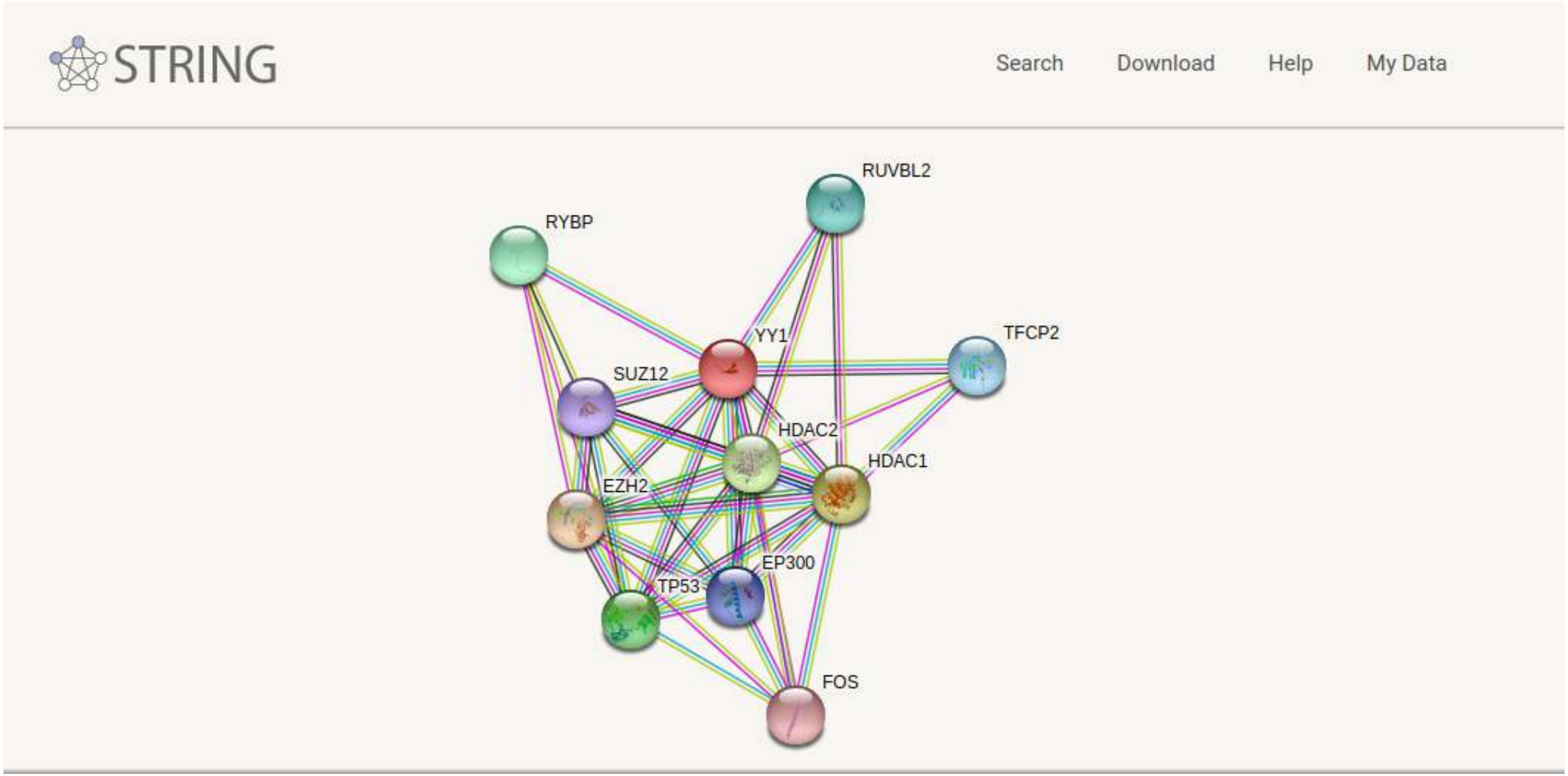






Также есть в виде Cytoscape app

STRING


STRING
Search
Download
Help
My Data
RUVBL2
RYBP
YY1
TFCP2
SUZ12
HDAC2
HDAC1
EZH2
EP300
TP53
FOS

Pathway Commons

База данных содержит 69 498 биологических путей

Pathway Commons ресурс, интегрирующий информацию из различных баз данных. Биологические пути, представленные в Pathway Commons загружены непосредственно из исходных баз данных. Реконструкция биологических путей в исходных базах данных могла проводиться как путем ручного извлечения информации из литературы, так и путем компьютерной автоматической реконструкции. Качество биологических путей в Pathway Commons зависит от качества путей исходных баз данных. Pathway Commons позволяет пользователям фильтровать данные по различным критериям, включая источник информации.

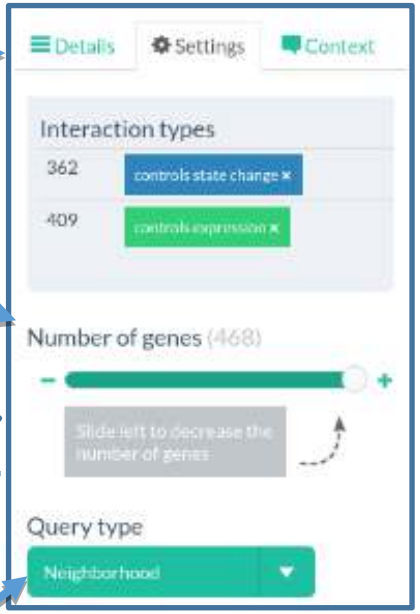
Интегрирует информацию из 24 баз данных

Reactome: 2007 pathways, 14427 interactions, 35835 participants
NCI Pathway Interaction Database: Pathway: 745 pathways, 14707 interactions, 10531 participants
PhosphoSitePlus: 27692 interactions, 15458 participants
HumanCyc: 302 pathways, 7102 interactions, 5896 participants
HPRD: 40595 interactions, 9844 participants
PANTHER Pathway: 272 pathways, 4700 interactions, 6703 participants
Database of Interacting Proteins: 8218 interactions, 4671 participants
BioGRID: 322538 interactions, 645241 participants
IntAct: 150549 interactions, 403729 participants
BIND: 35279 interactions, 74675 participants
CORUM:4401 participants
TRANSFAC: 427 pathways, 261624 interactions, 13276 participants
miRTarBase: 5 pathways, 51214 interactions, 12775 participants
DrugBank: 19297 interactions, 15854 participants
Recon X: 1 pathways, 10813 interactions, 8316 participants
Comparative Toxicogenomics Database: 32722 pathways, 390428 interactions, 61031 participants
KEGG Pathway: 122 pathways, 3566 interactions, 3355 participants
Small Molecule Pathway Database: 1206 pathways, 4701 interactions, 4863 participants
Integrating Network Objects with Hierarchies: 774 pathways, 5432 interactions, 17142 participants
NetPath: 27 pathways, 6347 interactions, 3266 participants
WikiPathways: 333 pathways, 9758 interactions, 9584 participants
ChEBI: All names
SwissProt: All names
UniChem: All names

Текстовое поле для ввода названия интересующего гена

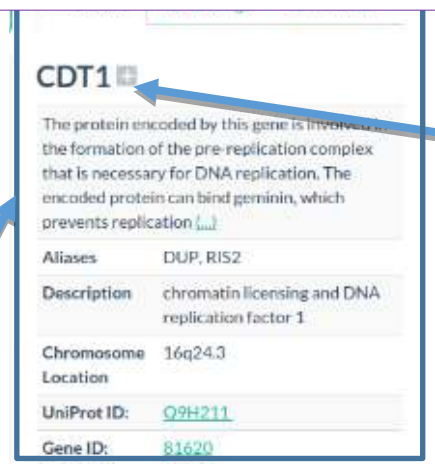
Добавление в сеть данных по изменениям, наблюдающимся при раке

Настройки для фильтрации сети по связям и вершинам



По умолчанию для заданного гена строится сеть, включающая всех соседей этого гена, однако, если у вас задано несколько генов, то можно просто увидеть взаимосвязи между ними.

При нажатии на вершину или связь в генной сети появляется окно с более детальной информацией по данному объекту.



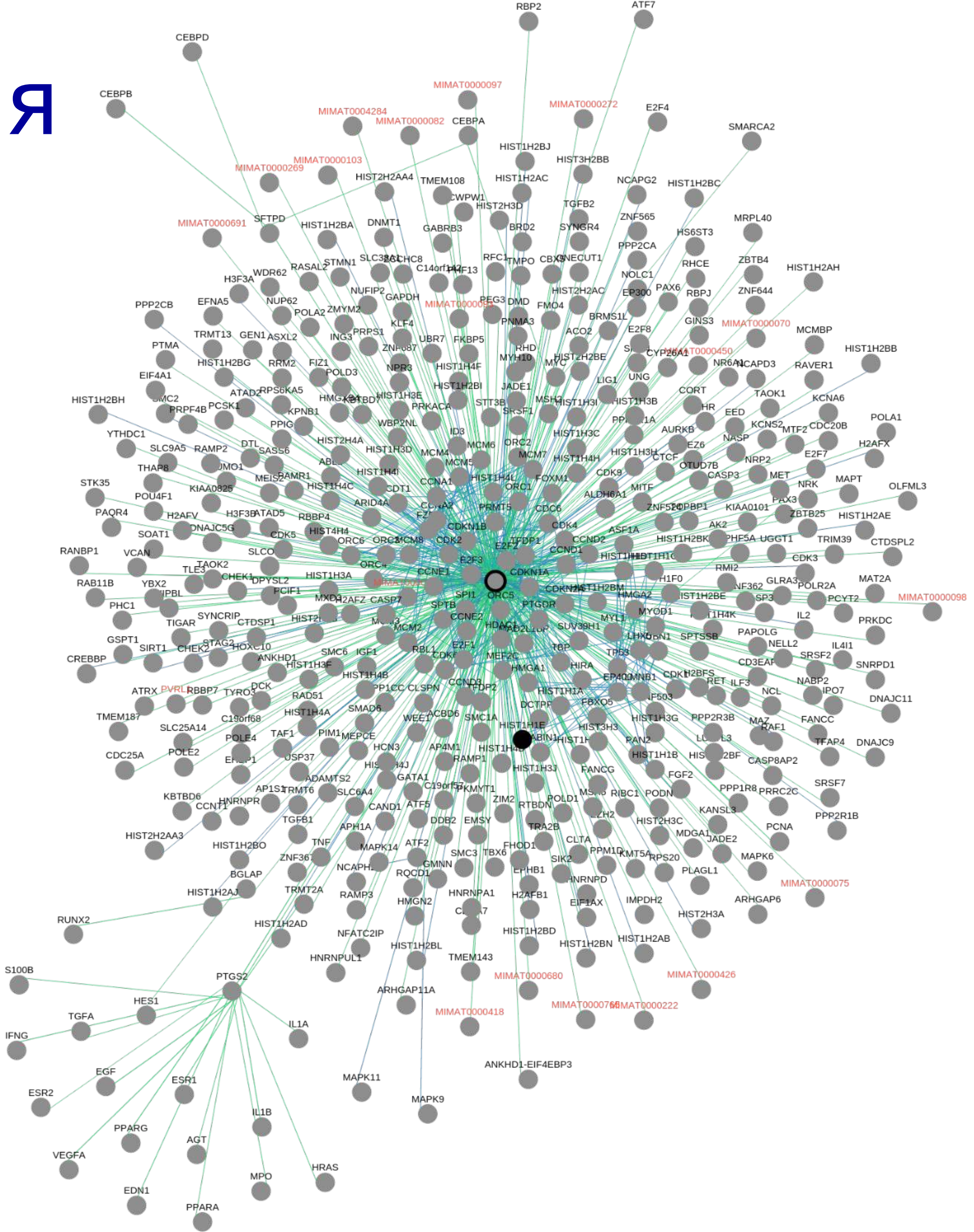
Добавление гена в список интересующих вас генов

Язык BioPAX

Пример генной сети для гена ретинобластомы RB1

Пример загрузки генной сети в формате SIF

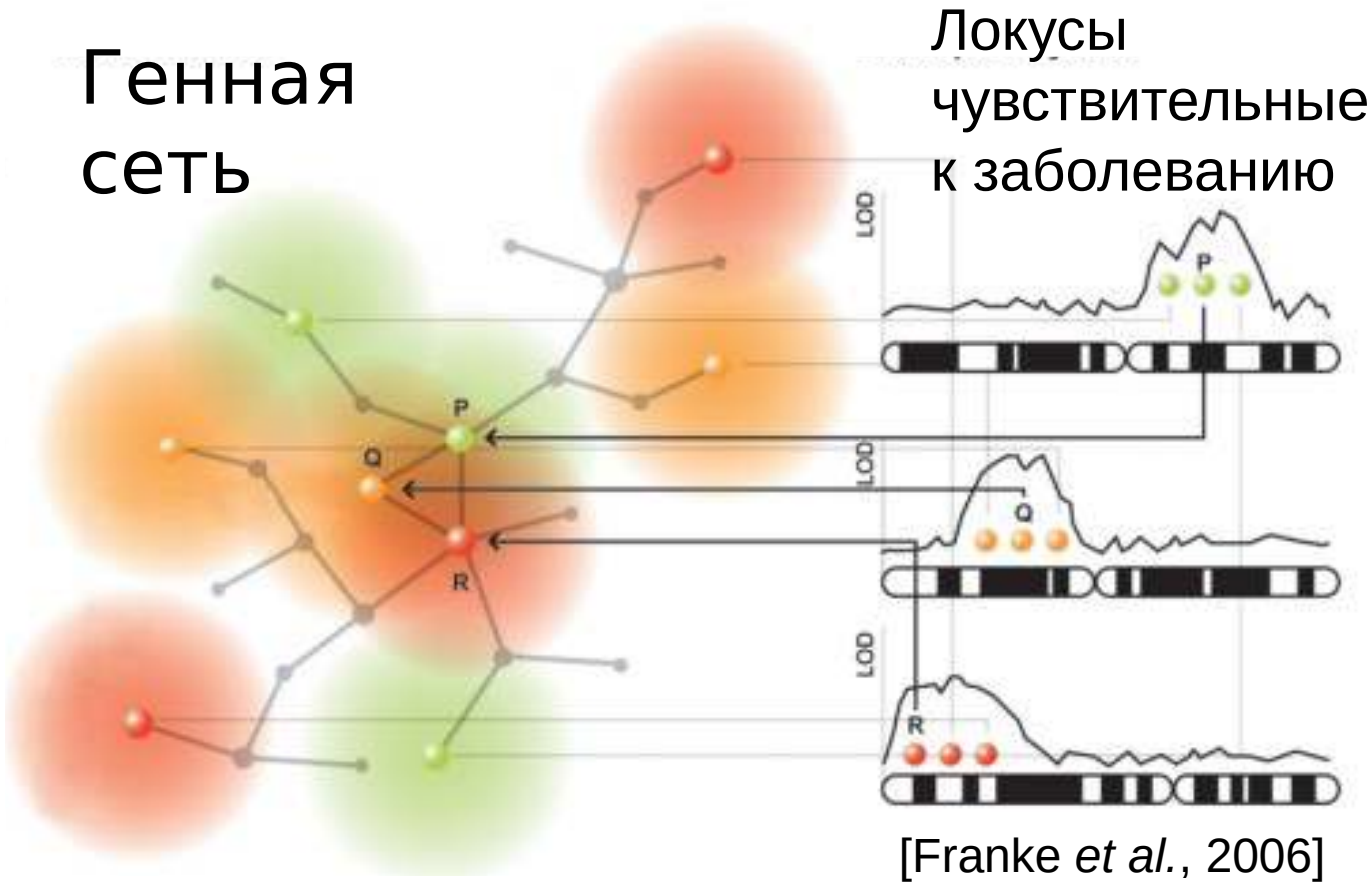
AATF in-complex-with E2F1
AATF in-complex-with E2F2
AATF in-complex-with E2F3
AATF in-complex-with RB1
AATF interacts-with RB1
AATF in-complex-with TFDP1
ABCD3 interacts-with DYRK1A
ABCD3 interacts-with DYRK1B
ABCD3 interacts-with RB1
ABHD10 interacts-with FOXK1



План лекции

1. Введение.
2. Обзор Интернет-ресурсов, содержащих разнообразную информацию о генных сетях.
3. Ресурсы, интегрирующие биологическую информацию из разнородных источников и представляющие ее в виде генных сетей: ANDSystem, STRING, GeneMania, Pathway Commons.
4. Практическое применение инструментов интеграции.

Поиск исходных вершин для реконструкции генных сетей



[Franke *et al.*, 2006]

https://www.sciencedirect.com/science/article/pii/S0002929707639226

1. Избавление от фактора сцепленности генов в хромосомах при использовании информации об ассоциациях аллелей с заболеванием

2. Нахождение генов, мутации в которых являются первопричинами заболеваний

STRING

SCIENTIFIC REPORTS

OPEN

Received: 24 February 2016
Accepted: 10 June 2016
Published: 01 July 2016

Alpha-tubulin enhanced renal tubular cell proliferation and tissue repair but reduced cell death and cell-crystal adhesion

Juthatip Manissorn, Supaporn Khamchun, Arada Vinaiphat & Visith Thongboonkerd

Adhesion of calcium oxalate (CaOx) crystals on renal tubular epithelial cells is a critical event for kidney stone disease that triggers many cascades of cellular response. Our previous expression proteomics study identified several altered proteins in MDCK renal tubular cells induced by CaOx crystals. However, functional significance of those changes had not been investigated. The present study thus aimed to define functional roles of such proteome data. Global protein network analysis using STRING software revealed α-tubulin, which was decreased, as one of central nodes of protein-protein interactions. Overexpression of α-tubulin (pcDNA6.2-TUBA1A) was then performed and its efficacy was confirmed. pcDNA6.2-TUBA1A could maintain levels of α-tubulin and its direct interacting partner, vimentin, after crystal exposure. Also, pcDNA6.2-TUBA1A successfully reduced cell death to almost the basal level and increased cell proliferation after crystal exposure. Additionally, tissue repair capacity was improved in pcDNA6.2-TUBA1A cells. Moreover, cell-crystal adhesion was reduced by pcDNA6.2-TUBA1A. Finally, levels of potential crystal receptors (HSP90, HSP70, and α-enolase) on apical membrane were dramatically reduced to basal levels by pcDNA6.2-TUBA1A. These findings implicate that α-tubulin has protective roles in kidney stone disease by preventing cell death and cell-crystal adhesion, but on the other hand, enhancing cell proliferation and tissue repair function.

Until now, kidney stone disease is still a public health problem in almost all areas around the world. The disease causes substantial suffering and ultimately end-stage renal disease (ESRD). Unfortunately, the disease mechanisms remain poorly understood. Calcium oxalate (CaOx) is the major chemical component found in clinical stones[1]. This type of the stones can be originated from supersaturation of calcium and oxalate ions, leading to crystallization inside renal tubular fluid or urine[2]. CaOx crystals can then nucleate to form "stone nidus" and adhere directly onto apical surface of renal tubular epithelial cells[3–5]. Adhesion of crystals onto the cells is a critical event, which triggers many cascades of cellular response, e.g. cytotoxicity, injury, proliferation and apoptosis, that ultimately lead to kidney stone formation[6,7]. CaOx crystals also evoke inflammatory processes that can lead to fibrosis, loss of nephron and eventually ESRD[8,9].

Even with the aforementioned knowledge, molecular mechanisms of the downstream cellular response remain largely unknown. From our previous expression proteomics study[7], we have identified a number of proteins with altered levels in MDCK renal tubular cells in response to CaOx crystals. Those altered proteins were involved in various biological processes, i.e. ubiquitination pathway, signal transduction, cellular structure, purine biosynthesis, metabolic enzyme, retinol biosynthesis, cellular transportation, protein degradation, RNA metabolism, RNA binding protein, cell surface antigen, nucleic acid metabolism, antioxidant enzyme, chaperone, carrier protein, and protein biosynthesis. However, functional significance of those altered proteins had not been investigated. In the present study, we thus performed global protein network analysis of those altered proteins. Subsequently, overexpression of a protein, which was one of the central nodes of such protein-protein interactions network, was performed. Moreover, functional investigations were performed to address functional significance of the central-node protein and its associated partners in kidney stone disease.

Medical Proteomics Unit, Office for Research and Development, Faculty of Medicine Siriraj Hospital, and Center for Research in Complex Systems Science, Mahidol University, Bangkok 10700, Thailand. Correspondence and requests for materials should be addressed to V.T. (email: thongboonkerd@dr.com)

GeneMania

RESEARCH Open Access

Gene expression network analyses in response to air pollution exposures in the trucking industry

Jen-hwa Chu[1*], Jaime E. Hart[2,3], Divya Chhabra[2], Eric Garshick[2,4], Benjamin A. Raby[2,5] and Francine Laden[2,3,6]

Abstract

Background: Exposure to air pollution, including traffic-related pollutants, has been associated with a variety of adverse health outcomes, including increased cardiopulmonary morbidity and mortality, and increased lung cancer risk.

Methods: To better understand the cellular responses induced by air pollution exposures, we performed genome-wide gene expression microarray analysis using whole blood RNA sampled at three time-points across the work weeks of 63 non-smoking employees at 10 trucking terminals in the northeastern US. We defined genes and gene networks that were differentially activated in response to $PM_{2.5}$ (particulate matter ≤ 2.5 microns in diameter) and elemental carbon (EC) and organic carbon (OC).

Results: Multiple transcripts were strongly associated ($p_{adj} < 0.001$) with pollutant levels (48, 260, and 49 transcripts for EC, OC, and $PM_{2.5}$, respectively), including 63 that were statistically significantly correlated with at least two out of the three exposures. These genes included many that have been implicated in ischemic heart disease, chronic obstructive pulmonary disease (COPD), lung cancer, and other pollution-related illnesses. Through the combination of Gene Set Enrichment Analysis and network analysis (using GeneMANIA), we identified a core set of 25 interrelated genes that were common to all three exposure measures and were differentially expressed in two previous studies assessing gene expression attributable to air pollution. Many of these are members of fundamental cancer-related pathways, including those related to DNA and metal binding, and regulation of apoptosis and also but include genes implicated in chronic heart and lung diseases.

Conclusions: These data provide a molecular link between the associations of air pollution exposures with health effects.

Keywords: Air pollution, Trucking industry, Gene expression, Network analysis

Background

Air pollution exposures, have been associated with a number of adverse health effects, including greater morbidity and mortality risks for cardiopulmonary diseases, and increased risk of lung cancer [1–6]. However, the underlying biological mechanisms have not been fully elucidated. Human studies of global changes in gene expression following controlled exposures [7], or using in vitro models [8, 9] have provided some insights in this regard, yet few studies have rigorously assessed the impact of air pollution on gene expression in real-life settings. For example, though observational studies have been conducted in individuals from geographic regions with differing levels of air pollution have suggested associations, [10] studies with more refined exposure measures have not been performed.

In this study, we characterized the cellular response induced by traffic-related air pollution exposures in a population of non-smoking US trucking industry employees. We performed genome-wide gene expression microarray analysis using whole blood RNA sampled at three time-points during the work week. We integrate

* Correspondence: jen-hwa.chu@yale.edu
[1]Section of Pulmonary, Critical Care and Sleep Medicine, Department of Internal Medicine, Yale University School of Medicine, New Haven, CT, USA
Full list of author information is available at the end of the article

Pathway Commons

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

A Multi-Method Approach for Proteomic Network Inference in 11 Human Cancers

Yasin Şenbabaoğlu[1]*, Selçuk Onur Sümer[1], Francisco Sánchez-Vega[1], Debra Bemis[1], Giovanni Ciriello[1¤], Nikolaus Schultz[2], Chris Sander[1]*

1 Computational Biology Program, Memorial Sloan Kettering Cancer Center, New York, New York, United States of America, 2 Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York, United States of America

¤ Current address: Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland
* senbabay@mskcc.org (YŞ); chris@sanderlab.org (CS)

OPEN ACCESS

Citation: Şenbabaoğlu Y, Sümer SO, Sánchez-Vega F, Bemis D, Ciriello G, Schultz N, et al. (2016) A Multi-Method Approach for Proteomic Network Inference in 11 Human Cancers. PLoS Comput Biol 12(2): e1004765. doi:10.1371/journal.pcbi.1004765

Editor: Christian von Mering, University of Zurich and Swiss Institute of Bioinformatics, SWITZERLAND

Received: June 7, 2015

Accepted: January 20, 2016

Published: February 29, 2016

Data Availability Statement: The pan-cancer RPPA dataset is available from The Cancer Proteome Atlas at http://app1.bioinformatics.mdanderson.org/tcpa/_design/basic/download.html.

Funding: CS and YŞ were supported by the National Institute of General Medical Sciences (P41GM103504) and the National Cancer Institute (U24CA143840). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Protein expression and post-translational modification levels are tightly regulated in neoplastic cells to maintain cellular processes known as 'cancer hallmarks'. The first Pan-Cancer initiative of The Cancer Genome Atlas (TCGA) Research Network has aggregated protein expression profiles for 3,467 patient samples from 11 tumor types using the antibody based reverse phase protein array (RPPA) technology. The resultant proteomic data can be utilized to computationally infer protein-protein interaction (PPI) networks and to study the commonalities and differences across tumor types. In this study, we compare the performance of 13 established network inference methods in their capacity to retrieve the curated Pathway Commons interactions from RPPA data. We observe that no single method has the best performance in all tumor types, but a group of six methods, including diverse techniques such as correlation, mutual information, and regression, consistently rank highly among the tested methods. We utilize the high performing methods to obtain a consensus network; and identify four robust and densely connected modules that reveal biological processes as well as suggest antibody–related technical biases. Mapping the consensus network interactions to Reactome gene lists confirms the pan-cancer importance of signal transduction pathways, innate and adaptive immune signaling, cell cycle, metabolism, and DNA repair; and also suggests several biological processes that may be specific to a subset of tumor types. Our results illustrate the utility of the RPPA platform as a tool to study proteomic networks in cancer.

Author Summary

Pan-cancer proteomic datasets from The Cancer Genome Atlas provide a unique opportunity to study the functions of proteins in human cancers. Such datasets, where proteins are measured in different conditions and where correlations are informative, can enable the discovery of potentially causal protein-protein interactions, which may in turn shed light on the function of proteins. However, it has been shown that the dominant correlations in

Анализ протеомных экспериментов, направленных на выявление белков в протеоме мочи чувствительных к различному содержанию NaCl в пище для человека

Период эксперимента	1- 35 сутки, или 1-5 неделя	36-70 сутки, или 6-9 неделя	71-75 сутки, или 10 неделя	76-98 сутки, или 11-14 неделя
Уровень потребления соли (г/сутки)	12	9	12	6

Примечание: норма потребления соли для человека находится в пределах 4-12 г/сутки.

Ларина И. М. и др. РЕКОНСТРУКЦИЯ АССОЦИАТИВНЫХ БЕЛКОВЫХ СЕТЕЙ, СВЯЗАННЫХ С ПРОЦЕССАМИ РЕГУЛЯЦИИ ОБМЕНА И ДЕПОНИРОВАНИЯ НАТРИЯ В ОРГАНИЗМЕ ЗДОРОВОГО ЧЕЛОВЕКА, НА ОСНОВЕ ИЗУЧЕНИЯ ПРОТЕОМА МОЧИ //Физиология человека. – 2012. – Т. 38. – №. 3. – С. 107-107.

Запись результатов протеомных экспериментов для проведения кластерного анализа

Недели эксперимента и уровень потребления соли													
1	2	3	4	5	6	7	8	9	10	11	12	13	14
12 г/сут					9 г/сут				12 г/сут	6 г/сут			

Динамика появления белков в моче

Белок А

1	0	0	0	0	0	1	1	1	1	1	0	1	1

Белок В

1	0	1	0	0	0	1	0	1	1	1	0	1	1

Запись результатов протеомных экспериментов для проведения кластерного анализа

Недели эксперимента и уровень потребления соли													
1	2	3	4	5	6	7	8	9	10	11	12	13	14
12 г/сут					9 г/сут				12 г/сут	6 г/сут			

Динамика появления белков в моче

Белок А

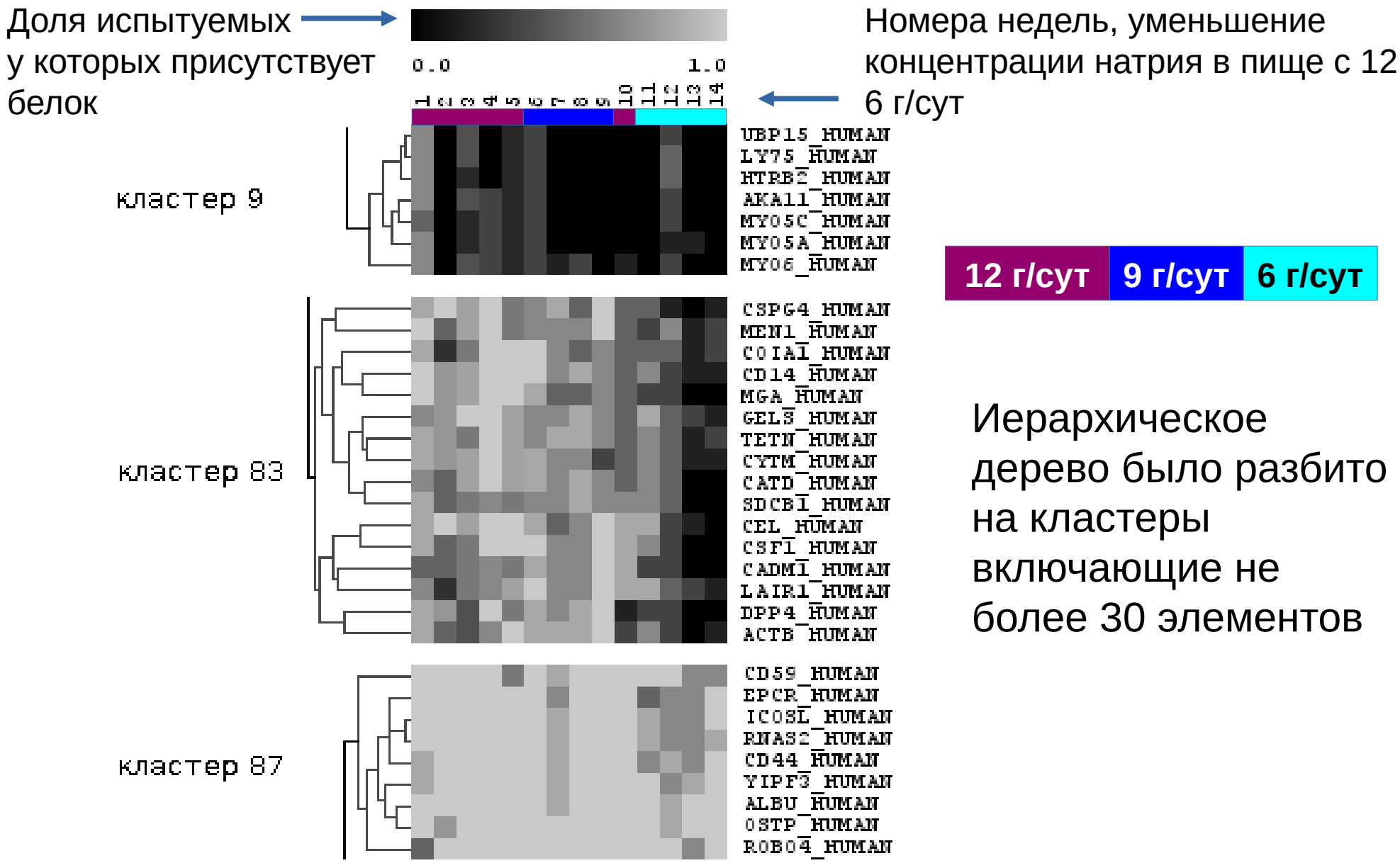
0,83	0,16	0	0	0,33	0,16	1	0,83	1	1	0,83	0	0,83	1

Белок В

1	0	0,83	0	0	0	1	0,16	1	1	0,66	0,16	1	1

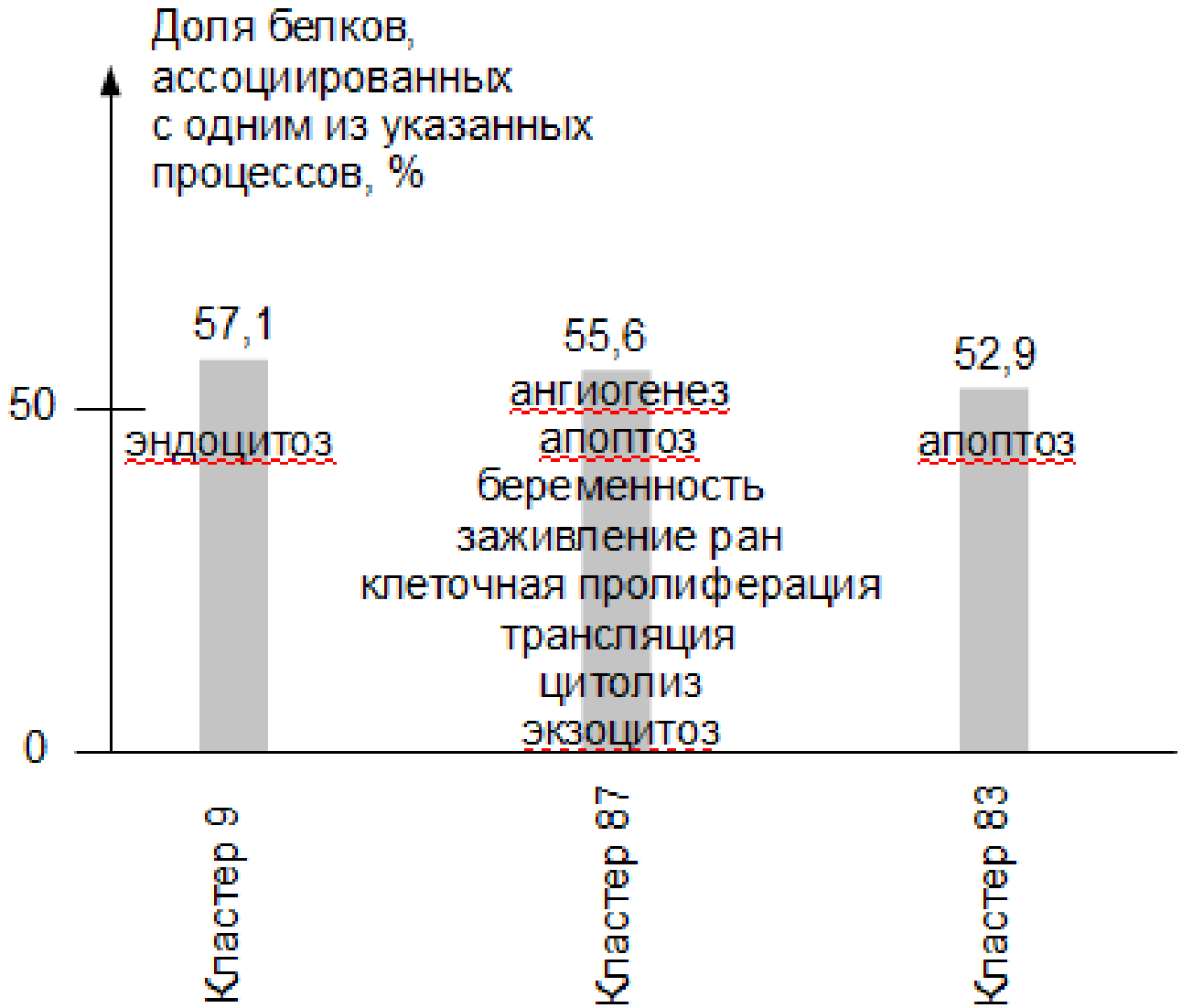
Манхэттенское расстояние между векторами: $D = \sum \left| A_i - B_i \right|$

Кластеризация проводилась в программе MeV 4.6.2 по методу Average linkage clustering [Sokal, Michener, 1958]

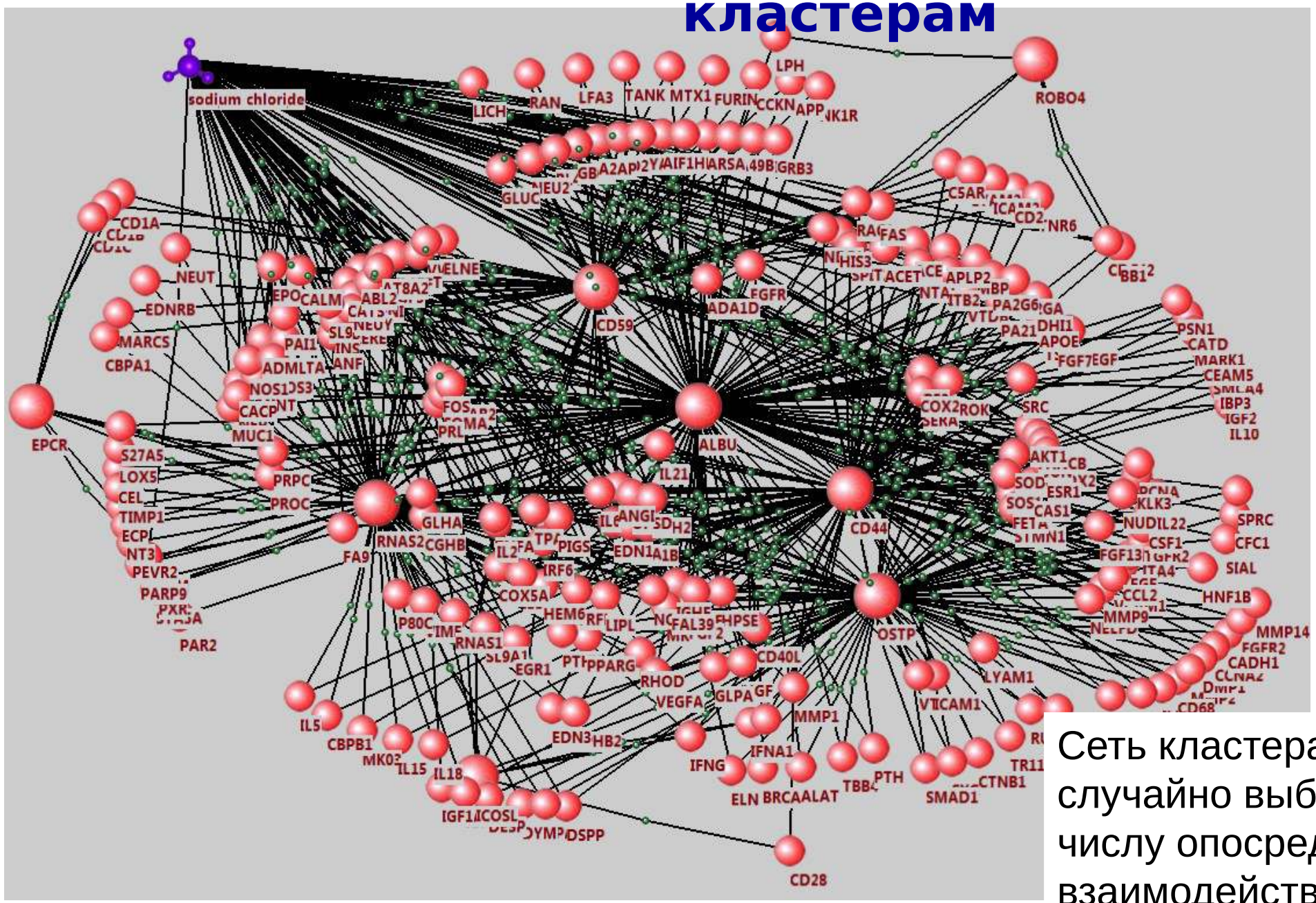


Иерархическое дерево было разбито на кластеры включающие не более 30 элементов

Поиск кластеров ассоциированных с



Построение сетей по найденным кластерам

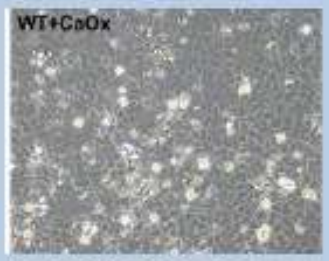


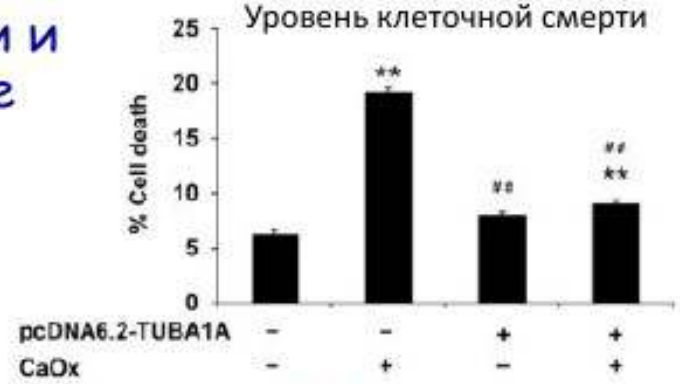
Сеть кластера 87 отличается от случайно выбранных сетей по числу опосредованных взаимодействий (p=0.0175)

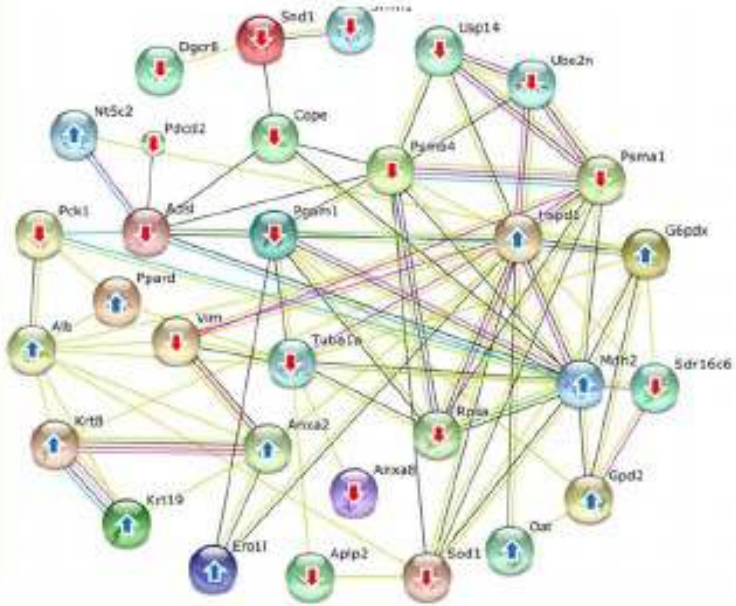
Способность альфа-тубулина усиливать процессы пролиферации и восстановления тканей и снижать клеточную смерть и связывание кристаллов оксалата кальция почечными трубчатыми клетками.

Адгезия кристаллов оксалата кальция (CaOx) на почечных канальцевых эпителиальных клетках наблюдается при мочекаменной болезни. Было показано, что экспрессия ряда белков достоверно меняется в клетках, индуцированных кристаллами CaOx. Анализ глобальной сети белок-белок взаимодействий с использованием системы STRING показал, что α-тубулин, экспрессия которого была снижена в эксперименте, является одним из центральных узлов сети белок-белковых взаимодействий. Для подтверждения ключевой роли α-тубулина была исследована клеточная линия pcDNA6.2-TUBA1A, в которой α-тубулин сверхэкспрессирован. Было показано, что для клеток pcDNA6.2-TUBA1A наблюдается снижение гибели и увеличение клеточной пролиферации после воздействия кристаллов CaOx. Кроме того, в клетках pcDNA6.2-TUBA1A была улучшена репаративная способность тканей и снижена клеточно-кристаллическая адгезия. Эти результаты говорят в пользу того, что α-тубулин играет ключевую роль в патогенезе почечнокаменной болезни.

Фазово-контрастная микроскопия почечных трубчатых клеток





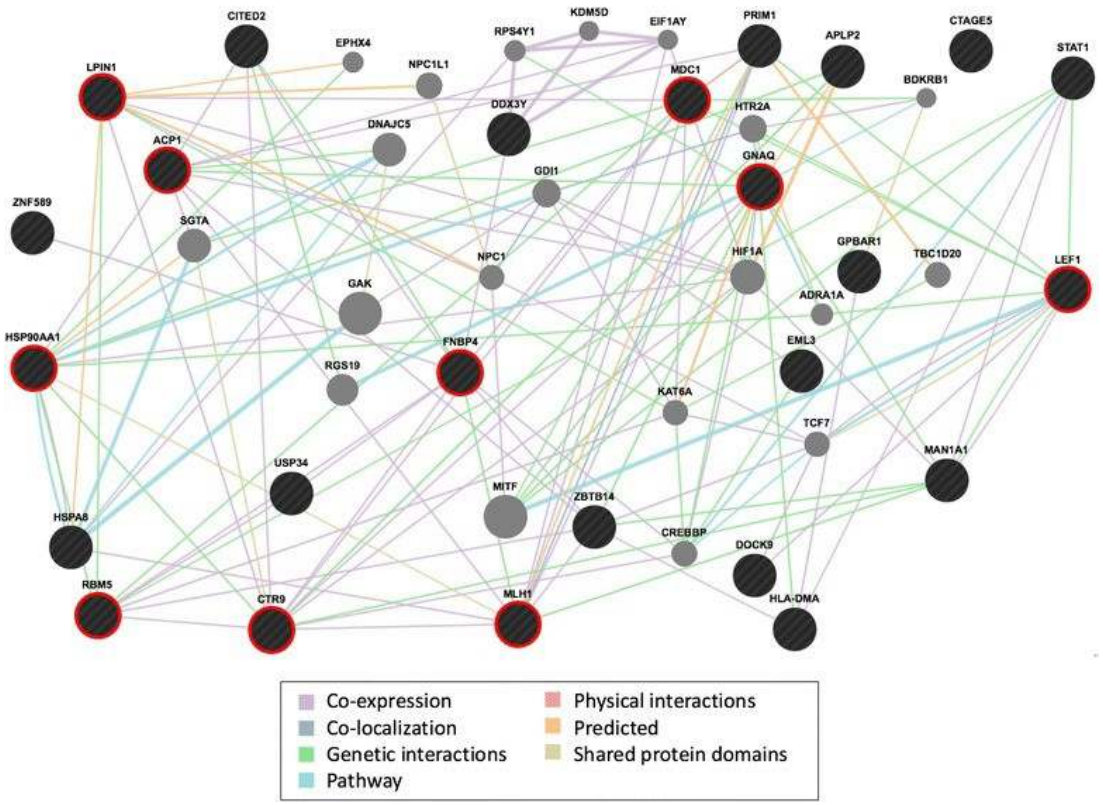


Сеть взаимодействий между белками с измененной экспрессией в почечных трубчатых клетках MDCK, индуцированных кристаллами CaOx, реконструированная с помощью системы STRING.

Manissorn J. et al. Alpha-tubulin enhanced renal tubular cell proliferation and tissue repair but reduced cell death and cell-crystal adhesion //Scientific Reports. – 2016. – Т. 6.

Анализ сети генов, экспрессия которых изменяется в ответ на воздействие загрязнения воздуха в автотранспортной отрасли.

Воздействие загрязненного воздуха связано с неблагоприятными последствиями для здоровья, включая увеличение сердечно-легочной заболеваемости и смертности, а также увеличение риска рака легких. Был проведен полно-геномный анализ уровней экспрессии РНК крови 63 некурящих сотрудников парковочных терминалов в США. Экспрессия ряда генов изменилась при воздействии трех загрязнителей (твердых микрочастиц, элементарного углерода и органического углерода). В том числе экспрессия 63 гена, статистически значимо изменилась в случае двух из трех загрязнителей. Многие из этих генов были вовлечены в ишемическую болезнь сердца, хроническую обструктивную болезнь легких (ХОБЛ), рак легких и другие заболевания, связанные с загрязнением. Анализ генной сети (реконструированной с использованием GeneMANIA) позволил выявить 25 взаимосвязанных генов, которые были общими для всех трех загрязнений. Многие из них ассоциированы с раком, а также вовлечены в связывание металлов и регуляцию апоптоза. Эти результаты описывают молекулярную взаимосвязь между воздействием загрязненного воздуха и неблагоприятными последствиями для здоровья.

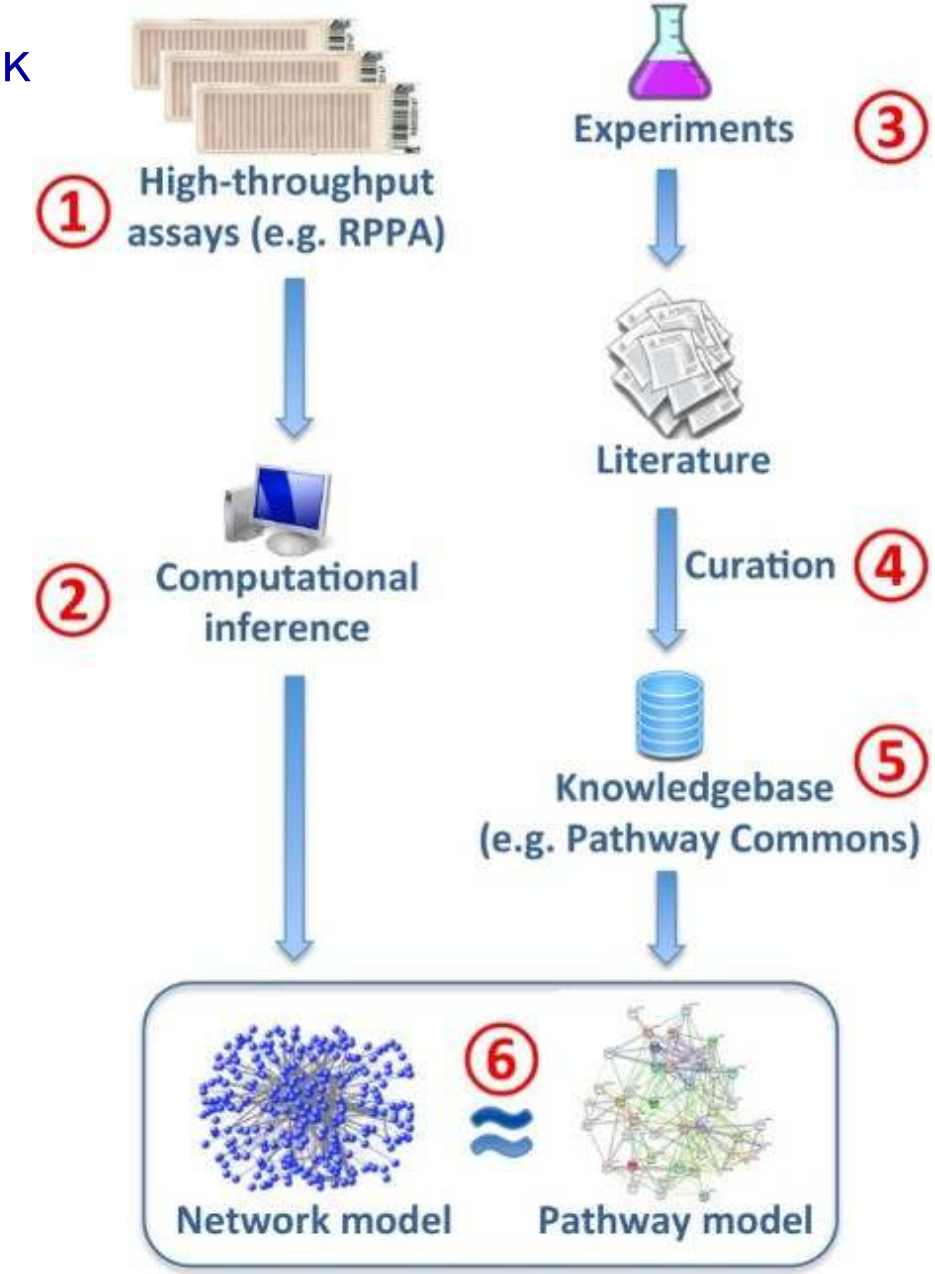


Сеть взаимосвязей между генами, экспрессия которых изменилась в ответ на воздействие загрязнения воздуха, реконструированная с использованием GeneMANIA. Анализ позволил выявить 25 взаимосвязанных генов, которые были общими для всех трех загрязнений.

Chu J. et al. Gene expression network analyses in response to air pollution exposures in the trucking industry //Environmental Health. – 2016. – Т. 15. – №. 1. – С. 101.

Мульти-метод для реконструкции сетей белок-белок взаимодействий, найденных в 11 видах раковых заболеваний человека.

Раковые клетки имеют особые характеристики уровней экспрессии белков и пост-трансляционных модификаций. Ресурс The Cancer Genome Atlas (TCGA) объединяет профили экспрессии белков 3467 образцов пациентов, страдающих от 11 типов опухолей. Эти протеомные данные могут быть использованы для компьютерной реконструкции сетей белок-белок взаимодействий (PPI) и последующего изучения общих характеристик различных типов опухолей. В этом исследовании было проведено сравнение эффективности 13 автоматических методов для реконструкции белок-белок взаимодействий с сетями белок-белок взаимодействий системы Pathway Commons, реконструированными вручную экспертами. Было показано, что ни один из методов не обладает наилучшими показателями во всех типах опухолей, но группа из шести методов, включая методы, основанные на корреляции, взаимной информации и регрессии, позволяют реконструировать сети, которые хорошо согласуются с сетями, построенными вручную экспертами.



Şenbabaoğlu Y. et al. A multi-method approach for proteomic network inference in 11 human cancers //PLoS Comput Biol. – 2016. – Т. 12. – №. 2. – С. e1004765.

План лекции

1. Введение.
2. Обзор Интернет-ресурсов, интегрирующих информацию по биомедицинской тематике.
3. Ресурсы, интегрирующие биологическую информацию из разнородных источников и представляющие ее в виде генных сетей: ANDSystem, STRING, GeneMania, Pathway Commons.
4. Практическое применение инструментов интеграции.