

# Лекция №5

## **Интеграция разрозненных источников информации по биомедицинской тематике**

м.н.с. лаб. гликогеномики  
Тийс Евгений Сергеевич,  
[tiys@bionet.nsc.ru](mailto: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.

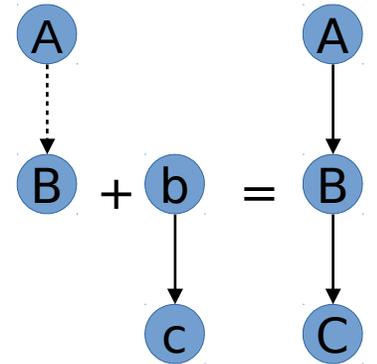
# План лекции

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 механизмов биологических процессов

Преимущества интеграции биологических данных:

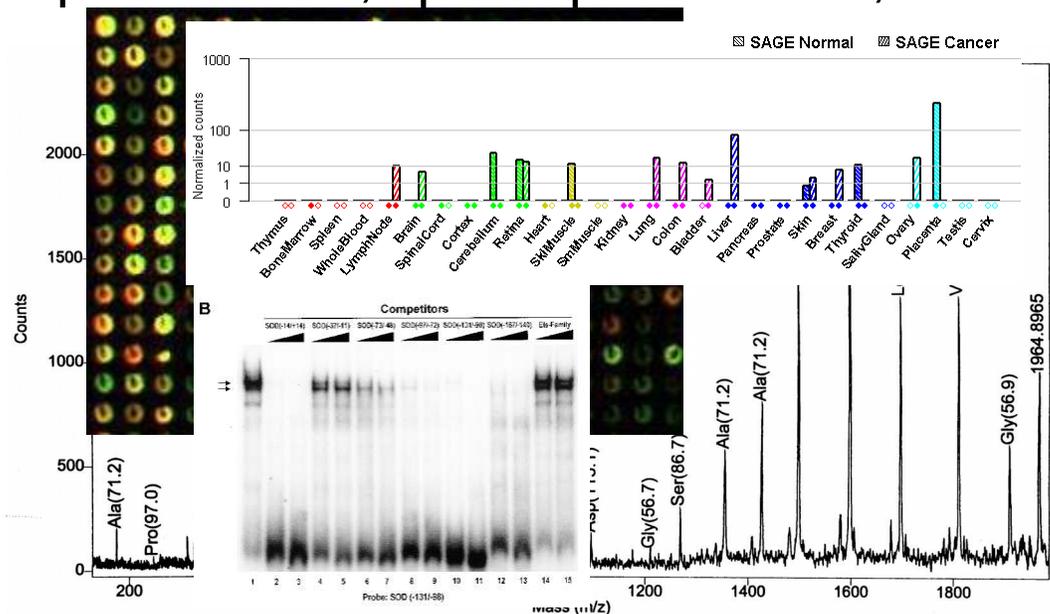
- Единый формат – возможность автоматизации обработки и сбор статистики.
- Единый способ графического и текстового представления – упрощение осмысления, отсутствие штрафов на переключения внимания.
- Возможность выполнения поискового запроса ко всем данным
- Установление эквивалентности объектов из различных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



# Процесс интеграции

Экспериментальные данные:

Протеомные, транскриптомные, метаболомные



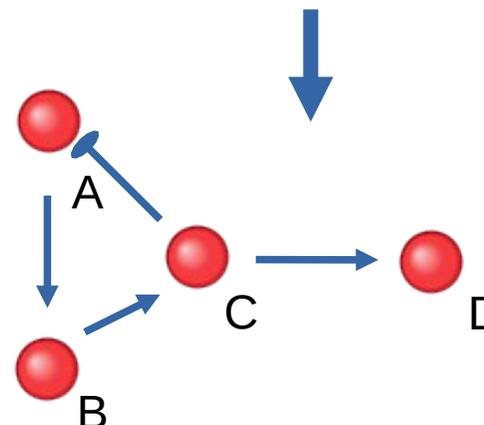
базы данных



PubMed

Скорость появления публикаций **1,4** в  
минуту

[http://www.nlm.nih.gov/bsd/medline\\_cit\\_counts\\_yr\\_pub.html](http://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html)

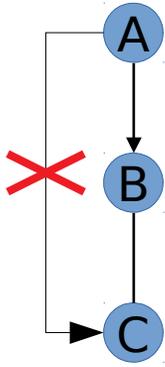


**Генная сеть** – группа координированно функционирующих генов, взаимодействующих друг с другом как через свои первичные продукты (РНК и белки), так и через разнообразные метаболиты и другие вторичные продукты функционирования генных сетей, которая контролирует какой-либо фенотипический признак организма.

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).



“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. - T. 5. - №. 10. - C. e12912.

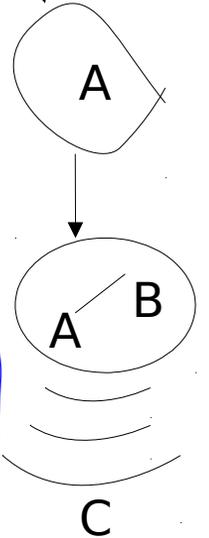


“A gene network is a directed labelled graph, where each node represents a gene and each arc represents a relation between the genes.”

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

“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.



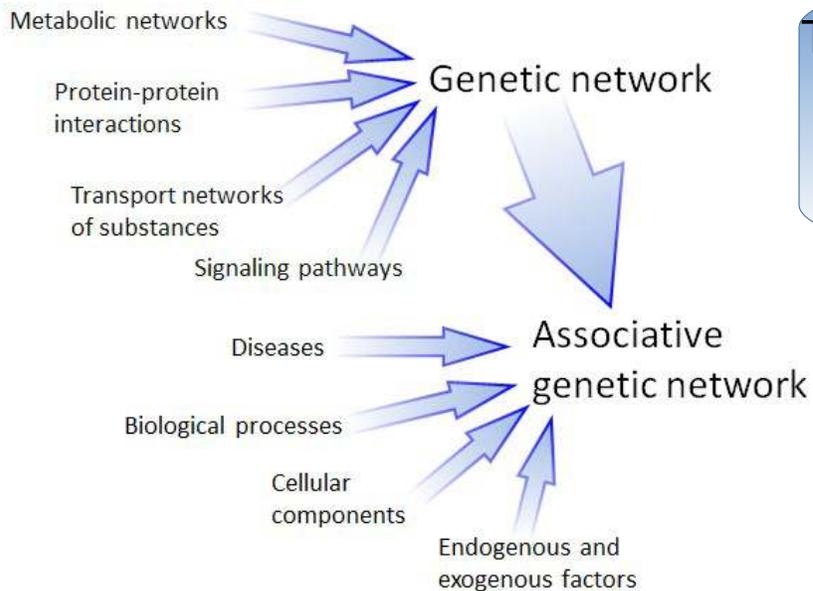
“A gene network is a collection of effective interactions, describing the multiple ways through which one gene affects all the others to which it is connected.”

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. - T. 2. - №. 3. - C. 257.

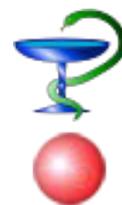
Ассоциативные генные сети являются молекулярно-генетическими сетями, ассоциированными с какими либо биологическими процессами, фенотипическими признаками или заболеваниями.

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).

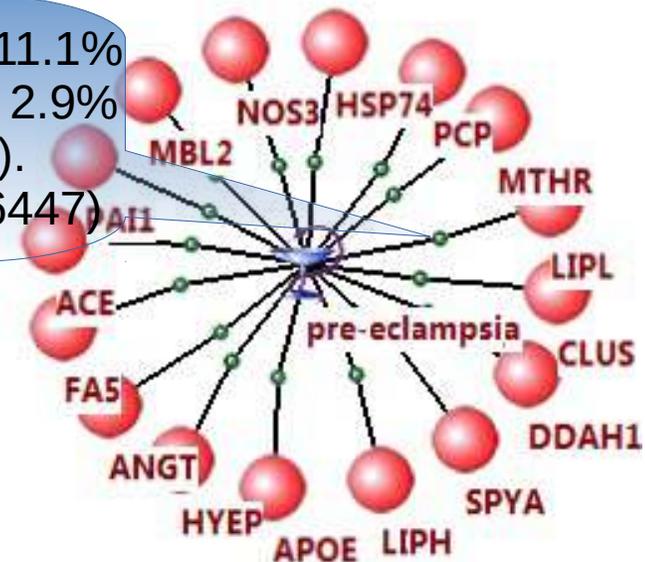


The N291S variant was identified in 11.1% of pre-eclampsics as compared with 2.9% of pregnancy controls ( $p = 0.008$ ). (Hubel CA et al. 1999, PMID: 10636447)



заболевание

белок



# 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)	<b>Protein Database</b>
BioSystems	Genome	<b>PubChemBioAssay</b>
Bookshelf	Genome Reference Consortium (GRC)	<b>PubChemCompound</b>
ClinicalTrials.gov	HIV-1	<b>PubChemSubstance</b>
<b>ClinVar</b>	HomoloGene	<b>PubMed</b>
CloneDB (formerly Clone Registry)	Influenza Virus	<b>PubMed Central (PMC)</b>
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	<b>RefSeqGene</b>
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
<b>Database of Short Genetic Variations (dbSNP)</b>	NCBI Help Manual	Third Party Annotation (TPA) Database
<b>GenBank</b>	NCBI Pathogen Detection Project	Trace Archive
<b>Gene</b>	NCBI Website Search	UniGene
<b>Gene Expression Omnibus (GEO) Database</b>	<b>Nucleotide Database</b>	UniGene Library Browser
<b>Gene Expression Omnibus (GEO) Datasets</b>	<b>Online Mendelian Inheritance in Man (OMIM)</b>	Viral Genomes
<b>Gene Expression Omnibus (GEO) Profiles</b>	PopSet	Virus Variation

# Результат поискового запроса к ресурсу NCBI по ключевому слову YY1

## Search NCBI databases

[Help](#)

### Results found in 33 databases for "yy1"

#### Literature

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

#### Health

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

#### Genomes

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

#### Genes

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

#### Proteins

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

#### Chemicals

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

# План лекции

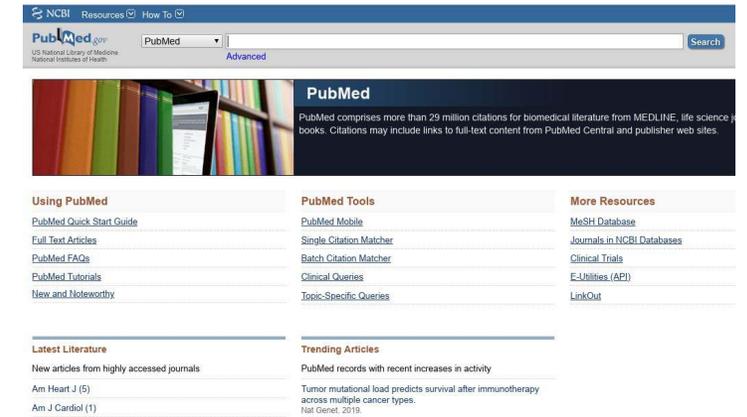
1. Введение.
2. Обзор Интернет-ресурсов, интегрирующих информацию по биомедицинской тематике.
3. Ресурсы, интегрирующие биологическую информацию из разнородных источников и представляющие ее в виде генных сетей: ANDSystem, STRING, GeneMania, Pathway Commons.
4. Практическое применение инструментов интеграции.

# 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.

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>



## Интерфейс PubMed

## 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**



How To Sign in to NCBI

PubMed   Create RSS Create alert Advanced Help

Format: Summary  Sort by: Best Match  Per page: 20  Send to  Filters: [Manage Filters](#)

Search results

Items: 1 to 20 of 3950 << First < Prev Page 1 of 198 Next > Last >>

[Functional characterisation and cell specificity of BvSUT1, the transporter that loads sucrose into the phloem of sugar beet \(Beta vulgaris L.\) source leaves.](#)

1. Nieberl P, Ehrl C, Pommerrenig B, Graus D, Marten I, Jung B, Ludewig F, Koch W, Harms K, Flügge UI, Neuhaus HE, Hedrich R, Sauer N. *Plant Biol (Stuttg)*. 2017 May;19(3):315-326. doi: 10.1111/plb.12546. Epub 2017 Feb 1. PMID: 28075052 [Similar articles](#)

Results by year

Download CSV

How To Sign in to NCBI

PubMed   Create RSS Create alert Advanced Help

Format: Summary  Sort by: Best Match  Per page: 20  Send to  Filters: [Manage Filters](#)

Search results

Items: 1 to 20 of 6911 << First < Prev Page 1 of 346 Next > Last >>

[Panax ginseng and Panax quinquefolius: From pharmacology to toxicology.](#)

1. Mancuso C, Santangelo R. *Food Chem Toxicol*. 2017 Sep;107(Pt A):362-372. doi: 10.1016/j.fct.2017.07.019. Epub 2017 Jul 8. Review. PMID: 28698154 [Similar articles](#)

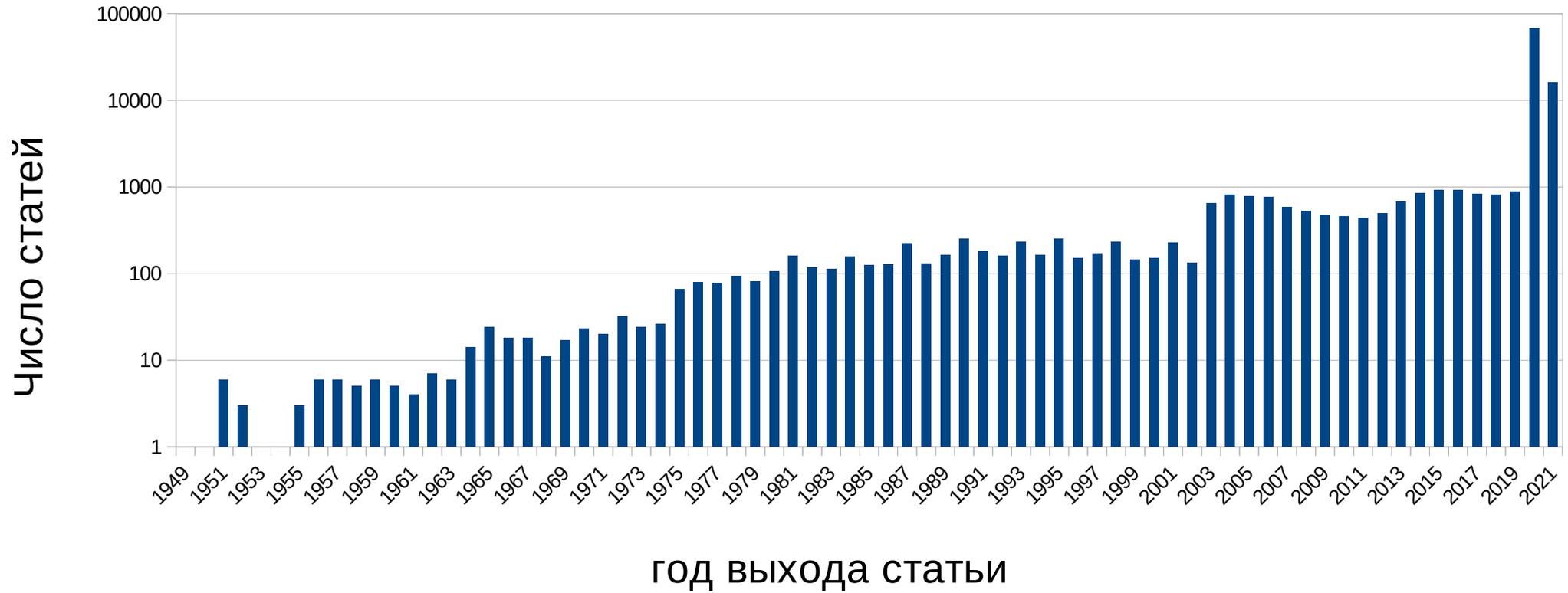
[Ginseng in Dermatology: A Review.](#)

2. Sabouri-Rad S, Sabouri-Rad S, Sahebkar A, Tayaranl-Najaran Z. *Curr Pharm Des*. 2017;23(11):1649-1666. doi: 10.2174/1381612822666161021152322. Review.

Results by year

Download CSV

# 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

The screenshot displays the HGNC website interface. At the top, the HGNC logo and name are visible. A search bar is present with the text "Search everything" and "Search symbols, keywords or IDs". Below the search bar, there is a navigation menu with links for Home, Downloads, Gene Families, Tools, Useful Links, About, Newsletters, Contact Us, Help, VGNC, and Request Symbol. The main content area is titled "Symbol Report: YY1". It contains a table with the following information:

APPROVED SYMBOL	YY1
APPROVED NAME	YY1 transcription factor
HGNC ID	HGNC:12856
PREVIOUS SYMBOLS & NAMES	-
LOCUS TYPE	gene with protein product
CHROMOSOMAL LOCATION	14q32.2
GENE FAMILY	INO80 complex Zinc fingers C2H2-type
HCOP	Orthology Predictions for YY1
SYNONYMS	DELTA, "INO80 complex subunit S", INO80S, NF-E1, UCRBP, "Yin and Yang 1 protein", YIN-YANG-1

Below the main report, there is an "External links" section. It includes a "HOMOLOGS" table with columns for Symbol and Database, listing homologs in Mus musculus and Rattus norvegicus. Other sections include "GENE RESOURCES", "NUCLEOTIDE SEQUENCES", "PROTEIN RESOURCES", and "CLINICAL RESOURCES", each with links to various databases and resources.

# 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.

Поиск

Окно для ввода запросов

Сохранение

Организмы

The screenshot shows the NCBI Gene database search results for the query 'yy1'. The search bar at the top contains 'yy1' and the search button is labeled 'Search'. The results are displayed in a table with columns for Name/Gene ID, Description, Location, and Aliases. The first four results are highlighted with red boxes:

Name/Gene ID	Description	Location	Aliases
YY1 ID: 7528	YY1 transcription factor [ <i>Homo sapiens</i> (human)]	Chromosome 14, NC_000014.9 (100238765..100279034)	DELTA, INO80S, NF-E1, UCRBP, YIN-YANG-1
Yy1 ID: 22632	YY1 transcription factor [ <i>Mus musculus</i> (house mouse)]	Chromosome 12, NC_000078.6 (108793269..108816632)	AW488674, NF-E1
Yy1 ID: 24919	YY1 transcription factor [ <i>Rattus norvegicus</i> (Norway rat)]	Chromosome 6, NC_005105.4 (132702443..132726848)	NF-E1, NMP-1, NMP1, UCRBP
YY1 ID: 826093	zinc finger (C2H2 type) family protein [ <i>Arabidopsis thaliana</i> (thale cress)]	Chromosome 4, NC_003075.7 (3764241..3766671, complement)	AT4G06634, AtYY1, Yin Yang 1

The search results also include a summary of 1415 items and a list of filters for taxonomic classification. The 'Search details' section shows the query: 'yy1[All Fields] AND alive[prop]'. The 'Recent activity' section shows the query: 'yy1 AND (alive[prop]) (1415)'.

# NCBI Gene - Информационная карточка гена

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 <a href="#">HGNC</a>
Official Full Name	histone deacetylase 2 provided by <a href="#">HGNC</a>
Primary source	<a href="#">HGNC:HGNC:4853</a>
See related	<a href="#">Ensembl:ENSG00000196591</a> <a href="#">MIM:605164</a>
Gene type	protein coding
RefSeq status	REVIEWED
Organism	<a href="#">Homo sapiens</a>
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 <a href="#">See more</a>
Orthologs	<a href="#">mouse</a> <a href="#">all</a>

### Table of contents

Summary

Genomic context

Genomic regions, transcripts, and products

Expression

Bibliography

Phenotypes

Variation

HIV-1 interactions

Pathways from PubChem

Interactions

General gene information

Markers, Related pseudogene(s), Homology, Gene Ontology

General protein information

NCBI Reference Sequences (RefSeq)

Related sequences

Additional links

# NCBI Nucleotide

- <https://www.ncbi.nlm.nih.gov/nucleotide>
- Коллекция нуклеотидных последовательностей из нескольких источников GenBank, RefSeq, TrEMBL 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

The screenshot shows the NCBI Nucleotide search interface. The search query is 'yy1'. The results page displays a list of 3830 nucleotide sequences. The first few results are:

- Synthetic construct Homo sapiens clone ccsbBroadEn\_07141 YY1 gene, encodes complete protein**  
1,374 bp linear other-genetic  
Accession: KJ897747.1 GI: 649121807
- Homo sapiens YY1 transcription factor (YY1), mRNA**  
3,159 bp linear mRNA  
Accession: NM\_003403.4 GI: 459683878
- Rattus norvegicus YY1 transcription factor (Yy1), mRNA**  
1,236 bp linear mRNA  
Accession: NM\_173290.1 GI: 27545349
- Mus musculus YY1 transcription factor (Yy1), mRNA**  
2,324 bp linear mRNA  
Accession: NM\_009537.3 GI: 118130338
- Danio rerio YY1 transcription factor a (yy1a), mRNA**  
2,903 bp linear mRNA  
Accession: NM\_212617.1 GI: 47086800
- Xenopus laevis YY1 transcription factor L homeolog (yy1.L), mRNA**  
1,924 bp linear mRNA  
Accession: NM\_001087615.1 GI: 148234201
- PREDICTED: Canis lupus familiaris YY1 transcription factor (YY1), mRNA**  
2,220 bp linear mRNA

The interface includes filters for Species, Molecule types, Source databases, Genetic compartments, Sequence length, Release date, and Revision date. The search results are sorted by Default order. The page also shows a 'Results by taxon' section with top organisms like Homo sapiens, Mus musculus, and Danio rerio. A 'Find related data' section is also present.

# NCBI Nucleotide. Карточка последовательности гена

## Особенности нуклеотидной последовательности

```
exon 14..1543
/gene="YY1"
/gene_synonym="DELTA; IN0805; NF-E1; UCRBP; YIN-YANG-1"
/inference="alignment:Splign:2.0.8"
1543..3159
/gene="YY1"
/gene_synonym="DELTA; IN0805; NF-E1; UCRBP; YIN-YANG-1"
/inference="alignment:Splign:2.0.8"
STS 1729..1852
/gene="YY1"
/gene_synonym="DELTA; IN0805; NF-E1; UCRBP; YIN-YANG-1"
/standard_name="RH1618"
/db_xref="UniSTS:42115"
STS 1954..2640
/gene="YY1"
/gene_synonym="DELTA; IN0805; NF-E1; UCRBP; YIN-YANG-1"
/standard_name="YY1_3892"
/db_xref="UniSTS:462950"
STS 2035..2173
/gene="YY1"
/gene_synonym="DELTA; IN0805; 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; IN0805; NF-E1; UCRBP; YIN-YANG-1"
polyA_site 2714
/gene="YY1"
/gene_synonym="DELTA; IN0805; NF-E1; UCRBP; YIN-YANG-1"
STS 2879..3086
/gene="YY1"
/gene_synonym="DELTA; IN0805; NF-E1; UCRBP; YIN-YANG-1"
/standard_name="RH68412"
/db_xref="UniSTS:51218"
STS 3001..3100
/gene="YY1"
/gene_synonym="DELTA; IN0805; 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; IN0805; NF-E1; UCRBP; YIN-YANG-1"
polyA_site 3132
/gene="YY1"
/gene_synonym="DELTA; IN0805; NF-E1; UCRBP; YIN-YANG-1"
ORIGIN
1 agggcgaacg ggcgagtggc agcgaggcgg ggcgggctga ggccagcggc gaagtctcgc
61 gaggccgggc ccgagcagag tgtggcggcg gcggcgagat ctgggctcgg gttgaggagt
121 tggattttgt gtggaaggag gcggaaggcg aggaggaagg ggaagcggga gcgccggccc
181 ggaggggcgg aggagggcgg gccaggcggg gcggttgccg cgagcggagg cgagcggggc
241 agccgagacg agcagcggcg gaggcggggc gacgagcggc gcaccgagcc gagggagcgg
301 gggaagcccc gccgcccggc cggcgcccgc cccttcccc gccgccccc ccctctccc
361 ccgcccgctc gccgcttcc tcctctgccc ttcttcccc acggcggccc gcctctcgc
421 ccgcccgccc gcagccgagg agccgaggcc gccgcgggcg tggcggcggg gcctcagcc
```

NCBI Reference Sequence: NM\_003403.4

FASTA Graphics

Go to: [ ]

LOCUS NM\_003403 3159 bp mRNA linear PRI 07-OCT-2016

DEFINITION Homo sapiens YY1 transcription factor (YY1), mRNA.

ACCESSION NM\_003403

VERSION NM\_003403.4

KEYWORDS RefSeq.

SOURCE Homo sapiens (human)

ORGANISM Homo sapiens

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 3159)

AUTHORS Tsang DP, Wu WK, Kang W, Lee YY, Wu F, Yu Z, Xiong L, Chan AW, Tong JH, Yang W, Li MS, Lau SS, Li X, Lee SD, Yang Y, Lai PB, Yu DY, Xu G, Lo KW, Chan MT, Wang H, Lee TL, Yu J, Wong N, Yip KY, To KF and Cheng AS.

TITLE Yin Yang 1-mediated epigenetic silencing of tumour-suppressive microRNAs activates nuclear factor-kappaB in hepatocellular carcinoma

JOURNAL J. Pathol. 238 (5), 651-664 (2016)

PUBMED 26800240

REMARK GeneRIF: YY1 overexpression contributes to EZH2 recruitment for H3K27me3-mediated silencing of tumour-suppressive microRNAs, thereby activating NF-kappaB signalling in hepatocarcinogenesis.

REFERENCE 2 (bases 1 to 3159)

AUTHORS Nieborak A and Gorecki A.

TITLE Significance of the pathogenic mutation T372R in the Yin Yang 1 protein interaction with DNA--thermodynamic studies

JOURNAL FEBS Lett. 590 (6), 838-847 (2016)

PUBMED 26910132

REMARK GeneRIF: Significance of the pathogenic mutation T372R in the Yin Yang 1 protein interaction with DNA--thermodynamic studies: the mutation does not affect the secondary structure of either zinc

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Articles about the YY1 gene

A High-Density Map for Navigating the Human Polycomb Complex [Cell Rep. 2016]

Significance of the pathogenic mutation T372R in the Yin Yang 1 p [FEBS Lett. 2016]

Regulation of Transcription Factor Yin Yang 1 by SET7/9-mediated Lysin [Sci Rep. 2016]

See all...

Pathways for the YY1 gene

UCH proteinases

Deubiquitination

TFAP2 (AP-2) family regulates transcription of growth factors and their receptors

See all...

Общая информация

Локус

Идентификатор

Организм и таксономия

Ссылки на литературу

Продолжение информационной карточки

# GeneCards

<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

## Информационная карточка гена

The screenshot displays the GeneCards website interface for the YY1 Gene. The page title is "YY1 Gene (Protein Coding) ★" with the subtitle "YY1 Transcription Factor". Key statistics shown include GCID: GC14P100238 and GIFIS: 64. The page features a navigation menu with categories like "Aliases", "Disorders", "Domains", "Drugs", "Expression", "Function", "Genomics", "Localization", and "Orthologs". Below the navigation, there are several service providers: EMD MILLIPORE (Proteins & Enzymes, Antibodies, Assays & Kits), GenScript (Genes Peptides, Proteins CRISPR), ORIGENE (Proteins Antibodies, Assays Genes, shRNA Primers, CRISPR), and Vigene Biosciences (Genes (adenoviral), Genes (lentiviral), miRNA shRNA (AAV)).

**Aliases for YY1 Gene**

YY1 Transcription Factor<sup>2 3 5</sup>  
INO80 Complex Subunit S<sup>2 3 4</sup>  
Delta Transcription Factor<sup>3 4</sup>  
Yin And Yang 1 Protein<sup>2 3</sup>  
INO80S<sup>3 4</sup>  
NF-E1<sup>3 4</sup>  
YY-1<sup>3 4</sup>

Transcriptional Repressor Protein YY1<sup>3</sup>  
Yin And Yang 1<sup>4</sup>  
YIN-YANG-1<sup>3</sup>  
Delta<sup>3</sup>  
UCRBP<sup>3</sup>

**External Ids for YY1 Gene**  
HGNC: 12856 Entrez Gene: 7528 Ensembl: ENSG00000100811 OMIM: 600013 UniProtKB: P25490

**Previous GeneCards Identifiers for YY1 Gene**  
GC14P098216, GC14P094520, GC14P098695, GC14P099774, GC14P100992, GC14P101002, GC14P080888, GC14P101020, GC14P101031, GC14P101051, GC14P101071, GC14P101112, GC14P101161, GC14P101221

Search aliases for YY1 gene in PubMed and other databases

**Summaries for YY1 Gene**

**Entrez Gene Summary for YY1 Gene**  
YY1 is a ubiquitously distributed transcription factor belonging to the GLI-Kruppel class of zinc finger proteins. The protein is involved in repressing and activating a diverse number of promoters. YY1 may direct histone deacetylases and histone acetyltransferases to a promoter in order to activate or repress the promoter, thus implicating histone modification in the function of YY1. [provided by RefSeq, Jul 2008]

# 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 паралограм
  - ✓Биологическим процессам
  - ✓Полиморфизмам
  - ✓Экспрессии
  - ✓Регуляции и др.

## Информационная карточка гена

www.ensembl.org/Homo\_sapiens/Gene/Summary?db=core;g=ENSG00000100811;r=14:100238298-100282792

Ensembl BLAST/BLAT | BioMart | Tools | Downloads | More

Human (GRCh38.p7) Location: 14:100,238,298-100,282,792 Gene: YY1

**Gene-based displays**

- Summary
  - Splice variants
  - Transcript comparison
  - Gene alleles
- Sequence
  - Secondary Structure
- Comparative Genomics
  - Genomic alignments
  - Gene tree
  - Gene gain/loss tree
  - Orthologues
  - Paralogues
  - Ensembl protein families
- Ontologies
  - GO: Biological process
  - GO: Molecular function
  - GO: Cellular component
- Phenotypes
- Genetic Variation
  - Variant table
  - Variant image
  - Structural variants
- Gene expression
- Regulation
- External references
- Supporting evidence
- ID History
  - Gene history

Configure this page

Custom tracks

Export data

Share this page

Bookmark this page

**Gene: YY1** ENSG00000100811

**Description** YY1 transcription factor [Source:HGNC Symbol;Acc:HGNC:12856]

**Synonyms** INO80S, hsa-mir-6764, UCRBP, YIN-YANG-1, NF-E1, DELTA

**Location** [Chromosome 14: 100,238,298-100,282,792](#) forward strand. GRCh38:CM000676.2

**About this gene** This gene has 7 transcripts ([splice variants](#)), [63 orthologues](#), [17 paralogues](#), is a member of [1 Ensembl protein family](#) and is associated with [1 phenotype](#).

**Transcripts** [Show transcript table](#)

**Summary**

**Name** [YY1](#) (HGNC Symbol)

**CCDS** This gene is a member of the Human CCDS set: [CCDS9957.1](#)

**UniProtKB** This gene has proteins that correspond to the following UniProtKB identifiers: [P25490](#)

**RefSeq** Overlapping RefSeq Gene ID [7528](#) matches and has similar biotype of protein\_coding

**Ensembl version** ENSG00000100811.11

**Other assemblies** This gene maps to [100,704,635-100,749,129](#) in GRCh37 coordinates. View this locus in the GRCh37 archive: [ENSG00000100811](#)

**Gene type** Known protein coding

**Annotation method** Annotation for this gene includes both automatic annotation from Ensembl and [Havana](#) manual curation, see [article](#).

**Alternative genes** **This gene corresponds to the following database identifiers:**  
**Havana gene:** [OTTHUMG00000150479](#)

**Annotation Attributes** overlapping locus [Definitions](#)

[Go to Region in Detail for more tracks and navigation options \(e.g. zooming\)](#)

64.50 kb Forward strand

Genes (Comprehensive.set....)

YY1-004 > protein coding

YY1-005 > retained intron

YY1-002 > TEC

RP11-638I2.6-001 > lincRNA

# 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 uc001ygu.3, другой - uc059fch.1
- Содержится информация по:
  - ✓ По интрон/экзонной структуре гена
  - ✓ Последовательности
  - ✓ Биологическим процессам
  - ✓ Полиморфизмам
  - ✓ Экспрессии
  - ✓ Регуляции и др.

## Информационная карточка гена

The screenshot displays the UCSC Genome Browser interface for the Human Gene YY1 (ENST00000262238.8). The top navigation bar includes links for Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, View, and Help. The main content area shows the gene's location on chromosome 14 (chr14:100,238,985-100,282,792) and provides a detailed description of the gene's function and structure. The description includes the RefSeq Summary (NM\_003403) and the GeneCode Transcript (ENST00000262238.8). The page index table lists various resources and links for the gene.

**Human Gene YY1 (ENST00000262238.8) Description and Page Index**

**Description:** Homo sapiens YY1 transcription factor (YY1), mRNA. (from RefSeq NM\_003403)

**RefSeq Summary (NM\_003403):** YY1 is a ubiquitously distributed transcription factor belonging to the GLI-Kruppel class of zinc finger proteins. YY1 may direct histone deacetylases and histone H3 to a promoter in order to activate or repress the promoter, thus implicating histone modification in the function of YY1. [provided by RefSeq, Jul 2008]. Sequence Note: This RefSeq record was created from transcript and genomic sequence data to make the sequence consistent with the reference genome assembly. The genomic coordinates used for the transcript record were based on transcript alignments. Public RefSeq record includes a subset of the publications that are available for this gene. Please see the Gene record to access additional publications. ##Evidence-Data-START## Transcript exon combination :: M77698.1, BC037308.1 [ECO:0000332] RNAseq introns :: single sam introns SAMEA1965299, SAMEA1966682 [ECO:0000348] ##Evidence-Data-END##

**GeneCode Transcript:** ENST00000262238.8

**Transcript (Including UTRs)**  
Position: hg38 chr14:100,238,985-100,282,792 Size: 43,808 Total Exon Count: 5 Strand: +

**Coding Region**  
Position: hg38 chr14:100,239,245-100,277,600 Size: 38,356 Coding Exon Count: 5

Page Index	Sequence and Links	UniProtKB Comments	MalaCards	CTD	RNA-Seq Expression
Microarray Expression	RNA Structure	Protein Structure	Other Species	GO Annotations	mRNA Descriptions
Pathways	Other Names	Methods			

Data last updated: 2016-03-28

**Sequence and Links to Tools and Databases**

Genomic Sequence (chr14:100,238,985-100,282,792)	mRNA (may differ from genome)	Protein (414 aa)			
Gene Sorter	Genome Browser	Other Species FASTA	Table Schema	BioGPS	CGAP
Ensembl	Entrez Gene	ExonPrimer	GeneCards	Gepis Tissue	HGNC
HPRD	Lynx	MGI	MOPED	neXtProt	OMIM
PubMed	Reactome	Stanford SOURCE	UniProtKB	Wikipedia	

**Comments and Description Text from UniProtKB**

ID: **YY1\_HUMAN**

**DESCRIPTION:** RecName: Full=Transcriptional repressor protein YY1; AltName: Full=Delta transcription factor; AltName: Full=IN subunit S; AltName: Full=NF-E1; AltName: Full=Yin and yang 1; Short=YY-1;

**FUNCTION:** Multifunctional transcription factor that exhibits positive and negative control on a large number of cellular and viral genes overlapping the transcription start site. Binds to the consensus sequence 5'-CCGCCATNTT-3'; some genes have been shown longer binding motif allowing enhanced binding; the initial CG dinucleotide can be methylated greatly reducing the binding affinity. transcription regulation is depending upon the context in which it binds and diverse mechanisms of action include direct activation

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.

genome.ucsc.edu/cgi-bin/hgGene?db=hg19&hgg\_gene=uc003adj.3

Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

### Human Gene MN1 (uc003adj.3) Description and Page Index

**Description:** Homo sapiens meningioma (disrupted in balanced translocation) 1 (MN1), mRNA.

**RefSeq Summary (NM\_002430):** Meningioma 1 (MN1) contains two sets of CAG repeats. It is disrupted by a balanced translocation (4;22) in a meningioma, and its inactivation may contribute to meningioma 32 pathogenesis. [provided by RefSeq, Jul 2008]. Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Gene record to access additional publications. ##Evidence-Data-START## Transcript exon combination :: X82209.2, CX866728.1 [ECO:0000332] RNAseq introns :: single sample supports all introns SAMEA2147975, SAMN03465404 [ECO:0000348] ##Evidence-Data-END##

**Transcript (Including UTRs)**  
**Position:** hg19 chr22:28,144,265-28,197,486 **Size:** 53,222 **Total Exon Count:** 2 **Strand:** -

**Coding Region**  
**Position:** hg19 chr22:28,146,903-28,196,531 **Size:** 49,629 **Coding Exon Count:** 2

<b>Page Index</b>	Sequence and Links	UniProtKB Comments	Genetic Associations	MalaCards	CTD
Gene Alleles	RNA-Seq Expression	Microarray Expression	RNA Structure	Protein Structure	Other Species
GO Annotations	mRNA Descriptions	Other Names	Model Information	Methods	

Data last updated: 2013-06-14

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

Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

### Human Gene MN1 (uc003adj.3) Description and Page Index

**Description:** Homo sapiens meningioma (disrupted in balanced translocation) 1 (MN1), mRNA.

**RefSeq Summary (NM\_002430):** Meningioma 1 (MN1) contains two sets of CAG repeats. It is disrupted by a balanced translocation (4;22) in a meningioma, and its inactivation may contribute to meningioma pathogenesis. [provided by RefSeq, Jul 2008]. Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Gene page to access additional publications. ##Evidence Data START## Transcript exon combination: Y82208.2, CY866728.1, F50002221, BNAseq intron: single sample

**Transcript (uc003adj.3)**  
Position: chr22:27748277-27801498

**Coding Region (uc003adj.3)**  
Position: chr22:27748277-27801498

#### Known Genes

<a href="#">MN1 (uc003adj.3) at chr22:27748277-27801498</a>	- Homo sapiens meningioma (disrupted in balanced translocation) 1 (MN1), mRNA. (from RefSeq NM_002430)
<a href="#">MN1 (uc062csm.1) at chr22:27750064-27796896</a>	- The sequence shown here is derived from an Ensembl automatic analysis pipeline and should be considered as preliminary data. (from UniProt H7C105)
<a href="#">MN1 (uc010gvg.6) at chr22:27750678-27791883</a>	- meningioma (disrupted in balanced translocation) 1 (from HGNC MN1)

**Page Index**  
Gene Allele

GO Annotations	mRNA Descriptions	Other Names	Model Information	Methods
----------------	-------------------	-------------	-------------------	---------

Data last updated: 2013-06-14

# GEO - Gene Expression Omnibus

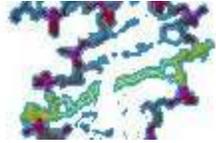
- <https://www.ncbi.nlm.nih.gov/geo/>
- Данные по экспрессии генов. Высокопроизводительные эксперименты по оценке уровней экспрессии генов (методы ДНК-микрочипов и RNA-seq).
- Идентификаторы для DataSet - GDSчисло и для DataSeries - GSEчисло. Например, серия экспериментов, посвященных анализу уровня 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

## Информационная карточка вариации

- <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.

**dbSNP** Short Genetic Variations

Reference SNP (rs) Report

[Switch to classic site](#)

**rs61992955**

**Organism** Homo sapiens

**Position** chr14:100274488

**Alleles** G>A / G>C

**Variation Type** SNV Single Nucleotide

**Frequency** C=0.04944 (6208)  
C=0.0509 (1576/3)  
C=0.030 (150/500)

**Variant Details** Genomic  
Sequenc  
GRCh37  
GRCh37  
GRCh38  
GRCh38  
YY1 RefS  
YY1 RefS

**Clinical Significance** Not Reported in ClinVar

**Gene : Consequence** YY1 : Intron Variant

**Publications** 0 citations

**Genomic View** [See rs on genome](#)

**Change**

NC_000014.8:g.100740825G>C
NC_000014.8:g.100740825G>A
NC_000014.9:g.100274488G>C
NC_000014.9:g.100274488G>A
NG_046908.1:g.40724G>C
NG_046908.1:g.40724G>A

В базе данных dbSNP для гена AGLB4 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 AGLB4 в большей степени подвержен мутагенезу?

### Human Gene AGLB4 (uc001cru.2) Description and Page Index

**Description:** Homo sapiens ATP/GTP binding protein-like 4 (AGLB4), mRNA.  
**Transcript (Including UTRs):**  
**Position:** chr1:48,998,527-50,489,626 **Size:** 1,491,100 **Total Exon Count:** 14 **Strand:** -  
**Coding Region:**  
**Position:** chr1:48,999,845-50,489,468 **Size:** 1,489,624 **Coding Exon Count:** 14

SNP

Display Settings: Summary, 20 per page, Sorted by SNP\_ID

Send to:

Search results

Items: 1 to 20 **221431**

<< First < Prev Page 1 of 11072 Next > Last >>

Find related data  
 Database:

Search details  
 AGLB4 [All Fields]

Recent activity

- rs17378497 has merged into rs355206 [Homo sapiens]
- 1. ATAAAACTGCTTTTCAGTTAAGG [A/C] AACCTGCTTTTGACCCCTTTCCAAA  
 Chromosome: 1:49206502  
 Gene: AGLB4 (GeneView)  
 Functional Consequence: intron variant  
 Validated: by 1000G, by 2hit 2allele, by cluster, by frequency  
 Global MAF: A=0.4091/2049  
 HGVS: NC\_000001.10:g.49672174C>A, NC\_000001.11:g.49206502C>A, NM\_001323573.1:c.413+39268G>T, NM\_001323574.1:c.413+39268G>T, NM\_001323575.1:c.377+39268G>T, NM\_032785.3:c.377+39268G>T, NR\_136623.1:n.410+39268G>T, XM\_005271284.1:c.413+39268G>T, XM\_011542308.2:c.413+39268G>T, XM\_011542310.2:c.413+39268G>T, XM\_017002595.1:c.377+39268G>T, XM\_017002596.1:c.377+39268G>T, XM\_017002597.1:c.377+39268G>T, XM\_017002598.1:c.377+39268G>T
- rs17379875 has merged into rs4926831 [Homo sapiens]
- 2. TGCTAAGCTCTGTGTTGAGCCTCA [C/T] GATTGTCAAATGAATGAGATATGG  
 Chromosome: 1:49824429  
 Gene: AGLB4 (GeneView)  
 Functional Consequence: intron variant  
 Validated: by 1000G, by 2hit 2allele, by cluster, by frequency, by hapmap  
 Global MAF: T=0.2883/1444  
 HGVS: NC\_000001.10:g.50290101C>T, NC\_000001.11:g.49824429C>T, NM\_001323573.1:c.157+26967G>A, NM\_001323574.1:c.157+26967G>A, NM\_001323575.1:c.157+26967G>A, NM\_032785.3:c.157+26967G>A, NR\_136623.1:n.315+26967G>A, XM\_005271284.1:c.157+26967G>A, XM\_011542308.2:c.157+26967G>A, XM\_011542310.2:c.157+26967G>A, XM\_017002595.1:c.157+26967G>A, XM\_017002596.1:c.157+26967G>A, XM\_017002597.1:c.157+26967G>A, XM\_017002598.1:c.157+26967G>A, XM\_017002599.1:c.157+26967G>A

### Human Gene GALE (uc001bhx.1) Description and Page Index

**Description:** UDP-galactose-4-epimerase  
**Transcript (Including UTRs):**  
**Position:** chr1:23,994,676-23,999,881 **Size:** 5,206 **Total Exon Count:** 12 **Strand:** -  
**Coding Region:**  
**Position:** chr1:23,995,026-23,998,084 **Size:** 3,059 **Coding Exon Count:** 10

SNP

Display Settings: Summary, 20 per page, Sorted by SNP\_ID

Send to:

Search results

Items: 1 to 20 **2843**

<< First < Prev Page 1 of 143 Next > Last >>

Find related data  
 Database:

Search details  
 GALE [All Fields]

Recent activity

- rs28940885 [Homo sapiens]
- 1. CCCAGCTGGCCCAAGGAGGCTGG [A/G] GTGGACAGCAGCCTTAGGGCTGGAC  
 Chromosome: 1:23796183  
 Gene: GALE (GeneView)  
 Functional Consequence: missense  
 Allele Origin: G(germline)/A(germline)  
 Clinical significance: Pathogenic  
 Validated: by 1000G, by cluster, by frequency  
 Global MAF: T=0.0020/10  
 HGVS: NC\_000001.10:g.24122673C>T, NC\_000001.11:g.23796183C>T, NG\_007068.1:g.9622G>A, NM\_000403.3:c.956G>A, NM\_01008216.1:c.956G>A, NM\_001127621.1:c.956G>A, NP\_000394.2:p.Gly319Glu, NP\_01008217.1:p.Gly319Glu, NP\_001121093.1:p.Gly319Glu, XM\_005245833.1:c.1066G>A, XM\_005245834.1:c.1066G>A, XM\_005245835.1:c.1066G>A, NP\_005245890.1:p.Gly356Ser, XP\_005245891.1:p.Gly356Ser, XP\_005245892.1:p.Gly356Ser, XP\_005245893.1:p.Gly292Ser, XP\_005245894.1:p.Gly255Glu
- rs28940884 [Homo sapiens]
- 2. AAGGGCCACATTGCGACCTTAAGGA [A/G] GCTGAAAGAACAGTGGCTGCCGG  
 Chromosome: 1:23796722  
 Gene: GALE (GeneView)  
 Functional Consequence: missense  
 Allele Origin: G(germline)/A(germline)  
 Clinical significance: Pathogenic  
 Validated: by 1000G, by cluster, by frequency  
 Global MAF: C=0.0082/44  
 HGVS: NC\_000001.10:g.24123212T>C, NC\_000001.11:g.23796722T>C, NC\_007068.1:g.9083A>G, NM\_000403.3:c.770A>G, NM\_01008216.1:c.770A>G, NM\_001127621.1:c.770A>G, NP\_000394.2:n.Iso257Arm, NP\_01008217.1:n.Iso257Arm

- GALE (2843) SNP
- AGLB4 (221431) SNP
- Titin (3871) SNP
- Titin TTN (1) SNP
- dystrophin DMD (6) SNP

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.

### Human Gene AGBL4 (uc001cru.2) Description and Page Index

Description: Homo sapiens ATP/GTP binding protein-like 4 (AGBL4), mRNA.

Transcript (Including UTRs):

Position: chr1:48,998,527-50,489,622 **Size: 1,491,100** total Exon Count: 14 Strand: -

Coding Region:

Position: chr1:48,999,845-50,489,468 **Size: 1,489,624** Coding Exon Count: 14

### Human Gene GALE (uc001bhx.1) Description and Page Index

Description: UDP-galactose-4-epimerase

Transcript (Including UTRs):

Position: chr1:23,994,676-23,999,882 **Size: 5,206** total Exon Count: 12 Strand: -

Coding Region:

Position: chr1:23,995,026-23,998,084 **Size: 3,059** Coding Exon Count: 10

SNP Search Results for AGBL4

Search results: 221,431 items

Search details: AGBL4 (Gene)

Recent activity: AGBL4 (221431), Tbx (3571), Tbx TTN (7), cytochrome P450 (5), cytochrome (7)

SNP Search Results for GALE

Search results: 2,843 items

Search details: GALE (Gene)

Recent activity: GALE (2843), AGBL4 (221431), Tbx (3571), Tbx TTN (7), cytochrome P450 (5)

- <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

The screenshot displays the ClinVar website interface for a specific variant. The browser address bar shows the URL: <https://www.ncbi.nlm.nih.gov/clinvar/variation/91950/#supporting-observatio>. The page title is "ClinVar" and the search bar contains "ClinVar". The variant details are as follows:

- Variant ID:** 91950
- Review status:** (0/4) no assertion criteria provided
- Clinical significance:** [Uncertain significance](#)
- Number of submission(s):** 1
- Allele(s):** NM\_003403.4(YY1):c.1115C>G (p.Thr372Arg)
- Allele ID:** 97428
- Variant type:** single nucleotide variant
- Cytogenetic location:** 14q32.2
- Genomic location:**
  - Chr14: 100277470 (on Assembly GRCh38)
  - Chr14: 100743807 (on Assembly GRCh37)
- Protein change:** T372R
- HGVS:**
  - NG\_046908.1:g.43706C>G
  - NM\_003403.4:c.1115C>G
  - NP\_003394.1:p.Thr372Arg
- Links:**
  - UniProtKB: [P25490#VAR\\_074172](#)
  - dbSNP: [386834266](#)
- NCBI 1000 Genomes Browser:** [rs386834266](#)
- Molecular consequence:** NM\_003403.4:c.1115C>G: missense variant [Sequence Ontology SO:0001583]

Additional information on the right side of the page includes:

- 1 Affected gene:** YY1 transcription factor (YY1) [Gene OMIM - Variation Viewer]
- Variant frequency in dbGaP:** NM\_003403.4(YY1):c.1115C>G (p.Thr372Arg) GRCh37 Chr14:100743807
- Called variants table:**

Sample count	Called variants	Potential variants
no data		0 of 4090
- Browser views:** RefSeqGene, Variation Viewer [GRCh38 - GRCh37], UCSC [GRCh38/hg38 - GRCh37/hg19]
- Related information:** dbSNP, Gene, MedGen, Related genes (specific)

# 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

rs2494262

[PMID 19685047] FcepsilonR1alpha gene -18483A>C polymorphism affects transcriptional activity through YY1 binding

[PMID 18846228] Genome-wide scan on total serum IgE levels identifies FCER1A as novel susceptibility locus.

[PMID 23525950] Single-nucleotide polymorphisms of allergy-related genes and risk of adult glioma

[PMID 25412950] Association of FCER1A genetic polymorphisms

Orientation	plus
Stabilized	plus
Make rs2494262(A;A)	
Make rs2494262(A;C)	
Make rs2494262(C;C)	
Reference	GRCh38
	38.1/141
Chromosome	1
Position	159283882
Gene	FCER1A
GWAS Ctig	rs2494262
GMAF	0.4601
Max	
Magnitude	

Population	(A;A)	(A;C)	(C;C)
CEU	~35%	~45%	~20%
HCB	~55%	~35%	~10%
JPT	~65%	~25%	~10%
YRI	~10%	~45%	~45%
ASW	~15%	~45%	~40%
CHB	~55%	~35%	~10%
CHD	~55%	~35%	~10%
GIH	~45%	~45%	~10%
LWK	~15%	~45%	~40%
MEX	~15%	~45%	~40%
MKK	~15%	~45%	~40%
TSI	~35%	~45%	~20%
AVG	~35%	~45%	~20%

# 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

The screenshot shows the GWAS Catalog search results for the SNP rs2766692. The page includes a search bar with the query 'rs2766692' and a search button. Below the search bar, there are navigation links for 'Home', 'Search', 'Diagram', 'Download', 'Documentation', and 'About'. The main content area displays the search results for rs2766692, including a table of studies and a section for associations. The table of studies lists the author (Kang SJ), date (2012-05-03), journal (Genes Brain Behav), title (Family-based genome-wide association study of frontal  $\theta$  oscillations identifies potassium channel gene KCNJ6), reported trait (Electroencephalographic traits in alcoholism), and association count (8). The associations section shows various metrics such as p-value, OR, Beta, CI, Region, Functional, and Reported gene(s).

GWAS Catalog

The NHGRI-EBI Catalog of published genome-wide association studies

rs2766692

Examples: breast cancer, rs7329174, Yang, 2q37.1, HBS1L

GWAS / Search / rs2766692

Refine search results

Show results for

- Studies 1
- Associations 1
- Catalog traits 9

Filter results by

p-value  $\leq 5 \times 10^{-8}$

Odds ratio from to

Beta coefficient

Search results for rs2766692

Download association results

Expand all studies

Studies

Author	Date	Journal	Title	Reported trait	Association count
Kang SJ (PMID: 22554406)	2012-05-03	Genes Brain Behav	Family-based genome-wide association study of frontal $\theta$ oscillations identifies potassium channel gene KCNJ6.	Electroencephalographic traits in alcoholism	8

Associations

RAF p-value OR Beta CI Region Functional Reported gene(s)

# Базы данных, содержащие информацию по белкам



# Uniprot (Universal Protein Resource)

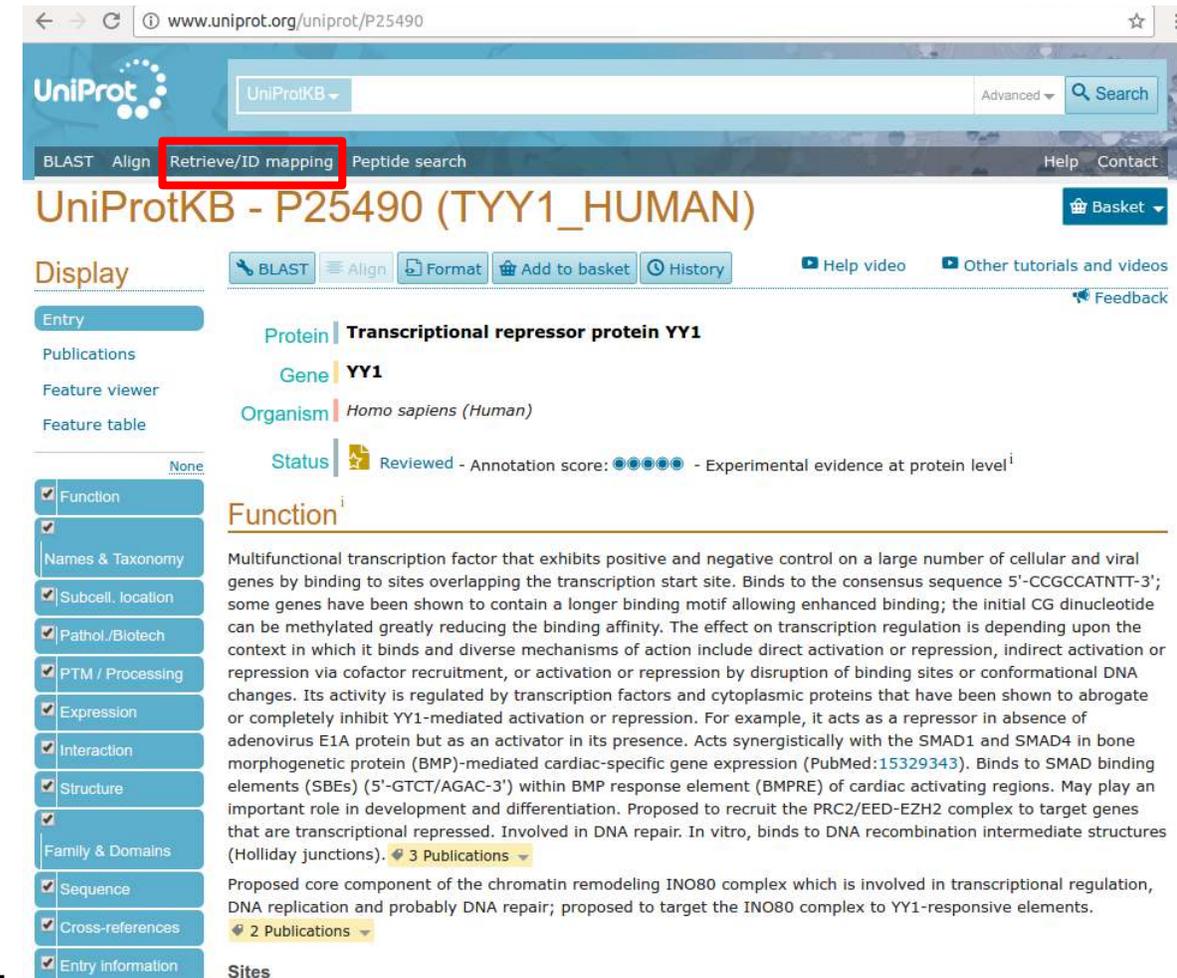
<http://www.uniprot.org/>

- 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



www.uniprot.org/uniprot/P25490

UniProtKB

BLAST Align **Retrieve/ID mapping** Peptide search Help Contact

UniProtKB - P25490 (TYY1\_HUMAN)

Display BLAST Align Format Add to basket History Help video Other tutorials and videos Feedback

Entry Protein | **Transcriptional repressor protein YY1**

Publications Gene | **YY1**

Feature viewer Organism | *Homo sapiens (Human)*

Feature table Status |  Reviewed - Annotation score: ●●●●●● - Experimental evidence at protein level<sup>i</sup>

None

Function

Names & Taxonomy

Subcell. location

Pathol./Biotech

PTM / Processing

Expression

Interaction

Structure

Family & Domains

Sequence

Cross-references

Entry information

**Function<sup>i</sup>**

Multifunctional transcription factor that exhibits positive and negative control on a large number of cellular and viral genes by binding to sites overlapping the transcription start site. Binds to the consensus sequence 5'-CCGCCATNTT-3'; some genes have been shown to contain a longer binding motif allowing enhanced binding; the initial CG dinucleotide can be methylated greatly reducing the binding affinity. The effect on transcription regulation is depending upon the context in which it binds and diverse mechanisms of action include direct activation or repression, indirect activation or repression via cofactor recruitment, or activation or repression by disruption of binding sites or conformational DNA changes. Its activity is regulated by transcription factors and cytoplasmic proteins that have been shown to abrogate or completely inhibit YY1-mediated activation or repression. For example, it acts as a repressor in absence of adenovirus E1A protein but as an activator in its presence. Acts synergistically with the SMAD1 and SMAD4 in bone morphogenetic protein (BMP)-mediated cardiac-specific gene expression (PubMed:15329343). Binds to SMAD binding elements (SBEs) (5'-GTCT/AGAC-3') within BMP response element (BMPRE) of cardiac activating regions. May play an important role in development and differentiation. Proposed to recruit the PRC2/EED-EZH2 complex to target genes that are transcriptionally repressed. Involved in DNA repair. In vitro, binds to DNA recombination intermediate structures (Holliday junctions).  3 Publications

Proposed core component of the chromatin remodeling INO80 complex which is involved in transcriptional regulation, DNA replication and probably DNA repair; proposed to target the INO80 complex to YY1-responsive elements.  2 Publications

Sites

# 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

GenPept

Send to:

Change region shown

Customize view

Analyze this sequence

Run BLAST

Identify Conserved Domains

Highlight Sequence Features

Find in this Sequence

Protein 3D Structure

Crystal Structure O Mtd1 Yy1 Complex PDB: 4C5I Source: Homo sapiens

LinkOut to external resources

A selection of literature about the proteins [GoPubMed Proteins]

MODBASE, Database of Comparative Pi [MODBASE, Database of Comparat..]

Transcript/Protein Information [PANTHER Classification System]

Protein Ontology Consortium [Protein Ontology Consortium]

reagents [ExactAntigen/Labome]

reagent reviews [ExactAntigen/Labome]

Recent activity

Turn Off Clear

YY1 transcription factor [Homo sapiens] Protein

(yy1) AND "Homo sapiens"[porgn] (160) Protein

NCBI Resources How To Sign in to NCBI

Protein Protein Search Help

Advanced

GenBank: AAH65366.1

Identical Proteins FASTA Graphics

Go to:

LOCUS AAH65366 414 aa linear PRI 15-JUL-2006

DEFINITION YY1 transcription factor [Homo sapiens].

ACCESSION AAH65366

VERSION AAH65366.1

DBSOURCE accession [BC065366.1](#)

KEYWORDS MGC.

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

REFERENCE 1 (residues 1 to 414)

AUTHORS Strausberg,R.L., Feingold,E.A., Grouse,L.H., Derge,J.G., Klausner,R.D., Collins,F.S., Wagner,L., Shenmen,C.M., Schuler,G.D., Altschul,S.F., Zeeberg,B., Buetow,K.H., Schaefer,C.F., Bhat,N.K., Hopkins,R.E., Jordan,H., Moore,T., Max,S.T., Wang,J., Heich,F.

Region /db\_xref="CDD:290200" /region name="C2H2 Zn finger" /note="C2H2 Zn finger [structural motif]" /db\_xref="CDD:275368" order(385,390,403,407) /site\_type="other" /note="Zn binding site [ion binding]" /db\_xref="CDD:275368" 1..414

Site /gene="YY1" /gene\_synonym="DELTA" /gene\_synonym="NF-E1" /gene\_synonym="UCRBP" /gene\_synonym="YIN-YANG-1" /coded\_by="BC065366.1:46..1290" /db\_xref="GeneID:7528" /db\_xref="HGNC:HGNC:12856" /db\_xref="MIM:600013"

CDS

ORIGIN

1 masgdtlyia tdgsempaei velheiev et ipvetiettv vgeeeeeedd dedggggdhdg 61 gggghghagh hhhhhhhhhh ppmialqplv tddptqvhhh qevilvqtre evvggddsdg 121 laedgfdedq ilipvpapag gddyieqtl vtvaagksg gggsssggg rvkkgggkks 181 gkksylsgga gaaggggpadp gnkkweqkv qiktlegefs vtmwssdekk diidhetvvee 241 qiigensppd yseymtgkkl ppggpgidl sdpkqlaefa rmkprkiked daprtiacph 301 kgctkmfrdn samrkhhlth gprvhcaec gkafvesskl krhqlvhtge kpfqctfegc 361 gkrfsldfnl rthvrihtgd rpyvcpfdc nkkfaqstnl kshilthaka knnq

# 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

www.rcsb.org/pdb/explore/explore.do?structureId=1DGB

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB Login

RCSB PDB An Information Portal to 127823 Biological Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or ligand Go

Advanced Search | Browse by Annotations | Search History (7) | Previous Results (6)

PDB-101 WORLDWIDE PROTEIN DATA BANK EMDatabank NUCLEIC ACID DATABASE StructuralBiology Knowledgebase Worldwide Protein Data Bank Foundation

Structure Summary 3D View Annotations Sequence Sequence Similarity Structure Similarity Experiment

Biological Assembly 1

1DGB  
HUMAN ERYTHROCYTE CATALASE  
DOI: 10.2210/pdb1dgb/pdb

Classification: **OXIDOREDUCTASE**  
Deposited: 1999-11-23 Released: 2000-02-11  
Deposition author(s): [Putnam, C.D.](#), [Arvai, A.S.](#), [Bourne, Y.](#), [Tainer, J.A.](#)  
Organism: [Homo sapiens](#)

Structural Biology Knowledgebase: 1DGB (>23 annotations) [SBKB.org](#)

Experimental Data Snapshot

Method: X-RAY DIFFRACTION  
Resolution: 2.2 Å  
R-Value Free: 0.227  
R-Value Work: 0.172

wwPDB Validation

Metric	Percentile Ranks	Value
Clashscore		10
Ramachandran outliers		0.3%
Sidechain outliers		3.2%
RSRZ outliers		1.2%

Literature

Download Primary Citation

Active and inhibited human catalase structures: ligand and NADPH binding and catalytic mechanism.  
[Putnam, C.D.](#), [Arvai, A.S.](#), [Bourne, Y.](#), [Tainer, J.A.](#)  
(2000) J.Mol.Biol. **296**: 295-309  
PubMed: 10656833 [Search on PubMed](#)

# Базы данных, содержащие информацию по микроРНК



miRDB

miRTarBase

# 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.

www.mirbase.org/cgi-bin/mirna\_entry.pl?acc=MI0003716

miRBase MANCHESTER 1824

Home Search Browse Help Download Blog Submit **mmu-mir-302b** Search

### Stem-loop sequence mmu-mir-302b

Accession MI0003716

Symbol [MGI:Mir302b](#)

Description Mus musculus miR-302b stem-loop

Gene family MIPF0000071; [mir-302](#)

Stem-loop

```
5' guucc a uu au ucugu au
   cuuc acu aacauggga gcuu cuc c
   ||||| ||| ||||| ||||| |||
3' gaag uga uuguaccuu ugaa gag g
   ---u a uu cg ---u aa
```

Get sequence

1633 reads, 84.5 reads per million, 35 experiments

Deep sequencing

Confidence Feedback: Do you believe this miRNA is real?

Comments

Mouse miR-302b was verified experimentally by Mineno et al using MPSS technology [1]. The MPSS protocol used provides 22nt sequences, but the true extents of the mature miRNA are not reliably obtained. The mature sequence shown here represents the most commonly cloned form from large-scale cloning studies [2]. The 5' end of the miRNA may be offset with respect to previous annotations.

Genome context

Coordinates (GRCm38) chr3: [127545228-127545301](#) [+]

Overlapping transcripts antisense [ENSMUST00000029588](#); Larp7-201; intron 8

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).



**MicroRNA and Target Gene Description:**

<b>miRNA Name</b>	hsa-miR-302b-3p	<b>miRNA Sequence</b>	UAAGUGCUUCAUGUUUUAGUAG
<b>Previous Name</b>	hsa-miR-302b	<b>Target Score</b>	100
<b>Target Score</b>	100	<b>Seed Location</b>	984, 1137, 1538, 1600, 3494
<b>NCBI Gene ID</b>	<a href="#">55432</a>	<b>GenBank Accession</b>	<a href="#">NM_018566</a>
<b>Gene Symbol</b>	YOD1	<b>3' UTR Length</b>	5171
<b>Gene Description</b>	YOD1 deubiquitinase		

**3' UTR Sequence**

```

1 CCTATGCATG AATGAGGGTT GAAGCCTACT ACCTCACACA TCCAGAAGGC TCTGGGTTTT
61 CCAATAAGCT ATGGTAACCC TAAAGAACAA AGGATACAAT GCTTGAACCA TCCTTTTAAC
121 TTA AAAACAC TAAGACTCTG AAATTCCTTG TTAAGATTAA AATTAGTGTG CAAGTTTACA
181 GATGTGTGTC TACAGTGGTA AACTGTACAT ACATGCCTCT TTCTGCTGGA GTGACAGAAT
241 AGGTGATGCT TGGACCTAC TGACACTGAC CTGAAGGTTG AGATTGACTA TTATAAAGTA
    
```



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

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

# miRTarBase

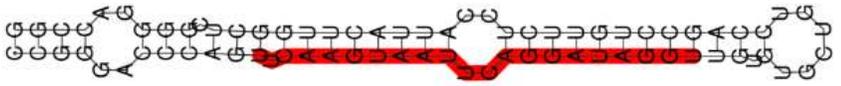
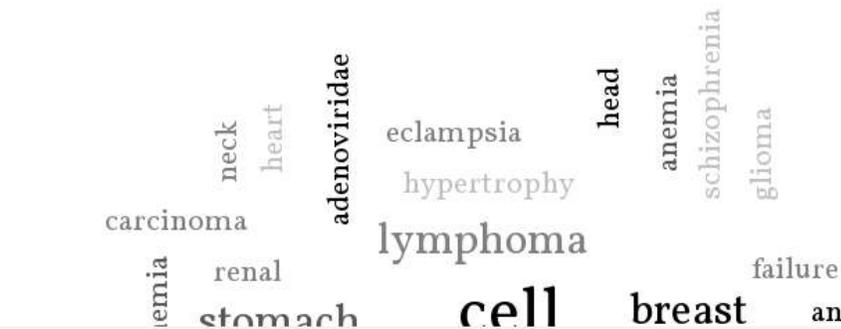
- <http://mirtarbase.mbc.nctu.edu.tw/>
- Содержит информацию по экспериментально подтвержденным взаимодействиям микроРНК-мишень для 23 организмов.
- Идентификаторы: микроРНК человека hsa-miR-26b-5p имеет идентификатор MIRT029499.
- Содержит информацию по
  - ✓ Мишеням (23054 генов для 4076 микроРНК)
  - ✓ Последовательности
  - ✓ Ассоциированным заболеваниям
  - ✓ Ссылки на публикации и др.

## Информационная карточка микроРНК.

**miRTarBase**

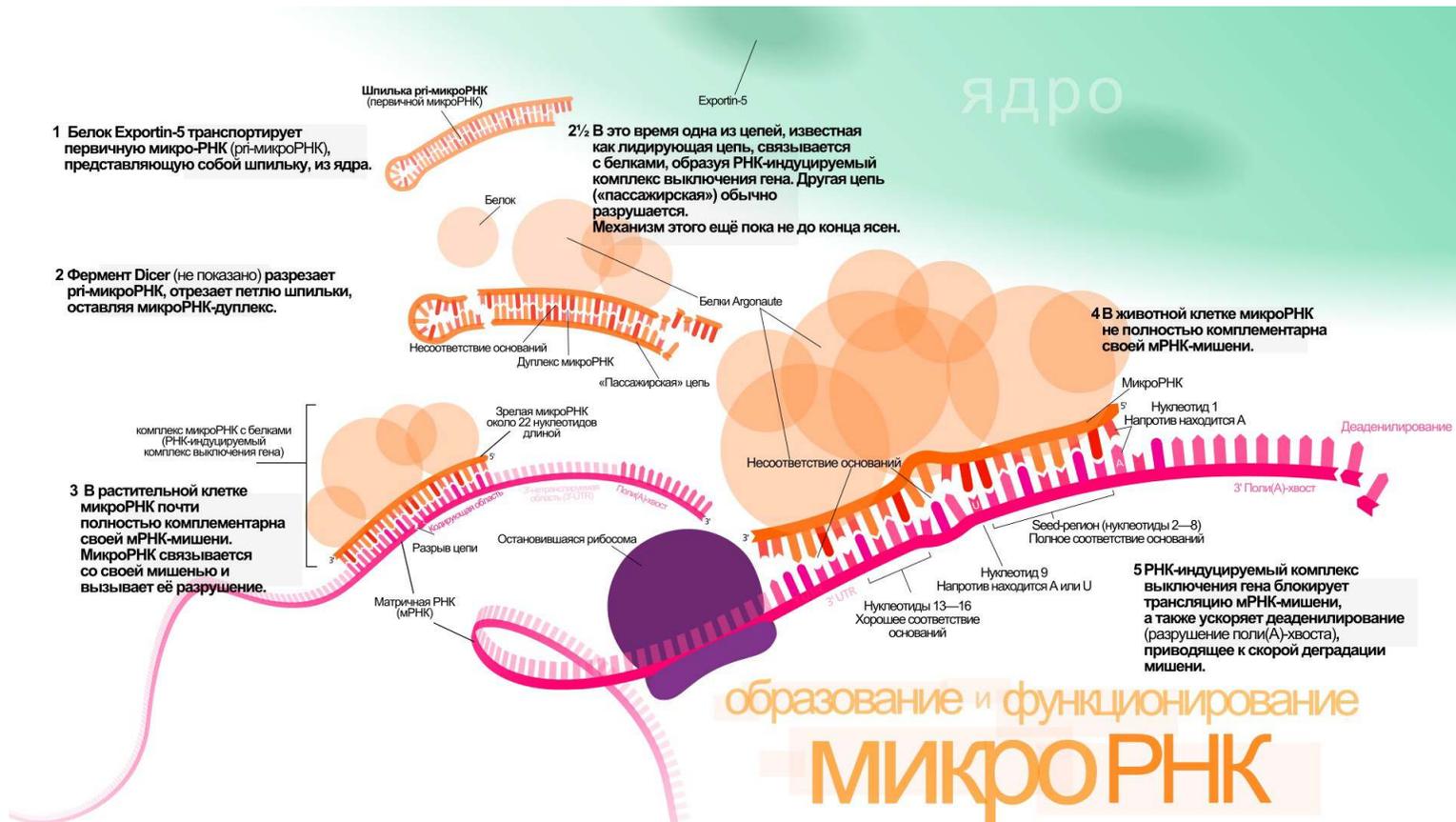
Accession ID: MIRT029499 [miRNA, hsa-miR-26b-5p :: MIR22HG, target gene]

miRNA Target Gene Evidences Expression TCGA Gene Set Enrichments Network ERROR Report

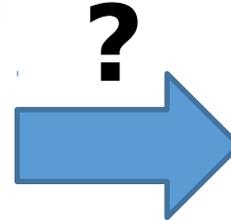
pre-miRNA Information	
pre-miRNA ID	hsa-mir-26bLinkOut: [miRBase]
Synonyms	MIRN26B, hsa-mir-26b, miR-26b, MIR26B
Description	Homo sapiens miR-26b stem-loop
Comment	The mature sequence shown here represents the most commonly cloned form from large-scale cloning
2nd Structure of pre-miRNA	
Disease	

МикроРНК (англ. microRNA, miRNA) — малые некодирующие молекулы РНК длиной 18—25 нуклеотидов (в среднем 22), принимающие участие в транскрипционной и посттранскрипционной 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

<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

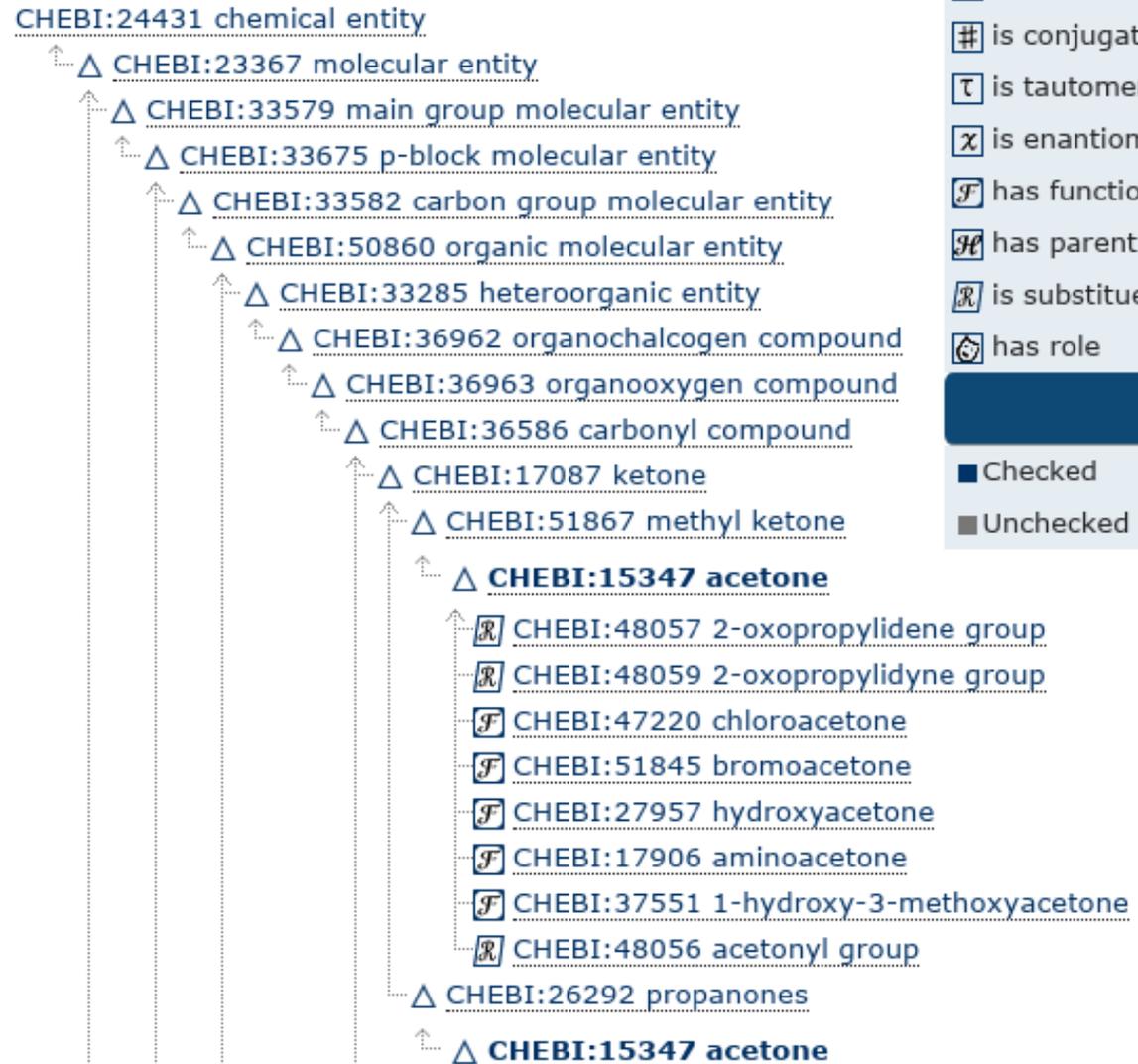
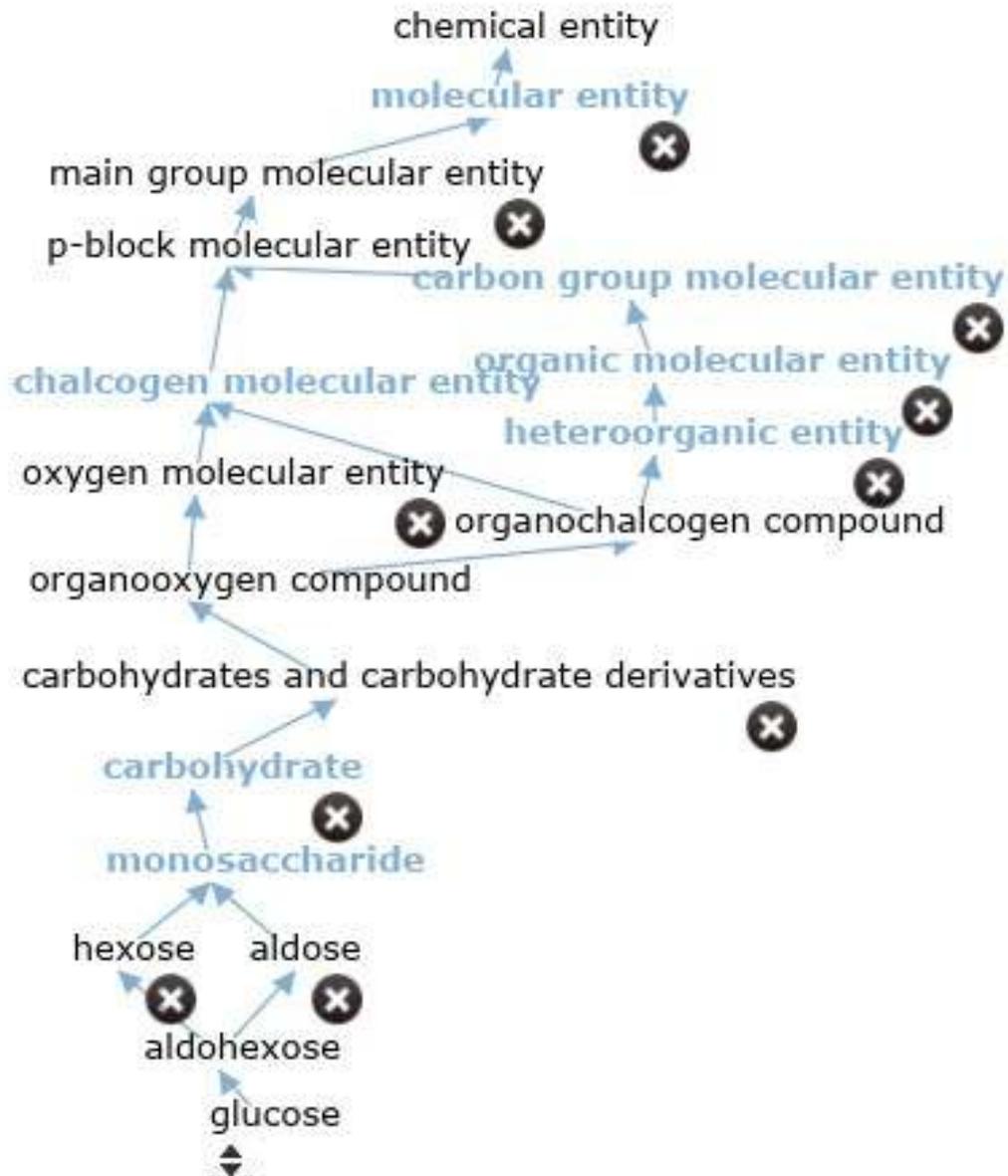
The screenshot shows the ChEBI website interface. At the top, there is a search bar with the text "CHEBI:153" and a "Search" button. Below the search bar, there are navigation links: Home, Advanced Search, Browse, Documentation, Download, Tools, About ChEBI, Preferences, and Submit. The main content area displays the ChEBI logo and the text "ChEBI:15347 - acetone". Below this, there are tabs for "Main", "ChEBI Ontology", "Automatic Xrefs", "Reactions", "Pathways", and "Models". The "Main" tab is selected, showing a chemical structure of acetone (H<sub>3</sub>C-C(=O)-CH<sub>3</sub>) and a table of information:

ChEBI Name	acetone
ChEBI ID	CHEBI:15347
Definition	A methyl ketone that consists of propane bearing an oxo group at C2.
Stars	☆☆☆ This entity has been manually annotated by the ChEBI Team.
Secondary ChEBI IDs	CHEBI:40571, CHEBI:2398, CHEBI:13708, CHEBI:22182
Supplier Information	ZINC00895111, eMolecules:474422

Below the table, there are links for "Download Molfile", "Find compounds which contain this structure", "Find compounds which resemble this structure", and "Take structure to the Advanced Search". At the bottom, there is a "Wikipedia" section with a "License" button and a paragraph of text: "Acetone (systematically named propanone) is the organic compound with the formula (CH<sub>3</sub>)<sub>2</sub>CO. It is a colorless, volatile, flammable liquid, and is the simplest ketone. Acetone is miscible with water and serves as an important solvent in its own right, typically for cleaning purposes in the laboratory. About 6.7 million tonnes were produced worldwide in 2010, mainly for use as a solvent and production of methyl methacrylate and bisphenol A. It is a common building block in organic chemistry. Familiar

# ChEBI

# Онтология



## Relationship Types

- $\Delta$  is a
- $\diamond$  has part
- $\boxed{b}$  is conjugate base of
- $\boxed{\#}$  is conjugate acid of
- $\boxed{\tau}$  is tautomer of
- $\boxed{\chi}$  is enantiomer of
- $\boxed{F}$  has functional parent
- $\boxed{H}$  has parent hydride
- $\boxed{R}$  is substituent group from
- $\boxed{\text{gear}}$  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

Надежный | <https://pubchem.ncbi.nlm.nih.gov/compound/180#section=Top>

NIH NLM National Center for Biotechnology Information

PubChem OPEN CHEMISTRY DATABASE

Search Compounds

Compound Summary for CID 180

Download Share Help

### Acetone

Cite this Record

STRUCTURE VENDORS PHARMACOLOGY LITERATURE PATENTS BIOACTIVITIES

PubChem CID:	180
Chemical Names:	Acetone; 2-propanone; Propanone; Dimethyl ketone; Methyl ketone; 67-64-1 <a href="#">More...</a>
Molecular Formula:	$C_3H_6O$ or $CH_3-CO-CH_3$ or $(CH_3)_2CO$
Molecular Weight:	58.08 g/mol
InChI Key:	CSCPPACGZOO CGX-UHFFFAOYSA-N
Substance Registry:	<a href="#">FDA UNII</a>
Safety Summary:	<a href="#">Laboratory Chemical Safety Summary (LCSS)</a>

Acetone is a colorless liquid used as a solvent and an antiseptic. It is one of the ketone bodies produced during ketoacidosis. [from MeSH](#)

Acetone is a colorless, volatile, flammable organic solvent. Acetone occurs naturally in plants, trees, forest fires, vehicle exhaust and as a breakdown product of animal fat metabolism. This agent may be normally present in very small quantities in urine and blood; larger amounts may be found in the urine and blood of diabetics. Acetone is toxic in high doses. (NCI04) [Pharmacology from NCI](#)

Acetone is one of the ketone bodies produced during ketoacidosis. Acetone is not regarded as a waste product of metabolism. However, its physiological role in biochemical machinery is not clear. A model for the role of acetone metabolism is presented that orders the events occurring in acetonemia in sequence: in diabetic ketosis or starvation, ketone body production (b-hydroxy-butyrate, acetoacetate) provides fuel for vital organs (heart, brain . .) raising the chance of survival of the metabolic catastrophe. However, when ketone body production exceeds the degrading capacity, the accumulating acetoacetic acid presents a new challenge to the pH regulatory system. Acetone production and its further degradation to C3 fragments fulfill two purposes: the maintenance of pH buffering capacity and provision of fuel for peripheral tissues. Since ketosis develops under serious metabolic circumstances, all the mechanisms that balance or moderate the effects of ketosis enhance the chance for survival. From this point of view, the theory that transportable C3 fragments can serve as

# 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

www.hmdb.ca/metabolites/HMDB01659

HMDB

TMIC The Metabolomics Innovation Centre | Specializing in ready to use metabolomics kits.

Showing metabocard for Acetone (HMDB01659)

Identification Taxonomy Ontology Physical properties Spectra Biological properties Concentrations Links References XML

Show Metabolites with Similar Structures

Record Information	
Version	3.6
Creation Date	2005-11-20 22:13:15 UTC
Update Date	2017-03-02 21:26:41 UTC
HMDB ID	HMDB01659
Secondary Accession Numbers	None

Metabolite Identification	
Common Name	Acetone
Description	Acetone is one of the ketone bodies produced during ketoacidosis. Acetone is not regarded as a waste product of metabolism. However, its physiological role in biochemical machinery is not clear. A model for the role of acetone metabolism is presented that orders the events occurring in acetonemia in sequence: in diabetic ketosis or starvation, ketone body production (b-hydroxy-butyrate, acetoacetate) provides fuel for vital organs (heart, brain .) raising the chance of survival of the metabolic catastrophe. However, when ketone body production exceeds the degrading capacity, the accumulating acetoacetic acid presents a new challenge to the pH regulatory system. Acetone production and its further degradation to C3 fragments fulfill two purposes: the maintenance of pH buffering capacity and provision of fuel for peripheral tissues. Since ketosis develops under serious metabolic circumstances, all the mechanisms that balance or moderate the effects of ketosis enhance the chance for survival. From this point of view, the theory that transportable C3 fragments can serve as additional nutrients is a novel view of acetone metabolism which introduces a new approach to the study of acetone degradation, especially in understanding its physiological function and the interrelationship between liver and peripheral tissues. (PMID 10580530). Acetone is typically derived from acetoacetate through the action of microbial acetoacetate

Базы данных, содержащие информацию  
по заболеваниям.

OMIM<sup>®</sup>

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:125853)

\* - 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

# 125853 ICD+

**DIABETES MELLITUS, NONINSULIN-DEPENDENT; NIDDM**

*Alternative titles; symbols*

**DIABETES MELLITUS, TYPE II; T2D**  
**NONINSULIN-DEPENDENT DIABETES MELLITUS**  
**MATURITY-ONSET DIABETES**

Other entities represented in this entry:

**INSULIN RESISTANCE, SUSCEPTIBILITY TO, INCLUDED**  
**DIABETES MELLITUS, TYPE 2, PROTECTION AGAINST, INCLUDED**

**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 (regex) или подстановочные знаки. Пример : `\w[A-Z]\d\w\.\w{2} \w{4}\.\w{2}`

The screenshot shows the ICD-11 for Mortality and Morbidity Statistics website. The search bar contains "Diabetes mellitus". The left sidebar shows a tree view of the classification, with "Diabetes mellitus" selected. The right sidebar shows details for "Diabetes mellitus", including its parent category "Endocrine diseases", a description "A metabolic disorder with hetero both.", and a list of "Coded Elsewhere" entries: "Diabetes mellitus in pregnancy" and "Neonatal diabetes mellitus".

**ICD-11 for Mortality and Morbidity Statistics** (December 2018)

Search  [ Advanced Search ] [Browse](#)

Foundation Id : <http://id.who.int/icd/en>

**Diabetes mellitus**

**Parent**

Endocrine diseases

**Description**

A metabolic disorder with hetero both.

**Coded Elsewhere**

- Diabetes mellitus in pregnancy
- Neonatal diabetes mellitus

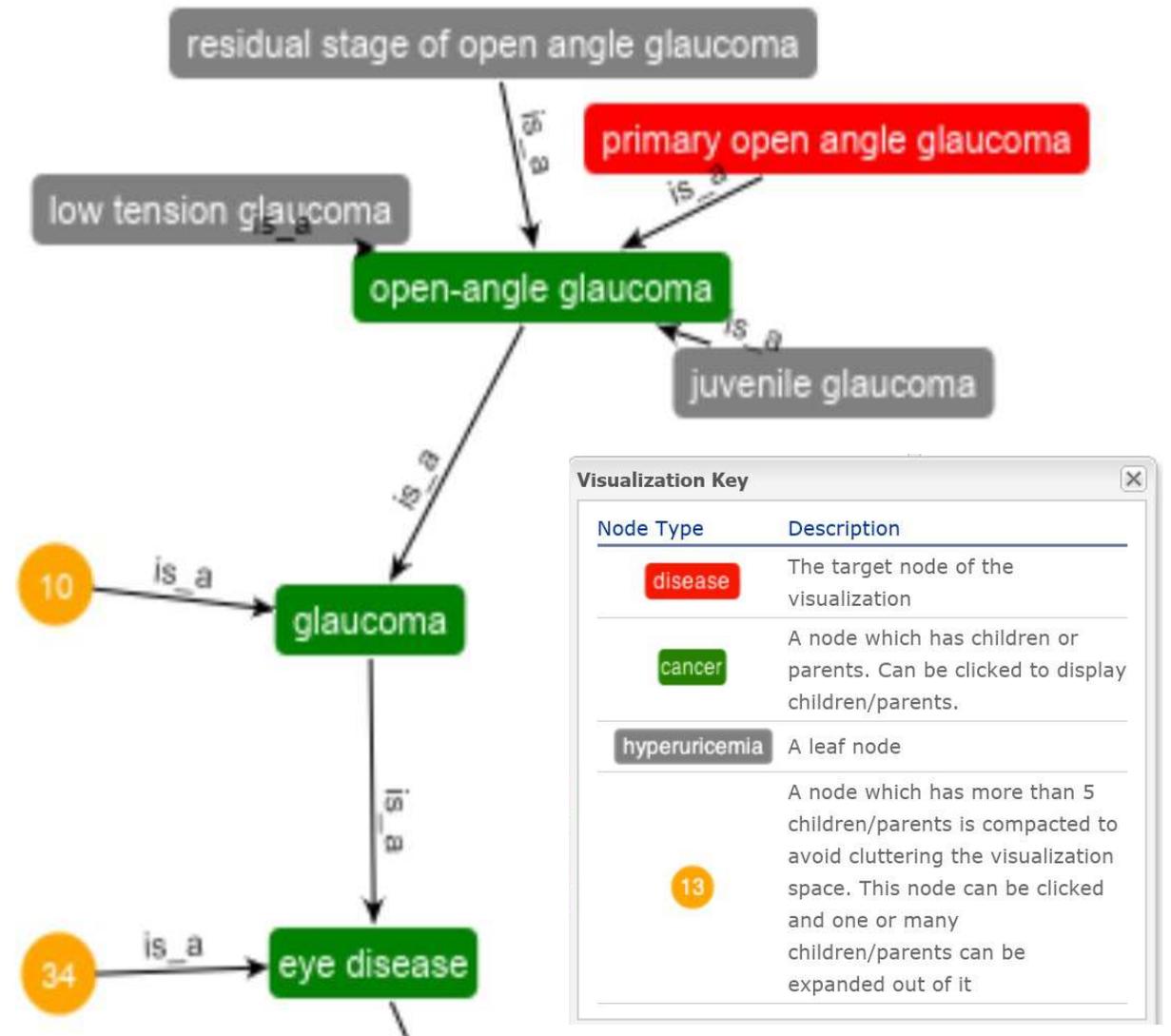
**ICD-11 - Mortality and Morbidity Statistics**

- ▶ 01 Certain infectious or parasitic diseases
- ▶ 02 Neoplasms
- ▶ 03 Diseases of the blood or blood-forming organs
- ▶ 04 Diseases of the immune system
- ▼ 05 Endocrine, nutritional or metabolic diseases
  - ▼ Endocrine diseases
    - ▶ Disorders of the thyroid gland or thyroid hormones system
    - ▼ Diabetes mellitus
      - 5A10 Type 1 diabetes mellitus
      - 5A11 Type 2 diabetes mellitus
      - 5A12 Malnutrition-related diabetes mellitus
      - ▶ 5A13 Diabetes mellitus, other specified type
      - 5A14 Diabetes mellitus, type unspecified
      - ▶ Acute complications of diabetes mellitus
    - ▼ JA63 Diabetes mellitus in pregnancy
      - JA63.0 Pre-existing type 1 diabetes mellitus in pregnancy
      - JA63.1 Pre-existing type 2 diabetes mellitus in pregnancy
      - JA63.2 Diabetes mellitus arising in pregnancy
      - JA63.Y Other specified diabetes mellitus in pregnancy
      - JA63.Z Diabetes mellitus in pregnancy, unspecified
    - ▶ KB60.2 Neonatal diabetes mellitus

# 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).

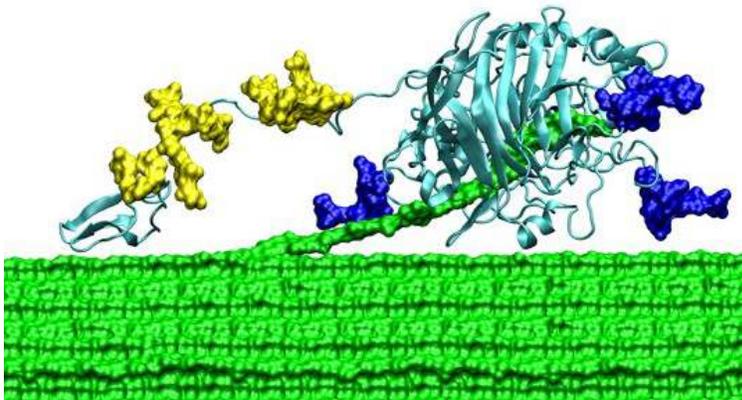




# Что такое биологический процесс?

## функция

Биохимическая  
активность



Molecular function  
(MF)

Биологическая  
цель



Biological process  
(BP)

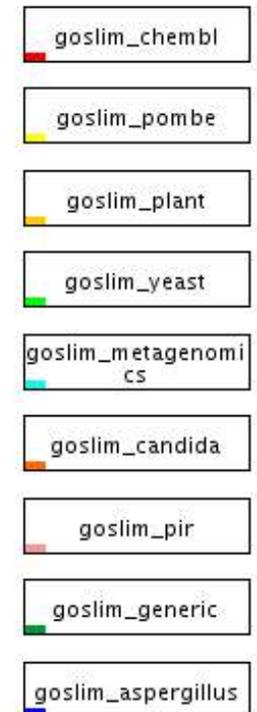
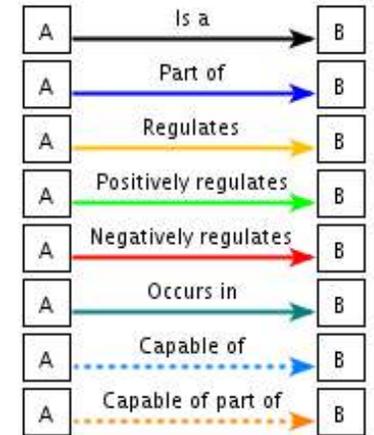
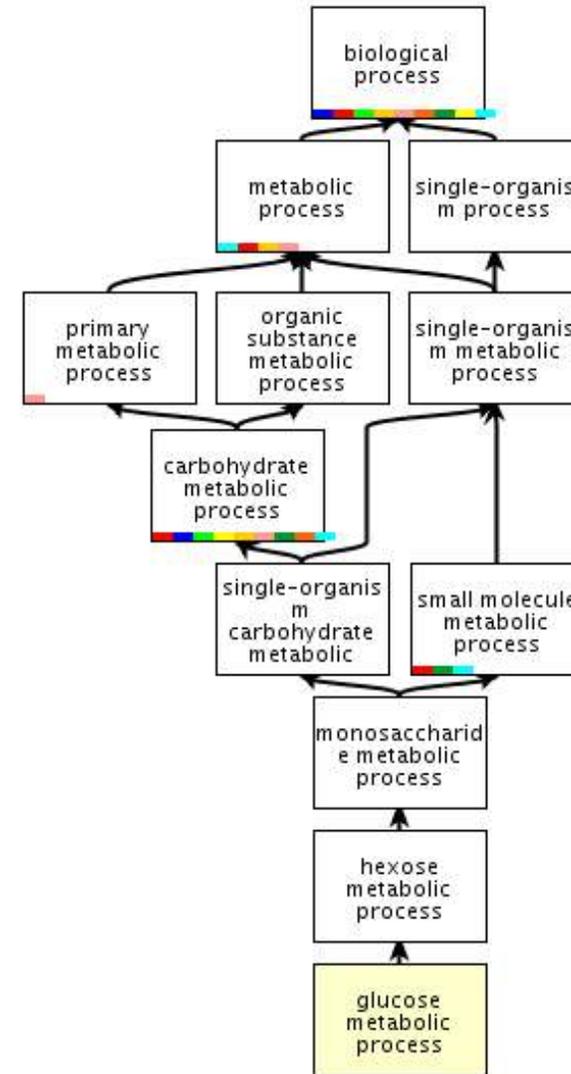
Структура



Cellular component  
(CC)

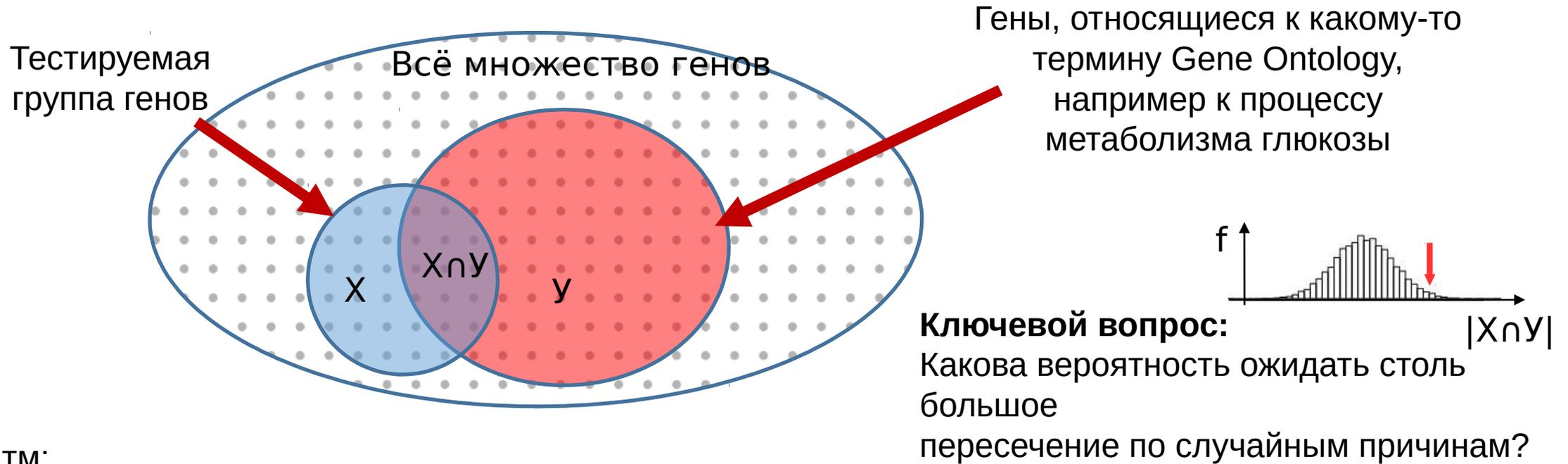
# Gene Ontology

- <http://www.geneontology.org/>
- Является онтологией, описывающей биологические процессы (BP), клеточные компоненты (CC) и молекулярные функции (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.



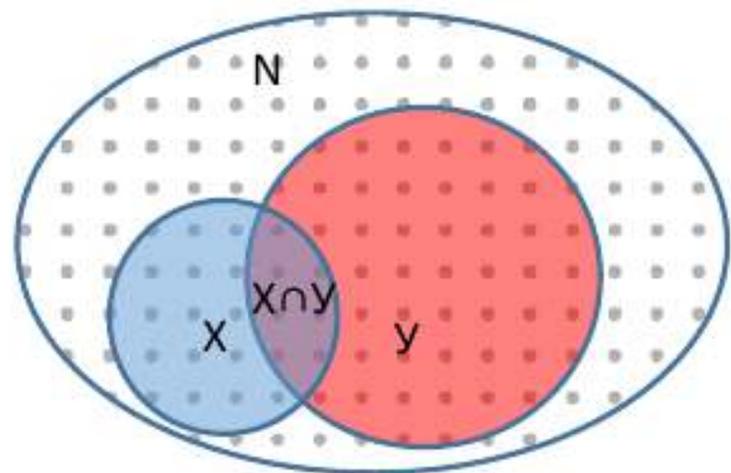
Алгоритм:

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  $X \cap Y$ , 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 сопряженности (2x2)

	Число генов в наборе X	Число генов не в наборе X	Полное число генов
Число генов в наборе Y	$ X \cap Y  = 5$	$ Y  -  X \cap Y  = 1$	$ Y  = 6$
Число генов не в наборе Y	$ X  -  X \cap Y  = 1$	$N -  X \cup Y  = 23500 - 6 - 6 + 5 = 23493$	$N -  Y  = 23500 - 6 = 23494$
Полное число генов	$ X  = 6$	$N -  X  = 23500 - 6 = 23494$	$N = 23500$



Что ожидается по случайным причинам?

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

Ген А, ген Б, ... ген В

...

Ген В, ген А, ... ген В

$$\rightarrow |X|!$$

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

Полное число вариантов  $X_{\text{случайный}}$

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

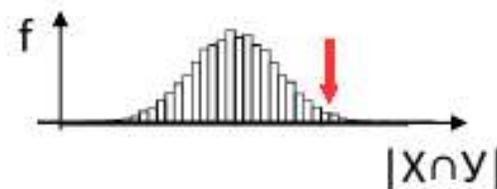
$$C_{|Y|}^{|X \cap Y|}$$

Сколькими способами можно заполнить оставшиеся места в  $X_{\text{случайный}}$  генами не из  $Y_{\text{случайный}}$

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

Какова ожидаемая частота (вероятность)?

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



$$p\text{-значение} = \frac{C_6^5 \cdot C_{23494}^1}{C_{23500}^6} + \frac{C_6^6 \cdot C_{23494}^0}{C_{23500}^6} = 6,03 \cdot 10^{-19}$$

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

**KEGG** PATHWAY: hsa00010 Help

<b>Entry</b>	hsa00010	Pathway
<b>Name</b>	Glycolysis / Gluconeogenesis - Homo sapiens (human)	
<b>Description</b>	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 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 glyceraldehyde-3P 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].	
<b>Class</b>	Metabolism; Carbohydrate metabolism <a href="#">BRITE hierarchy</a>	
<b>Pathway map</b>	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

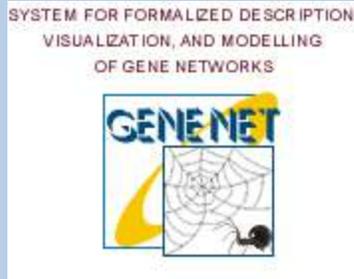
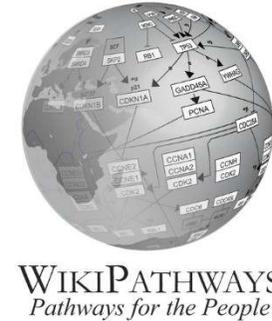
The screenshot shows the Reactome PathwayBrowser interface. On the left is a navigation pane with a tree view of biological processes, including "Metabolism of carbohydrates" and "Glucose metabolism". The main area displays a complex metabolic pathway diagram with various molecules represented by colored nodes and arrows indicating reaction directions. Several sub-pathways are highlighted with colored boxes and labels: "Digestion of dietary carbohydrate", "Galactose catabolism", "Fructose metabolism", "Pentose phosphate pathway (fructose monophosphate shunt)", "Glucose metabolism", "5-Phosphoribose 1-diphosphate biosynthesis", and "Catabolism of glucuronate to xylulose-5-phosphate". At the bottom, there is a summary section with a "Stable Identifier" (R-HSA-71387.2) and a "Summation" paragraph: "These pathways together are responsible for: 1) the extraction of energy and carbon skeletons for biosyntheses from dietary sugars and related molecules; 2) the short-term storage of glucose in the body (as glycogen) and its mobilization and synthesis of glucose from pyruvate during extended fasts." On the right side of the screenshot, there is a text box with the following text: "breakdown is a major source and conversion to glucose yield pyruvate, glycogen can be synthesized from p" and "inner mitochondrial mem".

# План лекции

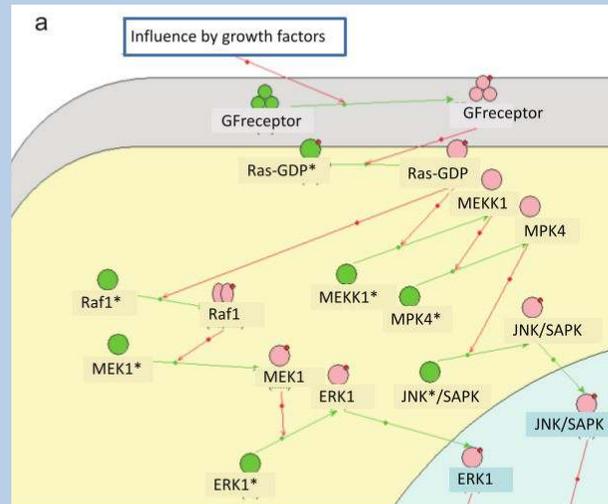
1. Введение.
2. Обзор Интернет-ресурсов.
3. Ресурсы, интегрирующие биологическую информацию из разнородных источников и представляющие ее в виде генных сетей: ANDSystem, STRING, GeneMania, Pathway Commons.
4. Практическое применение инструментов интеграции.

# 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.

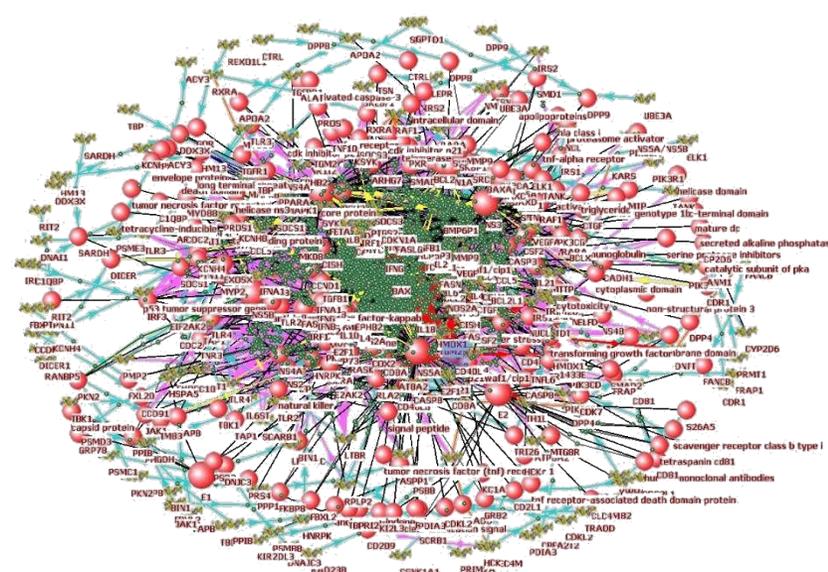
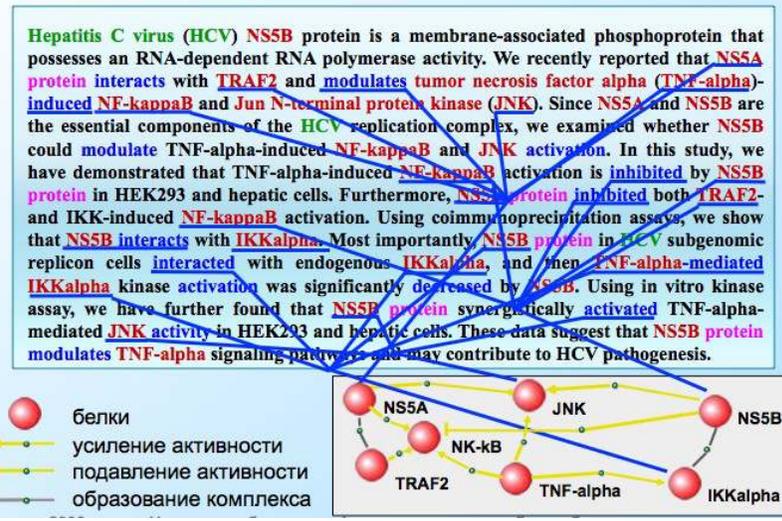
В ИЦиГ СО РАН разработана компьютерная технология автоматического извлечения знаний из текстов научных публикаций и международных патентов ANDSystem. Проведен автоматический анализ более 25 млн. научных публикаций и 10 млн. международных патентов.

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.





# STRING

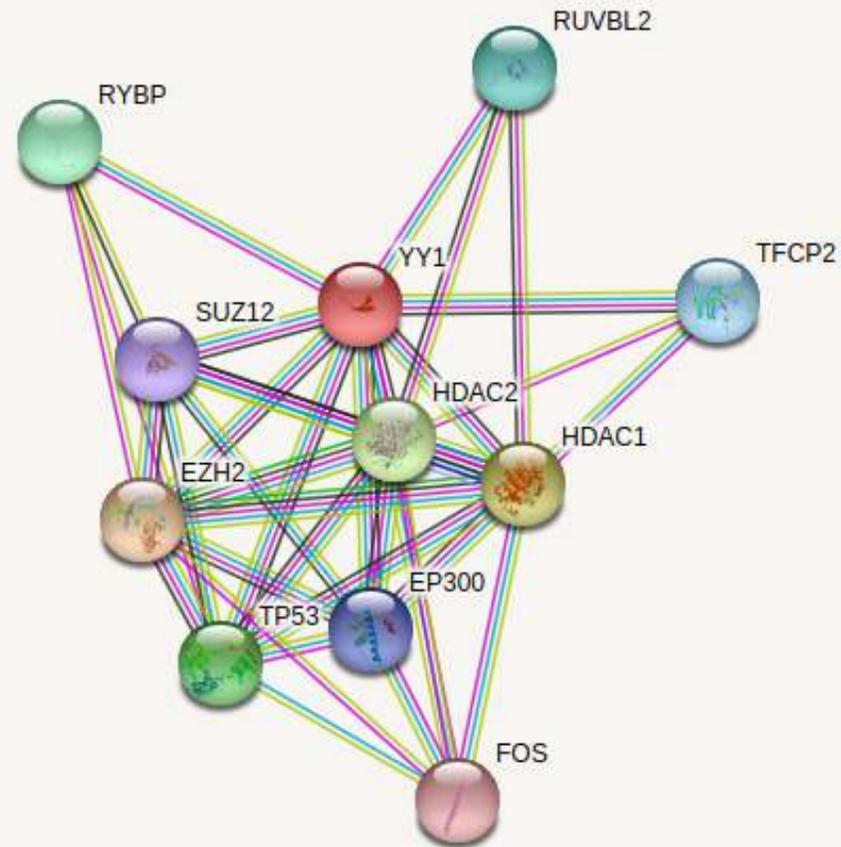


[Search](#)

[Download](#)

[Help](#)

[My Data](#)



# 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

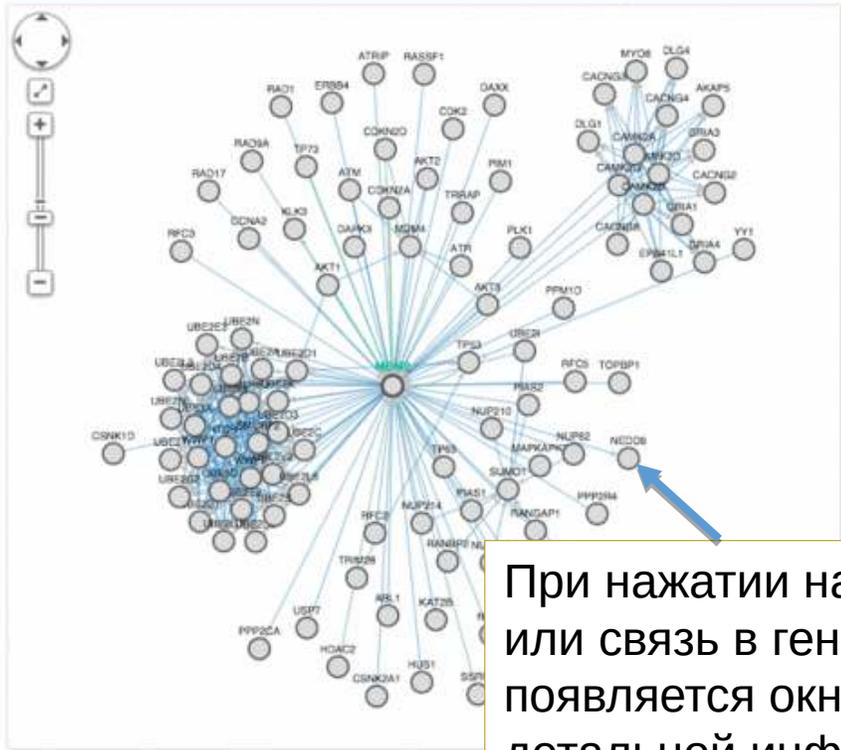
Добавление в сеть данных по изменениям, наблюдающимся при раке

Настройки для фильтрации сети по связям и вершинам

# PCViz Pathway Commons Network Visualizer

Genes of interest

MDM2 +



Details Settings Context

Click on one of the interactions or genes in the network to details...

Details Settings Context

Interaction types

- 362 controls state change
- 409 controls expression

Number of genes (468)

Slide left to decrease the number of genes

Query type

Neighborhood

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.

CDT1

The protein encoded by this gene is involved in the formation of the pre-replication complex that is necessary for DNA replication. The encoded protein can bind geminin, which prevents replication

Aliases: DUP, RIS2

Description: chromatin licensing and DNA replication factor 1

Chromosome Location: 16q24.3

UniProt ID: Q9H211

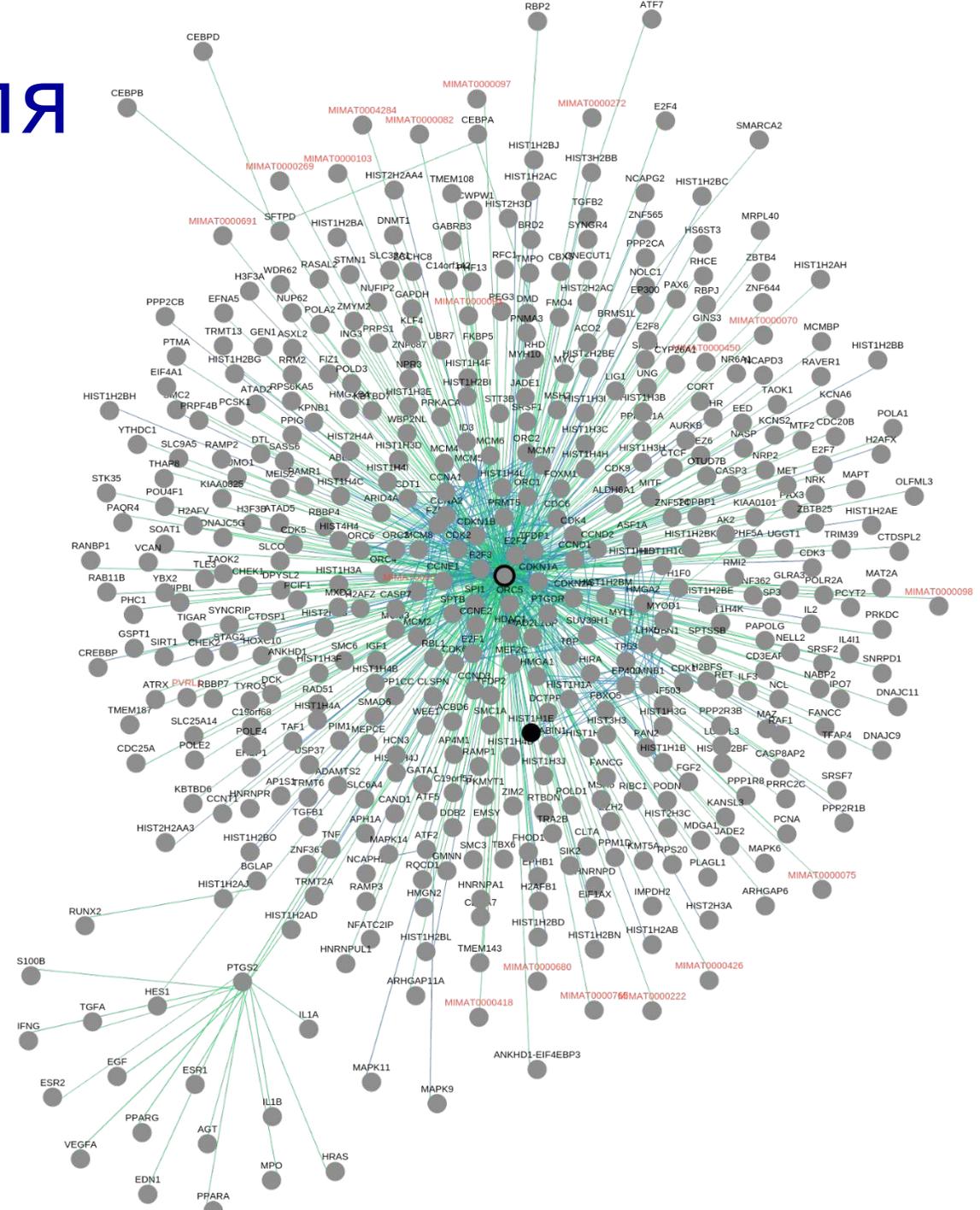
Gene ID: 81620

Добавление гена в список интересующих вас генов

# Пример генной сети для гена ретинобластомы RB1

Пример загрузки генной сети в формате SIF

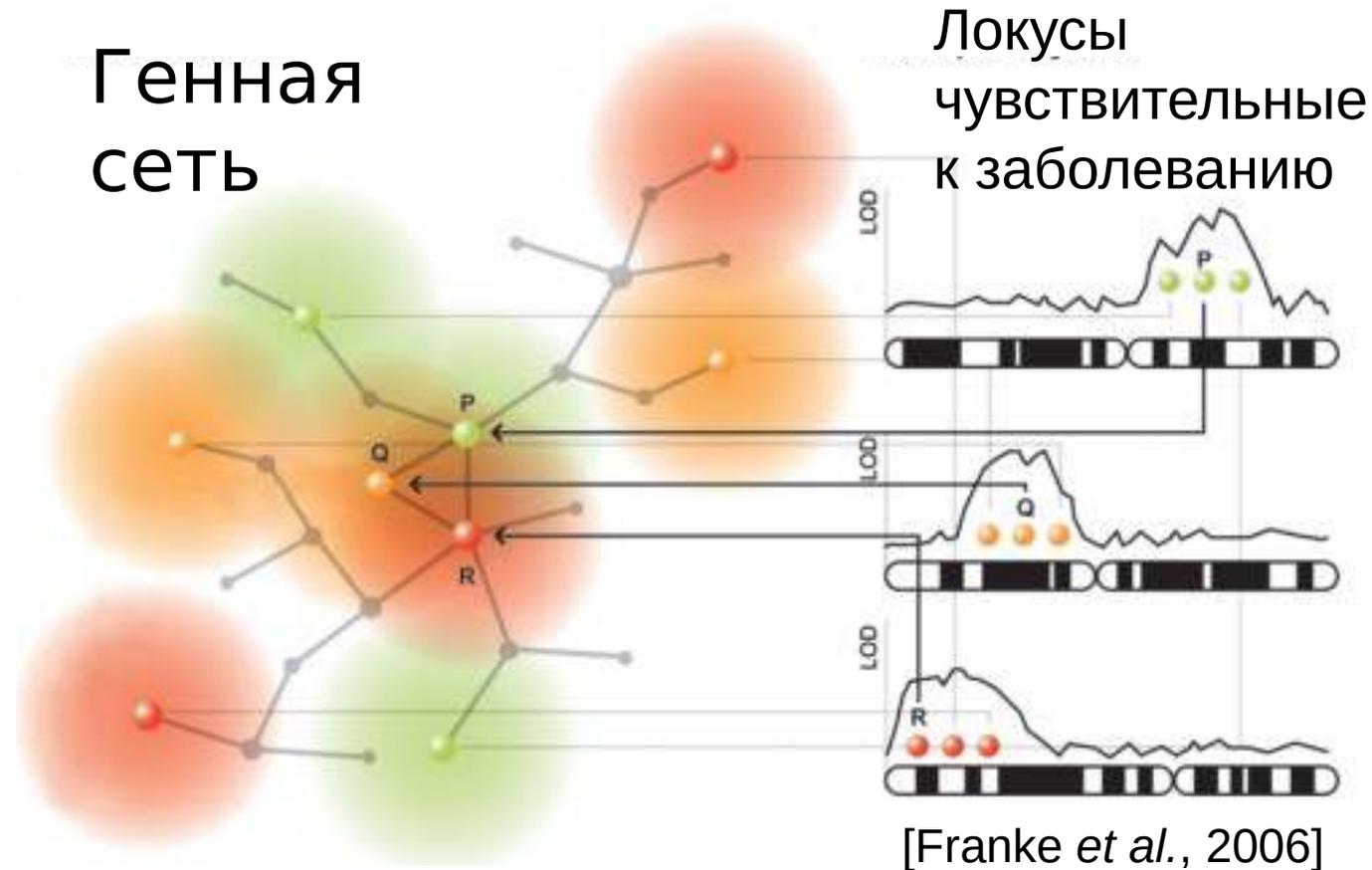
- AATF in-complex-with E2F1
- AATF in-complex-with E2F2
- AATF in-complex-with E2F3
- AATF in-complex-with RB1
- AATF interacts-with RB1
- AATF in-complex-with TFDP1
- ABCD3 interacts-with DYRK1A
- ABCD3 interacts-with DYRK1B
- ABCD3 interacts-with RB1
- ABHD10 interacts-with FOXK1



# 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



<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

OPEN

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

Juthatip Manisorn, Supaporn Khamchun, Arada Vinaiphath & VisithThongboonkerd

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  $\alpha$ -tubulin, which was decreased, as one of central nodes of protein-protein interactions. Overexpression of  $\alpha$ -tubulin (pcDNA6.2-TUBA1A) was then performed and its efficacy was confirmed. pcDNA6.2-TUBA1A could maintain levels of  $\alpha$ -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  $\alpha$ -enolase) on apical membrane were dramatically reduced to basal levels by pcDNA6.2-TUBA1A. These findings implicate that  $\alpha$ -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<sup>1</sup>. This type of the stones can be originated from supersaturation of calcium and oxalate ions, leading to crystallization inside renal tubular fluid or urine<sup>2</sup>. CaOx crystals can then nucleate to form "stone nidus" and adhere directly onto apical surface of renal tubular epithelial cells<sup>3-5</sup>. 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<sup>6,7</sup>. CaOx crystals also evoke inflammatory processes that can lead to fibrosis, loss of nephron and eventually ESRD<sup>8</sup>.

Even with the aforementioned knowledge, molecular mechanisms of the downstream cellular response remain largely unknown. From our previous expression proteomics study<sup>9</sup>, 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)

RESEARCH

Open Access



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

Jen-hwa Chu<sup>1</sup>, Jaime E. Hart<sup>2,3</sup>, Divya Chhabra<sup>2</sup>, Eric Garshick<sup>2,4</sup>, Benjamin A. Raby<sup>2,5</sup> and Francine Laden<sup>2,3,6</sup>

### 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<sub>2.5</sub> (particulate matter  $\leq 2.5$  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<sub>2.5</sub>, 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  
<sup>1</sup>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



© The Author(s). 2016 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.



OPEN ACCESS

**Citation:** Şenbabaoğlu Y, Sümer SO, Sánchez-Vega F, Bemis D, Criello 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 Meiring, University of Zurich and Swiss Institute of Bioinformatics, SWITZERLAND

**Received:** June 7, 2015

**Accepted:** January 20, 2016

**Published:** February 28, 2016

**Copyright:** © 2016 Şenbabaoğlu et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The pan-cancer RPPA dataset is available from The Cancer Proteome Atlas at [http://agp1.bioinformatics.mskcc.org/pan\\_cancer\\_rppa/download.html](http://agp1.bioinformatics.mskcc.org/pan_cancer_rppa/download.html).

**Funding:** CS and YS were supported by the National Institute of General Medical Sciences (P41GM103504) and the National Cancer Institute (L24CA14384). 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.

RESEARCH ARTICLE

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

Yasin Şenbabaoğlu<sup>1\*</sup>, Selçuk Onur Sümer<sup>1</sup>, Francisco Sánchez-Vega<sup>1</sup>, Debra Bemis<sup>1</sup>, Giovanni Criello<sup>1\*</sup>, Nikolaus Schultz<sup>2</sup>, Chris Sander<sup>1\*</sup>

**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 (YS); chris@sanderlab.org (CS)

### 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.

## 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
---	---	------	---	---	---	---	------	---	---	------	------	---	---

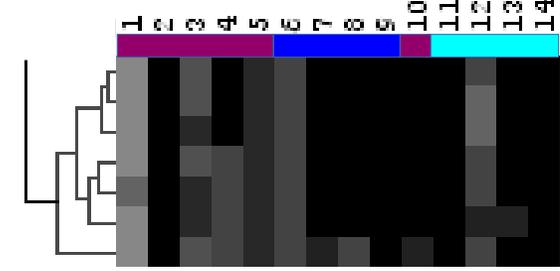
Манхэттенское расстояние между векторами:  $D = \sum |A_i - B_i|$

Доля испытуемых →  
у которых присутствует  
белок



← Номера недель, уменьшение  
концентрации натрия в пище с 12 г/сут до  
6 г/сут

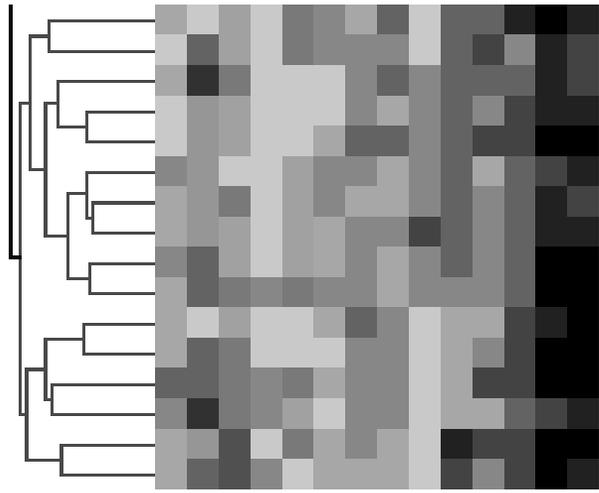
кластер 9



UBP15\_HUMAN  
LY75\_HUMAN  
HTRBE2\_HUMAN  
AKA11\_HUMAN  
MYO5C\_HUMAN  
MYO5A\_HUMAN  
MYO6\_HUMAN



кластер 83



CSPG4\_HUMAN  
MEN1\_HUMAN  
COL1A1\_HUMAN  
CD14\_HUMAN  
MGA\_HUMAN  
GELS\_HUMAN  
TETN\_HUMAN  
CYTM\_HUMAN  
CATD\_HUMAN  
SDCB1\_HUMAN  
CEL\_HUMAN  
CSF1\_HUMAN  
CADM1\_HUMAN  
LAIR1\_HUMAN  
DPP4\_HUMAN  
ACTB\_HUMAN

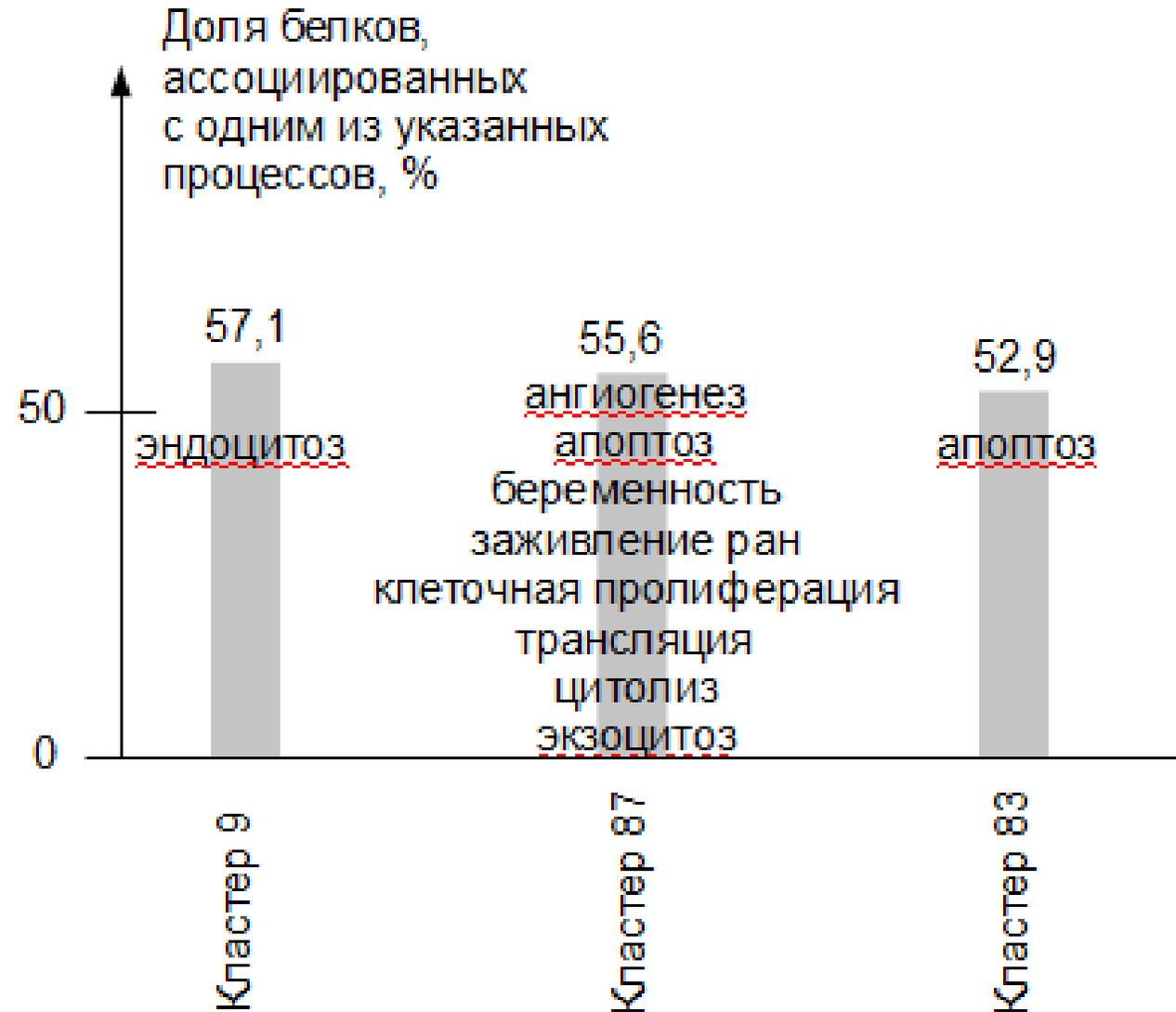
кластер 87



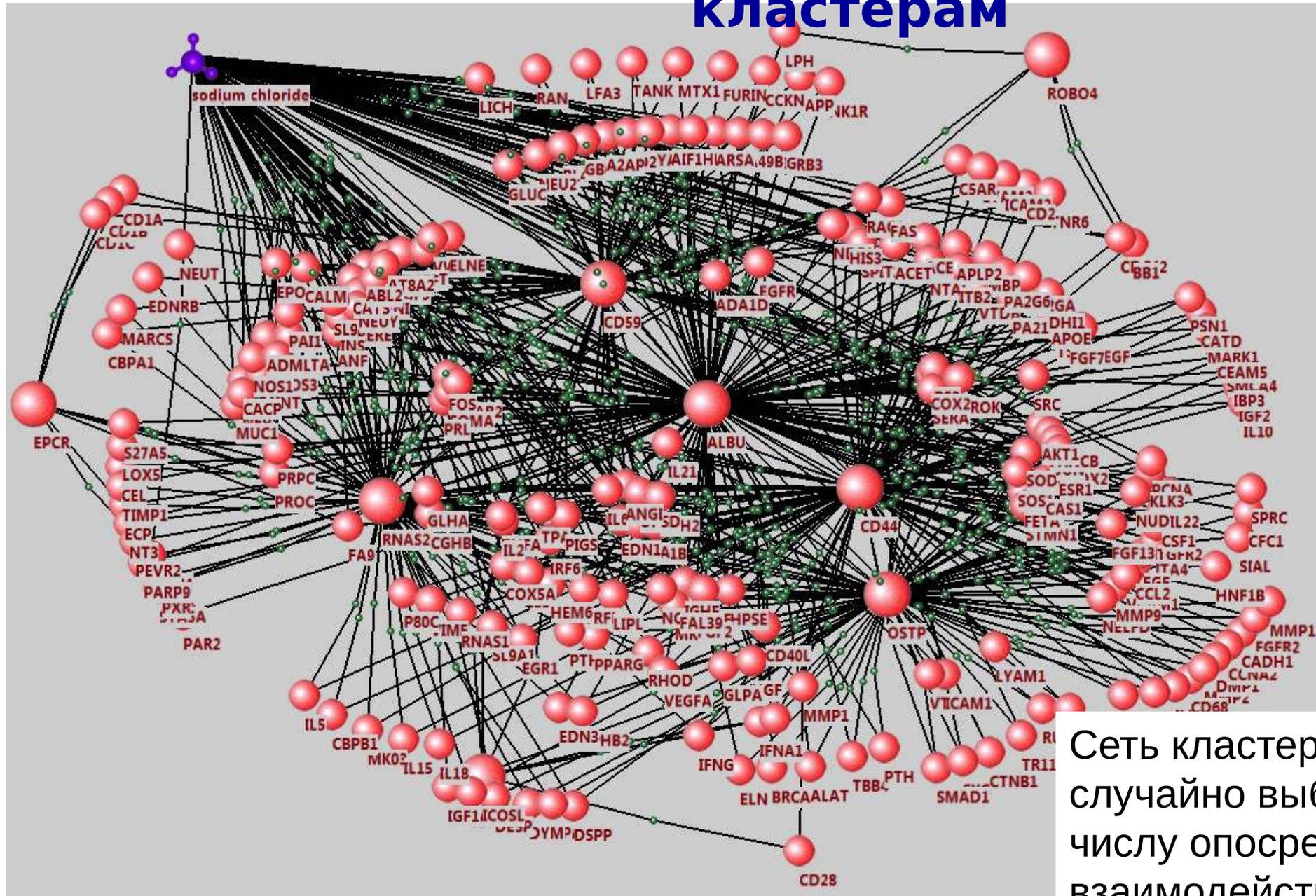
CD59\_HUMAN  
EPCR\_HUMAN  
ICOSL\_HUMAN  
RHAS2\_HUMAN  
CD44\_HUMAN  
YIPF3\_HUMAN  
ALBU\_HUMAN  
OSTP\_HUMAN  
ROBO4\_HUMAN

Иерархическое  
дерево было разбито  
на кластеры  
включающие не  
более 30 элементов

# Поиск кластеров ассоциированных с



# Построение сетей по найденным кластерам

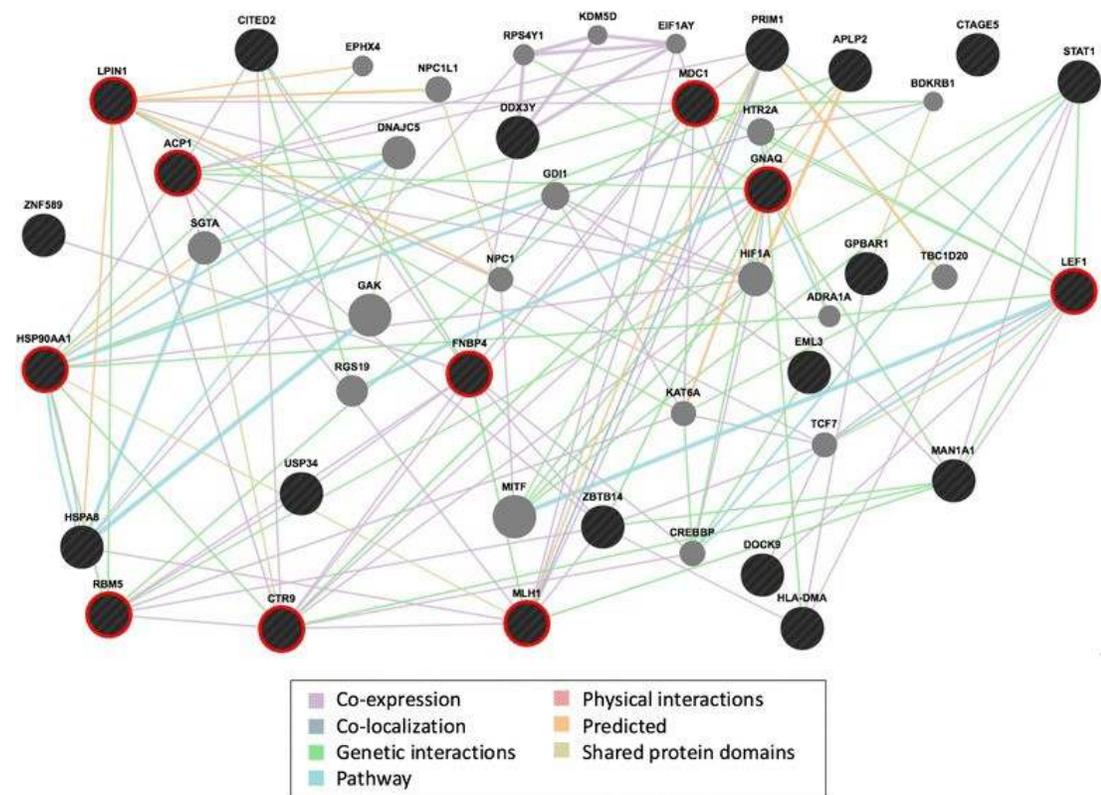


Сеть кластера 87 отличается от случайно выбранных сетей по числу опосредованных взаимодействий ( $p=0.0175$ )



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

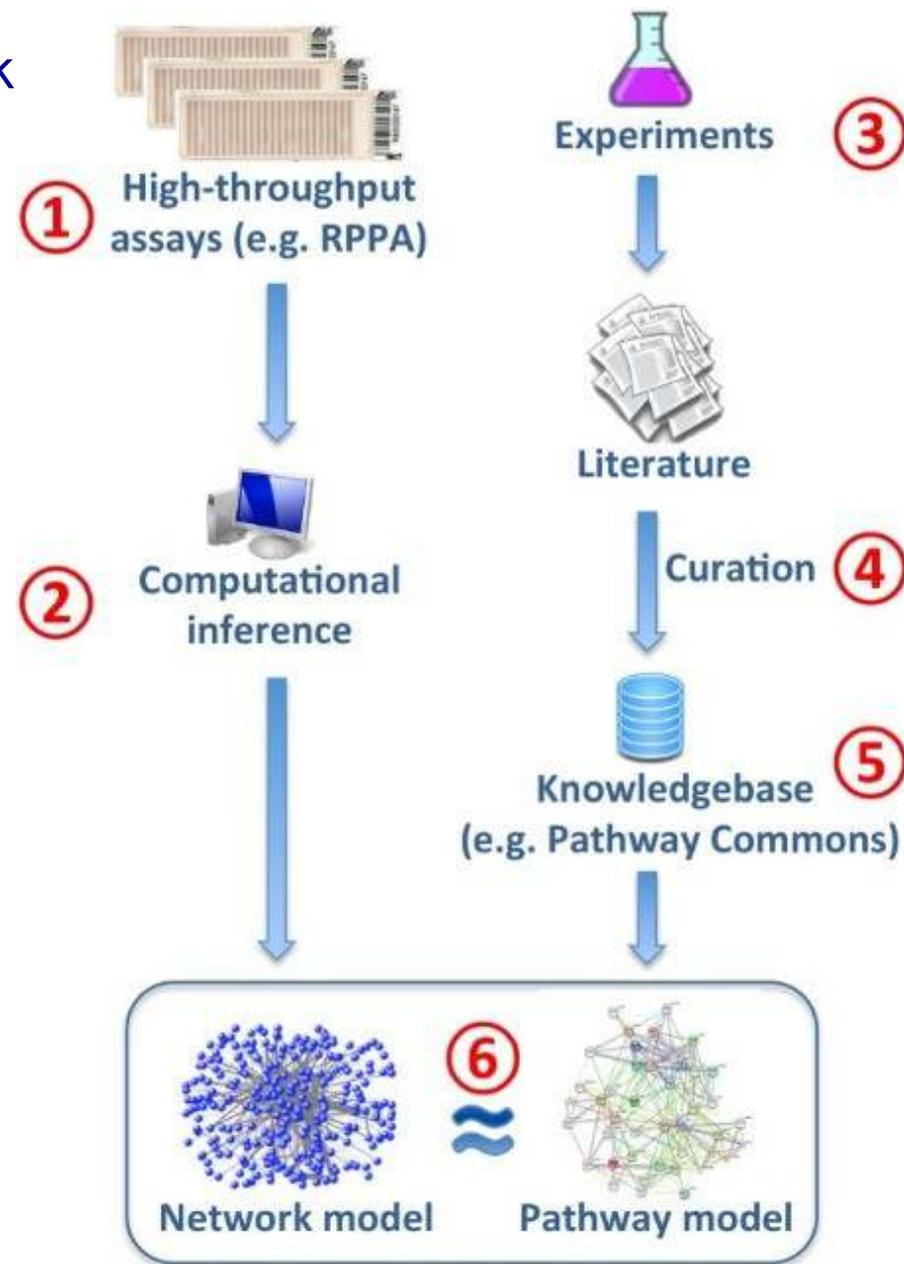
Воздействие загрязненного воздуха связано с неблагоприятными последствиями для здоровья, включая увеличение сердечно-легочной заболеваемости и смертности, а также увеличение риска рака легких. Был проведен полно-геномный анализ уровней экспрессии РНК крови 63 некурящих сотрудников парковочных терминалов 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.

## 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.



# 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.