Практикум 6

1. Описание входных данных

В качестве входных данных имеется текстовый файл со списком из 7 генов человека.

ALDH1A1 — альдегиддегидрогеназа семейства 1A1, фермент, участвующий в окислении альдегидов до карбоновых кислот. Играет важную роль в метаболизме ретиноевой кислоты, детоксикации токсичных альдегидов и защите клеток от окислительного стресса.

TKFC — триокиназа и FMN-циклаза, фермент углеводного обмена, участвующий в метаболизме фруктозы и глицеральдегида. Обеспечивает фосфорилирование промежуточных продуктов катаболизма фруктозы.

GLYCTK — глицераткиназа, фермент, связанный с метаболизмом глиоксилата и фруктозы. Катализирует образование фосфоглицерата и участвует в энергетическом обмене клетки.

AKR1B1 — альдокеторедуктаза семейства 1B1, фермент полиолового пути, восстанавливающий глюкозу до сорбитола. Связан с развитием диабетических осложнений и клеточным ответом на окислительный стресс.

SORD — сорбитолдегидрогеназа, фермент полиолового пути, превращающий сорбитол во фруктозу. Играет ключевую роль в метаболизме углеводов и поддержании клеточного энергетического баланса.

ALDOB — альдолаза B, фермент гликолиза и фруктозного обмена, обеспечивающий расщепление фруктозо-1-фосфата. Наиболее активно экспрессируется в печени, почках и тонком кишечнике.

KHK — кето-гексокиназа (фруктокиназа), фермент, катализирующий первый этап метаболизма фруктозы — образование фруктозо-1-фосфата. Имеет важное значение для утилизации фруктозы в печени.

Думаю, большая часть этих генов функционально связана между собой через процессы метаболизма фруктозы, полиолового пути и общего углеводного обмена. Также часть белков ассоциирована с нарушениями обмена веществ, диабетом и развитием метаболических заболеваний.

2. Групповой анализ (STRING)

Описание функционала сервиса

С помощью сервиса STRING можно выполнять несколько типов анализа:

- исследование белок-белковых взаимодействий (PPI)
- анализ коэкспрессии генов из заданного списка
- поиск обогащения по GO-терминам и метаболическим путям KEGG

Для оценки статистической значимости STRING использует статистические критерии, включая тест Фишера и критерий Колмогорова—Смирнова. Кроме того, сервис автоматически проводит поправку на множественное тестирование. Это отражается в параметре **False Discovery Rate (FDR)**, который

показывает ожидаемую долю ложноположительных результатов среди статистически значимых находок.

Запрос в STRING

Для анализа функционального обогащения был использован сервис STRING. В анализ были внесены следующие белки человека: **ALDH1A1**, **TKFC**, **GLYCTK**, **AKR1B1**, **SORD**, **ALDOB**, **KHK**. Анализ проводился для организма *Homo sapiens*. В ответ на мой запрос программа построила граф взаимодействий для всех 7 белков *(7 вершин и 11 рёбер)*.

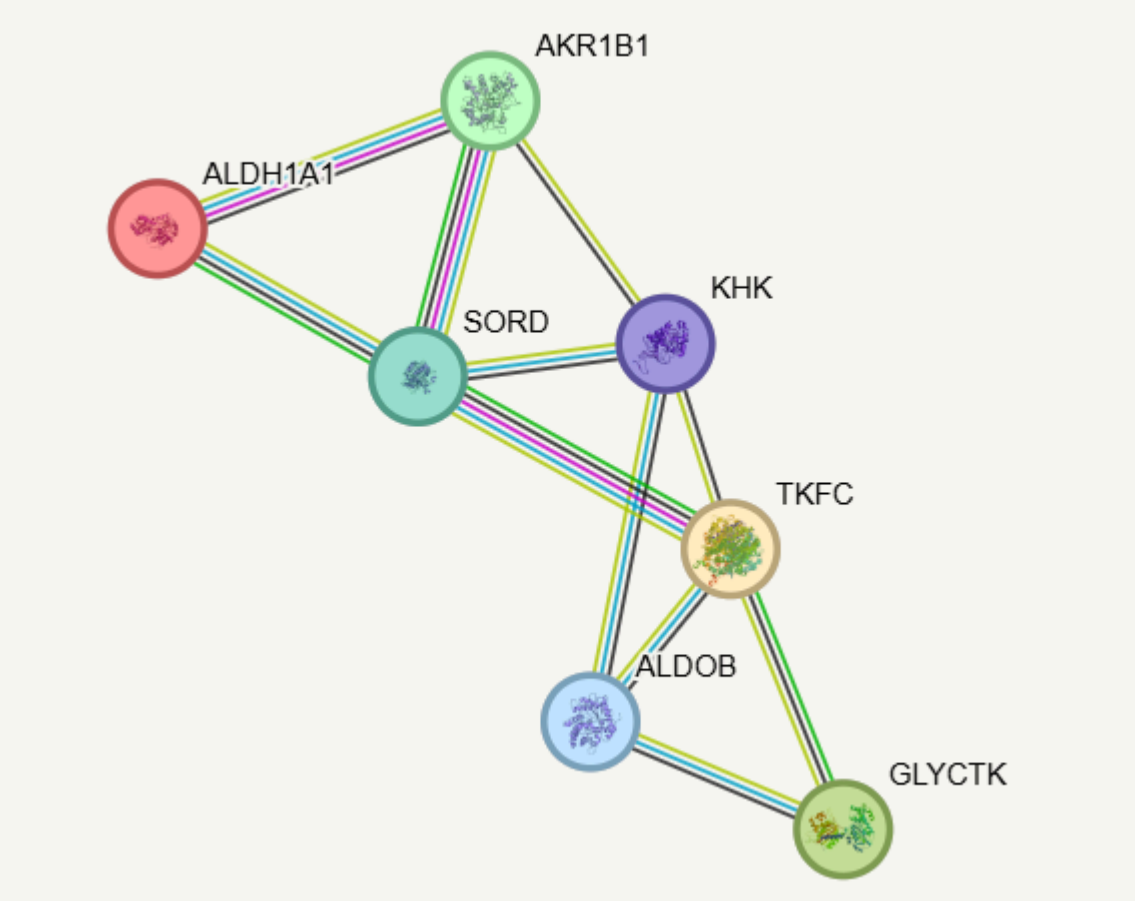


Рисунок 1. Визуализация взаимодействий внутри выборки белков.

На рисунке представлена сеть белок-белковых взаимодействий, построенная на основе данных STRING. Узлы соответствуют белкам, а линии между ними отражают наличие функциональных или экспериментально подтверждённых связей. Большое количество соединений между белками указывает на функциональную связанность исследуемого набора. Наиболее выраженные взаимодействия наблюдаются между белками, участвующими в метаболизме фруктозы, маннозы и полиоловом пути. Разноцветные линии обозначают различные типы взаимодействий, включая данные экспериментальных исследований, сведения из баз данных, коэкспрессию, совместную встречаемость в геномах и публикациях.

В разделе статистики указано, что общий локальный коэффициент кластеризации составил 0.69, то есть присутствует около 69% от максимально возможного количества локальных связей между белками. Значение PPI (protein–protein interactions) enrichment p-value = 3.33e−16, что свидетельствует о статистически значимо большем числе взаимодействий, чем ожидалось бы случайно. Следовательно, можно предполагать, что данные белки функционально связаны и участвуют в общих метаболических путях.

GO-анализ

Были проанализированы категории Gene Ontology (GO) и KEGG pathways. Статистическая значимость оценивалась с использованием FDR — p-value с поправкой на множественное тестирование методом Benjamini–Hochberg. Анализ позволил определить основные биологические процессы и метаболические пути, характерные для данного набора белков.

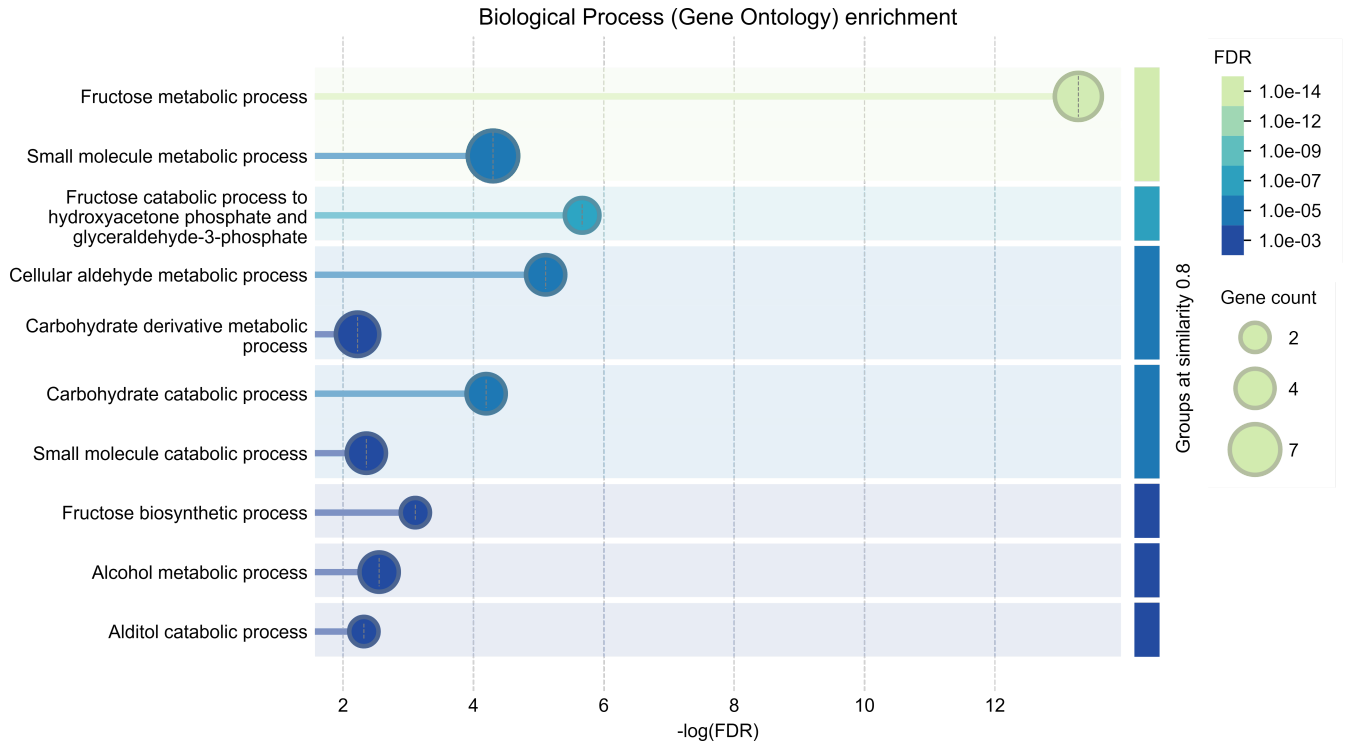


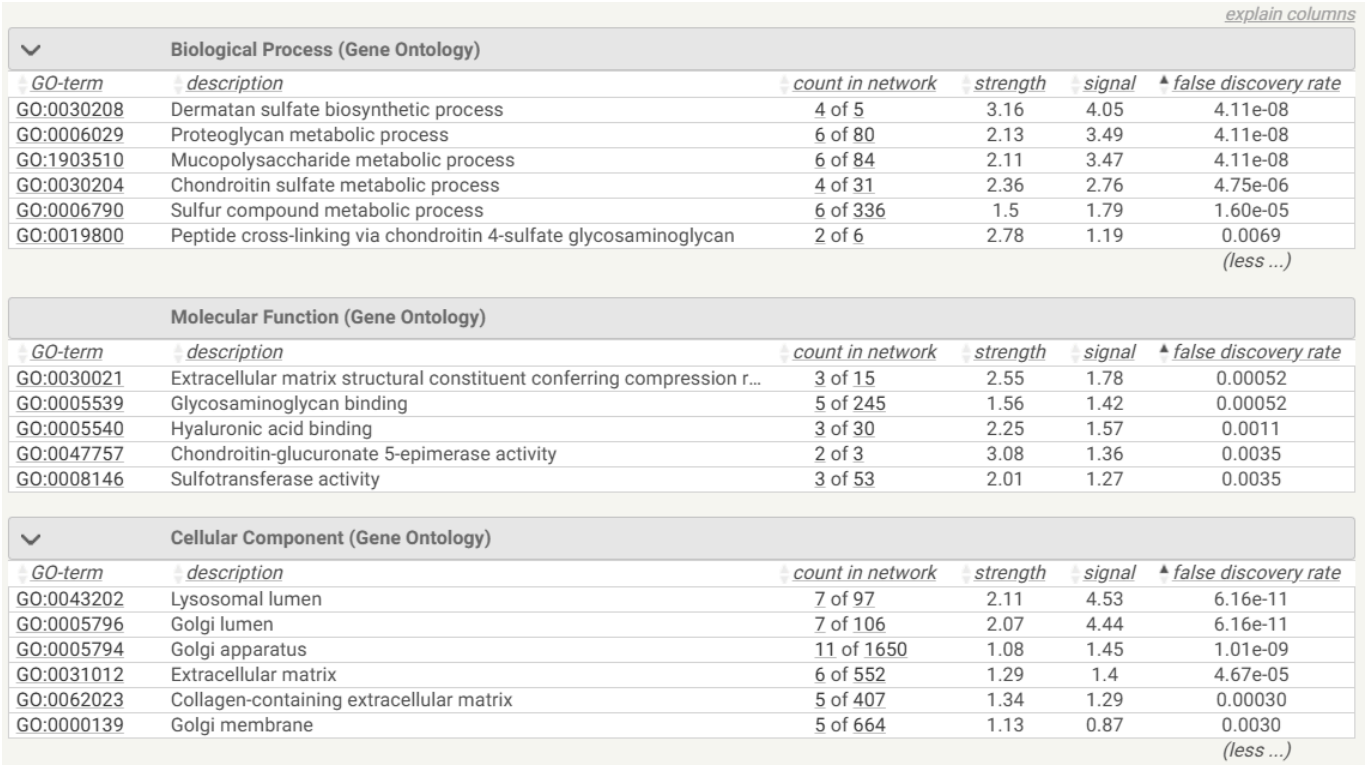
Рисунок 2. Граф функционального обогащения Gene Ontology для списка белков, построенный в STRING.

Цвет отражает значение FDR: более светлые оттенки соответствуют меньшему FDR и более значимому обогащению. Размер круга показывает число белков из исходного списка, попавших в соответствующий термин. Наиболее значимые термины связаны с обменом углеводных производных, гликозаминогликанов и протеогликанов.

KEGG Pathways					
pathway	*description*	*count in network*	*strength*	*signal*	*false discovery rate*
hsa00040	Pentose and glucuronate interconversions	2 of 33	2.23	1.49	0.0036
hsa00030	Pentose phosphate pathway	2 of 29	2.29	1.52	0.0034
hsa01200	Carbon metabolism	3 of 116	1.86	1.76	0.00062
hsa00561	Glycerolipid metabolism	3 of 59	2.16	2.28	0.00012
hsa01100	Metabolic pathways	7 of 1435	1.14	1.35	1.86e-06
hsa00051	Fructose and mannose metabolism	5 of 32	2.64	5.87	1.24e-10
					(less ...)

Рисунок 3. Обогащённые KEGG pathways для списка белков.

KEGG путей было получено 6. Наиболее значимый путь — Fructose and mannose metabolism, в который входит 5 из 7 белков исходного списка. Это указывает на основную функциональную тему набора — метаболизм фруктозы, маннозы и связанных углеводных соединений.



explain columns

Biological Process (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0030208	Dermatan sulfate biosynthetic process	4 of 5	3.16	4.05	4.11e-08
GO:0006029	Proteoglycan metabolic process	6 of 80	2.13	3.49	4.11e-08
GO:1903510	Mucopolysaccharide metabolic process	6 of 84	2.11	3.47	4.11e-08
GO:0030204	Chondroitin sulfate metabolic process	4 of 31	2.36	2.76	4.75e-06
GO:0006790	Sulfur compound metabolic process	6 of 336	1.5	1.79	1.60e-05
GO:0019800	Peptide cross-linking via chondroitin 4-sulfate glycosaminoglycan	2 of 6	2.78	1.19	0.0069

(less ...)

Molecular Function (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0030021	Extracellular matrix structural constituent conferring compression r...	3 of 15	2.55	1.78	0.00052
GO:0005539	Glycosaminoglycan binding	5 of 245	1.56	1.42	0.00052
GO:0005540	Hyaluronic acid binding	3 of 30	2.25	1.57	0.0011
GO:0047757	Chondroitin-glucuronate 5-epimerase activity	2 of 3	3.08	1.36	0.0035
GO:0008146	Sulfotransferase activity	3 of 53	2.01	1.27	0.0035

Cellular Component (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0043202	Lysosomal lumen	7 of 97	2.11	4.53	6.16e-11
GO:0005796	Golgi lumen	7 of 106	2.07	4.44	6.16e-11
GO:0005794	Golgi apparatus	11 of 1650	1.08	1.45	1.01e-09
GO:0031012	Extracellular matrix	6 of 552	1.29	1.4	4.67e-05
GO:0062023	Collagen-containing extracellular matrix	5 of 407	1.34	1.29	0.00030
GO:0000139	Golgi membrane	5 of 664	1.13	0.87	0.0030

(less ...)

Рисунок 4. Результаты GO-enrichment анализа в STRING для категорий Biological Process, Molecular Function и Cellular Component.

Всего было получено 12 GO-терминов. В Biological Process наиболее значимые термины связаны с метаболизмом протеогликанов, мукополисахаридов и сульфатированных гликозаминогликанов. В Molecular Function выделяются функции связывания гликозаминогликанов и гиалуроновой кислоты. В Cellular Component наиболее значимые категории связаны с лизосомальным просветом, аппаратом Гольджи и внеклеточным матриксом.

Gene Cooccurrace

Для дополнительной оценки функциональной связанности исследуемых белков был проведён анализ совместной встречаемости генов (gene cooccurrence) в различных таксономических группах с использованием STRING. Такой анализ позволяет определить, насколько часто соответствующие гены обнаруживаются совместно в геномах разных организмов. Совместная встречаемость генов обычно свидетельствует о функциональной связи между кодируемыми белками, их участии в общих метаболических путях или о координированной эволюции.

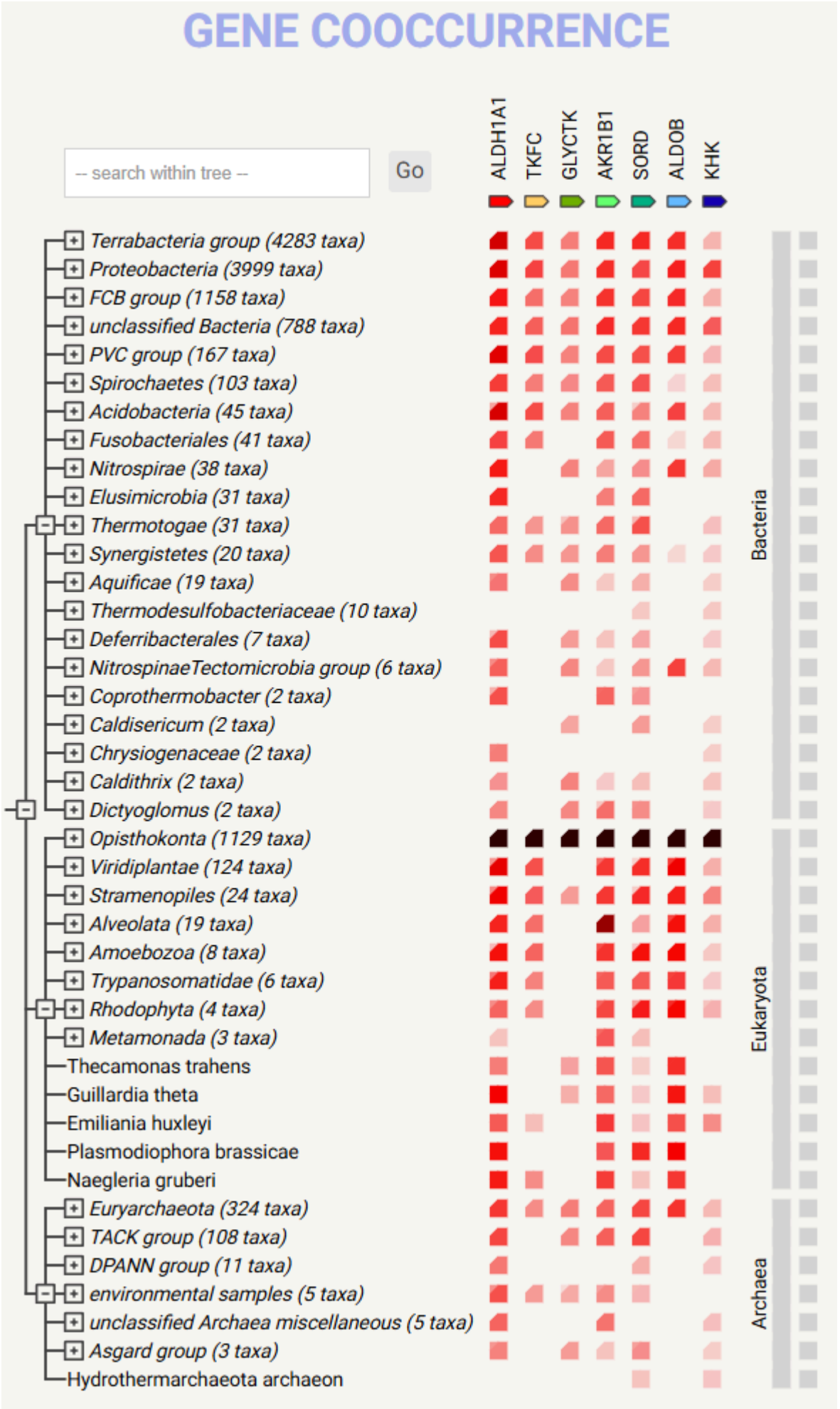

GENE COOCCURRENCE
-- search within tree --
Go
ALDH1A1
TKFC
GLYCTK
AKR1B1
SORD
ALDOB
KHK
Terrabacteria group (4283 taxa)
Proteobacteria (3999 taxa)
FCB group (1158 taxa)
unclassified Bacteria (788 taxa)
PVC group (167 taxa)
Spirochaetes (103 taxa)
Acidobacteria (45 taxa)
Fusobacteriales (41 taxa)
Nitrospirae (38 taxa)
Elusimicrobia (31 taxa)
Thermotogae (31 taxa)
Synergistetes (20 taxa)
Aquificae (19 taxa)
Thermodesulfobacteriaceae (10 taxa)
Deferribacterales (7 taxa)
NitrospinaeTectomicrobia group (6 taxa)
Coprothermobacter (2 taxa)
Caldisericum (2 taxa)
Chrysiogenaceae (2 taxa)
Caldithrix (2 taxa)
Dictyoglomus (2 taxa)
Opisthokonta (1129 taxa)
Viridiplantae (124 taxa)
Stramenopiles (24 taxa)
Alveolata (19 taxa)
Amoebozoa (8 taxa)
Trypanosomatidae (6 taxa)
Rhodophyta (4 taxa)
Metamonada (3 taxa)
Thecamonas trahens
Guillardia theta
Emiliania huxleyi
Plasmodiophora brassicae
Naegleria gruberi
Euryarchaeota (324 taxa)
TACK group (108 taxa)
DPANN group (11 taxa)
environmental samples (5 taxa)
unclassified Archaea miscellaneous (5 taxa)
Asgard group (3 taxa)
Hydrothermarchaeota archaeon
Bacteria
Eukaryota
Archaea

Рисунок 5. Анализ совместной встречаемости генов из списка Д. Эпштейна ALDH1A1, [redacted], GLYCTK, [redacted], ALDOB и KHK в различных таксономических группах.

На рисунке показана тепловая карта совместного присутствия исследуемых генов в геномах представителей Bacteria, Eukaryota и Archaea. Интенсивность окраски отражает степень совместной встречаемости генов: более тёмные клетки соответствуют более частому совместному обнаружению. Наиболее выраженная кооккуренция наблюдается у эукариотических организмов, что указывает на функциональную связанность исследуемых генов в путях углеводного обмена, прежде всего метаболизма фруктозы и связанных сахаров.

Gene coexpression

Для оценки возможной совместной регуляции исследуемых генов был проведён анализ коэкспрессии (gene coexpression) в STRING. Данный подход основан на сравнении профилей экспрессии генов в различных тканях, клетках или условиях. Если гены демонстрируют сходные паттерны экспрессии, это может свидетельствовать об их участии в общих биологических процессах, регуляторных путях или метаболических сетях. Анализ был выполнен как для данных *Homo sapiens*, так и для других организмов на основе перенесённых (transferred) данных.

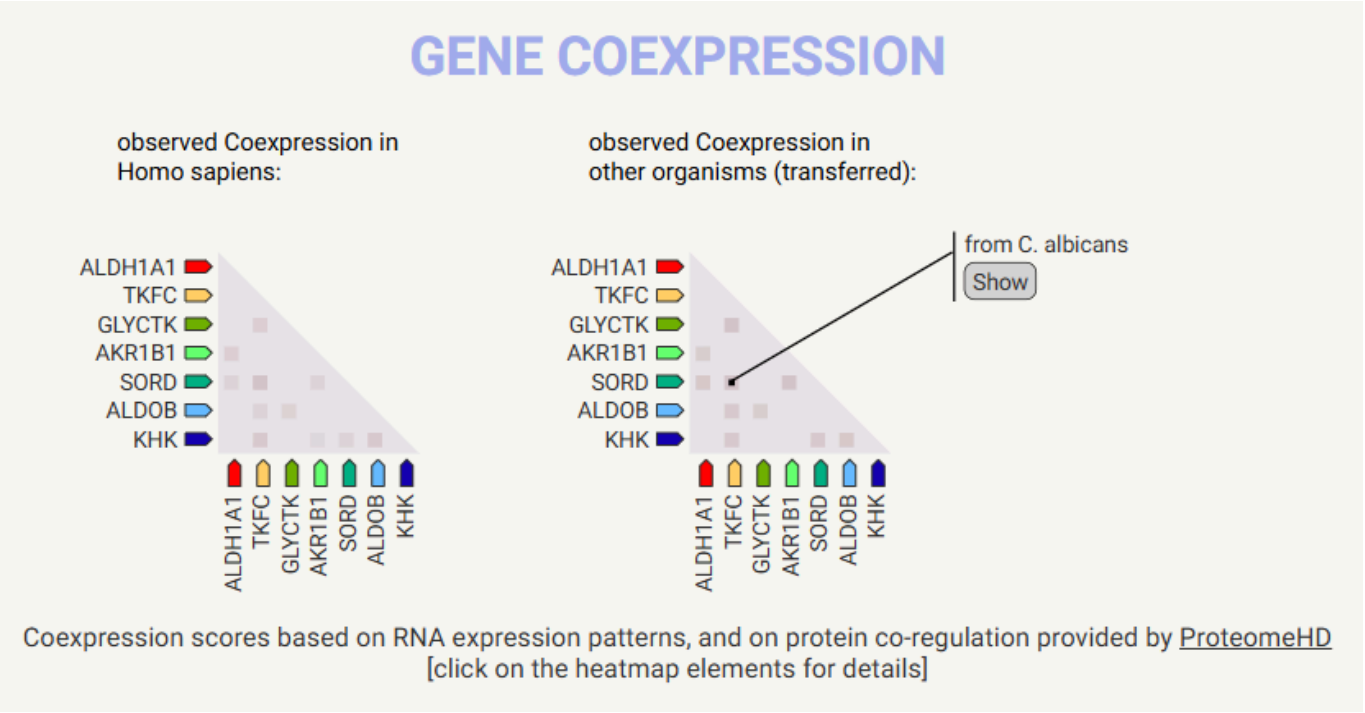


Рисунок 6. Анализ коэкспрессии генов.

На рисунке представлены матрицы коэкспрессии исследуемых генов для *Homo sapiens* и для других организмов на основе перенесённых данных STRING. Интенсивность окраски отражает степень сходства профилей экспрессии генов: более тёмные клетки соответствуют более выраженной коэкспрессии. Анализ основан на RNA-expression паттернах и данных о совместной регуляции белков. Совместные паттерны экспрессии особенно характерны для генов, участвующих в метаболизме фруктозы и полиоловом пути. Это согласуется с результатами KEGG- и GO-анализа, где основной темой набора также оказался углеводный обмен. При этом уровень коэкспрессии в целом выглядит умеренным, что может объясняться тем, что исследуемые ферменты экспрессируются в разных тканях и зависят от метаболического состояния клетки. Тем не менее наличие коэкспрессии поддерживает гипотезу о координированной работе данных генов в общих метаболических процессах.

Индивидуальный анализ (Human Protein Atlas)

Сервис позволяет решать следующий задачи:

- Проанализировать распределение белка по тканям
- Исследовать связь уровней экспрессии белков с выживаемостью пациентов при разных типах рака
- Выполнять анализ данных на уровне отдельных клеток
- Посмотреть на межбелковое и метаболическое взаимодействие

Для индивидуального анализа я выбрал ген KHK и соответствующий ему белок — кетогексокиназу.

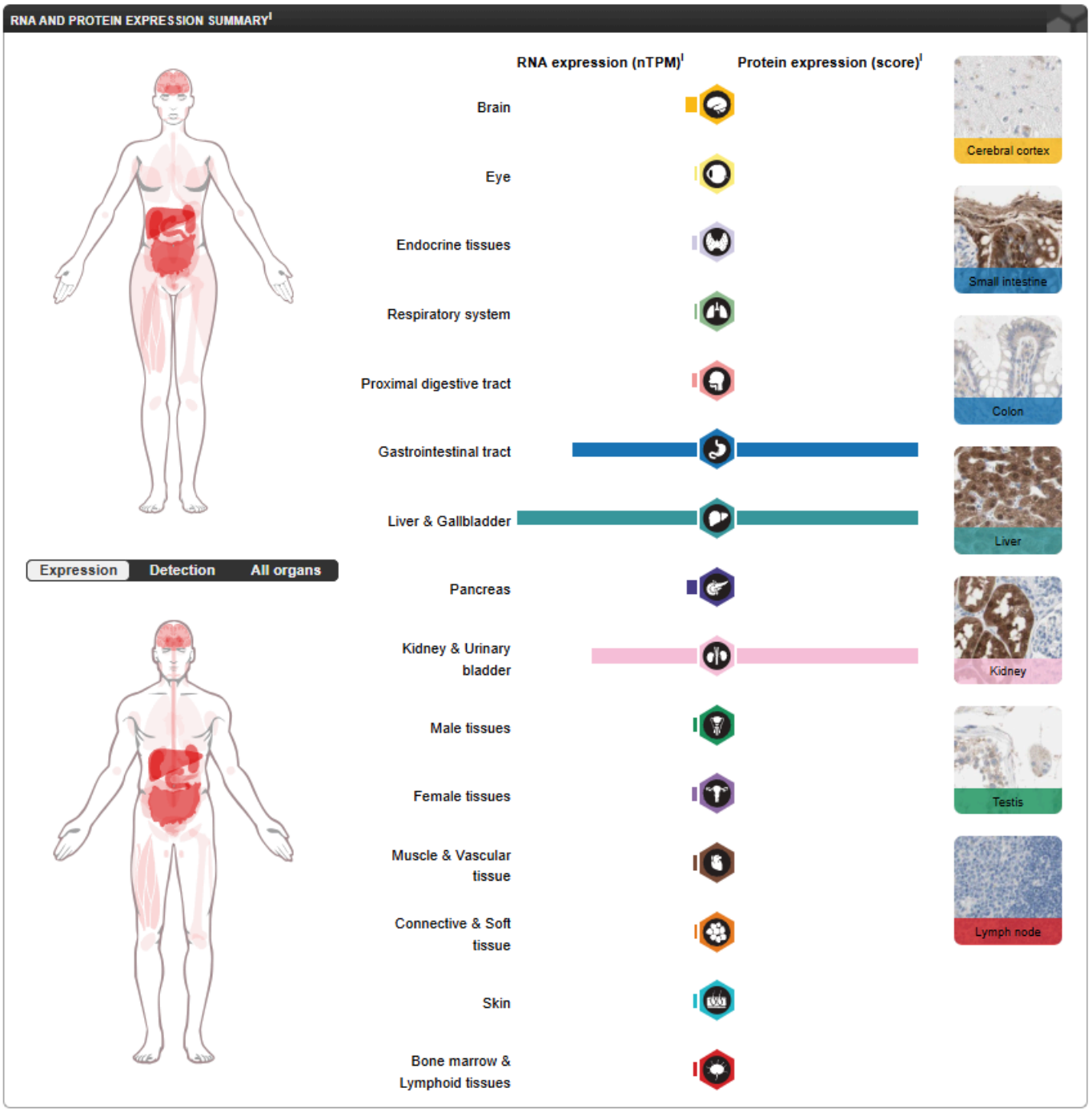


Рисунок 7. Тканеспецифическая экспрессия гена и белка KHK по данным Human Protein Atlas. Наибольший уровень экспрессии наблюдается в печени, почках и органах желудочно-кишечного

тракта, что соответствует функции KHK как ключевого фермента метаболизма фруктозы. Справа представлены примеры иммуногистохимического окрашивания белка в различных тканях человека.

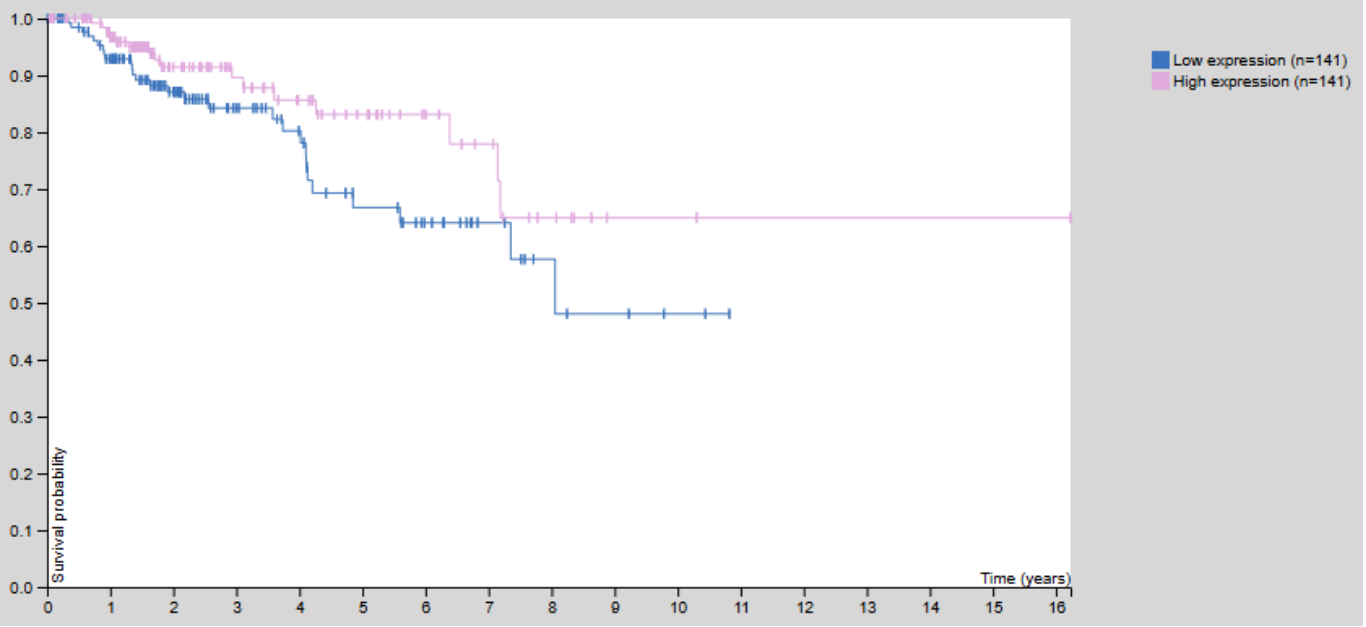


Рисунок 8. Кривые выживаемости.

Кривые выживаемости Kaplan–Meier для пациентов с различным уровнем экспрессии гена KHK. Показано сравнение групп с высоким и низким уровнем экспрессии KHK. Более высокая экспрессия гена ассоциирована с повышенной вероятностью выживания пациентов по сравнению с группой низкой экспрессии.

Рисунок 9. Межбелковое взаимодействие.

Сеть взаимодействий белка KHK по данным Human Protein Atlas. Для KHK показана ассоциация с белком LHX9 — транскрипционным фактором семейства LIM-homeobox. В отличие от STRING, данный граф отражает ограниченный набор ассоциаций, основанных на данных HPA и интегрированных источниках взаимодействий, а не полную метаболическую сеть ферментов.