Базы данных KEGG, GO и другие
Киселёва Ольга, ФББ, практикум 6

Задача практикума: с помощью базы данных или сервиса для одиночного и группового анализа генов исследовать возможности этой базы или сервиса, используя в качестве входных данных ID генов.

1. Краткое описание входных данных

Мне достался список (ссылка) из 25 ID генов человека. Их будем использовать в качестве входных данных для группового и одиночного анализа. Можно заметить белки с одинаковыми мнемониками, которые, вероятно, будут кластеризоваться по функции. Для анализа этого набора генов я воспользуюсь самой, на мой взгляд, универсальной и визуально приятной базой из рассматриваемых в курсе - STRING.

2. База данных для группового анализа - STRING

2.1. Возможности базы STRING:

- **Построение и анализ сетей белковых взаимодействий:** на основе данных о взаимодействии белков STRING строит графы, где узлами являются белки, а ребрами — связи между ними.
- **Анализ обогащения категорий:** позволяет статистически оценить насколько чаще гены из нашего списка попадают в категорию, чем если бы выбрали их из всех генов случайно. Результаты включают p-value, FDR (скорректированное значение) и количество генов, связанных с каждым термином.
- **Функциональная аннотация новых генов:** по функции уже охарактеризованных белков сети можно предположить функцию неизвестных или слабоизученных белков, связанных с ними

2.2. Анализ списка генов на взаимосвязь

a) Для запуска анализа были использованы следующие параметры:

- **Метод:** анализ обогащения (Multiple proteins) находит статистически значимые общие функции и пути для списка из нескольких белков.
- **Поправка на множественное тестирование** с использованием процедуры Бенджамина - Хохберга указана в столбце FDR (False Discovery Rate) или частота ложных открытий - эта мера описывает, насколько значимым является обогащение.

b) Полученная схема белок-белковых взаимодействий исследуемого набора представлена на рис. 1. Узлы графа - белки, цвета рёбер - уровень достоверности взаимодействия. Для более удобной визуализации карты я кластеризовала белки с помощью метода k-средних, теперь для каждого кластера можно проводить анализ обогащения ещё по отдельности.

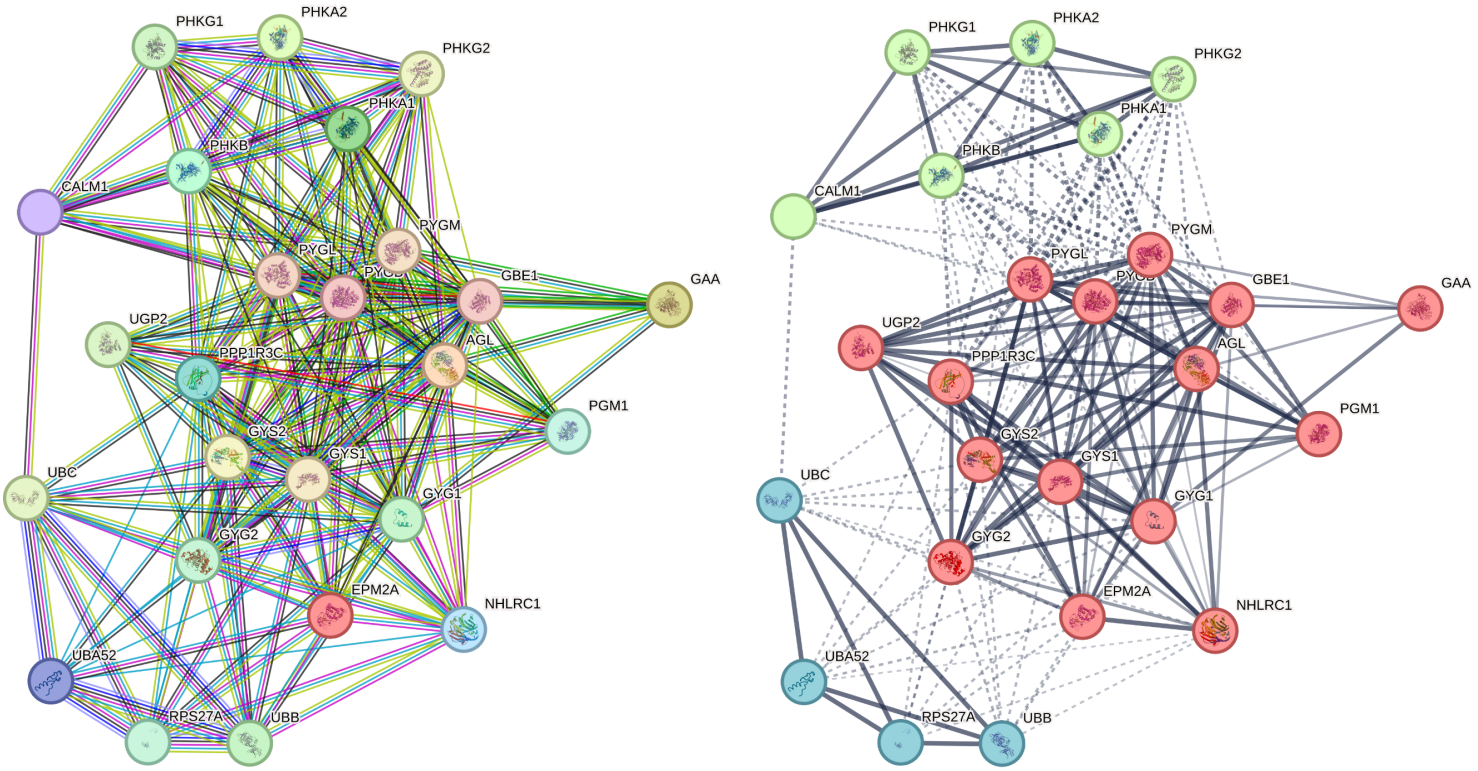

Рис. 1. Слева - сеть белковых взаимодействий для продуктов моего набора генов, где цвет линии обозначает тип доказательной базы взаимодействия (например, известные из курируемых баз данных - синие, по эксперименту - фиолетовые, извлеченные из статей - жёлтые и т.д.). Справа - та же сеть кластеризованная, где толщина линий показывает силу поддержки данных, пунктир - взаимодействия между кластерами.

По результатам кластерного анализа выделяются 3 функциональные группы генов: метаболизм гликогена, фосфорилазкиназный комплекс и комплекс убиквитинилирования.

color	cluster Id	gene count	description
	Cluster 1	15	Glycogen metabolism
	Cluster 2	6	1. Phosphorylase kinase complex 2. Glycogen breakdown (glycogenolysis)
	Cluster 3	4	1. Maturation of protein E 2. Protein tag

Далее хочется подробнее посомтреть, какие термины были обогащены. В окне «Analysis», где отображаются результаты обогащения по другим различных базам, таких как GO, KEGG, Reactome, DISEASES и прочим. Устройство «Analysis» изображено на рис. 3.

b) Таблицу с результатами анализа, отсортированную по FDR, можно получить интерактивно на сайте STRING (ссылка)

c) Выдача обогащенных терминов была такой:

Biological Process (Gene Ontology) - *18 GO-terms significantly enriched;*
Molecular Function (Gene Ontology) *22 GO-terms significantly enriched;*

Cellular Component (Gene Ontology) *14 GO-terms significantly enriched;*
KEGG Pathways *13 pathways significantly enriched;*
Reactome Pathways *183 pathways significantly enriched;*

Biological Process (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0005977	Glycogen metabolic process	19 of 53	2.45	12.95	9.82e-39
GO:0005978	Glycogen biosynthetic process	12 of 21	2.65	9.54	3.96e-25
GO:0016051	Carbohydrate biosynthetic process	13 of 133	1.89	5.34	5.48e-19
GO:0006091	Generation of precursor metabolites and energy	20 of 411	1.58	4.7	3.33e-26
GO:0005980	Glycogen catabolic process	6 of 12	2.6	4.31	3.47e-11

(more ...)

Molecular Function (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0035251	UDP-glucosyltransferase activity	5 of 12	2.52	3.18	2.15e-08
GO:0004689	Phosphorylase kinase activity	4 of 4	2.9	2.96	1.15e-07
GO:0016758	Hexosyltransferase activity	10 of 194	1.61	2.85	1.58e-10
GO:0031386	Protein tag	4 of 13	2.38	2.3	2.91e-06
GO:0102499	SHG alpha-glucan phosphorylase activity	3 of 3	2.9	2.02	1.97e-05

(more ...)

Рис. 2. Устройство выдачи поля «Analysis» по исследуемым генам. На рисунке представлены результаты GO-обогащения, представленность KEGG-путей и STRING. Результаты были отсортированы по колонке count in network. Первое число в данной колонке обозначает количество исследуемых белков с данной категорией, второе — общее количество известных белков с данным термином.

Анализ обогащения по категории Gene Ontology Biological Process показал, что исследуемые гены преимущественно вовлечены в процессы, связанные с углеводным обменом, причем наиболее представленным термином является «Glycogen metabolic process» (GO:0005977) — метаболизм гликогена. Для этого термина значительная часть белков из сети аннотированы с высоким уровнем достоверности (FDR = 9.82e-39, strength = 2.45).

Другие метаболические термины также значительно представлены. К ним относятся: синтез гликогена: (12 белков), распад гликогена (6 белков), синтез предшественников метаболизма и энергии (20 белков).

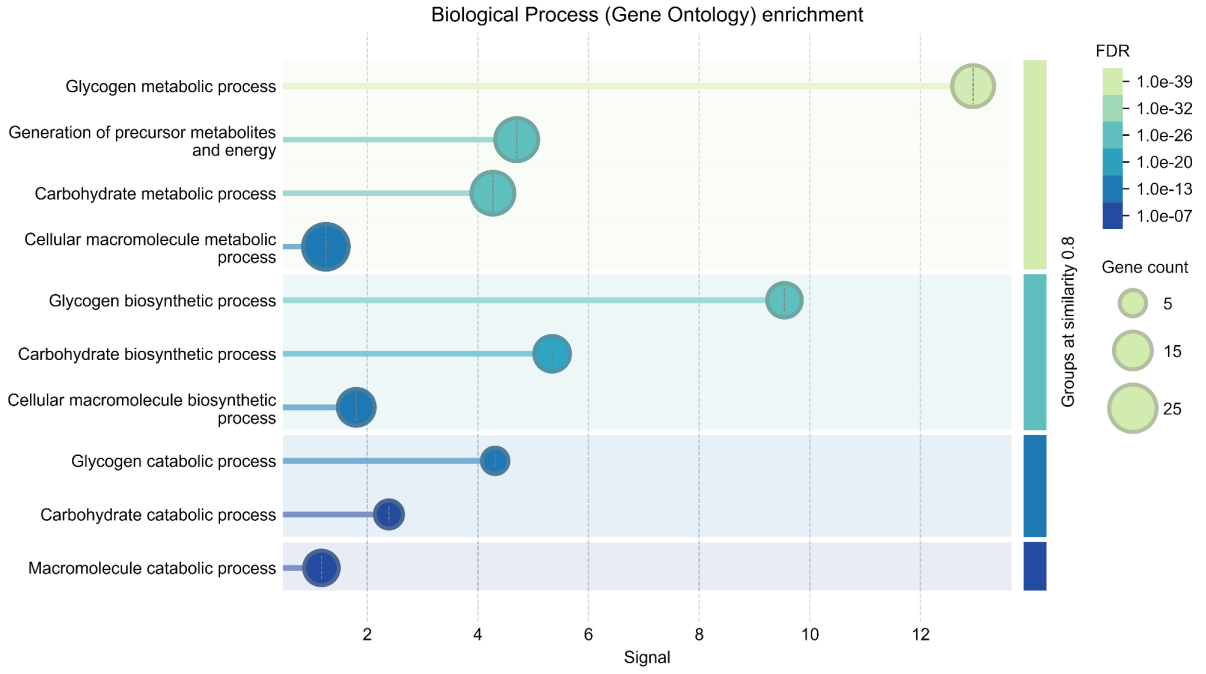

Рис. 3. Визуализация GO-обогащения Biological Process. Сортировка происходила по количеству генов в категории (gene count). По оси Y изображены GO-категории, по оси X — количество генов в данных категориях. Цветом обозначено значение FDR, размером точек — количество генов в категории.

Для более конкретного анализа набора обратимся к GO-обогащению молекулярных функций.

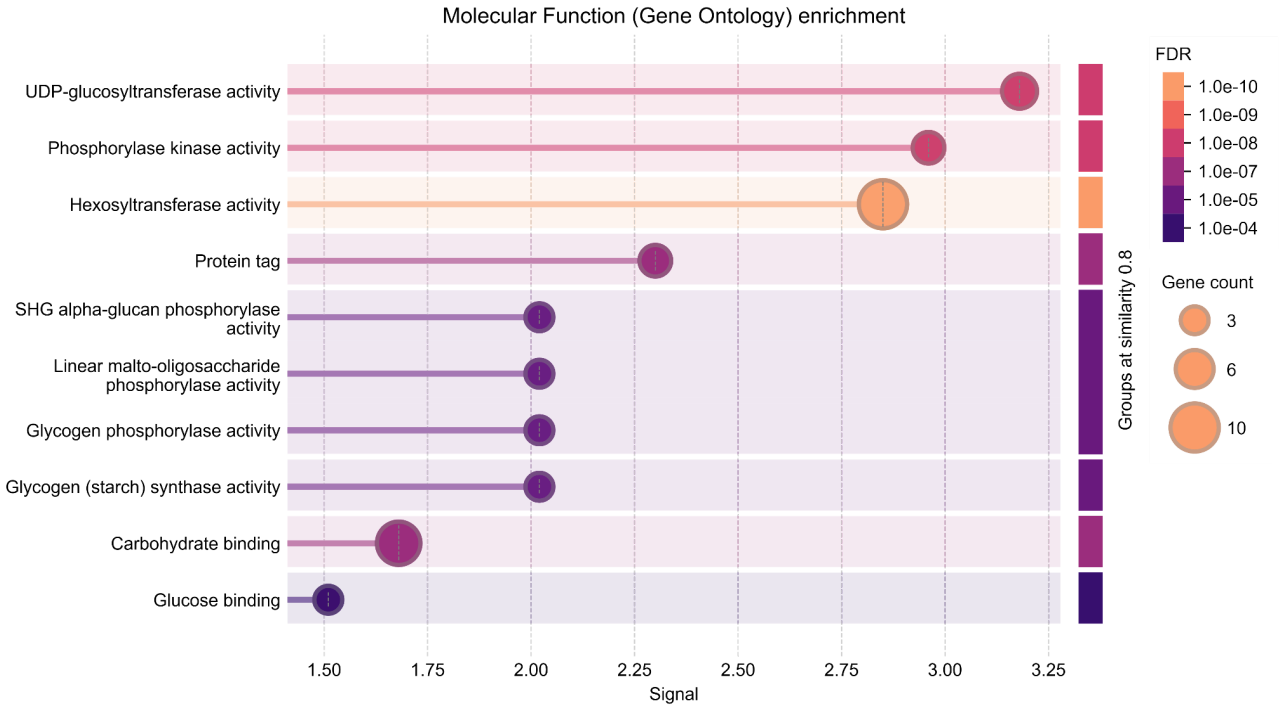

Рис. 4. Визуализация GO-обогащения Molecular Function. Подписи к рисунку такие же, как к рис.3

Наиболее обогащённым термином (наименьший FDR = 1.58e-10) является "Hexosyltransferase activity" (GO:0016758). Гораздо более показательна ассоциированная с ней "UDP-glucosyltransferase activity" (GO:0035251), которая специфична для использования UDP-глюкозы.

Фосфорилазы гликогена "SHG alpha-glucan phosphorylase activity" и "Linear malto-oligosaccharide phosphorylase activity" также обогащены.

Обогащение термина "Phosphorylase kinase activity" включает в себя полноценный регуляторный модуль, предназначенный для активации фосфорилазы и запуска каскада распада гликогена в ответ на внутриклеточные сигналы (например, повышение Ca^{2+}). Термин "Protein tag" ассоциирован с генами убиквитина, указывая на то, что белки метаболизма гликогена являются мишенями для убиквитинилирования.

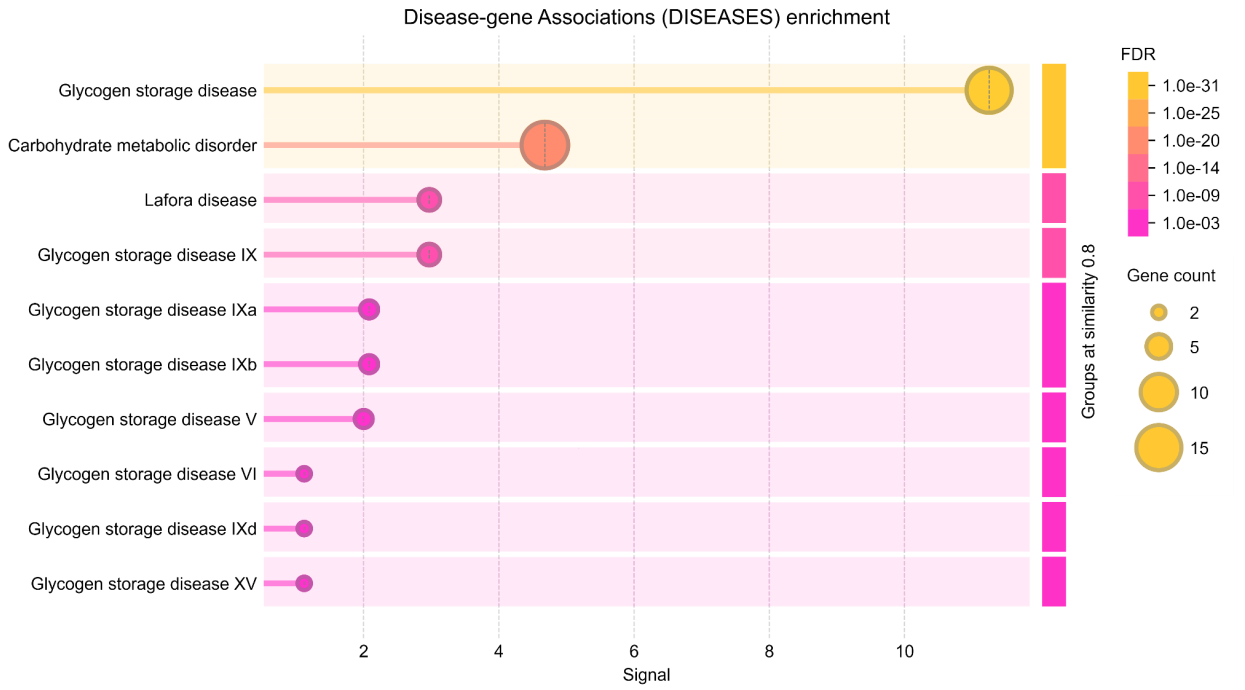

Рис. 5. Визуализация обогащения DISEASES. Подписи к рисунку такие же, как к рис. 3.

Анализ обогащения по базе данных DISEASES позволил подтвердить ассоциацию исследуемого генного набора с болезнями накопления гликогена "Glycogen storage disease". Более специфичный термин - Болезнь Лафора (Lafora disease). Дефицит ферментов “качества” гликогена приводит к накоплению токсичного аномального гликогена (телец Лафора) в нейронах, что клинически проявляется как тяжелая прогрессирующая миоклонус-эпилепсия и деменция. Болезнь вызвана потерей функции комплекса лафорин-малин, кодируемого генами EPM2A и NHLRC1.

Таким образом, можно сделать вывод, что исследуемый набор представляет собой группу генов, преимущественно вовлечённую в цитозольный синтез и распад гликогена, регуляцию этих процессов через фосфорилирование/дефосфорилирование и убиквитинирование, а также в ассоциированные с ними заболевания — гликогенозы и болезнь Лафора. На самом деле мне показалось удивительным, что благодаря одному сервису можно выделить довольно много информации о генах, что является несомненным плюсом STRING.

3. База данных для одиночного анализа - Human Protein Atlas

В качестве следующей базы данных я решила выбрать Human Protein Atlas. С помощью этого ресурса можно определять тканеспецифичную экспрессию генов, анализировать их локализацию в клетках, идентифицировать биомаркеры заболеваний, а также изучать экспрессию белков в норме и при патологиях, включая раковые опухоли.

Я решила взять в рассмотрение уже упомянутый ген EPM2A, который кодирует белок лафорин (laforin). Лафорин представляет собой фосфатазу, обладающую как гликан-фосфатазной, так и протеинтирозинфосфатазной активностью. Белок участвует в регуляции метаболизма гликогена, предотвращая его гиперфосфорилирование и образование токсичных агрегатов. Мутации в этом гене вызывают прогрессирующую миоклонус-эпилепсию 2 типа — болезнь Лафора . Мне

стало интересно, в каких тканях наблюдается экспрессия этого белка, в каких клеточных структурах он находится и ассоциирован ли он с какими-либо заболеваниями.

Анализ тканевой экспрессии

Согласно данным Human Protein Atlas, ген EPM2A демонстрирует широкую тканевую экспрессию. На основе RNA-секвенирования (HPA, GTEx, FANTOM5) установлено, что мРНК EPM2A детектируется во всех изученных тканях, повышенная экспрессия - в языке и скелетных мышцах. Данные по белковой экспрессии, полученные с помощью иммуногистохимии, уточняют эту картину: мышечная ткань, лимфоузлы, печень, кора мозга.

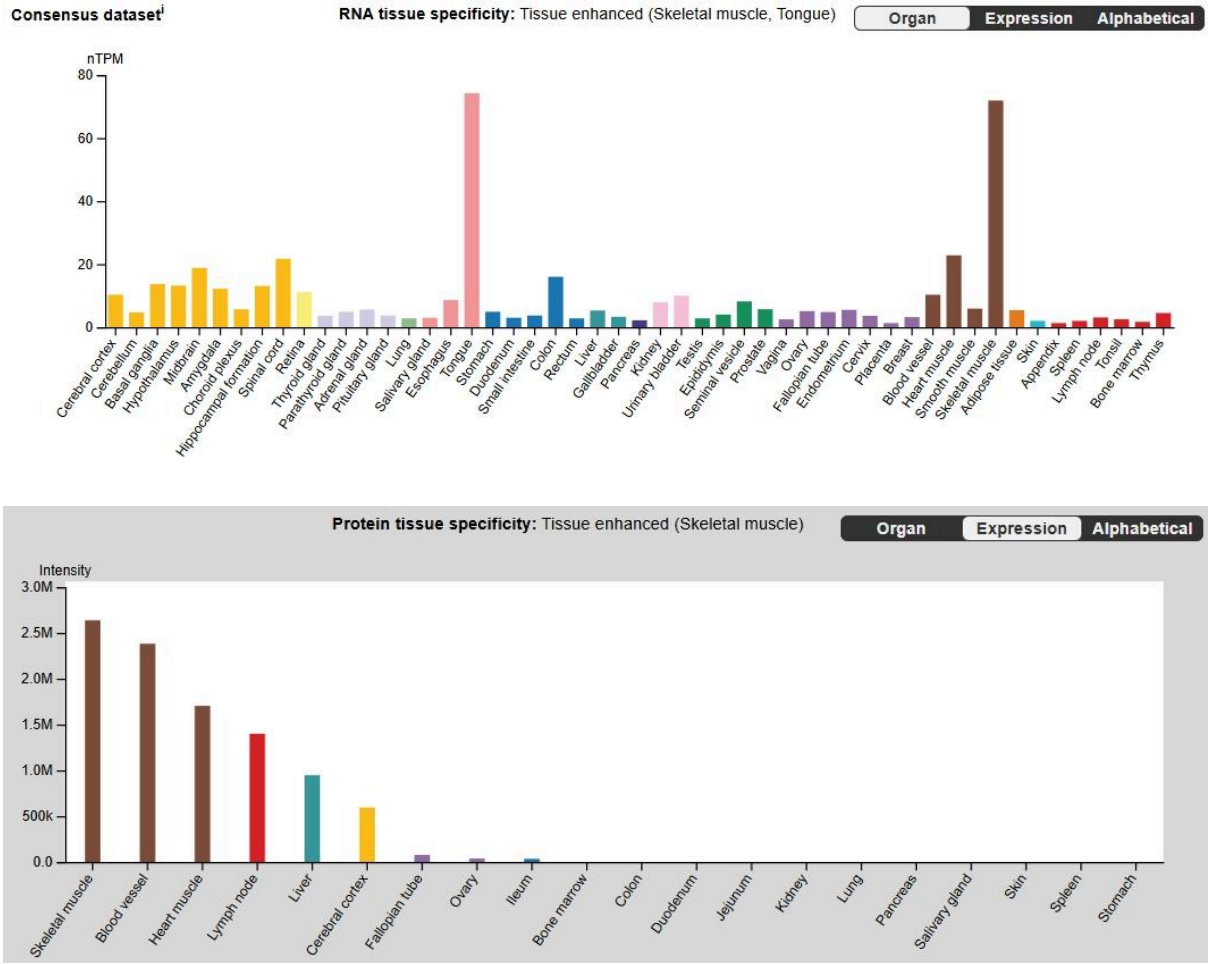

Рис. 6. Столбчатая диаграмма экспрессии РНК (верхняя) и белка (нижняя) гена EPM2A в различных тканях человека. По оси ординат отложено значение nTPM (нормализованное количество транскриптов, кодирующих белки, на миллион).

Клеточная локализация

Анализ клеточной локализации белка EPM2A проводился с использованием иммунофлуоресцентной микроскопии. Согласно анализу, лафорин локализуется преимущественно в нуклеоплазме (ядре), причём эта локализация характеризуется как «усиленная» (enhanced). Такая локализация в ядре соответствует его функции: лафорин может участвовать в регуляции экспрессии генов, взаимодействуя с транскрипционными факторами, помимо своей основной роли в цитоплазматическом метаболизме гликогена.

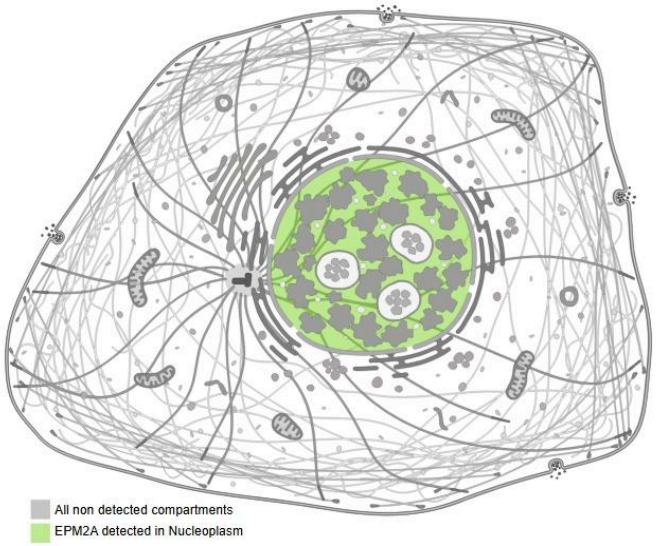

Рис. 7. Схематичное изображение клетки, на котором видно клеточную локализацию исследуемого белка EPM2A (зеленый цвет).

Анализ ассоциации с раковыми заболеваниями

Для оценки потенциала EPM2A в качестве биомаркера раковых заболеваний я проанализировала данные из базы The Cancer Genome Atlas (TCGA), представленные в Human Protein Atlas. На представленных графиках распределения уровней экспрессии белка по типам рака видно, что EPM2A не демонстрирует высокой специфичности к какому-либо одному типу опухолей. Белок детектируется в большинстве исследованных тканей, что коррелирует с его широкой экспрессией в норме. Таким образом, EPM2A не подходит на роль клинически значимого биомаркера для прогноза течения раковых заболеваний.

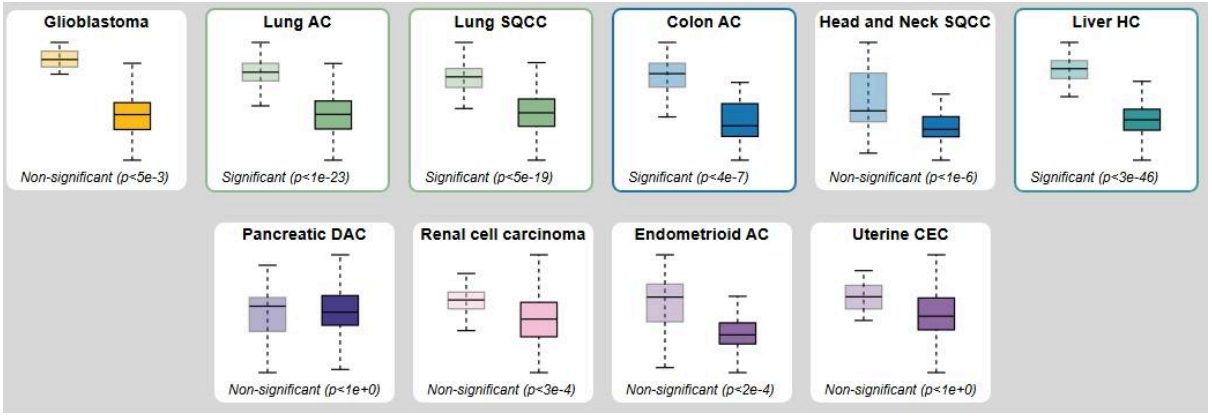

Рис. 8. Распределение уровней экспрессии белка по типам рака. В некоторых типах рака значимая повышенная экспрессия белка, в некоторых нет. Тем не менее, экспрессия EPM2A не тканеспецифична для одного типа рака

Для гена EPM2A удалось установить, что он экспрессируется во всех тканях человека, его белок лафорин локализуется преимущественно в ядре (нуклеоплазме), а сам ген не является специфическим биомаркером раковых заболеваний.