Отчет по практикуму 6

Блок 1. Входные данные

Входные данные представляли собой список из 127 идентификаторов генов человека.
Часто встречаются GPC*, B3GALT*, HS3ST*, CHST*, что связано с протеогликанами и гликозаминогликанами.
Поэтому можно предположить, что список в основном связан с компонентами внеклеточного матрикса.

Блок 2. База данных STRING для группового анализа

Для группового анализа я выбрала сервис STRING. Он удобен тем, что позволяет не только получить таблицы обогащения по GO, KEGG и Reactome, но и сразу построить белок-белковый граф.

С помощью STRING можно решать несколько задач.

1) STRING позволяет проверить, связаны ли белки из списка между собой функционально или список больше похож на случайный набор генов.
2) Можно определить, какие биологические процессы, молекулярные функции и клеточные компоненты статистически обогащены в данном списке.
3) Можно визуализировать белки в виде сети и примерно понять, какие функциональные группы есть внутри списка.

Для построения сети была выбрана full STRING network (по умолчанию) без добавления дополнительных интеракторов, чтобы анализировать только исходные гены. Остальные параметры также были по умолчанию. Ребра отображали связь между белками, в то время как их цвет показывал достоверность данного функционального взаимодействия. Значимость оценивалась по FDR (метод Беньямини-Хохберга), то есть с поправкой на множественное тестирование. Результаты были отсортированы по FDR.
Полная таблица с результатами
При пороге FDR < 0.05 были найдены термины GO обогащения, ссылки на статьи PubMed, пути KEGG и Reactome, Tissue Expression, Diseases, Protein Domains и не только.

На рисунке 1 видно, что большая часть белков входит в общую плотную сеть. Это значит, что список не выглядит случайным: многие белки связаны с похожими процессами.
Значение PPI enrichment p-value < 1.0e-16, показывает, что белки из списка имеют гораздо больше связей между собой, чем ожидалось бы для случайного набора белков такого размера.

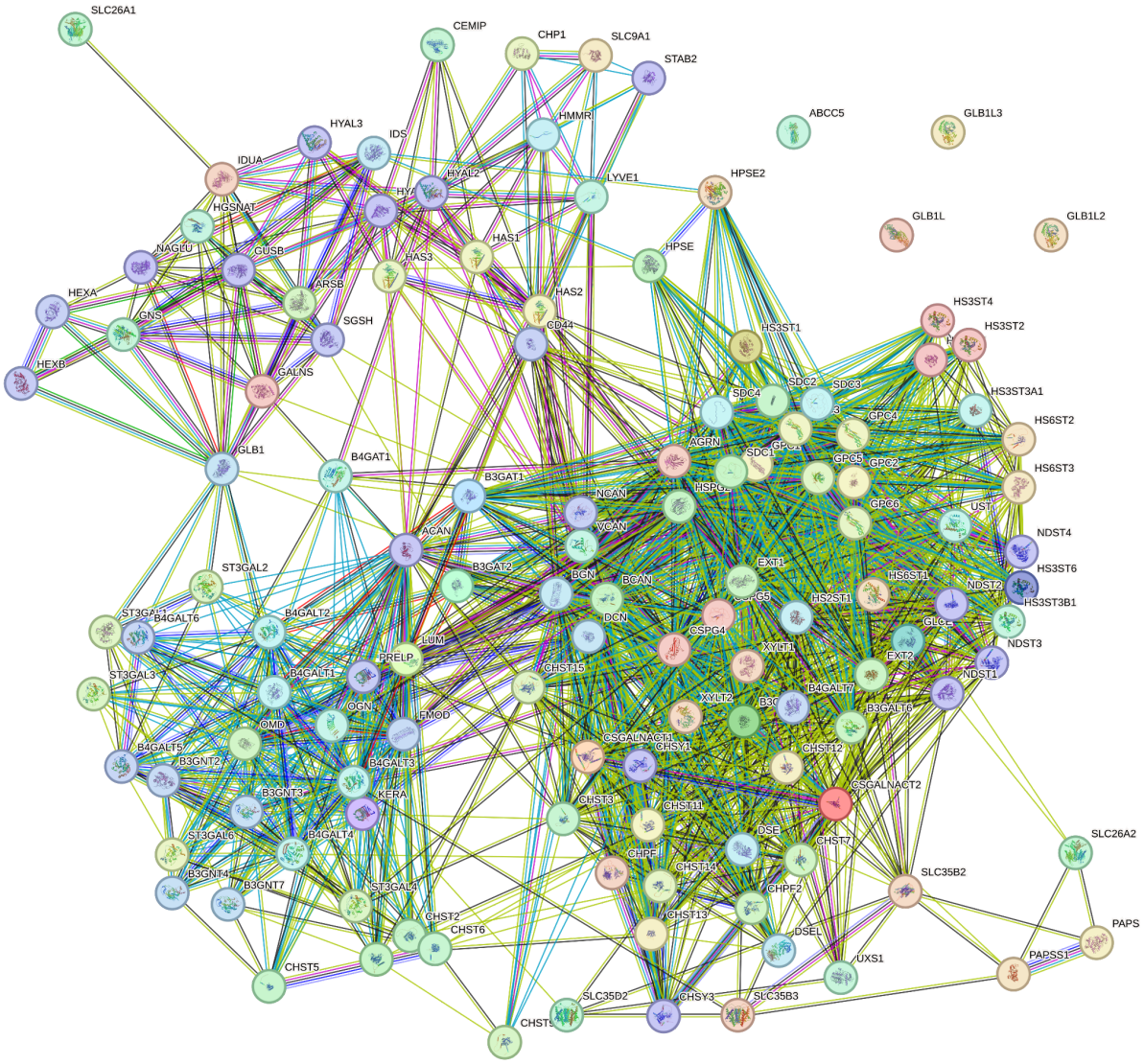

Рисунок 1. Сеть функциональных ассоциаций белков из исследуемого списка, построенная в STRING для *Homo sapiens*. Узлы соответствуют белкам, рёбра — функциональным ассоциациям между ними. Наличие плотной сети указывает на функциональную связанность списка.

По результатам анализа (рис 2) общей темой списка оказался метаболизм и биосинтез гликозаминогликанов, что также связано с протеогликанами. Это хорошо совпадает с исходным предположением. То есть эти гены кодируют ферменты, которые участвуют в сборке, удлинении и сульфатировании ГАГ-цепей, как мы можем видеть по Biological Process и Molecular Function.
В клеточной локализации тоже все достаточно логично (рис 2, 3). Среди значимых терминов оказались различные части аппарата Гольджи и лизосом (мембраны, просветы). Это хорошо объясняется биологией гликозаминогликанов.
Синтез и модификация протеогликанов происходят в секреторном пути, особенно в аппарате Гольджи. Поэтому появление терминов, связанных с Гольджи, ожидаемо. А лизосомальный просвет связан с другой частью списка — ферментами деградации гликозаминогликанов. То есть список содержит гены не только синтеза ГАГ, но и их деградацией

Так же, в сервисе STRING можно визуализировать данные рисунка 2, в график, как тот, что приведет для клеточной локализации на рисунке 3 и для путей Reactome на рисунке 4.

Biological Process (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0030203	Glycosaminoglycan metabolic process	77 of 115	2.02	21.09	1.46e-123
GO:0006022	Aminoglycan metabolic process	78 of 126	1.98	20.42	1.46e-123
GO:0006024	Glycosaminoglycan biosynthetic process	61 of 69	2.14	20.29	1.04e-100
GO:0006023	Aminoglycan biosynthetic process	62 of 73	2.12	20.19	1.18e-101
GO:1903510	Mucopolysaccharide metabolic process	59 of 84	2.04	17.86	7.61e-93

(more ...)

Molecular Function (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0008146	Sulfotransferase activity	29 of 53	1.93	8.95	3.76e-40
GO:0008194	UDP-glycosyltransferase activity	32 of 143	1.54	6.28	8.60e-35
GO:0034483	Heparan sulfate sulfotransferase activity	15 of 15	2.19	5.91	5.39e-23
GO:0016757	Glycosyltransferase activity	37 of 268	1.33	5.08	4.86e-34
GO:0016758	Hexosyltransferase activity	30 of 194	1.38	4.79	2.40e-28

(more ...)

Cellular Component (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0043202	Lysosomal lumen	39 of 97	1.79	10.06	3.60e-52
GO:0005796	Golgi lumen	28 of 106	1.61	6.34	1.56e-32
GO:0000139	Golgi membrane	68 of 664	1.2	5.46	2.08e-60
GO:0032580	Golgi cisterna membrane	19 of 89	1.52	4.18	3.66e-20
GO:0005794	Golgi apparatus	101 of 1650	0.98	3.95	4.15e-79

(more ...)

Рисунок 2. Наиболее значимые GO-термины по результатам анализа STRING. В категории Biological Process преобладают процессы метаболизма и биосинтеза гликозаминогликанов, в Molecular Function — сульфотрансферазная и гликозилтрансферазная активности, в Cellular Component — аппарат Гольджи и лизосомы. Значимость оценивалась по FDR.

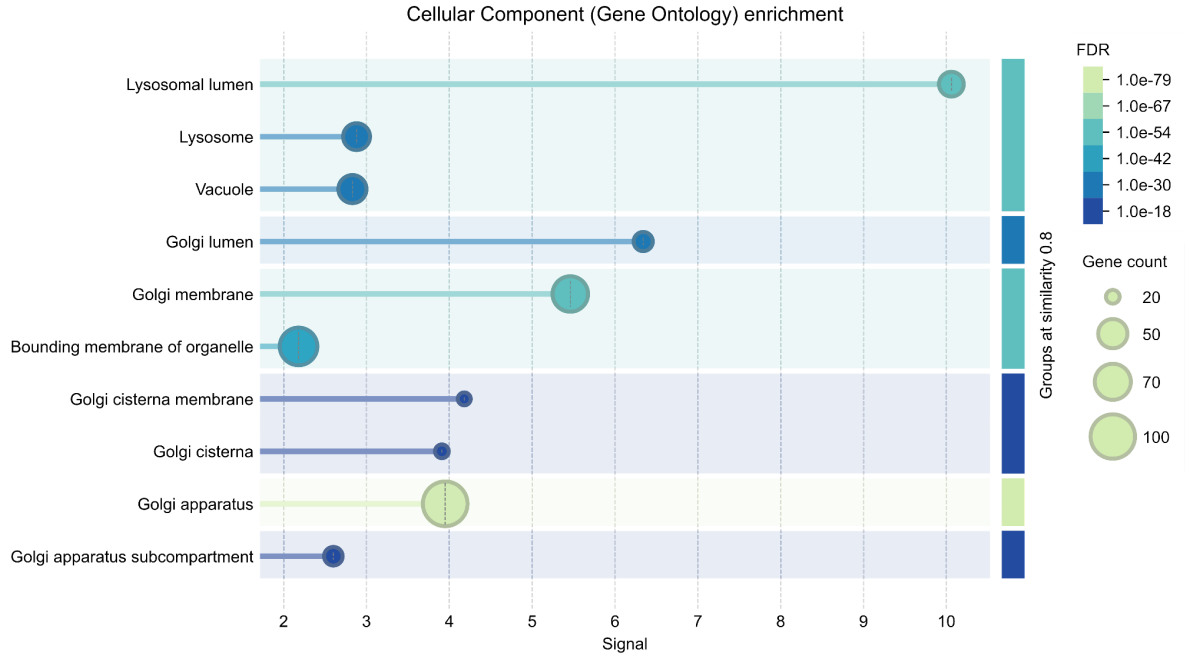

Рисунок 3. Обогащённые термины GO Cellular Component. Размер круга показывает число генов из списка, связанных с данным термином, а цвет — значение FDR. Среди наиболее значимых локализаций представлены аппарат Гольджи, мембрана/просвет Гольджи и лизосомальный просвет.

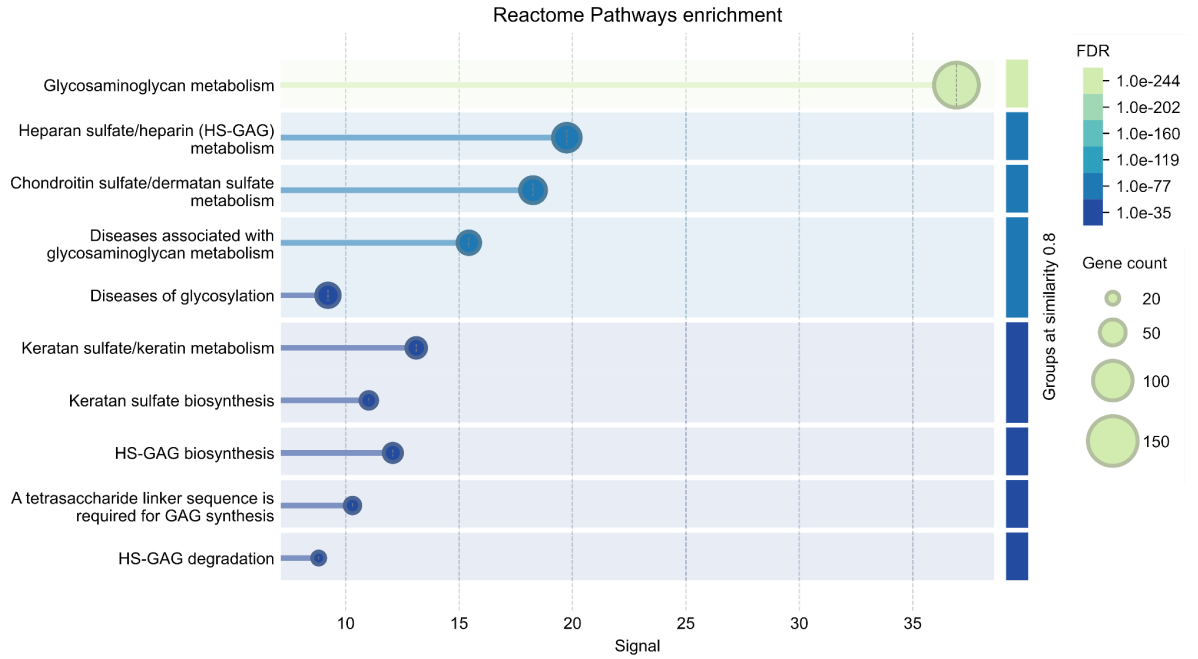

Рисунок 4. Обогащённые пути Reactome по результатам анализа STRING. Размер круга показывает число генов из списка, связанных с данным путём, а цвет — значение FDR. Наиболее значимые пути относятся к метаболизму гликозаминогликанов и протеогликанов.

Результаты Reactome подтверждают выводы GO-анализа. Наиболее значимые пути также связаны с метаболизмом, синтезом и деградацией гликозаминогликанов. Поэтому можно сказать, что одна и та же тема видна сразу в нескольких типах анализа: в GO-терминах, в клеточной локализации и в путях.

В целом результаты анализа показывают, что общая тема списка — гликозаминогликаны. Список выглядит биологически логичным, потому что разные результаты STRING указывают на одну и ту же систему: синтез, модификацию и деградацию гликозаминогликанов.

Блок 3. База данных THE HUMAN PROTEIN ATLAS для индивидуального анализа

Эта база позволяет посмотреть, где в организме и в каких клетках экспрессируется выбранный ген, а также оценить тканевую и клеточную локализацию соответствующего белка.
С помощью этой базы данных можно решать такие задачи как:

1) Анализ тканевой и клеточной специфичности экспрессии генов и белков человека.
2) Изучение белковой локализации по данным иммуногистохимии и транскриптомным данным.
3) Сравнивать экспрессии генов в нормальных и патологических тканях, например различные типы опухолей.

В качестве примера был выбран ген ACAN. Он входит в исходный список ID и кодирует белок аггрекан. Аггрекан является крупным протеогликаном внеклеточного матрикса.

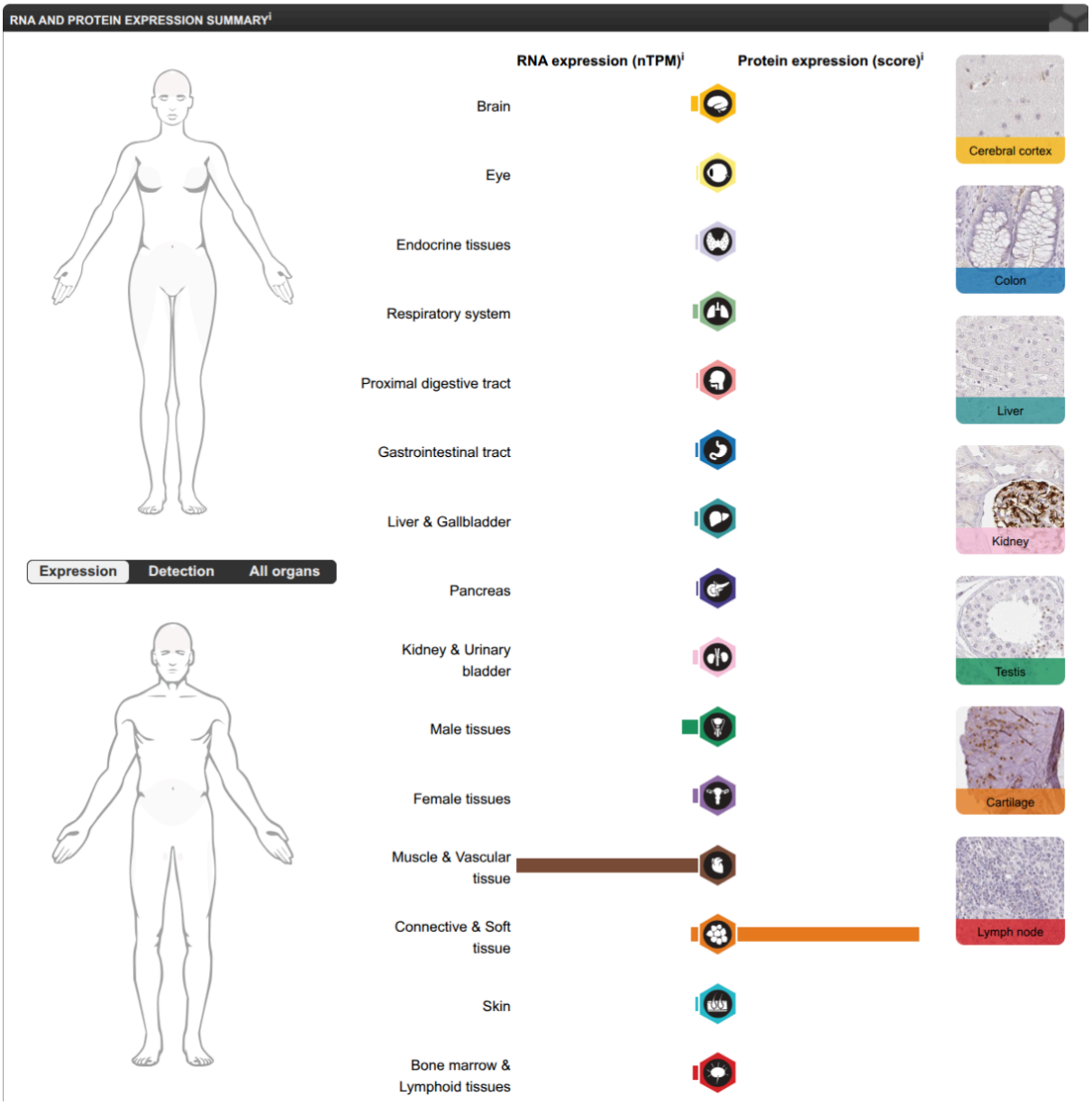

Рисунок 5. Сводка RNA и protein-экспрессии гена ACAN по данным Human Protein Atlas. На графике показано распределение РНК-экспрессии по группам тканей, а справа приведены примеры иммуногистохимического окрашивания белка в отдельных тканях. Наиболее заметная RNA-экспрессия ACAN наблюдается в соединительной ткани и мышечно/сосудистой ткани.

На рисунке 5 показана экспрессия гена ACAN в тканях человека. По данным Human Protein Atlas, наиболее выраженная РНК-экспрессия видна в группах соединительной ткани и мышечно-сосудистой ткани. Это хорошо согласуется с функцией ACAN, потому что аггрекан является компонентом внеклеточного матрикса, а такие белки особенно важны для соединительных тканей.

На рис 5 справа также показаны примеры иммуногистохимического окрашивания белка в разных тканях, в том числе в хондроцитах. Это тоже логично, потому что аггрекан является одним из характерных протеогликанов хрящевого матрикса. Таким образом, уже по тканевой экспрессии видно, что ACAN связан не с универсальными внутриклеточными процессами, а скорее со структурными тканями и внеклеточным матриксом, что также подтверждается на рис 6.

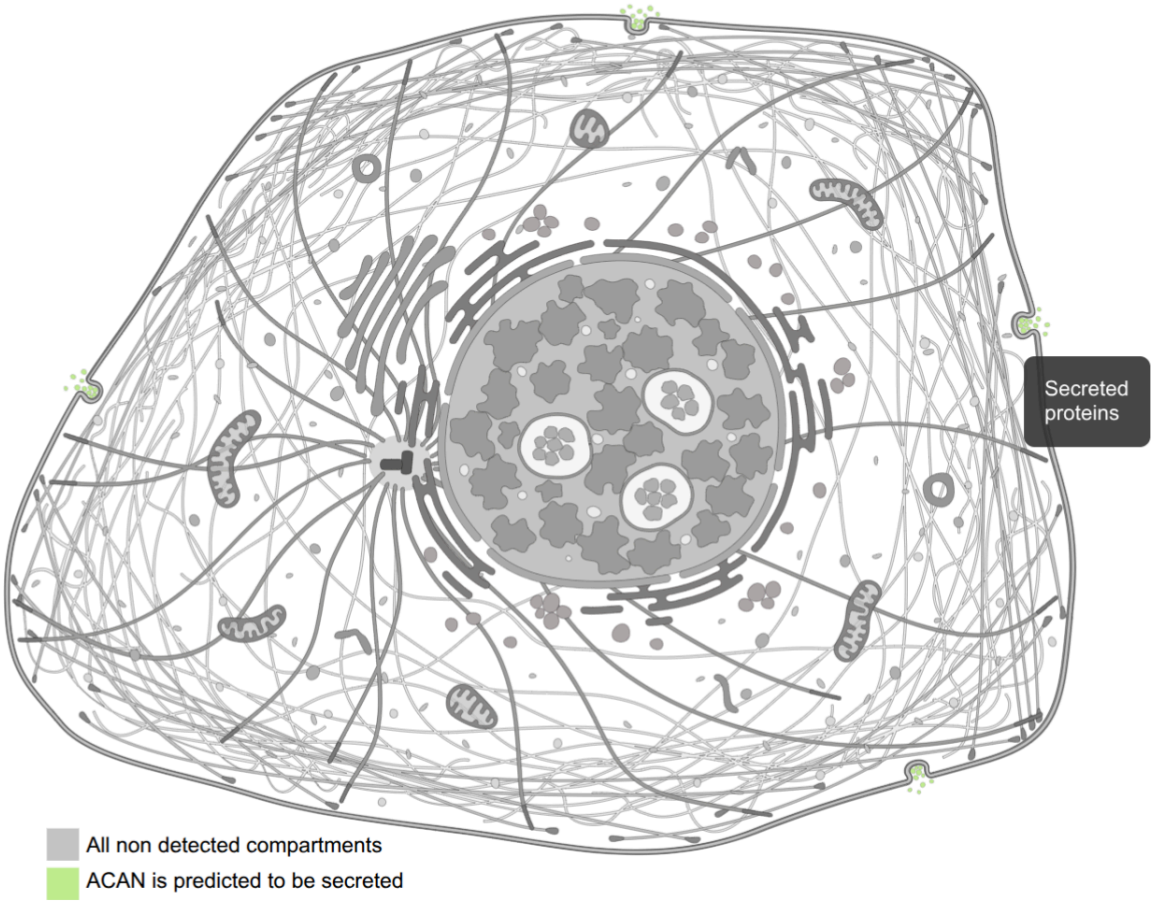

Рисунок 6. Предсказанная субклеточная локализация белка ACAN по данным Human Protein Atlas. ACAN отмечен как секретируемый белок. Такая локализация соответствует функции аггрекана как внеклеточного матриксного протеогликана.

На рисунке 6 показано, что белок ACAN предсказан как секретируемый. После синтеза и прохождения через секреторный путь он оказывается во внеклеточном матриксе, где выполняет структурную функцию.

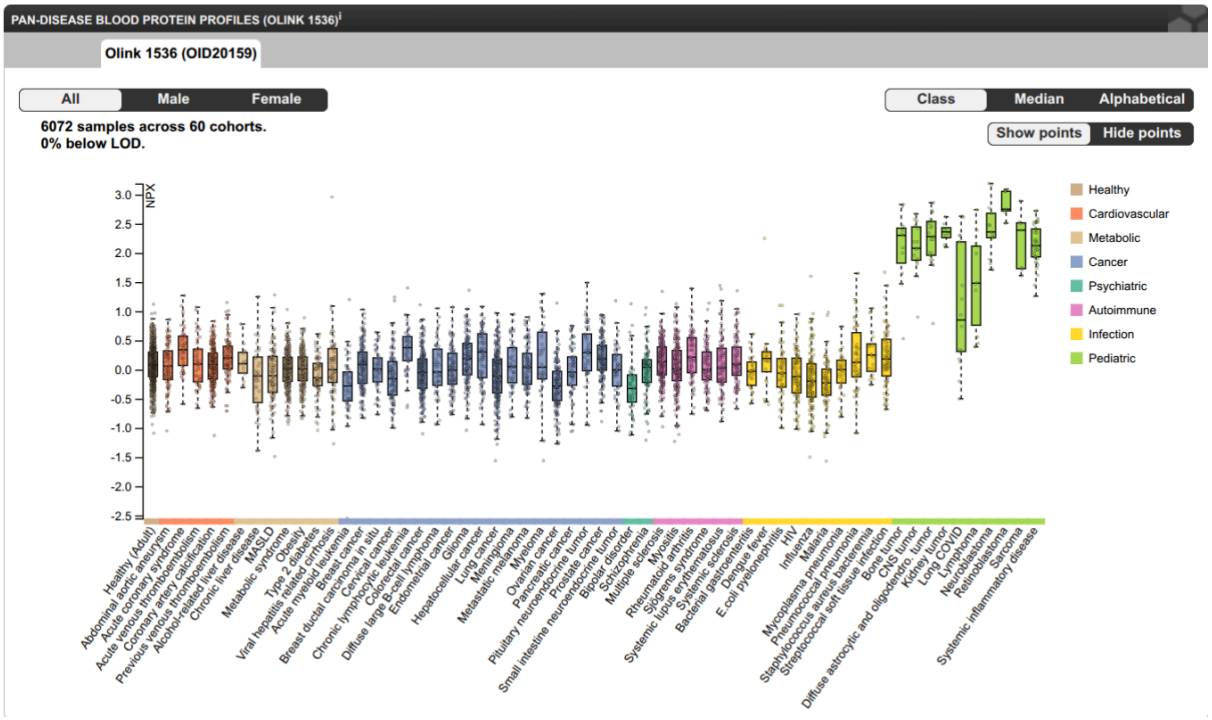

Рисунок 7. Уровень белка ACAN в крови при разных группах заболеваний по данным Human Protein Atlas и Olink. По оси Y показан относительный уровень белка, по оси X — здоровые образцы и группы заболеваний. Цвета обозначают классы заболеваний. Наиболее заметное повышение уровня ACAN наблюдается в педиатрических группах.

На рисунке 7 показан уровень белка ACAN в крови в разных группах здоровых людей и пациентов с различными заболеваниями. Наиболее выраженное отличие видно для педиатрических групп: в этих группах уровень ACAN заметно выше, чем у здоровых взрослых и большинства взрослых групп заболеваний.

Это наблюдение выглядит биологически осмысленным. ACAN кодирует аггрекан — протеогликан внеклеточного матрикса, особенно важный для хрящевой и соединительной ткани. У детей активно идут процессы роста и развития скелета, поэтому повышенный уровень ACAN в крови может отражать более активный обмен компонентов внеклеточного матрикса. При этом по одному такому графику нельзя быть уверенным, что это связано именно с возрастом и ростом, так как это может быть связано с особенностями конкретных педиатрических заболеваний.

Связь ACAN с результатами группового анализа

В предыдущем пункте по STRING общая тема списка была определена как метаболизм гликозаминогликанов. ACAN хорошо иллюстрирует эту тему, потому что он кодирует сам протеогликан — аггрекан, то есть непосредственно связан с ГАГ.

Данные Human Protein Atlas хорошо поддерживают такую интерпретацию. Тканевая экспрессия ACAN связана с соединительными тканями, белок предсказан как секретируемый, а значит, его функция действительно должна быть связана с внеклеточной средой. Это согласуется с результатами STRING, где список оказался связан с ГАГ, аппаратом Гольджи и внеклеточным матриксом.

Выводы

Таким образом, если групповой анализ STRING показал общую функциональную тему всего списка (гликозаминогликаны), то анализ ACAN в Human Protein Atlas показывает конкретный пример белка, который реализует эту тему на уровне ткани и клеточной локализации, т.е. хорошо вписывается в общую тему.