1.Описание входных данных

В качестве входных данных у нас имеется текстовый файл со списоком из 11 генов человека.
Ссылка: https://kodomo.fbb.msu.ru/FBB/year_24/lectures/lists_go/list14.txt
Вот наши гены с описанием получающихся с них белков:
VCAN - белок входит в состав внеклеточного матрикса тканей, включая глаза, кожу и сердце.
BCAN - специфический для головного мозга протеогликан хондроитинсульфата, играющий решающую роль в стабилизации структуры внеклеточного матрикса в центральной нервной системе
DSE - фермент, который превращает D-глюкуроновую кислоту в L-идуроновую кислоту, необходимую для синтеза компонента внеклеточного матрикса.
UST - белок, связан с заболеваниями молочной железы и острым некроз сетчатки.
DSEL - связан с синдром Эйкена и дисплазией Шнекенбекена.
NCAN - протеогликан хондроитинсульфата, находящийся во внеклеточном матриксе головного мозга и необходимый для адгезии нейронных клеток, роста нейритов и синаптической пластичности.
BGN - бигликан, небольшой богатый лейцином протеогликан, необходимый для структуры внеклеточного матрикса, сборки коллагеновых фибрилл, роста костей и клеточной сигнализации.
CHST14 - фермент, который необходим для образования важнейшего гликозаминогликана в соединительной ткани.
CSPG5 - специфический для головного мозга трансмембранный белок, имеющий решающее значение для развития нервной системы, в частности для роста нейритов, ветвления дендритов и синаптической пластичности.
CSPG4 - трансмембранный протеогликан, который выступает в качестве ключевого регулятора пролиферации, миграции и ангиогенеза клеток.
DCN - протеогликан, богатый лейцином, имеющий решающее значение для сборки коллагеновых фибрилл и структурной целостности внеклеточного матрикса.

Думаю все эти гены связаны между собой внеклеточным матриксом. А еще ассоциированы с разными видами рака.

2. Групповой сервис STRING

Данным сервисом можно решать следующие задачи:

- Анализировать белок-белковые взаимодействия
- Анализировать коэкспрессию генов из исследуемого списка
- Исследовать обогащение по GO терминам и KEGG путям

Покопавшись в интернете, нашли, что сервис использует тестирование Фишера для оценки статистической значимости обогащения функциональных категорий(те проверяет не произошло ли пересечение списка белков с известными путями случайным образом); гипергеометрический тест (для оценки того, насколько сильно определенные

биологические пути или функции обогащены в наборе генов по отношению к фоновому распределению); тест Колмогорова-Смирнова (при загрузке полногеномных данных, чтобы обнаружить значимые пути со смещенным распределением в заданном списке значений). Сервис также делает поправку на множественное тестирование, что следует из наличия False Discovery Rate по методу Бенджамини - Хохберга (BH FDR). Это популярный статистический алгоритм, который используется для корректировки p-значений при одновременной проверке множества гипотез. Он применяется для того, чтобы избежать ложных открытий (ошибок I рода), вызванных случайными совпадениями при проведении большого количества тестов.

Удобная ссылка на таблицу находок:
https://string-db.org/cgi/network?taskId=bclG12ktsEzu&sessionId=bx4iUL7Rvo4h

Неудобная скачанная и занесенная в Google sheets таблица находок:
Таблица находок

Если говорить о проведенном GO-анализе, то по биологическим процессам было получено шесть терминов. По молекулярным функциям пять терминов, а по клеточным компонентам шесть. KEGG путь был получен всего один - биосинтез гликозаминогликанов (Рис 1).

Biological Process (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0019800	Peptide cross-linking via chondroitin 4-sulfate glycosaminoglycan	2 of 6	2.78	1.19	0.0069
GO:0006790	Sulfur compound metabolic process	6 of 336	1.5	1.79	1.60e-05
GO:0030204	Chondroitin sulfate metabolic process	4 of 31	2.36	2.76	4.75e-06
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

(less ...)

Molecular Function (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
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
GO:0005540	Hyaluronic acid binding	3 of 30	2.25	1.57	0.0011
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

Cellular Component (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0000139	Golgi membrane	5 of 664	1.13	0.87	0.0030
GO:0062023	Collagen-containing extracellular matrix	5 of 407	1.34	1.29	0.00030
GO:0031012	Extracellular matrix	6 of 552	1.29	1.4	4.67e-05
GO:0005794	Golgi apparatus	11 of 1650	1.08	1.45	1.01e-09
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

(less ...)

Рис 1. Результаты GO-анализа, выполненного с использованием сервиса STRING, демонстрирующие распределение исследуемых генов по категориям биологических процессов (Biological Processes), молекулярных функций (Molecular Functions) и клеточных компонентов (Cellular Components), отражающих их предполагаемую функциональную роль и участие в клеточных механизмах.

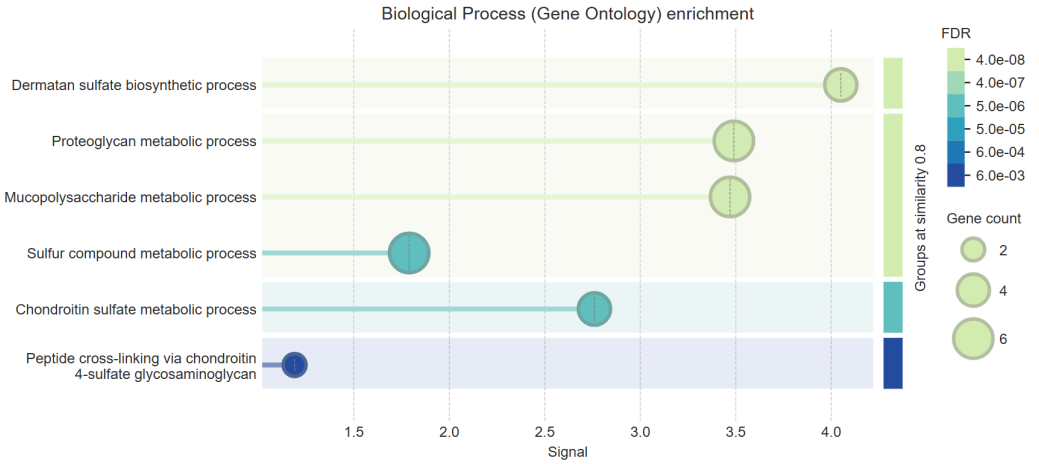


Рис 2. Визуализация находок GO анализа по биологическим процессам. Цвет точек отображает значение метрики FDR, а размер - количество генов, участвующих в соответствующем биологическом процессе.

Рассмотрим значения FDR на (Рис 2). Самые хорошие значения у дерматан-сульфат биосинтетического, протеингликанового и микополисахаридного метаболических процессов.Молекулярные функции(Рис1) соответствующие: поддержание структуры внеклеточного матрикса и связывание гликозаминогликана. Представлены наши белки в люмине клеток(рис 1), что логично.

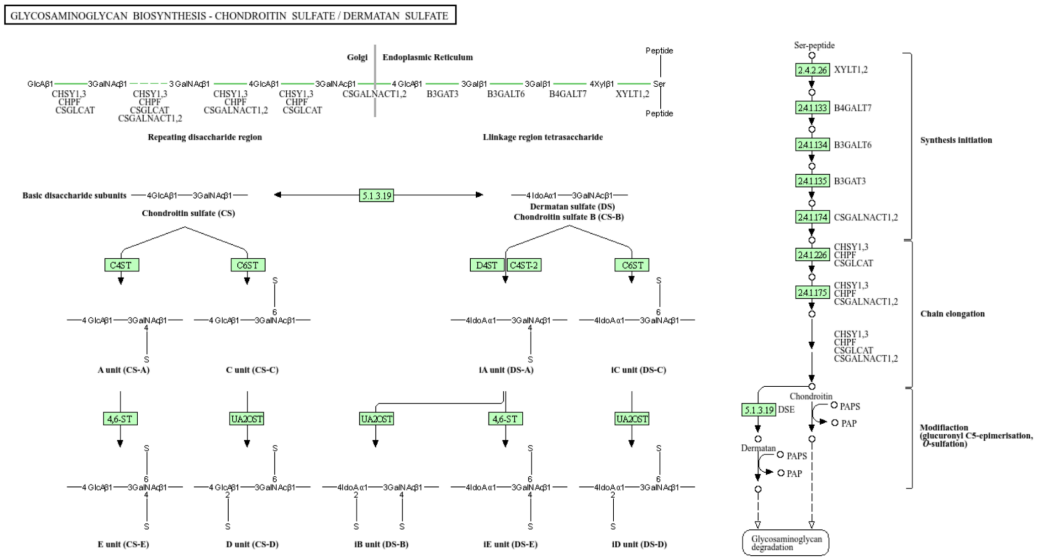


Рис 3. Единственный KEGG путь биосинтеза гликозаминогликанов, полученный сервисом STRING для нашего списка генов.

Обогащение пути биосинтеза гликозаминогликанов (chondroitin sulfate / dermatan sulfate), изображенное на (рис 3) обусловлено присутствием генов, кодирующих протеогликаны внеклеточного матрикса (VCAN, BCAN, NCAN, CSPG4/5, BGN, DCN), а также ферменты их постсинтетической модификации (DSE, DSEL, CHST14, UST).

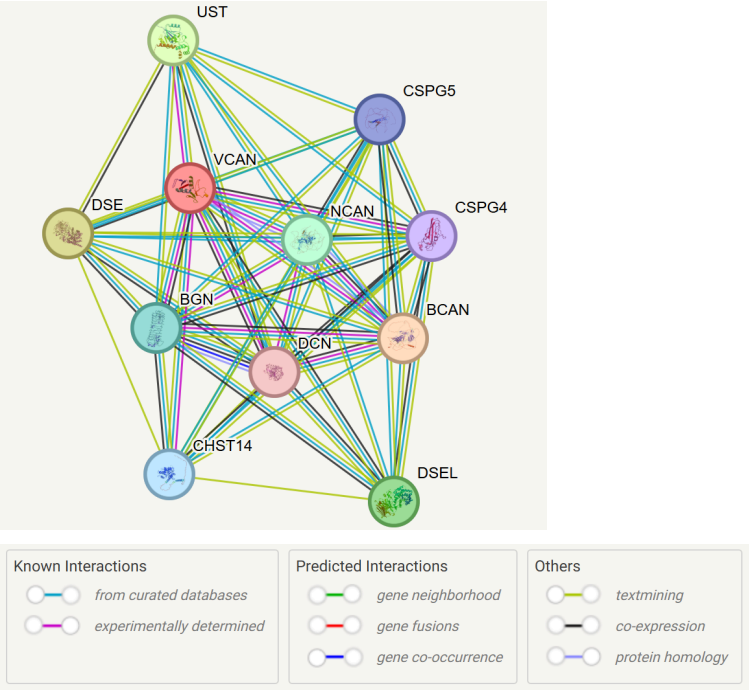


Рис 4. Визуализация взаимодействий внутри нашей выборки белков.

На (Рис 4) представлена плотная сеть взаимодействий. Большое количество розовых и голубых линий свидетельствует об экспериментальной подтвержденности взаимодействий. Салатовые линии показывают, что белки совместно упоминаются в статьях достаточно часто, чтобы это не было случайной находкой. Это подтверждает нашу находку KEGG путя, в котором участвуют все исследуемые гены.

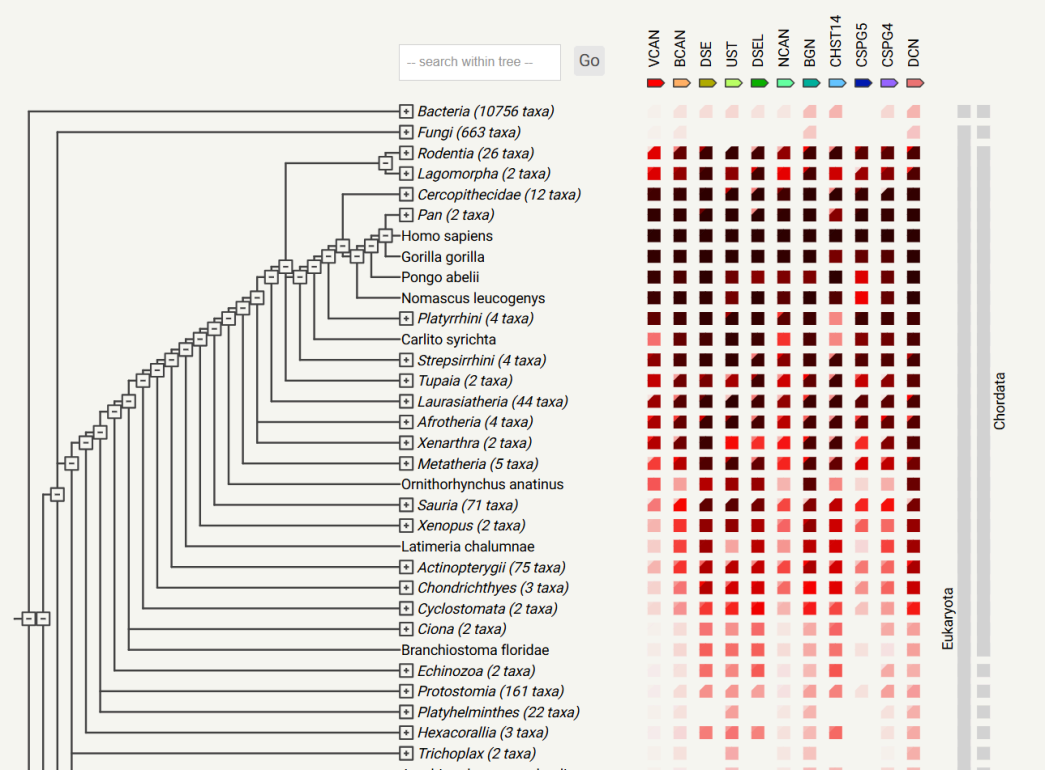


Рис 5. Семейства генов, закономерности встречаемости которых в разных геномах демонстрируют сходство. Чем темнее окраска, тем выше представленность семейства генов в данной группе организмов

Исследуемые гены консервативны и представлены у всех млекопитающих (интересно, что похожие по последовательности находки есть даже у грибов). В наибольшей степени они сходны у приматов, что видно по черной окраске на (Рис 5). Это отражает эволюционную важность метаболического пути, в котором данные гены имеют ключевую роль.

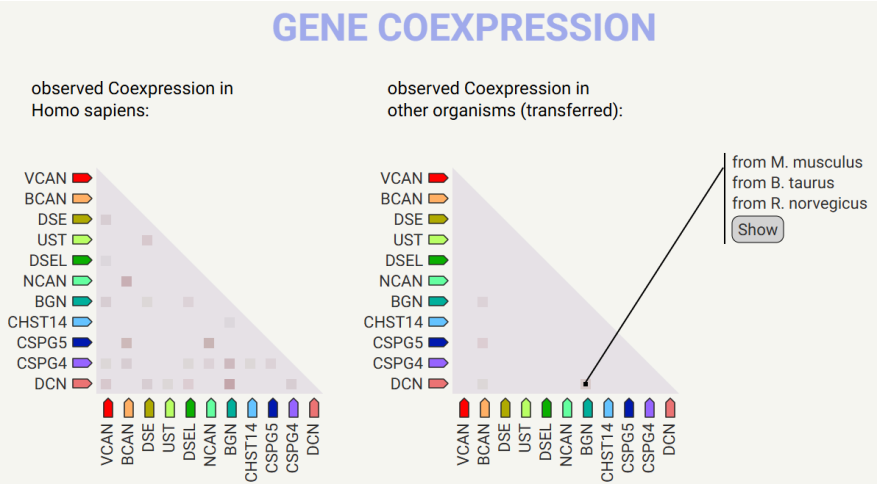


Рис 6. Коэкспрессия наших генов в людях и других организмах. Чем насыщеннее окраска, тем выше уровень коэкспрессии.

При том, что наши гены участвуют в одном метаболическом пути, уровень их коэкспрессии низок, что видно из (Рис 6). Не уверена, что так должно быть, возможно сервис по этой задаче обладает слишком маленьким количеством информации или эти белки работают последовательно.

Мы продемонстрировали, что исследуемые гены функционально объединены участием в формировании внеклеточного матрикса и метаболизме гликозаминогликанов, а нарушения их экспрессии могут быть связаны с опухолевыми и другими патологиями

3.Индивидуальный сервис Human Protein Atlas

Данным сервисом можно решать следующие задачи:

- Определять экспрессию белков во всех основных органах и тканях человека
- Изучать влияние уровней экспрессии белков на выживаемость пациентов с различными видами рака
- Анализировать экспрессию генов на уровне отдельных клеток (single-cell analysis), чтобы определить, в каких типах клеток экспрессируется исследуемый ген.

Будем работать с BCAN. Из результатов прошлого раздела видно, что этот ген кодирует протеогликан внеклеточного матрикса(обсуждение под рис 3). В пути биосинтеза гликозаминогликанов, полученном в прошлом пункте, он влияет на связывание воды, обеспечение структурной жесткости, участвует в регуляции воспалений.

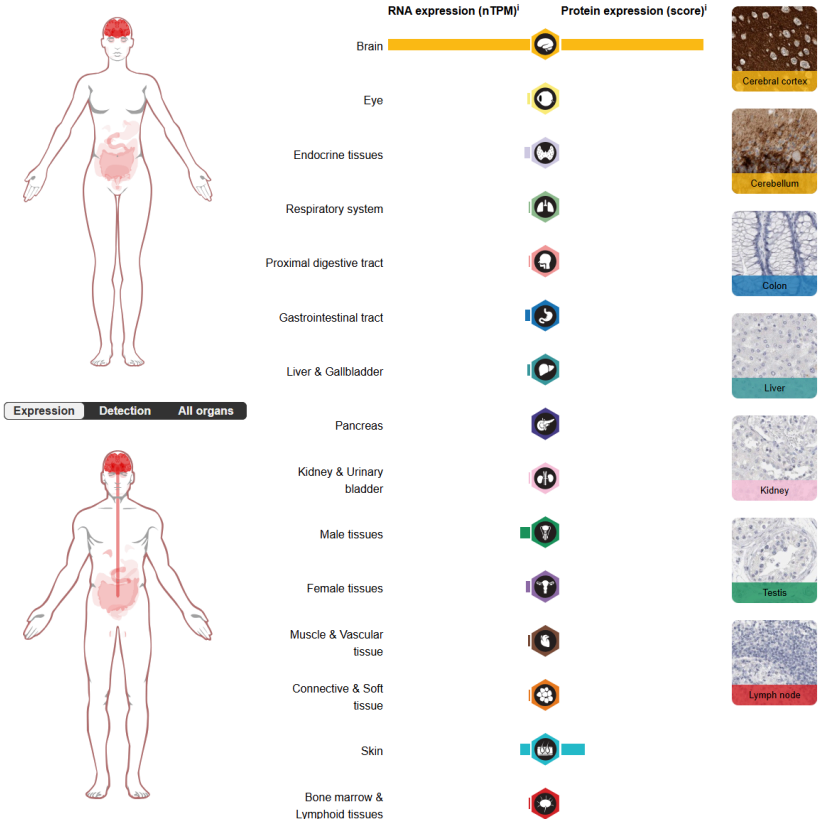


Рис 7. РНК и белковая коэкспрессия BCAN в разных частях тела человека. Левые столбцы относительно иконок органов соответствуют РНК экспрессии, правые - белковой.

По (Рис 7) бесспорный чемпион по РНК и белковой экспрессии - ЦНС. По схеме с манекенами видно, что у женщин высокая экспрессия, в отличие от мужчин,

наблюдается именно в головном мозге. Это может быть связано с тем, что выборка женщин, как правило, во многих исследованиях меньше выборки мужчин, и для них просто не хватает данных. Небольшой уровень экспрессии РНК наблюдается в тестикулах, яичниках, а также в коже.

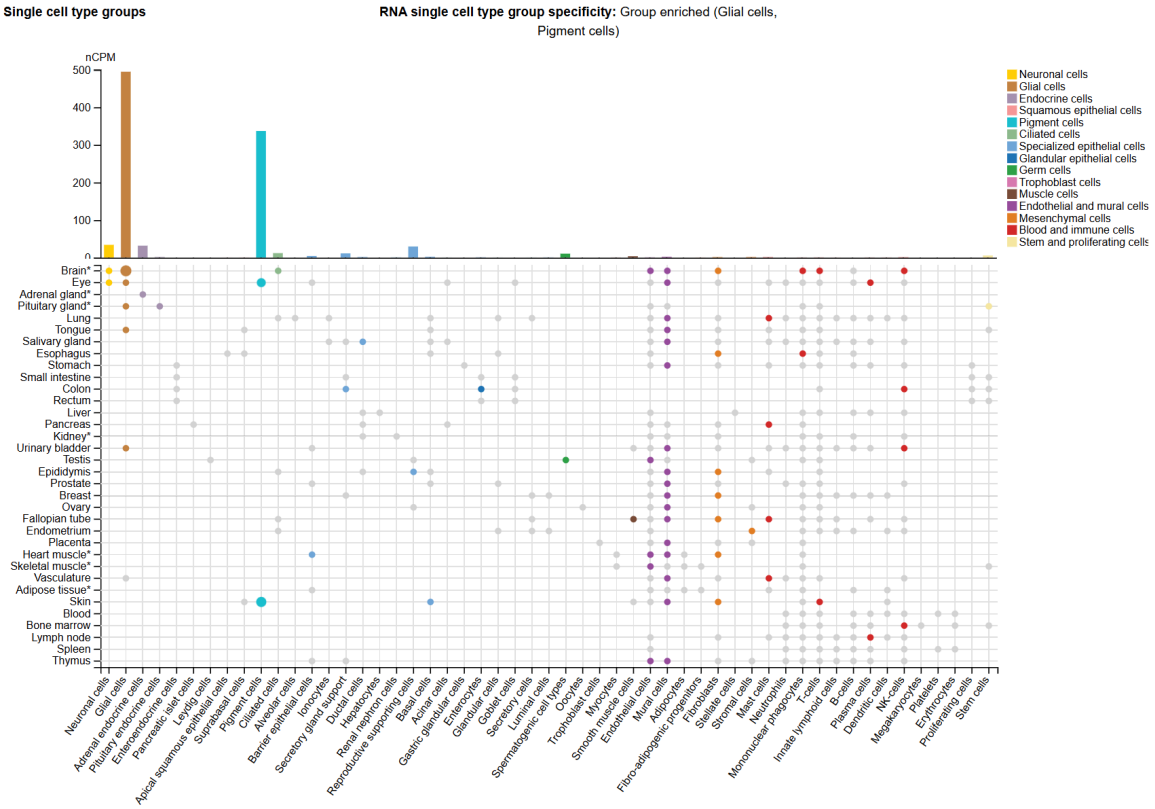
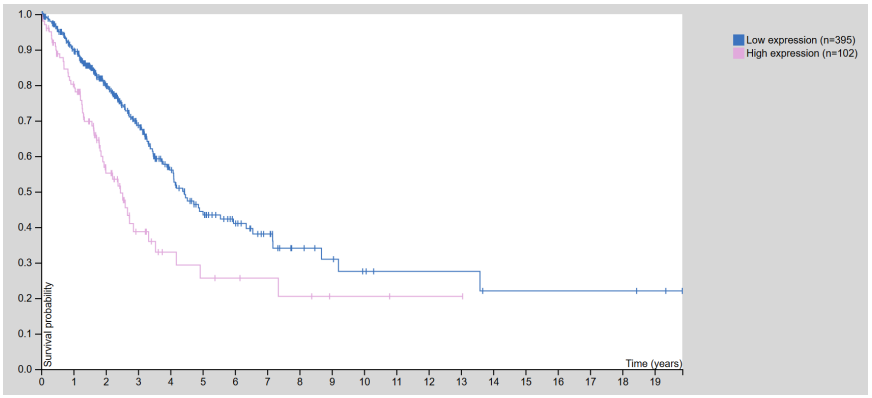


Рис 8. Обогащение экспрессии синглсельных групп в разных органах.
Из (Рис 8) видим, что обогащены глиальные и пигментные клетки. Глия потому что мозг, пигментные потому что кожа. Еще точек много для mural cells(выстилают поверхность кровяных сосудов), но они размером поменьше, поэтому насыщение низкое.



BCAN is potential prognostic, high expression is unfavorable in Lung Adenocarcinoma (TCGA)

Рис 9. Вероятность выжить с одним из видов рака легких в зависимости от уровня экспрессии BCAN с течением времени. Синяя линия отражает график вероятности для низкой экспрессии, розовая - для высокой.
Это мой любимый график в этом практикуме. BCAN обладает потенциальным прогностическим значением, высокая экспрессия неблагоприятна при аденокарциноме

легких (TCGA). Видим из (Рис 9), что вероятность подольше пожить с раком при высокой экспрессии нашего гена значительно ниже. Все потому что он влияет на синтез гликозаминогликанов. Гликозаминогликаны - это высокогидрофильные углеводные компоненты межклеточного матрикса, которые удерживают воду, обеспечивая тургор тканей, формируют структурный каркас соединительной ткани, выступают смазочным материалом в суставах, регулируют ионный обмен, свертывание крови и клеточную пролиферацию. Повышенный синтез протеогликанов делает матрикс более пластичным и удобным для инвазии.

Human Protein Atlas показал тканеспецифичную экспрессию BCAN, преимущественно в мозге, а также связь повышенной экспрессии гена с ухудшением прогноза при раке лёгких. Это подтверждает важную роль BCAN во внеклеточном матриксе и патологических процессах.

4.Выводы

В ходе работы были проанализированы гены, связанные с компонентами внеклеточного матрикса и метаболизмом гликозаминогликанов. С помощью сервиса STRING было показано, что исследуемые белки образуют тесно связанную сеть взаимодействий и участвуют преимущественно в процессах организации внеклеточного матрикса, биосинтеза дерматан- и хондроитинсульфатов. Также были выявлены значимые GO-термины и KEGG-путь, подтверждающие функциональное единство исследуемого набора генов.

Анализ гена BCAN в Human Protein Atlas показал его преимущественную экспрессию в ЦНС и отдельных типах клеток, прежде всего глиальных. Кроме того, повышенная экспрессия BCAN оказалась связана с неблагоприятным прогнозом при аденокарциноме лёгких, что указывает на его возможную роль в опухолевых процессах.

Таким образом, оба использованных ресурса дополняют друг друга и позволяют получить представление как о функциональных взаимодействиях генов и белков, так и об их тканевой и клеточной специфичности, а также возможной связи с заболеваниями человека.