Отчет по практикуму 6
Сорокин Илья

Этап 1. Описание входного списка данных

В качестве входных данных использовался файл (list44.txt) со списком символов генов человека в номенклатуре HGNC. Общее количество генов в списке составляет 39 штук. Сразу по этому списку можно заметить большое количество генов с одинаковой мнемоникой (ARS…, GBA…, NEU…).

Задача практикума заключается в опробовании различных возможностей баз данных с использованием предоставленного перечня генов.

Этап 2. База данных STRING

Для первой работы была выбрана база данных STRING, которая специализируется на белковых взаимодействиях и привлекла меня своей универсальностью. Помимо анализа белок-белковых взаимодействий, этот ресурс предоставляет широкие возможности: позволяет проводить анализ обогащения по Gene Ontology, KEGG и другим базам, а также обращаться к публикациям из PubMed.

При работе с базой данных STRING в разделе поиска по полю Multiple Proteins в качестве запроса был подан полный список мнемонических обозначений генов человека. Остальные параметры поиска использовались по умолчанию.

На выходе STRING формирует детализированную карту белок-белковых взаимодействий. Узлы графа соответствуют белкам и соединены ребрами, отражающими взаимодействия между ними. При этом цвета линий указывают на уровень достоверности каждого конкретного взаимодействия (то есть как было показано, что данные белки взаимодействуют). Полученная схема белок-белковых взаимодействий исследуемых белков представлена на рис. 1.

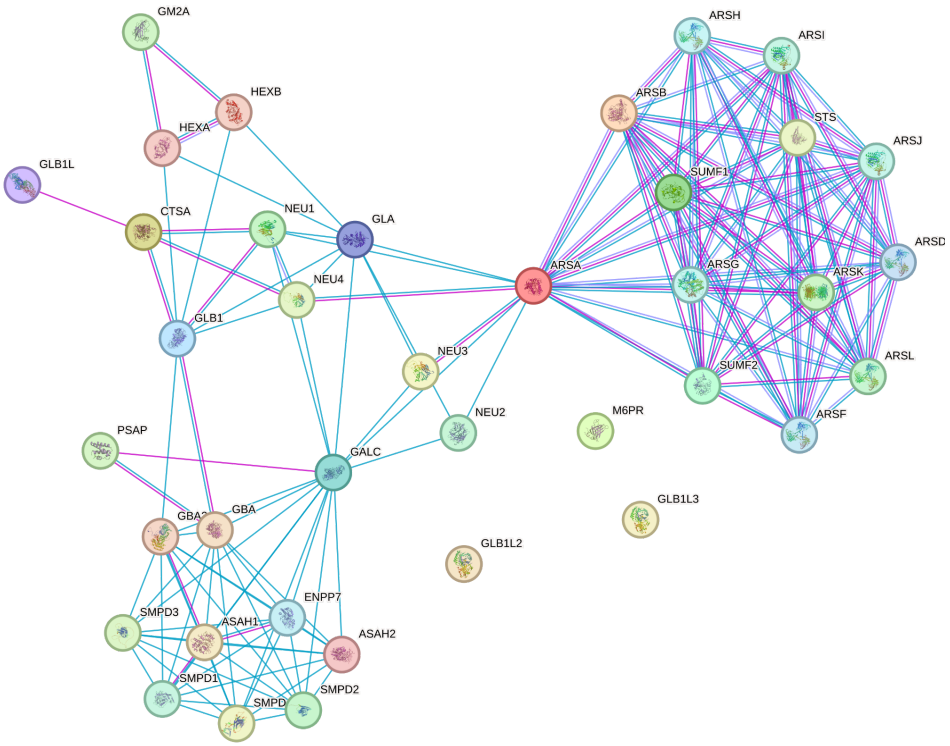
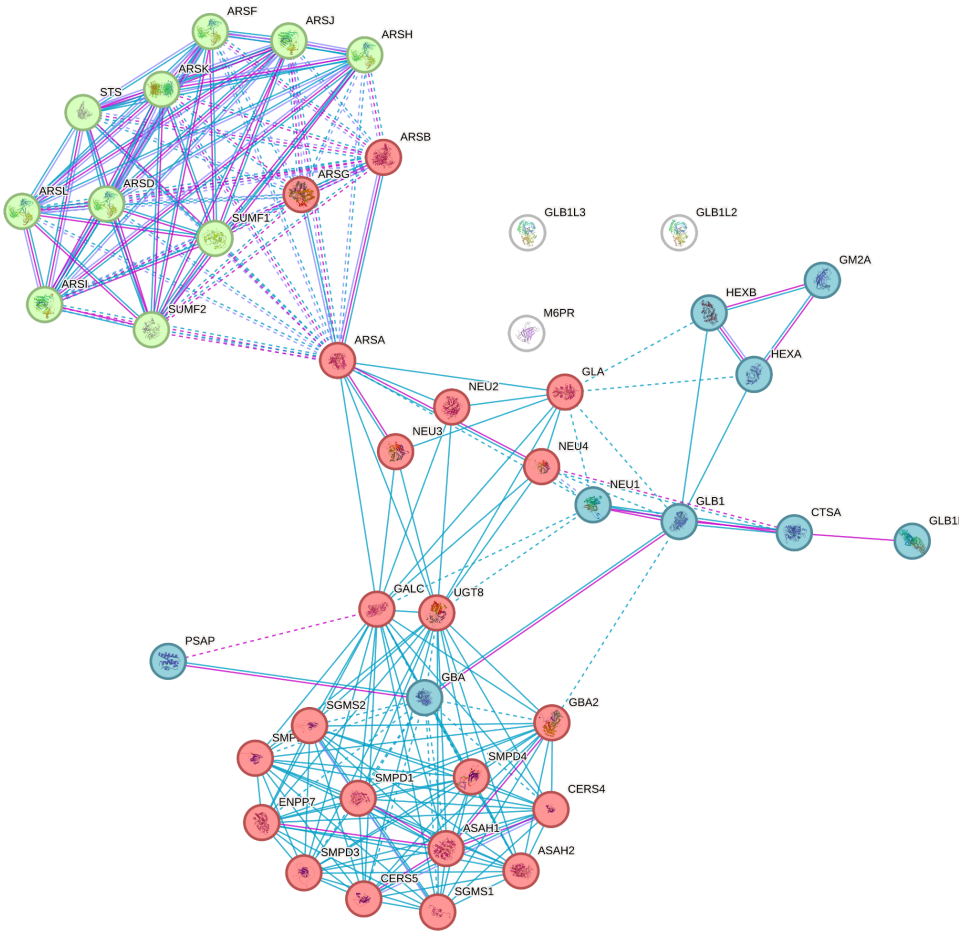


Рис. 1. Белок-белковые взаимодействия продуктов исследуемых генов в базе данных STRING. Для надежности данных на рисунке отражены только те взаимодействия, которые задокументированы в курируемых базах данных (синии линии) либо подтверждены экспериментально (фиолетовые линии).

Как можно видеть по рис. 1, в сети белок-белковых взаимодействий присутствуют как белки, которые много с кем взаимодействуют (белки с мнемониками ARS, SMPD и другие), так и три белка, не имеющие связей с другими узлами в данной сети (GLB1L3, GLB1L2, M6PR), которые аннотированны как Бета-галактозидаза-1-подобный белок 2, Бета-галактозидаза-1-подобный белок 3 и Катион-зависимый рецептор маннозо-6-фосфата соответственно.

В STRING есть прекрасная возможность разделить исследуемые белки на определенное количество кластеров. Я решил разделить исследуемые белки на 3 кластера, используя метод k-средних. Результат представлен на рис. 2.

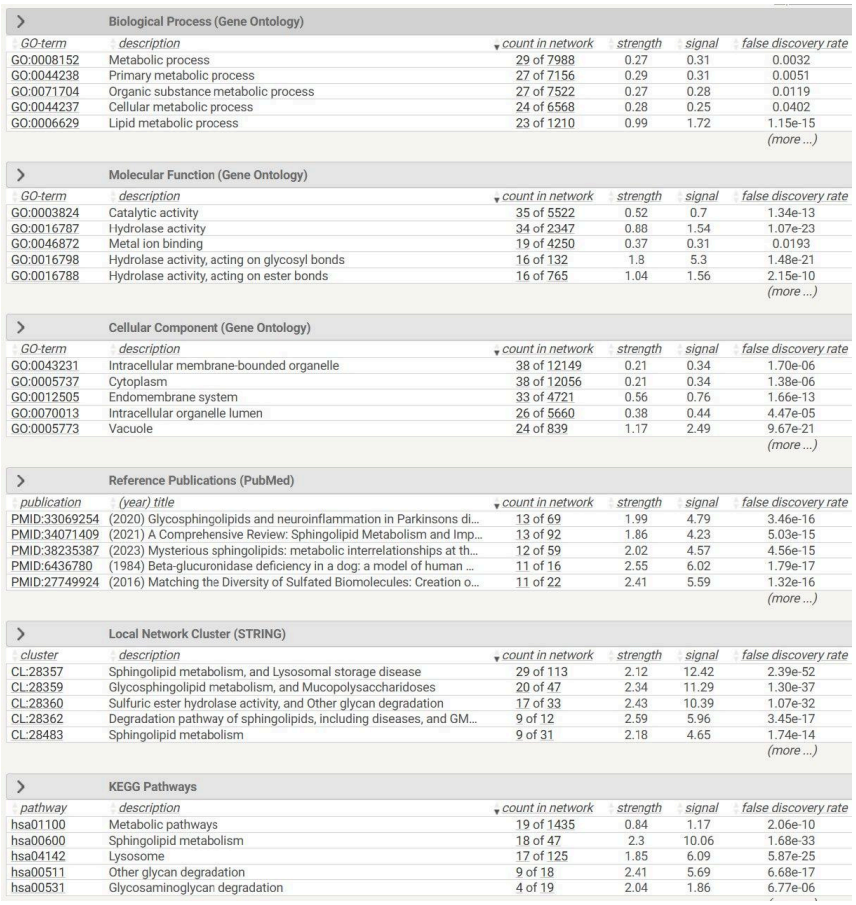


color	cluster Id	gene count	description
	Cluster 1	21	Sphingolipid metabolism
	Cluster 2	10	The activation of arylsulfatases
	Cluster 3	9	Glycosphingolipid metabolism

Рис. 2. Белок-белковые взаимодействия продуктов исследуемых генов в базе данных STRING по разделению на 3 кластера. Легенда изображена под графом.

По результатам на рис. 2 видно, что мы имеем дело с генами метаболизма сфинголипидов и гликосфинголипидов, а также четверть генов ответственны за активацию арилсульфатаз.

Продолжаем разбираться дальше. Мне стало интересно, какие гены входят в исследуемый набор и что они из себя в целом представляют. Данную информацию можно получить 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.



Biological Process (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0008152	Metabolic process	29 of 7988	0.27	0.31	0.0032
GO:0044238	Primary metabolic process	27 of 7156	0.29	0.31	0.0051
GO:0071704	Organic substance metabolic process	27 of 7522	0.27	0.28	0.0119
GO:0044237	Cellular metabolic process	24 of 6568	0.28	0.25	0.0402
GO:0006629	Lipid metabolic process	23 of 1210	0.99	1.72	1.15e-15
					(more ...)

Molecular Function (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0003824	Catalytic activity	35 of 5522	0.52	0.7	1.34e-13
GO:0016787	Hydrolase activity	34 of 2347	0.88	1.54	1.07e-23
GO:0046872	Metal ion binding	19 of 4250	0.37	0.31	0.0193
GO:0016798	Hydrolase activity, acting on glycosyl bonds	16 of 132	1.8	5.3	1.48e-21
GO:0016788	Hydrolase activity, acting on ester bonds	16 of 765	1.04	1.56	2.15e-10
					(more ...)

Cellular Component (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0043231	Intracellular membrane-bounded organelle	38 of 12149	0.21	0.34	1.70e-06
GO:0005737	Cytoplasm	38 of 12056	0.21	0.34	1.38e-06
GO:0012505	Endomembrane system	33 of 4721	0.56	0.76	1.66e-13
GO:0070013	Intracellular organelle lumen	26 of 5660	0.38	0.44	4.47e-05
GO:0005773	Vacuole	24 of 839	1.17	2.49	9.67e-21
					(more ...)

Reference Publications (PubMed)

publication	(year) title	count in network	strength	signal	false discovery rate
PMID:33069254	(2020) Glycosphingolipids and neuroinflammation in Parkinsons di...	13 of 69	1.99	4.79	3.46e-16
PMID:34071409	(2021) A Comprehensive Review: Sphingolipid Metabolism and Imp...	13 of 92	1.86	4.23	5.03e-15
PMID:38235387	(2023) Mysterious sphingolipids: metabolic interrelationships at th...	12 of 59	2.02	4.57	4.56e-15
PMID:6436780	(1984) Beta-glucuronidase deficiency in a dog: a model of human ...	11 of 16	2.55	6.02	1.79e-17
PMID:27749924	(2016) Matching the Diversity of Sulfated Biomolecules: Creation o...	11 of 22	2.41	5.59	1.32e-16
					(more ...)

Local Network Cluster (STRING)

cluster	description	count in network	strength	signal	false discovery rate
CL:28357	Sphingolipid metabolism, and Lysosomal storage disease	29 of 113	2.12	12.42	2.39e-52
CL:28359	Glycosphingolipid metabolism, and Mucopolysaccharidoses	20 of 47	2.34	11.29	1.30e-37
CL:28360	Sulfuric ester hydrolase activity, and Other glycan degradation	17 of 33	2.43	10.39	1.07e-32
CL:28362	Degradation pathway of sphingolipids, including diseases, and GM...	9 of 12	2.59	5.96	3.45e-17
CL:28483	Sphingolipid metabolism	9 of 31	2.18	4.65	1.74e-14
					(more ...)

KEGG Pathways

pathway	description	count in network	strength	signal	false discovery rate
hsa01100	Metabolic pathways	19 of 1435	0.84	1.17	2.06e-10
hsa00600	Sphingolipid metabolism	18 of 47	2.3	10.06	1.68e-33
hsa04142	Lysosome	17 of 125	1.85	6.09	5.87e-25
hsa00511	Other glycan degradation	9 of 18	2.41	5.69	6.68e-17
hsa00531	Glycosaminoglycan degradation	4 of 19	2.04	1.86	6.77e-06

Рис. 3. Устройство выдачи поля «Analysis» по исследуемым 39 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.

Также стоит упомянуть про то, что такое FDR, или false discovery rate, а также strength. В своих анализах STRING выполняет статистическую проверку для каждого термина внутри заданной коллекции путей (категории). STRING использует поправку Бенджамини — Хохберга для учета множественного тестирования гипотез [1]. Данная статистическая коррекция необходима для эффективного контроля уровня ложно-положительных результатов. Strength представляет собой десятичный логарифм отношения наблюдаемого количества белков к ожидаемому. То есть чем выше strength, тем сильнее обогащение представлено в вашем наборе по сравнению со случайным ожиданием.

Далее я буду приводить в гиперссылках результаты обогащения терминов по разным баз данных в виде tsv-таблиц, отсортированные по FDR.

Анализ обогащения по категории GO Biological Process показал, что исследуемые гены преимущественно вовлечены в какие-то метаболические процессы (GO:0008152), причем наиболее представленный термин — «липидный метаболический процесс» (GO:0006629), для которого 23 белка из сети аннотированы данным термином (FDR = 1.15e-15, strength = 0.99). Другие метаболические термины также присутствуют, например, самыми надежными терминами (с самым низким FDR) — GO:0030149 (Sphingolipid catabolic process), GO:0046514 (Ceramide catabolic process), GO:0046479 (Glycosphingolipid catabolic

process). Полученные данные однозначно свидетельствуют о том, что исследуемые белки играют центральную роль в процессах распада именно этих классов липидов.

На рис. 4 изображен график обогащения терминами GO Biological Process списка генов.

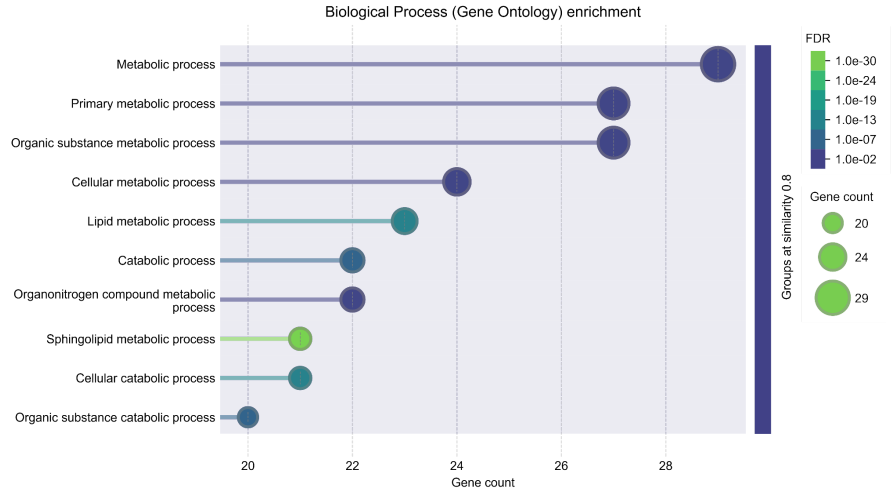


Рис. 3. Визуализация GO-обогащения Biological Process. Сортировка происходила по количеству генов в категории (gene count). По оси Y изображены GO-категории, по оси X — количество генов в данных категориях. Цветом обозначено значение FDR, размером точек — количество генов в категории.

Далее мне стало интересно, в каких клеточных компартментах данные белки наиболее представлены. Анализ категории Cellular Component выявил выраженную локализацию исследуемых белков в системе внутренних мембран то есть в лизосомах, ЭПР и аппарате Гольджи, (GO:0043202, Lysosomal lumen, FDR = 2.19e-21; GO:0005773, Vacuole, FDR = 9.67e-21; GO:0005788, Endoplasmic reticulum lumen, FDR = 5.64e-12). Такой результат мне показался довольно логичным, поскольку большинство ферментов деградации липидов функционирует именно в этих компартментах. На рис. 4 представлена визуализация обогащения категориями GO Cellular Component.

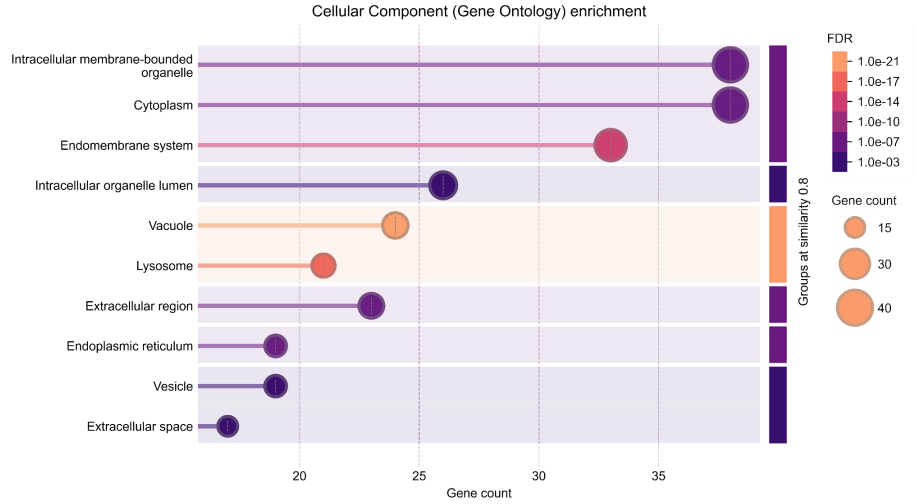


Рис. 4. Визуализация GO-обогащения Cellular Compartment. Подписи к рисунку такие же, как к рис. 3.

Согласно анализу обогащения по базе данных DISEASES, наиболее значимой категорией оказались «липидные болезни накопления» (Lipid storage disease,

DOID:9455, FDR = 9.18e-29): 18 белков из сети аннотированы этим термином при всего 80 белках в фоне. Также представлен термин «сфинголипидозы» (Sphingolipidosis, DOID:1927, FDR = 3.26e-26): 14 белков из 27 в фоне. Среди конкретных заболеваний наиболее представлены ганглиозидозы (Gangliosidosis), галактозиалидоз (Galactosialidosis), метахроматическая лейкодистрофия (Metachromatic leukodystrophy), болезнь Гоше (Gauchers disease) и болезнь Краббе (Krabbe disease) — для каждого из этих заболеваний в сети обнаружено от 3 до 6 белков при крайне малом числе фоновых белков (от 3 до 11), что говорит о высокой специфичности ассоциации исследуемых генов с данными патологиями. В целом полученные результаты подтверждают, что анализируемые белки преимущественно вовлечены в наследственные болезни обмена веществ, особенно сфинголипидов и других липидов, с преимущественным поражением нервной системы. На рис. 5 визуализировано обогащение терминами базы данных DISEASES.

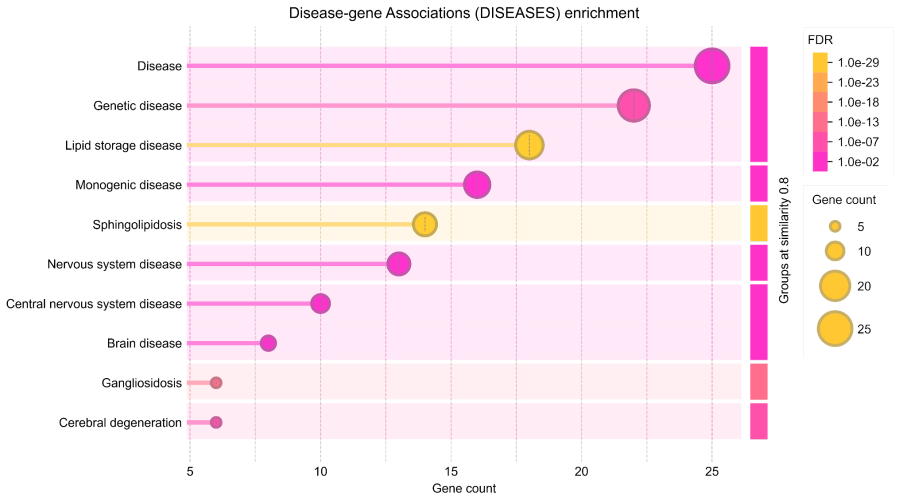


Рис. 5. Визуализация обогащения DISEASES. Подписи к рисунку такие же, как к рис. 3.

Таким образом можно сделать вывод, что исследуемый набор из 39 генов представляет собой группу генов, преимущественно вовлеченную в лизосомальный катаболизм сфинголипидов и гликосфинголипидов, а также вовлечены в связанные с ними заболевания. На самом деле мне показалось удивительным, что благодаря одному сервису можно выделить довольно много информации о генах, что является несомненным плюсом STRING.

Этап 3. Human Protein Atlas

В качестве следующей базы данных я решил 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 решил взять в рассмотрение ген **ArsB**. Белок данного гена удаляет сульфатные группы с хондроитин-4-сульфата (C4S) и регулирует его деградацию. Он участвует в регуляции клеточной адгезии и миграции; в центральной нервной системе является регулятором роста клеток, действуя через контроль уровня сульфатированных

гликозаминогликанов. Мне стало интересно, в каких тканях наблюдается наибольшая экспрессия этого белка, в каких клеточных структурах он находится и биомаркером каких раковых заболеваний он является.

На рисунке 6 изображена столбчатая диаграмма экспрессии РНК в различных тканях человека.

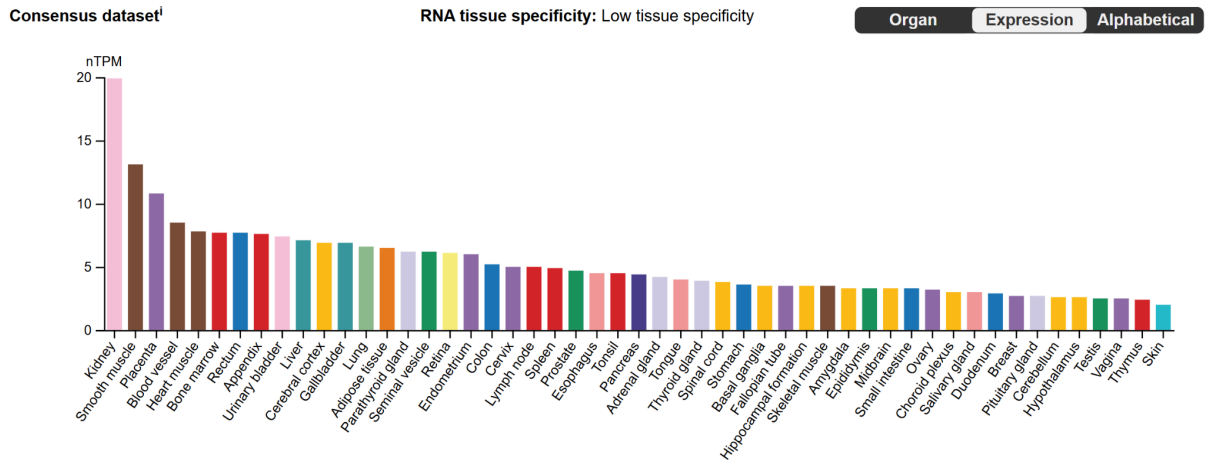


Рис. 6. Столбчатая диаграмма экспрессии РНК гена ArsB в различных тканях человека. По оси ординат отложено значение nTPM (нормализованное количество транскриптов, кодирующих белки, на миллион).

Значение nTPM, как написано на сайте, отражает количество транскриптов ArsB, нормализованное на один миллион общих транскриптов. Как я понял, это значение показывает, насколько активно конкретный ген экспрессируется в определенной ткани, клетке или органе. Чем выше nTPM, тем выше уровень экспрессии гена. Этот показатель позволяет сравнивать экспрессию разных генов как в пределах одного образца, так и между различными тканями. Как можно видеть по рис. 6., экспрессия гена ArsB обладает низкой тканеспецифичностью, но наибольшая экспрессия наблюдается в клетках печени.

Далее проанализируем клеточную локализацию выбранного белка. На рис. 7 зеленым цветом изображены компартменты клетки, в которых белок был детектирован.

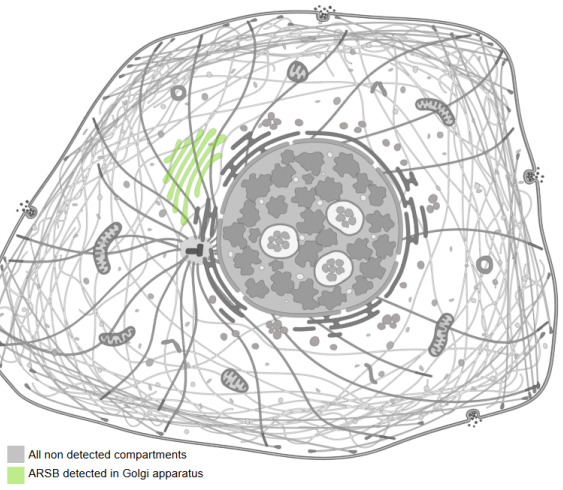


Рис. 7. Схематичное изображение клетки, на котором видно клеточную локализацию исследуемого белка ArsB (зеленый цвет).

Как видно по рисунку 4, исследуемый белок располагается в аппарате Гольджи. Данная локализация соответствует его функции, так как в аппарате Гольджи происходят процессы десульфатирования липидов.

На последок посмотрим, является ли исследуемый ген ArsB биомаркером каких-либо раковых опухолей. Данные представлены на рис. 8.

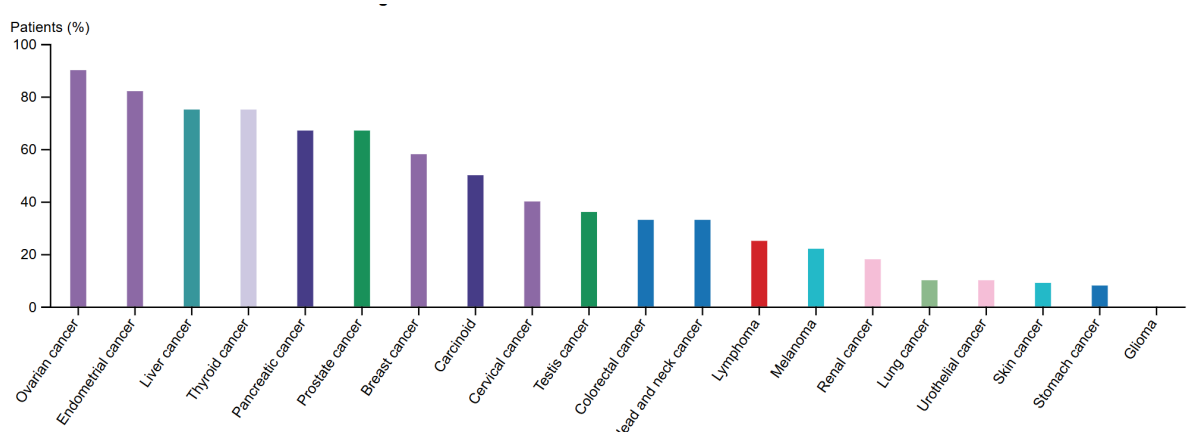


Рис. 8. Распределение уровней экспрессии белка по типам рака. По оси X — типы рака (по органам), по оси Y — процент пациентов с оверэкспрессией данного белка. Цветные столбцы показывают долю пациентов с высоким и средним уровнем экспрессии, белый столбец — с низким или недетектируемым уровнем

На основании рис. 8 можно заключить, что, к сожалению, ген ArsB не подходит на роль биомаркера типов рака: ArsB демонстрирует низкую сепцифичность к каким-то определенным типам рака и детектируется во многих из них. На самом деле это довольно логично, поскольку раннее мы узнали, что экспрессия гена ArsB не привязана к какому-либо одному органу или типу ткани.

Мне очень понравилась работа с этой базой данных особенно из-за обилия красивых картинок и большого количества информации, которую можно отсюда извлечь.

Список источников

1. Nucleic Acids Res. 2024 Nov 18;53(D1):D730–D737. doi: 10.1093/nar/gkae1113. The STRING database in 2025: protein networks with directionality of regulation Damian Szklarczyk et al.