Практикум №6

ПУНКТ №1

Для выполнения работы был использован файл List22.txt, содержащий список идентификаторов генов человека. Полный список доступен по ссылке.

Всего в анализируемый список входит 19 генов, большинство из них кодируют гликозилтрансферазы — ферменты, которые переносят остатки сахаров на молекулы акцепторы. Возможно все эти гены участвуют в биосинтезе гликосфинголипидов.

ПУНКТ №2a

Для группового анализа был выбран сервис DAVID (Database for Annotation, Visualization and Integrated Discovery).

Инструменты DAVID tools позволяют:

1. Определять обогащенные биологические темы — показывает, какие биологические процессы и пути KEGG статистически значимо встречаются в списке чаще, чем случайно.
2. Объединять избыточные термины в аннотации — DAVID группирует похожие GO-термины (удобно видеть одну общую тему, а не путаться во множестве одинаковых по смыслу формулировок)
3. Отображать взаимосвязи между генами и терминами в 2D — позволяет наложить гены на схему метаболического пути и увидеть, какие участки конвейера заняты

ПУНКТ №2b

Для определения общей «темы» списка генов был запущен анализ DAVID. Мы перешли на сайт DAVID, выбрали «Start Analysis», создали список наших генов и выбрали Functional Annotation Tool. Сервис DAVID использует статистический тест — точный тест Фишера. Этот тест позволяет ответить на вопрос: «Не случайно ли много наших генов оказались вместе в одной категории», то есть чем меньше p-value, тем более вероятно, что наши гены действительно участвуют вместе в каком-то процессе.

В качестве поправки на множественное тестирование использовался Benjamini. Эта поправка ужесточает p-value для более точного результата при проверке множества гипотез (в нашем случае получилось 47 находок (GO + KEGG) без кластеризации, то есть 47 гипотез о том, что каждый конкретный термин обогащен в нашем списке).

Для начала мы получили таблицу без кластеризации, чтобы увидеть более полную картину, описывающую наш список генов. Всего получилось 47 находок, 8 из них — KEGG pathways, остальные — термины GO (BP, CC, MF). Чтобы было удобнее разбираться в полученной информации, мы сфокусировались на KEGG-путях, так как они дают более конкретное представление о метаболических процессах.

После получения результатов (всего 8 находок) мы отсортировали их по значимости (p-value). Лучшей находкой оказался путь «Metabolic pathways», в котором участвуют все 19 генов. Однако это общее и широкое понятие, так что мы рассмотрели еще пару путей - Glycosphingolipid biosynthesis - ganglio series и Sphingolipid metabolism (по ссылке можно получить очень страшные картинки со множеством связей), p-value и оценка после поправки Бенджамини у них достаточно маленькие, что подтверждает статистическую значимость результатов.

Sublist	Category	Term	RT	Genes	Count	P-Value	Benjamini
☐	KEGG_PATHWAY	Metabolic pathways	RT	100.00%	19	9.12e-15	1.19e-13
☐	KEGG_PATHWAY	Glycosphingolipid biosynthesis - ganglio series	RT	36.84%	7	9.05e-14	5.88e-13
☐	KEGG_PATHWAY	Sphingolipid metabolism	RT	36.84%	7	5.02e-10	2.17e-9
☐	KEGG_PATHWAY	Glycosphingolipid biosynthesis - lacto and neolacto series	RT	31.58%	6	1.28e-9	4.15e-9
☐	KEGG_PATHWAY	Glycosphingolipid biosynthesis - globo and isoglobo series	RT	26.32%	5	1.62e-8	4.22e-8
☐	KEGG_PATHWAY	Glycosaminoglycan biosynthesis - keratan sulfate	RT	10.53%	2	2.62e-2	5.68e-2
☐	KEGG_PATHWAY	Mucin type O-glycan biosynthesis	RT	10.53%	2	6.61e-2	1.23e-1
☐	KEGG_PATHWAY	Ether lipid metabolism	RT	10.53%	2	9.25e-2	1.50e-1

Рис.1. Результаты анализа KEGG-путей, полученные с помощью сервиса DAVID.
Ссылка на результаты выдачи DAVID (KEGG):
https://davidbioinformatics.nih.gov/chartReport.html?annot=52

Все три лучшие по значимости KEGG-пути связаны с гликосфинголипидами — классом сложных мембранных липидов. Снаружи у них находится головка, по которой клетки узнают друг друга, передают сигналы и общаются с окружением. Эти молекулы важны для передачи импульсов между нейронами, памяти и работы нервной системы в целом.

Наиболее значимый путь — ganglio series (ганглиозидная серия). Ганглиозиды — это гликосфинголипиды, содержащие сиаловую кислоту. Основные функции: межклеточные взаимодействия (участвуют в адгезии клеток, формируют клеточные контакты), рецепторная функции (рецепторы для токсинов, гормонов, вирусов и бактерий).

Два других пути (лакто/неолакто и метаболизм сфинголипидов) показывают, что наш список генов охватывает не только ганглиозиды, но и другие ветви биосинтеза гликосфинголипидов.

Дополнительно мы проанализировали GO-термины с помощью кластеризации. Всего было получено два аннотированных кластера. Согласно результатам первого кластера, 18 из 19 наших генов участвуют в липидном метаболическом процессе, все 19 генов проявляют активность трансфераз и связаны с такой клеточной структурой, как мембрана. Это полностью соответствует роли генов как мембранных ферментов-трансфераз, работающих с липидными субстратами.

Sublist	Category	Term		RT	Genes	Count	%	P-Value	Benjamini
Term Cluster 1		Enrichment Score: 15.53			G				
	GOTERM_BP_DIRECT	lipid metabolic process		RT	94.74%	18	94.74	1.02e-22	4.23e-21
	GOTERM_MF_DIRECT	transferase activity		RT	100.00%	19	100.00	8.14e-19	1.51e-17
	GOTERM_CC_DIRECT	membrane		RT	100.00%	19	100.00	3.00e-7	8.25e-7
Term Cluster 2		Enrichment Score: 7.18			G				
	GOTERM_BP_DIRECT	carbohydrate derivative biosynthetic process		RT	31.58%	6	31.58	2.24e-12	4.64e-11
	GOTERM_MF_DIRECT	hexosyltransferase activity		RT	31.58%	6	31.58	7.67e-11	7.10e-10
	GOTERM_MF_DIRECT	N-acetyl-beta-D-glucosaminide beta-(1,3)-galactosyltransferase activity		RT	15.79%	3	15.79	6.39e-5	2.95e-4
	GOTERM_BP_DIRECT	protein O-linked glycosylation		RT	15.79%	3	15.79	1.82e-3	1.37e-2

Рис.2. Результаты GO анализа, полученные с помощью сервиса DAVID.
Ссылка на результаты выдачи DAVID (GO):
https://davidbioinformatics.nih.gov/term2term.html?annot=27%2C35%2C43

ПУНКТ №3

Для индивидуального анализа была выбрана база данных Human Protein Atlas. С ее помощью можно:

1. Посмотреть тканевую экспрессию — в каких органах белок работает.
2. Узнать субклеточную локализацию — в какой части клетки находится белок. Это помогает понять его функцию.
3. Найти связь с заболеваниями — какие болезни возникают при мутациях.

На примере *ST3GAL5* ниже мы проиллюстрировали решение 2 задачи (субклеточная локализация). Выбранный ген кодирует лактозилцерамид-альфа-2,3-сиалилтрансферазу. Этот белок переносит сиалильную группу с CMP-NeuAc на нередуцирующую терминальную галактозу гликофосфолипидов, образующих ганглиозиды, в основном участвует в биосинтезе ганглиозида GM3, но может использовать в качестве субстрата и другие гликолипиды (ссылка на страницу UniProt). *(если проще - ST3GAL5 кодирует фермент, который делает первый ганглиозид (GM3) из более простой молекулы-предшественника. Без него синтез остальных ганглиозидов невозможен)*

Для нормального выполнения своей функции фермент должен работать в аппарате Гольджи (+ связанными с ним везикулами), где происходит сборка гликосфинголипидов. Мы вбили в поиск HPA ID гена и во вкладке Subcellular указано, что белок локализован в везикулах. Таким образом, местоположение белка соответствует его функции.

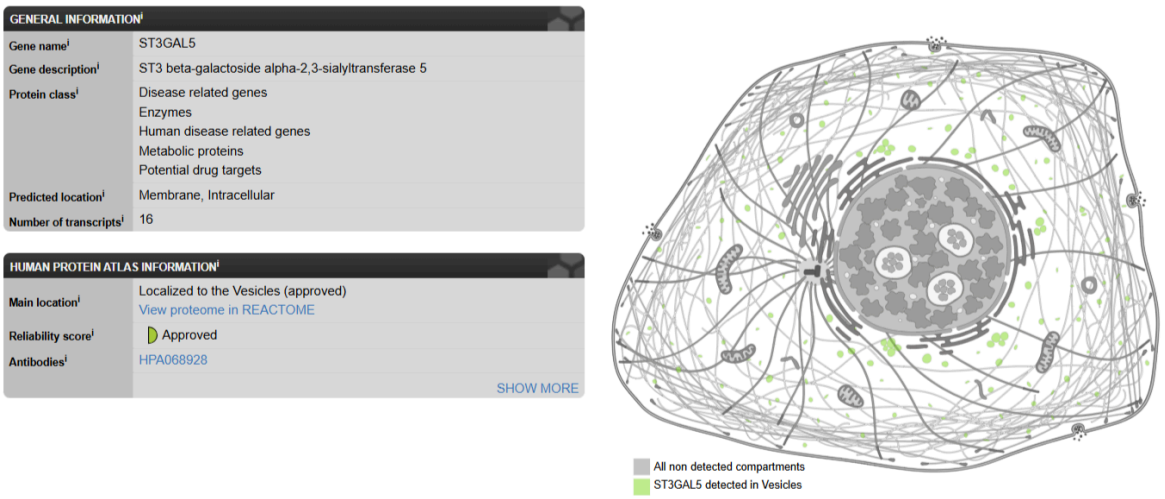

Рис.3. Клеточная локализация белка, кодируемого геном *ST3GAL5* по данным Human Protein Atlas. Ссылка на страницу: https://www.proteinatlas.org/ENSG00000115525-ST3GAL5/subcellular

Дополнительно мы решили посмотреть информацию о болезнях, связанных с выбранным белком. Использовали базу данных UniProt, с помощью которой можно получить разную информацию, например:

1. Узнать молекулярную функцию белка
2. Найти доменную структуру
3. Посмотреть с какими заболеваниями связан белок

Мы ввели идентификатор гена в UniProt и в разделе Disease & Variants нашли, что мутации в гене *ST3GAL5* вызывают редкое аутосомно-рецессивное заболевание, характеризующееся тяжелыми, рецидивирующими и резистентными судорогами в младенческом возрасте, задержкой физического развития, психомоторной заторможенностью, отставанием в развитии и корковой слепотой. У некоторых пациентов наблюдается глухота. У больных на туловище, лице и конечностях имеются участки гипо- или гиперпигментации (ссылка на страницу в UniProt).

Как уже было написано ранее - выбранный нами ген кодирует фермент, без работы которого синтез всех сложных ганглиозидов становится невозможен. В п.2 было сказано, что самым значимым обогащенным путем является именно ганглиозидная серия, то есть наш список генов работает на этот путь. Однако если ген, с которого начинается вся эта цепочка, ломается, то ганглиозиды не могут нормально синтезироваться, что приводит к различным неврологическим проблемам таким как: инфантильная эпилепсия, глубокая задержка развития, корковая слепота, глухота и тд.

Таким образом мы проанализировали список из 19 генов человека. Групповой анализ DAVID показал, что все эти гены статистически значимо обогащены по пути биосинтеза гликосфинголипидов (а наиболее значимый путь - ганглиозидна серия с p-value = 9.05e-14). Это означает, что предоставленный нам список является не просто случайным набором генов, а действительно связанной группой, работающей в едином процессе.

Индивидуальный анализ гена *ST3GAL5* подтвердил результаты группового анализа. HPA показал, что белок локализован в везикулах, которые связаны с АГ, где происходит синтез гликосфинголипидов. Локализация полностью соответствует функции фермента. Данные, полученные из базы UniProt тоже не противоречат групповому анализу, а наоборот подтверждают клиническую значимость этого пути.