Базы данных KEGG и GO

1. Входные данные

Входные данные представляют собой список из 176 ID генов.
Мы хотим проанализировать связь между предоставленными генами.

2. Групповой анализ

Цели

Сначала проведём групповой анализ полного списка ID на сайте geneontology.
Цель Gene Ontology обогащения:

Узнать какие функции/процессы встречаются в списке чаще, чем ожидалось случайно.
Сколько таких процессов
Какая общая тематика этих процессов

Поиск

Использовались параметры:

a) Точный тест Фишера
b) Поправка на множественное тестирование — False Discovery Rate (ожидаемая доля ложноположительных результатов среди всех найденных значимых результатов)
c) GO (geneontology.org) использует метод Benjamini–Hochberg для контроля FDR.
d) Поиск по категориям биологических процессов

Анализ

Выявлено выраженное и статистически значимое обогащение генов в процессах метаболизма липидов и жирных кислот.
Наиболее значимые термины включают:

- Метаболитические процессы жирных кислот
- Липидные метаболические процессы
- Метаболитические процессы карбоновых кислот

Нашлось 310 терминов GO с значениями FDR (False Discovery Rate) < 0.05. Все термины характеризуются низкими значениями FDR (False Discovery Rate до 10^{-249}) и высоким уровнем обогащения (Fold Enrichment до ~60).
Полный список из 310 значимых GO-терминов представлен в таблице выдачи обогащения.

Например, один из генов самой достоверной категории GO process – метаболические пути жирных кислот – это ген SLC27A4 из семейства переносчиков растворённых веществ 27: он участвует в транспорте жирных кислот через клеточную мембрану и их активации за счёт образования ацил-КоА..
Другой пример: ген из категории процессов биосинтеза карбоновых кислот кодирует POPTR_004G043700 - кетол-ацид редуктоизомераза- фермент из биосинтеза разветвлённых карбоновых кислот.

Вывод

В результате анализа списка генов можем сделать вывод, что наши гены перепредставлены в категориях, относящихся к метаболитическим процессам жирных кислот и липидов.

Далее сделаем анализ связи генов в сервисе STRING.

Цели

Получить таблицу обогащения – то, за что глобально отвечают гены из анализируемого списка

Проанализировать, сколько белков связаны между собой (сколько всего предсказанных связей)

Найти белки которые никак не связаны с остальными

Поиск

Оценка сети (PPI Enrichment): STRING делает сравнение с распределением в протеоме (p-value)

Статистический тест (Enrichment): точный тест Фишера (Fisher's Exact Test)

Методом оценки связи между генами STRING использует вероятностный комбинированный счет (combined score) [1] :

1) Из каждого канала убирается "априорная вероятность" — шанс того, что любые два случайных белка будут взаимодействовать (в STRING этот prior = 0.041)
2) Очищенные вероятности складываются
3) Априорная вероятность добавляется обратно только один раз

Поправка на множественное тестирование — метод Benjamini-Hochberg (FDR).

Анализ

STRING выявил 258 категорий Gene Ontology, 90 молекулярных функций и 31 клеточный компонент.

GENE COEXPRESSION использует экспериментальные данные из транскриптомики и протеомики и показывает имеют ли гены схожие паттерны экспрессии. Это выяснить не получилось, STRING отказывается считать сеть совместной экспрессии более чем для 50 белков.

GENE FUSION показывает схему того, могут ли в некоторых организмах два отдельных гена быть объединены в один. Ничего особенного этот анализ не выявил. Нашлось около 15 генов, которые могут попарно сливаться.

GENE COOCCURRENCE Показывает у каких организмов на филогенетическом древе чаще встречаются эти гены. Мы выяснили, что большинство генов встречаются у млекопитающих, хотя есть несколько, которые часто встречаются у бактерий: PCCB, HSD17B8, MMUT, CBR4. Они кодируют ферменты базовых метаболитических реакций.

Таким образом, эти анализы подтверждают, что гены из списка эволюционно консервативны и функционально связаны преимущественно у позвоночных.

Самое интересное и красивое это **NETWORK** . Он строит граф того, как связаны гены, используя данные из различных источников, включая экспериментальные данные, тексты научной литературы, вычислительные предсказания и базы метаболитических путей:

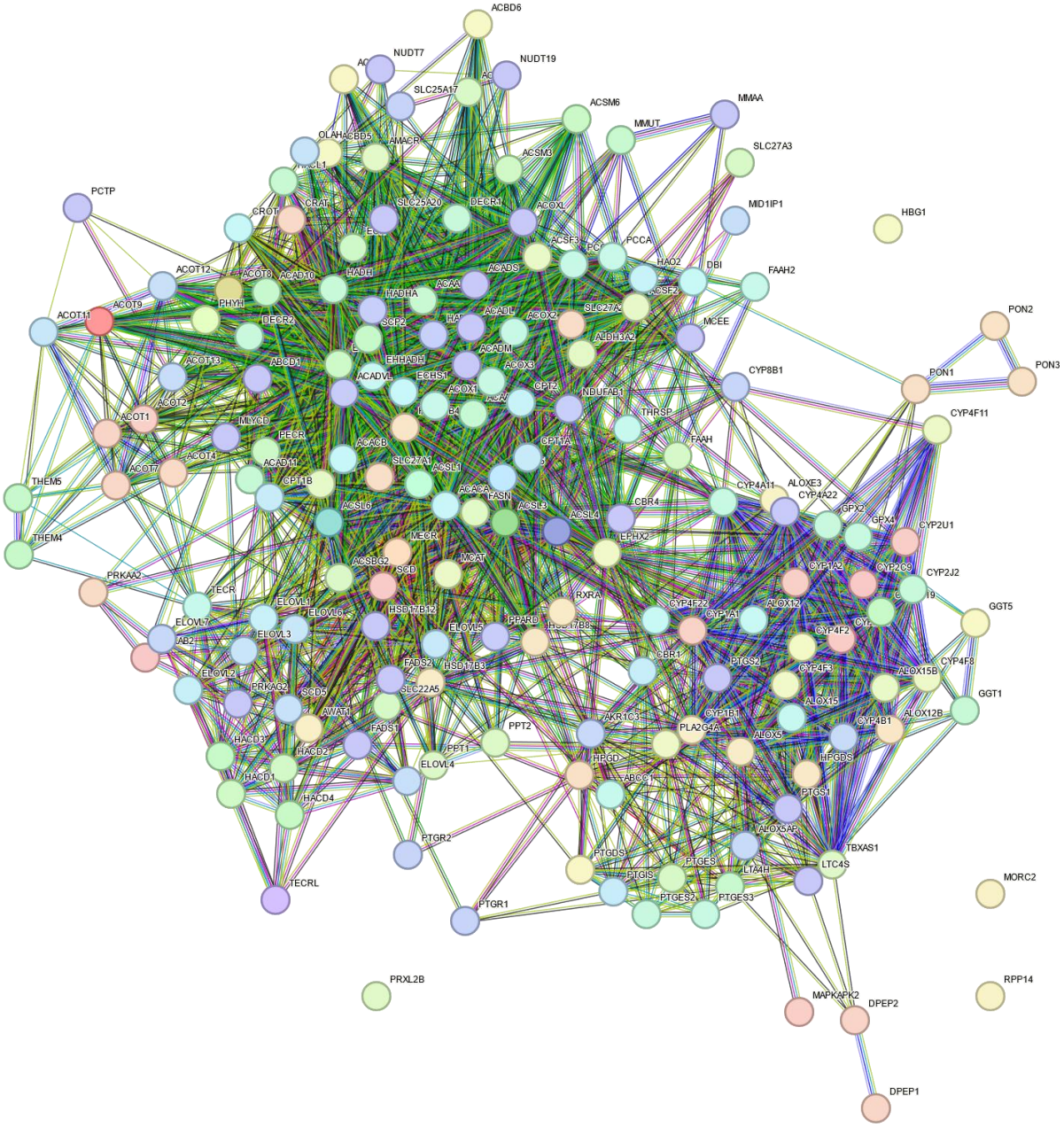


Рис. 1 Network - граф, полученный сервисом STRING по указанному списку генов

Всего в графе на **Рис.1** 2240 ребра, 174 вершины, а PPI enrichment p-value < 1.0e-16, то есть наши белки взаимодействуют друг с другом значительно больше, чем случайные белки из генома. Это говорит о том, что они работают вместе как единая система

Если попросить STRING разделить граф (**Рис.1**) на два кластера, то левый он охарактеризует как “гены метаболизма жирных кислот” (121 штука), а правый как “метаболизм арахидоновых кислот” (49 генов).

По цвету рёбер графа на **Рис.1** можем сказать, что в левом кластере много генов-соседей и которые вместе упоминаются в статьях (зелёные связи), а в правом высока совместная встречаемость в геномах и гомология (синие и фиолетовые связи).

Также есть множество белков-одиночек, которые не имеют какой бы то ни было связи с другими белками:

HBG1 – субъединица гемоглобина
RPP14 - Субъединица Рибонуклеазы P
PRXL2B - Простамид/простагландин F-синтаза
MORC2 - АТФаза MORC2, необходимая для эпигенетического сайленсинга с помощью комплекса HUSH

Чтобы понять, какие именно функции объединяют белки кластеров, обратимся к таблице обогащения **STRING** (Рис. 2):

Functional enrichments in your network

Note: some enrichments may be expected here (why?)

explain columns

Biological Process (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0006631	Fatty acid metabolic process	152 of 325	1.71	19.22	1.43e-222
GO:0032787	Monocarboxylic acid metabolic process	154 of 502	1.53	13.79	4.02e-203
GO:0019752	Carboxylic acid metabolic process	161 of 819	1.34	9.56	3.42e-189
GO:0044255	Cellular lipid metabolic process	156 of 918	1.27	8.33	1.31e-171
GO:0006629	Lipid metabolic process	164 of 1210	1.17	6.88	8.73e-171

(more ...)

Molecular Function (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0003824	Catalytic activity	162 of 5522	0.51	1.59	7.72e-65
GO:0016491	Oxidoreductase activity	75 of 731	1.05	4.31	2.93e-53
GO:0016627	Oxidoreductase activity, acting on the CH-CH group of donors	25 of 61	1.65	5.87	4.08e-28
GO:0005506	Iron ion binding	28 of 155	1.3	4.01	1.90e-23
GO:0016289	CoA hydrolase activity	17 of 24	1.89	5.21	9.61e-22

(more ...)

Cellular Component (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0005777	Peroxisome	35 of 143	1.43	5.68	8.67e-34
GO:0005782	Peroxisomal matrix	23 of 52	1.69	5.73	1.14e-26
GO:0005739	Mitochondrion	70 of 1681	0.66	1.83	7.38e-26
GO:0005737	Cytoplasm	169 of 12056	0.19	0.73	2.80e-22
GO:0005789	Endoplasmic reticulum membrane	54 of 1157	0.71	1.87	3.59e-21

(more ...)

Рис.2 Таблица с биологическими процессами и молекулярными функциями белков, кодирующихся представленными генами. Сортировка по FDR

Выдача биологическиих процессов

По данным STRING (**Рис.2**), белки из анализируемого списка относятся преимущественно к метаболизму жирных и карбоновых кислот. Самый значимый термин — Fatty acid metabolic process (GO:0006631, FDR = 1.43e-222): 152 из 174 белков сети аннотированы в этом процессе, притом что в геноме человека к нему отнесено 325 белков. Это означает, что почти половина всех известных участников метаболизма жирных кислот сосредоточена в нашем списке.

Cellular lipid metabolic process и Lipid metabolic process — ещё более широкие категории, куда входят не только жирные кислоты, но и фосфолипиды, стероиды, и другие липиды. То, что они тоже значимы, означает, что часть белков списка участвует в липидном метаболизме за пределами строго жирнокислотных путей.

Выдача клеточной локализации

Ферменты локализованы в пероксисомах, пероксисомальном матриксе, митохондриях, цитоплазме и эндоплазматическом ретикулуме (ЭПР) (**Рис. 2**) . Это не случайный набор компартментов — именно в них протекают все три основных пути окисления жирных кислот: β-окисление происходит в митохондриях и пероксисомах, α-окисление разветвлённых жирных кислот — в пероксисомах, ω-окисление — в ЭПР печени и почек (**Рис.3,4**)[2].

Выдача молекулярных функций

Среди молекулярных функций белков из списка STRING выявил (**Рис. 2**) оксидоредуктазную активность, CoA-гидратазную активность и связывание иона железа. Эти функции напрямую соответствуют ферментативным шагам окисления жирных кислот: CoA-гидратазы катализируют одну из ключевых реакций β-окисления, оксидоредуктазы обеспечивают оксиление и восстановление CH-CH связей, которые часто повторяются в остатках жирных кислот. Ионы железа необходимы для катализа этих реакций через окисление и восстановление Fe^{2+} (**Рис.3**). Таким образом, молекулярные функции белков согласованы с их клеточной локализацией и вместе описывают единый ферментативный модуль.

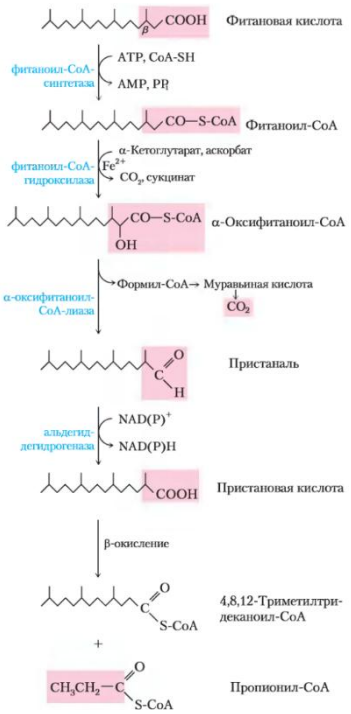


Рис. 3 Схема реакций α-окисления жирных кислот с разветвлённой цепью [2]

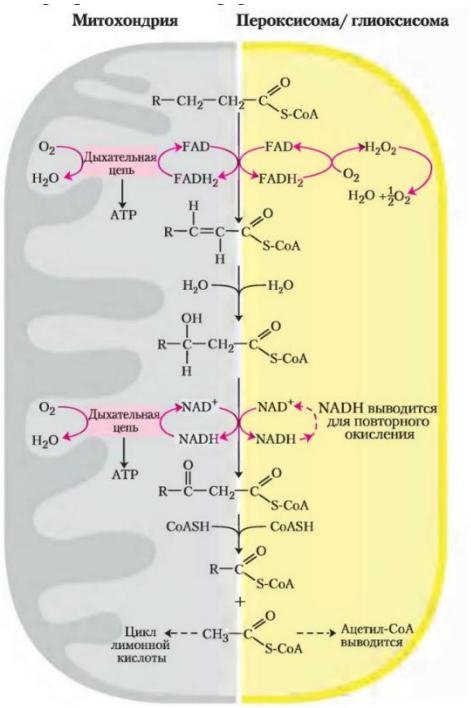


Рис. 4 Схема реакций β-оксиления жирных кислот [2]

Выводы

В результате анализа с помощью **STRING** мы выяснили что:

Большинство белков из списка (170) специфичны для позвоночных и образуют единую взаимосвязанную систему: один кластер объединяет ферменты общего метаболизма жирных кислот, второй — белки метаболизма арахидоновой кислоты, связанные с воспалительными процессами. .

Четыре белка (**HBG1**, **RPP14**, **PRXL2B**, **MORC2**) не имеют предсказанных взаимодействий с остальными, что указывает на их функциональную обособленность от липидного метаболизма.

3. Индивидуальный анализ гена из списка

Цель

1. Определение тканевой специфичности экспрессии гена — сервис позволяет установить, в каких тканях и клеточных типах исследуемый ген экспрессируется на максимальном уровне, что критически важно для интерпретации данных дифференциальной экспрессии.
2. Верификация субклеточной локализации белка — база данных предоставляет экспериментально подтвержденную информацию о том, в каких компартментах клетки (цитозоль, ядро, пероксисомы и др.) локализован целевой белок и мы можем сопоставить это с информацией полученной при анализе **STRING** (**Рис.2**)
3. С какими реакциями в каких компартментах ассоциирован ген.

Анализ

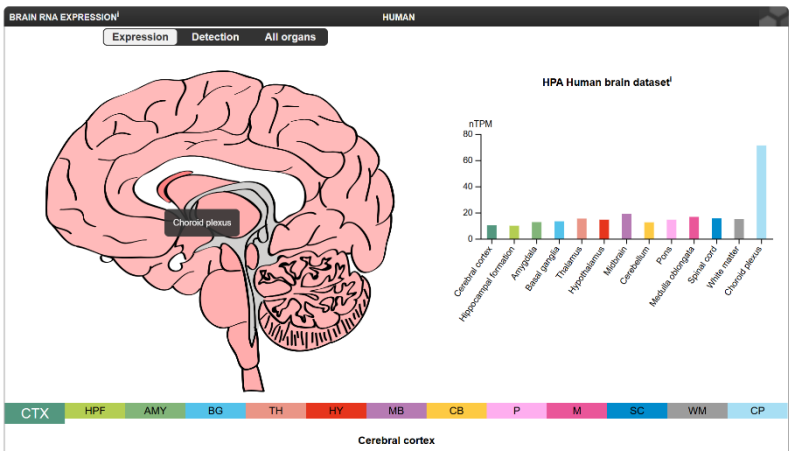
Попробуем узнать это на сайте The human protein atlas.

Поищем там информацию о белке EPHX2 (Bifunctional epoxide hydrolase 2 из GO пути lipid phosphatase activity), который является одним из мостов между левым и правым кластером в глобальном графе всех связей, полученном при анализе **STRING** (**Рис.1**). То есть продукт этого гена имеет связи и взаимодействия и с белками из кластера метаболизма жирных кислот, и с белками кластера метаболизма арахидоновых кислот. На примере белка EPHX2 попробуем узнать, в какой ткани может быть повышена экспрессия генов из списка.

N-концевой и C-концевой домены белка выполняют разные функции. C-концевой домен преобразует противовоспалительные и сосудорасширяющие липидные эпоксиды в менее активные диолы. N-концевой домен обладает липидфосфатазной активностью специфично для определённых кислот. Мутации в этом гене были связаны с семейной гиперхолестеринемией и сердечно-сосудистыми заболеваниями[3].

Поискав этот ген на Human proteins atlas, обнаруживаем:

- В графе **Protein interactions** указано **No protein interactions,** несмотря на огромное количество связей на графе на **Рис.1**
- Категория экспрессии: Согласно набору данных HPA и GTEx, ген классифицируется как "tissue enhanced" (ткане-усиленный) с максимальной экспрессией в печени . Что логично, ведь там происходит перерабатывание жирных кислот и обезвреживание эндогенных эпоксидов.
- Детализация по тканям: Помимо печени, повышенный уровень EPHX2 регистрируется в почках, предстательной железе, а также в железистых клетках других органов .
- Субклеточная локализация: Белок преимущественно локализован в цитозоле.



Интересно, что в мозге экспрессия гена повышена в Choroid plexus (сосудистое сплетение) - **Рис.5**. Это ворсинчатое образование в желудочках головного мозга, которое вырабатывает спинномозговую жидкость (ликвор). Оно состоит из кровеносных сосудов и эпителия, регулирует состав ликвора, обеспечивает барьер между кровью и мозгом, а также выводит нейротоксины - функции, похожие на функции печени.

Рис. 5 Отделы мозга с повышенной экспрессии гена EPHX2

METABOLIC SUMMARY

Gene EPHX2 is associated with 41 reactions in 4 different subsystems, and present in the compartments: Cytosol, Endoplasmic reticulum, Nucleus, Peroxisome. Provided by metabolicatlas.org

Pathway / Subsystem	Subsystem-associated compartments	# proteins	# metabolites	# reactions for this protein
Xenobiotics metabolism	**Cytosol**, Extracellular, **Endoplasmic reticulum**, **Peroxisome**	110	178	10
Arachidonic acid metabolism	**Cytosol**, **Peroxisome**, **Endoplasmic reticulum**, Extracellular, **Nucleus**, Mitochondria, Golgi apparatus	107	88	21
Omega-3 fatty acid metabolism	**Cytosol**, **Peroxisome**, **Nucleus**, **Endoplasmic reticulum**	52	73	6
Linoleate metabolism	**Cytosol**, **Endoplasmic reticulum**	63	42	4

Рис. 6 Сводка метаболитических процессов, связанных с продуктом гена EPHX2

В таблице на **Рис. 6** указано, что ген связан с 41 реакцией. Это пути ксенобионтного метаболизма и метаболизма жирных кислот, протекающие в цитозоле, внеклеточном пространстве, пероксисомах, митохондриях и ЭПР.

Вывод

Продукт гена EPHX2 вовлечён в 41 реакцию из путей ксенобиотического метаболизма и окисления жирных кислот; реакции протекают в цитозоле, пероксисомах, митохондриях и ЭПР — тех же компартментах, которые были выявлены как обогащённые при групповом анализе в STRING. Это подтверждает, что EPHX2 является типичным представителем функциональной группы генов данного списка и участвует в процессах детоксикации.

4. Выводы

Проведённый анализ списка из 176 генов с использованием GO-обогащения, STRING и индивидуального разбора показал, что данный набор генов представляет собой функционально связанную систему, вовлечённую в метаболизм липидов и жирных кислот.

В STRING показана плотная сеть взаимодействий (PPI enrichment p-value < 1e-16), разделяющаяся на кластеры метаболизма жирных и арахидоновой кислот; функционально белки представлены в основном ферментами окислительно-восстановительных реакций с локализацией в пероксисомах, митохондриях, эндоплазматическом ретикулуме и цитозоле, а индивидуальный анализ (на примере EPHX2) подтверждает тканевую специфичность и согласованность субклеточной локализации, что в совокупности указывает на то что это согласованная функциональная группа белков из липидного обмена.

5. Источники

1) FAQ STRING - How are the scores computed?

2) Лекция 9. Катаболизм жирных кислот

3) Cristiano Fava, Martina Montagnana, Elisa Danese, et al. **Homozygosity for the EPHX2 K55R polymorphism increases the long-term risk of ischemic stroke in men: a study in Swedes.** *Pharmacogenet Genomics (2010)* DOI: 10.1097/FPC.0b013e3283349ec9 PMID: 20065888