Практикум 6. Базы данных KEGG и GO

1. Входные данные

Мой файл: list43.txt. Он содержит 58 идентификаторов генов человека. Некоторые гены выглядят как семейство (например, `GLB1`, `GLB1L`, `GLB1L2`), что может указывать на их общую функцию.

2. Групповой анализ с помощью DAVID

Я выбрал программу DAVID (Database for Annotation, Visualization and Integrated Discovery) поскольку она умеет искать обогащение не только по GO-терминам, но и по KEGG-путям, доменам, ассоциациям с болезнями и по другим параметрам.
Результаты в формате csv:

- DAVIDChartReport_group43_2026-05-02.csv
- DAVIDFunctAnnotClusterReport_group43_2026-05-02.csv
- group43_DAVIDFunctionalAnnotationTable_2026-05-02.csv

2a. Примеры задач, решаемые с помощью DAVID

1. **Выявление перепредставленных функциональных категорий в списке генов.** Functional Annotation Chart показывает, по каким биологическим процессам, молекулярным функциям или клеточным компонентам наш список генов значимо перепредставлен относительно остального генома.
2. **Идентификация метаболических и сигнальных путей, в которые вовлечены продукты генов списка.** Категории `KEGG_PATHWAY`, `REACTOME` и другие позволяют установить, в каких канонических путях участвуют гены, и визуализировать их на картах KEGG.
3. **Кластеризация похожих терминов.** DAVID позволяет кластеризовать избыточные аннотационные термины в функциональные группы (Functional Annotation Clustering), облегчая интерпретацию результатов обогащения

2b. Общая «тема» списка генов по результатам обогащения

Входные данные и параметры анализа

В анализ был подан список из 58 генов человека (идентификаторы Gene Symbol, файл list43.txt). Использовался инструмент Functional Annotation Chart сервиса DAVID со следующими параметрами:

- **Тест на обогащение:** модифицированный точный тест Фишера.
- **Поправка на множественное тестирование:** Benjamini–Hochberg (FDR), а также более консервативная поправка Бонферрони; в качестве основного критерия

значимости использовался $\mathrm{FDR} < 0.05$.

- **Фоновый список:** все гены *Homo sapiens*, представленные в базе DAVID.

Общая характеристика результатов

В категории `GOTERM_BP_DIRECT` (BP – biological process) нашлось 63 значимых термина ($\mathrm{FDR} < 0.05$). Полная таблица результатов доступна в дополнительных материалах (файл DAVIDChartReport_group43_2026-05-02.csv), отсортированная по значению p-value с поправками. Кроме того, анализ по `KEGG_PATHWAY` показал значительное обогащение путей `lipid metabolism` и `Sphingolipid metabolism`, а **Functional Annotation Clustering** (файл DAVIDFunctAnnotClusterReport_group43_2026-05-02.csv) сгруппировал сходные термины в три основных кластера (Enrichment Score 11.21; 7.25; 2.27).

Лучшие находки

На рис. 1. показаны 15 путей с наименьшими FDR первые значимые обогащенные биологические процессы ($\mathrm{FDR} < 1 \times 10^{-20}$):

1. **lipid metabolic process** ($\mathrm{FDR} = 1.99 \times 10^{-41}$, $\mathrm{Fold\ Enrichment} = 18.1$) – 40 из 51 распознанного гена списка ассоциированы с этим термином. Это наиболее широкая категория, подтверждающая, что список почти целиком состоит из генов липидного метаболизма.
2. **glycosphingolipid biosynthetic process** ($\mathrm{FDR} = 6.40 \times 10^{-35}$, $\mathrm{Fold\ Enrichment} = 260$) – 17 генов (B3GALNT1, GAL3ST1, CERK, B3GALT4, FUT1/2, UGT8, UGCG, B3GNT5, ST3-семейство, ST6-семейство, B4GALT6, A4GALT). очень сильное обогащение: в списке находятся ключевые гликозилтрансферазы, которые собирают углеводную часть гликосфинголипидов.
3. **glycosphingolipid catabolic process** ($\mathrm{FDR} = 1.04 \times 10^{-26}$, $\mathrm{Fold\ Enrichment} = 383$) – 12 генов, среди которых GALC, NEU2, NEU3, GM2A, GLB1, GBA2, GBA3, GLA, ENPP7, SMPD1, M6PR, SUMF1. Почти все перечисленные белки являются лизосомальными гидролазами или их активаторами, участвующими в ступенчатой деградации гликосфинголипидов.
4. **sphingolipid metabolic process** ($\mathrm{FDR} = 5.87 \times 10^{-23}$, $\mathrm{Fold\ Enrichment} = 97.3$) еще раз подчеркивает фокус списка на сфинголипидах и ганглиозидах.

Кластерный анализ (Рис. 2.) объединил термины в группы:

- **Кластер** 1 (Enrichment Score 11.21) – ganglioside/oligosaccharide catabolic process.
- **Кластер** 2 (Enrichment Score 7.25) – ceramide/sphingomyelin metabolic process.
- **Кластер** 3 (Enrichment Score 2.27) – response to pH/estrogen и autophagy (с участием ARSA, GBA1, ARSB).
  Основная тема списка – лизосомальный метаболизм глико- и сфинголипидов, включающий как деградацию ганглиозидов, церамидов, сфингомиелина, так и их

биосинтез (Рис. 3). В список входят не только собственно ферменты-гидролазы (GLA, GBA1, GALC, HEXA/B, NEU1-4, ARSA и др.), но и вспомогательные белки.

Иллюстрации

Sublist	Category	Term	RT	Genes	Count	P-Value	Benjamini
☐	GOTERM_BP_DIRECT	lipid metabolic process	RT	68.97%	40	7.35e-44	2.28e-41
☐	GOTERM_BP_DIRECT	glycosphingolipid biosynthetic process	RT	29.31%	17	4.72e-37	7.32e-35
☐	GOTERM_BP_DIRECT	glycosphingolipid catabolic process	RT	20.69%	12	1.15e-28	1.19e-26
☐	GOTERM_BP_DIRECT	sphingolipid metabolic process	RT	25.86%	15	8.67e-25	6.72e-23
☐	GOTERM_BP_DIRECT	carbohydrate metabolic process	RT	27.59%	16	3.85e-19	2.39e-17
☐	GOTERM_BP_DIRECT	ganglioside catabolic process	RT	13.79%	8	3.74e-18	1.93e-16
☐	GOTERM_BP_DIRECT	oligosaccharide catabolic process	RT	12.07%	7	1.89e-13	8.39e-12
☐	GOTERM_BP_DIRECT	oligosaccharide biosynthetic process	RT	13.79%	8	2.97e-13	1.15e-11
☐	GOTERM_BP_DIRECT	carbohydrate derivative biosynthetic process	RT	10.34%	6	5.44e-10	1.87e-8
☐	GOTERM_BP_DIRECT	ceramide metabolic process	RT	10.34%	6	6.96e-9	2.16e-7
☐	GOTERM_BP_DIRECT	ganglioside biosynthetic process	RT	8.62%	5	1.24e-8	3.50e-7
☐	GOTERM_BP_DIRECT	ceramide biosynthetic process	RT	10.34%	6	6.29e-8	1.62e-6
☐	GOTERM_BP_DIRECT	sphingomyelin metabolic process	RT	6.90%	4	1.58e-7	3.26e-6
☐	GOTERM_BP_DIRECT	sphingomyelin catabolic process	RT	6.90%	4	1.58e-7	3.26e-6
☐	GOTERM_BP_DIRECT	galactose catabolic process	RT	6.90%	4	1.58e-7	3.26e-6
☐	GOTERM_BP_DIRECT	lipid storage	RT	8.62%	5	1.48e-6	2.86e-5
☐	GOTERM_BP_DIRECT	glycolipid biosynthetic process	RT	6.90%	4	3.43e-6	6.25e-5
☐	GOTERM_BP_DIRECT	glucosylceramide catabolic process	RT	5.17%	3	1.93e-5	3.32e-4
☐	GOTERM_BP_DIRECT	lysosome organization	RT	8.62%	5	2.30e-5	3.76e-4
☐	GOTERM_BP_DIRECT	oligosaccharide metabolic process	RT	6.90%	4	2.70e-5	4.19e-4
☐	GOTERM_BP_DIRECT	glycoside catabolic process	RT	5.17%	3	9.59e-5	1.42e-3
☐	GOTERM_BP_DIRECT	protein O-linked glycosylation via N-acetyl-galactosamine	RT	6.90%	4	1.46e-4	2.06e-3

Рис. 1. Первые 15 строк таблицы Functional Annotation Chart (`GOTERM_BP_DIRECT`). Экстремально низкие значения FDR для процессов метаболизма гликосфинголипидов.

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
Term Cluster 1		**Enrichment Score: 11.21**			G			
☐	GOTERM_BP_DIRECT	ganglioside catabolic process	RT	13.79%	8	13.79	3.74e-18	1.93e-16
☐	GOTERM_BP_DIRECT	oligosaccharide catabolic process	RT	12.07%	7	12.07	1.89e-13	8.39e-12
☐	GOTERM_BP_DIRECT	lipid catabolic process	RT	8.62%	5	8.62	3.42e-4	4.24e-3
Term Cluster 2		**Enrichment Score: 7.25**			G			
☐	GOTERM_BP_DIRECT	ceramide metabolic process	RT	10.34%	6	10.34	6.96e-9	2.16e-7
☐	GOTERM_BP_DIRECT	sphingomyelin metabolic process	RT	6.90%	4	6.90	1.58e-7	3.26e-6
☐	GOTERM_BP_DIRECT	sphingomyelin catabolic process	RT	6.90%	4	6.90	1.58e-7	3.26e-6
Term Cluster 3		**Enrichment Score: 2.27**			G			
☐	GOTERM_BP_DIRECT	response to pH	RT	5.17%	3	5.17	1.78e-4	2.30e-3
☐	GOTERM_BP_DIRECT	response to estrogen	RT	5.17%	3	5.17	7.57e-3	6.42e-2
☐	GOTERM_BP_DIRECT	autophagy	RT	5.17%	3	5.17	1.12e-1	5.32e-1

3 annotation clusters

Рис. 2. Результат Functional Annotation Clustering, показывающий три главных кластера с Enrichment Score и входящими в них терминами.

Рис. 3. Гистограмма распределения генов списка по ключевым KEGG-путям (Lysosome, Sphingolipid metabolism, Glycosphingolipid biosynthesis, Other glycan degradation и т.д.), построенная на основе файла group43_DAVIDFunctionalAnnotationTable_2026-05-02.csv.

Обсуждение результатов

Предоставленный список генов обладает цельной биологической темой: подавляющее большинство его членов кодируют белки, работающие в лизосомах и непосредственно участвующие в обмене сфинголипидов и гликанов. Такая слаженность редко бывает случайной, что объясняет большие значения обогащения (Fold Enrichment до 383).

С биологической точки зрения это семейство генов представляет собой связную группу, где:

- Гидролазы (GLA, GBA1, GALC, HEXA/B, NEU1-4, ARSA, ARSB) осуществляют поэтапное отщепление моносахаридных, сульфатных и сиаловых остатков от гликосфинголипидов в кислой среде лизосом.
- Активаторы и переносчики (GM2A, PSAP, M6PR) обеспечивают правильную доставку и презентацию субстратов для гидролаз.
- Ферменты модификации (SUMF1) активируют сульфатазы, перевлдя их в функциональную форму.
- Гликозилтрансферазы (FUT1/2, ST3-семейство, B3GALT4, B4GALNT1 и др.), напротив, участвуют в биосинтезе тех же самых глико- и сфинголипидов, что указывает на тонкий баланс между синтезом и деградацией, контролируемый данной группой генов.

Нарушения в этих генах в клинически проявляются как лизосомные болезни накопления что дополнительно подтверждает функциональную общность списка.

Таким образом, DAVID подтвердил лизосомно-сфинголипидную тему списка и разложил её на несколько более узких тем (распад ганглиозидов, обмен церамида, работа с олигосахаридами).

3. Индивидуальный анализ с помощью The Human Protein Atlas

a. Задачи, решаемые с помощью Human Protein Atlas

1. **Исследование профиля экспрессии мРНК и белка в нормальных тканях, органах и клеточных типах человека.** Для любого гена можно посмотреть, в каких тканях и насколько сильно экспрессируется РНК и белок, и узнать, насколько он тканеспецифичен.
2. **Анализ субклеточной локализации белков.** На основе данных конфокальной микроскопии с флуоресцентно-мечеными антителами сервис предоставляет сведения о преимущественной внутриклеточной локализации белка, дополнненные предсказаниями сигнальных пептидов и трансмембранных доменов
3. **Оценка клинической значимости гена: прогноз выживаемости при различных видах рака (раздел Pathology) и ассоциация с наследственными заболеваниями.** Human Protein Atlas интегрирует данные о прогностической ценности уровня мРНК/белка для многих форм злокачественных новообразований, а также содержит информацию о связи гена с моногенными заболеваниями (например, болезнь Фабри для GLA).

b. Пример решения задачи на основе гена GLA (исследование профиля экспрессии и клинической значимости)

Для иллюстрации выбрана задача исследования профиля экспрессии мРНК и белка в нормальных тканях на примере гена GLA

Согласно странице сводки, ген GLA характеризуется следующими ключевыми особенностями:

- **Экспрессия РНК в тканях:**
  Тканевая специфичность низкая. Атлас относит его к кластеру «Parathyroid gland – Mixed function», то есть чуть выше экспрессия в паращитовидной железе, но в целом его можно считать типичным «геном домашнего хозяйства».
- **Экспрессия белка:**
  Белок GLA демонстрирует гранулярное цитоплазматическое окрашивание, наиболее выраженное в железистых (секреторных) клетках. Это ожидаемо для лизосомалного фермента.

- **Клиническая значимость:**
  GLA является прогностическим маркером при раке печени (гепатоцеллюлярная карцинома) и может быть потенциальной мишенью для лекарств.

Можно сделать вывод, что фермент GLA присутствует во всех тканях (что ожидаемо для лизосомального белка), его белок локализуется в цитоплазматических гранулах, а нарушение его функции приводит к конкретной патологии (гепатоцеллюлярная карционма).

3c. Соотнесение характеристики GLA с результатами группового анализа (пункт 2)

Групповой анализ в DAVID вывел на первый план лизосомальный обмен глико-/сфинголипидов и организацию лизосомы:

- Молекулярная функция: GLA кодирует лизосомальную гидролазу (альфа-галактозидазу), которая отщепляет концевые альфа-галактозильные остатки от гликосфинголипидов, преимущественно от церамидтригексозида.
- Субклеточная локализация: Данные Human Protein Atlas (гранулярная цитоплазма, предсказанная внутриклеточная локализация) согласуются с обогащением термина **lysosome organization** ($\mathrm{FDR} = 3.28 \times 10^{-4}$) и путем KEGG **Lysosome**. GLA – классический лизосомальный фермент, транспортрруемый в лизосомы через маннозо-6-фосфатный рецептор (ген *M6PR* также присутствует в списке!).
- Биологический процесс: В категории GO *lipid metabolic process* ($\mathrm{FDR} = 1.99 \times 10^{-41}$) GLA участвует в деградации сфинголипидов, что связывает индивиуальную характеристику гена с общей темой всего списка.
- Патология: Болезнь Фабри (мутации GLA) – пример лизосомной болезни. Он находится в списке рядом с GBA1, HEXA, ARSA с такими же патологиями, так что группа подобрана цельно.

3d. Иллюстрации

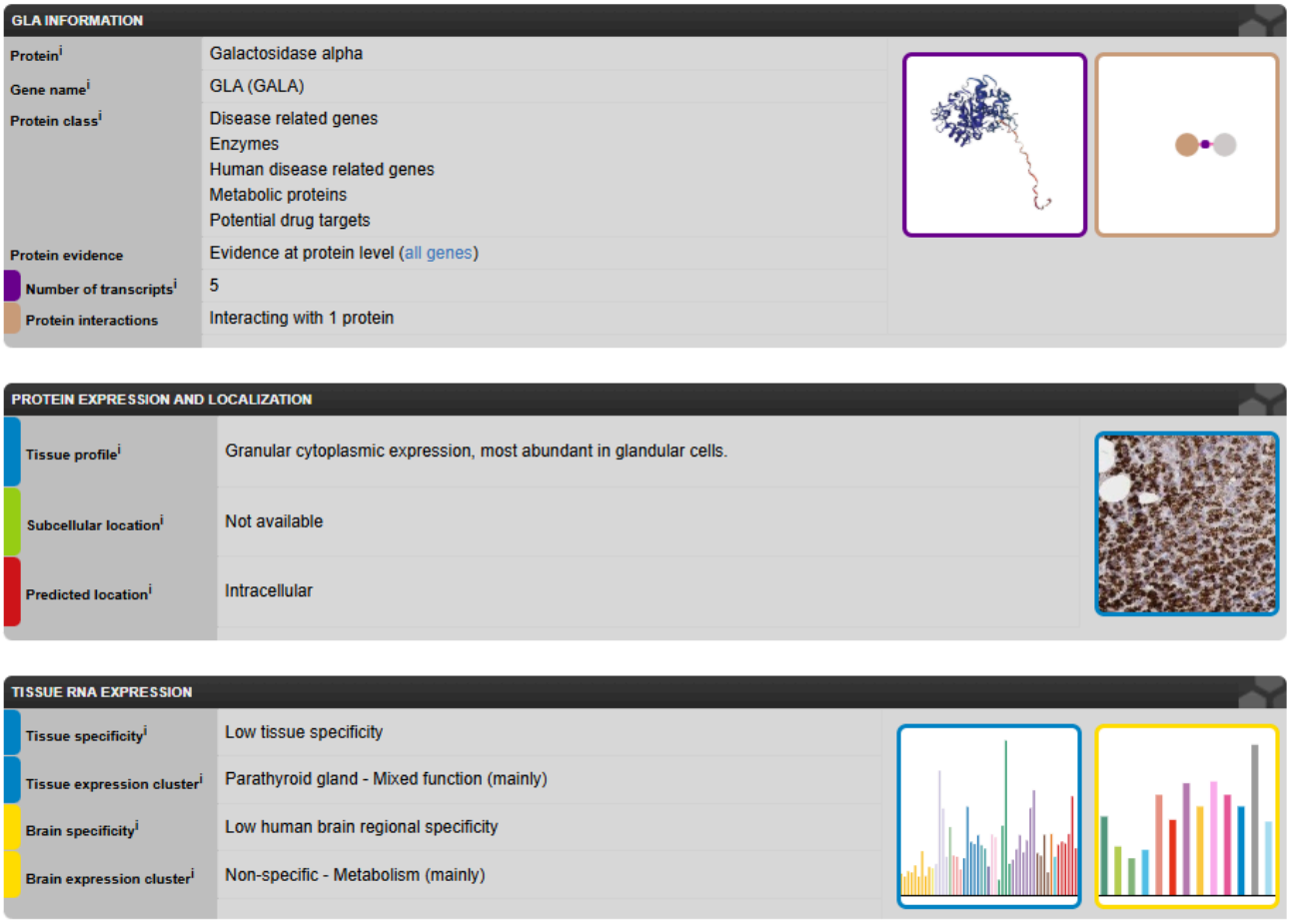

Рис. 4. Основная информация о GLA на вкладке Summary

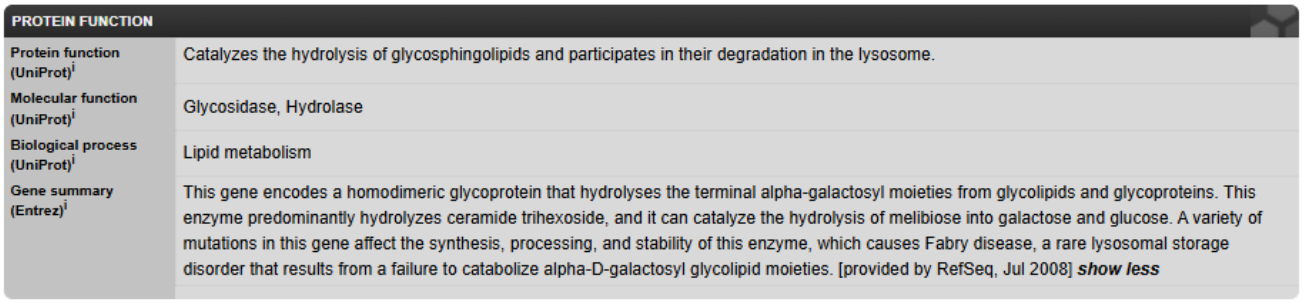

Рис. 5. Функции белка, кодируемого геном GLA, указанные на вкладке Summary

3e. Обсуждение

Данные Human Protein Atlas по GLA сходятся с групповым анализом. Ген экспрессируется везде, так как лизосомы есть в каждой клетке. Более высокая экспрессия в секреторных клетках, вероятно, связана с тем, что они активнее обновляют мембраны.

Если наложить это на результаты DAVID, видно, что GLA выполняет одну из важнейших ролей в деградации гликосфинголипидов. Стоит отметить, что в списке присутствует M6PR — рецептор, который направляет GLA и другие гидролазы в лизосомы.