Анализ набора генов человека

Рассматриваемый далее набор содержит 44 гена человека, в списке приведены HGNC-символьные обозначения генов.

На первый взгляд данный набор содержит большое количество генов ABC- и SLC-транспортеров, а также генов из семейства цитохромов P450 (CYP). Поэтому можно предположить, что список содержит преимущественно гены, связанные с метаболизмом липидов или ксенобиотиков.

Анализ набора генов с использованием инструмента Enrichr

Enrichr – один из инструментов, позволяющих проводить анализ обогащения наборов генов. Его преимуществом в сравнении с аналогами является работа с десятками различных библиотек (по данным с сайта Enrichr – 225 библиотек).

С помощью данного инструмента можно решать такие задачи, как:

- Анализ результатов РНК-секвенирования. Инструмент позволяет провести анализ обогащения терминов GO, KEGG и др. для списка дифференциально экспрессируемых генов.
- Определение клеточного состава образца по генам-маркерам. Значимые результаты в разделе “Cell Types” могут показать, для каких тканей характерная экспрессия генов из анализируемого набора.
- Определение потенциальной причины изменения экспрессии генов. В разделе “Transcription” результатов работы Enrichr можно узнать, какие транскрипционные факторы влияют на экспрессию генов из набора.

Инструмент использует точный тест Фишера (Fisher exact test) для расчета значимости обогащения. В качестве поправки на множественное тестирование применяется поправка Беньямини-Хохберга (Benjamini-Hochberg), менее строгая, чем поправка Бонферрони, что снижает вероятность ложноотрицательного результата.

Также Enrichr рассчитывает отношение шансов (Odds ratio), являющийся показателем того, насколько чаще гены из набора оказываются в конкретной категории по сравнению со всеми генами из генома. Также инструмент рассчитывает комбинированный показатель (Combined score) как абсолютное значение произведения логарифма p-value на z-score. Z-score в свою очередь рассчитывается в результате генерации множества случайных наборов генов того же размера, что и запрос, и подсчета для них числа генов, попадающих в определенную категорию, после чего для запроса и полученного распределения числа генов из набора в категории подсчитывается z-значение. Таким образом комбинированный показатель будет завышен для категорий, содержащих малое количество генов, и занижен для обширных категорий.

Для рассматриваемого набора генов был проведен анализ обогащения при помощи Enrichr.

Для GO: Biological Process в выдаче было получено 262 термина, из которых значимыми (adjusted p-value < 0.05) были 109.

Для GO: Cellular Component – 32 термина, 5 значимых.

Для GO: Molecular function – 82 термина, 56 значимых.

Для путей KEGG – 44 находки, 11 значимых.

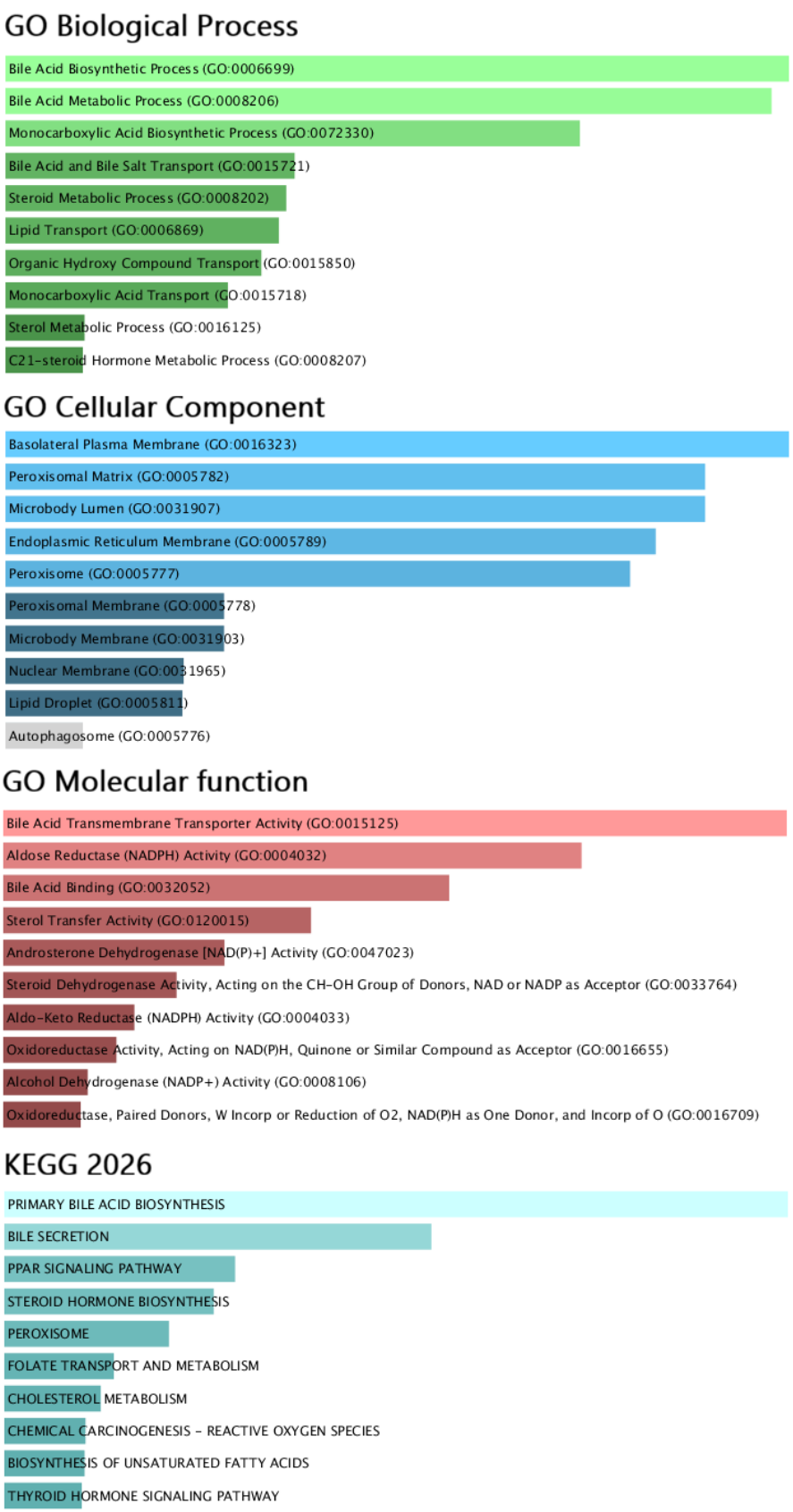

Рисунок 1. Обогащение категорий GO и путей KEGG в рассматриваемом наборе генов. Величина столбцов отражает статистическую значимость обогащения, оцениваемую по p-value.

Как можно видеть (Рис. 1), среди категорий биологических процессов наиболее значимое обогащение наблюдается для процессов метаболизма и транспорта желчных кислот и их солей. Среди категорий, описывающих локализацию белковых продуктов генов, обогащены те, что связаны с эндоплазматическим ретикулумом и пероксисомами, а также с базолатеральной мембраной клетки (это, вероятно, связано с наличием в наборе генов большого числа переносчиков).

Среди категорий GO: Molecular function и путей KEGG также оказались обогащены связанные с метаболизмом желчных кислот, стероидных соединений и соответствующими сигнальными путями (сигнальный путь PPAR).

Таким образом, результаты обогащения категорий GO, а также путей KEGG подтверждают изначальное предположение о связи рассматриваемого набора генов с метаболизмом липидов и уточняет его. По-видимому, данные гены принимают участие в метаболизме желчных кислот.

При этом такой набор генов ожидаемо является маркером клеток печени с высокой статистической значимостью (adjusted p-value = 0.0001) при анализе по базе данных Human Gene Atlas (Рис. 2). Всего в выдаче было получено 27 находок, из которых только 2 имели adjusted p-value < 0.05.

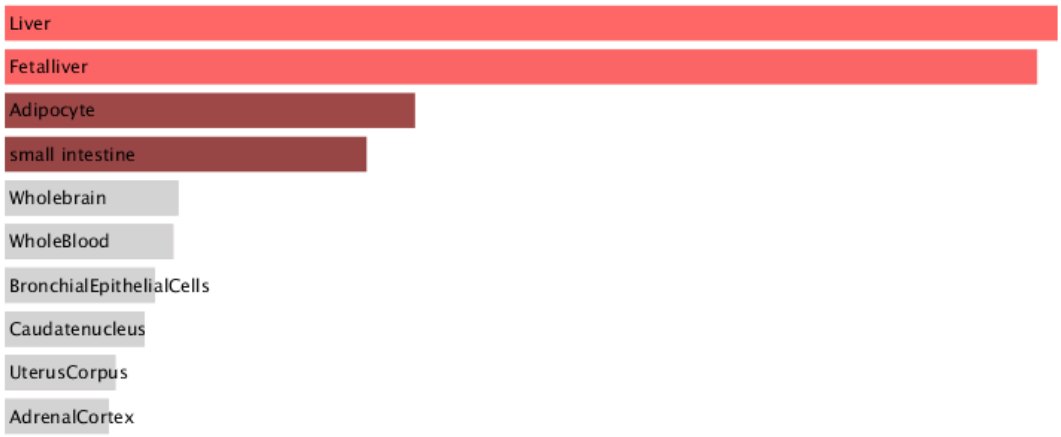

Рисунок 2. Результаты анализа обогащения типов тканей и клеток в рассматриваемом наборе генов. Величина столбцов отражает статистическую значимость обогащения, оцениваемую по p-value. Серый цвет отражает отсутствие статистической значимости (p-value < 0.05).

Также гены из набора рассматриваются в связи с такими заболеваниями, как холестатический синдром (Cholestasis), желчнокаменная болезнь (Cholelithiasis и Cholecystolithiasis) и стеаторрея (Steatorrhea) согласно базе данных DisGeNET (Рис. 3). В данном случае в выдаче было 1113 заболеваний, при этом только для 114 adjusted p-value < 0.05.

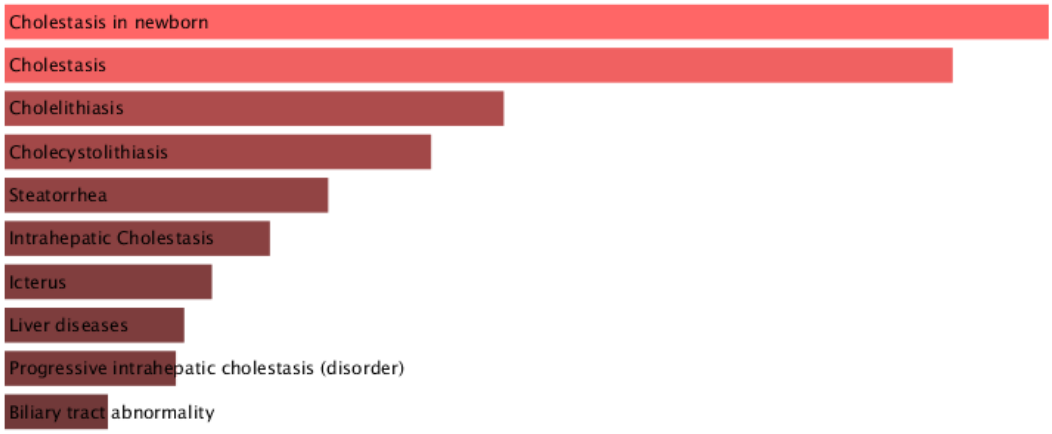

Рисунок 3. Результаты анализа обогащения заболеваний согласно данным из DisGeNET. Величина столбцов отражает статистическую значимость обогащения, оцениваемую по p-value.

Согласно данным из GWAS Catalog 2025 вариации в рассматриваемых генах могут быть ассоциированы с изменениями уровня билирубина в крови (Bilirubin Levels), миопатией, вызванной статинами (Statin-Induced Myopathy), особенностями фармакокинетики метотрексата (Methotrexate Pharmacokinetics) и др. (Рис. 4). В выдаче было 196 находок, при этом только 3 имели adjusted p-value < 0.05.

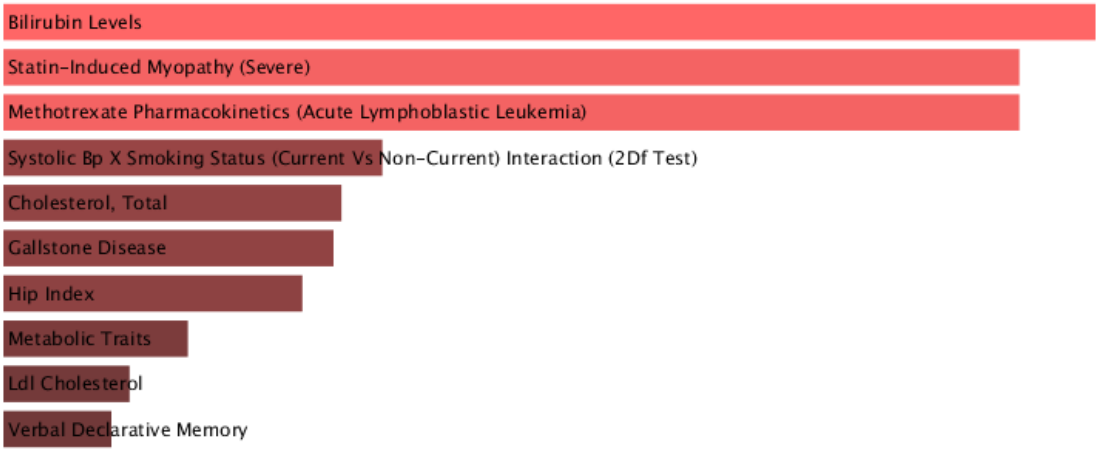

Рисунок 4. Результаты анализа обогащения ассоциаций согласно данным из GWAS Catalog 2025. Величина столбцов отражает статистическую значимость обогащения, оцениваемую по p-value.

Анализ информации о гене NR1H4, представленной в базе данных Human Protein Atlas

The Human Protein Atlas – база данных, объединяющая большое количество протеомных и транскриптомных данных. С ее помощью можно решать различные задачи, в том числе:

- Определение ткани, для которой специфична экспрессия рассматриваемого гена
- Определение субклеточной локализации белкового продукта гена
- Поиск биологических маркеров различных видов рака

Из рассматриваемого набора был выбран ген NR1H4. Для него был произведен поиск известной информации в базе данных The Human Protein Atlas.

По представленным из UniProt сведениям, данный ген кодирует белок, являющийся лиганд-активируемым транскрипционным фактором (ядерным рецептором желчных кислот) и регулятором экспрессии генов, отвечающих за синтез, транспорт и конъюгацию жирных кислот. Экспрессия NR1H4 в первую очередь характерная для печени (Liver) и кишечника (Small intestine, Duodenum, Colon и Rectum) (Рис. 5).

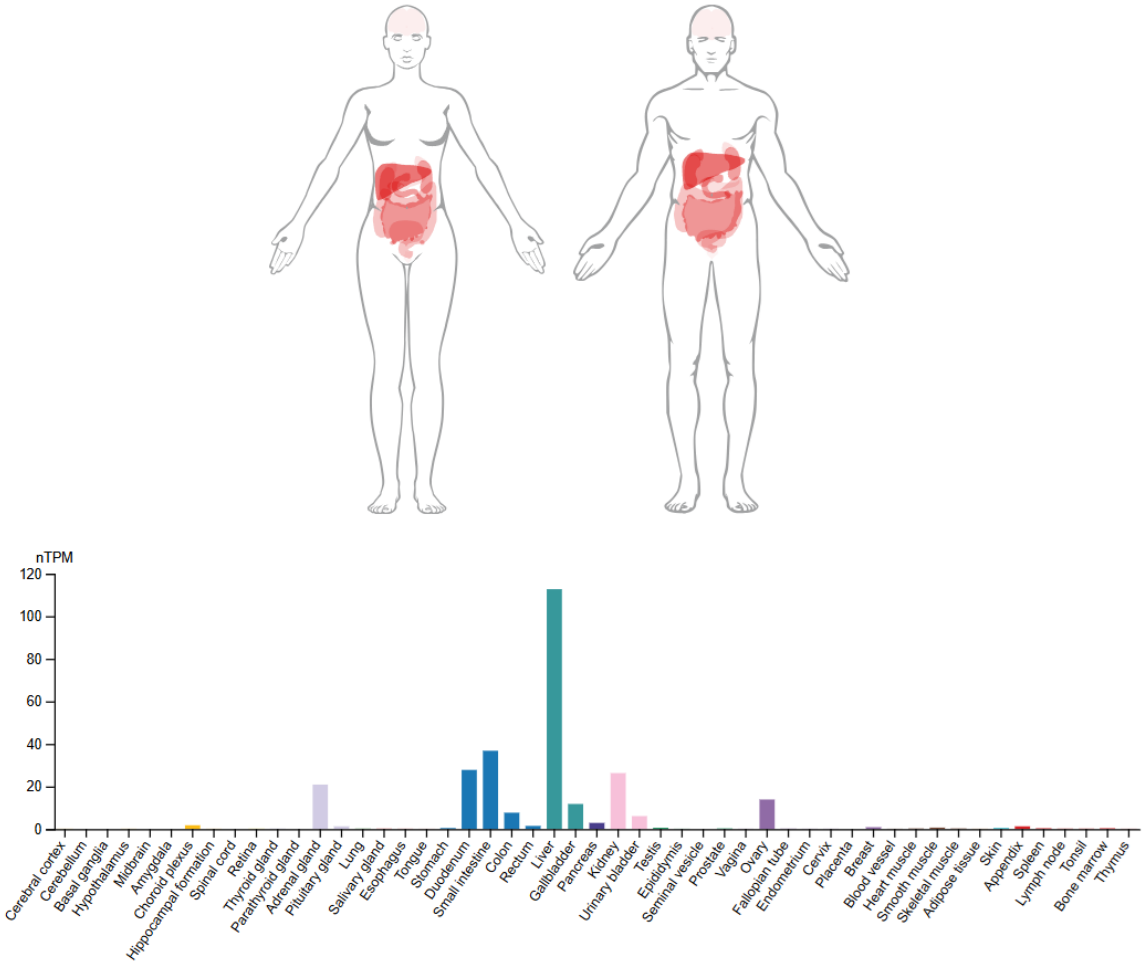

Рисунок 5. Данные по экспрессии гена NR1H4 в различных тканях и органах человека, приведенные в базе данных The Human Protein Atlas. Представлены графические изображения, показывающие районы экспрессии данного гена в женском (слева) и мужском (справа) организмах. Ниже приведены данные по уровням экспрессии в тканях.

Данный транскрипционный фактор локализуется в цитозоле, нуклеоплазме, ядерных спеклах и первичной ресничке (Рис. 6). В целом такая картина типична для ядерных рецепторов, так как они могут находиться как в связанном с ДНК, так и в растворенном в цитозоле состоянии. Первичная ресничка – важный сенсорный элемент клетки, поэтому локализация NR1H4 в ее области вполне объяснима.

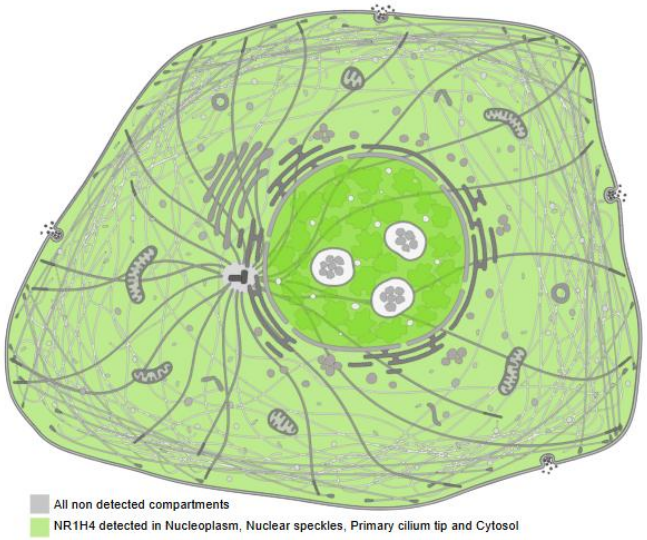

Рисунок 6. Схематичное представление локализации фактора транскрипции NR1H4 из базы данных The Human Protein Atlas. Зеленый цвет окрашивает компартменты и структуры клетки, в которых данный белок может быть обнаружен.

Для NR1H4 в базе данных Human Protein Atlas указано лишь одно взаимодействие – с коактиватором NCOA1 (Рис. 7)

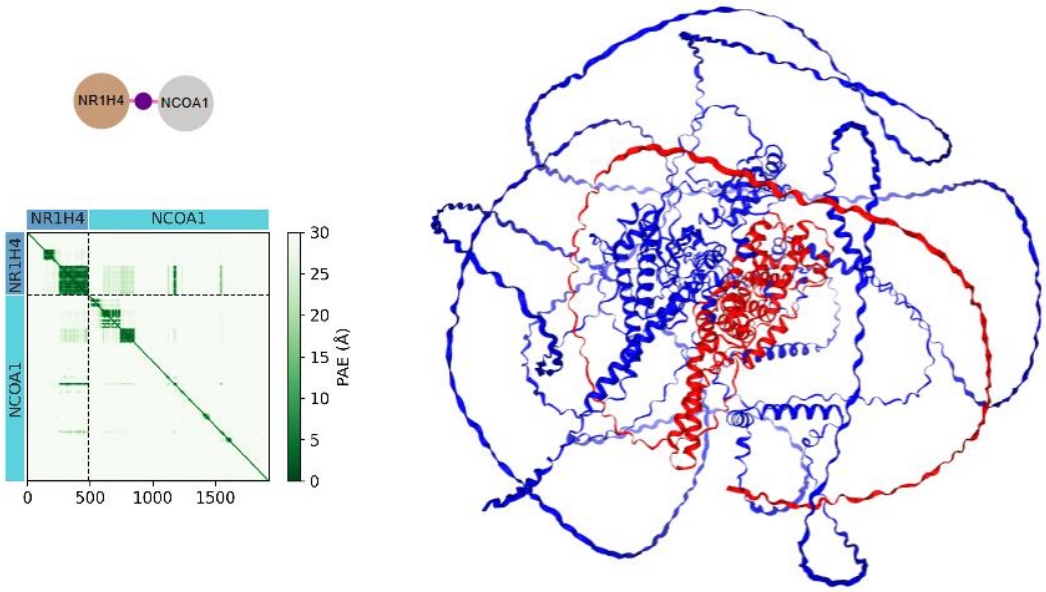

Рисунок 7. Данные по белок-белковым взаимодействиям фактора NR1H4, представленные в базе данных The Human Protein Atlas. В базе приведено лишь одно взаимодействие с участием данного белка – с коактиватором NCOA1. На рисунке приведена “карта контактов” для предсказанных структур белков и изображение их взаимодействия (красный – NR1H4, синий – NCOA1). Можно заметить многочисленные внутренние дезорганизованные регионы (IDR) в структурах белков (особенно NCOA1).

Этот факт настораживает, поскольку транскрипционные факторы обычно способны взаимодействовать как минимум с несколькими другими белками. К тому же из литературы известно, что NR1H4 связывается с ДНК в составе димера с белками RXR (гомологи NR1H4), а также взаимодействует с корепрессорами, например, с NCOR1, являющимся частью комплекса-ремоделера хроматина, осуществляющим деацетилирование гистонов.

Таким образом можно сделать вывод, что данные по взаимодействиям, представленные в Human Protein Atlas, неполны. В связи с этим имеет смысл обратиться за информацией в другие базы данных.

Поиск белков, взаимодействующих с NR1H4, при помощи STRING

STRING – агрегатор, позволяющий собрать информацию о белок-белковых взаимодействиях из разнообразных источников, с его помощью можно решать такие задачи, как:

- Построение сетей белок-белковых взаимодействий
- Анализ совместной встречаемости генов в различных организмах
- Анализ функционального обогащения в наборе генов

Для подтверждения предположения о том, что NR1H4 взаимодействует не только с NCOA1, был осуществлен поиск взаимодействий (известных из курируемых баз данных или экспериментальных данных), в которых участвует данный транскрипционный фактор.

Как оказалось, NR1H4 действительно взаимодействует с несколькими белками (Рис. 8), среди которых 2 коактиватора (NCOA1 и NCOA2), 2 корепрессора (NCOR1 и NCOR2), 3 белка RXR, а также белки SUMO1 (по-видимому, NR1H4 может подвергаться сумоилированию), PPARGC1A и MED1 (один из белков комплекса медиатора).

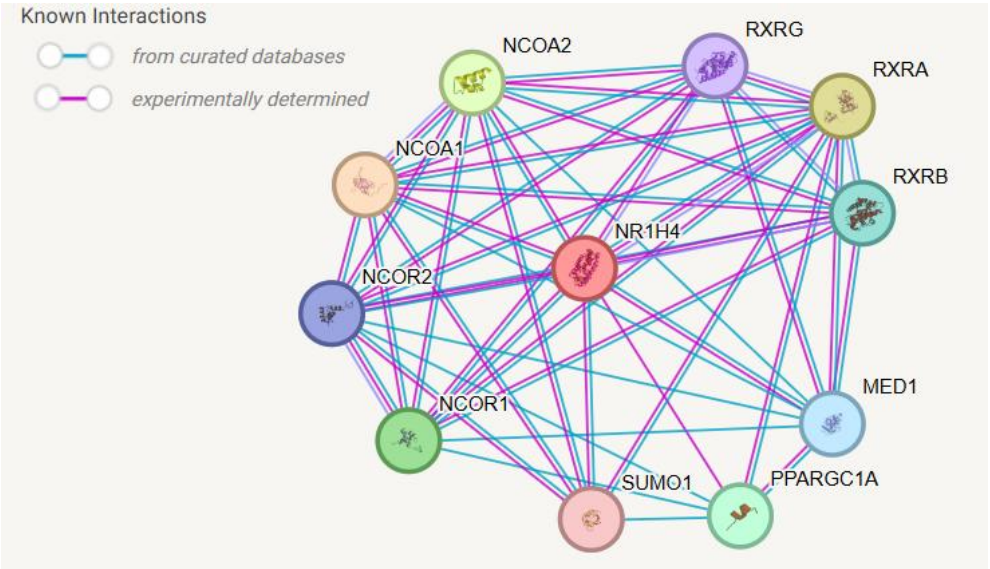

Рисунок 8. Схема белок-белковых взаимодействий для фактора транскрипции NR1H4, полученная при помощи STRING. Изображены только взаимодействия, описанные в курируемых базах данных или из экспериментальных данных. Также присутствуют ребра третьего типа (фиолетовые), соединяющие между собой гомологичные белки: NR1H4, RXRG, RXRA и RXRB.