Базы данных KEGG и GO

1. Входные данные

Входные данные представляют собой список из 22 ID генов.

Можно заметить несколько генов MMA* и TCN* (метаболизм B12), а также PRSS* и CTRB* (пищеварительные протеазы) в списке, что намекает на то, что в данном списке собраны гены, связанные с метаболизмом и транспортом витамина B12.

Цель - проанализировать связь между этими генами.

2. Групповой анализ

Для группового анализа я решил использовать сервис Enrichr.

Enrichr - это веб-сервис, позволяющий проводить анализ обогащения наборов генов. Он принимает на вход набор генов (список в нашем случае) и определяет, какие биологические пути, типы клеток, транскрипционные факторы или другие функциональные категории статистически значимо обогащены в этом списке.

Enrichr использует точный тест Фишера для для расчёта статистической значимости обогащения. Для учёта множественных сравнений (проверка множества категорий) применяется поправка Бенджамина-Хохберга, что рассчитывает q-value, которое контролирует долю ложных обнаружений (FDR). Важно, что поправка Бенджамина-Хохберга менее строгая, чем поправка Бонферрони, что снижает вероятность ложноотрицательного результата (пропустить истинно обогащённую категорию).

Enrichr также рассчитывает отношение шансов (Odds ratio), которое показывает, насколько чаще гены из нашего набора попадают в конкретную категорию по сравнению со случайным распределением среди всех генов генома.

Ещё рассчитывается комбинированный показатель (Combined score) по формуле: c = -log(p) * Odds ratio, где p - это p-value, полученное из точного теста Фишера. Комбинированный показатель позволяет объединить p-value и odds ratio, что позволяет более надёжно ранжировать результаты, чем каждый из показателей по отдельности. Также такая формула позволяет занизить оценку для обширных категорий генов и наоборот завысить оценку для категорий с малым количеством генов, что делает возможным обнаружение конкретных биологических процессов (более узких), а не только самых общих категорий.

С помощью Enrichr можно решать следующие задачи:

- Связать набор генов с заболеваниями или лекарственными чувствительностями. Подав список генов, ассоциированных с определенным состоянием (например, мутации в опухоли при терапии), можно определить через соответствующие библиотеки, какие болезни и классы препаратов связаны с этим набором.

- Определить биологические процессы, вовлечённые при дифференциальной экспрессии генов. При проведении RNA-seq между контрольной и экспериментальной группами можно получить список дифференциально экспрессированных генов. Enrichr позволяет провести анализ обогащения терминов GO и KEGG, помогая сформулировать гипотезу.

- Выявить транскрипционные факторы, регулирующие наблюдаемые изменения экспрессии генов. Например, если были получены гены определённого фенотипа, Enrichr может определить, какие транскрипционные факторы статистически значимо ассоциированы с этими генами.

Собственно, я провёл анализ обогащения при помощи Enrichr для моего списка генов.

Сначала посмотрим, в каких биологических путях задействованы наши гены. Смотреть будем по данным из базы Reactome за 2024 год. Мы получили вот такие результаты (рис.1): 68 путей, 25 значимых находок (с Adjusted P-value < 0,05). Видим, что лучшие находки связаны транспортом и метаболизмом витамина B12 (собственно мы так и предположили в начале анализа): Cobalamin (Cbl, Vitamin B12) Transport and Metabolism, Metabolism of Water-Soluble Vitamins and Cofactors или Defects in Cobalamin (B12) Metabolism. Знаем, что витамин B12 водорастворимый, поэтому пути связанные с водорастворимыми витаминами дополнительно подтверждают нашу первоначальную теорию.

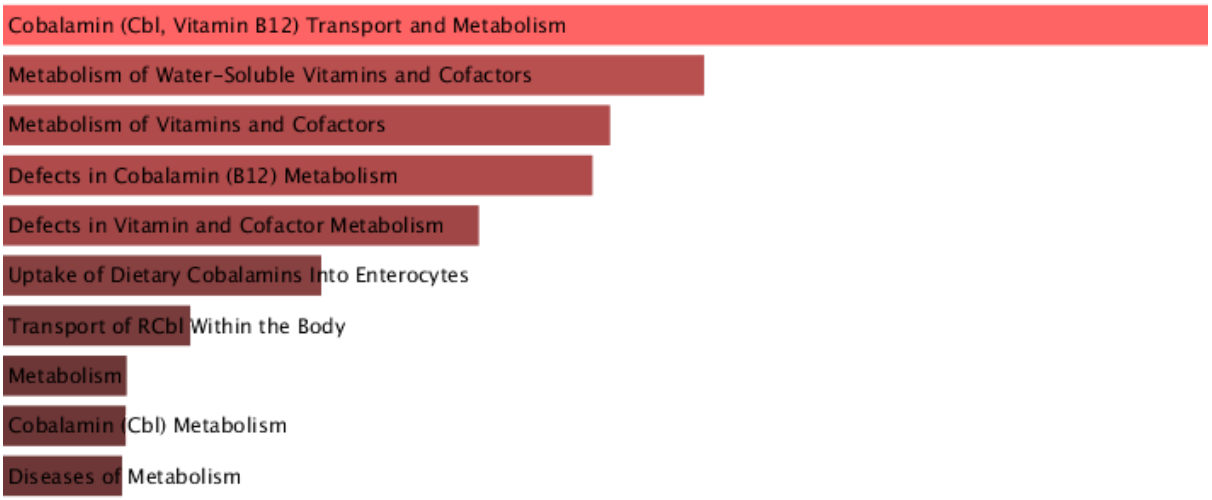

Рис.1. Результаты анализа обогащения биологических путей в рассматриваемом наборе генов. Величина столбцов отражает статистическую значимость обогащения, оцениваемую по p-value.

Теперь посмотрим три основные категории (рис.2.), с помощью которых Gene Ontology (GO) описывает любую функцию гена с разных, дополняющих друг друга сторон:

Для GO: Biological Process - 178 термина, 17 значимых.

Для GO: Cellular Component – 40 термина, 21 значимый.

Для GO: Molecular function – 39 термина, 24 значимых.

В результатах Biological Process видим основные процессы, связанные с метаболизмом и транспортом витамина B12 (кобаламина): Cobalamin Metabolic Process, Cobalamin Transport, Vitamin Transport. B12 - это азотсодержащее соединение, поэтому Nitrogen Compound Transport тоже подходит.

Результаты Cellular Component можем разделить на 2 обширные группы: первая - эндоцитоз-лизосомальная система (лизосома, её люмен и мембран; мембрана литической вакуоли; везикула, покрытая клатрином и её мембрана) - суть в том, что B12 захватывается через клатрин-опосредованный эндоцитоз, попадает в лизосому, откуда через мембранные транспортеры экспортируется в цитозоль, а затем часть его направляется в митохондрии; вторая - микроворсинки и щёточная кайма (пищеварительной системы) - характерные клеточные компоненты для процессов всасывания в кишечном эпителии и пищеварения.

Находки из анализа GO Molecular function можно также разделить на несколько групп. Здесь есть протеолитическая активность (Serine-type Endopeptidase Activity и Serine-type Peptidase Activity), характерная для пищеварительных ферментов поджелудочной железы, кофактор-связывающая активность (B12-зависимые ферменты) (Oxidoreductase Activity, Acting on Metal Ions, NAD or NADP as Acceptor, FAD Binding), связывание ионов и димеризация (Calcium Ion Binding, Protein Homodimerization Activity), необходимые для стабильной работы пищеварительных ферментов.

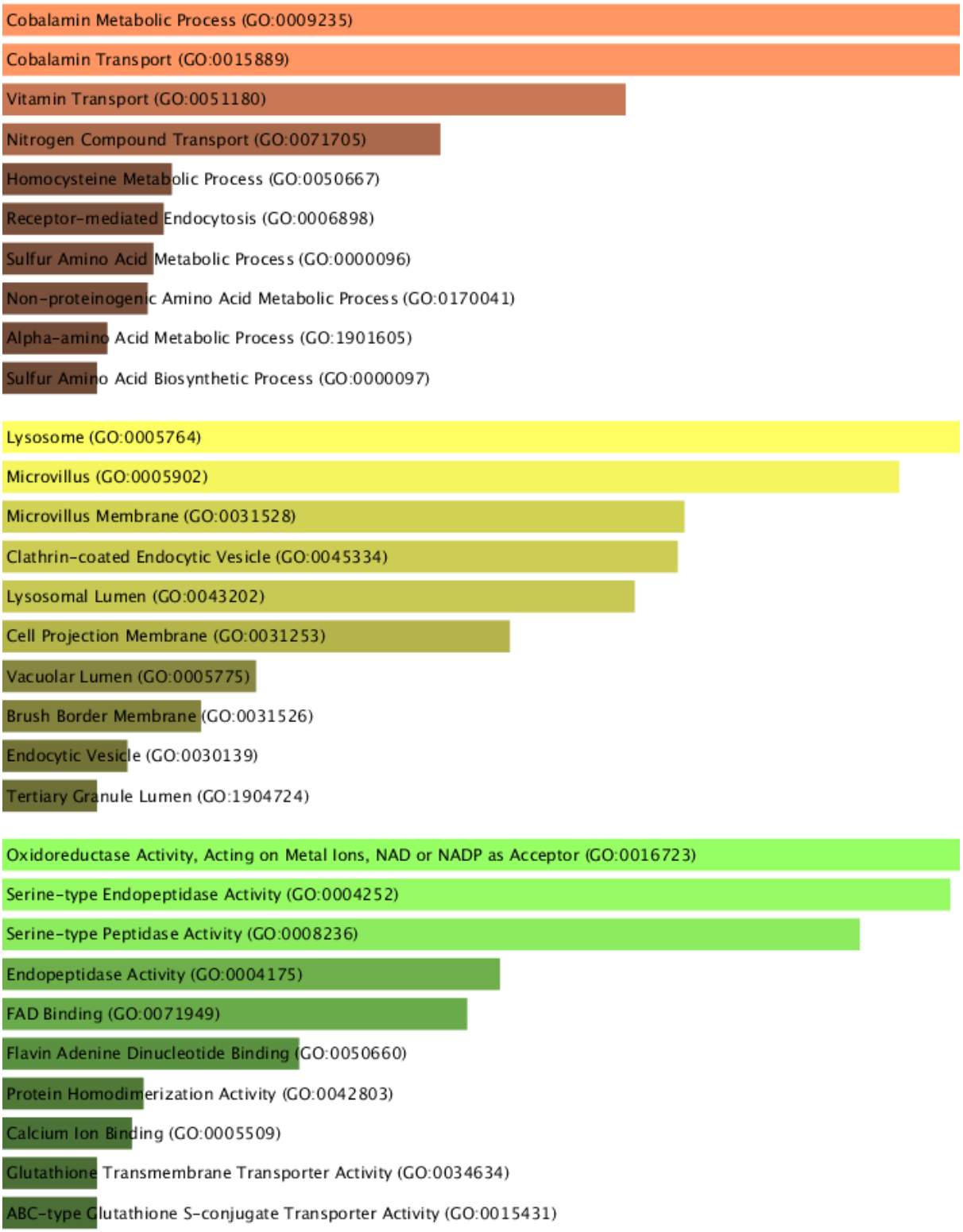

Рис.2. Обогащение категорий GO рассматриваемом наборе генов (оранжевый - Biological Process, жёлтый - Cellular Component, зелёный - Molecular function). Величина столбцов отражает статистическую значимость обогащения, оцениваемую по p-value.

После этого посмотрим на результаты из базы данных Orphanet Augmented за 2021 год (рис.3), хотим увидеть заболевания и патологии, с которыми связаны наши гены. Из 261 результата, 8 имеют одинаково хороший P-value, про них и поговорим. Видим синдром Имерслунда-Гресбека, связанный с нарушением всасывания витамина B12, и панкреатиты, связанные с нарушением работы протеаз в поджелудочной железе. Остальные находки я бы счёл за ложноположительные, так как они не имеют связи с моими генами.

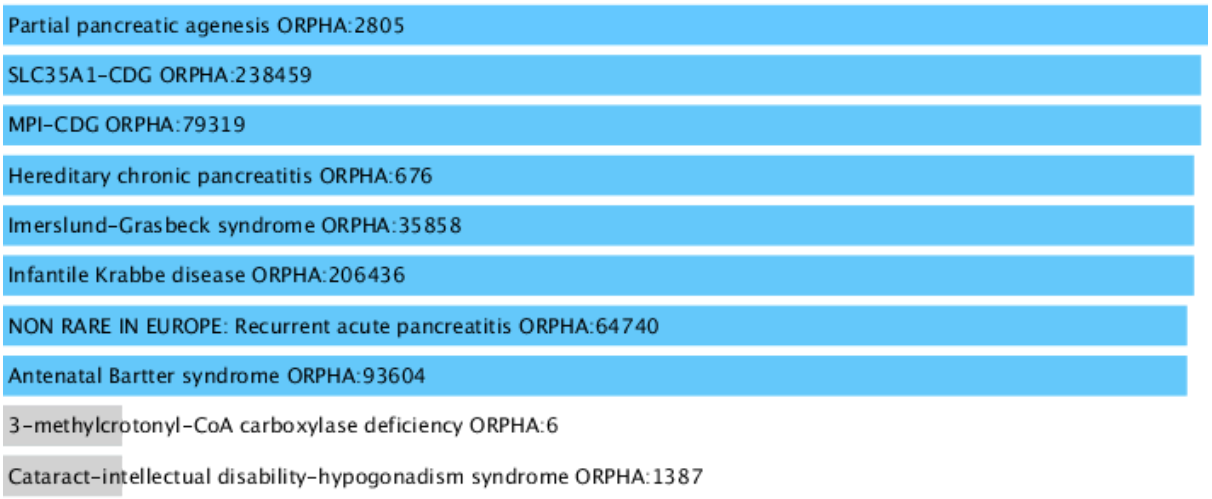

Рис.3. Заболевания, с которыми связаны рассматриваемые гены. Величина столбцов отражает статистическую значимость обогащения, оцениваемую по p-value.

И последнее на что посмотрим - это типы клеток, в которых экспрессируются гены из списка по базе данных CellMarker за 2024 год (рис.4). Всего 34 результата, 10 значимых. Здесь видим ацинарные клетки поджелудочной железы (которые как раз и секретируют пищеварительные ферменты и проксимальные канальцы почки (реабсорбция белков и витаминов, в том числе B12, из первичной мочи) как самые лучшие результаты поиска.

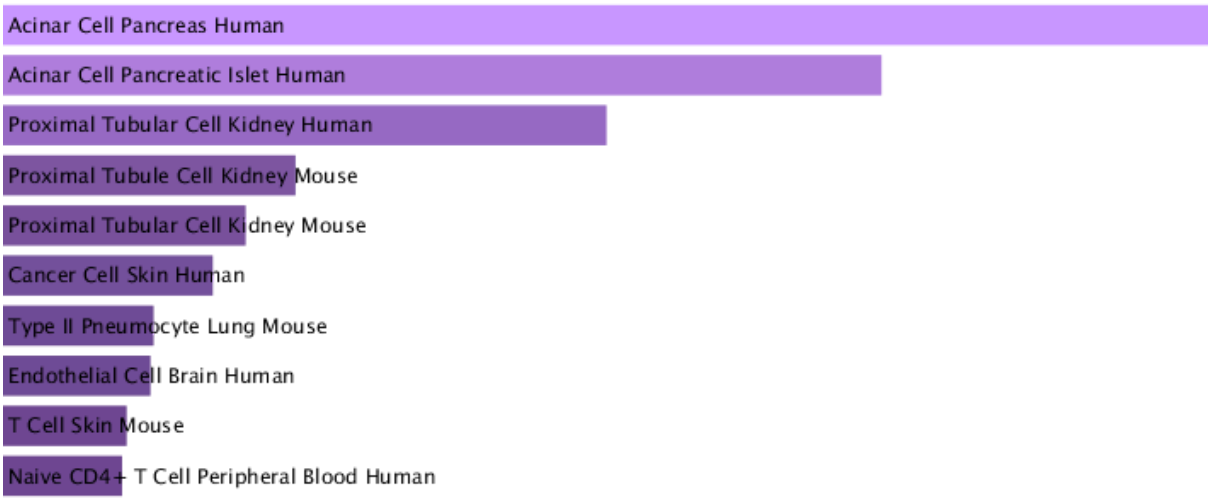

Рис.4. Типы клеток, в которых экспрессируются рассматриваемые гены. Величина столбцов отражает статистическую значимость обогащения, оцениваемую по p-value.

Выводы:

Проведённый анализ с помощью Enrichr с использованием нескольких библиотек (GO, Reactome, Orphanet, CellMarker) показал, что исследуемый набор генов функционально разделён на два независимых кластера.

Первый (я бы назвал его основным) кластер включает гены, обеспечивающие полный цикл метаболизма витамина B12: от захвата в кишечнике и внутриклеточного транспорта до использования в B12-зависимых ферментах. Эти гены локализованы в микроворсинках, эндоцитарных везикулах, лизосомах и митохондриях, а их мутации ассоциированы с синдромом Имерслунда-Гресбека и дефектами метаболизма кобаламина.

Второй кластер состоит из генов пищеварительных протеаз, экспрессирующихся преимущественно в ацинарных клетках поджелудочной железы и ассоциированных с панкреатитом.

3. Индивидуальный анализ гена из списка

Анализировать буду ген MMAA - кодирует белок MMAA (Metabolism of cobalamin associated A), отвечающий за транспорт витамина B12 в митохондрии, активацию фермента метилмалонил-КоА-мутазы (MMUT), который играет ключевую роль в метаболизме жиров и белков и помогает получить необходимую для работы активную форму B12.

Для индивидуального анализа гена MMAA я решил использовать сервис The human protein atlas.

The human protein atlas - это достаточно большая база данных, которая совмещает множество данных по локализации и функциям человеческих белков, объединяет данные на уровне тканей, клеток, субклеточных структур и отдельных молекул (лично меня HPA особенно привлекает своим интерфейсам, графиками и изображениями).С её помощью можно решать следующие задачи:

- Определять, в каких тканях и клетках экспрессируется интересующий нас белок и увидеть его субклеточную локализацию.
- Идентифицировать белки, которые экспрессируются строго в определенном типе клеток или ткани, но отсутствуют в других.
- Искать биомаркеры для диагностики и мониторинга рака.

Попробуем начать анализ с тканей, а затем углубляться в более мелкие структуры.

На рисунке 5 видим, что ген экспрессируется в различных тканях, но особенно высокая экспрессия наблюдается в печени (22,9 nTPM), также можно отметить ткани пищеварительной и мочевыделительной систем.

Сам белок высоко экспрессируется в почках, желудке, двенадцатиперстной кишке, тонком кишечнике, паращитовидной железе, коре головного мозга и в плаценте (рис.6).

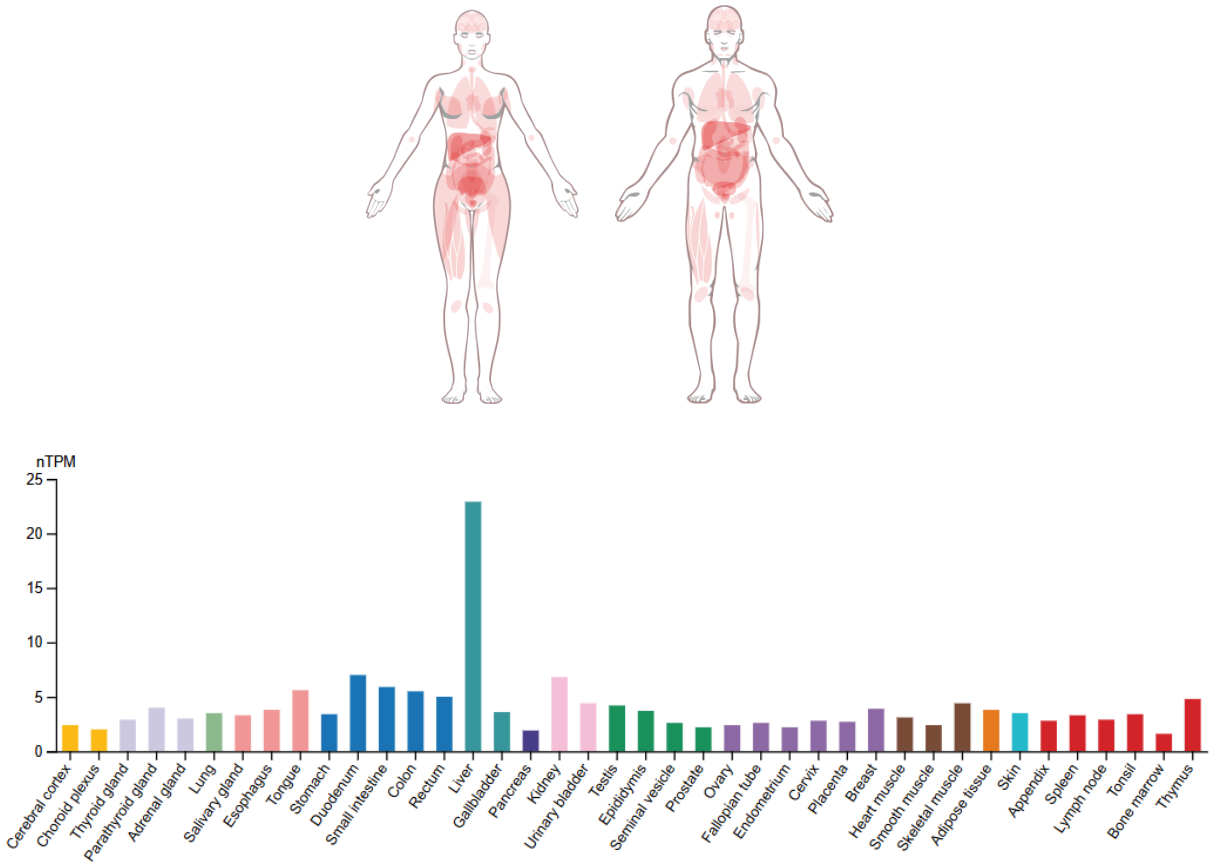

Рис.5. Данные по экспрессии гена MMAA в различных тканях и органах человека, приведенные в базе данных The Human Protein Atlas. Представлены графические изображения, показывающие районы

экспрессии данного гена в женском (слева) и мужском (справа) организмах. Ниже приведены данные по уровням экспрессии в тканях.

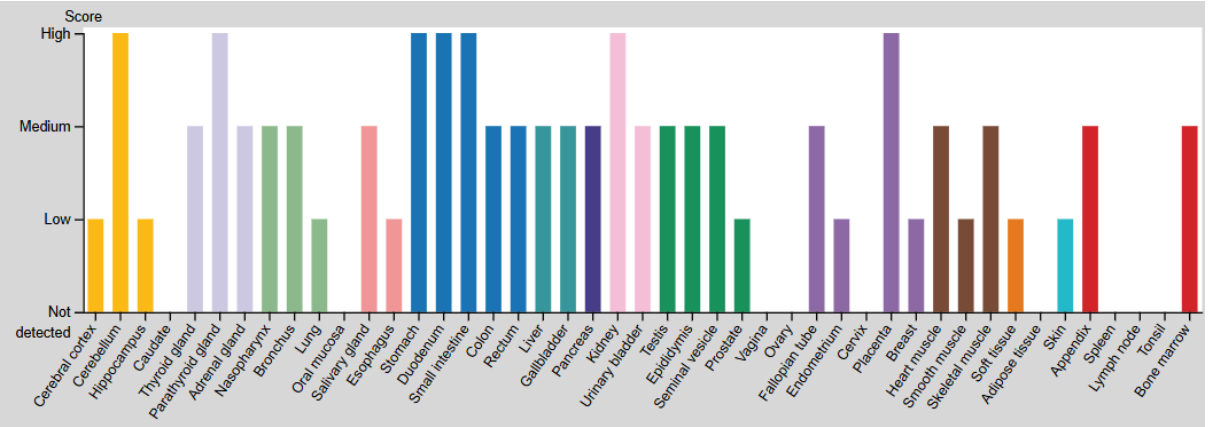

Рис.6. Данные по экспрессии белка MMAA в различных тканях и органах человека, приведенные в базе данных The Human Protein Atlas.

Далее видим, что данный ген имеет 6 транскриптов, то есть сам ген MMAA кодирует 6 изоформ мРНК. Белок MMAA локализуется в цитозоле (рис.7). И хотя раннее было сказано, что MMAA работает в митохондриях (транспорт B12, активация MMUT) - это не противоречие. Белок может синтезироваться в цитозоле и транспортироваться в митохондрии или иметь функции в обоих компартментах. Также видим, что уровень экспрессии белка может различаться от клетки к клетке в пределах одной популяции. Это возможно из-за разной метаболической активности клеток или разной потребности в продукции MMAA в зависимости от фазы метаболизма. Ещё можем заметить, что экспрессия белка не зависит от клеточного цикла. Это означает, что белок MMAA производится клеткой постоянно, а не только в какой-то конкретной фазе клеточного цикла. Поэтому, скорее всего, клетка всегда нуждается в уровне метаболизма B12.

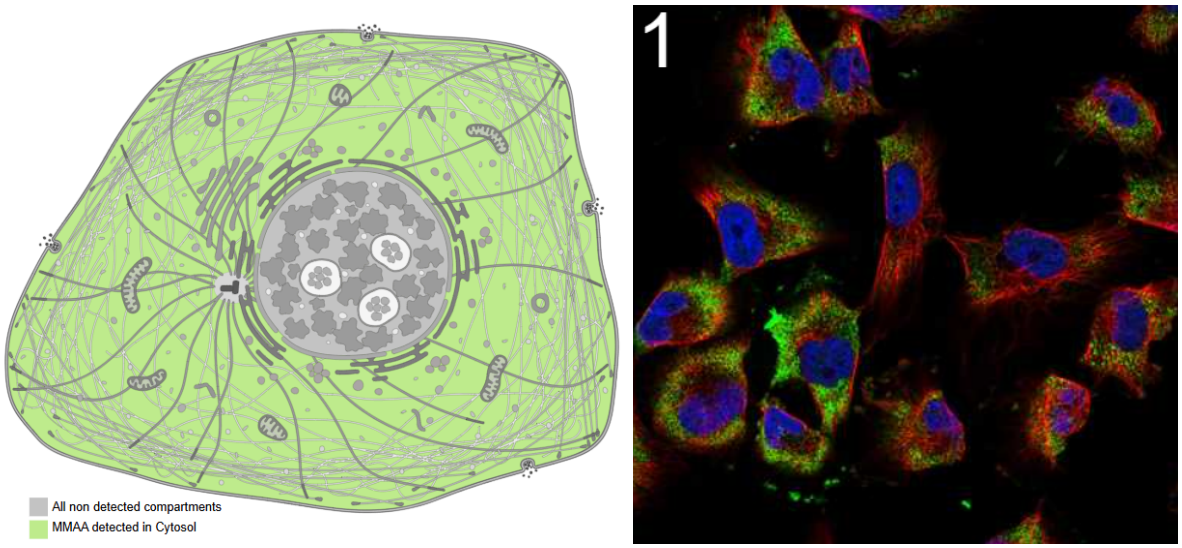

Рис.7. Изображение локализации белка MMAA в клетке. Модель локализации (слева) и и белок в реальных клетках (обозначен зелёным цветом, справа).

Теперь поговорим о взаимодействиях данного белка в клетке. Единственное взаимодействие MMAA - с белком MMUT (рис.8). При этом оценка взаимодействия достаточно высокая (ipTM: 0.69, pTM: 0.76), что говорит, что белки действительно связываются. Из литературы знаем, что MMAA связывается с MMUT и физически снижает скорость окисления кофактора, замедляя образование неактивной формы. Но также мы знаем, что MMAA связывается с другими белками (MMAB, MMACHC, MMADHC), которые здесь не указаны.

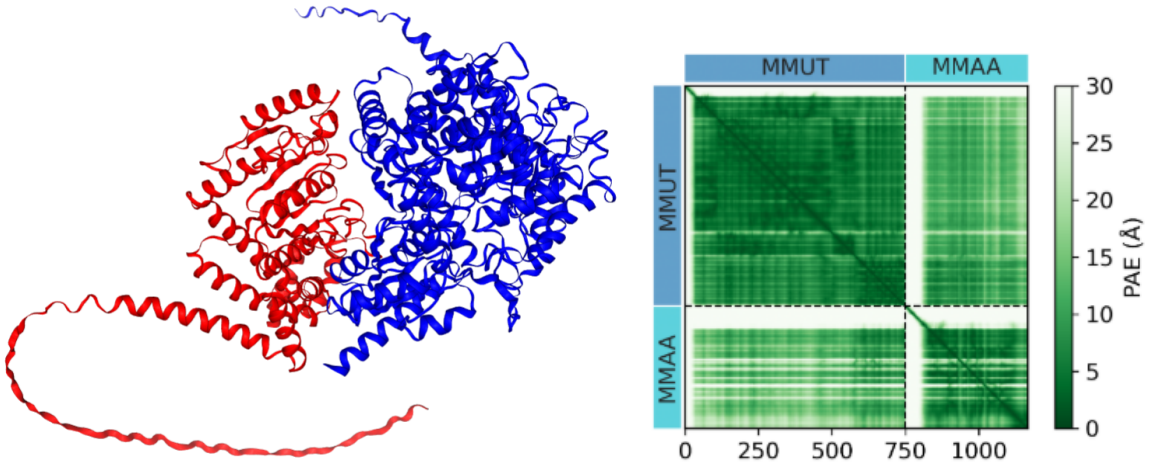

Рис.8. Взаимодействие MMAA с белком MMUT. 3D-модель взаимодействия белков (MMAA - красный, MMUT - синий, слева), карта контактов белков (справа).

Выводы:

Анализ гена MMAA с использованием с помощью базы данных The Human Protein Atlas позволил подтвердить его ключевую роль в метаболизме кобаламина. MMAA широко экспрессируется в тканях с высоким метаболизмом, а его белковый продукт преимущественно локализован в цитозоле. Наличие вариабельности экспрессии на уровне отдельных клеток указывает на гибкую регуляцию, в то время как независимость от клеточного цикла говорит о постоянной потребности клеток в MMAA. Высокое качество предсказанной модели взаимодействия с MMUT подтверждает их функциональную связь, однако стоит сказать о неполноте данных о взаимодействии белка в базе данных.