ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Межвузовская студенческая научная школа-конференция

Информационные технологии и системы

Биоинформатика

Федеральное государственное
бюджетное учреждение науки
Институт проблем
передачи информации
им. А. А. Харкевича
Российской академии наук

ИППИ РАН

Skoltech
Сколковский институт науки и технологий

18-21 февраля 2022

Оглавление

18.02.2022-21.02.2022

1. Сплайсинг

1.1 Анализ больших транкриптомных данных указывает на взаимодействие между альтернативным сплайсингом и интронным полиаденилированием

Авторы: М.А. Власенок[1], Д.Д. Первушин[1]

1- Сколковский Институт Науки и Технологий, Москва, Россия

Сплайсинг и полиаденилирование являются двумя важными этапами посттранскрипционной регуляции экспрессии эукариотических генов. Недавние исследования выявили множество случаев интронного полиаденилирования, приводящего к образованию дисфункциональных укороченных белков и соответствующим болезням человека. Однако, взаимодействие этих двух котранскриптционных процессов, как и связь между интронным полиаденилированием и альтернативным сплайсингом, в настоящее время мало изучено. Мы проанализировали большой массив данных RNA-seq, полученных в рамках проекта Genotype Tissue Expression (GTEx), чтобы одновременно идентифицировать и сопоставить тканеспецифическое использование сайтов полиаденилирования с тканеспецифическим альтернативным сплайсингом. Комбинация вычислительных методов, среди которых анализ коротких прочтений с внематричными аденинами (полиА-прочтения), выявила множество событий интронного полиаденилирования и полиаденилирования лариатов, что указывает на динамическую связь между двумя процессами.

1.2 Предсказание РНК-РНК взаимодействий по данным конформационного секвенирования РНК in situ

Авторы: С.Д. Маргасюк[1], С. Трифонова[2], Д. Первушин[1]

1 - Сколковский Институт Науки и Технологий, Москва, 2 - МГУ им. М.В. Ломоносова

Дальние взаимодействия в молекулах РНК, определяемые их структурой, играют важную роль в регуляции альтернативного сплайсинга, а также в патогенезе связанных с ним заболеваний. Наиболее перспективной высокопроизводительной технологией изучения глобальной структуры РНК является конформационное секвенирование РНК in situ, сочетающее лигирование близко расположенных участков молекул РНК, в том числе опосредованное РНК-связывающими белками, с глубоким секвенированием для глобального профилирования внутри- и межмолекулярных взаимодействий РНК. Основной целью данного

проекта является построение обобщенной модели для предсказания глобальной вторичной структуры пре-мРНК эукариотических транскриптов с учетом информации о конформационном секвенировании. Важный частный случай этой задачи – предсказание РНК-РНК контактов, которое во многом аналогично задаче анализа данных Hi-C, но также имеет ряд отличий, связанных с особенностями секвенирования РНК. Главное из них состоит в том, что РНК содержат интроны, которые соответствуют разрывам в чтениях РНК при картировании на геном, но РНК-РНК контакты также соответствуют разрывам в чтениях РНК при картировании на геном. Мы разработали вычислительный конвейер, основанный на двухпроходном картировании, который позволяет учесть одновременно оба типа разрывов.

1.3 Изучение специфичности взаимодействия РНК-связывающих белков с РНК

Авторы: С.Д. Маргасюк[1], О.В. Калинина[2], Д.Первушин[1]

1 - Сколковский Институт Науки и Технологий, Москва, 2 - Саарландский Университет, Саарбрюкен, Германия

РНК-связывающие белки (RBP) связываются с РНК с помощью РНК-связывающих доменов (RBD), которые распознают определённые элементы структуры или последовательности РНК. RBD может принадлежать к одной из нескольких групп родственных доменов, при этом специфичность внутри группы определяется небольшим числом специфичность-определяющих позиций (SDP). Логично предположить, что специфичность RBP должна определяться набором RBD, из которых он состоит, и аминокислотными остатками в SDP этих доменов. Нахождение SDP в группах родственных последовательностей – классическая задача биоинформатики, ранее решавшаяся для транскрипционных факторов. Целью данной работы является построение статистической модели, предсказывающей специфичность взаимодействия RBP с нуклеотидными последовательностями по информации о доменной архитектуре RBP и известным из литературы данным по футпринтингу RBP. Эта модель может быть обучена, например, на существующих данных CLIP. Модель, полученная в результате обучения, может быть более точной, чем существующие модели связывания; кроме того, она может быть применена для предсказания специфичности белков, для которых отсутствуют данные экспериментов футпринтинга.

1.4 Идентификация регуляторного непродуктивного сплайсинга в нормальных и раковых транскриптомах человека

Авторы: А.Г. Миронов[1], Д.Д. Первушин[1]

1 - Сколковский Институт Науки и Технологий, Москва, Россия

У эукариот существует механизм уничтожения транскриптов, содержащих преждевременные стоп-кодоны (PTC), называемый нонсенс-опосредованным распадом NMD). Многие РНК-связывающие белки (RBP) используют NMD для регуляции собственной экспрессии через отрицательную обратную связь, при которой белковый продукт гена связывается с кодирующей его мРНК и индуцирует в ней альтернативный сплайсинг, приводящий к появлению PTC. Этот механизм называется непродуктивным сплайсингом. RBP часто используют непродуктивный сплайсинг для ауто- и кросс-регуляции экспрессии друг друга и других белков, а нарушения этой регуляции могут приводить к образованию опухолей. Данный проект направлен на полнотранскриптомную идентификацию регуляторных событий непродуктивного сплайсинга, участвующих в тканеспецифичной и рако-специфичной регуляции экспрессии генов. С этой целью мы предлагаем биоинформатический анализ, объединяющий несколько общедоступных экспериментов по футпринтингу РНК-связывающих белков, отклику транскриптома на нокдаун, нокаут или суперэкспрессию RBP, секвенированию транскриптомов нормальных тканей человека из проекта Genotype Tissue Expression Project (GTEx), секвенированию транскриптомов опухолей из Атласа Ракового Генома (TCGA), и клиническому протеомному анализу опухолей (CPTAC). Предварительный анализ известных из литературы событий непродуктивного сплайсинга в генах PTBP2, GABBR1, FUS, HNRNPDL показал обнадеживающие результаты по анализу ткане- и опухолеспецифичности и поиску регуляторов. Предлагаемый проект позволит получить новое представление о биологии рака и может предложить новые мишени для генной терапии рака с помощью антисмысловых олигонуклеотидов.

2. Z-DNA

2.1 Сравнительный анализ предсказаний DeepZ для геномов мыши и человека

Автор: Н.С. Бекназарова младший научный сотрудник МЛБ ФКН НИУ ВШЭ, Москва, Россия

В современном мире, при стремительном росте количества информации, стало возможно использовать большие массивы данных для методов машинного обучение. Это явило отличные результаты для биоинформатических задач. Сделанные нами ранее подход машинного обучения DeepZ показал, что при помощи глубинного обучения можно добиться высокой предсказательной способности при предсказании регионов Z-ДНК человека. Тем не менее дальнейшее изучение поведение Z-ДНК необходимо и на других организмах, а так же сравнение разных организмов мужду собой. В этих целях мы получили экспериментальную разметку для генома мыши и изучили результаты ее с точки зрения машинного обучения. В результаты полученные модели смогли сделать полногеномную разметку для человека и мыши, которые были сравнены между собой. Также сами полученные модели были проинтерпретированны и результаты интерпретации моделей человека и мыши показали важные сходства и различия этих друх организмов. Из этих данных были полученные важные выводы о общих и различающихся ролях Z-ДНК, которые она играет в человеческом и мышином геномах.

2.2 Aire, Z-ДНК и организация хроматина

Автор: Ф.И. Павлов стажер-исследователь МЛБ ФКН НИУ ВШЭ, Москва, Россия
Научный руководитель: М.С. Попцова, МЛБ ФКН НИУ ВШЭ

Aire (autoimmune regulator) -- это транскрипционный фактор, который кодируется геном Aire и экспрессируется в клетках тимуса. Данный белок играет значимую роль в регуляторных процессах генома, так как активно вовлечен в процесс формирования толерантности иммунной системы. Мы знаем, что Aire отвечает за регуляцию тысяч генов, но при этом на текущий момент не имеем представления, за счет каких конкретно механизмов происходит эта регуляция.

Тем не менее, на сегодняшний день у нас есть свидетельства того, что вторичные структуры участвуют в регулировании процессов, протекающих в ДНК. Например, левозакрученная вторичная структура Z-DNA участвует в регуляции процесса транскрипции. В связи с этим нам интересно посмотреть на несколько вещей, Во-первых, участвует ли Z-ДНК в процессе регулирования экспрессии генов в тимусе. Во-вторых, оказывает ли Aire

воздействие на формирование Z-ДНК в тимусе.

В рамках исследования мы выделили четыре метода для определения регионов связывания Z-ДНК в геноме: аннотация регионов при помощи алгоритмического метода Z-Hunt (что дает чуть больше трети регионов Aire, которые потенциально могут образовывать Z-ДНК), аннотация генома с помощью алгоритма глубинного обучения DeepZ (примерно четверть Aire-пиков обогащена регионами из данной аннотации), экспериментально размеченные данные (обнаружилось значимое обогащение сразу для двух структур -- G4-квадруплексов и Z-ДНК), а также характерные для Z-ДНК регионы чередующейся пурин-пиримидиновой последовательности (в наших данных имеется значимое обогащение GC-повторами, причем значимость сравнима с экспериментальными данными для Z-ДНК).

В результате проведения bulk RNA-seq анализа экспрессии генов Aire-нокаута против дикого типа мы выделили группу генов, которые активируются Aire, а также соответствующую им группу ассоциированных Aire пиков. Наиболее высокая концентрация данного типа пиков обнаружилась в промоторах генов, которые активируются Aire. Данную подгруппу промотеров удалось более подробно проанализировать путем применения алгоритма Z-Hunt и составления профиля связывания Z-ДНК.

Таким образом, мы ожидаем, что в вопросе регулирования экспрессии генов нам удастся однозначно определить отличие между генами, которые активируют Aire, и остальными. Также одним из направлений работы является изучение механизмов взаимодействия результирующих данных с другими транскрипционными факторами и гистонами.

2.3 ADAR1 и ZBP1-индуцируемый некроптоз

Автор: А.Н. Федоров младший научный сотрудник МЛБ ФКН НИУ ВШЭ

ZBP1 - это один из ключевых сенсоров системы врожденного иммунитета. Его роль не до конца детализирована, но из структурных и экспериментальных данных ясно, что ZBP1 распознает Z-нуклеиновые кислоты с помощью доменов Zα и, в зависимости от их концентрации в клетке, вызывает клеточный иммунный ответ. Это, в конечном итоге, приводит к гибели клетки - апоптозу, некроптозу или пироптозу.

Одним из способов искусственно вызвать активацию ZBP1 является использование кураксина (CBL0137). Это мало исследованная молекула способна вызывать масштабные изменения трехмерной структуры хроматина, стимулировать формирование Z-ДНК и инициировать клеточную смерть, некроптоз. В литературе было продемонстрировано, что такая форма некроптоза является ZBP1-зависимой, однако детальный молекулярный

механизм ранее описан не был.

Благодаря анализу новых ChIP-seq экспериментов удалось установить, что сайты формирования Z-ДНК в эмбриональных мышиных клетках после обработки CBL0137 совпадают с сайтами связывания ZBP1. Это подтверждает гипотезу о Z-ДНК зависимой активации ZBP1. Более того, распределение сайтов связывания не случайно, они находятся преимущественно в 5`UTR потенциально активных LINE-1 повторов.

Таким образом, сформировано целостное представление о процессе CBL0137-зависимого некроптоза и роли Z-ДНК в нем: (1) кураксин взаимодействует с ДНК и вызывает B->Z конформационный переход в 5`UTR неповрежденных LINE-1 повторах; (2) ZBP1 посредством доменов Zα связывается с этими повторов и (3) инициирует имунный ответ, который и приводит к иммуногенной клеточной смерти, некроптозу.

3. Машинное обучение

3.1 Графовые нейронные сети для вторичных структур ДНК (1)

Автор: А.А.Войтецкий стажер-исследователь МЛБ ФКН НИУ ВШЭ, Москва, Россия
Научный руководитель: М.С. Попцова

В настоящее время машинное обучение активно применяется для анализа данных. Например, в биоинформатике его используют для исследования генома, его различных функциональных элементов. Машинное обучения помогает делать выводы из уже проведенных экспериментов, а также генерировать данные для проведения новых экспериментов. В этой работе мы исследовали графовые нейронные сети в задаче распознавания вторичных структур ДНК.

ДНК чаще всего существует в виде правозакрученной двойной спирали (B-ДНК), однако было обнаружено, что она может принимать и другие формы, такие как A-ДНК, Z-ДНК и G-квадруплексы, а также переходить от одной формы к другой. В данной работе предлагается использовать нейронные сети для распознавания участков ДНК, способных переходить из B-ДНК в Z-ДНК - левозакрученную спираль. Научная новизна работы заключается в реализации архитектуры GNN (Graph Neural Network), которая до этого не применялась к такой задаче, и последующем сравнении ее работы с уже исследованными CNN и RNN с помощью известных ML-метрик.

3.2 Графовые нейронные сети в задаче распознавания вторичных структур ДНК

Автор: А.К. Колчина стажер-исследователь МЛБ ФКН НИУ ВШЭ, Москва, Россия
Научный руководитель: М.С. Попцова

ДНК чаще всего существует в виде правозакрученной двойной спирали (B-ДНК), однако было обнаружено, что она может принимать и другие формы, такие как A-ДНК, Z-ДНК и G-квадруплексы, а также переходить от одной формы к другой. В данной работе предлагается использовать нейронные сети для распознавания участков ДНК, связанных с определенными гистоновыми метками и способных переходить из B-ДНК в Z-ДНК - левозакрученную спираль. Работа модели базируется на поиске мотивов нуклеотидов, характерных для конкретных гистоновых меток. Научная новизна работы заключается в том, чтобы реализовать архитектуру GNN (Graph Neural Network) и сравнить ее работу с уже исследованными CNN и RNN с помощью известных ML-метрик, и, в конце концов, сделать вывод о пригодности применения данного типа нейросетей в задаче распознавания вторичных

структур ДНК.

3.3 Применение трансформеров к задачам распознавания вторичных структур ДНК

Автор: А.В. Данилова стажер-исследователь МЛБ ФКН НИУ ВШЭ, Москва, Россия
Научный руководитель: М.С. Попцова

Данная работа направлена на выявление функциональных участков Z-ДНК с помощью такой модели глубокого обучения, как трансформер. Требуется решить задачу классификации последовательностей нуклеотидов на два класса: содержащие и не содержащие функциональные участки. В процессе работы планируется, во-первых, применить предобученную модель DNABERT, выполнив fine-tuning на подготовленном наборе данных, который представляет собой последовательности нуклеотидов. Во-вторых, построить собственную модель трансформера, обучить её на том же наборе данных и сравнить полученный результат с DNABERT. Далее планируется добавить к входным последовательностям эпигенетические данные и на этом наборе обучить модель трансформера. На последнем этапе будет проведена интерпретация полученного результата путем проецирования матрицы внимания, которое используется в модели трансформера, на входную последовательность данных.

3.4 Методы интерпретации искусственного интеллекта для задач геномики

Автор: Эмилия Шаймуратова студентка 4 курса ФКН НИУ ВШЭ, Москва, Россия
Научный руководитель: М.С. Попцова

Современные модели машинного обучения активно используются в биоинформатике. Точное предсказание - это всего лишь одно из необходимых требований к модели. Так как перед ученым стоит задача не только создать инструмент для работы с данными, но и понять биологический процесс, необходимо создать параллель между работой модели и процессами, протекающими в организме.

Часто современные алгоритмы машинного обучения сравнивают с черными ящиками, интерпретация параметров которых является сложной задачей. Необходима разработка методов интерпретации параметров обучаемых нейронных сетей.

Я буду рассматривать задачу интерпретации параметров нейронной сети на примере задачи предсказания участков Z-ДНК на основе омиксных данных. Целью работы будет

оценить вклад гистоновых меток при распознавании Z-ДНК. Для решения данной задачи будет использован метод интегрированных градиентов.

3.5 Доменная адаптация для аннотации геномов квадруплексами

Автор: П.В. Латышев стажер-исследователь МЛБ ФКН НИУ ВШЭ, Москва, Россия
Научный руководитель: М.С. Попцова

Одной из активно изучаемых вторичных структур ДНК являются G-квадруплексы, формирующиеся в насыщенных гуанином участках. Данные структуры могут играть важную биологическую роль, также известна их связь с различными заболеваниями, такими как рак. В настоящее время было представлено несколько экспериментальных методов картирования генома такими структурами (G4-Seq и G4-ChIP-Seq). К сожалению, в настоящий момент доступны данные только по ограниченному количеству организмов и клеточных линий. В данной работе изучаются модели, способные предсказывать формирование G-квадруплекса на определенном участке ДНК, а также способы применения таких моделей к геномным последовательностям других организмов, используя методы Unsupervised Domain Adaptation. В ходе работы была получена полногеномная аннотация G-квадруплексами генома мыши с использованием доменно-состязательной нейронной сети (DANN).

4. Хроматин

4.1 Топологический анализ данных Hi-C как метод определения трехмерной структуры хроматина

Авторы: В.А. Кузнецов[1], О.Д. Крюков[1], Е.Е. Храмеева[1]

1 - Центр наук о жизни, Сколковский Институт Науки и Технологий, Москва, Россия,

Хроматин имеет множество уровней структурной организации (Bonev, Cavalli 2016), каждый из которых влияет на важнейшие клеточные процессы, следовательно имеет биомедицинское значение (Misteli 2010). Методы изучения структуры хроматина можно условно разделить на 2 группы: микроскопические методы и методы определения конформации хромосом (3C). Хотя микроскопические методы и обладают рядом преимуществ, они лимитированы изучением конкретного региона и не позволяют оценить трехмерную структуру в целом. Результатом Hi-C, одного из 3C методов, является полногеномная карта контактов. Методы дальнейшего анализа Hi-C данных позволяют выявить основные компоненты структуры хроматина (Lieberman-Aiden et al 2009): компартменты, хромосомные территории, ТАДы, петли — но эти методы либо требуют больших вычислительных ресурсов, либо вносят определенные допущения, ограничивающие возможное поле поиска (Wolff et al 2020).

В нашем анализе мы решили применить один из методов топологического анализа — персистентные гомологии, для анализа Hi-C данных и поиска координат петель хроматина. Для решения поставленной задачи нами был разработан алгоритм перевода матрицы контактов в матрицу неевклидовых расстояний с выполнением аксиом метрики, топологического анализа облака точек с помощью библиотеки Eirene языка программирования Julia, фильтрации результата и визуализации. В дальнейшем нами планируется проведение тестирования на различных данных, кросс-валидации с другими методами поиска петель, определение воспроизводимости, чувствительности к шуму и разреженности.

Источники и литература:

1. Bonev, B., & Cavalli, G. (2016). Erratum: Organization and function of the 3D genome. Nature Reviews Genetics, 17(12), 772–772. https://doi.org/10.1038/nrg.2016.147

2. Misteli, T. (2010). Higher-order Genome Organization in Human Disease. Cold Spring Harbor Perspectives in Biology, 2(8), a000794–a000794. https://doi.org/10.1101/cshperspect.a000794

3. Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., &

Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science, 326(5950), 289–293. https://doi.org/10.1126/science.1181369

4. Wolff, J., Backofen, R., & Grüning, B. (2020). Loop detection using Hi-C data with HiCExplorer. https://doi.org/10.1101/2020.03.05.979096

4.2 Хроматиновые петли у D. discoideum

Авторы: И. Жегалова[1,2,3], С. Ульянов[2,4], А. Галицына[1], А. Лужин[4], И. Плетенев[1], Е. Храмеева[1]

1 - Сколковский институт науки и технологий, Москва, Россия, 2 – МГУ им Ломоносова, Москва, Россия, 3 - А.А. Харкевича Институт проблем передачи информации, Москва, Россия, 4 - Институт биологии гена РАН, Москва, Россия

Целью настоящего исследования является получение всесторонней картины изменений конформации хроматина, экспрессии генов и их регуляции при развитии Dictyostelium discoideum. Были получены данные из экспериментов Hi-C, ATAC-Seq и RNA-seq через 0, 2, 5 и 8 часов после индукции голодания.

Анализ карт Hi-C показал отсутствие ТАДов и классических петель, характерных для геномов млекопитающих, однако, на разрешении 2 килобазы видны яркие точки обогащенных контактов. Эти точки топологически, вероятно, представляют собой петли, хотя и заметно отличающиеся от петель млекопитающих. Было отмечено, что петли покрывают более половины генома диктиостелиума, имеют медианный размер 20 килобаз, а их положения весьма консервативны, с пересечением более 90% между стадиями.

Было показано, что уровень экспрессии внутри петли значительно превышает таковой в границах петель, также петли характеризуются конвергентным расположением генов вокруг оснований петель. Дальнейший анализ выявил гетерогенность петель по уровням экспрессии и сигналу меток активного хроматина. Удалось выделить четыре кластера в зависимости от их поведения.

Среди петель были выявлены также “вытянутые”, в дальнейшем названные el-loops. Такие петли характеризуются пониженной экспрессией основания, вдоль которого они вытянуты при симметричном для оснований сигнале меток активного хроматина. Наблюдается также обогощение в покрытии WGS в основаниях петель, что может сигнализировать о наличии ориджинов репликации, однако, необходимо провести дополнительные исследования этого вопроса.

Чтобы проверить, существует ли связь между уровнями экспрессии и уровнями инсуляции, мы выполнили кластеризацию RNA-seq и оценили соответствующие показатели

инсуляции в каждый момент времени для каждого кластера отдельно. Все кластеры имели противоположные тенденции в уровнях экспрессии и инсуляции, что свидетельствует о возможной связи этих параметров. Более того, изменения в уровнях экспрессии и инсуляции между стадиями развития статистически значимо коррелируют с коэффициентом -0.4.

Моделирование с использованием позиций конвергентых генов показало, что петли возникают на пересечении позиций конвергнтных генов. Вместе с предыдущими результатами, это свидетельствует о возможном механизме формирования петель у Dictyostelium discoideum по типу экструзии, при этом терминация экструзии зависит от активности транскрипции. Однако не исключена и регуляторная роль петель, для проверки этой гипотезы требуется дополнительные анализы.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-34-90058

4.3 Изменение дальних контактов хроматина D. discoideum при переходе к многоклеточной стадии развития

Авторы: И.А.Плетенев[1], И.В.Жегалова[1,2,3], А.А.Галицына[1], С.В.Ульянов[3,4], М.С.Гельфанд[1,2], Е.Е.Храмеева[1]

1 - Центр наук о жизни, Сколковский институт науки и технологий, Москва, Россия, 2 - Институт проблем передачи информации им. А.А. Харкевича, РАН, Москва, Россия, 3 - МГУ им. М.В. Ломоносова, Москва, Россия, 4 - Институт биологии гена, РАН, Москва, Россия, Б

Реакцией амебы D. discoideum на нехватку питательных веществ является переход от одноклеточной стадии развития к многоклеточной, биологические механизмы которого до конца не ясны. Одним из вероятных факторов может быть изменение трехмерной структуры хроматина, так как она играет существенную роль в развитии других эукариот (Zheng et al., 2019). Тем не менее, структура хроматина D. discoideum остается мало изученной.

В данной работе исследуются карты, полученные методом фиксации конформации хромосом Hi-C. Карты Hi-C фиксируют несколько временных точек перехода D. discoideum из одноклеточной к многоклеточной стадии развития. Этот переход сопровождается изменениями в структуре хроматина, в частности, контактов между локусами, находящимися на расстоянии более 1 Mb на одной хромосоме или на разных хромосомах. Ручная разметка выявила две группы таких дальних контактов: контакты, исчезающие после начала голодания, а также контакты специфичные для последнего исследуемого часа голодания. Большинство дальних контактов включают в себя три и более локуса, образуя сеть межхромосомных взаимодействий. Некоторые локусы, образующие дальние контакты, обогащены гистонами с

модификацией H3K9me3.

Изменение дальних хроматиновых контактов D. discoideum при переходе к многоклеточности указывает на возможную роль таких контактов в регуляции данного процесса, что могли бы подтвердить дальнейшие эксперименты.

Источники и литература:

1. Zheng H, Xie W. The role of 3D genome organization in development and cell differentiation. Nat Rev Mol Cell Biol. 2019 Sep;20(9):535-550. doi: 10.1038/s41580-019-0132-4.

4.4 Характеристика хроматиновых петель на разных стадиях клеточного цикла Saccharomyces cerevisiae

Авторы: К.А.Ульянов[1], И.В.Жегалова[1], Е.Е.Храмеева[1], М.С.Гельфанд[1]

1 - Сколковский институт науки и технологий, Москва, Россия

Исследования клеточного ядра показали, что хроматин внутри него не располагается беспорядочно, его архитектура следует определённым правилам и закономерностям, начиная от укладки хромосом в хромосомные территории, заканчивая формированием топологически ассоциированных доменов и петель. Известно, что выпетливание нити ДНК играет важную роль во многих процессах связанных с компактизацией хромосом, репликаией ДНК и регуляцией промотор-энхансерных взаимодействий.

В качестве основного механизма выпетливания хроматиновых фибрилл широко распространена модель экструзии петель (Alipour & Marko, 2012). Согласно ей, петли в интерфазном ядре образуются за счёт протягивания нити ДНК через молекулу когезина. Инсуляторный белок CTCF ограничивает движение когезинового кольца — связанные с ДНК в противоположных ориентациях молекулы CTCF действуют как якоря в основаниях петли (de Wit et al., 2015).

Модель экструзии хорошо изучена в геномах млекопитающих, однако хроматиновые петли можно обнаружить на Hi-C картах полученных для организмов, которые не имеют белка CTCF. Примером могут стать дрожжи Saccharomyces cerevisiae, у которых петли начинают формироваться с наступлением фазы S клеточного цикла и сохраняются вплоть до митоза (Costantino et al., 2020). Примечательно, что положение этих петель остаётся постоянным с прохождением клеточного цикла. Из результатов ChIP-seq анализа известно, что петли дрожжей также содержат когезин (Costantino et al., 2020).

Упомянутые выше особенности дрожжей позволяют предположить существование иных барьерных механизмов, альтернативных классической модели. например, последние работы показали, что РНК-полимераза II в определённых условиях способна блокировать

продвижение когензинового кольца вдоль нити ДНК (Brandão et al., 2019).

Задачей представленной работы является проведение комплексного анализа петель у Saccharomyces cerevisiae с целью выявления особенностей их организации и формирования. В фокусе внимания лежат данные по генной активности, открытости хроматина, положению ориджинов репликации и эпигенетических модификаций гистонов в районах оснований и внутренних участков петель.

Источники и литература:

1. Alipour, E., & Marko, J. F. (2012). Self-organization of domain structures by DNA-loop-extruding enzymes. Nucleic acids research, 40(22), 11202–11212. https://doi.org/10.1093/nar/gks925

2. de Wit, E., Vos, E. S., Holwerda, S. J., Valdes-Quezada, C., Verstegen, M. J., Teunissen, H., Splinter, E., Wijchers, P. J., Krijger, P. H., & de Laat, W. (2015). CTCF Binding Polarity Determines Chromatin Looping. Molecular cell, 60(4), 676–684. https://doi.org/10.1016/j.molcel.2015.09.023

3. Costantino, L., Hsieh, T. S., Lamothe, R., Darzacq, X., & Koshland, D. (2020). Cohesin residency determines chromatin loop patterns. eLife, 9, e59889. https://doi.org/10.7554/eLife.59889

4. Brandão, H. B., Paul, P., van den Berg, A. A., Rudner, D. Z., Wang, X., & Mirny, L. A. (2019). RNA polymerases as moving barriers to condensin loop extrusion. Proceedings of the National Academy of Sciences of the United States of America, 116(41), 20489–20499. https://doi.org/10.1073/pnas.1907009116

4.5 Проблемы исследования структуры хроматина по картам Hi-C

Отчет подгруппы теоретической биоинформатики структурной биологии хроматина

Авторы: А. Галицына Центр наук о жизни, Сколковский Институт Науки и Технологий, Москва, УНЦ Биоинформатика ИППИ РАН, Н. Быков Центр наук о жизни, Сколковский Институт Науки и Технологий, Москва, А. Кондрашина Центр наук о жизни, Сколковский Институт Науки и Технологий, Москва, К. Перевощикова УНЦ Биоинформатика ИППИ РАН, Факультет Биоинженерии и Биоинформатики МГУ, Д.Скрипка УНЦ Биоинформатика ИППИ РАН, Факультет Биоинженерии и Биоинформатики МГУ, А. Школиков Факультет Биоинженерии и Биоинформатики МГУ, М.С. Гельфан Центр наук о жизни, Сколковский Институт Науки и Технологий, Москва, УНЦ Биоинформатика ИППИ РАН, Факультет Биоинженерии и Биоинформатики МГУ

В последние годы мы наблюдаем расцвет области исследования трехмерной организации хроматина ядра. Принято считать, что он произошел благодаря изобретению метода высокопроизводительной фиксации конформации хромосом, или Hi-C [1]. Однако прорыв в понимании устройства структурной укладки ДНК объясняется не столько удобной и качественной экспериментальной техникой, сколько совершенно новым формальным представлением этой структуры (в виде контактной карты), и развитием сопутствующего аналитического аппарата.

Контактная карта ДНК представляет собой матрицу, строки и столбцы которой представляют собой участки ДНК одинакового размера, а в каждой ячейке записано число - количество контактов соответствующих участков, наблюдаемых в эксперименте. Такая карты может быть представлена в виде изображения, на которой выражены структуры компартментов (“шахматной доски”), топологических доменов (ТАДов, ярких квадратов вдоль диагонали) и петель (ярких точек обогащенных взаимодействий, например, промоторов и своих энхансеров). С другой стороны, контактная карта представляет собой матрицу, для анализа которой применим весь арсенал методов линейной алгебры и машинного обучения.

В нашей учебной подгруппе мы задались общей задачей проработки областей структурной биологии хроматина на стыке анализа данных Hi-C и алгоритмов биоинформатики.

Во-первых, мы разбираем метод поиска компартментов хроматина и показываем, что этот метод может быть применен для поиска более тонких структур хроматина - субкомпартментов. В проекте Дмитрия Скрипки мы предлагаем унифицированную схему проверки и подбора параметров алгоритмов поиска компартментов и субкомпартментов на симулированных картах Hi-C. Мы улучшаем существующий алгоритм поиска субкомпартментов, находим оптимальные параметры и планируем провести его валидацию на реальных данных Hi-C разных видов.

Во-вторых, мы предлагаем решение нового типа задачи вывода временной динамики ТАДов хроматина в зависимости от стадии развития эмбриона. В своем проекте Николай Быков решает эту задачу с помощью программы HiChew - гибкого алгоритма кластеризации границ ТАДов во временных рядах измерений карт Hi-C. Мы применяем этот метод для эмбриогенеза трех организмов и на основании этого предполагаем вывести общие принципы укладки хроматина в эмбриогенезе.

Инструменты, разработанные в ходе этого проекта (поиск ТАДов, поиск потенциальных перестроек как точек карты с выраженным градиентом) планируется использовать для проекта Анны Кондрашиной, в котором мы проводим анализ структуры нового (для области структурной биологии хроматина и эпигенетики) организма - гриба-базидиомицета Schizophyllum commune. Сборка генома этого организма несовершенна, и мы планируем улучшить ее с помощью Hi-C. В частности, нам потребуется провести анализ

данных Hi-C на уровне отдельно взятых ридов: для этого мы проведем поиск перестроек по методу анализа химерных ридов, разработанным Александрой Галицыной. Мы рассчитываем разработать метод детекции позиций геномных перестроек по данным Hi-C, который может пролить свет на уникальные эволюционные свойства генома этого гриба.

В-третьих, мы прорабатываем проблему промотор-энхансерных взаимодействий в проекте Кристины Перевощиковой и связываем данные Hi-C с данными новых экспериментов по пертурбации геномных регионов. Мы задаемся вопросам: связано ли формирование стабильного контакта хроматина на картах Hi-C с активностью энхансера по отношению к подчиненному промотору, расположенному в тех же геномных регионах? Предварительный анализ приводит к опровержению этой гипотезы.

Наконец, в проекте Алексея Школикова, мы ставим вопрос о кодировании структуры хроматина в последовательности ДНК. Мы тестируем возможности самых популярных и мощных нейронных сетей для решения этой задачи в нашем унифицированном фреймворке Chimaera. Так как структура хроматина неизбежно зависит от эпигенетического состояния клеток (а значит, и клеточного типа), мы не ставим задачу точного предсказания, но пытаемся определить ключевые факторы, которые влияют на укладку.

В заключение, в большинстве проектов мы используем общий подход для исследования возможностей и ограничения методов анализа данных Hi-C: сравнение работы алгоритмов для разных видов организмов. Это не только дает нам методологическую основу для тестирования алгоритмов, но и открывает перспективу анализа как эволюционных особенностей разных групп эукариот, так и общих законов формирования укладки хроматина.

Источники и литература:

1. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. science. 2009 Oct 9;326(5950):289-93.

4.6 Поиск химерной ДНК в последовательности ридов Hi-C

Автор: А. Галицына Центр наук о жизни, Сколковский Институт Науки и Технологий, Москва, УНЦ Биоинформатика ИППИ РАН

Картирование химерной ДНК и поиск участков стыковки разных молекул - проблема, часто встречающаяся при обработке данных секвенирования. Так, в RNA-Seq это прочтения, попадающие на сплайсированные сайты мРНК [1], в Hi-C - прочтения, попадающие на стыки

лигирования [2], а в обычном полногеномном секвенировании - риды, попадающие на перестройки [3].

Ранее мы разработали метод поиска химер для Hi-C одиночных клеток [4, 5] и показали, что аннотация стыков лигирования таких химер по сайтам рестрикции помогает добиться лучшего качества контактных пар, если в эксперименте была использована полимераза с повышенной частотой смены матрицы [4]. Теперь этот метод был имплементирован как одна из стратегий аннотации пар Hi-C в инструменте pairtools [6]*.

В качестве демонстрации мы применяем метод для поиска химер рибосомальной ДНК полногеномного секвенирования и показываем, что метод позволяет эффективно обнаруживать геномные перестройки. Недостатком подхода является традиционная проблема множественных картирований ДНК на повторы, которая приводит к невозможности интерпретации части химерных случаев**.

Мы применяем этот метод для ряда данных семейства фиксации конформации хромосом: Hi-C с разной длиной рида [7], Micro-C [8], а также MC-3C [9]. Мы показываем, что использование нового типа картирования химер позволяет увеличить количество интерпретируемых контактов (в случае MC-3C - в десятки раз).

Мы проводим качественный анализ полученных контактов и показываем, что такие пары имеют свойства шкалирования от геномного расстояния, похожие на обыкновенные пары. Мы разделяем контакты, подтвержденные сайтами рестрикции, и предположительные артефакты. Оказывается, что соотношение цис- к транс-контактам предположительных артефактов оказывается измененным.

Таким образом, предложенный метод не только увеличивает выход контактных пар методов фиксации конформации хромосом, но и улучшает их качество.

* - Часть работы выполняется в коллаборации с Антоном Голобородько (IMBA, Вена, Австрия) и коллективом Open2C.

** - Часть работы выполняется в коллаборации с Олегом Денисенко (University of Washington, США).

Источники и литература:

1. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature biotechnology. 2019 Aug;37(8):907-15.

2. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature methods. 2012 Oct;9(10):999-1003.

3. Lu N, Li J, Bi C, Guo J, Tao Y, Luan K, Tu J, Lu Z. ChimeraMiner: An Improved Chimeric Read Detection Pipeline and Its Application in Single Cell Sequencing. International journal of molecular sciences. 2019 Jan;20(8):1953.

4. Ulianov SV, Zakharova VV, Galitsyna AA, Kos PI, Polovnikov KE, Flyamer IM, Mikhaleva EA, Khrameeva EE, Germini D, Logacheva MD, Gavrilov AA. Order and stochasticity in the folding of individual Drosophila genomes. Nature Communications. 2021 Jan 4;12(1):1-7.

5. Kos PI, Galitsyna AA, Ulianov SV, Gelfand MS, Razin SV, Chertovich AV. Perspectives for the reconstruction of 3D chromatin conformation using single cell Hi-C data. PLoS computational biology. 2021 Nov 18;17(11):e1009546.

6. https://github.com/open2c/pairtools

7. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014 Dec 18;159(7):1665-80.

8. Hsieh TH, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ. Mapping nucleosome resolution chromosome folding in yeast by micro-C. Cell. 2015 Jul 2;162(1):108-19.

9. Tavares-Cadete F, Norouzi D, Dekker B, Liu Y, Dekker J. Multi-contact 3C reveals that the human genome during interphase is largely not entangled. Nature structural & molecular biology. 2020 Dec;27(12):1105-14.

4.7 Изменения трехмерной структуры хроматина в развитии

Авторы: Николай Быков, Александра Галицына, Михаил Гельфанд Сколковский институт науки и технологий, Москва, Россия

Ранее было показано, что трехмерная структура хроматина постепенно устанавливается во время раннего эмбриогенеза мухи [1], мыши [2, 3], рыбы [4, 5] и человека [6]. Таким образом, на разных стадиях раннего развития можно наблюдать разную архитектуру хроматина.

Главными архитектурными компонентами хроматина традиционно считаются компартменты и топологически ассоциированные домены (ТАДы) [7, 8]. Ранее было показано, что границы ТАДов устанавливаются во время раннего эмбриогенеза в локусах, обогащенных генами домашнего хозяйства и эпигенетическим сигналом [1, 9, 10]. Структурные изменения в границах ТАДов могут привести к нарушениям в экспрессии близлежащих генов, и, как следствие, к возникновению заболеваний развития или рака [11].

ТАДы демонстрируют разные скорости созревания в процессе эмбрионального развития [1]. Однако взаимосвязь между скоростью созревания ТАДов, эпигенетической

регуляцией, композицией генов и их экспрессией в раннем эмбриогенезе остается не до конца изученной [1, 12, 13].

Чтобы изучить эти взаимосвязи в ходе развития организма, можно искать границы ТАДов и впоследствии кластеризовать их в пространстве стадий эмбриогенеза. Так как не существует универсального общепринятого решения для поиска оптимальной разметки границ ТАДов и анализа динамики их инсуляции во времени [14, 15], нами было подготовлено обновление разработанной ранее программы HiChew [16]. В частности, мы разработали и реализовали алгоритм для поиска оптимальной разметки границ ТАДов, соответствующей ожидаемому размеру ТАДа, и последующей кластеризации динамики инсуляции найденных границ в пространстве стадий развития.

Мы протестировали алгоритм и показали, что динамика инсуляции в кластерах границ ТАДов коррелирует с геномными и эпигеномными характеристиками. В частности, в соответствии с нашими предыдущими выводами [16], скорость созревания границ ТАДов положительно коррелирует с обогащением генов домашнего хозяйства в локусах границ. Мы также показали, что динамика инсуляции границ ТАДов связана с обогащением эпигенетического сигнала ATAC-Seq в локусах этих границ. Кроме того, вместе с участниками проекта “Time series for DNA-DNA interactions” на ШМТБ-2021 мы применили алгоритм HiChew к данным Hi-C эмбриогенеза мухи, мыши, рыбы и человека, и показали, что во времени каждого эмбриогенеза есть доминирующий кластер ТАДов с преобладающей активацией хроматина.

Мы благодарим Ольгу Сигалову за предоставленные данные по эмбриогенезу мухи. Мы также благодарим участников проекта “Time series for DNA-DNA interactions” на ШМТБ-2021: Анну Калыгину и Дмитрия Скворцова – за плодотворное сотрудничество и ценный опыт. Проект поддержан грантом РФФИ 21--34--70051 для Екатерины Храмеевой. Программа HiChew доступна по ссылке https://github.com/encent/hichew.

Источники и литература:

1. C.B.Hug et al. (2017) Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription, Cell, 169:216–228.

2. Z.Du et al. (2017) Allelic reprogramming of 3D chromatin architecture during early mammalian development, Nature, 547:232–235.

3. Y.Ke et al. (2017) 3D chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis, Cell, 170:367–381.

4. C.L.Wike et al. (2021) Chromatin architecture transitions from zebrafish sperm through early embryogenesis, Genome research, 31.6:981-994.

5. A.Labudina, J.A.Horsfield (2021) The three-dimensional genome in zebrafish

development, Briefings in Functional Genomics.

6. X.Chen, Y.Ke, K.Wu et al. (2019) Key role for CTCF in establishing chromatin structure in human embryos, Nature, 576:306–310.

7. E.P.Nora et al. (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre, Nature, 485:381–385.

8. J.R.Dixon et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions, Nature, 485:376–380.

9. S.A.Blythe, E.F.Wieschaus (2015) Zygotic genome activation triggers the DNA replication checkpoint at the midblastula transition, Cell, 160:1169–1181.

10. S.V.Ulianov et al. (2016) Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains, Genome research, 26:70–84.

11. M.Yu, B.Ren (2017) The three-dimensional organization of mammalian genomes, Annual review of cell and developmental biology, 33:265–289.

12. C.P.Fulco et al. (2019) Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations, Nature genetics, 51:1664–1669.

13. R.Stadhouders et al. (2018) Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming, Nature genetics, 50:238–249.

14. N.Sauerwald, A.Singhal, C.Kingsford (2020) Analysis of the structural variability of topologically associated domains as revealed by Hi-C, NAR genomics and bioinformatics, 2.

15. M.Forcato et al. (2017) Comparison of computational methods for Hi-C data analysis, Nature methods, 14:679–685.

16. N.S.Bykov et al. (2020) HiChew: a Tool for TAD Clustering in Embryogenesis, International Symposium on Bioinformatics Research and Applications, 381–388.

4.8 Анализ структуры компартментов хроматина

Авторы: Дмитрий Скрипка Факультет биоинженерии и биоинформатики МГУ, Александра Галицына ИППИ РАН

Разбиение хроматина на два компартмента, видимое на Hi-C картах как паттерн шахматной доски, известно более десяти лет и продолжает активно изучаться. Каждый компартмент представляет собой пространственно сближенные регионы хромосом, которые могут располагаться в геноме как рядом, так и на значительных геномных расстояниях (в том числе на соседних хромосомах), что отличает их он топологически ассоциированных доменов (ТАДов) - локальных геномных структур. Разбиение на компартменты в целом соответствует делению хроматина на активный (эухроматин) и неактивный (гетерохроматин).[1, 2, 4]

Предложены разнообразные механизмы сегрегации хроматина на компартменты: разделение фаз, движущей силой которого является взаимодействие между участками одного типа, а также локализация неактивного компартмента на периферии за счет взаимодействием с ядерной ламиной.[3]

В последние годы было замечено, что компартменты имеют внутреннюю структуру. Так, Falk et al. отмечают выделенное расположение в ядре факультативного хроматина, а Spracklin et al. обнаружили 3 субкомпартмента гетерохроматина и2 субкомпартмента эухроматина, из которых только один представляет собой эффективно сближенные участки.

Целью нашей работы является разработка программы для поиска субкомпартментов по карте Hi-C, определение границ ее применимости и применение к новым организмам, на которых субкомпартменты ранее не были размечены.

Общая схема алгоритма поиска субкомпартментов, от которой мы исходили, аналогична описанной в [5]: производится поиск собственных векторов предобработанных карт Hi-C с наибольшими собственными значениями, координаты собственных векторов кластеризуются, полученные кластеры затем интерпретируются исходя из их положений на хромосоме с помощью HMM. Для тестирование и подбора параметров алгоритма мы разработали симулятор карт Hi-C, который позволяет сгенерировать “золотой стандарт” компартментов и соответствующую им карту Hi-C. Нам удалось добиться верного нахождения субкомпартментов в симулированной матрице, однако при степени выраженности субкомпартментов значительно превышающей выраженность компартментов.

На данный момент мы начали адаптацию алгоритма для реальных матриц Hi-C (в частности, дрозофилы).

В ходе дальнейшей работы мы планируем два ключевых улучшения предложенной схемы. Во-первых, мы усовершенствуем (как с помощью поиска параметров, так и замены и добавления промежуточных шагов) алгоритм для нахождения более слабых субкомпартментов. Во-вторых, мы оптимизируем его реализацию для реальных данных. Для оценки достоверности предсказания планируется использовать итеративный подход: мы будем брать исходные данные Hi-C, находить в них субкомпартменты, симулировать карту Hi-C с такими субкомпартментами, искать субкомпартменты в новосгенерированной матрице и наблюдать за схождением параметров алгоритма.В частности, мы будем наблюдать и сохранятьраспределение длин отдельных протяжённых регионов внутри субкомпартмента, чтобы использовать этоя как дополнительнуюя информацию после шага кластеризации для уточнения кластеров.

Источники и литература:

1. Lieberman-Aiden, Erez et al. “Comprehensive mapping of long-range interactions reveals folding principles of the human genome.” Science (New York, N.Y.) vol. 326,5950 (2009): 289-93.

doi:10.1126/science.1181369

2. Mirny LA, Imakaev M, Abdennur N. Two major mechanisms of chromosome organization. Curr Opin Cell Biol. 2019;58:142-152. doi:10.1016/j.ceb.2019.05.001

3. Heterochromatin drives organization of conventional and inverted nuclei Martin Falk, Yana Feodorova, Natasha Naumova, Maxim Imakaev, Bryan R. Lajoie, Heinrich Leonhardt, Boris Joffe, Job Dekker, Geoffrey Fudenberg, Irina Solovei, Leonid Mirny bioRxiv 244038; doi: https://doi.org/10.1101/244038

4. Irina Solovei, Katharina Thanisch, Yana Feodorova, How to rule the nucleus: divide et impera, Current Opinion in Cell Biology, Volume 40, 2016, Pages 47-59, ISSN 0955-0674,

https://doi.org/10.1016/j.ceb.2016.02.014.

5. Heterochromatin diversity modulates genome compartmentalization and loop extrusion barriers George Spracklin, Nezar Abdennur, Maxim Imakaev, Neil Chowdhury, Sriharsa Pradhan, Leonid Mirny, Job Dekker bioRxiv 2021.08.05.455340; doi: https://doi.org/10.1101/2021.08.05.455340

4.9 Роль структуры хроматина в промоторно-энхансерных взаимодействиях млекопитающих

Авторы: Кристина Перевощикова Факультет биоинженерии и биоинформатики МГУ, Александра Галицына ИППИ РАН

В настоящее время отсутствует единая модель, объясняющая природу промотор-энхансерных взаимодействий млекопитающих. Предполагается, что в возникновении промотор-энхансерных контактов основную роль играют два независимых фактора: “биохимическая совместимость” промотора с энхансером, вероятно, обусловленная наличием на промоторе и энхансере определенных сайтов посадки транскрипционных факторов, и структура хроматина.[1]

Роль “биохимической совместимости” между промоторами и энхансерами в настоящее время активно изучается, и имеющиеся данные говорят о том, что часть энхансеров активирует любые, помещенные рядом с ними промоторы в равной степени, в то время как некоторые энхансеры характеризуются селективностью по отношению к активируемым промоторам.[1,2] Также было показано, что степень активации транскрипции с расположенных во внехромосомных генетических конструкциях промоторов энхансерами не связана с частотой контактов этих двух элементов в пространстве, что заставляет думать о структуре хроматина и биохимической совместимости как о независимых факторах.[1]

Многие исследования подтверждают роль структуры хроматина в формировании и

функционировании промотор-энхансерных пар. На значимость структуры хроматина указывает в частности то, что большая часть пар промотор-энхансер расположена в пределах одного ТАДа.[3] Существуют экспериментальные доказательства того, что нарушение границы ТАДа, отделяющей активный энхансер от некоторой группы генов, может приводить к увеличению частоты контактов между энхансером и этими генами и, как следствие, к увеличению экспрессии этой группы генов по сравнению с клетками, где граница домена осталась интактна.[4,5] Однако эффект от нарушения границы домена на экспрессию не всегда однозначен. Иногда нарушение границы домена может оказывать незначительное влияние на функционирование промотор-энхансерных пар, локализованных внутри домена и препятствовать работе промотор-энхансерных пар, локализованных вне домена и разделенных этим самым доменом. [5,6]

Ввиду накопления большого количества данных о том, какие промоторы и энхансеры образуют между собой пары а также полногеномных карт контактов клеток млекопитающих в высоком разрешении мы решили провести биоинформатический анализ и систематизировать понимание роли структуры хроматина в регуляции экспрессии.

Таким образом, целью нашей работы является выяснение влияния структуры хроматина на формирование пар промотор-энхансер и на силу активации промотора энхансером.

В ходе работы были проанализированы 664 пары промотор - энхансер, выделенные путем dCAS9 CRISPRi скрининга на линии клеток K562.[3] Каждая пара описывается значением FC (fold change), характеризующим то, насколько меняется экспрессия гена при инактивации энхансера. Наш анализ показал, что FC пары промотор энхансер не зависит от того, сближены ли эти регуляторные элементы вследствие локализации на границах общей петли или нет. Более детальный анализ показал, что FC не связан и с частотой контактов между промотором и энхансером на Hi-C карте. Эти результаты, вероятно, указывают на то, что сила активации промотора энхансером слабо зависит от структуры хроматина, либо модель этой зависимости более сложная и требует учета дополнительных факторов, как, например, “биохимической совместимости” промотора и энхансера.

Наша дальнейшая работа предполагает определение роли структуры хроматина в формировании пар промотор энхансер, другими словами, мы будем искать ответ на вопрос, есть ли связь между структурой хроматина и тем, какой из нескольких альтернативных промоторов будет активироваться энхансером.

Источники и литература:

1. Martinez-Ara, M., Comoglio, F., van Arensbergen, J. & van Steensel, B. Systematic analysis of intrinsic enhancer-promoter compatibility in the mouse genome. (2021) doi:10.1101/2021.10.21.465269.

2. Bergman, D. T. et al. Compatibility logic of human enhancer and promoter sequences. bioRxiv 2021.10.23.462170 (2021) doi:10.1101/2021.10.23.462170.

3. Gasperini, M. et al. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell 176, 1516 (2019).

4. Xiao, J. Y., Hafner, A. & Boettiger, A. N. How subtle changes in 3D structure can create large changes in transcription. Elife 10, (2021).

5. Despang, A. et al. Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. Nature Genetics vol. 51 1263–1271 (2019).

6. Yokoshi, M., Segawa, K. & Fukaya, T. Visualizing the Role of Boundary Elements in Enhancer-Promoter Communication. Mol. Cell 78, 224–235.e5 (2020).

4.10 Модель глубокого обучения для предсказания карт Hi-C по нуклеотидной последовательности и поиска паттернов ДНК, значимых для предсказания

Авторы: А.С.Школиков Факультет биоинженерии и биоинформатики, Московский государственный университет, А.А.Галицына Центр наук о жизни, Сколковский институт науки и технологи1, УНЦ Биоиформатика ИППИ РАН

Hi-C - один из наиболее популярных методов изучения пространственной организации генома. Результатом эксперимента и последующей обработки данных секвенирования является карта, отражающая частоту контактов участков генома друг с другом. Полученные карты могут отражать такие структуры, как компартменты хроматина, петли и топологически ассоциированные домены. Было показано, что наличие некоторых из этих структур коррелирует с определенными паттернами в последовательности ДНК, такими как сайты связывания белков или локальные изменения GC-состава. Понимание закономерностей, лежащих в основе формирования пространственных структур в геномной ДНК, может предоставить большие возможности для изучения многих аспектов молекулярной биологии. Машинное обучение, используемое для предсказания карт Hi-C на основе последовательности ДНК, может помочь найти такие паттерны. Мы использовали модель глубокого обучения на основе нейронных сетей, включающую в себя сверточные слои и механизм внимания, которая принимает в качестве входных данных последовательность участка ДНК и предсказывает для него карту контактов. Модель отличается от использованных в других работах[1] для решения аналогичной задачи применением архитектуры, использующей механизм внимания (использована архитектура, применявшаяся для предсказания эпигенетических свойств по последовательности ДНК[2]) и использованием автоэнкодера (нейронной сети, которая обратимо преобразует данные в скрытое представление меньшей размерности) для кодирования карт Hi-C. Добавление механизма внимания позволило лучше учитывать дальние

взаимодействия. Автоэнкодер позволил ускорить обучение, сделать его более интерпретируемым и снизить чувствительность модели к шуму в картах Hi-C. Также новая архитектура модели легко масштабируется и может быть использована для предсказания по последовательностям большого диапазона длин. Для анализа модели были созданы инструменты поиска важных для предсказаний модели паттернов в ДНК, основанные на анализе сверточных фильтров, матриц внимания и градиентов модели. Модель была независимо обучена предсказывать карты Hi-C трех организмов: Drosophila melanogaster, Saccharomyces cerevisiae и Homo sapiens. Для всех организмов коэффициент корреляции Пирсона между предсказанными и истинными картами составил более 0,5. Для человека анализ модели четко показал важность мотивов связывания CTCF для предсказаний.

Источники илитература:

1. G. Fudenberg, D. R. Kelley, K. S. Pollard (2020) Predicting 3D genome folding from DNA sequence, Nature Methods, 17(11):1111-1117.

2. Žiga Avsec et al. (2021) Effective gene expression prediction from sequence by integrating long-range interactions, Nature Methods, 18, 1196–1203.

4.11 A proposal to query biological context for non-linearity in genotype data

Author: Elena Nabieva Skoltech; IITP

State-of-the-art methods for using genotype data for phenotype prediction are mostly linear1,2. While there is ongoing effort to apply to this task the methods of deep learning that have been very successful in other domains, such as image processing, these approaches have not yet shown definitive advantage over linear ones3. Perhaps their underwhelming performance in this area can be explained not just by the inherent linearity of genomic data, but also by the lack of exploitable genomic locality in those interactions that do exist. I propose to explore the possibility that the search for nonlinearity in genome-wide data can be informed by incorporating context-specific biological knowledge, such as the data on enhancer-promoter interactions or HiC.

Sources and literature:

1. Qian J. et al. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. PLOS Genetics 16(10): e1009141. https://doi.org/10.1371/journal.pgen.1009141

2. Medvedev A. et al. Human genotype-to-phenotype predictions: boosting accuracy with nonlinear models. medRxiv 2021.06.30.21259753; doi:

https://doi.org/10.1101/2021.06.30.21259753

3. Bellot P, de Los Campos G, Pérez-Enciso M. Can Deep Learning Improve Genomic Prediction of Complex Human Traits? Genetics. 2018 Nov;210(3):809-819. doi: 10.1534/genetics.118.301298.

4.12 Комплексный биоинформатический анализ последовательностей белков хроматина человека

Атовры: А.К. Грибкова Биологический факультет, МГУ, Москва, Шайтан А.К. Биологический факультет, МГУ, Москва, МЛБ ФКН НИУ ВШЭ

В регуляции молекулярных процессов в клеточном ядре задействовано огромное количество белков хроматина, которые выполняют разнообразные функции - узнают пост-трансляционные модификации гистонов, передвигают нуклеосомы, привлекают транскрипционную машинерию для экспрессии генов, изменяют степень конденсации ДНК и др. Не так давно было показано, что некоторые биохимические реакции в хроматине проходят в результате образования немембранных капель при процессе разделения жидких фаз (liquid-liquid phase separation).

Целью данной работы было выявление физико-химических и архитектурных особенностей белков хроматина и, в частности, белков различных функциональных классов хроматина. Для этого были предложены критерии классификации белков на белки хроматина, другие белки ядра и белки цитоплазмы из баз данных и литературных источников, а также с помощью разработанной функциональной иерархической классификации белков хроматина. Были проанализированы физико-химические свойства белков, доменная архитектура, фракции неупорядоченных регионов и регионов низкой сложности, общее количество белков хроматина разных классов в клетках человека.

Было показано, что по ряду параметров белки хроматина статистически значимо отличаются от белков других групп сравнения. Например, по фракциям заряженных и экстремально заряженных белков. При более детальном анализе функциональных групп было выявлено, что положительно заряженные белки преобладают в классах гистонов, HMG, и белков, модифицирующих ДНК; а отрицательно заряженные - в белках, связанных с транскрипцией, теломерными регионами, в гистоновых шаперонах, в белках, осуществляющих процессинг гистонов. Распределения фракций аминокислот также статистически значимо различаются, по сравнению с белками цитоплазмы: в хроматине выше медианные значения фракций полярных и небольших аминокислот, и ниже значения фракций гидрофобных, ароматических и алифатических аминокислот.

Особый интерес представляет доменная архитектура белков хроматина, позволяющая им дирижировать ядерными процессами. В белках хроматина статистически значимо выше медианное значение неупорядоченной фракции и меньше медианное значение доменной фракции, по сравнению с белками цитоплазмы. Не смотря на это, в хроматине преобладают фракции белков с общим количеством доменов больше 3-ех, по сравнению с другими группами. Однако это достигается не увеличением разнообразия доменов, а увеличением количества доменов одного типа, различия между которыми не способна уловить аннотация Pfam.

Таким образом, в работе выявлены физико-химические и архитектурные особенности белков хроматина, позволяющих им выполнять специфические функции в клеточном ядре и участвовать в разделении жидких фаз.

Исследование выполнено за счет гранта Российского научного фонда № 18-74-10006-П, https://rscf.ru/project/18-74-10006/.

4.13 Сравнение молекулярно-динамических моделей разных форм ДНК

Авторы: Г.М.Выходцев, А.К.Шайтан, МГУ им М.В.Ломоносова Москва, Россия; МЛБ ФКН НИУ ВШЭ, Москва, Россиия

Открытая Уотсоном и Криком конформация ДНК преобладает в живых организмах и называется B-ДНК. Но это не единственная биологически значимая форма ДНК.

Цель данной работы заключается в сравнительном изучении структуры и конформации B и Z формы ДНК методом молекулярной динамики, сравнении молекулярно-динамических моделей этих форм и создании базы данных для дальнейшего изучения биологических функций Z-ДНК. Для этого были рассчитаны траектории d(GC)6 ДНК дуплексов в B и Z форме длительностью в 1 микросекунду, было проведено сравнение геометрических параметров моделей ДНК с литературными значениями и между собой и сравнение стабильности этих моделей.

4.14 Разработка методов анализа данных секвенирования нового поколения в экспериментах по лентивирусной трансдукции мезенхимальных стволовых клеток

Авторы: Д.Д.Кожевникова[1,2], Д.В.Карпенко[3], А.Е.Бигильдеев[3], А.К. Шайтан[1,2]

[1]Биологический факультет, МГУ им М.В.Ломоносова Москва, Россия; [2] Международная лаборатория биоинформатики ФКН НИУ ВШЭ, Москва, Россия; [3]

Лаборатория физиологии кроветворения ФГБУ “НМИЦ Гематологии” Минздрава России

Работа посвящена исследованию пролиферативного и дифференцировочного потенциала популяции мезенхимных стволовых клеток (МСК) в красном костном мозге мышей.

Задача заключалась в анализе клонального состава (количества и размера клонов) клеток, составляющих строму очагов эктопического кроветворения. Клетки предварительно были маркированы лентивирусным вектором, который интегрируется в ДНК в условно-случайном месте, и тем самым обеспечивает наследуемый маркер всего клеточного потомства трансдуцированной клетки в виде сайта интеграции.

Пробоподготовка повторяла процедуру, описанную в статье [1] и была устроена таким образом, что после фрагментации ДНК ультразвуком к каждой молекуле ДНК пришивались линкеры, состоящие из нескольких участков: участка, общего для всех образцов, уникального для каждого образца, и баркода, уникального для каждой молекулы ДНК. Исходя из процедуры пробоподготовки была создана схема фильтрации данных и выявлена необходимость кластеризации ридов по последовательностям баркодов, находящихся в них, и по расположению сайтов интеграции в геноме. Код алгоритма обработки данных был написан на языке Python с использованием библиотек BioPython и HTSeq, выравниваие ридов производилось с помощью программы BWA. Визуализация кластеров последовательностей была создана с помощью библиотек Networkx и Pyvis.

Источники и литература:

1. Sherman E, et al. INSPIIRED: A Pipeline for Quantitative Analysis of Sites of New DNA Integration in Cellular Genomes. Mol Ther Methods Clin Dev. 2016 Dec 18;4:39-49. doi: 10.1016/j.omtm.2016.11.002.

4.16 Интегративное моделирование супрануклеосомной структуры хроматина

Авторы: Г.С.Тимохин [1,2], А.К.Шайтан [1,2]

[1] Международная лаборатория биоинформатики ФКН НИУ ВШЭ, Москва, Россия

[2] Биологический факультет, МГУ им. М.В. Ломоносова, Москва, Россия

Организация эукариотического хроматина на уровне компактизации нуклеосомной фибриллы – супрануклеосомном уровне – долгое время оставалась недоисследованной. При этом процессы, протекающие на этом уровне, имеют ключевое значение для регуляции транскрипции. Согласно современным представлениям, в ядрах клеток млекопитающих,

благодаря сложному взаимодействию процессов выпетливания и микрофазного разделения, начиная с супрануклеосомного уровня, осуществляется разделение хроматина на иерархически организованнные компартменты, в пределах которых регуляторные элементы сближаются с регулируемыми ими промоторами. Однако компактизация и компартментализация хроматина носит несистемный характер – характер компактизации определяется различными видоспецифическими и тканеспецифическими эпигенетическими факторами.

Целью данной работы было построение многофакторных моделей организации хроматина на супрануклеосомном уровне в нормальных эмбриональных стволовых (hESC) и малигнизированных (HeLa) клетках H.Sapiens и в клетках S.cerevisiae дикого типа на основании данных Micro-C, MNase-seq и ChIP-seq с использованием методов грубозернистого моделирования и дальнейшее сопоставление этих моделей на предмет выявления сходств и различий в организации супрануклеосомного хроматина между полученными в ходе работы и известными из литературы моделями.

В ходе работы было разработано программное обеспечение, позволяющее интегрировать данные Micro-C, содержащие информацию о пространственных контактах между всеми локусами генома на нуклеосомном (200 п.о) разрешении, данные MNase-seq, содержащие информацию о позиционировании нуклеосомных диад, и данные ChIP-seq, содержащие информацию о распределении по геному специфических модификаций гистоновых хвостов (в работе использовались данные о распределении метки гетерохроматина H3K9me3, метки промоторов H3K4me3 и метки энхансеров – H3K4me1), получая нуклеосомные контактные карты, аннотированные эпигенетическими метками. Разработанное в ходе работы программное обеспечение также позволяет переводить данные о частоте пространственных контактов между нуклеосомами и линкерными участками в данные о физических расстояних между ними и моделировать компактизацию хроматина с помощью методов грубозернистого моделирования, имплементированных в разработанный ранее нашей научной группой программный пакет. С помощью разработанного программного обеспечения были получены реконструкции организации супрануклеосомного хроматина в ядрах клеток hESC и HeLa H.sapiens и клеток дикого типа S.cerevisiae с учетом распределения эпигенетических меток H3K9me3, H3K4me3 и H3K4me1. Сопоставительный анализ распределения числа контактов нуклеосом с учетом их относительных позиций (N, N+1, N+2 и т.д) в полученных моделях показал, что дрожжевой хроматин и хроматин малигнизированных клеток человека (HeLa) на супрануклеосомном уровне релаксирован и слабо упорядочен, в то время как хроматин эмбриональных стволовых клеток человека более конденсирован, частично структурирован в регулярные петли и содержит глобулярные структуры, которые нами были идентифицированы как описанные ранее в литературе микроТАДы, так как их размеры соответствовали размерам описанных в литературе

микроТАДов и они колокализовались с пиками насыщенности метками энхансеров и промоторов, что согласуется с представлением о микроТАДах как о структурах, образующихся при сближении энхансеров и промторов в процессе выпетливания. При этом в хроматине клеток hESC не было детектировано локальных тетрануклеосомных двухстартовых спиралей, детектируемых, согласно литературе, в хроматине эмбриональных стволовых клеток M.musculus.

4.17 Исследование выполнено за счет гранта Российского фонда фундаментальных исследований № 20-34-70039

Authors: Sergey Isaev[1, 2], Rachelly Normand[3, 4, 5, 6], Peter Kharchenko[7]

[1]Research Institute of Personalized Medicine, National Center for Personalized Medicine of Endocrine Diseases, The National Medical Research Center for Endocrinology, Moscow, Russia, [2]Moscow Institute of Physics and Technology, Dolgoprudniy, Russia, [3]Center for Immunology and Inflammatory Diseases, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA, [4]Center for Cancer Research, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA, [5]Harvard Medical School, Boston, MA, USA, [6]Broad Institute of MIT and Harvard, Cambridge, MA, USA, [7]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

Single-cell multimodal omics allow measurements of different modalities (e.g. RNA and chromatin accessibility, or RNA and epitope abundance) for each cell at the same time [1]. It is unclear how to create some kind of embedding that would include information from each modality and would reflect the whole variety of cell states. Many methods have been developed to solve this problem [2, 3, 4]; however, due to the lack of ground truth (we do not know the actual cell identities), which approach for joint analysis of multimodal omics is the best remains a matter of debate.

We developed a metric for assessing the preservation of information in an embedding by comparing the neighbourhoods of cells in k-NN graphs of single modalities with their joint embedding.

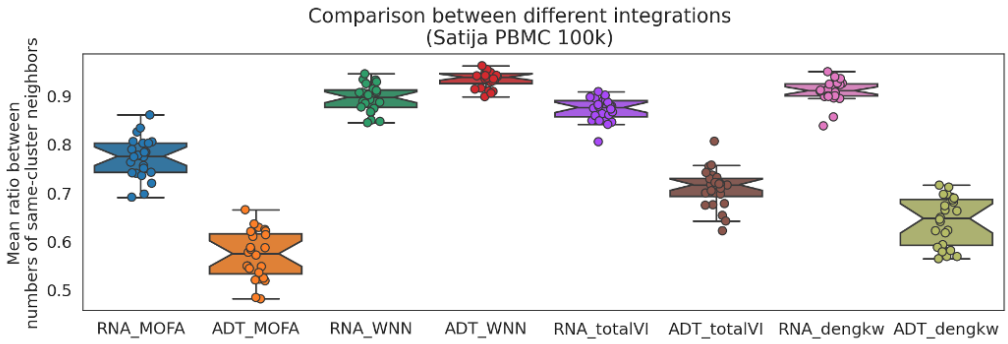

Figure 1. Comparison of preservation of cell neighborhood between different modalities and their integration.

Using this metric, various methods for integrating multimodal single-cell data were evaluated and the method that meets this criterion the best was determined to be WNN [2] (see Figure 1).

Acknowledgements

We thank Alexandra-Chloe Villani for useful discussions.

Sources and literature

1. Zhu C, Preissl S, Ren B. (2020) Single-cell multimodal omics: the power of many. Nat Methods 17:11-14.

2. Hao Y, Hao S, Andersen-Nissen E, et al. (2021) Integrated analysis of multimodal single-cell data. Cell 184:3573-3587.e29.

3. Argelaguet R, Arnol D, Bredikhin D, et al. (2020) MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol 21:111.

4. Gayoso A, Steier Z, Lopez R, et al. (2021) Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat Methods 18:272-282.

4.18 Эволюция Т-боксов

Авторы: Е.И. Григорашвили[1], М.С. Гельфанд[1,2]

1 - Сколковский институт науки и технологий, 2 - Институт проблем передачи информации им. Харкевича

У грамположительных бактерий экспрессия генов биосинтеза и транспорта аминокислот на уровне транскрипции часто регулируется рибопереключателями, получившими название “Т-боксы”. Т-бокс специфично связывается с незаряженной тРНК и формирует антитерминатор, стимулируя транскрипцию генов синтеза соответствующей аминокислоты (у некоторых Actinobacteria - инициацию трансляции). Обычно Т-бокс состоит

из двух консервативных РНК-шпилек: одна определяет специфичность Т-бокса, а вторая формирует антитерминатор. Т-боксы могут быть двойными или “полуторными”. Двойной Т-бокс представляет собой два тандемно расположенных Т-бокса. В полуторных Т-боксах за основной парой шпилек следует еще одна шпилька-антитерминатор.

Предполагается, что Т-боксы в основном эволюционируют путем дупликаций с последующей потерей одной из копий, сочетающейся с коэволюцией с регулируемыми генами. Это находит отражение в существовании двойных и полуторных Т-боксов. Благодаря дупликациям, Т-боксы способны менять свою специфичность. Показано, что Т-боксы способны вытеснять другие механизмов регуляции биосинтеза аминокислот (Vitreschak, 2008).

Описанные механизмы в основном были получены на сравнительно небольших данных (805 Т-боксов из 96 геномов). Для многих примеров замещения Т-боксами другой системы не ясна динамика процесса; не исследовано влияние горизонтального переноса генов и эпистатических взаимодействий на эволюцию Т-боксов; неизвестен механизм адаптации Т-боксов к модификациям в тРНК и формирования таксон-специфических вариаций структур Т-боксов; механизм действия двойных и полуторных Т-боксов также до сих пор не вполне ясен (Zhang, 2020).

Целью нашего исследования является заполнить вышеописанные пробелы, используя большее количество данных. На сегодняшний день в базе данных TBDB (Marchand, 2021) содержится более 20000 стандартных одинарных Т-боксов из 3632 бактериальных штаммов. Мы планируем дополнить эти данные двойными и полуторными Т-боксами и при помощи филогенетического анализа исследовать горизонтальный перенос генов, эпистатические взаимодействия, дупликации и смену специфичности в Т-боксах бактерий.

Источники и литература:

1. Marchand JA, Pierson Smela MD, Jordan THH, Narasimhan K, Church GM. TBDB: a database of structurally annotated T-box riboswitch:tRNA pairs. Nucleic Acids Res. 2021;49(D1):D229-D235. doi:10.1093/nar/gkaa721

2. Vitreschak AG, Mironov AA, Lyubetsky VA, Gelfand MS. Comparative genomic analysis of T-box regulatory systems in bacteria. RNA. 2008;14(4):717-735. doi:10.1261/rna.819308

3. Zhang J. Unboxing the T-box riboswitches-A glimpse into multivalent and multimodal RNA-RNA interactions. Wiley Interdiscip Rev RNA. 2020;11(6):e1600. doi:10.1002/wrna.1600

5. Мутации

5.1 Предсказание внутригенной компенсации функциональных несинонимичных замен

Авторы: Надежда Азбукина[1,2], Анастасия Жарикова[1,2], Александр Гресс[3], Ольга Калинина[3], Василий Раменски[1,2]

1 - МГУ им Ломоносова, Факультет бионженерии и биоинформатики, 2 - Национальный медицинский исследовательский центр терапии и профилактической медицины, 3 - Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI), Saarbrücken, Германия

Экспоненциальный рост накопленных генетических данных делает все более актуальной задачу их точной интерпретации с точки зрения клинической практики. Особенно остро стоит задача предсказания функционального эффекта аминокислотных замен у отдельных индивидуумов. Современные подходы, основанные на методах статистической генетики, при своем предсказании основываются на анализе спектра аминокислот, наблюдаемых на данной позиции в гомологичных белках. Однако порядка 10 % болезнетворных у человека мутаций наблюдается в норме у гомологов близких видов. Такой эффект объясняется тем, что в гомологичных белках замены существуют в другом контексте, который влияет на приспособленность. Этот феномен носит название эпистаза - нелинейности суммы эффектов отдельных мутаций.

В данной работе мы сфокусировались на определенном типе эпистаза – внутригенной компенсации, случае, когда одна из замен нарушает функцию белка, а другая – «спасает» этот эффект. Для этого было собрано порядка 670 тысяч пар несинонимичных аминокислотных замен, для каждой пары известен функциональные эффект мутаций по-отдельности и вместе. В качестве источников информации были использованы данные по глубокому мутационному сканированию (база maveDB), база PMD, составленная по статьям, и литературные данные.

Далее собранные мутации были охарактеризованы с использованием структурных и филогенетических признаков. Для картирования мутаций на структуру был использован метод StructMAn, который который для исходного белка находит гомологичные белки с известной пространственной структурой и затем картирует каждый остаток на наиболее репрезентативную структуру. Также StructMAn позволяет определить структурную локлаизацию остатка: гидрофобное ядро, интерфейсы взаимодействий с белками и низкомолекулярными лигандами, свободная поверхность. На основе полученных структур вычислялось изменение стабильности белков методом FoldX. Для получения филогенетических признаков использовались глубокие выравнивания базы PFAM. Для каждой мутации вычислялась чистота встречаемости исходной и мутированной аминокислот

в выравнивании, а также консервативность позиции. Для пар мутации вычислялась взаимная информация позиций, которая использовалась для оценки коэволюции аминокислотных остатков.

Было получено, что компенсаторы на белок-белковых интерфейсах имеют тенденцию находиться ближе к компенсируемой замене, чем компенсаторы, находящиеся в ядре (p-value <0.05). Также показано, что универсальные компенсаторы (компенсаторы, которые могут компенсировать более одной замены) являются более стабилизирующими, чем не универсальные (p-value <0.05), и не локальны. В данный момент ведется работа над разработка модели машинного обучения для предсказания вероятности замены быть компенсатором. Исследование выполнено при финансовой поддержке РФФИ и DFG в рамках научного проекта No 20-54-12008.

5.2 Биоинформатический анализ соматических мутаций в контексте образования G-квадруплексов в промоторе гена TERT и в других онкогенах

Авторы: В.В. Панова[1], К.А. Новоселов[1], Е.А. Кубарева[1], М.Э. Зверева[2], А.В. Алексеевский[1,3].

1 - Факультет биоинженерии и биоинформатики, МГУ имени М.В.Ломоносова, Москва, Россия, 2 - Химический факультет, МГУ имени М.В. Ломоносова, Москва Россия, 3 - НИИ Физико-химической биологии им. А.Н.Белозерского МГУ, МГУ имени М.В.Ломоносова, Москва, Россия, 4 - Научно-исследовательский институт системных исследований РАН, Москва, Россия

Активность гена теломеразы способствует развитию процесса онкогенеза [1].Мутации в промоторной области гена обратной транскриптазы теломеразы (TERT)наблюдаются при различных видах рака (Li et al., J. Biol. Chem., 2017). Промотор hTERT имеет GC богатую последовательность и в нем детектируются G-квадруплексы (G4) [2]. Полагают, что G-квадруплексы регулируют транскрипцию генов [3], что подтверждается полногеномными биоинформатическими исследованиями: в промоторных участках генов человека G4-мотивами обогащены участки в 200 п.н. перед стартами транскрипции. В нашей группе было экспериментально показано, что белки MutS и MutL из системы MMR (mismatch repair system) образуют прочные комплексы с G4, которые могут снизить эффективность репарации ДНК [4]. G4 структуры могут приводить к повреждению ДНК и блокировать механизмы репарации [5]. Мы предполагаем, что стабилизация G4 может способствовать возникновению «горячих точек» мутаций.

Целью данной работы является определение положения последовательностей G4 в

промоторах онкогенов и анализ мутаций в этих последовательностях. Первым анализируемым геном является TERT. Ищутся G4 последовательности в промоторах TERT млекопитающих и мутации в них.

В базе данных мРНК в NCBI найдено гены теломеразы TERT 157 млекопитающих. Для каждого организма из genbank скачаны фрагменты ДНК, включающие максимальную мРНК и участок в 1000 п.н. предшествующий старту транскрипции. Этот участок включает промоторную область гена. Старты транскрипции определены в соответствии с аннотацией мРНК. Старты трансляции определены в соответствии с аннотацией CDS. Проверка положения стартов трансляции проводилась путем множественного выравнивания аминокислотных последовательностей самых длинных изоформ белка TERT различных видов млекопитающих. . Для нескольких последовательностей, в которых имеются протяженные участки плохого выравнивания или крупные делеции в области двух доменов, и ещё нескольких последовательностей нетипичной для TERT длины (менее 900 аминокислотных остатков) планируется перепроверить правильность аннотации интрон-экзонной структуры в записях из Genbank. Были выкачаны последовательности промоторов по координатам начала транскрипции-1000 и начала транкрипции, и в этих последовательностях с помощью трех программ G4Hunter, QGRS mapper и pqsfinder (пакет R) найдены последовательности, образующие G-квадруплексы. Проводится анализ с помощью NPG-explorer квадруплексов у 27 видов приматов, чтобы выяснить, какая программа наиболее корректно определяет G-квадруплексы.

В ближайших планах на основе сравнения результатов разных алгоритмов поиска G-квадруплексов, определить или разработать оптимальный алгоритм их поиска,а также проверить и, при необходимости, исправить аннотации экзон-интронной структуры генов с недостоверным выравниванием N-конца их продуктов.

Работа поддержана Российским научным фондом (проект № 21-14-00161).

Источники и литература:

1. Yuan et al. Mechanisms underlying the activation of TERT transcription and telomerase activity in human cancer: old actors and new players. Oncogene, 2019, doi: 10.1038/s41388-019-0872-9

2. Bochman et al. DNA secondary structures: stability and function of G-quadruplex structures. Nat. Rev. Gen., 2012, doi: 10.1038/nrg3296

3. David et al. G-quadruplexes as novel cis-elements controlling transcription during embryonic development. Nucleic Acids Res., 2016, doi: 10.1093/nar/gkw011

4. Pavlova et al.Responses of DNA Mismatch Repair Proteins to a Stable G-Quadruplex Embedded into a DNA Duplex Structure. Int. J. Mol. Sci., 2020, doi: 10.3390/ijms21228773

5. Linke et al. The Relevance of G-Quadruplexes for DNA Repair. Int. J. Mol. Sci. 2021, doi: 10.3390/ijms222212599

5.3 APOBEC – индуцированный мутагенез снижен в большинстве областей ДНК, имеющих неканоническую структуру

Авторы: Геннадий Пономарев[1], Булат Фатыхов[2], Владимир Назаров[2], Руслан Абасов[3], Евгений Шваров[4], Нина-Вики Ландик[5], Александра Денисова[5], Альмира Червова[6], Михаил Гельфанд[1,7], Марат Казанов[1,7]

1 - Институт Проблем Передачи Информации им А.А.Харкевича, Москва, Россия, 2 - Московский Физико-Технический Институт, Москва, Россия, 3 - ФГБУ НМИЦ ДГОИ им. Дмитрия Рогачева Минздрава России, Москва, Россия, 4 - InterSystems Corporation, Cambridge MA, USA, 5 - Московский Государственный Университет им М.В.Ломоносова, Москва, Россия, 6 - Institut Pasteur, Paris, France, 7 - Сколковский Институт Науки и Технологий, Москва, Россия

Известно, что области генома человека, имеющие неканоническую структуру ДНК, имеют повышенную плотность соматических мутаций, однако вклад отдельных мутагенов остается неизученным. Мы обнаружили, что в геномах злокачественных опухолей человека мутагенез области прямых повторов, зеркальных повторов, коротких тандемных повторов и G-квадруплексов не обогащены APOBEC-индуцированными мутациями, и их количество даже снижено по сравнению с участками B-ДНК в образцах с высокой активностью APOBEC-индуцированного мутагенеза. В противоположность данному эффекту мы обнаружили повышенную плотность APOBEC-индуцированных мутаций в инвертированных повторах, и выяснили, что основной вклад в данное повышение вносят мутации в цитозинах на 3’-конце петель шпилек инвертированных повторов. В участках однонитевой ДНК G-квадруплекса нами была обнаружена пониженная плотность APOBEC-индуцированных мутаций, по сравнению с плотностью мутаций в гуанин-богатой части G-квадруплекса. В областях Z-ДНК мы обнаружили практически полное отсутствие APOBEC-мутагенеза, как и мутаций, вызванных ультрафиолетовым излучением, по причине отсутствия ди-пиримидиновых мишеней.

5.4 Анализ геномного контекста APOBEC-индуцированных мутаций в геномах злокачественных опухолей человека

Авторы: Александра Денисова[1], Марат Казанов[2,3]

1 – МГУ им М.В.Ломоносова, Москва, Россия, 2 - Институт Проблем Передачи Информации им А.А.Харкевича, Москва, Россия, 3 - Сколковский Институт Науки и Технологий, Москва, Россия

Цитидин дезаминазы семейства APOBEC, являясь компонентом врожденной иммунной системы человека, играют важную роль в защите организма от вирусов и мобильных генетических элементов. Дезаминирование цитозина на однонитевой ДНК ферментами APOBEC в контексте мотива TC может приводить к его замене на тимин или гуанин. Такие паттерны замен были недавно обнаружены в геномах некоторых типов злокачественных опухолей, что указывает на мутагенное воздействие цитидин дезаминаз семейства APOBEC. Анализ распределения мутаций вдоль генома с использованием мутационной подписи TC показывает необычное распределение APOBEC-индуцированных мутаций, а именно большую плотность мутаций в активно транскрибируемой части генома. Механизм, связанный с данным распределением, неизвестен, хотя предыдущие исследования установили некоторую корреляцию позиций мутаций с расположением транспозонов. В данном исследовании мы сократили набор мутационных данных отобрав для анализа только образцы с высокой активностью APOBEC-индуцированного мутагенеза. В таком случае, мутации в рассматриваемых образцах в мотиве TC с гораздо большей вероятностью индуцированы APOBEC-мутагенезом. Такой подход позволит нам более точно изучить геномных контекст APOBEC-индуцированных мутаций, рассмотрев корреляцию позиций мутаций с позициями различных элементов генома.

5.5 Сигнал положительного отбора при реактивации B клеток памяти свидетельствует о новых циклах созревания аффинности

Авторы: Артем Микелов[1,2,3], Евгения Алексеева[1], Екатерина Комеч[2,3], Дмитрий Староверов[2], Мария Турчанинова[2], Михаил Шугай[2,3], Дмитрий Чудаков[1,2,3], Георгий Базыкин[1,4], Иван Звягин[2,3]

1 - Сколковский Институт Наук и Технологий, Москва, 2 - Институт биоорганической химии им. академиков М.М. Шемякина и Ю.А. Овчинникова, 3 - РНИМУ им. Н.И. Пирогова, 4 - Институт Проблем Передачи Информации им. А.А. Харкевича, Москва

Для производства высоко-аффинных антител B клеточные клональные линии проходят через эволюционный процесс, основанный на циклах соматических гипермутаций и естественного отбора. Расширение клональных линий сопровождается дифференцировкой B клеток в антитело-производящие плазматические клетки и переключением изотипа иммуноглобулина. Распределение клеточных типов и изотипов в клональной линии отражает ее роль в иммунном ответе, однако эволюционные механизмы, стоящие за ее развитием, изучены мало. В данной работе мы изучили динамику иммунных репертуаров трех B-клеточных фракций: B клеток памяти, плазмабластов и плазматических клеток. Образцы

репертуаров были получены трижды на протяжении года от пятерых здоровых добровольцев. Среди наиболее крупных линий В клеток мы обнаружили два кластера с разными свойствами. Первый кластер представлял собой линии персистирующей памяти с преобладанием IgM изотипа. Второй кластер состоял из линий, преимущественно детектируемых в одной временной точке и сформированных антитело производящими типами с переключенным Ig G или Ig A изотипом. Кроме этого линии второго кластера произрастали из предка с большим числом соматических гипермутаций и зачастую имели В клетки памяти до временной точки их наибольшей представленности, таким образом демонстрируя признаки реактивации и расширения В клеточной памяти. Кроме состава два кластера В клеточных линий отличались по типу действующего в них естественного отбора. В эволюции персистирующая память преобладал отрицательный отбор, необходимый для поддержания структуры В клеточного рецептора, сформированной при созревании аффинности во время соответствующей инфекции. Напротив антитело-производящие линии имели отпечаток положительного отбора, свидетельствующий о новых циклах созревания аффинности при реактивации В клеточной памяти. Данные результаты демонстрируют адаптивность В клеточной памяти при повторных встречах с антигеном.

Благодарности. Мы благодарны нашим волонтерам за участие в исследовании. Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов 20-34-90153 (Е.А.) и Министерства Науки и Высшего Образования РФ в рамках проекта 075-15-2019-1789 (Д.Ч.).

5.6 Эволюция SARS-CoV2 ведет к избеганию Т клеточного ответа при долговременной инфекции на фоне иммуносупрессивной терапии

Авторы: Оксана Станевич[1,2], Евгения Алексеева[3], Артем Фадеев[1], Ксения Комиссарова[1], Анна Иванова[1], Тамара Симакова[6], Мария Сергеева[1], Кирилл Васильев[1], Анна-Полина Шурыгина[1], Марина Стукова[1], Ксения Сафина[3], Елена Набиева[3], Софья Гарушянц[4], Галка Клинк[4], Евгений Бакин[1,7], Юлия Забутова[5], Анастасия Холодная[2,5], Ольга Лукина[2], Ирина Скороход[5], Виктория Рябчикова[5], Надежда Медведева[5], Дмитрий Лиознов[1,2], Дарья Даниленко[1], Дмитрий Чудаков[3,8], Андрей Комиссаров[1], Георгий Базыкин[3,4*]

1 - НИИ гриппа им. А.А. Смородинцева, Санкт Петербург, 2 - Первый Санкт-Петербургский государственный медицинский университет им. И.П. Павлова, Санкт-Петербург, 3 - Сколковский Институт Наук и Технологий, Москва, 4 - Парсек Лаб, Санкт-Петербург, 5 - Институт Проблем Передачи Информации им. А.А. Харкевича, Москва, 6 - Институт Биоинформатика, Санкт-Петербург, 7- Городская больница №31, Санкт Петербург,

8 - РНИМУ им. Н.И. Пирогова

Долговременная эволюция SARS-CoV-2 в условиях подавленной иммунной системы может играть роль в возникновении новых вариантов коронавируса. Целый ряд работ продемонстрировал избегание вирусом гуморального иммунитета при эволюции внутри одного хозяина, однако данных об избеганием вируса Т клеточного иммунитета крайне мало. В данной работе мы сообщаем о случае длительного COVID-19 в иммуноподавленном пациенте с неходжкинской лимфомой на фоне терапии ритуксимабом и отсутствие нейтрализующих антител. За 318 дней болезни в геноме SARS-CoV-2 произошло 40 изменений, 34 из которых присутствовали к концу исследования. Среди приобретенных мутаций 12 уменьшали или предотвращали связывание известных иммуногенных CD8 эпитопов, свидетельствуя в пользу избегания цитотоксического Т-клеточного ответа. Две мутации с самым сильным эффектом, nsp3:T504A и nsp3:T504P, были экспериментально оценены в цитотоксическом анализе CD8 Т-клеток пациента. Обе мутации снижали реакцию CD8 Т клеток пациента, однако эффект nsp3:T504P был намного сильнее, что согласуется с там, что этот вариант и был зафиксирован в вирусной популяции. Данные наблюдения позволяют полагать, что ускользание от цитотоксического иммунного ответа может быть недооцененным фактором эволюции SARS-CoV-2 в человеческой популяции.

Благодарности. Мы благодарны пациенту за участие в исследовании. Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов 20-04-60556 (Г.Б.), 20-34-90153 (Е.А.) и Министерства Науки и Высшего Образования РФ в рамках проекта 075-15-2019-1789 (Д.Ч.).

5.7 Extensive analysis of epistatic coefficients from combinatorially complete datasets

Authors: María Carolina Erazo Muñoz, Dmitry Ivankov Skolkovo Institute of Science and Technology, Moscow, Russian Federation

Classical definitions of epistasis focus on interactions between different genes, and the biological implications of this. Statistical epistasis, as presented by Fisher, goes deeply into possible relations between genes from a mathematical point of view. The current understanding of epistasis also includes interactions between protein/DNA/RNA positions.

There are various methods to detect epistasis. One of them is to compute epistatic coefficients from combinatorially complete datasets, the datasets forming hypercubes in sequence space (1). From this point of view, the following methods can be applied:

Most commonly, researchers apply the Hadamard-Walsh transformation to obtain epistatic coefficients; however, "thermodynamic" transformations can also be applied (2). The interactions between positions are modelled by “house of cards” or NK model (3, 4).

Using this knowledge, we aim at an extensive analysis of epistatic coefficients of hypercubes obtained from random mutagenesis experiments. The objectives are: (i) to verify if thermodynamic and ensemble method results are statistically indistinguishable, (ii) if the epistatic picture corresponds to a House of cards model, and (iii) if the coefficients obtained by these distinct methods can be comparable to a K value in a modified NK model.

For this, all possible hypercubes were computed in 37 datasets obtained from random mutagenesis, using the HypercubeME_recursive.py algorithm (5). Afterwards, an algorithm was developed to compute the epistatic coefficients for these hypercubes, using both thermodynamic and Hadamard-Walsh approach, in the real landscape picture and the House of cards model compatible image. Correlation coefficients were computed between the two methods, whilst the development of an NK model adaptable to existing fitness values and its testing is in process.

It was found that ensemble and thermodynamic methods are not strongly correlated, as correlation and dimension have an inverse relationship from dimension 2 and on. Getting farther from one as dimension increases, not reaching values close to -1 either. The epistatic picture in the studied datasets does not correspond to a house of cards model.

Plans for testing different K values in the NK model representation and detection of presence or absence of sign epistasis in third and higher dimensional hypercubes. Overall, we show that there can be significant differences in epistatic coefficients calculated by Hadamard-Walsh and the "thermodynamic" form, both of which require attention.

Sources and literature:

1. Weinreich DM et al. (2013) Should evolutionary geneticists worry about higher-order epistasis? Curr. Opin. Gen. Dev., 23:700–707.

2. Poelwijk FJ, Krishna V, Ranganathan R (2016) The context-dependence of mutations: a linkage of formalisms. PLoS Comput. Biol., 12:e1004771.

3. Kryazhimskiy S et.al (2009) The dynamics of adaptation on correlated fitness landscapes. PNAS, 106:18638-18643.

4. Kauffman, S.; Levin, S. (1987). "Towards a general theory of adaptive walks on rugged landscapes". Journal of Theoretical Biology. 128 (1): 11–45. doi:10.1016/s0022-5193(87)80029-2

5. Esteban LA et al. (2020) HypercubeME: two hundred million combinatorially complete datasets from a single experiment. Bioinformatics, 36:1960–1962.

5.8 Исследование эпистаза с использованием композитных мутаций

Авторы: Евгений Зорин, Дмитрий Иванков Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow

Сложная взаимосвязь генотипа и фенотипа является давней проблемой биологии, важной как с фундаментальной, так и с прикладной точек зрения. Одним из основных факторов, препятствующих прямолинейному связыванию генотипа и фенотипа, является эпистаз - отклонение эффекта множественных мутаций от суммы их индивидуальных эффектов. Несмотря на накопленные генетические данные и прогресс в этой области, эпистаз все еще недостаточно изучен. Эпистаз можно оценить в различных генетических контекстах в комбинаторно полных структурах генотипов - гиперкубах - однако этого может быть недостаточно для отображения полной картины эпистатических взаимодействий. Потенциальным решением может быть изучение эффектов составных (композитных) мутаций в структурах, называемых гиперпрямоугольниками, где два генотипа могут быть соединены вне зависимости от количества мутаций, которые их соединяют. Данный доклад сообщает о разработке нового алгоритма для расчета гиперпрямоугольников и сравнения количества эпистаза, идентифицированного в гиперкубах и гиперпрямоугольниках, построенных из частей экспериментальных наборов данных.

Первой частью проекта было успешное написание нового алгоритма Hypercuboid ME для вычисления всех возможных составных мутационных путей от одной последовательности к другой в гиперкубе. Впоследствии, были построены гиперкубы и гиперпрямоугольники с использованием программ HypercuboidME существующий алгоритм) и Hypercuboid ME. Эпистатические коэффициенты a 1 2 были рассчитаны для каждой структуры путем применения треугольной матрицы к фитнес значениям каждой гиперструктуры в измерении N = 2. Для анализа были использованы части 1500 генотипов из баз данных, включающих сегменты 1 - 12 гена HIS 3 (белок IGPD) и WW. Из них были вычислены эпистатические коэффициенты a 12, отражающие парный эпистаз, и их распределения были сопоставлены между гиперкубами и гиперпрямоугольниками с помощью U-критерия Манна-Уитни (Mann – Whitney U test). Результаты показывают, что для всех баз данных, кроме WW-доменане, достаточно доказательств, чтобы отвергнуть нулевую гипотезу об одинаковости двух распределений. Для базы данных WW значение p является статистически значимым ($p< 0.01$), однако количество гиперструктур, использованных для анализа, на два порядка меньше, чем в других образцах; также, разница между двумя распределениями не кажется практически значимой. Интересный результат был обнаружен для исследуемых подмножеств сегментов 2,6, и 12 белка IGPD, где гиперкубы не были обнаружены, однако для гиперпрямоугольников был оценен эпистаз, что указывает на ценность использования гиперпрямоугольников для анализа комбинаторно неполных наборов данных.

Полученные данные свидетельствуют об отсутствии различий в величине эпистаза, выявляемого гиперкубами и гипер прямоугольниками для рассматриваемых баз данных. Однако использование гипер прямоугольников вместо гиперкубов все еще может быть полезно в комбинаторно неполных наборах данных, где может быть недостаточно генотипов для составления гиперкубов. В будущем может быть разработан более быстрый алгоритм для рассмотрения больших размеров выборки и, таким образом, для лучшей проверки исходной гипотезы.

5.9 Определение эпистатических взаимодействий в генахMycobacterium tuberculosis, ассоциированных с лекарственной устойчивостью

Авторы: Дарья Быкова[1,2], Геннадий Федонин[1,3,4], Алексей Неверов

1 - Центральный научно-исследовательский институт эпидемиологии Роспотребнадзора, Москва, Россия, 2 – МГУ им М.В. Ломоносова, Москва, Россия 3 - Московский физико-технический институт, 4 - Институт проблем передачи информации им. А.А. Харкевича РАН

Ежегодно примерно пол миллиона людей по всему миру заражаются штаммами Mycobacterium tuberculosis (MTB) с лекарственной устойчивостью [1]. В отличие от множества других бактериальных патогенов, возбудитель туберкулёза приобретает устойчивость к антибактериальным препаратам не за счёт горизонтального переноса генов, а вследствие накопления мутаций в геноме. Многие мутации, приводящие к лекарственной устойчивости, в условиях отсутствия антибиотика являются вредными, то есть снижают приспособленность бактерии [2]. В результате, возникновение мутаций устойчивости должно способствовать отбору компенсаторных мутаций, однако подобные компенсаторныемеханизмыдляMTBизученынедостаточно.Данный проект включает в себя две подзадачи. Во-первых, определить локусы в геноме MTB, приводящие к лекарственной устойчивости. Во-вторых, обнаружить эпистатические взаимодействия между сайтами, находящимися под отбором в резистентных линиях. В работе анализируются данные полногеномного секвенирования более 11 500 штаммов MTB, в анализ включены данные об устойчивости к 14 противотуберкулёзным препаратам. Для поиска мутаций, приводящихклекарственной устойчивости, используется метод, описанный ранее в работе [3]. Метод основан на восстановлении филогенетического дерева поимеющимся геномам, а также предковых генотипов и фенотипов устойчивости, и подсчёте числа событий, когда появление мутации в геноме сопровождается изменением фенотипа (т.е. приобретением или потерей резистентности). Для борьбы с ложными ассоциациями, вызванными совместным

применением нескольких препаратов, в работе учитываются корреляции между фенотипами устойчивости к различным препаратам. В дальнейшемпланируется проанализировать обнаруженные ассоциации с учётом поправок и, возможно, выявить новые кандидатные гены, вовлечённые в процесс развития резистентности. В ходе дальнейшей работы планируется выявить сайты, мутации в которых происходят в резистентных линиях значительно чаще по сравнению с чувствительными, и провести поиск эпистатических взаимодействий между парами таких сайтов.

Источники и литература:

1. Tackling the drug-resistant TB crisis. [cited 31 Jan 2022]. Available: https://www.who.int/activities/tackling-the-drug-resistant-tb-crisis

2. Vogwill T, MacLean RC. The genetic basis of the fitness costs of antimicrobial resistance: a meta-analysis approach. Evol Appl. 2015;8: 284–295.

3. Neverov AD, Popova AV , Fedonin GG, Cheremukhin EA, Klink GV , Bazykin GA. Episodic evolution of coadapted sets of amino acid sites in mitochondrial proteins. PLoS Genet. 2021;17: e1008711.

5.10 Signatures of selection against stop codon readthrough in populations of D. melanogaster and H. sapiens.

Authours: Vladimir Shikov[1], Georgii A. Bazykin[1,2], Olga Vakhrusheva[1]

1 - Skolkovo Institute of Science and Technology, 2 - Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute)

Termination of translation is an imperfect process, as the ribosome may sometimes add an amino acid at the stop codon instead of terminating the translation and continue translating past the stop codon in the 3’UTR, a phenomenon known as stop-codon readthrough. The three stop codons differ in the probabilities of spontaneous readthrough. UGA is considered the most readthrough-prone stop codon, followed in terms of readthrough rate by UAG and finally by UAA. In addition, the downstream nucleotides are known to influence stop codon readthrough rate, with UGA-C being the most error-prone combination.

Although the stop codon readthrough has been reported to play a role of a regulatory mechanism possibly contributing to the proteome diversity, most readthrough events appear to stem from molecular errors and to have deleterious effects arising from the synthesis and accumulation of aberrant proteins. As aberrant readthrough events are likely to negatively impact fitness, mutations resulting in leaky termination signals are expected to be under negative selection. Indeed, recent

studies have demonstrated that stop codon contexts associated with elevated readthrough rates are underrepresented in eukaryotic genomes and that highly expressed genes are depleted for readthrough-prone stop codon contexts both in yeast and fruit fly. Moreover, negative selection acting on stop codon usage has been inferred in both prokaryotes and eukaryotes from the data on interspecies divergence. However, it is unclear whether there is ongoing selection against inefficient translation termination in natural populations.

Here, we look for the signatures of ongoing negative selection against mutations resulting in less accurate termination contexts. We have already tested for negative selection on polymorphic variants resulting in switches between different stop codons (UAA <-> UGA, UAA <-> UAG) based on the allele frequency spectra in populations of Drosophila melanogaster and Homo sapiens. Our preliminary results suggest there is ongoing selection restricting transitions between the more and the less "accurate" stop-codons in both directions.

Next, we are planning to assess the strength of negative selection against mutations creating the least accurate stop codon context (UGA-C), and explore whether the degree of selective constraint on stop-codon usage is associated with different gene features such as gene expression levels and evolutionary conservation.

Finally, we will examine the potential effects of GC-biased gene conversion on the efficacy of selection against leaky stop codon contexts. GC-biased gene conversion favors GC-enriched alleles and is expected to tip the balance towards UGA and UAG codons and towards the most error-prone stop codon context UGA-C. However, no studies were conducted on the extent to which GC-biased conversion affects stop codon usage and usage of translation termination signals in general. To determine whether GC-biased gene conversion is shaping stop codon usage, one can use genomes with different rates of GC-biased gene conversion, as well as genomic regions varying in recombination frequency and/or GC content. Here, we will explore whether GC-biased gene conversion could affect efficacy of selection on translation termination signals making use of both genomes exhibiting different levels of GC-biased gene conversion (H. sapiens and D. melanogaster) and genomic regions exhibiting different recombination rates.

Optimal time – 5 minutes.

5.11 Эволюция сайтов сдвига рамки считывания в инфузории рода Euplotes

Authors: Sofya Gaydukova[1], Adriana Vallesi[2], Stephen M. Heaphy[3], John F Atkins[34], Pavel V. Baranov[2], Mikhail S. Gelfand[56], Mikhail Moldovan[5]

1 - Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia, 2 - Laboratory of Eukaryotic Microbiology and Animal Biology, School of Biosciences and Veterinary Medicine, University of Camerino, Camerino 62032, Italy, 3 - Schools

of Biochemistry and Microbiology, University College Cork, Cork, T12 XF62 Ireland, 4 - Department of Human Genetics, University of Utah, Salt Lake City, UT 84112, 5 - Skolkovo Institute of Science and Technology, Moscow, Russia, 6 - A. A. Kharkevich Institute for Information Transmission Problems RAS, Moscow, Russia.

Аминокислотные последовательности белков закодированы в нуклеотидах мРНК и декодируются с помощью генетического кода. Известно несколько различных вариантов генетического кода, однако в каждом из них любая аминокислота или стоп-кодон кодируется тремя нуклеотидами (триплетом). Исключение составляют инфузории рода Euplotes, где стоп-кодоны в консервативном коротком контексте в середине транскрипта определяют сдвиг рамки считывания, то есть ситуацию, когда рибосома пропускает несколько нуклеотидов мРНК. В Euplotes сдвиги рамки считывания происходят на +1 или +2 нуклеотида и имеют широкое распространение по транскриптому: сайты сдвига содержат, по разным оценкам, от 5 до 20% белок-кодирующих генов. Таким образом, генетический код рода Euplotes содержит семейства квадриплетов и квинтиплетов.

В данном исследовании мы изучили эволюцию сайтов сдвигов в Euplotes и предложили схему, объясняющую возникновение столь необычного свойства генетического кода. Было отсеквенировано девять траснкриптомов девяи представителей рода Euplotes. Далее, мы разработали процедуру, позволяющую предсказывать сайты сдвига рамки считывания с высокой точностью по транскриптомным данным. Мутации, приводящие к возникновению сайта сдвига в контексте AAATAR оказались слабовредными. При этом против сайтов в контекстах, отличных от AAATAR, действует сильный отбор. Также было показано, что, несмотря на слабовредность, сайты сдвига рамки в изученной группе накапливаются в настоящий момент и будут продолжать накапливаться в дальнейшем. Исходя из наших результатов, мы предложили модель изменения генетического кода через нейтральные, но в то же время необратимые, изменения в кодирующих последовательностях.

5.12 Поиск точки смены однопозиционного адаптивного ландшафта на филогенетическом дереве с помощью машинного обучения

Авторы: Галя Клинк[1], Георгий Базыкин[1,2]

1 - ИППИ РАН, 2 - Сколковский Институт Науки и Технологий, Москва

Существование и распространённость явления непостоянства однопозиционного адаптивного ландшафта (изменения приспособленности аминокислот в позициях белков) показаны многими научными группами на разных данных и с использованием различных методов. Однако, не существует универсального метода, способного определить для каждого сайта белка, когда в эволюции происходили изменения в его адаптивном ландшафте. Наличие

такого метода будет полезно как для фундаментальных исследований, так и для определения эффекта мутаций в разных видах. Мы предлагаем совместно использовать филогенетический анализ и машинное обучение для поиска событий смены адаптивного ландшафта. Метод, который мы разрабатываем, работает по следующей схеме: для каждого аминокислотного сайта белка на основе выравнивания и филогенетического дерева вычисляются признаки, по которым с помощью алгоритмов «Случайный лес» или «Градиентный бустинг» сайты классифицируются на «постоянные», где не происходила смена адаптивного ландшафта, и «переменные». Затем, по такому же или модифицированному набору признаков с помощью тех же алгоритмов проводится регрессия, в результате которой для каждого «переменного» сайта предсказывается величина Y. Пока мы предполагаем, что для каждого сайта событие изменения адаптивного ландшафта может наступить максимум один раз, и величина Y характеризует узел филогенетического дерева, на котором оно произошло. Y определяется как соотношение размеров меньшего и большего поддеревьев, на которые узел разбивает дерево. Последний этап – поиск узла филогенетического дерева по предсказанной величине Y. Для тренировки и валидации метода мы используем данные, полученные при симуляции эволюции аминокислотных последовательностей вдоль филогенетического дерева. Поскольку наша цель – создание универсального метода, мы используем два больших дерева с разной структурой: дерево митохондриальных белков животных (Klink et al., 2017; «митохондриальные симуляции») и дерево для пяти подтипов ВИЧ-1 (Klink et al., 2022; «вирусные симуляции»). На данный момент для этапа классификации нам удалось достичь чувствительности 90% при специфичности 90% для классификации сайтов в вирусных симуляциях методом, тренировавшимся на митохондриальных симуляциях.

5.13 Распределение совпадающих однонуклеотидных полиморфизмов (SNP) в кодирующих последовательностях человека и макаки резус

Авторы: Е.Правдолюбова, Г. Базыкин Центр Наук о Жизни, Сколковский институт науки и технологий, Сколково, Россия

Известно, что у родственных видов число, совпадающих SNP в гомологичных позициях по всему геному в несколько раз выше ожидаемого (показано на парах видов дрозофил и для человека и шимпанзе [1,2]). Этот избыток связывают с гетерогенностью скорости мутации. Такие SNP должны создавать обширный материал для последующих параллельных замен.

За последние годы количество данных о генетическом разнообразии выросло настолько, что можно сравнить распределение совпадающих SNP между синонимичными и несинонимичными сайтами в пределах кодирующих последовательностей и таким образом

оценить относительный вклад гетерогенности скоростей мутации и отбора. Для работы мы взяли данные по SNP 2504 человек [3] и 833 макак [4].

Для анализа мы взяли четырехкратно вырожденные (синонимичные) и невырожденные (несинонимичные) гомологичные позиции в кодонах, которые не содержат замен относительно реконструированной в IQ-tree предковой последовательности ни у человека, ни у макаки. На пуле всех SNP мы увидели превышение наблюдаемого числа совпадающих SNP относительно ожидаемого в 3.0 раз в синонимичных позициях и в 3.7 раз в несинонимичных позициях.

Эффект уменьшается до 2.6 и 3.4 после исключения из анализа CpG динуклеотидов. Эффект усиливается по мере добавления фильтра аллельных частот (без синглетонов он достигает 3.3 и 4.7 раз, без SNP частотой меньше 5% он достигает 4.0 и 7.4 раз).

Следует ожидать, что SNP из сайтов с повышенной скоростью мутации чаще становятся заменами. Мы наблюдаем, что если у одного из видов произошла замена, то вероятность SNP в гомологичной позиции другого вида повышена (в 2-4 раза выше, чем в случае позиций без замен).

Сейчас мы исследуем, при каких условиях можно наблюдать сходный эффект при повышении порога фильтра аллельных частот в нейтральных сайтах, с помощью симуляций в SLIM. В планах повторить анализ на расширенных данных по человеку и оценить долю несинонимичных сайтов, на которые отрицательный отбор действует сходным образом у обоих видов.

Благодарности. Федору Кондрашову и Ладе Исаковой за предоставленную базу данных кодонных выравниваний ортологичных генов позвоночных.

Истоиники и литература:

1. Hodgkinson, A., Ladoukakis, E., Eyre-Walker, A., 2009. Cryptic Variation in the Human Mutation Rate. PLoS Biol 7, e1000027. https://doi.org/10.1371/journal.pbio.1000027

2. Seplyarskiy, V.B., Kharchenko, P., Kondrashov, A.S., Bazykin, G.A., 2012. Heterogeneity of the Transition/Transversion Ratio in Drosophila and Hominidae Genomes. Molecular Biology and Evolution 29, 1943–1955. https://doi.org/10.1093/molbev/mss071

3. The 1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. Nature 526, 68–74. https://doi.org/10.1038/nature15393

4. Warren, W.C., Harris, R.A., Haukness, M., Fiddes, I.T., Murali, S.C., et al. Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. Science. 2020;370(6523):eabc6617. doi: 10.1126/science.abc6617.

5.14 Динамика dN/dS на малых эволюционных расстояниях

Авторы: Е. Белоусова[1], А. Столярова[2], А. Кондрашов[1,3], Г. Базыкин[2,4]

1 - Факультет биоинженерии и биоинформатики МГУ им. Ломоносова, Москва, Россия, 2 - Центр Наук о Жизни, Сколковский институт науки и технологий, Сколково, Россия, 3 - Мичиганский университет, Анн-Арбор, США, 4 - Институт проблем передачи информации РАН, Москва, Россия

Для измерения давления отбора на эволюционирующие последовательности используется показатель dN/dS. Изначально этот показатель был разработан для применения на больших эволюционных расстояниях, но зачастую используется и для видов, разошедшихся недавно, и в случаях, когда сложно определить границы видов. Существует теоретическое выражение для dN/dS в состоянии мутационно-отборного равновесия, однако до достижения равновесия установить зависимости между показателями dN, dS (то есть, эволюционным расстоянием) и dN/dS сложно [1, 2, 3].

Наше предположение состоит в том, что если на только что отделившуюся популяцию действует положительный отбор, то в ней положительные мутации могут сначала фиксироваться быстрее, чем нейтральные, что приведет к более высокому dN/dS, чем при этих же параметрах в условиях равновесия. По мере накопления нейтральных замен dN/dS будет падать до своего равновесного значения. Такая динамика dN/dS означала бы, что этот показатель ведет себя неединообразно во времени.

Теоретические зависимости dN и dS от времени получены в работе [4] с использованием теории Poisson Random Field, которая является расширением диффузионной теории, используемой для вывода равновесного значения dN/dS. Однако в [4] зависимость dN/dS от времени не рассматривается. Но при делении зависимости для dN на зависимость для dS мы получили убывающую гиперболическую зависимость с вертикальной ассимптотой в виде оси y и горизонтальной – в виде равновесного значения dN/dS.

В этой работе мы хотим проверить теорию из [4] с помощью симуляций эволюции. В нашей модели родительская популяция сначала эволюционирует до равновесного состояния, а затем разделяется на две дочерние популяции, и все зависимости мы рассчитываем с момента разделения популяций. По предварительным результатам динамика dN и dS, действительно, описывается теорией Poisson Random Field, и, похоже, dN/dS получается значительно выше равновесного теоретического значения на малых временах, что потенциально может упростить проблему детекции положительного отбора.

Источники и литература:

1. Kryazhimskiy, S., & Plotkin, J. B. (2008). The population genetics of dN/dS. PLoS genetics, 4(12), e1000304.

2. Rocha, E. P., Smith, J. M., Hurst, L. D., Holden, M. T., Cooper, J. E., Smith, N. H., & Feil, E. J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. Journal of theoretical biology, 239(2), 226–235.

3. Mugal, C. F., Wolf, J. B., & Kaj, I. (2014). Why time matters: codon evolution and the temporal dynamics of dN/dS. Molecular biology and evolution, 31(1), 212–231.

4. Sawyer, S. A., & Hartl, D. L. (1992). Population genetics of polymorphism and divergence. Genetics, 132(4), 1161–1176.

5.15 Отбор в межгенных участках дрожжей Saccharomyces cerevisiae

Авторы: Павел Шелякин[1,2], Михаил Гельфанд[3]

1 - Институт проблем передачи информации им. А.А.Харкевича РАН, 2 - Институт общей генетики им. Н.И.Вавилова РАН, 3 - Сколковский институт науки и технологий (Сколковский Институт Науки и Технологий, Москва)

Дрожжи S. cerevisiae являются одним из модельных объектов и важным для биотехнологии и промышленности организмом. Около 70% генома S. cerevisiae занимают гены, большинство из которых не имеет интронов. Оставшиеся 30% некодирующей ДНК содержат промотеры, сайты связывания белков и другие регуляторные элементы, поэтому доля «нефункциональной» ДНК ещё ниже. Одним из методов поиска функциональных элементов в межгенных участках является филогенетический футпринтинг - метод, основанный на анализе консервативных участков ДНК в множественном выравнивании родственных последовательностей.

В данной работе мы планируем применить филогенетический футпринтинг для поиска регуляторных элементов в межгенных участках дрожжей, а также описания интенсивности и типа отбора, действующего на них. На данный момент проведён отбор и аннотация геномов S. cerevisiae из базы данных GenBank и из проекта 1002 Yeast Genomes, несколькими методами построены ортологические группы генов, а межгенные участки из разных штаммов S. cerevisiae, расположенные между парами ортологичных генов объединены в ортологичные группы межгенных участков (ОГМ). Кроме того, из открытых баз данных собрана информация об известных промотерах, сайтах связывания белков, 5` и 3`-некодирующих участках.

На следующей стадии работы планируется адаптировать метод полуавтоматического поиска консервативных участков в множественных выравниваниях ОГМ, сопоставить их с известными функциональными последовательностями генома и провести анализ действующего на них отбора.

Благодарности. Дмитрий Скрипка и Александр Соколов принимали участие в отборе, обработке и анализе данных

5.16 Анализ действия отбора на мобильные элементы и homing-эндонуклеазы в митогеномах грибов группы Basidiomycota

Авторы: Скаков Иван[1], Георгий Базыкин[2]

1 - ФББ МГУ им. М.В.Ломоносова, Москва Россиия, 2 - Skoltech, Center of Life Sciences

Прокариоты известны компактностью своего генома и небольшим количеством (по сравнению с эукариотами) мобильных элементов в геноме. Митохондрии – потомки альфа-протеобактерий, у большинства эукариот митогеномы крайне малы. Однако митогеномы грибов, в частности группы Basidiomycota, значитиельно варьируют по размеру - от 25 кб (Cryptococcus neoformans) до 257 кб (Clavaria fumosa) - для сравнения, размер человеческого генома - 16,6 кб. При этом мобильные элементы в митохондрия встречаются нечасто - к примеру, в этой работе достоверно найден пока единственный ретротранспозон, в митогеноме Rhizoctonia solani. Основной вклад в увеличение размера грибных митогеномов вносят так называемы homing-(“самонаводящиеся”)-эндонуклеазы. По механизму размножения они больше всего похожи на ДНК-транспозоны, при этом встраиваются эти эндонуклеазы вместе с самосплайсирующимся итроном - и, таким образом, уходят от действия отрицательного отбора.

Целью этой работы стала оценка действия отбора на homing-эндонуклеазы и мобильные элементы в митохондриальном геноме, по возможности сравнение с действием отбора на их ядерные гомологи в том же организме.

В этой работе поиск homing-эндонуклеаз и мобильных элементов проводился в митохондриальных геномам грибов из группы Basidiomycota - это 155 доступных геномов 142 различных видов. На первом этапе все доступные геномы автоматические аннотировалися программой PROKKA, все найденные CDS далее прогонялись через HMMERscan, далее в работе анализировались только CDS с доменами homing-эндонуклеаз (ради интереса результаты сравнивалися с опубликованной аннотацией для данного CDS, если она была). Выходы программ-поисковиков мобильных элементов: RepeatModeller2 и EDTA, также фильтровались на наличие транспозонных доменов.

Далее планируется сгруппировать эндонуклеазы по семействам (основные семейства обнаруженных эндонуклеаз - LAGLIDADG и GIY-YIG), для каждого семейства оценить давление отбора - как независимо для каждого организма, так и в среднем для всех представителей Basidiomycota. Аналогичную работу надо проделать и с обнаруженными

митохондриальными мобильными элементами.

5.17 Estimating the number of cancer driver mutations through mutation bias

Authors: A. V. Stolyarova[1], G. A. Bazykin[1,2]

1 - Skolkovo Institute of Science and Technology, 2 - Institute for Information Transmission Problems, Russian Academy of Sciences In the course of adaptation, positive selection drives the spread and fixation of beneficial mutations.

However, the initial source of positively selected variants is the mutational process. Although new mutations are acquired randomly, some types of mutations are more likely to occur than others due to molecular mechanisms of DNA replication and repair. If standing genetic variation is low, adaptation is generally limited by waiting for new beneficial mutations, which become fixed quickly. In this case, the distribution of beneficial mutations accumulated in the course of adaptation is biased by the underlying distribution of mutation rates. An example of mutation-driven adaptive evolution is tumor development, which is initiated when a cell acquires a positively selected “driver” mutation. Indeed, for most cancer genes, the observed driver mutations are enriched in those mutation types that occur at a higher rate. On the other hand, multiple studies show that the distribution of driver mutations differs from that of passenger mutations: although they both are introduced by random mutation, the probability of spread of a driver mutation is defined by selection. Therefore, the strength of mutation bias among the observed driver mutations is defined by the relative contribution of the non-uniformity of mutation rate and selection. Saturation analysis of currently available data show that the catalog of driver mutations is incomplete, so that sequencing of more tumor samples increases the number of candidate drivers. We propose that by measuring the mutation bias we can characterise the set of potential driver mutations, both observed and unobserved, and infer selective constraints shaping the occurrence of these drivers in the tumor samples. In order to do it, we’ll compare the spectra of annotated drivers to the spectra of neutral mutations of the same type (i.e. 3-letter contexts). If the general number of potential drivers is large and therefore the impact of selection is weak, mutation rates will determine the observed frequencies of different types of driver mutations — such patterns have previously been shown for tumor suppressor genes, where many driver mutations are possible because most loss-of-function mutations can be drivers. The opposite situation is observed in oncogenes: here, a higher than expected fraction of driver mutations is observed at slowly mutating sites, indicating that selection limits the set of potential drivers. We are going to develop a mathematical framework to examine how the occurrence of different types of driver mutations and the recurrence of specific drivers depend on the number of potential drivers. In order to account for the non-uniformity of fitness effects of driver mutations, we’ll analyze the recurrence of specific

driver mutations: strongly beneficial mutations will be observed in multiple tumor samples even if the background mutation rate is low. Next, we'll use tumor sequencing data to estimate mutation bias in annotated drivers in multiple tumor types. We are going to apply our model to these data to estimate the number of potential drivers and the distribution of fitness effects among them based on the mutational spectra of the drivers and their recurrence. Similarly, we'll characterize the pool of potential drivers separately for oncogenes and tumor suppressor genes. By applying our method to tumor sequencing data, we aim to predict how many mutations in cancer genes are yet to be discovered and how many tumor samples are required to detect these mutations.

5.18 Mutation Annotation machine learning algorithm for predicting the clinical effect of single nonsynonymous nucleotide substitutions in cancer cells

Authors: Aleksandr Sarachakov, Viktor Svekolkin, Anna Parfenenkova, Anastasia Yuidina, Zoia Antysheva, Jessica H. Brown, Alexander Bagaev, Nathan Fowler and Mikhail Gelfand

Abstract

With the wealth of next-generation sequencing (NGS) produced over the last decade, distinguishing between pathogenic and neutral single nucleotide missense mutations is essential to understanding disease pathogenesis and cancer treatment selection and optimization. Current experimental methods, such as overexpression assays, siRNA, knock-out models, and studies in cultured cells or model organisms, can help uncover the phenotype of a set of mutations in genes but are not optimal for the identification of pathogenic mutations in thousands of potential candidates. In this study, we present a new mutation annotation tool, MutAnt, for the prediction of mutation pathogenicity that achieves superior performance by integrating state-of-the-art machine learning techniques with feature selection and hyperparameter tuning.

To develop MutAnt, a supervised machine learning algorithm was trained to classify pathogenic and neutral single nonsynonymous nucleotide substitutions. The training dataset was based on ClinVar v20170601 and ClinVar v2020124, the CancerGenomeInterpreter database of validated pathogenic mutations, and benign mutations from the database of Cancer Passenger Mutations (dbCPM). The Boruta algorithm with Shapley values was used for feature selection, and the Bayesian optimization was applied for the model hyperparameter tuning. Diverse validation methods were utilized in the development of MutAnt, such as cross-validation; the Database of Curated Mutations (DoCM); holdout on new mutations documented in the latest released version of ClinVar database, which contains validated disease-associated single nucleotide polymorphisms (SNPs) not previously annotated in existing algorithms; mutations with high allele frequency; test dataset III from VariBench; and correlation with the mean function score in a saturation genome

editing experiment for the BRCA gene.

MutAnt exhibited a stronger performance than other methods on all of the tested validation strategies, which was demonstrated by its accuracy (high f1-score, 0.95 to 0.99) and sensitivity-specificity value (ROC-AUC value, 0.93 to 0.99) in classifying mutations in each validation dataset. Furthermore, the present approach allowed for the prediction of novel mutations that are not covered by extensive population studies with higher accuracy than other methods. This was demonstrated by its prediction of the clinically relevant BRCA variants in the saturation genome editing experiment.

We also proposed a new pipeline for mutation calling from RNAseq experiments including MutAnt, which highly increases precision of a procedure.

In conclusion, MutAnt outperformed all compared mutation annotation algorithms in this study for the classification of clinically relevant benign and pathogenic mutations in the ClinVar dataset. Importantly, this approach allows prediction of novel mutations, subsequently allowing for the selection of targeted therapies for a broader set of patients.

5.19 Анализ полиморфизмов, ассоциированных с шизофренией, в группе больных и здоровых индивидуумов

А.Кононкова[1,2], И.Плетенев[1], М. Базаревич, А.Черкасов[1], О.Ефимова[1], Н.Кондратьев[3], Е.Храмеева[1]

1 - Сколковский институт науки и технологий, [2] ИППИ РАН, [3] ФГБУН НЦПЗ,

Известно, что нарушения психики относятся к мультифакторным заболеваниям и приводят к комплексным изменениям профиля различных биомолекул и структур клеток мозга. Поэтому интеграция омиксных данных различного типа и их анализ лежит в основе наиболее масштабных и современных исследований в данной области. Так, например, база данных SZDB содержит информацию о экспрессии, вариациях числа копий, метилировании, данные полногеномных ассоциативных исследований и проч. Структура хроматина и ее изменения при психических заболеваниях также могут послужить ценным источником дополнительной информации, однако подобные данные представлены лишь небольшим количеством здоровых образцов. Таким образом, изучение структуры хроматина в целом у больных шизофренией, а также анализ особенностей регуляции, которые могут проявляться на уровне изменения частот контактов отдельных геномных локусов, и интеграция с другими типами данных (экспрессия, генотипирование) является важной задачей при исследовании психических нарушений.

На первом этапе проекта были получены данные генотипирования для двух групп по шесть человек (6 здоровых образцов, 6 образцов больных шизофренией разного возраста и

пола). Предварительный анализ проводился на подгруппе полиморфизмов, ассоциированных с шизофренией согласно данным GWAS; были отобраны потенциально значимые (“надежные”) полиморфизмы из двух источников – Won et al, 2016 (данные PGC, метод CAVIAR) и Pardinas et al, 2019 (данные PGC2 и CLOZUK, метод FINEMAP).

В ходе предварительного анализа не выявлено различий в количестве полиморфизмов, ассоциированных с заболеванием, их значимости, отношении шансов (для полиморфизмов) и их пересечением с энхансерами в исследуемых группах.

Был применен альтернативный подход, основанный на определении двух групп полиморфизмов (Ш и К): в группе больных для каждого образца были отобраны те варианты, которые ассоциированы с заболеванием, но при этом не встречаются ни в одном контрольном образце, и, наоборот, в контрольной группе для каждого образца отбирались варианты, также ассоциированные с заболеванием, но не встречающиеся ни у одного больного. При таком анализе удалось обнаружить различия в двух группах: в группе больных подобных полиморфизмов оказалось больше, чем в контрольной группе. Была обнаружена связь между принадлежностью полиморфизмов к группе Ш/К и частоте пересечения с энхансерами (p-value 0.025 , хи-квадрат Пирсона). Пермутационный тест также показывает значимо более частое пересечение с энхансерами для группы Ш (p-value 0.01) и значимо менее частое пересечение для группы К (p-value 0.03). Однако взаимосвязь с энхансерами оказалась статистически значимой для исходной выборки “надежных” полиморфизмов, полученных в статье Pardinas et al, но не подтвердилась для данных из статьи Won et al.

В настоящий момент актуально сравнение различных методов получения списка полиморфизмов, которые могут быть напрямую связаны с определенным заболеванием (сравнение методов CAVIAR, FINEMAP и др). В перспективе планируется исследование особенностей структуры хроматина в образцах мозга больных шизофренией, сравнение с данными по экспрессии и генотипированию в тех же образцах.

6. Белки

6.1 Prediction of Protein Stability Change Upon Mutation using Deep Learning

Authors: Marina A. Pak, Dmitry N. Ivankov

Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow

Protein stability is a crucial characteristic of a protein and predicting its alteration upon amino acid substitution is essential for understanding of protein folding mechanisms as well as engineering of proteins with desired properties. Experimental estimation of protein stability change upon mutation, being cumbersome and limitedly executable, makes computational prediction of this parameter a relevant and challenging task.

To date, a great number of computational methods for prediction of protein stability change upon amino acid substitutions have been developed; nevertheless, they are still far from delivering the level of robustness for wide application for protein engineering. Among the most common drawbacks are overfitting, bias towards destabilizing mutations, violation of anti-symmetry principle. The proposed project is aimed to develop a computational tool for prediction of protein stability change upon mutation combining multiple approaches to overcome the limitations of modern predictors. The predictor is designed to be based on deep neural network trained on a large balanced and symmetrized dataset of artificial and experimental data of folding free energy changes upon mutation.

6.2 Mechanism of uranyl ion (UO_2^{2+}) molecular toxicity towards zinc finger-containing DNA-binding proteins revealed through in silico simulations

Authors: Egor S. Bulavko, Dmitry N. Ivankov

Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow, Russia

DNA-binding proteins are essential for many cellular processes including replication, transcription, gene expression, etc. To perform their functions, the special domains are utilized that allow DNA to be bound. Zinc fingers are among the most frequent eukaryotic DNA-binding domains, common for various transcription factors, chromatin remodelers and DNA reparation proteins (Laity et al., 2001).

It has long been known that one of the major consequences of uranium salts poisoning is DNA binding proteins inhibition (Hartsock et al., 2007). Recently it was shown that uranyl ion UO_2^{2+} – the most stable uranium form under physiological conditions – directly interacts with zinc fingers (Zhou et al., 2021), which impairs their DNA-binding activity. Nevertheless, the atomistic mechanism of

inhibition remains unresolved. The only thing we know is that, according to Zhou and coauthors, UO_2^{2+}/Zn^{2+} binding is competitive.

In this project we, following Zhou and colleagues, are working with PARP-1 zinc finger motif and trying to answer the following questions:

- Do uranyl and zinc ions have the similar binding sites?
- How does uranyl affect zinc finger's conformation and disrupts DNA-protein interactions?

For this purpose, we enroll the vast arsenal of computational chemistry and perform different types of simulations using various approaches, including quantum chemistry and classical molecular mechanics.

Источники и литература:

1. Hartsock W.J., Cohen J.D., Segal D.J. Uranyl acetate as a direct inhibitor of DNA-binding proteins // Chem. Res. Toxicol. 2007. V. 20, № 5. - P. 784–789.

2. Zhou X., Xue B., Medina S., Burchiel S.W., Liu K.J. Uranium directly interacts with the DNA repair protein poly (ADP-ribose) polymerase 1 // Toxicol. Appl. Pharmacol. 2021. V. 410, - P. 115360.

3. Laity J.H., Lee B.M., Wright P.E. Zinc finger proteins: new insights into structural and functional diversity // Curr. Opin. Struct. Biol. 2001. V. 11, № 1. - P. 39–46.

6.3 Классификация семейств ДНК-узнающих белковых доменов на основе структурных особенностей ДНК-белковых комплексов

Авторы: Вера Панова[1], Евгений Баулин [2], Анна Карягина [3,4,5], Андрей Алексеевский [3,6], Сергей Спирин [3,6,7].

1 - Факультет биоинженерии и биоинформатики, МГУ, Москва, Россия, 2 - Институт математических проблем биологии РАН – Филиал Федерального государственного учреждения "Федеральный исследовательский центр Институт прикладной математики им. М. В. Келдыша Российской академии наук", Пущино, Московская область, Россия, 3 – НИИ Физико-химической биологии им. А.Н.Белозерского МГУ, МГУ имени М.В.Ломоносова, Москва, Россия, 4 - Национальный исследовательский центр эпидемиологии и микробиологии имени Н. Ф. Гамалеи, Москва, Россия, 5 - Всероссийский НИИ сельскохозяйственной биотехнологии, Москва, Россия, 6 - Научно-исследовательский институт системных исследований РАН, Москва, Россия, 7 - Национальный исследовательский университет «Высшая школа экономики» , Москва, Россия

На настоящий момент в открытых базах данных доступно 6557 ДНК-белковых структур. Разные белки взаимодействуют с двойной спиралью ДНК разными способами. Создание классификации ДНК-белковых взаимодействий и ее интеграция в базу ДНК-белковых комплексов позволяет эффективно исследовать закономерности ДНК-белкового узнавания.

Цель данной работы - исправить ошибки в существующей версии NPIDB ([1], http://npidb.belozersky.msu.ru/), а далее, используя исправленные данные из NPIDB, создать классификацию структур ДНК-белковых комплексов на современном материале, а также описать структурные особенности и внутреннюю классификацию наиболее популярных семейств.

Классификация основана на принципах из работ [1,2], но вместо доменов, выделяемых в цепях белка согласно данным банка SCOP (поддержка которого прекращена в 2009 г.), классифицируются домены, выделяемые согласно банку Pfam [3]. Классификация основана на контактах между молекулами ДНК и белка. Рассматриваются водородные связи, гидрофобные взаимодействия и водяные мостики. Тип контакта — это пара контактирующих элементов структуры. В белках мы выделяем три элемента: спираль, бета-лист, поворот или неструктурированный сегмент, в ДНК тоже три элемента: большая бороздка, малая бороздка, сахарофосфатный остов). Способ взаимодействия домена белка определяется как список типов контактов структур этого домена с ДНК (структурой домена мы называем конкретную часть записи PDB с расшифровкой её структуры, т.е. техническую реплику, а доменом — участок белка). Класс взаимодействия для семейства доменов определяется как пересечение способов взаимодействия доменов [2].

Был написан комплекс программ на Python, который определяет способы взаимодействия для доменов и структур доменов, а также классы взаимодействия для семейств. Рассматривались только структуры белка с двойной спиралью ДНК (не менее шести комплементарных пар), решённые методом рентгеноструктурного анализа или криоэлектронной микроскопии, с разрешением менее 3Å. Всего на декабрь 2021 (последнее обновление NPIDB) доступно 4089 структур ДНК-белковых комплексов, удовлетворяющих нашим условиям. В белках этих комплексов выявлено 10262 контактирующих с ДНК структур 856 доменов. Семействам, представленным не менее чем тремя различными доменами с доступными структурами в комплексах с ДНК, приписаны классы взаимодействия, всего таких семейств 100. Всего на данный момент реализуется 37 классов взаимодействия (из 512 теоретически возможных). Самые популярные — (i) спираль с большой бороздкой и петля с большой бороздкой, (ii) спираль с большой бороздкой, (iii) петля с большой бороздкой (по 9 семейств в каждом из этих классов). Описано несколько конкретных семейств, в т.ч. числе шесть семейств, называемых “разнородными”, это семейства, у которых нулевой класс взаимодействия. Показана внутренняя классификация

“разнородных” семейств, определены причины, по которым данные семейства являются “разнородными”.

Источники и литература:

1. Kirsanov D.D. et al. NPIDB: nucleic acid-protein interaction database. Nucleic Acid Research 2013, 41 (D1):D517–D523, doi: 10.1093/nar/gks1199

2. Zanegina O.N. et al. An updated version of NPIDB includes new classifications of DNA-protein complexes and their families. Nucleic Acid Research 2016, 45 (D1):D144–D153, doi: 10.1093/nar/gkv1339.

3. Mistry, J., et al. (2020). Pfam: The protein families database in 2021. Nucleic Acids Research, 49(D1). doi:10.1093/nar/gkaa913

6.4 Изучение распространения и эволюции систем рестрикции-модификации семейства AlwI

Авторы: О.Л.Макарикова, А.В.Алексеевский МГУ имени М.В.Ломоносова, факультет биоинженерии и биоинформатики, Москва, Россия

Системы рестрикции-модификации (Р-М) - защитные системы бактерий и архей против чужеродной ДНК, представленные белками с метилтранферазной и эндонуклеазной активностями. Системы Р-М типа II в силу простого устройства и понятного механизма действия получили широкое применение в молекулярной биологии и генной инженерии, в частности, системы с эндонуклеазой рестрикции (ЭР) семейства AlwI, к которому относятся некоторые никирующие ЭР (никазы). Целью данной работы является изучение распространения и эволюции (горизонтальные переносы) систем Р-М, содержащих белки этого семейства.

Выравнивание последовательностей белков семейства Pfam RE_AlwI было размечено по доменам, определенным по пространственной структуре никазы N.BspD6I. Оказалось, что Pfam профиль семейства описывает ДНК-связывающий и линкерный домен (всего 290 а.к.о.) и частично каталитический домен (145 из 221 а.к.о.). По выравниванию каталитических доменов проверенной и не-избыточной выборки (87 последовательностей), полученной из 218 ЭР из REBASE (содержащей информацию о всех системах Р-М) с доменом RE_AlwI, был построен новый профиль семейства, названный cat_AlwI. Он был сравнен с профилем Pfam на 1097 полных протеомах прокариот из базы данных Uniprot, содержащих хотя бы одну ЭР с доменом RE_AlwI. Проверено, что 84 последовательности с доменом RE_AlwI, не найденные профилем cat_AlwI, найдены профилем RE_AlwI за счет сходства ДНК-связывающего и

линкерного доменов, а не каталитического.

По данным REBASE о системах Р-М с ЭР с доменом RE_AlwI найдено 20 классов систем Р-М, содержащих разные метилтрансферазы из 3-х семейств. Классом считается группа систем Р-М с одинаковыми каталитическими доменами ЭР и метилтрансфераз. Самым представленным оказался класс с одной ЭР и двумя метилтрансферазами семейства Pfam MethyltransfD12. Для этого класса были получены координаты в геноме для генов белков ЭР, найденных профилем cat_AlwI в упомянутых выше 1097 протеомах, и метилтрансфераз, найденных в тех же 1097 протеомах профилем Pfam MethyltransfD12. В результате были отобраны белки, находящиеся в одном локусе и наиболее вероятно ассоциированные с одной системой Р-М. Для таких белков ЭР (80 последовательностей) и метилтрансфераз (155 последовательностей) были построены филогенетические деревья с целью дальнейшего анализа эволюции этого класса систем Р-М.

В будущей работе планируется для каждого выделенного класса систем Р-М получить последовательности белков ЭР и метилтрансфераз на основании расположения в геноме соответствующих генов и провести их филогенетический анализ с целью определения эволюционных событий, таких как горизонтальные переносы систем в целом и отдельных генов между классами систем Р-М, содержащими ЭР семейства cat_AlwI, изучить разделение на клады для метилтрансфераз из систем Р-М с родственными ЭР для класса с двумя метилтрансферазами.

6.5 Редкие классы систем рестрикции-модификации

Авторы: Алина Кирпиченко, Андрей Алексеевский, МГУ имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия

В ходе эволюции прокариоты получили возможность эффективно отличать собственную ДНК от внедрившейся посторонней. Для этого ими были выработаны механизмы защиты. Один из них - системы рестрикции-модификации (РМ). Они образованы ферментами с двумя активностями: эндонуклеазы рестрикции (ЭР) и ДНК метилтрансферазы (МТ). ЭР типа II расщепляет ДНК в сайтах с определённой последовательностью, если сайты не метилированы. МТ обеспечивает различие между своей и чужой ДНК за счет переноса метильной группы на сайты с той же последовательностью. В данной работе рассмотрены системы РМ, содержащие ЭР типа II Pfam семейства Bse634I.

Помимо вертикального наследования прокариоты передают генетический материал с помощью горизонтального переноса. Он помогает приобретать новые метаболические возможности и адаптироваться к среде [1]. Цель работы – установить явное наличие

горизонтальных переносов и вертикального наследования систем РМ, содержащих ЭР Bse634I, и соответствующих генов МТ и ЭР отдельно. В базе данных Uniprot было найдено 83 последовательностей белков с доменом Bse634I. Для поиска всех гомологов ЭР семейства Bse634I использовался BLASTp и поиск по профилю семейства Bse634I из Pfam в трансляциях открытых рамок полных бактериальных геномов. Добавилось 17 белков. На основе анализа множественного выравнивания последовательностей, разметки его по вторичной структуре ЭР Bse634IR с известной 3D структурой и идентификации каталитического мотива PDx(10,51)[DE]x(1,13)K [2] были удалены вероятные ошибки аннотации. Осталось 82 ЭР. С помощью программы BLAST и баз данных были найдены 79 последовательностей МТ, входящих вместе с обнаруженными ЭР в одну систему РМ. Из найденных белков 76 МТ содержат консервативные мотивы, гомологичны и принадлежат одному Pfam семейству DNA_methylase (PF00145).

Найденные нами 76 систем РМ класса DNA_methylase/Bse634I присутствуют в 45 родах из 42 семейств бактерий. Такое распределение гомологичных систем по таксонам высокого уровня может говорить о том, что в эволюции система РМ недолго существует в одном виде, но часто переносится горизонтально. В ходе анализа филогенетических деревьев было выявлено предположительно 24 случая вертикального наследования систем РМ, 2 случая горизонтального переноса систем целиком. Рассмотрено 7 возможных горизонтальных переносов генов ЭР и МТ. Ранее в лаборатории студенткой Гусевой Е.А были изучены классы систем РМ, содержащие ЭР семейства RE_Tdell, также являющуюся редкой. В 6620 полных геномах найдено 45 ЭР семейства RE_Tdell, принадлежащих двум классам систем РМ: DNA_methylase/RE_TdeIII и N4_N6_Mtase/RE_TdeIII. В результате получено: 2 случая обмена МТ между классами, 11 случаев обмена генами внутри класса, 13 случаев горизонтального переноса целой системы, порядка 20 случаев вертикального наследования.

В базе данных Pfam было найдено около 40 семейств малочисленных (<100) ЭР. Представляет интерес и планируется изучение эволюции таких редких систем РМ.

Источники и литература:

1. Garushyants S.K., Gelfand M.S., Kazanov M.D. Horizontal gene 47 transfer and genome evolution in Methanosarcina // BMC Evol. Biol., Jun. 2015. vol.15. p.102.

2. Grazulis S. et al. Crystal structure of the Bse634I restriction endonuclease: comparison of two enzymes recognizing the same DNA sequence // Nucleic Acids Res., Feb. 2002. vol.30. No.4 pp. 876-85.

6.6 Универсальные доменные архитектуры белков их консервативность и эволюция

Авторы: Р. Ириоглов, студент 6 курса, А.В. Алексеевский, к. ф.-м. н., в. н. с., И.С. Русинов, н. с., ФББ, МГУ имени М.В. Ломоносова, Москва, Россия

Основной единицей эволюции белка является домен. Для белкового домена существует несколько определений, характеризующих его с разных сторон: структурный домен – стабильная компактная структурная единица белка, способная сворачиваться отдельно от других подобных единиц; функциональный домен – часть белка, выполняющая либо способствующая выполнению функции белка; эволюционный домен – относительно консервативная часть последовательности белка, встречающаяся как отдельно, так и вместе с другими подобными частями на одной белковой цепи и способная участвовать с ними в перестановках. Часто, но не всегда все 3 характеристики домена совпадают. Домены в белке расположены в определённом порядке, который называется доменной архитектурой. Домены и доменные архитектуры, встречающиеся у представителей всех трёх суперцарств живой природы: бактерий, архей и эукариот, мы называем универсальными. Понятие универсальности доменов и архитектур можно применять к суперцарствам или другим таксонам высокого уровня.

Целью текущей работы является исследование эволюционной консервативности универсальных доменных архитектур на примерах.

Для нахождения универсальных доменных архитектур была составлена таблица, в которой для каждого домена из Pfam было определено количество встреч в последовательностях, принадлежащих трём суперцарствам живой природы: Bacteria, Archaea и Eukaryota. Из данной таблицы были отобраны домены, встречающиеся во всех суперцарствах, т.е. универсальные домены, всего было отобрано 4219 доменов.

Моделью для исследования стал универсальный домен UVR (PF02151), который преимущественно входит в состав белков системы эксцизионной репарации нуклеотидов у прокариот, однако встречается и в других белках, в том числе и у эукариот. Основное участие в эксцизионной репарации принимают белки-компоненты мультиферментного комплекса, который носит название UvrABC эндонуклеаза. Субъединицами этого комплекса являются белки UvrA, UvrB и UvrC. Домен UVR белка UvrB может взаимодействать с доменом UVR, белка UvrC [1]. Две наиболее распространённые доменные архитектуры с доменом UVR ассоциированы именно с белками UvrB и UvrC. Это архитектуры ResIII, UvrB_inter, Helicase_C, UvrB, UVR и GIY-YIG, UVR, UvrC_RNaseH_dom, HhH_5, которые мы обозначаем arch_UvrB и arch_UvrC соответственно.

Изначально предполагалась стабильность данных архитектур, однако на основании кластеризации последовательностей домена UVR и средних эволюционных расстояний между

ними можно сделать вывод, что доменные архитектуры с этим доменом не выделяются в отдельные клады на филогенетическом дереве. Это может быть связано со сложной эволюционной историей. Дополнительную трудность вызывает небольшой размер домена (менее 40 а.к.).

На данный момент, для отбора универсальных доменных архитектур для дальнейшего исследования создана таблица, где для всех доменных архитектур базы данных Pfam, образованных из универсальных доменов, указано количество встреч у организмов каждого из 3 суперцарств. В дальнейшем планируется выделить из данной таблицы архитектуры с частотой встреч в каждом из суперцарств, превышающем определённый числовой порог и провести исследование их эволюционной консервативности

1. The C-terminal region of the UvrB protein of Escherichia coli contains an important determinant for UvrC binding to the preincision complex but not the catalytic site for 3'-

incision. Moolenaar GF, Franken KL, Dijkstra DM, Thomas-Oates JE, Visse R, van de Putte P, Goosen N. J. Biol. Chem. 270, 30508-15, (1995). PMID: 8530482

6.7 Предсказание эпитопов белков коронавируса SARS-CoV-2

Авторы: Мария Тутукина[1,2,3], Анна Казнадзей[1,4], Андрей Комаров[4], Татьяна Бессонова[3], Илья Мазо[4]

1 - Институт проблем передачи информации им. А.А.Харкевича РАН, 2 - Сколковский институт науки и технологий (Сколковский Институт Науки и Технологий, Москва), 3 - Институт биофизики клетки РАН (ФИЦ ПНЦБИ РАН), 4 - VirIntel, LLC

При попадании вируса в организм человека его иммунная система начинает производить специфические антитела. Участок вирусного белка, с которым непосредственно взаимодействует антитело, называется эпитоп. Эпитопы могут совпадать или различаться у разных штаммов вируса. Соответственно, важно понимать, что именно является эпитопом и как, например, мутации в нем отразятся на иммунной реакции организма. Кроме того, пептиды (короткие аминокислотные последовательности) проще синтезировать, чем белки - поэтому пептиды, соответствующие подтвержденным эпитопам, могут быть удобны в практическом применении в качестве основы антигенных экспресс-тестов и других систем.

В геноме коронавируса SARS-CoV-2 закодировано десять белков. Для четырех из них (шиповидного, S, нуклеокапсидного, N, мембранного, M и белка оболочки, E) мы определили потенциальные эпитопы с помощью инструмента BepiPred-2.0. Принцип работы данной программы основан на обучении алгоритма Random Forest на выборке из кристаллических структур антитело-антиген из базы данных Protein Data Bank с пятикратной процедурой кросс-

валидации. По полученным результатам были синтезированы пептиды. В некоторых случаях длина пептидов была впоследствии увеличена, и соответствующий эпитоп был исследован в двух вариантах, на коротком и длинном пептиде. Мы также исследовали пептиды, “сшитые” с белком GST (глутатион-сульфо-трансфераза) и отдельно в целях подбора условий их правильной конформации.

Иммуногенность потенциальных эпитопов была проанализирована с помощью иммуноферментного анализа (ИФА) на плашках. В качестве образцов была использована инактивированная сыворотка пациентов, переболевших подтвержденным COVID-19 или привитых вакцинами Sputnik V или КовиВак. В качестве отрицательных контролей была использована инактивированная сыворотка, собранная у пациентов до 2019 года.

Первым важным наблюдением было то, что химически синтезированные пептиды длиной 20-30 аминокислот не распознаются антителами сыворотки переболевших. Поэтому вся дальнейшая работа проводилась на пептидах, сшитых с N-конца с белком GST. Второе наблюдение заключалось в том, что ни один из эпитопов M-белка не давал статистически значимого сигнала ни в одном из вариантов, а в S-белке особенно иммуногенными являются эпитопы, входящие в состав рецептор-связывающего домена (RBD) и расположенные в области фуринового сайта на границе доменов S1/S2. Интересно, что чем длиннее пептид, тем больше паттерн его узнавания похож на таковой для RBD, независимо от того, где этот пептид расположен. Это, возможно, говорит о том, что при болезни и вакцинации указанными вакцинами антитела в равной степени образуются к разным участкам S-белка. В настоящее время мы в процессе завершения анализа свойств выбранных пептидов, а также работаем с пептидами, содержащими мутации, характерные для штаммов дельта и омикрон.

6.8 OrthoQuantum: веб-сервис для визуализации эволюционного репертуара эукариотических белков

Авторы: И.Ильницкий [1,2], А.Жарикова [1,2], А.Миронов [1,2]

1 - Факультет Биоинженерии и Биоинформатики МГУ имени Ломоносова, 2 - Институт проблем передачи информации имени А. А. Харкевича РАН

Обширные объемы данных секвенирования нового поколения и омиксных исследований привели к накоплению информации, которая дает представление об эволюционном ландшафте родственных генов/белков. Ранее, скоррелированные паттерны потери белков использовалась в исследованиях для определения роли митохондриальных белков, кодируемых ядром [1] и белков, связанных с ресничками эукариотических клеток [2].

Белки, которые функционируют вместе, например, члены одного и того же пути или белкового комплекса, часто демонстрируют сходные паттерны представленности в разных филогенетических кладах. Филогенетическое распределение белков может быть записано в виде бинарных векторов, т. е. строк, которые кодируют наличие или отсутствие ортологов в других геномах. Эти строки используются для расчета корреляционных матриц, необходимых для поиска эволюционных событий, таких как потеря генов, дупликации и неортологичные замены. На практике мы ожидаем, что пары белков, по крайней мере, с двумя-тремя коррелирующими событиями приобретения или потери гена, почти наверняка функционально связаны [3].

Некоторые аспекты таких филогенетических профилей требуют интерпретаций и решений специалистов. Поэтому мы разработали автоматизированную процедуру поиска и визуализации филогенетического распределения белков (и соответствующих ортологичных групп) и их взаимной корреляции. Веб-сервис в полной мере использует данные гомологии из OrthoDB, PantherDB и NCBI [4,5]. Веб-интерфейс позволяет пользователям выполнять запросы по названиям генов и идентификаторам UniProt, проверять аннотации ортологов с помощью дополнительного поиска BLAST, просматривать выходные данные в графическом формате и загружать результаты в формате .svg. Одной из его сильных сторон является возможность выбора между компактным набором видов с хорошим качеством сборки генома и полным набором видов, которые могут обеспечить лучшее разрешение для филогенетического профиля в больших таксономических группах, таких как Metazoa. Для анализа доступны более чем 1000 полностью секвенированных эукариотических геномов и предсказанных для них ортологов. OrthoQuantum находится в свободном доступе: http://orthoq.bioinf.fbb.msu.ru.

В нашем исследовании показан пример использования инструмента для всестороннего изучения эволюции хроматин-ассоциированных белков. Мы рассмотрели и сравнили филогенетические профили набора 720 белков из 150 эукариотических геномов. На основании полученных данных были сделаны следующие основные выводы: (i) в базе данных OrthoDB наблюдается неточная кластеризация на уровне эукариот: не все имеющиеся гомологи попадают в ортологичные группы растений и простейших; (ii) белки комплекса Polycomb PRC1 значительно различаются между животными и растениями, в то время как PRC2 более консервативен у всех эукариот; (iii) у трематоды Schistosoma japonicum полностью отсутствует аппарат биогенеза piRNAs и самих белков PIWI, что указывает на альтернативный путь сайленсинга транспозонов.

Источники и литература:

1. Pagliarini DJ, Calvo SE, Chang B, et al. A mitochondrial protein compendium elucidates complex I disease biology. Cell. 2008;134(1):112-123. doi:10.1016/j.cell.2008.06.016

2. Avidor-Reiss T, Maer AM, Koundakjian E, et al. Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. Cell. 2004;117(4):527-539. doi:10.1016/s0092-8674(04)00412-x

3. Barker D, Pagel M. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput Biol. 2005;1(1):e3. doi:10.1371/journal.pcbi.0010003

4. Zdobnov, E. M., Kuznetsov, D., Tegenfeldt, F., Manni, M., Berkeley, M., & Kriventseva, E. V. (2021). OrthoDB in 2020: evolutionary and functional annotations of orthologs. Nucleic acids research, 49(D1), D389–D393. https://doi.org/10.1093/nar/gkaa1009

5. Mi H, Ebert D, Muruganujan A, et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res. 2021;49(D1):D394-D403. doi:10.1093/nar/gkaa1106

6.9 Особенности эволюции С-концевого домена белков нуклеоплазминового семейства

Авторы: Вяльцев В.[1], Ильницкий И.[1], Жарикова А.[1,2], Миронов А.[1,2]

1 - Факультет Биоинженерии и Биоинформатики МГУ имени Ломоносова, 2 - Институт проблем передачи информации имени А. А. Харкевича РАН

Впервые белок, впоследствии классифицированный как нуклеоплазмин, был выделен из Xenopus laevis, и для него была показана функция шаперона гистонов.[1] Позже похожие белки были выделены из человека. Оказалось, что в человеке экспрессируются три белка – NPM1, NPM2 и NPM3, каждый из которых выполняет функцию шаперона гистонов за счет наличия областей из кислых аминокислот в последовательности.[2] Помимо этого, у каждого из них есть свои особенности, например, NPM1, он же нуклеофозмин, в основном локализуется в ядрышке, однако может перемещаться между ядром и цитоплазмой, для него были показаны взаимодействие с нуклеиновыми кислотами, участие в сборке рибосом, формировании ядрышка, репарации, митозе, репликации, транскрипции и апоптозе, а мутации в данном белке могут вызывать острую миелоидную лейкемию.[3] NPM1 в клетке существует в виде пентамера, часть мономеров из которого может заменяться на NPM3, самостоятельно роль гистонового шаперона не играет.[4] NPM2 в свою очередь важен для ремоделирования хроматина сперматозоидов и в ходе раннего эмбриогенеза, а также является основным компонентом ядрышка ооцита.[2]

Группы белков NPM1, NPM2 и NPM3 выделяют в основном для позвоночных животных. Белки, принадлежащие беспозвоночным, причисляют к группе NPM-like белков.[2] В белках нуклеоплазминового семейства встречается два домена: N-концевой нуклеоплазминовый домен, ответственный за олигомеризацию, по наличию которого

определяется принадлежность к семейству, и С-концевой домен, который, как принято считать, встречается в основном в NPM1 позвоночных и ответственен за взаимодействие с рРНК.[2] Однако при поиске в белковых базах данных нам удалось обнаружить белки нуклеоплазминового семейства из беспозвоночных, которые содержат С-концевой домен, в частности некоторые из таких белков принадлежат трихоплаксу, некоторым представителям стрекающих и иглокожих[5]. В этой работе мы хотим определить, как эволюционировал С-концевой домен, почему он присутствует в одних таксономических группах и отсутствует в других, какие эволюционные события могли привести к этому. Для этого мы используем методы построения филогенетических деревьев на основе причисленных к нуклеоплазминовому семейству последовательностей, полученных из белковых баз данных Pfam и UniProt. Руководствуясь общими представлениями о структуре белков нуклеоплазминового семейства, для анализа мы собрали 1080 белковых последовательностей, принадлежащих 651 организму. В данный момент идет работа над созданием корректного филогенетического дерева на основе последовательностей N-концевого домена, где будут отмечены последовательности, имеющие С-концевой домен.

Источники и литература:

1. Laskey, R., Honda, B., Mills, A. et al. Nucleosomes are assembled by an acidic protein which binds histones and transfers them to DNA. Nature 275, 416–420 (1978). https://doi.org/10.1038/275416a0

2. Frehlick, L. J., Eirín-López, J. M., & Ausió, J. (2007). New insights into the nucleophosmin/nucleoplasmin family of nuclear chaperones. BioEssays : news and reviews in molecular, cellular and developmental biology, 29(1), 49–59. https://doi.org/10.1002/bies.20512

3. Karimi Dermani, F., Gholamzadeh Khoei, S., Afshar, S., & Amini, R. (2021). The potential role of nucleophosmin (NPM1) in the development of cancer. J Cell Physiol, 236, 7832–7852. https://doi.org/10.1002/jcp.30406

4. Mitsuru Okuwaki, Ayako Sumi, Miharu Hisaoka, Ai Saotome-Nakamura, Satoko Akashi, Yoshifumi Nishimura, Kyosuke Nagata, Function of homo- and hetero-oligomers of human nucleoplasmin/nucleophosmin family proteins NPM1, NPM2 and NPM3 during sperm chromatin remodeling, Nucleic Acids Research, Volume 40, Issue 11, 1 June 2012, Pages 4861–4878, https://doi.org/10.1093/nar/gks162

5. Matoba, K., Matsumoto, Y., Hongo, T., Nagamatsu, Y., Sugino, H., Shimizu, T., Takao, T., Shimonishi, Y., & Ikegami, S. (2000). Chemical structure of nuclear proteins which are phosphorylated during meiotic maturation of starfish oocytes. Biochemistry, 39(21), 6390–6400. https://doi.org/10.1021/bi992759x

6.10 Биоинформатический анализ роторных АТФ-синтаз

Авторы: Литвин А.В.[1], Фенюк Б.А.[2,3], Гельфанд М.С.[1,4]

1 - Сколковский институт науки и технологий, Москва, Россия, 2 - Факультет биоинженерии и биоинформатики, МГУ им. М.В. Ломоносова, Москва, Россия, 3 - Научно-исследовательский институт физико-химической биологии им. А.Н. Белозерского, Москва, Россия, 4 - Институт проблем передачи информации им. А.А. Харкевича, Москва, Россия

Почти все прокариоты используют мембранный потенциал в качестве энергетического эквивалента и, следовательно, имеют генераторы потенциала и роторные АТФ-синтазы для использования этого потенциала. Роторные АТФ-синтазы производят АТФ из АДФ, за счет энергии мембранного потенциала, или могут функционировать в обратном направлении, как ионные насосы. АТФ-синтазы разнообразны — существует два типа общего структурного строения (F и A), ферменты имеют натриевую или протонную специфичность, различную стехиометрию (соотношение ионов и АТФ в реакции), однако все АТФ-синтазы обладают гомологичным ядром субъединиц. АТФазы F-типа встречаются в основном у бактерий,. ферменты A-типа часто встречаются у архей и иногда присутствуют у бактерий в качестве основной или второй АТФ-синтазы. Натриевые АТФ-синтазы считаются эволюционно более древними по сравнению с протонными, но сейчас натриевая биоэнергетика встречается редко [1].

Целью этой работы является изучение нестандартных моделей биоэнергетики — организмов с натриевыми биоэнергетикой, а также организмов, имеющих несколько роторных АТФ-синтаз.

Распространенность и структура натриевых АТФ-синтаз недостаточно изучена. Планируется провести поиск натриевых АТФаз в геномной базе данных GTDB и изучить их структурные особенности, а также геномные и фенотипические особенности их хозяев. Во-первых, это позволит описать распределение натриевых АТФаз по филумам прокариот. Во-вторых, известно, что натриевые F-АТФазы имеют специфический мотив в одной из мембранных субъединиц, которая связывает ионы во время транспорта (субъединица c в F-типе, субъединица K в A-типе). Однако в транспорте ионов участвует и другая субъединица (субъединица a в F-типе, субъединица I в A-типе), которая создает полуканалы с обеих сторон мембраны для ионов. Наша гипотеза заключается в том, что эта субъединица (a/I) участвует в обеспечении ионной специфичности. Предварительный анализ показал, что натриевые ферменты F-типа имеют такой мотив в субъединице a; ранее это не было известно. Дальнейшее исследование будет включать изучение генераторов натриевого потенциала у этих организмов и определение особенностей этих генераторов. Это позволит нам определить физиологические преимущества и экологические ниши натриевой биоэнергетики.

Некоторые прокариоты имеют несколько роторных АТФ-синтаз в геноме — предварительный анализ показывает, что такие случаи составляют до 10% организмов в базе

данных GTDB. Отдельный уже изученный случай составляют организмы, имеющие вторую АТФазу N-типа — АТФаза N-типа похожа на фермент F-типа, всегда является второй АТФазой геноме, и играет роль ионного насоса. Однако неясно, почему некоторые прокариоты имеют две или более АТФазы F и A-типа, которые обычно являются синтазами. Планируется изучить комбинации типов АТФаз и ионную специфичность, найти корреляции с генераторами потенциала и фенотипическими признаками. Это позволит нам найти преимущества и причины наличия двух или более роторных АТФаз.

Источники и литература:

1. A. Y. Mulkidjanian, M. Y. Galperin, K. S. Makarova, Y. I. Wolf, and E. V. Koonin, 'Evolutionary primacy of sodium bioenergetics', Biology Direct, vol. 3, no. 1, p. 13, 2008, doi: 10.1186/1745-6150-3-13.

6.11 Идентификация протеолитических сайтов на основе известной информации о техмерных структурах потенциальных белковых субстратов

Авторы: Е.В. Матвеев[1], М.Д. Казанов[1,2]

1 - Институт Проблем Передачи Информации им. А.А. Харкевича РАН, Москва, Россия, 2- Сколковский Институт Науки и Технологий, Москва, Россия

Идентификация белковых субстратов протеолитических ферментов крайне важна для выяснения механизмов многих молекулярных процессов, протекающих в живой клетке, как, например, апоптоз, пролиферация клеток, активация или деградация белков. Вычислительные предсказания событий протеолиза могут существенным образом сократить количество экспериментальных затрат, необходимых для идентификации белковых субстратов протеаз. Известно, что наряду со специфичностью протеазы по последовательности, трёхмерная структура субстратов также влияет на возможность протеолитического расщепления определённых участков белка. Однако большинство из существующих биоинформатических методов используют для предсказания сайтов протеолиза исключительно информацию о специфичности протеолитических ферментов по последовательности, а информация о трёхмерной структуре потенциального субстрата практически не используется. Насколько нам известно, на данный момент существует лишь один метод, который использует информацию о 3D структурах потенциальных субстратов протеаз, и при этом нет ни одного метода, который бы предсказывал подверженность определённых участков белка протеолитическому расщеплению только на основе информации о трёхмерной структуре. Чтобы восполнить данный пробел, мы разработали биоинформатический алгоритм для оценки

подверженности участков белков к ограниченному протеолизу, который основывается на информации об известных трёхмерных структурах белков, полученной из базы данных PDB. В основу алгоритма легли выводы о значимости структурных детерминант ограниченного протеолиза, полученные в наших предыдущих исследованиях на основе анализа данных экспериментально зафиксированных протеолитических событий из базы CutDB. Мы дополнили первоначальный датасет информацией о сайтах протеолитических расщеплений из базы данных MEROPS, а также использовали информацию о предсказанных структурах белков из новой базы AlphaFold DB. MEROPS также является базой данных, содержащей экспериментально подтверждённую информацию о протеолитических событиях. Поскольку эта база данных содержит информацию о сайтах протеолиза, полученных не только регуляторными протеазами, нами был тщательно проведён этап курирования обучающих наборов данных, относящихся к базе данных MEROPS. Полученный алгоритм может быть использован в связке с предсказаниями специфичности протеазы по последовательности, получаемых с помощью позиционно-весовых матриц специфичности (PWM). На основе данных эксперимента о расщеплении металлопротеиназой MMP9 белков E.coli, имеющей большое количество разрешенных трехмерных структур белков, нами также разрабатывается оптимальный подход к совмещению двух предсказаний – вероятностей расщепления пептидной связи с точки зрения структурных данных и специфичности протеазы по последовательности.

6.12 Оптимизация планирования эксперимента по анализу аллельно-специфической экспрессии генов

Авторы: Ася Менделевич [1], Андрей Миронов [2], Александр Гимельбрант [3]

1 - Skoltech, 2 - МГУ, НИУ ВШЭ, ИППИ РАН, 3 - Altius Institute for Biomedical Sciences

В предыдущих работах мы показали, что для надежного анализа аллельно-специфической экспрессии необходимо делать технические реплики библиотек для РНК секвенирования. Этот подход позволяет адекватно учитывать значительную техническую вариабельность РНК-сека и дает возможность точной оценки дифференциальной аллельно-специфической экспрессии. Создание параллельных технических реплик для каждого образца достигает цели, но многократно увеличивает стоимость эксперимента. Здесь мы показываем, что эквивалентный результаат можно получить без увеличения числа библиотек путем добавки дополнительной РНК (spike-in) выделенной из другого организма (гетерозиготных мыши или круглого червя C.elegans) в каждый образец. Оценка овердисперсии хорошо переносится с добавочной компоненты на представляющие интерес образцы.

Дополнительным преимуществом данного метода является его экстраполируемость на данные одноклеточного секвенирования (в разработке).

7. РНК-хроматин

7.1 Анализ специфичности РНК-ДНК контактов

Авторы: Анастасия Александровна Жарикова[1], Андрей Александрович Миронов[2]

1 - МГУ им. М.В.Ломоносова, Москва, Россия, 2 - ИППИ Москва, Россия

Значительная часть генома эукариот транскрибируется с образованием большого количества разнообразных РНК, включая мРНК и различные длинные и короткие некодирующие РНК. Молекулы РНК могут выполнять свои функции не только в цитоплазме, но и оставаясь в ядре клетки, где они активно участвуют в процессах регуляции транскрипции, а также ремоделирования и поддержания пространственной структуры хроматина [1]. Классическими примерами таких РНК могут служить XIST, HOTAIR, MALAT1, TERC и другие [2].

На сегодняшний день существует целый спектр разработанных ранее методик, позволяющих выявить локусы ДНК, с которыми взаимодействует одна или несколько заранее известных РНК [3]. За последние 5 лет появилось сразу несколько независимо разработанных методов, с помощью которых в рамках одного эксперимента можно полногеномно определить РНК-ДНК интерактом: MARGI, ChAR-Seq и GRID-Seq, а также метод Red-C, предложенный нами ранее [4], включающий эксперимент и алгоритм анализа, охватывающий все этапы от фильтрации сырых чтений до аннотации отобранного пула РНК-ДНК контактов.

При анализе любых результатов массового секвенирования неизбежно возникает проблема неспецифических данных. Разработка экспериментальных и биоинформатических подходов, позволяющих бороться с такими артефактами, дает возможность получать более качественные и достоверные результаты. В данной работе мы предлагаем несколько вариантов нормировок, созданных специально для анализа данных РНК-ДНК интерактома, которые учитывают уровень экспрессии отдельных РНК, расстояние между местом синтеза РНК и ее контактами, специфические контакты мРНК. Разработанные подходы позволяют снизить уровень “шума” в данных и могут быть применены к любым существующим на сегодняшний день полногеномным экспериментам по изучению РНК-хроматиновых взаимодействий.

Сегодня мы располагаем огромным арсеналом методов с использованием технологий секвенирования нового поколения, позволяющих изучать трехмерную структуру хроматина, экспрессию РНК, определять сайты связывания факторов транскрипции, а также восстанавливать сети взаимодействий нуклеиновых кислот с белками и между собой и многие другие. Большой объем данных по самым разным организмам и клеточным линиям находится в открытом доступе.

Всесторонний анализ большого количества накопленных данных, полученных в результате разнообразных экспериментов, несомненно, позволит более целостно взглянуть на работу клетки, увидеть, как эти данные соотносятся друг с другом, подтверждают друг друга или, наоборот, противоречат друг другу.

Благодарности: бесконечно благодарю студентов и сотрудников группы А.А. Миронова за регулярные плодотворные семинары, обсуждения и помощью в работе, а также Андрея Сигорских и Александру Галицыну за вклад в проект в целом.

Источники и литература:

1. Engreitz, J.M.: Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. Nature Reviews Molecular Cell Biology, 17, 756-770 (2016)

2. Quinn J.J., Chang H.Y.: Unique features of long non-coding RNA biogenesis and function. Nature Reviews Genetics. 2016 Jan;17(1):47-62. doi: 10.1038/nrg.2015.10

3. Engreitz, J.M.: RNA-RNA Interactions Enable Specific Targeting of Noncoding RNAs to Nascent Pre-mRNAs and Chromatin Sites. Cell 159(1), 188-199 (2014).

4. Gavrilov, A.A., Zharikova, A.A, Galitsyna, A.A., Luzhin, A.V., Rubanova, N.M., Golov, A.K., Petrova, N.V., Logacheva, M.D., Kantidze, O.L., Ulianov, S.V., Magnitov, M.D., Mironov, A.A., and Razin, S.V. Studying rna-dna interactome by red-c identifies noncoding rnas associated with various chromatin types and reveals transcription dynamics. Nucleic Acids Research, 2020 Jul 9;48(12):6699-671

7.2 Оценка специфичности РНК в контексте взаимодействий РНК-ДНК-хроматиновых взаимодействий

Авторы: Е.Н. Питиков[1,2] ФББ МГУ, А.А.Миронов[1,2], Д.Д.Пензар[1,2,3]

1 – ФББ, Москва, Россия, 2 - МГУ им Ломоносова, Москва, Россия, 3 - Институт проблем передачи информации им. А. А. Харкевича РАН, Москва, Россия

Изучение строения и функций различных декодирующих хроматин-ассоциированных РНК является важной задачей молекулярной биологии, однако все существующие методы имеют высокий уровень шума в выходных данных. Данная работа предлагает подход, основанный на методиках машинного обучения для решения этой проблемы.

В работе предпринята попытка предсказать специфичность ДНК-РНК контакта, опираясь на его к-мерный состав (использовались к-меры от 1 до 6). Составлены обучающие и тестовые выборки на основе типа РНК, ее хроматинового потенциала и количества близких контактов. На них были обучены модели на основе бэггинга, бустинга и случайного леса.

Модели на основе бэггинга показали максимальную точность и этот подход стал основным в дальнейшей работе.

Обнаружено, что предсказания моделей заметно отличаются в зависимости от выбранного подхода к оценке специфичности РНК при составлении выборок.

Так, при использовании хроматинового потенциала как критерия специфичности белок-кодирующие РНК и днРНК получили схожее число специфичных контактов. Подход, основанный на числе близких контактов давал значительно больший процент специфичных контактов днРНК, что позволяет считать его более точным.

Были проверены самые важные к-меры. Обнаружено, что большая часть из них - ГЦ-богатые. Особенно часто встречались паттерны типа $GG(N)_{2-4}GG$.

Программой Queseq был проведен пикколлинг специфичных контактов. Контакты днРНК в среднем получали более значимые пики, но количество пиков у белок-кодирующих РНК было выше, что можно объяснить соотношением данных типов РНК в клетке.

7.3 Определение статистически значимых РНК-хроматиновых взаимодействий в данных «все-против-всех»

Авторы: Дмитрий Евгеньевич Мыларщиков[1], Андрей Александрович Миронов[1,2]

1 - МГУ им. М.В.Ломоносова, Москва, Россия, 2 - Институт проблем передачи информации им. А. А. Харкевича РАН, Москва, Россия

Хроматин-ассоциированные РНК (хаРНК) участвуют в регуляции транскрипции генов и организации архитектуры ядра. Для детального изучения функциональных ролей хаРНК разработаны молекулярно-биологические методы определения РНК-хроматиновых взаимодействий.

Группа методов «один-против-всех» (RAP, CHART, ChIRP, CHOP) позволяет установить участки ДНК, с которыми взаимодействует одна конкретная РНК, с высоким покрытием. Данные этой группы методов соответствуют модели данных ChIP-seq, поэтому для определения статистически значимых взаимодействий (или «пиков») можно использовать соответствующие программы, например, MACS2.

Группа методов «все-против-всех» (GRID, iMARGI, RADICL, Red-C) позволяет установить РНК-хроматиновые взаимодействия для всех РНК и всей ДНК сразу, однако для каждой РНК покрытие хроматина получается не очень большим. Эти данные не соответствуют модели данных ChIP-seq, поэтому для определения «пиков» необходимы оригинальные алгоритмы. Однако разработанные два алгоритма неточно описывают статистическую природу данных. Так, для данных «все-против-всех» наблюдается скейлинг –

степенное убывание частоты контактов каждой конкретной РНК с участками родной хромосомы при удалении от родного гена (явление, которое наблюдается и для ДНК-ДНК контактов по данным Hi-C). Алгоритм, предложенный для анализа данных GRID, учитывает только эндогенный фоновый сигнал, но не скейлинг. А алгоритм, предложенный для анализа данных RADICL, не учитывает ни скейлинг, ни эндогенный фон.

Мы предлагаем алгоритм, который учитывает и эндогенный фон, и скейлинг. Он совмещает подходы для анализа данных ChIP-seq и Hi-C и вычисляет пики для индивидуальных РНК. Алгоритм состоит из четырёх частей:

1. Бинирование генома. Для каждой РНК подбирается индивидуальный размер бина, в которых будут суммироваться ДНК-части контактов.

2. Выбор эндогенного фона. Для этого выбираются РНК, которые не любят контактировать с неродными хромосомами (то есть, транс), и берутся их транс-контакты. Контакты агрегируются в 1Кб бины и нормируются до вероятностей.

3. Расчёт параметров фоновой модели. Для биномиальной модели N_i – число всех контактов данной РНК i, а p_{ij} – это фоновая вероятность одного контакта РНК i попасть в ДНК-бин j. Для ДНК-бинов на неродной хромосоме p_{ij} берётся из эндогенного фона, а для ДНК-бинов на родной хромосоме p_{ij} есть произведение эндогенного фона и скейлинга. Скейлинг предпосчитан заранее так, чтобы он соответствовал этой процедуре. В конце фоновые вероятности нормируются так, чтобы сумма p_{ij} по j равнялась 1.

4. Вычисление p-value и поправка на множественное тестирование Бенджамини-Хохберга (FDR < 0.05).

Для валидации алгоритма проведено сравнение пиков РНК MALAT1, полученных предложенным алгоритмом из данных GRID, оригинальным алгоритмом для GRID и пиков из данных RAP («один-против-всех»). Предложенный алгоритм определял на 43% больше пиков RAP, чем алгоритм для GRID. Вместе с тем, ложно-положительных пиков тоже было больше.

Планируется доработать процедуру выбора эндогенного фона для улучшения сходимости результатов с данными RAP. После этого планируется функциональный анализ найденных пиков для каждой РНК для каждого доступного эксперимента «все-против-всех».

7.4 Поиск участков, обогащенных РНК-хроматиновыми контактами, в данных “один-против-всех”

Авторы: С.В. Кузнецов[1] , Г.К. Рябых[2,3], А.А. Миронов[2,3]

1 – BostonGene, 2 - МГУ им Ломоносова, Москва, Россия, 3 - Институт проблем передачи информации им. А. А. Харкевича РАН, Москва, Россия

Основной задачей в обработке данных “один-против-всех” является определение пиков – участков генома, обогащенных контактами данной РНК с хроматином. Полученные пики можно связать с регуляторными областями генов и другими геномными областями, тем самым функционально охарактерезовать некодирующую РНК [1]. Данная работа посвящена сравнительному анализу определенных унифицированным подходом пиков во всех доступных данных “один-против-всех” (более 120 датасетов) и опубликованных данных.

Источники и литература

1. Ryabykh G.K., Mylarshchikov D.E., Kuznetsov S.V., Sigorskikh A.I., Ponomareva T.Y., Zharikova A.A. and Mironov A.A., “RNA-chromatin interactome. What? Where? When?”, Molecular Biology, 2022

7.5 Сравнительный анализ РНК-хроматиновых карт контактов клеток

А.М. Васильев[1], Г.К. Рябых[1,2], А.А. Миронов[1,2]

1 - МГУ им Ломоносова, Москва, Россия, 2 - Институт проблем передачи информации им. А. А. Харкевича РАН, Москва, Россия

РНК-хроматиновые взаимодействия влияют на многие процессы в клетке. Экспериментальные подходы, позволяющие исследовать РНК-хроматиновый интерактом, делятся на два класса: «один-против-всех» и «все-против-всех». Первый определяет взаимодействия конкретной РНК, в то время как методы второго класса определяют взаимодействия всех РНК, экспрессирующихся в клетке, со всеми локусами ДНК [1]. Так как для каждой отдельной РНК методы «все-против-всех» выявляют значительно меньше контактов, чем методы типа «один-против всех», был проведен сравнительный анализ карт контактов из обоих классов методов.

Источники и литература

Ryabykh G.K., Mylarshchikov D.E., Kuznetsov S.V., Sigorskikh A.I., Ponomareva T.Y., Zharikova A.A. and Mironov A.A., “RNA-chromatin interactome. What? Where? When?”, Molecular Biology, 2022.

7.7 Поиск дуплексов в РНК-хроматиновых данных

Авторы: И.К. Марков[1], Г.К. Рябых[1,2], А.А. Миронов[1,2]

1 - МГУ им Ломоносова, Москва, Россия, 2 - Институт проблем передачи информации им. А. А. Харкевича РАН, Москва, Россия

На текущий момент существует два класса методов: «один-против-всех» и «все-против-всех», которые позволяют определить РНК-хроматиновые взаимодействия. Характерной и в то же время удивительной чертой данных, полученных методами «все-против-всех», является высокая доля контактов, приходящаяся на мРНК. Во всех работах, и в частности в Red-C [1], считается, что мРНК в основной своей массе не должны специфически контактировать с хроматином, поэтому их контакты использовали в качестве фоновой модели.

Так как РНК может образовывать с ДНК R-петлевую структуру и таким образом регулировать экспрессию генов [2], мы попытались определить, образуют ли мРНК R-петли в РНК-хроматиновых данных? Если да, то действительно ли доля таких РНК среди всех мРНК будет небольшой? Для этого была применена программа, основанная на FASTA алгоритме.

Источники и литература:

1. Alexey A Gavrilov, Anastasiya A Zharikova, Aleksandra A Galitsyna, Artem V Luzhin, Natalia M Rubanova, Arkadiy K Golov, Nadezhda V Petrova, Maria D Logacheva, Omar L Kantidze, Sergey V Ulianov, Mikhail D Magnitov, Andrey A Mironov, Sergey V Razin, Studying RNA–DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics, Nucleic Acids Research, Volume 48, Issue 12, 09 July 2020, Pages 6699–6714

2. Piroon Jenjaroenpun, Thidathip Wongsurawat, Sawannee Sutheeworapong, Vladimir A. Kuznetsov, R-loopDB: a database for R-loop forming sequences (RLFS) and R-loops, Nucleic Acids Research, Volume 45, Issue D1, January 2017, Pages D119–D127

7.8 Анализ вторичных структур хроматин ассоциированных РНК

Авторы: О.Д. Богомаз[1], Г.К. Рябых[1,2], А.А. Миронов[1,2]

1 - МГУ им Ломоносова, Москва, Россия, 2 - Институт проблем передачи информации им. А. А. Харкевича РАН, Москва, Россия

Известно, что некодирующие РНК (нкРНК) вовлечены в регуляцию экспрессии генов. Часть нкРНК обладает стабильной вторичной структурой, однако до сих пор многие нкРНК не изучены на предмет наличия вторичных структур и их функциональной роли [1]. Косвенным признаком, указывающим на функциональность вторичной структуры, является ее консервативность, поэтому целью данной работы является поиск и изучение консервативных вторичных структур хроматин-ассоциированных некодирующих РНК (с помощью программы RNAsurface), а также установление зависимости между структурированностью и степенью контактированности с хроматином того или иного домена нкРНК.

Источники и литература:

1. Zampetaki A., Albrecht A., Steinhofel K. (2018) Long Non-coding RNA Structure and Function: Is There a Link? Front. Physiol. 9:1201

7.9 Сравнительный анализ данных экспериментов Red-C и Hi-C

Авторы: Д.С. Звездин[1], А.А. Жарикова[1], А.А. Миронов[1]

1 - МГУ им Ломоносова, Москва, Россия

Некодирующие РНК, присутствующие в ядре клетки, выполняют много различных функций, в частности, среди них, есть те, которые участвуют в регуляции процессов, происходящих в хроматине, и в организации его структур. Метод Red-C [1] позволяет фиксировать взаимодействия РНК с ДНК. Метод Hi-C [2] позволяет определить число контактов между различными участками хромосом, по которому можно судить об их пространственной близости. Наличие большого количества контактов между участками ДНК, говорит о том, что они могут образовывать структурные элементы хроматина (например, TADы). Если РНК имеет много контактов с участками ДНК, которые образуют структуры хроматина, то она, вероятно, может участвовать в их организации.

В данном проекте проводится анализ РНК, склонных контактировать с хроматином, на наличие повышенного числа контактов со сближенными участками хромосом. Выявление таких РНК проводится с использованием точного теста Фишера. Для контактов каждой РНК с каждой хромосомой производится подсчет числа пар контактов, произошедших со сближенными и с отдаленными локусами хромосом. Вывод о пространственной близости локусов делается на основании Hi-C карты для данной хромосомы. Затем та же процедура проводится на фоновой модели, в качестве которой используется Hi-C карта, полученная сдвигом координат исходной карты на несколько миллионов нуклеотидов. Из полученных значений составляется таблица сопряженности и рассчитывается тест Фишера.

Таким образом были проанализированы 351 РНК, склонные контактировать с хроматином, для 248 были получены статистически значимые результаты для контактов хотя бы с одной из хромосом.

Источники и литература:

1. Alexey A Gavrilov, Anastasiya A Zharikova, Aleksandra A Galitsyna, Artem V Luzhin, Natalia M Rubanova, Arkadiy K Golov, Nadezhda V Petrova, Maria D Logacheva, Omar L Kantidze, Sergey V Ulianov, Mikhail D Magnitov, Andrey A Mironov, Sergey V Razin. Studying RNA–DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics. Nucleic Acids Research, 2020.

2. Erez Lieberman-Aiden 1, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 2009.

7.10 Сравнительный анализ данных РНК-ДНК и РНК-белковых взаимодействий. Данные fRIP-seq

Авторы: Д.А.Хлебников [1], А.А.Миронов[1,2]

1 - МГУ им Ломоносова, Москва, Россия, 2 - Институт проблем передачи информации им. А. А. Харкевича РАН, Москва, Россия

Некодирующие РНК принимают участие во многих процессах в клеточном ядре, включая регуляцию транскрипции генов. Взаимодействие РНК с ДНК чаще всего опосредуется некоторым белком. В настоящее время существует множество лабораторных методов по выделению фракций РНК, взаимодействующих с макромолекулами белков или ДНК [1-5], найдены многие некодирующие РНК, которые достоверно участвуют в регуляторных процессах (MALAT1, NEAT1, XIST) [6]. В данной работе были проанализированы результаты эксперимента fRIP-Seq [4], определяющего РНК-белковые контакты для 24 белков-регуляторов хроматина, и полногеномного метода Red-C [5], определяющего РНК-ДНК контакты.

Данные секвенирования результатов fRIP-Seq для клеток линии K562 были получены из соответствующего репозитория БД NCBI GEO (ID: GSE67963), данные Red-C (также для клеток линии K562) получены в виде таблицы контактов. Картирование данных fRIP-Seq было проведено на геном человека (GRCh38.p13) программой hisat2. Аннотирование данных fRIP-Seq проводилось с помощью пакета программ bedtools на версию разметки человеческого геном Gencode Human Release 37. Подсчёт сигналов контактов РНК-белок или РНК-ДНК экспериментов проведён с помощью написанных в процессе работы программ на языках Python 3 и R. Статистические тесты также проводились в R.

В процессе работы создан программный конвейер на языках bash, Python, R, позволяющий рассчитать сигналы контактов экспериментов, а также найти комплексы опосредованных белком РНК-ДНК взаимодействий. Результаты проведённых тестов предполагают ассоциированность 19 исследуемых белков с РНК, достоверно ассоциированных с хроматином (caRNA). Исследованы пересечения групп caRNA, контактирующих с каждым из белков, возможность кластеризации и коррелированность

caRNA по интенсивности контакта с белками. Дальнейшая работа в данном направлении предполагает валидацию троек взаимодействий РНК-белок-ДНК с помощью данных ChIP-Seq, сравнение полученных по каунтам ридов контактов и полученных при помощью РНК-белкового пик-коллинга, сравнение троек РНК-белок-ДНК для разных клеточных линий и организмов.

Автор выражает благодарность Жариковой А. А. за предоставление результатов эксперимента Red-C.

Источники и литература:

1. Sridhar B., Rivas-Astroza M., Nguyen T.C., Chen W. et al.: Systematic Mapping of RNA-Chromatin Interactions In Vivo. Curr Biol (2017) 27:602-609

2. Li X., Zhou B., Chen L., Gou L.T. et al.: GRID-seq reveals the global RNA-chromatin interactome. Nat Biotechnol (2017) 35:940-950

3. Bonetti A., Agostini F., Suzuki A.M., Hashimoto K. et al.: RADICL-seq identifies general and cell type-specific principles of genome-wide RNA-chromatin interactions. Nat Commun (2020)11:1018-1018

4. G Hendrickson, D., Kelley, D.R., Tenen, D. et al. Widespread RNA binding by chromatin-associated proteins. Genome Biol 17, 28 (2016).

5. Alexey A Gavrilov, Anastasiya A Zharikova, Aleksandra A Galitsyna, Artem V Luzhin, Natalia M Rubanova, Arkadiy K Golov, Nadezhda V Petrova, Maria D Logacheva, Omar L Kantidze, Sergey V Ulianov, Mikhail D Magnitov, Andrey A Mironov, Sergey V Razin, Studying RNA–DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics, Nucleic Acids Research, Volume 48, Issue 12, 09 July 2020,

6. Zhang P, Wu W, Chen Q, Chen M. Non-Coding RNAs and their Integrated Networks. J Integr Bioinform. 2019;16(3):20190027.

7.11 Графовый анализ РНК-хроматиновых взаимодействий

Авторы: С.В. Козюлина[1], А.А.Миронов[1,2], Д.Д.Пензар[1], А.А.Жарикова[1]

1 - МГУ им Ломоносова, Москва, Россия, 2 - Институт проблем передачи информации им. А. А. Харкевича РАН, Москва, Россия

На данный момент большое количество работ приписывают разнообразным некодирующим РНК важную роль в регуляции генов и ядерной организации. Хотя механизмы такой регуляции пока не изучены, известно, что многие подобные РНК взаимодействуют с хроматином. Следовательно, для изучения механизмов регуляции необходимо определить локусы геномных ДНК, с которыми взаимодействуют интересующие РНК.

Методы, позволяющие достичь этой цели, условно делят на два типа. Первые позволяют найти геномные сайты взаимодействий одной заранее известной РНК за один эксперимент (так называемые методы “один-против-многих”), вторые же дают возможность искать сайты взаимодействий всех клеточных РНК (методы “все-против-всех”) [1]. Поскольку в ходе эксперимента “все-против-всех” исследователь получает больший объём информации, чем в ходе единичного эксперимента “один-против-многих”, а также использование стратегии “все-против-всех” не требует обязательного предварительного знания о конкретных РНК для включения их в анализ, совершенствование методов “все-против-всех” является одним из наиболее перспективных направлений развития.

На данный момент существует 6 методов, реализующих стратегию “все-против-всех”, при этом между ними есть существенные различия как в дизайне экспериментальной части, так и в способах биоинформатической обработки. В силу этих различий присутствует большое количество факторов, затрудняющих прямое сравнение эффективности методов.

В настоящий момент группа биоинформатики ФББ МГУ работает над проектом разработки оптимального протокола биоинформатической обработки данных “все-против-всех”, который бы позволял эффективно находить регуляторные взаимодействия РНК-хроматин, а также был бы унифицированным для различных вариаций протоколов “все-против-всех”.

Представляемая работа является частью этого проекта и заключается в проведении графового анализа матрицы контактов, полученной в ходе эксперимента [2]. В ходе выступления будут представлены результаты анализа карт контактов, полученных в ходе эксперимента Red-C [3], по ряду параметров: степень вершины, сумма весов вершины, степень посредничества и центральность собственного вектора.

Источники и литературы:

1. Masaki Kato, et al. Genome-Wide Technologies to Study RNA–Chromatin Interactions. Noncoding RNA. 2020 Jun; 6(2): 20.

2. Georgios A Pavlopoulos, et al. Bipartite graphs in systems biology and medicine: a survey of methods and applications. Gigascience. 2018 Apr 1;7(4):1-31.

3. Alexey A. Gavrilov, et al. Studying RNA-DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics. Nucleic Acids Res. 2020 Jul 9; 48(12): 6699–6714.

7.12 Множественное картирование NGS данных

Авторы: Е.А.Бердникович, А. А.Жарикова, А.А.Миронов[1,2]

1 - МГУ им Ломоносова, Москва, Россия, 2 - Институт проблем передачи информации им. А. А. Харкевича РАН, Москва, Россия

На данный момент, существует большое количество пайплайнов, позволяющих анализировать NGS данные, полученные из разнообразных геномных исследований. К сожалению, большое количество алгоритмов не учитывают риды, картированные больше одного раза. Такое допущение при обработке данных не дает полноценного представления об организации хроматина и не принимает во внимание повторяющиеся последовательности, которые являются частью генома [1,2].

Также уже существуют некоторое количество подходов, позволяющих учитывать множественное картирование, например mHi-C, Kallisto, Salmon, mmquant.

В данном исследовании используется картировщик hisat2 для данных ДНК-РНК интерактома линии клеток K562 человека. Параметры картировщика позволяют выровнять рид не более 1000 раз. На выходе образуются риды, картированные 1 и более раз с не более чем 2 допустимыми ошибками.

Планируется использовать протокол mHi-C [3] для полученных ридов. Суть определения позиции рида в пересчете вероятности из-за местного количества контактов.

Источники и литературы:

1. Sun, J. H., Zhou, L., Emerson, D. J., Phyo, S. A., Titus, K. R., Gong, W., … Phillips-Cremins, J. E. (2018). Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries. Cell. doi:10.1016/j.cell.2018.08.005

2. Cournac, A., Koszul, R., & Mozziconacci, J. (2015). The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. Nucleic Acids Research, 44(1), 245–255. doi:10.1093/nar/gkv1292

3. Zheng Y, Ay F, Keles S.. 2019. Generative modeling of multi-mapping reads with mHi-C advances analysis of Hi-C studies. Elife 8:1–29. doi: 10.7554/eLife.38070.

7.13 RNA-Chrom – не только база данных РНК-хроматиновых взаимодействий, но и аналитический онлайн-инструмент

Авторы: Г.К. Рябых[1,2], А.А.Миронов[1,2]

1 - МГУ им Ломоносова, Москва, Россия, 2 - Институт проблем передачи информации им. А. А. Харкевича РАН, Москва, Россия

Как известно, кодирующие и некодирующие транскрипты могут выполнять свои функции не только в цитоплазме, но и в ядре клетки, где активно участвуют в процессах регуляции транскрипции, а также в ремоделировании и поддержании пространственной структуры хроматина. Среди методов, изучающих механизмы взаимодействия некодирующих

РНК с хроматином, модификаторами хроматина или с другими белками, можно выделить два класса методов: “один-против-всех” (RAP, CHART-seq, ChIRP-seq, dChIRP-seq, ChOP-seq, CHIRT-seq) и “все-против-всех” (MARGI, GRID-seq, ChAR-seq, iMARGI, RADICL-seq, Red-C) [1].

Нами была разработана база данных, которая содержит не только обработанные универсальным пайплайном данные типа “один-против-всех” (более 120 датасетов) и “все-против-всех” (более 20), но и подробные метаданные обо всех экспериментах. Главным ее преимуществом по сравнению с другими базами данных является то, что она является не только хранилищем данных, но и позволяет быстро и легко провести базовый или расширенный анализ хранимых в ней данных (https://rnachrom2.bioinf.fbb.msu.ru).

Источники и литература:

1. Ryabykh G.K., Mylarshchikov D.E., Kuznetsov S.V., Sigorskikh A.I., Ponomareva T.Y., Zharikova A.A. and Mironov A.A., “RNA-chromatin interactome. What? Where? When?”, Molecular Biology, 2022

7.14 Изучение методами биоинформатики сигналов разрывной транскрипции при образовании субгеномных РНК коронавирусов

Авторы: А.А.Черкашина[1], А.В.Алексеевский[1]

1 - МГУ им Ломоносова, Москва, Россия

Геном коронавирусов состоит из единственной РНК положительной полярности. У коронавирусов и, более широко, вирусов порядка Nidovirales, трансляция поздних генов происходит с субгеномных sgRNA. Для каждого позднего гена sgRNA своя. Поздний ген закодирован в (+) РНК вируса на 3'-конце после гена полипротеина. (-) sgRNA синтезируются с (+) РНК вируса в процессе транскрипции, т.е. репликации (+) RNA вируса с пропуском огромного участка (своего для каждого гена). (-) sgRNA реплицируется в sgRNA и снабжается кэпом. Таким образом sgRNA состоит из короткой лидерной цепи РНК, содержащей кэп, первый ген в ней - ген соответствующего позднего белка, на 3’ конце поли-А, закодированная в РНК вируса. В большинстве нидовирусов разрывная транскрипция регулируется сигналами, называемыми TRS. Последовательность CS из 6-и нуклеотидов из TRS в конце лидера совпадает с CS из TRS перед каждым геном позднего белка. Известно, что генно-инженерные мутации CS меняют экспрессию соответствующих поздних белков (S Zuniga et al., 2020).

Цель проекта - изучить природные мутации в последовательностях CS и TRS семейства коронавирусов. Для изучения мутаций TRS необходимо иметь способ определения CS и TRS

в геноме любого коронавируса, поэтому была создана программа (SAE:Search-align-exceptions) для определения CS с использованием только последовательности генома и координат всех генов. Удалось получить правильное определение CS (совпадающее с литературой) по аннотации у ряда вирусов. Далее был проведён анализ 430 377 геномов SARS CoV-2, загруженных в базу данных NCBI на 15.09.21, проведенный моей программой CoronaLab, показал наличие мутаций не менее чем в 0,1 % CS в предполагавшихся ранее консервативных участках, что вызывает особый интерес и предполагает дальнейшее исследование. На рис.1 продемонстрированы последовательности, включающие в себя лидерную CS (ACGAAC в референсе) и содержащие в себе или в TRS мутации. Каждая последовательность входит в группу из нескольких геномов.

В дальнейшем планируется выполнение анализа, аналогичного приведенному в данной работе, на некоторых других, родственных SARS Cov-2, вирусах с последующим соотнесением мутаций с положениями поздних генов в геноме, инициаторными кодонами гена и, при наличии транскриптомных данных - с экспрессией генов.

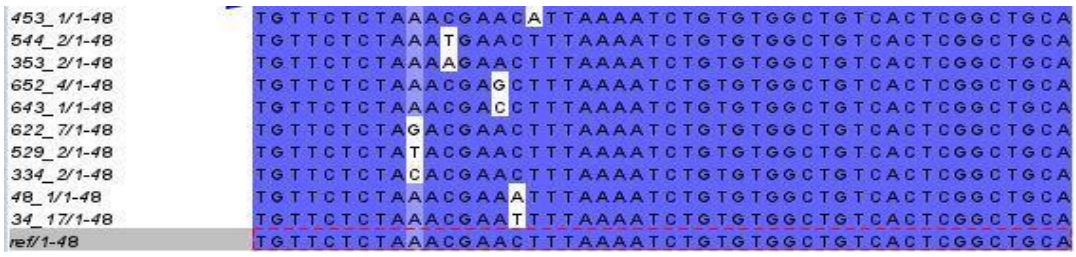

Рис.1. Мутации в лидерной последовательности CS и TRS. Имя посл.: (номер группы)_(её размер)

8. Транскрипция

8.1 Функциональный и филогенетический анализ кассеты генов Escherichia coli, участвующей в деградации сульфохиновозы и лактозы

Авторы: А.АРыбина [1], А.Д.Казнадзей [2], М.Н.Тутукина [1,2,3], М.С. Гельфанд [1,2]

1 - Сколковский институт науки и технологий, Москва, Россия, 2 - Институт проблем передачи информации им. А.А. Харкевича РАН, Москва, Россия, 3 - Институт биофизики клетки РАН, Пущино, Россия

Геном Escherichia coli K12 MG1655 содержит кассету из девяти генов, кодирующих ферменты катаболизма сульфохиновозы (Denger et al., 2014). Ранее в нашей лаборатории с использованием методов сравнительной геномики было выдвинуто предположение, что данная кассета (далее yih-кассета), также принимает участие в деградации лактозы: её состав у E. coli аналогичен с составом кассеты бактерий класса Bacilli, отвечающей за катаболизм этого дисахарида (Kaznadzey et al., 2018). Была обнаружена значительная активация четырех генов yih-кассеты, кодирующих альдолазу, изомеразу, киназу и фактор транскрипции, во время роста культуры на лактозе (Kaznadzey et al., 2018). В данной работе нами была поставлена цель более подробно исследовать возможную роль yih генов в альтернативном пути катаболизма лактозы.

Согласно результатам филогенетического анализа, yih-кассета преимущественно присутствует в геномах бактерий из порядка Enterobacteriales и представлена в виде “короткого” (например, yihTUVW у E. coli Nissle 1917) или “длинного” вариантов (в основном, ompLyihOPQRSTUVW как у E. coli K12 MG1655). На филогенетических белковых деревьях наблюдаются два отдельных кластера в соответствие с тем, в каком варианте кассеты расположен ген, кодирующий соответствующий белок. Белки, закодированные генами из разных типов кассеты, могут отличаться по своей функции и специфичности.

Чтобы сравнить оба типа кассеты, лабораторный штамм E. coli K12 MG1655 и пробиотический штамм E. coli Nissle 1917 были подвергнуты ряду экспериментов: измерению экспрессии yih-генов методом qRT-PCR во время роста культур на глюкозе, лактозе и сульфохиновозе и анализу кривых роста. Кроме того, был проведено РНК-секвенирование обоих штаммов. Согласно полученным результатам qRT-PCR, экспрессия yih-генов длинной кассеты значительно возрастает во время роста культуры на сульфохиновозе, на лактозе наблюдалась менее значительная активация. Нами не было обнаружено значительной разницы в экспрессии yih-генов короткой кассеты при изменении главного источника углерода (сульфохиновоза, лактоза и глюкоза). Данные РНК-секвенирования в целом согласуются с описанными выше результатами. На основании полученных наблюдений мы можем предположить, что длинная кассета необходима для деградации сульфохиновозы и лактозы, а

короткая кассета, возможно, участвует в другом метаболическом пути.

Источники и литература:

1. Denger, K., Weiss, M., Felux, A., Schneider, A., Mayer, C., Spiteller, D., Schleheck, D. (2014). in the biogeochemical sulphur cycle. Nature, 507(7490), 114–117.https://doi.org/10.1038/nature12947

2. Kaznadzey, A., Shelyakin, P., Belousova, E., & Eremina, A. (2018). The genes of the sulphoquinovose catabolism in Escherichia coli are also associated with a previously unknown pathway of lactose degradation. Scientific Reports, (February), 1–12. https://doi.org/10.1038/s41598-018-21534-3

8.2 Странные дела биопленок кишечной палочки

Авторы: Т.А.Бессонова[1], Костарева О.С.[2], Озолинь О.Н.[1], Тутукина М.Н.[1,3,4]

1 - Институт биофизики клетки РАН ФИЦ ПНЦБИ РАН, Пущино, Россия, 2 - Институт белка РАН, Пущино, Россия, 3 - Сколковский Институт науки и технологий, Москва, Россия, 4 - Институт проблем передачи информации РАН, Москва, Россия

Проблема глобальной резистентности бактерий к антибиотикам решается не только путём поиска новых антибиотических веществ, но и путём расшифровки механизмов формирования биопленок на уровне регуляции бактериальной транскрипции. Одной из наших задач является поиск ключевых регуляторных белков, контролирующих процесс образования биопленок кишечной палочкой и веществ, способных предотвратить этот процесс за счет изменения активности регуляторов.

Решение бактерии сформировать ей с другими бактериями биопленку или нет - это сложный и комплексный механизм, включающий в себя так же системы кворум-сенсинга, хемотаксиса и формирования фимбрий и пили. Изучая лабораторный штамм E. coli и его регуляторы транскрипции, мы обнаружили, что некоторые их них могут существенно влиять на способность кишечной палочки формировать биопленки в условиях in vitro. Например, cAMP-CRP, в том числе предсказанный как регулятор формирования биопленок, влияет на их развитие по-разному - в зависимости от аэрации и присутствия в среде глюкозы. Его эффект может быть и опосредованным, потому что CRP является одним из важнейших глобальных регуляторов метаболизма E. coli, в регулон которого входят и гены, кодирующие белки формирования биопленок. UxuR и YjjM - другие, более локальные, регуляторы, которые, согласно нашим транскриптомным данным, должны подавлять флагеллярные гены и активировать образование биопленок. И если для YjjM эти наблюдения полностью подтвердились в физиологических экспериментах, то для UxuR эксперименты показывают

неоднозначные результаты: уменьшение роста биопленок в делеционном мутанте по гену uxuR происходит не всегда и не имеет четкой зависимости от условий роста культуры. Недавно проведенный протеомный анализ удивил нас гипертрофированным синтезом Fli – или Flg – белков в первом и полным отсутствием синтеза оных во втором повторе. Это частично перекликается с транскриптомными данными, в которых в одном эксперименте из трех экспрессии флагеллярных генов не наблюдалось, и раньше считалось нами ошибкой. Сопоставив неоднородность наших результатов по формированию биопленок in vitro с результатами коллег из IST Austria и Медицинского Университета Вены, у которых сильная разница в формировании биопленок наблюдалась даже для разных колоний одного штамма, на данный момент мы предполагаем, что на формирование биопленок E. coli могут влиять не только температура и аэрация, но и популяционный состав колоний, как это иногда происходит в кишечнике хозяина.

Благодарности – Mia Juracic за идею попробовать растить биопленки из отдельных колоний, Michi Lang за обмен результатами и опытом в изучении биопленок. Ольге Бочкаревой за анализ клинических изолятов E. coli.

8.3 Влияние антисмысловой и дивергентной транскрипции на экспрессию генов у Escherichia coli

Авторы: Артемий Дахновец[1,2], Ольга Озолинь[2], Мария Тутукина[2,3,4]

1 - Московский Государственный Университет им М.В. Ломоносова, Москва, Россия, 2 - Институт биофизики клетки РАН (ФИЦ ПНЦБИ РАН), 3- Сколковский институт науки и технологий, Москва, Россия, 4 - Институт проблем передачи информации им. А.А. Харкевича РАН, Москва, Россия

Согласно центральной догме молекулярной биологии реализация генетической информации осуществляется по пути ДНК-РНК-белок. Процессы реализации генетической информации включают в себя транскрипцию и трансляцию. В ходе этих процессов осуществляется синтез молекул различных РНК, некоторые из которых станут матрицей для синтеза белка. Для обеспечения согласованности этапов экспрессии генов существуют разнообразные регуляторные механизмы, позволяющие выключать или активировать работу генов в клетке в ответ на изменившиеся условия окружающей среды. Основными регуляторными агентами в клетках бактерий являются разнообразные белки, среди которых регуляторы транскрипции, белки нуклеоида, факторы промоторной специфичности, а также регуляторы процессивности и терминаторы. Тем не менее существует ещё один класс регуляторов — это малые некодирующие РНК, среди которых особый интерес представляют

антисмысловые РНК. В настоящей работе мы планируем с помощью анализа данных полногеномного секвенирования транскриптомов выявить гены кишечной палочки, экспрессия которых может контролироваться с помощью антисмысловой или дивергентной транскрипции. В качестве модельного организма использовался штамм Escherichia coli K-12 MG1655.

К настоящему моменту мы оценили дифференциальную экспрессию генов кишечной палочки с помощью анализа данных полногеномного секвенирования тотальной РНК, выделенной из клеток Escherichia coli K-12 MG1655, выращенных на глюкозе и D-глюкуронате. Анализ данных заключался в подсчёте выравненных на референсный геном ридов для обеих цепей ДНК с помощью программы FeatureCounts. В результате было обнаружено, что транскрипция в антисмысловой области генов cbl и csrC, кодирующих регулятор транскрипции cys-оперона и малую регуляторную РНК, соответственно, в клетках, выращенных на глюкозе, осуществлялась очень эффективно, тогда как в клетках, растущих на D-глюкуронате ее не наблюдалось. У таких клеток антисмысловая транскрипция наблюдалась у гена yihL, кодирующего предполагаемый регулятор транскрипции. Кроме того, зависимость от источника углерода наблюдалась для антисмысловой транскрипции в генах регуляторов сахарного метаболизма – cAMP-CRP, ExuR, YjjM. Для еще одного регулятора, UxuR, было найдено много малых РНК, в том числе, антисмысловых, секретируемых клеткой во внешнюю среду. Но данных по эффективности антисмысловой транскрипции в присутствии различных источников углерода у нас пока не было. Поэтому, мы подготовили библиотеки для секвенирования тотальной РНК, внутриклеточных малых РНК и секреторных малых РНК свободноживущих клеток кишечной палочки и клеток в составе биоплёнок, выращенных на глюкозе и D-глюкуронате. Мы планируем проанализировать спектр внутриклеточных РНК и РНК, секретируемых кишечной палочкой при росте на различных источниках углерода, а также выявить потенциальные белки-партнеры антисмысловых РНК, секретируемых клеткой, и оценить необходимость формирования дуплексов для функционирования таких РНК.

Исследования поддержаны грантом РНФ 18-14-00348П.

8.4 Консервативность неконсенсусных нуклеотидов в сайтах связывания факторов транскрипции

Авторы: Е. Белоусова[1], М. Гельфанд[1,2,3]

1 - МГУ им. Ломоносова, Москва, Россия, 2 - Центр Наук о Жизни, Сколковский институт науки и технологий, Сколково, Россия, 3 - Институт проблем передачи информации РАН, Москва, Россия

Принято считать, что в сайтах связывания факторов транскрипции консервативны те нуклеотиды, которые критичны для связывания фактора с сайтом, то есть консенсусные. Однако ранее коллеги в работе [1] наблюдали в сайтах транскрипционных факторов в бактериальных геномах, что консервативными могут быть не только консенсусные нуклеотиды. Иногда консервативны некоторые неконсенсусные нуклеотиды или некоторые позиции в сайте с любым из неконсенсусных нуклеотидов. В данной работе мы хотим найти хорошо выраженные консервативные неконсенсусные нуклеотиды и позиции с ними, сравнить их консервативность с нейтрально эволюционирующими элементами и попытаться объяснить это явление.

Потенциально возможных объяснений такому явлению два: во-первых, рассматриваемые сайты могут перекрываться с еще неизвестными сайтами транскрипционных факторов, и наблюдаемые консервативные неконсенсусные нуклеотиды могут быть консенсусными для неизвестного сайта. Во-вторых, неконсенсусные нуклеотиды могут поддерживать конкретный немаксимальный уровень связывания фактора с сайтом (то есть, вес сайта), и, таким образом, являться дополнительным звеном регуляции транскрипции.

Мы рассматривали геномы 47 видов Shewanellaceae для 6 транскрипционных факторов из семейства GntR, и для того, чтобы сравнить консервативность неконсенсусных нуклеотидов в сайтах транскрипционных факторов с нейтрально эволюционирующими элементами, мы вычислили коэффициенты отбора, действующего на позиции с неконсенсусными нуклеотидами и таковые для синонимичных позиции близлежащих генов (аналогично работе [2]). Таким образом мы, действительно, выявили паттерны стабилизирующего отбора, действующего на неконсенсусные нуклеотиды. Затем мы ввели другую меру консервативности (так же, как коллеги в [2]) и также выявили фракции неконсенсусных нуклеотидов, более консервативных, чем нуклеотиды в синонимичных позициях лежащих рядом генов. Мы также показали, что неконсенсусные нуклеотиды в пределах одного сайта эпистатически взаимодействуют таким образом, чтобы сохранять вес сайта перед конкретным геном в некоторых пределах.

Далее в этой работе планируется найти консервативные неконсенсусные нуклеотиды на большем объеме данных и попытаться объяснить это явление с помощью моделирования регуляции транскрипции.

Источники и литература:

1. Kotelnikova, E. A., Makeev, V. J., and Gelfand, M. S. (2005) Evolution of transcription factor DNA binding sites. Gene 347, 255–263.

2. Denisov, S. V., Bazykin, G. A., Sutormin, R., Favorov, A. V., Mironov, A. A., Gelfand, M. S., and Kondrashov, A. S. (2014) Weak negative and positive selection and the drift load at splice sites. Genome Biol. Evol. 6, 1437–1447.

8.5 Коэволюция промоторов и факторов транскрипции

Авторы: Иосиф Финкельберг [1], Михаил Гельфанд [2,3], Михаил Молдован [2]

1 - МГУ им. Ломоносова, Москва, Россия, 2 - Сколковский институт науки и технологий, Сколково, Россия, 3 - Институт проблем передачи информации РАН, Москва, Россия

Самый распространенный способ регуляции транскрипции – это регуляция путем связывания транскрипционного фактора (ТФ) с нуклеотидной последовательностью в промоторной области гена. При этом, один ТФ может функционально связываться как с одной, так и с множеством последовательностей. В последнем случае ТФ называется плейотропным. Для плейотропных ТФ с экспериментально определенными сайтами связывания разработана физико-химическая модель, предсказывающая энергию связывания ТФ с сайтом. Предсказания основываются на частотах встречаемости нуклеотидов в различных позициях сайта 1. Эта модель используется в частности для изучения эволюции сайтов связывания: J.Berg, S.Willmann и M. Lassig предположили, что коэффициент отбора против мутации, приводящей к потере функциональности в сайте, пропорционален вероятности связывания промоторной последовательности с ТФ. Вероятность связывания при этом имеет сигмоидную зависимость от энергии связывания 2. Нелинейность этой зависимости приводит к эпистазу внутри любого поднабора мутаций, что было показано на реальных данных. Кроме того, с помощью симуляций, Майкл Линч объяснил разнообразие последовательностей функциональных сайтов связывания ТФ через большое число разрешенных в любой момент времени нейтральных мутаций, не приводящих к потере функциональности сайта 3. Однако во всех предыдущих исследованиях предполагалась абсолютная консервативность факторов транскрипции и не учитывался вклад коэволюции ТФ с сайтами их связывания. В своем исследовании, мы планируем изучить влияние мутаций в сайт-связывающих последовательностях бактериальных ТФ на эволюцию соответствующих им промоторов. Влияние мутаций на приспособленность будет оцениваться с помощью физико-химической модели, предложенной ранее (J.Berg, S.Willmann и M. Lassig). Также планируется оценить влияние плейотропности фактора, определяемой как число обслуживаемых им промоторов, на его консервативность и на консервативность промоторов. В работе планируется использование множества факторов транскрипции и соответствующих промоторных последовательностей из базы RegPrecise.

Источники и литература:

1. Berg, O. G.; von Hippel, P. H. Selection of DNA Binding Sites by Regulatory Proteins. Journal of Molecular Biology 1987, 193 (4), 723–743. https://doi.org/10.1016/0022-2836(87)90354-8.

2. Berg, Willmann and Lässig. Adaptive Evolution of Transcription Factor Binding Sites. BMC Evol Biol 2004, 4 (1), 42. https://doi.org/10.1186/1471-2148-4-42.

3. Lynch, M.; Hagner, K. Evolutionary Meandering of Intermolecular Interactions along the Drift Barrier. Proc Natl Acad Sci USA 2015, 112 (1), E30–E38. https://doi.org/10.1073/pnas.1421641112.

8.6 Предсказание кодона по нуклеотидному окружению и транслированной аминокислоте

Авторы: Алексей Шевкопляс, Зоя Червонцева, Михаил Сергеевич Гельфанд[1,2]

1 - Сколковский институт науки и технологий, Сколково, Россия, 2 - Институт проблем передачи информации РАН, Москва, Россия

Синонимичные кодоны встречаются в мРНК с разной частотой. Известно, что кодоны, тРНК с антикодонами к которым чаще встречаются в клетке, встречаются в стабильно высокоэкспрессированных генах. Предположительно, на выбор кодона среди синонимичных может влиять и вторичная структура мРНК. Задача состоит в том, чтобы обучить нейросеть по нуклеотидному окружению триплета и аминокислоте, закодированной данным кодоном, предсказывать, какие именно три нуклеотида стоят на этом месте. Отдельно от этого интересно изучить необычные места с редкими или неожиданными кодонами и посмотреть на связь со вторичной структурой мРНК. Используются выравнивания гомологичных генов дрожжей.

8.7 Коэволюция транскрипционных факторов семейства ArsR и их сайтов связывания

И.А. Суворова[1]

1 - Институт проблем передачи информации РАН, Москва, Россия

ArsR – одно из широко распространенных у прокариот семейств регуляторных белков. Большинство из них контролирует экспрессию генов, обеспечивающих гомеостаз ионов тяжелых металлов, однако ряд транскрипционных факторов этого семейства регулирует такие процессы как образование биопленок, вирулентность, образование пигментов и токсинов (Busenlehner et al., 2003, Mac Aogáin et al., 2012). ArsR-регуляторы имеют в своем составе высоко консервативный ДНК-связывающий HTH домен и связываются в виде гомодимеров с 12-2-12 инвертированными повторами (Busenlehner et al., 2003).

В работе были идентифицированы 4814 сайтов связывания для 2408 факторов транскрипции семейства ArsR в 682 прокариотических геномах, доступных в базе MicrobesOnline.

Для полученной выборки факторов транскрипции было построено филогенетическое дерево, выделены ортологические группы. Для каждой ортологической группы были идентифицированы мотивы связывания. Мотивы связывания выявлялись методом филогенетического футпринтинга, для ранее исследованных регуляторов предсказанные мотивы связывания были дополнительно верифицированы с помощью литературных данных, а также данных базы RegPrecise. Построенные с помощью обучающей выборки предсказанных сайтов PWM использовались для полногеномного поиска дополнительных потенциальных сайтов связывания соответствующих регуляторов.

С целью предсказания ДНК-белковых взаимодействий регуляторов семейства ArsR был проведен анализ корреляций аминокислот ДНК-связывающих HTH доменов транскрипционных факторов и нуклеотидов соответствующих сайтов связывания. Корреляционный анализ проводился при помощи программы Prot-DNA-Korr (Korostelev et al., 2016). Корреляции рассчитывались для каждой пары столбцов выравниваний аминокислотных последовательностей HTH доменов регуляторов и нуклеотидных последовательностей сайтов связывания, результат визуализировался в виде карт интенсивности. Мотивы четной и нечетной длины были выравнены при помощи пробелов в центрах сайтов связывания. В связи с симметричной структурой анализируемых мотивов связывания и, следовательно, соответствующих карт интенсивности, корреляции показаны для G/C или A/T пар.

Общая консенсусная последовательность для всех 4814 исследованных сайтов связывания факторов транскрипции семейства ArsR представляет собой A/T-богатый палиндром WYAYWWMRRYA-W1-2-TRYYKWWRTRW (Рисунок 1). При этом были выделены несколько групп мотивов, отличающиеся от общего консенсуса в ряде нуклеотидных позиций:

i) WYWYWTCGWY-W3-4-RKCGAWRYRW (821 фактор транскрипции, 1746 сайтов связывания),

ii) WYAYWTMAAN -W4-NTTKAWRTRW (330 факторов транскрипции, 735 сайтов связывания),

iii) WYWNKYAAMY-W3-4-RKTTRMNWRW (448 факторов транскрипции, 778 сайтов связывания),

iv) WYAYWTGMRYA-W2-TRYKCAWRTRW (371 фактор транскрипции, 627 сайтов связывания),

v) WYANWTWMGN -W3-4-NCKWAWNTRW (383 факторов транскрипции, 754 сайтов связывания),

vi) WYAWWWTAT-W6-ATATWWTRW (55 факторов транскрипции, 174 сайтов связывания).

Несмотря на отсутствие однозначной и четкой корреляции между типом мотива и расположением соответствующих регуляторов на филогенетическом дереве факторов транскрипции ArsR, наблюдается заметная тенденция к кластеризации: большая часть регуляторов, имеющих мотивы типа iii) и iv) располагаются на отдельных ветвях, факторы транскрипции с мотивами типа i) и ii) преимущественно совместно кластеризованы.

Была показана также определенная функциональная специализация. К группе i) относится значительная часть регуляторов устойчивости к мышьяку, факторов транскрипции, регулирующих споруляцию, а также ряд факторов транскрипции, участвующих в ответе на окислительный стресс, обеспечивающих устойчивость к антибиотикам и ксенобиотикам (функции предсказаны на основании геномного контекста, или же описаны ранее). Группа ii) также включает ряд регуляторов устойчивости к мышьяку, и подавляющее большинство регуляторов метаболизма метионина. В группу iii) входит большая часть регуляторов устойчивости к кадмию, устойчивости к тепловому шоку, а также регуляторы биосинтеза железо-серных кластеров. Группа iv) включает большую часть регуляторов устойчивости к цинку, кобальту и никелю. Группа v) содержит ряд факторов транскрипции, контролирующих устойчивость к мышьяку и кадмию, а также нетипичный для семейства ArsR регулятор метаболизма арабинозы. Группа vi, небольшая и четко кластеризованная на филогенетическом дереве, включает регуляторы образования биопленок и пигментов.

Были показаны корреляции для пар нуклеотидов в позициях 13/29 сайтов и аминокислотных остатков в позиции 21 HTH домена, при этом, в соответствии со статистически значимыми предпочтениями, Gln, Leu и Lys, вероятно, участвуют в специфических взаимодействиях с парой A/T, тогда как Arg и Glu – с парой G/C.

Для аминокислотных остатков в позиции 22 была найдена корреляция с нуклеотидами в позициях 15/27, показано предпочтение пары A/T для Ala, Cys, Gln и Pro, а также пары G/C для Asn и Ser.

Для аминокислотных остатков в позиции 23 была показана корреляция с нуклеотидами в позициях 16/26, где Ala и Lys статистически значимо предпочитают A/T, а Gln, Glu и Thr значимо коррелируют с парой G/C.

Корреляции также найдены для пар нуклеотидов 17/25 и аминокислотных остатков в позиции 26, для His, Phe и Tyr выявлено статистически значимое предпочтение пары A/T, для Arg, Gln и Lys в этой позиции домена предпочтительна G/C пара.

Следует отметить, что позиции аминокислот, для которых выявлены корреляции, хорошо соответствуют позициям аминокислот, участвующих в формировании контактов с ДНК (по данным мутагенеза или структурным данным для ДНК-белковых комплексов) для регуляторов NolR (Lee et al., 2014), HlyU (Liu et al., 2011), CzrA (Arunkumar et al., 2009). Таким

образом, данные анализа корреляций хорошо согласуются с ранее полученными данными о специфичности связывания регуляторов семейства ArsR (Таблица 1).

Кроме того, все аминокислоты, для которых показаны корреляции, расположены на распознающей α-спирали HTH домена, что также согласуется с ранее полученными результатами анализа корреляций для других семейств регуляторов HTH-типа, где было показано, что большая часть предсказанных ДНК-белковых контактов формируется аминокислотами распознающей α-спирали (Korostelev et al., 2016, Suvorova and Gelfand, 2019, 2021).

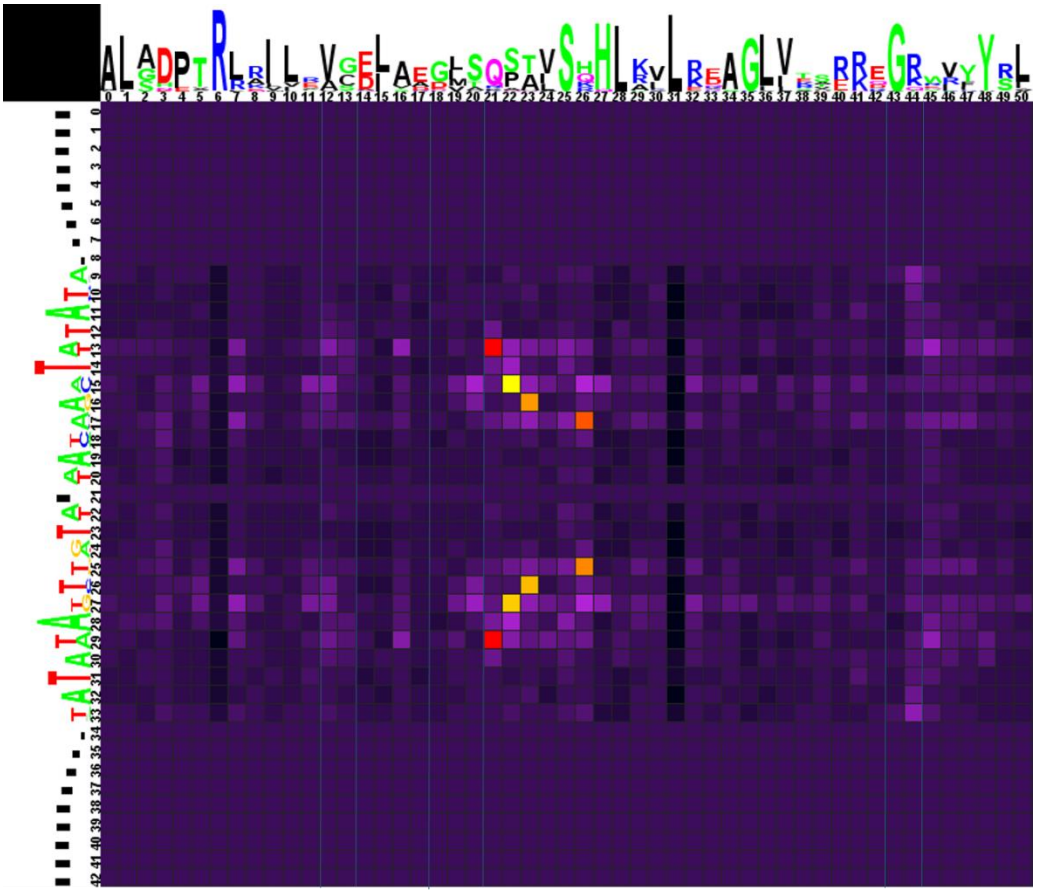

Рис 1. Карта интенсивности корреляций аминокислот HTH-доменов транскрипционных факторов семейства ArsR и нуклеотидов соответствующих сайтов связывания

Диаграммы Logo ДНК-связывающих HTH-доменов и сайтов связывания показаны, соответственно, сверху и слева от карты интенсивности. Общая высота символов в каждой позиции соответствует информационному содержанию, тогда как высота конкретного символа пропорциональна частоте встречаемости аминокислоты/нуклеотида в данной позиции. Уровень корреляции показан цветом и изменяется по градиенту от желтого до красного для статистически значимо (выше автоматически определяемого порога) коррелирующих пар аминокислот и нуклеотидов; прочие пары показаны фиолетово-черным. Голубые линии показывают места, в которых из множественного выравнивания доменов были вырезаны пробелы и плохо выравненные участки.

Таблица 1

Результаты анализа корреляций аминокислот HTH-доменов транскрипционных факторов семейства ArsR и нуклеотидов сайтов связывания

Позиция		Предпочтение	
Аминокислота	Нуклеотид	A/T пара	G/C пара
21*	13/29	Gln, Leu, Lys	Arg, Glu
22*	15/27	Ala, Cys, Gln, Pro	Asn, Ser
23	16/26	Ala, Lys	Gln, Glu, Thr
26*	17/25	His, Phe, Tyr	Arg, Gln, Lys

* - позиции, для которых показано участие в формировании контактов с ДНК у ряда транскрипционных факторов семейства ArsR

Источники и литература:

1. LS Busenlehner, MA Pennella, DP Giedroc. The SmtB/ArsR family of metalloregulatory transcriptional repressors: Structural insights into prokaryotic metal resistance. FEMS Microbiol Rev. 2003 Jun;27(2-3):131-43. doi: 10.1016/S0168-6445(03)00054-8.

2. M Mac Aogáin, M J Mooij, RR McCarthy, E Plower, YP Wang, ZX Tian, A Dobson, J Morrissey, C Adams, F O'Gara. The non-classical ArsR-family repressor PyeR (PA4354) modulates biofilm formation in Pseudomonas aeruginosa. Microbiology (Reading). 2012 Oct;158(Pt 10):2598-2609. doi: 10.1099/mic.0.058636-0. Epub 2012 Jul 19.

3. YD Korostelev, IA Zharov, AA Mironov, AB Rakhmaininova, MS Gelfand (2016). Identification of Position-Specific Correlations between DNA-Binding Domains and Their Binding Sites. Applic. MerR Family Transc. Fact. 11:e0162681. 10.1371/journal.pone.0162681

4. SG Lee, HB Krishnan, JM Jez. Structural basis for regulation of rhizobial nodulation and symbiosis gene expression by the regulatory protein NolR. Proc Natl Acad Sci U S A. 2014 Apr 29;111(17):6509-14. doi: 10.1073/pnas.1402243111. Epub 2014 Apr 14.

5. M Liu, M Rose, JH Crosa. Homodimerization and binding of specific domains to the target DNA are essential requirements for HlyU to regulate expression of the virulence gene rtxA1, encoding the repeat-in-toxin protein in the human pathogen Vibrio vulnificus. J Bacteriol. 2011 Dec;193(24):6895-901. doi: 10.1128/JB.05950-11. Epub 2011 Oct 21.

6. AI Arunkumar, GC Campanello, DP Giedroc. Solution structure of a paradigm ArsR family zinc sensor in the DNA-bound state. Proc Natl Acad Sci U S A. 2009 Oct 27;106(43):18177-82. doi: 10.1073/pnas.0905558106. Epub 2009 Oct 12.

7. IA Suvorova, MS Gelfand (2019). Comparative Genomic Analysis of the Regulation of Aromatic Metabolism in Betaproteobacteria. Front. Microbiol. 10:642. 10.3389/fmicb.2019.00642

8. IA Suvorova, MS Gelfand. Comparative Analysis of the IclR-Family of Bacterial Transcription Factors and Their DNA-Binding Motifs: Structure, Positioning, Co-Evolution, Regulon Content. Front Microbiol. 2021 Jun 10;12:675815. doi: 10.3389/fmicb.2021.675815. eCollection 2021.

8.8 Reconstruction of sugar metabolism regulons in Thermococci

Authors: Natalia V. Sernova, Dmitry A. Rodionov, Institute for Information Transmission Problems, Moscow, Russia

Pyrococcus furiosus is an extremely thermophilic, strictly anaerobic, sugar-utilizing archaeon microorganism that may grow up to 103°C on a variety of sugars including starch, laminarin, maltose, trehalose, cellobiose and beta-glucan oligosaccharides, but does not succeed on crystalline cellulose, xylan or monosaccharides. In this study, we applied a subsystems-based approach combining comparative genomics, metabolic recosntruction and genome-scale transcription factor binding site prediction to reconstruct regulatory networks for carbohydrate metabolism in P. furiosus and related Thermococcales. The complete genomes of 30 Pyrococcus and Thermococcus species were selected for ortholog mapping and comparative analysis. The functional gene assignments, genome context analysis, comparative analysis of orthologous genes and DNA upstream regions, gene co-occurrence analysis and protein similarity searches were performed in the SEED environment. We also used genome annotations from Swiss-Prot, KEGG, and TCDB databases, and published experimental data on gene function and expression profiles. Members of the TrmB family act as global transcriptional regulators for the activation or repression of sugar ABC transporters and central sugar metabolism in Thermococcales species. We identified DNA-binding motifs and reconstructed regulons for several orthologous groups from the TrmB family in Thermococcales. TrmBL1 is a global regulator of the carbohydrate metabolism that co-regulates large sets of genes involved in the starch/maltose utilization, glycolysis and gluconeogenesis. TrmBL1 negatively control the majority of its targets involved in glycolytic pathways but activates a few genes involved in gluconeogenesis. The recosntructed TrmBL1 regulon is highly conserved across analyzed Thermococcales genomes, however we also found a few species-specific targets. Several other members of the TrmB family with identified DNA motifs and recosntructed regulons are local regulators of specific sugar utilization gene loci. We further identified potential new DNA motifs for cellobiose and b-glucans utilization genes, these genes are known from early experimental work and expression data to be upragulated on cellobiose. However, specific regulator associated with this motif remains to be identified. The published whole-genome gene expression and transcriptional start-sites obtained for P. furiosus grown on different carbon sources (glucose, maltose, cellobiose) were used for validation of reconstructed transcriptional regulons.

9. Метагеном

9.1 Применение геномной реконструкции для анализа продукции короткоцепочечных жирных кислот микробиотой кишечника человека

Авторы: Мария С. Фролова [1], Станислав Н. Яблоков [2], Дмитрий А. Родионов[2]

1 - Институт биофизики клетки РАН - обособленное подразделение ФИЦ «Пущинский научный центр биологических исследований РАН», Пущино, Россия, 2 - Институт проблем передачи информации им. А.А. Харкевича РАН, Москва, Россия

Короткоцепочечные жирные кислоты (КЖК), включая ацетат, формиат, пропионат и бутират, являются конечными продуктами ферментации пищевых волокон и гликанов микробиотой кишечника человека. КЖК, произведенные в кишечнике, имеют огромное значение для физиологии и здоровья хозяина, играют ключевую роль в нейроэндокринной и иммунной системах. Прогнозирование метаболического потенциала кишечной микробиоты важно для понимания влияния диеты и метаболитов на здоровье человека. Мы провели детальную метаболическую реконструкцию путей синтеза КЖК и лактата в 2856 бактериальных геномах, представляющих штаммы более 800 известных видов кишечных бактерий. Методами сравнительной геномики был проведен поиск генов для 48 метаболических ферментов из реконструированных путей синтеза КЖК. по крайней мере одним из четырех возможных путей синтеза бутирата. Реконструированные пути синтеза бутирата и пропионата были обнаружены у 359 и 826 бактерий, соответственно, тогда как пути синтеза ацетата, формиата и лактата были обнаружены у большинства изученных бактерий. Мы классифицировали изученные геномы в соответствии с их упрощенными бинарными фенотипами, кодирующими способность («1») или неспособность («0») данного организма продуцировать каждый тип КЖК. Полученные бинарные фенотипы, объединенные в матрицу бинарных фенотипов, использовались для оценки метаболических потенциалов синтеза КЖК в метагеномных образцах кишечной микробиоты человека (в формате 16S и WGS). В итоге рассчитывался кумулятивный индекс каждого фенотипа сообщества (CPI,%) как сумма произведения бинарных фенотипов на относительную численность для каждого вида бактерий (или для каждой операционной таксономической единицы из 16S образца).

Данный подход был использован для оценки способности бактериальных сообществ продуцировать КЖК для трех больших наборов 16S данных микробиома кишечника человека из American Gut Project (AGP), из проекта британских близнецов UK twins (UKT) и набора данных кишечной микробиоты сообщества охотников-собирателей из Танзании (Hadza). Анализ датасетов AGP и UKT показал значительное сходство между количеством и составом продуцентов КЖК (высокое количество продуцентов ацетата и формиата и низкое - для

продуцентов бутирата и лактата). Интересно, что по сравнению с европейскими и американскими микробиомами кишечные микробиомы африканских охотников-собирателей имеют пониженное количество бутиратных, пропионатных, D- и L-лактатных продуцентов. Для того, чтобы продемонстрировать потенциал фенотипического профилирования КЖК для диагностики заболеваний, мы расчитали CPI-индексы на WGS метагеномных данных при воспалительных заболеваниях кишечника, количество бутиратных продуцентов при болезни Крона был достоверно ниже, чем в здоровой группе. Также, мы применили наш подход для оценки CPI-индексов КЖК для данных когорты детей раннего возраста с высоким риском развития сахарного диабета 1 типа. Кроме того, мы связали предсказанные метаболические возможности КЖК с опубликованными концентрациями КЖК как для фекальных образцов, так и для образцов ферментации in vitro из предыдущих исследований. Полученные значения CPI показали отсутствие или низкую корреляцию с экспериментально измеренными концентрациями бутирата и пропионата в образцах фекалий в исследованиях in vivo, что можно объяснить эффективным всасыванием КЖК в толстом кишечнике. Напротив, значения CPI хорошо коррелируют с измеренной концентрацией SCFAs, полученной в экспериментах по ферментации фекального инокулята in vitro. Наконец, мы проанализировали дифференциальное представление отдельных генов пути КЖК в нескольких наборах WGS метагеномных данных. Полученная коллекция генов и путей фенотипов КЖК позволяет прогнозировать метаболические профили в метагеномных наборах данных и улучшать методологию in silico для изучения кишечных микробиомов.

9.2 Genomics-based reconstruction of aromatic amino acid degradation pathways in the human gut microbiome

Authors: German A. Ashniev[1], Dmitry A. Rodionov[1,2]

1 - Institute for Information Transmission Problems, Moscow, Russia, 2- Sanford-Burnham-Prebys Medical Discovery Institute, La Jolla, CA, USA

Human gut microbiome (HGM) represents a set of bacterial commensals living in lumen of human colon. HGM consists of a vast spectrum of diversified microorganisms that participate in fermentation of dietary fibers and proteins, synthesis of vitamins, amino acids and other metabolic products. Primary and secondary metabolites mediate cross-talk between HGM and the immune and nervous systems of the host. Genome-scale reconstruction of microbial biochemical pathways for degradation and transformation of dietary and host-derived metabolites is a critical task for our understanding of HGM function. The subsystem approach to genome annotation and metabolic

reconstruction techniques allows one to map known enzymes to biochemical pathways in model species, propagate the experimentally-confirmed functional annotations to other genomes, and identify candidate for missing genes for pathway gaps. We present a bioinformatic analysis of amino acid metabolism in the reference set of 2,856 genomes representing over 800 known species HGM bacteria. Previously we applied this approach for reconstruction of biosynthetic pathways for proteinogenic amino acids in the reference HGM genomes and predictive profiling of amino acid synthesis phenotypes in HGM metagenomic samples. Here, we continued this effort to reconstruct diverse biochemical pathways for tryptophan, phenylalanine and tyrosine utilization in the reference HGM genomes. We described 10 distinct pathways for degradation of aromatic amino acids producing 24 metabolites many of which are absorbed by the host and participate in gut-brain axis. The reconstructed degradation pathways usually involve either a single biochemical step, or 2-3 consequent enzymes and were found in ~11% of HGM genomes including the Fusobacterium, Clostridium, Escherichia, Lachnospiraceae and Peptostreptococcus genera.

9.3 Влияние керосина на микробиомы различных почв

Авторы: Павел Шелякин[1,2], Иван Семенков[3], Мария Тутукина[1,4,5], Дарья Николаева[1,4], Анна Шарапова[3], Юлия Сарана[4], Сергей Леднев[3], Михаил Гельфанд[1,4], Павел Кречетов[3], Татьяна Королёва[3]

1 - Институт проблем передачи информации им. А.А.Харкевича РАН, Москва, Россия, 2 - Институт общей генетики им. Н.И.Вавилова РАН, Москва, Россия, 3 - Географический факультет МГУ им. М.В.Ломоносова, Москва, Россия, 4 - Сколковский институт науки и технологий (Сколковский Институт Науки и Технологий, Москва), 5 - Институт биофизики клетки РАН (ФИЦ ПНЦБИ РАН)

Углеводороды и в частности керосин, служащие топливом для большинства транспортных средств, являются одними из приоритетных загрязнителей окружающей среды. При этом ответ микробных сообществ на загрязнение почв керосином изучен мало. В настоящей работе мы моделировали влияние утечки керосина на состав почвенного микробиома в лабораторном и полевом экспериментах. В лабораторном эксперименте предварительно отобранные образцы дерново-подзолистой и песчаной пустынной почвы находились в регулярно проветриваемых банках в контролируемых условиях. Полевой эксперимент проводили на площадках с дерново-подзолистой и торфяной болотной верховой почвами, находящимися под естественной растительностью в пределах Сатинской учебно-научной базы географического факультета МГУ. В обоих экспериментах в почву вносили 5 нагрузок керосина (1, 5, 10, 25, 100 г/кг почвы + контроль). В отобранных через 3, 90, 180 и

360 дней образцах измеряли концентрацию керосина и физико-химические свойства почвы, а также оценивали состав бактериального сообщества, секвенируя две пары вариабельных участков V3-V4 и V4-V5 16S рибосомальной РНК.

В незагрязненной песчаной пустынной почве с щелочной средой, низким содержанием доступного фосфора и органического вещества преобладали Actinobacteriota, Firmicutes, Nitrospirota, Planctomycetota и в меньшей степени – Acidobacteriota и Verrucomicobacteriota. Напротив, в сильно кислой болотной почве, богатой органическим веществом и доступным фосфором, Acidobacteriota являлись одним из доминантных таксонов, а Actinobacteriota – минорным. Дерново-подзолистая почва занимала промежуточное положение как по физико-химическим свойствам, так и по составу микробиома.

При неограниченных аэрации, дренаже, миграции бактерий и веществ в условиях полевого эксперимента почвы очищалась от керосина быстрее, чем в лабораторном эксперименте. Через год после загрязнения во всех исследованных почвах следы керосина (менее 1,4 г/кг) детектированы только в образцах с самой высокой изначальной нагрузкой.

Микробиомы всех изученных почв схожим образом реагировали на одинаковые нагрузки керосина и изменялись тем сильнее, чем больше было загрязнение. В сильно загрязненных почвах бактериальные сообщества становились менее разнообразными, увеличивалась доля Proteobacteria и уменьшалась доля Acidobacteriota и Actinobacteriota. При этом росла совокупная доля анаэробных бактерий, метаболизирующих углеводороды. В сильно загрязнённых образцах болотной почвы уже через 6 месяцев после добавления керосина разнообразие микробного сообщества восстановилось, а таксономический состав стал более похож на сообщество в контрольных образцах, в то время как микробиомы дерново-подзолистой и песчаной пустынной почвы не вернулись в исходное состояние даже через год, несмотря на то, что почва уже не содержала керосин.

Благодарности: Работа выполнена в рамках грантов РФФИ №19-29-05206 (анализ данных) и №18-29-13011 (обработка метагеномных данных). Секвенирование осуществлено в Центре геномики Сколковского института науки и технологий.

9.4 Микробиологические индикаторы стадий постагрогенного восстановления почв сосняков

Авторы: О.В.Шопина [1,2], И.Н. Семенков [1,2]

1 - Центр по проблемам экологии и продуктивности лесов РАН, Москва, Россия, 2 - МГУ им. М.В.Ломоносова, Москва, Россия,

Следы былой распашки прослеживаются во многих даже условно коренных лесных

сообществах на протяжении длительного времени.

Для оценки изменений микробного сообщества почв в ходе постагрогенных сукцессий сосновых лесов в пределах национального парка «Смоленское Поозерье» изучено 6 стадий восстановления лесов на 18 ключевых участках (тройная повторность). Стадия С0 представлена распахиваемыми в настоящий момент полями. С1 соответствует молодым залежным лугам. Следующие стадии соответствуют разновозрастным сосновым лесам: С2 – очень молодому (возраст древостоя менее 30 лет), С3 – молодому (пока не исследован нами), С4 – среневозрастному (50 – 70 лет). Леса стадий С5 и С6 имели близкий возраст (70 – 120 лет), но отличались по степени восстановления почвенных горизонтов после распашки.

Из верхних 5 мм гумусового горизонта почв отбирали образцы для дальнейшего выделения ДНК с помощью набора DNeasy PowerSoil (Qiagen) и анализа вариабельного участка V3-V4 гена 16S рибосомной РНК.

Микробные сообщества почв средне- и старовозрастных сосняков (индекс Шеннона на уровне родов был 3,6-4,1) более однородны по сравнению с обнаруженными в почвах лугов (4,2-4,7) и молодых лесов (4,3-4,6). На уровне родов самыми распространенными были Acidobacteriae Subgroup 2, Tepidisphaerales WD2101 soil group и Candidatus Udaeobacter.

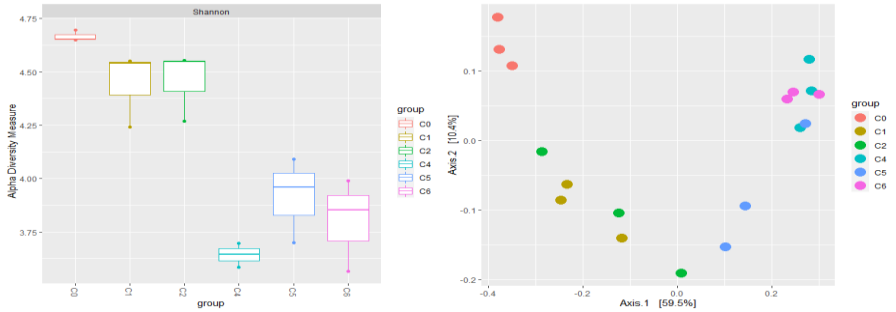

Рис. 1. Оценки изменений микробного сообщества почв

По бета-разнообразию, оцененному с помощью расстояния Брея-Кертиса и методу главных координат (PCoA), выявлено три группы образцов, соответствующих культивируемым в настоящий момент почвам (С0), почвам залежных лугов и очень молодых сосняков (С1 и С2), а также средне- и старовозрастных лесов (С4, С5 и С6).

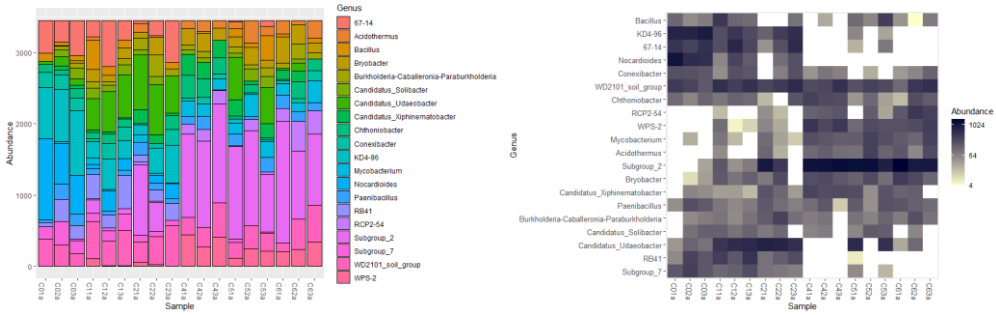

Рис. 2. Оценки изменений

Из 20-ти самых распространенных родов индикационными для луговых стадий и стадии молодого леса были Chloroflexi KD4-96, Solirubrobacterales 67-14, Nocardioides, Pyrinomonadaceae RB41 и Holophagae Subgroup 7. Среди 20-ти рассмотренных родов не

удалось выделить индикационные для стадий средне- и старовозрастных лесов, так как они в целом отличались меньшим разнообразием почвенного микроценоза.

Благодарности. Исследование выполнено в рамках проекта РНФ № 21-74-20171 совместно с центром геномики Сколковского института науки и технологий.

9.5 Выявление растений-гипераккумуляторов тяжелых металлов и металлоидов

Авторы: Г.В.Клинк [1], И.Н. Семенков [2]

1 - Институт проблем передачи информации им. А.А.Харкевича РАН, Москва, Россия, 2 - МГУ им. М.В.Ломоносова, Москва, Россия

Человечество производит всё больше отходов и загрязняет окружающую среду разнообразными поллютантами, среди которых приоритетное значение имеют тяжелые металлы и металлоиды (ТММ). Важным аспектом реабилитации загрязненных территорий является поиск наиболее оптимальных вариантов по соотношению время- и трудозатрат. Одним из динамично развивающихся направлений ремедиации загрязненных территорий является использование биологических объектов – активированных углей и растений-гипераккумуляторов.

До сих пор применяется экстенсивный путь поиска растений-гипераккумуляторов, основанный на сопряженном анализе химического состава почв и растений, произрастающих на территориях с повышенным геохимическим фоном: месторождения, техногенные геохимические аномалии вокруг разнообразных объектов. Выявленные таким образом растения-гипераккумуляторы толерантны к высоким содержаниям ТММ за счет биохимических адаптаций, способствующих переводу повышенных количеств токсичных для обычных растений веществ в менее токсичные соединения. При этом в растениях-гипераккумуляторах не наблюдаются какие бы то ни было симптомы микроэлементозов (некрозы, пониженная фитомасса, угнетенность и проч) – заболеваний, связанных с повышенным содержанием в субстрате и в них самих.

Мы планируем разработать эффективный подход к поиску растений-гипераккумуляторов с помощью методов статистического анализа больших объемов данных, эволюционного анализа, машинного обучения и биогеохимии ландшафтов, базируясь на имеющихся в литературе данных о таксономии и свойствах семенных растений и уровнях содержания в них ТММ.

Благодарности. Исследование выполнено при поддержке Междисциплинарной научно-образовательной школы Московского государственного университета имени М.В.Ломоносова «Будущее планеты и глобальные изменения окружающей среды».

9.6 Метагеномный анализ ксилобионтных грибов и бактерий в валежных стволах лиственных и хвойных деревьев разной степени деструкции после массового ветровала в полидоминантном широколиственном лесу

Авторы: Павел Шелякин[1,2], Мария Тутукина[1,3,4], Анна Рыбина[3], Сергей Волобуев[5], Максим Бобровский[6], Лариса Ханина[7]

1 - Институт проблем передачи информации им. А.А.Харкевича РАН, Москва, Россия, 2 - Институт общей генетики им. Н.И.Вавилова РАН, Москва, Россия, 3 - Сколковский институт науки и технологий, Москва, Россия, 4 - Институт биофизики клетки РАН (ФИЦ ПНЦБИ РАН, 5 - Ботанический институт им. В.Л.Комарова РАН, 6 - Институт физико-химических и биологических проблем почвоведения РАН, 7 - Институт математических проблем биологии РАН - филиал Института прикладной математики им. М.В.Келдыша РАН

Валежная или мёртвая древесина играет важную роль в круговороте питательных веществ, так как является хранилищем запаса углерода и ряда макро- и микроэлементов, которые становятся вновь доступными для биоты в результате разложения древесины. За разложение древесины в первую очередь отвечает сложное сообщество ксилобионтных грибов и бактерий. Структура данного сообщества зависит от множества факторов, таких как тип древесины или климатические условия, и может различаться на разных стадиях разложения. В настоящей работе мы планируем описать структуру и свойства бактериального и грибного сообществ в валежной древесине 8 древесных пород (Quercus robur, Fraxinus excelsior, Tilia cordata, Ulmus glabra, Acer platanoides, Populus tremula, Betula pendula и Picea abies) на пяти стадиях разложения. Отбор проб будет проводиться в заповеднике “Калужские засеки”. Леса заповедника являются уникальными по своему многовидовому составу, возрасту деревьев и относительно большой территории; они являются "референсными территориями" для охраны лесного биоразнообразия и экосистемного управления лесами в подзоне умеренных лесов.

В рамках настоящего проекта будет проведён ДНК-баркодинг макромицетов, а именно будут секвенированы последовательности внутренних транскрибируемых спейсеров (ITS1 и ITS2), прилегающих к генам рибосомальной РНК, у грибов, отобранных непосредственно с разлагающихся деревьев в заповеднике и о которых отсутствует информация в референсных базах данных. Это позволит создать набор эталонных последовательностей для дальнейшего анализа.

Состав бактериального и грибного сообществ будет оценён с помощью секвенирования ампликона вариабильного региона V4 гена 16S рибосомальной РНК и ампликона ITS2 соответственно. На основе полученных данных будет показано, как структура сообществ зависит от породы дерева и как она меняется в процессе разложения. Кроме того, будут оценены метаболические возможности сообществ и выделены группы бактерий и грибов,

демонстрирующие одинаковые, либо противоположные паттерны встречаемости в образцах.

На данный момент проведено секвенирование пилотного набора из двух образцов, обработанных двумя наборами для выделения ДНК, и на основе полученных данных скорректирована методика для дальнейшего анализа. Полученные сообщества содержат представителей с известной способностью к разложению древесины. Видна смена грибного сообщества, состоящего в основном из представителей Ascomycota на первой стадии разложения, на сообщество с доминированием представителей Basidiomycota на третьей стадии. Также наблюдается рост разнообразия бактериального сообщества, при этом уменьшается относительное количество представителей семейства Acidobacteriaceae и класса Bacili и растёт количество представителей классов Actinobacteria и Thermoleophilia.

Работа поддержана проектом РНФ 22-24-01063.

9.7 Различия в структуре пангенома бактерий-специалистов и бактерий-генералистов

Авторы: Д.Д.Николаева [1,2], С.К.Гарушянц [1], М.С.Гельфанд [1,2]

1 - Институт проблем передачи информации им. А.А.Харкевича РАН, Москва, Россия, 2 - Сколковский институт науки и технологий, Москва, Россия

Пангеном - это совокупность белок-кодирующих генов, присутствующих в некотором наборе геномов. Для описания структуры пангенома выделяют “универсальный геном” - гены, которые встречаются почти во всех рассматриваемых штаммах, и “периферию” - гены, встречающиеся лишь у небольшого количества штаммов. Известно, что у разных видов бактерий соотношение размера универсального генома к размеру периферии различается, и какие именно факторы определяют это соотношение, до сих пор остается непонятным. Мы предположили, что одним из определяющих факторов может являться количество экологических ниш, в которых встречается данный вид. Так, виды бактерий, способные существовать в разнообразных условиях (виды-генералисты), должны, с одной стороны, иметь гены, необходимые для приспособления к конкретному местообитанию, а, с другой стороны, могут взаимодействовать с бóльшим количеством бактерий других видов, от которых могут получать более разнообразные гены в результате горизонтального переноса. На другом конце спектра находятся виды-специалисты, которые обитают в небольшом количестве экологических ниш, хорошо приспособлены к ним и поэтому, вероятно, генетически более однородны.

Для получения информации о представленности бактерий в разнообразных местообитаниях использовались данные Earth Microbiome Project (EMP) [1] - крупной

коллекции метагеномных образцов из множества экологических ниш, подготовленных согласно единому экспериментальному протоколу. Чтобы отобрать виды бактерий-генералистов и специалистов для построения и анализа пангеномов, нами была разработана классификация местообитаний, объединяющая схожие по таксономическому составу образцы EMP. На отобранных видах бактерий нам удалось показать, что доля периферии в пангеноме вида больше у генералистов, чем у специалистов, тогда как для доли универсального генома в пангеноме вида наблюдается обратное соотношение, что подтверждает изначальную гипотезу.

Еще один вопрос, вызывающий дискуссию в научном сообществе, заключается в определении того, какая часть пангенома является ответственной за приспособление к генерализму: универсальный геном или периферия. Во множестве работ были описаны гены периферии, дающие возможность бактериям приспосабливаться к новым нишам или хозяевам. Однако, в одной из недавних работ [2] было показано, что генералистов отличают именно гены универсального генома, которые отвечают за адаптацию к большому количеству местообитаний. Анализ функций генов, входящих в универсальный геном или периферию и отличающихся между отобранными нами видами-генералистами и специалистами, позволит нам дать свой ответ на этот вопрос.

Источники и литература:

1. Thompson, Luke R., et al. "A communal catalogue reveals Earth’s multiscale microbial diversity." Nature 551.7681 (2017): 457-463.

2. Maistrenko, Oleksandr M., et al. "Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity." The ISME journal 14.5 (2020): 1247-1259.

9.8 Адаптивная эволюция бактерий кишечного микробиома человека под воздействием пищевых волокон

Авторы: Д.Д.Николаева [1,2], М.С.Гельфанд [1,2], Jens Walter[3]

1 - Институт проблем передачи информации им. А.А.Харкевича РАН, Москва, Россия 2 - Сколковский институт науки и технологий, Москва, Россия 3 - University College Cork, Ireland

Состав и состояние кишечного микробиома оказывают большое влияние на здоровье и самочувствие людей. Несмотря на активное изучение микробиома кишечника человека, лишь небольшое число исследований посвящены тому, как эволюционируют бактерии кишечного микробиома. Поскольку эти бактерии обитают в относительно стабильной среде кишечника

человека на протяжении десятков тысяч лет, ожидается, что их эволюция определяется, главным образом, отрицательным отбором, а все возникающие замены являются нейтральными. Действительно, в ранних исследованиях эволюции кишечного микробиома не было обнаружено следов положительного отбора. Тем не менее, микробиом отдельного индивида в течение жизни хозяина может быть подвержен изменениям в среде обитания в результате воздействия антибиотиков, взаимодействия с другими бактериями и с иммунной системой хозяина, либо перемен в характере питания и образа жизни хозяина. Эти факторы могут оказывать давление отбора на кишечный микробиом индивида, являясь причиной адаптивных эволюционных процессов в микробиоме кишечника человека. На примере бактерии-комменсала кишечника человека Bacteroides fragilis было показано наличие сигналов, свидетельствующих об адаптивной эволюции бактерии в рамках индивидуальных микробиомов здоровых людей [1]. Однако дизайн этого эксперимента не позволил определить, какие именно факторы оказали давление отбора на популяции B. fragilis в индивидуальных микробиомах, поэтому вопрос о том, действует ли адаптивная эволюция на бактерий кишечного микробиома при наличии давления отбора, остается открытым.

Установлено, что на кишечный микробиом можно воздействовать с помощью питания, что может приводить к изменению соотношения присутствующих в микробиоме видов бактерий. Интересно проверить, будет ли длительное воздействие продукта питания не только вызывать изменения в составе кишечного микробиома, но и оказывать давление отбора, вызывая адаптивные процессы в популяциях бактерий. Для проверки этой гипотезы будут использованы данные проекта FYBER [2] — интервенционного клинического испытания, цель которого — выявить закономерности ответа кишечного микробиома на прием пищевых волокон участниками испытания в течение шести недель. Поскольку в этом проекте пробы для метагеномного секвенирования отбирались до начала испытания, а затем после первой и после шестой недель испытания, эти данные позволят разделить экологические и эволюционные последствия приема пищевых волокон. Так, отличия между нулевой и первой точками могут соответствовать изменениям состава микробиома, тогда как отличия между первой и шестой неделями в собранных из метагенома полных геномах одних и тех же видов, будучи обнаруженными, могут отражать адаптации бактерий к пищевым волокнам. Выбор пищевых волокон для этого испытания также повышает вероятность возникновения адаптивного ответа у бактерий кишечного микробиома: части испытуемых был предложен гуммиарабик, который не является основным компонентом питания большинства людей, тогда как другая группа участников принимала устойчивый крахмал четвертого типа, который становится устойчивым при химической обработке и не встречается в природе. Следовательно, есть все основания ожидать, что бактерии кишечного микробиома человека будут адаптироваться к длительному воздействию ранее незнакомых пищевых волокон.

Источники и литература:

1. Zhao, Shijie, et al. "Adaptive evolution within gut microbiomes of healthy people." Cell host & microbe 25.5 (2019): 656-667.

2. https://clinicaltrials.gov/ct2/show/NCT02322112

9.9 Бактериальные кластеры в пространстве вагинальных метагеномов

Авторы: Наталия Драненко[1], Фарида Еникеева[2], Михаил Гельфанд[1,3]

1 - Институт проблем передачи информации имени А. А. Харкевича РАН, Москва, Россия, 2 - Université de Poitiers, Poitiers, Poitou-Charentes, France; 3 - Сколковский институт науки и технологий, Сколково, Россия

Постановка задачи:

Каждый метагеном может быть рассмотрен как точка в пространстве метагеномов. Мы рассматриваем в этом пространстве вагинальные бактериальные сообщества. В таких сообществах были обнаружены кластеры, разделяющиеся по доминирующему виду Lactobacillus или его отсутствию [1]. Целью текущего исследования является анализ вагинальных метагеномов на данных [1] различными способами, проверка на них разных подходов и дальнейшее расширение датасета.

Результаты:

Был проведён анализ результатов кластеризации после проекции на первые 2 и 3 компоненты UMAP на данных [1]. Было обнаружено четыре кластера: в трёх доминирующим видом является тот или иной вид Lactobacillus (L. gasseri, L. crispatus или комбинация L. inners и L. fornicalis) и один кластер с малым содержанием любых бактерий рода Lactobacillus. Хотя в кластере, имеющим в качестве доминантных видов L. inners и L. fornicalis можно провести границу, отделив один вид от другого, она не такая строгая, как для других кластеров. Эти результаты согласуются с результатами [1] с точностью до разделения смешанного кластера. В то же время альтернативный вариант анализа с использованием спектральной кластеризации не дал результатов.

Дальнейшие планы:

В дальнейшем мы планируем анализ пространства метагеномов для других данных: вагинальных метагеномов беременных, для которых доступны метаданные о своевременности родоразрешения, возрасте, приёме лекарств, влияющих на микробиом.

Источники и литература:

1. Ravel, J., Gajer, P., Abdo, Z., Schneider, G., Koenig, S., McCulle, S., Karlebach, S., Gorle, R., Russell, J., Tacket, C., Brotman, R., Davis, C., Ault, K., Peralta, L. and Forney, L., 2010. Vaginal

microbiome of reproductive-age women. Proceedings of the National Academy of Sciences, 108(Supplement_1), pp.4680-4687.

10. Зоопарк

10.1 De novo сборка генома и изучение адаптаций экстремального галотолерантного комара-звонца - Baeotendipes noctivagus (Diptera: Chironomidae)

Авторы: Н.М.Шайхутдинов [1,2], Н.Е.Гоголева [2], О.С.Козлова [2], Н.В.Шадрин [3], Е.В.Ануфриева [3], В.А.Яковенко [3], А.А.Пржиборо [4], Г.Р.Газизова [2], Е.И.Шагимарданова [2], О.А.Гусев [2,5,6], Г.А. Базыкин [1,7]

1 - Центр наук о жизни, Сколковский институт науки и технологий, Москва, Россия, 2 - Институт фундаментальной медицины и биологии, Казанский (Приволжский) федеральный университет, Казань, Россия, 3 - Институт биологии южных морей им. А.О. Ковалевского РАН, Севастополь, 4 - Зоологический институт РАН, Санкт-Петербург, Россия, 5 - Высшая школа медицины, Университет Джунтендо, Токио, Япония, 6 - Центр интегративных медицинских наук RIKEN, Йокогама, Япония, 7- Институт проблем передачи информации им. А.А. Харкевича РАН, Москва, Россия

Изучение геномов немодельных организмов, адаптированных к жизни в экстремальных условиях обитания, представляет собой развивающееся направление в современной геномике, которое позволяет изучить разнообразные физиологические, молекулярные и эволюционные адаптации организмов к окружающей среде. За последнее десятилетие было опубликовано большое количество исследований по секвенированию и ресеквенированию геномов немодельных двукрылых, обитающих в экстремальных условиях [1-5]. Из всех семейств двукрылых - комары-звонцы (Diptera: Chironomidae) имеют наибольшее количество видов, адаптированных к экстремальным абиотическим условиям (низкие и высокие температуры, гипоксия, высокая соленость, низкое и высокое pH, и нехватка воды) [6].

В данной работе был собран геном рекордсмена среди галотолерантных насекомых - Baeotendipes noctivagus, комара-звонца, который может выдерживать широкий диапазон солености от 20 г/л до 300 г/л в личиночной стадии. Была осуществлена гибридная сборка генома de novo с помощью геномного сборщика - Wengan, используя парные короткие прочтения длинной 250 п.о., полученные на платформе Illumina HiSeq 2500, и длинные прочтения, полученные с помощью нанопорового секвенирования (N50 прочтений = 15Kbp) на платформе MinION Mk1C. Общая длина полученной сборки после деконтаминации от бактериальных контигов составила 172.3 Мб, N50 = 2.1 Мб. Геномная сборка была отполирована с помощью программы HyPo с использованием тех же данных, что и при сборке генома. Полнота сборки по BUSCO после этапа полировки с использованием базы данных diptera_odb10 составила 94.4%, что позволяет проводить дальнейший биоинформатический анализ. Данные полиА РНК-секвенирования, полученные из личинок, куколок и имаго были использованы для аннотации генома, которая была выполнена с помощью программы

BRAKER2.

Для определения пула генов, связанных с галотолераностью были получены транскриптомные данные в режиме одиночных прочтений длинной 50 п.о. из индивидуальных личинок 6 гипергалинных озер Крымского полуострова с разной соленостью. Были обнаружено, что в ответ на повышение солености увеличивает свою экспрессию ген Eglp1, связанный с транспортом глицерина - главного осморегулятора в клетках, а также увеличивают свою экспрессию гены гемоглобинов (Hb), которые представлены 45 паралогами в геноме B. noctivagus, что больше, чем у других комаров-звонцов. Дополнительные паралоги является примером эволюционной адаптации на гипоксию, которая наблюдается в гипергалинных озерах.

В данной работе было показано, что технологии секвенирования третьего поколения значительно упрощают и ускоряют сборку генома немодельных насекомых. Также предварительно были определены молекулярные адаптации, позволяющие комару жить при экстремальной солености.

Работа выполнена при поддержке гранта Российского научного фонда № 20-44-07002.

Источники и литература:

1. Gusev O, Suetsugu Y, Cornette R, Kawashima T, Logacheva MD, Kondrashov AS, Penin AA, Hatanaka R, Kikuta S, Shimura S, et al.: Comparative genome sequencing reveals genomic signature of extreme desiccation tolerance in the anhydrobiotic midge. Nat Commun 2014, 5:4784.

2. Mazin PV, Shagimardanova E, Kozlova O, Cherkasov A, Sutormin R, Stepanova VV, Stupnikov A, Logacheva M, Penin A, Sogame Y, et al.: Cooption of heat shock regulatory system for anhydrobiosis in the sleeping chironomid Polypedilum vanderplanki. Proc Natl Acad Sci U S A 2018, 115:E2477–E2486.

3. Kelley JL, Peyton JT, Fiston-Lavier A-S, Teets NM, Yee M-C, Johnston JS, Bustamante CD, Lee RE, Denlinger DL: Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. Nat Commun 2014, 5:4611.

4. Sun X, Liu W, Li R, Zhao C, Pan L, Yan C: A chromosome level genome assembly of Propsilocerus akamusi to understand its response to heavy metal exposure. Mol Ecol Resour 2021, 21:1996–2012.

5. Kim S, Oh M, Jung W, Park J, Choi H-G, Shin SC: Genome sequencing of the winged midge, Parochlus steinenii, from the Antarctic Peninsula. Gigascience 2017, 6:1–8.

6. Pinder LCV: Biology of Freshwater Chironomidae. Annu Rev Entomol 1986, 31:1–23.

7. Hinton HE: A new Chironomid from Africa, the larva of which can be dehydrated without injury. In Proceedings of the Zoological Society of London. . Wiley Online Library; 1951:371–380.

8. Wharton DA: Life at the Limits: Organisms in Extreme Environments. Cambridge University Press; 2002.

9. Sogame Y, Kikawada T: Current findings on the molecular mechanisms underlying anhydrobiosis in Polypedilum vanderplanki. Curr Opin Insect Sci 2017, 19:16–21.

10. Convey P, Block W: Antarctic diptera: Ecology, physiology and distribution. Eur J Entomol 2013, 93:1–13.

11. Shaikhutdinov N, Gusev O: Chironomid Midges (Diptera) Provide Insights into Genome Evolution in Extreme Environments. Curr Opin in Insect Sci 2022, 49:101–7.

10.2 De novo сборка генома и поиск геномных адаптаций у комара Dasyhelea calycata (Diptera: Ceratopogonidae)

Авторы: Дмитрий А. Фёдоров[1], Нурислам М. Шайхутдинов[1,2], Георгий А. Базыкин[1]

1 - Центр наук о жизни, Сколковский Институт Науки и Технологий, Москва, Россия, 2 - Институт фундаментальной медицины и биологии, Казанский (Приволжский) федеральный университет, Казань, Россия

Живые организмы адаптируются к различным условиям среды используя различные физиологические и поведенческие механизмы. Ранее было показано, что комары, живущие в экстремальных условиях, обладают широким спектром геномных адаптаций, таких как дупликации семейств генов, сокращение интронов и геномные перестройки (Shaykhutdinov, Gusev 2022). Dasyhelea calycata сталкивается с экстремальными значения солености на личиночной стадии. При использовании коротких парных прочтений (Illumina) и длинных прочтений (Oxford Nanopore) нами был собран геном D. calycata. 45 полученных контигов покрывают 97% собранного генома. Размер генома D. calycata составляет 103 Мб, что близко к минимально наблюдаемым значениям у Culicomorpha – 89.7 Мб (Belgica antarctica) (Kelley et al. 2014). По нашим данным в геноме D. calycata можно найти сходные адаптации с B. antarctica, например, уменьшение размера интронов, при стабильной длине экзонов. Как и в случае с B. antarctica у D. calycata показано уменьшение эффективной численности популяции в последних поколениях, что может обуславливать сокращение размеров генома благодаря действию отбора. При сравнительном анализе геномов двух известных мокрецов – D. calycata и Culicoides sonorensis с другими известными комарами выявлена потеря значительной части генов.

Источники и литература:

1. Kelley, J. L., Peyton, J. T., Fiston-Lavier, A.-S., Teets, N. M., Yee, M.-C., Johnston, J. S., Bustamante, C. D., Lee, R. E., & Denlinger, D. L. (2014). Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. Nature Communications, 5(1), 4611. https://doi.org/10.1038/ncomms5611

2. Shaikhutdinov, N., & Gusev, O. (2022). Chironomid midges (Diptera) provide insights into genome evolution in extreme environments. Current Opinion in Insect Science, 49, 101–107. https://doi.org/10.1016/j.cois.2021.12.009

10.3 Характеристика внутри- и межсортового разнообразия высококопийной фракции генома гречихи

Авторы: С.Р.Прокопчук, М.Д.Логачева, И.В.Киров, А.Н.Фесенко

Гречиха обыкновенная (Fagopyrum esculentum) является важной сельскохозяйственной зерновой культурой, принадлежащей к семейству Polygonaceae. Гречиха – облигатная перекрестно опыляемая культура, механизм защиты от самоопыления которой находится на морфологическом уровне, и заключается в явлении гетеростилии. Так как для перекрёстно опыляемых культур сортом является популяция растений с различными геномами, а не чистые линии, то вопрос о генетическом разнообразии и различиях между геномами растений внутри сорта и генетическими различиями между сортами-популяциями, является достаточно важным аспектом в изучении генома гречихи.

Известно, что большая часть генома растений представлена фракцией повторяющихся последовательностей, которая отличается высокой вариабельностью и быстрой изменчивостью, а в случае Fagopyrum esculentum, существует предположение о том, что мобильные элементы играют значимую роль в одомашнивании гречихи. Об этом говорят исследования, в которых сравнивались растения F. esculentum с диким родственником F. tataricum. Оказалось, что увеличение размера генома F. esculentum является следствием транспозиционной активности. [1] С прикладной точки зрения, изучение репитома гречихи может стать основой для создания новых сортов методами маркерной и геномной селекции.

В ходе исследования предполагается провести анализ данных секвенирования 40 различных сортов гречихи обыкновенной, а затем сравнить составы репитомов между собой. Для этого по каждому сорту планируется провести независимый запуск RepeatExplorer в версии для командной строки. При дальнейшей обработке полученных с помощью RepeatExplorer данных будут отобраны те повторяющиеся элементы, которые покажут вариабельность между сортами. В результате планируется установить качественный и количественный состав мобильных генетических элементов в каждом сорте и выявить различия между ними внутри и между сортами.

Источники и литература:

1. Penin A. A. et al. High-resolution transcriptome atlas and improved genome assembly of common buckwheat, Fagopyrum esculentum //Frontiers in plant science. – 2021. – Т. 12.

10.4 Поиск перестроек в геноме Schizophyllum Commune

Авторы: А.А.Кондрашина, А.Галицына, М.С.Гельфанд[1,2]

1 - Институт проблем передачи информации имени А. А. Харкевича РАН, Москва, Россия, 2 - Сколковский институт науки и технологий, Сколково, Россия

Hi-C эксперименты широко используются для определения пространственной структуры хроматина [1][3], в частности, компартментов и доменов (ТАДов) хроматина. Границы ТАДов, например, могут быть обнаружены как соседние участки генома с пониженной частотой контактов между собой (иными словами, повышенной инсуляцией). Однако такое свойство может быть присуще также участкам с возможными хромосомными перестройками [10], что позволяет, в частности, улучшать качество сборки геномов. Помимо инсуляции, перестройки генома можно искать с помощью аннотации дальних контактов [2].

Мы заметили, что в методе Hi-C часто секвенируются химерные молекулы ДНК, происходящие из контактирующих фрагментов рестрикции. Однако не все химеры появляются в результате работы протокола Hi-C, некоторые из них могут отражать существующие перестройки в геноме. Поэтому мы предлагаем использовать новый способ для поиска перестроек - аннотацию разрывов ДНК по ридам Hi-C.

В качестве исследуемого объекта был выбран древесный гриб Schizophyllum commune, известный своей высокой нуклеотидной вариабельностью (0.13-0.20, у человека этот показатель - 0.03) [4]. В частности, нам интересно исследовать проблему геномных перестроек этого организма, выяснить, как гипервариабельность влияет на структуру хроматина, а также на обмен областями хромосом при кроссинговере. Так, например, в статье [8] было показано, что кроссинговер в гипервариабельных видах порисходит преимущественно между схожими областями, что также может быть связано с архитектурой хроматина. Мы поставили своей задачей исследование структуры хроматина S. commune.

Цель данной работы - изучение возможностей метода Hi-C для определения перестроек в организме S. commune и, в дальнейшем, исследование передачи этих перестроек потомству.

Методы. В начале результаты эксперимента Hi-C были закартированы на различные доступные геномы гриба S. commune при помощи стандартного пайплайна (distiller-nf [7]). Далее возможные перестройки были определены при помощи пакета pairtools и составлена статистика для выравниваний на разные геномы. Альтернативно, перестройки были определены при помощи пайплайнов juicer [6] и 3d-DNA [5].

План исследования. В ходе работы планируется применить стандартные методы поиска перестроек в геноме (juicer [6] + 3d-DNA [5]) к разным доступным вариантам сборки генома, составить список геномных перестроек, полученных с помощью указанных программ. Так как этот пайплайн чувствителен к качеству сборки генома [9], [10], за референс для валидации можно взять перестройки, полученные при картировании результатов эксперимента на геном

из базы данных RefSeq H4-8, как на наиболее точный.

Кроме того, в планах применить методы computer vision для поиска перестроек по тепловым картам. Такой способ менее надежный, как показали предыдущие исследования хроматина в Danio rerio, но, возможно, окажется менее ресурсозатратным.

Источники и литература:

1. Douglas J. Chapski, Manuel Rosa-Garrido, Nan Hua, Frank Alber and Thomas M. Vondriska. Spatial Principles of Chromatin Architecture Associated With Organ-Specific Gene Regulation.

2. Louise Harewood, Kamal Kishore, Matthew D. Eldridge, Steven Wingett, Danita Pearson, Stefan Schoenfelder, V. Peter Collins and Peter Fraser1. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumors

3. Claire Hoencamp et.al. 3D genomics across the tree of life reveals condensin II as a determinant of architecture type

4. Baranova M.A. et. al. Extraordinary Genetic Diversity in a Wood Decay Mushroom

5. https://github.com/aidenlab/3d-dna

6. https://github.com/aidenlab/juicer

7. https://github.com/open2c/distiller-nf

8. Vladimir B. Seplyarskiy, Maria D. Logacheva, Aleksey A. Penin, Maria A. Baranova, Evgeny V. Leushkin, Natalia V. Demidenko, Anna V. Klepikova, Fyodor A. Kondrashov, Alexey S. Kondrashov, and Timothy Y. James. Crossing-Over in a Hypervariable Species Preferentially Occurs in Regions of High Local Similarity

9. Neva C. Durand, James T. Robinson, Muhammad S. Shamim, Ido Machol, Jill P. Mesirov, Eric S. Lander, and Erez Lieberman Aiden. 2016. “Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom.” Cell Systems

10. Olga Dudchenko, Muhammad S. Shamim, Sanjit Batra, Neva C. Durand, Nathaniel T. Musial, Ragib Mostofa, Melanie Pham, et al. 2018. “The Juicebox Assembly Tools Module Facilitates de Novo Assembly of Mammalian Genomes with Chromosome-Length Scaffolds for under $1000.”

10.5 Филогенетическое дерево трутовых грибов

Авторы: Е.Правдолюбова[1], Otto Miettenen[2], Г.Базыкин[1]

1 - Центр Наук о Жизни, Сколковский институт науки и технологий, Сколково, Россия, 2 - Finnish Museum of Natural History – LUOMUS

Это задача на начальной стадии (мы еще не видели данных с финской стороны).

Возможно, к середине февраля что-то изменится.

Разные способы строить филогенетические деревья у грибов довольно часто дают противоречивые результаты. Яркий пример – работа [1], авторы который перебрали самые разные подходы и самые разные способы фильтровать данные, и показали, что если учитывать известные ловушки (например, притяжение длинных ветвей), то не удается получить достаточно весомые свидетельства в пользу любой из 3 возможных топологий 3 основных ветвей базидиомицетов (при этом результаты анализов на мультифуркацию отрицательные).

Данные – несколько сотен геномов грибов из порядков Гименохетовые, Полипоровые и insertae sedis. Еще одна проблема связана с контаминацией: не все образцы были выделены в культуру, некоторые образцы – плодовые тела, собранные в лесах, в этом случае следует ожидать контаминации, в том числе близкородственной, потому что трутовики иногда растут на трутовиках. Данные собраны до контигов, но есть мысль, что этот шаг надо переделать и сначала деконтаминировать риды.

Желаемый результат – это дерево с хорошо разрешенными кладами, которые можно интерпретировать как семейства и рода. Если будет использовано немного локусов, то топология дерева на уровне этих клад должна быть довольно устойчивой и не меняться от последующего добавления локусов.

Вопросы на обсуждение: какие подходы лучше использовать для того, чтобы обнаружить близкородственную контаминацию? Какие подходы стоит попробовать и каких ошибок избегать?

Источники и литература:

1. Arun N Prasanna, Daniel Gerber, Teeratas Kijpornyongpan, M Catherine Aime, Vinson P Doyle, Laszlo G Nagy, Model Choice, Missing Data, and Taxon Sampling Impact Phylogenomic Inference of Deep Basidiomycota Relationships, Systematic Biology, Volume 69, Issue 1, January 2020, Pages 17–37, https://doi.org/10.1093/sysbio/syz029

11. Транскриптом

11.1 Поиск возможностей анализа низкоэкспрессируемых РНК по данным РНК-секвенирования единичных клеток

Авторы: Мария С. Точилкина, Анна А. Валяева, Евгений В. Шеваль, Андрей А. Миронов[1,3]

1 - Сколковский институт науки и технологий, Сколково, Россия

В настоящее время метод секвенирования РНК единичных клеток (scRNA-seq) популярен, поскольку в отличие от секвенирования группы клеток (bulk) он предоставляет данные экспрессии генов, не усреднённые по популяции, а индивидуальные для каждой клетки. Такой подход позволяет выявлять различия клеточных популяций и эволюционные отношения между клетками. Тем не менее методы scRNA-seq имеют и свои недостатки. Из-за невысокой глубины секвенирования, по сравнению с bulk RNA-seq методами, детектированный уровень экспрессии многих генов может быть равен 0 в данных scRNA-seq. Подобный результат обусловлен либо биологическими причинами, либо дропаут эффектом - техническим несовершенством платформ приготовления библиотек, которые неспособны обработать все транскрипты.

К проблеме дропаут эффекта можно подойти по-разному. С одной стороны, можно использовать более чувствительные методы подготовки scRNA библиотек. Было показано, что full-length подходы имеют более низкую вероятность дропаут эффекта по сравнению с UMI-подходами. Тем не менее они также имеют свои недостатки, такие как высокая стоимость, меньшее количество обрабатываемых клеток, и проблема дропаута тут полностью не решается. С другой стороны, можно использовать биоинформатические методы для восстановления пропущенных значений. Такая задача называется scRNA-seq data imputation. Существуют различные методы решения этой задачи, но их эффективность может варьироваться в зависимости от характеристик данных (используемой платформы подготовки библиотеки, размера библиотеки и т.д.). Более того, такие методы импутации не всегда улучшают дальнейший анализ данных.

Низкоэкспрессируемые гены наиболее подвержены дропаут эффекту. Так, сравнение данных scRNA-seq мышиной трахеи [1], полученных full-length и UMI подходами, показало, что доля клеток, экспрессирующих ACE2 – рецептор входа коронавируса SARS-CoV-2 в клетку, значительно выше в full-length данных. Это может указывать на то, что популяция эпителиальных клеток, потенциально чувствительных к инфекции SARS-CoV-2, может быть недооценена UMI-методами scRNA-seq, которые благодаря своей высокой производительности используются чаще, чем full-length.

Таким образом, проблема изучения экспрессии низкоэкспрессируемых генов по

данным scRNA-seq остаётся актуальной. Для её решения нужно подбирать подходящий метод импутации, возможно комбинировать датасеты, полученные разными платформами, или дополнительно использовать данные bulk-секвенирования.

Источники и литература:

1. Montoro DT, Haber AL, Biton M, et al (2018) A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. Nature 560:319–324

11.2 Биоинформатический анализ полу-экстрагируемых РНК

Авторы: Анна Александровна Валяева, Анастасия Александровна Жарикова, Евгений Валерьевич Шеваль, Андрей Александрович Миронов, МГУ им. М.В.Ломоносова, Москва, Россия

Стандартные методы экстракции РНК не позволяют выделить всю клеточную РНК: определенные фракции РНК остаются в связанном состоянии и вместе с белками осаждаются из раствора либо по другим причинам плохо детектируются. Усиленный дополнительной механической гомогенизацией метод экстракции РНК позволяет обогатить выделяемую фракцию так называемыми полу-экстрагируемыми РНК, которые могут быть как белок-кодирующими, так и некодирующими РНК, в том числе архитектурными РНК (например, NEAT1).

В данной работе мы провели эксперимент по выделению РНК стандартным методом и методом с дополнительной гомогенизацией клеточного лизата в шприце (использовалась клеточная линия HeLa). Выделение РНК проводилось на колонках и с помощью Тризола, затем было проведено секвенирование образцов. Для поиска РНК, проявляющих свойство полу-экстрагируемости, был адаптирован пайплайн для анализа дифференциально экспрессируемых генов. Полученны е риды были картированы на референсный геном человека GRCh38 с помощью программы HISAT2 и уникально картированные на гены (а также на экзоны и интроны) риды были подсчитаны с помощью программы HTSeq count. Анализ дифференциальной экстракции РНК (по аналогии с дифференциальной экспрессией) был произведен с помощью пакета R DESeq2.

Проведенный анализ показал, что РНК 1083 генов проявляют свойство полу-экстрагируемости (ДЭ, padj<0.05, FC>1.5). В 1.5 и более раз эффективнее при усиленной экстракции выделились РНК 536 белок-кодирующих генов и 497 длинных некодирующих РНК (в том числе NEAT1 и MALAT1). GO анализ, однако, не позволил выделить какие-то закономерности в функциональных ролях продуктов найденных белок-кодирующих РНК. Вероятно, свойство полу-экстрагируемости связано с какими-то особенностями молекул РНК, а не с тем, что эти молекулы кодируют.

Нами было замечено, что суммарная длина интронных участков у ДЭ генов больше, чем у неДЭ. Поэтому далее мы попытались проанализировать, в каких случаях полуэкстрагируемость определялась интронами, в каких экзонами. Для этого был проведен анализ дифференциальной экспрессии/экстракции отдельно по ридам, картирующимся на интроны или экзоны генов, в результате которого были получены соответствующие списки ДЭ генов. Оказалось, что в большинстве случаев полуэкстрагируемость генов связана с полуэкстрагируемостью именно интронов (634 гена отличались по покрытию интронов, FC>1.5).

При анализе экспрессии повторов, проведенном с помощью программы TEtranscript, была выявлена полу-экстрагируемость повторов семейств LINE и LTR.

11.3 Pupae recapitulate the embryonic expression program in holometabolous insects

Authers: Aleksandra Ozerova, Mikhail Gelfand, Skolkovo Institute of Science and Technology, Moscow, Russia

Gene expression in multicellular organisms demonstrates large variations in early and late embryogenesis while being constrained in the mid-embryogenesis [1]. This pattern is called the developmental hourglass, the stage of the lowest observed diversity known as the phylotypic stage. The phylotypic stage accounts for establishment of the body plan and the initiation of the organ formation. The hourglass pattern has been observed in the development of many species across various kingdoms including plants, animals, and fungi on the transcriptome level.

Some organisms pass several crucial developmental stages during their lifespan, and thus the hourglass model should not be restricted to embryogenesis. For example, seed germination and vegetative-to-reproductive transition in plants is known to follow the pattern of dissimilarity - similarity - dissimilarity in Arabidopsis thaliana [2]. In insects undergoing radical metamorphosis, the existence of the stage with maximum conservation in the middle of the pupae development is a reasonable hypothesis that has not been tested yet.

Metamorphosis is usually a motionless stage, when no active feeding is observed and the insect body is enclosed into the chrysalis. These physiological properties as well as undergoing processes resemble embryogenesis since at pupal stage organs and systems of the imago are formed.

Study on the mosquito Polypedilum vanderlanki showed the inversion of the transcriptional profile back to the embryonic one during the pupal stage [3]. Therefore recapitulation of the embryonic expression program could be hypothesized. We have performed comprehensive analysis on the datasets available in the public domain to check the similarity between pupae and embryo on the level of transcriptome.

Several datasets among the collected ones indeed have shown an increased similarity between the embryonic and pupal stages on the gene expression level when compared with embryo-larvae transcriptome pairs. At that, gradual changes during life span would yield transcriptomes similar to the previous and upcoming stage; on the contrary, the pupal gene expression resembles the embryonic rather than the larval pattern.

This study was partially supported by RFBR (18-29-13001).

Sources and literature:

1. Kalinka, A.T., Varga, K.M., Gerrard, D.T., Preibisch, S., Corcoran, D.L., Jarrells, J., Ohler, U., Bergman, C.M., and Tomancak, P. (2010). Gene expression divergence recapitulates the developmental hourglass model. Nature 468, 811–814.

2. Drost, H.-G., Bellstädt, J., Ó'Maoiléidigh, D.S., Silva, A.T., Gabel, A., Weinholdt, C., Ryan, P.T., Dekkers, B.J.W., Bentsink, L., Hilhorst, H.W.M., et al. (2016). Post-embryonic Hourglass Patterns Mark Ontogenetic Transitions in Plant Development. Mol. Biol. Evol. 33, 1158–1163.

3. Mazin, P.V., Shagimardanova, E., Kozlova, O., Cherkasov, A., Sutormin, R., Stepanova, V.V., Stupnikov, A., Logacheva, M., Penin, A., Sogame, Y., et al. (2018). Cooption of heat shock regulatory system for anhydrobiosis in the sleeping chironomid Polypedilum vanderplanki. Proc. Natl. Acad. Sci. U. S. A. 115, E2477–E2486.

11.4 Gene expression signature of cell reprogramming demonstrates longevity and rejuvenation effects independent of the loss of cellular identity

Authors: Dmitrii Kriukov [1], Ekaterina Khrameeva[1], Sergey Dmitriev[2], Vadim Gladyshev[3] Alexander Tyshkovskiy[2,3]

1 - Skolkovo Institute of Science and Technology, Moscow, Russian, 2 - Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow, Russian, 3 - Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School

Transient somatic cell reprogramming has been proposed as a perspective rejuvenation intervention. However, its association with the mechanisms of aging and longevity at molecular level remains unclear. Here we conducted a meta-analysis of time-series gene expression data and identified robust transcriptomic signature of cell reprogramming. We observed significant co-regulation of genes perturbed by reprogramming and established lifespan-extending interventions, including those involved in DNA repair and inflammation. On the other hand, age-related gene expression changes were reversed during reprogramming, as confirmed by multi-tissue transcriptomic aging clock. Importantly, the longevity and rejuvenation effects induced by reprogramming factors were mainly independent from the loss of cellular identity. Overall, this work

suggests that new interventions mimicking the rejuvenation effect of reprogramming without posing the risk of neoplasy can be developed and proposes specific genes responsible for this process.

11.4 Регуляция экспрессии генов при трансдифференцировке фибробластов сердца в миофибробласты в результате фиброза сердца

Авторы: Даниил Бобровский[1], Екатерина Храмеева[2]

1 - Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия, 2 - Сколковский институт науки и технологий, Москва, Россия

Фиброз миокарда - разрастание соединительной ткани в сердце, чаще всего возникающее после гибели кардиомиоцитов в результате инфаркта, которое ассоциировано с систолической и диастолической дисфункцией, аритмогенезом и другими заболеваниями [1]. При фиброзе происходит трансдифференцировка фибробластов в особые секреторные и сократительные клетки, называемые миофибробластами [2]. Изучение изменений в организации хроматина и, следовательно, в регуляции экспрессии генов, происходящих в ходе трансдифференцировки, представляет большой интерес как в целях поиска мишеней для терапии, так и для более фундаментального и полного понимания путей, по которым может происходить трансдифференцировка клеток.

Мы анализируем изменения в организации хроматина после симуляции инфаркта на культуре кардиофибробластов крысы. Данные по ацетилированию гистона H3 в позиции K27 были получены с помощью метода иммунопреципитации хроматина с последующим секвенированием (ChIP-seq), а данные по степени открытости хроматина – с помощью секвенирования участков генома, доступных для транспозазы (ATAC-seq).

В прошлом году мы провели предварительные исследования, однако выявленные недостатки протоколов сделали необходимым проведение новых экспериментов.

На предварительных данных были установлены участки генома, в которых после симуляции инфаркта изменялся характер ацетилирования гистона H3K27 или степень открытости хроматина и было показано, что гены, активирующиеся при трансдифференцировке фибробластов в миофибробласты in vitro, участвуют в тех же сигнальных путях, которые in vivo приводят к фиброзу сердца, а также с помощью анализа обогащения мотивов были выявлены транскрипционные факторы, чьи сайты связывания находятся в активных энхансерах.

Сейчас мы приступили к анализу данных, полученных по улучшенному экспериментальному протоколу и надеемся получить более полную картину и детальнее

изучить изменения, происходящие при трансдифференцировке.

Благодарности:

Мы благодарим группу Рамона Бирнбаума из Университета Бен-Гурион (Негев, Израиль) за предоставленные экспериментальные данные.

Источники и литература:

1. Nikolaos G. Frangogiannis, Cardiac fibrosis: Cell biological mechanisms, molecular pathways and therapeutic opportunities, Molecular Aspects of Medicine, Volume 65, 2019, Pages 70-99, ISSN 0098-2997, https://doi.org/10.1016/j.mam.2018.07.001.

2. Hinz B, Phan SH, Thannickal VJ, Galli A, Bochaton-Piallat ML, Gabbiani G. The myofibroblast: one function, multiple origins. Am J Pathol. 2007 Jun;170(6):1807-16. doi: 10.2353/ajpath.2007.070112. PMID: 17525249; PMCID: PMC1899462.

11.5 Исследование функций Sirt6 в тканях мозга на метаболомном уровне в контексте клеточной линии GT1.7

Авторы: Анна Мишина[1], Екатерина Храмеева[1], Дмитрий Смирнов[1], Дебора Тойбер[2]

1 - Сколковский Институт Науки и Технологий, Москва, Россия, 2 - Университет имени Давида Бен-Гуриона в Негеве, Израиль

Sirt6 принадлежит семейству сиртуинов, эволюционно консервативных НАД-зависимых белков, обладающих деацетилазной, АДФ-рибозилтрансферазной и деацилазной активностью. Спустя 16 лет после открытия первого представителя семейства в S. cerevisiae (Rine et al., 1979) было проведено исследование (Guarente et al., 1995) , где обнаружилось, что сиртуин Sir2 является ключевым регулятором продолжительности жизни и клеточного старения дрожжей. С тех пор интерес к этому семейству белков распространился на гомологов Sir2 в более сложных биологических оргазмах, и в настоящем научном проекте исследуется функциональный репертуар Sirt6, одного из семи представителей семейства сиртуинов у млекопитающих.

Нашими колобораторами из лаборатории Деборы Тойбер был проведен метаболомный анализ на мышиной нейронной клеточной линии GT1.7 с двумя группами образцов: Wild Type/нокаут Sirt6. В результате статистического анализа было найдено 15 метаболитов значимо различающихся по концентрации между двумя группами. К найденному списку дифференциальных метаболитов далее применялся Enrichment Analysis, а также была осуществлена интеграция результатов метаболомного анализа с ранее полученными данными транскриптомного анализа с той же концепцией, но на образцах ткани мозга мышей. Применение двух различных подходов позволило выявить 6 метаболических путей

потенциально ассоциированных с функцией Sirt6 в клетках мозга. Далее мы сравнили полученные метаболические тренды с известными (Rabow, Z. et al., 2021) изменениями в ткани гипоталамуса характерными для стареющих мышей. Сравнение продемонстрировало повторение тренда для 7 из 11 дифференциальных аминокислот. Остальные 4 противоречащих метаболита оказались участниками метаболических путей потенциально ассоциированных с деятельностью изучаемого нами сиртуина.

На текущем этапе нашей работы весь фокус сосредоточен на анализе литературы с целью валидировать наши результаты, а также предположить детали и механизмы функционирования Sirt6.

Источники и литература:

1. Rine J, Strathern JN, Hicks JB, Herskowitz I. A suppressor of mating-type locus mutations in Saccharomyces cerevisiae: evidence for and identification of cryptic mating-type loci. Genetics. 1979 Dec;93(4):877-901.

2. Kennedy BK, Austriaco NR Jr, Zhang J, Guarente L. Mutation in the silencing gene SIR4 can delay aging in S. cerevisiae. Cell. 1995 Feb 10;80(3):485-96.

3. Ding J, Ji J, Rabow Z, Shen T, Folz J, Brydges CR, Fan S, Lu X, Mehta S, Showalter MR, Zhang Y, Araiza R, Bower LR, Lloyd KCK, Fiehn O. A metabolome atlas of the aging mouse brain. Nat Commun. 2021 Oct 15;12(1):6021.

12. Эпидемиология

12.1 Молекулярная эпидемиология ВКЭ в России

Авторы: Сафина К.Р.[1,2], Карань Л.С.[3], Неверов А.Д.[3], Базыкин Г.А.[2,1]

1 - Институт проблем передачи информации им. А.А.Харкевича РАН, Москва, Россия,

2 - Сколковский Институт Науки и Технологий, Москва, Россия, 3 ФБУН ЦНИИ эпидемиологии Роспотребнадзора

Вирус клещевого энцефалита (ВКЭ) - флавивирус, приводящий к развитию клещевого энцефалита, инфекционного заболевания центральной нервной системы. ВКЭ эндемичен для многих стран Европы и Азии, что во многом обусловлено ареалом обитания основных хозяев вируса - нескольких видов клещей рода Ixodes. Распространенность вируса со временем меняется под влиянием климатических и социо-экономических изменений; изучение этих процессов представляет очевидный практический интерес.

В рамках данного проекта планируется отсеквенировать и проанализировать коллекцию из 303 образцов ВКЭ, собранных в 1937-2015 годах преимущественно на территории России. В настоящий момент полные геномы получены для 203 образцов. Филогенетический анализ показал, что отсеквенированные образцы относятся к четырем основным генотипам ВКЭ: Европейскому, Сибирскому, Дальневосточному, а также к так называемой Байкальской группе, или 886-подобным штаммам (Рис. 1).

Мы планируем получить для образцов коллекции подробные метаданные (в частности, детальные данные по местам изоляции образцов) и провести филогеографический анализ, чтобы описать изменение представленности генотипов ВКЭ на территории России со временем, выявить паттерны миграции вируса и потенциально оценить вклад различных процессов (н-р, изменение климата, строительство Байкало-Амурской магистрали) в распространение вируса.

Этот проект был поддержан РНФ (номер проекта 21-74-20160).

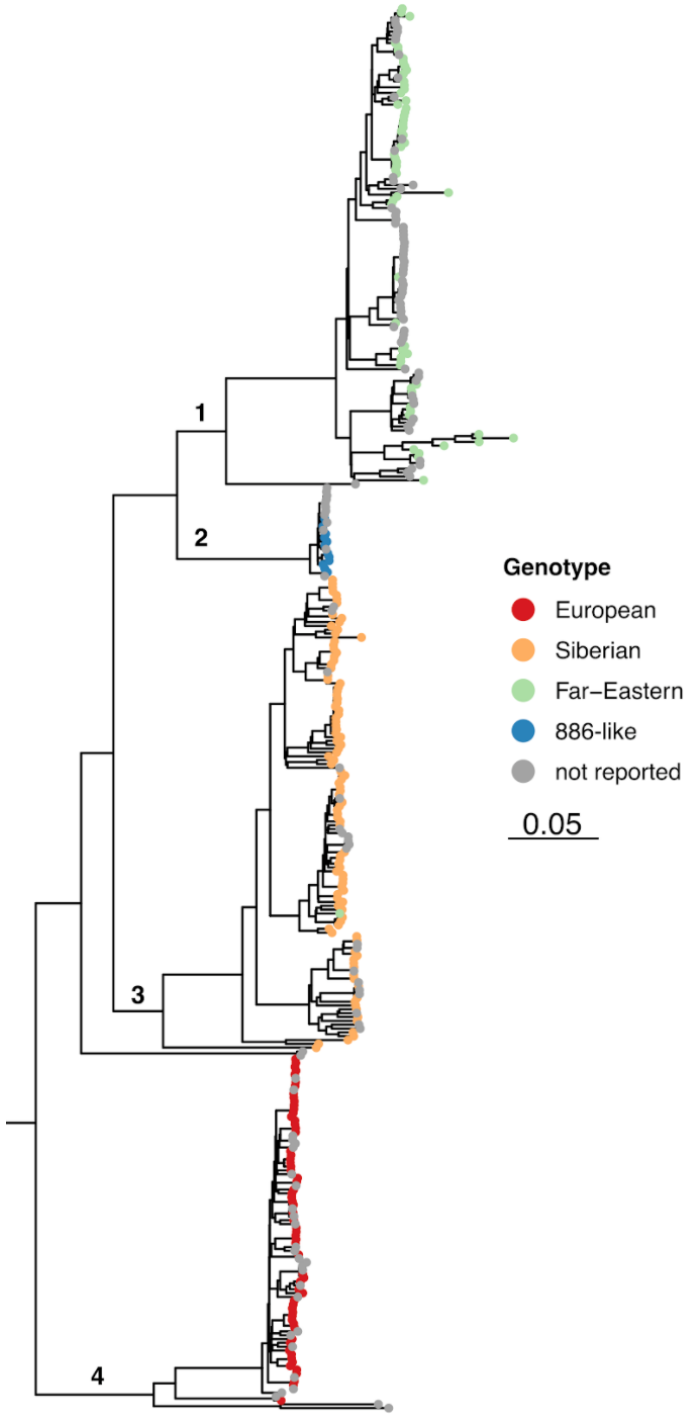

Рис. 1. Филогенетическое дерево ВКЭ, полученное в данной работе. В целях визуализации из дерева были исключены образцы, для которых не был известен ни год, ни регион изоляции. Цветом показаны генотипы образцов, непроаннотированные образцы отмечены серым цветом

12.2 Поиск пространственно-временных особенностей эволюции SARS-CoV2 филогенетическими методами

Авторы: Галя Клинк[1], Георгий Базыкин[1,2]

1 - Институт проблем передачи информации им. А.А.Харкевича РАН, Москва, Россия,

2 - Сколковский Институт Науки и Технологий, Москва, Россия

Более двух лет прошло с тех пор, как SARS-CoV2 впервые был замечен как патоген человека. За это время эволюционное преимущество в разных странах получали как эндемичные, так и глобально распространяющиеся линии. Сейчас в базе данных GISAID лежат геномы для более … вирусных образцов, собранных с начала эпидемии до нынешнего времени по всему миру. Такие данные могут позволить понять с помощью филогенетического анализа, как менялся отбор, действующий на вирус, во времени и пространстве. Помимо этого, можно искать и изучать клады с определенными свойствами: например, с ускоренной или замедленной эволюцией, с избытком редких мутаций. В первую очередь нужно построить надёжное эволюционное дерево и восстановить порядок мутаций на его ветках. Мы будем делать это с помощью пакета UShER (Turakhia et al. 2021). Поскольку ошибки секвенирования порождают множество ложных мутаций в последовательностях геномов SARS-CoV2, важным пунктом работы будет разработка алгоритма фильтрации образцов и позиций в выравнивании.

12.3 Анализ распространения SARS-CoV2 в регионах РФ

Авторы: Галя Клинк[1], Дмитрий Биба[2], Георгий Базыкин[1,2]

1 - Институт проблем передачи информации им. А.А.Харкевича РАН, Москва, Россия,

2 - Сколковский Институт Науки и Технологий, Москва, Россия

Поиск закономерностей распространения вариантов SARS-CoV2 по России - интересная и важная задача. Мы хотим понять, чем определяется степень сходства регионов России по составу вирусных вариантов. Основная наша проблема – неравномерность секвенирования вирусных образцов во времени и пространстве. Мы определяем степень сходства регионов двумя независимыми способами: как интенсивность миграции между регионами, оцененная программой BEAST (Bouckaert et al., 2019) и как нормированный индекс Жаккарда (Jaccard, 1912). Первые результаты показывают, что европейские регионы России сильнее похожи по составу трансмиссионных линий, чем азиатские.

12.4 Повлияло ли закрытие границ в большинстве стран мира в начале 2020 на распространение коронавируса?

Авторы: Дмитрий Биба[1], Guy Baele[2], Георгий Базыкин[1]

1 - Сколковский Институт Науки и Технологий, Москва, Россия 2 - Rega Institute, KU Leuven

В марте 2020 большинство государств предприняло ряд мер по ограничению мобильности граждан, в том числе ограничение на перелёты между странами (закрытие границ). По задумке это должно было задержать распространение SARS-CoV-2 и дать время на подготовку тем странам, до которых пандемия ещё не дошла. Очевидным побочным эффектом был удар по экономике. Цель этой работы – понять, замедлилось ли распространение коронавируса между странами в результате предпринятых мер. Для этого мы используем филогенетическое дерево коронавируса, построенное на основе последовательностей, собранных в период декабрь 2019 – май 2020. С помощью методов дискретной филогеографии, имплементированных в пакете BEAST, мы оцениваем скорости переносов коронавируса между разными частями света до и после закрытия границ. Мы также проводим субсэмплирования и используем симулированные данные для демонстрации устойчивости результатов.

Благодарности: Галина Клинк, Ксения Сафина, Владимир Щур, Вадим Спирин, Mandev Gill, Nena Bollen, Samuel Hong.

12.5 Распространенность гаплотипов гена APOE в российской популяции

Авторы: Василий Раменский[1,3], Александра Ершова[1], Мария Зайченока[2], Анастасия Жарикова[1,3], Юрий Вяткин[1,4], Алексей Мешков[1]

1 - Национальный медицинский исследовательский центр терапии и профилактической медицины, Россия, 2 - Московский физико-технический институт, Центр образовательных программ по биоинформатике, 3 - МГУ им Ломоносова, Москва, Россия, 4 - Новосибирский государственный университет, Россия

Было проведено исследование вариантов генома в 242 клинически важных генах в условно здоровых неродственных индивидуумах из г. Иваново и области. Целью этого исследования было описание спектра вариантов, в первую очередь, известных и потенциально болезнетворных, в выборке клинически значимых генов у представителей российской популяции. Итоговая выборка включала 1685 человек: 1056 женщин с медианным возрастом 52 года на момент включения в исследование и 629 мужчин с медианным возрастом 44 года. Для проведения исследования была разработана панель генов, включающая кодирующие экзоны 242 белок-кодирующих генов, связанных с сердечно-сосудистыми заболеваниями и высоким риском ранней или внезапной сердечной смерти. Обработка данных секвенирования и оценка контроля качества выполнялись с помощью специально разработанного на основе GATK 3.8 конвейера, следующего «лучшим практикам», разработанным в Broad Institute

вместе с программой GATK.

Аполипопротеин Е (APOE) представляет собой белок длиной 299 остатков, который играет ключевую роль в регуляции уровня липидов в плазме крови. У человека хорошо изучены три наиболее распространенных гаплотипа APOE, определяемые двумя несинонимичными вариантами rs429358 и rs7412 и образующие изоформы e2, e3 и e4 белка APOE. В данной работе была определена распространенность основных клинически значимых гаплотипов гена APOE в российской популяции на примере ивановской выборки. Показано, что носительство гаплотипа e2 ассоциировано с понижением уровна липопротеинов низкой плотности относительно наиболее распространенного генотипа e3/e3 и повышением уровня триглицеридов в случае гомозиготного носительства e2. Для носителей гаплотипа e4 характерно повышение уровня липопротеинов низкой плотности.

Исследование было рассмотрено и одобрено независимым этическим комитетом Национального медицинского исследовательского центра Терапии и профилактической медицины (протокол № 07-03/12 от 03.07.2012) и проведено в соответствии с принципами Хельсинкской декларации. Участники предоставили письменное информированное согласие на участие в исследовании. Работа поддержана Госзаданием № 121021100127-8.

12.6 Properties of variation in populations of different effective sizes

Authors: Yunna S. Petrusenko[1], Alexey S. Kondrashov[2]

1 - Faculty of Biology, M. V. Lomonosov Moscow State University, Moscow, Russian Federation, 2 - Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA

Abstract:

As a key parameter in population genetics effective population size (Ne) reflects an intensity of genetic drift that in its turn defines the effectiveness of natural selection and the tendency to accumulate slightly deleterious mutations in particular cases. However, the signatures characterizing the power of natural selection often are not obvious; this obstructs our understanding of the population fate and its mutation load regarding other populations within the same species or those related to other species. Studying deleteriousness of alleles in the context of conservative genes across species of different Ne, we assume to reveal some correlations demonstrating how single nucleotide polymorphisms (SNPs) within the orthologs of the selected populations correspond a conservation of appropriate sites. Here we compare Homo sapiens, Drosophila melanogaster and Schizophyllum commune as species which populations are characterized by different genetic diversity and Ne, and there are available genomic data for variation analysis. To establish conservation values for each site

in one-to-one orthologs trios for the selected species, we consider orthologs of organisms being outgroups for the lines of a common ancestor of animals and fungi. Thus, based on common conservation values, various metrics showing properties of variation in the given populations could be counted and, eventually, support our hypothesis that in the population with smaller Ne, a human population in our case, natural selection is weakened and therefore actually provides long-term fixation of slightly deleterious variants in the genome.

12.7 Генетические корреляты социальной структуры в Эстонии

Авторы: Иван Кузнецов[1], Георгий Базыкин[1,2], Юрий Аульченко[3,4]

1 - Сколковский институт науки и технологий, Москва, Россия, 2 - Институт проблем передачи информации им. А. А. Харкевича РАН, Москва, Россия, 3 - Институт цитологии и генетики СО РАН, Новосибирск, Россия, 4 - PolyKnomics BV, Хертогенбос, Нидерланды

Популяция человека имеет генетическую структуру. Различия частот аллелей между субпопуляциями могут быть следствием случайного дрейфа, естественного отбора, миграций и смешения популяций. Субпопуляции имеют также фенотипические отличия, имеющие под собой как генетическую, так и средовую компоненты. Фенотип, в свою очередь, может оказывать влияние на вероятность и направление миграций, как и на взаимодействие между субпопуляциями, частично определяя генетическую структуру.

Abdellaoui с коллегами в недавней работе1 показали, что средние значения полигенных оценок для некоторых признаков, связанных с социоэкономическим положением и здоровьем человека, распределены по регионам Великобритании неравномерно. Особенно выделяется на фоне остальных полигенная оценка для уровня образования. Регионы с относительно низкими средними значениями полигенной оценки для уровня образования значительно перекрываются с областями, в которых ранее активно велась добыча угля. Сейчас эти области отличаются от других своей относительной экономической отсталостью. Кроме того, отмечаются различия между средними значения полигенных оценок в группах людей, мигрировавших из/в бывшие угледобывающие регионы и оставшихся в них или за их пределами. Данные наблюдения могут являться следствием социальной стратификации в Великобритании 1.

В своей работе совместно с коллегами из Тартуского университета мы намерены провести анализ географического распределения признаков, а также значений полигенных оценок для них в эстонской популяции. Планируется использовать данные Эстонского Биобанка, содержащего информацию о генотипе и фенотипе более чем 150 тыс. человек, что составляет более 10% населения Эстонии 2,3. Данная работа позволит определить, в какой

степени следы социальной стратификации присутствуют в генетической структуре популяции Эстонии.

Источники и литература:

1. Abdellaoui, A. et al. Genetic correlates of social stratification in Great Britain. Nat Hum Behav 3, 1332–1342 (2019).

2. Leitsalu, L. et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int. J. Epidemiol. 44, 1137–1147 (2015).

3. Kivisild, T. et al. Patterns of genetic connectedness between modern and medieval Estonian genomes reveal the origins of a major ancestry component of the Finnish population. Am. J. Hum. Genet. 108, 1792–1806 (2021)._

13. Бактерии

13.1 Поиск новых случаев фазовых вариаций у бактерий

Авторы: Елизавета Силиг[1], Ольга Бочкарева[2]

1 - ФББ МГУ, Москва, Россия, 2 - Institute of Science and Technology (IST Austria)

Фазовые вариации - один из адаптивных процессов, в ходе которого бактерии претерпевают периодические обратимые перемены в фенотипе. Причиной этих изменений служат генетические изменения, ассоциированные с определенными видоспецифичными локусами в геноме. Показано, что фазовые вариации аспространены у бактерий, выживающих в организме хозяина и способствуют “обману” их иммунитета в постоянно изменяющихся условиях.

Целью нашей работы является поиск и изучение новых случаев фазовых вариаций и их механизмов с помощью алгоритма PaReBrick. В результате работы будут систематизированы как уже описанные случаи фазовой вариации в патогенах человека, так и предсказаны новые механизмы.

13.2 Гомологичная рекомбинация в базовом геноме Vibrio

Авторы: Афасижев Р.Н.[1], Гельфанд М.С.[1]

1 - Институт проблем передачи информации им А.А.Харкевича Российская академия наук, Москва, Россия

Известно, что при сравнении двух геномов однонуклеотидные полиморфизмы распределены не случайно, а формируют группы. Такие группы снипов часто являются следами события гомологичной рекомбинации. Мы разбиваем геном на участки по 1 кб и рассматриваем распределения полиморфизмов в таких блоках. Построена модель, описывающая такое распределение полиморфизмов, которая является взвешенной суммой распределения Пуассона и распределения Эрланга, для описания вклада клонально унаследованных и рекомбинантно перенесенных однонуклеотидных полиморфизмов, соответственно. С помощью фиттинга модели получены параметры, показывающие: соотношение вклада клональных и рекомбинантных полиморфизмов, расстояния расхождения пар геномов. Модель была применена для оценки рекомбинации между штаммами Vibrio. Были продемонстрированы отличия между штаммами Vibrio на примере выборки из видов Vibrio.

13.3 Организация многокомпонентных бактериальных геномов

Авторы: Наталия Драненко[1], Ярослав Деменчук[2], Александра Родина[3], Вера Емельяненко[4], Ольга Бочкарева[4]

1 - Институт Проблем Передачи Информации РАН, 2 - Национальный исследовательский университет «Высшая школа экономики», 3 - «Частная школа ЦОДИВ», 198323, Санкт-Петербург (Горелово), 4 - Institute of Science and Technology (IST Austria)

Большинство известных нам бактерий содержит только один крупный репликон – бактериальную хромосому. Однако известны бактерии, у которых несколько крупных репликонов. Структуры геномов таких бактерий довольно разнообразны, встречаются как 2 или 3 крупных репликона, которые принято считать дополнительными хромосомами, так и большее число более мелких репликонов, которые считают мегаплазмидами. Мы изучили состав и размеры репликонов в геномах различных бактериальных родов и выявили корреляционные зависимости в этих данных.

Цель работы – узнать, как формируется размер разных типов репликонов в таких геномах и какие факторы влияют на эти зависимости.

Благорадарности: проект начат на Школе Молекулярной и Теоретической Биологии 2021.

13.4 B. mallei: adaptation to intracellular lifestyle

Authors: Ariadna A. Semenova[1], Alexey Zabelkin[2], Olga O. Bochkareva[3]

1 - FBB MSU, Moscow, Russia, 2 - Computer Technologies Laboratory, ITMO University, St Petersburg 197101, Russia, 3 - Institute of Science and Technology (IST Austria), 3400 Klosterneuburg, Austria

Burkholderia is a genus of Gram-negative, ubiquitous bacteria that includes a large number of human, animal, and plant pathogens. Their genomes consist of two replicons: a primary chromosome and a chromid, an essential circular replicon evolved from plasmid. B. mallei relatively recently separated from the ancestral species of extracellular pathogen B. pseudomallei and transited to intracellular parasitism. In this work we investigated patterns of such adaptation in primary and secondary replicons and the role of different IS families in this process. Our results show that an adaptation to intracellular lifestyle is accompanied by 1) decrease of IS elements variety; 2) growth of copy number of specific IS families. Accumulation of multiple identical repeats explains the high rate of genome reduction.

Благорадарности: проект начат на Школе Молекулярной и Теоретической Биологии 2019.

13.5 Chromosome-encoded IpaH ubiquitin ligases indicate non-human enteroinvasive Escherichia

Authors: Natalia O. Dranenko[1], Maria Tutukina[1,2,3], Mikhail S. Gelfand[1,2], Fyodor A. Kondrashov[4], Olga O. Bochkareva[4#]

1 - A.A. Kharkevich Institute for Information Transmission Problems, Moscow, Russia, 2 - Skolkovo Institute of Science and Technology, Moscow, Russia, 3 - Institute of Cell Biophysics, Russian Academy of Sciences, FRC PSCBR RAS, Moscow Region, Pushchino, Russia, 4 - Institute of Science and Technology (IST Austria), Klosterneuburg, Austria

Until recently, Shigella and enteroinvasive Escherichia coli were thought to be primate-restricted pathogens. The base of their pathogenicity is the type 3 secretion system (T3SS) encoded by the pINV virulence plasmid, which facilitates host cell invasion and subsequent proliferation. A large family of T3SS effectors, E3 ubiquitin-ligases encoded by the ipaH genes, have a key role in the Shigella pathogenicity through the modulation of cellular ubiquitination that degrades host proteins. However, recent genomic studies identified ipaH genes in the genomes of Escherichia marmotae, a potential marmot pathogen, and an E. coli extracted from fecal samples of bovine calves, suggesting that non-human hosts may also be infected by these strains, potentially pathogenic to humans.

We performed a comparative genomic study of the functional repertoires in the ipaH gene family in Shigella and enteroinvasive Escherichia from human and predicted non-human hosts. We found that fewer than half of Shigella genomes had a complete set of ipaH genes, with frequent gene losses and duplications that were not consistent with the species tree and nomenclature. Non-human host IpaH proteins had a diverse set of substrate-binding domains and, in contrast to the Shigella proteins, two variants of the NEL C-terminal domain. Inconsistencies between strains phylogeny and composition of effectors indicate horizontal gene transfer between E. coli adapted to different hosts. These results provide a framework for understanding of ipaH-mediated host-pathogens interactions and suggest a need for a genomic study of fecal samples from diseased animals.

Acknowledgements:

The project was initiated with Aygul Minnegalieva and Yulia Yakovleva at the Summer School of Molecular and Theoretical Biology (SMTB-2020), supported by the Zimin Foundation. We thank Inna Shapovalenko, Daria Abuzova, Elizaveta Kaminskaya, and Dmitriy Zvezdin for their contribution to the project during SMTB-2020. We also thank Peter Vlasov for fruitful discussions.

13.6 Composition of metabolic loci in bacterial genomes

Authors: Evgenia Khodzhaeva[1], Zoe Chervontseva[1] Mikhail Gelfand[1,2]

1 - Institute for Information Transmission Problems, Moscow, Russia, 2 - Skolkovo Institute of Science and Technology, Skolkovo, Moscow Region, Russia