УДК 575.112

РАЗРАБОТКА АЛГОРИТМОВ ПОСТРОЕНИЯ ПРИЗНАКОВ И АНАЛИЗ МЕХАНИЗМОВ ОНКОГЕНЕЗА С ИСПОЛЬЗОВАНИЕМ ДАННЫХ HI-C.

Дравгелис В.А. (Университет ИТМО), **Забелкин А.А.** (Университет ИТМО)
Научный руководитель – кандидат физико-математических наук, ординарный доцент Алексеев Н.В.
(Университет ИТМО)

Аннотация: В данной работе проводится анализ пространственной расположенности геномных перестроек в человеческом геноме основанный на данных Hi-C. Осуществлен поиск возможных паттернов и зависимостей расположения перестроек и пространственной организации генома.

Введение.

Злокачественные опухоли всегда были частью человеческого опыта, первые упоминания рака молочной железы обнаружены в древнеегипетском папирусе датированным 1600 годом до н.э. Самая же древняя опухоль обнаруженная у человека имеет возраст более 1.7 млн. лет и была обнаружена в кости стопы.

На данный момент во всём мире рак вызывает 3,7 млн новых случаев и 1,9 млн смертей ежегодно, являясь второй из основных причин заболеваемости и смертности в Европе, третьей в России и причиной 13% смертей в мире. Рак может поразить человека любого возраста, но чем старше человек, тем больше вероятность, поэтому это одна из основных причин смерти в развитых странах с большой продолжительностью жизни.

Но почему настолько долго известная и распространённая болезнь до сих пор с трудом поддается лечению и остаётся не до конца изученной? Всё дело в том, что рак может поражать все виды биологических тканей. А этиология рака остаётся не до конца изученной, потому что факторы возникновения рака имеют стохастический характер и лишь увеличивают вероятность его возникновения. Так же, из-за того, что это по сути своей последствия соматических мутаций генома, то существует бесчисленное множество вариантов того, во что клетка может эволюционировать в ходе онкогенеза, что вызывает необходимость индивидуального подхода к лечению каждого случая.

По всем этим причинам исследования рака очень сложны и комплексны, но от этого они не теряют своей актуальности. Исследования причин возникновения рака (онкогенеза) вносят существенный вклад в борьбу с этой болезнью. В данной работе анализируются раковые клетки для выявления фундаментальных закономерностей онкогенеза. В частности исследуется одно из геномных событий, которое с высокой вероятностью может приводить к превращению обычной клетки в раковую. Это событие – хромотрипсис.

Основная часть. Хромотрипсис – это катастрофическое событие в ядре клетки, которое представляет собой массивное разрушение хромосомы и затем сборку её в случайном порядке. Если после этого клетка не умирает, то у нее высокая вероятность стать раковой. Существуюшие на данный момент алгоритмы детекции хромотрипсиса не учитывают расположение генома в трёхмерном пространстве, а ограничиваются только его одномерной проекцией. Поэтому для дальнейшего построения собственного алгоритма, было принято решение добавить к данным информацию о пространственной организации хромосомы. Для этого используются матрицы пространственной близости полученные методом Hi-C.

Результатом метода Hi-C является матрица частот обнаружения контактов локусов хромосомы или хромосом рядом друг с другом. И в итоге чем ближе были определенные локусы хромосомы друг к другу, тем сильнее сигнал Hi-C они имеют.

В работе были использованы следующие данные: матрицы Hi-C в формате cool для нескольких рассматриваемых типов рака, список хромосомных перестроек и таблицы с разметкой ShatterSeek для них, полученные из проекта PanCancer.

Для выполнения дальнейшего анализа потребовалось нормализовать матрицы Hi-C, чтобы уменьшить влияние различных шумов и снизить влияние паттерна главной диагонали на результаты. Благодаря этому произошёл переход от абсолютных значений к значениям observed/expected с ожидаемым средним равным единице. Для того, чтобы определить в каком разрешении использовать матрицы Hi-C и какого размера перестройки рассматривать, был проведён quality control, результатом которого стал выбор разрешения 400 тысяч баз на бин для матриц Hi-C и перестройки длиной более 60 тысяч баз.

Была выдвинута и проверена гипотеза, о том, что перестройки происходят на находящихся рядом в пространстве локусах. Для этого на основе изначальных перестроек, из их брейкпоинтов, были получены комбинации двух видов. Комбинации из брейкпоинтов одного пациента и разных пациентов. Было выяснено, что медианные и средние значения сигнала Hi-C множества изначальных перестроек и множества комбинаций одного пациента имеют большие значения чем у множества комбинаций разных пациентов и чем ожидаемое среднее. Это подтверждает изначальную гипотезу и говорит о том, что перестройки образуют пространственные кластеры в рамках одного пациента.

Далее перестройки были поделены на два множества – хромотрипсисные и нехромотрипсисные на основании разметки ShatterSeek. После этого из них так же были составлены комбинации только уже большего числа типов. Были получены комбинации из двух хромотрипсисных брейкпоинтов, из двух нехромотрипсисных и из одного хромотрипсисного и одного нехромотрипсисного. Так же они делятся на составленные из брейкпоинтов одного пациента и разных. Полученные значения среднего и медианного сигнала Hi-C показали, что чисто хромотрипсисные комбинации и чисто нехромотрипсисные комбинации образуют множества с более высоким сигналом Hi-C, точно также как комбинации одного пациента, что может говорить о том, что хромотрипсисные перестройки формируют отдельный пространственный кластер.

Выводы. На текущий момент получены следующие выводы: перестройки происходят на близких в пространстве локусах и образуют пространственные кластеры. Так же внутри этих кластеров могут существовать отдельные кластеры делящиеся по признаку хромотрипсиса. Полученные данные предполагается использовать в дальнейшем для построения алгоритма детекции хромотрипсиса, который будет учитывать не только данные о координатах перестроек, но и данные из матриц Hi-C.

Дравгелис В.А. (автор) Подпись

Забелкин А.А. (автор) Подпись

Алексеев Н.В. (научный руководитель) Подпись