Федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

На правах рукописи

Челошкина Ксения Сергеевна

Подходы машинного обучения для анализа разрывов раковых геномов

Резюме

диссертации на соискание учёной степени

кандидата компьютерных наук

Научный руководитель:
кандидат физико-
математических наук,
Попцова Мария Сергеевна

Москва – 2023

Диссертационная работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский университет «Высшая школа экономики».

Научный руководитель: Попцова Мария Сергеевна, к.ф.-м.н., доцент факультета компьютерных наук, заведующий лабораторией «Международная лаборатория биоинформатики», Национальный исследовательский университет «Высшая школа экономики».

ТЕМА ДИССЕРТАЦИИ

Обнаружение и лечение рака являются первостепенными задачами науки и медицины 21 века. Сложность решения этих задач обусловлена сложностью процессов развития рака и гетерогенностью раковых мутаций в геноме. Обычно раковый геном имеет большое число мутаций. Раковые геномы характеризуются наличием точечных мутаций нуклеотидов и небольшими (длиной в несколько нуклеотидов) делециями и вставками, называемыми «indels». Другое характерное свойство раковых геномов – это формирование разрывов, которые ведут к значительным геномным перестановкам (вставкам, делециям, тандемным дубликациям, транслокациям) от нескольких десятков до миллионов нуклеотидов. Эти изменения лишают раковый геном стабильности и разрушают различные механизмы нормального функционирования клетки, такие как деление, рост и дифференциация.

Для изучения мутационных процессов в раковых геномах с целью определения биомаркеров и генов-драйверов были созданы раковые геномные консорциумы для организации сбора данных раковых геномов. Благодаря усилиям международных консорциумов The Cancer Genome Atlas (TCGA) и International Cancer Genome Consortium (ICGC) были задокументированы сотни тысяч раковых разрывов для разных типов рака [1], [2]. В настоящий момент консорциум Pan-Cancer Analysis of Whole Genomes (PCAWG) представил комплексный анализ более чем 2500 раковых геномов по 38 типам рака [3]. Вышеупомянутые консорциумы предоставили публичный доступ к данным для ученых со всего мира с целью проведения исследований по изучению рака.

Одновременно с данными о раковых геномах стали доступны омиксные данные, а именно полногеномные карты различных эпигенетических признаков (метилирование, доступность хроматина, модификации гистонов), а также данные об альтернативных формах ДНК (Z-ДНК, квадруплексы, триплексы, структуры стебель-петля). Исторически сформировалось несколько научных направлений для изучения геномов в разрезе разных научных областей, имеющих в названии общий суффикс «омика»: геномика, протеомика, метаболомика, транскриптомика. Все вместе эти направления

называются «омиксными» и нацелены на получение комплексного представления о структуре и функционировании генома.

Однако несмотря на большое количество доступных раковых данных, мутагенез раковых разрывов еще не был достаточно изучен и качество предсказания раковых разрывов моделями машинного обучения было намного ниже, чем для точечных раковых мутаций.

Целью данного исследования является комплексное изучение раковых разрывов с помощью методов машинного обучения. Для достижения этой цели были поставлены следующие задачи:

1. Собрать данные и провести обзор state-of-the-art методов для предсказания раковых разрывов и точечных мутаций.

2. Определить правила для идентификации областей повышенной плотности раковых разрывов, разработать и применить пайплайн машинного обучения для их предсказания.

3. Предложить методы определения важности признаков для предсказания областей повышенной плотности раковых разрывов.

4. Исследовать «случайность» появления раковых разрывов.

5. Проверить, могут ли методы обучения на позитивных и неразмеченных примерах (PU-learning) повысить качество модели.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

Основные положения, выносимые на защиту:

1. Были выделены области повышенной плотности раковых разрывов и предложена методология для их предсказания с помощью методов машинного обучения.

2. Данная методология была протестирована на реальных данных. Были разработаны модели машинного обучения для предсказания раковых разрывов и оценки вклада омиксных данных, которые превзошли другие известные на тот момент модели машинного обучения.

3. С помощью разработанного подхода на основе машинного обучения было обнаружено тканеспецифичное влияние квадруплексов и структур стебель-петля на наличие областей повышенной плотности раковых разрывов.

4. С помощью разработанного подхода для оценки важности индивидуальных признаков и групп признаков было обнаружено, что неканонические структуры ДНК и транскрипционные факторы являются наиболее важными предикторами областей повышенной плотности раковых разрывов для всех типов рака.

5. С помощью разработанного подхода было продемонстрировано, что области более высокой плотности разрывов более отличимы от остальных участков раковых геномов, чем участки с меньшей плотностью разрывов.

6. Мы протестировали два PU-learning подхода и обнаружили, что включение в модель информации о неопределенности в разметке областей повышенной плотности раковых разрывов не позволяет повысить качество модели.

Личный вклад автора представлен в виде анализа и визуализации данных, разработки подходов машинного обучения, имплементаций кода, написания текстов научных публикаций. Мария Попцова предложила концепцию исследования и ставила задачи.

ПУБЛИКАЦИИ И АПРОБАЦИЯ РАБОТЫ

Публикации повышенного уровня:

1. Cheloshkina K, Poptsova M. Tissue-specific impact of stem-loops and quadruplexes on cancer breakpoints formation. BMC cancer. 2019 Dec;19(1):1-7.

2. Cheloshkina K, Poptsova M. Comprehensive analysis of cancer breakpoints reveals signatures of genetic and epigenetic contribution to cancer genome rearrangements. PLoS computational biology. 2021 Mar 1;17(3):e1008749.

3. Cheloshkina K, Bzhikhatlov I, Poptsova M. Cancer Breakpoint Hotspots Versus Individual Breakpoints Prediction by Machine Learning Models. International Symposium on Bioinformatics Research and Applications 2020 Dec 1 (pp. 217-228). Springer, Cham.

Публикации стандартного уровня:

1. Cheloshkina K, Poptsova M. Understanding cancer breakpoint determinants with omics data. Integr Cancer Sci Therap. 2020;7(1):10-5761.

2. Cheloshkina K, Bzhikhatlov I, Poptsova M. Randomness in Cancer Breakpoint Prediction. Journal of Computational Biology. 2021 Jun 15.

Для всех представленных публикаций автор является первым автором и ответственным за дизайн и имплементацию подходов машинного обучения. Мария Попцова предложила концепцию исследования и ставила задачи, в написании статей участвовали автор и Мария Попцова.

Прочие публикации:

1. Cheloshkina, K. (2021). Ranking Weibull Survival Model: Boosting the Concordance Index of the Weibull Time-to-Event Prediction Model with Ranking Losses. In: Kovalev, S.M., Kuznetsov, S.O., Panov, A.I. (eds) Artificial Intelligence. RCAI 2021. Lecture Notes in Computer Science(), vol 12948. Springer, Cham. https://doi.org/10.1007/978-3-030-86855-0_4

СОДЕРЖАНИЕ РАБОТЫ

В данной секции показаны основные результаты и сформулированы выводы.

1. Существующие подходы для предсказания разрывов в раковых геномах

Раковые геномы характеризуются нестабильностью и подвергаются многочисленным перестановкам, что приводит к возникновению таких структурных вариантов, как делеции, вставки, транслокации и вариации

числа копий. В течение последних 20 лет несколько консорциумных проектов по исследованию раковых геномов (The Cancer Genome Atlas (TCGA) [1], International Cancer Genome Consortium (ICGC) [2] и ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Project [3]) опубликовали данные о расположении точечных и структурных мутациях в тысячах раковых геномов. Определение ключевых факторов возникновения раковых мутаций открывает возможности для понимания геномики заболевания, однако гетерогенность мутаций раковых геномов усложняет анализ.

В этой области одной из важнейших задач является понимание факторов и механизмов, лежащих в основе мутагенеза. Ниже описаны исследования, в которых методы статистики и машинного обучения применяются для предсказания точечных мутаций и разрывов при раке.

Алгоритмические подходы для предсказания точечных мутаций и разрывов при раке

До машинного обучения исследователи использовали статистические методы, чтобы идентифицировать зависимости между различными геномными факторами. Затем машинное обучение стало широко использоваться для различных задач, таких как предсказание целевой переменной (задача классификации или регрессии), отбор важных признаков, уменьшение размерности признакового пространства.

Для первого обширного исследования плотности мутаций были использованы агрегированные до уровня 1 Мб данные об экспрессии генов, времени репликации, состоянии гетерохроматина (H3K9me3) и данные о репарации ошибочно спаренных нуклеотидов (измеренные через микросателлитную нестабильность) [6]. С помощью теста Манна-Уитни было показано, что процесс репарации ошибочно спаренных нуклеотидов оказывает влияние на вариацию концентрации мутаций: для 72% геномных окон разница в частоте мутаций оказалась значимой. Кроме того, PCA анализ плотности мутаций индивидуальных геномов продемонстрировал ее различие между геномами разных видов рака

(посредством сравнения весов PCA-компонент с помощью теста Манна-Уитни).

Комплексный анализ некодирующих точечных мутаций и вставок-делеций в 212 геномах с раком желудка был проведен в [7]. Для определения наиболее информативных эпигенетических признаков для задачи моделирования плотности соматических мутаций авторы использовали логистическую регрессию с лассо-регуляризацией для предсказания наличия мутации в нуклеотиде по среднему сигналу признаков. Кроме того, логистическая регрессия была использована для предсказания вероятности мутаций для каждого пациента, где вероятности были использованы для идентификации областей разрывов повышенной плотности с помощью пуассоновской биномиальной регрессии. Авторы обнаружили 34 таких области, 11 из которых были расположены в сайтах связывания CTCF. Тест Манна-Уитни показал, что расстояние от сайтов связывания CTCF до ближайшего разрыва - вариации числа копий - меньше в раковых геномах по сравнению с геномами, не имеющими мутаций.

В [8] исследовалось наличие неканонических структур ДНК в окрестностях разрывов. Анализ расстояния между мотивами, формирующими квадруплексы, и разрывами в нестабильных сайтах, ассоциирующихся с 11 генами, показал тесную взаимосвязь квадруплексов и разрывов почти в 70% генов, вовлеченных в геномные перестановки при лимфоидном раке.

Анализ почти 700 000 раковых разрывов из 26 типов рака показал, что регионы генома с разрывами обогащены квадруплексами с помощью теста Манна-Уитни [9]. Используя тот же метод, была продемонстрирована зависимость между хотспотами разрывов и гипометилированием.

Комплексный статистический анализ разрывов (транслокаций и делеций) в раковых геномах подтвердил значительную взаимосвязь между разрывами и неканоническими структурами ДНК для большого набора данных (около 20 000 транслокаций и 46 000 делеций) [10]. С помощью t-теста Стьюдента сравнивалось количество потенциальных

неканонических структур в средних секциях транслокаций и делеций и было показано, что повторы могут часто наблюдаться в транслокациях, в то время как поли(A)-хвост – в делециях.

Статистический анализ частотности мест связывания белка и открытого хроматина был проведён для выборки из 147 сэмплов, представляющих 8 типов рака и 14 600 структурных мутаций [11]. Для анализа были использованы 457 ChiP-seq экспериментов из базы ENCODE, 125 DNase I и 24 FAIRE эксперимента. Было найдено, что окрестности разрывов обогащены местами связывания белка и открытого хроматина, что было показано с помощью двустороннего t-теста для логарифма "отношения шансов" для регионов далеких и близких от разрывов.

Полногеномное сравнение распределения ультрафиолетовых повреждений ДНК с распределениями эпигенетических признаков было проведено для рака кожи [12]. В статье рассматривались отклонения распределения УФ-повреждений ДНК от медианы по геному для разных состояний хроматина, а также корреляция с модификациями гистонов, с помощью чего было показано, что гетерохроматин более уязвим для ультрафиолетовых повреждений ДНК.

Методы машинного обучения для предсказания раковых точечных и структурных мутаций

Одним из первых комплексных исследований раковых точечных мутаций методами машинного обучения было исследование [4], которое включало данные модификаций гистонов, сайты связывания CTCF и Pol-II, скорость рекомбинации, скорость репликации, позиции нуклеосом, плотность генов и уровень консервации. С помощью линейной регрессии авторы предсказывали плотность раковых точечных мутаций на уровне 1 Мб и достигли значения 55% для коэффициента детерминации R^2. Используя анализ важности признаков, было установлено, что один единственный признак – модификация гистона H3K9me3 – объясняет 40% дисперсии плотности точечных мутаций.

В другом исследовании [5] модель на основе алгоритма случайного леса смогла объяснить до 86% дисперсии плотности точечных раковых

мутаций для различных типов рака. Было обнаружено, что тканеспецифичные эпигенетические признаки (доступность хроматина, модификации гистонов и время репликации) улучшает качество моделей. Моделирование также показало, что может быть решена и обратная задача – профиль плотности мутаций может помочь в определении типа рака.

Предсказание раковых точечных мутаций оказалось более легкой задачей, чем предсказание раковых разрывов. В [13] авторы использовали линейную регрессию и алгоритм случайного леса для предсказания плотности как раковых точечных мутаций, так и разрывов на уровне 500 Кб, используя данные о неканонических структурах ДНК, модификациях гистонов и времени репликации, как отдельно, так и совместно. В зависимости от типа рака комбинация неканонических структур ДНК и эпигенетических факторов объяснила 43-76% дисперсии плотности точечных мутаций, в то время как использование всех рассмотренных признаков смогло объяснить только 10% дисперсии плотности разрывов для всех типов рака за исключением рака груди (18%).

Линейная регрессия частоты разрывов-транслокаций совместно для различных типов лейкемии [14] по данным плотности хроматина, генов и сайтов связывания CTCF показала качество на уровне 18-39% скорректированного R^2, где плотность хроматина оказалась наиболее важным фактором.

Взаимосвязь между разрывами и участками генома, обогащенными генами, была исследована в [15], где с помощью линейной регрессии по количеству генов предсказывалось количество разрывов. Авторы получили 40% R2 и показали, что взаимосвязь является значимой как для повторяющихся, так и для неповторяющихся разрывов.

В [18] авторы предложили подход на основе алгоритмов машинного обучения для предсказания двунитевых разрывов, которые были сгенерированы с помощью методов DSBcapture [16] и BLESS [17]. Используя такие признаки, как плотности гистоновых меток, данные DNase-seq, параметры структуры ДНК, сайты связывания CTCF и p63 на уровне 1 Кб, модель случайного леса, предсказывающая, является ли

локация двунитевым разрывом или нет, достигла 97% ROC AUC, однако такую завышенную точность можно объяснить спецификой использованного метода генерации двунитевых разрывов.

В [19] авторы изучали зависимость между экспрессией генов и метилированием CpG участков вблизи разрывов. Используя линейную регрессию, они обнаружили, что наличие разрыва в окрестности до $\pm$ 1 пар оснований изменяет метилирование.

2. Предлагаемый подход для предсказания раковых разрывов с помощью методов машинного обучения

Предыдущие попытки предсказания раковых разрывов показали, что эта задача не решается стандартными способами. Поэтому была поставлена цель разработать подход на основе машинного обучения, который смог бы учесть и справиться со всеми недостатками предыдущих подходов.

Предобработка данных: выбор уровня агрегации

Для исследования [20] мы использовали общедоступные данные из the International Cancer Genome Consortium (ICGC) Data Portal (версия 25)[21]. В данных были представлены 10 типов рака (груди, кости, мозга, крови, простаты, кожи, поджелудочной железы, печени, яичников, матки). Набор данных содержал всего 2 234 полных генома и 487 425 разрывов, при этом наибольшее количество геномов было доступно для рака груди (644) в то время, как для рака мозга и рака матки – всего лишь 72 и 16 геномов соответственно.

В качестве признаков для предсказания раковых разрывов мы рассмотрели наиболее часто встречающиеся неканонические структуры ДНК – структуру стебель-петля и квадруплексы – которые являются известными источниками хромосомной нестабильности. Так как первостепенной целью является разработка и реализация подхода на основе машинного обучения для задачи предсказания раковых разрывов, мы ограничили набор признаков этими двумя потенциально важными

факторами. Аннотация генома человека структурой стебель-петля (трех размеров – 6-15, 15-30, 16-50 нуклеотидов) была получена с ресурса «ДНК пунктуация» [22], в то время как аннотация генома квадруплексами была получена с помощью регулярных выражений [23]. Входные данные (как признаки, так и целевая переменная), были представлены в табличном формате, где одна строка описывает один экземпляр исследуемого объекта с отметкой стартовой и конечной позиции этого экземпляра в геноме. Для поиска паттернов в данных мы провели агрегацию данных: весь геном был разбит на непересекающиеся окна фиксированной длины. Так как нет априорных знаний об оптимальной длине окна, мы рассмотрели 6 вариантов (далее называемые уровни агрегации): 10, 20, 50, 100, 500 Кб и 1 Мб. Далее данные были агрегированы до этих уровней, так что для каждого экземпляра было найдено соответствующее окно, к которому он относится исходя из его стартовой и конечной позиций, после чего для каждого окна была посчитана плотность разрывов и покрытие неканоническими структурами. Плотность разрыва в данном окне была рассчитана как отношение количества разрывов в этом окне к общему количеству разрывов в геноме. Покрытие признаком в данном окне равнялось отношению суммарной длины всех структур в этом окне (без перекрытий) к длине окна.

Так как рак является гетерогенным заболеванием, очень важно выявить общие паттерны, присущие множеству сэмплов. Ранее было введено понятие повторяющихся разрывов [24], где авторы зафиксировали набор наиболее часто встречаемых раковых разрывов. В качестве альтернативного подхода в данной работе мы ввели понятие областей с повышенной плотностью разрывов, руководствуясь следующими соображениями: для каждого типа рака были найдены перцентили распределения плотности разрывов (1%, 0.5%, 0.1%, 0.05% и 0.01%), и для каждого перцентиля (называемого далее тип маркировки) геномные окна были промаркированы как области с повышенной плотностью разрывов, если значение плотности разрывов превышает или равно этому порогу. После анализа количества положительных примеров (областей с повышенной плотностью разрывов) для каждого

типа рака, уровня агрегации и типа маркировки мы получили итого 236 наборов данных для моделирования, при этом наборы данных с маленьким количеством положительных примеров были исключены из исследования. Стоит отметить, что с точки зрения машинного обучения во всех указанных типах маркировки задача классификации областей с повышенной плотностью разрывов является крайне несбалансированной, что необходимо учесть при разработке подхода.

В дополнение к указанным типам рака мы рассмотрели так называемый «общий» раковый профиль, используя плотности разрывов для каждого типа рака и мировую статистику о распространенности этих типов рака в качестве весов для типов рака.

Работа с экстремальной несбалансированностью классов

Для адаптации подхода к работе с несбалансированными классами в выборке мы протестировали несколько схем сэмплирования выборок, алгоритмов машинного обучения и техник балансировки классов. Так как в этом случае наиболее важной метрикой качества являются полнота и точность, мы использовали их среднее гармоническое F1 для оценки качества модели. Конечной целью перебора являлось достижение минимального переобучения, максимального качества на тестовой выборке и минимальная дисперсия метрик качества на тестовых выборках. К описанным данным были применены следующие методы:

- Схемы сэмплирования выборок: разбиение на обучающую и тестовую выборки (50%), leave-one-out кросс-валидация (LOOCV), 15 раз повторенная кросс-валидация с 3 фолдами. Был выбран последний метод, так как он позволяет получить распределение метрик качества с оценкой наихудшего, наилучшего и среднего качества модели.

- Алгоритм машинного обучения: логистическая регрессия, случайный лес. Случайный лес (со стратификацией) в случае с высоким дисбалансом классов показывает большее переобучение, чем логистическая регрессия.

- Техники балансировки классов: стратификация, пересэмплирование (oversampling), SMOTE. По сравнению со стратификацией пересэмплирование повысило качество классификации областей с повышенной плотностью разрывов, в то время как SMOTE не показал прироста.

По окончании экспериментов пайплайн для обучения моделей был финализирован на следующих компонентах: логистическая регрессия с пересэмплированием, повторно обученная 15 раз на кроссвалидации с 3мя фолдами (частями выборки) и z-score нормализацией признаков. Данный пайплайн был применен к каждому из 236 датасетов.

Прирост полноты и точности как метрики для оценки качества модели

Для оценки качества модели классификации в случае несбалансированных классов обычно используются такие метрики, как полнота и точность. Так как только 0,01 - 1% выборки представлен положительными примерами, задача классификации является сложной. Для оценки того, целесообразно ли использование модели машинного обучения, мы ввели такие производные метрики, как прирост полноты и прирост точности. Если случайный алгоритм размечает n% выборки в качестве положительных примеров, тогда полнота модели будет также приблизительно равна n% (случайный выбор n% примеров из выборки в среднем приводит к выбору n% выборки от каждого класса). Тогда мы можем оценить, лучше ли модель машинного обучения, чем случайный алгоритм, найдя отношение полноты модели к перцентилю вероятности (доле помеченных моделью положительных примеров от всей выборки). Это отношение называется далее приростом полноты и интерпретируется следующим образом: если прирост полноты больше единицы, это означает, что модель машинного обучения превосходит случайный алгоритм, и чем выше это значение, тем больше предсказательная сила модели. Иначе, если прирост полноты менее единицы, это означает, что модель машинного обучения не может уловить паттерны в данных. Похожим образом прирост полноты

рассчитывается как отношение точности модели к пропорции положительных примеров в выборке.

В дополнение к среднему и медианному ROC AUC на тестовой выборке мы приводим показатели прироста полноты для различных порогов в соответствии с фиксированными перцентилями распределения вероятности 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 97%, 99%.

Выбор лучшей модели для каждого типа рака

Предложенный подход был применен ко всем 236 наборам данных для 3 наборов признаков: только структуры стебель-петля, только квадруплексы, квадруплексы и структуры стебель-петля. Было замечено, что влияние этих признаков тканеспецифично: нет такого набора признаков, который демонстрировал наилучшие результаты для всех типов рака, хотя направление влияния признаков (положительное или отрицательное) сохранялось. Таблица 1 показывает метрики качества лучших моделей среди различных уровней агрегации, типов маркировки и спецификаций модели для каждого типа рака. Таблица показывает, что наилучшее качество удалось достичь для рака груди и костей с ROC AUC 0.94 и 0.86 и приростом полноты 8 и 10 соответственно. В то же время нашлись типы рака с относительно низкими метриками качества: ROC AUC 0.63 и 0.61 и прирост полноты 2.05 и 1.71 для рака простаты и поджелудочной железы.

Тип рака	**Количество геномов**	**Количество разрывов**	**Количество наборов данных**	**Прирост полноты (лучшая модель)**	**Медианный тестовый ROC AUC (лучшая модель)**
Мозга	72	1 564	20	5,00	67%
Крови	118	2 330	20	4,00	67%
Костей	117	2 546	20	10,00	86%
Матки	16	6 782	19	4,00	65%
Печени	255	22 324	21	5,71	73%

Простаты	212	48 126	23	2,05	55%
Кожи	190	54 688	23	4,00	64%
Яичников	115	71 446	22	6,67	68%
Поджелудочной железы	495	85 769	22	1,71	57%
Грудь	644	191 850	23	10,00	94%
Общий профиль			23	6,67	72%

Таблица 1 Статистики наборов данных: тип рака, количество геномов, количество разрывов, количество наборов данных для моделирования, лучшие метрики качества (для всех уровней агрегации и типов маркировки).

Разработанный подход на основе машинного обучения позволил увидеть тканеспецифичность воздействия рассматриваемых признаков. Протестированный на небольшом количестве признаков, данный подход может быть применен к более широкому набору факторов для обогащения и углубления исследования.

3. Предсказание раковых разрывов на основе омиксных данных

Использование омиксных данных для предсказания раковых разрывов

С появлением технологий секвенирования омиксные данные стали ценным источником информации для моделей машинного обучения. Омиксные данные - это совокупность данных разных научных направлений – транскриптомики, эпигеномики, метаболомики, липидомики и других, которые позволяют квантифицировать характеристики, описывающие функциональные свойства генома. Подходы машинного обучения, учитывающие различные омиксные факторы, могут помочь в более системном понимании генов-драйверов, определяющих развитие рака. До настоящего исследования только две

группы признаков - неканонические структуры ДНК и модификации гистонов - были протестированы в качестве предикторов на большом наборе данных. Добавление других групп омиксных данных в модель машинного обучения может помочь в поиске большего количества или более сильных факторов, влияющих на формирование раковых разрывов [25].

Так как ранее не было исследований, которые учитывали бы максимально доступные признаки для предсказания раковых разрывов, мы осуществили такое широкомасштабное исследование, используя разработанный нами ранее подход на основе машинного обучения для получения более высокого качества моделей по сравнению с предыдущими результатами. Наше исследование [26] включало такие признаки как неканонические структуры ДНК (квадруплексы, стебель-петля, повторы, Z-ДНК), модификации гистонов (HM), метилирование ДНК, транскрипционные факторы (TF), доступность хроматина (HDNase), топологически ассоциированные домены (TAD) и регионы генома (гены, экзоны, интроны, нетранслируемые области).

Включение трансформаций признаков в модель

Исследование было ограничено одним уровнем агрегации генома - 100 КБ, который является наиболее распространенным в предыдущих исследованиях и продемонстрировал в прошлом высокое качество моделей. Для признаков помимо стандартного расчета покрытия в геномном окне (далее “локальные признаки”) были рассчитаны различные трансформации данных признаков для проверки возможности улучшения качества модели: бинарные флаги наличия признака в окне, индикаторы локального (1-10 соседей) и глобального максимума значения покрытия признака в окне и «глобальные признаки» (покрытие окна 1 Мб). Добавление флагов наличия признака не дало никакого улучшения, в то время как модель, использующая только эти признаки, продемонстрировала значительное снижение качества (~ -0,13 ROC AUC в среднем по всем типам рака), которое могло быть лишь частично компенсировано добавлением индикаторов максимума (~ -0,03 ROC AUC в среднем), что указывает на то, что

точное значение покрытия признаками играет значительную роль в предсказании скоплений разрывов. Напротив, добавление глобальных признаков к локальным продемонстрировало среднее повышение ROC AUC на 0,03 в целом, хотя эффект немного отличается для разных типов рака. Это повышение качества может быть объяснено тем фактом, что сочетание глобальных и локальных признаков позволяет модели оценить "аномальность" каждого окна с точки зрения покрытия признаков относительно его ближайшего окружения. Основываясь на данных результатах, глобальные признаки были добавлены в финальный набор признаков для построения моделей. Более того, как показано ниже, глобальные признаки вошли в число наиболее важных признаков для моделей.

Результаты

В исследовании [26] мы немного изменили пайплайн для построения моделей машинного обучения. Во-первых, мы заменили схему семплирования с 15-кратной кросс-валидации с тремя частями на более естественную, но с той же механикой действия схему 30-кратного повторного разбиения выборки на обучающую и тестовую с 30% данных в тестовой выборке. Во-вторых, мы заменили модель логистической регрессии моделью случайного леса, поскольку набор признаков увеличился и стал более разнообразным. Используя этот пайплайн машинного обучения и омиксные данные, представленные в виде локальных и глобальных признаков, мы построили модели для предсказания скоплений раковых разрывов (99% / 99.5% / 99.9%) на уровне агрегации 100 кб.

Финальное качество моделей для каждого типа рака представлено в Таблице 2. Продемонстрированные результаты можно сравнить с двумя существующими работами по прогнозированию раковых разрывов: нашей предыдущей работой [20] и работой Georgakopoulos-Soares et al. [13].

По сравнению с нашим предыдущем исследованием [20] – как показано в Таблице 2 – модели на основе омиксных данных продемонстрировали более высокое качество для всех типов рака, кроме

рака костей, как по медиане ROC AUC на тестовой выборке, так и по аплифту полноты. Ранее медиана тестового ROC AUC превышала 70% только для рака костей, в то время как модели на основе омиксных данных достигли 69-86 % медианы тестового ROC AUC для всех типов рака, кроме рака кожи и костей. Кроме того, медиана аплифта полноты также значительно увеличилась: медианное значение этого показателя по всем типам рака составляет 2,6 в абсолютном выражении или +77,5% относительно, хотя PR AUC все еще остается относительно низким - от 0,3% до 4,8%.

Другим сопоставимым исследованием было исследование Georgakopoulos-Soares et al. [13], в котором авторы решали задачу регрессии по прогнозированию плотности раковых разрывов по неканоническим структурам ДНК и модификациям гистонов. Используя предоставленные авторами данные, мы, во-первых, воспроизвели их исследование, а во-вторых, добавили наши признаки (локальные и глобальные признаки, полученные по всем имеющимся омиксным данным) в модель прогнозирования плотности разрывов в разрезе геномных окон длиной 500 кб, используя такие же преобразования признаков и целевой переменной, как в оригинальной работе [13]. Оригинальная модель случайного леса достигла максимального значения 18% R^2, в то время как за счет пополнения исходных данных в источниках (из-за временной разницы в проведении исследований) и добавления омиксных данных удалось улучшить этот показатель до 34%.

Тип рака	**Омиксная модель Прирост полноты (лучшая модель, 0.03 перцентиль вероятности)**	**Омиксная модель Медиана ROC AUC на тестовой выборке (лучшая модель)**	**Омиксная модель Средний PR AUC на тестовой выборке (лучшая модель)**	**Модель на неканонических структурах ДНК, 100 Кб Прирост полноты (лучшая модель, 0.03 перцентиль вероятности)**	**Модель на неканонических структурах ДНК, 100 Кб Медиана ROC AUC на тестовой выборке (лучшая модель)**

Крови	5,7	75%	0,3%	2,5	65%
Костей	5,1	64%	1,8%	6,0	80%
Мозга	8,0	75%	0,5%	5,0	67%
Груди	7,6	86%	0,6%	6,7	65%
Печени	7,8	73%	0,6%	4,0	66%
Яичников	5,0	69%	2,9%	2,1	59%
Поджелудочной железы	16,7	76%	4,8%	1,7	57%
Простаты	4,3	73%	0,4%	2,0	56%
Кожи	2,6	57%	1,5%	2,2	56%
Матки	4,0	69%	1,3%	4,0	62%

Таблица 2 Качество моделей предсказания областей с повышенной плотностью раковых разрывов с помощью омиксных данных и сравнение с предыдущими результатами [20] для моделей на уровне агрегации 100 Кб.

4. Подход к анализу важности омиксных признаков

Так как в исследовании [26] было использовано большое количество разных (по функции в геноме) признаков, то важной задачей является анализ важности признаков для определения ключевых факторов, влияющих на формирование областей с повышенной плотностью раковых разрывов. С этой целью сначала было оценена важность групп признаков, после чего был проведен анализ для определения наиболее важных индивидуальных признаков.

Важность групп признаков

Рассматриваемые в исследовании [26] признаки могут быть объединены в следующие группы по своему происхождению: неканонические структуры ДНК (non-B), модификации гистонов (HMs), метилирование, транскрипционные факторы (TF), доступность хроматина (HDNase), топологически ассоциированные домены (TADs), и геномные регионы. Для изучения влияния каждой группы признаков на качество предсказания скоплений разрывов мы обучили модели

машинного обучения отдельно на каждой группе признаков. Для ранжирования групп признаков было найдено максимальное значение среднего прироста полноты для перцентиля вероятности 0.03 среди всех моделей для каждого типа маркировки и типа рака, а затем значение среднего прироста полноты для каждой группы признаков было нормализовано на это значение. Так как результаты немного отличались для типов маркировки 99% и 99.5%, то данный коэффициент был усреднен на уровне типа рака. Результаты (Рис.1) показали, что лучшая (по качеству модели) группа признаков значительно (на 25%) превосходит другие группы признаков почти для всех типов рака, и это значение достигает 50% для 3 видов рака (крови, поджелудочной железы и предстательной железы). В качестве лучшей группы признаков для 5 видов рака (печени, кожи, предстательной железы, яичников, молочной железы) выступают транскрипционные факторы, неканонические вторичные структуры ДНК - для 2 видов рака (головного мозга и костей). Неканонические вторичные структуры ДНК всегда входят в топ-3 важных групп для всех видов рака, транскрипционные факторы присутствуют в списке топ-3 в 8 типах рака (за исключением крови и головного мозга). Если в качестве метрики качества использовать ROC AUC, то результаты окажутся похожими. Кроме того, из обученных моделей была получена важность признаков для лучших групп признаков - неканонических структур и транскрипционных факторов. Мы установили, что среди всех неканонических структур квадруплексы и прямые повторы являются наиболее влиятельными факторами для всех типов рака, в то время как короткие повторы, дубликаты и Z-ДНК также вносят вклад в повышение качества модели. Что касается факторов транскрипции, то наиболее важные признаки разнятся для разных типов рака, однако CTCF, GABPA, RXRA, SP1, MAX и NR2F2 чаще присутствуют в наборе важных признаков.

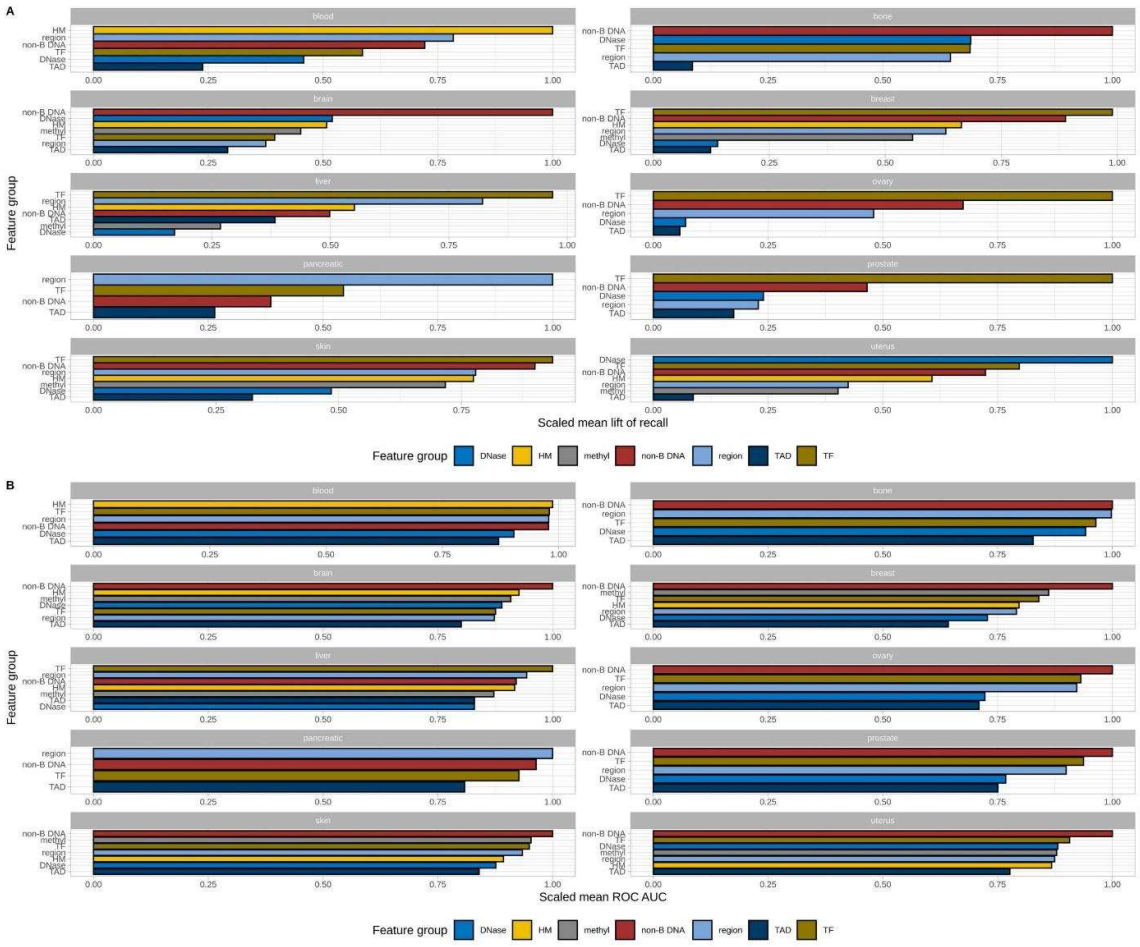

Рисунок 1. Ранжирование групп признаков по качеству модели на основе двух метрик качества - прироста полноты и ROC AUC. **A.** Средний прирост полноты для перцентиля вероятности 0.03 для каждого типа рака и группы признаков, отмасштабированный и усреднённый для типов маркировки 99% и 99.5%. **B.** Средний ROC AUC для каждого типа рака и группы признаков, отмасштабированный и усреднённый для типов маркировки 99% и 99.5%.

Важность индивидуальных признаков

Для того, чтобы определить наиболее важные признаки для предсказания областей с повышенной плотностью раковых разрывов для каждого типа рака, была проведена процедура отбора признаков методом Борута. В результате было найдено итого 50 важных признаков, где для каждого типа рака было выбрано от 5 до 23 признаков (Рис.2). Данный список включает, по большей части, признаки из группы неканонических структур ДНК, транскрипционных факторов и геномных регионов, что коррелирует с результатами, полученными при анализе важности групп признаков.

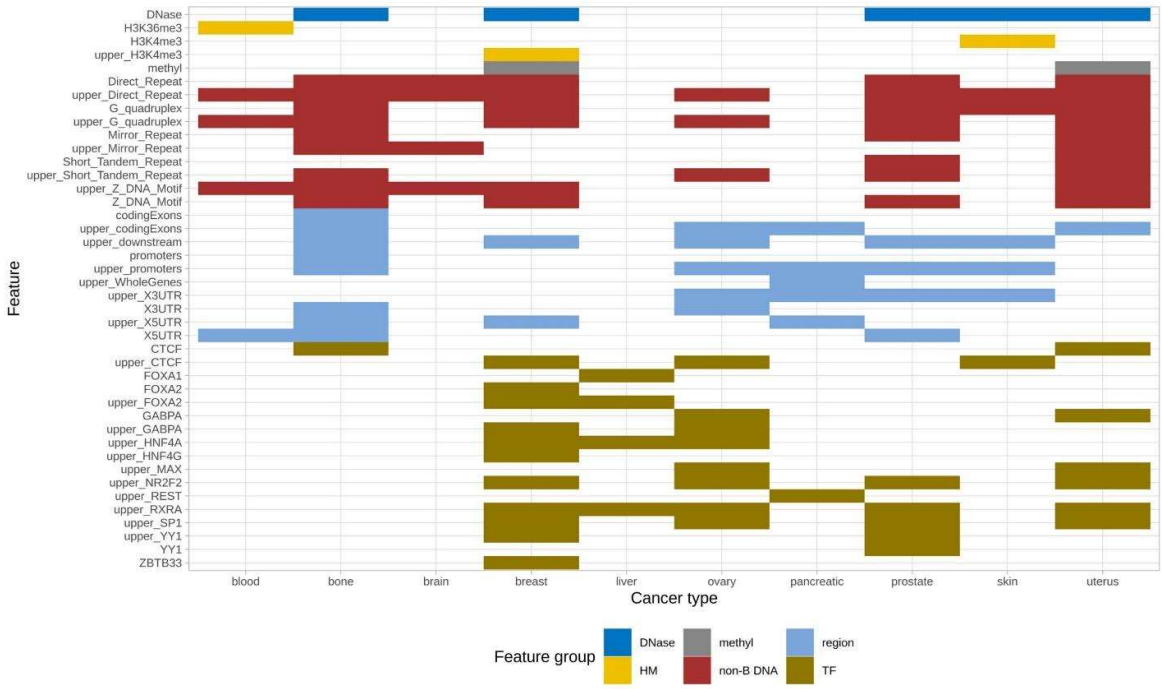

Рисунок 2. Отобранные с помощью метода Борута наборы признаков для каждого типа рака.

В предыдущем исследовании [27] нами было показано, что нет такого единого порога плотности разрывов для идентификации областей с повышенной плотностью раковых разрывов (типа маркировки), который показал бы наилучшие результаты для всех типов рака. Более того, было продемонстрировано, что чем выше порог плотности разрывов, тем выше дисперсия модели машинного обучения для предсказания таких областей. По этой причине анализ важности индивидуальных признаков с помощью метода Борута был проведен для типа маркировки, использующего самый низкий из рассматриваемых порог плотности 99% для получения наиболее стабильных результатов.

Для каждого из 30-и разбиений на обучающую и тестовую выборку для каждого типа рака был применен следующий алгоритм отбора признаков.

Метод является итеративным и на каждой итерации рассматривается только набор важных признаков, найденный на предыдущей итерации, при этом на первой итерации набор важных признаков составляют все признаки. На каждой итерации к набору важных признаков добавляется набор «теневых» (shadow) признаков. Данные теневые признаки

представляют собой перемешанные значения реальных признаков, таким образом, что для каждого важного признака есть один теневой признак, полученный перемешиванием его значений. Затем обучается модель случайного леса на этом расширенном датасете, и для каждого признака рассчитывается метрика важности (Mean Decrease Accuracy) и ее z-score. Новый набор важных признаков на этой итерации формирует те исходные признаки, которые имеют z-score, превышающий максимальный z-score теневых признаков. Алгоритм переходит к следующей итерации, если осталось более 5 признаков и проведено менее 10 итераций.

С помощью данного алгоритма были выбраны важные признаки для каждого типа рака для типа маркировки порогом плотности разрывов 99%. При анализе качества моделей, обученных на данных наборах важных признаков, были найдены 3 типа рака (поджелудочной железы, простаты и груди), для которых получено более низкое качество, чем на наборе всех доступных признаков. Для этих типов рака была проведена процедура аддитивного добавления признаков (forward feature selection) и определены 1,1 и 2 признака соответственно, добавление которых в модель позволяет получить сравнимое качество. Финальные наборы признаков для каждого типа рака представлены на Рис.2.

Данный анализ показывает, что признаки только четырех групп (неканонические структуры ДНК, транскрипционные факторы, геномные области и HDNase) являются наиболее важными по критерию попадания в список отобранных признаков по крайней мере в 300 из 3000 проанализированных наборов данных для всех типов рака по методу Борута (3000 = 10 типов рака * 30 разбиений на обучающую и тестовую выборки * 10 итераций). Топ-5 признаков составляют прямые повторы и квадруплексы – локальные и глобальные - и транскрипционный фактор SP1, далее идут такие признаки, как Z-DNA, короткие тандемные повторы, зеркальные повторы, транскрипционные факторы RXRA, NR2F2, GABPA, CTCF, геномные регионы 5' UTR, кодирующие экзоны, 3' UTR, промотеры, а также HDNase, чье влияние варьируется при сравнении разных типов рака. Стоит отметить, что большинство важных признаков представлены в виде глобальных признаков.

Результаты

Мы предложили подход к анализу важности омиксных данных и с его помощью определили, что неканонические структуры и транскрипционные факторы являются наиболее важными признаками для предсказания хотспотов раковых разрывов, что подтверждается как с помощью анализа важности индивидуальных признаков, так и их групп. Если выделить наиболее важные признаки, то наибольший вклад вносят такие признаки, как квадруплексы и повторы, а также транскрипционные факторы CTCF, GABPA, RXRA, SP1, MAX и NR2F2.

5. Подход к анализу случайности разрывов в раковых геномах

Из-за известной гетерогенности раковых геномов для того, чтобы найти зависимости в формировании раковых разрывов мы исследовали повторяющиеся разрывы - области с повышенной плотностью разрывов. Тем не менее, важно понимать, насколько хорошо предсказуемы индивидуальные разрывы и есть ли зависимость между порогом для плотности разрывов и качеством ML-модели для обнаружения соответствующих скоплений разрывов. Для ответа на этот вопрос была проведена серия экспериментов [26].

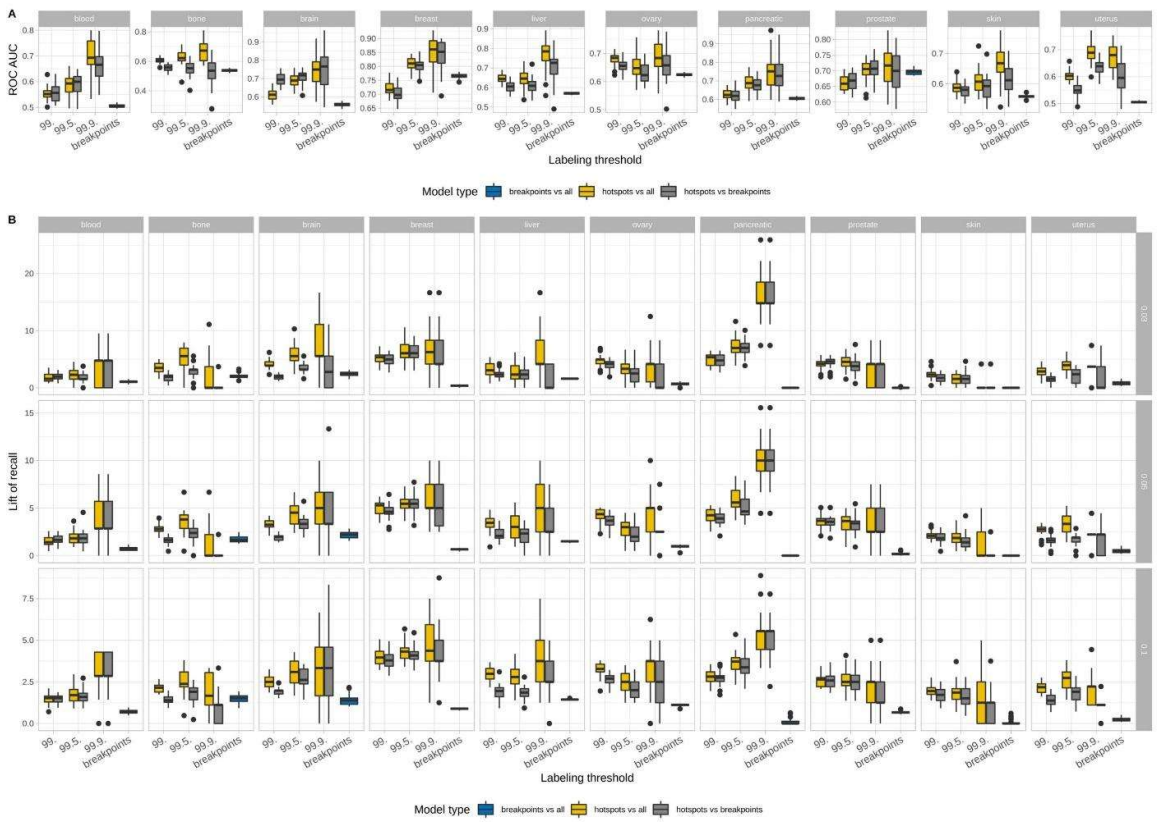

Рисунок 3. Сравнение моделей для предсказания индивидуальных разрывов и их областей с повышенной плотностью разрывов. Прирост полноты представлен для следующих перцентилей вероятности: 0.03, 0.05, 0.1

На первом этапе разработанный пайплайн был применён к задаче предсказания индивидуальных раковых разрывов (Рис.3). Результаты показывают, что для большинства видов рака индивидуальные разрывы неотличимы от случайных местоположений в геноме (ROC AUC около 50% или немного выше, а средний аплифт полноты от 0 до 2,5), однако для рака молочной железы, яичников и предстательной железы даже отдельные разрывы можно предсказать с ROC AUC 65-75 % (тем не менее, прирост полноты остается очень низким).

На втором этапе были построены модели машинного обучения, различающие хотспоты разрывов от остальных регионов генома, содержащих разрывы. Кроме того, модели, которые могут отличать скопления разрывов от индивидуальных разрывов, по качеству сравнимы с моделями, предсказывающими скопления разрывов в геноме, и достигают 85% ROC AUC для рака молочной железы и более 70% для мозга, печени, поджелудочной железы и простаты, подтверждая, что местоположения скоплений разрывов сильно отличаются от

местоположений отдельных разрывов по рассматриваемым признакам. Результаты представлены на Рис. 3.

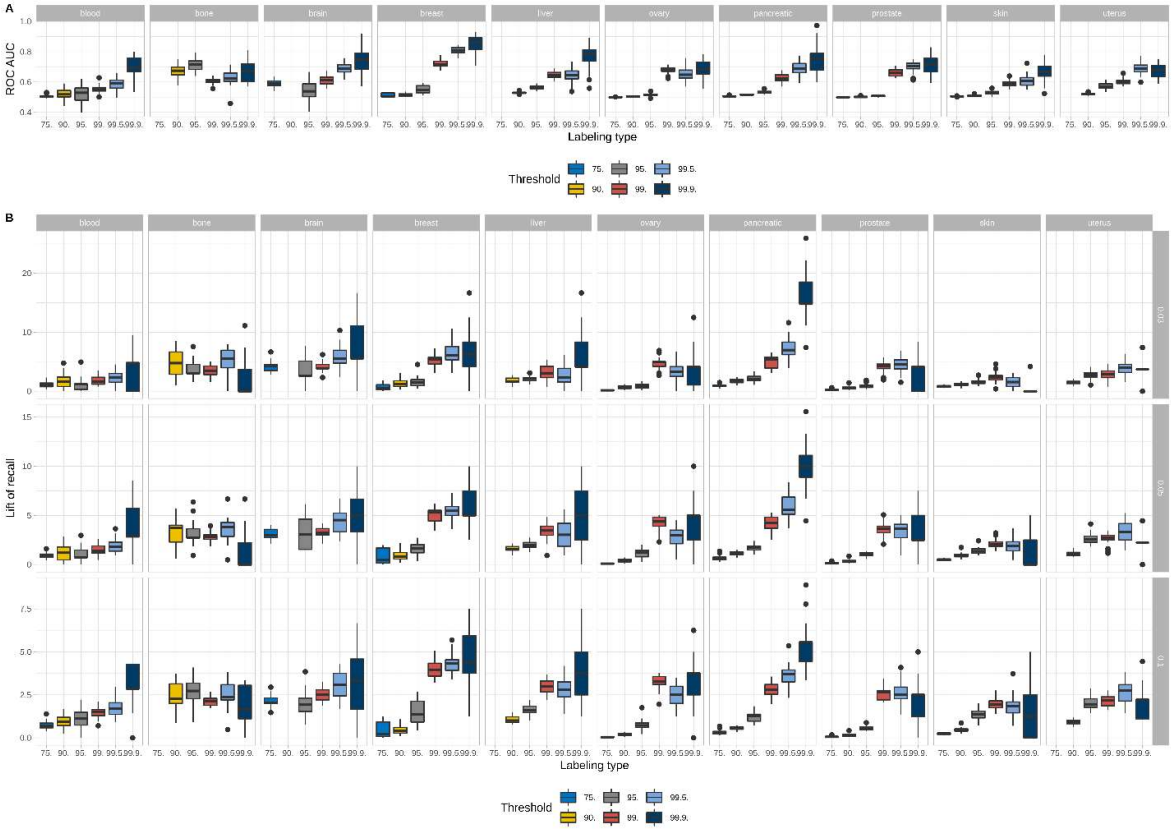

Рисунок 4. Сравнение различных критериев маркировки хотспотов разрывов. Прирост полноты представлен для следующих перцентилей вероятности: 0.03, 0.05, 0.1

На последнем этапе мы протестировали другие пороги - 75%, 90%, 95% - для плотности разрывов для маркировки областей с повышенной плотностью разрывов и построили модели машинного обучения для обнаружения регионов генома, менее насыщенных разрывами, чем ранее, среди всех остальных регионов (Рис. 4). Было показано, что увеличение порога плотности разрывов для маркировки областей с повышенной плотностью приводит к повышению качества модели предсказания соответствующих областей. В среднем, понижение порога на 5% связано с уменьшением прироста полноты в два раза и ROC AUC на 15%. Кроме того, для низких порогов плотности областей с повышенной плотностью разрывов абсолютные значения ROC AUC достигают 60% только для рака кости, в то время как среднее по всем типам рака находится на уровне 54%, со средним приростом полноты 1,6. Исходя из этого, можно сделать вывод о том, что области более высокой плотности разрывов

больше отличаются от других участков генома по изучаемым признакам, чем области меньшей плотности.

6. Учёт достаточности данных в модели

В случае, когда целевая переменная зависит от распределения определённой переменной, могут возникать проблемы с достаточностью выборки. Например, если данные о раковых разрывах нерепрезентативны, то разметка генома областями повышенной плотности разрывов может быть неправильной, так как из-за недостаточности данных плотность разрывов может быть недооценена, и тогда некоторые области будут не размечены как таковые. В этом случае можно добавить в модель информацию об этой неуверенности с помощью подхода PU-learning (обучение на позитивных и неразмеченных примерах) [28, 29].

Для проверки достаточности выборки был использован метод PU-Learning. Подход предполагает, что в выборке есть примеры, помеченные как положительные, а все остальные примеры имеют неизвестную метку (могут быть положительными или отрицательными). Задача состоит в том, чтобы присвоить метку всем неразмеченным примерам с учетом известных положительных меток и знания распределений признаков. В статье мы использовали алгоритм PU-learning [28, 29] и применили этот алгоритм для всех наборов данных для каждого типа рака.

Общая идея алгоритма состоит в следующем. Первично обучается модель классификации, в которой все неразмеченные примеры выступают в качестве отрицательных. Далее истинные метки неразмеченных примеров итеративно обновляются до сходимости. На каждой итерации текущей моделью формируется предсказание для всех примеров и на основе 10 и 90 перцентилей распределения вероятности положительного класса, а также ширины интервала определенности epsilon формируются границы, такие что если вероятность примера выше верхней границы, то пример относится к надежно-положительным примерам (reliable positives, RP), а если ниже нижней границы – к

надежно-отрицательным примерам (reliable negatives, RN). На каждой итерации для обучения модели используются только надежно-размеченные примеры и исходные известные положительные примеры, и процесс проводится до сходимости. Алгоритм имеет гиперпараметр epsilon, определяющий ширину интервала определенности. Данный алгоритм был реализован в 2 режимах: RP (итеративно обновляется и набор положительных, и набор отрицательных примеров) и RN (итеративно обновляется только набор отрицательных примеров, набор положительных примеров зафиксирован как набор исходных известных положительных примеров).

Данный подход был применён ко всем наборам данных для каждого типа рака в исследовании [30]. На Рис.5 для каждого типа рака представлен доверительный интервал для средней разницы в приросте полноты для режима RP и RN. Так как качество моделей обоих типов было оценено на одной и той же исходной тестовой выборке, положительная разница в приросте полноты для RP и RN на тестовой выборке означает, что смещенные распределения признаков для областей с повышенной плотностью разрывов (полученные в процессе PU-обучения) лучше описывают тестовую выборку, чем изначальные распределения признаков для этих же областей, поэтому добавление дополнительных положительных примеров даёт хороший сигнал для идентификации областей с повышенной плотностью разрывов. Было замечено, что знак различия зависит от количества разрывов, доступных для типа рака. С одной стороны, стабильный положительный эффект наблюдается для типов рака, которые входят в топ-5 типов рака с наименьшим количеством разрывов. В частности, наилучшие результаты достигнуты для рака головного мозга, который имеет минимальное количество разрывов. На основе этих данных можно сделать вывод о том, что включение дополнительных положительных примеров в случае зашумленных данных (разметка целевой переменной может быть зашумленной из-за нерепрезентативных данных о разрывах) помогает повысить качество модели с помощью PU-learning. С другой стороны, стабильный негативный эффект наблюдается для типов рака, которые входят в топ-4 типа рака с максимальным количеством

разрывов. Это может быть объяснено тем фактом, что в случае достаточного количества данных дополнительные положительные примеры вносят шум.

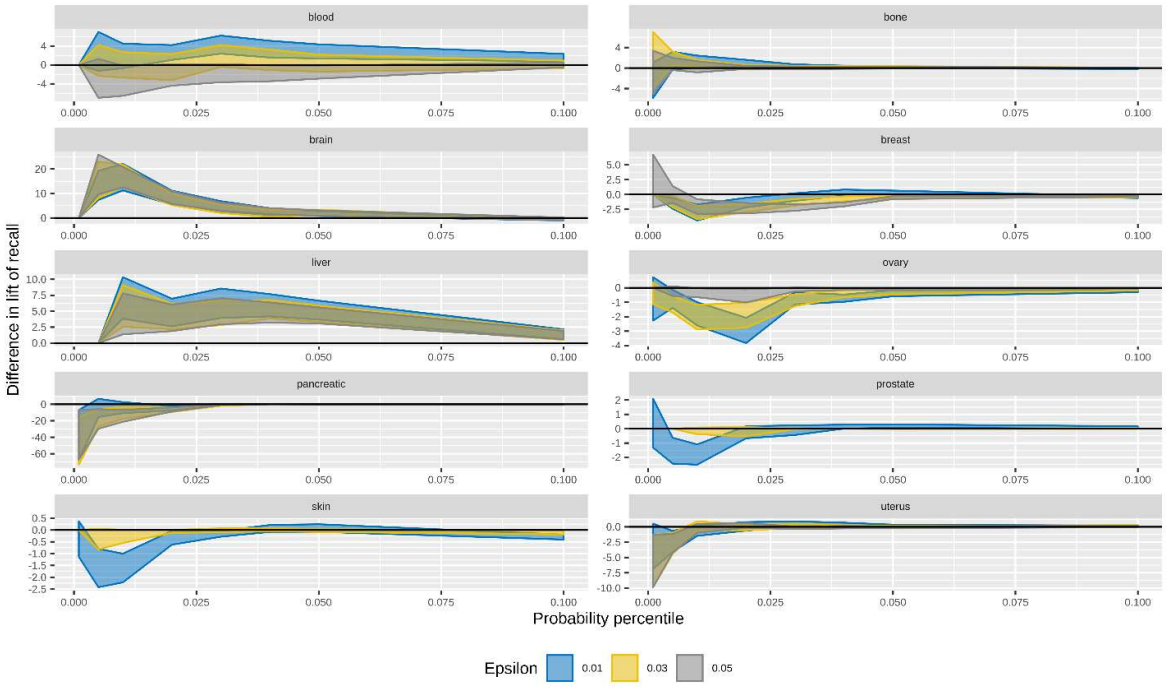

Рисунок 5. Доверительные интервалы для средней разницы в приросте полноты между RP и RN типами для алгоритма PU-learning для различных перцентилей вероятности.

Тем не менее, PU-learning не смог улучшить результаты, полученные с помощью обучения классической модели бинарной классификации для рассматриваемых порогов вероятностей (0.03, 0.05, 0.1 перцентили распределения): результаты практически идентичны.

ВЫВОДЫ

В диссертации был разработан подход на основе машинного обучения для исследования областей с повышенной плотностью раковых разрывов. Данная диссертация внесла вклад в область исследования генетики рака, определив ключевые факторы, связанные с образованием в геноме областей с повышенной плотностью разрывов, и исследовав случайность их возникновения.

Данные о распределении раковых разрывов по 10 типам рака совместно с большим набором омиксных данных позволили провести комплексный анализ областей с повышенной плотностью разрывов. Был

разработан и имплементирован подход на основе машинного обучения для задачи предсказания областей с повышенной плотностью раковых разрывов. Был проведён анализ важности признаков и обнаружены две группы признаков - неканонические структуры ДНК и транскрипционные факторы - важные для предсказания хотспотов для всех типов рака. Была проведена серия экспериментов, направленная на изучение случайности появления раковых разрывов. Помимо этого, был протестирован подход для учёта в модели неуверенности в данных с помощью PU-learning.

Разработанные модели, обученные на омиксных данных, показывают самое высокое качество предсказания областей с повышенной плотностью раковых разрывов, из всех известным нам исследований.

СПИСОК ЛИТЕРАТУРЫ

1. Nakagawa, H., et al. (2015). "Cancer whole-genome sequencing: present and future." Oncogene 34(49): 5943-5950.

2. Nakagawa, H. and M. Fujita (2018). "Whole genome sequencing analysis for cancer genomics and precision medicine." Cancer Sci 109(3): 513-522.

3. Consortium, I. T. P.-C. A. o. W. G. (2020). "Pan-cancer analysis of whole genomes." Nature 578(7793): 82-93.

4. Schuster-Böckler B, Lehner B (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature 488: 504-507.

5. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, et al. (2015) Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature 518: 360- 364.

6. Supek F, Lehner B (2015) Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature 521: 81-84.

7. Guo YA, Chang MM, Huang W, Ooi WF, Xing M, et al. (2018) Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. Nat Commun 9: 1520.

8. Katapadi VK, Nambiar M, Raghavan SC (2012) Potential G-quadruplex formation at breakpoint regions of chromosomal translocations in cancer may explain their fragility. Genomics 100: 72-80.

9. De S, Michor F (2011) DNA secondary structures and epigenetic determinants of cancer genome evolution. Nat Struct Mol Biol 18: 950-955.

10. Bacolla A, Tainer JA, Vasquez KM, Cooper DN (2016) Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. Nucleic Acids Res 44: p. 5673-88.

11. Grzeda KR, Royer-Bertrand B, Inaki K, Kim H, Hillmer AM, et al. (2014) Functional chromatin features are associated with structural mutations in cancer. BMC Genomics 15: 1013.

12. García-Nieto PE, Schwartz EK, King DA, Paulsen J, Collas P, et al. (2017) Carcinogen susceptibility is regulated by genome architecture and predicts cancer mutagenesis. EMBO J 36: 2829-2843.

13. Georgakopoulos-Soares I, Morganella S, Jain N, Hemberg M, Nik-Zainal S (2018) Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. Genome Res 28: 1264-1271.

14. Lin CY, Shukla A, Grady JP, Fink JL, Dray E, et al. (2018) Translocation Breakpoints Preferentially Occur in Euchromatin and Acrocentric Chromosomes. Cancers (Basel) 10. [Crossref]

15. Mitelman F, Johansson B, Mertens F, Schyman T, Mandahl N (2019) Cancer chromosome breakpoints cluster in gene-rich genomic regions. Genes Chromosomes Cancer 58: 149-154. [Crossref]

16. Lensing SV, Marsico G, Hänsel-Hertsch R, Lam EY, Tannahill D, et al. (2016) DSBCapture: in situ capture and sequencing of DNA breaks. Nat Methods 13: 855- 857.

17. Crosetto N, Mitra A, Silva MJ, Bienko M, Dojer N, et al. (2013) Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. Nat Methods 10: 361-365. [Crossref]

18. Mourad R, Ginalski K, Legube G, Cuvier O (2018) Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution. Genome Biol 19: 34. [Crossref]

19. Zhang Y, Yang L, Kucherlapati M, Hadjipanayis A, Pantazi A, et al. (2019) Global impact of somatic structural variation on the DNA methylome of human cancers. Genome Biol 20: 209. [Crossref]

20. Cheloshkina, K., Poptsova, M. (2019). Tissue-specific impact of stem-loops and quadruplexes on cancer breakpoints formation. BMC cancer, 19(1), 1-17.

21. The Cancer Genome Atlas (TCGA). Available from: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

22. DNA Punctuation Project. Available from: http://www.dnapuncutation.org

23. Huppert JL, Balasubramanian S. Prevalence of quadruplexes in the human genome. Nucleic Acids Res. 2005;33(9):2908–16

24. Mitelman F, Johansson B, Mertens F. Mitelman database of chromosome aberrations and gene fusions in cancer. 2019. http://cgap.nci.nih.gov/Chromosomes/Mitelman

25. Cheloshkina K, Poptsova M (2020) Understanding cancer breakpoint determinants with omics data. Integr Cancer Sci Therap 7: DOI: 10.15761/ICST.1000333

26. Cheloshkina, K., & Poptsova, M. (2021). Comprehensive analysis of cancer breakpoints reveals signatures of genetic and epigenetic contribution to cancer genome rearrangements. PLOS Computational Biology, 17(3), e1008749.

27. Cheloshkina, K., Bzhikhatlov, I., & Poptsova, M. (2020, December). Cancer Breakpoint Hotspots Versus Individual Breakpoints Prediction by Machine Learning Models. In International Symposium on Bioinformatics Research and Applications (pp. 217-228). Springer, Cham.

28. Liu, B., Lee, W.S., Yu, P.S., et al. 2002. Partially supervised classification of text documents. In ICML. 387–394

29. Liu, B., Dai, Y., Li, X., et al. 2003. Building text classifiers using positive and unlabeled examples. In Third IEEE International Conference on Data Mining. IEEE179–IEEE186

30. Cheloshkina, K., Bzhikhatlov, I., & Poptsova, M. (2021) Randomness in cancer breakpoint formation. Journal of Computational Biology.