УДК 575.112:[616.155.392::616.155.392]

БИОИНФОРМАТИЧЕСКИЙ ПАЙПЛАЙН ПО ИДЕНТИФИКАЦИИ ПЕПТИДОВ, ПРИГОДНЫХ ДЛЯ CAR-T ТЕРАПИИ ЛЕЙКОЗОВ

В.В. Гринев, И.Н. Ильюшёнок, В.А. Сучек, Е.В. Гузова

Белорусский государственный университет, Кафедра генетики, биологический факультет, пр. Независимости, д. 4, 220050, г. Минск, Беларусь, grinev_vv@bsu.by

Статья описывает разработку многокомпонентного пайплайна, нацеленного на идентификацию иммуногенных пептидов, пригодных для CAR-T терапии лейкозов человека. Принцип идентификации основан на реконструкции аномальных открытых рамок считывания мРНК из данных RNA-Seq. Поиск пептидов осуществляется либо через сборку транскриптома, либо на основе уже имеющегося эталонного набора транскриптов. Пайплайн сфокусирован на пептидах, которые образуются за счёт альтернативного сплайсинга нормальных генов, а не за счёт нарушения последовательности генов.

***Ключевые слова*:** лейкозные клетки; биоинформатика; альтернативный сплайсинг РНК; иммуногенные пептиды; CAR-T терапия.

BIOINFORMATICS PIPELINE FOR IDENTIFICATION OF IMMUNOGENIC PEPTIDES SUITABLE FOR CAR-T THERAPY OF LEUKEMIA

V.V. Grinev, I.M. Ilyushonak, V.A. Suchok, K.V. Huzava

Department of Genetics, the Faculty of Biology, Belarusian State University, Nezavisimosti Avenue-4, 220050, Minsk, Belarus.
Corresponding author: grinev_vv@bsu.by

The work describes the development of a multicomponent pipeline aimed at identifying immunogenic peptides suitable for CAR-T therapy of human leukemia. A principle of identification is based on reconstruction of abnormal open reading frames in mRNAs, captured by RNA-Seq. Searching of peptides can be performed by transcriptome assembly or with using a reference set of known RNAs. Our approach focused on peptides, derived from alternatively spliced RNAs of normal genes, nor mutated sequences of genes.

***Keywords*:** leukemic cells; bioinformatics; alternative RNA splicing; immunogenic peptides; CAR-T therapy.

Введение

Экспертное сравнение новых методов лечения лейкозов, которые в настоящее время проходят клинические испытания или уже одобрены для применения, показывает, что одним из наиболее перспективных подходов считается CAR-T терапия. Инструментом такой терапии являются генети-

чески модифицированные аллогенные или аутологичные цитотоксические Т-лимфоциты, экспрессирующие химерные антигенные рецепторы (CAR, от англ. Chimeric Antigen Receptor) [1].

В качестве молекулярной мишени, присутствующей на поверхности лейкозных клеток, модифицированные Т-лимфоциты могут распознавать как опухоль-ассоциированные, так и опухоль-специфические антигены. Предпочтительным является второй класс антигенов, которые появляются в результате мутирования нормальных генов. Однако разнообразие таких антигенов, равно как и их иммуногенность, ограничены.

В то же время хорошо известно, что существенным источником разнообразия протеома клеток человека является альтернативный сплайсинг. Более того, известно, что протекание сплайсинга в лейкозных клетках нарушено, а также изменена функциональная активность систем контроля качества получаемых молекул РНК. По нашему мнению, все эти изменения в системах, обеспечивающих формирование транскриптома лейкозных клеток, могут приводить к появлению новых аминокислотных последовательностей, в том числе с высокой иммуногенностью. В связи с этим мы задались целью разработать биоинформатический пайплайн для идентификации потенциально иммуногенных пептидов, появляющихся в лейкозных клетках из-за изменений в работе системы сплайсинга РНК.

1. Методология исследования

Основные этапы пайплайна, нацеленного на идентификацию в клетках острого миелоидного лейкоза потенциально иммуногенных пептидов, представлены на рисунке 1. Пайплайн объединил два подхода в достижении поставленной цели. Первый подход предусматривает реконструкцию открытых рамок считывания (ORFs, от англ. Open Reading Frames) в транскриптах, подвергшихся альтернативному сплайсингу, на основе эталонных аннотаций транскриптома человека и информации о фазах кодонов в мультиэкзонных генах, кодирующих белки. В этом подходе учитываются все типы альтернативного сплайсинга, идентифицируемые в изучаемых лейкозных клетках, в том числе сохранение интронов.

Альтернативный подход основан на сборке транскриптома целевых клеток. Сборка ведется без опоры (*de novo*) и с опорой на эталонный геном. В каждом из этих случаев реконструированные молекулы РНК классифицируются на некодирующие и кодирующие, после чего в последних определяются координаты ORFs и их сиквенсы.

Из представленного описания и рисунка 1 видно, что конечным результатом применения любого из выше указанных подходов является список и сиквенсы ORFs, идентифицированных в транскриптоме изучае-

мых клеток. В дальнейшем сиквенсы этих ORFs подвергаются трансляции *in silico*, а полученные белки валидируются биоинформатически (относительно баз данных с известными белками человека) и экспериментально (по данным протеомного анализа). На последнем этапе каждый из идентифицированных белков проходит биоинформатическую проверку в качестве потенциального источника иммуногенных пептидов.

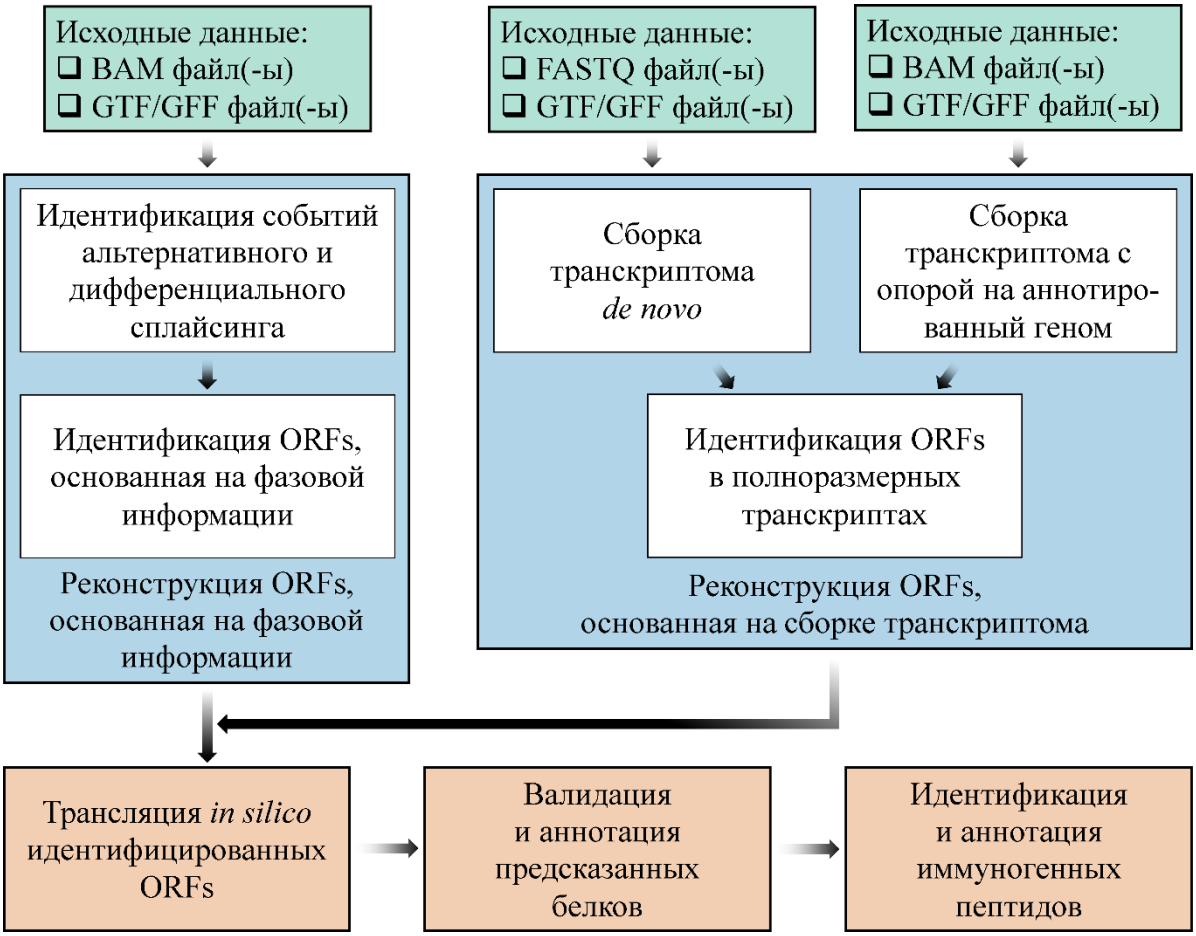


Рисунок 1 – Основные этапы пайплайна по идентификации иммуногенных пептидов лейкозных клеток человека

2. Результаты и их обсуждение

Ключевым компонентом нашего пайплайна является идентификация всех ORFs в транскриптоме изучаемых клеток. В контексте решения такой задачи прямой подход, основанный на использовании информации о фазах кодонов, имеет ряд преимуществ. Одно из таких преимуществ – высокая чувствительность детекции альтернативных событий сплайсинга. Так, опубликованная нами ранее [2] методология прямого обнаружения сохраняемых интронов в данных RNA-Seq на порядок чувствительнее,

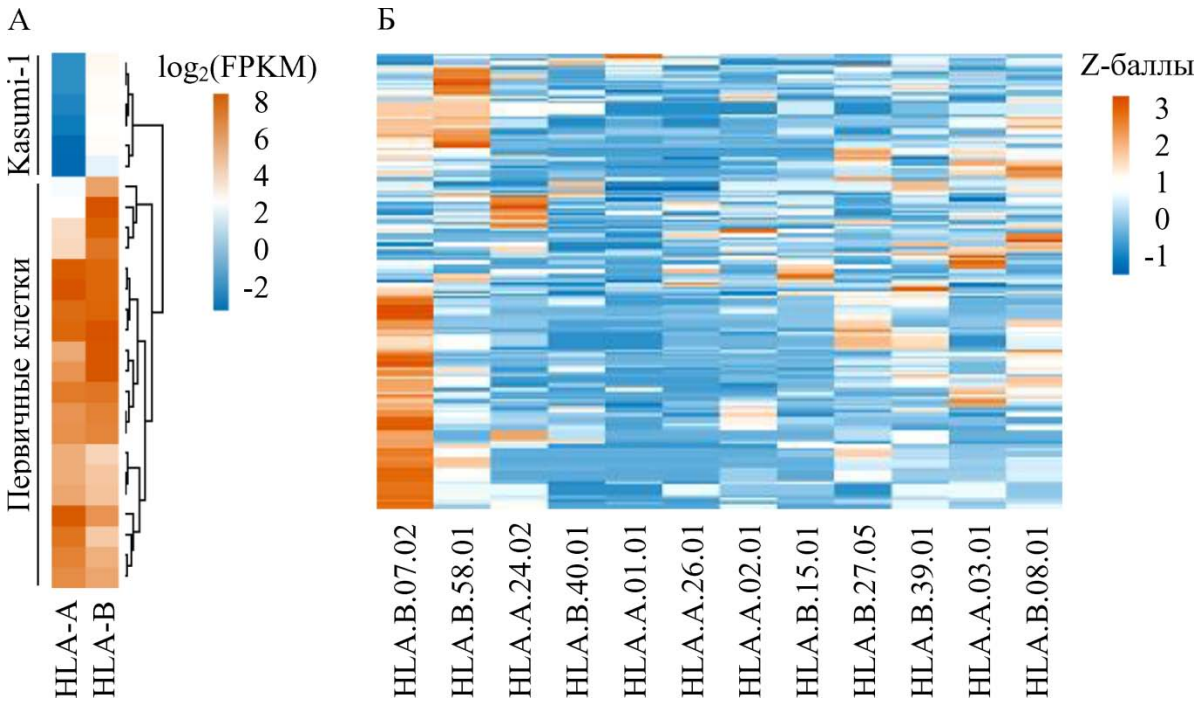
чем обнаружение таких же событий сплайсинга, но через сборку транскриптома. Дальнейшее «встраивание» интронов, классифицированных как сохраненные, в фазовый контекст аннотированных ORFs и позволяет реконструировать новые нуклеотидные последовательности, кодирующие белки. В то же время такой подход имеет и свои недостатки. В частности, текущий уровень наших знаний о детерминации сплайсинга не позволяет, основываясь только на эталонных аннотациях и информации об обнаруженном альтернативном событии сплайсинга, надежно реконструировать весь транскрипт. Следовательно, и реконструкция аминокислотной последовательности также будет носить локальной характер.

Второй подход в идентификации ORFs в лейкозных клетках основан на сборке всего транскриптома. Поскольку в постнатальный период иммуногенными могут быть только новые пептиды, то мы сосредоточились на тех транскриптомных сборщиках, которые позволяют собирать не только известные, но и новые транскрипты. Для этого мы воспользовались хорошо зарекомендовавшими себя сборщиками Cufflinks и StringTie. Оба сборщика работают с картированными ридами и эталонным геномом. Это, с одной стороны, повышает надежность сборки, но, с другой стороны, ограничивает репертуар обнаруживаемых новых транскриптов. Поэтому в качестве дальнейшей альтернативы мы рассматриваем использование сборщиков типа Trinity, которые реконструируют транскриптом *de novo*, напрямую из исходных некартированных ридов.

Помимо выбора сборщика транскриптома критически важным компонентом второго подхода является алгоритм классификации собранных транскриптов на некодирующие и кодирующие и обнаружения истинных ORFs в последнем классе транскриптов. Этому вопросу мы уделили особое внимание и разработали программный пакет ORFhunteR. Этот пакет использует векторизацию признаков транскриптов и мета-классификатор по типу случайного леса деревьев принятия решения для надежного определения координат ORFs в молекулах мРНК [4]. В настоящее время пакет ORFhunteR интегрирован с пайплайном по идентификации новых пептидов в лейкозных клетках.

На этапе валидации с помощью биоинформатического анализа проводился поиск тех белков (или их фрагментов), которые не имеют гомологии или же имеют низкую гомологию с известными белками человека. Этот анализ реализовывался с помощью стандартных инструментов выравнивания BLAST и белковых баз данных. Дальнейшая проверка реальной экспрессии белков в изучаемых клетках проводилась с помощью данных протеомного анализа. Количество обнаруживаемых нами таким способом белков варьировало в зависимости от типа изучаемых клеток и ме-

тода реконструкции/обнаружения ORFs. Так, в клетках линии Kasumi-1 при сборке транскриптома с помощью Cufflinks обнаруживается 207 потенциально новых полипептидных последовательностей.



А) Экспрессия антигенов HLA класса I в положительных по транслокации t(8;21)(q22;q22) клетках модельной линии Kasumi-1, а также первичных клетках пациентов с острым миелоидным лейкозом.
Б) Теплокарта аффинности связывания пептидов новых белков с антигенами HLA класса I.

Рисунок 2 – Идентификация иммуногенных пептидов в протеоме лейкозных клеток человека.

Наконец, на последнем этапе проводилось моделирование взаимодействия пептидов, высвобождаемых из идентифицированных белков, с антигенами HLA класса I. Интересно, что эти антигены экспрессируются на разных уровнях в клетках перевиваемых лейкозных линий и первичных клетках, выделенных из красного костного мозга или периферической крови больных лейкозами (рисунок 2А). В нашем пайплайне аффинность связывания пептидов с антигенами HLA класса I рассчитывается с помощью алгоритма NetMHCpan. Этот алгоритм позволяет фрагментировать белок на пептиды длиной от 8 до 14 аминокислот и оценить их взаимодействие с 10386 аллелями антигенов HLA.A, HLA.B, HLA.C, HLA.D, HLA.E и HLA.G. Как видно на рисунке 2Б, часть из обнаруженных белков, не имеющих гомологии или же с низкой гомологией к известным

белкам человека, содержат пептиды, которые с высокой аффинностью связываются с некоторыми аллелями антигенов HLA класса I. Эти пептиды можно рассматривать как потенциально иммуногенные мишени для антигенных рецепторов CAR T-клеток.

Таким образом, нами разработан прототип биоинформатического пайплайна, нацеленного на поиск потенциально иммуногенных пептидов, пригодных для CAR-T терапии лейкозов человека. Инновационным в нашем пайплайне является поиск пептидов не среди мутантных белков, а среди белков, образующихся в результате аберрантного сплайсинга, в частности, идущего с включением в зрелые молекулы РНК интронов, кодирующих совершенно новые аминокислотные последовательности. Дальнейшее развитие этого пайплайна будет нацелено на включение дополнительных методов идентификации ORFs, а также биоинформатическую кросс-валидацию и экспериментальную валидацию иммуногенности обнаруживаемых пептидов.

Библиографические ссылки

1. Larson R.C., Maus M.V. Recent advances and discoveries in the mechanisms and functions of CAR T cells // Nature Reviews Cancer. 2021. № 21(3). P. 145–161. DOI: 10.1038/s41568-020-00323-z.
2. Grinev V.V., Barneh F., Ilyushonak I.M., Nakjang S., Smink J., van Oort A., Clough R., Seyani M., McNeill H., Reza M., Martinez-Soria N., Assi S.A., Ramanouskaya T.V., Bonifer C., Heidenreich O. RUNX1/RUNX1T1 mediates alternative splicing and reorganises the transcriptional landscape in leukemia // Nature Communications. 2021. № 12(1). P. 520. DOI: 10.1038/s41467-020-20848-z.
3. Pertea M., Pertea G.M., Antonescu C.M., Chang T.C., Mendell J.T., Salzberg S.L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads // Nature Biotechnology. 2015. № 33(3). P. 290–295. DOI: 10.1038/nbt.3122.
4. Grinev V.V., Yatskou M.M., Skakun V.V., Chepeleva M.K., Nazarov P.V. ORFhunteR: an accurate approach for the automatic identification and annotation of open reading frames in human mRNA molecules // Software Impacts. 2022. № 12. P. 1–4. DOI: 10.1016/j.simpa.2022.100268.