

ТРАНСКРИПТОМНО-ПРОТЕОМНЫЙ ПОДХОД К ИДЕНТИФИКАЦИИ НОВЫХ БЕЛКОВ ЛЕЙКОЗНЫХ КЛЕТОК ЧЕЛОВЕКА

Е.В. Гузова¹, И.Н. Ильюшёнко¹, Ф. Барне², Х.Л. Марин-Рубио³,
О. Хайденрайх², В.В. Гринев¹

1) Белорусский государственный университет, Минск, Беларусь

2) Центр детской онкологии им. Принцессы Максимы, Утрехт,
Нидерланды

3) Ньюкаслский университет, Ньюкасл, Великобритания

E-mail: k.guzowwa0099@gmail.com

В настоящей работе описан подход, позволяющий осуществлять поиск неизвестных на сегодняшний день белков, синтезирующихся в лейкозных клетках человека. В его основе лежит биоинформатический анализ данных полнотранскриптомного секвенирования указанных клеток, а также результатов протеомного анализа. Реализация такого подхода потенциально может позволить расширить репертуар неоантигенов, пригодных для противолейкозной терапии.

Ключевые слова: лейкозные клетки; транскриптом; сплайсинг РНК; протеом; неоантигены.

Среди существующих в настоящее время направлений лечения лейкозов наиболее перспективными признаны адаптивная Т-клеточная терапия и вакцинация. Главным фактором, сдерживающим их широкое применение, является низкое разнообразие молекулярных мишеней. Известно, что в роли таковых могут выступать экспрессируемые лейкозными клетками опухоль-ассоциированные или опухоль-специфические антигены. Последние, так же называемые неоантигенами, представляют наибольший интерес. Рассматриваемые главным образом как продукт экспрессии нормальных генов, мутировавших при опухолевой трансформации клеток, такие антигены характеризуются ограниченным репертуаром и зачастую недостаточной иммуногенностью.

Между тем доказано, что основным источником разнообразия транскриптома и, следовательно, протеома клеток человека является альтернативный сплайсинг РНК. Более того, многочисленные исследования показали, что его протекание, равно как и функционирование систем контроля качества РНК, в лейкозных клетках нарушено. Это, в свою очередь, позволяет сделать предположение о возможности опосредованного подобными aberrациями появления новых аминокислотных последовательностей, в том числе иммуногенных [1]. На основании изложенных фактов была поставлена цель разработать подход, который базировался бы на реконструкции полноразмерных РНК лейкозных клеток человека и позволял

с опорой на результаты протеомного анализа идентифицировать экспрессирующиеся в указанных клетках неизвестные на сегодняшний день белки. Предполагается, что, по меньшей мере, часть из них будет способна выступать в роли неоантигенов.

Обобщенная схема алгоритма поиска новых белков, экспрессирующихся в лейкозных клетках человека, представлена на рисунке ниже. В предлагаемом подходе может быть выделено два крупных блока. Первый включил в себя ряд этапов работы на уровне транскриптома исследуемых клеток и начинался с подготовки файлов, необходимых для реконструкции полноразмерных РНК. Непосредственно сборка транскриптома осуществлялась с использованием двух хорошо зарекомендовавших себя программных средств: *StringTie* и *Trinity* [4]. Первое работает с картированными прочтениями, а также эталонными аннотациями известных генов и транскриптов. Среди последовательностей РНК, реконструируемых *StringTie*, могут обнаруживаться как уже известные транскрипты, так и транскрипты, структурно не соответствующие эталонным, однако разнообразие последних ограничено. С целью нивелировать этот недостаток для выполнения рассматриваемого этапа также использовался сборщик *Trinity*, позволяющий собирать полноразмерные РНК *de novo*.



Основные этапы алгоритма поиска новых белков, экспрессирующихся в лейкозных клетках человека

Результаты работы указанных сборщиков наряду с корректно реконструированными транскриптами содержали некоторое количество шума и потому требовали проведения фильтрации. Последняя осуществлялась по

ряду параметров, выбор которых зависел от того, каким программным средством была собрана последовательность. Так, в случае реконструкции с помощью *Trinity*, фильтрация основывалась на отборе молекул в зависимости от уровня их экспрессии. Обработка результатов работы *StingTie* включала в себя учет более широкого спектра показателей. Наряду с уровнем экспрессии оценивалась протяженность транскриптов, длина экзонов и интронов в их составе, принадлежность генов той или иной хромосоме и цепь ДНК. Для установления пороговых значений количественных параметров фильтрации были построены и проанализированы соответствующие графики плотности вероятности.

Среди реконструированных транскриптов наибольший интерес представляли белок-кодирующие, поэтому на следующем этапе с использованием функционала R/Bioconductor-пакета *ORFhunteR* [3] проводилась классификация всех РНК в соответствии с наличием у них кодирующего потенциала. В транскриптах, которые были отнесены к классу кодирующих белок, устанавливались координаты основных открытых рамок считывания (ОРС). Следующий шаг заключался в трансляции *in silico* обнаруженных ОРС. Именно он обеспечивал переход ко второму блоку подхода, посвященному работе на протеомном уровне.

На следующем этапе была использована нейронная сеть *DIA-NN* [2], которая позволила провести валидацию предсказанных *in silico* белков путем поиска подтверждающих их экспрессию пептидов по данным, полученным в ходе масс-спектрометрического (МС) анализа трипсинолизованной белковой фракции лизата целевых клеток. Поскольку описываемый подход нацелен на поиск новых белков, важно было выяснить, какие из исследуемых последовательностей, чья экспрессия подтверждалась пептидами МС-анализа, несут в себе фрагменты, специфические для лейкозных клеток. В качестве таковых в данном случае рассматривались фрагменты, которые не обнаруживались в последовательностях известных на сегодняшний день белков. Последние были взяты из таких баз данных, как UniProtKB/Swiss-Prot, Consensus CDS и NCBI RefSeq.

Далее идентифицированные уникальные фрагменты, а также выравнивающиеся относительно них пептиды анализировались с использованием *blastp*. По результатам данного анализа создавался набор новых белков исследуемых клеток. Включались в него те белки, уникальные фрагменты которых, равно как и пептиды, подтверждающие их, имели идентичностью менее 100 % к известным белкам человека.

Основные результаты, полученные в ходе реализации описанного подхода с использованием в качестве исходных данных результатов полнотранскриптомного секвенирования клеток линии Kasumi-1, представлены в таблице ниже.

Основные результаты реализации подхода

	Способ сборки транскриптома	
	<i>de novo</i>	с опорой на референсный геном
Общее количество уникальных предсказанных белков	12405	25812
Количество белков, прошедших валидацию данными МС-анализа	3313	11963
Количество идентифицированных новых белков	16	11

Примечательно, что среди новых белков, предсказание которых проводилось на основе сборки полноразмерных РНК с помощью *StringTie*, 5 кодировались псевдогенами. Остальные же белки, как и следовало ожидать, были получены в результате трансляции *in silico* ОРС, идентифицированных в транскриптах, структурно не соответствовавших эталонным изоформам РНК.

Таким образом, разработан подход, который позволяет обнаруживать неизвестные в настоящее время белки лейкозных клеток человека и может рассматриваться как часть пайплайна, нацеленного на идентификацию неоантигенов указанных клеток. Его дополнение алгоритмом оценки иммуногенности обнаруживаемых новых белков позволит расширить репертуар молекулярных мишеней, ограниченность которого в настоящее время сдерживает развитие противолейкозной терапии.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. A genome-wide aberrant RNA splicing in patients with acute myeloid leukemia identifies novel potential disease markers and therapeutic targets. / Adamia S., Haibe-Kains B., Pilarski P.M. [et al.] // Clin Cancer Res. 2013. Vol. 20. P. 1135-1145. DOI: 10.1158/1078-0432.CCR-13-0956
2. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. / Demichev V., Messner C. B., Vernardis S. I. [et al.] // Brief Communication. 2020. Vol. 17. P. 41-44. DOI: 10.1038/s41592-019-0638-x
3. 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
4. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. / Pertea M., Pertea G.M., Antonescu C.M. [et al.] // Nature Biotechnology. 2015. Vol. 33. P. 290-295. DOI: 10.1038/nbt.3122