ДЕТЕКЦИЯ ДИФФЕРЕНЦИАЛЬНЫХ СПЛАЙСИНГОВЫХ СОБЫТИЙ В ТРАНСКРИПТОМЕ КЛЕТОК ЧЕЛОВЕКА С ПОМОЩЬЮ ВЫСОКОПРОИЗВОДИТЕЛЬНОГО СЕКВЕНИРОВАНИЯ

В. В. Гринев[1], И. Н. Ильюшёнок[1], S. Накжайн[2], К. Бонифер[3], О. Хайденрайх[4]

[1]*Кафедра генетики, Белорусский государственный университет*
Минск, Беларусь
[2]*Группа биоинформационной поддержки, Ньюкаслский университет*
Ньюкасл, Великобритания
[3]*Институт рака и геномных наук, Бирмингемский университет*
Бирмингем, Великобритания
[4]*Северный институт по исследованию рака, Ньюкаслский университет*
Ньюкасл, Великобритания
e-mail: grinev_vv@bsu.by

На основе данных полнотранскриптомного секвенирования осуществлена сборка транскриптома клеток, различающихся по уровню экспресии гибридного онкогена RUNX1-RUNX1T1, построены карты сплайсинговых событий, показана высокая эффективность линейного моделирования при оценке дифференциального проявления таких событий. Проведена всесторонняя функциональная аннотация дифференциальных сплайсинговых событий, идентифицированы структурные различия в организации белков, кодируемых дифференциально экспрессирующимися изоформами РНК.

Ключевые слова: RNA-seq; сборка транскриптома; дифференциальный сплайсинг; аннотирование; мета-классификаторы.

DETECTION OF THE DIFFERENTIAL SPLICING EVENTS IN TRANSCRIPTOME OF HUMAN CELLS USING NEXT-GENERATION SEQUENCING

V. V. Grinev[1], I. M. Ilyushonak[1], S. Nakjang[2], C. Bonifer[3], O. Heidenreich[2]

[1]*Department of Genetics, Belarusian State University*
Minsk, Belarus
[2]*Bioinformatics Support Unit, University of Newcastle*
Newcastle upon Tyne, UK
[3]*Institute of Cancer and Genomic Sciences, University of Birmingham*
Birmingham, UK
[4]*Northern Institute for Cancer Research, University of Newcastle*
Newcastle upon Tyne, UK

The transcriptomes of cells with different expression of the fusion oncogene RUNX1-RUNX1T1 were assembled from high-throughput sequencing data. Splicing

maps were also inferred from primary sequencing data and the differential splicing events were identified with linear modeling. Moreover, a comprehensive functional annotation of identified differential splicing events was carried out. This annotation confirmed that the differentially expressed RNA isoforms may encode functionally distinct proteins with unique conserved domain structures.

Keywords: RNA-seq; assembling of transcriptome; differential splicing; annotation; meta-classifiers.

Результаты исследований, проведенных в различных научных лабораториях мира в последние годы, указывают, что первичные транскрипты не менее чем 90 % генов человека подвергаются альтернативному сплайсингу [2]. Альтернативный сплайсинг может протекать как в одной и той же клетке, так и в клетках разных типов и разной тканевой принадлежности, а также на разных стадиях развития или в разных условиях существования. В конечном счете альтернативный сплайсинг значительно увеличивает сложность транскриптома и протеома клеток и расширяет функциональные и адаптационные возможности живых организмов.

Разработка и внедрение в исследовательскую практику инструментальных методов высокопроизводительного анализа, в частности, полнотранскриптомного секвенирования, предоставляет уникальные возможности глубже и на системном уровне понять особенности организации и функционирования транскриптома клеток человека в норме и при различных заболеваниях. Однако чтобы воспользоваться открывшимися возможностями, нужны новые подходы в обработке больших массивов данных и интерпретации результатов.

Мы апробировали комплексный аналитический подход для идентификации дифференциальных (проявляющих себя в разной степени) сплайсинговых событий по данным высокопроизводительного секвенирования. Объектом наших изысканий был транскриптом клеток линии Kasumi-1, которая является лабораторной моделью положительной по транслокации *t*(8; 21)(q22; q22) формы острого миелоидного лейкоза. При этом использовались клетки, где идет экспрессия гибридного онкогена RUNX1-RUNX1T1, играющего важную роль в развитии данной формы лейкоза, и без экспрессии этого онкогена.

СБОРКА ТРАНСКРИПТОВ

Первый этап нашей аналитической работы был нацелен на сборку транскриптов из RNA-seq ридов. Для этого использовался сборщик Cufflinks [5], который был выбран как наиболее эффективный и релевантный поставленной задаче. Алгоритм работы сборщика включает реконструкцию ациклического орграфа (экзонного графа), извлечение из этого орграфа ориентированных путей, соответствующих полноразмерным транскриптам, и взвешивание полученных транскриптов.

Вершинами экзонного графа являются экзоны, а дугами – сплайсинговые события, которые объединяют экзоны, разделенные на уровне геномной ДНК и пре-мРНК интронами, в единую непрерывную последовательность зрелой мРНК. Нуклеотидная последовательность каждого из экзонов реконструировалась из групп перекрывающихся ридов, объединенных в контиги. Сплайсинговые события идентифицировались с помощью алгоритма глобального картирования seed-and-vote. Подтверждением таких событий были риды, для успешного картирования которых их необходимо было разделить на два (реже три) сегмента, причем разные сегменты картировались по раз-

ным участкам генома (экзонам гена). Для каждого гена, экспрессирующегося в изучаемых клетках, реконструировался отдельный экзонный граф.

Молекулой РНК в экзонном графе является ориентированный путь, выходящий из 5'UTR-экзона (истока) и заканчивающийся в 3'UTR-экзоне (стоке). Для каждого экзонного графа в соответствии с теоремой Дилуорса извлекалось такое минимальное подмножество ориентированных путей, которое максимально полно описывало все риды, картированные по данному гену. Эти пути взвешивались по количеству ридов, их подтверждающих, как описано в работе Trapnell C. [5]. Полученные промежуточные результаты подвергались глубокому парсингу и фильтрации с помощью оригинального *R*-кода, после чего с помощью алгоритма Cuffdiff [6] путям, прошедшим фильтрацию, переназначались веса, а окончательные данные сохранялись в виде GFF3-файлов. Пример результатов сборки транскриптов представлен на рис. 1.

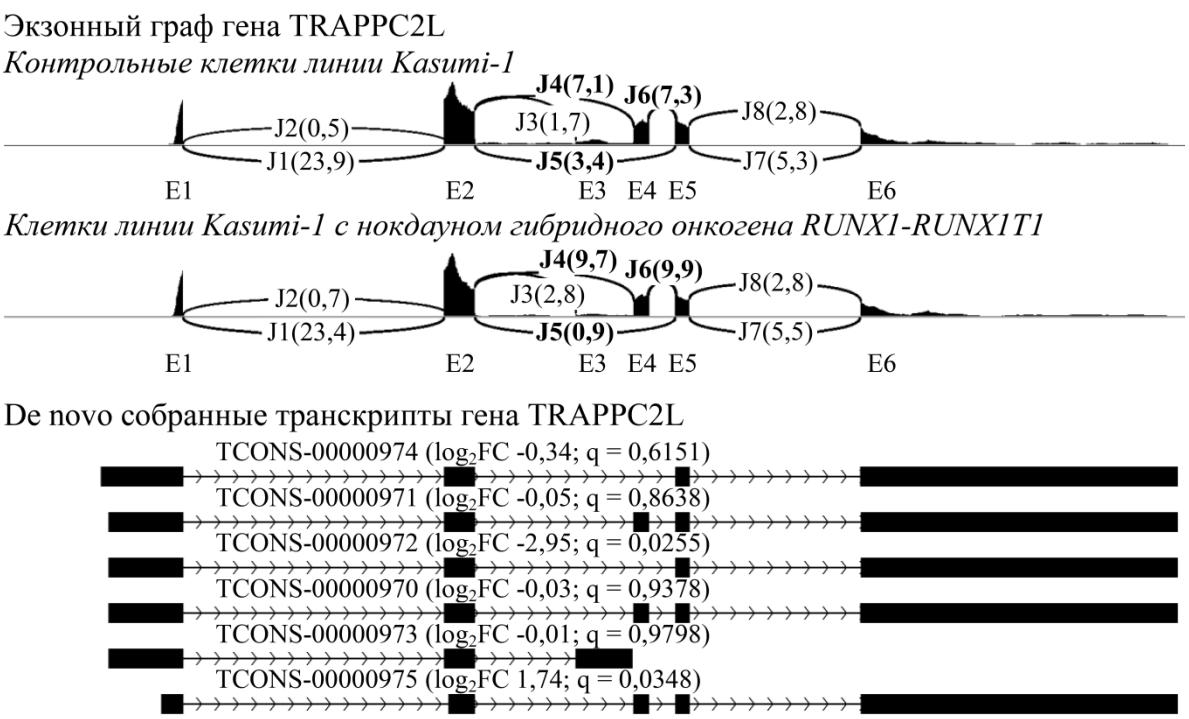

Рис. 1. Сборка транскриптов гена TRAPPC2L человека. На экзонных графах экзоны и сплайсинговые события обозначены как E и J, а в скобках указано нормализованное количество ридов, подтверждающих то или иное сплайсинговое событие. Для собранных транскриптов в скобках указаны количественные различия их содержания в двух состояниях клеток линии Kasumi-1 и статистическая значимость наблюдаемых различий

Теоретически сборка полноразмерных транскриптов и их количественное сравнение является наиболее правильным решением задачи по идентификации дифференциальных сплайсинговых событий в двух разных типах или состояниях клеток. Однако этот подход имеет ряд недостатков, что ограничивает его применимость. В частности, использованный нами сборщик Cufflinks позволяет реконструировать не менее двух мультиэкзонных транскриптов на ген только для 70 % проанализированных генов. Кроме того, в тех случаях, когда с гена считывается много близких по структуре изоформ РНК, алгоритм может собрать несколько альтернативных наборов транс-

криптов. При этом количественная оценка экспрессии таких изоформ нередко ненадежна и нуждается в экспериментальной верификации.

ИДЕНТИФИКАЦИЯ ДИФФЕРЕНЦИАЛЬНЫХ СПЛАЙСИНГОВЫХ СОБЫТИЙ С ПОМОЩЬЮ ЛИНЕЙНОГО МОДЕЛИРОВАНИЯ

На втором этапе работы мы сосредоточились на прямом поиске дифференциальных сплайсинговых событий, без реконструкции полноразмерных транскриптов. По нашему мнению, такой подход позволяет уточнить и расширить результаты первого этапа, описанного выше, хотя при этом и не предоставляет информации о структуре дифференциально экспрессирующихся изоформ РНК.

Для успешной реализации второго этапа мы провели картирование RNA-seq ридов с помощью элайнера Subjunc и собрали данные обо всех потенциальных сплайсинговых событиях в виде матрицы сплайсинга. Далее количественные данные последовательно нормализовались относительно размера RNA-seq библиотек, преобразовывались в количество ридов на миллион (CPM, акроним от англ. counts per million), $\log_2$-трансформировались [1, 3] и использовались для расчета параметров линейных моделей limma [4].

В итоге линейное моделирование позволило нам установить кратность различий (или logFC, от англ. logarithm of fold changes) по встречаемости сплайсинговых событий в транскриптоме клеток линии Kasumi-1, находящихся в двух состояниях (с и без нокдауна гибридного онкогена RUNX1-RUNX1T1), и рассчитать ассоциированные статистические метрики. С помощью такого подхода нам удалось идентифицировать 95 генов, мРНК которых содержат сплайсинговые события, различающиеся в двух состояниях клеток линии Kasumi-1, что на 30 % больше, чем при сборке полноразмерных транскриптов (рис. 2).

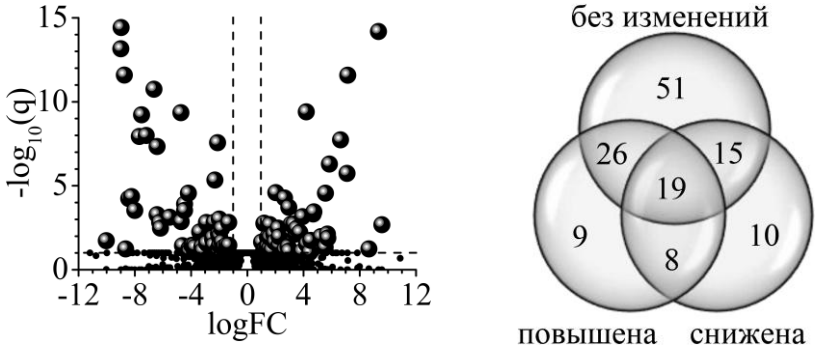

Рис. 2. График-«вулкан» распределения сплайсинговых событий между двумя состояниями клеток (слева) и диаграмма Венна распределения доменных структур белков, кодируемых транскриптами с постоянной или меняющейся экспрессией (справа)

ФУНКЦИОНАЛЬНАЯ АННОТАЦИЯ ИЗОФОРМ РНК И СПЛАЙСИНГОВЫХ СОБЫТИЙ

На третьем, завершающем, этапе мы составили всестороннюю аннотацию (описание) идентифицированных изоформ РНК и сплайсинговых событий. При этом основной упор делался на функциональные особенности, но учитывались также и структурные признаки анализируемых объектов. Так, РНК были описаны по 51-му призна-

ку, касающемуся особенностей их нуклеотидной последовательности, и по 5588-му признаку, описывающему кодируемые ими белки. Кроме того, особое внимание было уделено доменной структуре белков, кодируемых изоформами РНК.

Столь подробная аннотация понадобилась, что бы в последующем найти различия между дифференциально и недифференциально экспрессирующимися изоформами РНК. Для этого использовался метод главных компонент (стандартный вариант и его «ядерные» модификации), а также мета-классификаторы (random forests, GBM и SVM). В конечном итоге мы обнаружили, что ряд дифференциально экспрессирующихся изоформ РНК кодируют структурно и функционально различающиеся белки по сравнению с изоформами, экспрессия которых постоянна (рис. 2).

Кроме того, отдельно были аннотированы не изоформы РНК, а сплайсинговые события, для чего были задействованы признаки пяти классов – от особенностей нуклеотидной последовательности экзонов до эпигенетических признаков. В дальнейшем, с опорой на такую комплексную аннотацию, было показано, что дифференциальный сплайсинг не является случайным процессом. Так, мы наблюдали нелинейную отрицательную зависимость между вероятностью того, что сплайсинговое событие будет дифференциально представлено в транскриптоме клеток с разным статусом гибридного онкогена RUNX1-RUNX1T1, и расстоянием до таких эпигенетических маркеров, как сайты гиперчувствительности к ДНКазе I, модифицированные гистоны H3K9Ac или пики РНК полимеразы II.

Таким образом, данные полнотранскриптомного секвенирования содержат информацию о дифференциальных сплайсинговых событиях, отличающих транскриптомы сравниваемых клеток. Однако для извлечения такой информации необходим комплексный аналитический подход, который должен включать как разные методы идентификации сплайсинговых событий, так и их всестороннюю аннотацию.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor / S. Anders [et al.] // Nat. Prot. 2013. Vol. 8. № 9. P. 1765–1786.
2. Genomics of alternative splicing: evolution, development and pathophysiology / E. R. Gamazon [et al.] // Hum. Gen. 2014. Vol. 133. № 6. P. 679–687.
3. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts / C. W. Law [et al.] // Gen. Biol. 2014. Vol. 15. P. R29.
4. Limma powers differential expression analyses for RNA-sequencing and microarray studies / M. E. Ritchie [et al.] // Nuc. Acid. Res. 2015. Vol. 43. № 7. P. e47.
5. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks / C. Trapnell [et al.] // Nat. Prot. 2012. Vol. 7. № 3. P. 562–578.
6. Differential analysis of gene regulation at transcript resolution with RNA-seq / C. Trapnell [et al.] // Nat. Biotechnol. 2013. Vol. 31. № 1. P. 46–53.