ИДЕНТИФИКАЦИЯ СТЕПЕННОГО ЯДРА В ТРАНСКРИПТОМЕ ЧЕЛОВЕКА

В. В. Гринев[1], И. Н. Ильюшёнок[1], П. В. Назаров[2], Л. Валлар[2], О. Хайденрайх[3]

[1]*Кафедра генетики, Белорусский государственный университет*
Минск, Беларусь
[2]*Лаборатория геномных исследований, Люксембургский институт здоровья*
Штрассен, Люксембург
[3]*Северный институт по исследованию рака, Ньюкаслский университет*
Ньюкасл, Великобритания
e-mail: grinev_vv@bsu.by

Работа нацелена на выяснение закономерностей сплайсинга в клетках человека. Обнаружено, что на уровне транскриптома большинство сайтов сплайсинга и экзонов задействовано лишь в единичных сплайсинговых событиях. В то же время существенно меньшая их часть вовлечена во множество сплайсинговых событий. Без учета действия внешних и внутренних стохастических факторов такое неодинаковое использование сайтов сплайсинга и экзонов может быть описано с помощью экспоненциальных моделей или степенного закона с экспоненциальным обрезом. Если же такие факторы учтены, то вычленяется устойчивый компонент распределения, который описывается степенным законом. Этот компонент был назван степенным ядром транскриптома человека.

Ключевые слова: сплайсинг; закономерности сплайсинга; статистические модели; степенной закон.

IDENTIFICATION OF THE POWER-LAW CORE IN HUMAN TRANSCRIPTOME

V. V. Grinev[1], I. M. Ilyushonak[1], P. V. Nazarov[2], L. Vallar[2], O. Heidenreich[3]

[1]*Department of Genetics, Belarusian State University*
Minsk, Belarus
[2]*Genomics Research Laboratory, Luxembourg Institute of Health*
Strassen, Luxembourg
[3]*Northern Institute for Cancer Research, University of Newcastle*
Newcastle upon Tyne, UK

The study focuses on evaluation of the splicing patterns in human cells. It was found that at the level of entire transcriptome most splice sites and exons are involved only in single splicing events. At the same time, a small part of splice sites and exons are used in wide range of splicing events. Such unequal usage of splice sites and exons can be plausible described with exponential statistical models or power-law with an

exponential cut-off. Furthermore, we also observed a pure power-law component in empirical distributions if influence of stochastic external and internal factors was excluded. We called this component a power-law core of human transcriptome.

Keywords: splicing; splicing patterns; statistical models; power-law.

Конститутивный и альтернативный сплайсинг является фундаментальным процессом, протекающим во всех без исключения клетках эукариот и приводящим к образованию зрелых функциональных РНК-продуктов [4]. Однако, несмотря на столь высокую значимость и почти 40-летнюю историю изучения данного процесса, принципы (правила) комбинаторики экзонов во время сплайсинга до сих пор не установлены. Так, расшифровка «кода сплайсинга» как наиболее важное достижение последних лет в этой области [1] позволяет объяснить лишь некоторые особенности локального сплайсинга, но не закономерности глобальной комбинаторики экзонов в целых транскриптах и всем транскриптоме.

Недавно нами было обнаружено, что поведение экзонов гибридного онкогена RUNX1-RUNX1T1, экспрессирующегося в клетках положительной по транслокации *t*(8; 21)(q22; q22) формы острого миелоидного лейкоза человека, во время сплайсинга следует степенному закону [3]. Однако вопрос о том, является ли это особенностью только данного онкогена или же это характерно и для других генов человека и транскриптома в целом остался за рамками того исследования. В настоящем исследовании мы предприняли попытку найти ответ на поставленный вопрос.

Для этого мы воспользовались данными секвенирования транскриптома клеточной линии Kasumi-1, которая является лабораторной моделью положительной по транслокации *t*(8; 21)(q22; q22) формы острого миелоидного лейкоза. В нашей работе были задействованы клетки, находящиеся в одном из двух состояний: 1) клетки, обработанные неспецифическими короткими интерферирующими РНК (контрольные siRNA), что сохраняло неизмененной экспрессию онкогена RUNX1-RUNX1T1; 2) клетки, в которых с помощью специфических коротких интерферирующих РНК (анти-RUNX1-RUNX1T1 siRNA) экспрессия указанного онкогена была подавлена. Далее из данных полнотранскриптомного секвенирования была извлечена информация о сплайсинговых событиях, протекающих на уровне зарождающихся РНК, а также при окончательном созревании РНК. Кроме того, на основе этих же данных была проведена сборка и идентификация полноразмерных зрелых транскриптов, присутствующих в транскриптоме изучаемых клеток.

Информация о сплайсинговых событиях была использована нами для реконструкции графов сплайсинга. Графы сплайсинга являются ациклическими орграфами с 5’ и 3’ сайтами сплайсинга в качестве вершин и экзонами и интронами в качестве дуг. С помощью стандартного топологического анализа мы рассчитали степени выходов для каждой вершины типа 5’ сайта сплайсинга (5’ss SSCI, от англ. combinatorial index of 5’ splice site), а также степени входов для 3’ сайтов сплайсинга (3’ss SSCI, от англ. combinatorial index of 3’ splice site). В случае с графами сплайсинга физический смысл таких степеней довольно прост – они являются количественной мерой разнообразия сплайсинговых событий, в которые вовлечен данный сайт сплайсинга (или количества уникальных попарных комбинаций между данным сайтом и другими сайтами сплайсинга).

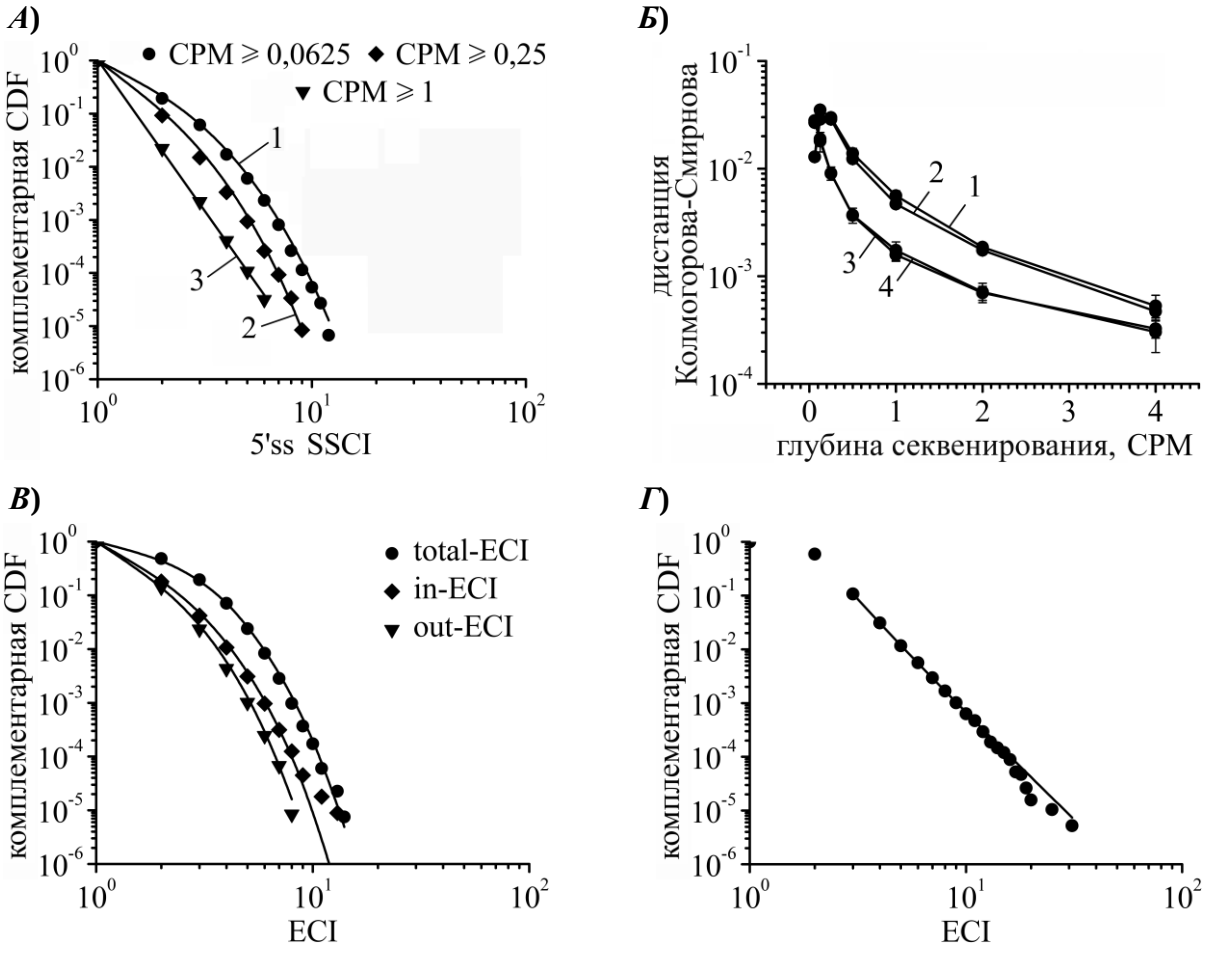
Помимо этого мы провели реконструкцию экзонных графов, воспользовавшись собранными из RNA-seq ридов полноразмерными РНК. Экзонный граф также являет-

ся ациклическим орграфом, однако, в отличие от графа сплайсинга, вершины в нем представлены целыми экзонами, а дуги – сплайсинговыми событиями, которые объединяют экзоны в последовательность зрелой РНК. Анализ структуры таких графов позволяет дать количественную оценку вовлеченности экзонов в разнообразные сплайсинговые события: по аналогии с сайтами сплайсинга экзонам могут быть присвоены степени как мера такой вовлеченности. При этом для каждого экзона мы рассчитывали три варианта степеней: степень входа (in-ECI, от англ. in-degree exon combinatorial index), степень выхода (out-ECI, от англ. out-degree exon combinatorial index) и общая степень (как сумма in-ECI и out-ECI; total-ECI, от англ. total-degree exon combinatorial index).

Дальнейший анализ полученных степеней сайтов сплайсинга и экзонов показал, что мы имеем дело с сильно скошенными вправо распределениями с коэффициентом асимметрии от 1,65 до 19,24 (в среднем 5,61). Поскольку такого рода распределения могут быть описаны целым семейством статистических моделей, то мы предприняли попытку определить, какая из них наиболее полно соответствует наблюдаемым данным. Для этого мы отобрали семь самых подходящих моделей (степенной закон, степенной закон с экспоненциальным обрезом, а также такие распределения как экспоненциальное, растянутое экспоненциальное, Юла-Саймона, Пуассона и логнормальное) и воспользовались разработками Newman M. E. J. с соавт. [2, 5] для подгонки этих моделей и проверки гипотез об их соответствии экспериментальным данным.

На уровне сайтов сплайсинга наиболее точно исходные экспериментальные данные описывает модель растянутого экспоненциального распределения. Однако, по мере удаления из набора данных тех сплайсинговых событий, которые подтверждаются небольшим количеством ридов (эти события встречаются редко и могут быть техническими артефактами или шумом, возникающем при работе внутриклеточной системы сплайсинга), экспоненциальный компонент уменьшается и в конечном итоге наиболее соответствующей экспериментальным данным становится модель степенного закона (рисунок, часть *А*). Причем описанная ситуация наблюдалась нами при работе со всеми наборами данных – от зарождающихся РНК до зрелых РНК, а также с и без нокдауна онкогена. Следует лишь отметить, что в случае с зарождающимися РНК необходима более глубокая фильтрация исходных данных, что бы достигнуть той же точности подгонки степенного закона к экспериментальным данным, как и в случае со зрелыми РНК (рисунок, часть *Б*).

На уровне экзонов наиболее точно исходные экспериментальные данные описывает степенной закон с экспоненциальным обрезом (рисунок, часть *В*). Опять же, по мере фильтрации и удаления транскриптов с низким уровнем экспресии экспоненциальный компонент и здесь уменьшается, и мы получаем степенной закон как наиболее подходящую статистическую модель. Особо следует отметить, что, как и в случае с сайтами сплайсинга, распределения экзонов по степеням имеют одинаковый вид в двух состояниях клеток – с нокдауном онкогена RUNX1-RUNX1T1 и без него. Это особо интересно тем, что при нокдауне указанного онкогена транскриптом клетки существенно перестраивается – статистически значимо меняется экспрессия 794 генов, а еще у 87 генов меняется характер сплайсинга их РНК.



Вовлеченность сайтов сплайсинга и экзонов в сплайсинг:

А) Взаимосвязь между встречаемостью в транскриптах, выраженной в количестве подтверждающих это ридов на 1 млн всех ридов (CPM, от англ. counts per million), и характером распределения 5’ сайтов сплайсинга по вовлеченности в различные сплайсинговые события. В верхней части рисунка указаны пороговые значения CPM, использованные для фильтрации экспериментальных данных. Каждому набору данных соответствует своя теоретическая кривая, рассчитанная с помощью наиболее правдоподобной статистической модели: *1* – растянутая экспоненциальная, *2* – степенной закон с экспоненциальным обрезом, *3* – степенной закон;

Б) Влияние фильтрации сплайсинговых событий против низких значений CPM на точность подгонки степенного закона под экспериментальные данные. На рисунке показаны обобщенные результаты анализа трех независимых секвенирований (как арифметическая средняя ± стандартное отклонение) зарождающихся РНК (кривые *1* и *2*) и зрелых РНК (кривые *3* и *4*), взятых из клеток, обработанных либо контрольными siRNA (кривые *1* и *3*), либо анти-RUNX1-RUNX1T1 siRNA (кривые *2* и *4*);

В) Характер распределения экзонов, присутствующих в зрелых транскриптах контрольных клеток линии Kasumi-1, по вовлеченности в различные сплайсинговые события. Каждый набор экспериментальных данных наиболее точно описывается с помощью степенного закона с экспоненциальным обрезом;

Г) Характер распределения экзонов, присутствующих в транскриптах из GenBank, по вовлеченности в различные сплайсинговые события. Экспериментальные данные наиболее точно описываются с помощью степенного закона с $x_{min} = 3$.

Для проверки универсальности наблюдаемого характера распределения сайтов сплайсинга и экзонов по степеням в транскриптоме человека мы использовали данные из GenBank, а также транскрипционные модели генов человека, созданные автоматическими и полуавтоматическими (с ручной проверкой) системами аннотирования NCBI RefSeq, UCSC Genome Browser, Ensembl, VEGA, AceView и ECgene. Наиболее соответствующей этим данным моделью является степенной закон с экспоненциальным обрезом, хотя в некоторых случаях (в том числе для разных релизов одних и тех же баз данных) приемлемы и альтернативные модели, в частности, растянутое экспоненциальное и логнормальное распределение. Однако при удалении из транскриптов концевых экзонов, границы которых, как правило, определены ненадежно, степенной закон становится наиболее подходящей статистической моделью и в этих случаях (рисунок, часть *Г*).

Таким образом, глобально, на уровне транскриптома, большинство сайтов сплайсинга и экзонов задействовано лишь в единичных сплайсинговых событиях. В то же время существенно меньшая их часть вовлечена во множество разнообразных сплайсинговых событий, выступая в роли своеобразных хабов всего процесса сплайсинга. Без учета стохастических факторов, имеющих как внешнее (методические и технические артефакты), так и внутреннее (стохастичность в работе самой системы сплайсинга клетки) происхождение, такое неодинаковое использование сайтов сплайсинга и экзонов в сплайсинге может быть описано с помощью экспоненциальных моделей или степенного закона с экспоненциальным обрезом. Если же такие факторы учтены, то вычленяется более устойчивый компонент распределения, который описывается степенным законом. Этот компонент нами был назван степенным ядром, или кором, транскриптома человека.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Barash Y. Deciphering the splicing code / Y. Barash [et al.] // Nature. 2010. Vol. 465. P. 53–59.

2. Clauset A., Shalizi C. R., Newman M. N. J. Power-law distributions in empirical data // SIAM Rev. 2009. Vol. 51. P. 661–703.

3. Decoding of exon splicing patterns in the human RUNX1-RUNX1T1 fusion gene / V. V. Grinev [et al.] // Int. J. Biochem. Cell Biol. 2015. Vol. 68. P. 48–58.

4. Structural basis of pre-mRNA splicing / J. Hang [et al.] // Science. 2015. Vol. 349. № 6253. P. 1191–1198.

5. Newman M. E. J. Power laws, Pareto distributions and Zipf's law // Contemp. Phys. 2005. Vol. 46. P. 323–351.