Биомедицинская химия, 2017 том 63, вып. 5, с. 413-417.

БИОИНФОРМАТИЧЕСКИЙ ПРОТОКОЛ ДЛЯ ОБРАБОТКИ NGS-ДАННЫХ И ИДЕНТИФИКАЦИИ МУТАЦИЙ В СОЛИДНЫХ ОПУХОЛЯХ ЧЕЛОВЕКА

К.Ю. Цуканов[1], А.Ю. Красненко[1], Д.А. Плахина[1], Д.О. Коростин[2], А.В. Чуров[3], О.С. Дружиловская[2]*, Д.В. Ребриков[2], В.В. Ильинский[1]

[1]ООО "Генотек", Москва.
[2]Институт общей генетики имени Н.И. Вавилова, Москва; эл. почта: dep-nano@yandex.ru
[3]Институт биологии Карельского научного центра Российской академии наук, Петрозаводск

Целью исследования была разработка и тестирование протокола для биоинформатической обработки NGS-данных для эффективного поиска мутаций в геноме солидных опухолей. Согласно разработанному протоколу на начальном этапе проводили оценку качества прочтений нуклеотидов. Нуклеотиды с качеством прочтения ниже 10 удаляли с 3′-конца с помощью программы Cutadapt. Далее прочтения картировали на референсный геном hg19 (GRCh37.p13) с помощью программы BWA. Дедупликацию ридов выполняли специализированной программой SAMtools. Для распознавания однонуклеотидных вариантов применяли MuTect. Комплексно оценивали функциональный эффект мутаций на основе алгоритма, включающего аннотирование и оценку патогенности с помощью программного решения SnpEff, а также анализа таких баз данных, как COSMIC, dbNSFP, Clinvar, OMIM. Дополнительно для оценки эффекта на функцию кодируемого белка применяли утилиты SIFT и Poly-Phen2. Информацию о частоте мутаций получали на основе данных проектов 1000 Genomes и ExAC, а также собственной базы данных частот. Для проведения тестирования протокола был проведен анализ 18 образцов биопсии опухолей молочной железы. Секвенирование образцов проводили на платформе Illumina. Для таргетного обогащения кодирующих регионов генома использовали набор реагентов MYbaits Onconome KL v1.5 Panel ("MYcroarray," США). По результатам биоинформатической обработки данных секвенирования обнаружено множество мутаций в генах BRCA1, BRCA2, ATM, CDH1, CHEK2, TP53, в том числе мутации-драйверы, оказывающие влияние на аминокислотную последовательность кодируемого белка. Таким образом, предлагаемый нами биоинформатический протокол позволяет эффективно выполнять автоматическую обработку NGS-данных образцов опухолей и обнаруживать мутации. Данный протокол для биоинформатической обработки данных впервые апробирован на выборке образцов опухолей из российской популяции. Для подтверждения эффективности и точности предлагаемого протокола в дальнейшем необходимо тестирование алгоритма на данных секвенирования различных форм опухолей.

Ключевые слова: мутация, биоинформатический протокол, высокопроизводительное секвенирование, биоинформатическая обработка NGS-данных

DOI: 10.18097/PBMC20176305413

ВВЕДЕНИЕ

Мутации, обнаруживаемые в геноме опухолевых клеток, представлены двумя типами: мутации-драйверы и мутации-пассажиры. В общем случае мутации первого типа являются триггерами, которые нарушают нормальную работу клеток и переводят их в состояние неконтролируемого деления. С клинической точки зрения выявление драйверных мутаций является ключевым в подборе противоопухолевой терапии, так как с большей вероятностью приведёт к снижению частоты рецидивов. Большинство мутаций являются соматическими и играют важную роль в развитии *de novo* резистентности к противоопухолевым химиотерапевтическим средствам. Как следствие, многие исследования направлены на профилирование мутаций в образцах опухолей с применением методов высокопроизводительного секвенирования (NGS, next generation sequencing), позволяя идентифицировать потенциальные мишени для таргетных препаратов, выявлять молекулярно-генетические онкомаркеры и разрабатывать протоколы для персонализированной терапии злокачественных опухолей.

Однако при исследовании мутаций на основе методов NGS возникают определенные технические трудности: 1) сложность определения генетических изменений с низкой частотой мутантного аллеля ввиду высокой опухолевой генетической гетерогенности, клонального разнообразия и вариаций числа копий; 2) распознавание мутаций от артефактов, возникающих в результате ошибок на этапе секвенирования; 3) определение и отличие соматических и герминальных мутаций; 4) трудности анализа опухолевых образцов, содержащих долю нормальных опухолевых клеток [1].

Для решения разнообразных задач онкогеномики сегодня создаётся большое количество программного обеспечения, позволяющего обрабатывать и интерпретировать результаты NGS. Для понимания механизмов патогенеза опухолей и разнообразия репертуара мутаций в опухолях необходима разработка биоинформатических алгоритмов, позволяющих эффективно идентифицировать мутации и преодолевать вышеуказанные технические барьеры. В результате ряда исследований было проведено

* - адресат для переписки

сравнение производительности и точности различного программного обеспечения для биоинформатической обработки данных, а также созданных на их основе протоколов [2, 3].

Оказалось, что степень соответствия данных, полученных с применением различных программных средств, как правило, низкая, так как в основе такого программного обеспечения лежат различные математические алгоритмы. Многие научные группы отмечают низкий уровень воспроизводимости результатов анализа различными методами [4].

В результате применения доступного программного обеспечения часто возникает множество ложноположительных результатов, либо в результате фильтрации данных может происходить потеря истинных данных – важных мутаций, определяющих биологические характеристики опухолевых клеток, – так называемых мутаций-драйверов [5].

Таким образом, большинство исследователей сталкивается с проблемой выбора наиболее подходящего алгоритма для биоинформатической обработки данных.

В данной работе нами представлен протокол для биоинформатической обработки данных секвенирования образцов ДНК опухолей и идентификации мутаций. Проведено тестирование протокола на выборке образцов рака молочной железы. По предварительным результатам, разработанный протокол продемонстрировал высокую производительность и может использоваться для эффективного поиска мутаций-драйверов, функциональной аннотации мутаций, а также для подбора средств для противоопухолевой терапии.

МЕТОДИКА

Материал исследования

Для проведения исследования и получения NGS-данных была собрана коллекция образцов опухолей (биопсийный материал) пациентов со злокачественными новообразованиями молочной железы в возрасте от 25 до 76 лет. Все пациенты поступали на обследование в "Национальный медицинский исследовательский центр онкологии имени Н.Н. Блохина" Минздрава России. Биопсийный материал был представлен образцами опухолевой ткани и прилежащей к опухоли нормальной ткани молочной железы, что подтверждено результатами гистологического исследования. Всего был проведен анализ образцов 18 опухолей молочной железы. От всех пациентов получены информированные согласия на проведение исследования.

Выделение ДНК и контроль качества

Выделение ДНК из опухолевой ткани было проведено с помощью набора DNeasy Blood and Tissue Kit ("Qiagen", США). Для этого измельченные образцы опухолевой ткани после добавления буфера ATL обрабатывали протеиназой K и инкубировали при 56°C до полного лизиса. Затем добавляли последовательно 200 мкл буфера AL и 96% этанола. Полученную смесь переносили на спин-колонки для связывания в результате центрифугирования при 8000 g в течение 1 мин. Затем образцы промывали центрифугированием, добавляя сначала 500 мкл буфера AW1 (1 мин, 6000g), а затем 500 мкл буфера AW2 (3 мин, 20000 g) для полного удаления этанола. Для элюции ДНК колонки дважды обрабатывали по 30 мкл буфера Low-TE, инкубировали и центрифугировали согласно протоколу производителя. Контроль качества полученной ДНК проводили на приборе Qubit 3.0, а также с помощью агарозного гель-электрофореза.

Получение NGS данных. Секвенирование таргетной панели онкогенов

Для получения NGS-данных применяли секвенирование таргетной панели онкогенов в образцах биопсийного материала опухолей молочной железы. Из полученных образцов ДНК были приготовлены библиотеки с применением наборов реагентов NEBNext Ultra DNA Library Prep Kit for Illumina ("New England Biolabs", США) согласно протоколу производителя.

Двойное баркодирование библиотек проводили с помощью полимеразной цепной реакции (ПЦР) с применением наборов реагентов NEBNext Ultra DNA Library Prep Kit for Illumina ("New England Biolabs") и NEBNext Multiplex Oligos for Illumina (Dual Index Primers Set 1, "New England Biolabs").

Контроль качества полученных библиотек фрагментов ДНК проводили на приборе Agilent Bioanalyzer 2100 ("Agilent Technologies", США) с помощью набора High Sensitivity Kit в соответствии с протоколом компании-производителя.

Далее образцы эквимолярно пулировались по 9 штук. Для таргетного обогащения кодирующих регионов генома использовали набор MYbaits Onconome KL v1.5 Panel ("MYcroarray", США). Контроль качества полученных пулов проводили на приборе Agilent Bioanalyzer 2100 с помощью набора реагентов High Sensitivity Kit ("Agilent Technologies") по протоколу производителя.

Секвенирование проводили на геномном анализаторе HiSeq 2500 System ("Illumina", США). Подготовку образцов и запуск осуществляли согласно протоколам Illumina: разведение образцов (Denature and Dilute Libraries for HiSeq Clustering); запуск (Sequencing in High Output Mode).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Протокол для биоинформатического анализа данных

Мы разработали протокол для поиска и аннотации однонуклеотидных вариантов в геноме опухолей на основе анализа данных высокопроизводительного секвенирования ДНК. Биоинформатическая обработка полученных в результате секвенирования ДНК опухолей данных выполняется оригинальным программным решением, объединяющим нижеперечисленные шаги. На начальном этапе была проведена оценка качества прочтений. Для этого

последовательности нуклеотидов с 3′-конца, имеющие качество прочтения ниже 10 обрезались с помощью программы Cutadapt [6]. Полученные прочтения картировали на референсный геном hg19 (GRCh37.p13) с помощью пакета программ BWA [7]. Фильтрацию данных – удаление повторяющихся фрагментов ДНК, выполняли специализированной командой rmdup в составе программной платформы SAMtools [8].

Распознавание коротких вариантов (мутаций) осуществляли с применением программы MuTect [9]. В качестве значимых вариантов рассматривали последовательности ДНК, число покрытий которых в результате процедуры секвенирования составило не менее 12.

Кроме того, оценивали функциональный эффект мутации на основе комплексного алгоритма. С этой целью проводили аннотирование вариантов и предсказание их влияния на кодируемый белок на основе анализа геномных координат фрагментов с помощью программного решения SnpEff [10]. Для оценки патогенности и консервативности выявляемых генетических вариантов использовали данные, которые извлекали из таких баз данных, как COSMIC (Catalogue of Somatic Mutations In Cancer) [11], dbNSFP [12], Clinvar [13], OMIM (Online Mendelian Inheritance In Man) [14]. Дополнительно для предсказания возможной патогенности мутаций, оценки эффекта на функцию кодируемого белка, применяли утилиты SIFT (Sorting Intolerant From Tolerant) [15] и PolyPhen2 [16]. Информацию о частоте мутаций получали на основе данных проекта 1000 Genomes [17] и консорциума ExAC [18].

Тестирование протокола

Для апробации предлагаемого нами протокола был проведён анализ данных, полученных в результате высокопроизводительного секвенирования образцов 18 опухолей молочной железы. Среди онкогенов, изученных в составе таргетной панели, в результате биоинформатической обработки данных наибольшее количество генетических вариантов обнаружено в генах BRCA1, BRCA2, ATM, CDH1, CHEK2, TP53.

По результатам исследований в вышеуказанных генах было обнаружено множество мутаций-драйверов. В дальнейшем были отобраны и аннотированы только мутации со значительным функциональным эффектом. Всего с применением разработанного протокола было обнаружено 15 мутаций, оказывающих влияние на аминокислотную последовательность кодируемых белков (таблица).

В каждом из исследованных образцов опухолей молочной железы обнаружена минимум одна мутация, функциональный эффект которой был подтверждён в ходе аннотации с применением программного обеспечения и баз данных. Из 15 мутаций 7 присутствуют в одной из наиболее информативных баз мутаций – COSMIC, что указывает на высокую вероятность их значимости для развития рака груди.

Кроме того, нами был проведён поиск на наличие таргетных противоопухолевых лекарственных препаратов. Аннотация проведена с применением баз данных My Cancer Genome (www.mycancergenome.org), а также ресурса Genomics of Drug Sensitivity in Cancer (GDSC, www.cancerrxgene.org). Результаты аннотации представлены в таблице.

Таблица. Однонуклеотидные варианты, обнаруженные в геномах опухолей молочной железы с применением протокола для биоинформатической обработки данных NGS

Ген	Мутация	Экзон	Аминокислотная замена	Эффект мутации	Таргетный препарат
TP53	c.524G>A	5	p.Arg175His	Патогенная	SAR405838|MI-773
	c.469G>T	5	p.Val157Phe	Патогенная	SAR405838|MI-773
	c.743G>A	7	p.Arg248Gln	Патогенная	SAR405838|MI-773
BRCA1	c.1865C>T	10	p.Ala622Val	Вероятно патогенная	-
	c.384G>A	6	p.Met128Ile	Вероятно патогенная	-
	c.54G>T	2	p.Met18Ile	Вероятно патогенная	-
BRCA2	c.5070A>C	11	p.Lys1690Asn	Вероятно патогенная	-
	c.4828G>A	11	p.Val1610Met	Вероятно патогенная	-
ATM	c.4258C>T	29	p.Leu1420Phe	Вероятно патогенная	KU-55933, CP466722
	c.1192G>C	9	p.Asp398His	Вероятно патогенная	KU-55933, CP466722
	c.5558A>T	37	p.Asp1853Val	Вероятно патогенная	KU-55933, CP466722
	c.146C>G	3	p.Ser49Cys	Патогенная	KU-55933, CP466722
CDH1	c.790C>T	6	p.Gln264*	Вероятно патогенная	-
	c.1342C>T	10	p.Gln448*	Вероятно патогенная	-
CHEK2	c.1289C>T	12	p.Thr430Ile	Вероятно патогенная	Rabusertib|AZD7762

ЗАКЛЮЧЕНИЕ И ВЫВОДЫ

Применение высокопроизводительного секвенирования позволило существенно улучшить наши знания об основных закономерностях возникновения и развития опухолей человека. Идентификация мутаций в опухолях на основе анализа NGS-данных открывает новые перспективы для разработки персонализированных подходов к диагностике и терапии опухолей, позволяя учитывать особенности геномов опухолей при подборе противоопухолевых лекарственных препаратов.

В настоящей работе мы разработали оригинальный биоинформатический алгоритм и провели его тестирование в результате анализа таргетной панели онкогенов методом высокопроизводительного секвенирования ДНК опухолей молочной железы. Предлагаемый нами протокол, представленный комплексом инструментов для биоинформатической обработки данных, может применяться на всех этапах анализа NGS-данных, включая оценку качества прочтений, аннотацию однонуклеотидных генетических вариантов и поиск подходящих таргетных противоопухолевых препаратов (с учётом генетических особенностей индивидуальных опухолей). В основе работы алгоритма лежит применение платформы MuTect, позоляющей проводить одновременный анализ образцов опухолевой и нормальной ткани, что обеспечивает высокую чувствительность и специфичность анализа.

По результатам проведенного исследования протокол для биоинформатической обработки данных высокопроизводительного секвенирования впервые апробирован на выборке образцов опухолей из российской популяции. Было обнаружено множество мутаций в образцах опухолей, а также выявлены некоторые ключевые мутации-драйверы.

Таким образом, предлагаемый нами протокол позволяет выполнять автоматическую обработку NGS-данных образцов опухолей. В перспективе применение данного программного обеспечения может быть использовано для эффективной диагностики онкозаболеваний и помощи в подборе таргетной терапии.

Однако для подтверждения эффективности и точности предлагаемого протокола в дальнейшем необходимо тестирование алгоритма на выборке, состоящей из бóльшего числа образцов опухолей. Необходима проверка указанного протокола на данных, полученных в результате секвенирования различных форм рака.

БЛАГОДАРНОСТИ

Работа выполнена при финансовой поддержке государства в лице Минобрнауки России (идентификатор соглашения RFMEFI60716X0152).

ЛИТЕРАТУРА

1. *Ding L., Wendl M.C., Koboldt D.C., Mardis E.R.* (2010) Hum. Mol. Genet., **19**, 188-196.
2. *Spencer D.H., Tyagi M., Vallania F., Bredemeyer A.J., Pfeifer J.D., Mitra R.D., Duncavage E.J.* (2014) J. Mol. Diagn., **16**, 75-88.
3. *Xu H., DiCarlo J., Satya R.V., Peng Q., Wang Y.* (2014) BMC Genomics, **15**, 244.
4. *O'Rawe J., Jiang T., Sun G., Wu Y., Wang W., Hu J. et al.* (2013) Genome Med., **5**, 28.
5. *Kim S.Y., Jacob L., Speed T.P.* (2014) BMC Bioinformatics, **15**, 154.
6. *Martin M.* (2011) EMBnet J., **17**, 10-12.
7. *Li H., Durbin R.* (2009) Bioinformatics, **25**, 1754-1760.
8. *Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N. et al.* (2009) Bioinformatics, **25**, 2078-2079.
9. *Cibulskis K., Lawrence M.S., Carter S.L., Sivachenko A., Jaffe D., Sougnez C. et al.* (2013) Nat. Biotechnol., **31**, 213-219.
10. *Cingolani P., Platts A., Wang le L., Coon M., Nguyen T., Wang L., Land S.J., Lu X., Ruden D.M.* (2012) Fly (Austin), **6**, 80-92.
11. *Forbes S.A., Beare D., Bindal N., Bamford S., Ward S., Cole C.G. et al.* (2016) Curr. Protoc. Hum. Genet., **91**, 10.11.1-10.11.37. DOI: 10.1002/cphg.21.
12. *Liu X., Wu C., Li C., Boerwinkle E.* (2016) Hum. Mut., **37**, 235-241.
13. *Landrum M.J., Lee J.M., Benson M., Brown G., Chao C., Chitipiralla S. et al.* (2016) Nucl. Acids Res., **44**, D862-D868.
14. *Stenson P.D., Ball E.V., Mort M., Phillips A.D., Shiel J.A., Thomas N. et al.* (2003) Hum. Mut., **21**, 577-581.
15. *Ng P.C., Henikoff S.* (2003) Nucl. Acids Res., **31**, 3812-3814.
16. *Adzhubei I., Jordan D.M., Sunyaev S.R.* (2013) Curr. Protoc. Hum. Genet. DOI: 10.1002/0471142905.hg0720s76.
17. *Auton A., Brooks L.D., Durbin R.M., Garrison E.P., Kang H.M., Korbel J.O. et al.* (2015) Nature, **526**, 68-74.
18. *Lek M., Karczewski K.J., Minikel E.V., Samocha K.E., Banks E., Fennell T. et al.* (2016) Nature, **536**, 285-291.

Поступила: 31. 08. 2017.
Принята к печати: 14. 09. 2017.

A BIOINFORMATIC PIPELINE FOR NGS DATA ANALYSIS AND MUTATION CALLING IN HUMAN SOLID TUMORS

K.Yu. Tsukanov[1], A.Yu. Krasnenko[1], D.A. Plakhina[1], D.O. Korostin[2], A.V. Churov[3], O.S. Druzhilovskaya[2], D.V. Rebrikov[2], V.V. Ilinsky[1]

[1]"Genotek Ltd", Moscow, Russia
[2]Vavilov Institute of General Genetics, Moscow, Russia; e-mail: dep-nano@yandex.ru
[3]Institute of Biology of Karelian Research Centre, Petrozavodsk, Russia

We aimed to develop a pipeline for the bioinformatic analysis and interpretation of NGS data and detection of a wide range of single-nucleotide somatic mutations within tumor DNA. Initially, the NGS reads were submitted to a quality control check by the Cutadapt program. Low-quality 3′-nucleotides were removed. After that the reads were mapped to the reference genome hg19 (GRCh37.p13) by BWA. The SAMtools program was used for exclusion of duplicates. MuTect was used for SNV calling. The functional effect of SNVs was evaluated using the algorithm, including annotation and evaluation of SNV pathogenicity by SnpEff and analysis of such databases as COSMIC, dbNSFP, Clinvar, and OMIM. The effect of SNV on the protein function was estimated by SIFT and PolyPhen2. Mutation frequencies were obtained from 1000 Genomes and ExAC projects, as well as from our own databases with frequency data. In order to evaluate the pipeline we used 18 breast cancer tumor biopsies. The MYbaits Onconome KL v1.5 Panel ("MYcroarray") was used for targeted enrichment. NGS was performed on the Illumina HiSeq 2500 platform. As a result, we identified alterations in BRCA1, BRCA2, ATM, CDH1, CHEK2, TP53 genes that affected the sequence of encoded proteins. Our pipeline can be used for effective search and annotation of tumor SNVs. In this study, for the first time, we have tested this pipeline for NGS data analysis of samples from patients of the Russian population. However, further confirmation of efficiency and accuracy of the pipeline is required on NGS data from larger datasets as well as data from several types of solid tumors.

Key words: mutation, bioinformatic pipeline, high-throughoutput sequencing, bioinformatic analyses of NGS data