

Многофункциональный программный конвейер обработки данных HTS, полученных с использованием высокомультиплексных праймерных панелей

Гуков Б.*, Стеценко И., Мацвай А., Шипулин Г.

Федеральное государственное бюджетное учреждение «Центр стратегического планирования и управления медико-биологическими рисками здоровью» Федерального медико-биологического агентства, Москва, Россия

* bgukov@cspfmba.ru

Ключевые слова: NGS; патоген; метагеномика; программа; вирус

Мотивация и цель: Вирусы считаются самой распространенной формой жизни на нашей планете. В настоящее время известно почти 13 миллионов геномных последовательностей вирусов, и это число с каждым годом растет экспоненциально. При этом считается, что большинство известных вирусов, вызывающих заболевания у человека, первоначально имели зоонозный характер и были переданы людям из популяций животных. Поэтому важность мониторинга природных резервуаров инфекций не нуждается в пояснении. Здесь обращают на себя особое внимание рукокрылые, которые на сегодняшний день считаются одним из самых важных природных резервуаров потенциально зоонозных инфекций, поскольку отличаются исключительным видовым разнообразием – они составляют примерно 20 % всех видов млекопитающих. Главное оружие в борьбе с будущими зоонозами – это превентивный контроль, мониторинг и оценка путей распространения известных патогенов, а также выявление новых патогенов в природных очагах инфекций. Выявление и изучение свойств новых, не описанных ранее вирусов тоже является важной задачей, поскольку достоверно неизвестно, какой именно патоген окажется способным вызвать следующую эпидемию зоонозного заболевания. Применение в данной области секвенирования, помимо выявления этиологического фактора заболевания, позволяет провести генотипирование любого локуса, реконструкцию полного генома для любого типа образца и любого типа патогена. Применение высокомультиплексных праймерных панелей широкого скрининга для таргетного обогащения позволяет повысить эффективность исследования, существенно снизив его себестоимость, в том числе за счет уменьшения количества данных секвенирования, необходимого для анализа одного образца. Однако, в силу частого использования при секвенировании «метода дробовика», существующие для решения поставленной задачи программы разработаны с расчетом на анализ именно таких образцов, что не позволяет произвести качественный и быстрый анализ данных другого типа. При этом за счет малого количества данных, необходимых для анализа каждого образца, количество образцов в одном запуске может быть значительно больше, чем при исследовании с использованием «метода дробовика», что приводит к тому, что время, необходимое на биоинформатическую обработку данных, начинает составлять значительную долю общего времени, необходимого для проведения исследования. Кроме того, существующие решения не специализируются на анализе образцов, полученных от животных, и

не позволяют провести обработку эндогенных примесей, в лучшем случае отфильтровывая их, без извлечения дополнительной информации.

Методы и алгоритмы: Программа *pathogenid* предназначена для анализа нуклеиновых кислот патогенов и предоставляет следующий функционал: фильтрация исходных прочтений по качеству и длине, объединение парных прочтений в один файл в случае парно-концевого секвенирования, обработка прочтений на предмет принадлежности к геному хозяина, кластеризация последовательностей с целью сокращения времени анализа и разделения объектов различных групп между собой, построение консенсусных последовательностей, сравнение схожести прочтений с последовательностями из референсных нуклеотидных баз данных, а в случае анализа нуклеиновых кислот вирусов и сравнение консенсусных последовательностей с базой данных аминокислотных последовательностей, аннотация прочтений, схожих с референсными последовательностями, по таксономической базе данных, расчет относительного и абсолютного содержания нуклеиновых кислот, интересующих пользователя в контексте задачи, в исследуемом образце, создание отчета о результатах анализа образцов в форматах Excel-таблицы и JSON. Программа обрабатывает данные, поступающие с высокопроизводительных платформ секвенирования как второго, так и третьего поколений.

Результаты: Было разработано программное обеспечение *pathogenid*, адаптированное для задачи изучения виroma природных резервуаров инфекций. Программа апробирована на выборке из 713 биологических образцов, полученных от рукокрылых (*Chiroptera*), представителей 29 различных видов, в основном рода *Myotis* ($N = 429$, 60.17 %), пойманных по 16 субъектам Российской Федерации. Наибольшее количество образцов было собрано в Амурской ($N = 165$) и Алтайской ($N = 111$) областях. Исследование проводилось с использованием комбинации из семи праймерных панелей, состоящих суммарно из 40 415 первичных структур, нацеленных на обнаружение 52 родов вирусов, относящихся к 23 различным семействам. В результате анализа данных высокопроизводительного секвенирования были обнаружены последовательности нуклеиновых кислот 217 известных групп вирусов и 343 предположительно новых вирусов животных. Среди обнаруженных семейств чаще остальных встречалось семейство *Coronaviridae*. Представители семейства были обнаружены в 348 образцах, что составляет 48.1 % от всех проанализированных образцов. Полученные результаты, а также факты присутствия нуклеиновых кислот новых вариантов вирусов были подтверждены метагеномными исследованиями. Мы показали, что с использованием того же набора данных возможно эффективное определение таксономической принадлежности носителя вируса, что может быть важной задачей метагеномных исследований. Определение таксономии хоста возможно при условии наличия соответствующей геномной информации в открытых источниках. В среднем анализ одного образца, включая поиск гомологичных аминокислотных последовательностей, занимает 15 минут.

Выводы: В процессе работы была разработана программа *pathogenid* – универсальный биоинформатический конвейер обработки данных секвенирования, адаптированный для использования в задачах анализа виroma природных резервуаров инфекций. Данная программа подходит для образцов любой природы (при соответствующем методе экстракции), патогенов любой природы, обнаружения новых вирусов, анализа результатов секвенирования на любой платформе второго и третьего поколения. При наличии известных геномных данных животного-носителя,

pathogenid позволяет одновременно определять таксономическую принадлежность хоста. Пайплайн настроен для таксономической классификации патогенов с точностью до вида/штамма. Мы апробировали данный алгоритм анализа в проекте, включающим более 700 образцов, и подтвердили полученные результаты методом метагеномного секвенирования. Таким образом, мы можем проводить массовый скрининг биологического материала, обрабатывая большое количество образцов в короткие сроки, проводя при этом скрининг сразу на все значимые группы патогенов одновременно, выявляя в том числе новые, не известные ранее вирусы.

Финансирование: Исследование выполнено при поддержке государственного задания № 388-00084-24-00.

Multifunctional software pipeline for processing HTS data obtained using highly multiplexed primer panels

Gukov B.*, Stetsenko I., Matsvay A., Shipulin G.

Federal State Budgetary Institution "Centre for Strategic Planning and Management of Biomedical Health Risks" of the Federal medical and biological agency, Moscow, Russia

* bgukov@cspfmba.ru

Key words: NGS; pathogen; metagenomics; programm, virus

Motivation and Aim: Viruses are considered as the most prevalent form of life on our planet. Currently, almost 13 million genomic sequences of viruses are known and this number is growing exponentially every year. It is widely believed that the majority of known viruses responsible for human diseases originated from zoonotic transmission, meaning they were initially transmitted to humans from animal populations. Therefore, the significance of monitoring natural reservoirs of infections is self-evident. Bats (Chiroptera) merit particular attention in this regard, as they are now recognized as one of the most significant natural reservoirs of potentially zoonotic infections. Their exceptional species diversity, comprising roughly 20 % of all mammal species, underscores their importance in this context. The primary strategy in combating future zoonotic diseases involves preventive control, monitoring, and assessing the spread of both known pathogens and the identification of new pathogens in natural infection foci. Additionally, the identification and investigation of new, previously unknown viruses are crucial tasks, because pathogen, which will be capable of causing the next epidemic of zoonotic disease, cannot be accurately predicted. The utilization of sequencing goes beyond merely identifying the causative agent of the disease. It enables genotyping of any locus and reconstruction of the complete genome for any sample type and pathogen species. The use of broad screening panels for targeted enrichment can increase the efficiency of the study, significantly reducing its cost. The widespread adoption of the shotgun method in metagenomics has resulted in the specialization of existing software solutions for analyzing such samples. Consequently, high-quality and rapid analysis of other types of data is difficult. Additionally, current solutions lack specialization in analyzing animal samples and struggle to effectively manage endogenous contaminants, typically resorting to filtering without extracting additional information.

Methods and Algorithms: The "pathogenid" is pipeline intended for the analysis of nucleic acids of pathogens and provides the following functionality: filtering of source

reads by quality and length; combining paired reads into one file in the case of paired-end sequencing; processing of reads for belonging to the host genome; clustering of sequences in order to reduce analysis time and separate objects of different groups from each other; construction of consensus sequences; comparison of the similarity of reads with sequences from reference nucleotide databases; comparison of consensus sequences with a database of amino acid sequences, annotation of reads similar to reference sequences in a taxonomic database, calculation of the relative and absolute contents of nucleic acids of interest to the user in the context of the problem in the sample under study, creation of a report on the results of analysis of samples in Excel table and JSON formats. The program processes data coming from high-throughput second- and third-generation sequencing platforms.

Results: Pathogenid software has been developed and adapted for the task of studying the virome of natural reservoirs of infections. The program was tested using a set of 713 biological samples collected from Chiroptera, including representatives of 29 different species, mainly the genus *Myotis* (N = 429, 60.17 %), collected in 16 regions of the Russia. The largest number of samples were collected in the Amur (N = 165) and Altai (N = 111) regions. The study was conducted using a combination of seven primer panels, consisting of a total of 40,415 primary structures, aimed at detecting 52 genera of viruses belonging to 23 different families. Nucleic acids from 217 known groups of viruses and 343 presumably new animal viruses were discovered. The results obtained, as well as the facts of the presence of nucleic acids of new variants of viruses, were confirmed by metagenomics. Among all results, the *Coronaviridae* family was most common. Representatives of the *Coronaviridae* family were found in 348 samples, which is 48.1 % of all analyzed samples. We demonstrate that, utilizing the same dataset, it is possible to efficiently determine the taxonomic identity of the virus host, which can be an important goal of metagenomic studies. The determination of host taxonomy was achievable given the availability of suitable genomic information from open sources. On average, the analysis of one sample, inclusive of searching for homologous amino acid sequences, took 15 minutes.

Conclusion: Pathogenid – a universal bioinformatics pipeline for processing sequencing data, adapted for use in analyzing the virome of natural reservoirs of infections. This program is suitable for samples of any nature (with the appropriate extraction method), pathogens of any nature, detection of new viruses, analysis of sequencing results on any second and third generation platform. Given known genomic data from the host animal, pathogenid allows the taxonomic affiliation of the host to be simultaneously determined. The pipeline is configured for taxonomic classification of pathogens with species/strain accuracy. We tested this analysis algorithm in a project involving more than 700 samples and confirmed the results obtained using metagenomic sequencing. Thus, we are capable of conducting mass screening of biological material, processing a large number of samples swiftly. This enables simultaneous screening for all significant groups of pathogens, including the identification of new, previously unknown viruses.

Funding: The study was supported by the state assignment No. 388-00084-24-00.