**ОБРАБОТКА ДЛИННЫХ ЧТЕНИЙ ТРАНСКРИПТОМНОГО СЕКВЕНИРОВАНИЯ С ПРИМЕНЕНИЕМ ОБЛАЧНОЙ ВЫЧИСЛИТЕЛЬНОЙ ПЛАТФОРМЫ AMAZON WEB SERVICES**

Шаповалова В. В.1, Радько С. П.2, Птицын К. Г. 2, Краснов Г. С. 2, Наход К. В.\* 2, Конаш О. С. 2, Виноградина М. А. 2, Пономаренко Е. А. 2, Дружиловский Д. С. 2, Лисица А. В. 2

1 - Центр стратегического планирования и управления медико-биологическими рисками здоровью , 119121, г. Москва, ул. Погодинская, д. 10, стр.1

2 - Научно-исследовательский институт биомедицинской химии имени В. Н. Ореховича 119121, Россия, Москва, ул. Погодинская, д. 10, стр.8

\*g-s2011@mail.ru

**ПРИЛОЖЕНИЕ №1**

**Пошаговая инструкция для обработки результатов количественного анализа транскриптома человека с применением нанопорового секвенатора**

***ЭТАП 1***– Запуск и подключение к виртуальной машине

***ЭТАП 2*** - Выполнение программ на платформе AWS

***ЭТАП 3*** – Перенос результатов вычислений и выключение виртуальной машины

***ЭТАП 1*** *– Запуск и подключение к виртуальной машине*

Шаг 1. Запуск виртуальной машины на сервере AWS (Amazon Web Strvices).

Используя веб-браузер, перейдите на страницу предоставления доступа к ресурсам AWS. Для этого укажите в поисковой строке браузера (Google, Yandex и др.) “Amazon AWS” и перейдите по соответствующей ссылке. Найдите на веб-странице кнопку-ссылку [Вход в консоль] и введите авторизующую информацию, которая предоставляется авторами в виде дополнительных материалов к статье по запросу (рис. 1-1). На появившемся после авторизации экране выберите последовательно ссылки [EC2] (рис. 1-2) и далее – [Running Instances] (рис. 1-3). Затем «кликните» виртуальную машину Mr. First (ONT) (рис. 1-4) и с использованием опции верхнего меню [Actions] произведите запуск виртуальной машины, через выбор [instance state] как показано на рис. 2-5.

**Рисунок 1**. Запуск виртуальной машины на платформе Amazon Web Services.

**(1)** Ввод авторизующей информации для входа в пользовательскую консоль. **(2)** Консоль пользователя AWS, в которой нужно выбрать гиперссылку EC2 (Elastic Compute Cloud) для перехода к облачным ресурсам. **(3)** В списке ресурсов нужно выбрать [Running Instance], чтобы перейти к списку виртуальных машин для выбора машины Mr.First **(4).**

**Рисунок 2.** Запуск, выключение виртуальной машины и смена ее конфигурации Запуск (5) и выключение (6) виртуальной машины в меню [Instance state]. **(6)** Изменение конфигурации виртуальной машины (7) в меню [Instance setting].

Обратите внимание, что изменение конфигурации виртуальной машины может быть осуществлено только после ее выключения, затем требуется повторный запуск. При выключении соединение с виртуальным компьютером PuTTY/WinSCP обрывается и его нужно возобновить (см. далее).

Шаг 2. Соединение с виртуальным компьютером.

Для подключения к выбранной виртуальной машине. платформы Amazon запустите на компьютере с ОС Windows программу-терминал PuTTy (рис. 3-1). Соединение с удаленным компьютером потребует ввести (а) адрес удаленной машины (адрес “хоста”), который копируется из консоли управления облачными сервисами (рис. 3-2, рис. 3-3). и (б) путь к файлу-ключу, которые обеспечивает защиту доступа к удаленной машине (рис. 4-4, рис. 4-5). Файл-ключ был сгенерирован при создании машины Mr. First (ONT) и наименован bober.ppk (доступен в составе дополнительных материалов к данной статье, рекомендуется его скопировать на рабочий стол Windows). При загрузке терминала ввести данные по запросу login: «ubuntu». **(рис. 5-6)**.

В появившемся терминале на черном фоне появится приглашение для ввода («промт») в виде знака «$», указывая на готовность терминала к работе (рис. 5-7).

**Рисунок 3.** Подключение к файловой системе удаленной машины с использованием программы PuTTy (<https://www.putty.org/>). **(1)** Выбор виртуальной машины, находящейся в режиме «Running», переход в пункт меню [Connect]. **(2**) В окне [Connect to instance] скопировать строку, содержащую временный IP-адрес виртуальной машины. **(3)** Скопированный адрес нужно вставить в поле “Host Name (or IP address)” программы PuTTy.

**Рисунок 4.** Подключение к файловой системе удаленной машины с использованием программы PuTTy (<https://www.putty.org/>). **(4)** В иерархическом меню программы PuTTy выбрать подпункт Connections > SSH > Auth и, используя кнопку [Browse], указать в поле «Public key file for authentication» имя файла открытого ключа .ppk **(5)**, который входит состав дополнительный материалов к статье (предоставляется по запросу).

 **Рисунок 5.** Подключение к файловой системе удаленной машины с использованием программы PuTTy (<https://www.putty.org/>). **(6)** При загрузке терминала ввести данные по запросу login: «ubuntu». **(7)** Терминал готов к работе.

Установите соединение [connect] с использованием файлового менеджера WinSCP (рис. 6-1). Для этого следует указать имя “хоста” (рис. 6-2, 6-3). и путь к файл-ключу (bober.ppk, см. рис. 6-4, рис. 6-5), вокне авторизации **(6-6)** с использованием кнопки[войти] (6-7) открывается окно файлового менеджера для работы с файлами на сервере AWS (рис. 6-8).

**Рисунок. 6.** Подключение к файловой системе удаленной машины с использованием программы WinSCP ([https://winscp.net](https://winscp.net/eng/docs/lang:ru)). **(1)** Установка соединения с использованием опции [connect] **(2)** Копирование адреса удаленной машины (адрес «хоста») из консоли управления облачными сервисами в **(3)** систему управления программы WinSCP.

**Рисунок. 7.** Подключение к файловой системе удаленной машины с использованием программы WinSCP ([https://winscp.net](https://winscp.net/eng/docs/lang:ru)**).(4)** В иерархическом меню программы WinSCP выбрать подпункт аутентификация и, используя кнопку […], и указать имя файла открытого ключа .ppk **(5)**, который входит состав дополнительный материалов к статье (предоставляется по запросу). Указание файла-ключа bober.ppk **(5)** вокне авторизации **(6)** с использованием кнопки[войти] **(7)** открывается доступ к дискам данных на локальном сервере и на виртуально машине AWS**.(8).**

При подключении с использованием утилит PuTTy/WinSCP появляется предупреждение об отсутствии файла-ключа в кэше, которое нужно игнорировать выбирая опцию “Да”.

Шаг 3. Проверка наличия свободного виртуального дискового пространства на виртуальной машине.

Команда операционной системы (ОС) Linux “df” позволяет получить информацию о наличии свободного дискового пространства в директории work\, которое представляет собой виртуальное хранилище данных (диск, или volume в терминологии Amazon Web Services), подключенное к виртуальной машине. Запуск команды “df -h” осуществляется в терминале PuTTy и выводит на экран информацию о наличии свободного пространства на дисках (параметр “-h”, указываемый после команды “df” означает “human readable”, то есть результат выполнения команды выводится в интуитивно понятном пользователю виде). Если рабочий диск, в частности директория work\, занята более чем на 60%, то следует ее освободить от результатов предыдущих запусков пайп-лайна (см. раздел “Завершающий этап”).

Шаг 3а. Копирование данных с использованием файлового менеджера WinSCP.

Для копирования данных рекомендуется использовать файловый менеджер WinSCP (см. рис. 8). Перед копированием данных на облачную платформу AWS, определите размер и расположение исходной папки с результатами секвенирования на локальном компьютере (или на СХД организации) и назовите ее условно fast5. Для копирования нужно переместить в графическом интерфейсе WinSCP папку с файлами из файловой панели локального компьютера в директорию work/ файловой панели удаленной виртуальной машины. Копирования 100 Гбайт займет в среднем 8-10 часов, поэтому целесообразно ставить копирование на ночь, при этом используя наиболее дешевую конфигурацию удаленного компьютера, например, t2.micro (~1 цент в час), плюсом к этому идет оплата за передачу данных. После окончания процедуры передачи файлов на виртуальную машину AWS, используйте утилиту PuTTy для оценки объема занимаемого виртуального дискового пространства в директории work/ командой “du - h” и убедитесь, что размер скопированной папки соответствует исходному оригиналу.



**Рисунок 8.** Экран программы WinSCP: перенос файлов формата fast5 с локального компьютера или СХД организации в директорию work/fast5 на виртуальной машине облачной платформы Amazon Web Services.

***ЭТАП 2*** *- Выполнение программ на платформе AWS*

Программа WinSCP используется, чтобы загрузить в директорию work\ на виртуальный диск облачной машины AWS данные в формате fast5. Субдиректория work\fast5\ указывается в качестве входного параметра программе guppy\_basecaller, а результат выполнения «бейз-коллинга» генерируется в субдиректорию guppy\_results. Потом данные анализируются программой контроля качества секвенирования, а далее, файлы формата fastQ, прошедшие валидацию и находящиеся в субдиректории guppy\_results/pass/, объединяются в один файл bigfile.fastQ командой “cat”. Этот файл в свою очередь является входным для процедуры выравнивания ридов на транскриптом программой Minmap2. Результат картирования, в виде файла aln.sam передается для статистического анализа качества выравнивания программой Samtools stats с последующим запуском программы оценки количества транскриптов программой Salmon quant.

На выходе выполнения конвейера обработки ONT-данных генерируется набор директорий и файлов, перечисленных в разделе описания результатов (см. табл. 2 в статье).

Шаг 4. «Бейз-коллинг».

Перед выполнением процедуры «бейз-коллинг» переведите виртуальную вычислительную машину в конфигурацию p3.2xlarge (рис. 2-7).

В окне терминала перейдите в рабочую директорию, используя команду «cd» (change directory):

$ cd /home/ubuntu/work/

В директории work/, где ранее была подготовлена папка с файлами в формате fast5, запустите программу guppy\_basecaller (команда C1), или в режиме тестирования используете команду (C1a), в которой вместо папки fast5 указана папка fast5.tutor:

|  |  |  |
| --- | --- | --- |
| **(C1)\*** | :**~/work**$ guppy\_basecaller –i fast5 –s guppy\_results  -qscore\_filtering –flowcell FLOW-MIN106 -kit SQKRNA002  -device auto | \  \ |
| **(C1a)\*** | :**~/work**$ guppy\_basecaller -i fast5.tutor -s guppy\_results  -qscore\_filtering --flowcell FLOW-MIN106 -kit SQKRNA002  -device auto | \  \ |
| \*для копирования текста исполняемой команды используйте Приложение №2. | | |

Обратите внимание на значки “backslash” (“\”) в составе команды. Данное обозначение указывает, что команда является единой строкой, но разнесена по нескольким строчкам для удобства представления в статье. Для выполнения протокола скопируйте в окно терминал текст команды из приложения №1, далее перейдите в окно терминала программы PuTTy и кликните правой кнопкой «мыши», что является аналогом функции Paste.

Командой (С1) с указанием параметра -i (input) задаем в качестве входной субдиректорию fast5, которая содержит ранее скопированные выходные файлы, полученные от нанопорового секвенатора, а в качестве субдиректории, в которую требуется «сохранить» обработанные данные (параметр -s, save), указываем субдиректорию guppy\_results (см. С1).

ВНИМАНИЕ! В стандартной системе символов, ASCII (American Standard Code for Information Interchange) есть два похожих символа - “минус” (короткая горизонтальная черточка) и “тире” (длинная горизонтальная черточка). В командной строке операционной системы семейства Unix используются только знаки “минус”, а при выполнении процедуры «бейз-коллинг» программа Guppy\_basecaller выводит «индикатор прогресса», отражающий процент обработанных данных (рис. 9). После выполнения, создается субдиректория guppy\_results/, в pass/ субдиректории которой находятся файлы формата fastQ, необходимые для выполнения дальнейших шагов (рис. 10). (Справочно: чтобы просмотреть список файлов в субдириктории guppy\_results/pass можно использовать команду «ls guppy\_results/pass»).



**Рисунок 9.** Индикатор прогресса выполнения процедуры «бейз-коллинг».

После завершения работы guppy\_basecaller, выведите виртуальную машину из режима p3.2xlarge и переведите в сберегающий режим t2.2large (или р3.micro) см. (рис. 2-6).



**Рисунок 10.** После процедуры «бейз-коллинг» образуется субдиректория guppy\_results.

Шаг 5 (опциональный). Оценка качества результатов процедуры «бейз-коллинг».

Оценка качества (QC, quality control) производится с использованием скрипта MiniIONQC.R, разработанного на языке программирования R. (см. команду C2). Этот я зык был прдустановлен в ходе подготовки виртуальной машины.

|  |  |
| --- | --- |
| **(C2)\*** | :**~/work**$ MinIONQC.R \  -i guppy\_results\sequencing\_summary.txt –o qc\_results |

\*для копирования текста исполняемой команды используйте Приложение №2.

На вход команде (С2) подается файл sequencing\_summary.txt, полученный в результате выполнения программы guppy\_basecaller, в качестве входной директории, в которой будут размещены результаты контроля качества, qc\_results (рис. 11).



**Рисунок 11.** Директория qc\_results, в которой размещаются результаты контроля качества.

Шаг 6. Картирование (Minimap2) и статистика картирования (samtools stats)

Чтобы продолжить работу, необходимо объединить набор файлов, получившихся в результате работы программы guppy\_basecaller, в один файл формата fastQ. Для этого применяется стандартная команда “cat” операционной системы Linux/Ubuntu (см. С3). В результате выполнения команды С3, файлы, размещенные в папке guppy\_results/pass/ будут объединены в один файл, который мы назвали bigfile.fastQ (рис. 12). Используйте команду “ls –hl”, чтобы проверить, что файл ,был успешно создан и имеет размер, сопоставимый с размером генома человека (несколько гигабайт, как минимум). Справочно: геном человека занимает объем примерно 3 гигабайта.

|  |  |
| --- | --- |
| **(C3)\*** | :**~/work**$ cat guppy\_results/pass/\*.fastq > bigfile.fastQ |

**Рисунок 12.** Файл bigfile.fastQ в результате выполнения команды “cat”.

|  |  |
| --- | --- |
| **(C3)\*** | :**~/work**$ cat guppy\_results/pass/\*.fastq > bigfile.fastQ |

\*для копирования текста исполняемой команды используйте Приложение №2.

Находясь в директории work/, запустите программу Minimap2 с целью картирования результатов секвенирования на референсный транскриптом, как указано в команде (C3). В качестве исходных параметров для программы Minimap2 указывается наименование входного файла с «ридами» в формате fastq и референсный транскриптом – gencode.v32.transcripts.fa. Референсный транскриптом размещен в папке ref.

|  |  |
| --- | --- |
| **(C4)\*** | :**~/work**$ minimap2 -ax map-ont -N 100 \ ref/gencode.v32.transcripts.fa bigfile.fastQ > aln.sam |

**Рисунок 13.** Файл aln.sam, полученный в результате картирования результатов секвенирования на референсный транскриптом.

|  |  |
| --- | --- |
| **(C4)\*** | :**~/work**$ minimap2 -ax map-ont -N 100 \ ref/gencode.v32.transcripts.fa bigfile.fastQ > aln.sam |

\*для копирования текста исполняемой команды используйте Приложение №2.

Команда (С4) обозначает вызов программы картирования (minimap2) и определяет для этой программы параметры выполнения (в частности, параметр –ax указывает, что картировать нужно “риды”, полученные по технологии оксфордской нанопоры, map-ont), далее указывается на файл транскриптома в FASTA-формате (gencode.v32.transcripts.fa), который расположен в субдиректории work/ref. Затем указывается имя файла с прочтениями секвенирования bigfile.fastQ, который был получен в результате выполнения команды (C3). Еще раз напомним, что символ \ в синтаксисе команд означает, что строка должна быть единой, как приведено в Приложении №2.

Шаг 7 (опциональный). Статистика результатов картирования.

Результаты картирования генома (предыдущий шаг, команда minimap2) направляются в файл aln.sam (рис. 13). Проверьте командой “ls –l”, что такой файл создан, и он не нулевого размера. Далее выполните команду (C5), позволяющую получить статистическую справку о результатах картирования генома. Для этого применяется программа Samtools с параметром “stats”. В качестве входного файла указывается aln.sam, (сгенерированный в результате картирования, см. (С4), а результат помещается в файл samtools\_stats.txt (рис. 14).

|  |  |
| --- | --- |
| **(C5)\*** | :**~/work**$ samtools stats aln.sam > samtools\_stats.txt |

\*для копирования текста исполняемой команды используйте Приложение №2.

**Рисунок 14.** Файл samtools\_stats.txt, содержащий статистические сведения о результатах картирования нанопоровских прочтений на транскриптом человека.

Шаг 8. Квантификация транскриптов.

Выполните команду (C6), предназначенную для запуска программы salmon с базовым параметром “quant”. Результат исполнения программы будет размещен в рабочей директории work/ в папке salmon\_quant.out (рис. 15),. Дождитесь завершения выполнения команды и убедитесь, что результирующая папка была создана с помощью команды «ls –hl”.

|  |  |
| --- | --- |
| **(C6)\*** | :**~/work**$ salmon quant -t ref/gencode.v32.transcripts.fa -l U \  -a aln.sam -o salmon\_qnt.out |

**Рисунок 15.** Результат исполнения программы salmon в папке salmon\_quant.out.

Рассмотрим параметры выполнения команды (С6), принимая во внимание, что символ «\» означает, что команда выполняется без переносов строки (см. Приложение №1). Параметр «–t» указывает на файл, содержащий транскриптом человека; в данном случае: ref/gencode.v32.transcripts.fa, где ref/ это субдиректория в директории work/, в которой размещен данный файл в FASTA формате. Следующий параметр «–а» указывает на файл формата SAM (Sequence Alignment Map), который был создан в результате работы программы minimap2 на этапе выполнения команды (С4). Наконец, параметр «–о» (output) необходим, чтобы указать субдиректорию, в которую должны быть помещены результаты работы программы Salmon.

Шаг 9. Выборка по белок-кодирующим генам.

Ключевым объектом в субдиректории salmon\_quant\_out является файл quant.sf (рис. 13), который содержит в табличном виде наименование гена, транскрипта и соответствующее ему количество ридов, выраженное либо в абсолютных, либо в нормированных (TPM - transcripts per million) единицах. Файл имеет размер более 2 Гб, так как содержит информацию о разных типах транскриптов, включая длинные некодирующие РНК, рибосомальные и транспортные РНК и др. Поэтому перед загрузкой файла целесообразно сократить его размер, отобрав транскрипты, соответствующие белок-кодирующим генам. Для этого используем команду “grep” (из директории work/). Выполнение команды “grep” позволяет отобрать из результирующего файла work/salmon\_qnt.out/quant.sf только те строки, которые относятся к белок-кодирующим генам (рис. 16).

~/work$ grep “protein\_coding” salmon.qnt.out/quant.sf > quant\_protcoding.sf

**Рисунок 16.** Выполнение команды grep "protein\_coding" и сведения о размере выходного файла quant\_protcoding.sf.

***ЭТАП 3*** *– Перенос результатов вычислений и выключение виртуальной машины*

Шаг 10. Копирование файлов.

Используйте утилиту WinSCP, чтобы скопировать файлы на локальный компьютер или на диск организации, отведенный для хранения информации (рис. 5). Не все из них нуждаются скачивании, для дальнейшей работы вполне достаточно скачать файл quant\_proteincoding.sf, объем которого при анализе транскриптома человека с использованием нано поры составляет всего 12 мегабайт. (рис. 16).

Шаг 11. Завершение работы.

С использованием сайта AWS выберите виртуальную машину MrFirst ONT и далее используйте меню, чтобы остановить работу с использованием пункта меню «instate state» (рис. 2-6).