2019-10-28

Практическое занятие по обработке данных NGS.

Цель нашего занятия - взяв сырые данные с секвенатора (в формате FASTQ), получить из них набор вариантов (отличий от референсного генома) в файле формата VCF.

Основные стадии процесса:
1. Проверить качество данных (FastQC). Убедиться, что секвенирование прошло нормально и с данными можно работать.
2. Осуществить выравнивание на референсный геном.
3. Предобработать файл с выравниванием (BAM): маркировать дубликатные прочтения, перевыровнять регионы инделов, рекалибровка качества.
4. Определение вариантов, аннотация и фильтрация.

Занятие будет проходить на платформе Galaxy (http://usegalaxy.org), содержащей интерактивный интерфейс приложений. В реальности большинство операций на потоке выполняются из интерфейса командной строки.

Данные и доступ на платформу.

Для начала, стоит войти в свой аккаунт в системе Galaxy (или зарегистрироваться, если у Вас еще нет своего аккаунта). Это можно сделать на вкладке Login or Register. Данные можно загрузить в Galaxy двумя основными путями.

Мы будем работать с данными секвенирования одного из образцов, которые анализировались в Институте биоинформатики. Эти данные ограничены до нескольких участков экзома, чтобы с ними было легче и удобнее работать. Этот фрагмент описан в специальном формате типа BED, который называется “MODY_Genes.bed”. Данные и BED-файл доступны по ссылке. Данные секвенирования доступны в виде четырех файлов: образец был секвенирован на двух дорожках секвенатора и с использованием парных прочтений.

https://drive.google.com/open?id=105JgjKB69LGNCC-79fW0sA4R0oB2CtfH

Загрузите данные и BED-файл к себе на компьютер. Далее загрузите их в Galaxy, для этого:

1. *Выбрать пункт меню **“Get Data”**, а затем - **“Upload file”**. В*

*появившемся окне можно загрузить файлы **(все четыре .fastq.gz и BED)** напрямую по ссылке (пункт **"Paste/Fetch Data"** внизу окна) или со своего компьютера (пункт **"Local file"**). Для Fastq.gz файлов выберите тип файлов fastqsanger.gz. Нажать **"Start"**, дождаться окончания загрузки файлов. Проверить, что файлы появились в меню справа.*

Контроль качества (Quality Control/QC).

Оценим качество секвенирования. Для этого запустим программу FastQC, которая создаст графический отчет с различными характеристиками.

*2. Выбрать пункт меню "FASTQ Quality Control", далее "FastQC". Указать только первый пункт (short read data). Нажать **"Execute" (во всех пунктах далее подразумевается под словом "запустить")**. Дождаться окончания работы программы.*

*3. Выбрать в окне справа отчет **FastQC (webpage)**. Открыть его, нажав на пиктограмму с изображением глаза.*

Выравнивание прочтений (read alignment/mapping).

Выравнивание мы будем производить при помощи инструмента BWA MEM. Этот пакет является наиболее стандартным решением для работы с медицинскими данными экзомного или геномного секвенирования ввиду наибольшей аккуратности в оценке качества выравнивания.

*4. Для запуска программы выберите **"Mapping",** затем выберите в списке **BWA MEM**.*

*5. **Внимание! Эту стадию нужно проводить отдельно для каждой пары ридов (L001.R1/2 и L002.R1/2).** Задайте необходимые параметры: выберите референсный геном **"Human (Homo sapiens) (b37): hg_g1k_v37"**; средний размер вставки - 100 н. В поле **"Set Read Groups"** выберите опцию **"Set Read Groups (Picard style)"**. Заполните поля по образцу (используйте вместо X в Lane_X цифру, соответствующую номеру дорожки L00**1**/L00**2**).*

Read Group Identifier	*Sample_X.Run_1.Lane_X*
Read Group Sample Name	*Sample_X*
Library Name	*Sample_X.Library_1*
Platform/technology used to produce the reads (PL)	*ILLUMINA*

Platform Unit *HWI-XXXXX*

Остальные параметры - по умолчанию. Запустите программу, дождитесь окончания работы программы.

*6. Объедините BAM-файлы для каждой дорожки (L001/L002). Для этого выберите пункт **“SAM/BAM”** в меню слева, затем “**Merge BAM files**”. В качестве входных данных выберите результаты выравнивания данных **с обеих дорожек!***

*7. Отсортируйте BAM-файл по координате. Для этого выберите пункт **“SAM/BAM”** в меню слева, затем **“Sort”.** Запустите сортировку на выводе из предыдущего пункта с параметрами по умолчанию. Дождитесь окончания сортировки.*

Предобработка выравнивания.

*8. Выберите на вкладке слева опцию **“Picard”**, внутри которой - **“MarkDuplicatesWithMateCigar”**. Запустите на результате выполнения п. 7 (отсортированный BAM-файл) с настройками по умолчанию. Дождитесь окончания работы программы. Откройте файл со статистикой дубликатных прочтений (**metrics**). Какой процент прочтений являются дубликатными?*

В данном примере мы ограничимся лишь маркировкой дубликатных прочтений - это быстро, но очень важно для статистических выводов в дальнейшем.

*9. (опционально) Еще раз проверим качество секвенирования (подтвердим подозрения из п. 3). Для этого запустим еще одну программу из пакета Picard (**“NGS: Picard”**) - **CollectInsertSizeMetrics**. Найдите данный плагин в списке опций Picard в меню слева и запустите на результатах п. 6 (отсортированный BAM-файл) с настройками по умолчанию. Дождитесь окончания работы программы. Выберите в меню справа файл типа **“pdf”**. Откройте его нажатием на соответствующую пиктограмму.*

*10. (п. 10 и п. 11 важны для подготовки заключения в рамках экзаменационного задания, но на практическом занятии повторять их на своем компьютере не обязательно) Проанализируем покрытие целевых регионов с использованием **deepTools**.*

*Выберите пункт "**plotEnrichment**", в поле для ввода BAM-файла выберите результат стадии (8). В качестве целевых регионов выберите BED-файл **MODY_Genes.bed**. [в рамках своего экзаменационного задания используйте соответствующий Вашему заданию BED-файл] В пункте "**Save percentages to a file**" отметьте Yes.*

*11. Проанилизируем покрытие целевых регионов с использованием инструмента "**BedCov**" в меню "**SAM/BAM**". В опции "**Genemic intervals (in BED format)**" укажите файл **MODY_Genes.bed**. В качестве входного BAM-файла выберите результат стадии (8). Запустите инструмент. По окончанию работы Вы должны получить файл табличного формата. Воспользуйтесь инструментом "**Summary Statistics**" в меню "**Statistics**". В качестве входных данных выберите результат стадии "BedCov". В пункте "**Column or expression**" введите "**c4 / (c3 - c2)**". Запустите программу. На выходе Вы получите набор общих показателей покрытия целевых регионов.*

*Также полезно узнать общее количество прочтений в файле. Это можно сделать при помощи программы "**FlagStat**" в меню "**SAM/BAM**".*

Определение вариантов.

Для определения вариантов воспользуемся пакетом FreeBayes - в UseGalaxy это практически единственный доступный инструмент для анализа вариантов. Перед процедурой определения вариантов проведем также выравнивание вставок/выпадений в BAM-файле по левому краю.

*12. Выберите пункт меню "**Variant calling**" и найдите инструмент "**BamLeftAlign**". В поле "**Select alignment file in BAM format**" выберите файл, полученный в результате шага (8). Выберите референсный геном **Human (Homo sapiens) (b37): hg_g1k_v37**. Запустите инструмент.*

*13. Выберите пункт меню **"Variant calling"**. Найдите подпункт **"FreeBayes"**.*

*14. Выберите референсный геном **Human (Homo sapiens) (b37): hg_g1k_v37**. В поле **"Choose parameter selection level"** выберите **"Simple diploid calling with filtering and coverage"**. В качестве входных данных используйте результат стадии (12). Укажите минимальное покрытие по своему усмотрению в появившемся поле.*

Запустите FreeBayes и дождитесь окончания работы программы.

*15. (эту стадию на занятии можно не выполнять, но она пригодится при выполнении экзаменационных заданий). Выберите пункт "**VcfAllelicPrimitive**" в меню "**VCF/BCF**". В качестве входных данных используйте результат стадии (14). Запустите инструментю*

Аннотация вариантов.

*16. Стадию аннотации также можно выполнить в веб-версии программного пакета ANNOVAR (http://wannovar.wglab.org/). Для этого воспользуйтесь файлом-результатом стадии (15). Загрузите файл в поле **Input File**. Укажите дополнительную информацию (e-mail, название образца, сборку генома - **hg19**) и нажмите кнопку "**Submit**". По окончанию вычислений результат аннотации можно будет просмотреть в веб-форме или скачать к себе на компьютер.*

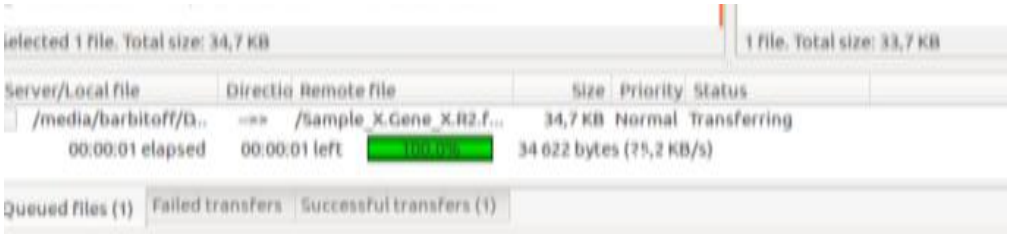
Дополнительная информация:

Загрузка файлов большого размера при помощи протокола FTP

1. Загрузите клиент FileZilla, установите его и запустите.
2. В верхней части окна клиента введите в соответствующие поля значения:

Host	ftp://usegalaxy.org
Username	Ваше имя пользователя
Password	Ваш пароль

3. Нажмите Quickconnect. Дождитесь подключения (будет Status: Directory listing of “/” succesful под полями адреса.
4. В нижней части окна клиента FileZilla отобразится содержание диркеторий на Вашем компьютере (Local site, слева) и на сервере Galaxy (Remote site, справа). Перетащите файлы .fastq.gz из директории на Вашем компьютере на удаленный сервер. Дождитесь окончания загрузки (прогресс будет отображаться в нижней части окна клиента):



5. В интерфейсе Galaxy выберите Get Data -> Upload Files -> Choose FTP file. В появившемся списке выберите загруженные Вами файлы. Перед импортом в рабочую среду не забудьте указать тип файла (см. основную часть)!

GATK Best Practices workflow:
https://software.broadinstitute.org/gatk/best-practices/

Оригинальные публикации BROAD Institute:

Van der Auwera, G.A. Carneiro, M.O., Hartl, C., Poplin, R., Angel, G. del, Levy-Moonshine, A., et al. (2013) From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr Protoc Bioinformatics* 11.10.1-11.10.33

DePristo, M.A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J.R., Hartl, C., et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43:** 491–498