Проект по биоинформатике: Анализ данных секвенирования ДНК

Казанцева Варвара Денисовна

13 июня 2024 г.

Содержание

1 Введение

Секвенирование ДНК является одной из ключевых технологий современной молекулярной биологии и генетики. Она позволяет определять последовательность нуклеотидов в ДНК, что имеет важное значение для понимания генетической основы различных биологических процессов и заболеваний.

В рамках данного проекта будет выполнен анализ данных секвенирования ДНК с целью выявления патогенных вариантов, которые могут быть связаны с развитием рака молочной железы у пациентов с наследственными мутациями синдрома Линча. Исследование основано на данных из проекта с идентификатором SRP229462, в котором рассматривается следующая гипотеза: мутации в генах, ассоциированных с синдромом Линча, могут быть напрямую связаны с развитием рака молочной железы.

2 Постановка задачи

Целью данного проекта является анализ данных секвенирования ДНК пациентов с раком молочной железы с наследственными мутациями синдрома Линча. Для этого были поставлены следующие задачи:

1. Изучить статью: "Мутации герминативной линии синдрома Линча при раке молочной железы".
2. Выбрать десять случайных образцов ДНК из данных, исследуемых в выбранной статье, и загрузить их в формате FASTQ.
3. Провести анализ, включающий в себя:
   - Контроль качества данных секвенирования;
   - Выравнивание ридов на референсный геном человека;
   - Поиск вариантов при помощи VarScan;
   - Описание вариантов в VEP.
4. Проанализировать наиболее заметные варианты, выявленные в ходе исследования, и обосновать их потенциальное влияние на прогрессирование заболевания.

3 Обзор статьи

Статья "Мутации герминативной линии синдрома Линча при раке молочной железы"авторов Алексея Никитина, Дарьи Чудаковой, Рафаэля Еникеева, Дины Сакаевой, Максима Дружкова, Лейлы Шигаповой, Ольги Бровкиной, Елены Шагимардановой, Олега Гусева, Марата Гордиева посвящена исследованию связи мутаций, связанных с синдромом Линча, и раком молочной железы.

Исследование основывается на предположении, что герминативные мутации в генах системы репарации неспаренных оснований (MMR), вызывающие синдром Линча, также могут быть связаны с раком молочной железы (РМЖ). В рамках исследования было проведено целевое секвенирование нового поколения генов MMR (MLH1, MSH2, MSH6, EPCAM и PMS2) у 711 пациентов с наследственным РМЖ, 60 пациентов со спорадическим РМЖ и 492 здоровых доноров.

Результаты исследования показали, что 69 пациентов (9.7%) с наследственным РМЖ имели хотя бы одну герминативную мутацию в генах MMR, из них 32 пациента (4.5%) имели мутации, определенные как патогенные или вероятно патогенные, а 26 пациентов (3.6%) не имели патогенных мутаций в генах системы обнаружения и репарации повреждений ДНК (DDR). В группе здоровых доноров были обнаружены только две мутации в генах MMR (0.4%).

Авторы приходят к выводу, что мутации, связанные с синдромом Линча, чаще встречаются у пациентов с наследственным РМЖ по сравнению со здоровыми донорами. В частности, была выявлена связь между наследственным РМЖ и мутациями c.1321G>A в MLH1, c.260C>G и c.2178G>C в MSH2, c.3217C>T в MSH6, c.1268C>G и c.86G>C в PMS2. Эти данные указывают на необходимость включения патогенных мутаций, связанных с синдромом Линча, в генетические тесты для пациентов с наследственным РМЖ.

Исследование подчеркивает важность учета этнических особенностей при оценке генетического риска и выявляет необходимость более глубокого изучения популяционно-специфических аспектов наследственного РМЖ. Авторы также обсуждают возможность пересмотра классификации некоторых мутаций в соответствии с рекомендациями Американского колледжа медицинской генетики и геномики и Ассоциации молекулярной патологии.

Таким образом, данное исследование предоставляет новые данные о биологии синдрома Линча и рака молочной железы и подчеркивает важность включения мутаций, связанных с синдромом Линча, в генетические тесты для оценки риска наследственного РМЖ.

4 Выбор проб

Для получения данных из Sequence Read Archive (SRA) я использовала веб-инструмент SRA-Explorer. Процесс скачивания проб включал следующие шаги:

1. Переход на сайт SRA-Explorer: https://sra-explorer.info/.
2. Поиск данных по идентификатору стати, в данном случаи `SRP229462`.
3. Выбор проб 10 проб случайным образом (отмечаем галочками рядом с выбранной пробой).
4. Генеруем скрипт на языке Bash с ссылоками для скачивания: Сгенерированный скрипт использует команду `curl` для загрузки каждого файла из SRA и сохранения его с соответствующим именем:

```
#!/usr/bin/env bash
curl -L ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/094/SRR10427694/
SRR10427694_1.fastq.gz -o SRR10427694_MMR_genes_in_breast_cancer_1.fastq.gz
...
```

5. Создаем файл с этим скриптом `load_samples.sh` и даем права на выполнение: `chmod +x load_samples.sh`, а затем запускаем: `./load_samples.sh`.

5 Анализ проб

5.1 Контроль качества

Для начала были предприняты следующие действия:

- Изменение прав доступа к файлам с расширением `.fastq.gz`, делая их исполняемыми (`+x`): `chmod +x *.fastq.gz`.
- Запуск программы FastQC для анализа качества проб: `fastqc *.fastq.gz`.

Затем были рассмотрены следующие показатели:

1. Основные статистические показатели (Basic Statistics)
2. Качество последовательности на каждой позиции (Per base sequence quality)
3. Оценка качества последовательности (Per sequence quality scores)
4. Содержание последовательности на каждой позиции (Per base sequence content)
5. Содержание GC в последовательности (Per sequence GC content)
6. Содержание N на каждой позиции (Per base N content)
7. Распределение длин последовательностей (Sequence Length Distribution)
8. Уровни дупликации последовательностей (Sequence Duplication Levels)
9. Перепредставленные последовательности (Overrepresented sequences)
10. Содержание адаптеров (Adapter Content)

В результате оценки качества были выявлены несоответствия в следующих направлениях:

- **Качество последовательности на каждой позиции** - это показатель того, как распределено качество прочтения каждого нуклеотида по всем ридам.

  Синяя линия – средние значения. Красная полоса – медиана. Жёлтый ящик – 25-й и 75-й перцентили. Усы – 10-й и 90-й перцентили.

  Оцениваются Phred-значения 25-го квартиля и медианы по всем позициям.

  Warning (предупреждение — данные могли быть и лучше, требует внимания): 25-й квартиль меньше 10, но больше 5 и медиана меньше 25, но больше 20.

  Failure (провал, дальнейший анализ не рекомендуется проводить без устранения проблем): 25-й квартиль меньше 5 и медиана меньше 20.

- **Содержание последовательности на каждой позиции** - это показатель доли каждого нуклеотида (A, T, G, C) в каждой позиции по всем ридам. Превалирование одних оснований над другими говорит о том, что, возможно, в библиотеке есть большое количество копий одного и того же участка. Однако, это может быть обусловлено специфичной фрагментацией.

  Warning: разница между долями A и T или G и C составляет от 10 до 20% в любой позиции.

  Failure: разница между долями A и T или G и C превышает 20% в любой позиции.

- **Содержание GC в последовательности** - это сравнение распределения ридов по их GC-составу с нормой.

  Warning: доля отклонившихся от нормального распределения ридов лежит в пределах от 15 до 30%.

  Failure: доля отклонившихся от нормального распределения ридов превышает 30%.

- **Распределение длин последовательностей** - это распределение ридов по длине.

  Warning: не все риды одинаковой длины.

  Failure: есть риды с длиной 0.

- **Перепредставленные последовательности** - это показатель, представленный в виде таблицы с (суб)последовательностями (от 20 п.о.), которые встречаются в более чем 0,1% ридов.

  Warning: в таблице есть хоть один hit, при этом все они встречаются в менее 1% ридов.

  Failure: в таблице есть последовательности, встречающиеся в более 1% ридов.

Таблица с количеством проб, на которых были предупреждения и провалы по вышеописанным показателям:

Показатель	Warning	Failure
Качество последовательности на каждой позиции	1	6
Содержание последовательности на каждой позиции	11	9
Содержание GC в последовательности	14	2
Распределение длин последовательностей	20	0
Перепредставленные последовательности	3	0

Попробуем улучшить качество, используя команду `fastp`. Для этого:
Создадим папку `data` и поместим туда исходные файлы.
Создадим папку `trimmed` для новых файлов.
Сгенерируем их при помощи fastp, указав, что нам нужно:

- Взять исходные файлы (`-i` и `-I`);
- Попытаться устранить адаптеры в парноконцевых ридах, если адаптеры есть (`-detect_adapter_for_pe`);
- Проанализировать часто встречающиеся последовательности (`-overrepresentation_analysi`
- Скорректировать нуклеотиды на пересечении парных ридов (`-correction`);
- Обрезать низкокачественные участки ридов "справа"(где они, как правило, плохие) (`-cut_right`).

Также укажем пути к отчётам (`-html` и `-json`), файлам, которые получатся на выходе (`-o` и `-O`).

Получившийся скрипт:

```
#!/usr/bin/env bash
mkdir data
mkdir trimmed

mv *.fastq.gz data/

for i in SRR10427304 SRR10427694 SRR10427700 SRR10427706 SRR10427716
SRR10427720 SRR10427722 SRR10427725 SRR10427728 SRR10427732; do
    fastp --detect_adapter_for_pe \
      --overrepresentation_analysis \
      --correction --cut_right --thread 2 \
      --html trimmed/${i}.fastp.html \
      --json trimmed/${i}.fastp.json \
      -i data/${i}_MMR_genes_in_breast_cancer_1.fastq.gz \
      -I data/${i}_MMR_genes_in_breast_cancer_2.fastq.gz \
      -o trimmed/${i}_1.fastq.gz -O trimmed/${i}_2.fastq.gz
done
```

Далее создадим папку `trimmed/result/` и повторно оценим качество с помощью команды `fastqc *.fastq.gz`.

Результаты:

Показатель	**Warning**	**Failure**
Качество последовательности на каждой позиции	0	0
Содержание последовательности на каждой позиции	14	5
Содержание GC в последовательности	14	0
Распределение длин последовательностей	20	0
Перепредставленные последовательности	0	0

Получилось значительно повысить качество последовательности на каждой позиции и убрать все перепредставленные последовательности. Также в некоторых случаях удалось уменьшить разницу между долями нуклеотидов.

5.2 Выравнивание ридов

В описанной выше статье выдвигалось предположение, что герминативные мутации в генах системы репарации неспаренных оснований (MMR), вызывающие синдром Линча, также могут быть связаны с раком молочной железы (РМЖ). В рамках исследования было проведено целевое секвенирование нового поколения генов MMR (MLH1, MSH2, MSH6, EPCAM и PMS2) у 711 пациентов с наследственным РМЖ, 60 пациентов со спорадическим РМЖ и 492 здоровых доноров. Мы поступим аналогично и рассмотрим референсную последовательность 2-й хромосомы человека, содержащую гены MSH2 и MSH6. Возможно, мы увидим мутации в этих генах, что может говорить о возникновении синдрома Линча.

Идентификатор 2-й хромосомы — NC_000002.12. Загрузим её последовательность в формате FASTA с помощью EDirect: `efetch -db nuccore -id NC_000002.12 -format fasta > chr2.fasta`.

Далее, подготовим референсную последовательность:

1. Сгенерируем сопутствующие вспомогательные файлы при помощи программы BWA: `bwa index chr2.fasta`.

2. Создадим индекс, используя Samtools: `samtools faidx chr2.fasta`.

3. Воспользуемся GATK4 и создадим словарь: `gatk CreateSequenceDictionary -R chr2.fasta`.

Наконец, получим SAM-файлы выравниваний, переведём их в более компактный формат BAM, а затем воспользуемся GATK для сортировки и Samtools для индексации получившихся BAM-файлов:

```
#!/usr/bin/env bash

# Создание папок для хранения промежуточных и итоговых файлов
mkdir -p aligned
mkdir -p sorted

# Выравнивание ридов с использованием BWA и преобразование SAM в BAM
for i in SRR10427304 SRR10427694 SRR10427700 SRR10427706 SRR10427716 \
SRR10427720 SRR10427722 SRR10427725 SRR10427728 SRR10427732; do
    bwa mem chr2.fasta \
        trimmed/${i}_1.fastq.gz \
        trimmed/${i}_2.fastq.gz \
        > aligned/${i}.sam
    samtools view -bS aligned/${i}.sam > aligned/${i}.bam
done

# Сортировка BAM-файлов с использованием GATK SortSam и индексация
# с помощью Samtools
for i in SRR10427304 SRR10427694 SRR10427700 SRR10427706 SRR10427716 \
SRR10427720 SRR10427722 SRR10427725 SRR10427728 SRR10427732; do
    gatk SortSam -I aligned/${i}.bam -O sorted/${i}_sorted.bam -SO coordinate
    samtools index sorted/${i}_sorted.bam
done
```

Риды успешно выровнялись на 2 хромосому.

5.3 Поиск вариантов с использованием VarScan

Для поиска вариантов в генах MMR была использована утилита VarScan. Сначала были получены pileup-файлы для каждого образца, а затем VarScan был запущен для поиска вариантов.

Для каждого образца был создан pileup-файл с использованием инструмента Samtools mpileup. Этот файл содержит информацию о выравнивании ридов на референсную последовательность.

```
#!/usr/bin/env bash

# Получение pileup-файлов с использованием Samtools mpileup
for i in SRR10427304 SRR10427694 SRR10427700 SRR10427706 SRR10427716 \
SRR10427720 SRR10427722 SRR10427725 SRR10427728 SRR10427732; do
    samtools mpileup -B -f chr2.fasta \
        sorted/${i}_sorted.bam > ${i}.mpileup
done
```

VarScan был запущен для каждого pileup-файла.

```
#!/usr/bin/env bash

# Запуск VarScan для поиска вариантов
for i in SRR10427304 SRR10427694 SRR10427700 SRR10427706 SRR10427716 \
SRR10427720 SRR10427722 SRR10427725 SRR10427728 SRR10427732; do
    varscan mpileup2snp ${i}.mpileup > ${i}.vcf
done
```

Этот скрипт проходит через каждый образец, использует VarScan для поиска вариантов в соответствующем pileup-файле и создает VCF-файлы с обнаруженными вариантами.

5.4 Описание вариантов в VEP

После получения файлов формата VCF, содержащих варианты, мы воспользовались инструментом VEP (Variant Effect Predictor) на веб-сервере Ensembl для аннотации и оценки воздействия этих вариантов на гены.

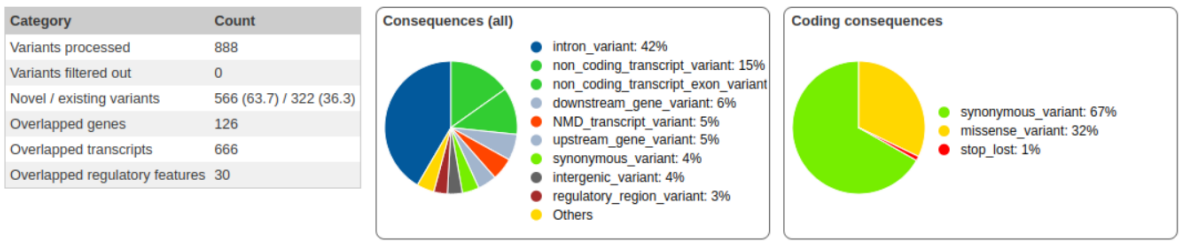

Рис. 1: Пример результатов аннотации вариантов с использованием VEP (проба 1)

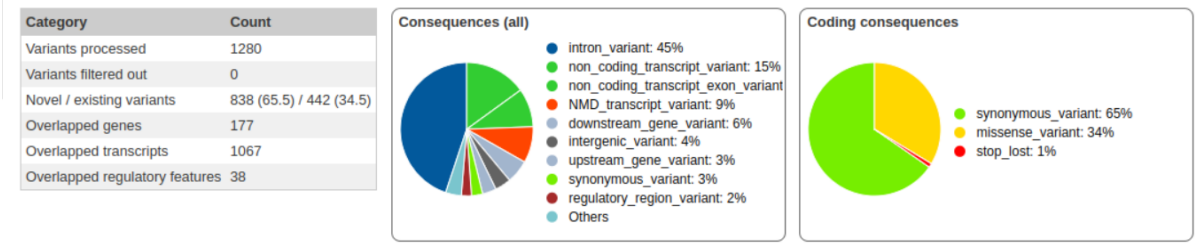

Рис. 2: Пример результатов аннотации вариантов с использованием VEP (проба 2)

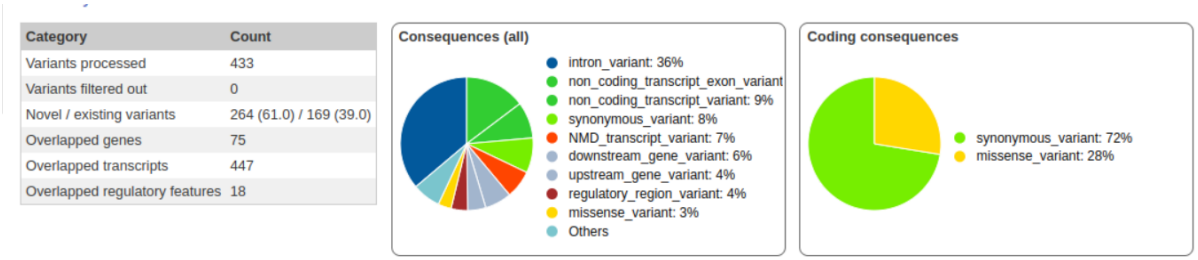

Рис. 3: Пример результатов аннотации вариантов с использованием VEP (проба 3)

На рисунках 1,2 и 3 показаны примеры части результатов аннотации вариантов, полученных с использованием VEP.

6 Анализ полученных вариантов

Гены MSH2 и MSH6 находятся на 2-й хромосоме человека. Их координаты, основанные на данных Ensembl, представлены ниже:

Ген MSH2

- Начальная позиция: примерно 47,477,249
- Конечная позиция: примерно 47,537,546

Ген MSH6

- Начальная позиция: примерно 47,493,546
- Конечная позиция: примерно 47,523,856

В этих границах были рассмотрены результаты анализа и обнаружены мутации. Например, была обнаружена мутация в интроне гена MSH6 на хромосоме 2 человека. Эта мутация характеризуется заменой исходного нуклеотида T на альтернативный нуклеотид G на позиции 47795842. Интронные мутации могут влиять на сплайсинг и регуляцию экспрессии гена, что может иметь последствия для функционирования белка и общего состояния клетки.

7 Заключение

В ходе данной работы был выполнен анализ данных секвенирования ДНК с целью выявления патогенных вариантов, связанных с развитием рака молочной железы у пациентов с наследственными мутациями синдрома Линча.

Сначала были выбраны и загружены десять случайных образцов ДНК из доступных данных, после чего проведен контроль качества и предобработка данных.

Затем был выполнен анализ на наличие вариантов с использованием VarScan, что позволило выявить потенциально патогенные мутации в генах системы репарации неспаренных оснований (MMR), связанные с синдромом Линча.

Наконец, полученные варианты были аннотированы и оценены с помощью Variant Effect Predictor (VEP) для определения их воздействия на гены.

Этот анализ дал дополнительное понимание связи между наследственными мутациями в генах MMR и развитием рака молочной железы, подтверждая необходимость включения таких мутаций в генетические тесты для оценки риска наследственного рака молочной железы. Полученные результаты могут иметь важное значение для развития индивидуализированных стратегий профилактики и лечения данного заболевания.