

Домашнее задание по лекции №9

Тема: «Анализ и интерпретация NGS данных (Dna-seq) в диагностике наследственных заболеваний»

Описание: домашнее задание выполняла на сервере лаборатории МГНЦ (eod-wgs), поэтому я буду прикреплять скриншоты из команд терминала с промежуточным результатом в виде картинок. Открытые варианты можно посмотреть в [Google Colab](https://colab.research.google.com/).

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1. Опишите основные этапы проведения секвенирования в "мокрой" части лаборатории.

Одни из основных этапов проведения секвенирования в мокрой части включают: **выделение ДНК (РНК), подготовка библиотеки и секвенирование** (запуск на платформе).

Фрагментация ДНК может быть воспроизведена различными способами, но глобально подразделяют на *ферментативную* (использование каких-то эндонуклеаз рестрикции) или *ультразвуковую* (соникация, к примеру). После вносятся **адаптерные последовательности**, содержащие чаще всего **баркоды** (для того, чтобы отличать образцы друг от друга) и после подготовленные библиотеки измеряются по концентрации и отправляются на секвенирование.

Кажется очень абстрактно, но я могу описать подготовку библиотек **полногеномного бисульфитного секвенирования (WGBS)**, которыми я занималась:

Выделение геномной ДНК: фенол-хлороформная экстракция из любой доступной ткани, например периферическая кровь. На выходе измерить концентрацию ДНК (нг/мкл) и чистоту (соотношение длин волн A260/280 и A260/230) ¹

Растворить необходимое количество ДНК в MQ, чтобы получился требуемый под протокол одинаковый объем ДНК для библиотек на вход
--

Гидролиз XmaI: один из примеров ферментативной фрагментации ДНК. Получаем на выходе пул фрагментов ДНК

Затупление концов частичное фрагментом Klenow exo- , где добавление
--

¹ Это важно!! Потому что даже на самом первом этапе гидролиза фермент может не до конца порезать геномную ДНК, и дальнейшие этапы с дорогими реагентами просто в помойку можно (было...)

5-methyl-dСТР нужно для частичного заполнения концов ДНК-фрагментов, полученных после гидролиза XmaI. Это предотвращает самолигирование ДНК-фрагментов и адаптеров и создает липкие **не палиндромные концы**, которые **предотвращают самозамыкание фрагментов**

Лигирование фрагментов с адаптерами: аннелированный синтетический адаптер с ДНК-фрагментами

Ник-трансляция. создание одноцепочечных разрывов DNase I, удаление нуклеотидов с 5'-разрыва и добавляются новые к 3' разрыву, используя Cme. То есть олигонуклеотиды адаптеров заменялись на последовательности с метилированными цитозинами для предотвращения деаминации цитозинов в последующем преобразовании бисульфитом.²

Очистка для инактивации полимеразы протоколом Agencourt AMPure XP

Селекция по длине (Pippin Prep), “дорожка в кассете”. Здесь уже оценивается качество полученной библиотеки и отбираются оптимальной длины фрагменты. Например, Ion Torrent требует до 200 п.н. максимум длину, но и короткие (до 100 п.н.) тоже плохо.

Промывка от бромистого этидия AMPureXP

Бисульфитная конверсия "Отмывка от бисульфита с carrier RNA"
 Отработка циклов BS qPCR
Амплификация библиотеки, циклов
 Финальная отмывка и измерение **конечной концентрации библиотеки**
HS (нг/мл)

2. Какие платформы для секвенирования вы знаете? Кратко опишите не менее трех платформ, указав их основные характеристики и особенности.

Технология	Чтения	Длина чтений	Скорость	Точность	Описание метода
Sanger	по одному чтению	400-900 bp	несколько часов на чтение	>99.99%	метод, основанный на дидактических цепях, с использованием флуоресцентных меченых нуклеотидов для

² Очень дорого заказывать адаптеры сразу с метилированным цитозином, поэтому дополнительный этап

					определения последовательности ДНК
Illumina/Solexa	массово параллельное	50-300 bp	несколько дней на геном	98-99.9%	использует флуоресцентные метки для определения нуклеотидов во время синтеза, чтения фиксируются камерой
Roche 454	массово параллельное	400-600 bp	несколько дней на геном	99.9%	Пиросеквенирование: определение последовательности на основе высвобождения пирофосфата при инкорпорации нуклеотидов
SOLiD	массово параллельное	50-75 bp	несколько дней на геном	99.94%	использует лиганды для чтения двух оснований одновременно, улучшая точность
PacBio SMRT	одномолекулярное	10,000-15,000 bp	несколько часов на геном	87-99.99 %	использует одномолекулярное времяпролетное секвенирование (SMRT) для прямого чтения длинных цепей ДНК
Oxford Nanopore MinION	одномолекулярное	До 1,000,000 bp	несколько часов на геном	90-98%	использует поры белка для чтения нуклеотидов по мере прохождения через мембрану.
Illumina NovaSeq	массово параллельное	100-300 bp	Несколько дней на геном	99.9%	усиленная версия Illumina/Solexa с улучшенной пропускной способностью и скоростью секвенирования

Примечание: табличку составляла сама, поэтому ссылку не привела

3. Опишите стандартный биоинформатический пайплайн для анализа данных NGS (DNA-seq). Кратко укажите цель каждого из них.

- 1. Контроль качества исходных данных (FastQ).** Задача оценить качество сырых данных (FastQ файлы), включая качество чтений, уровень ошибок, наличие адаптеров и других артефактов.
Инструменты: FastQC, MultiQC
- 2. Удаление адаптеров и обрезка низкокачественных областей.** Удалить адаптеры и обрезать низкокачественные концы чтений, чтобы повысить точность последующих этапов анализа.

Инструменты: Cutadapt, Fastp

- 3. Выравнивание (Alignment).** Выравнивание чтений на референсный геном для идентификации местоположений, где были сделаны изменения (варианты)

Инструменты: BWA, Bowtie2, STAR (для RNA-seq), Minimap2

- 4. Обработка выровненных данных.** Например, удаление дублирующих чтений и улучшение качества выравнивания.

Инструменты: MarkDuplicates, Samtools

- 5. Вызов вариантов (Variant Calling).** Идентификация однонуклеотидных полиморфизмов (SNPs) и инделов из выровненных данных.

Инструменты: GATK HaplotypeCaller, Samtools mpileup, FreeBayes

- 6. Аннотация вариантов.** Обогащение информации о вариантах, добавление функциональной информации (например, какие гены или экзоны затронуты) и сопоставление с известными базами данных, такими как ClinVar, dbSNP.

Инструменты: SnpEff, ANNOVAR

- 7. Фильтрация вариантов.** Отбор надежных и потенциально значимых вариантов, исключая низкое качество, низкую глубину покрытия

Инструменты: GATK VariantFiltration, bcftools, vcfutils.

- 8. Качественная оценка и визуализация данных.** Оценка качества вызванных вариантов и визуализация для дальнейшей интерпретации.

Инструменты: IGV (Integrative Genomics Viewer), UCSC Genome Browser, bedtools.

- 9. Статистический и функциональный анализ.** *Инструменты: PLINK, SNPTEST, GWAS tools, R/Bioconductor*

- 4. Кратко опишите основные этапы пайплайна GATK для анализа наследуемых коротких инделов (indels) и однонуклеотидных полиморфизмов (SNPs)**

GATK (Genome Analysis Toolkit) используется для анализа наследуемых коротких инделов (indels) и однонуклеотидных полиморфизмов (SNPs) и включает в себя следующие шаги:

1. **Качество и фильтрация исходных данных (FastQ):** выполняется предварительная фильтрация сырых данных (например, с FastQC или тримминг адаптеров). Если много образцов еще multiQC.
2. **Выравнивание с использованием BWA.** FastQ файлы выравниваются на референсный геном с помощью программы BWA или Bowtie2. Результат: BAM файл с выровненными прочтениями
3. **Обработка выравнивания:** MarkDuplicates удаляются дублирующие чтения, IndelRealigner, чтобы исправить возможные ошибки выравнивания, и можно выполнить перекалибровку качества баз с использованием Base Recalibrator.
4. **Вызов вариантов (SNPs и инделов)** с использованием HaplotypeCaller или DeepVar из выровненных BAM файлов после удаления дубликатов. Выход - vcf файл.
5. **Варианты аннотируются** с использованием инструментов, таких как SnpEff или ANNOVAR.
6. **Фильтрация вариантов по качеству** (Variant Filtration, например, качество, глубина покрытия, флаг PASS)
7. **Анализ и интерпретация данных.** Визуализация результатов с помощью таких инструментов, как IGV (Integrative Genomics Viewer)

Прежде чем приступить к работе с **VAM** файлом, мне требовалось посмотреть на сам файл: отображается ли основная информация и целый файл, так как предыдущие этапы подготовки файла (контроль качества, тримминг, выравнивание) были выполнены не мной. Это можно сделать при помощи команды `samtools view` с выводом первых 4 строк:

[illegible]

Рисунок 1. Содержимое ВАР-файла

QNAME: идентификатор прочтения (например, SRR22359471.36921505)
FLAG: флаги выравнивания, дают информацию о выравнивании и парности прочтений (99, 147, 83, 163 и тд)
RNAME: референсное имя хромосомы (например, chr1)
POS: позиция выравнивания на референсе
MAPQ: Качество выравнивания (например, 41). Более высокие значения указывают на лучшее качество выравнивания
CIGAR: Строка CIGAR, описывающая операции выравнивания, такие как совпадение (M), вставка (I), удаление (D), мягкое отсечение (S) и т.д.
RNEXT: Референсное имя для следующего прочтения в паре
PNEXT: Позиция следующего прочтения в паре
TLEN: Шаблоновая длина
SEQ: Последовательность нуклеотидов прочтения
QUAL: Качество нуклеотидов прочтения

После такой инфы можно оценить качество выравниваний, например по MAPQ и CIGAR. После маркирования дубликатов с помощью MarkDuplicates, можно анализировать метрики дубликатов, искать аномальные выравнивания, такие как инверсии, транслокации и тд.

ЧАСТЬ ПЕРВАЯ. Маркирование дубликатов в bam-файле с помощью MarkDuplicates из GATK

- 1) Прежде чем маркировать дубликаты, я установила GATK при помощи команды `wget`
`wget https://github.com/broadinstitute/gatk/releases/download/4.6.1.0/gatk-4.6.1.0.zip`
- 2) Команда распаковки архива: `unzip gatk-4.6.1.0.zip`
- 3) Для маркирования дубликатов я использовала команду с использованием функции **JavaScript**:

```
java -jar picard.jar MarkDuplicates \  
    I=sample_15.bam \  
    O=sample_15_marked.bam \  
    M=sample_15_metrics.txt \  
    CREATE_INDEX=true
```


- **unpaired_read_duplicates**: дубликаты среди одиночных прочтений: 24,251.
- **read_pair_duplicates**: дубликаты среди пар прочтений: 7,071,456. Это может быть следствием технических артефактов или особенностей подготовки библиотеки.
- **read_pair_optical_duplicates**: оптические дубликаты среди пар прочтений: 0. Отсутствие оптических дубликатов указывает на то, что проблемы с дублированием связаны не с техническими ошибками считывания, а, возможно, с этапом подготовки библиотеки или характеристиками исходного материала.
- **percent_duplication**: процент дублирования: 0.177995 или 17.8%. Процент дублирования 17.8%, значит почти пятая часть данных состоит из дубликатов.
- **estimated_library_size**: оценочный размер библиотеки: 98,144,242, вообще достаточный объем данных для проведения дальнейших анализов.

bin	coverage	all_sets	non_optical_sets
1.0	1	26702071	26702071
2.0	1.666827	5076620	5076620
3.0	2.111485	787868	787868
4.0	2.407995	114087	114087
5.0	2.605716	15835	15835
6.0	2.737562	2168	2168
7.0	2.82548	295	295
8.0	2.884106	47	47
9.0	2.9232	10	10
10.0	2.949268	3	3
11.0	2.966652	1	1
12.0	2.978243	2	2
13.0	2.985973	0	0
14.0	2.991127	0	0
15.0	2.994564	2	2
16.0	2.996856	0	0
17.0	2.998384	0	0
18.0	2.999403	0	0
19.0	3.000083	0	0
20.0	3.000536	0	0
21.0	3.000838	0	0
22.0	3.00104	0	0
23.0	3.001174	0	0
24.0	3.001264	0	0
25.0	3.001324	1	1
26.0	3.001363	0	0
27.0	3.00139	0	0
28.0	3.001408	1	1
29.0	3.00142	0	0
30.0	3.001427	1	1
31.0	3.001433	1	1
32.0	3.001436	0	0
33.0	3.001439	0	0
34.0	3.00144	0	0
35.0	3.001441	0	0
36.0	3.001442	1	1

**Гистограмма покрытий
(Coverage Histogram)**

ЧАСТЬ ВТОРАЯ. Коллинг вариантов с помощью **DeepVariant** в режиме **WES**

```

sudo apt -y update
sudo apt-get -y install docker.io
sudo docker pull google/deepvariant:"1.6.1"

```


Для проведения коллинга воспользовалась инструкцией, приведенной по [ссылке](#). Для этого требовалось предварительно установить [Docker](#)³.

Для его установки воспользовалась следующей инструкцией:

1) Установление зависимостей:

```
sudo apt update \
sudo apt install apt-transport-https ca-certificates curl
software-properties-common
```

2) Добавила GPG ключ для официального Docker репозитория:

```
curl -fsSL https://download.docker.com/linux/ubuntu/gpg |sudo apt-key add -
```

3) Docker репозиторий в АРТ источники:

```
sudo add-apt-repository "deb [arch=amd64]
https://download.docker.com/linux/ubuntu $(lsb_release -cs) stable"
```

4) Установка Docker: `sudo apt install docker-ce`

5) Добавила в своего пользователя в группу Docker, чтобы запускать команды Docker без sudo: `sudo usermod -aG docker aimanalieva`

Теперь, когда Docker установлен, я провела **DeepVariant** для анализа экзомки (WES). Для этого скрипт описан ниже:

```
mkdir -p output
mkdir -p output/intermediate_results_dir
BIN_VERSION="1.6.1"
sudo docker run \
-v "${PWD}/input":"/media/HEAP-EPI/Diplome_gi_Imanalieyva_Amina/NGS/input" \ \
-v "${PWD}/output":"/media/HEAP-EPI/Diplome_gi_Imanalieyva_Amina/NGS/deepvariant/output" \
-v "${PWD}/reference":"/media/HEAP-EPI/Diplome_gi_Imanalieyva_Amina/NGS/input/reference" \
google/deepvariant:"${BIN_VERSION}" \
/opt/deepvariant/bin/run_deepvariant \
--model_type WES \
--ref /reference/GRCh38_no_alt_analysis_set.fasta \
--reads /input/sample_15_marked.bam \
--regions /input/idx_capture_novogene.grch38.bed \
--output_vcf /output/sample_15_marked.vcf.gz \
--output_gvcf /output/sample_15_marked.g.vcf.gz \
--num_shards 4
```

Примечание: выражаю большую благодарность моему коллеге Тимуру Кулагину за помощь и консультацию, так как на сервере лабы не получалось настроить Tensor Flow для запуска Docker.

³ Docker — платформа для контейнеризации, позволяющая запускать приложения в изолированных, переносимых средах.

ЧАСТЬ ТРЕТЬЯ. Фильтрация VCF файла.

Флаг 'PASS' означает, что данный вариант успешно прошел все фильтры и считается надежным для дальнейшего анализа. Например, это помогает отобрать варианты, прошедшие дополнительные или специфические фильтрации (например, генные варианты с низким качеством или низким покрытием).

Input file: sample_15_marked

Output file: output.vcf

Команда для фильтрации:

```
awk -F '\t' '{if($0 ~ /#/ ) print; else if($7 == "PASS") print}'  
output/sample_15_marked.vcf > output.vcf
```

```
(base) etcetera_mi@AMINA-161002:~$ tail output.vcf  
chrX    154652556    .    C    A    69    PASS    .    GT:G  
Q:DP:AD:VAF:PL  1/1:67:204:1,203:0.995098:69,72,0  
chrX    154653251    .    C    G    69    PASS    .    GT:G  
Q:DP:AD:VAF:PL  1/1:65:89:0,89:1:68,67,0  
chrX    154653499    .    C    T    32.8    PASS    .    GT:G  
Q:DP:AD:VAF:PL  1/1:16:6:0,6:1:32,15,0  
chrX    154680220    .    C    G    56    PASS    .    GT:G  
Q:DP:AD:VAF:PL  0/1:46:154:65,87:0.564935:56,0,46  
chrX    154766321    .    G    T    70.2    PASS    .    GT:G  
Q:DP:AD:VAF:PL  1/1:69:252:0,252:1:70,76,0  
chrX    154791839    .    C    A    50    PASS    .    GT:G  
Q:DP:AD:VAF:PL  0/1:50:179:102,77:0.430168:50,0,63  
chrX    154929926    .    T    G    50.1    PASS    .    GT:G  
Q:DP:AD:VAF:PL  0/1:49:76:43,33:0.434211:50,0,57  
chrY    14840455    .    T    C    28.2    PASS    .    GT:G  
Q:DP:AD:VAF:PL  1/1:24:9:0,9:1:28,26,0  
chrY    14840467    .    T    C    24.8    PASS    .    GT:G  
Q:DP:AD:VAF:PL  1/1:23:10:0,10:1:24,28,0  
chrY    14840785    .    C    T    22.9    PASS    .    GT:G  
Q:DP:AD:VAF:PL  1/1:16:29:0,29:1:22,16,0
```

Рисунок 4. Отфильтрованный файл по 'PASS'

ЧАСТЬ ЧЕТВЕРТАЯ. Аннотация с помощью SnpEff.

Аннотация с помощью SnpEff помогает интерпретировать результаты, полученные из данных о генетических вариантах. Он предоставляет тип варианта (например, замена нуклеотида, вставка, делеция), frame-shift, сдвиг стоп-кодона или какие-то миссенс-варианты. Часто использует аннотирование вариантов с использованием с использованием базы данных по типу **dbSNP**, **ClinVar**, **Ensembl**. Он также может давать информацию о возможных эффектах вариантов на регуляцию генов. Это может включать: изменения в промоторах, энхансерах или других регуляторных элементах, влияние на сплайсинг РНК.

Скрип приведен ниже:

4.1 Добавить аннотацию SnpEff

1) Скачать SnpEff

```
wget https://snpeff.blob.core.windows.net/versions/snpEff_latest_core.zip  
unzip snpEff_latest_core.zip
```

2) Скачать базу данных hg38 для SnpEff

```
java -jar snpEff/snpEff.jar download -v hg38
```

3) Аннотация файла с флагом -canon на hg38

```
java -jar snpEff/snpEff.jar -canon hg38 output.vcf > annotated_output.vcf
```

```
(base) etcetera_mi@AMINA-161002:~$ tail output.vcf  
chrX      154652556      .      C      A      69      PASS      .      GT:G  
Q:DP:AD:VAF:PL  1/1:67:204:1,203:0.995098:69,72,0  
chrX      154653251      .      C      G      69      PASS      .      GT:G  
Q:DP:AD:VAF:PL  1/1:65:89:0,89:1:68,67,0  
chrX      154653499      .      C      T      32.8      PASS      .      GT:G  
Q:DP:AD:VAF:PL  1/1:16:6:0,6:1:32,15,0  
chrX      154680220      .      C      G      56      PASS      .      GT:G  
Q:DP:AD:VAF:PL  0/1:46:154:65,87:0.564935:56,0,46  
chrX      154766321      .      G      T      70.2      PASS      .      GT:G  
Q:DP:AD:VAF:PL  1/1:69:252:0,252:1:70,76,0  
chrX      154791839      .      C      A      50      PASS      .      GT:G  
Q:DP:AD:VAF:PL  0/1:50:179:102,77:0.430168:50,0,63  
chrX      154929926      .      T      G      50.1      PASS      .      GT:G  
Q:DP:AD:VAF:PL  0/1:49:76:43,33:0.434211:50,0,57  
chrY      14840455      .      T      C      28.2      PASS      .      GT:G  
Q:DP:AD:VAF:PL  1/1:24:9:0,9:1:28,26,0  
chrY      14840467      .      T      C      24.8      PASS      .      GT:G  
Q:DP:AD:VAF:PL  1/1:23:10:0,10:1:24,28,0  
chrY      14840785      .      C      T      22.9      PASS      .      GT:G  
Q:DP:AD:VAF:PL  1/1:16:29:0,29:1:22,16,0  
(base) etcetera_mi@AMINA-161002:~$ tail annotated_output.vcf | head -n 3  
chrX      154652556      .      C      A      69.0      PASS      ANN=A|synony  
mous_variant|LOW|CTAG2|CTAG2|transcript|NM_020994.5|protein_coding|2/2|c.345  
G>T|p.Pro115Pro|409/988|345/633|115/210||      GT:GQ:DP:AD:VAF:PL      1/1:  
67:204:1,203:0.995098:69,72,0  
chrX      154653251      .      C      G      69.0      PASS      ANN=G|missen  
se_variant|MODERATE|CTAG2|CTAG2|transcript|NM_020994.5|protein_coding|1/2|c.  
265G>C|p.Glu89Gln|329/988|265/633|89/210||      GT:GQ:DP:AD:VAF:PL      1/1:  
65:89:0,89:1:68,67,0  
chrX      154653499      .      C      T      32.8      PASS      ANN=T|missen  
se_variant|MODERATE|CTAG2|CTAG2|transcript|NM_020994.5|protein_coding|1/2|c.  
17G>A|p.Arg6Gln|81/988|17/633|6/210||      GT:GQ:DP:AD:VAF:PL      1/1:16:6:0,6  
:1:32,15,0
```

Вот для сравнения я привела на картинке, как отличается вывод информации до добавления аннотации (output.vcf) и после добавления аннотации (annotated_output.vcf).

4.2 Добавить информацию из ClinVar

Скачать базу данных ClinVar с индексированным файлом:

wget https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/clinvar.vcf.gz

wget https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/clinvar.vcf.gz.tbi

Скрипт для аннотации на клинвар:

```
java -Xmx1g -jar /snpEff/SnpSift.jar annotate -v clinvar.vcf  
annotated_output.vcf > clinvar_annotation.vcf
```

ЧАСТЬ ПЯТАЯ. Фильтрация патогенных и вероятно патогенных вариантов

Отфильтруйте Pathogenic и Likely pathogenic варианты.

- **Патогенные варианты:** варианты, которые могут быть связаны с заболеваниями
- **Рискованные варианты:** варианты, которые могут увеличивать вероятность возникновения заболевания, но еще не полностью подтверждены

После аннотации на патогенные и вероятно-патогенные варианты полученного **vcf.файла** получила **17 вариантов**:

```
[100] !grep -i -e 'Likely_pathogenic' -e 'ClinSig-Pathogenic' -e 'ClinSig-Pathogenic' clinvar_annotation.vcf > result.vcf  
[100] !cat result.vcf  
chr1 976215 1320832 A G 69.8 PASS ANH-G[misense_variant|MODERATE|PERM1|PERM1|transcript[NM_001291366.2|protein_coding[4/4|c.2330T>C|p.Val777Ala|2377/3394|2330/2373|777/790|],6|downstream_gene_variant|MODIFIER|PL  
chr1 226735804 1705897 G T 66.4 PASS ANH-T[misense_variant|MODERATE|ITPKB|ITPKB|transcript[NM_002221.3|protein_coding[2/8|c.1655C>A|p.Pro526Gln|1995/6162|1655/2841|552/946|],ALLELEID=1704217|CLNDISDB=MONDO:  
chr1 226735237 1705899 A C 68.7 PASS ANH-C[misense_variant|MODERATE|ITPKB|ITPKB|transcript[NM_002221.3|protein_coding[2/8|c.1222T>A|p.Ser488Ala|1152/6162|1222/2841|488/946|],ALLELEID=1704219|CLNDISDB=MONDO:  
chr1 226737174 1705896 ACTGCCGCTG A 68.1 PASS ANH-A[disruptive_inframe_deletion|MODERATE|ITPKB|ITPKB|transcript[NM_002221.3|protein_coding[2/8|c.276_284delCAGCGGCAG|p.Ser93_Ser95del|624/6162|276/284|92/946|]  
chr2 127426980 1394296 T G 61.4 PASS ANH-G[misense_variant|MODERATE|PROC|PROC|transcript[NM_000312.3|protein_coding[7/9|c.541T>G|p.Phe181Val|635/1780|541/1386|181/461|],6|upstream_gene_variant|MODIFIER|HNR4  
chr6 24848382 128343 C T 78.2 PASS ANH-T[misense_variant|MODERATE|ALDEA1|ALDEA1|transcript[NM_170740.1|protein_coding[9/11|c.538C>T|p.His180Tyr|566/5170|538/1847|180/548|],AF_ESP=0.37121|AF_EXAC=0.3155  
chr6 32836310 12181 A T 42.4 PASS ANH-T[misense_variant|MODERATE|CYP21A2|CYP21A2|transcript[NM_000540.9.3|protein_coding[1/8|c.188A>T|p.His53Tol|156/2080|188/1483|63/495|],1|downstream_gene_variant|PC  
chr8 186679222 1678542 A G 63.5 PASS ANH-G[misense_variant|MODERATE|QOR1|QOR1|transcript[NM_001198532.1|protein_coding[3/16|c.236A>G|p.Lys79Arg|335/4476|236/2625|79/874|],ALLELEID=1678259|CLNDISDB=MedGen:C  
chr12 102852922 92748 C T 69.6 PASS ANH-T[synonymous_variant|LOW|PAH|PAH|transcript[NM_000277.3|protein_coding[7/13|c.735G>A|p.Val245Val|849/3759|735/1359|245/452|],AF_ESP=0.18153|AF_EXAC=0.29058|AF_TGP=0.1  
chr14 81091070 314093 G C 59.6 PASS ANH-C[misense_variant|splice_region_variant|MODERATE|TSR1|TSR1|transcript[NM_000369.3|protein_coding[5/10|c.294G>C|p.Gly112>G|454/4300|394/2259|112/764|],AF_EXAC=0.808  
chr16 69711242 16889 G A 67.0 PASS ANH-A[misense_variant|MODERATE|NQO1|NQO1|transcript[NM_000893.3|protein_coding[6/6|c.559C>T|p.Pro187Ser|680/2521|559/825|187/274|],A|upstream_gene_variant|MODIFIER|SMORC  
chr17 80385145 39780 G A 64.3 PASS ANH-A[misense_variant|MODERATE|BNF213|BNF213|transcript[NM_001256071.3|protein_coding[60/68|c.14429G>A|p.Arg481Lys|14588/21874|14429/15624|481/5287|],A|intron_variant|MOD  
chr18 23539813 21136 G A 64.6 PASS ANH-A[splice_region_variant|synonymous_variant|LOW|NPC1|NPC1|transcript[NM_000271.5|protein_coding[18/25|c.2793C>T|p.Asp831Asn|2956/4768|2793/3837|831/1278|],AF_ESP=0.442  
chr18 57571588 562 A G 78.7 PASS ANH-G[intron_variant|MODIFIER|FECH|FECH|transcript[NM_001012515.3|protein_coding[3/10|c.333_487C>T|],AF_ESP=0.83737|AF_EXAC=0.10730|AF_TGP=0.13730|ALLELEID=15601|CLNDI  
chr19 1615797 1165836 G A 56.8 PASS ANH-A[misense_variant|MODERATE|TCF3|TCF3|transcript[NM_003200.5|protein_coding[17/19|c.1475C>T|p.Ala492Val|1830/4735|1475/1965|492/654|],AF_ESP=0.01346|AF_EXAC=0.05265|AF_TGP=0  
chr19 12896366 575670 T C 74.4 PASS ANH-C[misense_variant|MODERATE|GDIH|GDIH|transcript[NM_000159.4|protein_coding[8/12|c.797T>C|p.Met266Ile|874/1852|797/1317|266/438|],C|downstream_gene_variant|MODIFIER|S  
chr19 44980684 17864 T C 58.1 PASS ANH-C[misense_variant|MODERATE|APCE|APCE|transcript[NM_00130288.2|protein_coding[4/4|c.466T>C|p.Cys150Arg|539/1240|466/1032|156/343|],C|downstream_gene_variant|MODIFIER|S
```

Патогенных вариантов получилось всего 9, при этом, очень интересный вариант с делецией (всего один)

ЧАСТЬ ШЕСТАЯ. Генетические варианты, ассоциированные с фенотипом

Из проведенного анализа и полученных 17 вариантах я решила остановиться на некоторых патогенных вариантах (из девяти я выбрала после анализа 6, другие не оч были показаны в ИБС) и одним потенциально-патогенным, делеция которого приводит к изменению рамки считывания.

Likely Pathogenic

ITPKB (chr1, позиция 226737174)

- **disruptive inframe deletion** c.276_284delCAGCGGCAG
- **белковая мутация:** p.Ser93_Ser95del — потеря аминокислот серина в позиции 93 и 95 в белке

Ген *ITPKB* (инозитол-трифосфаткиназа В) участвует в **клеточном метаболизме и может быть связан с регуляцией клеточного сигнала**.

Тип вариации: Microsatellite, что указывает на возможную микросателлитную нестабильность, которая может быть связана с различными заболеваниями, включая рак и сердечно-сосудистые заболевания.

CLNDISDB: Связь с несколькими медицинскими идентификаторами и заболеваниями, такими как миелопролиферативная неоплазия, что может указывать на редкие заболевания с возможной сердечно-сосудистой компонентой. *Частота встречаемости варианта VAF (Variant Allele Frequency) 0.6304.*

Показано на NCBI, что вариант в гене может повлиять на метаболизм клеток и процессы передачи сигналов, что в свою очередь может повлиять на развитие болезней, таких как ишемическая болезнь сердца, особенно если этот вариант влияет на регуляцию клеточного ответа в сосудистой ткани или сердечной мышце.

Pathogenic

PROC (chr 2, позиция 127426090)

- **missense_variant (T > G)**
- **c.541T>G – изменение аминокислоты в белке: p.Phe181Val (замена фенилаланина на валин в позиции 181)**

Ген **PROC** (протеин C) участвует в антикоагулянтной системе организма и играет важную роль в регуляции свертывания крови. **CLNDISDB:** вариант ассоциирован с различными заболеваниями: **thrombophilia due to protein C deficiency, autosomal dominant** — тромбофилия, вызванная дефицитом протеина C, в доминантной форме, **thrombophilia due to protein C deficiency, autosomal recessive** — тромбофилия, вызванная дефицитом протеина C, в рецессивной форме, **OMIM:176860** и **Orphanet:745** — заболевания, связанные с нарушениями свертывания крови и тромбообразованием. В аннотации также указано наличие **upstream_gene_variant** для гена **MIR4783**, что может свидетельствовать о влиянии на регуляцию гена, расположенного выше по течению, но это не оказывает значительного воздействия на основной вариант в PROC. Недавние исследования показывают, что дефицит протеина C может увеличивать склонность к образованию тромбов, что, в свою очередь, повышает риск ишемической болезни сердца, инсульта и

других сосудистых заболеваний. **Миссенс-мутация**, как в этом варианте, может привести к аномальному функционированию протеина C, что нарушает его антикоагулянтные свойства, увеличивая вероятность тромбообразования.

ALDH5A1 (chr 6, позиция 24503362)

- **missense_variant: C > T**
- мутация с.538C>T приводит к замене аминокислоты гистидина (His) на тирозин (Tyr) в позиции 180 (p.His180Tyr)

ALDH5A1 — этот ген кодирует белок, называемый альдегиддегидрогеназа 5A1, который участвует в метаболизме гамма-аминомасляной кислоты (ГАМК) и играет важную роль в детоксикации токсичных метаболитов. **CLNDISDB**: вариант ассоциирован с несколькими заболеваниями, включая дефицит сукцинат-семальдегидрогеназы, нарушениями в метаболизме и неврологическими расстройствами. Хотя прямое связанное с ишемической болезнью сердца (ИБС) не установлено, нарушения в метаболизме, связанные с *ALDH5A1*, могут оказывать влияние на систему, участвующую в воспалении и регуляции сосудистого тонуса, что в свою очередь может быть связано с развитием заболеваний, таких как ИБС. Мутации в генах, таких как *ALDH5A1*, могут нарушать сосудистую функцию, увеличивая риск сердечно-сосудистых заболеваний, особенно если мутация влияет на клеточные процессы, связанные с воспалением или метаболизмом липидов.

CYP21A2 (chr 6, позиция 32038610)

- **missense_variant: A > T**
- замена аминокислоты гистидина (His) на лейцин (Leu) в позиции 63 (p.His63Leu)

CYP21A2 — этот ген кодирует **21-гидроксилазу**, фермент, который участвует в синтезе стероидных гормонов в надпочечниках, таких как кортизол и альдостерон. Недостаток этого фермента приводит к нарушению гормонального баланса и может вызвать конгенитальную гиперплазию надпочечников. Прямое отношение между **CYP21A2** и ишемической болезнью сердца (ИБС) не установлено, так как этот ген в основном ассоциирован с эндокринными нарушениями и гиперплазией надпочечников. Однако, изменения в метаболизме стероидов и воспаления,

связанные с дефицитом 21-гидроксилазы, могут косвенно влиять на сердечно-сосудистую систему. Например, стресс и хроническое воспаление, характерные для дефицита гормонов, могут повышать риски развития сердечно-сосудистых заболеваний, включая ИБС.

NQO1 (chr 16, позиция 69711242)

- **missense_variant: G > A**
- замена аминокислоты **пролина (Pro)** на **серин (Ser)** в позиции 187 (**p.Pro187Ser**)

NQO1 кодирует фермент **NAD(P)H:quinone oxidoreductase 1**, который участвует в детоксикации организма, защищая клетки от повреждений, вызванных окислительным стрессом. Он также играет роль в метаболизме различных химических веществ, включая токсины и канцерогены. Вариант имеет статус **Conflicting classifications of pathogenicity** (противоречивые классификации патогенности), что указывает на различные мнения о его патогенности. Вариант описан, как связанный с дефицитом или дисфункцией *NQO1*, увеличением риска развития рака легких, повышенной восприимчивостью к токсичности бензола, ухудшением прогноза для выживаемости при раке молочной железы после химиотерапии. В случае *NQO1* может нарушиться его способность детоксифицировать клетки. Прямое отношение между *NQO1* и ишемической болезнью сердца (ИБС) не установлено, но изменение активности *NQO1* может повлиять на уровень окислительного стресса в клетках и тканях. Окислительный стресс связан с воспалением и повреждением сосудов, что может быть фактором риска для ИБС. Также, нарушение метаболизма токсинов и канцерогенов может повысить риск сердечно-сосудистых заболеваний через воздействия на организм в целом.

RNF213 (chr 17, позиция 80385145)

- **missense_variant: G > A**
- **intron_variant:** для гена *RNF213-AS1*
- замена аминокислоты **аргинина (Arg)** на **лизин (Lys)** в позиции 4810 (**p.Arg4810Lys**)

Связан с **болезнью Мойя-Мойя** (Moyamoya disease), кодирует белок, который участвует в различных клеточных процессах, включая регулирование клеточного цикла и поддержание сосудистого здоровья. Белок *RNF213* играет важную роль в поддержании сосудистой

целостности и функциональности, и его нарушение может быть связано с сосудистыми расстройствами, которые могут также повлиять на сердце.

FECH (chr 18, позиция 57571588)

- **intron_variant**

FECH кодирует ключевой фермент в пути биосинтеза гемоглобина, который катализирует последнюю стадию синтеза гема. Мутация **c.333-48T>C** находится в интронной области, что может повлиять на правильное сплайсирование РНК или нарушить регуляцию экспрессии гена, что может привести к дефициту фермента и нарушениям в метаболизме порфиринов. Вариант ассоциирован с эритропоэтической протопорфирией: накоплением протопорфирина в эритроцитах и других тканях, нарушения в метаболизме порфиринов, как в случае эритропоэтической протопорфирии, могут косвенно повлиять на сосудистую систему, поскольку накопление порфиринов может привести к повреждению сосудов. В редких случаях такие нарушения могут быть связаны с заболеваниями, влияющими на кровоснабжение, что теоретически может повысить риск ишемической болезни сердца.

APOE (chr 19, позиция 44908684)

- **missense_variant: T > C**

- замена **цистеин (Cys)** на **аргинин (Arg)** на позиции 156

АРОЕ участвует в метаболизме липидов и холестерина, а также в функционировании мозга. Он играет ключевую роль в транспортировке холестерина в нейронах. Вариант ассоциирован с несколькими заболеваниями: болезнь Альцгеймера), ответ на антикоагулянтное лечение (варфарин), нарушения обмена липидов и холестерина. Прямое отношение между АРОЕ и ишемической болезнью сердца (ИБС) в контексте данного варианта не установлено. Однако известно, что АРОЕ играет ключевую роль в метаболизме липидов, и определенные аллели АРОЕ могут увеличивать риск атеросклероза и, как следствие, ИБС, особенно АРОЕ **ε4**. Этот вариант **p.Cys156Arg** может потенциально изменять функциональные характеристики белка АРОЕ, что влияет на липидный обмен и может быть связано с риском атеросклероза и сердечно-сосудистых заболеваний.

