NGS Resequencing

Лев Мазаев | мАДБМ18

Этап 1: установка ПО и скачивание файлов

Так как задание будет выполняться локально, установим все необходимые программы с помощью conda:

```
conda install -c bioconda bwa picard samtools igv
conda install -c bioconda gatk
```

snpEff скачаем и будем использовать отдельно, чтобы можно было выделять память, так как версия из conda вылетает из-за нехватки памяти.

Затем нужно дополнительно скачать `GenomeAnalysisTK.jar` и зарегистрировать:

```
gatk3-register ~/GenomeAnalysisTK.jar
```

Теперь скачаем файлы, а именно `chr13.fa` и парные чтения. Так как версия локально установленной программы bwa может отличаться от той, с помощью которой создавался индекс, скачивать его не будем, а создадим свой.

Проверка файлов прочтений с помощью FastQC показала, что с ними все в порядке.

Этап 2: выравнивание и дедупликация

Создадим индекс bwa:

```
bwa index -p chr13_idx chr13.fa
```

И произведём выравнивание (id ERR232255 из заголовков ридов):

```
bwa mem -M -R
"@RG\tID:ERR232255_chr13\tLIB:ERR232255\tPL:ILLUMINA\tSM:ERR232255" -t 8
chr13_idx reads_1.fastq.gz reads_2.fastq.gz | samtools view -@ 8 -b >
unsorted.chr13.bam
```

Посмотрим, что получилось:

```
(bio) leo@MS7922:~/BioData/NGS_HW9$ samtools flagstat unsorted.chr13.bam
3855858 + 0 in total (QC-passed reads + QC-failed reads)
2734 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
3855554 + 0 mapped (99.99% : N/A)
3853124 + 0 paired in sequencing
1926562 + 0 read1
1926562 + 0 read2
3800516 + 0 properly paired (98.63% : N/A)
3852618 + 0 with itself and mate mapped
202 + 0 singletons (0.01% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Исходно у нас было 1,926,562 пары прочтений, то есть 3,853,124 прочтений. Как видим по статистике флагов, **закартировались 99.99% прочтений**, при этом в строке с пометкой secondary подсчитаны случаи вторичного выравнивания. То есть есть 2734 случая, когда прочтение выровнялось 2 и более раз. Отсортируем и проиндексируем файл:

```
samtools sort -@ 8 unsorted.chr13.bam > sorted.chr13.bam
samtools index -@ 8 sorted.chr13.bam
```

Теперь произведём дедупликацию:

```
picard MarkDuplicates I=sorted.chr13.bam O=dedup.sorted.chr13.bam
M=dedup_metrics.txt
samtools index -@ 8 dedup.sorted.chr13.bam
```

Судя по результатам в метрике здесь **4.38% дуплицированных прочтений**. Статистика по флагам:

```
(bio) leo@MS7922:~/BioData/NGS_HW9$ samtools flagstat dedup.sorted.chr13.bam
3855858 + 0 in total (QC-passed reads + QC-failed reads)
2734 + 0 secondary
0 + 0 supplementary
168927 + 0 duplicates
3855554 + 0 mapped (99.99% : N/A)
3853124 + 0 paired in sequencing
1926562 + 0 read1
1926562 + 0 read2
3800516 + 0 properly paired (98.63% : N/A)
3852618 + 0 with itself and mate mapped
202 + 0 singletons (0.01% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

показывает тоже самое (если поделить 168,927 на число прочтений).

Этап 3: поиск и аннотация вариантов

Сперва нужно создать индекс и словарь для хромосомы:

```
samtools faidx chr13.fa
picard CreateSequenceDictionary R=chr13.fa O=chr13.dict
```

Теперь воспользуемся HaplotypeCaller из пакета GATK. При этом удалять дупликаты из `bam`-файла не нужно, так как DuplicateReadFilter также запускается в процессе.

```
gatk3 -R chr13.fa -T HaplotypeCaller -I dedup.sorted.chr13.bam -L
chr13:32889617-32973809 -stand_call_conf 30 -o vars.vcf -bamout reas.bam
```

Получили нужный `vars.vcf`, а также результат пересборки - `reas.bam`. Теперь snpEff, чтобы аннотировать мутации:

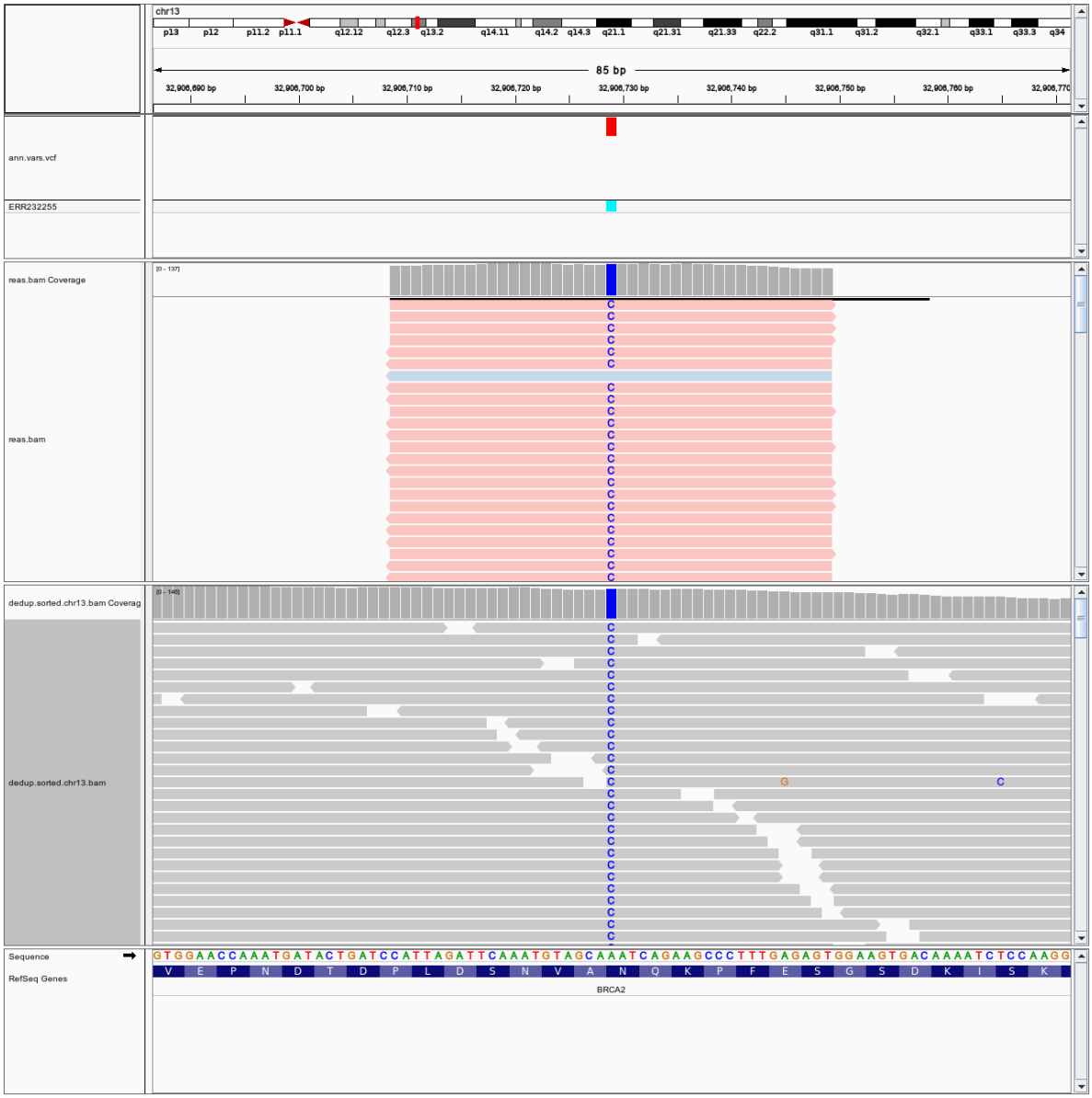
```
java -Xmx4g -jar ./snpEff/snpEff.jar ann -t -v hg19 vars.vcf > ann.vars.vcf
```

Теперь получим список мутаций с эффектом 'HIGH' или 'MODERATE':

```
egrep '(HIGH|MODERATE)' ann.vars.vcf > list1.txt
```

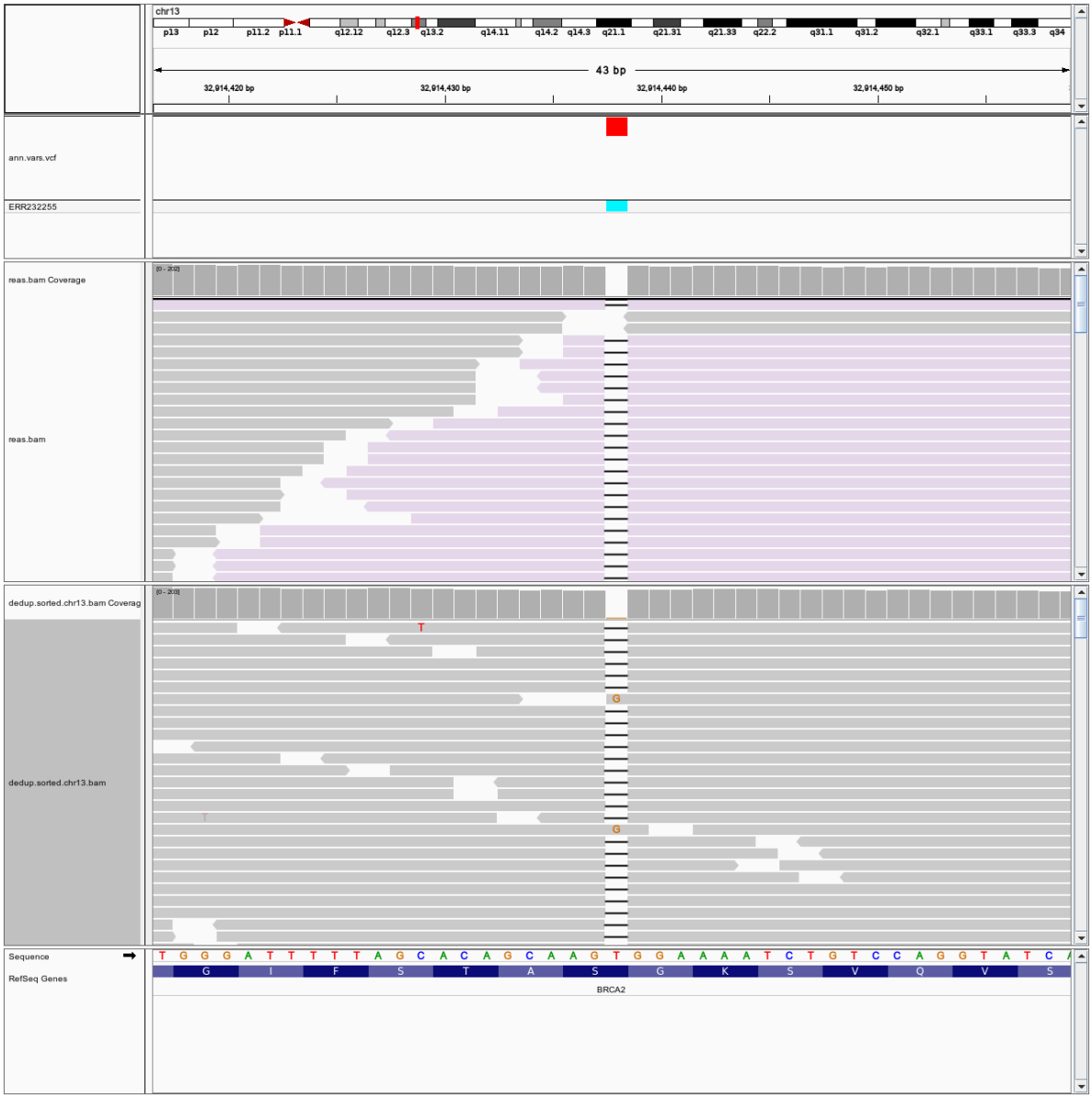
> chr13 32906729 . A C 4427.77 .
> AC=2;AF=1.00;AN=2;DP=129;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;
> QD=34.59;SOR=1.270;ANN=C|missense_variant|MODERATE|BRCA2|BRCA2|transcript|NM
> _000059.3|protein_coding|10/27|c.1114A>C|p.Asn372His|1341/11386|1114/10257|372/3
> 418|| GT:AD:DP:GQ:PL 1/1:0,128:128:99:4456,385,0
>
> chr13 32914437 . GT G 6888.73 .
> AC=2;AF=1.00;AN=2;DP=175;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;
> QD=34.24;SOR=0.827;ANN=G|frameshift_variant|HIGH|BRCA2|BRCA2|transcript|NM_000
> 059.3|protein_coding|11/27|c.5946delT|p.Ser1982fs|6173/11386|5946/10257|1982/3418
> ||;LOF=(BRCA2|BRCA2|1|1.00) GT:AD:DP:GQ:PL 1/1:0,173:173:99:6926,521,0
>
> chr13 32929387 . T C 2858.77 .
> AC=2;AF=1.00;AN=2;DP=89;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;Q
> D=32.12;SOR=0.809;ANN=C|missense_variant|MODERATE|BRCA2|BRCA2|transcript|NM_
> 000059.3|protein_coding|14/27|c.7397T>C|p.Val2466Ala|7624/11386|7397/10257|2466/3
> 418|| GT:AD:DP:GQ:PL 1/1:0,89:89:99:2887,267,0

Как видим, таких мутаций нашлось только, две из них - несинонимичные замены, а третья - сдвиг рамки считывания, в данном случае делеция. Визуализирум с помощью IGV (для сборки гаплотипов `reas.bam` включим окрашивание по тэгу HC):



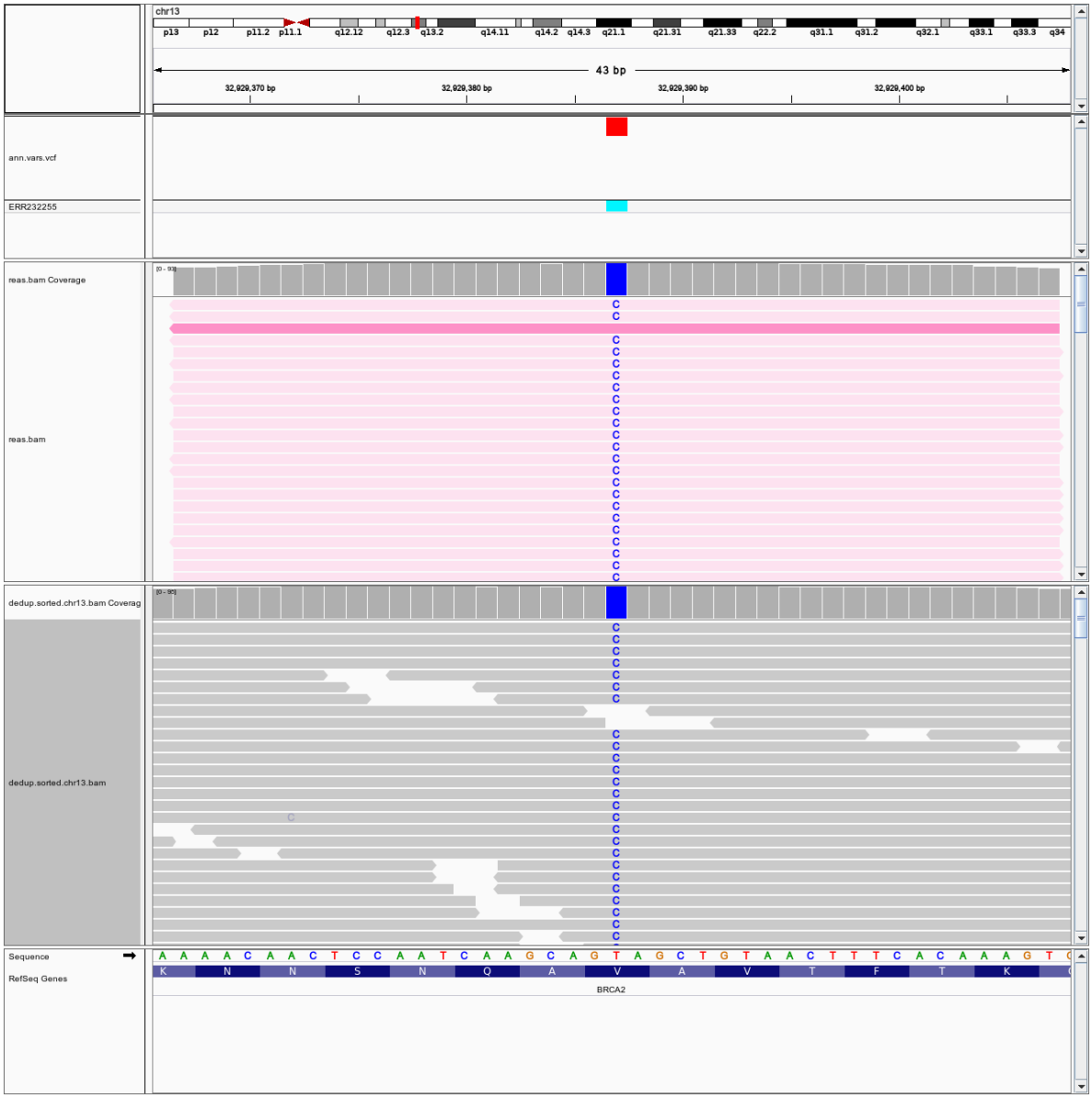
Это первая из мутаций, замена референсного аденина на цитозин, которая приводит к замене аспарагина на гистидин. Видна и разница между `bam`-файлами: в пересобранном (`reas.bam`) имеется дополнительное "искусственное" прочтение (129 + 1, в то время как в оригинале - 129) для референсной аллели, к тому же все риды обрезаны до некоторого окна и пересортированы.

Следующая мутация - делеция: в референсе GT, альтернатива - G (удаление тимина), в некоторых прочтениях также GG:



Видна разница между файлами: в пересобранном отсутствуют прочтения, в которых GT заменилось на GG, есть только варианты с делецией и референс. Эта мутация приводит к замене серина на аргинин, а также сдвигает рамку считывания.

Следующая мутация - замена референсного тимина на цитозин (валин заменяется на аланин):



Разница между файлами - как в первом случае + несколько ридов отфильтровано (92 -> 90+1).

Если запустить HaplotypeCaller без параметра stand_call_conf, то получается тоже самое.

Этап 4*. Более "жесткая" фильтрация вариантов

Воспользуемся следующим руководством.

Сперва извлечём SNPs в файл `raw_snps.vcf`:

```
gatk3 -T SelectVariants -R chr13.fa -V vars.vcf -selectType SNP -o
raw_snps.vcf
```

Теперь профильтруем все варианты с указанными в руководстве параметрами:

```
gatk3 -T VariantFiltration -R chr13.fa -V raw_snps.vcf --filterExpression "QD
< 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0"
--filterName "generic_filter" -o filtered_snps.vcf
```

Повторим обе операции для инделов:

```
gatk3 -T SelectVariants -R chr13.fa -V vars.vcf -selectType INDEL -o
raw_indels.vcf
gatk3 -T VariantFiltration -R chr13.fa -V raw_indels.vcf --filterExpression
"QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0" --filterName
"generic_indel_filter" -o filtered_indels.vcf
```

Следом SnpSift, чтобы оставить только те варианты, которые получили пометку PASS в столбце:

```
cat filtered_indels.vcf | java -jar ./snpEff/SnpSift.jar filter "(FILTER =
'PASS')" > indels_pass.vcf
cat filtered_snps.vcf | java -jar ./snpEff/SnpSift.jar filter "(FILTER =
'PASS')" > snps_pass.vcf
```

И запустим snpEff на полученных файлах `snps_pass.vcf` и `indels_pass.vcf`:

```
java -Xmx4g -jar ./snpEff/snpEff.jar ann -t -v hg19 indels_pass.vcf >
ann.indels.vcf
java -Xmx4g -jar ./snpEff/snpEff.jar ann -t -v hg19 snps_pass.vcf >
ann.snps.vcf
```

Наконец, найдём варианты с эффектом HIGH или MODERATE:

```
egrep '(HIGH|MODERATE)' ann.indels.vcf > list2.txt
egrep '(HIGH|MODERATE)' ann.snps.vcf >> list2.txt
```

Получили список тех же вариантов, они все прошли проверку с тем набором параметров, что мы использовали:

> chr13 32914437 . GT G 6888.73 PASS AC=2;AF=1.00;AN=2;DP=175;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=34.24;SOR=0.827;ANN=G|frameshift_variant|HIGH|BRCA2|BRCA2|transcript|NM_000059.3|protein_coding|11/27|c.5946delT|p.Ser1982fs|6173/11386|5946/10257|1982/3418||;LOF=(BRCA2|BRCA2|1|1.00) GT:AD:DP:GQ:PL 1/1:0,173:173:99:6926,521,0
>
> chr13 32906729 . A C 4427.77 PASS AC=2;AF=1.00;AN=2;DP=129;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=34.59;SOR=1.270;ANN=C|missense_variant|MODERATE|BRCA2|BRCA2|transcript|NM_000059.3|protein_coding|10/27|c.1114A>C|p.Asn372His|1341/11386|1114/10257|372/3418|| GT:AD:DP:GQ:PL 1/1:0,128:128:99:4456,385,0
>
> chr13 32929387 . T C 2858.77 PASS AC=2;AF=1.00;AN=2;DP=89;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=32.12;SOR=0.809;ANN=C|missense_variant|MODERATE|BRCA2|BRCA2|transcript|NM_000059.3|protein_coding|14/27|c.7397T>C|p.Val2466Ala|7624/11386|7397/10257|2466/3418|| GT:AD:DP:GQ:PL 1/1:0,89:89:99:2887,267,0