

Введение в анализ NGS

Продолжение

Анастасия Жарикова

azharikova89@gmail.com

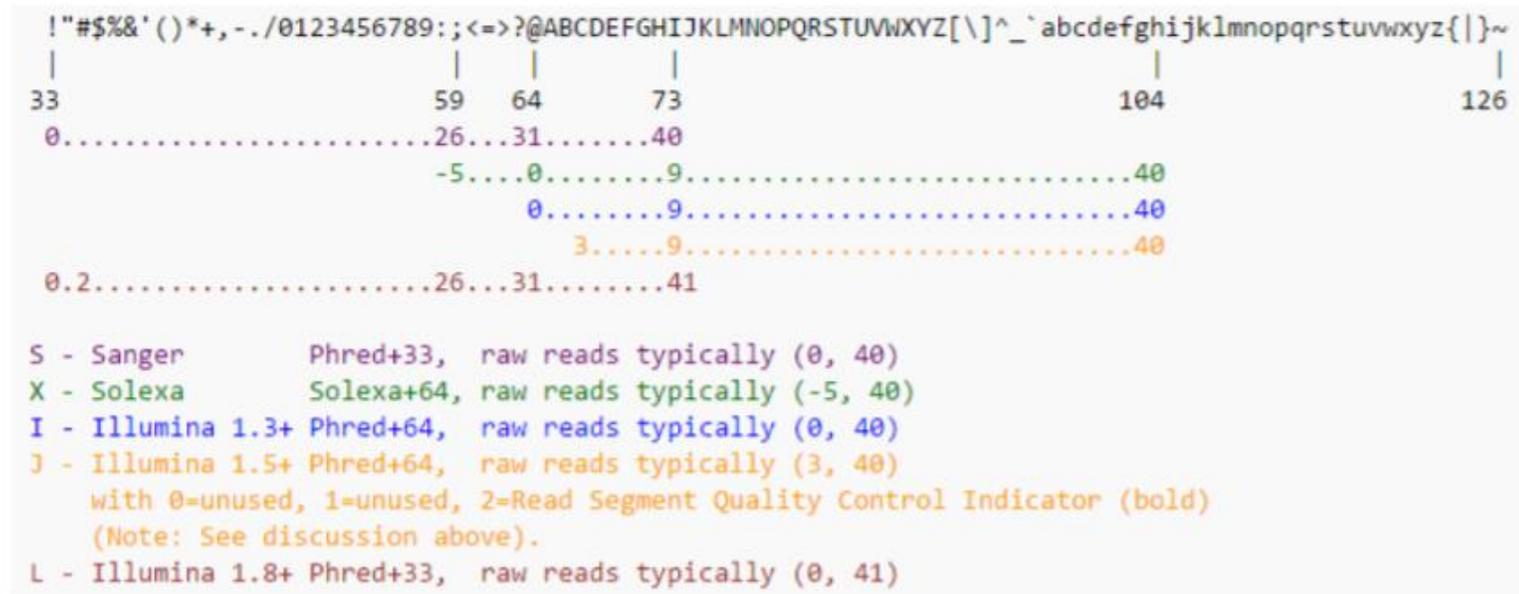
Мы пройдем много форматов файлов
для хранения данных

Все они будут на коллоквиуме

<https://genome.ucsc.edu/FAQ/FAQformat.html>

Fastq формат

```
@HWI-ST992:147:D22HDACXX:3:1112:14175:15297 2:N:0:GGCTAC
Последовательность TAATGGCTTTTCCAAAACGCTCCACTCTTAAAGATGTGTATAAGAGACAGCAACAACAATTA
+
Качество 8??DDDBEDHFFHJJJJJJJAFGIIIIIGIGEEGIIIIHBFGEEGCGIJIFFIDIIJJIIII
```



Что делать?

Нужно удалить «плохие» фрагменты чтений:

- Адаптеры
- Нуклеотиды с неудовлетворительным качеством (< 20)

Trimmomatic

<http://www.usadellab.org/cms/?page=trimmomatic>

В результате получаем только те чтения, качество которых нас устраивает

С ними можно смело работать дальше!

Что делать дальше?

Дано:

- «очищенные» чтения хорошего качества (fastq)
- Последовательность референсного генома (fasta)

Задача:

Каждому чтению найти свое место на геноме -
картирование

Картирование

Программы:

- bowtie
- bwa
- hisat2

Есть много других!

Шаг 0. Подготовка референса: индексирование
Для каждой программы свой индекс!

Шаг следующий – картирование чтений на референс
Получаем .sam или .bam

sam

Содержит заголовок и информацию о картировании чтений

<http://samtools.github.io/hts-specs/SAMv1.pdf>

```
SRR2776256.15395984      0      chr12    9822304 60      100M    *      0
0      AGATCACTCATAGAAACTGGAGGCAAAAATGCATGACAGTAACAATGTGGAGAAAGACATTACACCATCTGAA
TTGCCTGCAAAGCCAGGTAAGAAGCTGG      ?@DFFFDHHHHHJIJIHEGFAGHEG;FCFDFHI<GIJCFFDH?<<00
?98929/0.=B:8B78CC=CCEAAH=)=ECCB;7B;>@362@;@@C@CD359      AS:i:-4 XN:i:0 XM:i:1
XO:i:0 XG:i:0 NM:i:1 MD:Z:83C16      YT:Z:UU NH:i:1
SRR2776256.23192736     16      chr12    9822307 60      100M    *      0
0      TCACTCATAGAAACTGGAGGCAAAAATGCATGACAGTAACAATGTGGAGAAAGACATTACACCATCTGAATTG
CCTGCAAACCCAGGTAAGAAGCTGGGCT      CCCC>;CEECEEEC@=DBC>ACHEHCD@=;G@GGGEHF=C<>IHFFGB
HGCDDGHGDFD?HGHEGGHFFGFA>GFH@HFADCHEHNBFFHHFFDDDD@@@      AS:i:0 XN:i:0 XM:i:0
XO:i:0 XG:i:0 NM:i:0 MD:Z:100      YT:Z:UU NH:i:1
```

SRR2776256.15395984 – ID чтения

chr12 9822304 - хромосома и координата, куда «легло» чтение

100M – CIGAR: сжато кодирует информацию о выравнивании чтения

NM:i – расстояние до генома

NH:i – количество картирований для данного чтения

samtools

<http://www.htslib.org/doc/samtools.html>

Этот пакет поможет отсортировать и индексировать .bam, узнать покрытие фрагмента генома и многое другое

Читайте мануал и подсказки к заданию!

Помните, что bam файлы должны быть отсортированы по координате и индексированы

Дублированные чтения

Бывают ПЦР-дубли и оптические
Дубли можно удалять, можно маркировать

```
$ samtools view foo.bam
SRR7012201.2594959      0      chr1    3000061 30      50M    *      0      0
SRR7012201.2594959      0      chr1    3000061 30      50M    *      0      0
(base)

$ samtools markdup foo.bam - | samtools view
SRR7012201.2594959      0      chr1    3000061 30      50M    *      0      0
SRR7012201.2594959     1024   chr1    3000061 30      50M    *      0      0
(base)
```

IGV

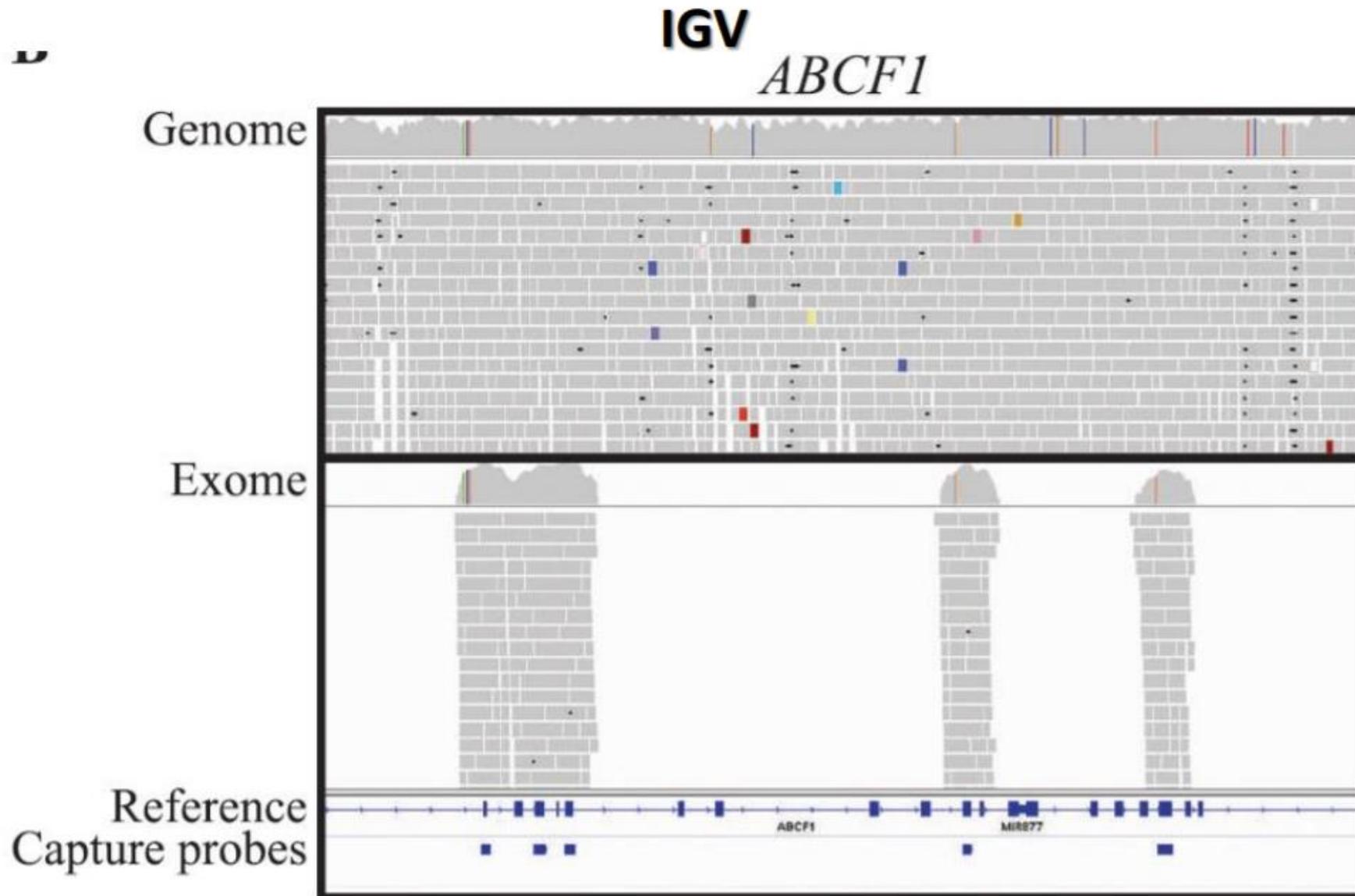
<https://software.broadinstitute.org/software/igv/download>

Программа для визуализации bam файлов с выравниваниями чтений на референсный геном.

Нужно установить себе на компьютер.

Принимает на вход bam (сортированный) + .bam.bai

https://www.youtube.com/watch?v=E_G8z_2gTYM&t=76s

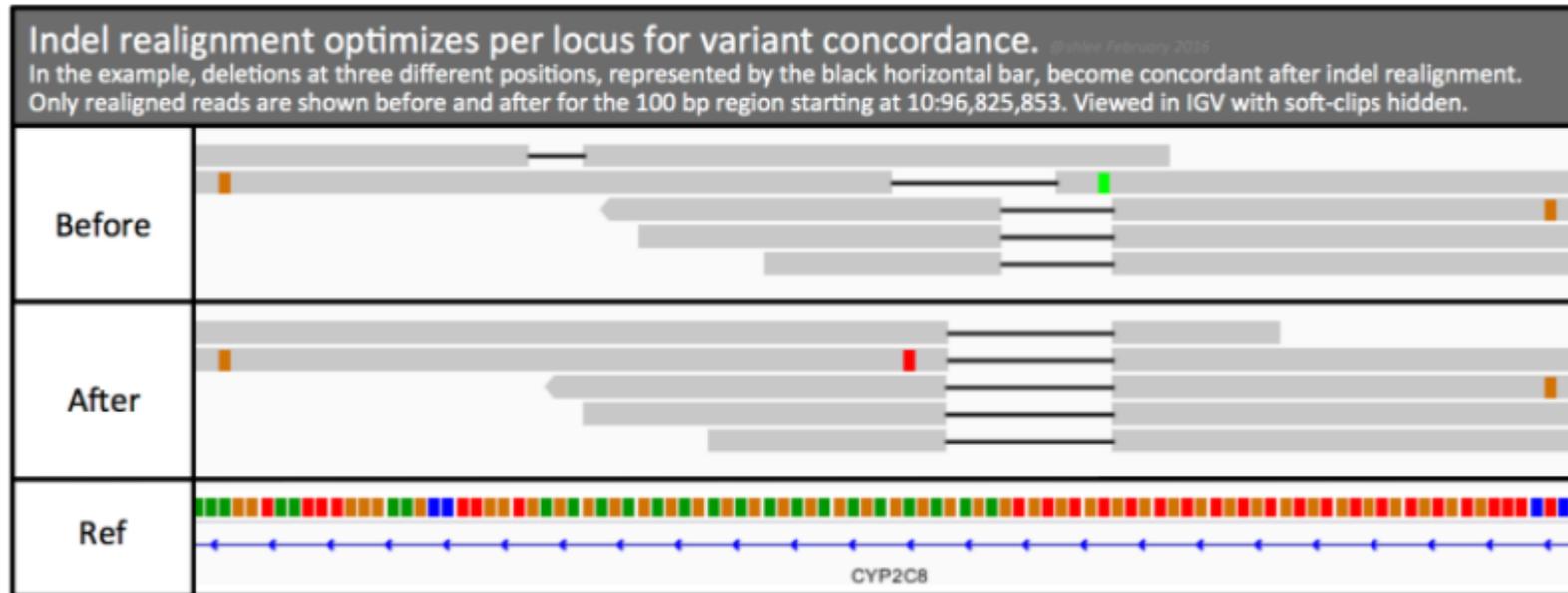


Emily M. Coonrod, Jacob D. Durtschi, Rebecca L. Margraf, and Karl V. Voelkerding (2013) Developing Genome and Exome Sequencing for Candidate Gene Identification in Inherited Disorders: An Integrated Technical and Bioinformatics Approach. Archives of Pathology & Laboratory Medicine: March 2013, Vol. 137, No. 3, pp. 415-433.

GATK3

<https://gatk.broadinstitute.org/hc/en-us/sections/360007226651-Best-Practices-Workflows>

Indel realignment



Поиск вариантов

.bam

.gvcf

.vcf

.filt.vcf

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
20 10001567 . A <NON_REF> . . END=10001616 GT:DP:GQ:MIN_DP:PL 0/0:38:99:34:0,101,11
20 10001617 . C A,<NON_REF> 493.77 . BaseQRankSum=1.632;ClippingRankSum=0.000;DP=38;Excess
20 10001618 . T <NON_REF> . . END=10001627 GT:DP:GQ:MIN_DP:PL 0/0:39:99:37:0,105,15
20 10001628 . G A,<NON_REF> 1223.77 . DP=37;ExcessHet=3.0103;MLEAC=2,0;MLEAF=1.00,0.00;RAW_
20 10001629 . G <NON_REF> . . END=10001660 GT:DP:GQ:MIN_DP:PL 0/0:43:99:38:0,102,12
```

```
chr1 28563 . A G 139.90 PASS AC=2;AF=1.00;AN=2;DP=5;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ
chr1 49298 . T C 515.77 PASS AC=2;AF=1.00;AN=2;DP=17;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;M
chr1 52238 . T G 716.77 PASS AC=2;AF=1.00;AN=2;DP=22;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;M
chr1 55926 . T C 120.90 PASS AC=2;AF=1.00;AN=2;DP=5;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ
chr1 61442 . A G 314.77 PASS AC=2;AF=1.00;AN=2;DP=10;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;M
chr1 61947 . C T 357.77 PASS AC=1;AF=0.500;AN=2;BaseQRankSum=3.01;ClippingRankSum=0.00;DP=33;Exces
chr1 61987 . A G 703.77 PASS AC=1;AF=0.500;AN=2;BaseQRankSum=0.426;ClippingRankSum=0.00;DP=42;Exces
chr1 61989 . G C 703.77 PASS AC=1;AF=0.500;AN=2;BaseQRankSum=0.125;ClippingRankSum=0.00;DP=41;Exces
chr1 69511 . A G 358.77 PASS AC=2;AF=1.00;AN=2;DP=13;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;M
chr1 83084 . T A 204.80 PASS AC=2;AF=1.00;AN=2;DP=7;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ
```

```
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA000
20 14370 rs6054257 C A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 CT:CQ:DP:HQ 010:48:1:51,51 110:48:8:51,51 1/1:4
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 CT:CQ:DP:HQ 010:49:3:58,50 011:3:5:65,3 0/0:4
20 1110696 rs6040355 A C,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB CT:CQ:DP:HQ 112:21:6:23,27 211:2:0:18,2 2/2:3
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T CT:CQ:DP:HQ 010:54:7:56,60 010:48:4:51,51 0/0:6
20 1234567 microsat1 GTC C,GTCT 50 PASS NS=3;DP=9;AA=G CT:CQ:DP 0/1:35:4 0/2:17:2 1/1:4
```

Основные шаги

- Проверка качества чтений
- Триммирование
- Проверка качества триммированных чтений
- Картирование чтений на геном (sam)
- Конвертация sam в bam
- Сортировка bam
- Индексирование bam
- Поиск вариантов (vcf)
- Фильтрация вариантов

?Аннотация?

ANNOVAR

<https://doc-openbio.readthedocs.io/projects/annovar/en/latest/>

refgene

dbsnp

1000 genomes

GWAS

Clinvar

Далее – клиническая интерпретация

Bedtools

<http://bedtools.readthedocs.io/en/latest/index.html>

<https://media.readthedocs.org/pdf/bedtools/latest/bedtools.pdf>

Очень хороший инструмент для работы с геномными
интервалами

Более 35 опций + параметры