Ресеквенирование
Поиск полиморфизмов у человека

Анастасия Жарикова

17 ноября 2020

azharikova89@gmail.com

http://enseqlopedia.com/

DNA Sequencing Methods

- Protein-Protein Interaction
  - PD-Seq
  - ProP-PD/PDZ-Seq
- Sequence Rearrangements
  - 2b-RAD
  - CPT-seq
  - ddRADseq
  - Digenome-seq
  - EC-seq
  - hyRAD
  - RAD-Seq
  - Rapture
  - RC-Seq
  - Repli-Seq
  - SLAF-seq
  - TC-Seq
  - Tn-Seq/INSeq
  - Bubble-Seq
  - NSCR
  - NS-Seq
  - Rep-Seq/Ig-Seq/MAF
- DNA Break Mapping
  - BLESS
  - DSB-Seq
  - GUIDE-seq
  - HTGTS
  - LAM-HTGTS
  - Break-seq
  - SSB-Seq
- DNA Protein Interactions
  - DNaseI Seq or DNase-Seq
  - Pu-seq
  - 3-C/Capture-C/Hi-C
  - 4C-seq
  - 5C
  - ATAC-Seq/Fast-ATAC
  - CATCH_IT
  - Chem-seq
  - ChIA-PET
  - ChIPmentation
  - ChIP-Seq/HT-ChIP/ChIP-exo/Mint-ChIP
  - DamID
  - DNase I SIM
  - FAIRE-seq/Sono-Seq
  - FIT-Seq
  - HiTS-FLIP
  - MINCE-seq
  - MNase-Seq/MAINE-Sequcleo-Sequc-seq
  - MPE-seq
  - NG Capture-C
  - NOMe-Seq
  - ORGANIC
  - PAT-ChIP
  - PB_seq
  - SELEX or SELEX-seq / HT-SELEX
  - THS-seq
  - UMI-4C
  - X-ChIP-seq
- Epigenetics
  - Aba-seq
  - BisChIP-Seq/ChIP-BS-Seq/ChIP-BMS
  - BSAS
  - BSPP
  - BS-Seq/Bisulfite-Seq/WGBS
  - CAB-Seq
  - EpiRADseq
  - fCAB-seq
  - fC-CET
  - fC-Seal
  - hMeDIP-seq
  - JBP1-seq
  - MAB-seq
  - MBDCap-seq/MethylCap-Seq/MiGS
  - MeDIP-Seq/DIP-seq
  - MIRA
  - MRE-Seq and Methyl-Seq
  - oxBS-Seq
  - PBAT
  - redBS-Seq/caMAB-seq
  - RRBS-Seq
  - RRMAB-seq
  - TAB-Seq
  - TAmC-Seq
  - T-WGBS
- Low-Level DNA Detection
  - Safe-SeqS
  - scAba-seq
  - scATAC-Seq (Cell index variation)
  - scATAC-Seq (Microfluidics variation)
  - scBS-Seq
  - scM&T-Seq

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
TAATGGCTTTTCCAAAACGCTCCACTCTTAAAGATGTGTATAAGAGACAGCAACAACAATTA
+
8??DDDBEDHHFHJJJJIJAFGIIIIIGIGEEGIIIIHBFGGEEGCGIJIFFIDIIJJIIII
```

Последовательность

Качество

```
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                          |    |        |                              |                     |
33                         59   64       73                            104                   126
 0........................26...31.......40
                           -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Fastq формат

```
@NB551509:7:HHJTJBGXC:1:11101:2231:1116 1:N:0:TGACCA
CATTACGGAATGTATCATCTTCTGAATGTGAACCACATCAGATGCAATACAGAGAAACACACACTCTCCAGGCAC
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@NB551509:7:HHJTJBGXC:1:11101:7127:1116 1:N:0:TGACCA
TTTTTTTCCCCCTCATTACTTTGCTTTTAGCTCACTCCTTGCAGGAATCTTTCCAGCTGCCTACCTAGCCCTTCC
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEE/EEEEEEEEEEEEEEEEEEEEEEEEEEEA
@NB551509:7:HHJTJBGXC:1:11101:2059:1116 1:N:0:TGACAA
CAAATATATTAGACCTTGTCCTGATTTGGAGTATGGCAAAAAATGTGCCATATCATATTCTTACCAAAACATTTG
+
AAAAAEEEEEEAEEEAE/EEAEEEEEEEEEAEEEEEAEEEEEEA/6A/AE/EEAEEE6EEEE/EEEEEE6E/EEE
@NB551509:7:HHJTJBGXC:1:11101:3510:1116 1:N:0:TGACCA
AATGGTTAGAGGTTCTAAATCTTGGGACACGCAGCAAGGAGAAGCAGATGCTTCTGGATTTATGGTATTATATA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEE
@NB551509:7:HHJTJBGXC:1:11101:8048:1117 1:N:0:TGACCA
CCCCCTTCTACAGCTTATAGAGTGTTGGATCCAGGACTGTCAGTCTCTGGAGATCCCAATCGATCCTTCCTTC
+
AAAAAEEEEEE/EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@NB551509:7:HHJTJBGXC:1:11101:5801:1117 1:N:0:TGACCA
CAAAACCTATAACATATTGTATACATATATATAATATATAAACACACATACACAATATAGACTTATCTTGCTCTT
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

Качество чтений

P – вероятность ошибки

Q – параметр качества. (Phred Quality Score)

$$\mathbf{Q} = -10\log_{10}\mathbf{P}$$

Вероятность ошибки	Q
0.001 (*точность 99,9%*)	30
0.01 (*точность 99%*)	20
0.1 (*точность 90%*)	10

Типичные значения Q от 1 до 40

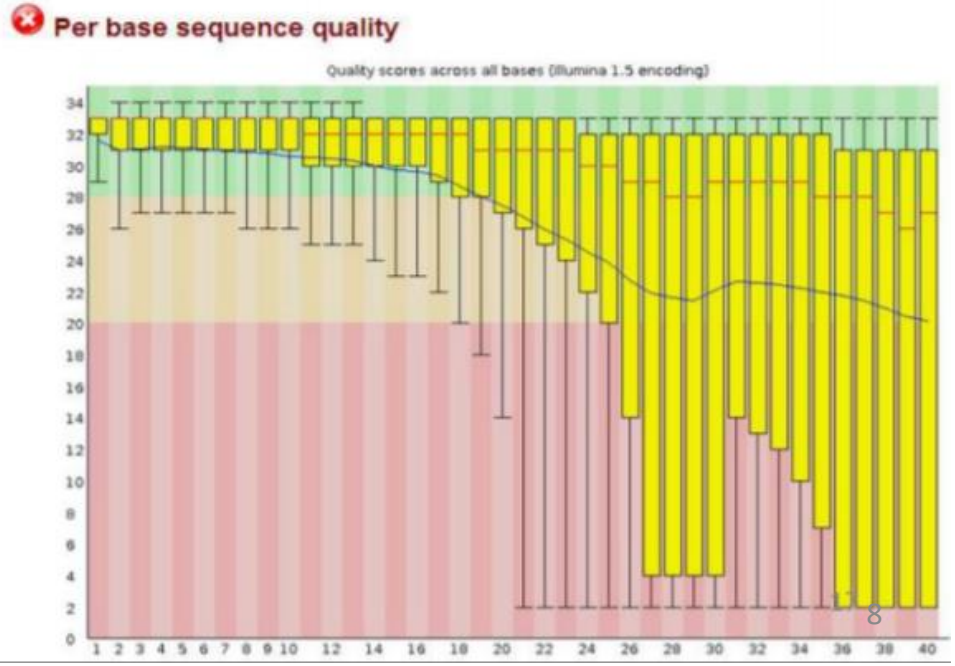
Q>20 – «хорошее качество»

Пересчет качества в вероятность ошибки

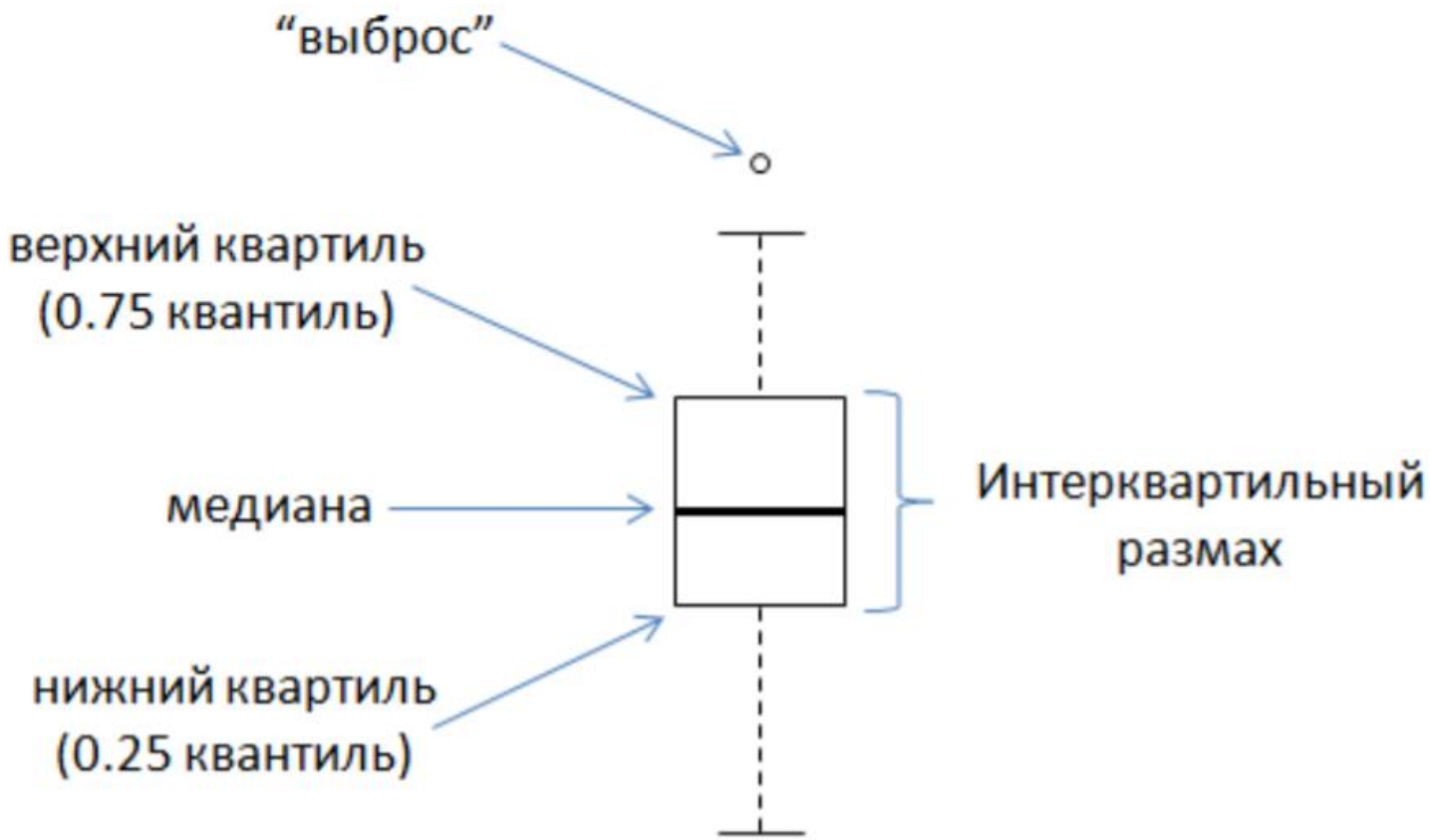
Phred Quality Score	Символ	Вероятность ошибки	Точность
10	+	1/10	90%
20	5	1/100	99%
30	?	1/1000	99,9%
40	I	1/10 000	99,99%
50	S	1/100 000	99,999%
60	]	1/1 000,000	99,9999%

fastQC


Per base sequence quality
Quality scores across all bases (Illumina 1.5 encoding)
38
36
34
32
30
28
26
24
22
20
18
16
14
12


Per base sequence quality
Quality scores across all bases (Illumina 1.5 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40

«Ящик с усами» / диаграмма размахов / boxplot



fastQC

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Программа FasqQC стоит на kodomo

Версию с графическим интерфейсом можно поставить на свой компьютер.

На сайте отличное руководство!

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       AGATCACTCATAGAAACTGGAGGCAAAATGCATGACAGTAACAATGTGGAGAAAGACATTACACCATCTGAA
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       TCACTCATAGAAACTGGAGGCAAAATGCATGACAGTAACAATGTGGAGAAAGACATTACACCATCTGAATTG
CCTGCAAACCCAGGTAAGAAGCTGGGCT     CCCC>;CEECEEEC@=DBC>ACHEHCD@=;G@GGGEHF=C<>IHFFGB
HGCDDGHGDFD?HGHEGGHFFGFA>GFH@HFADCHEHHBFHHHFFDDDD@@@     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

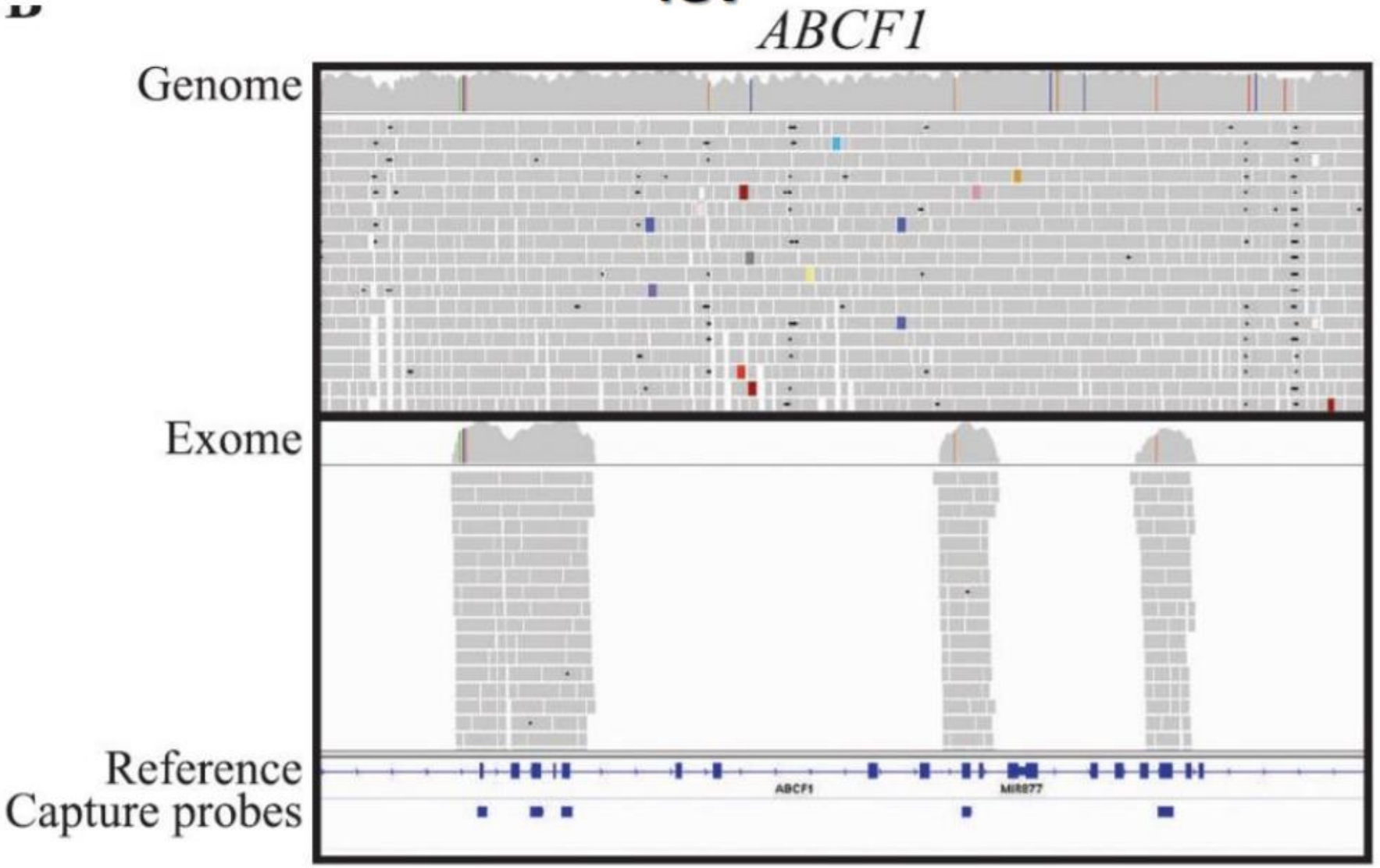
Дублированные чтения

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
```

IGV

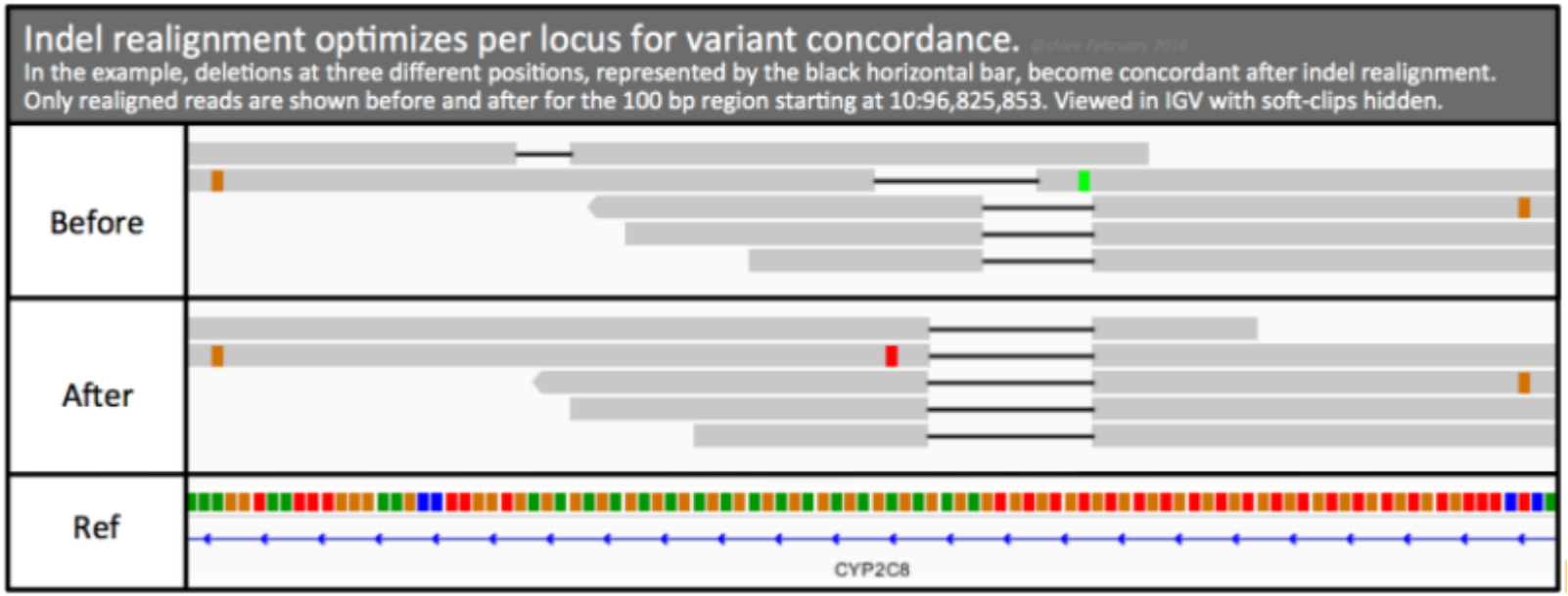


Emily M. Coonrod, Jacob D. Durtschi, Rebecca L. Margraf, and Karl V. Voelkerding (*2013*) Developing Genome and Exome Sequencing for Candidate Gene Identification in Inherited Disorders: An Integrated Technical and Bioinformatics Approach. Archives of Pathology & Laboratory Medicine: March 2013, Vol. 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

```
#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
```

.vcf

```
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 397.77  PASS  AC=1;AF=0.500;AN=2;BaseQRankSum=3.01;ClippingRankSum=0.00;DP=33;Excess
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
```

.filt.vcf

```
##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 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:4
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:4
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:3
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:6
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:4
```

Программный сценарий

Идея

Все команды должны быть в одном месте.
Указаны версии всех используемых программ.
На вход подаются чтения, на выходе получаем файл для дальнейшей обработки.
Воспроизводимость.
Есть неизменяемая часть пайплайна и вариативная.
Можно использовать с любыми входными файлами указанного формата на любой машине.
Сохраняйте «логи» программ.
Архивируйте и индексируйте.