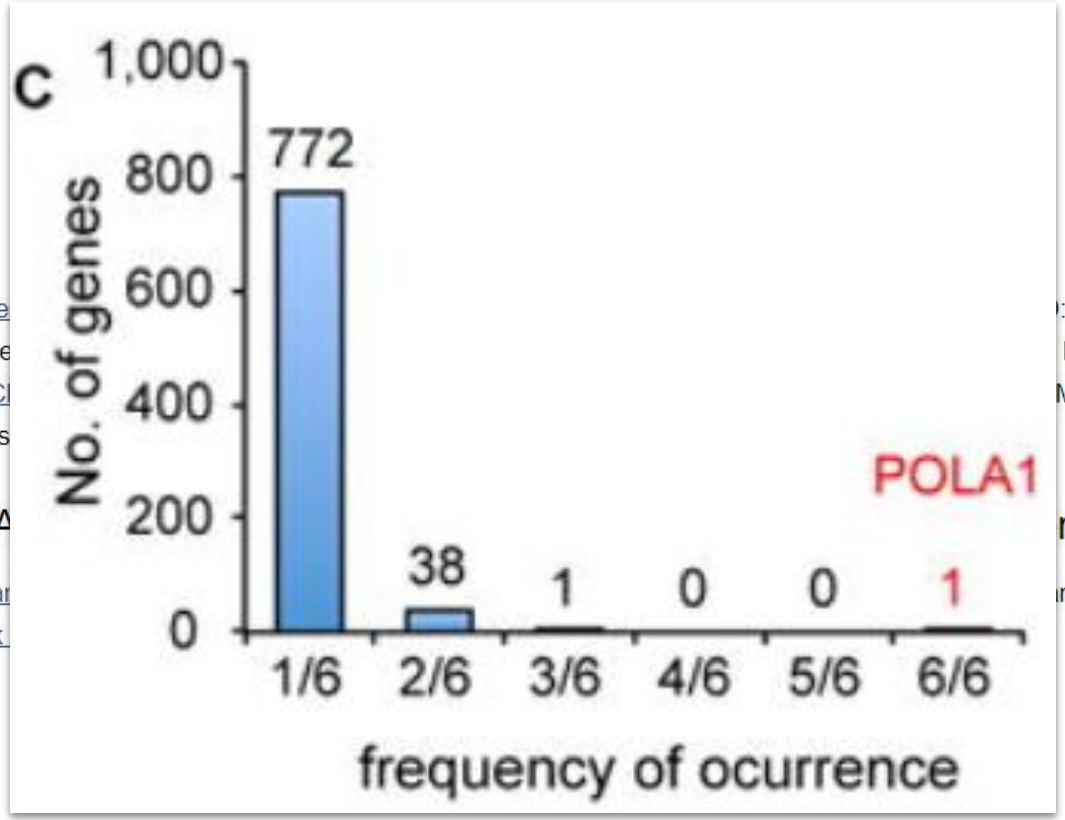
Особенность задания - оно имеет решение :)



Nat Che
Publishe
Nat Cl
Publis

: PMC4912453
NIHMS771697
MID: 27182663

The A
rase α

Ting Har
Deepak
nd

Общая схема решения

картирование

На какие места референсного генома похожи риды в сырых файлах?

поиск SNP

В каких позициях референс и наши данные отличаются??

аннотирование SNP

Какие гены находятся в этих позициях и как SNP может влиять на них?

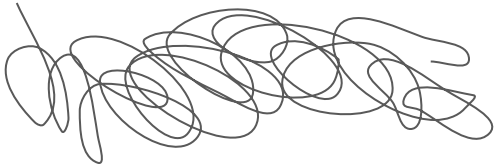
Анализ результатов

Есть ли какие-то очевидные кандидаты??

Красивый ответ

Исходные данные

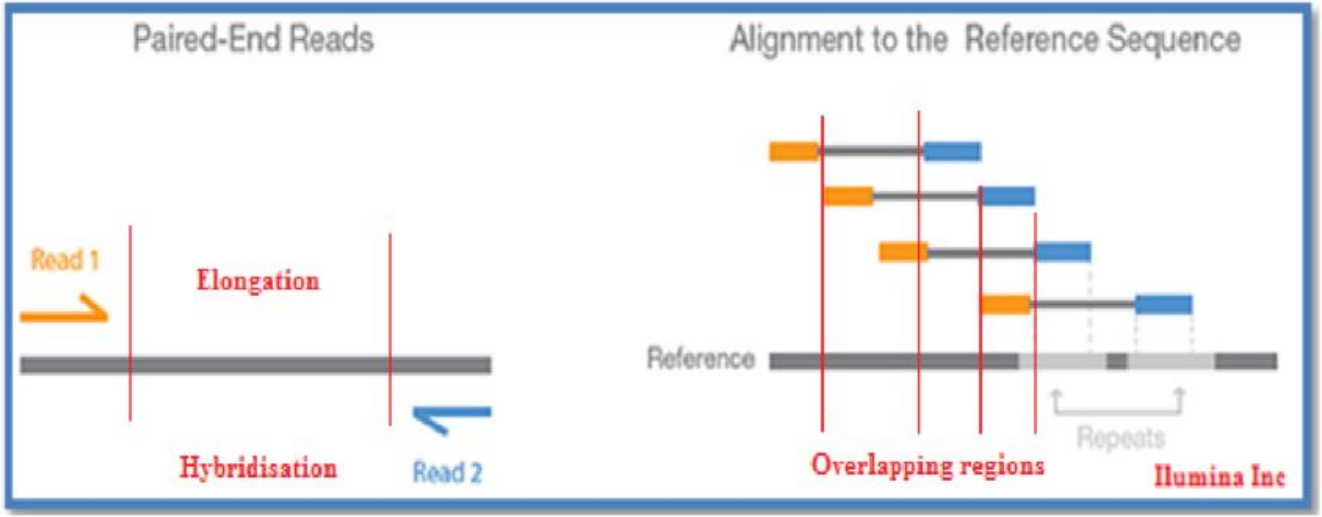
ДНК



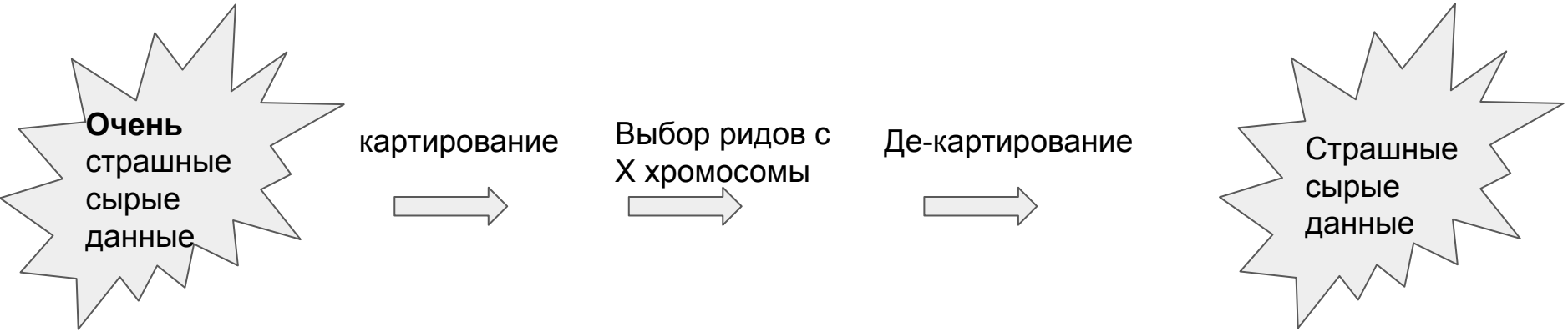
Много кусочков днк

Пришивание адаптеров

Чтение ДНК с обоих адаптеров



Исходные данные - подготовка

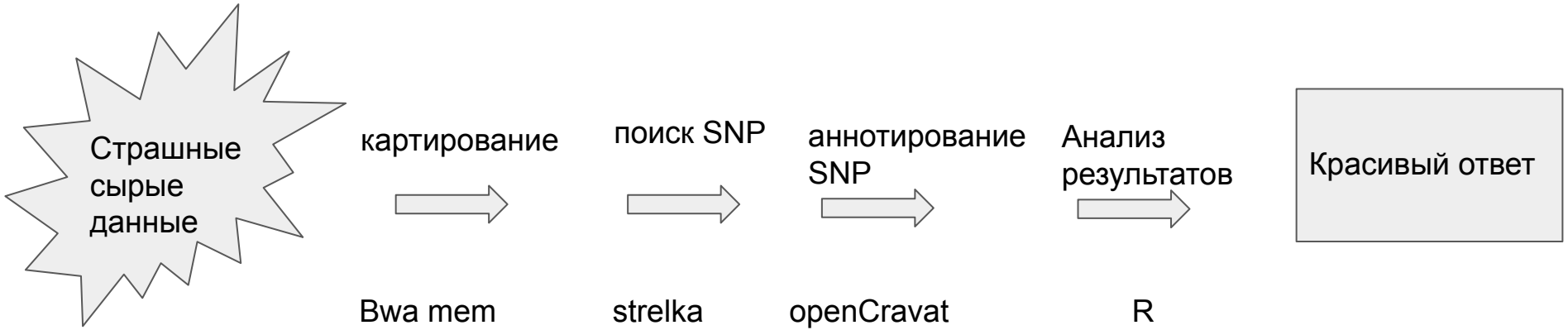


fastq

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Инструменты для реализации общего решения можно подобрать совершенно разные!!

UGENE
Galaxy
….



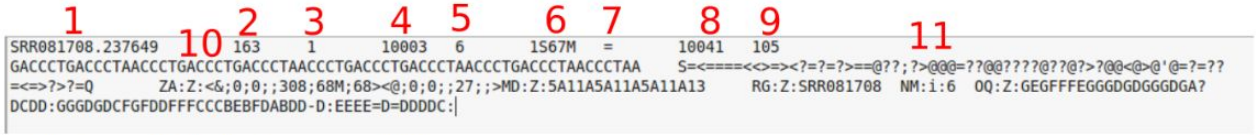
Картирование (bwa mem)

```
#!/bin/bash

while read p; do
        echo $p

        echo "#########################Started mapping#######################"

       bwa mem -R "@RG\tID:4\tSM:$p\tPL:illumina\tLB:lib1\tPU:unit1"    ./Homo_sapiens_assembly38.fasta  "$p"_1.fastq  "$p"_2.fastq -t 25 > "$p".sam

        echo "####################Deleting fastq files###################"
        #rm "$p"_1.fastq
        #rm "$p"_2.fastq

        echo "#################### Converting files to Bam and sort ###################"

        /samtools view -Su "$p".sam | samtools sort -@25 - -o "$p".sorted.bam

        echo "##############Deleting sam files###############"
        rm "$p".sam

        echo "#################mark duplicates################"

        java -Dpicard.useLegacyParser=false  -Xmx16G -jar ../picard.jar  MarkDuplicates\
         -I "$p".sorted.bam\
         -O "$p".MD.bam\
         -METRICS_FILE metrics.txt\
         -CREATE_INDEX true

        echo "#############Deleting sorted.bam################"
        rm "$p".sorted.bam

done < samp_list
```

samp_list

resist_1
resist_2
resist_3
resist_4
resist_5
resist_6
sensitive_1
sensitive_2
sensitive_3
sensitive_4
sensitive_5
sensitive_6

Результат картирования (SAM/BAM)



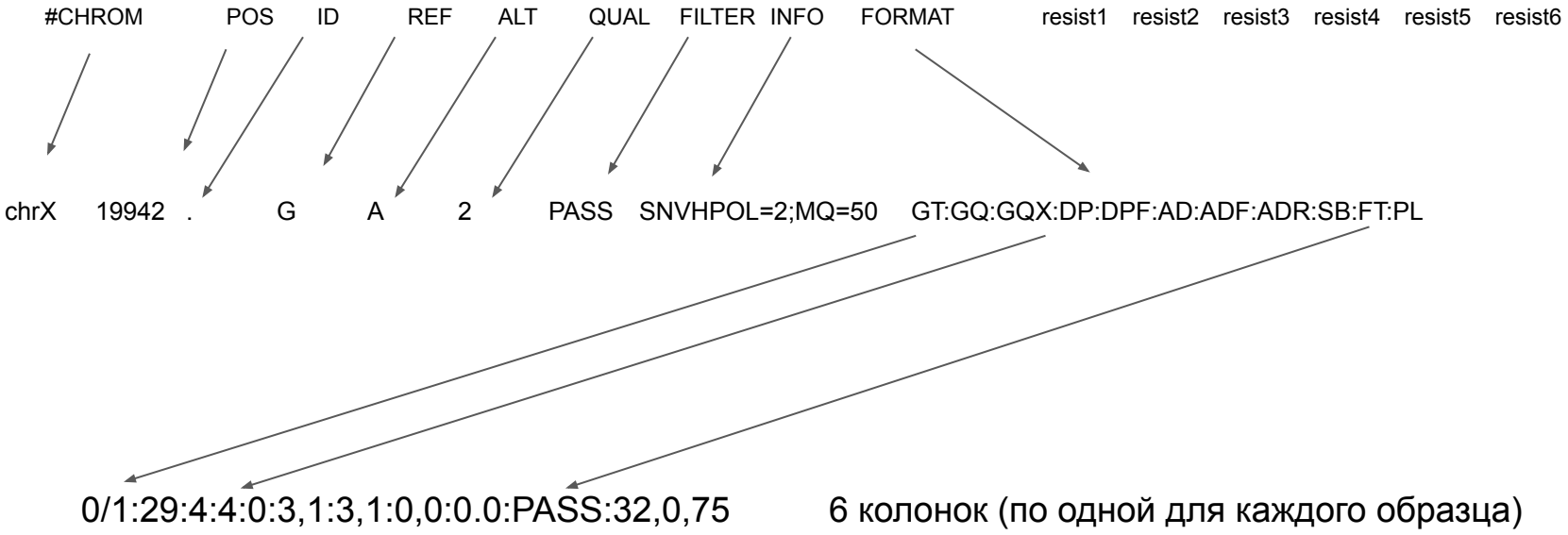
Column	Official Name	Brief
1	QNAME	Sequence ID
2	FLAG	Sequence quality expressed as a bitwise flag
3	RNAME	Chromosome
4	POS	Start Position
5	MAPQ	Mapping Quality
6	CIGAR	Describes positions of matches, insertions, deletions w.r.t reference
7	RNEXT	Ref. name of mate / next read
8	PNEXT	Postion of mate / next read
9	TLEN	Observed Template length
10	SEQ	Sequence
11	QUAL	Base Qualities

https://bioinformatics-core-shared-training.github.io/cruk-summer-school-2017/Day1/Session5-alignedReads.html

Поиск SNP (strelka)

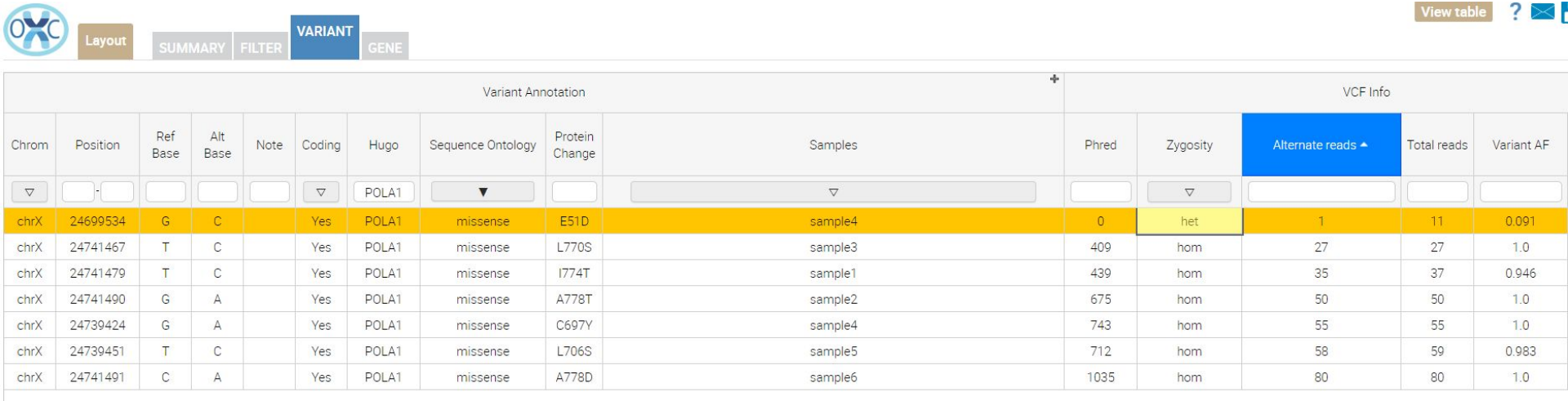
```
#!/bin/bash

# configuration

./strelka-2.9.10.centos6_x86_64/bin/configureStrelkaGermlineWorkflow.py \
--bam ./sensitive_1.MD.bam \
--bam ./sensitive_2.MD.bam \
--bam ./sensitive_3.MD.bam \
--bam ./sensitive_4.MD.bam \
--bam ./sensitive_5.MD.bam \
--bam ./sensitive_6.MD.bam \
--referenceFasta ./Homo_sapiens_assembly38.fasta \
--runDir ./vcf

# execution on a single local machine with 20 parallel jobs
./vcf/runWorkflow.py -m local -j 20
```

Результаты поиска SNP (VCF формат)



Аннотирование SNP (open cravat)

Результат - тот же VCF формат с дополнительными колонками, содержащими аннотации (название гена, эффект мутаций и тд)



Layout | SUMMARY | FILTER | VARIANT | GENE — View table

Variant Annotation										VCF Info				
Chrom	Position	Ref Base	Alt Base	Note	Coding	Hugo	Sequence Ontology	Protein Change	Samples	Phred	Zygosity	Alternate reads ▲	Total reads	Variant AF
▽	-				▽	POLA1	▼		▽		▽			
chrX	24699534	G	C		Yes	POLA1	missense	E51D	sample4	0	het	1	11	0.091
chrX	24741467	T	C		Yes	POLA1	missense	L770S	sample3	409	hom	27	27	1.0
chrX	24741479	T	C		Yes	POLA1	missense	I774T	sample1	439	hom	35	37	0.946
chrX	24741490	G	A		Yes	POLA1	missense	A778T	sample2	675	hom	50	50	1.0
chrX	24739424	G	A		Yes	POLA1	missense	C697Y	sample4	743	hom	55	55	1.0
chrX	24739451	T	C		Yes	POLA1	missense	L706S	sample5	712	hom	58	59	0.983
chrX	24741491	C	A		Yes	POLA1	missense	A778D	sample6	1035	hom	80	80	1.0