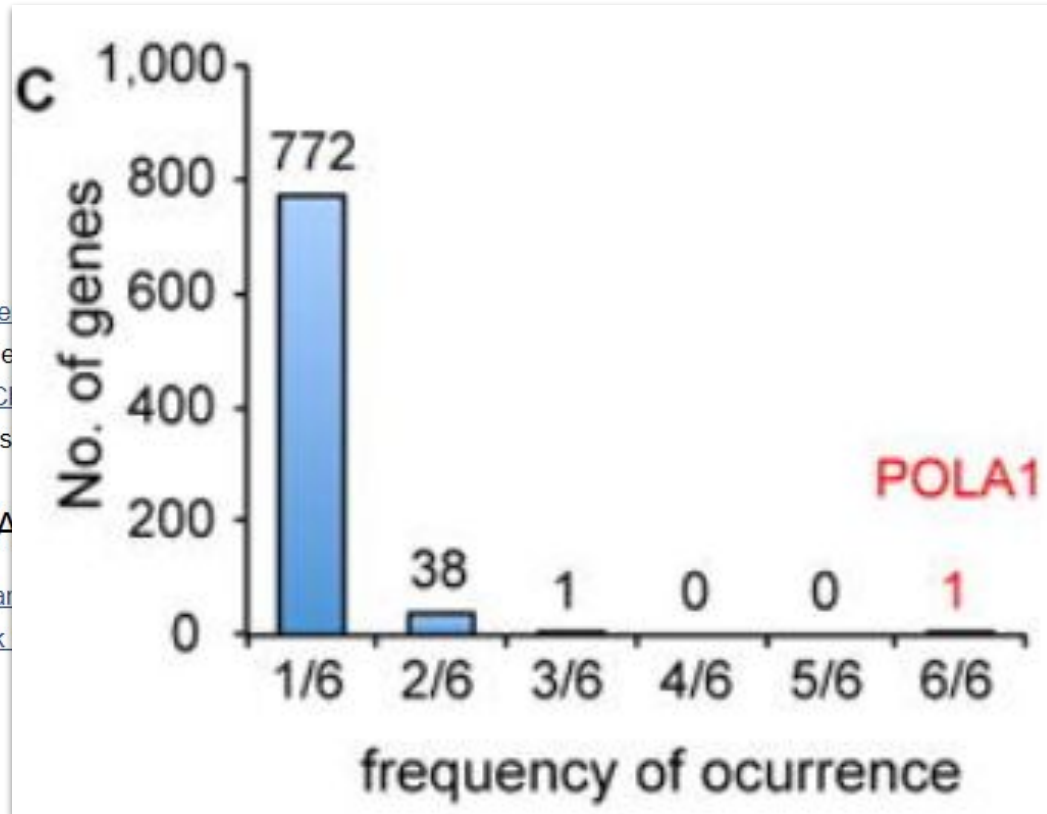


Особенность задания - оно имеет решение :)

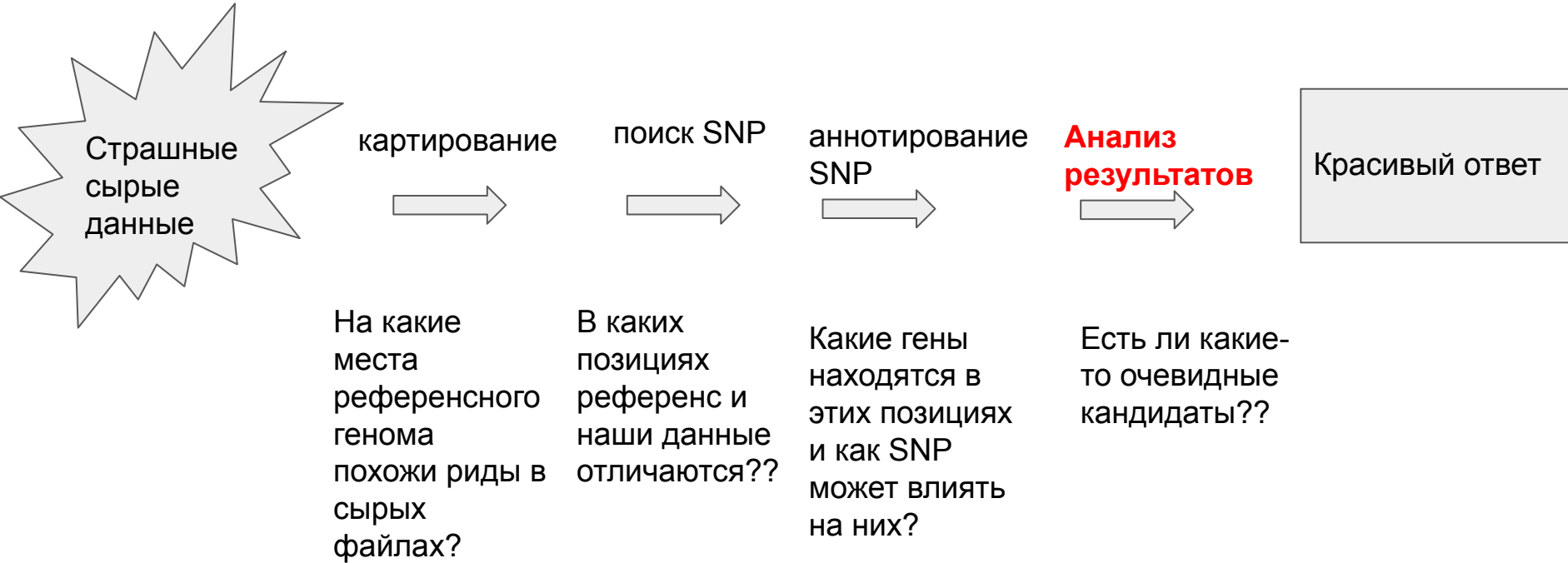


Author Manu

Nat Che
Publishe
Nat C
Publis
The A
Ting_Har
Deepak

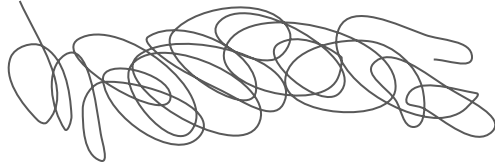
: PMC4912453
NIHMS771697
MID: 27182663
rase α
ind

Общая схема решения



Исходные данные

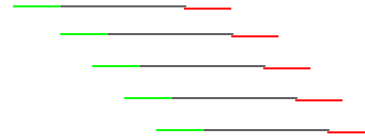
ДНК



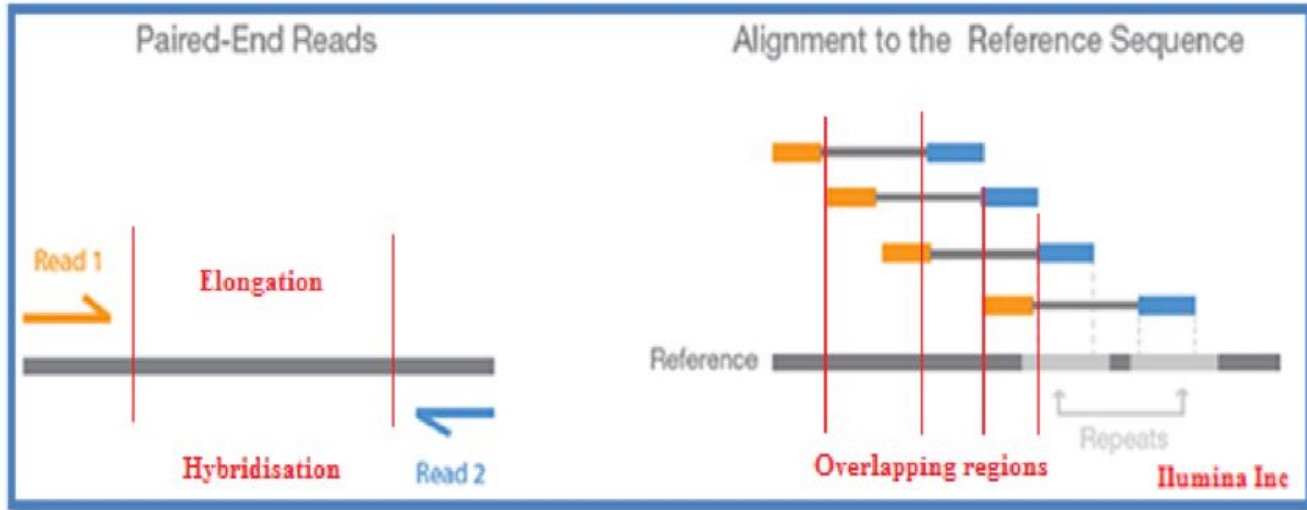
Много кусочков днк



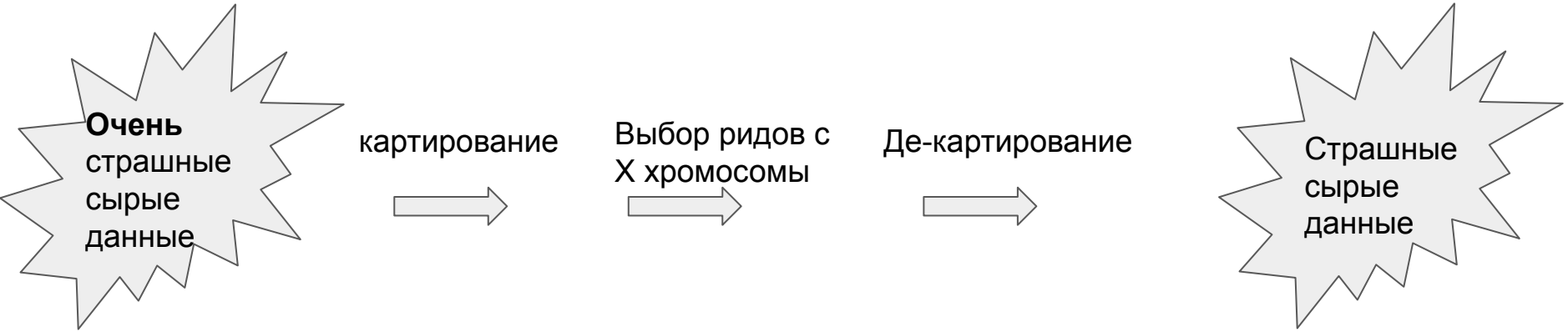
Пришивание адаптеров



Чтение ДНК с обоих адаптеров



Исходные данные - подготовка




fastq

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

Инструменты для реализации общего решения можно подобрать совершенно разные!!

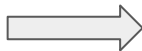
UGENE
Galaxy

....



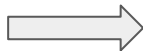
Страшные
сырые
данные

картирование



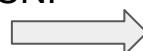
Bwa mem

поиск SNP



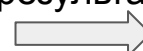
strelka

аннотирование
SNP



openCravat

Анализ
результатов



R



Красивый ответ

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)

```
1      10 2 3 4 5 6 7 8 9      11
SRR081708.237649.163.1.10003.6.1567M.=.10041.105
GACCTGACCCTAACCTGACCTGACCTAACCTGACCTGACCTAACCTGACCTAACCTAA S=<====<>><?=?>==@??;?>@??=?@??@??@??@?>?@<@'@=?=?
=<>?>?=@ ZA:Z:<@;0;0;308;68M;68><@;0;0;27;;>MD:Z:5A11A5A11A5A11A13 RG:Z:SRR081708 NM:i:6 OQ:Z:GEGFFEGGGDGGGGDGA?
DCDD:GGGDGDCFGFDDFFCCCBEBFDABDD-D:EEEE=D=DDDDC:|
```

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	Position of mate / next read
9	TLEN	Observed Template length
10	SEQ	Sequence
11	QUAL	Base Qualities

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)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	resist1	resist2	resist3	resist4	resist5	resist6
chrX	19942	.	G	A	2	PASS	SNVHPOL=2;MQ=50	GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL						
0/1:29:4:4:0:3,1:3,1:0,0:0.0:PASS:32,0,75									6 колонок (по одной для каждого образца)					

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	Zygoty	Alternate reads	Total reads	Variant AF
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