基因组重测序分析

生信帮
云南农业大学
云南省生物大数据重点实验室
2023年8月20日

目录

基因组重测序介绍

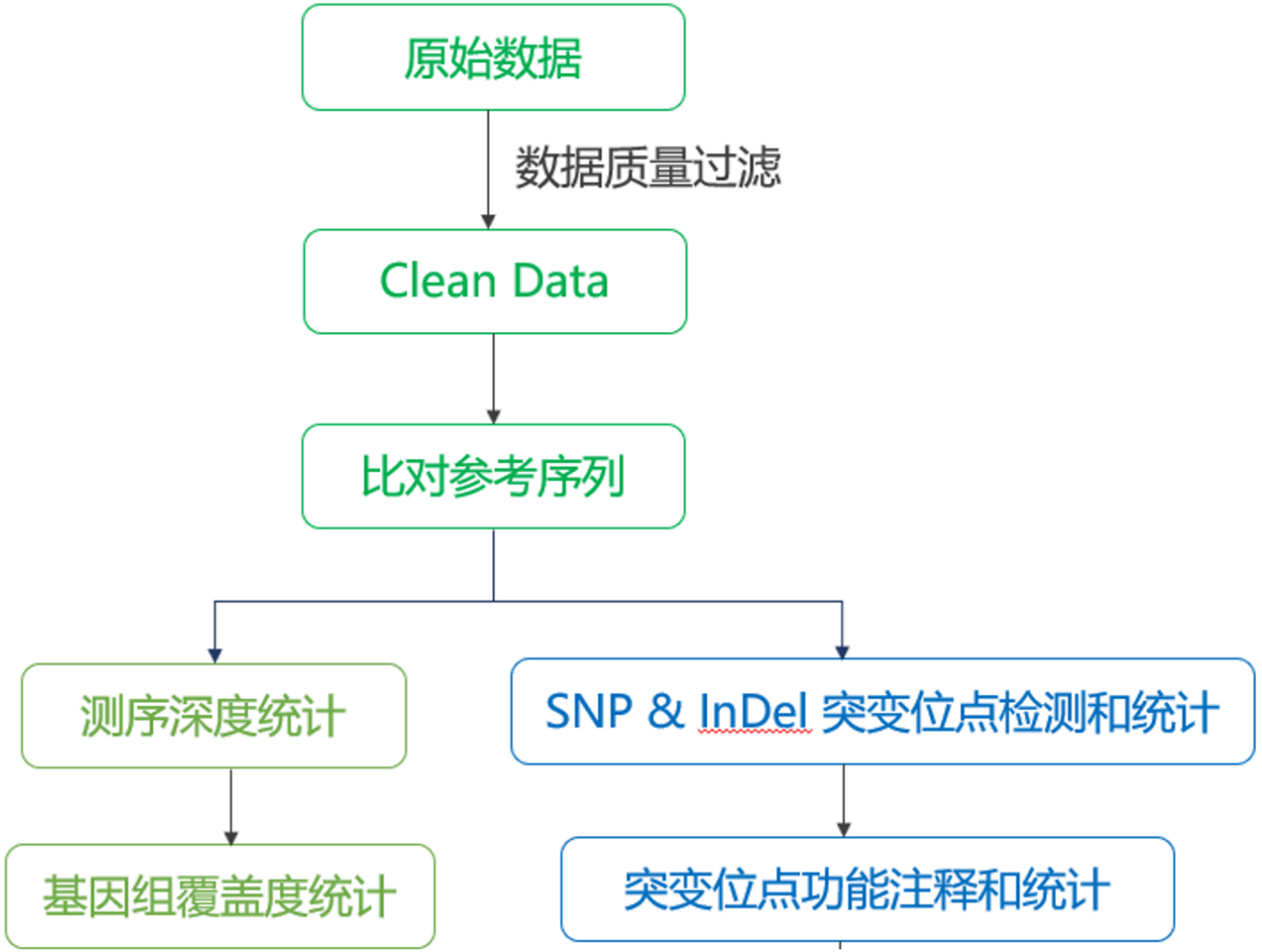
全基因组测序分为**从头测序**(de novo sequencing)**和重测序**(re-sequencing)。

从头测序(de novo)不需要任何参考基因组信息即可对某个物种的基因组进行测序，利用生物信息学分析方法进行拼接、组装，获得该物种的基因组序列图谱，从而推进该物种的后续研究。

基因组重测序 是对有参考基因组物种的不同个体进行的基因组测序，并在此基础上对个体或群体进行相关分析。

基因组重测序主要通过将序列比对到参考基因组上，通过检测发现**单核苷酸多态性位点**(SNPs)、**拷贝数变异**(CNV)、**插入/缺失**(Indel)等变异类型，以较低的价格将单个参考基因组信息扩增为生物群体的遗传特征。全基因组重测序在人类疾病和动植物育种研究中广泛应用。

基因组重测序数据分析


原始数据
数据质量过滤
Clean Data
比对参考序列
测序深度统计
基因组覆盖度统计
SNP & InDel 突变位点检测和统计
突变位点功能注释和统计

测序数据比对参考基因组

重测序下机数据一般为双端150bp的序列，数据格式为fastq格式；原始测序数据需要先进行数据质控，获得高质量的干净的测序数据，再进行后续分析；

首先需要将质控后的clean序列比对到参考基因组上，通过序列比对，可以找到大量的单核苷酸多态性（SNP），插入缺失（InDel，Insertion/Deletion）和结构变异（SV，Structure Variation）位点。

测序数据比对参考基因组

比对软件：BWA

1. 简介

BWA (Burrows-Wheeler Aligner)是一个用于DNA序列比对的快速、高效、可扩展的软件工具。它可以将几GB的序列数据集与参考序列进行比对。BWA的比对速度和准确性都非常高，被广泛应用于基因组学、转录组学、表观遗传学等领域的研究中。

BWA的主要特点包括：

1. 高效快速：BWA采用了BWT算法和一些优化策略，可以在较短的时间内完成大规模的比对任务。
2. 准确性高：BWA采用了一些质量控制策略，如配对信息、序列长度等，可以提高比对的准确性。
3. 多功能：BWA支持不同类型的序列比对，包括单端比对、双端比对、重测序、局部比对等。
4. 可扩展性强：BWA可以处理多个比对任务，支持多线程和分布式计算，能够高效地处理大规模的数据集。

测序数据比对参考基因组

bwa有三个不同的算法：

BWA-backtrack: 是用来比对 Illumina 的序列的，reads 长度最长能到 100bp；

BWA-SW: 用于比对 long-read ，支持的长度为 70bp-1Mbp；同时支持剪接性比对。；

BWA-MEM: 都支持较长的read长度，同时都支持剪接性比对（split alignments)，但是BWA-MEM是更新的算法，也更快，更准确，且 BWA-MEM 对于 70bp-100bp 的 Illumina 数据来说，效果也更好些。

测序数据比对参考基因组

2. 用法

（1）对参考基因组建立索引

命令示例：`bwa index genome.fa`

（2）比对参考基因组

命令示例：`bwa mem genome.fa sampleA.R1.fq sampleA.R2.fq > sampleA.sam`

输出结果处理：用samtools对比对结果进行排序并转换成bam格式

sam文件格式详解

sam格式用来存储测序reads与参考序列比对结果信息的一种文件格式。

比对情况展示：

```
Coor      12345678901234  5678901234567890123456789012345
ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                           ttagctTAGGC
-r001/2                                      CAGCGGCAT
```

sam文件格式详解

云南省生物大数据重点实验室
生信帮

sam格式文件内容：

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
+r001/1  99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
 +r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
 +r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
 +r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
 -r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
-r001/2 147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

1 2 3 4 5 6 7 8 9 10 11 12

sam文件格式详解

sam格式：header表头信息

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
```

@HD，格式的版本、比对序列的排列顺序
@SQ，参考序列说明 （SN：ref name, LN 是参考序列的长度）
@PG，使用的比对程序说明

https://www.samformat.info/sam-format-header

sam文件格式详解

sam格式：flag含义

flag：位标识，每一个数字代表一种比对情况，这里的值是符合情况的数字相加总和。

0 单端测序序列(SE)
1 (0x1) read paired read是pair中的一条（read表示本条read，mate表示pair中的另一条read）
2 (0x2) read mapped in proper pair pair一正一负完美的比对上
4 (0x4) read unmapped 这条read没有比对上
8 (0x8) mate unmapped mate没有比对上
16 (0x10) read reverse strand 这条read反向比对
32 (0x20) mate reverse strand mate反向比对
64 (0x40) first in pair 这条read是read1
128 (0x80) second in pair 这条read是read2
256 (0x100) not primary alignment 第二次比对
512 (0x200) read fails platform/vendor quality checks 比对质量不合格
1024 (0x400)read is PCR or optical duplicate read是PCR或光学副本产生
2048 (0x800)supplementary alignment 辅助比对结果

查询flag值的网站：https://broadinstitute.github.io/picard/explain-flags.html

sam文件格式详解

sam格式：CIGAR

简要比对信息表达式（Compact Idiosyncratic Gapped Alignment Report），其以参考序列为基础，使用数字加字母表示比对结果。

op	Description
M	Alignment match (can be a sequence match or mismatch
I	Insertion to the reference
D	Deletion from the reference
N	Skipped region from the reference
S	Soft clip on the read (clipped sequence present in <seq>)
H	Hard clip on the read (clipped sequence NOT present in <seq>)
P	Padding (silent deletion from the padded reference sequence)

sam文件格式详解

sam格式：第12列

SAM格式的第12列是一些可选字段，如下：

AS:i 匹配的得分

XS:i 第二好的匹配的得分

YS:i mate 序列匹配的得分

XN:i 在参考序列上模糊碱基的个数

XM:i 错配的个数

XO:i gap open的个数

XG:i gap 延伸的个数

NM:i 经过编辑的序列

YF:i 说明为什么这个序列被过滤的字符串

YT:Z

MD:Z 代表序列和参考序列错配的字符串

https://www.cnblogs.com/emanlee/p/5366610.html

sam文件格式详解

雲南農業大學 Yunnan Agricultural University
云南省生物大数据重点实验室


```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001    99 ref  7 30 8M2I4M1D3M = 37   39 TTAGATAAAGGATACTG *
r002     0 ref  9 30 3S6M1P1I4M * 0    0 AAAAGATAAGGATA     *
r003     0 ref  9 30 5S6M       * 0    0 GCCTAAGCTAA        * SA:Z:ref,29,-,6H5M,17,0;
r004     0 ref 16 30 6M14N5M    * 0    0 ATAGCTTCAGC        *
r003  2064 ref 29 17 6H5M       * 0    0 TAGGC              * SA:Z:ref,9,+,5S6M,30,1;
r001   147 ref 37 30 9M         = 7  -39 CAGCGGCAT          * NM:i:1
```

Header section

Alignment section

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

sam文件格式详解

不同格式所占存储大小

SAM	6.7G
BAM	1.1G
CRAM	0.5G

samtools软件简介

view

-b

-h

-H

-S

例子:

1) 将sam文件转换成bam文件

```
$ samtools view -bS abc.sam > abc.bam
$ samtools view -b -S abc.sam -o abc.bam
```

2) 提取比对到参考序列 1 的比对结果

```
$ samtools view -bF 4 abc.bam > abc.F.bam
```

3)提取paired reads中两条reads都比对到参考序列上的比对结果，只需要把两个4+8的值12作为过滤参数即可

```
$ samtools view -bF 12 abc.bam > abc.F12.bam
```

4) 提取没有比对到参考序列上的比对结果

```
$ samtools view -bf 4 abc.bam > abc.f.bam
```

5) 提取bam文件中比对到caffold1上的比对结果，并保存到sam文件格式

```
$ samtools view abc.bam scaffold1 > scaffold1.sam
```

6) 提取scaffold1上能比对到30k到100k区域的比对结果

```
$ samtools view abc.bam scaffold1:30000-100000 > scaffold1_30k-100k.sam
```

7) 根据fasta文件，将 header 加入到 sam 或 bam 文件中

```
$ samtools view -T genome.fasta -h scaffold1.sam > scaffold1.h.sam
```

samtools软件简介

Sort：对bam文件进行排序

tview：直观显示比对情况

merge：合并多个bam文件

flagstat：给出bam文件比对结果

index：对bam产生.bai的索引文件

depth：每个碱基测序深度

faidx：对fasta产生.fai的文件

……

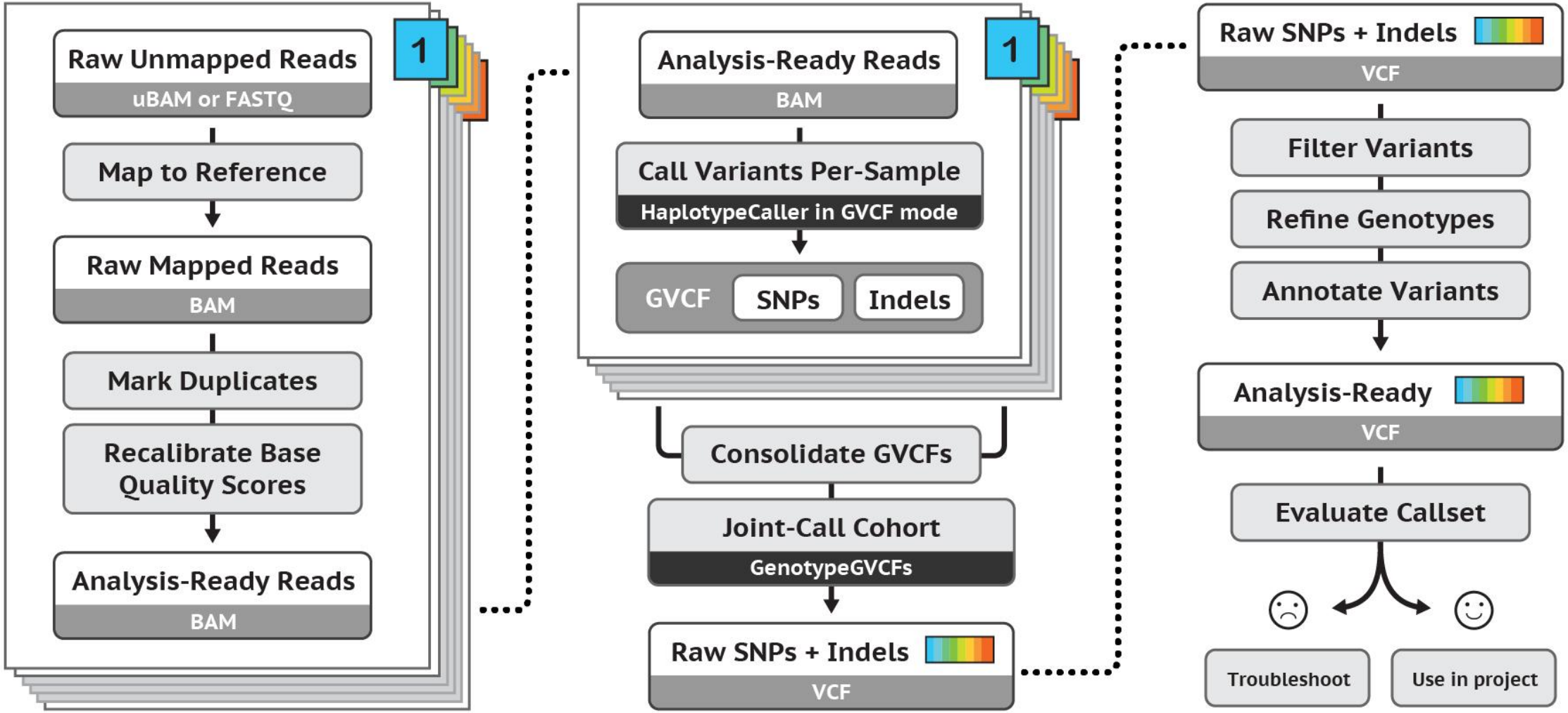
GATK变异位点检测

GATK是Genome Analysis Toolkit的缩写，是用来处理高通量测序数据的一套软件。最初，GATK被设计用来分析人类基因组和外显子，主要用来寻找SNP和indel。后开，GATK的功能越来越丰富，增加了short variant calling、计算copy number（CNV）和结构变异（SV）等新功能。同时，GATK也越来越广泛地应用于其他物种的数据分析中。现在，GATK已经成为了基因组和转录组分析过程中，寻找变异的行业标准。

https://github.com/broadinstitute/gatk

GATK变异位点检测

GATK分析pipeline



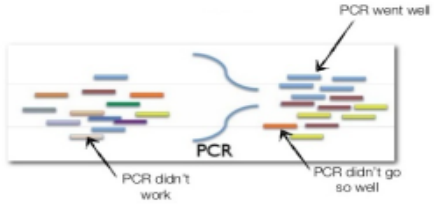
GATK变异位点检测

云南省生物大数据重点实验室
生信帮

第一步：标记重复（Duplicates Marking）

How do duplication events arise?

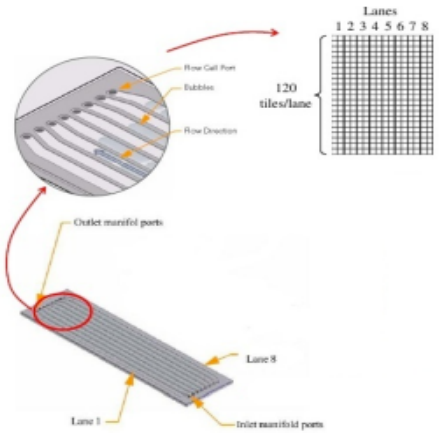
PCR duplicates



Optical duplicates

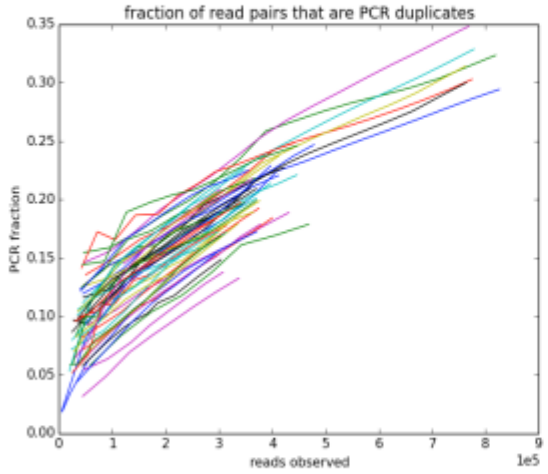
Read names have the following form:

```
@identifier:lane:tile:x:y
```

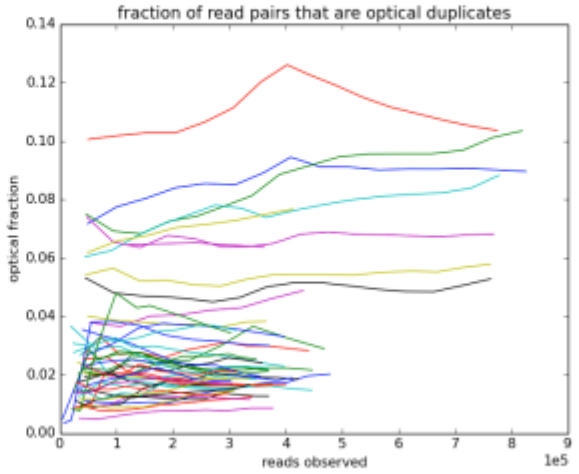


Optical and PCR duplication events arise at different rates as a sequencing experiment proceeds

PCR duplicates



Optical duplicates



GATK变异位点检测

第一步：标记重复（Duplicates Marking）

在制备文库的过程中，由于PCR扩增过程中会存在一些偏差，也就是说有的序列会被过量扩增。这样，在比对的时候，这些过量扩增出来的完全相同的序列就会比对到基因组的相同位置。而这些过量扩增的reads并不是基因组自身固有序列，不能作为变异检测的证据，因此，要尽量去除这些由PCR扩增所形成的duplicates。

重复序列产生的途径，有以下两种：

1. PCR duplicates，PCR根据一份模板，扩增出多份拷贝，来源于同一模板的多份拷贝之间就是PCR重复序列；

2. Optical duplicates，illumina测序仪的基本单位是flowcell，测序反应在flowcell上发生和进行，高密度的flowcell使得测序的通量显著提升，也带来了序列重复读取的问题。虽然比例非常低，但是也需要考虑进来。

去重复的过程是给这些序列设置一个flag以标志它们，方便GATK的识别。还可以设置REMOVE_DUPLICATES=true 来丢弃duplicated序列。对于是否选择标记或者删除，对结果应该没有什么影响，GATK官方流程里面给出的例子是仅做标记不删除。**最主要目的就是尽量减小文库构建时引入文库的PCR bias**

使用的工具模块：MarkDuplicates

GATK变异位点检测

第二步：变异检测（生成单样本的gvcf文件）

HaplotypeCaller 是 GATK 检测变异（SNP/INDEL）的核心模块，主要通过单倍型的局部重组来实现准确的 SNP 和 INDEL 检测。它能够处理非二倍体生物以及混合的实验数据，还能够正确处理可变剪切，因此可以适用RNAseq数据。

GATK-HaplotypeCaller 模块进行 SNP/indel 检测的基本工作流程包含四个主要步骤：

1) 识别活跃区域

2) 通过重组装活跃区域确定单体型

3) 确定每个read的单倍型的似然值

4) 确定基因型。

使用的工具模块： HaplotypeCaller

GATK变异位点检测

第三步：变异检测（多个样本的gvcf合并）

两个工具模块：

- GenomicsDBImport：新方法，速度快，但目前一次只能处理一条染色体；
- CombineGVCFs：旧方法，速度慢，但是可以一次全部合并；

第四步：变异检测（ GenotypeGVCFs ）

多个样本joint genotyping用模块GenotypeGVCFs, 目前GenotypeGVCFs只支持以下三种形式的输入文件：

1）a single single-sample GVCF

2）a single multi-sample GVCF created by CombineGVCFs

3）a GenomicsDB workspace created by GenomicsDBImport.

GATK变异位点质控

1. 提取不同的变异类型（ SelectVariants ）

使用GATK的SelectVariants模块将snp和indel两种变异类型分开为不同的文件，便于后续分析。

2. 位点过滤（ VariantFiltration）

质控是指通过一定的标准，最大可能地剔除假阳性的检测结果，并尽可能地保留最多的正确数据。

首选的质控方案是GATK VQSR， 原理是利用自身数据和已知变异位点集的overlap，通过GMM模型构建一个分类器来对变异数据进行打分，从而评估每个位点的可行度。

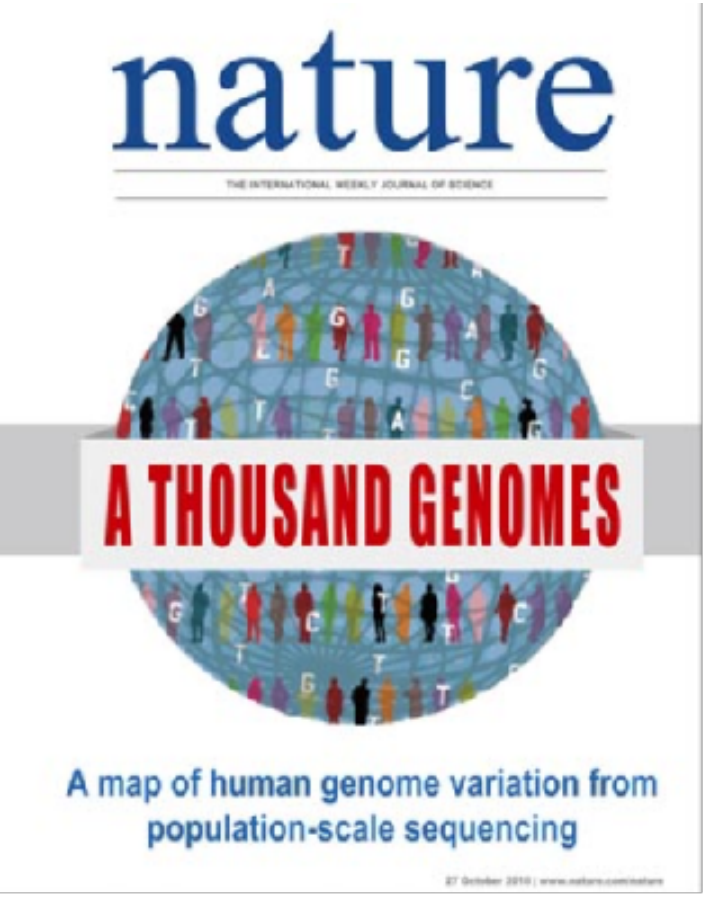
两个前提条件：

1. 已知变异集（Hapmap、OMNI，1000G和dbsnp等这些国际性项目的数据可以作为高质量的已知变异集）；
2. 新检测结果中要有足够多到变异位点。

不满足这两个条件就只能用硬过滤方法进行过滤了（通过人为设定一个或者若干个指标阈值，如QUAL，然后把所有不满足阈值的变异位点直接过滤掉的方法）

Vcf格式详解

1000 Genomes Project



检测到的突变信息如何保存?

VCF (Variant Call Format)

http://samtools.github.io/hts-specs/VCFv4.3.pdf

Vcf格式详解

雲南農業大學 Yunnan Agricultural University
云南省生物大数据重点实验室

Infomation

Header

Variants

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
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 1/1:43:5:.,.
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   0/0:41:3
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   2/2:35:4
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 0/0:61:2
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       1/1:40:3
```

Vcf格式详解

雲南農業大學 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮

#CHROM	POS	ID	REF	ALT	QUAL	FILTER
chr1	873762	.	T	G	5231.78	PASS
chr1	877664	rs3828047	A	G	3931.66	Filter
chr1	888888	.	G	T,C	2120.39	PASS

INFO

AC=1;AF=0.25;AN=4;DP=398;……
AC=2;AF=0.50;AN=4;DP=185;……
AC=1,2;AF=0.25,0.50;AN=4;DP=58;……

FORMAT	S1	S2
GT:AD:DP:GQ:PL	0/0:173,141:282:99:255,0,255	0/1:45,38:83:99:255,0,255
GT:AD:DP:GQ:PL	1/1:0,105:94:99:255,255,0	0/0:80,0:80:99:255,255,0
GT:AD:DP:GQ:PL	0/2:15,0,6:21:99:143,189,670,0,481,463	1/2:25,12,0:37:99:330,0,800,405,836,1241

Vcf格式详解

Ref : AAGCTAGCTACGATGCTAGCATCGATCGTACGATCGATCGAT

S1 : AACCTAGCAACGATGCTGGCATCGATCGTACGATCGATCGAT
AAGCTAGCAACGATGCTCGCATCGATCGTACGATCGATCGAT

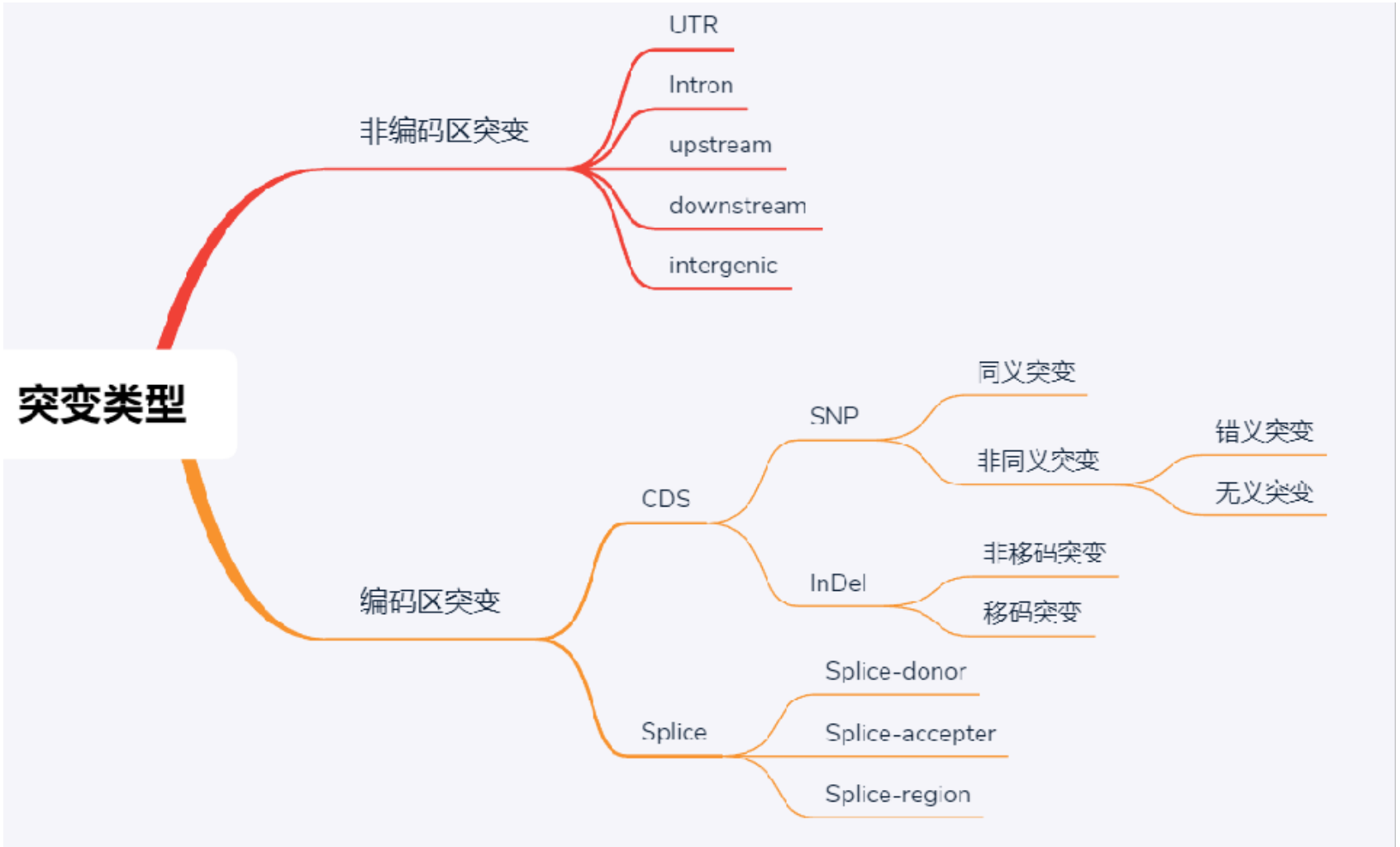
GT (genotype)	0/1	1/1	1/2
AD (allele depth)	15,23	0,39	0,10,20
			GT 0,GT 1, GT 2
DP (depth)	38	39	30

AC(allele count),AF (allele frequency),AN (allele total number) 都是根据 GT 统计的

变异位点注释统计

通常在获得SNP或INDEL变异位点后，我们肯定想知道这些突变位点在基因组上出现的位置、突变的类型以及危害程度。常用的注释软件有Annovar、snpEff等，本次培训详细讲解snpEff 注释软件。

注释分析的前提：基因组有对应的基因注释文件（即gff或gtf文件）。



变异位点注释统计

注释结果文件snp.eff.summary.html解读：

1. 概要

从上往下依次是：基因组（物种名）、注释日期、注释命令、警告信息、错误信息、输入文件行数、变异位点数（过滤之前）、非变异位点数（与参考基因组碱基一致）、变异位点数（过滤之后）、具有ID的变异位点数、非双等位基因组SNP位点数、Number of effects、参考基因组总长度、参考基因组有效长度、变异率（参考基因组有效长度/变异位点数）

Summary

Genome	tair10
Date	2023-08-19 13:53
SnpEff version	SnpEff 4.3t (build 2017-11-24 10:18), by Pablo Cingolani
Command line arguments	SnpEff -stats result/3.annotation/snp.eff.summary.html tair10 result/2.gatk/filter/snp.filter.vcf.gz
Warnings	8,010
Errors	0
Number of lines (input file)	16,107
Number of variants (before filter)	16,235
Number of not variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	16,180
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	128
Number of effects	37,423
Genome total length	119,668,634
Genome effective length	119,668,634
Variant rate	1 variant every 7,396 bases

变异位点注释统计

注释结果文件snp.eff.summary.html解读：

2. 每条染色体上的变异率

从左往右：染色体编号、长度、变异位点数、变异率（多少个碱基中有一个变异位点）

Variants rate details

Chromosome	Length	Variants	Variants rate
1	30,427,671	3,472	8,763
2	19,698,289	2,074	9,497
3	23,459,830	4,141	5,665
4	18,585,056	3,186	5,833
5	26,975,502	3,229	8,354
MT	367,808	77	4,776
Pltd	154,478	1	154,478
Total	**119,668,634**	**16,180**	**7,396**

变异位点注释统计

注释结果文件snp.eff.summary.html解读：

3. 变异类型

包括：SNP（单核苷酸多态性）、MNP（多核苷酸多态性）、INS（插入变异）、DEL（缺失变异）、MIXED（混合变异）、INV(倒位变异）、DUP（重复变异）、BED（易位变异）、INTERVAL（间隔变异）

Number variants by type

Type	Total
SNP	534,659
MNP	0
INS	0
DEL	0
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	534,659

变异位点注释统计

注释结果文件snp.eff.summary.html解读：

4. 有效影响数量

影响大小：HIGH>MODERATE>LOW>MODIFIER

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	362	0.967%
LOW	2,124	5.676%
MODERATE	3,739	9.991%
MODIFIER	31,198	83.366%

变异位点注释统计

注释结果文件snp.eff.summary.html解读：

5. 功能分级有效数

MiSSENSE（错义突变）、NONSENSE（无义突变）、SILENT（沉默突变）

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	4,180	44.783%
NONSENSE	122	1.307%
SILENT	5,032	53.91%

Missense / Silent ratio: 0.8307

变异位点注释统计

注释结果文件snp.eff.summary.html解读：

6. 按类型和区域划分的有效变异数和百分比

第一张图左边为按类型划分有效变异数，包括（从上往下）：3’ 端UTR变异、5’ 端UTR提前启动子获得变异、5’ 端UTR变异、下游基因变异、起始密码子编码变异、基因间隔区、内含子变异、剪接受体变异、剪接供体变异、剪接区域变异、起始缺失、起始保留变异、终止获得、终止缺失、终止保留变异、同义变异、上游基因变异。

第一张图右边为按区域划分有效变异数，包括（从上往下）：下游、外显子、间隔区、内含子、剪接位点受体、剪接位点供体、剪接位点区域、上游、3’ UTR区、5’ UTR区。

第二张图按区域划分有效变异数的百分比可视化图。

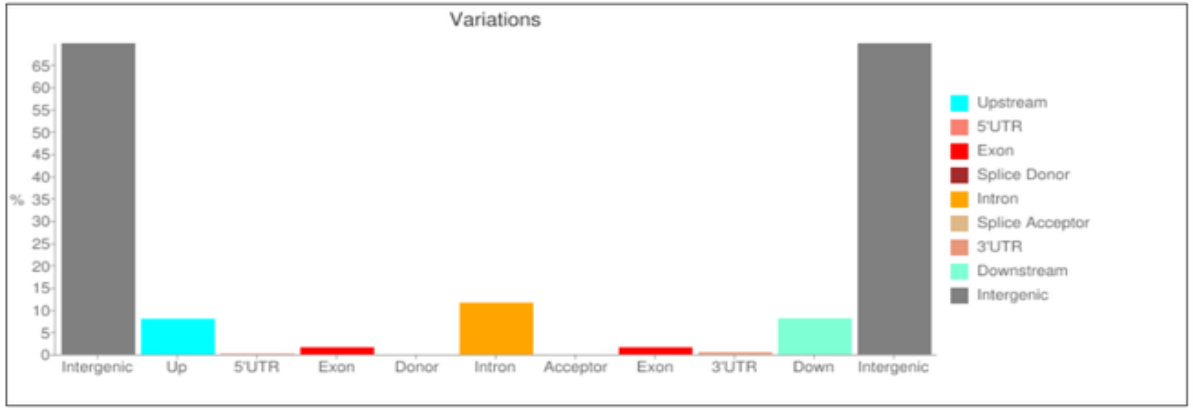
Number of effects by type and region

Type

Type (alphabetical order)	Count	Percent
3_prime_UTR_variant	3,147	0.492%
5_prime_UTR_premature_start_codon_gain_variant	156	0.024%
5_prime_UTR_variant	899	0.141%
downstream_gene_variant	51,944	8.128%
initiator_codon_variant	2	0%
intergenic_region	445,867	69.769%
intron_variant	74,729	11.694%
missense_variant	4,142	0.648%
splice_acceptor_variant	18	0.003%
splice_donor_variant	15	0.002%
splice_region_variant	735	0.115%
start_lost	12	0.002%
start_retained_variant	1	0%
stop_gained	122	0.019%
stop_lost	24	0.004%
stop_retained_variant	6	0.001%
synonymous_variant	6,031	0.944%
upstream_gene_variant	51,208	8.013%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	51,944	8.138%
EXON	10,234	1.603%
INTERGENIC	445,867	69.851%
INTRON	74,160	11.618%
SPLICE_SITE_ACCEPTOR	18	0.003%
SPLICE_SITE_DONOR	15	0.002%
SPLICE_SITE_REGION	659	0.103%
UPSTREAM	51,208	8.022%
UTR_3_PRIME	3,147	0.493%
UTR_5_PRIME	1,055	0.165%



变异位点注释统计

注释结果文件snp.eff.summary.html解读：

7. SNP位点碱基变异表

可以看出SNP中碱基的转换比较多（A腺嘌呤、C胞嘧啶、G鸟嘌呤、T胸腺嘧啶）

Base changes (SNPs)

	A	C	G	T
A	0	23,697	79,492	26,752
C	24,019	0	13,841	99,522
G	99,446	14,162	0	23,729
T	26,822	79,633	23,544	0

变异位点注释统计

注释结果文件snp.eff.summary.html解读：

8. 转换、颠换数量与比值

碱基颠换 (transversion): 是指在碱基置换中嘌呤与嘧啶之间的替代，而转换 (transition)则是一个嘌呤被另一个嘌呤，或者是一个嘧啶被另一个嘧啶替代。

Ts/Tv (transitions / transversions)

Note: Only SNPs are used for this statistic.
Note: This Ts/Tv ratio is a 'raw' ratio (ratio of observed events).

Transitions	8,952,325
Transversions	4,407,771
Ts/Tv ratio	2.031

All variants:

Sample , J19039, J19044, J19061, J19068, ZJ18007, ZJ18008, ZJ18010, ZJ18014, ZJ18015, ZJ18018, ZJ18019, ZJ18021, ZJ18025, ZJ18028, ZJ18031, ZJ18032, ZJ18034, ZJ18037, ZJ18040, ZJ18041, ZJ18042, ZJ19001,
Transitions , 197655, 194070, 244113, 228524, 194593, 215783, 181523, 195040, 216218, 204765, 214264, 187846, 179546, 217641, 203156, 214749, 196403, 170192, 171032, 183944, 209418, 120774, 136141, 173099
Transversions , 97174, 96014, 120538, 113263, 96006, 106206, 89607, 96115, 106252, 100628, 105195, 92309, 88054, 106709, 99998, 106019, 97073, 84259, 84498, 90695, 103030, 59420, 67026, 84956, 93228, 97381,
Ts/Tv , 2.034, 2.021, 2.025, 2.018, 2.027, 2.032, 2.026, 2.029, 2.035, 2.035, 2.037, 2.035, 2.039, 2.040, 2.032, 2.026, 2.023, 2.020, 2.024, 2.028, 2.033, 2.033, 2.031, 2.038, 2.024, 2.043, 2.051, 2.030, 2.032

Only known variants (i.e. the ones having a non-empty ID field):

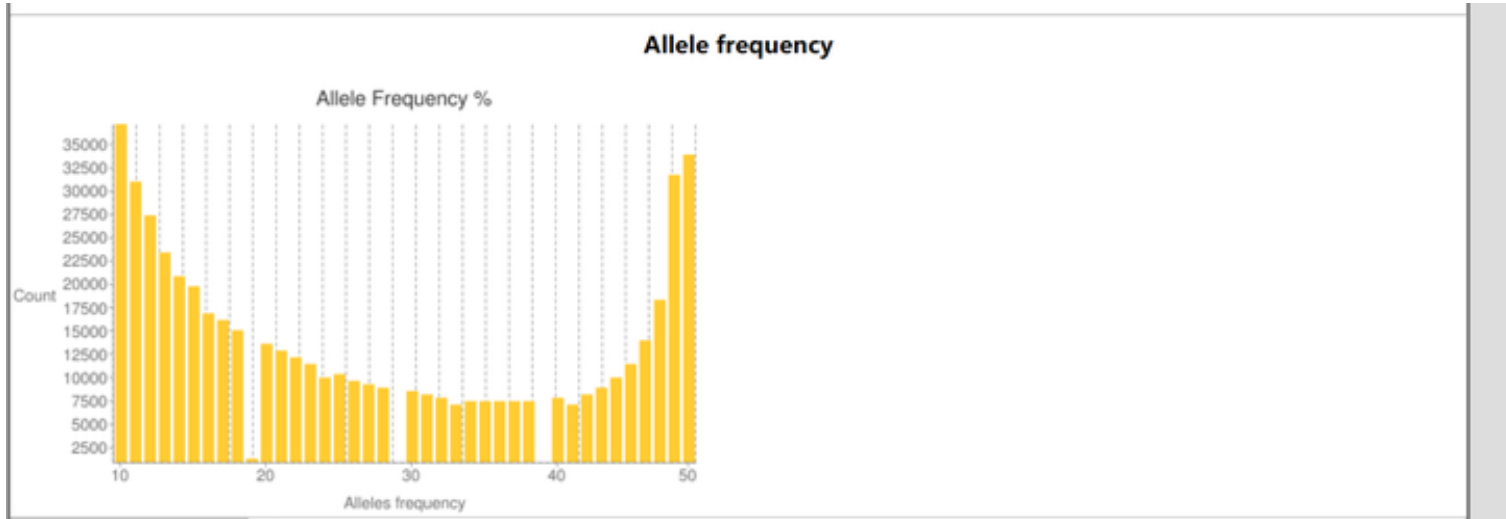
Sample , J19039, J19044, J19061, J19068, ZJ18007, ZJ18008, ZJ18010, ZJ18014, ZJ18015, ZJ18018, ZJ18019, ZJ18021, ZJ18025, ZJ18028, ZJ18031, ZJ18032, ZJ18034, ZJ18037, ZJ18040, ZJ18041, ZJ18042, ZJ19001,
Transitions , 197655, 194070, 244113, 228524, 194593, 215783, 181523, 195040, 216218, 204765, 214264, 187846, 179546, 217641, 203156, 214749, 196403, 170192, 171032, 183944, 209418, 120774, 136141, 173099
Transversions , 97174, 96014, 120538, 113263, 96006, 106206, 89607, 96115, 106252, 100628, 105195, 92309, 88054, 106709, 99998, 106019, 97073, 84259, 84498, 90695, 103030, 59420, 67026, 84956, 93228, 97381,
Ts/Tv , 2.034, 2.021, 2.025, 2.018, 2.027, 2.032, 2.026, 2.029, 2.035, 2.035, 2.037, 2.035, 2.039, 2.040, 2.032, 2.026, 2.023, 2.020, 2.024, 2.028, 2.033, 2.033, 2.031, 2.038, 2.024, 2.043, 2.051, 2.030, 2.032

变异位点注释统计

注释结果文件snp.eff.summary.html解读：

9. 等位基因频率

横坐标为等位基因频率、纵坐标为数量。



Min	10
Max	50
Mean	27.431
Median	24
Standard deviation	14.003
Values	10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,50
Count	37100,31284,27388,23655,20824,19968,16993,16430,15132,1315,13907,13040,12262,11492,10036,10409,9928,9520,8883,890,8710,8420,786

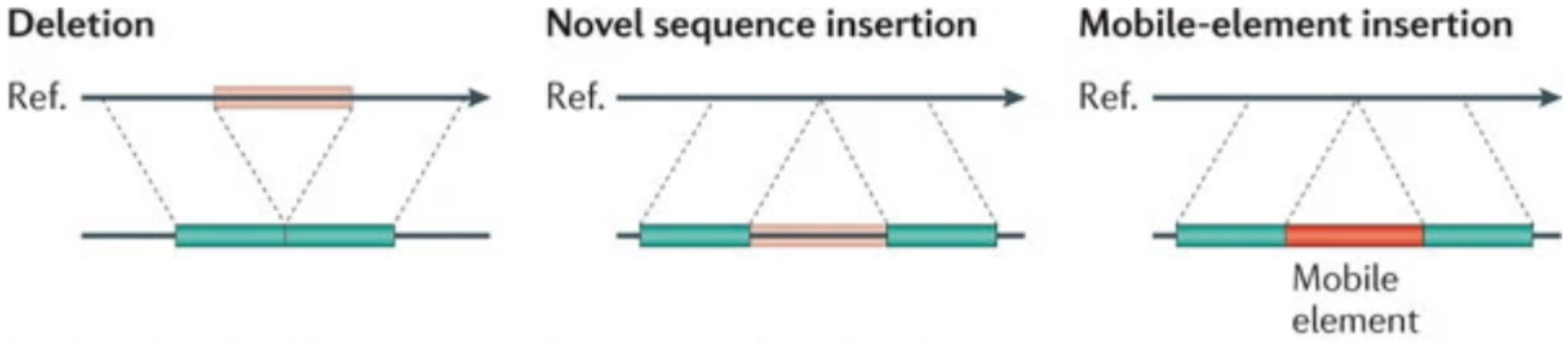
结构变异（SV）检测

结构变异的定义

结构变异（structural variation，SV）是指基因组上长片段（1kb以上的）的序列变化和位置关系变化。包括长片段的插入删除（Big Indel），染色体倒位（Inversion），易位（Translocation），串联重复（Tandem repeat），拷贝数变异（copy number variation, CNV）等，有些标准还包括简单重复序列（SSR），散在的重复（Interspersed duplications）和其他重复序列。

结构变异的类型

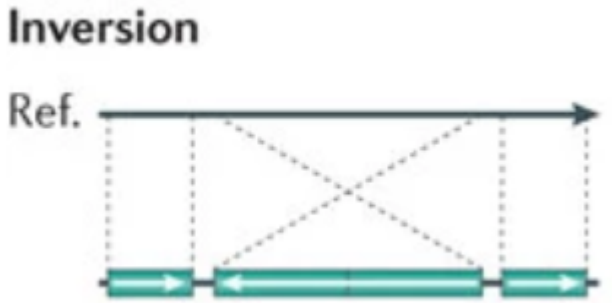
1. 大于50bp的长片段序列插入删除（Big Indel）；



结构变异（SV）检测

结构变异的类型

2. 长度大于1000bp的染色体倒位（Inversion）；

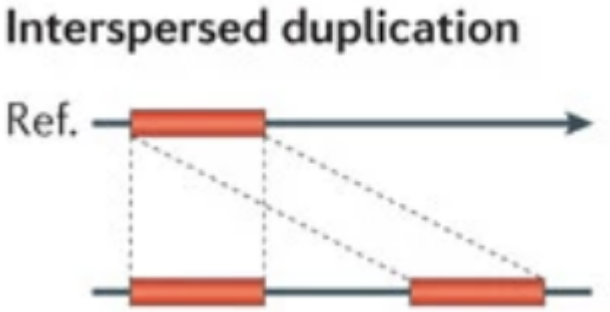


3. 长度大于1000bp的染色体内部或染色体间的序列易位（Translocation）；

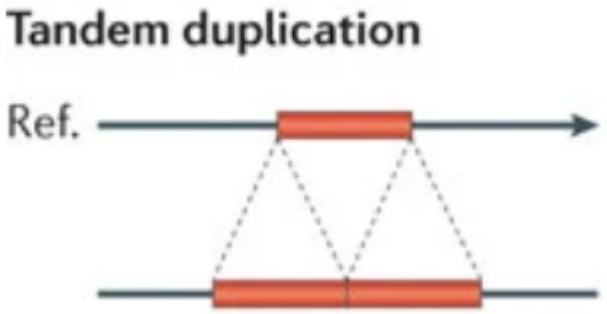
结构变异（SV）检测

结构变异的类型

4. 拷贝数变异（copy number variation, CNV）；



5. 串联重复（Tandem repeat）；



6. 其它更为复杂的嵌合性变异；

结构变异（SV）检测

检测结构变异的方法

1. Read-pair (RP)：通过对双端测序reads的距离（即插入序列长度的分布）或位置关系（即双端reads的正反链情况）的异常值分析来检测结构变异；
2. Split-read (SR)：通过对双端测序reads中一条reads能比对上，另一条不能比对上的情况进行分析来检测结构变异；
3. Read-depth (RD)：利用read的mapping深度来检测基因组拷贝数变异（Copy number variantion，简称CNV）的方法；
4. Assembly (AS)：通过三代测序和de novo assembly来检测大长度和复杂结构的变异；

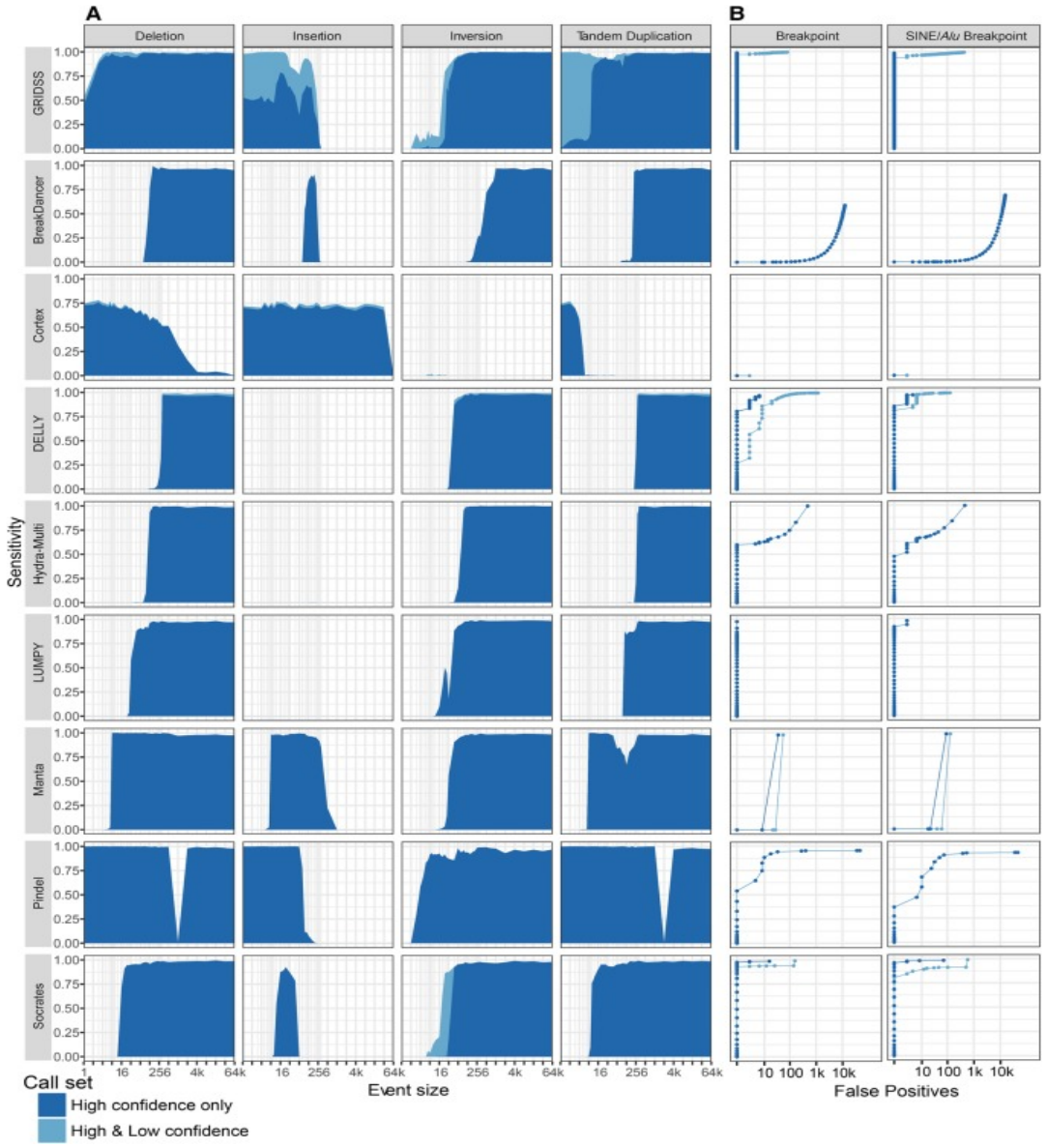
此外还可能利用 SV 范围内 snv 变异辅助确定，比如二倍体 DUP 上的杂合变异 reads 的 ref/alt 倾向于 1/2 或 2/1, 而二倍体 DEL 范围内应该不存在杂合变异。各检测原理都有各自的优势，其中 AS 是相对最为准确的原理。文献综述推荐使用多种检测原理的软件同时进行检测，最后综合各原理的检测结果进行筛选、过滤，以此提高检测灵敏度和准确度。

结构变异（SV）检测

现有公认的 GATK 是检测小变异（snv + indel）的 “最佳实践” ，但是对于结构变异（SV），尽管有相当多软件能进行 SV 检测，但目前业内没有共识的检测软件和流程。

对文献中常用的比较可靠的 SV 检测软件进行比较全面的测试，从结果上看，多款软件都有各自的优势，如 gridss 和 lumpy 检测的断点位置最准确， manta、Socrates 检测的 SV 更全面，测试结果如图：

经验上来说，最常用的SV 检测软件： **lumpy**，manta



结构变异（SV）检测

lumpy是一款基于概率框架检测结构变异(structure variants)的软件, 它根据read-pair, split-read, read-depth和其他先验知识寻找基因组上可能的结构变异。

lumpy基于paired-end reads比对后得到的三类信息来推断SV：

① 局部异常的测序深度；

② 不一致的比对；

③ 断裂的比对(split-read alignment)。

局部异常的测序深度比较容易理解，平均20X测序的地方，如果深度大于50X，意味着存在着拷贝数变异，如果深度程度非常低，可能意味着这里存在 大片段缺失。不一致的联配和断裂的联配能够提供的信息更多，如果基因组一个区域齐刷刷的截断，就意味着这个区域可能存在插入/缺失。当然也有其他可能，当两个read在不同链或者不同染色体时，可能是易位或倒置。

注意事项：

- SV 的假阳性结果特别多，需要根据一定的条件进行过滤
- 断点明确的 SV 可信度更高

生信帮，答你所问

雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室