使用 *GATK* 分析二代测序数据

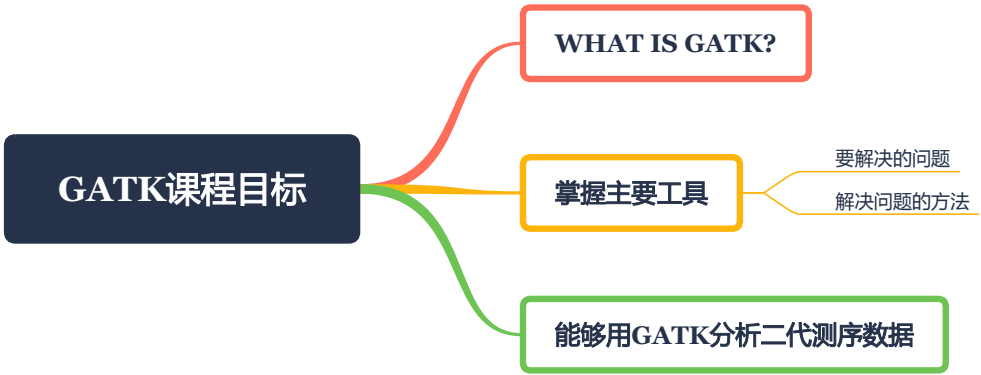
—— *Genome Analysis Toolkit*

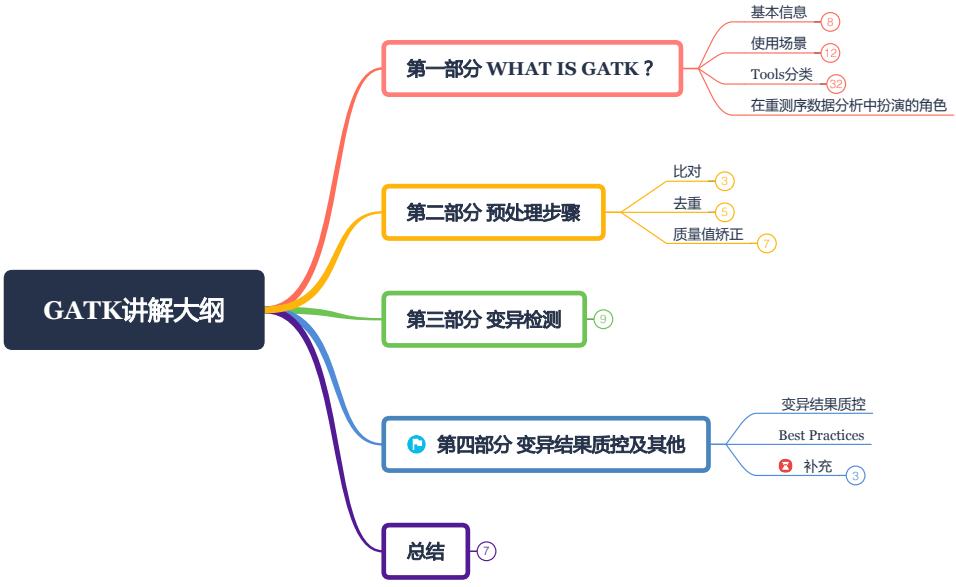
黄志博

Email: huangzhibo@genomics.cn

华大基因大数据中心

2018 年 6 月 14 日


GATK课程目标
WHAT IS GATK?
掌握主要工具
要解决的问题
解决问题的方法
能够用GATK分析二代测序数据


GATK讲解大纲
第一部分 WHAT IS GATK？
基本信息 8
使用场景 12
Tools分类 32
在重测序数据分析中扮演的角色
第二部分 预处理步骤
比对 3
去重 5
质量值矫正 7
第三部分 变异检测 9
第四部分 变异结果质控及其他
变异结果质控
Best Practices
补充 3
总结 7

Part I

What is GATK?

1. 简介
2. 使用场景
3. Tools 分类
   - 按运行环境分类
   - 按功能分类
4. 在 NGS 数据分析中扮演的角色?
   - Best Practices*

简介

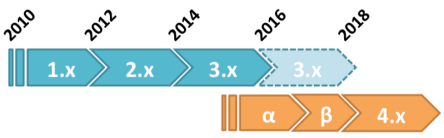
Genome Analysis Toolkit

用于分析高通量测序数据的命令行工具集

简介

Genome Analysis Toolkit

用于分析高通量测序数据的命令行工具集



开发者： Broad Institute

最新版本： GATK4

License： BSD 3-clause

官方网站： https://software.broadinstitute.org/gatk/

使用场景

GATK 可以干什么?

使用场景

GATK 可以干什么?

- 人
  - germline
  - somatic
- 其他物种
- 宏基因组

谁在用 GATK?

GATK 是行业金标准?

功能相似的工具集

- SOAPgaea
- sentitieon
- edico
- ...

GATK 包含多少工具?

GATK 包含多少工具?

>200 个

GATK 包含多少工具?

>200 个

工具太多！我们来分分类...

按运行环境分类

- Non-Spark Tools
- Spark-based Tools

```
Short Variant Discovery:                Tools that perform variant calling and genotyping for short variants (SNPs, SNVs and Indels)
    CombineGVCFs                        Merges one or more HaplotypeCaller GVCF files into a single GVCF with appropriate annotations
    GenomicsDBImport                    Import VCFs to GenomicsDB
    GenotypeGVCFs                       Perform joint genotyping on one or more samples pre-called with HaplotypeCaller
    HaplotypeCaller                     Call germline SNPs and indels via local re-assembly of haplotypes
    HaplotypeCallerSpark                (BETA Tool) HaplotypeCaller on Spark
    Mutect2                             Call somatic SNVs and indels via local assembly of haplotypes
    ReadsPipelineSpark                  (BETA Tool) Takes unaligned or aligned reads and runs BWA (if specified), MarkDuplicates, BQSR,
```

Spark 是什么?

Spark 一种大数据计算框架

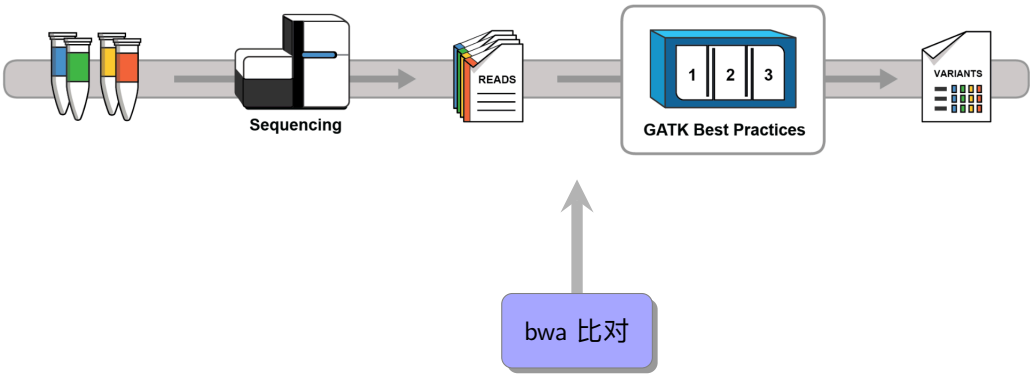
多线程与并行化

多线程 利用一台计算机的多个 CPU 核
并行化 利用多台计算机


按功能分类
Base Calling
生成FastQ的步骤，限于Illumina测序平台
Coverage Analysis
Diagnostics and Quality Control
Intervals Manipulation
对区域文件(bed)进行操作，其他软件bedtools
Reference
对参考序列（FastA）进行操作，做索引...
Read Data Manipulation
对测序数据测序数据（FastQ/BAM）进行操作，其他软件bwa，samtools
Short Variant Discovery
短变异检测
Copy Number Variant Discovery
拷贝数变异检测
Structural Variant Discovery
结构变异检测
Variant Evaluation and Refinement
评估
Variant Filtering
变异结果质控
Variant Manipulation
对VCF进行操作，合并、排序、格式转换...，其他软件bcftools
Metagenomics
宏基因组分析

在 NGS 数据分析中扮演的角色?

在 NGS 数据分析中扮演的角色?

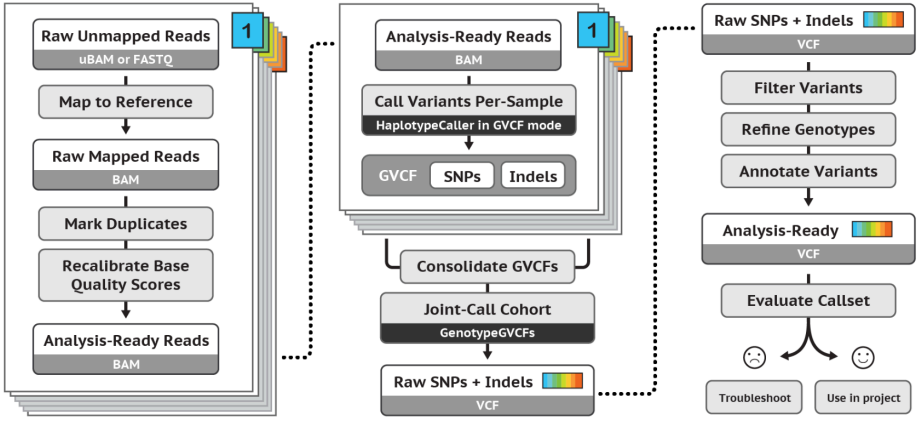


What is GATK Best Practices?

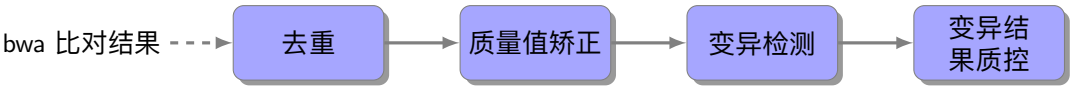
- Germline short variant discovery (SNPs + Indels)
- Somatic short variant discovery (SNVs + Indels)
- (开发中) RNAseq short variant discovery (SNPs + Indels)
- (开发中) Germline copy number variant discovery (CNVs)
- (开发中) Somatic copy number variant discovery (CNVs)

我们要讲的...

Identify germline short variants (SNPs and Indels) in one or more individuals to produce a joint callset in VCF format.



哪些是我们必须掌握的?



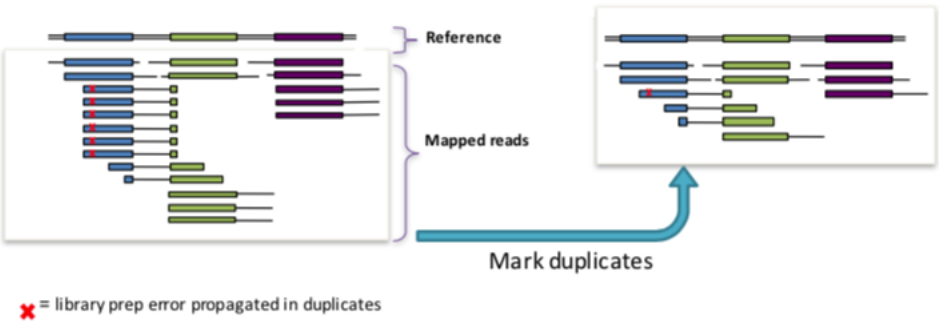
Part II

预处理步骤

5. 去重
   - 什么是重复 reads?
   - dup reads 的来源
   - 判断 dup reads 的方法

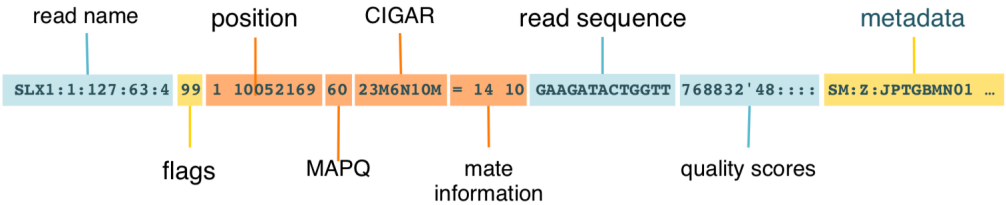
6. 质量值矫正

什么是重复 reads?


Reference
Mapped reads
Mark duplicates
= library prep error propagated in duplicates

标记重复 reads

- 在 BAM 文件的 flags 字段中 Mark dup



bam flags

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

请判断哪条 read 被标记为 dup 了

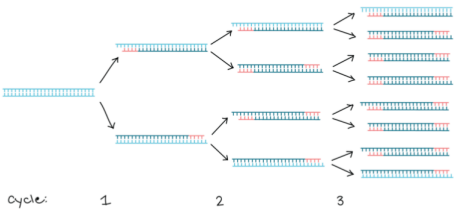
read name	flags	reference name	position	MAPQ	CIGAR	...
r001	2048	chr17	34932982	0	68H32M	...
r002	2064	chr17	37651638	0	35M65H	...
r004	0	chr17	41222824	60	100M	...
r005	1024	chr17	41222824	60	100M	...
r006	16	chr17	41222824	60	100M	...
r008	1024	chr17	41222825	60	100M	...
r009	0	chr17	41222825	60	100M	...
r010	1040	chr17	41222825	60	100M	...

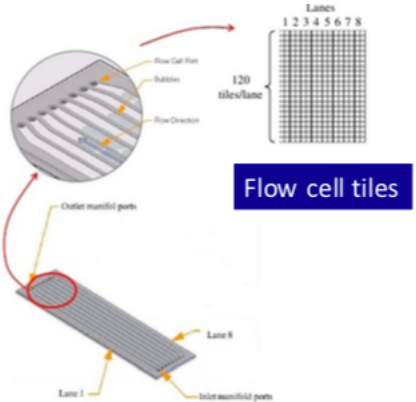
请判断哪条 read 被标记为 dup 了

read name	flags	reference name	position	MAPQ	CIGAR	...
r001	2048	chr17	34932982	0	68H32M	...
r002	2064	chr17	37651638	0	35M65H	...
r004	0	chr17	41222824	60	100M	...
r005	1024	chr17	41222824	60	100M	...
r006	16	chr17	41222824	60	100M	...
r008	1024	chr17	41222825	60	100M	...
r009	0	chr17	41222825	60	100M	...
r010	1040	chr17	41222825	60	100M	...

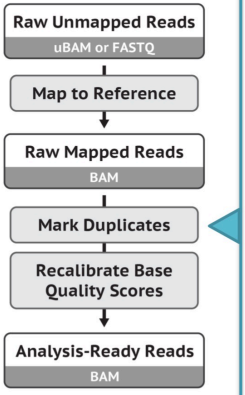
重复 reads 的来源

- 文库 dup
  - 在实验建库阶段，由 PCR 扩增引入的
- 光学 dup
  - 在测序仪检测碱基荧光信号时，由 flow cell tiles 中相邻的簇引起的



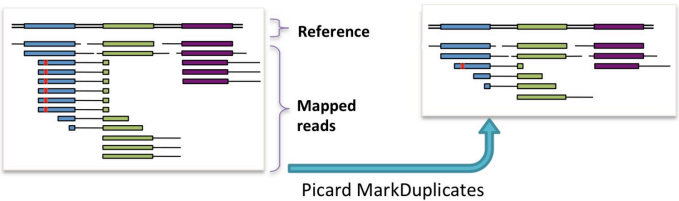


不去重会如何?



Duplicates = **non-independent measurements** of a sequence fragment

-> Must be removed to assess support for alleles correctly



✖ = sequencing error propagated in duplicates

判断 dup reads 的方法

- 相同起始位置
  - 需判断 start position 在哪儿?
  - 对于 paired-end (PE) 数据，两条 reads 有对应相同的 start position
- 同一组 dup reads，保留质量值高的一条，标记其余

start position 必须是 unclipped 5‘端 start position

(可通过 flags + CIGAR 来判断)

- 比对软件有时会剪切（clip）掉末端的碱基
- 反向 reads 在 bam 中的 start position 是 3’ 端，需用 5‘端替代

判断 dup reads 的方法

Pos	1	2	3	4	5	6	7	8	9
Ref	T	A	G	C	C	G	A	T	C
r1	T	A	G	C	C	G	A		
r2	T	A	G	C	C	G	A		
r3	T	A	—	C	CA	G	A		
r4	T	A	G	C	C	H	H		
r5	T	A	G	C	C	G	A	T	C
r6	S	S	G	C	C	G	A		
r7			G	C	C	G	A		

蓝色　比对到正链
橙色　比对到负链
灰色　bases 被剪切
下划线　read 的 5’ 端

判断 dup reads 的方法

Pos	1	2	3	4	5	6	7	8	9
Ref	T	A	G	C	C	G	A	T	C
r1	T	A	G	C	C	G	A		
r2	T	A	G	C	C	G	A		
r3	T	A	—	C	CA	G	A		
r4	T	A	G	C	C	H	H		
r5	T	A	G	C	C	G	A	T	C
r6	S	S	G	C	C	G	A		
r7			G	C	C	G	A		

蓝色　比对到正链

橙色　比对到负链

灰色　bases 被剪切

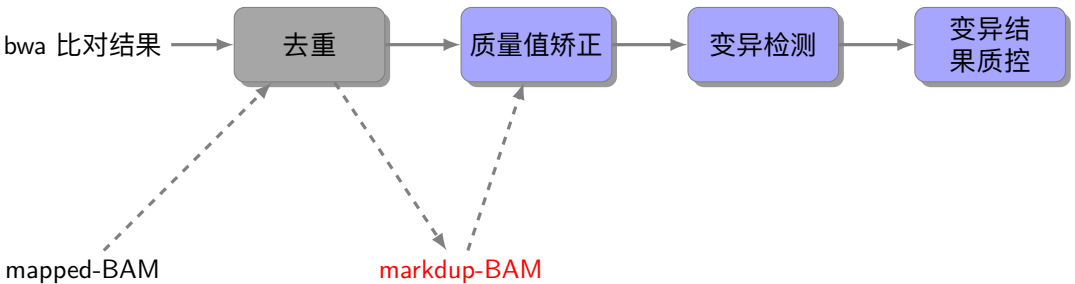
下划线　read 的 5' 端

- r1,r3,r5,r6 是一组重复
- r2,r4 是一组重复
- r7 是一组重复

去重相关工具

- MarkDupcation
- UnmarkDuplicates

去重之后

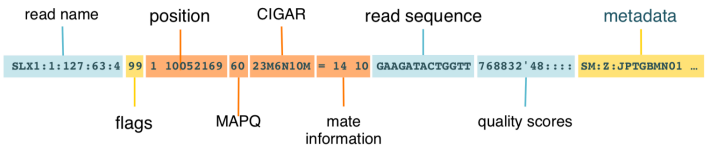


质量值?

测序仪检测的每个碱基的可信度

FASTQ 的第 4 行
BAM 文件的第 11 列

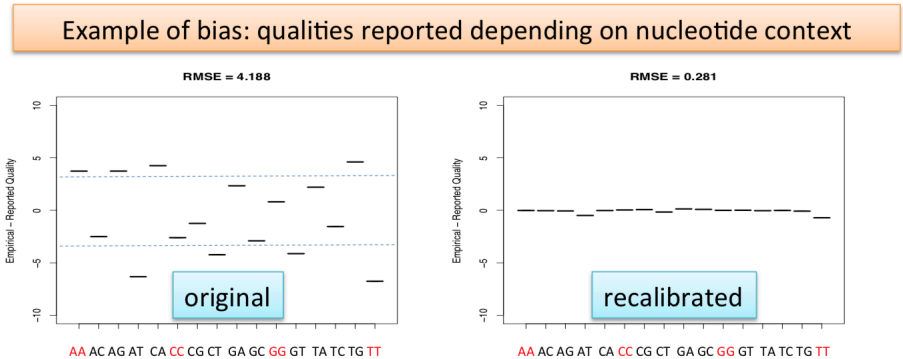
```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```



https://en.wikipedia.org/wiki/FASTQ_format

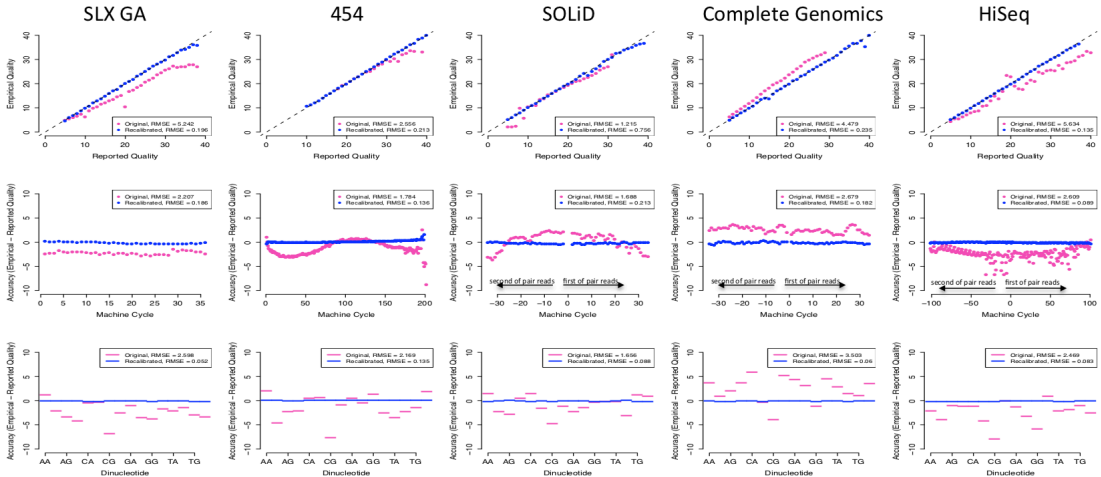
碱基质量值为什么需要矫正?

- 碱基质量值对下游分析影响很大
- 原始数据的碱基质量值普遍存在系统性误差

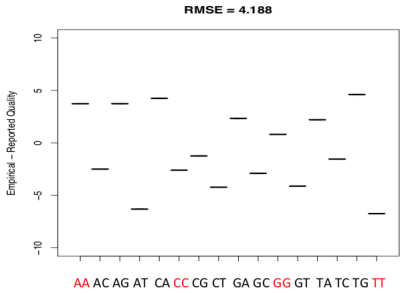


RMSE：即均方根误差，为了说明样本的离散程度

不同测序仪有不同误差模式

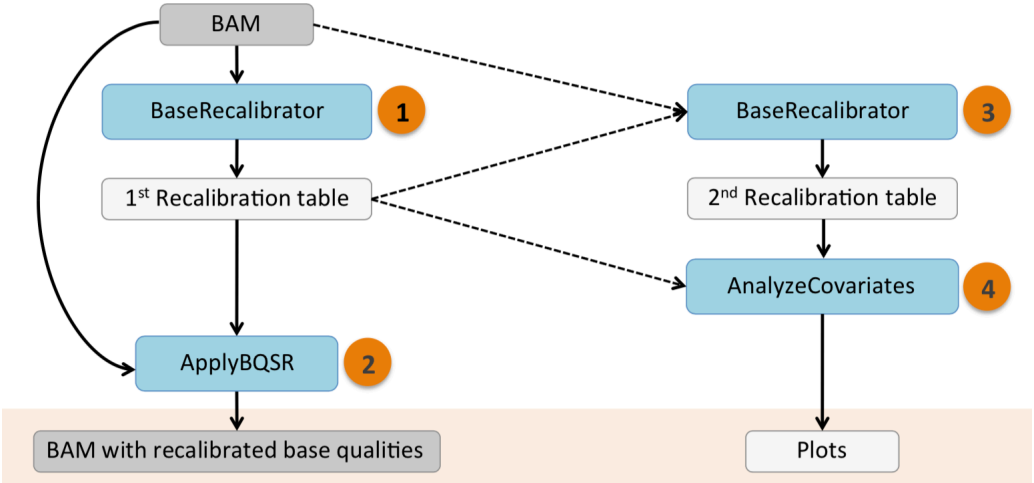


系统性误差相关协变量

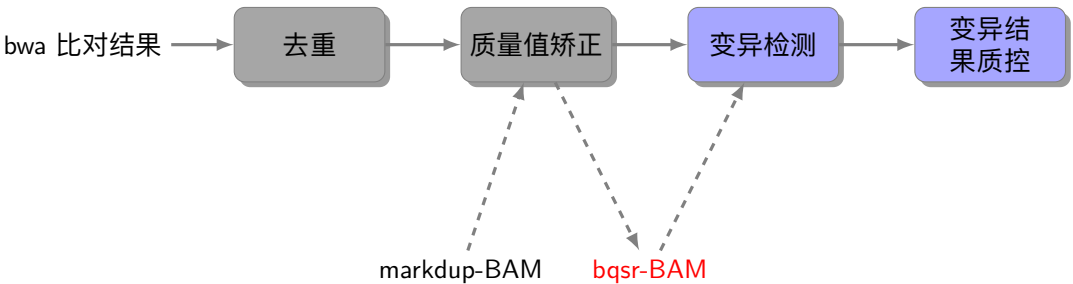


- Read group sample
- 碱基在 read 中的位置
  - 位于 pair reads 的哪条? 具体位置.
- Sequence context
- Reported base quality score

质量值矫正及其质控 *


BAM
BaseRecalibrator
1
1st Recalibration table
ApplyBQSR
2
BAM with recalibrated base qualities
BaseRecalibrator
3
2nd Recalibration table
AnalyzeCovariates
4
Plots

质量值矫正之后



Part III

变异检测

7. 变异检测
   - 单个样本的变异检测
   - GVCF 存在的意义
   - 群体变异检测

什么是变异检测 (Variant Discovery)?

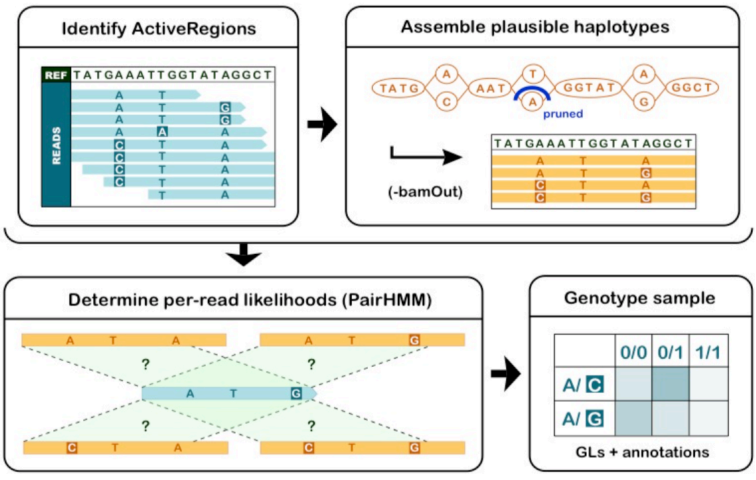
从预处理后的 BAM 文件中计算出变异信息

可检测的变异类型

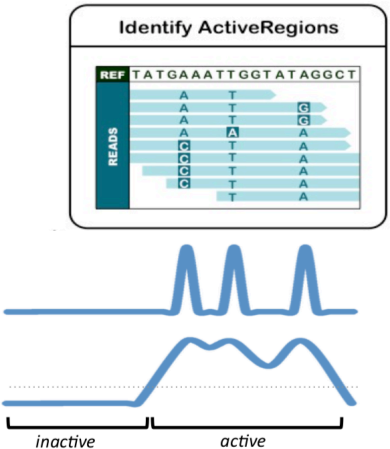
- GATK 可以检测出 SNP & INDEL

Type	What is means	Example
SNP	Single-Nucleotide Polymorphism	A/C
InDel	Insertion and Deletion	A/AGT, AC/C
CNV	Copy number variations	*
SV	Structural Variants	*

计算方法

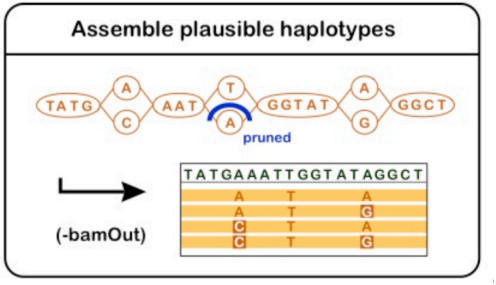

Identify ActiveRegions
REF
TATGAAATTGGTATAGGCT
READS
Assemble plausible haplotypes
TATG
AAT
GGTAT
GGCT
pruned
(-bamOut)
TATGAAATTGGTATAGGCT
Determine per-read likelihoods (PairHMM)
Genotype sample
0/0
0/1
1/1
A/C
A/G
GLs + annotations

Step1: Identify activeRegions



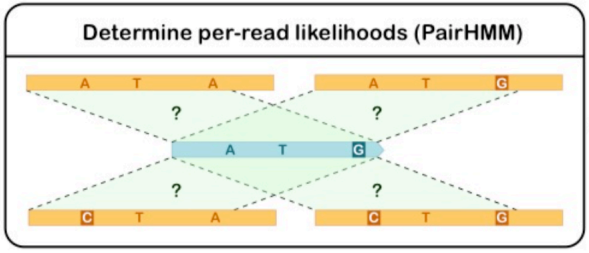
- 沿参考序列滑动窗口
- 统计 mismatches, indels and soft-clips

Step2: Assemble plausible haplotypes



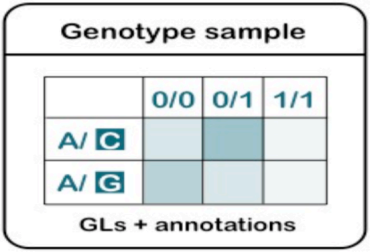
- Local realignment via graph assembly
- Align haplotypes to reference using Smith-Waterman

Step 3: Score haplotypes using PairHMM



- PairHMM aligns each read to each haplotype
- Uses base qualities as the estimate of error

Step 4: Genotype calls



- 判断每个位点的 allele(s) 组成
- 利用贝叶斯法则

Call variants per-sample

需要用到的 Tools

HaplotypeCaller

执行 HaplotypeCaller 的两种模式

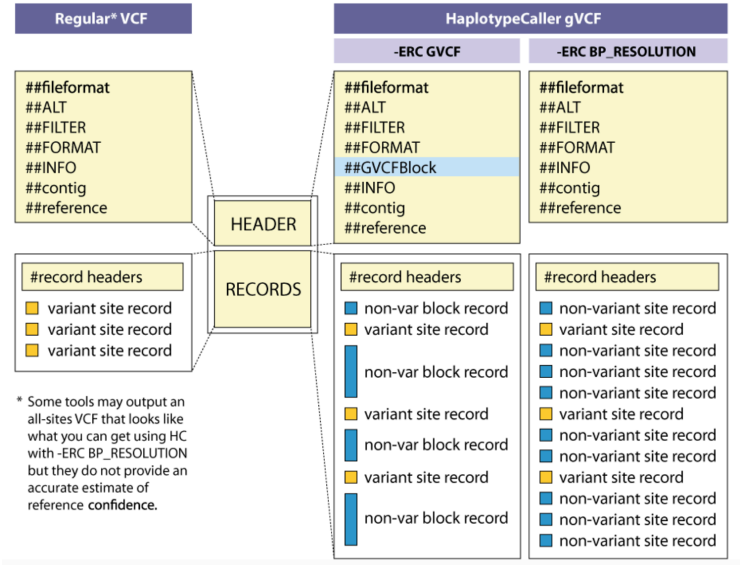
Basic mode

```
gatk HaplotypeCaller \
    --reference chr17.fa \
    --input demo.bqsr.bam \
    --output demo.hc.vcf
```

GVCF mode

```
    --output demo.g.vcf \
    --emit-ref-confidence GVCF
```

GVCF 文件格式 --包含更多信息的 VCF 格式



详见 https://gatkforums.broadinstitute.org/gatk/discussion/4017/what-is-a-gvcf-and-how-is-it-different-from-a-regular-vcf

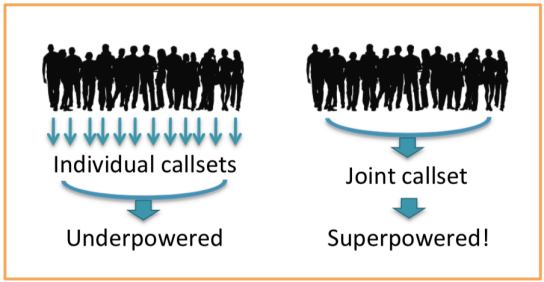
我们只想要变异信息

- GVCF 只是中间格式
- 我们需要用 GenotypeGVCFs 对 GVCF 进行 re-genotype

```
gatk GenotypeGVCFs \
   -R ref.fasta \
   -V input.g.vcf \
   -O output.vcf
```

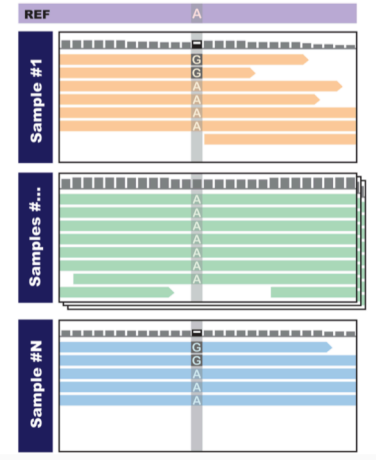
GVCF 存在的意义?

群体变异检测



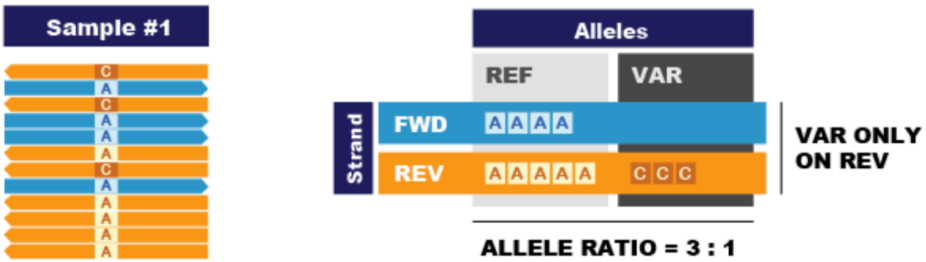
- 一个家庭或同一种群的数据可提供更有价值的信息
  - rarity of variants
  - de novo mutations
  - ethnic background

群体变异检测的优点



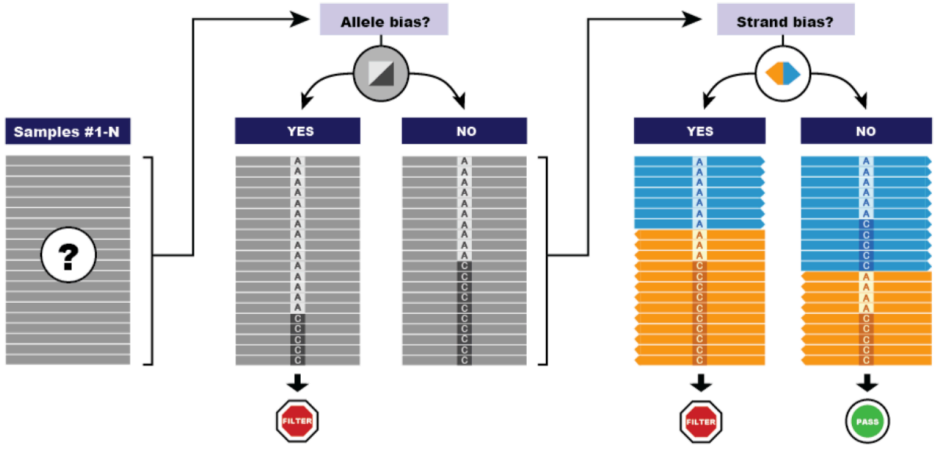
- 深度过低或有干扰信息，单样本 Call 变异时难以 Call 出可信变异
- Joint analysis 可以参考其他样本的 reads 信息

突变位点对正负链的偏好性



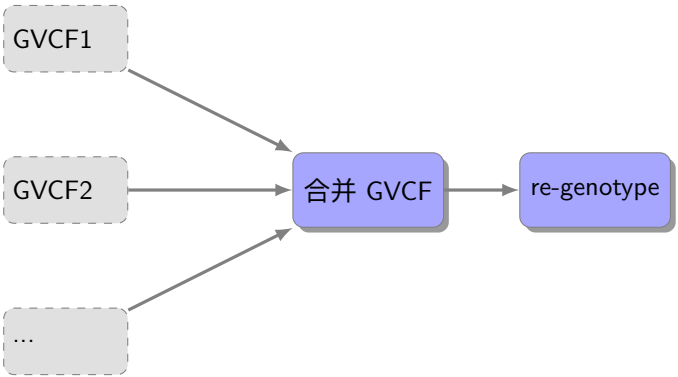
Single sample showing strand and allelic biases – would you call it?

突变位点对正负链的偏好性



Joint analysis 可减少系统性偏差

如何进行群体变异检测?



- Run GenotypeGVCFs to re-genotype samples with multi-sample model

合并 GVCF 的 Tools

用 CombineGVCFs:

```
gatk CombineGVCFs \
    -R reference.fasta \
    -V sample1.g.vcf \
    -V sample2.g.vcf \
    -O combined.g.vcf
```

- 一次最多合并 200 个样本

用 GenomicsDBImport:

```
gatk GenomicsDBImport \
    -R reference.fasta \
    -V sample1.g.vcf \
    -V sample2.g.vcf \
    -L chr20 \
    --genomicsdb-workspace-path gvcfs_db
```

- 可合并大规模样本，但必须指定区域

用 GenotypeGVCFs 进行重新分型 (re-genotype)

输入 single- or multi-sample GVCF:

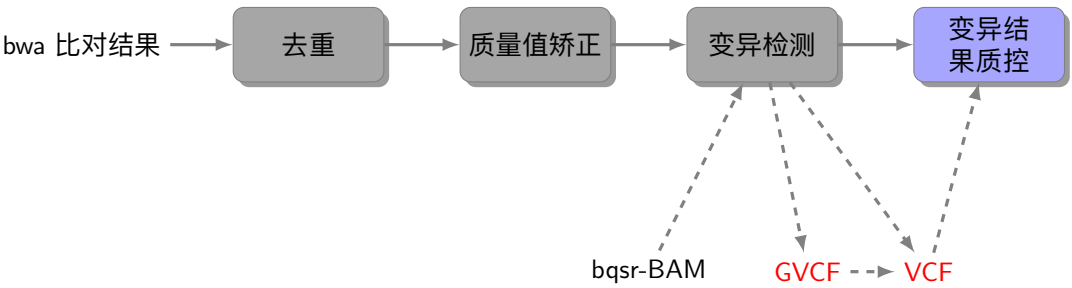
```
gatk GenotypeGVCFs \
    -R reference.fasta \
    -V variants.g.vcf \
    -O final_variants.vcf
```

输入 GenomicsDB:

```
gatk GenotypeGVCFs \
    -R reference.fasta \
    -V gendb://gvcfs_db \
    -O final_variants.vcf
```

- 最终输出的 VCF 包含多个样本的变异信息

变异检测之后



Part IV

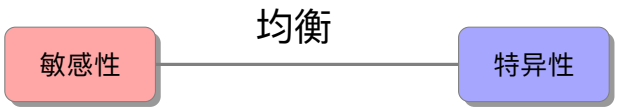
变异结果质控及其他

8. 变异结果质控和过滤

9. 精度评估 *

10. 流程化 *

变异结果进行质控和过滤的目的



- 变异检测算法是相对宽松的，通常容许较多假阳性
- 过滤方法
  - 用二分阈值进行硬过滤
  - Variant “recalibration”using machine learning

硬过滤

硬过滤定义

根据 VCF 中各项指标的阈值，进行暴力过滤.

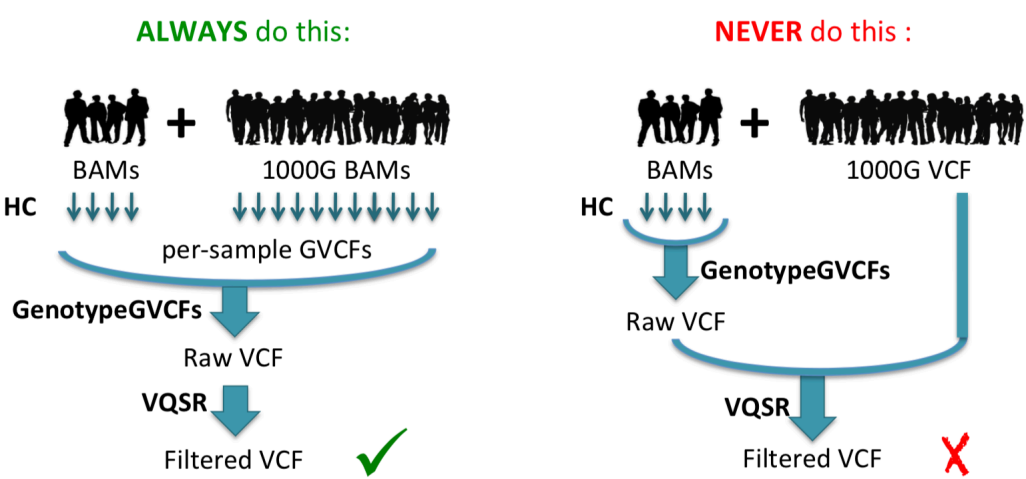
示例: QD<2.0 || MQ<40.0 || GQ>20.0 || HaplotypeScore>13.0 || MQRankSum<-12.5 || ReadPosRankSum<-8.0

指标含义在 VCF 的 header 信息中有定义

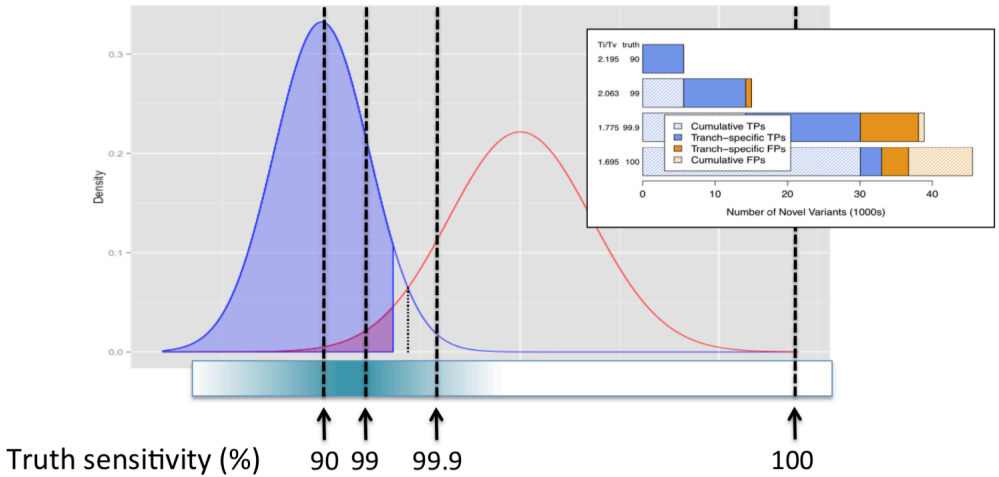
使用机器学习的方法过滤（VariantRecalibrator）

- 需要 high-confidence known sites 参考集进行训练
- 需要不少于 30 个样本同时进行

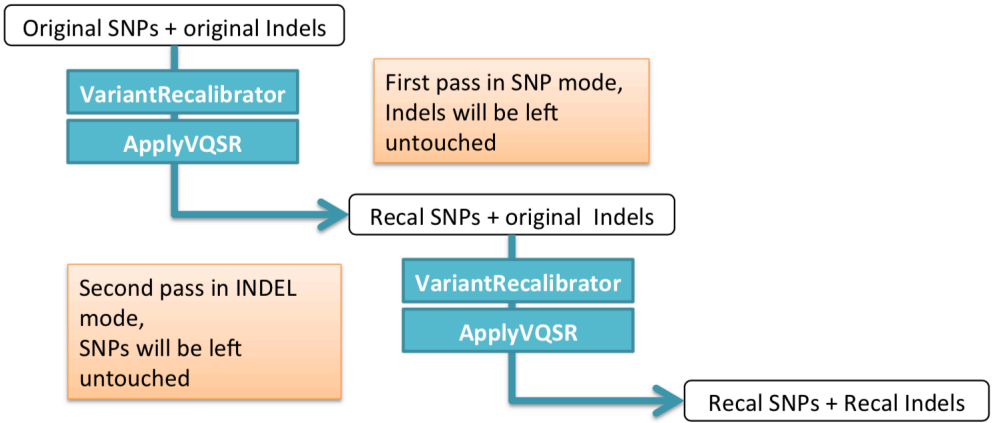
known sites 参考集: https://software.broadinstitute.org/gatk/download/bundle


ALWAYS do this:
BAMs
1000G BAMs
HC
per-sample GVCFs
GenotypeGVCFs
Raw VCF
VQSR
Filtered VCF
NEVER do this :
BAMs
1000G VCF
HC
GenotypeGVCFs
Raw VCF
VQSR
Filtered VCF

敏感度阈值


Density
0.3
0.2
0.1
0.0
Ti/Tv truth
2.195 90
2.063 99
1.775 99.9
1.695 100
Cumulative TPs
Tranch−specific TPs
Tranch−specific FPs
Cumulative FPs
0
10
20
30
40
Number of Novel Variants (1000s)
Truth sensitivity (%)
90
99
99.9
100

VariantRecalibrator 操作方法



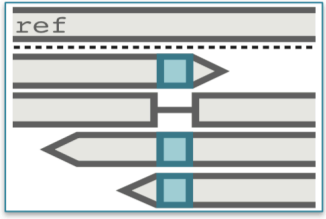
Pro-tip: Run VQSR twice in succession according to this workflow.

如何判断变异检测结果的好坏?

- 通过统计 VCF 的一些指标判断
  - 变异数目
  - indel ratio
  - titv ratio
  - ...

推荐软件 rtg vcfstat

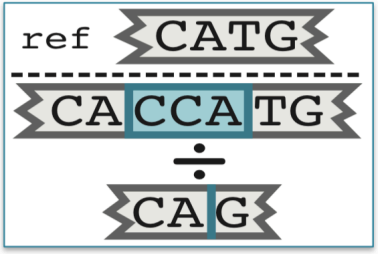
Number of Indels & SNPs



Sequencing Type	# of Variants (in 1 sample)
WGS	~4.4 M
WES	~21 k

- Variants = Indels + SNPs

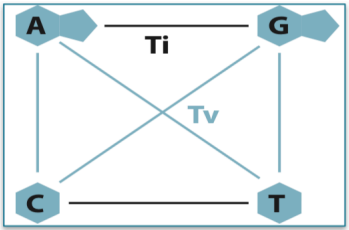
indel Ratio



Variant prevalence	Indel Ratio
Common	~1
Rare	0.2-0.5

- Ratio of insertions to deletions
- Varies by allele frequency: common ("know") vs. rare ("nove")

TiTv Ratio (Transitions/Transversions)

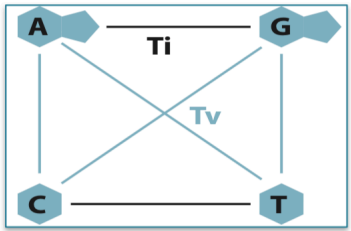


转换 (transition) 指嘌呤置换嘌呤或嘧啶置换嘧啶

颠换 (transversion) 指嘌呤置换嘧啶或嘧啶置换嘌呤

Sequencing Type	TiTv Ratio
WGS	2.0-2.1
WES	3.0-3.3

TiTv Ratio (Transitions/Transversions)



转换 (transition) 指嘌呤置换嘌呤或嘧啶置换嘧啶

颠换 (transversion) 指嘌呤置换嘧啶或嘧啶置换嘌呤

Sequencing Type	TiTv Ratio
WGS	2.0-2.1
WES	3.0-3.3

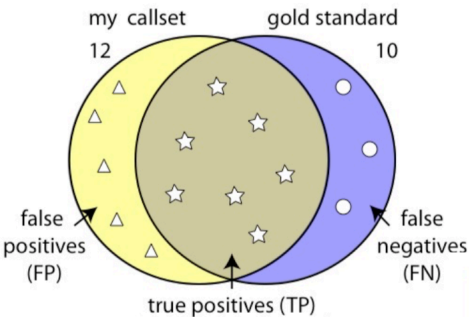
- 仅适用于 SNP
- 如果突变是随机的，ratio 应该是 0.5
- 低的 titv ratio 指示高的假阳性

如何评估自己搭建的一套分析流程

两个维度：

- 性能
- 精度 (准确性)

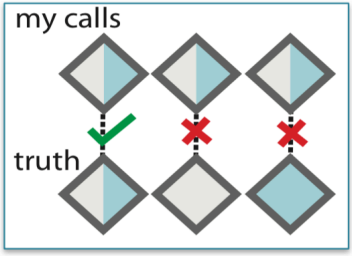
精度评估



精度评估需要什么?

- 参考样本
  例如由美国国家标准与技术研究所（NIST）提供的样本 NA12878
- 参考样本的变异标准集

Genotype 一致性



- 与标准集 genotype 不一致的突变是假阳性

精度评估的主要指标

TP：真阳性位点 (True positives)，在标准集中存在，test.vcf 中也存在的变异数.

FN：假阴性位点 (False negatives), 在标准集中存在，test.vcf 中不存在的变异数.

FP：假阳性位点 (False positives), 在标准集中不存在，test.vcf 中存在的变异数.

Precision (PPV)： $Precision = \frac{TP}{TP+FP}$

Sensitivity (Recall): $Sensitivity = \frac{TP}{TP+FN}$

F-measure：Precision 和 Sensitivity 的调和平均数, $Fmeasure = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity}$

FDR：False Discovery Rate, $FDR = \frac{FP}{FP+TP}$

根据 F-measure 的值来判断结果优劣，值越高越好

精度评估软件

- RTG vcfeval
- GATK3 VariantEval
- Picard CVCM

RTG vcfeval 结果示例

```
rtg vcfeval:
Threshold  True-pos  False-pos  False-neg  Precision  Sensitivity  F-measure
-------------------------------------------------------------------------------
  109.820   3180942      15954      12007     0.9950       0.9962     0.9956
     None   3184920      28556       8029     0.9911       0.9975     0.9943
```

Threshold：作为结果最优情况下的过滤阈值，可以是变异的 QUAL(VCF 第六列) 值等指标.

流程化 (Pipelining)*

--将我们的分析过程写成可重复利用的流程

有什么好处：

- 减少重复工作
- 增强结果可重现性
- 减少人工错误
- 提高分析效率

来看看 Broad Institute 提供的自动化流程框架

WDL

What is WDL?

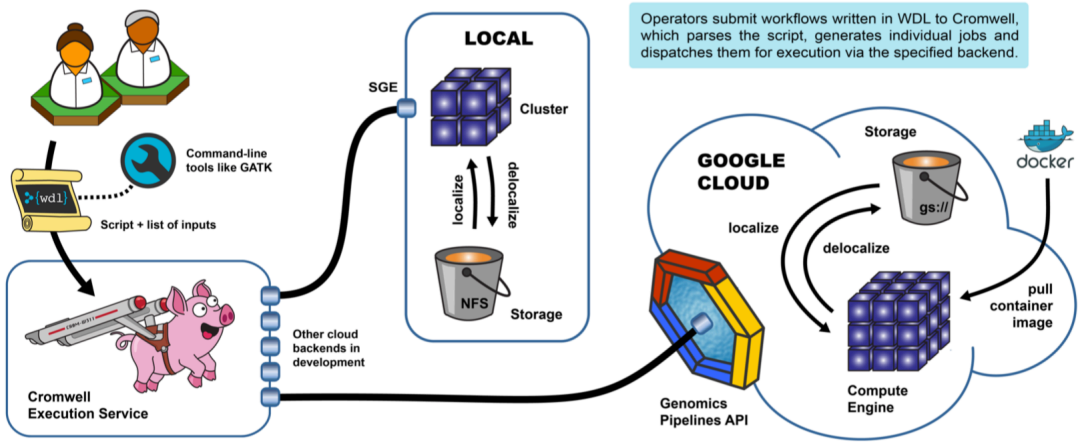
Cromwell

What is Cromwell?

WDL = Workflow Definition Language

http://www.openwdl.org/

Cromwell = Execution Service



Summary

- What is GATK?
- 我们用它干什么?
- 主要步骤有哪些?