SNP Calling

北京基因组研究所
Husn Group：连明

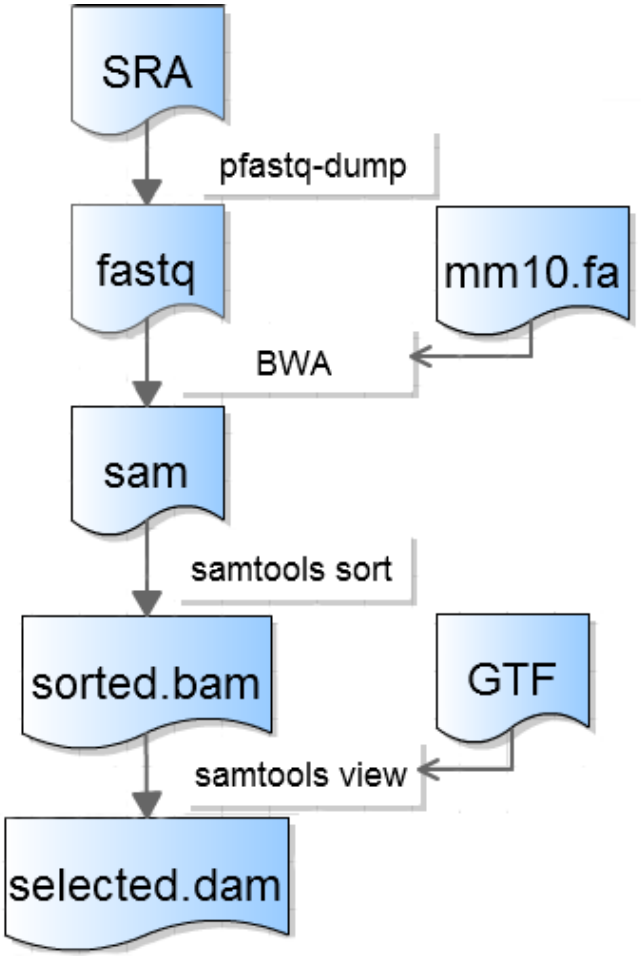
目录

- 我们的课题
- SNP Calling的常规步骤
- GATK4流程
- GATK4流程里的那些坑

我们的课题
——目标基因在小鼠各品系的reference

- 下载小鼠（mouse）各品系的重测序数据
- Mapping到小鼠标准参考基因组（GRCm38/mm10）上
- 富集落在目标基因区域的reads
- 将富集到的reads跑snp calling流程（目前选择GATK4流程）得到目标基因区域的snp&indel
- 替换标准参考基因组上的snp&indel（RGAAT）

我们的课题
——目标基因在小鼠各品系的reference

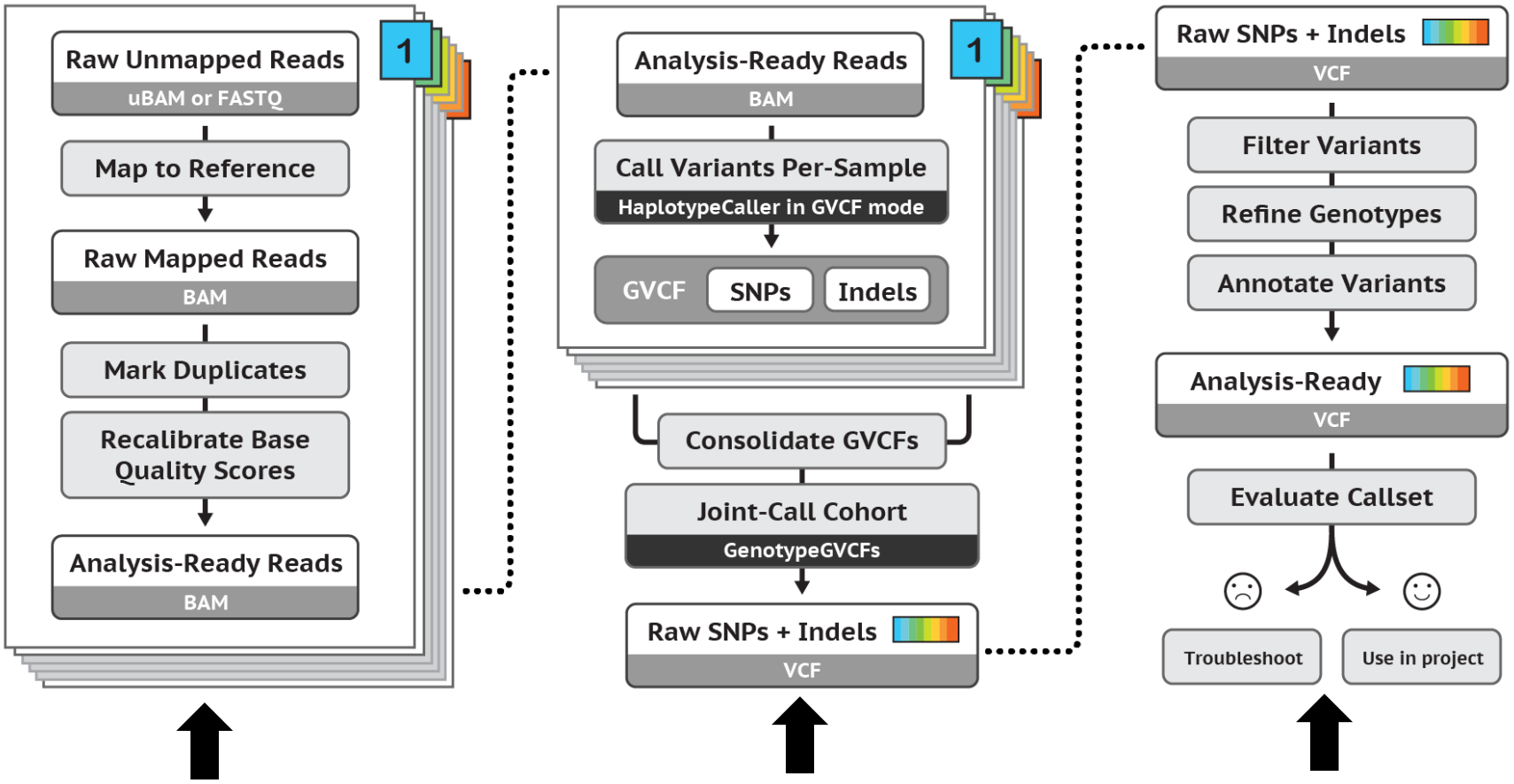


目标区域reads富集流程

SNP Calling的常规步骤

- Mapping to reference
  - BWA √
  - Bowtie2
- Call SNP
  - GATK √
  - Samtools+bcftools
  - Freebayes
- Variants Annotation
  - Annovar
  - SnpEff
  - VEP

GATK4流程



snp calling前的预处理

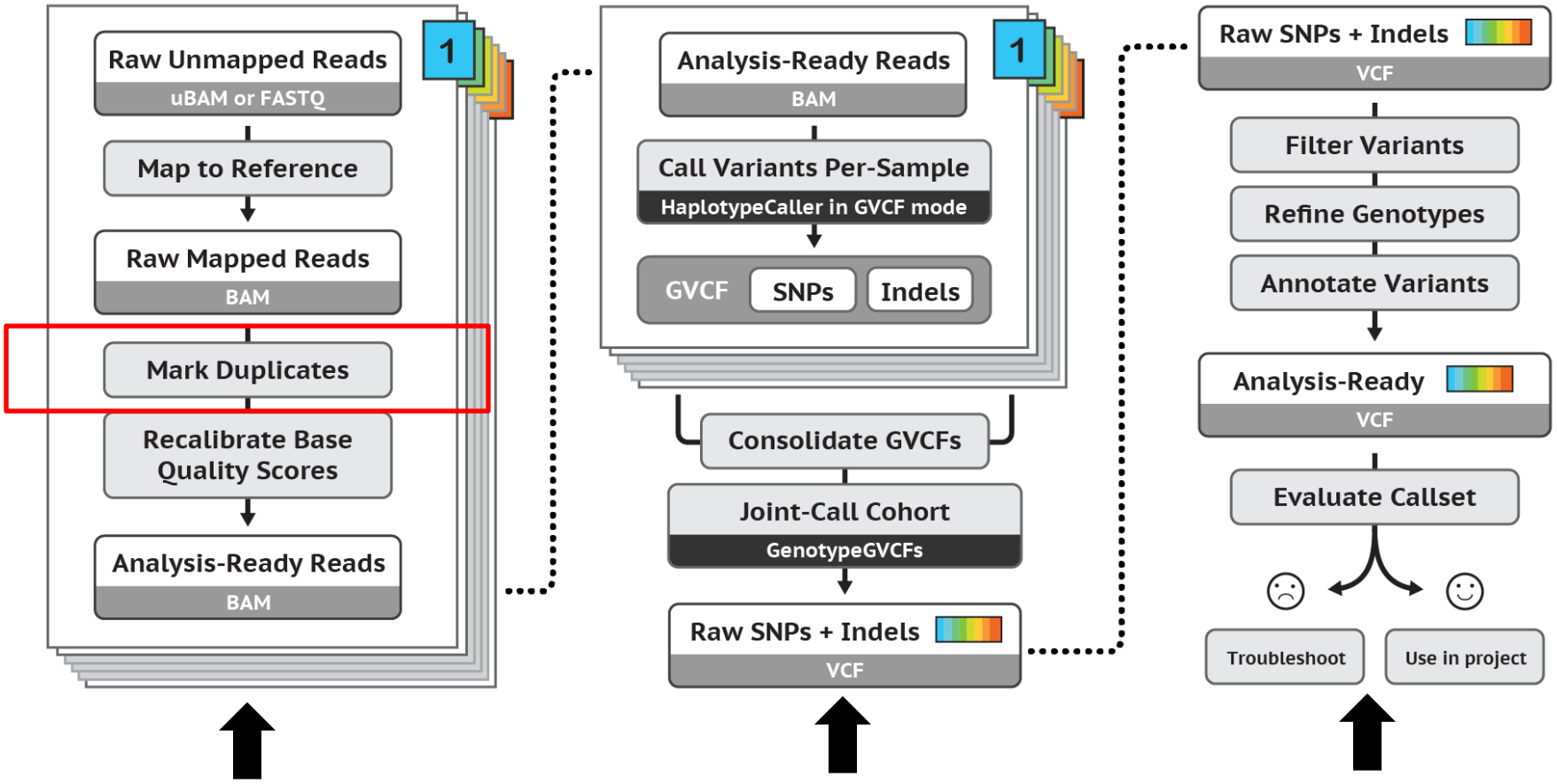
snp calling

过滤snp与注释

该部分不涉及操作命令，具体代码与更详细的内容请移步GitHub笔记

传送门：https://github.com/Ming-Lian/NGS-analysis/blob/master/call-snp.md

GATK4流程
--PCR bias的影响

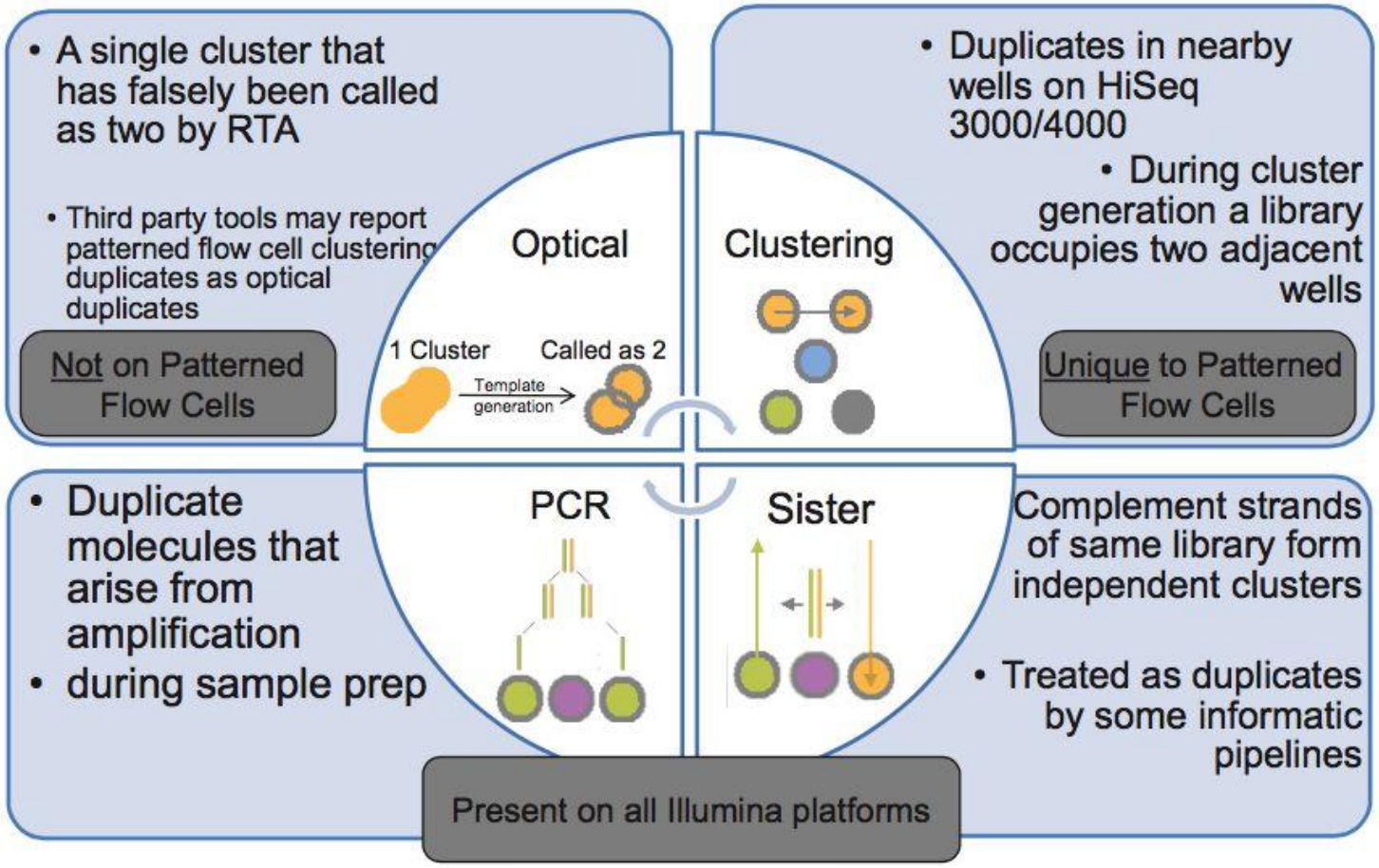


snp calling前的预处理

snp calling

过滤snp

GATK4流程
--PCR bias的影响

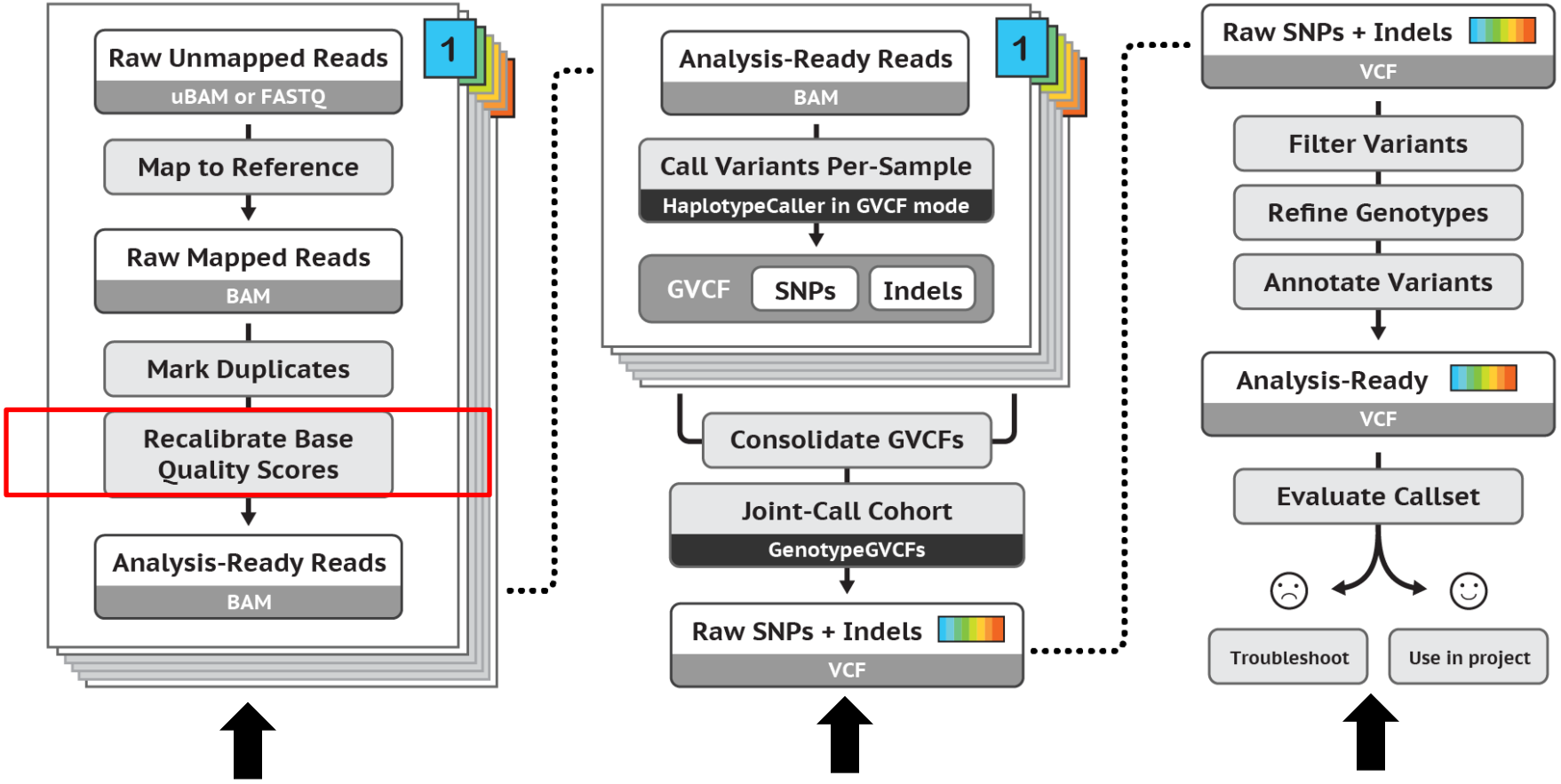


NGS中duplicate产生的原因

GATK4流程
--PCR bias的影响

- PCR反应过程中也会带来新的碱基错误。 发生在前几轮的PCR扩增发生的错误会在后续的PCR过程中扩大， 同样带来假的变异
- 对于真实的变异， PCR反应可能会对包含某一个碱基的DNA模版扩增更加剧烈（这个现象称为PCR Bias） 。 因此， 如果反应体系是对含有reference allele的模板扩增偏向强烈， 那么变异碱基的信息会变小， 从而会导致假阴

GATK4流程
--碱基质量值校正



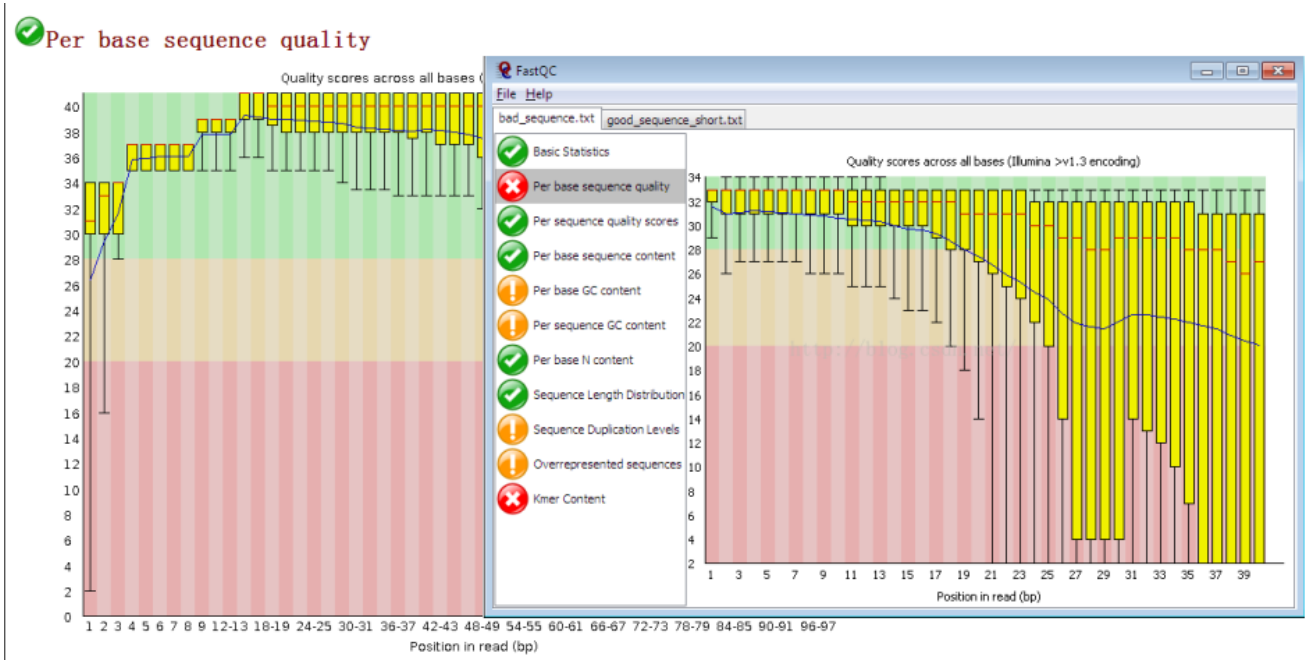
snp calling前的预处理

snp calling

过滤snp

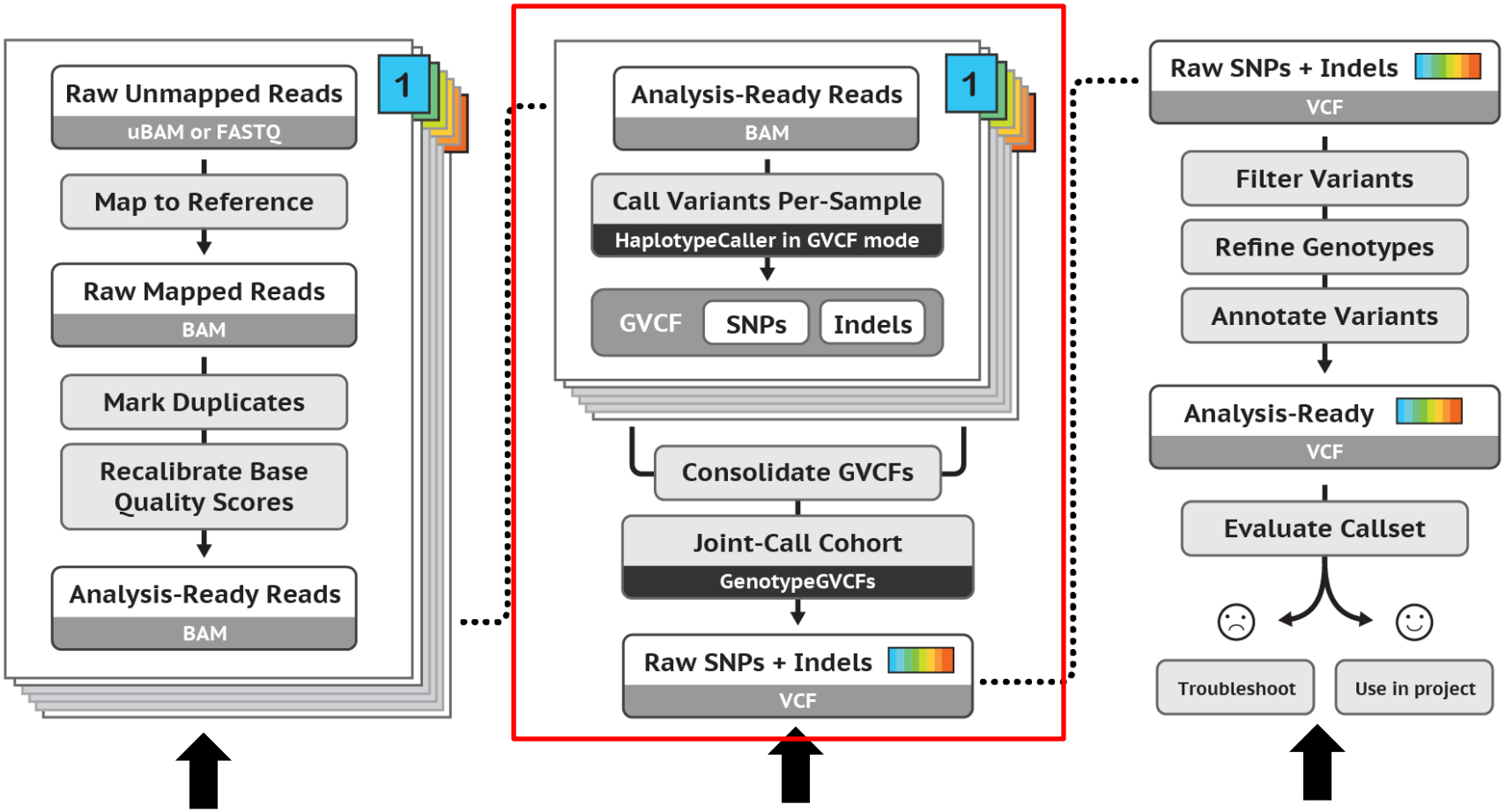
GATK4流程
--碱基质量值校正

- 测序仪在测序过程中会给出每个碱基质量值的评估
- 一般两端质量值低且不稳定，中间质量值高——存在系统误差



用机器学习的方法进行碱基质量校正

GATK4流程
-- snp calling的策略



snp calling前的预处理

snp calling

过滤snp

GATK4流程
-- snp calling的策略

- single sample calling：

  每一个sample的bam file都进行单独的snp calling，然后每个sample单独snp calling结果再合成一个总的snp calling的结果

- joint calling：

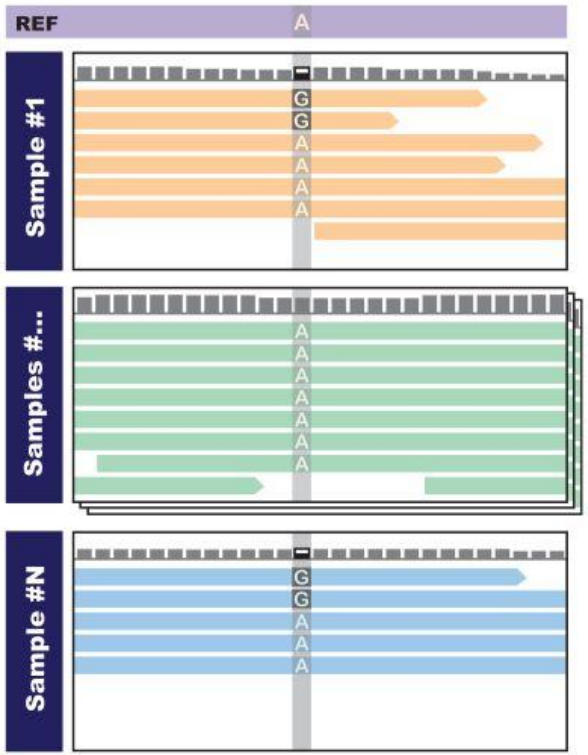
  所有samples的BAM files一起call 出一个包含所有samples 变异信息的output

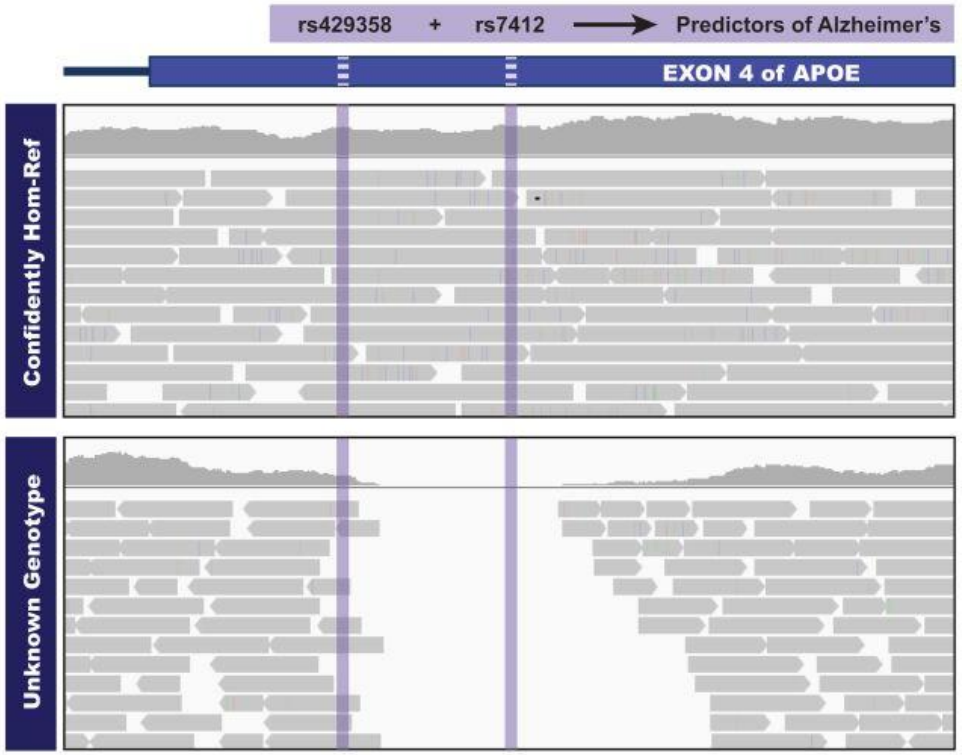
GATK4流程

-- snp calling的策略

多样本时，推荐使用joint calling

原因一：对于低频率的变异具有更高更好的检测sensitivity



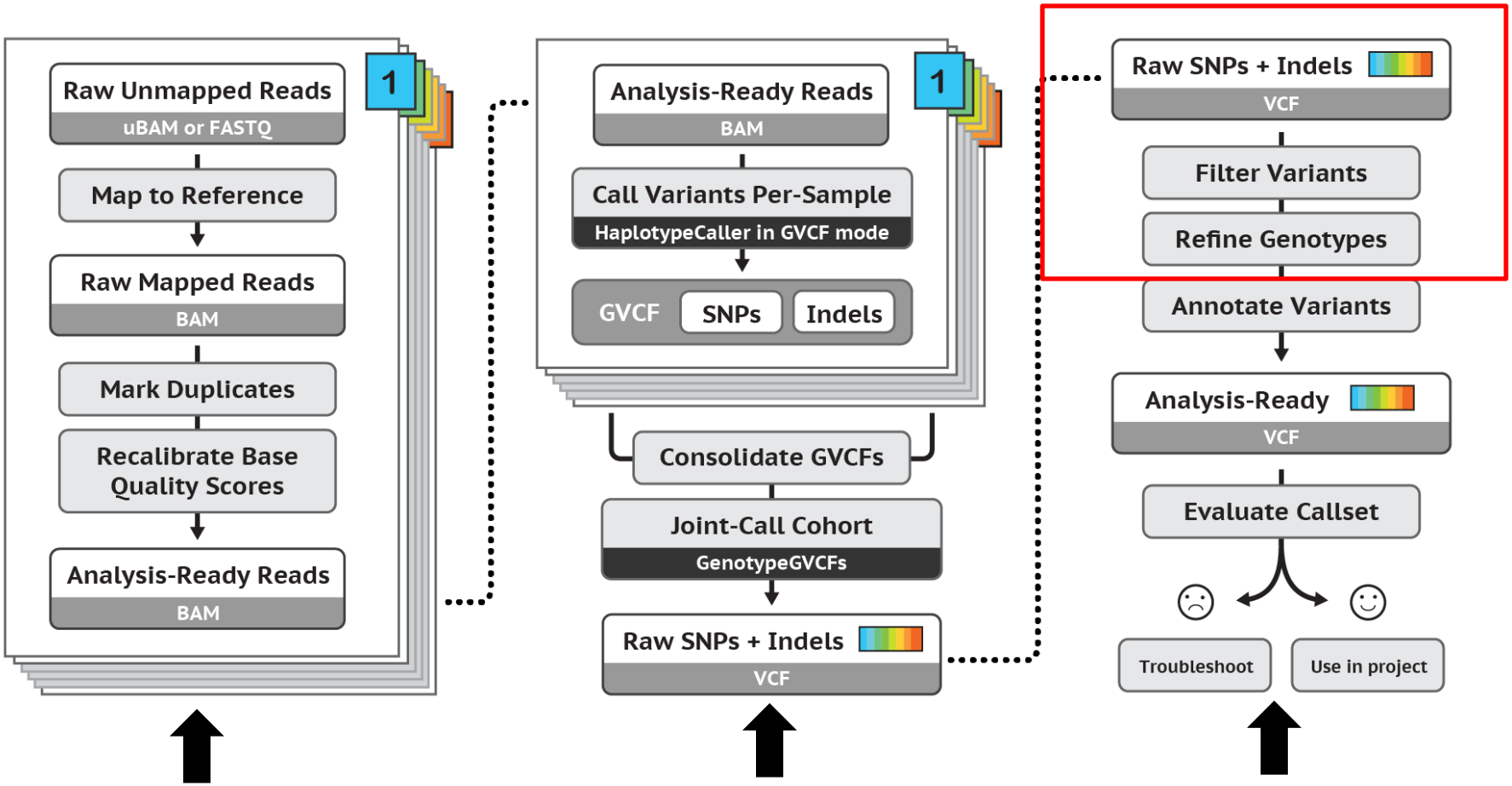


GATK4流程
-- snp calling的策略

- 原因二：

现在使用过滤变异的方法例如VQSR等利用的统计模型，都基于一个比较大的samples size。joint calling 这种方法可以提供足够的数据，确保过滤这一步是统一应用于所有samples的

GATK4流程
-- snp位点过滤



snp calling前的预处理

snp calling

过滤snp

GATK4流程
-- snp位点过滤

- 通过质量校正来过滤（Filter Variants by Variant (Quality Score) Recalibration）

  用机器学习的方法基于已知的变异位点对caller给出的原始variant quality score 进行校正 (VQSR)

  执行变异校正需要满足两个条件：

  （1）大量高质量的已知variants作为训练集，而这对于许多的物种是不满足的

  （2）数据集不能太小，因为它需要足够的数据集来识别good vs. bad variants

GATK4流程
-- snp位点过滤

- 直接过滤 (hard-filtering)

通过卡阈值来筛选高质量的snp&indel

GATK4流程里的那些坑
-- Mapping

- 建index

  用cat按照染色体的顺序拼接起来，因为GATK后面的一些步骤对染色体顺序要求非常变态，如果下载整个hg19，很难保证染色体顺序是1-22，X,Y,M

- reference chromosome/contig 命名方式

  所有涉及到chromosome/contig名的地方都要用统一的命名方法，**chr+染色体号** 或 **染色体号**，尤其要注意作为reference的VCF文件

- read goup 标签

  mapping的时候需要给输出的SAM/BAM文件添加read goup 标签

GATK4流程里的那些坑

-- Mapping

```
@HD     VN:1.0  GO:none SO:coordinate
@SQ     SN:1    LN:249250621    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/hu
man_g1k_v37.fasta    M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ     SN:2    LN:243199373    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/hu
man_g1k_v37.fasta    M5:a0d9851da00400dec1098a9255ac712e
@RG     ID:ERR000162    PL:ILLUMINA     LB:g1k-sc-NA12776-CEU-1 PI:200  DS:SRP000031    SM:NA12776      CN:
SC
@RG     ID:ERR000252    PL:ILLUMINA     LB:g1k-sc-NA12776-CEU-1 PI:200  DS:SRP000031    SM:NA12776      CN:
SC
@RG     ID:ERR001684    PL:ILLUMINA     LB:g1k-sc-NA12776-CEU-1 PI:200  DS:SRP000031    SM:NA12776      CN:
SC
@RG     ID:ERR001685    PL:ILLUMINA     LB:g1k-sc-NA12776-CEU-1 PI:200  DS:SRP000031    SM:NA12776      CN:
SC
@PG     ID:GATK TableRecalibration      VN:v2.2.16      CL:Covariates=[ReadGroupCovariate, QualityScoreCova
riate, DinucCovariate, CycleCovariate], use_original_quals=true, defau
t_read_group=DefaultReadGroup, default_platform=Illumina, force_read_group=null, force_platform=null, solid
_recal_mode=SET_Q_ZERO, window_size_nqs=5, homopolymer_nback=7, except on_if_no_tile=false, pQ=5, maxQ=40,
smoothing=137       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:b4eb
71ee878d3706246b7c1dbef69299
@PG     ID:bwa  VN:0.5.5
ERR001685.4315085       16      1       9997    25      35M     *       0       0       CCGATCTCCCTAACCCTAA
CCCTAACCCTAACCCT     ?8:C7ACAABBCBAAB?CCAABBEBA@ACEBBB@?     XT:A:U  XN:i:4    X0:i:1  X1:i:0  XM:i:2  XO:i
:0  XG:i:0  RG:Z:ERR001685  NM:i:6  MD:Z:0N0N0N0N1A0A28     OQ:Z:>>:>2>>>>>>>>>>>>>>>>>>>>?>>>>>??>???>
ERR001689.1165834       117     1       9997    0       *       =       9997    0       CCGATCTAGGGTTAGGGTT
AGGGTTAGGGTTAGGG     >7AA<@@C?@?B?B?>>9?B?>>A?B???BAB??@     RG:Z:ERR001689    OQ:Z:>:<<8<<<><<><><<<7<>>>?
>>??>???????
```

GATK4流程里的那些坑
-- snp calling

- VCF index

  在进行碱基质量校正，snp calling和变异质量值校正时都要用到已知的变异位点信息作为reference，需要提供reference VCF文件，该文件需要做好index

- 指定call snp的区域

  可以通过-L参数指定call snp的区域，这些区域称为interval，若有多个intervals可以提供一个文件保存这些interval list，

  注意： **interval list 必须按照reference的顺序排好序**