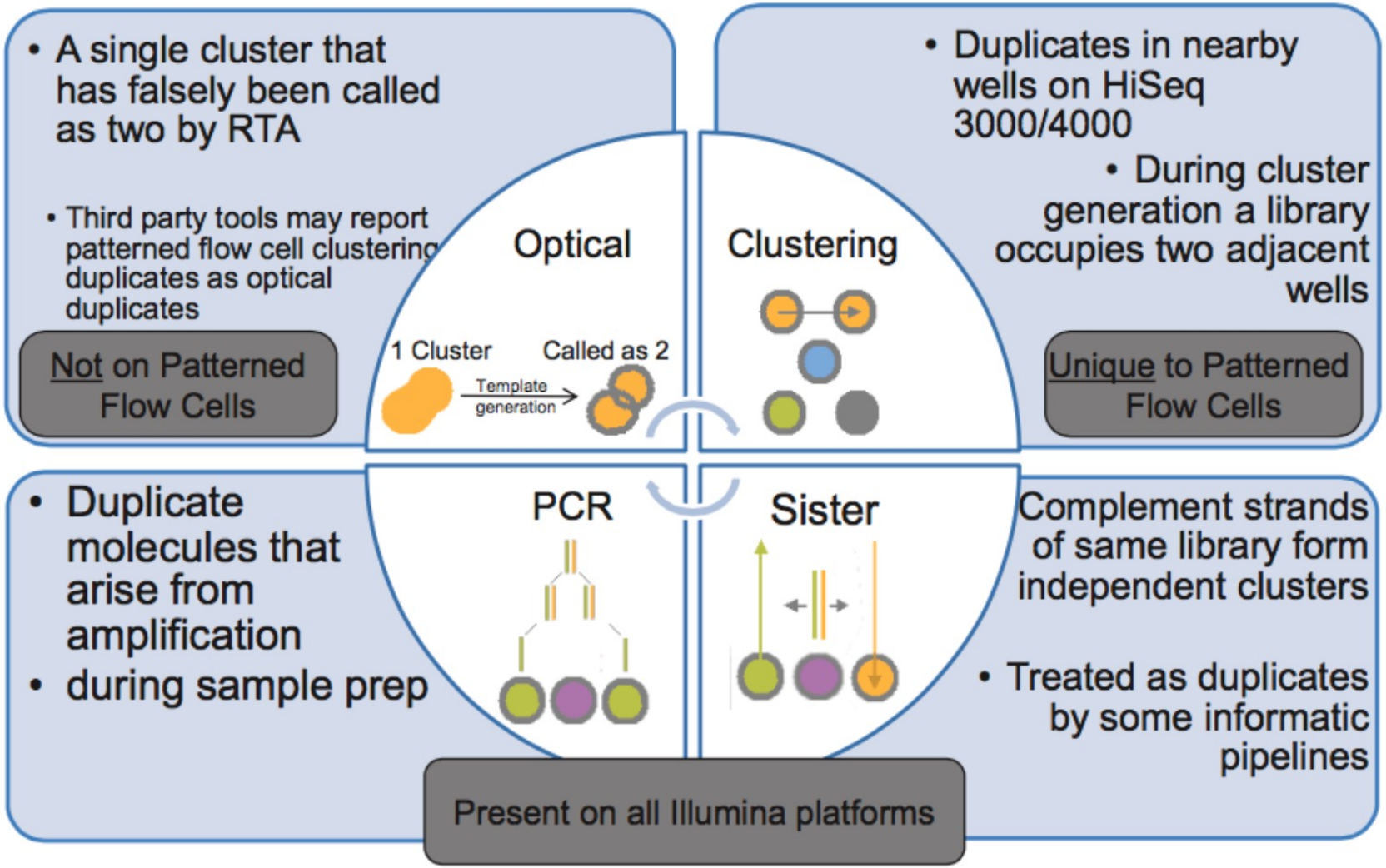
BerryGenomics
贝瑞基因

群体重测序-variants calling

王鹏 2020/11/1

Illumina中的重复reads



1）PCR重复：同一个DNA片段会产生多个不同的拷贝

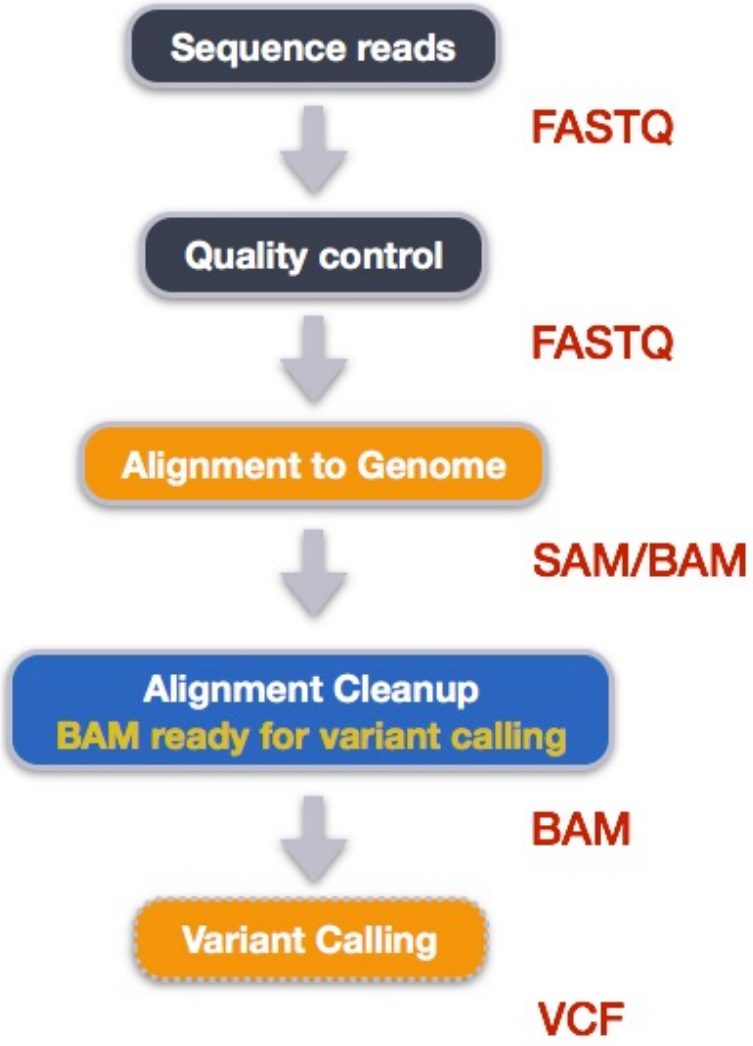
2）簇（Cluster）重复：Flowcell上桥式PCR建库过程中，一个cluster的序列结合到另一个cluster中

3）光学（Optical）重复：测序时，由于光波衍射导致重影出现相同reads

4）姐妹（Sister）重复：两条互补链结合Flowcell上的引物，测序后生成反向互补的reads

https://ming-lian.github.io/2019/02/08/call-snp/ https://www.biostars.org/p/229842/

Call变异的流程



Step1: Index the reference genome

Bwa/Samtools/bcftools/GATK

Step2: Align reads to reference genome

Step3: Sort BAM file by coordinates

Step4: Variant calling

- Calculate the read coverage of positions in the genome
- Detect the single nucleotide polymorphisms (SNPs)
- Filter and report the SNP variants in variant calling format (VCF)

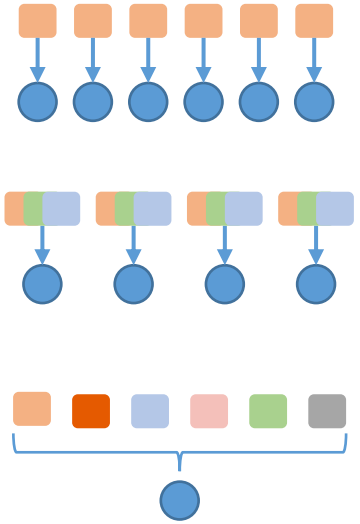
Step5: Assess the alignment (visualization) - optional step

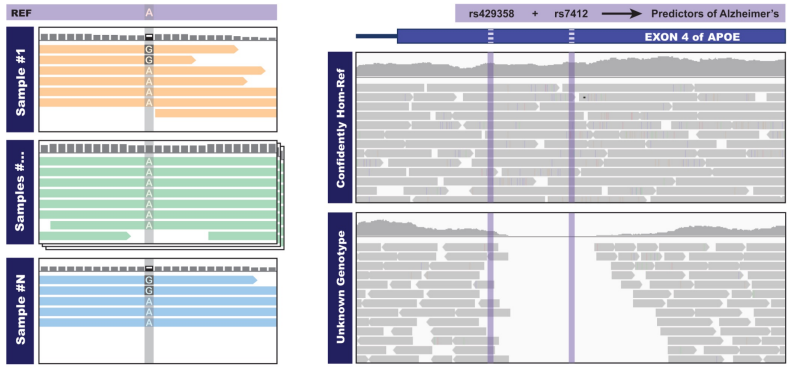
https://datacarpentry.org/wrangling-genomics/04-variant_calling/index.html

群体Call变异的方法

https://gatk.broadinstitute.org/hc/en-us/articles/360035890431

- **single sample calling**: sample BAMs are analyzed individually, and individual call sets are combined in a downstream processing step
- **batch calling**: sample BAMs are analyzed in separate batches, and batch call sets are merged in a downstream processing step
- **classic joint calling**: variants are called simultaneously across all sample BAMs, generating a single call set for the entire cohort.





✓ 1. Clearer distinction between homozygous reference sites and sites with missing data

✓ 2. Greater sensitivity for low-frequency variants

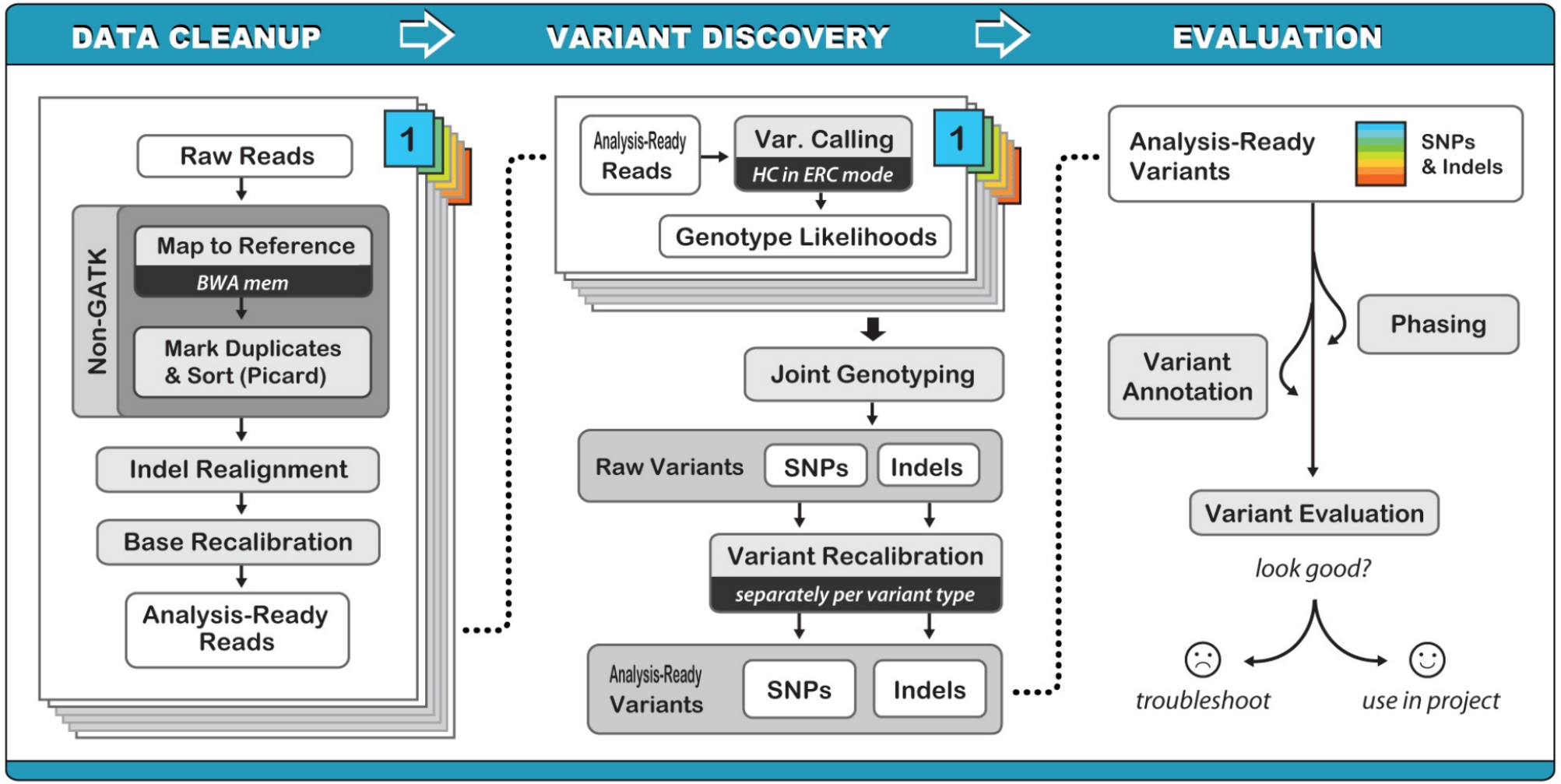
✓ 3. Greater ability to filter out false positives

☹ **Scaling & infrastructure**

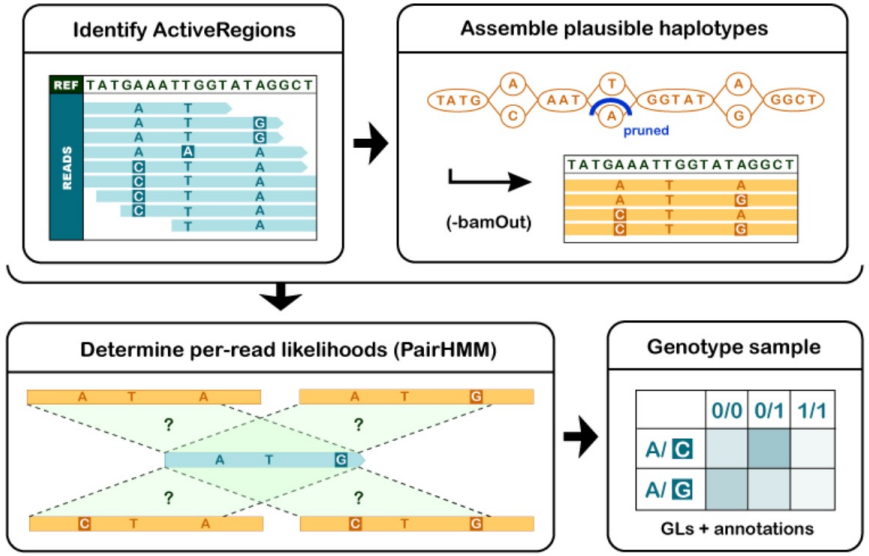
☹ **The N+1 problem**

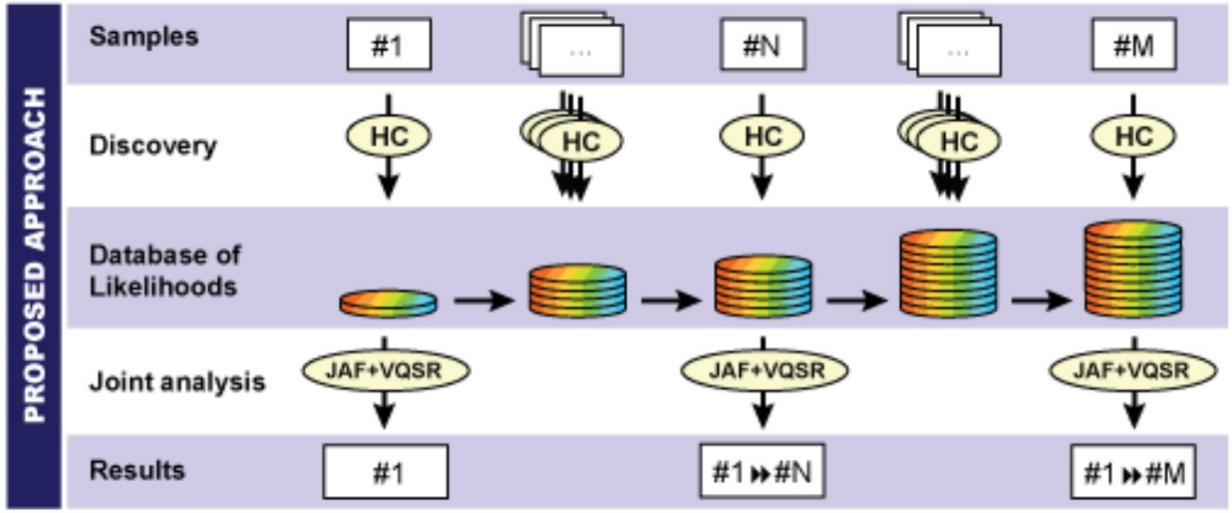
https://ming-lian.github.io/2019/02/08/call-snp/

HaplotypeCaller群体变异检测



HaplotypeCaller群体变异检测





Step1: call variants individually (gVCF)

Step2: combine multiple gVCFs using GenomicsDB

Step3: joint genotyping to get SNPs and inDels

Step4: filtering based on Quality Score

All case and control samples

~3M variants

	Site	Variant	Sample 1	Sample 2	...	Sample N
SNP	*1:1000*	*A/C*	*0/0 0,10,100*	*0/1 20,0,200*	...	*0/0 0,100,255*
Indel	1:1050	T/TC	0/0 0,10,100	0/0 0,20,200	...	1/0 255,0,255
SNP	*1:1100*	*T/G*	*0/0 0,10,100*	*0/1 20,0,200*	...	*0/0 0,100,255*
	...	...	...	...	...	...
SNP	*X:1234*	*G/T*	*0/1 10,0,100*	*0/1 20,0,200*	...	*1/1 255,100,0*

Genotypes:
0/0 ref
0/1 het
1/1 hom-alt

Likelihoods:
A/B/C phred-scaled probability of hom (A), het (B), hom-alt (C) genotypes given NGS data

BerryGenomics
贝瑞基因

Thank You!

官 方 网 站

官 方 微 信

www.berrygenomics.com