Topic 7: Variant calling using GATK

Biol 525D - Bioinformatics for Evolutionary Biology 2020

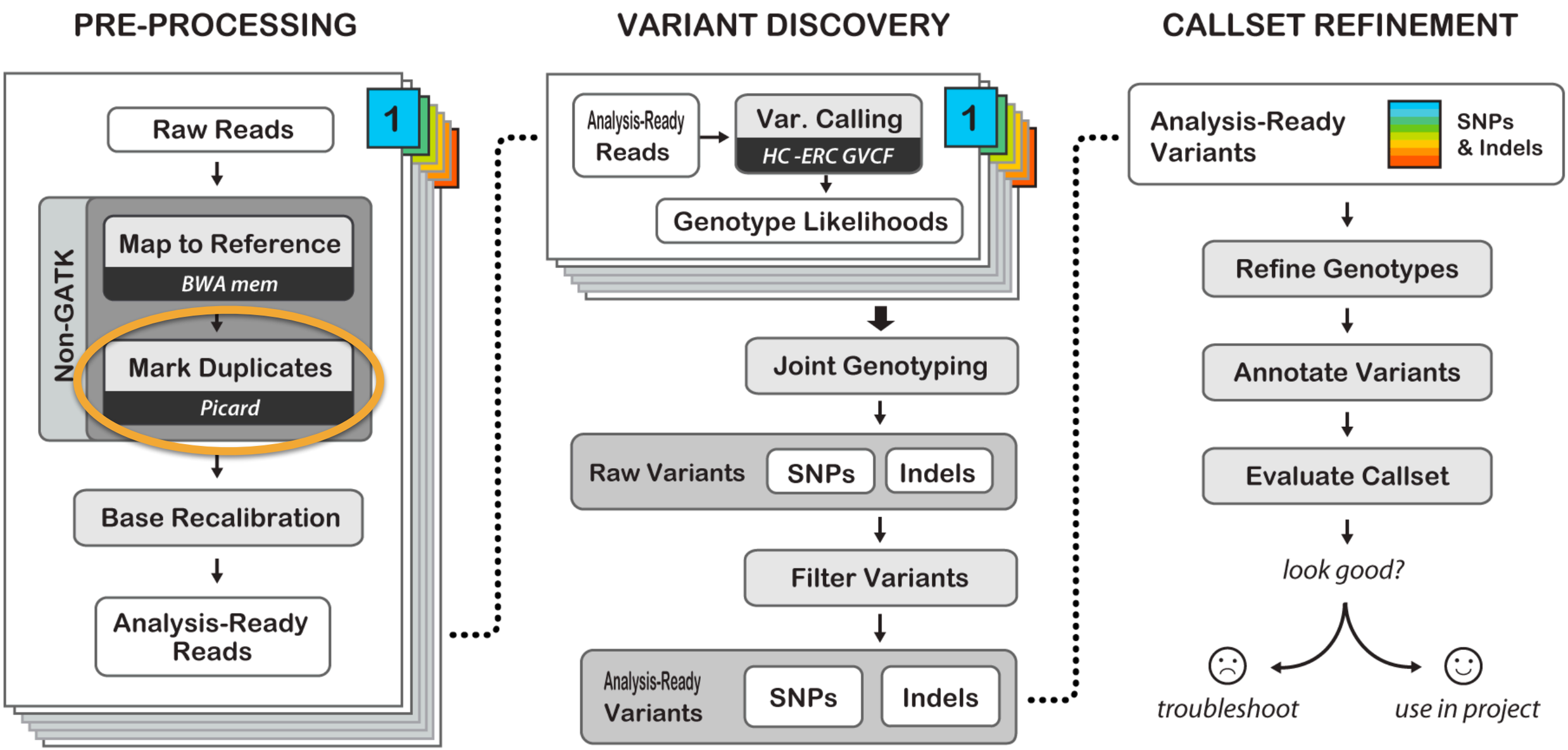
Learning Goals

- Define the steps involved in SNP calling and what they are doing.
- Understand the reason for haplotype based SNP calling.
- Define the N+1 problem in genotyping.
- Understand approaches to variant filtering.

Alignment + Variant Calling in concept

1. get some reference system
2. index it
3. align
4. mark duplicates
5. call variants
6. filter variants

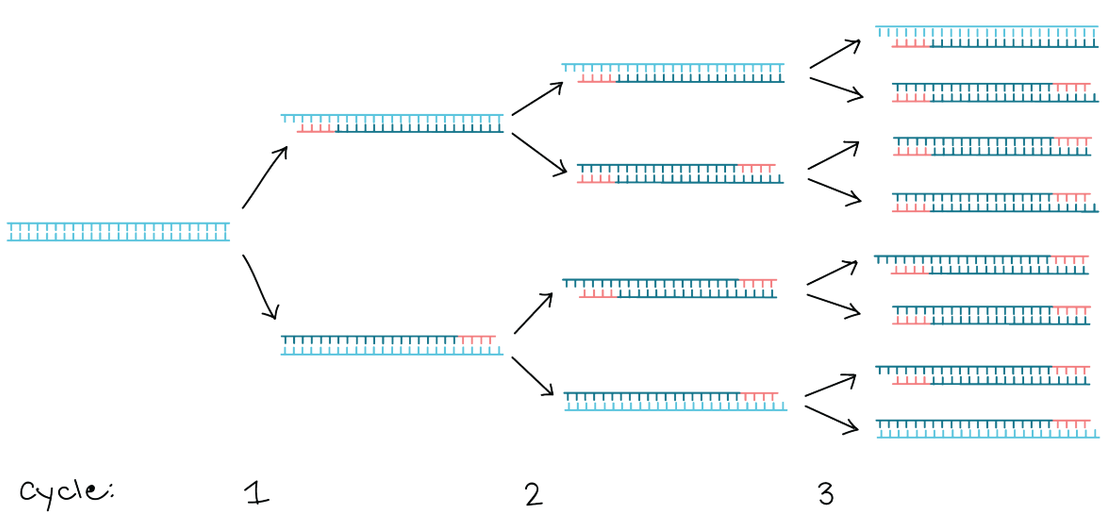
GATK Best Practises for Variant Discovery in DNAseq



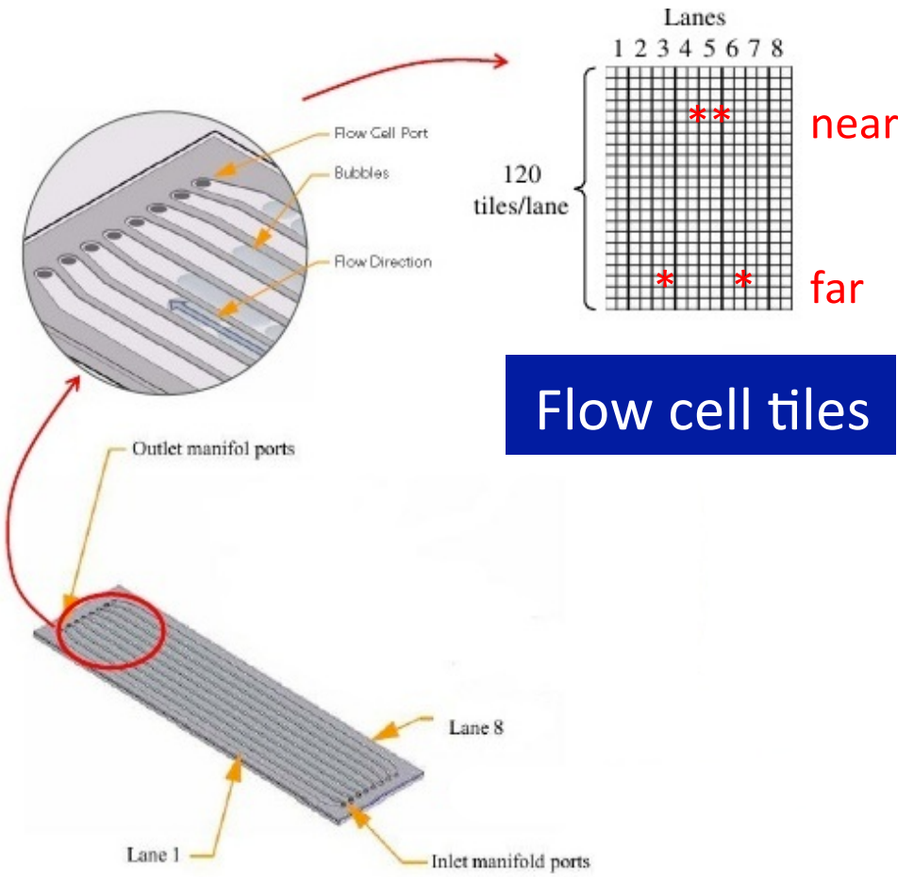
Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

Where does the duplication come from?

- **PCR DUPLICATES**
  - **Increases with cycles**
- **OPTICAL DUPLICATES**
  - **Are nearby clusters on a flow cell lane**



https://www.khanacademy.org/science/biology/biotech-dna-technology/dna-sequencing-pcr-electrophoresis/a/polymerase-chain-reaction-pcr

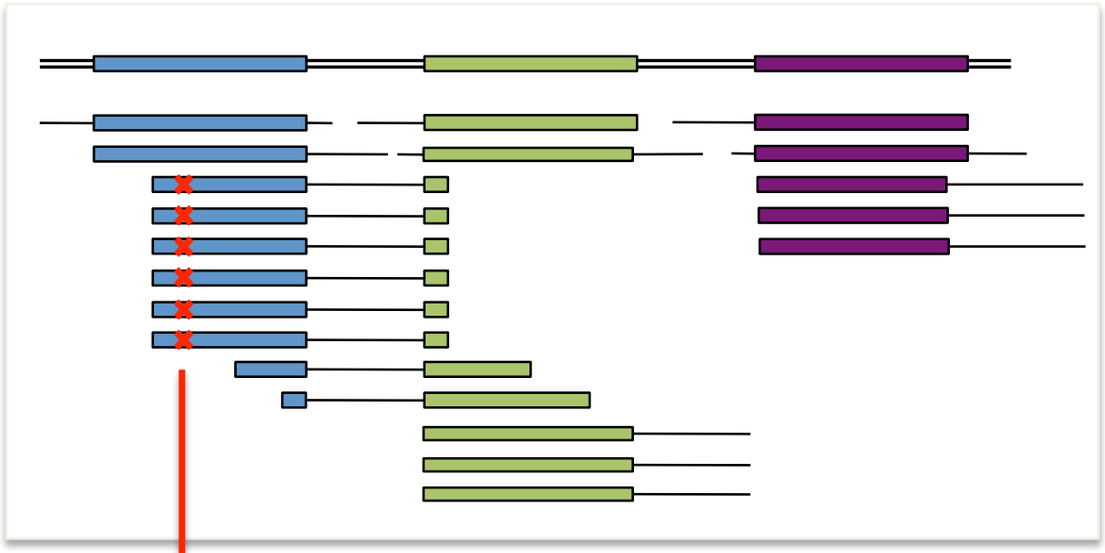


Flow cell tiles

http://www.slideshare.net/jandot/next-generation-sequencing-course-part-2-sequence-mapping
http://www.slideshare.net/cosentia/illumina-gaiix-for-high-throughput-sequencing

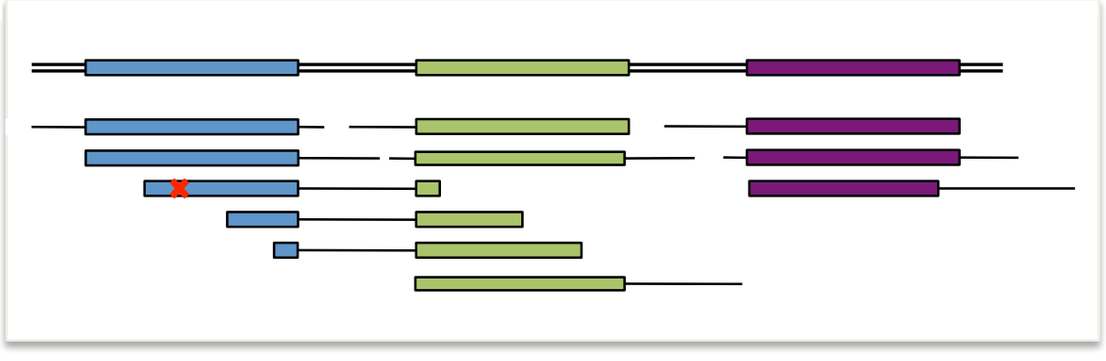
The reason why duplicates are bad

✖ = sequencing error propagated in duplicates



FP variant call (bad)

After marking duplicates, the GATK will only see :

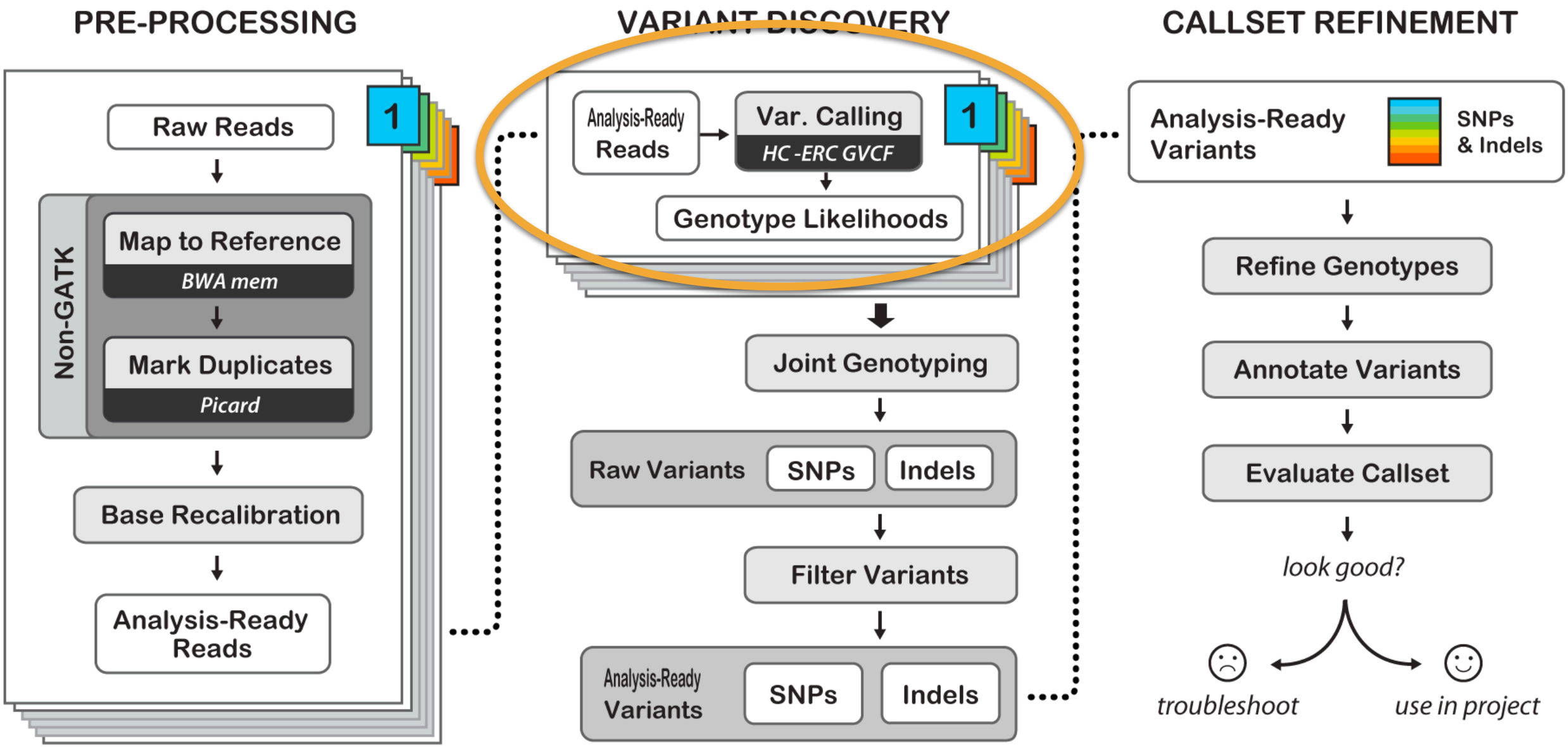


… and thus be more likely to make the right call

Easy to identify: duplicate reads have the same starting position and same CIGAR string

Why wouldn't we do this for GBS?

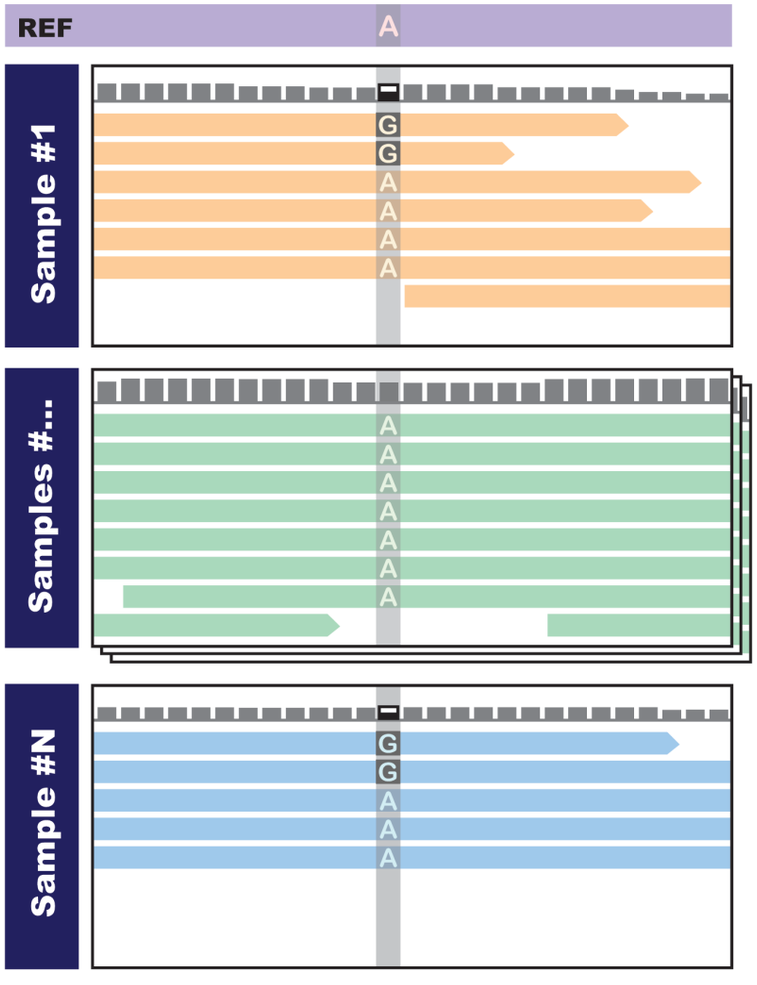
Variant Discovery



Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

Real mutations are hidden in the noise

Joint discovery empowers discovery at difficult sites

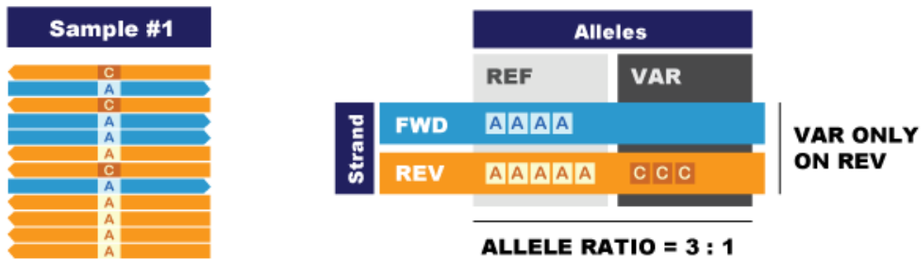


- If we analyze Sample #1 or Sample #N alone we are not confident that the variant is real

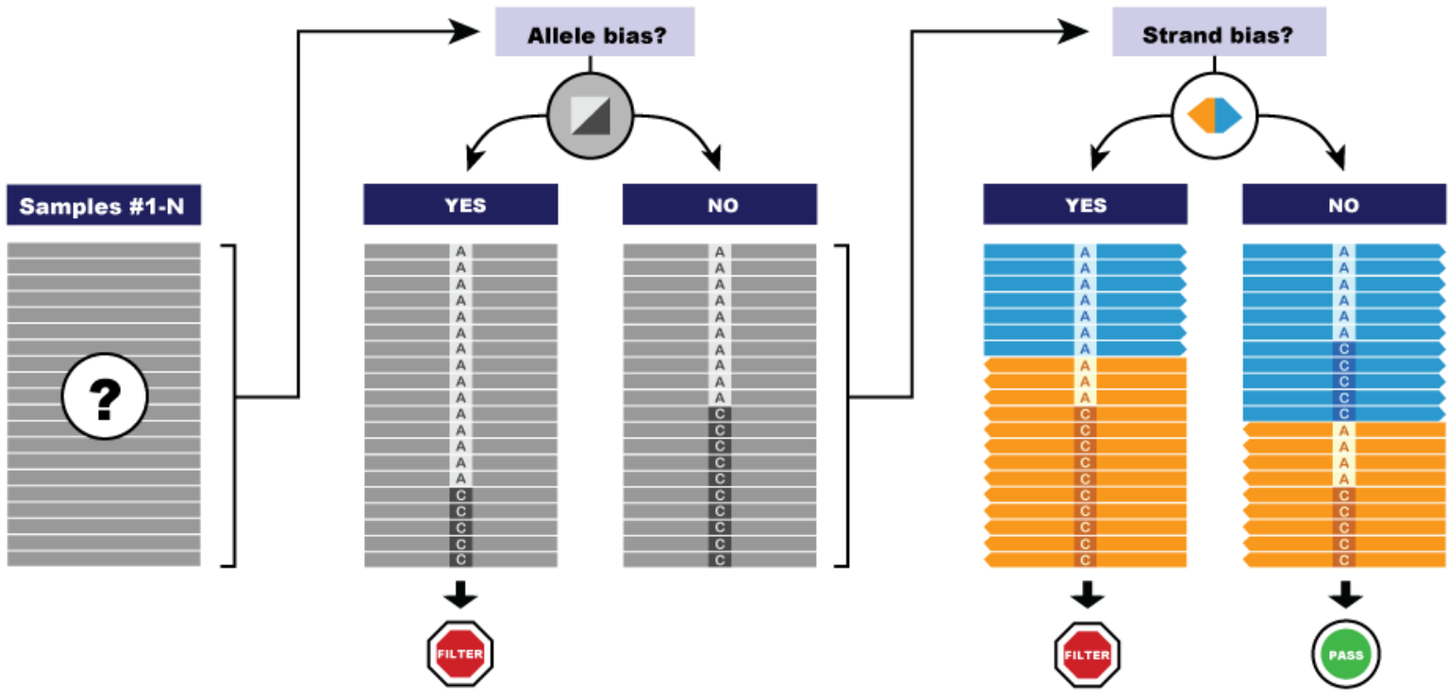
- If we see both samples then we are more confident that there is real variation at this site in the cohort

Joint discovery helps resolve bias issues

A. Single sample showing strand and allelic biases

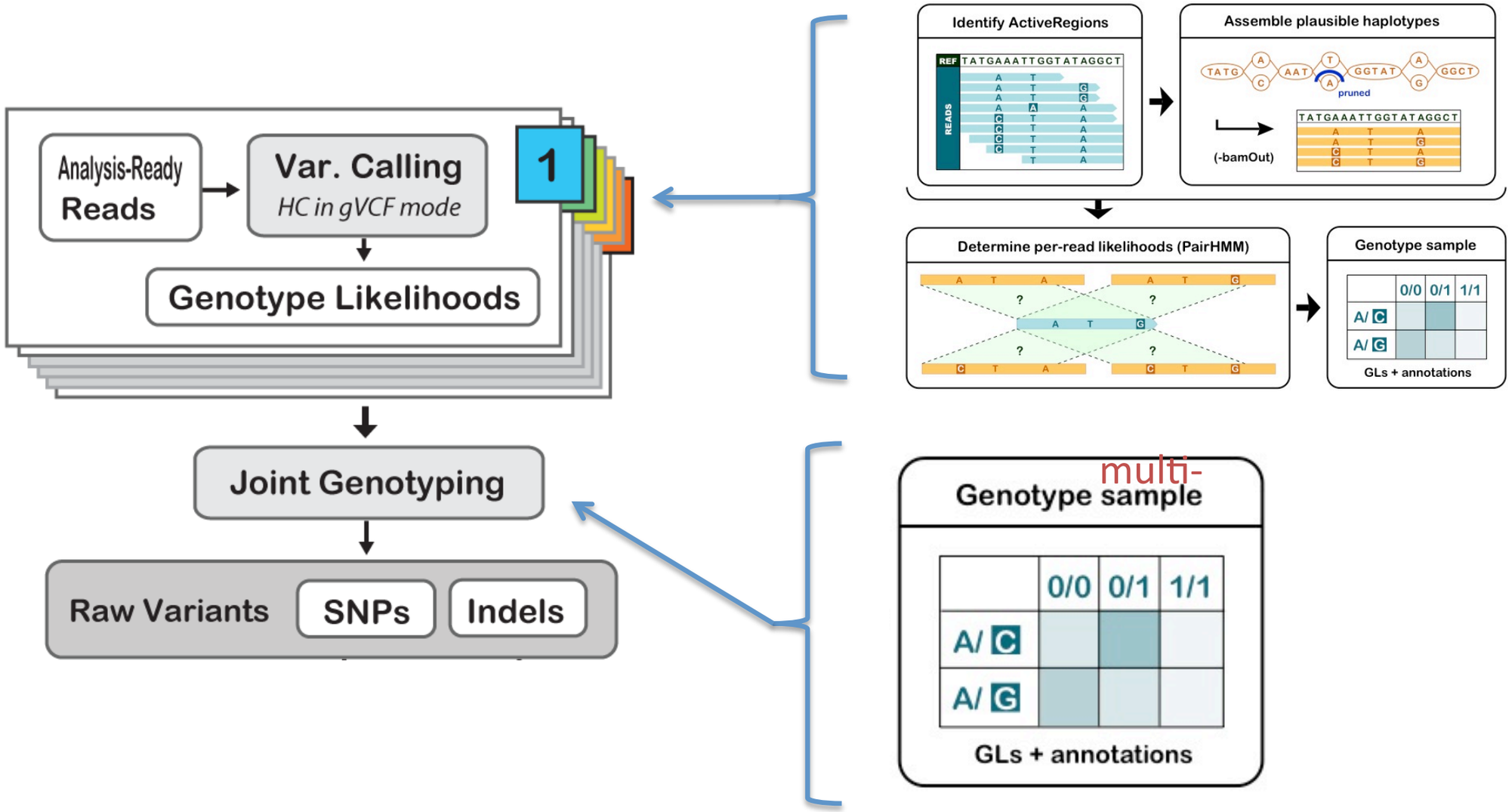


B. Decision process using evidence from multiple samples to filter out sites showing systematic biases



Joint Genotyping

Add a joint analysis step to take advantage of cohort / pop genetics data



Variant callers in GATK

- UnifiedGenotyper

  Call SNPs and indels separately by considering each variant locus independently

  - Accepts any ploidy
  - Pooled calling

- HaplotypeCaller

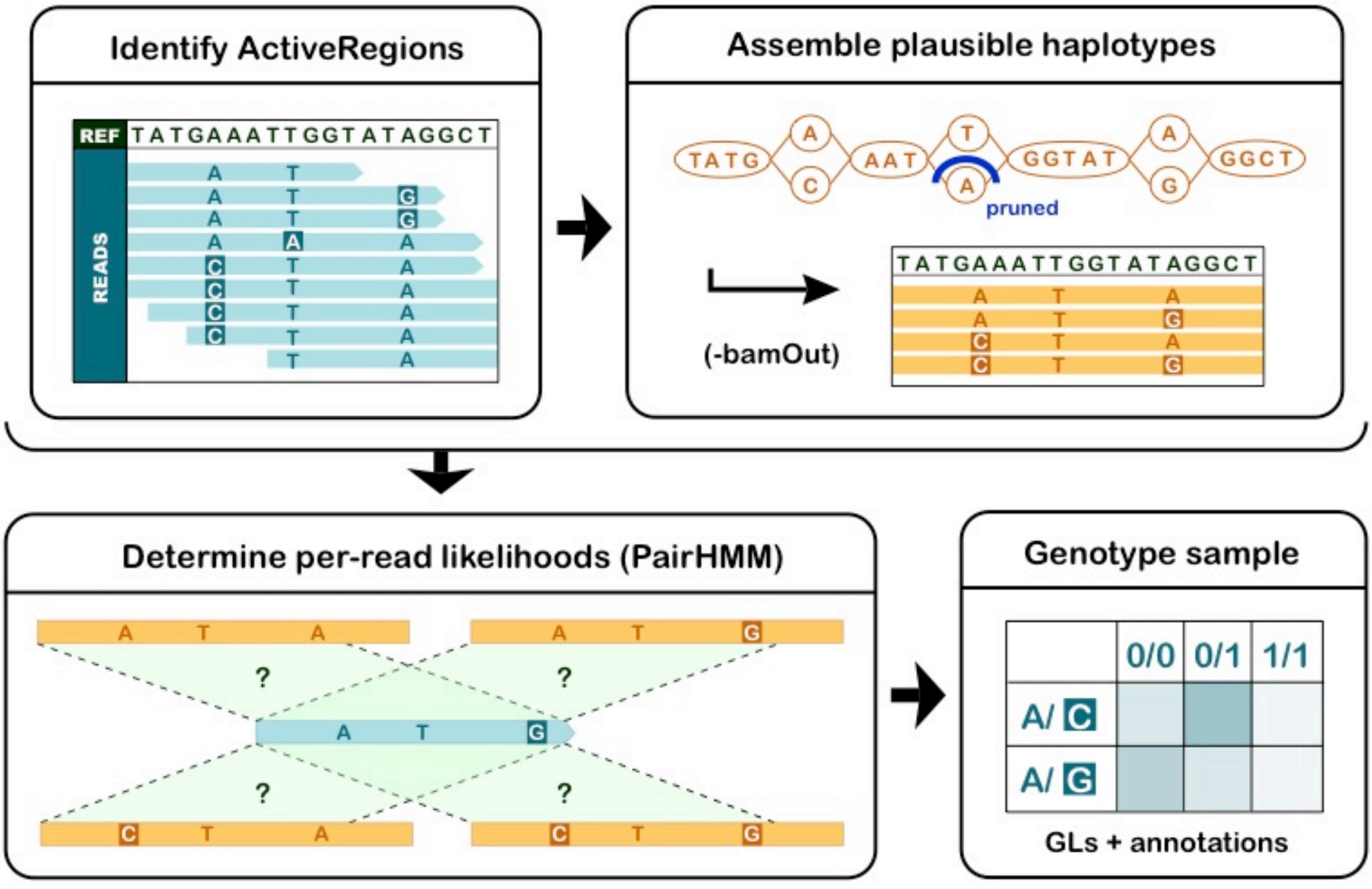
  Call SNPs, indels, and some SVs simultaneously by doing local re-assembly and considering haplotypes

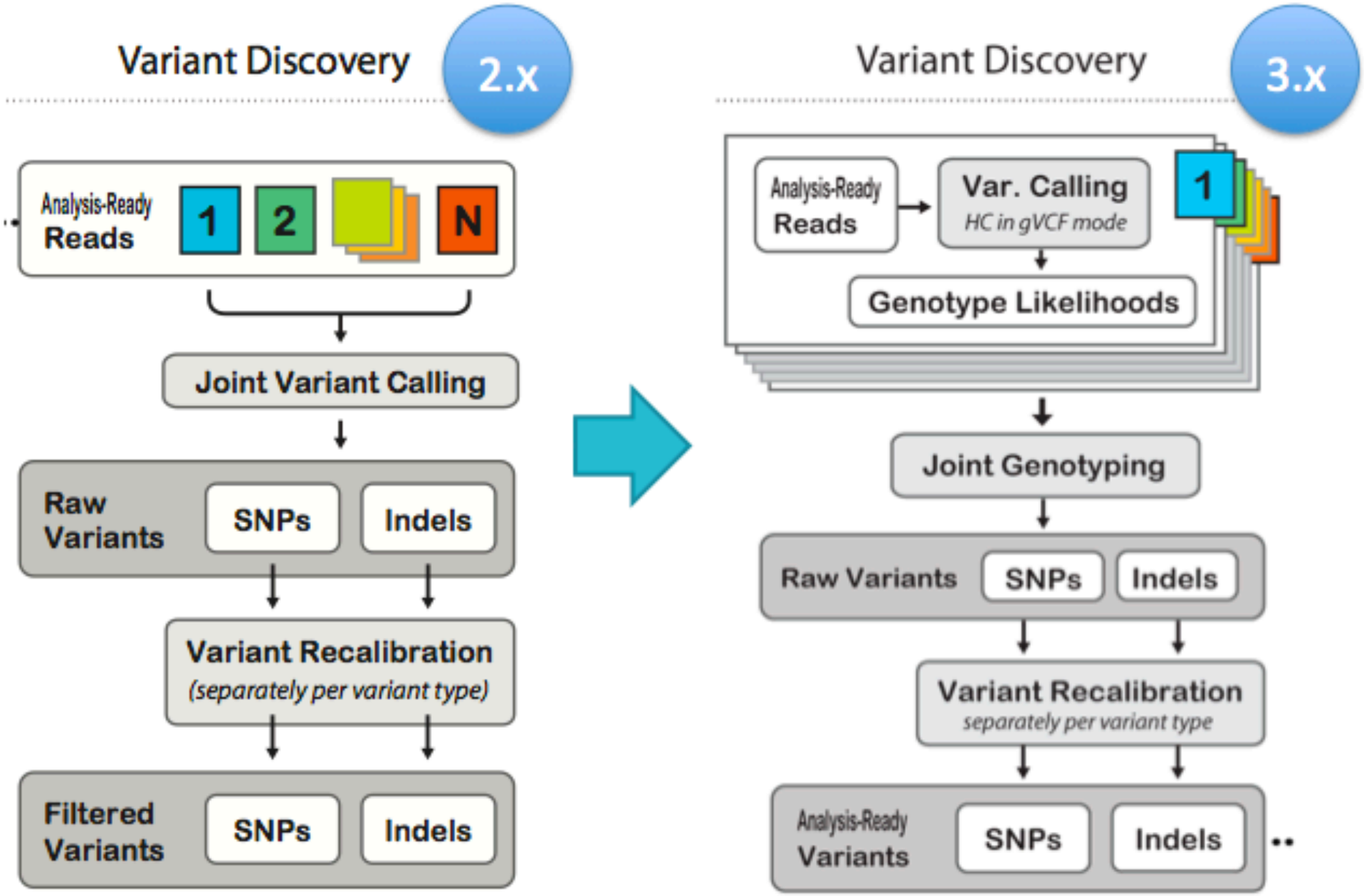
  - More accurate (esp. indels)
  - Reference confidence model
  - Replaces UG

HaplotypeCaller method overview

- Call SNPs, indels, and some SVs simultaneously by doing local re-assembly and considering haplotypes
  - Determine if a region has **potential variation**
  - Make **deBruijn assembly graph** of the region
  - Paths in the graph = **potential haplotypes** to evaluate
  - Calculate **data likelihoods** given the haplotypes (PairHMM)
  - **Identify variants** on most likely haplotypes
  - Compute **allele frequency distribution** to determine most likely allele count, and emit a variant call if appropriate
  - If emitting a variant, **assign genotype** to each sample

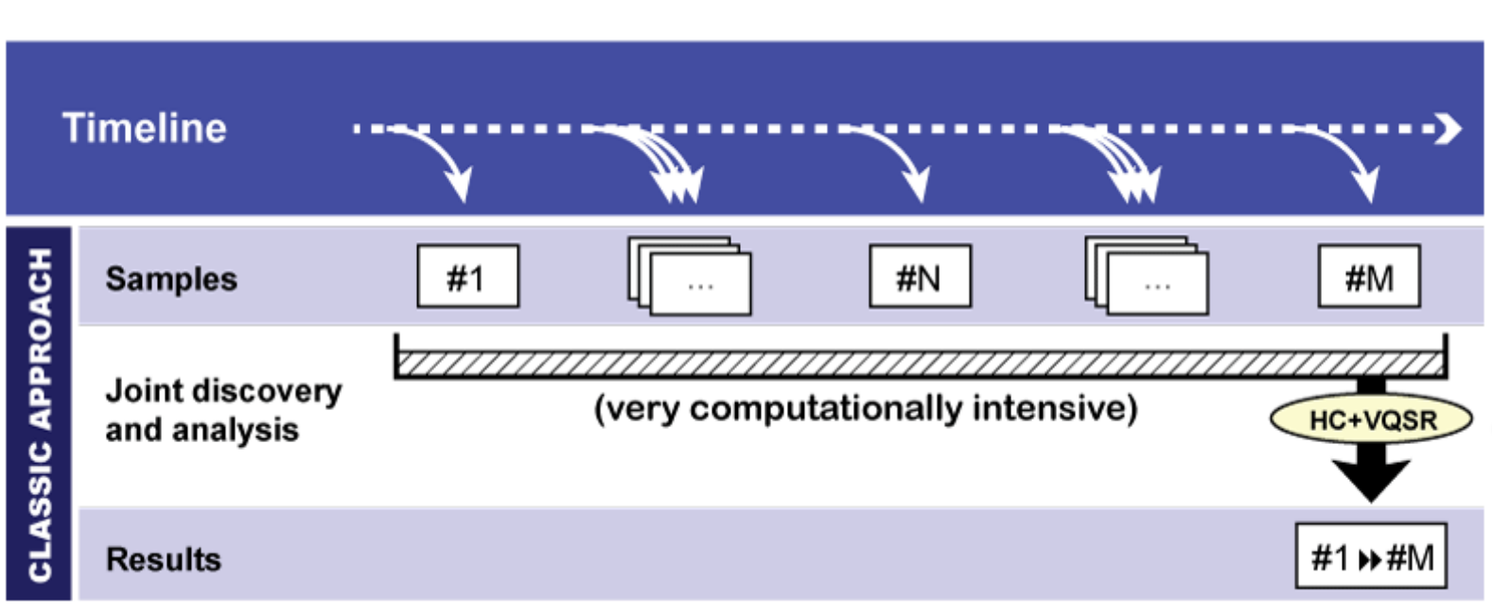
HC method illustrated

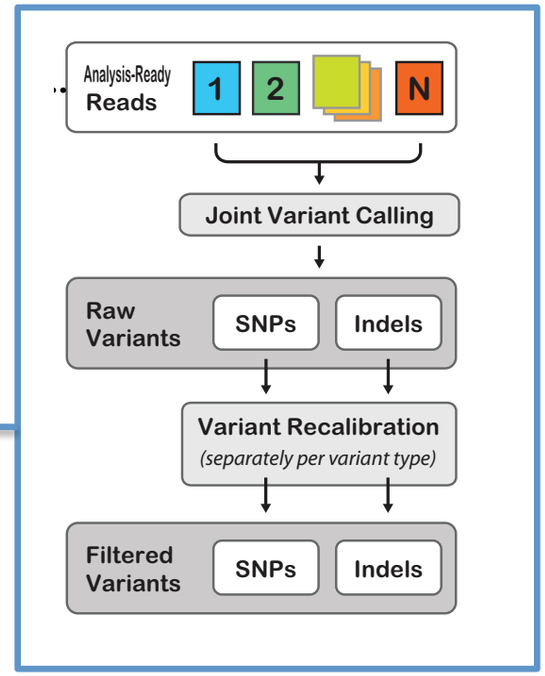


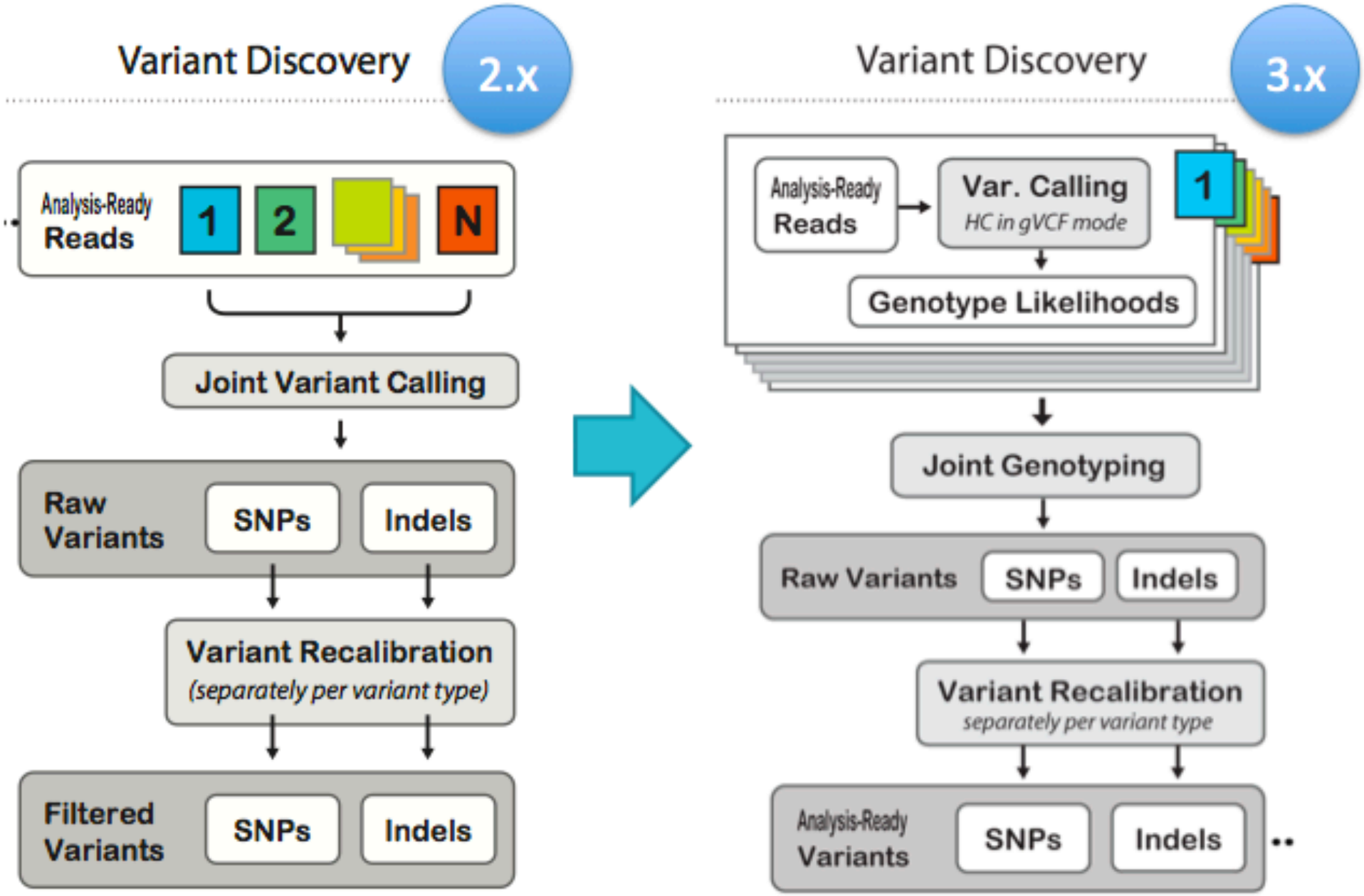

Variant Discovery
2.x
Analysis-Ready Reads
1
2
N
Joint Variant Calling
Raw Variants
SNPs
Indels
Variant Recalibration
(separately per variant type)
Filtered Variants
SNPs
Indels
Variant Discovery
3.x
Analysis-Ready Reads
Var. Calling
HC in gVCF mode
1
Genotype Likelihoods
Joint Genotyping
Raw Variants
SNPs
Indels
Variant Recalibration
separately per variant type
Analysis-Ready Variants
SNPs
Indels

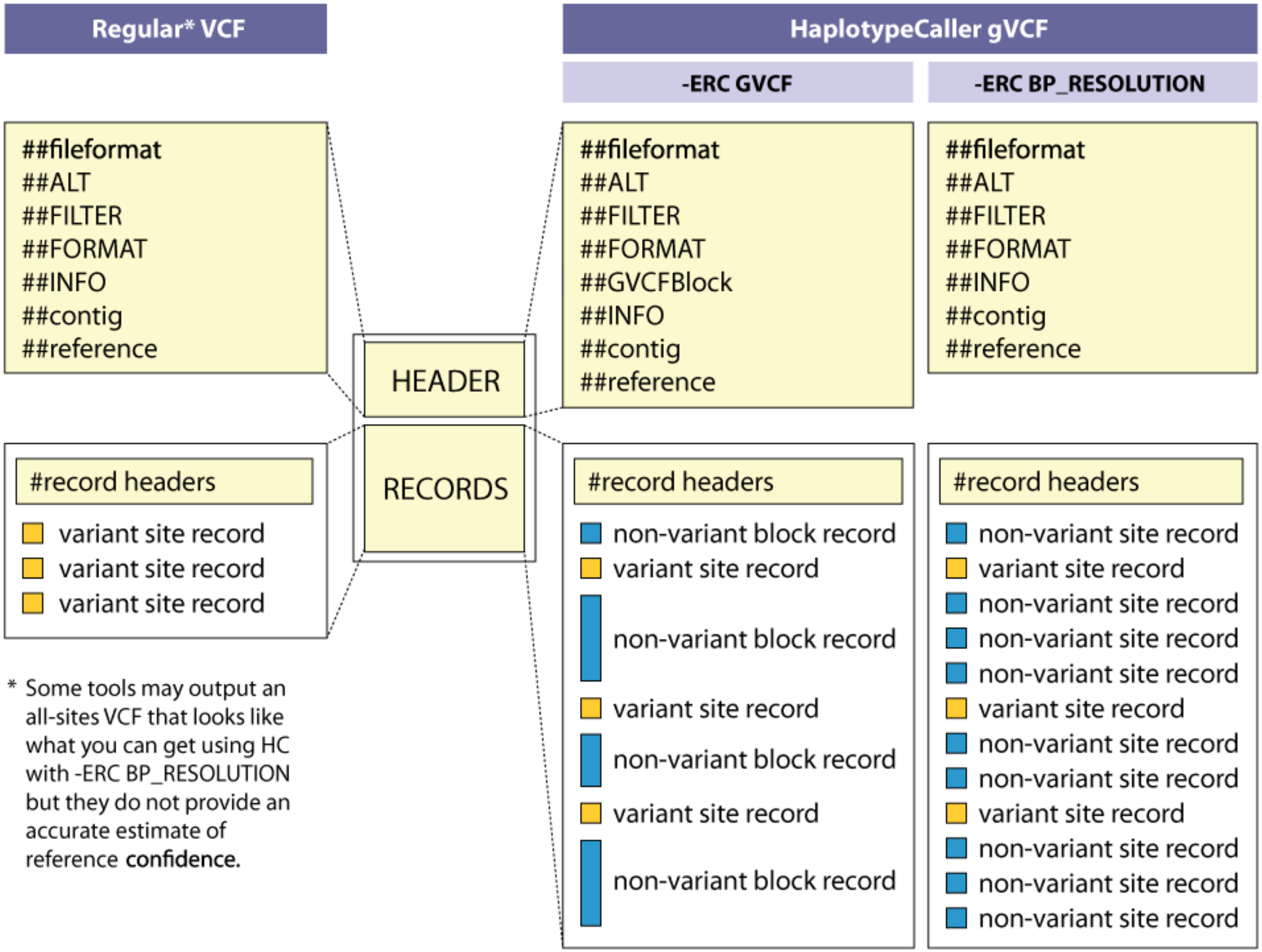
Problems with the “all together” approach

- Computing costs
- The “N+1 problem”



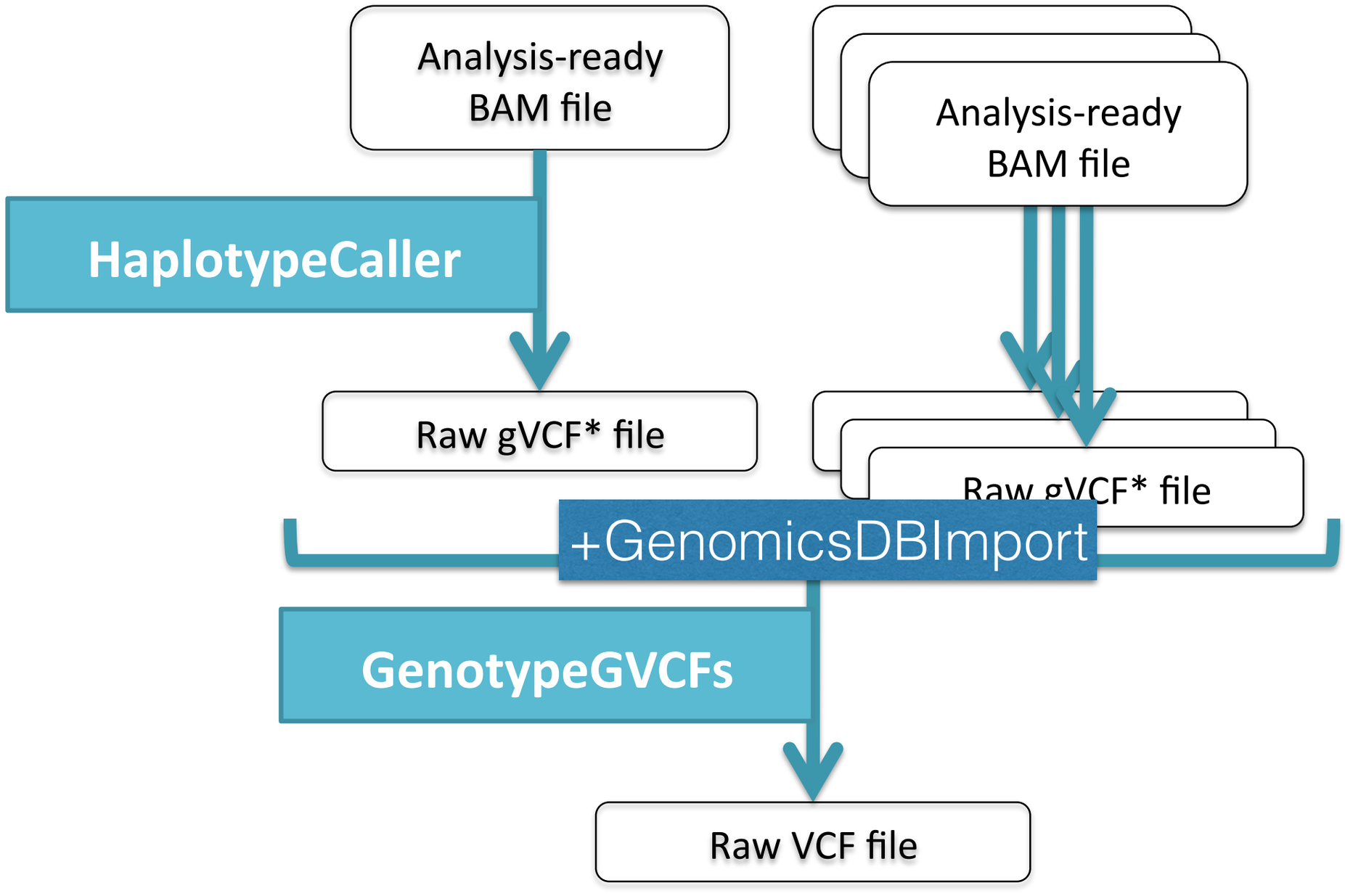



Variant Discovery
2.x
Analysis-Ready Reads
1
2
N
Joint Variant Calling
Raw Variants
SNPs
Indels
Variant Recalibration
(separately per variant type)
Filtered Variants
SNPs
Indels
Variant Discovery
3.x
Analysis-Ready Reads
Var. Calling
HC in gVCF mode
1
Genotype Likelihoods
Joint Genotyping
Raw Variants
SNPs
Indels
Variant Recalibration
separately per variant type
Analysis-Ready Variants
SNPs
Indels



* Some tools may output an all-sites VCF that looks like what you can get using HC with -ERC BP_RESOLUTION but they do not provide an accurate estimate of reference **confidence.**

Variant calling + joint genotyping workflow



VCF (Variant Call Format)

Headers

Variants

```
##INFO=<ID=MLEAC,Number=A,Type=Integer,Description="Maximum likelihood expectation (MLE) for the allele counts (not necessarily t
##INFO=<ID=MLEAF,Number=A,Type=Float,Description="Maximum likelihood expectation (MLE) for the allele frequency (not necessarily
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=RAW_MQandDP,Number=2,Type=Integer,Description="Raw data (sum of squared MQ and total depth) for improved RMS Mapping Q
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias"
##INFO=<ID=SOR,Number=1,Type=Float,Description="Symmetric Odds Ratio of 2x2 contingency table to detect strand bias">
##contig=<ID=chr_1,length=5000000>
##contig=<ID=chr_2,length=5000000>
##source=GenomicsDBImport
##source=GenotypeGVCFs
##source=HaplotypeCaller
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  Chinook.p1.i1.r400000   Chinook.p1.i2.r400000   Chinook.p
chr_1   163     .       T       C       179.59  .       AC=3;AF=0.375;AN=8;BaseQRankSum=-4.310e-01;DP=16;ExcessHet=1.0474;FS=0.00
chr_1   196     .       T       C       686.01  .       AC=8;AF=1.00;AN=8;DP=17;ExcessHet=3.0103;FS=0.000;MLEAC=7;MLEAF=0.875;MQ=
chr_1   296     .       T       A       514.38  .       AC=6;AF=1.00;AN=6;DP=14;ExcessHet=3.0103;FS=0.000;MLEAC=6;MLEAF=1.00;MQ=6
chr_1   726     .       A       C       714.91  .       AC=6;AF=1.00;AN=6;DP=20;ExcessHet=3.0103;FS=0.000;MLEAC=6;MLEAF=1.00;MQ=6
chr_1   755     .       T       A       987.52  .       AC=6;AF=1.00;AN=6;DP=29;ExcessHet=3.0103;FS=0.000;MLEAC=7;MLEAF=1.00;MQ=6
chr_1   804     .       T       C       173.03  .       AC=1;AF=0.125;AN=8;BaseQRankSum=-1.097e+00;DP=28;ExcessHet=3.0103;FS=2.63
chr_1   1052    .       G       T       1106.76 .       AC=8;AF=1.00;AN=8;DP=29;ExcessHet=3.0103;FS=0.000;MLEAC=8;MLEAF=1.00;MQ=6
chr_1   1420    .       G       A       1181.88 .       AC=8;AF=1.00;AN=8;DP=30;ExcessHet=3.0103;FS=0.000;MLEAC=8;MLEAF=1.00;MQ=6
chr_1   1492    .       C       G       645.47  .       AC=6;AF=0.750;AN=8;DP=26;ExcessHet=0.3218;FS=0.000;MLEAC=6;MLEAF=0.750;MQ
chr_1   1886    .       A       G       475.50  .       AC=4;AF=0.500;AN=8;BaseQRankSum=-4.310e-01;DP=22;ExcessHet=2.4304;FS=0.00
chr_1   1939    .       A       T       1122.43 .       AC=8;AF=1.00;AN=8;DP=29;ExcessHet=3.0103;FS=0.000;MLEAC=8;MLEAF=1.00;MQ=6
chr_1   3434    .       A       G       691.97  .       AC=6;AF=1.00;AN=6;DP=18;ExcessHet=3.0103;FS=0.000;MLEAC=6;MLEAF=1.00;MQ=6
chr_1   3462    .       A       C       543.54  .       AC=6;AF=1.00;AN=6;DP=14;ExcessHet=3.0103;FS=0.000;MLEAC=6;MLEAF=1.00;MQ=6
chr_1   3851    .       T       C       504.65  .       AC=4;AF=0.500;AN=8;DP=20;ExcessHet=0.1902;FS=0.000;MLEAC=4;MLEAF=0.500;MQ
chr_1   4139    .       A       T       1007.38 .       AC=8;AF=1.00;AN=8;DP=26;ExcessHet=3.0103;FS=0.000;MLEAC=8;MLEAF=1.00;MQ=6
chr_1   4267    .       A       G       303.58  .       AC=3;AF=0.375;AN=8;BaseQRankSum=-1.036e+00;DP=25;ExcessHet=1.0474;FS=0.00
chr_1   4455    .       G       C       187.46  .       AC=2;AF=0.250;AN=8;DP=20;ExcessHet=0.3218;FS=0.000;MLEAC=2;MLEAF=0.250;MQ
chr_1   4750    .       G       A       443.30  .       AC=2;AF=0.250;AN=8;DP=31;ExcessHet=0.3218;FS=0.000;MLEAC=2;MLEAF=0.250;MQ
chr_1   4780    .       G       A       144.69  .       AC=2;AF=0.250;AN=8;BaseQRankSum=1.28;DP=32;ExcessHet=0.3218;FS=0.000;MLEA
chr_1   5139    .       G       T       1078.75 .       AC=8;AF=1.00;AN=8;DP=28;ExcessHet=3.0103;FS=0.000;MLEAC=8;MLEAF=1.00;MQ=6
chr_1   5354    .       G       C       327.28  .       AC=3;AF=0.375;AN=8;BaseQRankSum=1.53;DP=26;ExcessHet=1.0474;FS=2.059;MLEA
```

Variant annotations provide key information to identify and remove artifacts!

VCF record for an A/G SNP at 22:49582364

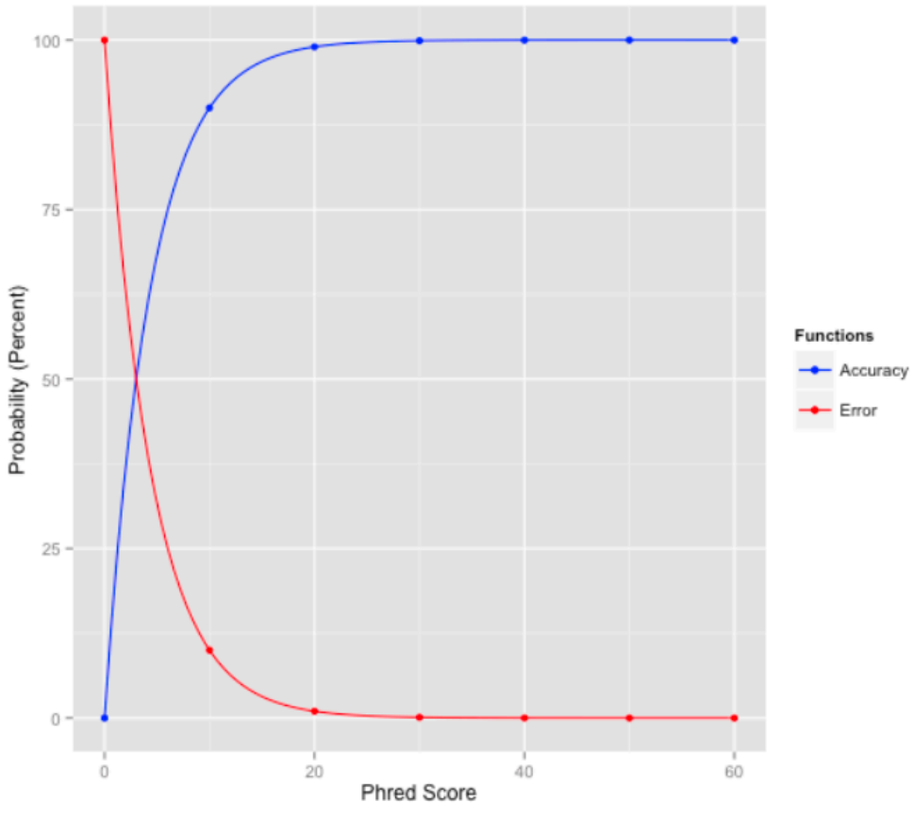
```
22 49582364    .    A    G    198.96    .
   AC=3;
   AF=0.50;
   AN=6;
   DP=87;
   MLEAC=3;
   MLEAF=0.50;
   MQ=71.31;
   MQ0=22;
   QD=2.29;
   SB=-31.76
   GT:DP:GQ    0/1:12:99.00    0/1:11:89.43    0/1:28:37.78
```

INFO field

AC	No. chromosomes carrying alt allele	MLEAF	Max likelihood AF
AN	Total no. of chromosomes	MQ	RMS MAPQ of all reads
AF	Allele frequency	MQ0	No. of MAPQ 0 reads at locus
DP	Depth of coverage	QD	QUAL score over depth
MLEAC	Max likelihood AC	SB	Estimated strand bias score

Phred-Scaled Quality Scores

$$Q = -10 \log E$$



Phred Quality Score	Error	Accuracy (1 - Error)
10	1/10 = 10%	90%
20	1/100 = 1%	99%
30	1/1000 = 0.1%	99.9%
40	1/10000 = 0.01%	99.99%
50	1/100000 = 0.001%	99.999%
60	1/1000000 = 0.0001%	99.9999%

Variant Filtering

- By default, GATK is very permissive. It will output false positive sites!

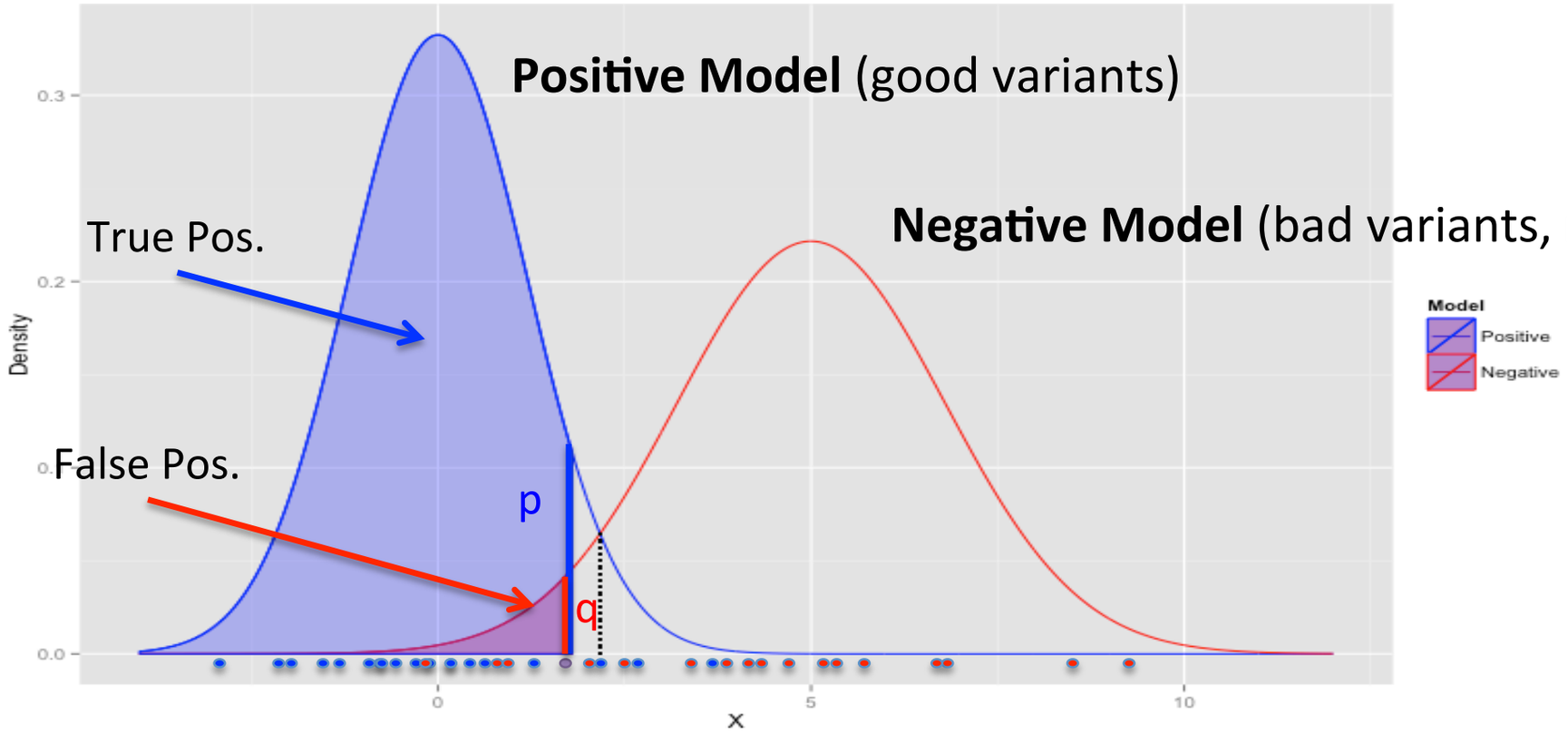
- Three(?) ways of filtering:
  - Variant recalibration (VQSR, not to be confused with BQSR!)
  - Intersect of diff. program SNP call sets
  - **Hard filtering**

How variant recalibration works

Train on high-confidence known sites to determine the probability that other sites are true or false

- Assume annotations tend to form **Gaussian clusters**
- Build a "Gaussian mixture model" from annotations of **known variants** in our dataset
- Score **all variants** by where their annotations lie relative to these clusters
- Filter base on **sensitivity to truth set**

Actually two models: positive and negative

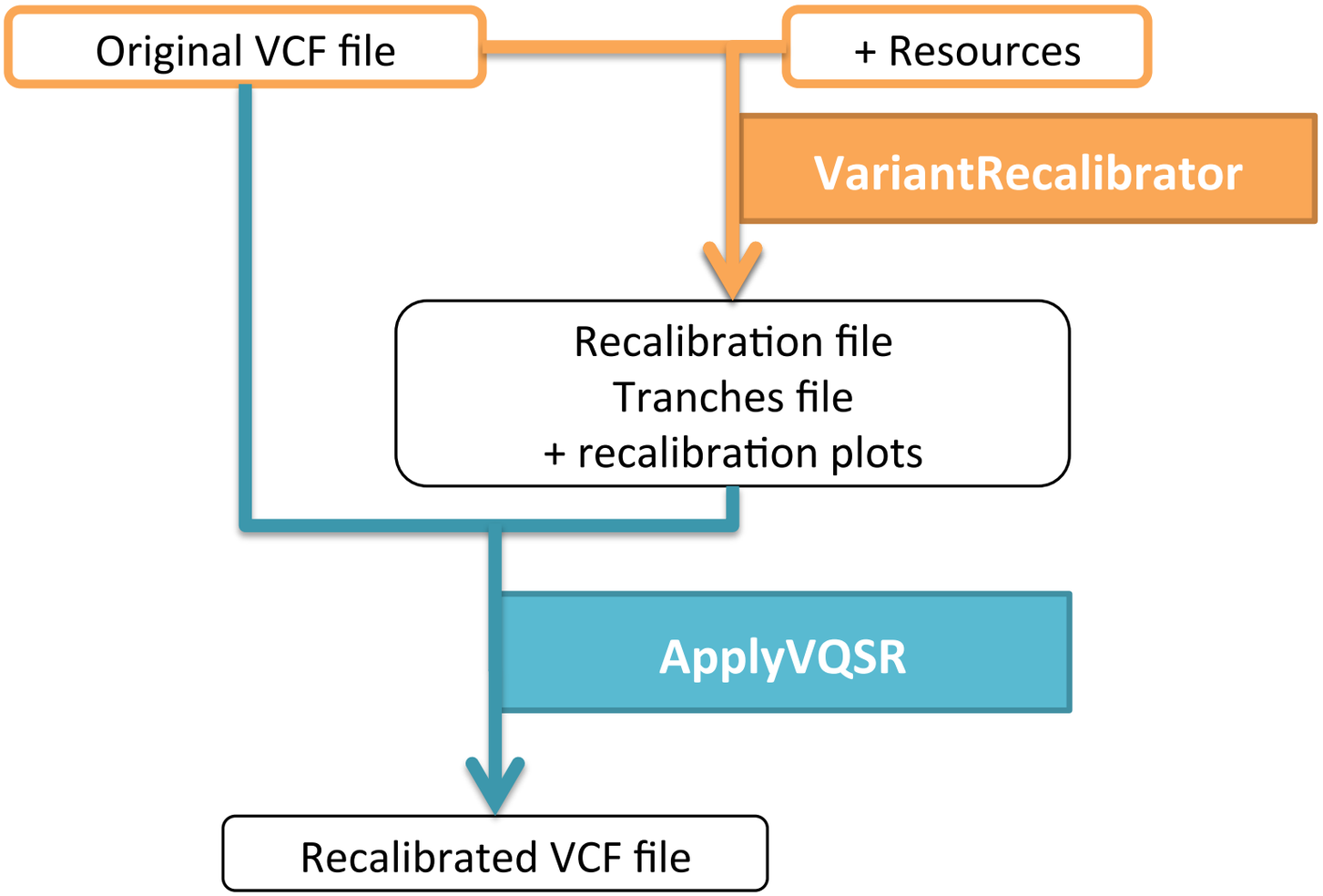


VQSLOD(x) = Log(p(x)/q(x))

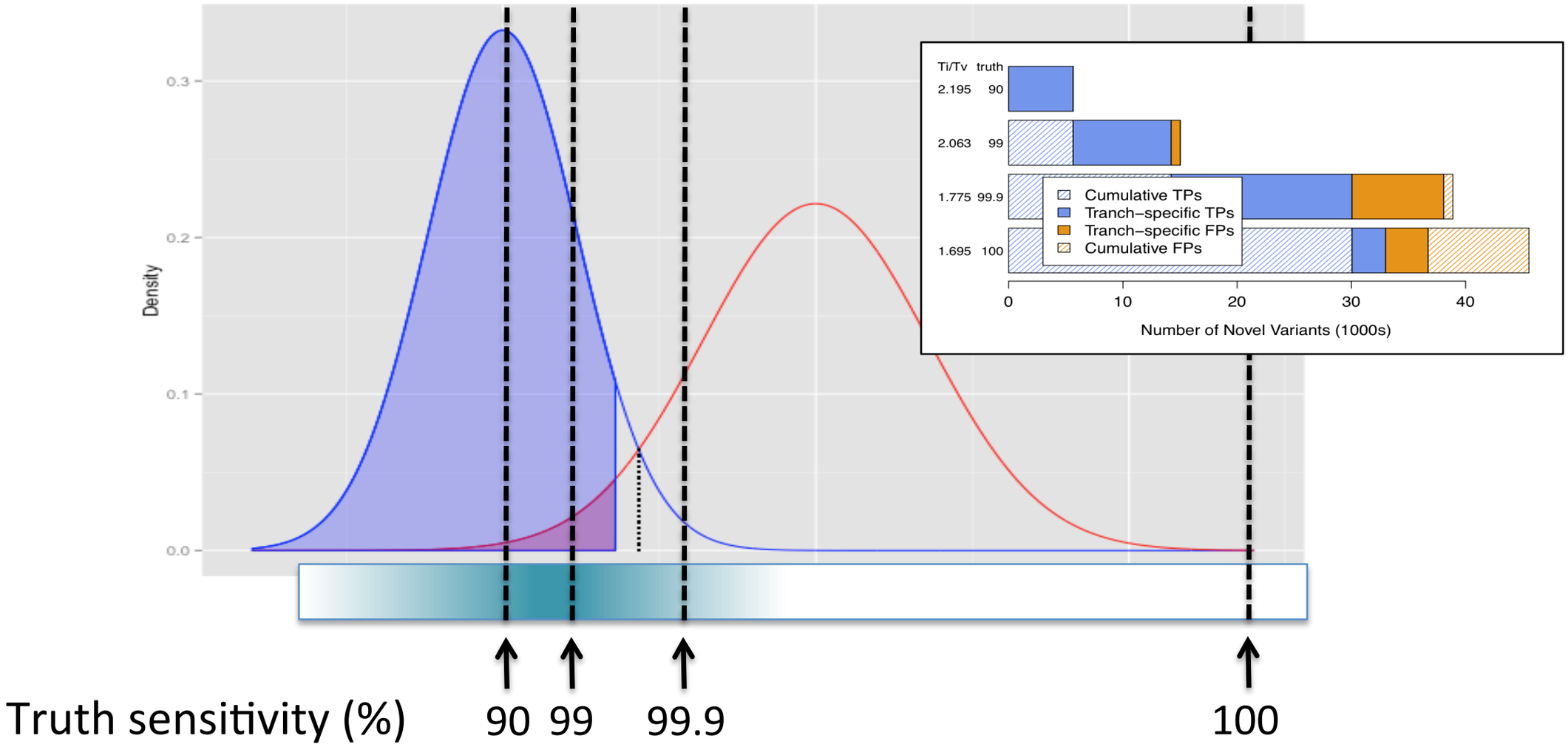
a new quality score (variant quality score log-odds) that takes into account various properties of the variant context not captured in the QUAL score

Done for each annotation X
then integrated into single overall VQSLOD

Step 1: VariantRecalibrator



Tranches : slices of sensitivity threshold values



Lower tranche = More stringent filtering

Where to get truth set?

- Reference sets (e.g. 1000 genomes) - but! Not available in most systems…

- Take your dataset, call SNPs, hard filter heavily, then use those as a truth set for recalibrating the unfiltered dataset.

Alternative Filtering

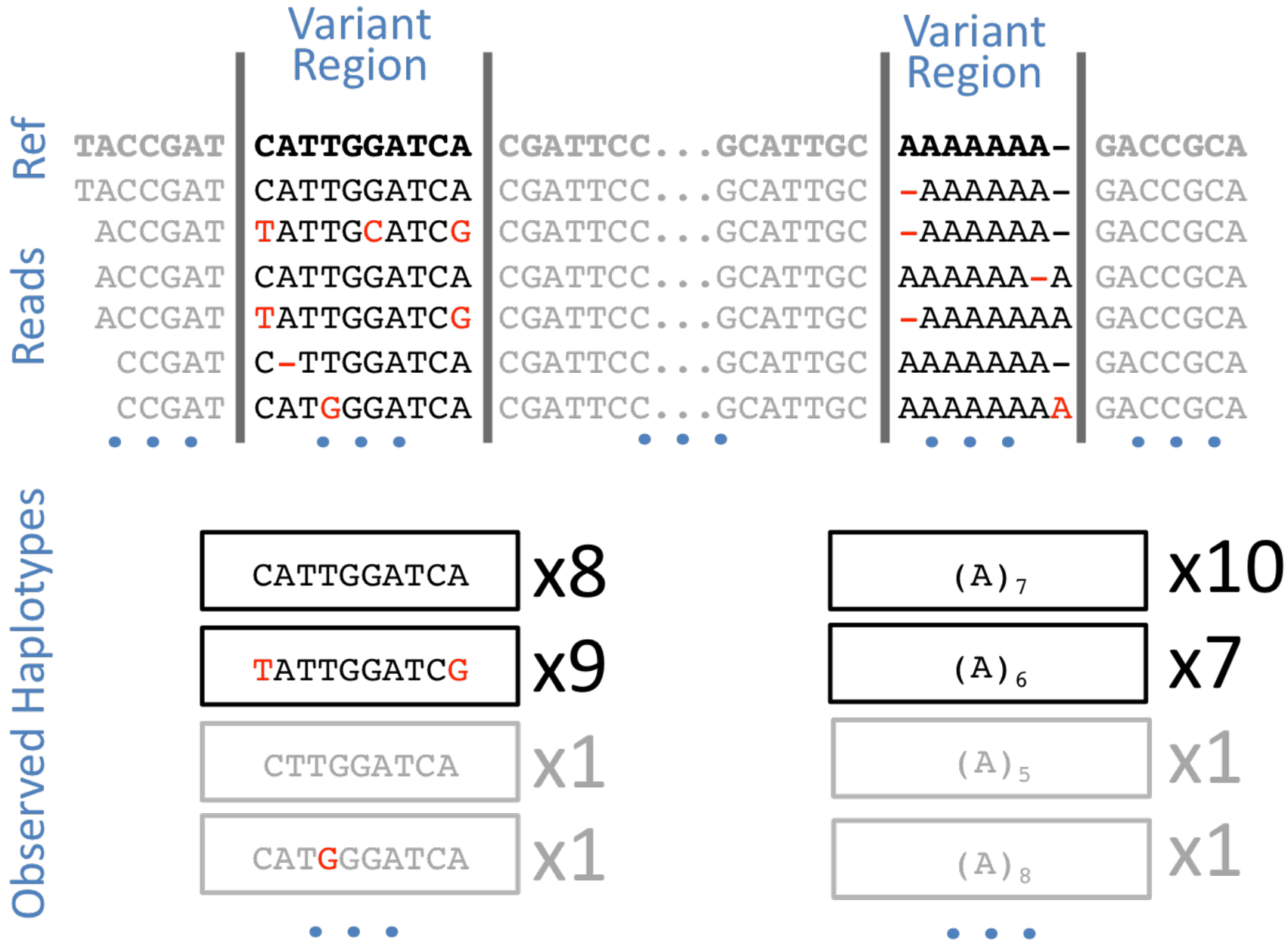
- Call SNPs using multiple programs and look for variants called in multiple programs, or combine the information in each.
    - e.g. BAYSIC, FREEBAYES, SAMTOOLS

- Time intensive…

FreeBayes

- Free and open source.
- Uses literal sequences of bases and haplotypes to call SNPs so is less affected by local alignment issues.
- Does not have “gvcf” n+1 method, although complicated work around exists.
- Nonetheless, generally faster than GATK, although RAM intensive.
- Outputs tons of annotations for *hard* filtering

FreeBayes


Variant Region
Variant Region
Ref
Reads
TACCGAT CATTGGATCA CGATTCC...GCATTGC AAAAAAA- GACCGCA
TACCGAT CATTGGATCA CGATTCC...GCATTGC -AAAAAA- GACCGCA
ACCGAT TATTGCATCG CGATTCC...GCATTGC -AAAAAA- GACCGCA
ACCGAT CATTGGATCA CGATTCC...GCATTGC AAAAAA-A GACCGCA
ACCGAT TATTGGATCG CGATTCC...GCATTGC -AAAAAAA GACCGCA
CCGAT C-TTGGATCA CGATTCC...GCATTGC AAAAAAA- GACCGCA
CCGAT CATGGATCA CGATTCC...GCATTGC AAAAAAAA GACCGCA
Observed Haplotypes
CATTGGATCA x8
TATTGGATCG x9
CTTGGATCA x1
CATGGATCA x1
(A)7 x10
(A)6 x7
(A)5 x1
(A)8 x1

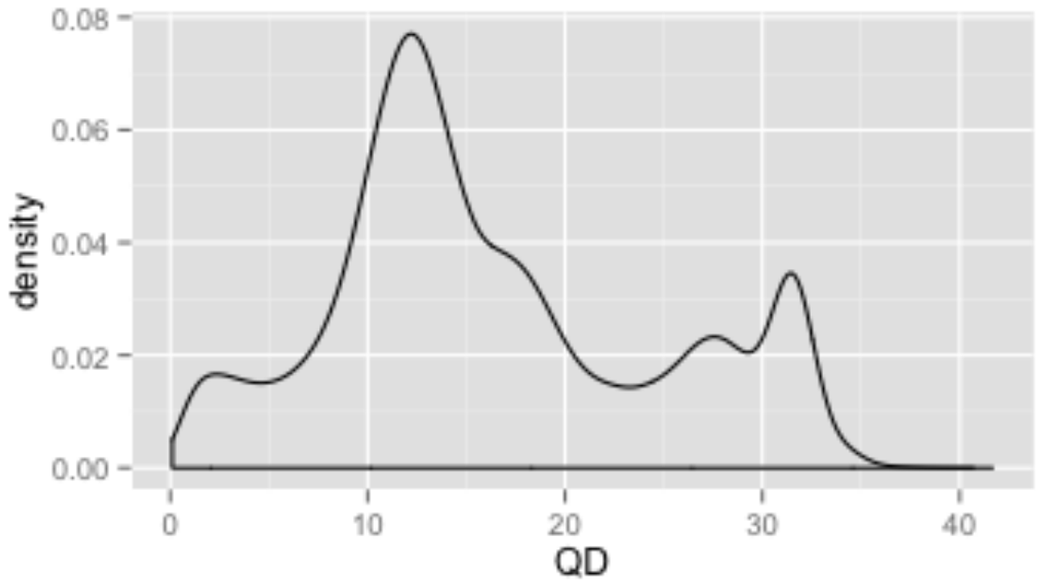
ANGSD

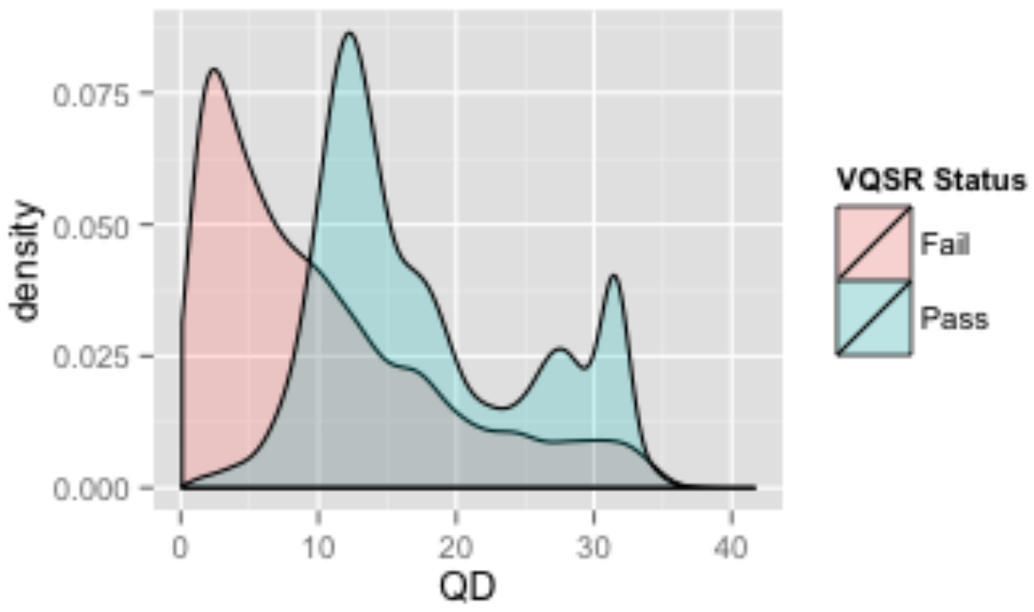
- Calls SNPs based on reads per site, no realignment.
- **Outputs genotype likelihoods.**
- Links with algorithms that use likelihood.
- Questionable with high diversity systems.
- Great for low coverage data**

Hard Filters

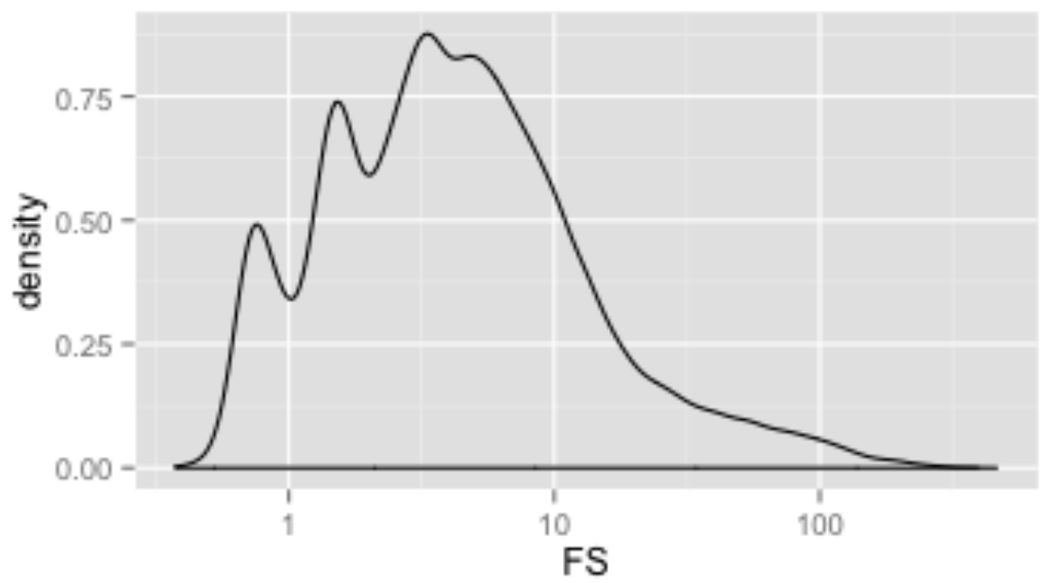
- User defined thresholds for each site. But where to make cut offs?
  - Mapping quality high enough
  - Depth above a minimum but not too high
  - Minor allele frequency above a minimum
  - Heterozygosity not too high

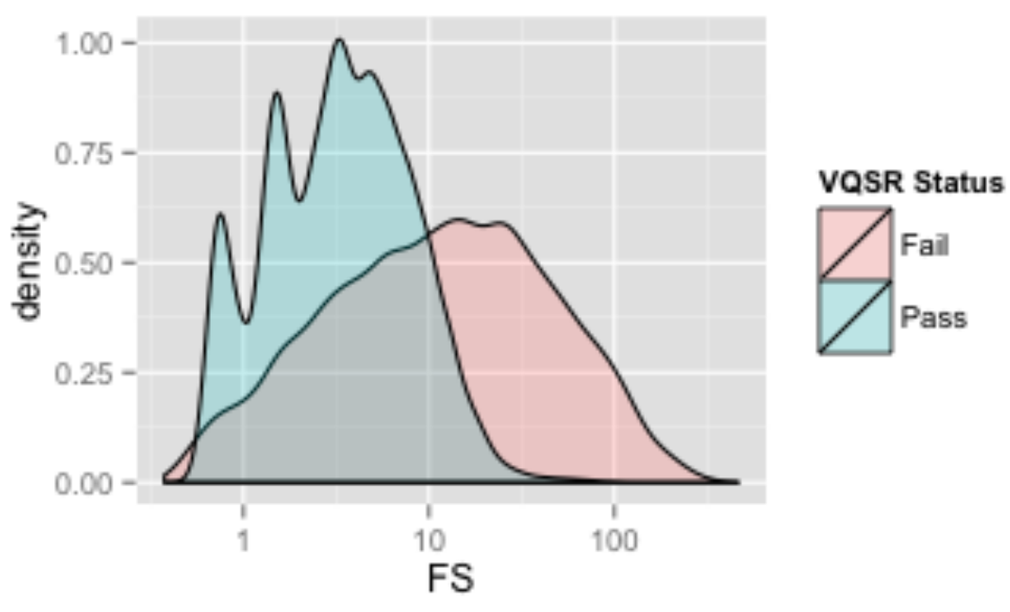
QualityByDepth





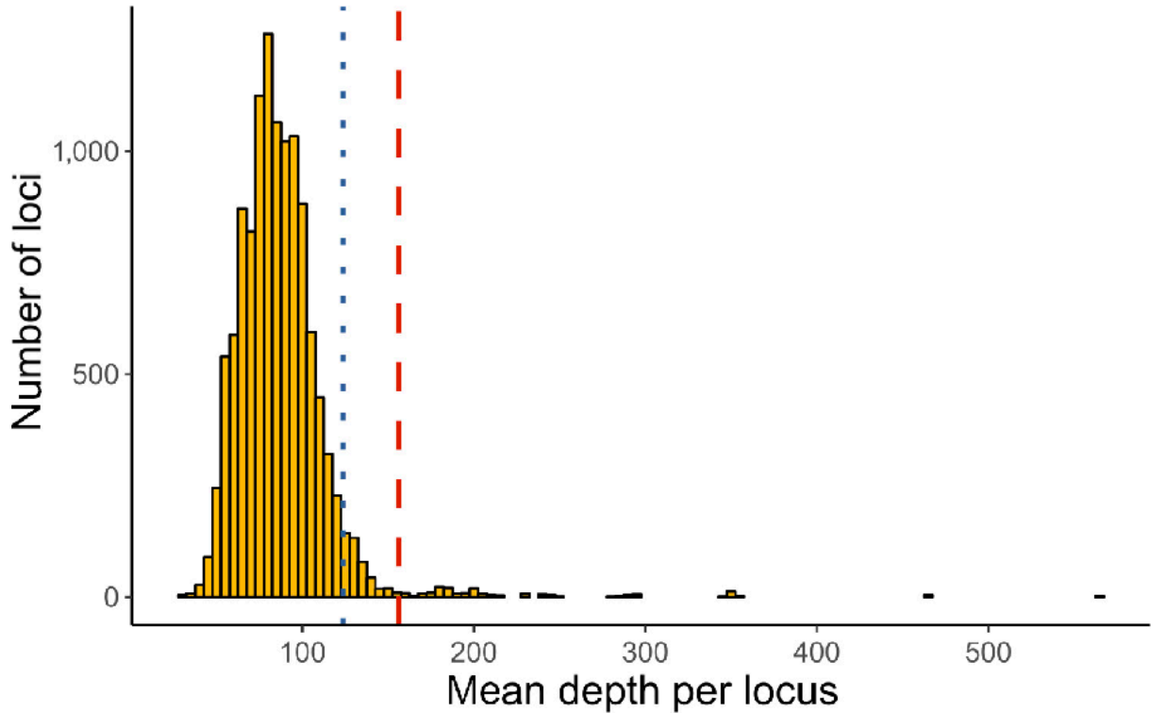
FisherStrandRatio (Strand Bias)

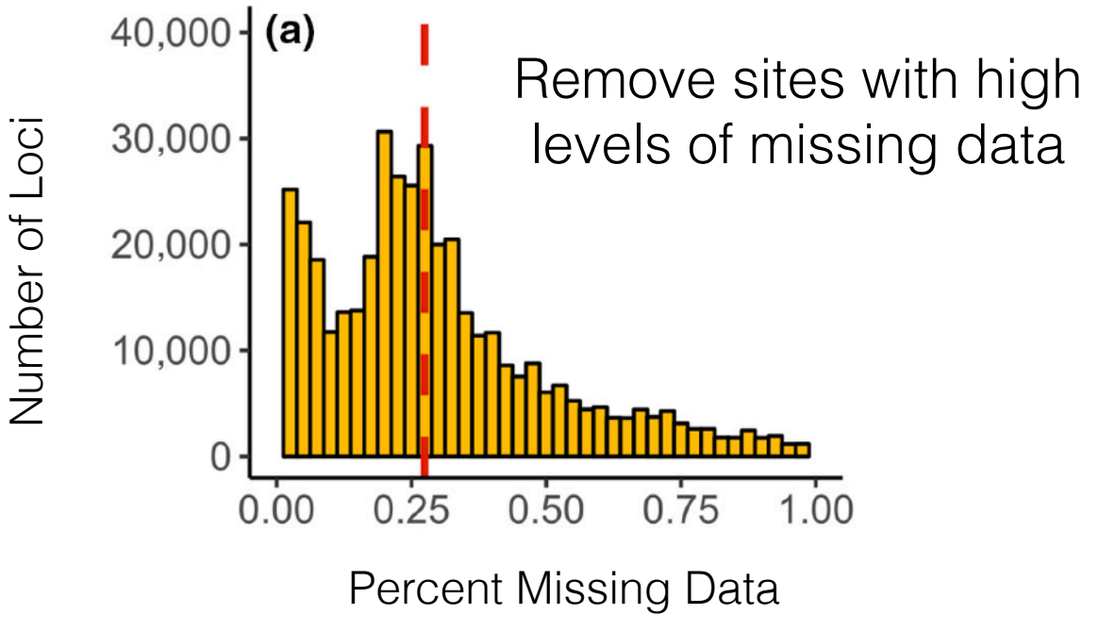




Visualizing your data to inform filtering cutoffs

Identify rare, duplicated loci



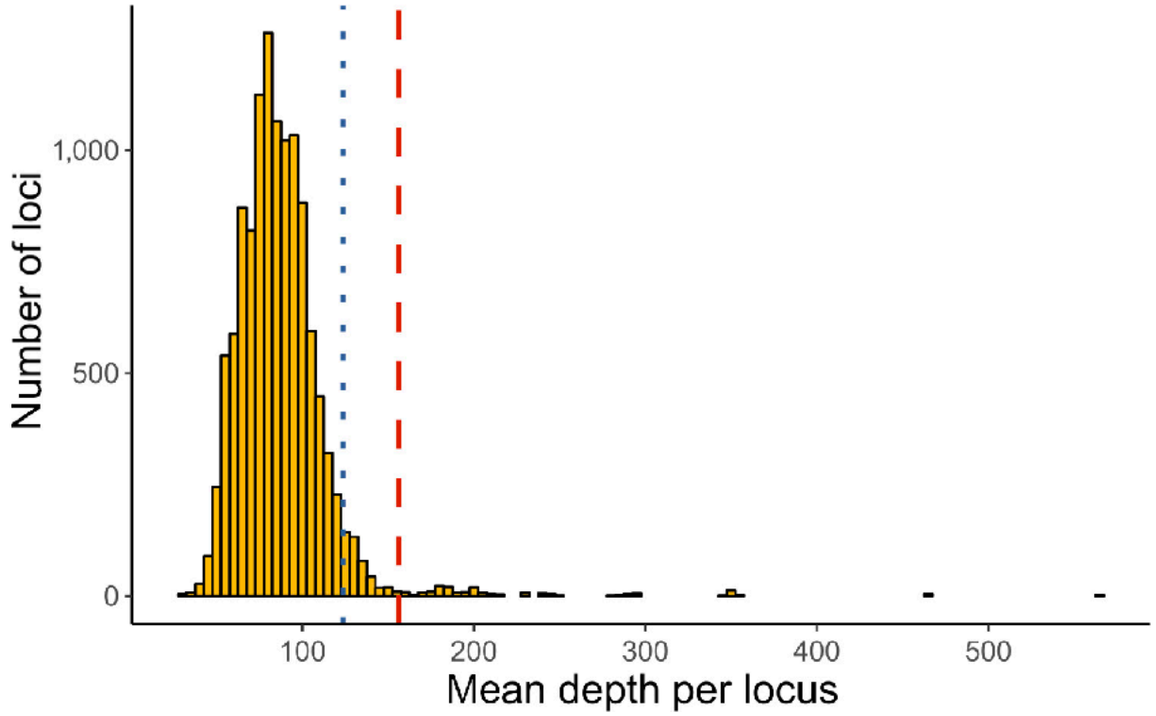


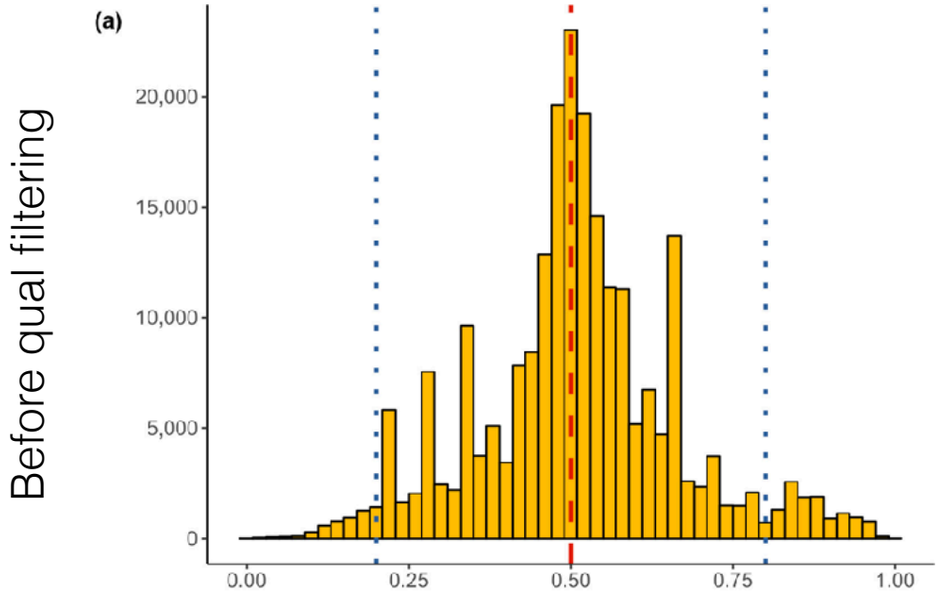
Remove sites with high levels of missing data

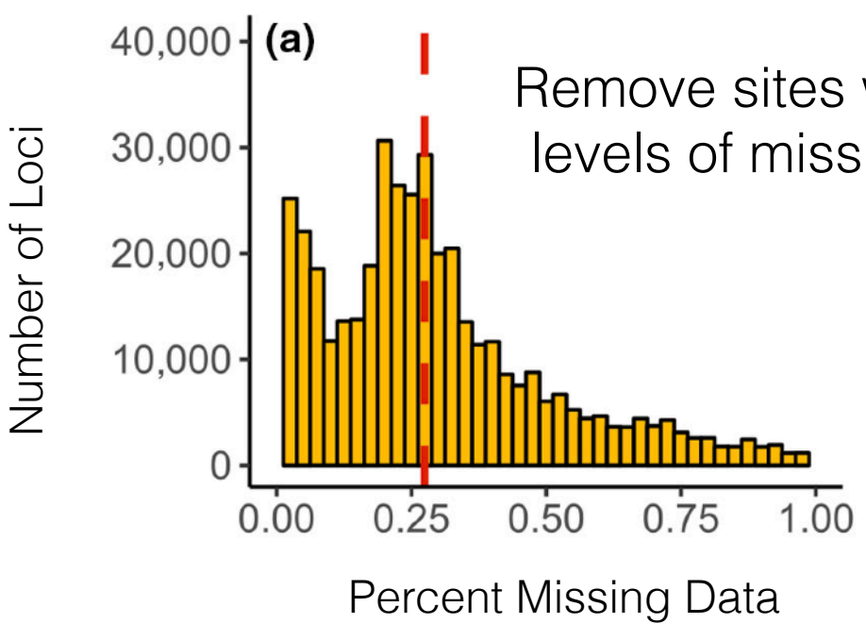
O'leary et al., 2018

Visualizing your data to inform filtering cutoffs

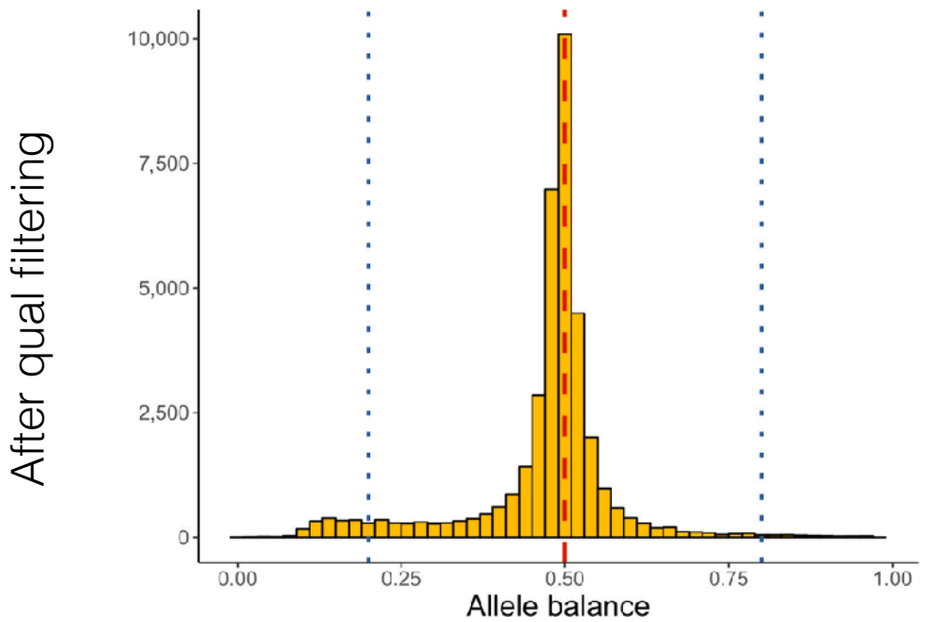
Identify rare, duplicated loci







Remove sites with high levels of missing data



O'leary et al., 2018