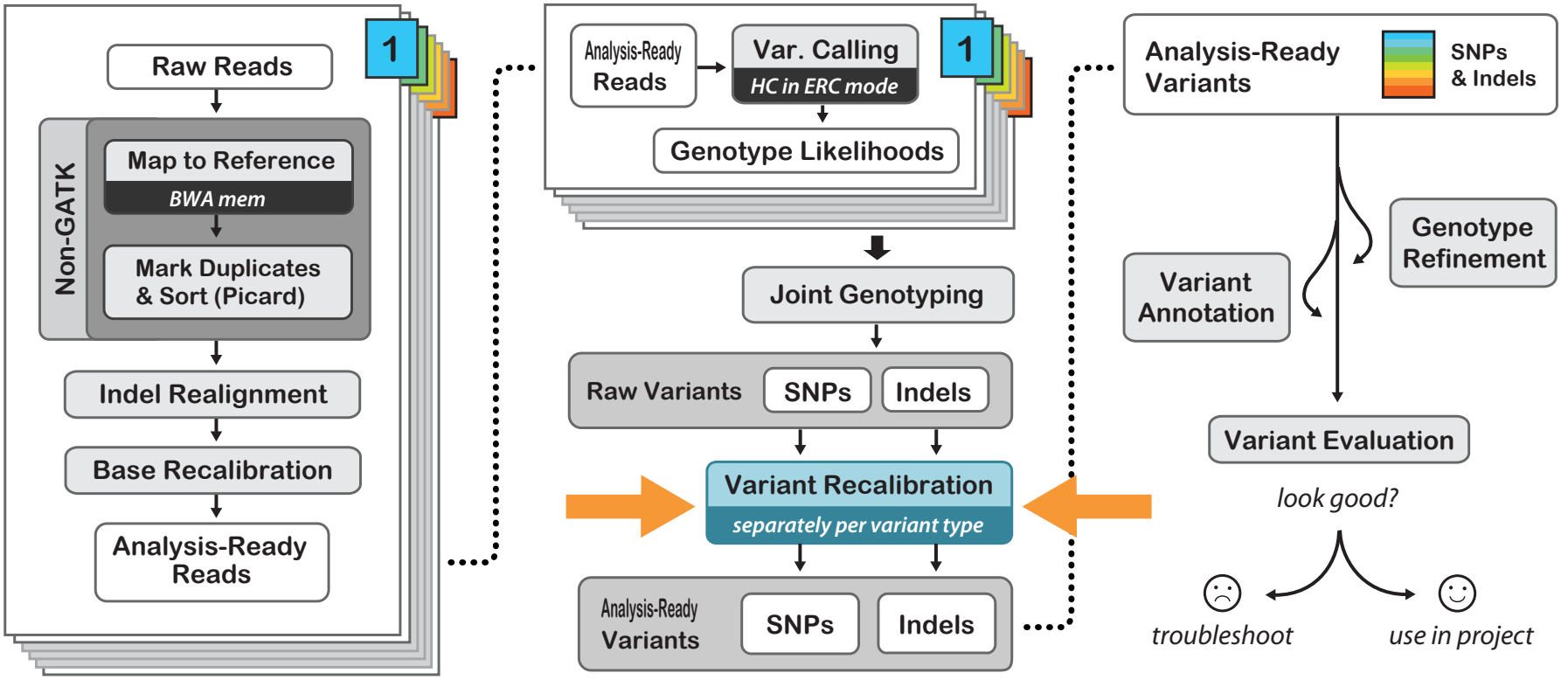
Variant Quality Score Recalibration

Assigning accurate confidence scores to each putative mutation call

You are here in the GATK Best Practices workflow for **germline variant discovery**



Raw, high-sensitivity callsets contain many false positives

- Mutation calling algorithms are very permissive by design
- How to filter?
  - Hand-tuned hard-filtering requires time and expertise
  - Better to learn what the filters should be from the data itself
- Must enable analysts to trade off sensitivity and specificity depending on project goals

☑ **Building a model of what true genetic variation looks like will allow us to rank-order variants based on their likelihood of being real**

From annotations to mixture models

- Each variant has a diverse set of statistics associated with it.
- These annotations tend to form Gaussian clusters
- We can fit a “Gaussian mixture model” to the annotations known variants in our dataset.
- Any new variant can be scored by evaluating the associated annotations in this model.

Variant annotations are the “features” of the model

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    0/1:11:89    0/1:28:37
```

INFO field

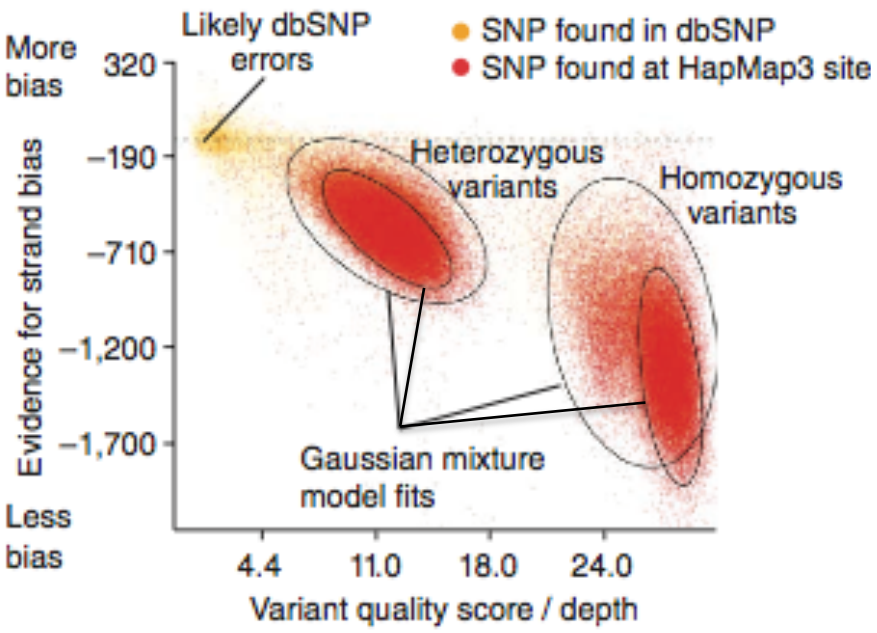
AC	No. chromosomes carrying alt allele	MLEAF	Max likelihood AF
AN	Total no. of chromosomes	MQ	RMS MAPQ of all reads
AF	Allele frequency	MQ0	No. of MAPQ 0 reads at locus
DP	Depth of coverage	QD	QUAL score over depth
MLEAC	Max likelihood AC		

Note that VQSR will only look at INFO annotations;
sample-level annotations (genotype, AD etc) are not used.

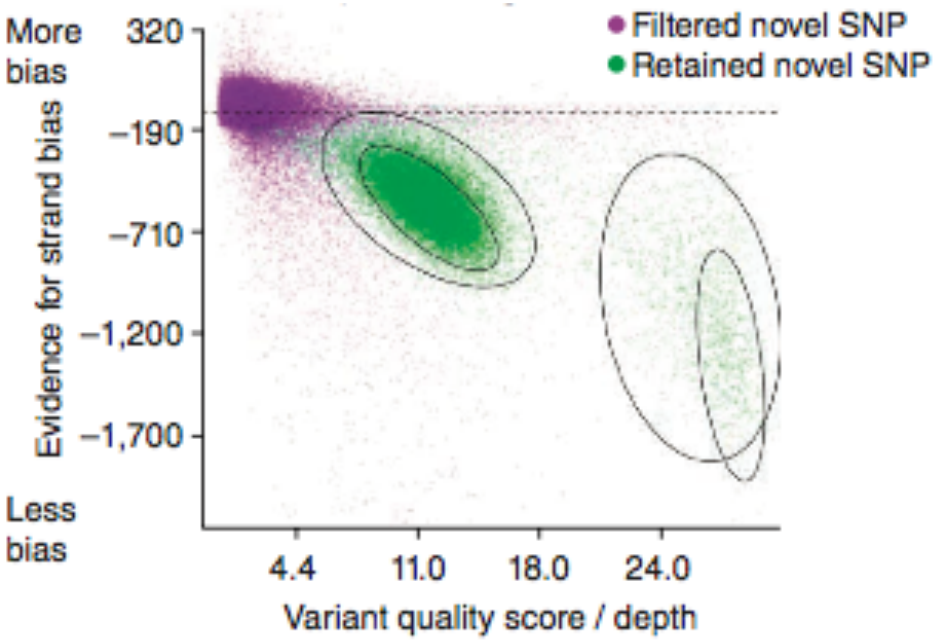
Two steps: (1) train a model then (2) apply to callset

Basic idea: training on high-confidence known sites to determine the probability that other sites are true

(1) Train model using HapMap



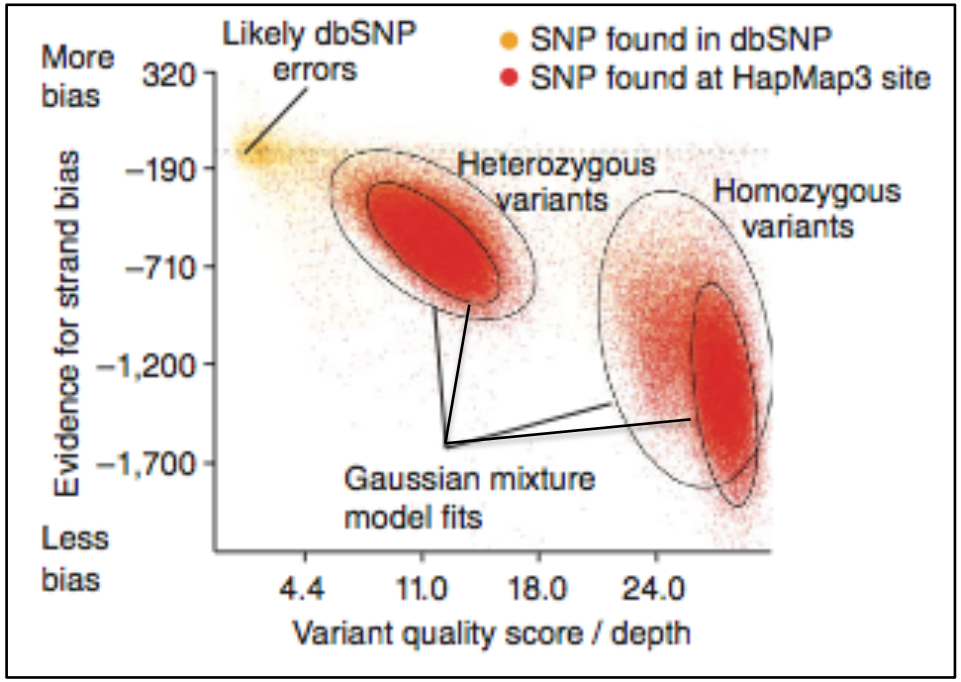
(2) Apply model to callset



Modified from DePristo et al. Nature Genetics. 2011

(1) Training the model

(1) Train model using e.g. HapMap



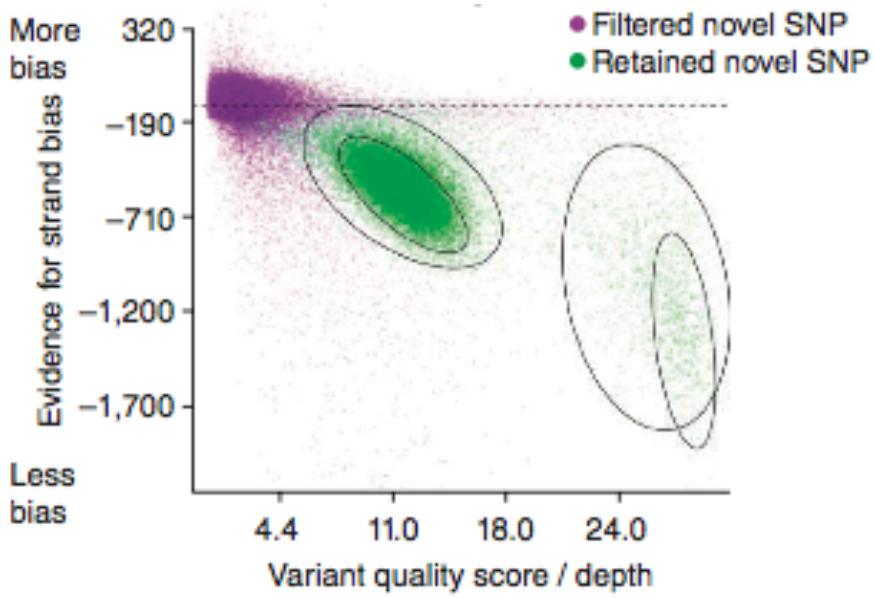
- We choose a training set
- Variants that are both in the training set and in our callset are selected.
- We train the model using the annotations of the selected variants

- This tells us **what good variants look like**
- A similar model for the variants in our callset that least look like good variants is also created (bad model, no biscuit!)
- All variants can now be ranked based on the ratio between their scores in the good model and the bad model (= VQSLOD)

(2) Applying the model to our callset

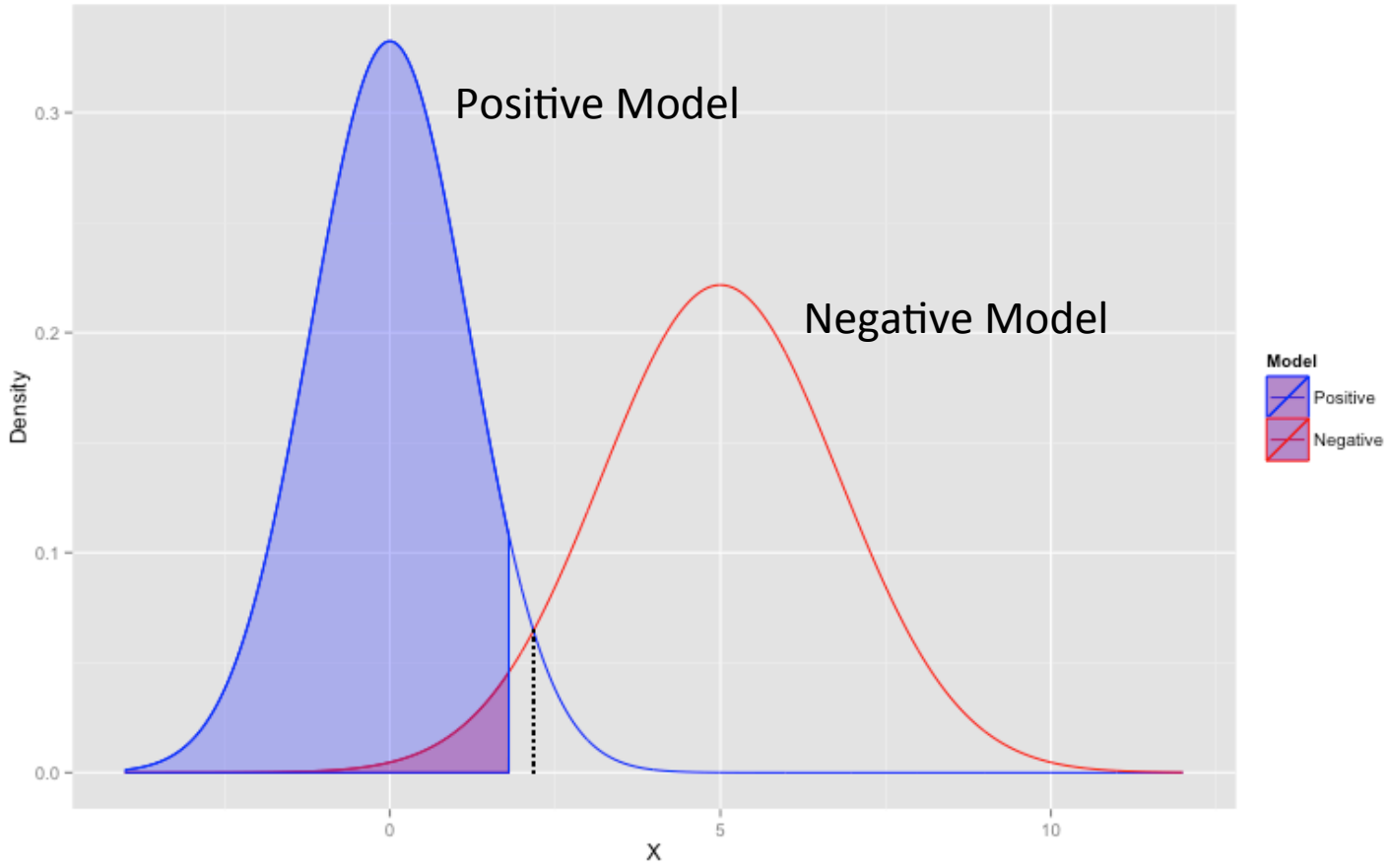
- Using the ranking produced by the model, filtering variants is as easy as setting a single threshold value
- Any variants whose score falls below the threshold is filtered out

(2) Apply model to callset



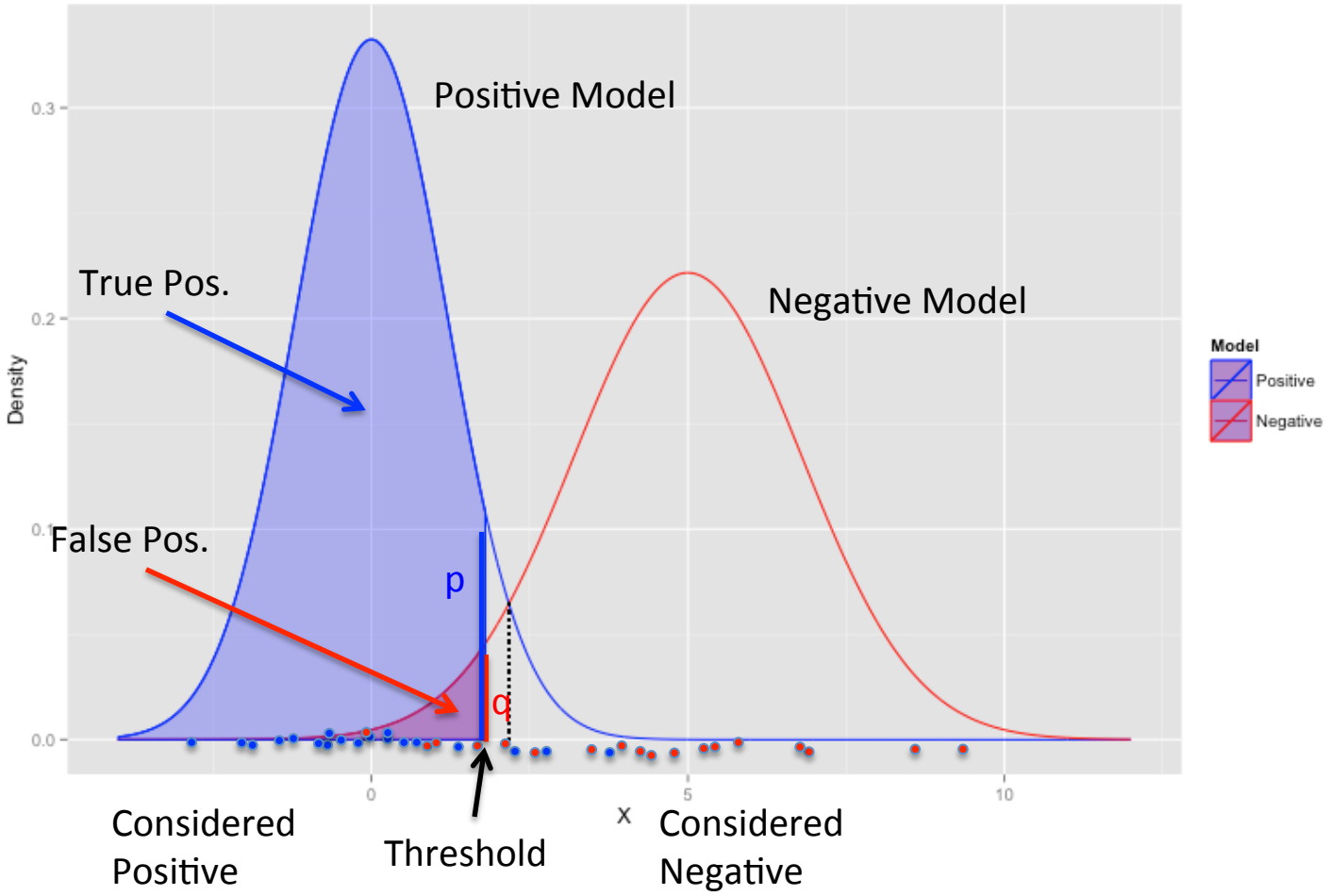
But how do we set that threshold?

There are in fact two components to the model



- A **negative model** is also built during training
- It represents the probability of variants to be **false positives**

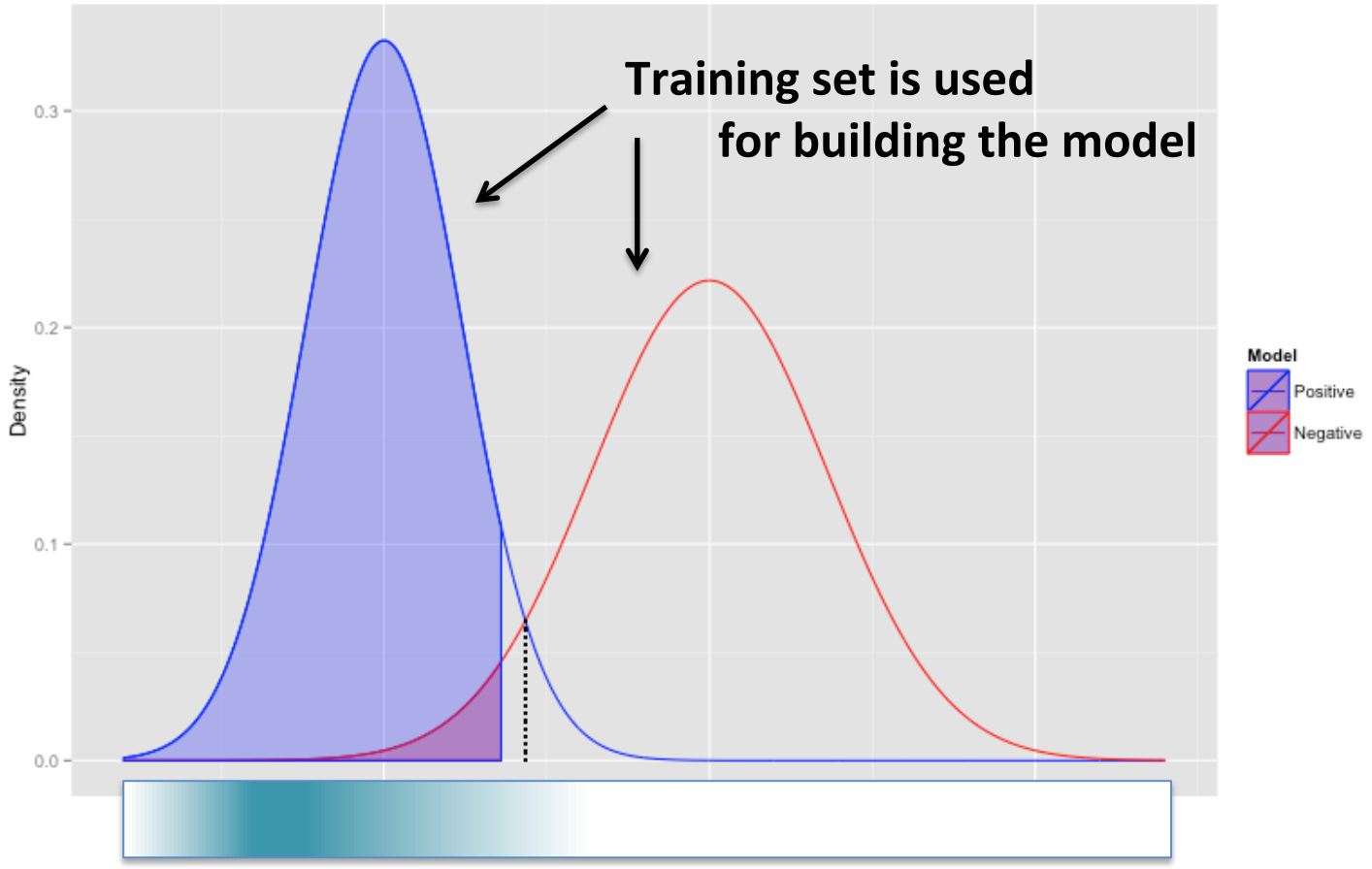
The VQSLOD threshold is a tradeoff between TP and FP



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

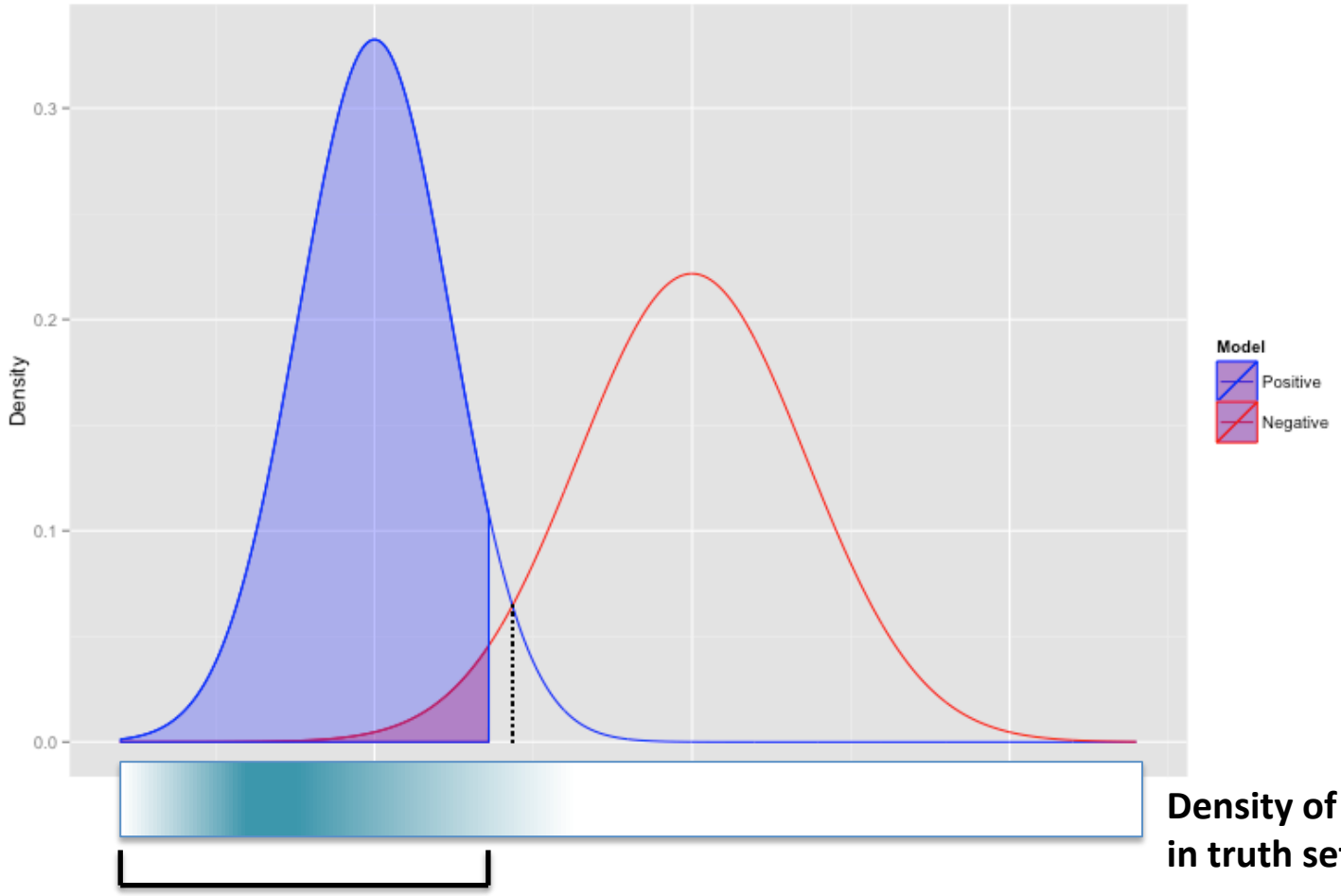
(VQSLOD is distinct from QUAL!)

Role of training and truth resources



Truth set is used for translating VQSLOD values into sensitivity "tranches"

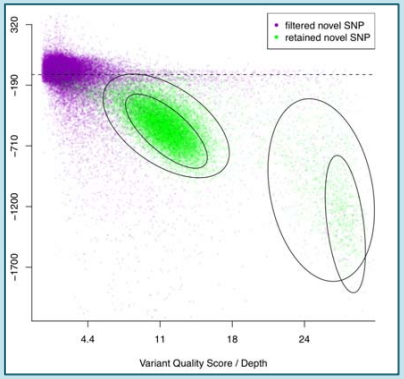
We set the threshold based on **sensitivity to truth data**



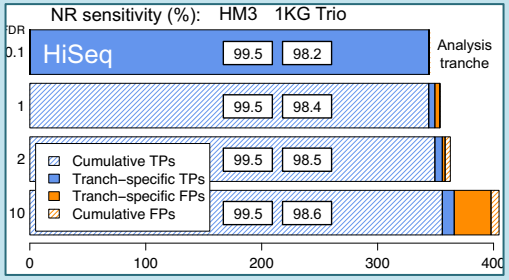
What threshold do we need to set to capture X % of the sites in the truth set?

Variant Recalibration steps & tools

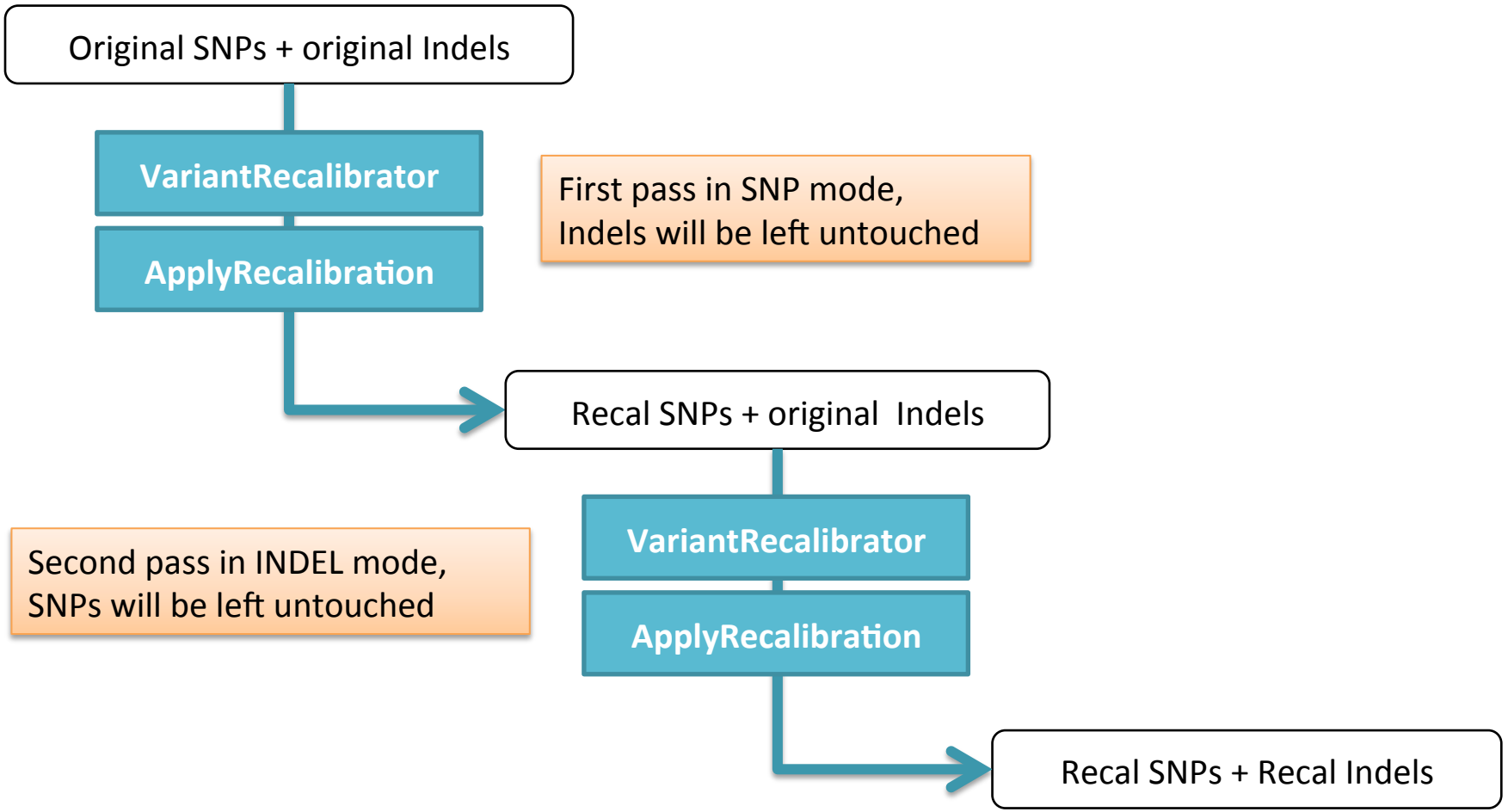
- Build and Apply the models (from resources and callset)
  - ➔ **VariantRecalibrator**



- Use VQSLOD to filter variants and write a new annotated VCF
  - ➔ **ApplyRecalibration**

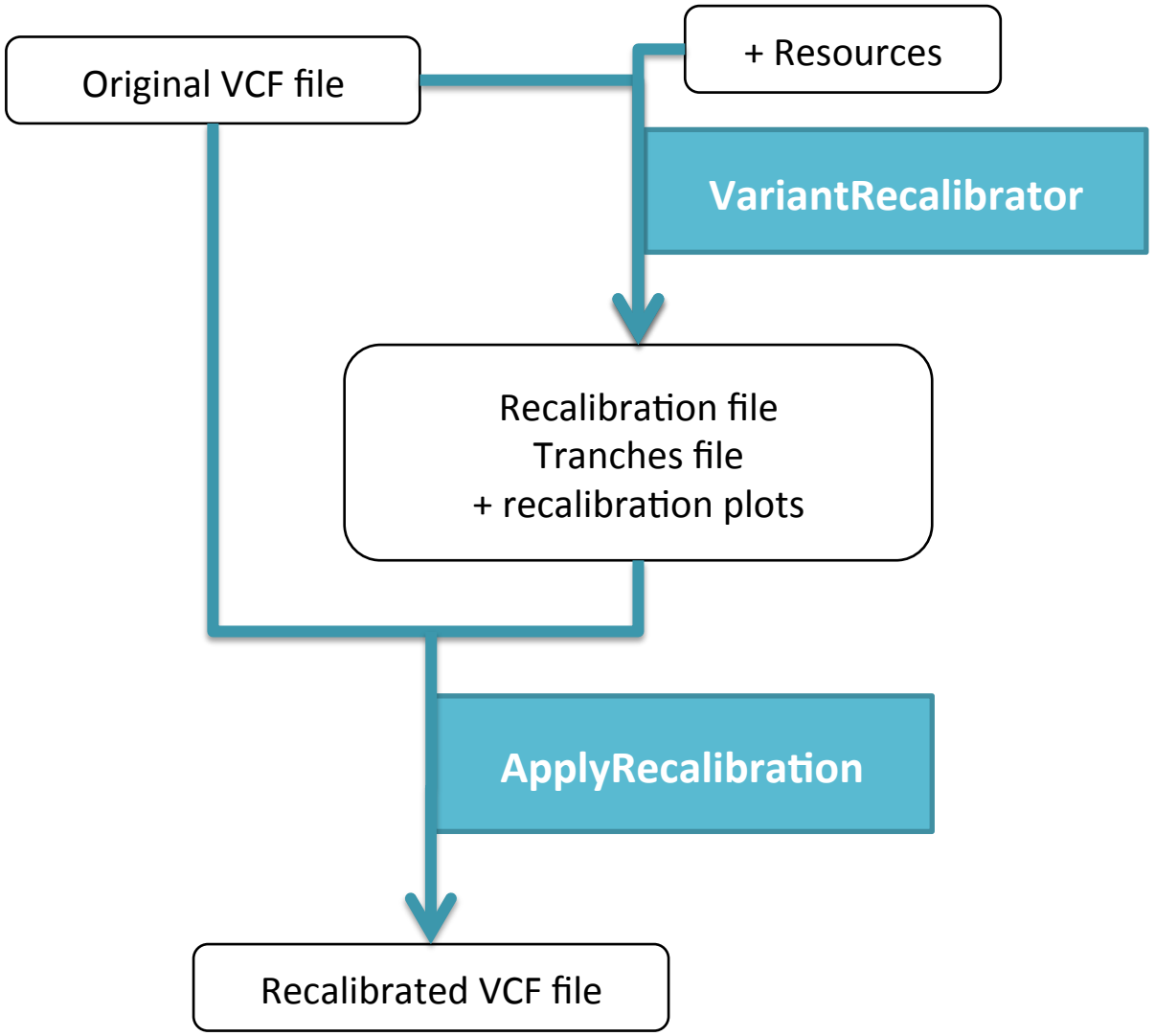


NOTE: SNPs and Indels must be recalibrated separately!

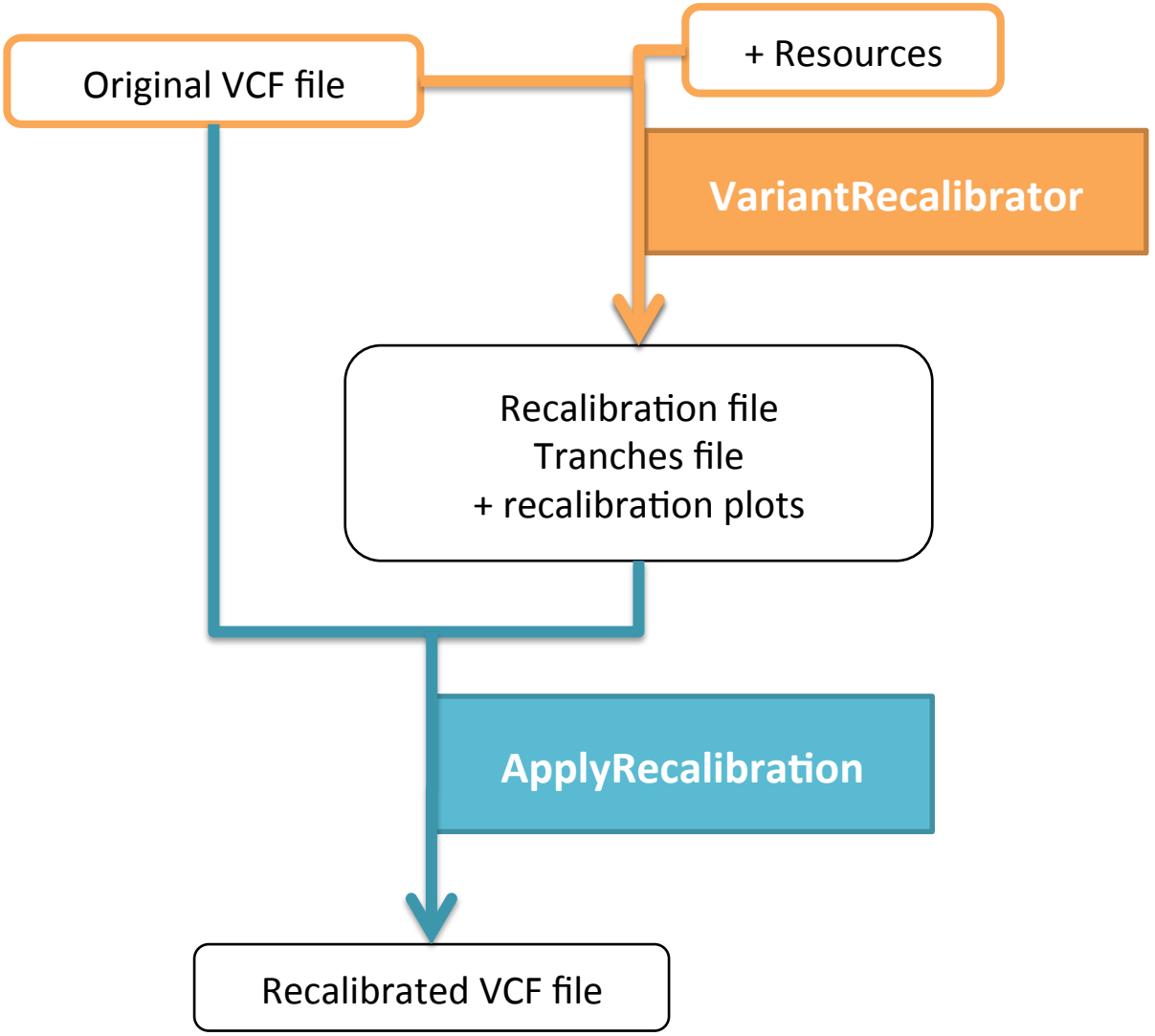


Pro-tip: Run VQSR twice in succession according to this workflow.
That way you avoid having to split them, recalibrate and combine them again.

Variant Recalibration workflow



Variant Recalibration workflow



TOOL TIPS

VariantRecalibrator

- Build the Gaussian mixture model using the variants in the input callset which overlap the training data

```
java –jar GenomeAnalysisTK.jar –T VariantRecalibrator \
        –R human.fasta \
        –input raw.SNPs.vcf \
        –resource: {see next slide} \
        –an DP –an QD –an FS –an MQRankSum {…} \
        –mode SNP \
        –recalFile raw.SNPs.recal \
        –tranchesFile raw.SNPs.tranches \
        –rscriptFile recal.plots.R
```

SNP example – see documentation for indel recommendations

TOOL TIPS

VQSR resources

```
-resource:hapmap,known=false,training=true,truth=true,prior=15.0
hapmap_3.3.b37.sites.vcf
-resource:omni,known=false,training=true,truth=false,prior=12.0
omni2.5.b37.sites.vcf
-resource:1000G,known=false,training=true,truth=false,prior=10.0
1000G.b37.sites.vcf
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
dbsnp_137.b37.vcf
```

SNP example – see documentation for indel recommendations

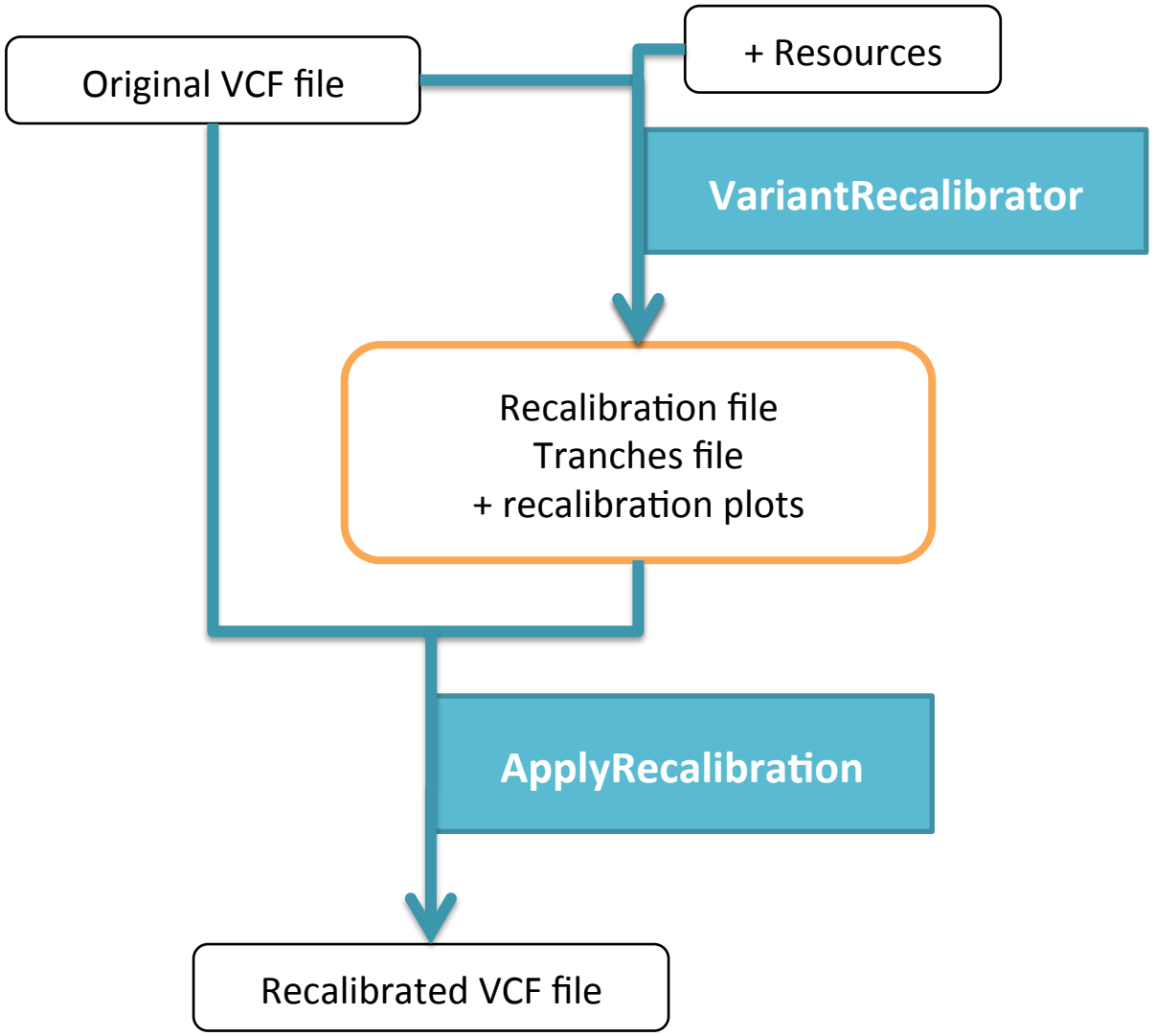
TOOL TIPS

Understanding “resources”

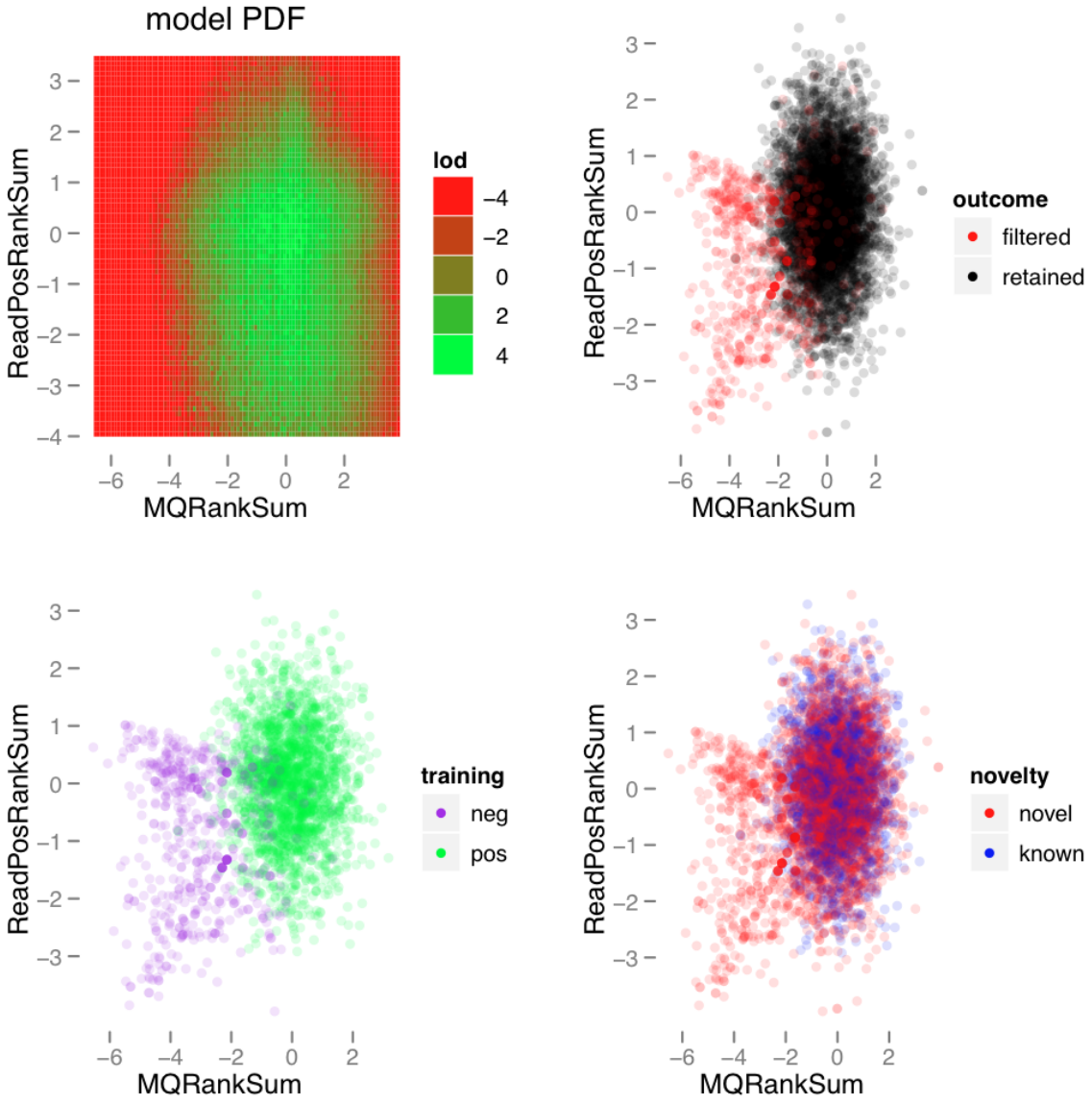
- **Prior –** Phred-scaled estimate of data accuracy

- **Training –** use input variants that overlap with these training sites to build the model
- **Truth -** use these truth sites to determine where to set the cutoff in VQSLOD sensitivity.
- **Known –** only for reporting purposes, not used in any calculations

Variant Recalibration workflow

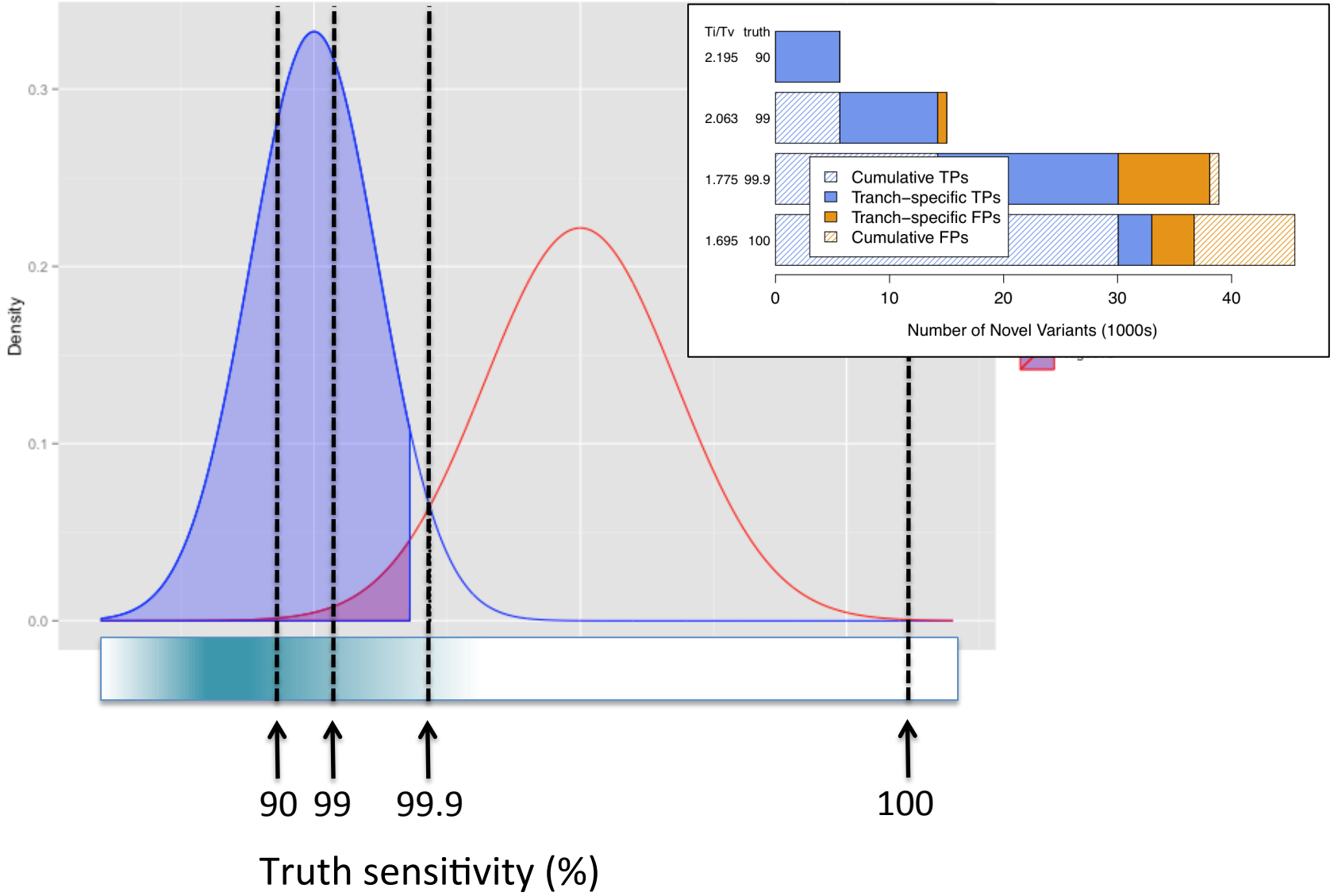


Recalibration plots show aspects of the model
(for each possible pair of annotations)

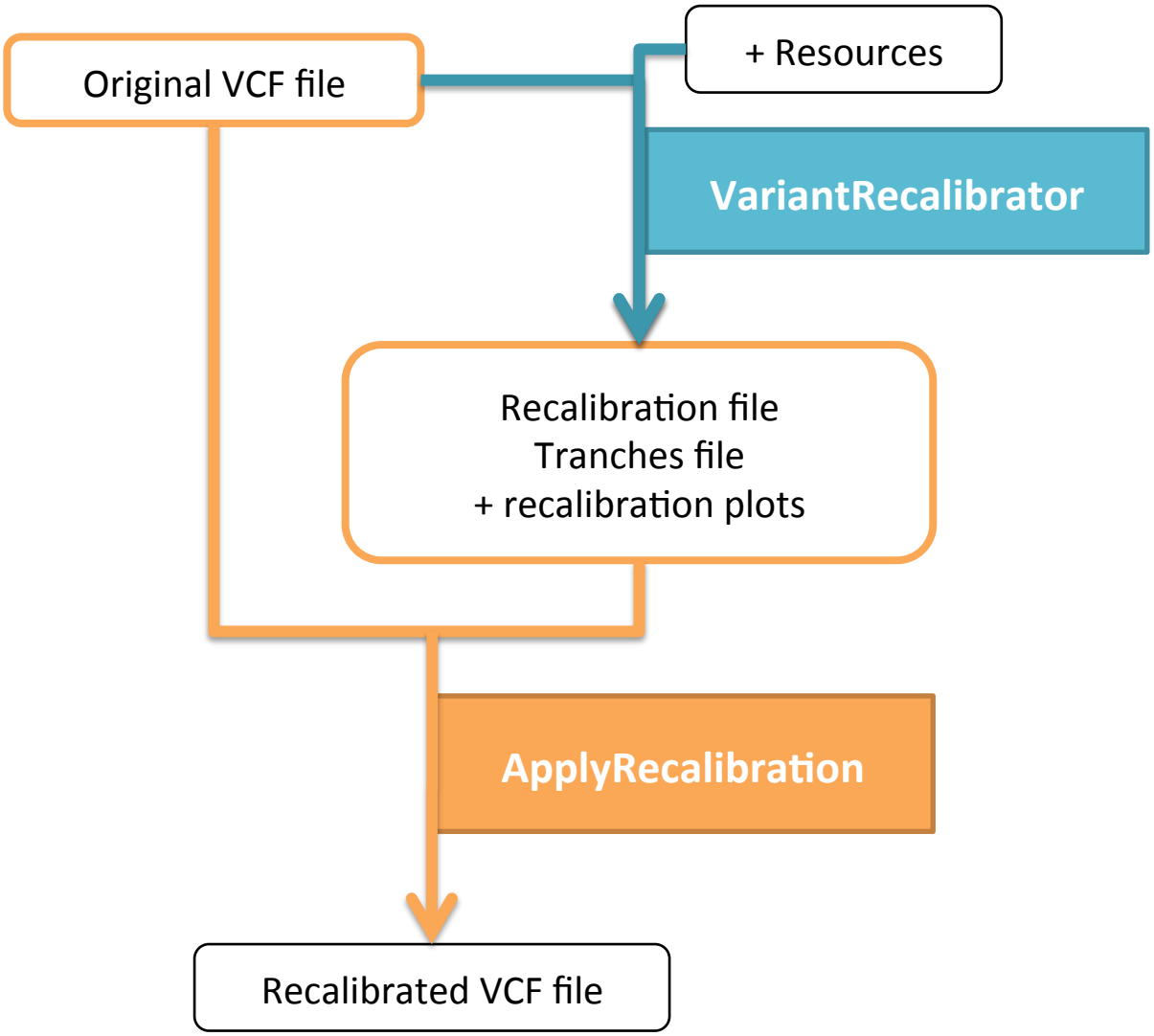


Tranches are slices of the data

corresponding to pre-set sensitivity threshold values



Variant Recalibration workflow



TOOL TIPS

ApplyRecalibration

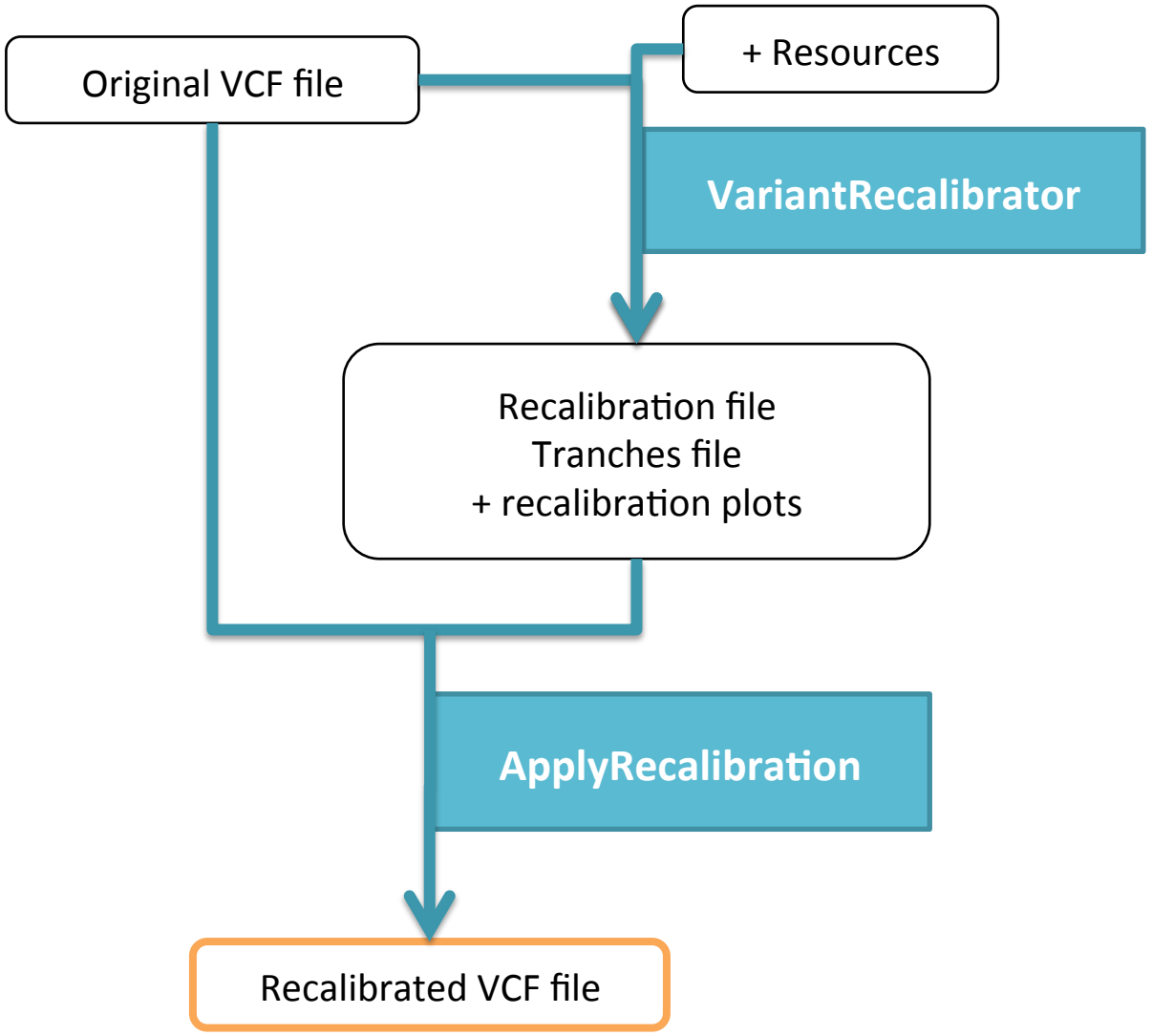
- Executes the desired sensitivity / specificity tradeoff by applying filters to the input callset
- Creates a new, filtered, analysis-worthy VCF file.

```
java –jar GenomeAnalysisTK.jar –T ApplyRecalibration \
        –R human.fasta \
        –input raw.vcf \
        –mode SNP \
        –recalFile raw.SNPs.recal \
        –tranchesFile raw.SNPs.tranches \
        –o recal.SNPs.vcf \
        –ts_filter_level 99.0
```

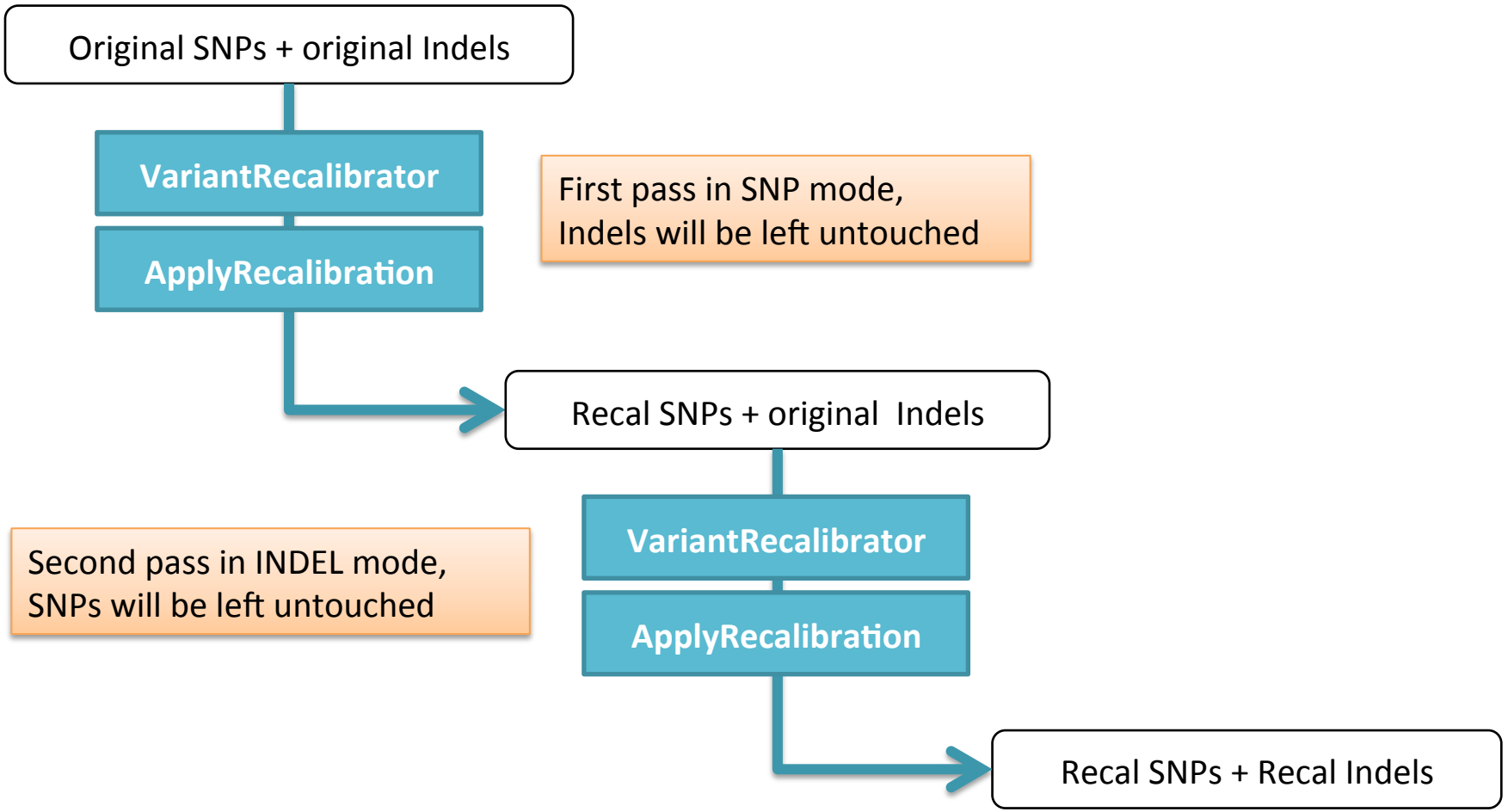
- Additionally every variant is now annotated with its VQSLOD score.

SNP example – see documentation for indel recommendations

Variant Recalibration workflow



NOTE: SNPs and Indels must be recalibrated separately!



Pro-tip: Run VQSR twice in succession according to this workflow.
That way you avoid having to split them, recalibrate and combine them again.

VQSR output VCF (vs. Hard Filter)

- Before VQSR (input vcf):

#CHROM	POS	FILTER	INFO
1	10146	.	AC=1;DP=32;FS=9.208; MQ=31.96;MQRankSum=0.085;...
1	10403	.	AC=1;DP=64;FS=1.645;MQ=41.86;MQRankSum=1.87;...
1	234313	.	AC=1;DP=239;FS=12.675;MQ=38.19;MQRankSum=-0.122;...

- After VQSR (output vcf):

#CHROM	POS	FILTER	INFO
1	10146	VQSRTrancheINDEL99.30to99.50	AC=1...;NEGATIVE_TRAIN_SITE;VQSLOD=-1.328;culprit=SOR
1	10403	PASS	AC=1;...;QD=0.60; VQSLOD=0.794;culprit=QD
1	234313	VQSRTrancheSNP99.90to100.00	AC=1;...;POSITIVE_TRAIN_SITE;VQSLOD=-5.356;culprit=MQ

- Hard filtering vcf:

#CHROM	POS	FILTER	INFO
1	10146	PASS	AC=1;DP=32;FS=9.208; MQ=31.96;MQRankSum=0.085;...
1	10403	INDEL_Filter	AC=1;DP=64;FS=1.645;MQ=41.86;MQRankSum=1.87;...
1	234313	SNP_Filter	AC=1;DP=239;FS=12.675;MQ=38.19;MQRankSum=-0.122;...

Did the recalibration work properly?

- Common error modes:
  - "No data found"
    -> too few variants in callset, see docs for workarounds
  - "Annotation X not found for any variant"
    -> annotation not present in file, use VariantAnnotator
    -> if related to InbreedingCoefficient probably less than 10 samples, cannot use this annotion
  - Screwed-up tranche plots, low novel TiTv
    -> wrong dbSNP version, use the one in the bundle

NOTES

Variant Recalibration (VQSR) on WEx data

- Smaller number of variants per sample compared to WGS

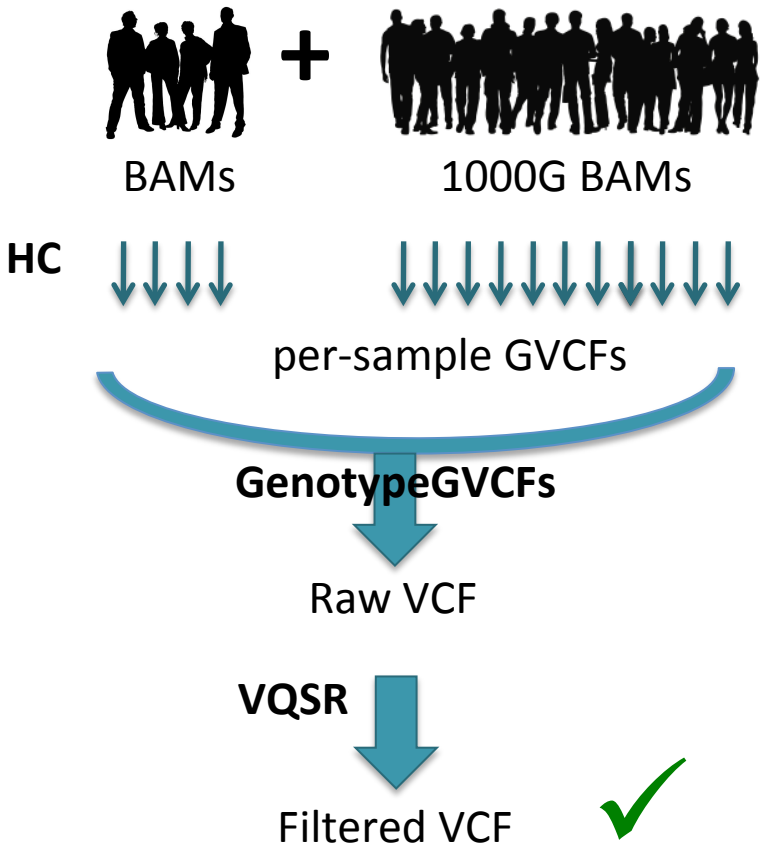
  -> typically insufficient to build a robust recalibration model if running on only a few samples

- ➢ Analyze samples jointly in cohorts of at least 30
- ➢ If necessary, add exomes from 1000G Project

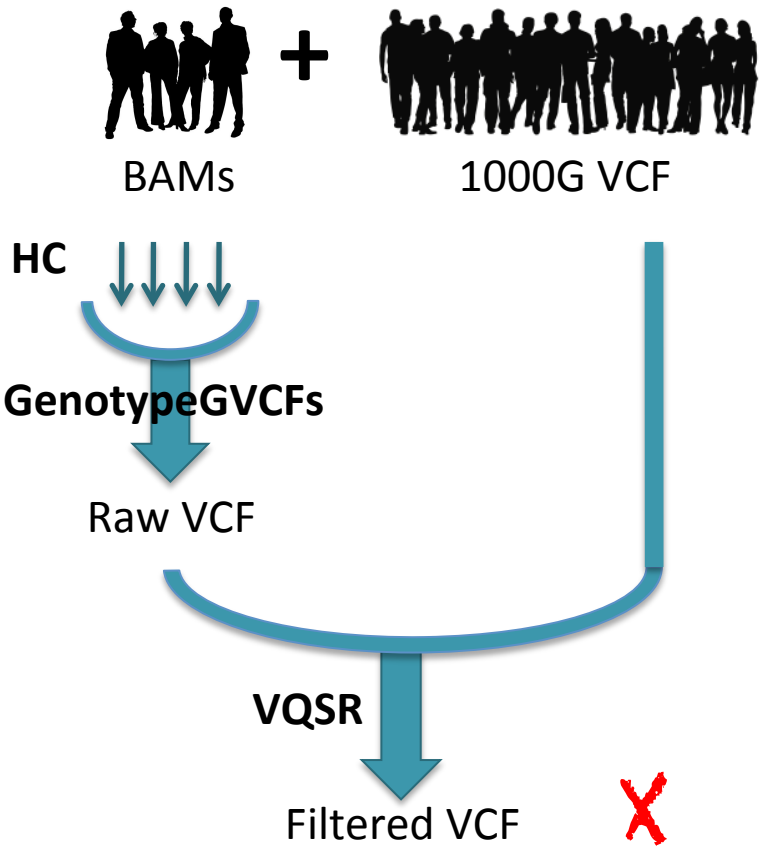
- What to look for in samples for padding a cohort:
  - Similar technical generation (technology, capture, read length, depth)
  - Similar ethnic background

How to add exomes from 1000G to your analysis

- **ALWAYS** do this:



- **NEVER** do this :

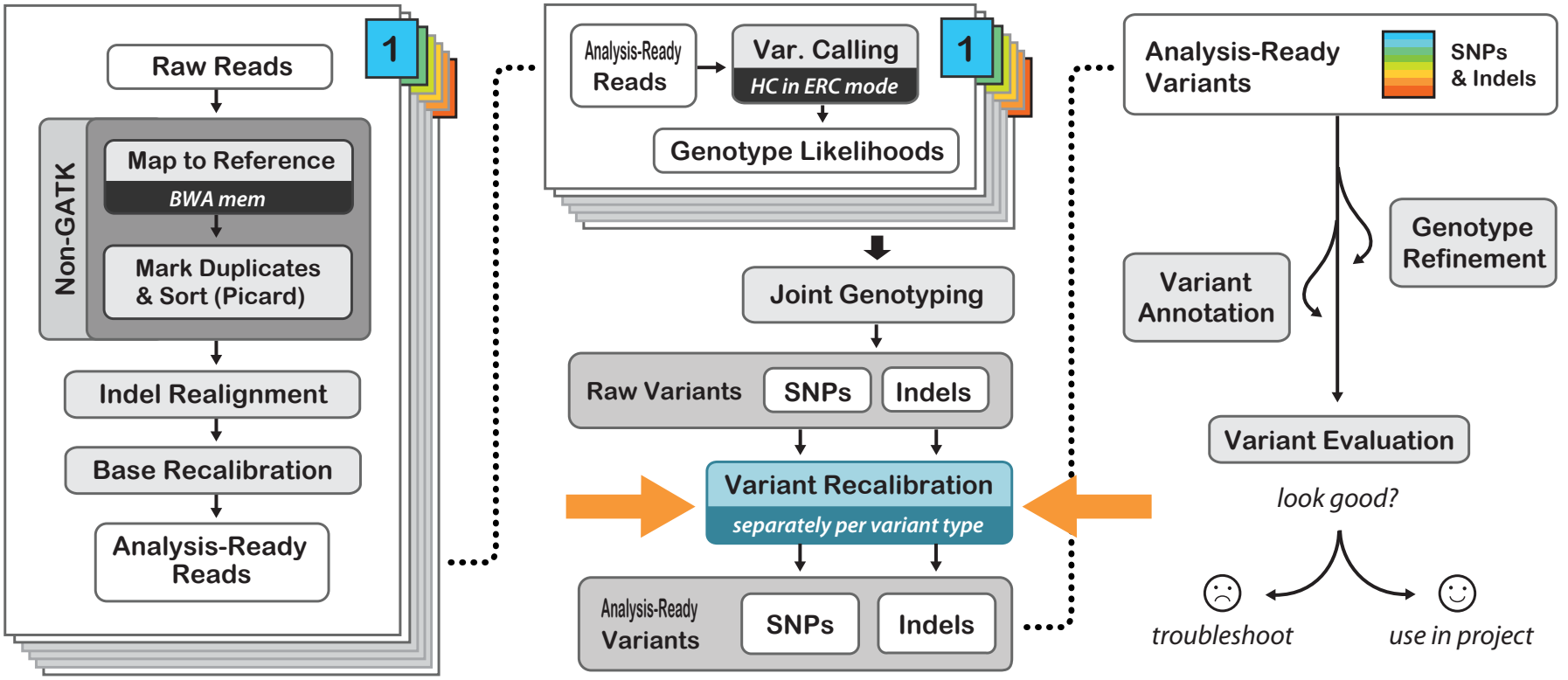


When should you NOT run VQSR?

- Non-human organisms where known resources are unavailable or insufficiently curated
- RNAseq data → see RNAseq-specific filtering
- Cohort is too small and no other samples are available for "padding" the cohort

→ Use manual filtering recommendations instead

You are here in the GATK Best Practices workflow for **germline variant discovery**



Further reading

http://www.broadinstitute.org/gatk/guide/best-practices

http://www.broadinstitute.org/gatk/guide/article?id=39

http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_variantrecalibration_VariantRecalibrator.html

http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_variantrecalibration_ApplyRecalibration.html