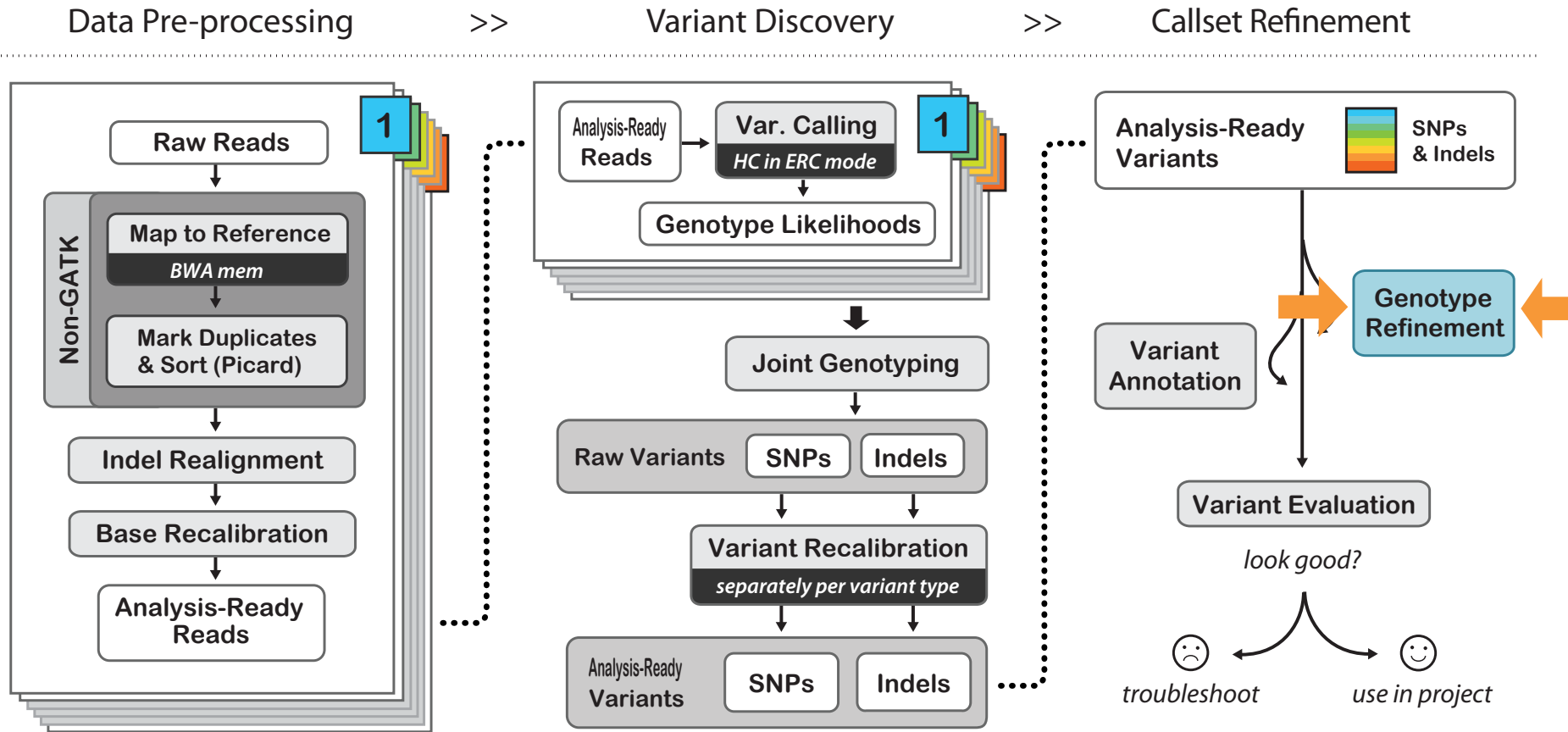


# Genotype Refinement Workflow

Using additional data to improve  
genotype calls and likelihoods

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



# Why care about genotypes?

- Medical geneticists need genotypes for patients
  - Do any patients have two copies of a LOF mutation?
  - Are the parents of a diseased child likely to have more afflicted children?
- Population geneticists need genotypes for association studies
  - How does the number of copies of an allele affect the phenotype?

# Variant call vs. Genotype call

- **Variant call**
  - = there is variation at this site in one or more samples
- **Genotype call**
  - = these are the haplotypes present in this sample

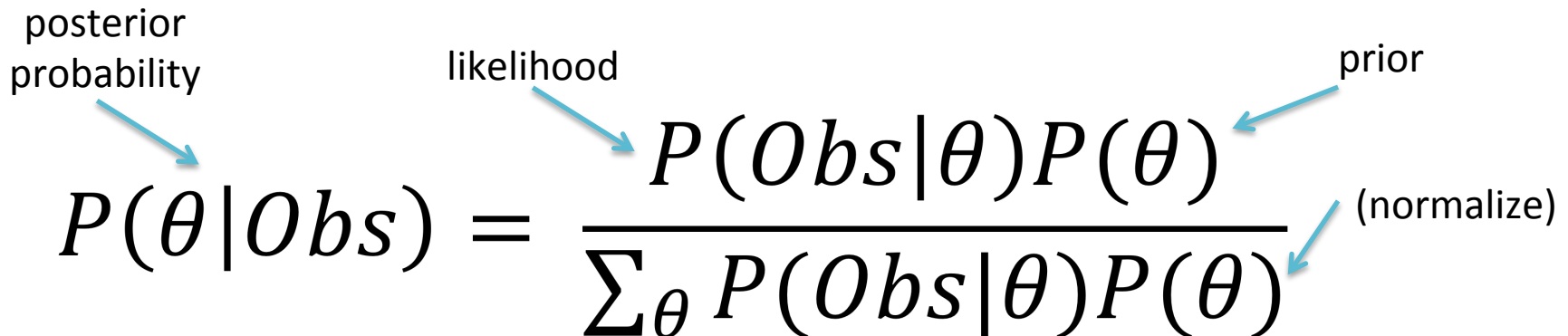
# Genotype call quality is important!

- Some sites/samples have poor genotype calls
  - Can be ambiguous due to low confidence
  - Might be entirely wrong!
- Can additional (independent) data improve genotype calls?
  - Use high quality data (like 1000G) as priors
  - Use pedigree (if available)
  - Calculate posterior genotype probabilities

# Review of Bayes's Rule

Given that your coworker just walked in with an umbrella, what is the probability that it is raining?

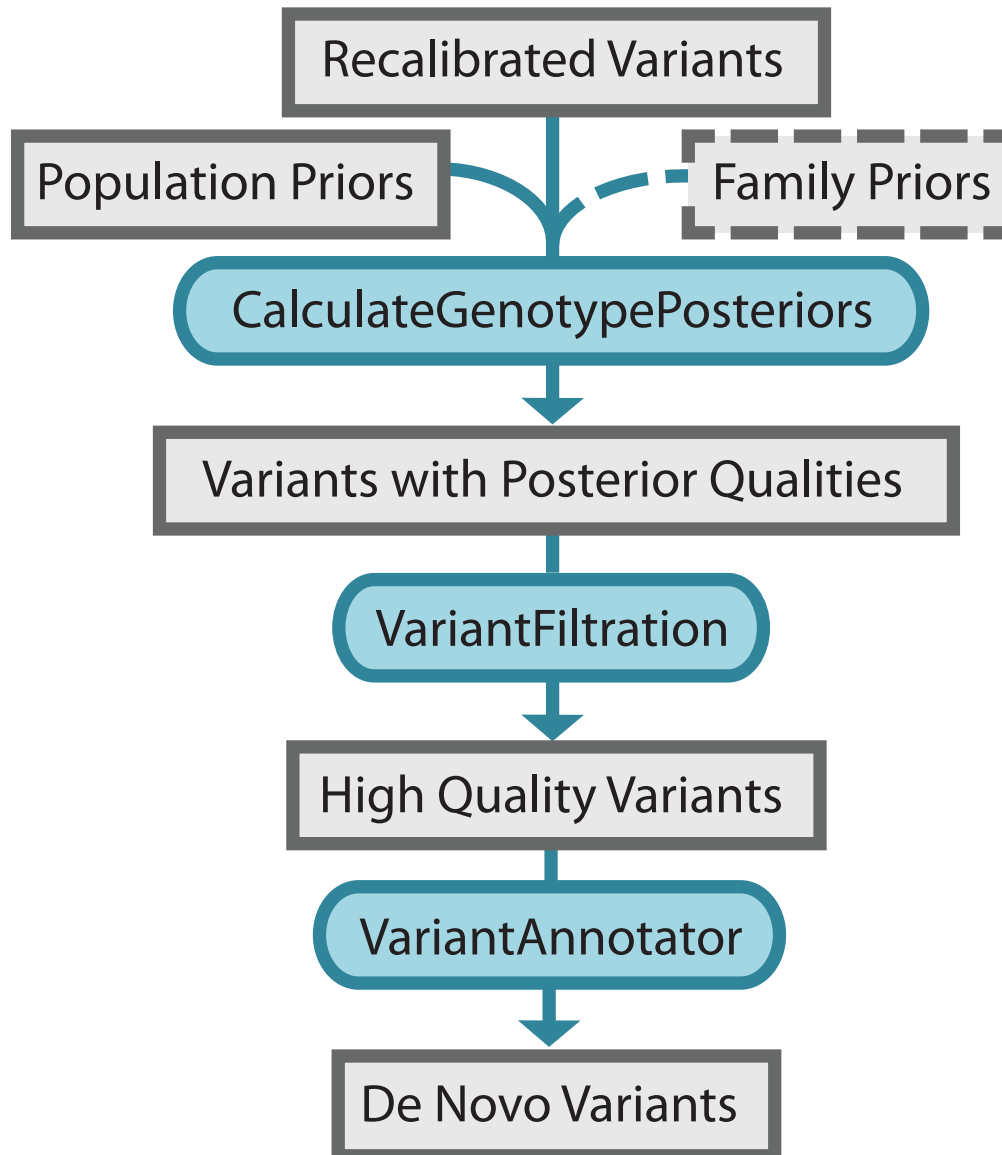
- Observation = umbrella
- $\Theta$  = probability of rain



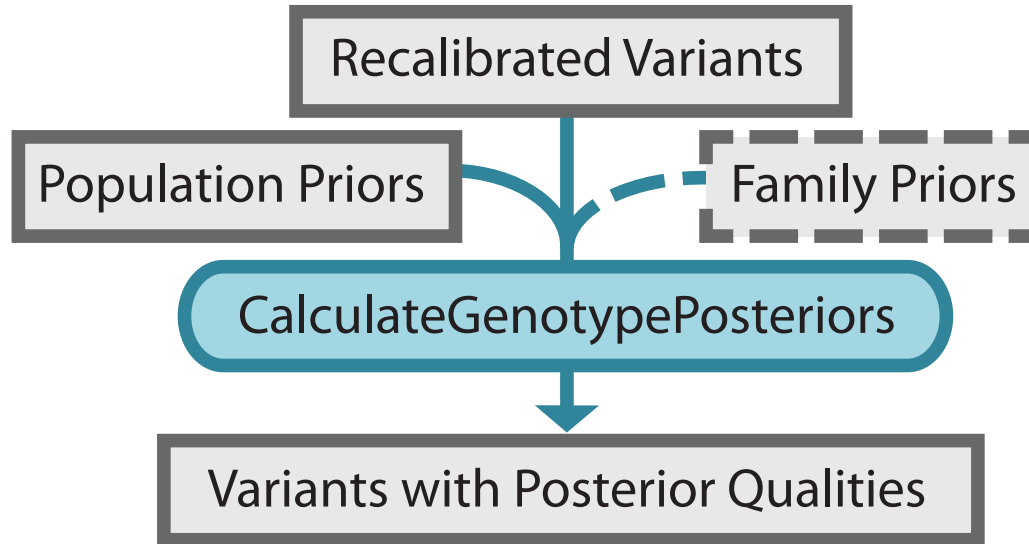
The diagram shows the Bayes's Rule equation with four blue arrows pointing to its components: 'posterior probability' points to  $P(\theta | Obs)$ , 'likelihood' points to  $P(Obs|\theta)$ , 'prior' points to  $P(\theta)$ , and '(normalize)' points to the denominator  $\sum_{\theta} P(Obs|\theta)P(\theta)$ .

$$P(\theta | Obs) = \frac{P(Obs|\theta)P(\theta)}{\sum_{\theta} P(Obs|\theta)P(\theta)}$$

# Genotype Refinement Workflow



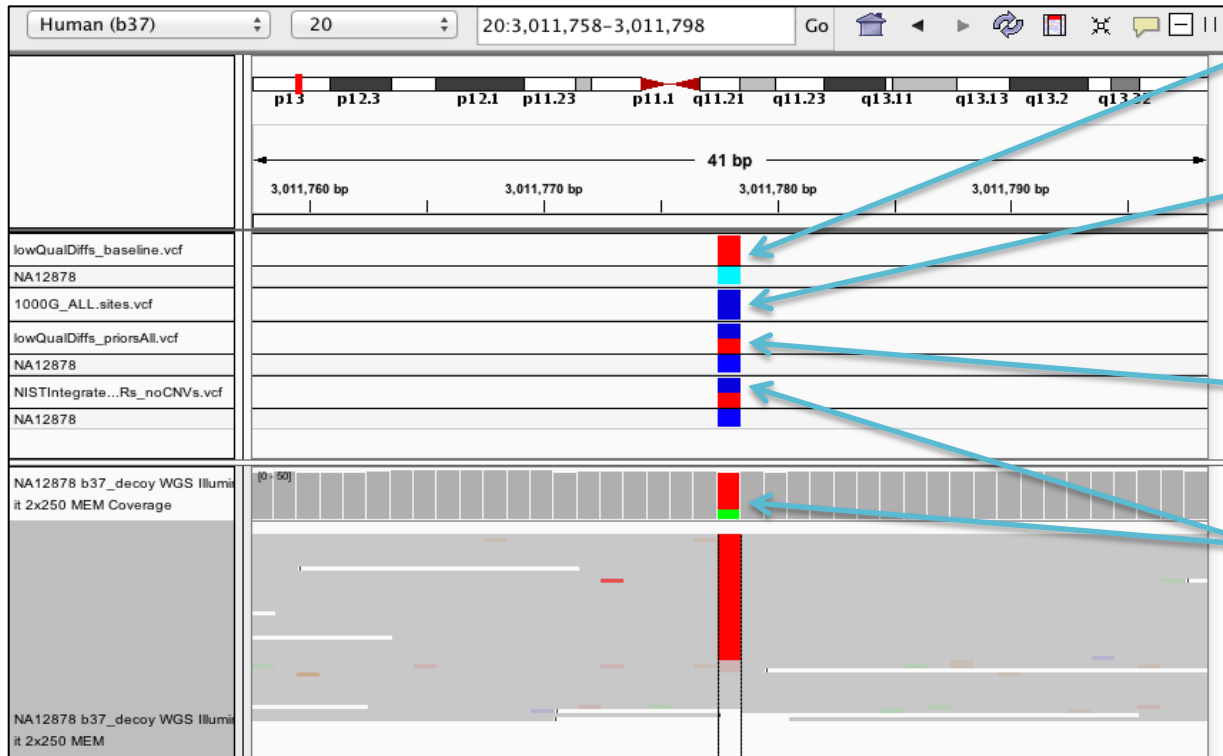
# CalculateGenotypePosteriors



```
java -jar GenomeAnalysisTK.jar \  
-T CalculateGenotypePosteriors \  
-R reference.fasta \  
-V input.vcf \  
-ped family.ped \  
-supporting population.vcf \  
-o output.vcf
```



# Case 1: HOM\_VAR Call w/ Low Frequency Priors



1) Baseline HOM\_VAR call

2) Priors w/low allele frequency applied

3) Posterior genotype called HET

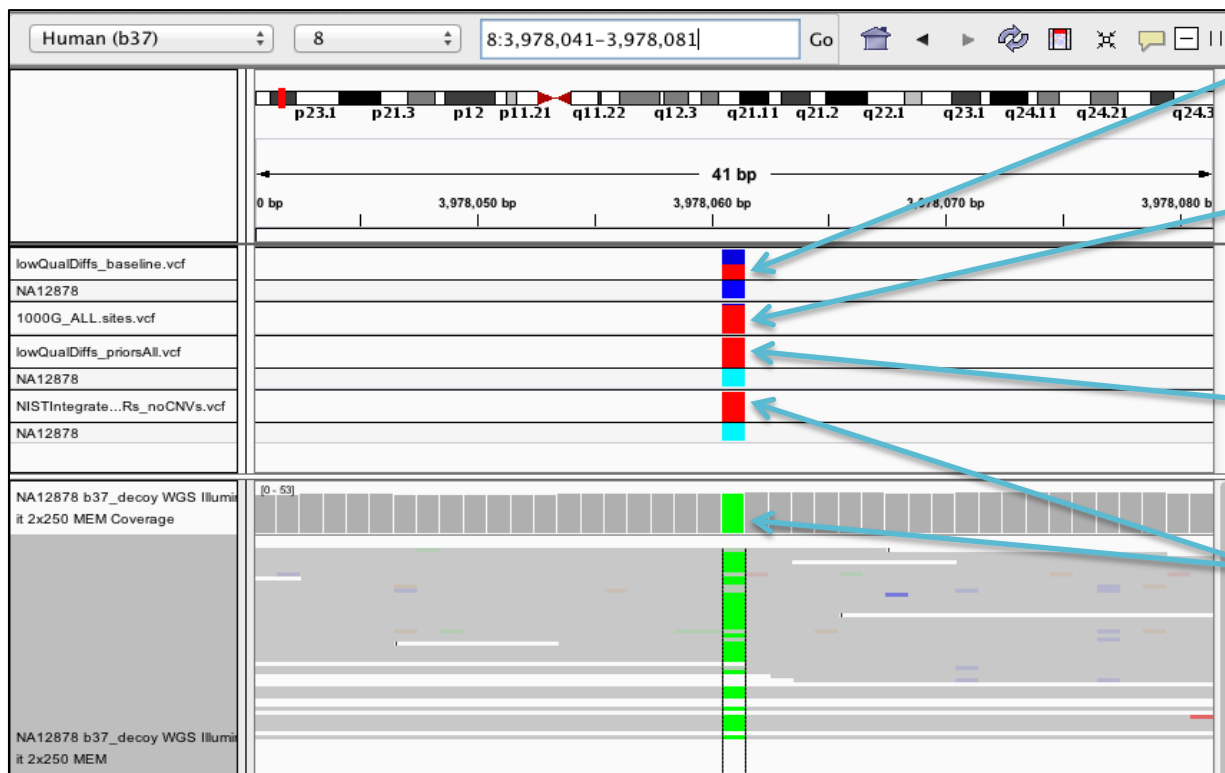
4) In agreement w/NIST and BAMs

Likelihoods x Priors = Posterior Probabilities

$[895, 3, 0]$      $AF=0.002$      $[868, 0, 27]$   
 [HOM\_REF, HET, HOM\_VAR]                      [HOM\_REF, HET, HOM\_VAR]

Genotype corrected  
 Confidence improved  
 from Q3 to Q27

# Case 2: HET Call with High Frequency Priors



1) Baseline HET call

2) Priors w/high allele frequency applied

3) Posterior genotype called HOM\_VAR

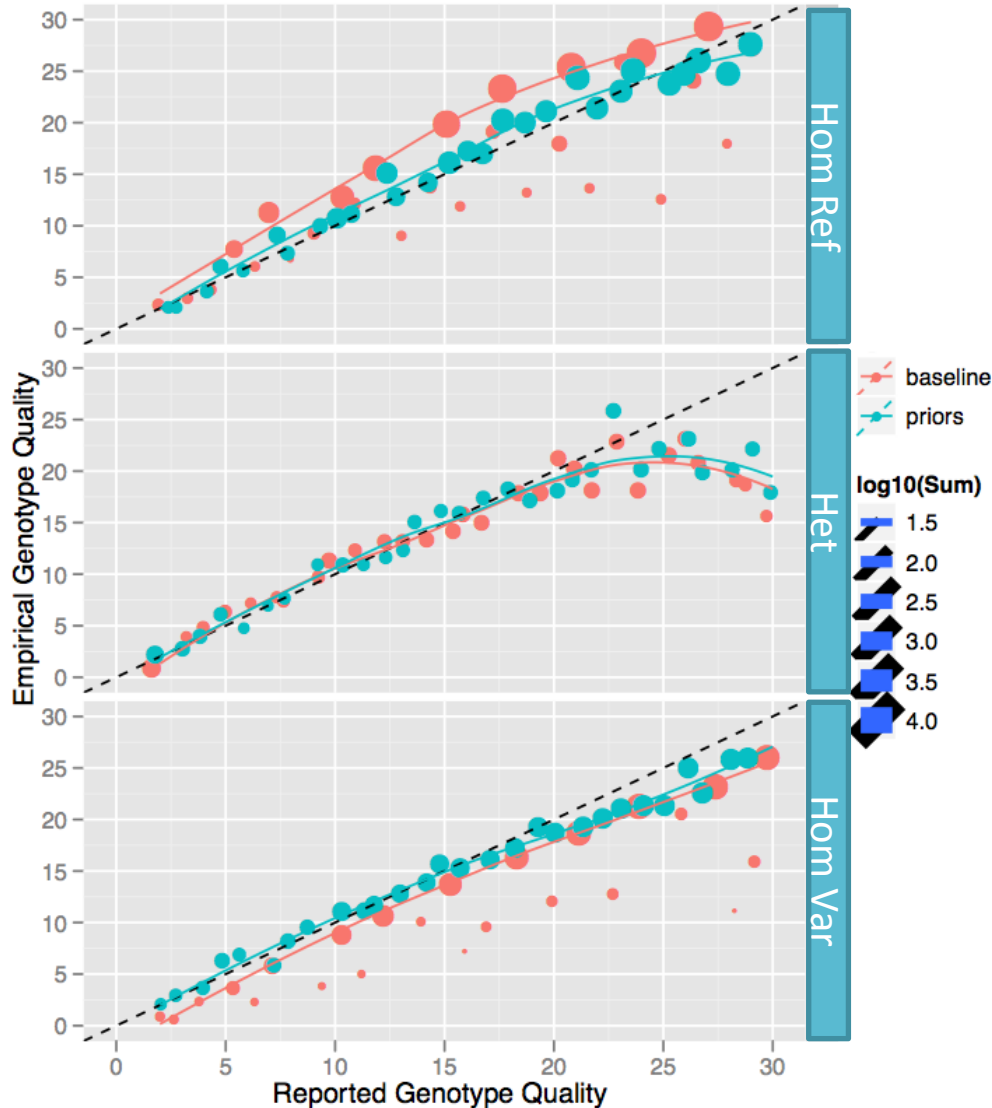
4) In agreement w/NIST and BAMs

Likelihoods x Priors = Posterior Probabilities

$[894, 0, 0]$      $AF=0.987$      $[932, 16, 0]$   
 [HOM\_REF, HET, HOM\_VAR]                      [HOM\_REF, HET, HOM\_VAR]

Genotype corrected  
Confidence improved  
from Q0 to Q16

# Population priors improve genotype confidence

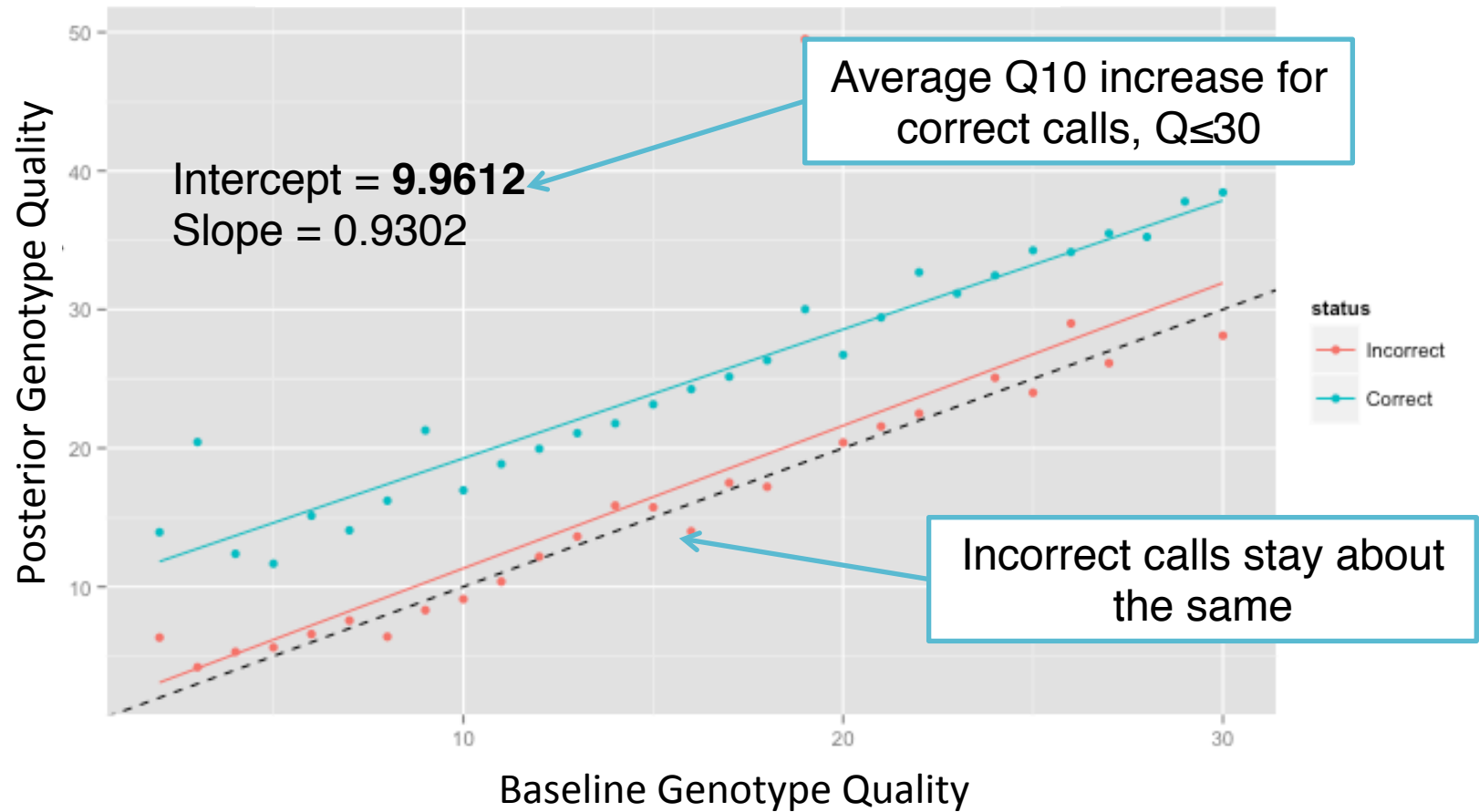


**Baseline** HomRef calls are under confident, but **posterior** calls are more accurate

**Baseline** HomVar calls are over confident, but **posterior** calls are improved

# Assessing confidence and correctness

## Homozygous Reference Calls



# Parental genotypes inform child genotypes

- Child can only inherit alleles present in parents
- Parent genotypes determine possible child genotypes (assuming no mutations)

Child	Mother	Father
HR	HR	HR
HR	HR	HET
HR	HET	HR
HR	HET	HET

Child	Mother	Father
HET	HET	HET
HET	HR	HET
HET	HET	HR
HET	HV	HET
HET	HET	HV
HET	HR	HV
HET	HV	HR

Child	Mother	Father
HV	HV	HV
HV	HV	HET
HV	HET	HV
HV	HET	HET

- HaplotypeCaller gives  $P(G_C|D_C)$
- Given trio data we can derive  $P(G_C|D_C, D_M, D_F)$

# Bayesian priors applied to trios

- Recall Bayes's Rule: 
$$P(\theta|Obs) = \frac{P(Obs|\theta)P(\theta)}{\sum_{\theta} P(Obs|\theta)P(\theta)}$$
- Establish genotype configuration probabilities

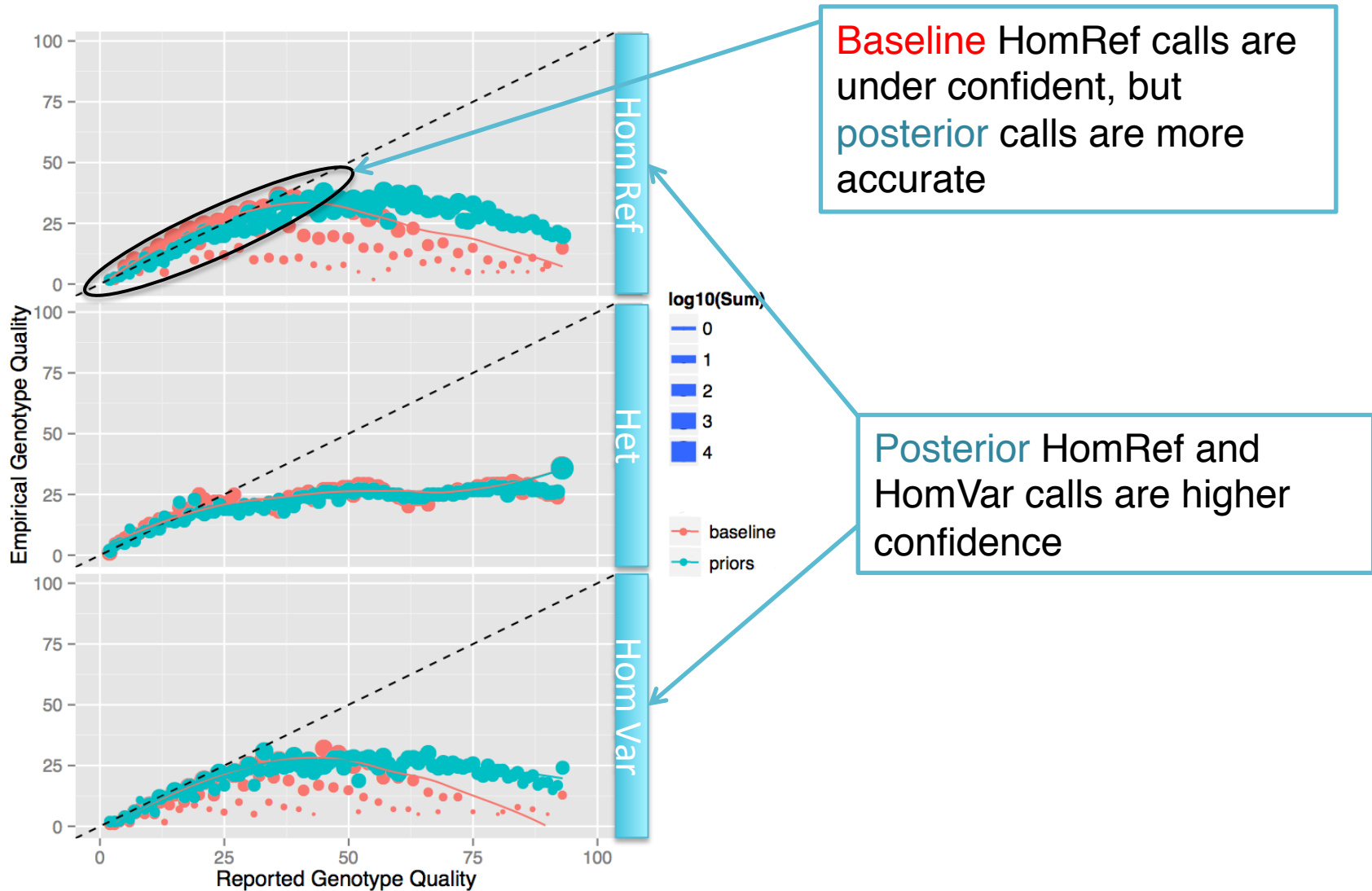
$$P(G_M, G_F, G_C) = P(\vec{G}) \begin{cases} \mu, 1MV \\ \mu^2, 2MVs \\ 1 - 10\mu - 2\mu^2, non - MV \end{cases}$$

- Apply family priors

$$P(G_C = HR | \vec{D}) = \frac{L_C(G_C = HR) \sum_{G_F, G_M} L_F(G_F) L_M(G_M) P(\vec{G})}{\sum_{\vec{H}} P(\vec{D} | \vec{H}) P(\vec{H})}$$

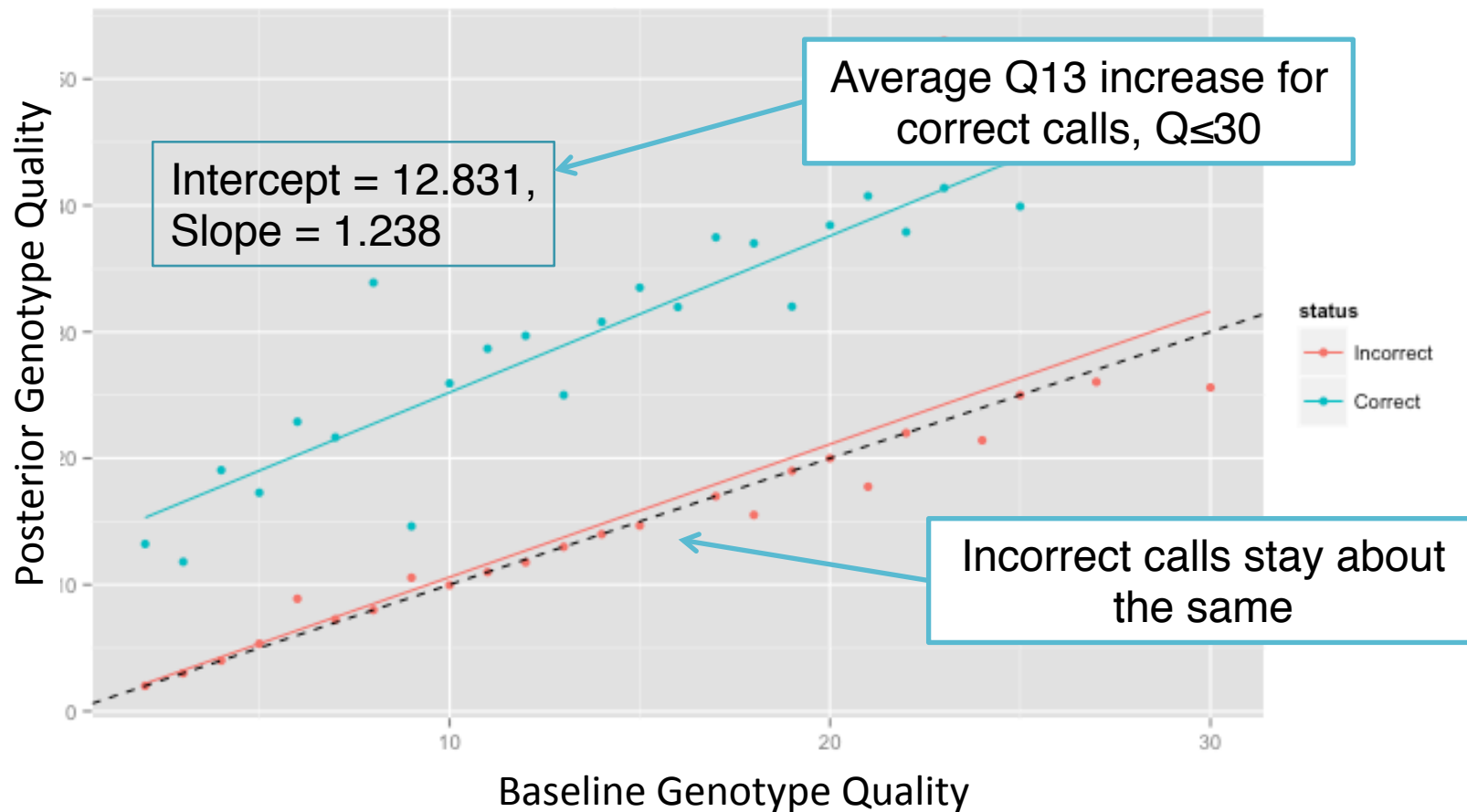
posterior  $\rightarrow$   $P(G_C = HR | \vec{D})$   
 likelihood  $\rightarrow$   $L_C(G_C = HR)$   
 apply prior  $\rightarrow$   $\sum_{G_F, G_M} L_F(G_F) L_M(G_M) P(\vec{G})$   
 normalize  $\rightarrow$   $\sum_{\vec{H}} P(\vec{D} | \vec{H}) P(\vec{H})$

# Family priors improve genotype confidence



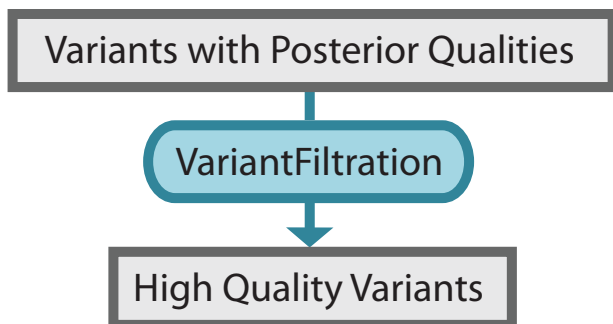
# Assessing confidence and correctness

## Homozygous Reference Calls





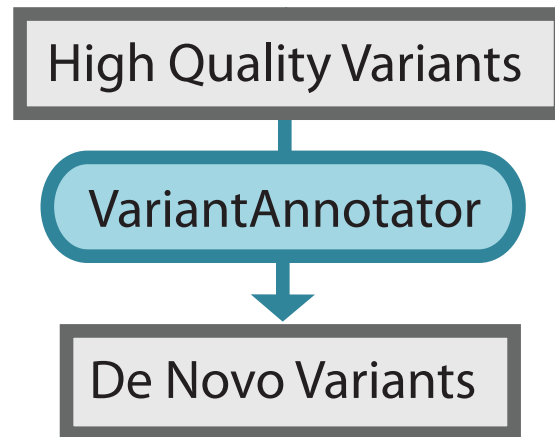
# Filter low confidence GQs



```
java -jar GenomeAnalysisTK.jar \  
-T VariantFiltration \  
-R reference.fasta \  
-V input.vcf \  
--filterExpression "GQ<20" \  
--filterName "lowGQ" \  
-o output.vcf
```

- Use VariantFiltration to filter ambiguous, low-confidence calls
- Recommended threshold is GQ = 20
  - GQ 20 is Phred-scaled 99% confidence
- Restrict further analysis to high-quality data

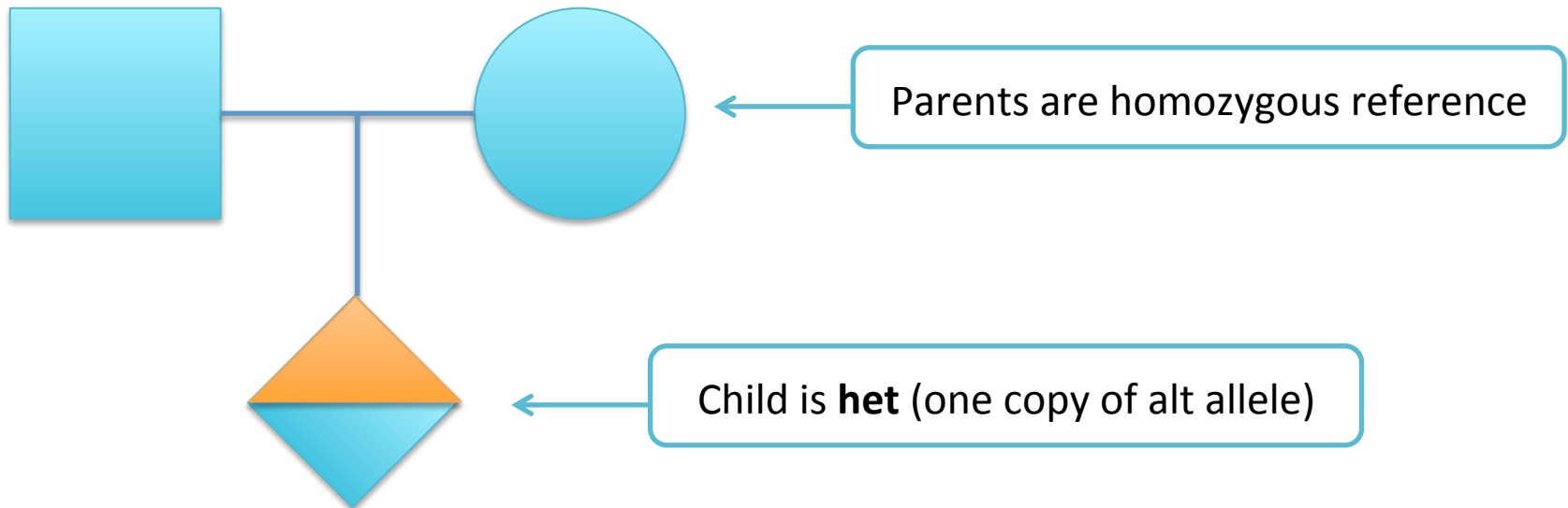
# VariantAnnotator



```
java -jar GenomeAnalysisTK.jar \  
-T VariantAnnotator \  
-R reference.fasta \  
-V input.vcf \  
-A PossibleDeNovo \  
-o output.vcf
```

# What are *De Novo* mutations?

- Culprits in many rare Mendelian disorders
- ~30 *de novo* mutations occur per human genome

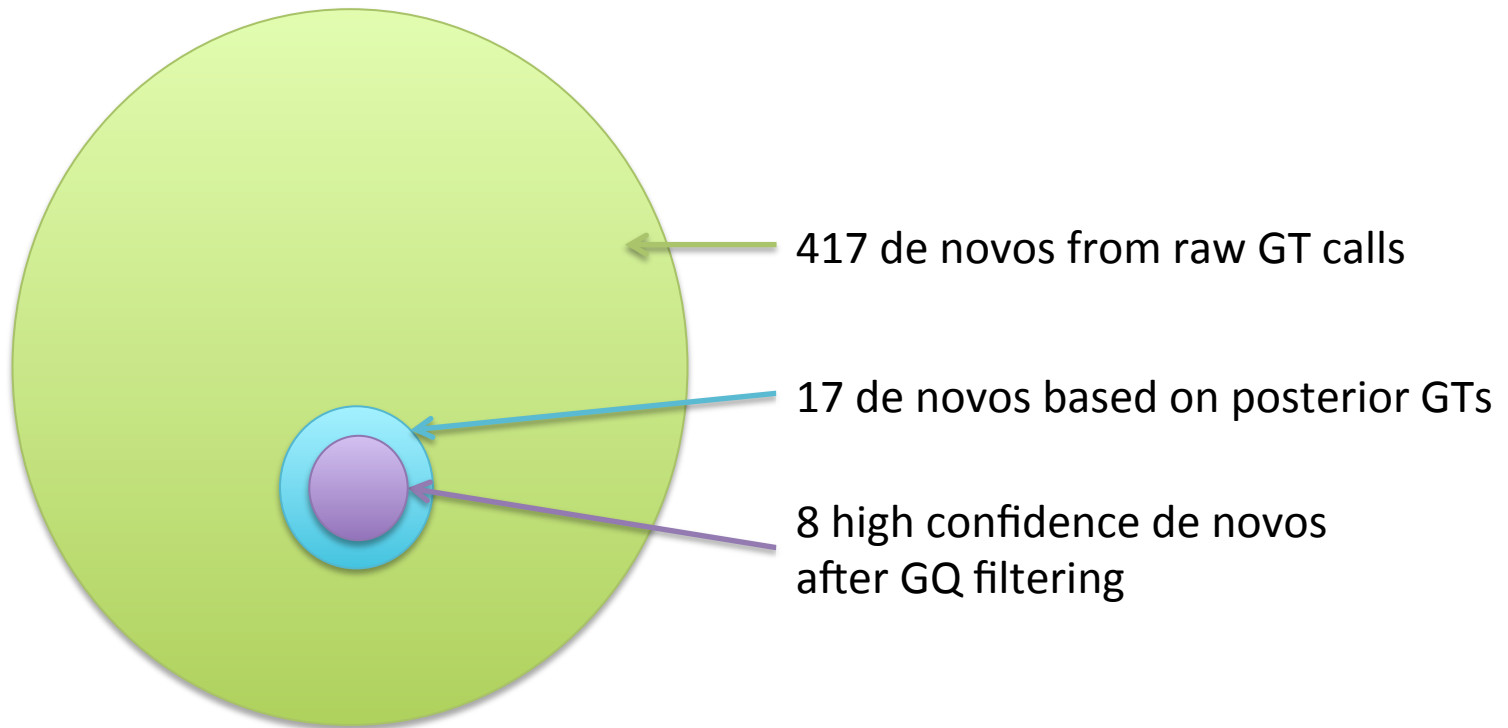


# Properties of sequenced De Novos

- Novelty
  - Child has only alt allele in trio, not inherited
- Rarity
  - Allele frequency across all samples sequenced is low
- Confidence
  - Set GQ threshold for parents and child
  - (GQ improvement tools help A LOT here!)

# Example of a clinical case

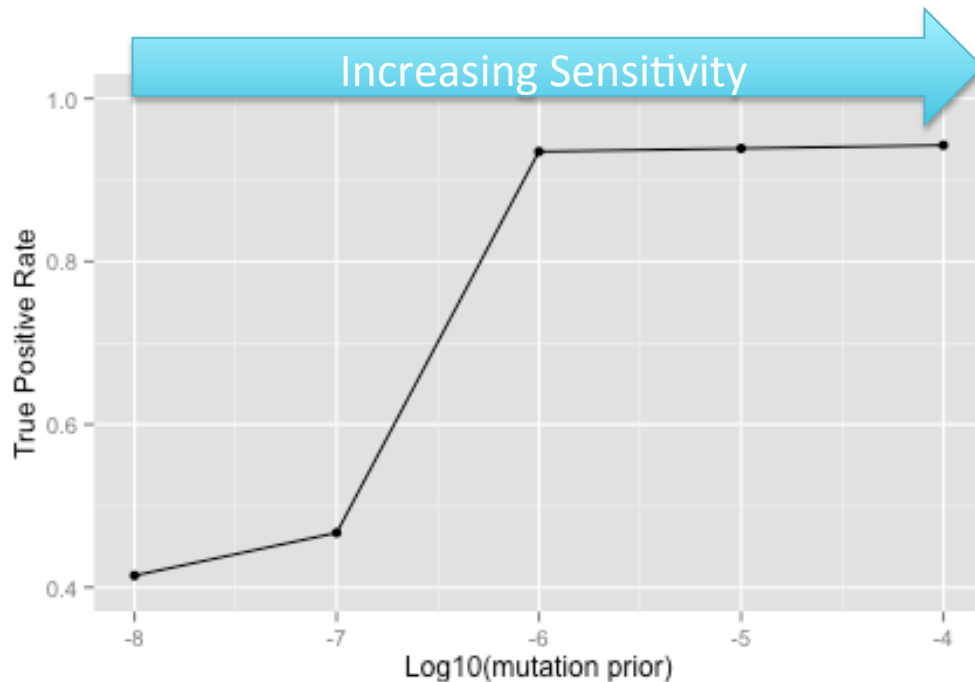
- Real clinical data
- Suspected de novo mutation in offspring



# Priors can be tuned for sensitivity

Mutation prior is a parameter in genotype configuration probability:

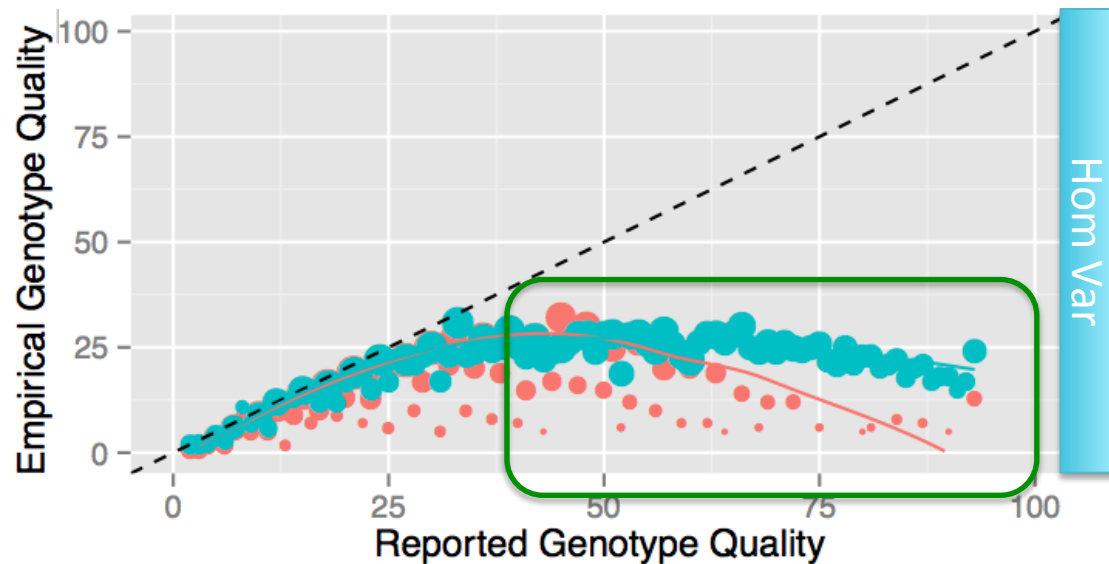
$$P(G_M, G_F, G_C) = P(\vec{G}) \begin{cases} \mu, 1MV \\ \mu^2, 2MV_s \\ 1 - 10\mu - 2\mu^2, non - MV \end{cases}$$



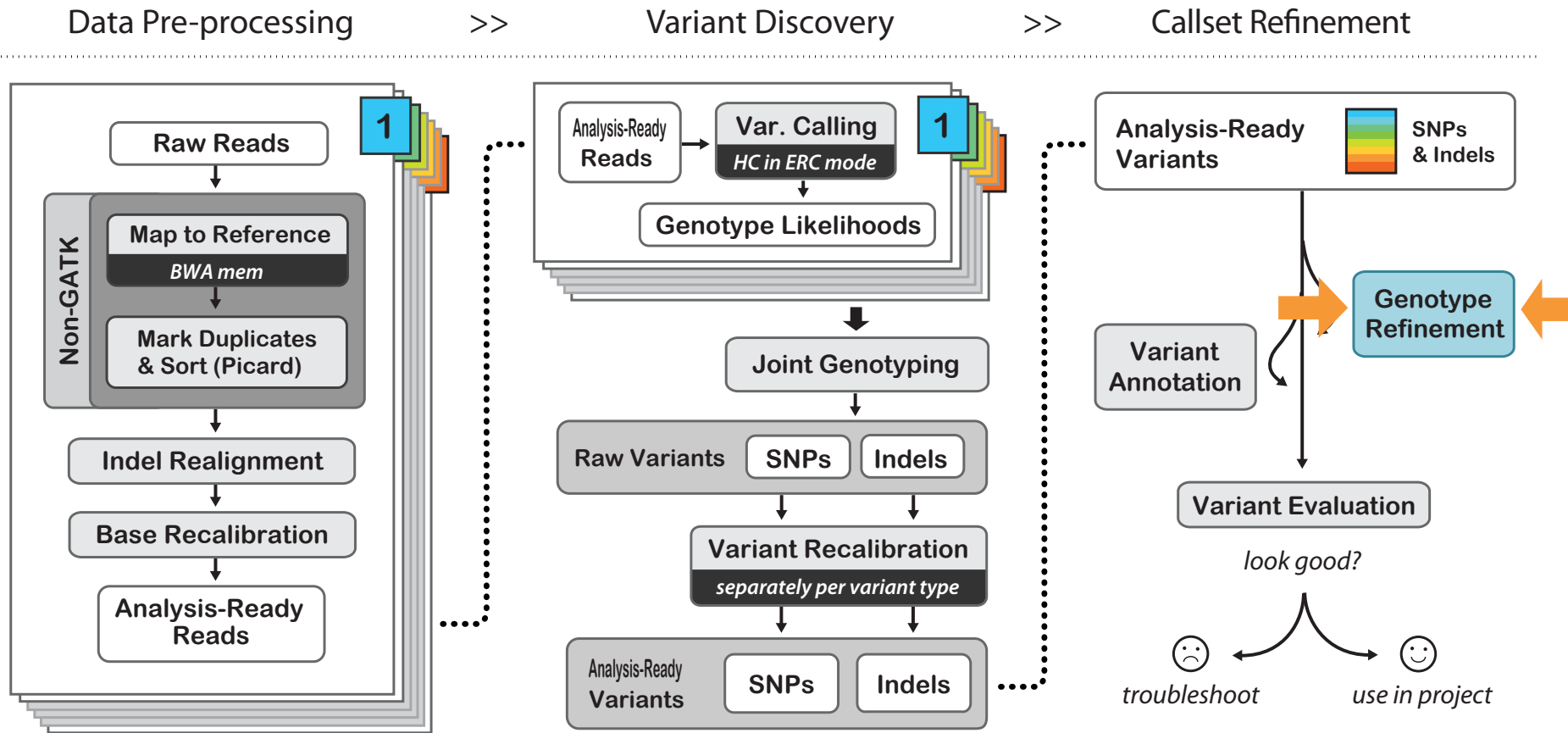
Sensitivity and specificity can be tuned as in VQSR

# Genotype refinement yields more high-quality genotypes

- Initial genotype calls may be ambiguous or wrong
- Applying population + family priors improves confidence
- More high confidence genotypes -> more data for downstream analysis!



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





## Further reading

<http://www.broadinstitute.org/gatk/guide/>

[https://www.broadinstitute.org/gatk/gatkdocs/  
org\\_broadinstitute\\_gatk\\_tools\\_walkers\\_variantutils\\_CalculateGenotypePosteriors](https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_variantutils_CalculateGenotypePosteriors)

<http://www.broadinstitute.org/gatk/guide/article?id=4723>

<http://www.broadinstitute.org/gatk/guide/article?id=4726>