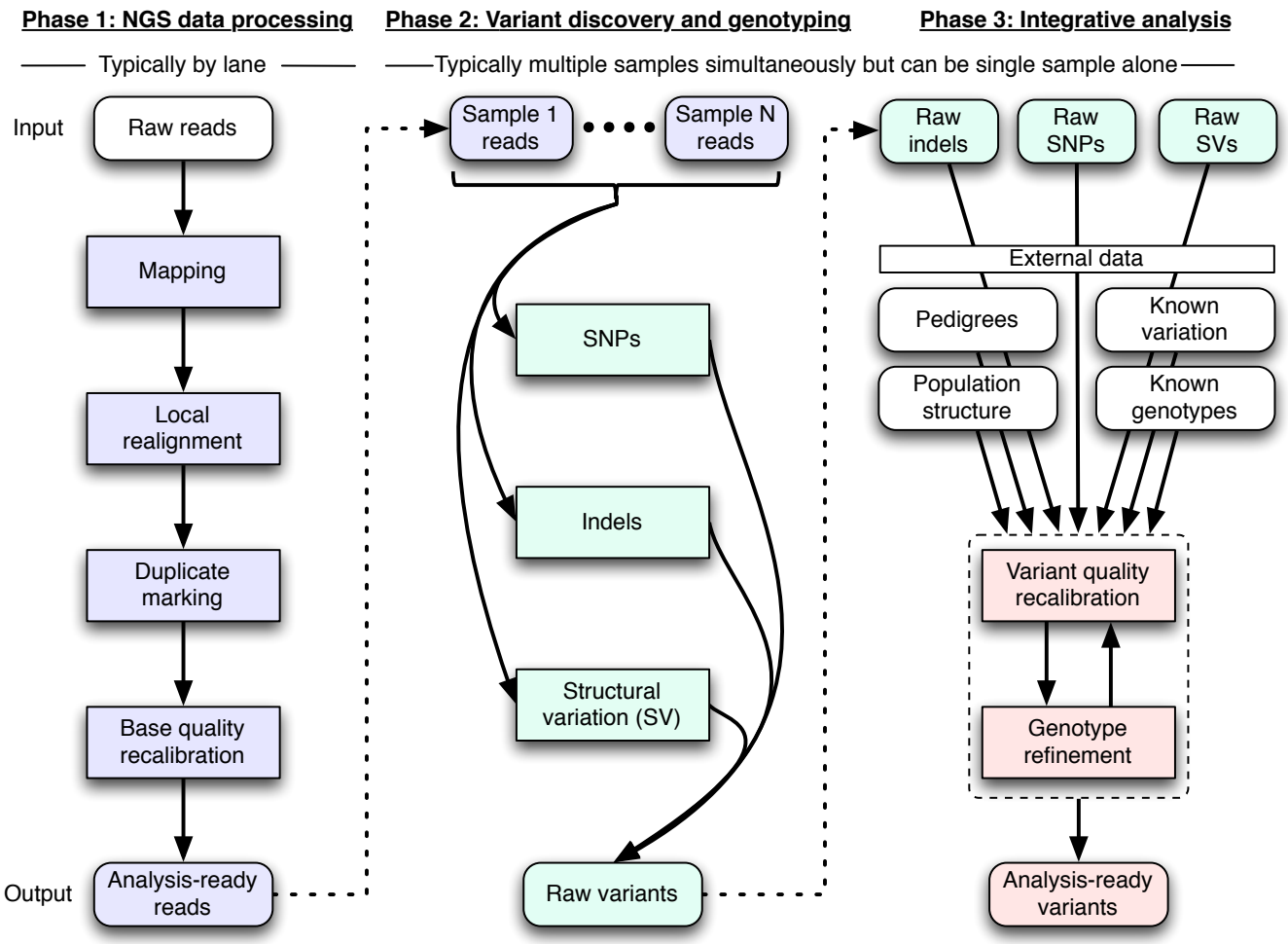
MPG NGS workshop:
SNP calling and error modeling

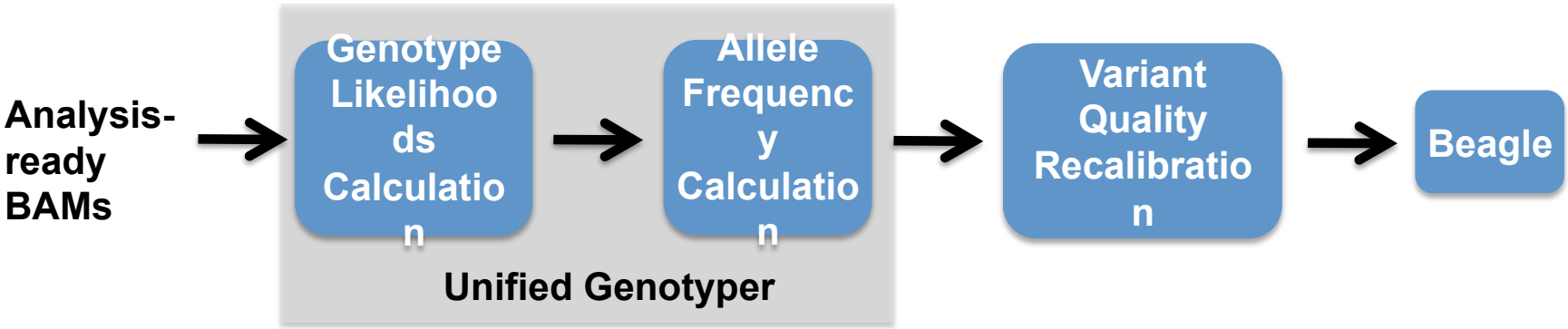
February 2011

Ryan Poplin
Genome Sequencing and Analysis
Medical and Population Genetics

The paradigm today

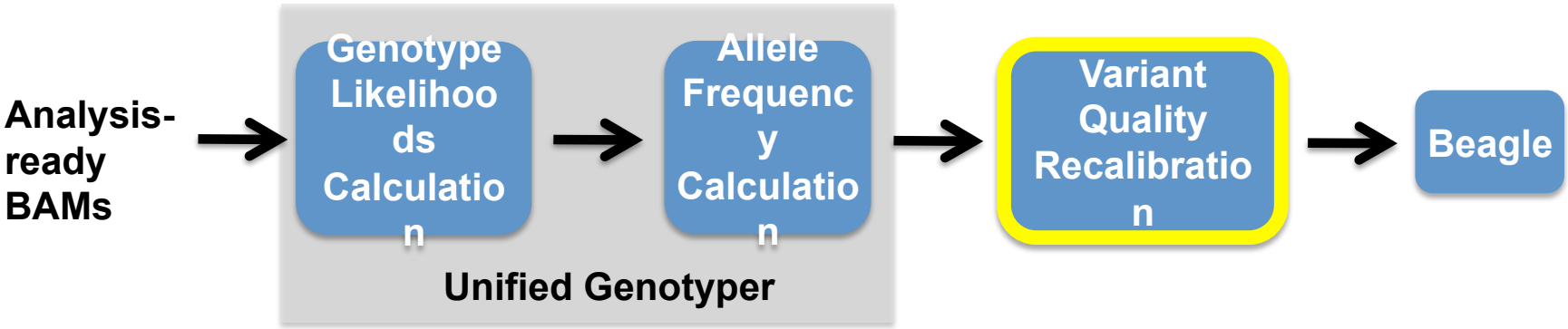


Step 2: SNP discovery



- We note that we no longer use any hard-filters (proximity to indel calls, clustered SNPs, etc.) at any point in the process.
- Unified Genotyper math and command lines discussed in previous meetings.
  (see Appendix for full details)

Step 3: SNP discovery



- The variant quality recalibration process has gone through a major overhaul recently. Most notably, we have removed any dependency on Ti/Tv in the calculation. This and further changes are highlighted in the following slides.

- Outline:
  - **Quick Variant Recalibration overview**
  - Contrastive clustering walkthrough
  - Ti/Tv-free quality thresholding or commitment-free probabilistic callsets

Variant annotations provide signal with which to remove artifacts!

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

```
22 49582364    .    A    G    198.96    .
   AB=0.67;
   AC=3;
   AF=0.50;
   AN=6;
   DP=87;
   Dels=0.00;
   HRun=1;
   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	AB	Allele balance of ref/alt in hets
AN	Total no. of chromosomes	HRun	Length of longest contiguous homopolymer
AF	Allele frequency	MQ	RMS MAPQ of all reads
DP	Depth of coverage	MQ0	No. of MAPQ 0 reads at locus
QD	QUAL score over depth	SB	Estimated SB score

Variant Quality Score Recalibration Model

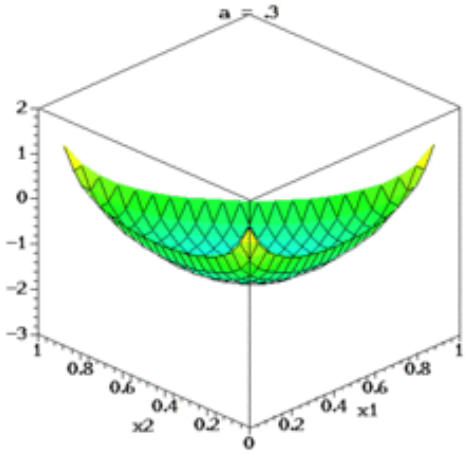
Gaussian Mixture Model trained on annotated variants, find MAP using VBEM:

$$p(\vec{c}) = \sum_{z} p(z)p(\vec{c} \mid z) = \sum_{k=1}^{K} \pi_k p(\pi_k) N(\vec{c} \mid \vec{\mu}_k, \Sigma_k) p(\vec{\mu}_k, \Sigma_k)$$

Dirichlet distribution

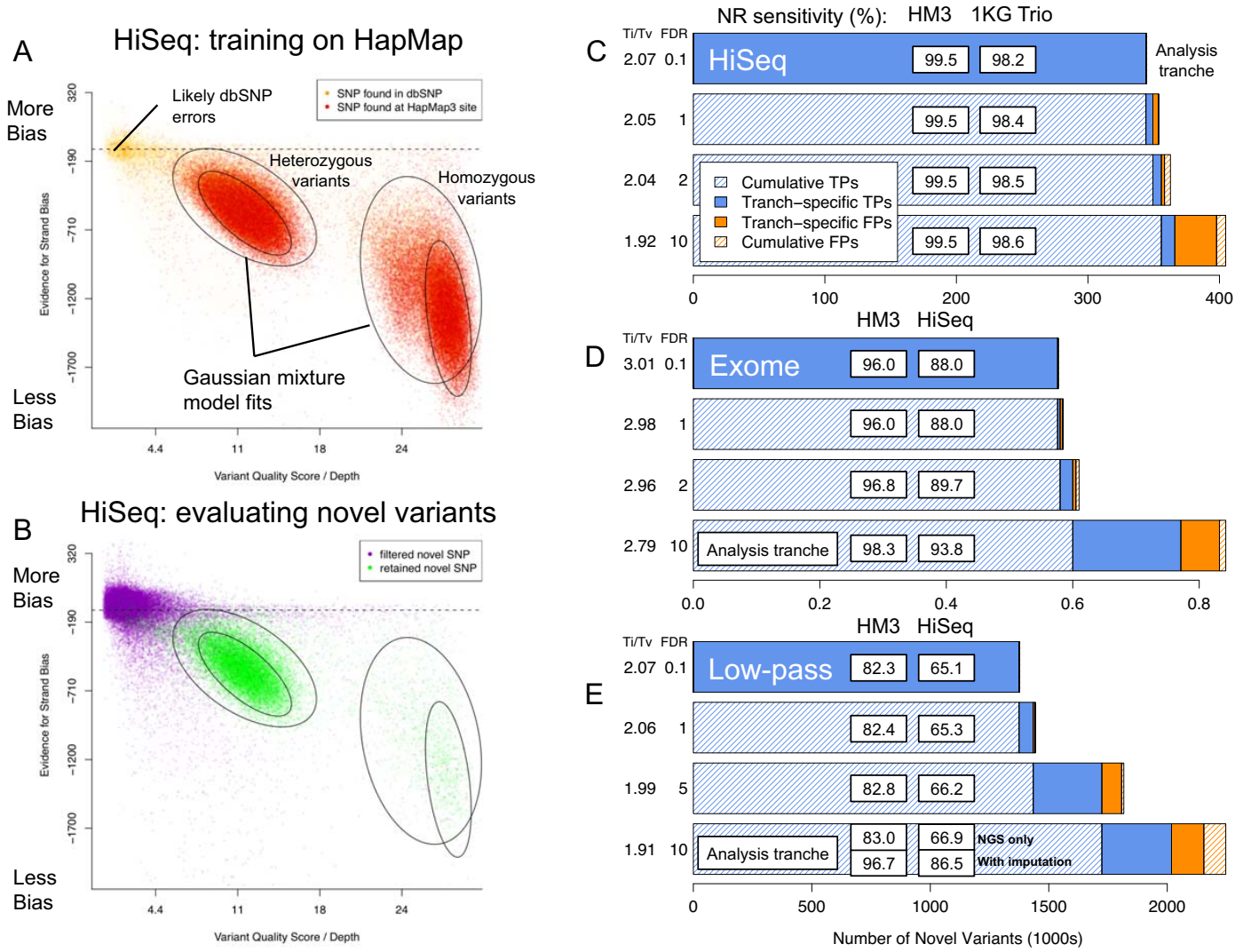
Normal – inverse Wishart distribution

Prior expectation is sparse set

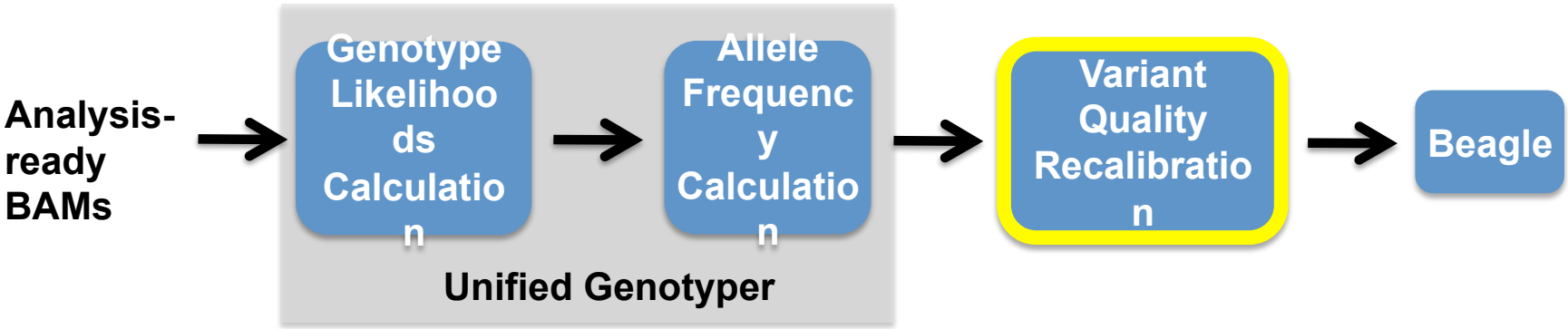


Prior expectation is the empirical mean and empirical covariance of the data.
Bias away from singularities.

Variant Quality Score Recalibration: training on highly confident known sites to determine the probability that other sites are true



Step 3: SNP discovery



- The variant quality recalibration process has gone through a major overhaul recently. Most notably, we have removed any dependency on Ti/Tv in the calculation. This and further changes are highlighted in the following slides.

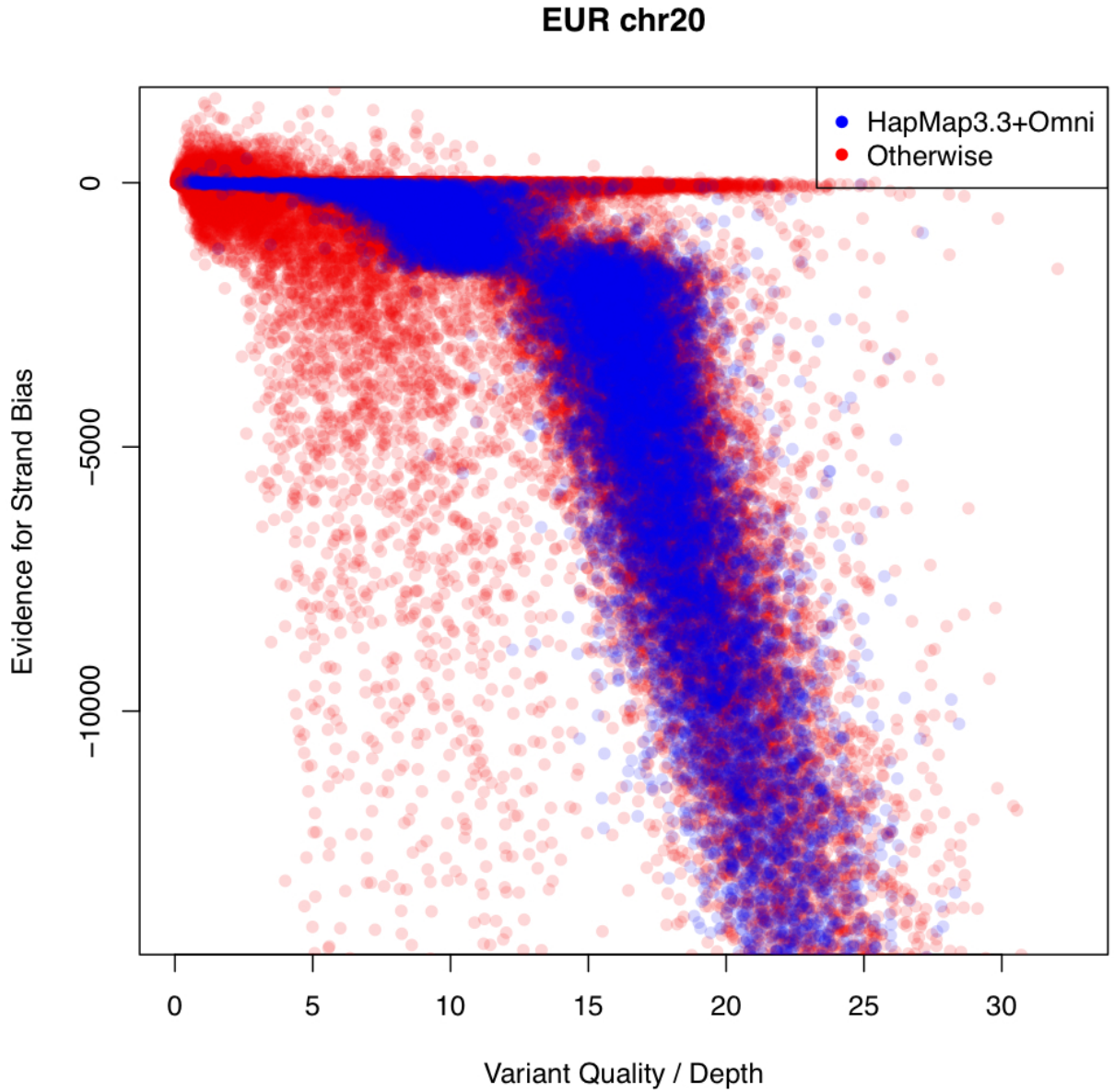
- Outline:
  - Quick Variant Recalibration overview
  - **Contrastive clustering walkthrough**
  - Ti/Tv-free quality thresholding or commitment-free probabilistic callsets

Running the Variant Quality Score Recalibrator

- Wiki page has full list of command lines broken out by the various steps in the process
- Wiki page also has links to all the data sets we recommend using as training data
- In a few weeks this whole process will be condensed into two much easier to use steps

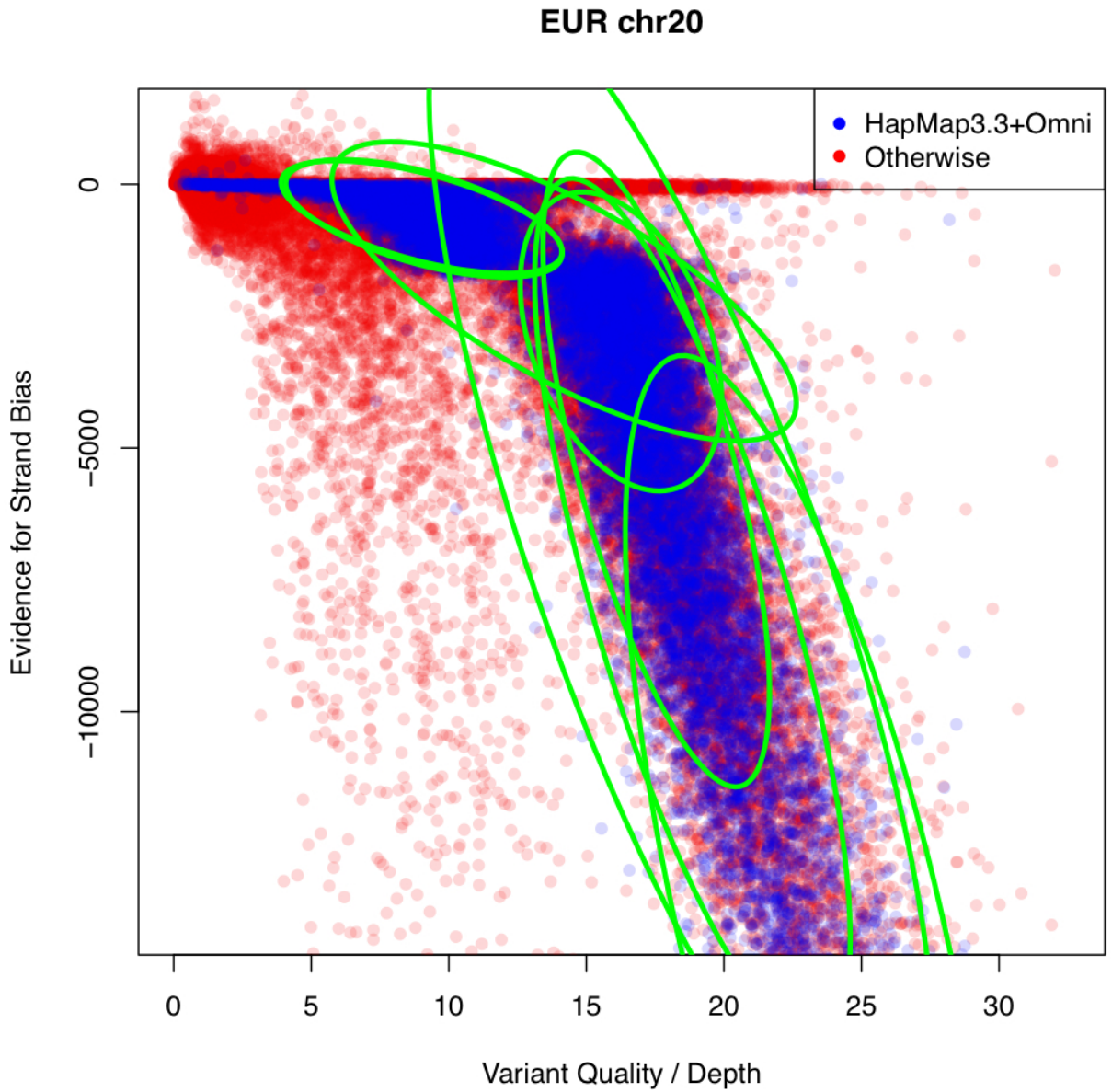
See http://www.broadinstitute.org/gsa/wiki/index.php/Variant_quality_score_recalibration

Contrastive VQSR Clustering Walkthrough



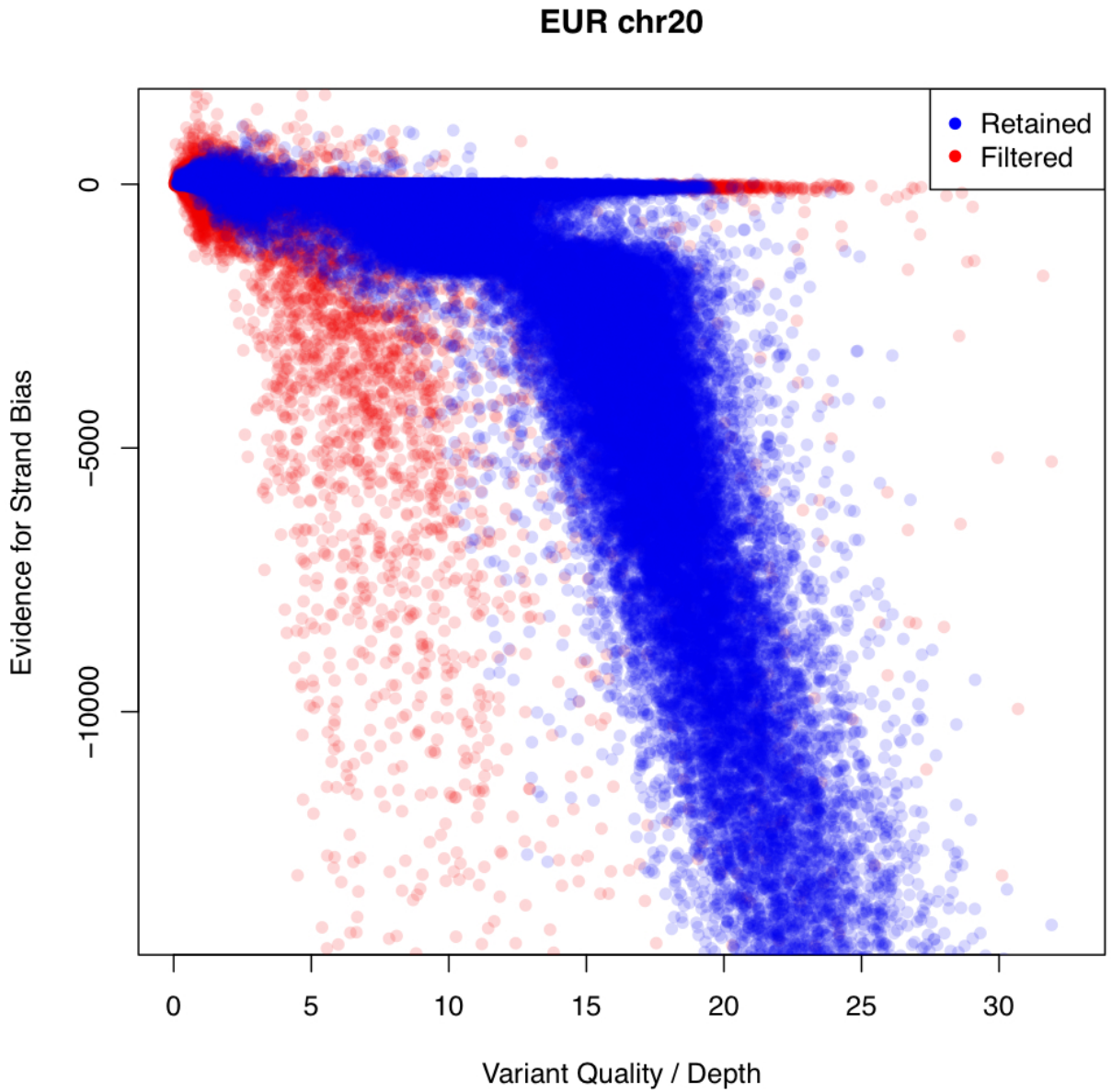
First partition the data into a training set by looking at sites which overlap with HapMap3.3 and the Omni chip.

Contrastive VQSR Clustering Walkthrough



Using Variational Bayes EM algorithm learn probability distribution over the training set.

Contrastive VQSR Clustering Walkthrough

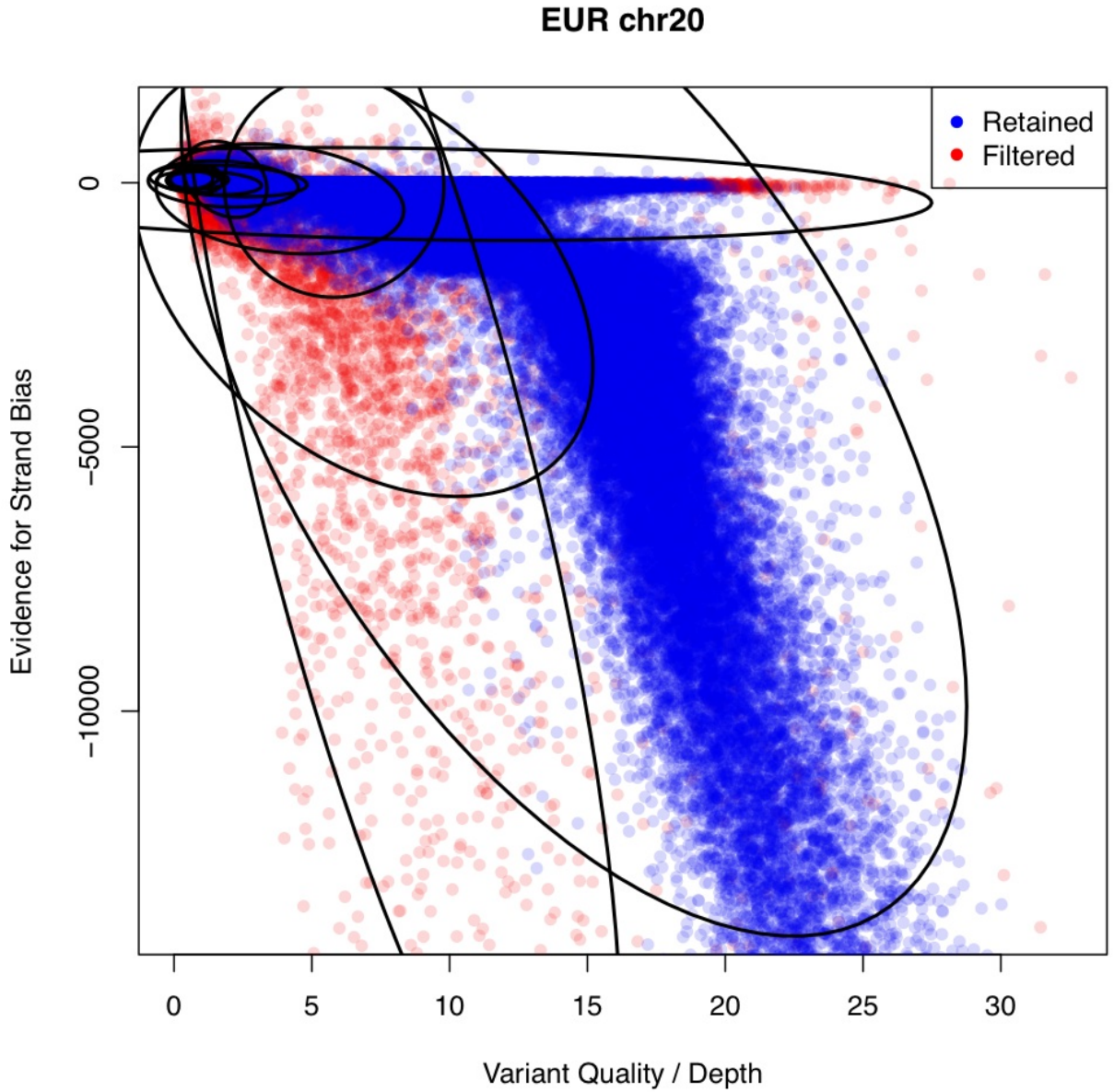


Assign a probability to each variant based on how well it clusters with the training set.

Unfortunately a sizeable number of seemingly good variants fall outside of the main clusters.

Furthermore, all clusters are essentially two-sided tests but most annotations are really only one-sided.

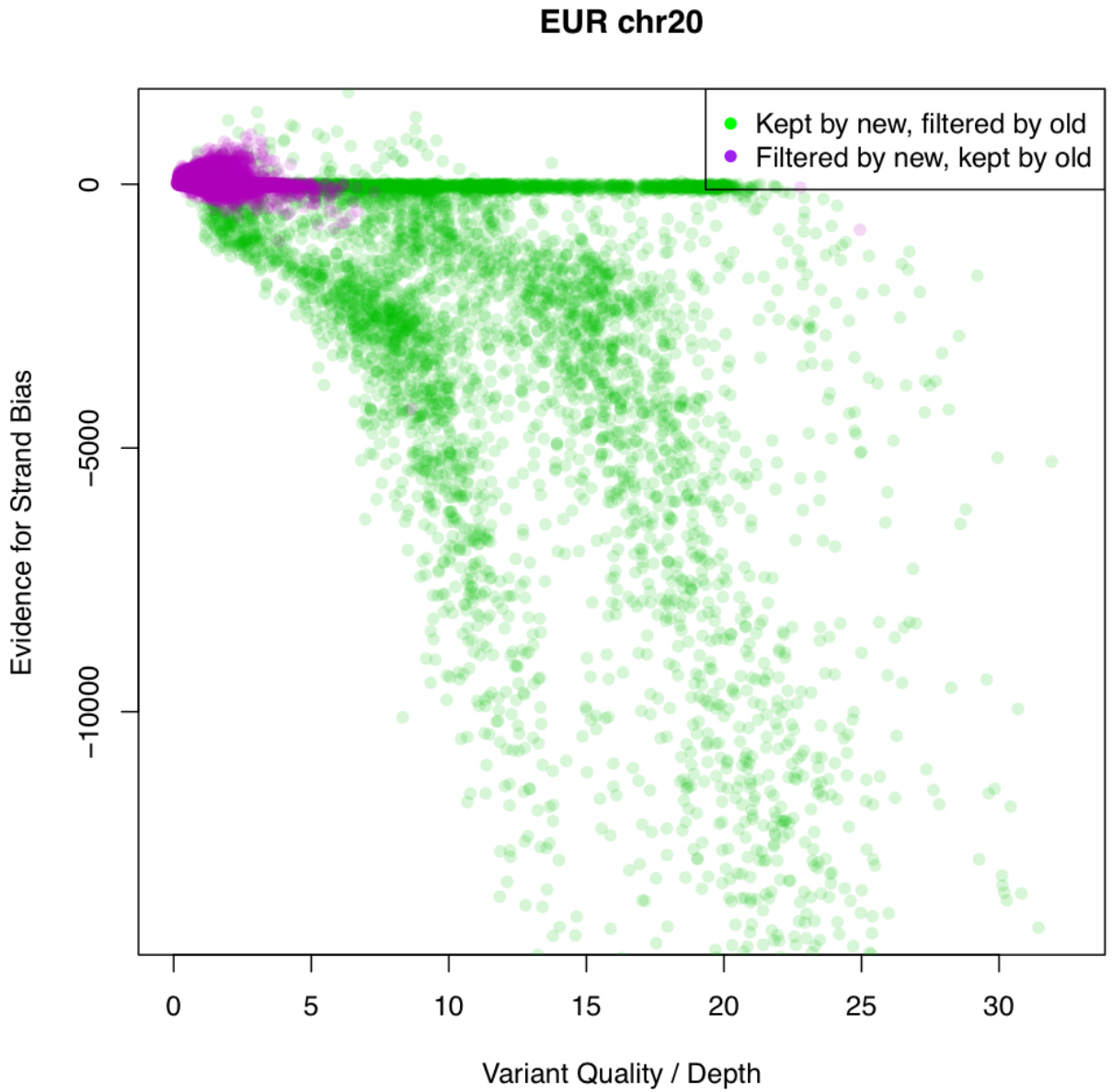
Contrastive VQSR Clustering Walkthrough



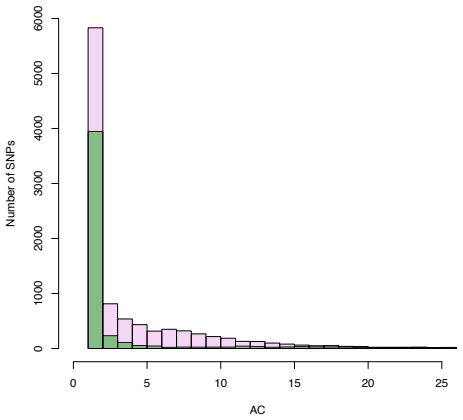
Solution: Train a second set of clusters based on the bottom 10% of variants which had the worst LOD.

This model for the bad variants allows for contrastive evaluation. New LOD score becomes difference between the good model and the bad model.

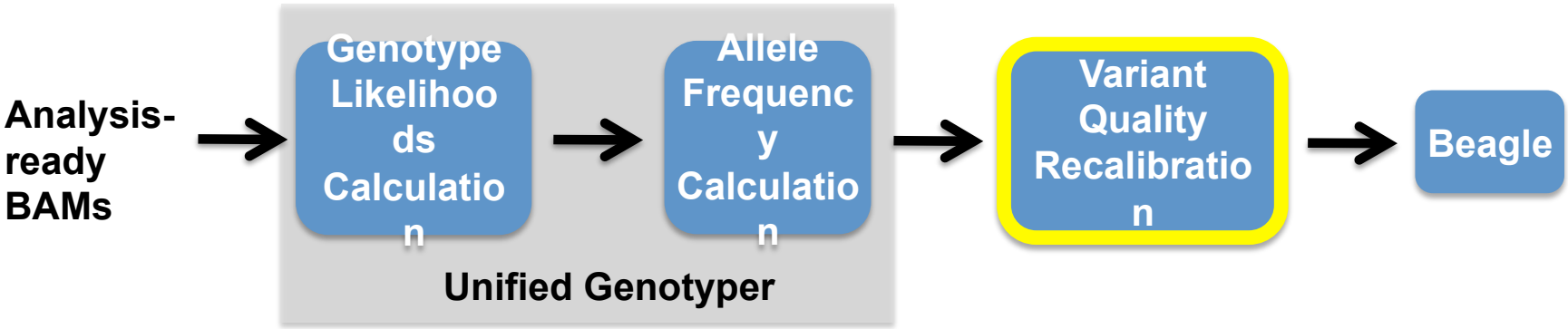
Contrastive VQSR Clustering Walkthrough



Contrastive VQSR clustering allows us to rescue the variants which fall outside of the main clusters but which also don't fit the model for bad variants.



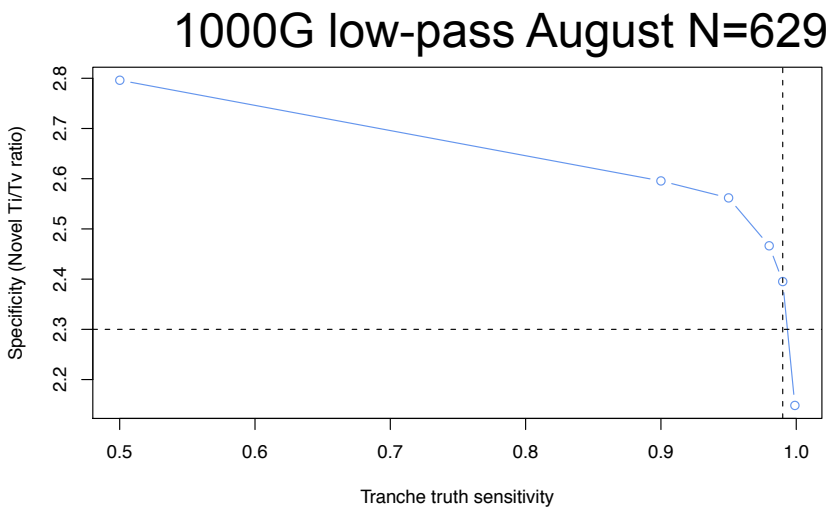
Step 3: SNP discovery

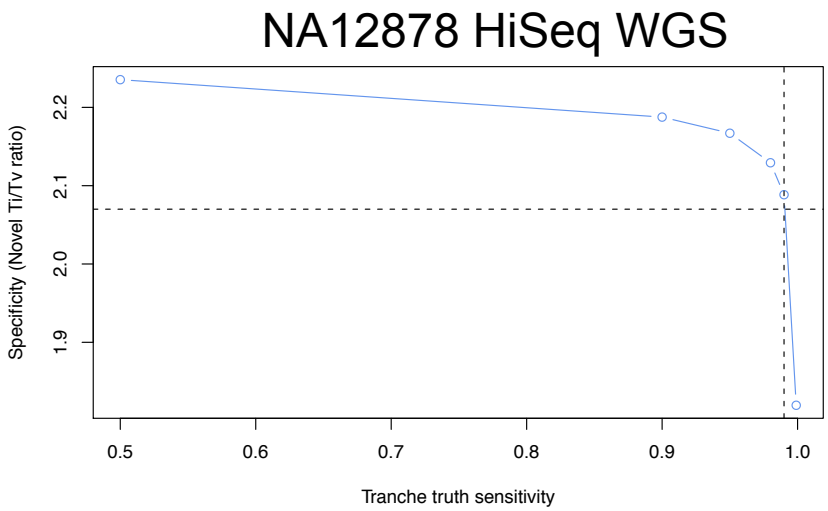


- The variant quality recalibration process has gone through a major overhaul recently. Most notably, we have removed any dependency on Ti/Tv in the calculation. This and further changes are highlighted in the following slides.

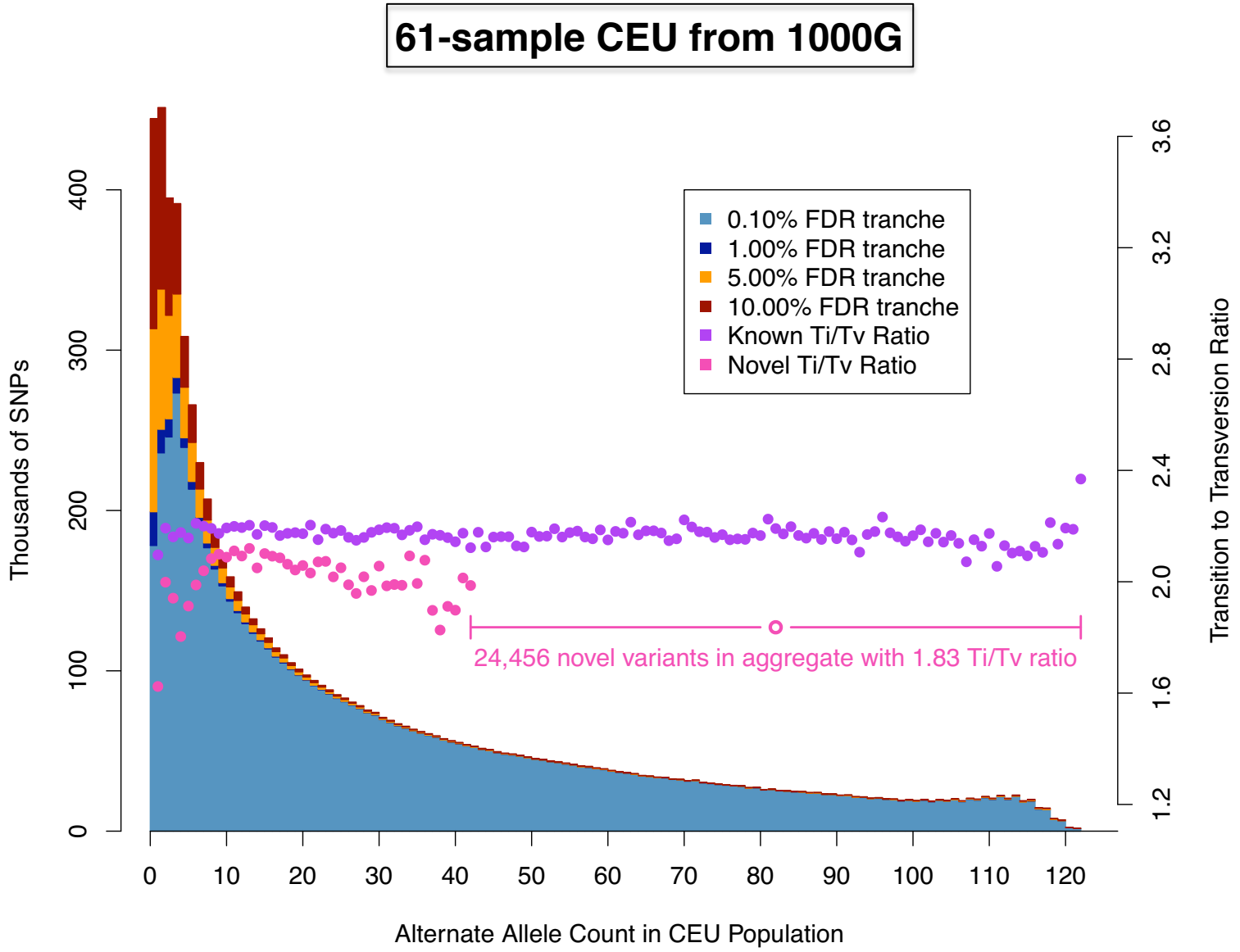
- Outline:
  - Quick Variant Recalibration overview
  - Contrastive clustering walkthrough
  - **Ti/Tv-free quality thresholding or commitment-free probabilistic callsets**

Sensitivity vs. specificity plots with the new Ti/Tv-less approach look good





The low confidence *tranches* are comprised of the low frequency events (most likely FPs)



Broad discovered the most variants at very high quality levels in 1000G chr20 bake-off exercise

# samples	Center	Total # variants	dbSNP % (129)	# knowns	Known ti/tv	# novels	Novel ti/tv	Includes genotype refinement?
1004	Broad	**765,365**	24.82	190,000	2.36	575,365	2.37	No
1004	BC	**733,155**	25.34	185,787	2.37	547,368	2.32	No
1004	Sanger	**728,374**	25.31	184,341	2.36	544,033	2.36	No
1004	UMich	**721,250**	26.46	190,871	2.33	530,379	2.35	Yes
1004	Oxford	**660,024**	27.44	181,095	2.38	478,929	2.38	Yes
1004	BCM	**605,274**	29.98	181,444	2.33	423,830	2.29	Yes
1004	NCBI	**601,907**	29.26	176,150	2.39	425,757	2.57	No

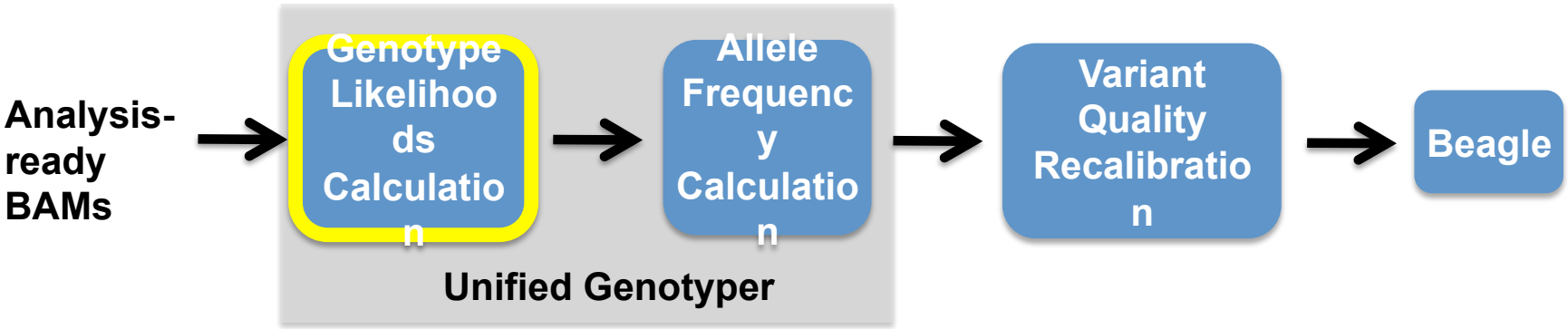
Final Thoughts

- Our data processing pipeline produces really good SNP calls. The same pipeline is used for whole exome and WGS, both deep and low-pass sequencing. Short indel calls too!
- Anything can be used as truth data. Validation assays, several 1000G callsets, or auto-generate your own by subsetting to the highest quality SNPs
- There is no reason to decide between high sensitivity or high specificity. Just use a probabilistic callset.
- The tools are available to all:

http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit

Appendix

Step 2a: SNP discovery



- The genotype likelihoods calculation now takes overlapping read pairs (where bases are not independent observations) into account, which we term “fragment-based calling”.

GATK single sample genotype likelihoods

Bayesian model

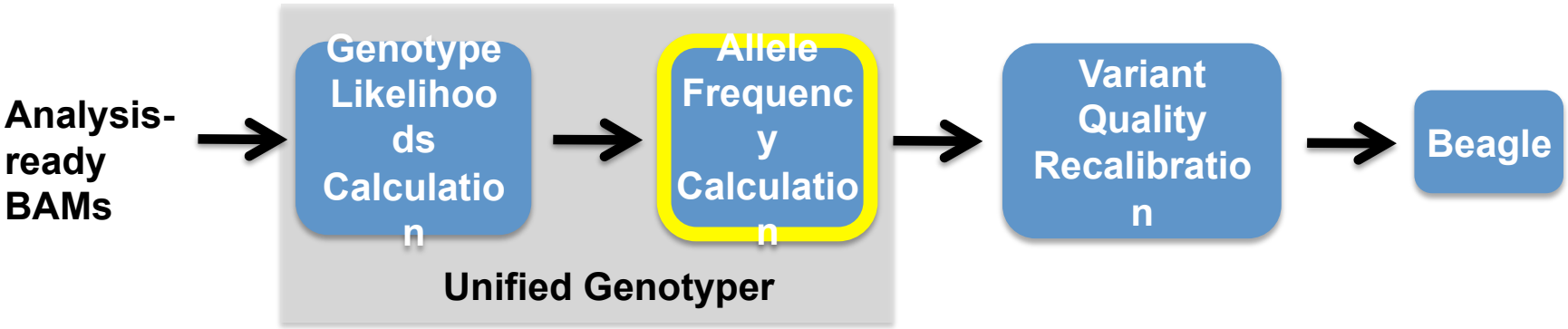
Likelihood for the genotype | Prior for the genotype | Likelihood of the data given the genotype | Independent base model

$$L(G \mid D) = P(G)P(D \mid G) = \prod_{b \in \{good\_bases\}} P(b \mid G)$$

- Priors applied during multi-sample calculation; P(G) = 1
- Likelihood of data computed using pileup of bases and associated quality scores at given locus
- Only "good bases" are included: those satisfying minimum base quality, mapping read quality, pair mapping quality, NQS
- *P(b | G)* uses calibrated base quality score
- *L(G|D)* computed for all 10 genotypes

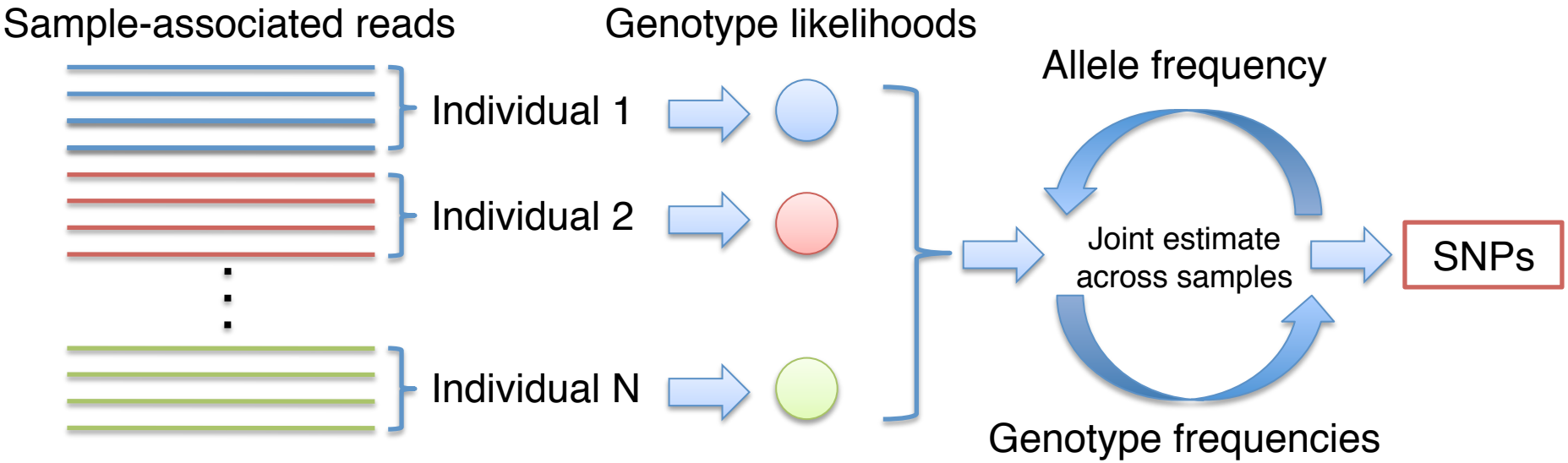
See http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper for more information

Step 2b: SNP discovery



- We now use Heng Li's Exact model to calculate P(AF>0) instead of our previous heuristic grid search model.

We apply a generalization of the single sample SNP caller for multi sample data



- This approach allows us to combine weak single sample calls to discover variation among several samples with high confidence

See http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper for more information

Running the Unified Genotyper

```
java -Xmx2048m -jar GenomeAnalysisTK.jar
    -R /broad/1KG/reference/human_b37_both.fasta
    -T UnifiedGenotyper
    -B:dbsnp,VCF dbsnp_132_b37.vcf
    -o NA19240.raw.vcf
    -stand_call_conf 30
    --heterozygosity 1.000000e-03
    -I NA19240.SLX.bam
```

Minimum phred-scaled confidence required to emit a SNP

1 het per 1000 reference bases on average for a Yoruban

BAM file containing NA19240 SLX reads

Long string of variant annotations (more info in a few slides)

Raw VCF calls (NA19240.raw.vcf)

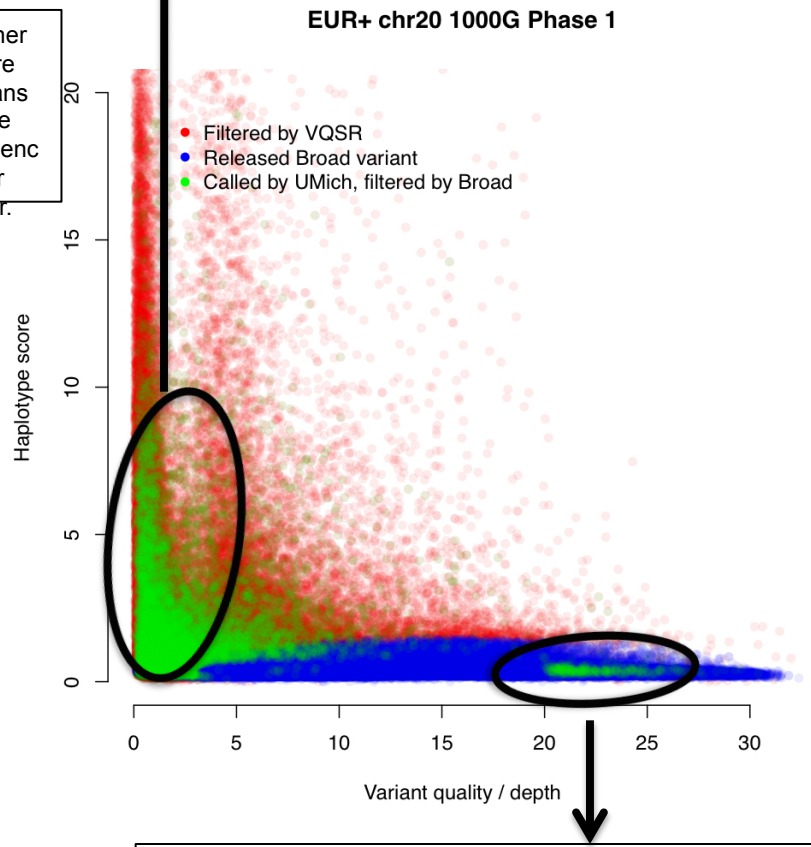
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA19240
1	36496	.	T	A	53.13	.	<ATTRIBUTES>	GT:DP:GQ	1/0:6:84.70
1	45162	rs10399749	C	T	331.37	.	<ATTRIBUTES>	GT:DP:GQ	0/1:27:99.00
1	48677	.	G	A	399.86	.	<ATTRIBUTES>	GT:DP:GQ	1/0:25:99.00

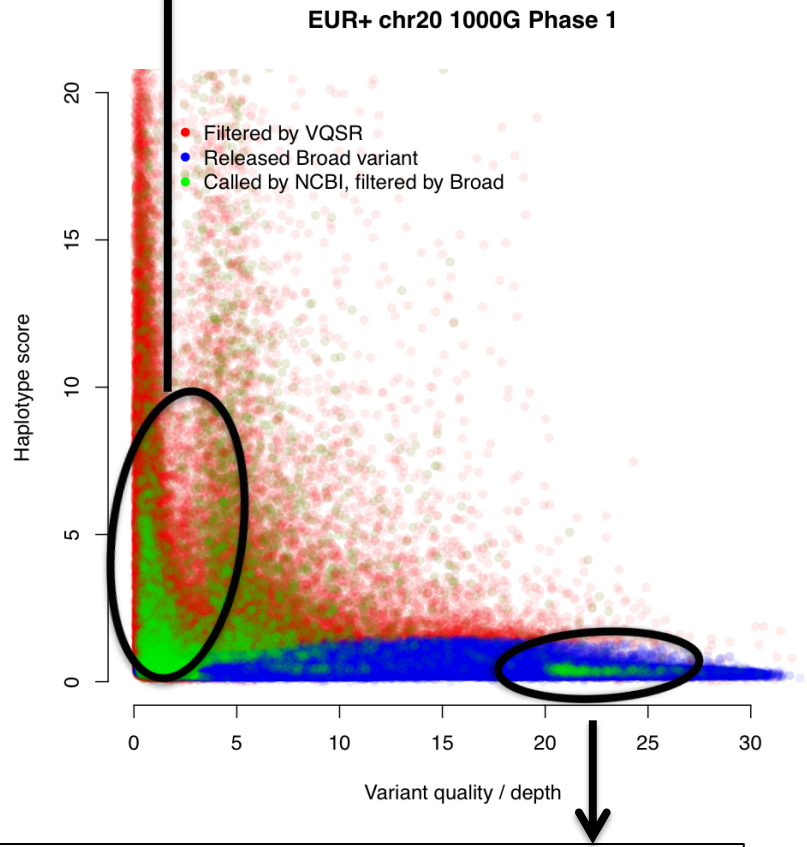
See http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper for more information

Variants with bad Haplotype Scores often exhibit good Ti/Tv ratios and are included in other centers' callsets, but are likely FPs

Bad sites being called by other centers but correctly filtered by the Broad.

Higher score means more evidence for error.





These sites are potentially bad in other SNP annotation dimensions.