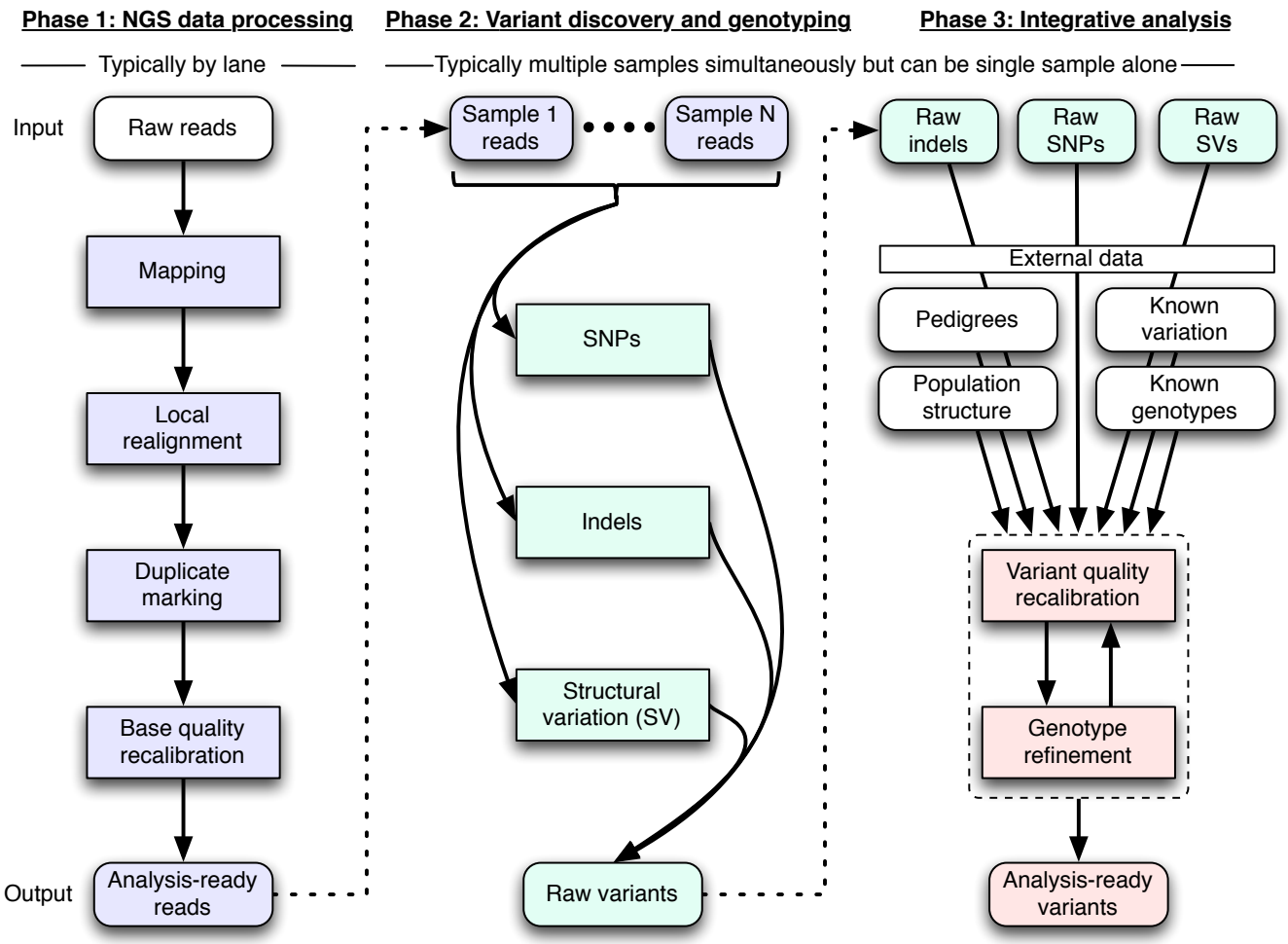
Variant Quality Score Recalibration

Ryan Poplin
Genome Sequencing and Analysis
Program in Medical and Population Genetics

301 Binney Floor Meeting
June 30, 2011

Variant discovery workflow



Outline

- Variants as points in a point cloud can be modeled using a Gaussian mixture model
- Details of individual component statistics
- Exposition of results
  - 1000 Genomes Project SNPs and indels
  - MPG Exome SNPs

Outline

- Variants as points in a point cloud can be modeled using a Gaussian mixture model
- Details of individual component statistics
- Exposition of results
  - 1000 Genomes Project SNPs and indels
  - MPG Exome SNPs

Variant annotations provide signal with which to remove artifacts!

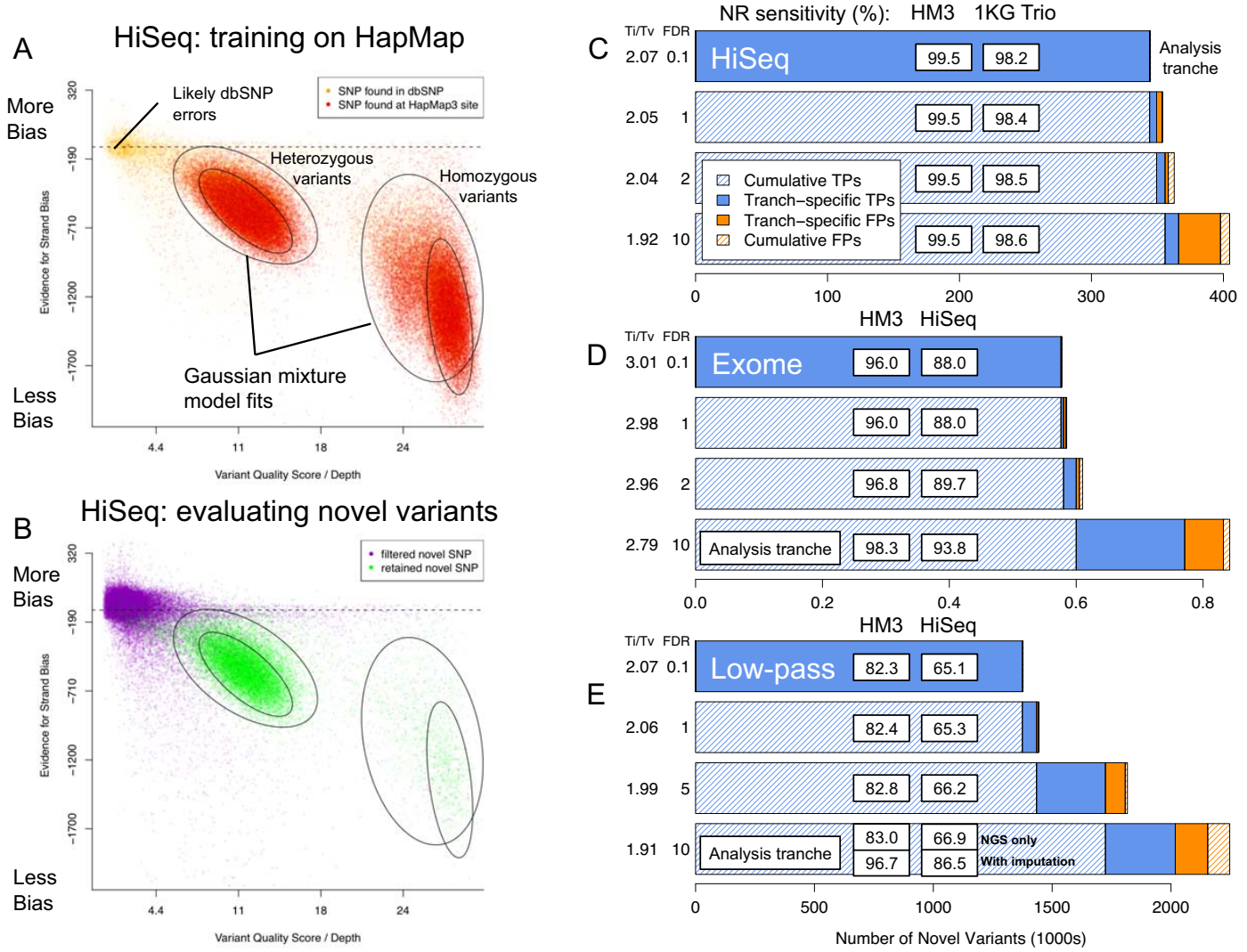
VCF record for an A/G SNP at 22:49582364

```
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 strand bias score

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



DePristo et al. Nature Genetics. 2011

Variant Quality Score Recalibration Model

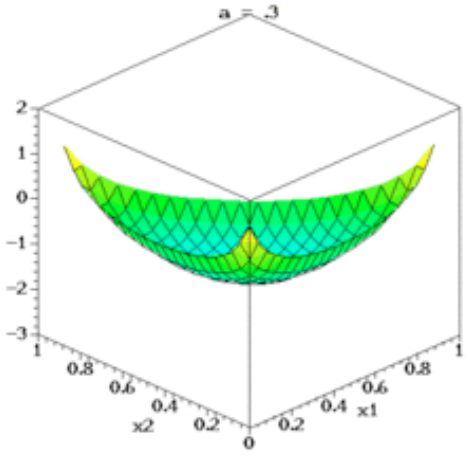
Gaussian Mixture Model trained on annotated variants, find MAP using VBEM:

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

Dirichlet distribution

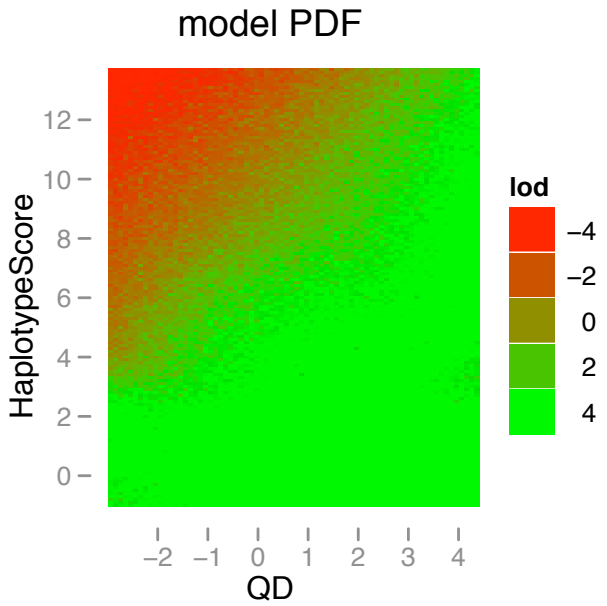
Normal – inverse Wishart distribution

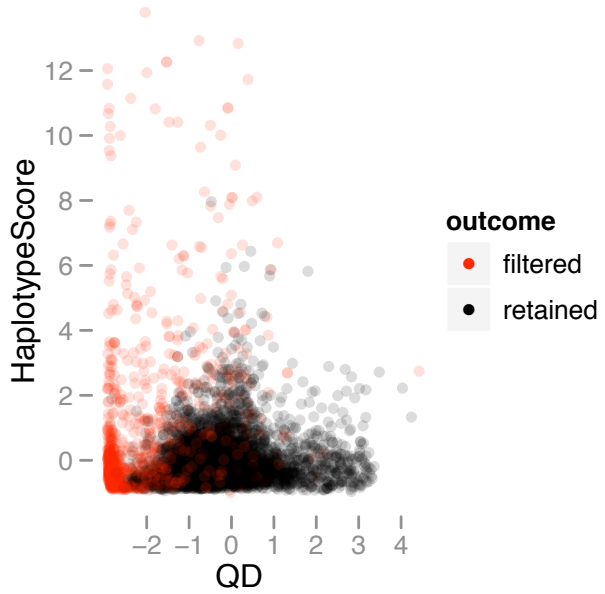
Prior expectation is sparse set



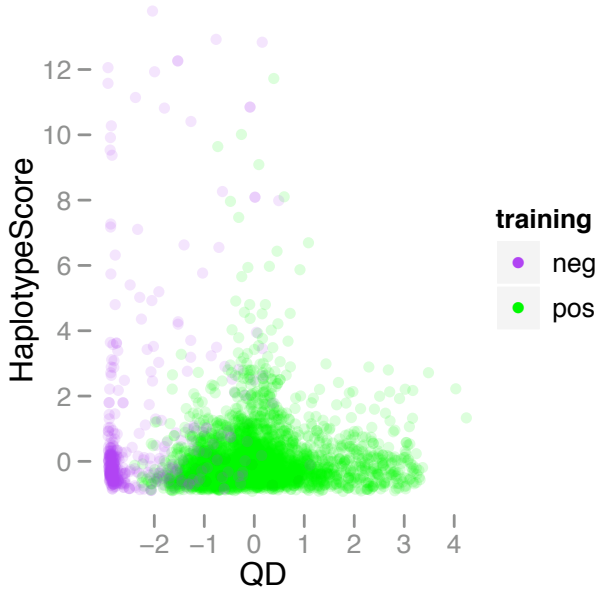
Prior expectation is the empirical mean and empirical covariance of the data. Bias away from singularities.

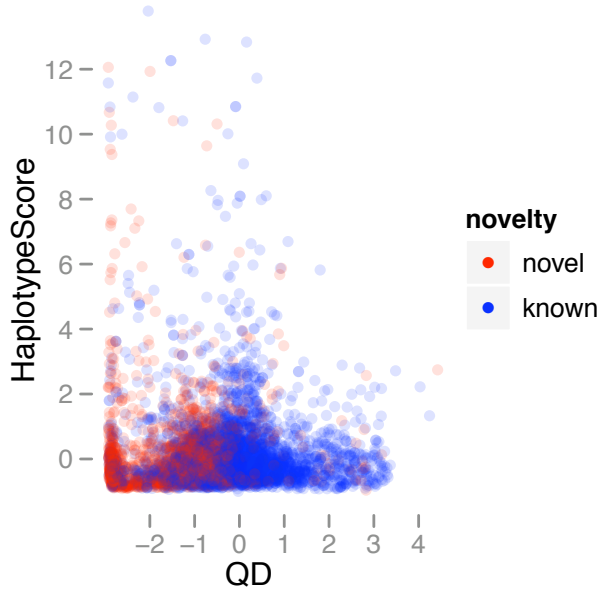
Mixture model plots generated by VQSR





Contrastive evaluation is an important feature of the algorithm





Running the Variant Quality Score Recalibrator

- Wiki page has example command lines and a tutorial using HiSeq data
- Wiki page also has links to all the data sets we recommend using for training data
- Tips for interpreting the model plots and advice for non-human data

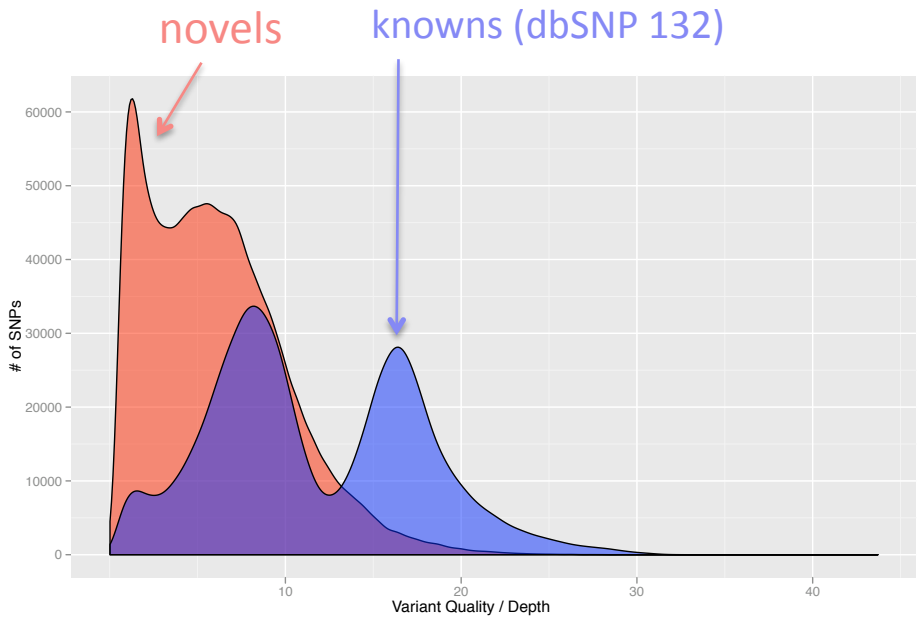
See http://www.broadinstitute.org/gsa/wiki/index.php/Variant_quality_score_recalibration

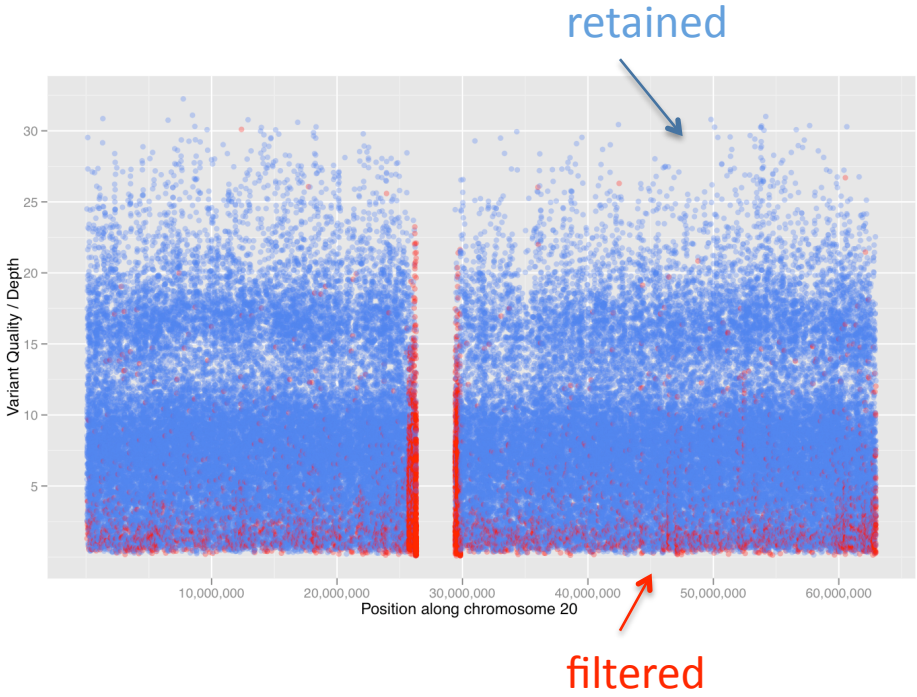
Outline

- Variants as points in a point cloud can be modeled using a Gaussian mixture model
- Details of individual component statistics
- Exposition of results
  - 1000 Genomes Project SNPs and indels
  - MPG Exome SNPs

QD: Variant Quality / Depth

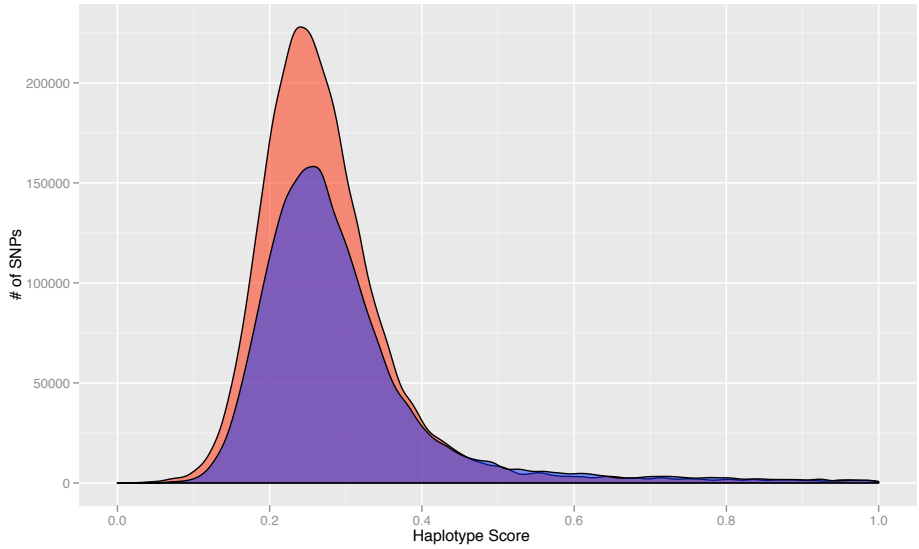
- Confidence in the site being variant should increase with increasing depth

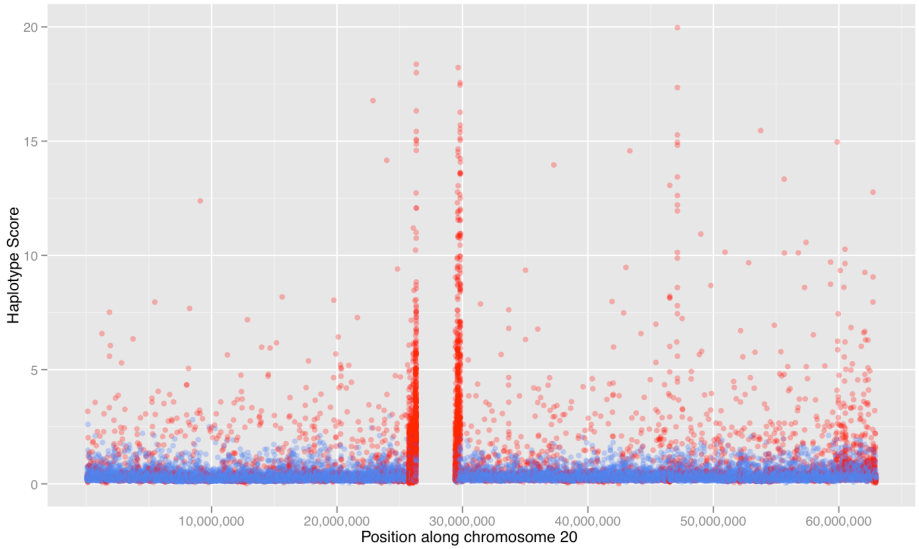




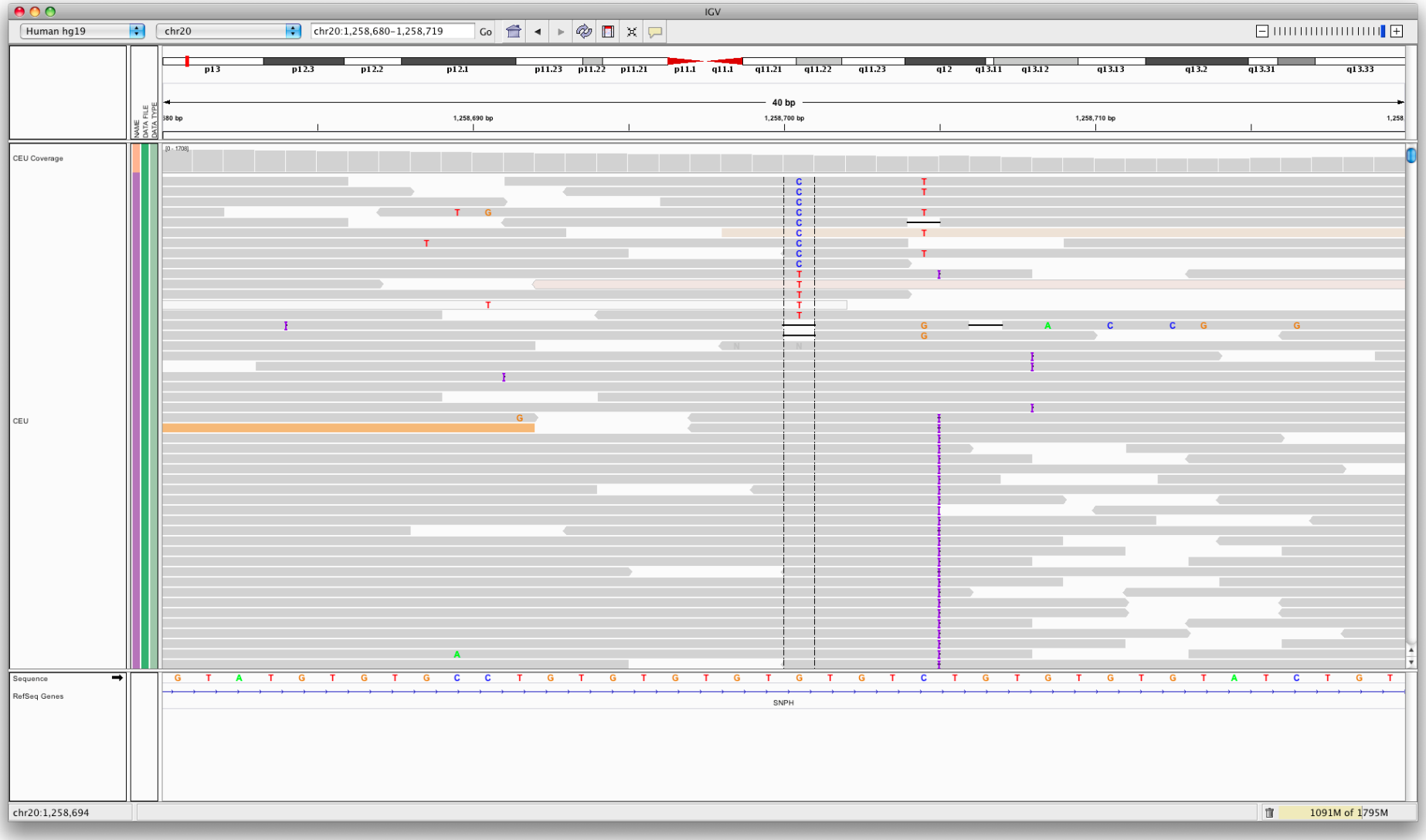
Haplotype Score

- Probability that the reads in a window around the variant can be explained by at most two haplotypes





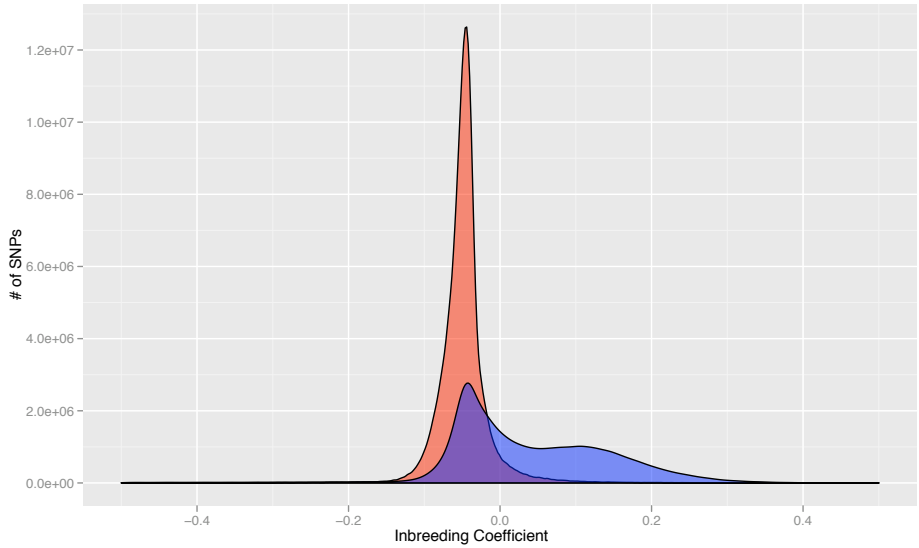
SNP removed by HaplotypeScore

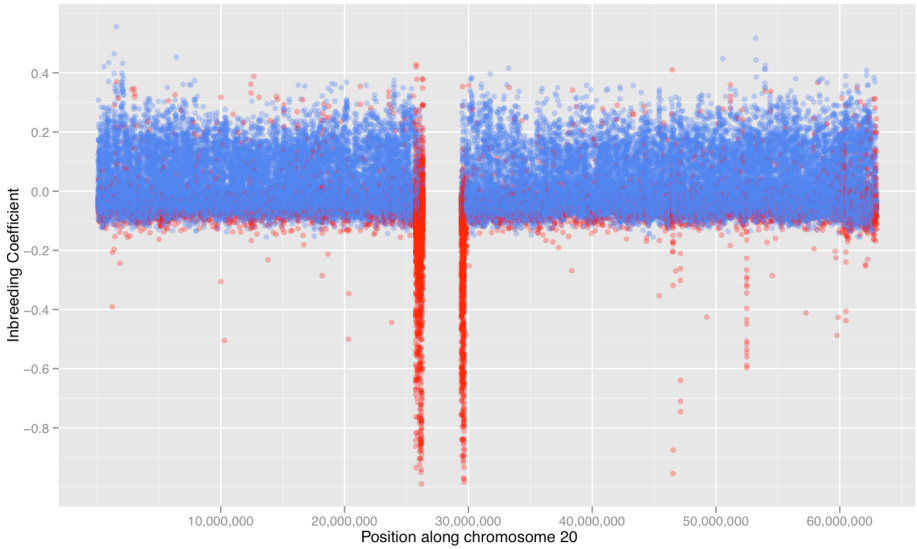


Inbreeding Coefficient

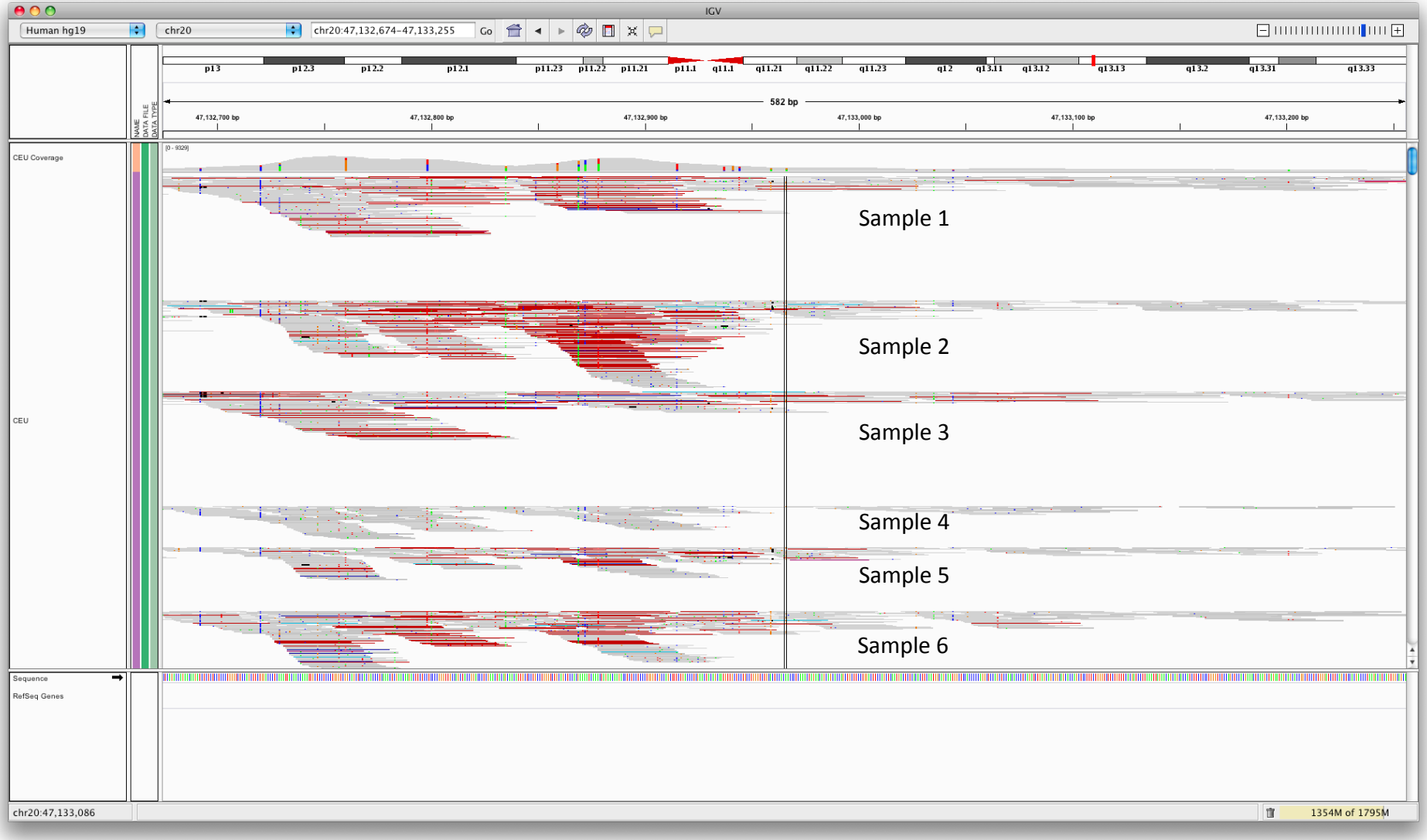
- Extreme Hardy-Weinberg violations (especially all heterozygous sites) are likely false positives

$$F = \frac{\mathrm{E}(f(\mathrm{Aa})) - \mathrm{O}(f(\mathrm{Aa}))}{\mathrm{E}(f(\mathrm{Aa}))} = 1 - \frac{\mathrm{O}(f(\mathrm{Aa}))}{\mathrm{E}(f(\mathrm{Aa}))}$$



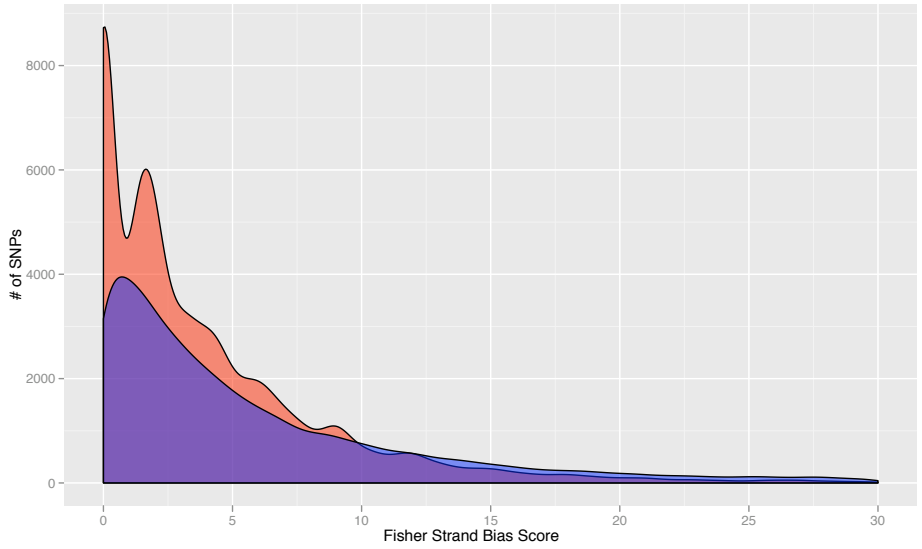


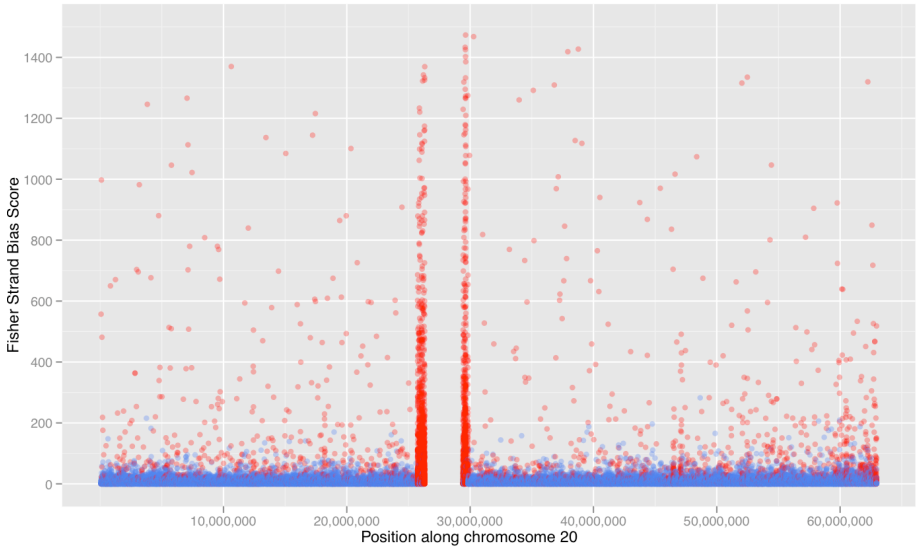
All-het SNPs removed by Inbreeding Coefficient



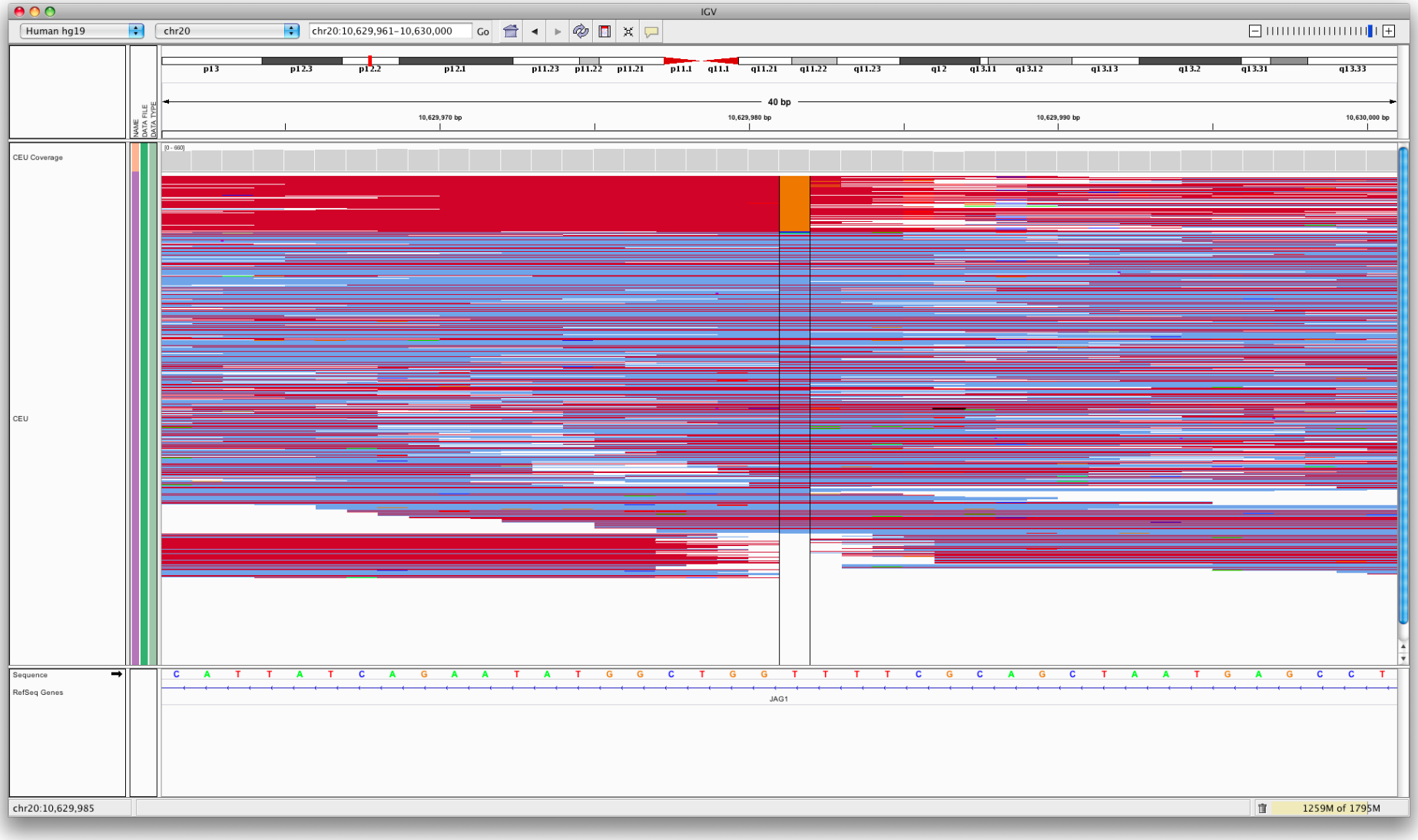
FS: Fisher Exact Test of Read Strand

- If the reference-carrying reads are balanced between forward and reverse strands then the alternate-carrying reads should be as well



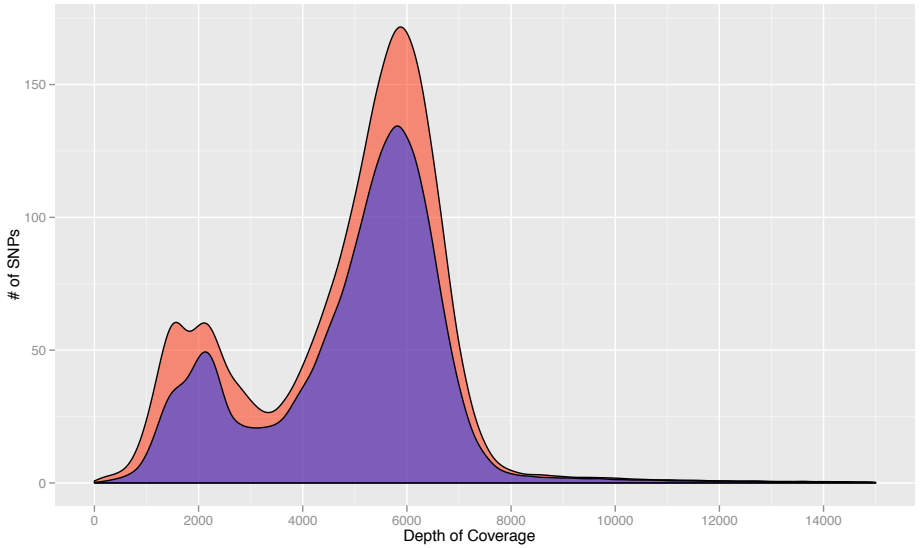


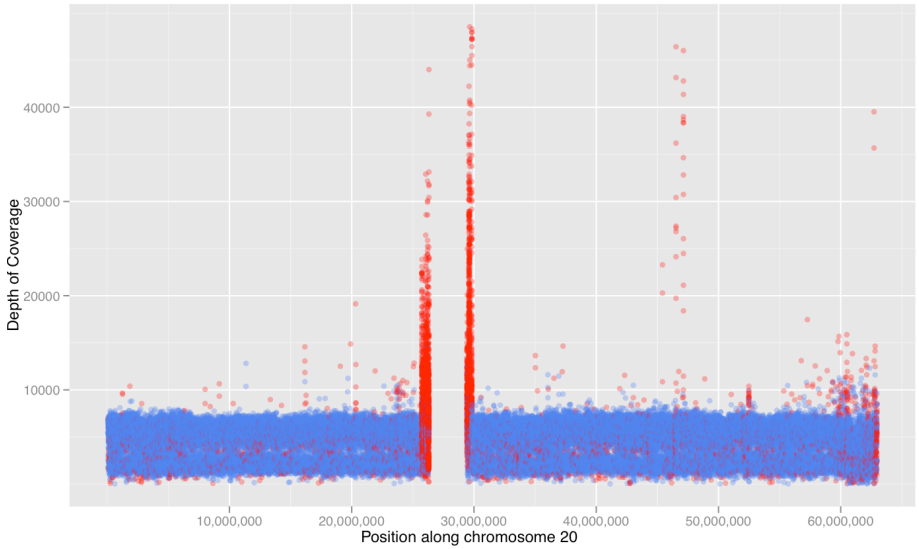
Classic strand biased SNP removed by FS



DP: Depth of Coverage

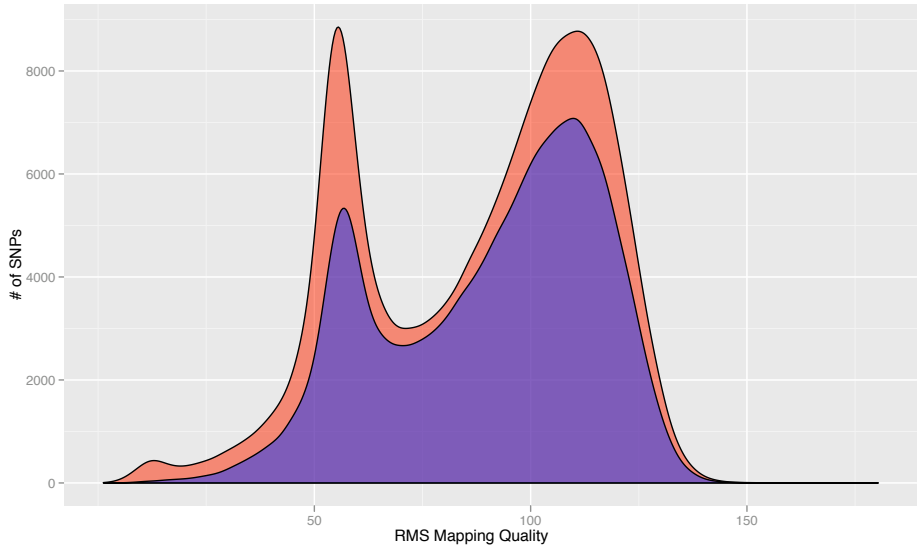
- Excessive piling up of reads is indicative of poorly mapped region

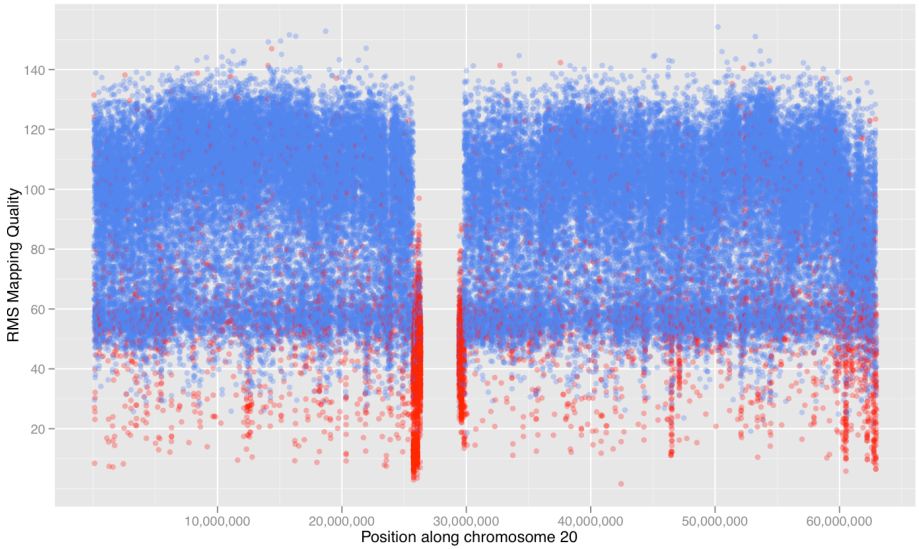




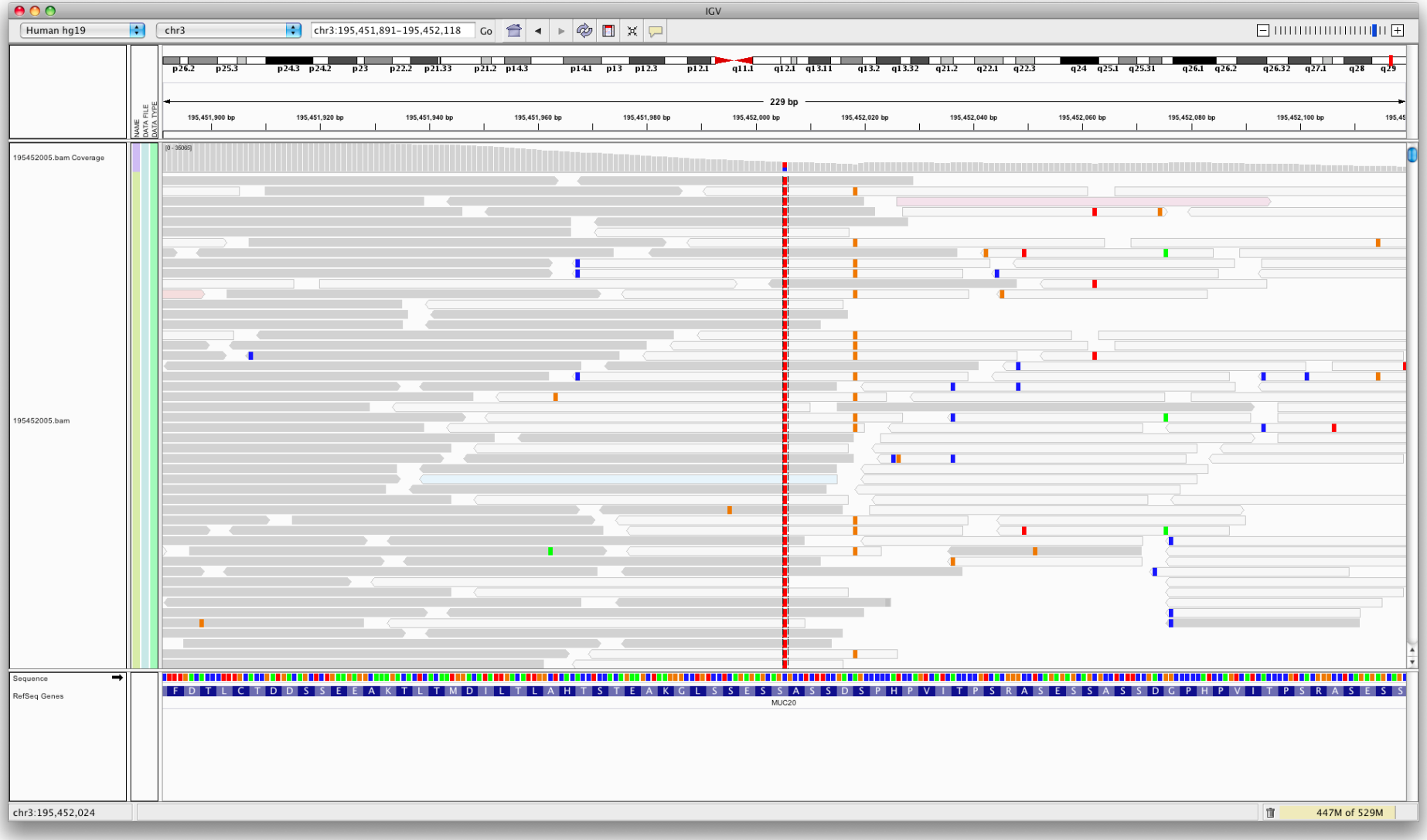
MQ: RMS Mapping Quality

- Regions of excessively low mapping quality are ambiguously mapped and variants called within are suspicious



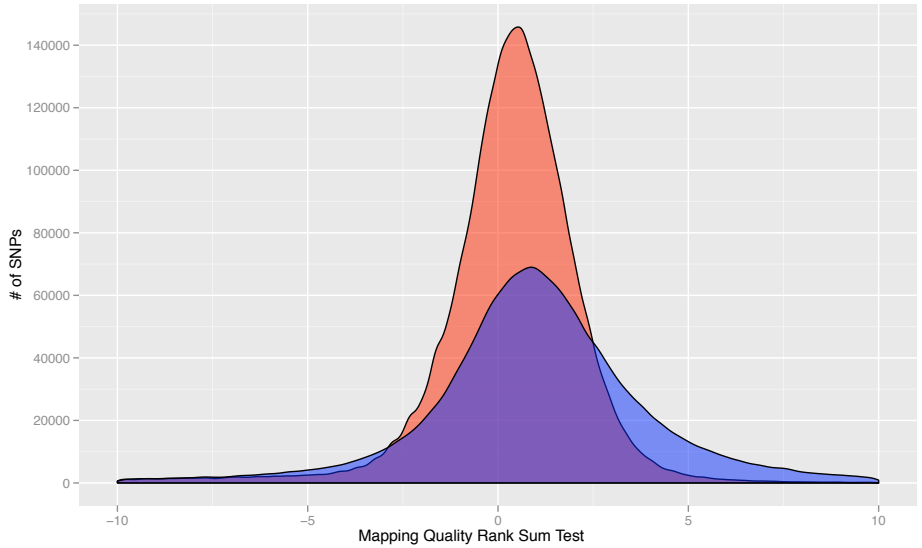


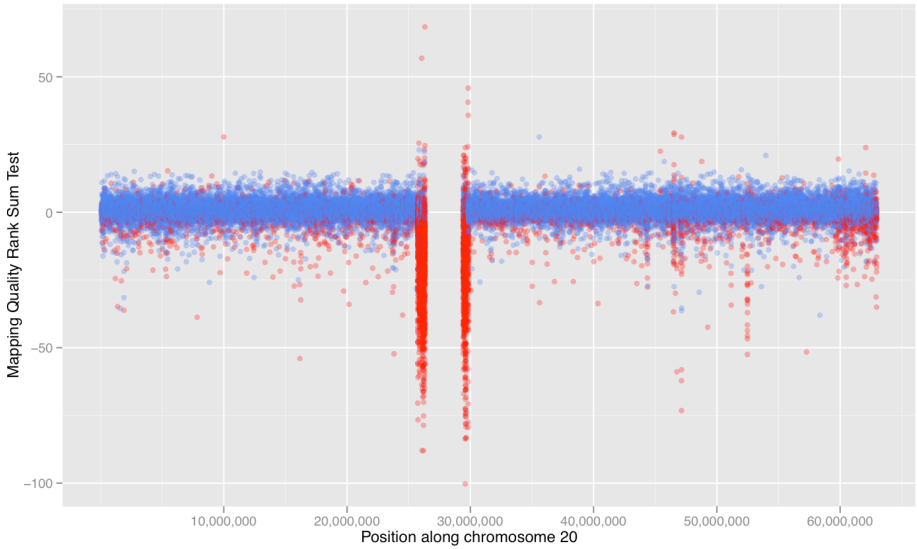
Removed SNP in bad MQ region



Mapping Quality Rank Sum Test

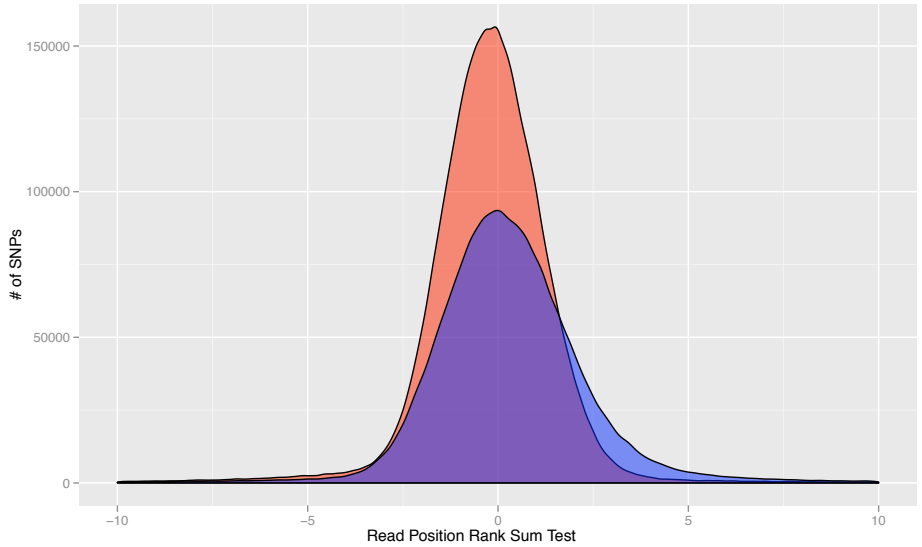
- If the alternate bases are more likely to be found on reads with lower MQ than reference bases then the site is likely mismapped

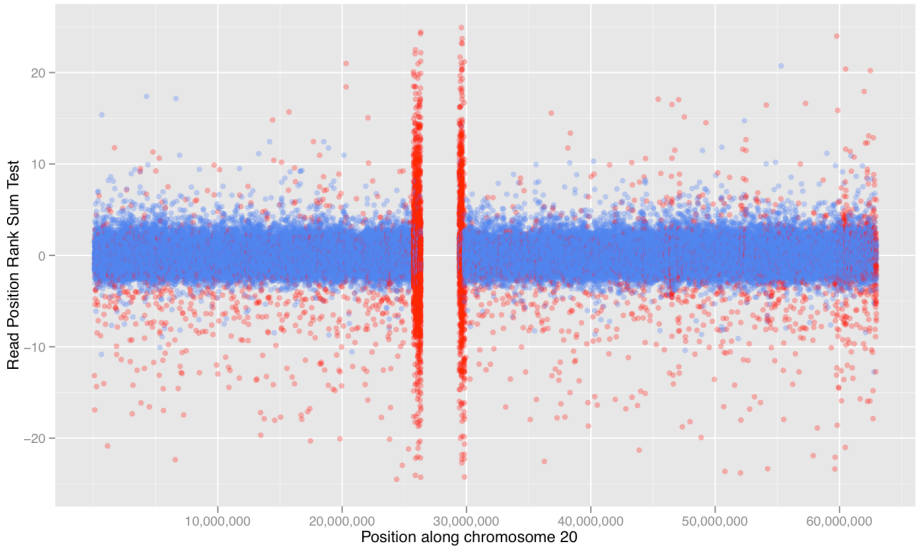




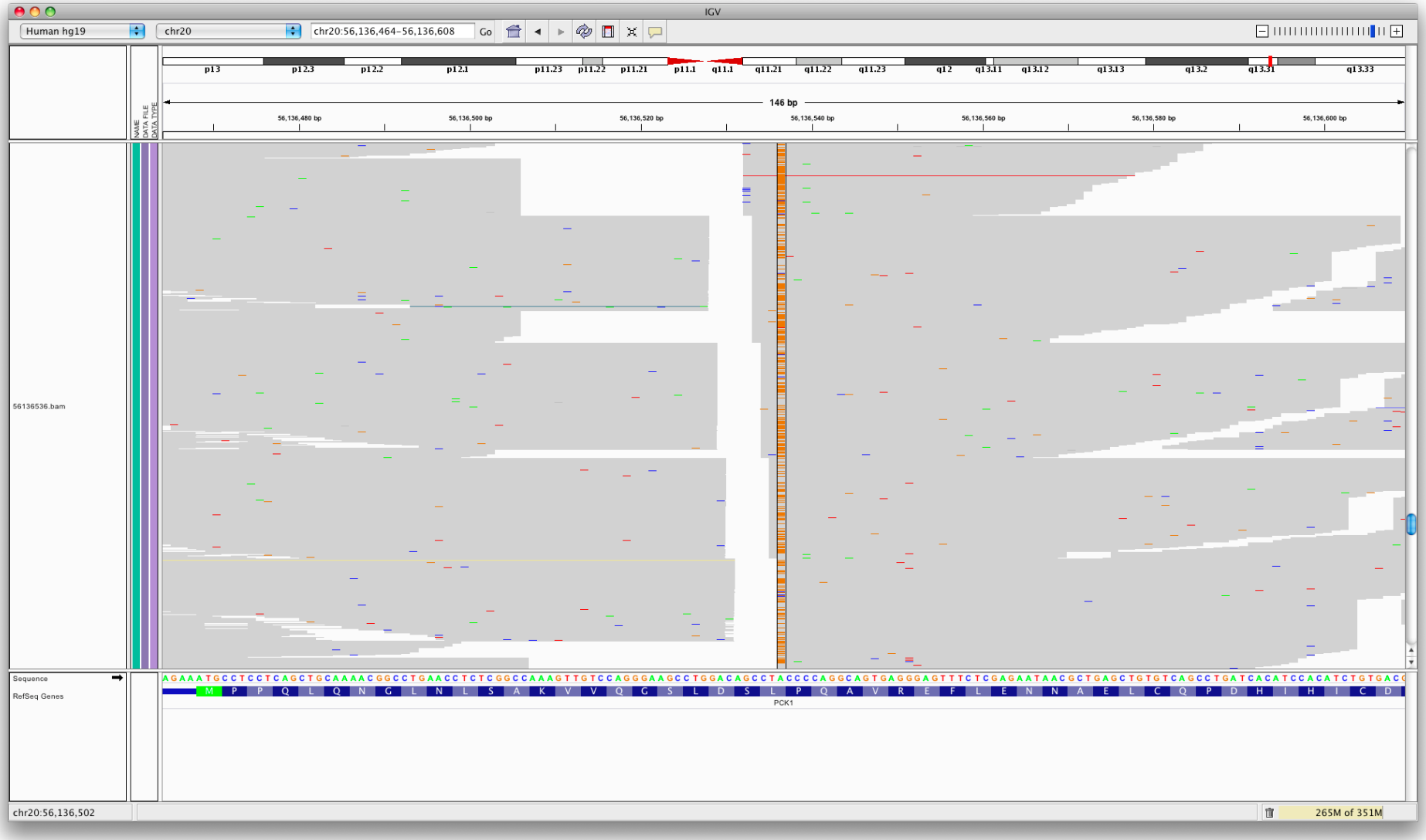
Read Position Rank Sum Test

- If the alternate bases are biased towards the beginning or end of the reads then the site is likely a mapping artifact





SNP removed by ReadPosRankSum



Other Statistics …

- And any other statistics anyone thinks of which can identify data processing or mapping artifacts in NGS data

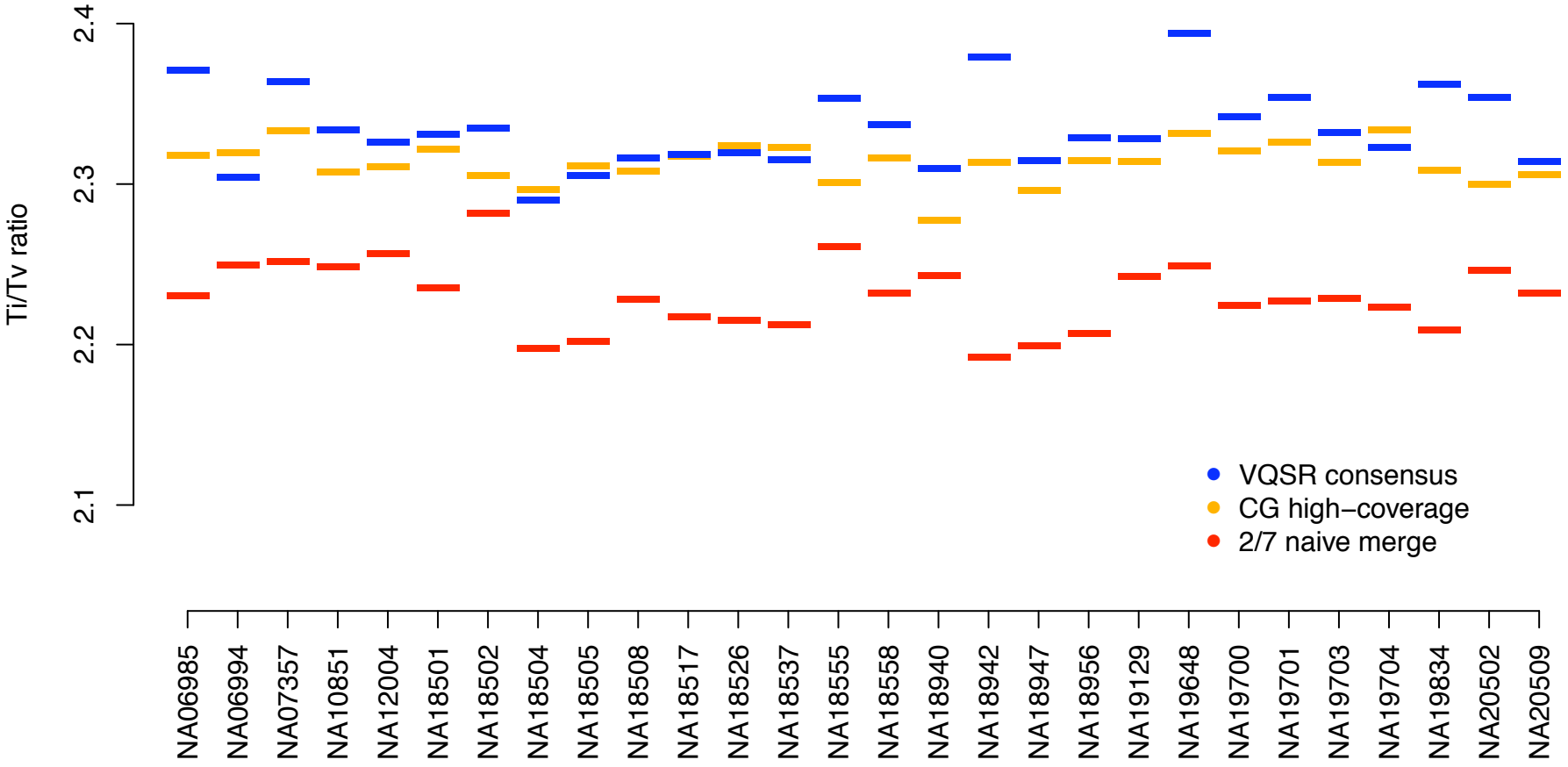
Outline

- Variants as points in a point cloud can be modeled using a Gaussian mixture model
- Details of individual component statistics
- Exposition of results
  - 1000 Genomes Project SNPs and indels
  - MPG Exome SNPs

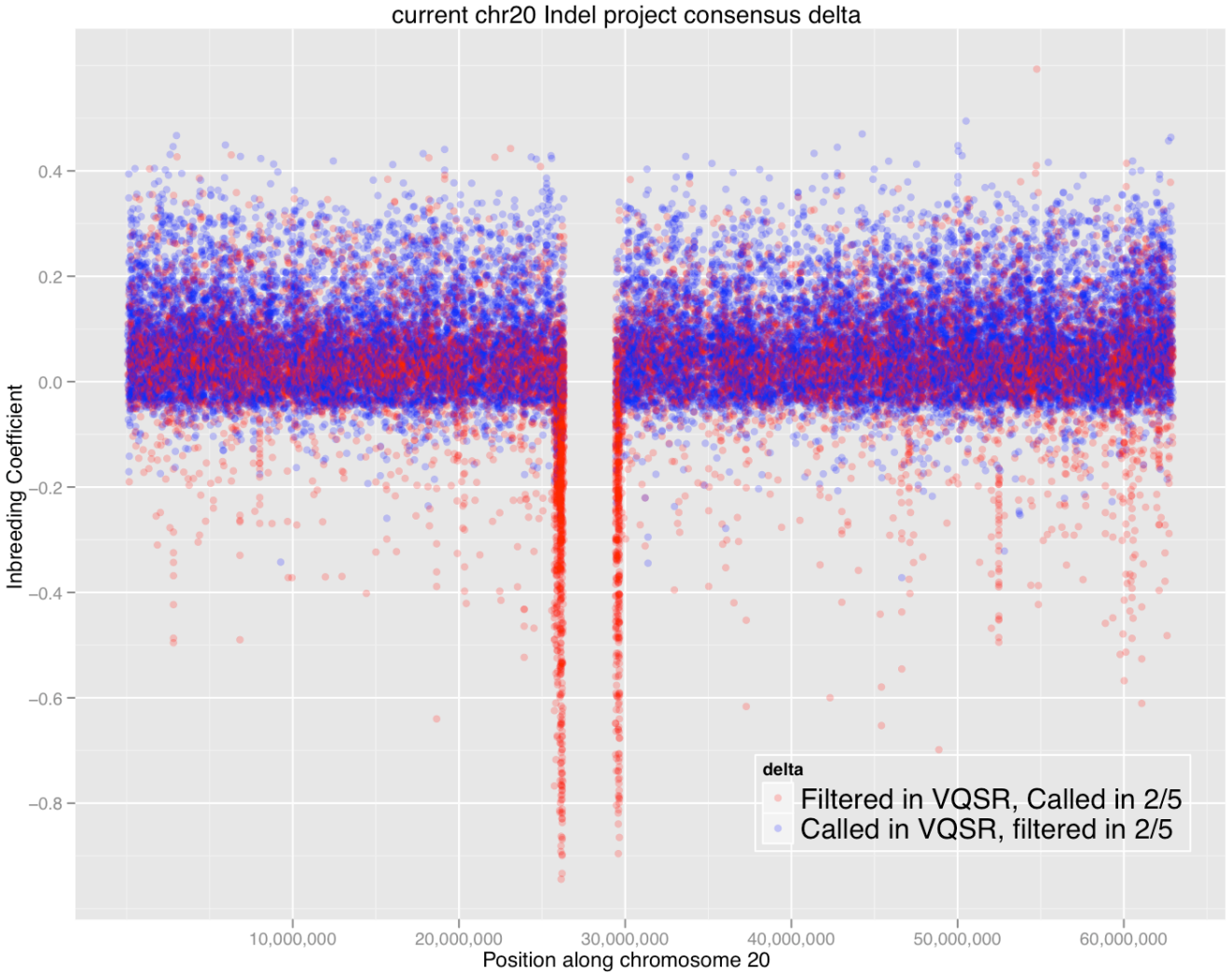
VQSR consensus outperforms previous merging strategy

Called In	Total # variants	dbSNP% (129)	# novels	Novel ti/tv	Omni poly sensitivity	Omni mono false discovery
Union	46.26M	19.39%	37.29M	1.998	98.94% 2.09M / 2.12M	16.31% 9,739 / 59,721
★ 2 of 7	39.11M	22.24%	30.41M	2.153	98.55% 2.09M / 2.12M	11.23% 6,707 / 59,721
3 of 7	35.69M	23.62%	27.26M	2.219	98.09% 2.08M / 2.12M	3.66% 2,184 / 59,721
4 of 7	32.55M	24.82%	24.48M	2.263	97.39% 2.06M / 2.12M	1.82% 1,085 / 59,721
5 of 7	28.45M	26.72%	20.85M	2.286	95.93% 2.03M / 2.12M	1.06% 634 / 59,721
Intersection	24.02M	27.57%	17.40M	2.317	89.23% 1.89M / 2.12M	0.76% 457 / 59,721
★ VQSR Project Consensus	38.88M	21.92%	30.36M	2.154	98.41% 2.08M / 2.12M	2.11% 1,261 / 59,721

Per-sample Ti/Tv is worse in 2/7 merge versus project consensus (suggesting that there are shared high-frequency errors)



VQSR-based Indel callset better than 1000G Project callset



Guillermo del Angel

VQSR Exomes - Data and Definitions

- T2D exome batch 005 with 100 EUR samples

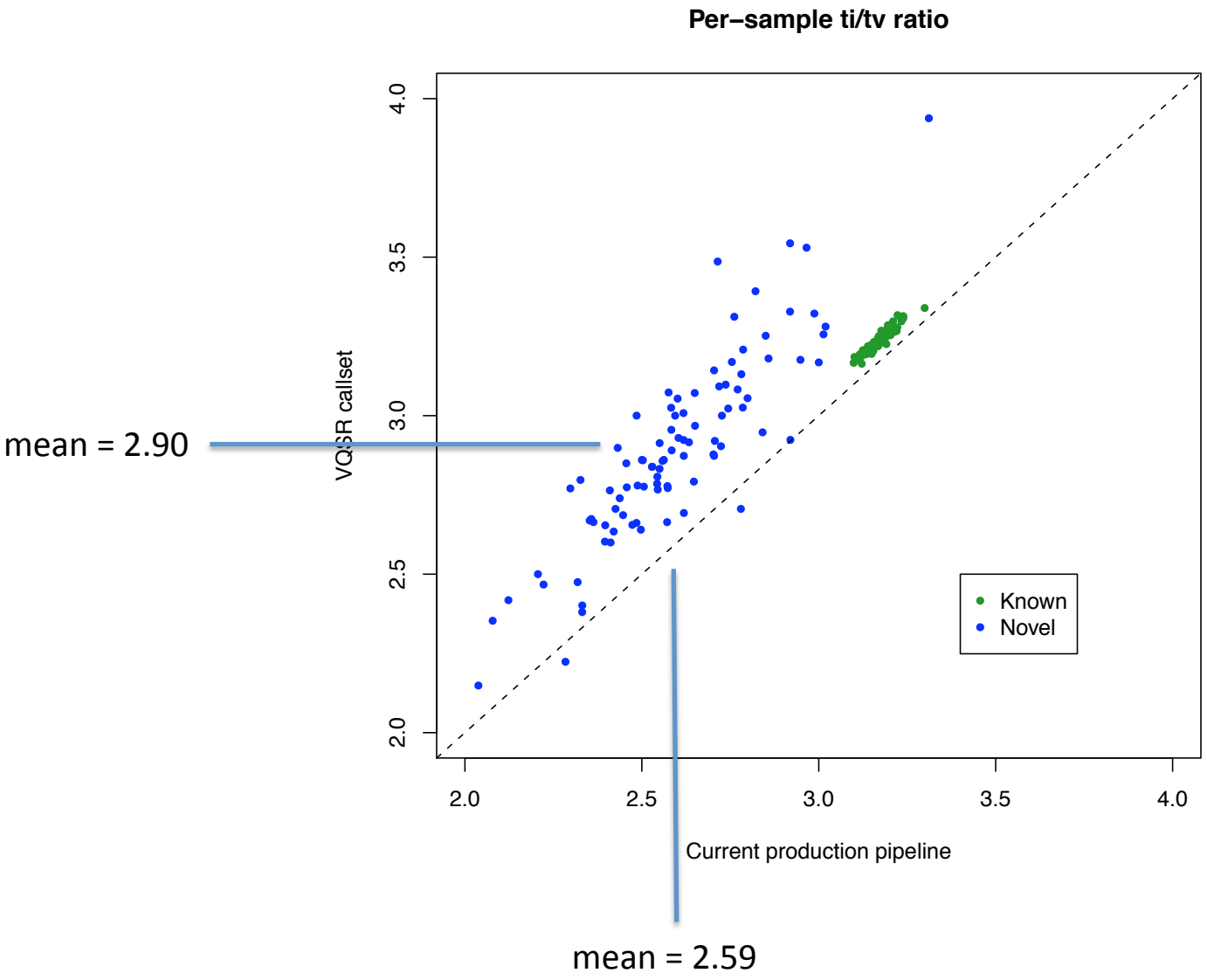
Current hand filtered pipeline

- SNPs at indel mask
- SnpCluster filter
- QD < 5.0
- SB >= 0.10
- Hrun >= 4

Example VQSR pipeline

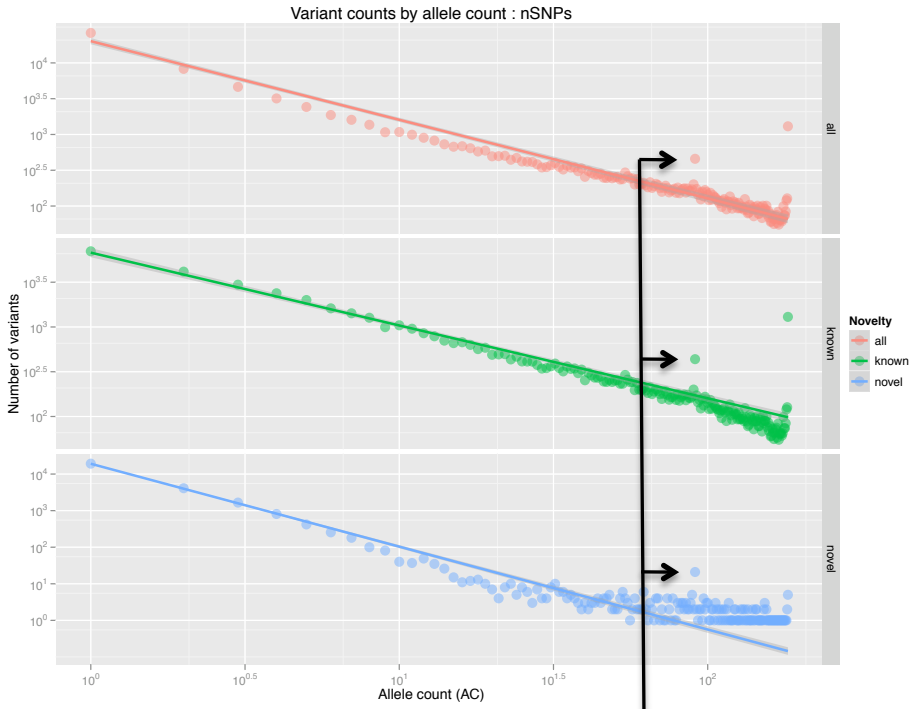
- QD
- HaplotypeScore
- MQRankSum
- ReadPosRankSum
- FS
- InbreedingCoeff
- MQ

Per-sample ti/tv is greatly improved

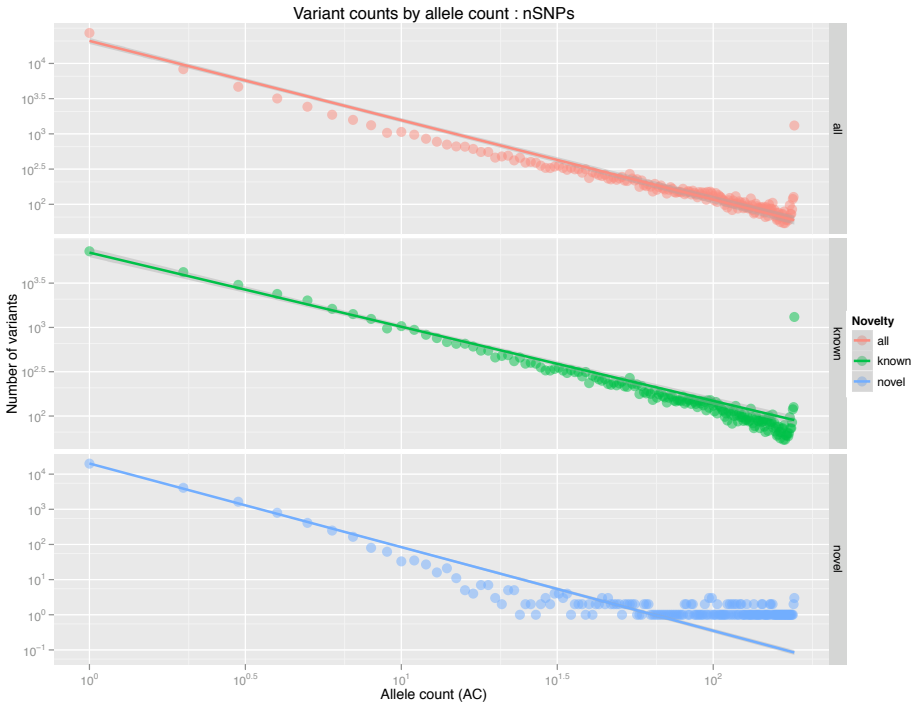


Problem with all-het calls is corrected

Current hand filtered pipeline

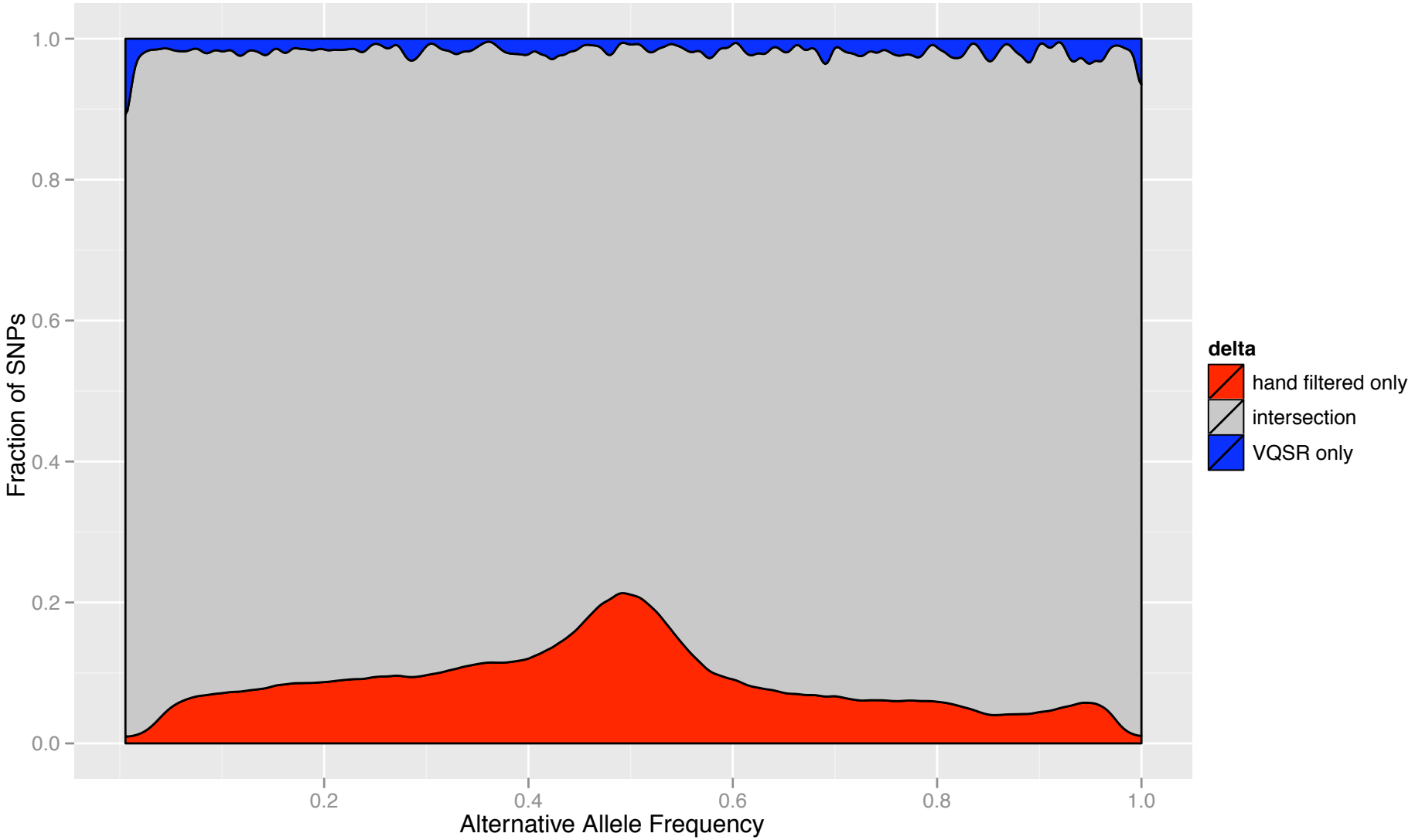


Proposed VQSR pipeline

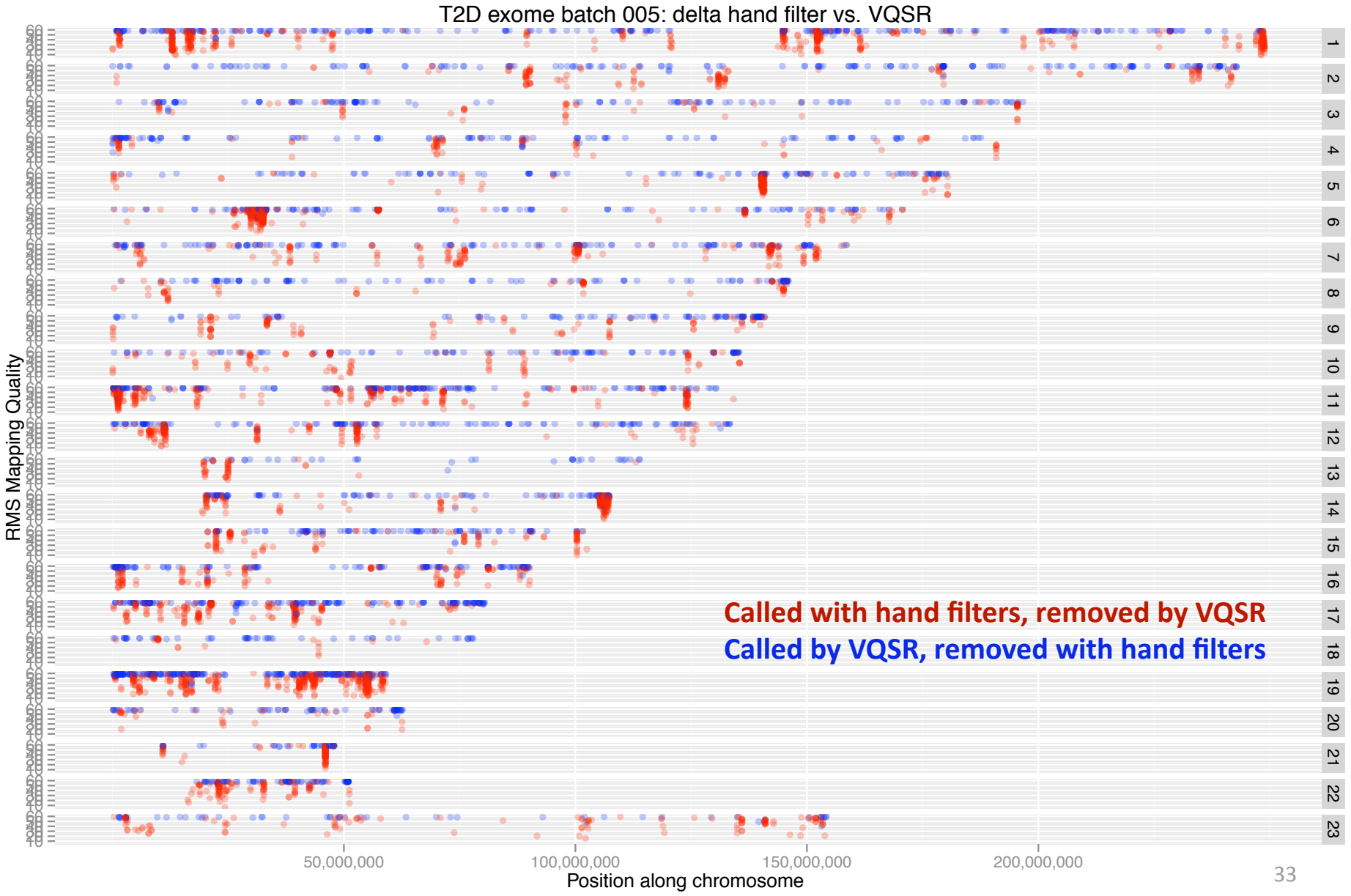


False all-het calls appear as an excess of AF=0.5 variants.
This phenomenon is corrected by the VQSR.

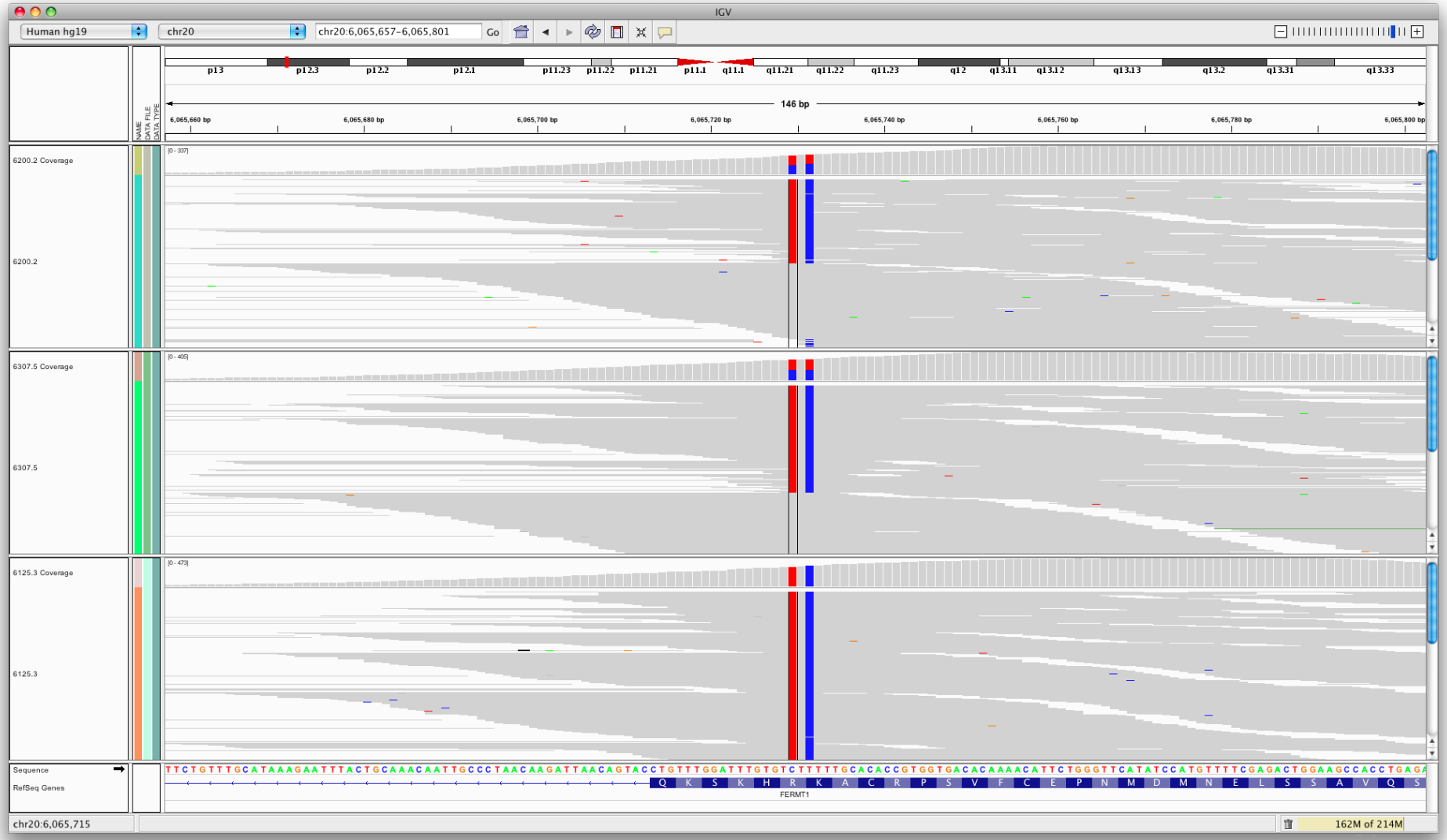
Problem with all-het calls is corrected



Removed SNPs are often near the centromere or telomere



Nonsynonymous MNP
Red SNP removed via HRun filter but kept by VQSR



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!

- Very happy to investigate any new ideas as additional statistics for the model

- 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