VARIANT FILTERING

The rationale for filtering

- QUAL is a basic measure of variant quality but it assumes that reads are correctly aligned and there are no systematic base call errors

- What causes errors in variant calling?
  - sequencing errors >> should be accounted for by base quality + recalibration + marking of duplicates
  - Incorrect alignment >> **Re-alignment step should have reduced this problem but not eliminated it**

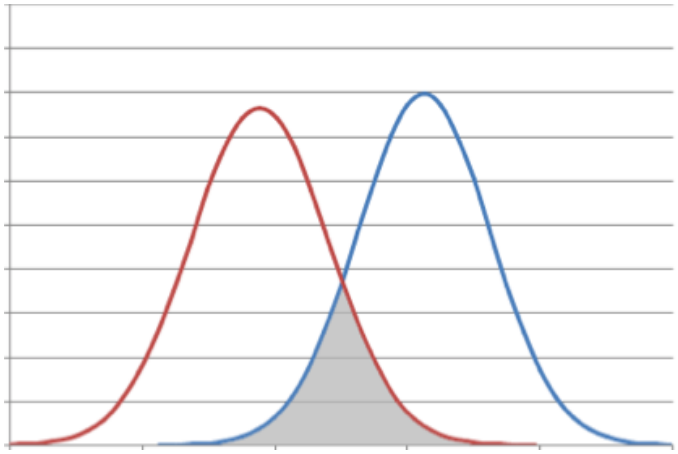
- **Thus although QUAL (which depends on Mapping Quality of reads and Base qualities) is a useful measure, there will still be FP with high QUAL.**
  - **and we wish to reduce the number of FP**

- Tell tale signs of suspicious variants
  - poorly mapped reads (ambiguity)
    - MQ: Root Mean Square of MAPQ of all reads at locus
    - MQ0: Number of MAPQ 0 reads at locus
  - biased support for the **REF** and **ALT** alleles
    - MQRankSum: Mapping quality rank sum test
    - ReadPosRankSum: Read position rank sum test
    - Strand bias and FS

INFO fields – important for filtering

- **QD:** variant quality score over depth of variant allele
  - Confidence in the site being variant should increase with increasing depth
- **MQ:** RMS MAPQ of all reads at locus
  - Regions of excessively low mapping quality are ambiguously mapped and variants called within are suspicious
- **MQ0:** number of MAPQ 0 reads at locus
- **MQRankSum:** 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
- **Haplotype score:** Probability that the reads in a window around the variant can be explained by at most two haplotypes
- **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
- **ReadPosRankSum:** 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

Details of MQRankSum



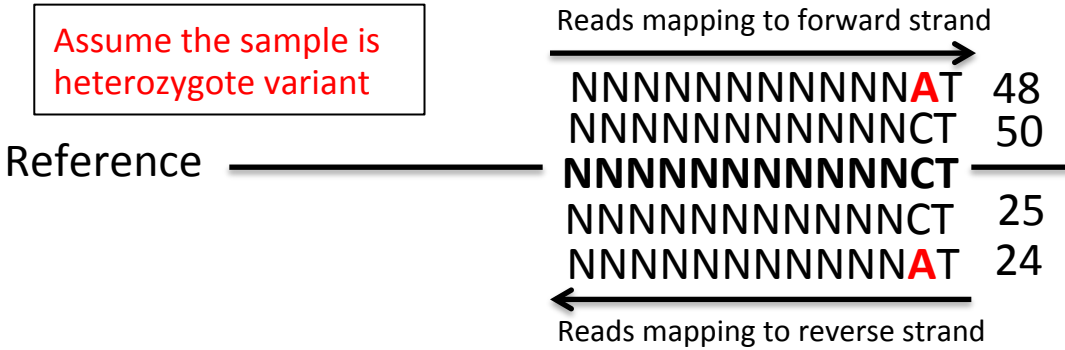
Rank sum test are used to test if two samples come from the same population when the distribution is unknown.

- We test the null hypothesis that the two samples of MQ are the same
- If we reject this null hypothesis, the variant site is likely to be a False Positive

Strand bias - optional

Strand bias **is NOT** about more reads mapping to one of the strands than the other

Assume the sample is heterozygote variant



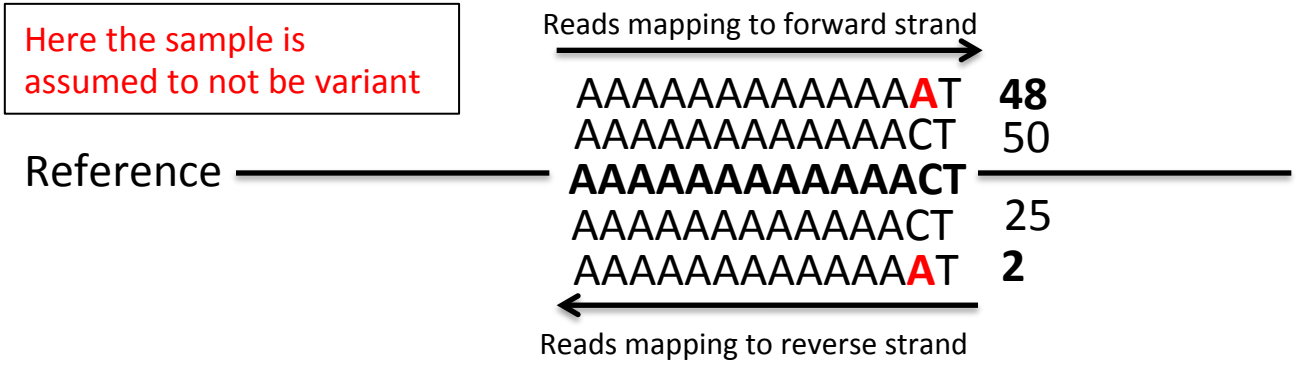
		Fw	Rev
Ref	C	50	25
Alt	A	48	24

Clearly more reads mapping to FW than Rev

But, Fw/Rev ratio is same for Ref allele and Alt: 50/25 = 48/24

This **IS** strand bias

Here the sample is assumed to not be variant



		Fw	Rev
Ref	C	50	25
Alt	A	48	2

50/25 != 48/2

Hard vs. soft filtering

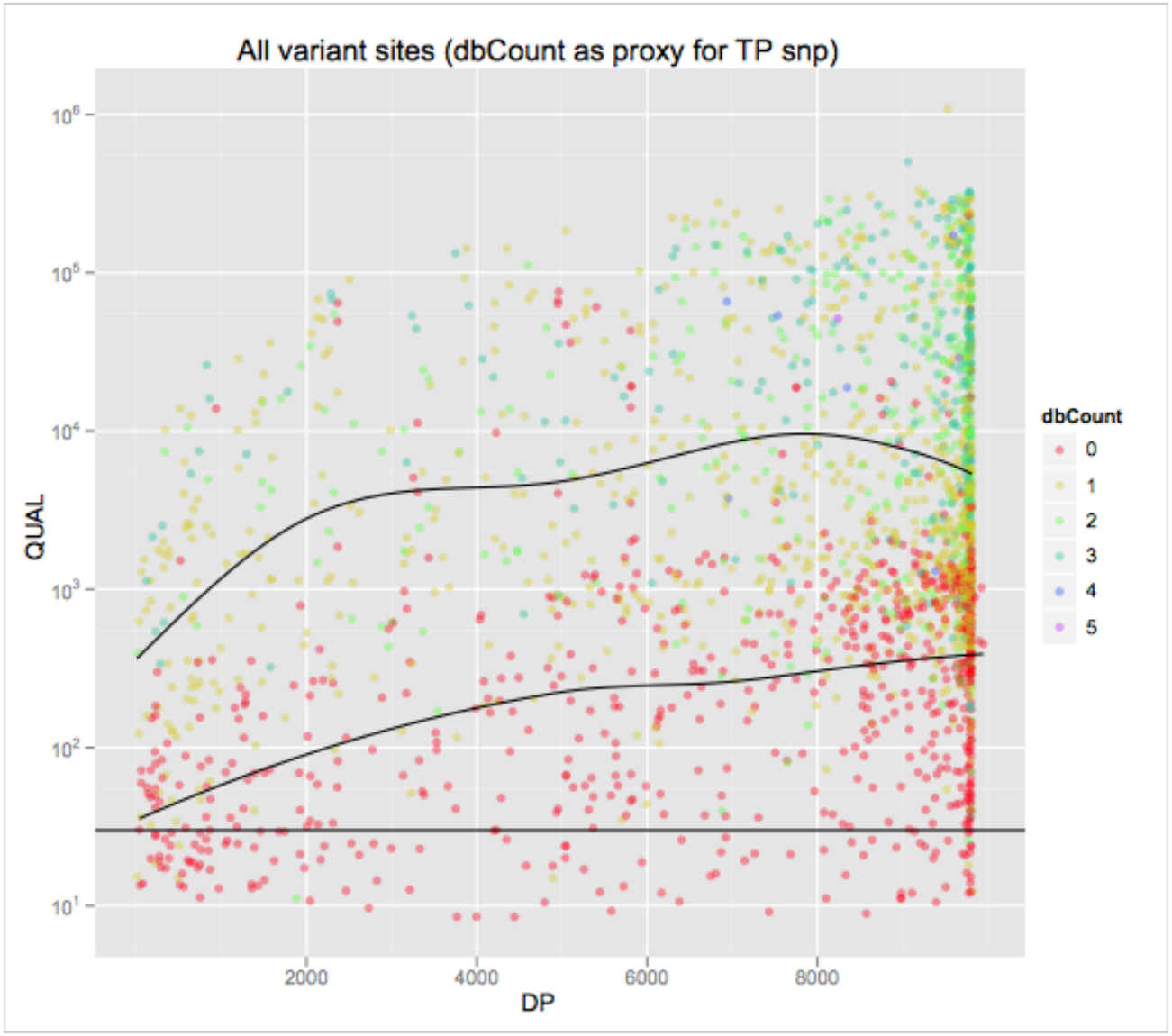
- Can set thresholds for these INFO fields and request that all thresholds are passed for a variant to be considered valid

- Which fields to you use and where do you set the thresholds?
  - use datasets of known SNPs and compare their INFO fields to those likely FP variants
  - fields that provide a good separation can be used as filters

- Disadvantage of **hard filtering**
  - works with hard cut-offs

- Variant Quality Score Recalibration (GATK) or **soft filtering**
  - **use machine learning to learn the features of true variants and distinguish them from false positives**

QUAL provides OK but not great separation



Red: likely false positive SNP

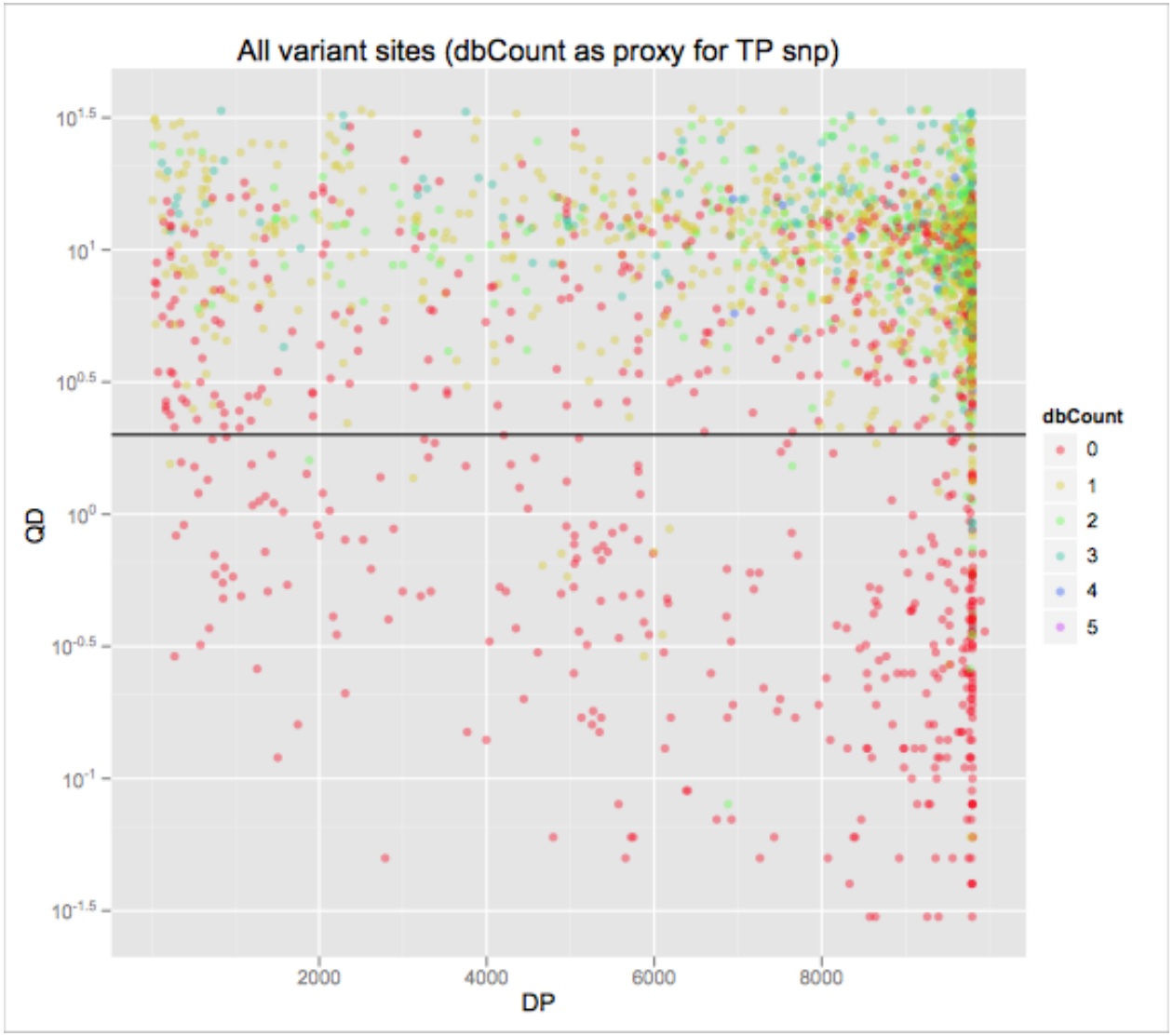
Other colour: likely true positive SNP

Note how:

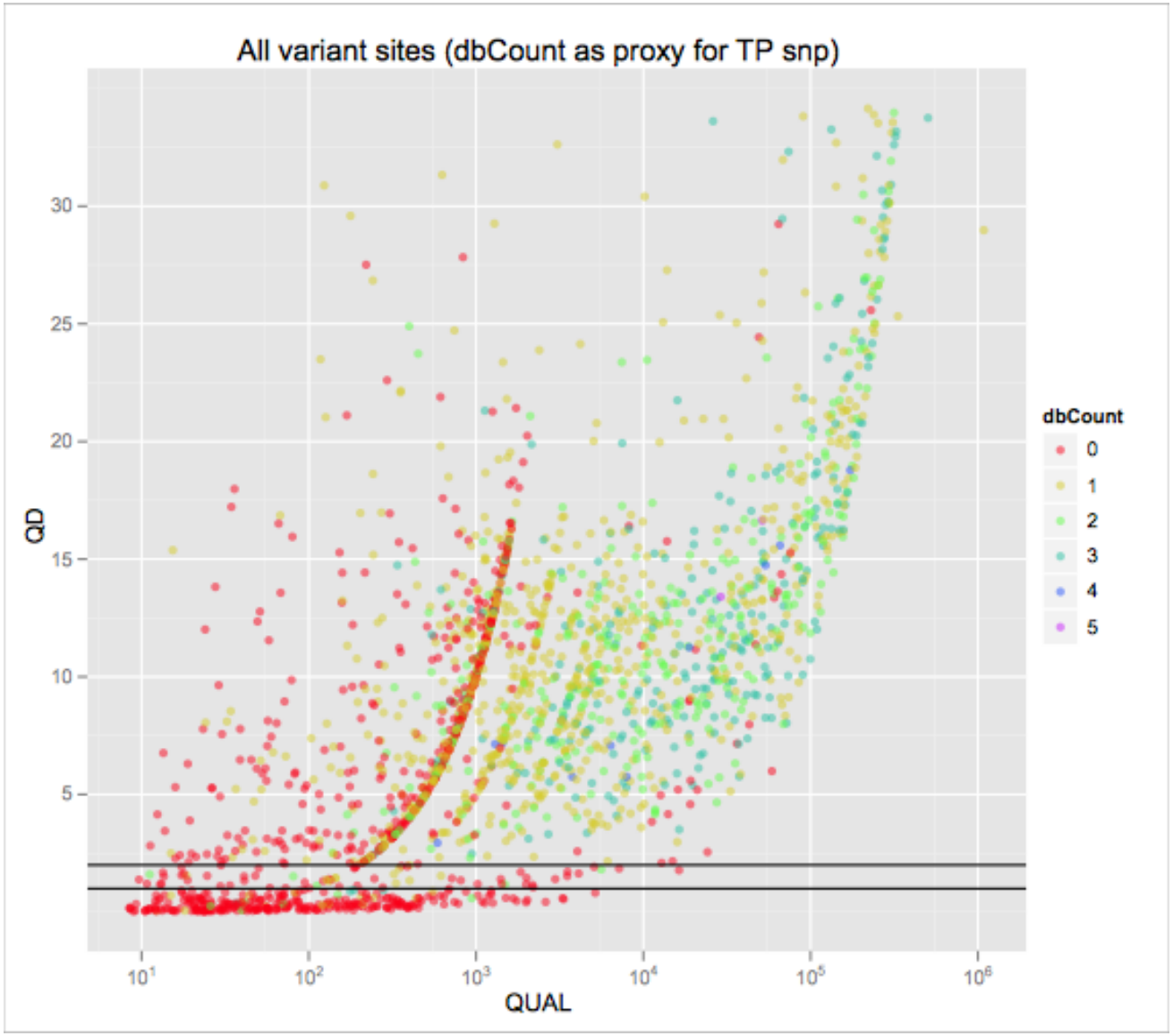
- QUAL increases with depth
- QUAL for known SNPs (green) is higher than for unknown SNPs (red) at any given depth

Note: in these plots an exremely high depth of sequencing was used

QD vs depth: much better separation



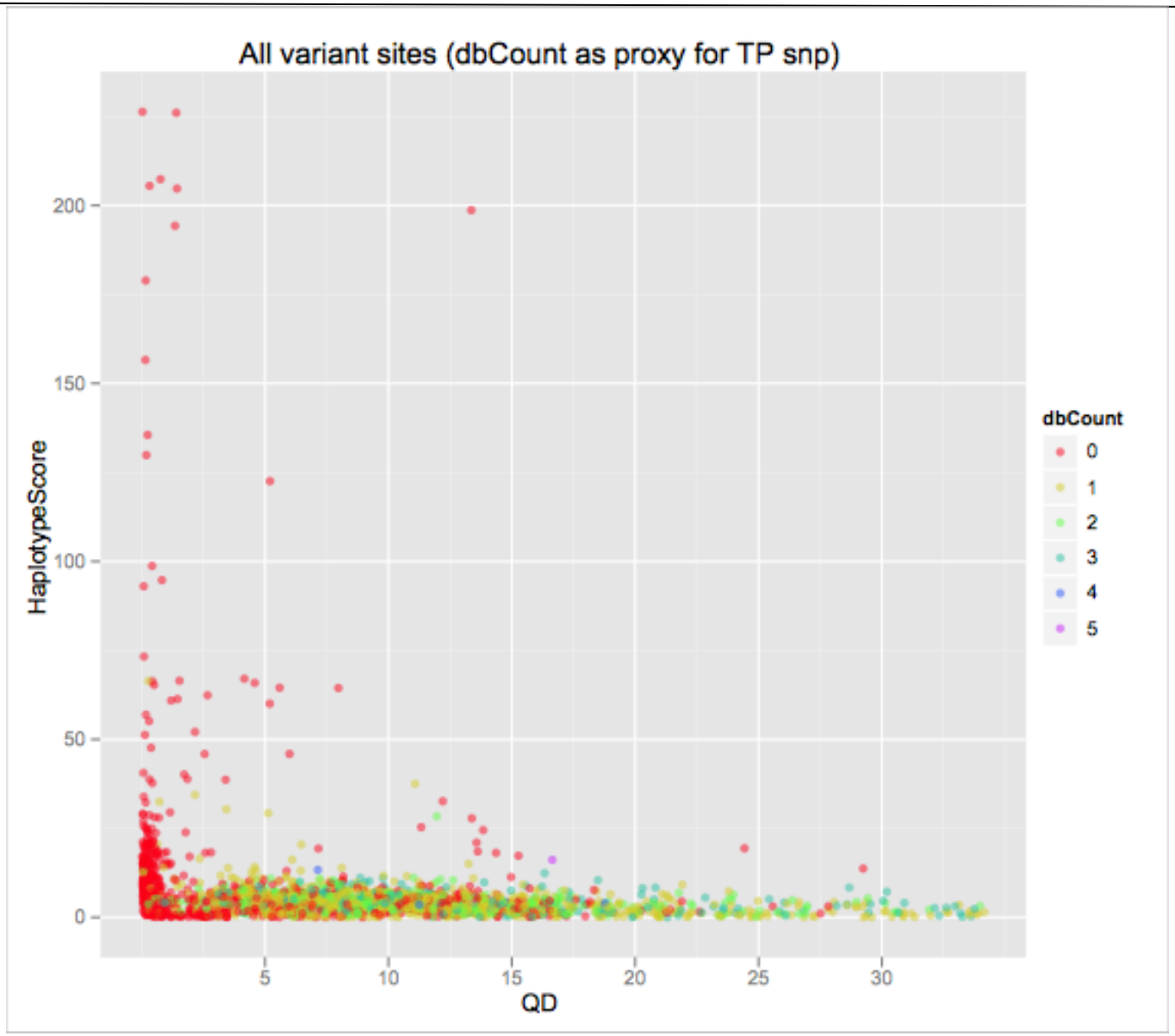
Qual vs QD



Illustrates advantages of soft filtering:

- learning features of true SNPs
- no fixed level cutoffs

Haplotype Score vs QD: advantage of combining



No known SNPs even with QD >3 have high HaplotypeScore, so can remove more likely FP SNPs by filtering on HaplotypeScore

Can require SNPs to have HaplotypeScore less than 20.

practical 03_advancedPipeline.bash

- Now we can try out all the refinements we have discussed

- Please make an effort to progress execute **rapidly** and **without error**
  - **pause to understand new commands (realignment, filtering)**

practical 031_generatingReports.bash

- Let us see if all those refinements made any difference.

- Talk the students through the improvements in IGV