# NCBI dbGaP

## Genotype Quality Analysis

Genome Variation Working Group

# Processing genotypes

Applying software provided by Goncalo Abecasis for FNIH GAIN

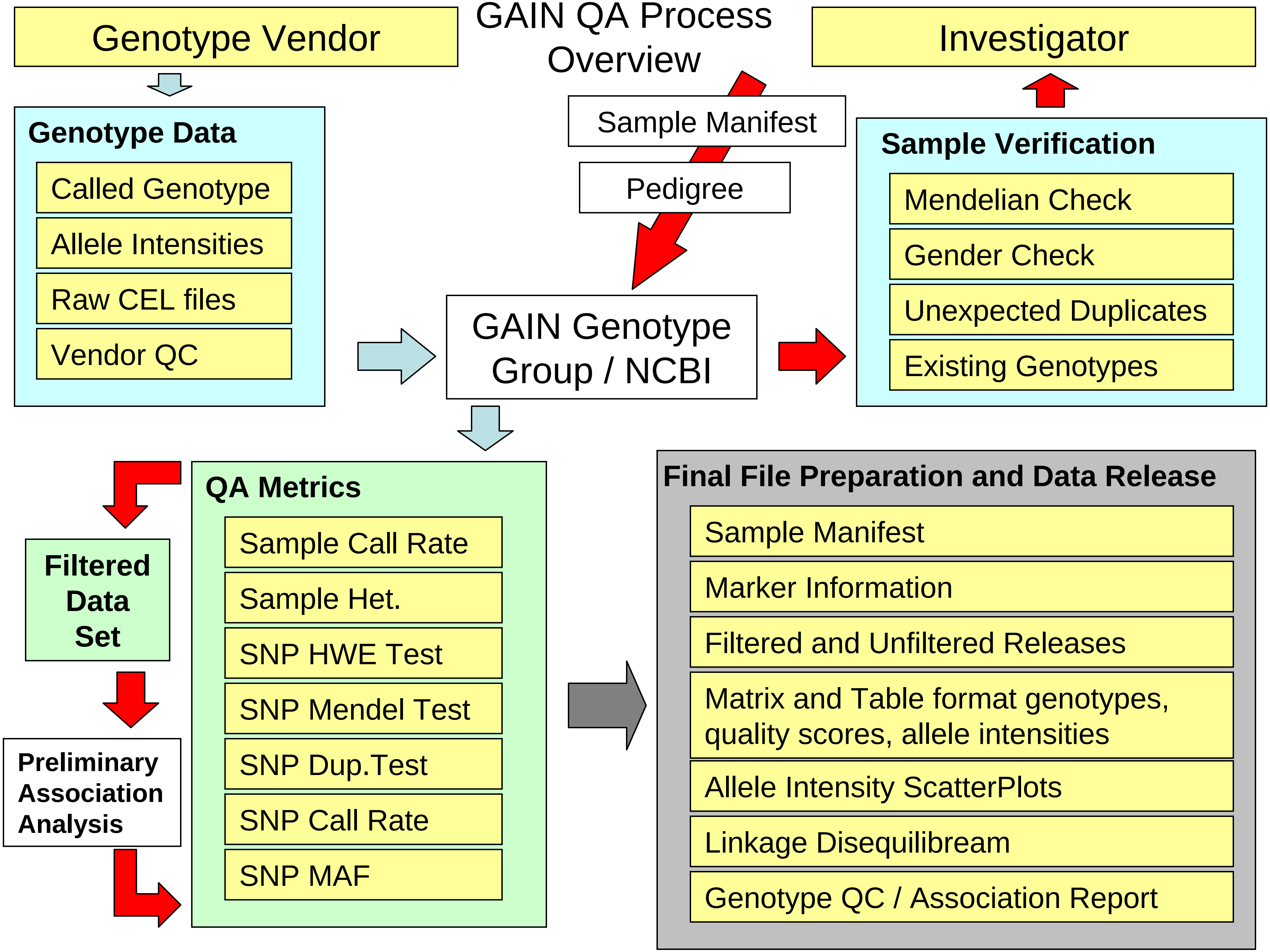
1) **Verify Transferred Dataset**

- Verify counts of individuals, duplicate, failed samples, consent groups
- Verify all components of dataset: raw data (CEL files), normalized intensity, genotypes, quality scores, marker information

2) **Sample Quality Metrics:**

- Mendelian Error check in families
- Gender agreement with manifest
- Identification of unexpected duplicate samples
- Call rate per sample
- Average Heterozygosity per sample
- Verify with existing genotypes if available

# GAIN QA Process Overview



Genotype Vendor

**Genotype Data**

- Called Genotype
- Allele Intensities
- Raw CEL files
- Vendor QC

Sample Manifest

Pedigree

GAIN Genotype Group / NCBI

Investigator

**Sample Verification**

- Mendelian Check
- Gender Check
- Unexpected Duplicates
- Existing Genotypes

**Filtered Data Set**

**Preliminary Association Analysis**

**QA Metrics**

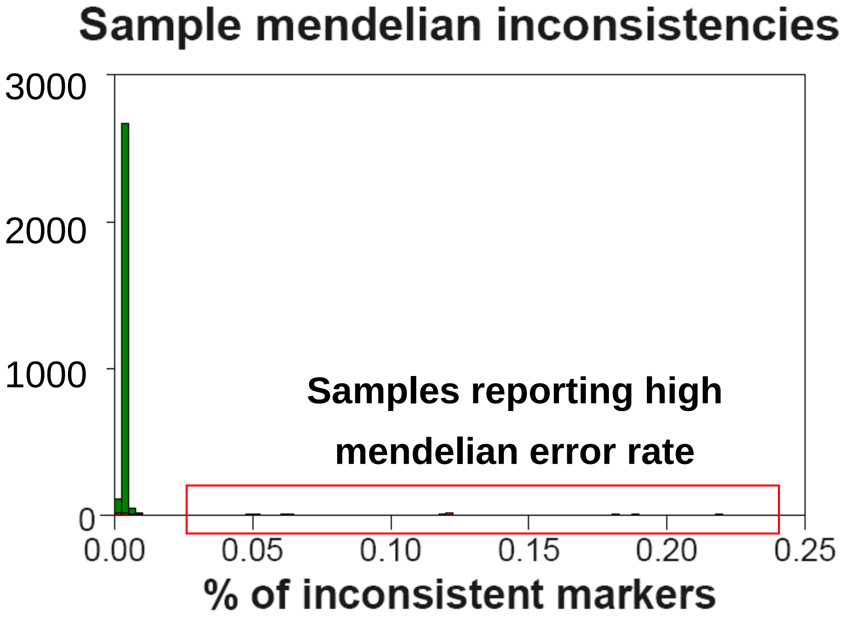
- Sample Call Rate
- Sample Het.
- SNP HWE Test
- SNP Mendel Test
- SNP Dup.Test
- SNP Call Rate
- SNP MAF

**Final File Preparation and Data Release**

- Sample Manifest
- Marker Information
- Filtered and Unfiltered Releases
- Matrix and Table format genotypes, quality scores, allele intensities
- Allele Intensity ScatterPlots
- Linkage Disequilibream
- Genotype QC / Association Report

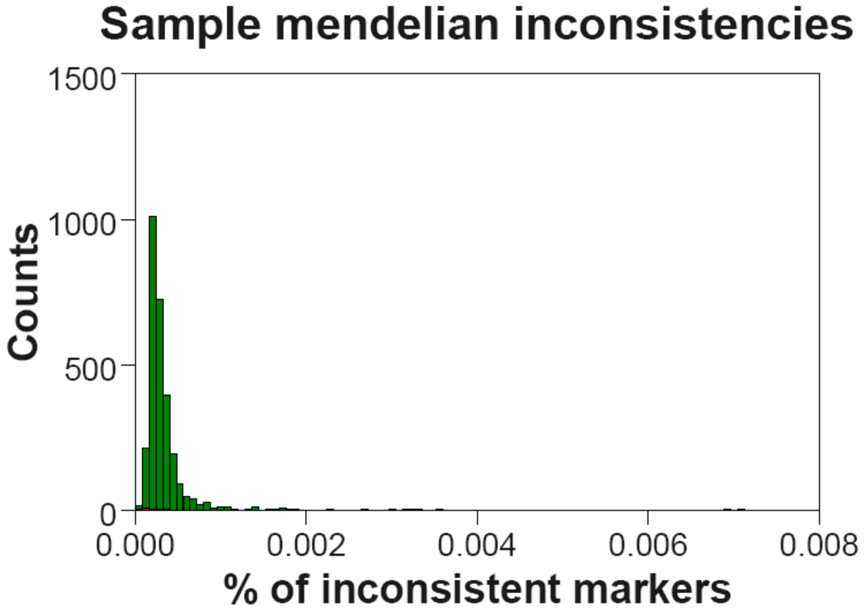
# Mendelian Errors in Trios per Sample

*Prior to Sample QC*

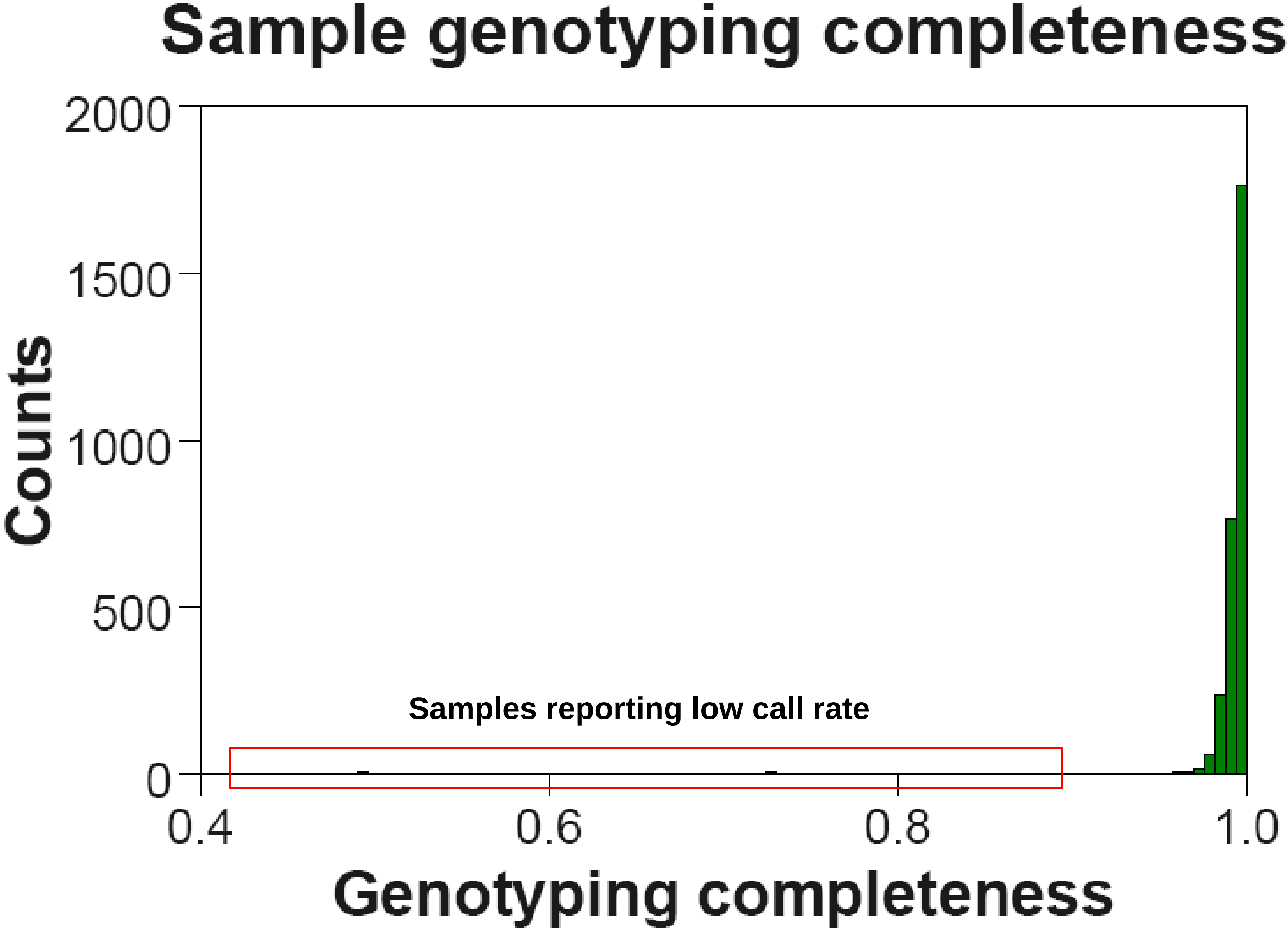


Samples reporting high mendelian error rate

*Following Sample QC*

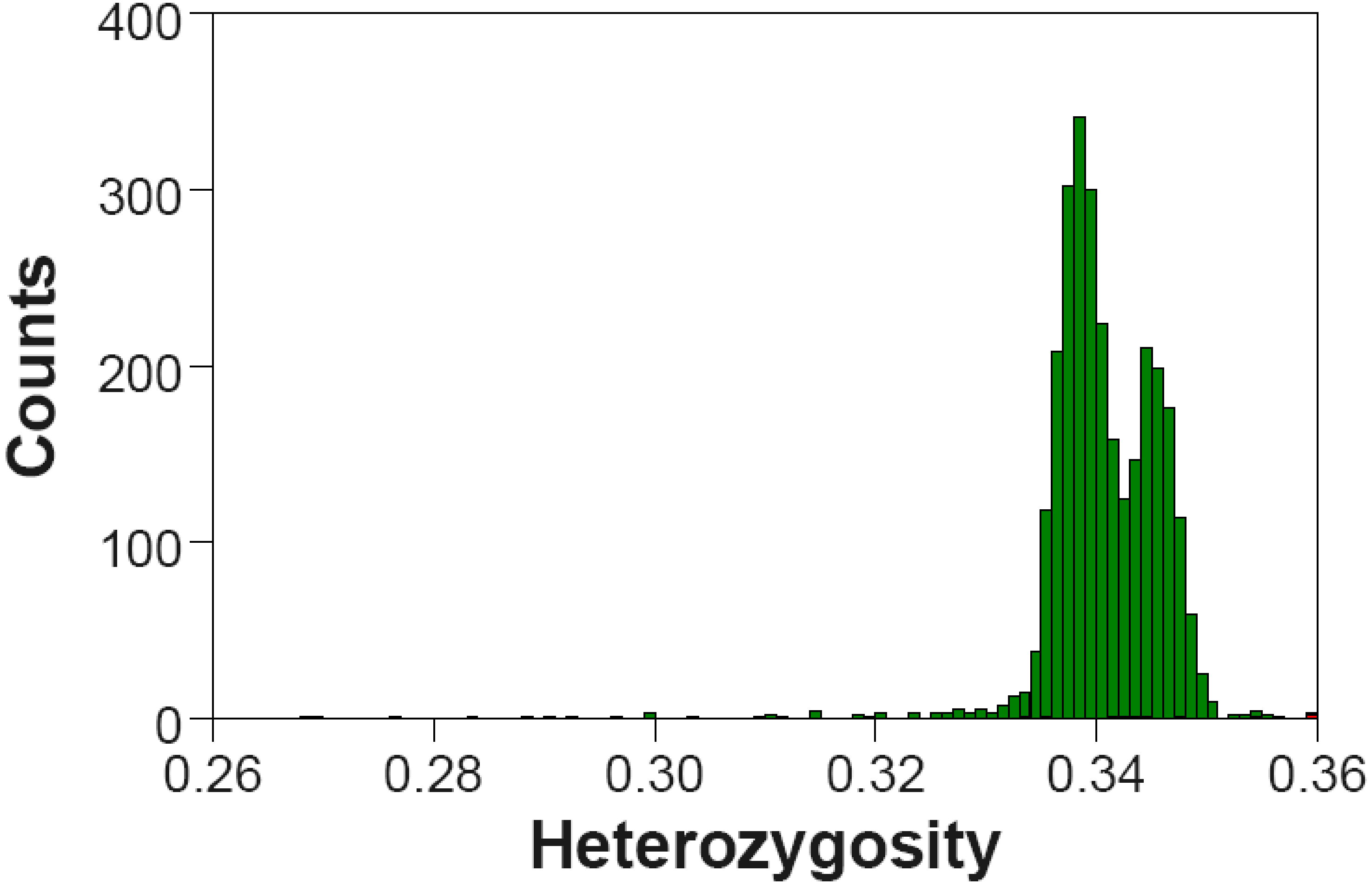


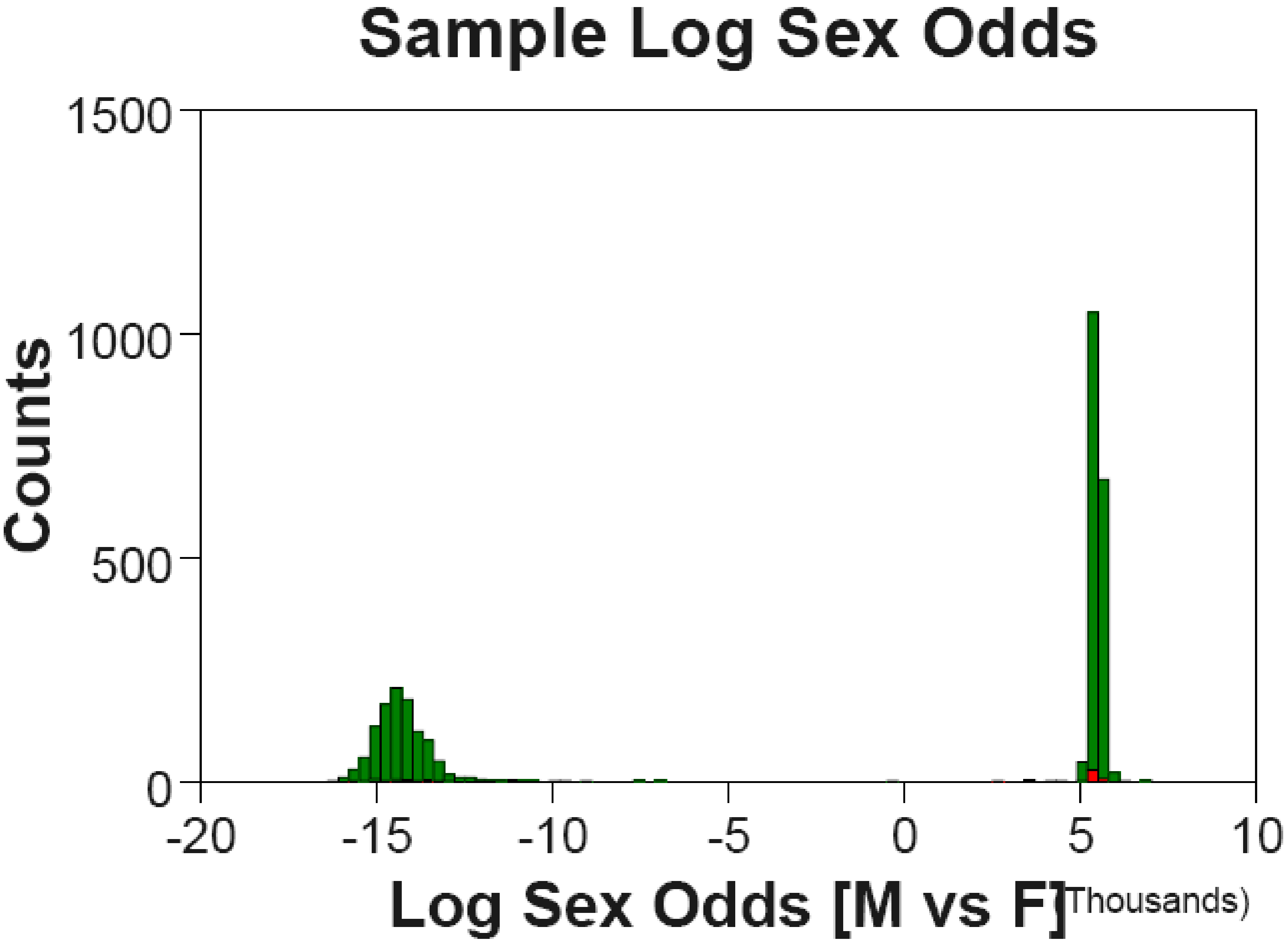
*Note difference in X-axis scale above



Samples reporting low call rate

# Sample heterozygosity





Sample Log Sex Odds
Counts
1500
1000
500
0
-20
-15
-10
-5
0
5
10
Log Sex Odds [M vs F]
Thousands)

# Processing Genotypes

3) **SNP Quality Metrics**

- **Tolerances to be reviewed and set for each study:**

  Mendelian error rate per marker

  HWE test, by population

  Call Rate per marker

  Duplicate Error Rate per marker

  Plate/Batch effect test

  Concordance with HapMap for control HapMap samples

- **Above tolerances define constraints for “filtered subset”**

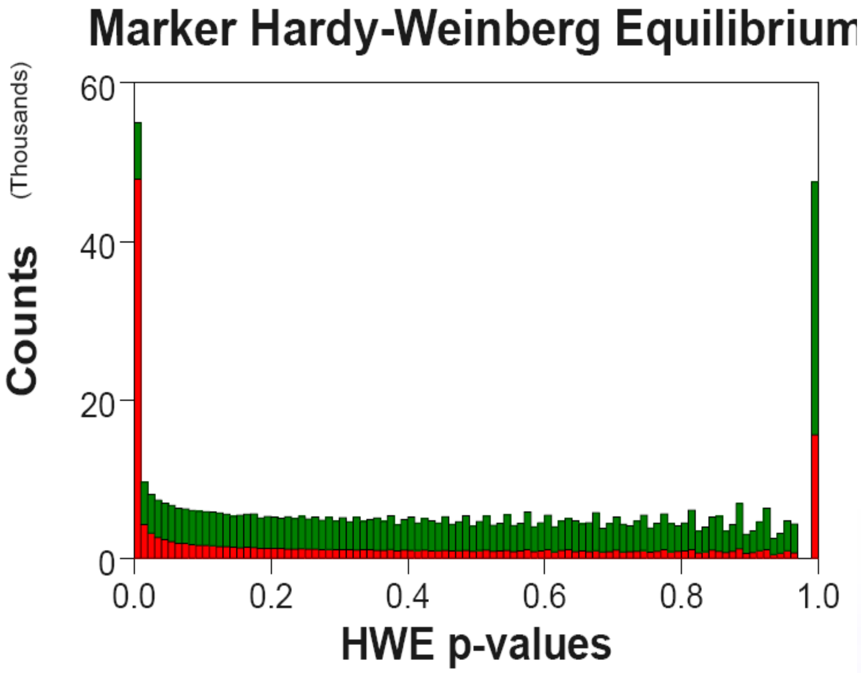
  Set a genotype quality score threshold for accepting a call

  Set a minimum minor allele frequency for reliable genotype calls

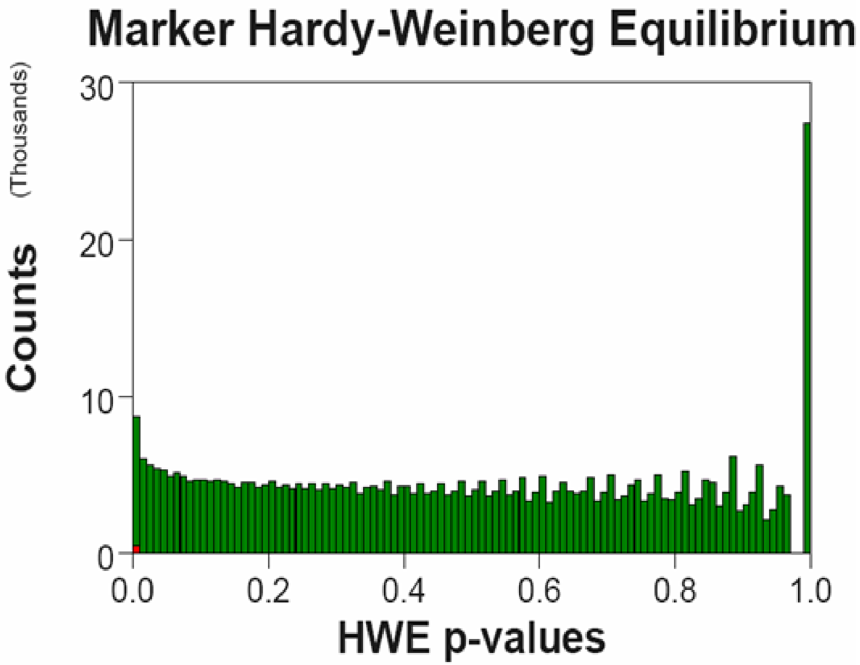
  Conduct preliminary association test to review top hits for potential quality issues that might be filtered out by adjusting QC thresholds

# HWE test : pvalue < 0.000001 threshold used

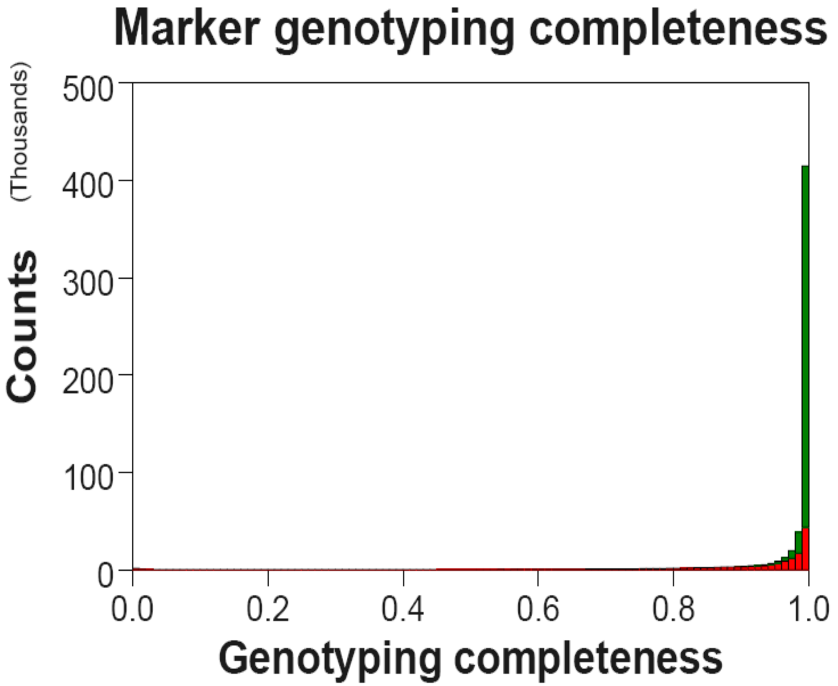
*Prior to SNP QC*

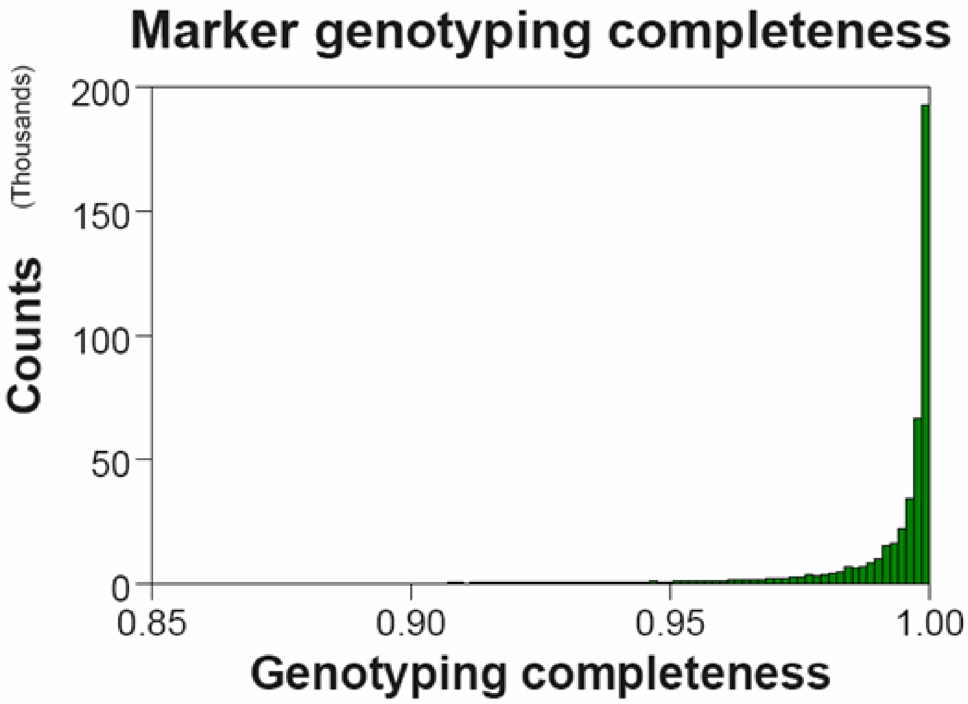


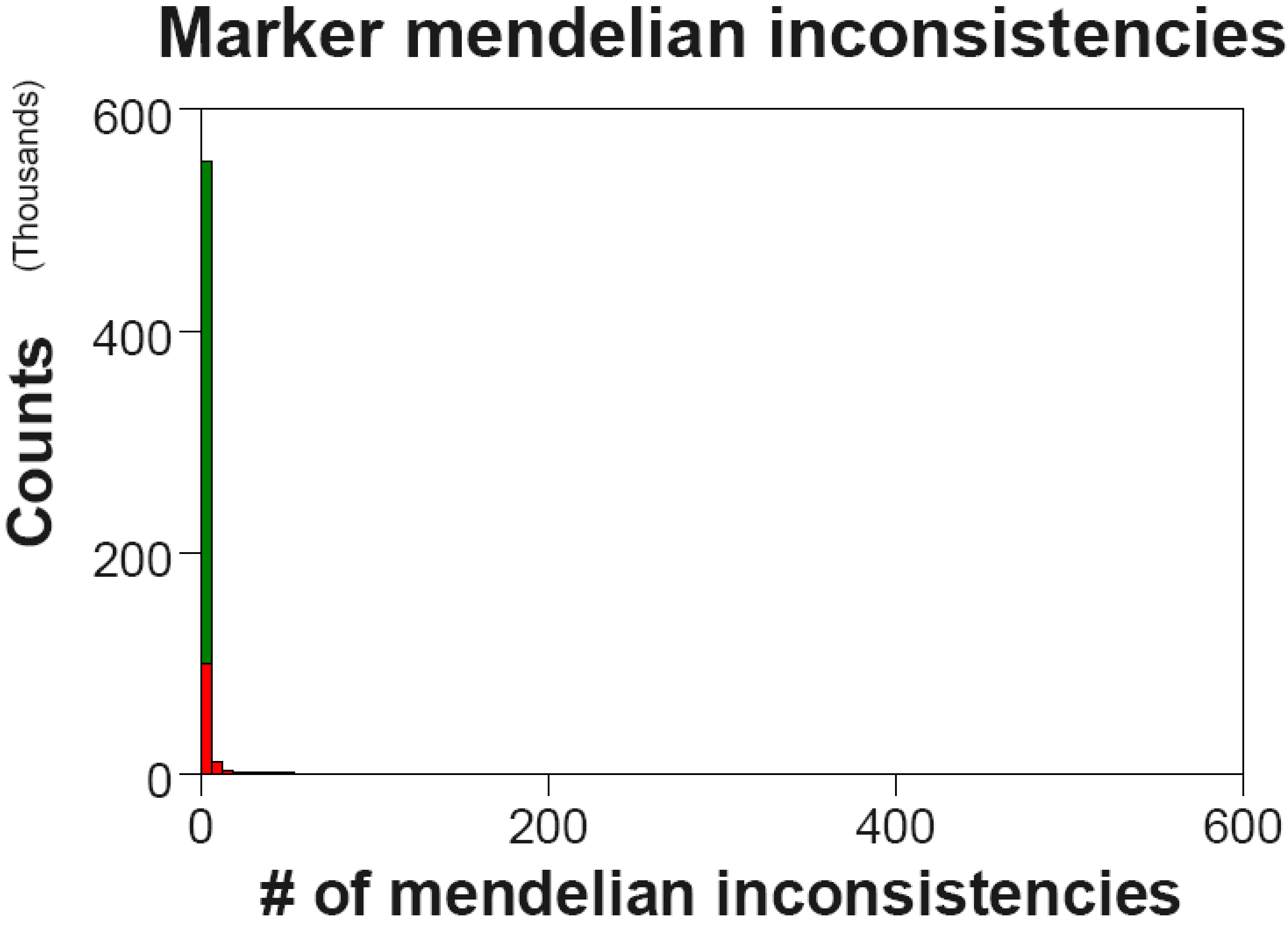
*Filtered SNP set*

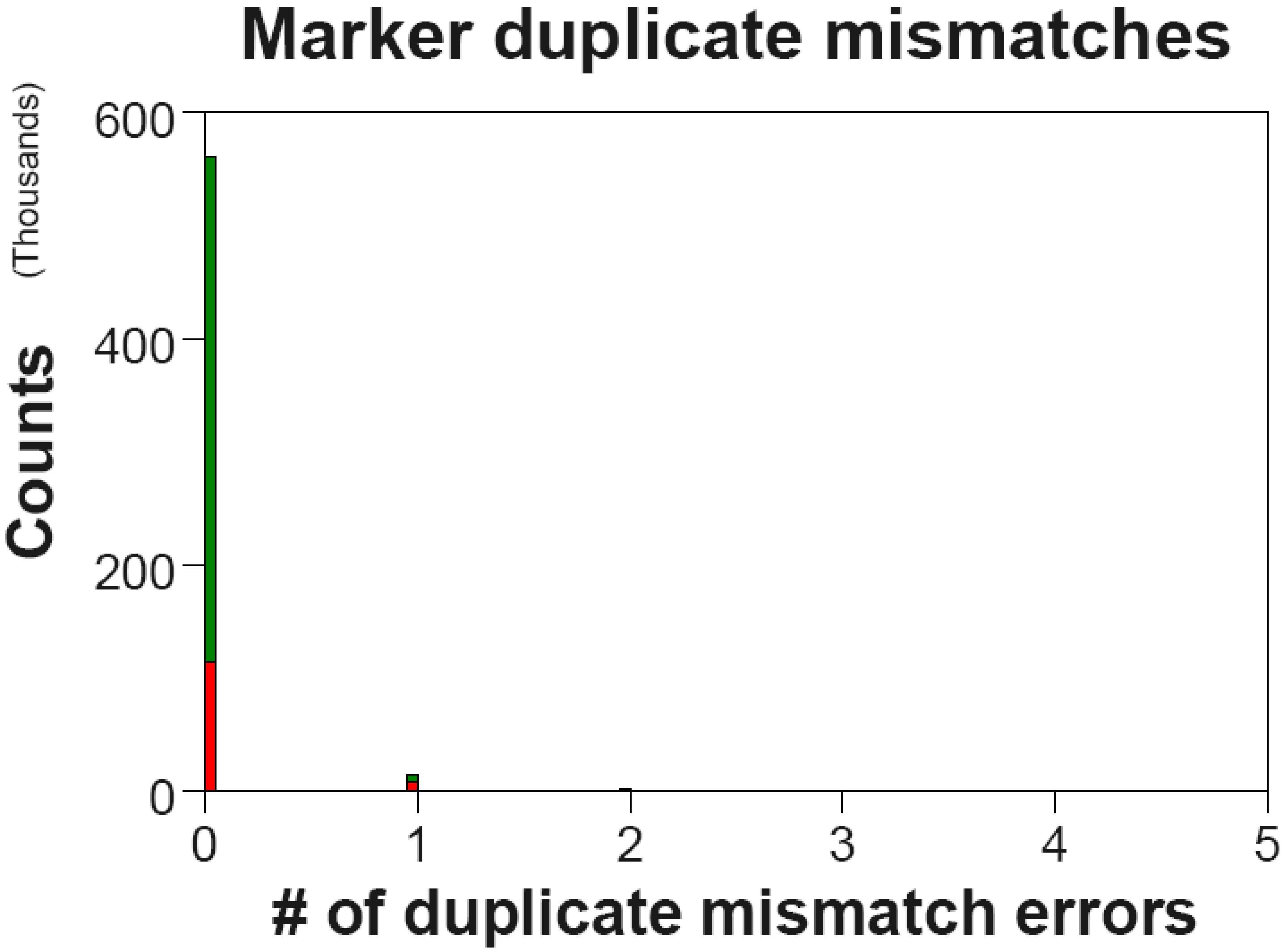


# Genotyping Call Rate



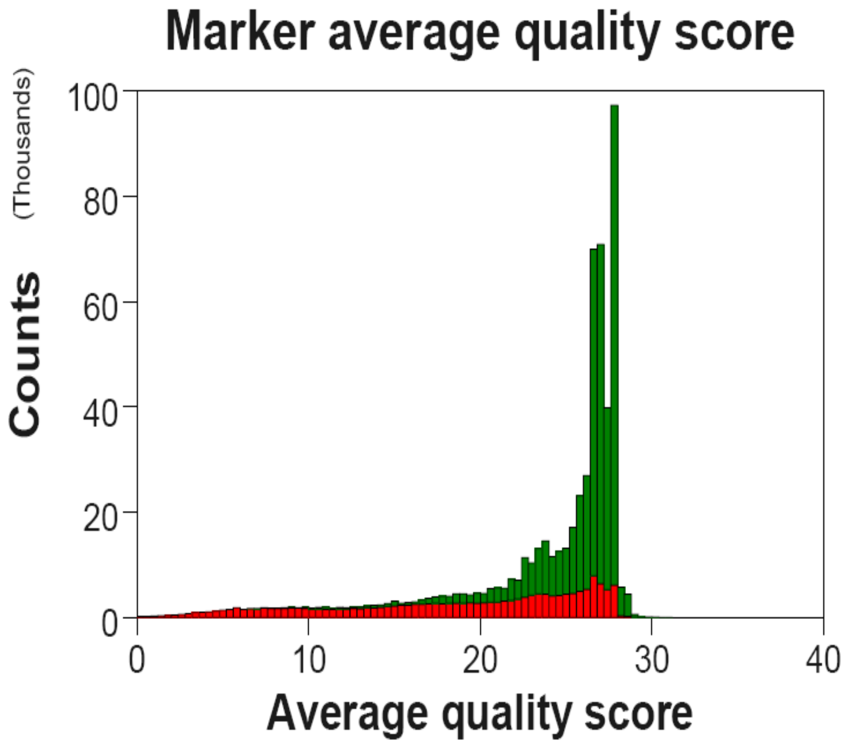



Marker mendelian inconsistencies
Counts
(Thousands)
600
400
200
0
0
200
400
600
# of mendelian inconsistencies

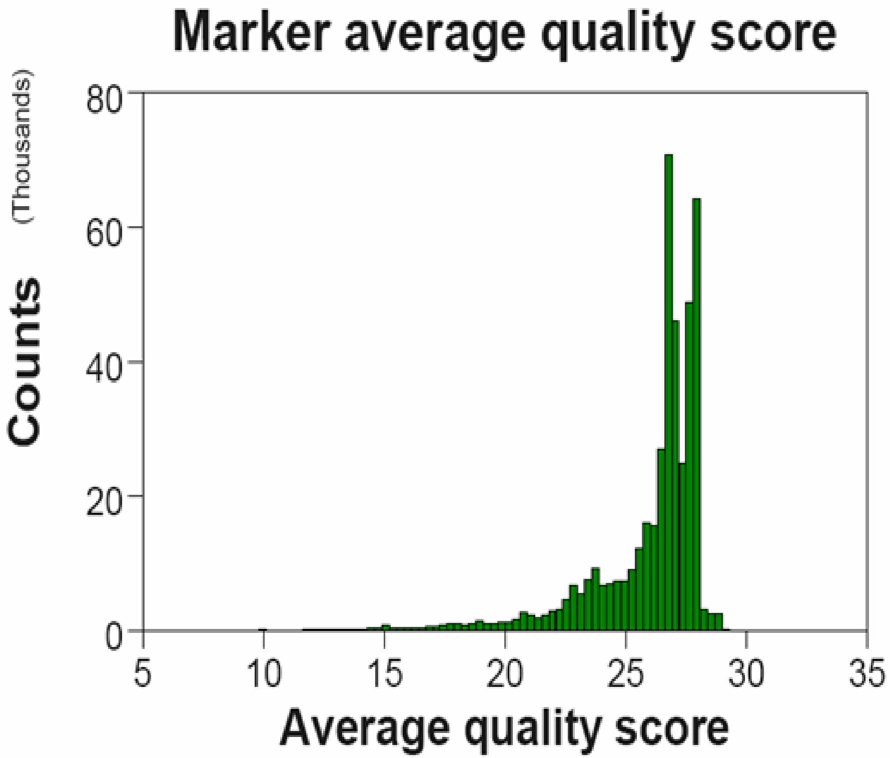

Marker duplicate mismatches
Counts
(Thousands)
600
400
200
0
0
1
2
3
4
5
# of duplicate mismatch errors

# Filtered set of SNPs based on QC metrics eliminates SNPs with low average genotype quality scores

*Prior to SNP QC*

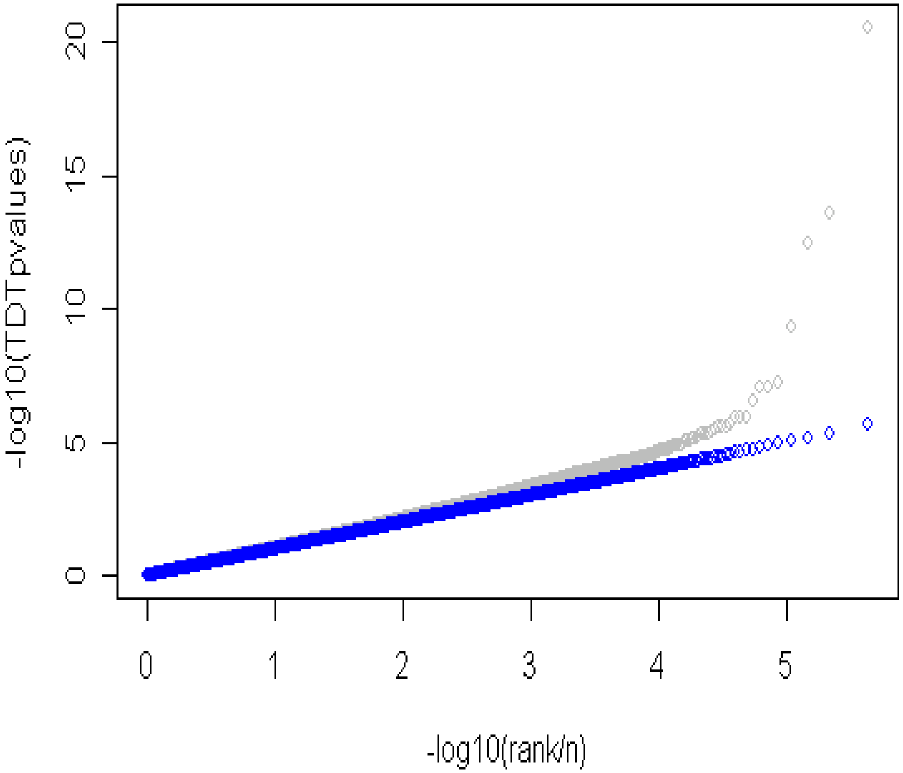


*Filtered SNP set*

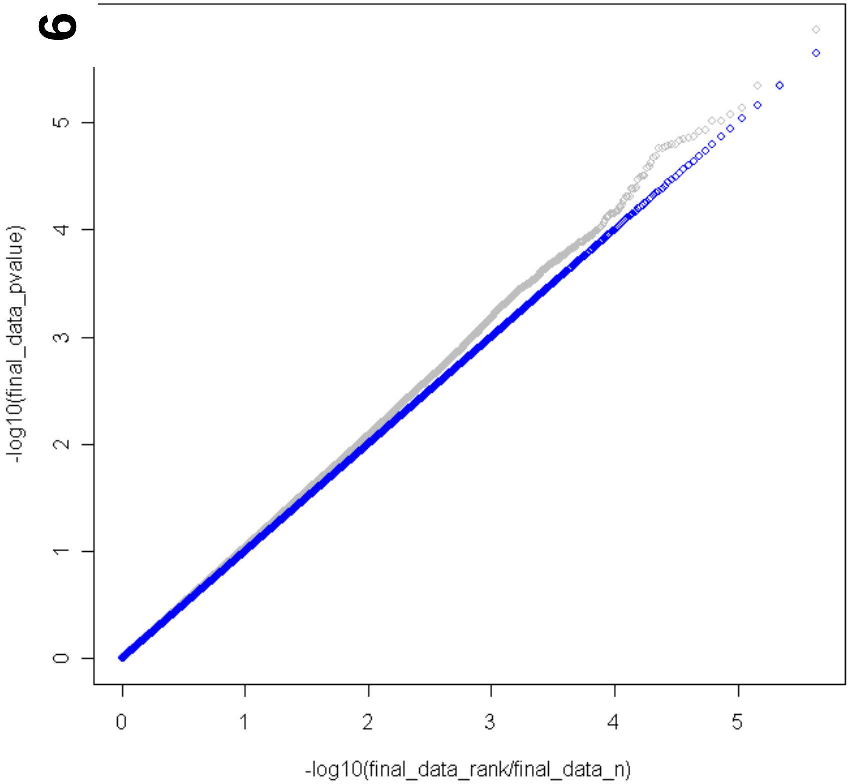


## Comparison of qq-plots before and after elimination of SNPs with low call rate and low MAF illustrates utility of preliminary association tests in calibrating quality control thresholds:

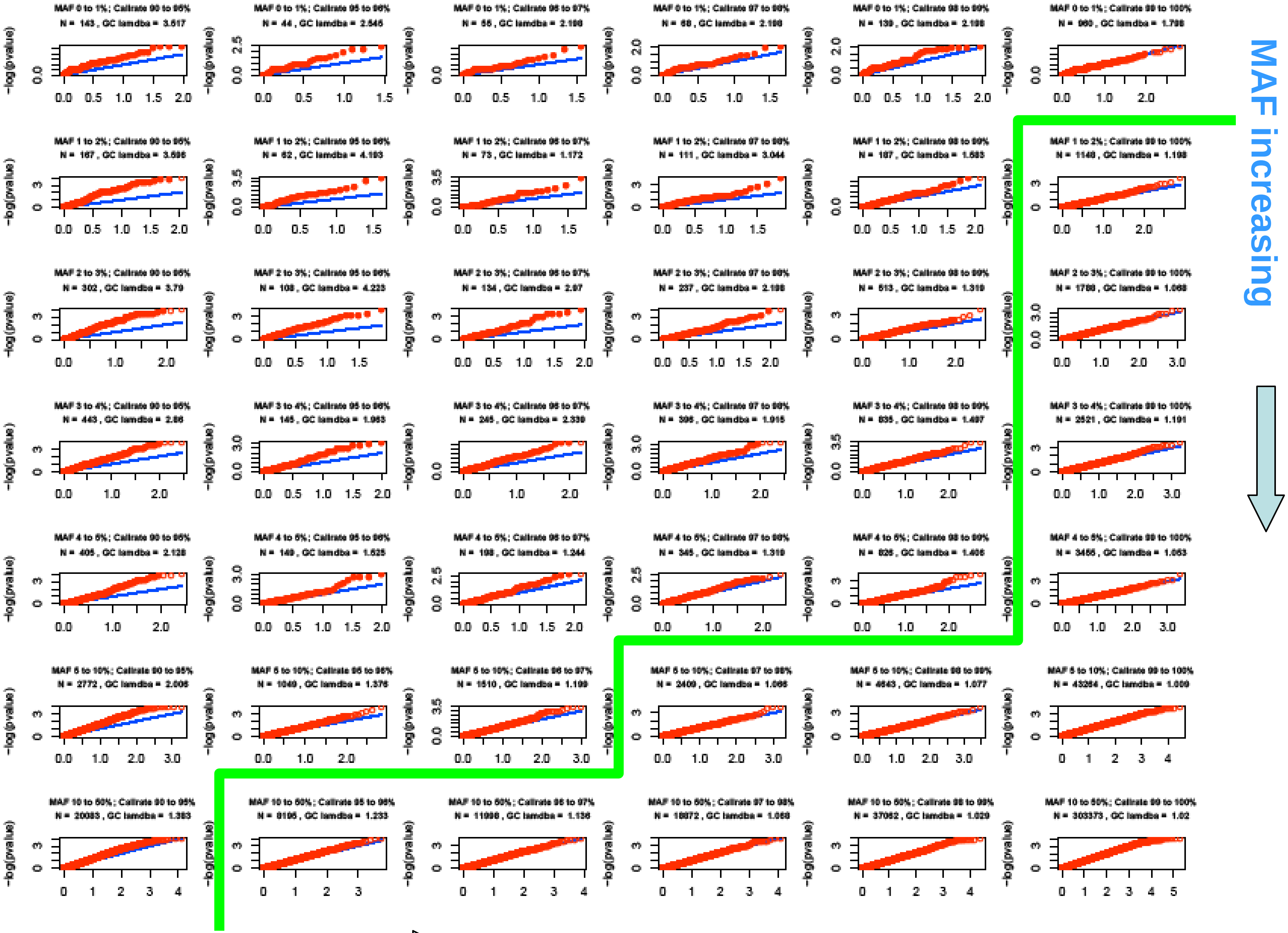
MAF > 1%, Call rate > 90%



0.01<=MAF<0.05 and call rate >= 99%
0.05<=MAF<0.10 and call rate >= 97%
0.10>=MAF and call rate >= 95%



# SNPs excluded from Filtered Dataset



MAF increasing

**Call Rate Increasing**

SNPs included in Filtered Dataset