A minicourse on

Genomewide association studies (GWAS) Part I

Matti Pirinen

FIMM, University of Helsinki

December 4th 2012

Published Genome-wide associations 07/2012
p<5e-8 for 18 trait categories

Digestive system disorder
Cardiovascular disorder
Metabolic disorder
Immune system disorder
Neurological disorder
Liver enzyme measurement
Lipid or lipoprotein measurement
Inflammatory marker measurement
Hematological measurement
Body measurement
Cardiovascular measurement
Other measurement
Chemical compound
Biological process
Cancer
Other disease
Other trait
Trait mapping in progress

Contents:

1.Concepts And Rationale

2.Technology and Data

3.Statistics and Analysis

4.Results and Hot topics

EMBL-EBI

National Human Genome Research Institute

www.genome.gov/gwastudies
www.ebi.ac.uk/fgpt/gwas/

Concepts

- Individuals (humans, mice, flies, grass)
- Phenotypes, traits
  - Quantitative (height, cholesterol levels)
  - Discrete (case-control, Parkinson's disease)
- EXPLAIN
  - Understanding, mechanisms, therapeutics
  - Example: SORT1 (Musunuru 2010)
- PREDICT
  - Early intervention, agriculture

Y=μ+G+E+(GxE)

- Y phenotype, μ population mean
- E nvironment
  - Chemicals, temperature, food, physical activity...
- G enetics
  - DNA, A,C,G,T, 3e+9 bases, 22 chrs + X + Y, diploid, meiosis
  - Genes → proteins, 2e+4, 1-2 % of DNA
  - Single nucleotide polymorphisms (SNPs)
- GxE not in this course

Linkage disequilibrium

- Non-independence of alleles at (two) SNPs in population
- r^2 and D', (Wray et al. 2011, p4 BOX 2)
- Look at 1000 Genomes browser: rs4988235
- LD → SNPs tag each other in a population sample
- Basis for gene mapping with a SNP panel

Heritability

- $Y=G+E$
- Simplest model without measured variants or environment
  - $G \sim N(0,s^2)$, cor(full-sibs)>cor(half-sibs)
  - $E \sim N(0,t^2)$, cor(household)>0
  - $H=var(G)/var(Y)=s^2/(s^2+t^2)$, heritability
  - var(Y) can be decomposed to s^2 and t^2 with variance component models, but environment is a potential confounder

Example: ACE twin model

- Y=A+C+E
  - A, additive genetic component
  - C, common environment for twins
  - E, non-shared environment for twins
- From (many pairs of) monozygotic (identical) twins AND dizygotic twins, var(A), var(C) and var(E) can be estimated → estimate of heritability as 2(cor(MZ)-cor(DZ))
  - But with strong assumptions!

Linkage analysis

- Families
  - Does trait correlate with genetic sharing at some parts of the genome?
  - Parametric Linkage analysis in pedigrees
  - Affected sib-pairs (non-parametric)
  - Need for families restricts sample size and thus size of genetic effects that can be found
  - Localisation is coarse since large blocks of genome are linked in close relatives

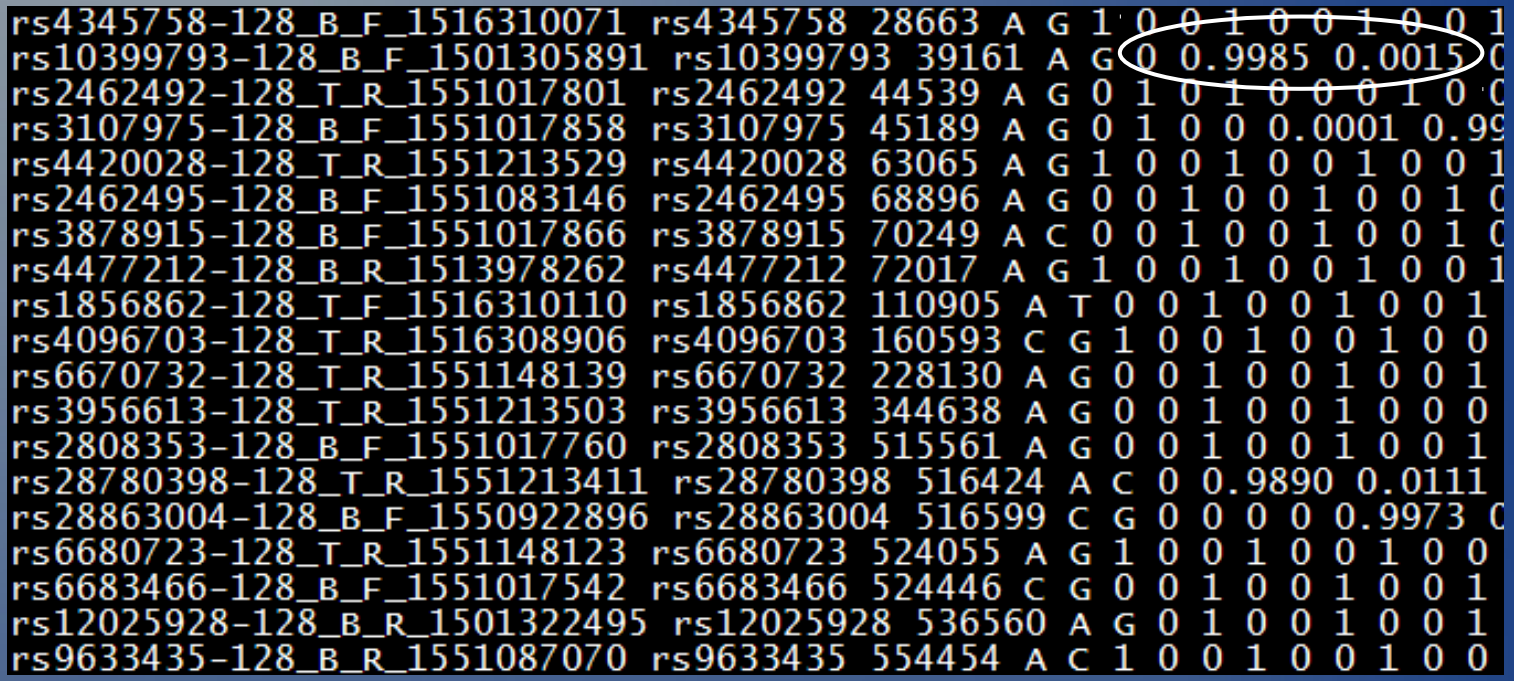
Association analysis

- If a large panel of SNPs can be genotyped, association between each SNP and the trait can be tested
  - Role of LD (HapMap project)
- Risch 2000, Fig 4
- Only samples from (homogeneous) population (not families) needed
- Genome coverage and localisation depend on #SNPs and LD in the population

Genotyping

- Chips
  - YOUTUBE: “Microarray method for genetic testing”
  - YOUTUBE: “DNA chips and microarrays”
  - Currently 50-200 euros per sample
- Intensities → genotype calls (Vukcevic p.36)
- Problems (Vukcevic p.38)

Genotype probabilities



The data comes with 5 types of “header” information

We report the probability for each of the three genotypes

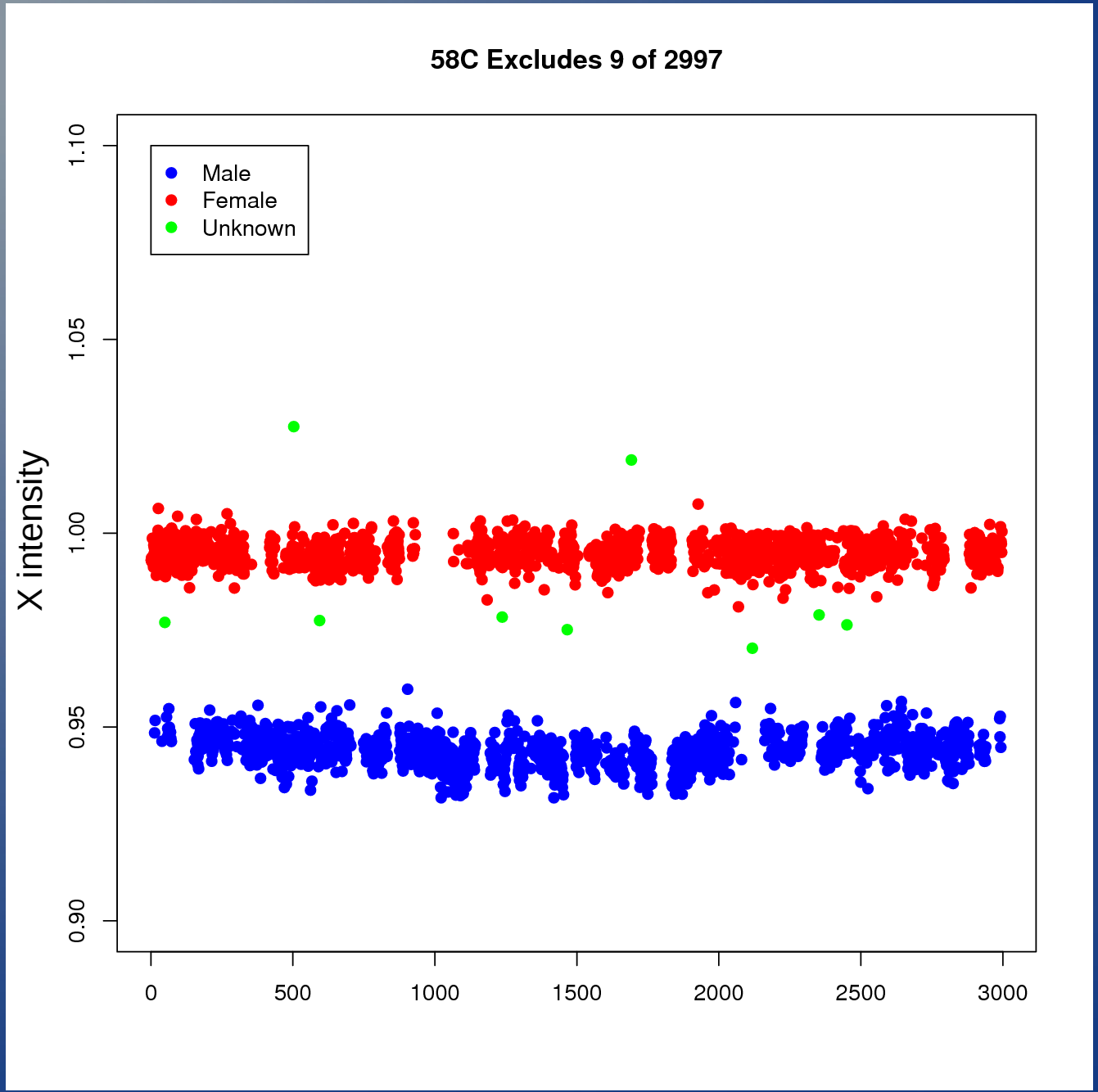
P(AA) = 0 P(AG) = 0.9985 P(GG) = 0.0015

A null call is represent by three zeros

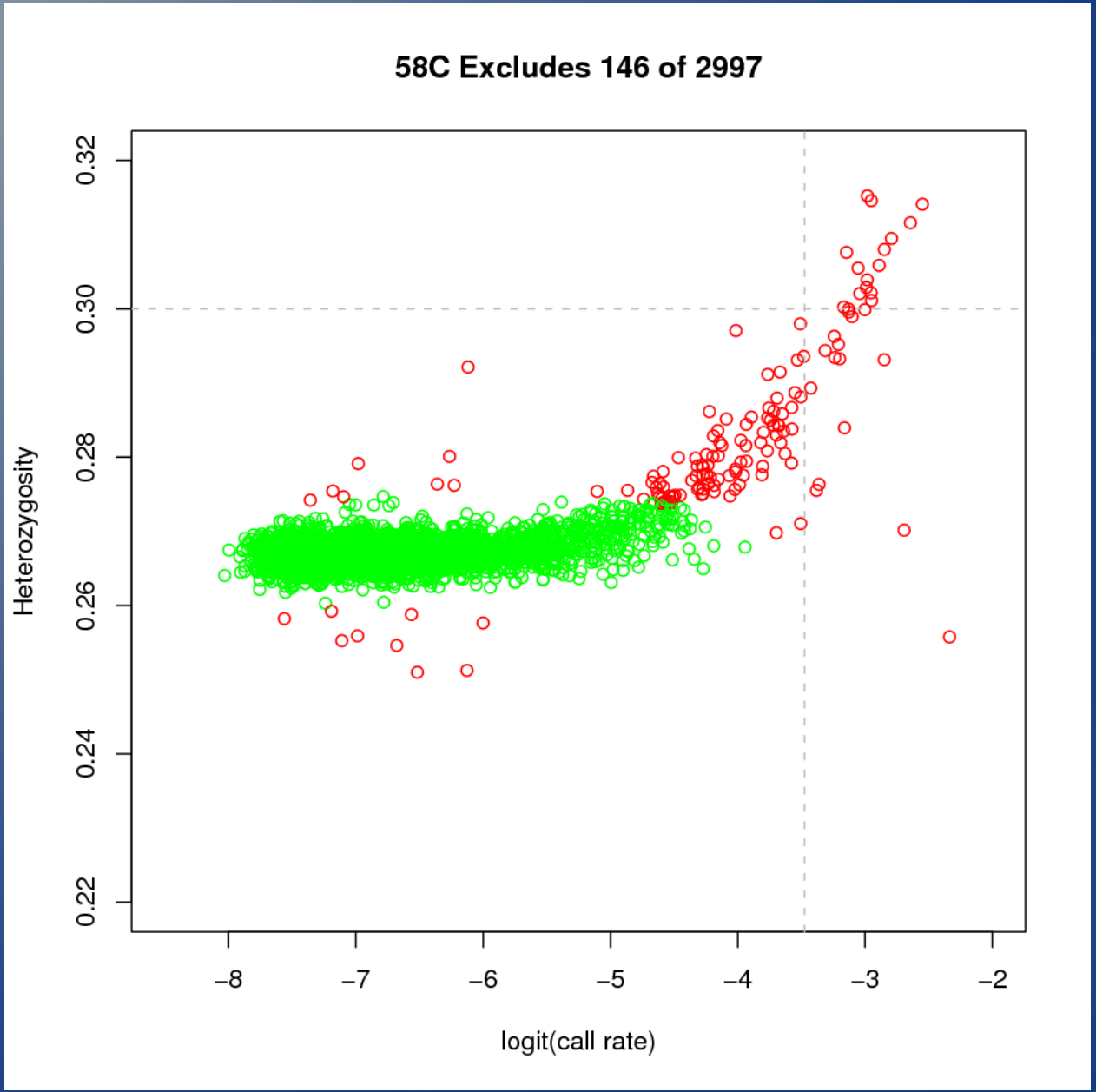
Quality control

- Remove individuals and SNPs that show low quality
- Anderson et al. 2010
- Individuals: Sex discrepancies, missingness, heterozygosity, relatedness, ancestry
- SNPs: missingness, deviation from Hardy-Weinberg equilibrium, low minor allele frequency

Gender


58C Excludes 9 of 2997
X intensity
1.10
1.05
1.00
0.95
0.90
Male
Female
Unknown
0
500
1000
1500
2000
2500
3000

Individual QC



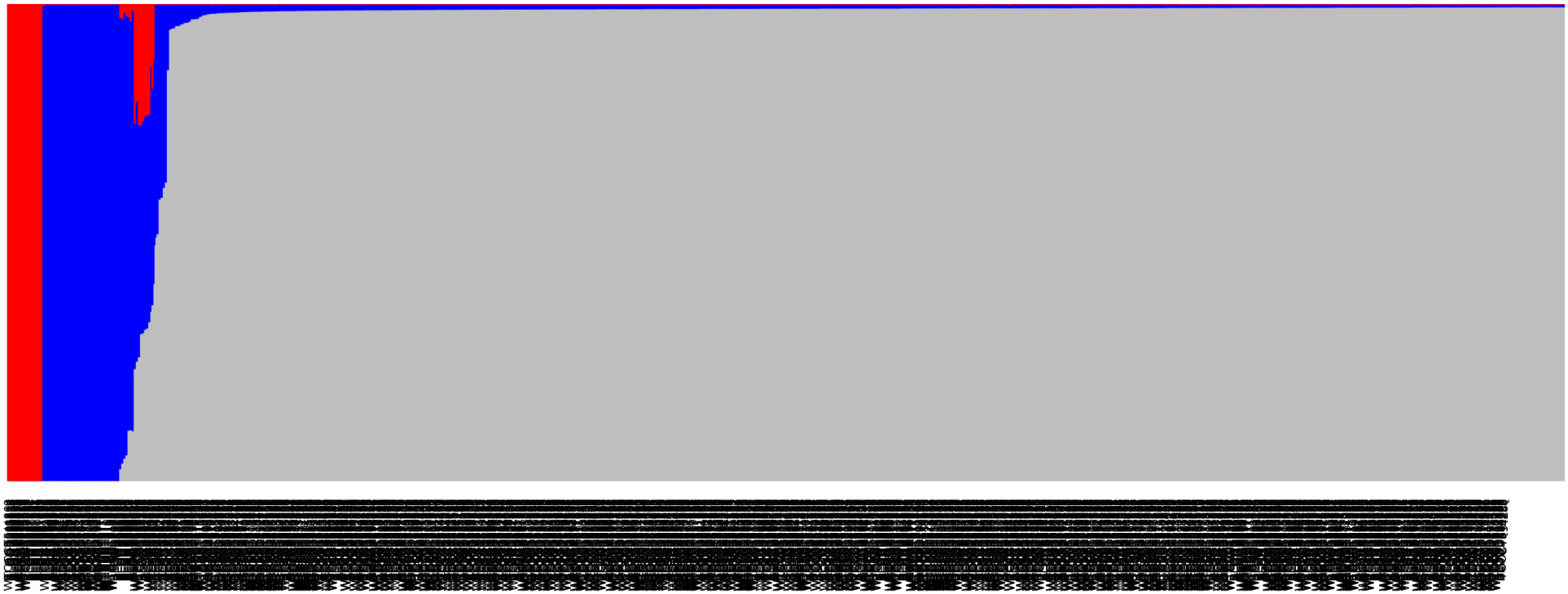
Relatedness

We look for relatedness in two stages:

- Calculate genome-wide allele sharing between all pairs
- For each individual estimate the portion of their genome IBD 0,1 and 2 with their closest relative

We exclude individuals until there is no pair of individuals with more than 5% IBD

Relatedness

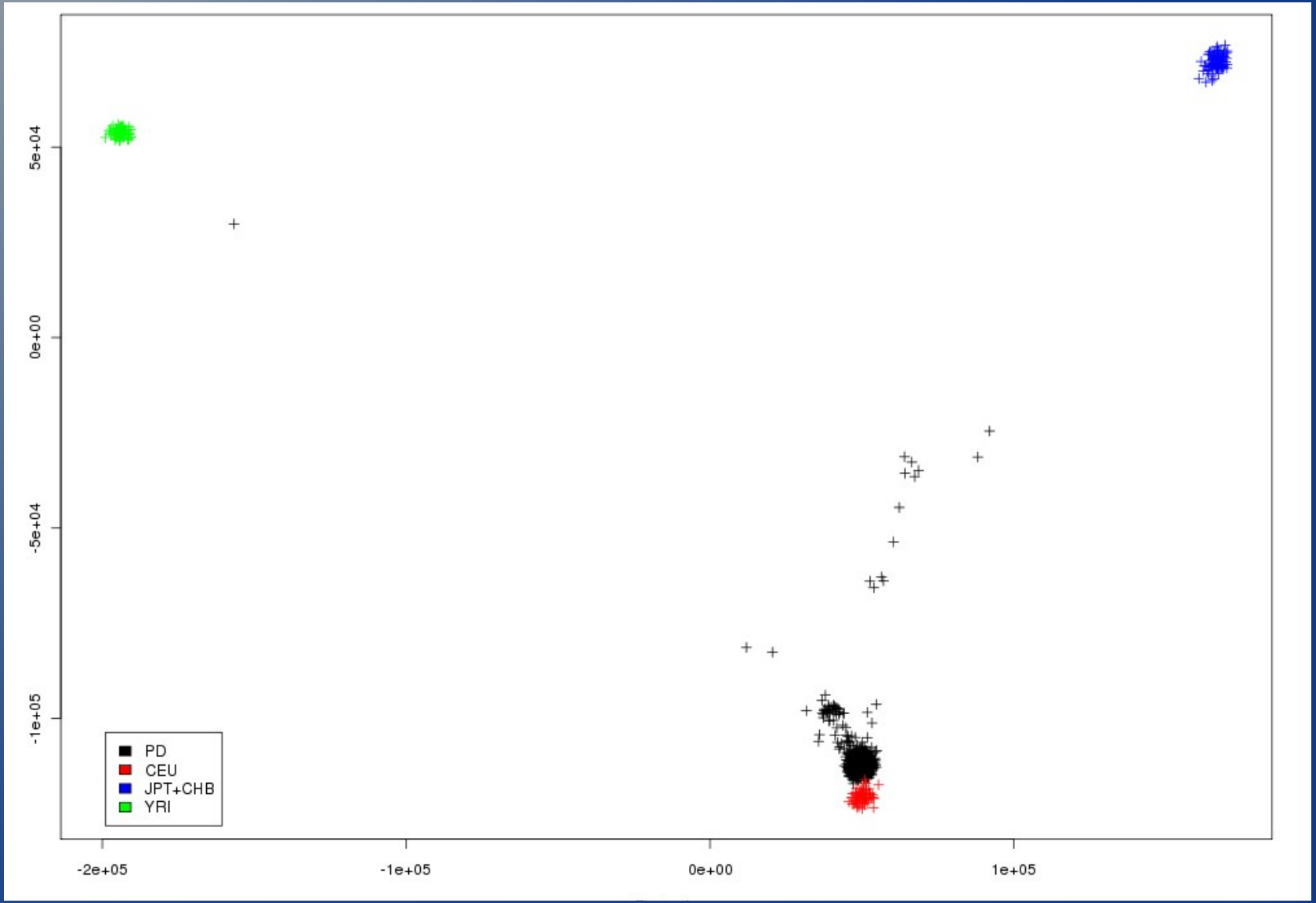


Population structure

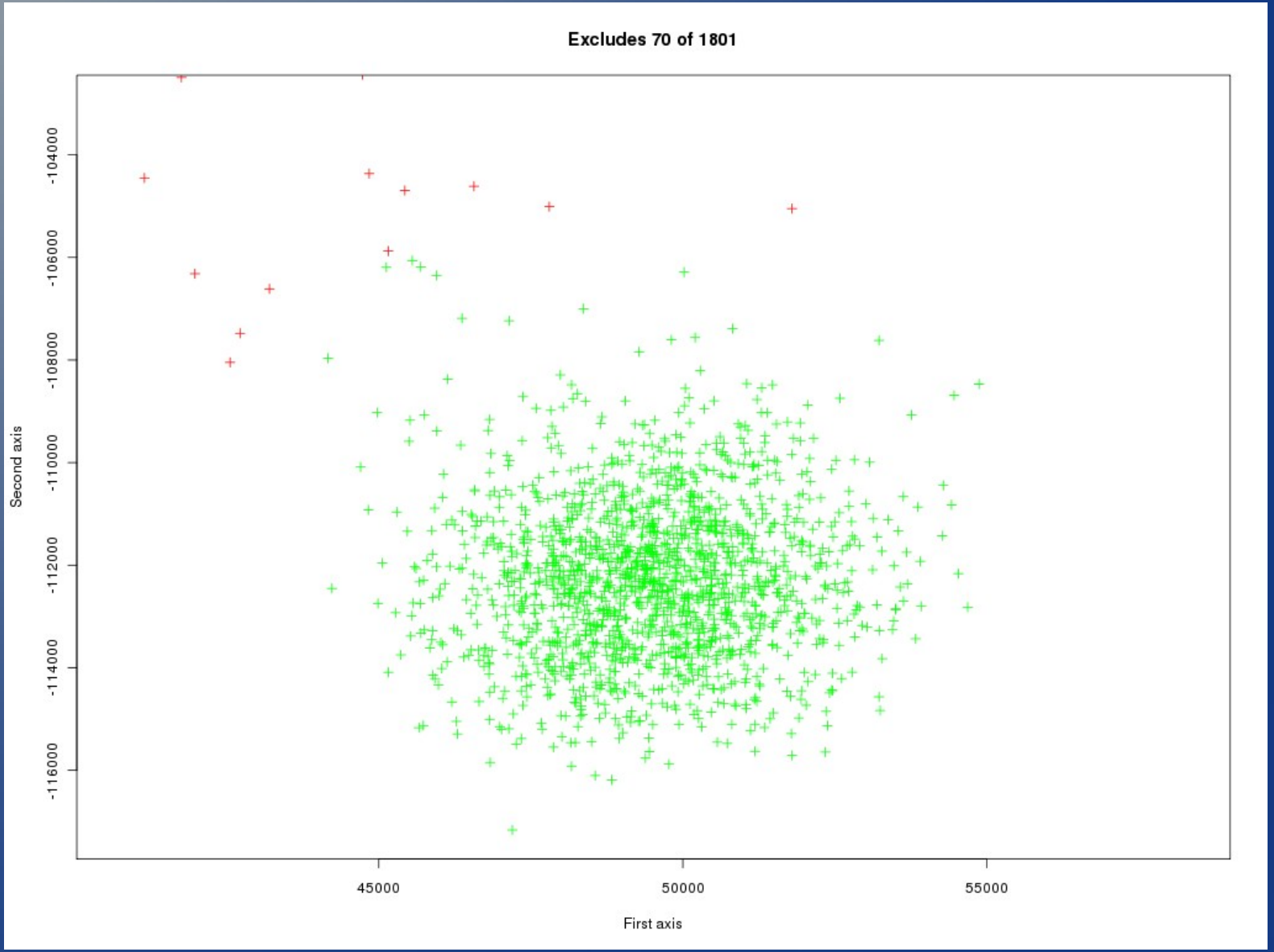
For all collections we project individuals onto axis of variation which are chosen to explain the diversity in HapMap

We cluster individuals and exclude those who’s ancestry outlier with respect to the rest of the sample

Population structure



Population structure



QC genotype calls

The optimal approach is either to model the calling errors or to look at all cluster plots

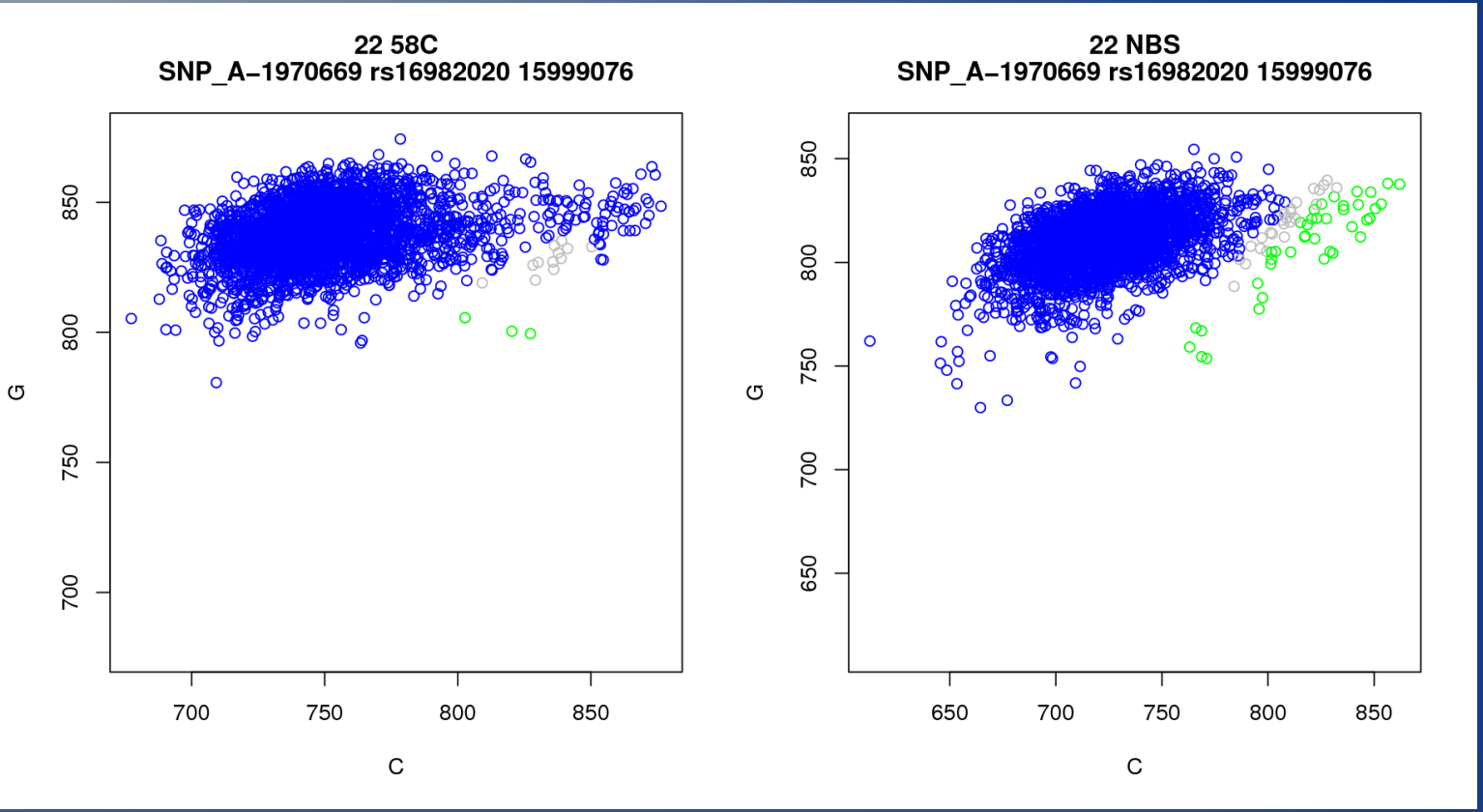
SNP filtering is then a short cut

The level of SNP filtering is therefore a trade-off

Software for looking at cluster plots using binary data
http://sourceforge.net/projects/evoker/

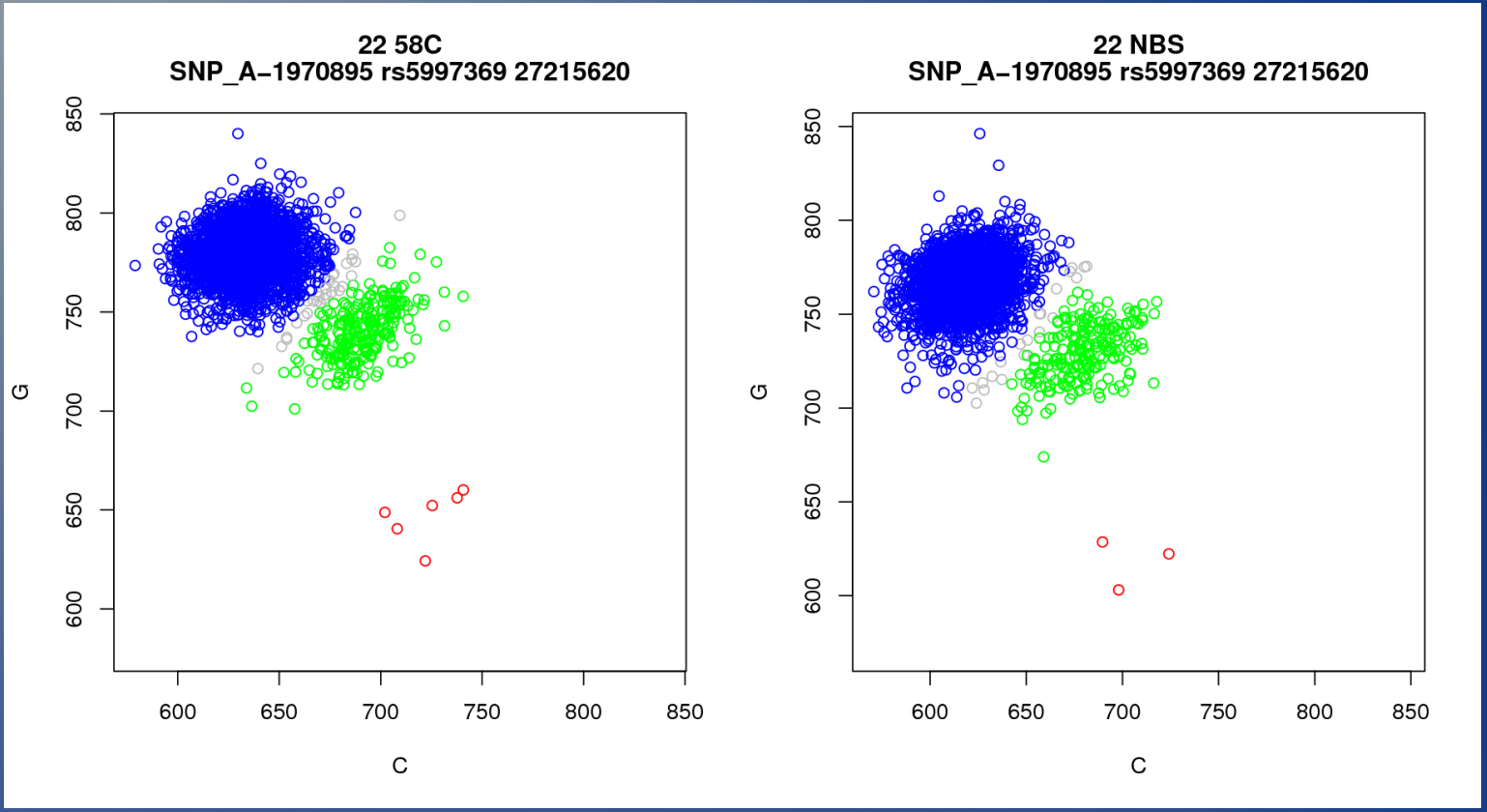
(Jeff Barrett)

Minor allele frequency



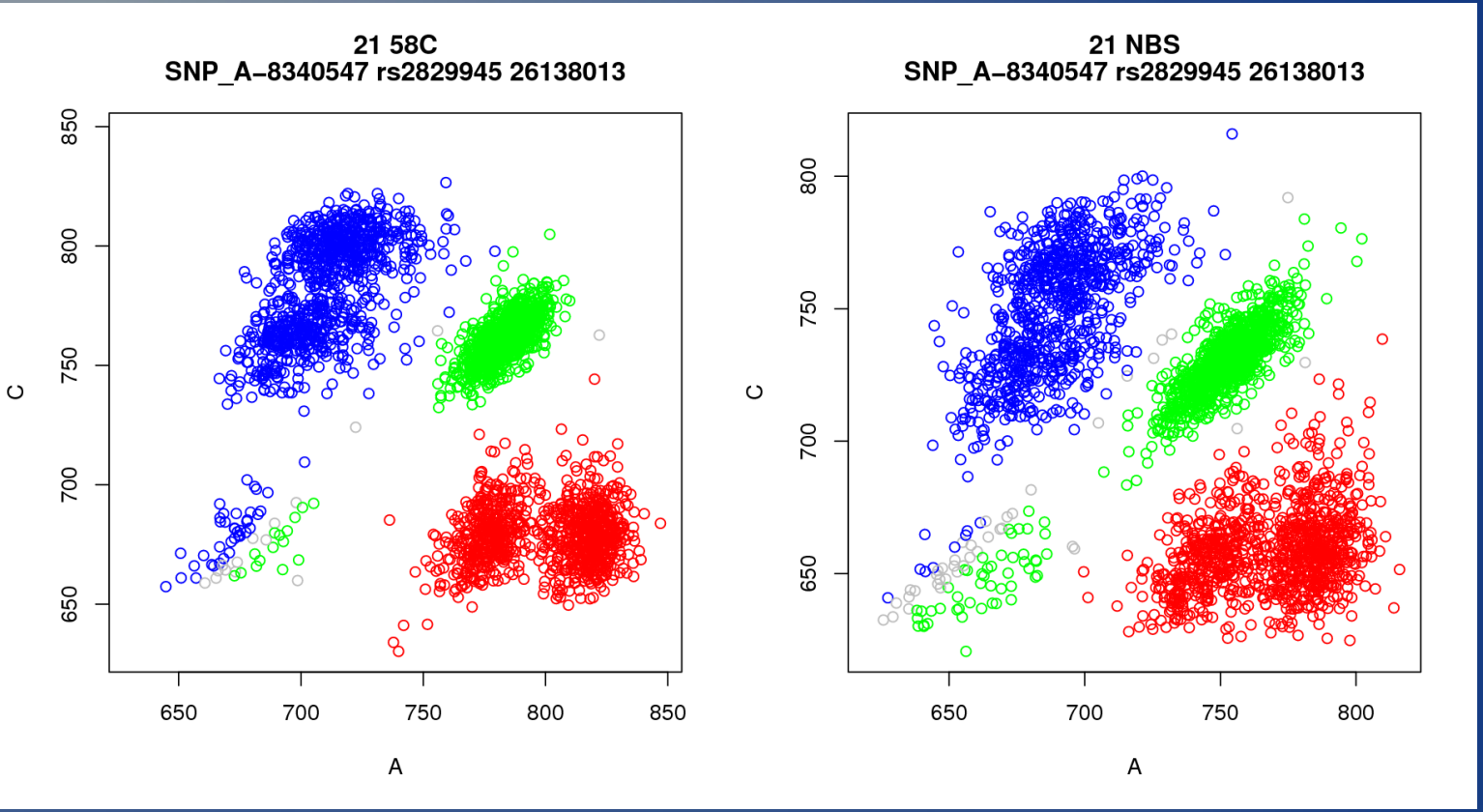
Genotype calling algorithms like to model the extra clusters

Low Information



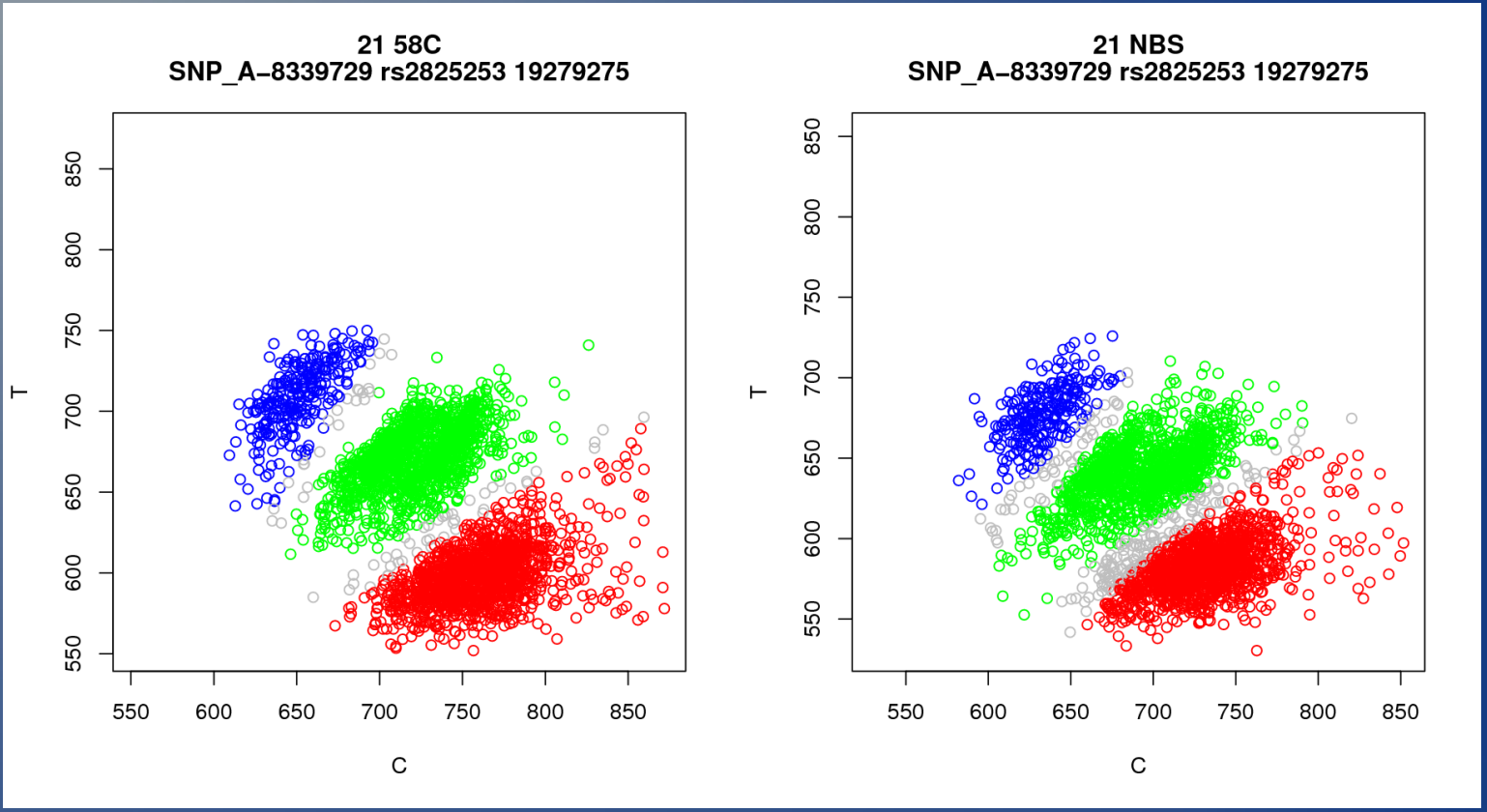
Information is good a identify SNPs with insufficient resolution of genotypes

Hardy-Weinberg



Clustering errors, or putative CNP often distorts HW equilibrium

Missingness



High missingness is also indicative of clustering failure or poor cluster resolution

SNP exclusions

The application of SNP exclusions is aimed at capturing three classes:

- Poor clustering

- Low signal to noise

- Non-SNP like variation (known CNVs)

Guide to looking at cluster plots in the Supplementary material of WTCCC 2007 paper

Information metric

A natural way to assess the utility of the calls at a SNP is to assess the information about the allele frequency. (info column in SNPTEST output)

Compare two measures of information

- Using expectation of genotype calls

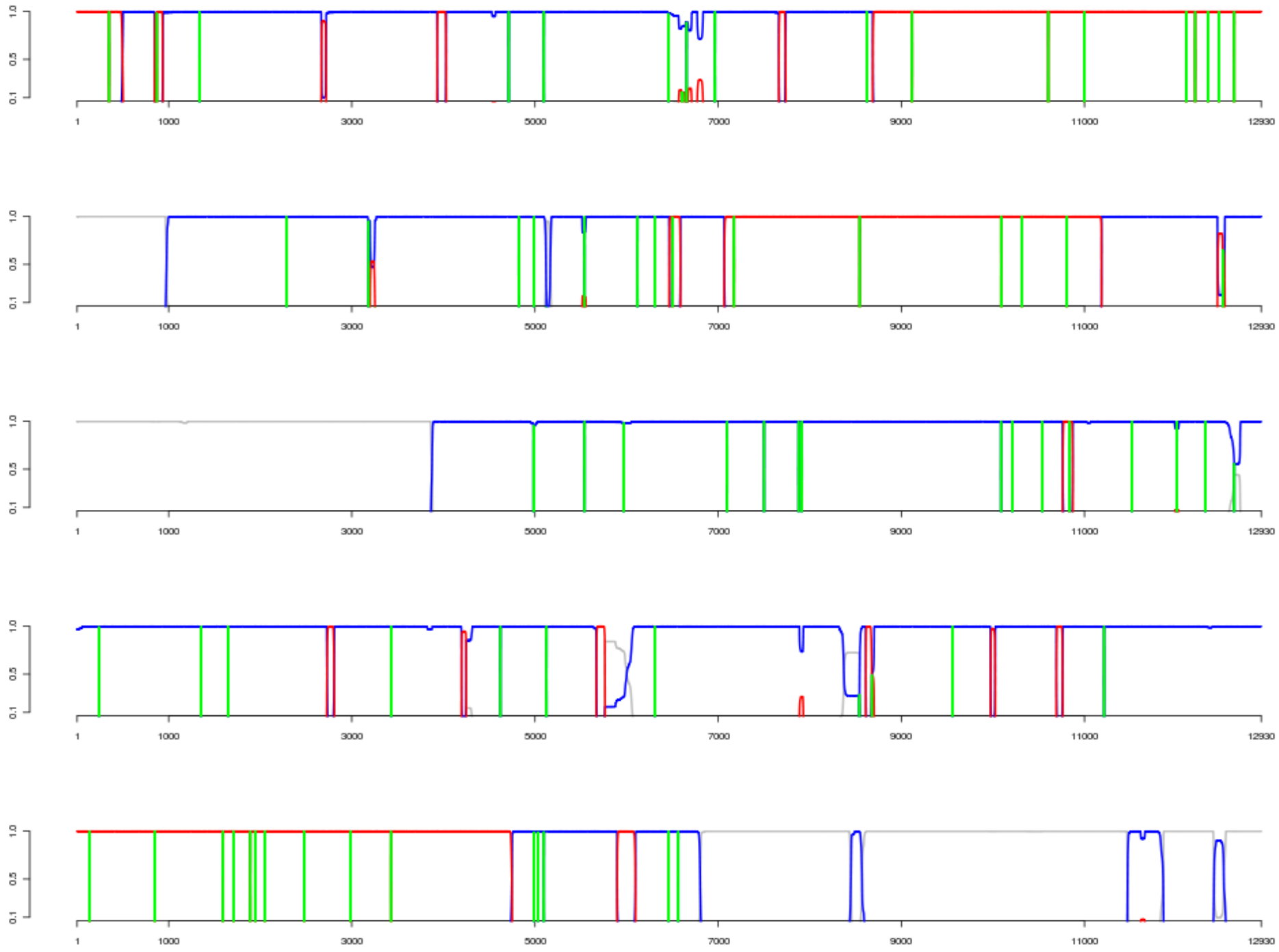
- Accounting for probabilistic genotypes

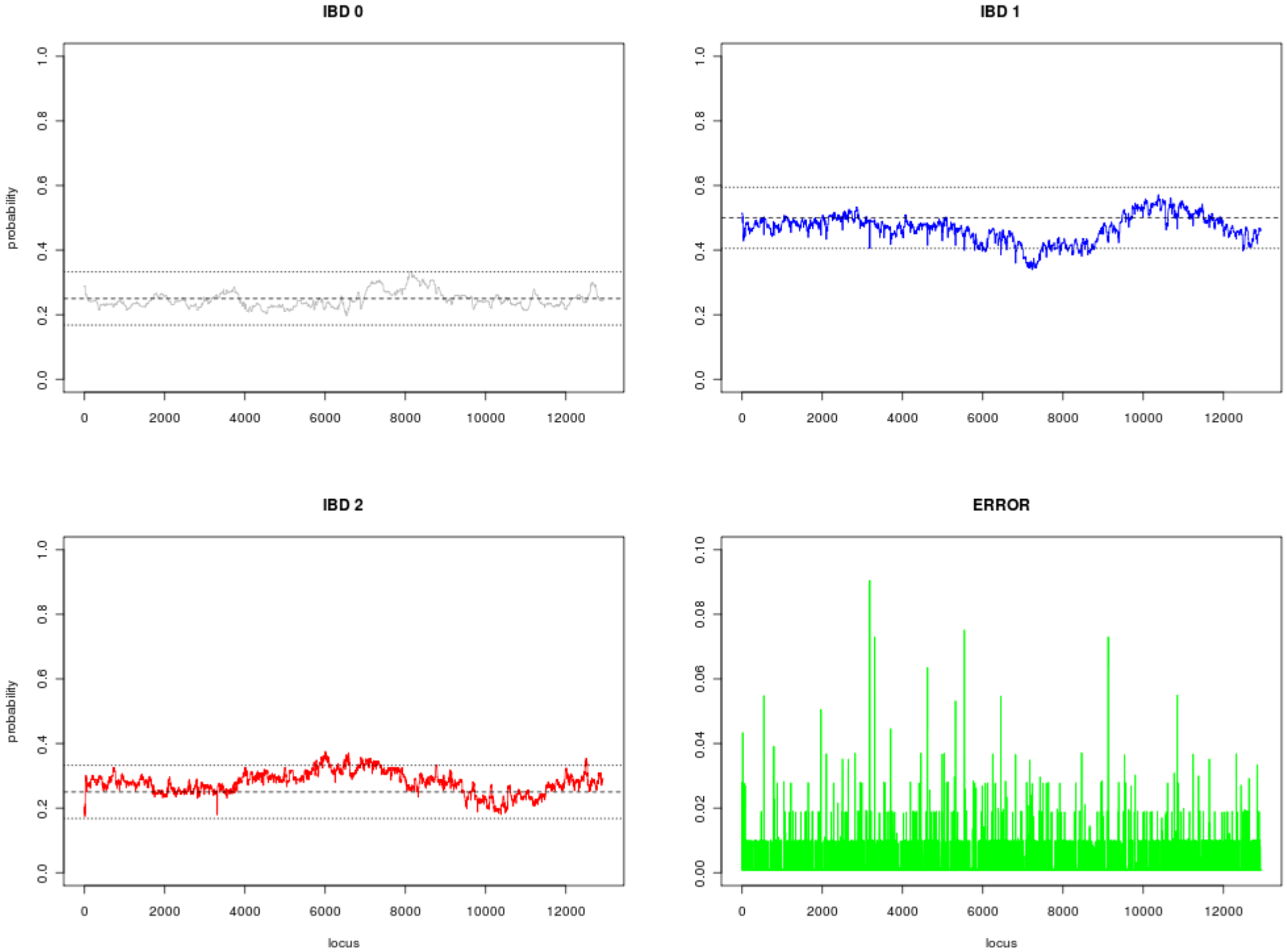
SNP QC

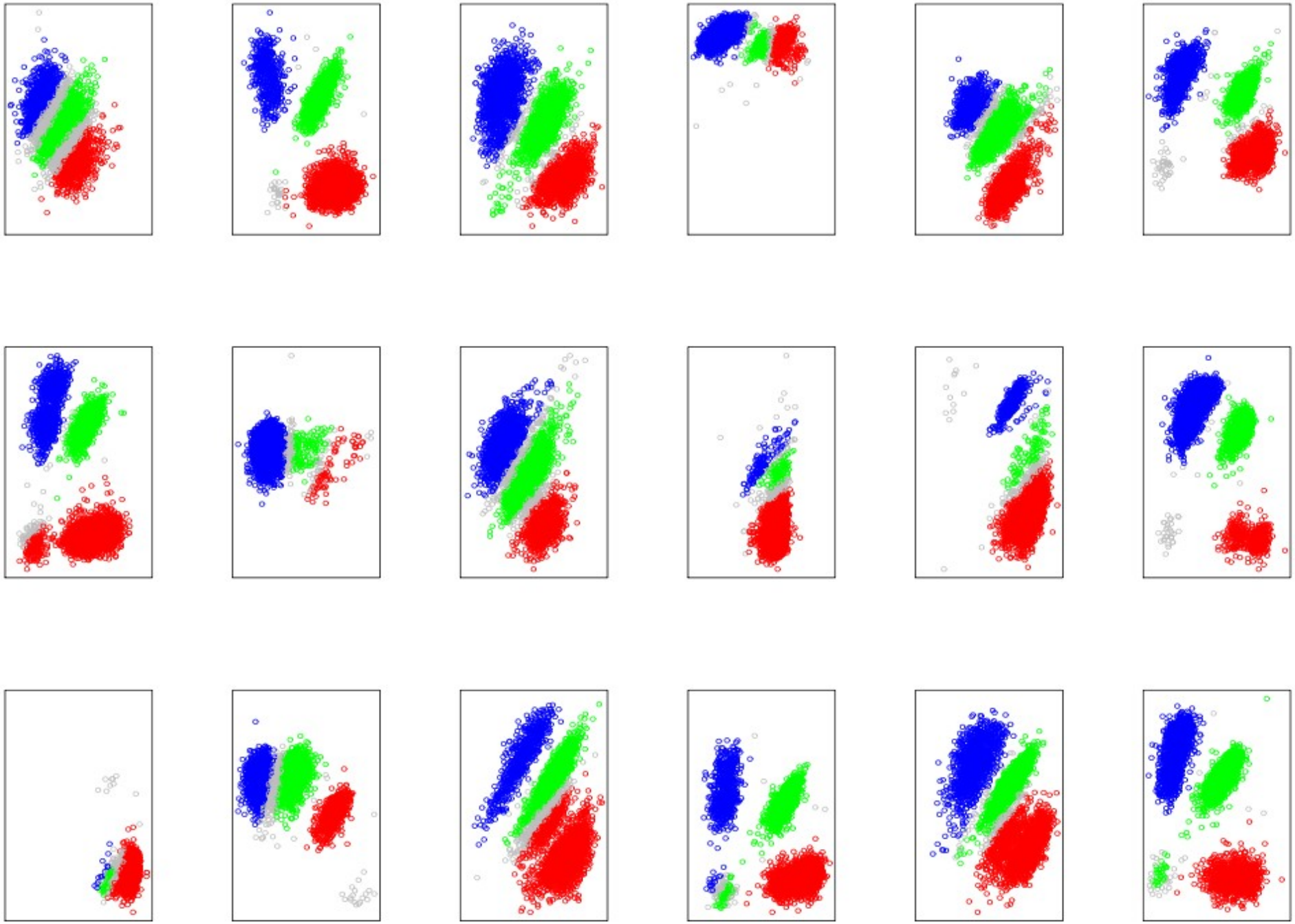
Most poor cluster plots can be identified by these simple filters

We have chosen a set of SNP filters on the basis of the control-control comparison

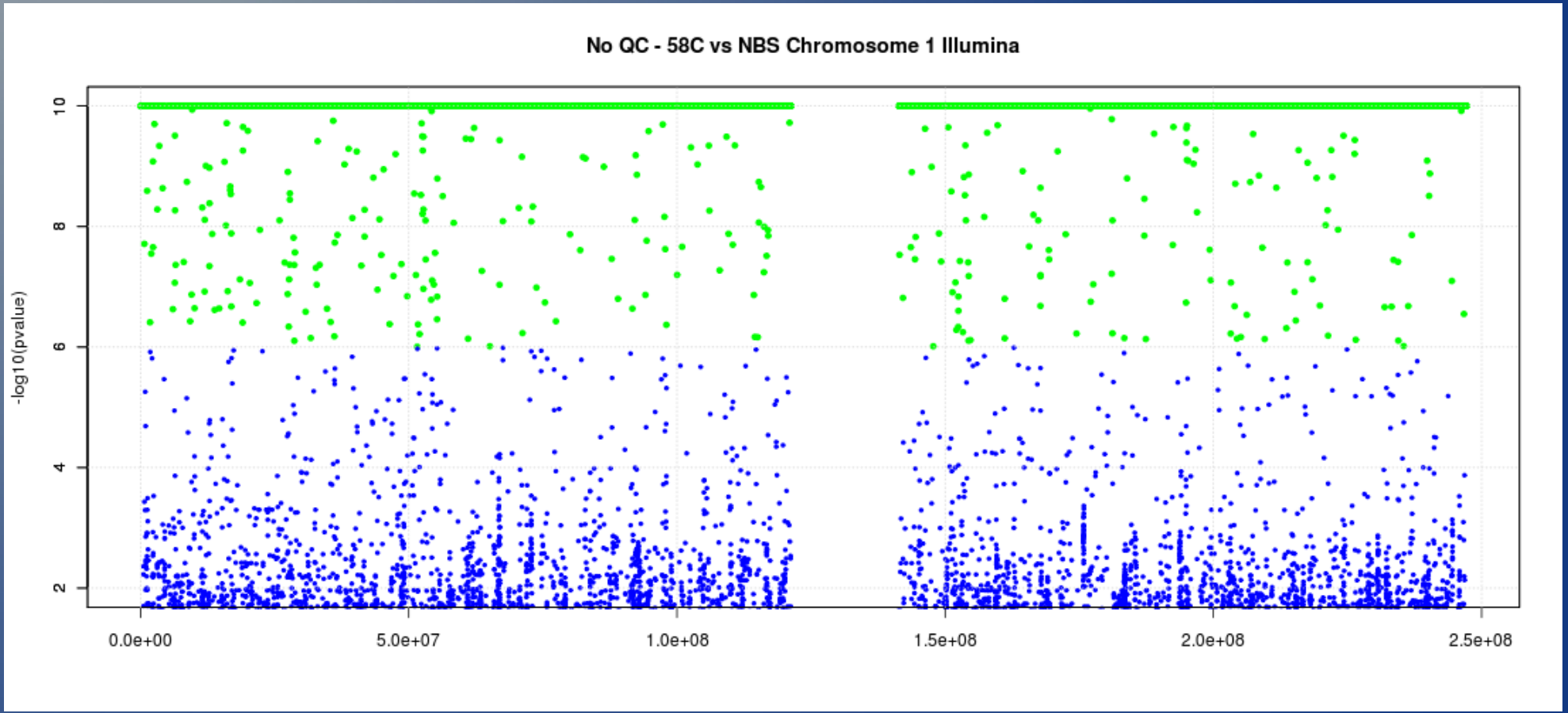
Filters depend on data quality and requires an iterative approach for each data set…


1.0
0.5
0.1
1
1000
3000
5000
7000
9000
11000
12930


IBD 0
IBD 1
IBD 2
ERROR
probability
locus
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.02
0.04
0.06
0.08
0.10
0
2000
4000
6000
8000
10000
12000

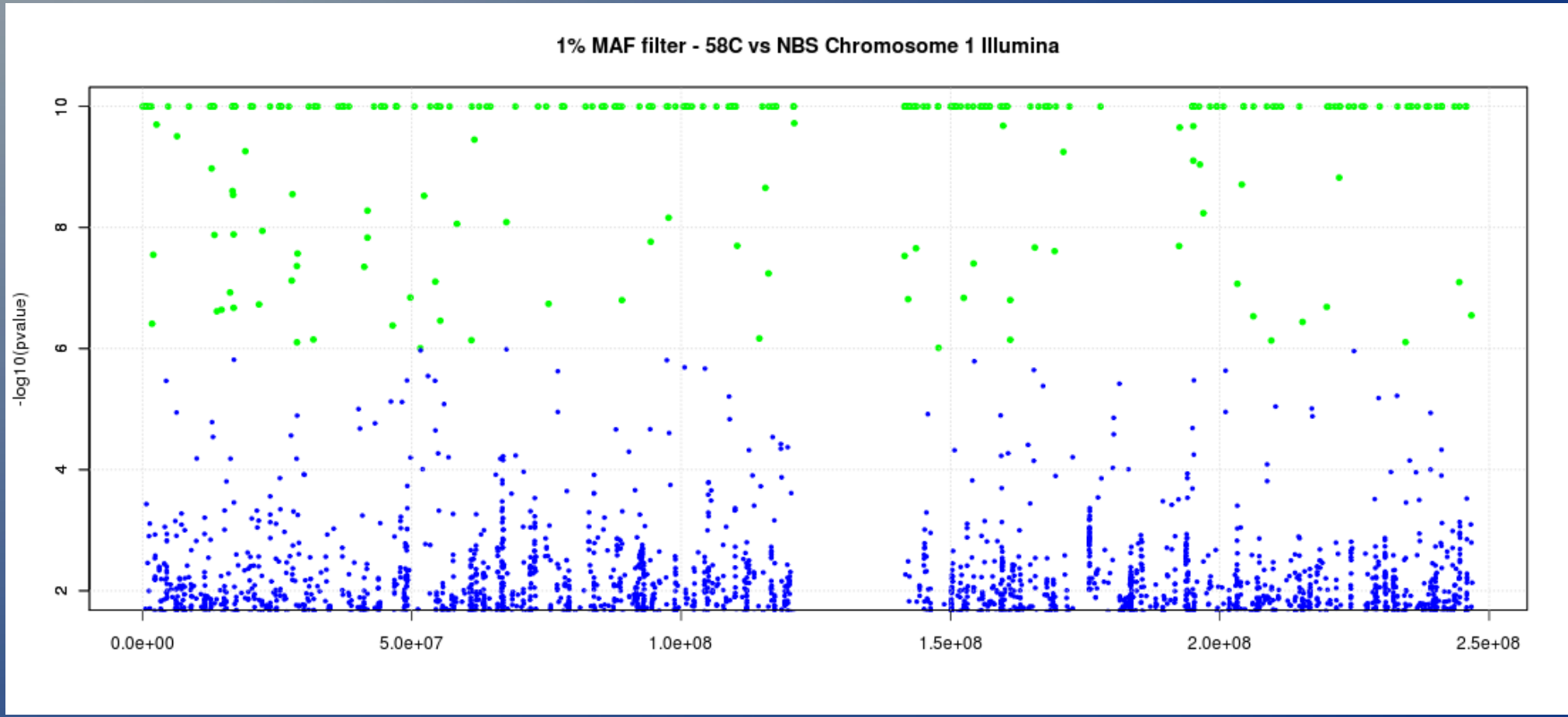


Testing for association



100% of SNPS

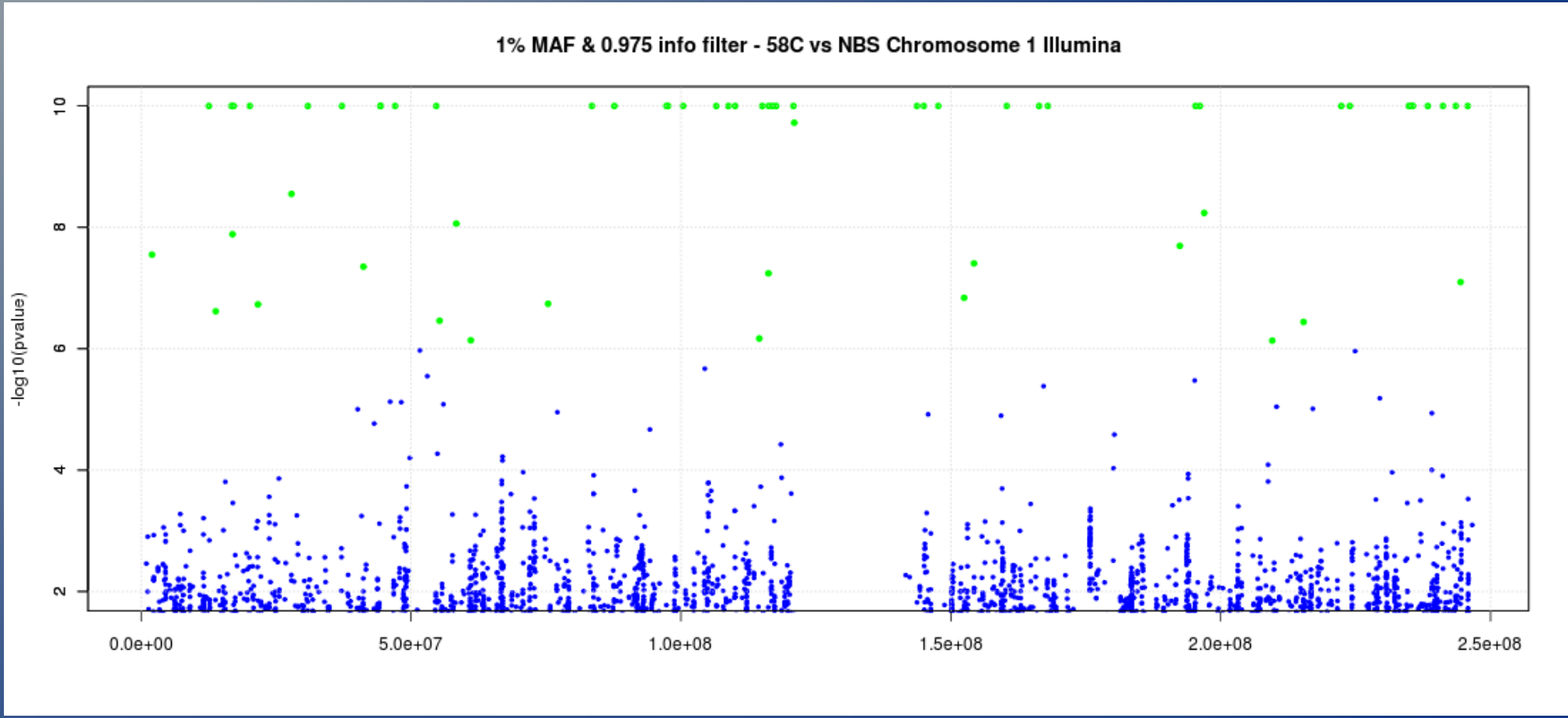
Testing for association



80.69% of SNPS

1% MAF

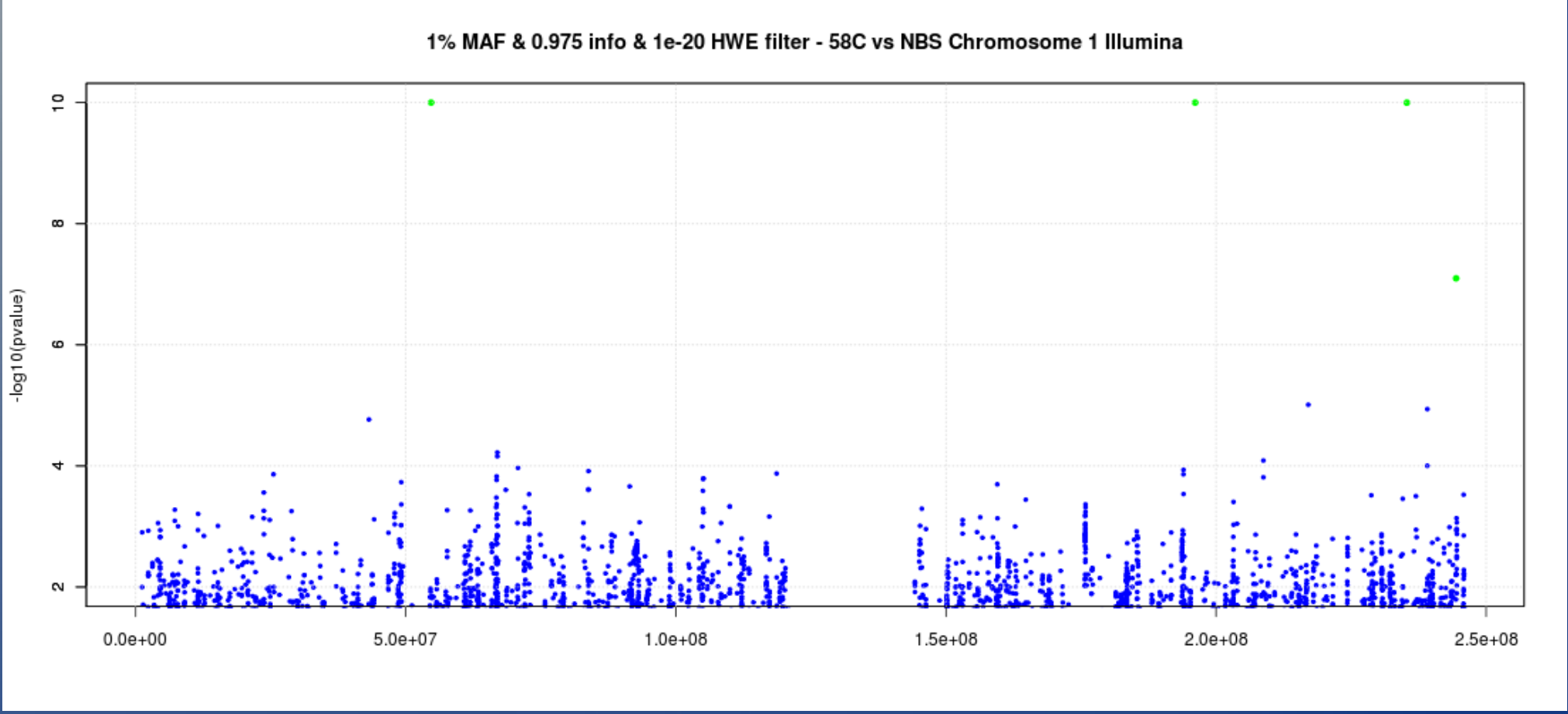
Testing for association



78.36% of SNPS

MAF > 1% & info > 0.975

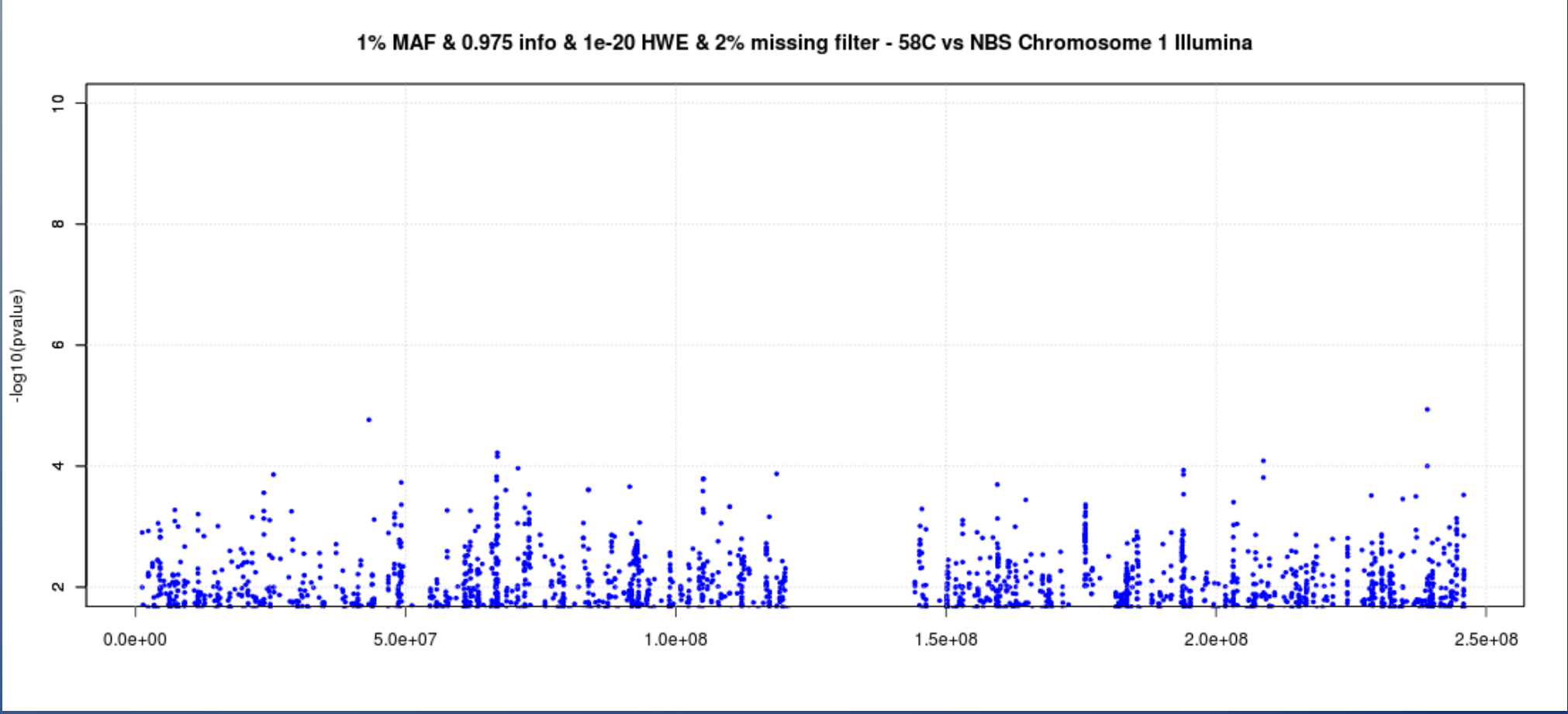
Testing for association



78.36% of SNPS

MAF > 1% & info > 0.975 & HW < 1e-20

Testing for association



77.92% of SNPS

MAF > 1% & info > 0.975 & HW < 1e-20 miss < 2%

A “clean” data set

Applying these filters results in a typically only 10 to 100 signals of association (e.g. $p < 1e\text{-}7$)

Comparing number of exclusions depends heavily on low MAF (which depends on population etc)

Important to apply filters to each set of collections or calls separately

Of those remaining…still lots of checking to do

Fingerprint markers, gender checks

Genotype QC, intensity outlier

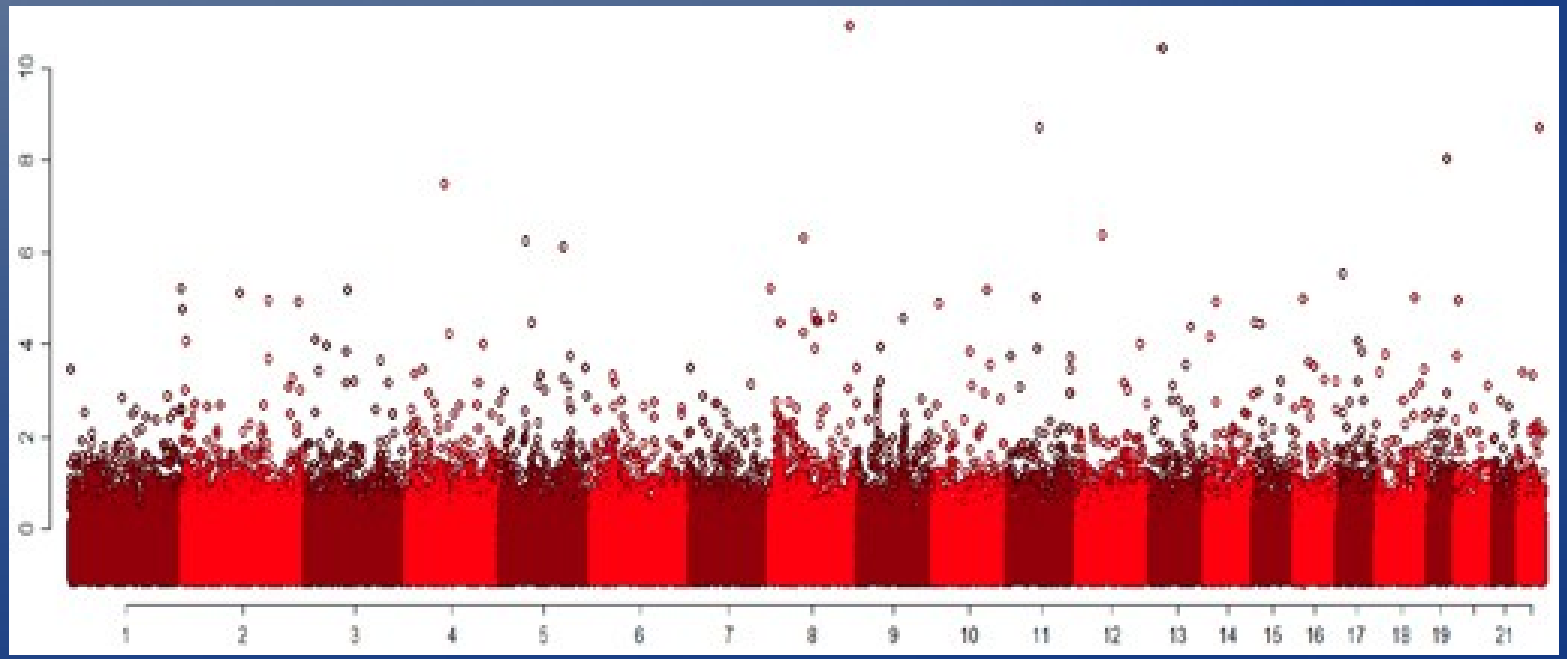
Given an individual's disease status, their genotype at a SNP, is independent of, and identically distributed to, other samples conditional on the underlying genotype frequencies in cases and controls

Duplicate checks, relatedness

Population structure analysis

A retracted paper

- Sebastiani et al. “Genetic signatures of exceptional longevity in humans” Science July 2010
  - Was retracted in July 2011 because QC had not been done properly!



Sebastiani et al. 2010 Science