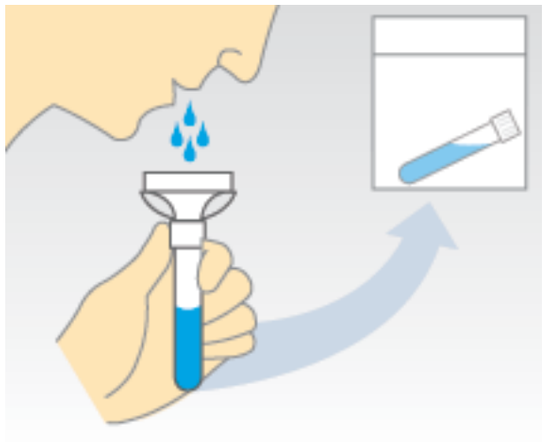
Genotype Calling

using Unsupervised Learning

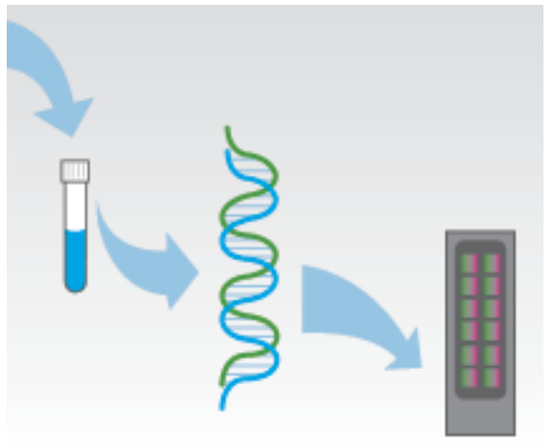
by Terry Huang

Motivation

DNA sample kit



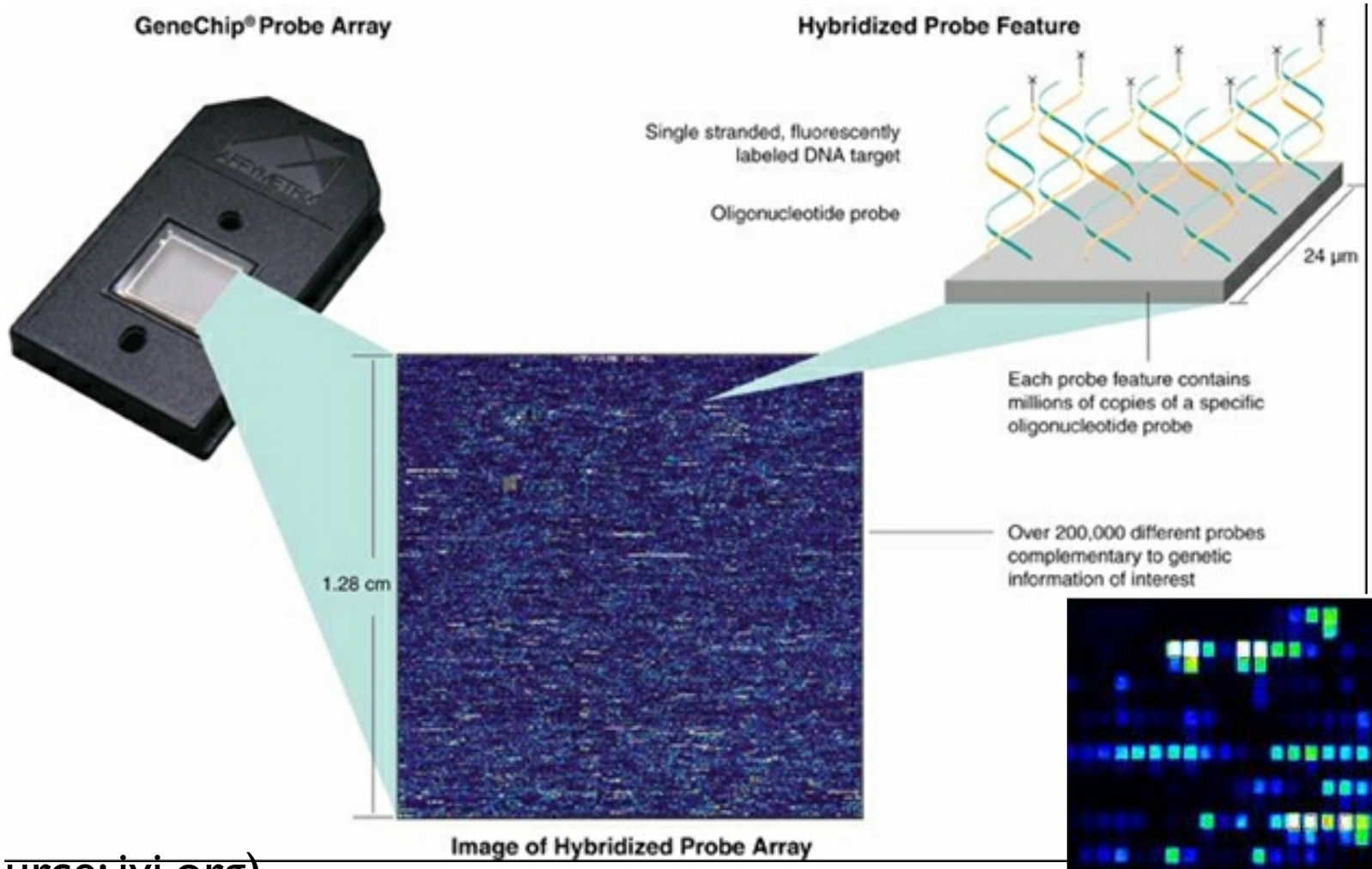
Provide sample



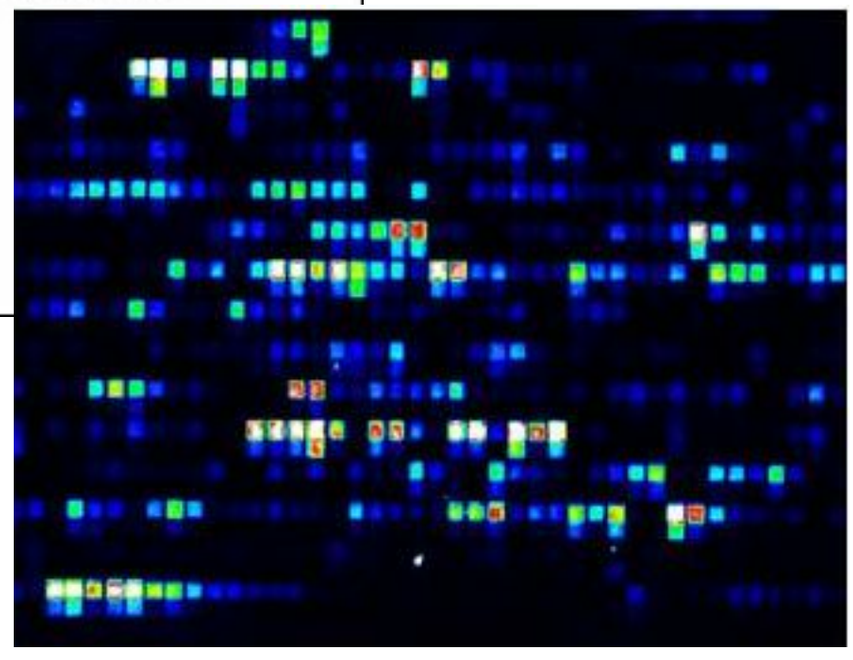
Analyze sample

(Source: 23andme.com)

SNP Microarray



(Source: jyi.org)

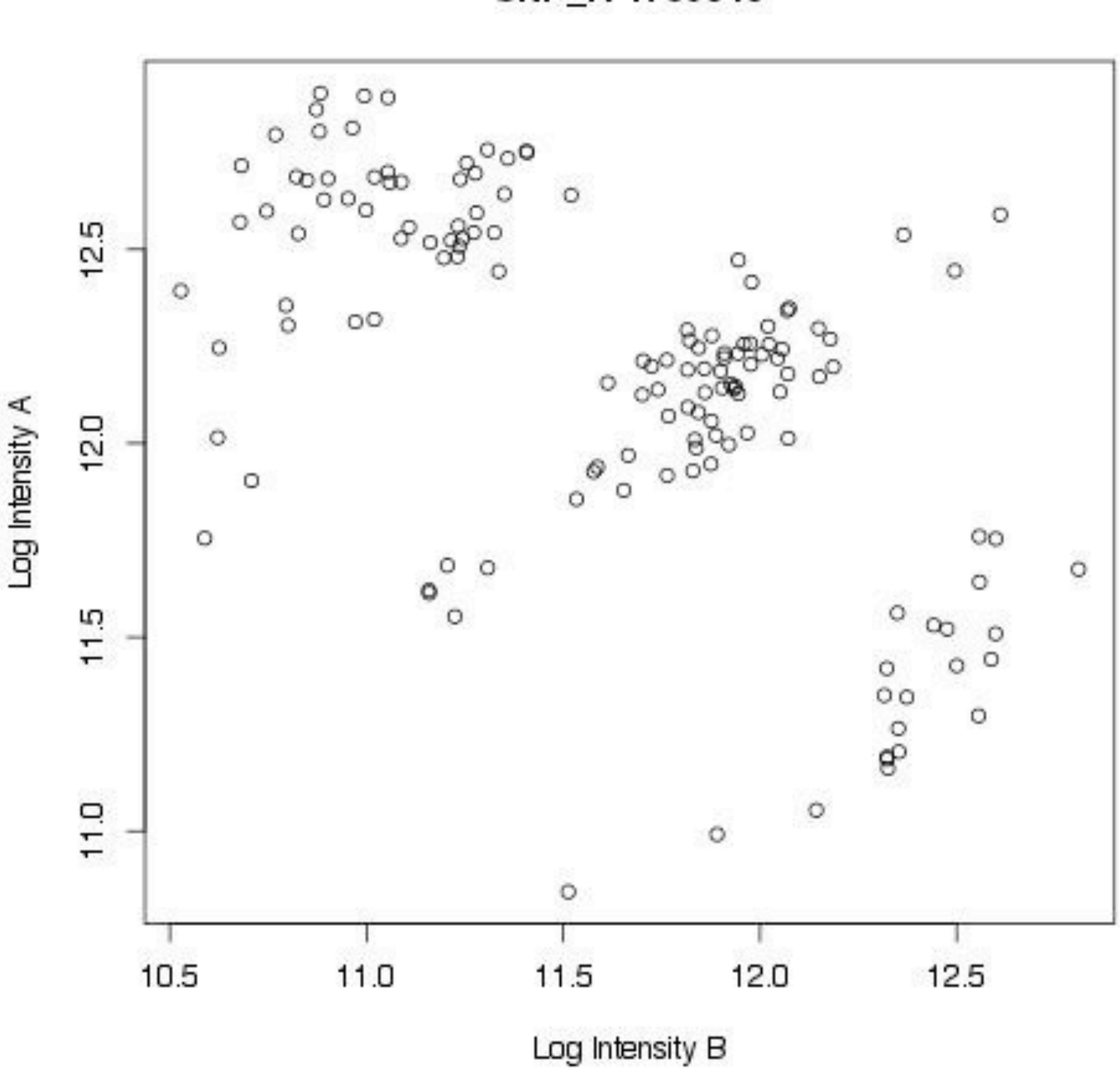


(Source: umass.edu)

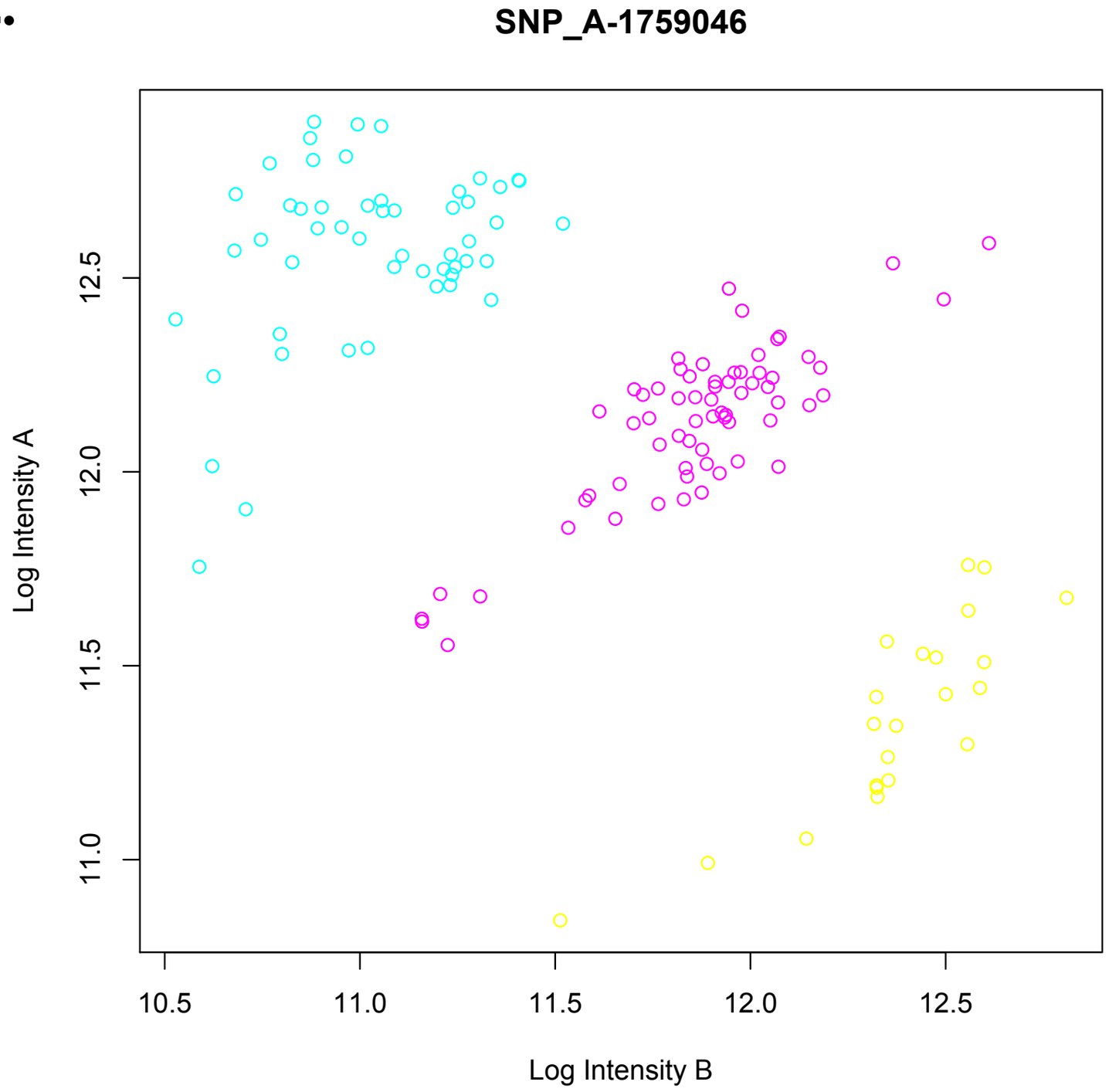
Computational Problem

Given a set of observations (x_1, x_2, ...x_n),
partition the observations into K sets,
such distance between the each observation
and cluster mean is minimal.

Input:


SNP_A-1759046
Log Intensity A
12.5
12.0
11.5
11.0
10.5
11.0
11.5
12.0
12.5
Log Intensity B

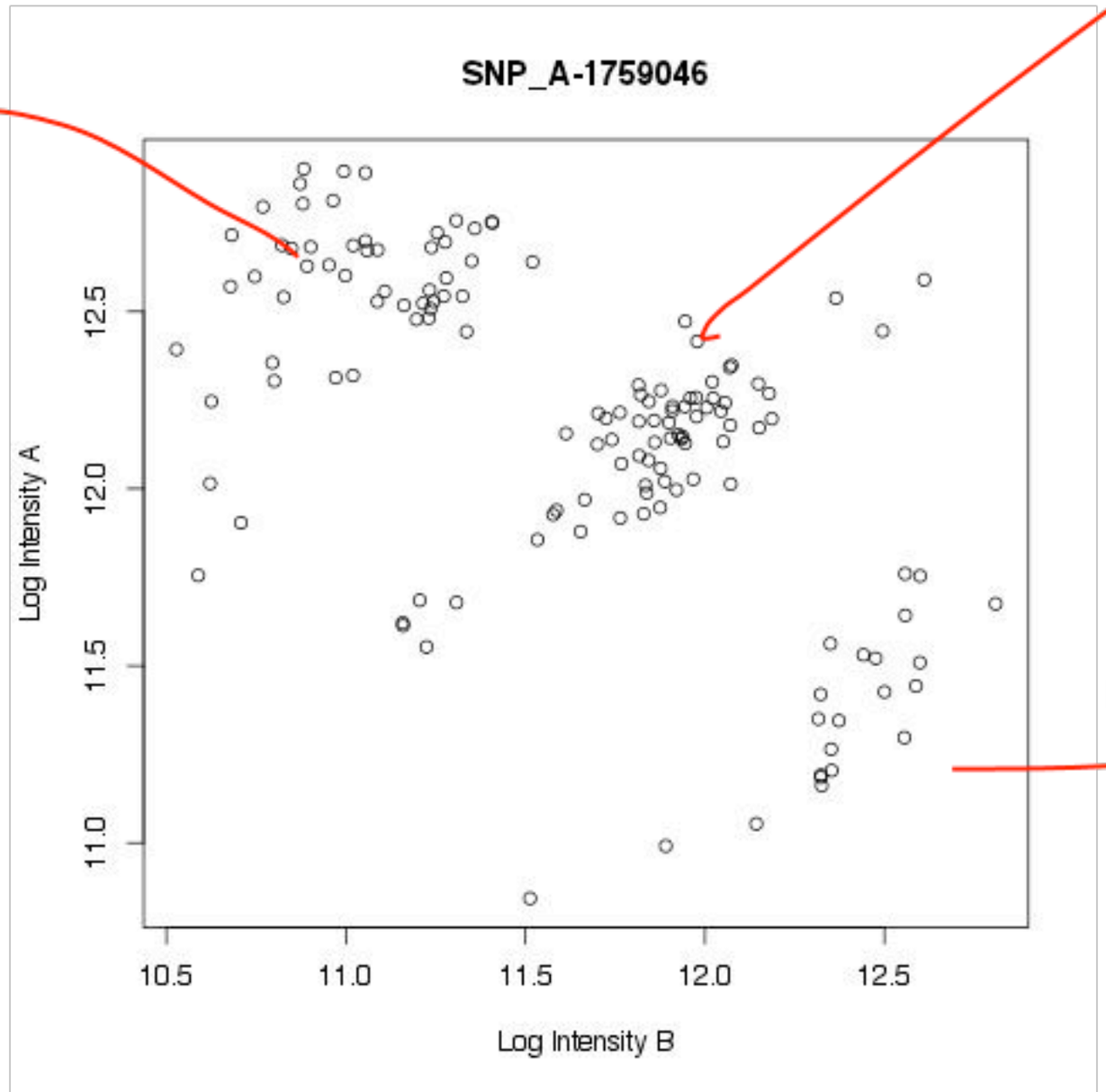
Output:


SNP_A-1759046
Log Intensity A
12.5
12.0
11.5
11.0
10.5
11.0
11.5
12.0
12.5
Log Intensity B

Baseline Method

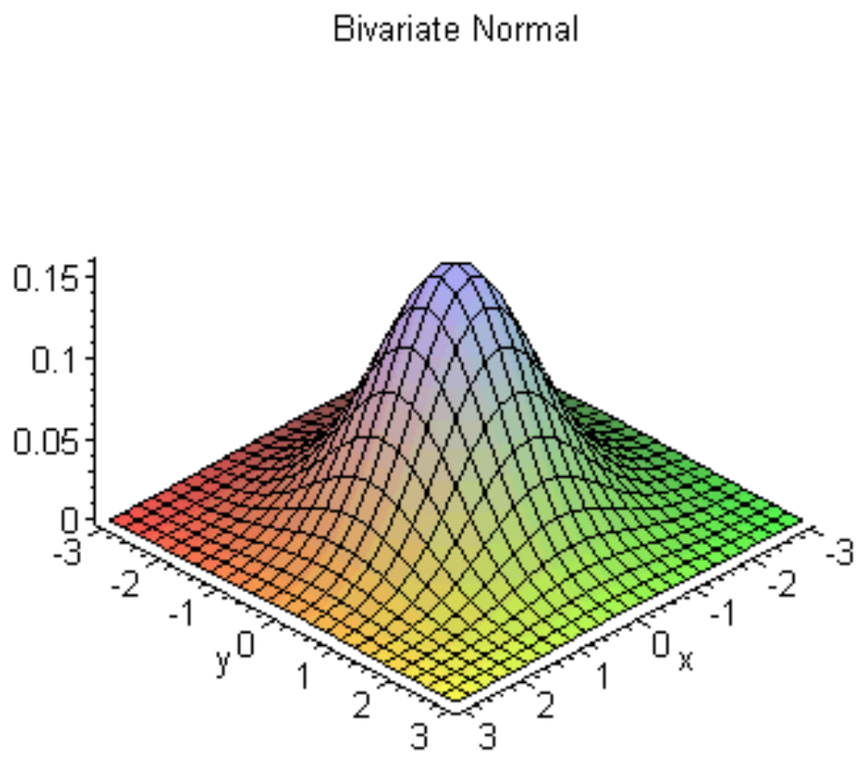
log A / log B ≈ 1

log A / log B > 1



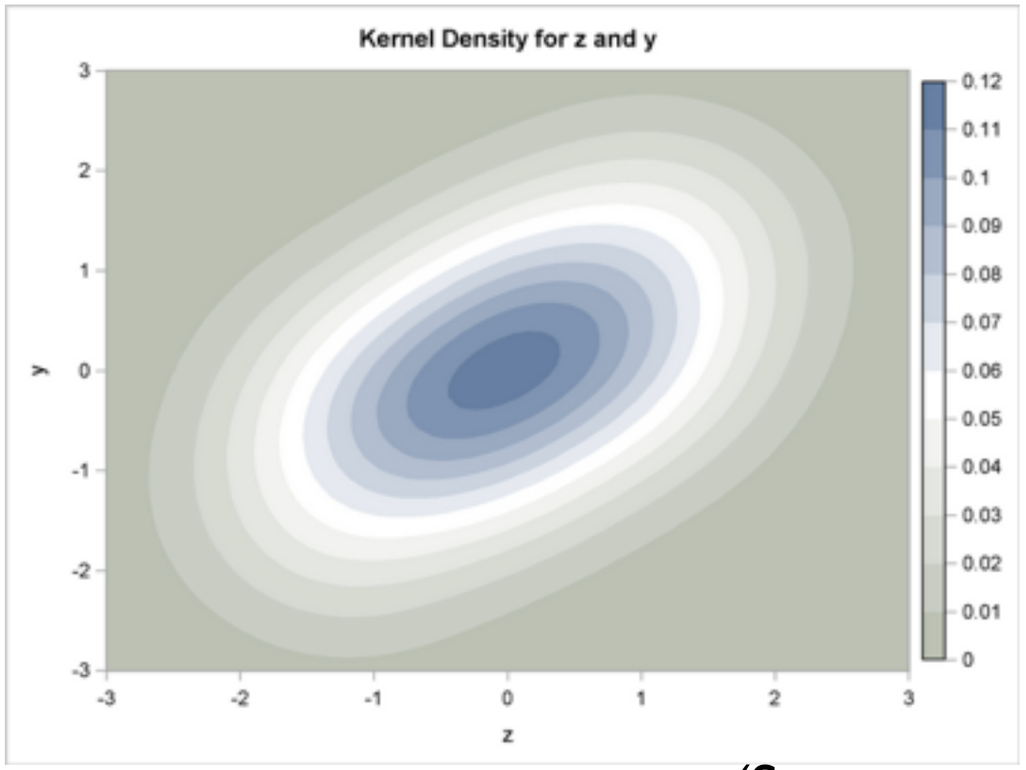
log A / log B < 1

Bivariate Gaussian



(Source: kenyon.edu)

in 3D



(Source: sas.com)

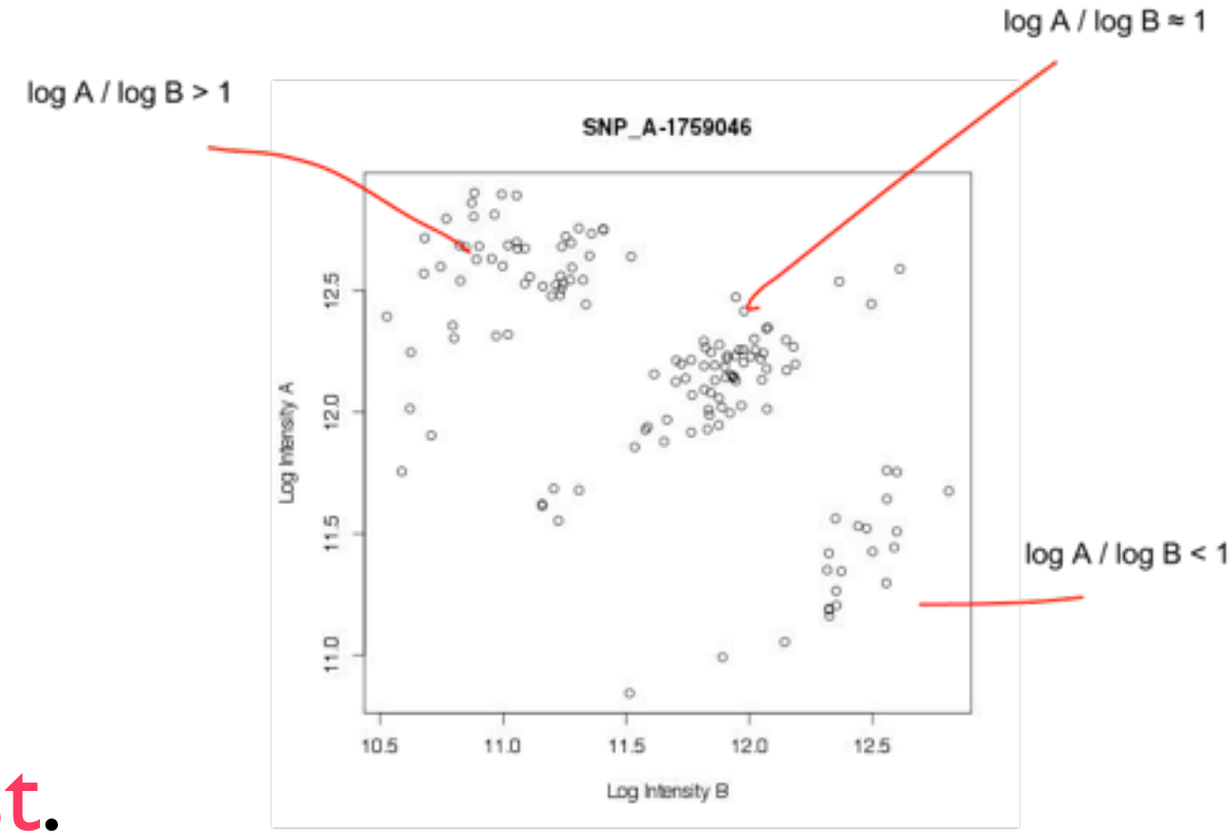
in 2D (level curves)

Baseline Method

In different SNPs, plot looks similar, but ratios may be different.

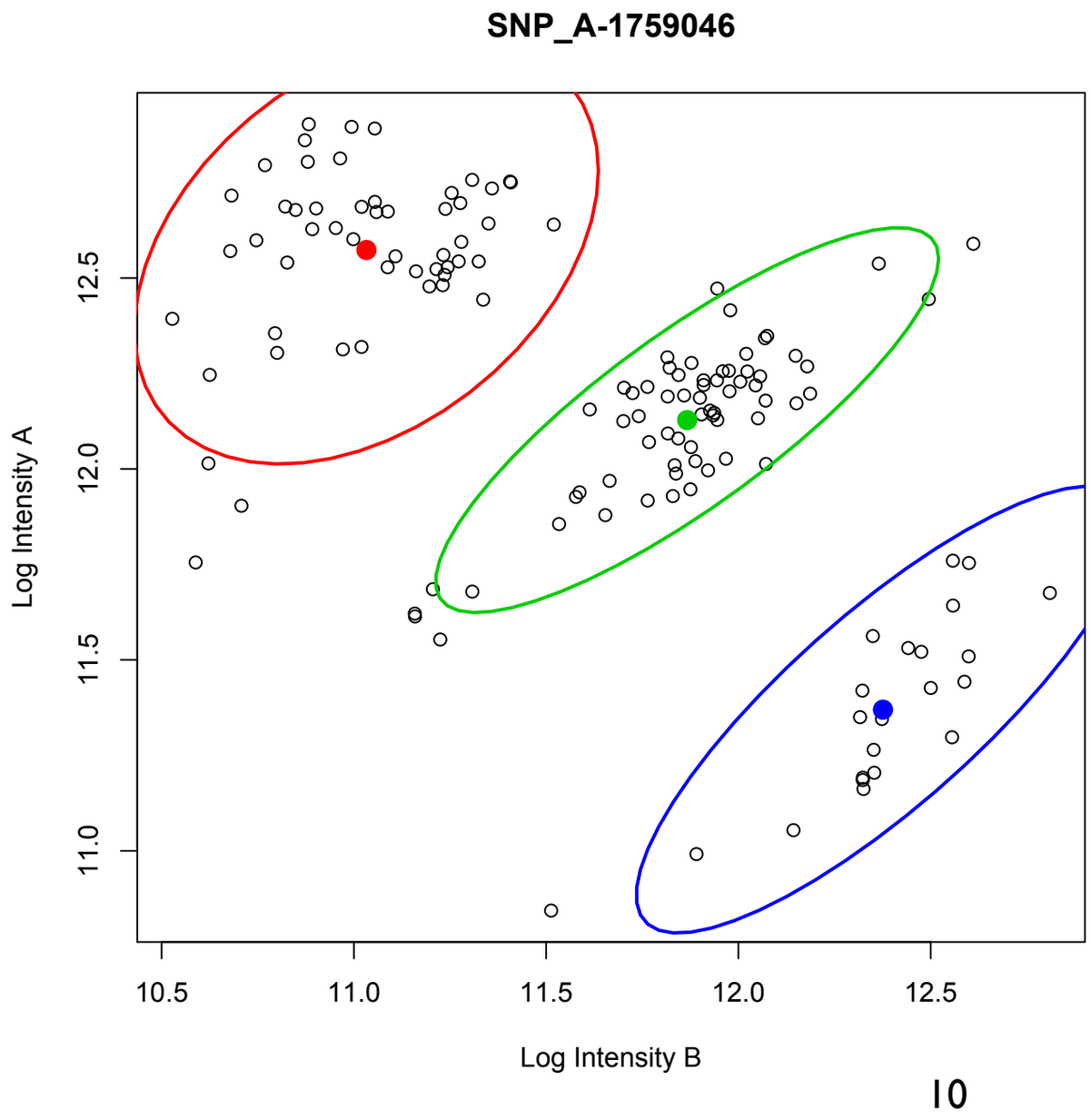
Ratio-method is not robust.

Terrible for borderline points.



My Method:

Gaussian Mixture Model



Data is generated around a mean.

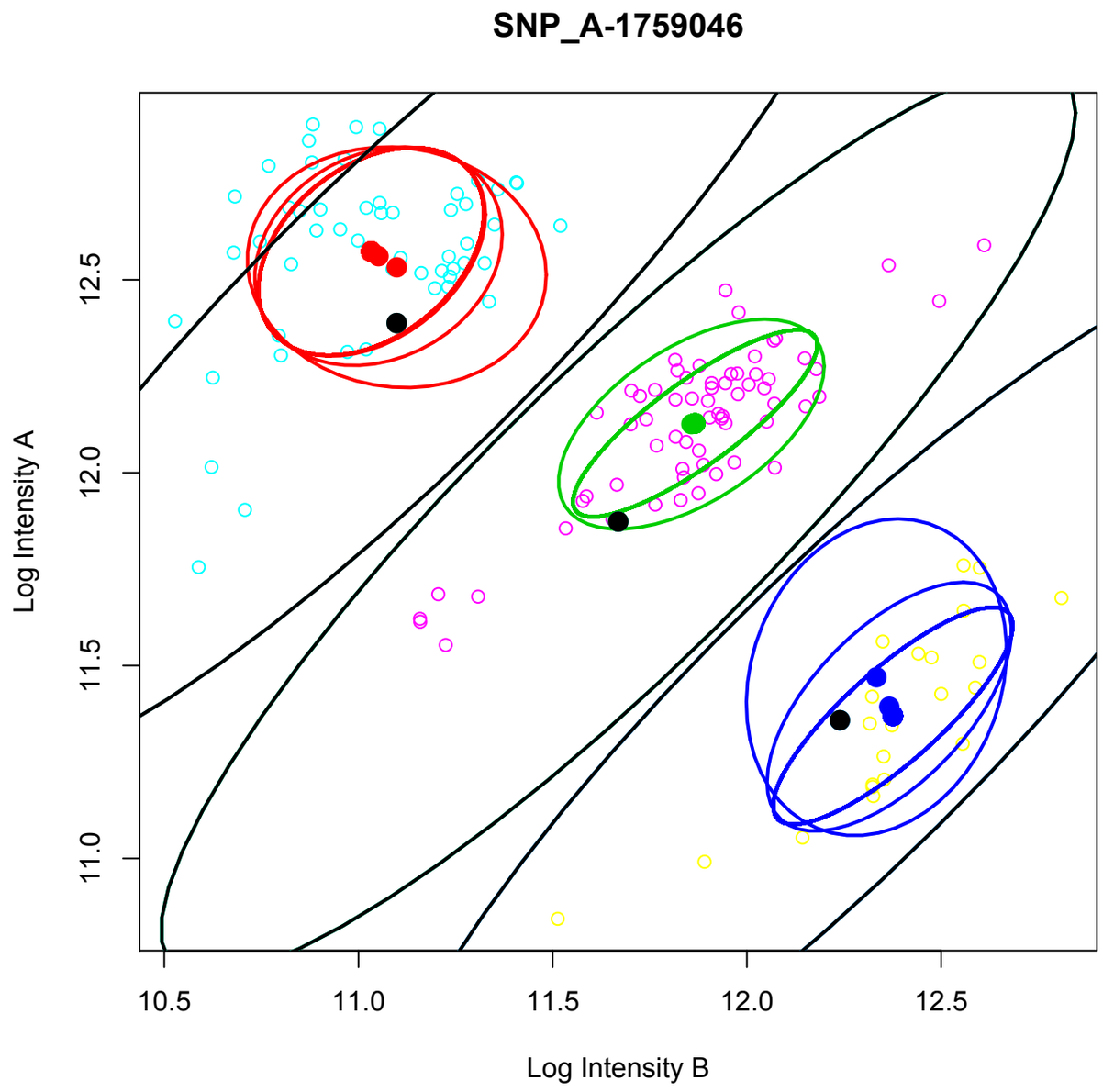
Generated at a range around mean (variance).

Gaussian Mixture Model

Randomly guess initial mean.

Iteratively move the bivariate gaussian curve.

Run EM algorithm until convergence.



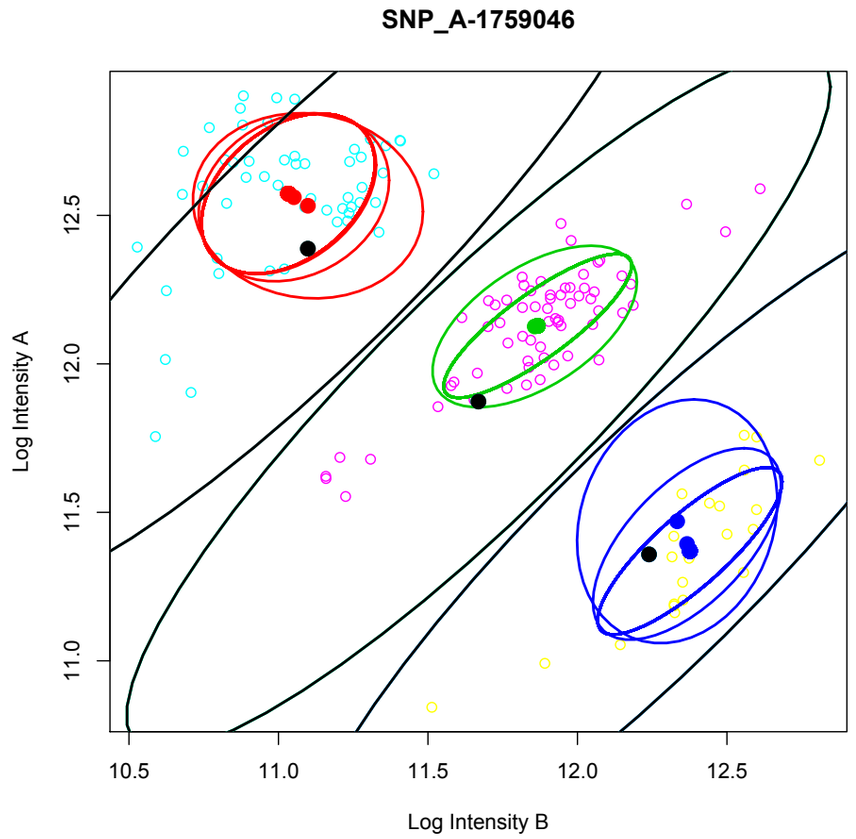
Expectation Maximization Algorithm

Expectation Step

Choose what cluster each point belongs to.

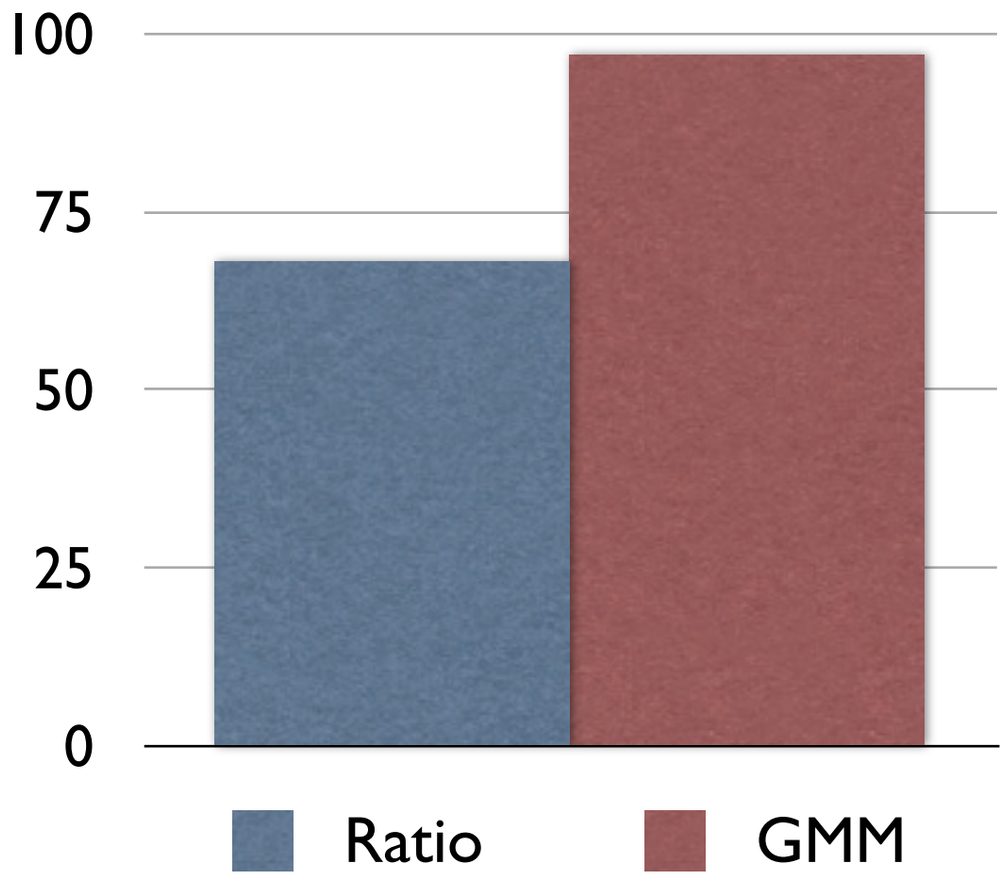
Maximization Step

Calculate the new mean and variance for each cluster.



Results

Accuracy of Methods



GMM outperforms baseline method.

Results run on data obtained from HapMap.

Over 1000 SNPs

Future Work

Automatically detecting number of clusters.

Can automatically detect copy number variation.

Use non-parametric bayesian methods such as the chinese restaurant process.