Distances and clustering

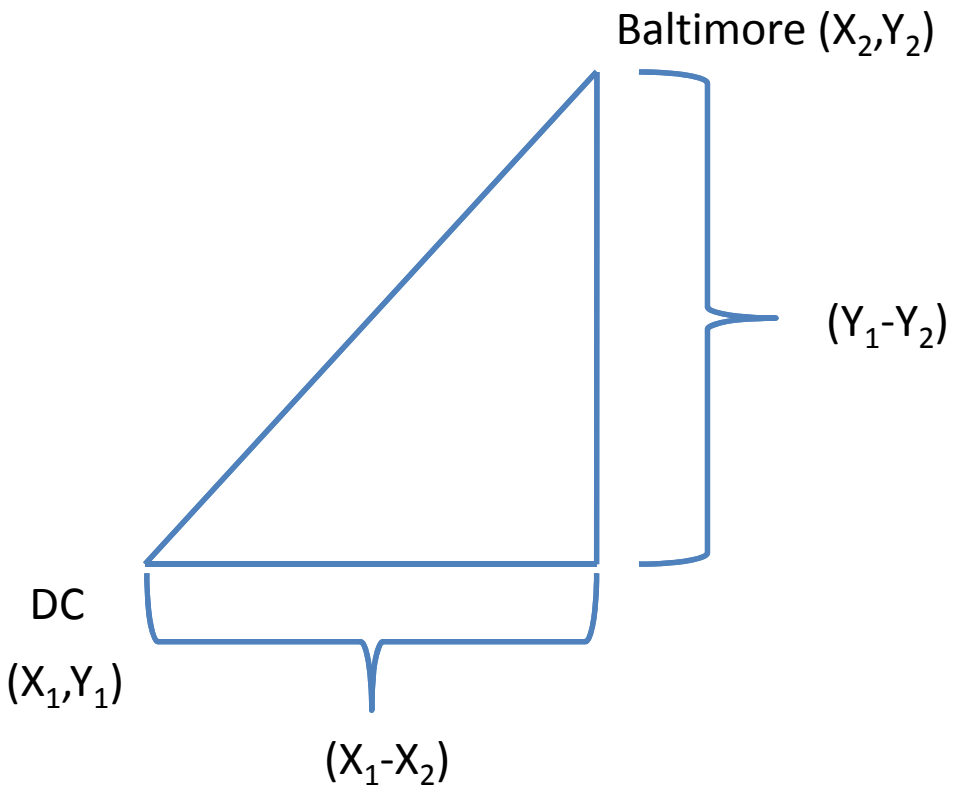
Illustrative example: genotyping arrays

Clustering

Distance

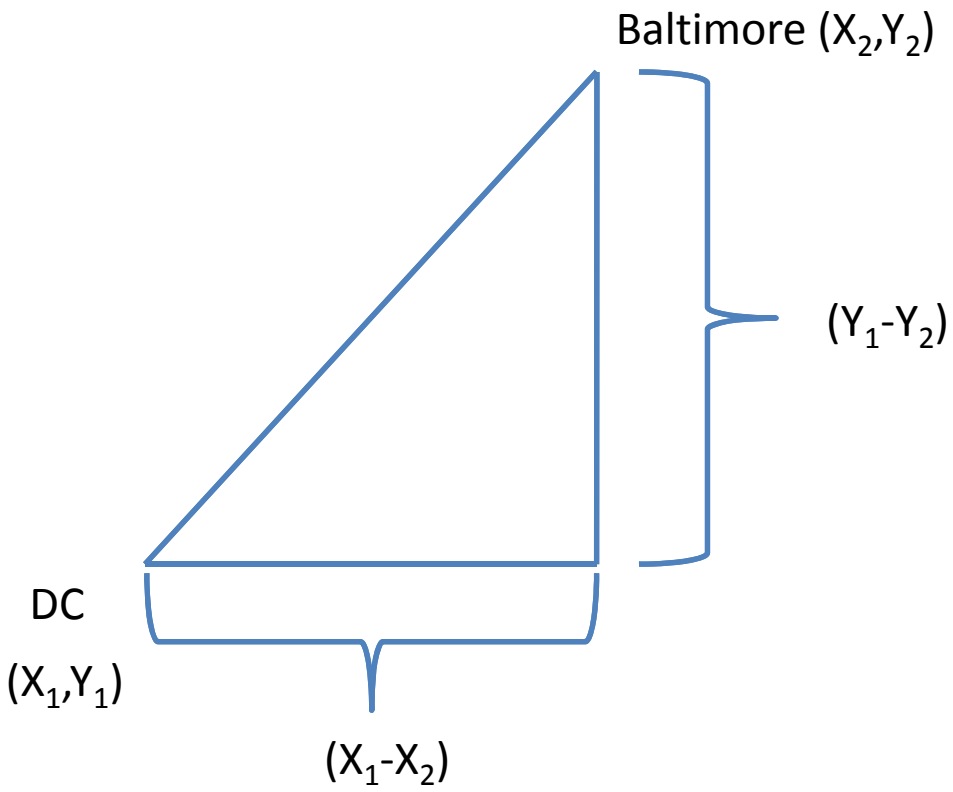
- Clustering organizes things that are *close* into groups
- What does it mean for two genes to be close?
- What does it mean for two samples to be close?
- Once we know this, how do we define groups?

Distance



Distance=

$$\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$$

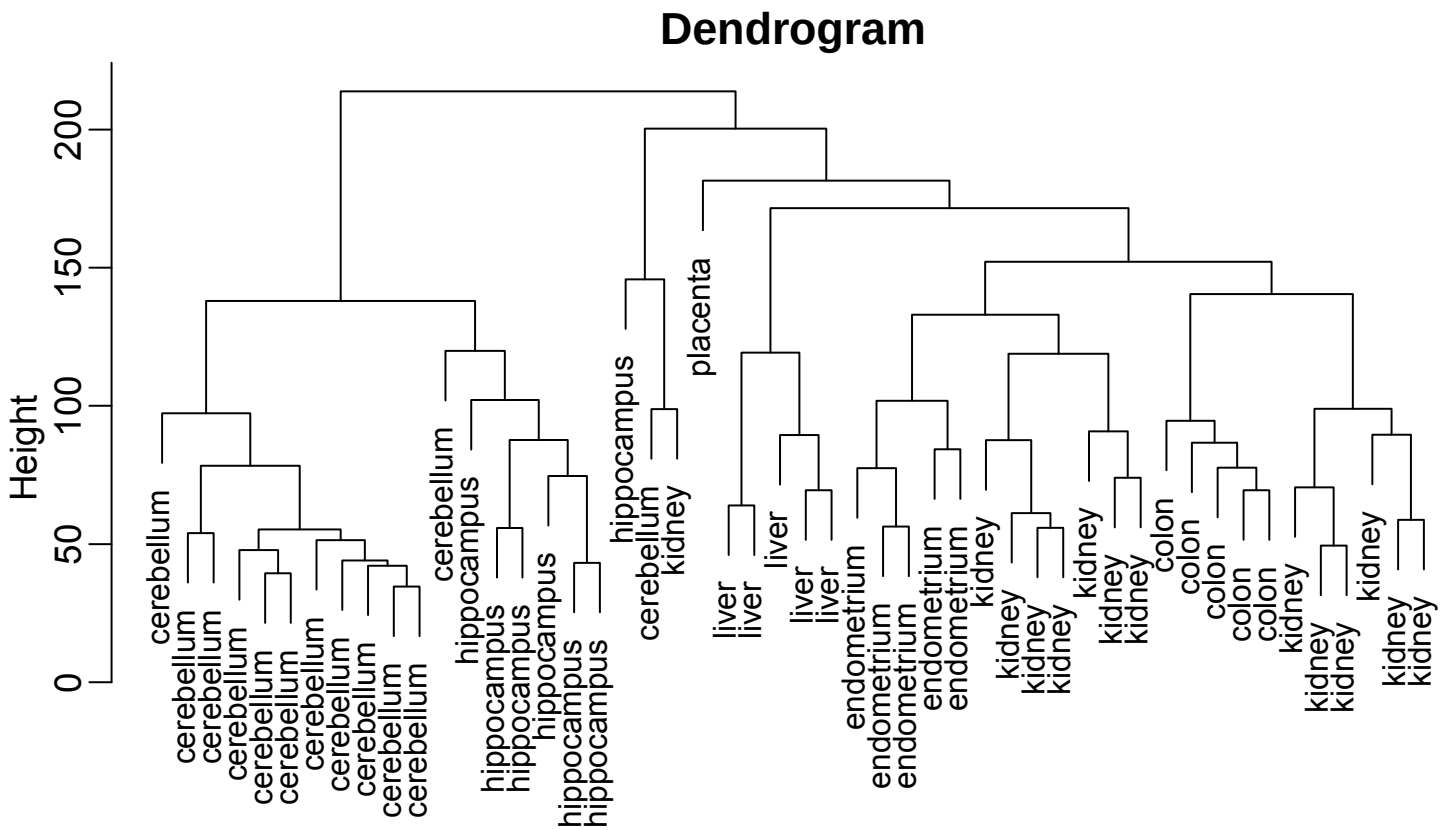


In general

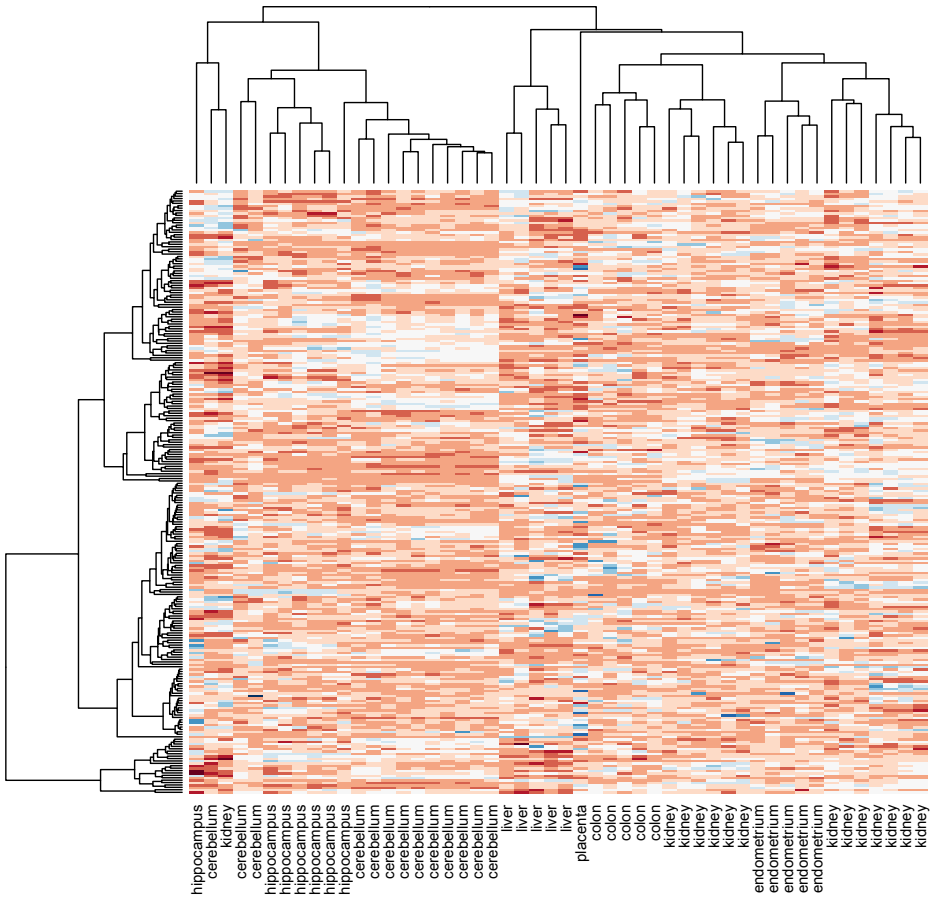
$$\sqrt{(A_1 - A_2)^2 + (B_1 - B_2)^2 + \square \ + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2}$$

We can’t draw in 26 dimensions

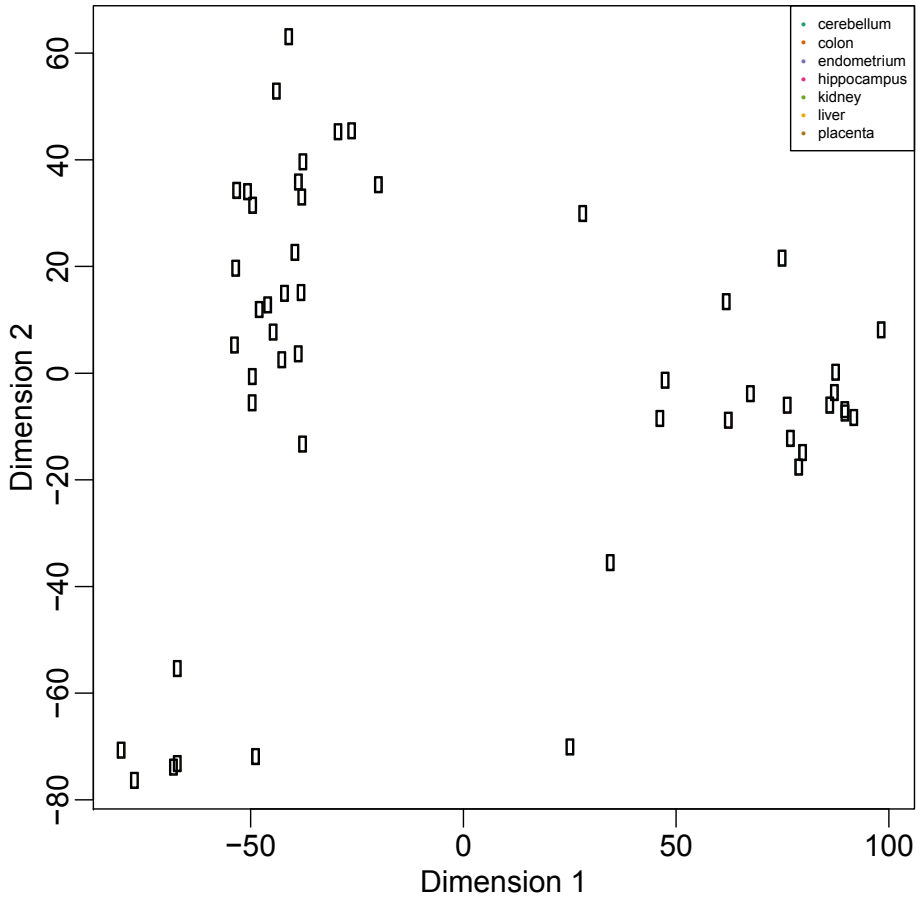
Dendrogram


Dendrogram
Height
0
50
100
150
200
cerebellum
cerebellum
cerebellum
cerebellum
cerebellum
cerebellum
cerebellum
cerebellum
cerebellum
cerebellum
hippocampus
hippocampus
hippocampus
hippocampus
hippocampus
hippocampus
cerebellum
kidney
placenta
liver
liver
liver
liver
endometrium
endometrium
endometrium
endometrium
kidney
kidney
kidney
kidney
kidney
kidney
kidney
colon
colon
colon
colon
colon
kidney
kidney
kidney
kidney
kidney
kidney

Heatmap


hippocampus
cerebellum
kidney
cerebellum
cerebellum
hippocampus
hippocampus
hippocampus
hippocampus
hippocampus
hippocampus
cerebellum
cerebellum
cerebellum
cerebellum
cerebellum
cerebellum
cerebellum
cerebellum
liver
liver
liver
liver
placenta
colon
colon
colon
colon
colon
kidney
kidney
kidney
kidney
kidney
kidney
endometrium
endometrium
endometrium
endometrium
endometrium
kidney
kidney
kidney
kidney
kidney
kidney

Multidimensional Scaling



Using clustering to genotype

Human Variation

Single Nucleotide Polymorphism (SNP)

Genomic DNA:

SNP

A

TACATAGCCATCGGTANGTACTCAATGATGATA

G

Three genotypes

AA

Mom

TACATAGCCATCGGTAAGTACTCAATGATGATA
ATGTATCGGTAGCCATTCATGAGTTACTACTAT

Dad

TACATAGCCATCGGTAAGTACTCAATGATGATA
ATGTATCGGTAGCCATTCATGAGTTACTACTAT

AG

Mom

TACATAGCCATCGGTAAGTACTCAATGATGATA
ATGTATCGGTAGCCATTCATGAGTTACTACTAT

Dad

TACATAGCCATCGGTAGGTACTCAATGATGATA
ATGTATCGGTAGCCATCCATGAGTTACTACTAT

GG

Mom

TACATAGCCATCGGTAGGTACTCAATGATGATA
ATGTATCGGTAGCCATCCATGAGTTACTACTAT

Dad

TACATAGCCATCGGTAGGTACTCAATGATGATA
ATGTATCGGTAGCCATCCATGAGTTACTACTAT

Affymetrix SNP chip terminology

SNP

Genomic DNA: TACATAGCCATCGGTANGTACTCAATGATGATA

A

G

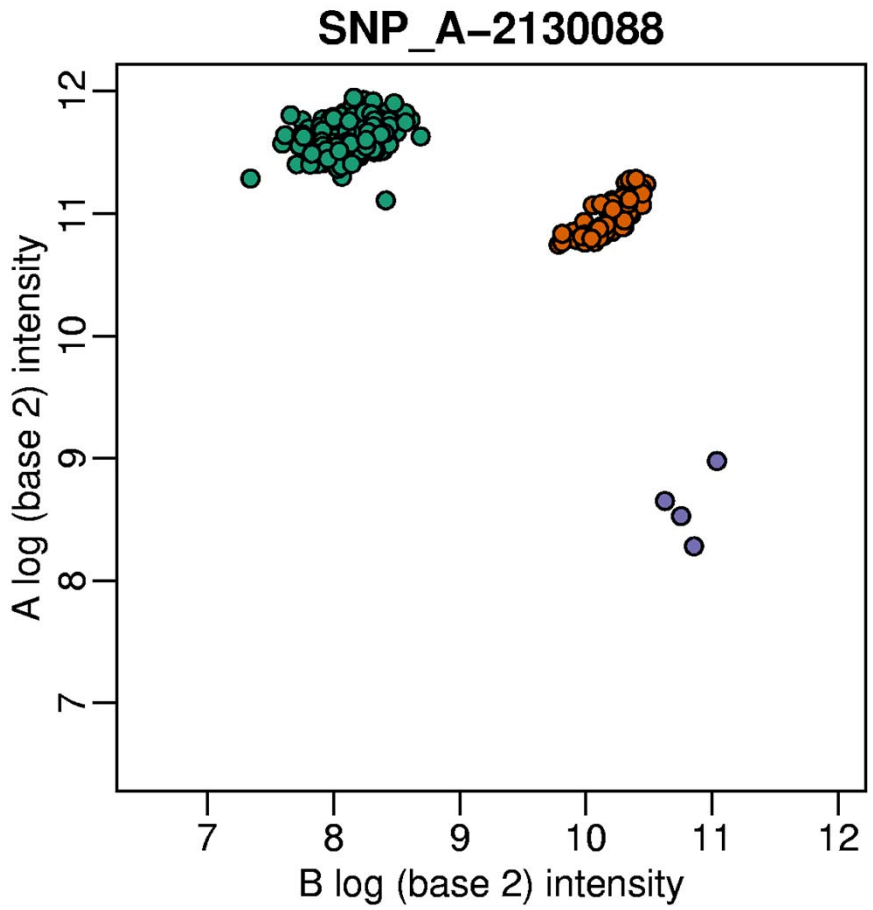
PM probe for Allele A: ATCGGTAGCCATTCATGAGTTACTA

PM probe for Allele B: ATCGGTAGCCATCCATGAGTTACTA

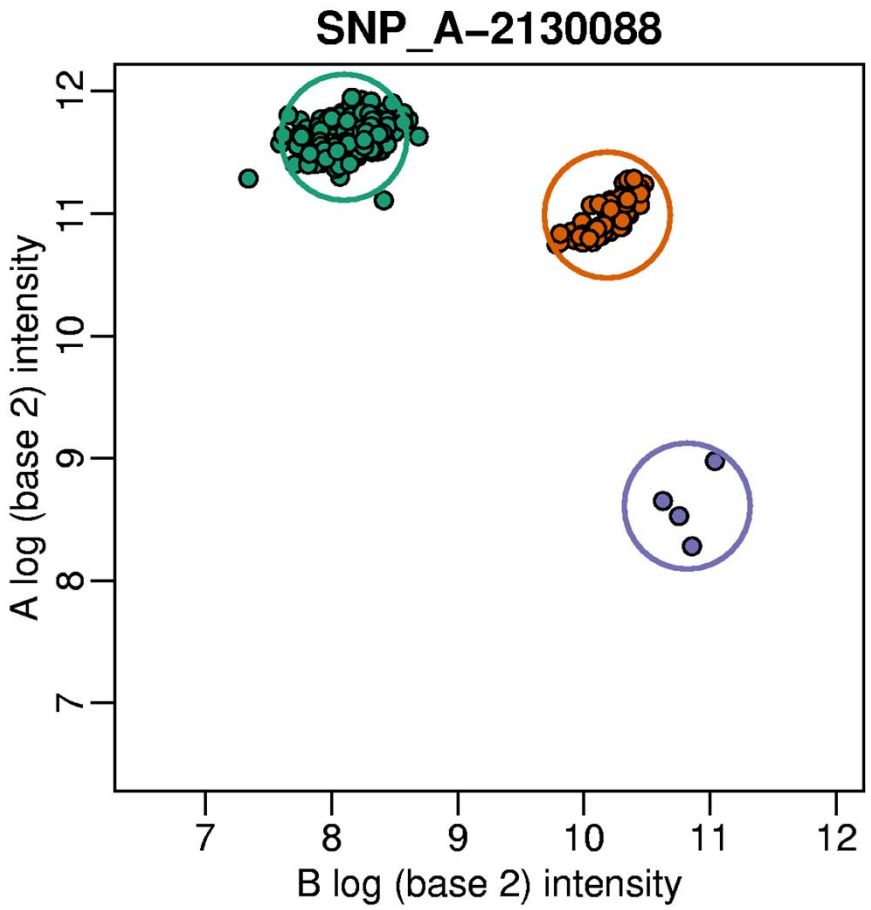
Genotyping: answering the question about the two copies of the chromosome on which the SNP is located:

Is a person AA , AG or GG at this Single Nucleotide Polymorphism?

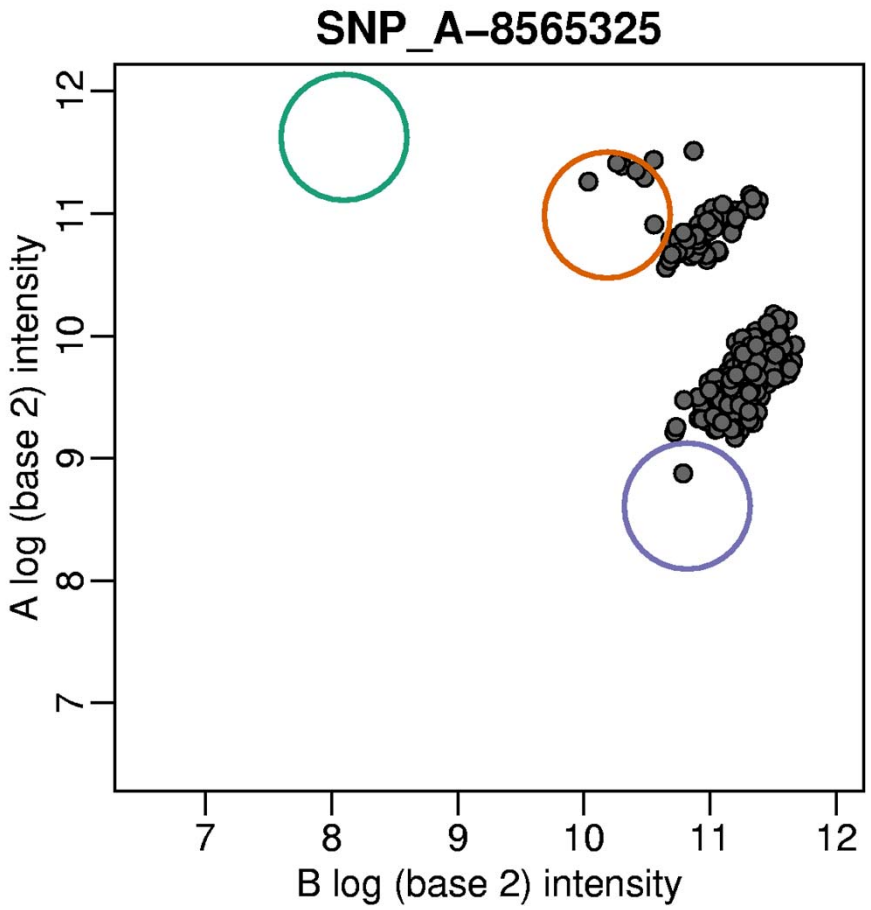
Infer genotype from data



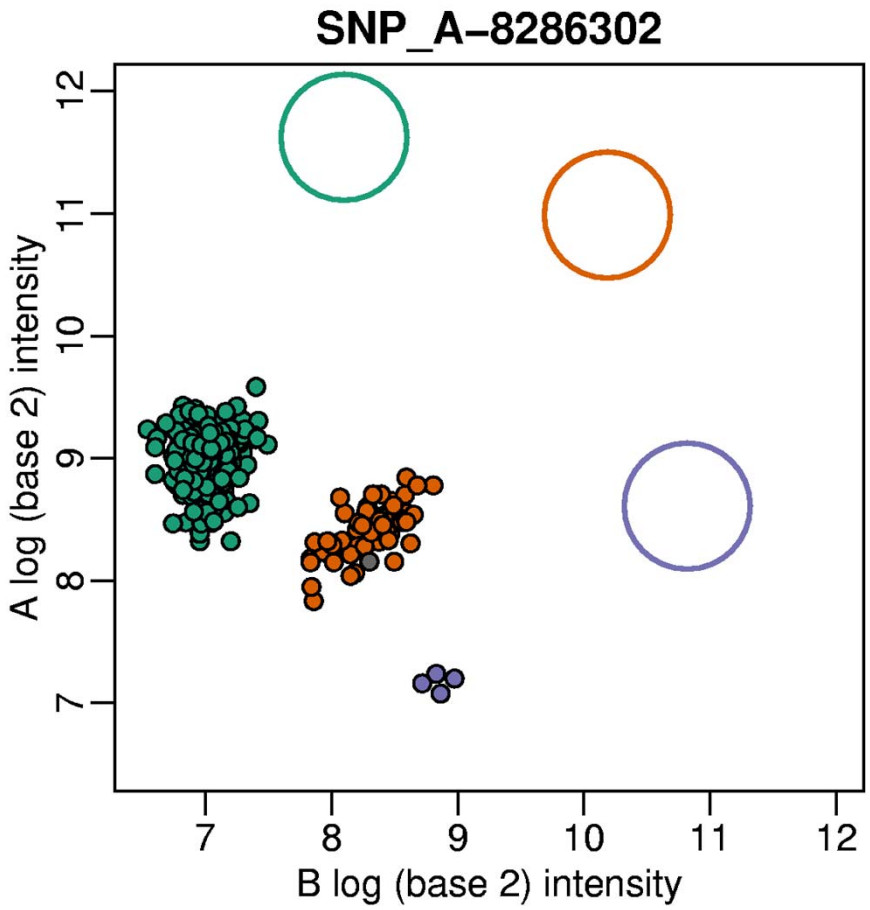
Using regions



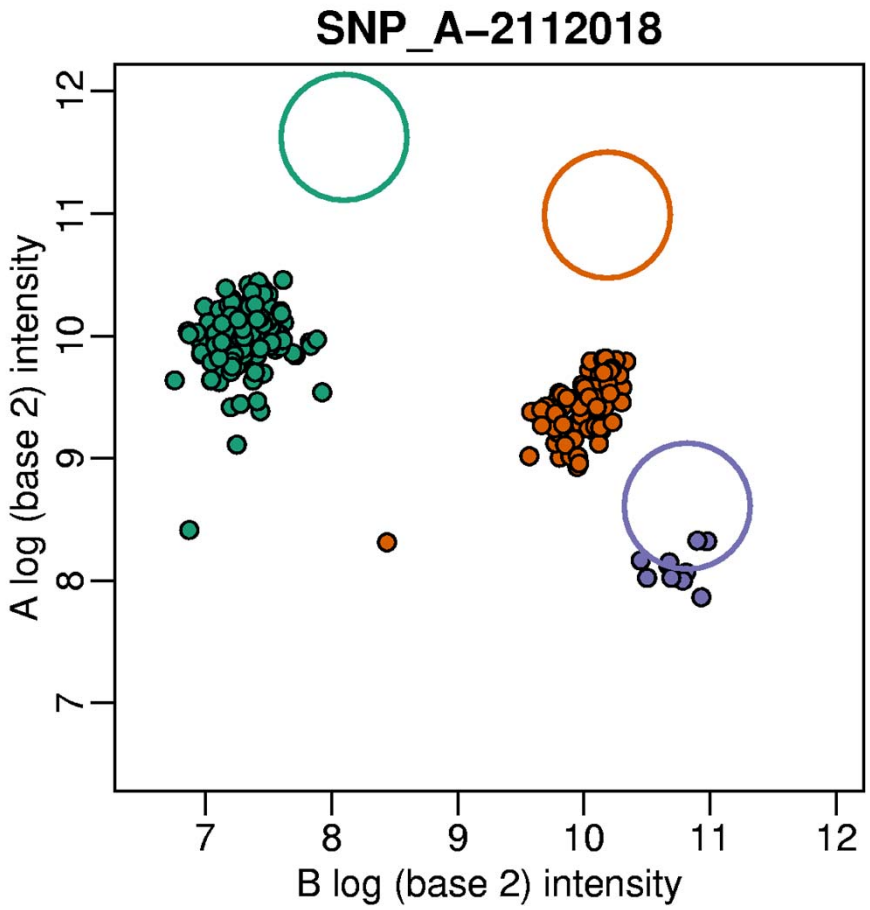
Probe effect



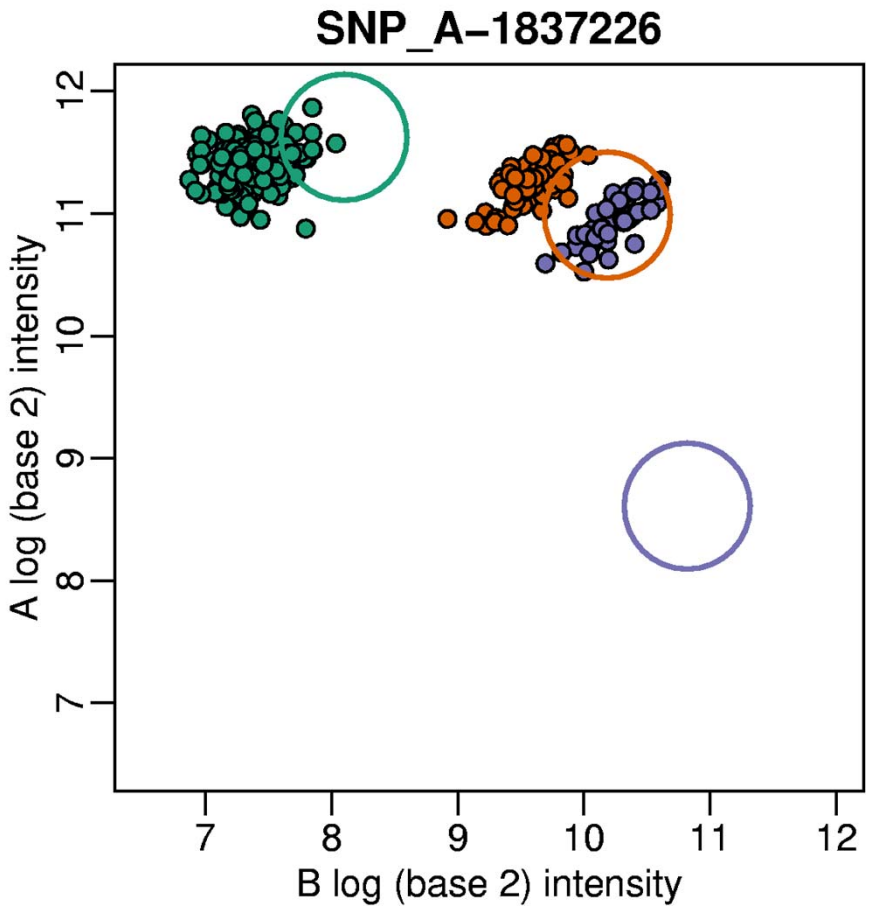
Probe effect



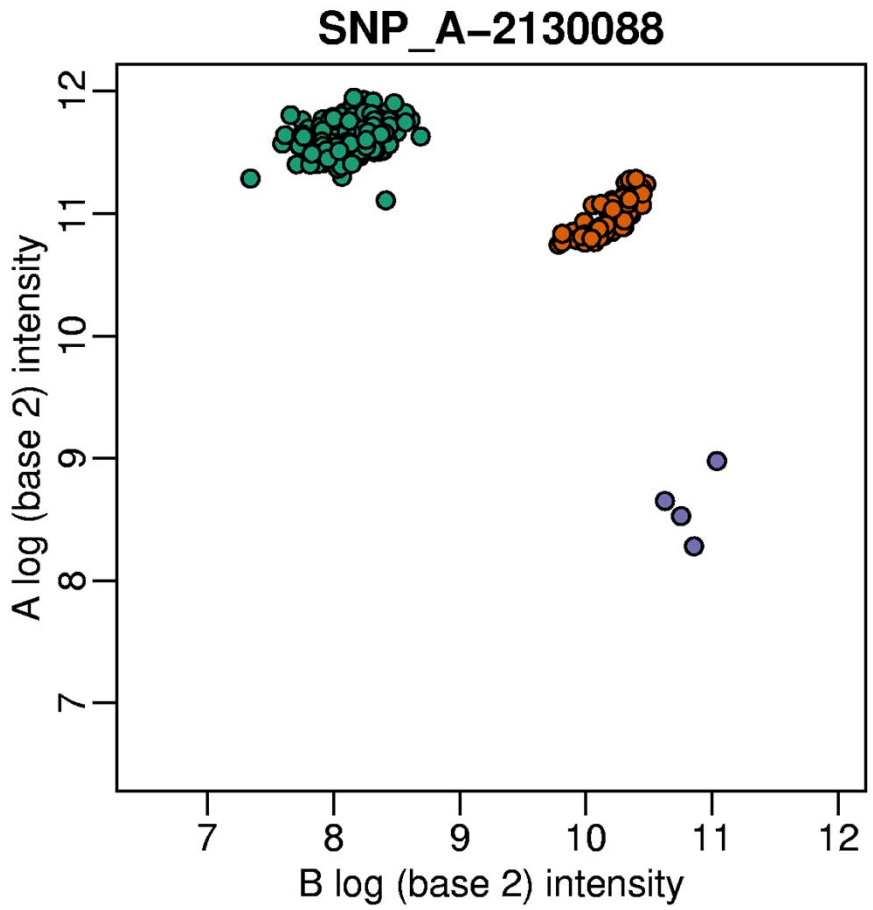
Probe effect



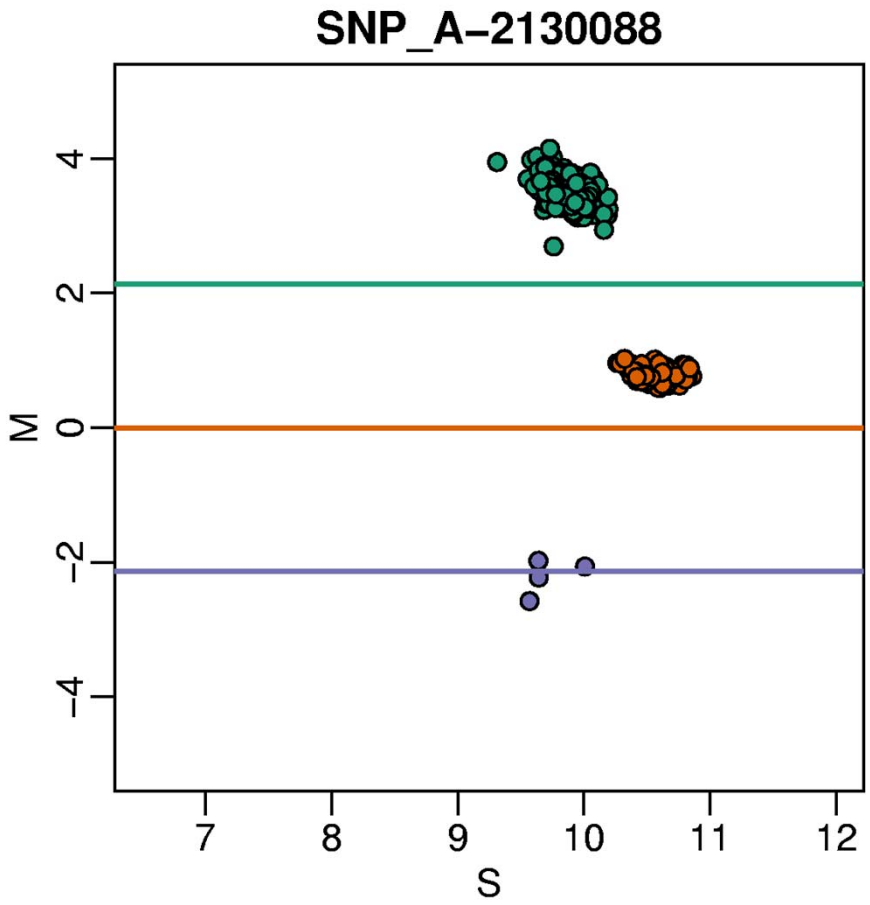
Probe effect



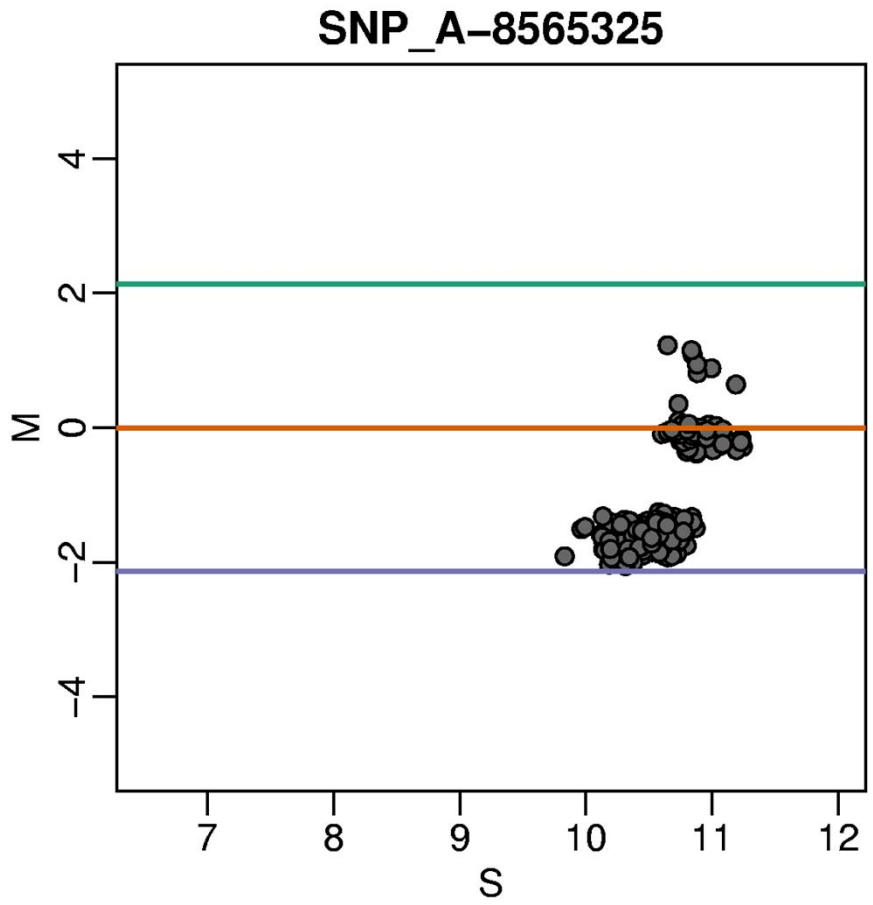
Most information is in *M*



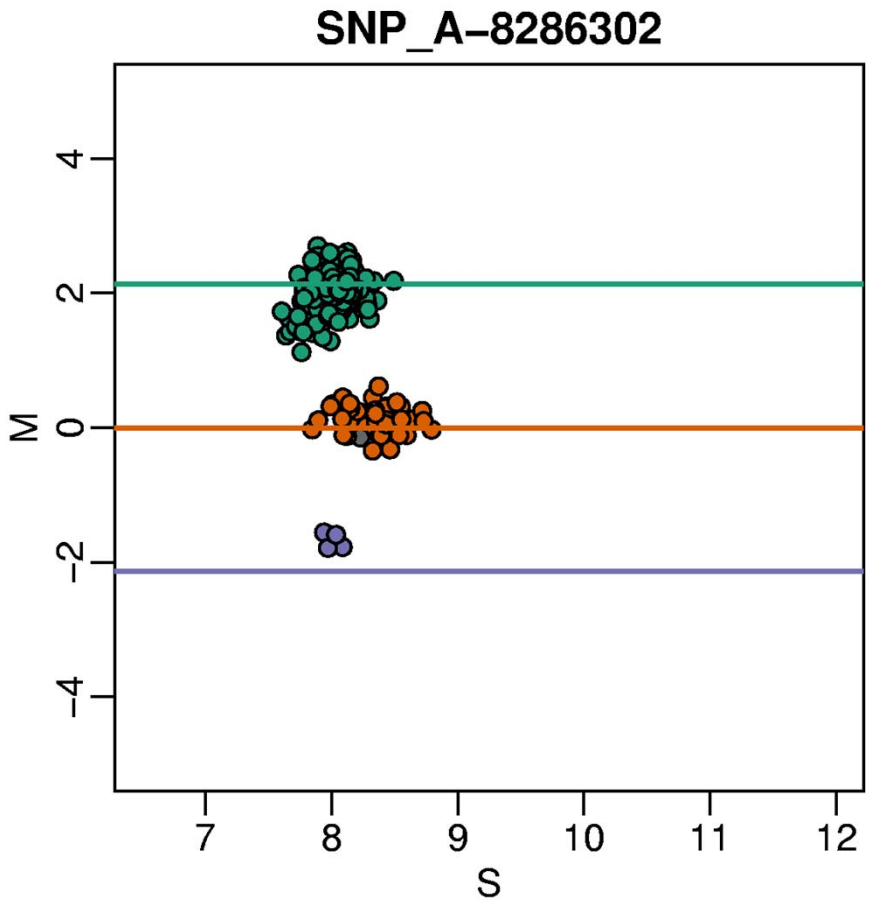
Most information is in *M*



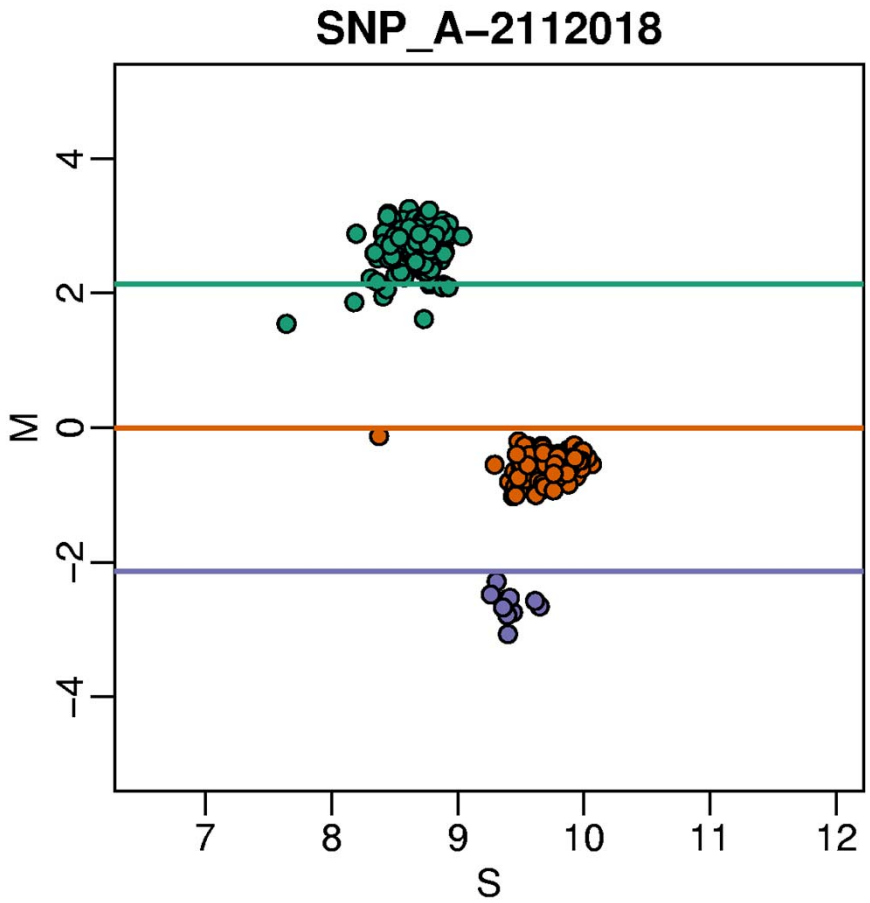
M more stable than *S*



M more stable than *S*

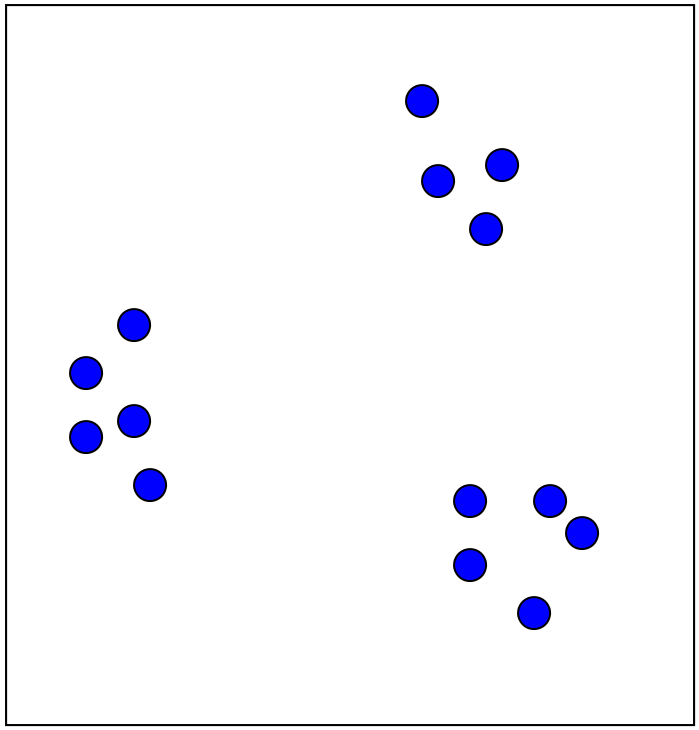


M more stable than *S*



K-means

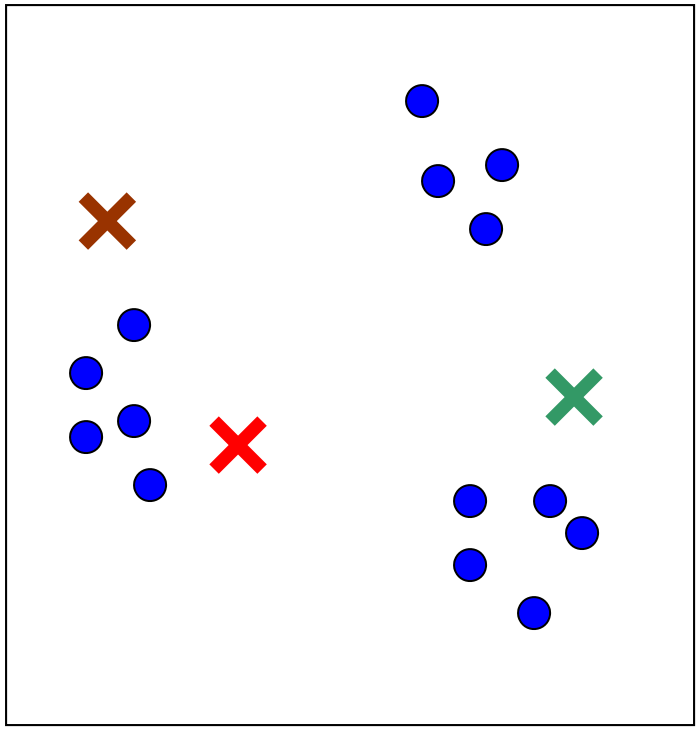
- We start with some data
- Interpretation:
  - We are showing expression for two samples for 14 genes
  - We are showing expression for two genes for 14 samples
- This is simplifaction



Iteration = 0

K-means

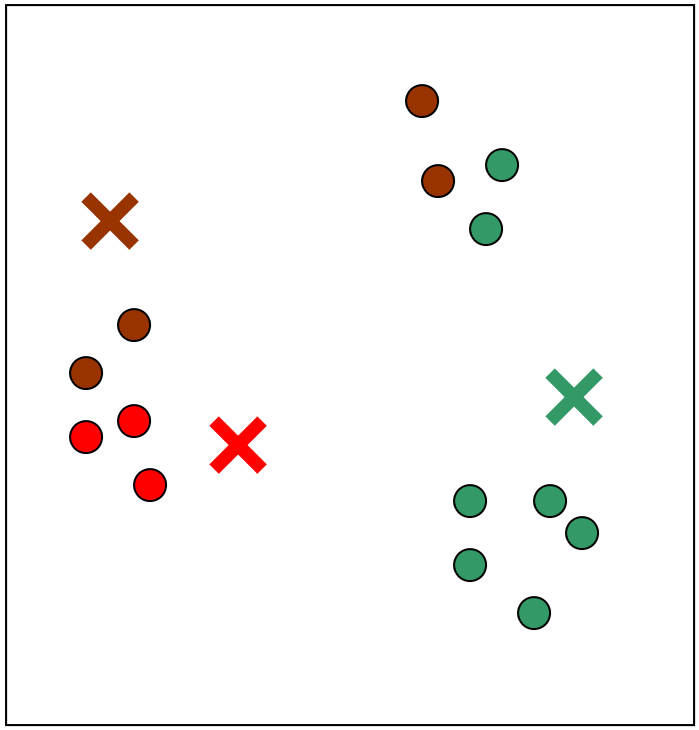
- Choose K *centroids*
- These are starting values that the user picks.
- There are some data driven ways to do it



Iteration = 0

K-means

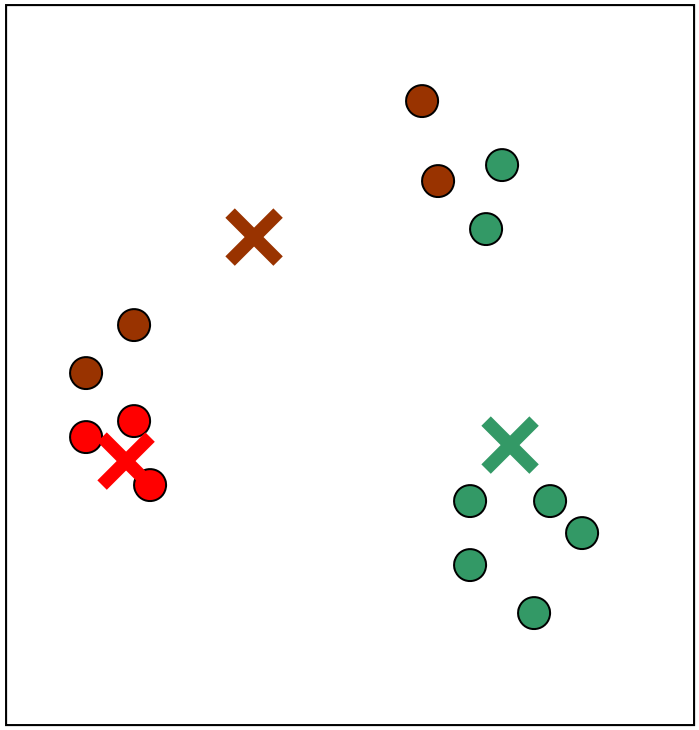
- Make first *partition* by finding the closest centroid for each point
- This is where distance is used



Iteration = 1

K-means

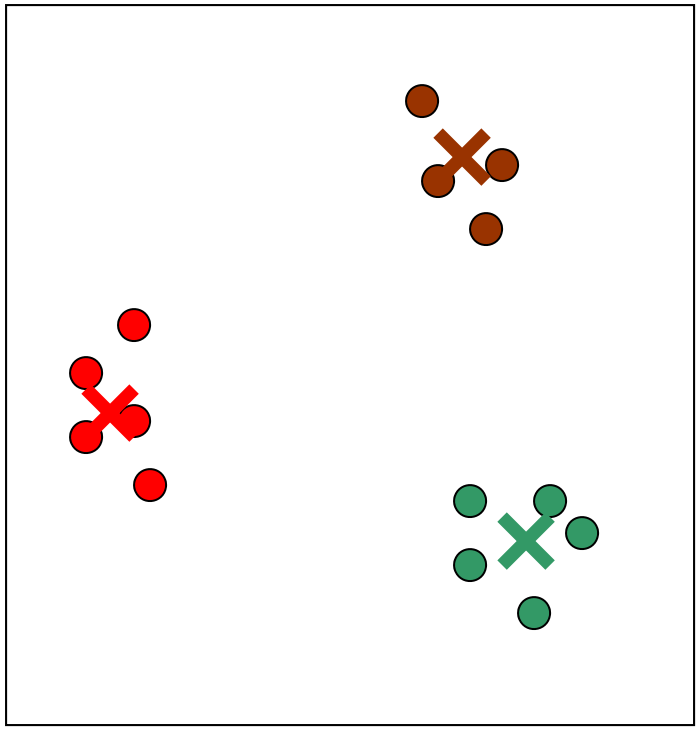
- Now re-compute the centroids by taking the *middle* of each cluster



Iteration = 2

K-means

- Repeat until the centroids stop moving or until you get tired of waiting



Iteration = 3

K-means Limitations

- Final results depend on starting values
- How do we chose K? There are methods but not much theory saying what is best.
- Where are the pretty heatmpas and dendrograms?

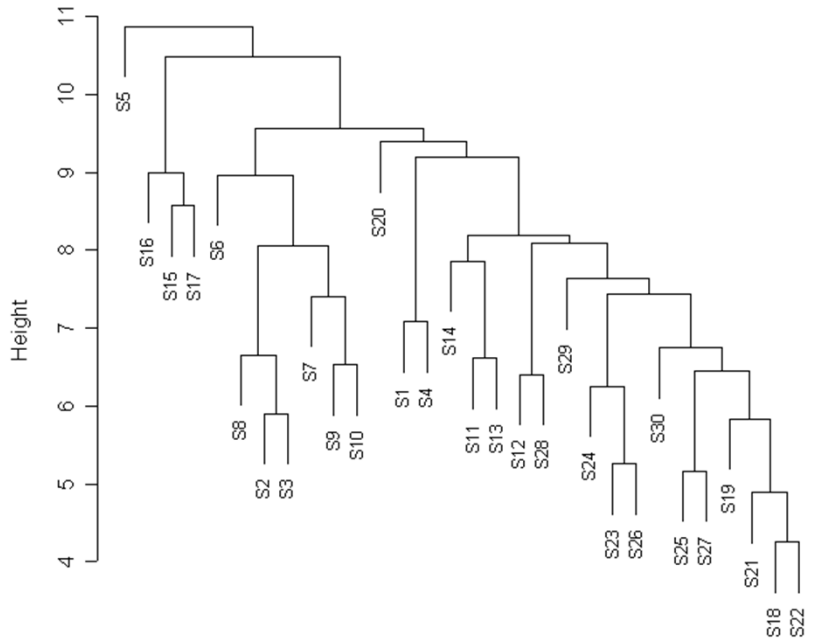
Hierarchical

- Divide all points into 2. Then divide each group into 2. Keep going until you have groups of 1 and can not divide further.

- This is divisive or top-down hierarchical clustering. There is also agglomerative clustering or bottom-up

Dendrograms

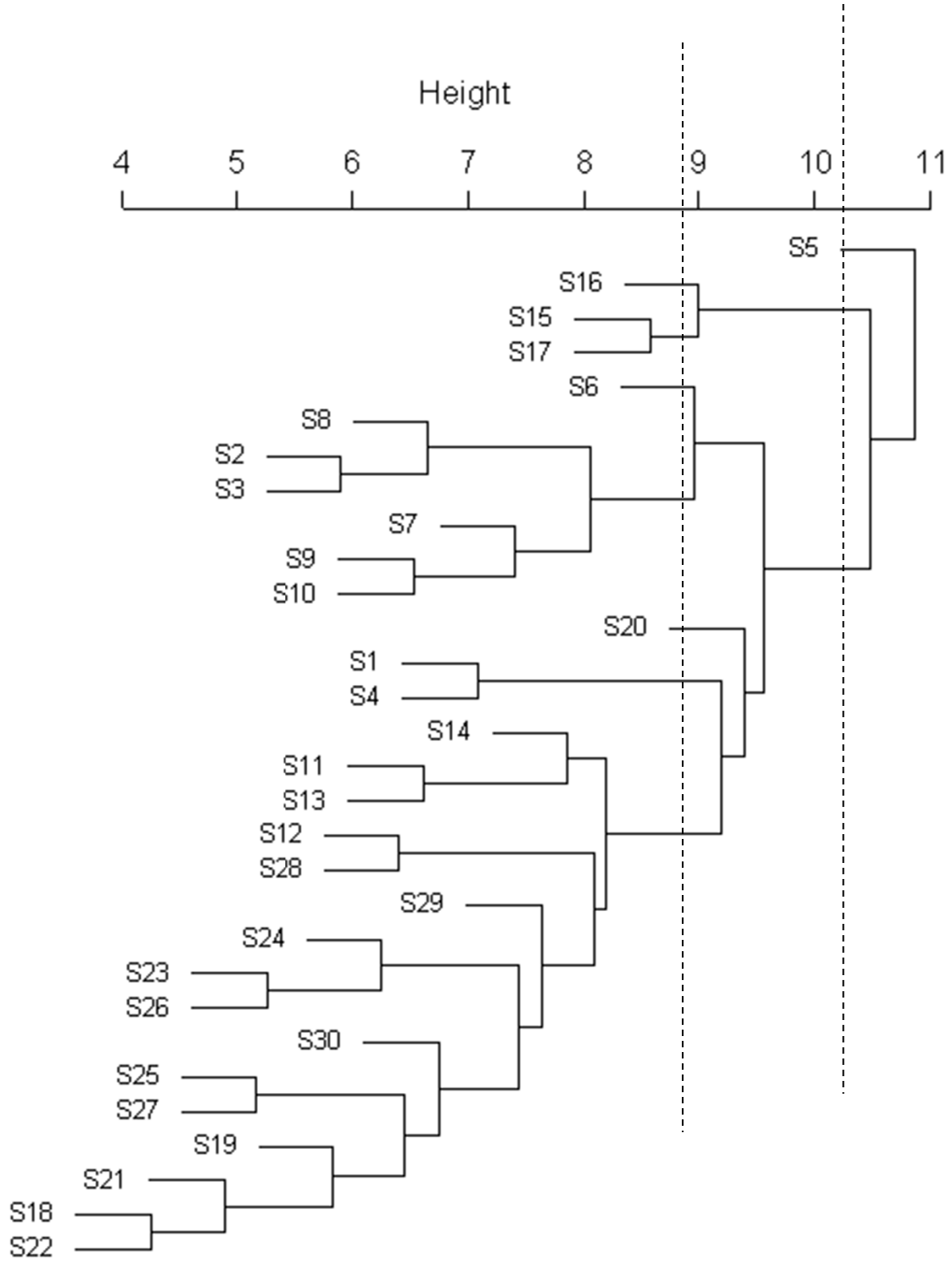
- We can then make dendrograms showing divisions
- The y-axis represents the distance between the groups divided at that point



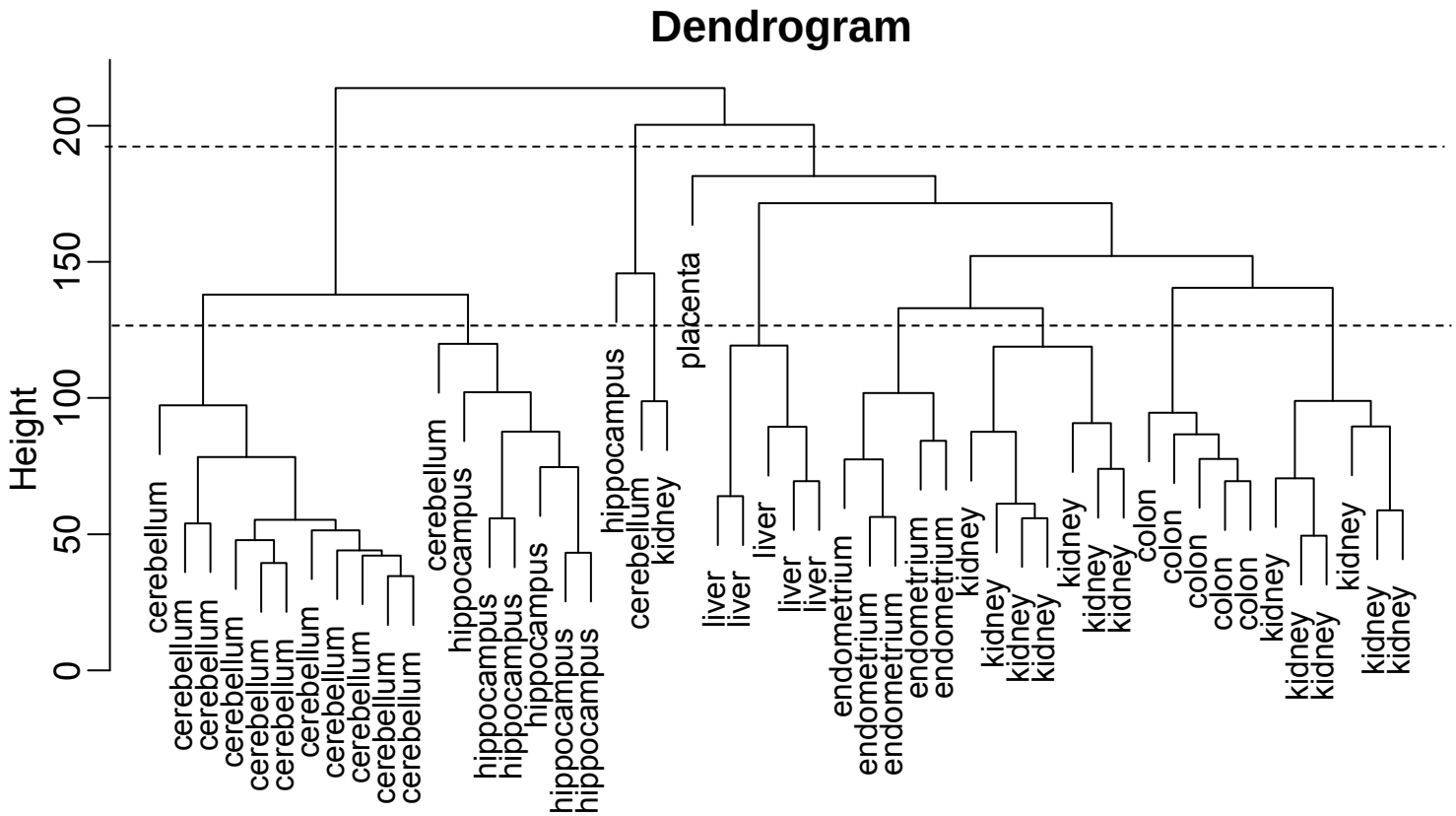
Note: Left and right is assigned arbitrarily.
Look at the height of division to find out distance.
For example, S5 and S16 are very far.

But how do we form actual clusters?

We need to pick a height

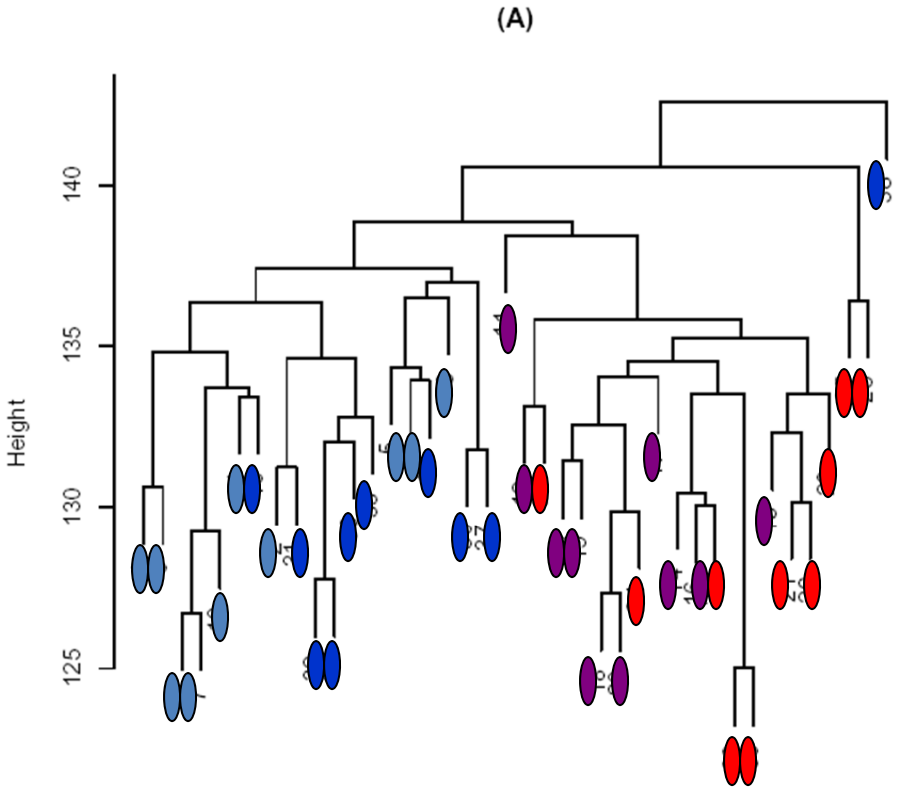

Height
4
5
6
7
8
9
10
11
S5
S16
S15
S17
S6
S8
S2
S3
S7
S9
S10
S20
S1
S4
S14
S11
S13
S12
S28
S29
S24
S23
S26
S30
S25
S27
S19
S21
S18
S22

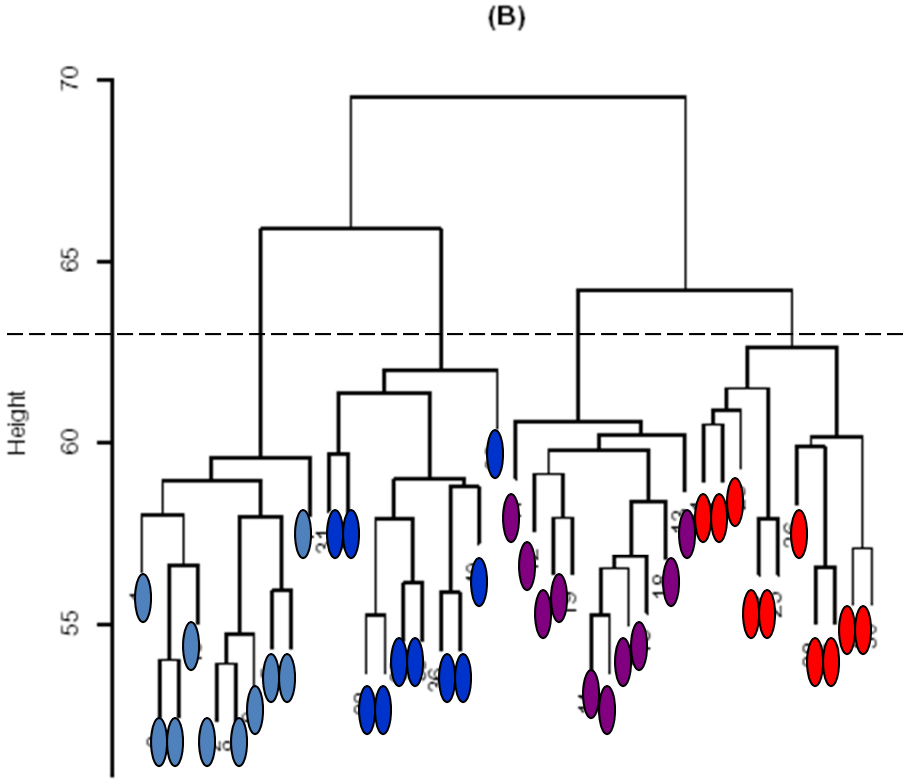
Dendrogram


Dendrogram
Height
0
50
100
150
200
cerebellum
cerebellum
cerebellum
cerebellum
cerebellum
cerebellum
cerebellum
cerebellum
cerebellum
cerebellum
hippocampus
hippocampus
hippocampus
hippocampus
hippocampus
hippocampus
cerebellum
kidney
placenta
liver
liver
liver
liver
endometrium
endometrium
endometrium
endometrium
kidney
kidney
kidney
kidney
kidney
kidney
kidney
colon
colon
colon
colon
colon
kidney
kidney
kidney
kidney
kidney
kidney

Note: distances susceptible to noise

Simulated Data with 4 clusters: 1-10, 11-20, 21-30, 31-40





A: 450 relevant genes plus 450 “noise” genes.

B: 450 relevant genes.