1. Find a research article for your presentation and final exam
2. Prepare your presentation
3. No class for the week of Midterm exam
4. http://sysbio.unl.edu/Teaching/BIOS497897_2014/

Transcriptome

Lecture 2

Outline

- Multiple Testing Procedures
- Data Visualization, Distance Measures
- Clustering
- Gene Annotation and Enrichment Analysis

The problem

- After differential expression testing (from RNA-seq or Microarray assay), a list of P-value is obtained, one for each gene.

- Most investigators want to
  - Identify the genes that are differentially expressed
  - Estimate the proportion of errors in the list of selected "differentially expressed genes"

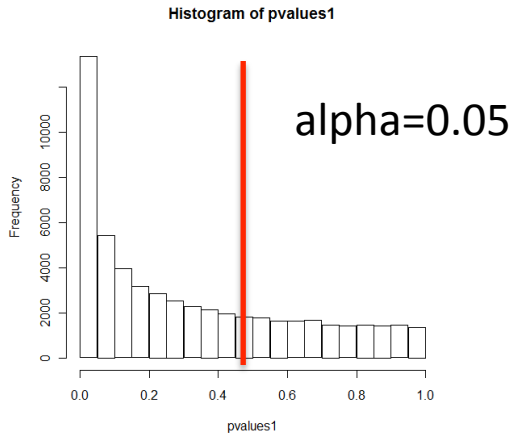
A naïve solution

- Since genes with small p-values are likely to be differentially expressed, why don't we just use the traditional (pre-specified) alpha = 0.05 to decide?
  - ❑ Yes?
  - ☒ No? But why?

What does this mean to microarray data?

- The result is that we obtain one p-value for each gene

	T 1	T 2	T 3	N 1	N 2	N 3	T-statistics	P-value
G 1							T1	P1
G 2							T2	p2
…							…	….
G 20000							T20000	P20000



- 20,000 p-values…
- If we use alpha=0.05 to decide differentially expressed genes, 5% of the 20,000 genes would then be selected by chance
- That means 1000 genes would be false positives…

A naïve solution

- Since genes with small p-values are likely to be differentially expressed, why don't we just use the traditional (pre-specified) $\alpha = 0.05$ to decide?
  - ❑Yes?
  - ❑No! 20,000x0.05 = 1000 false positives!
    - If the investigator is interested in selecting 100 genes for downstream analysis, they could all be false positives by chance!
  - ❑ Other solutions?

The solutions

- To select differentially expressed genes, we need to do multiple testing (multiplicity) corrections
  - Familywise Error Rate (FWER), such as Bonferroni correction and Holm's method: adjust the p-value threshold from alpha to alpha/(number of genes)
  - Control False Discovery Rate: algorithm proposed by Benjamini & Hochberg
  - Re-sampling techniques (i.e., Permutation P-values)

Familywise Error Rate (FWER)

- Traditionally statisticians have focused on controlling FWER when conducting multiple tests.
- FWER is defined as the probability of one or more false positive results:

$$FWER = P(V > 0).$$

- Controlling FWER amounts to choosing the significance cutoff c so that FWER is less than or equal to some desired level α.

The Bonferroni Method

- The Bonferroni Method is the simplest way to achieve control of the FWER at any desired level α.
- Simply choose $c = \alpha / m$.
- With this value of *c* for each individual test, the FWER will be no larger than α for any family of *m* tests.

Bonferroni correction

	T 1	T 2	T 3	N 1	N 2	N 3	T-statistics	P-value
G 1	y1	y2	y3	y4	y5	y6	T1	0.012
G 2	y1	y2	y3	y4	y5	y6	T2	0.045
...								
G 20000								

- using $\alpha = 0.05$ we reject the null hypothesis that the expression of gene 1 (2) is not changed in tumor versus normal tissue.
- In the other words, gene 1 (2) is differentially expressed genes between tumor and normal tissues.

Bonferroni correction

- However, the probability that either the expression difference observed for gene 1 (p=0.012) or the expression difference observed for gene 2 (p=0.045) under null hypothesis is 0.012+0.045 = 0.057 (>0.05!).

- Using an overall p-value alpha = 0.05, we have no evidence to reject the null hypothesis that the expression of either gene 1 or gene 2 has no change in tumor versus normal tissues.
  - Here overall p-value is the probability of making at least 1 mistake in the two performed tests.
  - Hence, the α=0.05 is not stringent enough for each test.

Bonferroni correction

- The Bonferroni rule
  - To guarantee that the probability of making at least 1 mistake in the two performed tests is not larger than alpha, we need to use for each test alpha/2 as significance level
  - To guarantee that he probability of making at least 1 mistake in the ten performed tests is not larger than alpha, we need to use for each test alpha/10 as significance level

Bonferroni correction

- For microarray, we need to, successively until the last gene, calculated the difference between group means, divided by the global standard error; obtain T20000 and P20000

	T 1	T 2	T 3	N 1	N 2	N 3	T-statistics	P-value
G 1							T1	0.012
G 2							T2	0.045
…							…	…
G 20000	y1	y2	y3	y4	y5	y6	T2000	P20000

$\overline{Y_T}$ $\overline{Y_N}$

S

- Result: 20,000 p-values need to be combined to give an overall conclusion of how many genes are differentially expressed.

Bonferroni correction

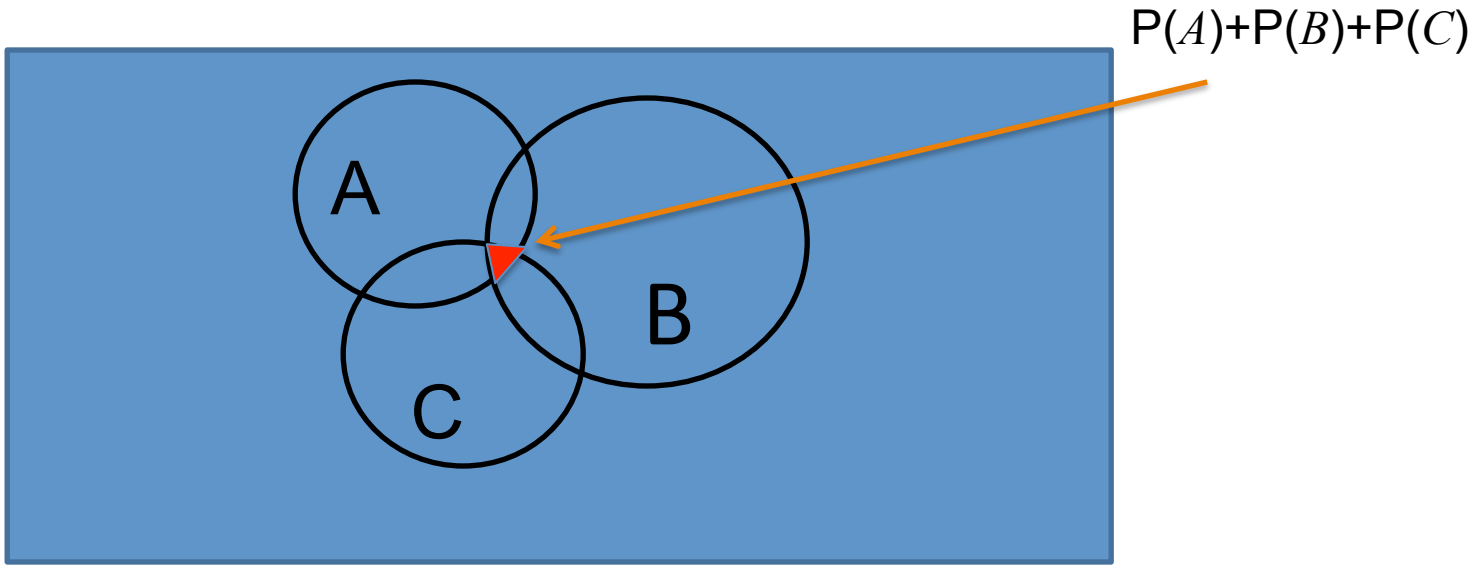
- Hence, under Bonferroni rule, we need to use a significance level of alpha/20000 for each gene .
  - Simply choose $c = \alpha / m$.
  - $\alpha = 0.05 \Rightarrow c = \alpha/20000 = 0.0000025$
  - In other words, under Bonferroni rule, we will select a gene as differentially expressed if its P-value < 0.0000025. This will guarantee the probability of making at least 1 mistake in the 20000 performed tests is not larger than 0.05.
    - More specifically, out of the genes selected, there is only very small chance (5%) that at least one of them is a false positive
  - Is this too tough (stringent, conservative)?
    - ❑Yes (if few genes' p-values are less than α/200000: Game Over…)

Weak Control vs. Strong Control

- A method provides *weak control* of an error rate for a family of *m* tests if the FWER control at level α is guaranteed **only** when all null hypotheses are true (i.e. when $m=m_0$ so the global null hypothesis is true).

- A method provides *strong control* of an error rate for a family of *m* tests if the FWER control at level α is guaranteed for **any** configuration of true and non-true null hypotheses (including the global null hypothesis)

Bonferroni's method can achieve strong control



Assuming the rectangle has probability 1, the three circles, A, B, C, represents three events. The probability P(AUBUC), i.e., the probability of A or B or C, is smaller than P(A)+P(B)+P(C).

Holm's Method for Controlling FWER at Level α

- Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ denote the m p-values ordered from smallest to largest. (need to sort all P-values first)
- Find the largest integer k so that

  $p_{(i)} \leq \alpha / (m-i+1)$ for all $i=1,\ldots,k$.

  (when you see it first time)
- set $c = p_{(k)}$ (reject the nulls corresponding to the smallest k p-values).
- If no such k exists, set $c = 0$ (declare nothing significant).

An Example

- Suppose we conduct 5 tests and obtain the following *p*-values for tests 1 through 5.

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

- Which tests' null hypotheses will you reject if you wish to control the FWER at level 0.05?

- Use both the Bonferroni method and the Holm method to answer this question.

Solution

Test	T1	T2	T3	T4	T5
P-value	0.042	0.001	0.031	0.014	0.007

- The cutoff for significance is $c = 0.05/5=0.01$ using the Bonferroni method. Thus we would reject the null hypothesis for tests 2 and 5 with the Bonferroni method.

T2: $0.001 \leq 0.05/(5-1+1)=0.01$
T5: $0.007 \leq 0.05/(5-2+1)=0.0125$
T4: $0.014 \leq 0.05/(5-3+1)=0.0167$
T3: $0.031 > 0.05/(5-4+1)=0.025$
T1: $0.042 \leq 0.05/(5-5+1)=0.05$

These calculations indicate that Holm's method would reject null hypotheses for tests 2, 5, and 4.

Adjusted p-value

- P-value: the probability to observe more or equally extreme data under the null hypothesis.
- Alternatively, a *p*-value for an individual test can be defined as the smallest significance level (tolerable type 1 error rate) for which we can reject the null hypothesis. For example, if p-value is 0.045, this null hypothesis will be rejected if α=0.05 but note rejected if α=0.04. The smallest α to reject this null hypothesis is 0.45 (p-value).
- The ***adjusted p*-value** for one test in a family of tests is the smallest significance level for which we can reject the null hypothesis for that one test and all others with smaller *p*-values.

Adjusted p-values

- FWER: the *adjusted p*-value for one test in a family of tests is the smallest FWER (α) for which we can reject the null hypothesis for that one test and all others with smaller *p*-values.
- Bonferroni: the null hypothesis will be rejected if unadjusted *p-value* ≤ *α/m*. So the smallest α that can lead to rejection will be *m* x *p-value*, i.e., the adjusted p-value is the raw p-value times *m*.
- Holms: adjusted p-value for *i*-th ordered p-value is

$$p_{(i)} \times (m - i + 1)$$

- The advantage of adjusted p-values: they can be compared directly with α.

Example

Test	T1	T2	T3	T4	T5
Raw P-value	0.042	0.001	0.031	0.014	0.007
Bonferroni adjusted	0.21	0.005	0.155	0.07	0.035

Reject hypotheses 2 and 5 for Bonferroni's method

Holms

0.001*(5-1+1)=0.005
0.007*(5-2+1)=0.028
0.014*(5-3+1)=0.042 $\alpha<0.05$
0.031*(5-4+1)=0.062
0.042*(5-5+1)=0.042

These calculations indicate that Holm's method would reject null hypotheses for tests 2, 5, and 4.

A Conceptual Description of FWER

- Suppose a scientist conducts 100 independent microarray/RNA-seq experiments.
- For each experiment, the scientist produces a list of genes declared to be differentially expressed by testing a null hypothesis for each gene.
- Each list that contains one or more false positive results is considered to be in error.
- The FWER is approximated by the proportion of 100 lists that contain one or more false positives.

FWER Too Conservative for Microarrays/RNA-seq?

- Suppose that one of the 100 gene lists consists of 500 genes declared to be differentially expressed.
- Suppose that one of those 500 genes is not truly differentially expressed but that the other 499 are.
- This list is considered to be in error and such lists are allowed to make up only a small proportion of the total number of lists if FWER is to be controlled.
- However such a list seems quite useful from the scientific viewpoint. Perhaps it is not so important to control FWER for most high throughput experiments.

The solutions with R

> results=topTable(fit2, number=20,
adjust.method="xxx")

> results=topTags(fit2, number=20,
adjust.method="xxx")

adjust.method: "holm", "hochberg", "hommel",
"bonferroni", "BH", "BY", "fdr", "none"

The solutions

- To select differentially expressed genes, we need to do multiple testing (multiplicity) corrections
  - Familywise Error Rate (FWER), such as Bonferroni correction and Holm's method: adjust the p-value threshold from alpha to alpha/(number of genes)
  - Control False Discovery Rate: algorithm proposed by Benjamini & Hochberg
  - Re-sampling techniques (i.e., Permutation P-values)

FDR (False Discovery Rate)

- The investigators, after spending thousands of dollars, want to obtain a list of selected genes
- As Bonferroni correction is very strict, only a few genes might be selected
- As an alternative solution, we can choose to control the proportion of false positives out of selected genes.
- FDR is an alternative error rate that can be useful for high throughput experiments.

False Discovery Rate (FDR)

- FDR was introduced by Benjamini and Hochberg (1995) and is formally defined as

  E(Q) where Q=V/R if R>0 and Q=0 otherwise.

- Controlling FDR amounts to choosing the significance cutoff *c* so that FDR is less than or equal to some desired level α.

- More specifically, if we want to control at most 5% false positives, which genes should be selected?

FDR (False Discovery Rate)

The results of statistics test

The real status of data

	Negative	Positive	Total
Truly unchanged	**True Negative (U)**	**False Positive (V) Type I Error**	**M0**
Truly differentially expressed	**False Negative (T) Type II error**	**True Positive (S)**	**M-M0**
Total	**M-R**	**R**	**M**

- U: number of true negatives; S: number of true positives
- T: number of false negatives; V: number of false positives
- In our microarray example, M=20000 genes
- R is known (i.e., how many genes are called positive by statistics tests)

FDR (False Discovery Rate)

The results of statistics test

The real status of data

	Negative	Positive	Total
Truly unchanged	**True Negative (U)**	**False Positive (V) Type I Error**	**M0**
Truly differentially expressed	**False Negative (T) Type II error**	**True Positive (S)**	**M-M0**
Total	**M-R**	**R**	**M**

- FDR is defined as the expected proportion of false positives (type I errors) among all rejected null hypotheses

$$FDR = E(Q) \quad \text{with} \quad Q = V / R \quad \text{if} \quad R > 0$$
$$Q = 0 \quad \text{if} \quad R = 0$$

FDR (False discovery rate): How?

- The Benjamini & Hochberg procedure to control FDR :
  - For each gene (out of a total of n), perform one test
  - Obtain n P-values: $p_1, p_2, \ldots, p_n$
  - Sort the obtained P-values: $p_{(1)}, p_{(2)}, \ldots, p_{(n)}$
  - To control the FDR at q , we will reject all genes with p-values $p \leq p_{(j)}$, where j is the largest index for which

$$p_{(j)} \leq \frac{qj}{n}$$

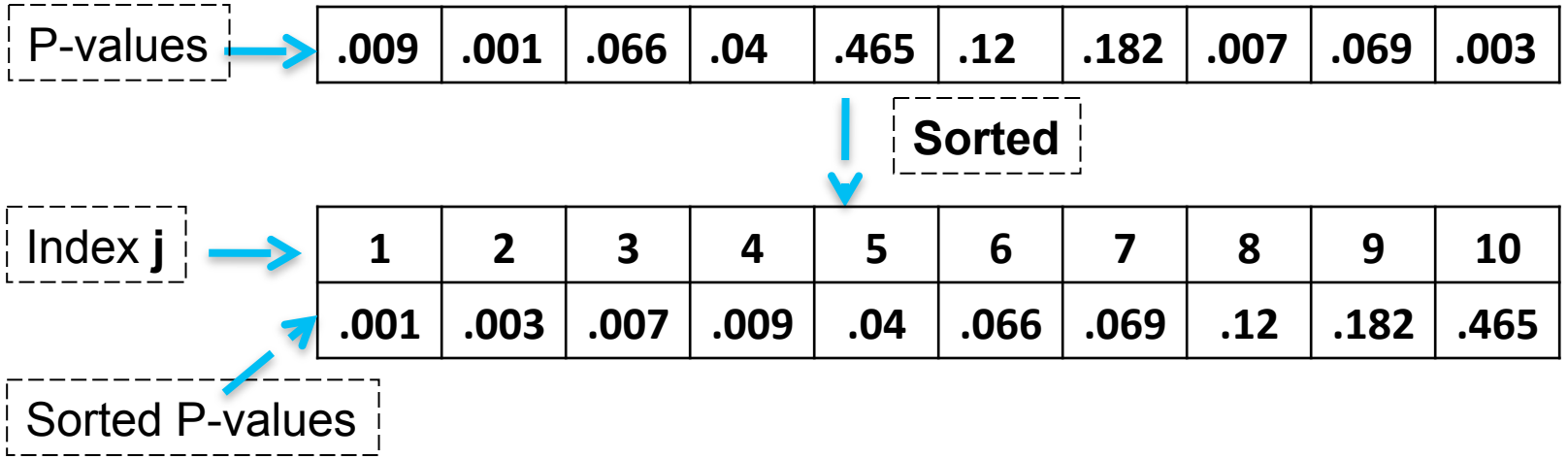
FDR (False Discovery Rate): An Example of 10 genes

- Aim: To control the FDR at level of 5%

P-values →	.009	.001	.065	.04	.454	.123	.172	.007	.68	.003

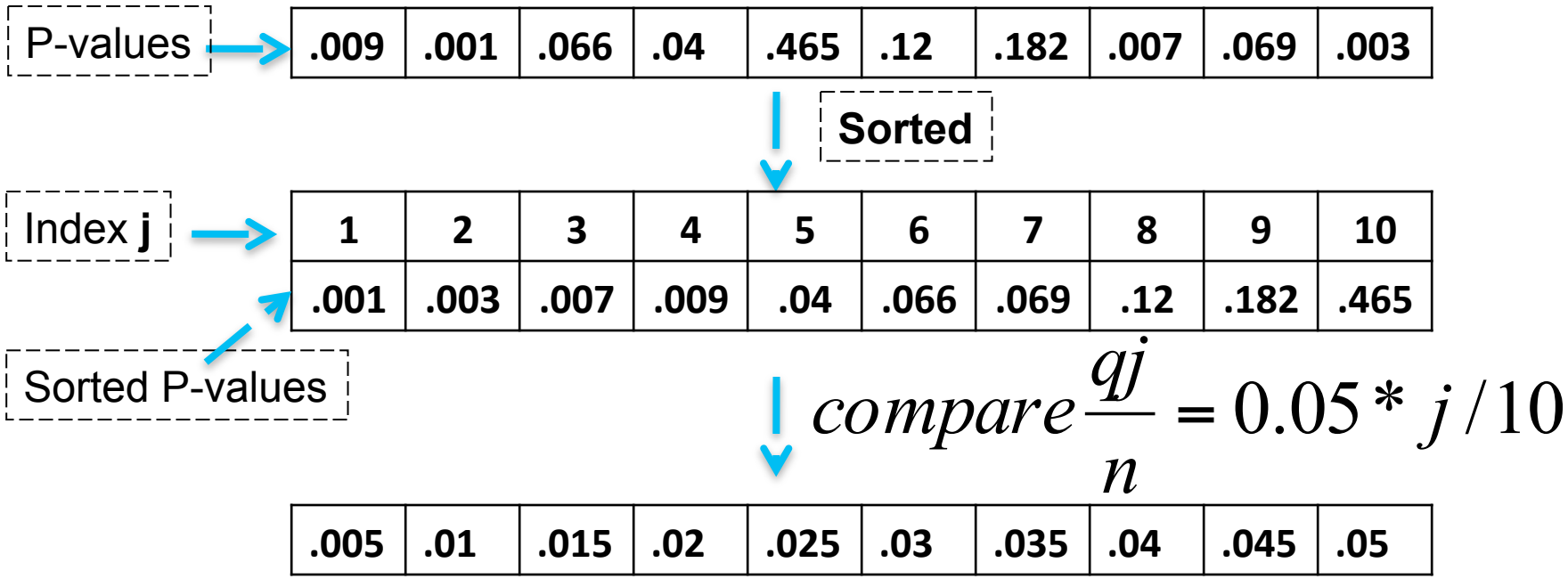
FDR (False Discovery Rate): An Example of 10 genes

- Aim: To control the FDR at 5% (q =0.05)



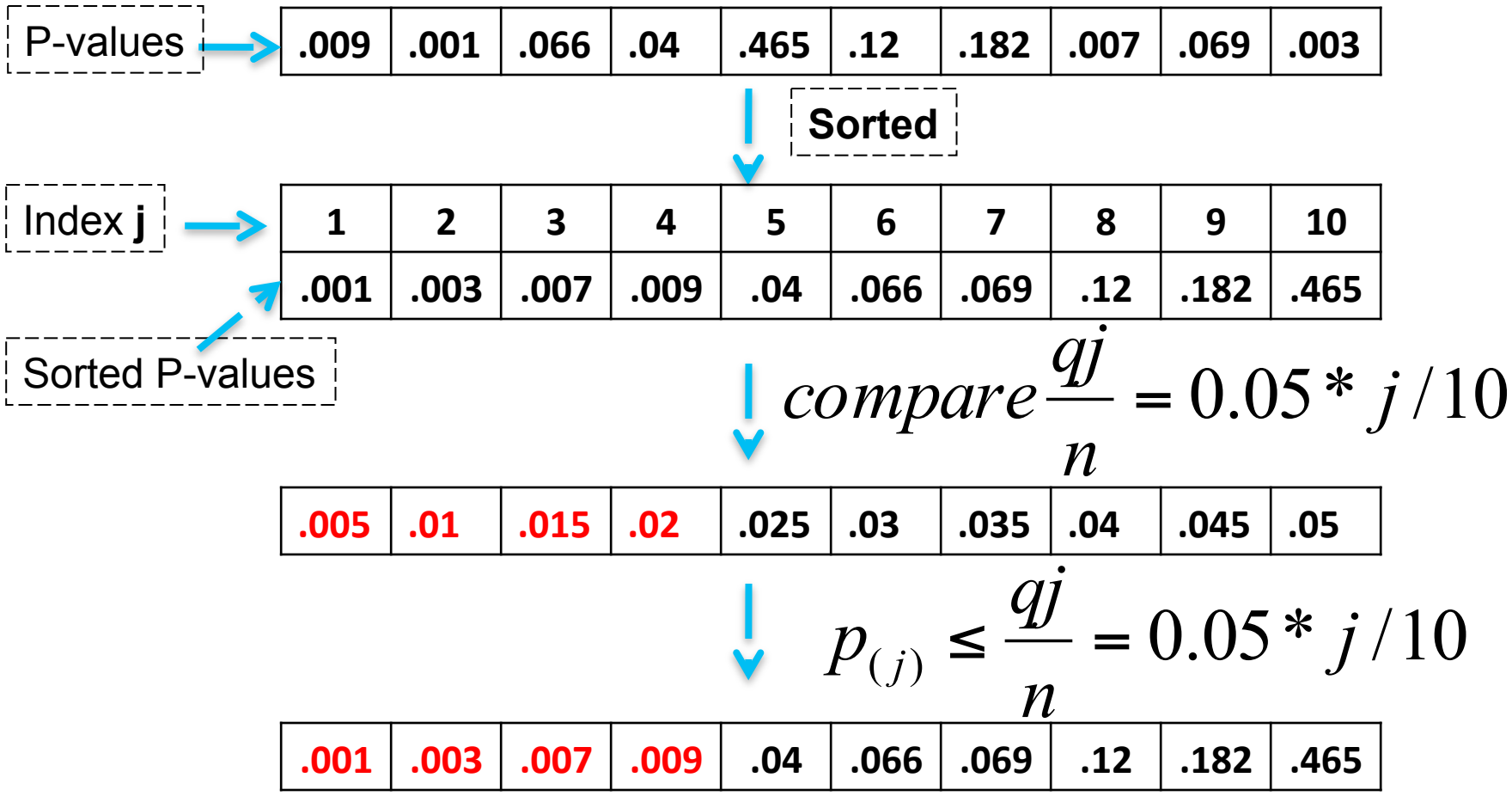
FDR (False Discovery Rate): An Example of 10 genes

- Aim: To control the FDR at 5% (q =0.05)



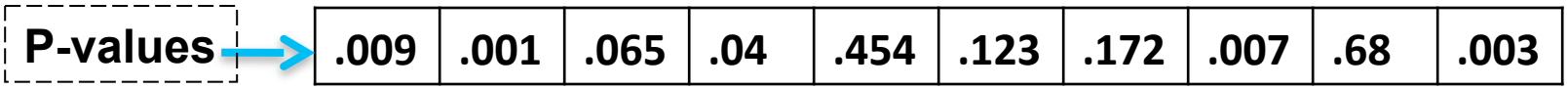
FDR (False Discovery Rate): An Example of 10 genes

- Aim: To control the FDR at 5% (q =0.05)



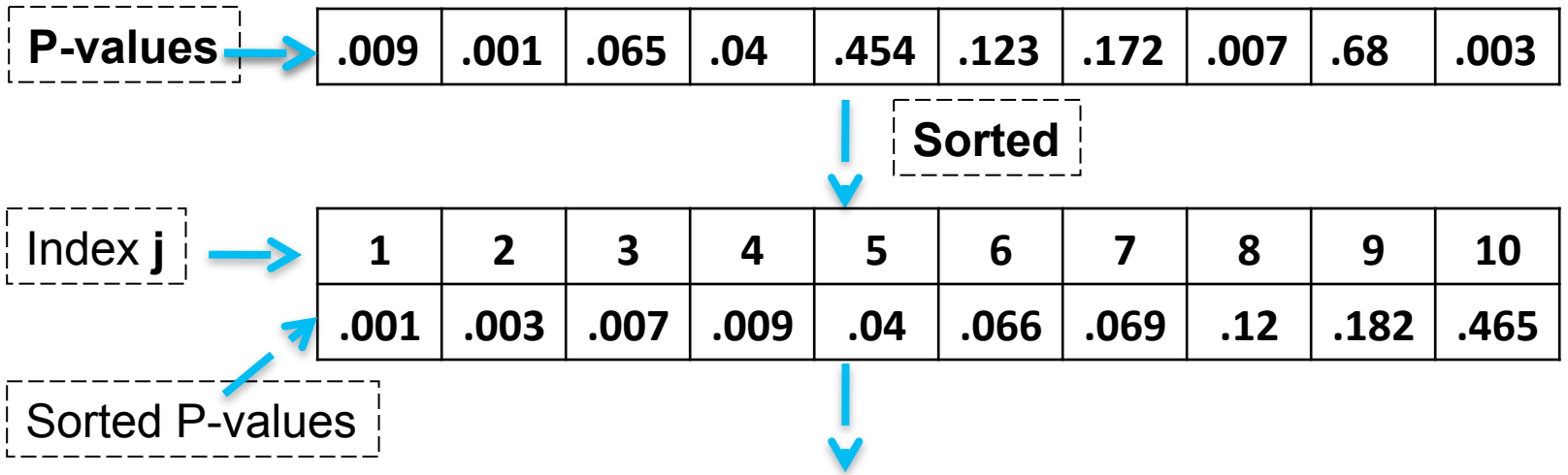
How about Bonferroni correction? The Same Example of 10 genes

- Aim: Use Bonferroni correction, alpha=0.05



How about Bonferroni correction? The Same Example of 10 genes

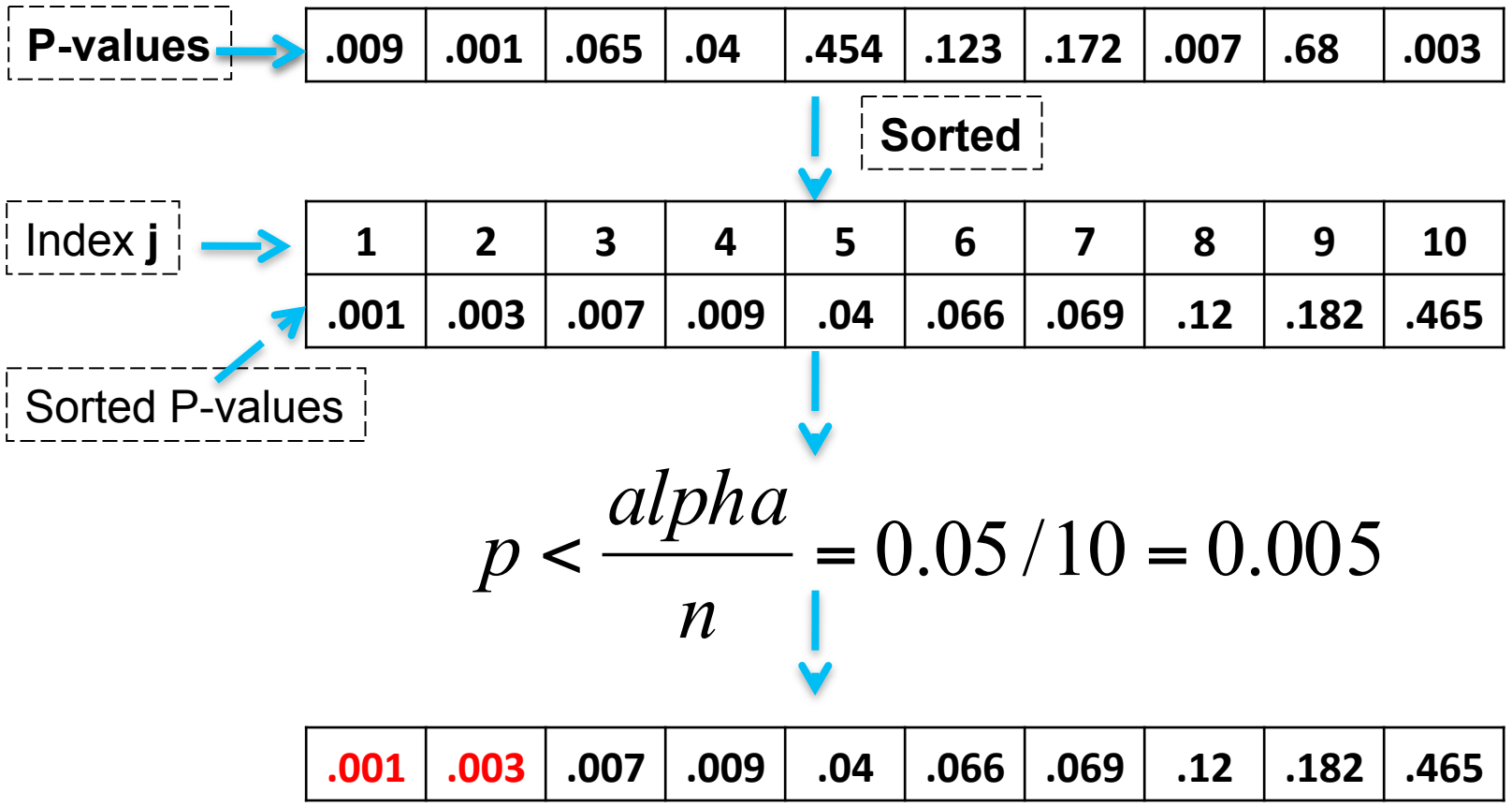
- Aim: Use Bonferroni correction, alpha=0.05



$$compare \frac{alpha}{n} = 0.05/10 = 0.005$$

How about Bonferroni correction? The Same Example of 10 genes

- Aim: Use Bonferroni correction, alpha=0.05



Adjusted p-values (q-values)

- If we use FDR as the significance threshold, the adjusted *p*-value for one test in a family of tests is the smallest FDR for which we can reject the null hypothesis for that one test and all others with smaller *p*-values.

- In FDR setting, adjusted p-values are also called q-values. q-value is derived in an empirical Bayes setting, but it is equivalent to adjusted p-value in practice.

The adjusted p-value or *q*-value for a given test fills the blanks in the following sentences:

- "If I set my cutoff for significance *c* equal to this *p*-value, I must be willing to accept a false discovery rate of _____."
- "To reject the null hypothesis for this test and all others with smaller *p*-values, I must be willing to accept a false discovery rate of _____."
- "To include this gene on my list of differentially expressed genes, I must be willing to accept a false discovery rate of _____."

Computation and Use of *q*-values

- Let $q_{(i)}$ denote the *q*-value that corresponds to the ith smallest *p*-value $p_{(i)}$.

- $\hat{q}_{(i)} = \min \{ p_{(k)} m_0 / k : k = i,\ldots,m \}$.

- To produce a list of genes with estimated FDR ≤ α, include all genes with *q*-values ≤ α.

The solutions with R

```
> results=topTable(fit2, number=20,
   adjust.method="fdr", lfc=1)
> results=topTags(fit2, number=20,
   adjust.method="fdr”)
```

adjust.method: "holm", "hochberg", "hommel",
"bonferroni", "BH", "BY”, "fdr", "none”

“fdr”

The solutions

- To select differentially expressed genes, we need to do multiple testing (multiplicity) corrections
  - Familywise Error Rate (FWER), such as Bonferroni correction and Holm's method: adjust the p-value threshold from alpha to alpha/(number of genes)
  - Control False Discovery Rate: algorithm proposed by Benjamini & Hochberg
  - Re-sampling techniques (i.e., Permutation P-values)

A typical re-sampling procedure

- For each gene g,
  - Step1: Perform test and obtain its absolute statistics: $\left|t_g\right|$
    - E.g., perform t-test and obtain t-statistics
  - Step2: Repeat the following sub-steps n times:
    - Randomly permute the sample labels (e.g., through bootstrap)
    - For each permutation, compute the new statistics, $\left|t_{g(i)}\right|$ where i=1,…n
  - Step3: The permutation p-value of gene g is the number of times $\left|t_{g(i)}\right| > \left|t_g\right|$, divided by the number of permutation (n).
  - Step4: Select the gene whose permutation p-value is less than the pre-specified significance level, alpha (e.g., 0.05) or make further multiple testing corrections.

P-value again…

- The meaning of a p-value from a permutation procedure differs from the meaning of a model-based p-value.
  - The model-based p-value is the probability of the test statistic, assuming that the gene levels in both the treatment and control groups follow the model (e.g., a Normal distribution).
  - A permutation-based p-value tells how rare that test statistic is, among all the random partitions of the actual samples into pseudo-treatment and pseudo-control groups.

Concluding Remarks

- In many cases, it will be difficult to separate the many of the differentially expressed genes from the non-differentially expressed genes.

- Genes with a small expression change relative to their variation will have a *p*-value distribution that is not far from uniform if the number of experimental units per treatment is low.

- To do a better job of separating the differentially expressed genes from the non-differentially expressed genes we need to use good experimental designs with more replications per treatment.

Outline

- Multiple Testing Procedures
- Data Visualization, Distance Measures
- Clustering
- Gene Annotation and Enrichment Analysis

Data visualization: why do we care?

- A (good) picture is worth of thousands of words
  - “The adage "**A picture is worth a thousand words**" refers to the idea that a complex idea can be conveyed with just a single still image. It also aptly characterizes one of the main goals of **visualization**, namely making it possible to absorb large amounts of data quickly.”
  - “Figures are often the most important part of a scientific paper, so please take some time making really good quality figures.” – Joe Wolfe, in "Writing and publishing a scientific paper".

The Visualization

- MA plot
- Volcano plot
- Heatmap
- Dendrogram

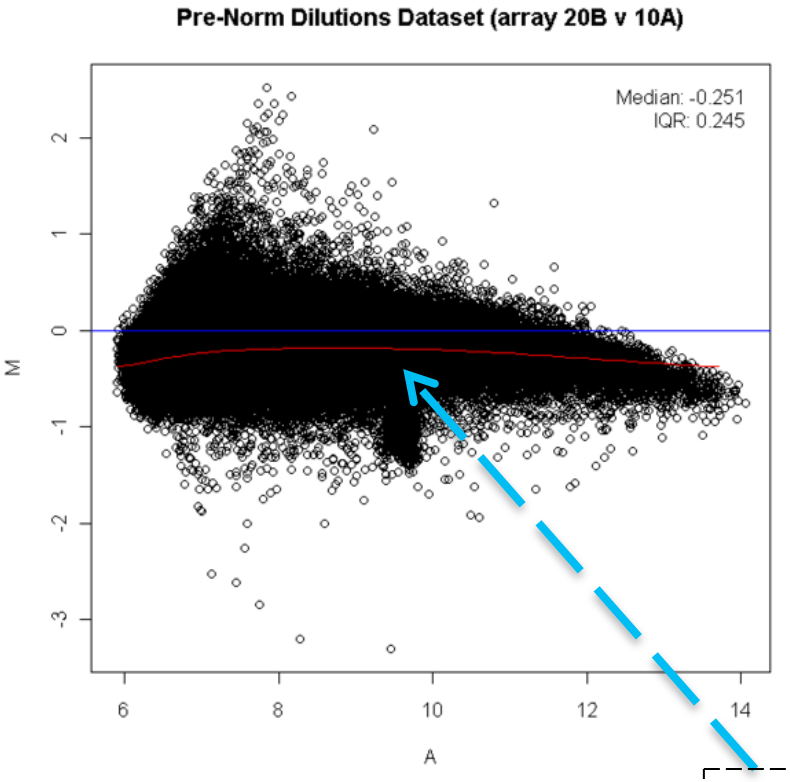
MA plot

- Under R

```
> y <- (exprs(Dilution)[, c("20B", "10A")])

> ma.plot( rowMeans(log2(y)), log2(y[, 1])-log2(y[, 2]), cex=1 )

> title("Pre-Norm Dilutions Dataset (array 20B v 10A)")
```

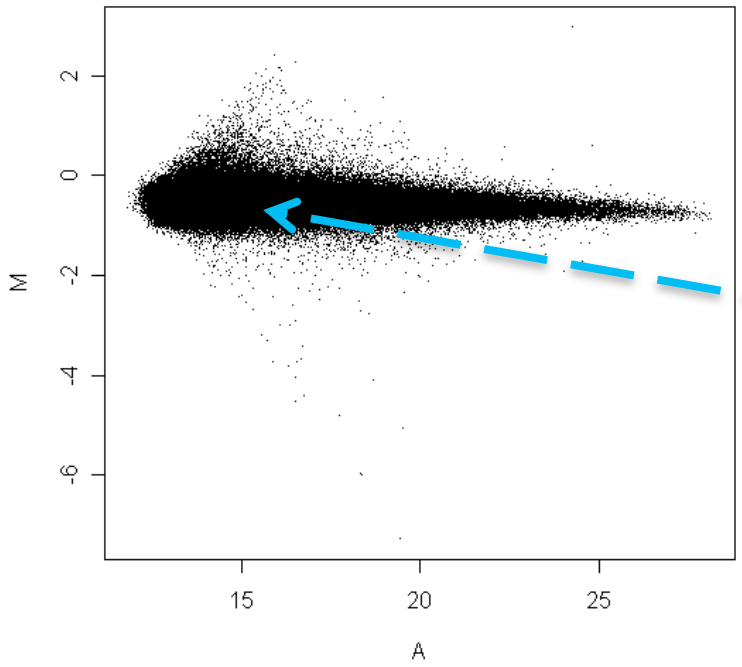


Not zero!

- M is the log2 intensity ratio for a probe in the two chips
- A is the average log2 intensity for a probe in the two chips
- The MA plot gives a quick overview of the distribution of the data. The general assumption is that most of the genes would not see any change in their expression. Therefore the majority of the points on the y axis (M) would be located at 0, since Log(1) is 0.

MA plot

- Limitation: Only informative with a small / intermediate number of observations



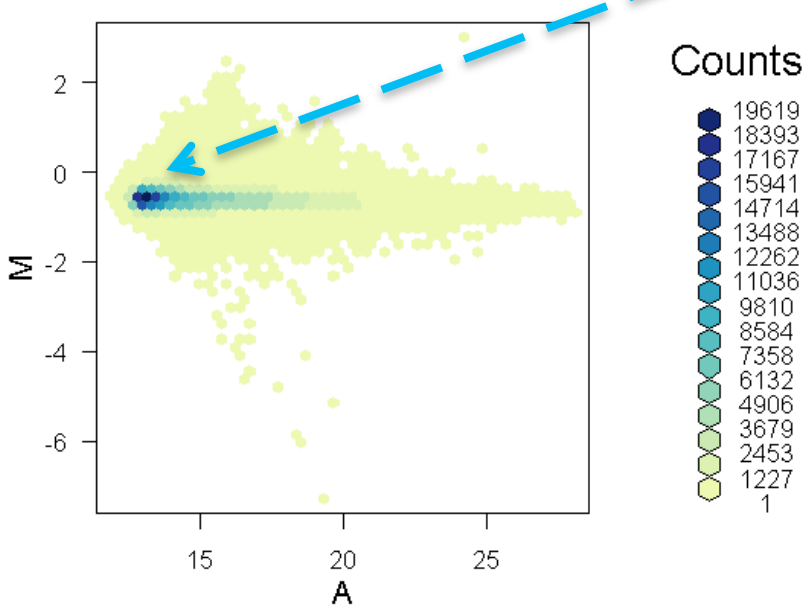
Can you tell where most of these 409,600 points are??
❑Yes?
❑No?

Data from two Affymetrix GeneChips with 409, 600 probes each

MA plot

- Solution - Hexagonal binning algorithom (Carr et al 1987)
- **hexbin** package

Can you tell where most of these 409,600 points are??




```
>library("hexbin")
>library("geneplotter")
>library("RColorBrewer")
>hb <- hexbin(x,xbins=50)
>plot(hb,colramp =
>colorRampPalette(brewer.pal
+ (9, "YlGnBu")[-1]))
```

Data from two Affymetrix GeneChips with 409, 600 probes each

The Visualization

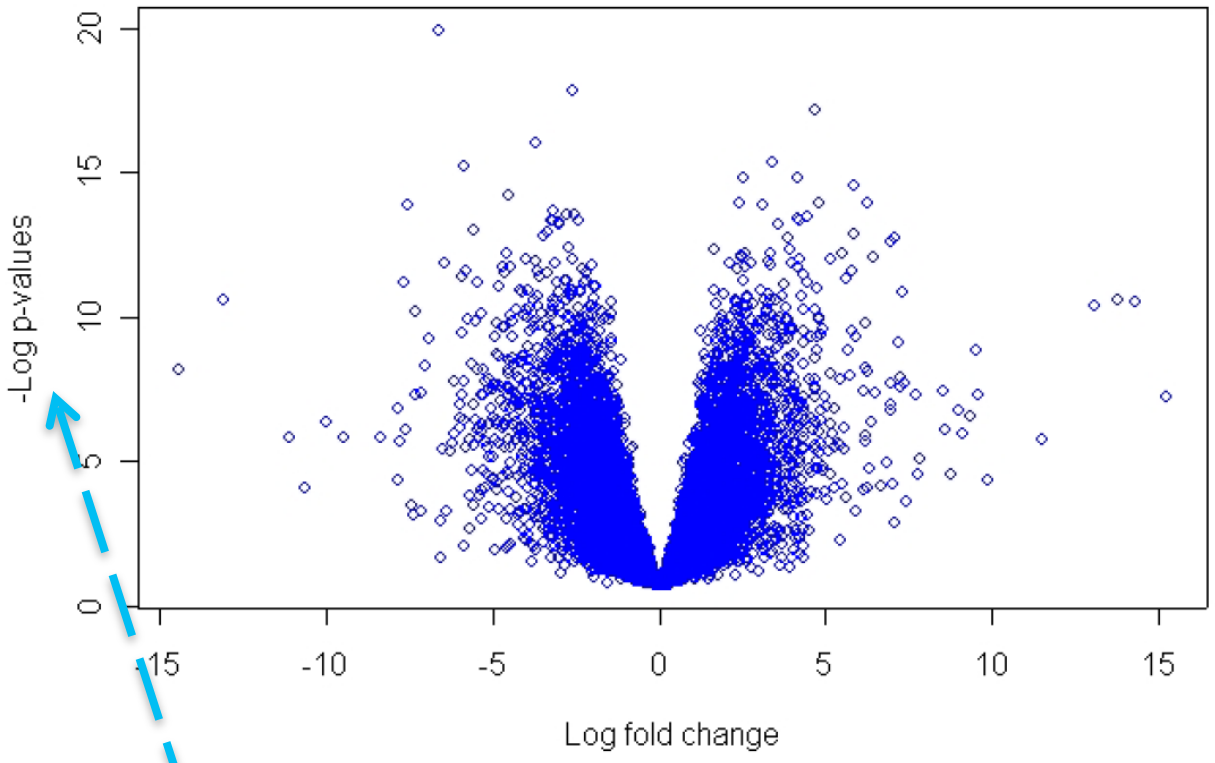
- MA plot
- Volcano plot
- Heatmap
- Dendrogram

Volcano plot

- In statistics, a **volcano plot** is a type of scatter-plot that is used to quickly identify changes in large datasets composed of replicate data.

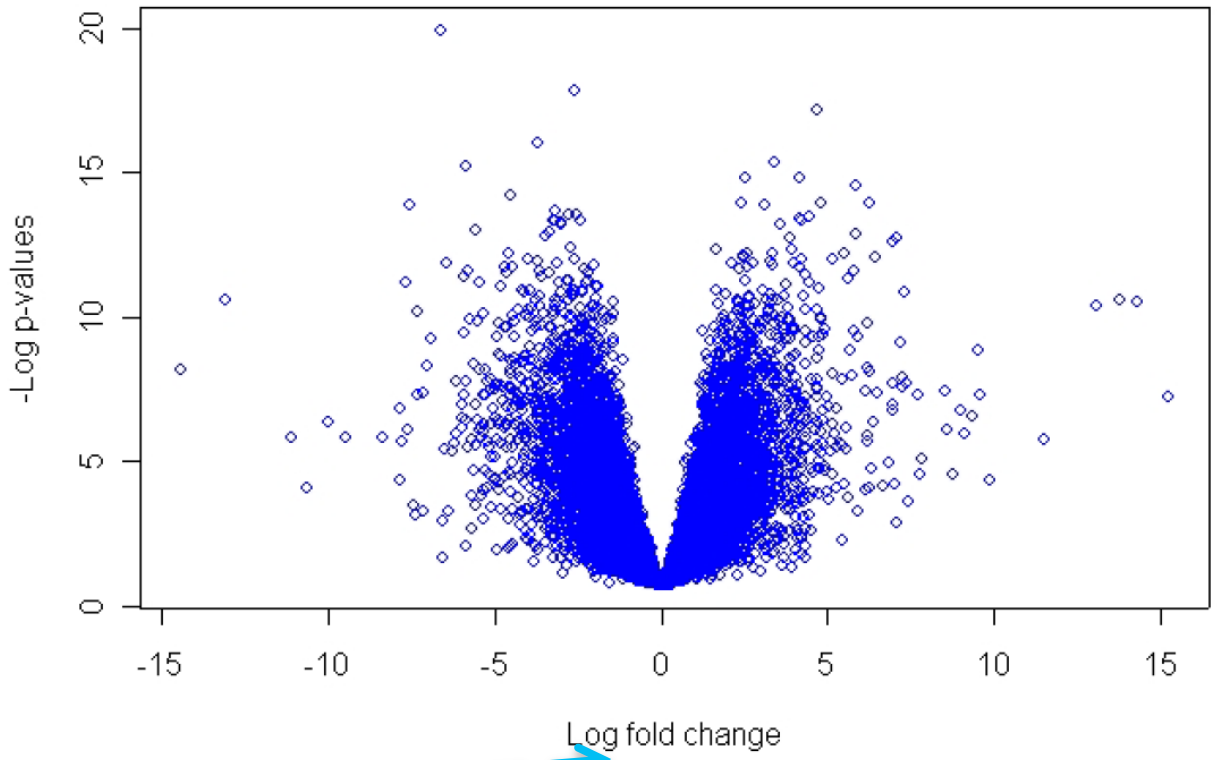
- It plots significance versus fold-change on the y- and x-axes, respectively.

Volcano plot



- A volcano plot is constructed by plotting the negative log of the p-value on the y-axis (usually base 10). This results in data points with low p-values (highly significant) appearing towards the top of the plot.

Volcano plot



- The x-axis is the log of the fold change between the two conditions. The log of the fold-change is used so that changes in both directions (up and down) appear equidistant from the center.

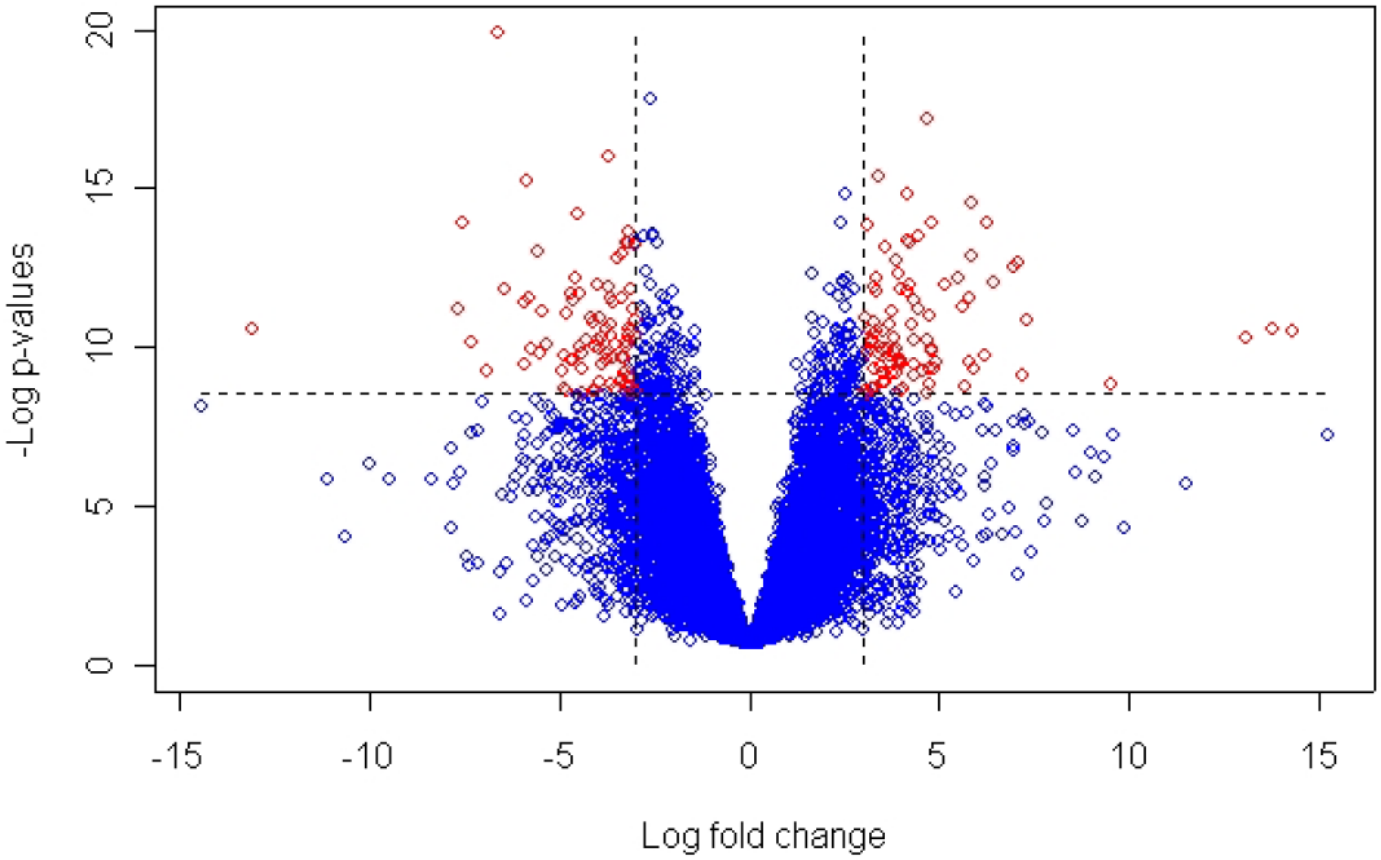
Volcano plot: why?

- P-values (on the y-axis) measure expression change in terms of probability
  - It has statistical meaning (significance)

- Fold-changes (on the x-axis) measure expression change in terms of magnitude
  - It has biological meaning (e.g., P53 is 3-fold up in tumor versus normal)

- A combination of these two gives us best information to select differentially expressed genes

Volcano plot

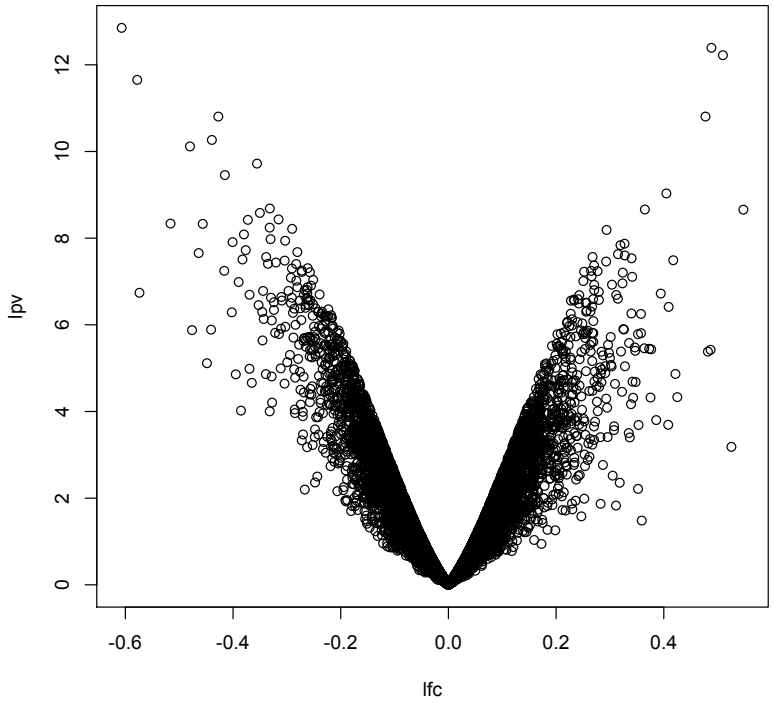


- If we combine fold-changes and P-values to select the differentially expressed genes, the list of genes are both "biologically interesting" and "statistically significant"

Volcano plot: R

```
> results=topTable(fit2, number=20, adjust.method="fdr",
    lfc=1)
```

```
> results=topTable(fit2, number=20000,
    adjust.method="fdr", lfc=0)
> lfc=results[,2]
> lpvalue=-log2(results[,5])
> plot(lfc,lpvalue)
```



The Visualization

- MA plot
- Volcano plot
- Heatmap
- Dendrogram

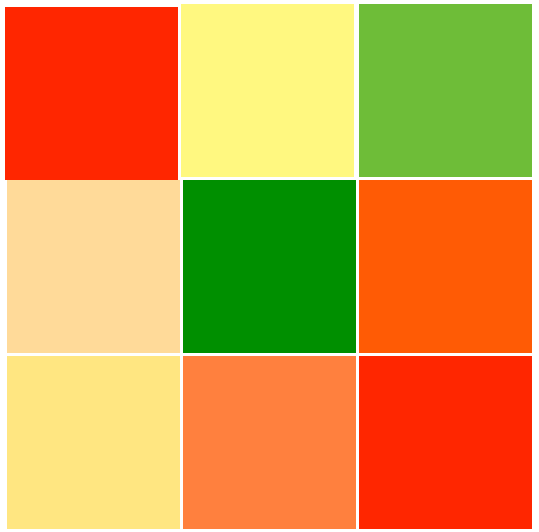
Heatmap

- A **heat map** is a graphical representation of data where the values taken by a variable in a two-dimensional map are represented as colors

- In Microarray, it plots the level of expression of many genes (in y-axis) across a number of samples (in x-axis)
  - The data is in the form of matrix

Heatmap

10	5	2
7	0	9
6	8	10



Heatmap: Example

- **ALL** data (Lymphoblastic leukemia study):
  - 12625 probes (genes)

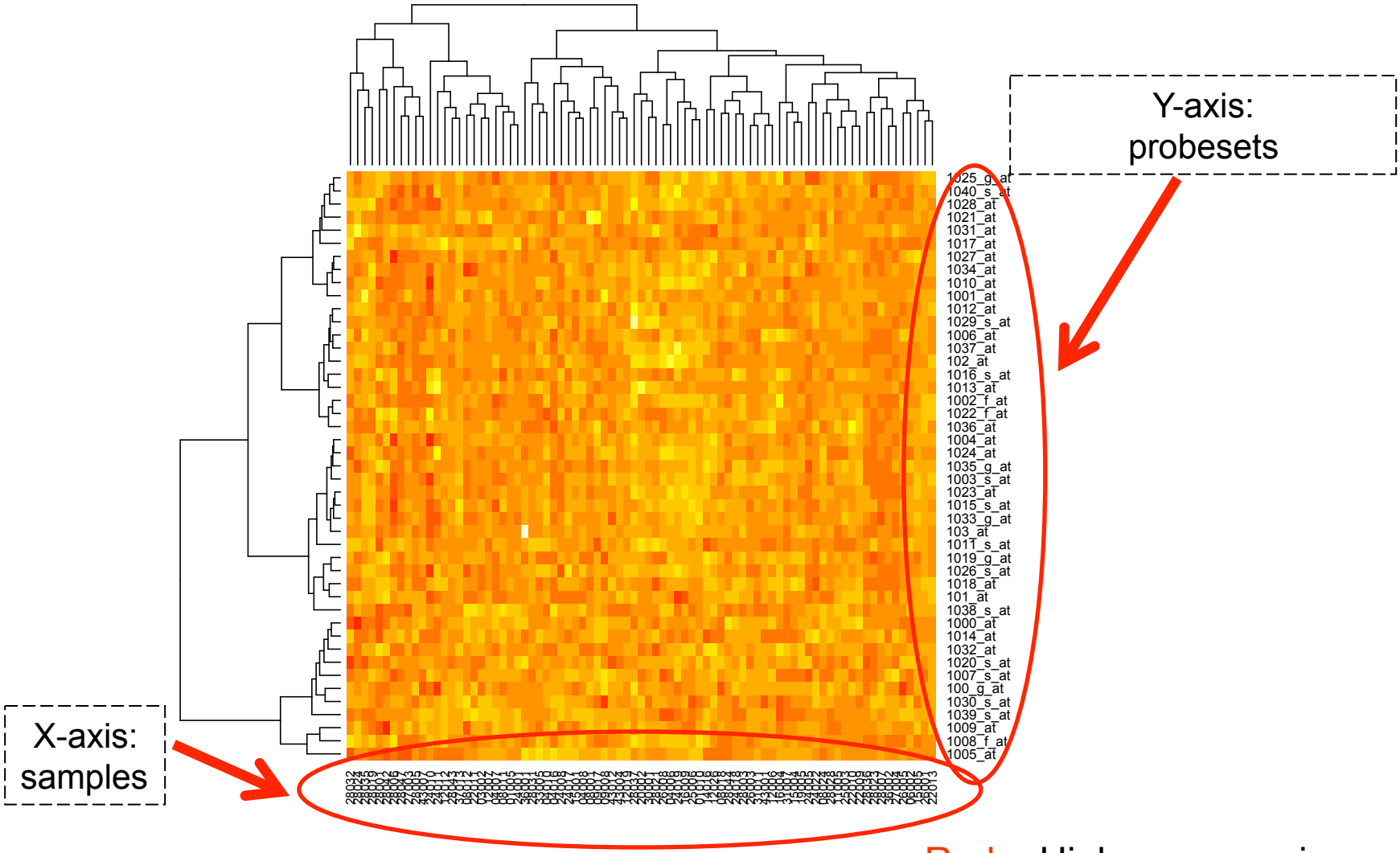
```
> exprs(ALLhm)
          ALL1/AF4 04006 E2A/PBX1 08018 ALL1/AF4 15004 ALL1/AF4 16004 ALL1/AF4 19005 AL
1007_s_at       6.816397       7.151422       6.822427       6.709222       6.798443
1044_s_at       4.570669       7.019295       4.892009       4.889920       4.339371
1065_at         8.475419       6.880097       9.939768       9.140339       9.579710
1081_at         8.631929      10.443100       8.487560       7.823037       9.879712
1134_at         7.854585       9.238699       7.559106       7.837794       7.864575
1140_at         8.039748       5.798014       6.791144       6.733774       7.276141
```

Heatmap: Example

- **ALL** data (Lymphoblastic leukemia study):
- source("http://www.bioconductor.org/biocLite.R")
- biocLite("ALL")

```
> library("ALL”)
> data(ALL)
> p=exprs(ALL)[1:45,1:81]
> heatmap(p)
```

Heatmap: Example

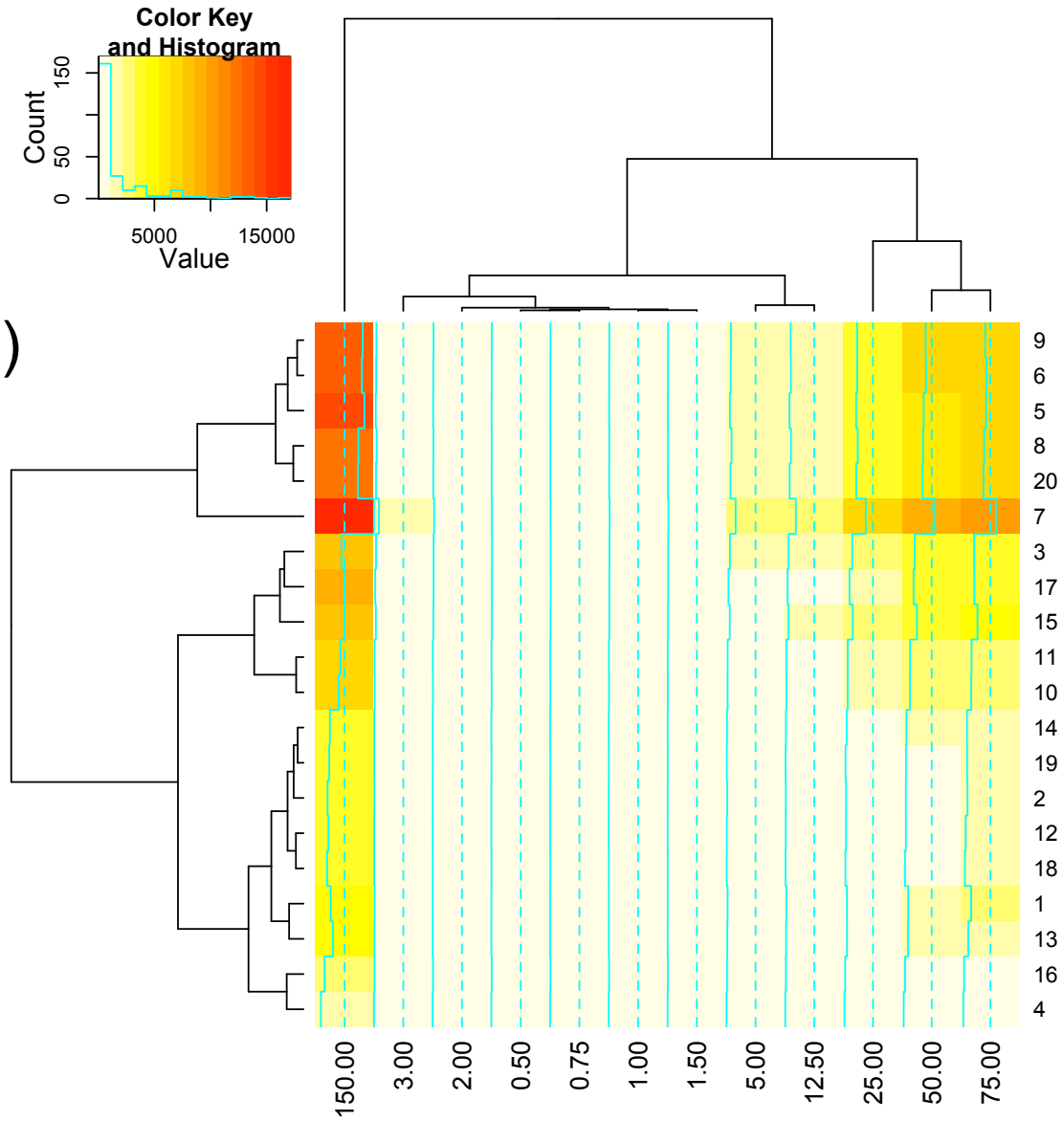


Red = Higher expression
Yellow = Lower Expression

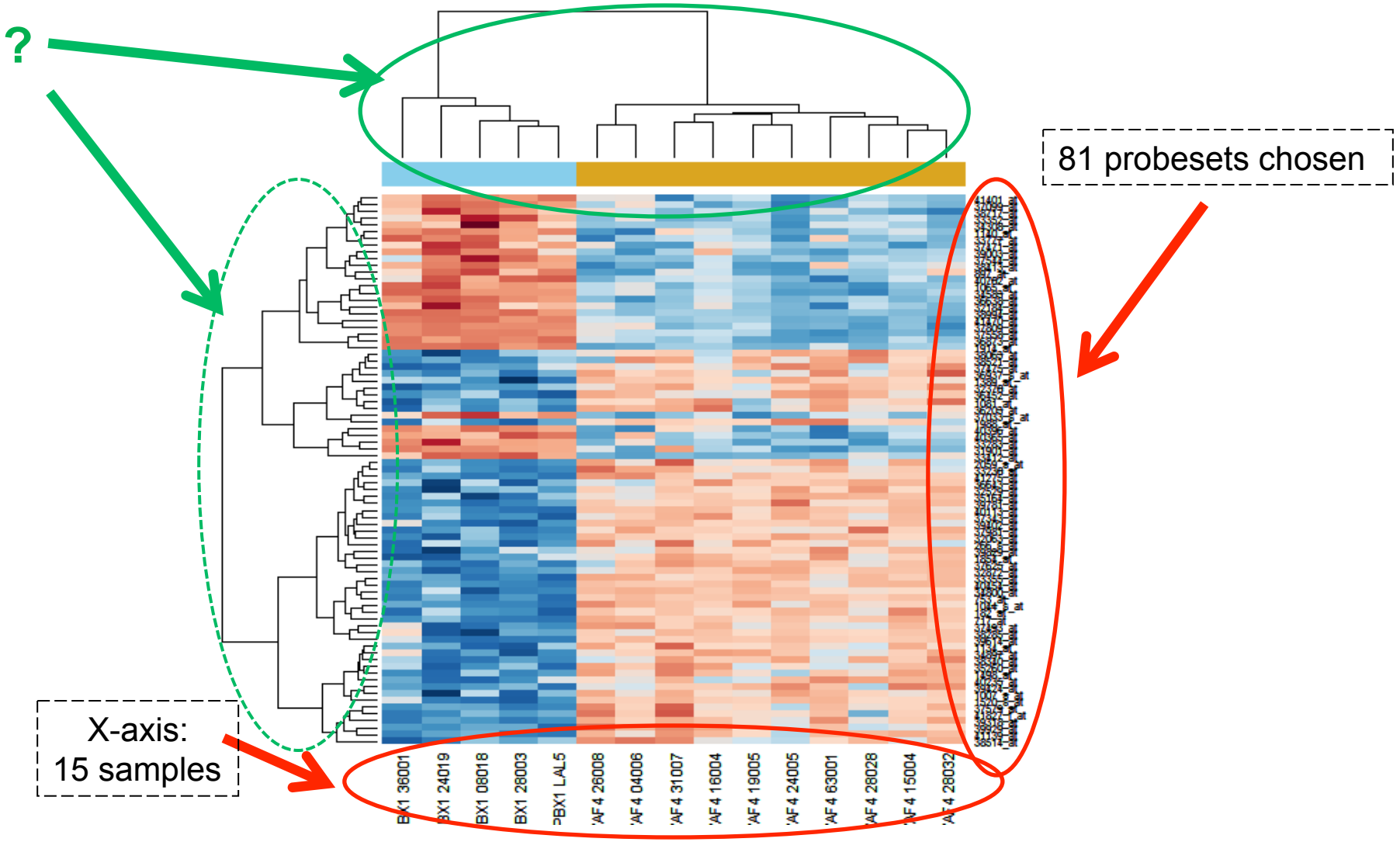
Heatmap.2()

#Name: heatmap.2
#Enhanced Heat Map

> library(gplots)
> heatmap.2(yourdata)



Heatmap: Example



Brown = Higher expression
Blue = Lower Expression

The Visualization

- MA plot
- Volcano plot
- Heatmap
- Dendrogram