BIOINFORMATICS in METABOLOMICS

Asst.Prof.Dr. Gökmen ZARARSIZ[1,2]

gokmenzararsiz@hotmail.com

[1]Erciyes University, Faculty of Medicine, Department of Biostatistics, Kayseri

[2]Turcosa Analytics Solutions Ltd.Co., Kayseri

November 4, 2016, Sivas

TURCOSA

Vahap Eldem
İstanbul Üniversitesi,
Biyoloji Bölümü,
Ankara

Selçuk Korkmaz
Hacettepe Üniversitesi,
Biyoistatistik ABD,
Ankara

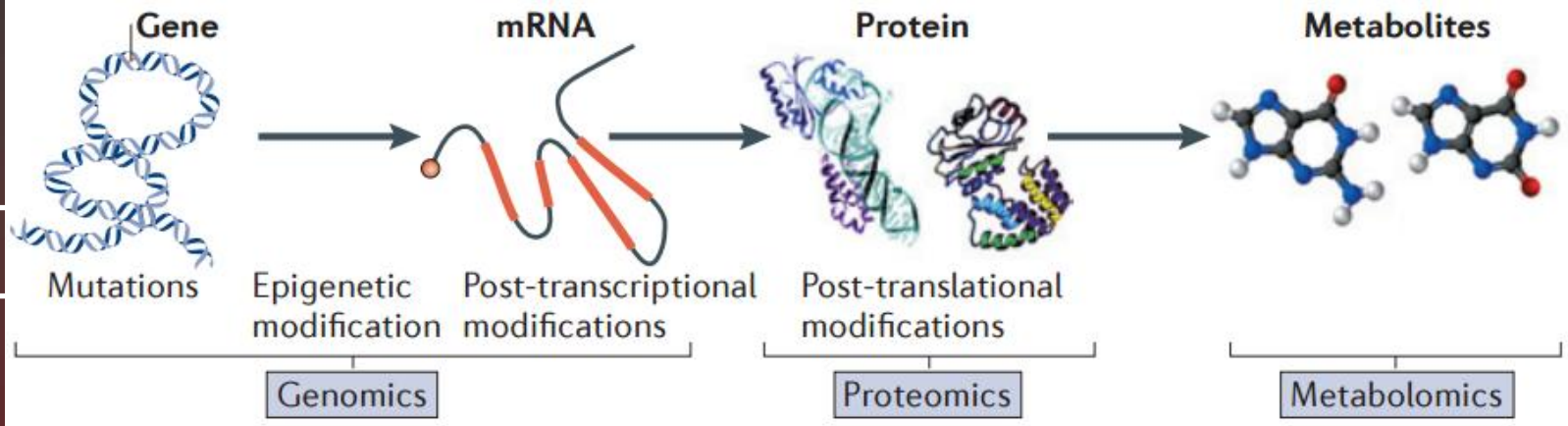
Dinçer Göksülük
Hacettepe Üniversitesi,
Biyoistatistik ABD,
Ankara

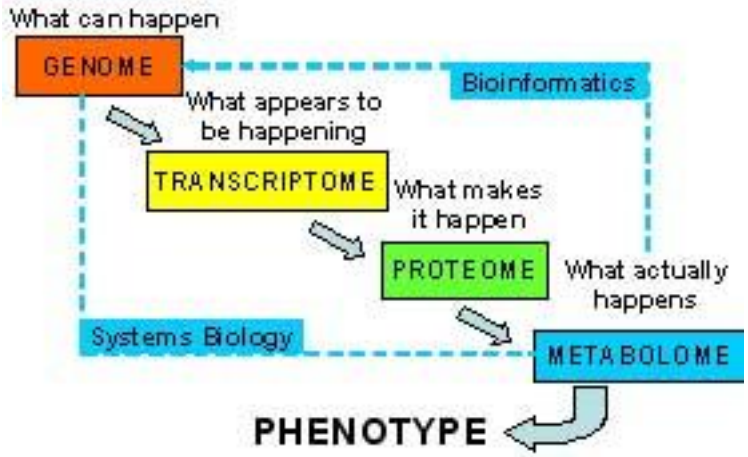
Ahmet Öztürk
Erciyes Üniversitesi,
Biyoistatistik ABD,
Kayseri

www.biosoft.hacettepe.edu.tr

www.turcosa.com.tr

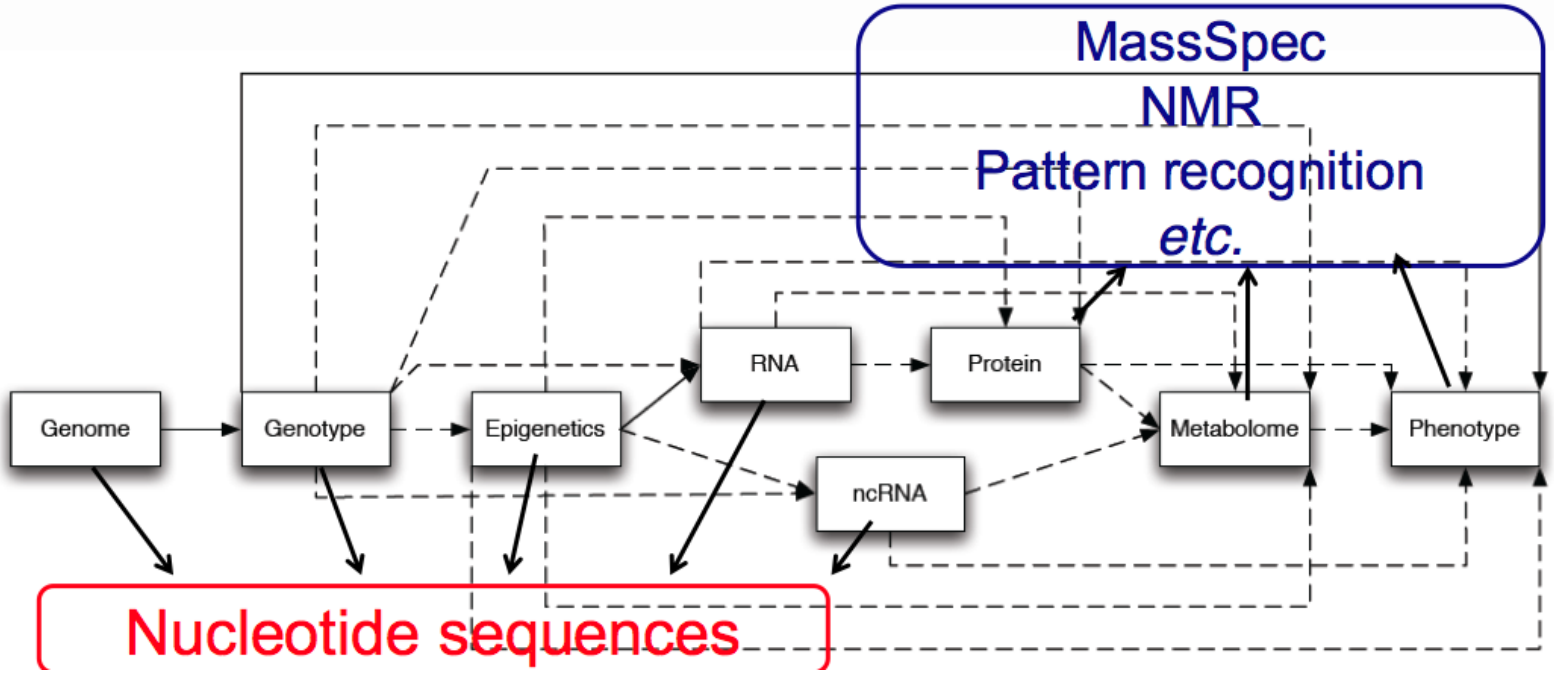
-OMICS SCIENCE





Patti et al. (2012) Innovation: Metabolomics: the apogee of the omics trilogy. Nature Reviews

-OMICS SCIENCE



METABOLOMICS

- Comprehensive analysis of all small-molecule compounds that can be found in biological samples, such as cells, tissues or biofluids
- Development / adaption of novel biostatistical & bioinformatics algorithms
- Development of softwares

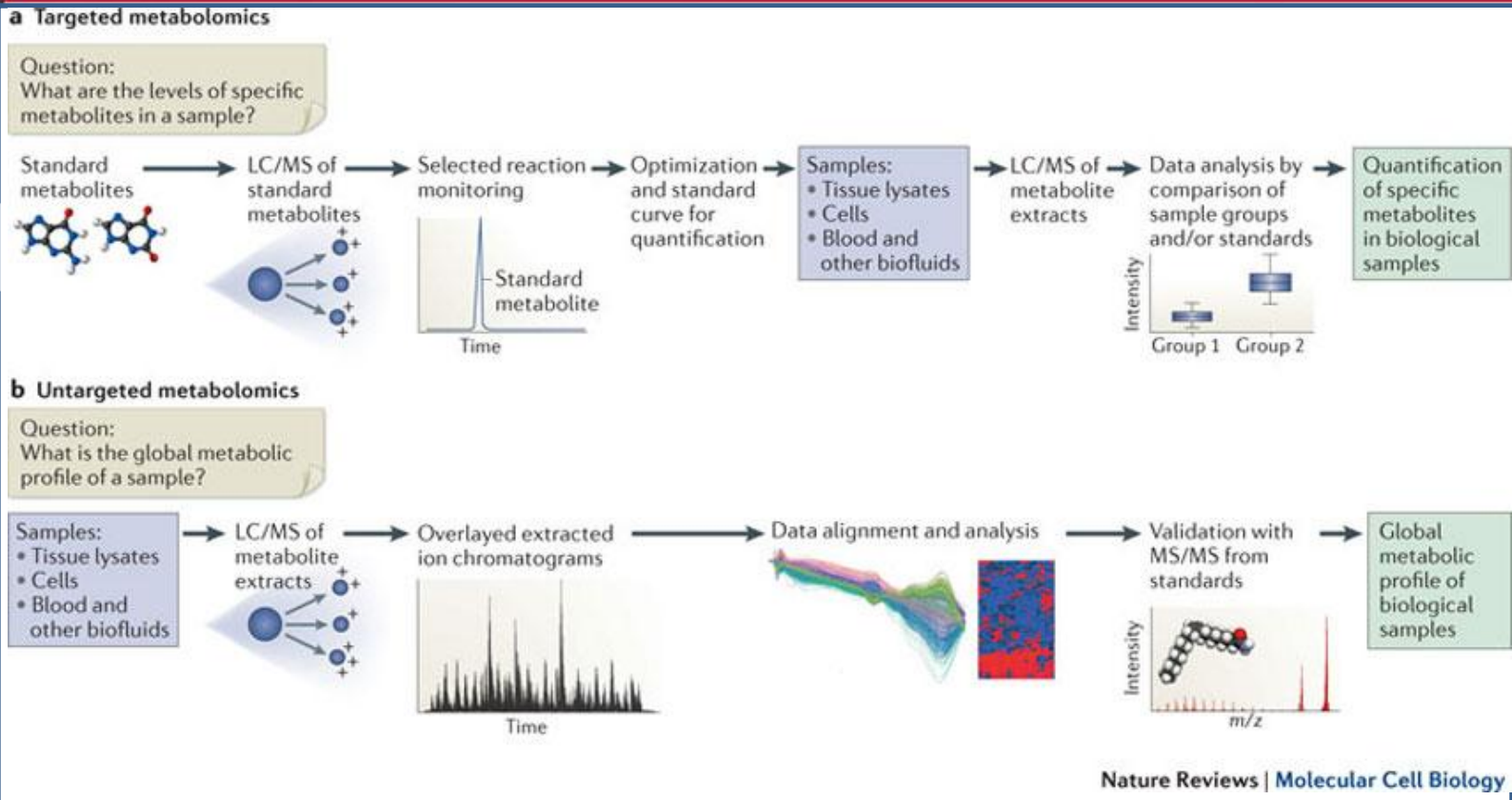
DATA COLLECTION

- **Separation Techniques**
  - Gas chromatography (GC)
  - Liquid chromatography (GC)
  - Capillary Electropheresis (CE)
- **Detection Techniques**
  - Nuclear Magnetic Resonance Spectrometry (NMR)
  - Mass Spectrometry (MS)
- **Hyphenated Techniques**
  - Gas Chromatography – Mass Spectrometry (GC-MS)
  - Liquid Chromatography – Mass Spectrometry (LC-MS)
  - Liquid Chromatography –Nuclear Magnetic Resonance (LC-NMR)

TARGETED / UNTARGETED METABOLOMICS

- Targeted (quantitative)
  - Measurement of specified list of metabolites on interested pathway(s)
  - Compound identification and quantification
  - Accurate identification and quantification of complex mixtures ??
- Untargeted (chemometric)
  - Simultaneous measurement of as many metabolites as possible from biological samples
  - Novel metabolites
  - False positives or noises ??

TARGETED / UNTARGETED METABOLOMICS



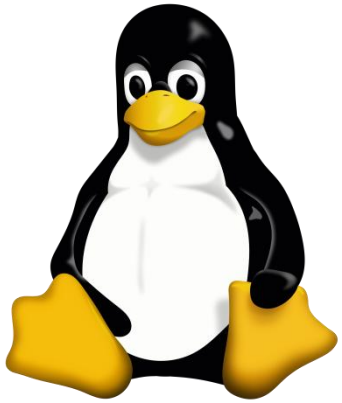
Patti et al. (2012) Innovation: Metabolomics: the apogee of the omics trilogy. Nature Reviews

DATA ANALYSIS

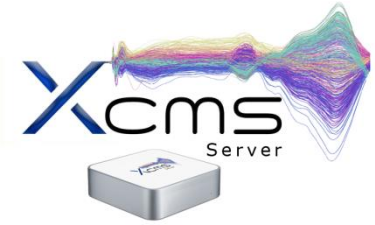
- Preprocessing
  - Handling missing values
  - Filtering
  - Normalization
  - Transformation
- Univariate and multivariate analysis
  - Class comparison (univariate analysis, biomarker discovery)
  - Class discovery (clustering)
  - Class prediction (classification)
  - Dimension reduction
- Downstream analysis
  - Metabolite set enrichment analysis
  - Pathway analysis

SOFTWARES


Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS




Statistics
MetaboAnalyst
3.0
Biomarkers
Pathways
Functions


XCMS
Server

XCMS

- Bioconductor package for processing LC/GC-MS spectra
- Highest average precision and recall
- Superfast
- Supported formats: NetCDF, mzXML, mzData

XCMS

```
> library(xcms)
> cdfpath <- system.file("cdf", package = "faahKO")
> cdffiles <- list.files(cdfpath, recursive = TRUE, full=T) # input files (step 1)
> xset <- xcmsSet(cdffiles)  # peak picking (step 2)
> xsg <- group(xset)    # peak alignment (step 3.1)
> xsg <- retcor(xsg)    # retention time correction (step 3.2)
> xsg <- group(xsg)     # re-align (step 3.3)
> xsg <- fillPeaks(xsg)  # filling in missing peak data (step 4)
> dat <- groupval(xsg, "medret", "into")  # get peak intensity matrix (step 5)
> dat <- rbind(group = as.character(phenoData(xsg)$class), dat)  # add group label
> write.csv(dat, file= ‘MyPeakTable.csv’)  # save the data to CSV file
>
```

MISSING VALUES

Control 1	Control 2	Control 3	Treated 1	Treated 2	Treated 3
464	428	483	418	-	425
283	278	303	314	288	318
235	-	277	269	259	217
534	505	-	491	478	-

- Filtering
- MAR & MCAR
- Missing value estimation
- KNN method

FILTERING

- Non-informative features
- Strongly recommended for untargeted metabolomics
- Multiple testing problem
- Interquartile ranges, coefficient of variation

NORMALIZATION

Metabolite	Cancer			Control		
	G1iNS1	G144	G166	G179	CB541	CB660
Acetoacetic acid	4.00	6.00	2.60	2.00	3.00	1.30
Beta alanine	1.90	1.80	2.00	0.95	0.90	1.00
Creatine	2.40	1.80	1.60	1.20	0.90	0.80
Dimethyglycine	12.60	14.40	10.00	6.30	7.20	5.00
Fumaric acide	5.80	3.00	22.40	2.90	1.50	11.20
Glysine	20.00	16.80	44.44	10.00	8.40	22.22
Homocysteine	0.66	0.20	2.12	0.33	0.10	1.06
[...]	...	...	...	...	...	...
TOTAL	2000	1800	2400	1000	900	1200

NORMALIZATION

- Remove systematic variation between experimental conditions unrelated to the biological differences (i.e. dilutions, mass)
- By sum
- By reference compound (creatinine, internal standard)
- By reference sample (e.g. probabilistic quotient normalization)

TRANSFORMATION

- To bring variances of all features close to equal
- Logarithmic transformation
- Box-Cox transformation
- Scaling (e.g. pareto scaling)
- Additional z-score transformation (standardization)

DATA MATRIX

control 1	control 2	.....	control 30	cancer 1	cancer 2	.....	cancer 30
9.249132	9.771213	9.390076	9.395176	8.583321	9.296368	8.821702	7.876008
6.989496	5.84592	6.063214	4.995175	5.143495	5.426189	6.116481	5.011464
4.549009	5.298832	4.028992	4.730776	3.661116	4.268401	4.078334	4.109569
7.042218	7.156791	6.516016	6.4736	6.785386	6.871651	6.612583	6.447812
2.842815	3.210668	3.168886	3.203355	3.055105	3.258568	3.068973	3.149365
6.076624	6.255116	5.53142	7.186467	6.117253	5.925629	6.542273	6.440859
4.001927	4.408226	4.426111	4.218325	4.424755	4.085715	3.99024	4.258238
4.011074	4.147679	3.506027	3.450706	3.771826	3.546628	3.643631	3.816385
6.374999	7.199643	5.660234	8.143042	5.13446	7.064966	7.252155	7.269149
.....	.....	.....	.....	.....	.....	.....	.....
3.710801	3.787264	3.713254	3.393635	3.646768	3.556236	3.573936	3.861748

- $n < p$ ⟶ curse of dimensionality

n: number of samples

p: number of features (e.g.metabolites)

Traditional statistical algorithms are not applicable

CLASS-BASED ANALYSIS

Class comparison

- Which metabolites are statistically significant among :
  - Tumor/non-tumor samples
  - Disease subtypes
  - Stage of the diseases

- What roles do these metabolites have in related pathways?

Class discovery

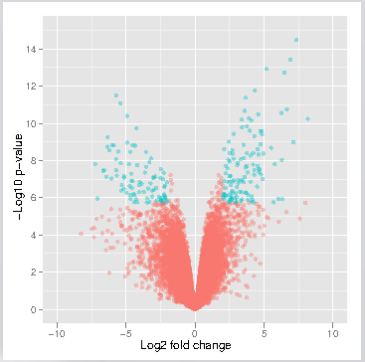
- Can we find metabolite clusters that have similar functions or find sample clusters to detect disease subtypes?

Class prediction

- Can I use this data to predict a phenotype?

CLASS-COMPARISON

- fold-changes
- *t*-tests, ANOVA
- sam (significance analysis of microarrays and metabolites)
- limma (linear models for microarrays and metabolites)
- volcano plots



MULTIPLE TESTING

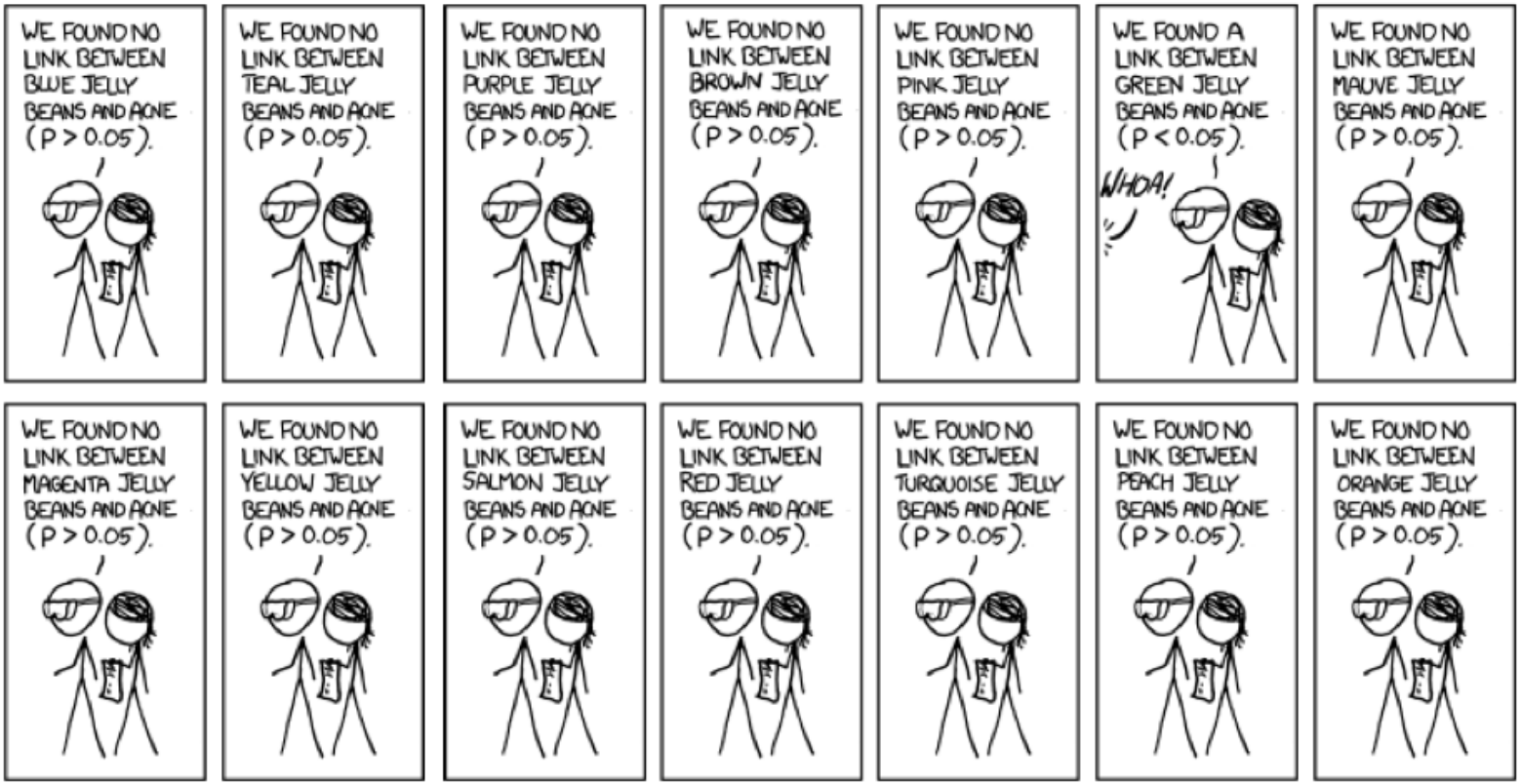
?

Acne

- Which colour of the jelly beans cause to acne?

http://training.bioinformatics.ucdavis.edu/docs/2012/05/RNA/_downloads/the-statistics-of-rna-seq.pdf

MULTIPLE TESTING



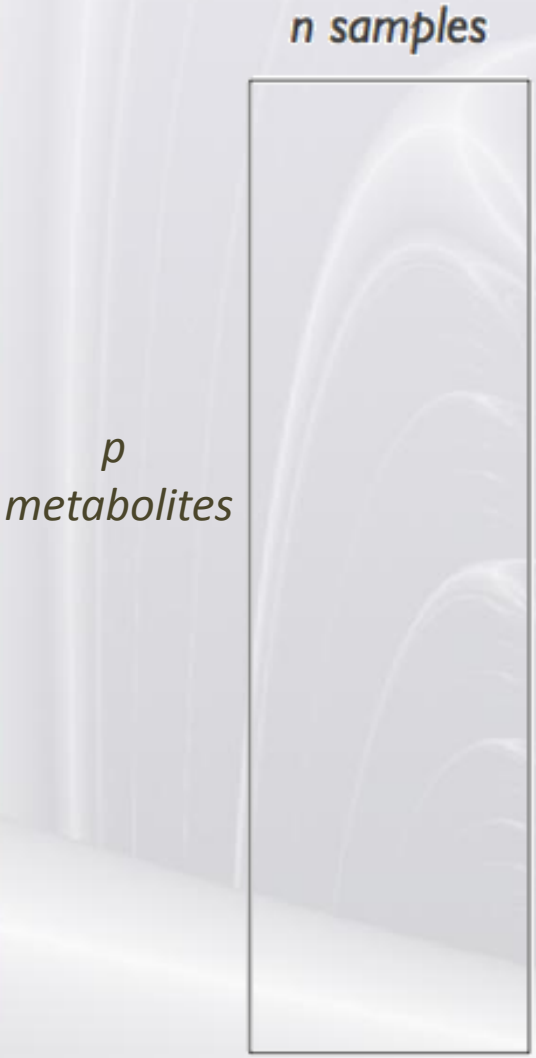
http://training.bioinformatics.ucdavis.edu/docs/2012/05/RNA/_downloads/the-statistics-of-rna-seq.pdf

MULTIPLE TESTING

Why does this matter?

http://training.bioinformatics.ucdavis.edu/docs/2012/05/RNA/_downloads/the-statistics-of-rna-seq.pdf

MULTIPLE TESTING ADJUSTMENT

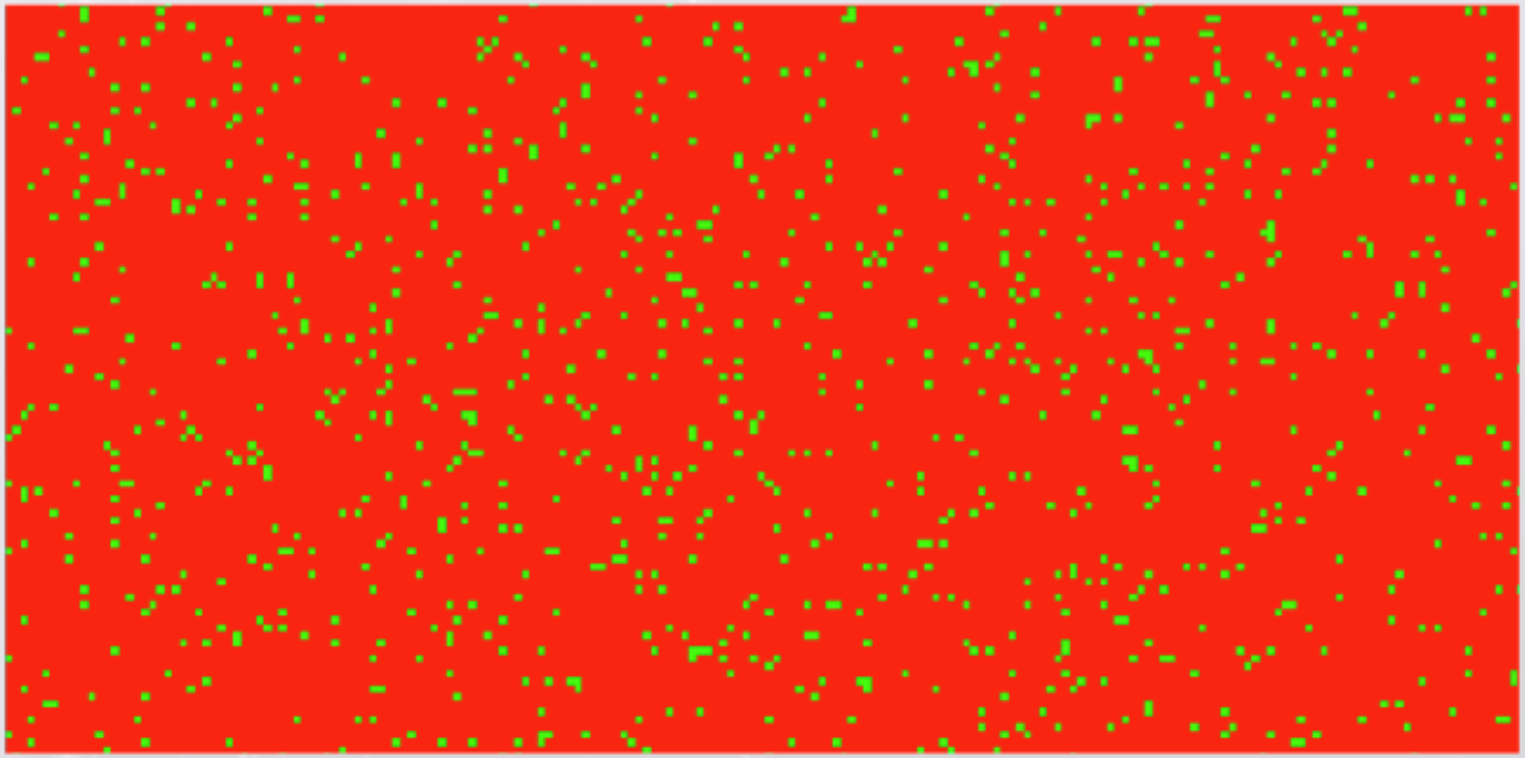


We're doing p simultaneous tests!

$H_1, H_2, H_3, \ldots, H_p$

http://training.bioinformatics.ucdavis.edu/docs/2012/05/RNA/_downloads/the-statistics-of-rna-seq.pdf

MULTIPLE TESTING ADJUSTMENT



20,000 simultaneous t-tests on random normal data from the same distribution. There are 1,009 green points (false positives), making up 0.05 of the comparisons (at $\alpha = 0.05$).

http://training.bioinformatics.ucdavis.edu/docs/2012/05/RNA/_downloads/the-statistics-of-rna-seq.pdf

MULTIPLE TESTING ADJUSTMENT

- *m* metabolites, *m* comparison tests
- Assume we are testing $H_1, H_2, \ldots, H_m$
- m_0: # of true hypotheses, *R*: # number of rejected hypotheses

Decision	Actual Situation		Total
	Null True	Alternative True	
Not called significant	U	T	m - R
Called significant	V	S	R
Total	m_0	$m - m_0$	m

- *V*: # of type I errors (false positives)

BONFERRONI APPROACH

- Simplest method to control type I error
- Reject any hypothesis with p value $\leq \alpha/m$
- Conservative
- Not logical to apply for high dimensional metabolomics data
- If we have 1000 features to compare, only features which have p values less than 0.05 / 1000 = 0.00005 are statistically significant.

BENJAMINI and HOCHBERG FDR

- FDR (False discovery rate) = V / R
- FDR is designed to control the proportion of false positives among the set of rejected hypotheses (R)

Decision	Actual Situation		Total
	Null True	Alternative True	
Not called significant	U	T	m - R
Called significant	V	S	R
Total	m_0	$m - m_0$	m

BENJAMINI and HOCHBERG FDR

- To control FDR at level δ:

1. Order the unadjusted *p*-values: $p_1 \leq p_2 \leq \ldots \leq p_m$
2. Then find the test with the highest rank *j*, for which the *p* value p_j, is less than or equal to $(j/m)x\delta$
3. Declare the tests of rank 1,2, ..., *j* as significant

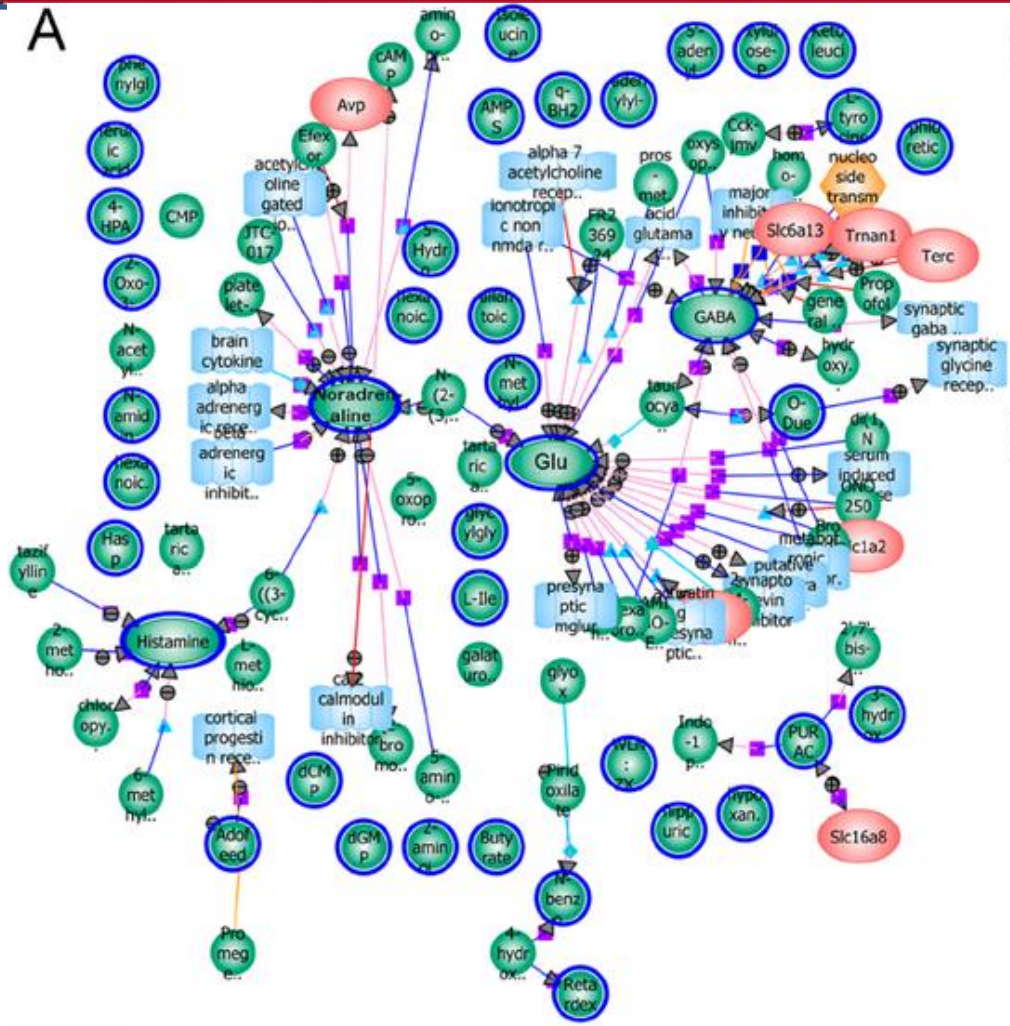
$$p(j) \leq \delta \frac{j}{m}$$

BENJAMINI and HOCHBERG FDR

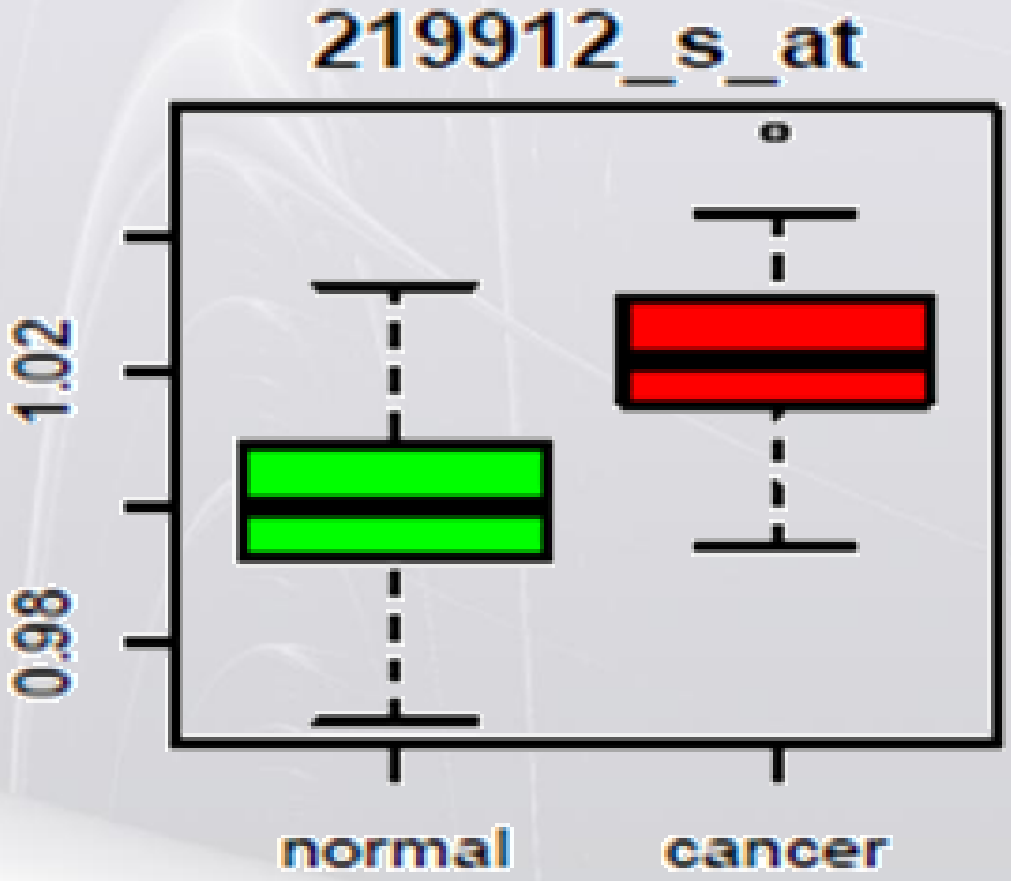
Controlling the FDR at $\delta = 0.05$

Rank (j)	P-value	(j/m)× δ	Reject H_0 ?
1	0.0008	0.005	1
2	0.009	0.010	1
3	0.165	0.015	0
4	0.205	0.020	0
5	0.396	0.025	0
6	0.450	0.030	0
7	0.641	0.035	0
8	0.781	0.040	0
9	0.900	0.045	0
10	0.993	0.050	0

PATHWAY ANALYSIS

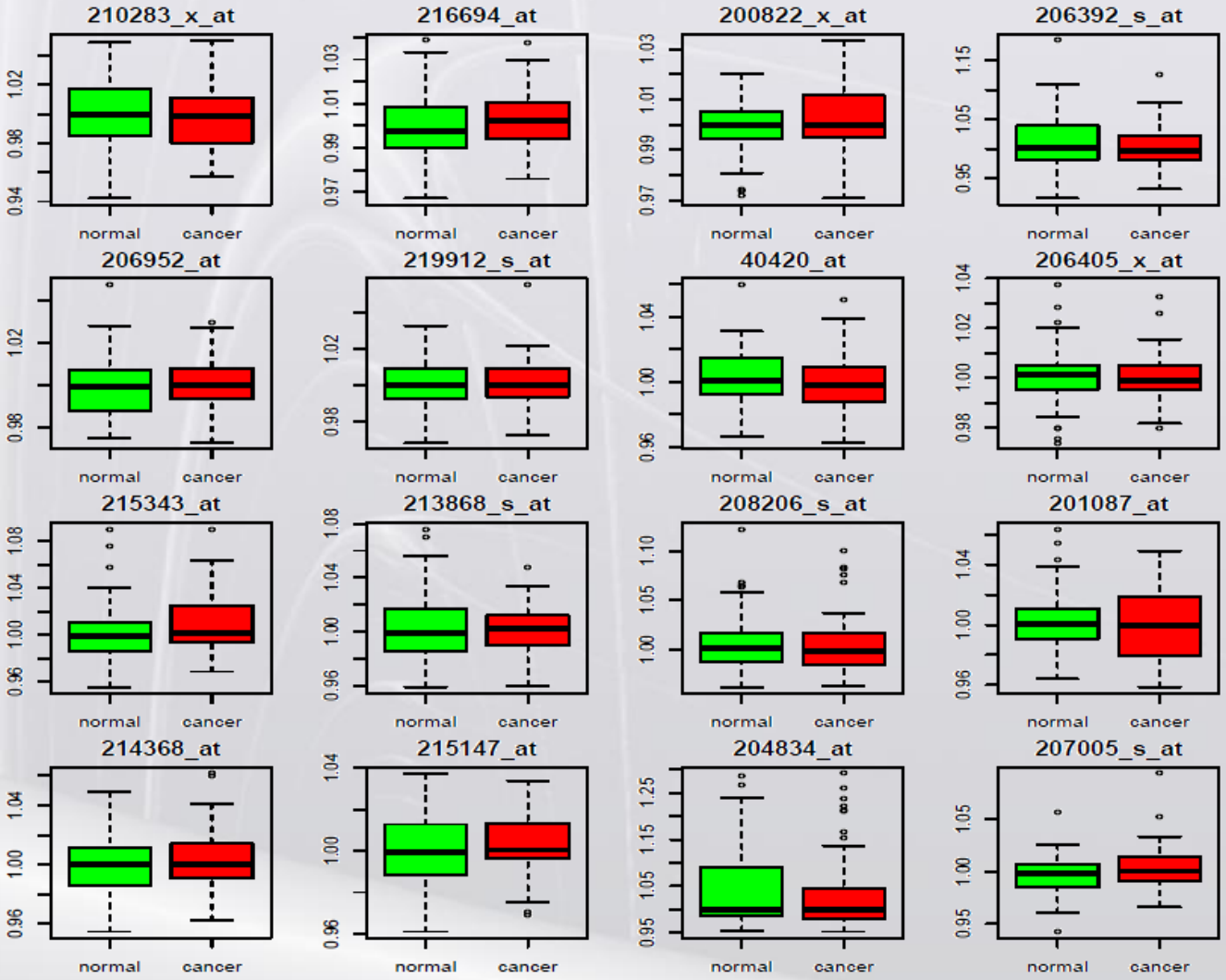


CLASSIFICATION



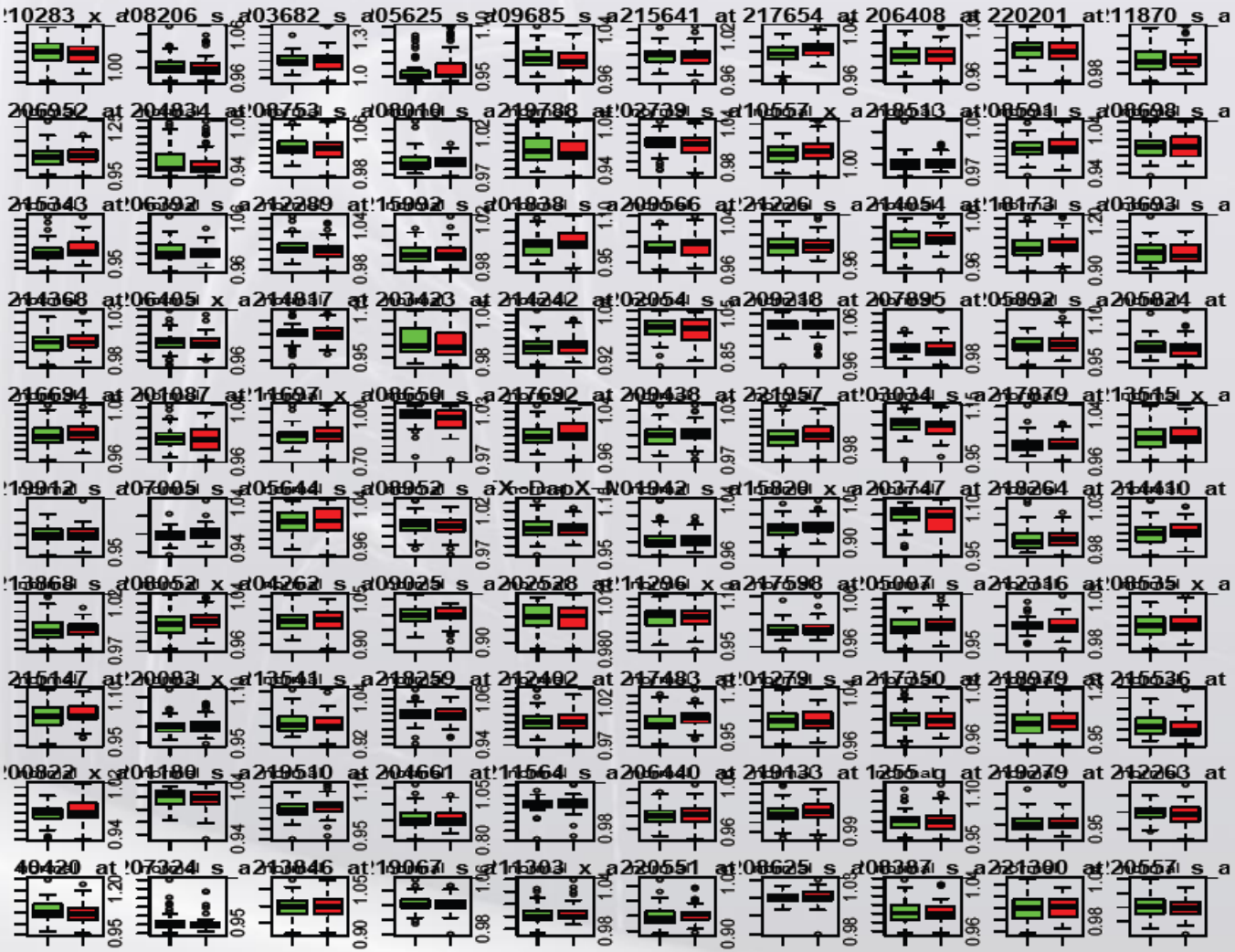
Chen G (2013). Machine Learning for Bioinformatics using R. BGI Bioinformatics Workshop

CLASSIFICATION



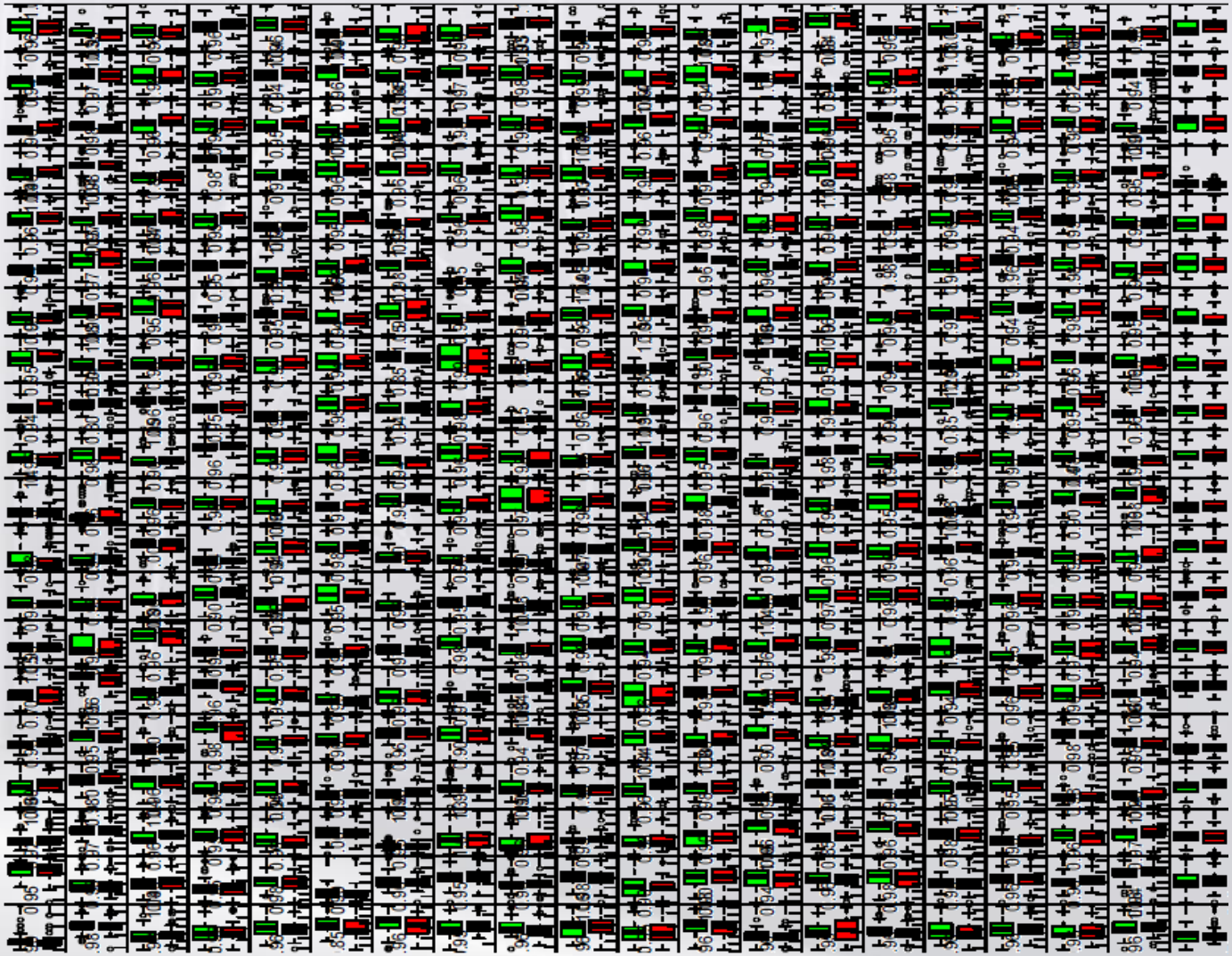
Chen G (2013). Machine Learning for Bioinformatics using R. BGI Bioinformatics Workshop

CLASSIFICATION



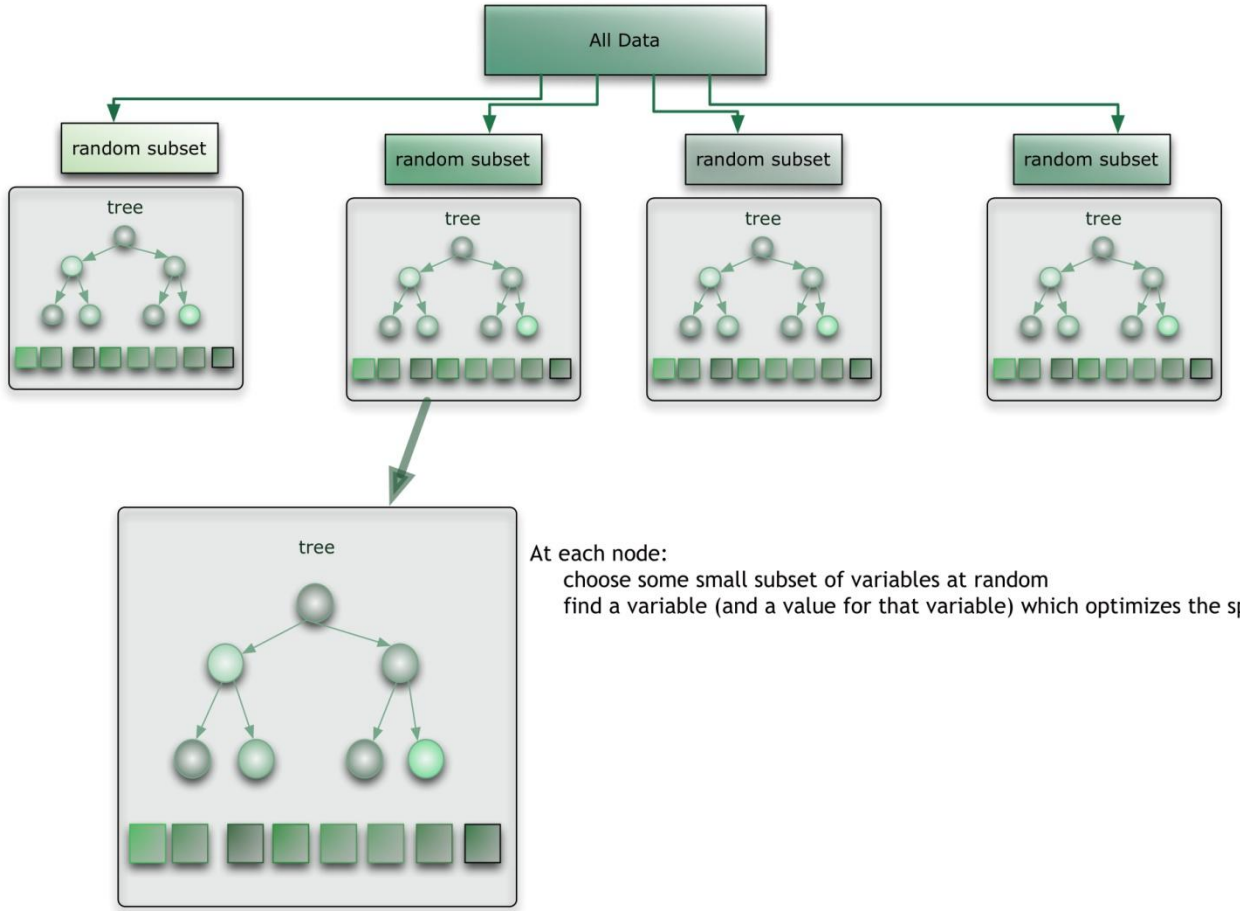
Chen G (2013). Machine Learning for Bioinformatics using R. BGI Bioinformatics Workshop

CLASSIFICATION



Chen G (2013). Machine Learning for Bioinformatics using R. BGI Bioinformatics Workshop

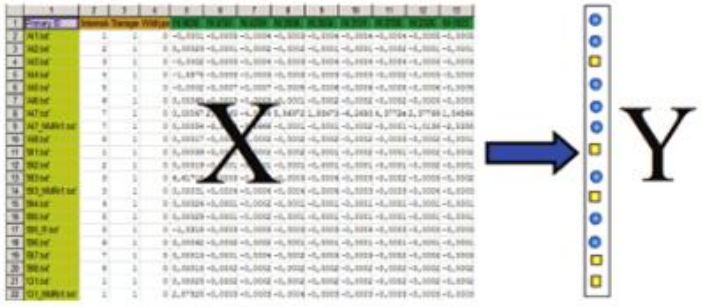
RANDOM FORESTS

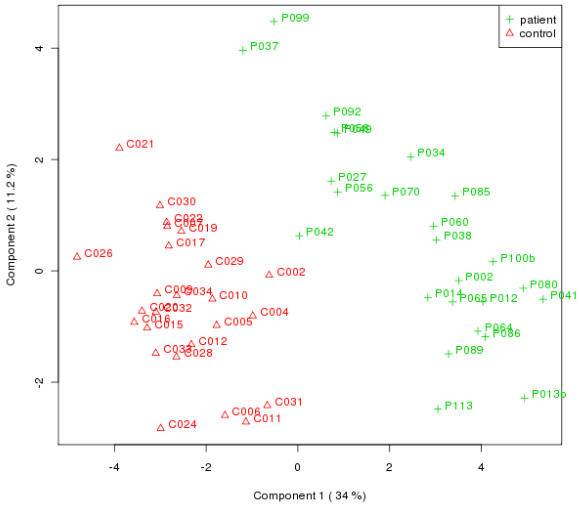


At each node:
choose some small subset of variables at random
find a variable (and a value for that variable) which optimizes the split

PLS-DA

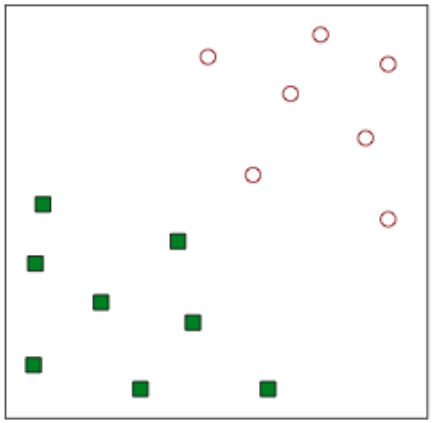
- *De facto* standard for chemometric analysis
- A supervised method that uses multiple linear regression technique to find the direction of maximum covariance between a data set (X) and the class membership (Y)



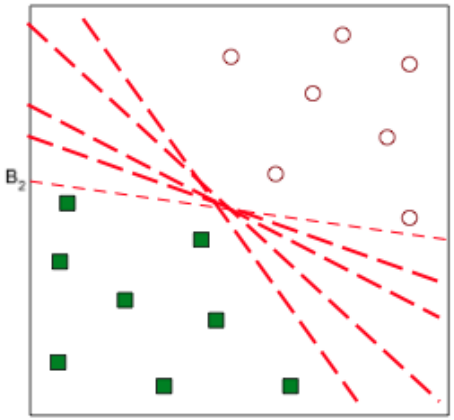


SUPPORT VECTOR MACHINES

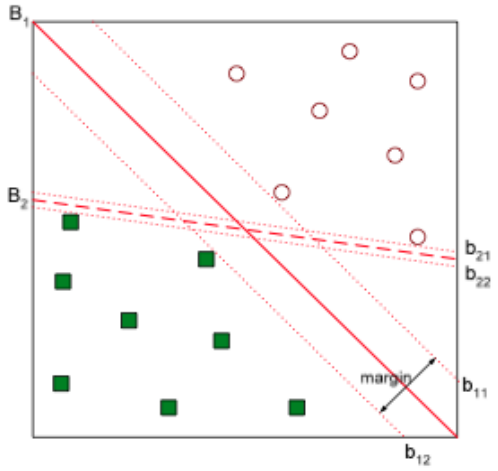
1



2

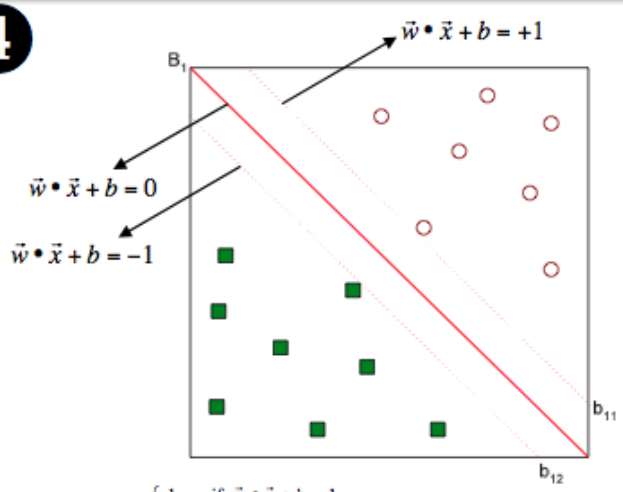


3



Tan P.N. et al. (2005)

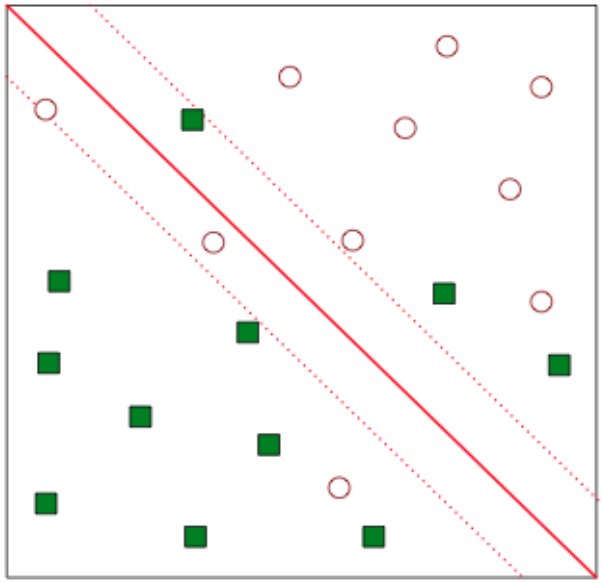
4

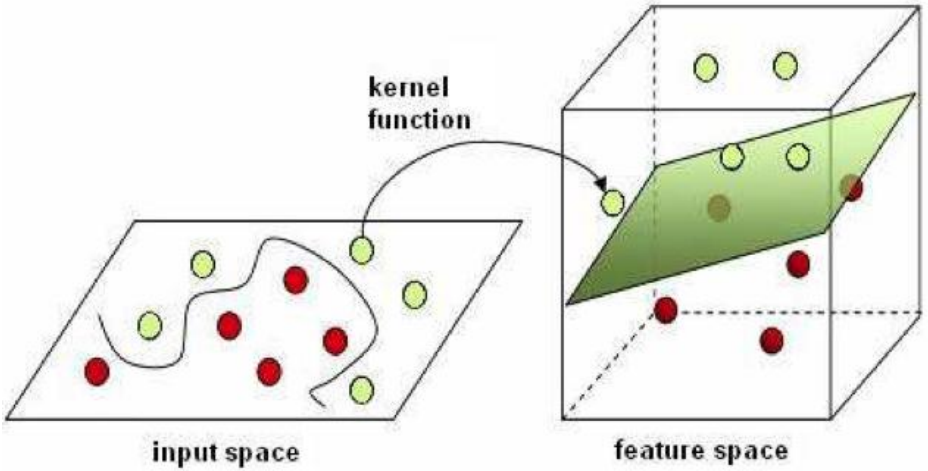


$$f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x} + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x} + b \leq -1 \end{cases}$$

$$\text{Margin} = \frac{2}{\|\vec{w}\|^2}$$

SUPPORT VECTOR MACHINES



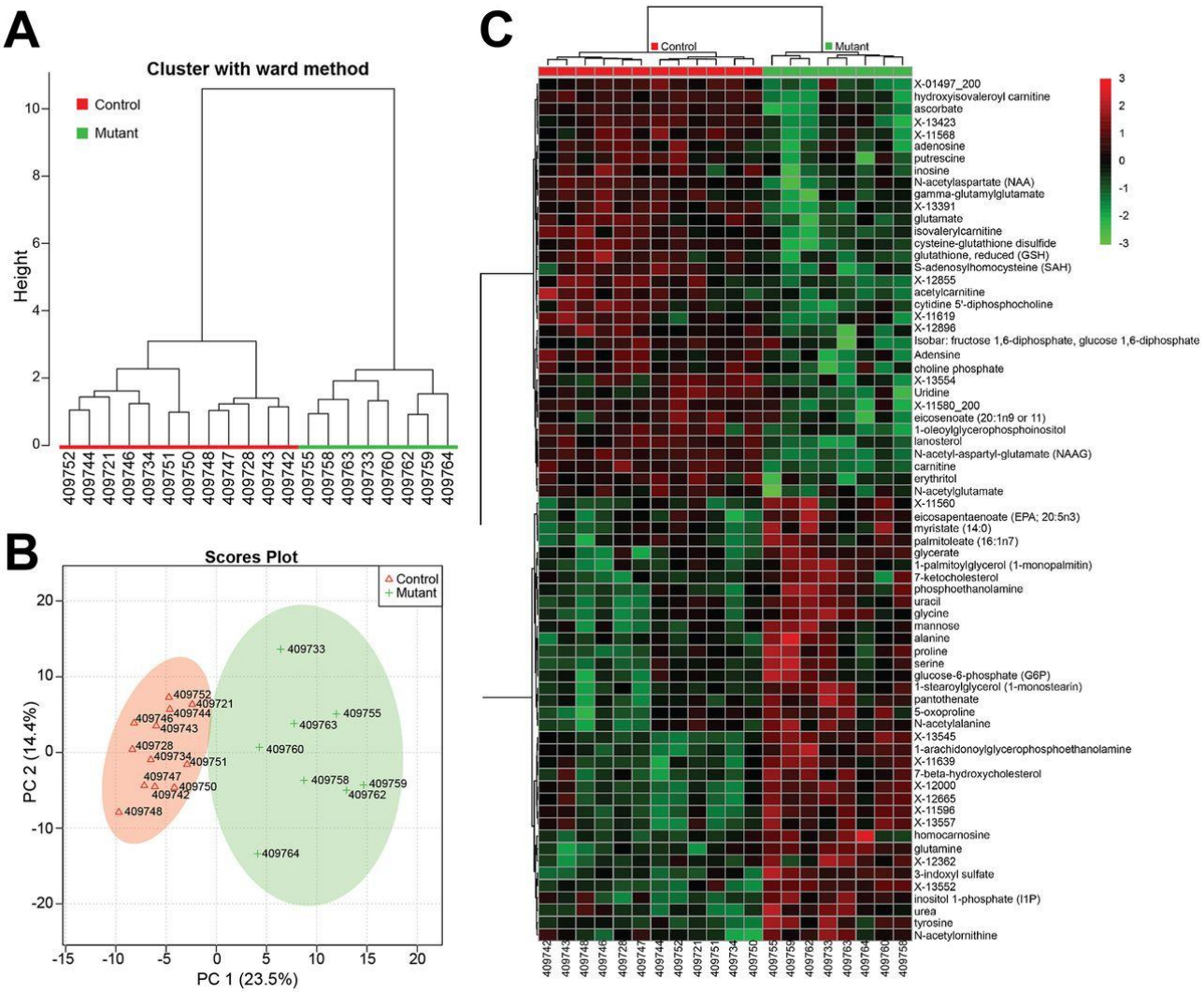


Slack variables (ξ) allow misclassifying some points

Kernel functions (radial-based, polynomial, etc.) are used for nonlinear classification

CLUSTERING

GÖKMEN ZARARSIZ



Bais et al. (Biology Open, 2016)

REFERENCES

1. Bais et al. (2016) Metabolite profile of a mouse model of Charcot–Marie–Tooth type 2D neuropathy: implications for disease mechanisms and interventions. *Biology Open*, doi: 10.1242/bio.019273.
2. Benjamini and Hochberg (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 289-200.
3. Breiman L (2001) Random forests. *Mach. Learn*. 45, 5–32.
4. Cortes and Vapnik (1995) Support vector networks. *Mach. Learn*. 20(3), 273-297.
5. Dieterle et al. (2006) Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in NMR metabonomics. *Anal. Chem*. 78, 4281–4290.
6. Dunn and Ellis (2005) Metabolomics: current analytical platforms and methodologies. *Trends Analyt. Chem*. 24, 285–294.
7. Ewald JC (2009) High-throughput quantitative metabolomics: workflow for cultivation, quenching, and analysis of yeast in a multiwell format. *Anal. Chem*. 81, 3623–3629.
8. Fiehn O (2002) Metabolomics—the link between genotypes and phenotypes. Plant. Mol. Biol. 48, 155–171.
9. Kuhn M (2008) Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5).
10. Patti et al. (2012) Innovation: Metabolomics: the apogee of the omics trilogy. Nature Reviews
11. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical *Applications in Genetics and Molecular Biology*, 3(3).
12. Smith CA et al. (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem*. 78, 779–787.
13. Tusher VG et al (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci*. USA 98, 5116–21.
14. van den Berg et al. (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* 7, 142.
15. Wang T. et al. (2009) Automics: an integrated platform for NMR-based metabonomics spectral processing and data analysis. *BMC Bioinformatics* 10, 83.
16. Wishart DS (2007) Current Progress in computational metabolomics. *Brief. Bioinform*. 8, 279–293.
17. Wishart DS (2008) Quantitative metabolomics using NMR. *Trends Analyt. Chem*. 27, 228–237.
18. Xia J et al. (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res. 37,* W652–W66.
19. Xia and Wishart (2010) MSEA: A web-based tool to identify biologically meaningful patterns in quantitative metabolomics data. *Nucleic Acids Res*. 38, W71–W77.
20. Zararsiz G, et al. (2014) MLSeq: Machine learning interface for RNA-Seq data. *R package version 1.1.5.*