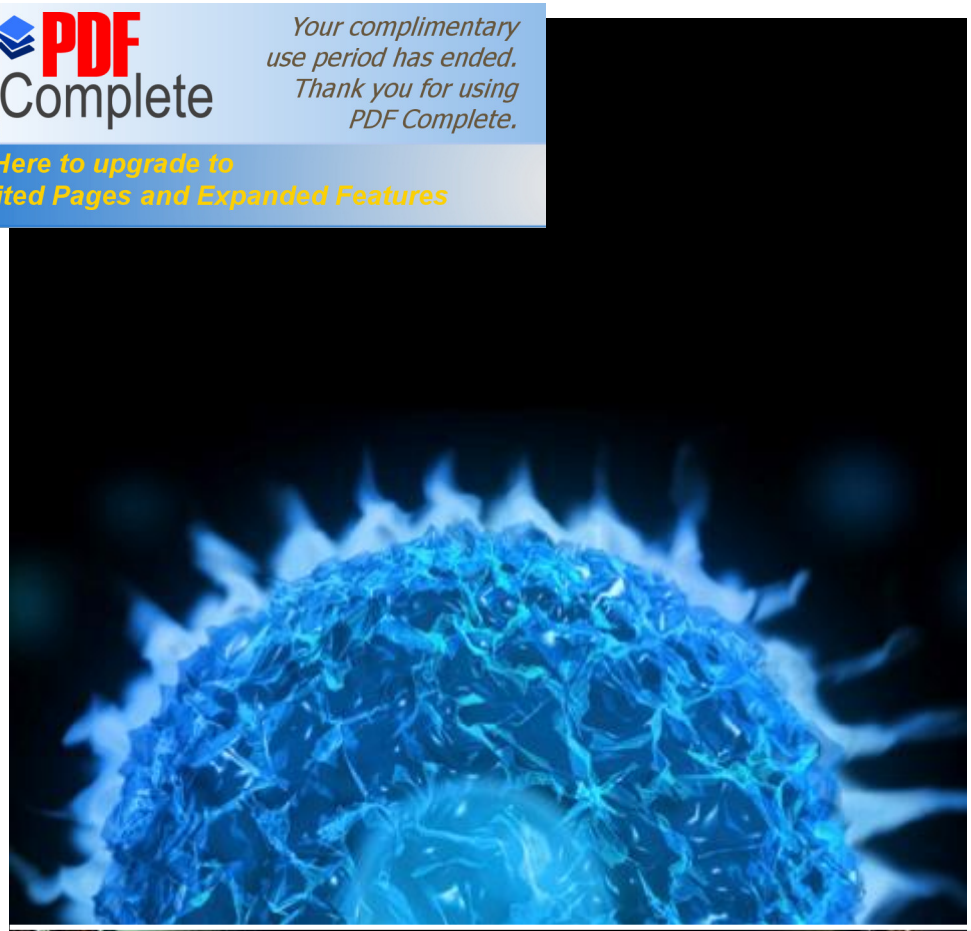


CORRESPONDENCE ANALYSIS APPLICATION IN CLASS COMPARISON STUDIES

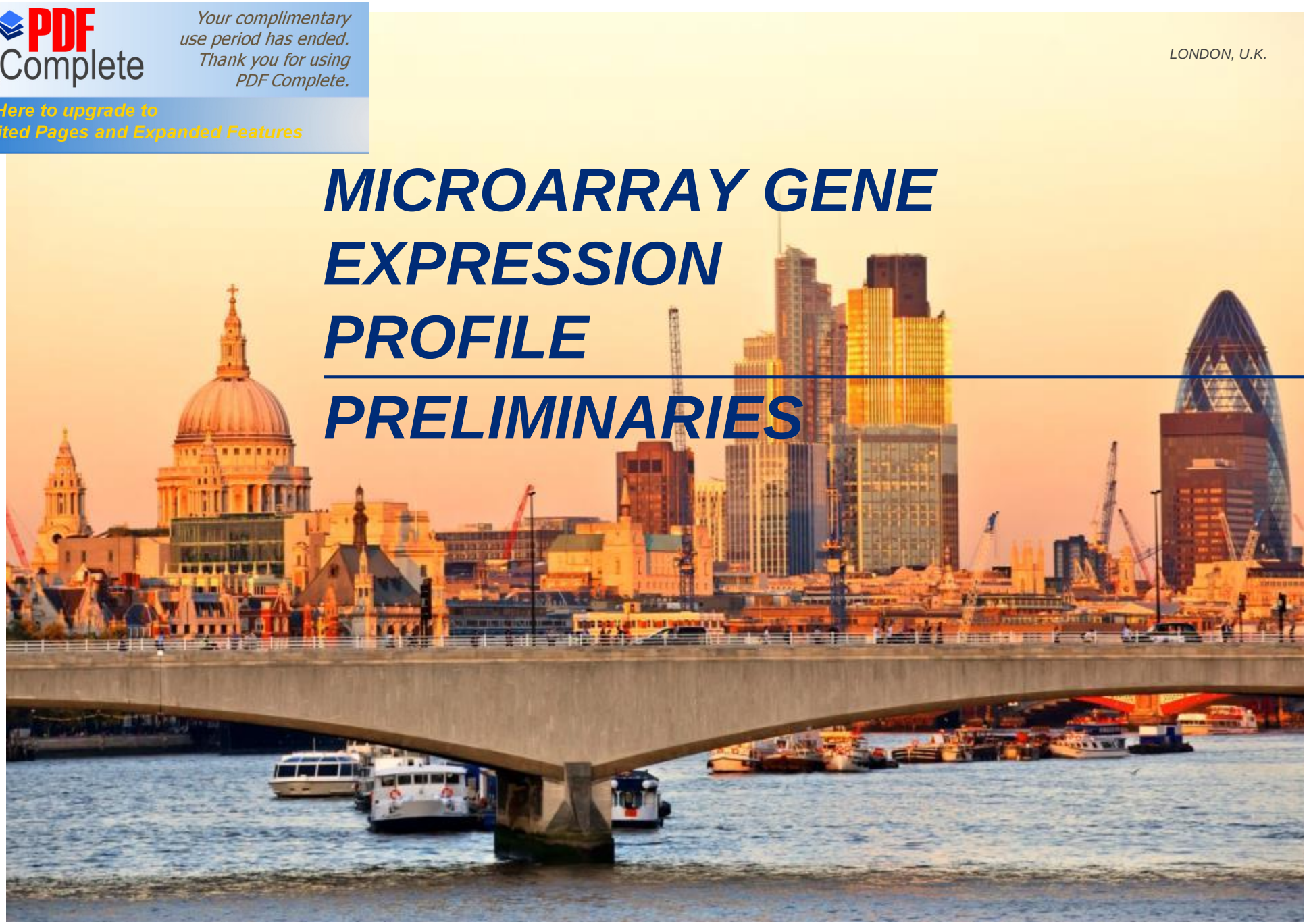
PhUSE London 2014

Edyta Winciorek

PAREXEL
YOUR JOURNEY. OUR MISSION.™

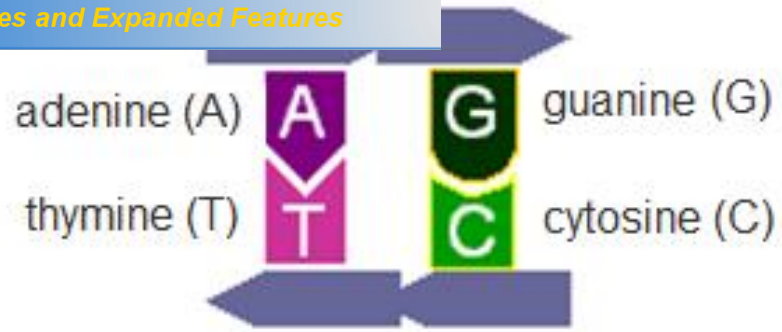
FROM GENE EXPRESSION TO EXPRESSION CARTOGRAPHY

1. MICROARRAY GENE EXPRESSION PROFILE ï PRELIMINARIES
2. TYPES OF DNA MICROARRAY EXPERIMENTS
3. EXAMPLE - CORRESPONDENCE ANALYSIS APPLICATION IN CLASS COMPARISON STUDIES
4. CONCLUSION



MICROARRAY GENE EXPRESSION PROFILE

PRELIMINARIES

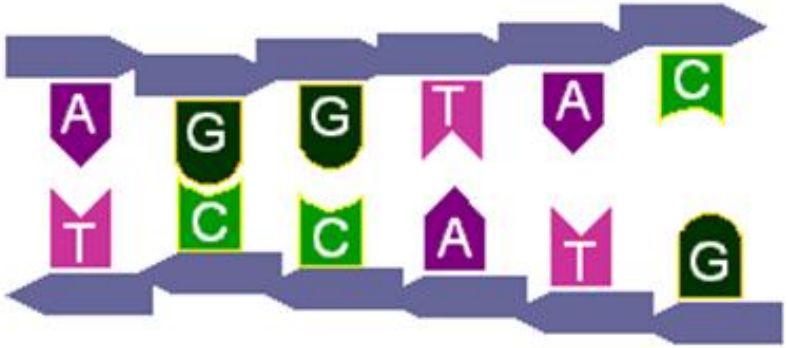


DNA double-stranded structure is formed through the hybridization of these complementary single stranded chains.

HYBRIDIZATION

Hybridization is the process of joining two complimentary strands of DNA or one of DNA and RNA to form a **double-stranded** molecule.

Given these characteristics, it is possible to detect a target gene by hybridizing it with DNA that has a complementary sequence.

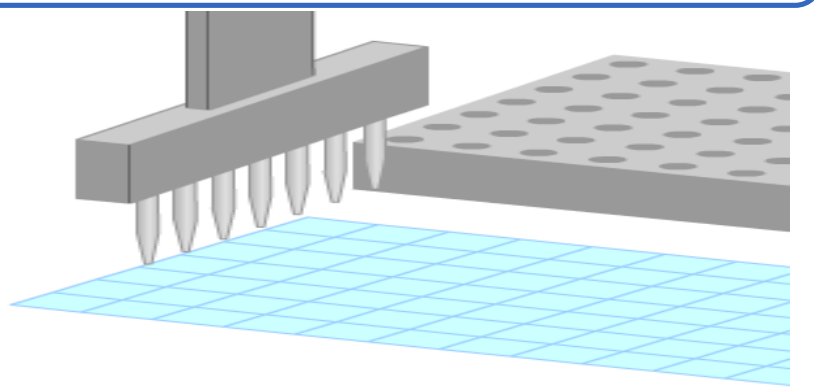


Source: http://www.olympus-global.com/en/news/2000b/nr000926oligoe.jsp

...RNA from the ...lls

Reverse transcribed mRNA to given unique cDNA population

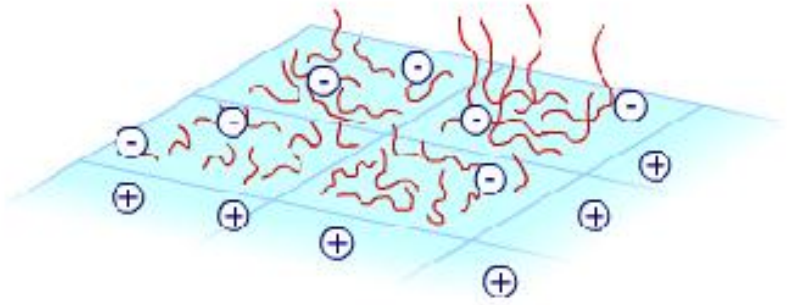
Embed population onto specially coated glass slides



The slides are coated with positively charged polysine. DNA is negatively charged, so that cDNA stick to the slide through an ionic interaction.

DNA ARRAY

glass slide where **single-stranded** DNA (called **probe)** with various sequences are printed on the surface of the substrate in a localized features that are arranged in a grid-like pattern.



Source*: http://www.dnalc.org/view/15992-DNA-microarrays.html*

PARATION AND LABELLING

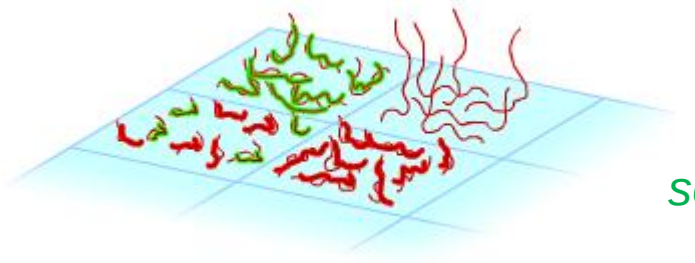
The cDNA from ill tissue (e.g. tumor) are labelled with red fluorescent tag Cyanine 5 (Cy5).

The cDNA from normal (reference) tissue are labelled with green tag Cyanine 3 (Cy3).

Tagged cDNA arrays are incubated, which bound the matching genes printed on the arrays.

If the gene is expressed in both cells, the sequence is yellow.

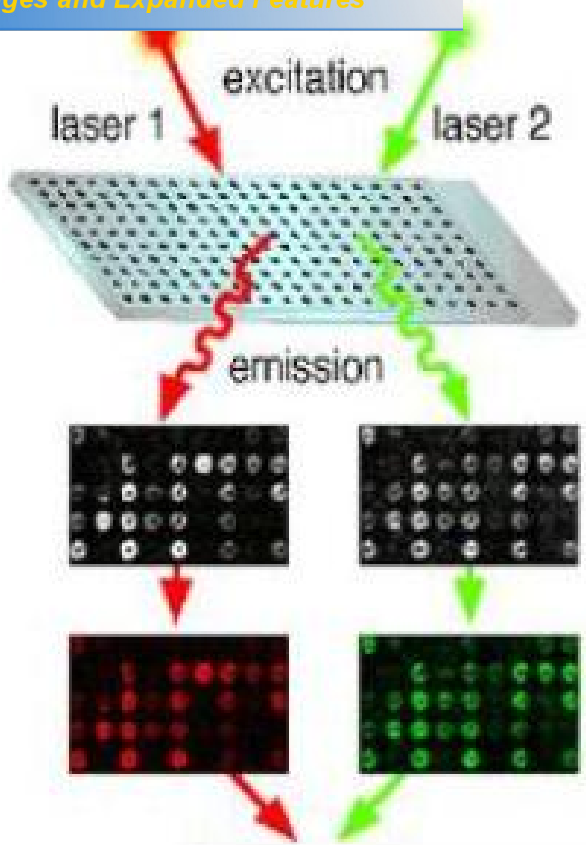
If the gene is expressed only in tumour cells, the sequence is red.



If the gene is expressed only in normal cells, the sequence is green.

Source: Jon Pollack 2011: *Microarrays and Analysis of Hybridization Data* , Genomic Medicine.

SION



Cy3	Cy5	Cy5/Cy3	log2(Cy5/Cy3)
200	10000	50.00	5.64
4800	4800	1.00	0.00
9000	300	0.03	-4.91

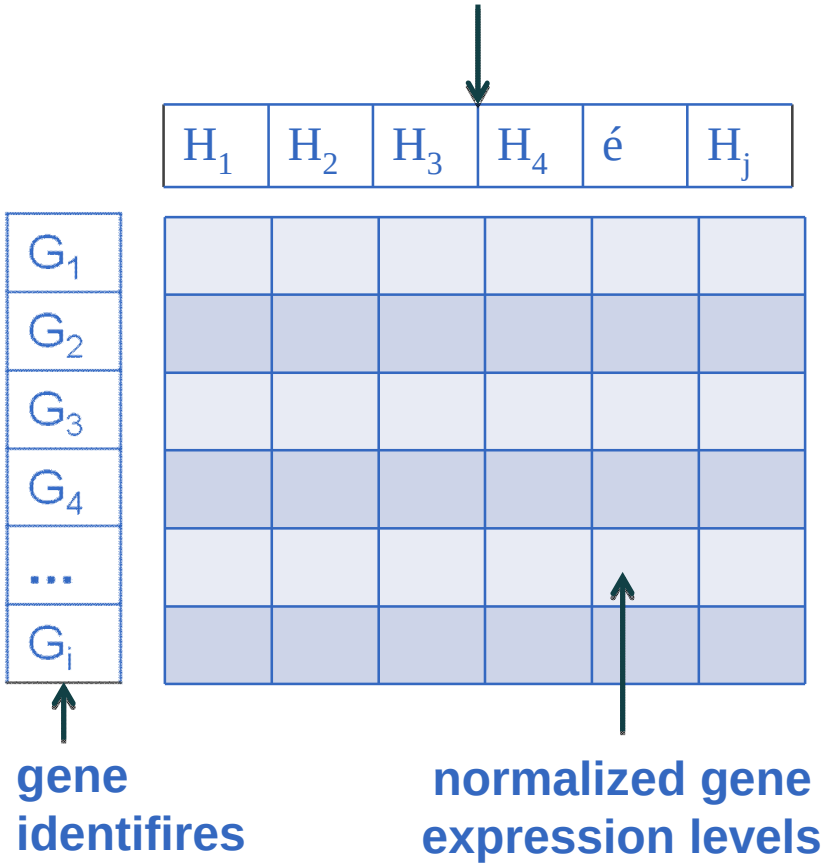
The next step after hybridization is to generate an image using laser-induced fluorescent imaging.

The amount of fluorescence measured at each sequence specific location is directly proportional to the amount of mRNA with complementary sequence present in the sample analyzed

transformed into numbers and will be the basis the statistical

Source: Jon Pollack 2011: *Microarrays and Analysis of Hybridization Data* , Genomic Medicine.

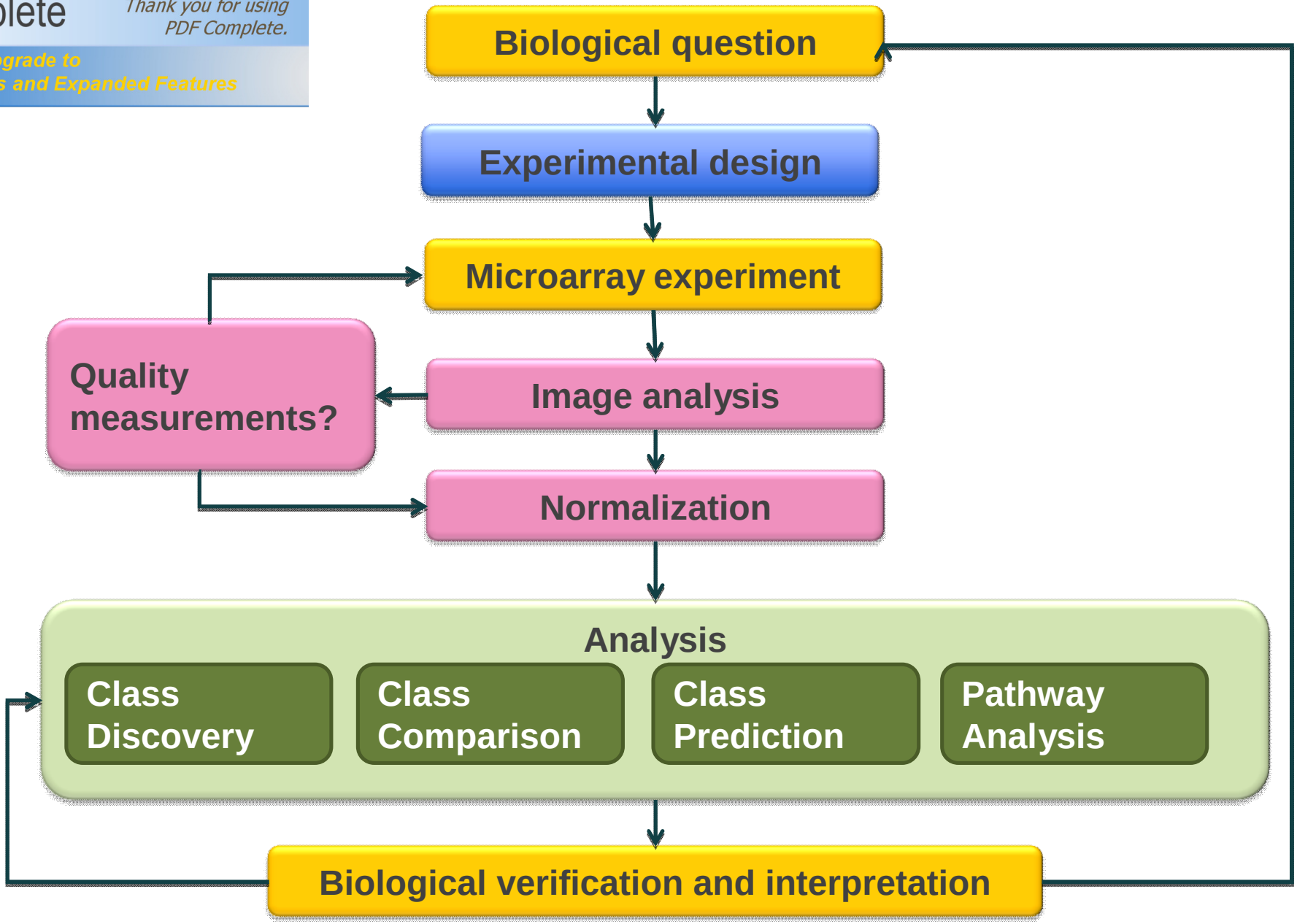
notation



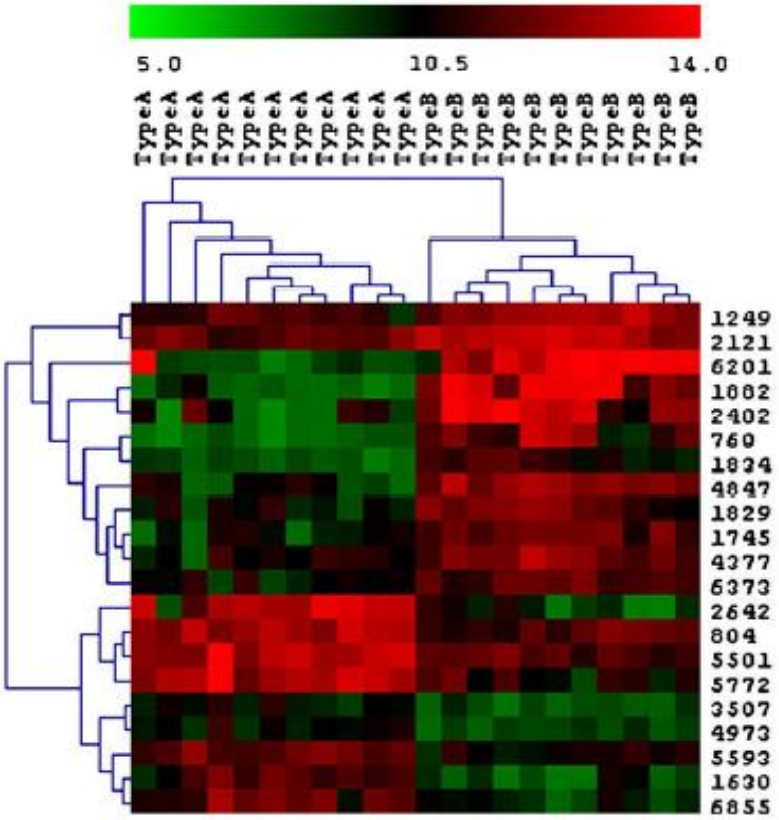
GENE EXPRESSION MATRIX

I genes and J hybridizations are collected into the IxJ matrix N with elements n_{ij} - the gene expression level for each gene G_i in hybridizations H_j.

Gene expression matrix needs to be preprocessed, for example the logarithm of the raw intensity values is taken or normalization of data is performed.



VERY (CLUSTERING)



Discovery of co-regulated groups of genes of 2 types of patients A and B.

Source: Tarca, A. L., Romero, R., Draghici, S. (2006). Analysis of microarray experiments of gene expression profiling. American journal of obstetrics, 195, no. 2, 373ï 388.

- Unsupervised machine learning method such as hierarchical clustering, k-means clustering or self-organizing maps.
- Identification of the genes that are similarly expressed.
- Detection of spatial or temporal expression patterns.
- Dimension reduction of the gene expression matrix.

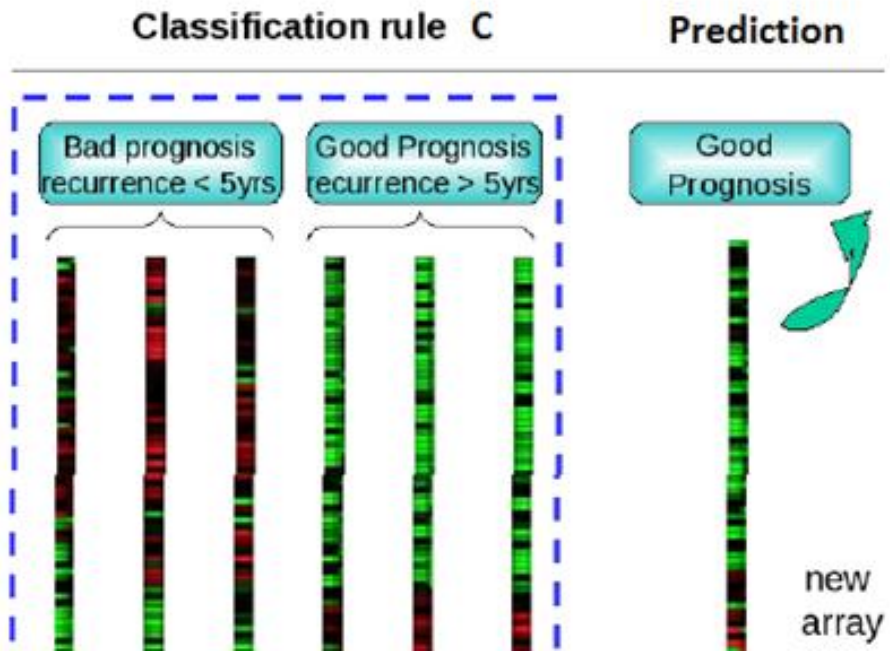
CTION (CLASSIFICATION)

Samples class membership basing on its gene expression profile using supervised machine learning method such as discriminant analysis.

Determine mathematical model well describing the classification rule used to distinguish the pre-defined classes.

Estimate the parameters of the mathematical function used in this model.

Estimate the accuracy of the predictor.

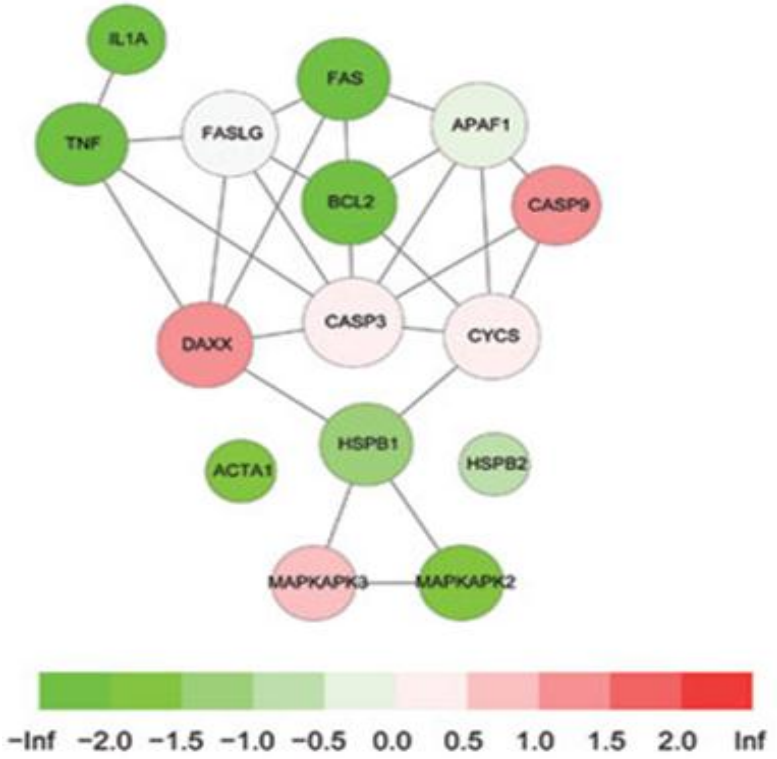


Class Prediction example: assignment of type to a new sample of gene expression matrix.

Source: Sánchez, A. and Ruíz de Villa, M. (2008). *A Tutorial Review of Microarray Data Analysis*

ALYSIS

tation of the list of genes selected in microarray analysis experiment.



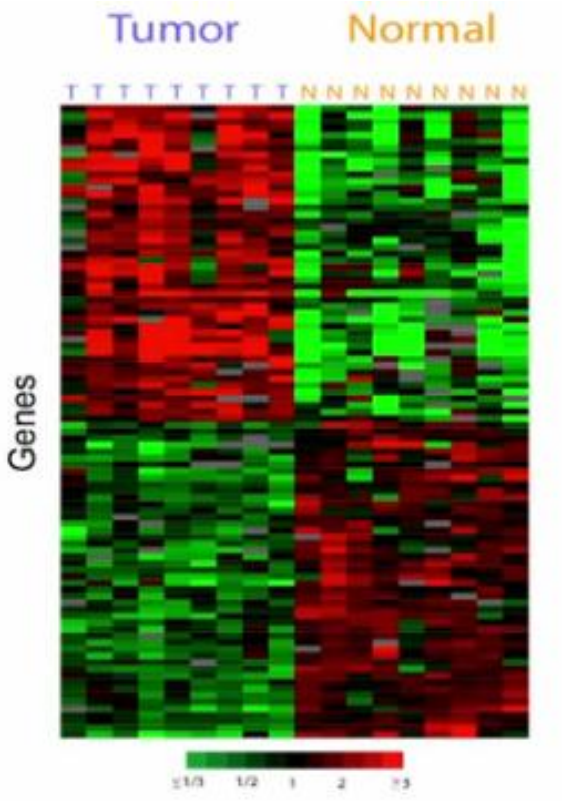
Gene functional association network: Node color reflects expression changes in microarray. Red: upregulated expression. Green: downregulated expression. Gray: not covered by microarray data.

Gene functional association networks for selected pathways.

Source: Zhaoyuan Fang, Weidong Tian and Hongbin Ji 2011: *A network-based gene-weighting Approach for pathway analysis.*

...ARISON (*DIFFERENTIAL ANALYSIS)*

...e gene expression levels of gens between groups of patients using such methodology as Student's t-test, ANOVA, survival analysis, PCA, Correspondence Analysis. .



Distinction of Tumor vs. Normal 1 000 genes .

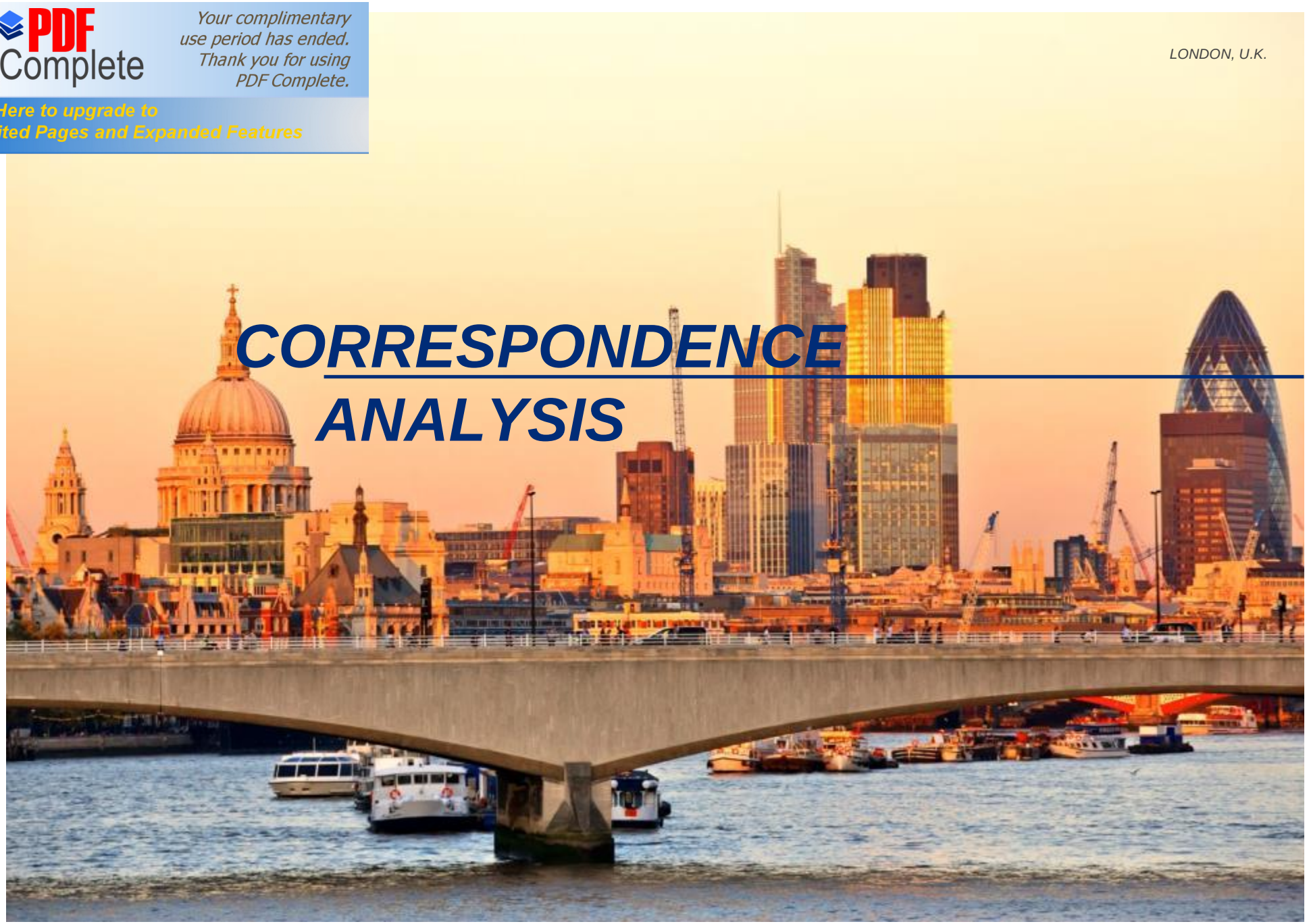
null hypothesis	*alternative hypothesis*
• given gene on the array is not differentially expressed between the two conditions under study	• the expression level of that gene is different

Student's t-test: $\frac{\bar{X}_T - \bar{X}_N}{\sqrt{var_T/n_T - var_N/n_N}}$

50 genes identified at P<0.05

Are there significant?

Source: Jon Pollack 2011: *Microarrays and Analysis of Hybridization Data* , Genomic Medicine

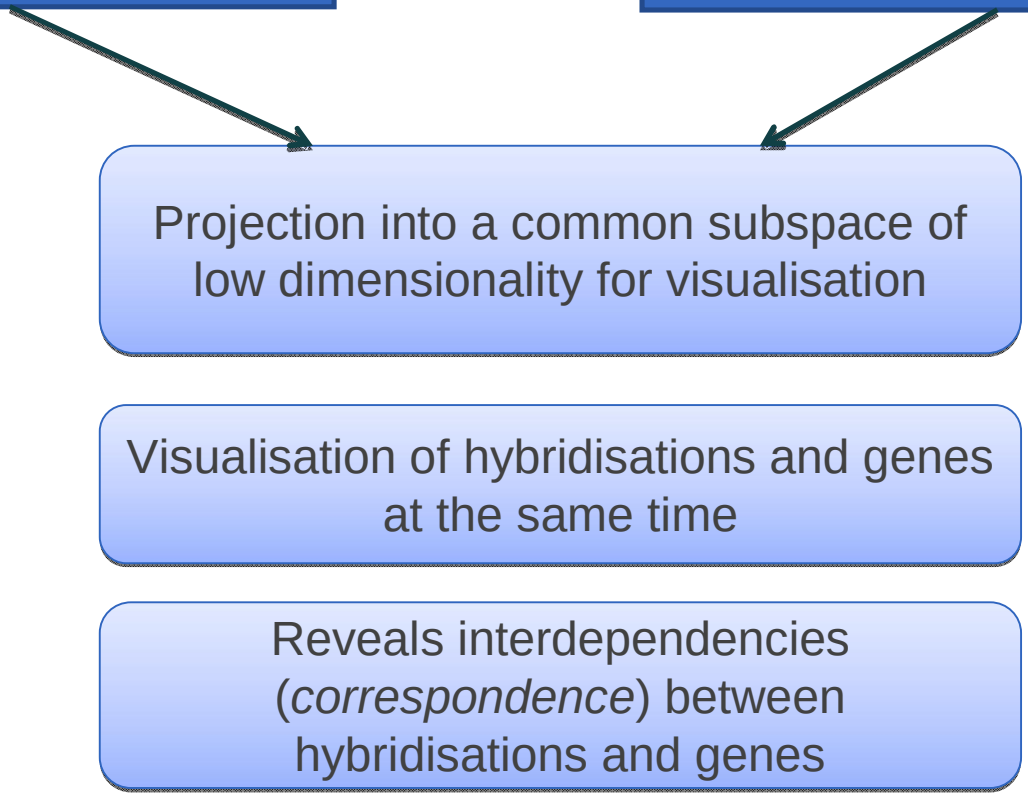


CORRESPONDENCE ANALYSIS

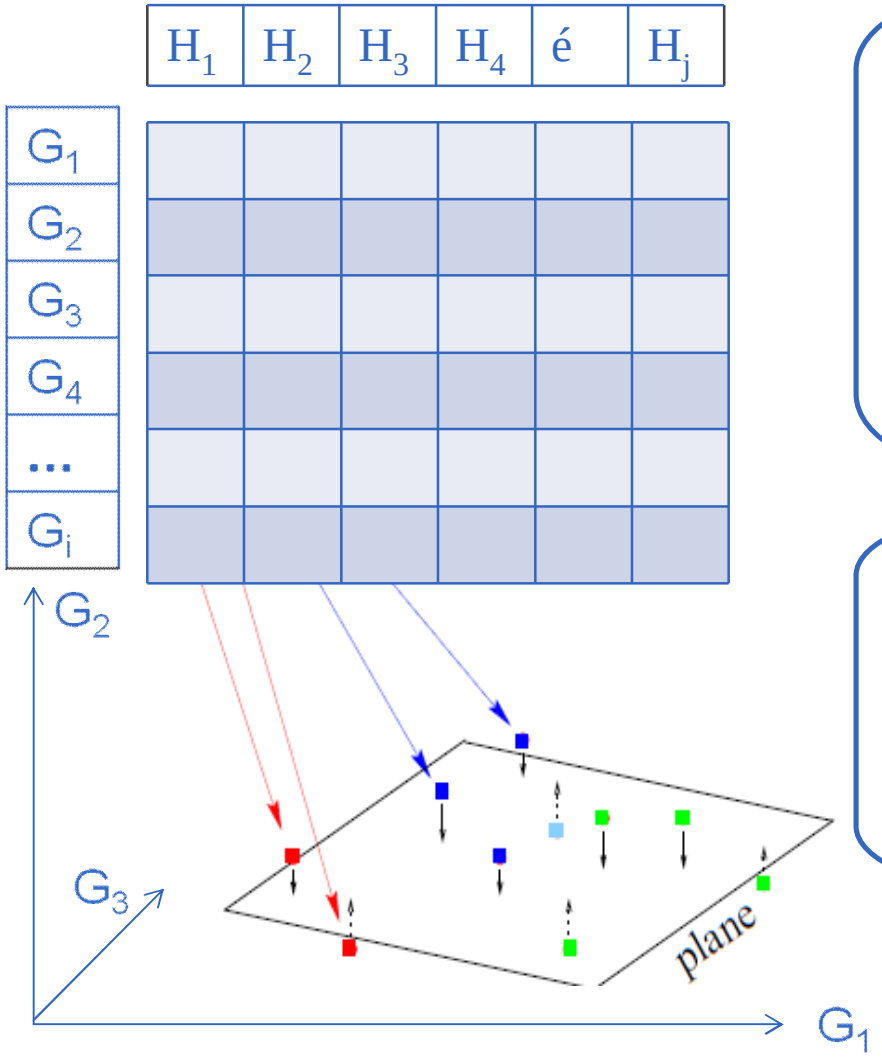
ENCE ANALYSIS

_n_ gene profiles: vectors in _m_-dimensional experiment space

_m_ hybridisation profiles: vectors in _n_-dimensional experiment space



ENCE ANALYSIS



The hybridizations are represented in *n*-dimensional gene space (here n=3). The plane is selected such that the distance of hybridization vectors to the plane is minimal, thus conserving point-to-point distances among those vector points as well as possible.

Genes and tissues are typically classified using correlations of gross expression level. The net relationship between a pair of genes may be measured by partial correlation.

SP

```
ons > ;
TABLES <row-variables, > column-variables ;
VAR variables ;
BY variables ;
ID variable ;
SUPPLEMENTARY variables ;
WEIGHT variable ;
```

TABLES statement instructs PROC CORRESP to create a contingency table from raw, categorical data.

VAR statement instructs PROC CORRESP to read an existing contingency table.

ID statement (only with VAR statement) labels the rows of the tables with the ID values and places the ID variable in the output data set.

BY statement separate analyses on observations in groups defined by the BY variables.

SUPPLEMENTARY statement specifies variables that are to be represented as points in the joint row and column space but that are not used in determining the locations of the other, active row and column points of the contingency table.

WEIGHT Statement specifies weights for each observation and indicates supplementary observations for simple correspondence analyses with VAR statement input.

...n et al. (1999): series of 62 Affimetrix ...riments upon normal (N) and cancerous (T) colon tissue.

Correspondence Analysis - plangoa

The CORRESP Procedure

Inertia and Chi-Square Decomposition

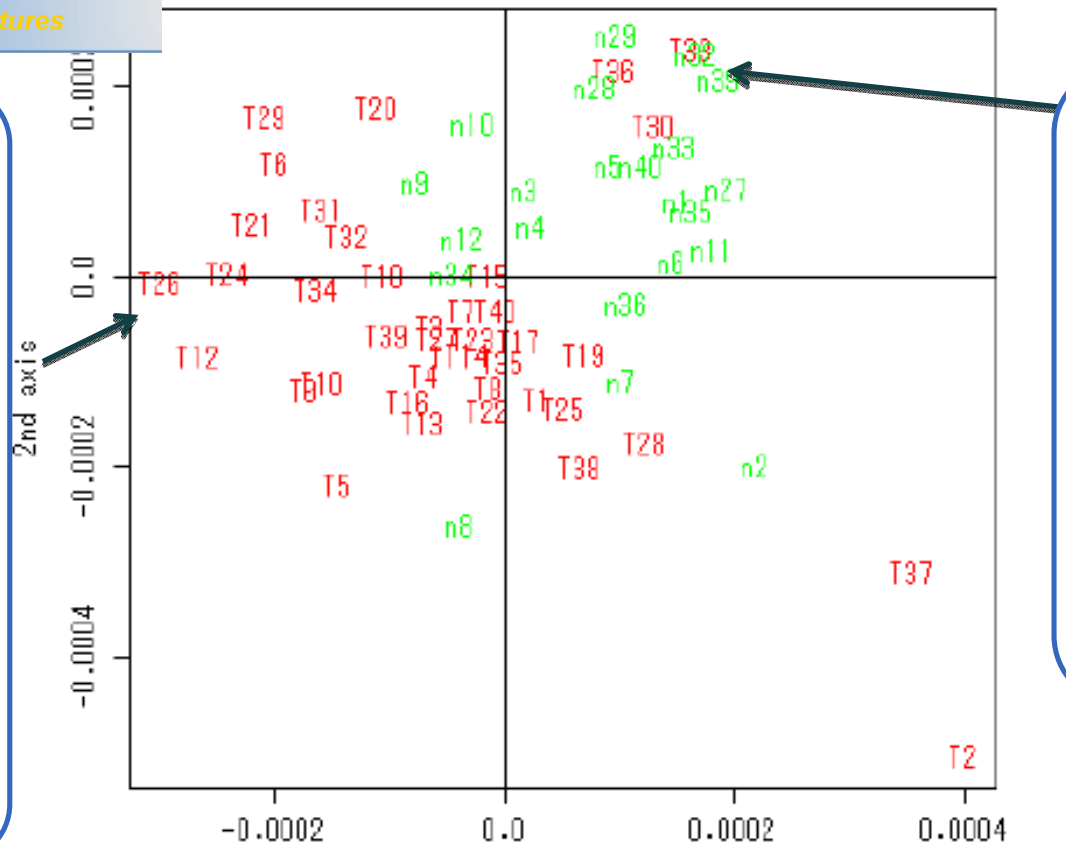
Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	11 22 33 44 55
0.42252	0.17852	2668.35	53.27	53.27	**********
0.39571	0.15658	2340.44	46.73	100.00	**********
Total	0.33510	5008.79	100.00		

Degrees of Freedom = 6342

The total $\chi 2$-statistic, which is a **measure of the association between the rows and columns** is 5008.79 and is explained equally for both the dimensions ï i.e. about 53.27% explain Dimension 1 and 46.73% explain Dimension 2. This indicates that the association between the row and column categories is essentially two dimensional.

T OF GENES

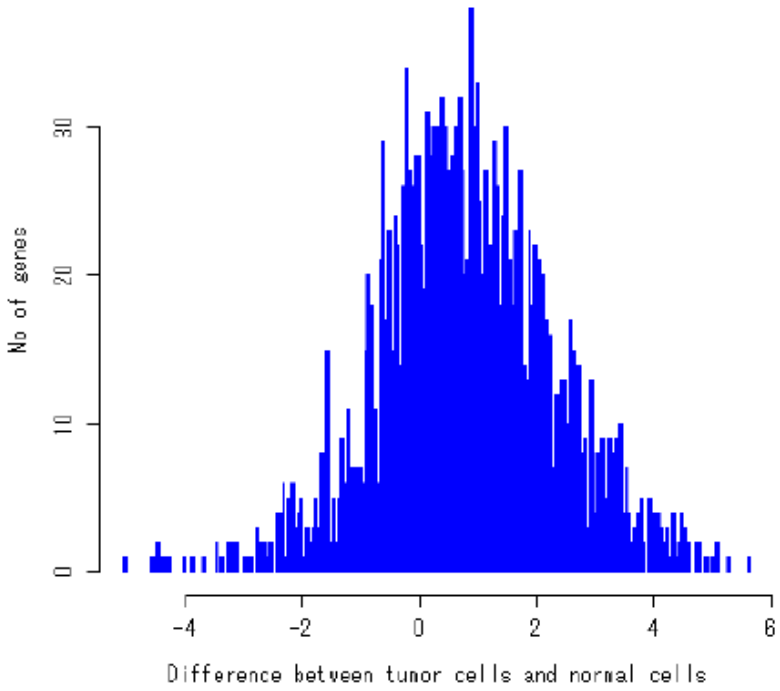
The distant pair of T33 and T26 shows a low correlation of gene expression, although they are of the same tissue type

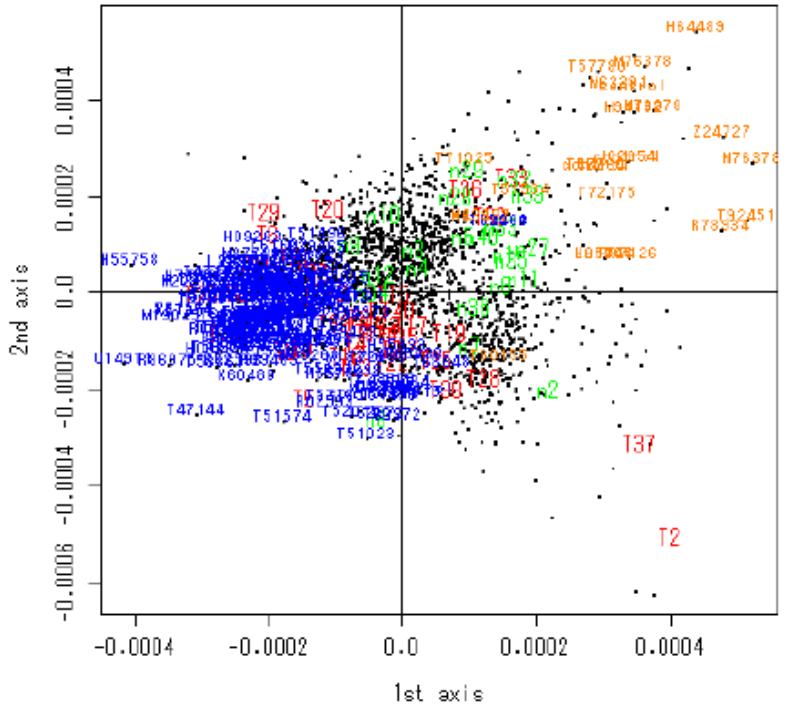


Tissues T33 and n39 are located close together, although they are of different tissue types and from different individuals.

The normal cells are mostly distributed in the upper-right region, whereas the tumor cells are distributed in the lower-left region, so the visual separation is moderately good.

OF GENES





The distribution of the z-scores (i.e. difference divided by s.d.) between the mean expression level of a gene in tumor cells and that in normal cells.

- **Orange accession numbers are attached to the genes with z-scores larger than 3,**
- **blue accession numbers are attached to the genes with z-scores smaller than *ī* 3**

MICROARRAYS DNA EXPERIMENTS :

Enables the researchers to monitor the expression levels of thousands of genes simultaneously.

Expression matrix can be used to detect the new subclasses of diseases, protect clinically important outcomes, such as the response to therapy and survival.

Problem: large number of genes vs. relatively small number of experiments.

CORRESPONDENCE ANALYSIS:

No parametrisation needed.

Projection into a common subspace hybridisations and genes.

Dimensionality reduction.

The result of experiments can be used in medicine for comparing clinically relevant groups (e.g., healthy vs diseased).

THANK YOU