Statistical Analysis of Gene Expression Data

박태성 교수

서울대학교 통계학과

E-mail: tspark@statcom.snu.ac.kr

Abstract

cDNA microarray technology allows the monitoring of expression levels for thousands of genes simultaneously. Many statistical analysis tools become widely applicable to the analysis of cDNA microarray data. In this talk, we consider a two-way ANOVA model to differentiate genes that have high variability and ones that do not. Using this model, we detect genes that have different gene expression profiles among experimental groups. The two-way ANOVA model is illustrated using cDNA microarrays of 3,800 genes obtained in an experiment to search for changes in gene expression profiles during neuronal differentiation of cortical stem cells.

CV

1999년-현재: 서울대 통계학과 부교수, 교수

1992년-1999년: 한국외대통계학과 조교수, 부교수

1991년-1992년: 미국립보건원 연구원

1990년-1991년: 아이오와 연구원

1990년: 미시간대 생물통계학과 박사

1986년: 서울대 통계학과 석사

1984년: 서울대 계산통계학과 학사

Statistical Analysis of Gene Expression Data

Taesung Park
Deptartment of Statistics, Seoul National University
tspark@stats.snu.ac.kr

Acknowledgments

Statistical Analysis

- Sung-Gon Lee
- Seungmook Lee
- Sung Hyun Kang
- Ho Sik Choi
- Deptartment of Statistics
  Seoul National University

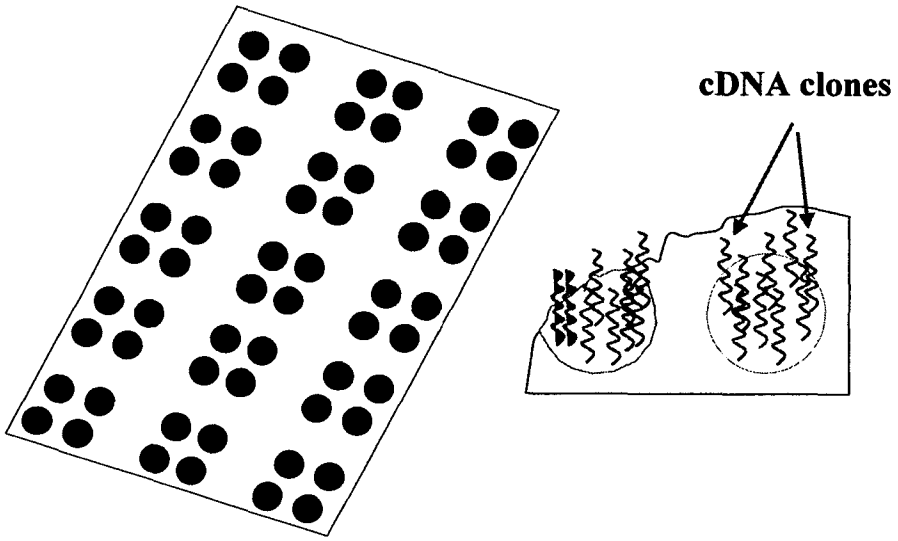
Lab

- Yong-Sung Lee
- Dong-Hyun Yoo
- Mi-Yoon Chang
- Department of Biochemistry
  Hanyang University College of Medicine

Outline

- Overview
  - cDNA microarrays
  - Statistical issues
- Example
- Normalization
- Statistical Analysis
  - Model
  - Statistical Test

cDNA Microarrays



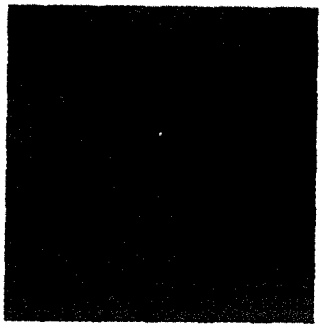
From T. Speed

cDNA Microarrays

Compare the genetic expression in two samples of cells

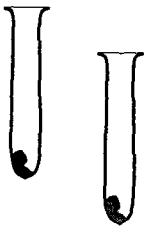
PRINT

cDNA from one gene on each spot



SAMPLES

cDNA labelled red/green



e.g. treatment control

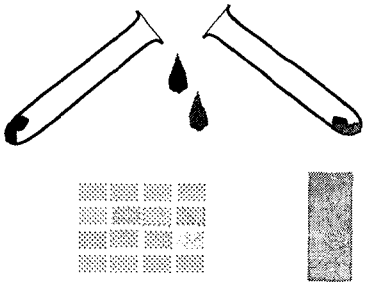
normal tumor tissue

From T. Speed

cDNA Microarrays

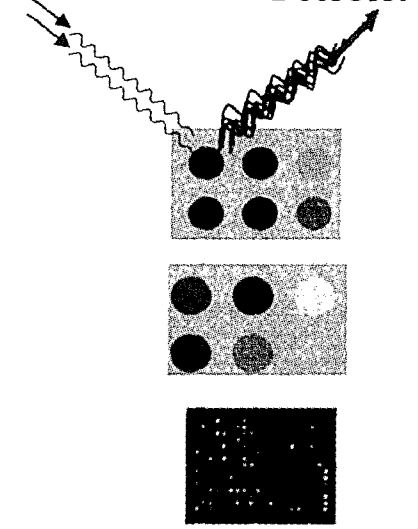
HYBRIDIZE

Add equal amounts of labelled cDNA samples to microarray.

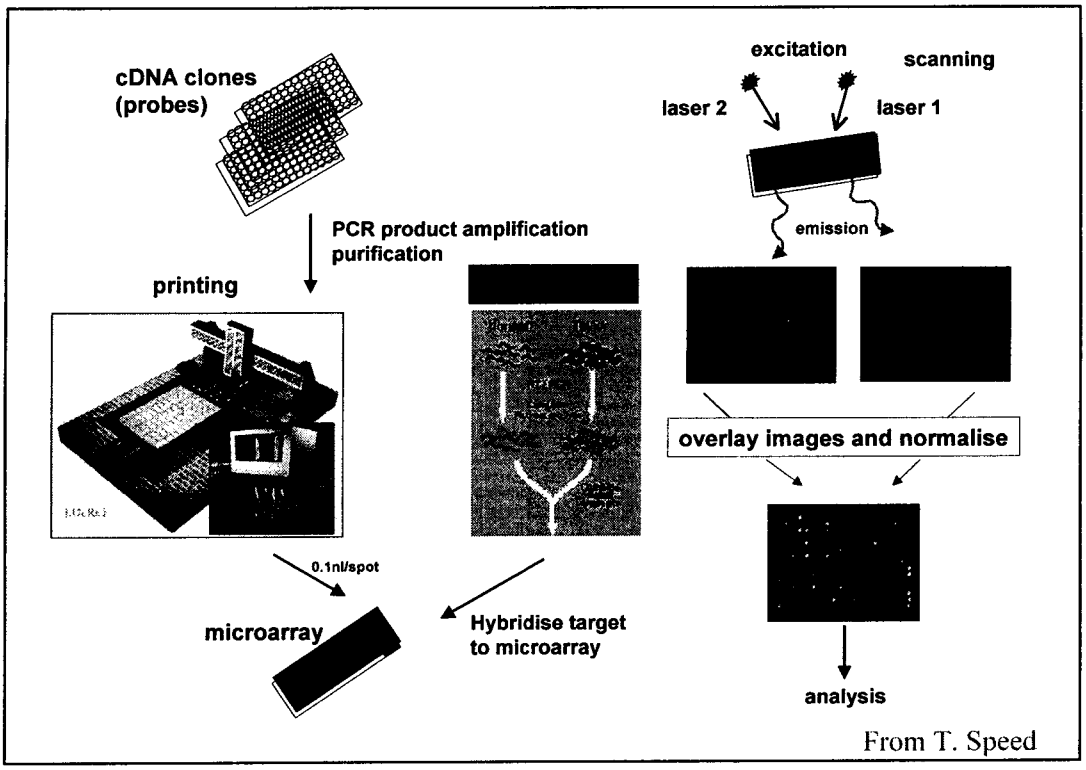


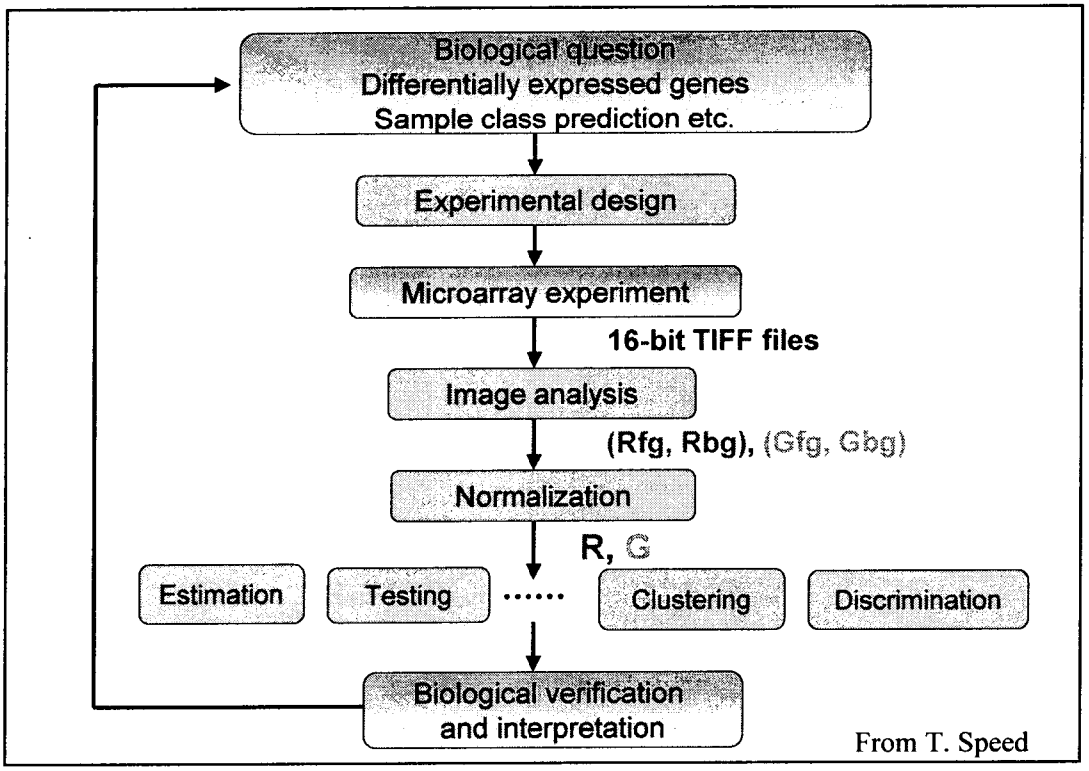
SCAN

Laser **Detector**



From T. Speed


cDNA clones
(probes)
PCR product amplification
purification
printing
0.1nl/spot
microarray
Hybridise target
to microarray
excitation
scanning
laser 2
laser 1
emission
overlay images and normalise
analysis
From T. Speed


Biological question
Differentially expressed genes
Sample class prediction etc.
Experimental design
Microarray experiment
16-bit TIFF files
Image analysis
(Rfg, Rbg), (Gfg, Gbg)
Normalization
R, G
Estimation
Testing
.....
Clustering
Discrimination
Biological verification
and interpretation
From T. Speed

cDNA Microarray Experiments

mRNA levels compared in many different contexts

- Different tissues, same organism (brain v. liver)
- Same tissue, same organism (tumor v. non-tumor)
- Same tissue, different organisms
- Time course experiments (effect of ttt, development)
- Other special designs (e.g. to detect spatial patterns).

From T. Speed

Statistical Issues

- Image analysis
  - Segmenting : fixed circles, adaptive Circle, adaptive shape ...
  - Quantifying : spot intensities, background values
- Normalization: within and between slides
- Quality: of images, of spots, of (log) ratios
- Testing
  - Which genes are (relatively) up/down regulated?
  - Assigning p-values to tests/confidence to results.

Statistical Issues, ctd

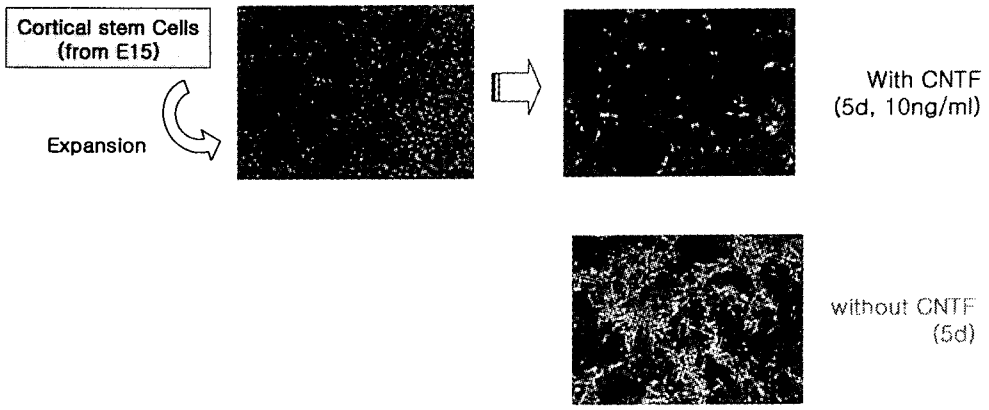
- Design of experiments: planing, sample size
- Analysis
  - Discrimination and allocation of samples
  - Clustering, classification: of samples, of genes
  - Analysis of time course experiments

Example

Differentiation of Neuronal Stem Cells

- Do genes change in their expression rates during neuronal differentiation?
  - Cortical neuronal stem cells were isolated from E15 rat fetus and expanded under the presence of bFGF.
  - After expansion, differentiation to neuronal cells was induced by removing bFGF and the expression patterns were compared to those of astrocyte differentiation. .

Fate of Cortical Stem Cells



CNTF: cilliary neuritrophic factor

cDNA Samples

- Over 3,800 genes obtained from developing fetal rat brain were immobilized on a glass chip and fluorescence-labeled target cDNA were hybridized
- cDNAs from cortical stem cells with CNTF
  - labeling red(R) channel
- cDNAs from cortical stem cells without CNTF
  - labeling red(R) channel
- Reference cDNAs
  - labeling green(G) channel

Image of 3.8K Rat Chip

Example

- 3,800 genes
- Two experimental groups
  - No CNTF (control)
  - CNTF (treatment)
- Six different time sequences
  - (0, 1day, 2day, 3day, 4day, 5day)
- Three replications
- 36 slides

Normalization

- Goal of Normalization

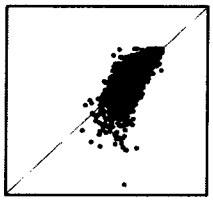
  R vs G, R < G

  ① To control noises (variation)

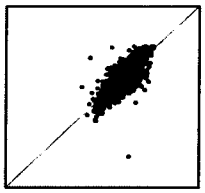
  ② To reduce biases

  ③ To allow comparison of multiple slides

Exampe 1



Exampe 2



Normalization

- Which Genes to Use ?

  ① All genes on the array

  - Only a small proportion of genes are expected to be differentially expressed
  - All remaining genes are expected to have constant expression

  ② Constantly expressed genes

  - Housekeeping genes

  ③ Controls

  - Spiked controls

Normalization

Within-slide Normalization

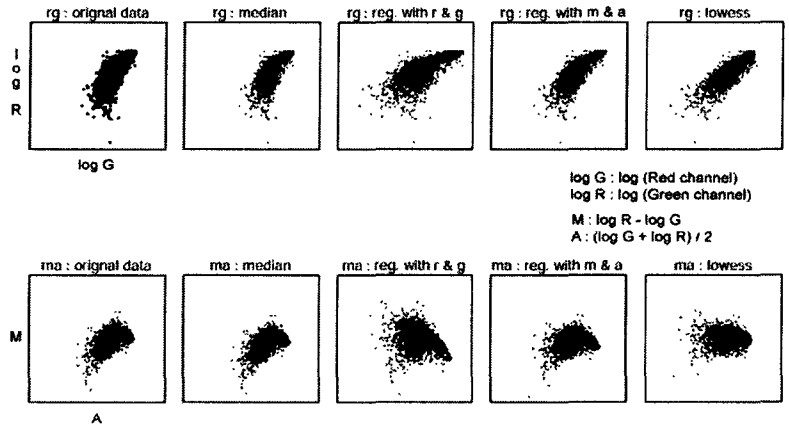
1. Global Normalization
   a. Using Median of log ratio
      - c = Median of M (= log R/G)
      - Modified log ratio=log R/G – c
   b. Using Regression between R vs G
      - Fitting: $G=b_0+b_1 R$
      - Modified log ratio=log $R/((G-b_0)/ b_1)$

Normalization, ctd

2. Intensity Dependent Normalization
   c. Using Regression between Ratio vs Intensity
      - M = log R/G, A = ½ log R*G
      - Fitting: $m(A)= b_0 + b_1 A$
      - Modified log ratio=log R/G-m(A)
   d. LOWESS (Locally Weighted Estimation)
      - Fitting lowess curve: c(A)
      - Modified log ratio = log R/G – c(A)

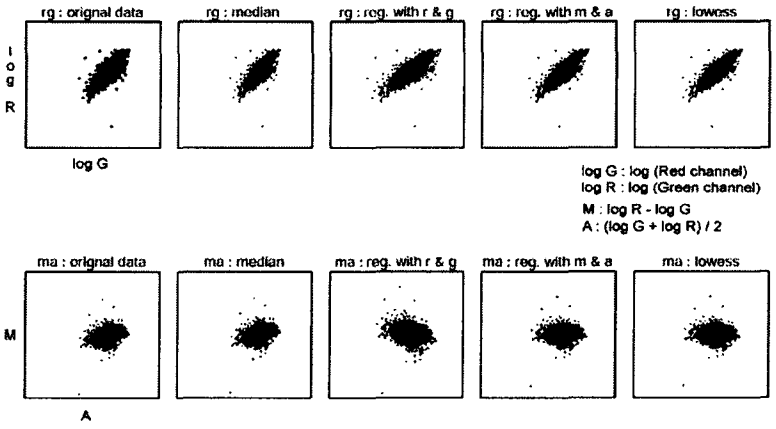
Normalization

- Example 1

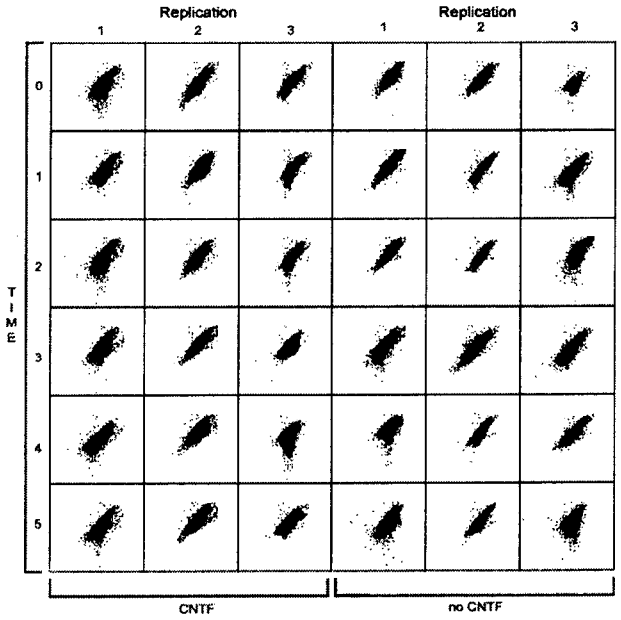


Normalization

- Example 2

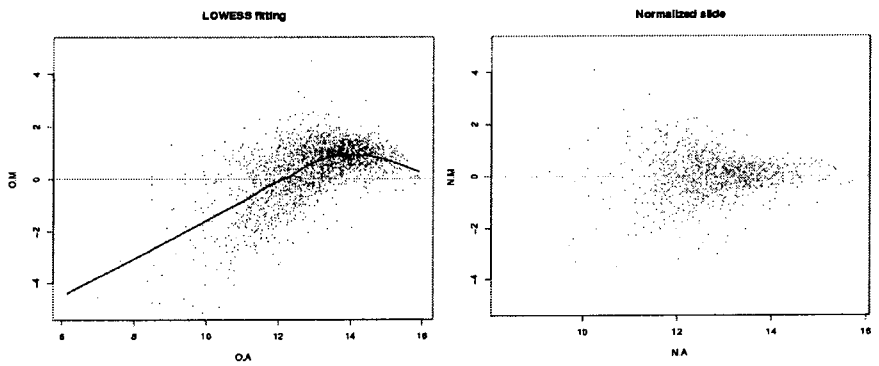


36 Orignal Slides

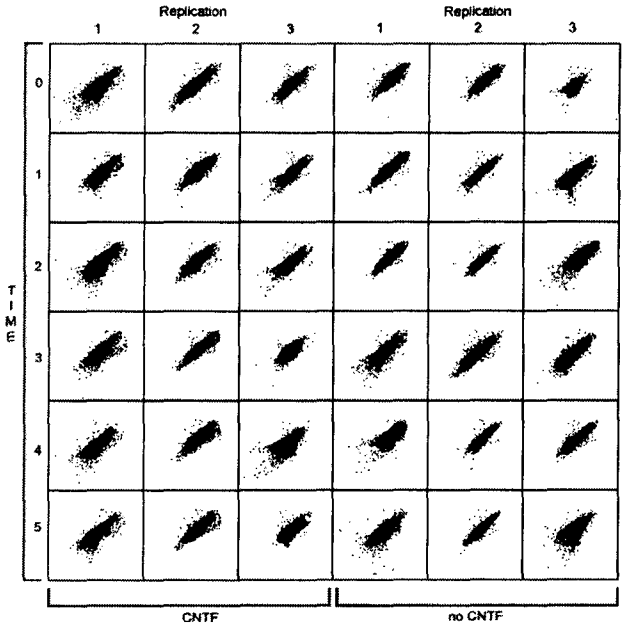


Normalization

- LOWESS



Normalization of 36 Slides



Statistical Analysis

- Find differentially expressed genes between CNTF and no CNTF group
  - Statistical test for each gene
- Need to adjust for time effect
- Statistical models: two-way ANOVA model
  - Two factors
    - Factor 1: experimental group
    - Factor 2: time
  - Models
    - *M*1: Main effect model
    - *M*2: Interaction model

Statistical Analysis

- Model

$$M1: \log R/G_{ijkl} = \mu_l + \alpha_{il} + \beta_{jl} + \varepsilon_{ijkl}$$

$$M2: \log R/G_{ijkl} = \mu_l + \alpha_{il} + \beta_{jl} + (\alpha\beta)_{ijl} + \varepsilon_{ijkl}$$

  - α : group effect (CNTF and no CNTF)
  - β : time effect
  - ε : error
    - i : CNTF(1) vs no CNTF(2)
    - j : time(0, 1, 2, 3, 4, 5)
    - k : replication (1, 2, 3)
    - l : gene number
- Test for α and $(\alpha\beta)$

Statistical Analysis

- Multiple Hypotheses Test
  - Type I error
    - Single test : $Pr(H_1|H_0)=\alpha$
    - Multiple tests :

$$P(H_1 \mid H_0)=1-(1-Pr(\text{single } H_1 \mid H_0))^n = 1-(1-\alpha)^n$$

1) F-test
   - Assume log ratio has a Normal distribution
   - F-value : F-test
2) Permutation test
   - No predefined distribution

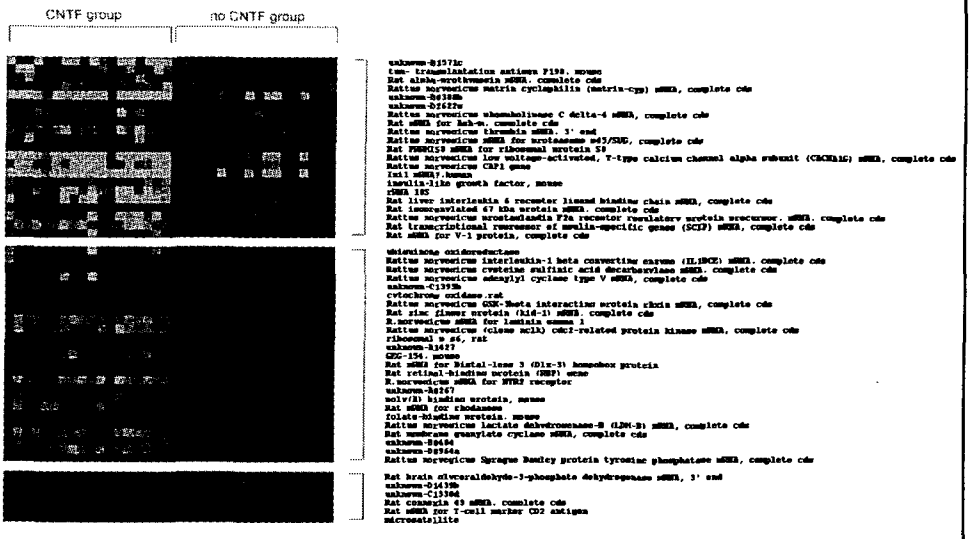
Statistical Analysis – Genes with Significant Group effect

Gene name	F-test		Permutation Test	
	Un-adjusted	Adjusted	Un-adjusted	Adjusted
R.norvegicus mRNA for laminin gamma 1	5.11E-18	1.96161E-14	<= 1.0E-5	<= 1.0E-5
Rat membrane guanylate cyclase mRNA, complete cds	1.34E-16	5.14763E-13	<= 1.0E-5	<= 1.0E-5
unknown-B0484	1.98E-16	7.6013E-13	<= 1.0E-5	<= 1.0E-5
R.norvegicus mRNA for NTR2 receptor	6.06E-16	2.32404E-12	<= 1.0E-5	<= 1.0E-5
Rattus norvegicus (clone nclk) cdc2-related protein kinase mRNA, complete cds	2.36E-15	9.05589E-12	<= 1.0E-5	<= 1.0E-5
GEG-154, mouse	1.83E-14	7.00048E-1.1	<= 1.0E-5	<= 1.0E-5
unknown-A1427	6.51E-13	2.49429E-09	<= 1.0E-5	<= 1.0E-5
poly(A) binding protein, mouse	9.3E-13	3.56316E-09	<= 1.0E-5	<= 1.0E-5
(more 45 genes)	<1.0E-07	< 0.05	0	<= 0.0064

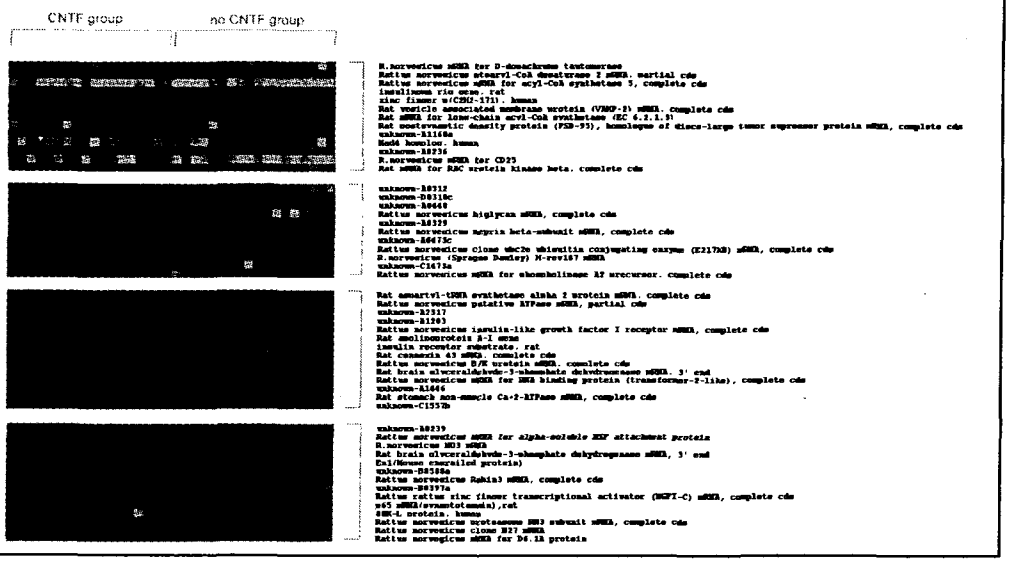
Statistical Analysis – Genes with Some Interaction Effect

Gene Name	F-test	
	Unadjusted	Adjusted
Rattus norvegicus B/K protein mRNA, complete cds	0.000215	0.827024478
Rattus norvegicus clone N27 mRNA	0.000477	1
R.norvegicus (Sprague Dawley) H-rev107 mRNA	0.000693	1
R.norvegicus NO3 mRNA	0.000698	1
Rattus norvegicus biglycan mRNA, complete cds	0.001039	1
Rattus norvegicus mRNA for D6.1A protein	0.001459	1
Rat connexin 43 mRNA, complete cds	0.001633	1
(more 45 genes)	< 0.001	

Clustering Using Genes with Significant Main Effect



Clustering Using Genes with Some Interaction Effect



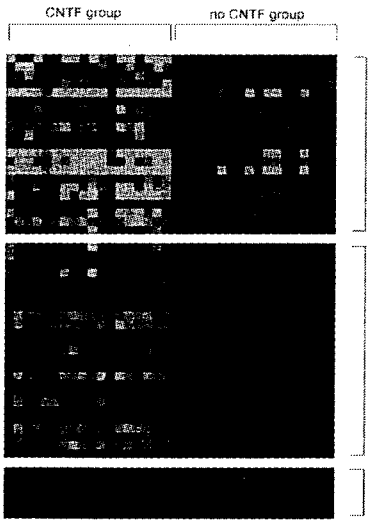
Statistical Analysis – Genes with Significant Group effect

Gene name	F-test		Permutation Test	
	Un-adjusted	Adjusted	Un-adjusted	Adjusted
R.norvegicus mRNA for laminin gamma 1	5.11E-18	1.96161E-14	<= 1.0E-5	<= 1.0E-5
Rat membrane guanylate cyclase mRNA, complete cds	1.34E-16	5.14763E-13	<= 1.0E-5	<= 1.0E-5
unknown-B0484	1.98E-16	7.6013E-13	<= 1.0E-5	<= 1.0E-5
R.norvegicus mRNA for NTR2 receptor	6.06E-16	2.32404E-12	<= 1.0E-5	<= 1.0E-5
Rattus norvegicus (clone nclk) cdc2-related protein kinase mRNA, complete cds	2.36E-15	9.05589E-12	<= 1.0E-5	<= 1.0E-5
GEG-154, mouse	1.83E-14	7.00048E-11	<= 1.0E-5	<= 1.0E-5
unknown-A1427	6.51E-13	2.49429E-09	<= 1.0E-5	<= 1.0E-5
poly(A) binding protein, mouse	9.3E-13	3.56316E-09	<= 1.0E-5	<= 1.0E-5
(more 45 genes)	<1.0E-07	< 0.05	0	<= 0.0064

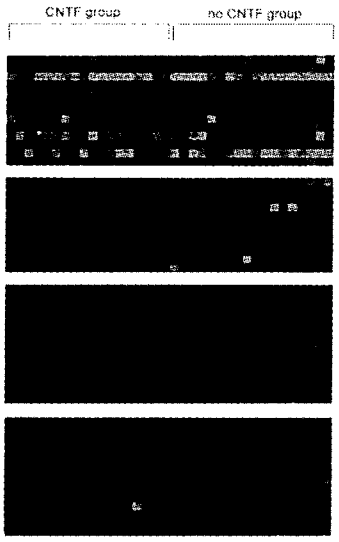
Statistical Analysis – Genes with Some Interaction Effect

Gene Name	F-test	
	Unadjusted	Adjusted
Rattus norvegicus B/K protein mRNA, complete cds	0.000215	0.827024478
Rattus norvegicus clone N27 mRNA	0.000477	1
R.norvegicus (Sprague Dawley) H-rev107 mRNA	0.000693	1
R.norvegicus NO3 mRNA	0.000698	1
Rattus norvegicus biglycan mRNA, complete cds	0.001039	1
Rattus norvegicus mRNA for D6.1A protein	0.001459	1
Rat connexin 43 mRNA, complete cds	0.001633	1
(more 45 genes)	< 0.001	

Clustering Using Genes with Significant Main Effect



Clustering Using Genes with Some Interaction Effect



Summary

Statistical Analysis for Microarray Data

- Normalization
- Two-way ANOVA Model
- Statistical Test
  - F-test: unadjusted p-value, adjusted p-value
  - Permutation test: unadjusted p-value, adjusted p-value
- Cluster Analysis

Some Statistical Research Microarray Data Analysis

- *Experimental design* : Churchill & Kerr
- *Image analysis*: Zuzan & West, ...
- *Data visualization*: Carr *et al*
- *Estimation*: Ideker *et al, ...*
- *Multiple testing*: Westfall & Young , Storey, ...
- *Discriminant analysis*: Golub et al,...
- *Clustering*: Hastie & Tibshirani, Van der Laan, Fridlyand & Dudoit, ...
- *Empirical Bayes*: Efron *et al*, Newton *et al*,....
- *Multiplicative models*: Li &Wong
- *Multivariate analysis*: Alter *et al*
- *Genetic networks*: D'Haeseleer *et al and more*

From T. Speed