Analysis for Gene Expression Data of the NCI 60 Cancer Cell Lines Using MCMC on a Hierarchical Effects Model

Jae K. Lee

University of Virginia

(Formerly at Lab of

Molecular Pharmacology, NCI)

November 11, 2000

IPAM Functional Genomics Workshop

OUTLINE

- NCI Large Screening Program of 60 Cancer Cell Lines
- Statistical Issues in High Throughput Array Data
- Hierarchical Effects Model Using MCMC
- Web-Based Interactive Analysis Tool (GS-HEM)

60 Cancer Cell Lines (12 reference pool lines)

- Developmental Therapeutics Programs (DTP), NCI

CELL LINES OF THE NCI DRUG SCREEN

Colon
- β COLO 205
- HCC-2998
- HCT-116
- HCT-15
- HT29
- KM12
- SW-620

Ovarian
- IGROV1
- β OVCAR-3
- β OVCAR-4
- OVCAR-5
- OVCAR-8
- SK-OV-3

Melanoma
- β LOX IMVI
- MALME-3M
- M14
- SK-MEL-2
- SK-MEL-28
- SK-MEL-5
- UACC-257
- UACC-62

Lung
- A549/ATCC
- EKVX
- HOP-62
- HOP-92
- β NCI-H226
- NCI-H23
- NCI-H322M
- NCI-H460
- NCI-H522

Renal
- 786-0
- A498
- ACHN
- β CAKI-1
- RXF 393
- SN12C
- TK-10
- UO-31

Prostate
- β PC-3
- DU-145

CNS
- SF-268
- SF-295
- SF-539
- β SNB-19
- SNB-75
- U251

Leukemia
- CCRF-CEM
- β HL-60(TB)
- β K-562
- MOLT-4
- RPMI-8226
- SR

Breast
- β MCF7
- NCI/ADR-RES
- MDA-MB-231/ATCC
- β HS 578T
- MDA-MB-435
- MDA-N
- BT-549
- T-47D

The NCI Cancer Drug Discovery - Development Pipeline

~300,000 cmpds

Inventory

> 68,000 cmpds

60-Cell line screen

In vivo studies

Clinical Trials

>100 descriptors

Database S

(>100 3-D Molecular Structure)

~300K cmpds

~68K cmpds

>30,000,000 numbers

60 cell lines

Database A

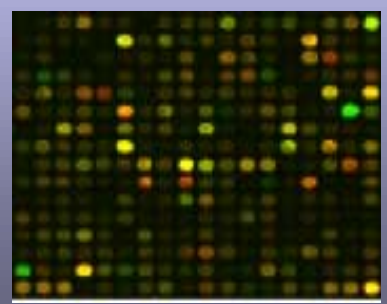
(Activity Patterns)

>4,000,000 numbers

Database T

(Molecular Targets)

~150 specific targets



mRNA expression database
-cDNA microarray
-Oligonucleotide array

60 Cancer Cell Line Screening Data

- **Activity Data (A)**
  - Drug Potency Activity; GI_{50}, TGI
    - 29,026 open compounds (Sep. '99)
    - 6,205 drugs tested more than once (Jan. '00)
    - 118 mechanism of action drug compounds
- **Target Data (T)**
  - Protein (41 protein expression data)
  - cDNA Hybridization Expression
    - Microarray (9,706 genes x 60 cells)
    - Oligonucleotide Array (6,800 genes x 60 cells)

OUTLINE

- NCI Large Screening Program of 60 Cancer Cell Lines
- Statistical Issues in High Throughput Array Data
- Hierarchical Effects Model Using MCMC
- Web-Based Interactive Analysis Tool (GS-HEM)

Analyzing Gene Expression Array Data

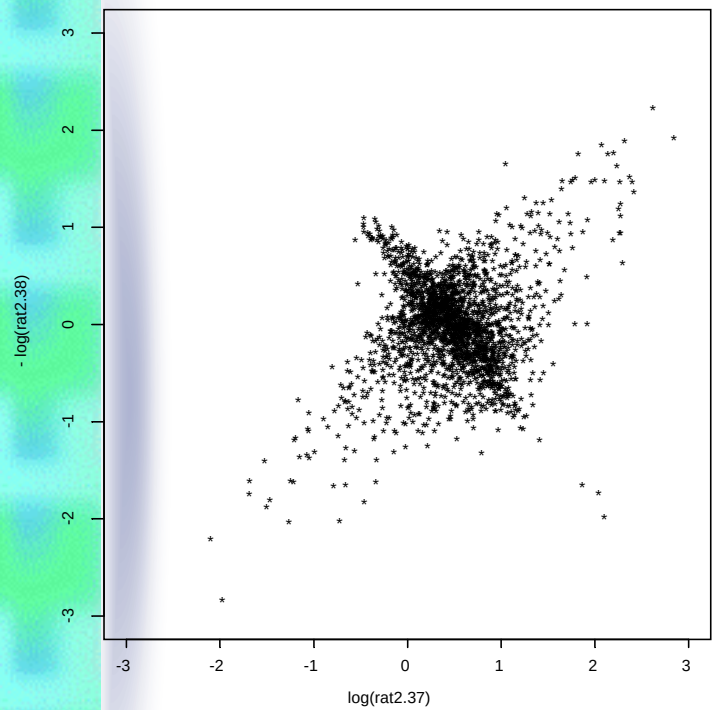
Need pre-processing and pre-screening to avoid (large number of) irrelevant results from various artifacts

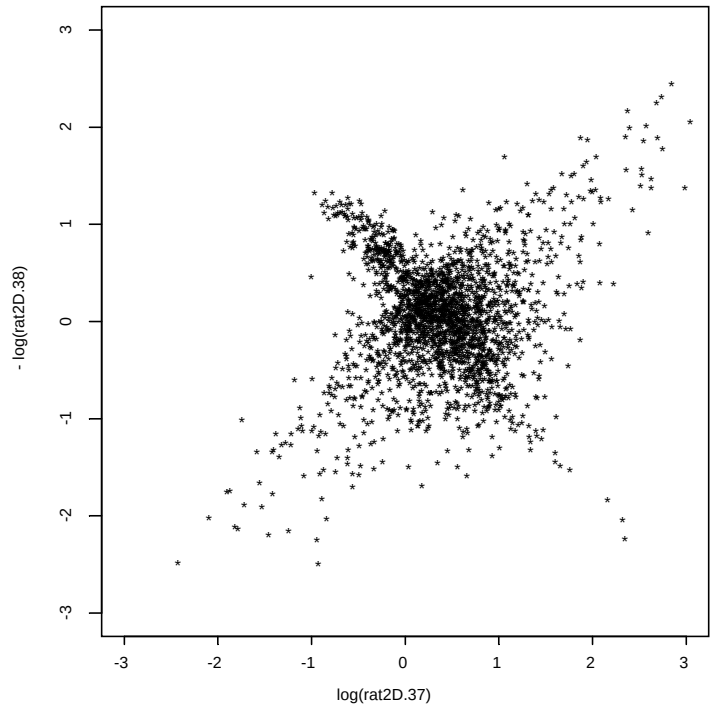
- **quality control:**
  - optimization criteria
    - normalization (scaling or centering factors)
    - reproducibility (ratio, Chen et al. 97; ave. difference, Affymetrix)
    - sensitivity (power for identifying differential expression levels)
  - thresholding or missing values
    - maximum number of informative data points
- **subsetting: depends on inference goals to identify(e.g., Tibshirani et al., 2000)**
  - genes with distinctive patterns in specific cases
  - genes with major expression variations

Myths in Gene Chip Study

- **Myth 1: Can do with each single hybridization**
- **Question: how can we do array study for 3-4 biological factors using 4 chips?**
- **Answer: don't do it!**
  - Enormous false positive findings, say 10^{-2} from single chip -> from duplicates $10^{-2}x10^{-2}$
  - Waste of time for a week for statisticians, for several months for biologists
- **Myth 2: Can do without a statistical design**
  - statistical factors of variability
    - gene
    - variety: types of sample, treatment, time
    - individual sample
    - array
    - dye (microarray)

Reciprocally Labeled Pairs of Microarray data





Analyzing Array Data (continued)

- **replication and experimental design (blocking)**
  - replicates of genes on a chip and/or of treatments on several chips, especially for interaction
  - blocking errors from individual, array, and dye (not interested in identifying them separately, but need to have replicates to "factor them out" together)

Example: Experimental design on an array study

- **Microarray study on** comparing a treatment effect at two different time points with two individual replicates

Chip 1		Chip 2		Chip 3		Chip 4	
Cy3	Cy5	Cy3	Cy5	Cy3	Cy5	Cy3	Cy5
I1-T1	Ref	Ref	I1-T2	Ref	I2-T1	I2-T2	Ref

- Replicates for arrays, dyes, individuals are shared.
- Treatment and time point factors are separately replicated from individual, array, and dye factors.

OUTLINE

- NCI Large Screening Program of 60 Cancer Cell Lines
- Statistical Issues in High Throughput Array Data
- Hierarchical Effects Model Using MCMC
- Web-Based Interactive Analysis Tool (GS-HEM)

Estimating Interaction Effects

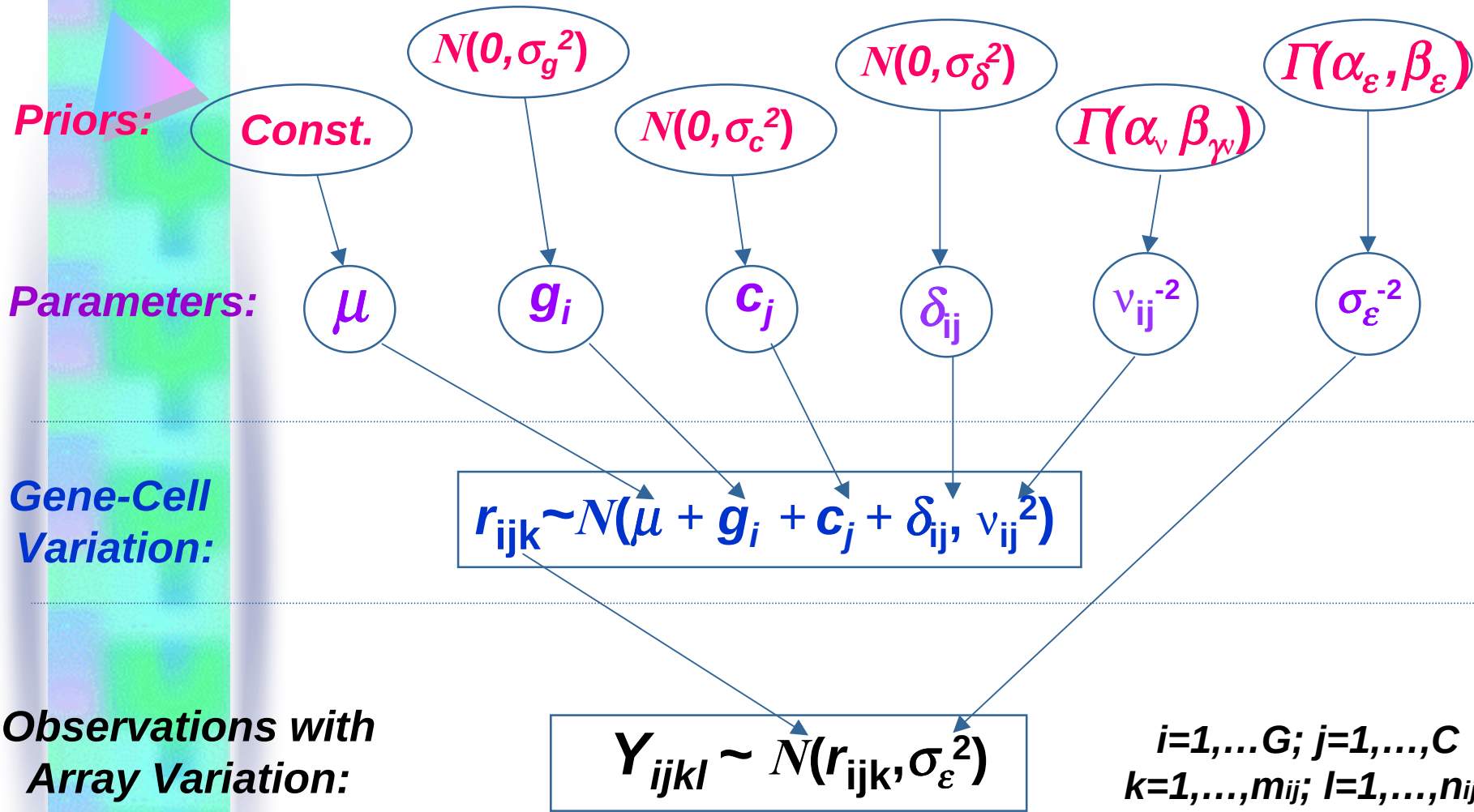
- **Need to identify interaction expression levels**
- **Linear Model for interaction** (Kerr and Churchill, 2000)

$$Y_{ijk} = G_i + C_j + (GC)_{ij} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0, \sigma^2)$$

- **Problem in large screening:**
  - astronomical number of interaction parameters
    - e.g., Two-way interaction model with 9 x 10,000 levels requires estimation of more than 100,000 parameters
- **Multiple layers and correlated structure of random variation**
- **Unbalanced and missing data structure (GLM, EM)**
- **Small sample bias in variance estimation (REML)**

Hierarchical Effects Model

- **Model**
  - Layer 1: Array experimental variability:
    $Y_{ijkl} = r_{ijk} + \varepsilon_{ijkl}$, $\varepsilon_{ijkl} \sim N(0, \sigma^2)$, given G_{ijk}
  - Layer 2: Biological variability:
    $r_{ijk} = G_i + C_j + \delta_{ij} + \alpha_{ijk}$, $\alpha_{ijk} \sim N(0, \nu_{ij}^2)$
  - Priors for parameters



DAG (Directed Acyclic Graph) for Bayesian Hierarchical Effects Model (HEM)

- **Why hierarchical?**
  - Experimental reasons:
    - Chronological experimental procedure
    - Several different hierarchical layers of errors
  - Statistical reasons:
    - Want to decompose error variation into several components, while utilizing all variation information among replicates
    - Estimate interaction effects of over-parameterized models, especially for large data sets, taking into account unbalanced, missing data structure
    - Predict cases when no experimental data available
- **Need computational statistical tools on complex hierarchical models -> MCMC**

Posterior Distributions of Parameters and Missing Data

$$\pi(\mu|rest) = \text{Normal}\left(\sum_{i,j} \frac{\sum_k (r_{i,j,k} - g_i - c_j - \delta_{i,j})}{\frac{\sigma^2_{r_{i,j}}}{m_{1,1}} + \cdots + \frac{\sigma^2_{r_{G,C}}}{m_{G,C}}}, \quad \left(\frac{m_{1,1}}{\sigma^2_{r_{1,1}}} + \cdots + \frac{m_{G,C}}{\sigma^2_{r_{G,C}}}\right)^{-1}\right)$$

$$\pi(g_i|rest) = \text{Normal}\left(\sum_j \frac{\sum_k (r_{i,j,k} - \mu - c_j - \delta_{i,j})/\sigma^2_{r_{i,j}}}{\frac{m_{i,1}}{\sigma^2_{r_{i,1}}} + \cdots + \frac{m_{i,C}}{\sigma^2_{r_{i,C}}} + \frac{1}{\sigma^2_g}}, \quad \left(\frac{m_{i,1}}{\sigma^2_{r_{i,1}}} + \cdots + \frac{m_{i,C}}{\sigma^2_{r_{i,C}}} + \frac{1}{\sigma^2_g}\right)^{-1}\right),$$

$$\pi(c_j|rest) = \text{Normal}\left(\sum_i \frac{\sum_k (r_{i,j,k} - \mu - g_i - \delta_{i,j})/\sigma^2_{r_{i,j}}}{\frac{m_{1,j}}{\sigma^2_{r_{1,j}}} + \cdots + \frac{m_{G,j}}{\sigma^2_{r_{G,j}}} + \frac{1}{\sigma^2_c}}, \quad \left(\frac{m_{1,j}}{\sigma^2_{r_{1,j}}} + \cdots + \frac{m_{G,j}}{\sigma^2_{r_{G,j}}} + \frac{1}{\sigma^2_c}\right)^{-1}\right),$$

$$\pi(\delta_{i,j}|rest) = \text{Normal}\left(\frac{\sigma^2_\delta}{\sigma^2_\delta + \sigma^2_{r_{i,j}}/m_{i,j}} \sum_k \frac{r_{i,j,k} - \mu - g_i - c_j}{m_{i,j}}, \quad \left(\frac{1}{\sigma^2_\delta} + \frac{1}{\sigma^2_{r_{i,j}}/m_{i,j}}\right)^{-1}\right),$$

$$\pi(r_{i,j,k}|rest) = \begin{cases} \text{Normal}(\mu + g_i + c_j + \delta_{i,j}, \sigma^2_{r_{i,j}}), & \text{if missing} \\ \text{Normal}\left(\frac{\sigma^2_{r_{i,j}}}{\sigma^2_{r_{i,j}} + \frac{\sigma^2_r}{n_{i,j,k}}} \sum_l \frac{y_{i,j,k,l}}{n_{i,j,k}} + \frac{\frac{\sigma^2_r}{n_{i,j,k}}}{\sigma^2_{r_{i,j}} + \frac{\sigma^2_r}{n_{i,j,k}}}(\mu + g_i + c_j + \delta_{i,j}), \quad \left(\frac{1}{\sigma^2_{r_{i,j}}} + \frac{n_{i,j,k}}{\sigma^2_r}\right)^{-1}\right), & \text{otherwise} \end{cases}$$

$$\pi(\sigma^{-2}_{r_{i,j}}|rest) = \text{Gamma}\left(\frac{n_{i,j}}{2} + \alpha_r, \quad \sum_k \frac{(r_{i,j,k} - \mu - g_i - c_j - \delta_{i,j})^2}{2} + \beta_r\right),$$

$$\pi(\sigma^{-2}_r|rest) = \text{Gamma}\left(\frac{N}{2} + \alpha_r, \quad \sum_{i,j,k,l} \frac{(y_{i,j,k,l} - r_{i,j,k})^2}{2} + \beta_r\right).$$

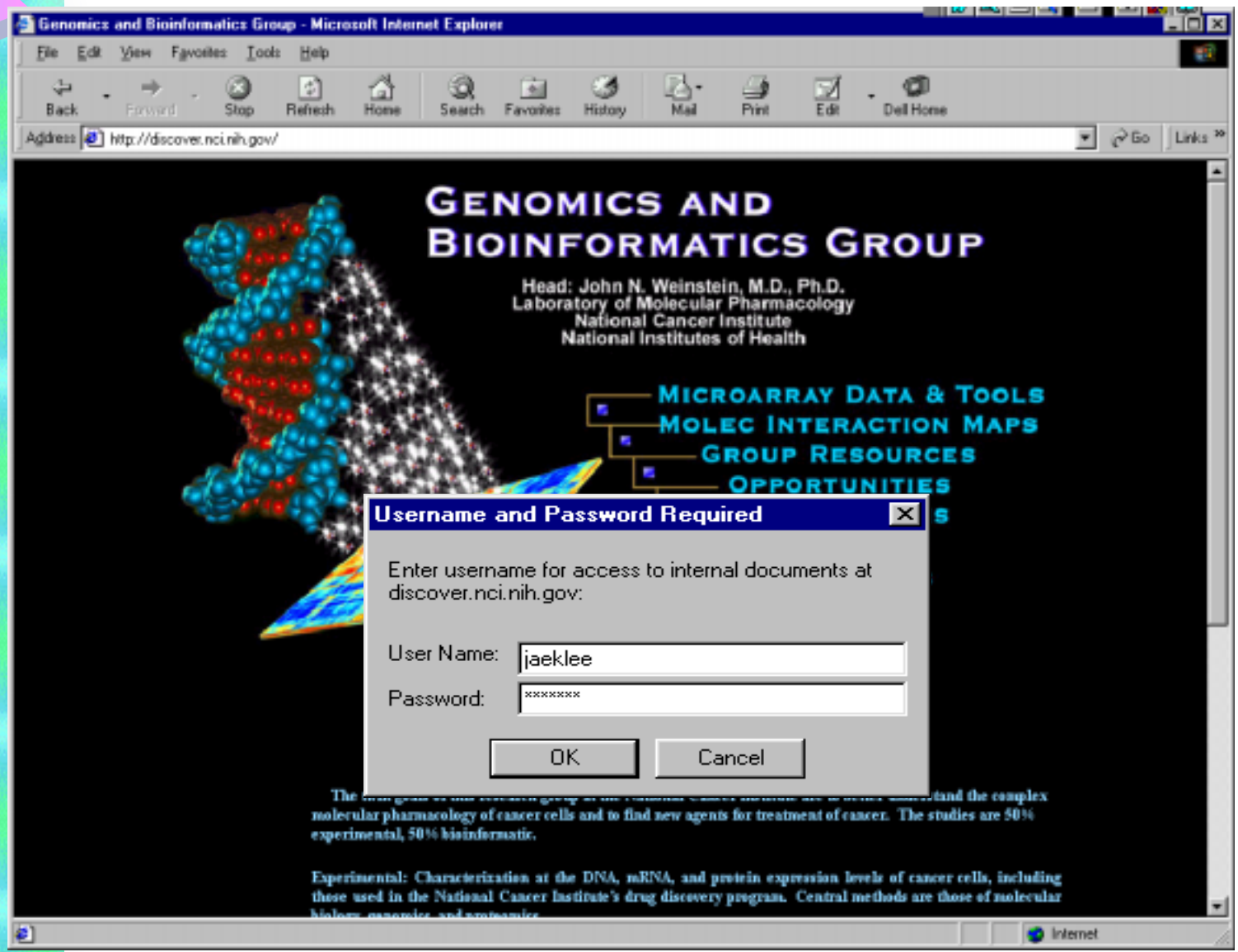
OUTLINE

- NCI Large Screening Program of 60 Cancer Cell Lines
- Statistical Issues in High Throughput Array Data
- Hierarchical Effects Model Using MCMC
- **Web-Based Interactive Analysis Tool (GS-HEM)**

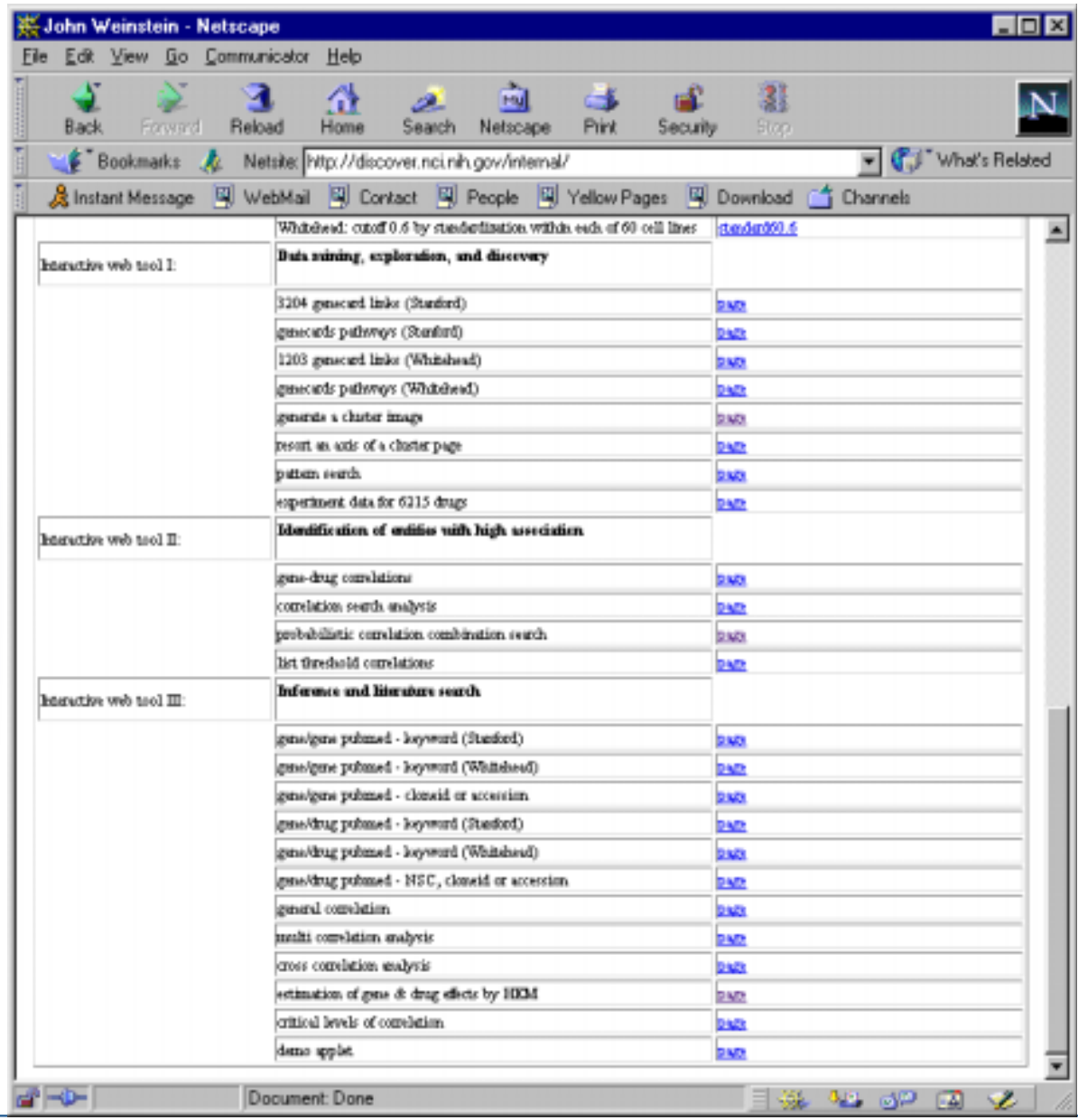
GS-HEM: Interactive Web-Based Tool

- **Innovations in our web-based interactive tools**
  - Effective interaction & collaboration
  - Efficient investigation on various combinations of interest
  - Accessible independently from platforms and locations
  - Fully utilize a statistical package, S-PLUS in both statistical and graphic (visualization) methods
    - minimize redundant costs for statistical development
    - incorporate other programs and software

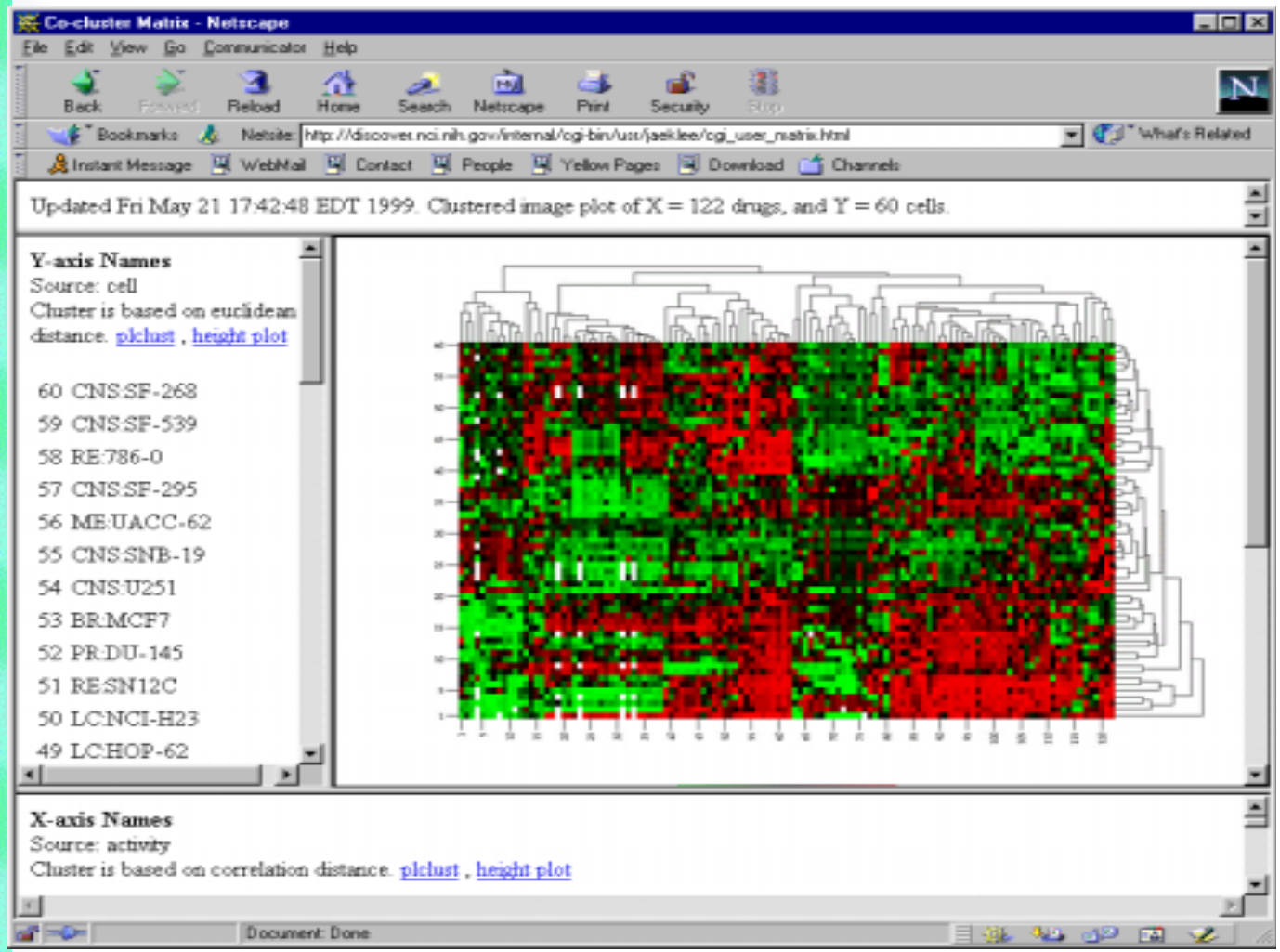
World Wide Web and On-line Analysis



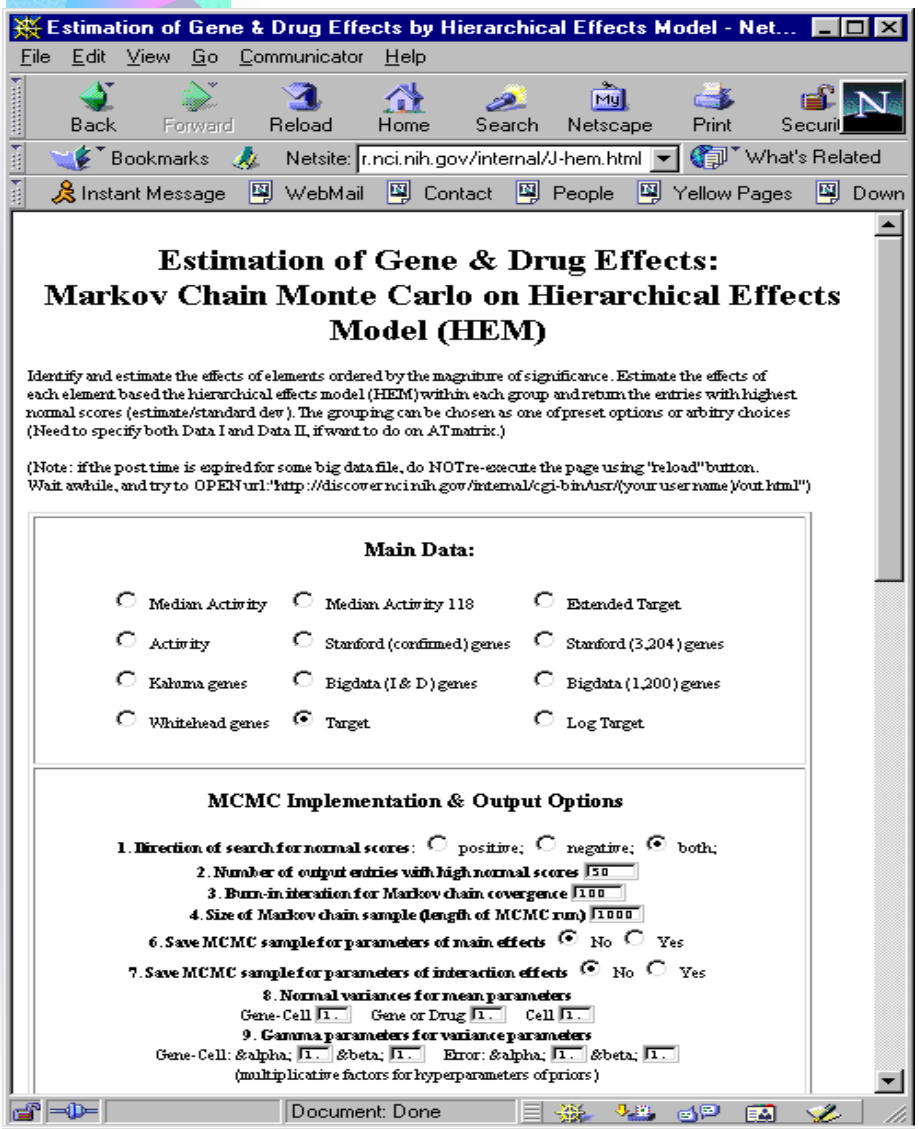
Interactive Analysis Pages

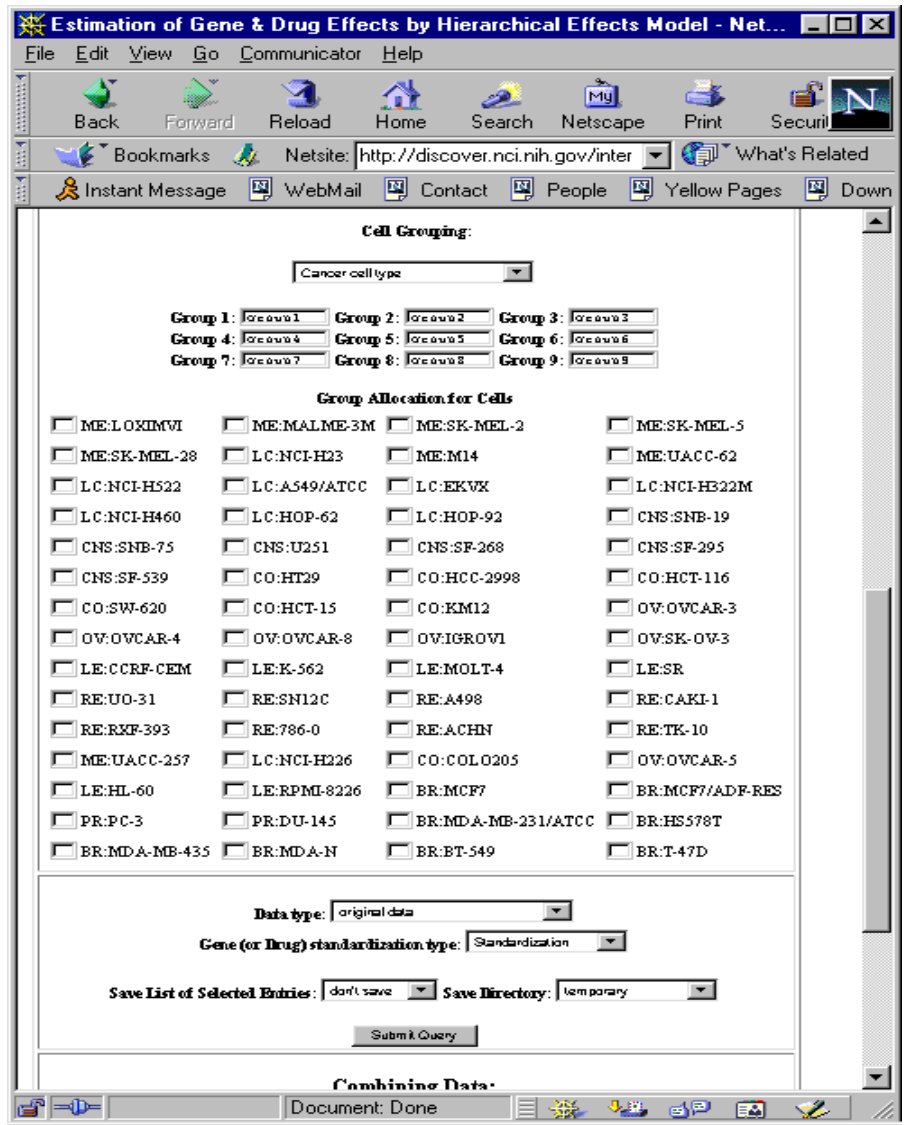


Cluster-image analysis result

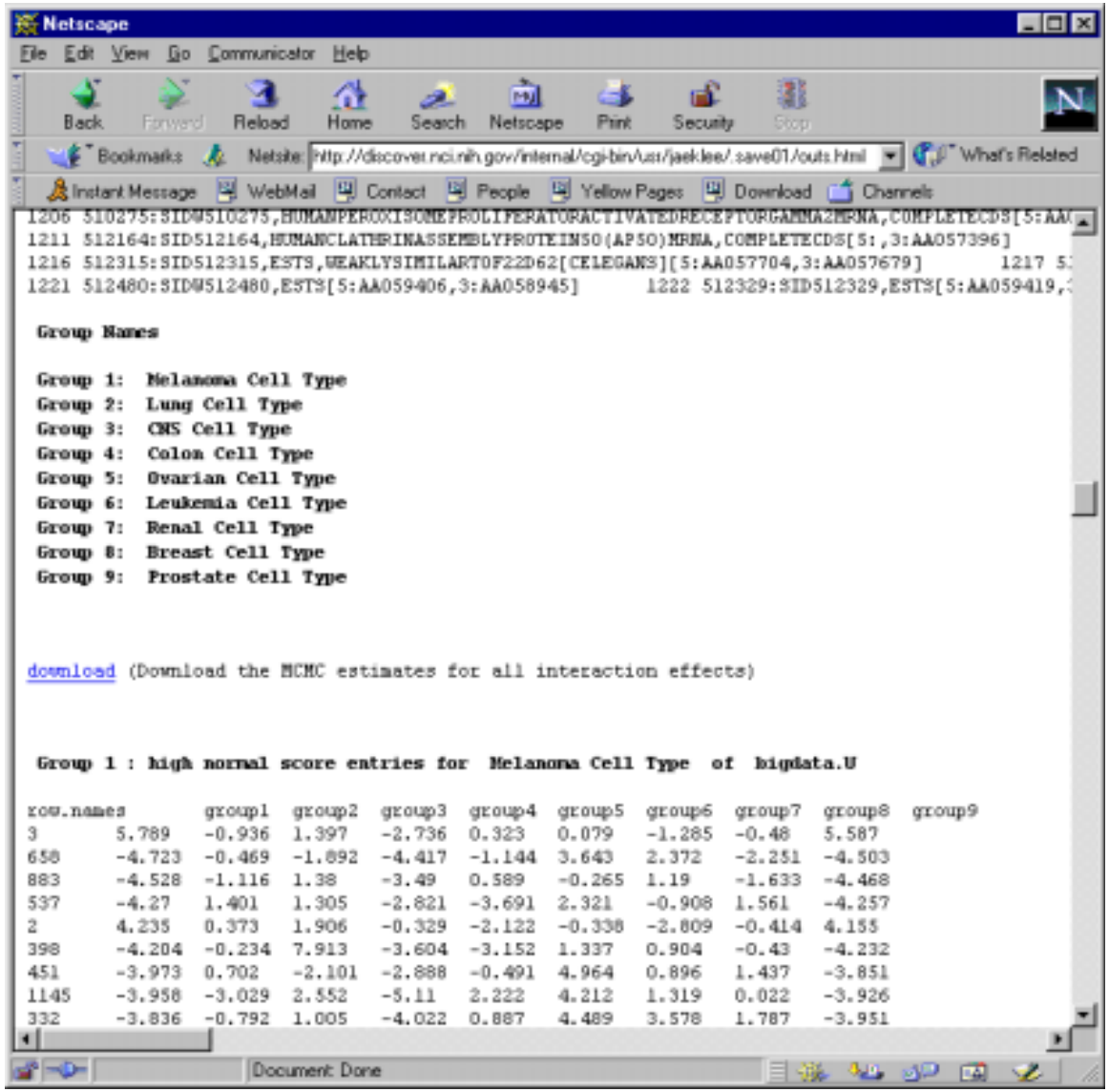


Estimation of Gene & Drug Effects by GS-HEM

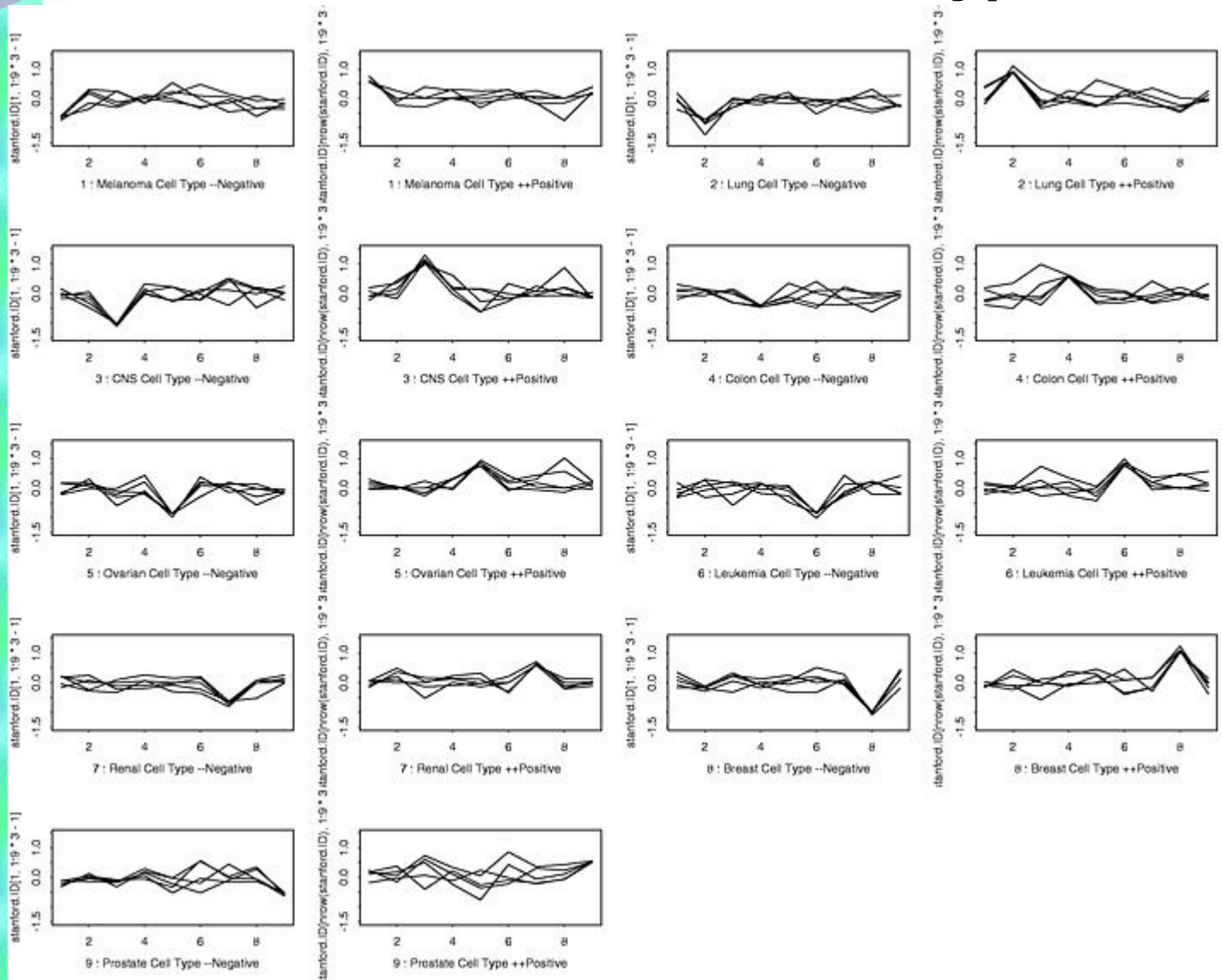




HEM Result: Estimation of Gene-Cell Interaction Effects



GS-HEM Results on Microarray data for 9 Cancer Cell Line Types



Cancer tissue specific genes

- **Group 4: Colon Cell Type**
  - 264347, Transforming growth factor beta
  - 489884, Human insulin-like growth factor binding protein 5 (IGFBP5)
  - 469842, Homo sapiens mRNA for fatty acid binding protein, complete cds
- **Group 5: Ovarian Cell Type**
  - 183950:THY-1 MEMBRANE GLYCOPROTEIN PRECURSOR
  - 489235:HADHB Hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase
  - 285784, ESTs, Highly similar to TUBULIN BETA CHAIN [Schizosaccharomyces pombe]
- **Group 6: Leukemia Cell Type**
  - 510301, GLUTAMINYL-TRNA SYNTHETASE
  - 361247:TISSUE FACTOR PATHWAY INHIBITOR 2 PRECURSOR Chr.7
  - 470385, Homo sapiens placental bikunin mRNA, complete cds
- **Group 8: Breast Cell Type**
  - 248955, Human mitochondrial 1,25-dihydroxyvitamin D3 24-hydroxylase mRNA
  - 236338:TP53 Tumor protein p53 (Li-Fraumeni syndrome) Chr.17
  - 429540:ELONGATION FACTOR TU, MITOCHONDRIAL PRECURSOR Chr.16

COLLABORATORS

- **LMP, NCI**
  - Larry Smith
  - Lorrie Tanabe
  - Uwe Scherf
  - William Reinhold
  - Yi Zhou
  - John Weinstein
- **Stanford**
  - Douglas Ross
  - Michael Eisen
  - Patrick Brown
  - David Botstein
- **Whitehead**
  - Donna Slonim
  - Jane Staunton
  - Todd Golub
  - Pablo Tamayo
  - Erik Lander

Reproducibility: Microarray MDA-MB-435 vs. MDA-N