

**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

A vertical decorative bar on the left side of the slide, featuring a colorful, abstract pattern of green, blue, and purple. At the top of this bar is a large, semi-transparent triangle with a gradient from blue to purple.

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

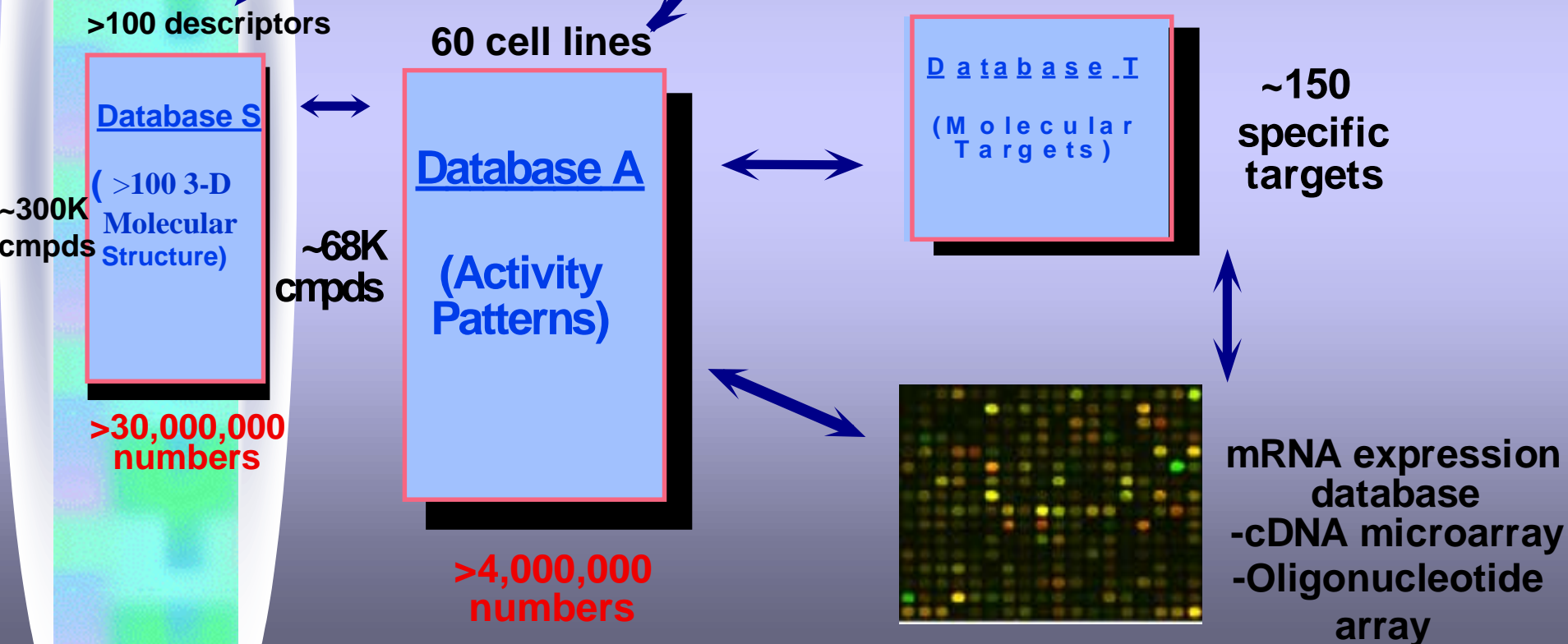
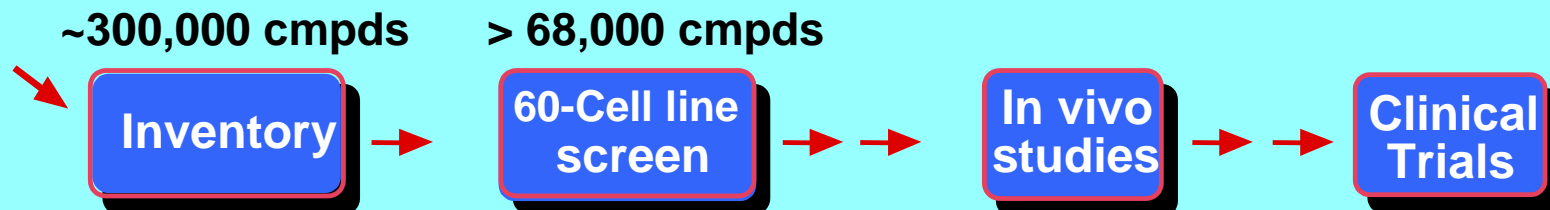
Leukemia

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





60 Cancer Cell Line Screening Data

- **Activity Data (A)**

- Drug Potency Activity; GI₅₀, 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)

A vertical decorative bar on the left side of the slide, featuring a colorful, abstract pattern of green, blue, and purple. At the top of this bar is a large, semi-transparent triangle with a gradient from blue to purple.

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

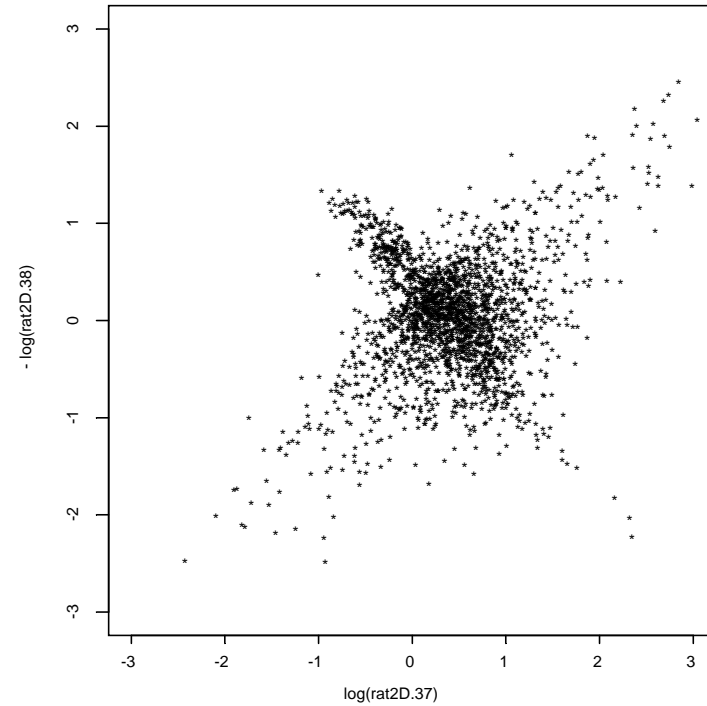
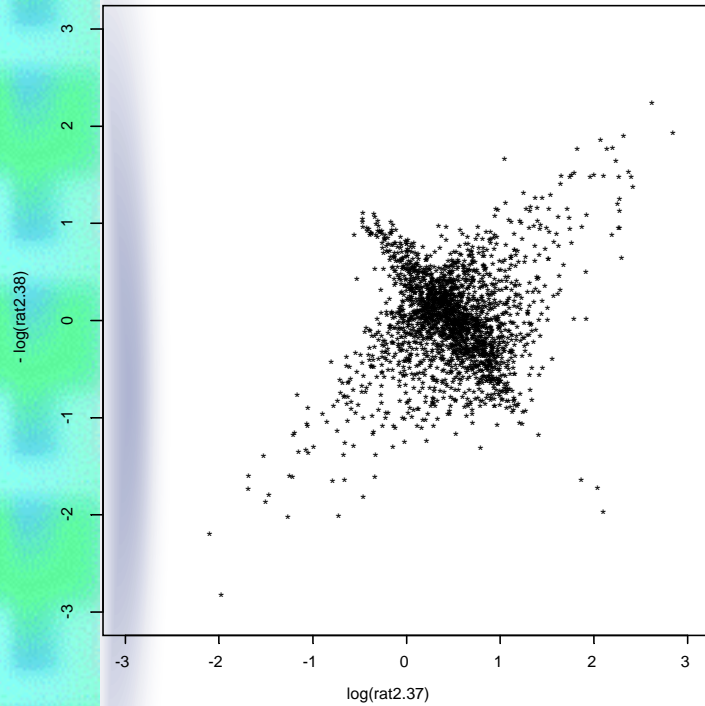
- **subsetting:** depends on inference goals to identify (e.g., Tibshirani et al., 2000)

- 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 $10^{-2} \times 10^{-2}$
 - 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

Cy3 Cy5

I1-T1 Ref

Chip 2

Cy3 Cy5

Ref I1-T2

Chip 3

Cy3 Cy5

Ref I2-T1

Chip 4

Cy3 Cy5

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**

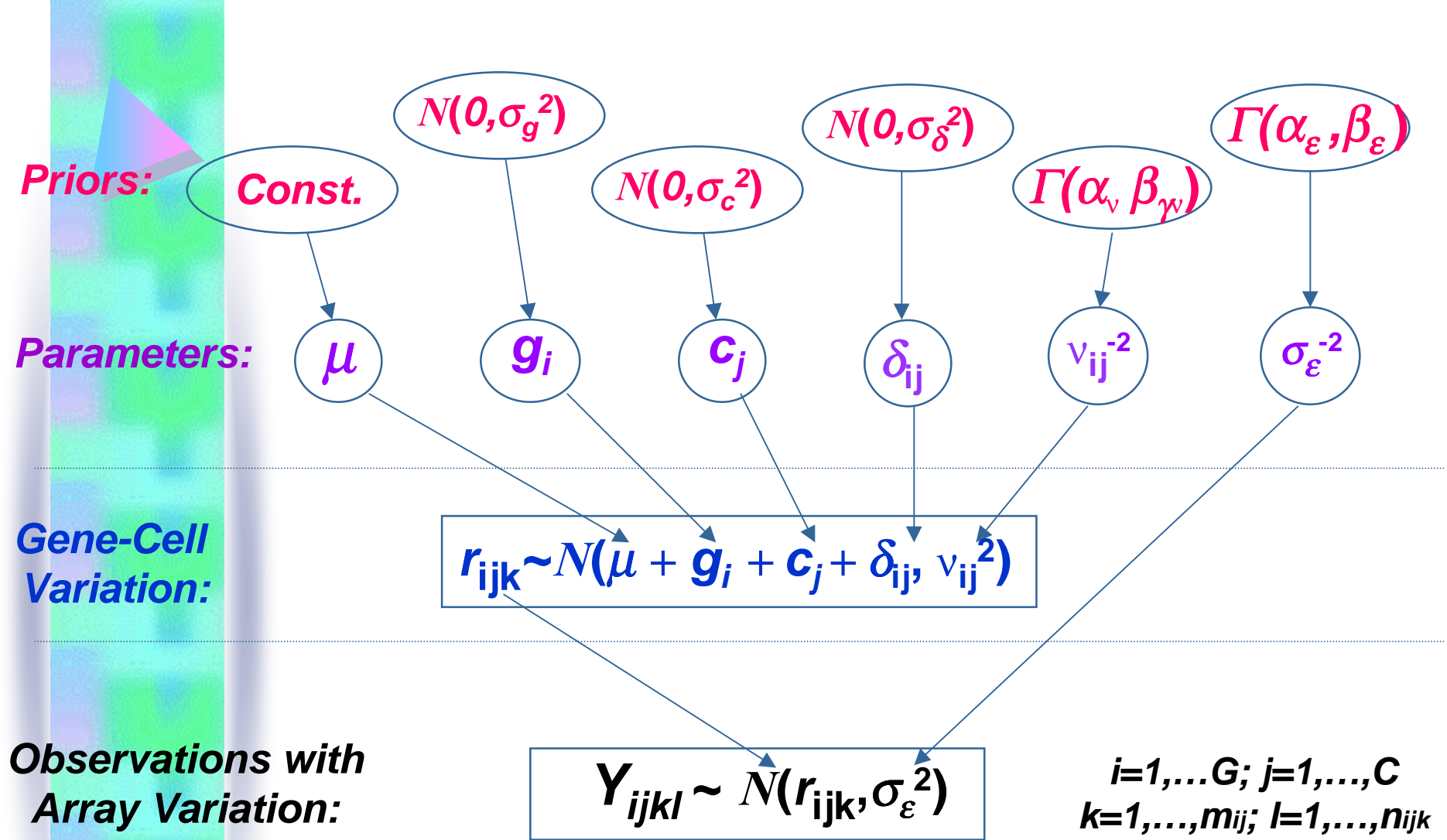
- Layer 1: Array experimental variability:

- $$Y_{ijkl} = r_{ijk} + \varepsilon_{ijkl}, \quad \varepsilon_{ijkl} \sim N(0, \sigma^2), \text{ given } G_{ijk}$$

- Layer 2: Biological variability:

- $$r_{ijk} = G_i + C_j + \delta_{ij} + \alpha_{ijk}, \quad \alpha_{ijk} \sim N(0, v_{ij}^2)$$

- 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

$$\pi(\mu | \text{rest}) = \text{Normal}\left(\sum_{i,j} \frac{\sum_k (r_{i,j,k} - \mu - c_j - \delta_{i,j})}{\frac{n_{i,j}}{\sigma_{r_{i,j}}^2} + \frac{n_{G,C}}{\sigma_{G,C}^2}}, \left(\frac{n_{i,j}}{\sigma_{r_{i,j}}^2} + \dots + \frac{n_{G,C}}{\sigma_{G,C}^2}\right)^{-1}\right)$$

$$\pi(g_i | \text{rest}) = \text{Normal}\left(\sum_j \frac{\sum_k (r_{i,j,k} - \mu - c_j - \delta_{i,j}) / \sigma_{r_{i,j}}^2}{\frac{n_{i,j}}{\sigma_{r_{i,j}}^2} + \frac{n_{G,C}}{\sigma_{G,C}^2} + \frac{1}{\sigma_g^2}}, \left(\frac{n_{i,j}}{\sigma_{r_{i,j}}^2} + \dots + \frac{n_{G,C}}{\sigma_{G,C}^2} + \frac{1}{\sigma_g^2}\right)^{-1}\right),$$

$$\pi(c_j | \text{rest}) = \text{Normal}\left(\sum_i \frac{\sum_k (r_{i,j,k} - \mu - g_i - \delta_{i,j}) / \sigma_{r_{i,j}}^2}{\frac{n_{i,j}}{\sigma_{r_{i,j}}^2} + \frac{n_{G,C}}{\sigma_{G,C}^2} + \frac{1}{\sigma_c^2}}, \left(\frac{n_{i,j}}{\sigma_{r_{i,j}}^2} + \dots + \frac{n_{G,C}}{\sigma_{G,C}^2} + \frac{1}{\sigma_c^2}\right)^{-1}\right),$$

$$\pi(\delta_{i,j} | \text{rest}) = \text{Normal}\left(\frac{\sigma_{r_{i,j}}^2}{\sigma_{r_{i,j}}^2 + \sigma_{\delta_{i,j}}^2} \sum_k \frac{r_{i,j,k} - \mu - g_i - c_j}{n_{i,j}}, \left(\frac{1}{\sigma_{\delta_{i,j}}^2} + \frac{1}{\sigma_{r_{i,j}}^2}\right)^{-1}\right),$$

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

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

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

A vertical decorative bar on the left side of the slide, featuring a colorful, abstract pattern of green, blue, and purple. At the top of this bar is a 3D-style triangle pointing to the right, with a blue top face, a purple left face, and a grey shadow.

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

Genomics and Bioinformatics Group - Microsoft Internet Explorer

Address <http://discover.nci.nih.gov/>

GENOMICS AND BIOINFORMATICS GROUP

Head: John N. Weinstein, M.D., Ph.D.
Laboratory of Molecular Pharmacology
National Cancer Institute
National Institutes of Health

- MICROARRAY DATA & TOOLS
- MOLEC INTERACTION MAPS
- GROUP RESOURCES
- OPPORTUNITIES

Username and Password Required

Enter username for access to internal documents at discover.nci.nih.gov:

User Name:

Password:

OK Cancel

The mission of the Genomics and Bioinformatics Group is to understand the complex molecular pharmacology of cancer cells and to find new agents for treatment of cancer. The studies are 50% experimental, 50% bioinformatic.

Experimental: Characterization at the DNA, mRNA, and protein expression levels of cancer cells, including those used in the National Cancer Institute's drug discovery program. Central methods are those of molecular biology, genomics, and proteomics.

Internet

Interactive Analysis Pages

John Weinstein - Netscape
File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Stop

Bookmarks Netsite: <http://discover.nci.nih.gov/internal/> What's Related

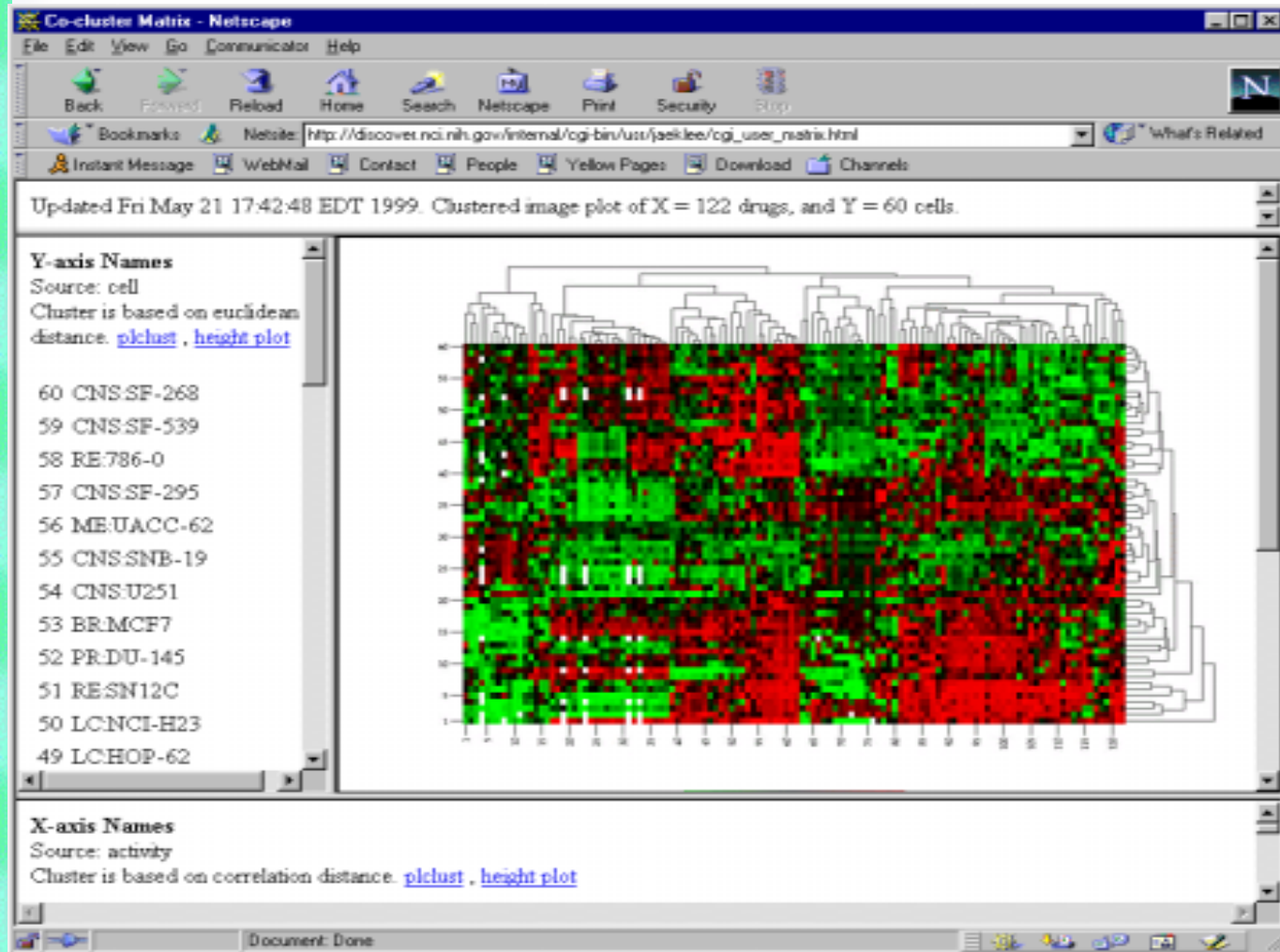
Instant Message WebMail Contact People Yellow Pages Download Channels

Whitehead: cutoff 0.6 by standardization within each of 60 cell lines [Standard0.6](#)

Interactive web tool I:	Data mining, exploration, and discovery	
	3204 gene:drug links (Stanford)	RND
	gene:drug pathways (Stanford)	RND
	1203 gene:drug links (Whitehead)	RND
	gene:drug pathways (Whitehead)	RND
	generates a cluster image	RND
	resort an axis of a cluster page	RND
	pattern search	RND
	experiment data for 6215 drugs	RND
Interactive web tool II:	Identification of entities with high association	
	gene-drug correlations	RND
	correlation search analysis	RND
	probabilistic correlation combination search	RND
	list threshold correlations	RND
Interactive web tool III:	Inference and literature search	
	gene/gene pubmed - keyword (Stanford)	RND
	gene/gene pubmed - keyword (Whitehead)	RND
	gene/gene pubmed - cloneid or accession	RND
	gene/drug pubmed - keyword (Stanford)	RND
	gene/drug pubmed - keyword (Whitehead)	RND
	gene/drug pubmed - HSC, cloneid or accession	RND
	general correlation	RND
	multi correlation analysis	RND
	cross correlation analysis	RND
	estimation of gene & drug effects by HEM	RND
	critical levels of correlation	RND
	demo applet	RND

Document Done

Cluster-image analysis result



Estimation of Gene & Drug Effects by GS-HEM

Estimation of Gene & Drug Effects by Hierarchical Effects Model - Net...

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security

Bookmarks Netsite: r.nci.nih.gov/internal/J-hem.html What's Related

Instant Message WebMail Contact People Yellow Pages Down

Estimation of Gene & Drug Effects: Markov Chain Monte Carlo on Hierarchical Effects Model (HEM)

Identify and estimate the effects of elements ordered by the magnitude of significance. Estimate the effects of each element based the hierarchical effects model (HEM) within each group and return the entries with highest normal scores (estimate/standard dev). The grouping can be chosen as one of preset options or arbitrary choices (Need to specify both Data I and Data II, if want to do on ATmatrix.)

(Note: if the post time is expired for some big data file, do NOT re-execute the page using 'reload' button. Wait awhile, and try to OPEN url: 'http://discover.nci.nih.gov/internal/cgi-bin/Ausr/(your username)/out.html')

Main Data:

Median Activity Median Activity 118 Extended Target
 Activity Stanford (confirmed) genes Stanford (3,204) genes
 Kalama genes Bigdata (I&D) genes Bigdata (1,200) genes
 Whitehead genes Target Log Target

MCMC Implementation & Output Options

1. Direction of search for normal scores: positive; negative; both;

2. Number of output entries with high normal scores:

3. Burn-in iteration for Markov chain coverage:

4. Size of Markov chain sample (length of MCMC run):

6. Save MCMC sample for parameters of main effects: No Yes

7. Save MCMC sample for parameters of interaction effects: No Yes

8. Normal variances for mean parameters
Gene-Cell: Gene or Drug: Cell:

9. Gamma parameters for variance parameters
Gene-Cell: α : β : Error: α : β :
(multiplicative factors for hyperparameters of priors)

Document: Done

Estimation of Gene & Drug Effects by Hierarchical Effects Model - Net...

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security

Bookmarks Netsite: http://discover.nci.nih.gov/inter What's Related

Instant Message WebMail Contact People Yellow Pages Down

Cell Grouping:

Cancer cell type:

Group 1: Group 2: Group 3:
 Group 4: Group 5: Group 6:
 Group 7: Group 8: Group 9:

Group Allocation for Cells

<input type="checkbox"/> ME:LOXIMVI	<input type="checkbox"/> ME:MALME-3M	<input type="checkbox"/> ME:SK-MEL-2	<input type="checkbox"/> ME:SK-MEL-5
<input type="checkbox"/> ME:SK-MEL-28	<input type="checkbox"/> LC:NCI-H23	<input type="checkbox"/> ME:M14	<input type="checkbox"/> ME:UACC-62
<input type="checkbox"/> LC:NCI-H522	<input type="checkbox"/> LC:A549/ATCC	<input type="checkbox"/> LC:EKVX	<input type="checkbox"/> LC:NCI-H322M
<input type="checkbox"/> LC:NCI-H460	<input type="checkbox"/> LC:HOP-62	<input type="checkbox"/> LC:HOP-92	<input type="checkbox"/> CNS:SNB-19
<input type="checkbox"/> CNS:SNB-75	<input type="checkbox"/> CNS:U251	<input type="checkbox"/> CNS:SF-268	<input type="checkbox"/> CNS:SF-295
<input type="checkbox"/> CNS:SF-539	<input type="checkbox"/> CO:HT29	<input type="checkbox"/> CO:HCC-2998	<input type="checkbox"/> CO:HCT-116
<input type="checkbox"/> CO:SW-620	<input type="checkbox"/> CO:HCT-15	<input type="checkbox"/> CO:KM12	<input type="checkbox"/> OV:OVCA-3
<input type="checkbox"/> OV:OVCA-4	<input type="checkbox"/> OV:OVCA-8	<input type="checkbox"/> OV:IGROV1	<input type="checkbox"/> OV:SK-OV-3
<input type="checkbox"/> LE:CCRF-CEM	<input type="checkbox"/> LE:K-562	<input type="checkbox"/> LE:MOLT-4	<input type="checkbox"/> LE:SR
<input type="checkbox"/> RE:UO-31	<input type="checkbox"/> RE:SN12C	<input type="checkbox"/> RE:A498	<input type="checkbox"/> RE:CAKI-1
<input type="checkbox"/> RE:RXF-393	<input type="checkbox"/> RE:786-0	<input type="checkbox"/> RE:ACHN	<input type="checkbox"/> RE:TK-10
<input type="checkbox"/> ME:UACC-257	<input type="checkbox"/> LC:NCI-H226	<input type="checkbox"/> CO:COLO205	<input type="checkbox"/> OV:OVCA-5
<input type="checkbox"/> LE:HL-60	<input type="checkbox"/> LE:RPMI-8226	<input type="checkbox"/> BR:MCF7	<input type="checkbox"/> BR:MCF7/ADP-RES
<input type="checkbox"/> PR:PC-3	<input type="checkbox"/> PR:DU-145	<input type="checkbox"/> BR:MDA-MB-231/ATCC	<input type="checkbox"/> BR:HS578T
<input type="checkbox"/> BR:MDA-MB-435	<input type="checkbox"/> BR:MDA-N	<input type="checkbox"/> BR:BT-549	<input type="checkbox"/> BR:T-47D

Data type:

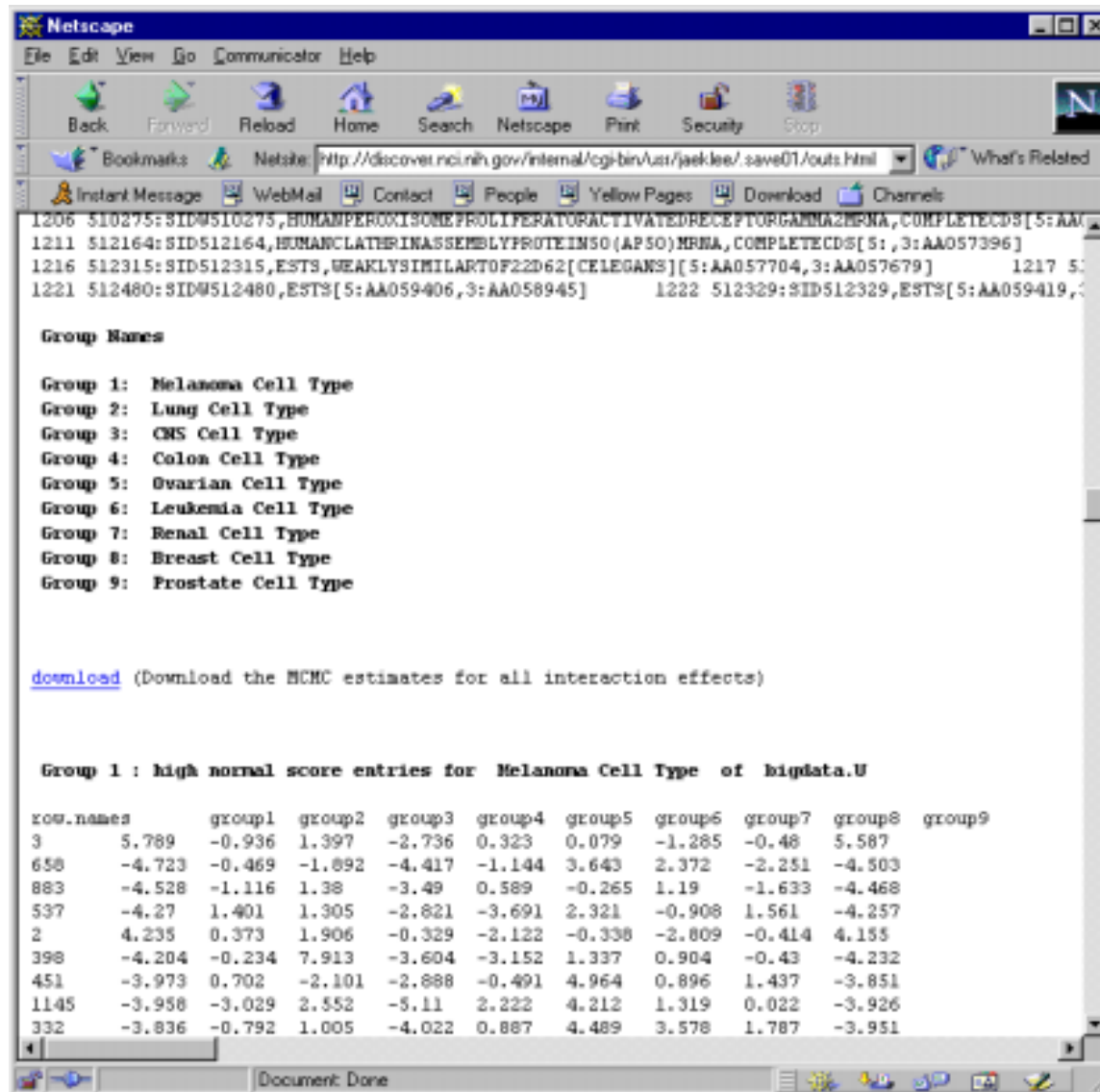
Gene (or Drug) standardization type:

Save List of Selected Entries: Save Directory:

Combining Data:

Document: Done

HEM Result: Estimation of Gene-Cell Interaction Effects



The screenshot shows a Netscape browser window with the following content:

Address bar: <http://discover.nci.nih.gov/internal/cgi-bin/ust/jaek.lee/save01/out5.html>

Text content:

```
1206 510275:SID510275,HUMANPEROXISOMEPROLIFERATORACTIVATEDRECEPTORGAMMA2MRNA,COMPLETECDS[5:AA057396]
1211 512164:SID512164,HUMANCLATHERINASSEMBLYPROTEIN50(AP50)MRNA,COMPLETECDS[5:3:AA057396]
1216 512315:SID512315,ESTS,WEAKLYSIMILARTOP22D62[CELEGANS][5:AA057704,3:AA057679] 1217 512315:SID512315,ESTS,WEAKLYSIMILARTOP22D62[CELEGANS][5:AA057704,3:AA057679]
1221 512480:SID512480,ESTS[5:AA059406,3:AA058945] 1222 512329:SID512329,ESTS[5:AA059419,3:AA058945]
```

Group Names

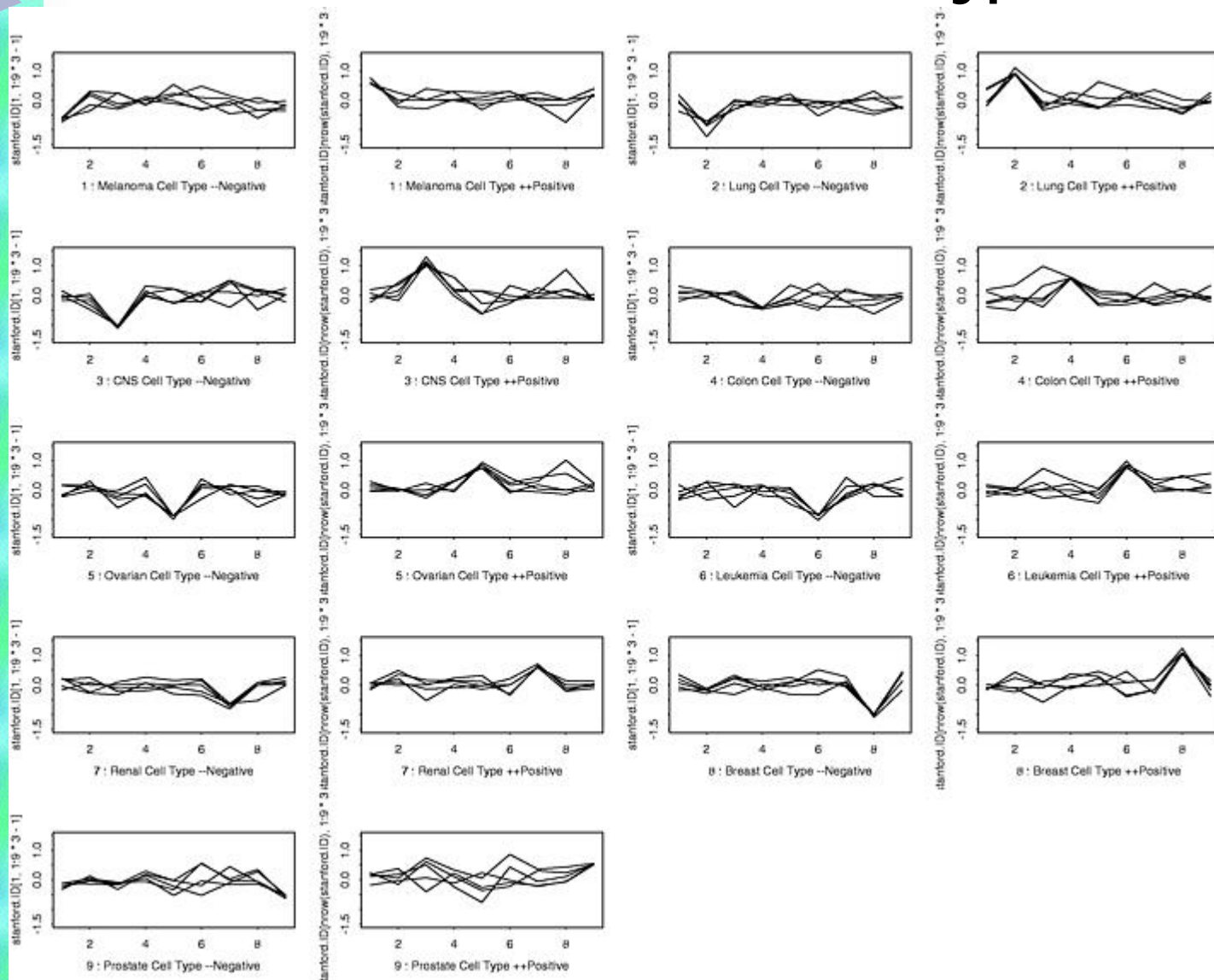
- Group 1: Melanoma Cell Type
- Group 2: Lung Cell Type
- Group 3: CNS Cell Type
- Group 4: Colon Cell Type
- Group 5: Ovarian Cell Type
- Group 6: Leukemia Cell Type
- Group 7: Renal Cell Type
- Group 8: Breast Cell Type
- Group 9: Prostate Cell Type

[download](#) (Download the MNC estimates for all interaction effects)

Group 1 : high normal score entries for Melanoma Cell Type of bigdata.U

row.names	group1	group2	group3	group4	group5	group6	group7	group8	group9
3	5.789	-0.936	1.397	-2.736	0.323	0.079	-1.285	-0.48	5.587
658	-4.723	-0.469	-1.892	-4.417	-1.144	3.643	2.372	-2.251	-4.503
883	-4.528	-1.116	1.38	-3.49	0.589	-0.265	1.19	-1.633	-4.468
537	-4.27	1.401	1.305	-2.821	-3.691	2.321	-0.908	1.561	-4.257
2	4.235	0.373	1.906	-0.329	-2.122	-0.338	-2.809	-0.414	4.155
398	-4.204	-0.234	7.913	-3.604	-3.152	1.337	0.904	-0.43	-4.232
451	-3.973	0.702	-2.101	-2.888	-0.491	4.964	0.896	1.437	-3.851
1145	-3.958	-3.029	2.552	-5.11	2.222	4.212	1.319	0.022	-3.926
332	-3.836	-0.792	1.005	-4.022	0.887	4.489	3.578	1.787	-3.951

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

