Dealing with Microarray Data

NIDDK Meeting

10 July 2001

TIGR
The Institute for Genomic Research

Science is built with facts as a house is with stones – but a collection of facts is no more a science than a heap of stones is a house.

– Jules Henry Poincare

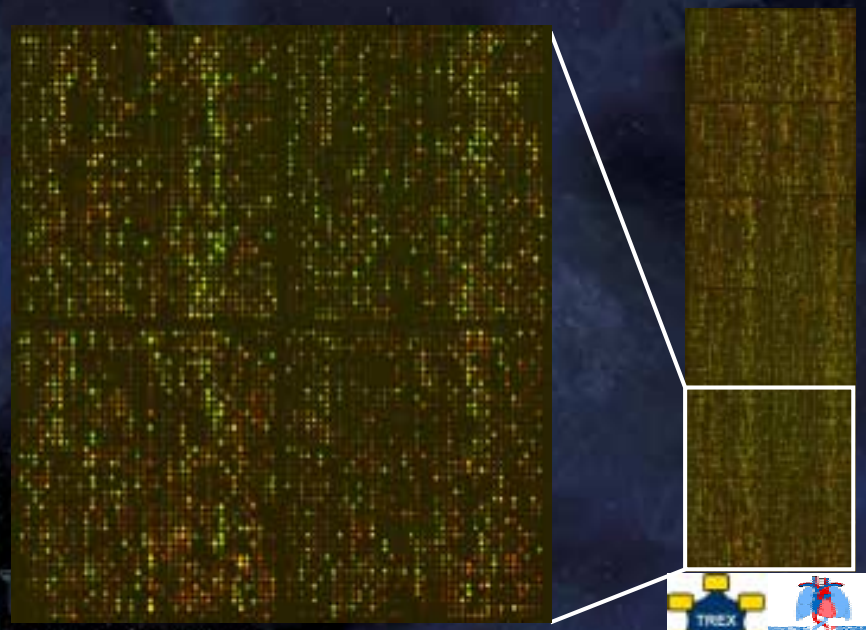
TIGR
The Institute for Genomic Research

The Golden Age of Genomics

- ~50 Microbial Genomes have been sequenced, at least 50 more are on the way
- Yeast, *C. elegans*, *Arabidopsis*, *Drosophila* and other Eukaryotic models are finished or well advanced
- A "working draft" of the Human Genome Sequence is now available, with mouse and rat to follow
- More than 8,000,000 ESTs are available; more than 3,600,000 from humans

TIGR
The Institute for Genomic Research

Hybridization to a 19,200 Element Human Array

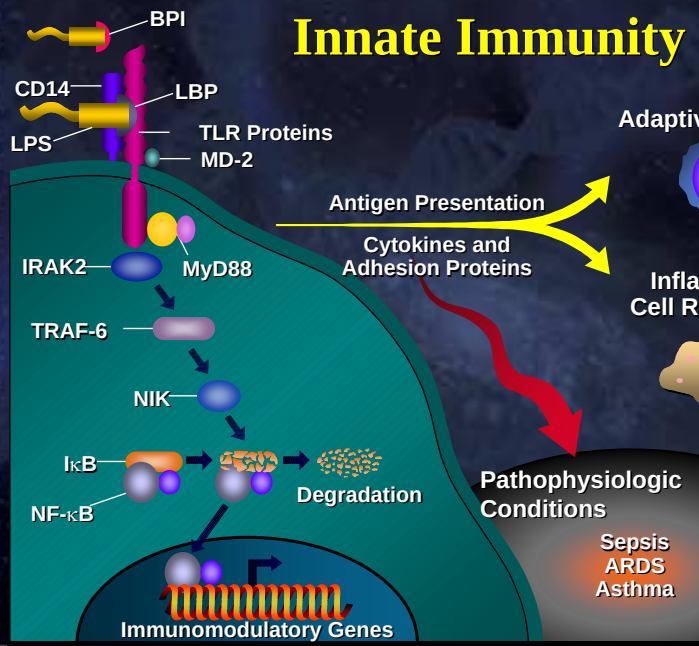


TIGR
The Institute for Genomic Research

Named Differentially Expressed Genes

Priti Hegde

Innate Immunity



David Schwartz

Adapted from Godowski. *NEJM* 1999; 340:1835

Microarray Expression Profiling of Rodent Models of Human Disease

The Theme of our PGA:

Examining Gene/Environment Interactions in Rodent Models of Human Disease Using cDNA Microarrays to link Phenotype to Genotype

See <http://pga.tigr.org> for information, data, tools, and Visiting Scientist program

Challenges in HLB Research

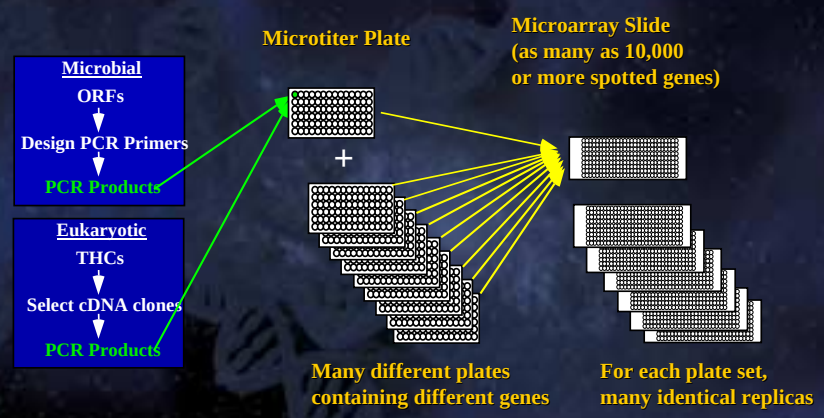
- Identifying appropriate models
- Identifying genes involved in HLB disorders
- Placing these genes into relevant pathways
- Separating genetic and environmental components
- Developing an understanding of the mechanisms

Expression Profiling: cDNA Microarrays

- **The Promise:**
  - Genome Scale Data on Gene Expression Patterns
- **The Challenges:**
  - Deduce Gene Function from Expression
  - Elucidate Functional Pathways from Function and Expression
- **The Solutions:**
  - Development of Robust Protocols for large-scale correlation studies
  - Relational Database for Data Storage
  - Data Visualization and Mining Tools
  - Tools for modeling pathways

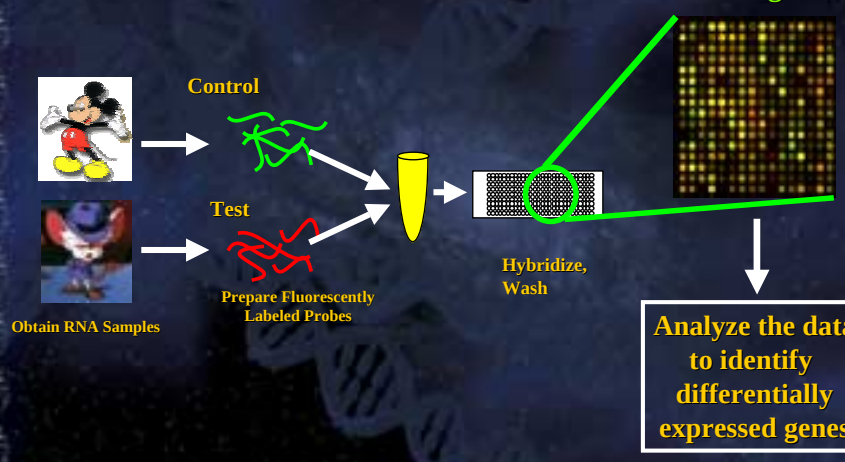
TIGR
The Institute for Genomic Research

Microarray Overview I



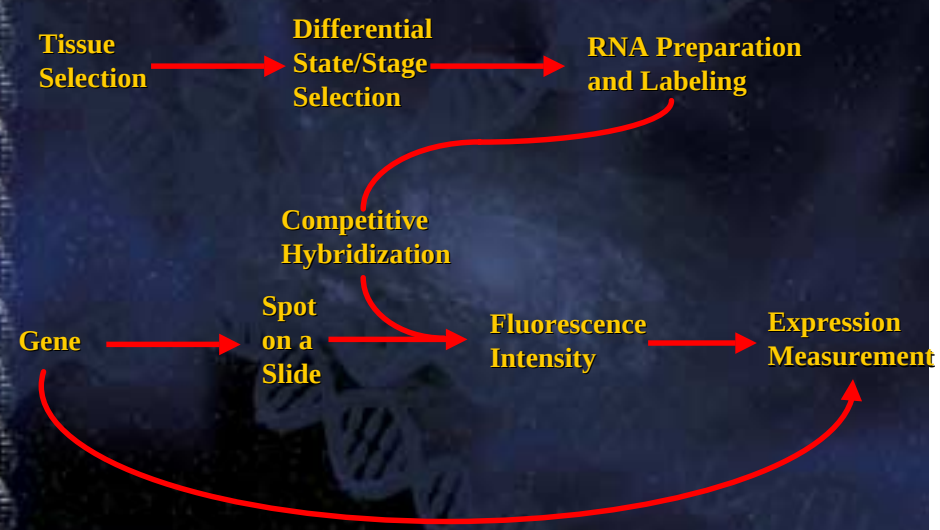
TIGR
The Institute for Genomic Research

Microarray Overview II



TIGR
The Institute for Genomic Research

Microarray Expression Analysis



TIGR
The Institute for Genomic Research

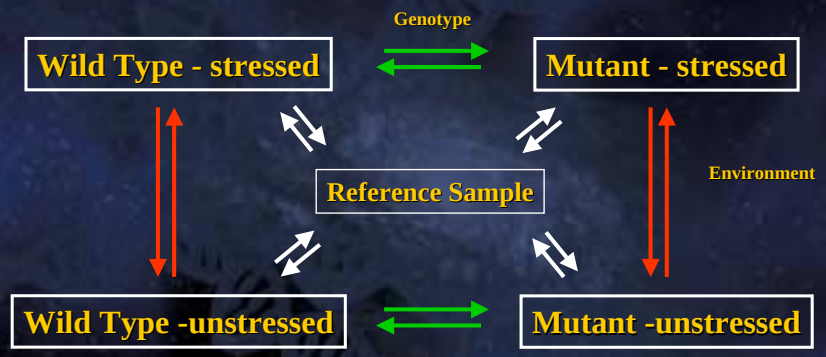
The Beast: Microarray Robot from Intelligent Automation

<http://www.ias.com>

Axon GenePix 4000B Scanner

<http://www.axon.com/GN_GenePix4000.html>

Basic Experimental Paradigm



Developmental Goals

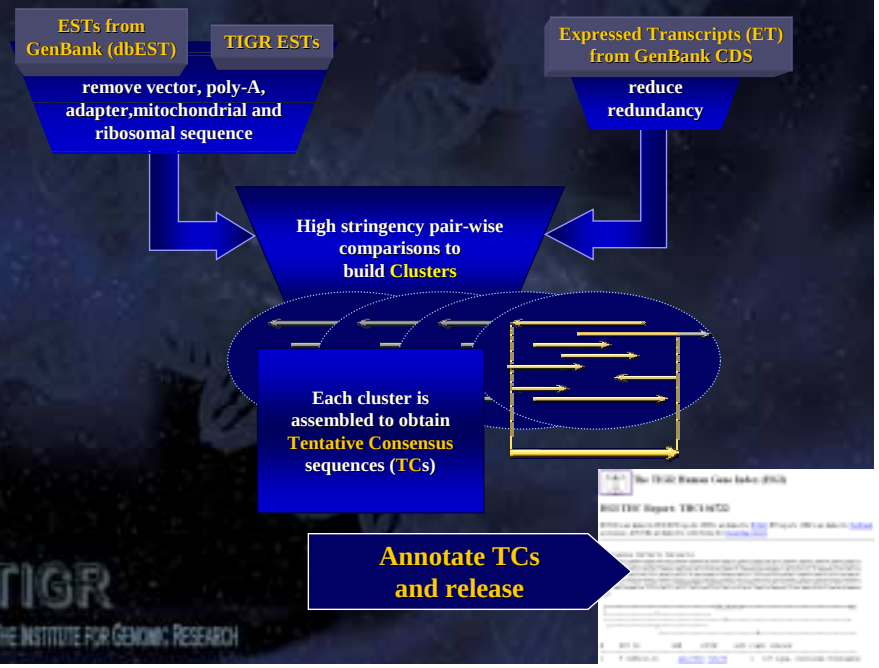
- Tools for linking Genes, ESTs, cDNAs, and Sequences
- Microarray Quality Control Protocols and Reagents
- Novel Analysis Techniques and Tools
- Phenotyping Pipelines
- Microarray Resources in Mouse and Rat
- Expression Profiles for Disease Phenotypes of HLBS interest

Developmental Goals

- Tools for linking Genes, ESTs, cDNAs, and Sequences
- Microarray Quality Control Protocols and Reagents
- Novel Analysis Techniques and Tools
- Phenotyping Pipelines
- Microarray Resources in Mouse and Rat
- Expression Profiles for Disease Phenotypes of HLBS interest

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

Gene Index Assembly process



TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

TIGR Gene Indices home page

www.tigr.org/tdb/tgi.shtml

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

The Mouse Gene Index <http://www.tigr.org/tdb/mgi>

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

A TC Example

Click here to see the live web page at TIGR

Ancillary Information

- Annotation based on gene content
- Annotation based on protein similarity searches
- "Opposite End" information
- Links to mapping information
- Links to Tentative Orthologue Groups (TOGs)

- Links to Gene Ontology (GO) Terms (in release)
- Links to EC Numbers (in release)
- Links to completed genomic sequence (in release)

- Links to species-specific databases – MGD, RGD (in testing)

Can ESTs be used to identify orthologues?

Property	Rat-Human Mean	(SD)	Mouse-Human Mean	(SD)	Mouse-Rat Mean	(SD)
5'UTR						
% Identity	68.4	(13.0)	69.7	(12.9)	84.5	(12.9)
Mutation distance K	0.486	(0.260)	0.493	(0.273)	0.212	(0.224)
CDS						
% Identity (protein)	88.0	(11.8)	86.4	(12.3)	94.5	(6.3)
% Identity (DNA)	85.9	(6.0)	85.2	(6.5)	93.8	(3.2)
Syn. distance K_s	4.60	(0.245)	0.468	(0.169)	0.166	(0.061)
Nonsyn. distance K_a	0.078	(0.095)	0.090	(0.102)	0.031	(0.040)
3' UTR						
% Identity	70.1	(11.4)	71.0	(12.2)	86.3	(8.9)
Mutation distance K	0.435	(0.212)	0.447	(0.225)	0.164	(0.152)

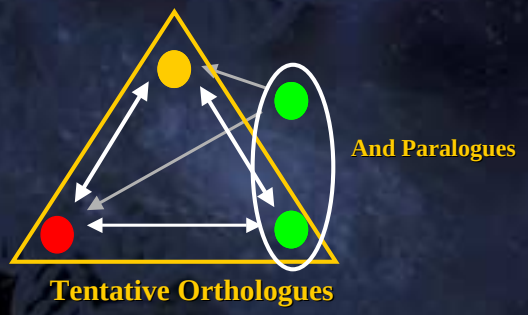
Analysis of 1,880 rodent-human orthologue pairs:
1,212 rat-human pairs, 1,138 mouse-human pairs, and 470 shared by all three.
Makałowski and Boguski (1998) *Proc. Natl. Acad. Sci. USA* 95, 9407-9412

Can ESTs be used to identify orthologues?

Property	Rat-Human Mean	(SD)	Mouse-Human Mean	(SD)	Mouse-Rat Mean	(SD)
5'UTR						
% Identity	68.4	(13.0)	69.7	(12.9)	84.5	(12.9)
Mutation distance K	0.486	(0.260)	0.493	(0.273)	0.212	(0.224)
CDS						
% Identity (protein)	88.0	(11.8)	86.4	(12.3)	94.5	(6.3)
% Identity (DNA)	85.9	(6.0)	85.2	(6.5)	93.8	(3.2)
Syn. distance K_s	4.60	(0.245)	0.468	(0.169)	0.166	(0.061)
Nonsyn. distance K_a	0.078	(0.095)	0.090	(0.102)	0.031	(0.040)
3' UTR						
% Identity	70.1	(11.4)	71.0	(12.2)	86.3	(8.9)
Mutation distance K	0.435	(0.212)	0.447	(0.225)	0.164	(0.152)

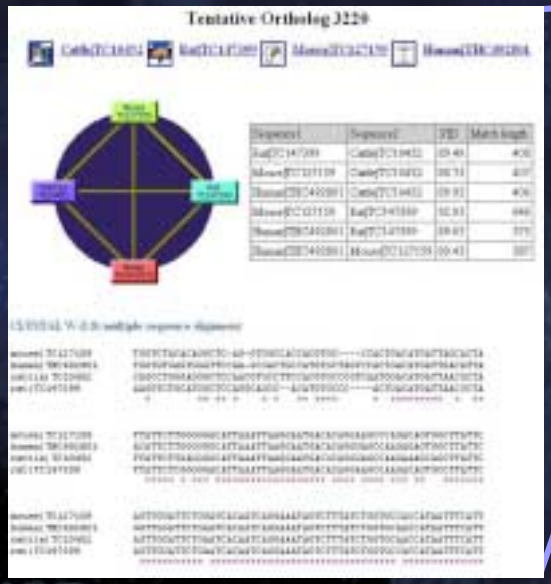
Analysis of 1,880 rodent-human orthologue pairs:
1,212 rat-human pairs, 1,138 mouse-human pairs, and 470 shared by all three.
Makałowski and Boguski (1998) *Proc. Natl. Acad. Sci. USA* 95, 9407-9412

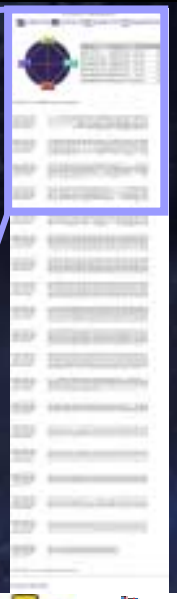
Building TOGs: Reflexive, Transitive Closure

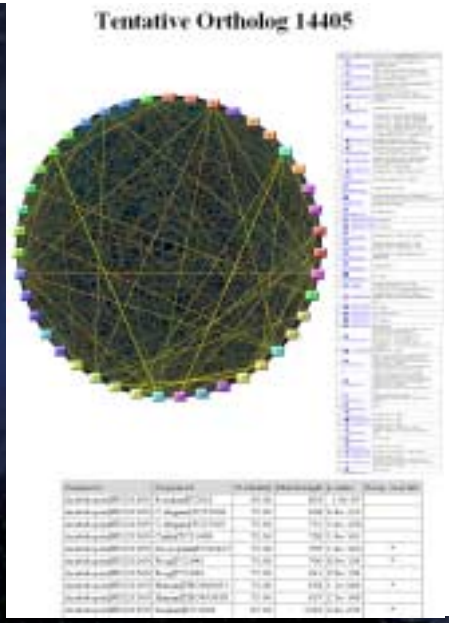


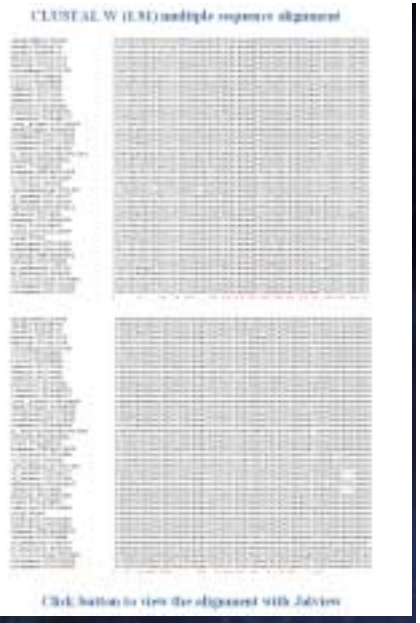
The TIGR Gene Indices <http://www.tigr.org.tdb/tdb/tgi.html>

TOGA: An Sample Alignment: bithoraxoid-like protein

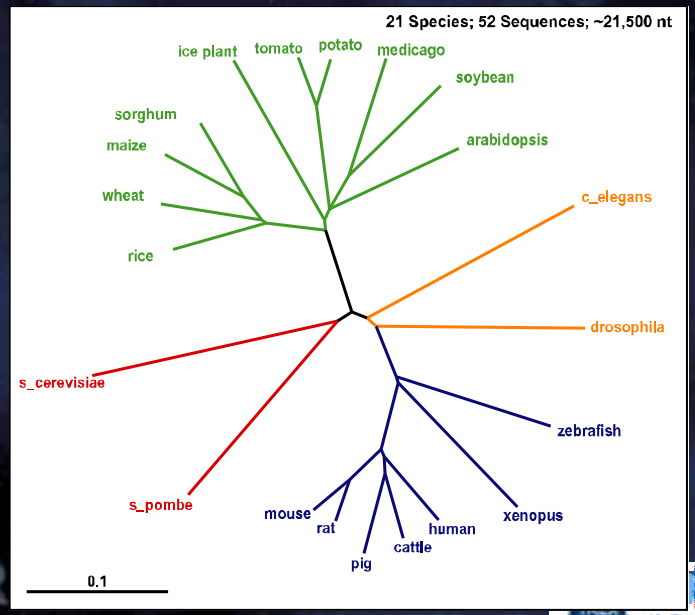


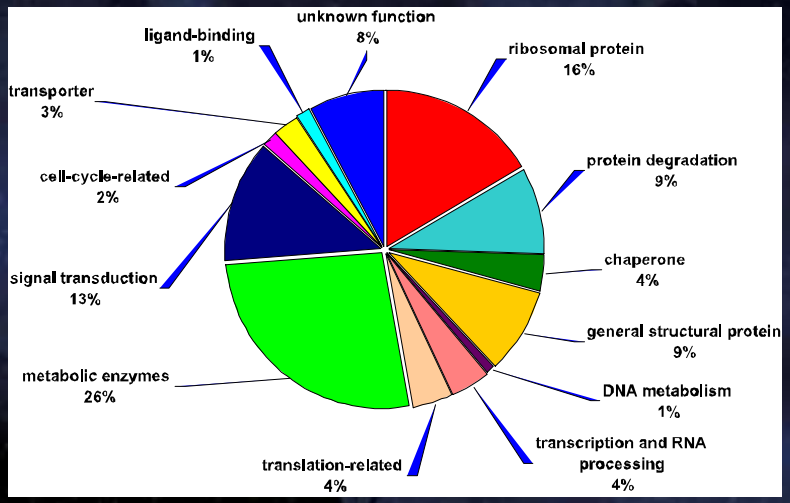




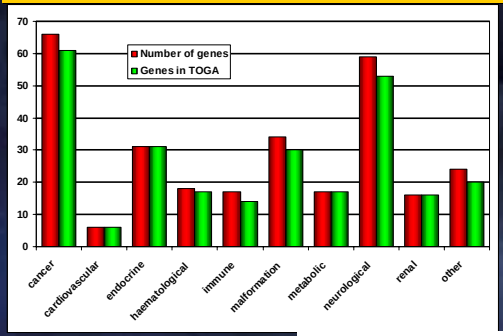


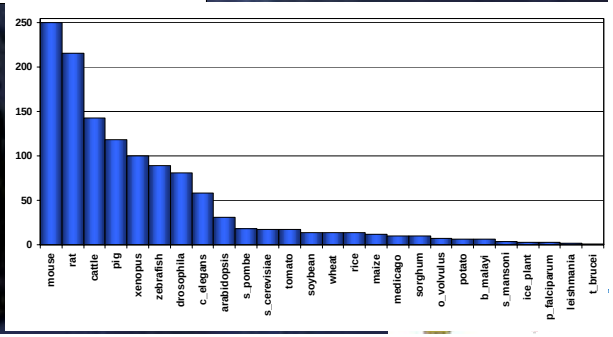
Relationships in TOGA





Representation of 288 Human Disease Genes in TOGA





Arabidopsis thaliana: Alignments with Chromosome II

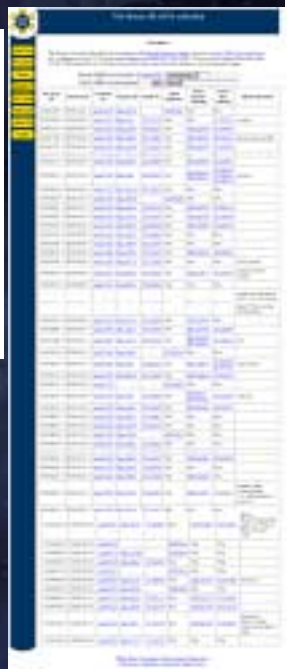
http://www.tigr.org/tdb/at/alignTC.html

Links from Public Clone Sets to TGI

http://pga.tigr.org/tools.shtml

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

Links from Public Clone Sets to TGI



http://pga.tigr.org/tools.shtml

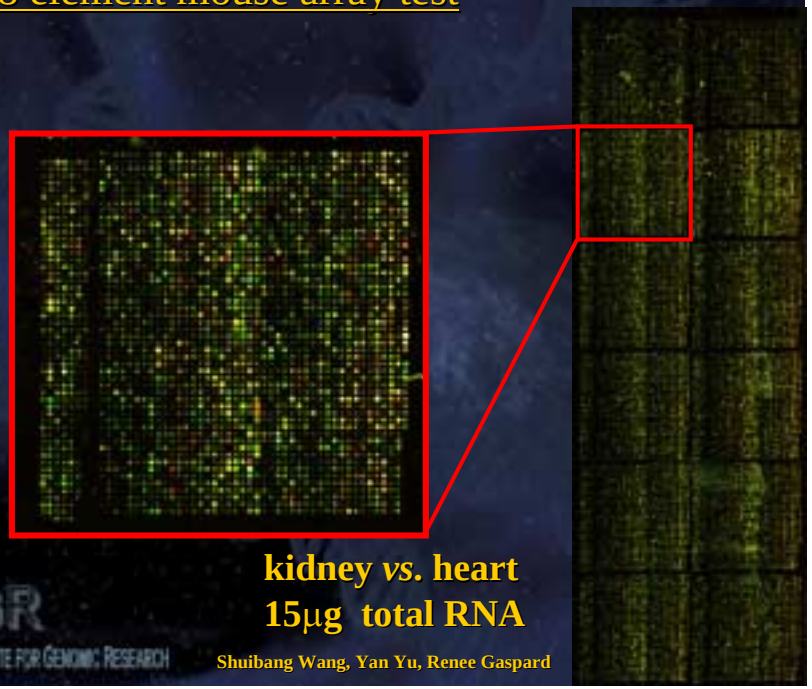
TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

RESOURCERER

http://pga.tigr.org/tools.shtml

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

32,448 element mouse array test

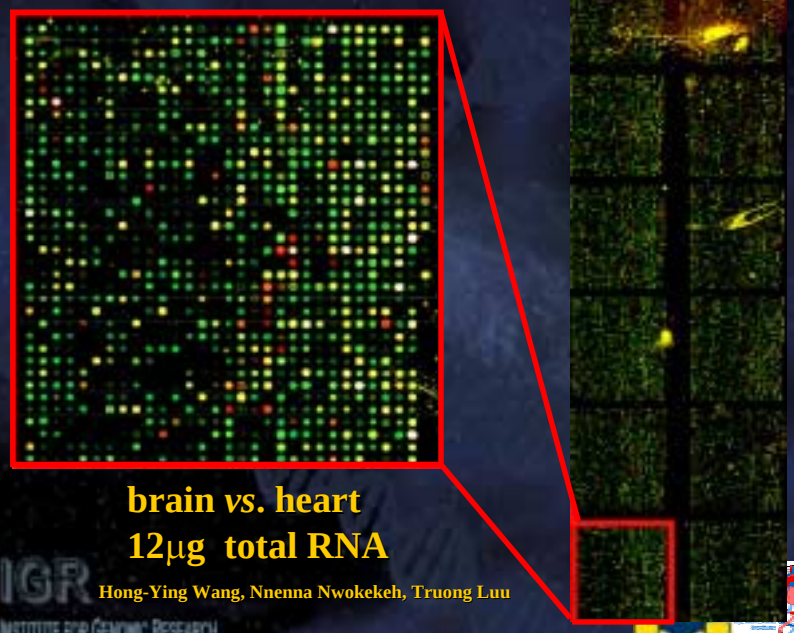


kidney *vs.* heart
15μg total RNA

Shuibang Wang, Yan Yu, Renee Gaspard

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

13,536 element rat array test

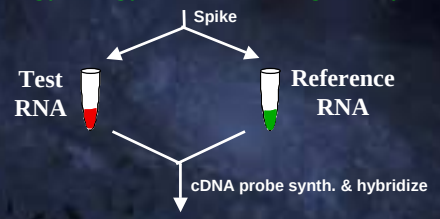


brain *vs.* heart
12μg total RNA

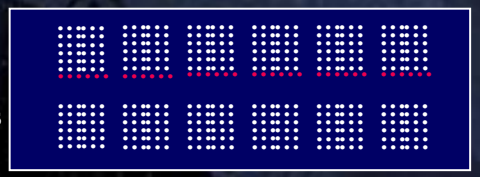
Hong-Ying Wang, Nnenna Nwokekeh, Truong Luu

Spiking Test Samples with *A.thaliana* Control RNAs

Spiked 2-fold change (copies/cell)					Spiked 3-fold change (copies/cell)				
RCA	Cab	rbcL	LTP4	LTP6	XCP2	RPC1	NAC1	TIM	PRK
2	10	60	100	300	3	15	60	150	300
1	5	30	50	150	1	5	20	50	100

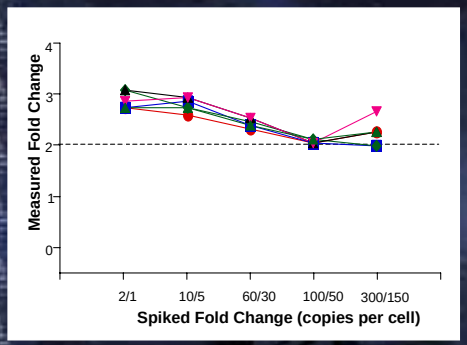


Array containing DNA controls

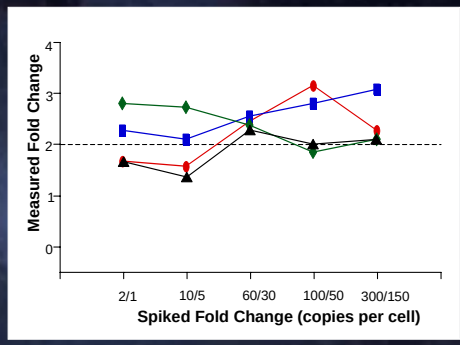


Assessing Sensitivity of Microarrays at 2:1 ratios

Intra-slide variability across 6 grids:

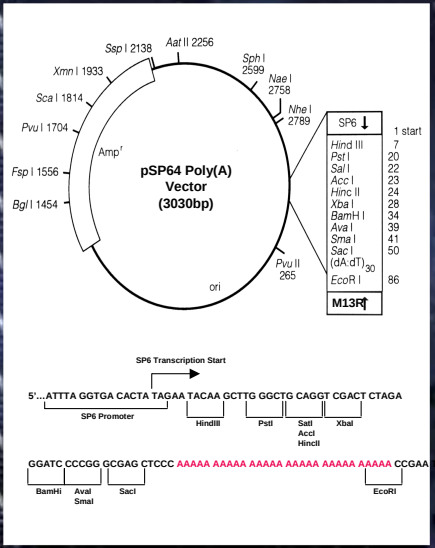


Inter-slide variability:



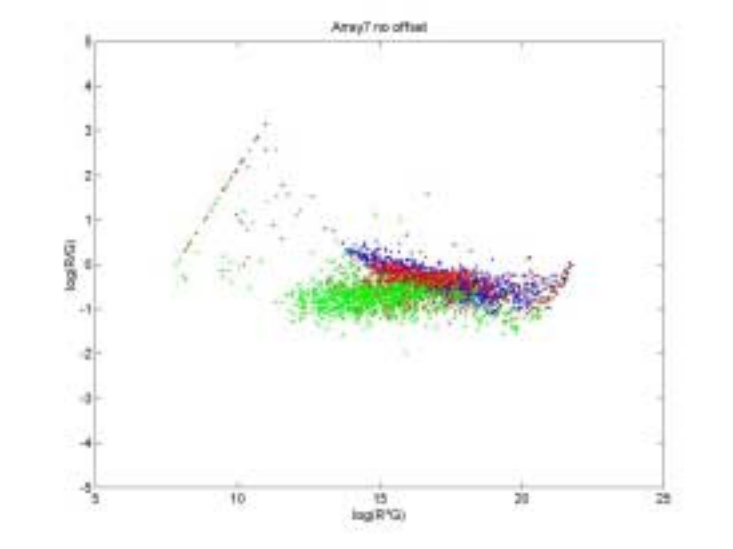
- Rat cDNA chip
- Total RNA from rat brain and heart
- A. thaliana RNA spiking controls: CAB, RCA, rbcL, LTP4, LTP6
- Hong-Ying Wang & Renae Malek

Resource: A. *thaliana* DNA clones for spiking



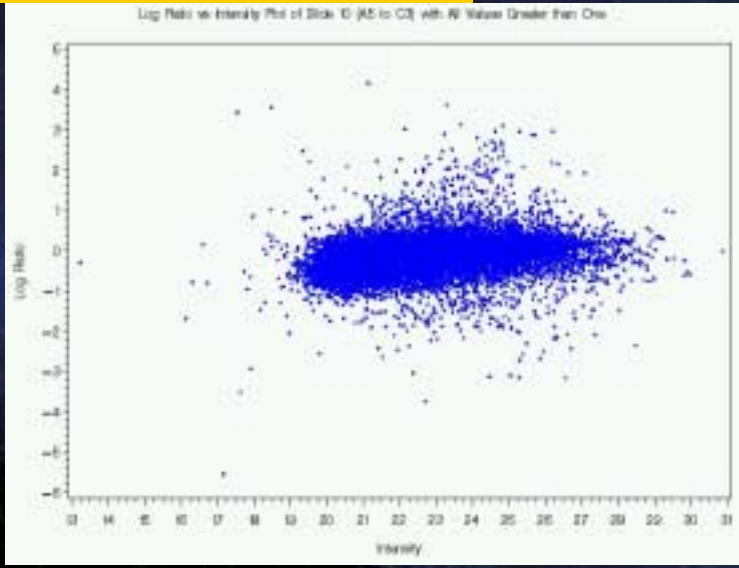
- chlorophyll a/b binding protein (Cab)
- RUBISCO activase (RCA)
- ribulose-1,5-bisphosphate carboxylase/oxygenase (RbcL)
- lipid transfer protein 4 (LTP4)
- lipid transfer protein 6 (LTP6)
- papain-type cysteine endopeptidase (XCP2)
- root cap 1 (RCP1)
- NAC1
- triosphosphate isomerase (TIM)
- ribulose-5-phosphate kinase (PRKase)

Bad Data from Parts Unknown

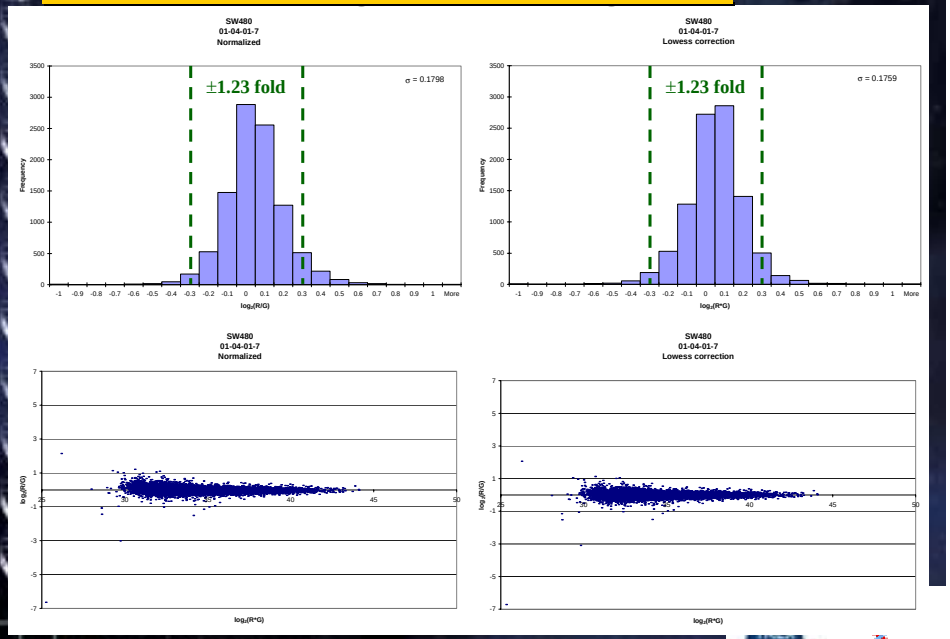


Gary Churchill

Good Data from TREX

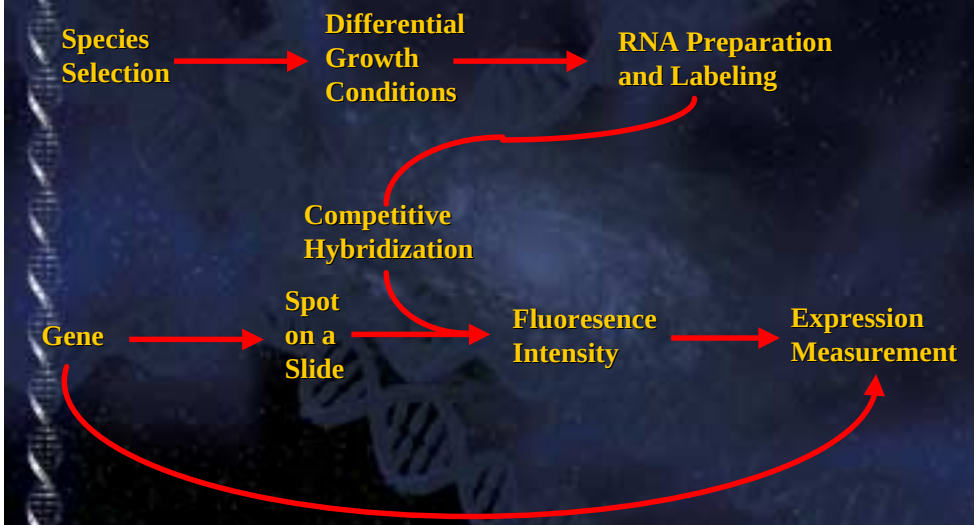


Normalization using local linear regression

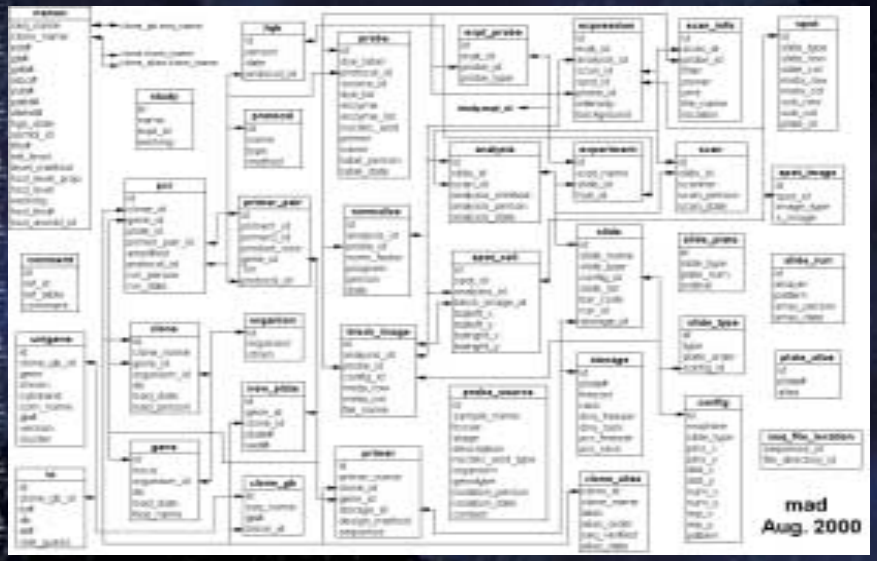


Ivana Yang , John Quackenbush

Microarray Expression Analysis

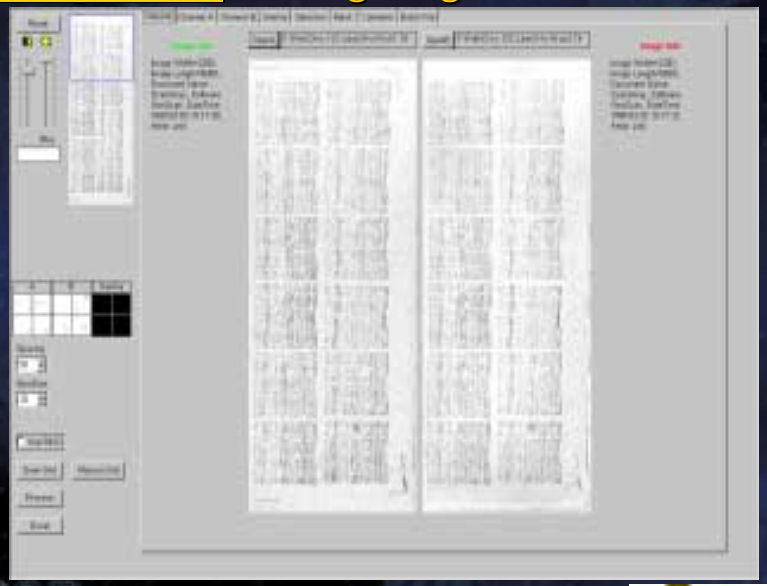


MAD Microbial Array Database Schema



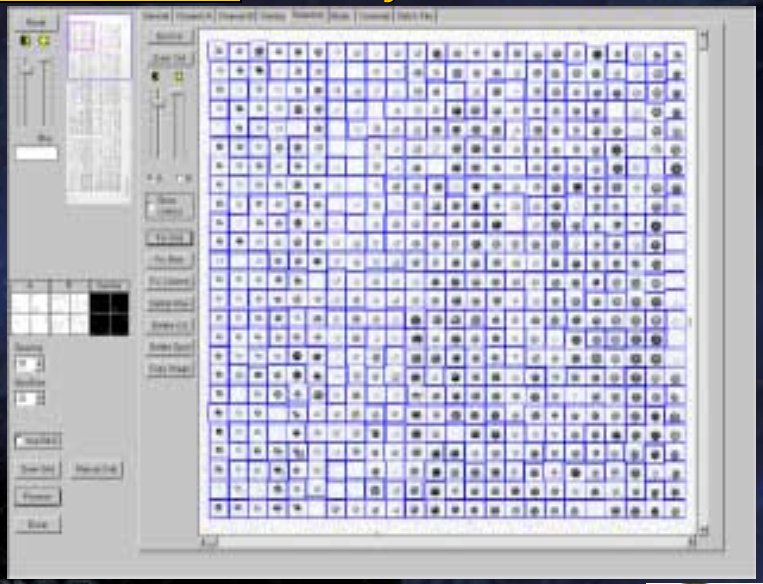
TIGR
The Institute for Genomic Research

TIGR Spotfinder Loading Image Data



TIGR
The Institute for Genomic Research

TIGR Spotfinder Grid Adjustment



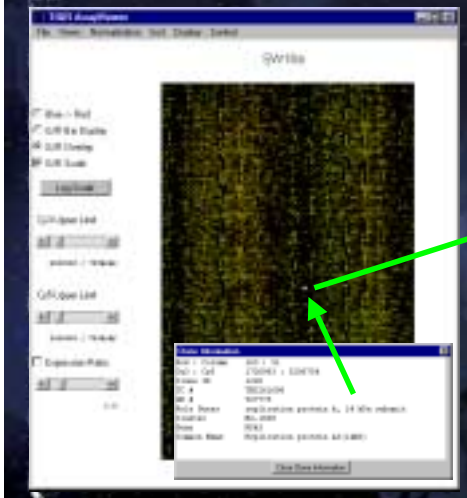
TIGR
The Institute for Genomic Research

Data Analysis Issues

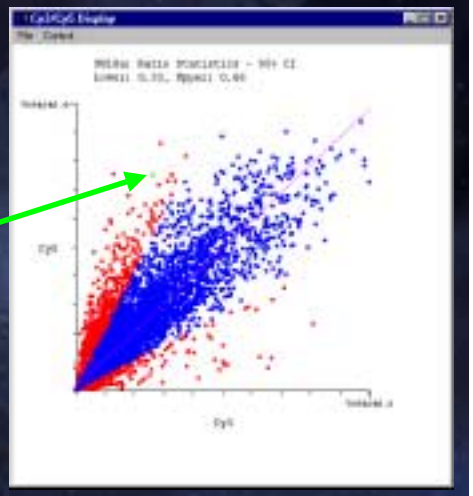
- Presentation
- Multiple Views
- Normalization
- Identification of Differentially Expressed Genes
- Multiple Experiments

TIGR
The Institute for Genomic Research

TIGR MultiExperiment Viewer

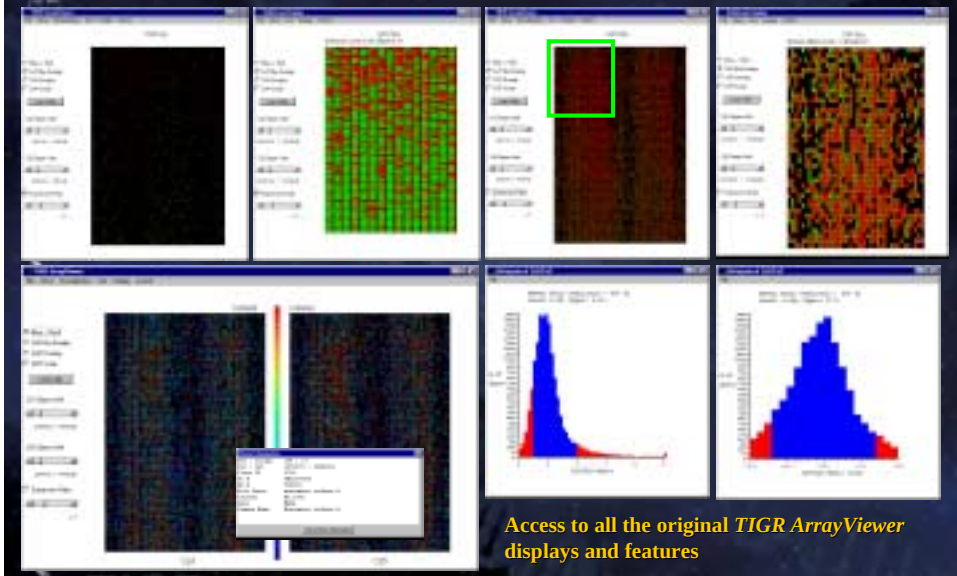


Detail view of a single slide with links to database information about the underlying genes



Highlighted spots in scatter plot track genes being examined in the display on the left

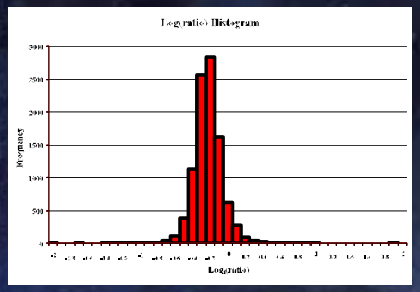
TIGR MultiExperiment Viewer

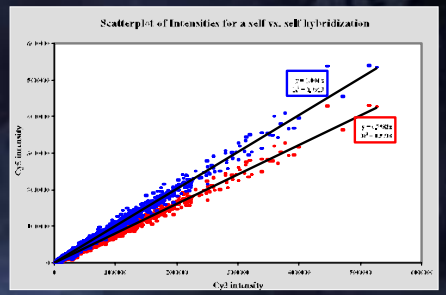


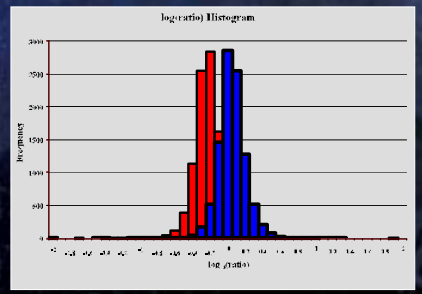
Access to all the original *TIGR ArrayViewer* displays and features

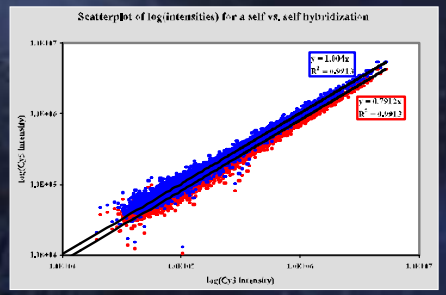
Why Normalize Data?

- Goal is to measure ratios of gene expression levels
  $(\text{ratio})_i = R_i/G_i$
  where R_i/G_i are, respectively , the measured intensities for the *i*th spot.
- In a self-self hybridization, we would expect all ratios to be equal to one:
  $R_i/G_i = 1$ for all *i*. But they may not be.
- Why not?
  - Unequal labeling efficiencies for Cy3/Cy5
  - Noise in the system
  - Differential expression
- Normalization brings (appropriate) ratios back to one.









Multiple Experiments?

- Goal is identify genes (or experiments) which have "similar" patterns of expression
- This is a problem in data mining
- "Clustering Algorithms" are most widely used
- Types
  - Agglomerative: Hierarchical
  - Divisive: *k*-means, SOMs
  - Others: Principal Component Analysis (PCA)
- All depend on how one measures distance

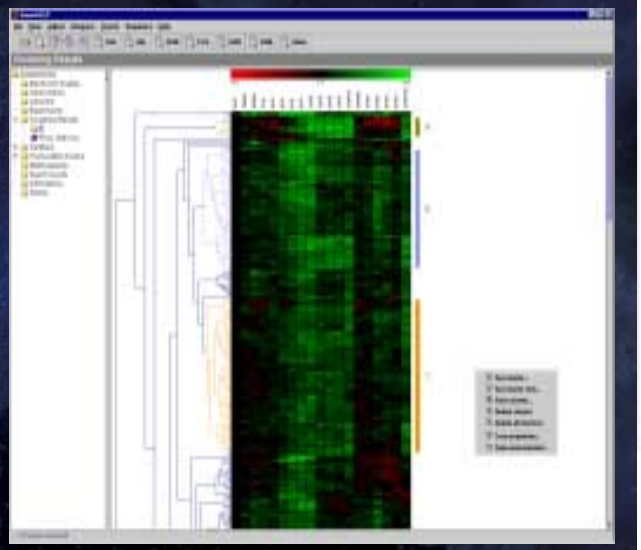
Expression Vectors

- Crucial concept for understanding clustering
- Each gene is represented by a vector where coordinates are its values log(ratio) in each experiment
  - $x = \log(\text{ratio})_{expt1}$
  - $y = \log(\text{ratio})_{expt2}$
  - $z = \log(\text{ratio})_{expt3}$
  - etc.
- For example, if we do six experiments,
  - $\text{Gene}_1 = (-1.2, -0.5, 0, 0.25, 0.75, 1.4)$
  - $\text{Gene}_2 = (0.2, -0.5, 1.2, -0.25, -1.0, 1.5)$
  - $\text{Gene}_3 = (1.2, 0.5, 0, -0.25, -0.75, -1.4)$
  - etc.

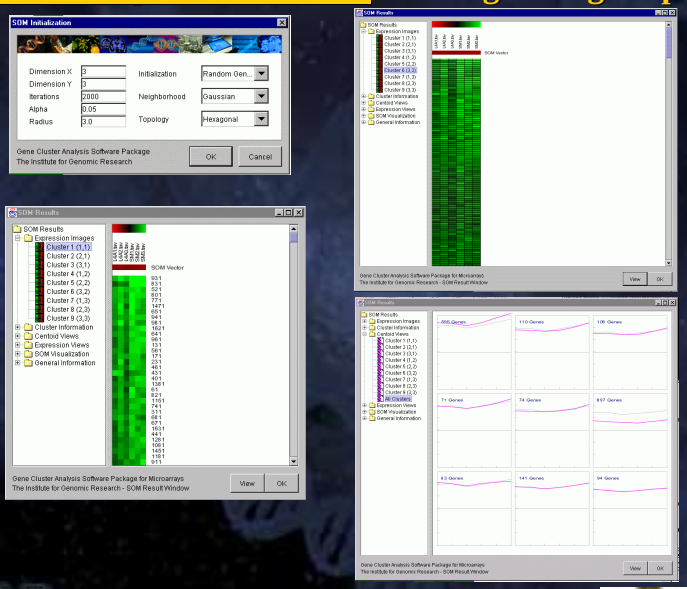
Distance metrics

- Distances are measured "between" expression vectors
- Distance metrics define the way we measure distances
- Many different ways to measure distance:
  - Euclidean distance
  - Pearson correlation coefficient(s)
  - Manhattan distance
  - Mutual information
  - Kendall's Tau
  - etc.
- Each has different properties and can reveal different features of the data

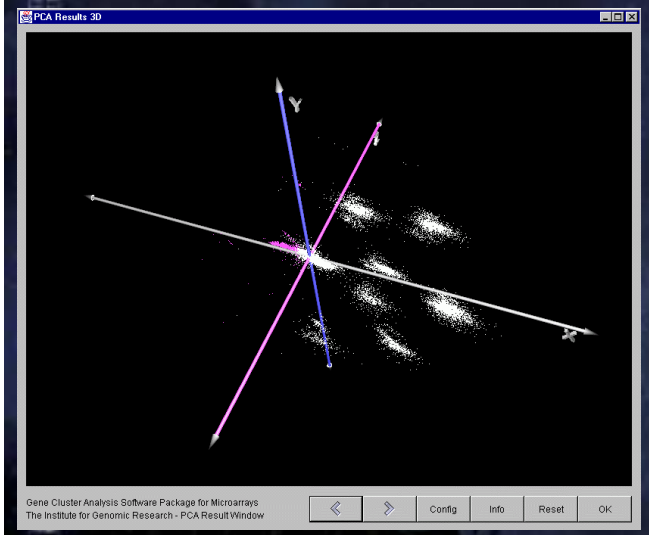
TIGR MultiExpriment Viewer: Hierarchical Clustering

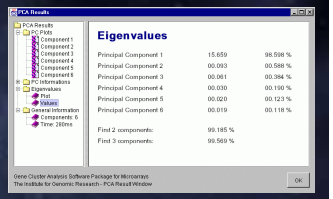


TIGR MultiExpriment Viewer: Self Organizing Maps



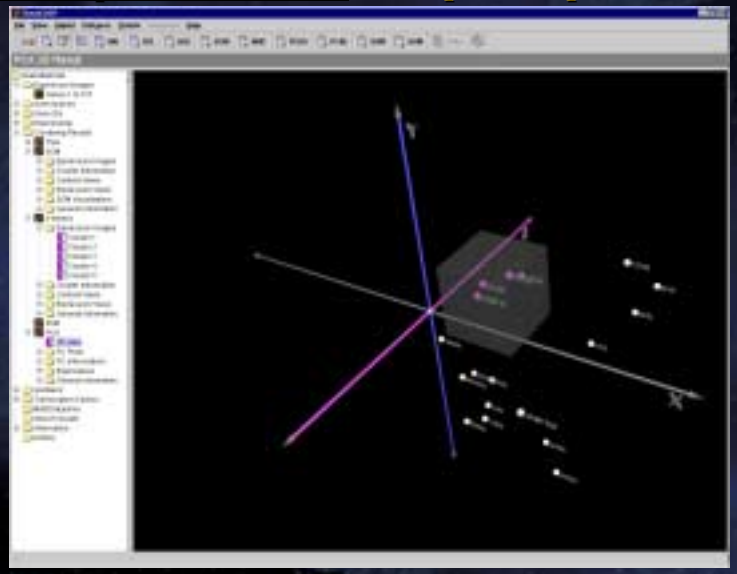
TIGR MultiExperiment Viewer:Principal Component Analysis



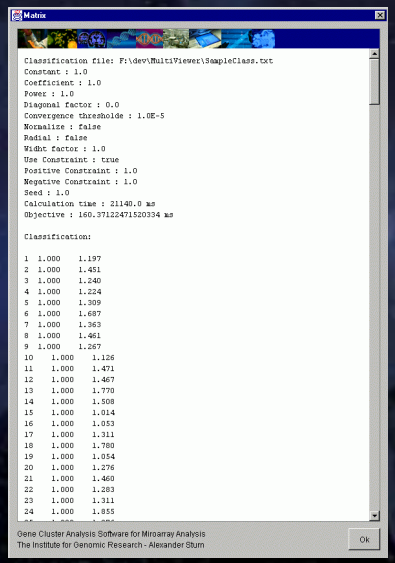


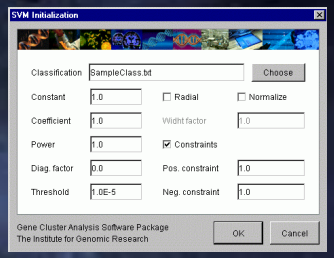
Three dimensional model of clustered data created using Principal Component Analysis

TIGR MultiExperiment Viewer: Principal Component Analysis



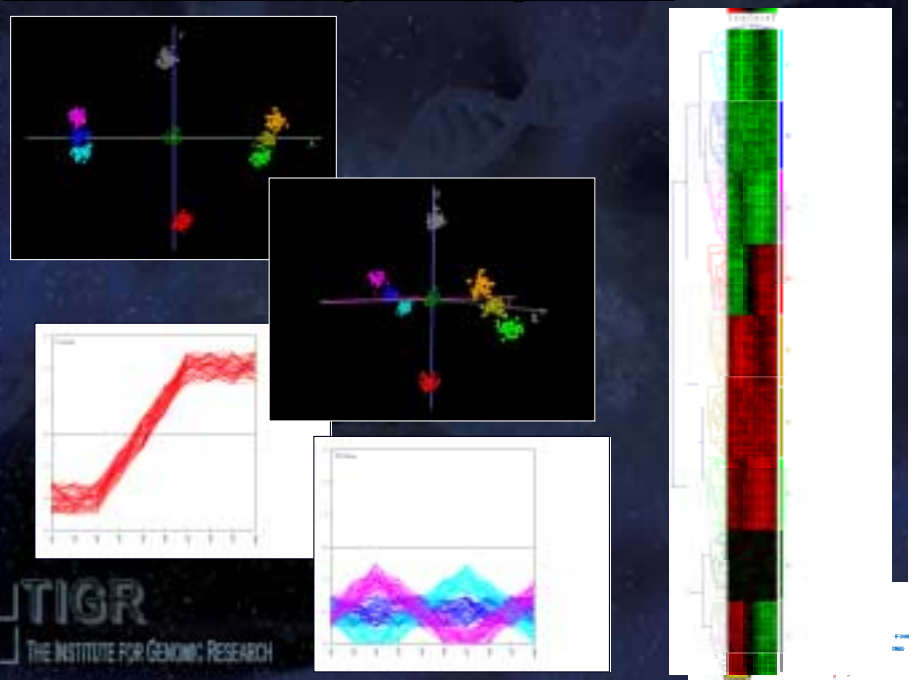
TIGR MultiExperiment Viewer: Support Vector Machines





Classification scheme proposed by a Support Vector Machine

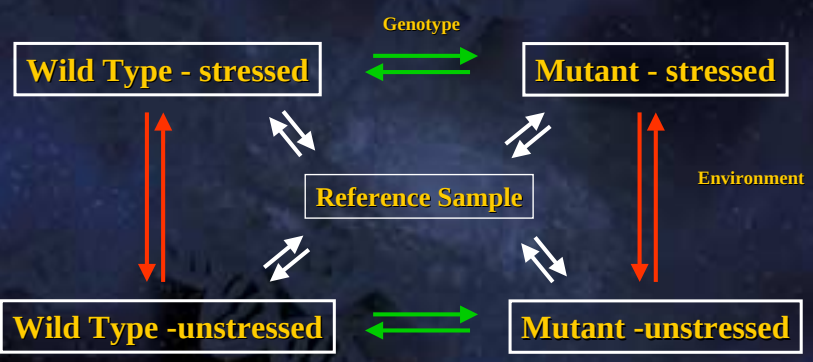
TIGR MeV: Tracking across experiments



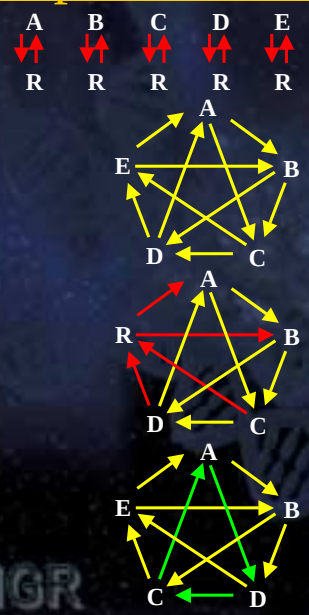
Developmental Goals

- Tools for linking Genes, ESTs, cDNAs, and Sequences
- Microarray Quality Control Protocols and Reagents
- Novel Analysis Techniques and Tools
- Phenotyping Pipelines
- Microarray Resources in Mouse and Rat
- Expression Profiles for Disease Phenotypes of HLBS interest

Basic Experimental Paradigm



Loops and Reference Designs



10 hybs — Standard flip-dye expt

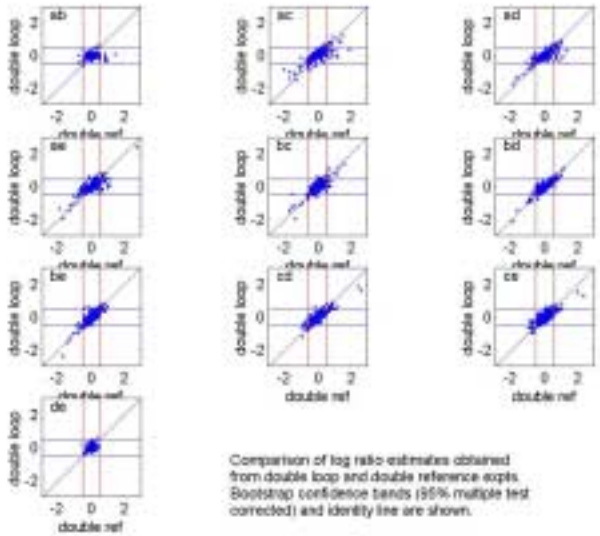
10 hybs — Proposed loop expt

0 new hybs — Proposed loop expt with reference to provide direct comparison with reference

3 new hybs — Proposed loop expt with reference to show invariance w.r.t order

S. Wang , J. Quackenbush, G. Churchill

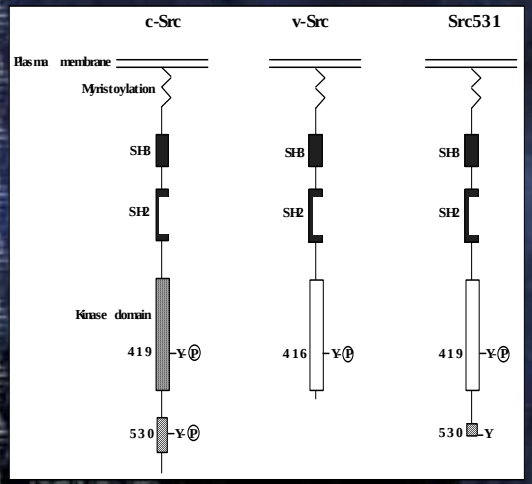
Loops and Reference Designs



Comparison of log ratio estimates obtained from double loop and double reference expts. Bootstrap confidence bands (95% multiple test corrected) and identity line are shown.

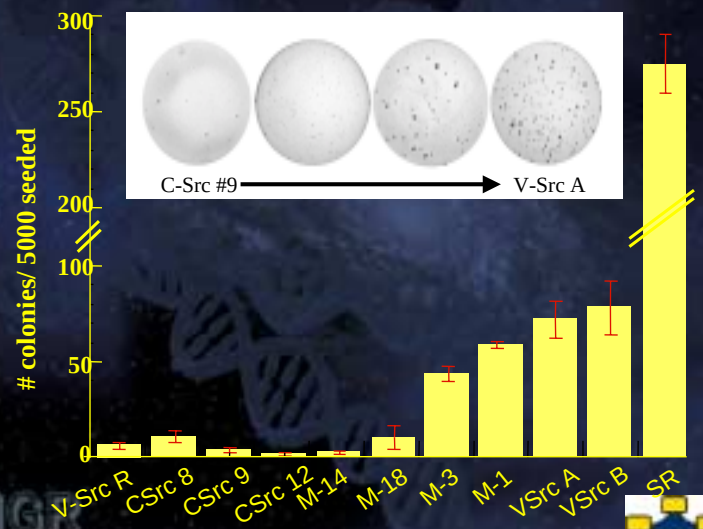
Shuibang Wang , John Quackenbush, Gary Churchill

An Activating Mutation of Src at Codon 531

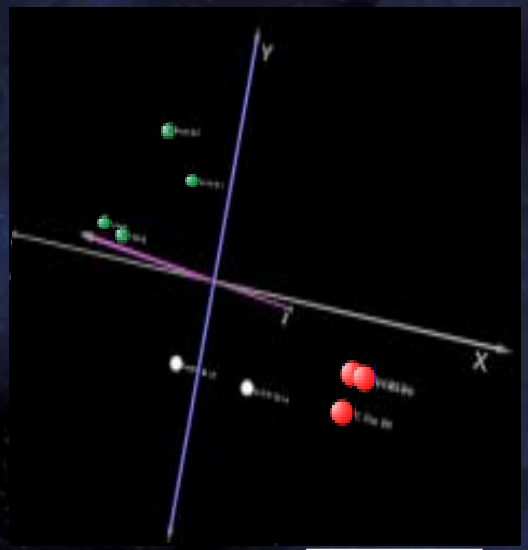


- c-Src
  - Non-receptor tyrosine kinase
  - Activated by growth factor receptors
  - Wide-spread tissue distribution
  - Cell growth & differentiation
  - Implicated in human colon cancer
- V-Src
  - Unregulated kinase activity
- Src531
  - Elevated kinase activity

Soft Agar Assay

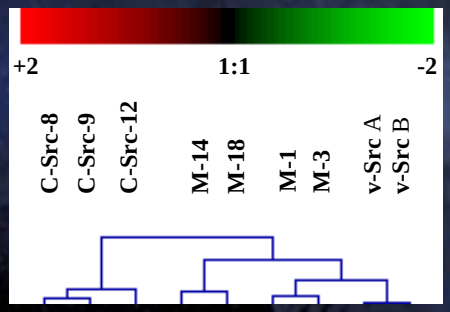


Principle Component Analysis of Cell Lines



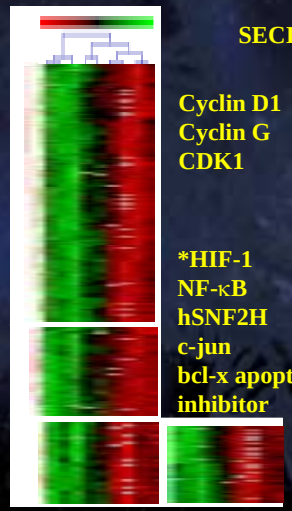
NH Lee, RL Malek

Hierarchical Clustering



NH Lee, RL Malek

Transformation Fingerprint



SECIS binding protein 2

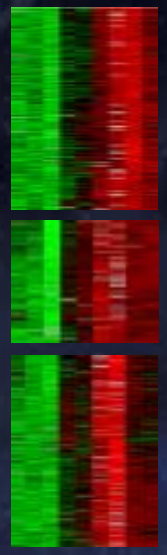
Cyclin D1
Cyclin G
CDK1

stanniocalcin
RYB-a
arylsulfatase A
cathepsin B
IGF-I and IGF-II

*HIF-1
NF-κB
hSNF2H
c-jun
bcl-x apoptosis inhibitor

FA synthase
stearyl-CoA desaturase
epoxide hydrolase
propionyl-CoA carboxylase
peroxisomal enoyl-CoA hydratase

MEK kinase
JAK3
A-raf
N-ras
ERK5
p190-B
ragB
Rho8
Ran/TC4
β-PAK
GDNFR alpha



NH Lee, RL Malek

Visiting Scientist Program

Goal: To provide access to the reagents, techniques, and analysis tools developed through this PGA

- Solicit applications from external investigators
- PGA steering committee will review applications and select participants
- Seek input from NHLBI staff and PGA-CC
- PGA Investigators will work with Applicants to refine experimental design
- Selected candidates will be invited to spend one (or more) weeks at TIGR generating array data

Acknowledgments

The TIGR Gene Index Team
Jennifer Cho
Svetlana Karamycheva
Yudan Lee
Babak Parvizi
Geo Pertea
Razvan Sultana
Jennifer Tsai
John Quackenbush
Joseph White

Emeritus
Ingeborg Holt (TGI)
Feng Liang (TGI)
Kristie Abernathy (μA)
Sonia Dharap(μA)
Julie Earle-Hughes (μA)
Cheryl Gay (μA)
Priti Hegde (μA)
Rong Qi (μA)

H. Lee Moffitt Center/USF
Timothy J. Yeatman

TIGR PGA Collaborators
Norman Lee
Renae Malek
Hong-Ying Wang
Truong Luu
Nnenna U. Nwokekeh

PGA Collaborators
Gary Churchill (TJL)
Greg Evans (NHLBI)
Harry Gavaras (BU)
Howard Jacob (MCW)
Anne Kwitek-Black (MCW)
Allan Pack (Penn)
Beverly Paigen (TJL)
Luanne Peters (TJL)
David Schwartz (Duke)

<johnq@tigr.org>

TIGR Human/Mouse/Arabidopsis Expression Team
Emily Chen
Renee Gaspard
Jeremy Hasseman
Heenam Kim
John Quackenbush
Erik Snesrud
Shiubang Wang
Ivana Yang
Yan Yu
Baoping Zhao

Array Software Hit Team
Jerry Li
John Quackenbush
Alex Saeed
Vasily Sharov
Alexander Sturn
Joseph White

Assistant
Mary Mulholland

Funding provided by the Department of Energy and the National Science Foundation

Funding provided by the National Cancer Institute, the National Heart, Lung, Blood Institute, and the National Science Foundation

TIGR Faculty, IT Group, and Staff