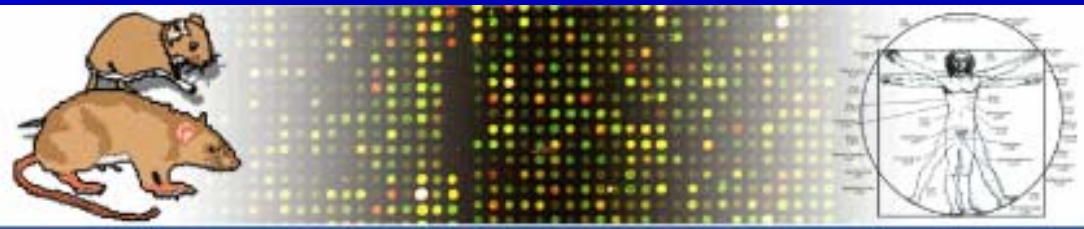
Expression to Biology Extracting Meaning from Microarrays

Montreal Microarray Symposium

18 March 2004



Microarray Expression Profiling of Rodent Models of Human Disease

The Institute for Genomic Research | Duke University | Boston University | The Jackson Laboratory | Medical College of Wisconsin | University of Pennsylvania

Acknowledgments

<johnq@tigr.org>

The TIGR Gene Index Team

Foo Cheung
Svetlana Karamycheva
Yudan Lee
Babak Parvizi
Geo Pertea
John Quackenbush
Razvan Sultana
Jennifer Tsai
Joseph White

Emeritus

Jennifer Cho (TGI)
Emily Chen (μA)
Ingeborg Holt (TGI)
Feng Liang (TGI)
Kristie Abernathy (μA)
Sonia Dharap (μA)
Julie Earle-Hughes (μA)
Cheryl Gay (μA)
Jeremy Hasseman (μA)
Priti Hegde (μA)
Heenam Kim (μA)
Lara Linford (μA)
Rong Qi (μA)
Erik Snesrud (μA)
Shuibang Want (μA)
Ivana Yang (μA)
Yan Yu (μA)

H. Lee Moffitt Center/USF

Timothy J. Yeatman
Greg Bloom

PGA Collaborators

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

TIGR PGA Collaborators

Norman Lee
Renae Malek
Hong-Ying Wang
Truong Luu
Bobby Behbahani

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

TIGR Human/Mouse/Arabidopsis Expression Team

Adriana Ahumada
Tove Andersson
Joanne Emerson
Bryan Frank
Molly Freeman
Renee Gaspard
Nadeeza Ishmael
Ka Yin (Simon) Kwong
Jennie Larkin
Fenglong Liu
John Quackenbush
Yonghong Wang
Yan Yu

Array Software Hit Team

Nirmal Bhagabati
John Braisted
Tracey Currier
Jerry Li
Wei Liang
John Quackenbush
Alexander I. Saeed
Vasily Sharov
Mathangi Thiagarajan
Joseph White

Assistant

Aseye Aboagye

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 Henri Poincare

Levels of Biological Information

'omics

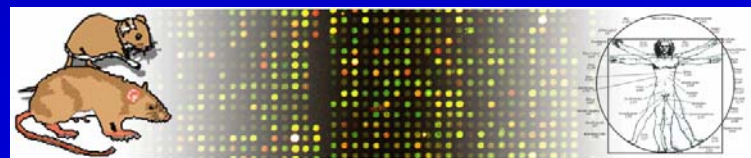
DNA
mRNA
Proteins
Informational
Informational
Cells
Organs
Individuals
Populations
Ecologies

Genomics
Functional Genomics
Proteomics
Metabolomics
ms Biology
llular Biology
Medicine
Medicine
Genetics
Ecology

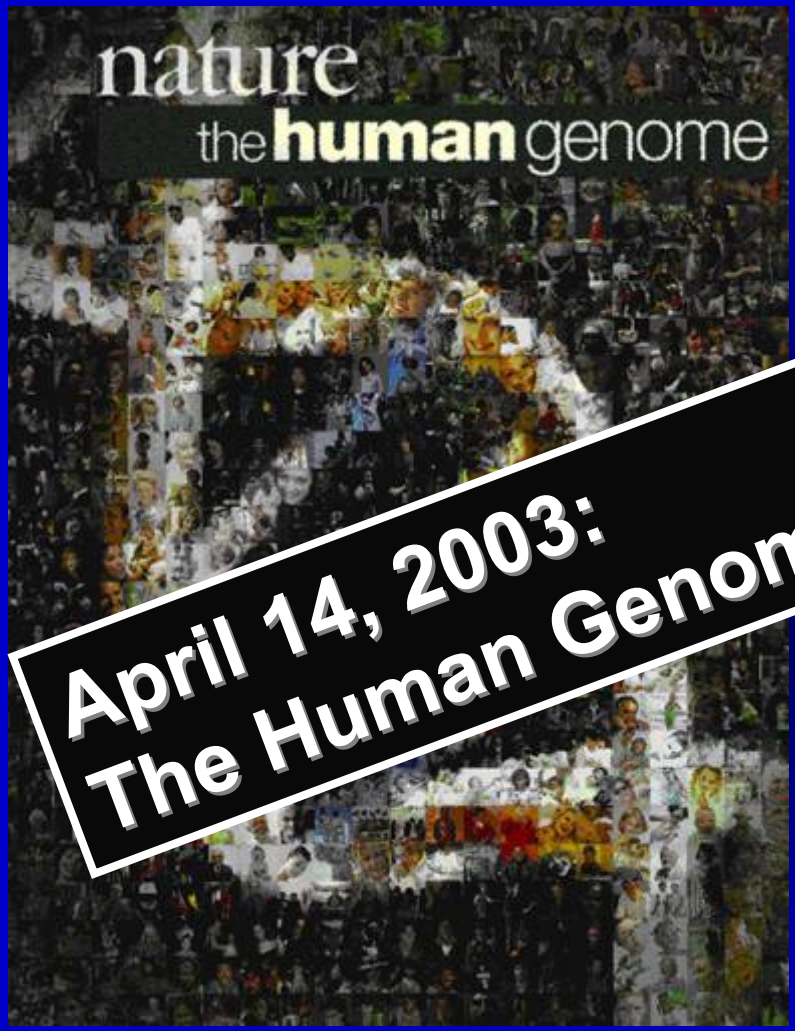
The Future!

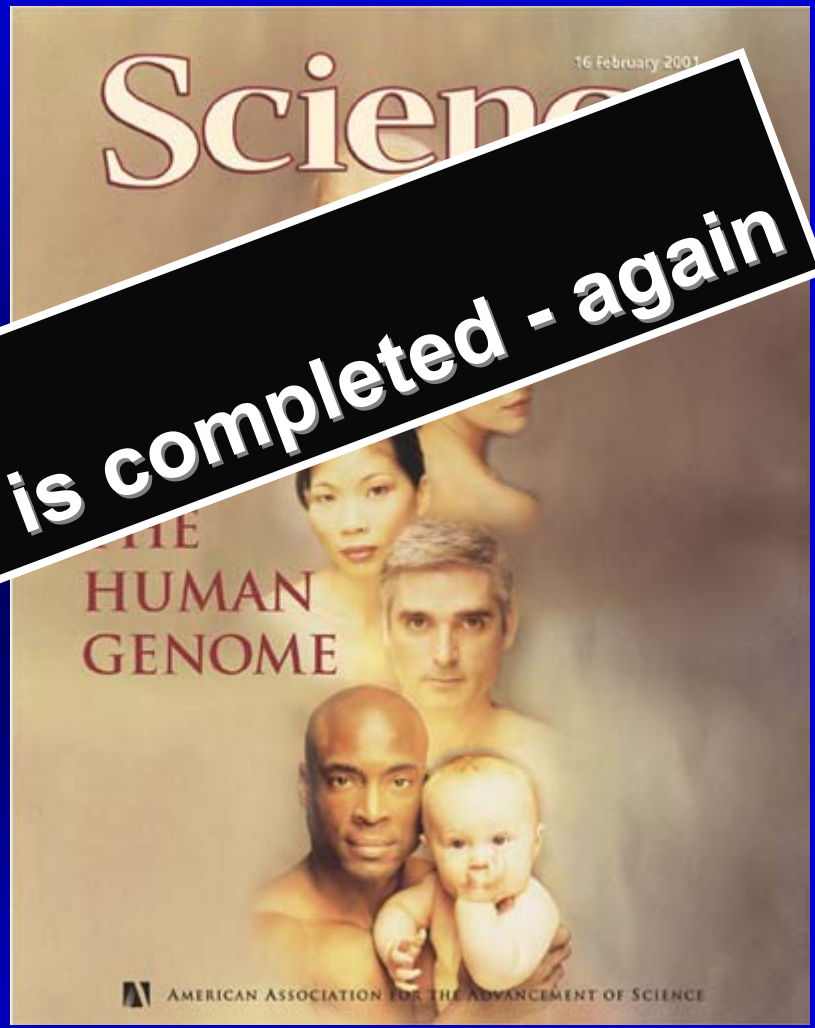
Traditional Biology

National Heart, Lung, and Blood Institute
PROGRAMS FOR GENOMIC APPLICATIONS



February 2001: Completion of the Draft Human Genome





April 14, 2003:
The Human Genome is completed - again

Public HGP

Celera Genomics

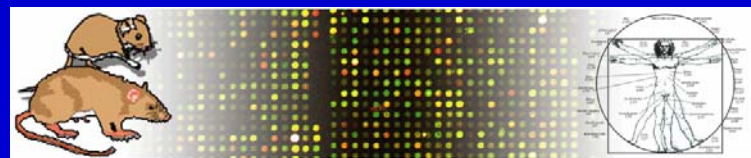
But what does *finished* mean???

National Heart, Lung, and Blood Institute
PROGRAMS FOR GENOMIC APPLICATIONS

Where are the pressing questions?

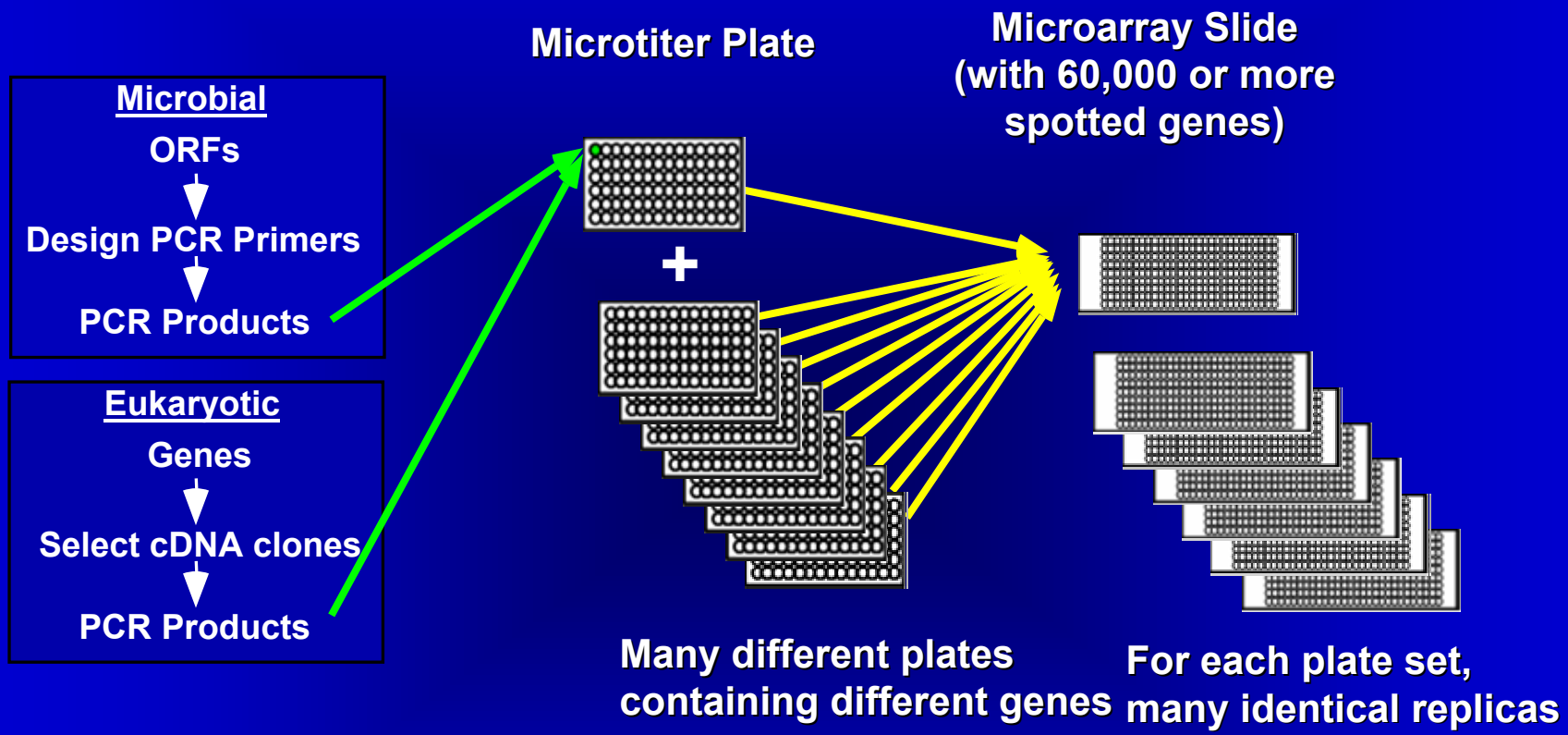
- Can we find the genes and assign them functions?
- Can we predict protein structures and functions?
- Can we reconstruct metabolic, signaling, and other pathways?
- Can we reconstruct informational networks?
- Can we link genotype to phenotype?
- Can we use genotype/phenotype to predict clinically-relevant outcome?
- Can we use cross-species comparisons to learn something?

National Heart, Lung, and Blood Institute
PROGRAMS FOR GENOMIC APPLICATIONS



Microarray Analysis

Microarray Overview I



The Beast: Microarray Robot from Intelligent Automation

<http://www.ias.com>

Microarray Overview II

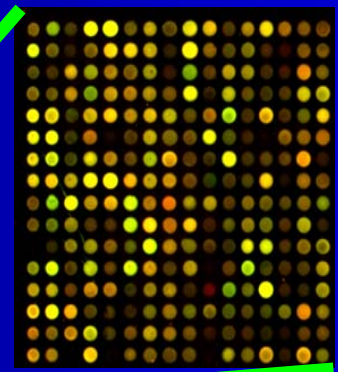
Control

Test

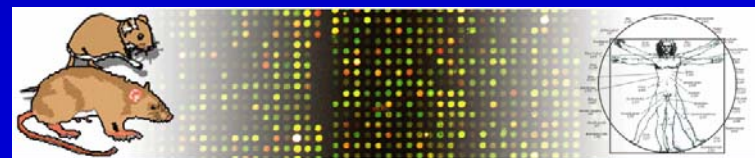
Prepare Fluorescently Labeled Probes

Hybridize, Wash

Measure Fluorescence in 2 channels red/green

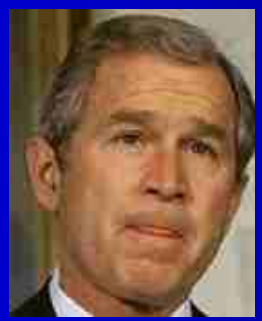


Analyze the data to identify patterns of gene expression



Microarray Overview II

Weed



Bush

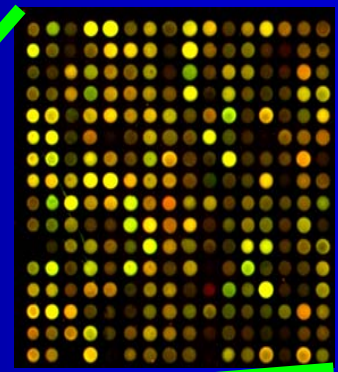
Control

Test

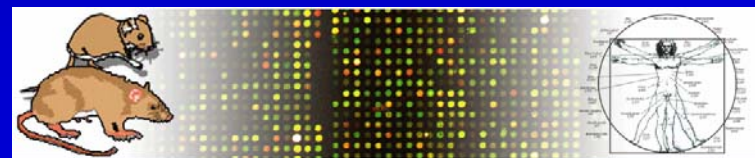
Prepare Fluorescently Labeled Probes

Hybridize, Wash

Measure Fluorescence in 2 channels red/green

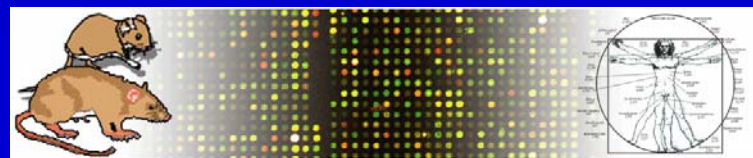


Analyze the data to identify patterns of gene expression



General Microarray Strategy

- Choose an experimentally interesting and tractable model system
- Design an experiment with comparisons between related variants
- Include sufficient biological replication to make good estimates
- Hybridize and collect data
- Normalize and filter
- Mine data for biological patterns of expression
- Integrate expression data with other ancillary data such, including genotype, phenotype, the genome, and its annotation



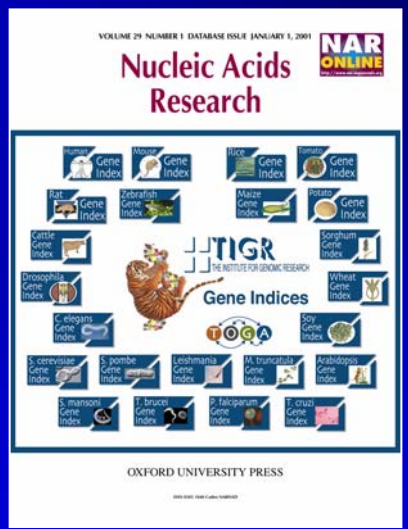
Annotating and Comparing Arrays

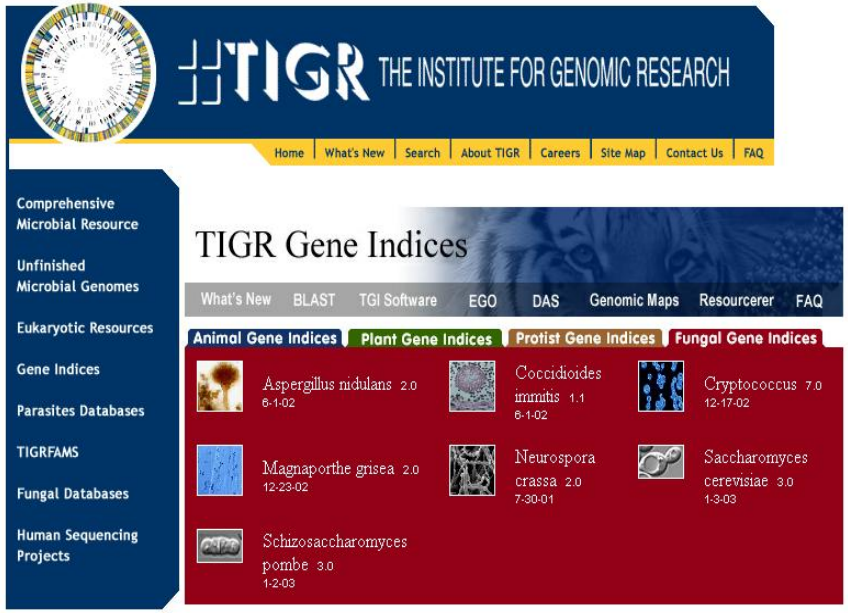
TIGR Gene Indices home page

www.tigr.org/tdb/tgi

~70 species

>19,000,000 sequences





The TIGR Gene Index Project is supported in part by funding from the US Department of Energy, Grant #DE-FG02-99ER62852, and the US National Science Foundation, Grant #DBI-9983070. Additional funds are provided by the US National Science Foundation through grants #DBI-9813392 and #DBI-9975866.

The TIGR Gene Indices are built using:
- **megablast** [Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller (2000), "A greedy algorithm for aligning DNA sequences", J Comput Biol 2000; 7(1-2):203-14]
- **CAP3**, developed by Dr. Xiaoqiu Huang
- **Paracel TranscriptAssembler**[tm], from Paracel Inc.
- **DNA-Protein Search program (dps)** developed by Dr. Xiaoqiu Huang

The ORF annotation of TCs is done using:
- **ESTScan** [Iseli, C. Jongeneel, C.V., and Bucher, P. (1999) ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. ISMB 7: 138-148.]
- **DIANA-EST** Hatzigeorgiou AG, Fiziev P, Reczko M. Related Articles DIANA-EST: a statistical analysis. Bioinformatics. 2001 Oct;17(10):913-9.
- **framefinder** Expressed Sequence Tag Analysis Tools Etc (c) Guy St.C. Slater 1996-1999. Human Genome Mapping Project RC, Hinxton, Cambridge, UK.

The Genome mappings are done using:
- **blest** (C) Webb Miller - see reference
- **Scout**[tm], from Paracel Inc.
- **gap2** developed by Dr. Xiaoqiu Huang

The expression profiles of the ESTs are scored using:
- **'R' statistic** (Stekel, Git, and Falciani (2000) The comparison of gene expression from multiple cDNA libraries. Genome Research 10:2055-2061).

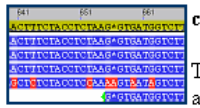
National Heart, Lung, and Blood Institute PROGRAMS FOR GENOMIC APPLICATIONS

TGICL Tools are available – with more coming

Geo Pertea
Razvan Sultana
Valentin Antonescu

TGI Clustering tools (TGICL): a software system for fast clustering of large EST datasets

This package automates clustering and assembly of a large EST/mRNA dataset. The clustering is performed by a slightly modified version of NCBI's *megablast* , and the resulting clusters are then assembled using CAP3 assembly program. TGICL starts with a large multi-FASTA file (and an optional peer quality values file) and outputs the assembly files as produced by CAP3. Both clustering and assembly phases can be parallelized by distributing the searches and the assembly jobs across multiple CPUs, as TGICL can take advantage of either SMP machines or PVM (Parallel Virtual Machine) clusters. Here is a link to the README file which comes with the package.

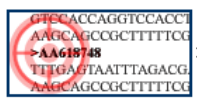


clview : an assembly file viewer.

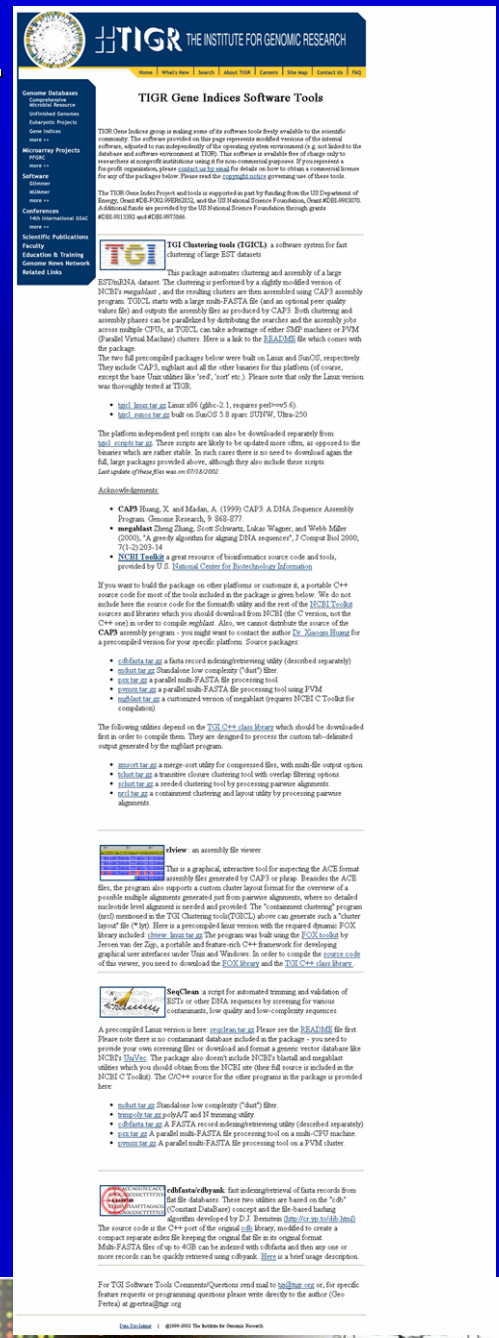
This is a graphical, interactive tool for inspecting the ACE format assembly files generated by CAP3 or phrap. Beasides the ACE files, the program also supports a custom cluster layout format for the overview of a possible multiple alignments generated just from pairwise alignments, where no detailed nucleotide level alignment is needed and provided. The "containment clustering" program (nrcl) mentioned in the TGI Clustering tools(TGICL) above can generate such a "cluster layout" file (*.lyt). Here is a precompiled linux version with the required dynamic FOX library included: clview_linux.tar.gz The program was built using the FOX toolkit by Jeroen van der Zijp, a portable and feature-rich C++ framework for developing graphical user interfaces under Unix and Windows. In order to compile the source code of this viewer, you need to download the FOX library and the TGI C++ class library .

SeqClean :a script for automated trimming and validation of ESTs or other DNA sequences by screening for various contaminants, low quality and low-complexity sequences.

A precompiled Linux version is here: seqclean.tar.gz Please see the README file first. Please note there is no contaminant database included in the package - you need to provide your own screening files or download and format a generic vector database like NCBI's UniVec. The package also doesn't include NCBI's blastall and megablast utilities which you should obtain from the NCBI site (their full source is included in the NCBI C Toolkit). The C/C++ source for the other programs in the package is provided here:



cdbfasta/cdbyank: fast indexing/retrieval of fasta records from flat file databases. These two utilities are based on the "cdb" (Constant DataBase) concept and the file-based hashing algorithm developed by D.J. Bernstein (http://cr.yp.to/djb.html) The source code is the C++ port of the original cdb library, modified to create a compact separate index file keeping the original flat file in its original format. Multi-FASTA files of up to 4GB can be indexed with cdbfasta and then any one or more records can be quickly retrieved using cdbyank. Here is a brief usage description.



National Heart, Lung, and Blood Institute — PROGRAMS FOR GENOMIC APPLICATIONS

☺ Available with source

Gene Index Assembly process

ESTs from GenBank (dbEST)

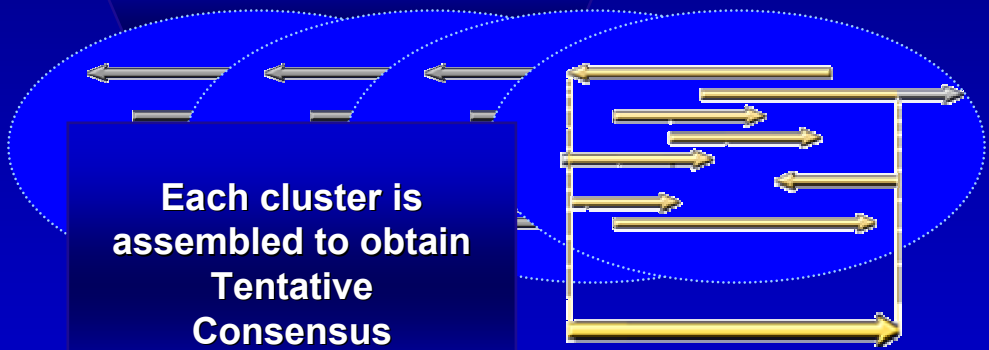
TIGR ESTs

Expressed Transcripts (ET) from GenBank CDS

remove vector, poly-A, adapter,mitochondrial and ribosomal sequence

reduce redundancy

High stringency pair-wise comparisons to build Clusters



Each cluster is assembled to obtain Tentative Consensus sequences (TCs)

Annotate TCs and release

National Heart, Lung, and Blood Institute PROGRAMS FOR GENOMIC APPLICATIONS

A TC Example

>**TC161360** TC25195 TC29362 TC33731 TC40754 TC149101

TGAGGCTCACAAGAACTTTTATTCTTTTTAAATAGACACTAAAATTATCTCCTAGTCATGAGAAATTGGTAAAGACTAAT
TATTTGAGAATCTGACGATGACTAATGTAAAATCATTAAGGAAATGAATTTCAGAGAGGGGAAACTTTTCAAATTGAATA
CTGCATTTAAAACTTTTCAGCTTGACACTCCTCCTCCCACCTCCCCATCCTCTCCAGGCATAGCGGTATCTTCTTTAGCT
TAGGGTACCTTCTATGGAGAAGAATGGATATGGAGAATCGTGCTGTGGCTTGTAAGTGGGCAGAACTTAGTAAAGACCTA
CTGGATGAGTAAACTCCTTGGGAGCATGTGTCAGATAGGTAGGAATAGCTCAATATGACTGGATGTGCCACTATTCAAAC
ACAGGTTAGTATTATGTGGCAGAAGCATCCCATTGTTTGATGTTATGGAGAAGAAAGAAACATCCAAGGTGGAGTATCCA
TTGCAGGCACTGACACAAAGTTTTATTTACTTAGAGCTTGTTTTGAAGACCACACAGGGAAAAGGTGCTACTTCCAGTTT
CTTTGTAATAACAGGAAATAACTCCCACCGGTAGCCTCTAATAAAATAGAAATATTCCAAGGAGTGAAACTTAAGCTGTT
CATATACCCATAATGCCTAGAAGCAGACTTGTCATGGTATCTGAATGATAGGCTATGGTGAGATCTTTTTAGGCTAACAG
TGTTTCTAGGTCAGGTGCTAGCATCCCTGCTCAGGAACAGGGGTGGGAAAGTATGGTGGCCTGAGATTTAGGATTTTAAC
TGTGTTGTTTTTAAGCATGATCTTTGTGTGGTAATTTATAGTGCATATAAGATGTGTTTGTGGTGCATCTTATAACTTTC
CAGCTAATTGCATATTAATGTCACGACTAGTTTTCCAAATGATGTAAGATTCTGGGTGTCTTTATTCATATGGTGTCAAA
TCCAATTCTGACTCTTTTGTGATTCAGCACATTGCTCACAAAGTATATACTTTGATATAGCTTATACAGGCATATGGGCA
TAGATAATTTGGTTATTTCAAACATTTCTAGAATTTGAAGAGCTGGGTTTGACTCAAGTCCCCACAACTATTATTGAATA
TCTTGGCTTGCTTCTTACATTCAACTAAATTACAAAATAGTGATAGGTTTTTTGTTTTTTGTTTTTTTGTTTTTTTTCTG
TGAATCATGAAAGACCAAATAAACTACTCTGCCATTTGGAAAGAATCTCAAAGCACAGCTGTTGCTTAGGATTTGAGATC
TGGGAACCCATTACATTTCTGTCTCACCATTCTTTTTTCTCTTGACTAAAGAATACAAATACATAAACGATGTGACCAGA
AGCCAAGAAACTGGAGATGGGCAAAGTTAAGACATGAAACTTGCTTATGTGTAAGCTATGCTTTTTGTACAGAGACAGAC
TTAGGACTAGGCCTTTCACAGCTTCCCAAGTGGGACACAGAAGCTTGAAGAACATGCCATATTTTGTGCCTTCCCCACAG
GCATATAGGTGTCTCATTTTCTCTCATTATTAAAACAATTCTCATTAAACAATTTCTCTCATATTCTTTAAGACAATCAT
GCAAAAGGGACTCCACCAAACATGAAGAAGGAGTCAGAGAGTTCAGAGATGAACAGAACAGTCATGTGTCTTTTCCCTCT
GCTGTAACTGCTGATGGAGGAAATGCGGAAAAACCATTACCACTACTGAGACACACATCCAAGGAAAAGGTGAATGAGAA
TTGGATTCTTCAGCGTTTCTGCCAGTGTAAGACTCCCTAAGCAAGGCAAAAGACAATCTTTTATAATCACTGCATTCTTC
AAATGTGAAGGAGAGCTAGATGTGGCTTCTCTATGCAAAGTTAAGCTGTGAAAATATGGAGAATAGATTGTGGAAGGCCA
CAAGAGATGAGGGTAACTATGTGTCTTGGAAGCTCTCTAAACAATAATGCTTAGAAAAGAAAAGAGTTTAAGGAGCGATC
TCATTTTCCACTGGCAAGGATTCCTGCTCTTGGGGAGGTGCTGGAAAATTTATTTCTGCTTCTTCTTCTTCTTCCTCCTG
CACTGGAATAATTTCAATTTCCTCTTCATTCTTTGGTTGGGAAGATACTCCACTTAGCTCAATGATTTCTTCTTTCATTT
TAATCGTCTGCTCATTTTTTTCTCCAGCTGACAGCAGAACCACATTAGATTTGGATCTGGTACACATTCTTTTCCACTCA
TTCTCCACAATACCAGCTGTCACAAGTTTCTGGAAGAAGGCAGAGATCAGCATCGCCGACAGAATGCCCAAGAACACAGA
CTGAATGCTGTTACAGTACTGTGTAGATGGGGAGTTTTTCTCTGAGGAATTAGATGGTTCACAGTCGTAGATATCAACAT
ACGGCTTGGAAGTTTGAATAAGCTCCAGTCTTCTCATTTTTAAAAAATGAGAAAGTGTCATGTTAAGTATGTCCATGATT
GAAAGAATTATTCCAGAAATGGCAGCAAAGAGGCTTAGAGAGCTCATTATCACTTTTGCTTTGACCAAACTCTTCCTGGA
GGTTTTTTCTGCTGCAGCTGCCAGGAGTGATCCTGAAATAATGTACATAATGCCTCCCCAGAGAGGGTACCATACACTCA
AACAGATGGGTGCGAAGACCCCTGTGGGGATCATCAGCAGTCCCCCCAGGGTAATATGGAAGAGGCCATTCATGATTTGG
ACAGCCCCCAAAGCCTTTGATTCCCTCATGAAGAAGCTTTGTGTGGGGCCCACCAGTGAAGATGTCCTTTTGAGGTTCAC
TTTTGGAGCAGGTTGCATGGCGAGGGGACCTTTTGTAGGCTCTGCTGGGAAAGGTCCACTCATTGAGTTTTCAAACGCCT
GGGTCTCCAAGGCTCCAAGGCTCCAAGGCTTAAGGGCCAAATTAGCCACTCACAGCTGAGATCAGTTGGCCTCTTCTATC
TCAAGTACTTGAAGATAGCTACAGTGTACTTACATATAATAATAAATAAATCTTTAAAAAAAAAAAAAAAAAAAGGAATTC

-cell differentiation antigen

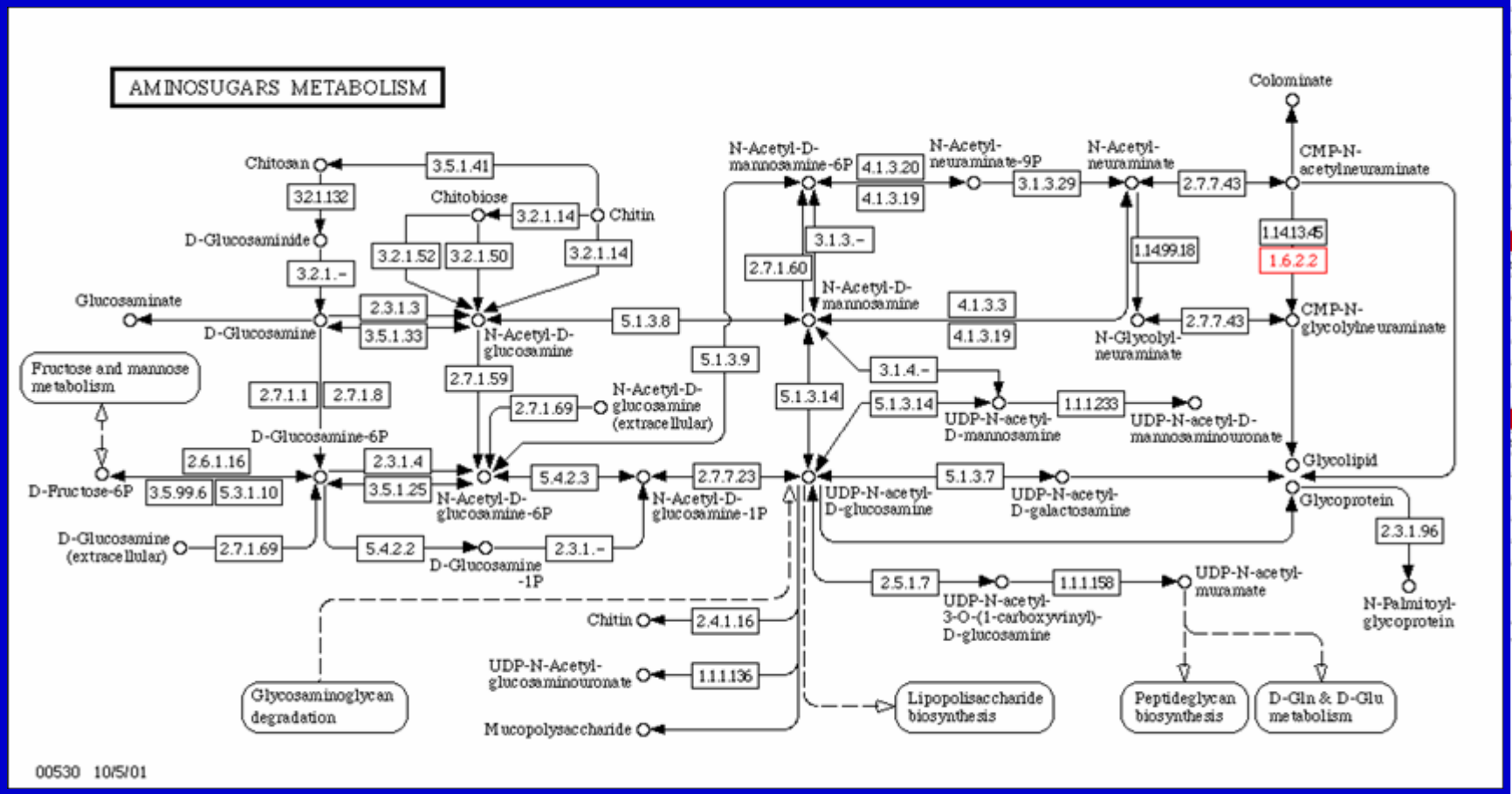
(B-LYMPHOCYTE

(Homo
1][X07203 CD20 receptor

Iouse Genome Database

GO Terms and EC Numbers

Position of term GO:0004128



(I) enzyme (GO:0003824) +
(I) oxidoreductase (GO:0016491) +
(I) oxidoreductase\, acting on NADH or NADPH (GO:0016651) +
(I) oxidoreductase\, acting on NADH or NADPH\, NAD or NADP as acceptor (GO:0016652) +
(I) cytochrome b5 reductase (GO:0004128) o

National Heart, Lung, and Blood Institute PROGRAMS FOR GENOMIC APPLICATIONS

Babak Parvizi

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

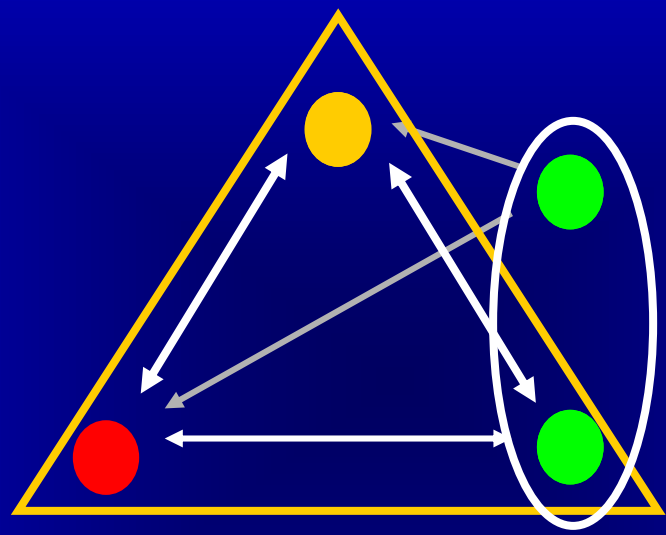
Eukaryotic Gene Orthologs

The **Eukaryotic Gene Orthologs (EGO)**, is a database for orthologous genes in eukaryotes. EGO is generated by pair-wise comparison between the Tentative Consensus (TC) sequences that comprise the TIGR Gene Indices from individual organisms. The reciprocal pairs of the best match were clustered into individual groups and multiple sequence alignments were displayed for each group. The EGO database can be accessed through the **SEARCH** function. The release notes for the current EGO can also be referenced.

National Heart, Lung, and Blood Institute PROGRAMS FOR GENOMIC APPLICATIONS

Dan Lee, Ingeborg Holt

Building TOGs: Reflexive, Transitive Closure



And Paralogues

Tentative Orthologues

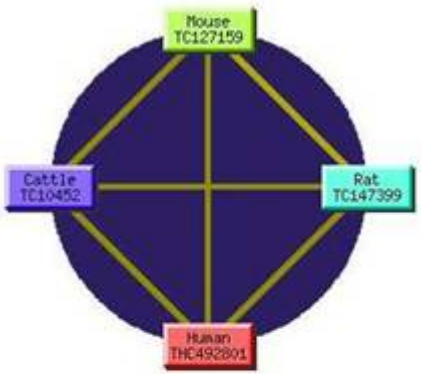
National Heart, Lung, and Blood Institute PROGRAMS FOR GENOMIC APPLICATIONS

Thanks to Woytek Makałowski and Mark Boguski

TOGA: An Sample Alignment: bithoraxoid-like protein

Tentative Ortholog 3220

Cattle|TC10452 Rat|TC147399 Mouse|TC127159 Human|THC492801



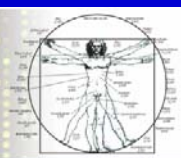
Sequence1	Sequence2	PID	Match length
Rat|TC147399	Cattle|TC10452	89.49	408
Mouse|TC127159	Cattle|TC10452	88.73	407
Human|THC492801	Cattle|TC10452	89.93	406
Mouse|TC127159	Rat|TC147399	92.93	646
Human|THC492801	Rat|TC147399	89.63	375
Human|THC492801	Mouse|TC127159	89.43	387

CLUSTAL W (1.8) multiple sequence alignment

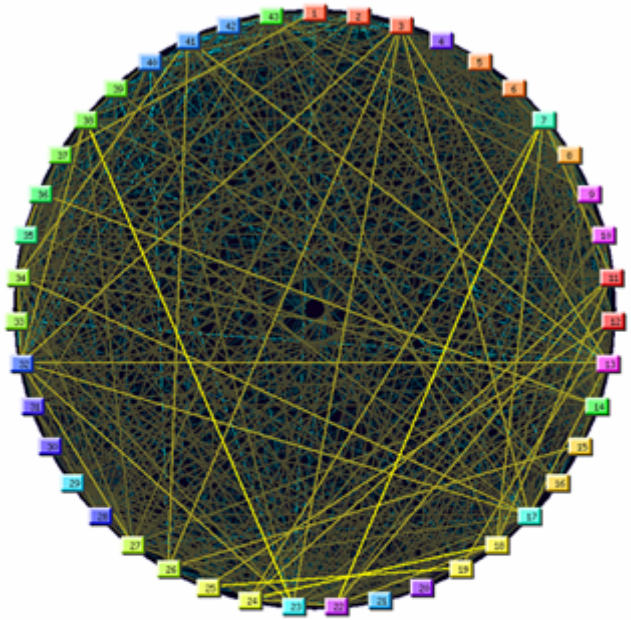
```
mouse|TC127159      TGGTCTACACAGGCTC-AG-GTGGCCACCACGTGC----CCACTGACATGATTAGCACTA
human|THC492801     TGGTGTGAGTGGGTTCCAA-GCGACTGCCATGTGCTAGTCCACTGACATGATTGACATTA
cattle|TC10452      CAGCCTGGGAGGGCTCCAACGTGCCTTCCACGTGCCCGTCAATGGACATGATTAACGCTA
rat|TC147399        AAGGTCTGCATGGCTCCAGGCAGCC--ACATGTGCCC-----ACTGACATGATTAACGCTA
                      *        ** ** *    * *   ** ****      *  *********  *  **

mouse|TC127159      TTATTCTTGGGGGACATTAAATTAAGGAATGACACAGGAAGCCCAGACAGTGGCTTATTC
human|THC492801     ACATTCTTGGGGGGCATTAAATTAAGGAATGACACAGGGAGCCAAGAGAGTGGCTTATTC
cattle|TC10452      TTATTCTTGAGGGGCATTAAATTAAGGAATGACGCAGGGAGCCAAGAAAGCAGCTTATTC
rat|TC147399        TTATTCTCGGGGGACATTAAATTAAGGAATGACACAGGAAGCCAAGACAGTGCCTTATTC
                      ***** * *** ********************* **** **** *** **   *******

mouse|TC127159      AGTTGGATTCTGGATCACAATCAGGAAATAGTCTTTATCTGGTGCCACCATAATTTCATT
human|THC492801     GGTTGGATTCTGAATCACAATCAGGAAATAGTCTTTATCTGGTGCAACCATAATTTCATT
cattle|TC10452      AGTTGGATTCTGAATCACAATCAGGAAATAGTCTTTATCTGGTGCAACCATAATTTCATT
rat|TC147399        AGTTGGATTCTGAATCACAATCAGGAAATAGTCTTTATCTGGTGCCACCATAATTTCATT
                    *********** ********************************** **************
```

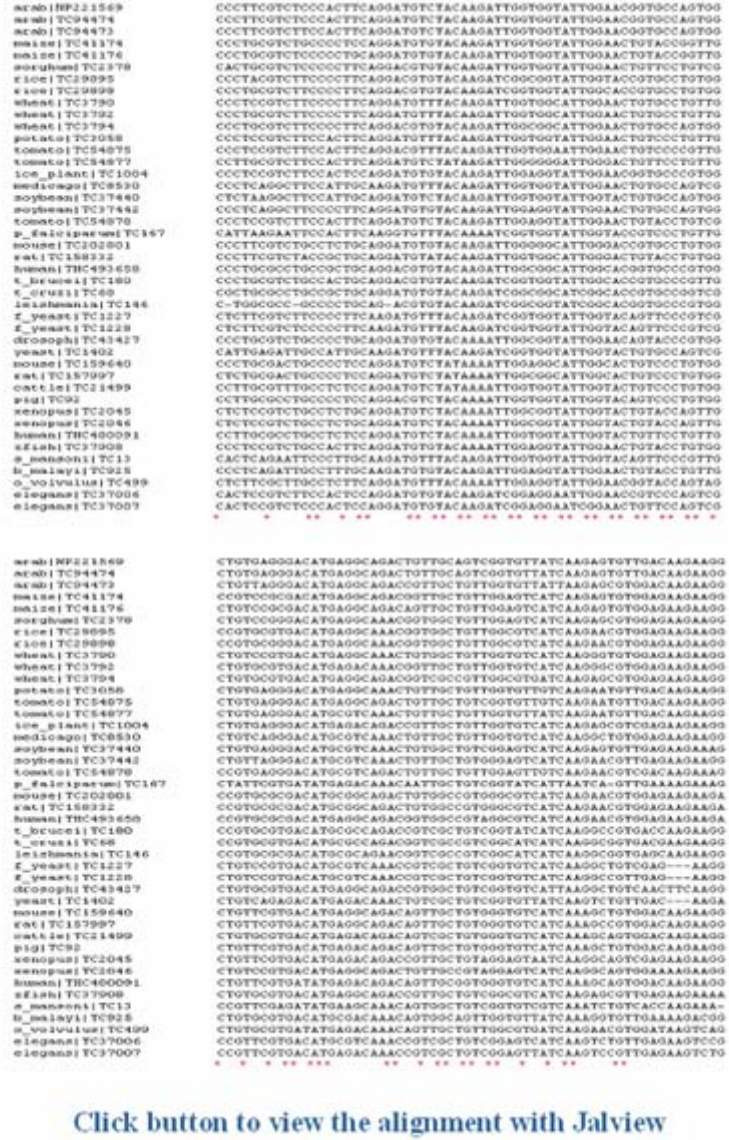


Tentative Ortholog 14405



Sequence1	Sequence2	% Identity	Match length	p-value	Recip. best hits
Arabidopsis|NP221569	B.malayi|TC925	68.00	803	1.0e-87	
Arabidopsis|NP221569	C.elegans|TC37006	75.00	638	6.4e-153	
Arabidopsis|NP221569	C.elegans|TC37007	73.00	731	3.4e-158	
Arabidopsis|NP221569	Cattle|TC21499	72.00	738	1.9e-141	
Arabidopsis|NP221569	Drosophila|TC43427	75.00	709	1.4e-160	*
Arabidopsis|NP221569	Frog|TC2045	75.00	700	8.9e-154	*
Arabidopsis|NP221569	Frog|TC2046	77.00	641	9.8e-154	
Arabidopsis|NP221569	Human|THC480091	75.00	638	1.1e-284	*
Arabidopsis|NP221569	Human|THC493658	75.00	637	2.5e-148	
Arabidopsis|NP221569	Iceplant|TC1004	87.00	1346	2.4e-236	*

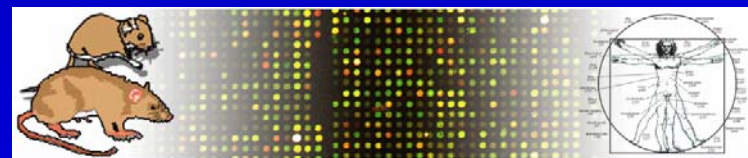
CLUSTAL W (1.81) multiple sequence alignment



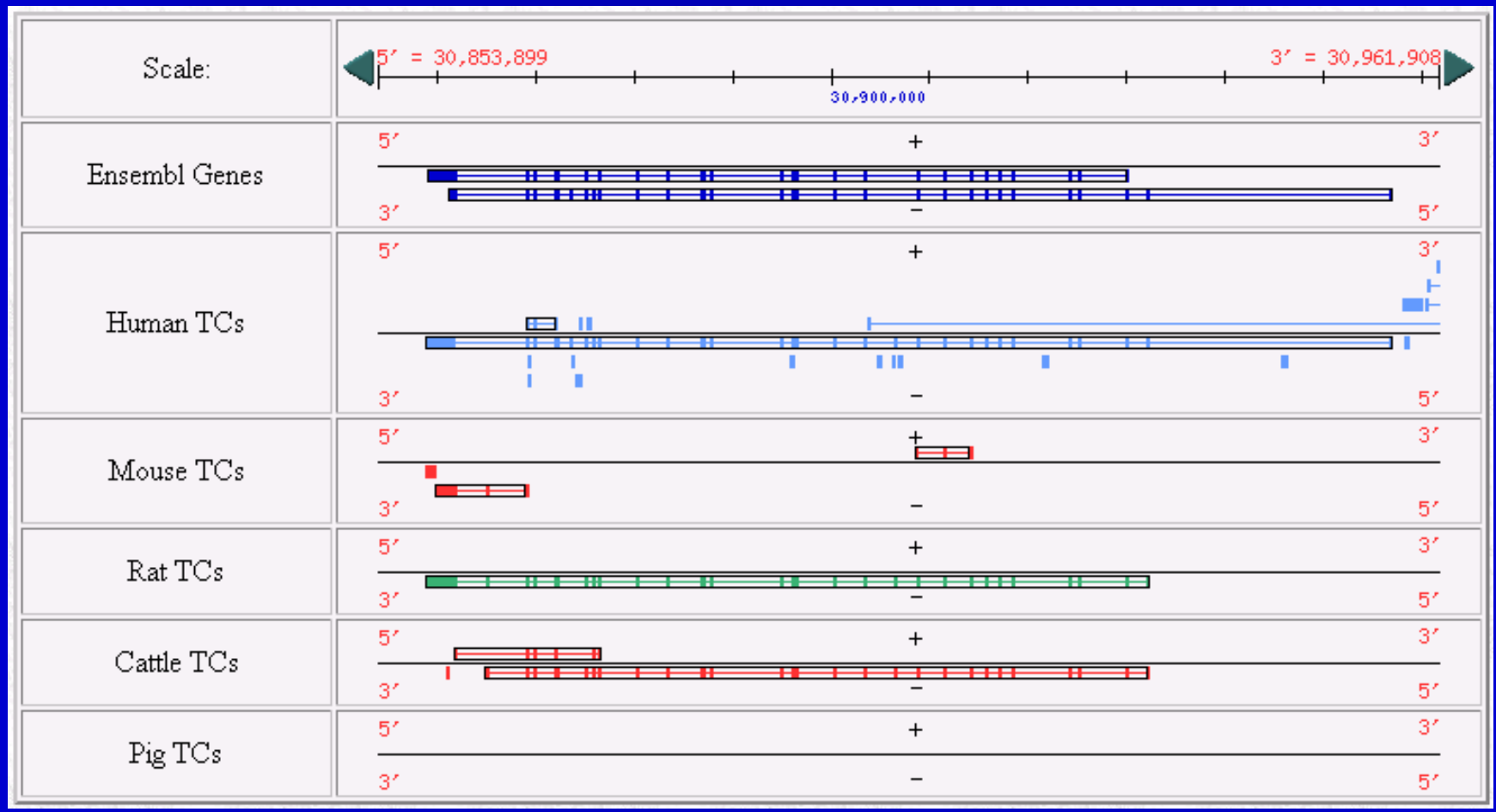
Click button to view the alignment with Jalview

National Heart, Lung, and Blood Institute

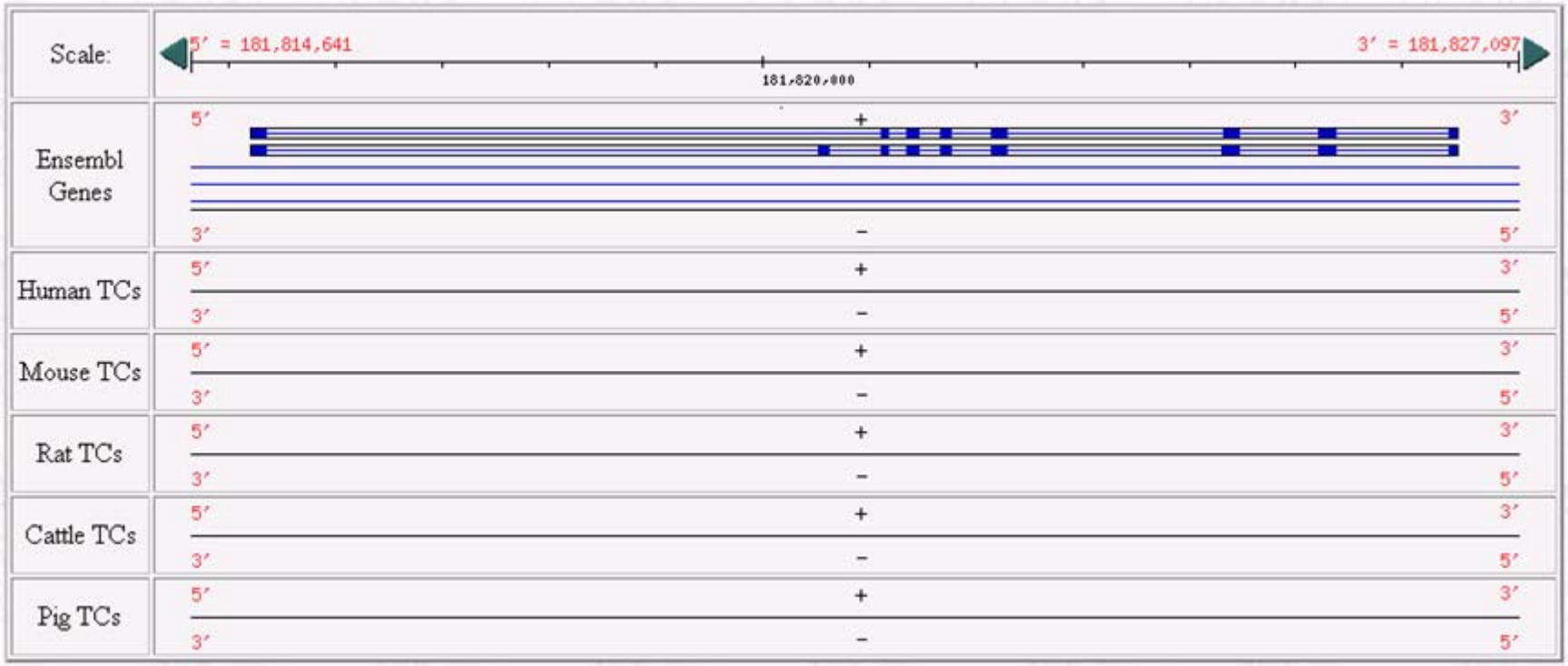
PROGRAMS FOR GENOMIC APPLICATIONS



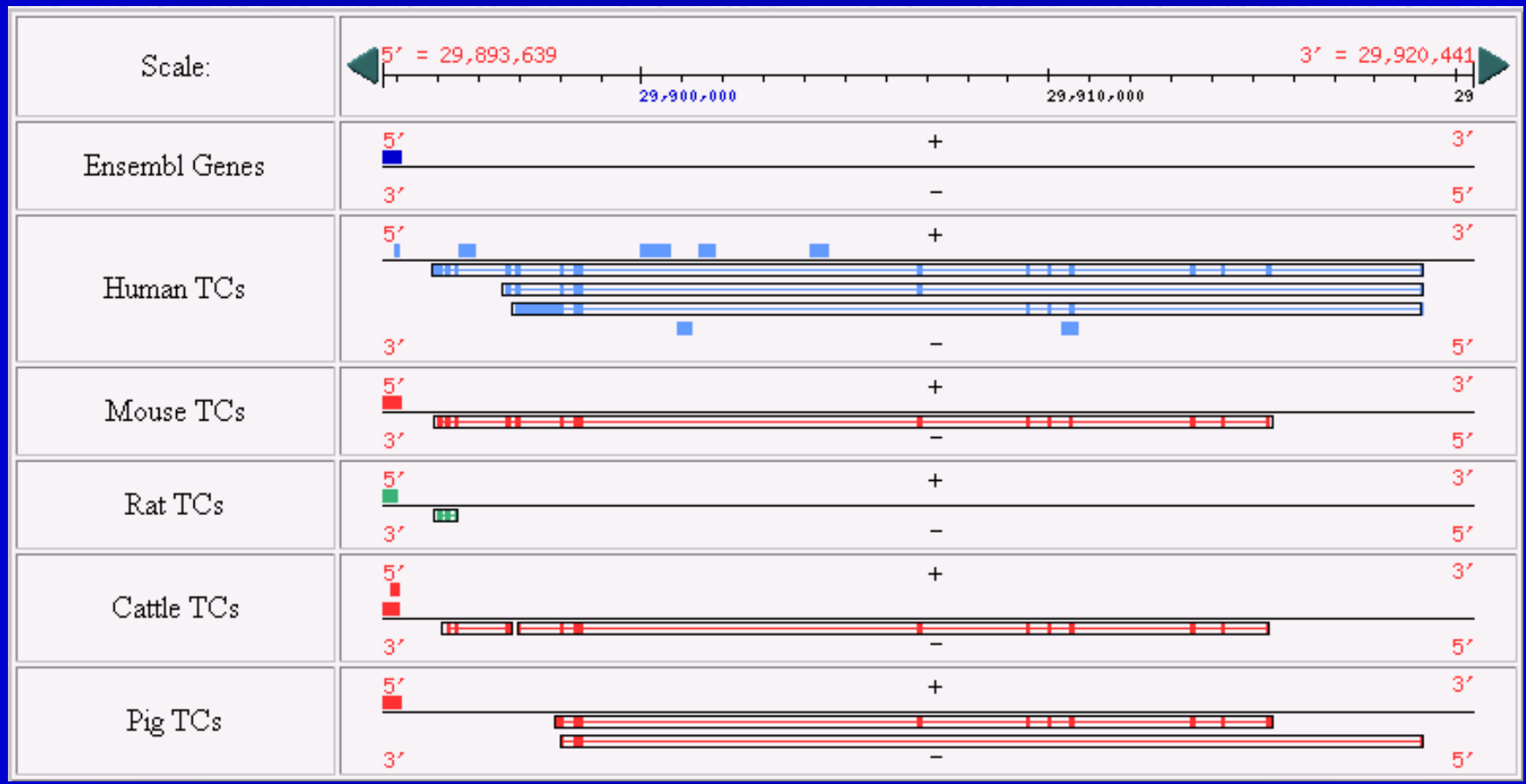
Gene Finding in Humans is easy!



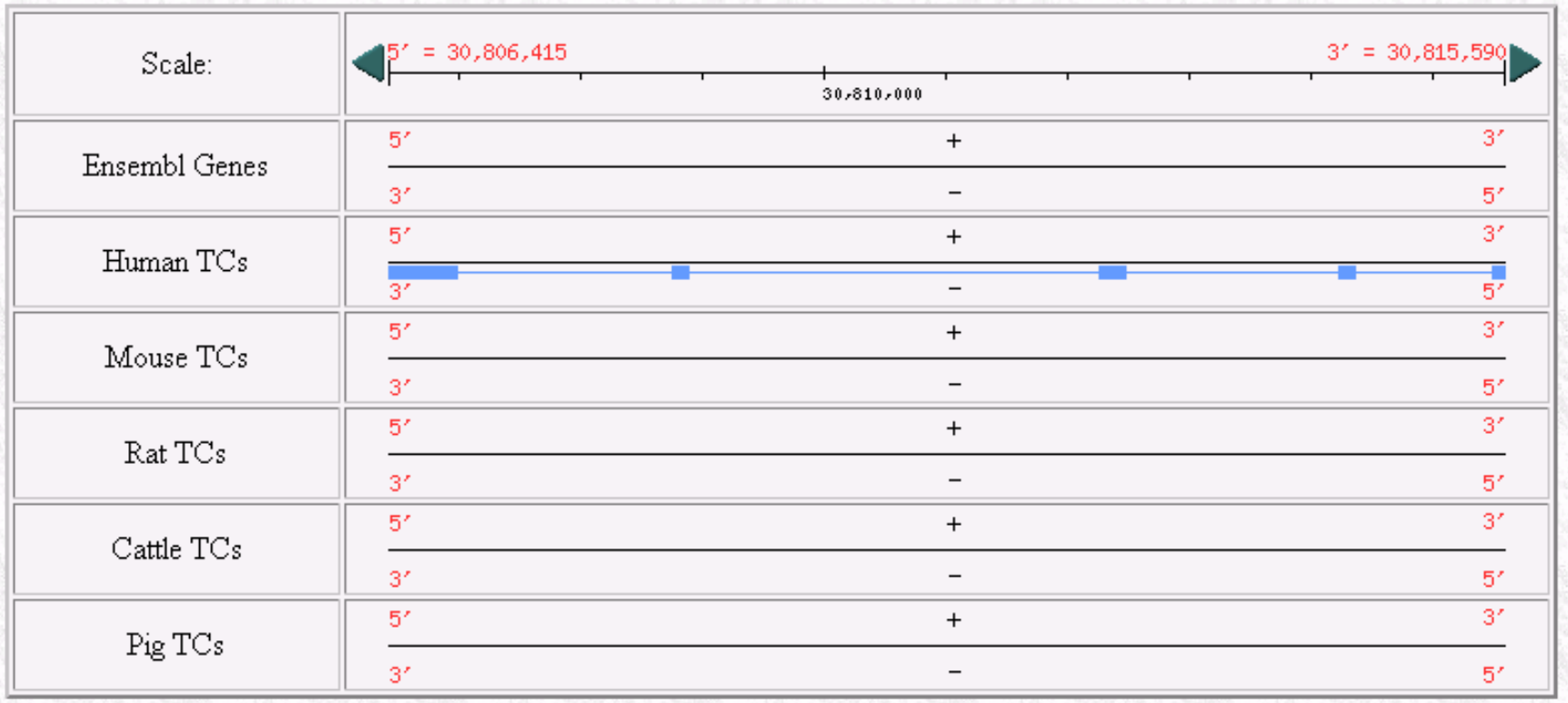
Gene Finding in Humans is easy?



Gene Finding in Humans is difficult?



Gene Finding in Humans is difficult?



A genome and its annotation is *only* a hypothesis that must be tested.

RESOURCERER

Jennifer Tsai

http://pga.tigr.org — The Institute for Genomic Research's Program for Genomic Application

Microarray Expression Profiling of Rodent Models of Human Disease

The Institute for Genomic Research | Duke University | Boston University | The Jackson Laboratory | Medical College of Wisconsin | University of Pennsylvania

- What's New
- Description
- Contacts
- Data
- Analysis Tools
- Targets
- Protocols
- Outreach
- NHLBI PGA Participants

email the PGA_webmaster

RESOURCERER 6.0

READ ME | GB Search | Gen Marker Search | BLAST Search

RESOURCERER(Genome Biology 2001 PDF) provides annotation based on the TIGR Gene Indices (TGI) for commonly available microarray resources, including widely used clone sets and Affymetrix GeneChip Arrays. RESOURCERER also allows comparisons between resources from the same species using either the TGI or UniGene and between species using the EGO database.

- Physical map for Rat Sets NEW
- Rat Genetic Marker Search NEW
- Array search based on physical chromosome coordinates (under Gen Marker Search). NEW
- Addition -- Human: Illumina_Human, Agilent_Human1A. NEW
- Addition -- Mouse: NIA_7.4K, FANTOM1, FANTOM2. NEW

Comments are welcome

Select a single resource in "Data Set A" (while leaving "Data Set B" as "None") to generate hyperlinked annotation.	**Data Set A:**	Human: affy_HG-U95Av2
To Compare Two Resources: Select both, choose the basis for comparison (EGO or UniGene), and the type of comparison to perform (Intersection, A_unique, or B_unique).	**Data Set B:**	None
	◉ EGO ○ UniGene	
	◉ Intersection ○ A_unique ○ B_unique	GetTable

What's New | Description | Contacts | Data | Analysis Tools
Targets | Protocols | Outreach | NHLBI PGA Participants

National Heart, Lung, and Blood Institute — PROGRAMS FOR GENOMIC APPLICATIONS

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

RESOURCERER: An Example

BMP & affy_HG-U95Av2

Based On: EGO

There a**956** rows in this table. Download | Jump to page 1

Page 1**100** is currently displayed. Next

Dataset A	Rearray ID	Clone Nau	GenB Acc	BMAP TC	Dataset B	Probe ID	Clone Name	GenB Acc	affy_HG-U95Av2 TC
BMAP	1-a-4	UI-M-AQ0-aaa-[illegible]-0-UI	AI835193	TC654714	affy_HG-U95Av2	38871_at		AJ006288	THC1358632
BMAP	1-a-7	UI-M-AQ0-aaa-[illegible]-0-UI	AI835201	TC662425	affy_HG-U95Av2	41449_at		AJ000534	THC1264723
BMAP	1-a-9	UI-M-AQ0-aaa-[illegible]-0-UI	AI835206	TC774982	affy_HG-U95Av2	40425_at		M57730	THC1251303
BMAP	1-b-1	UI-M-AQ0-aaa-[illegible]-0-UI	AI835214	TC754936	affy_HG-U95Av2	36542_at		AF030409	THC1252352
BMAP	1-b-3	UI-M-AQ0-aaa-[illegible]-0-UI	AI835221	TC773710	affy_HG-U95Av2	33577_at		AC004079	THC1357542 THC1228447 THC1359621 THC1395026
BMAP BMAP	1-b-3 1-b-3	UI-M-AQ0-aaa-[illegible]-0-UI UI-M-AQ0-aaa-[illegible]-0-UI	AI835221 AI835221	TC773710 TC773710	affy_HG-U95Av2 affy_HG-U95Av2	41321_s_at 41322_s_at	IMAGE-965756 IMAGE-2517632	AA528077 AI816034	THC1357542 THC1357542

GO:0008656 (caspase activator)
GO:0016064 (humoral defense mechanism (sensu Vertebrata))
GO:0016066 (cellular defense response (sensu Vertebrata))

leukemia/lymphoma 10 [Mus musculus]

National Heart, Lung, and Blood Institute PROGRAMS FOR GENOMIC APPLICATIONS

RESOURCERER: Using Genetic Markers

Mapped to Mouse Genome: Chr16

Range: 23628083-28355680

There are **14** rows in this table. Download | Jump to page 1

Page **1** of **1** is currently displayed.

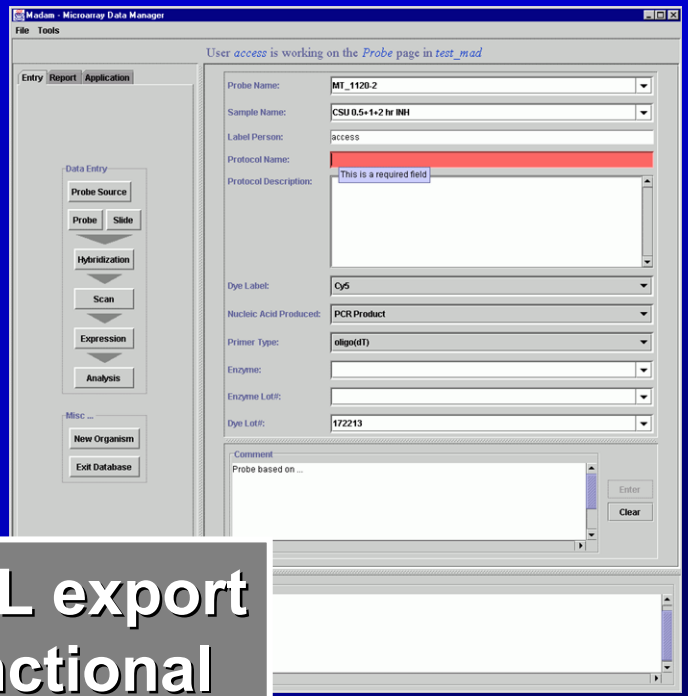
Marker Name	UniSTS ID	Genetic Map	Data Set	GenBank Acc	TIGR TC	Chr Left	Chr Right	TGI Annotation
.MMHAP12FLA1.seq	122645					23628083	23628267	
			TIGR_25K_Mouse_Set	AI848192	TC689579	23651614	23652949	preprosomatostatin; somatostatin [Mus musculus]
			TIGR_25K_Mouse_Set	AI853264	TC673044	23687727	23688118	
			TIGR_25K_Mouse_Set	AU017440	TC763498	24433266	24434248	
			TIGR_25K_Mouse_Set	AU024130	TC721291	24754238	24758505	
			TIGR_25K_Mouse_Set	AW553532	TC749004	25732192	25760355	homologue to GP|7022921|dbj|BAA91769 unnamed protein product {Homo sapiens}, partial (77%
			TIGR_25K_Mouse_Set	AU014937	TC717150	25804670	25813571	
			TIGR_25K_Mouse_Set	AI854418	TC676007	27152177	27210459	RIKEN cDNA 2610529H0 gene [Mus musculus]
			TIGR_25K_Mouse_Set	AW544111	TC669191	27212167	27214345	
			TIGR_25K_Mouse_Set	AU017969	TC671593	27728646	27729800	
			TIGR_25K_Mouse_Set	AI837800	TC683108	27916927	28202210	fibroblast growth factor-relate protein FGF-12A; fibroblas growth factor homologous factor 1; fibroblast growth factor 12 [Mus musculus]
			TIGR_25K_Mouse_Set	AI847382	TC683108	27916927	28202210	fibroblast growth factor-relate protein FGF-12A; fibroblas growth factor homologous factor 1; fibroblast growth factor 12 [Mus musculus]
01.MMHAP38FLA1.seq	122654					28355515	28355680	

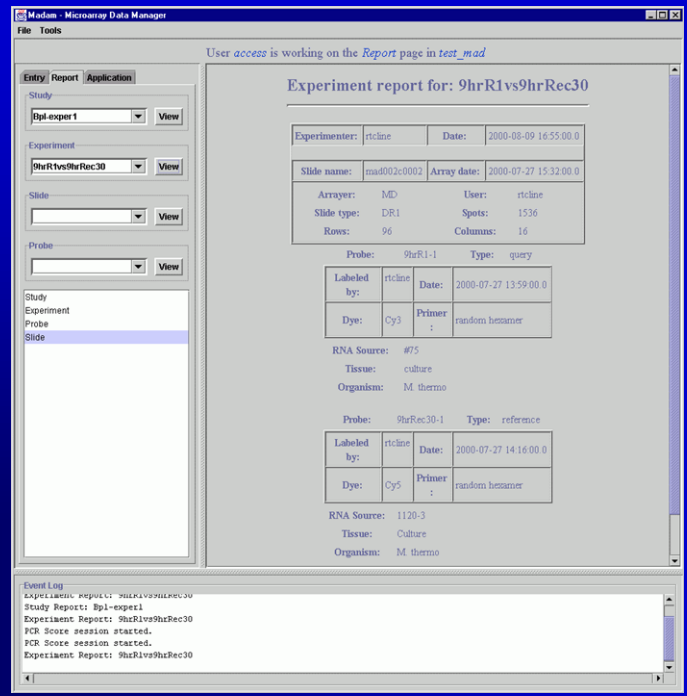
National Heart, Lung, and Blood Institute PROGRAMS FOR GENOMIC APPLICATIONS

Recently added: QTL-based searches

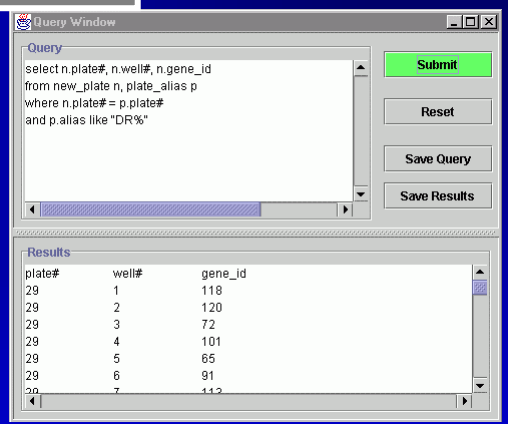
Tools for Array Analysis

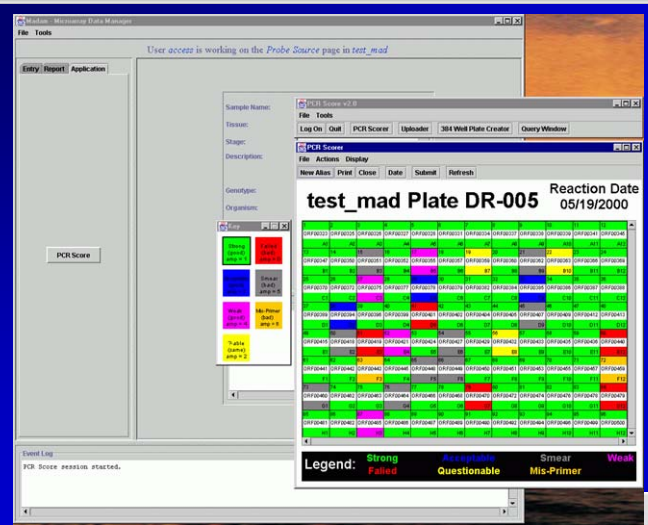
MADAM: Microarray Data Manager





MAGE-ML export now functional



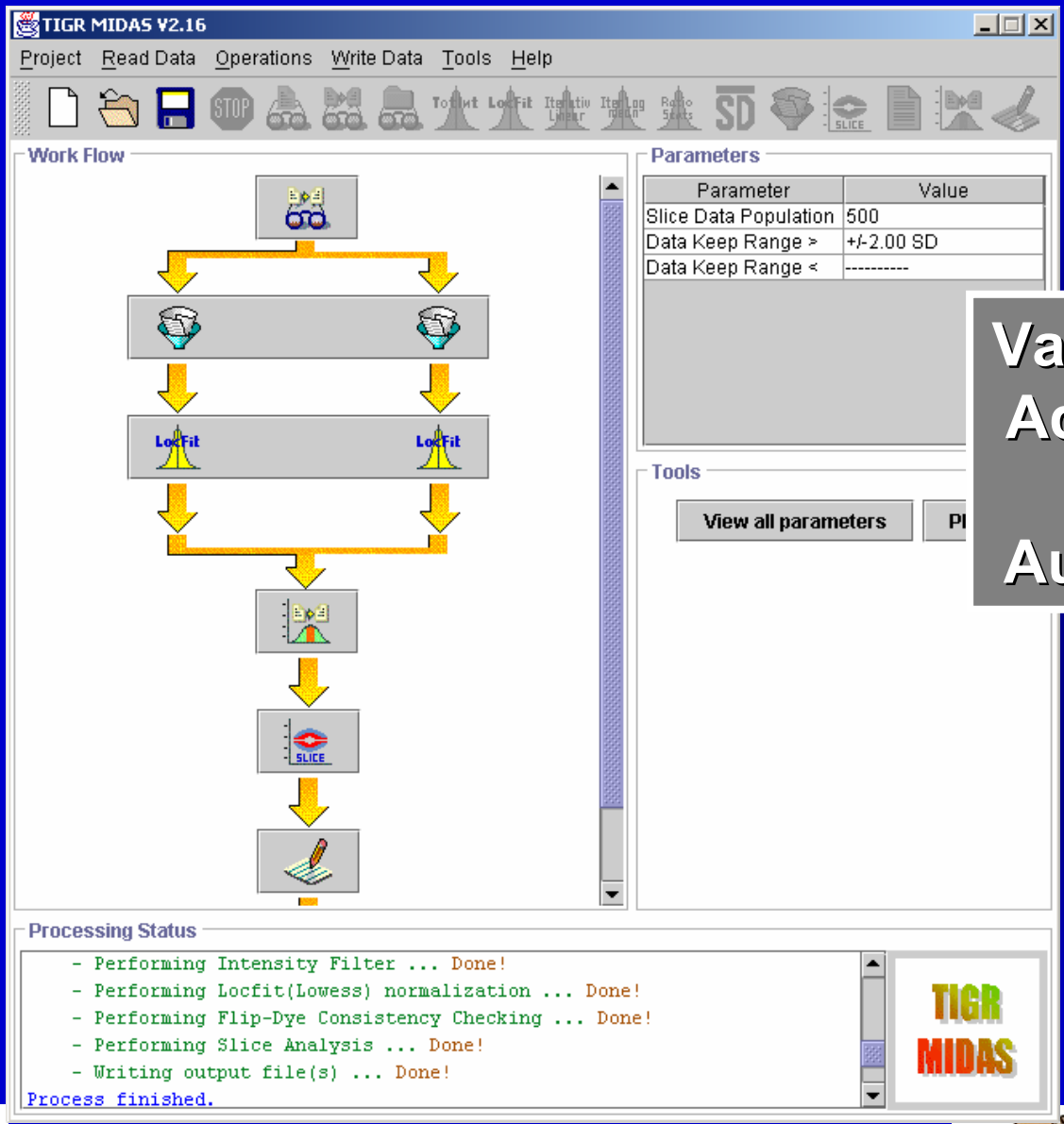


Joseph White
Jerry Li
Alexander Saeed
Vasily Sharov
Syntek Inc.

☺ Available with OSI source and MySQL

MIDAS: Data Analysis

Wei Liang

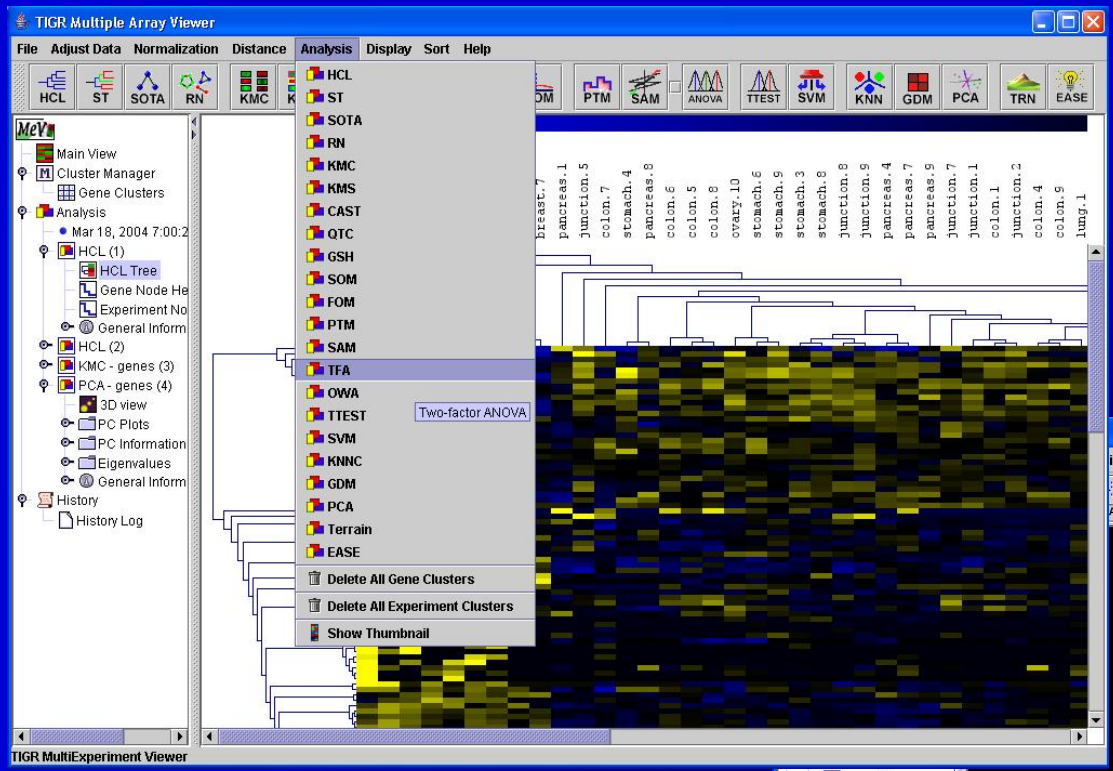


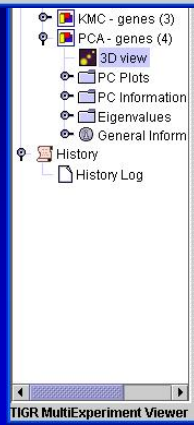
Variance Stabilization, Adding Error Models, MAANOVA, Automated Reporting

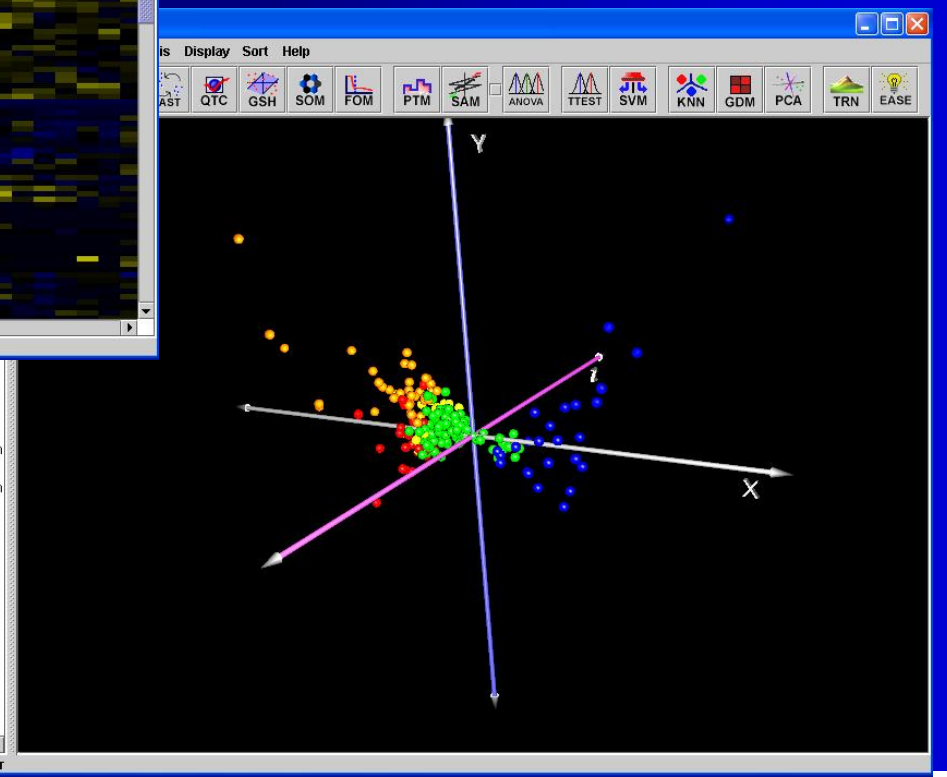
☺ **Available with source**

MeV: Data Mining Tools

Alexander Saeed
Alexander Sturn
Nirmal Bhagabati
John Braisted
Syntek Inc.
Datanaut, Inc.







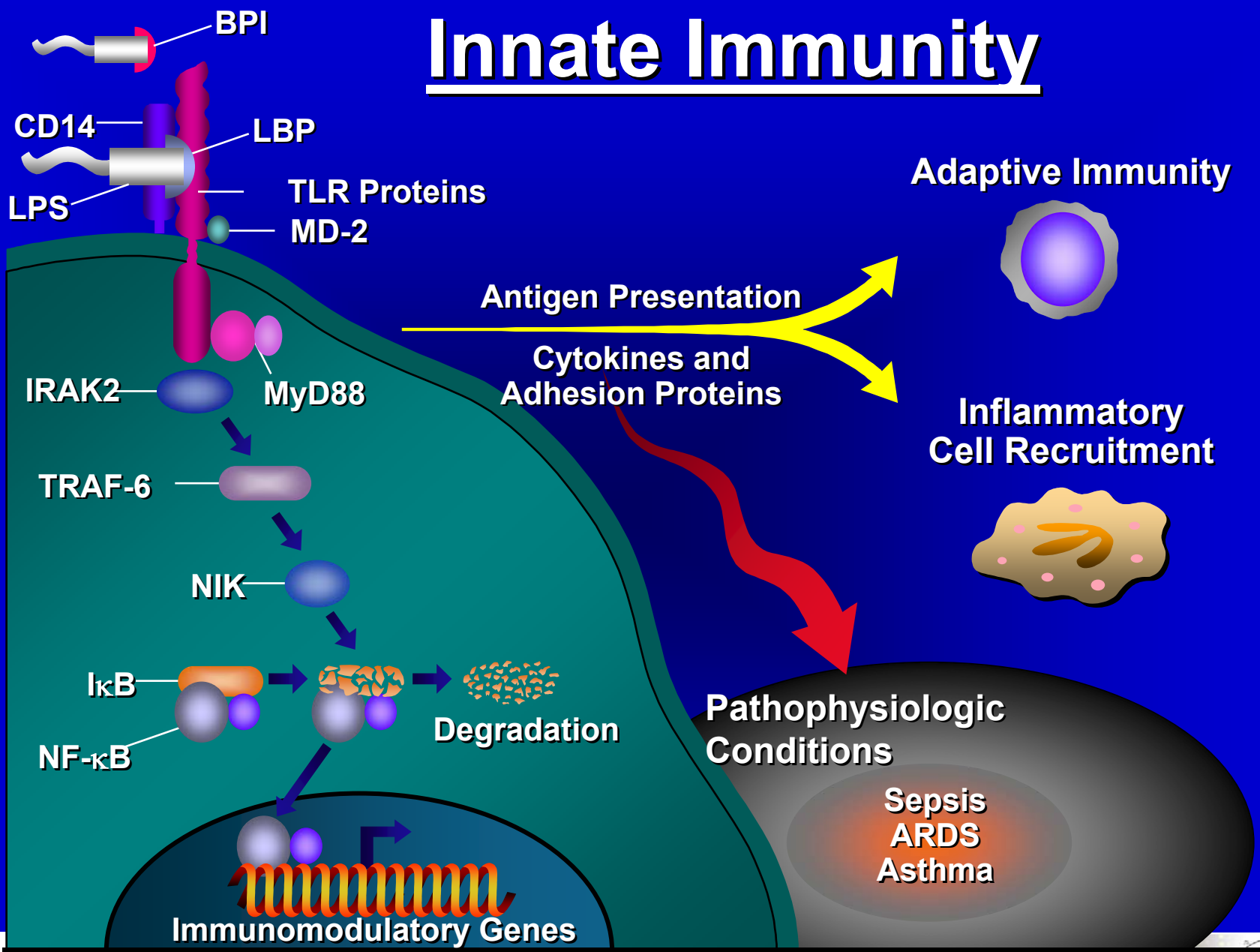
☺ Available with OSI source

Science


National Heart, Lung, and Blood Institute
PROGRAMS FOR GENOMIC APPLICATIONS

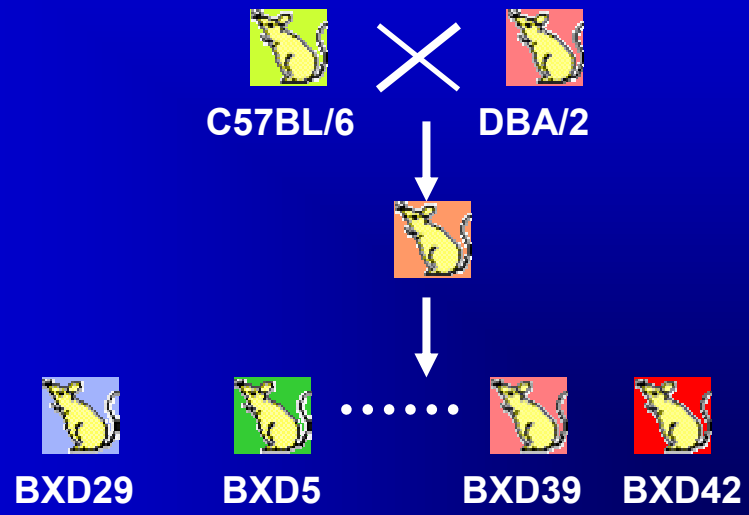
Integrating Expression with other data

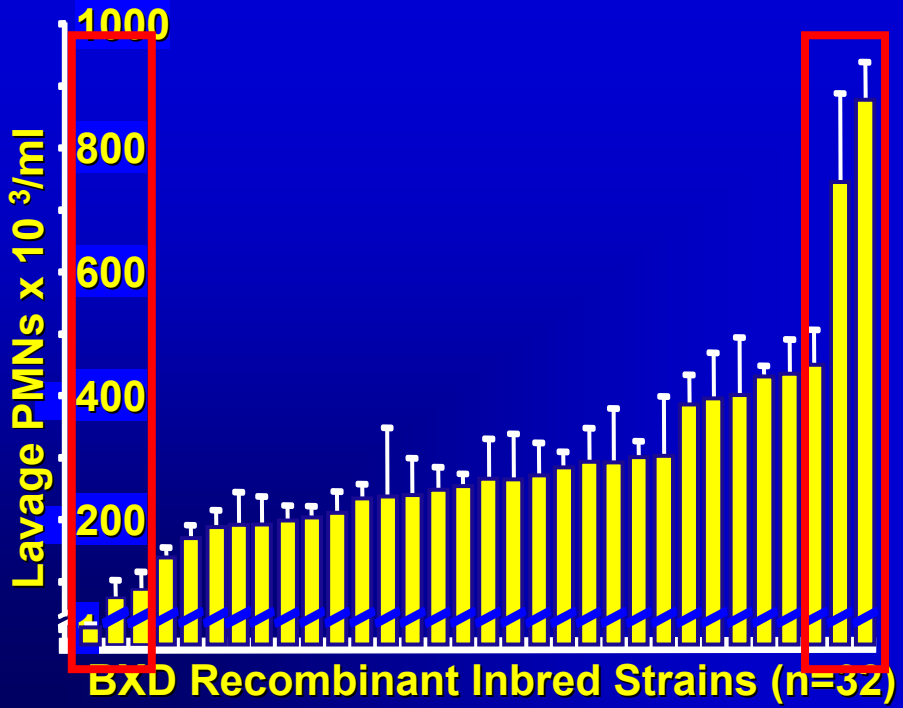
Innate Immunity



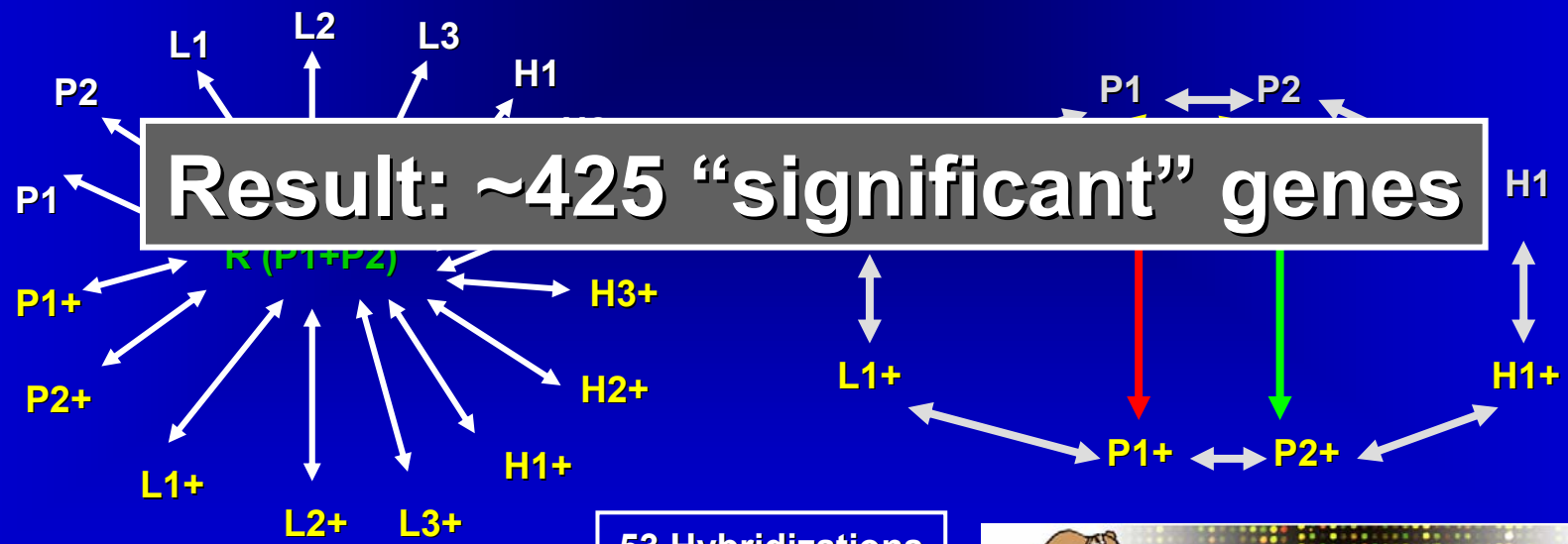
Adapted from Godowski. *NEJM* 1999; 340:1835

Examples





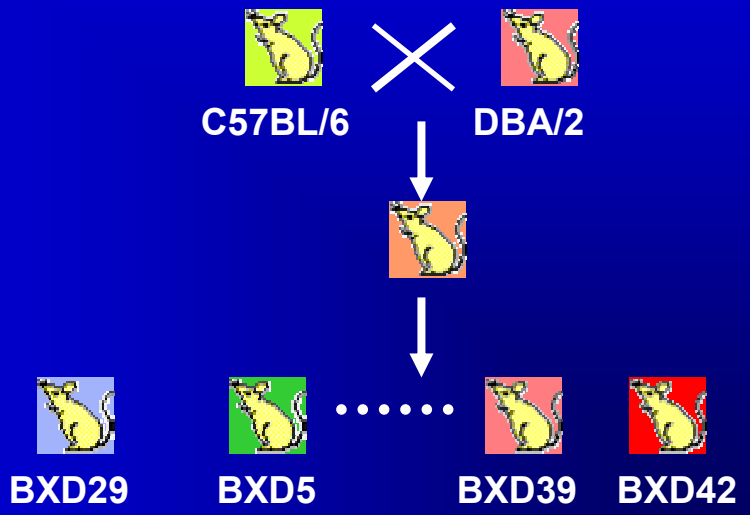
Result: ~425 "significant" genes

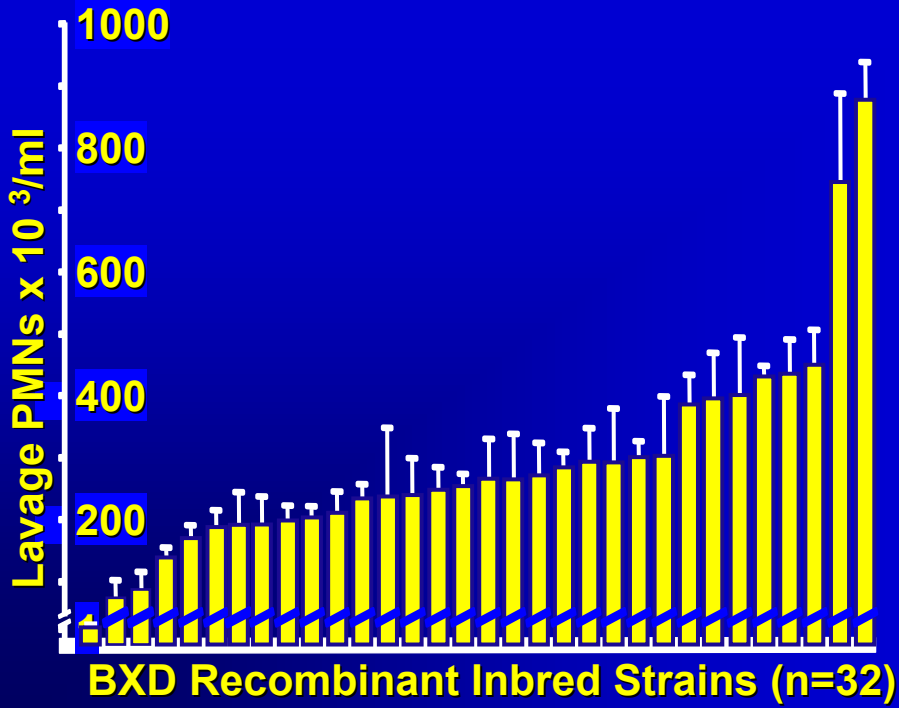


53 Hybridizations

Don Cook, Shuibang Wang, David Schwartz

Examples





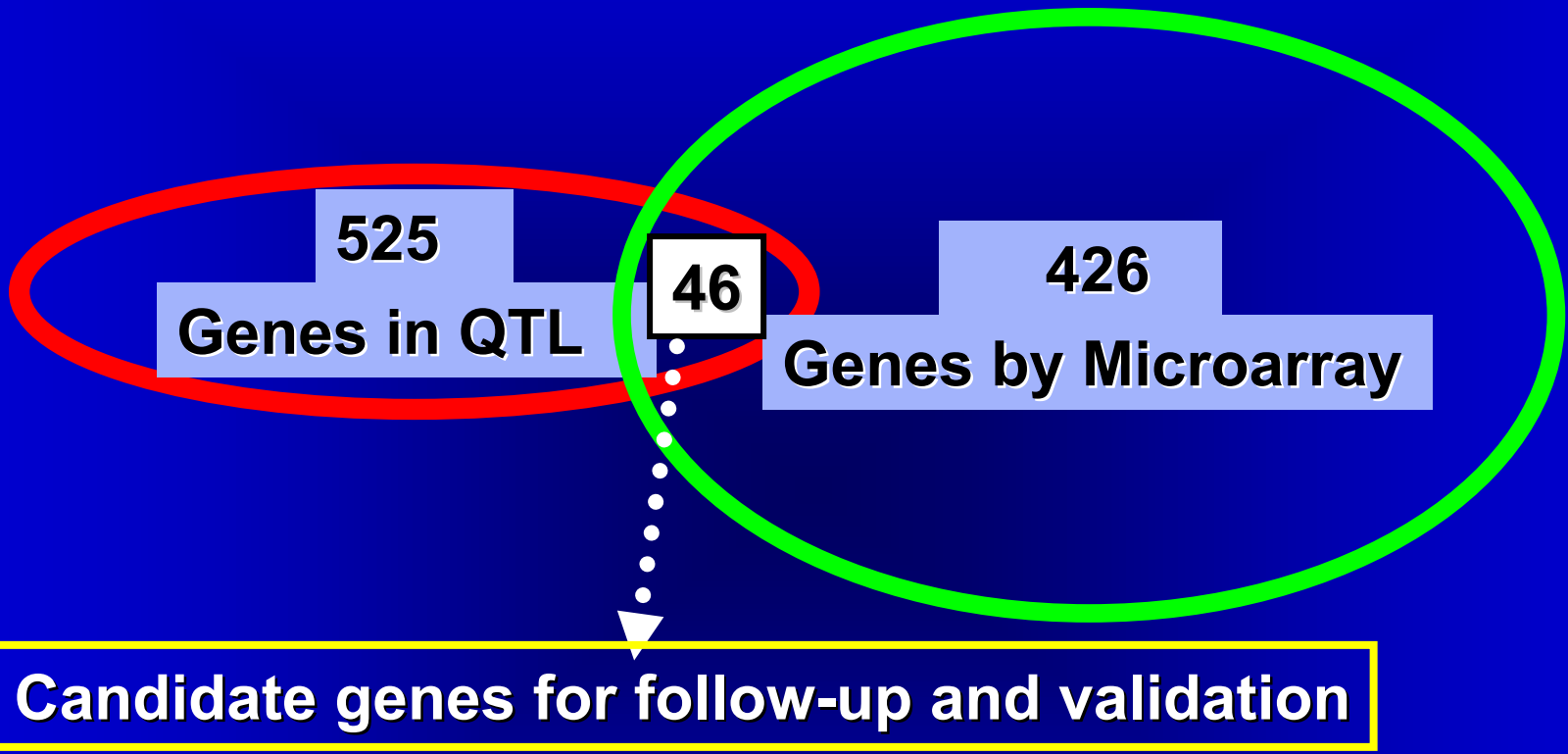
IDEA: Build QTL Maps and use those to filter expression data

Goal: Find differentially expressed genes genetically linked to response

National Heart, Lung, and Blood Institute PROGRAMS FOR GENOMIC APPLICATIONS

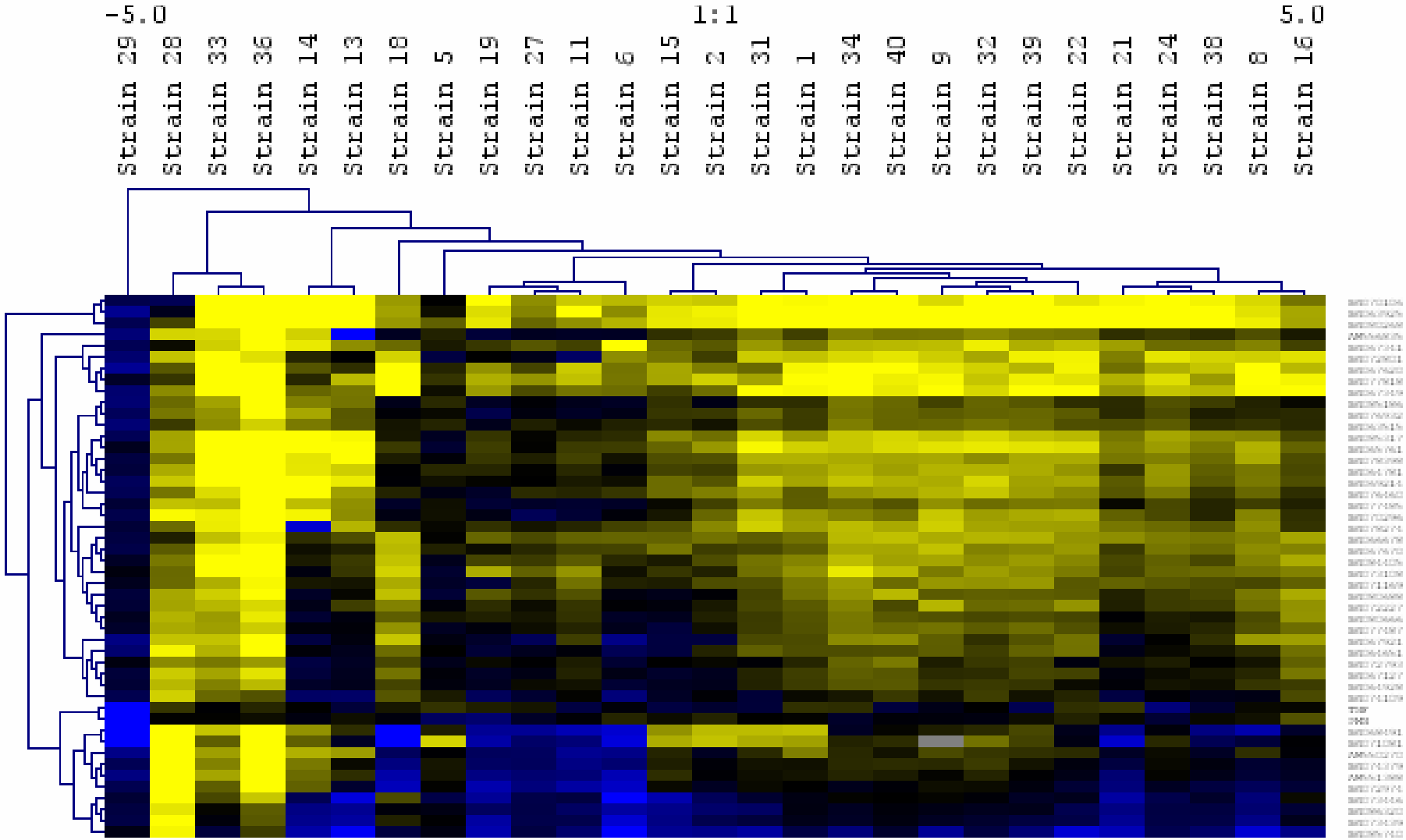
Don Cook, Shuibang Wang, David Schwartz

Microarray Expression-QTL Consensus Candidate Genes



National Heart, Lung, and Blood Institute
PROGRAMS FOR GENOMIC APPLICATIONS
Don Cook, Shuibang Wang, David Schwartz

Candidate Gene Set for LPS response

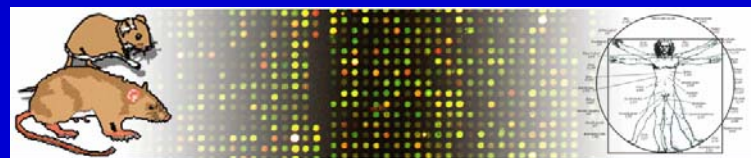


Don Cook, Bryan Frank, David Schwartz

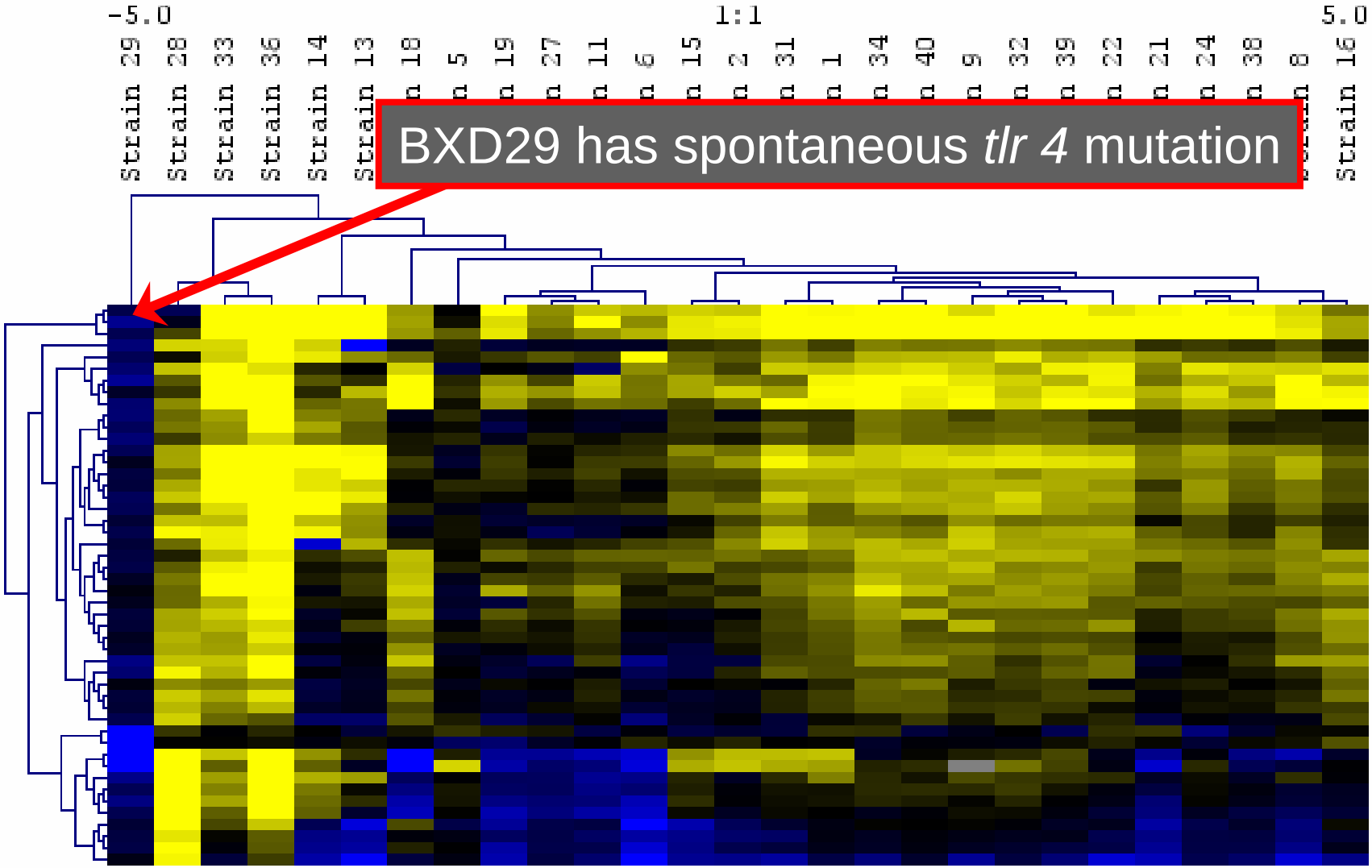
What have we learned?

- Well...
  - Genetics and Expression can be powerful when used together
  - This yields genes that are differentially expressed but also genetically linked to traits
  - The expression fingerprint itself can also be used as a quantitative trait - eQTLs

- But...
  - QTL+Expression may miss regulatory and other genes whose expression does not change but which contain mutations that are causative



Candidate Gene Set for LPS response



Don Cook, Bryan Frank, David Schwartz

Sleep Deprivation Studies in Mouse

National Heart, Lung, and Blood Institute
PROGRAMS FOR GENOMIC APPLICATIONS

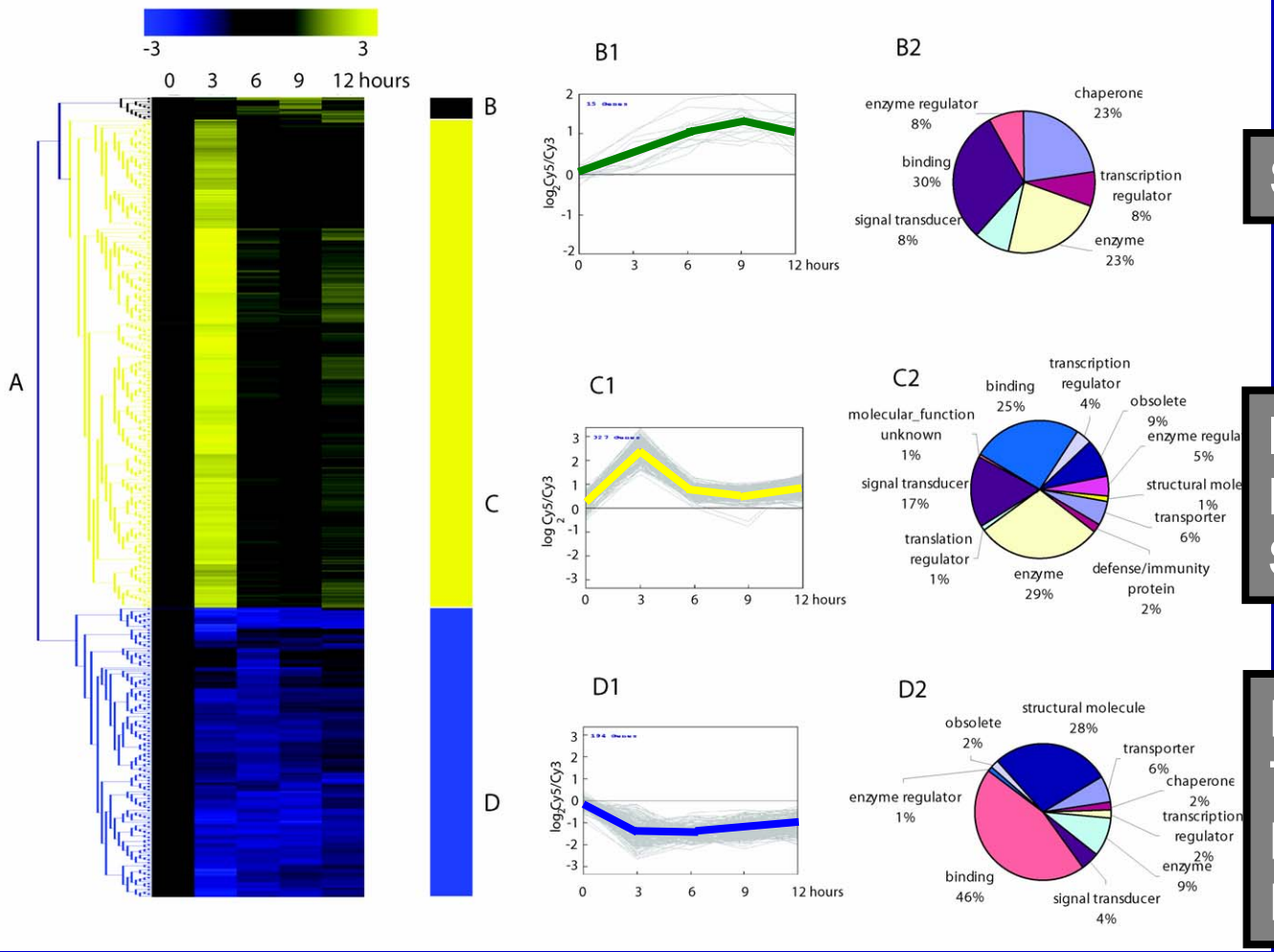
Mirek Mackiewicz, Yonghong Wang, Allan Pack

Experimental Paradigm

- Compare gene expression between sleeping and sleep-deprived mice in cortex and hypothalamus
- Perform 3 biological replicates
- Normalize and filter data and use data mining techniques to select distinct patterns of gene expression
- Use Gene Ontology (GO) assignments to classify genes by cellular localization, molecular function, biological process
- Use GO analysis to develop an understanding of response

Mirek Mackiewicz, Yonghong Wang, Allan Pack

Differential Expression in Cortex



Stress Response

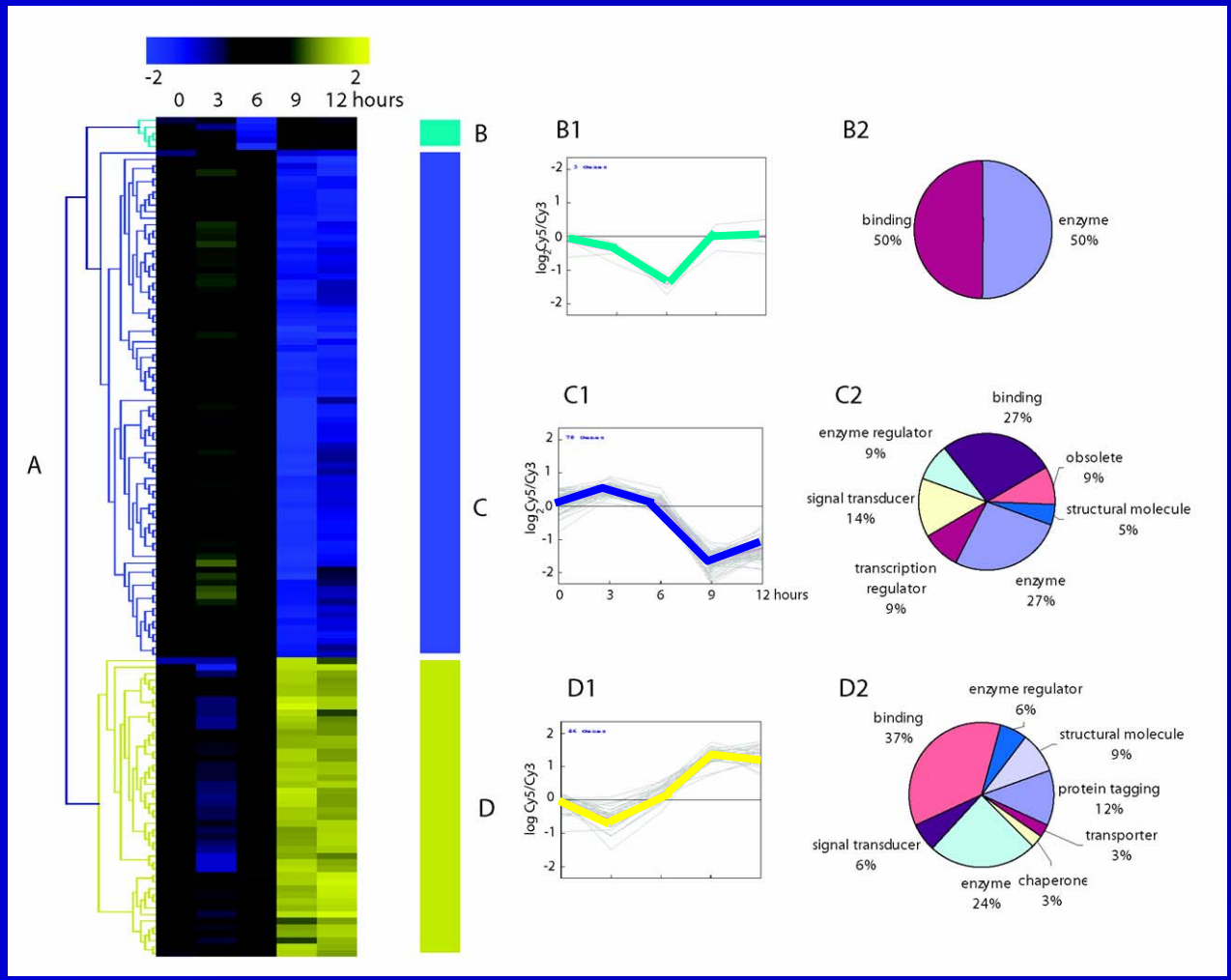
Intermediate Metabolism and Signal Transduction

Energy Metabolism Transcription; Mitochondrial and Ribosomal Proteins

National Heart, Lung, and Blood Institute — PROGRAMS FOR GENOMIC APPLICATIONS

Mirek Mackiewicz, Yonghong Wang, Allan Pack

Differential Expression in Hypothalamus



Sleep signaling

National Heart, Lung, and Blood Institute — PROGRAMS FOR GENOMIC APPLICATIONS

Mirek Mackiewicz, Yonghong Wang, Allan Pack

EASE Analysis of GO terms

Cortex – Up-regulated Genes		
GO Class	***GO Category***	***p-value***
GO Cellular Component	endoplasmic reticulum	6.06×10^{-03}
GO Molecular Function	heat shock protein activity	8.78×10^{-04}
	pyruvate dehydrogenase (lipoamide) phosphatase activity	3.17×10^{-03}
	chaperone activity	7.38×10^{-03}
	mitochondrial inner membrane	3.70×10^{-03}
GO Molecular Function	structural constituent of ribosome	6.46×10^{-39}
	RNA binding activity	4.83×10^{-21}
	cytochrome c oxidase activity	9.79×10^{-04}
	hydrogen ion transporter activity	1.88×10^{-03}

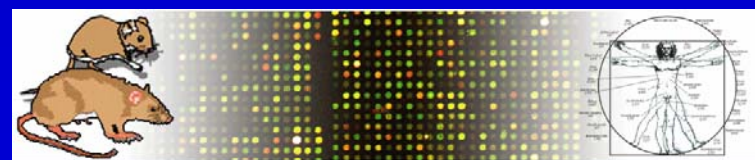
Themes:
General biological trends based on representation of functional roles on the array

Problem:
Requirement of functional class assignment limits utility for discovery of new functional networks

Thanks to Doug Hosack and Glynn Dennis, NIAID

What have we learned?

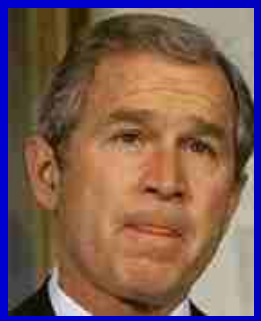
- Well...
  - Functional assignments provide a powerful filter on the measured expression
- But...
  - Functional classes are not functions
  - Significant additional work is necessary to translate these classes to pathways and responses



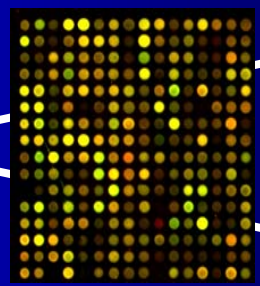
Predicting Outcome

work in collaboration with
Timothy J. Yeatman
H. Lee Moffitt Cancer Center

The Classification Problem

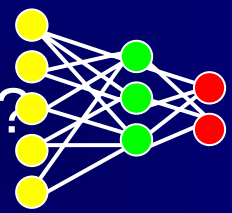


Sample



Profile Expression

???



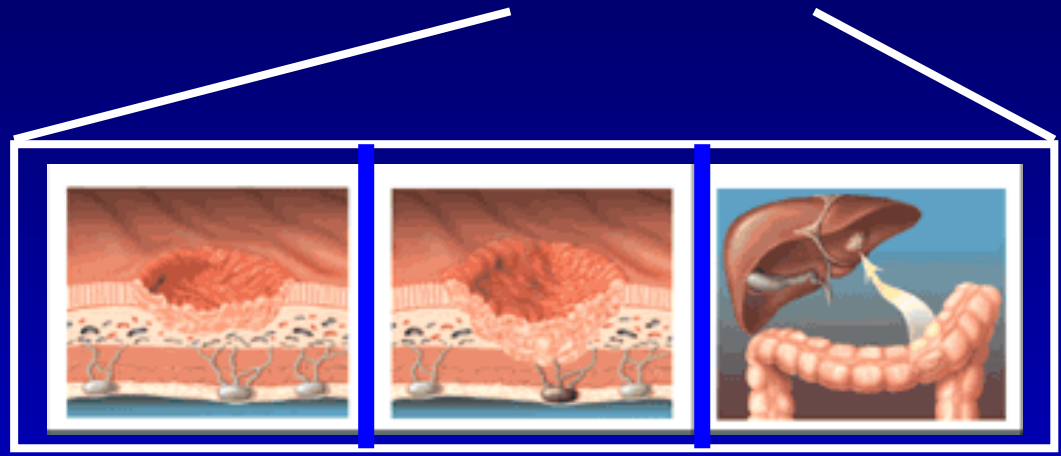
ANN
KNN
QDA
SVM
etc.

Class?

RNA and Protein Correlations

Colon Cancer Progression

Normal → Adenoma → Primary Tumor → Metastasis



	Dukes' B	Dukes' C	Dukes' D
5 year survival	70-85%	25-60%	5%

Timothy Yeatman, Simon Kwong, Greg Bloom, Steve Eschrich, Leigh Anderson

Experimental Design

- 6 groups: normal, adenoma, Dukes B, Dukes C, Dukes D, metastases
- 10 samples per group
- Microarray:
  - 2 color 32k array
  - Samples hybridized to common reference
  - Dye-reversal replicates performed
  - Total RNA, no amplification
- Microarray:
  - 2D-PAGE followed by mass-spectrometry

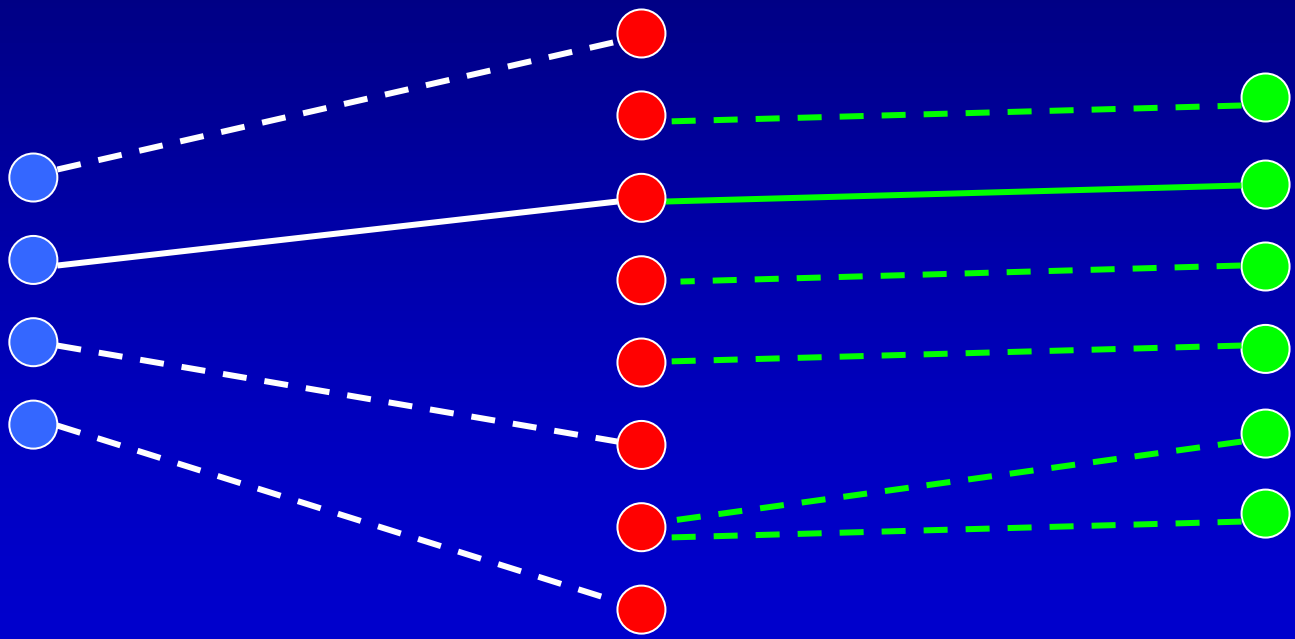
Linking genes to proteins

Array Probes are cDNAs, not genes

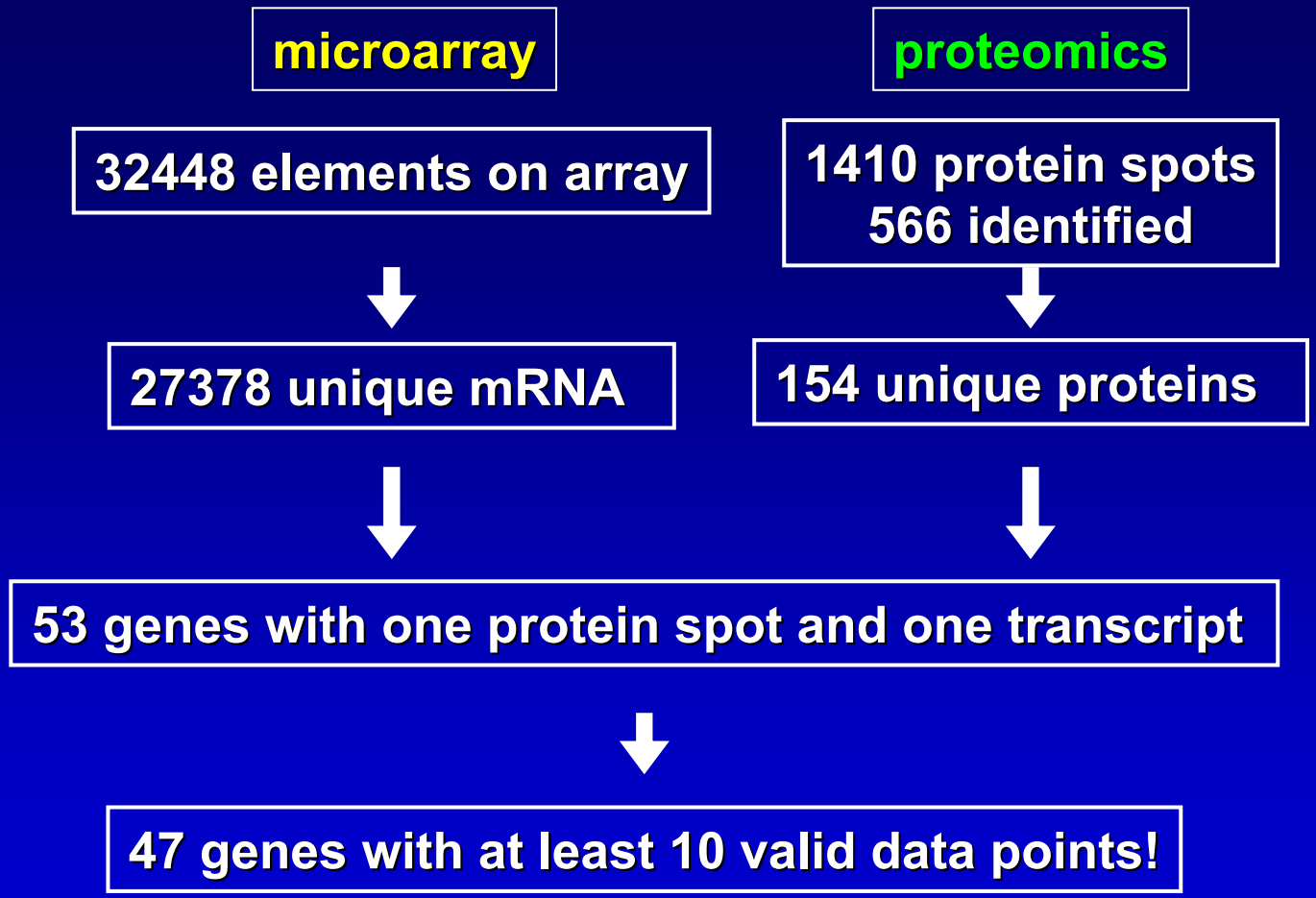
Protein Sequence
SwissProt

TIGR
Tentative Consensus (TC)

cDNA Sequence
Genbank



Linking genes to proteins

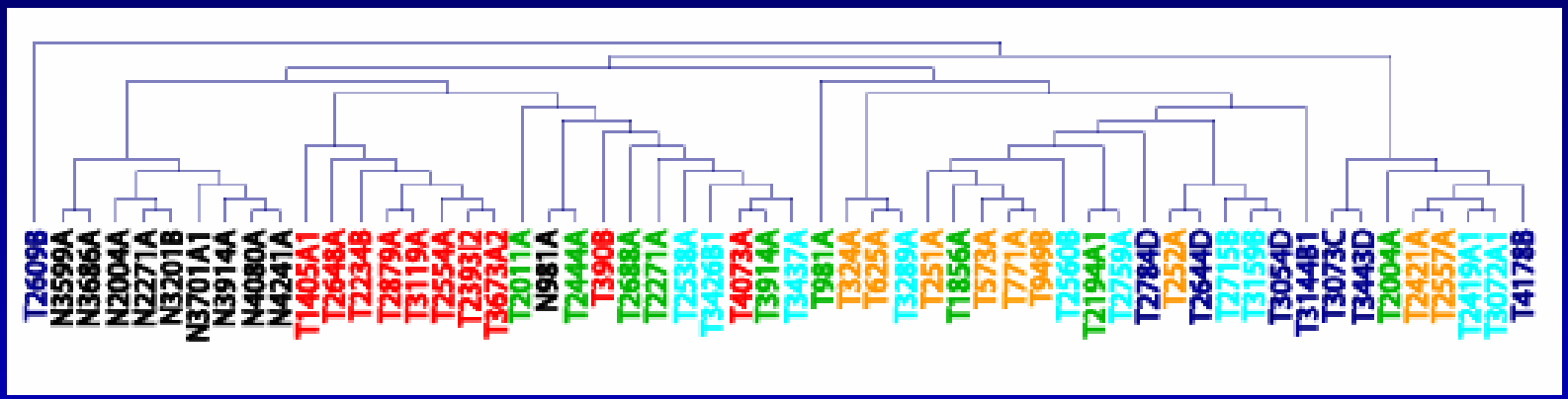


Timothy Yeatman, Simon Kwong, Greg Bloom, Steve Eschrich, Leigh Anderson

“Highly” correlated genes (12/47)

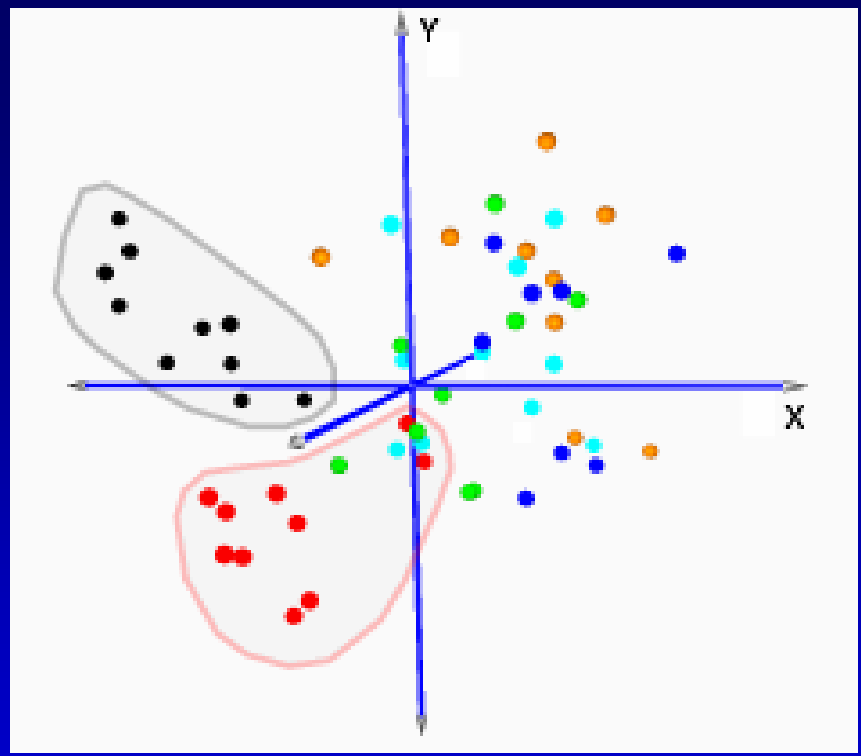
Description	r	p values
Proteasome activator complex subunit 2	0.682	2.53E-08
Metastasis inhibition factor nm23	0.579	8.14E-06
Proliferating cell nuclear antigen (PCNA)	0.567	1.08E-05
Proliferation-associated protein 2G4	0.558	1.88E-05
Fatty acid-binding protein, liver (L-FABP)	0.797	3.79E-05
Peroxiredoxin 4	0.517	7.18E-05
Creatine kinase, B chain	0.545	9.40E-05
Mitochondrial 3-hydroxyisobutyrate dehydrogenase	0.808	2.35E-04
Cadherin-17 precursor	0.516	2.09E-03
Argininosuccinate lyase	0.408	3.29E-03
F-actin capping protein alpha-2 subunit	0.336	0.01
Mitochondrial acyl-CoA dehydrogenase	0.487	0.03

Hierarchical Clustering by mRNA expression



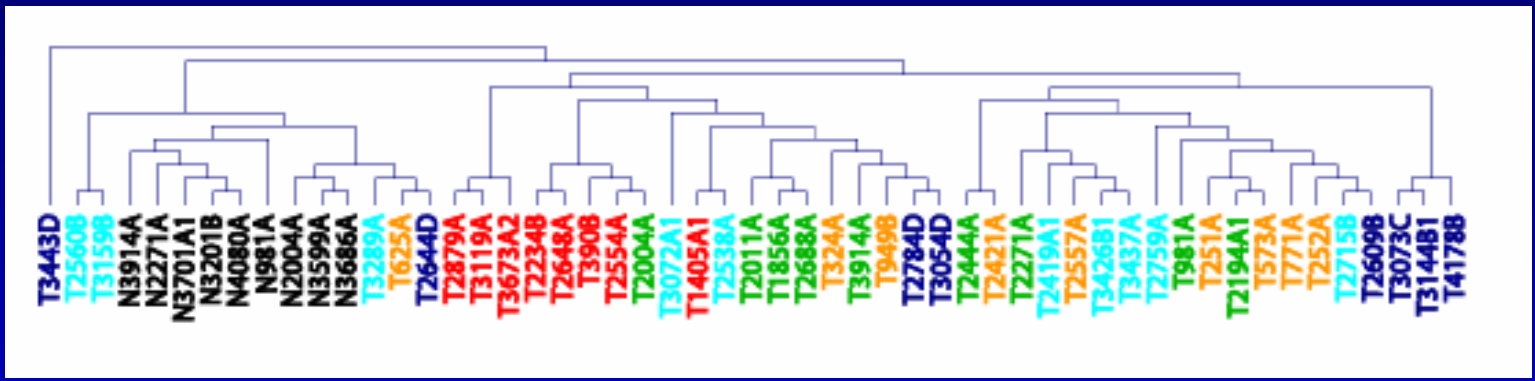
Timothy Yeatman, Simon Kwong, Greg Bloom, Steve Eschrich, Leigh Anderson

Microarray PCA



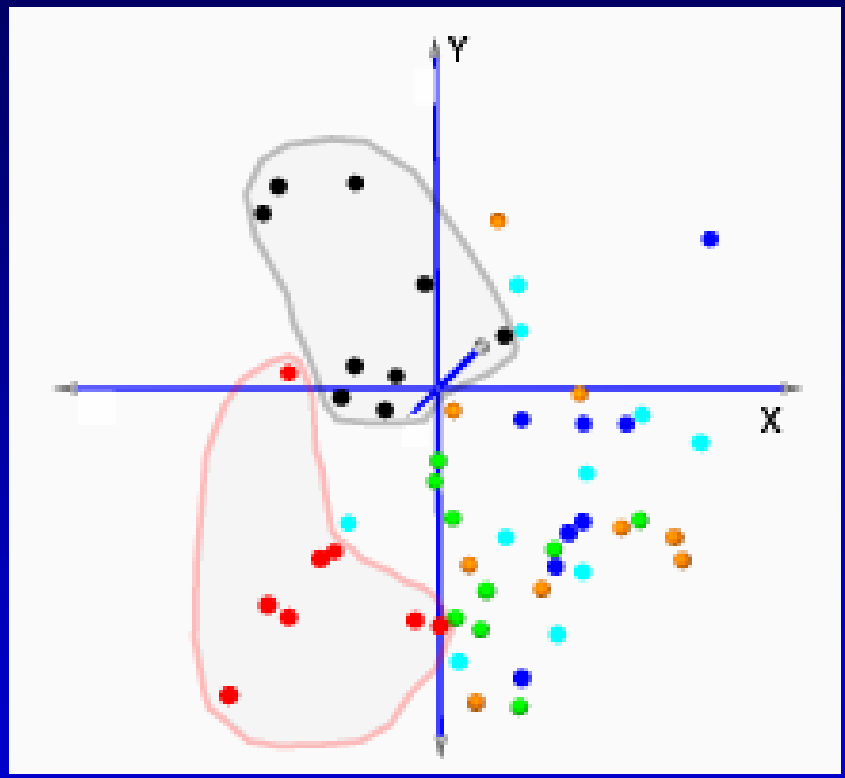
normal adenoma Duke B Duke C Duke D metastasis

Hierarchical Clustering by protein expression



Timothy Yeatman, Simon Kwong, Greg Bloom, Steve Eschrich, Leigh Anderson

Proteomics PCA



normal adenoma Duke B Duke C Duke D metastasis

What have we learned?

- RNA and proteins do not correlate well
- There are numerous possible technical explanations for this – requires further analysis
- But RNA and protein paint the same fundamental picture of the disease – primary tumor fingerprints do not respect clinical staging classes

Predicting Survival

work in collaboration with
Timothy J. Yeatman
H. Lee Moffitt Cancer Center

Survival Classification Set

- Start with 73 Dukes' B and C tumors with survival data
- Profile expression on arrays
- Use SAM with Censored Survival Analysis to select significant genes
- Use unsupervised k-means support to separate tumors into 2 classes

Ivana Yang, Steve Eschrich, Greg Bloom, Tim Yeatman

SAM SIGNIFICANT GENES

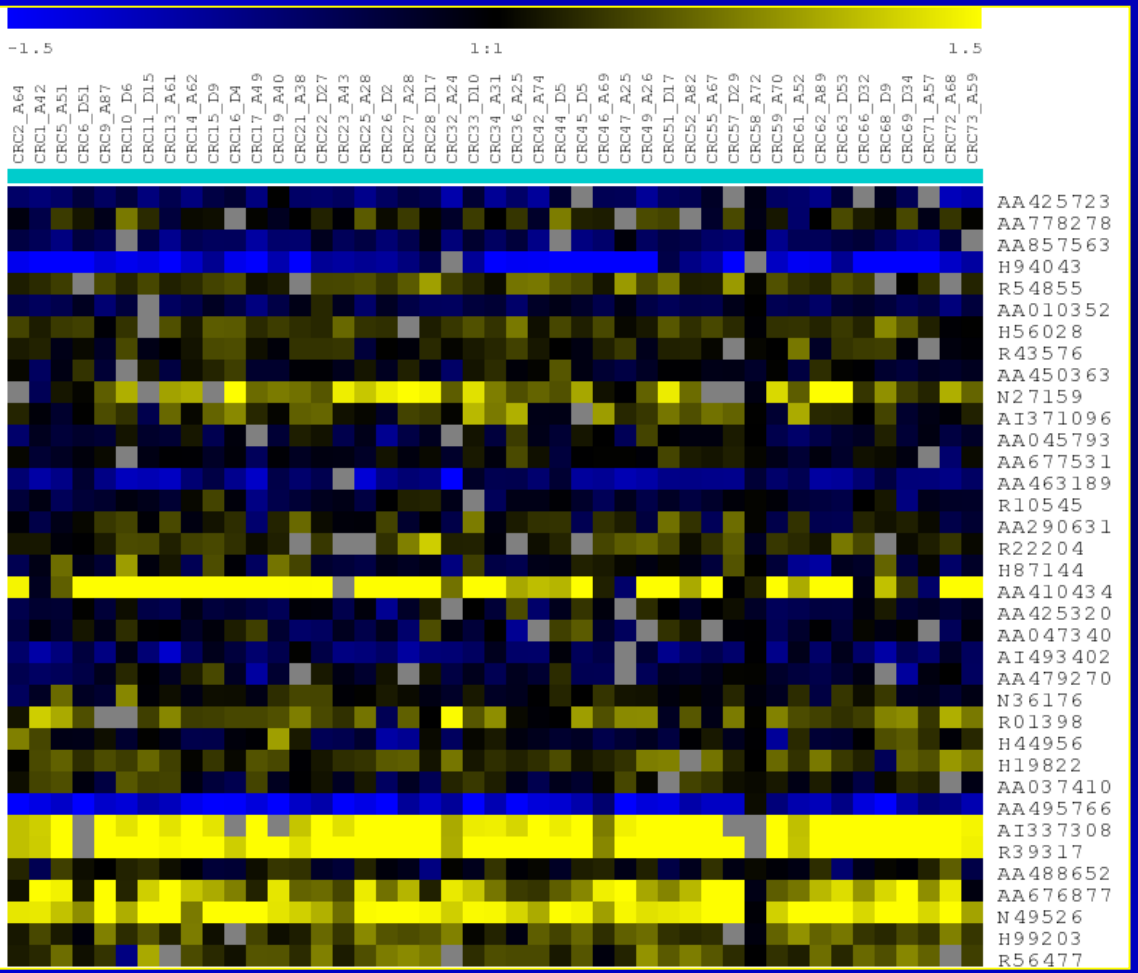
Original row	Clone Name	GB#	TC#	Putative Role
23667	Image:488033	AA045793	THC915844	DnaJ homolog subfamily B member 9 (Microvascular endothelial differentiation gene-1 protein) (Mdg
4697	Image:460395	AA677531	THC925057	Similar to hypothetical protein FLJ22625 {Homo sapiens}
11029	Image:128457	R10545		
17380	Image:130854	R22204	THC987381	
5152	Image:242797	H94043	THC1013186	yes-associated protein homolog DKFZp586I1419.1 - human (fragment)
28052	Image:2043415	AI371096	THC1022305	Death-associated protein kinase 1 (EC 2.7.1.-) (DAP kinase 1). [Human] {Homo sapiens}
9637	Image:700461	AA290631	THC918268	unnamed protein product {Homo sapiens}
25310	Image:796878	AA463189	THC942191	brain-specific GTP-binding protein {Homo sapiens}
26499	Image:269815	N27159	THC899426	Inhibin beta A chain precursor (Activin beta-A chain) (Erythroid differentiation protein) (EDF).
6925	Image:773278	AA425320	THC915844	DnaJ homolog subfamily B member 9 (Microvascular endothelial differentiation gene-1 protein) (Mdg
14102	Image:252453	H87144	THC958545	CGI-86 protein {Homo sapiens}
6146	Image:509516	AA047340	THC908165	unnamed protein product {Homo sapiens}
29198	Image:2115602	AI493402	THC1014756	transcription factor SL1 - human
28743	Image:754250	AA479270	THC862179	KIAA1253 protein {Homo sapiens}
30718	Image:272694	N36176	THC960848	hypothetical protein {Homo sapiens}
10497	Image:753428	AA410434	THC986960	Similar to RIKEN cDNA 1110014B07 gene {Homo sapiens}
1922	Image:123742	R01398	THC1024446	unknown {Homo sapiens}
18930	Image:183200	H44956	THC863066	Fumarylacetoacetase (EC 3.7.1.2) (Fumarylacetoacetate hydrolase) (Beta-diketonase) (FAA). [Human]
20064	Image:24067	R39317	THC987467	putative {Mus musculus}
24682	Image:172495	H19822	THC986991	hypothetical protein {Homo sapiens}
29976	Image:2062345	AI337308	THC987466	protein-tyrosine kinase EPHB2v {Homo sapiens}
27256	Image:897107	AA676877	THC1021656	mitochondrial citrate transport protein {Homo sapiens}
28497	Image:768316	AA495766	THC925935	RCC1-like G exchanging factor RLG [imported] - human
30633	Image:321271	AA037410	THC864588	rho GTPase activating protein 8 isoform 1 {Homo sapiens}
23089	Image:243549	N49526	THC1031960	Myb proto-oncogene protein (C-myb). [Human] {Homo sapiens}
21529	Image:741474	AA401111	THC897309	Glucose-6-phosphate isomerase (EC 5.3.1.9) (GPI) (Phosphoglucose isomerase) (PGI)
9651	Image:416280			
20264	Image:843263	AA488652	THC862216	ribosomal protein L2 {Homo sapiens}
3184	Image:509458	AA056375	THC899187	
17742	Image:261829	H99203	THC968032	Ubiquitin carboxyl-terminal hydrolase 7 (EC 3.1.2.15) (Ubiquitin thiolesterase 7)
30132	Image:41170	R56477	THC911668	
19840	Image:178818	H49455	THC882357	Apical-like protein (APXL protein). [Human] {Homo sapiens}
26573	Image:179276	H50323	THC862010	fatty-acid synthase (EC 2.3.1.85) (version 2) - human
26161	Image:810133	AA464251	THC950559	

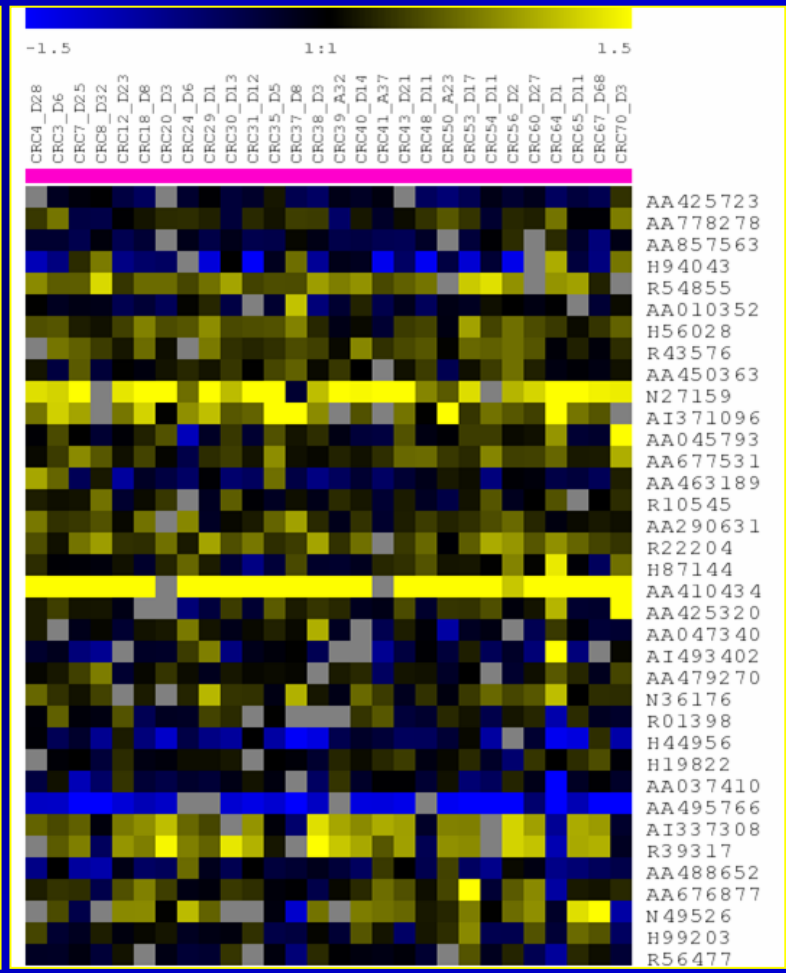
Red are positive and green are negative genes. A positive score means that higher expression is associated with higher risk, ie, shorter survival!!!

National Heart, Lung, and Blood Institute — PROGRAMS FOR GENOMIC APPLICATIONS

Ivana Yang, Steve Eschrich, Greg Bloom, Tim Yeatman

Clustering based on *k*-means support



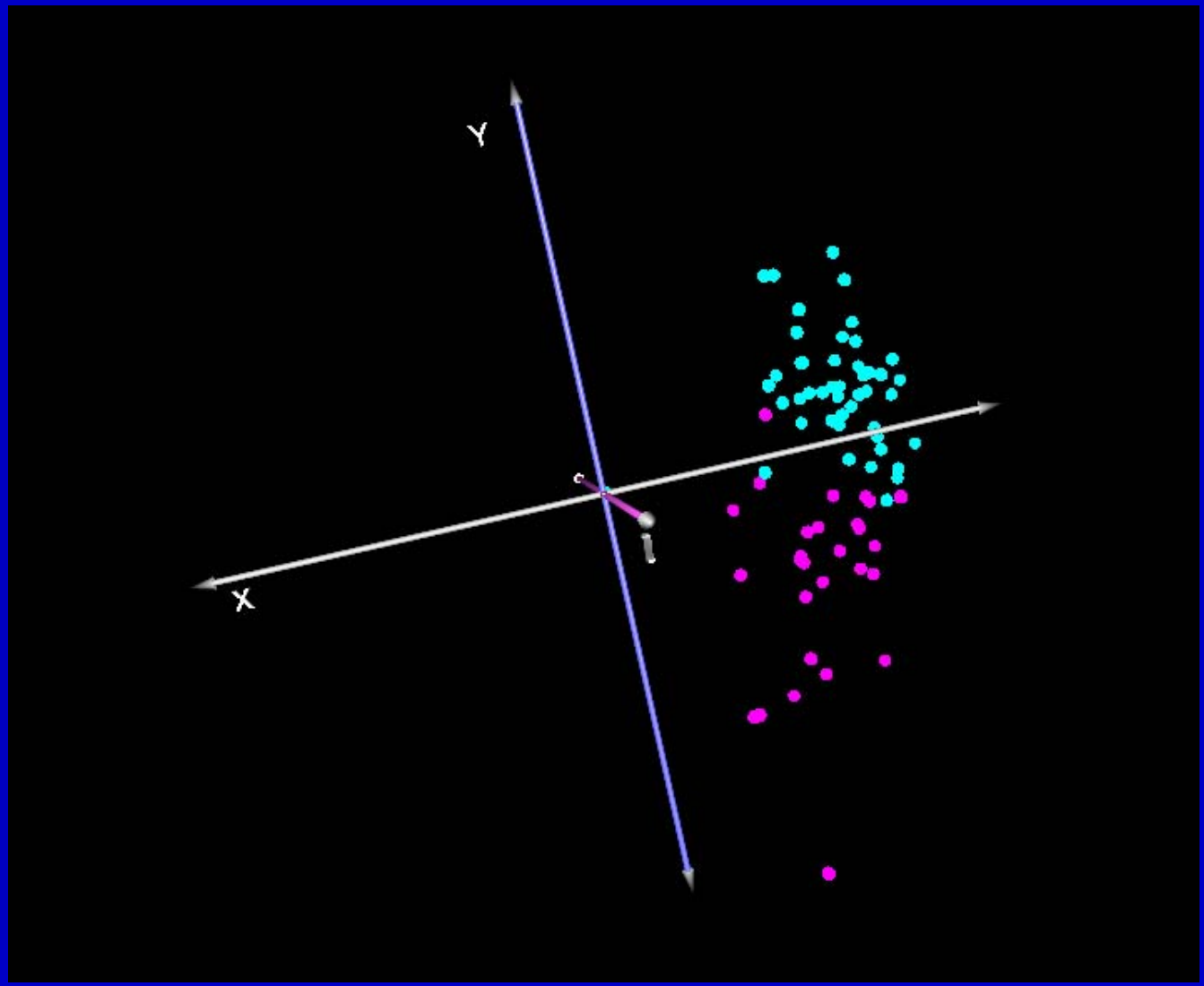


Good Prognosis

Poor Prognosis

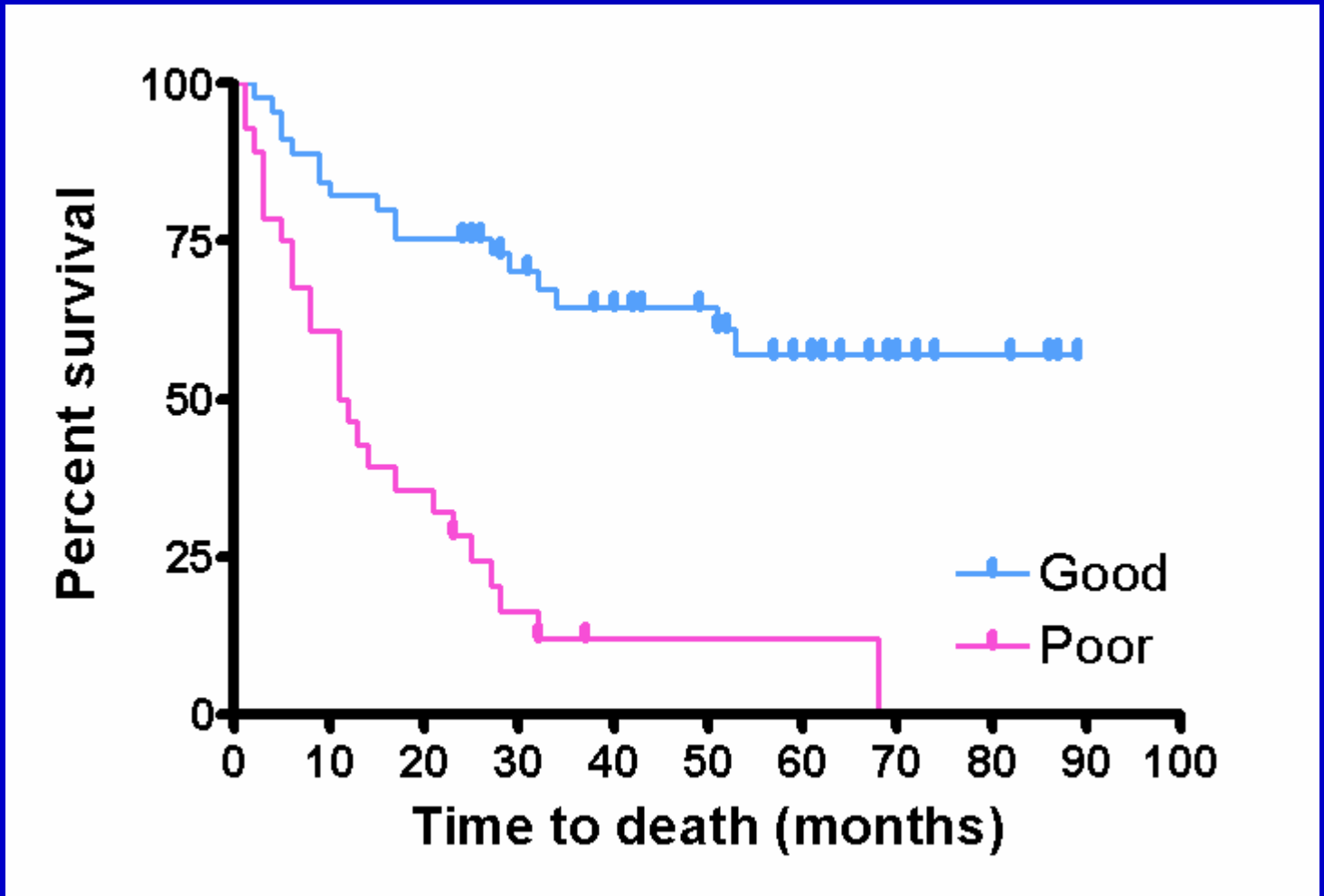
National Heart, Lung, and Blood Institute — PROGRAMS FOR GENOMIC APPLICATIONS

PCA analysis of survival samples



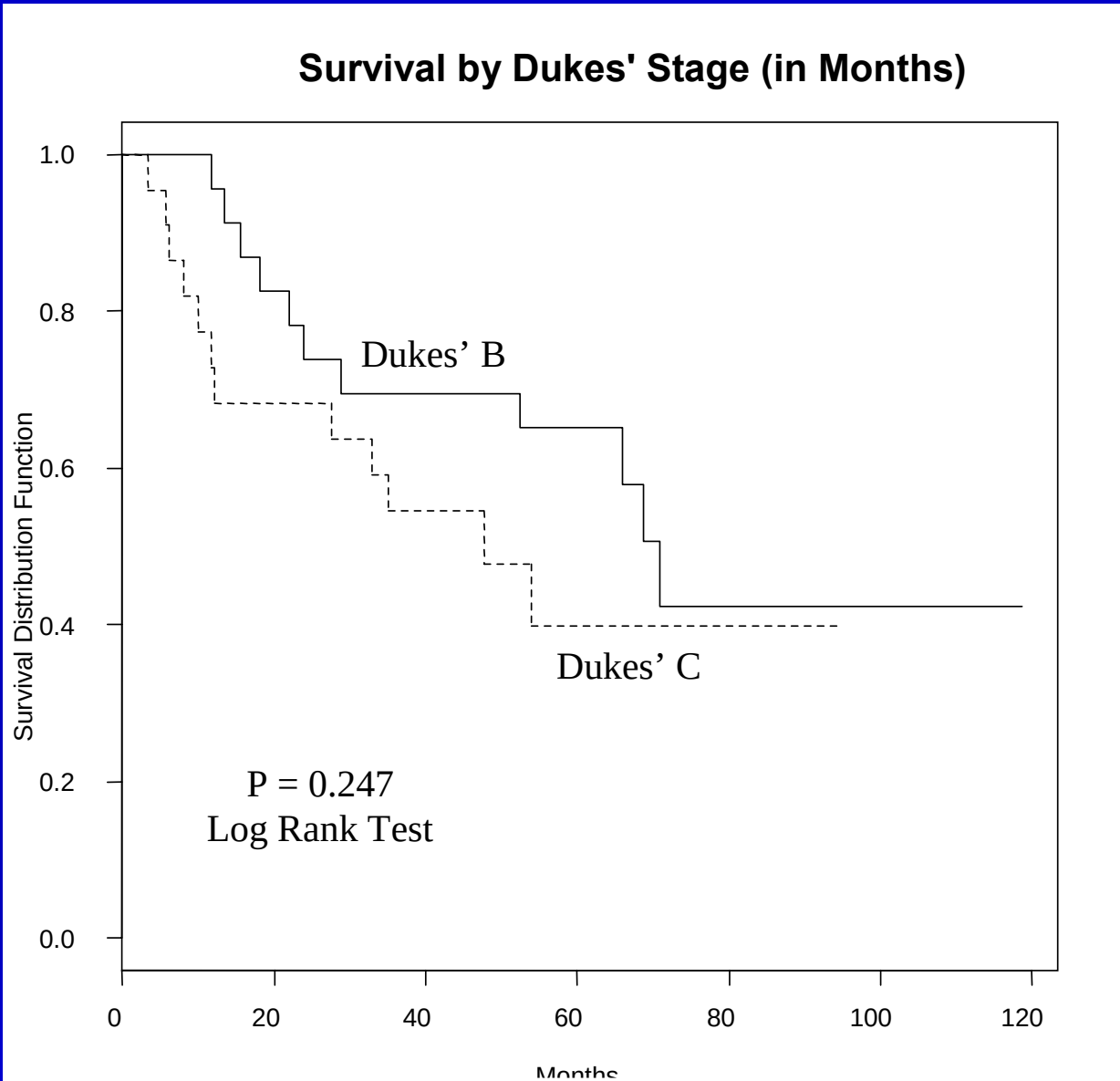
Ivana Yang, Steve Eschrich, Greg Bloom, Tim Yeatman

Survival Classes based on *k*-means clustering



Ivana Yang, Steve Eschrich, Greg Bloom, Tim Yeatman

Survival based on Clinical Stage



Steve Eschrich, Ivana Yang, Greg Bloom, Tim Yeatman

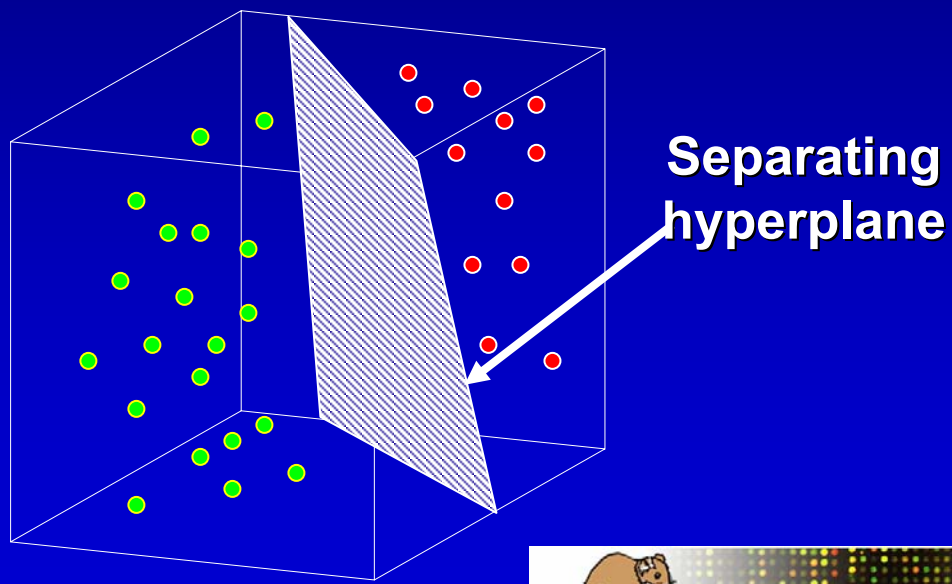
The Challenge

- Supervised gene selection followed by unsupervised clustering works quite well
- We anticipate that supervised approaches would perform better
- Use *complete* leave one out cross-validation to classify samples and to identify a core set of genes for classification
  - Iteratively leave out one sample and redo gene selection, algorithm training, and testing

Ivana Yang, Steven Eschrich, Greg Bloom, Tim Yeatman

SVM Classification

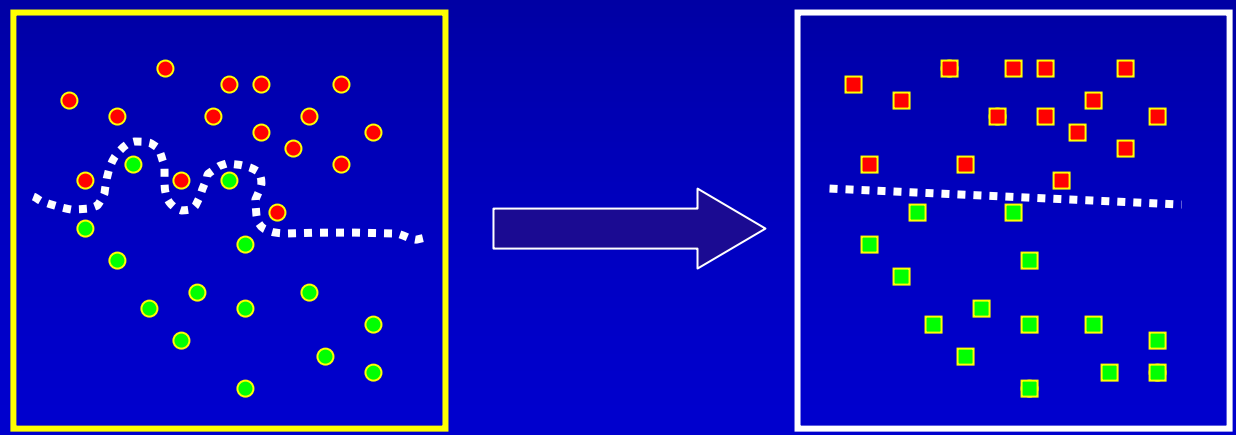
- SVM attempts to find an optimal separating hyperplane between members of the two initial classifications.



SVM Kernel Construction

The expression data can be transformed to a higher dimensional space (feature space) by applying a kernel function.

This transformation can have the effect of allowing a "separating hyperplane" to be found.



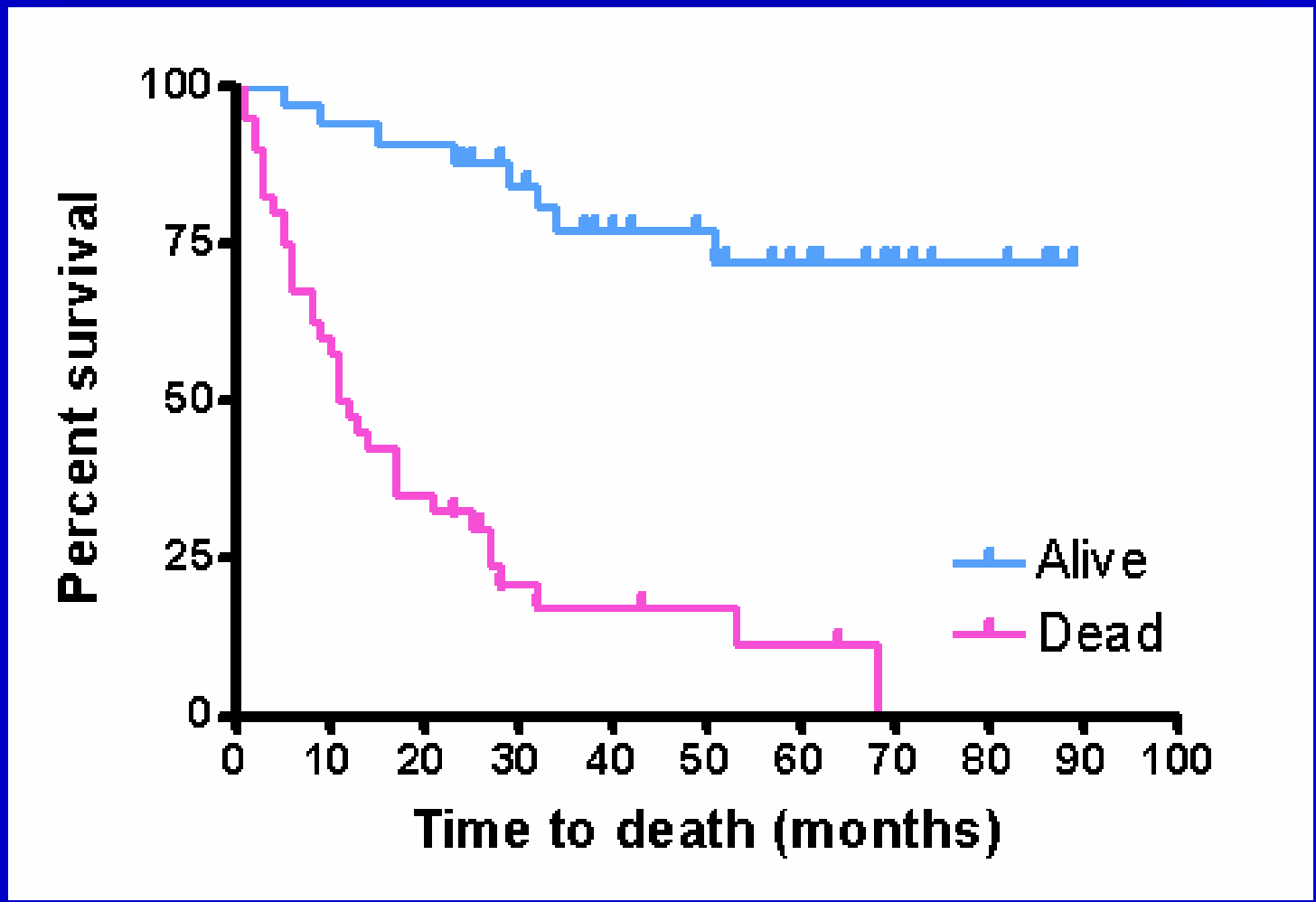
Consensus genes

GB#	TIGR TC#	Putative Role
R01398	THC1024446	unknown {Homo sapiens}
H44956	THC863066	Fumarylacetoacetase (EC 3.7.1.2) (Fumarylacetoacetate hydrolase) (Beta-diketonase) (FAA).
H19822	THC986991	hypothetical protein {Homo sapiens}
AA037410	THC864588	rho GTPase activating protein 8 isoform 1 {Homo sapiens}
AA495766	THC925935	RCC1-like G exchanging factor RLG [imported] - human
AI337308	THC987466	protein-tyrosine kinase EPHB2v {Homo sapiens}
R39317	THC987467	putative {Mus musculus}
AA488652	THC862216	ribosomal protein L2 {Homo sapiens}
R56477	THC911668	
N49526	THC1031960	Myb proto-oncogene protein (C-myb). [Human] {Homo sapiens}
H99203	THC968032	Ubiquitin carboxyl-terminal hydrolase 7 (EC 3.1.2.15) (Ubiquitin thiolesterase 7)
AA401111	THC897309	Glucose-6-phosphate isomerase (EC 5.3.1.9) (GPI) (Phosphoglucose isomerase) (PGI)
AA448641	THC889589	transcription factor {Homo sapiens}

National Heart, Lung, and Blood Institute PROGRAMS FOR GENOMIC APPLICATIONS

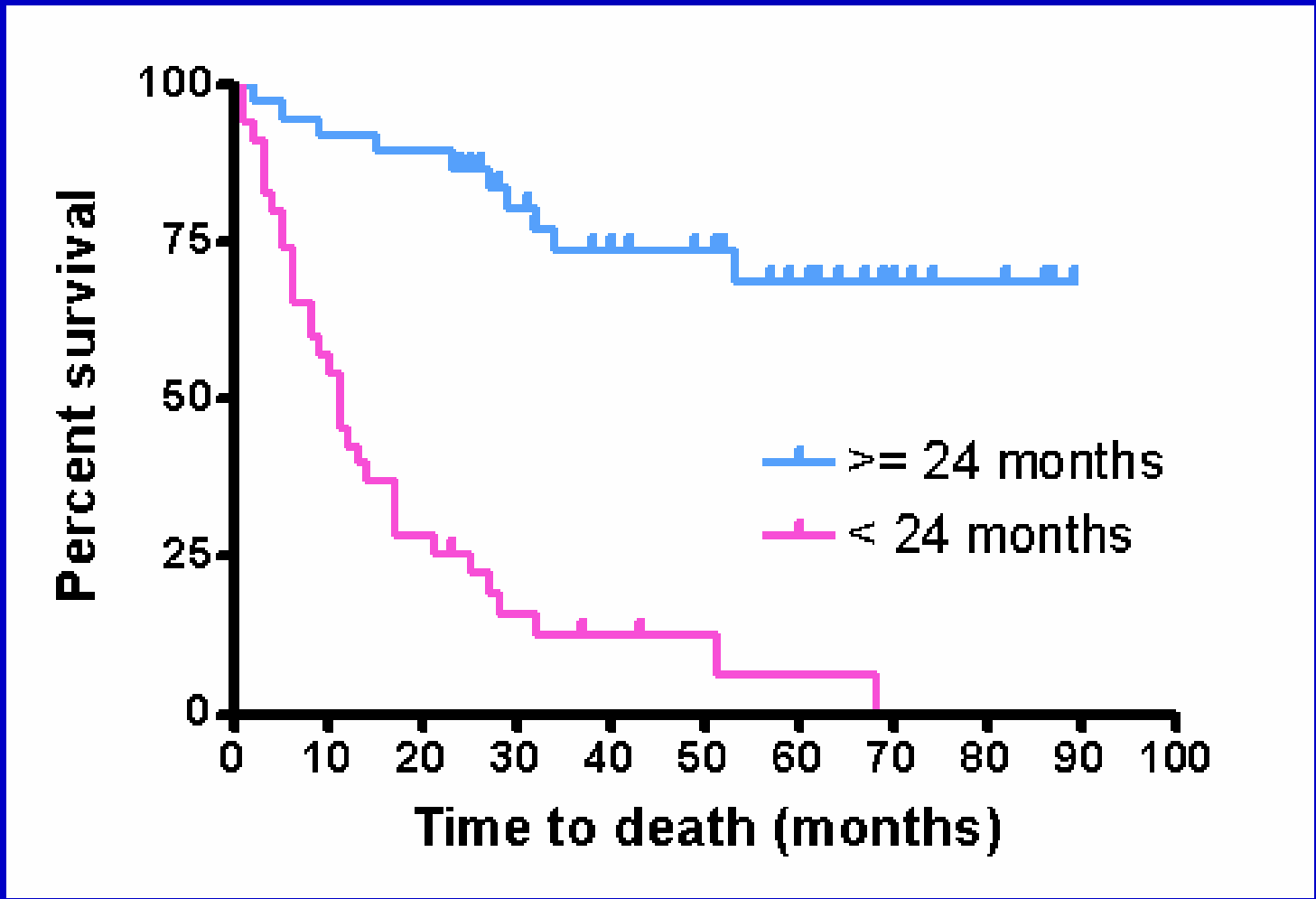
Ivana Yang, Steven Eschrich, Greg Bloom, Tim Yeatman

Survival Classes based on SVM – alive/dead



Ivana Yang, Steven Eschrich, Greg Bloom, Tim Yeatman

Survival Classes based on SVM – Two year cutoff



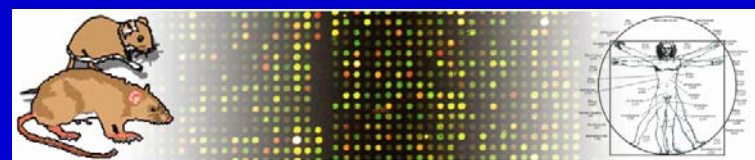
Ivana Yang, Steven Eschrich, Greg Bloom, Tim Yeatman

What have we learned?

- Microarray fingerprints can provide clinically important clues about disease progression
- Many of these require additional validation
- Many key genes are not obviously related to the underlying biology
- More data are needed to validate these findings
  - This approach and the classification gene set have since been validated using an *independent* test set and an Affymetrix GeneChip™ data set

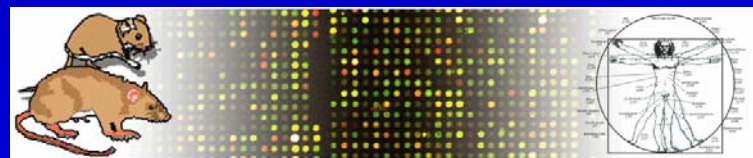
Where are we going?

- There is still a role for biology!
- We are approaching a time in which we can begin to look at cells and organisms holistically.
- We also need to begin to think about integrating diverse data types in an intelligent way.
- This must include cross-species comparisons and inclusion of environmental effects.
- We may soon be in a position to begin development of a theoretical biology.
- Theoretical biology will require a transition from a Deterministic to a Stochastic approach.



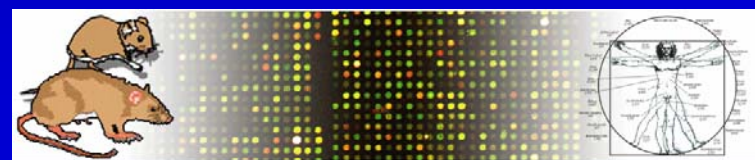
A theory has only the possibility of being right or wrong. A model has a third possibility; it may be right but irrelevant.

– Manfred Eigen



Nobody in the game of football
should be called a genius.
A genius is somebody like Norman Einstein.

- Joe Theisman,
Former Washington Redskins quarterback



Acknowledgments

<johnq@tigr.org>

The TIGR Gene Index Team

Foo Cheung
Svetlana Karamycheva
Yudan Lee
Babak Parvizi
Geo Pertea
John Quackenbush
Razvan Sultana
Jennifer Tsai
Joseph White

Emeritus

Jennifer Cho (TGI)
Emily Chen (μA)
Ingeborg Holt (TGI)
Feng Liang (TGI)
Kristie Abernathy (μA)
Sonia Dharap (μA)
Julie Earle-Hughes (μA)
Cheryl Gay (μA)
Jeremy Hasseman (μA)
Priti Hegde (μA)
Heenam Kim (μA)
Lara Linford (μA)
Rong Qi (μA)
Erik Snesrud (μA)
Shuibang Want (μA)
Ivana Yang (μA)
Yan Yu (μA)

H. Lee Moffitt Center/USF

Timothy J. Yeatman
Greg Bloom

PGA Collaborators

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

TIGR PGA Collaborators

Norman Lee
Renae Malek
Hong-Ying Wang
Truong Luu
Bobby Behbahani

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

TIGR Human/Mouse/Arabidopsis Expression Team

Adriana Ahumada
Tove Andersson
Joanne Emerson
Bryan Frank
Molly Freeman
Renee Gaspard
Nadeeza Ishmael
Ka Yin (Simon) Kwong
Jennie Larkin
Fenglong Liu
John Quackenbush
Yonghong Wang
Yan Yu

Array Software Hit Team

Nirmal Bhagabati
John Braisted
Tracey Currier
Jerry Li
Wei Liang
John Quackenbush
Alexander I. Saeed
Vasily Sharov
Mathangi Thiagarajan
Joseph White

Assistant

Aseye Aboagye