



# **Extracting Meaningful Information from Microarray Data**

**John Quackenbush**  
**NIDDK/NHLBI Microarray Workshop**  
**22 January 2003**

# The “Golden Age of Microarrays”

- **Print an array**
- **Get RNA samples**
- **Hybridize the samples to the array**
- **Find a long list of differences (Table 1)**
- **Select a small subset of “interesting genes” (Table 2)**
- **Write about those differences**
- **Submit paper to Nature/Science/PNAS and prepare for press conference**

**With apologies to Roger Bumgarner**

# Microarrays Today

- **Choose an Experimental System**
- **Design an Experiment**
- **Collect RNA samples**
- **Do the hybridizations, with replication, both biological and technical**
- **Collect, normalize, filter, transform, and analyze the data to identify significant differences**
- **Mine the literature and use other available information to form hypotheses based on the data**
- **Test those hypotheses**
- **Publish if you can**

# What's the difference

- **The cost of the assays have fallen and they have become more robust**
- **Our ability to generate data has increased dramatically**
- **Our sophistication in both experimental design and data analysis have evolved significantly**
- **The expectations from the community for data analysis and validation have increased**
- **The challenge of how one can extract meaning from a list of experimentally significant genes remains**

# Levels of Biological Information

'omics

DNA

Genomics

mRNA

Functional Genomics

Proteins

Proteomics

Informational Pathways

Metabolomics

Information

Systems Biology

Cells

**The Future!**

Cellular Biology

Organs

Medicine

Individuals

Medicine

Populations

Genetics

Ecologies

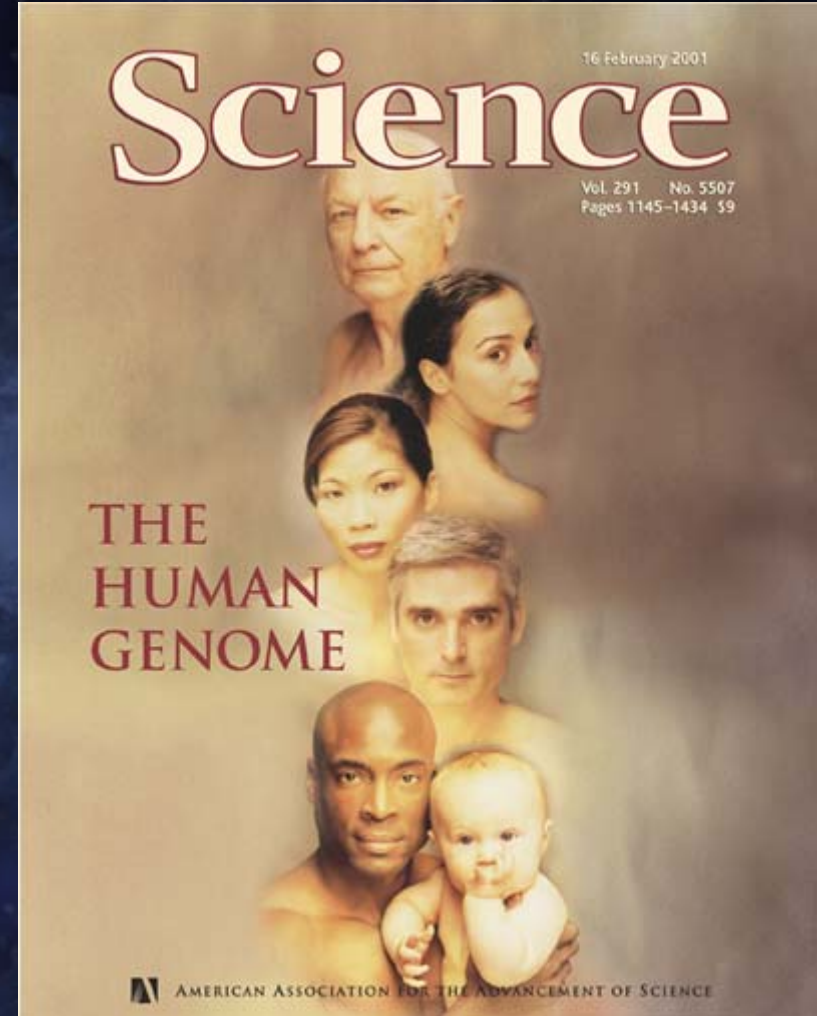
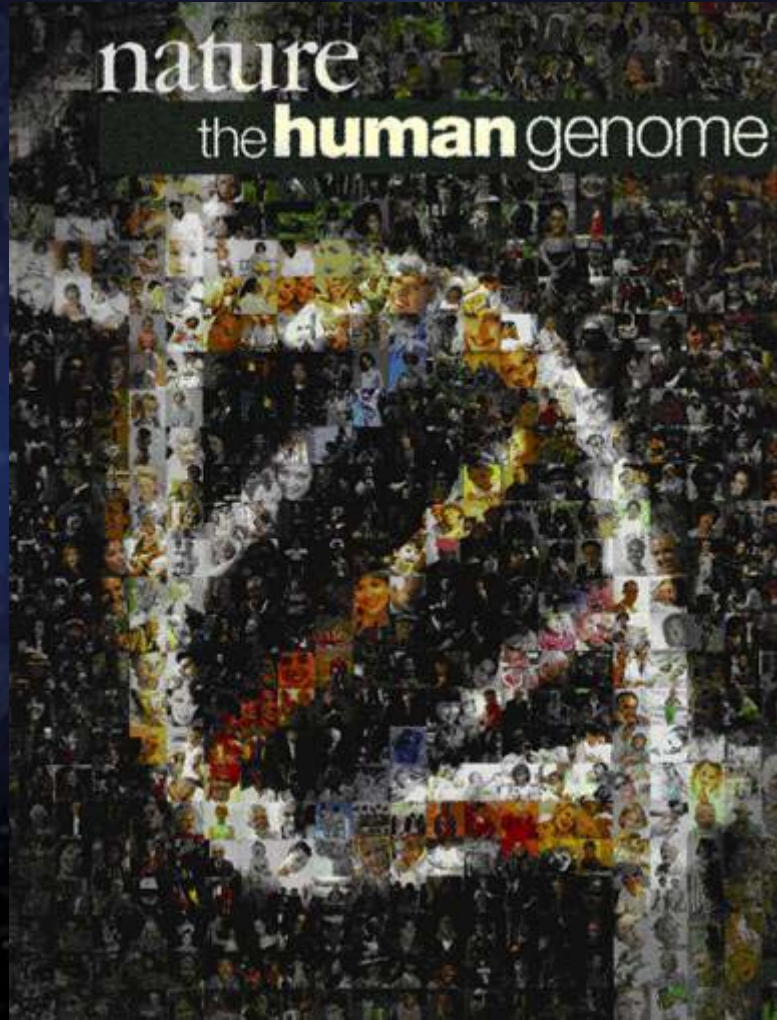
Ecology

Traditional  
Biology

# The challenges today

- **How do we best design and analyze the experiments to identify the most significant set of candidate genes?**
- **How can we leverage the existing biological knowledge base to extract information about the patterns of gene expression we see?**
- **Can we link expression data to the genome, to genetic and QTL maps, and to other related resources?**
- **Can we reconstruct metabolic and signaling pathways and networks?**
- **Can we use arrays to make clinically relevant predictions?**

# February 2001: Completion of the Draft Human Genome



**TIGR** Public HGP  
THE INSTITUTE FOR GENOMIC RESEARCH

**Celera Genomics**

**But what does *finished* mean???**



# How do we use sequence data?

 TIGR

THE INSTITUTE FOR GENOMIC RESEARCH



# The Golden Age of Genomics

- ~100 Microbial Genomes have been sequenced, at least 100 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 and Mouse Genome Sequences have been completed and the first Rat assembly is now available
- More than 13,000,000 Expressed Sequence Tags (ESTs) are available; more than 5,500,000 from humans

# TIGR Gene Indices home page

[www.tigr.org/tdb/tgi.shtml](http://www.tigr.org/tdb/tgi.shtml)

>60 species

>12,000,000 sequences

VERLAG ZU NÜRNBERG DRUCKERBEIHE (JANUARI 1. 2001)

## Nucleic Acids Research

**NAR ONLINE**

**TIGR THE INSTITUTE FOR GENOMIC RESEARCH**

**Gene Indices**

OXFORD UNIVERSITY PRESS

**TIGR THE INSTITUTE FOR GENOMIC RESEARCH**

Home > Databases > Gene Indices

Comprehensive Microbial Resource  
Unfinished Microbial Genomes  
Eukaryotic Resources  
Gene Indices  
Parasites Databases  
TIGRFAMs  
Fungal Databases  
Human Sequencing Projects

### TIGR Gene Indices

What's New | BLAST Search | TGI Software | FAQ

Integrating data from international EST sequencing, genome sequencing and gene research projects, Gene Index is an analysis of transcribed sequences represented in the world's public EST data.

**EGO** - linking orthologous genes across eukaryotic organisms  
**Genomic Maps** - mapping TIGR sequences to eukaryotic genomes  
**RESOURCECENTER** - virus resequencing, mammalian microarray resources  
**DAS** - providing distributed annotations for completed genomes

#### Animal Gene Indices


#### Plant Gene Indices


#### Ferret Gene Indices


#### Fungal Gene Indices


**Cattle 1.0** — Most recent version number  
— Date of most recent update

The TIGR Gene Index Project is supported in part by funding from the US Department of Energy, Grant #DE-FG02-95ER22722, and the US National Science Foundation, Grant #IBN-9812035. Additional funds are provided by the US National Science Foundation through grants #IBN-9812035 and #IBN-9973566.

The TIGR Gene Indices are built using:  
 • **seqalign** (Chang, Chang, Stein Schwartz, Lukar Wagner, and Wirth Mide (2000), "A greedy algorithm for aligning DNA sequences"),  
 • **EMAP** (Chen, 2000, 11, 2, 263-14),  
 • **CAPI**, developed by Dr. Zhaoyu Huang,  
 • **Parasit Transcription (ParasitT)**, from ParasitLar,  
 • **BINA Protein Search program (BPS)** developed by Dr. Zhaoyu Huang

The ORF annotation of TGI is done using:  
 • **ESTScan** (Jin, C. Engstrom, CV, and Buchler, P. (1999) ESTScan: A program for detecting, evaluating, and re-annotating potential coding regions in EST sequences. *BMC* 1: 13-24),  
 • **BIANNA EST** (Huangping AO, Fuzuo F, Yuchun M. *Related Article* DEANA-EST: a statistical analysis. *Bioinformatics* 2001 Oct 17;17(13):15-3),  
 • **GeneIndex** (Expressed Sequence Tag Analysis Tools Set) (© Copyright 1999-2000 Human Genome Mapping Project, BC, Montreal, Canada, QC)

The Orfmap mappings are done using:  
 • **Map2** (Wirth Mide), see reference  
 • **SeqMap**, from ParasitLar,  
 • **gag2** developed by Dr. Zhaoyu Huang

The expression profiles of the ESTs are scored using:  
 • **W** (Wirth Mide, Chih, Oh, and Falciano (2000) The comparison of gene expression from multiple cDNA libraries. *Oncogene Research* 10:2053-2061)

The Institute | ©1999-2002 The Institute for Genomic Research

# Gene Index Assembly process

**ESTs from GenBank (dbEST)**

**TIGR ESTs**

**Expressed Transcripts (ET) from GenBank CDS**

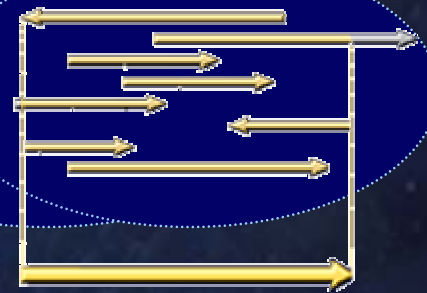
remove vector, poly-A, adapter, mitochondrial and ribosomal sequence

reduce redundancy

High stringency pair-wise comparisons to build **Clusters**

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

**Annotate TCs and release**



The TIGR Human Gene Index (HGI)

HGI THC Report: THC104722

EST IDs are linked to HGI EST reports. HTPs are linked to [EGAD](#) HT reports. GB#s are linked to [GenBank](#) accessions. ATCC#s are linked to order forms for [transposon clones](#).

TC104722

#	EST ID	GB#	ATCC#	left	right	library
1	F 580612.n1	AAL1548	314376	1	347	Lang, Oncology Research
2	F 580702.n1	302654	218082	4	394	Fetal tissue/epithel. INFIL, Bioma
3	A 587143Z	AAL4578	146947	8	145	Hell Bladder II
4	D 580571L18	557353		82	403	Actin polyA+, Clontech (#9372)

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



## TIGR Mouse Gene Index

### About the Index

- [Current Release - Version 4.0 Release Date - July 15, 2000](#)
- [Development and Goals](#)
- [Information About the Current Release](#)

### Search the Index by

- [Nucleotide or Protein Sequence](#)
- [Identifier \(TC, EF, EST, GB\)](#)
- [Transcript, cDNA Library Name, or cDNA Library Identifier\(s\)](#)
- [Gene Product Name \(Example: insulin\)](#)
- [Search by Radiation Hybrid Map Location \(coming soon\)](#)

### Data Availability

- MGI 3.0 is available free of charge only to researchers at non-profit institutions using it for non-commercial purposes. Please go to our [licensing agreement](#) and follow the instructions there to obtain the MGI data file.
- Please Note: Users of previous releases of the MGI can use TC# from that release to find the corresponding TC# in the current release by using the [TC# search function](#).
- If you represent a for-profit organization, please [contact us by email](#) for details on how to obtain a commercial license for any of the data files described below.
- Please read the [copyright notice](#) governing use of this data.
- Data
  - A fasta file containing the complete, minimally redundant Mouse Index.
  - A fasta file containing the complete set of TC sequences in the Index with previous TC identities in the definition line.
  - A file containing the TC IDs and the ESTs that comprise them.

### Help

- [Data Definitions and Protocols](#)
- [Frequently Asked Questions page for the Gene Indices](#)
- [NAR, 2000 paper: 'The TIGR Gene Indices'](#)
- Send mail to [www@tigr.org](mailto:www@tigr.org) for WWW specific Comments/Questions.
- Send mail to [mgi@tigr.org](mailto:mgi@tigr.org) for MGI Comments/Questions.

### Attributions

A significant number of ESTs used to construct this index were generated by:



[Washington University School of Medicine, Genome Sequencing Center.](#)



[The National Institute on Aging \(NIA\).](#)



[RIKEN \(The Institute of Physical and Chemical Research\).](#)



Last modified on: August 11, 2000

# A TC Example

>TC161360 TC25195 TC29362 TC33731 TC40754 TC149101  
TGAGGCTCACAAAGAACTTTTATTCCTTTTTAAATAGACACTAAAAATTATCTCCTAGTCATGAGAAAATTGGTAAAGACTAAT  
TATTTGAGAATCTGACGATGACTAATGTAATAATCATTAAAGGAAATGAATTTTCAGAGAGGGGAAAACCTTTTCAAATTTGAATA  
CTGCATTTAAAACCTTTTCAGCTTGACACTCCTCCTCCACCTCCCCATCCTCTCCAGGCATAGCGGTATCTTCTTTAGCT  
TAGGGTACCTTCTATGGAGAAGAAATGGATATGGAGAATCGTGCTGTGGCTTGTAAAGTGGGCAGAACTTAGTAAAGACCTA  
CTGGATGAGTAACTCCTTGGGAGCATGTGTCCAGATAGGTAGGAAATAGCTCAATATGACTGGATGTGCCACTATTCAAAC  
ACAGGTTAGTATTATGTGGCAGAACTCCCATTTGTTTGTATTTAGGAGAAGAAAAGAAACATCCAAGGTGGAGTATCCA  
TTGCAGGCCTGACACAAAAGTTTTATTTACTTAGAGCTTGTTTTTGAAGACCACACAGGGGAAAAGGTGCTACTTCCAGTTT  
CTTTGTAATAACAGGAAAATACTCCCACCGGTAGCCTCTAATAAAAATAGAAATATTCGAAGGAGTGAACCTTAAAGCTGTT  
CATATACCCATAATGCCTAGAAGCAGACTTGTCAATGGTATCTGAAATGATAGGCTATGGTGAGATCTTTTTAGGCTAACAG  
TGTTTTCTAGGTCAGGTGCTAGCATCCCTGCTCAGGAACAGGGGTGGGAAAAGTATGGTGGCCCTGAGATTTAGGATTTTAAAC  
TGTGTTGTTTTTAAAGCATGATCTTTTGTGTGGTAAATTTATAGTGCATATAAAGATGTGTTTTGTGGTGCATCTATAACTTTT  
CAGCTAATTTGCATATTAATGTCACGACTAGTTTTCCAAATGATGTAAGATTCTGGGTGCTTTTATTCATATGGTGTCAA  
TCCAATTCGACTCTTTTGTGATTGAGCATTGCTCACAAAAGTATATACTTTGATATAGCTTATACAGGCATATGGGCA  
TAGATAATTTGGTTATTTCAAACATTTCTAGAATTTGAAGAGCTGGGTTTGAAGTCCCACTTATTATTGAATA  
TCTTGGCTTGCTTCTTCACTTCACTAAATTAACAAAATAGTATAGGTTTTTTGTTTTTTTGGTTTTTTTGGTTTTTTTTCG  
TGAATCATGAAAAGACCAATAAATTAACCTACTCTGCCATTTGGAAGAATCTCAAAGCACAGCTGTGCTTAGGATTTGAGATC  
TGGGAACCCATTACATTTCTGTCTCACCATTCTTTTTCTCTTGACTAAAAGAATACAAATACATAAACGATGTGACCAGA  
AGCCAAGAACTGGAGATGGGCAAGTTAAGACATGAACTTGTCTTATGTGTAAGCTATGCTTTTTGTACAGAGACAGAC  
TTAGGACTAGGCCCTTTCACAGCTTCCCAAGTGGGACACAGAAGCTTGAAGAACATGCCATATTTTGTGCCCTTCCACAG  
GCATATAGGTGCTCATTTTCTCTCATTATTAACAATTTCTCATTAAACAATTTCTCATATTTCTTAAAGACAATCAT  
GCAAAAGGGACTCCACCAACATGAAGAGGAGTCAGAGAGTTTCAGAGATGAACAGAAGCAGTCATGTGCTTTTTCCCTCT  
GCTGTAAGTCTGATGGAGGAAATGCGGAAAAACCATTACCCTACTGAGACACACATCCAAGGAAAAGGTGAATGAGAA  
TTGGATTCTTCAGGTTTTCTGCCAGTGTAAAGACTCCCTAAGCAAGGCAAAAGACAATCTTTTATAATCACTGCATTCTC  
AAATGTGAAGGAGAGCTAGATGTGGCTTCTCTATGCAAAAGTTAAGCTGTGAAAAATATGGAGAATAGATTGTGGAAAGGCCA  
CAAGAGATGAGGGTAACTATGTGCTTGGAAAGCTCTCTAACAATAATGCTTAGAAAAAGAAAAGAGTTTAAAGGAGGCATC  
TCATTTTCCACTGGCAAGGATTCCTGCTCTGGGGAGGTGCTGGAAAAATTTATTTCTGCTTCTTCTTCTTCTTCTTCTCCTG  
CACTGGAATAAATTTCAATTTCTCTTCTTCTTCTTGGTTGGGAAGATACTCCACTTAGCTCAATGATTTCTTCTTCTTCTT  
TAATCGTCTGCTCATTTTTTCTCCAGCTGACAGCAGAACCACATTAGATTGGATCTGGTACACATTTCTTCCACTCA  
TTCTCCACAATACCAGCTGTCACAAGTTTCTGGAAAGAGGCAGAGATCAGCATCGCCGACAGAAATGCCCAAGAACACAGA  
CTGAAATGCTGTTACAGTACTGTGTAGATGGGGAGTTTTCTCTGAGGAATTAGATGGTTACAGTCGATAGATATCAACAT  
ACGGCTTGAAGTTTGAATAAGCTCCAGTCTTCTCATTTTTTAAAAAATGAGAAAAGTGTGATGTTAAGTATGTCCATGATT  
GAAAGAATTTCCAGAAATGGCAGCAAAGAGGCTTAGAGAGCTCATTATCACTTTTGCTTTGACCAAACTCTTCTGGA  
GGTTTTTCTGCTGACGCTGCCAGGAGTATCCTGAAAATAATGTACATAATGCCCTCCCAAGAGAGGGTACCATACACTCA  
AACAGATGGGTGGCAAGACCCCTGTGGGGATCATCAGCAGTCCCCCAAGGGTAAATATGGAAGAGGCCATTATGATTTGG  
ACAGCCCCAAAGCCTTTGATTCCTCATGAAGAAGCTTTGTGTGGGGCCACCAGTGAAGATGTCTTTTGGAGGTTAC  
TTTTGGAGCAGGTTGATGGCAGGGGACCTTTTGTAGCCTCTGCTGGGAAAAGTCCACTCATTGAGTTTTCAAACGCC  
GGTCTCCAAGGCTCCAAGGCTCCAAGGCTTAAAGGGCCAAATTAGCCACTCACAGCTGAGATCAGTTGGCCCTTCTTATC  
TCAAGTACTTGAAGATAGTACAGTGTACTTACATATAATAATAATAAATCTTTAAAAAAGGAAATTC

-cell differentiation antigen

(B-LYMPHOCYTE

(Homo  
1|X07203 CD20 receptor

Mouse Genome Database

The screenshot shows the TIGR BLAST search interface. At the top, it says "The TIGR Mouse Gene Index (TGI)" and "BLAST Report: TC149101". Below this, there's a table of search results. The first result is for "CD20 receptor" (Gene ID: 100000000) with a score of 1000.0 and an E-value of 0.0. The alignment shows a perfect match between the query sequence and the database sequence. The TIGR logo and "THE INSTITUTE FOR GENOMIC RESEARCH" are visible at the bottom left of the screenshot.

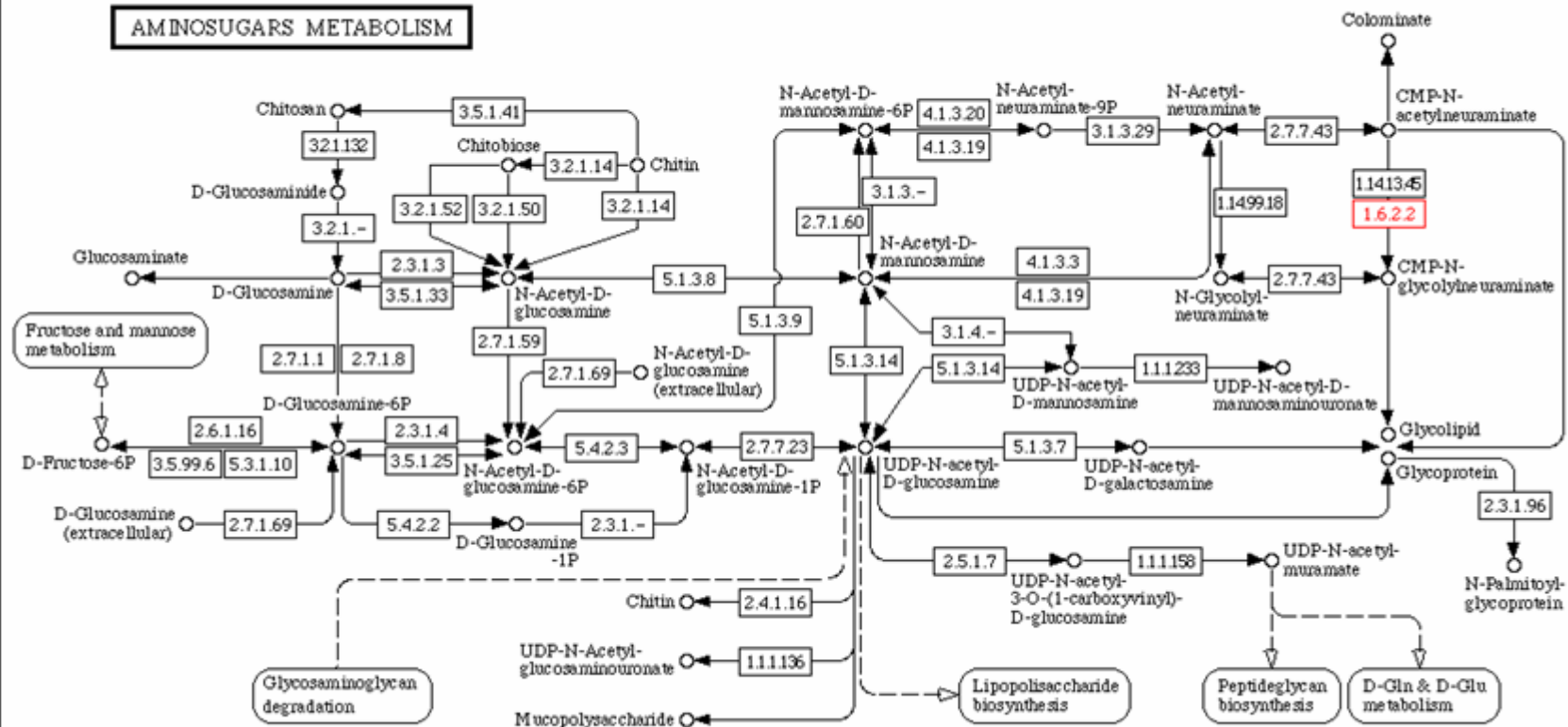
[Click here](#) to see the live web page at TIGR

# GO Terms and EC Numbers



## Position of term GO:0004128

Function	GO Assignments % of 12819 total TC/NP with GO Assignments
	31.21%
	2.52%
	1.92%



00530 10/5/01

- (I) enzyme (GO:0003024) +
- (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

last modified on: September 05, 2001

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

**TIGR THE INSTITUTE FOR GENOMIC RESEARCH**

Home | Database | Gene Indices | News | What's New | Search | About TIGR | Contact | Site Map | Home | TIGR

**TIGR Gene Indices**

What's New | BLAST Search | TGI Software | FAQ

Integrating data from international EST sequencing, genome sequencing and gene research projects, Gene indices are an analysis of transcribed sequences represented in the world's public EST data.

EGO - linking orthologous genes across eukaryotes

Genomic Maps - mapping TC sequences to eukaryotic genomes

RESOURCEER - cross referencing transcribed sequences

DAS - providing distributed annotations for eukaryotic genomes

**Animal Gene Indices**

Arabidopsis thaliana 10	Bovine 20	Catfish 11	Cattle 70
Canary 10	Chicken 10	Canary 10	Drosophila 70
House fly 10	Human 10	Monkey 10	Mouse 10
Octopus 10	Orchid 10	Tig 10	Tat 10
Schistosoma 10	Xenopus 10	Zebrafish 10	

**Plant Gene Indices**

Arabidopsis 10	Baker 10	Chlamydomonas 10	Cotton 10
Ice Plant 10	L. japonica 10	Malt 10	Melting 10
Peanut 10	Rice 10	Rye 10	Soybean 10
Soybean 10	Tobacco 10	Wheat 10	

**Fungal Gene Indices**

Cryptosporidium 10	Intestine 10	Stoma 10	Leishmania 10
Hemipha 10	Plasmodium 10	Plasmodium 10	Plasmodium 10
Caricotea 10	Thelazia 10	Trypanosoma 10	Trypanosoma 10
Trichomonas 10	Trichomonas 10		

**Fungal Gene Indices**

Aspergillus 10	Coccidioides 10	Oglossina 10	Magnaporthe 10
Microspora 10	Schistosoma 10	Schistosoma 10	

← Most recent version number  
← Date of most recent update

The TIGR Gene Indices Project is supported in part by funding from the US Department of Energy, Grant #DE-FG02-95ER22010, and the US National Science Foundation, Grant #IBN-9537070. Additional funds are provided by the US National Science Foundation through grants #IBN-9537070 and #IBN-9537070.

The TIGR Gene Indices are built using:  
 - **megalign** (Zheng Zhang, Guohui Sheng, Luke Wagner, and Walsh Miller (2003), "A greedy algorithm for aligning DNA sequences"),  
 - **ClustalW** (Thompson, D.J., Gibson, T.J., and Plewniak, F. (1994) *BIOSIS*: A program for detecting, embedding, and reconstructing conserved coding regions in EST sequences  
 ISMB'94, 133-142).  
 - **BLAST** (Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) *Journal of Molecular Biology* 215:403-410).  
 - **BLAST** (Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) *Journal of Molecular Biology* 215:403-410).  
 - **BLAST** (Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) *Journal of Molecular Biology* 215:403-410).  
 - **BLAST** (Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) *Journal of Molecular Biology* 215:403-410).

The TIGR Gene Indices are also using:  
 - **BLAST** (Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) *Journal of Molecular Biology* 215:403-410).  
 - **BLAST** (Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) *Journal of Molecular Biology* 215:403-410).  
 - **BLAST** (Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) *Journal of Molecular Biology* 215:403-410).  
 - **BLAST** (Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) *Journal of Molecular Biology* 215:403-410).

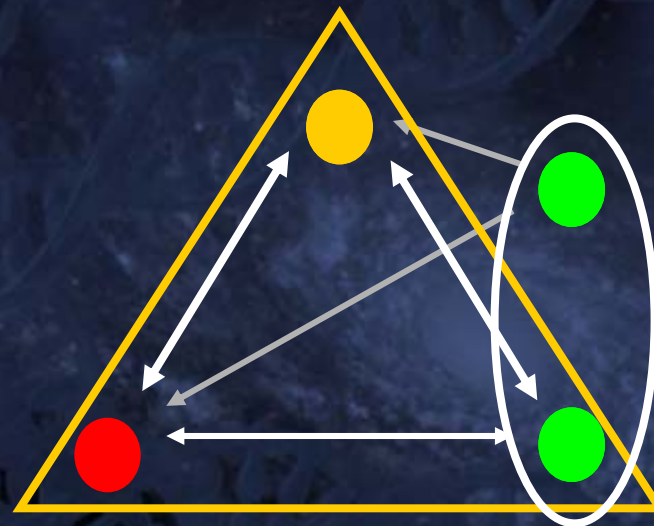
The expression profiles of the ESTs are used using:  
 - **W** (Wagner, L., and Plewniak, F. (2000) The comparison of gene expression from multiple cDNA libraries. *Genome Research* 10:355-363).

Site Information | ©1999-2002 The Institute for Genomic Research

## 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.

# Building TOGs: Reflexive, Transitive Closure



**And Paralogues**

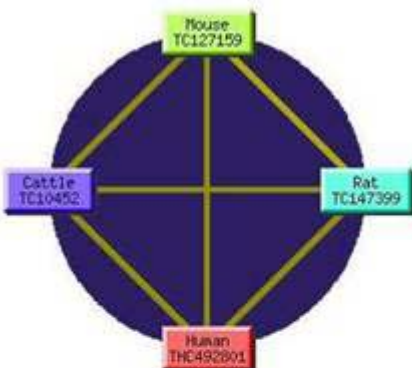
**Tentative Orthologues**



# TOGA: An Sample Alignment: bithoraxoid-like protein

## Tentative Ortholog 3220

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



Sequence 1	Sequence 2	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

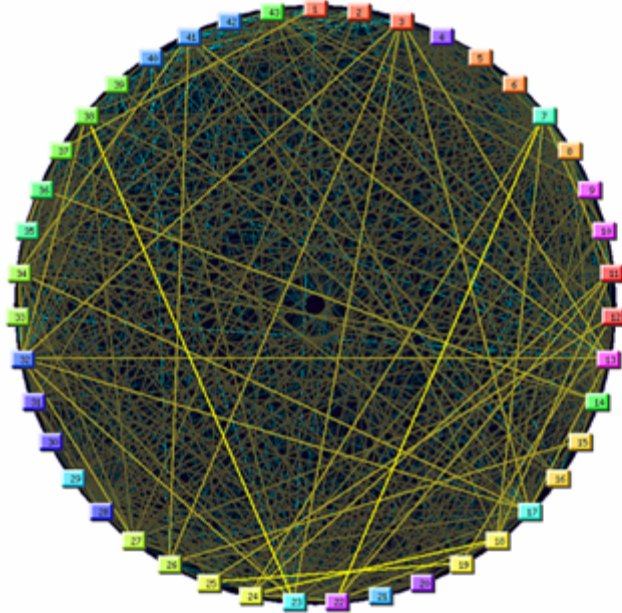
```

mouse| TC127159      TGGTCTACACAGGCTC-AG-GTGGCCACCACGTGC----CCACTGACATGATTAGCACTA
human| THC492801    TGGTGTGAGTGGGTTCCAA-GCGACTGCCATGTGCTAGTCCACTGACATGATTGACATTA
cattle| TC10452     CAGCCTGGGAGGGCTCCAAACGTGCC TTCACGTGCCCGTCAATGGACATGATTAACGCTA
rat| TC147399      AAGGTCTGCATGGCTCCAGGCAGCC--ACATGTGCC--ACTGACATGATTAACGCTA
                    *   ** *   * *   * * * * *   *   * * * * *   *   *

mouse| TC127159      TTATTCITGGGGGACATTAATAAAGGAATGACACAGGAAGCCAGACAGTGGCTTATTC
human| THC492801    ACATTCTITGGGGGGCATTAAATTAAGGAATGACACAGGGAGCC AAGAGAGTGGCTTATTC
cattle| TC10452     TTATTCITGAGGGGCATTAAATTAAGGAATGACGCAGGGAGCCAAGAAAGCAGCTTATTC
rat| TC147399      TTATTCITGGGGGACATTAATAAAGGAATGACACAGGAAGCC AAGACAGTGGCTTATTC
                    * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

mouse| TC127159      AGTTGGATTCTGGATCACAATCAGGAAATAGTCTTTATCTGGTGCCACCATAATTTTCATT
human| THC492801    GGTGGATTCTGAATCACAATCAGGAAATAGTCTTTATCTGGTGCAACCATAATTTTCATT
cattle| TC10452     AGTTGGATTCTGAATCACAATCAGGAAATAGTCTTTATCTGGTGCAACCATAATTTTCATT
rat| TC147399      AGTTGGATTCTGAATCACAATCAGGAAATAGTCTTTATCTGGTGCCACCATAATTTTCATT
                    * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
    
```

# Tentative Ortholog 14405



Species	Accession Number	Gene Name
Arabidopsis thaliana	NP221569	At1g07000
Arabidopsis thaliana	TC94474	At1g07000
Arabidopsis thaliana	TC41176	At1g07000
Arabidopsis thaliana	TC3799	At1g07000
Arabidopsis thaliana	TC3792	At1g07000
Arabidopsis thaliana	TC3058	At1g07000
Arabidopsis thaliana	TC54875	At1g07000
Arabidopsis thaliana	TC94877	At1g07000
Arabidopsis thaliana	TC1004	At1g07000
Arabidopsis thaliana	TC8510	At1g07000
Arabidopsis thaliana	TC37440	At1g07000
Arabidopsis thaliana	TC37442	At1g07000
Arabidopsis thaliana	TC30280	At1g07000
Arabidopsis thaliana	TC15932	At1g07000
Arabidopsis thaliana	TC180	At1g07000
Arabidopsis thaliana	TC69	At1g07000
Arabidopsis thaliana	TC146	At1g07000
Arabidopsis thaliana	TC122	At1g07000
Arabidopsis thaliana	TC43427	At1g07000
Arabidopsis thaliana	TC1402	At1g07000
Arabidopsis thaliana	TC15940	At1g07000
Arabidopsis thaliana	TC1499	At1g07000
Arabidopsis thaliana	TC92	At1g07000
Arabidopsis thaliana	TC2045	At1g07000
Arabidopsis thaliana	TC2046	At1g07000
Arabidopsis thaliana	THC480091	At1g07000
Arabidopsis thaliana	TC37908	At1g07000
Arabidopsis thaliana	TC13	At1g07000
Arabidopsis thaliana	TC925	At1g07000
Arabidopsis thaliana	TC499	At1g07000
Arabidopsis thaliana	TC37006	At1g07000
Arabidopsis thaliana	TC37007	At1g07000
Arabidopsis thaliana	TC94474	At1g07000
Arabidopsis thaliana	TC94473	At1g07000
Arabidopsis thaliana	TC41176	At1g07000
Arabidopsis thaliana	TC3799	At1g07000
Arabidopsis thaliana	TC3792	At1g07000
Arabidopsis thaliana	TC3058	At1g07000
Arabidopsis thaliana	TC54875	At1g07000
Arabidopsis thaliana	TC94877	At1g07000
Arabidopsis thaliana	TC1004	At1g07000
Arabidopsis thaliana	TC8510	At1g07000
Arabidopsis thaliana	TC37440	At1g07000
Arabidopsis thaliana	TC37442	At1g07000
Arabidopsis thaliana	TC30280	At1g07000
Arabidopsis thaliana	TC15932	At1g07000
Arabidopsis thaliana	TC180	At1g07000
Arabidopsis thaliana	TC69	At1g07000
Arabidopsis thaliana	TC146	At1g07000
Arabidopsis thaliana	TC122	At1g07000
Arabidopsis thaliana	TC43427	At1g07000
Arabidopsis thaliana	TC1402	At1g07000
Arabidopsis thaliana	TC15940	At1g07000
Arabidopsis thaliana	TC1499	At1g07000
Arabidopsis thaliana	TC92	At1g07000
Arabidopsis thaliana	TC2045	At1g07000
Arabidopsis thaliana	TC2046	At1g07000
Arabidopsis thaliana	THC480091	At1g07000
Arabidopsis thaliana	TC37908	At1g07000
Arabidopsis thaliana	TC13	At1g07000
Arabidopsis thaliana	TC925	At1g07000
Arabidopsis thaliana	TC499	At1g07000
Arabidopsis thaliana	TC37006	At1g07000
Arabidopsis thaliana	TC37007	At1g07000

Sequence 1	Sequence 2	% Identity	Match length	p-value	Recip. best hits
Arabidopsis NP221569	B. malay 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	Icelandic TC1004	87.00	1346	2.4e-236	*

```

arabid|NP221569      CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
arabid|TC94474      CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
arabid|TC94473      CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
mal1zei|TC41176     CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
arabid|TC3799      CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
arabid|TC3792      CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
arabid|TC3058      CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
komaat|TC54875     CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
komaat|TC94877     CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
icepln|TC1004      CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
medicago|TC8510   CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
arabid|TC37440     CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
arabid|TC37442     CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
mouse|TC30280      CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
kat|TC15932        CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
mouse|TC180        CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
k_drome|TC69       CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
m_zhuanian|TC146   CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
e_yeast|TC122      CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
drosophila|TC43427 CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
yeast|TC1402       CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
mouse|TC15940      CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
kat1zei|TC1499    CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
p191|TC92         CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
arabid|TC2045      CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
arabid|TC2046      CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
human|THC480091    CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
e_fly|TC37908      CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
b_bombus|TC13      CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
b_malay|TC925      CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
o_voivulua|TC499  CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
elegans|TC37006    CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC
elegans|TC37007    CCCTTCGTCCTCCCACTTCAGGATGTCACAAAGATTGGTGTTATTGGAAAGCGGTGCCAGTGC

```

```

arabid|NP221569      CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
arabid|TC94474      CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
arabid|TC94473      CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
mal1zei|TC41176     CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
arabid|TC3799      CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
arabid|TC3792      CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
arabid|TC3058      CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
komaat|TC54875     CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
komaat|TC94877     CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
icepln|TC1004      CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
medicago|TC8510   CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
arabid|TC37440     CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
arabid|TC37442     CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
mouse|TC30280      CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
kat|TC15932        CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
mouse|TC180        CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
k_drome|TC69       CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
m_zhuanian|TC146   CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
e_yeast|TC122      CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
drosophila|TC43427 CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
yeast|TC1402       CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
mouse|TC15940      CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
kat1zei|TC1499    CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
p191|TC92         CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
arabid|TC2045      CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
arabid|TC2046      CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
human|THC480091    CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
e_fly|TC37908      CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
b_bombus|TC13      CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
b_malay|TC925      CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
o_voivulua|TC499  CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
elegans|TC37006    CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG
elegans|TC37007    CTGTGAGGGACATGAGGGCAGCCTTTGCAAGTGGTGTATCAAGAGTTTGGACAAAGAAAGG

```

[Click button to view the alignment with Jalview](#)

# RESOURCERER

Jennifer Tsai



## RESOURCERER 4.0

What's New

Description

Data

Targets of Discovery

PGA Tools

Outreach

Contact List

Links



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.

- Genome mapping information for human and mouse sets. [NEW](#)
- Links to Affymetrix for affy sets. [NEW](#)
- Links to [Mouse Genome Informatics](#) for mouse sets. [NEW](#)
- Improved GenBank Accession Search function. [NEW](#)

[README](#)

*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:

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:

- EGO  
 UniGene  
 Intersection  
 A\_unique  
 B\_unique

To Search RESOURCERER Using GenBank Accessions: Select Species and Supply Accession Numbers.

Human

Rat

Mouse

Two options: Upload a text file with accession numbers, or write in the text area.

Human  
 Rat  
 Mouse

[What's New](#) | [Description](#) | [Data](#) | [Targets of Discovery](#)  
[PGA Tools](#) | [Outreach](#) | [Contact List](#) | [Links](#) | [Home](#)

TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

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

# RESOURCERER: An Example



NIABMAP

NIA+BMAP & affy\_HG-U95A

Based On: EGO

There are 10776 rows in this table.

Download

Jump to page

1

Page 1 of 216 is currently displayed. Next

Dataset A	Rearray ID	Clone Name	GenB Acc	NIA TC	Dataset B	Probe ID	Clone Name	GenB Acc	affy_HG-U95A TC
NIA	H3001A04	C0001C05	<a href="#">BG062931</a>	<a href="#">TC511431</a>	affy_HG-U95A	<a href="#">37668 at</a>		<a href="#">M69039</a>	<a href="#">THC888982</a>
NIA	H3001A07	C0001D11	<a href="#">BG062933</a>	<a href="#">TC536824</a>	affy_HG-U95A	<a href="#">38060 at</a>		<a href="#">AI541336</a>	<a href="#">THC889116</a>
NIA	H3001A08	C0001E03	<a href="#">BG062934</a>	<a href="#">TC501927</a>	affy_HG-U95A	<a href="#">41439 at</a>		<a href="#">AJ001381</a>	<a href="#">THC976851</a>
NIA	H3001A12	C0002A05	<a href="#">BG062937</a>	<a href="#">TC501957</a>	affy_HG-U95A	<a href="#">39093 s at</a>		<a href="#">Y12059</a>	<a href="#">THC881255</a>
NIA	H3001A12	C0002A05	<a href="#">BG062937</a>	<a href="#">TC501957</a>	affy_HG-U95A	<a href="#">39094 at</a>	IMAGE-2481377	<a href="#">AI991631</a>	<a href="#">THC881255</a>
NIA	H3001B02	C0002C06	<a href="#">BG063016</a>	<a href="#">TC457292</a>	affy_HG-U95A	<a href="#">35600 at</a>		<a href="#">AB023967</a>	<a href="#">THC1004576</a>
NIA	H3001B08	C0002F11	<a href="#">BG063022</a>	<a href="#">TC457696</a>	affy_HG-U95A	<a href="#">1402 at</a>		<a href="#">M16038</a>	<a href="#">THC995041</a>
							<small>UBIQUINONE dehydrogenase (ubiquinone) GO:0006120 (complex I</small>		<small>(ubiquinol:ubiquinone reductase) [Mus musculus]</small>

# RESOURCER: Genome M



## Genetic Marker Search



### To Search RESOURCER Using C

This genetic to physical mapping tool uses genetic markers from Uni

Mapped to Mouse Genome: chr16  
Range: 23628083-28355680

There are 78 rows in this table. Page 1 of 2 is currently displayed.

Marker Name	UniSTS ID	Genetic Map	Data Set	GenBank Acc	TIGR TC	Chr Left	Chr Right	TGI Annotation
01.MMHAP12FLA1.seq	122645	16				23628083	23628267	
			BMAP TIGR_25K_Mouse_Set	<a href="#">AI848192</a>	<a href="#">TC640765</a>	23040198	23652949	preprosomatostatin; somatostatin [Mus musculus]
			Agilent	<a href="#">AA051655</a>	<a href="#">TC640765</a>	23040198	23652949	preprosomatostatin; somatostatin [Mus musculus]
			RIKEN_20K	<a href="#">AV149991</a>	<a href="#">TC640765</a>	23040198	23652949	preprosomatostatin; somatostatin [Mus musculus]
			affy_MG-U74A affy_Mu11KB affy_MG-U74Av2	<a href="#">X51468</a>	<a href="#">TC640765</a>	23040198	23652949	preprosomatostatin; somatostatin [Mus musculus]
			OPB106_M02054	<a href="#">X810202</a>	<a href="#">TC640765</a>	23040198	23652949	preprosomatostatin; somatostatin [Mus musculus]
			Agilent	<a href="#">AA6558</a>	<a href="#">TC264821</a>	23726084	23750009	mouse BCL6; BCL6
			affy_Mu11KB affy_MG-U74A affy_MG-U74Av2	<a href="#">X51468</a>	<a href="#">TC640765</a>	23040198	23652949	mouse BCL6; BCL6
			affy_MG-U74C affy_MG-U74C2 Agilent	<a href="#">AA61805</a>	<a href="#">TC268173</a>	24043447	24043308	
			affy_Mu11KB	<a href="#">X71142</a>	<a href="#">TC268173</a>	24156457	24157478	
			affy_MG-U74A affy_MG-U74Av2	<a href="#">AA83205</a>	<a href="#">TC268173</a>	24156457	24157478	
			Y937W	<a href="#">AA62286</a>	<a href="#">TC268173</a>	24156924	24157424	
			SLA	<a href="#">X318561</a>	<a href="#">TC268173</a>	24157346	24152833	mouse to Q913701 [Spilopsocus

## Mapped to Mouse Genome: chr16

Range: 23628083-28355680

There are 78 rows in this table: [Download](#) [Jump to page](#)   
Page 1 of 2 is currently displayed. [Next](#)

Marker Name	UniSTS ID	Genetic Map	Data Set	GenBank Acc	TIGR TC	Chr Left	Chr Right	TGI Annotation
01.MMHAP12FLA1.seq	122645	16				23628083	23628267	
			BMAP TIGR_25K_Mouse_Set	<a href="#">AI848192</a>	<a href="#">TC640765</a>	23040198	23652949	preprosomatostatin; somatostatin [Mus musculus]
			Agilent	<a href="#">AA051655</a>	<a href="#">TC640765</a>	23040198	23652949	preprosomatostatin; somatostatin [Mus musculus]
			RIKEN_20K	<a href="#">AV149991</a>	<a href="#">TC640765</a>	23040198	23652949	preprosomatostatin; somatostatin [Mus musculus]
			affy_MG-U74A affy_Mu11KB affy_MG-U74Av2	<a href="#">X51468</a>	<a href="#">TC640765</a>	23040198	23652949	preprosomatostatin; somatostatin [Mus musculus]

# Mapping Genes to Genomes



**TIGR** THE INSTITUTE FOR GENOMIC RESEARCH

Home | Databases | Gene Indices | News | What's New | Search | About TIGR | Contact | Site Map | Contact Us | Feedback

Comprehensive Microbial Resource  
Unfinished Microbial Genomes  
Eukaryotic Resources  
Gene Indices  
Protein Databases  
TIGRFAMs  
Fungal Databases  
Human Sequencing Projects

**TIGR Gene Indices**

What's New | BLAST Search | TGI Software | FAQ

Integrating data from international EST sequencing, genome sequencing and gene research projects, Gene indices are an analysis of transcribed sequences represented in the world's public EST data.

**EGO** EGO - linking orthologous genes across eukaryotic organisms  
**Genomic Maps** Genomic Maps - mapping TC sequences to eukaryotic genomes  
**RESOURCEER** RESOURCEER - cross referencing mammalian microarray resources  
**DAS** DAS - providing distributed annotations for completed genomes

**Animal Gene Indices**

Arabidopsis thaliana 1.0 6/1/02	Brugia malayi 2.0 3/1/02	Canis 1.1 3/24/02	Cattle 1.0 6/1/02
C. elegans 2.0 6/1/02	Chicken 2.0 6/1/02	C. nematode 1.0 6/1/02	Drosophila 2.0 6/1/02
Honey bee 2.0 6/1/02	Human 4.0 6/1/02	Mus musculus 4.0 6/1/02	Mouse 2.0 6/1/02
O. latipes 1.0 6/1/02	C. vicina 2.0 7/24/02	P. galeatus 4.0 7/24/02	R. norvegicus 2.0 6/1/02
S. cerevisiae 4.0 6/1/02	X. laevis 1.1 6/1/02	Z. marmoratus 1.0 6/1/02	

**Plant Gene Indices**

Arabidopsis 2.0 6/1/02	B. distachyon 4.0 6/1/02	C. sinensis 1.0 2/24/02	C. glutinosa 1.0 6/1/02
I. plantaginea 2.0 6/1/02	M. domestica 1.1 6/1/02	M. sativa 1.0 6/1/02	M. truncatula 2.0 6/1/02
P. sativum 1.1 6/1/02	R. glaberrima 2.0 6/1/02	T. aestivum 1.0 6/1/02	Z. mays 4.0 6/1/02
S. tuberosum 1.0 6/1/02	T. monandrum 1.0 6/1/02	W. arabidopsis 4.0 6/1/02	

**Protist Gene Indices**

C. parvum 1.0 6/1/02	D. discoideum 2.0 6/1/02	I. trophanta 1.0 6/1/02	L. major 1.0 6/1/02
M. mus musculus 2.0 6/1/02	P. falciparum 2.0 6/1/02	P. falciparum 2.0 6/1/02	P. falciparum 2.0 6/1/02
S. cerevisiae 2.0 6/1/02	T. brucei 1.0 6/1/02	T. brucei 1.0 6/1/02	T. brucei 1.0 6/1/02
T. brucei 2.0 6/1/02	T. brucei 1.0 6/1/02		

**Fungal Gene Indices**

A. nidulans 2.0 6/1/02	C. glabrata 1.1 6/1/02	O. sativa 1.0 6/1/02	M. grisea 1.0 6/1/02
N. crassa 1.0 6/1/02	S. cerevisiae 2.0 6/1/02	S. cerevisiae 2.0 6/1/02	

**Legend:**  
 Most recent version number  
 Date of most recent update

The TIGR Gene Indices are built using:

- **seqalign** (Chang, Chang, Stein Schwartz, Lukar Wagner, and Wirth Mader (2000), "A greedy algorithm for aligning DNA sequences", J. Comput Biol. 2000, 13: 215-241)
- **CAPI**, developed by Dr. Xiaojin Huang
- **Parent Transcripts (AncestorSet)**, from ParentSet
- **BNA Protein Search program (bpe)** developed by Dr. Xiaojin Huang

The ORF annotation of TCs is done using:


- **ESTScan** (Jin, C. Longaker, C.Y. and Buchler, P. (1999) ESTScan: A program for detecting, evaluating, and reorientating potential coding regions in EST sequences. *BIOINFORMATICS* 15: 145-151)
- **BIANNA EST** (Hindupatla A.G., Frazee P., Ryzhko M. Related Article: DEANA-EST - a statistical analysis. *Bioinformatics* 2001 Oct 15;17(10):1213-9)
- **GeneFinder** (Expressed Sequence Tag Analysis Tools Set) (©) Cop. R.L.C. Stone 1996-1999. Human Genome Mapping Project, BC, Montreal, Canada, UK.

The Orf2 mappings are done using:

- **blast** (©) Wirth Mader, see reference
- **blast** (©) from ParentSet
- **gpg2** developed by Dr. Xiaojin Huang

The expression profiles of the ESTs are scored using:

- **W**, modified (Chick, Ok, and Falciano (2001) The comparison of gene expression from multiple cDNA libraries. *Oncogene Research* 10:2053-2061)



**Genome Views - Genomic Maps**

- Alignments of plants TCs and ESTs with arabidopsis chromosomes**
- Alignments of plants TCs and ESTs with rice BACs**
- Alignments of mammalian TCs and ESTs with the human "Golden Path" chromosomes**
- Alignments of model organisms TCs with C.elegans chromosomes**
- Alignments of model organisms TCs with Drosophila chromosomes**
- Alignments of model organisms TCs with Yeast chromosomes**
- Alignments of model organisms TCs with Mouse (MGSC v3) chromosomes**
- Alignments of Indica and Japonica TCs against the Indica Contigs**
- Alignments of O.latipes TCs against the Fugu Genome (Scaffolds)**

Send mail to TIGR Search | Site Map

TIGR Databases | What's New | About TIGR | TIGR Faculty  
 TIGR Gene Indices | Conferences, Education and Training  
 TIGR Software | Career Opportunities | Related Links


Frequently Asked Questions  
 ©1999-2000 TIGR

**Razvan Sultana**  
**Foo Cheung**

**TIGR**  
 THE INSTITUTE FOR GENOMIC RESEARCH

# Human: Annotation of the Golden Path

Human chromosome chr20



Click to position view

5':  3':

Scale click does:

Zoom in  Zoom out

Map click does:

- Zoom (as above)
- Display alignment
- Show sequence details
- Show paralog hits
- Link to GoldenPath

gene/TC info


Sequence: THCS83405  
(length=2846 nt)

	5'	3'
genome:	24840457	24841825
exon:	95	1453

identity: 100% similarity: 100%  
score: 2735 paralogs: no

Annotation:

Scale: 5' = 34,826,487 3' = 34,851,665



Ensembl Genes	5'	+	3'
Human TCs	5'	+	3'
Mouse TCs	5'	+	3'
Rat TCs	5'	+	3'
Cattle TCs	5'	+	3'
Pig TCs	5'	+	3'

**TIGR**  
THE INSTITUTE FOR GENOMIC RESEARCH

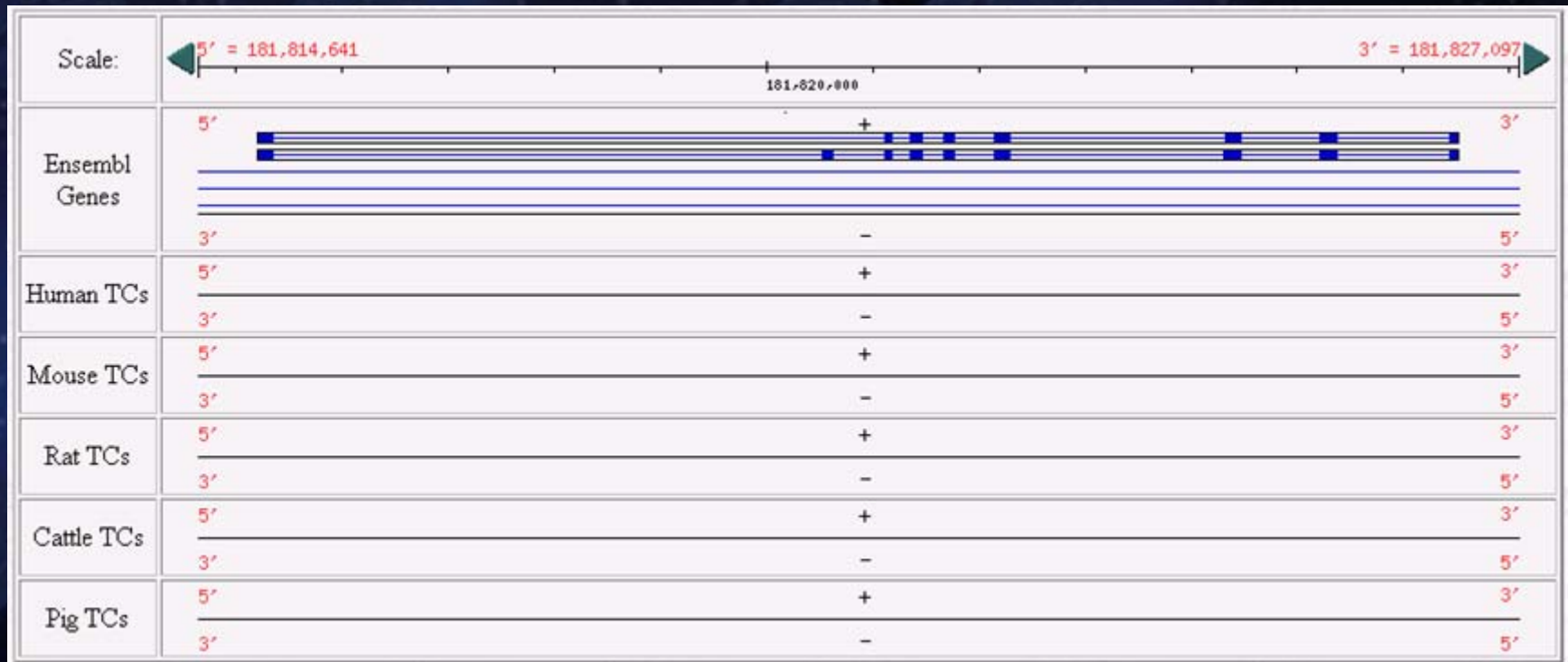
Razvan Sultana

# Gene Finding in Humans is easy!

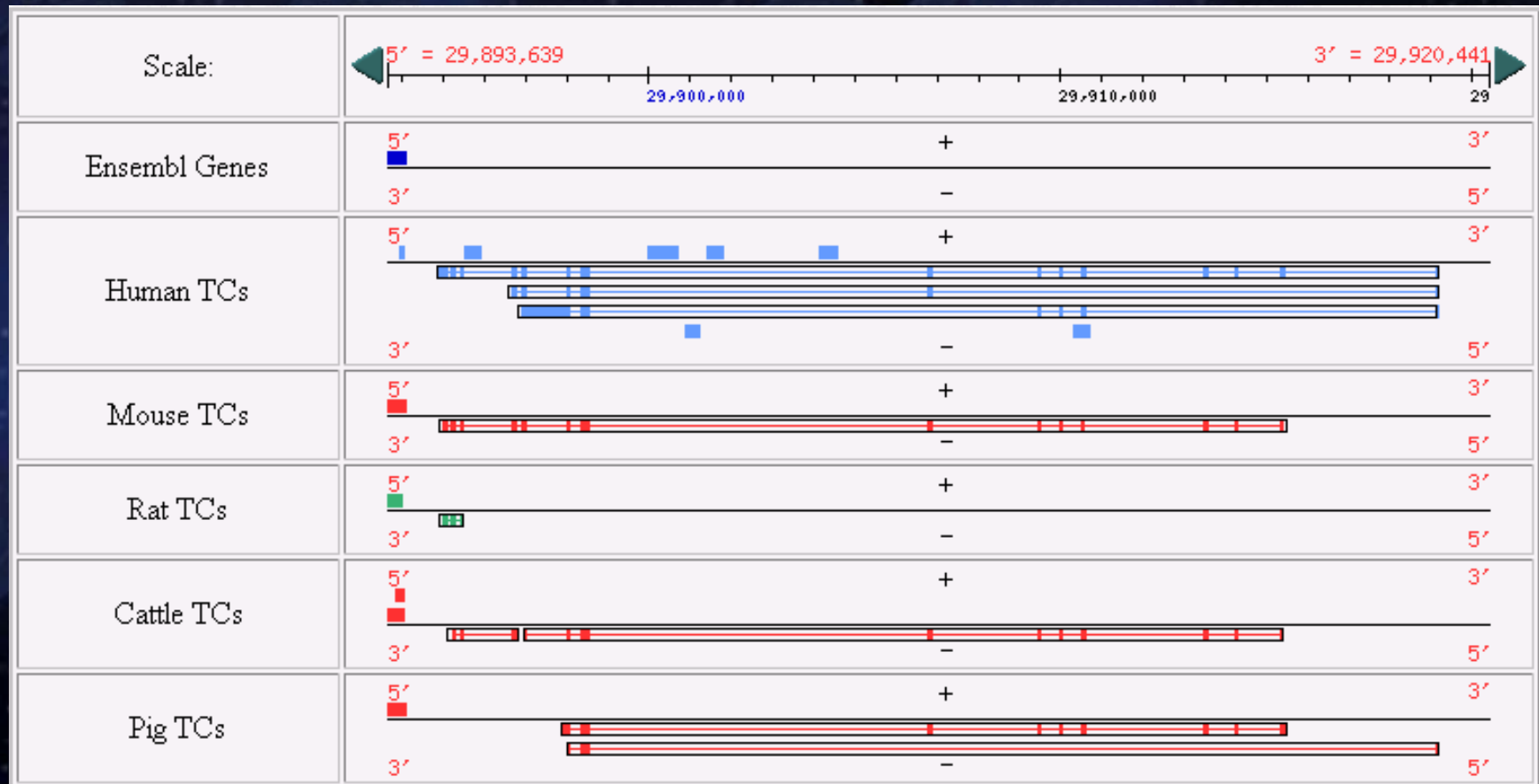




# Gene Finding in Humans is easy?



# Gene Finding in Humans is difficult?



# Gene Finding in Humans is difficult?

Scale:		
Ensembl Genes	5' + 3'	3' - 5'
Human TCs	5' + 3'	3' - 5'
Mouse TCs	5' + 3'	3' - 5'
Rat TCs	5' + 3'	3' - 5'
Cattle TCs	5' + 3'	3' - 5'
Pig TCs	5' + 3'	3' - 5'

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

**11,658**

**Razvan Sultana**



**What do we need to know to use  
expression data?**

**TIGR**

THE INSTITUTE FOR GENOMIC RESEARCH



## Minimum information about a microarray experiment (MIAME)—toward standards for microarray data

Alvis Brazma<sup>1</sup>, Pascal Hingamp<sup>2</sup>, John Quackenbush<sup>3</sup>, Gavin Sherlock<sup>4</sup>, Paul Spellman<sup>5</sup>, Chris Stoeckert<sup>6</sup>, John Aach<sup>7</sup>, Wilhelm Ansorge<sup>8</sup>, Catherine A. Ball<sup>4</sup>, Helen C. Causton<sup>9</sup>, Terry Gaasterland<sup>10</sup>, Patrick Glenisson<sup>11</sup>, Frank C.P. Holstege<sup>12</sup>, Irene F. Kim<sup>4</sup>, Victor Markowitz<sup>13</sup>, John C. Matese<sup>4</sup>, Helen Parkinson<sup>1</sup>, Alan Robinson<sup>1</sup>, Ugis Sarkans<sup>1</sup>, Steffen Schulze-Kremer<sup>14</sup>, Jason Stewart<sup>15</sup>, Ronald Taylor<sup>16</sup>, Jaak Vilo<sup>1</sup> & Martin Vingron<sup>17</sup>

Microarray analysis has become a widely used tool for the generation of gene expression data on a genomic scale. Although many significant results have been derived from microarray studies, one limitation has been the lack of standards for presenting and exchanging such data. Here we present a proposal, the Minimum Information About a Microarray Experiment (MIAME), that describes the minimum information required to ensure that microarray data can be easily interpreted and that results derived from its analysis can be independently verified. The ultimate goal of this work is to establish a standard for recording and reporting microarray-based gene expression data, which will in turn facilitate the establishment of databases and public repositories and enable the development of data analysis tools. With respect to MIAME, we concentrate on defining the content and structure of the necessary information rather than the technical format for capturing it.

**MAGE-ML – XML-based data exchange format**  
<<http://www.mged.org>>

# SOPs are available

Coming: Data QC SOP

## cDNA/template prep

THE INSTITUTE FOR GENOMIC RESEARCH <i>Standard Operating Procedure</i>		
TITLE: MICROARRAY C-DNA CLONE GROWTH AND TEMPLATE MINIPREPPING		PAGE: 1 of 3
SOP #: M001	REVISION LEVEL: 1	EFFECTIVE DATE: 6/21/01
AUTHOR: <i>Jeremy Hasselman</i>	PRIMARY REVIEWERS: <i>Emily Chen, John Quackenbush, Ivana Tang</i>	
<b>1. PURPOSE</b> This protocol describes clone handling, plate replication, and DNA template preparation in a 96 well format.		
<b>2. SCOPE</b> This procedural format is utilized by Human Colon Cancer and Mouse microarray		

## PCR purification

THE INSTITUTE FOR GENOMIC RESEARCH <i>Standard Operating Procedure</i>		
TITLE: MICROARRAY PCR, PURIFICATION, AND STORAGE		PAGE: 1 of 3
SOP #: M002	REVISION LEVEL: 1	EFFECTIVE DATE: 6/21/01
AUTHOR: <i>Jeremy Hasselman</i>	PRIMARY REVIEWERS: <i>Emily Chen, John Quackenbush, Ivana Tang</i>	
<b>1. PURPOSE</b> This protocol describes PCR amplification of eukaryotic cDNA plasmid inserts, gel electrophoresis, purification, and storage of PCR products.		
<b>2. SCOPE</b>		

## Printing

THE INSTITUTE FOR GENOMIC RESEARCH <i>Standard Operating Procedure</i>		
TITLE: MAKING MICROARRAY PRINTING PLATES (IN DMSO)		PAGE: 1 of 1
SOP #: M003	REVISION LEVEL: 1	EFFECTIVE DATE: 6/21/01
AUTHOR: <i>Jeremy Hasselman</i>	PRIMARY REVIEWERS: <i>Emily Chen, John Quackenbush, Ivana Tang</i>	
<b>1. PURPOSE</b> This protocol describes the method for making microarray printing plates in a 96 well format. (The same procedure applies to a 384 well format as well.)		
<b>2. MATERIALS</b>		

## RNA labeling

THE INSTITUTE FOR GENOMIC RESEARCH <i>Standard Operating Procedure</i>		
TITLE: AMINOALLYL LABELING OF RNA FOR MICROARRAYS		PAGE: 1 of 7
SOP #: M004	REVISION LEVEL: 1	EFFECTIVE DATE: 6/21/01
AUTHOR: <i>Jeremy Hasselman</i>	PRIMARY REVIEWERS: <i>Emily Chen, Erik Snesrud, Ivana Tang</i>	
<b>1. PURPOSE</b> This protocol describes the labeling of eukaryotic RNA with aminoallyl labeled nucleotides via first strand cDNA synthesis followed by a coupling of the aminoallyl groups to either Cyanine 3 or 5 (Cy 3/Cy5) fluorescent molecules.		
<b>2. SCOPE</b>		

## Hybridization

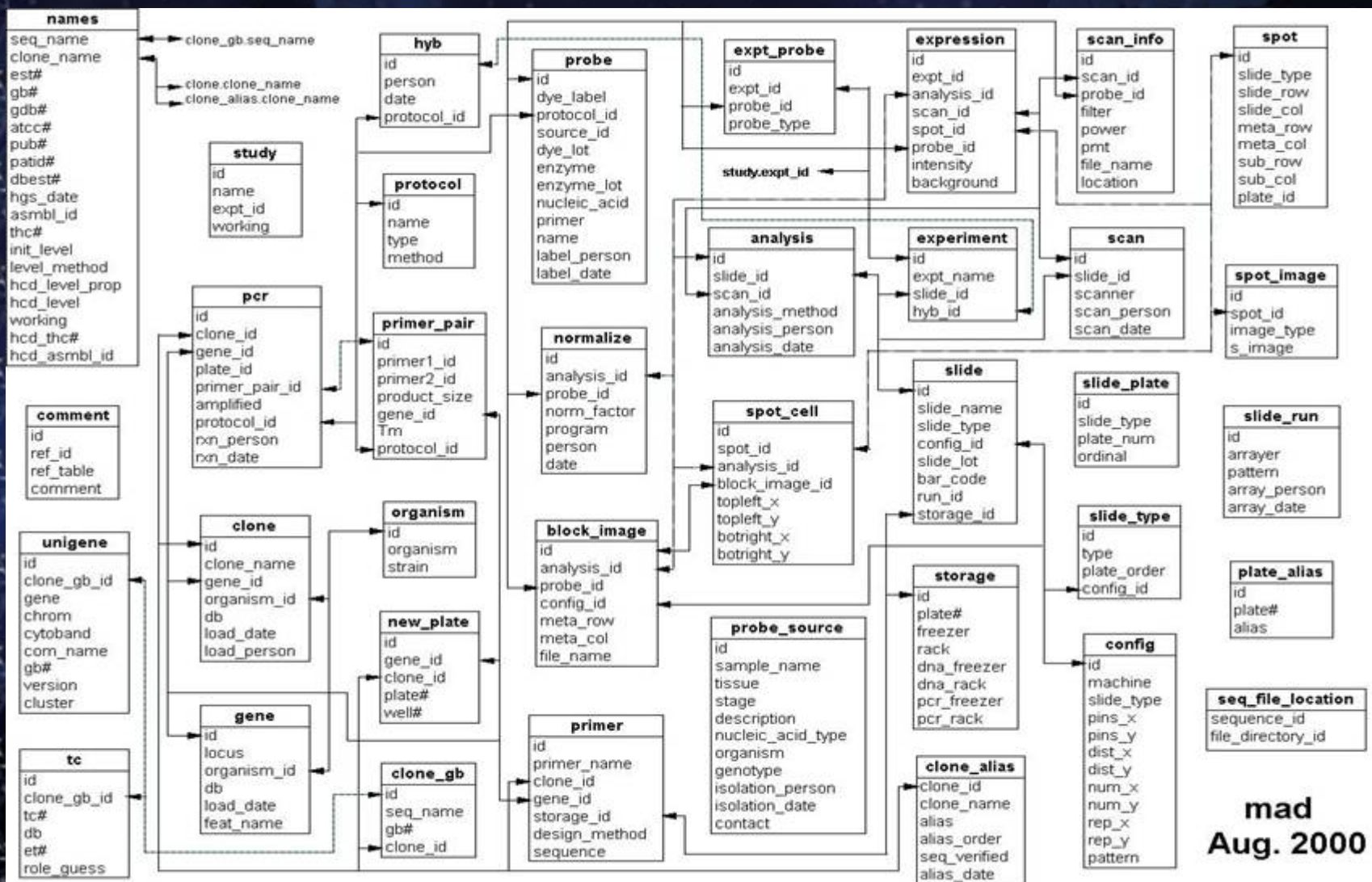
THE INSTITUTE FOR GENOMIC RESEARCH <i>Standard Operating Procedure</i>		
TITLE: MICROARRAY LABELED PROBE HYBRIDIZATION		PAGE: 1 of 5
SOP #: M005	REVISION LEVEL: 1	EFFECTIVE DATE: 6/21/01
AUTHOR: <i>Jeremy Hasselman</i>	PRIMARY REVIEWERS: <i>Emily Chen, Erik Snesrud, Ivana Tang</i>	
<b>1. PURPOSE</b> This protocol describes the hybridization of a Cy labeled cDNA probe (mix of Cy3 and Cy5) onto coated slide spotted with PCR amplified cDNA.		
<b>2. SCOPE</b>		

TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

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

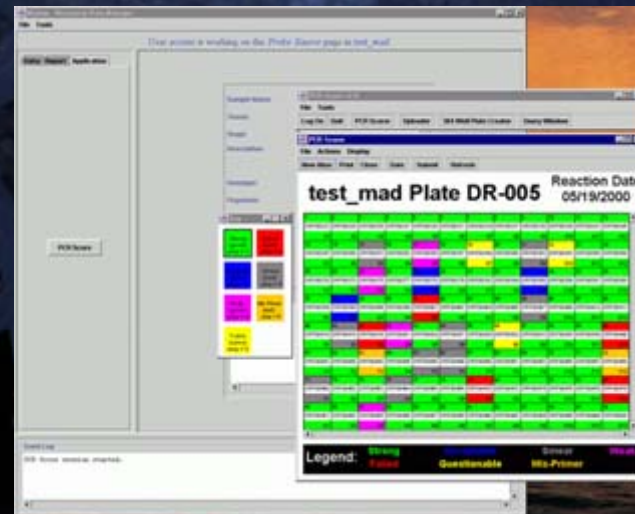
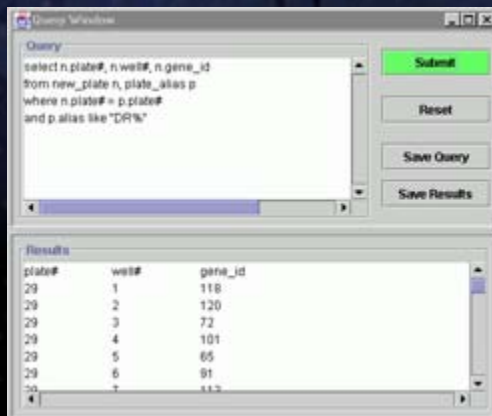
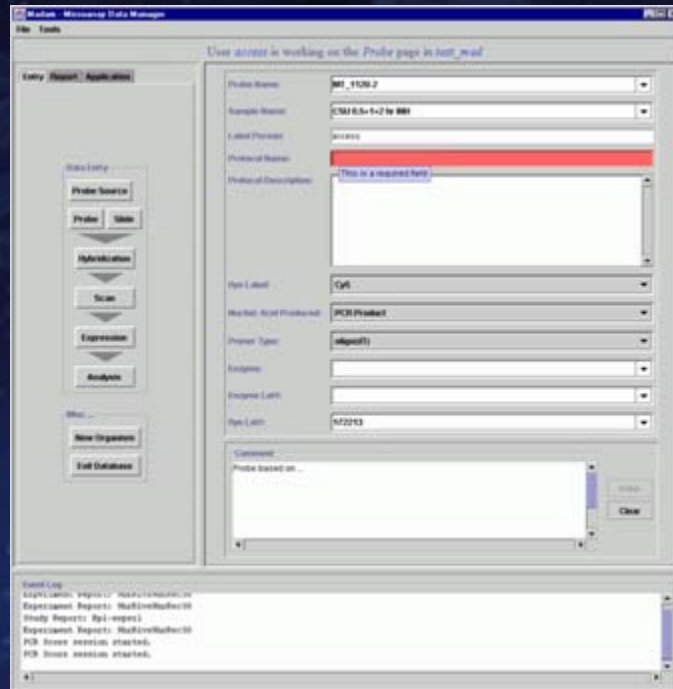
# MAD Microarray Database Schema



mad  
Aug. 2000

# MADAM: Microarray Data Manager

Alex Saeed  
Vasily Sharov  
Jerry Li  
Joe White



Available with source and MySQL database



# MIDAS: Normalization and Filtering

Wei Liang

The screenshot displays the TIGR MIDAS V2.0 software interface. At the top, there is a menu bar with options: Project, Read Data, Operations, Write Data, Tools, and Help. Below the menu is a toolbar with various icons for file operations and data processing. The main workspace is divided into three sections:

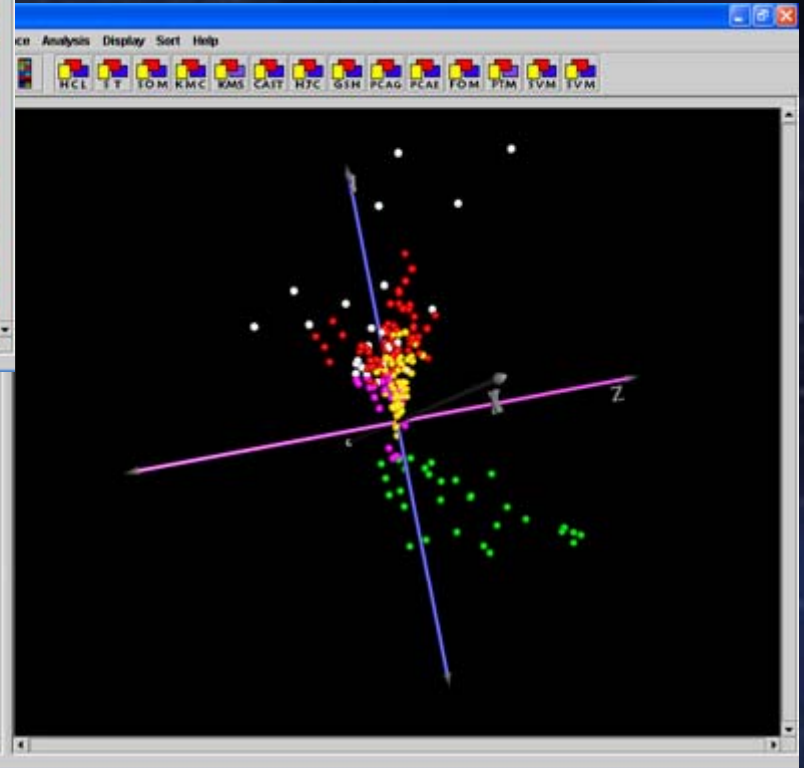
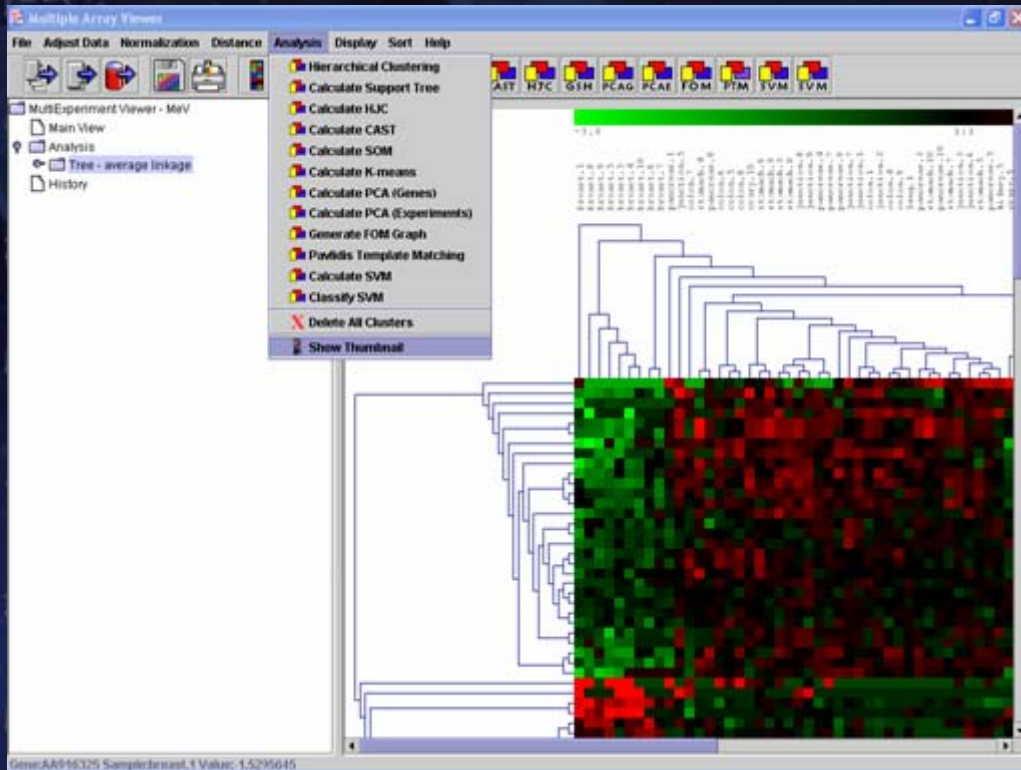
- Work Flow:** A vertical flowchart showing the sequence of operations. It starts with a 'Load OTO' step, followed by two parallel paths that merge, then 'LocFit' (two parallel steps), another merge, and finally 'Slicer' and 'Plot' steps.
- Parameters:** A table with two columns: 'Parameter' and 'Value'.

Parameter	Value
Virtual Trim	<input checked="" type="checkbox"/>
Re-number	<input type="checkbox"/>
Output Trimmed Data	<input type="checkbox"/>
- Tools:** A button labeled 'View all parameters'.

At the bottom, the 'Processing Status' section displays the message: 'Welcome to use TIGR MIDAS.' and the TIGR MIDAS logo.

# MeV: Data Mining Tools

Alex Saeed  
Alexander Sturn  
Nirmal Bhagabati  
Syntek Inc.  
Datanaut, Inc.



# MeV: Metabolic pathway analysis is coming

View Structural and Genome Information for the selected metabolic pathways

Metabolic Pathway	Gene	EC Number	Gene	EC Number
serine biosynthesis, serine biosynthesis	serC1	4.2.1.3	serC2	4.2.1.3
serine biosynthesis, serine biosynthesis	serC3	4.2.1.10	serC4	4.2.1.10
serine biosynthesis, serine biosynthesis	serC5	4.2.1.14	serC6	4.2.1.14
serine biosynthesis, serine biosynthesis	serD	2.3.1.22	serE	2.3.1.22
serine biosynthesis, serine biosynthesis	serG	4.2.1.20	serH	4.2.1.20
serine biosynthesis, serine biosynthesis	serI	4.2.1.26	serJ	4.2.1.26
serine biosynthesis, serine biosynthesis	serL	4.2.1.32	serM	4.2.1.32
serine biosynthesis, serine biosynthesis	serN	4.2.1.33	serO	4.2.1.33
serine biosynthesis, serine biosynthesis	serP	4.2.1.34	serQ	4.2.1.34
serine biosynthesis, serine biosynthesis	serR	4.2.1.35	serS	4.2.1.35

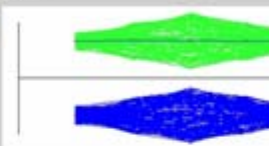
View expression pattern for selected genes

Save selected clusters of genes

FROM GENOME TO METABOLISM

1. Click here to view metabolic information for ALL genes
2. View Metabolic Information for Selected Clusters of Genes

Cluster 1  
Cluster 2



Click here to view metabolic information for highlighted genes

FROM METABOLISM TO GENOME

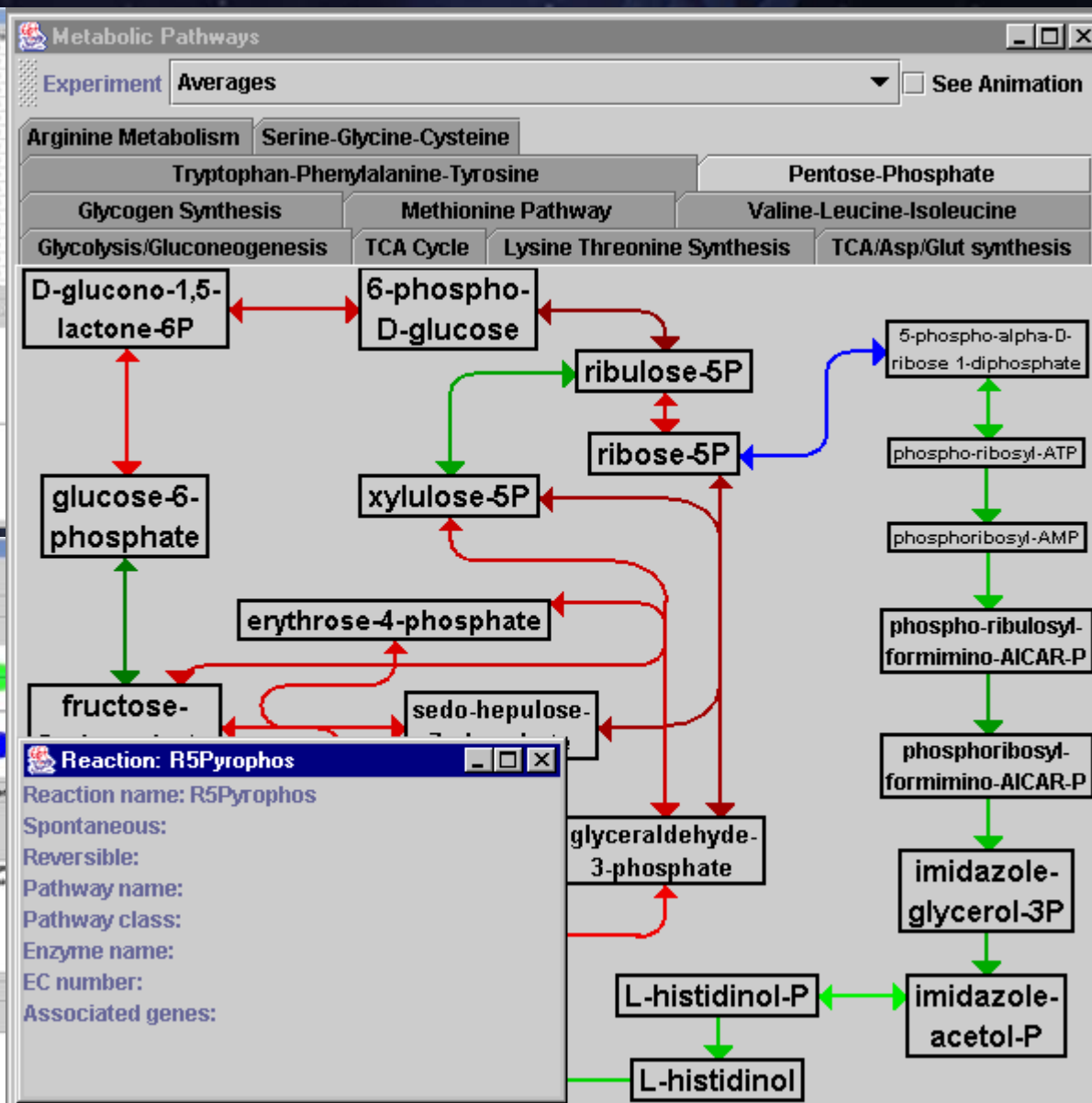
1. View Gene Expression for Selected Metabolic Pathways
2. View Gene Expression for Selected Metabolic Reactions

Pathway Name	Pathway Class
alanine biosynthesis	amino acid biosynthesis
arginine biosynthesis	amino acid biosynthesis
aspartate and asparagine biosynthesis	amino acid biosynthesis
aspartate biosynthesis	amino acid biosynthesis
lysine catabolism	amino acid degradation
glucose 1-phosphate metabolism	carbohydrate metabolism

View Selected Pathways

Reaction Strichs	Enzyme	EC Number
tyract1.0	1-phosphotransferase	2.7.1.56
tyract2.0	1-phosphotransferase	2.7.1.56
tyract1.0	transketolase	2.2.1.1
tyract1.0	transketolase	2.2.1.1
tyract1.0	transketolase	2.2.1.1
tyract1.0	transketolase	2.2.1.1
tyract1.0	transketolase	2.2.1.1

View Selected Reactions





# What can we do with expression data?

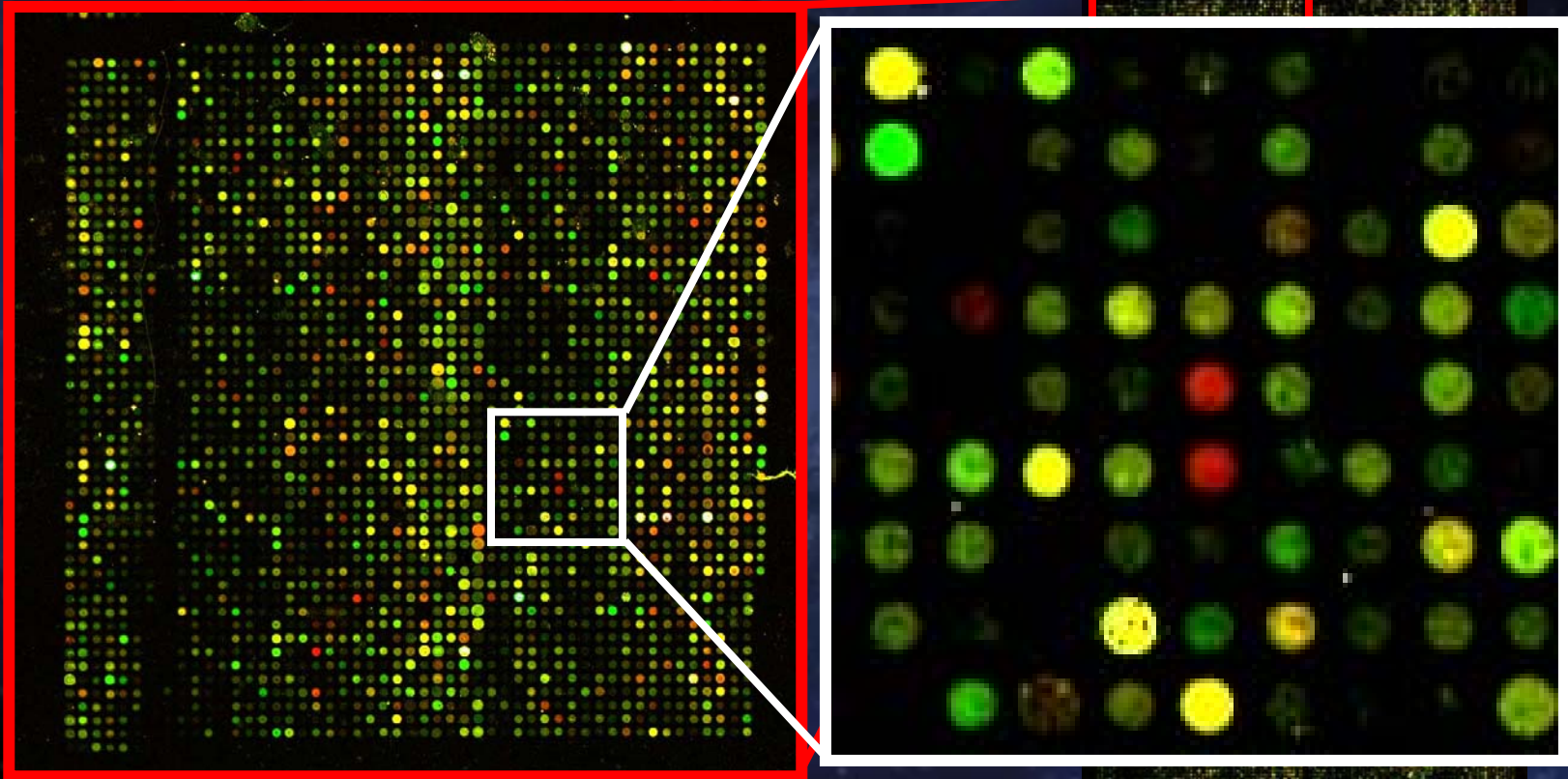
**Link to Genetics**

**TIGR**

THE INSTITUTE FOR GENOMIC RESEARCH

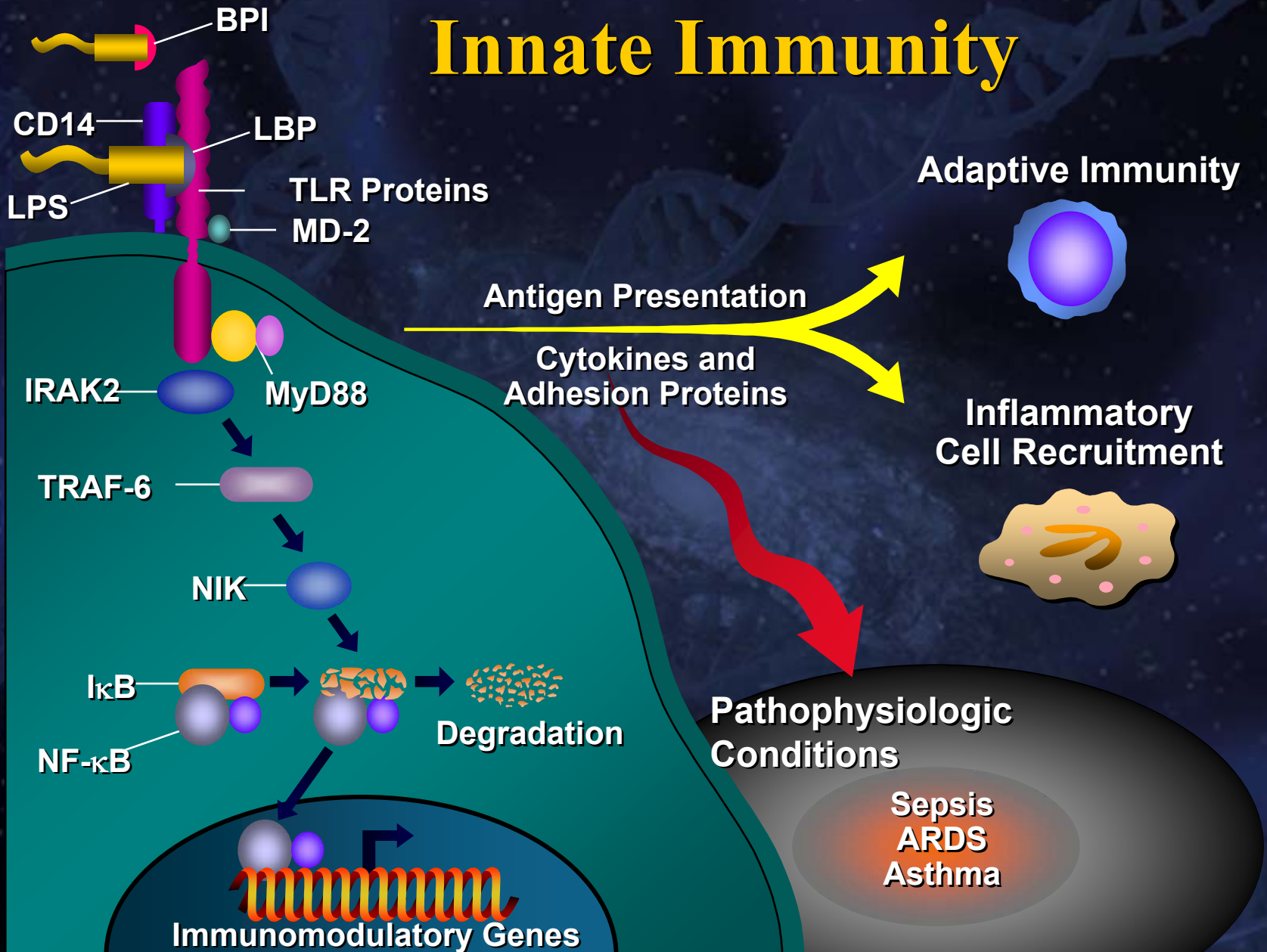
# 32,448 element mouse array

Thanks to M. Ko (NIA) and B. Soares (BMAP)

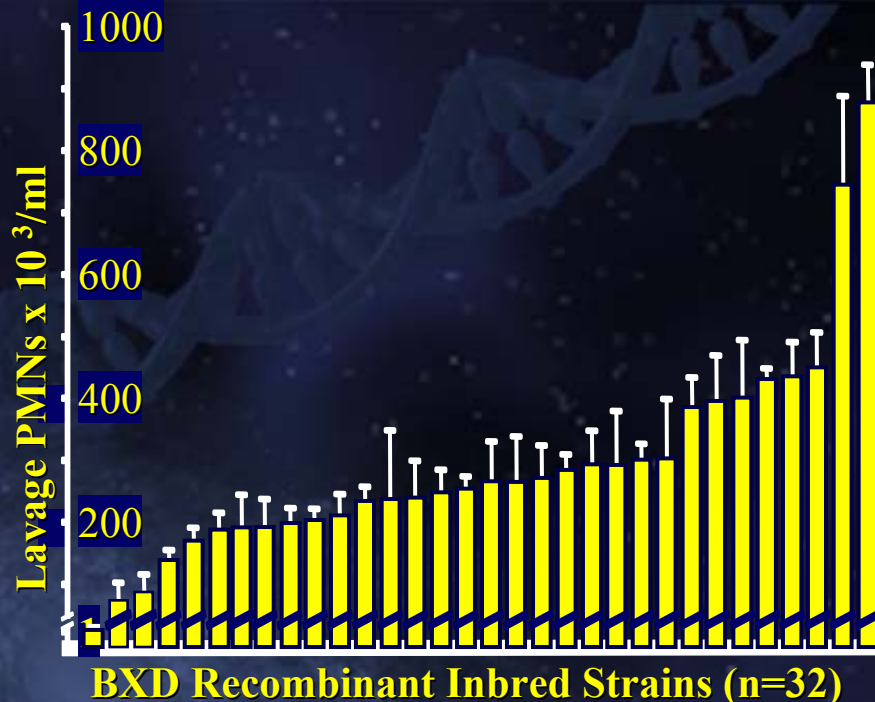


**kidney vs. heart**  
**15 $\mu$ g total RNA**

# Innate Immunity



# Examples

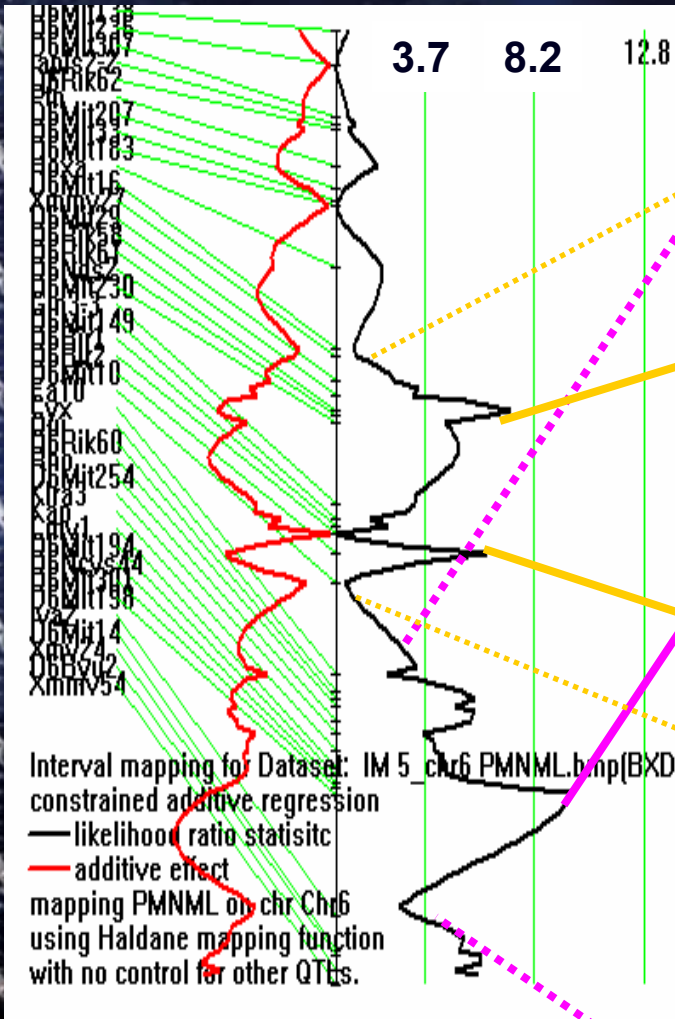


F2 progeny show QT (PMN and TNF $\alpha$  levels) after inhaling endotoxin LPS. Mice are classified as low to high responders and used to generate a QTL map.

**Goal: Identify mouse candidate genes that regulate lung response to inhaled lipopolysaccharide**

# Trait: PMN Chromosome 6

cM 61.0 - 66.0



Marker	Stat	cM
Cyx	2.2	50.0
	3.4	
Qui	5.2	50.5
D6Rik60	5.7	
Rho	5	51.5
	5.7	
D6Mit254	5.6	
	5	
Klra3	3.0	
	4	
Kap	4.1	
	4.5	
Cmv1	4.4	
D6Mit194	4.8	61.5
D6Ncvs44	5.5	64.0
D6Mit301	7.6	64.0
D6Mit198	9.7	63.9
	9.6	
	9.5	
	9.3	
	9	
	8.7	
	8.3	
	7.9	
	7.5	
	7	
	6.5	
	6.1	
	5.6	
	5.1	
	4.6	
	4.2	
	3.8	
	3.4	
	3	
	2.7	
Iva2	3.1	69.5

Marker	Stat	cM
D6Mit16	0.7	30.5
Xmmv27	0.9	31.5
	1.4	
	1.9	
	2.2	
D6Mit29	2.4	36.5
	3.4	
D6Rik58	3.2	38.5
	5.4	
D6Rik61	7.2	39.0
D6Nds2	6.4	39.5
D6Mit230	4.6	43.0
	4.8	
	4.9	
	5	
	5.1	
	5.1	
	5	
	4.8	
	4.5	
	4.3	
	4	
	3.7	
	3.4	
Gln3-3	3.3	46.0
	2.9	
D6Mit149	1.4	46.3
D6Bir1	2.2	47.0
D6Bir2	0	48.0
	0.8	
	2.8	
D6Mit10	6.2	48.7
	4.7	
	3.3	
	2	
	1	
Ea10	0.4	49.0

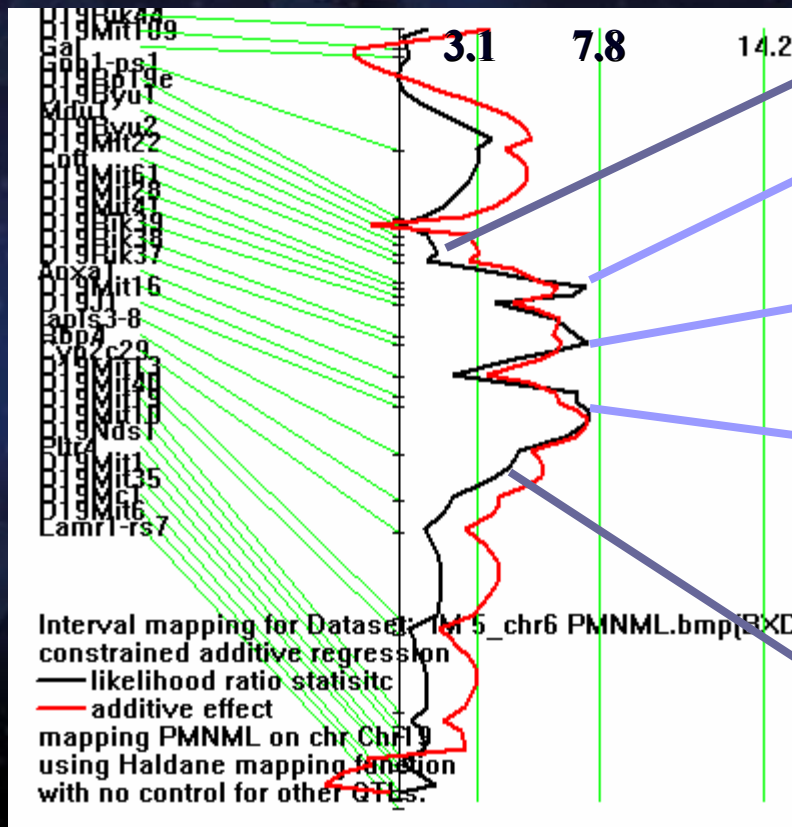
cM37.0-43.0

cM46.0-50.0



# Trait: PMN

## Chromosome 19

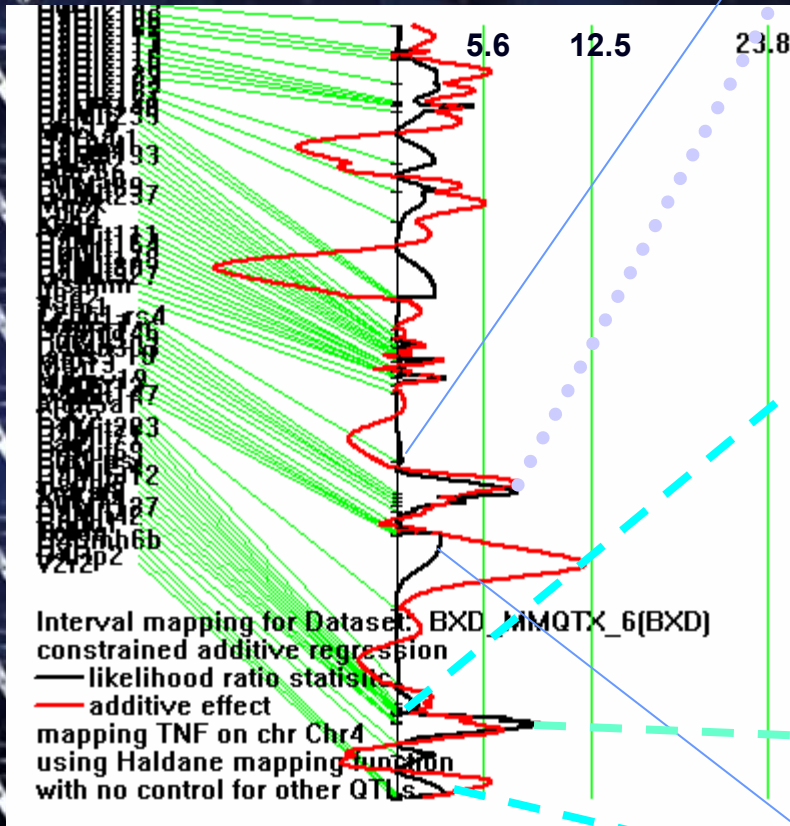


Marker	Stat	cM
Cntf	1.2	7.0
	3	
D19Mit61	5.5	9.0
D19Mit28	7.3	12.0
D19Mit41	6.8	16.0
D19Rik39	3.8	17.0
	5	
	6.1	
D19Rik38	6.8	14.0
D19Rik37	7.4	15.0
	6.1	
	4.6	
Anxa1	2.1	18.0
	4.2	
D19Mit16	7	15.0
D19J1	7	24.0
	7.4	
	7.5	
	7.1	
lpls3-8	6.5	30.0
	4.7	
	4.6	
	4.2	
	3.7	
	3.1	
Rbp4	2.2	38.0

**cM 9.0 - 17.0**

**cM 22.0 - 30.0**

# Trait: TNF Chromosome 4



Marker	Stat	cM
Tyrp1	0.4	38.0
Ccnb1-rs4	7.5	50.6
Ms15-1	7.2	50.6
	6.3	
	5.2	
	2.6	48.4
	3.3	
	3.8	
	4.2	
	4.3	
D4Mit146	4.3	53.6
Pmv19	4.4	52.7
	3.9	
	3.2	
D4Mit303	1.9	48.5
	1.8	
	1.7	
	1.6	
	1.4	
	1.1	
	0.9	
lpls3-10	1.2	52.7

Iva1	0	56.7
	0.3	
Mpmv19	1.4	53.5
D4Mit12	1.7	57.6
D4Mit147	2.9	38.0

Marker	Stat	cM
Lck	0.1	59.0

Cxv2	1.1	62.0
	1.2	
	1.3	
	1.4	
	1.4	
	1.4	
D4Mit203	1.3	60.0
	1.3	
D4Mit71	0.7	61.9
Lag	1.4	65.7
D4Mit69	1.2	63.4
	0.8	
Ela1-ps	0.3	66.1
D4Mit54	1.5	66.0
D4Mit312	4.4	69.8
Hspg2	4	71.4
	5.2	
	6.2	
	7.1	
	7.4	
Tnfrsf8	6.6	75.5
Xmv14	8.5	76.4
D4Mit127	7.9	77.5
D4Mit42	9.2	81.0

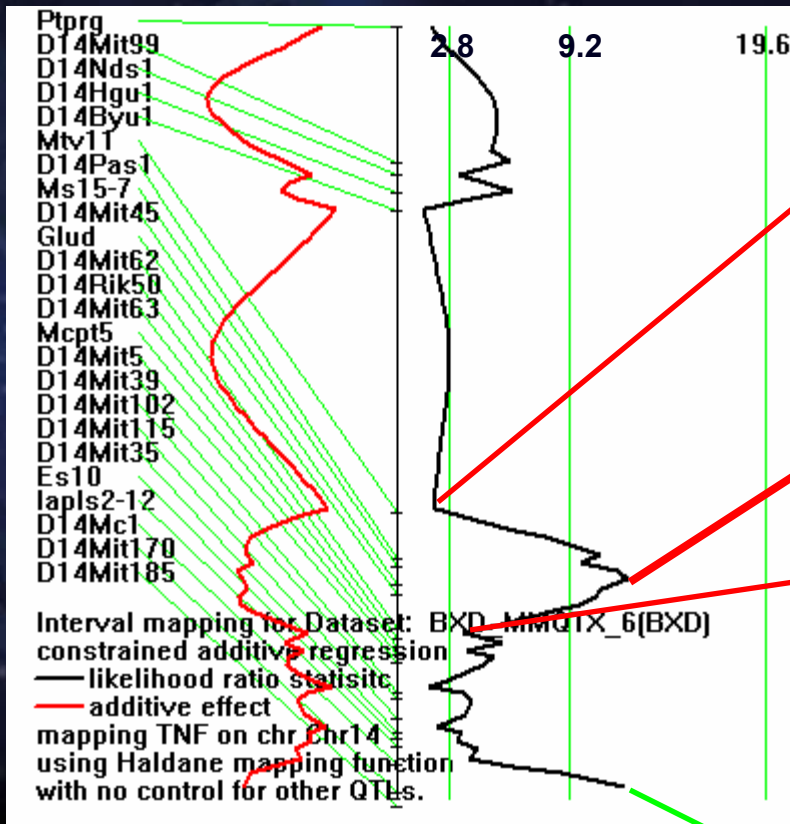
D4Smh6b	3.3	82.0
	3.2	
Dvl	2.5	82.0
	1.5	
D4Rp2	0.8	81.0

cM 48.0 - 52.0

cM 69.0 - 82.0

# Trait: TNF

## Chromosome 14



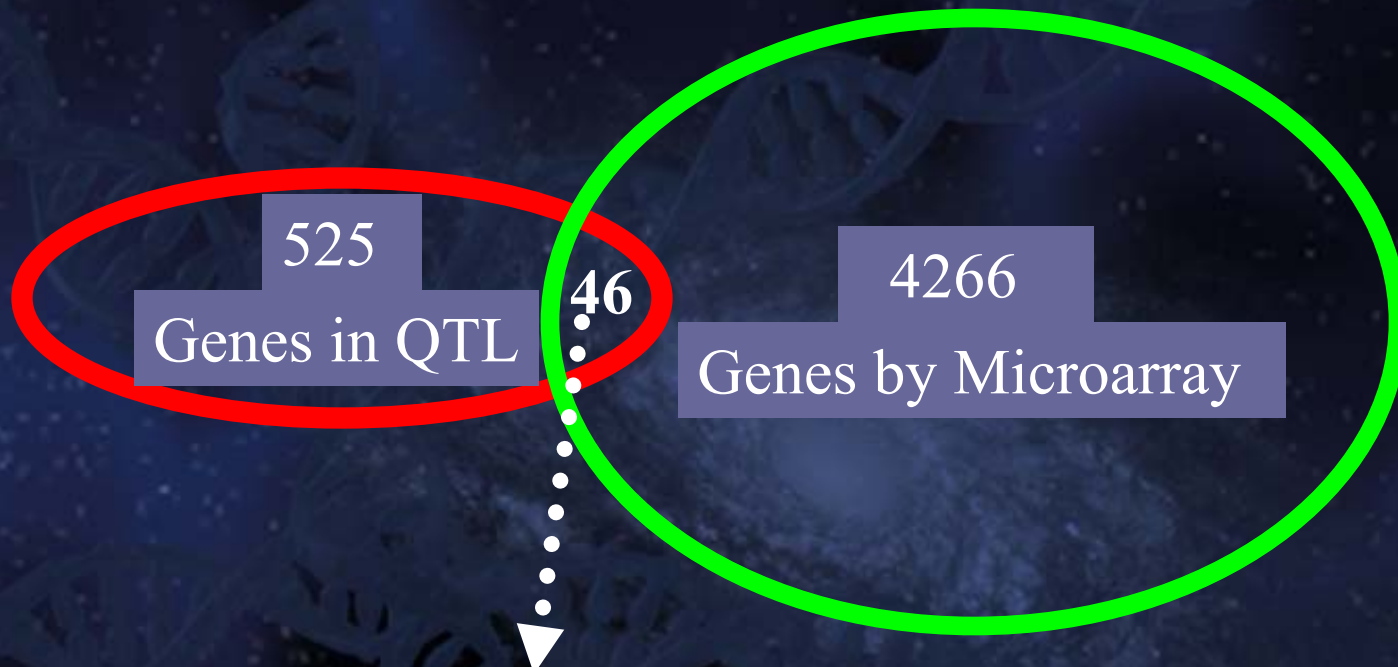
Marker	Stat	cM
Mtv11	2.1	16.0
	3.3	
	4.8	
	6.4	
	7.9	
	9.2	
D14Pas1	10.8	15.0
Ms15-7	9.9	16.5
	11.8	
D14Mit45	12.3	12.5
Glud	10.9	15.5
	10.6	
	9.7	
	8.4	
D14Mit62	5.6	18.5

cM 10.0 - 18.0

Marker	Stat	cM
D14Mc1	3.9	55.0
	6.3	
	8.7	
D14Mit170	12	63.0

cM 61.0 - 65.0

# Microarray Expression-QTL Consensus Candidate Genes



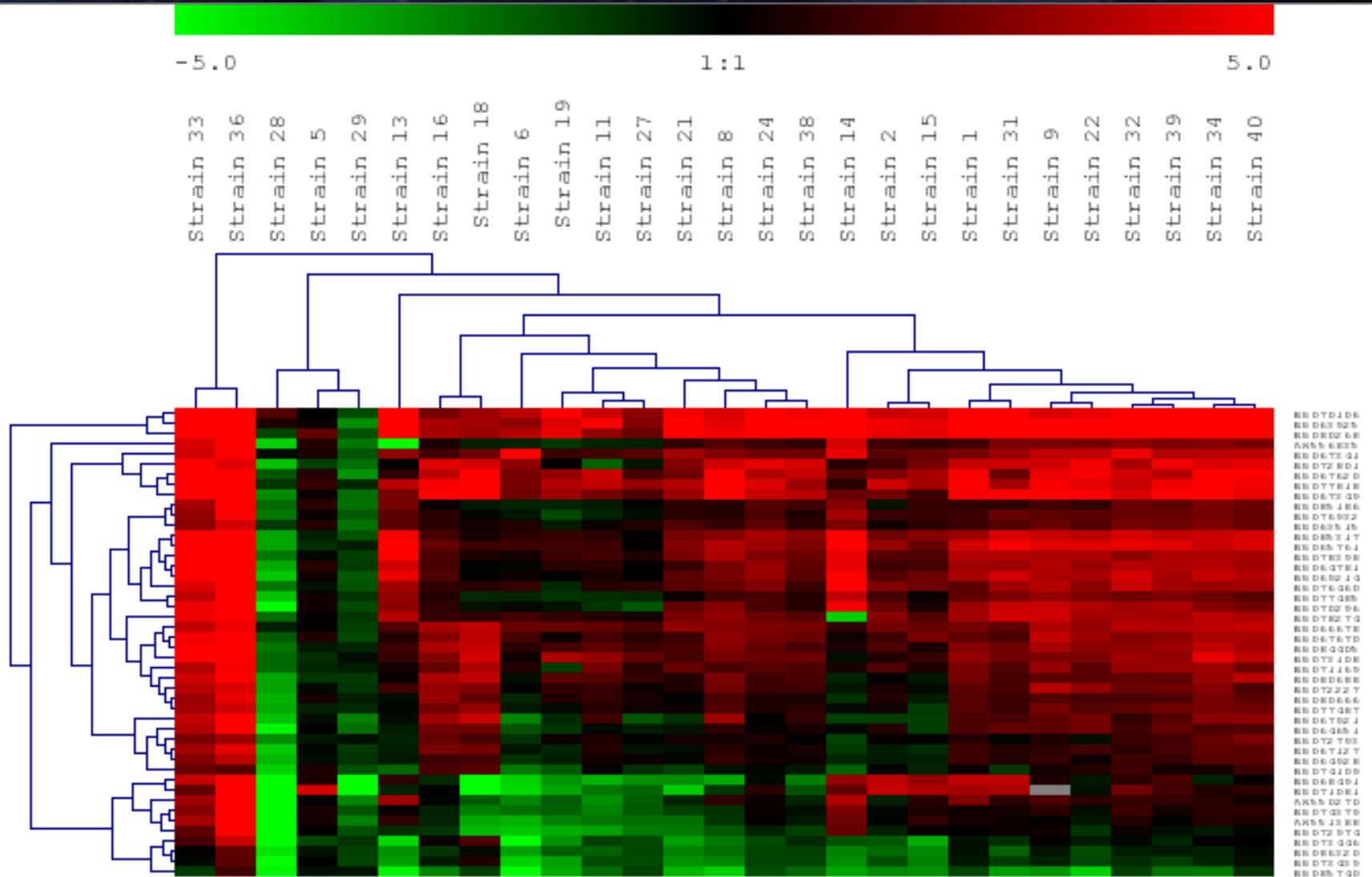
**Functional study of individual genes: Quantitative complementation of a QTL and analysis of association of SNP with QT**

**David Schwartz  
Gary Churchill  
Shuibang Wang  
Don Cook  
Gabe Howles  
Yonghong Wang  
Yan Yu  
Renee Gaspard**

# Going beyond the arrays

- We looked for mutations in the *tol4* receptor and found one in the BXD29 nonresponsive strain
- We used a variety of additional markers falling within our candidate QTL regions and genotyped our BXD strains. We were able to further refine the maps in those regions.
- We selected a set of candidate genes based on the arrays, the maps, and the functional roles of the genes and validated them by RT-PCR.
- We are now doing RNAi studies in cell culture to validate our hypotheses

# Candidate Gene Set for LPS response



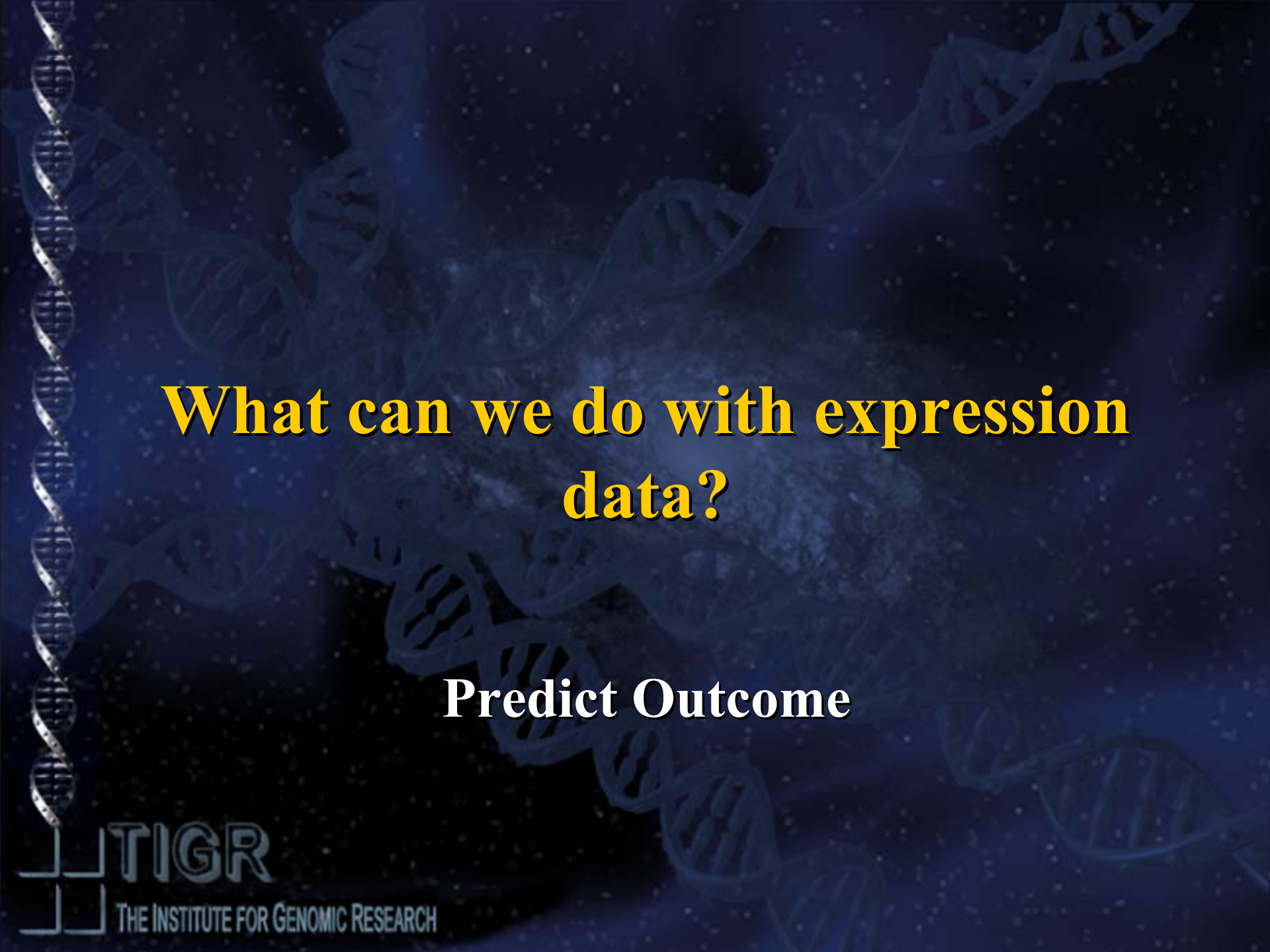
# Isn't this great?

## ■ Yes!

- Expression data and QTL data can be combined to find genes that are differentially expressed *and* that are likely to be important for the phenotype.
- The expression fingerprint itself can be used as a quantitative trait for genetic mapping.

## ■ No!

- We are likely to miss regulatory and signaling genes where sequence polymorphism may contribute to the phenotype.



# What can we do with expression data?

**Predict Outcome**

**TIGR**

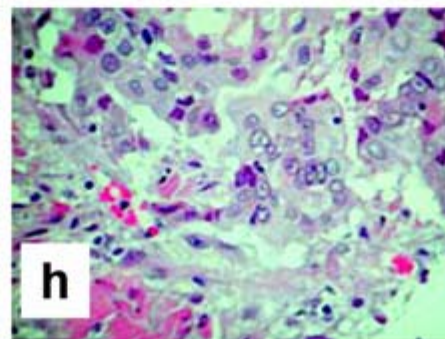
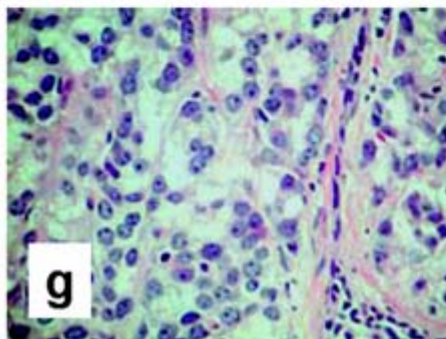
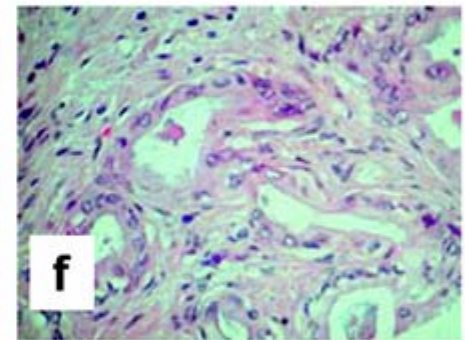
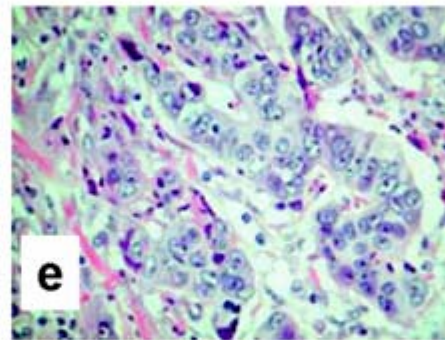
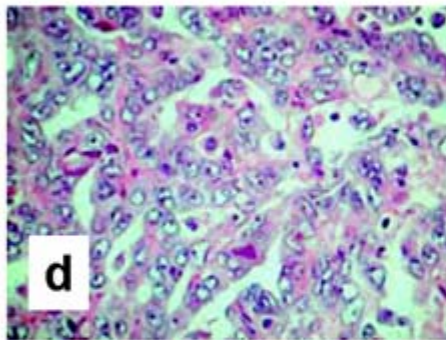
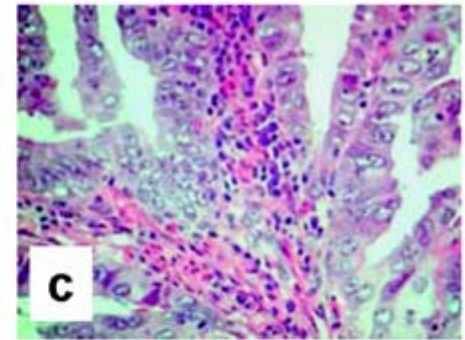
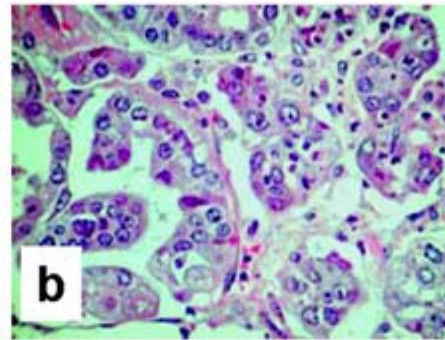
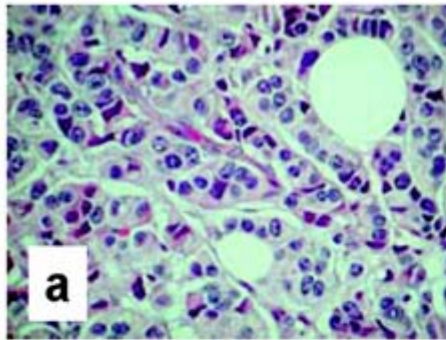
THE INSTITUTE FOR GENOMIC RESEARCH



# The problem

- Patients present with tumors, many of which are indistinguishable.
- Histology can provide some information, but these have little predictive power.
- Microarrays provide a “fingerprint” that can serve as a phenotypic measure that may be linked to outcome.
- This is a huge problem in data mining.

# The problem in pictures: Adenocarcinomas



a = Breast carcinoma  
b = Ovarian carcinoma  
c = Esophageal carcinoma  
d = Gastric carcinoma  
e = Colon carcinoma  
f = Pancreatic carcinoma  
g = Renal Cell carcinoma  
h = Lung carcinoma

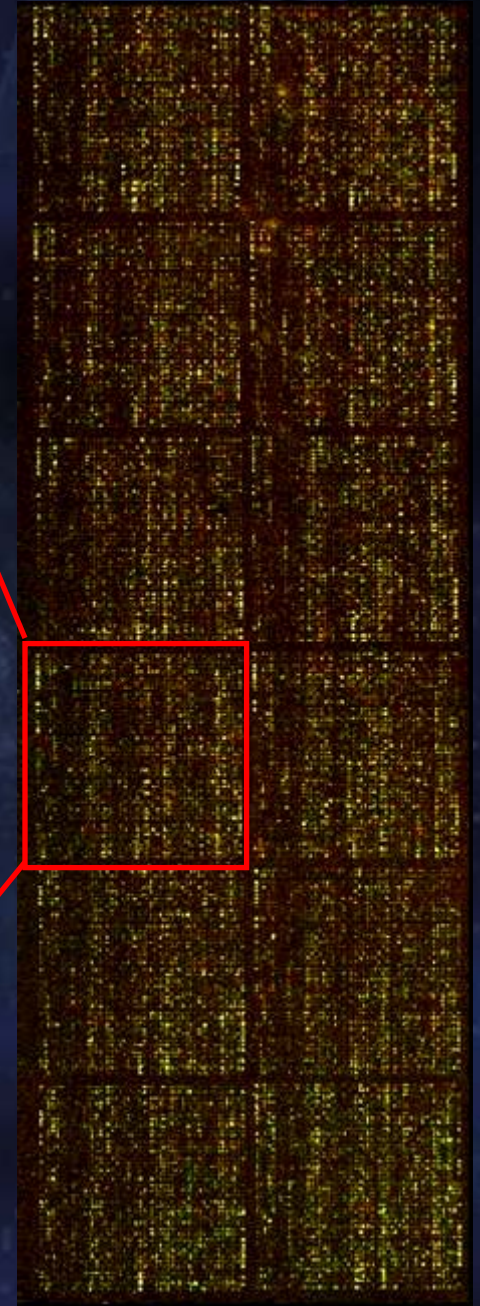
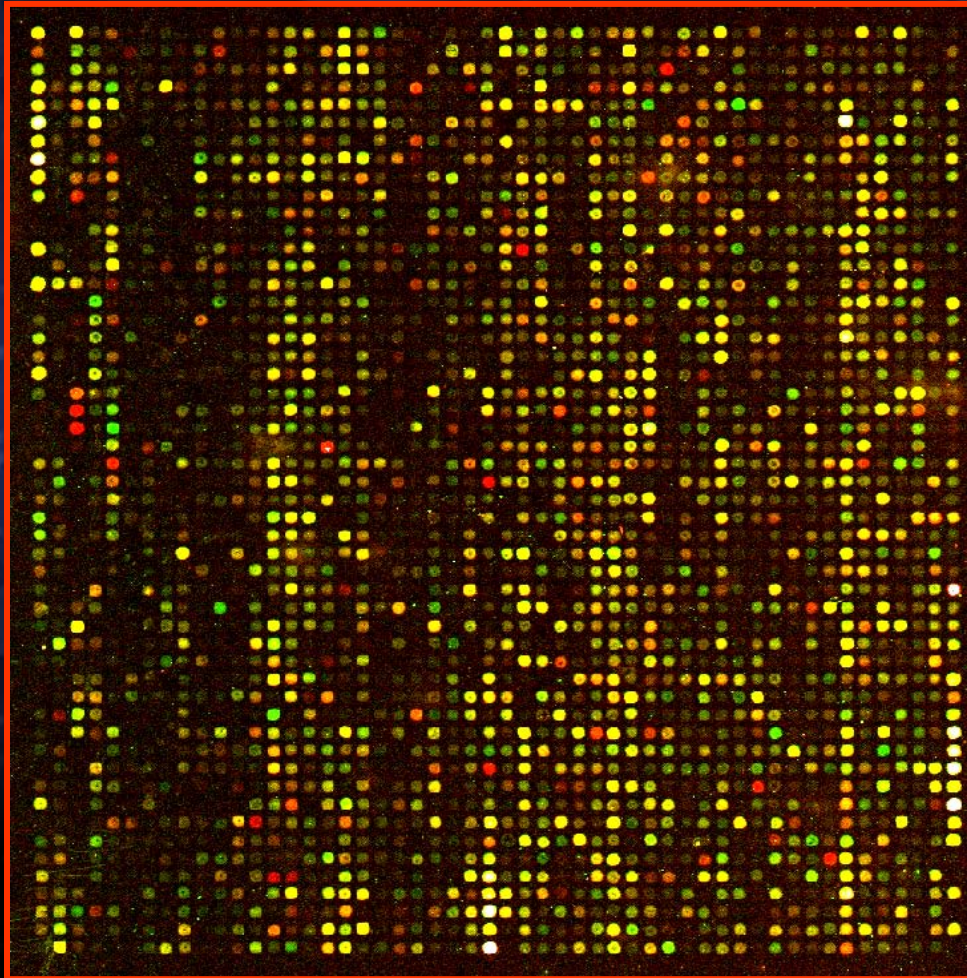


# Development of Molecular Organ Classifiers

 TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

# 32k Human Arrays



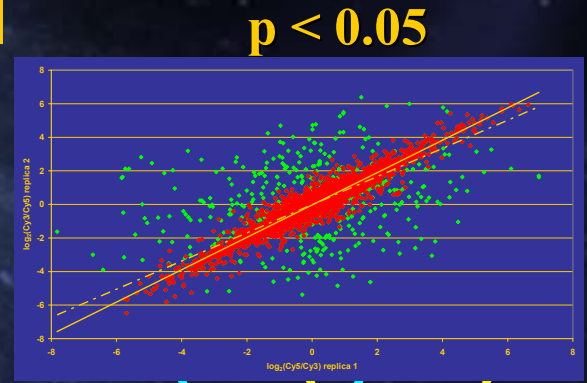
# TIGR Multi-Organ Cancer Classifier

77 tumor samples  
144 hybridization assays

individual array normalization

flip-dye replica  
consistency check

Statistical filtering of genes  
(Kruskal-Wallis H-test)  
685 genes



ovary lung

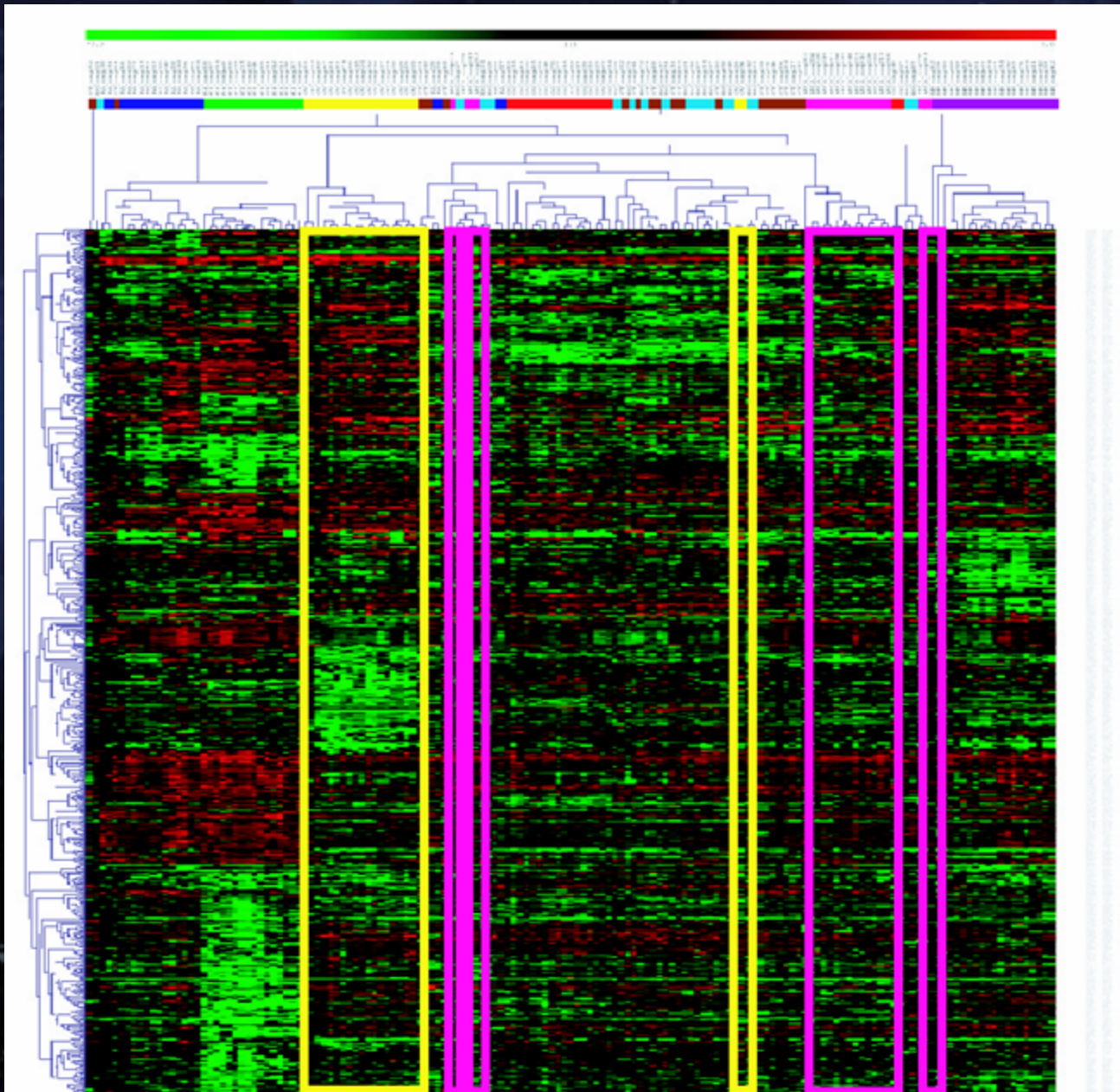
hierarchical clustering  
( Pearson correlation)

UNSUPERVISED  
CLASSIFICATION

Artificial neural network  
training and validation

SUPERVISED  
CLASSIFICATION

# Hierarchical Clustering of TIGR cDNA Data



— Ovary  
— Lung

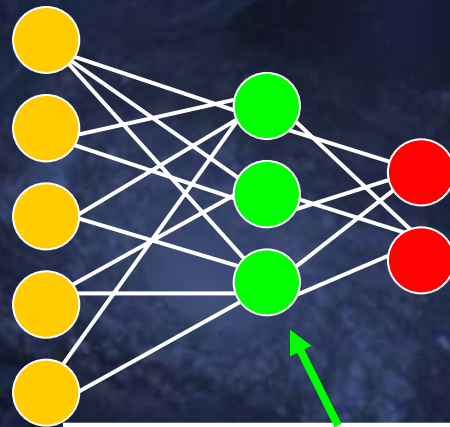
# Tissue-Specific Genes?

GenBank/TC/Role guess	Breast	Ovary	Colon	Stomach	EG junction	Pancreas	Lung	Kidney
AA280924 THC622568 carbonyl reductase (NADPH);	0.02	0.08	-0.29	-0.04	0.07	-0.11	1.08	0.15
AA429895 THC603542 MRP3; multidrug resistance protein 3	-0.26	-0.50	-0.15	-0.08	0.19	-0.01	0.87	0.09
AA056377 null null	-0.10	-0.10	0.19	-0.08	-0.11	0.00	1.31	-0.06
AA453898 THC650423 gal beta (1-3/1-4) GlcNAc alpha-2,3 sialyltransferase	0.01	0.05	-0.53	0.29	0.88	0.01	1.08	-0.60
AA505045 THC583342immunoglobulin heavy chain V(H)5 pseudogene	0.40	0.08	-0.14	0.39	0.58	0.22	1.49	0.25
AA972350 THC575287 pulmonary surfactant-associated proteolipid protein	-0.04	-0.04	-0.64	-0.05	-0.06	0.07	1.57	-0.18
AA864840 THC701676 KIAA1334 protein; novel retinal pigment epithelial cell protein	-0.11	0.06	-0.02	0.02	0.19	0.21	0.89	-0.01
AA677165 null null	2.22	-0.09	0.88	0.13	0.05	0.78	0.00	0.40
R54193 null null	2.20	0.13	-0.17	-0.01	-0.04	-0.11	0.04	-0.44
AA456975 null null	2.02	0.00	0.18	0.01	1.05	0.00	0.17	0.00
W90128 null null	1.90	-0.38	-0.40	-0.08	-0.02	-0.13	0.37	-0.86
R91803 THC570494 arylamine N-acetyltransferase; N-acetyltransferase-1	1.64	0.12	0.07	0.23	0.08	0.00	-0.02	0.00
AI628353 THC670392 KIAA0882 protein	1.63	0.06	-0.40	-0.04	0.14	0.11	0.06	0.10
H02231 THC615337 dJ483K16.1; long chain polyunsaturated fatty acid elongation enzyme	1.32	-0.11	-0.24	-0.55	-0.72	-0.27	0.40	-0.10
H29407 THC623058 LIV-1, breast cancer/estrogen regulated	1.25	-0.50	-0.34	-0.04	-0.30	-0.17	-0.10	-0.19
AA446839 THC570204 E1B 19K/Bcl-2-binding protein	-0.10	0.44	-0.15	-0.85	-0.85	-0.32	0.13	2.11
AA682423 THC688753 monoamine oxidase B	-0.16	-0.12	-0.57	-0.37	-0.43	0.09	-0.08	2.05
W84778 THC569476 NADH:ubiquinone oxidoreductase	0.08	-0.29	-0.86	-0.34	-0.25	0.09	-0.26	2.27
W85851 null null	-0.39	-0.32	-0.05	-0.46	-0.41	-0.46	-0.39	3.10
AA700054 THC602328 adipophilin	-0.78	-0.21	0.86	-0.53	-0.41	-0.70	-0.48	2.81
AA670144 null null	-0.16	-0.42	0.07	-0.39	-0.38	-0.13	-0.11	1.90
AA504943 THC601867 crystallin, alpha B	-0.13	-0.37	-0.44	-0.41	-0.42	-0.26	-0.54	3.21
W30988 THC682953 hepatic angiopoietin-related protein	-0.36	-0.78	-0.12	-0.37	-0.48	-0.18	-0.18	3.09
AI341427 null null	-0.29	1.31	-0.14	-0.39	-0.18	0.12	-0.10	-0.42
AA427924 null null	-0.68	2.30	-0.69	-0.72	-0.37	0.09	-0.21	0.36
H09099 THC702730 KIAA0762 protein; VSGP/F-spondin	-0.37	1.93	-1.41	-0.46	-0.43	0.21	0.09	0.55
AA001444 null null	-0.24	1.24	-0.24	-0.24	-0.20	0.14	-0.19	-0.34
AA865464 THC601987 retinoic acid induced gene E	-0.18	0.81	-0.47	0.07	0.12	-0.08	-0.25	-0.03
AA872323 THC583082 unnamed protein product	-0.13	0.95	-0.44	-0.23	-0.43	0.15	0.08	0.07
AI364369 THC616465 PBX1a; homeobox-containing protein; pre-B-cell leukemiaTF 1	0.40	1.48	-0.12	-0.46	-0.56	-0.05	-0.21	-0.59
N54596 null null	0.43	1.36	-0.61	0.18	-0.65	0.08	-0.09	-0.88
AI382830 THC590306 procarboxypeptidase B	-0.21	-0.21	-0.21	-0.21	-0.21	1.44	-0.21	-0.21
AA703660 THC564885 Histone H2A related	-0.23	-0.08	-0.13	-0.23	-0.16	1.35	-0.24	-0.35
AA155695 THC669946 transcobalamin I	-0.19	-0.24	-0.18	0.08	-0.24	1.19	-0.23	-0.24
AA284528 THC608762 trypsinogen II	-0.15	-0.13	0.07	-0.27	-0.13	1.21	-0.27	-0.41
NG8543 null null	-0.29	-0.14	-0.19	-0.13	-0.17	1.06	-0.08	-0.07
R38933 THC694690 plasminogen activator, tissue	0.25	0.02	-0.43	0.04	-0.06	1.09	-0.32	-0.74
AA425422 null null	0.08	-0.15	0.02	-0.10	-0.18	0.82	-0.21	-0.35

# Neural Networks and Cancer

## Input data:

A list of genes with expression levels



## Output data:

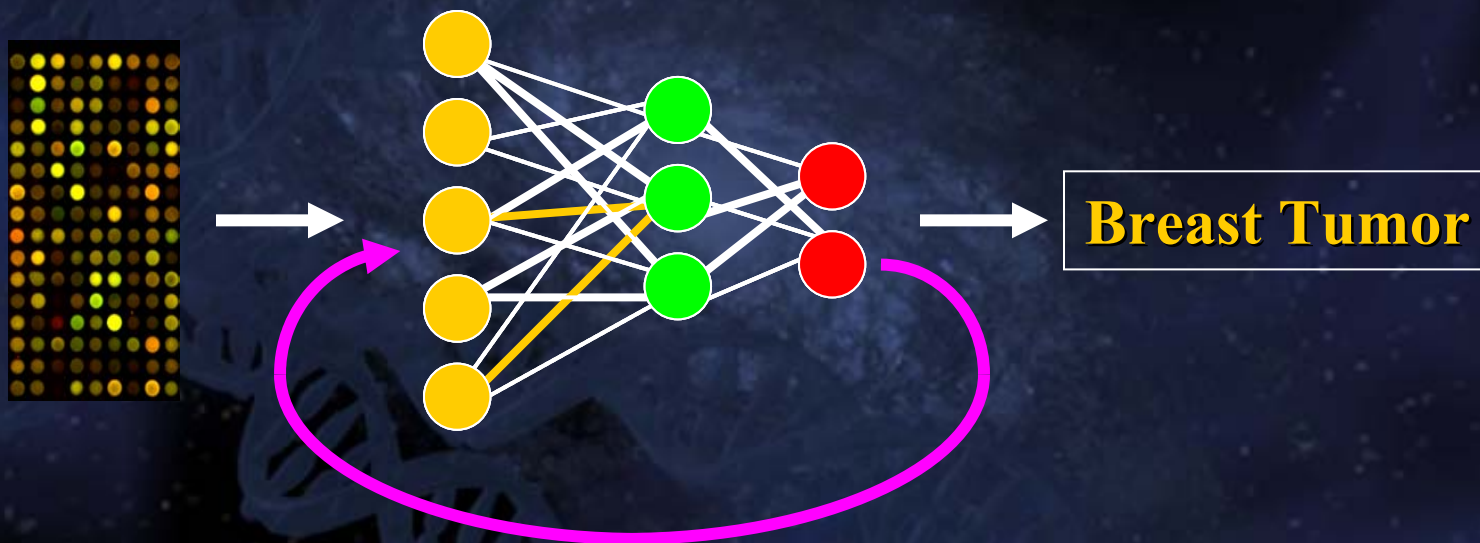
A tumor type call

“hidden layers” allow complex connections

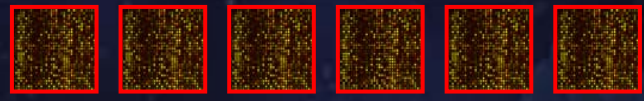


# Neural Networks and Cancer

Training:  
Adjusts weights  
and connections



# Data Acquisition



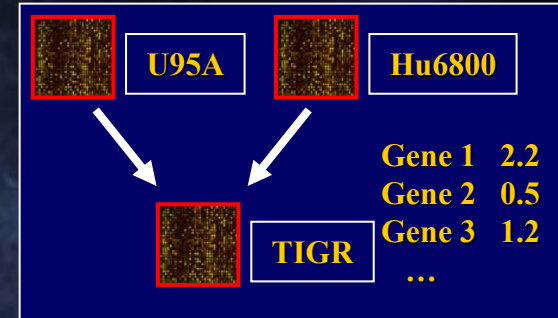
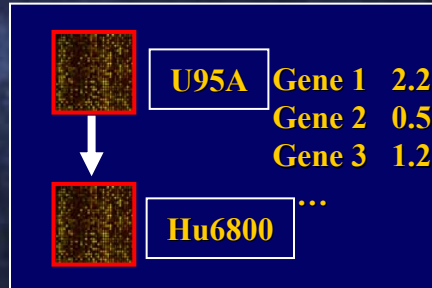
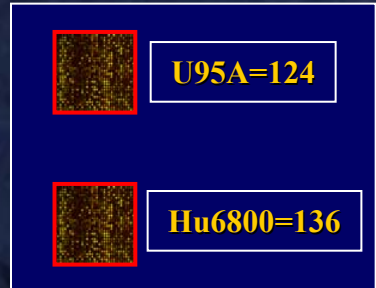
## Microarray Database

# Normalization and Scaling

Average Across Chips using Reference

Gene-by-Gene using Reference

Gene-by-Gene using Reference



# Statistical Screening

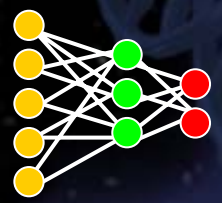
All Normalized and Scaled Genes

Kruskal-Wallis Bonferoni  $f(x)$

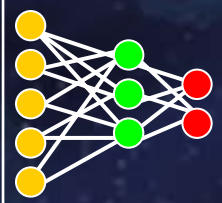
Correlative Gene Subset

# Neural Network Training and Validation

Training Set  
Tumor 1  
Tumor 2  
Tumor 3  
Tumor 4  
Tumor 5  
...  
Tumor n



Test Set  
Tumor 1  
Tumor 2  
Tumor 3  
Tumor 4  
Tumor 5  
...  
Tumor n



Classifier

# Summary

- **We collected 540 expression profiles**
  - **21 tumor types**
  - **95% of all cancers**
- **10 Independent Classifiers**
  - **75% of data for training, 25% for test**
  - **Average ~88% accuracy**
- **Classifier has been validated on an independent set of colon cancer samples and mets with 90% accuracy**
- **Web based Classifier available**
  - **So far, 7 of 8\* in classification**

# Isn't this great?

## ■ Yes!

- **Demonstration of cross-platform comparisons**
- **Represents a potential tool to assist in clinical diagnosis**

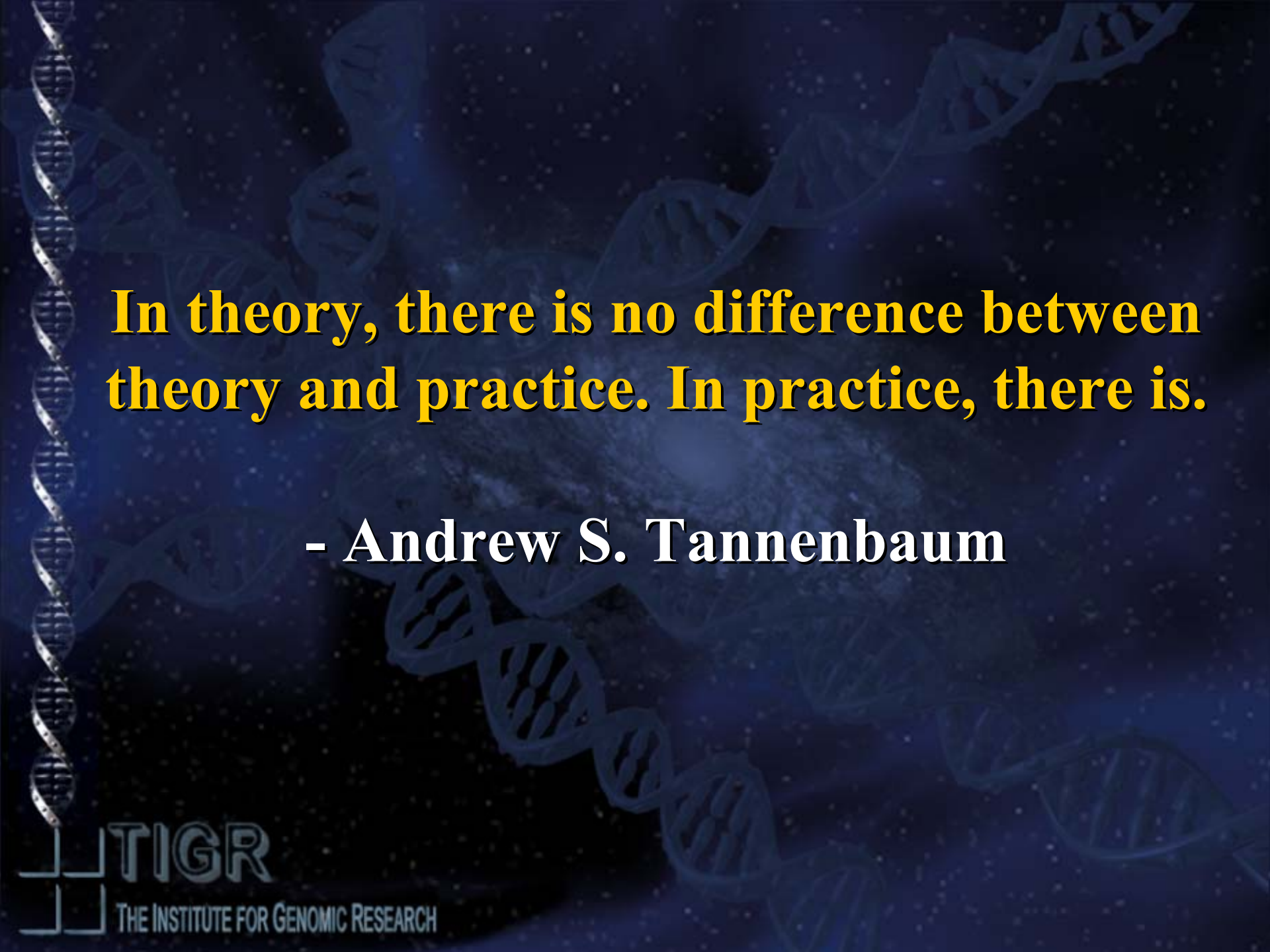
## ■ No!

- *Much* work remains to be done to actually demonstrate that this has true clinical relevance
- What we really want to do is extend this to survival and response to therapeutics
- We need *Much* more data

● **One tumor, one chip™**

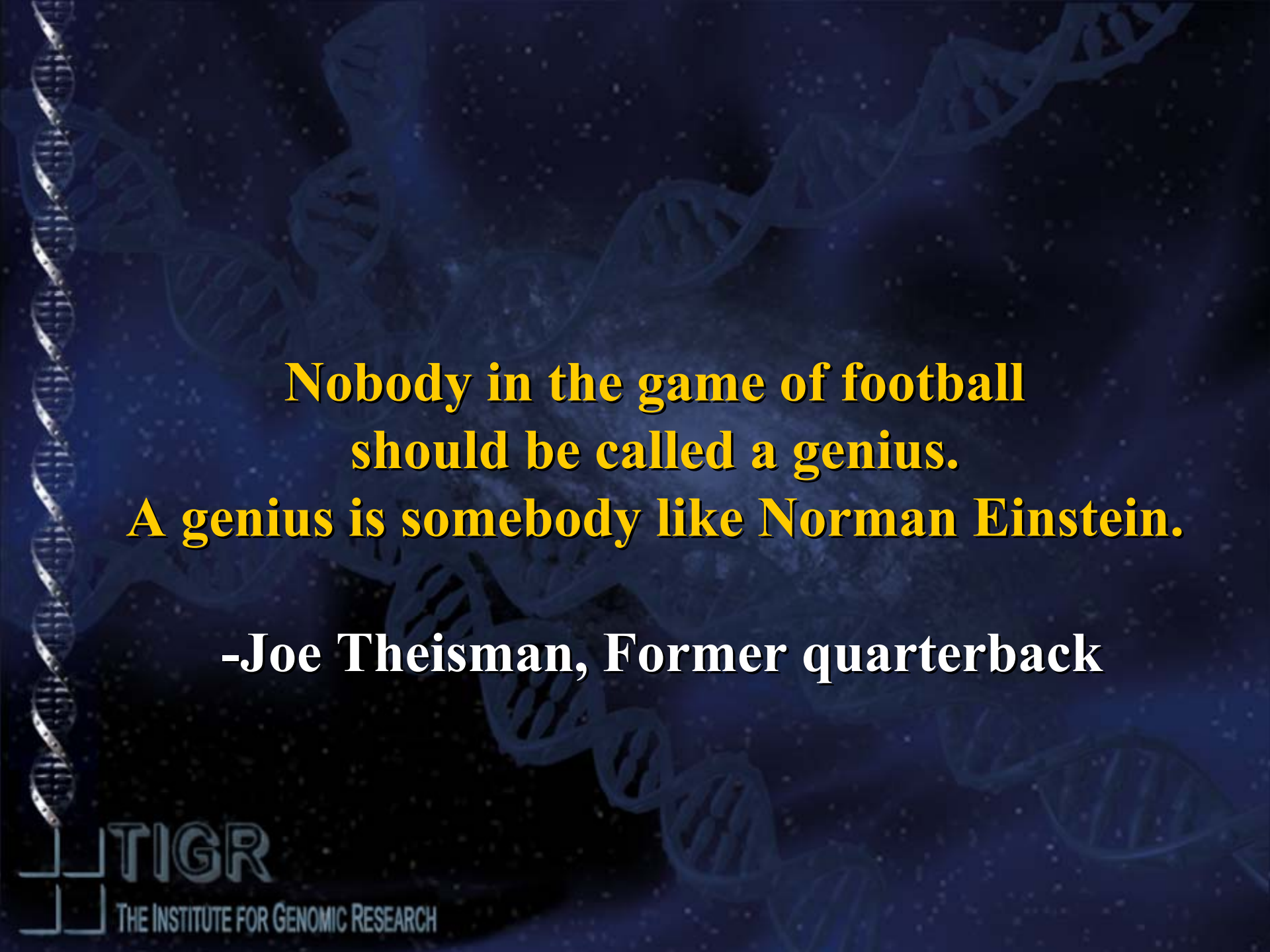
# Where are we going?

- **Array analysis has matured significantly in the past few years.**
- **The bottleneck in array studies is rapidly becoming data analysis and interpretation.**
- **The challenge now is to intelligently integrate expression data with other sources of biological knowledge to turn our gene sets into something approaching biology.**
- **Array-based technologies are poised to make the transition from the laboratory to the clinic.**
- **What we need are consistent means of collecting and archiving the data so that more comprehensive data mining can take place.**

The background is a dark blue gradient with a subtle pattern of DNA double helix structures. A faint, semi-transparent image of a person's face is visible in the upper left quadrant. The main text is centered and rendered in a bold, yellow, sans-serif font.

**In theory, there is no difference between theory and practice. In practice, there is.**

**- Andrew S. Tannenbaum**



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

**-Joe Theisman, Former quarterback**

# Acknowledgments

[<johnq@tigr.org>](mailto:johnq@tigr.org)

## The TIGR Gene Index Team

Foo Cheung

Svetlana Karamycheva

Yudan Lee

Babak Parvizi

Geo Pertea

Razvan Sultana

Jennifer Tsai

John Quackenbush

Joseph White

### Emeritus

Jennifer Cho (TGI)

Ingeborg Holt (TGI)

Feng Liang (TGI)

Kristie Abernathy ( $\mu$ A)

Sonia Dharap ( $\mu$ A)

Julie Earle-Hughes ( $\mu$ A)

Cheryl Gay ( $\mu$ A)

Priti Hegde ( $\mu$ A)

Rong Qi ( $\mu$ A)

Erik Snesrud ( $\mu$ A)

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

## H. Lee Moffitt Center/USF

Timothy J. Yeatman

Greg Bloom

### PGA Collaborators

Gary Churchill (TJL)

Greg Evans (NHLBI)

Harry Gavaras (BU)

Howard Jacob (MCW)

Anne Kwitek (MCW)

Allan Pack (Penn)

Beverly Paigen (TJL)

Luanne Peters (TJL)

David Schwartz (Duke)

### TIGR PGA Collaborators

Norman Lee

Renaë Malek

Hong-Ying Wang

Truong Luu

Bobby Behbahani

## TIGR Human/Mouse/Arabidopsis

### Expression Team

Emily Chen

Bryan Frank

Renee Gaspard

Jeremy Hasseman

Heenam Kim

Lara Linford

Simon Kwong

John Quackenbush

Shuibang Wang

Yonghong Wang

Ivana Yang

Yan Yu

### Array Software Hit Team

Nirmal Bhagabati

John Braisted

Tracey Currier

Jerry Li

Wei Liang

John Quackenbush

Alexander I. Saeed

Vasily Sharov

Mathangi Thalaragian

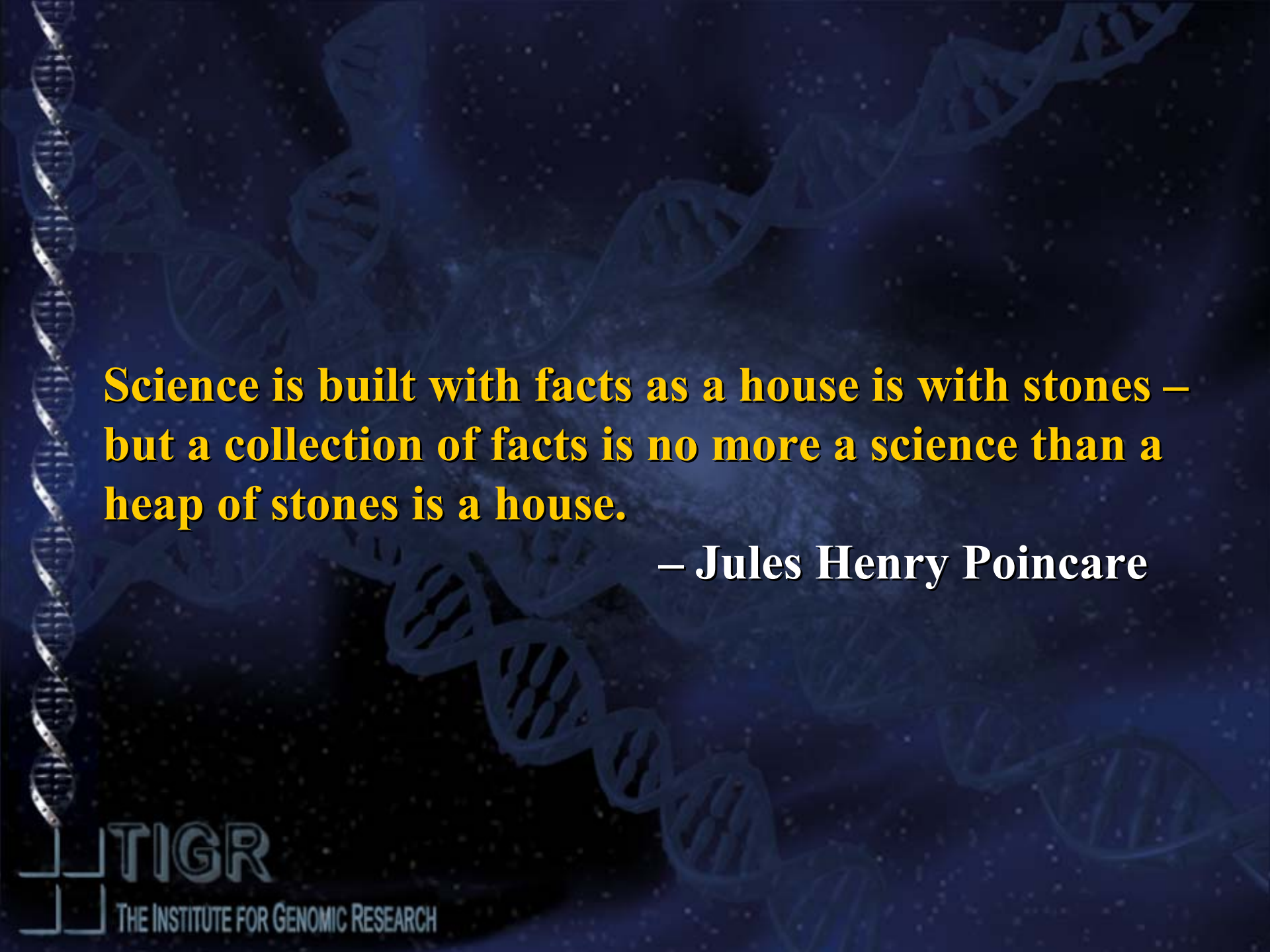
Joseph White

### Assistant

Mary Mulholland

TIGR Faculty, IT Group, and Staff





**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