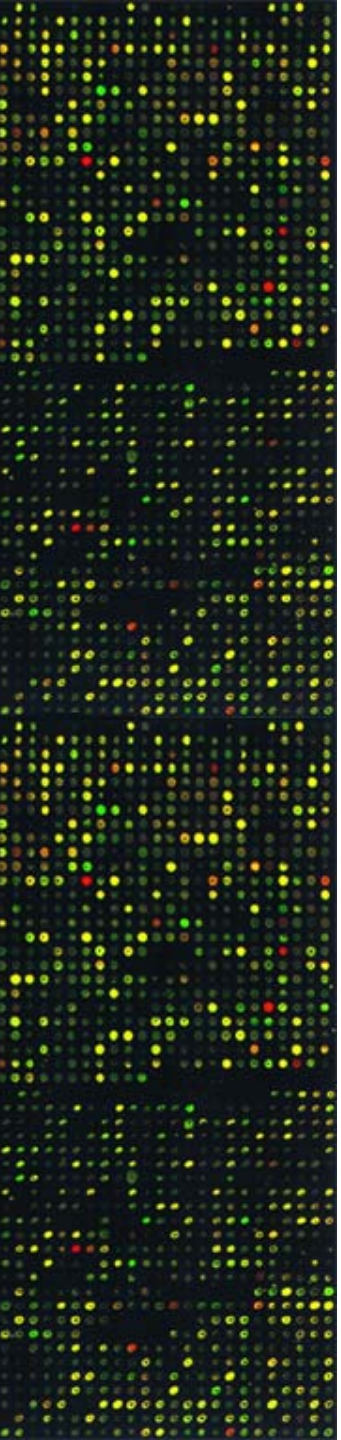
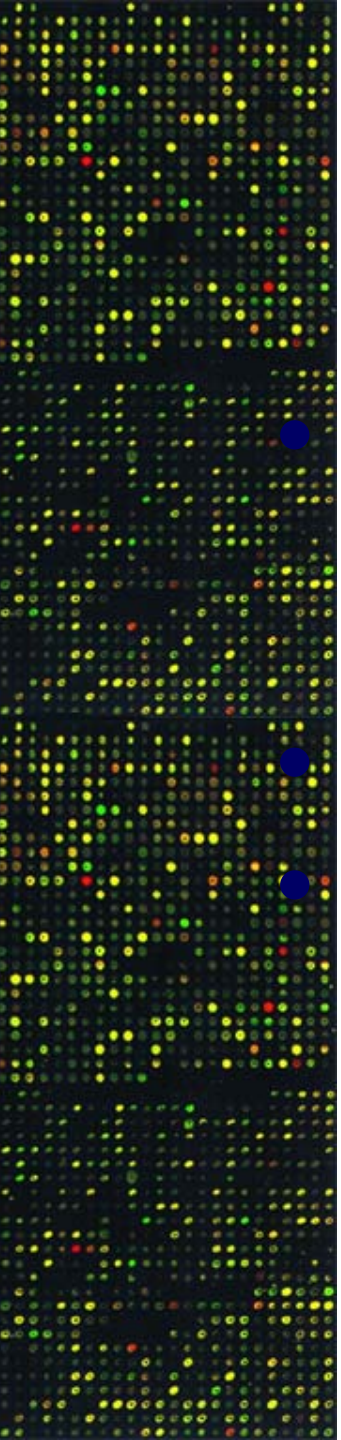


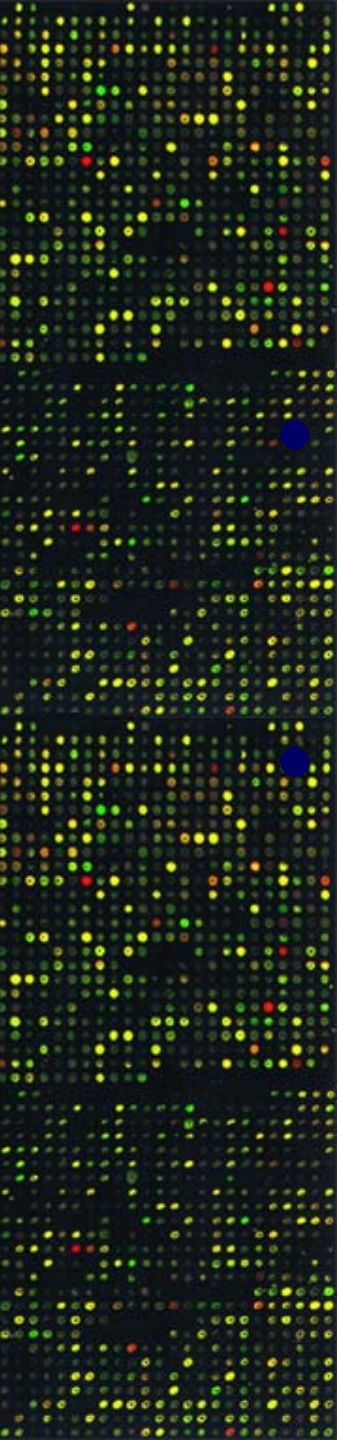
Leveraging the Literature For Gene Expression Analysis

Daniel R. Masys, M.D.
Director of Biomedical Informatics
Associate Professor of Medicine
University of California, San Diego



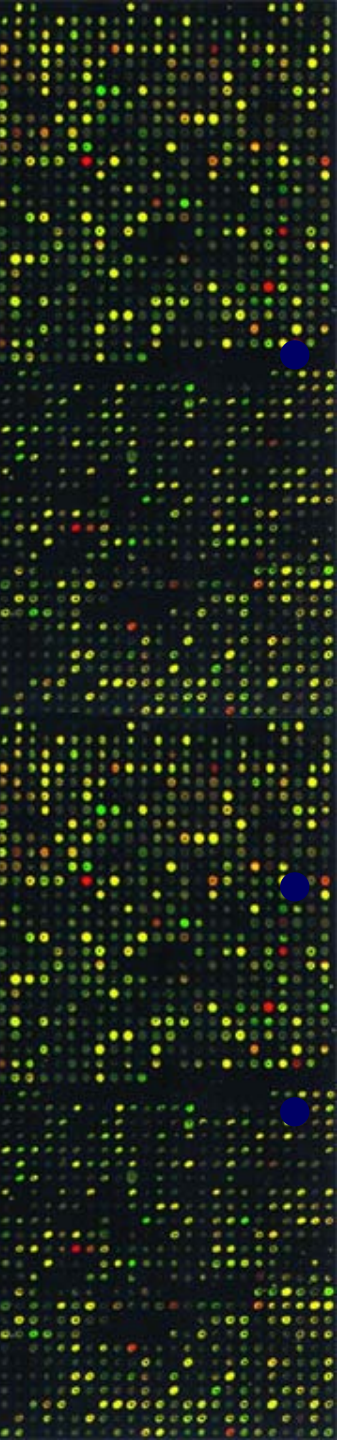
Characteristics of Array Data

- Voluminous – tens of thousands of variables with relatively few observations of each
- Noisy
- Methods designed to detect patterns and associations always find patterns and associations



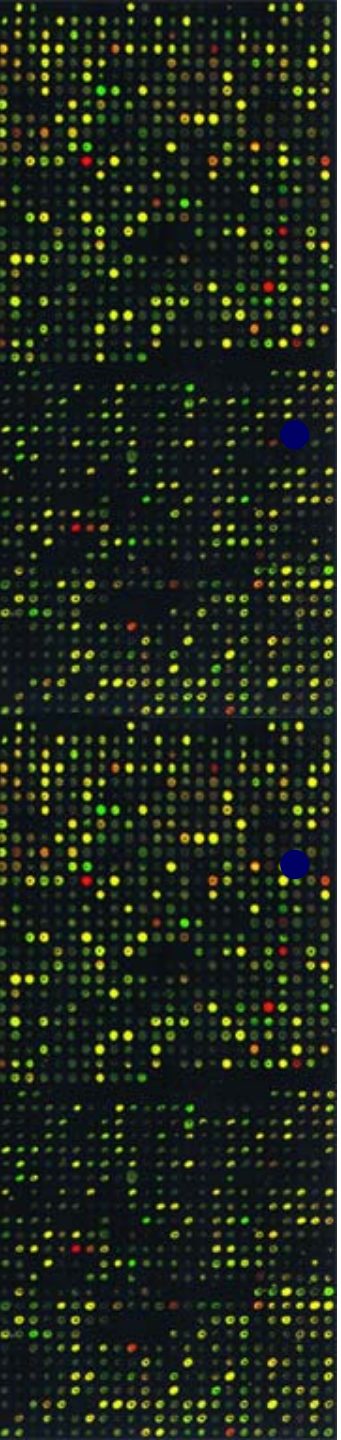
General approaches to microarray analysis

- Quantitative analysis: what are the similarities among genes based on numerical values for expression levels
- Semantic analysis: what do those quantitative patterns mean in terms of biology?



Challenges of microarray interpretation

- Most mathematical approaches to grouping genes (whether informed by biological knowledge or not) yield gene expression clusters that must then be manually inspected and evaluated
- Unusual for a researcher to recognize all genes in a cluster
- Genes may be clustered because of a variety of functional similarities, some not apparent to the viewer

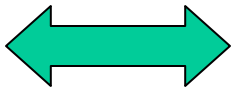


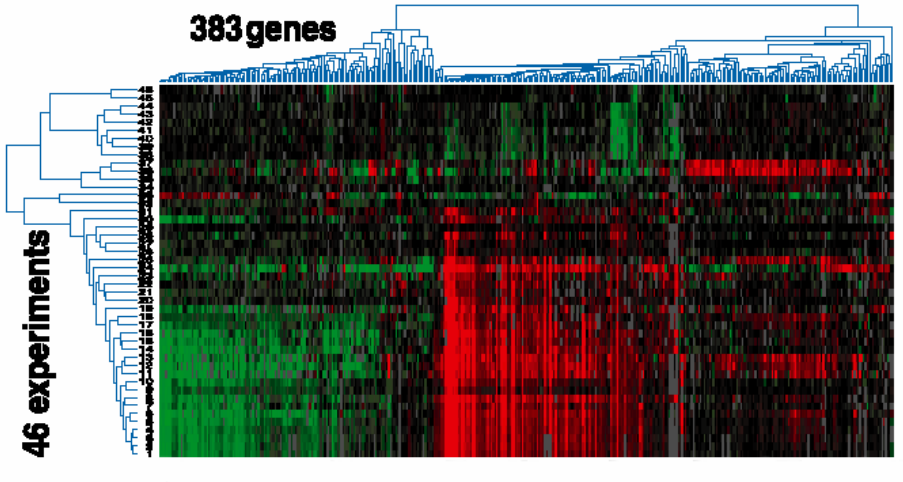
Data Mining

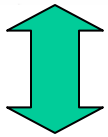
- Data Mining is the process of finding new and potentially useful knowledge from data, by finding patterns and associations
- Generally uses methods to join heterogeneous data sources using linking methods

A central issue: how to detect useful associations

The Biomedical Literature







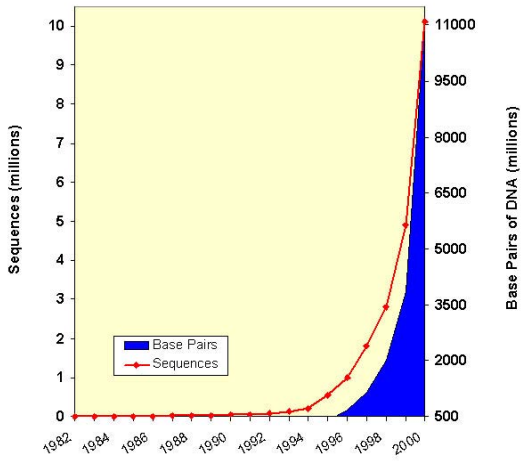
My Expression Data

MEDLINE:

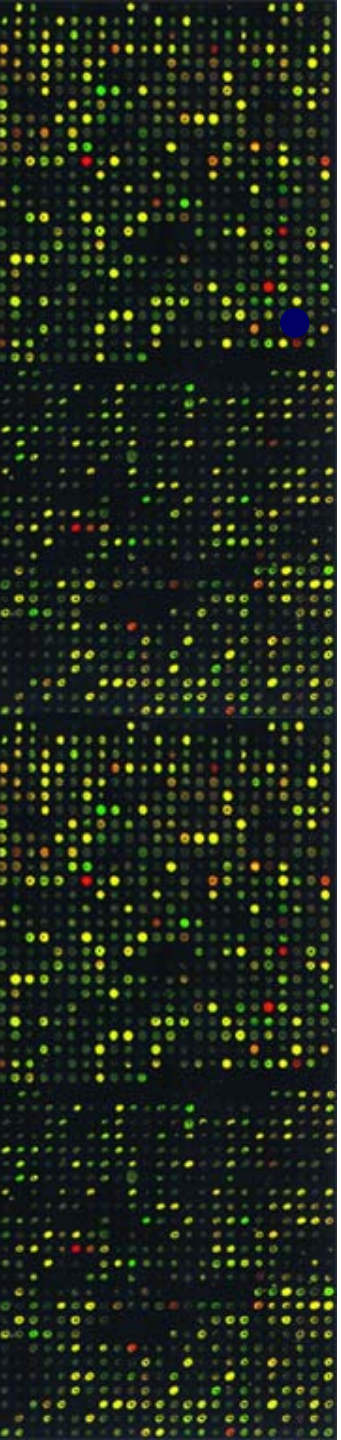
12 Million citations

Growing at 400,000 new articles per year

20.2 Billion DNA base pairs as of September 2002

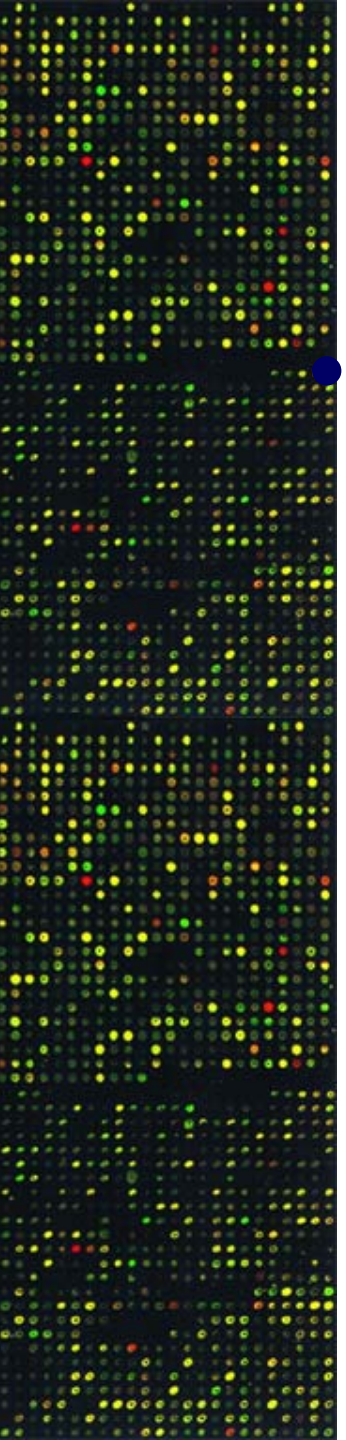


Genomics Databases



Data Mining Approaches

- Codes and Unique Identifiers
  - Require consistent standards and a responsible organization
  - Good for known entities but not new discoveries



Sources of codes and Unique Identifiers for data mining

- NCBI, publicly available
  - GenBank - individual gene sequences and partial sequences
  - Unigene - functionally similar gene units based on sequence similarity
  - LocusLink - cross references of gene Ids and names
  - RefSeq - Reference sequences
  - OMIM: Online Mendelian Inheritance in Man - interface between clinical and molecular genetics

HGNC Database Links

Human Other species Sequence Mutation, biochemical & other Gene/Protein families Nomenclature

Genome: Human

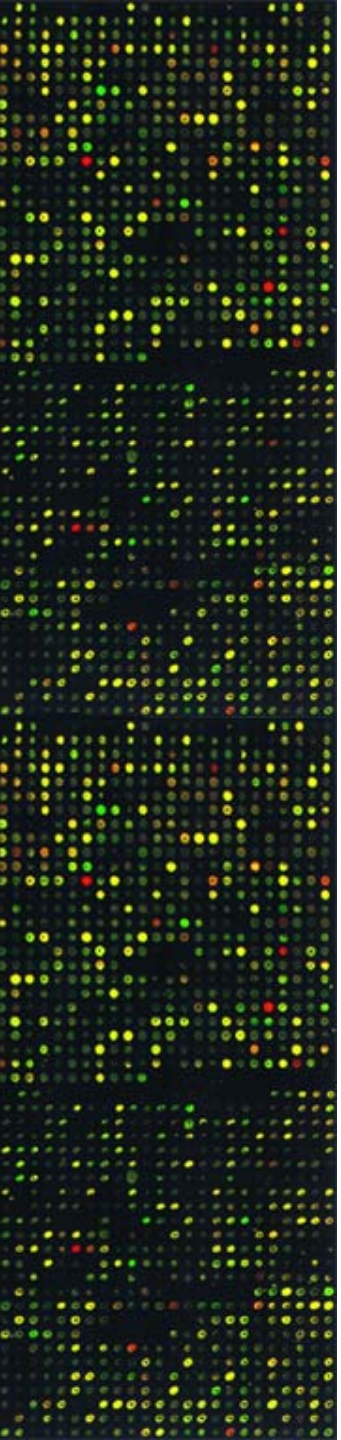
- HUGO Gene Nomenclature Committee Homepage
- GDB Genome Database (USA)
- OMIM Online Mendelian Inheritance in Man (USA)
- GENATLAS
- GeneCards integrated database
- LocusLink interface
- Ensembl annotation database
- Human Genome Project Working Draft
- euGenes: Human

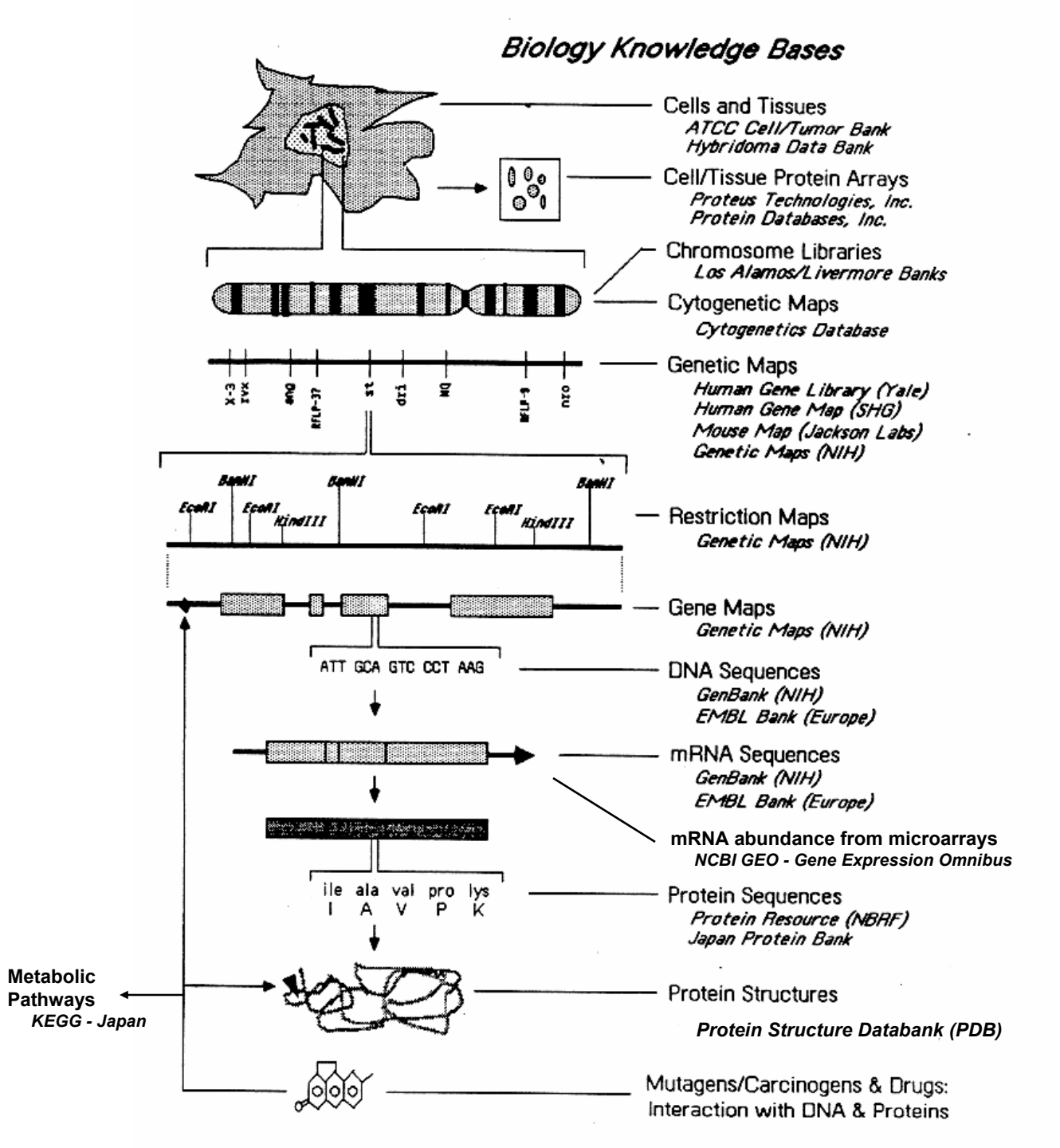
Genome: Other Species and Comparative

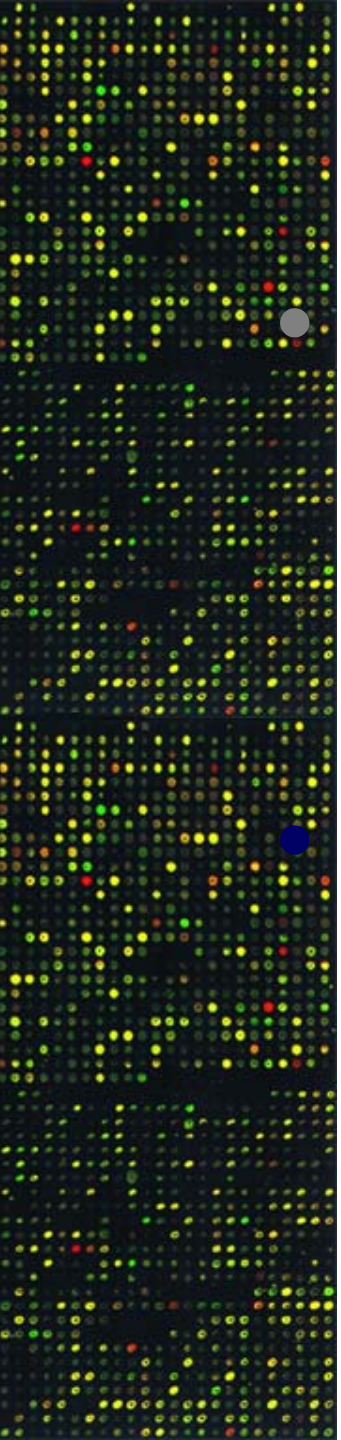
. good resource is Nucleic Acids Research 2002: Database Issue

Vertebrates

- MGD Mouse Genome Database (USA) or HGMP-MGD Mirror (UK)
- ARKdb species databases includes: Cat, Chicken, Cow, Deer, Horse, Pig, Salmon, Sheep, Tilapia, Turkey
- OMIA Online Mendelian Inheritance In Animals
- RATMAP
- Livestock genome databases ARK (UK) and in USA
- BOVMAP (France)
- DogMap
- The Swine Genome Map (USA) or PiGMaP project (UK)
- Chicken Genome Mapping Project (UK)
- Fugu project
- Zebrafish

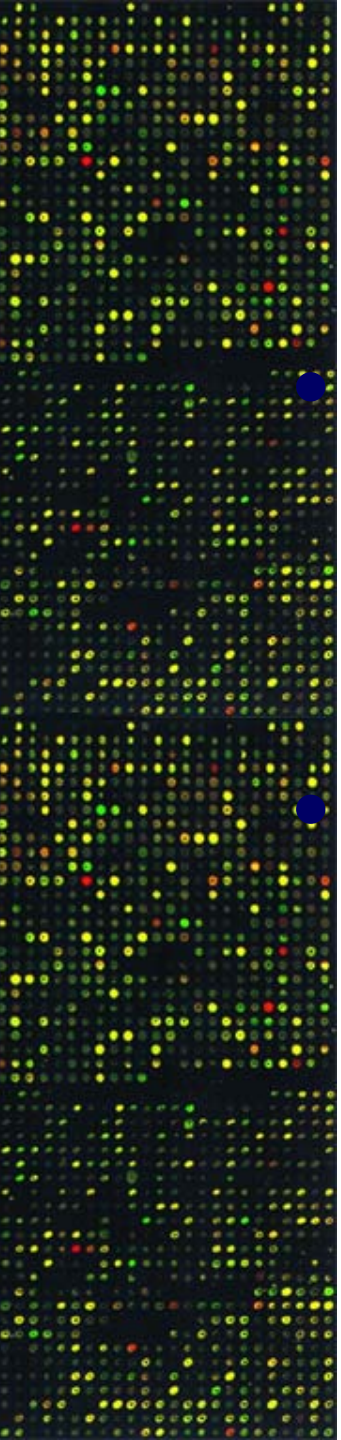



Biology Knowledge Bases
Cells and Tissues
ATCC Cell/Tumor Bank
Hybridoma Data Bank
Cell/Tissue Protein Arrays
Proteus Technologies, Inc.
Protein Databases, Inc.
Chromosome Libraries
Los Alamos/Livermore Banks
Cytogenetic Maps
Cytogenetics Database
Genetic Maps
Human Gene Library (Yale)
Human Gene Map (SHG)
Mouse Map (Jackson Labs)
Genetic Maps (NIH)
BamHI
EcoRI
HindIII
Restriction Maps
Genetic Maps (NIH)
Gene Maps
Genetic Maps (NIH)
ATT GCA GTC CCT AAG
DNA Sequences
GenBank (NIH)
EMBL Bank (Europe)
mRNA Sequences
GenBank (NIH)
EMBL Bank (Europe)
mRNA abundance from microarrays
NCBI GEO - Gene Expression Omnibus
ile ala val pro lys
I A V P K
Protein Sequences
Protein Resource (NBRF)
Japan Protein Bank
Metabolic Pathways
KEGG - Japan
Protein Structures
Protein Structure Databank (PDB)
Mutagens/Carcinogens & Drugs:
Interaction with DNA & Proteins



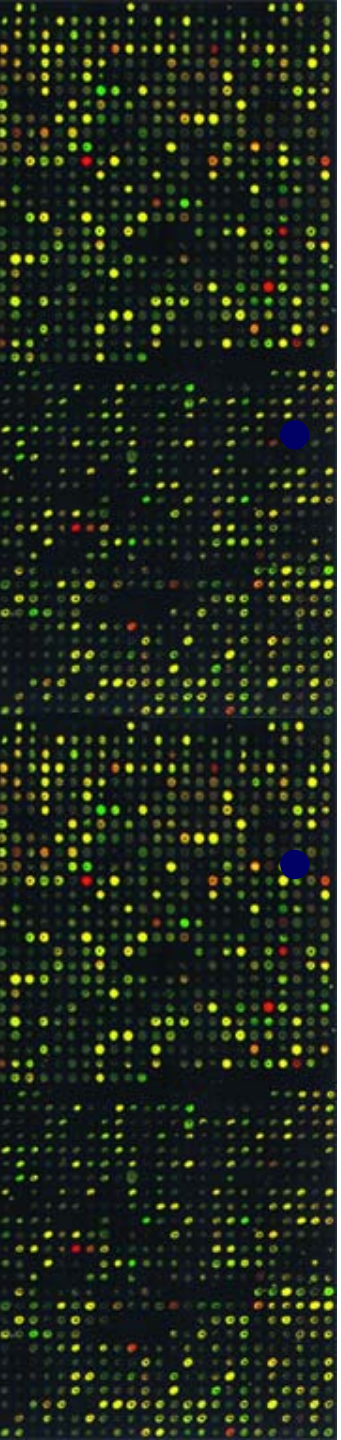
Data Mining Approaches

- Codes and Unique Identifiers
  - Require consistent standards and a responsible organization
  - Good for known entities but not new discoveries
- Language-based linkages
  - Names and abbreviations (e.g., HTLV1)
  - Keywords and terms (e.g., "infectious diseases"
  - Computational linguistics (e.g., automated reading of the literature)



Limits to data mining

- Synonymy: many ways to refer to the same object or concept
  - "The boundless chaos of living speech…"
- Polysemy: a word or concept may have multiple meanings
  - e.g., insulin is a gene, a protein, a hormone, a therapeutic agent
  - "CAT" - Hugo approved gene symbol for catalase



Linking Gene Expression results to the published literature

- Since 1987 National Library of Medicine has made GenBank accession numbers searchable keywords for retrieving articles describing specific genes
- Enables data mining to characterize gene groups by the distribution of keywords from the literature that has been published about the genes in the group

Linking Gene Expression results to the published literature

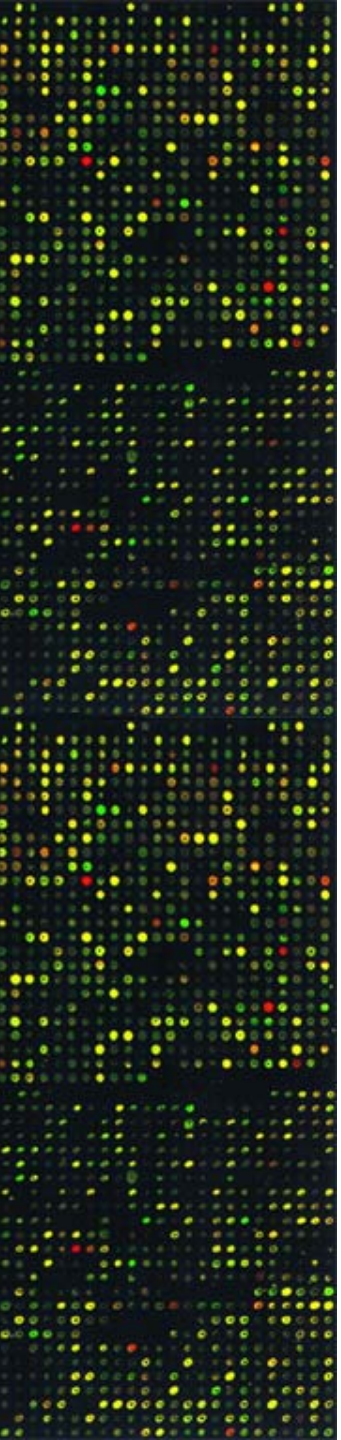
GenBank Accession List

Published MEDLINE citations

Combined list of keyword descriptors:
Medical Subject Heading (MeSH terms)
IUPAC Enzyme Nomenclature Registry Numbers

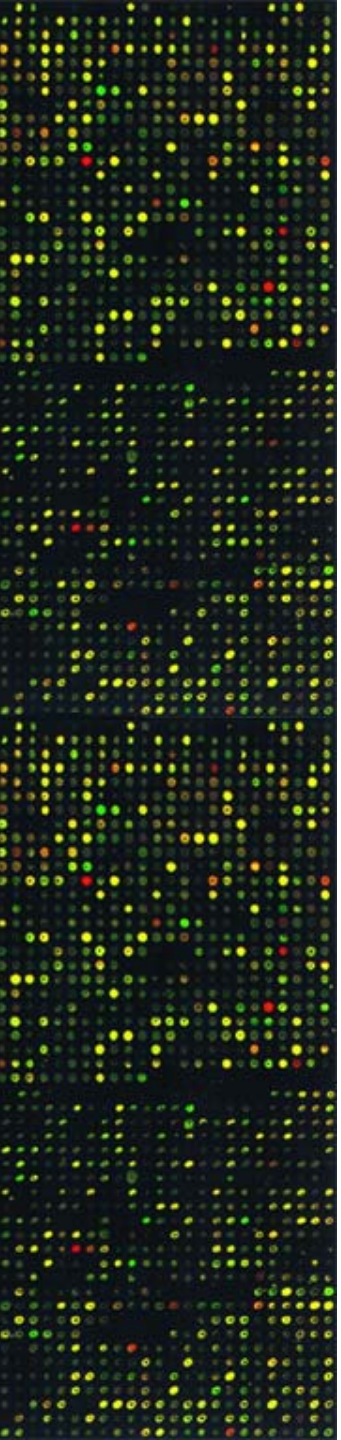
Medical Subject Headings (MeSH) Vocabulary

- 19,000 main concepts (300,000 synonyms)
- 103,500 chemical terms
- Arranged in 16 different concept hierarchies
- Include a separate hierarchy of IUPAC Enzyme Commission Registry Numbers



MeSH terminology concept hierarchies

- Anatomy
- Organisms
- Diseases
- Chemicals & Drugs
- Analytical Techniques
- Psychiatry & Psychology
- Biological Sciences
- Physical Sciences
- Anthropology & Social Sciences
- Technology, Food
- Information Science
- Humanities
- Persons
- Healthcare
- Geographic locations



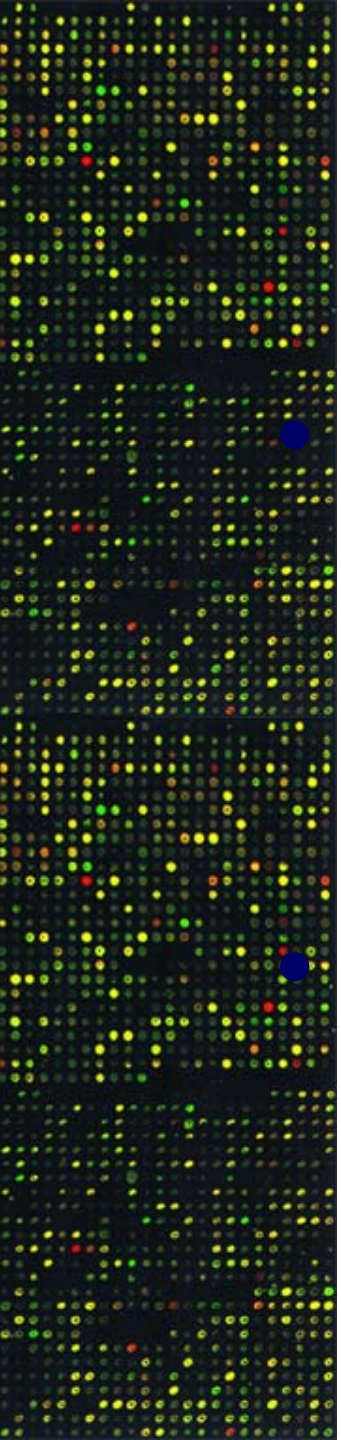
Sample MeSH "is-a" hierarchies

Diseases

- Nervous System Diseases
  - Demyelinating diseases
    - Multiple Sclerosis

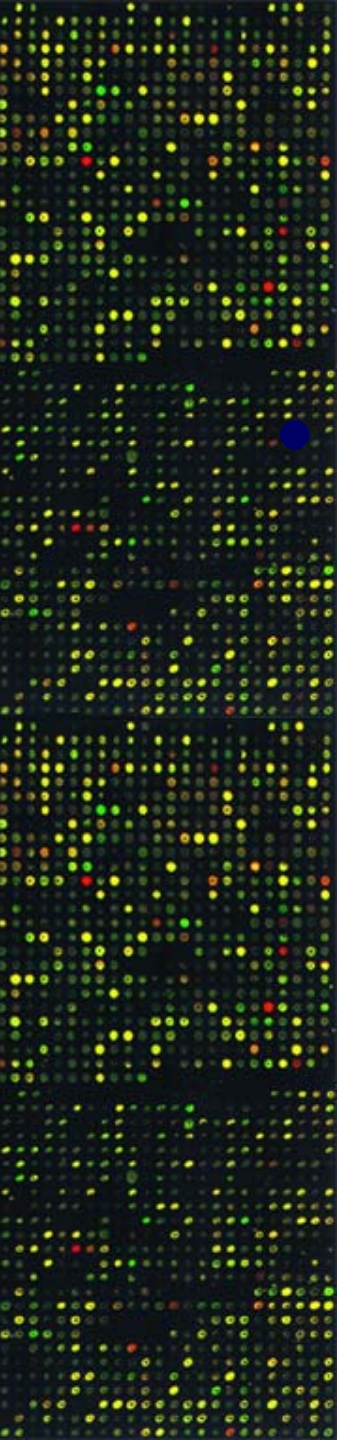
Enzymes

- Complement Activating Enzymes
  - Endopeptidases
    - Plasminogen Activators
      - Pancreatic Elastase



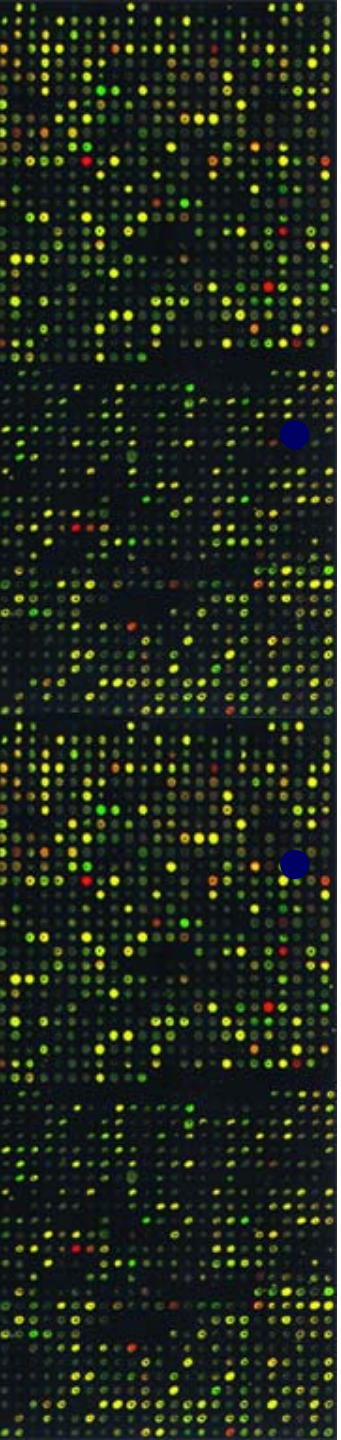
Why use hierarchies?

- Human indexer variability (r value = 0.6 for correlation of main indexing terms assigned to a given publication by different indexers, r=0.4 for minor keywords)
- Biological questions vary in scope – some detailed, some general



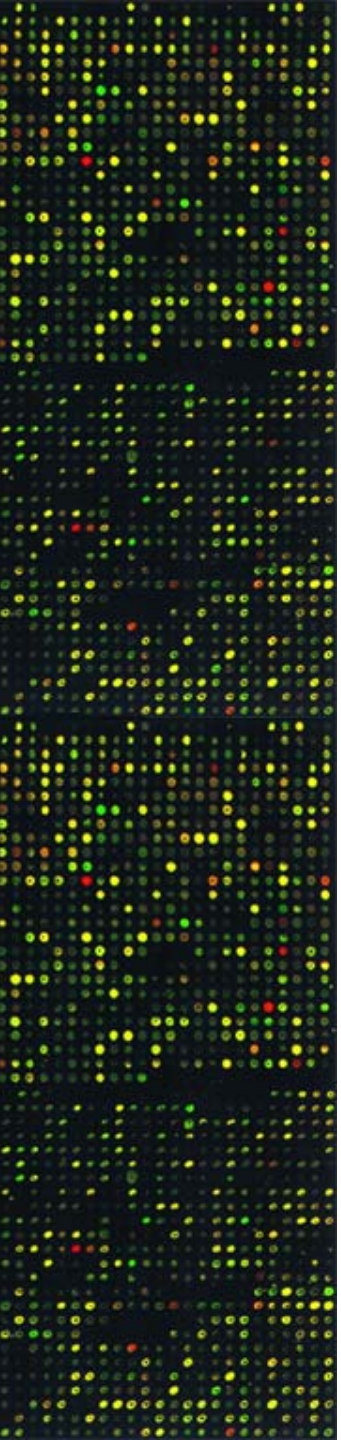
Methods

- Database of constructed of 159,345 array identifiers and corresponding GenBank accession numbers for:
  - GeneChip^{R} HuGeneFL, Cancer G100, U95a and Mu11K arrays (Affymetrix, Santa Clara, CA)
  - Human UniGEM^{TM} V 2.0 Clone Lists (Incyte Genomics, Palo Alto, CA)
  - Cluster identifiers from NCBI UniGene.



Methods, cont'd

- GenBank and other genomic database accession numbers identified in MEDLINE XML format citation tapes provided by NLM
- Citations processed to extract MeSH keywords, chemical terms, and Enzyme Commission Registry numbers

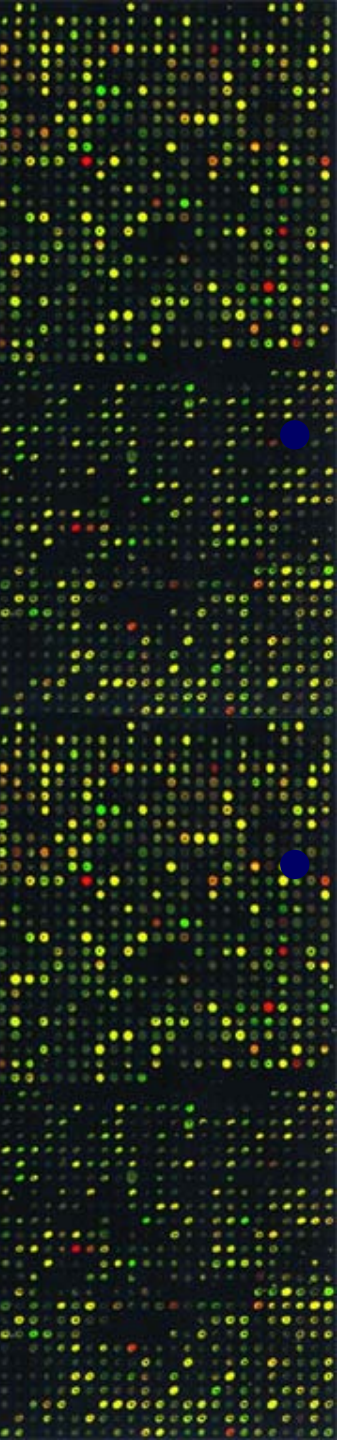


Literature Links Database as of January, 2003

- 159,345 array identifiers
- 79,855 unique Genbank Accession numbers
- 92,848 unique literature citations with one or more GenBank accession numbers
- 397,941 total links between a citation and a GenBank accession number
- 816,607 MeSH terms
- 348,455 Enzyme Registry terms

Sample match Results

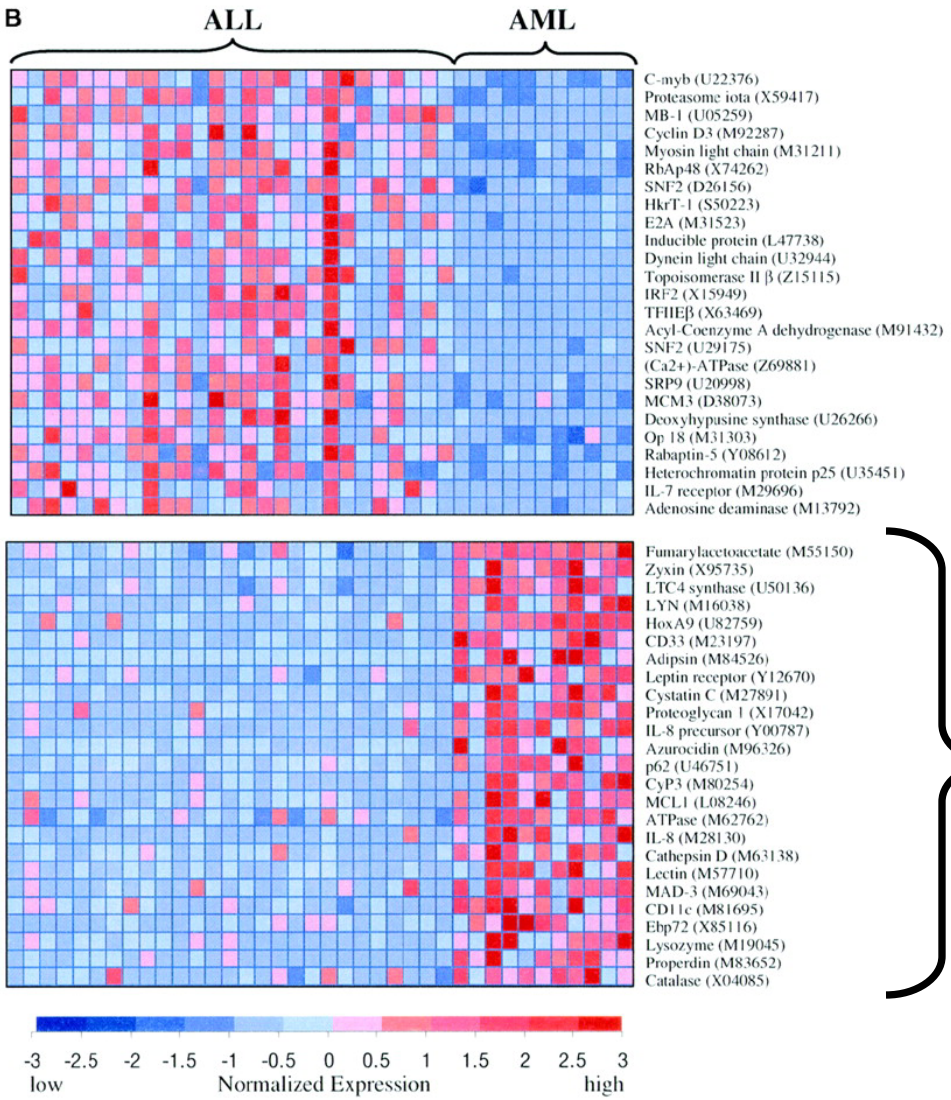
Array Name	Array IDs	GenBank Accessn Nrs	Cita-tions	Unique Cita-tions	Loci with no match	MeSH terms	Registry Number terms	EC Nrs	Total Index Terms	Fraction of array with 1 or more matching citations
Affy-HuFL	8693	6941	8771	6866	1551	54455	26498	5190	80953	77.6
Affy-U95a	8075	6547	8461	6679	1383	53038	25879	5097	78917	78.8
Affy-Cancer	2643	2223	3179	2553	452	20801	10197	2275	30998	79.6
Incyte Unigem v2	8820	8717	3586	2654	6357	23534	11676	2241	35210	27.0
Totals	**37051**	**14197**	**10378**	**8106**	7612	66054	32079	6174	**98133**	46.3



Methods, cont'd

- Web-accessible application built that accepts files containing groups of gene names and their associated expression values
- Creates keyword hierarchy summaries and detail pages with hyperlinks to GeneCard, Entrez, and PubMed citations

Hierarchical Keyword Analysis: An Example



AML-predictive genes

Golub TR, et al. (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science. 286(5439):531-7.

Golub - AML predictive genes

Results Available:

- List of genes that did and did not match
- Hierarchy of Keywords from literature associated with these genes
- Direct keyword matches in descending frequency

Hierarchy of Keywords from literature associated with these genes

Terms representing the largest 10th percentile of matches are shown in red. Numbers in {} brackets are P value estimates representing likelihood that this number of keyword matches would occur by chance. You may wish to bookmark this page or save it and any linked pages that you access on your own computer

Subject Keyword Areas	Term Matches
Enzyme Registry Numbers	29
Anatomy	43
Organisms	19
Diseases	17
Chemicals and Drugs	169
Analytical Techniques	15
Biological Sciences	72
Physical Sciences	6

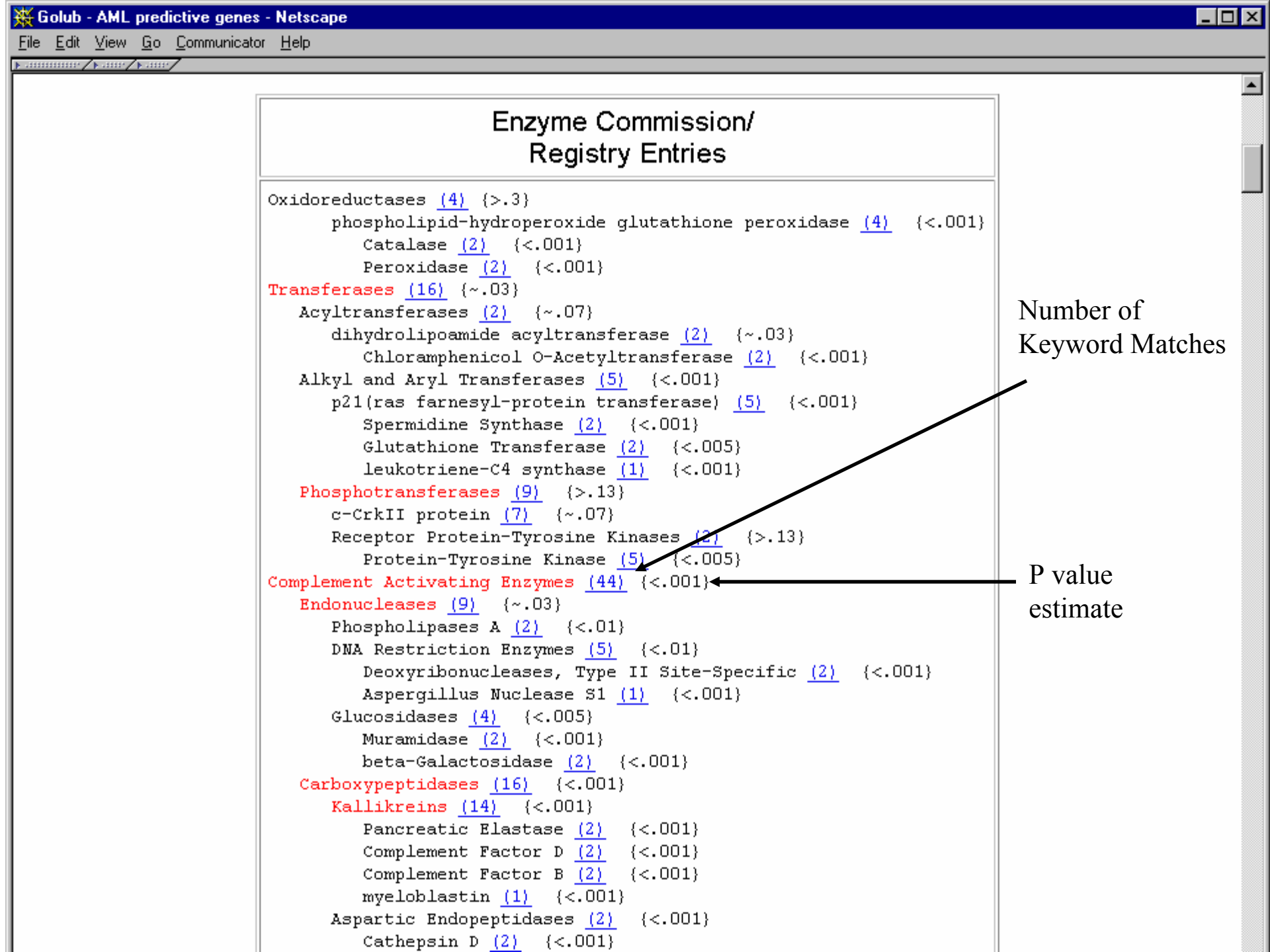
Golub - AML predictive genes

Genes that did and did not match one or more literature citations

Return to analysis results summary

Genes that matched

Accession	MEDLINE link	Citation
L08246	93234528	Buchan HL, Craig RW, Kozopas KM, Yang T, Zhou P: **MCL1, a gene expressed in programmed myeloid cell differentiation, has sequence similarity to BCL2.** Proc Natl Acad Sci U S A 1993 Apr 15;90(8):3516-20
	94193015	Irie S, Krajewski S, Reed JC, Sato T: **Cloning and sequencing of a cDNA encoding the rat Bcl-2 protein.** Gene 1994 Mar 25;140(2):291-2
	96256809	Afonso CL, Kutish GF, Neilan JG, Rock DL: **An African swine fever virus Bc1-2 homolog, 5-HL, suppresses apoptotic cell death.** J Virol 1996 Jul;70(7):4858-63
M16038	87172710	Fukushige S, Matsubara K, Miyajima N, Semba K, Sukegawa J, Toyoshima K, Yamamoto T, Yamanashi Y: **The yes-related cellular gene lyn encodes a possible tyrosine kinase similar to p56lck.** Mol Cell Biol 1987 Jan;7(1):237-43
M19045	88134189	Nakahama K, Toibana A, Yoshimura K: **Human lysozyme: sequencing of a cDNA, and expression and secretion by Saccharomyces cerevisiae.** Biochem Biophys Res Commun 1988 Jan 29;150(2):794-801
M23197	89009814	Seed B, Simmons D: **Isolation of a cDNA encoding CD33, a differentiation antigen of myeloid progenitor cells.** J Immunol 1988 Oct 15;141(8):2797-800


Golub - AML predictive genes - Netscape
File Edit View Go Communicator Help
Enzyme Commission/
Registry Entries
Oxidoreductases (4) {>.3}
phospholipid-hydroperoxide glutathione peroxidase (4) {<.001}
Catalase (2) {<.001}
Peroxidase (2) {<.001}
Transferases (16) {~.03}
Acyltransferases (2) {~.07}
dihydrolipoamide acyltransferase (2) {~.03}
Chloramphenicol O-Acetyltransferase (2) {<.001}
Alkyl and Aryl Transferases (5) {<.001}
p21(ras farnesyl-protein transferase) (5) {<.001}
Spermidine Synthase (2) {<.001}
Glutathione Transferase (2) {<.005}
leukotriene-C4 synthase (1) {<.001}
Phosphotransferases (9) {>.13}
c-CrkII protein (7) {~.07}
Receptor Protein-Tyrosine Kinases (2) {>.13}
Protein-Tyrosine Kinase (5) {<.005}
Complement Activating Enzymes (44) {<.001}
Endonucleases (9) {~.03}
Phospholipases A (2) {<.01}
DNA Restriction Enzymes (5) {<.01}
Deoxyribonucleases, Type II Site-Specific (2) {<.001}
Aspergillus Nuclease S1 (1) {<.001}
Glucosidases (4) {<.005}
Muramidase (2) {<.001}
beta-Galactosidase (2) {<.001}
Carboxypeptidases (16) {<.001}
Kallikreins (14) {<.001}
Pancreatic Elastase (2) {<.001}
Complement Factor D (2) {<.001}
Complement Factor B (2) {<.001}
myeloblastin (1) {<.001}
Aspartic Endopeptidases (2) {<.001}
Cathepsin D (2) {<.001}
Number of
Keyword Matches
P value
estimate

Diseases

Neoplasms (5) {>.13}
- Cysts (1) {<.001}
  - Kidney, Cystic (1) {<.001}
    - Kidney, Polycystic (1) {<.001}
- Neoplasms by Histologic Type (4) {>.6}
  - Leukemia (4) {<.001}
    - Leukemia, Hairy Cell (1) {<.001}
    - Leukemia, Myeloid (3) {<.001}
      - Leukemia, Myelomonocytic, Acute (1) {<.001}
      - Leukemia, Nonlymphocytic, Acute (1) {<.005}
        - Leukemia, Myelocytic, Acute (1) {<.001}

Urologic and Male Genital Diseases (2) {>.3}
- Urogenital Diseases (1) {>.13}
  - Urogenital Abnormalities (1) {<.005}
    - Kidney, Polycystic (1) {<.001}
- Urologic Diseases (1) {>.13}
  - Kidney Diseases (1) {~.07}
    - Kidney, Cystic (1) {<.001}
      - Kidney, Polycystic (1) {<.001}

Female Genital Diseases and Pregnancy Complications (1) {>.6}
- Genital Diseases, Female (1) {>.6}
  - Urogenital Diseases (1) {>.13}
    - Urogenital Abnormalities (1) {<.001}
      - Kidney, Polycystic (1) {<.001}

Hemic and Lymphatic Diseases (4) {>.3}
- Hematologic Diseases (1) {>.6}
  - Bone Marrow Diseases (1) {~.03}
    - Myelodysplastic Syndromes (1) {~.03}
      - Leukemia, Myeloid (1) {<.005}
- Lymphatic Diseases (3) {<.005}
  - Lymphoproliferative Disorders (3) {<.005}
    - Leukemia, Hairy Cell (1) {<.001}
    - Leukemia, Myeloid (2) {<.001}
      - Leukemia, Nonlymphocytic, Acute (1) {<.001}

Leukemia C4.557.337

GeneCards Link	Accession # (Entrez Link)	Citation (PubMed link)	Description
GeneCard	M31303	92011487	M31303 Human oncoprotein 18 (Op18) gene, complete cds
GeneCard	M31523	90150282	M31523 Human transcription factor (E2A) mRNA, complete cds
GeneCard	S50223	93043304	S50223 HKR-T1=Kruppel-like zinc finger protein [human, MOLT 4 T-cells, mRNA, 798 nt]
GeneCard	L08246	93234528	L08246 Human myeloid cell differentiation protein (MCL1) mRNA
GeneCard	X17042	90016819	X17042 Hematopoeitic Proteoglycan core protein

Return to Diseases index

Leukemia, Lymphocytic C4.557.337.428

GeneCards Link	Accession # (Entrez Link)	Citation (PubMed link)	Description
GeneCard	M31523	90150282	M31523 Human transcription factor (E2A) mRNA, complete cds
GeneCard	S50223	93043304	S50223 HKR-T1=Kruppel-like zinc finger protein [human, MOLT 4 T-cells, mRNA, 798 nt]

Return to Diseases index

Leukemia, B-Cell C4.557.337.428.500

GeneCards Link	Accession # (Entrez Link)	Citation (PubMed link)	Description
GeneCard	M31523	90150282	M31523 Human transcription factor (E2A) mRNA, complete cds

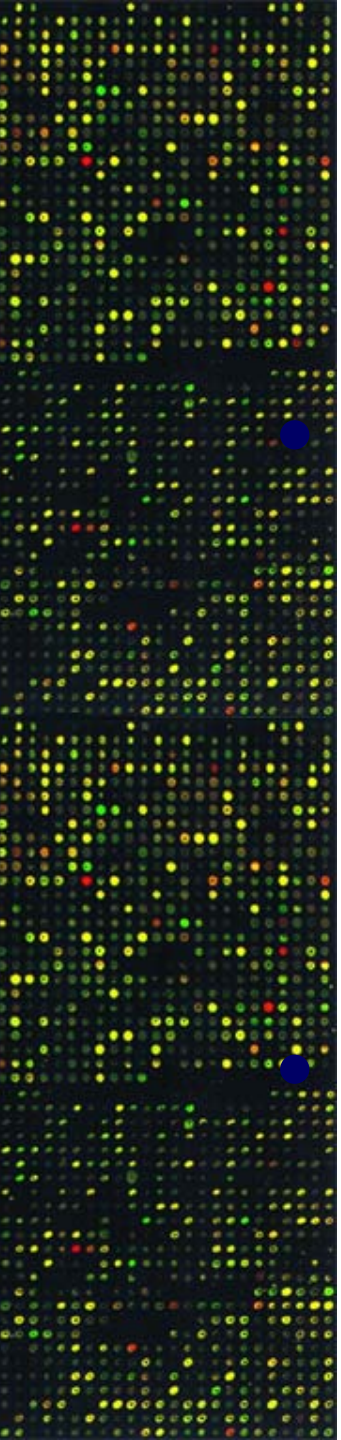
Return to Diseases index

Leukemia, B-Cell, Acute C4.557.337.428.500.100

Direct keyword matches in descending frequency

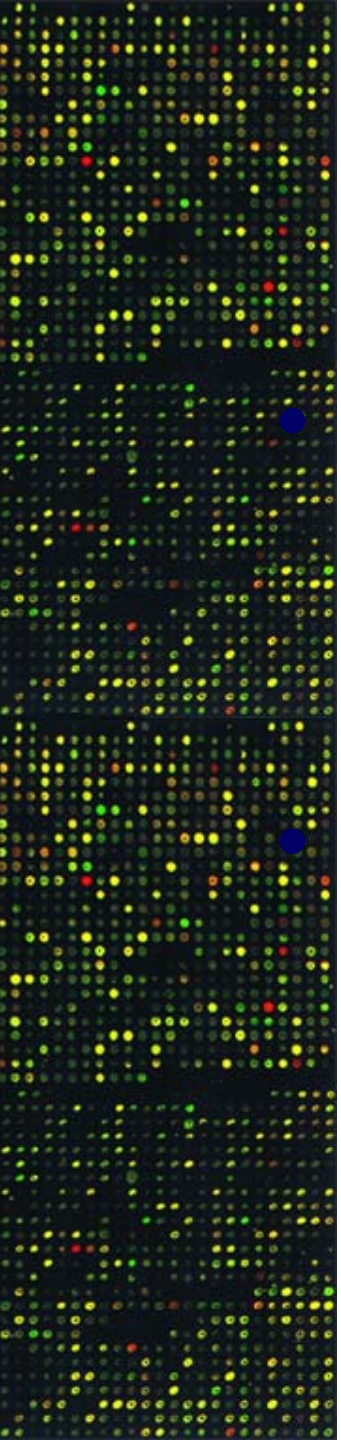
Return to analysis results summary

Matches	MeSH Term
25	Carrier Proteins
25	DNA-Binding Proteins
21	Sequence Homology, Amino Acid
16	Recombinant Proteins
16	Transcription Factors
14	Saccharomyces cerevisiae
13	Sequence Homology, Nucleic Acid
12	Promoter Regions (Genetics)
11	Fungal Proteins
10	Gene Expression Regulation
10	Genes, Structural
10	Genetic Markers
9	Adenosine Deaminase
9	Membrane Glycoproteins
9	Nuclear Proteins
9	Proto-Oncogene Proteins
9	Transcription, Genetic
9	Transfection
9	Tumor Cells, Cultured
8	DNA Probes
8	Hela Cells
8	Membrane Proteins
7	DNA, Neoplasm



HAPI Keyword Analysis of Golub, et. al. data shows:

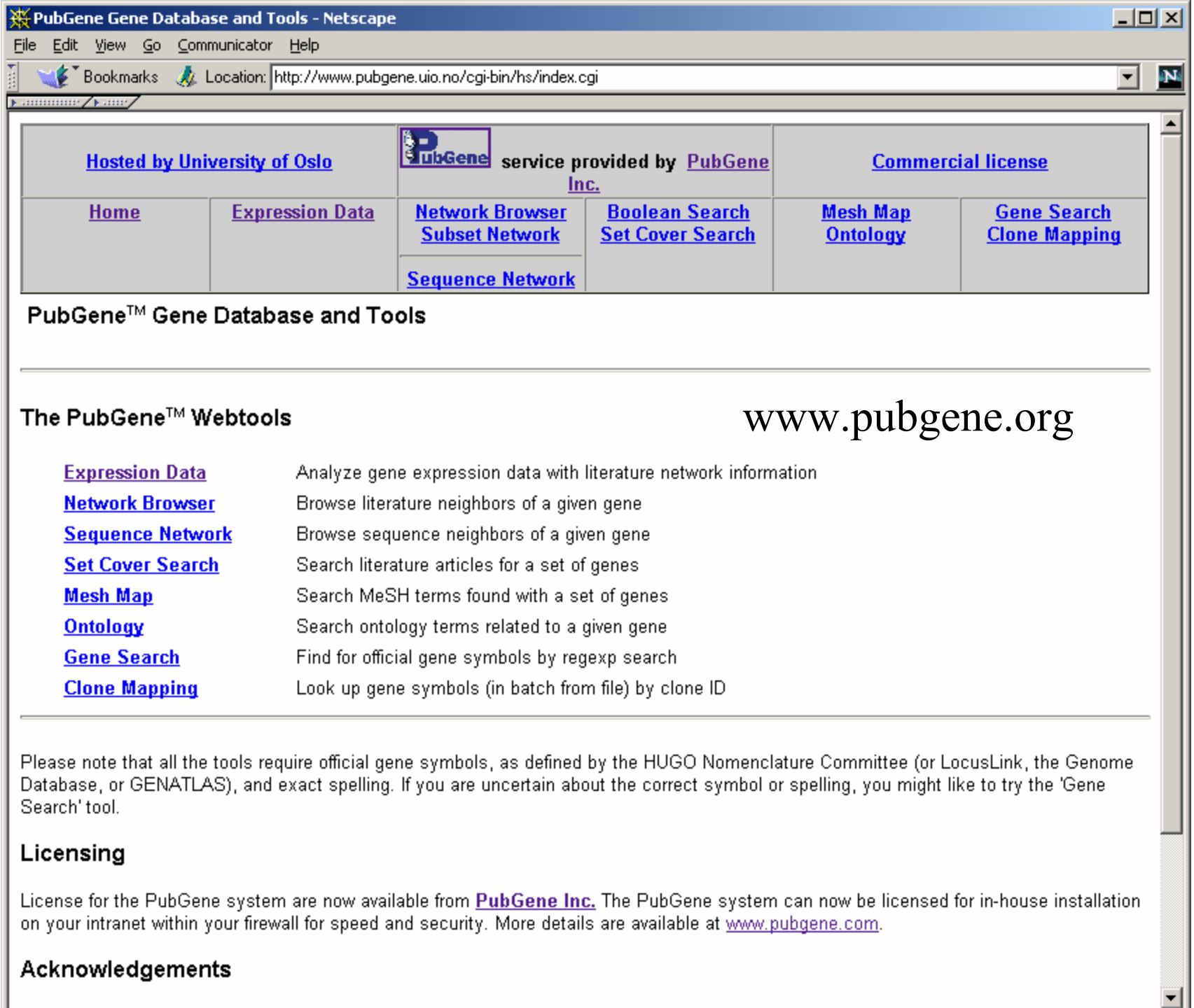
- In AML 'plasminogen activators' occur as a high frequency keyword, (potentially correlates with defibrination syndromes and other hemostatic abnormalities that are associated with AML but not with ALL)
- ALL-predictive genes also associated with inherited combined immunodeficiency



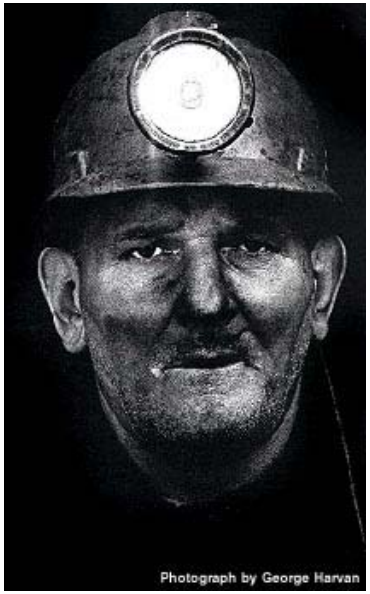
Data Mining of Literature-associated keywords

- Strengths
  - Shows potential similarities in multiple contexts
  - May yield unexpected biological insights
  - Results improve over time as new literature published
- Limitations/Weaknesses
  - Genes & ESTs with no linked literature do not participate in the keyword analysis
  - Older, well-characterized genes over-represented vs. new genes
  - Best used as adjunct to other clustering methods; mapping keywords of all genes looks like "all of known biology"

Current prototype available at http://array.ucsd.edu


PubGene Gene Database and Tools - Netscape
File Edit View Go Communicator Help
Bookmarks Location: http://www.pubgene.uio.no/cgi-bin/hs/index.cgi
Hosted by University of Oslo
PubGene service provided by PubGene Inc.
Commercial license
Home
Expression Data
Network Browser
Subset Network
Sequence Network
Boolean Search
Set Cover Search
Mesh Map
Ontology
Gene Search
Clone Mapping
PubGene™ Gene Database and Tools
The PubGene™ Webtools
www.pubgene.org
Expression Data Analyze gene expression data with literature network information
Network Browser Browse literature neighbors of a given gene
Sequence Network Browse sequence neighbors of a given gene
Set Cover Search Search literature articles for a set of genes
Mesh Map Search MeSH terms found with a set of genes
Ontology Search ontology terms related to a given gene
Gene Search Find for official gene symbols by regexp search
Clone Mapping Look up gene symbols (in batch from file) by clone ID
Please note that all the tools require official gene symbols, as defined by the HUGO Nomenclature Committee (or LocusLink, the Genome Database, or GENATLAS), and exact spelling. If you are uncertain about the correct symbol or spelling, you might like to try the 'Gene Search' tool.
Licensing
License for the PubGene system are now available from PubGene Inc. The PubGene system can now be licensed for in-house installation on your intranet within your firewall for speed and security. More details are available at www.pubgene.com.
Acknowledgements

Data Mining



- A miner leads a tough life, but once in a while you strike it rich
- The meek shall inherit the Earth, but not its mineral rights

- J. Paul Getty

Acknowledgements

HAPI
High-density
Array
Pattern
Interpreter

Jacques Corbeil, Ph.D.
UCSD Cancer Center

Igor Klacansky, Ph.D.
UCSD Cancer Center

Michael Gribskov, Ph.D.
Computational Biology Unit
San Diego Supercomputer Center

J. Lynn Fink
San Diego Supercomputer Center

John B. Welsh, M.D., Ph.D.
Novartis Research Foundation

Supported by:
NCI "Molecular Characterization of Prostate Cancer" grant
5 U01 CA84998-02