

Web based tools for transcriptome research

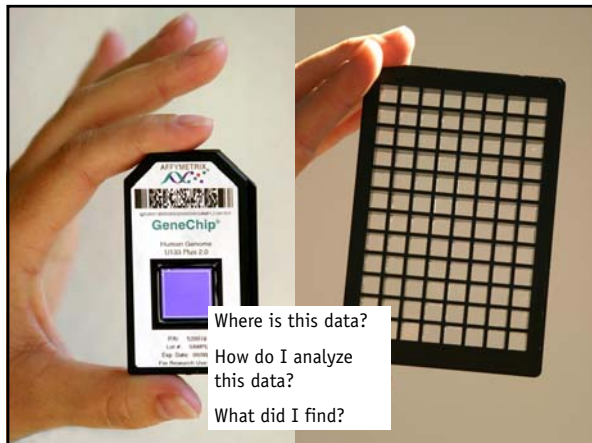
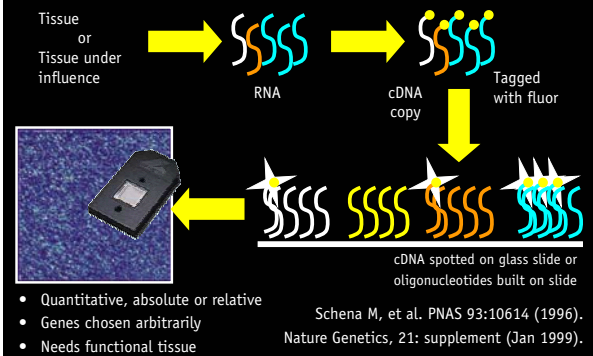
Atul Butte, MD, MS
atul_butte@harvard.edu

Children's Hospital Informatics Program
www.chip.org

Children's Hospital • Boston
Harvard Medical School
Massachusetts Institute of Technology



RNA expression detection chips



Agenda

- Finding web-accessible data
- Web-downloadable analysis tools
- Rediscovering what you've found
- Discovery portals: binding data and the tools that operate on them

The following lists and URLs are available in
Butte AJ. The use and analysis of microarray data.
Nature Reviews Drug Discovery, 1:951, December 2002.
so don't struggle to write them down...

Web-accessible data

- Great amounts of publicly available microarray data are available on the Internet
- *National Center for Biotechnology Information Gene Expression Omnibus*
 - <http://www.ncbi.nlm.nih.gov/geo/>
 - Over 16,000 microarrays from over 700 types of microarrays.
- *Whitehead Institute Center for Genome Research*
 - <http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi>
 - Microarrays from 12 publications involving many types of cancer, including some clinical measurements associated with each sample.

Web-accessible data

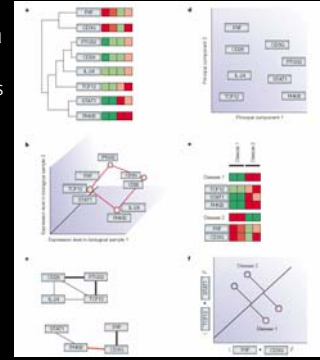
- *Children's National Medical Center (HopGenes Program in Genomic Applications)*
 - <http://microarray.cnmcresearch.org/pgadatatable.asp>
 - Over 1400 microarrays from 45 experiments, including many human diseases, muscular dystrophy, dermatomyositis, and heart failure, as well as mouse, rat and dog models of spinal cord injury, pulmonary disease, and heart failure.
- *Human Gene Expression Index*
 - <http://www.hugeindex.org/databases/index.html>
 - 121 microarrays from 19 normal human tissues.

Web-accessible data

- **Stanford Microarray Database**
 - <http://genome-www5.stanford.edu/MicroArray/SMD/>
 - Over 4000 microarrays measured across 15 species, from 173 publications.
- **CardioGenomics Program in Genomic Applications**
 - <http://cardiogenomics.med.harvard.edu/public-data.html>
 - 142 microarrays involving mouse models of cardiac development and signal transduction, including measurements made in time-series.
- **TREX Program in Genomic Applications**
 - <http://pga.tigr.org/data.shtml>
 - 565 microarrays from mouse and rat models of sleep, infection, hypertension, and pulmonary disease.

Existing clustering techniques

- Several algorithms have already been developed for knowledge discovery and data-mining of RNA expression data sets
- **Hierarchical clustering:** Eisen MB, PNAS 95:14863. Look at the neighbors of unknown genes/samples.
- **Self-organizing maps:** Tamayo P, PNAS 96:2907. Similar items are near each other.
- **Relevance networks:** Butte AJ, PNAS 97:12182. Networks of positive and negative association.
- **Principal components:** Raychaudhuri S, PSB 2000. How can I best spread my samples?
- **Nearest neighbors:** Golub TR, Science 286:531. What genes best predict my outcome individually?
- **Support vector machines:** Brown MP, PNAS 97:262. What combination of genes best splits my outcome?

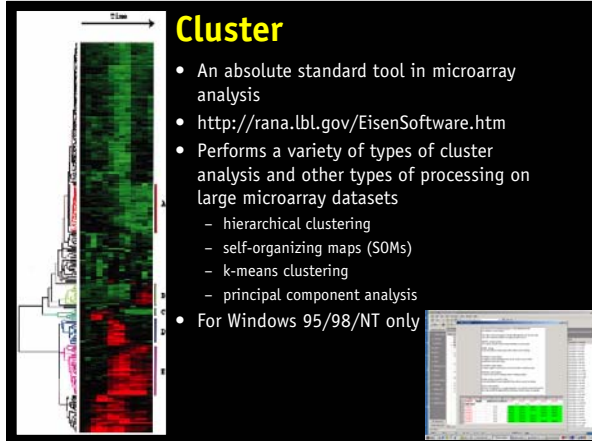


Butte AJ. Nature Reviews Drug Discovery, 1:951.

Tell Me More

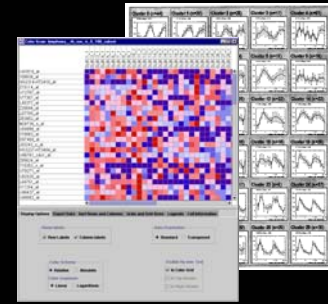
Cluster

- An absolute standard tool in microarray analysis
- <http://rana.lbl.gov/EisenSoftware.htm>
- Performs a variety of types of cluster analysis and other types of processing on large microarray datasets
 - hierarchical clustering
 - self-organizing maps (SOMs)
 - k-means clustering
 - principal component analysis
- For Windows 95/98/NT only



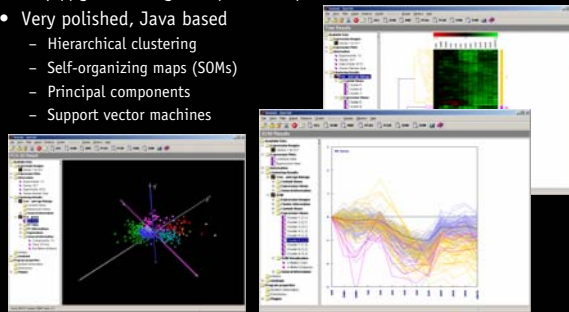
GeneCluster 2.0

- Whitehead Institute
- <http://www-genome.wi.mit.edu/cancer/software/genecluster2/gc2.html>
- Java based
 - Nearest neighbor
 - Self-organizing maps (SOMs)
 - Marker analysis



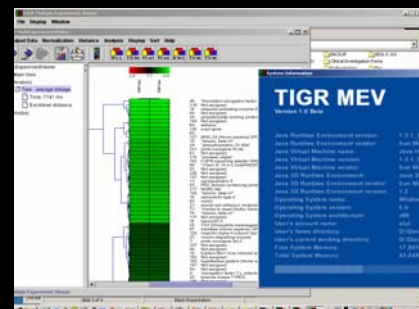
Genesis

- Free for the moment
- <http://genome.tugraz.at/Software/GenesisCenter.html>
- Very polished, Java based
 - Hierarchical clustering
 - Self-organizing maps (SOMs)
 - Principal components
 - Support vector machines



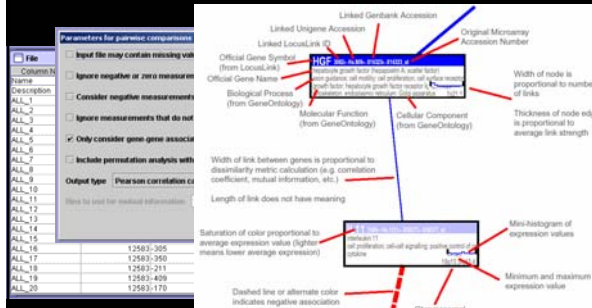
TIGR MEV

- Multi-experiment viewer
- <http://www.tigr.org/softlab>
- Extensive documentation
- Java-based



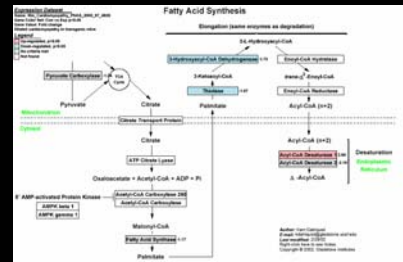
Relnet

- Java-based software to make relevance networks
- <http://www.chip.org/relnet>

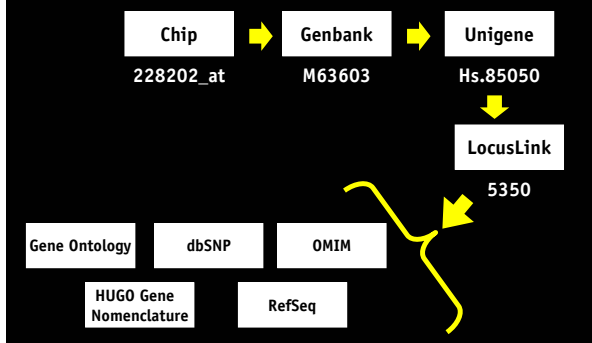


GenMAPP

- Gene Microarray Pathway Profiler
- Paints pathway pictures in the context of microarray measurements
- <http://www.genmapp.org>
- Can help translate list of genes into implicated pathways



Accession Numbers

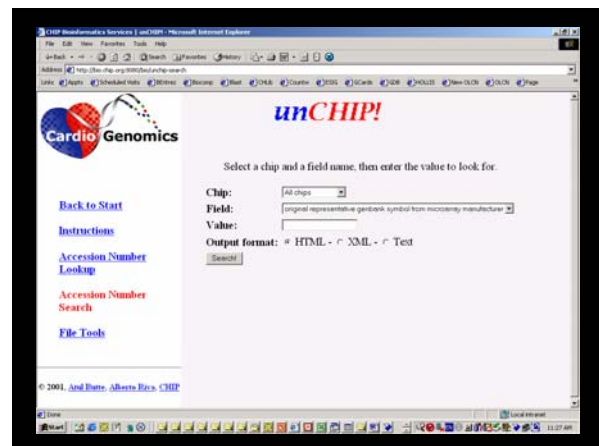


Looking up results

- Database Referencing of Array Genes Online
 - <http://www.kennedykrieger.org/pevsnerlab/dragon.htm>
- Unchip
 - <http://www.unchip.org>
- Resourcerer
 - http://pga.tigr.org/tigr-scripts/nhgi_scripts/resourcerer.pl
- Netaffx
 - <http://www.netaffx.com>

www.unchip.org

- “Before and after” example
 - 33640_at: Cluster incl. Y14768: Homo sapiens DNA, cosmid clone
 - Unchip: Allograft inflammatory factor 1
- Search for “95654_at”
 - valyl-tRNA synthetase 2, heat shock protein cognate 70, etc.
- Search for “M35416_at”
 - Affymetrix: RALB V-ral simian leukemia viral oncogene homolog B
 - Edition 2: v-ral simian leukemia viral oncogene homolog B (ras related; GTP binding protein)
 - Edition 3: stathmin 1/oncoprotein 18
- Can find official names, symbols, and synonyms for accessions
- Can search for expression by functional domain or meaning
- Available as database and web-site, at www.unchip.org
- Updated periodically

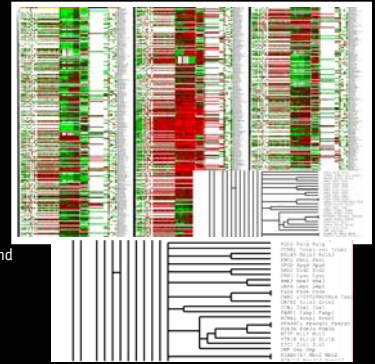


Discovery portals

- Binding data and the tools that operate on them
- Physgen
 - <http://pga.mcw.edu>
- PGAGene
 - <http://pgagene.chip.org>
- NCBI Geo
 - <http://www.ncbi.nlm.nih.gov/geo/>

Integrating data across species

- NHLBI Programs in Genomics Applications (PGA)
- 11 funded nationwide
- PGAGene: Integrated gene-centric view
- Available today at pgagene.chip.org
- 5.9 million pieces of data from over 1,200 publicly available microarrays
- Combining 893 genes measured in rat, mouse and human, across 997 arrays
- Programming by Kyungjoon Lee



HomoloExpress

- **HomoloGene**: orthologs and homologs across 19 organisms
- **GEO**: public gene expression data repository
- GEO does not provide any capability to allow searching across species using gene homology
- **HomoloExpress** allows a cross-species, gene-specific search of data within GEO
- homoloxp.chip.org

HomoloExpress

- HomoloExpress has over 4,000 samples
- 22.7 million individual expression measurements
- Mapped to over 60,000 ortholog relations
- Over 70 microarray platforms for 5 species
- Mouse beta adult hemoglobin major chain (*Hbb-b1*) has most measurements (18,950 array measurements)
- Gastric lipase (*Lipf*) ortholog family has the most measurements across species (19,444 array measurements)

PhysGen - Microsoft Internet Explorer

http://qiga.msu.edu/qiga/strain_desc.asp

PhysGen Home > Feedback > News > Overview > Components > Data > Protocols > Collaborations > PSA Query Links > Search all PSA > PSA Inventory > Other Links

Strain & Genotype Data Description

Scroll down a little...

The SS and BN strains were chosen because of their relevance to hypertension, cardiac and vascular disorders, hyperlipidemia, renal disease and insulin resistance. The SS and BN strains are the parents for developing the SS BN consomics. The SS BN consomics provide a means to validate the mapping studies, generate new genetic mapping data and most importantly will provide the community with renewable animal resources that allow rapid narrowing of genomic regions of interest and ultimate testing of genes to functional pathways of importance to the IHLB. The FHH strain both complements and contrasts with the SS with respect to heart, lung and blood function. The FHH BN consomics will be developed later in the program from the FHH and BN parental strains. The SFRD is an outbred rat commonly used in the scientific community. The GH Genetically Hypertensive Rat is an inbred rat that develops hypertension, cardiac hypertrophy and vascular disease. In a complete panel of consomic rats, each chromosome is replaced one at a time (strain A to strain B), so that the contribution of genes on each chromosome can be assessed by phenotyping the consomic strain for the traits of interest on an inbred background. Consomic strains overcome the confounding effects of heterogeneous genomic backgrounds in which the phenotypic noise may make the detection of weak QTLs difficult. Consomics also provide a renewable source of animals and a platform upon which subsequent genetic studies and/or physiological studies can be used to study a QTL on the chromosome.

Strain Name Strain Description Strain Type References & Links

PhysGen - Microsoft Internet Explorer

http://qiga.msu.edu/qiga/strain_desc.asp

physiological studies can be used to study a QTL on the chromosome.

Strain Name	Strain Description	Strain Type	References & Links
BN	Brown Norway, BN-SaNHsd Mcwi	PARENTAL	BN genotype test result, RGD BN strain report
CDF	Fischer CDF(F-344)/CjBR Charles River Laboratories	INBRED	CDF A-teloid genotype of the rat
CDRIGS	CD# IGS (Cdc)		CDRIGS genotype test result
FHH	Fawn-Hooded H		FHH genotype test result, RGD FHH strain report
FHH-1BNMcwi	A FHH genomic background with a BN chromosome 1 introgressed	CONSOOMIC	genotype test result
GH	GH Genetically Hypertensive Rat. The MCW colony was derived from the University of Otago colony, registered pure line GH/Otag (Festing MFW, 1979, Inbred Strains in Biomedical Research, Macmillan London) as described originally by Smak [Smak, FH and Had WH (1958) Inherited hypertension in rats. Nature (London): 192:747-8]	INBRED	GH genotype test result, RGD GH strain report
LEW	Lewis (LEW CjBR) Charles River Laboratories	INBRED	LEW detailed genotype of the region, RGD LEW strain report

I remember hearing that the Brown Norway has a problem with diabetes...

Strain Report - Microsoft Internet Explorer

RGD RAT GENE DATABASE

HOME MAPS GENES ESTS QTLs SSLPS SEQUENCES STRAINS REFERENCES HELP

Strain Report

Symbol BN
Strain BN
Type inbred

Scroll down a little...

See Also: BN/Cj, BN/Cj-Ix, BN/Ka, BN/NHsdMcwi, BN/Sa, BN/SaNHsd

Basic details

Genetic markers BN a,b,h(i)
Coat Color BN Brown, BN/Ka Brown, BN/NHsdMcwi Brown
Inbred BN F71(Pt)
Generations BN/Cj-Ix F>20

Strain Report - Microsoft Internet Explorer

Strain QTL data:

Chr	Symbol	Name	Trait	Subtrait
BN				
X	Niddm16	Non-insulin-dependent diabetes mellitus QTL 16	Diabetes Mellitus	
X	Cia18	CIA Severity QTL 18	Arthritis Severity	Maximum Score
X	Cia19	CIA Severity QTL 19	Arthritis Severity	Maximum Score
1	Niddm13	Non-insulin-dependent diabetes mellitus QTL 13	Diabetes Mellitus	body weight
1	Dme1	Diabetes Mellitus OLETF QTL 1	Body Weight	
1	Xhs2	X-ray Hypersensitivity QTL 2	Hypersensitivity	X-ray hypersensitivity
1	BpQTLcluster1	Blood Pressure QTL cluster 1	Blood Pressure	
1	Niddgk1	Non-insulin-dependent diabetes mellitus GK QTL 1	Glucose	Tolerance
2	Bp19	Blood Pressure QTL 19	Blood Pressure	
2	Cia7	CIA Severity QTL 7	Arthritis Severity	Maximum score
2	Niddgk2	Non-insulin-dependent diabetes mellitus GK QTL 2	Insulin level	Fasting
2	Bp17	Blood Pressure QTL 17	Blood Pressure	
2	Bp18	Blood Pressure QTL 18	Blood Pressure	Salt-loaded Systolic
3	Bp20	Blood Pressure QTL 20	Blood Pressure	
3	BpQTLcluster4	Blood Pressure QTL cluster 4	Blood Pressure	Systolic
4	Niddgk3	Non-insulin-dependent diabetes mellitus GK QTL 3	Insulin	Tolerance

Hmm, I was right. Here's a QTL for diabetes...

RGD QTL Report: Niddgk2 - Microsoft Internet Explorer

RGD RAT GENE DATABASE

HOME MAPS GENES ESTS QTLs SSLPS SEQUENCES STRAINS REFERENCES HELP

QTL Report: Niddgk2

Symbol Niddgk2
Full Name Non-insulin-dependent diabetes mellitus GK QTL 2
Chromosome 2
LOD 4.1

Mapping Crosses GK x BN
Strains used BN GK

Trait Insulin level
Subtrait Fasting

Here's the marker for the peak for the QTL...

Mapping Data:

Flank & Peak Marker Locations:	Marker	Symbol	Map Name	Chr.	Position
Peak	D2Wox23	-	-	2	-

RGD SSLP Report: D2Wox23 - Microsoft Internet Explorer

RGD RAT GENE DATABASE

HOME MAPS GENES ESTS QTLs SSLPS SEQUENCES STRAINS REFERENCES HELP

SSLP Report: D2Wox23

Symbol D2Wox23
Expected Size bp
Previous symbol(s) RS RATP9KA
Associated Genes S100a4
Associated Sequences Primer pair
Oligo 1 (5'-3') GATGAGAGATTCTGTGACGAG
Oligo 2 (5'-3') TTTTCAGTTTATCTCTGTGTC

Mapping Data (?)

Map Locations for:

Marker	Symbol	Map Name	Chr.	Position
SSLP	D2Wox23	-	2	-

Strain Variations (?)

And here's how to test for that marker. Interesting... there's a gene here.

Let's recap: in rats, using these PCR primers, we can extract a stretch of DNA. The length of this DNA differs in different strains of rats. And differences in the length of this DNA statistically correspond with differences in fasting insulin level. There happens to be a gene in this stretch.

Let's learn more about that QTL region...
NCBI: www.ncbi.nlm.nih.gov

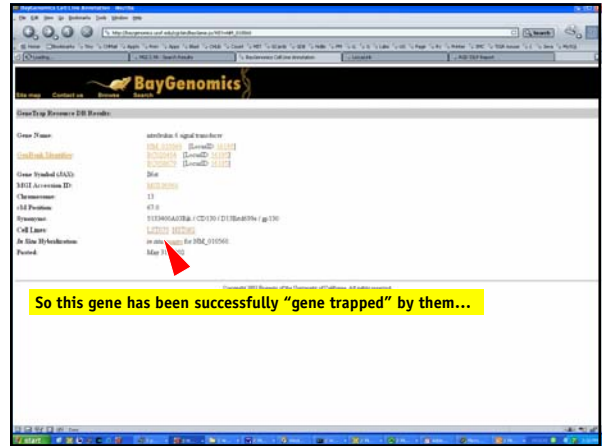
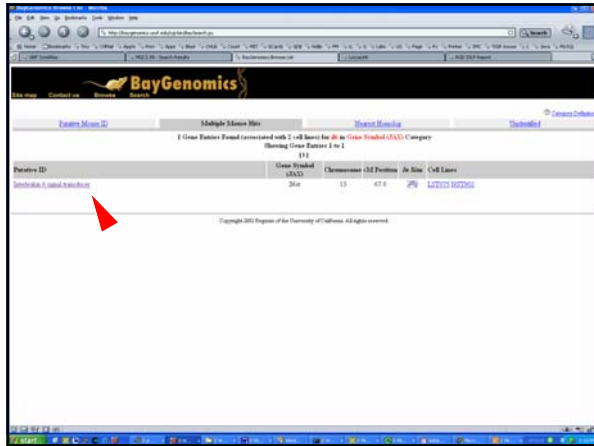
Let's look at Genomes, and the rat specifically...

Where is that gene S100A4 in the rat genome?

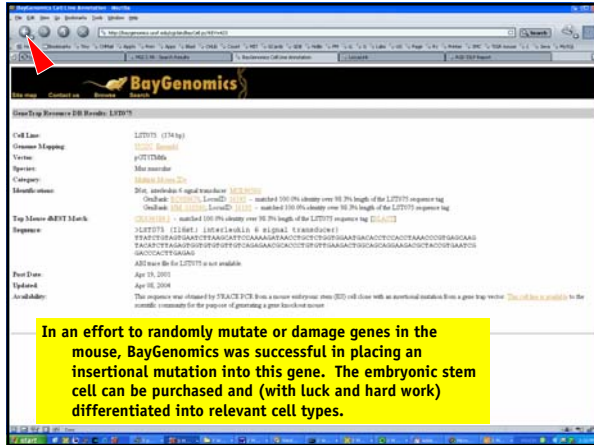
Here it is on chromosome 2...

Zoom out a little...

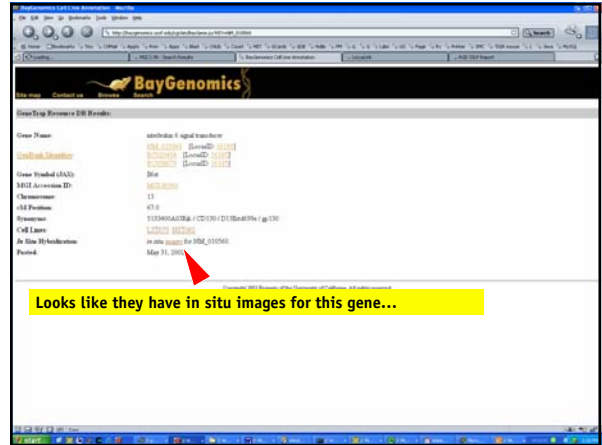
That gene is near several other interesting genes...



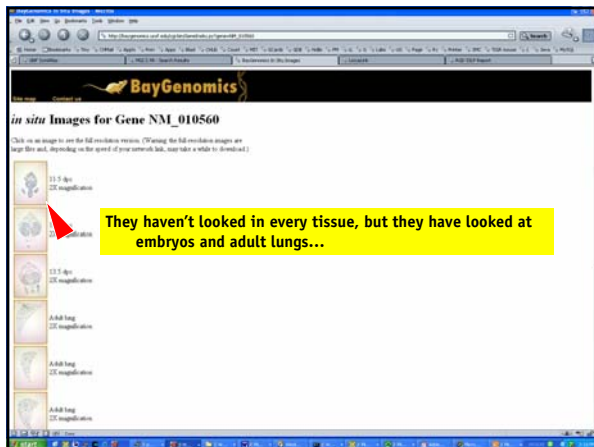
So this gene has been successfully "gene trapped" by them...



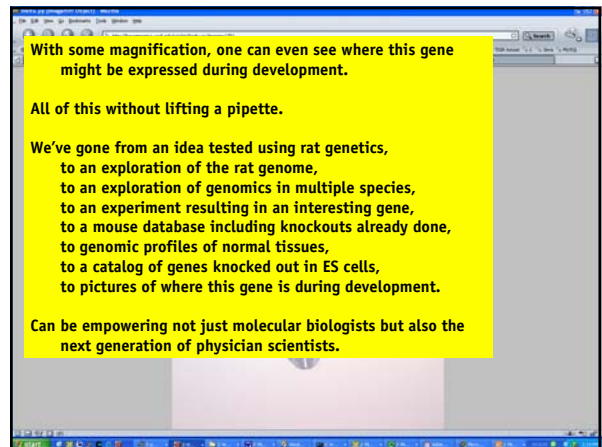
In an effort to randomly mutate or damage genes in the mouse, BayGenomics was successful in placing an insertional mutation into this gene. The embryonic stem cell can be purchased and (with luck and hard work) differentiated into relevant cell types.



Looks like they have in situ images for this gene...



They haven't looked in every tissue, but they have looked at embryos and adult lungs...



With some magnification, one can even see where this gene might be expressed during development.

All of this without lifting a pipette.

We've gone from an idea tested using rat genetics, to an exploration of the rat genome, to an exploration of genomics in multiple species, to an experiment resulting in an interesting gene, to a mouse database including knockouts already done, to genomic profiles of normal tissues, to a catalog of genes knocked out in ES cells, to pictures of where this gene is during development.

Can be empowering not just molecular biologists but also the next generation of physician scientists.

Using Genomics to Diagnose

- Difficulty distinguishing between leukemias
- Microarrays can find genes that help make the diagnosis easier

Golub TR. Science 286:531, 1999.

Using Genomics to Predict

- Patients with seemingly the same B-cell lymphoma
- Looking at pattern of activated genes helped discover two subsets of lymphoma
- Big differences in survival

Alizadeh AA. Nature 403:503, 2000. Nature Medicine 9:9.

Using Genomics to Treat

Sesti F. PNAS 97:10613, 2000

A common polymorphism associated with antibiotic-induced cardiac arrhythmia

Federico Sesti*, Geoffrey W. Abbott*, Jian Wei*, Katherine T. Murray*, Sanjeev Sakseena*, Peter J. Schwartz*, Silvia G. Priori*, Dan M. Roden†, Alfred L. George, Jr.†, and Steve A. H. Goldstein*†

*Departments of Medicine and Cellular and Molecular Physiology, Brown Center for Molecular Medicine, Yale University School of Medicine, New Haven, CT 06510; †Departments of Medicine and Pharmacology, Vanderbilt University, Nashville, TN 37232; †Robert Wood Johnson Medical School, Piscataway, NJ 07053; and †Department of Cardiology, University of Texas and Prentiss Lee Moore MDC, Pecos, New Mexico 87854

Edited by Vincent T. Marchesi, Yale University School of Medicine, New Haven, CT, and approved July 6, 2000 (received for review May 16, 2000)

- Genes will help us determine which drugs to use in particular disease subtypes
- Genes will help us predict those who get side-effects

EFFECT OF A SINGLE AMINO ACID CHANGE IN MHC CLASS I MOLECULES ON THE RATE OF PROGRESSION TO AIDS

XIAOJIANG GAO, PH.D., GEORGE W. NELSON, PH.D., PETER KARACIO, B.A., MAUREEN P. MARTIN, M.D., JOHN PHAIR, M.D., RICHARD KASLOW, M.D., JAMES J. GOEDERT, M.D., SUSAN BUCHBINDER, M.D., KEITH HOOTS, M.D., DAVID VLAHOV, PH.D., STEPHEN J. O'BRIEN, PH.D., AND MARY CARRINGTON, PH.D.

1668 • N Engl J Med, Vol. 344, No. 22 • May 31, 2001 • www.nejm.org

After microarrays comes wafers...

- Chromosome 21 has 21 million base-pairs
- Each 5 inch square wafers (Perlegen) hold 60 million probes
- Can sequence an entire chromosome in one experiment
- Can sequence all SNPs within a human in 10 days
- Each scan takes up around 10 terabytes

Patil N. Science 2001, 294:1719.

Many physicians do not know how to use the genome

IN PRACTICE: Genetics: Blind Spot in Medical Training

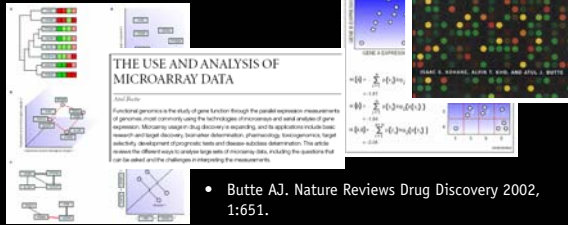
The explosion of knowledge about what role genetics plays in disease has altered the way medical care is being provided at the most basic levels, and many health professionals may not be up to the task.

Take Home Points

- Thousands of microarray results and tools are available on the Internet; annotations are harder to find.
- Due to rapidly changing information, one is never truly finished analyzing a microarray data set.
- Web-based discovery portals are usable and empower investigators to integrate across genomic data.

Microarrays for an Integrative Genomics

- One of the first books on microarray analysis and experimental design
- Barnes and Noble, Borders, Amazon: US\$32-40



Use and Analysis of Microarray Data

Diabetes Genome Anatomy Project

PEOPLE PROJECTS CON ANNOTATION STATEMENT RESOURCES

Goals of the DGAP Project

Normal Anatomy

- Identify the normal gene/protein expression program induced to insulin in muscle, fat and liver of rodents and humans and its differentiating cultured adipocytes?
- What elements are responsible of all target tissues and which are specific to each tissue?
- Can functional sub-programs of insulin regulated gene/protein expression be identified related to distinct branches of insulin action or specific compartments of the cell?
- Are expression profiles reproducible and does array analysis detect the majority of insulin-regulated genes or is subtraction cloning more sensitive?

Morbid Anatomy

- How do the gene/protein expression programs affected in diabetes and insulin resistant states?
- Which changes are related to the insulin resistance of type 2 diabetes or obesity, which to the insulin deficiency of type 1 diabetes, and which to the metabolic abnormalities of both forms of diabetes?
- Can expression profiles be defined that allow metabolic staging of the disease?
- Are array analysis detect the majority of diabetes-regulated genes or are other approaches more sensitive?

Genetic Variation

- How variable are the sequences of insulin and diabetes responsive genes?
- Do these sequence variations contribute to altered expression or function?

www.diabetesgenome.org

Collaborators and Support

- C. Ronald Kahn and Yu-Hua Tseng
- Mary Elizabeth Patti
- Dietrich Stephan / TGen
- Lois Smith / Children's Hospital
- Louis Kunkel / Children's Hospital
- Seigo Izumo / Beth Israel Deaconess NHLBI Program of Genomic Applications Framingham Heart Study
- Scott Weiss / BWH Channing NHLBI Program of Genomics Applications
- David Rowitch / Dana Farber
- Towia Libermann / Beth Israel
- Terry Strom / Beth Israel Deaconess
- NIH: NIDDK, NLM, NINDS, NHLBI, NIDDK, NIAID, NHGRI, NCI, NIGMS
- Lawson Wilkins NovoNordisk Award
- Merck / MIT Fellowship
- Genentech Foundation Fellowship
- Endocrine Fellows Foundation



Bioinformatics at the Children's Hospital Informatics Program www.chip.org

Fellows

- Dominic Alloco

Post-doctoral fellows

- Sangeeta English

Students

- Kyungjoon Lee

Staff

- Maung Min

Previous Students

- Maneesh Yadav
- Ling Bao
- Jinyun Chen
- Aaron Homer

- Isaac Kohane, Director
- Alvin Kho
- Peter Park
- Marco Ramoni
- Alberto Riva
- Asher Schachter

Atul Butte, MD
atul_butte@harvard.edu