

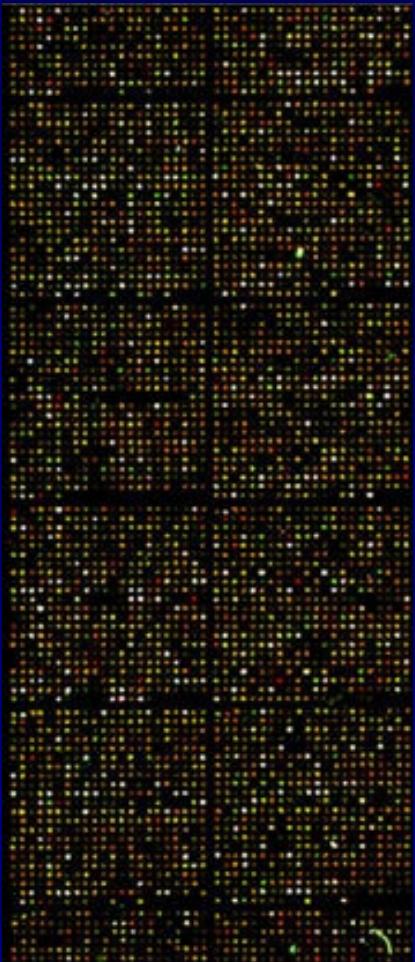
Microarray Data processing

Mikhail Dozmorov, Ph.D.,
Arthritis & Clinical Immunology Department
Oklahoma Medical Research Foundation
Mikhail-Dozmorov@omrf.org
January 17, 2013

Topics

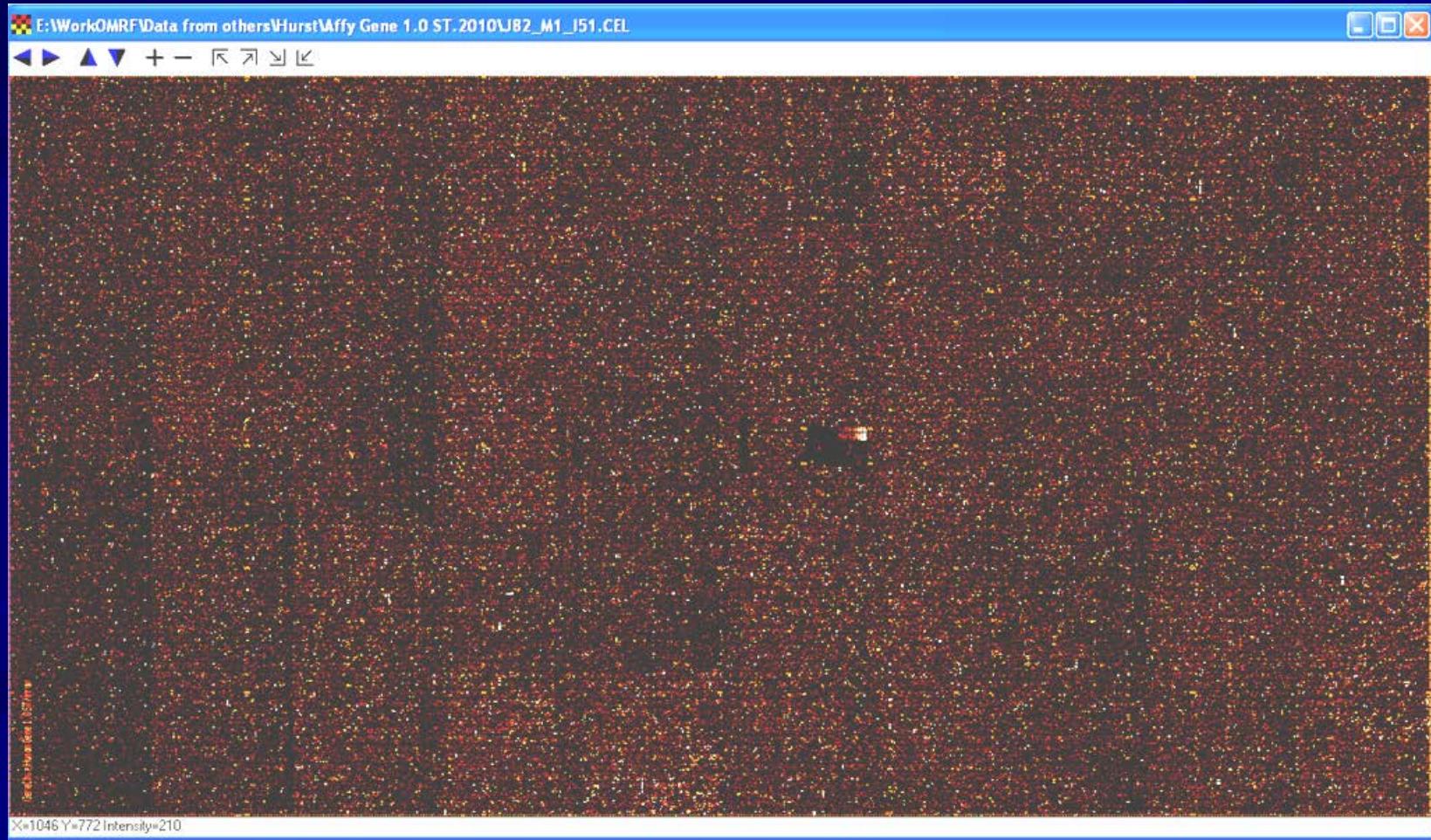
- Microarrays - new bioinformatics challenges
- Steps in microarray analysis
- Finding genes of interest
- Tools
- Time course analysis
- Exon array analysis

Microarrays



- Microarrays are a standardized technology for measuring the expression level of thousands of genes
- This present new statistical problems because the data are very high dimensional with very little replication (50,000 genes/10 replicates)

Microarray data format



Microarray data format

■ Raw data files



■ Numerical data

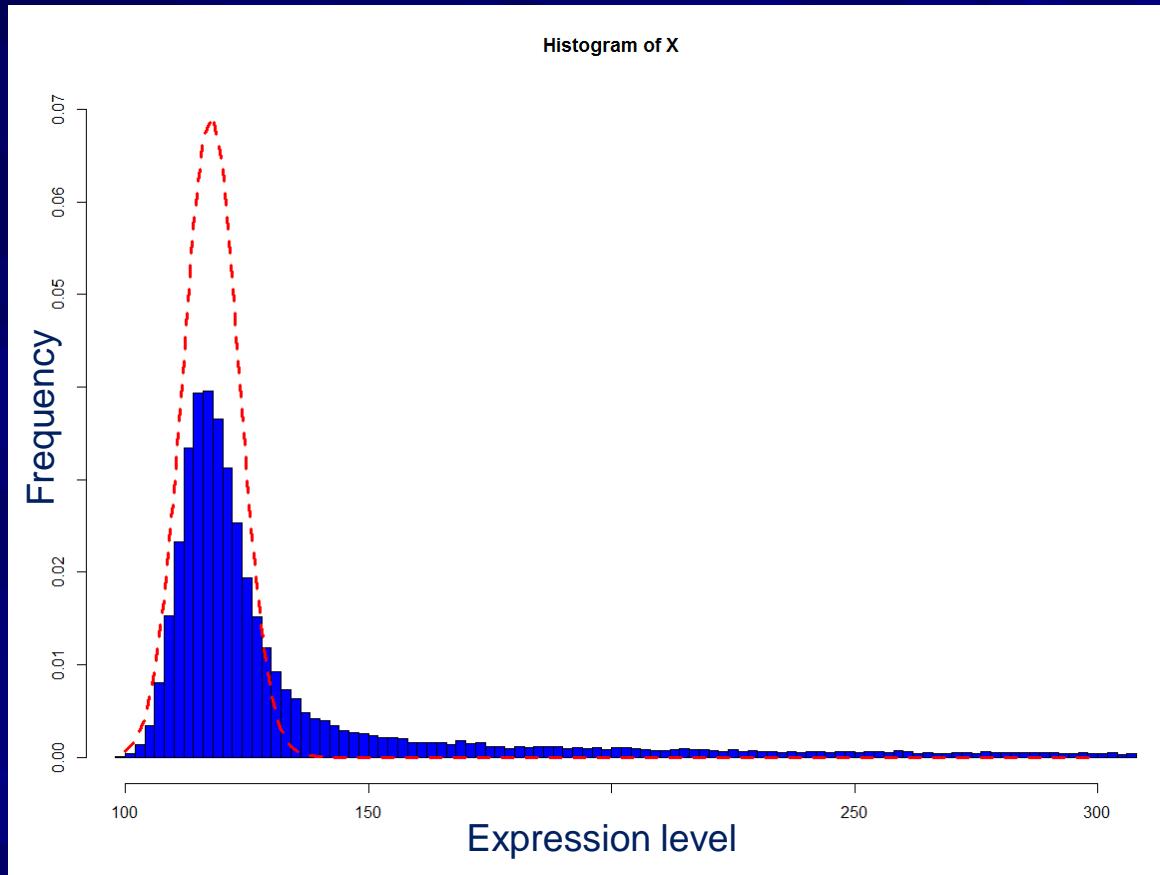
10	PROBE_ID	SYMBOL	504006100	504006100	504006100	504006100	504006100	504006100	5040061003_D.AVG_Signal		
11	ILMN_176	7A5	72.52097	67.3662	67.43669	60.3264	71.92233	60.27511			
12	ILMN_205	A1BG	87.25451	99.02529	88.08109	90.66778	102.9849	82.62512			
13	ILMN_173	A1BG	69.82148	72.1731	75.76444	65.86164	69.98101	65.98466			
14	ILMN_238	A1CF	76.60316	74.71108	70.72761	62.49332	72.25649	61.8455			
15	ILMN_180	A1CF	69.98457	79.65968	71.0647	64.39922	68.589	66.16525			
16	ILMN_177	A1CF	75.66752	74.74212	76.4988	69.93951	80.94523	75.94659			
17	ILMN_232	A26A1	82.16264	81.50092	81.0835	76.16673	88.16877	73.71473			
18	ILMN_167	A26A1	68.3719	69.37737	68.59947	65.18973	76.77369	66.23912			
19	ILMN_177	A26B1	69.36679	75.55676	66.57854	66.24841	64.60722	70.11909			
20	ILMN_173	A26C1B	74.72653	76.26254	78.0079	75.51186	80.17199	71.94649			
21	ILMN_165	A26C3	80.52338	88.53018	74.48835	88.8325	102.4532	84.26585			
22	ILMN_171	A26C3	63.41418	65.67715	58.444	66.48427	64.93271	57.78495			

Steps in microarray data analysis

Data preprocessing

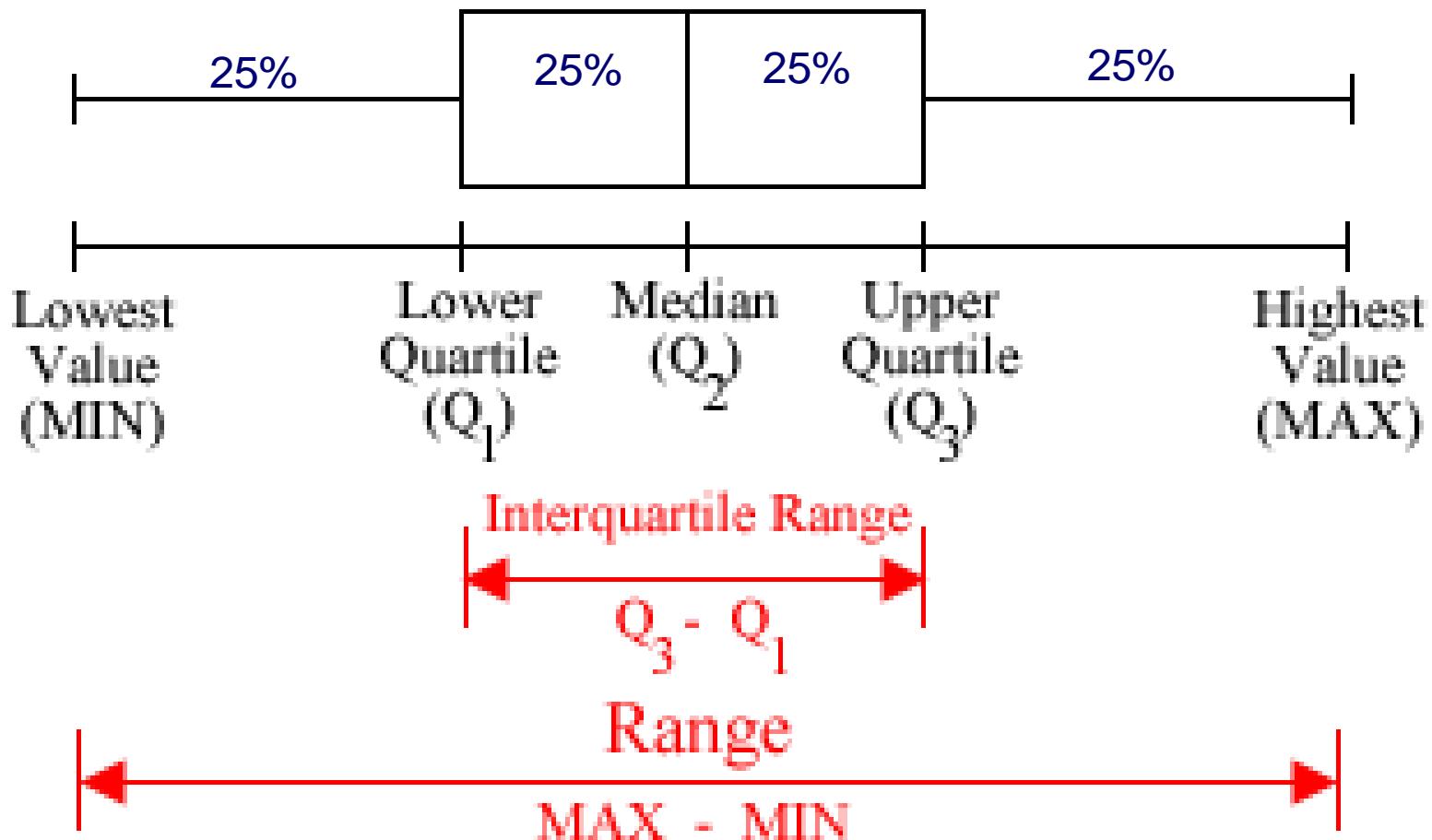
- The probe-level data must be cleaned and processed to obtain biologically meaningful measurement:
 1. Background correction: eliminate signals due to non-specific binding;
 2. Normalization: make multiple arrays comparable;
 3. Preprocessing: Log2 transformation, missing values imputation, etc.

Background correction

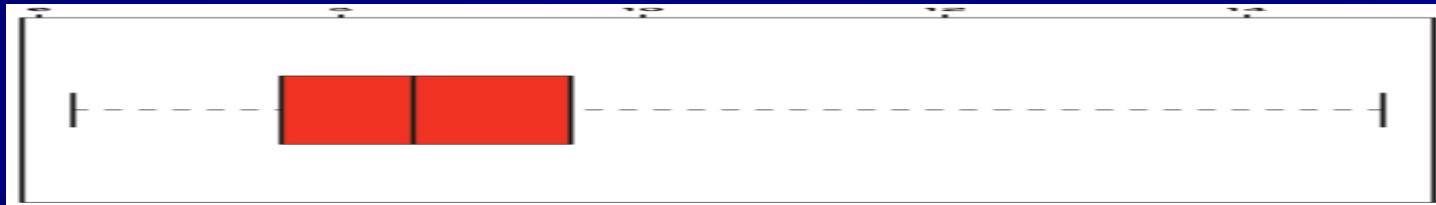
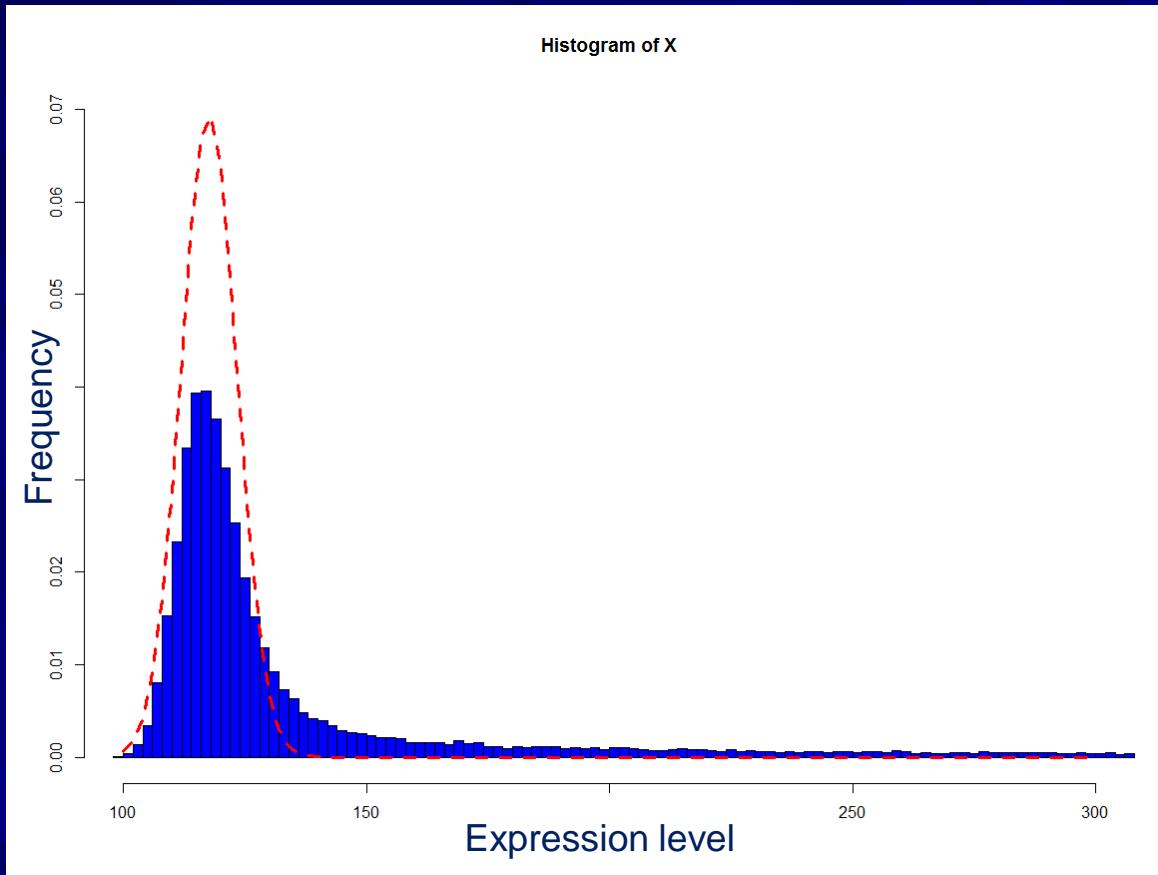


- Noise level – uncertainty about zero point

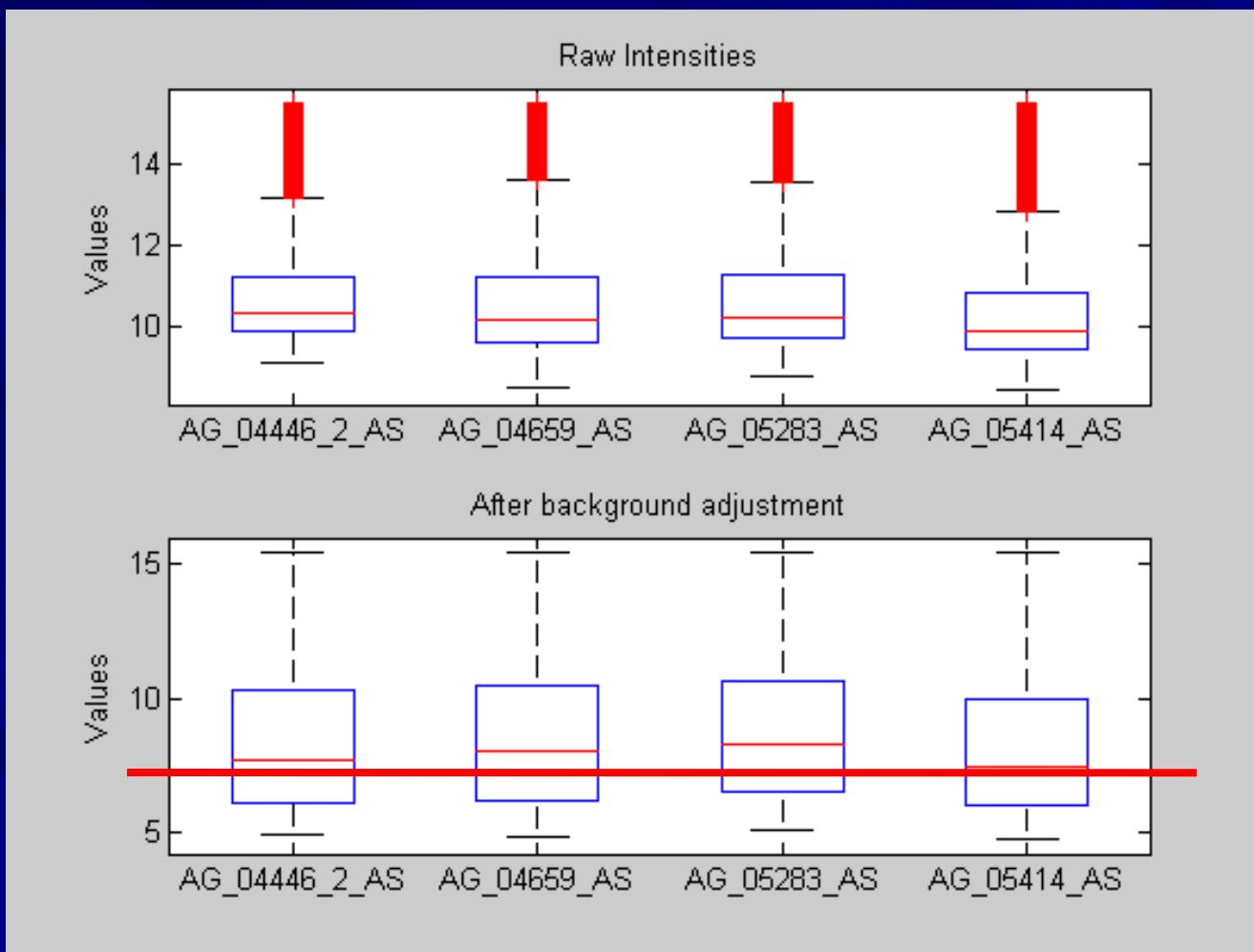
Box and whisker plot



Background correction



Background adjusted data



Why Normalize?

■ Sources of variation between multiple high-density oligonucleotide arrays:

➤ Biological (e.g., diseased vs. normal)

➤ Non-biological:

- Total RNA preparation, amplification
- Sample labeling differences
- Hybridization parameters
- Scanner differences
- Image analysis

■ Goal: make multiple arrays comparable.

Normalization Assumptions

- Most transcripts are not differentially expressed in response to a given stress.
 - Expression ratio of typical spot: Test/Control= 1
 - Unchanging spots are less interesting.
 - Outliers are biologically relevant.

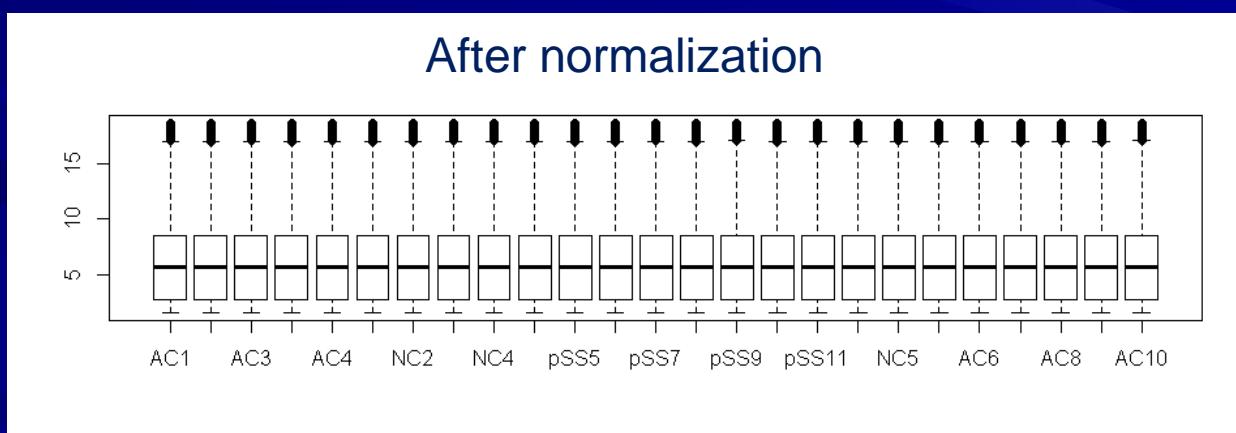
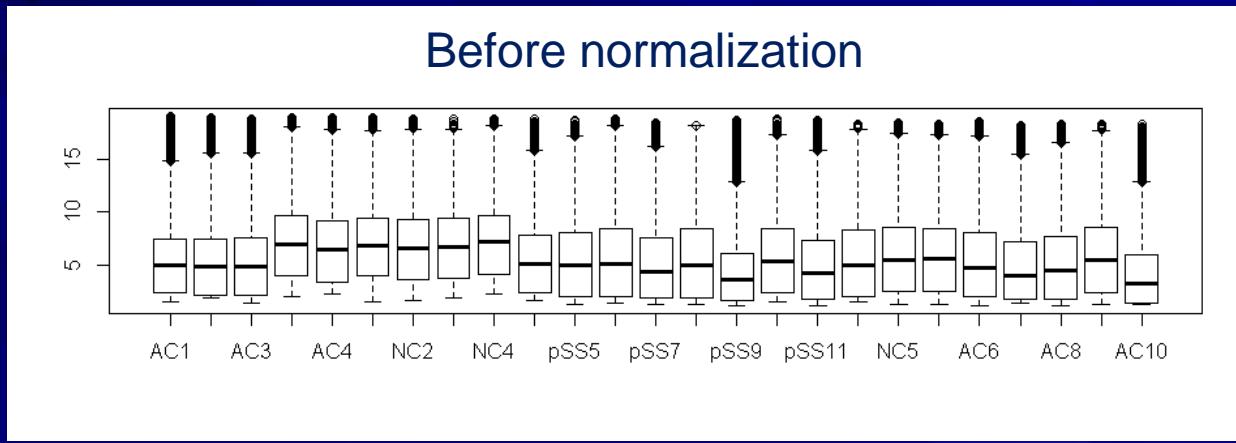
Normalization methods

■ Between-array normalization.

- **Scale:** Simply to scale the log-ratios to have the same median-absolute-deviation across arrays
- **MAS5, RMA:** Affymetrix
- **Cyclic Loess:** Affymetrix, CodeLink
- **Quantile normalization:** Ensures that the intensities have the same empirical distribution across arrays and across channels.
- **VSN:** Variance Stabilization Normalization. Illumina

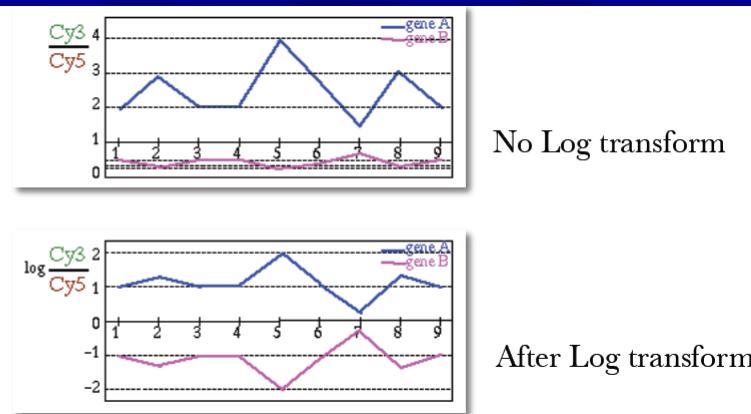
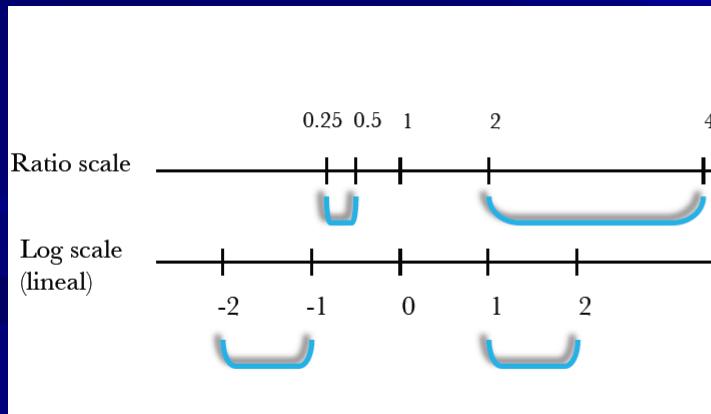
Quantile normalization

- Make distribution of data equal across all samples.



Log2 transformation

- Common technique used for two-color arrays (one-color as well)
- Log ratio transformation - convert data to a linear scale



Finding genes of interest

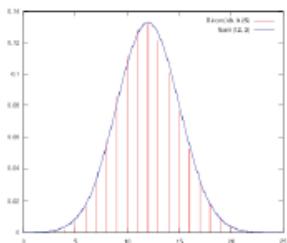
Nine steps for hypothesis testing

1. State the problem (What are differences between healthy and sick).
2. State the null (no differences) and alternative (diff) hypothesis.
3. Choose the level of significance (p-value=0.05).
4. Find the appropriate statistical model and test statistics (Welch t-test)
5. Calculate the appropriate test statistic.
6. Determine the p-value of the test statistic.
7. Compare the p-value with the chosen significance level (p<0.05?).
8. Reject (Yes, p<0.05) or do not reject H_0 based on the test above.
9. Answer the question in step 1.

Parametric and non-parametric test statistics

Parametric methods

- Assume that the data follow normal distribution.



$$N(\mu=12, \sigma=3)$$

T test.

Test difference in means between 2 independent populations with equal variances. Welch T-test for unequal variances.

Paired T test.

T test for paired data (blocks of 2 elements).

Example: Treatment in right arm, left arm as control

ANOVA.

Analysis of variance, for more than 2 populations.

Non parametric methods

- Appropriate when normality cannot be assumed.
- More robust (less sensitive to outliers).
- Less sensitive than parametric methods to detect significant changes.
- They order the data by expression, and use the rank to test.

Ex. Gene 63; 4 treatments and 5 controls; rank 1,2,3 ,4,5,6,7,8,9

Mann-Whitney test.

Test for differences in medians between two independent populations.

Wilcoxon Signed Rank test.

Non-parametric test equivalent to the paired T test for paired samples (test if median of paired differences is zero)

Kruskal-Wallis.

Non-parametric test equivalent to ANOVA for more than 2 populations.

Multiple testing problem

- With thousands of genes on a microarray we're not testing one hypothesis, but many hypotheses – one for each gene
- Analysis of 20,000 genes using commonly accepted significance level $\alpha=0.05$ will identify 1,000 differentially expressed genes simply by chance

Adjustment for multiple testing

- Control of Type I Family Wise Error Rate (FWER) - conservative
- Bonferroni
- Holm's Bonferroni Step-Down
- Westfall & Young permutation
- Control of False Discovery Rate (FDR) – less conservative
- Benjamini & Hochberg, or Benjamini & Yekutieli

Microarray data storage

ArrayExpress

<http://www.ebi.ac.uk/microarray-as/ae/>

The screenshot shows the homepage of the ArrayExpress website. At the top, there is a navigation bar with links for 'Databases', 'Tools', 'EBI Groups', 'Training', 'Industry', 'About Us', 'Help', 'Site Index', and 'Give us feedback'. Below the navigation bar, the main content area features the 'ARRAYEXPRESS' logo. A brief description explains that the Archive is a database of functional genomics experiments, including gene expression, and can be queried and downloaded according to MIAME and MINSEQE standards. It also mentions the 'Gene Expression Atlas' which contains curated and re-annotated data. The page is divided into several sections: 'Experiments Archive' (15261 experiments, 424790 assays), 'Gene Expression Atlas' (1504 experiments, 44436 assays, 9346 conditions), 'News' (with links to 'Global Expression Space' and 'A global map of human gene expression'), and 'Links' (with links to various EBI resources like User Survey, Help, Submit Data, Programmatic Access, Software Downloads, EFO, Bioconductor Package, Quality Metrics, Scientific Advisory Board, and Functional Genomics Group). The footer includes links for Terms of Use, EBI Funding, Contact EBI, and a copyright notice for the European Bioinformatics Institute 2010.

EMBL-EBI EB-eye Search All Databases Go Reset ? Advanced Search Give us feedback

Databases Tools EBI Groups Training Industry About Us Help Site Index

ARRAYEXPRESS

The **ArrayExpress Archive** is a database of functional genomics experiments including gene expression where you can^I query and download data collected to **MIAME** and **MINSEQE** standards. **Gene Expression Atlas** contains a subset of curated and re-annotated Archive data which can be queried for individual gene expression under different biological conditions across experiments.

Experiments Archive
15261 experiments, 424790 assays

Experiment, citation, sample and factor annotations

Browse experiments Advanced query interface

Submitter/reviewer login ArrayExpress Query Help

Gene Expression Atlas
1504 experiments, 44436 assays, 9346 conditions

Genes Conditions
Any species

[Gene Expression Atlas Home](#)

News

- 22 Apr 2010 - **Global 'Expression Space'**
EBI-Helsinki Team Integrates Array Data from Thousands of Samples to Map Global 'Expression Space'...[more](#)
- 09 Apr 2010 - **A global map of human gene expression**
By integrating gene expression data from a large variety of human tissue samples, a global map of human gene expression is produced. For more details, please see the [Nature Biotechnology](#) [PDF - 676KB] or [EMBL press release](#) [PDF - 148KB].

[ArrayExpress User Survey](#)
[Help | Training | FAQ | Citing](#)
[Submit Data](#) (array based and re-sequencing)
[Programmatic Access | FTP Access](#)
[Software Downloads and Statistics](#)
[EFO | Bioconductor Package | Quality Metrics](#)
[ArrayExpress Scientific Advisory Board](#)
[Functional Genomics Group](#)

Terms of Use | EBI Funding | Contact EBI | © European Bioinformatics Institute 2010. EBI is an Outstation of the European Molecular Biology Laboratory.

Gene Expression Omnibus

<http://www.ncbi.nlm.nih.gov/geo/>

The screenshot shows the GEO homepage with a blue header featuring the NCBI logo and the GEO logo. The main content area includes a navigation menu, a central query/browsing interface, and a site contents sidebar.

GEO navigation

QUERY

- DataSets
- Gene profiles
- GEO accession
- GEO BLAST

BROWSE

- DataSets
- GEO accessions
 - Platforms
 - Samples
 - Series

Site contents

Public data

Platforms	7,844
Samples	476,338
Series	18,780

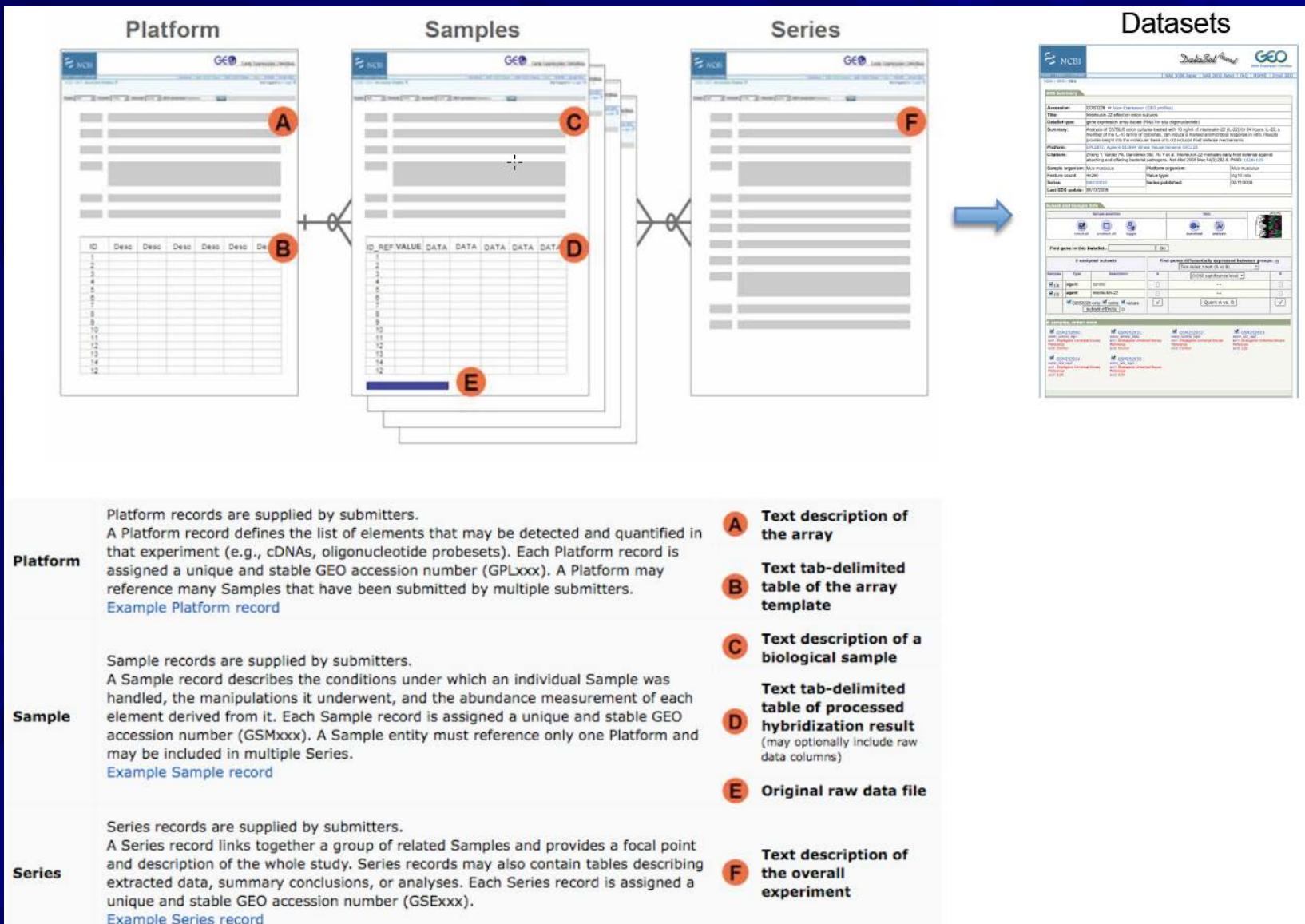
Documentation

- Overview | FAQ | Find
- Submission guide
- Linking & citing
- Journal citations
- Programmatic access
- DataSet clusters
- GEO announce list
- Data disclaimer
- GEO staff

Query & Browse

- Repository browser
- Submitters

GEO Database Organization



Tools to help you out in
microarray data analysis

Methods for data preprocessing and finding genes of interest

- The R Project for Statistical Computing,
<http://www.r-project.org>
Bioconductor, open source software for
bioinformatics <http://www.bioconductor.org>
- Matlab - The Language of Technical
Computing <http://www.mathworks.com/>
- Online tools

Gene IDs converters

Clone/Gene ID converter

<http://idconverter.bioinfo.cnio.es/IDconverter.php>

Input

- ▶ Organism: Human
- ▶ Identifier: Gene Name (HUGO)
- ▶ List of Identifiers:
* Put only one id by line.

Get your results on a spreadsheet file.
 Get your results on a text file (tab separated).

▶ Choose the information that you would like to include in your report:

Gene Level

<input checked="" type="checkbox"/> UniGene Cluster Taken from UniGene	<input type="checkbox"/> UniGene Taken from UniGene	<input checked="" type="checkbox"/> Gene Description Taken from UniGene	<input checked="" type="checkbox"/> Gene Name Taken from UniGene	<input checked="" type="checkbox"/> NCBI Taken from UniGene	EntrezGene Taken from UniGene
<input checked="" type="checkbox"/> Ensembl Gene Taken from Ensembl	<input type="checkbox"/> NCBI Taken from Ensembl	<input type="checkbox"/> RefSeq_RNA Taken from Ensembl	<input type="checkbox"/> RefSeq_peptide Taken from Ensembl	<input type="checkbox"/> NCBI Taken from Ensembl	CCDS Taken from Ensembl

Gene Location

<input type="checkbox"/> Ensembl	<input type="checkbox"/> UCSC	<input type="checkbox"/> UCSC*	<input checked="" type="checkbox"/> e! If nonexistent, uses UCSC	<input type="checkbox"/> UCSC	<input type="checkbox"/> UCSC*
----------------------------------	-------------------------------	--------------------------------	---	-------------------------------	--------------------------------

Clone Level

<input type="checkbox"/> CNIO ID CNIO Clone Tracking Database	<input checked="" type="checkbox"/> UniGene Taken from UniGene	<input checked="" type="checkbox"/> Clone ID Taken from UniGene	<input checked="" type="checkbox"/> NCBI Taken from UniGene	GenBank Accession Taken from UniGene	<input type="checkbox"/> Affymetrix ID Taken from Ensembl
--	---	--	--	---	--

Protein Level

<input type="checkbox"/> SwissProt Taken from Ensembl	<input type="checkbox"/> Embl Taken from Ensembl	<input type="checkbox"/> PDB Taken from Ensembl	<input type="checkbox"/> PDB id Taken from Ensembl	<input type="checkbox"/> IPI Taken from Ensembl
--	---	--	---	--

Functional Level

<input type="checkbox"/> GO Taken from Ensembl	<input type="checkbox"/> KEGG Pathways Taken from KEGG	<input type="checkbox"/> R Taken from Reactome	Reactome Pathway Taken from Reactome	<input type="checkbox"/> R Taken from Reactome	Reactome Reaction Taken from Reactome
---	---	---	---	---	--

PubMed id
Taken from NCBI

Information about genes

GeneCards - comprehensive database
about genes

<http://www.genecards.org>

The screenshot shows the GeneCards homepage. At the top right, it says "Version 3". To the right, there is a logo for "WEIZMANN" with a stylized "W" icon. The main title "GeneCards®" is displayed prominently in orange. Below the title, the subtitle "The Human Gene Compendium" is shown. A navigation bar at the bottom has four items: "Search (homepage)", "GeneCards Guide", "Suite", and "Terms and Conditions". A "Provided by" link is also visible. A search bar at the bottom allows users to search by keyword(s) and includes a dropdown menu and a "Search" button.

Version 3

WEIZMANN

GeneCards®

The Human Gene Compendium

Provided by

Search (homepage) GeneCards Guide Suite Terms and Conditions

Search the GeneCards human gene database

Search by keyword(s) for Search

Information about genes

GATA4 Gene
protein-coding GC08P011599

GATA binding protein 4
Symbol approved by the HUGO Gene Nomenclature Committee (HGNC) database

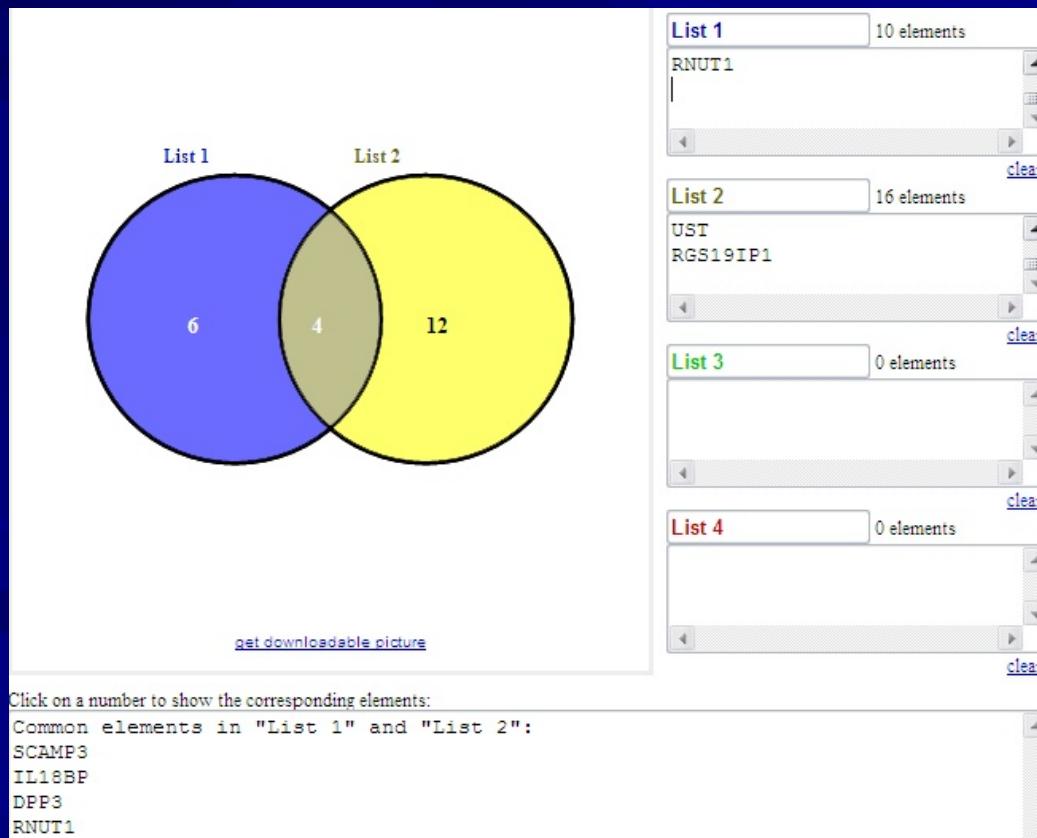
invitrogen Gene
Antibodies / RNAi / Pathways

GATA4 Aliases & Descriptions (According to ¹ HGNC, ² Entrez Gene, ³ UniProt/Swiss-Prot, ⁴ UniProt/TrEMBL, ⁵ OMIM, ⁶ GeneLoc , and/or ⁷ Ensembl) About This Section <input type="button" value="Jump to Section..."/>	Aliases MGC126629 ² OTTHUMP00000116010 ² Search outside databases for aliases for GATA4 Previous GC identifiers: GC08P011628 GC08P011599	Description GATA bindin GATA-bindin GATA-bindin Jump to Section...	<input type="button" value="Jump to Section..."/> Aliases for GATA4 Databases for GATA4 Disorders for GATA4 Domains/Families for GATA4 Drugs/Compounds for GATA4 Expression for GATA4 Function for GATA4 Location for GATA4 Medical News for GATA4 Orthologs for GATA4 Paralogs for GATA4 Pathways/Interactions for GATA4 Proteins for GATA4 Research Articles for GATA4 SNPs for GATA4 Search Box for GATA4 Services for GATA4 Summaries for GATA4 Technologies for GATA4 Transcripts for GATA4 TOP
GATA4 Summaries  (According to Entrez Gene, Wikipedia's Gene Wiki, UniProt/Swiss-Prot, and/or UniProt/TrEMBL) About This Section <input type="button" value="Jump to Section..."/>	EntrezGene summary for GATA4:  This gene encodes a member of the GATA family. Members of this family recognize the GATA motif which is thought to regulate genes involved in embryonic development and function. Mutations in this gene have been reported to cause various diseases. UniProt/Swiss-Prot: GATA4_HUMAN , P43694  Function: Transcriptional activator. Binds to the consensus sequence 5'-AGATAG-3'. Acts as a transcriptional activator of ANF in cooperation with NKX2-5 (By similarity)	 Gene Wiki entry for GATA4 	ctors. Members of this family are involved in initiation and regulation of transcription.
GATA4 Genomic Location	Genomic View: UCSC Golden Path with GeneCards custom track		

Comparison of gene lists

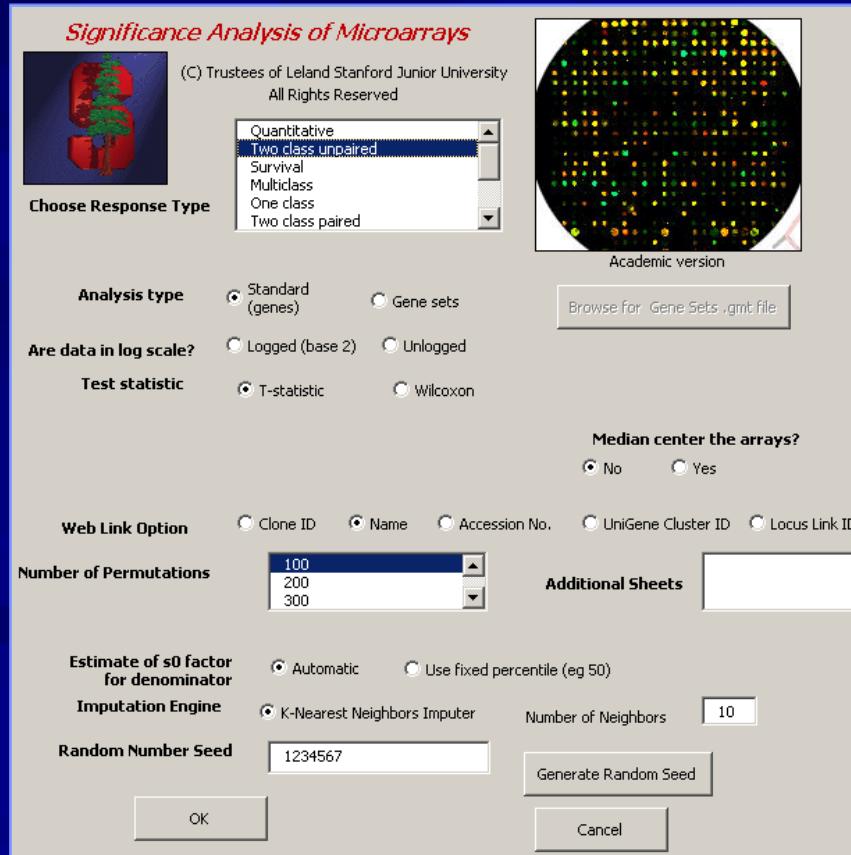
VENNY - An interactive tool for comparing lists with Venn Diagrams

<http://bioinfogp.cnb.csic.es/tools/venny/index.html>



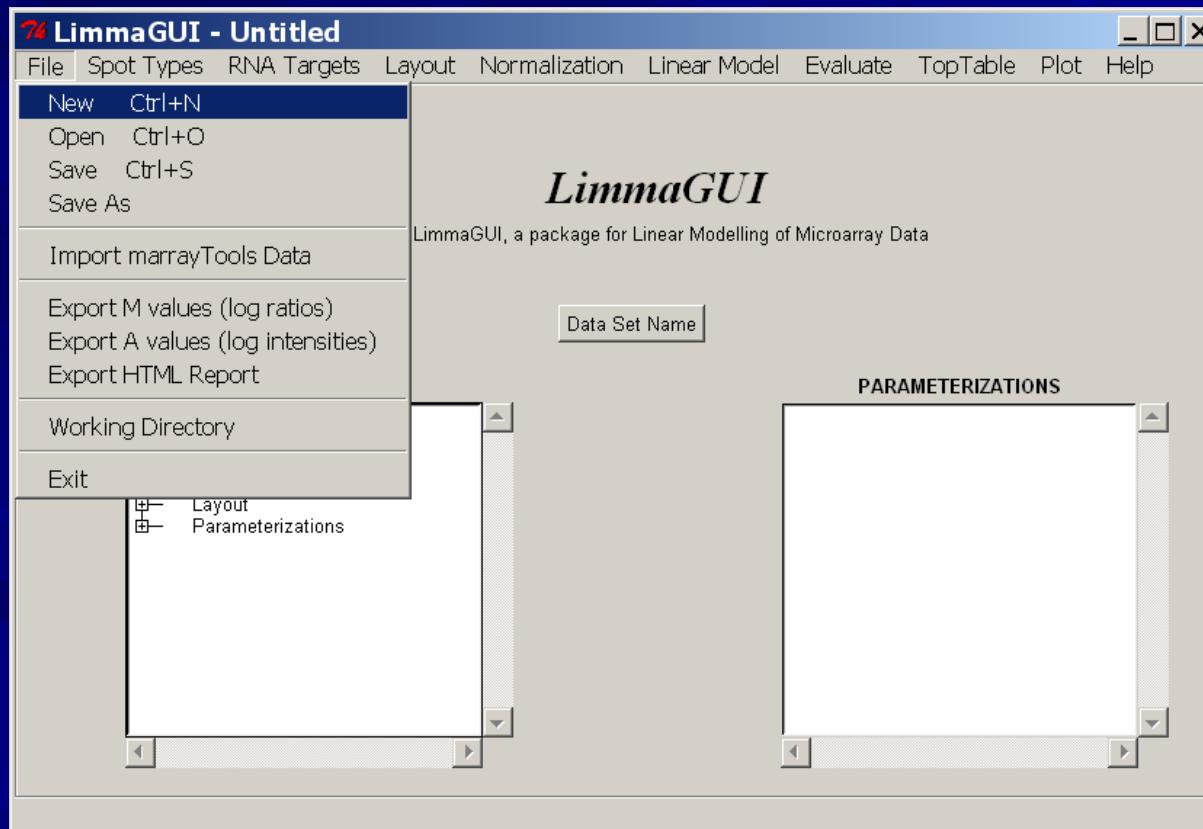
Identification of differentially expressed genes

- SAM – significance analysis of microarrays <http://www-stat.stanford.edu/~tibs/SAM/>



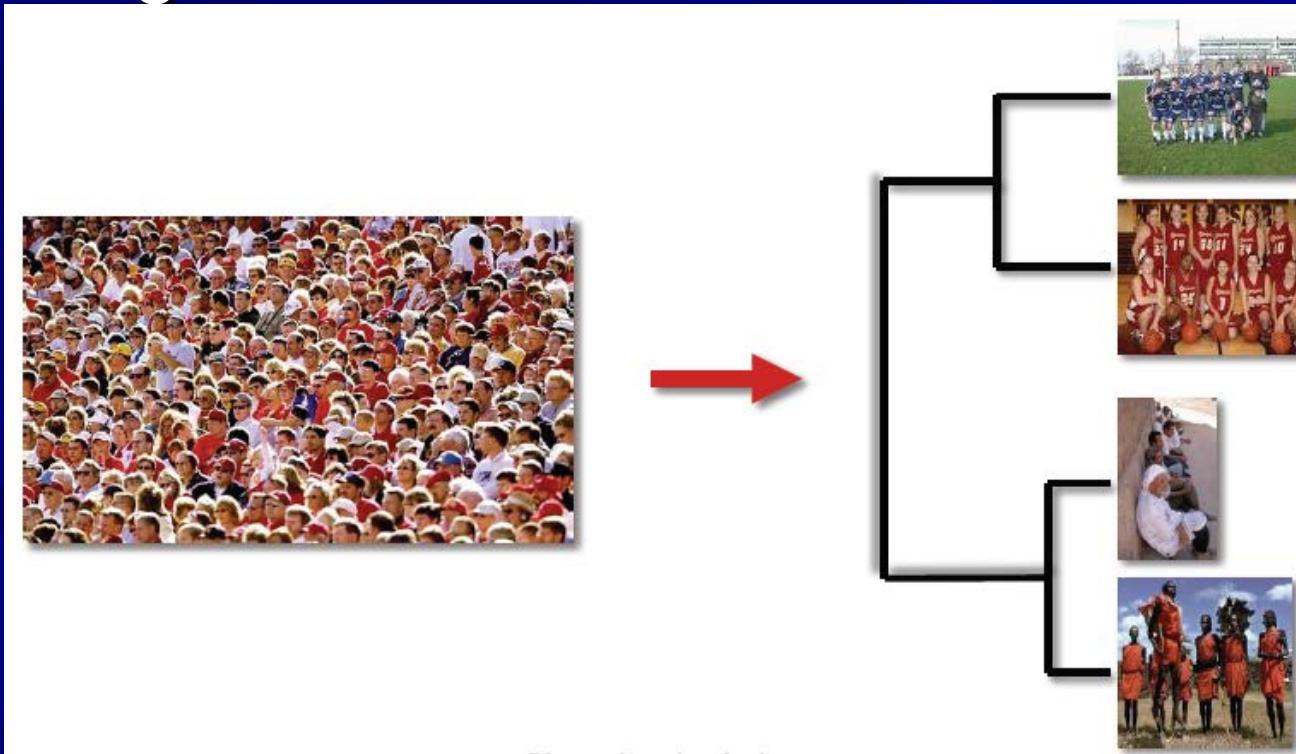
Identification of differentially expressed genes

■ LIMMA – Linear Modeling of Microarray Data (Bioconductor)



Clustering

- Partitioning of a data set into subsets.
- A cluster is a group of relatively homogeneous cases or observations

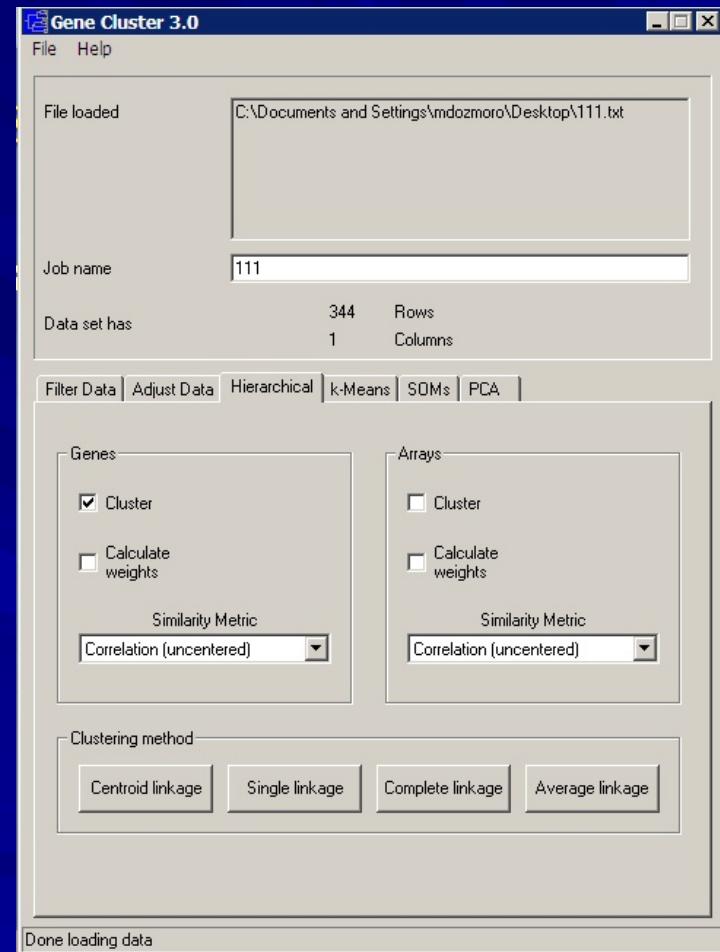


Clustering methods

- ❖ Hierarchical methods:
 - Agglomerative:
 - UPGMA (Sneath and Sokal, 1973)
- ❖ Divisive:
 - SOTA (Herrero et al. 2001)
 - Dlvisive ANAlysis clustering (Kaufman & Rousseeuw, 1990)
 - Gene Shaving (Hastie et al, 2000)
- ❖ • Non-hierarchical methods:
 - kmeans (Hartigan and Wong 1979)
 - kmedians (Hartigan and Wong 1979)
 - SOM (Kohonen 1979, Tamayo et al 1999)
 - Fuzzy c-means (Dougherty et al. 2002)
 - Probabilistic clustering (Bhattacharjee et al. 2001)

Clustering and visualization of genes of interest

GeneCluster 3.0 -
Next version of
GeneCluster
program developed
by Michael Eisen

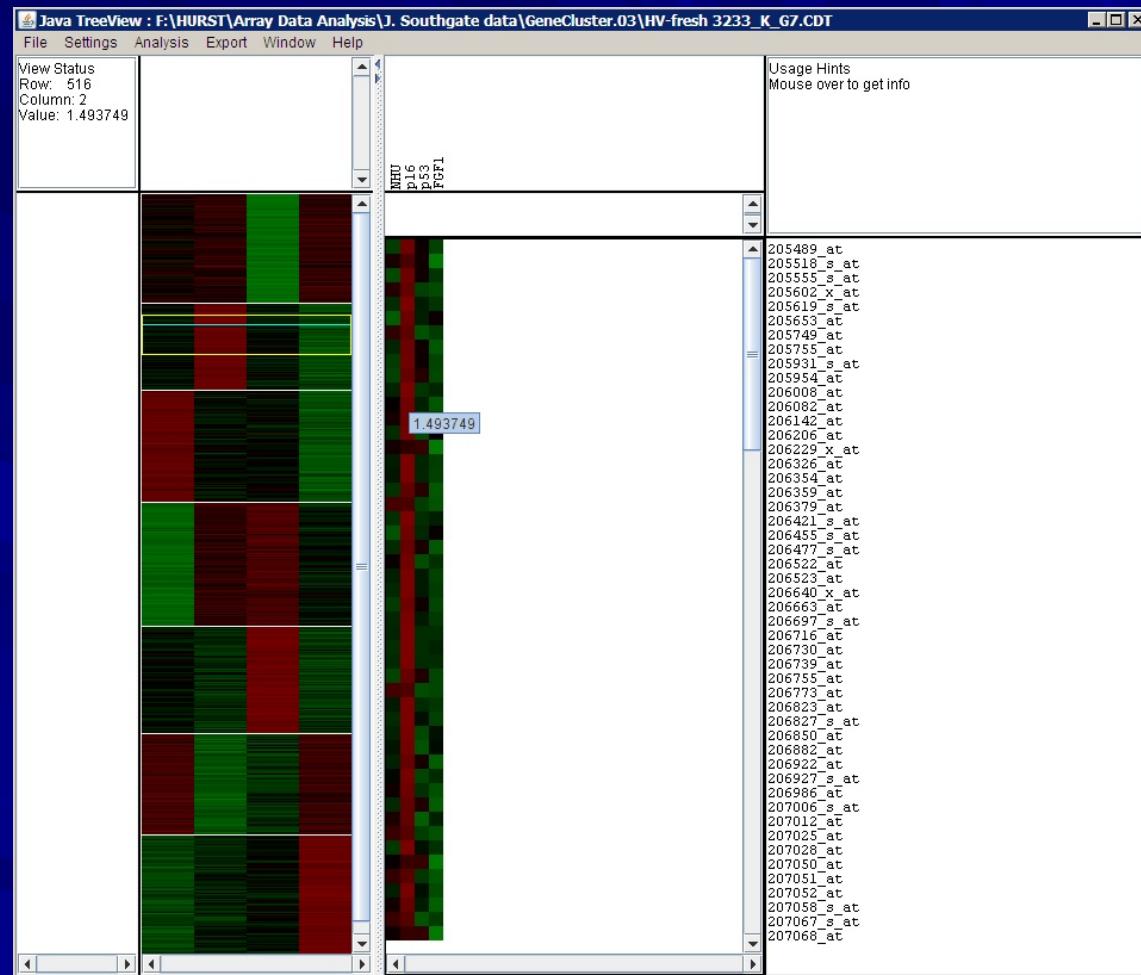


<http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster>

Clustering and visualization of genes of interest

Java TreeView –
Advanced version
of visualization
program
developed by
Michael Eisen

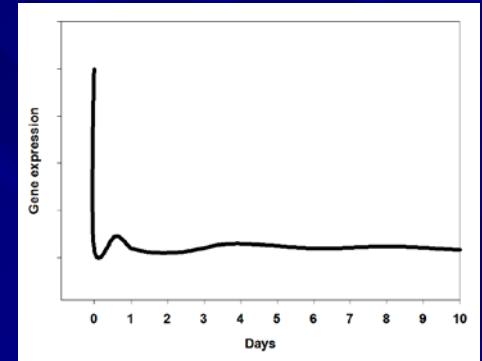
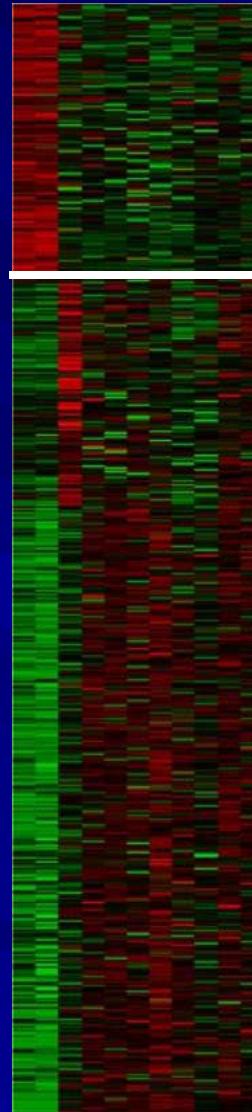
Red/Green –
high/low expression
level, respectively



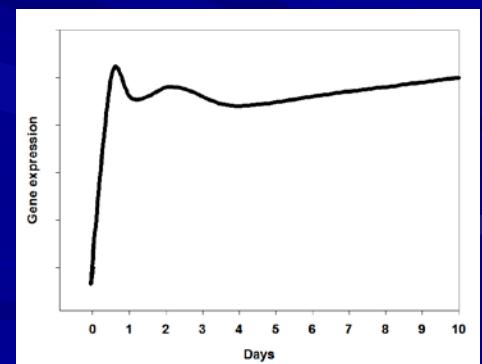
Clustering interpretation

- Hierarchical clustering – genes with similar expression profiles clustered together

Red/Green – high/low expression level, respectively



Genes turned off



Genes turned on

Gene Ontology

- Aim - to unify the representation of genes and gene product' attributes across all species
- The three categories of GO
 - Molecular Function
 - Biological Process
 - Cellular Component

GO tree structure

```
graph TD; Root[GO:0003673 : Gene Ontology (65883)] --> BP[GO:0008150 : biological process (44405)]; Root --> CP[GO:0009987 : cellular process (32672)]; Root --> MF[GO:0003674 : molecular function (53910)]; BP --> Behavior[GO:0007610 : behavior (357)]; BP --> CP; Behavior --> Unknown[GO:0000004 : biological process unknown (7877)]; CP --> CellCommunication[GO:0007154 : cell communication (5384)]; CP --> CellDeath[GO:0008219 : cell death (744)]; CP --> CellDifferentiation[GO:0030154 : cell differentiation (464)]; CP --> CellGrowth[GO:0008151 : cell growth and/or maintenance (28802)]; CP --> CellMotility[GO:0006928 : cell motility (911)]; CP --> MembraneFusion[GO:0006944 : membrane fusion (257)]; MF --> Death[GO:0016265 : death (793)]; MF --> Development[GO:0007275 : development (4615)]; MF --> Obsolete[GO:0008371 : obsolete (1581)]; MF --> PhysiologicalProcesses[GO:0007582 : physiological processes (31124)]; MF --> ViralLifeCycle[GO:0016032 : viral life cycle (115)]; MF --> CellularComponent[GO:0005575 : cellular component (32869)]; MF --> MolecularFunction[GO:0003674 : molecular function (53910)];
```

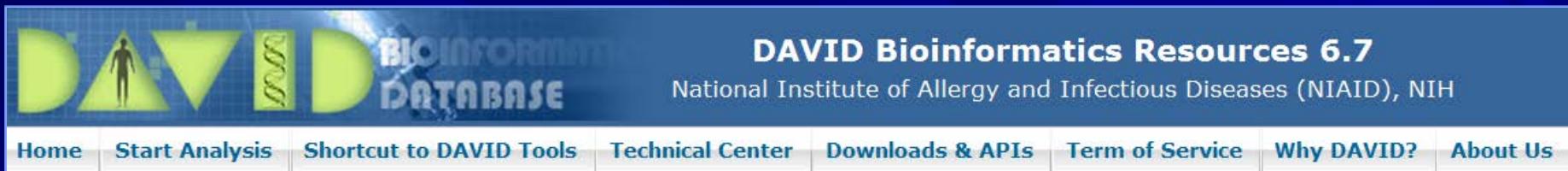
IS_A
relation

PART_OF
relation

Gene Ontology analysis

DAVID - The Database for Annotation,
Visualization and Integrated Discovery

<http://david.abcc.ncifcrf.gov/>



- Find and cluster genes with overrepresented ontologies
as compared with pre- or user-defined background

Gene ontology enrichment

DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Functional Annotation Clustering

Current Gene List: demolist1
Current Background: Alkaliphilus metallireducens
154 DAVID IDs

Help and Manual

Options Classification Stringency Medium ▾

Rerun using options Create Sublist

5 Cluster(s)

Annotation Cluster 1	Enrichment Score: 0.41	G		Count	P_Value	Benjamin
<input type="checkbox"/> GOTERM_BP_FAT	regulation of transcription, DNA-dependent	RT	■	8	2.6E-1	1.0E0
<input type="checkbox"/> GOTERM_BP_FAT	regulation of RNA metabolic process	RT	■	8	2.6E-1	1.0E0
<input type="checkbox"/> GOTERM_BP_FAT	regulation of transcription	RT	■	8	3.1E-1	1.0E0
<input type="checkbox"/> INTERPRO	Signal transduction response regulator, receiver region	RT	■	3	3.8E-1	1.0E0
<input type="checkbox"/> SMART	REC	RT	■	3	4.4E-1	1.0E0
<input type="checkbox"/> GOTERM_MF_FAT	two-component response regulator activity	RT	■	3	4.8E-1	1.0E0
<input type="checkbox"/> GOTERM_BP_FAT	two-component signal transduction system (phosphorelay)	RT	■	3	7.3E-1	1.0E0

Download File

Gene interactions

PubGene - networks, associations and more
<http://www.pubgene.org>

Bio Networks Bio Associations Sequence Homology My Workspace Login

Browse literature or sequence neighbours.

Organism: Homo sapiens Select one or "all organisms"

Gene / Protein *: TGFB Terms separated by comma, eg. *BRCA1*, *LEP*

and/or

Biological term: prostate cancer One term, eg. *Alzheimer Disease*. [Supported categories](#)

Network displays genes/proteins important for the biological term

[Submit](#) [Advanced Options](#) | [Network Parameters](#)

1 Choose a match and get to a literature network

2 Retrieve literature for the biological relationship

3 Annotate your gene network with biological association

4 Click on a gene to browse literature through networks

Smith-Waterman as fast as BLAST

The diagram illustrates a gene interaction network centered around the gene *BRCA2*. Nodes are represented by circles: red for proteins and black for other entities like genes or complexes. Edges represent interactions with numerical weights. Key nodes include *RB1*, *TP53*, *ATM*, *RAD50*, *BCL2*, *ESR1*, *BRCA1*, and *PGR*. Annotations provide biological context for these interactions, such as *tumor suppressor*, *on receptor binding activity*, *DNA repair protein*, and *caspase-3 activity*. A sidebar shows a list of pattern matches for *BRCA2* and its interacting proteins. A small inset shows a zoomed-in view of the network. At the bottom right, a bar chart compares the performance of the Smith-Waterman algorithm to BLAST.

Gene interactions

iHOP - Information Hyperlinked over Proteins

<http://www.ihop-net.org/UniPub/iHOP/>

The screenshot displays the iHOP interface with a central network graph and four main sections: PHYSIOLOGY, INTERACTION, PATHOLOGY, and PHENOTYPE.

- PHYSIOLOGY:** Describes the structure and dynamics of SH3 domains from iHOP. It mentions that the iHOP database contains 1,500 distinct proteins and 1,500 distinct protein domains, with 13,000 distinct interactions. A figure shows a network of interactions between KIF1B and its domains.
- INTERACTION:** Focuses on CD4, a protein involved in T-cell receptor signaling. It discusses the phosphorylation of CD4 by CD42 kinase and its role in T-cell activation.
- PATHOLOGY:** Discusses microbial pathogens and cytoskeletal function. It highlights how iHOP integrates data from various sources like PubMed and UniProt to provide a comprehensive view of pathogenesis.
- PHENOTYPE:** Shows a network of genes related to PubMe (PubMed) and KIF1B, illustrating how iHOP links gene interactions to observable phenotypes.

Search Gene: A search bar for entering gene names.

Gene Model: A link to the gene model interface.

Developer's Zone: A link to the developer's zone.

How to cite iHOP: Instructions for citing the iHOP resource.

Contact: Contact information.

Links: External links.

Help: Help documentation.

COCOON: COCOMONDO CONSORTIUM logo.

Hoffmann, R., Valencia, A. A Gene Network for Navigating the Literature. Nature Genetics 36, more than 2,700 organisms, 110,000 genes, 22.3 million sentences. ...always up to date – every day.

Search for a gene synonym or accession number... (Click here for an example: SNF1)

in

Pathway analysis

www.genome.jp/kegg/pathway.html

BioGPS - your Gene P... PubMed home CloneGene ID Convers... DAVID Functional Ann... GeneCards V3 - Huma...



KEGG PATHWAY Database
Wiring diagrams of molecular interactions, reactions, and relations

KEGG2 PATHWAY BRITE DISEASE DRUG KO GENES GENOME LIGAND DBGET

Select prefix Enter keywords

Pathway Maps

KEGG PATHWAY is a collection of manually drawn pathway maps (see [new maps](#), [change history](#), and [last updates](#)) representing our knowledge on the molecular interaction and reaction networks for:

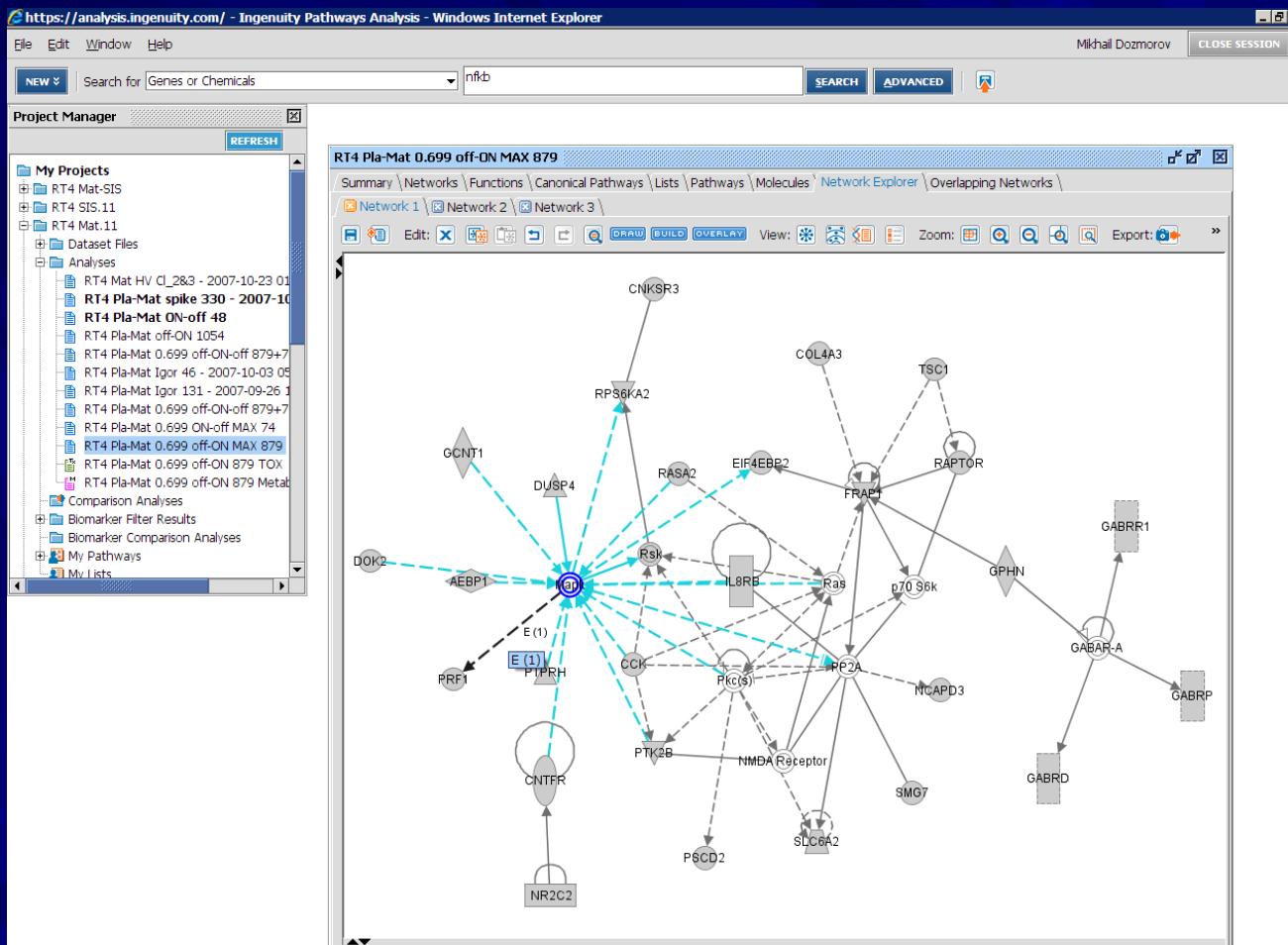
- 0. Global Map**
- 1. Metabolism**
Carbohydrate Energy Lipid Nucleotide Amino acid Other amino acid Glycan Cofactor/vitamin Terpenoid/PK Other secondary metabolite Xenobiotics Overview
- 2. Genetic Information Processing**
- 3. Environmental Information Processing**
- 4. Cellular Processes**
- 5. Organismal Systems**
- 6. Human Diseases**

and also on the structure relationships (KEGG drug structure maps) in:

- 7. Drug Development**

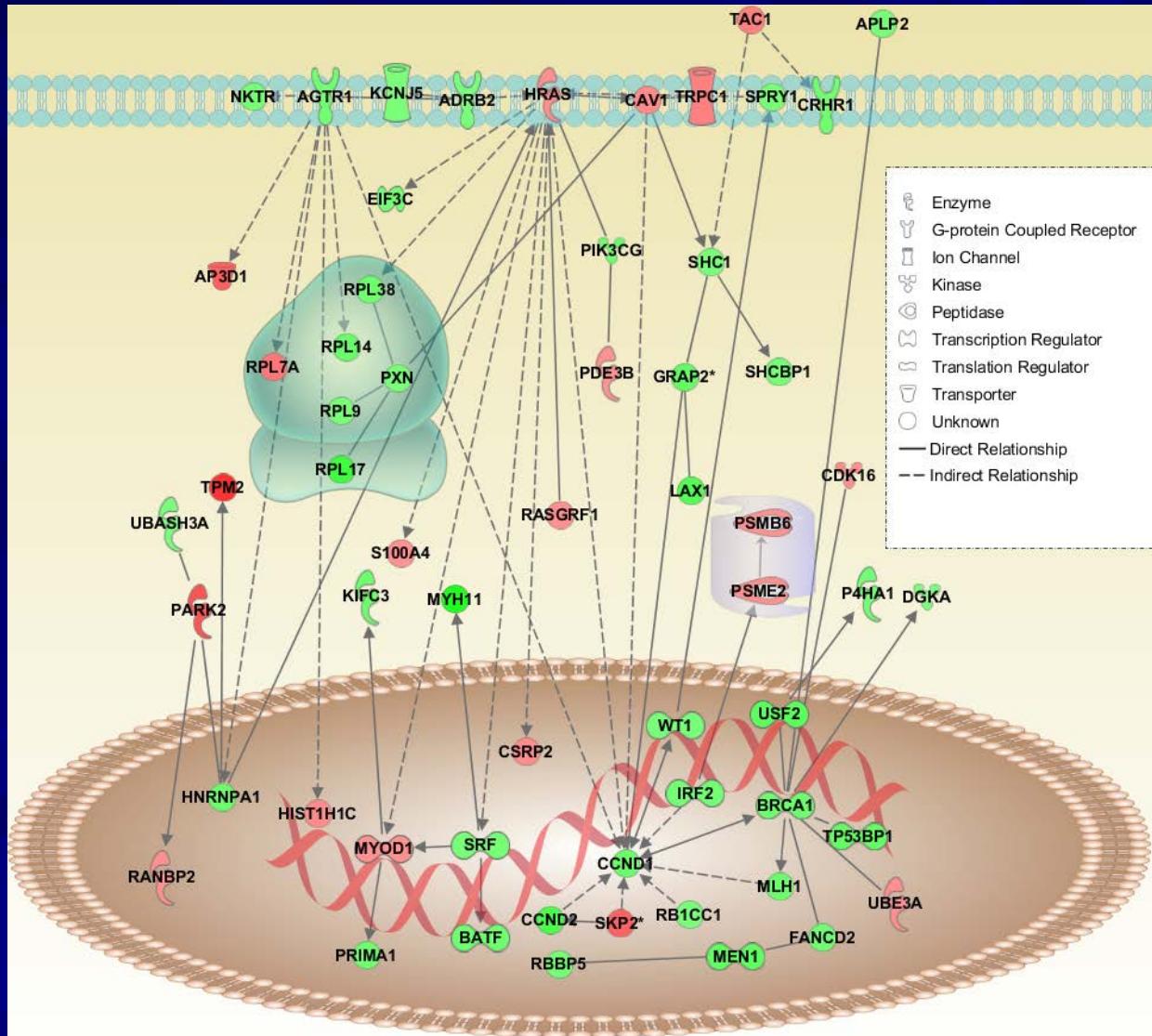
KEGG Atlas may now be used to examine any of the KEGG pathway maps.

Ingenuity Pathway Analysis



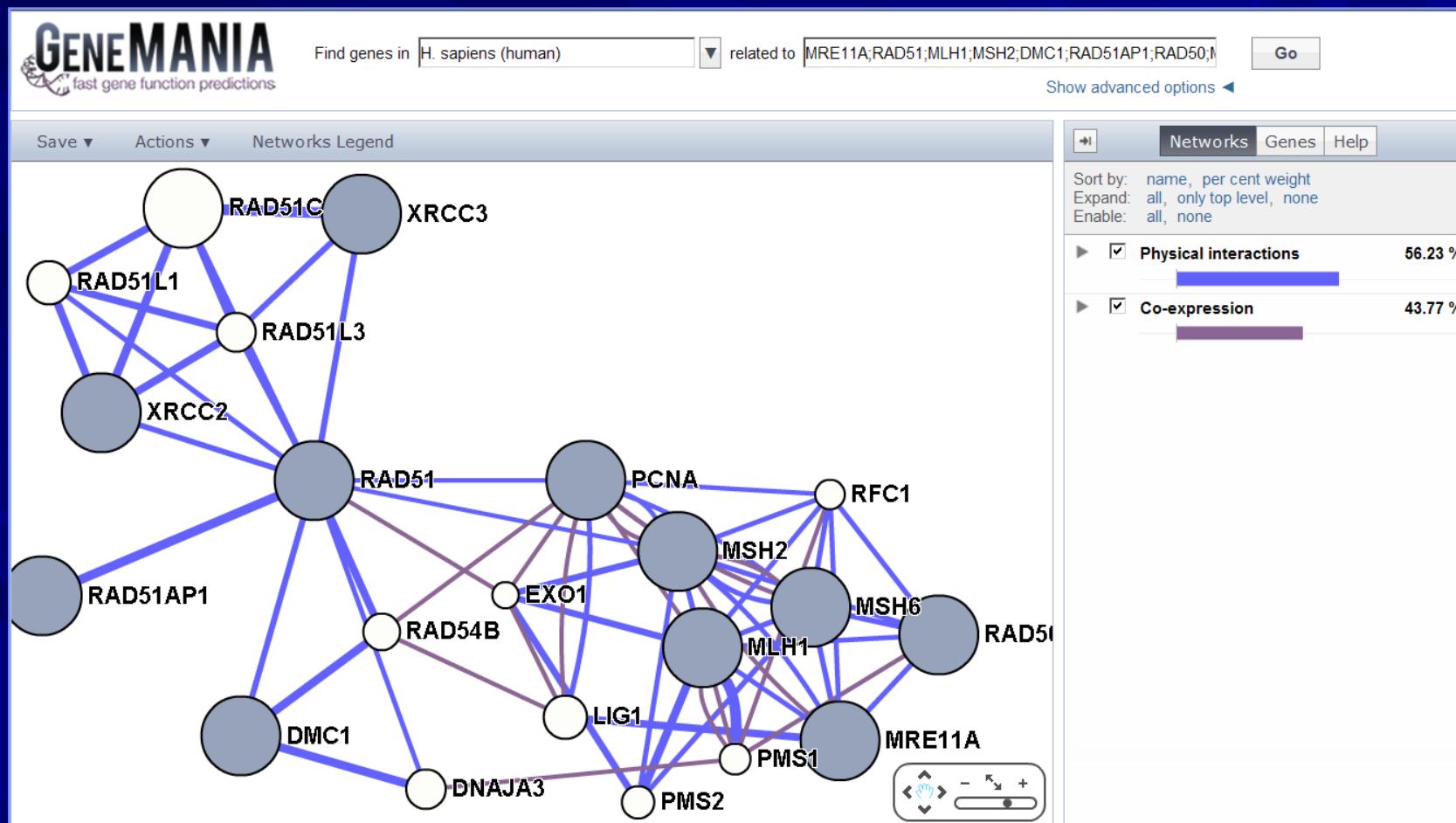
■ INGENUITY – analysis and visualization of overrepresented gene interactions <http://www.ingenuity.com/>

Ingenuity network

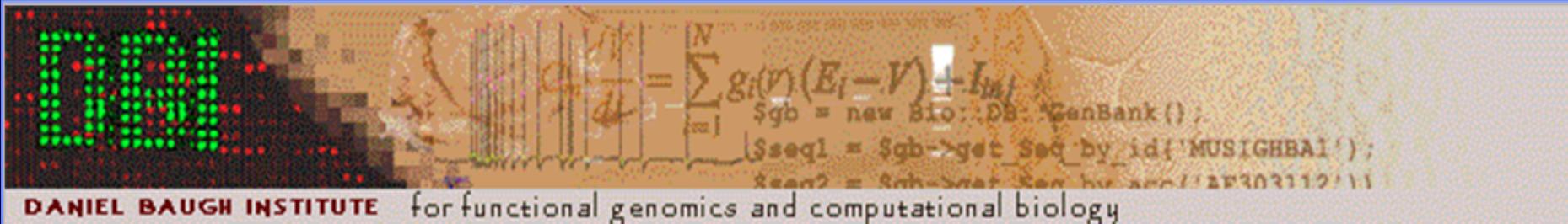


Gene interactions

GeneMania - networks, visualization, and more <http://genemania.org>



Promoter Analysis



DANIEL BAUGH INSTITUTE

for functional genomics and computational biology

PAINT Home

PAINT Modules:

Feasnet Builder

Feasnet Analysis and Visualization

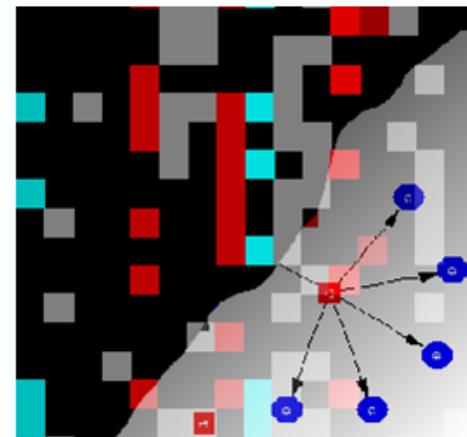
Documentation

Examples

DBI Home

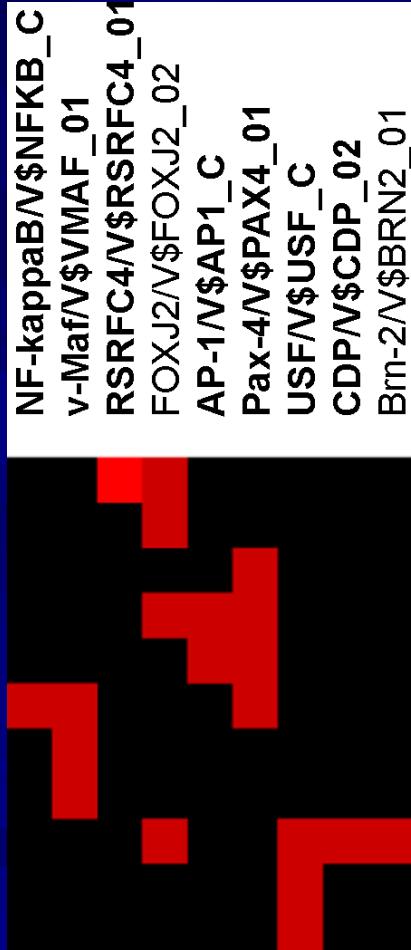
PAINT: Promoter Analysis and Interaction Network Toolset (V 4.0-pre)

Highly parallel gene expression analysis has led to analysis of gene regulation, in particular co-regulation, at a system level. PAINT was developed to provide the biologist a computational tool to integrate functional genomics data, for example from microarray-based gene expression analysis, with genomic sequence data to carry out transcriptional regulatory network analysis. TRNA. TRNA combines bioinformatics, used to identify and analyze gene regulatory regions, and statistical significance testing, used to rank the likelihood of the involvement of individual transcription factors, with visualization tools to identify transcription factors likely to play a role in the biology under study. In addition this tool can output results in several different formats such as Cytoscape format, IGB format, for use with modeling and simulation. As well as Primer3 output primers for hypothesized feasnet.



- PAINT - finds common promoters upstream of TSSs of the genes in a list. <http://www.dbi.tju.edu/dbi/tools/paint/>

P.A.I.N.T. Analysis



(NM_003854) IL1RL2
(BC001279) MFAP3L
(NM_025239) PDCD1LG2
(AY043465) FCRL2
(NM_033438) SLAMF9
(BC001279) MFAP3L
(AF361746) ESAM
(NM_014143) CD274
(NM_004233) CD83
(AL049370)
(NM_024036) LRFN4

Red –
statistically
significant TREs

TFACTS Analysis

The screenshot shows the TFACTS Analysis software interface. On the left, the 'Catalogue Selection' panel includes a dropdown menu set to 'Sign-Less' with an 'Add Data' button, and fields for 'Input Data' and 'Analysis Name (optional)'. Below these are sections for 'UP genes' (listing ADORA3, ADPRH, PARP1, PARP4) and 'DOWN genes' (empty). Under 'Negative control: number of random selections', the value '100' is selected. Thresholds are set to P-value: 0.05, E-value: 0.05, Q-value (Storey): 0.05, FDR (Benjamini-Hochberg): 0.05, RC (Random Control %): 5, and Target Genes(#): 1. On the right, the 'TFACTS Manual' page features a red header bar with a logo. The main content area is titled 'CONTENTS:' and lists numbered links to various manual sections.

Catalogue Selection:

Sign-Less

Input Data:

Analysis Name (optional) :

UP genes :

ADORA3
ADPRH
PARP1
PARP4

DOWN genes :

Negative control: number of random selections:
 50 100 500 1000

Thresholds:

P-value: 0.05

E-value: 0.05

Q-value (Storey): 0.05

FDR (Benjamini-Hochberg): 0.05

RC (Random Control %): 5

Target Genes(#): 1

TFACTS Manual

CONTENTS:

1. [DESCRIPTION](#)
2. [AUTHORS](#)
3. [INPUT](#)
 1. [Supported organisms](#)
 2. [Gene specification](#)
4. [ANNOTATED SIGNATURES](#)
 1. [Data sources](#)
 2. [Target gene signatures](#)
5. [METHOD](#)
 1. [Comparison Statistics](#)
 2. [Statistical significance](#)
 3. [Example](#)
6. [PUBLICATIONS](#)
7. [REFERENCES](#)

- TFACTS – finds common TFs regulating target genes. Manually curated. <http://www.tfacts.org/>

TF enrichment results

Regulated Transcription Factors:

Transcription Factor	P.value	E.value	Q.value	FDR control (B-H)	Intersection	Target genes	Random Control (%)
SRF	2.00e-5	1.16e-3	2.05e-6	8.62e-4	3	9	0
SP1	3.00e-5	1.74e-3	2.05e-6	1.72e-3	11	417	13
CREB1	7.00e-5	4.06e-3	3.19e-6	2.59e-3	7	169	7
TBP	3.10e-4	1.80e-2	1.04e-5	3.45e-3	4	53	2
TEAD1	3.80e-4	2.20e-2	1.04e-5	4.31e-3	2	5	0

Time course analysis

Time course data analysis

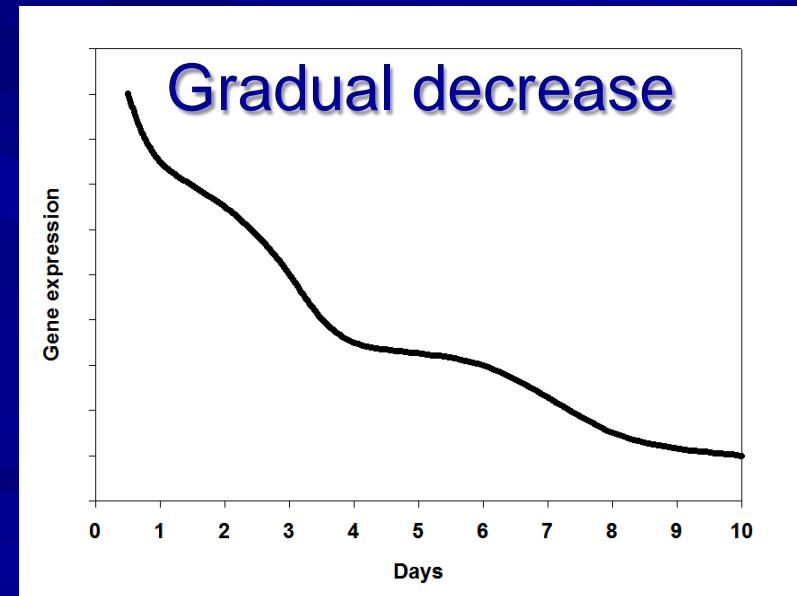
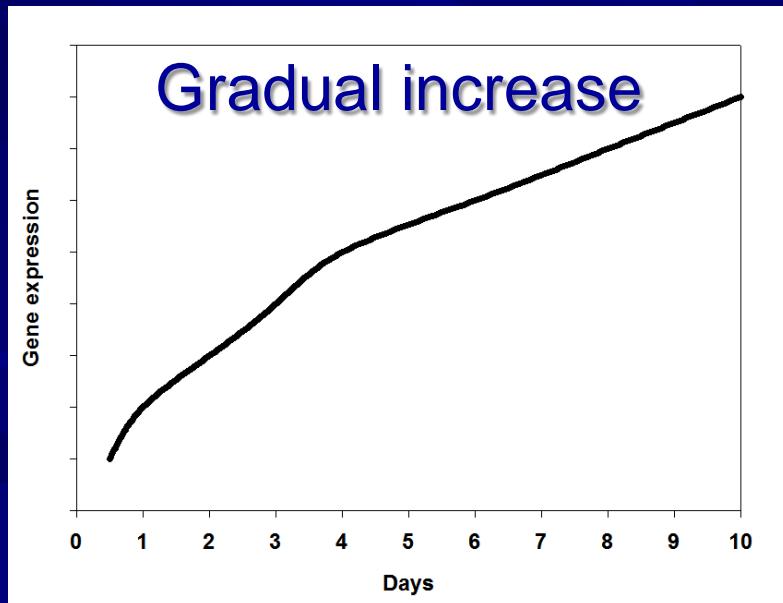
- Gene expression in dynamics.



- Analysis of profiles, not differences.
- Profiles may be more complex than just increase or decrease in expression level.

Gradual changes

- We expect to observe increase or decline in gene expression with maturation



SAM Analysis of Gene Expression Changes

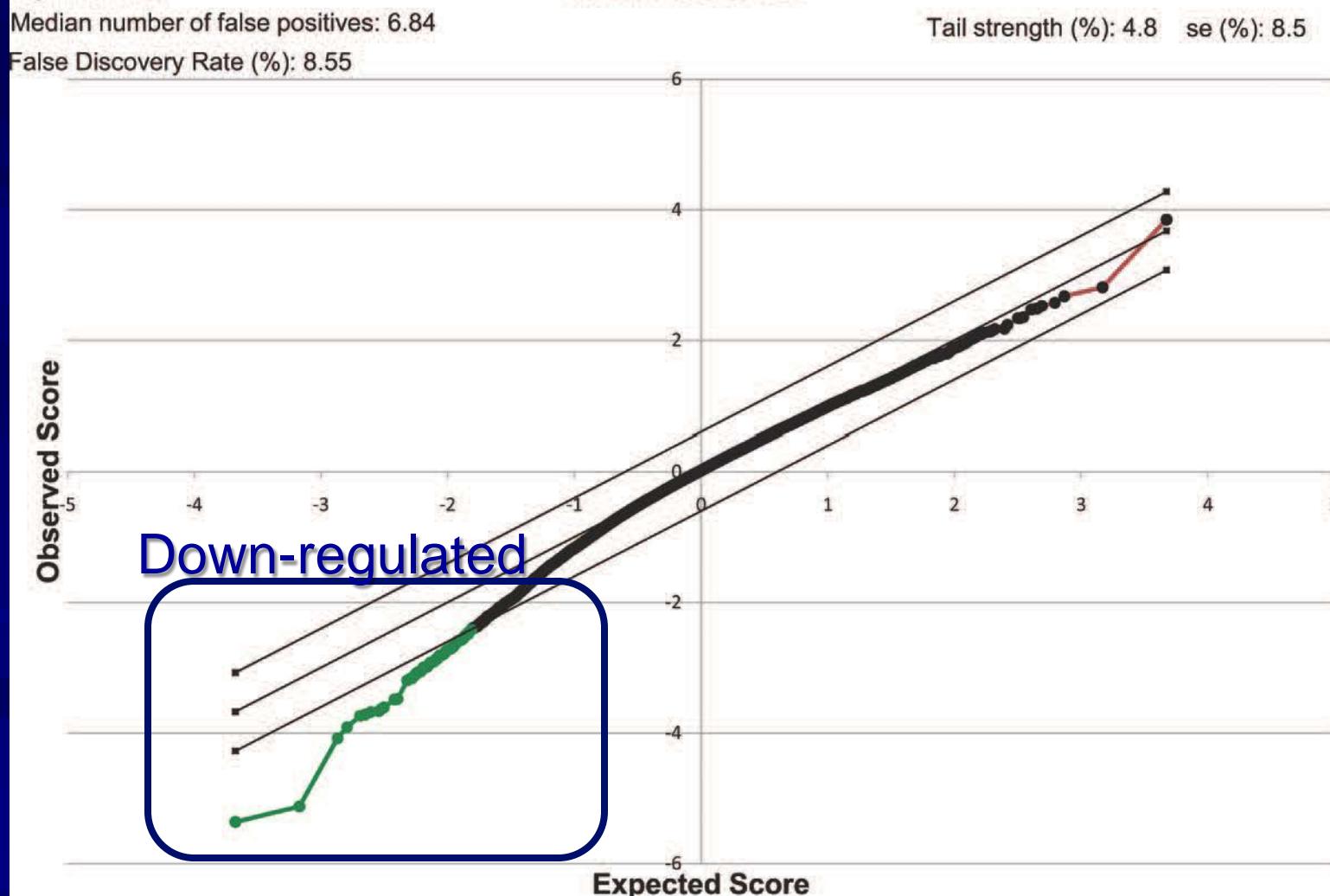
Significant: 69

Median number of false positives: 6.84

False Discovery Rate (%): 8.55

SAM Plotsheet

Tail strength (%): 4.8 se (%): 8.5

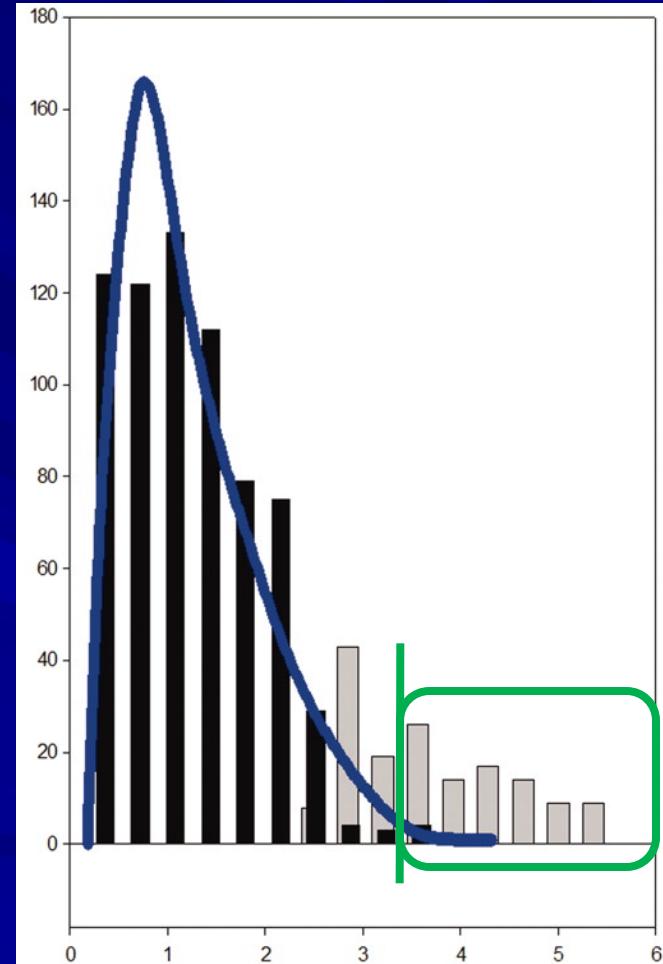


Hypervariable genes

- Hypervariable genes – whose variability is significantly more than background variability.
- Thousands of genes will exhibit technical variability, and hypervariable will contain biological meaning.
- Variability tested with F-statistics.

Identifying Hyper Variable (HV) Genes

- Histogram of relative SD shows right-skewed trend.
- Blue line - F-distribution fit to model technical variability V_t
- Green line shows HV gene cutoff.



What about analysis of exon arrays?

Affymetrix gene and exon arrays

- Affymetrix GeneChip Exon 1.0 ST

- Wide coverage

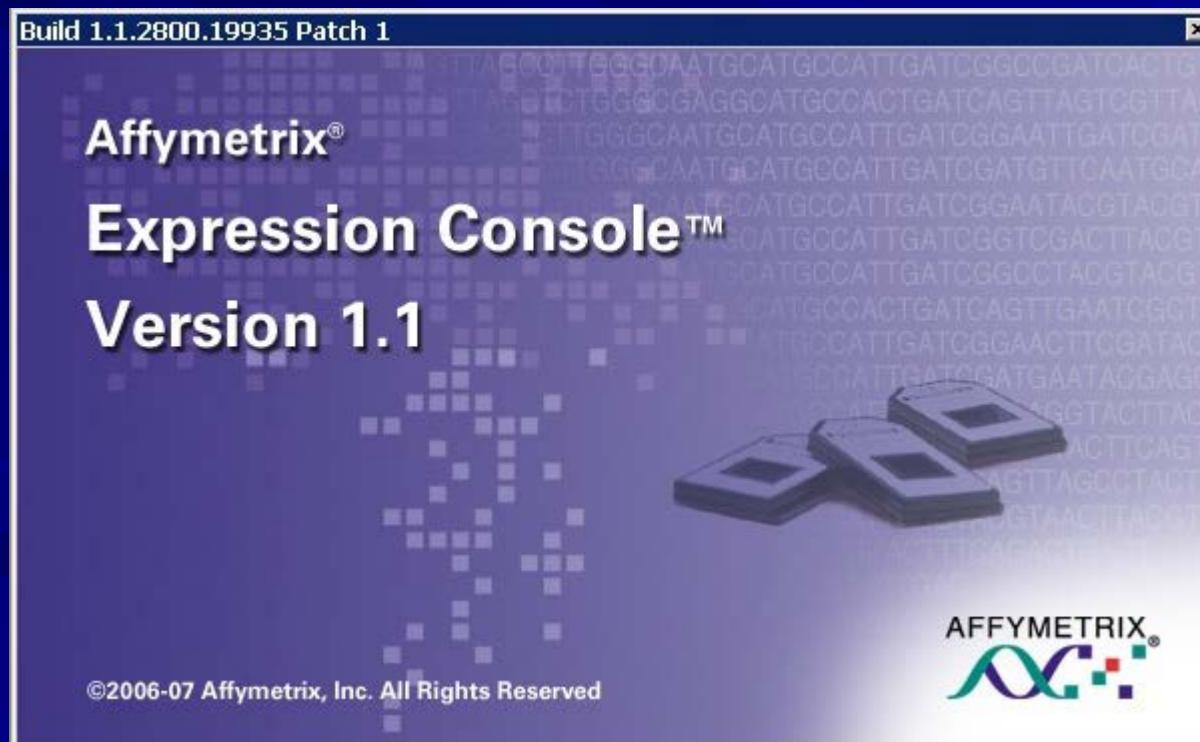
- Well annotated genes plus gene prediction sets

- Over 1.4 million probe sets

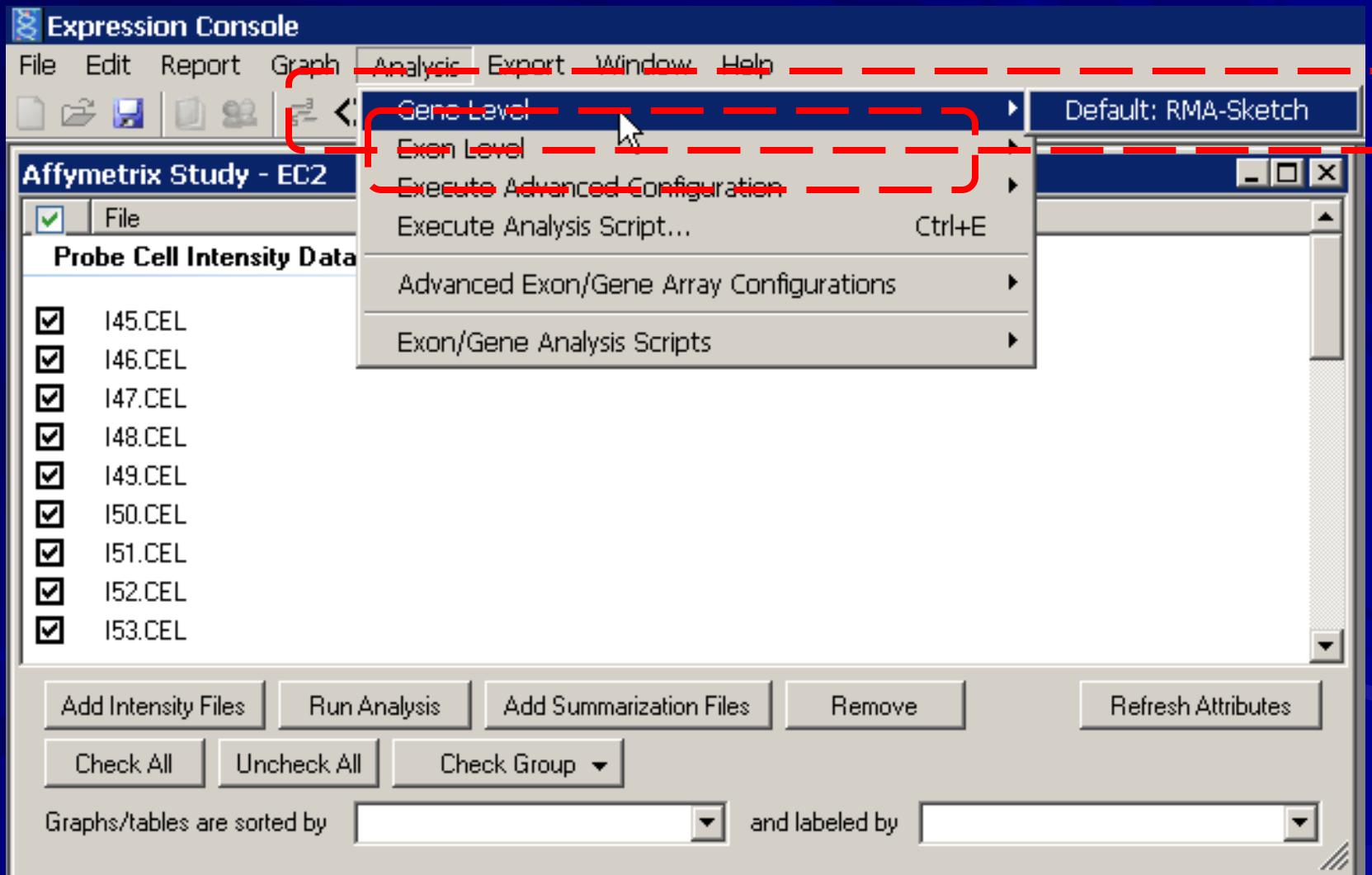
Affymetrix Expression Console (EC)

Affymetrix Expression Console Software

Affymetrix® Expression Console™ software supports Probe Set summarization and CHP file generation for both 3' Expression (GeneChip® Human Genome U133 Plus 2.0 Array) and Exon Arrays (GeneChip Human Exon 1.0 ST Array).



Gene/Exon level summarization



Gene/Exon level summarization

	A	B	C	D	E
1		Probe Set	H2132.JBV	H2133.JBV	H2134.JBV
2	1	7896736	7.196188	7.683278	7.728511
3	2	7896738	4.384084	4.567221	4.524869
4	3	7896740	4.729739	4.329544	4.110299
5	4	7896742	9.275142	9.514589	9.103162
6	5	7896744	9.930458	9.603415	9.982117
7	6	7896746	9.735742	9.511417	9.735742
8	7	7896748	9.988933	10.03924	10.1205
9	8	7896750	4.926641	4.68431	5.537951
33253	33252	8180413	10.23673	10.28367	10.43993
33254	33253	8180414	6.282938	6.263641	6.263641
33255	33254	8180415	9.1038	9.365929	9.203348
33256	33255	8180416	6.064609	6.064609	6.217036
33257	33256	8180417	9.841396	9.971924	9.867586
33258	33257	8180418	9.557718	9.589325	9.304793

Gene level summarization

	A	B	C	D	E
1	Probe Set	H2132.rma	H2133.rma	H2134.rma	H2135.rma
2	7896737	7.196188	7.683278	7.728511	7.396492
3	7896739	4.384084	4.567221	4.524869	4.213489
4	7896741	4.729739	4.329544	4.110299	4.344719
5	7896743	9.275142	9.514589	9.103162	9.329247
6	7896745	9.930458	9.603415	9.982117	9.675047
7	7896747	9.735742	9.511417	9.735742	10.09383
257424	8180019	7.097994	6.800755	7.049932	7.164893
257425	8180021	4.681778	4.247624	4.510065	5.093358
257426	8180023	2.581063	2.52183	2.348945	3.003194
257427	8180024	9.807602	9.71832	9.506336	9.64273
257428	8180025	6.679294	7.457044	6.010935	6.893983
257429	8180026	8.294184	8.149136	8.020469	8.334627
257430	8180027	5.965128	6.603519	6.008313	6.771286
257431	8180028	6.868145	6.344553	6.423243	6.866771

Exon level summarization

Exon array analysis

Aroma.affymetrix R package <http://aroma-project.org/>

aroma-project.org

Current versions:

aroma.affymetrix 2.8.0

aroma.cn 1.0.0

3rd party extensions

What's new?

Dec 22, 2012: aroma.affymetrix v2.8.0 released.

Nov 5, 2012: PSCBS v0.30.0 released.

Jan 14, 2012: OSX 10.6 & 10.7 users: If R crashes, install [the affxparser OSX binary bug fix](#).

An open-source R framework for your microarray analysis

FIRMGene <http://bioinf.wehi.edu.au/folders/firmagene/>

BMC Bioinformatics

Methodology article

Differential splicing using whole-transcript microarrays
Mark D Robinson^{*1,2,3} and Terence P Speed³

Analysis

Exon Array Analyzer (now accepts Gene Arrays) <http://eaa.mpi-bn.mpg.de/index.php>

BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 24 2009, pages 3323–3324
doi:10.1093/bioinformatics/btp577

Gene expression

Exon Array Analyzer: a web interface for Affymetrix exon array analysis

Pascal Gellert, Shizuka Uchida and Thomas Braun*



Exon Array Analyzer

Welcome

The Exon Array Analyzer allows to process CEL files from Affymetrix, Inc. GeneChip® Exon 1.0 ST Arrays to identify alternative splicing.

Exon array analysis results

Exon Level		Gene Level	
Code:	60a5de17c7ffc5037f4798c426deed630013		
Probe set ID			
Transcript cluster ID			
Exon Accession			
Gene Annotation			
P-Value	<0.01		
Splice Index	<-1		
Splice Index	>1		
Sort by	Probe Set ID		
<input type="button" value="Reset"/>	<input type="button" value="View"/>	<input type="button" value="Download"/>	
Search Result (0) show/hide			
No results returned			

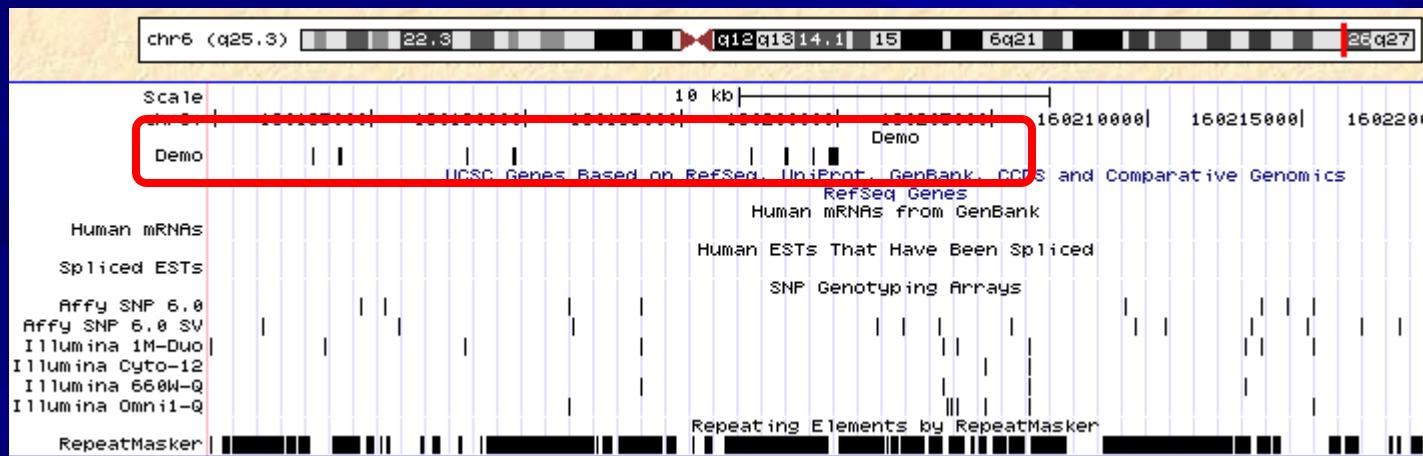
Exon Level		Gene Level	
Code:	60a5de17c7ffc5037f4798c426deed630013		
Transcript cluster ID			
Exon Accession			
Gene Annotation			
P-Value	<0.01		
Fold Change	<-1		
Fold Change	>1		
Sort by	TTest		
<input type="button" value="Reset"/>	<input type="button" value="View"/>	<input type="button" value="Download"/>	
Search Result (79) show/hide			
Transcript Cluster ID	Symbol	Entrez Gene ID	Description
7895163	-	-	-
7946071	OR51B5	282763	olfactory receptor, fam

Interpretation of exon results

UCSC genome browser <http://genome.ucsc.edu/>

browser position	chr6:160183128-1602000047
browser dense all	
track name="Demo"	
chr6	160183128 160183158
chr6	160183975 160184084
chr6	160188051 160188144
chr6	160189545 160189654
chr6	160197199 160197230
chr6	160198348 160198422
chr6	160199210 160199290
chr6	160199717 160200047

BED format



Interpretation of exon results

FAST DB - alternative splicing and their interpretation <http://www.fast-db.com>

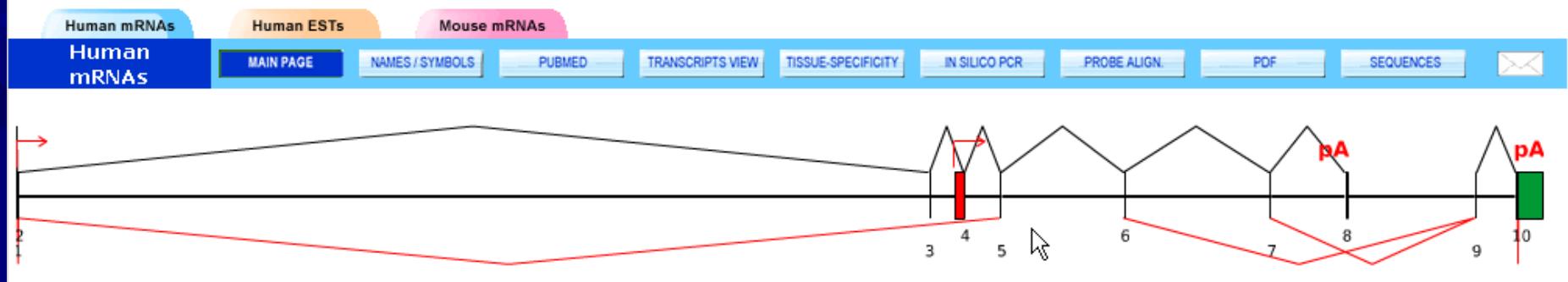
BMC Bioinformatics 

Database Open Access

A new advance in alternative splicing databases: from catalogue to detailed analysis of regulation of expression and function of human alternative splicing variants

Pierre de la Grange, Martin Dutertre, Margot Correa and Didier Auboeuf*

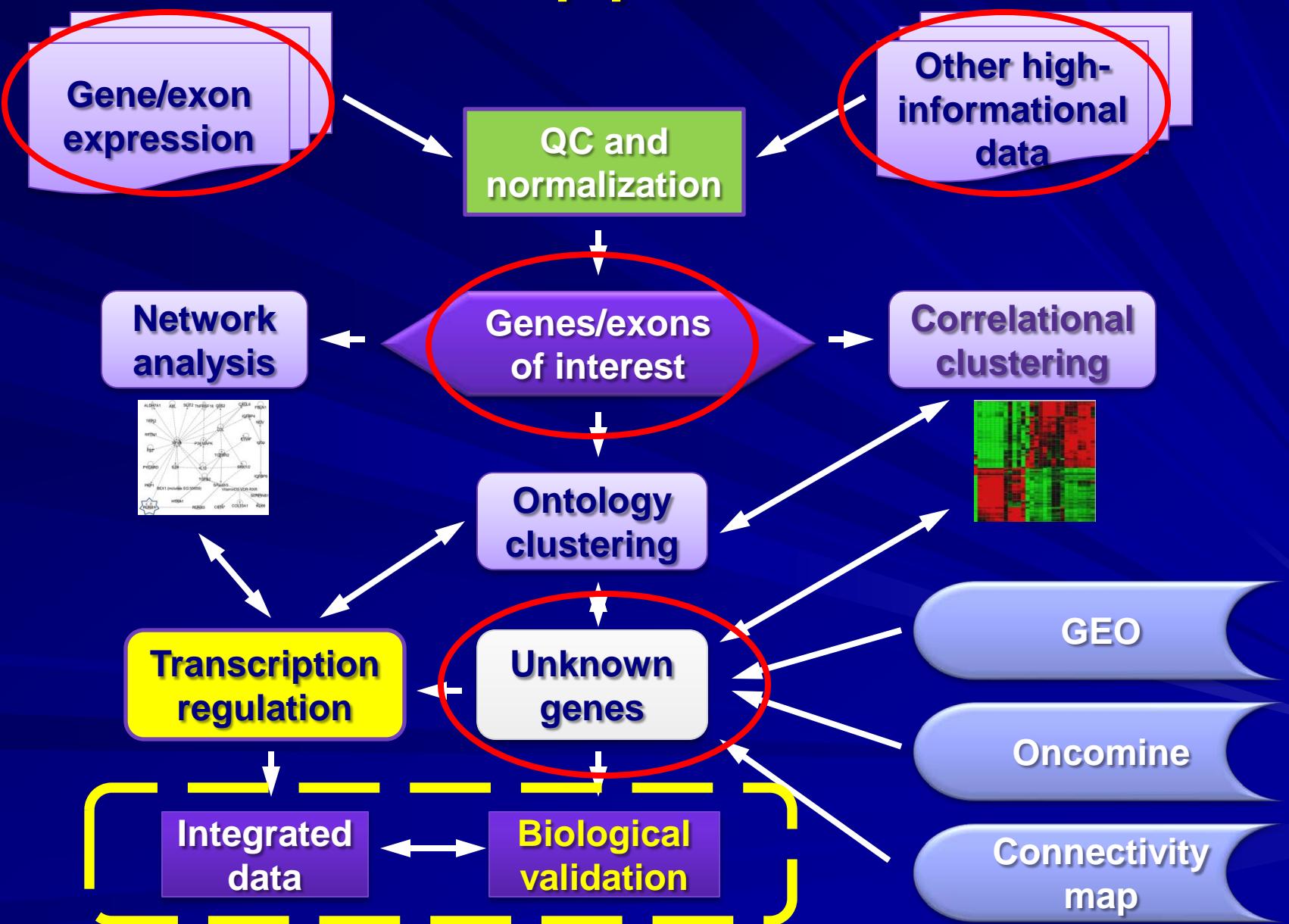
**runt-related transcription factor 1 (acute myeloid leukemia 1; aml1 oncogene)
(ID 2368)**



Summary

- Formulate test and alternative hypotheses before doing microarray experiments
- Normalize and pre-process data
- Select appropriate statistical method
- Use different tools to get answers from all aspects of the data

The Approach



Thank you!

Microarray Data processing

Mikhail Dozmorov, Ph.D.,
Arthritis & Clinical Immunology Department
Oklahoma Medical Research Foundation
Mikhail-Dozmorov@omrf.org
January 17, 2013