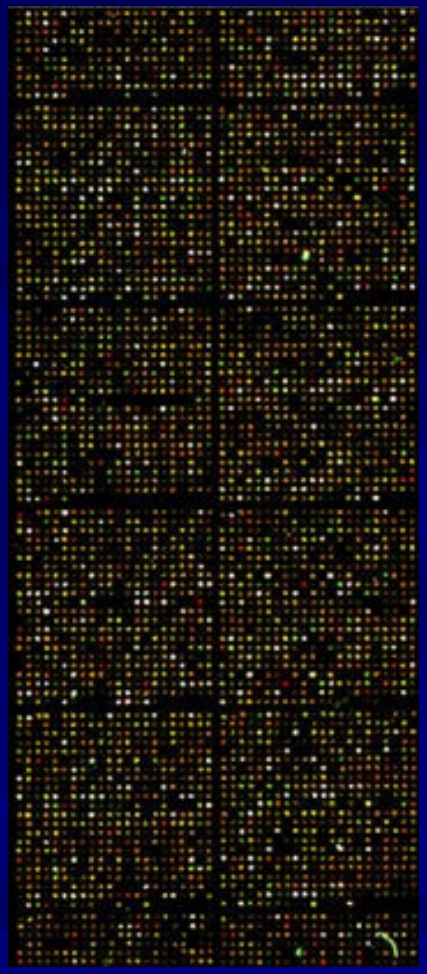
Microarray Data processing

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

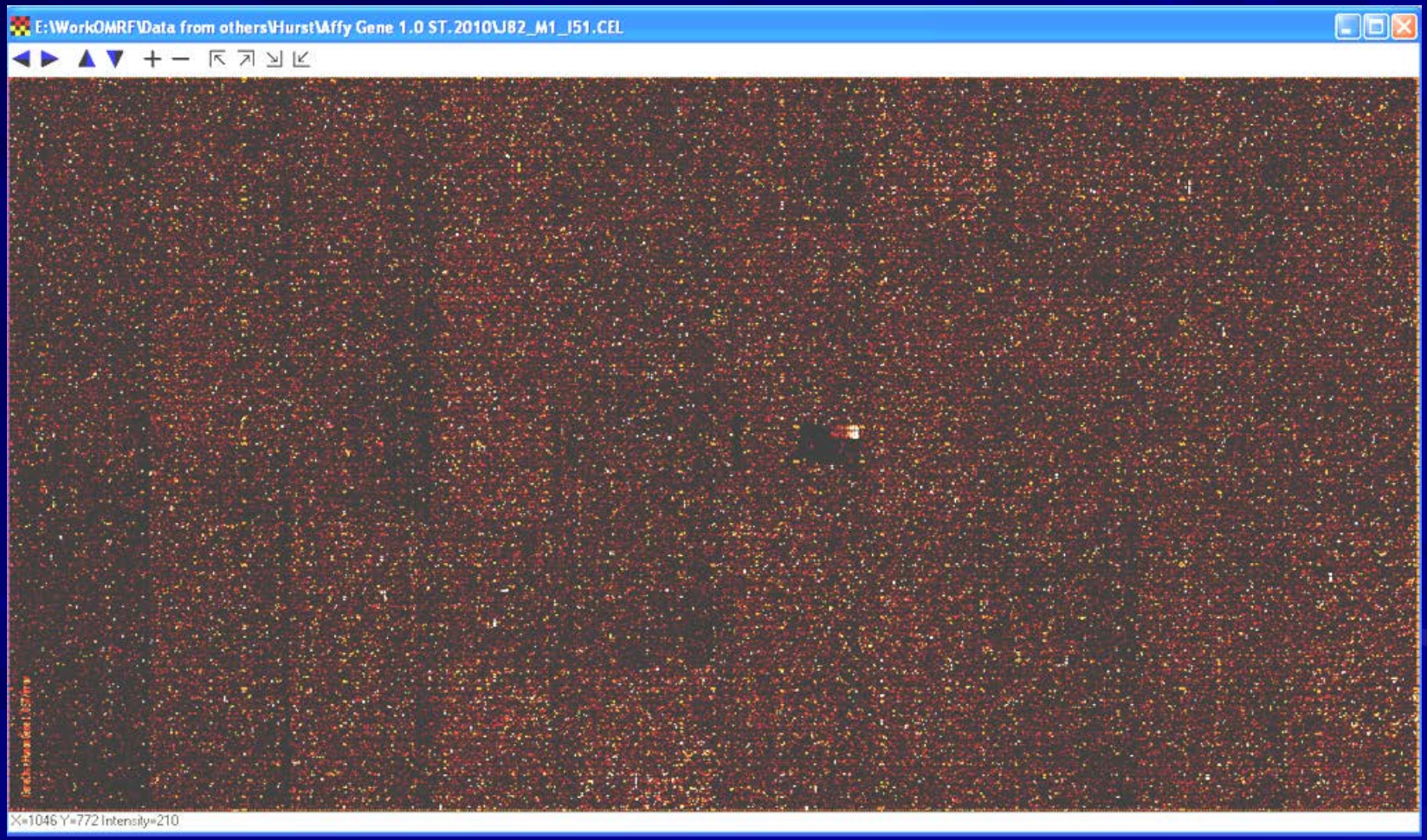
Microarrays



- Microarrays are a standardized technology for measuring the expression level of thousands of genes

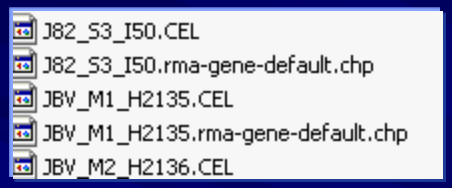
- This present new statistical problems because the data are very high dimensional with very little replication (50,000 genes/10 replicates)

Microarray data format



Microarray data format

- Raw data files



- Numerical data

10	PROBE_ID	SYMBOL	504006100	504006100	504006100	504006100	504006100	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.[illegible]835	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

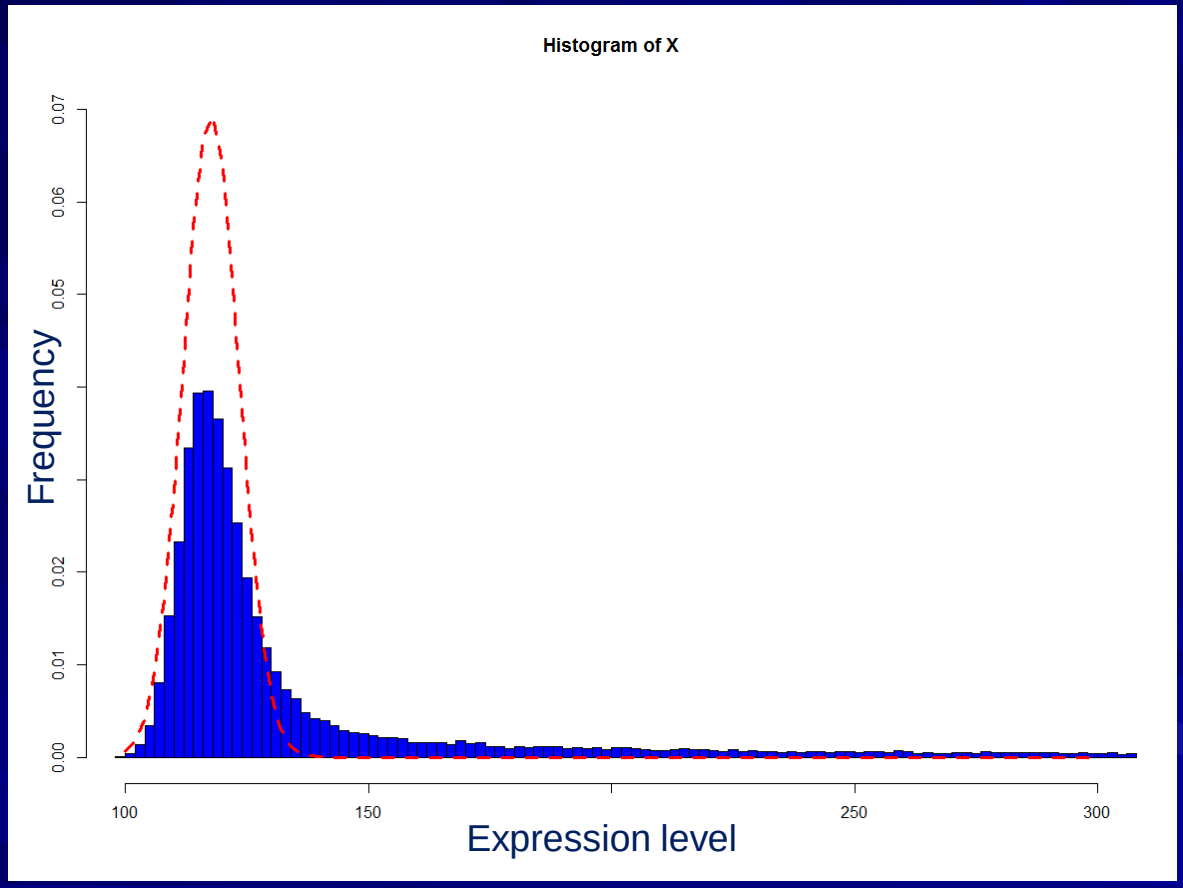
Steps in microarray data analysis

Data preprocessing

- The probe-level data must be cleaned and processed to obtain biologically meaningful measurement:

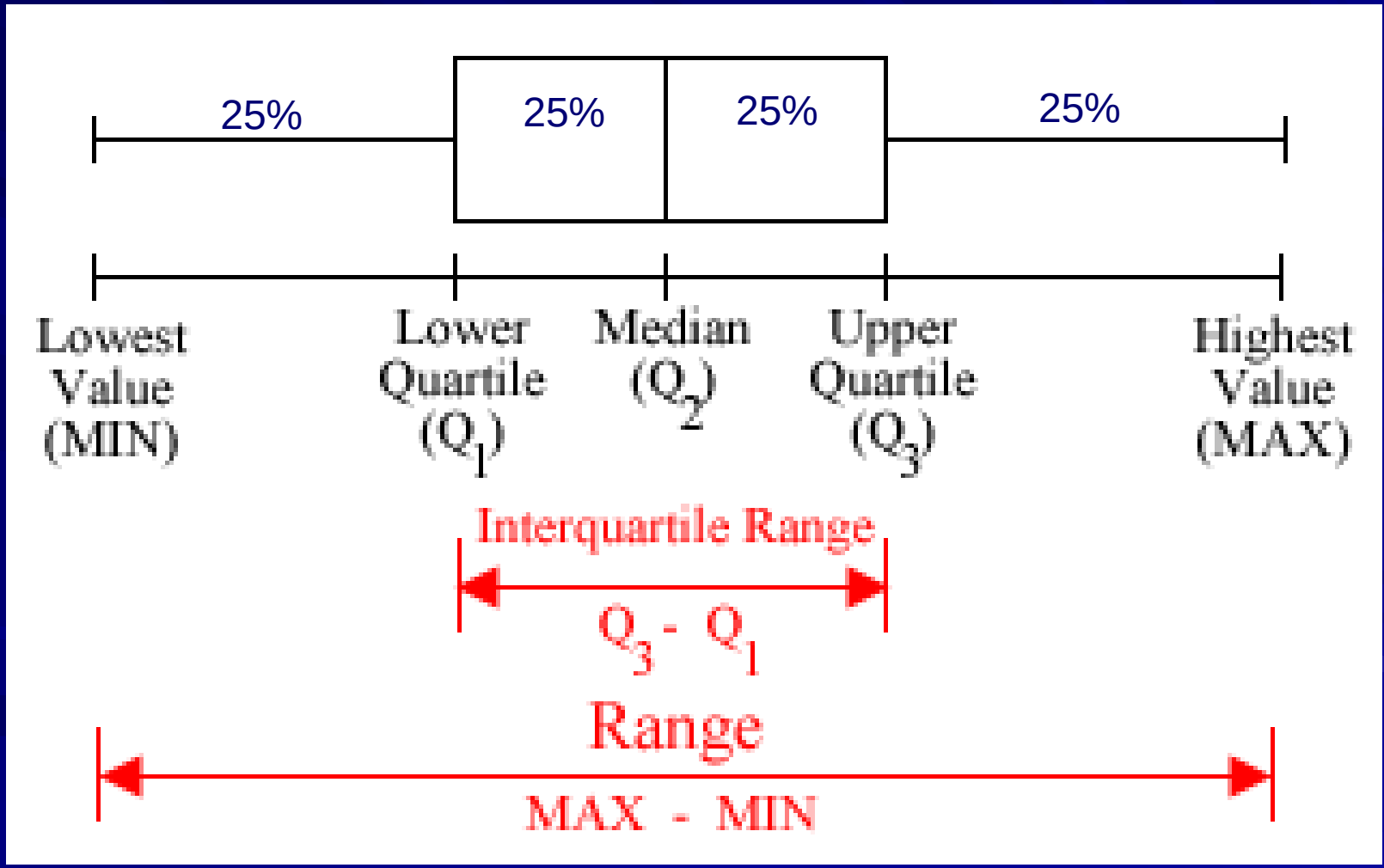
1. Background correction: eliminate signals due to non-specific binding;
2. Normalization: make multiple arrays comparable;
3. Preprocessing: Log2 transformation, missing values imputation, etc.

Background correction

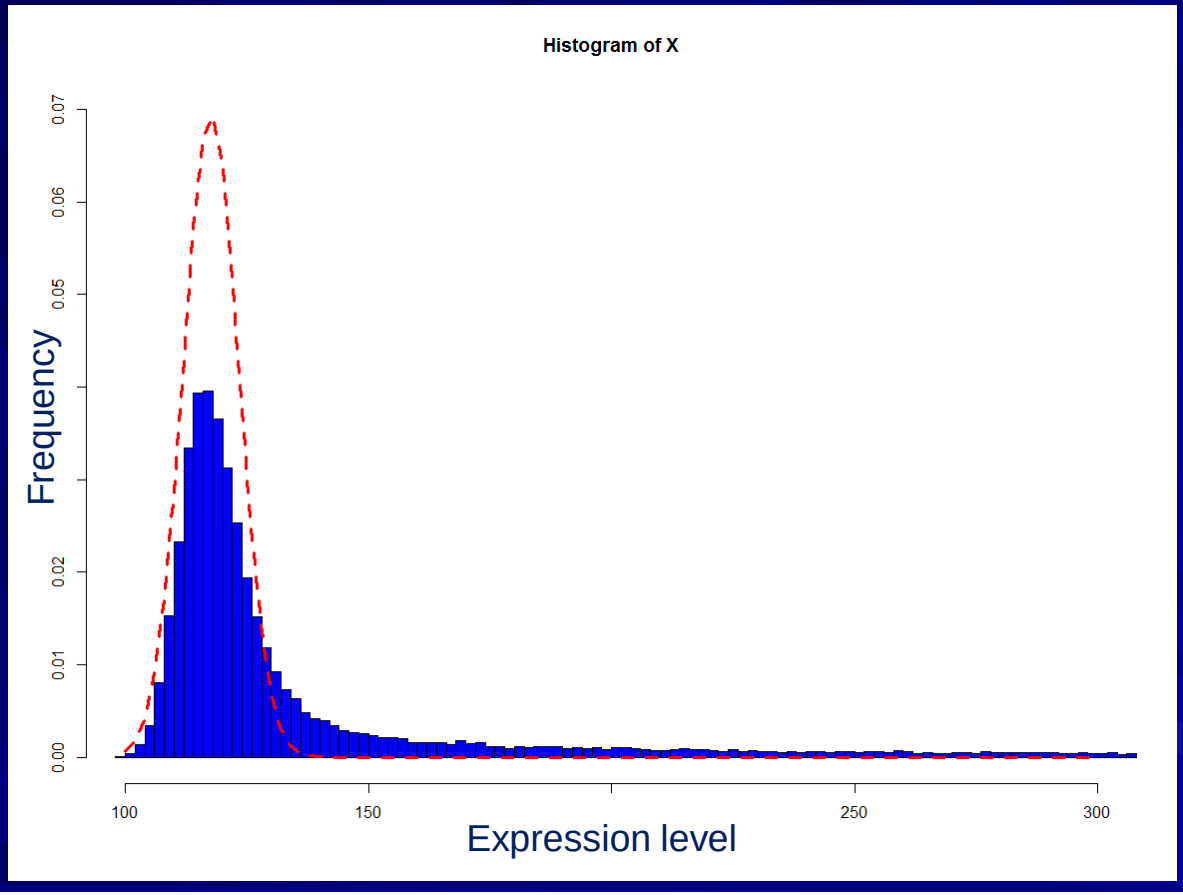


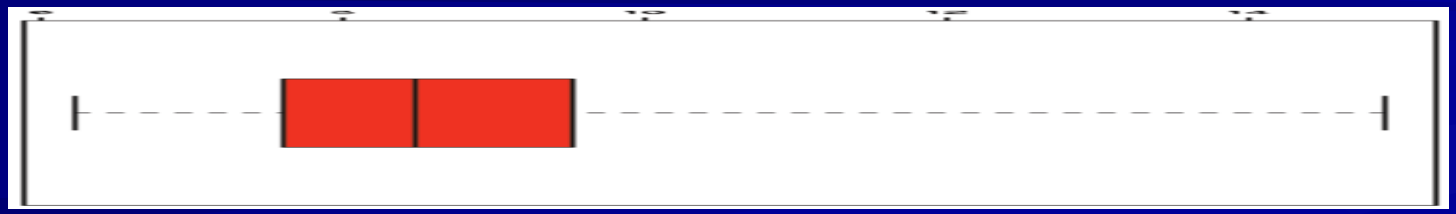
- Noise level – uncertainty about zero point

Box and whisker plot

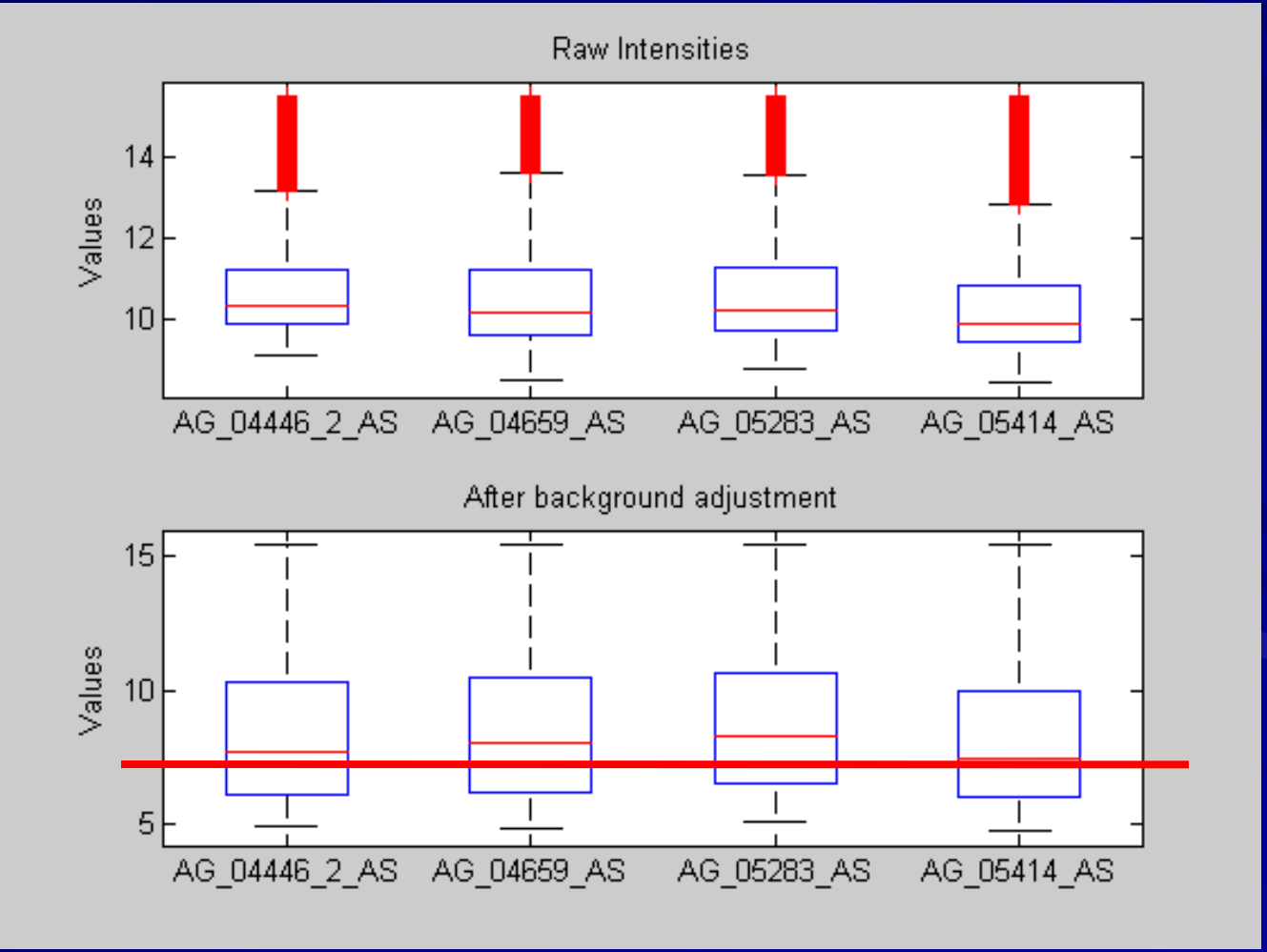


Background correction





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.

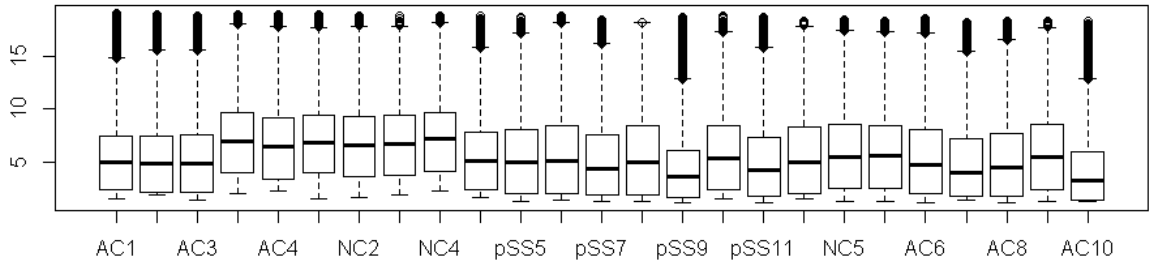
Normalization methods

- 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

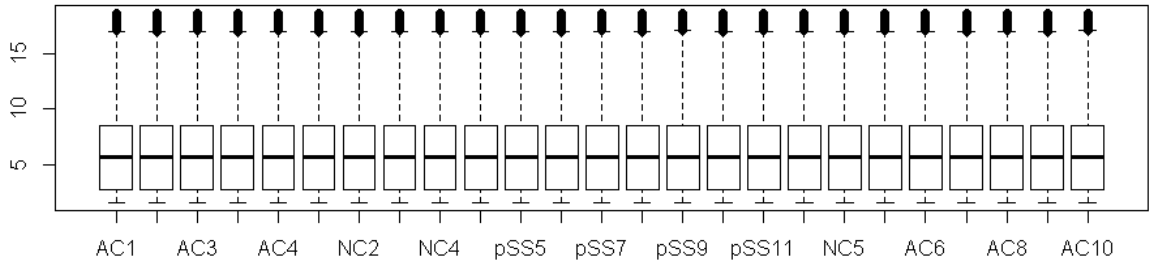
Quantile normalization

- Make distribution of data equal across all samples.

Before normalization

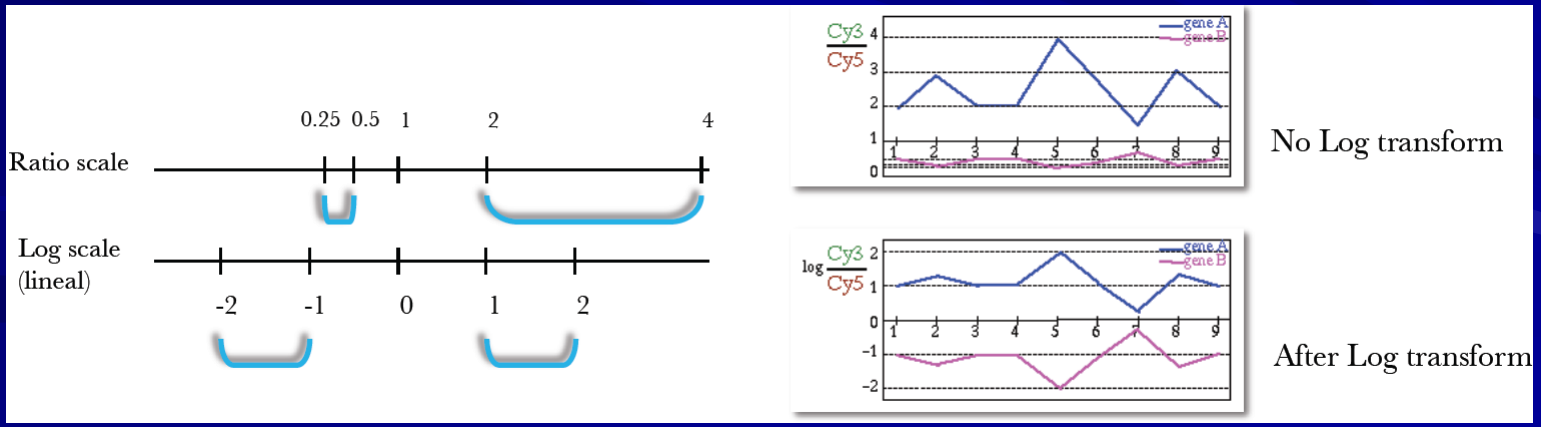


After normalization



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

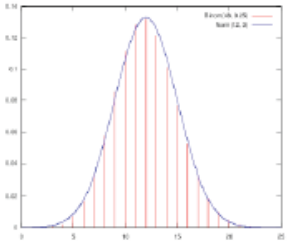
Nine steps for hypothesis testing

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 Ho 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(μ=12,σ=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 cannont 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

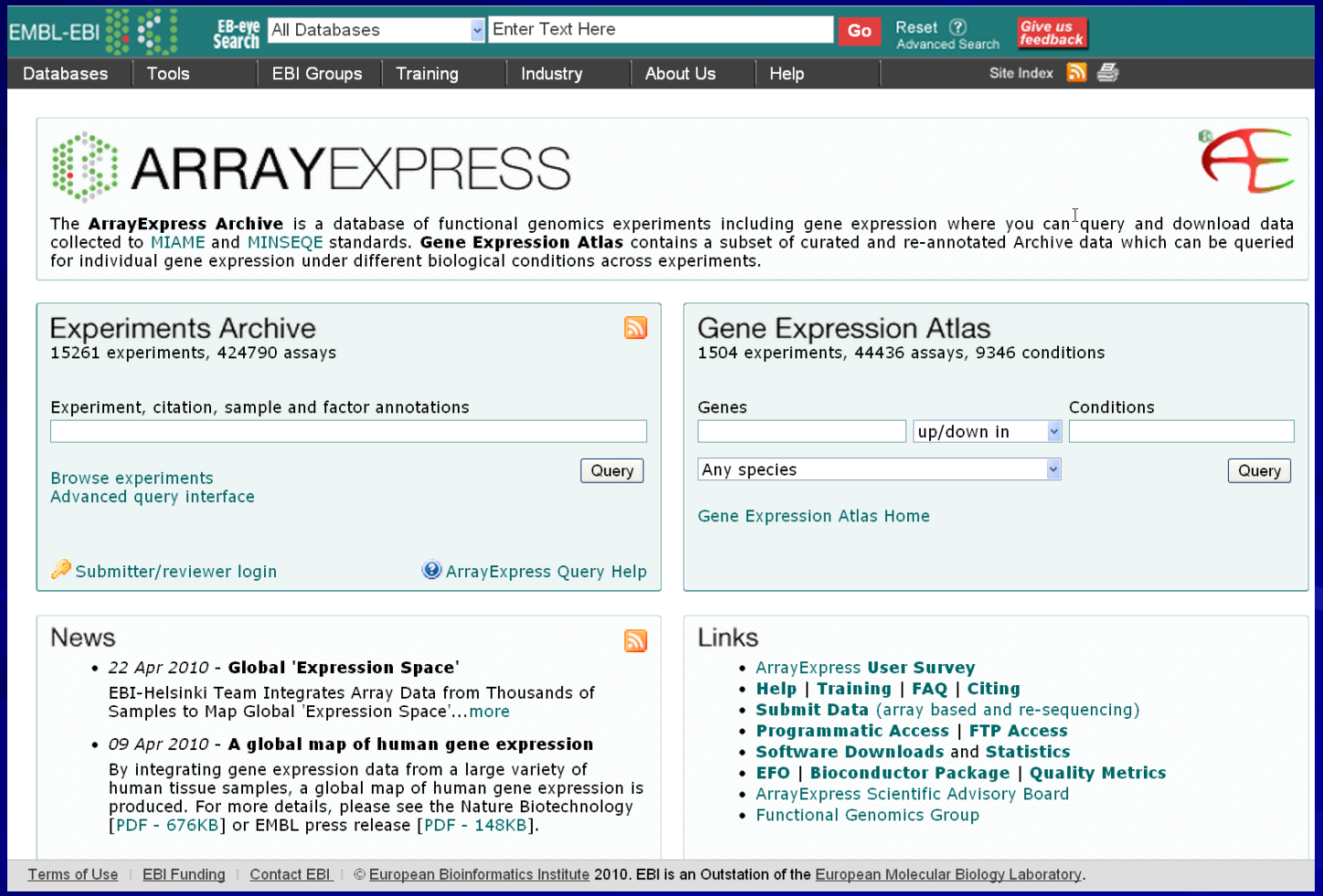
Adjustment for multiple testing

- 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

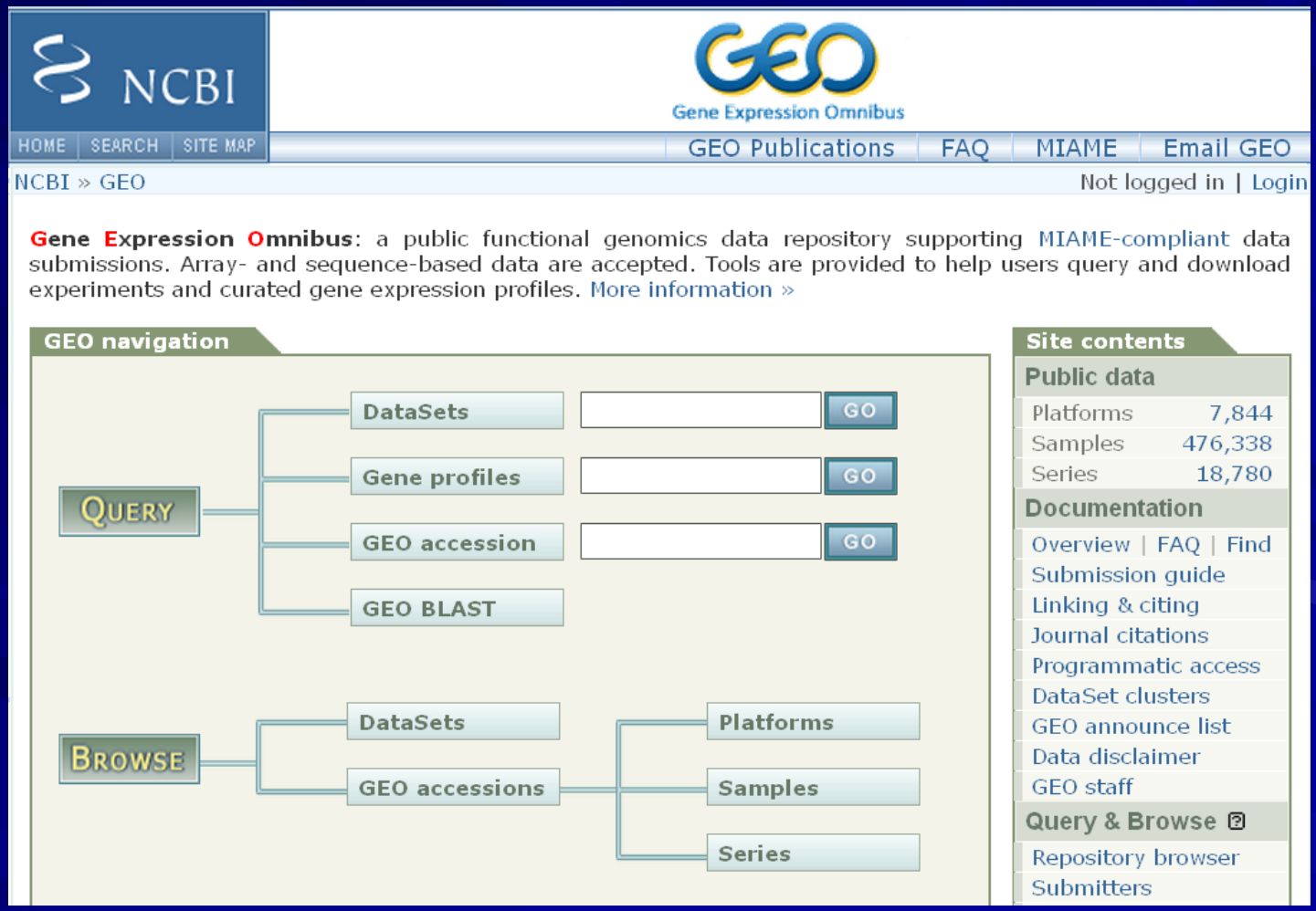
ArrayExpress

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

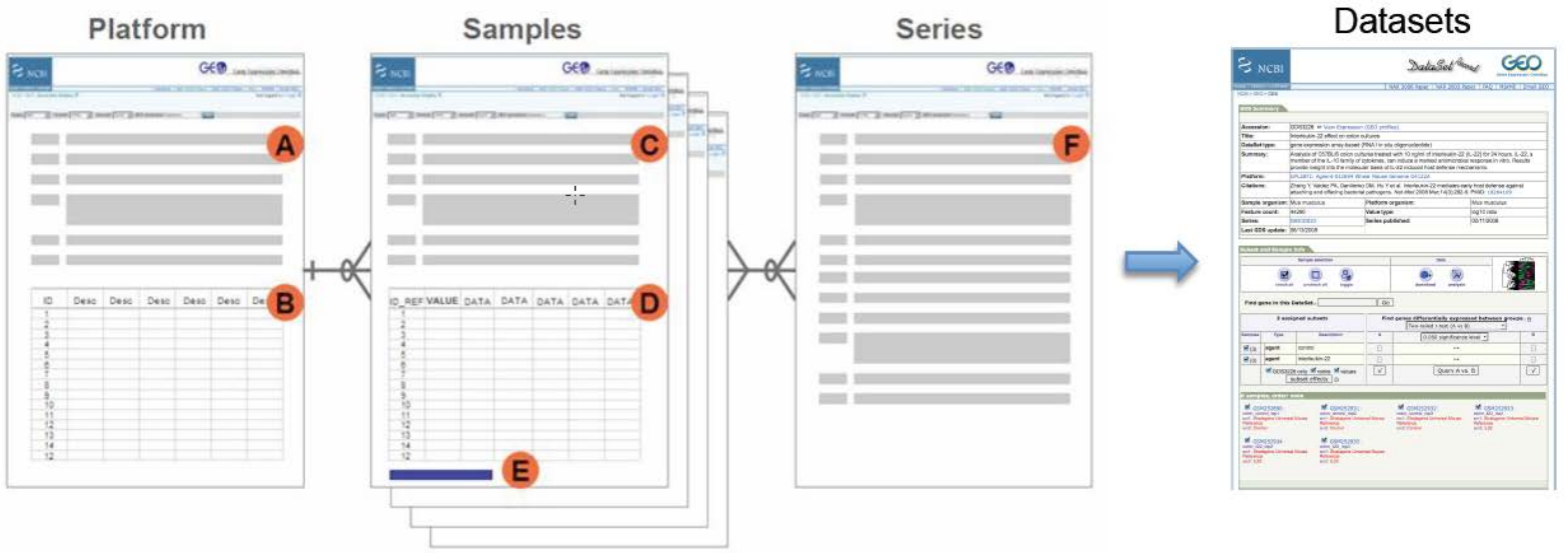


Gene Expression Omnibus

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



GEO Database Organization



Platform	Platform records are supplied by submitters. A Platform record defines the list of elements that may be detected and quantified in that experiment (e.g., cDNAs, oligonucleotide probesets). Each Platform record is assigned a unique and stable GEO accession number (GPLxxx). A Platform may reference many Samples that have been submitted by multiple submitters. Example Platform record	A	**Text description of the array**
		B	**Text tab-delimited table of the array template**
Sample	Sample records are supplied by submitters. A Sample record describes the conditions under which an individual Sample was handled, the manipulations it underwent, and the abundance measurement of each element derived from it. Each Sample record is assigned a unique and stable GEO accession number (GSMxxx). A Sample entity must reference only one Platform and may be included in multiple Series. Example Sample record	C	**Text description of a biological sample**
		D	**Text tab-delimited table of processed hybridization result** (may optionally include raw data columns)
		E	**Original raw data file**
Series	Series records are supplied by submitters. A Series record links together a group of related Samples and provides a focal point and description of the whole study. Series records may also contain tables describing extracted data, summary conclusions, or analyses. Each Series record is assigned a unique and stable GEO accession number (GSExxx). Example Series record	F	**Text description of the overall experiment**

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

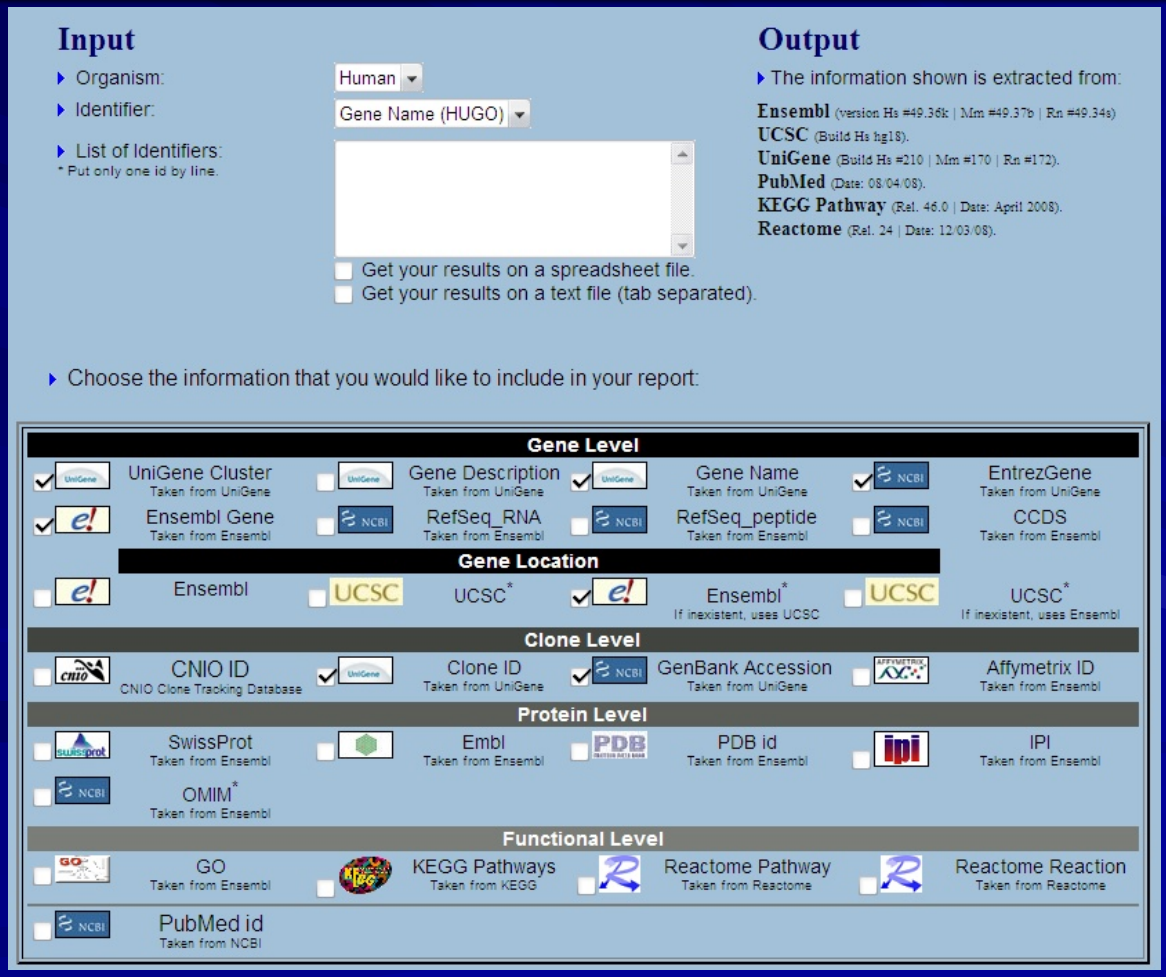
- Matlab - The Language of Technical Computing http://www.mathworks.com/

- Online tools

Gene IDs converters

Clone/Gene ID converter

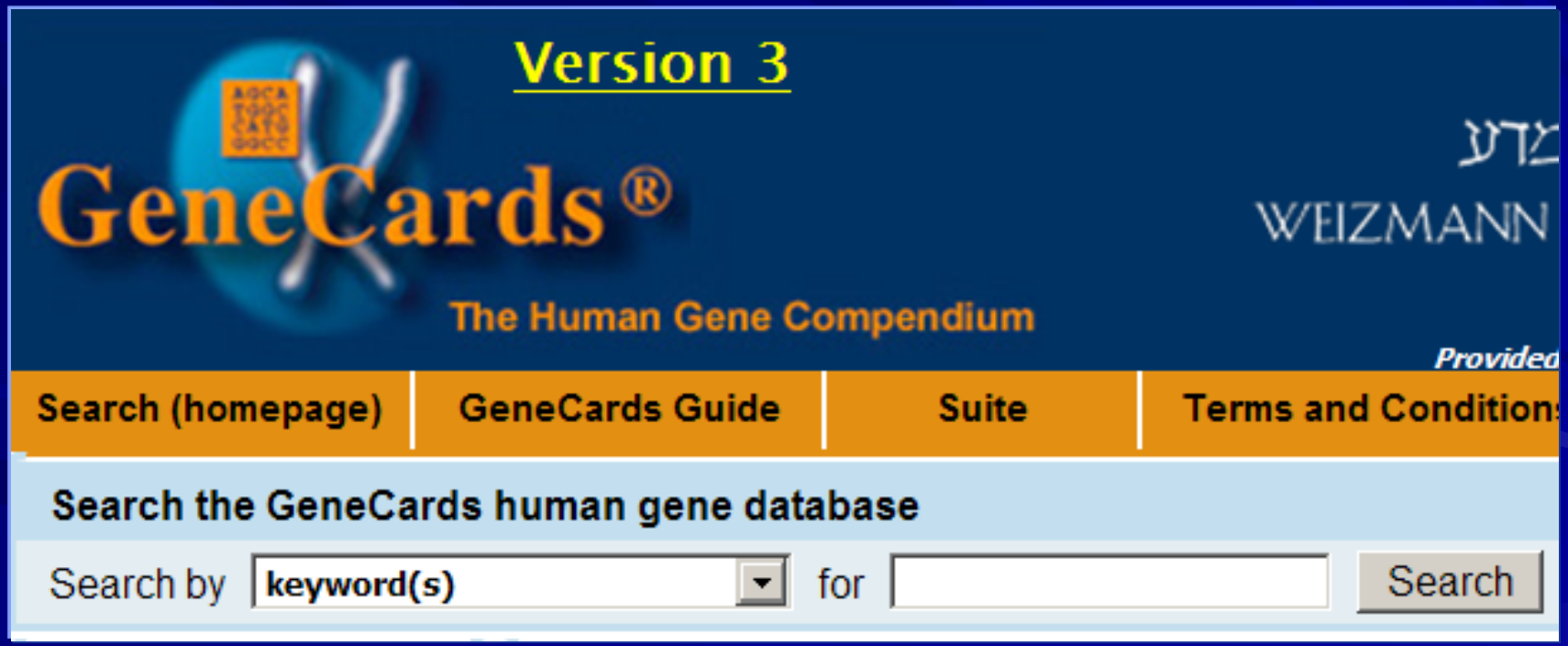
http://idconverter.bioinfo.cnio.es/IDconverter.php



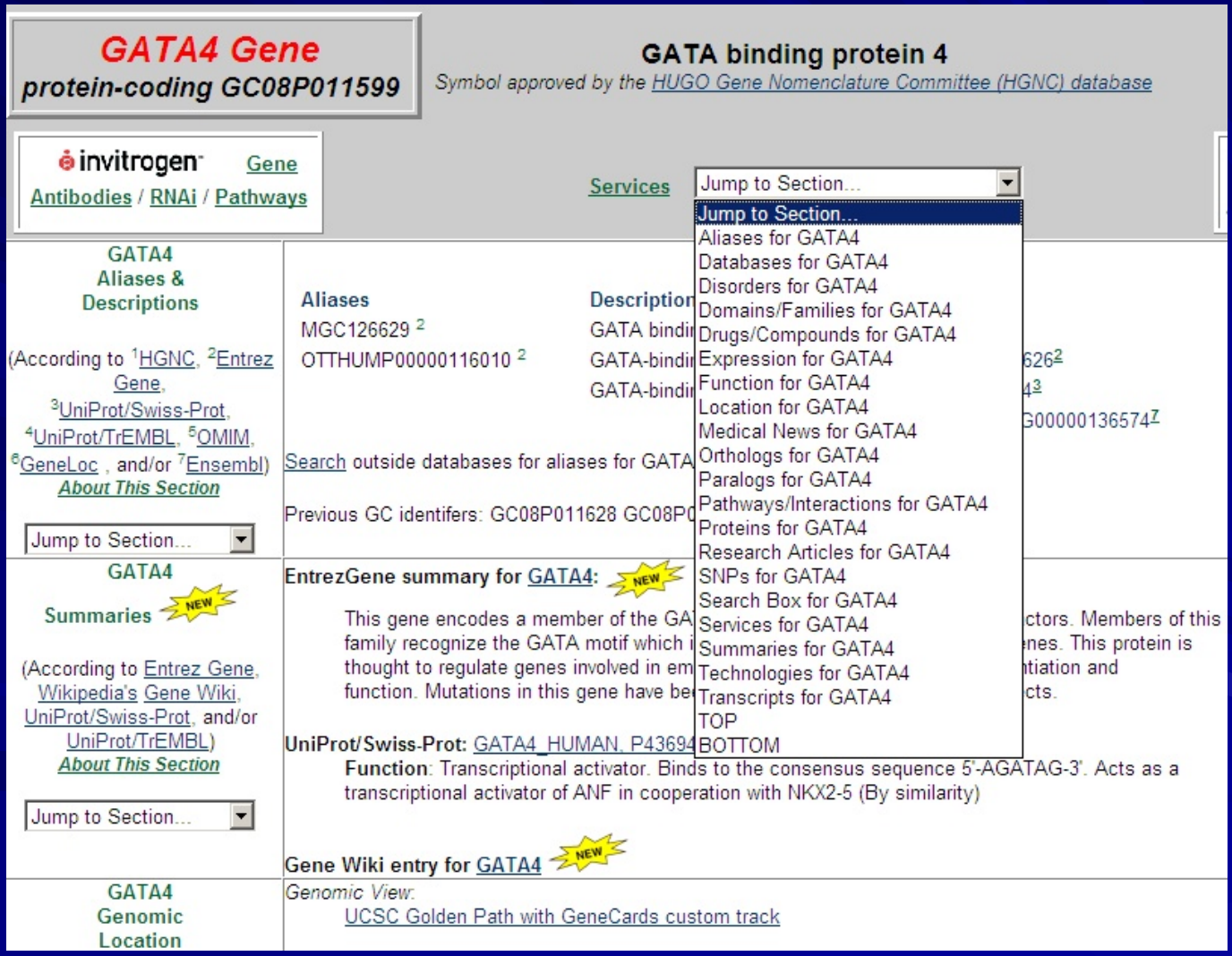
Information about genes

GeneCards - comprehensive database about genes

http://www.genecards.org



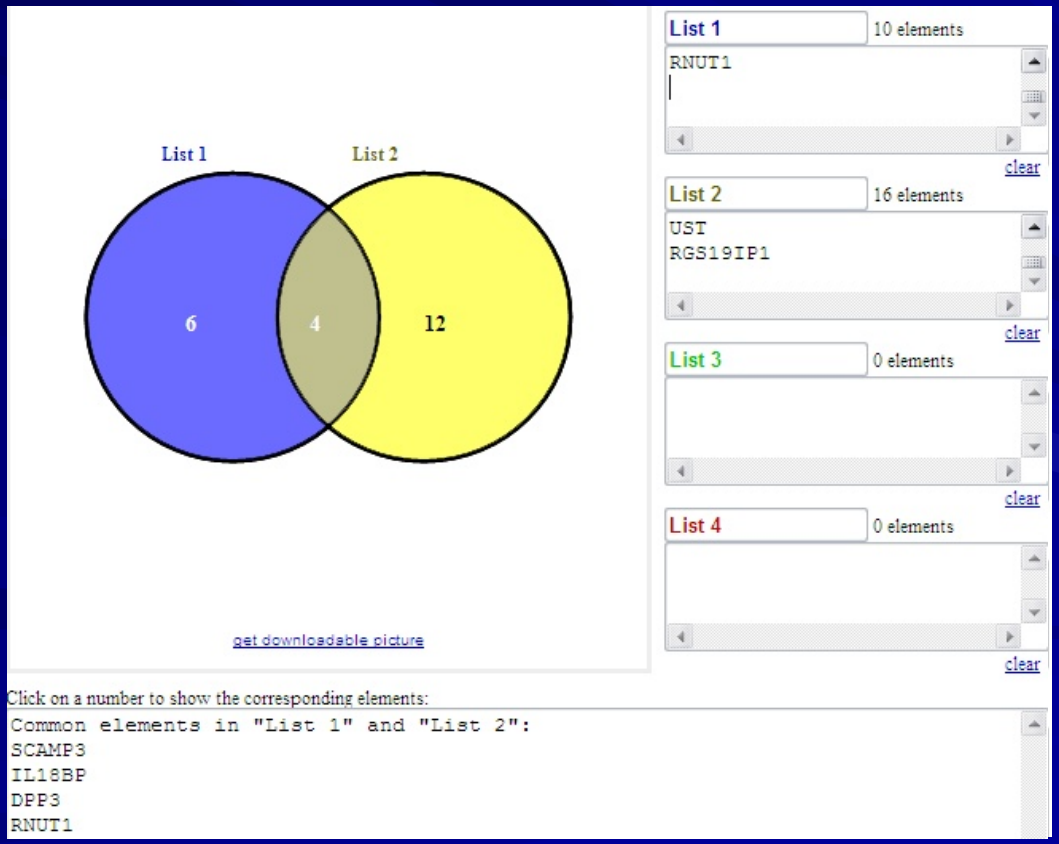
Information about genes



Comparison of gene lists

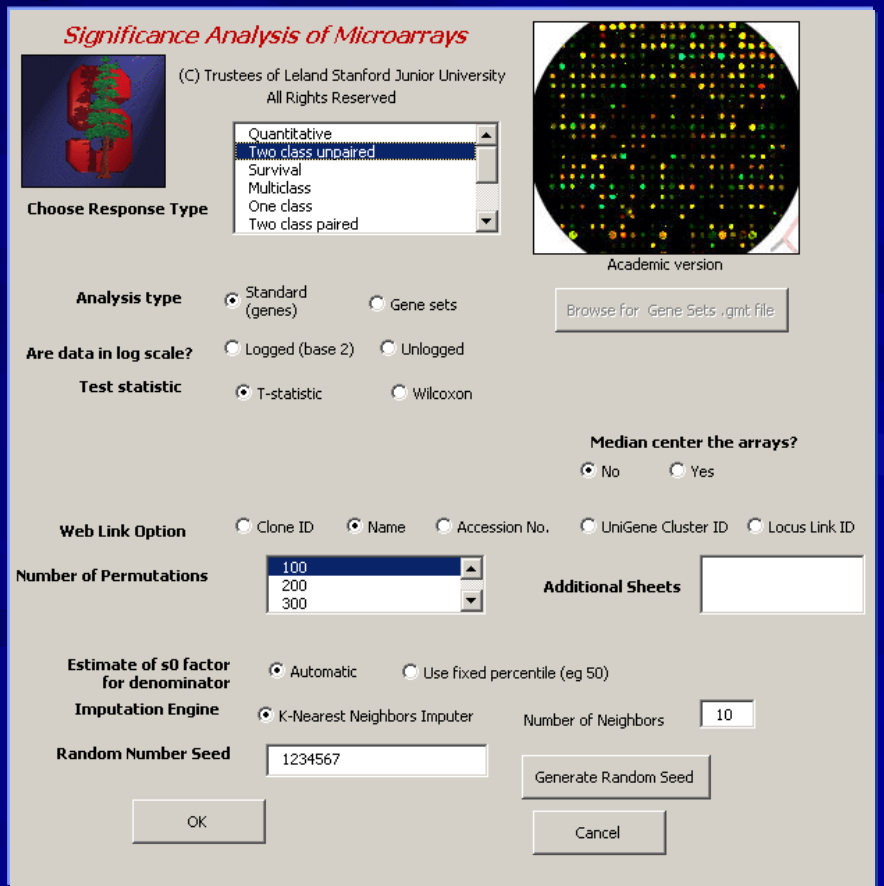
VENNY - An interactive tool for comparing lists with Venn Diagrams

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



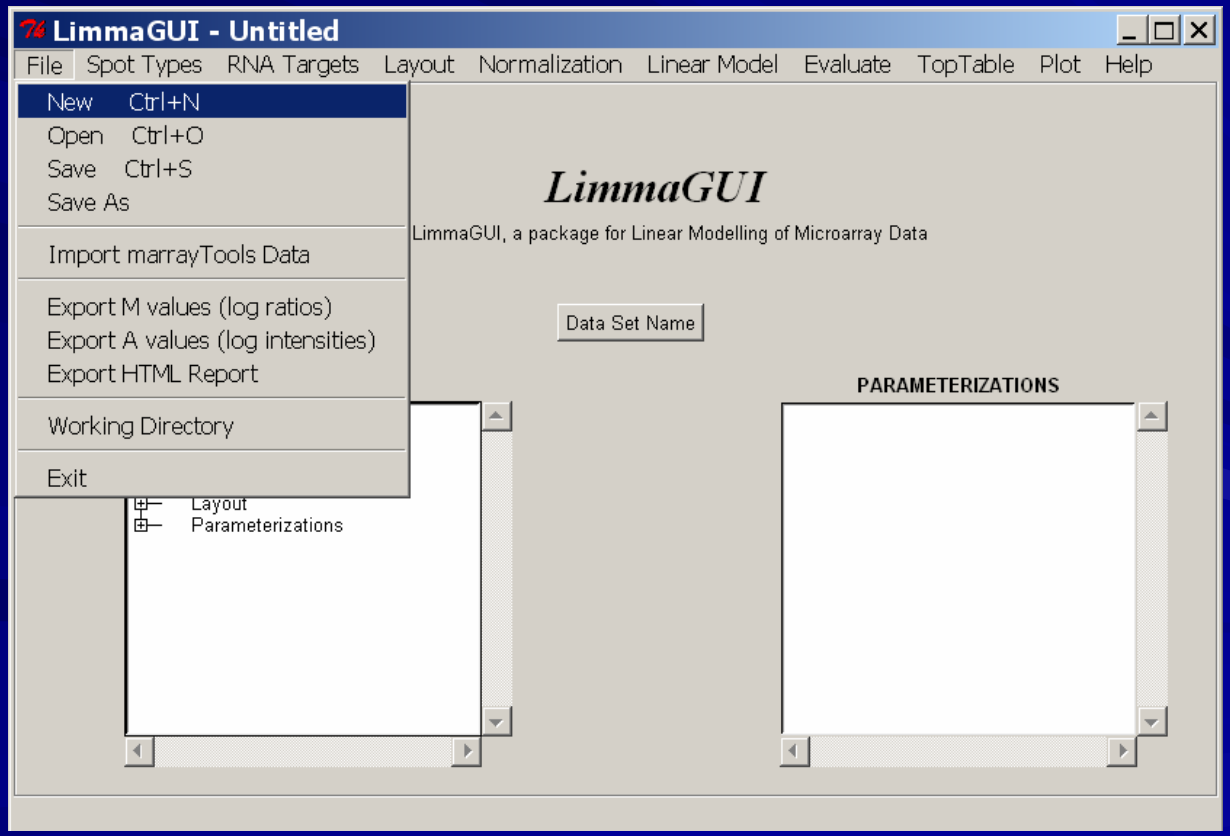
Identification of differentially expressed genes

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



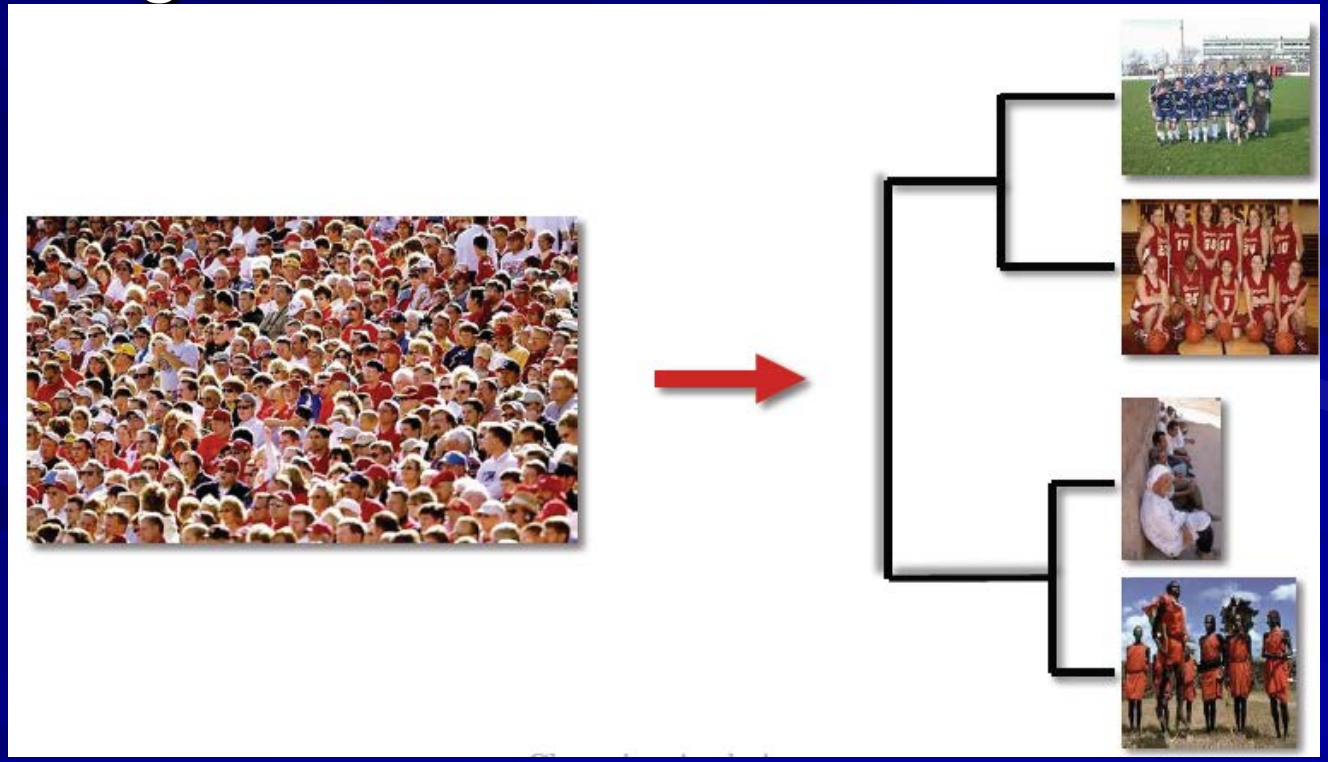
Identification of differentially expressed genes

- 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)

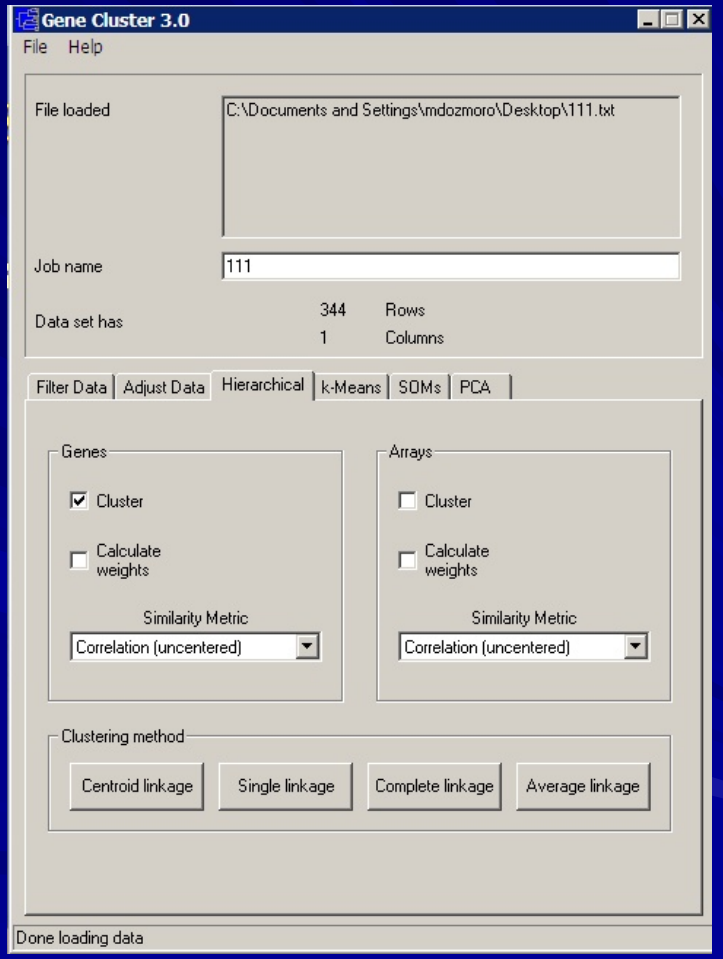
- Divisive:
- SOTA (Herrero et al. 2001)
- DIvisive ANAlysis clustering (Kaufman & Rousseeuw, 1990)
- Gene Shaving (Hastie et al, 2000)

- • 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



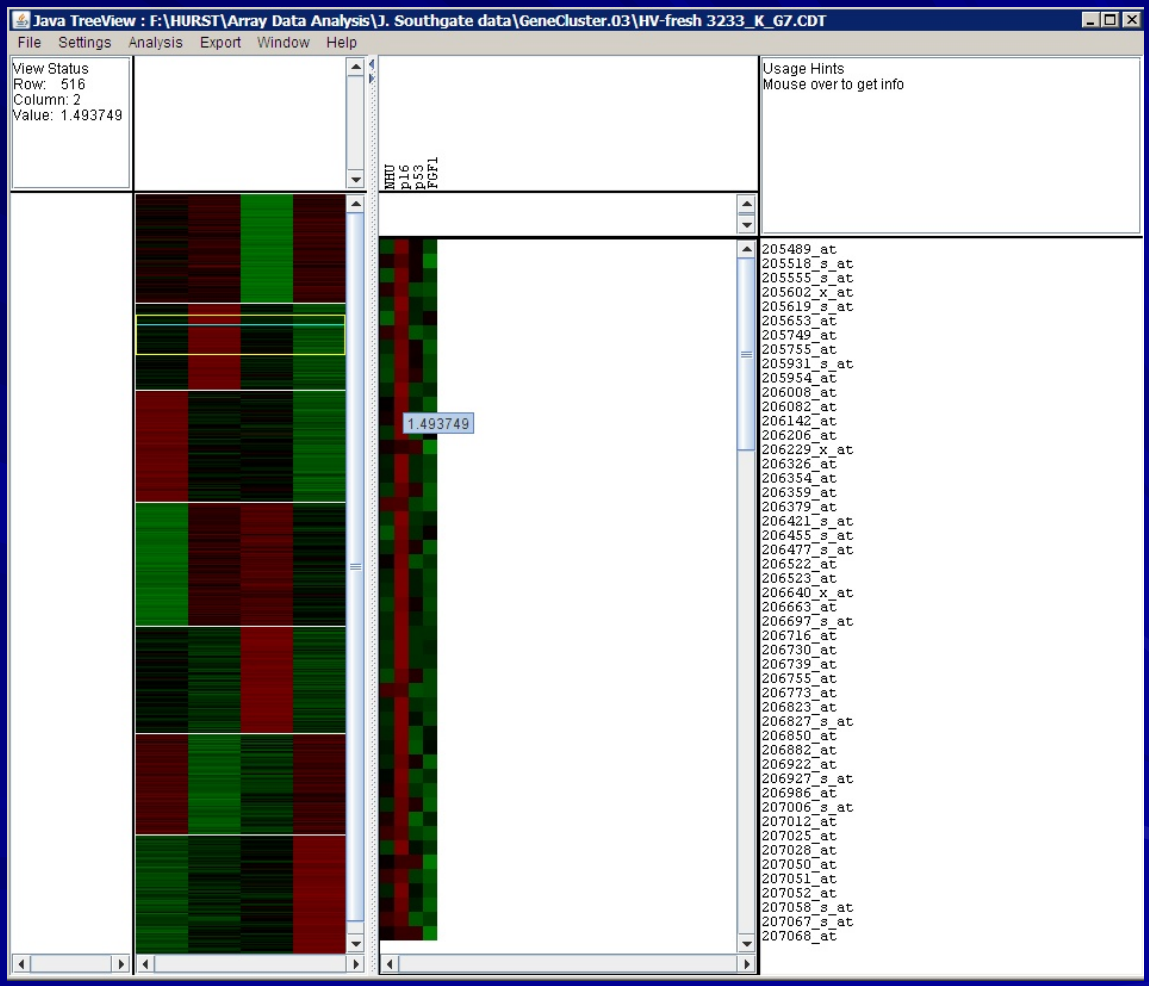
http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster

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

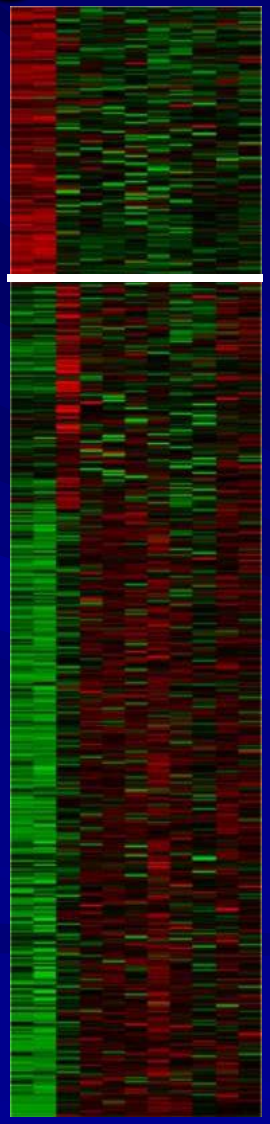


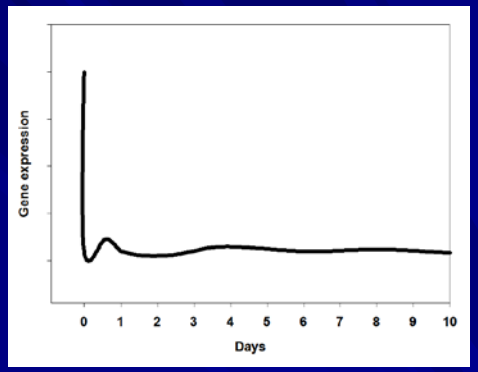
http://sourceforge.net/projects/jtreeview

Clustering interpretation

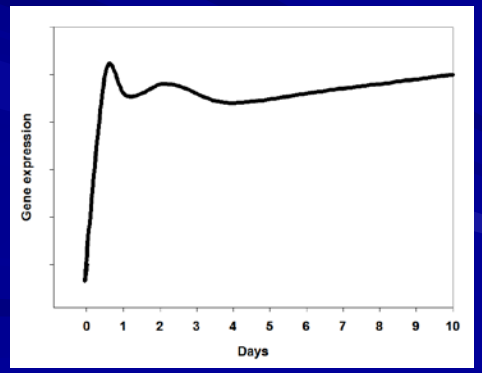
- Hierarchical clustering – genes with similar expression profiles clustered together

Red/**Green** – high/low expression level, respectively





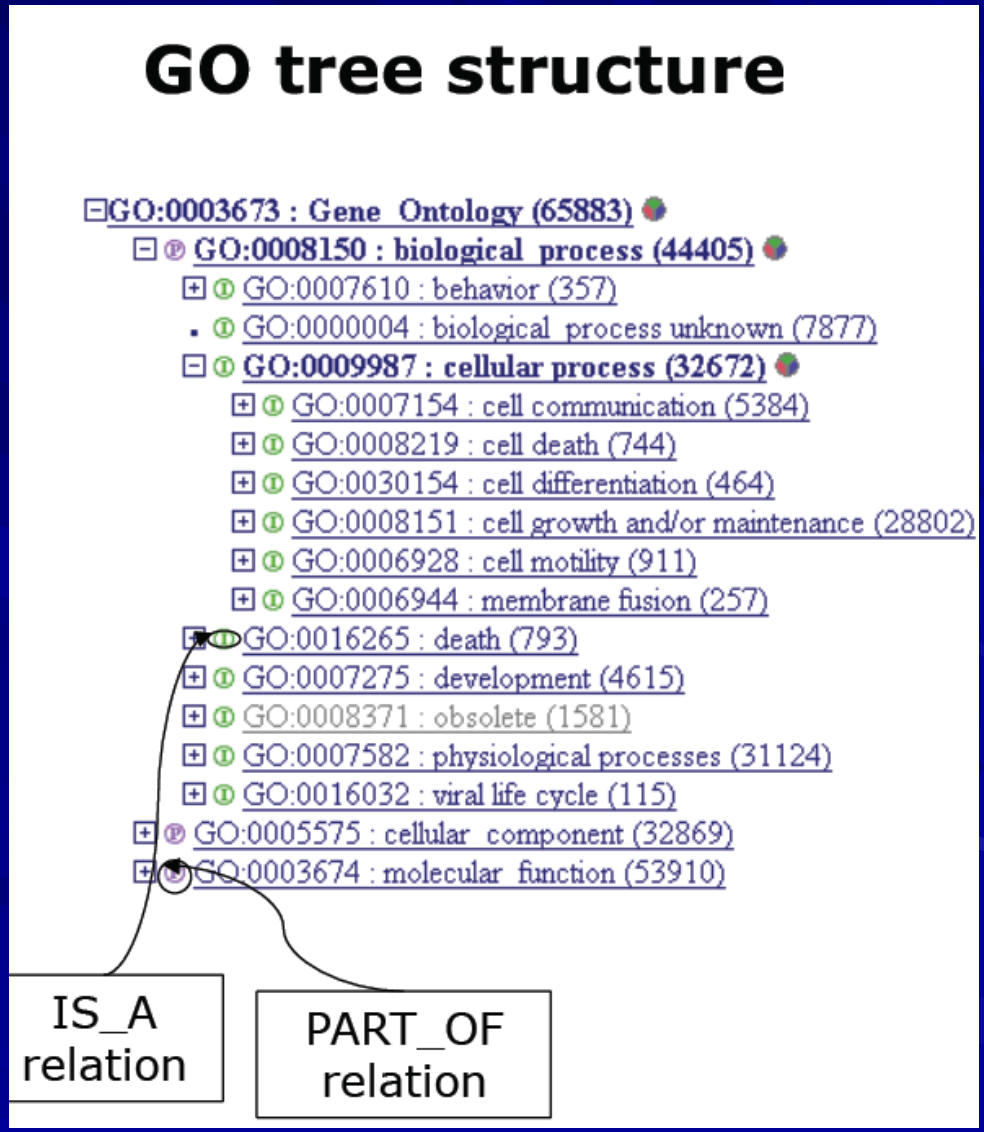
Genes turned off



Genes turned on

Gene Ontology

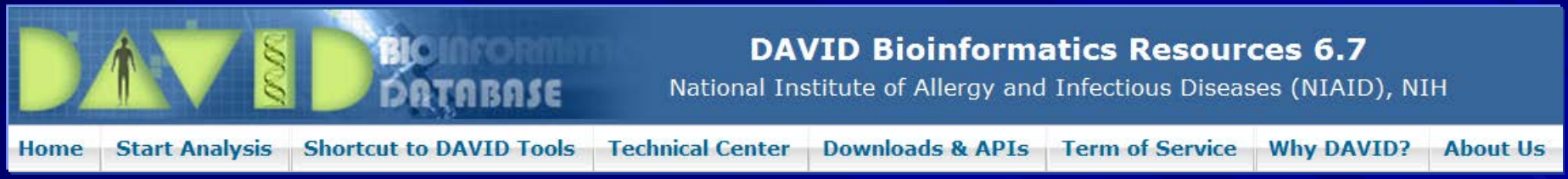
- Aim - to unify the representation of genes and gene product' attributes across all species
- The three categories of GO
  - Molecular Function
  - Biological Process
  - Cellular Component



Gene Ontology analysis

DAVID - The Database for Annotation, Visualization and Integrated Discovery

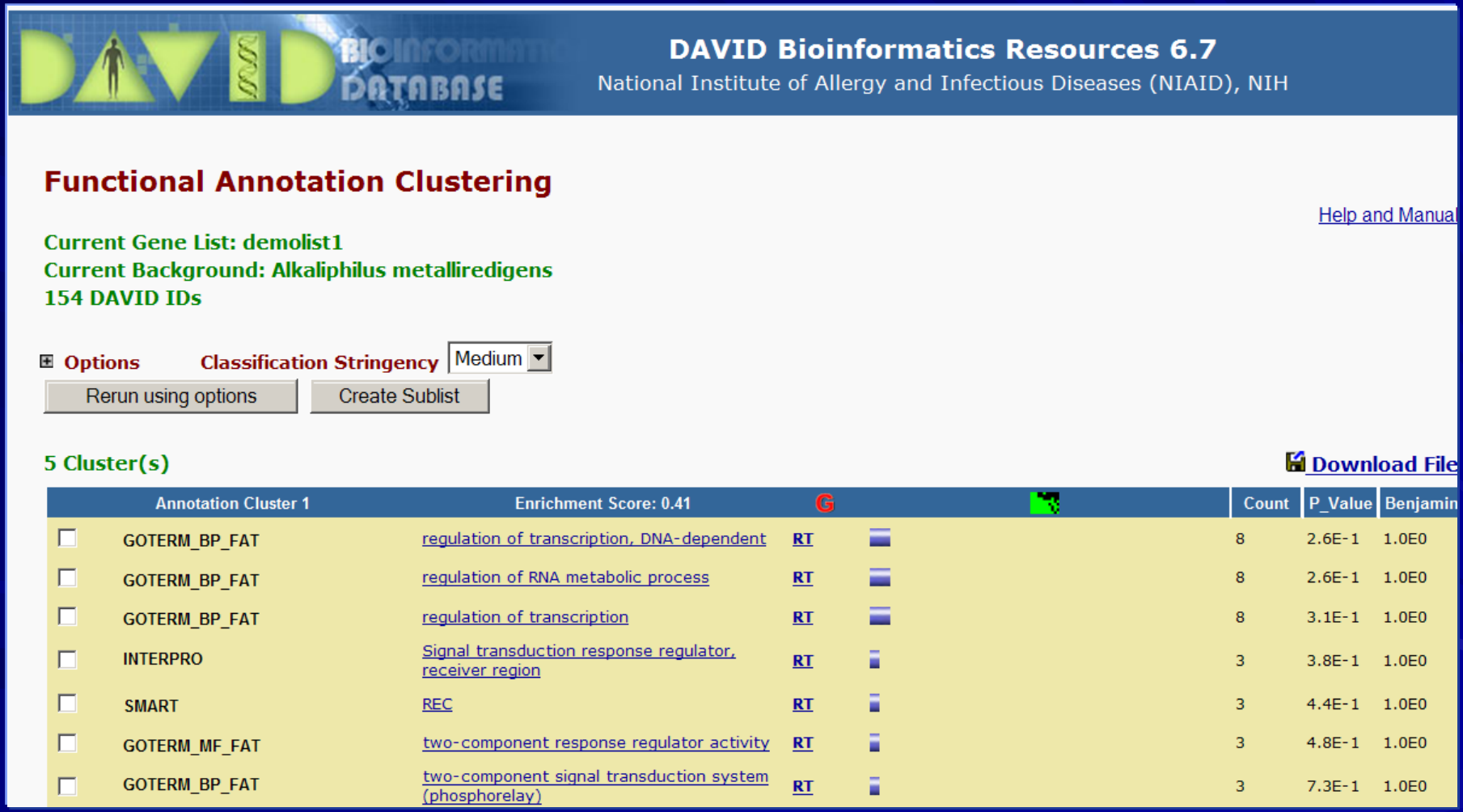
http://david.abcc.ncifcrf.gov/



- 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

Help and Manual

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

Options Classification Stringency: Medium

Rerun using options Create Sublist

5 Cluster(s)

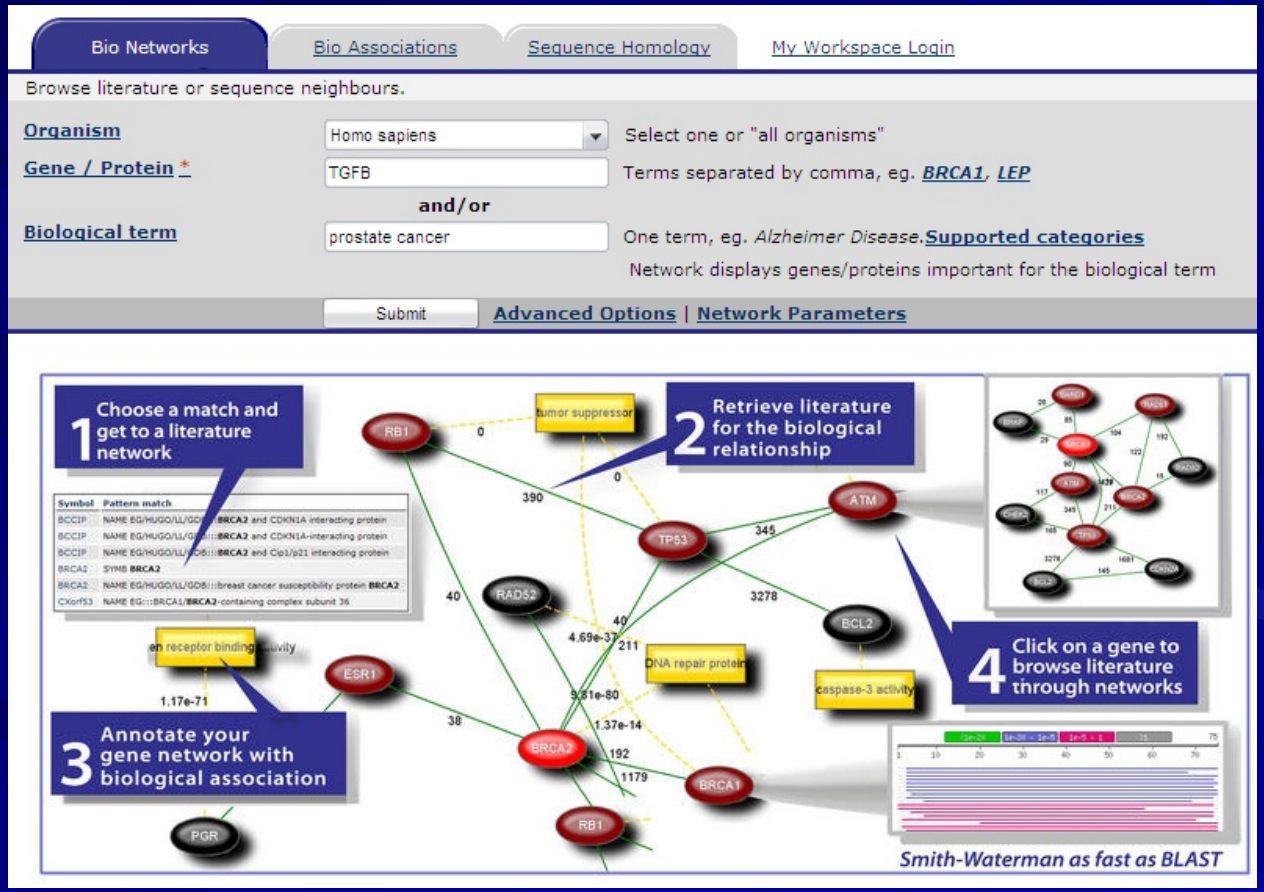
Download File

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

Gene interactions

PubGene - networks, associations and more

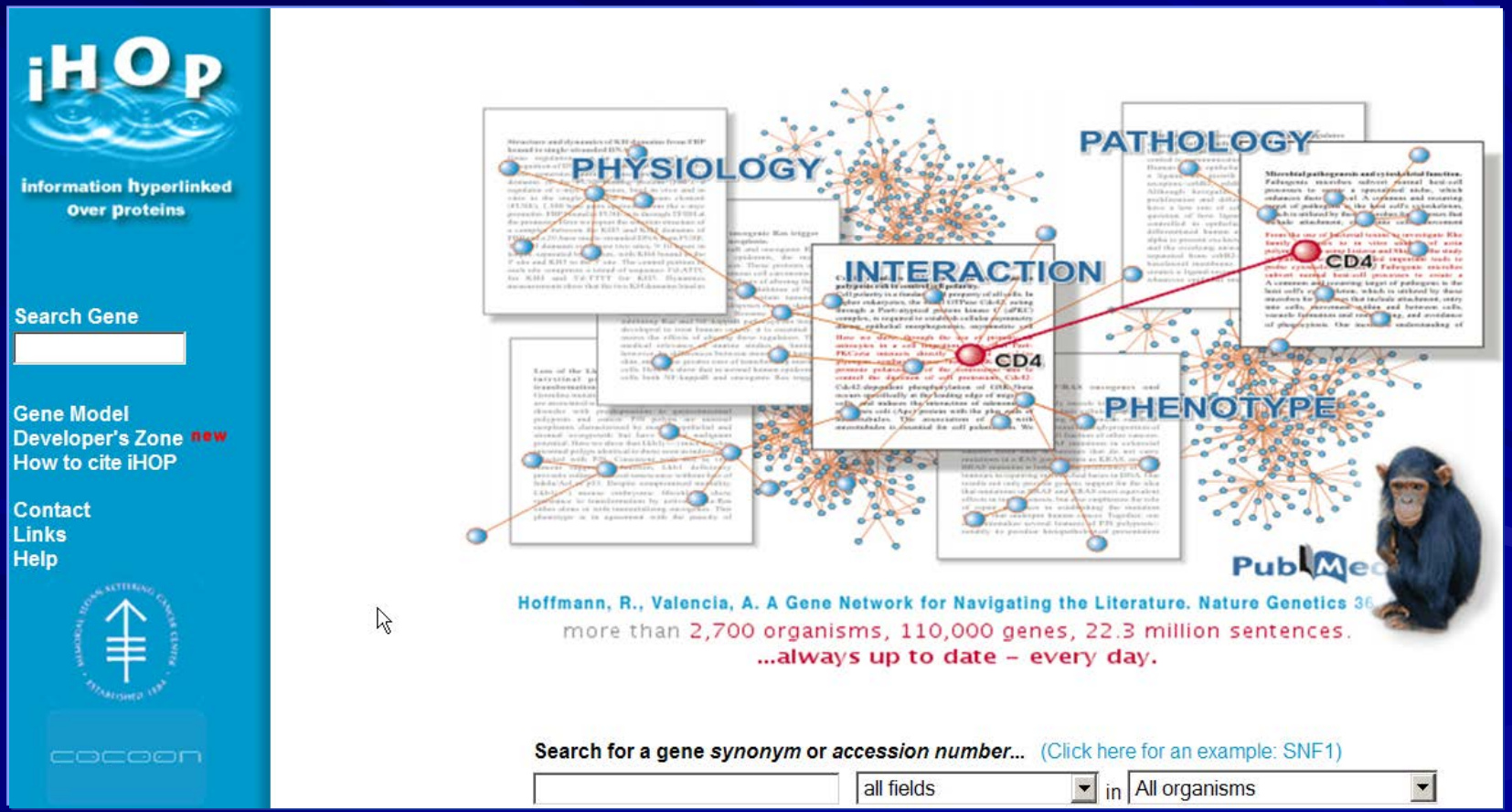
http://www.pubgene.org



Gene interactions

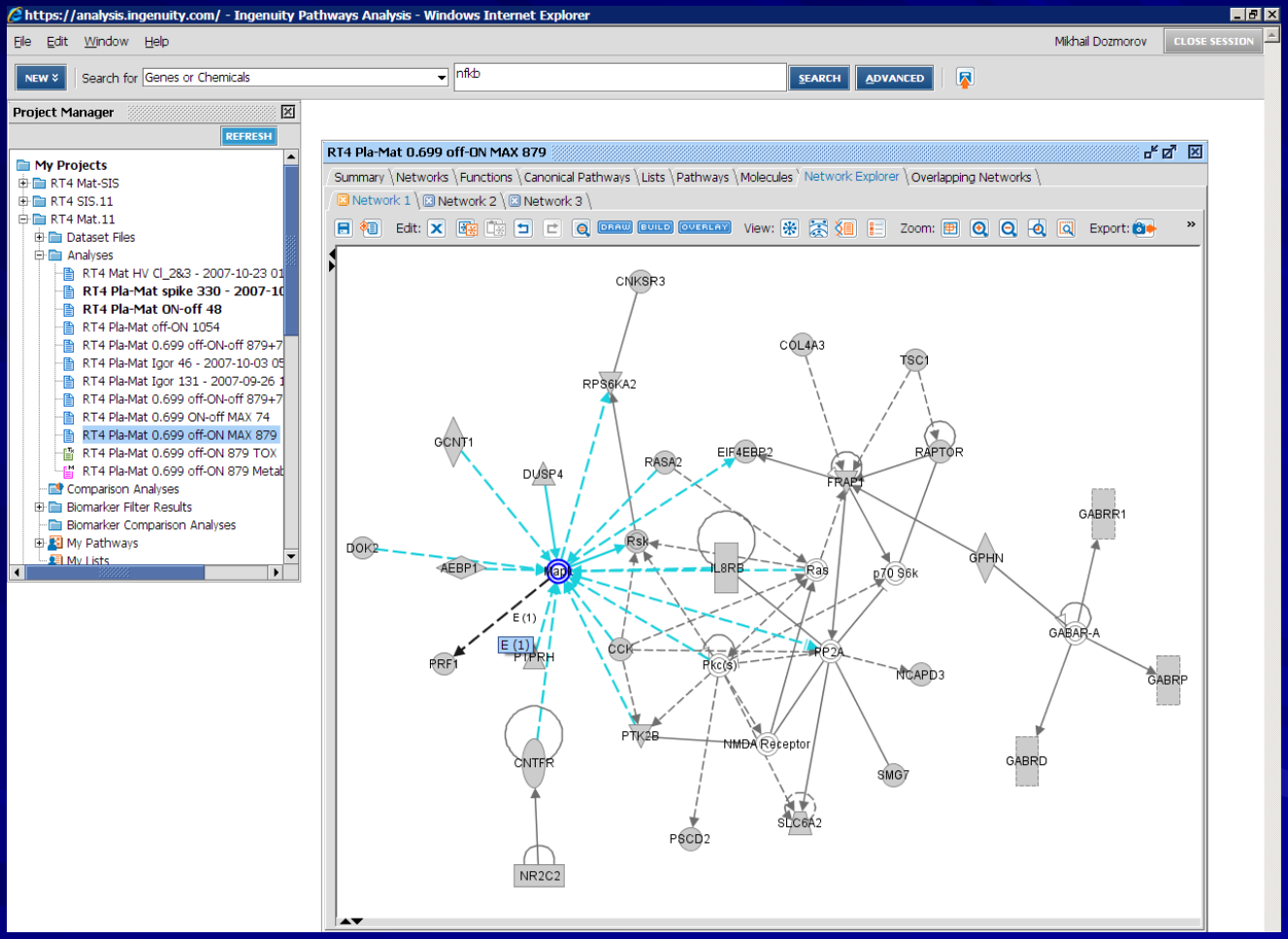
iHOP - Information Hyperlinked over Proteins

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



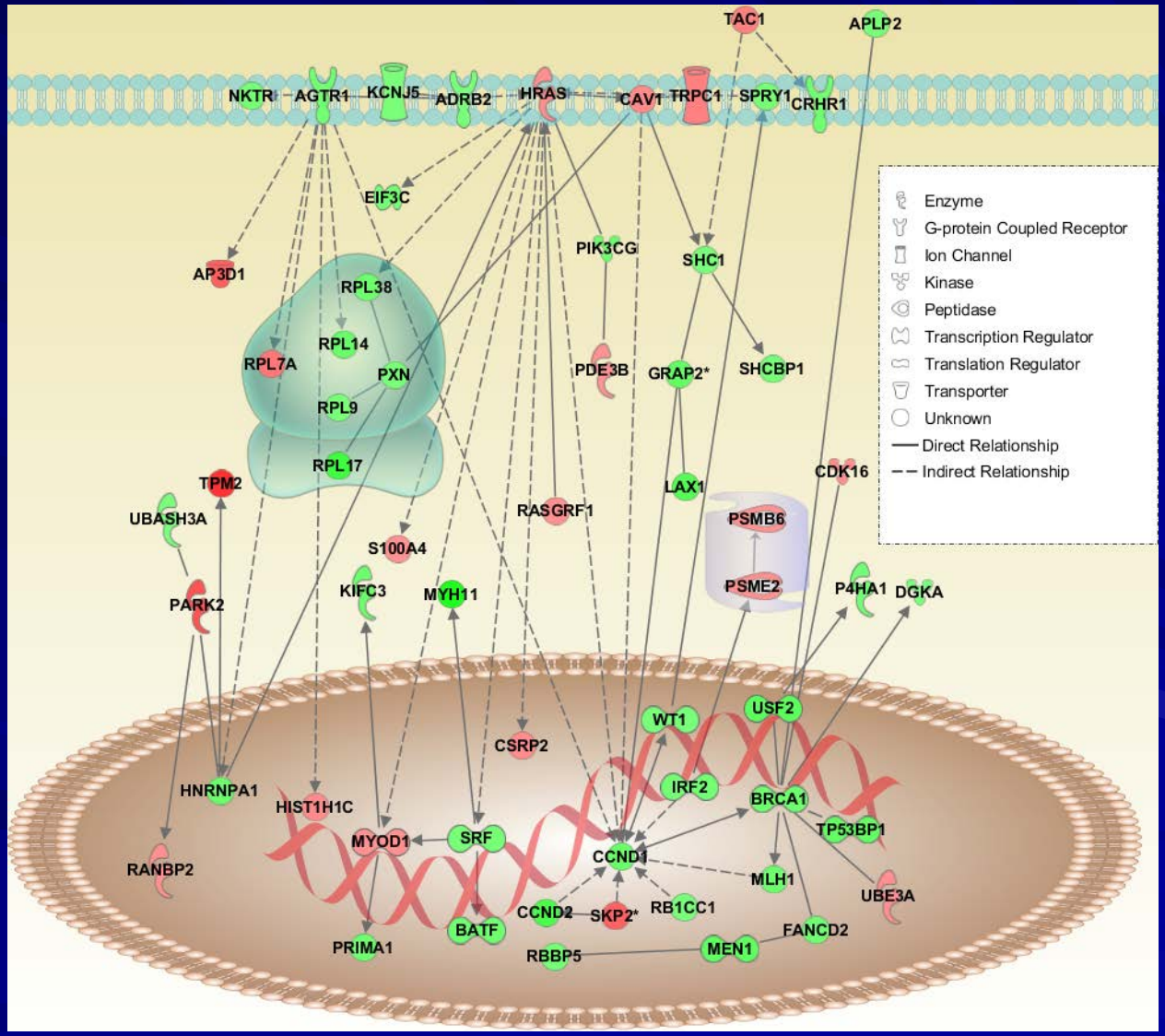
Pathway analysis

Ingenuity Pathway Analysis



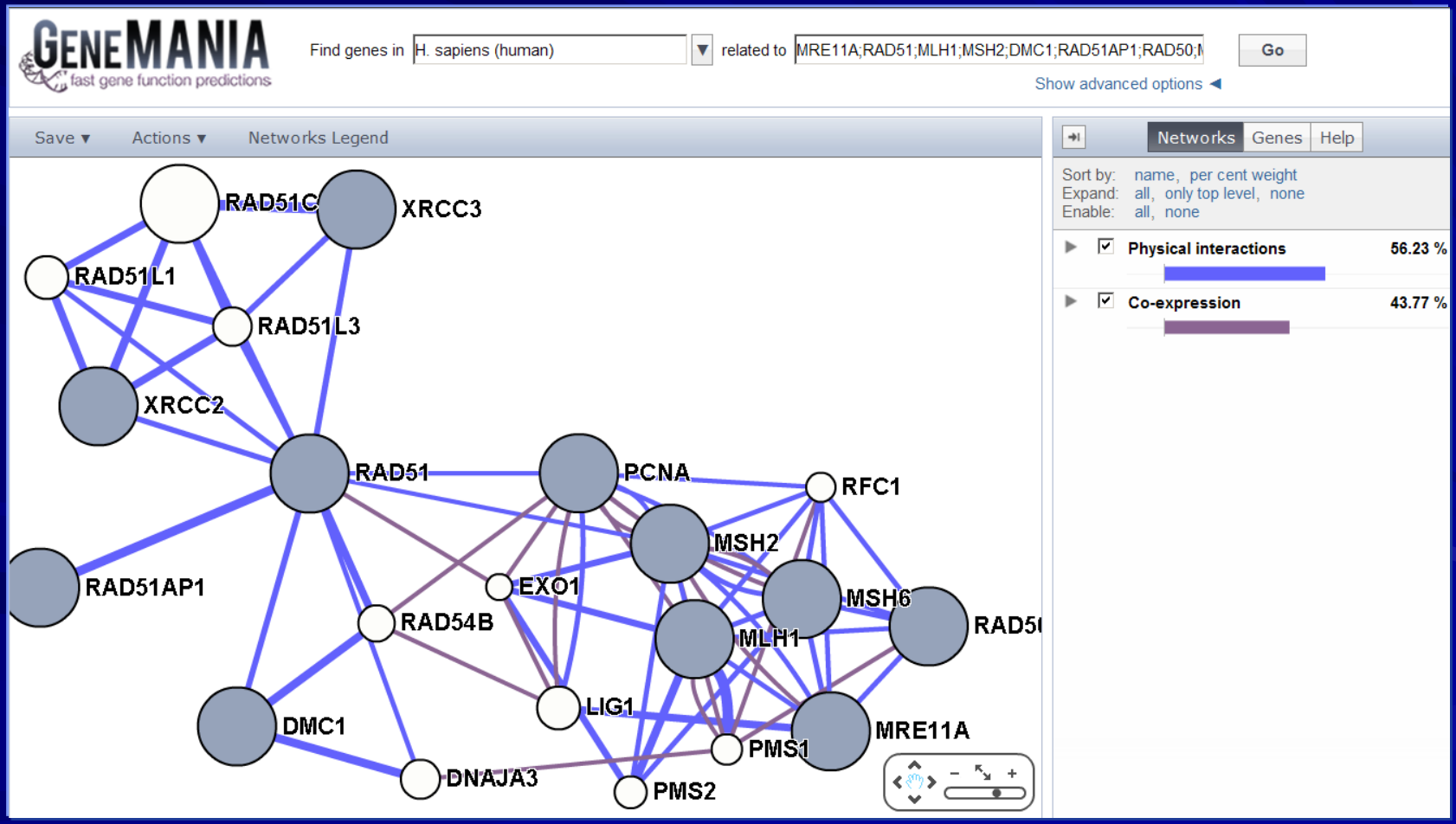
- INGENUITY – analysis and visualization of overrepresented gene interactions http://www.ingenuity.com/

Ingenuity network

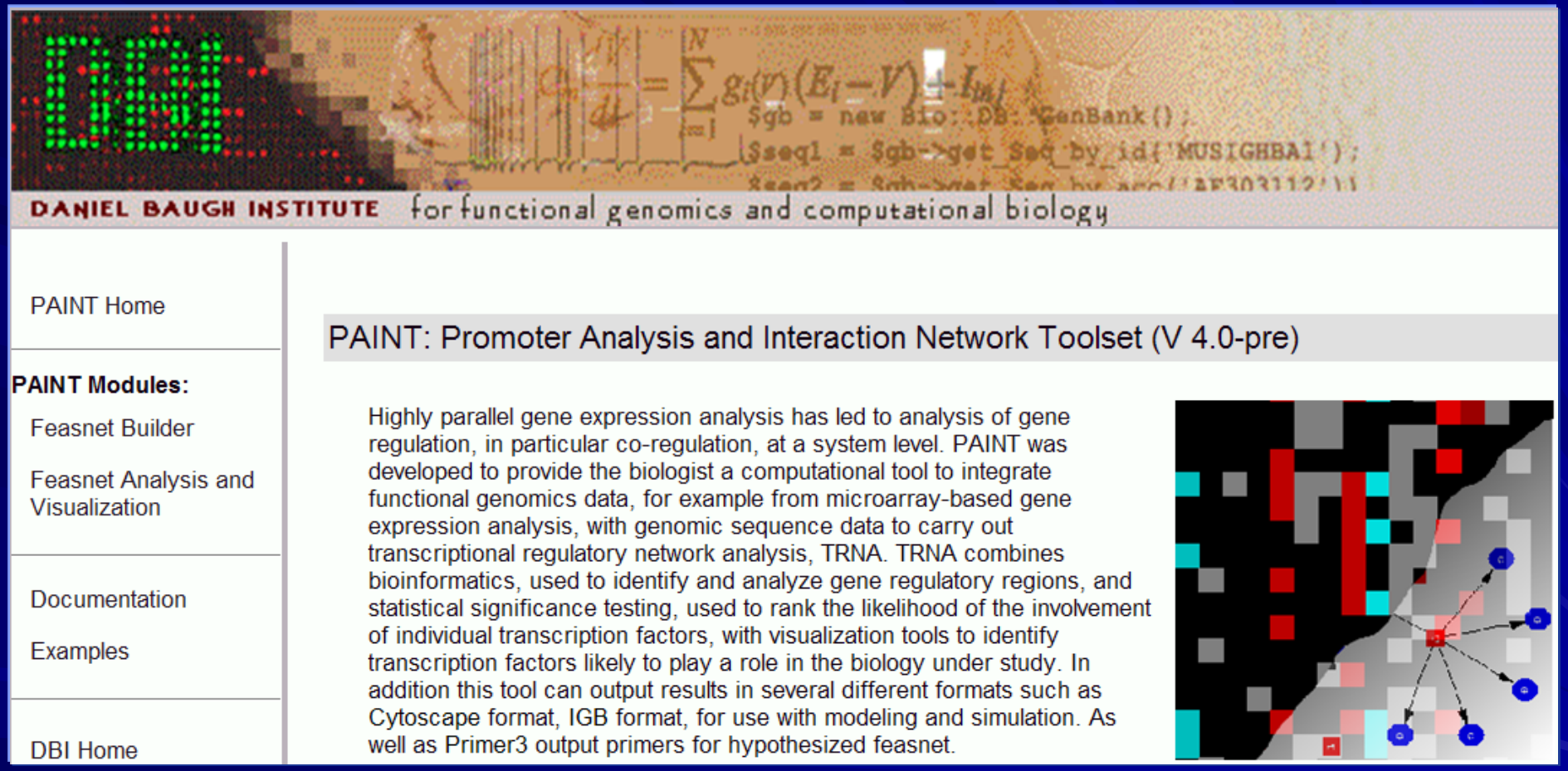


Gene interactions

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

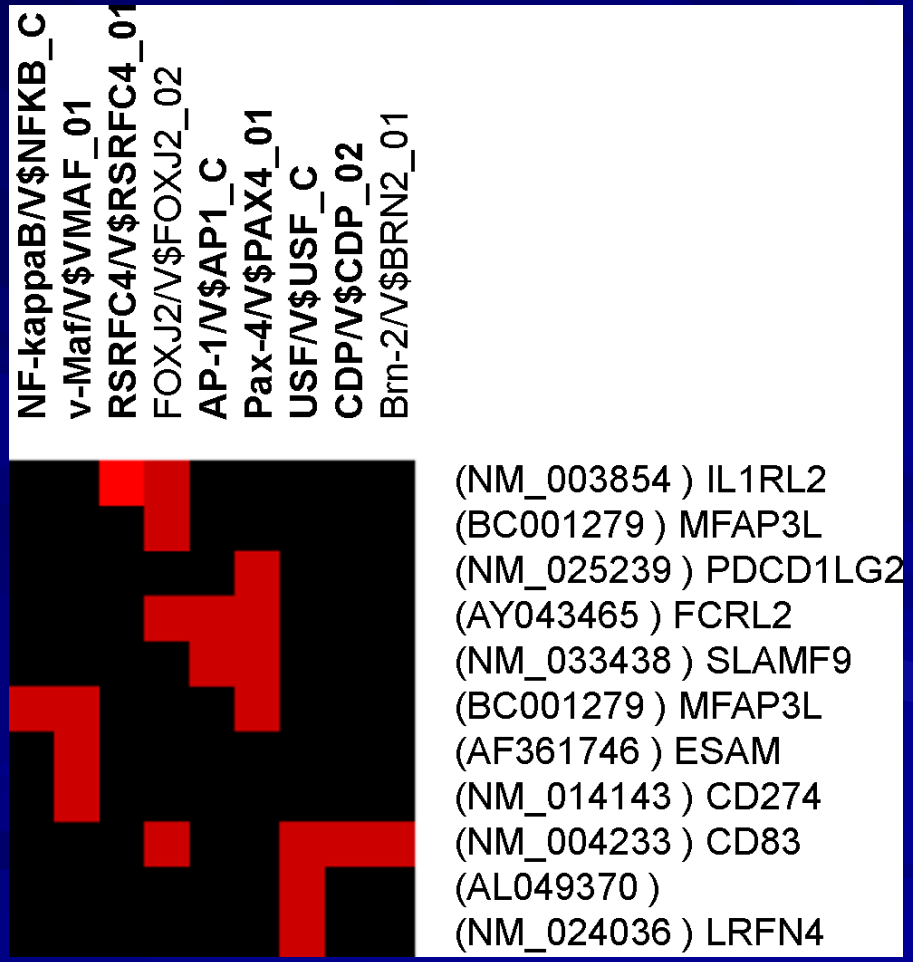


Promoter Analysis



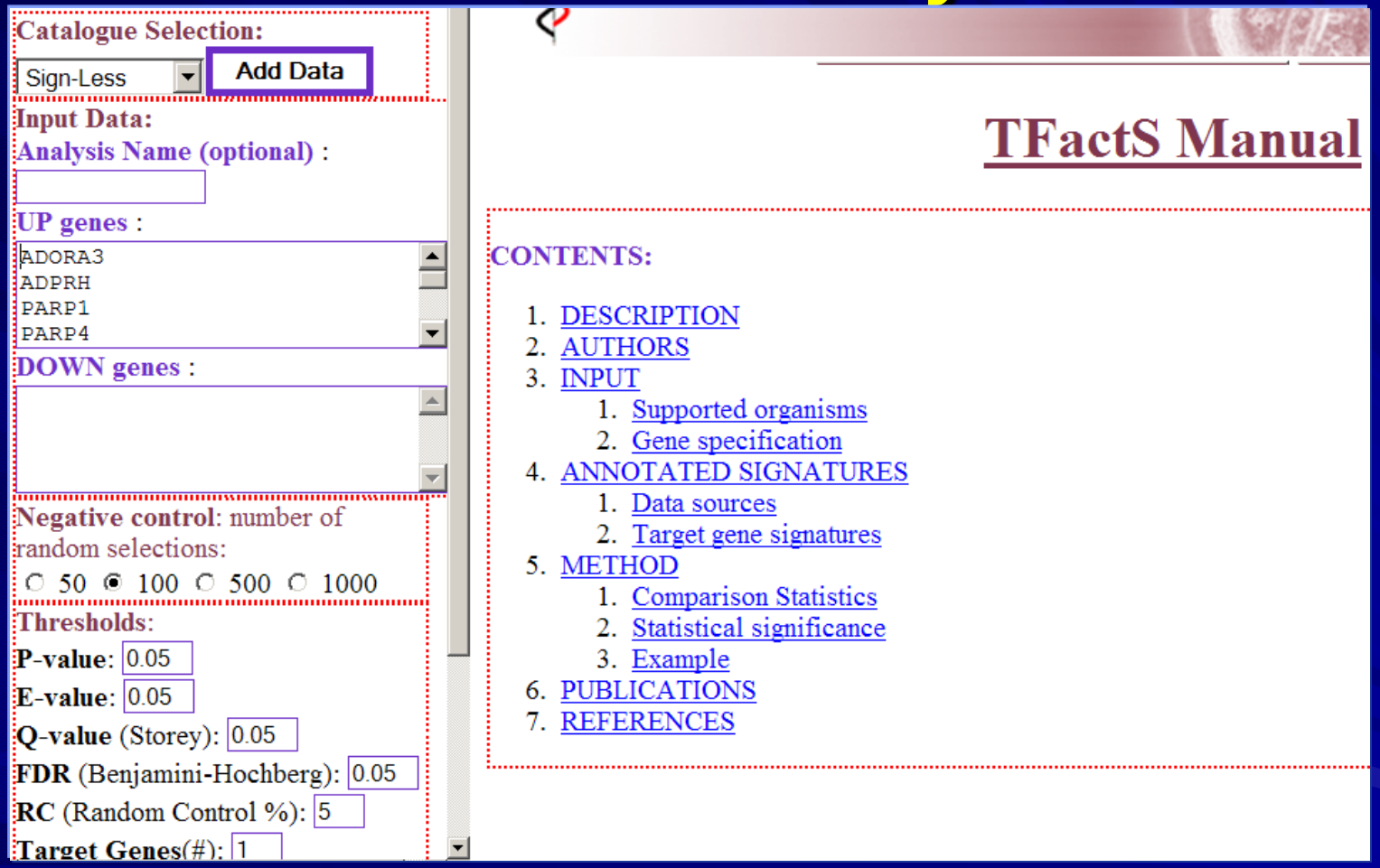
- PAINT - finds common promoters upstream of TSSs of the genes in a list. http://www.dbi.tju.edu/dbi/tools/paint/

P.A.I.N.T. Analysis



Red – statistically significant TREs

TFACTS Analysis



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

TF enrichment results

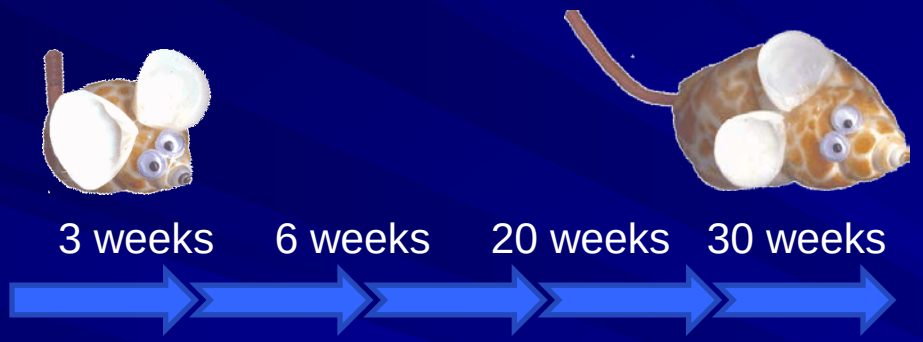
Regulated Transcription Factors:

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.

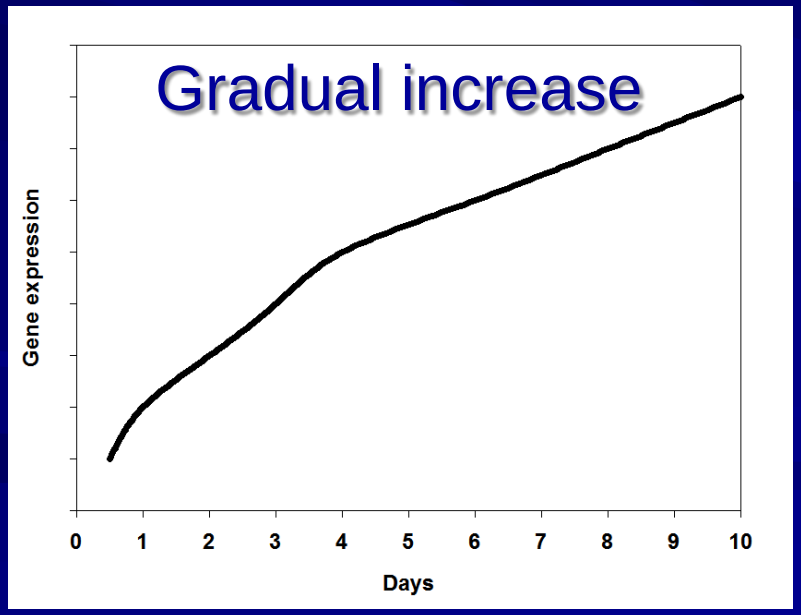


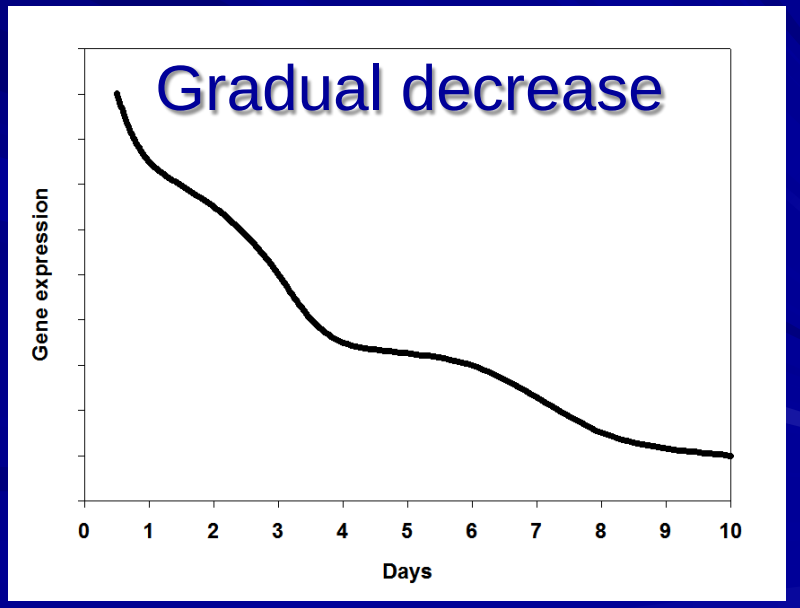
- Analysis of profiles, not differences.

- Profiles may be more complex than just increase or decrease in expression level.

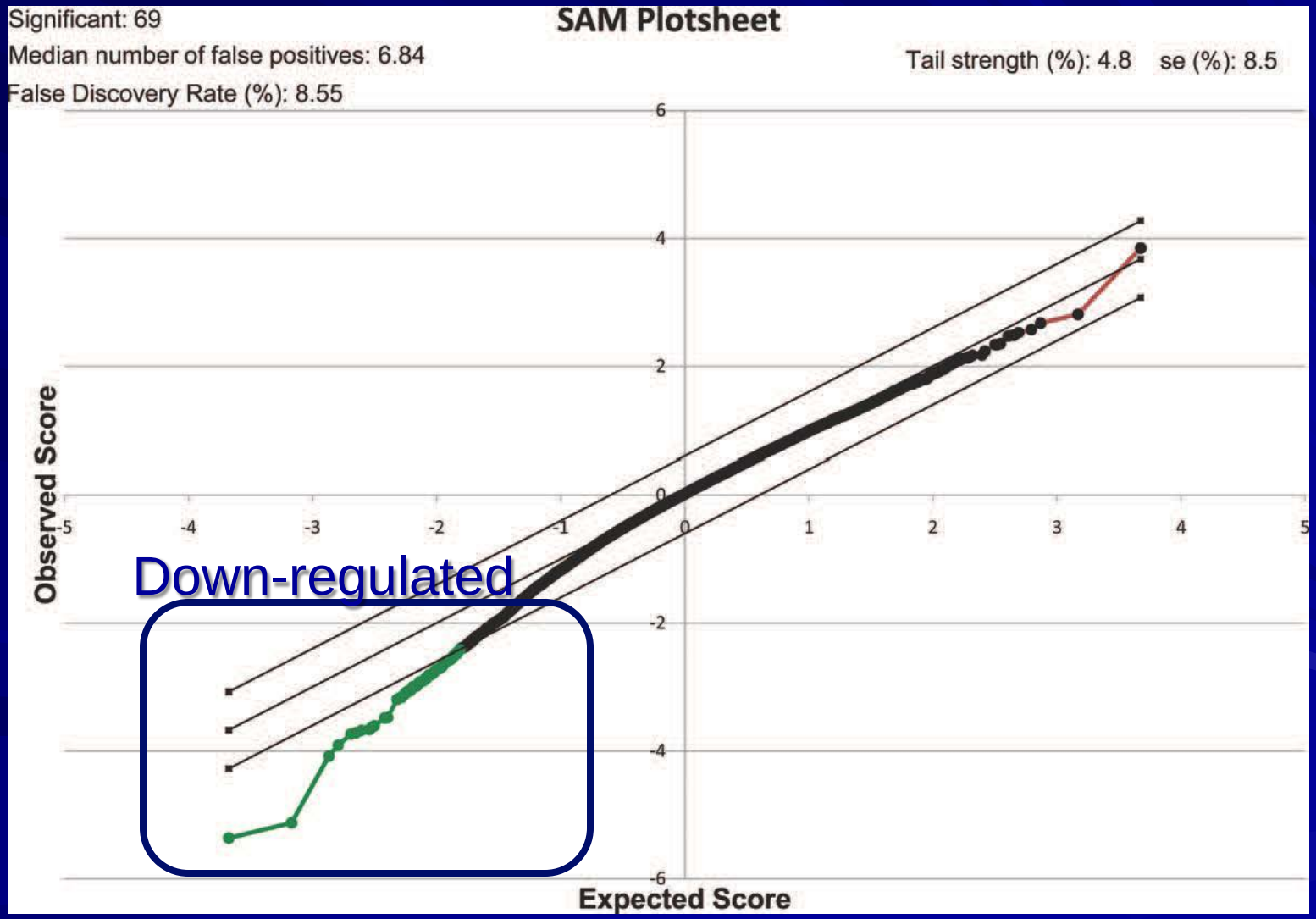
Gradual changes

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





SAM Analysis of Gene Expression Changes

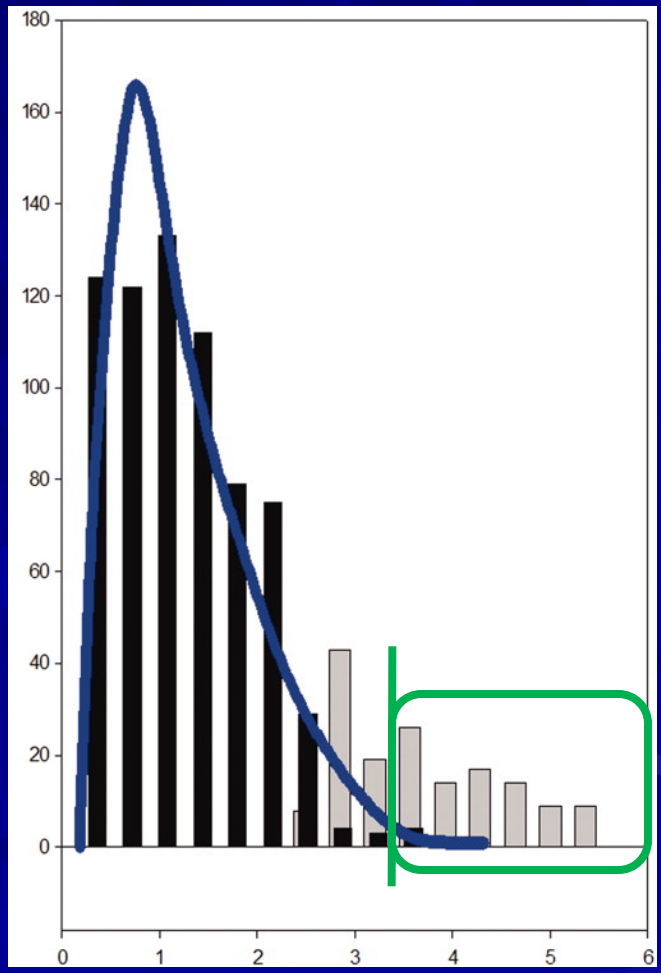


Hypervariable genes

- 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?

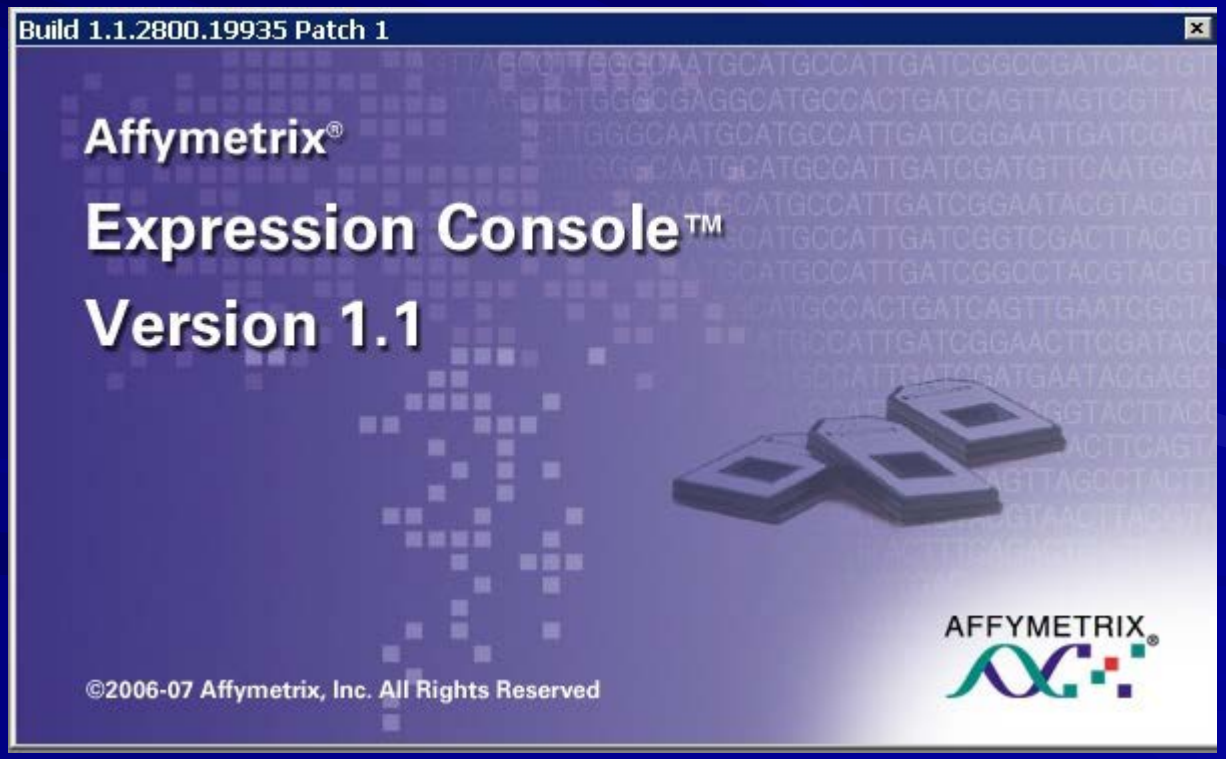
Affymetrix gene and exon arrays

- Affymetrix GeneChip Exon 1.0 ST
  - Wide coverage
  - Well annotated genes plus gene prediction sets
  - Over 1.4 million probe sets

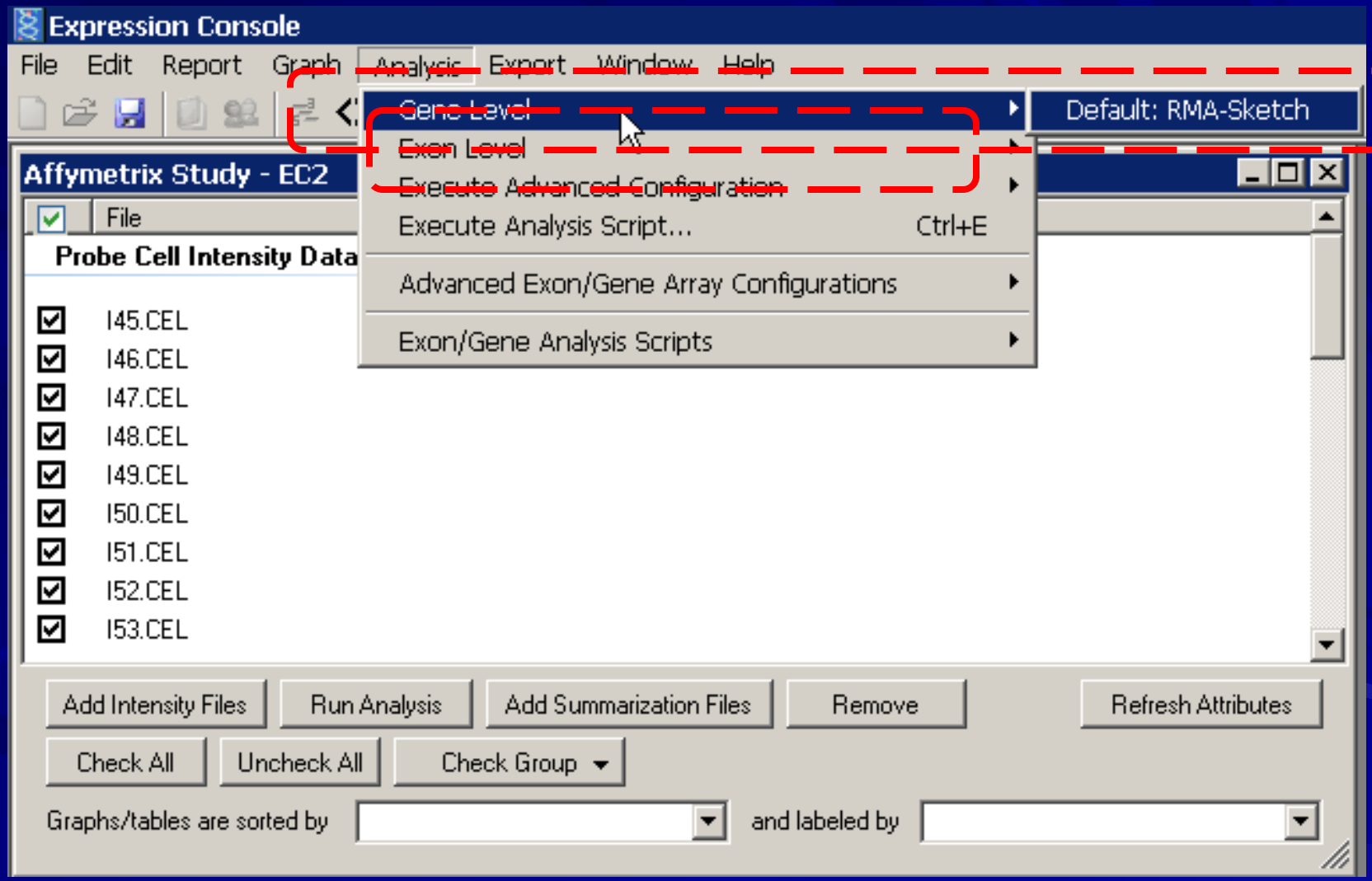
Affymetrix Expression Console (EC)

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

EC1.studyRMA - GeneRMA-GENE-DEFAULT with Annotations.xls

	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
[illegible]	33256	8180417	9.841396	9.971924	9.867586
33258	33257	8180418	9.557718	9.589325	9.304793

Gene level summarization

EC2.studyFullRMA - EXON-DEFAULT-DABG - NormReg.xlsx

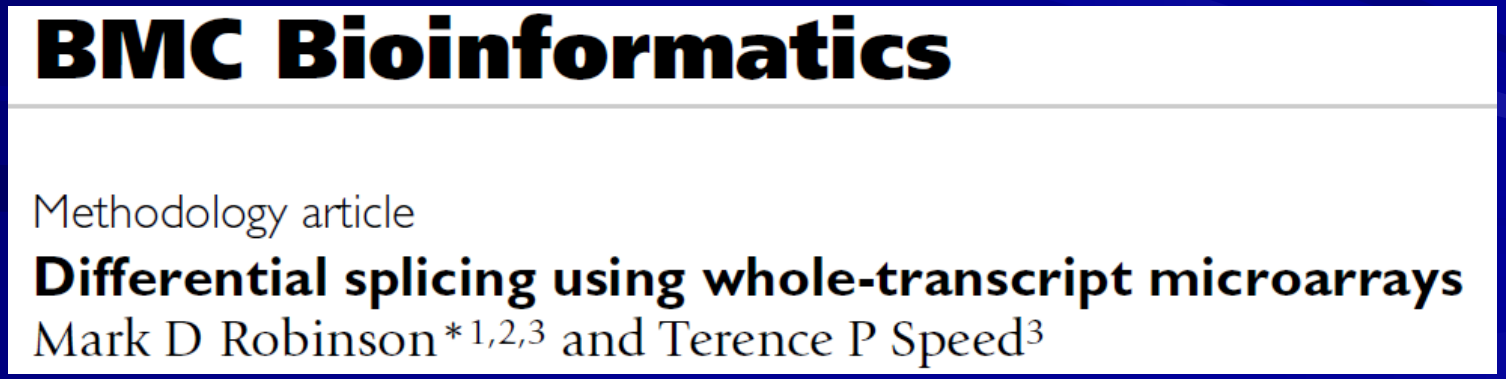
	A	B	C	D	E
1	Probe Set	H2132.rma	H2133.rma	H2134.rma	H2135.rma
2	7896737	7.196188	7.683278	7.728511	7.396492
3	7896739	4.384084	4.567221	4.524869	4.213489
4	7896741	4.729739	4.329544	4.110299	4.344719
5	7896743	9.275142	9.514589	9.103162	9.329247
6	7896745	9.930458	9.603415	9.982117	9.675047
7	7896747	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
[illegible]	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/

FIRMAGene http://bioinf.wehi.edu.au/folders/firmagene/



Analysis

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

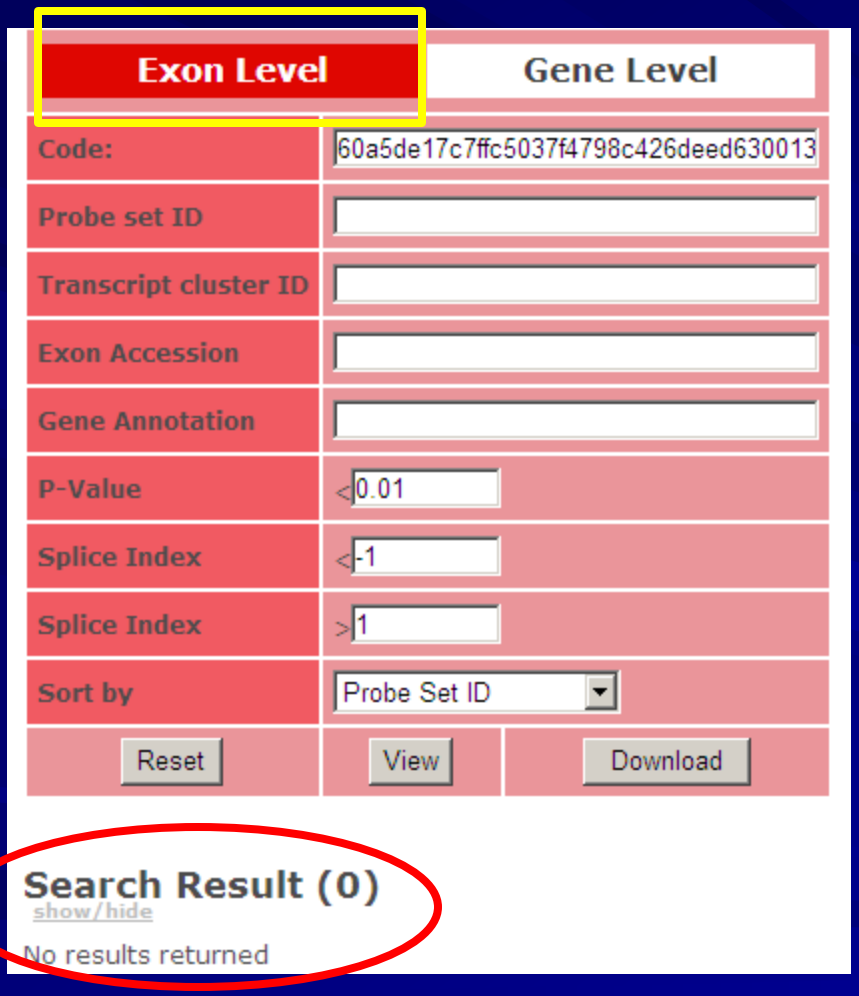
BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 24 2009, pages 3323–3324
doi:10.1093/bioinformatics/btp577

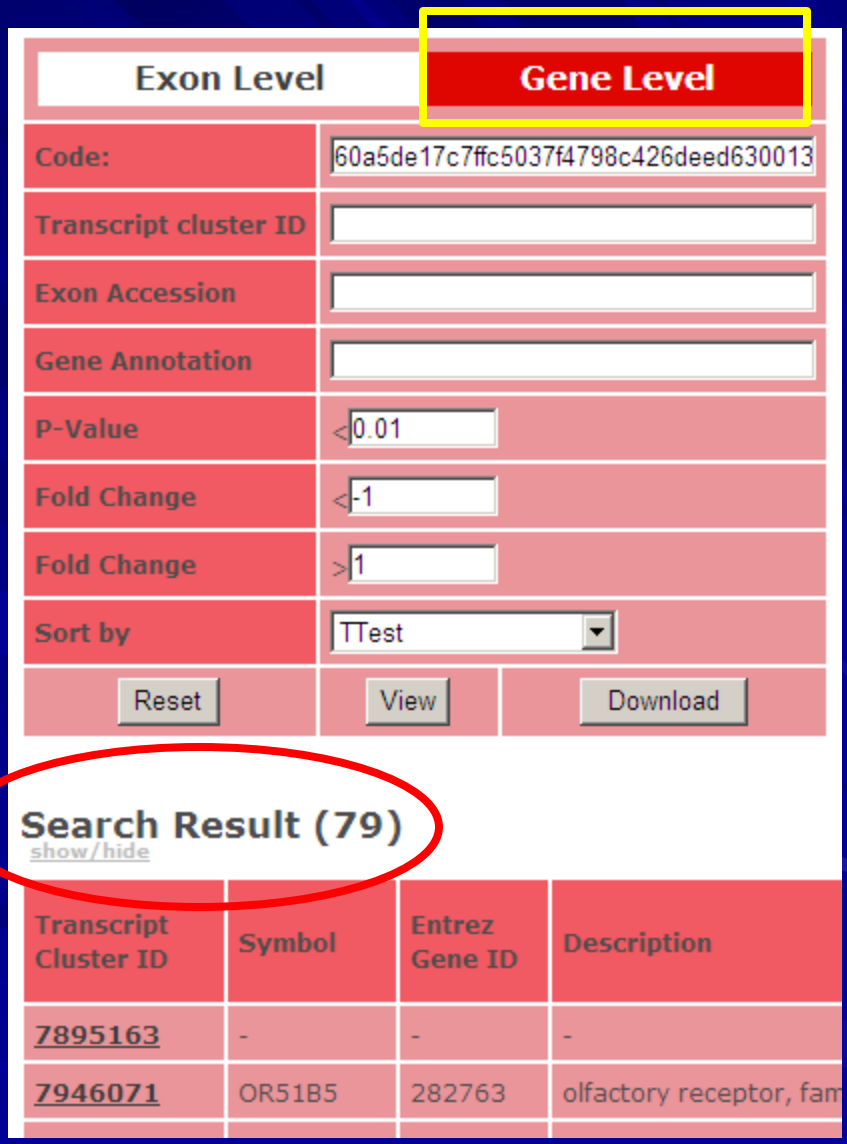
Gene expression

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

Pascal Gellert, Shizuka Uchida and Thomas Braun*

Exon array analysis results



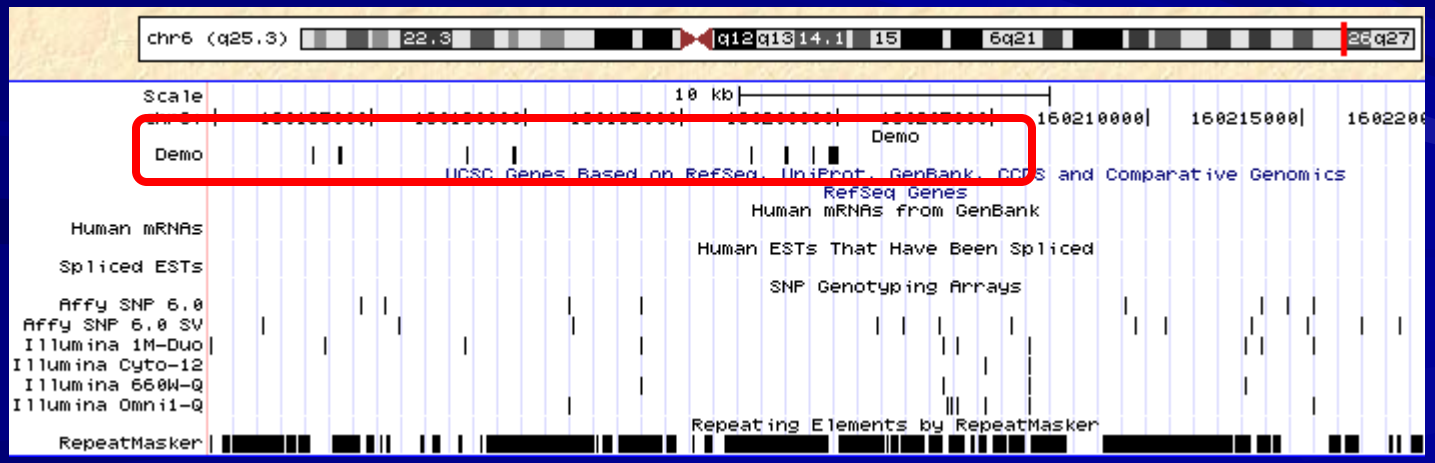


Interpretation of exon results

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

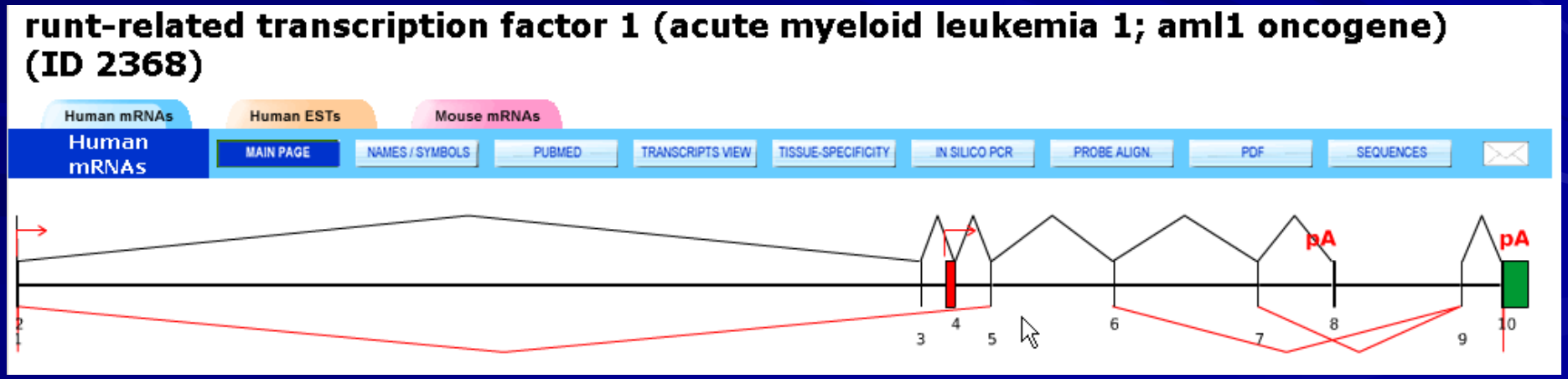
BioMed Central

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*



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

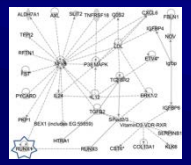
Gene/exon expression

Other high-informational data

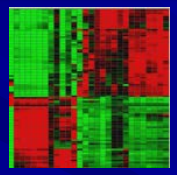
QC and normalization

Genes/exons of interest

Network analysis



Correlational clustering



Ontology clustering

Transcription regulation

Unknown genes

GEO

Oncomine

Connectivity map

Integrated data

Biological validation

Thank you!

Microarray Data processing

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