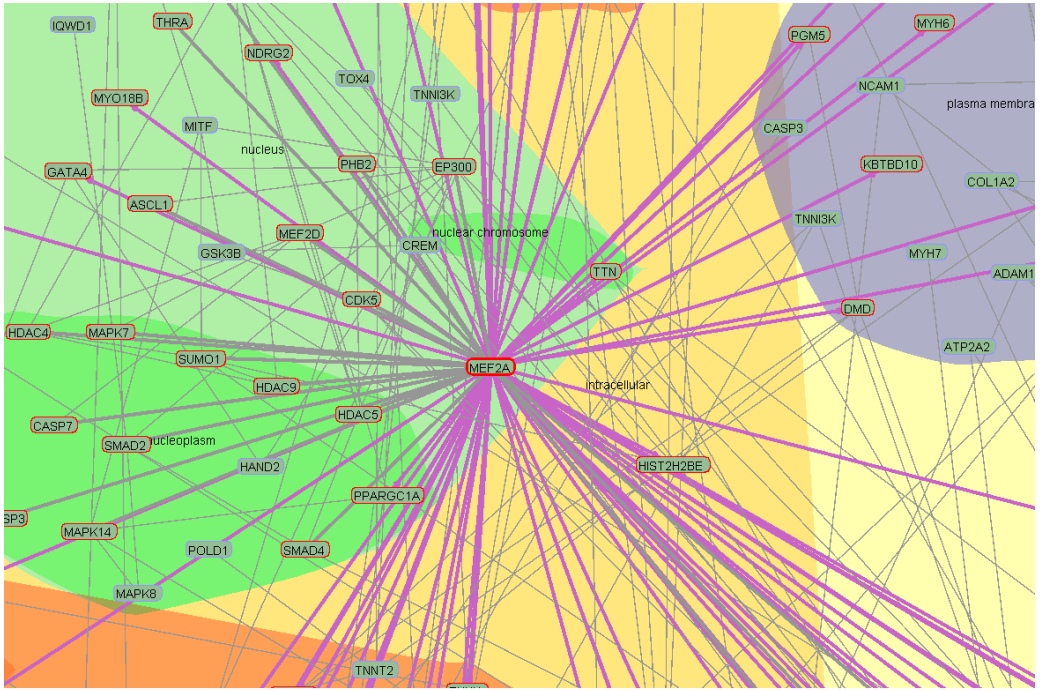
Mining microarray data using TranscriptomeBrowser



Cyrille Lepoivre
Inserm U928/TAGC, Marseille, France

lepoivre@tagc.univ-mrs.fr

A wealth of gene expression data

Gene Expression Omnibus (NCBI)

- 18 527 experiments
- 470 592 samples
- 7 779 platforms

Array Express (EBI)

- 13 268 experiments
- 370 011 samples

Other databases

- NASCArray (Arabidopsis thaliana)
- RED (Rice)
- SGD (Yeast)
- SMD
- ChipDB
- ExpressDB
- MCHiPS
- ...

Data mining approaches

Data browsers

Web sites

- Gene Expression Atlas
- BioGPS
- GEO profiles / GEO datasets
- Oncomine
- GeneChaser

Tasks

- Display individual gene expression profiles
- View clustered datasets
- Search for differentially expressed genes

Custom analysis

Tools

- Bioconductor Libraries
- TMeV
- GSEA
- Genomics Portals
- ...

Analysis types

- Normalization
- Differential expression analysis
- Clustering
- Annotations enrichment
- Network reconstruction
- Data integration

GEO datasets (NCBI)

NCBI

DataSet Browser (Curated)

Gene Expression Omnibus

Search for GDS3709[ACCN] Search Clear Show All Advanced Search

DataSet Record GDS3709: Expression Profiles | Data Analysis Tools | Sample Subsets

Title:	Cigarette smoke effect on the oral mucosa
Summary:	Analysis of oral mucosae from 40 cigarette smokers and 40 age and gender matched never-smokers. Results provide insight into the carcinogenic effects of cigarette smoke.
Organism:	*Homo sapiens*
Platform:	GPL570: [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
Citation:	Boyle JO, Gümüs ZH, Kacker A, Choksi VL et al. Effects of cigarette smoke on the human oral mucosal transcriptome. *Cancer Prev Res (Phila Pa)* 2010 Mar;3(3):266-78. PMID: 20179299

Reference Series:	GSE17913	**Sample count:**	79
Value type:	count	**Series published:**	2010/02/15

Cluster Analysis

Download

Data Analysis Tools

Find genes

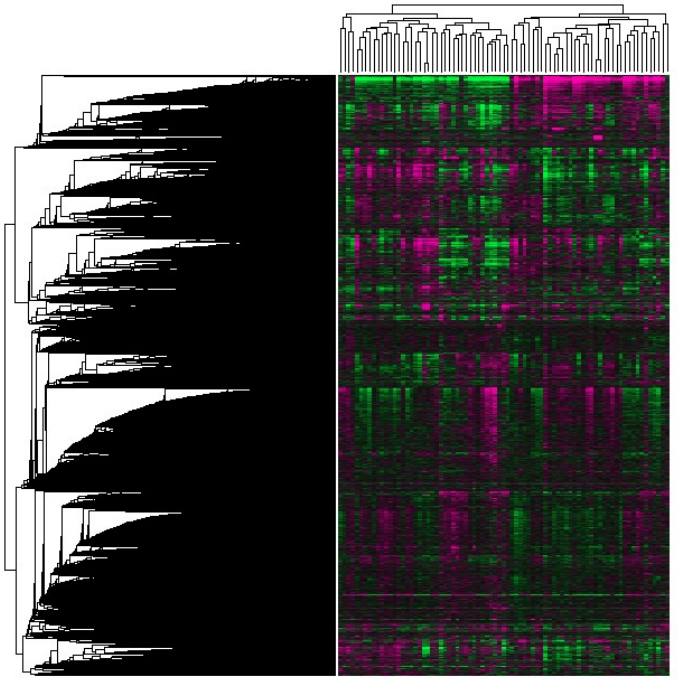
- Compare 2 sets of samples
- Cluster heatmaps
- Experiment design and value distribution

Find gene name or symbol:

Find genes that are up/down for this condition(s): gender, agent Go

GDS3709 Cigarette smoke effect on the oral mucosa [Homo sapiens]

Clustering: Uncentered Correlation UPGMA Colors: High Low Full image: 54675 x 79 spots



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

Gene Expression Atlas (EBI)

CD3E JSON XML

Homo sapiens

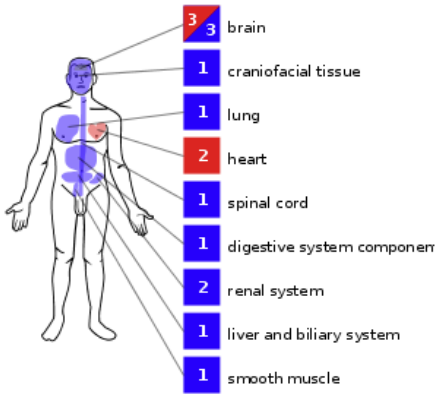
CD3E is differentially expressed in 103 experiments [107 up/97 dn]: 22 organism parts: thymus [5 up/0 dn], amygdala [0 up/3 dn], ...; 32 disease states: normal [5 up/6 dn], promyelocytic leukemia HL-60 [2 up/0 dn], ...; 29 cell types, 98 cell lines, 18 compound treatments and 16 other conditions.

Synonyms	CD3E, T3E
Orthologs	NP_001101610.1 (Rattus norvegicus) CD3E_BOVIN (Bos taurus) Cd3e (Mus musculus) (Compare orthologs)
UniProt Accession	P07766
Gene-Disease Assocation	Immunodeficiency due to defect in CD3-epsilon, Severe combined immunodeficiency, T cell-negative, B-cell/natural killer-cell positive
Gene Ontology Term	SH3 domain binding, T cell receptor binding, external side of plasma membrane, integral to plasma membrane, protein kinase binding, ...
InterPro Term	CD3 epsilon chain, Immunoglobulin subtype 2, Phosphorylated immunoreceptor signaling ITAM, Immunoglobulin-like fold
Search EB-eye	ENSG00000198851

⊞ Show more properties

Experimental Factors

Organism part
studied in E-GEOD-2665, E-TABM-145a, E-AFMX-5, E-MEXP-1600, E-AFMX-6, ... (11 experiments)



number of published studies, where the gene is over/under expressed compared to the gene's overall mean expression level in each study

show this factor only>>

Cell line
studied in E-TABM-321, E-MEXP-149, E-TABM-157, E-MEXP-440, E-MEXP-1014, ... (15 experiments)

Factor Value	U/D	Experiments
293	1	E-GEOD-1880
600mpe	1	E-TABM-157
A549	1	E-GEOD-4127
Abc-1	1	E-GEOD-4127
Au565	1	E-TABM-157
Bc-1	1	E-GEOD-1880

92 more value(s).
show this factor only>>

Expression Profiles

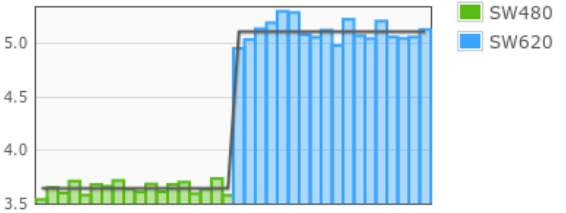
1 2 3 4 5 ... 20 21

103 experiments showing differential expression

E-MEXP-1014: Transcription profiling of two human colon cancer cell lines treated with n-3 PUFA docosahexaenoic acid for 3 different time points

Experimental Factors

Time | Cell Line | Dose | Compound Treatment



Show expression profile / experiment details

The TranscriptomeBrowser project : motivations

Limitations of other approaches

Data browsers :

- Often gene-centered
- Often only dual analysis
- Lack of filters
- Limited to a few particular questions
- Lack of synthesis / high level view

Advanced tools / programming languages :

- Often requires bioinformatics skills
- Time-consuming

Primary objectives of TranscriptomeBrowser

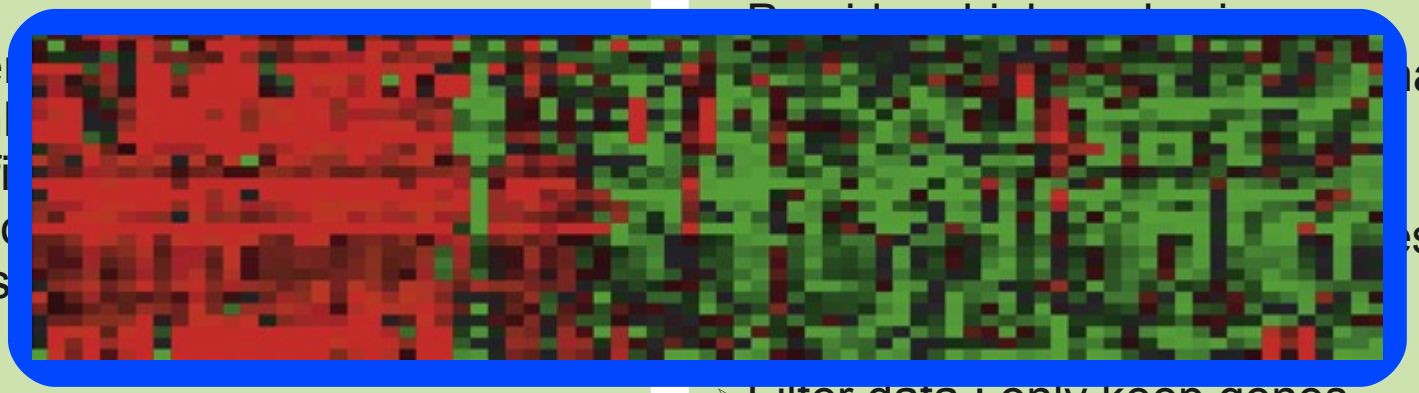
- Provide a high-scale view relying on **transcriptional signatures** and facilitating meta-analysis
- Re-analyse and organise expression data from public databases (GEO)
- Filter data : only keep genes with interesting changes in expression
- **Unsupervised** extraction of clusters of co-expressed genes
- Perform exhaustive annotation of clusters using a large panel of annotation sources.

The TranscriptomeBrowser project : motivations

Limitations of other approaches

Data browsers :

- Often ge
- Often onl
- Lack of fi
- Limited t
- Lack of s



Advanced tools / programming languages :

- Often requires bioinformatics skills
- Time-consuming

Primary objectives of TranscriptomeBrowser

- atures
- ession data
- Filter data : only keep genes with interesting changes in expression
- **Unsupervised** extraction of clusters of co-expressed genes
- Perform exhaustive annotation of clusters using a large panel of annotation sources.

The TranscriptomeBrowser project : motivations

Limitations of other approaches

Data browsers :

- Often gene-centered
- Often only dual analysis
- Lack of filters
- Limited to a few particular questions
- Lack of synthesis / high level view

Advanced tools / programming languages :

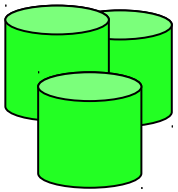
- Often requires bioinformatics skills
- Time-consuming

Primary objectives of TranscriptomeBrowser

- Provide a high-scale view relying on **transcriptional signatures** and facilitating meta-analysis
- Re-analyse and organise expression data from public databases (GEO)
- Filter data : only keep genes with interesting changes in expression
- **Unsupervised** extraction of clusters of co-expressed genes
- Perform exhaustive annotation of clusters using a large panel of annotation sources.

TranscriptomeBrowser pipeline

GEO Repository



GSE2004 GSE1456

Retrievial and preparation of µarray datasets.

More than 8 samples ?

Yes

6544 datasets (260,000 samples)

No

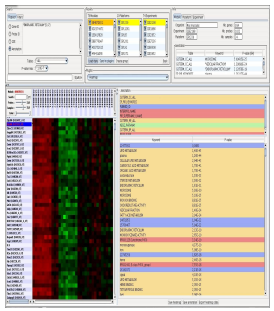
Stop

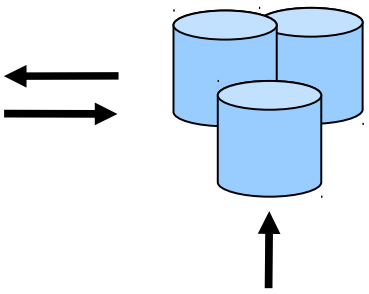
Apply Transcriptional Signature (TS) extraction to each Experiment using our locally developed algorithm (DBF-MCL).

For each experiment a set of transcriptional signatures that contain co-regulated genes are obtained. The number of selected genes and the number of TS obtained depend on individual experiments.

Storage
- Experiment metaData
- Probe metaData
- TS composition
- TS related expression Value

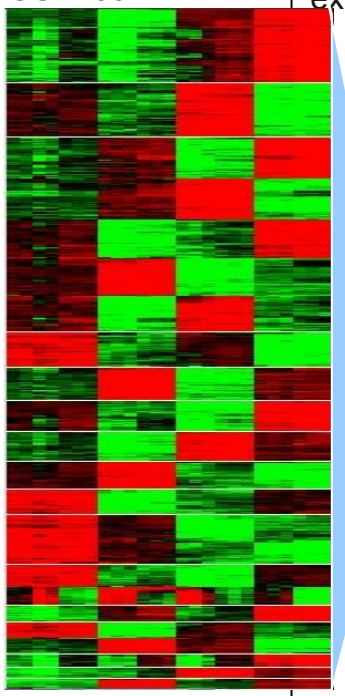
Tbrowser interface

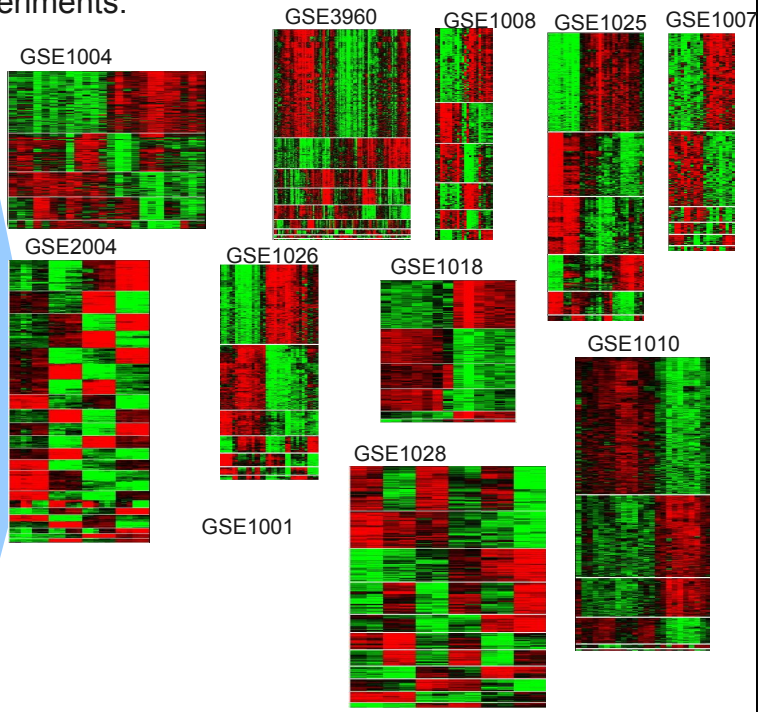




Experiment GSE2004

TS1<-
TS2<-
TS3<-
TS4<-
TS5<-
TS6<-
...





Functional Enrichment analysis
Each Transcriptional Signature is tested for functional Enrichment using terms from the DAVIDknowledgeBase (KEGG, GO, BIOCARTA, INTERPRO,...)

The algorithm behind TranscriptomeBrowser

DBF-MCL

- The **set of genes** in an **experiment** (a set of samples) is seen as a **graph**.
- Goal : to clean a dataset from uninteresting genes and then to extract **clusters** of co-expressed genes from the resulting dataset.
- **DBF** : density-based filtering
- **MCL** : Markov clustering
- Availability : R package on Bioconductor

RTools4TB

Data mining of public microarray data through connections to the TranscriptomeBrowser database.

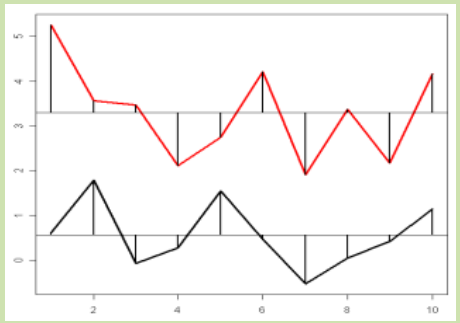
TranscriptomeBrowser (TBrowser) hosts a large collection of transcriptional signatures (TS) automatically extracted from the Gene Expression Omnibus (GEO) database. Each GEO experiment (GSE) was processed so that a subset of the original expression matrix containing the most relevant/informative genes was kept and organized into a set of homogeneous signatures. Each signature was tested for functional enrichment using annotations terms obtained from numerous ontologies or curated databases (Gene Ontology, KEGG, BioCarta, Swiss-Prot, BBID, SMART, NIH Genetic Association DB, COG/KOG...) using the DAVID knowledgebase. The RTools4TB package can be used to perform complexe queries to the database. Thereby, RTools4TB can be helpful (i) to define the biological contexts (i.e, experiments) in which a set of genes are co-expressed and (ii) to define their most frequent neighbors. In addition, RTools4TB comes with a new algoritm, "Density Based Filtering And Markov Clustering" (DBF-MCL), whose goal is to partition large and noisy datasets. DBF-MCL is a tree-step adaptative algorithm that (i) find elements located in dense areas (ie. clusters) (ii) uses selected items to construct a graph and (iii) performs graph partitioning using MCL. This algorithm is implemented in the RTools4TB package although it requires a UNIX-like systems.

Author Aurelie Bergon, Fabrice Lopez, Julien Textoris, Samuel Granjeaud and Denis Puthier
Maintainer Aurelie Bergon

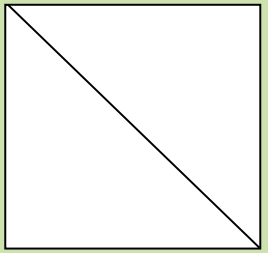
http://www.bioconductor.org/packages/2.6/bioc/html/RTools4TB.html

DBF-MCL step 1 : selecting informative genes

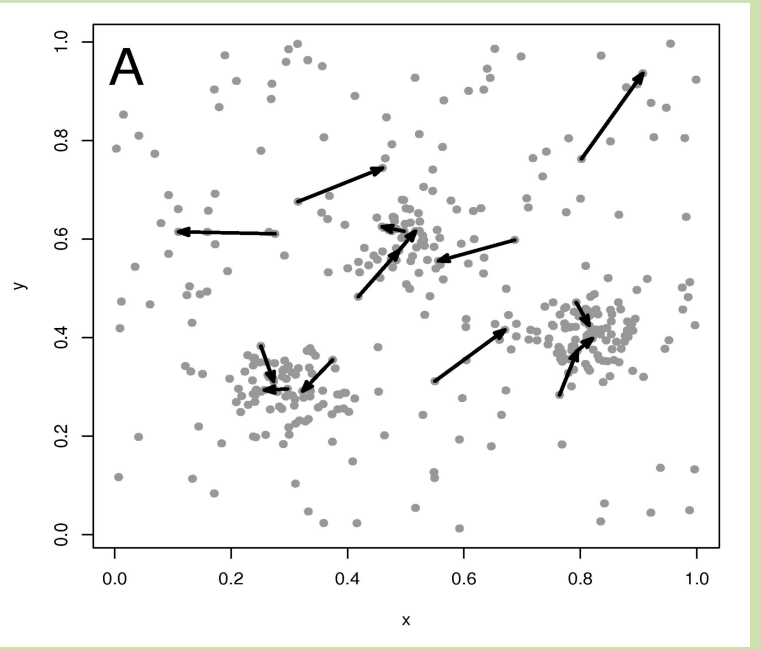
Gene expression profiles



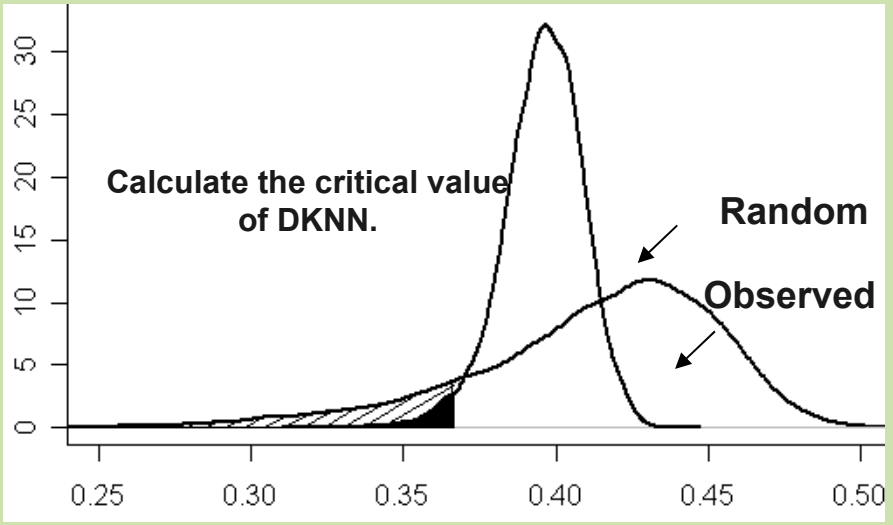
Gene-gene distance matrix (n x n)



Calculate for gene *1..n* the distances with their kth nearest neighbors (DKNN)



Calculate simulated DKNN values. Calculate FDR for each observed DKNN value



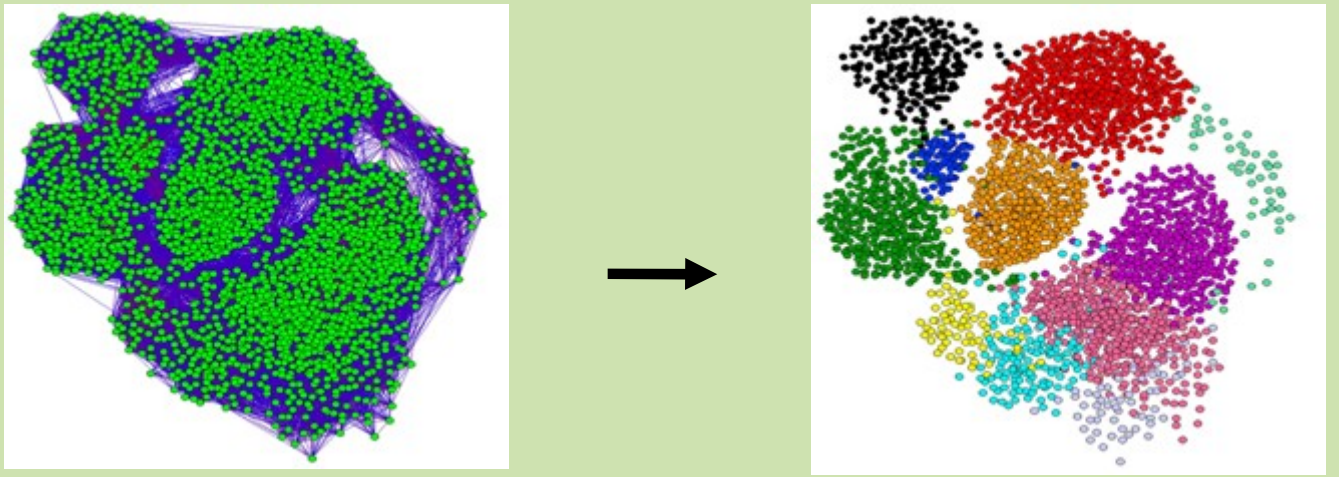
DBF-MCL steps 2 & 3 : filtering and partitionning

Graph construction :

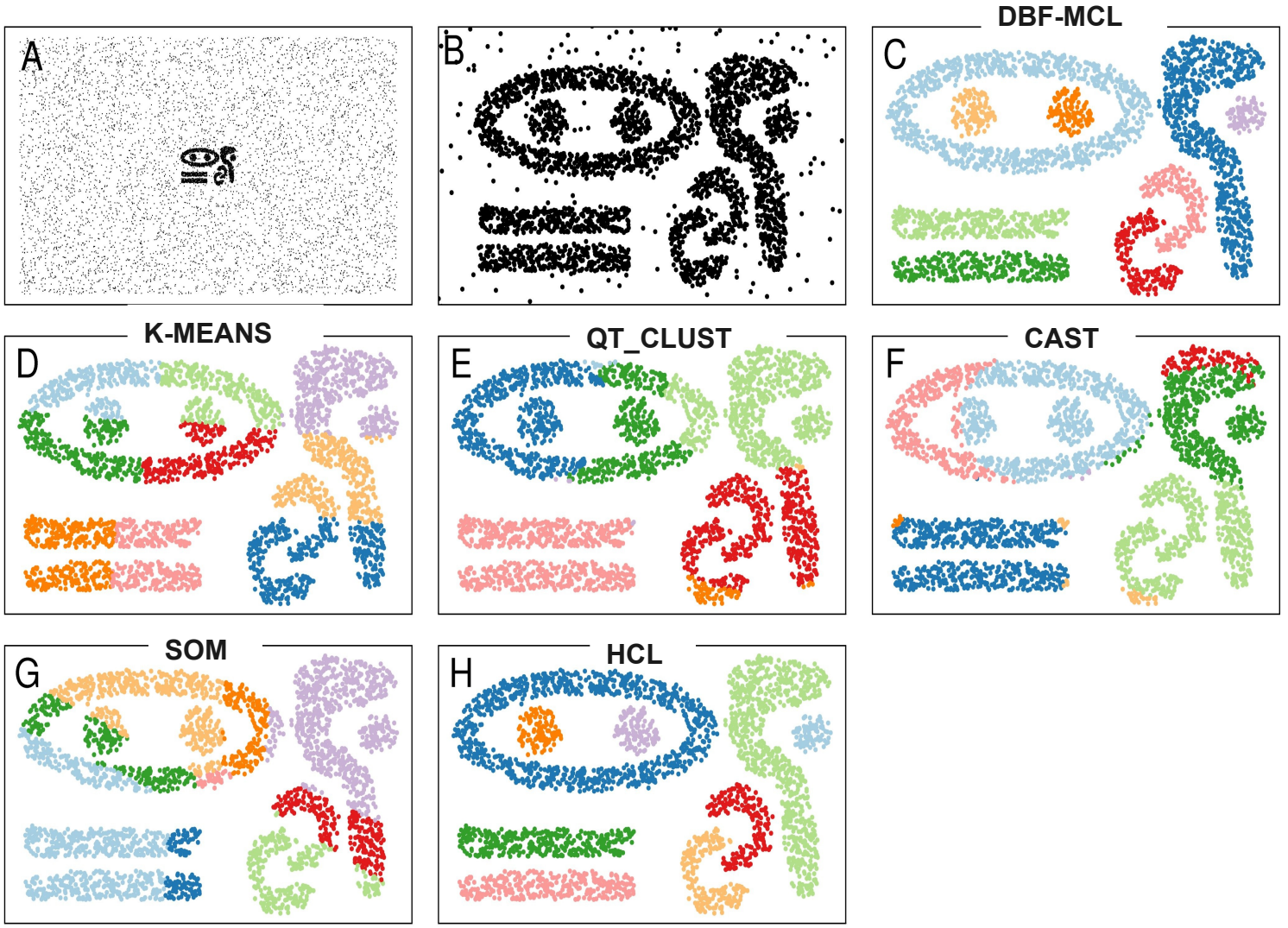
- Only genes that belong to dense regions (DKNN < cut-off) are conserved for analysis
- Nodes = genes
- An edge exists between two genes if one of them belongs to the k-nearest neighbors of the other.

Graph partitionning:

- Markov clustering (MCL, Stijn van Dongen)



DBF-MCL performance on a test dataset



DBF-MCL performance on a real dataset

GSE1456
Public on May 31, 2006

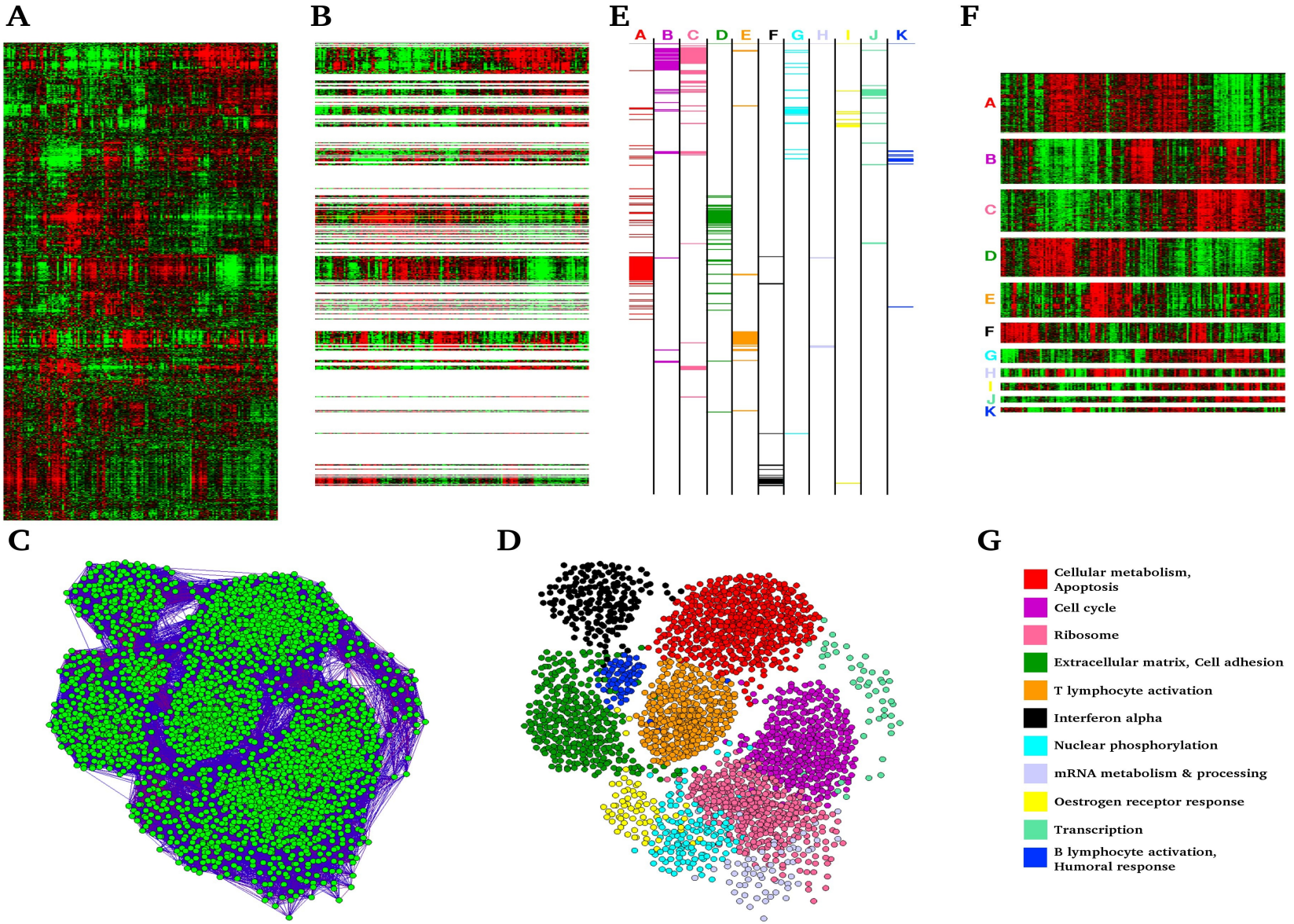
Title
Gene expression of breast cancer tissue in a large population-based cohort of Swedish patients

Platform
GPL96: Affymetrix GeneChip Human Genome U133 Array Set HG-U133A

Type
Breast cancer, expression profiling, predictive gene signature, molecular classification of cancer

Summary
Tissue material was collected from all breast cancer patients receiving surgery at Karolinska Hospital from 1994-1996 (n=159 tumors)

DBF-MCL performance on a real dataset



TBrowserDB

TranscriptomeBrowser hosts a mySQL relational database containing transcriptional signatures derived from:

- 101 platforms
- 54 species
- 6 544 experiments
- 244 692 samples

→ **40 151 Transcriptional Signatures**

→ 21 210 053 gene expression profiles stored

→ 508 465 296 expression values

Programmatic access to the database available through a SOAP **webservice :**

http://tagc.univ-mrs.fr/services/TBService/TBService.wsdl

Annotations

501528 terms derived from 50 sources of gene annotation

DISEASE
OMIM_PHENOTYPE
OMIM_ID
DISEASE
PHENOTYPE
GENETIC_ASSOCIATION_DB

HIV
HIV_INTERACTION
HIV_INTERACTION_CATEGORY

MICROARRAY
GENESIGDB
MSIGDB

GENOMIC LOCATION
CHROMOSOME
CYTOBAND

LITERATURE
PUBMED_ID
HIV_INTERACTION_PUBMED

MOTIFS
PICTAR_4WAYS
PICTAR_CHICKEN
PICTAR_DOG
TARGETSCAN
TFBS_CONSERVED
PICTAR_5WAYS
TARGETSCAN_WORM
TARGETSCAN_DROSO
TARGETSCAN_ALL

GENE ONTOLOGY
GOTERM_MF_ALL
GOTERM_CC_ALL
GOTERM_BP_ALL

PATHWAYS
KEGG_REACTION
NCICB_CAPATHWAY
KEGG_COMPOUND
PANTHER_FAMILY
REACTOME
BBID
PANTHER_PATHWAY
BIOCARTA
PANTHER_SUBFAMILY
PANTHER_TERM_BP
KEGG_PATHWAY
PANTHER_TERM_MF
WIKIPATHWAY

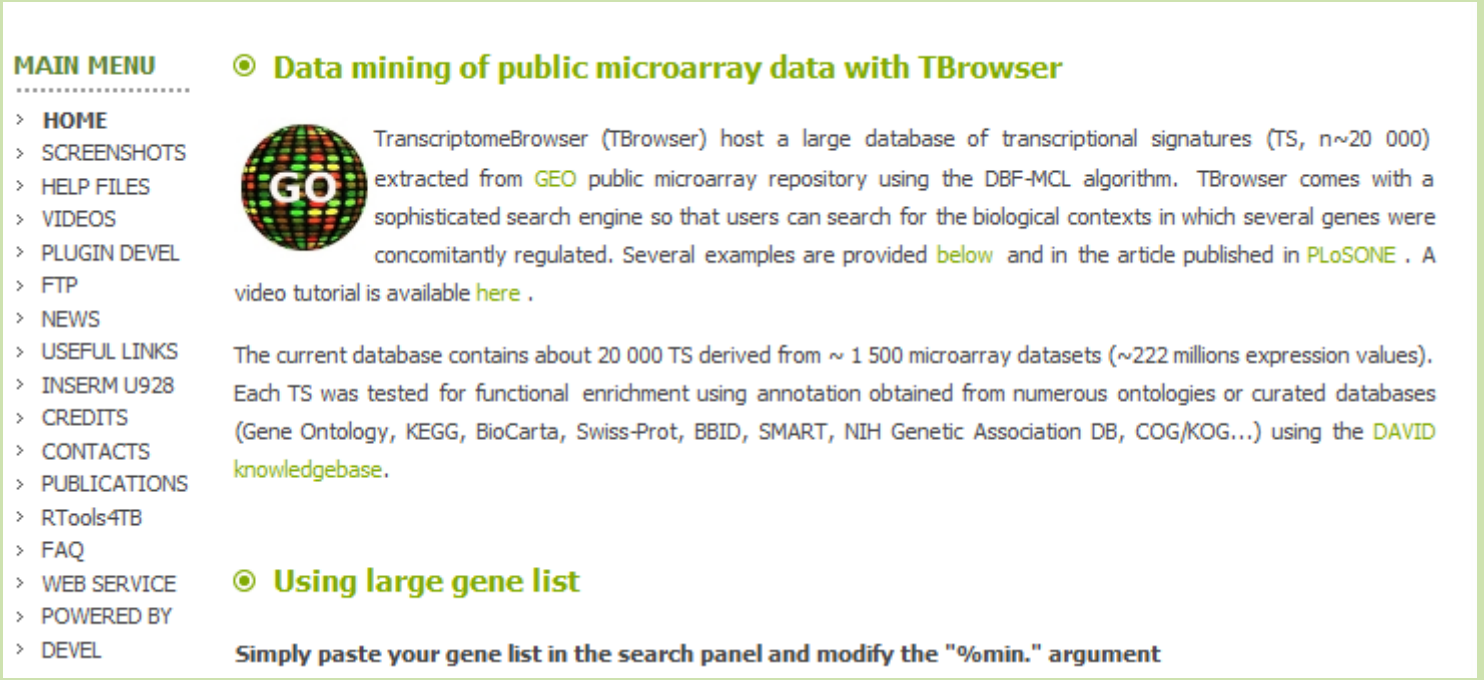
PROTEIN DOMAINS
SMART_NAME
COG_KOG_NAME
TIGRFAMS_NAME
PRODOM_NAME
SCOP_ID
PROSITE_NAME
COG_KOG_ONTOLOGY
PIR_SUPERFAMILY_NAME
PFAM_NAME
KEA
SP_PIR_KEYWORDS

→ Systematic annotation enrichment analysis of transcriptional signatures

→ Search with annotation keywords

Availability

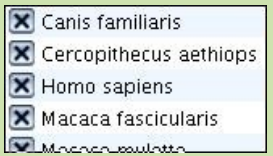
http://tagc.univ-mrs.fr/tbrowser/



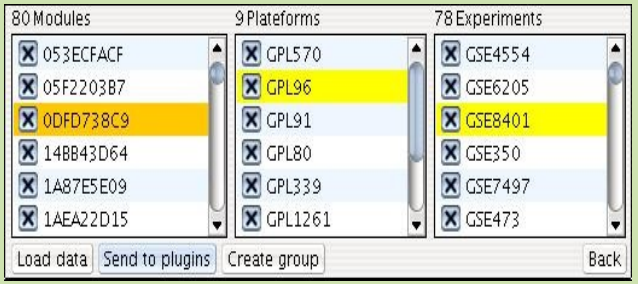
TranscriptomeBrowser: A Powerful and Flexible Toolbox to Explore Productively the Transcriptional Landscape of the Gene Expression Omnibus Database. F. Lopez *et al. PloS ONE. 2008*

TBrowser interface

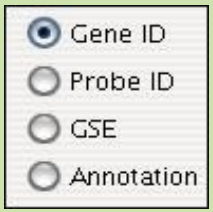
Filters: TS size, Species

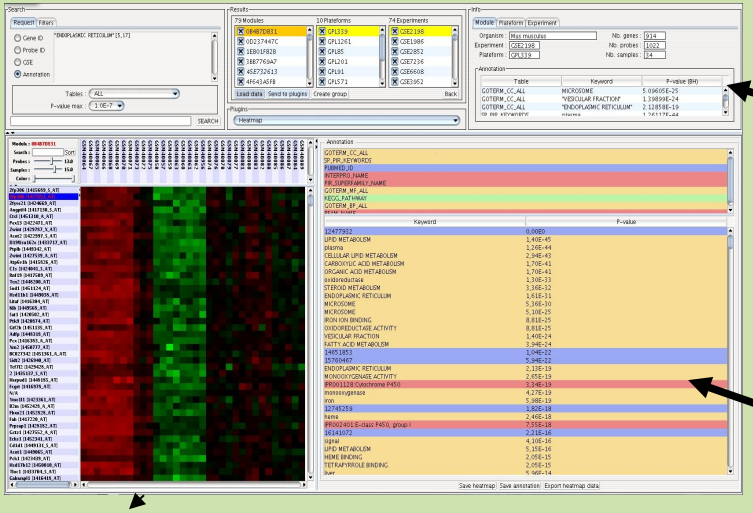


Current TS

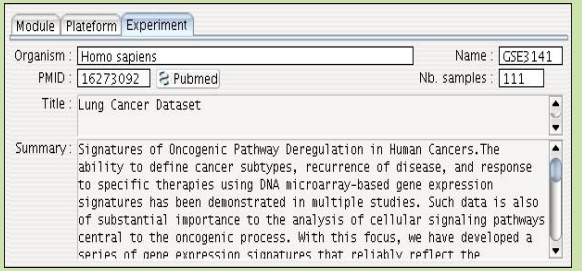


Query :
Gene symbol,
Probe name,
Experiment,
Functionnal Annotation

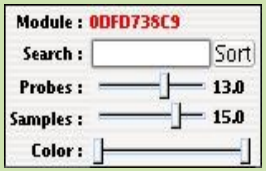




Informations: TS, experiments, plateform



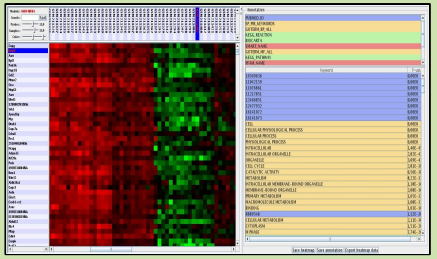
Heatmap parameters



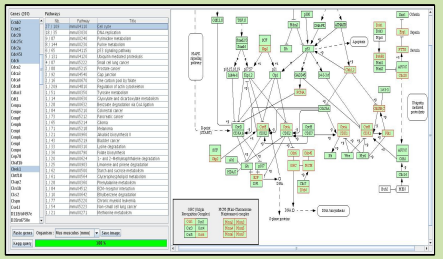
Functional enrichment

Currently available plugins

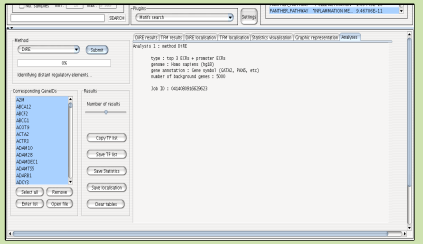
Heatmap



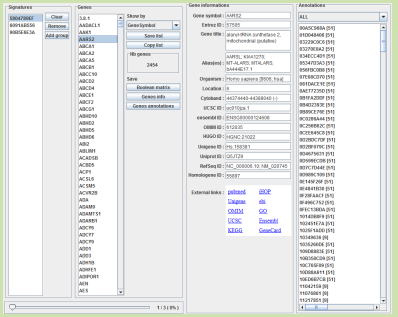
KEGG search



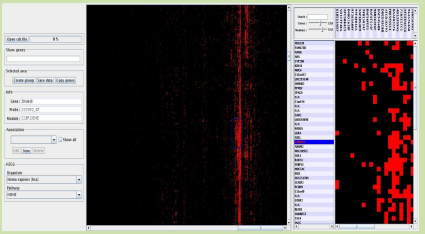
TBMotifSearch



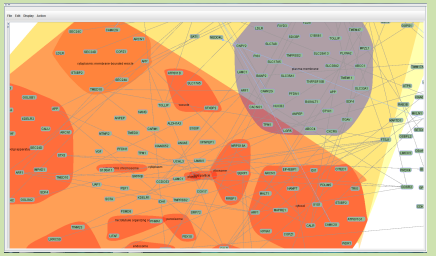
TBneighborhood



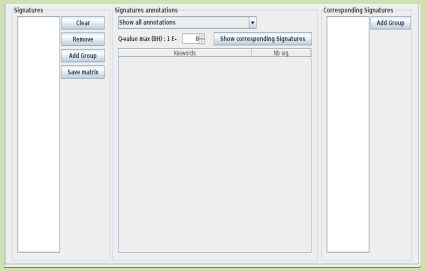
TBMap



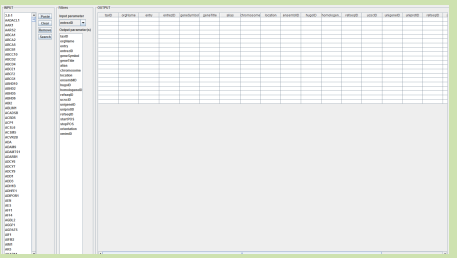
InteractomeBrowser



AnnotationOverview



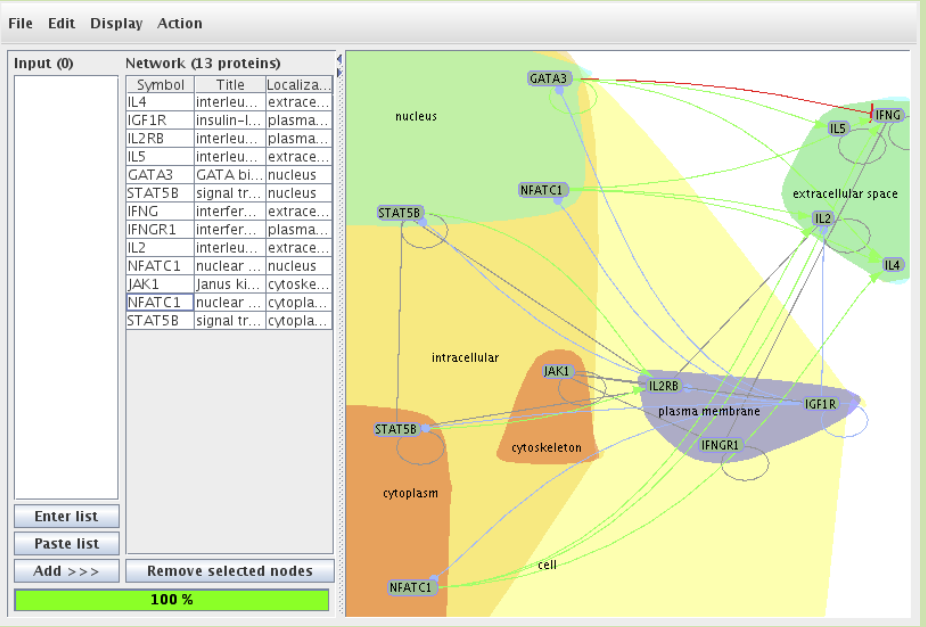
TBConvertor



Integration of TS data with interactome data

Plugin InteractomeBrowser

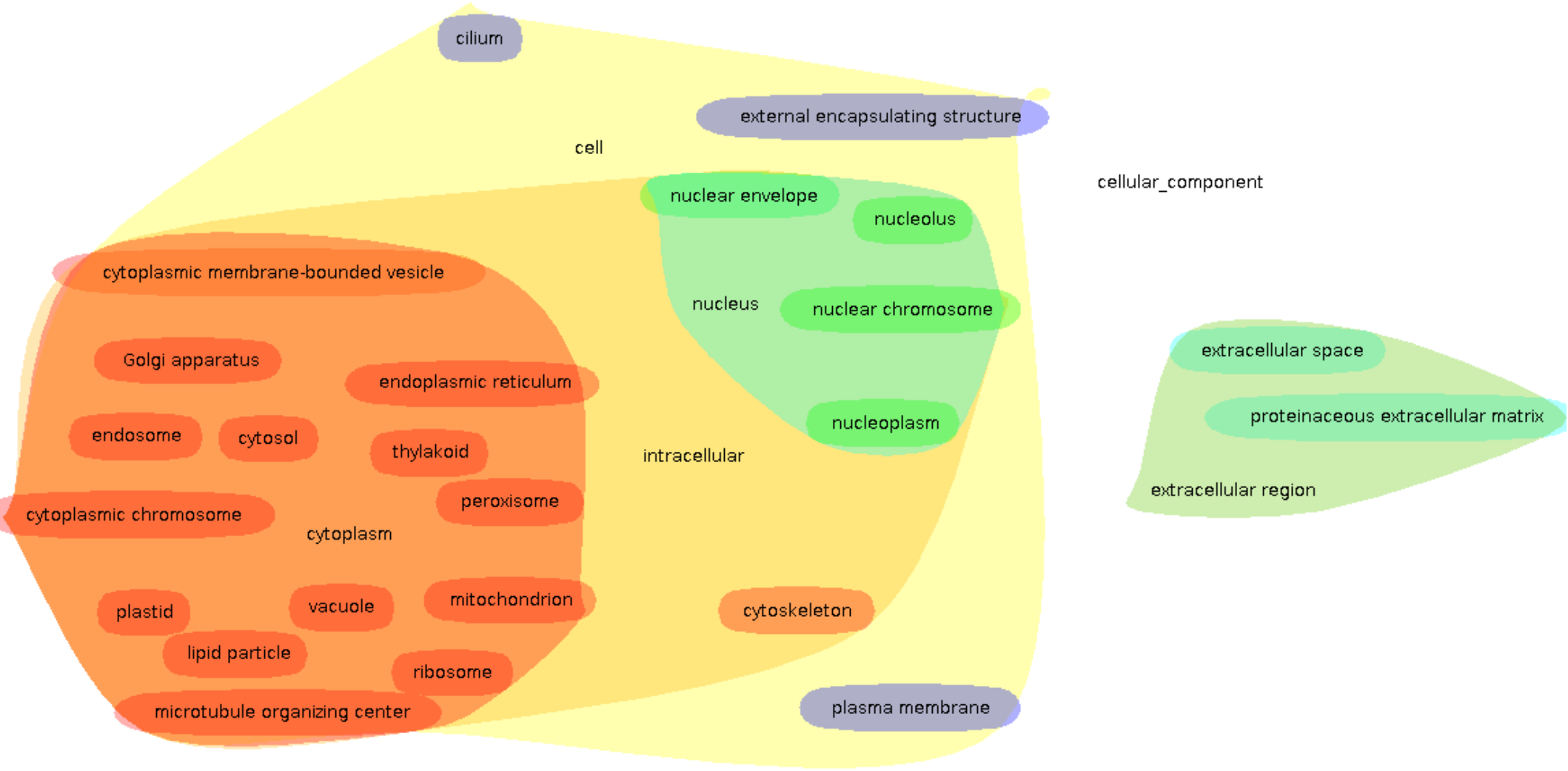
- Visualization of interactions of various nature, mined from several databases :
  - Protein-protein physical interactions : *HPRD*, *Intact*
  - Proven regulations of genes by transcription factors (Tfs) : *OregAnno, LymphTF-DB*
  - Kinase-substrate relationships : *KEA*
  - Potential binding of a TF in the promoter of a gene : *cisRED, TFBSconserved*



- Tools for network building and editing
- Cell-compartment -based layout
- "On-the-fly" enrichment analysis
- Cytoscape export
- Recently used in :

Sex-Related Differences in Gene Expression Following Coxiella burnetii Infection in Mice: Potential Role of Circadian Rhythm, Textoris J, Ban LH, Capo C, Raoult D, Leone M, et al., 2010, PLoS ONE

InteractomeBrowser : cell compartments -based layout



Uses of TranscriptomeBrowser

- Characterize the function of a gene (thanks to annotations of signatures in which this gene is found)
- Find potential partners of a gene (frequently co-expressed genes)
- Find biological contexts in which several genes are co-expressed
- Find genes associated with a combination of biological annotation terms (ex : Cancer + regulation of transcription)

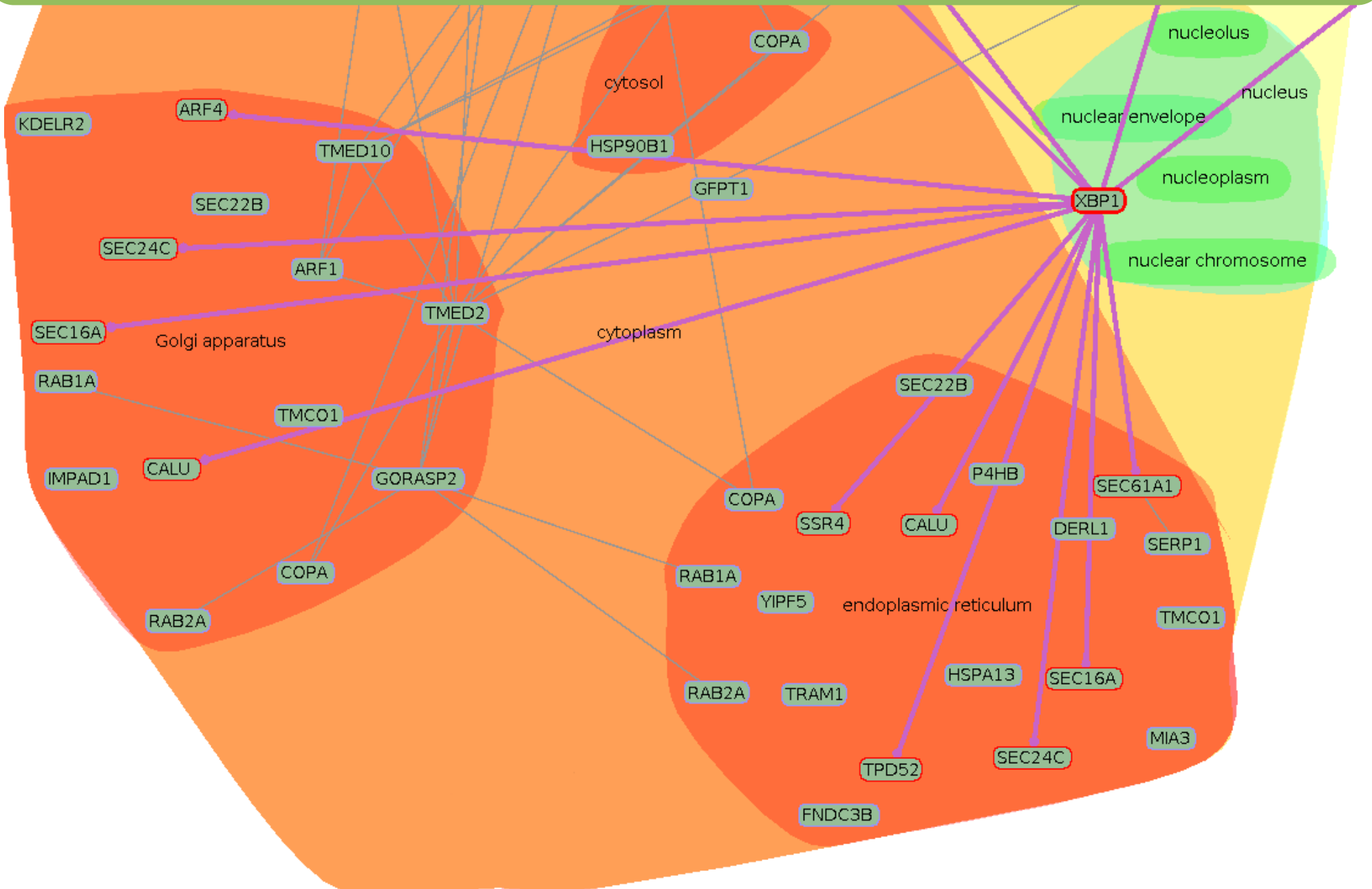
Uses of TranscriptomeBrowser

- Characterize the function of a gene (thanks to annotations of signatures in which this gene is found)
- Find potential partners of a gene (frequently co-expressed genes)
- Find biological contexts in which several genes are co-expressed
- Find genes associated with a combination of biological annotation terms (ex : Cancer + regulation of transcription)

- Example : search for targets of XBP1.
  - Known : XBP1 is activated upon accumulation of unfolded proteins in the endoplasmic reticulum (ER stress)
  - Search for signatures enriched in the term "V$XBP1" from TFBS (p-value < 10^{-10})

  → Result : 1 signature, highly enriched in genes involved in "protein folding" and "protein transport", and also in genes of the endoplasmic reticulum (GO terms)

Example : XBP1 targets Visualisation with InteractomeBrowser



Conclusions

- A extensible and unified data mining suite
- Meta-analysis of thousands of microarray samples
- A synthetic view of the current knowledge on transcriptome
- Data integration helps making new hypothesis on gene functions and expression networks

Perspectives

- Allow the import of user-provided :
    - microarray data
    - annotations
- Integrate more data : Chip-seq data, RNA seq data, ...
- Develop new plugins for the analysis of TS
- Integration with GINsim

TBrowser Team and collaborations

TAGC, Marseille

- Denis Puthier
- Aurélie Bergon
- Fabrice Lopez
- Samuel Granjeau
- Julien Textoris
- Jean Imbert

ENS, Paris

- Denis Thieffry

IML, Marseille

- Elisabeth Remy
- Gilles Didier

CRG, Barcelona

- Thomas Graf
- Eric Kallin