EXPression ANalyzer and DisplayER

Tom Hait
Aviv Steiner
Igor Ulitsky
Chaim Linhart
Amos Tanay
Seagull Shavit
Rani Elkon

Adi Maron-Katz
Dorit Sagir
Eyal David
Roded Sharan
Israel Steinfeld
Yossi Shiloh
Ron Shamir

Ron Shamir's Computational Genomics Group

Rani Elkon's Group

Schedule

- Data, preprocessing, grouping (14:15-14:30)
- Hands-on part I (14:30-14:40)
- Grouping analysis (14:40-14:55)
- Hands-on part II (14:55-15:10)
- Enrichment analysis (15:10-15:20)
- Hands-on part III (15:20-15:30)
- ChIP-seq and GSEA (15:30-15:45)
- Hands-on part IV (15:45-16:00)

- **EXPANDER – an integrative package for analysis of gene expression and NGS data**

- **Built-in support for 18 organisms:**

  human, mouse, rat, chicken, fly, zebrafish, C.elegans, yeast (s. cereviciae and s. pombe), arabidopsis, tomato, listeria, leishmania, E. coli (two strains), aspargillus, rice. And v.vinifera (grape)

- Demonstration on human CAL51 cell line experiments*:
  - RNA-Seq data, which contains expression profiles measured in several time points after IR-induction.
  - P53 ChIP-Seq data after 2 hours of IR-induction.

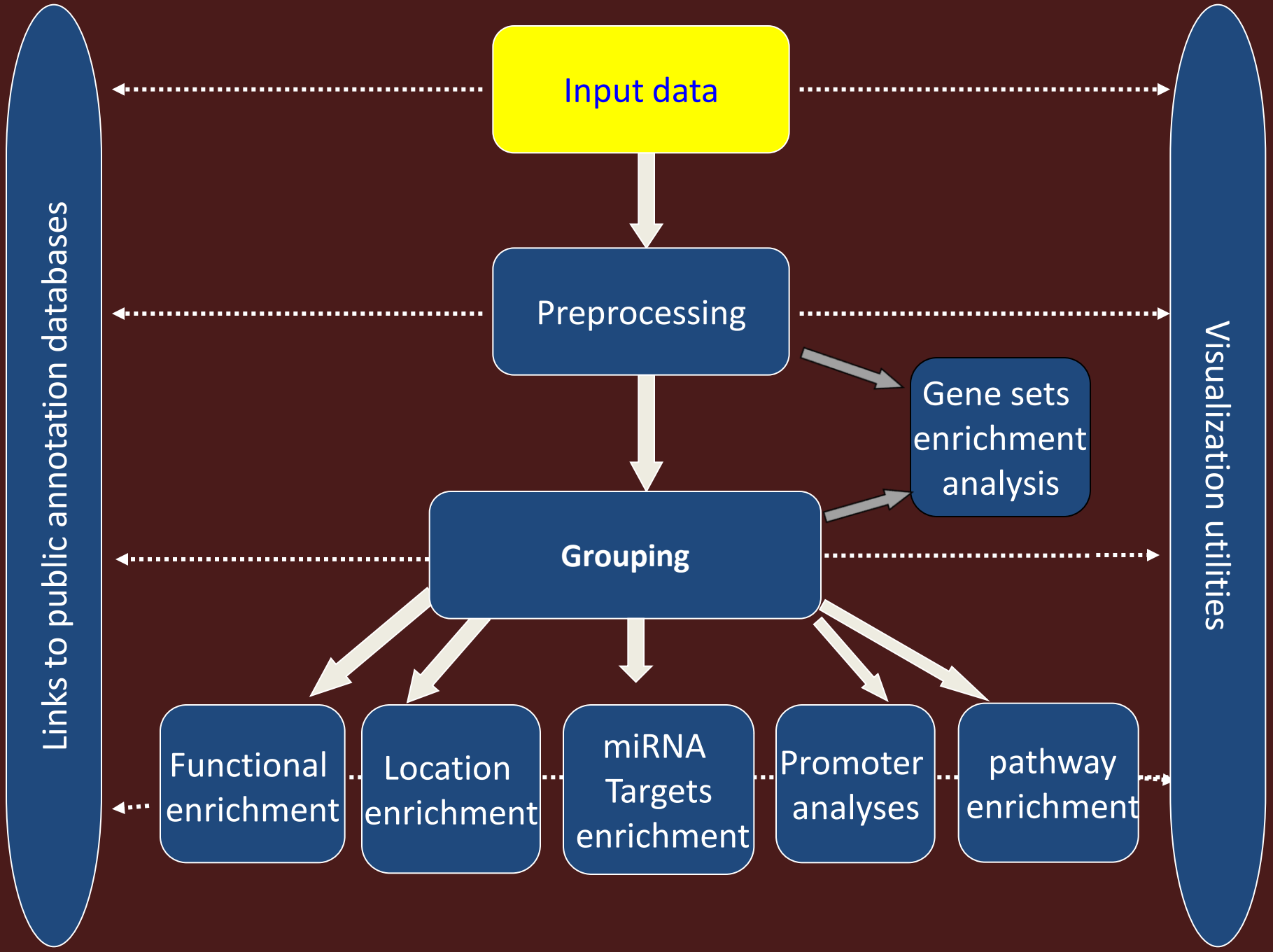
*Data from Rashi-Elkeles, Warnatz and Elkon et al, 2014, Science Signaling, DOI: 10.1126/scisignal.2005032

EXPANDER status

- 829 citations since 2003
- 63 citations since 2017
- 18,062 downloads since 2003
- 914 since 2017

What can it do?

- **Low level analysis**
  - Data adjustments (missing values, merging, divide by base, log)
  - Normalization
  - Probes & condition filtering
- **High level analysis**
  - Group detection (supervised clustering, differential expression, clustering, bi-clustering, network based grouping).
  - Ascribing biological meaning to patterns via enrichment analysis


Links to public annotation databases
Input data
Preprocessing
Gene sets enrichment analysis
Grouping
Functional enrichment
Location enrichment
miRNA Targets enrichment
Promoter analyses
pathway enrichment
Visualization utilities

EXPANDER – Data

- Expression matrix (probe-row; condition-column)
  - One-channel data (e.g., Affymetrix)
  - Dual-channel data, in which data is log R/G (e.g. cDNA microarrays)
  - '.cel' files
  - RNA-Seq counts OR absolute/relative intensities data
- ChIP-Seq data: **in BED or GFF3 formats**
- ID conversion file: maps probes to genes
- Gene groups data: defines gene groups
- Gene ranks data: defines gene ranking for GSEA
- Network information (e.g. PPI network) - .sif format

First steps with the data – load, define, preprocess

- **Load dialog box , “Data menu”, “Preprocessing menu”**
- Data definitions
  - Defining condition subsets
  - Data type & scale (log)
  - Define genes of interest
- Data Adjustments
  - Missing value estimation (KNN or arbitrary)
  - Flooring
  - Condition reordering
  - Merging conditions
  - Merging probes by gene IDs
  - Assigning genes to ChIP-Seq peaks
  - Divide by base
  - Log data (base 2)

Data preprocessing

- **Normalization=** removal of systematic biases
  - **Quantile** = equalizes distributions
  - **Lowess** (locally weighted scatter plot smoothing) = a non linear regression to a base array
- Visualizations to inspect normalization:
  - box plots
  - Scatter plots (simple and M vs. A)

$M=\log_2(A1/A2)$

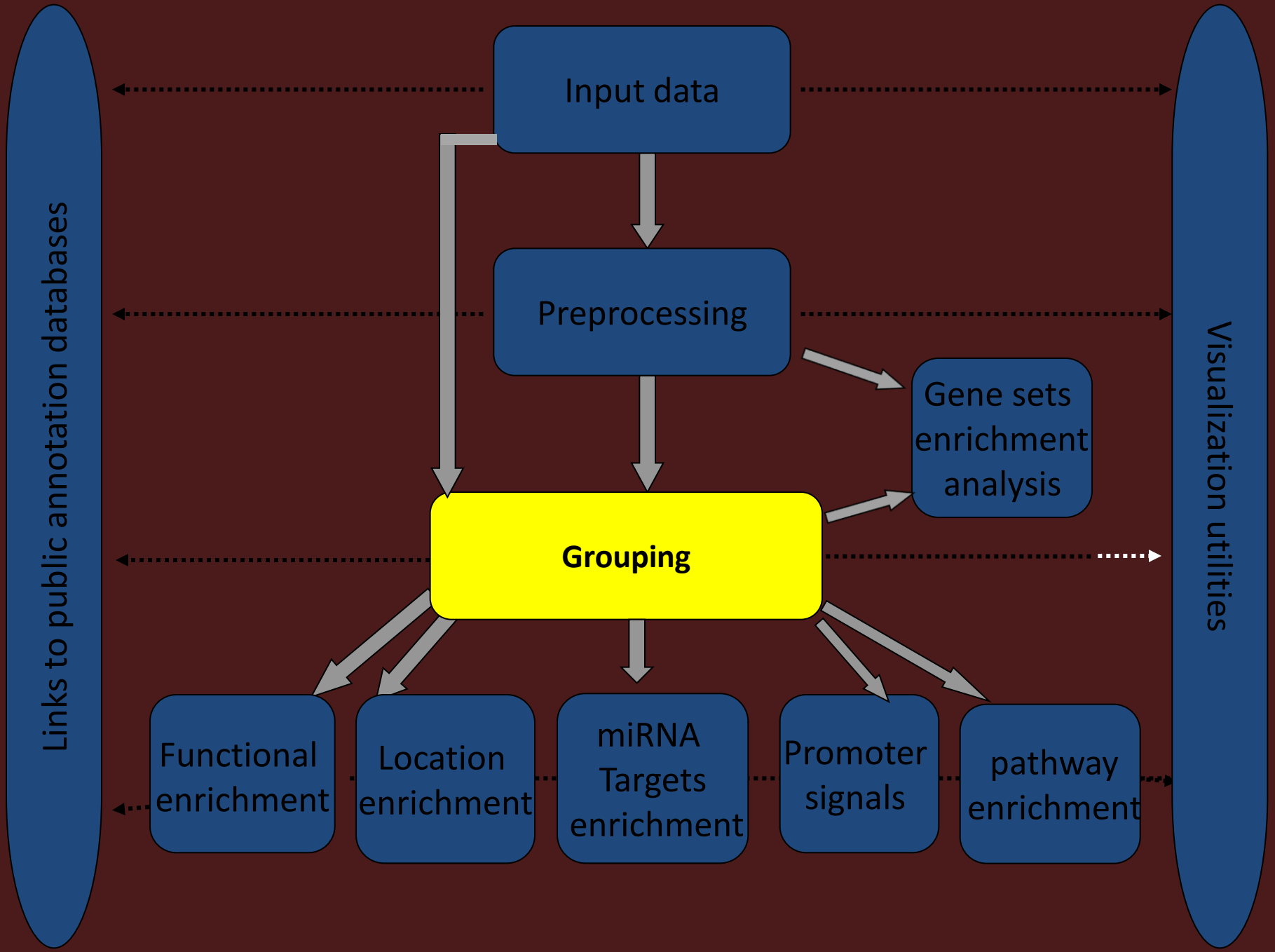
$A = 0.5*\log_2(A1*A2)$

Data preprocessing

- Probe filtering

  Focus downstream analysis on the set of “responding genes”
  - Fold-Change
  - Variation
  - Statistical tests: T-test, SAM (Significance Analysis of Microarrays), edgeR and deseq2 (RNAseq count data)
  - It is possible to define “VIP genes”.
- Standardization : Mean=0, STD=1 (visualization)
- Condition filtering
- Order of operations

Hands-on (1-2)


Links to public annotation databases
Input data
Preprocessing
Gene sets enrichment analysis
Grouping
Functional enrichment
Location enrichment
miRNA Targets enrichment
Promoter signals
pathway enrichment
Visualization utilities

Supervised Grouping

- Differential expression:
  a) Under normality assumption: t-test, SAM
  b) No normality assumption (RNA Seq data): Wilcoxon rank sum test , Negative binomial (edgeR/DESeq2)

- Similarity group (correlation to a selected probe/gene)

- Rule based grouping (define a pattern)

Unsupervised grouping - cluster Analysis

Partition into distinct groups, each with a particular expression pattern

- *co-expression → co-function*
- *co-expression → co-regulation*

Partition the genes attempts to maximize:

- **Homogeneity** within clusters
- **Separation** between clusters

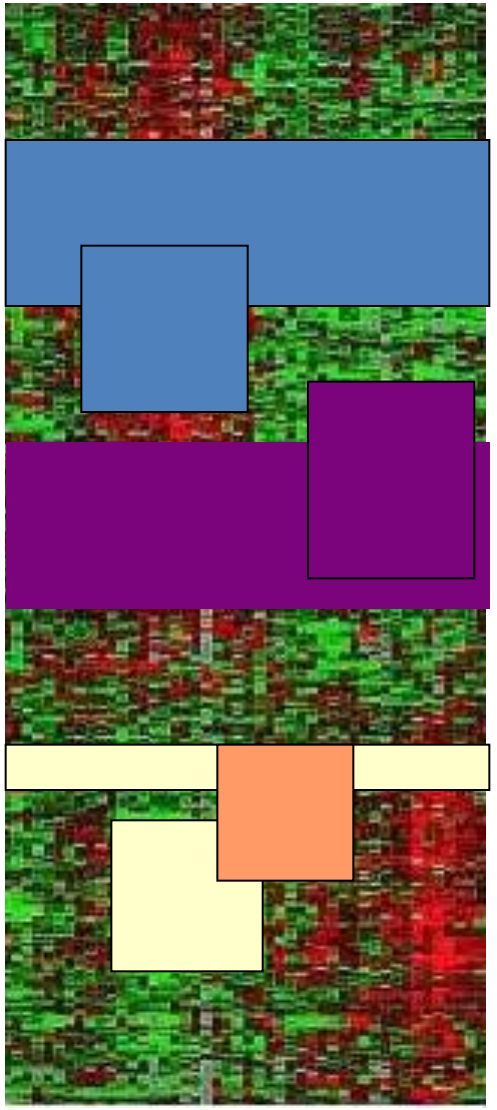
Cluster Analysis within Expander

- Implemented algorithms:
  - CLICK, K-means, SOM, Hierarchical
- Visualization:
  - Mean expression patterns
  - Heat-maps
  - Chromosomal positions
  - Network sub-graph **(Cytoscape integration)**
  - PCA
  - Clustered heat map

Biclustering

Clustering seeks global partition according to similarity across **ALL** conditions >> becomes too restrictive on large datasets.

- Relevant knowledge can be revealed by identifying genes with common pattern across a subset of the conditions
- Novel algorithmic approach is needed: *Biclustering*



Biclustering II

* *Bicluster* = subset of genes with similar behavior under a subset of conditions

  Computationally challenging: has to consider many combinations

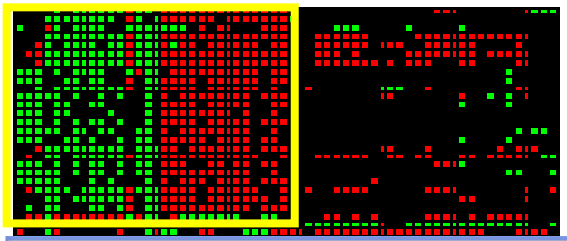
  Biclustering methods in EXPANDER:
  ISA (Iterative Signature Algorithm) - Ihmels et.al Nat Genet 2002

  SAMBA = Statistical Algorithmic Method for Bicluster Analysis (A. Tanay, R. Sharan, R. Shamir *RECOMB 02)*

Drawbacks/ limitations:

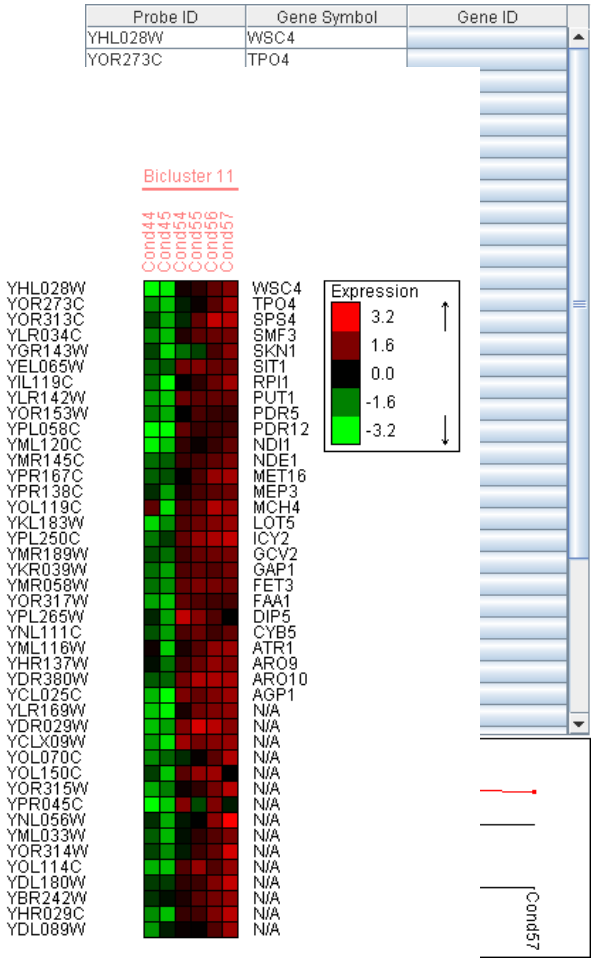
- Useful only for over 20 conditions
- Parameters
- How to asses the quality of Bi-clusters

Biclustering Visualization

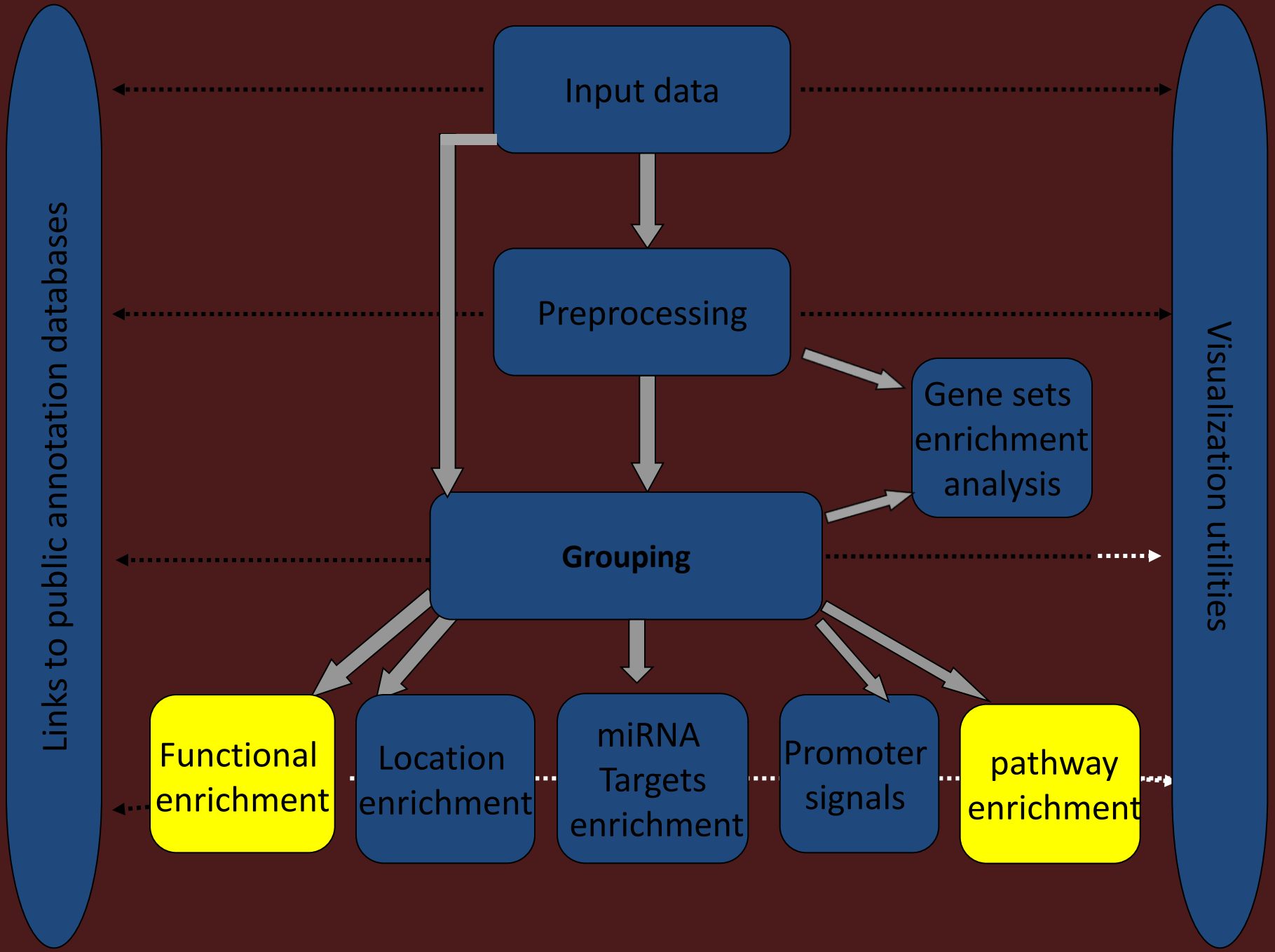


File Options

Bic num	Bic Score	# Conditions	
1	230.475	7	37
2	495.917	7	76
3	433.011	6	76
4	248.93	7	38
5	461.604	7	68
6	177.526	7	22
7	116.431	5	34
8	374.292	9	27
9	331.779	10	46
10	320.373	6	55
11	417.158	7	75
12	286.944	6	55
13	144.321	12	13
14	201.665	10	23
15	435.368	7	68
16	680.887	8	81
17	148.6	10	12
18	200.601	14	19
19	206.593	6	44
20	372.519	10	53
21	575.098	5	107
22	387.311	6	71
23	213.658	4	56
24	196.952	14	15
25	458.704	12	45
26	522.565	12	42
27	179.189	7	27



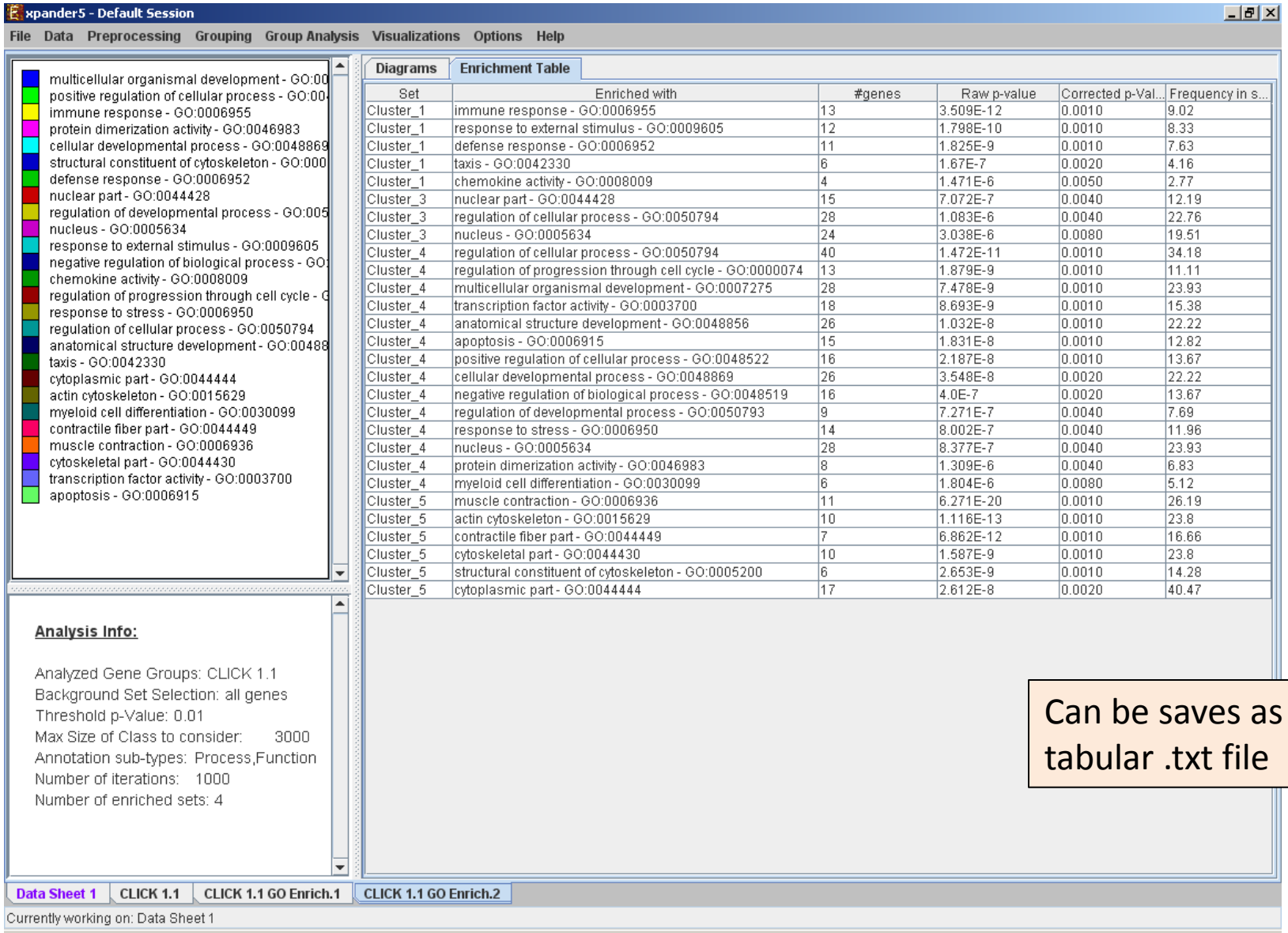
Hands-on (3-4)


Links to public annotation databases
Input data
Preprocessing
Gene sets enrichment analysis
Grouping
Functional enrichment
Location enrichment
miRNA Targets enrichment
Promoter signals
pathway enrichment
Visualization utilities

Functional enrichment analysis - Ascribing functional meaning to gene groups

- Gene Ontology (GO) annotations for all supported organisms
- TANGO: Apply statistical tests that seek over-represented GO functional categories in the groups

Functional Enrichment - Visualization



Set	Enriched with	#genes	Raw p-value	Corrected p-Val...	Frequency in s...
Cluster_1	immune response - GO:0006955	13	3.509E-12	0.0010	9.02
Cluster_1	response to external stimulus - GO:0009605	12	1.798E-10	0.0010	8.33
Cluster_1	defense response - GO:0006952	11	1.825E-9	0.0010	7.63
Cluster_1	taxis - GO:0042330	6	1.67E-7	0.0020	4.16
Cluster_1	chemokine activity - GO:0008009	4	1.471E-6	0.0050	2.77
Cluster_3	nuclear part - GO:0044428	15	7.072E-7	0.0040	12.19
Cluster_3	regulation of cellular process - GO:0050794	28	1.083E-6	0.0040	22.76
Cluster_3	nucleus - GO:0005634	24	3.038E-6	0.0080	19.51
Cluster_4	regulation of cellular process - GO:0050794	40	1.472E-11	0.0010	34.18
Cluster_4	regulation of progression through cell cycle - GO:0000074	13	1.879E-9	0.0010	11.11
Cluster_4	multicellular organismal development - GO:0007275	28	7.478E-9	0.0010	23.93
Cluster_4	transcription factor activity - GO:0003700	18	8.693E-9	0.0010	15.38
Cluster_4	anatomical structure development - GO:0048856	26	1.032E-8	0.0010	22.22
Cluster_4	apoptosis - GO:0006915	15	1.831E-8	0.0010	12.82
Cluster_4	positive regulation of cellular process - GO:0048522	16	2.187E-8	0.0010	13.67
Cluster_4	cellular developmental process - GO:0048869	26	3.548E-8	0.0020	22.22
Cluster_4	negative regulation of biological process - GO:0048519	16	4.0E-7	0.0020	13.67
Cluster_4	regulation of developmental process - GO:0050793	9	7.271E-7	0.0040	7.69
Cluster_4	response to stress - GO:0006950	14	8.002E-7	0.0040	11.96
Cluster_4	nucleus - GO:0005634	28	8.377E-7	0.0040	23.93
Cluster_4	protein dimerization activity - GO:0046983	8	1.309E-6	0.0040	6.83
Cluster_4	myeloid cell differentiation - GO:0030099	6	1.804E-6	0.0080	5.12
Cluster_5	muscle contraction - GO:0006936	11	6.271E-20	0.0010	26.19
Cluster_5	actin cytoskeleton - GO:0015629	10	1.116E-13	0.0010	23.8
Cluster_5	contractile fiber part - GO:0044449	7	6.862E-12	0.0010	16.66
Cluster_5	cytoskeletal part - GO:0044430	10	1.587E-9	0.0010	23.8
Cluster_5	structural constituent of cytoskeleton - GO:0005200	6	2.653E-9	0.0010	14.28
Cluster_5	cytoplasmic part - GO:0044444	17	2.612E-8	0.0020	40.47

Can be saves as a tabular .txt file

Pathway analysis

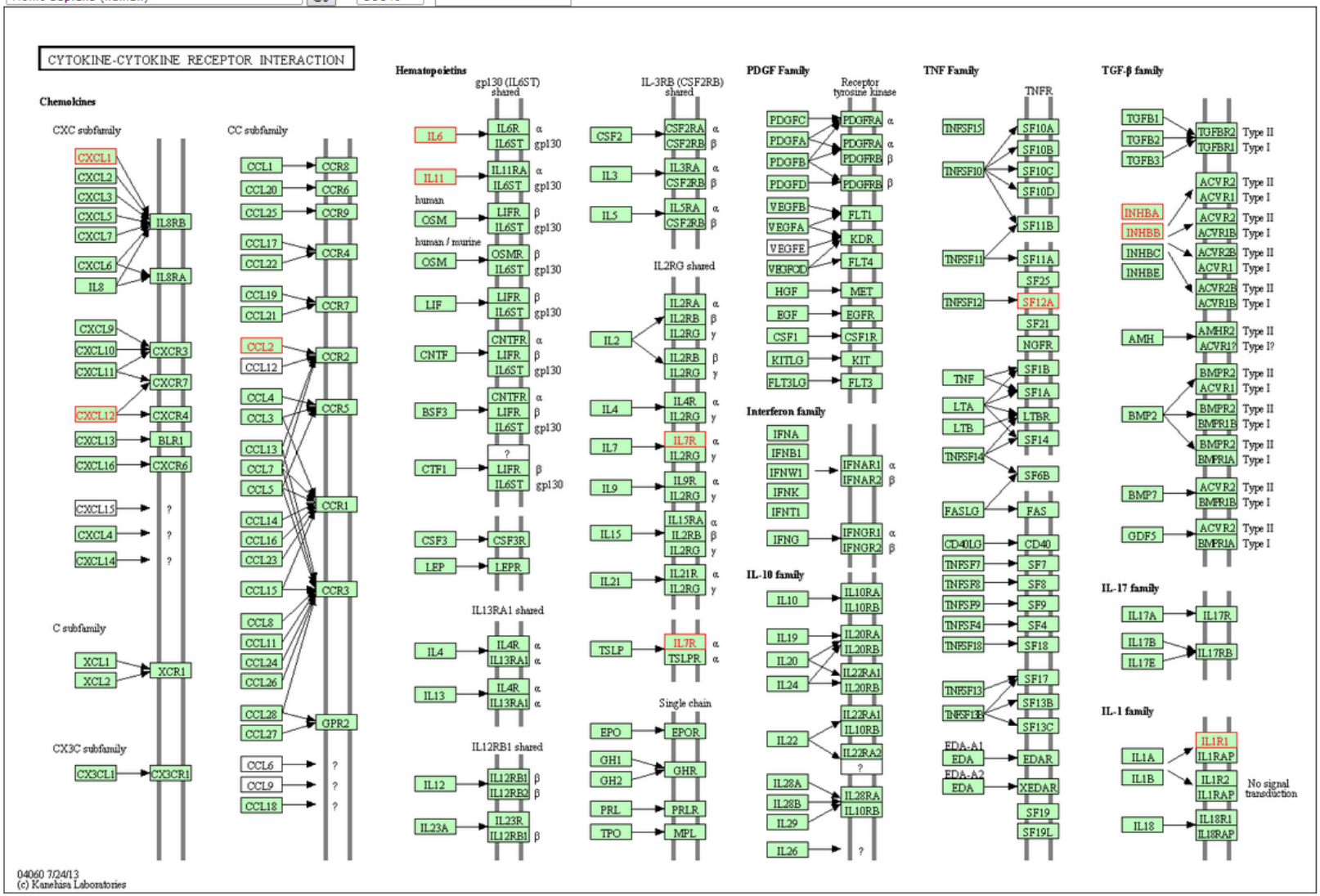
- Searches for biological pathways that are over-represented in gene groups
- **KEGG: Kyoto Encyclopedia of Genes and Genomes** (mainly metabolic),all 18 orgs
- **WikiPathways** – various biological pathways(~20 species, 1765 pathways) – open resource
- Statistical hyper-geometric (HG) cumulative distribution score + multiple testing correction

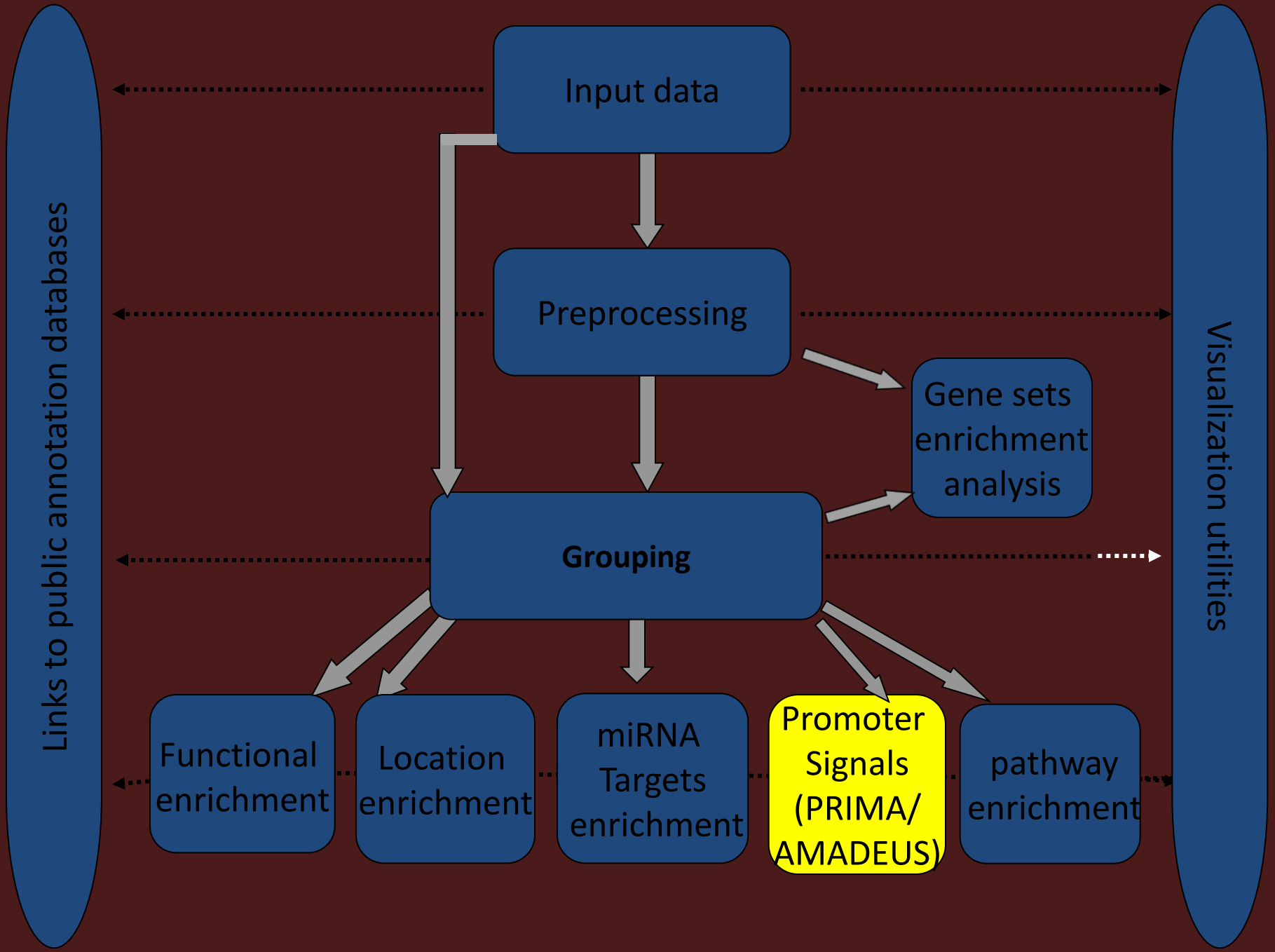
www.genome.jp/kegg-bin/show_pathway?hsa04060+6347+6387+3625+3624+2919+3575+51330+3589+3554+3569

KEGG **Cytokine-cytokine receptor interaction - Homo sapiens (human)**

[Pathway menu | Organism menu | Pathway entry | Download KGML | Show description | User data mapping]

Homo sapiens (human) Go 100%




Links to public annotation databases
Input data
Preprocessing
Gene sets enrichment analysis
Grouping
Functional enrichment
Location enrichment
miRNA Targets enrichment
Promoter Signals (PRIMA/ AMADEUS)
pathway enrichment
Visualization utilities

Inferring regulatory mechanisms from gene expression data

Assumption:

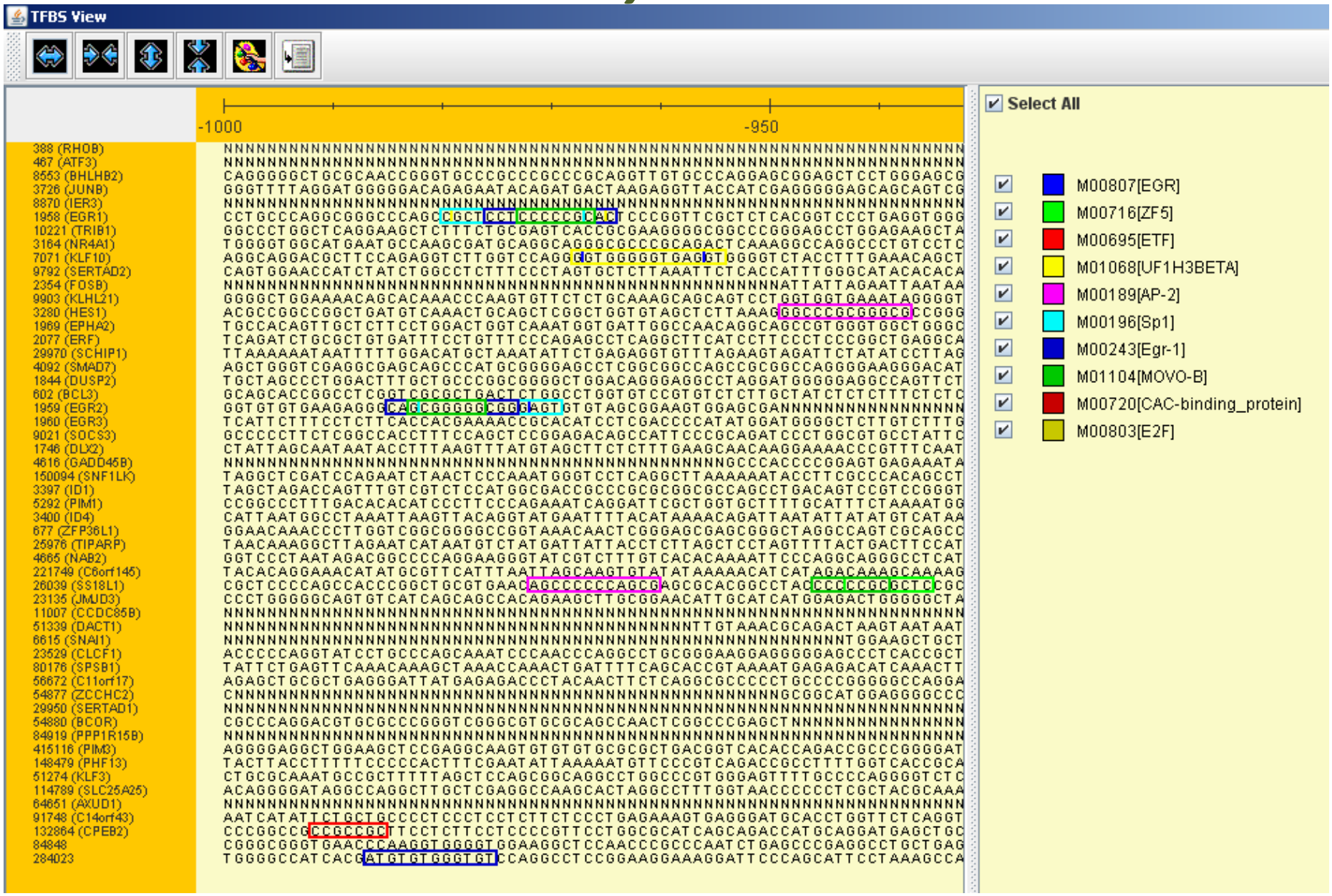
co-expression → transcriptional co-regulation → *common cis-regulatory promoter elements*

- Computational identification of *cis*-regulatory elements over-represention
- **PRIMA** - **PR**omoter **I**ntegration in **M**icroarray **A**nalysis (Elkon, et. Al, Genome Research , 2003)
- **AMADEUS** – novel motif enrichment analysis

PRIMA – general description

- *Input:*
  - *Target set* (e.g. co-expressed genes)
  - *Background set* (e.g. all genes on the chip)
- *Analysis:* Detects TFs with high target set prevalence
- TF binding site models – TRANSFAC DB
- Default: From -1000 bp to 200 bp relative the TSS

Promoter Analysis - Visualization

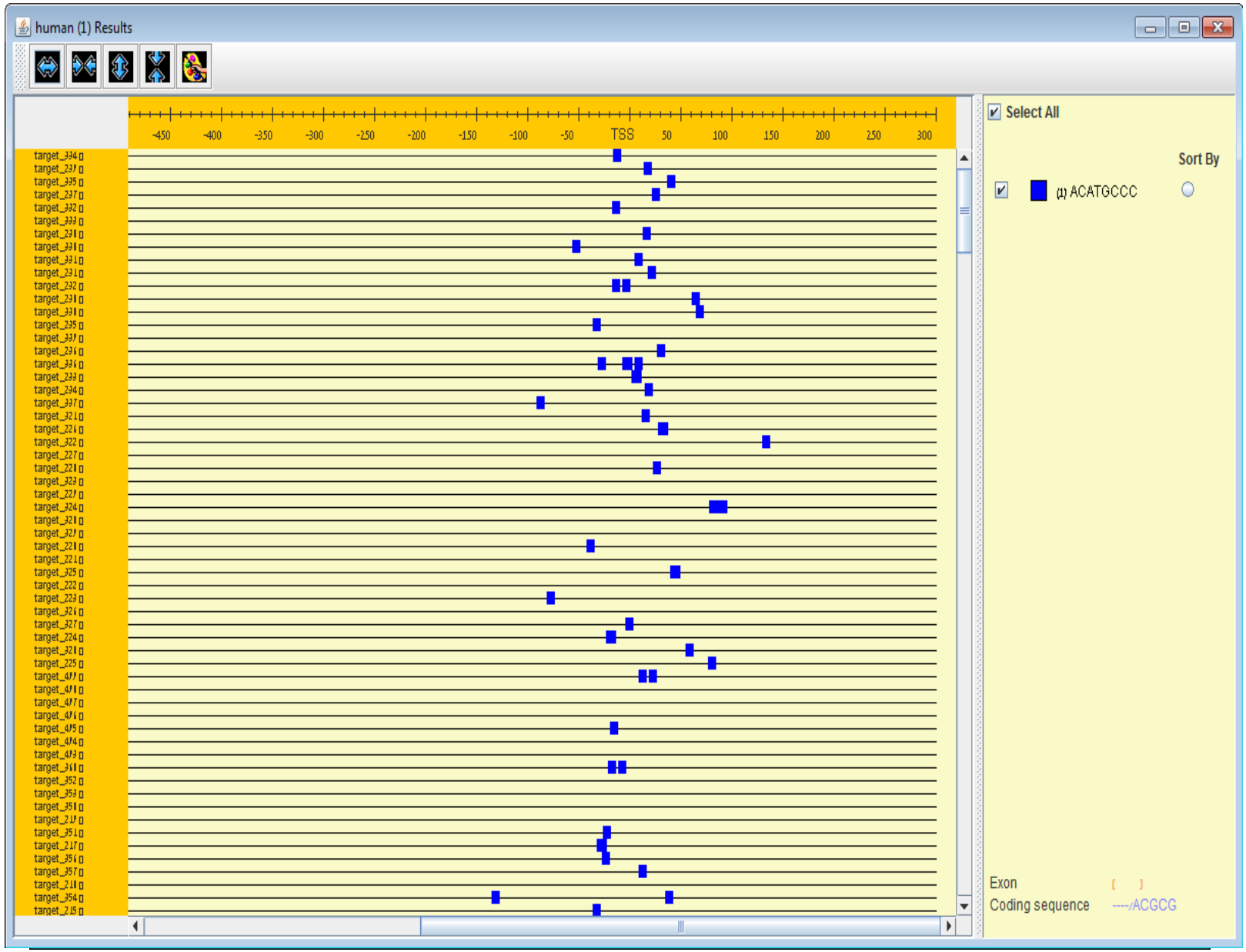


Amadeus

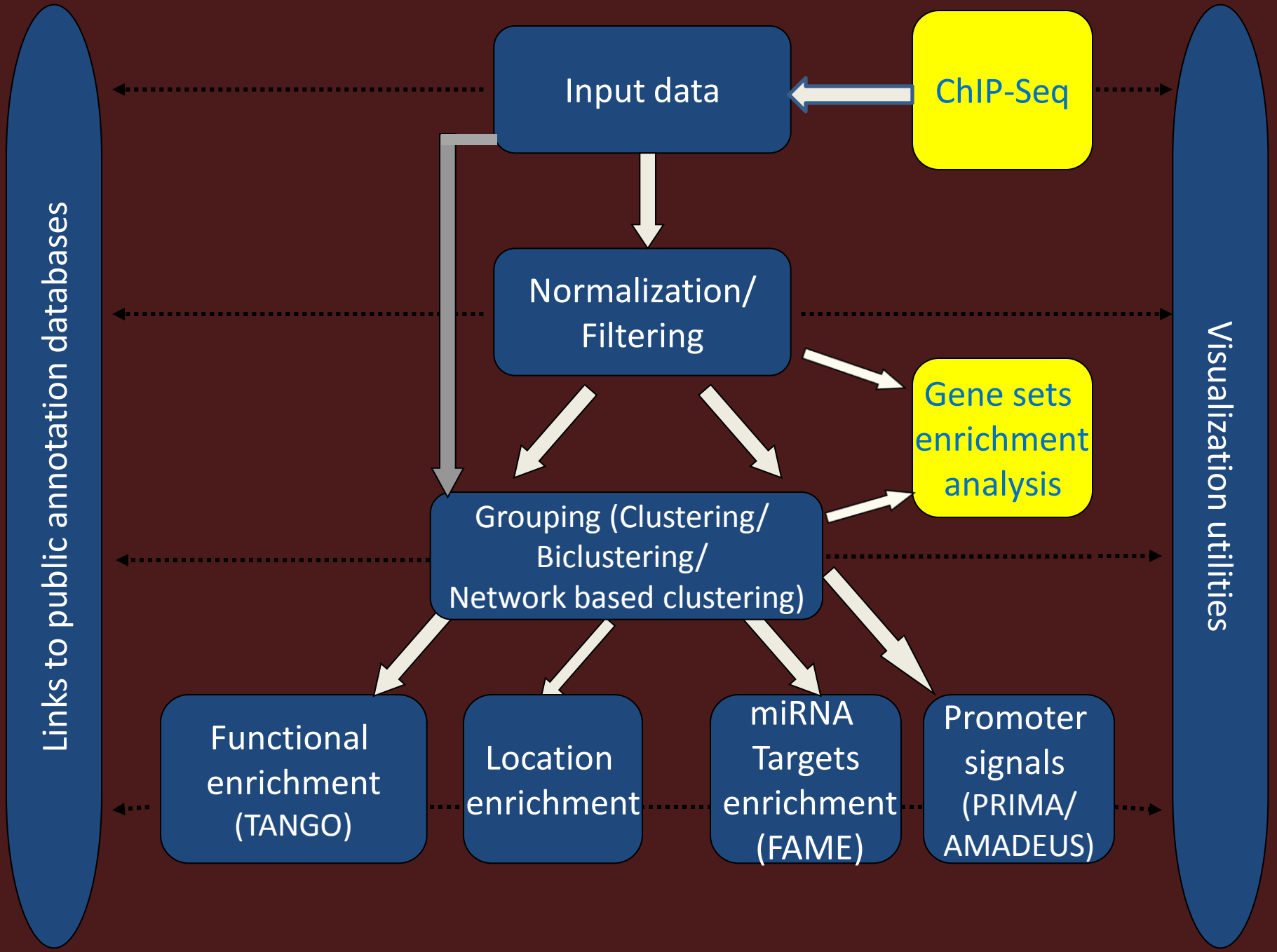
A Motif Algorithm for Detecting Enrichment in multiple Species

- **Supports diverse motif discovery tasks:**
  1. Finding **over-represented** motifs in one or more given **sets of genes**.
  2. Identifying motifs with **global spatial features** given **only** the genomic **sequences**.
- **Possible Gene-sets:**
  1. Identified gene sets clusters vs. all genes promoters.
  2. ChIP-Seq peaks sequences using Expander built-in FASTA sequences generation.

AMADEUS on ChIP-Seq peaks



Hands-on (4-7)


Links to public annotation databases
Input data
ChIP-Seq
Normalization/
Filtering
Gene sets
enrichment
analysis
Grouping (Clustering/
Biclustering/
Network based clustering)
Functional
enrichment
(TANGO)
Location
enrichment
miRNA
Targets
enrichment
(FAME)
Promoter
signals
(PRIMA/
AMADEUS)
Visualization utilities

Gene Sets Enrichment analysis

- Goal: Determine whether an a priori defined set of genes shows concordance with a biological pattern (e.g. differences between two phenotypes)
- Gene set sources:
  - ✓ MSigDB (Broad molecular signature database)
  - ✓ KEGG
  - ✓ Wiki pathways
- Gene rank sources
  - ✓ Phenotype labels
  - ✓ Imported
  - ✓ Selected condition
- Significance estimated with permutations
- FDR correction for multiple comparisons

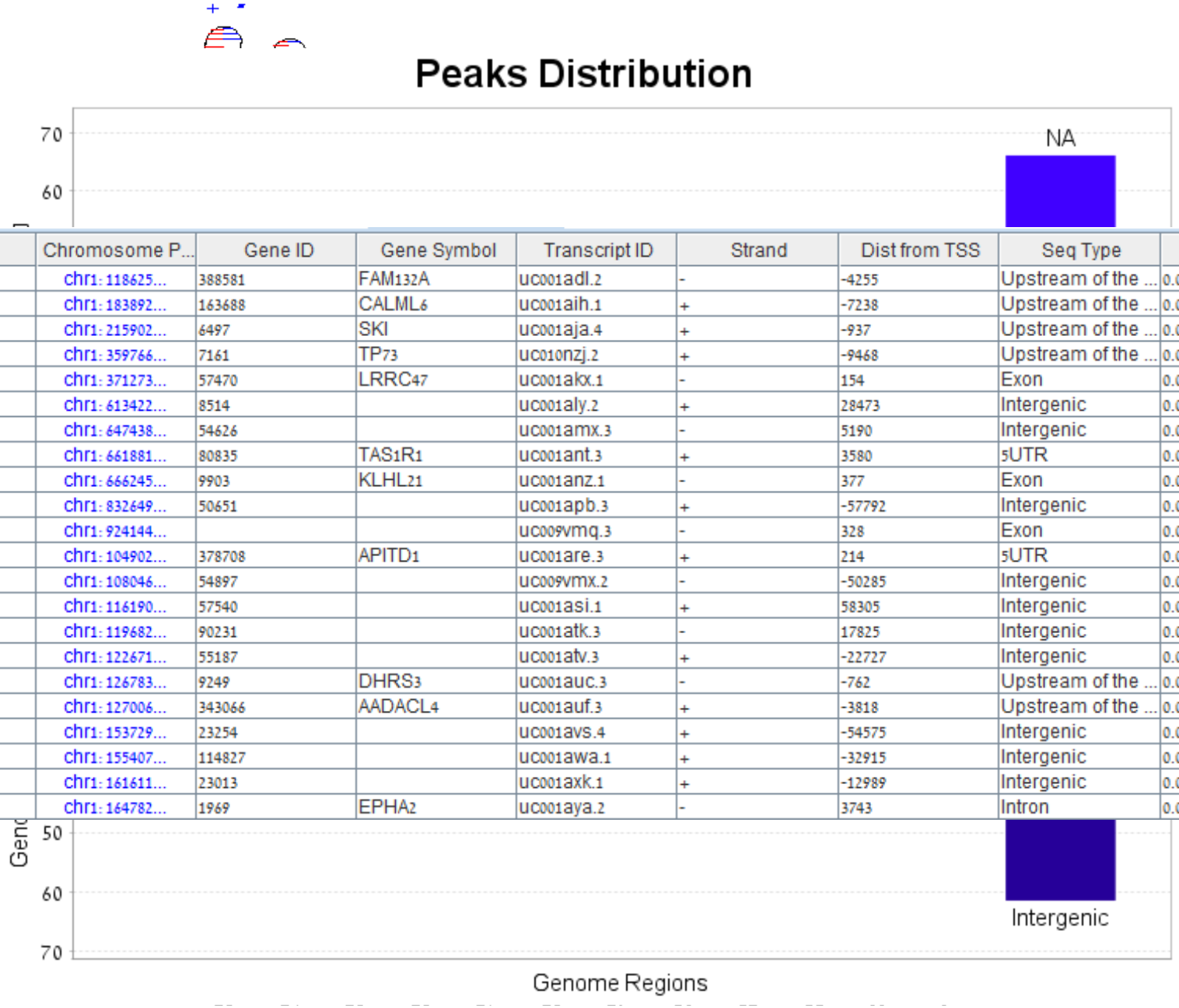
Gene Sets Enrichment visualization

ChIP-Seq enrichment analysis

- Searches for over-representation of genes closest to ChIP-Seq data peaks
- Uses hyper geometric test
- Multiple testing correction (Bonferroni)
- Enrichment results visualization (same as other group analysis results)

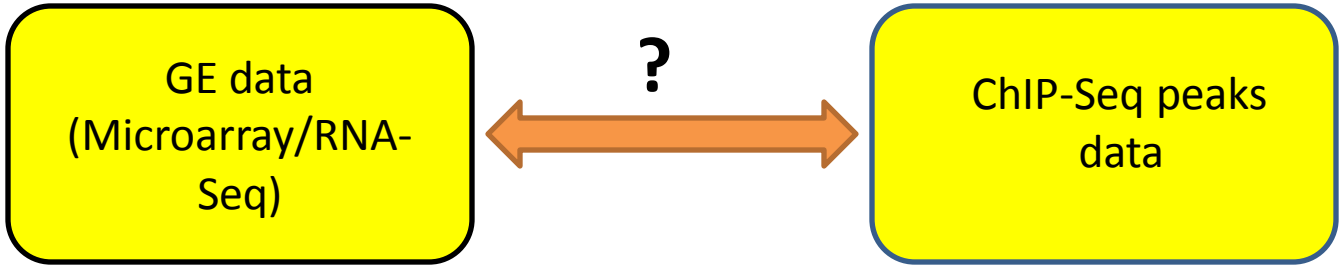
ChIP-Seq visualization

- Peaks to genomic region distributions
- Closest gene to peak chromosome visualization
- Peaks enrichment in genomic regions
- Peaks annotation table including closest gene and genomic region (e.g., 5UTR, Exon etc)



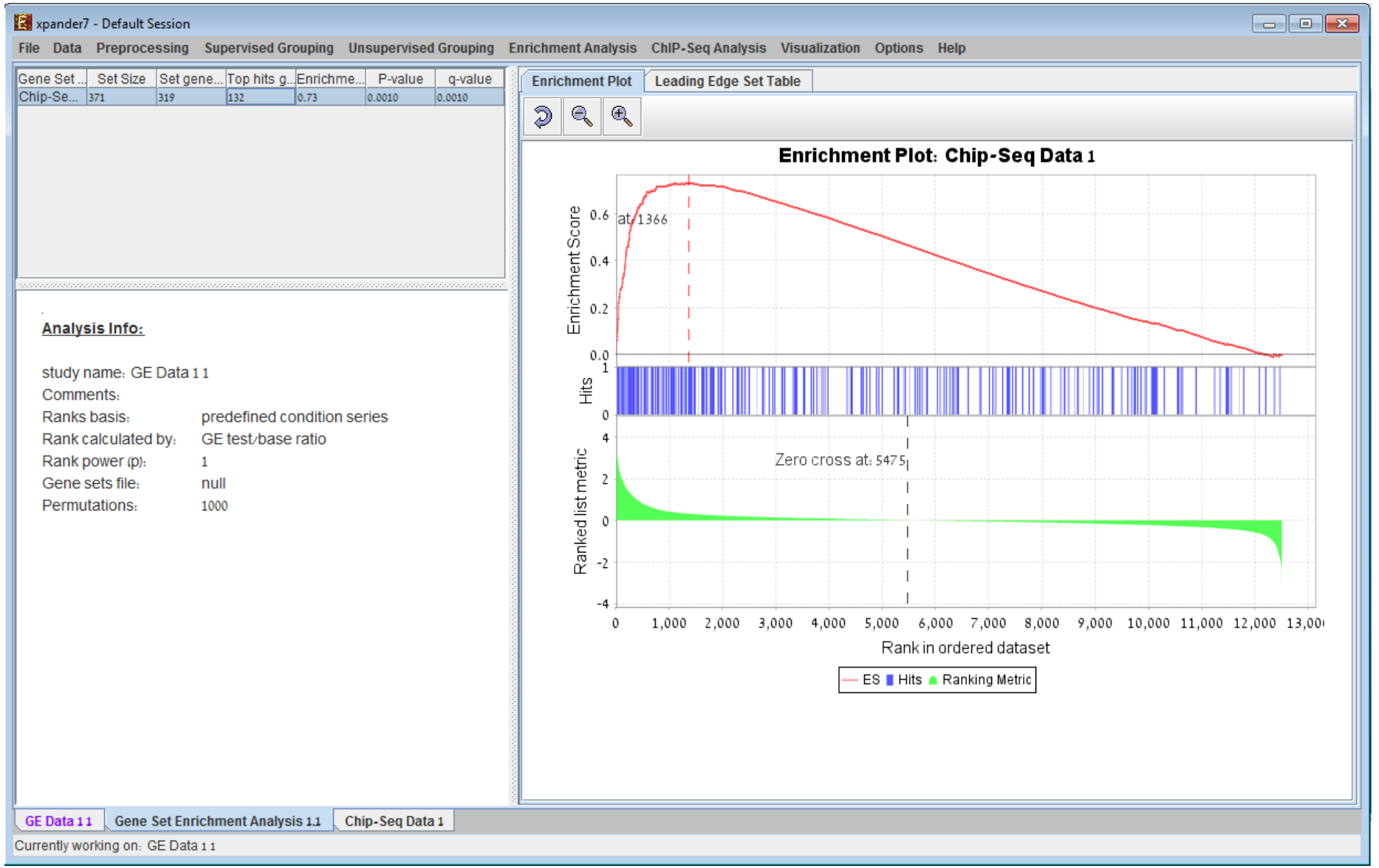
Peak ID	Chromosome P...	Gene ID	Gene Symbol	Transcript ID	Strand	Dist from TSS	Seq Type	Intensity
1	chr1: 118625...	388581	FAM132A	uc001adl.2	-	-4255	Upstream of the ...	0.0
2	chr1: 183892...	163688	CALML6	uc001aih.1	+	-7238	Upstream of the ...	0.0
3	chr1: 215902...	6497	SKI	uc001aja.4	+	-937	Upstream of the ...	0.0
4	chr1: 359766...	7161	TP73	uc010nzj.2	+	-9468	Upstream of the ...	0.0
5	chr1: 371273...	57470	LRRC47	uc001akx.1	-	154	Exon	0.0
6	chr1: 613422...	8514		uc001aly.2	+	28473	Intergenic	0.0
7	chr1: 647438...	54626		uc001amx.3	-	5190	Intergenic	0.0
8	chr1: 661881...	80835	TAS1R1	uc001ant.3	+	3580	5UTR	0.0
9	chr1: 666245...	9903	KLHL21	uc001anz.1	-	377	Exon	0.0
10	chr1: 832649...	50651		uc001apb.3	+	-57792	Intergenic	0.0
11	chr1: 924144...			uc009vmq.3	-	328	Exon	0.0
12	chr1: 104902...	378708	APITD1	uc001are.3	+	214	5UTR	0.0
13	chr1: 108046...	54897		uc009vmx.2	-	-50285	Intergenic	0.0
14	chr1: 116190...	57540		uc001asi.1	+	58305	Intergenic	0.0
15	chr1: 119682...	90231		uc001atk.3	-	17825	Intergenic	0.0
16	chr1: 122671...	55187		uc001atv.3	+	-22727	Intergenic	0.0
17	chr1: 126783...	9249	DHRS3	uc001auc.3	-	-762	Upstream of the ...	0.0
18	chr1: 127006...	343066	AADACL4	uc001auf.3	+	-3818	Upstream of the ...	0.0
19	chr1: 153729...	23254		uc001avs.4	+	-54575	Intergenic	0.0
20	chr1: 155407...	114827		uc001awa.1	+	-32915	Intergenic	0.0
21	chr1: 161611...	23013		uc001axk.1	+	-12989	Intergenic	0.0
22	chr1: 164782...	1969	EPHA2	uc001aya.2	-	3743	Intron	0.0

Integration between different technologies

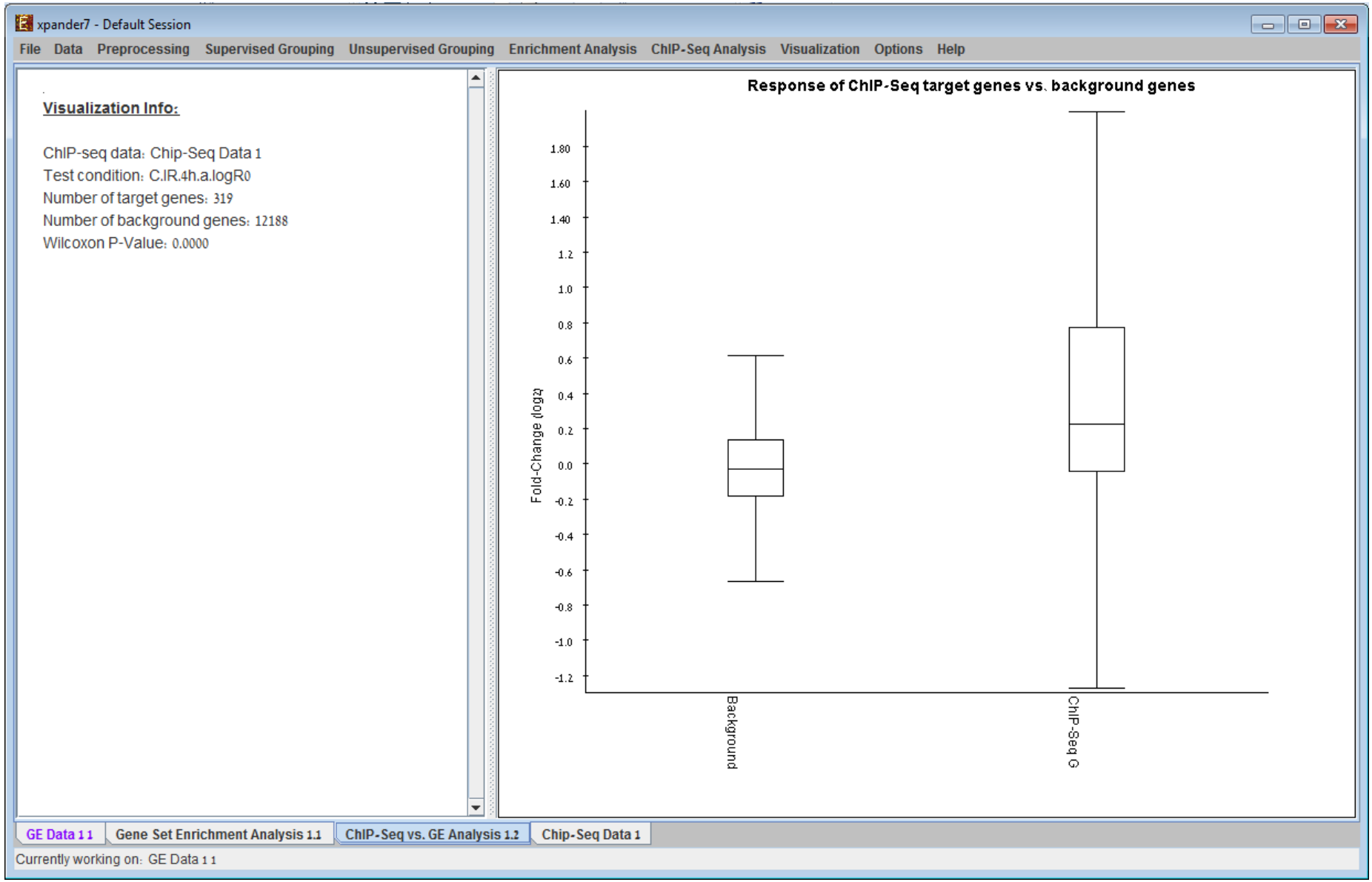


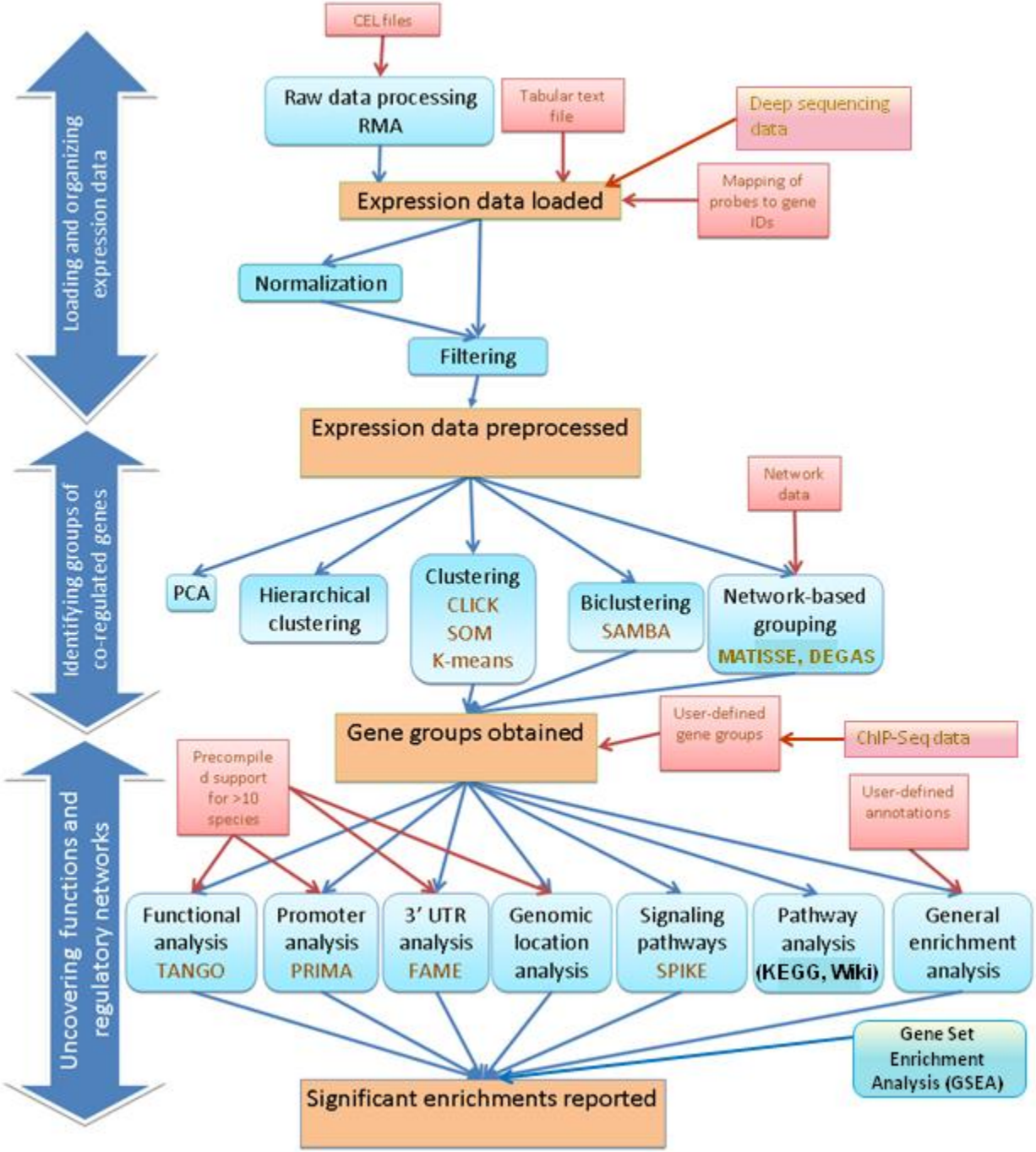
- ChIP-Seq vs. GE analysis:
  - GSEA – ChIP-Seq target genes as a single set
  - ChIP-Seq enrichment of GE's clusters
  - ChIP-Seq target genes distribution in GE
  - Expander enrichment tools (e.g., TANGO, PRIMA) – select ChIP-Seq target genes as a single cluster

GSEA – ChIP-Seq vs. GE



Rank distribution – ChIP-Seq vs. GE




Loading and organizing expression data
Identifying groups of co-regulated genes
Uncovering functions and regulatory networks
CEL files
Raw data processing RMA
Tabular text file
Deep sequencing data
Mapping of probes to gene IDs
Expression data loaded
Normalization
Filtering
Expression data preprocessed
Network data
PCA
Hierarchical clustering
Clustering CLICK SOM K-means
Biclustering SAMBA
Network-based grouping MATISSE, DEGAS
Gene groups obtained
User-defined gene groups
ChIP-Seq data
Precompiled support for >10 species
User-defined annotations
Functional analysis TANGO
Promoter analysis PRIMA
3' UTR analysis FAME
Genomic location analysis
Signaling pathways SPIKE
Pathway analysis (KEGG, Wiki)
General enrichment analysis
Gene Set Enrichment Analysis (GSEA)
Significant enrichments reported

Hands-on (8-10)