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. cerevisiae 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
Input data

Preprocessing

Grouping

Functional enrichment
Location enrichment
miRNA Targets enrichment
Promoter analyses
pathway enrichment

Gene sets enrichment analysis

Links to public annotation databases

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)
Input data

Preprocessing

Grouping

Functional enrichment
Location enrichment
miRNA Targets enrichment
Promoter signals
pathway enrichment

Gene sets enrichment analysis

Links to public annotation databases

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

- $\text{co-expression} \rightarrow \text{co-function}$
- $\text{co-expression} \rightarrow \text{co-regulation}$

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
Relevant knowledge can be revealed by identifying genes with common pattern across a subset of the conditions.

- 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 assess the quality of Bi-clusters
# Biclustering Visualization

![Biclustering Visualization](image)

<table>
<thead>
<tr>
<th>Bic num</th>
<th>Bic Score</th>
<th># Conditions</th>
<th>Project ID</th>
<th>Gene Symbol</th>
<th>Gene ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>230,475</td>
<td>7</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>2</td>
<td>456,917</td>
<td>7</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>3</td>
<td>433,011</td>
<td>6</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>4</td>
<td>249,352</td>
<td>7</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>5</td>
<td>431,604</td>
<td>7</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>6</td>
<td>177,526</td>
<td>7</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>7</td>
<td>116,431</td>
<td>5</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>8</td>
<td>374,202</td>
<td>9</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>9</td>
<td>331,779</td>
<td>10</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>10</td>
<td>320,373</td>
<td>6</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>11</td>
<td>417,158</td>
<td>7</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>12</td>
<td>256,944</td>
<td>6</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>13</td>
<td>144,321</td>
<td>12</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>14</td>
<td>201,665</td>
<td>10</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>15</td>
<td>435,369</td>
<td>7</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>16</td>
<td>650,657</td>
<td>8</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>17</td>
<td>148,6</td>
<td>10</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>18</td>
<td>200,601</td>
<td>14</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>19</td>
<td>206,653</td>
<td>6</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>20</td>
<td>372,519</td>
<td>10</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>21</td>
<td>575,098</td>
<td>5</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>22</td>
<td>387,311</td>
<td>6</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>23</td>
<td>213,658</td>
<td>4</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>24</td>
<td>156,952</td>
<td>14</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>25</td>
<td>458,704</td>
<td>12</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>26</td>
<td>522,585</td>
<td>12</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
<tr>
<td>27</td>
<td>179,158</td>
<td>7</td>
<td>YHL025W</td>
<td>YHL025W</td>
<td>YHL025W</td>
</tr>
</tbody>
</table>
Hands-on (3-4)
Input data

Preprocessing

Grouping

Functional enrichment

Location enrichment

miRNA Targets enrichment

Promoter signals

Gene sets enrichment analysis

Pathway enrichment

Visualization utilities

Links to public annotation databases
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

Can be saved 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
Cytokine-cytokine receptor interaction - Homo sapiens (human)
Input data

Preprocessing

Grouping

Functional enrichment

Location enrichment

miRNA Targets enrichment

Promoter Signals (PRIMA/AMADEUS)

Gene sets enrichment analysis

Links to public annotation databases

Visualization utilities

Links to public annotation databases
Inferring regulatory mechanisms from gene expression data

Assumption:

\[ \text{co-expression} \rightarrow \text{transcriptional co-regulation} \rightarrow \text{common cis-regulatory promoter elements} \]

• Computational identification of \textit{cis}-regulatory elements over-representation

• \textbf{PRIMA} - \textbf{PR}omoter \textbf{I}ntegration in \textbf{M}icroarray \textbf{A}nalysis (Elkon, et. Al, Genome Research , 2003)

• \textbf{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)
Input data

Normalization/Filtering

ChIP-Seq

Gene sets enrichment analysis

Grouping (Clustering/Biclustering/Network based clustering)

Links to public annotation databases

Visualization utilities

Functional enrichment (TANGO)

Location enrichment

miRNA Targets enrichment (FAME)

Promoter signals (PRIMA/AMADEUS)
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

Enrichment Plot: Notch Signaling Pathway

Zero cross at: 4491

Rank in ordered dataset

Enrichment Score

Hits

Ranked list metric

ES  Hits  Ranking Metric
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)
# Peaks Distribution

<table>
<thead>
<tr>
<th>Peak ID</th>
<th>Chromosome Position</th>
<th>Gene ID</th>
<th>Gene Symbol</th>
<th>Transcript ID</th>
<th>Strand</th>
<th>Dist from TSS</th>
<th>Seq Type</th>
<th>Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>chr1: 118625...</td>
<td>388581</td>
<td>FAM132A</td>
<td>uc001adl.2</td>
<td>-</td>
<td>-4255</td>
<td>Upstream of the TSS</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>chr1: 183892...</td>
<td>163688</td>
<td>CALML6</td>
<td>uc001ahl.1</td>
<td>+</td>
<td>-7238</td>
<td>Upstream of the TSS</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>chr1: 215902...</td>
<td>6497</td>
<td>SKI</td>
<td>uc001ajg.4</td>
<td>+</td>
<td>-937</td>
<td>Upstream of the TSS</td>
<td>0.0</td>
</tr>
<tr>
<td>4</td>
<td>chr1: 359766...</td>
<td>7161</td>
<td>TP73</td>
<td>uc001onz.2</td>
<td>+</td>
<td>-9468</td>
<td>Upstream of the TSS</td>
<td>0.0</td>
</tr>
<tr>
<td>5</td>
<td>chr1: 371273...</td>
<td>57470</td>
<td>LRRC47</td>
<td>uc001aix.1</td>
<td>-</td>
<td>154</td>
<td>Exon</td>
<td>0.0</td>
</tr>
<tr>
<td>6</td>
<td>chr1: 613422...</td>
<td>8514</td>
<td></td>
<td>uc001aly.2</td>
<td>+</td>
<td>28473</td>
<td>Intergenic</td>
<td>0.0</td>
</tr>
<tr>
<td>7</td>
<td>chr1: 647438...</td>
<td>54626</td>
<td></td>
<td>uc001amx.3</td>
<td>-</td>
<td>5190</td>
<td>Intergenic</td>
<td>0.0</td>
</tr>
<tr>
<td>8</td>
<td>chr1: 661881...</td>
<td>80835</td>
<td>TAS1R1</td>
<td>uc001ant.3</td>
<td>+</td>
<td>3580</td>
<td>5'UTR</td>
<td>0.0</td>
</tr>
<tr>
<td>9</td>
<td>chr1: 666245...</td>
<td>9903</td>
<td>KLHL21</td>
<td>uc001anv.2</td>
<td>+</td>
<td>377</td>
<td>Exon</td>
<td>0.0</td>
</tr>
<tr>
<td>10</td>
<td>chr1: 832649...</td>
<td>50651</td>
<td></td>
<td>uc001apb.3</td>
<td>+</td>
<td>-57792</td>
<td>Intergenic</td>
<td>0.0</td>
</tr>
<tr>
<td>11</td>
<td>chr1: 924144...</td>
<td>39184</td>
<td></td>
<td>uc009vmq.3</td>
<td>-</td>
<td>328</td>
<td>Exon</td>
<td>0.0</td>
</tr>
<tr>
<td>12</td>
<td>chr1: 104902...</td>
<td>378708</td>
<td>APITD1</td>
<td>uc001are.3</td>
<td>+</td>
<td>214</td>
<td>5'UTR</td>
<td>0.0</td>
</tr>
<tr>
<td>13</td>
<td>chr1: 108046...</td>
<td>54897</td>
<td></td>
<td>uc009vmx.2</td>
<td>-</td>
<td>-50285</td>
<td>Intergenic</td>
<td>0.0</td>
</tr>
<tr>
<td>14</td>
<td>chr1: 116190...</td>
<td>57540</td>
<td></td>
<td>uc001as1.1</td>
<td>+</td>
<td>59305</td>
<td>Intergenic</td>
<td>0.0</td>
</tr>
<tr>
<td>15</td>
<td>chr1: 119682...</td>
<td>90231</td>
<td></td>
<td>uc001atk.3</td>
<td>-</td>
<td>17825</td>
<td>Intergenic</td>
<td>0.0</td>
</tr>
<tr>
<td>16</td>
<td>chr1: 122671...</td>
<td>55187</td>
<td></td>
<td>uc001atv.3</td>
<td>+</td>
<td>-22727</td>
<td>Intergenic</td>
<td>0.0</td>
</tr>
<tr>
<td>17</td>
<td>chr1: 126783...</td>
<td>9249</td>
<td>DHRS3</td>
<td>uc001auc.3</td>
<td>-</td>
<td>-762</td>
<td>Upstream of the TSS</td>
<td>0.0</td>
</tr>
<tr>
<td>18</td>
<td>chr1: 127006...</td>
<td>343066</td>
<td>AADACL4</td>
<td>uc001auf.3</td>
<td>+</td>
<td>-3818</td>
<td>Upstream of the TSS</td>
<td>0.0</td>
</tr>
<tr>
<td>19</td>
<td>chr1: 153729...</td>
<td>23254</td>
<td></td>
<td>uc001avs.4</td>
<td>+</td>
<td>-54575</td>
<td>Intergenic</td>
<td>0.0</td>
</tr>
<tr>
<td>20</td>
<td>chr1: 159407...</td>
<td>114827</td>
<td></td>
<td>uc001awx.1</td>
<td>+</td>
<td>-32915</td>
<td>Intergenic</td>
<td>0.0</td>
</tr>
<tr>
<td>21</td>
<td>chr1: 161611...</td>
<td>23013</td>
<td></td>
<td>uc001axk.1</td>
<td>+</td>
<td>-12989</td>
<td>Intergenic</td>
<td>0.0</td>
</tr>
<tr>
<td>22</td>
<td>chr1: 164782...</td>
<td>1969</td>
<td>EPHA2</td>
<td>uc001aya.2</td>
<td>-</td>
<td>3743</td>
<td>Intron</td>
<td>0.0</td>
</tr>
</tbody>
</table>

---

**Genome Regions**

- **Intergenic**
Integration between different technologies

- GE data (Microarray/RNA-Seq) vs. ChIP-Seq peaks data

  - 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

Analysis Info:
- study name: GE Data 11
- Comments:
- Ranks basis: predefined condition series
- Rank calculated by: GE test/base ratio
- Rank power "p": 1
- Gene sets file: null
- Permutations: 1000

Enrichment Plot: Chip-Seq Data 1
- Enrichment Score at rank 1,366
- Zero cross at rank 5,475

Enrichment Analysis: Chip-Seq Analysis
- Gene Set Enrichment Analysis 11
Rank distribution – ChIP-Seq vs. GE

Visualization Info:
ChIP-seq data: Chip-Seq Data 1
Test condition: Chr4.th.a.logRq
Number of target genes: 319
Number of background genes: 12188
Wilcoxon P-Value: 0.0000
Hands-on (8-10)