Bioinformatics opportunities in Genomics and Genetics

Case Study: Prediction of novel gene functions of *NSF1/YPL230W* in ***Saccharomyces Cerevisiae*** via search for maximally interconnected sub-graph

Kyrylo Bessonov

PhD candidate

GiGa / Montefiore Institute

June 5th 2013

Outline

- Current challenges in bioinformatics
- Applications
  - Sequence alignments
  - Gene function prediction
    - Building gene expression network conditioned on *NSF1* .
    - Interconnected correlation clustering method (ICC)
  - Genome Wide Association Studies (GWAS)
    - Biological vs statistical epistasis
    - Brief into into Model Based Multifactor Dimensionality Reduction (MB-MDR) algorithm

Bioinformatics / Computational Biology

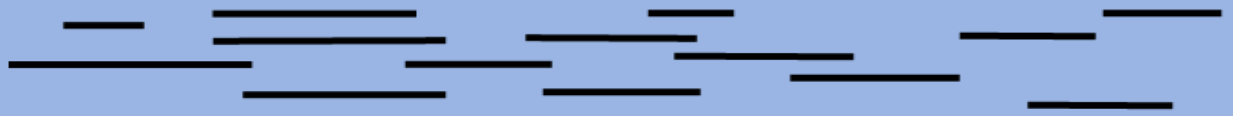
- **Definition:**

  Bioinformatics / Computational Biology - a field of biology concerned with the development of techniques for the collection and **manipulation of biological data**, and the use of such data to make biological discoveries or predictions.

- **Sub-fields / links:**
  - Genetics / Genomics
  - Molecular Biology and even,
  - Structural Biology / Molecular Dynamics

Bioinformatics Challenges

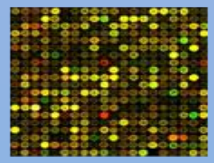
1. Genome Assembly

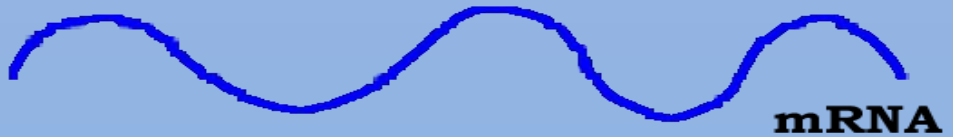


2. Genome Annotation
   - Gene search
   - Gene function
   - Lietrature search

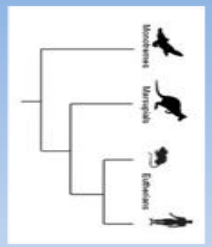
3. Sequence Alignments

4. Gene Expression



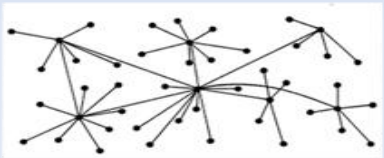


5. Evolution Theory

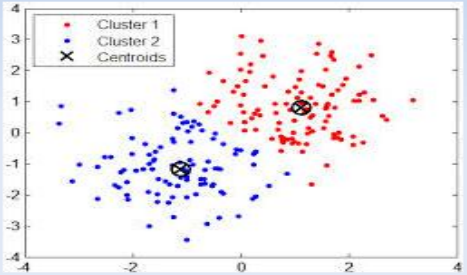


6. Database storage
"Big Data" management

7. Regulatory network analysis
   - inference
   - properties analysis



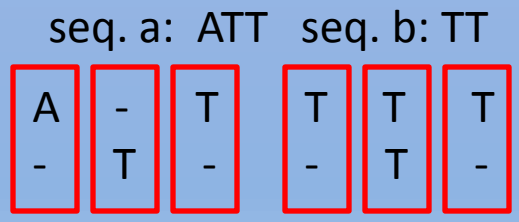
8. Cluster discovery



9. Classificiation of data
   - building of accurate classifiers

1) Sequence Alignments

- **Problem:** optimal alignment of sequences
  - **Aim:** find sequence homology / functional motifs
  - **Challenge:** many possible alignments
    - Not enough computational power
    - Naïve algorithm is very inefficient (O^{exp})
    - Example: for sequence of length 15, need to consider
      - Possibilities # = (insertion, deletion, gap)^{15} = 3^{15} = 1,4*10^7



- **Solution:** Brilliant idea to use dynamic programming
  - Define scoring rules to find the most optimal alignment
  - Create of alignment matrix
  - Avoid duplicate calculations; compute sub-alignment once

1) Local alignment

- Given alignment scoring function, find optimal alignment for sequence *a* and *b*

$$S = \sum_{1}^{n} s_{ij}$$

n = smallest sequence length

s=score at position *i,j*

S = overall alignment score

- Scoring scheme:

$-\ s(a_i, b_j) = +2$ if $a_i = b_j$,

$-\ s(a_i, b_j) = -1$ if $a_i \neq b_j$ and

$-\ s(a_i, -) = s(-, b_j) = -2$

sequence ***a***

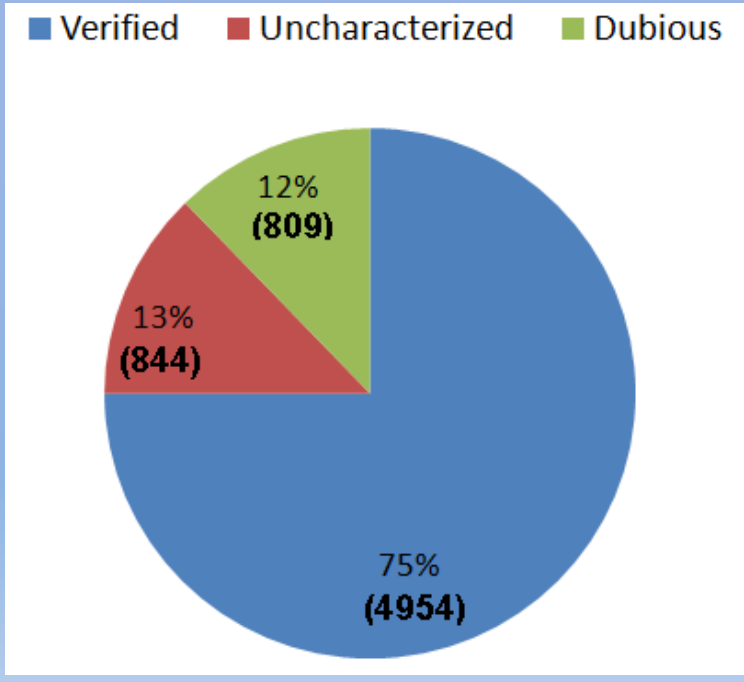
Sequence ***b***			**A**	**T**	**T**
		0	0	0	0
	T	0	0	2	0
	T	0	0	2	**4**

Case Study:

Prediction of novel gene functions of *NSF1/YPL230W* in ***Saccharomyces Cerevisiae*** via search for maximally interconnected sub-graph

2) Gene function prediction

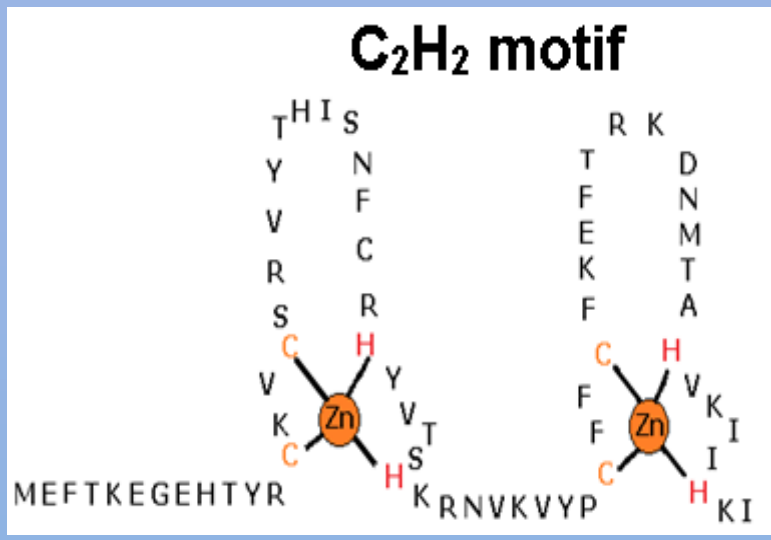
- ~25% of *S.cerevisiae* ORFs/genes are poorly characterized
- Even well annotated genomes (yeast / human) have lost of genes with unknown function
- Important for biology advances and posterior analysis / interpretation
- **E.g.** *NSF1/YPL230W* gene functions are far being explored under different context

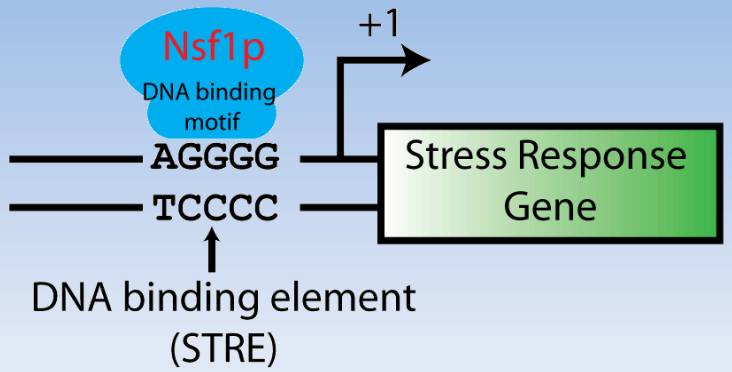


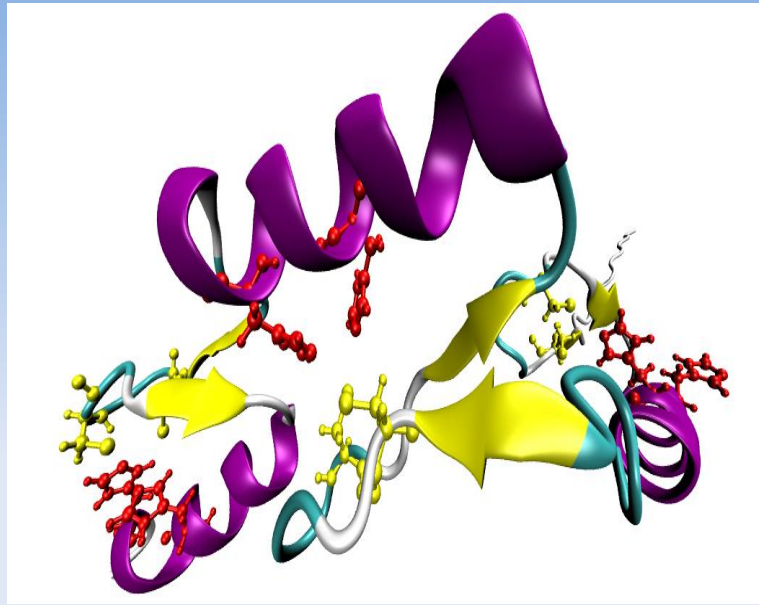
S.cerevisiae genome annotation
Source: SGD database

Nutrient and Stress Factor 1 (NSF1)

- *NSF1* encodes a C_2H_2 zinc finger transcription factor (TF)
- *NSF1* binds to **S**tress **R**esponse **E**lement (**5′-CCCCT-3′**) sequence [1]
  - involved in stress responses

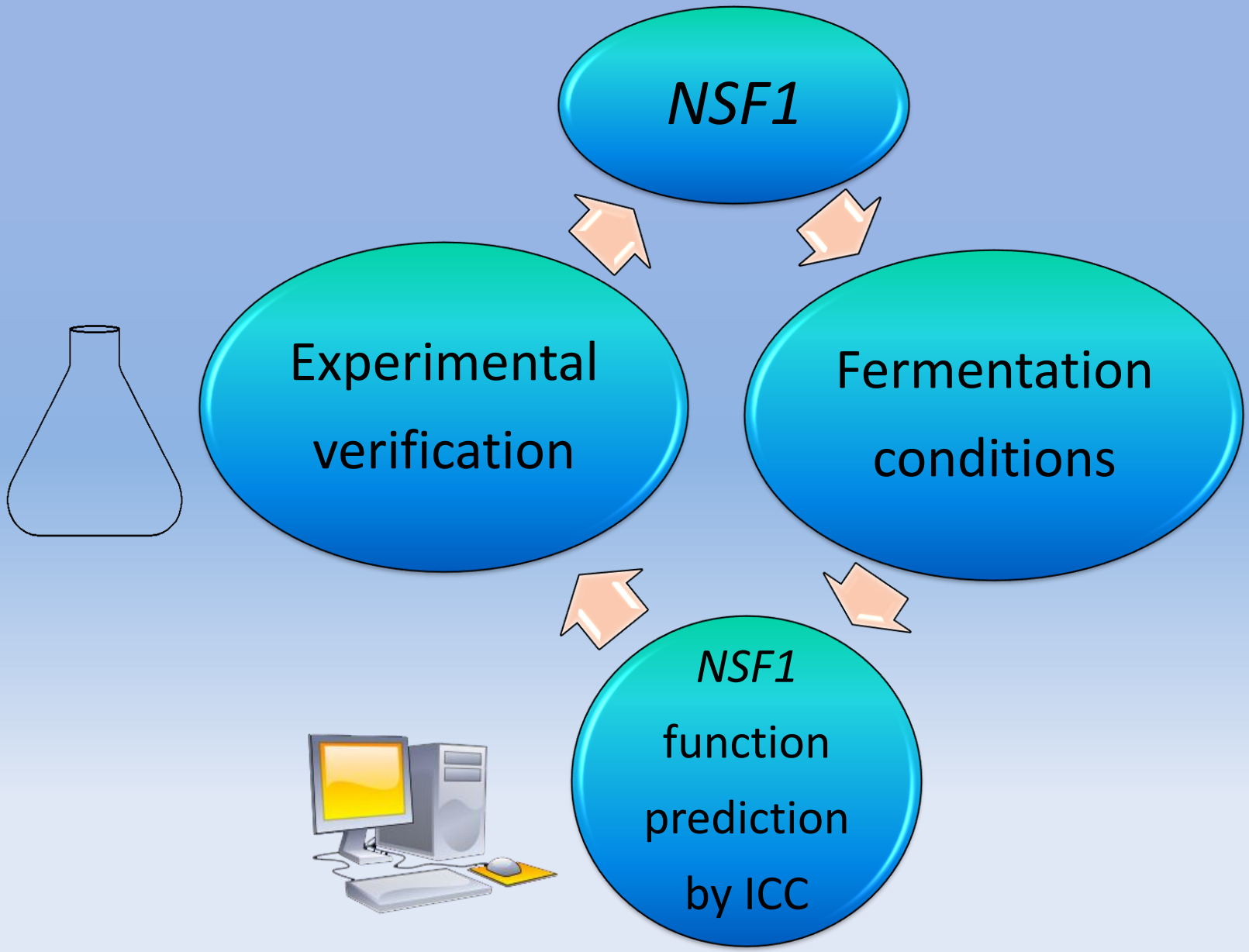






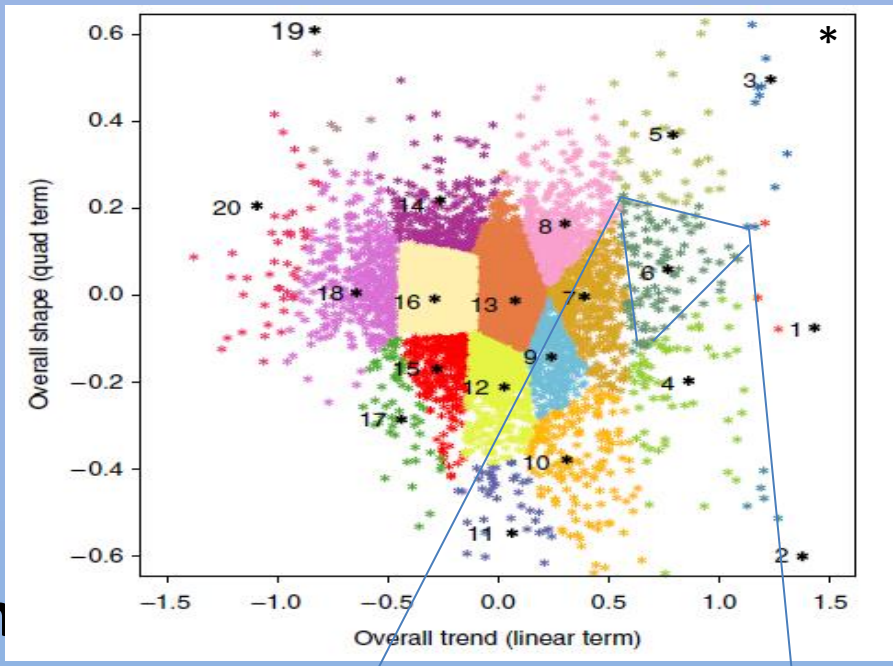
Putative Nsf1p structure based on Human Kruppel-Like Factor 5 (54%

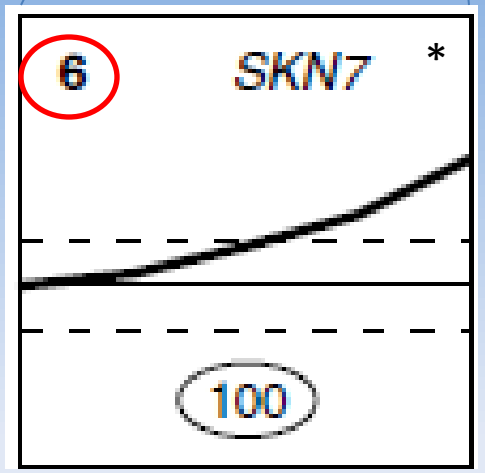
Experimental design



NSF1 known roles

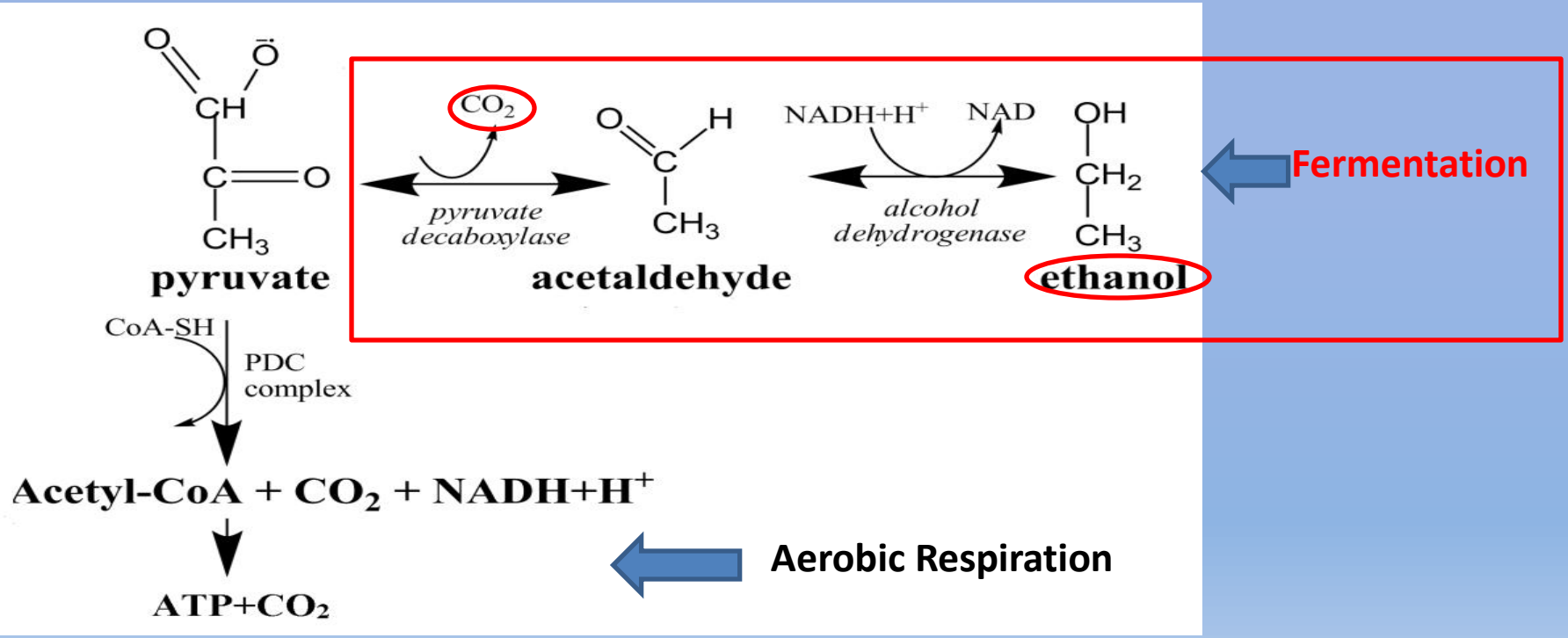
- *NSF1* previoulsy roles:
  - nutrient adaptation
  - energy metabolism
  - osmotic stress (e.g. salt)
- *NSF1* is part of so called **Ferm**
  Stress Response (FSR)[3] clusters 1-6
  - 223 genes 4 -80 fold <u>induced</u> (↑) at the end of a fermentation
- *NSF1* is part of <u>cluster 6</u> as on the right
- *NSF1* roles are unknown under **fermentation conditions**





***** Marks et. al 2008 “Dynamics of the yeast transcriptome during wine fermentation reveals a novel fermentation stress response” *FEMS* Yeast Res 8 (2008) 35–52

Fermentation – Production of Ethanol and CO_2



- **Fermentation:** under **low oxygen** conditions and **high glucose** levels (>0.8 mM) yeast ferments → CO_2 + **Ethanol**

Fermentation Stresses

- **Nutrient limitation**
  - fermentable carbon sources depletion (glucose, fructose)
  - depletion of nitrogen, phosphate and sulfur sources
- **Osmotic Stress and Ethanol Toxicity**
  - initial high levels of fermentable sugars (> 200 g/L)
  - increasing ethanol levels
- **Acidic conditions**
  - grape must pH ~ 3.5

Expression data / profiles

- Microarrays acquire **genome-wide** gene expression data[15]
- Gene expression data could be represented as:
  - time series (time-course experiments)

$$X = \begin{bmatrix} T_1 & T_2 & T_t \\ x_{11} & x_{12} & x_{1t} \\ x_{21} & x_{22} & x_{2t} \\ x_{n1} & x_{n2} & x_{nt} \end{bmatrix}$$

⇐ Gene expression **profile A**

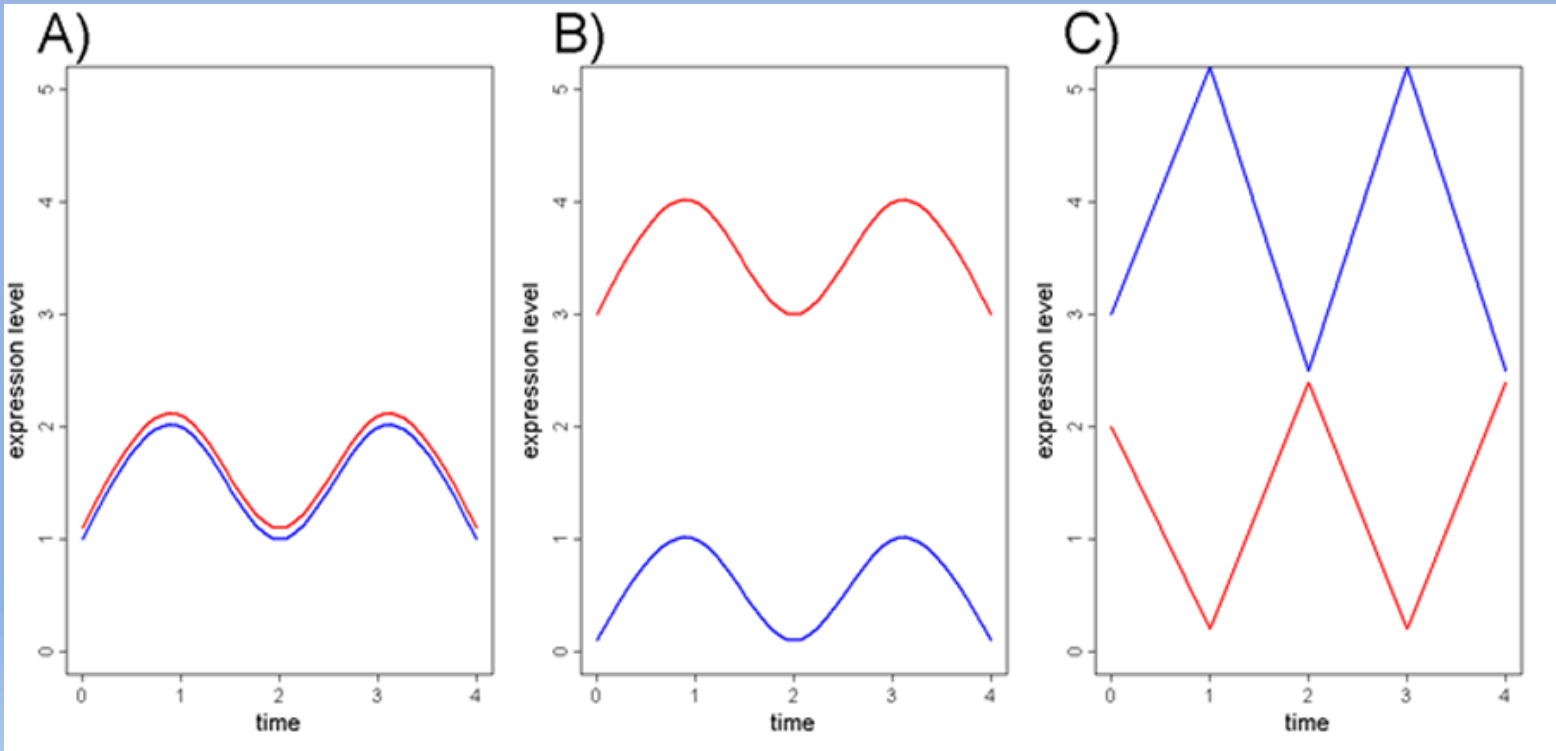
⇐ Gene expression **profile B**

- Given a_i and b_i corresponding to **profile A** and **B** across n points the Pearson's Correlation Coefficient (**PCC**) [14] is:

$$PCC(r) = \frac{\sum_{i=1}^{n}(a_i - \bar{a}) * (b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2} * \sqrt{\sum_{i=1}^{n}(b_i - \bar{b})^2}}$$

- **PCC** measures **similarity of expression profiles**

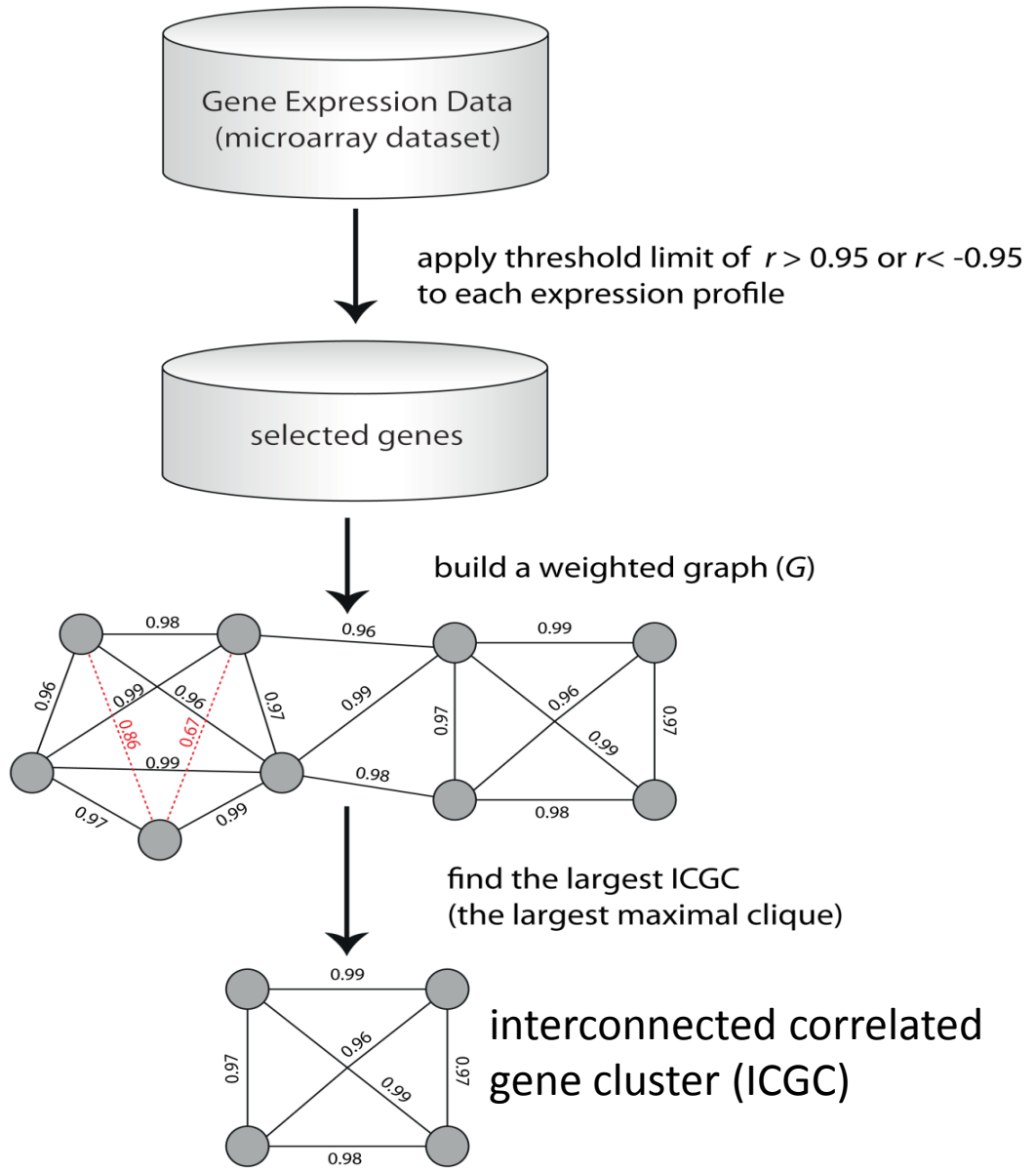
PCC as measure of similarity



- **A)** PCC = 1 **B)** PC = 1 **C)** PC = -1
- PCC compares expression profiles based on shape and not absolute values

Interdependent Correlation Clustering

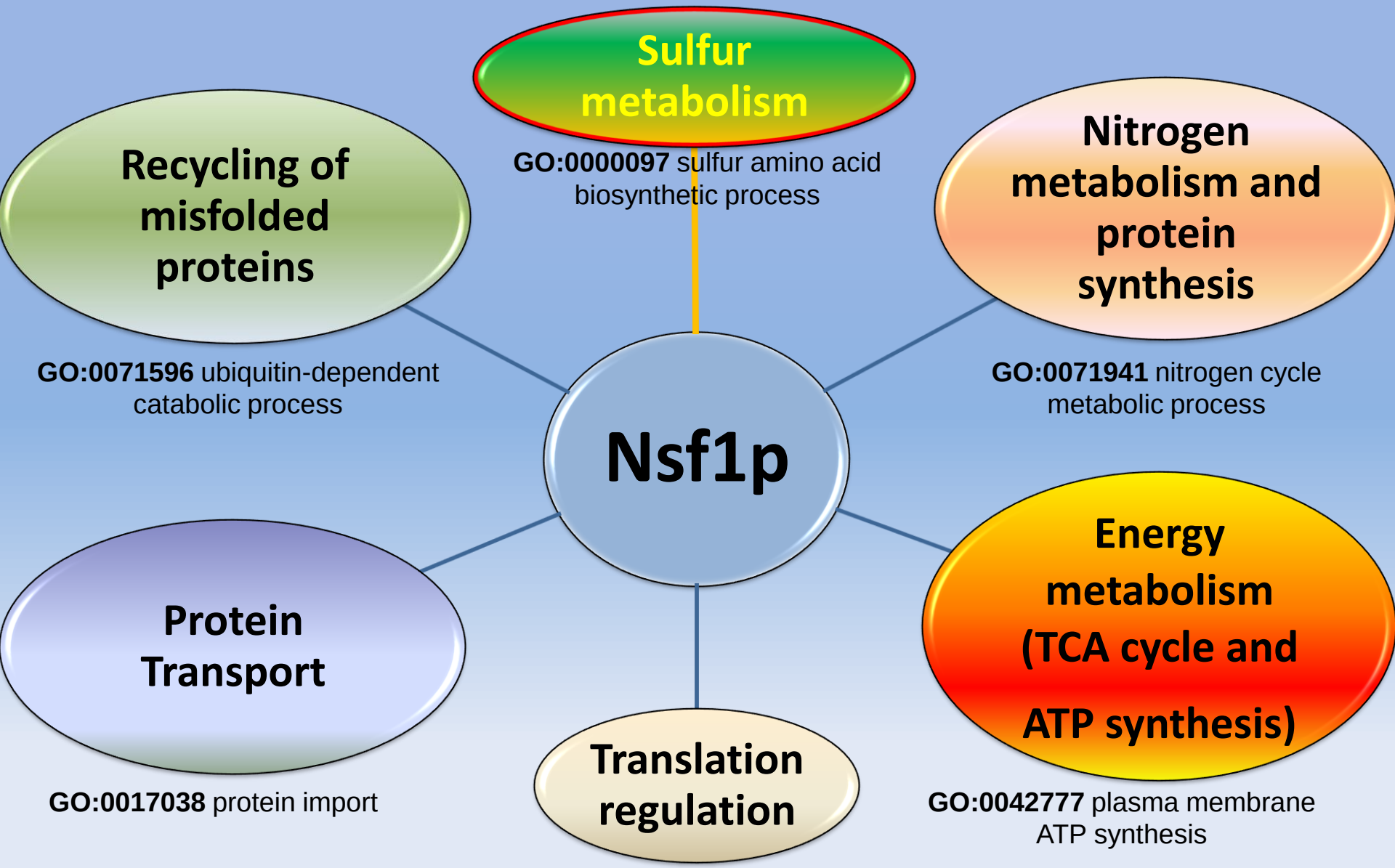
- **Step 1:** Compare each individual gene expression profile to the target gene (*NSF1*)
- **Step 2:** Select genes highly positively and negatively correlated to the target gene (*NSF1*) passing the threshold $r<-0.95$ or $r>0.95$. Store the selected genes in the *selected genes*.
- **Step 3:** Build a weighted graph (*G*). Assign $E^{+}=1$ [link] if the PCC value between corresponding vertices meets the threshold of $r<-0.95$ or $r>0.95$; otherwise assign $E^{-}=0$ [no link].
- **Step 4:** Find the maximally interconnected sub-group of nodes, the ICGC, in *G* using the Born-Kerbosch heuristic algorithm.



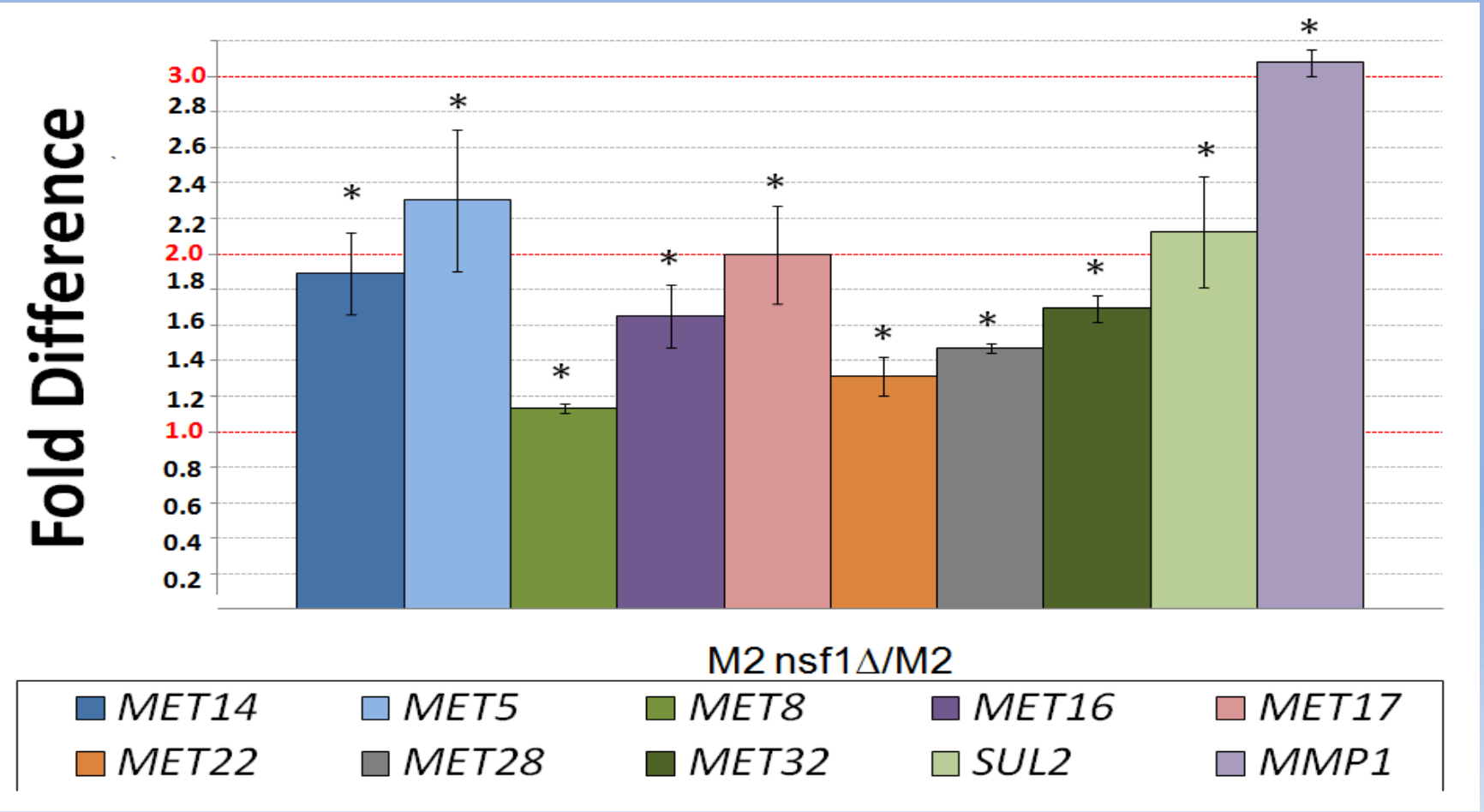
The Advantages of Proposed ICC Method

1. The complex data is transformed into a weighted graph using simple PCC a similarity measure
2. The ICGC represents a very tight cluster of:
   - interconnected genes that reinforce each other
   - stringent criteria for inclusion of additional genes
   - conditioned on target gene
3. Could be applied even on very small datasets and time series data (e.g. time course experiments)

ICC significant GO terms by category

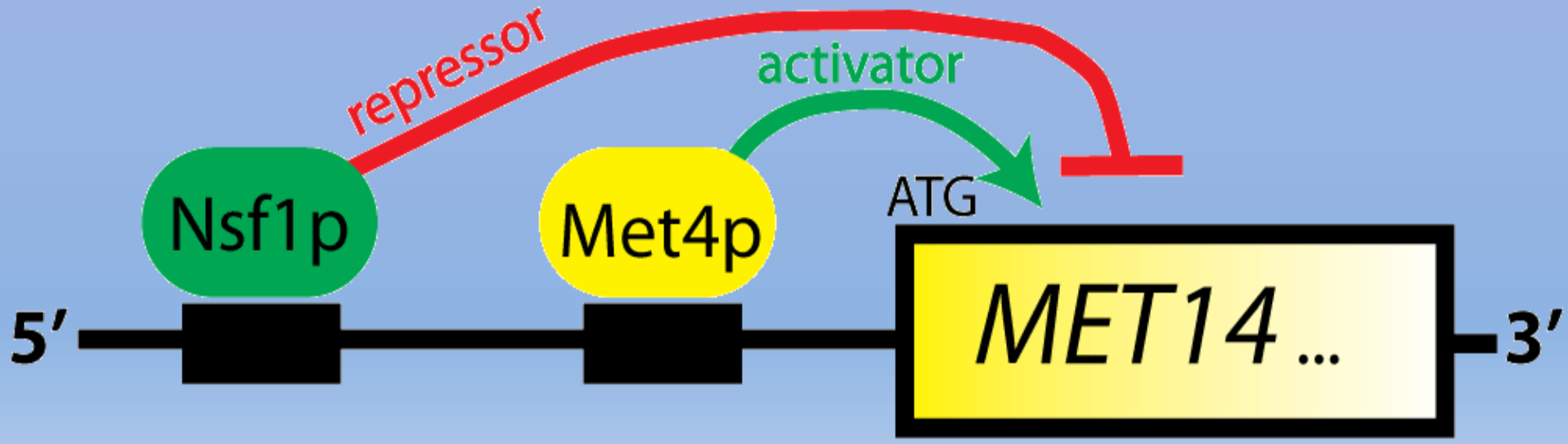


Differentially Expressed Genes (DEGs)



- Two-sample paired t-test at α = 0.05 found numerous sulfur metabolism DEGs between M2 vs M2*nsf1Δ* yeast groups at 85% glucose fermented time point
- sulfur metabolism related DEGs were all higher in M2 *nsf1Δ*

Met4p and Nsf1p

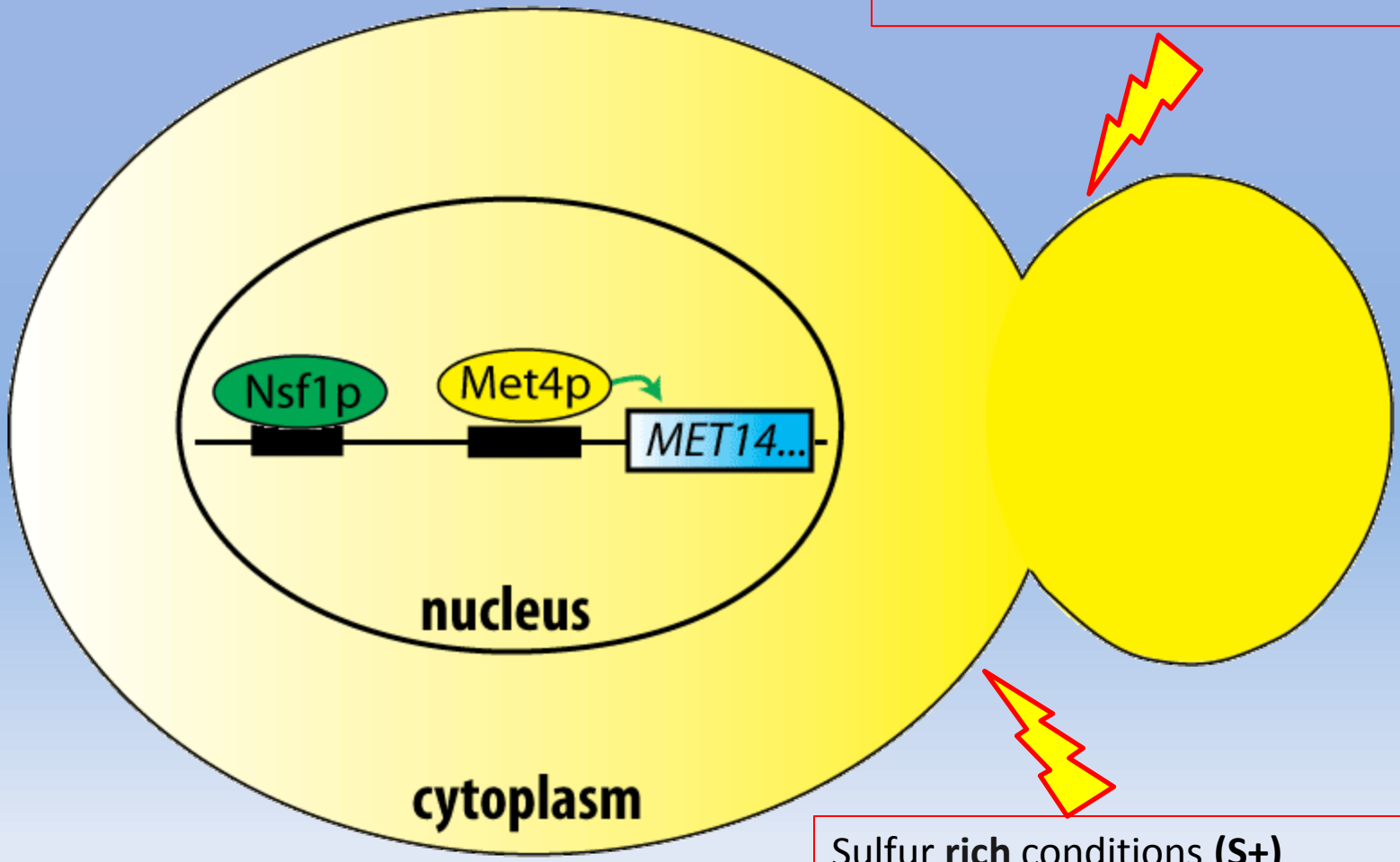


M2*nsf1Δ* / M2 (*nsf1Δ* mutant compared to wild-type):

- Nsf1p functions as a **negative regulator** of some *MET* genes
- Does Nsf1p and Met4p together regulate sulfur metabolism?

Summary of microscopy and RT-PCR results

Sulfur **limiting** conditions **(S-)**



Sulfur **rich** conditions **(S+)**

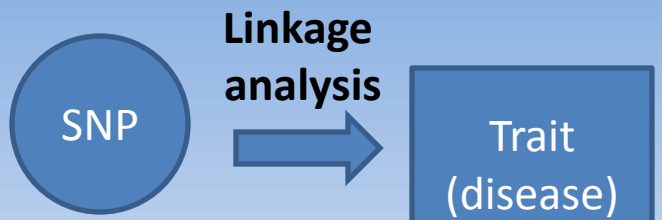
Conclusions

- The genes described in ICC provide a relevant ***NSF1*** **co-expression functional neighbourhood**
- *NSF1* participates in regulation of sulfur metabolism genes
- The results might have industrial applications
- Find **multiple clusters (ICCs)** to build
  - **functional gene networks**
  - the functional networks **based on organism fitness scores** are already available [7]

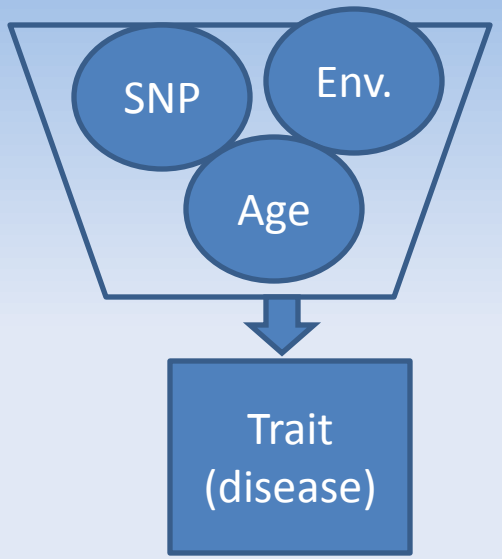
Genome Wide Association Studies (GWAS)

GWAS

- GWAS given markers distributed genome-wide find searches for those that are linked to a trait (e.g. disease, status, disease severity)
- Aims to identify mechanisms driving complex diseases (traits)
- **Single factor model**

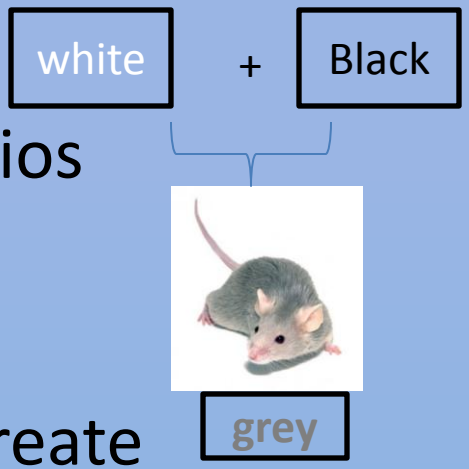


- **Complex disease traits**
  - Considers various factors
    - **GxE:** Gene x Environment factors
    - **GxG:** Gene x Gene analysis



Epistasis definitions

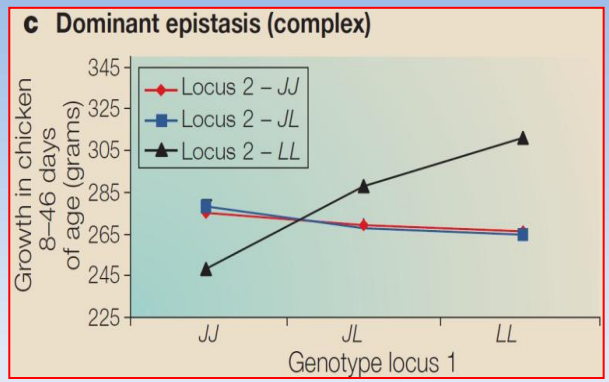
Biological epistasis:



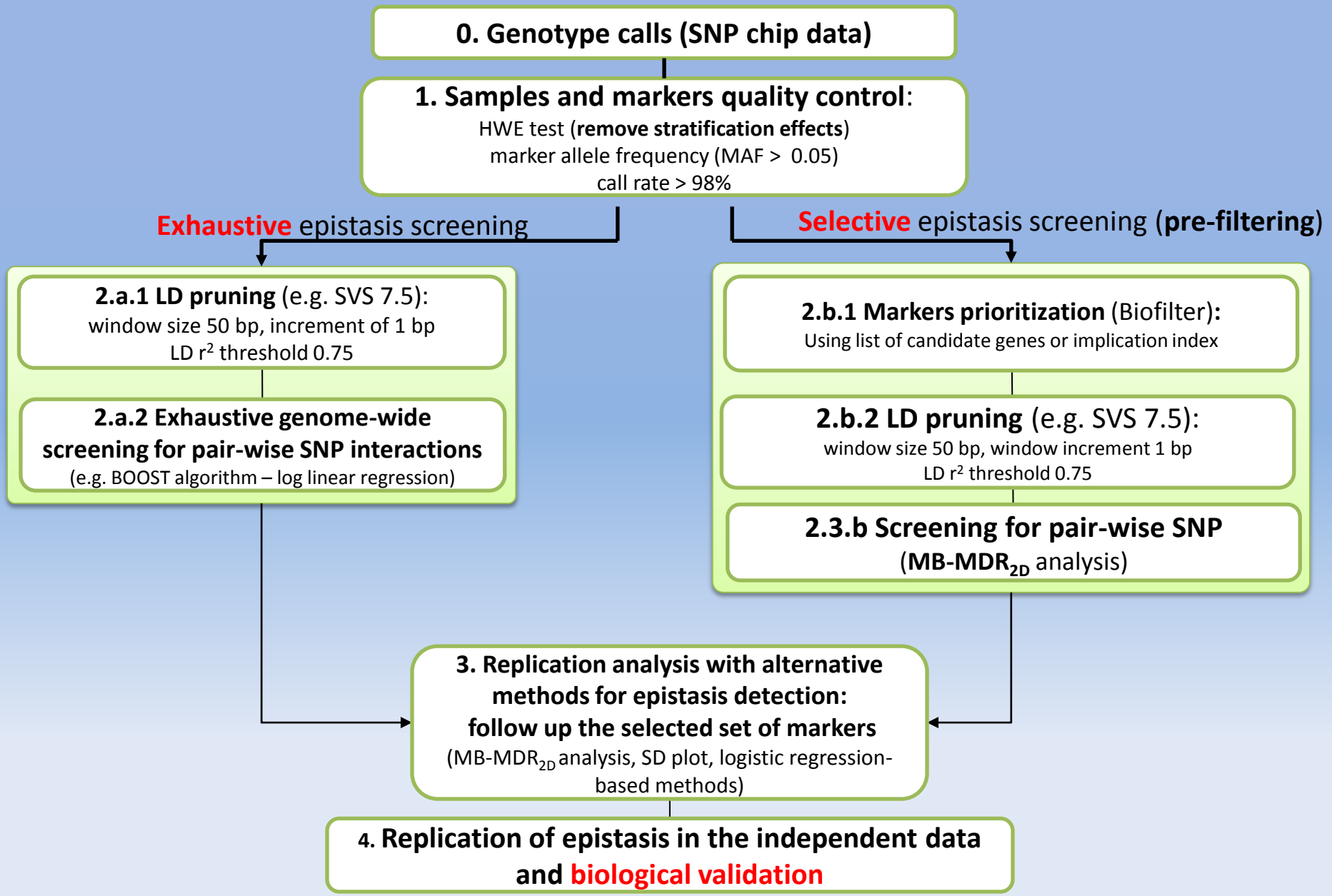
- Distortions of Mendelian segregation ratios (e.g. 1:2:1 and 9:3:3:1) due to one gene masking the effects of another
- Whenever two or more loci interact to create new phenotypes

Statistical epistasis (Fisher, 1918):

- Deviation from **a model** of additive multiple effects for quantitative traits. When two (or more) loci contribute to a single phenotype in additive manner.

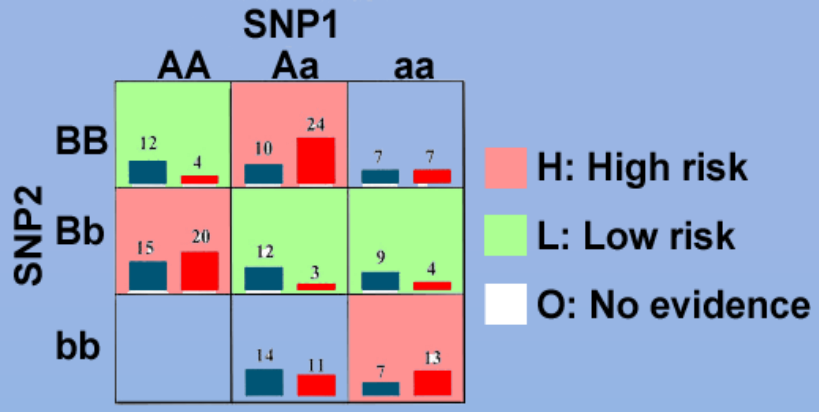


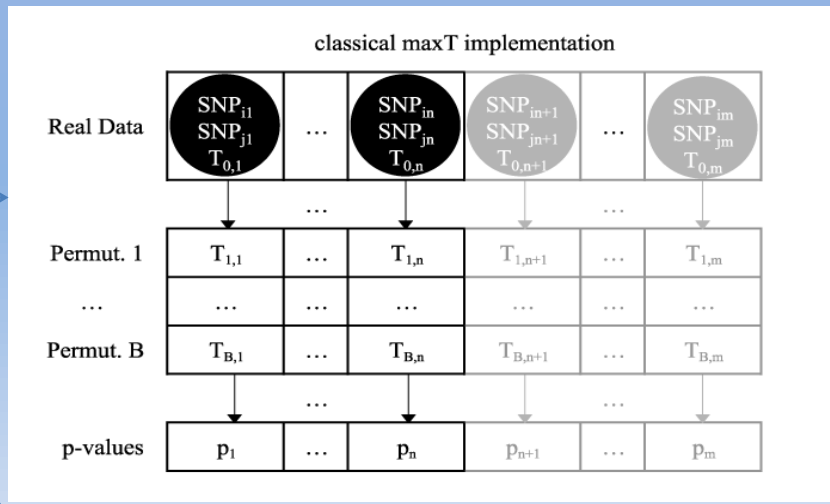
Genome Wide Association Interaction protocol

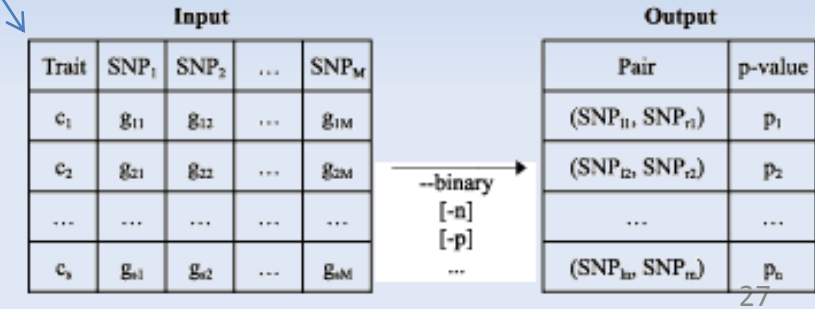


MB-MDR for epistasis detection

- Model-based multifactor dimensionality reduction
  - A tool to screen for pairwise epistatic loci interaction correcting for main effects
- Take two loci and label each cell as {H, L or O} based on cases to controls threshold.
- Compute the X^2 test-statistics between {H} vs {L,O} groups and correct for main effects for all pairs of SNPs ($j = 1…m$).
- Sort X^2 -values where $T_{01} > T_{02} > … > T_{0m}$ based on max T algorithm
- Permute data. Re-calculate X^2 test-statistic
- Compute the multiple-testing adjusted p-values based on distribution of X^2 test-statistic for each pair of SNPs.







Advantages / Limitations of MB-MDR

Advantages	Limitations
• Good control of false positive rates	• Challenging to process more than 100K SNP (computational time)
• Non-parametric. Does not make any assumption about the genetic inheritance model of SNPs	• Requires special file creation (can not accept yet PED or binary files)
• Can handle missing values	• Can not apply yet on GWAS scale only on pre-filtered data
• Runs much faster than original MBR implementation in R	
• Can analyze both binary and continuous trait variables (e.g. gene expression, disease status)	
• Can do GxG and GxE analysis	

Acknowledgements / Lab members

- Kristel van Steen

Principal Investigator

- Elena Gusareva

GWAS workflows

- Jestinah Mahachie

Statistician

- François Lishout

IT/MB-MDR coder

- Bärbel Maus

Statistician

References

- Mahachie John, J, Cattaert, T, Van Lishout, F, Gusareva, E, & Van Steen, K. (2012, January 05). **Lower-Order Effects Adjustment in Quantitative Traits Model-Based Multifactor Dimensionality Reduction**. PLoS ONE.
- Kyrylo Bessonov, Christopher J Walkey, Barry J Shelp, Hennie JJ van Vuuren ,David Chiu, George van der Merwe, **Functional analyses of NSF1 in wine yeast using Interconnected Correlation Clustering and Molecular Analyses**, PLOS One 2012 [submitted under PCOMPBIOL-D-12-01509]
- Marks et. al 2008 **Dynamics of the yeast transcriptome during wine fermentation reveals a novel fermentation stress response**". *FEMS* Yeast Res 8 (2008) 35–52

Merci pour

votre attention