Computational motif discovery in promoters of *Prunus persica* co-regulated genes under various abiotic-stresses

Ksouri N¹, Contreras-Moreira B²,³, Gogorcena Y¹

¹ Department of Pomology, Estación Experimental de Aula Dei-CSIC, Zaragoza, Spain
² Laboratory of Computational and Structural Biology, Department of Genetics and Plant Production, Estación Experimental de Aula Dei CSIC, Zaragoza, Spain
³ Fundación ARAID, Zaragoza, Spain
1. INTRODUCTION

Dra. Yolanda Gogorcena
Lab. Genetics, Genomics and Breeding of Fruits and Grapevine

INAT_TUNISIA

Master degree
(Plant Breeding_IAMZ, Spain)

PhD student
EEAD-CSIC, Spain

Agronomic engineer
INAT_TUNISIA

Dr. Bruno Contreras-Moreira
Laboratory of Computational and Structural Biology
1. INTRODUCTION

Plants regulate their development and stand the environmental changes by turning on the transcription machinery.

Simplified model of transcriptional regulation. Figure adapted from Shlyueva (2014)
2. OBJECTIVES

➢ Testing different clustering algorithms to define the co-expressed genes.

➢ Ab initio motif discovery for detecting the putative transcription factor binding sites.
2. MATERIALS AND METHODS

A. RNA-seq workflow

1. NCBI SRA database
2. Ensembl-plants
3. Prunus persica reference genome
5. Quality control
6. Trimmomatic
7. Pseudo-alignment
8. Kallisto
9. Gene expression profiling
10. Sleuth
11. Gene annotation
12. NCBI-balstn

Tissues: fruit, leaf, stigma and root
Conditions: cold, drought, hyper-hydricity and ripening

- 100 bootstraps
- Wald test Q-value < 0.01 and |b| > 1
- E-value threshold > 10^-5
- Identity percentage > 98.

A. RNA-seq workflow

NCBI SRA database

Ensembl-plants

Prunus persica reference genome

Exp1 Exp2 Exp3 Exp4 Exp5 Exp6 Exp7 Exp8

Quality control

Trimmomatic

Pseudo-alignment

Kallisto

Gene expression profiling

Sleuth

Gene annotation

NCBI-balstn
2. MATERIALS AND METHODS

B. Clustering approaches

K-means clustering
- Elbow method
- Silhouette method
- Gap statistic method

Hierarchical clustering
- Euclidean distance
- Ward.D linkage method

WGCNA analysis

Pairwise correlation

Network construction

Modules analysis

I authorize the use and reproduction of any content, data and digital images in this presentation by CIHEAM-BARI, or anyone authorized by CIHEAM BARI
2. MATERIALS AND METHODS

C. Ab Initio Motif Discovery: RSAT workflow

Motif discovery

(Bianchi et al. 2015)

(Korku, Schippers, and Walther 2014)

50 random modules (negative control)

Motif clustering

Cluster 1

Discovered motif

Familial binding profile

Motif comparison

Logo alignment

footprintDB plants

39 genes
33 genes
22 genes

N modules

500 +200

1500 +200

-500
3. RESULTS AND DISCUSSION

A. Differentially expressed genes (DEGs) screening

<table>
<thead>
<tr>
<th>Project ID</th>
<th>Chilling</th>
<th>Cold</th>
<th>Drought</th>
<th>Hyper-hydricity</th>
<th>Ripening</th>
<th>Nº of DEGs</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRJNA271307</td>
<td></td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td>2601</td>
</tr>
<tr>
<td>PRJNA288567</td>
<td></td>
<td></td>
<td>√</td>
<td></td>
<td></td>
<td>6447</td>
</tr>
<tr>
<td>PRJNA248711</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>PRJEB12334</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>350</td>
</tr>
<tr>
<td>PRJNA252780</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>406</td>
</tr>
<tr>
<td>PRJNA323761</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1118</td>
</tr>
<tr>
<td>PRJNA328435</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2963</td>
</tr>
<tr>
<td>PRJNA397885</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2429</td>
</tr>
</tbody>
</table>

11335 DEGs identified with Wald test (WT) under Q-value < 0.01 and |b| > 1
3. RESULTS AND DISCUSSION

B. Clustering approaches

✓ Module analysis (WGCNA)

Running time of clustering indices and approaches

45 modules (clusters); module size = 20

Gap statistic

Elbow

G from ward.D hierarchical clustering

tterns of 11335 DEGs and 28 samples

Running time of clustering indices and approaches
3. RESULTS AND DISCUSSION

C. Cis regulatory sequences prediction

Gene expression:

**CASE 1**

**CASE 2**

- **From -1500nt to +200nt**
  - Putative TFBM: AIL6
  - Family: AP2/ERF
  - Ncor = 0.69

- **From -500nt to +200nt**
  - Putative TFBM: HYS
  - Family: bZip
  - Ncor = 0.67

- **From -500nt to +200nt**
  - Putative TFBM: HYS
  - Family: bZip
  - Ncor = 0.97
4. Conclusions

- Out of 45 modules, 12 were associated with overrepresented regulatory motifs, classified into 6 families: AP2/ERF (3 motifs), bZip (3 motifs), Myb (3 motifs), CAMTA (1 motif), BES1 (1 motif) and WRKY (1 motif).

- Adapting a convenient approach for contracting gene co-expression network is the first step toward deciphering the complex plant’s regulatory code.

- The upstream sequence length is the main criteria for detecting significant patterns in biological sequences as any variation of promoter length affect considerably the relevance of the potential detected motif.
Thank You

Get in Touch
Najla Ksouri

Estación Experimental de Aula Dei-Consejo Superior de Investigaciones Científicas, Zaragoza, Spain
nksouri@ead.csic.es
+34602129331