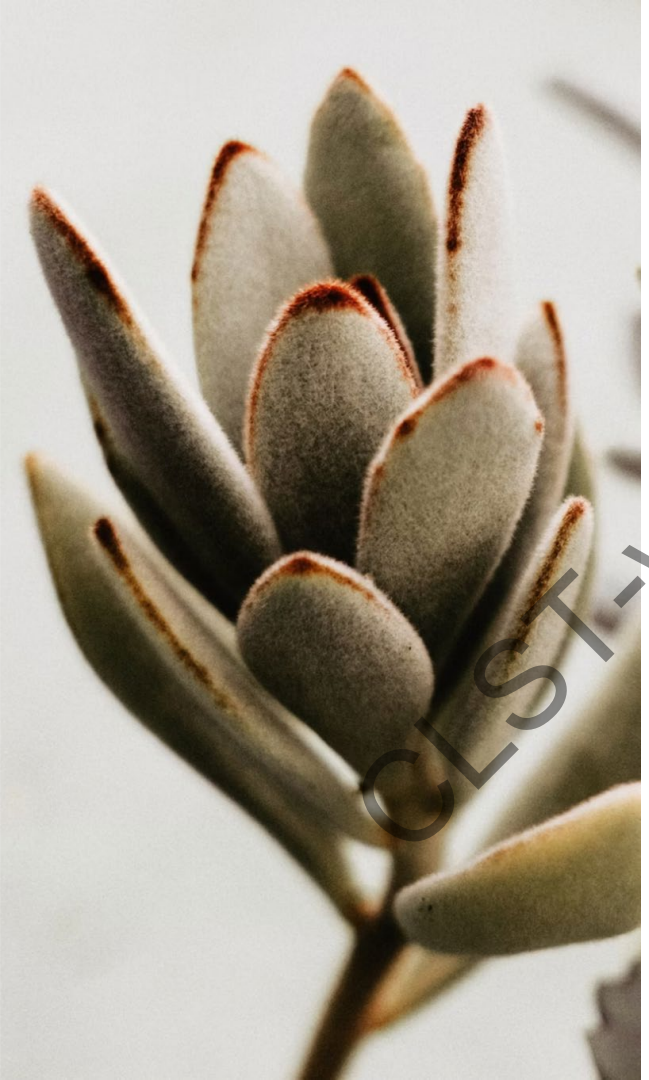
Time-Series Gene Expression Data of Kalanchoe Leaves

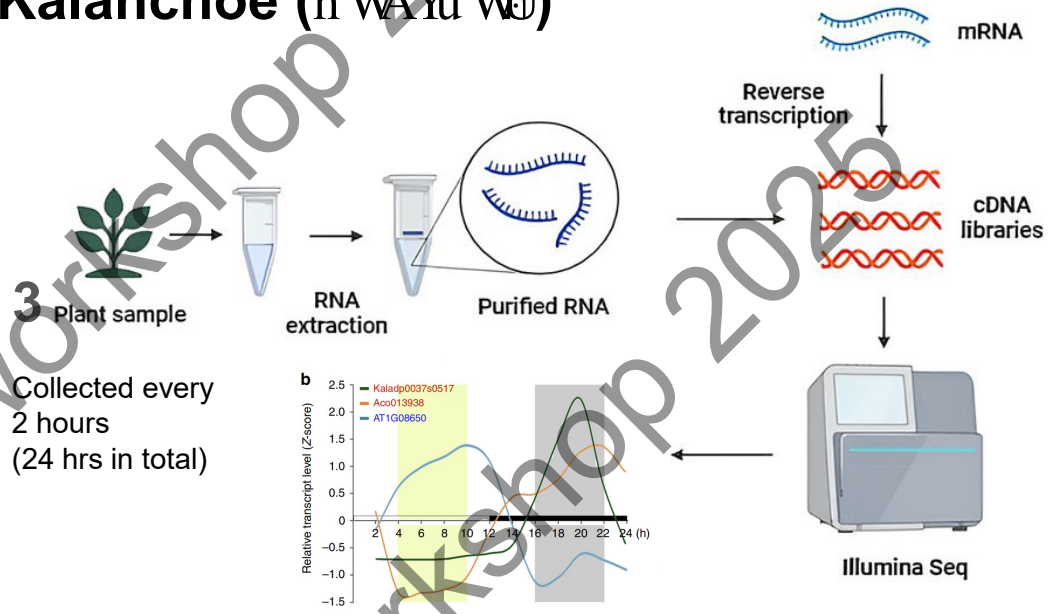
Group 4
Tai, Ning, Kong, Sin, Austin and Inging

CLST winter school 2025

24-26 January 2025



Transcriptomic data from Kalanchoe (ħ ŴÁ Ỳũ Ŵẽ)



Akhtar *et al.*, (2024), Yang *et al.*, (2017)

Factors that affect gene expression

1) CAM plant
2) Circadian rhythm
3) Flowering process
4) Stomatal movement

Objectives

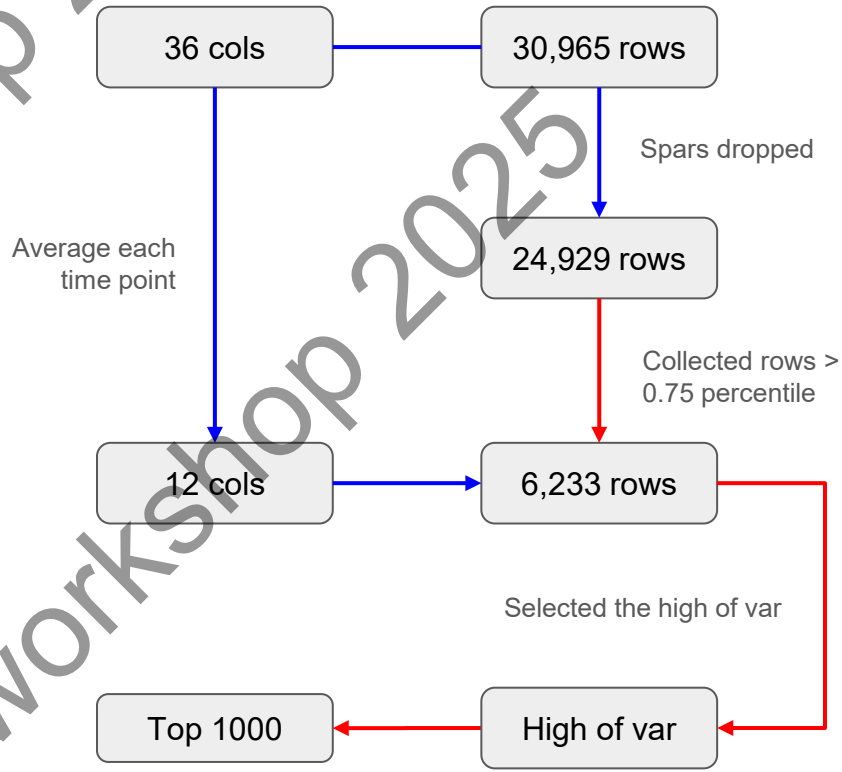
- **To identify the gene expression pattern** of Kalanchoe leaves
- **To cluster gene expression profiles** using K-mean with Shape Based Distance (SBD)
- **To investigate the biological functions** of the identified genes

Data Preprocessing and Feature selection

Original dataset

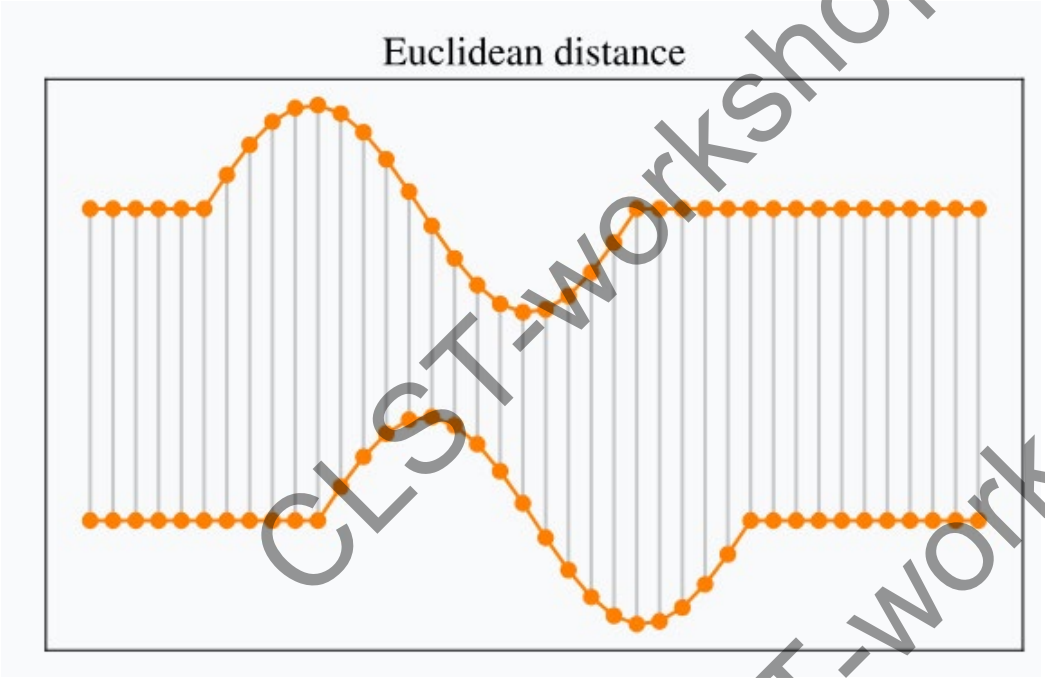
- **36 columns**: T00, T02, …, T22, (12x3 reps)
- **30,965 rows**: transcript genes

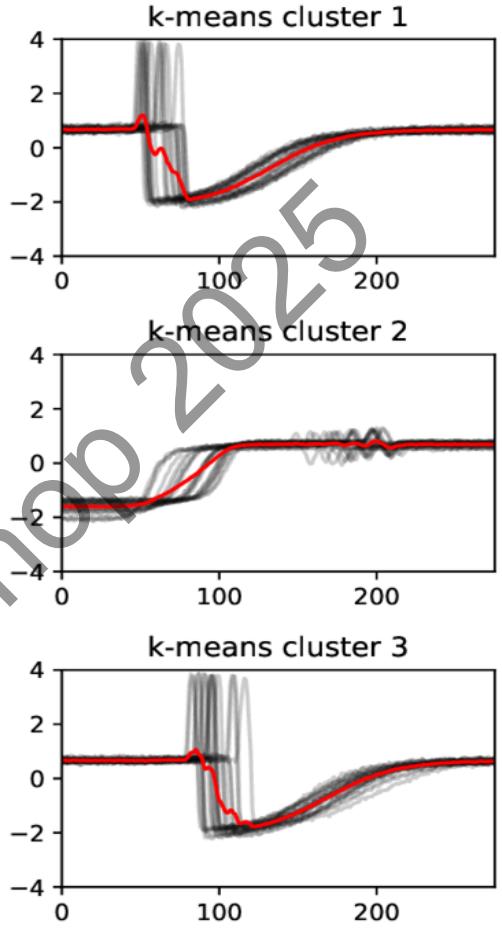
Transcript	Leaf_T08_Rep1	Leaf_T08_Rep2	Leaf_T08_Rep3
Kaladp0498s0001.1	0.803114	1.113	0.482016
Kaladp0011s0858.1	4.32901	6.74984	4.80492
Kaladp0071s0450.1	10.3723	16.1585	15.8322
Kaladp0022s0096.1	0.000109075	0	0
Kaladp0039s0681.1	1.50795	1.86168	2.66518
Kaladp0015s0111.1	13.9264	12.7251	6.23029
Kaladp0040s0680.1	17.561	16.5151	15.8198
Kaladp0046s0151.1	3.00518	4.22122	3.00002
Kaladp0045s0453.1	1.86634	0.539132	0.923603
Kaladp0025s0001.1	0	0	0
Kaladp0977s0008.1	1.06256	0.992991	0.947584



Data preprocessing Feature selection

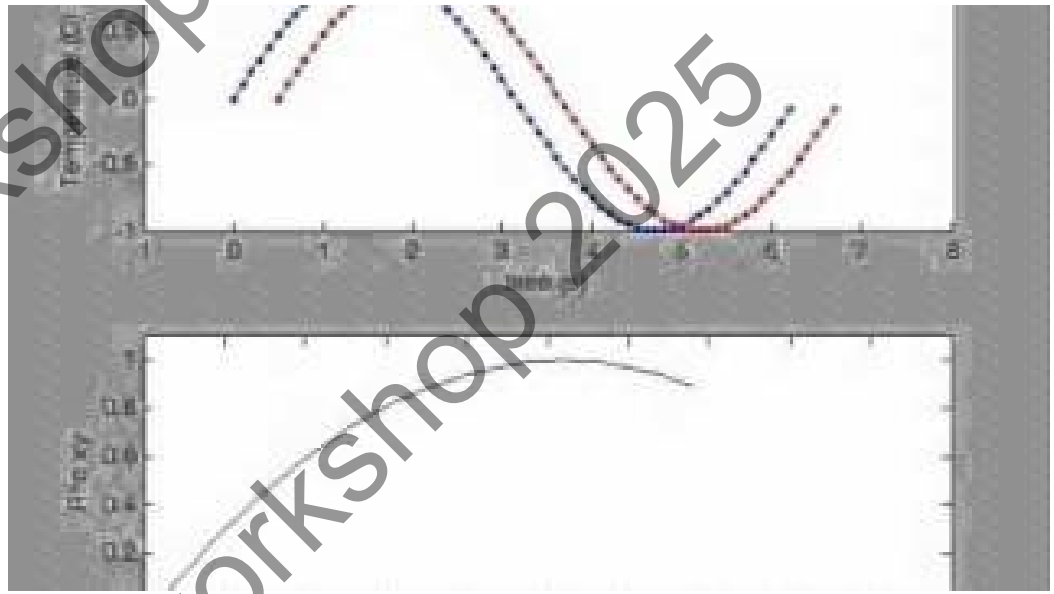
Problem of Euclidean K-Means





Shape-based Distance (SBD)

Cross-correlation is the measurement of how well two independent signals resemble each other

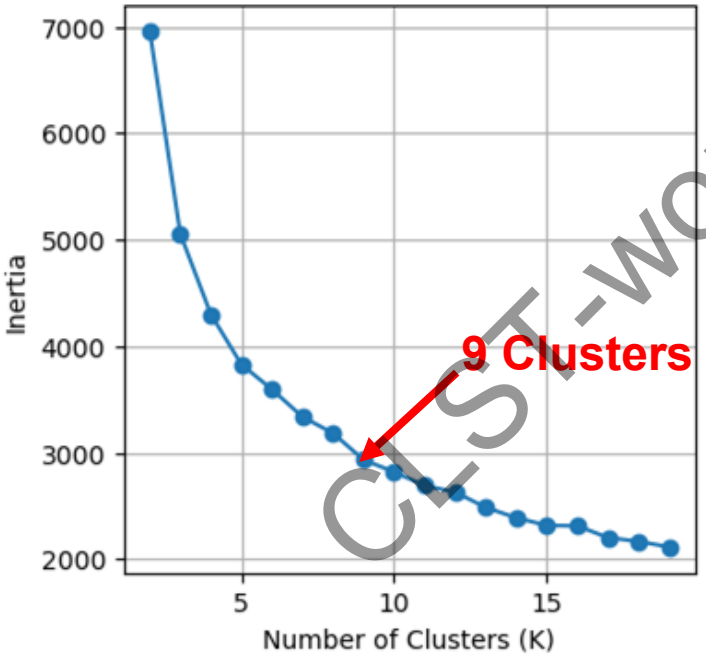


$$(f \star g)(\tau) = \int_{-\infty}^{\infty} f^*(t) \cdot g(t+\tau)\, dt$$

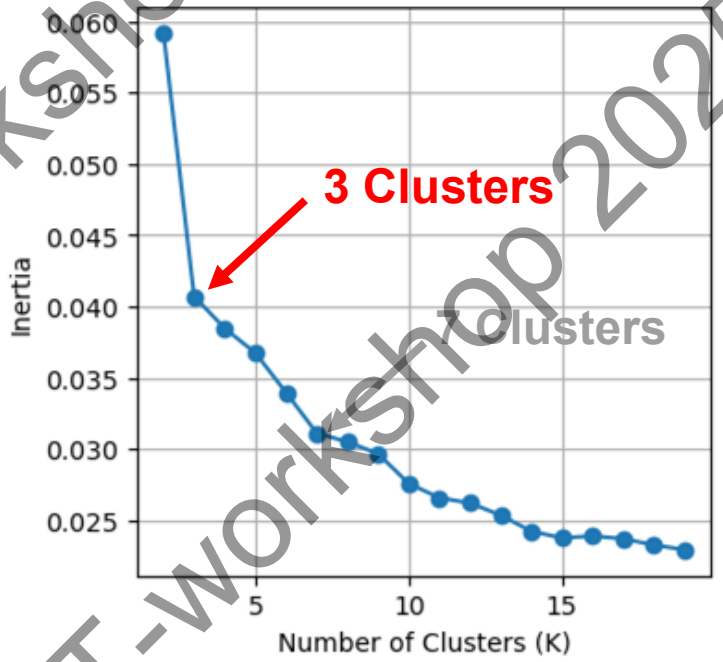
$$\mathrm{SBD}(x, y) = 1 - \max_{w} \left(\frac{\mathrm{CC}_w(x, y)}{\sqrt{\mathrm{R}_0(x, x) \cdot \mathrm{R}_0(y, y)}} \right)$$

Choose Number of Clusters (Elbow Method)

Euclidean Distance (ED)

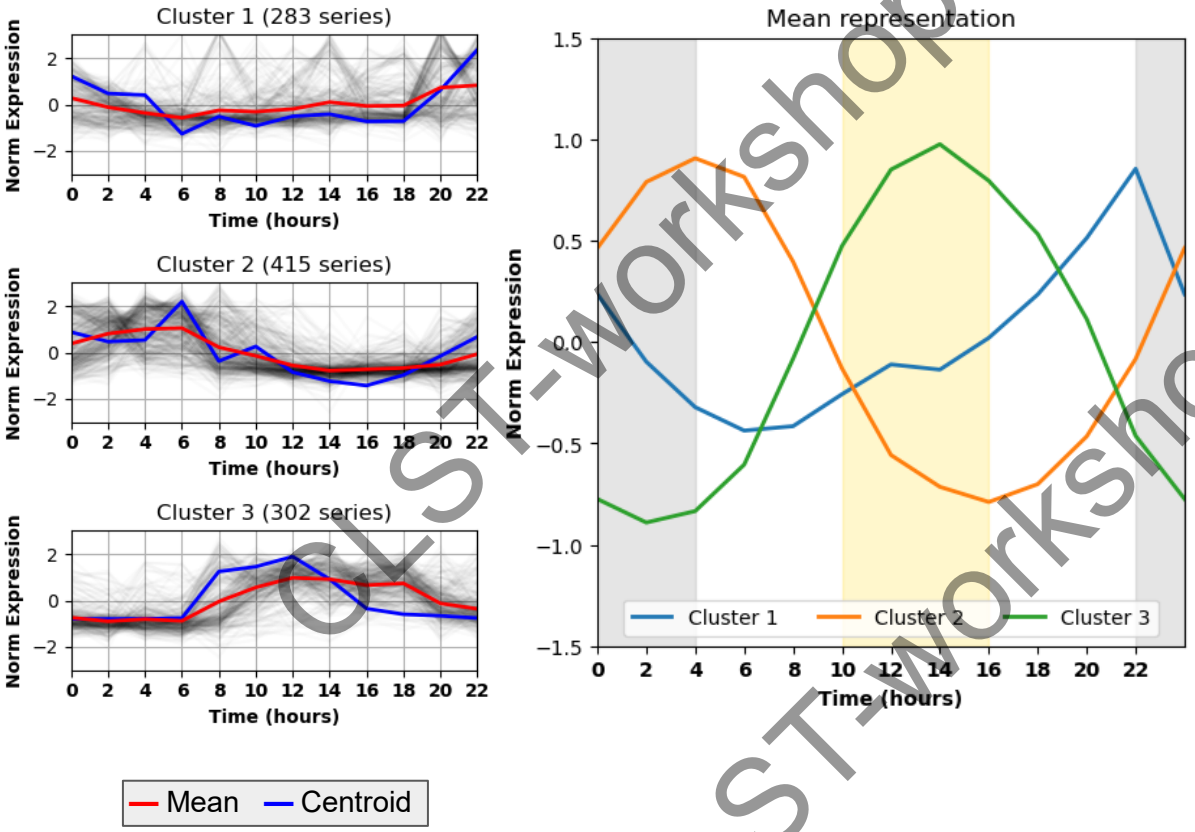


Shape-based Distance (SBD)

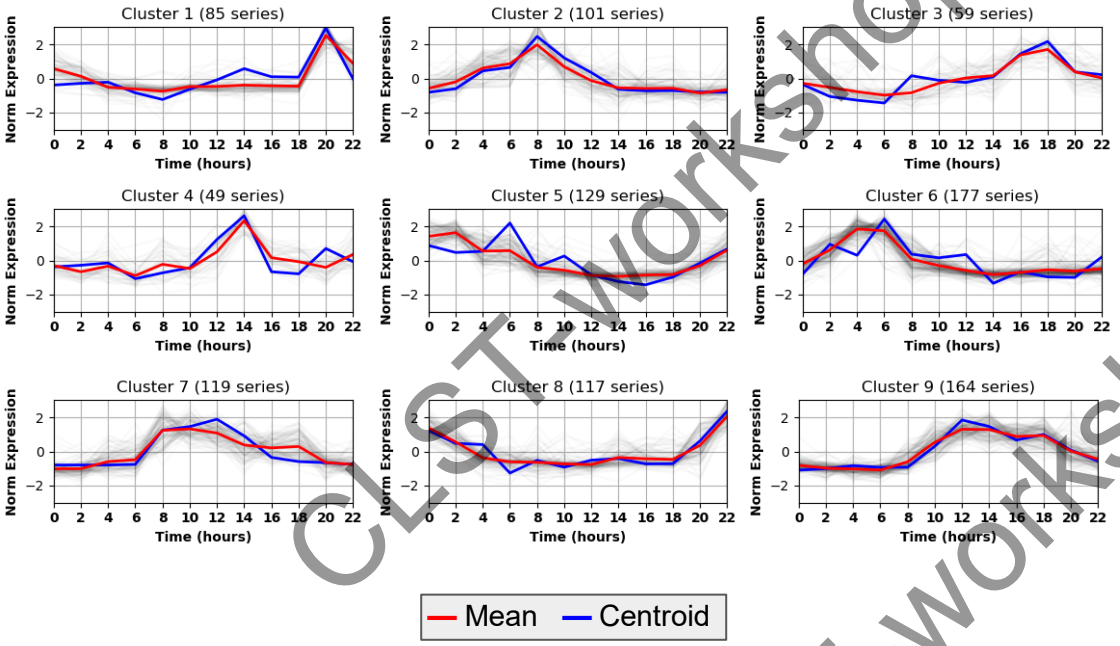


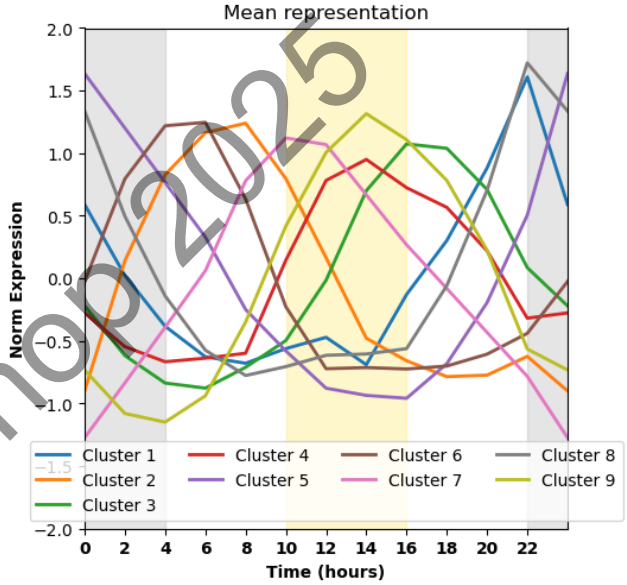
Inertia = Sum of squared distances of samples to their closest cluster center.

Clustering (SBD)



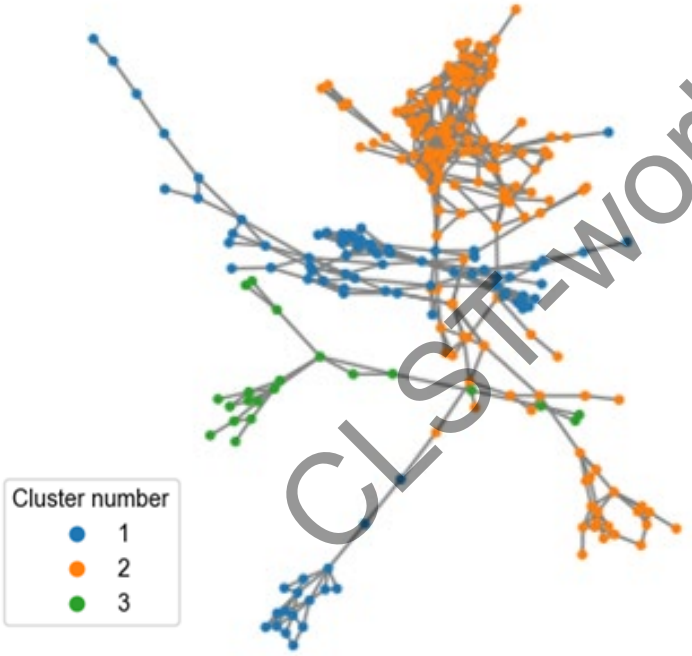
Clustering (ED)



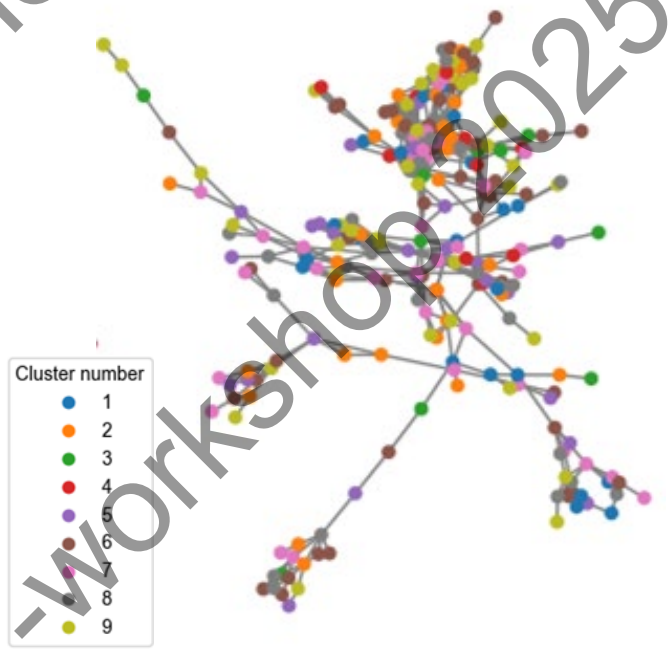


Network Construction

SBD (3 clusters)

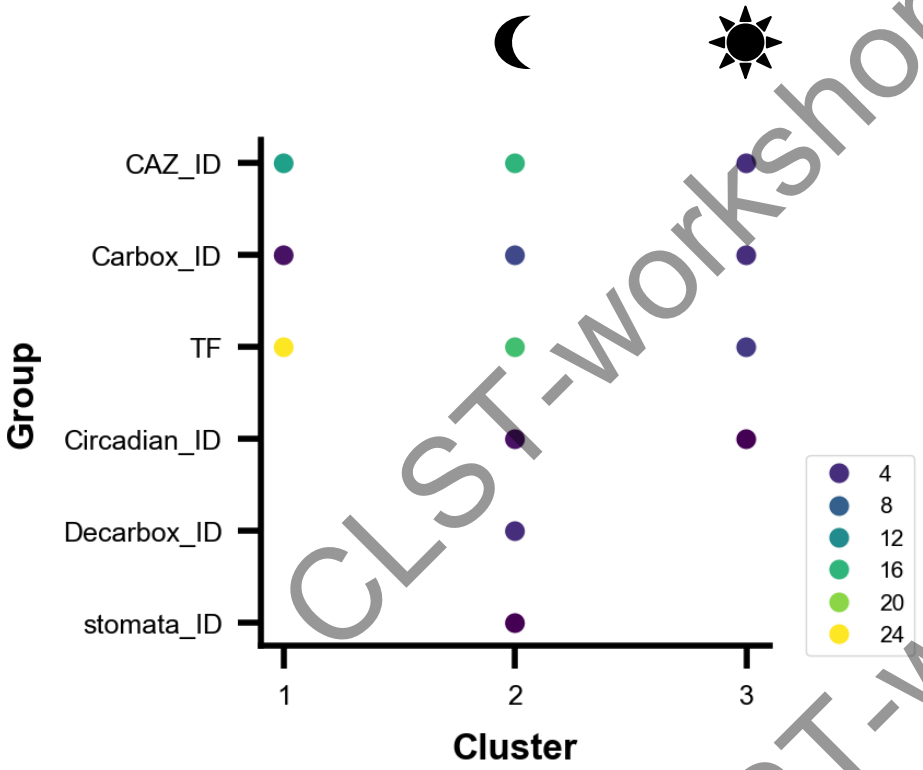


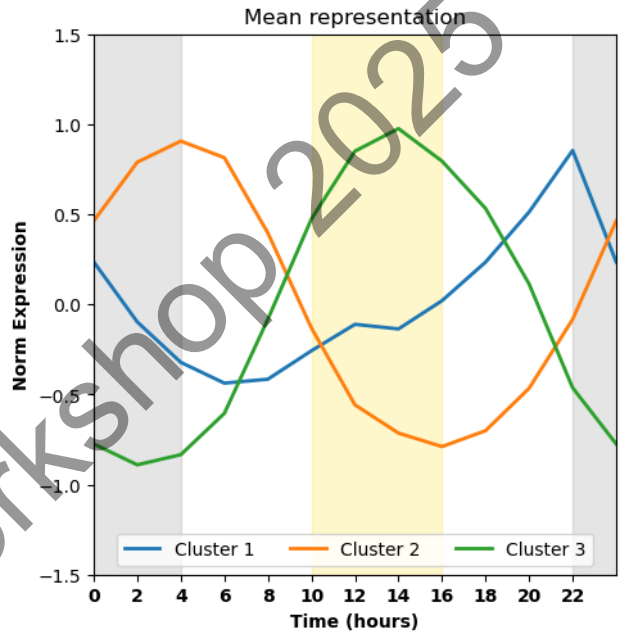
ED (9 clusters)



Pearson correlation on expression level between different transcripts (cutoff 0.98)

Temporal Dynamics of Gene Modules

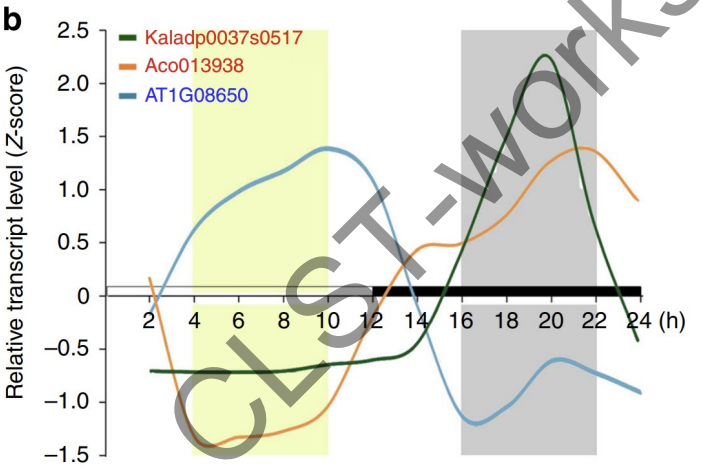




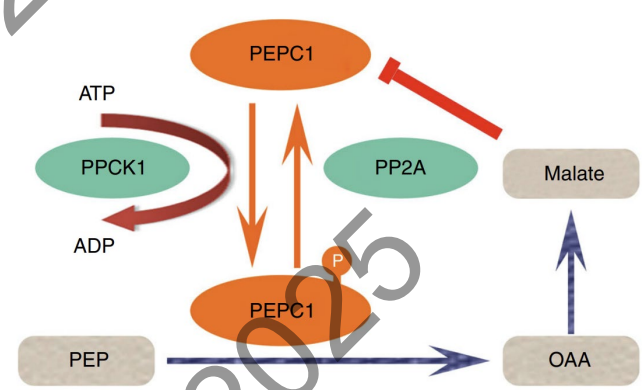
Gene Function

CO2 Fixation

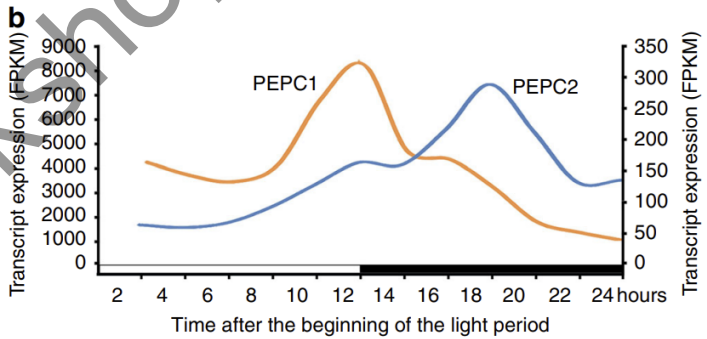
- PEPC, PPCK
- CAM fixes CO2 during night time



Green line depicts PPCK1 diel expression, enriched in midnight



Two phosphoenolpyruvate carboxylase (PEPC) Pathway

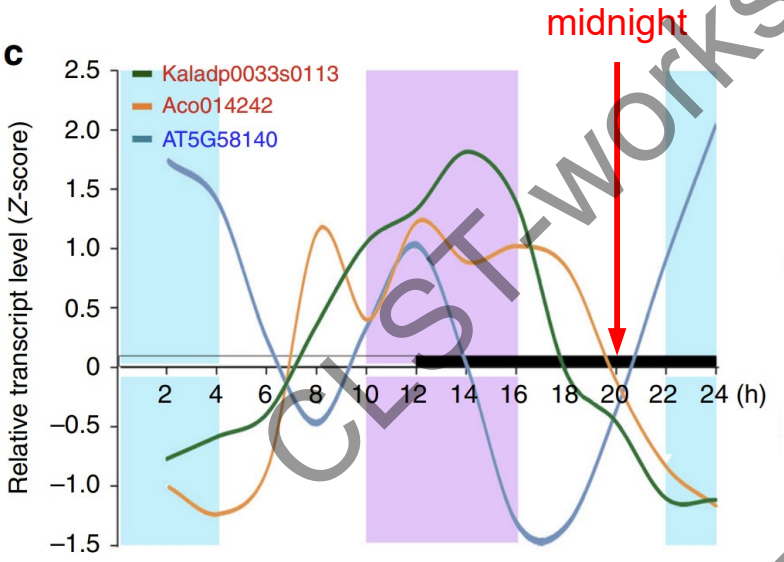


PEPC1 and PEPC2 diel expression

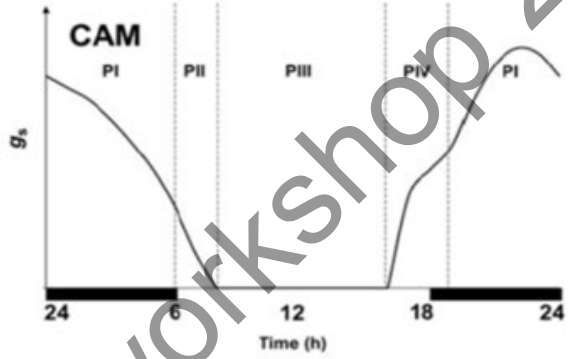
Gene Function

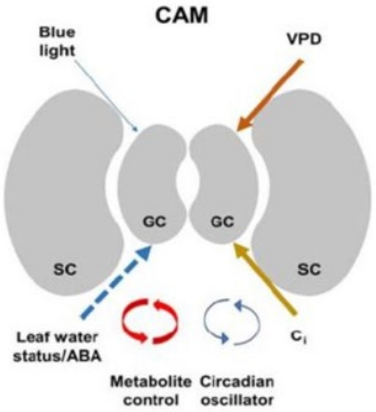
Stomatal Movement

- Response to blue light
- phototropin 2 (PHOT2)



Green line depicts PHOT1 diel expression, enriched in dusk





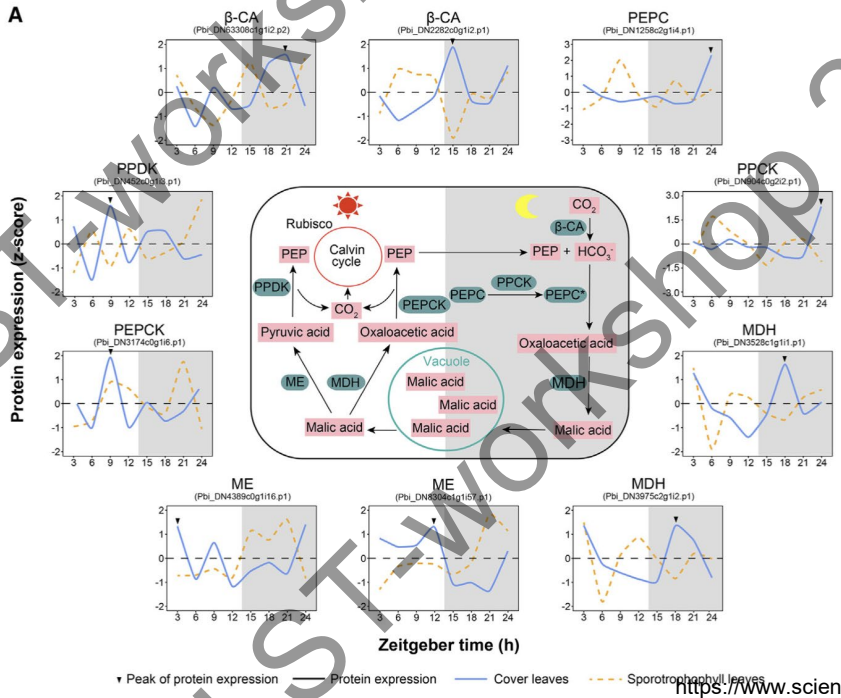
CAM stomatal opening

https://plantae.org/update-stomatal-biology-of-cam-plants/

Gene Function

Carbohydrate metabolism

- CAZymes: regulates carbohydrate synthesis, metabolism, and transport including 6 sub classes: glycoside hydrolases (GHs), glycosyltransferases (GTs), polysaccharide lyases, carbohydrate esterases, auxiliary activities, and carbohydrate-binding modules.



https://www.sciencedirect.com/science/article/pii/S2590346223001050

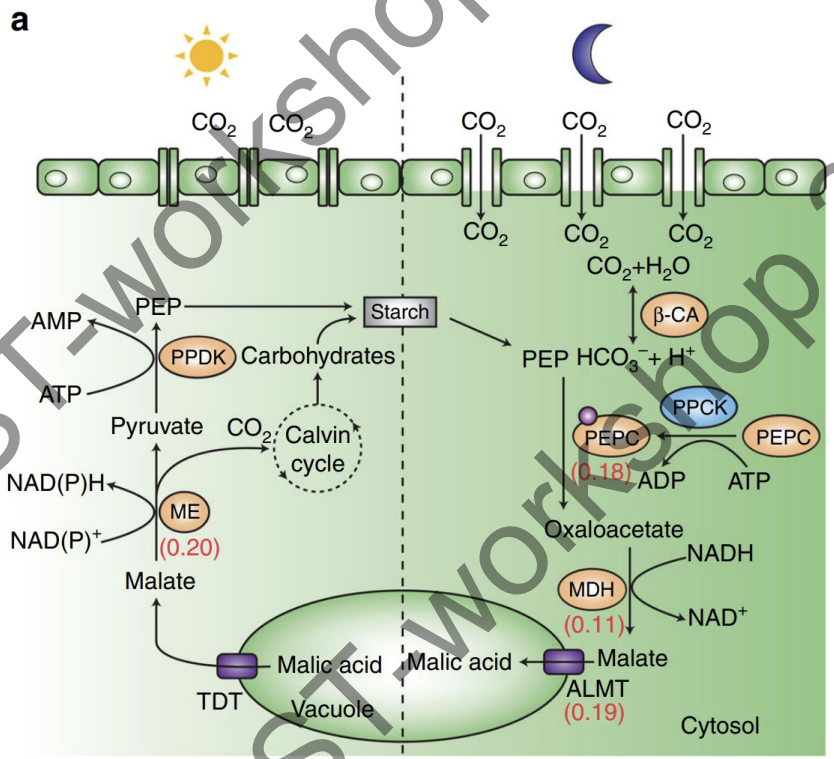
Conclusions

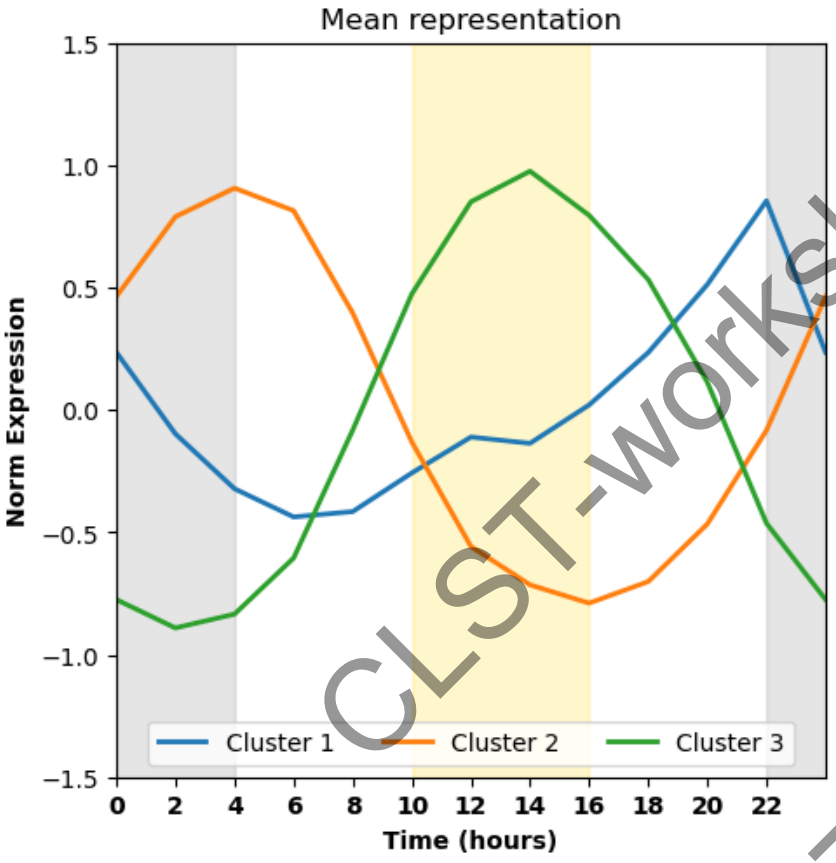
- We can **find and group the gene expression patterns** of Kalanchoe leaves using K-mean with Shape Based Distance (SBD)
  - 3 Clusters (Day, Night, Late Night)
- We can **explain the biological functions** of the identified genes related with the time series
  - Night
    - Carbon fixation – active during the night
    - Stramal movement – active during the night
  - 24-hour
    - Carbohydrate metabolism

Supplementary (?) figures

CLST-workshop 2025

Pathway





Cluster 1 (Late evening)

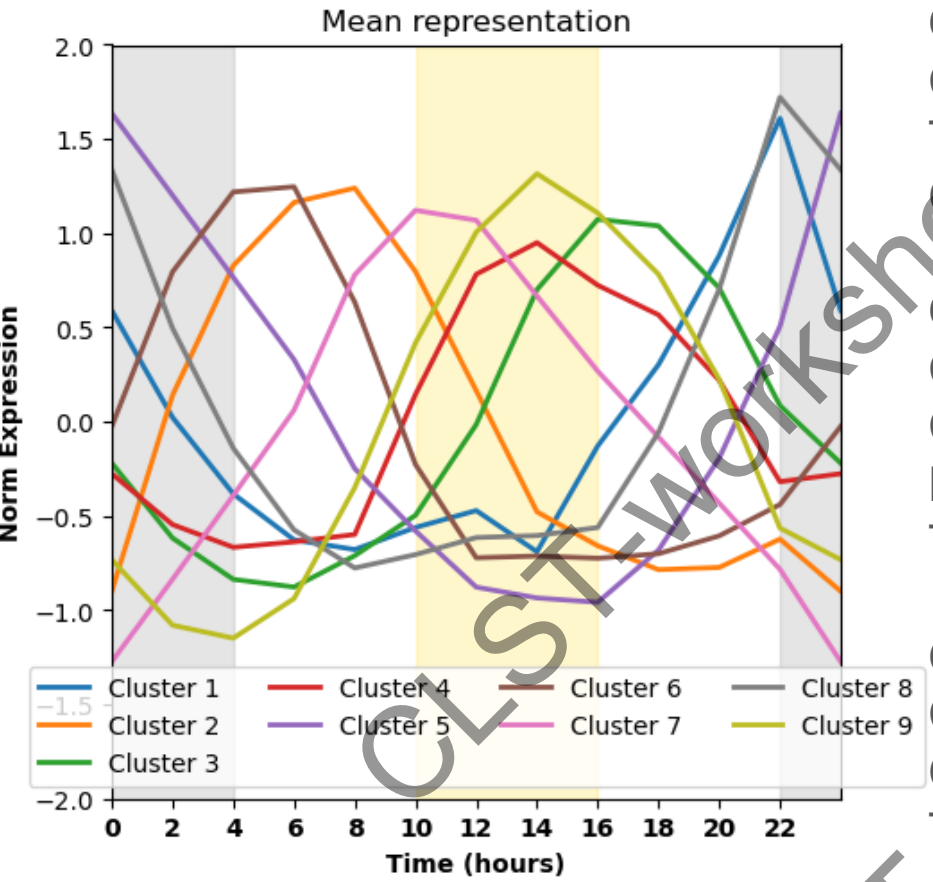
CAZ_ID (AA2, CE8, Cellulose_synt, GH19, GH9, GT1, GT2, GT75)
Carbox_ID (MDH, PEPC1)
TF (AP2/ERF-ERF, B3-ARF, C2C2-GATA, C3H, CAMTA, NAC, OFP, Tify, WRKY, bHLH)

Cluster 2 (Night)

CAZ_ID (AA1, CE13, CE8, Cellulose_synt, GH1, GH14, GT4, GT5)
Carbox_ID (NADP-ME, PEPC2, PPCK1, Î²-CA)
Circadian_ID (RVE1, RVE8)
Decarbox_ID (NADP-ME, PPDK, TDT)
TF (AP2/ERF-ERF, BES1, C2C2-CO-like, DBB, MADS-MIKC, MYB-related, PLATZ, Tify, bZIP)
stomata_ID (OST2/AHA1)

Cluster 3 (Day)

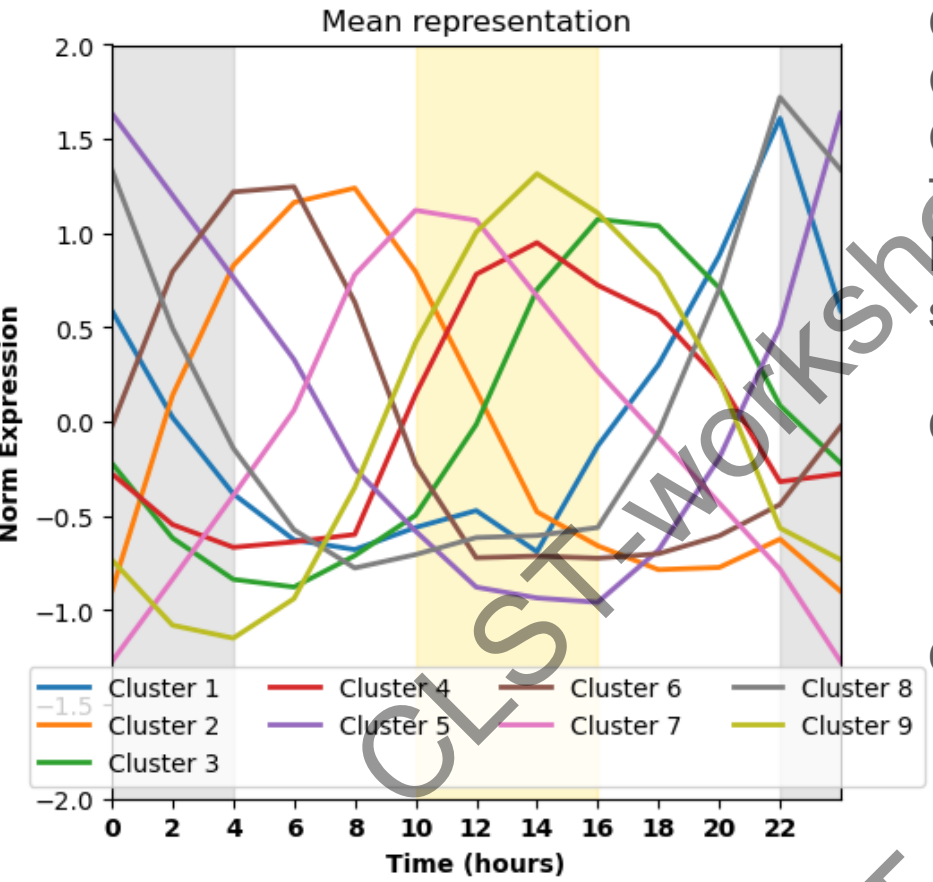
CAZ_ID (CBM45,GH13, GH29, GT35)
Carbox_ID (ALMT6, MDH, PPCK2)
Circadian_ID (ELF4)
TF (AP2/ERF-ERF, C2C2-CO-like, HB-HD-ZIP, HB-KNOX)



Cluster 1
CAZ_ID (GT1)
TF (AP2/ERF-ERF, B3-ARF, C2C2-GATA, C3H, CAMTA, NAC, Tify, WRKY, bHLH)

Cluster 2
CAZ_ID (GT5)
Carbox_ID (Î²-CA)
Decarbox_ID (PPDK, TDT)
TF (AP2/ERF-ERF,DBB)

Cluster 3
CAZ_ID (AA2,CBM45,GH13,GT35)
Carbox_ID (MDH, PEPC1,PPCK2)
TF (C2C2-CO-like,OFP)



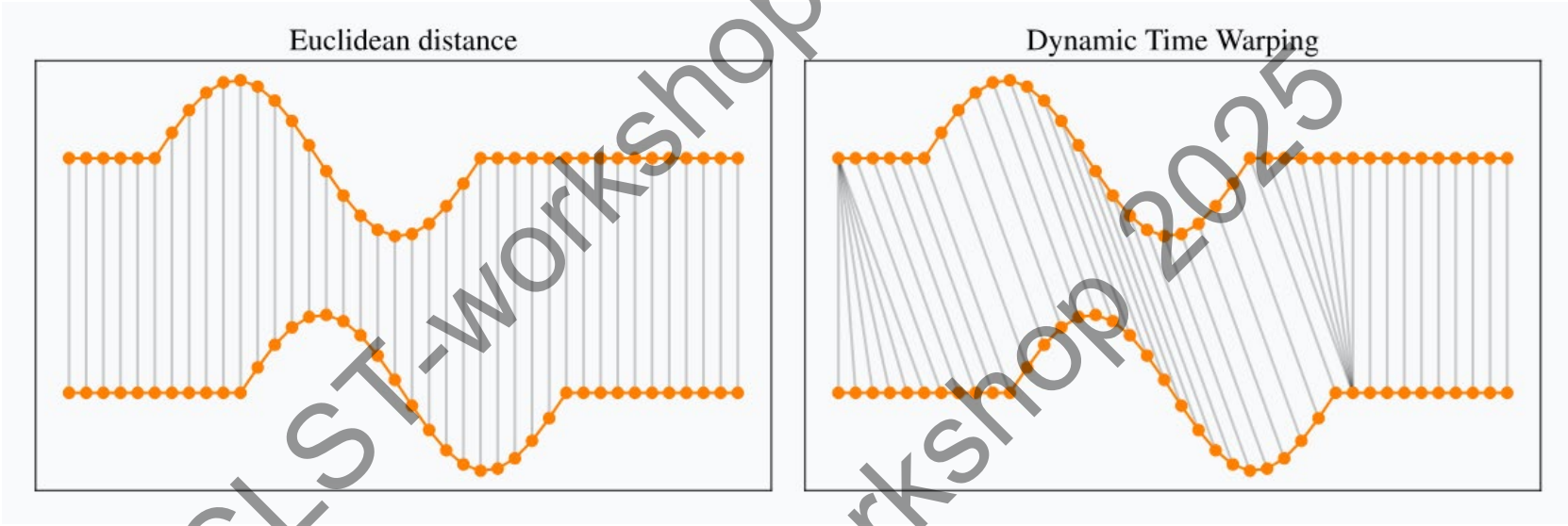
Cluster 5

CAZ_ID (CE13, Cellulose_synt, GH14)
Carbox_ID (PPCK1, Î²-CA)
TF (AP2/ERF-ERF, MYB-related, PLATZ,Tify,bZIP)
stomata_ID (OST2/AHA1)

Cluster 6

Cluster

Dynamic Time Warping (DTW)



DTW finds the optimal alignment between two sequences by minimizing the **cumulative distance** between them.

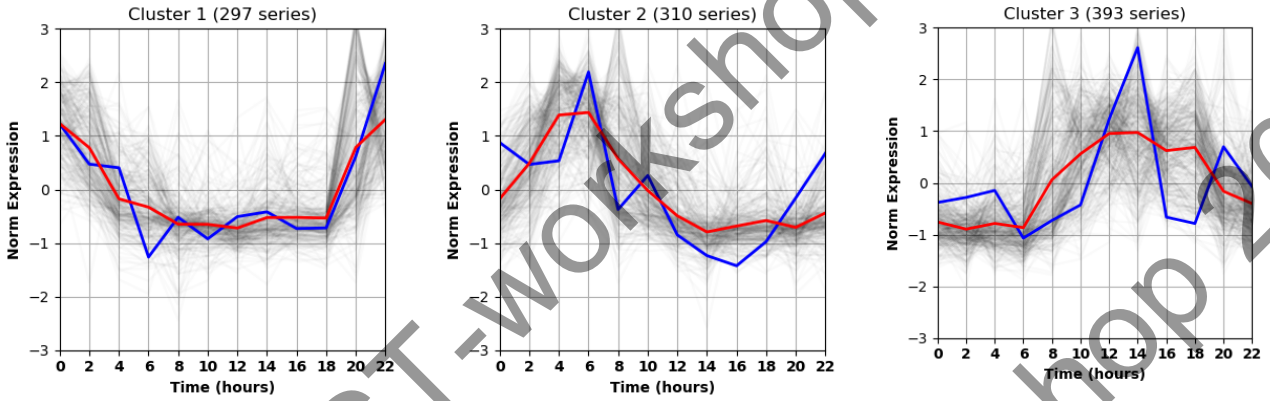
Euclidean

$$D(i,j) = d(i,j) + \min \begin{cases} D(i-1,j), \\ D(i,j-1), \\ D(i-1,j-1). \end{cases}$$

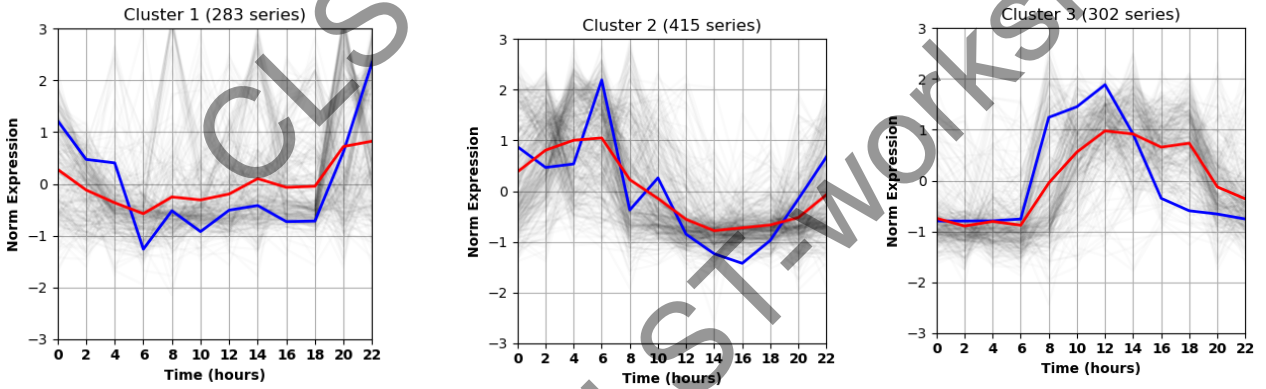
DTW vs SBD

	Dynamic-Time Warping	**Shape-based Distance**
Focus	Local Similarity	Global Similarity
Invariance	Distortions	Scaling and shifting
Alignment	Explicit	Implicit (Cross-correlation)
Cost	High (Dynamic Programming)	Faster than DTW

Euclidean vs SBD



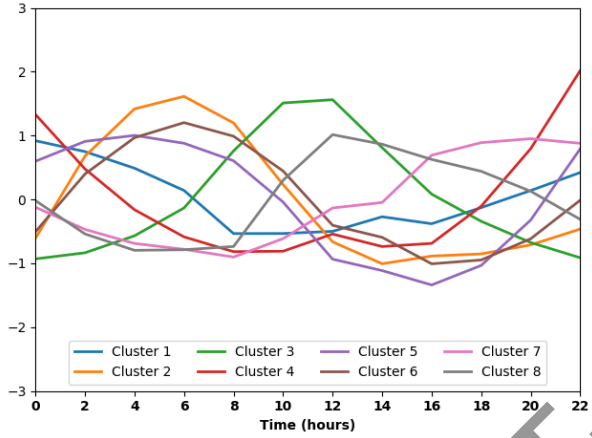
ED



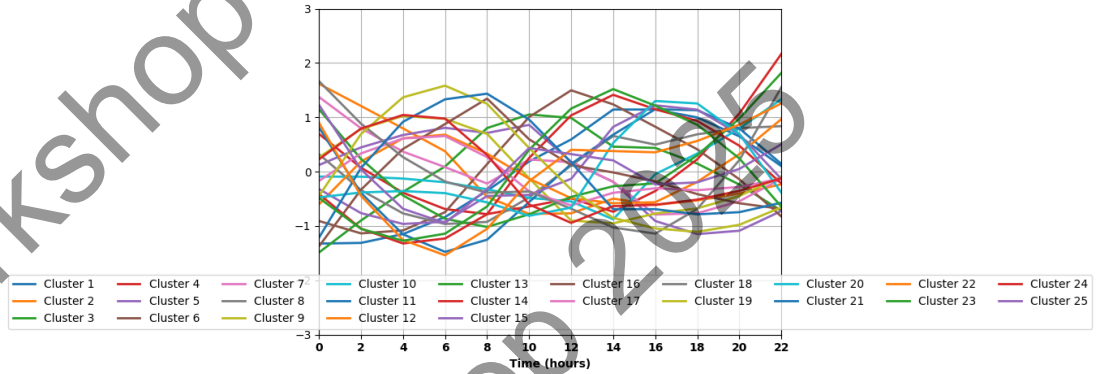
SBD

For compare only

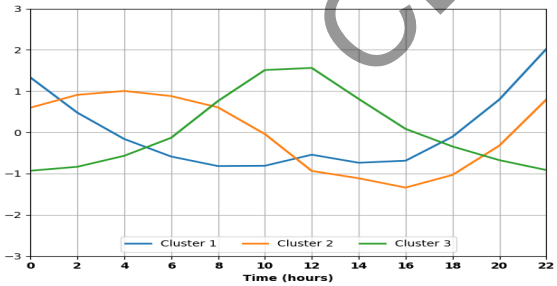
KMeanDTW_1K_8Cluster



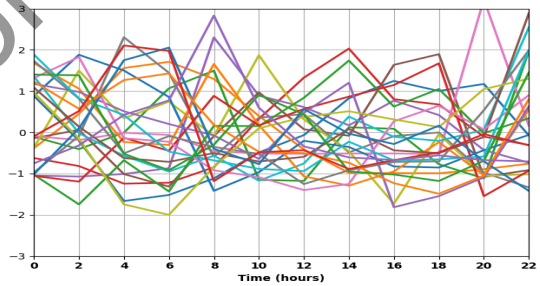
KMeanDTW_6K_25Cluster

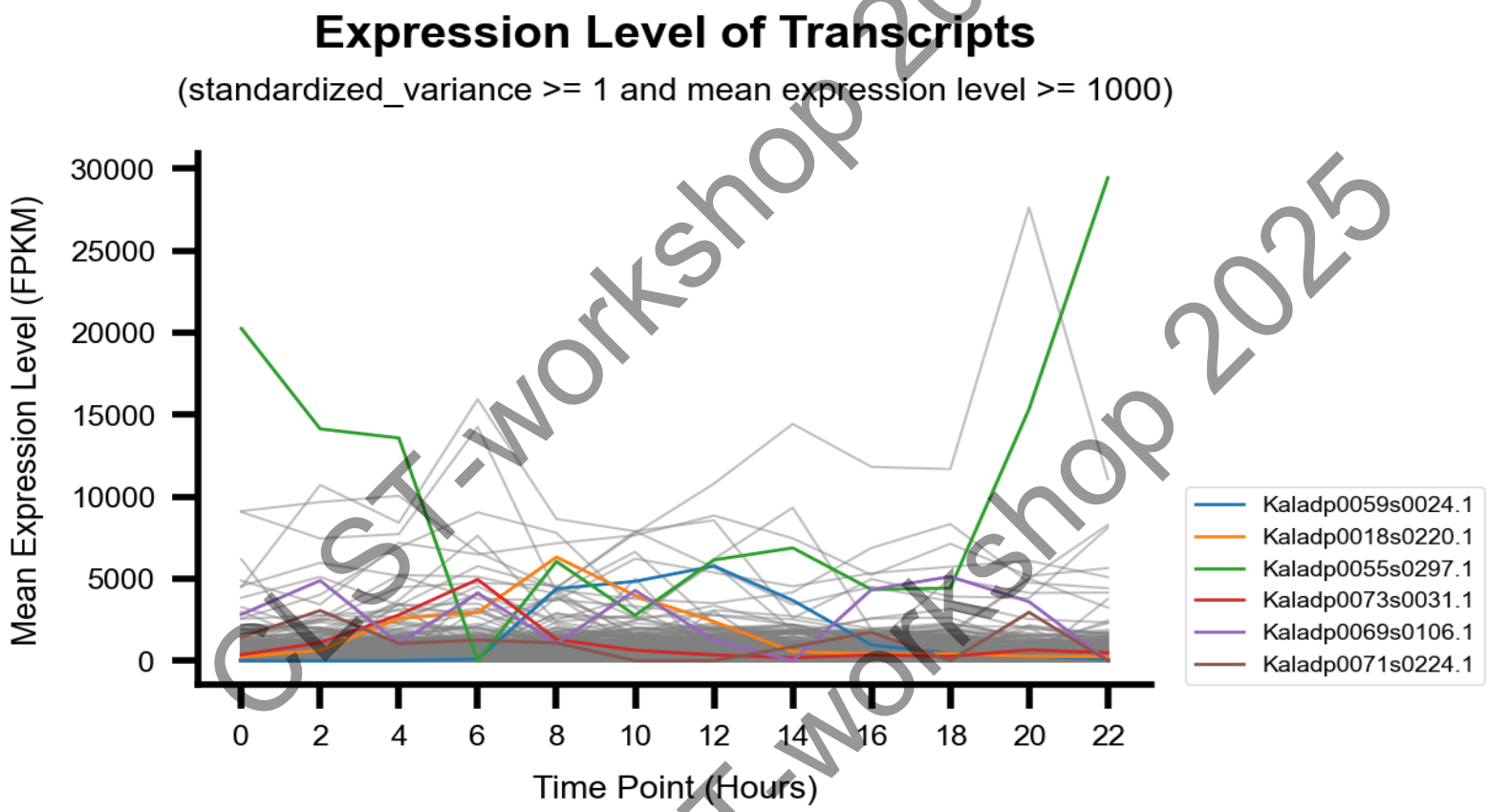


KShape_1K_3Cluster



KShape_6K_25Cluster





Kaladp0059s0024.1 - chlorophyll A/B binding protein
https://phytozome-next.jgi.doe.gov/phytomine/report.do?id=771740824

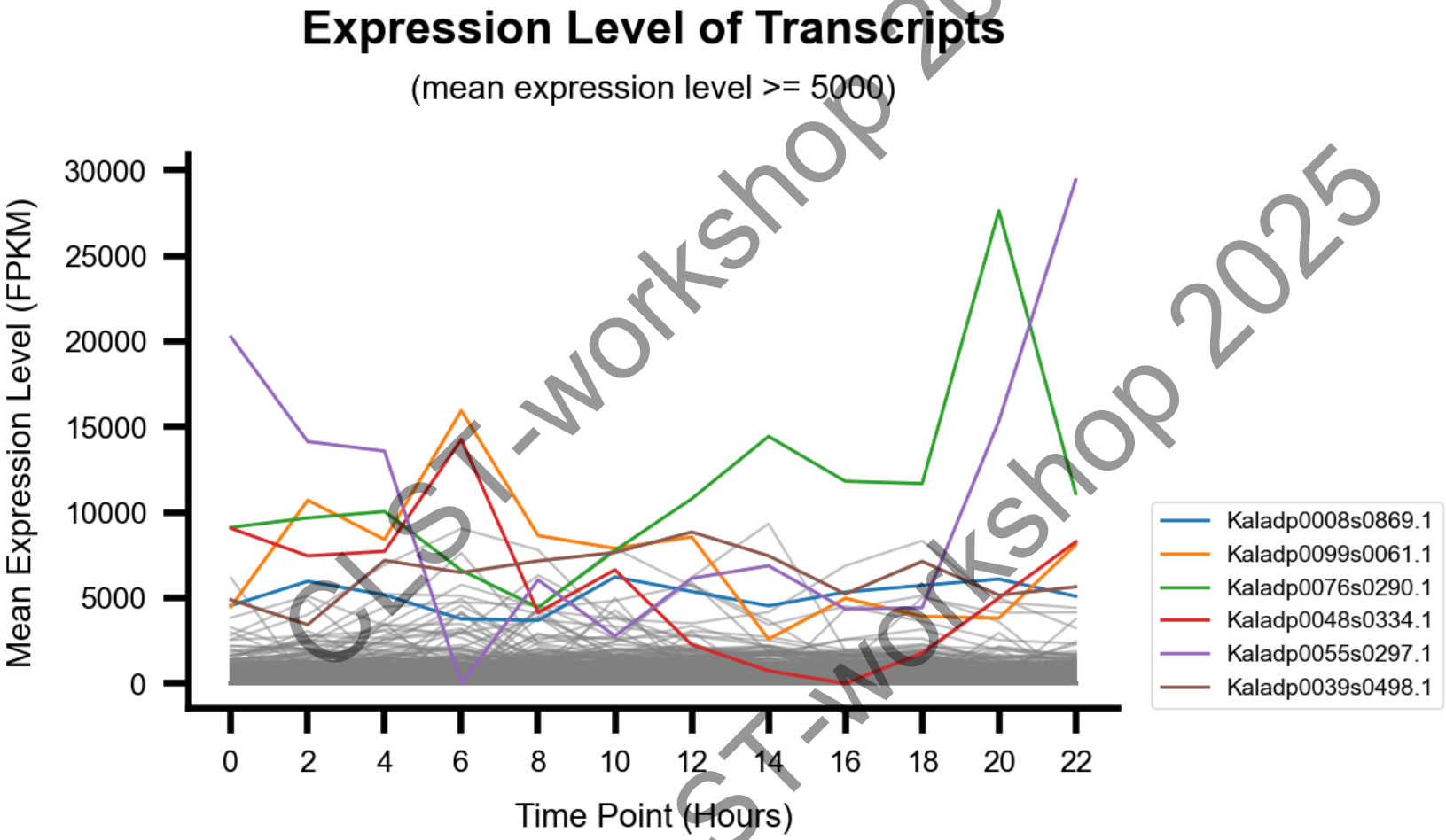
Kaladp0018s0220.1 - Fructose-bisphosphate aldolase (glycolysis)
https://phytozome-next.jgi.doe.gov/phytomine/report.do?id=772102882

Kaladp0055s0297.1 - 5-methyltetrahydropteroyltriglutamate--homocysteine S-methyltransferase / Tetrahydropteroylglutamate-homocysteine transmethylase
https://phytozome-next.jgi.doe.gov/phytomine/report.do?id=771650717

Kaladp0073s0031.1 - S-adenosyl-l-methionine decarboxylase leader peptide
https://phytozome-next.jgi.doe.gov/phytomine/report.do?id=772100337

Kaladp0069s0106.1 - unknown
https://phytozome-next.jgi.doe.gov/phytomine/report.do?id=772005627

Kaladp0071s0224.1 - unknown
https://phytozome-next.jgi.doe.gov/phytomine/report.do?id=771944717



Kaladp0008s0869.1 - **Metallothionein, a metal binding protein -> also respond to stress (?)**
https://phytozome-next.jgi.doe.gov/phytomine/report.do?id=772294356

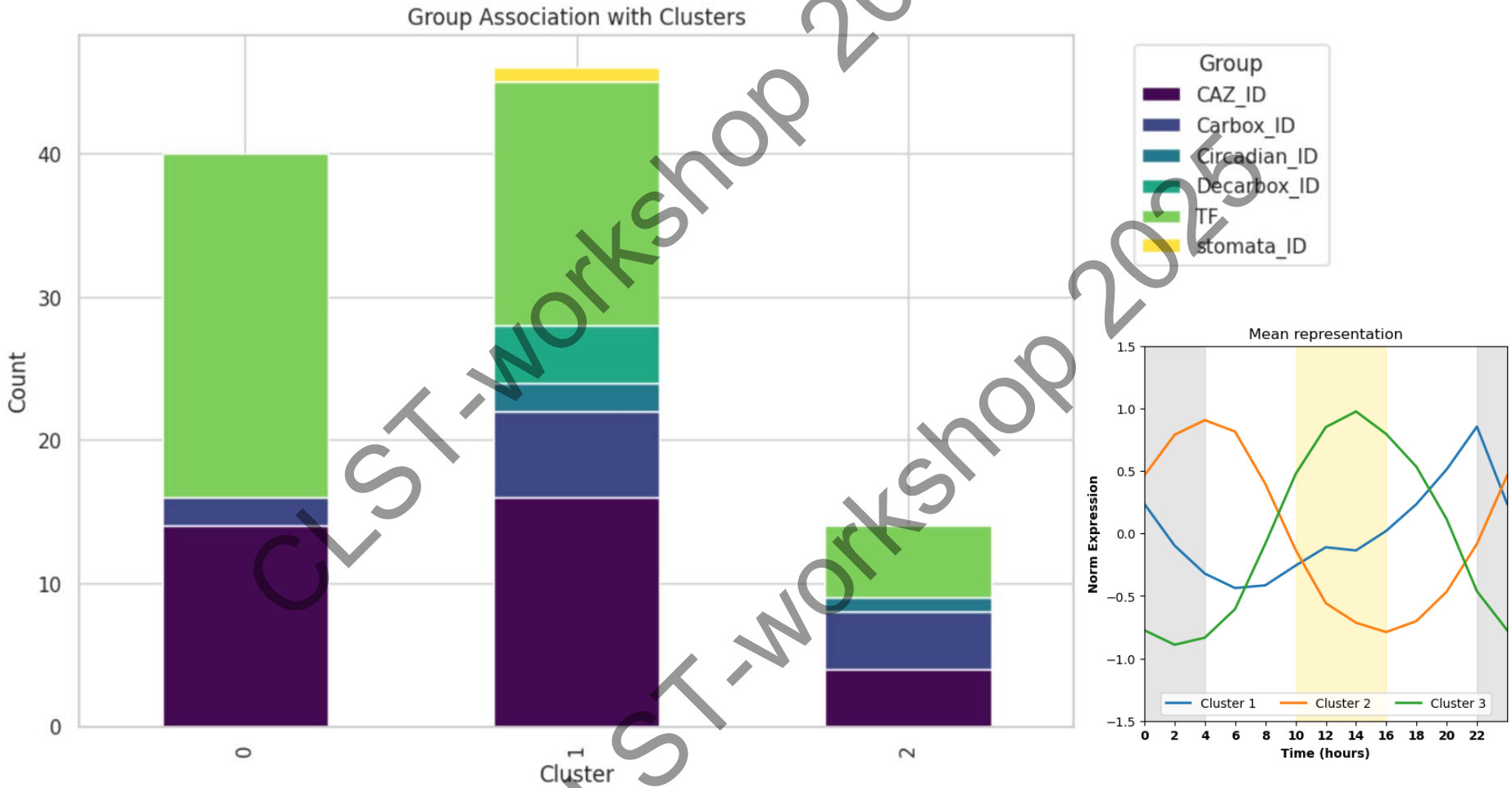
Kaladp0099s0061.1 - **unknown**
https://phytozome-next.jgi.doe.gov/phytomine/report.do?id=772619069

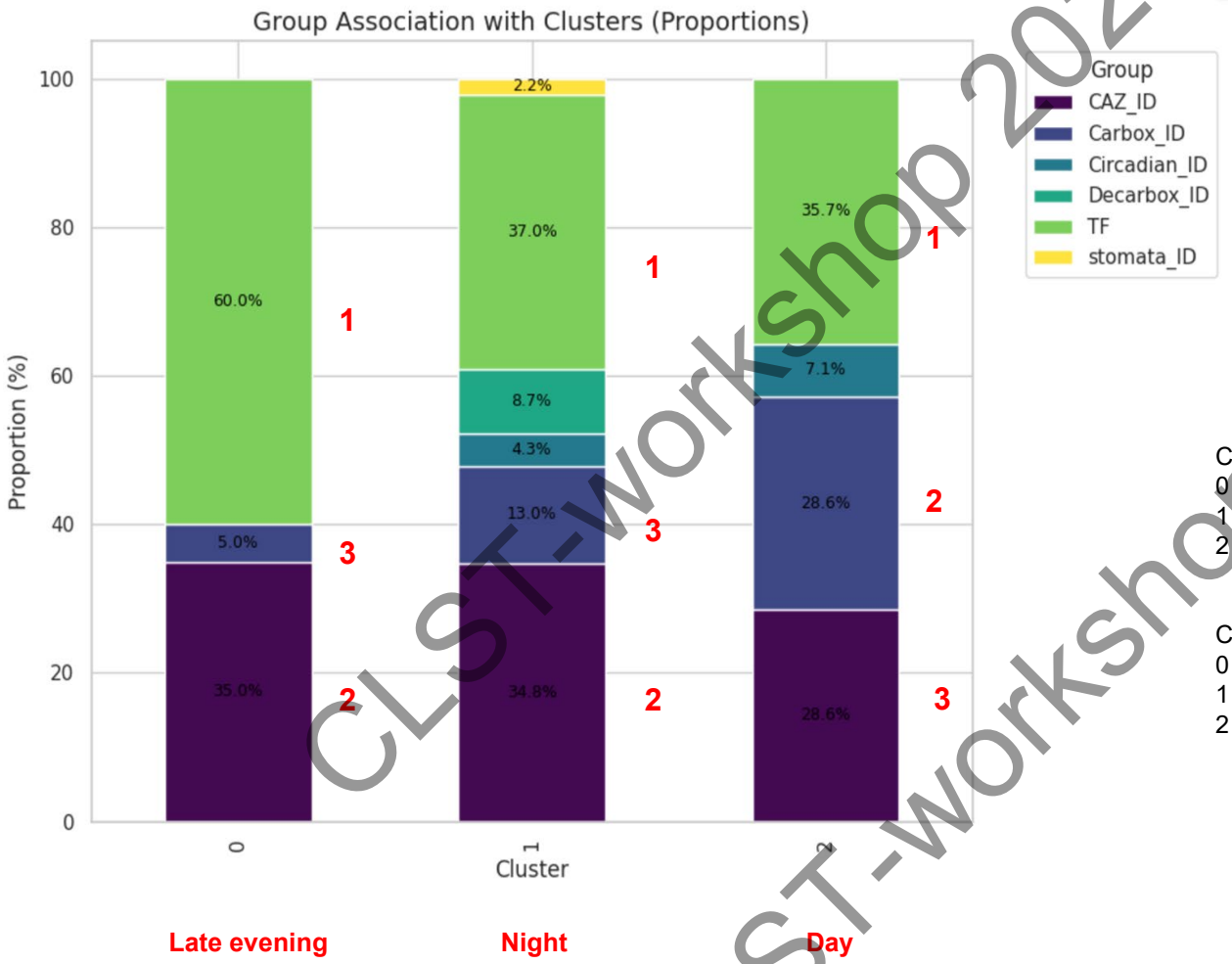
Kaladp0076s0290.1 - **Late embryogenesis abundant protein (LEA_3) -> response to stress**
https://phytozome-next.jgi.doe.gov/phytomine/report.do?id=772025521

Kaladp0048s0334.1 - **unknown**
https://phytozome-next.jgi.doe.gov/phytomine/report.do?id=771668982

Kaladp0055s0297.1 - 5-methyltetrahydropteroyltriglutamate--homocysteine S-methyltransferase / Tetrahydropteroylglutamate-homocysteine transmethylase -> Amino acid (methionine biosynthesis)
https://phytozome-next.jgi.doe.gov/phytomine/report.do?id=771650717

Kaladp0039s0498.1 - **ribulose-bisphosphate carboxylase small chain (rbcS)**
https://phytozome-next.jgi.doe.gov/phytomine/report.do?id=771841183



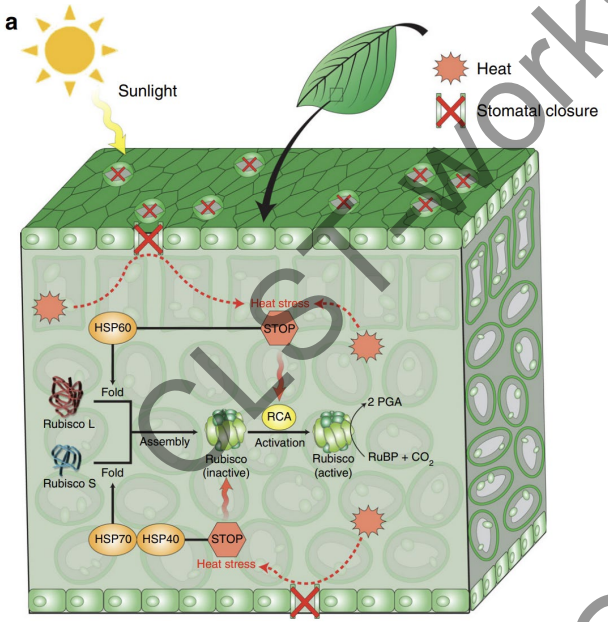


Cluster Counts Before Removing NaN:
0 283
1 418
2 302

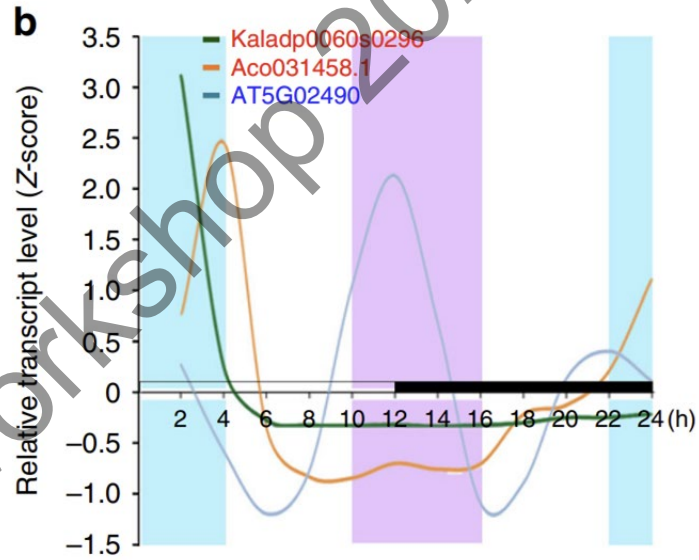
Cluster Counts After Removing NaN:
0 40
1 46
2 14

Gene Function

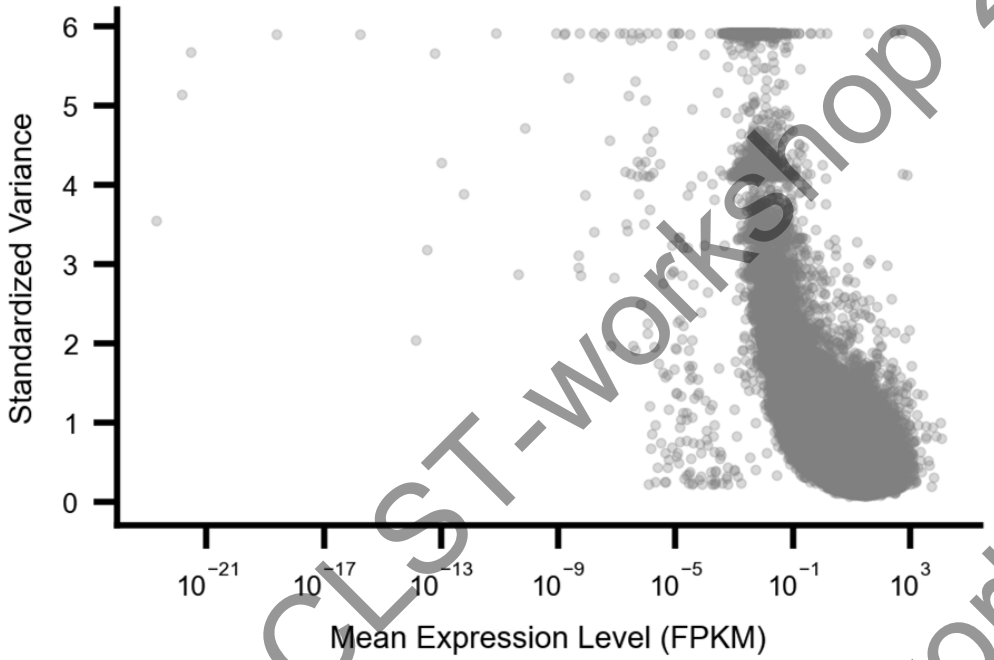
4. Heat Tolerance
   - Stomata close during light period may cause internal heat load on the leaves -> HSP stabilizes protein during heat stress
   - HSP60, HSP70, HSP40



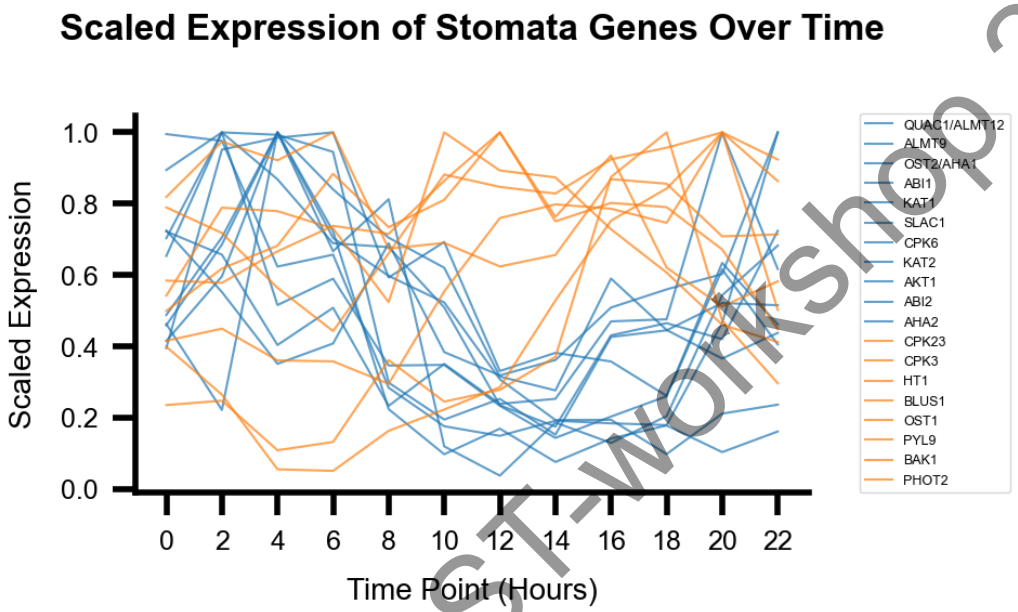
Mechanisms of HSP in heat tolerance

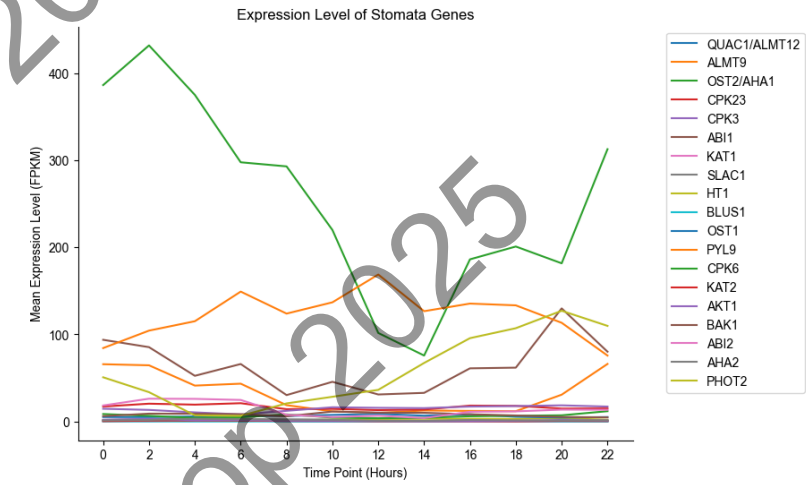


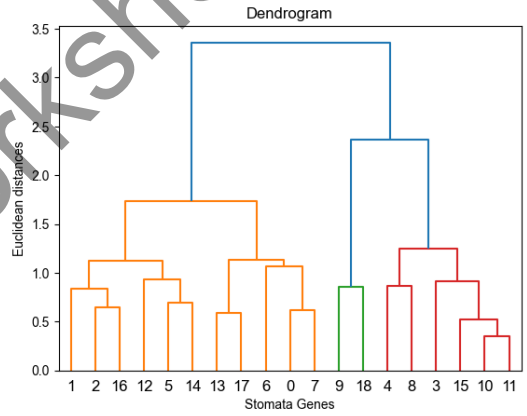
Green line depicts HSP70 diel expression, enriched in dawn

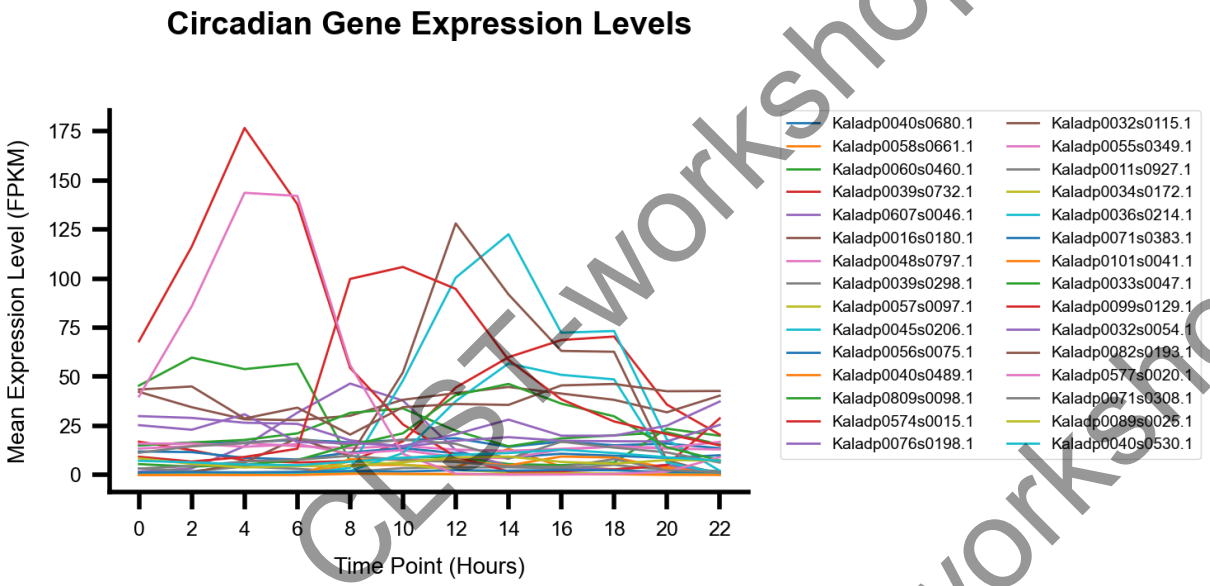


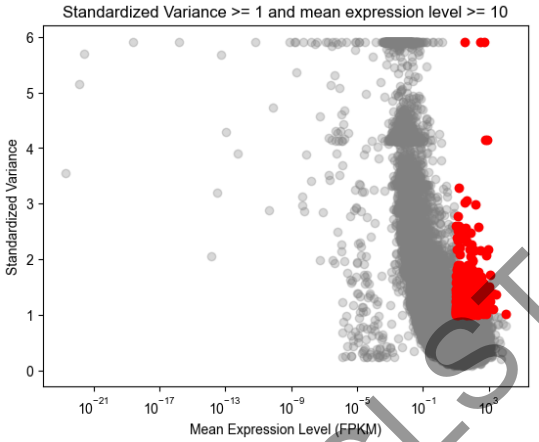
Standardized variance = variance/mean

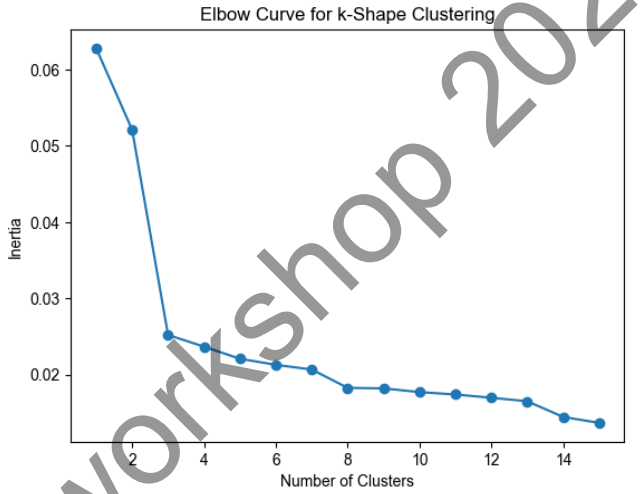


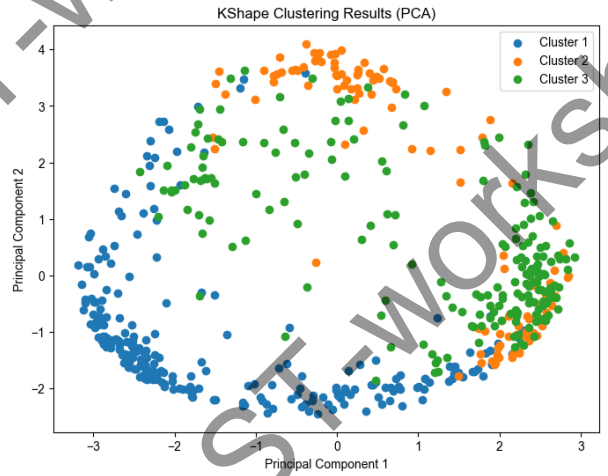


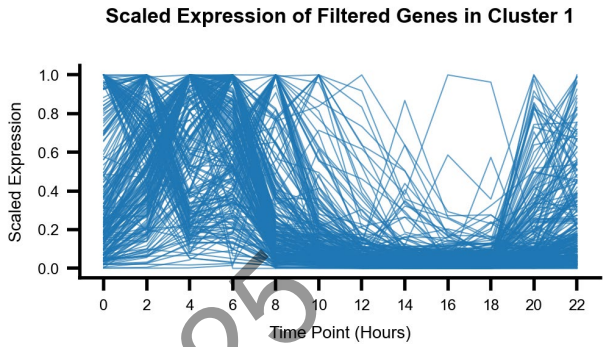


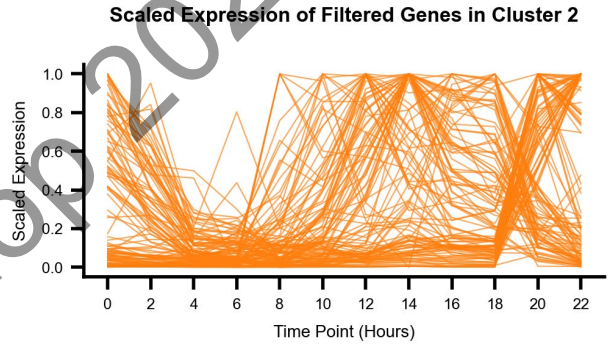


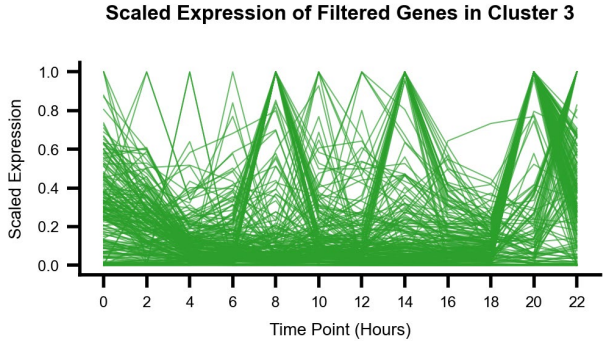


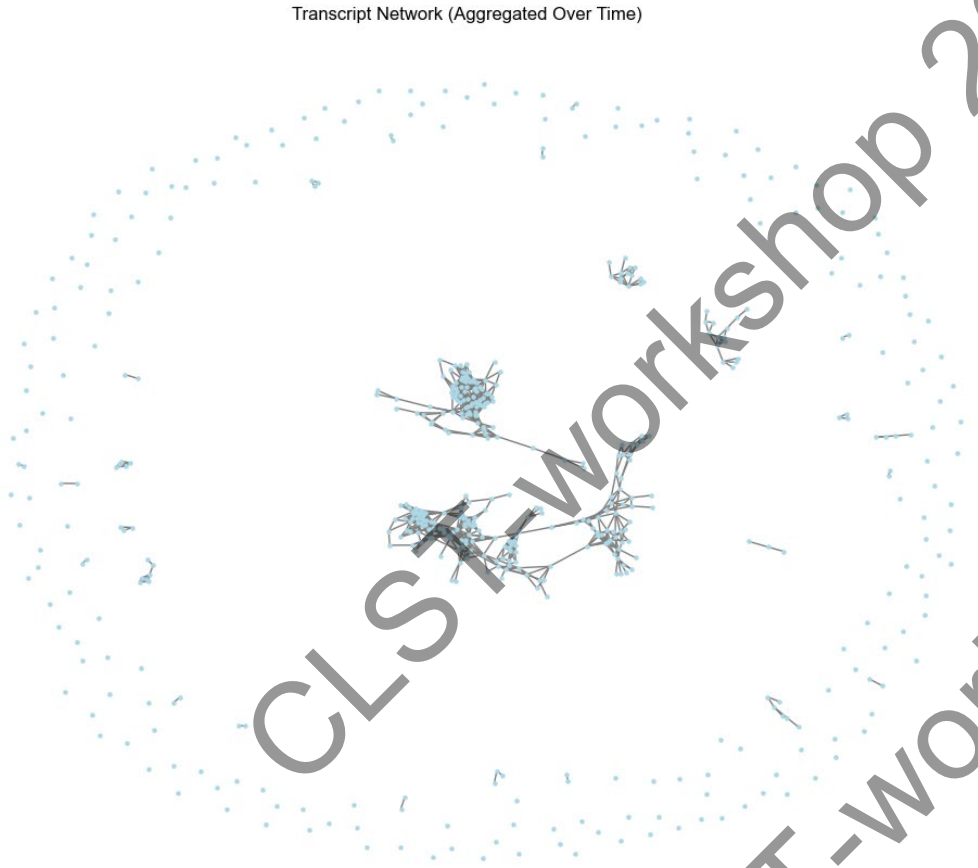


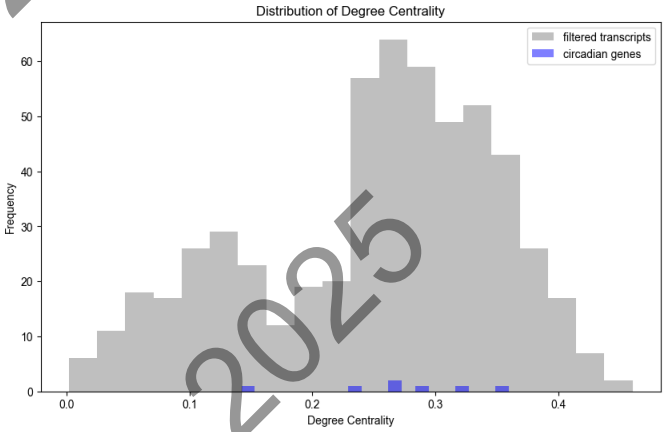


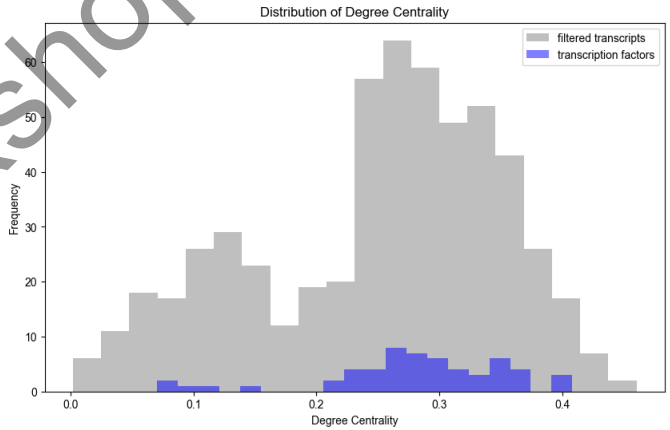


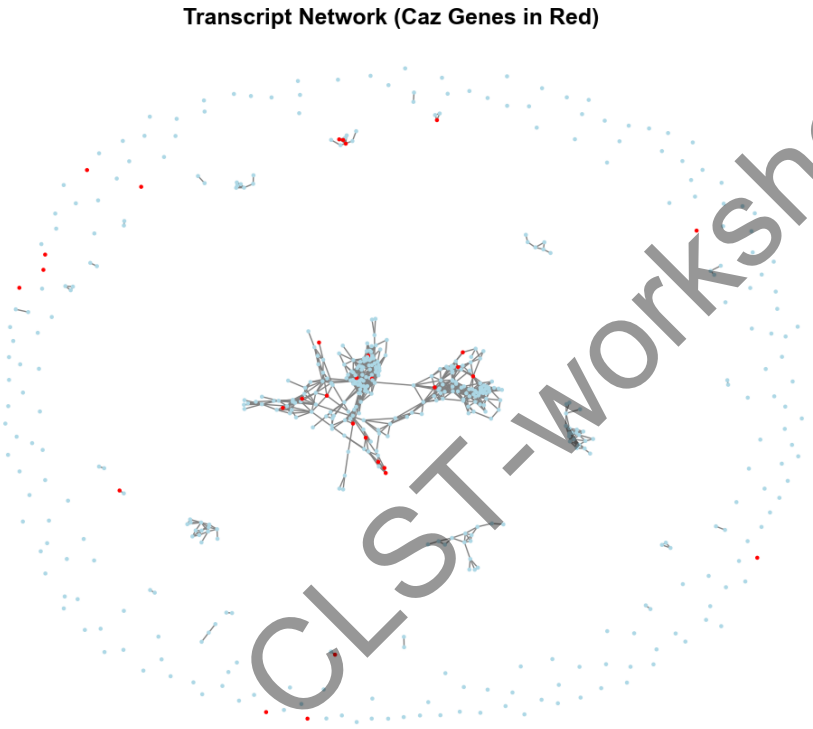


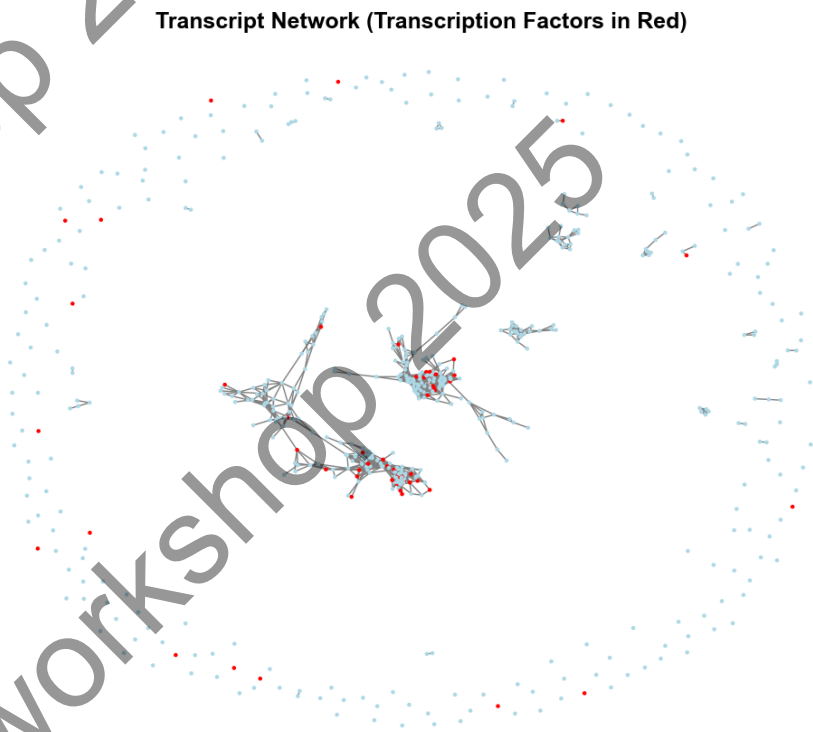


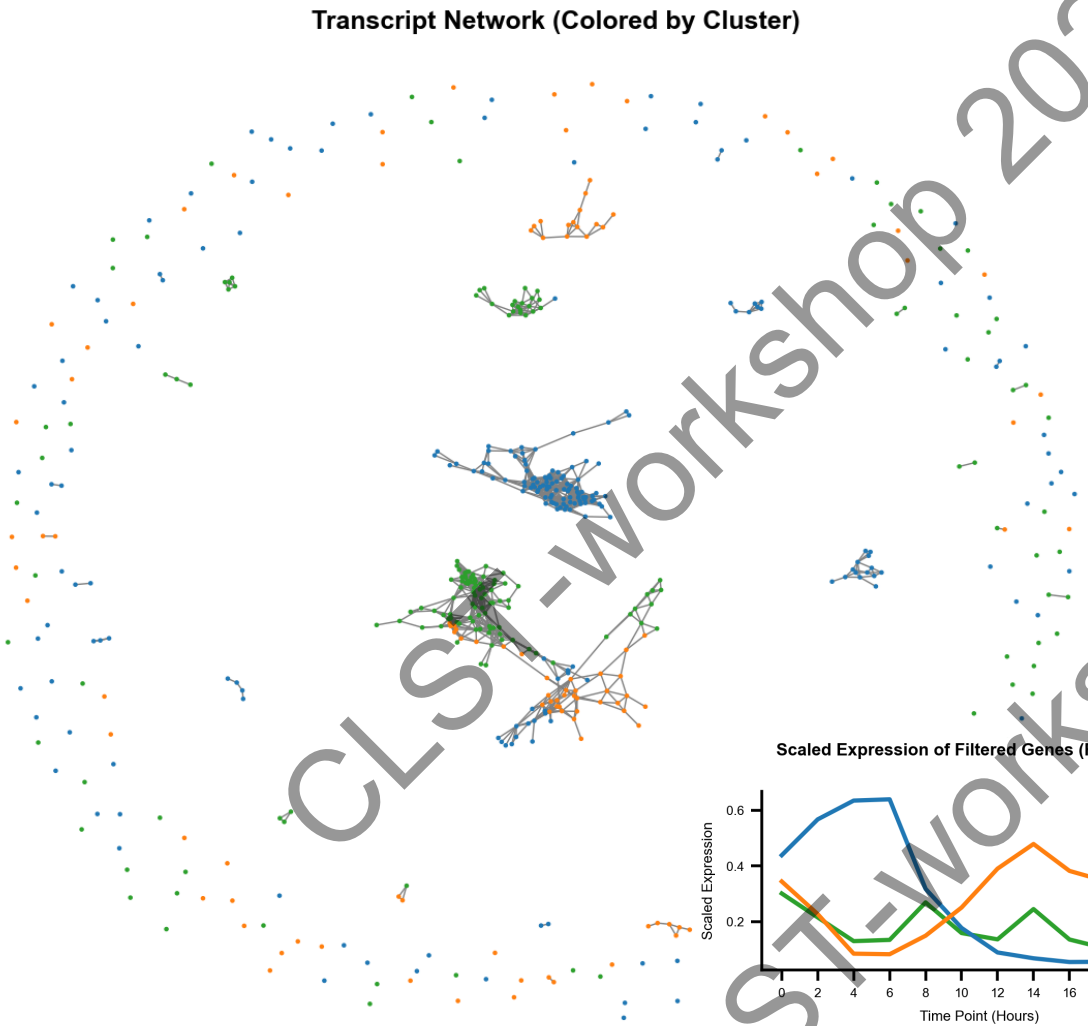


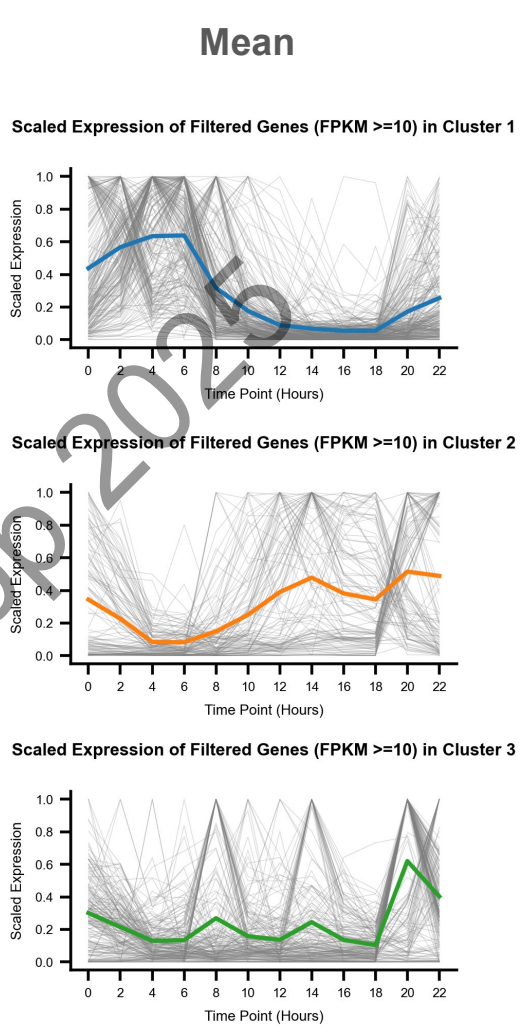


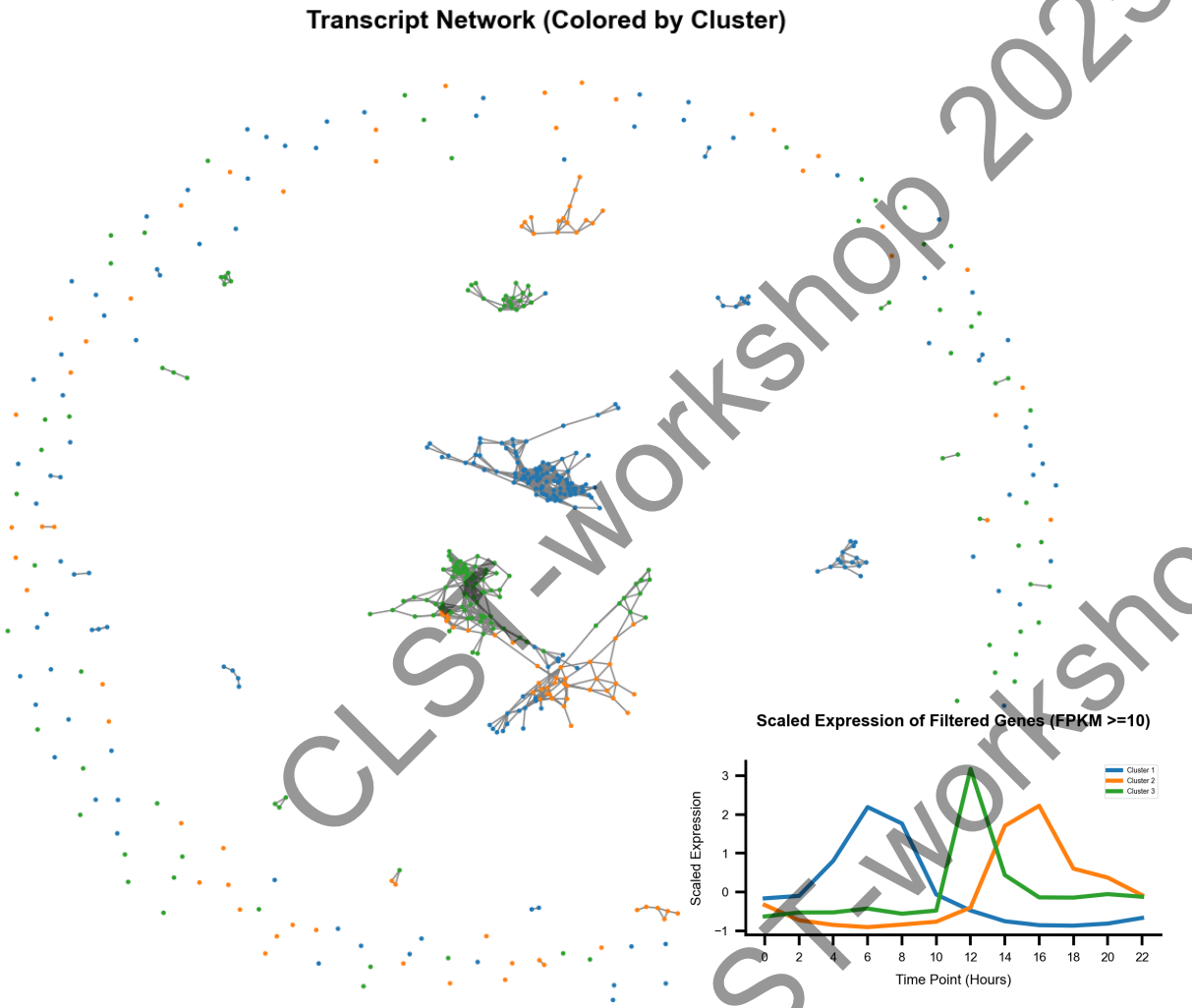




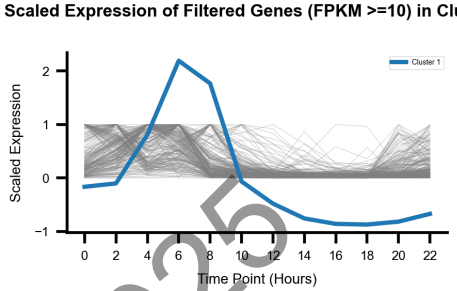


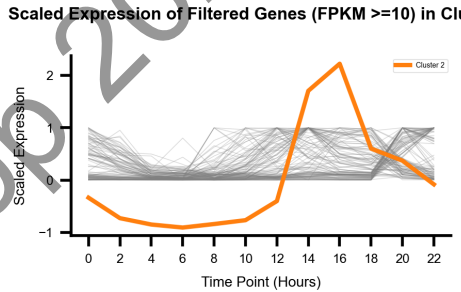


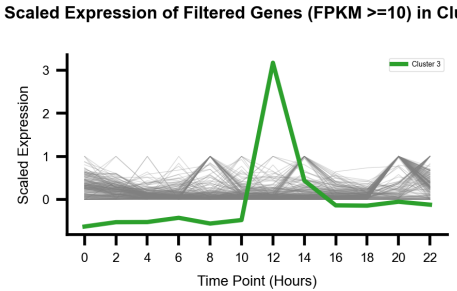




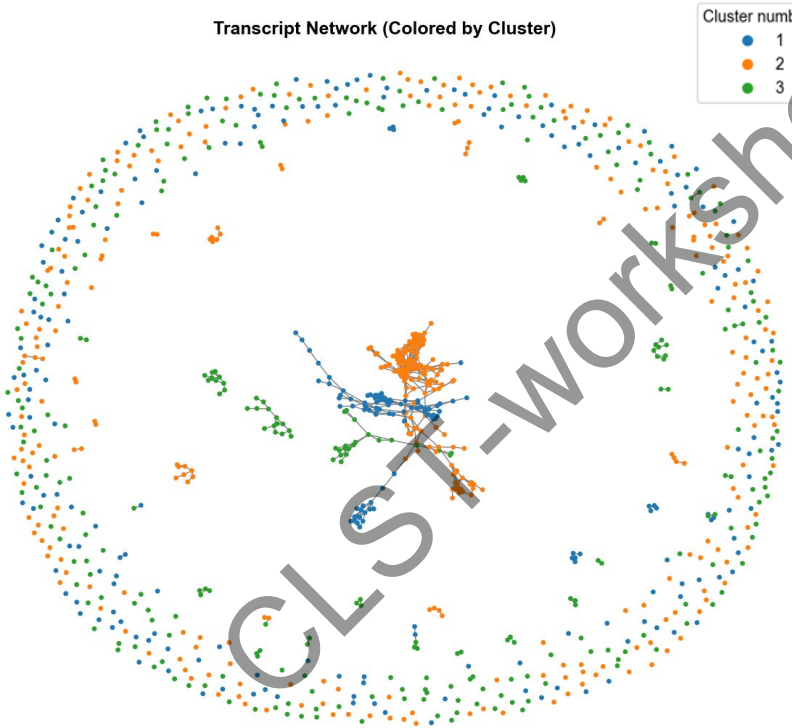
Centroid

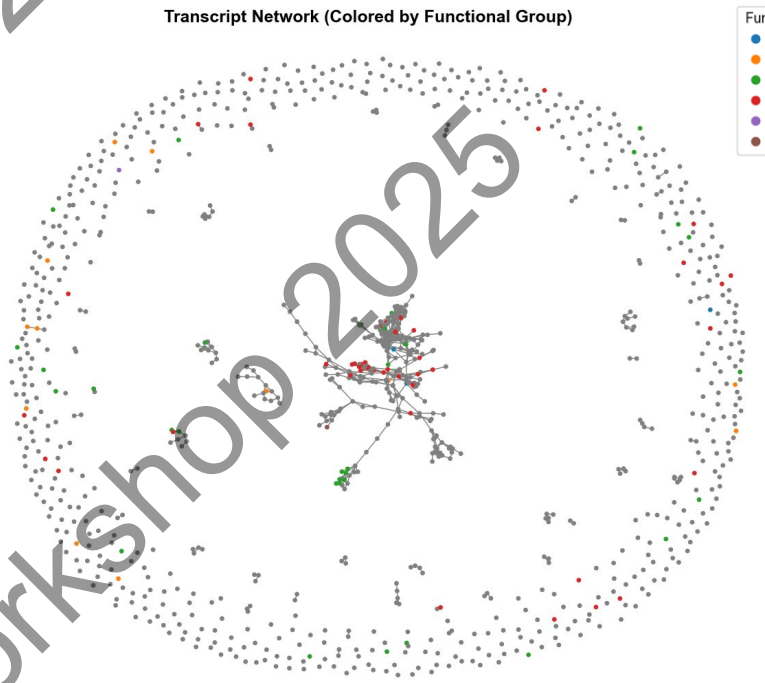


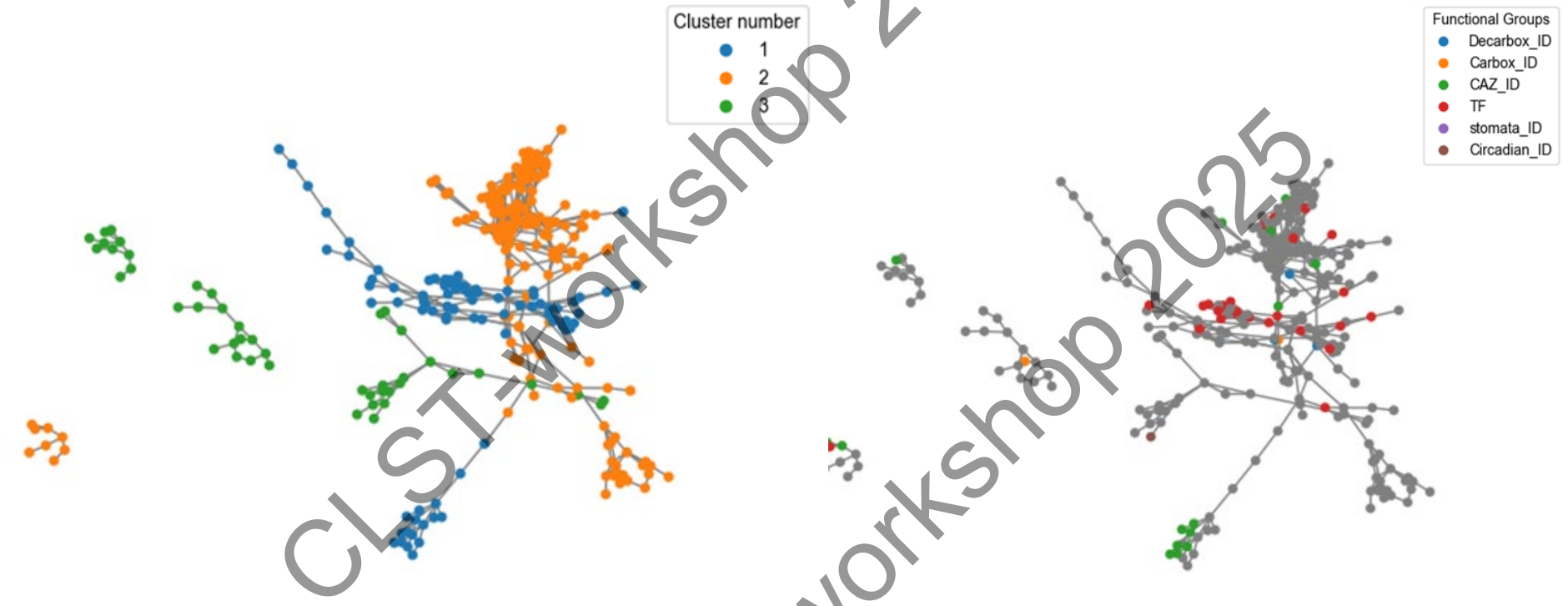




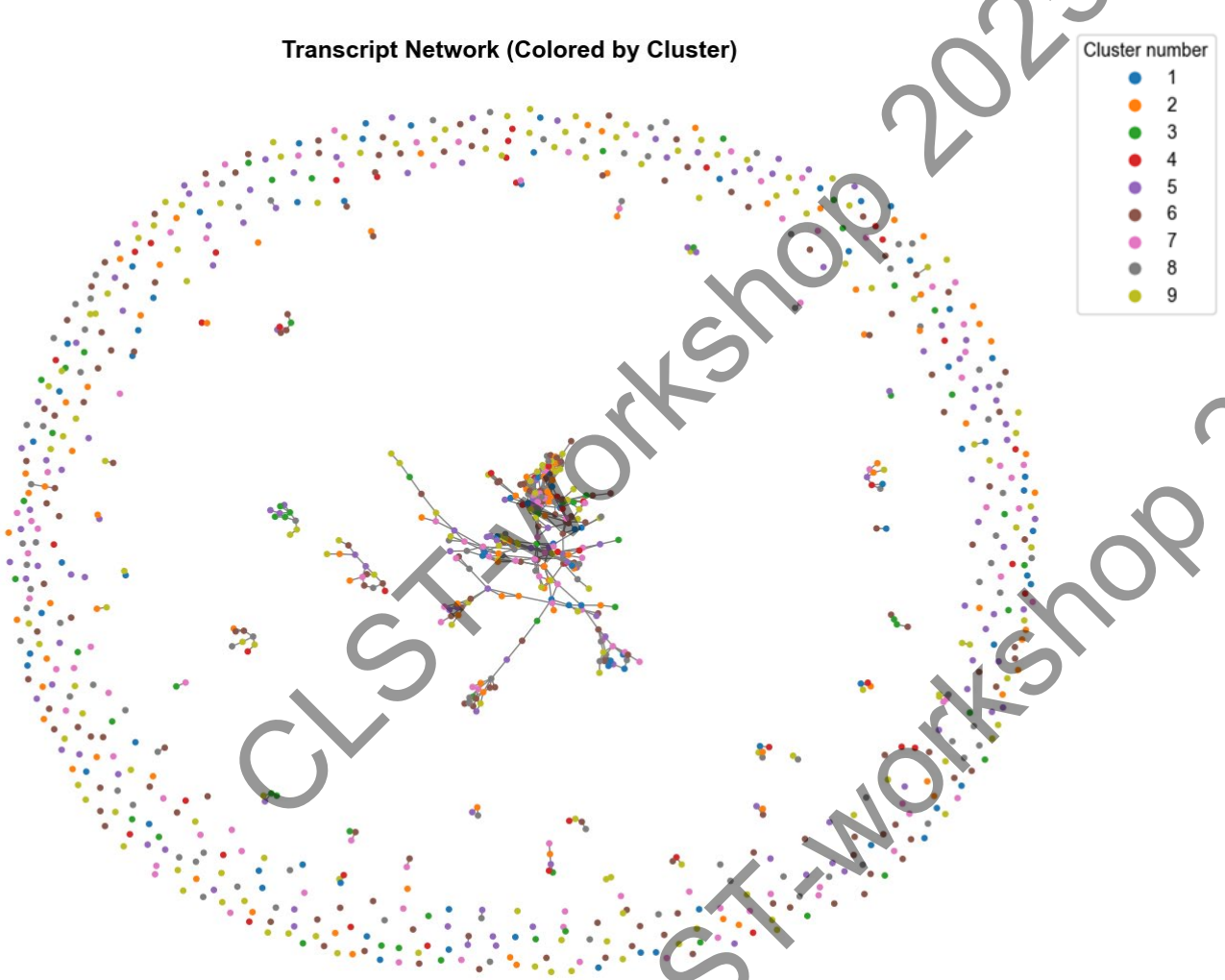
1k 3 cluster

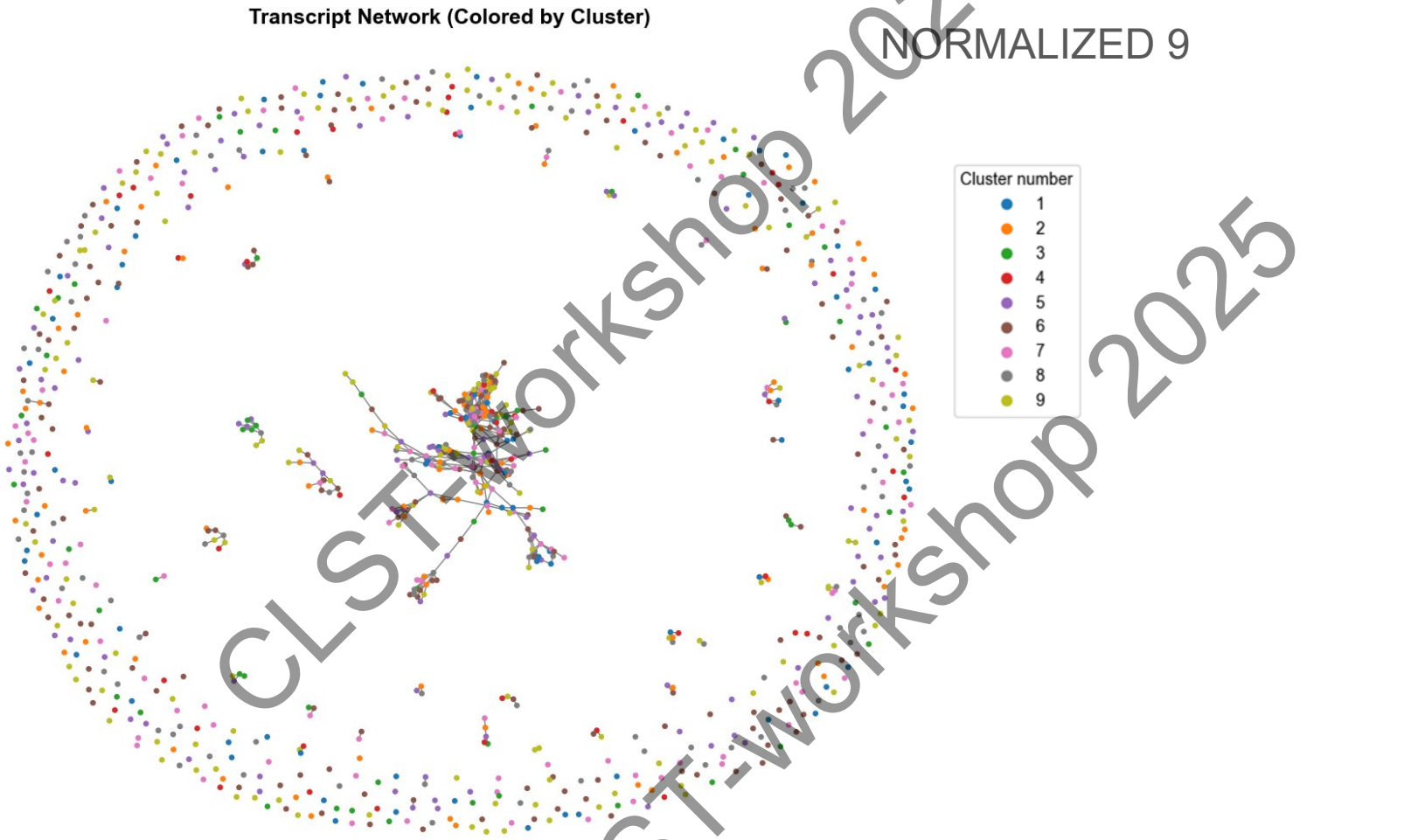






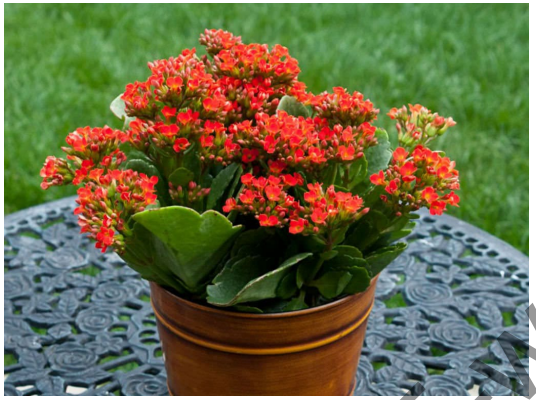
Carboxylation in cluster 2
Decarboxylation in cluster 1
Carbohydrate metabolism in cluster 1



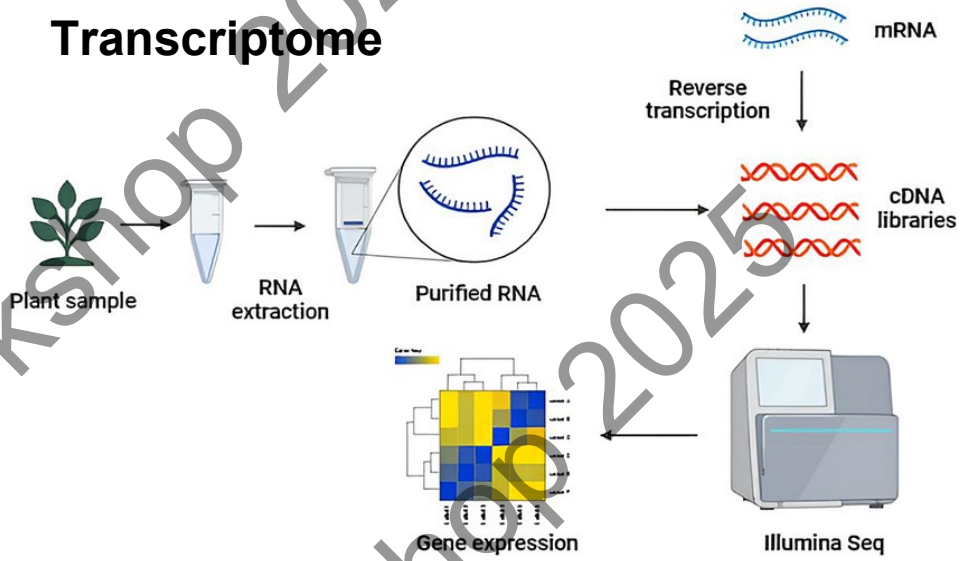


NORMALIZED 9

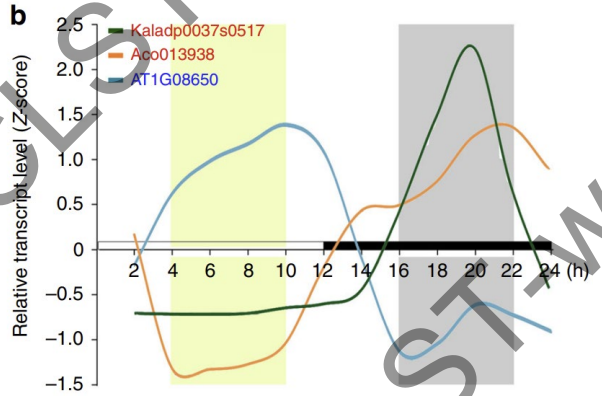
Kalanchoe (ĆŭŢĞńŰŭ)



Transcriptome



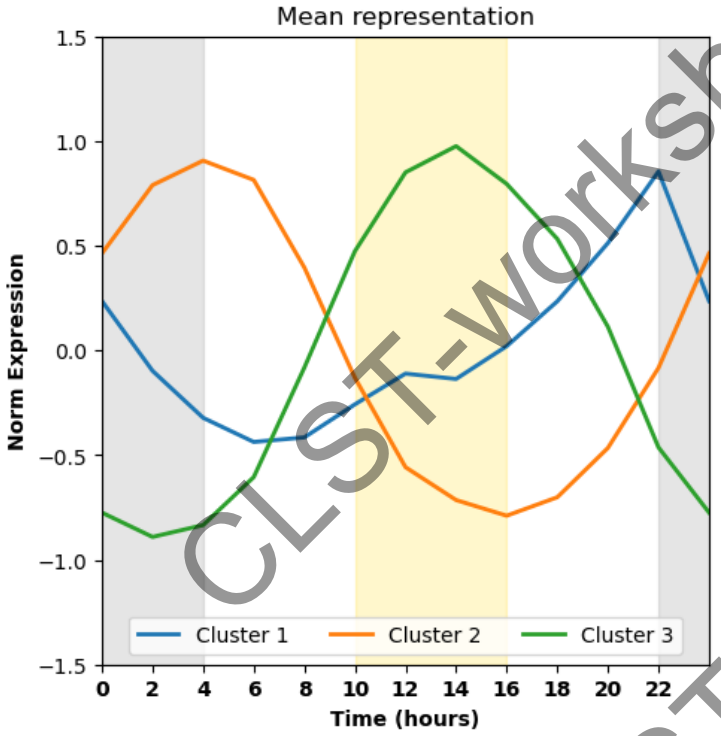
TIme-series

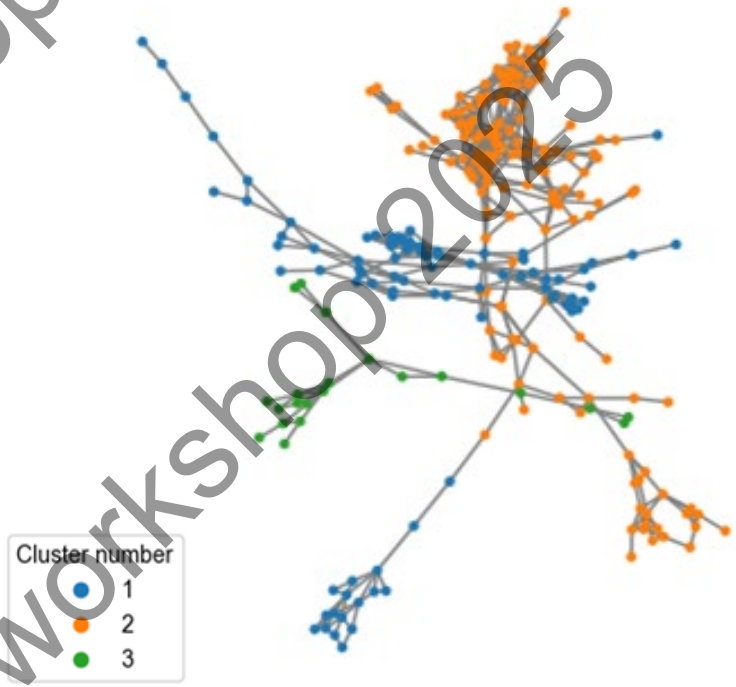


Factors that affect Gene Expression

- CAM Plant
- Circadian Rhythm
- Flowering Process
- Stomatal Movement

Network Construction (from SBD)





Network Construction (from ED)