# ManiNetCluster: a manifold learning approach to reveal the functional links between gene networks

- A biological application of multi-view learning

**Nam D. Nguyen**

Department of Computer Science

Stony Brook University

# Content

## Introduction

- Biological multi-view data & comparative analysis
- Manifold learning

## ManiNetCluster

- Non-linear network embedding & alignment
- Multi-layer network clustering
- Discovering functional links between gene networks

## Results

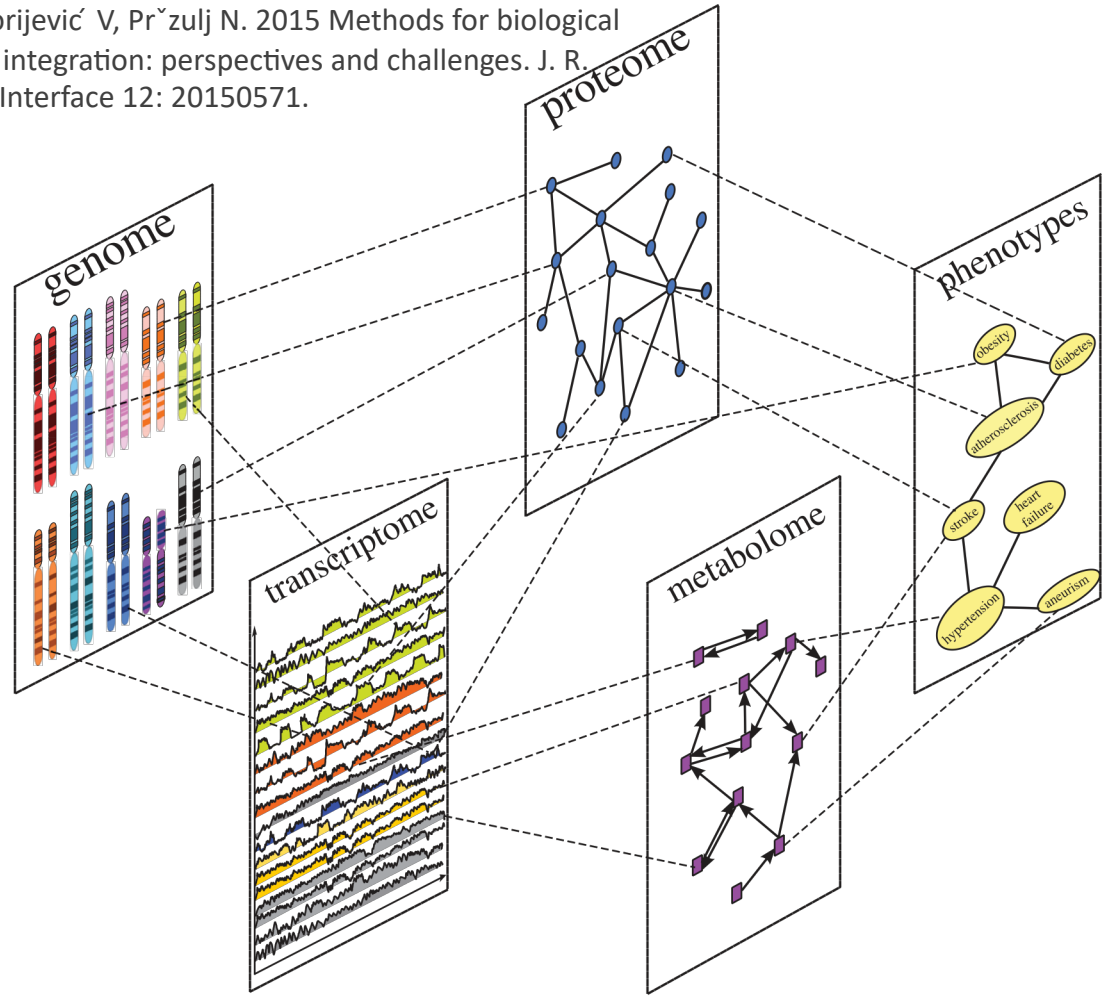
- Aligning cross-species developmental gene networks
- Identifying gene modules, including **function links** between light and dark condition in green algae

## Discussion & future work

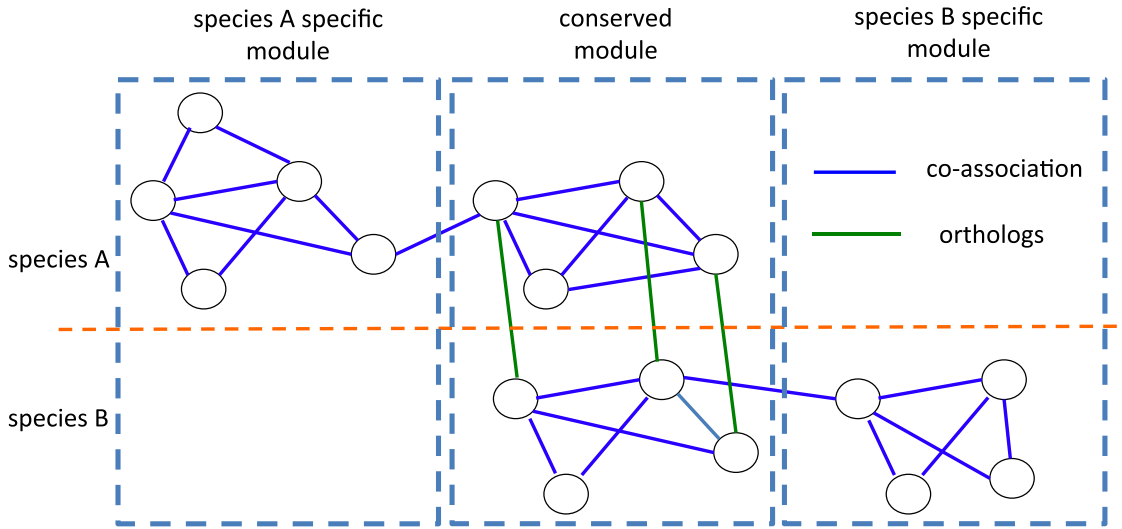
# Content

# Biological **multi-view** data integration

Gligorijević V, Pr˘zulj N. 2015 Methods for biological data integration: perspectives and challenges. J. R. Soc. Interface 12: 20150571.




Multi-omics data




Yan, Koon-Kiu, et al. "OrthoClust: an orthology-based network framework for clustering data across multiple species." *Genome biology* 15.8 (2014): R100.

Comparative genomics

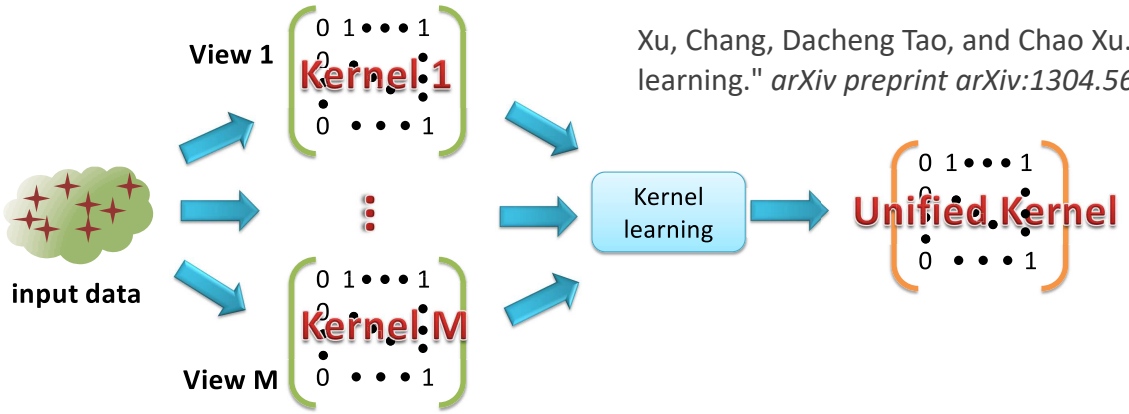
# Multi-view learning

**Multiple kernel learning**

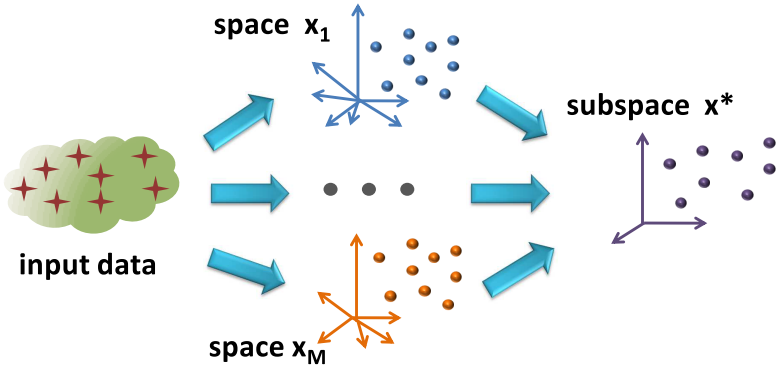
- learn a combination of predefined kernels
- not data dependent
- expensive
  - **non-linear**

**Subspace learning**

- obtains a common latent subspace
- **data dependent** → more general
- **non-expensive**
- CCA (Canonical Correlation Analysis)
  - linear

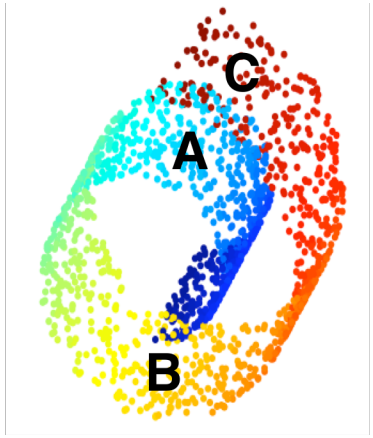



Xu, Chang, Dacheng Tao, and Chao Xu. "A survey on multi-view learning." *arXiv preprint arXiv:1304.5634* (2013)



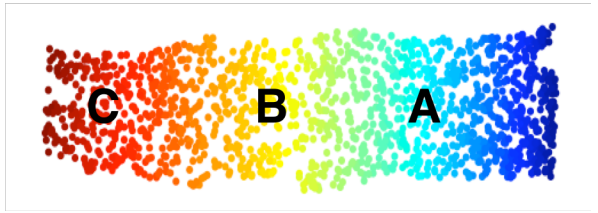

**ManiNetCluster** is a *subspace learning*, **non-linear** and non-expensive
(by employing **manifold learning**) to discover functional links across multiple views

# Manifold learning




High-dimensional data

(Manifold Learning)




Low-dimensional embedding

A $d$ dimensional manifold $M$ is embedded in a $m$ dimensional space, and there is and explicit mapping $f$: $\mathbb{R}^d \rightarrow \mathbb{R}^m$ where $d \leq m$. Given samples $x_i \in \mathbb{R}^m$ with noise

$$x_i = f(\tau_i)$$

→ **find $f(.)$ or $\tau_i$** from given $x_i$

# Content

Introduction

- Biological multi-view data & comparative analysis
- Manifold learning

ManiNetCluster

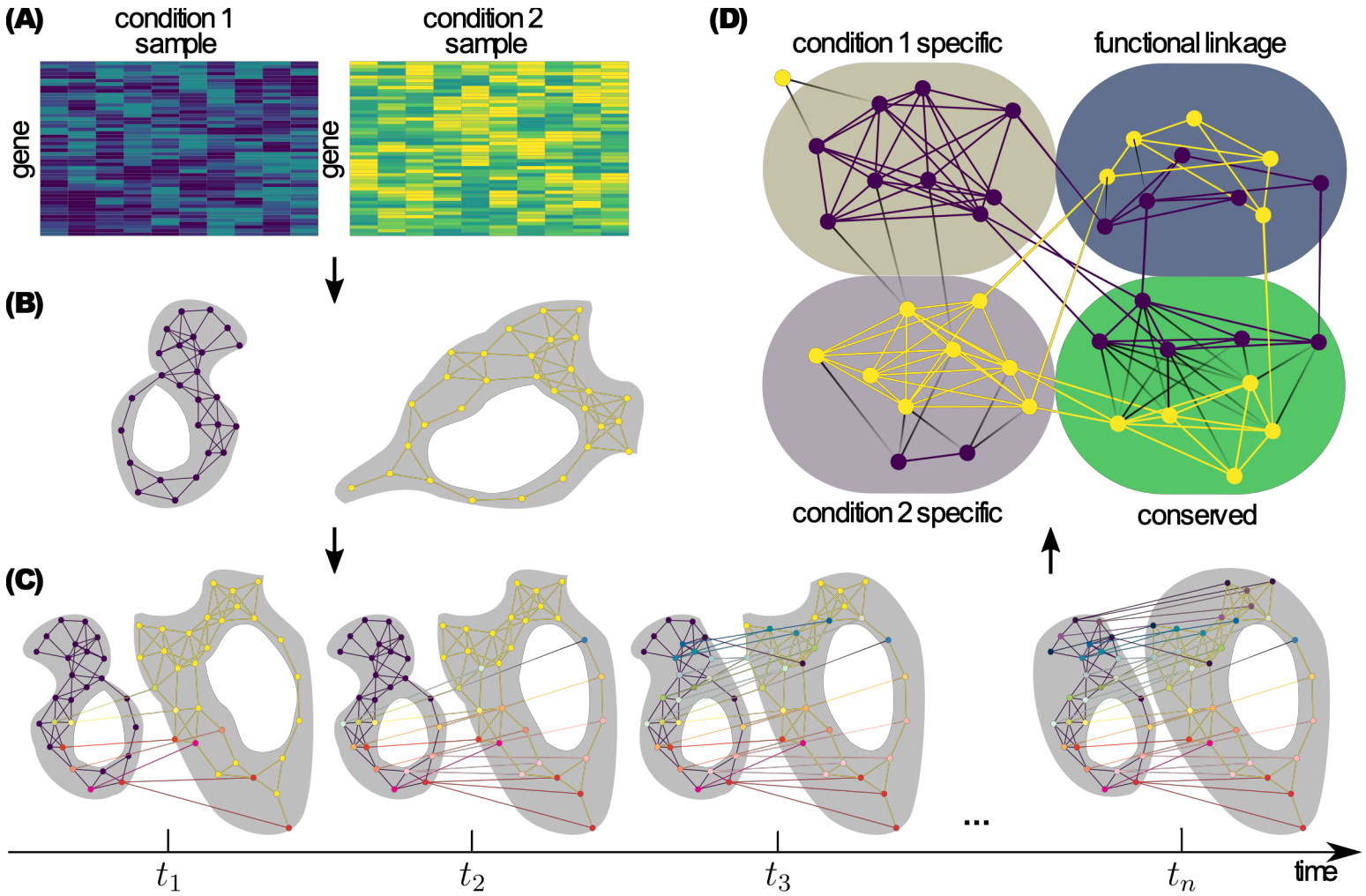
- Non-linear network embedding & alignment
- Multi-layer network clustering
- Discovering functional links between gene networks

Results

- Aligning cross-species developmental gene networks
- Identifying gene modules, including **function links** between light and dark condition in green algae

Discussion & future work

# ManiNetCluster




Tool available at https://github.com/namtk/ManiNetCluster

Nguyen et al, doi: https://doi.org/10.1101/470195

# Manifold alignment for 2 gene networks

2 gene expression profiles $X = [x_1, \dots, x_m], x_i \in \mathbb{R}^p$ $Y = [y_1, \dots, y_n], y_j \in \mathbb{R}^q$ $x_i \leftrightarrow y_i \; for \; i \in [1, l]$

find mapping function $\boldsymbol{f}, \boldsymbol{g}$ to minimize the cost function

$$\sum_{i,j} \|f(x_i) - g(y_j)\|^2 W^{i,j} + \sum_{i,j} \|f(x_i) - f(x_j)\|^2 W_X^{i,j} + \sum_{i,j} \|g(y_i) - g(y_j)\|^2 W_Y^{i,j}$$

global consistency

local smoothness

- extract and optimally aligns local geometry to minimize overall differences
- is a generalization of CCA $\sum_{i,j} \|f(x_i) - g(y_j)\|^2$
- can be interpreted as a manifold regularization

$$\sum_{i,j} \|f(x_i) - g(y_j)\|^2 W^{i,j} + tr(\mathbb{f}^T L_X \mathbb{f}) + tr(\mathbb{g}^T L_Y \mathbb{g})$$

# Aligned network clustering to reveal the functional links

cluster the embedded datasets simultaneously using k-medoids

→ gene modules $C_1, \ldots, C_n$

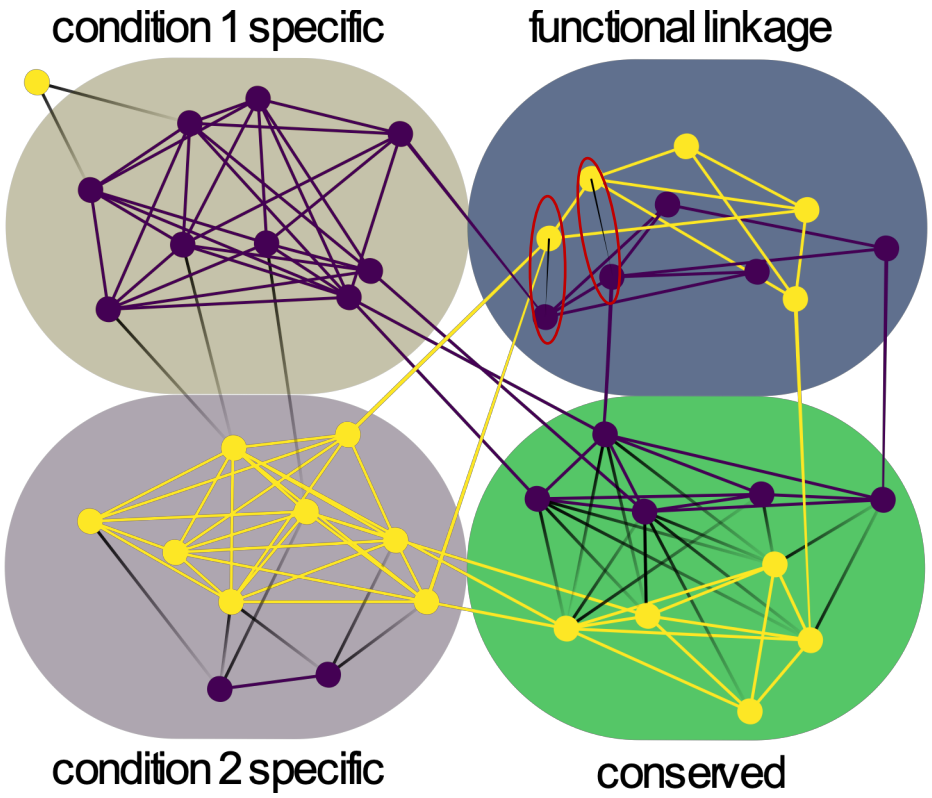
**Jaccard similarity** $J(C_i) = \frac{|X' \cap Y'|}{|X' \cup Y'|}$

**condition number** $\kappa(C_i) = \frac{|X'|}{|Y'|}$

→4 types of gene modules

- Conserved modules ← $J(C_i)$ is high
- View 1 specific modules ← $J(C_i)$ is low, and $\kappa(C_i) \gg 1$
- View 2 specific modules ← $J(C_i)$ is low, and $\kappa(C_i) \ll 1$
- Functional linkage modules ← $J(C_i)$ is low, and $\kappa(C_i) \approx 1$

**Functional linkage score** $S(C_i) = 1 - \frac{\frac{|1-\kappa(C_i)|}{max(\kappa(C_i))} + \frac{J(C_i)}{max(J(C_i))}}{\max\left(\frac{|1-\kappa(C_i)|}{max(\kappa(C_i))} + \frac{J(C_i)}{max(J(C_i))}\right)}$



# Content

Introduction

- Biological multi-view data & comparative analysis
- Manifold learning

ManiNetCluster

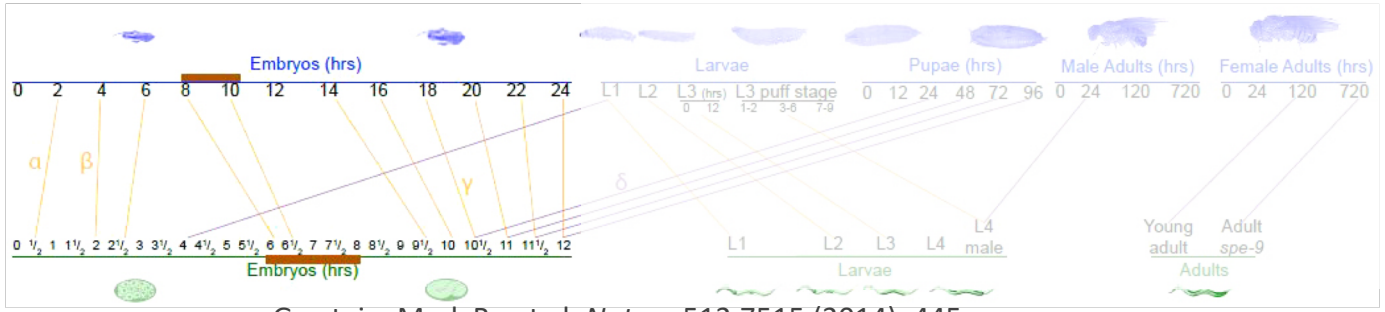
- Non-linear network embedding & alignment
- Multi-layer network clustering
- Discovering functional links between gene networks

Results

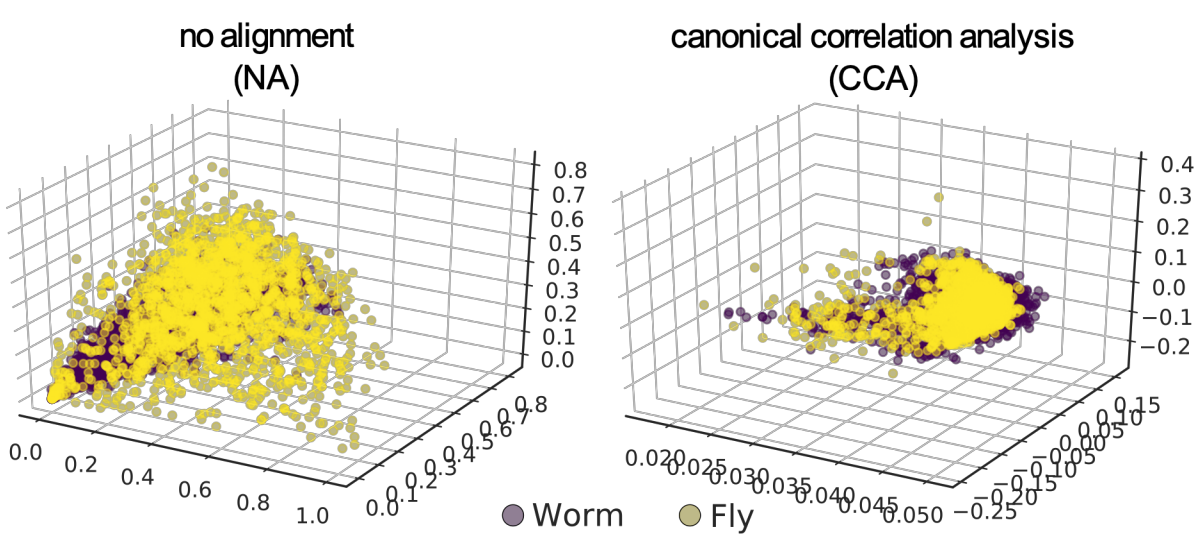
- Aligned cross-species developmental gene networks
- Identified gene modules, including **function links** between light and dark condition in green algae

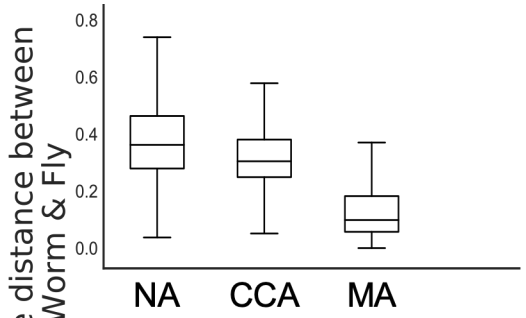
Discussion & future work

# Aligning cross-species gene networks




Gerstein, Mark B., et al. *Nature* 512.7515 (2014): 445.

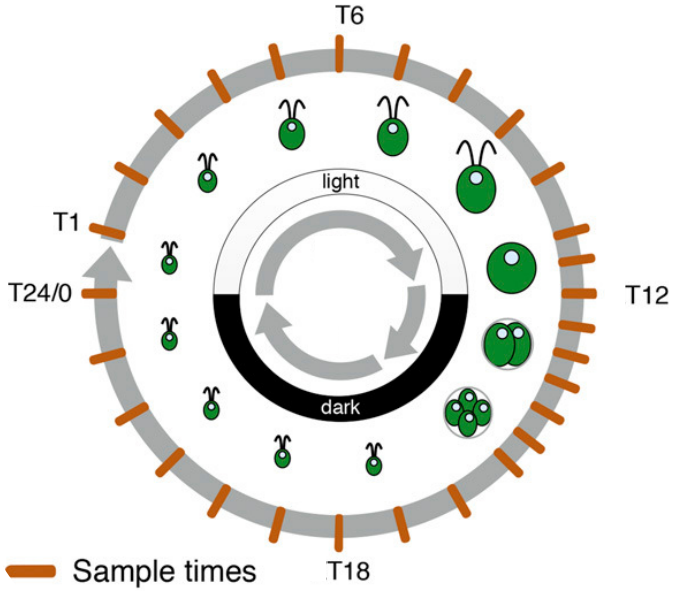


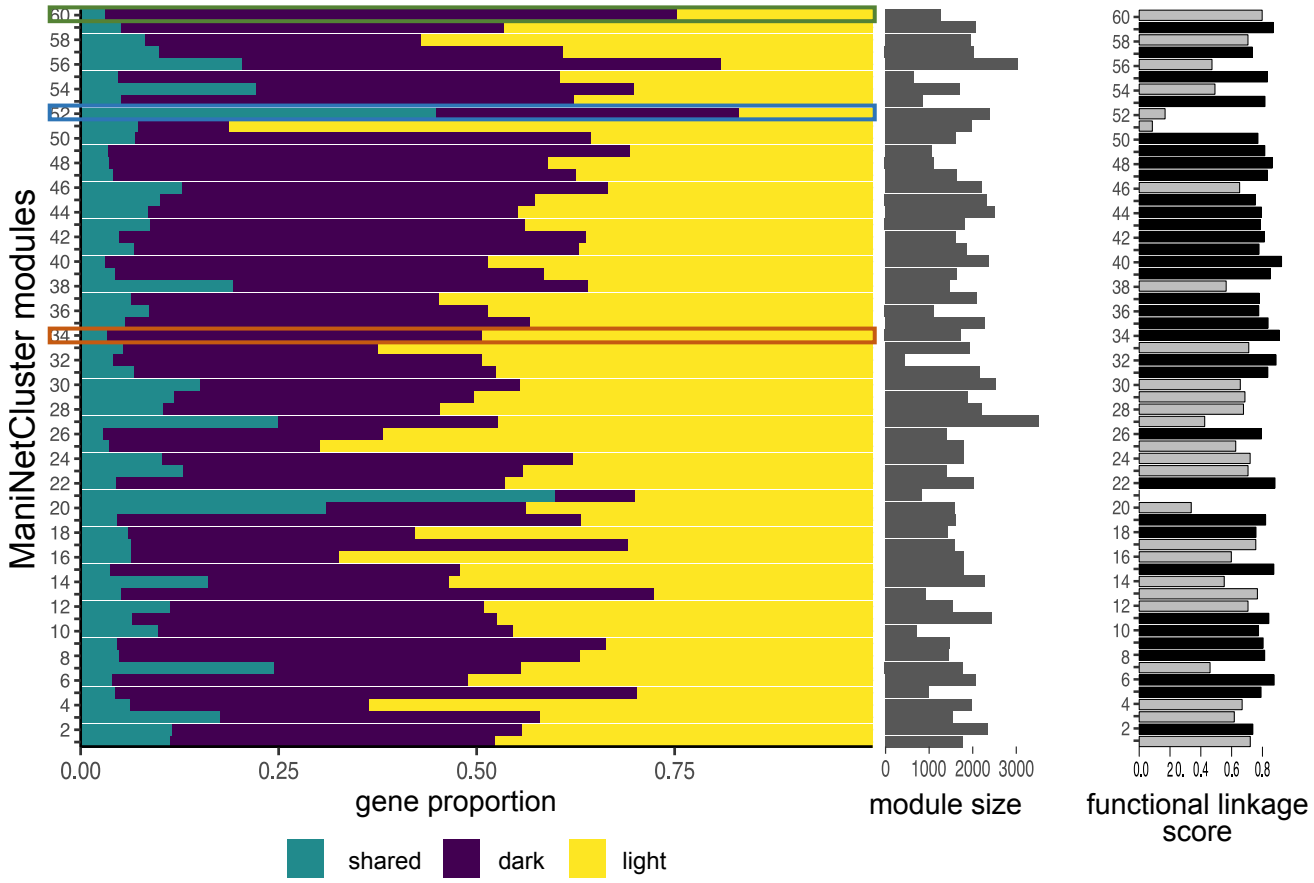

NA = no alignment
CCA = canonical correlation analysis
MA = manifold alignment

# Identifying gene modules between conditions in green algae

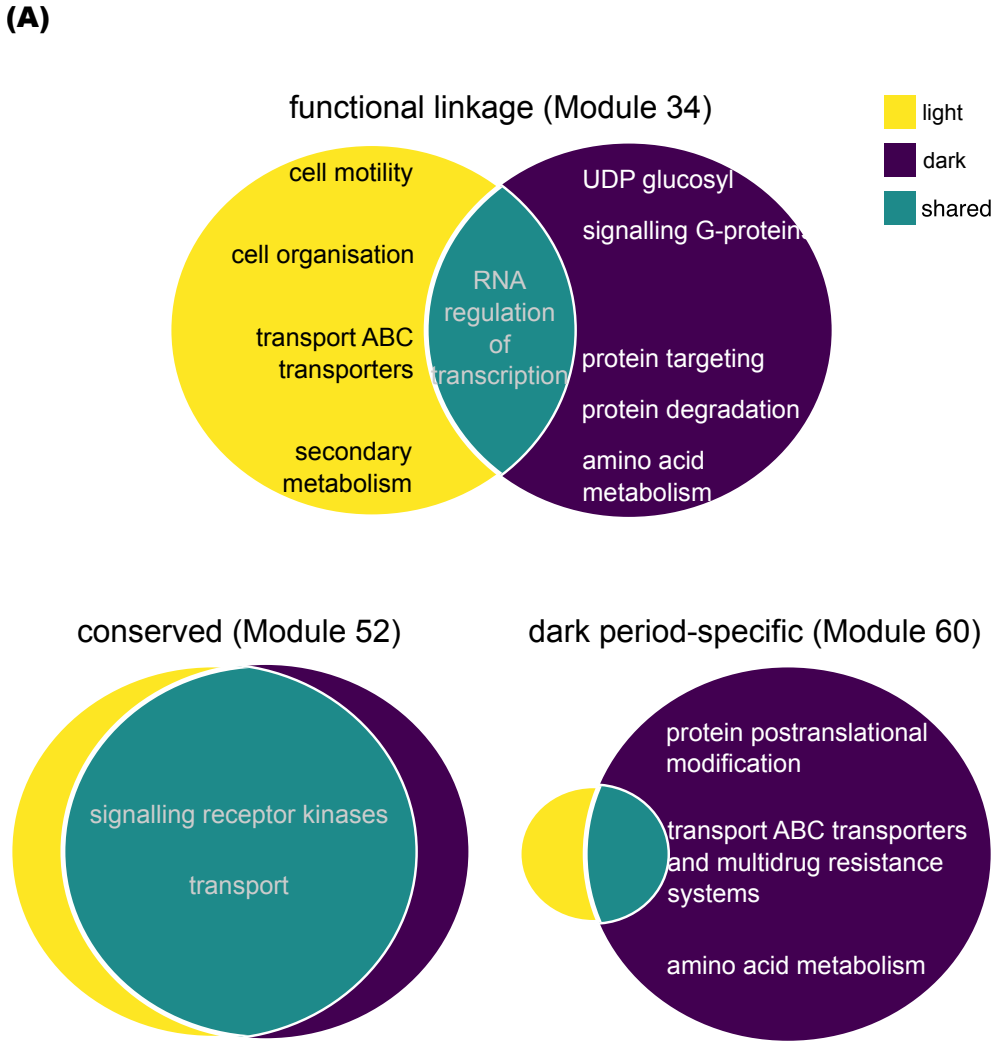
J. M. Zones, I. K. Blaby, S. S. Merchant, J. G. Umen, High-Resolution Profiling of a Synchronized Diurnal Transcriptome from Chlamydomonas reinhardtii Reveals Continuous Cell and Metabolic Differentiation. *Plant Cell* **27**, 2743-2769 (2015)

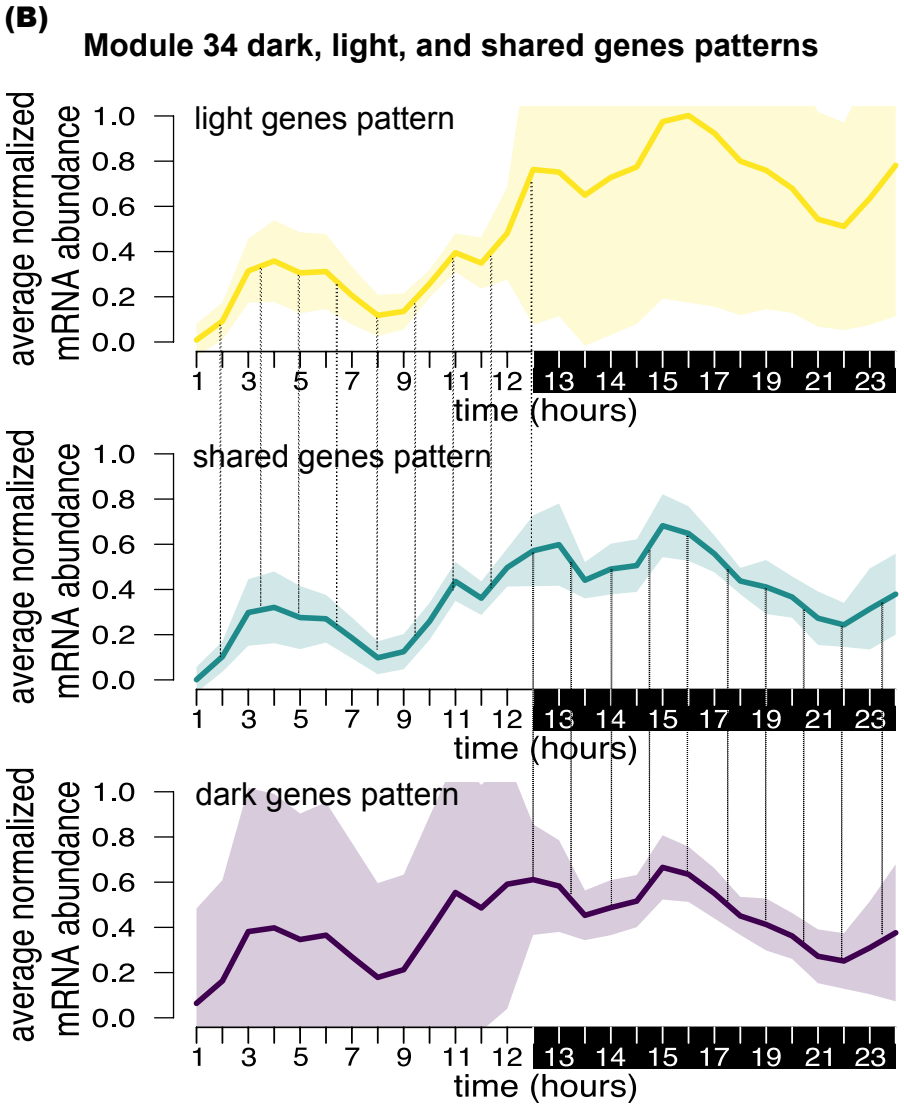



Chlamydomonas reinhardtii (algae)
over a 24hr period
17737 genes → 17695 after data cleaning

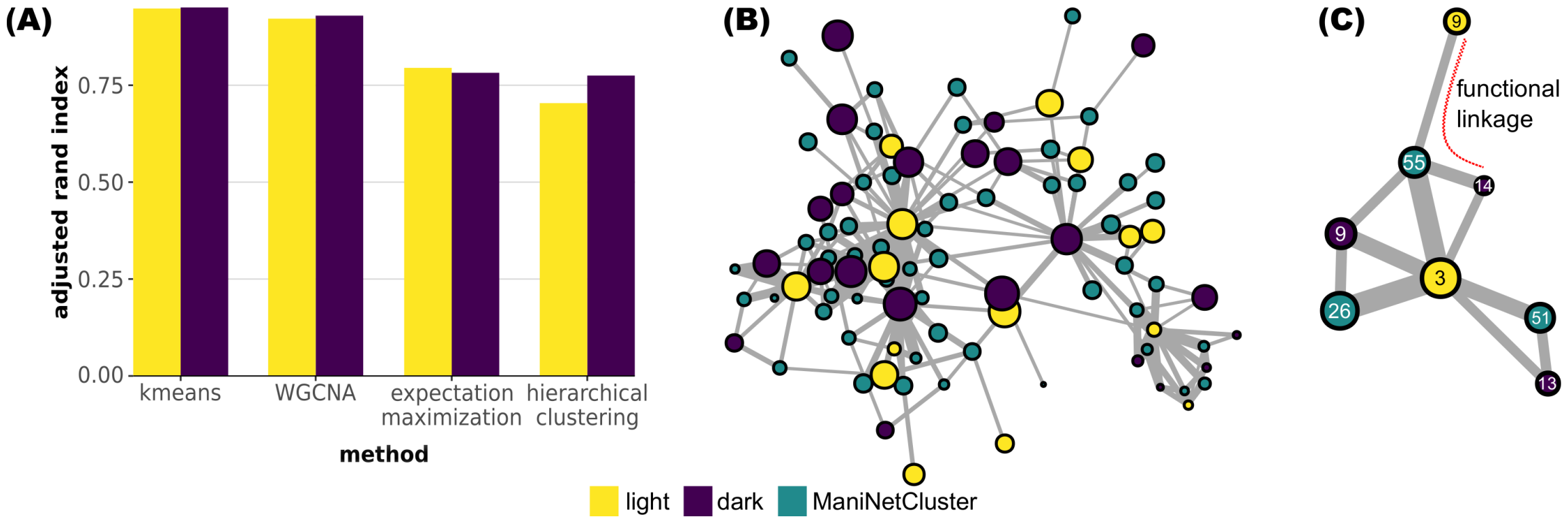


# Functional linkage modules of green algae between conditions



# Comparison of ManiNetCluster with other clustering methods




Single-view clustering methods, e.g. WGCNA, cannot automatically identify the links across network

# Content

# Discussion & future work

**ManiNetCluster:**

- **multi-view learning**
- **simultaneously cluster** across different species/conditions
- discovers of **functional linkage**
- **outperforms** linear methods, e.g., **CCA**
- is **consistent** with single-view clustering methods, but can **infer cross-view information**

**Future work**

- single cell omics
- more NGS data (ChIP-seq, ATAC-seq, etc.)
- diseases (Brain disorders, cancers, etc.)

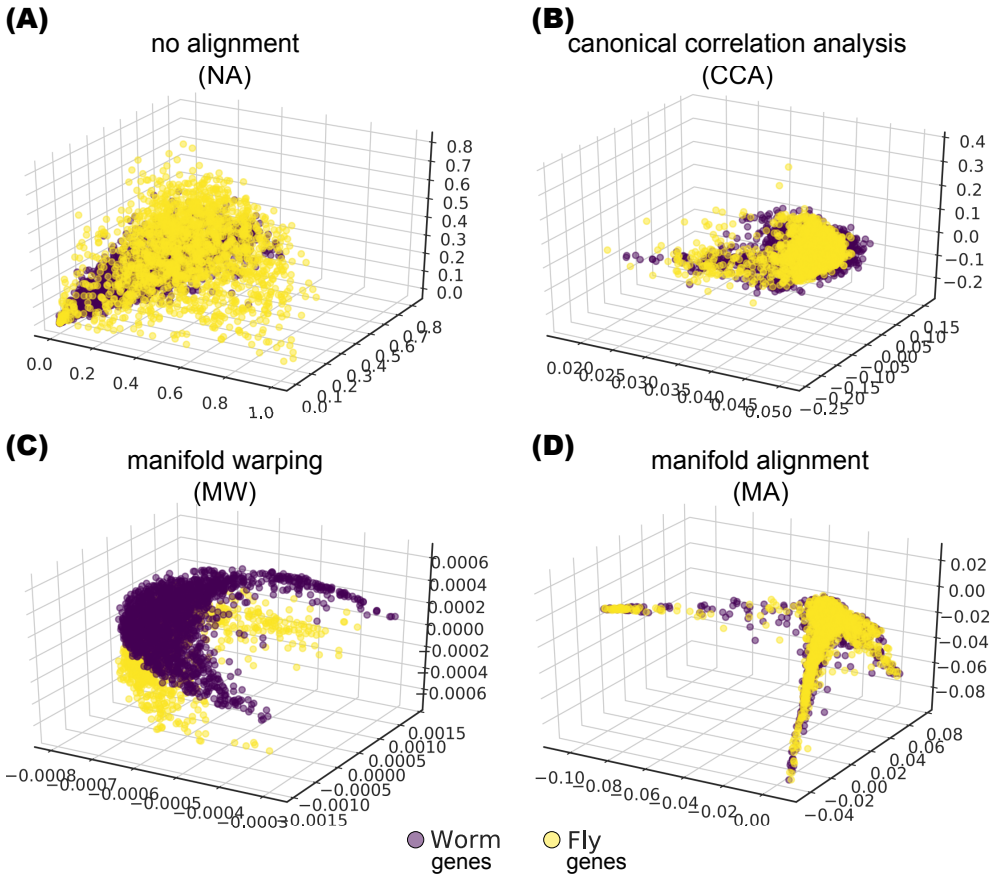
# Acknowledgment

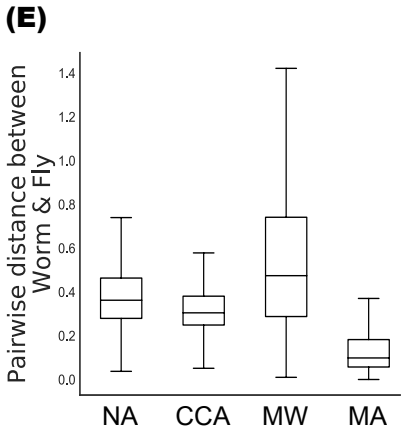

- Dr. Daifeng Wang & Dr. Ian Blaby
- This work was supported by a Stony Brook University/Brookhaven National Laboratory seed grant to Daifeng Wang and Ian K.Blaby.
- NSF travel award ICIBM 2019.


Daifeng Wang Lab (daifengwanglab.org) is moving to Univ. of Wisconsin - Madison.
Postdoc positions available (daifeng.wang@wisc.edu)

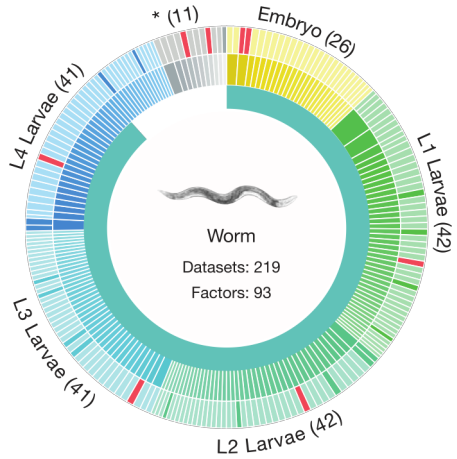

Thank you!

# Aligning cross-species gene networks

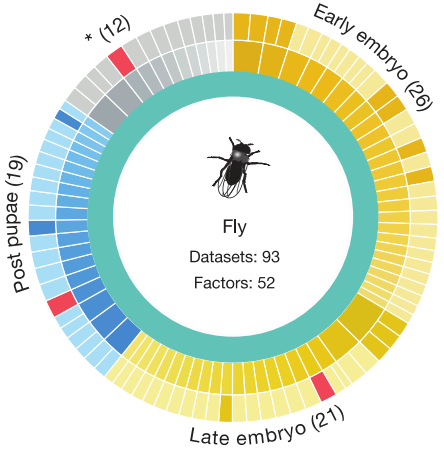



NA = no alignment
CCA = canonical correlation analysis
MW = manifold warping
MA = manifold alignment




- 33 stages: 0, 0.5, 1, …, 12 hours, L1, L2, Young Adults, Adults
- 20377 genes → 18555 genes (removed low expression)
- 1925 ortholog




- 30 stages: 0, 2, 4, …, 22 hours, L1-L4, Pupaes, Adults
- 13623 genes → 11265 genes (removed low expression)
- 1882 ortholog