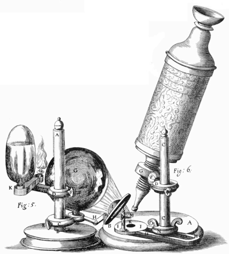
Quantitation in Colocalization Analysis: Beyond “Red+Green=Yellow"

Ming Yuan

Department of Statistics
Columbia University

ming.yuan@columbia.edu
http://www.columbia.edu/~my2550



(Joint work with Shulei Wang)

Seeing Is Believing

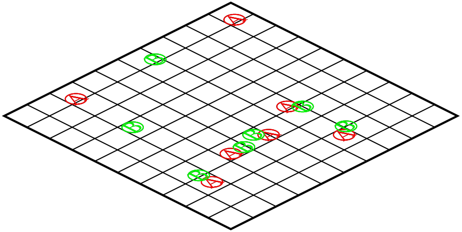
- High contrast
- High specificity (targeted molecules)
- High throughput
- Quantitative (fluorescence intensity, fluorescence lifetime etc.)
- High resolution ($\sim 20nm$)
- Dynamic (monitoring biological events for 24 hour)
- ……

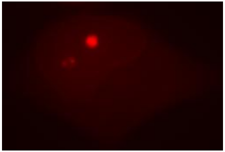
[*Zebrafish*; Cutrale, F. et al., 2017]

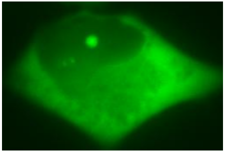
[*Drosophila*; Chhetri, R.K. et al., 2015]

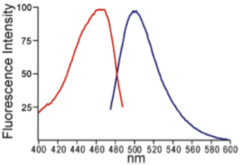
Outline

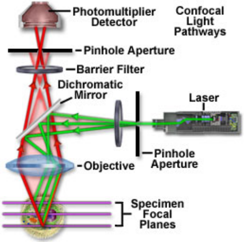
Dual Chanel Fluoresence Imaging





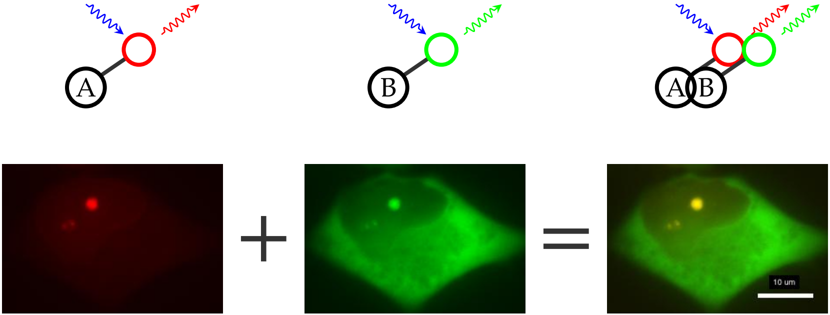






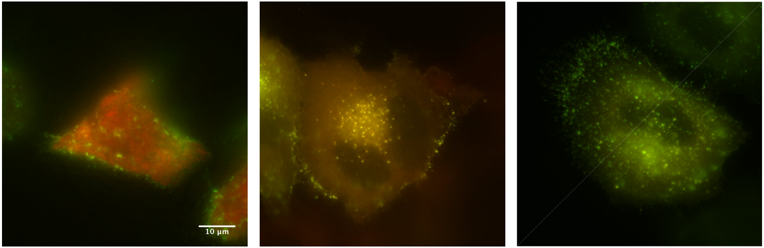
Colocalization

Interaction between bio-molecules → Binding physically → Colocalization between channels



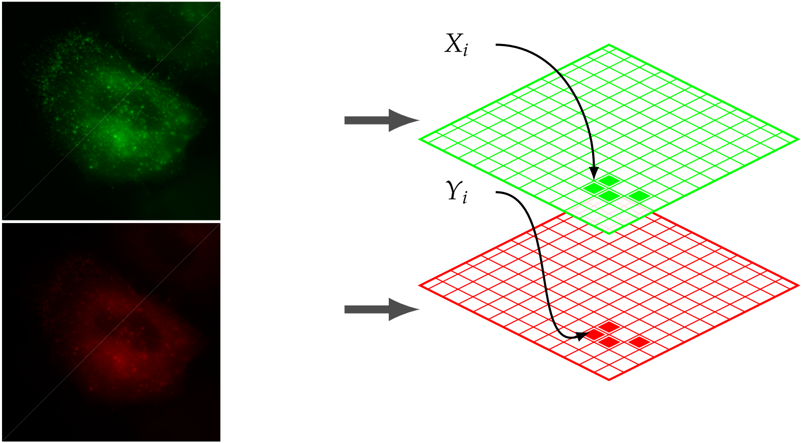
Qualitative v.s. Quantitative

Red + Green = Yellow?



- Subjective, susceptible to cross-talk, and etc.
- Time consuming, labor intensive

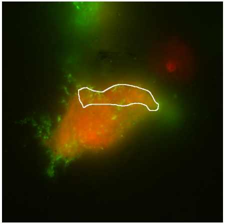
Pixel Based Modeling



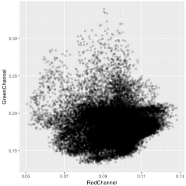
(Pioneered by Manders and Co., 1990s)

Current Pipeline for Colocalization

Select region of interest

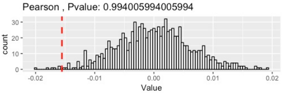


Make scatter plot



Calculate colocalization index: $r = -0.016$

Evaluate statistical significance



(see, e.g., Bolte and Cordeliéres, 2006, Dunn et al., 2011)

But

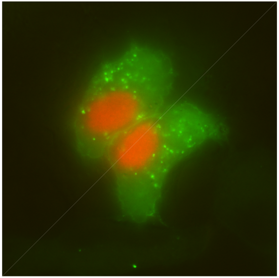
- How to choose region of interest?
- How to choose colocalization coefficient?
- How to evaluate statistical significance?
- How to do so in a computationally efficient way?

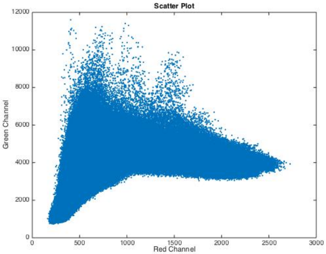
Our goal: a general *statistical/computational* framework for colocalization that is

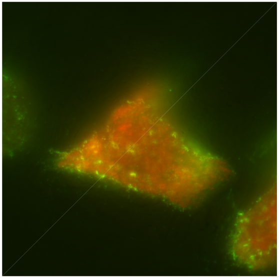
- Automated
- Statistically valid
- Computationally efficient
- Flexible and powerful

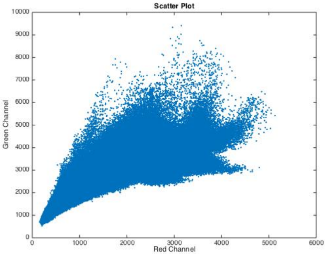
Overview

Background or Signal?

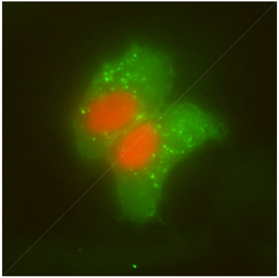


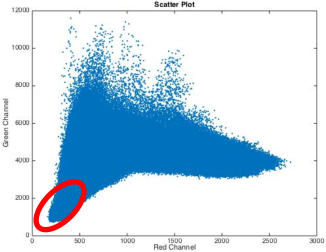


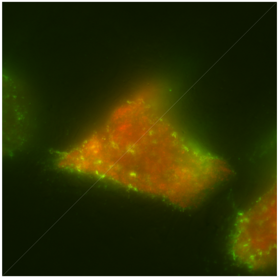


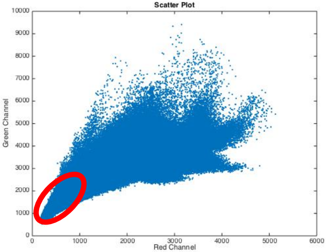


Background or Signal?

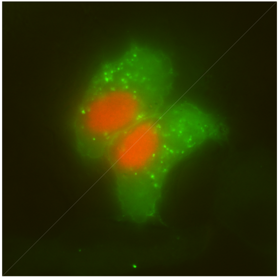


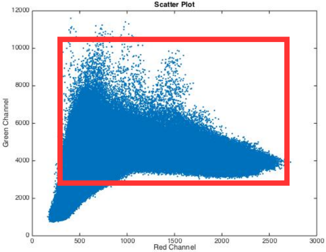


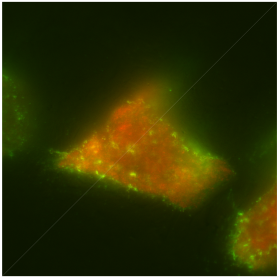


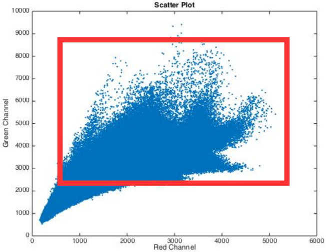


Background or Signal?









What are we measuring?

Pearson's correlation coefficient:

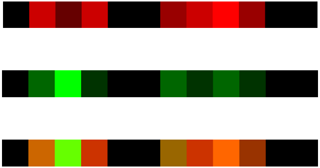
$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}}$$

Manders' colocalization coefficients:

$$M_1 = \frac{\sum_i X_i I_{(Y_i>0)}}{\sum_i X_i}, M_2 = \frac{\sum_i Y_i I_{(X_i>0)}}{\sum_i Y_i}$$

Correlation:

Co-occurrence:



Colocalization as Tail Dependence

- Positively quadrant dependence (PQD, for short) (Lehmann, 1966):

$$\mathbb{P}(X > x, Y > y) \geq \mathbb{P}(X > x)\mathbb{P}(Y > y)$$

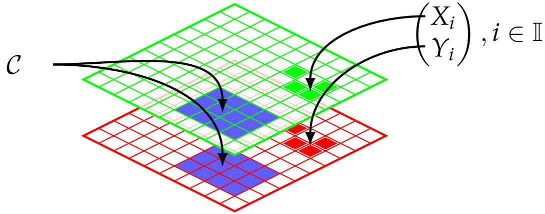
- Colocalization manifested as correlated co-occurrent signals:

$$\begin{aligned}\text{co} - \text{occurence } ({\color{red}V}): & \quad S(\eta_x, \eta_y) - S(\eta_x, -\infty)S(-\infty, \tau_y) \\ \text{correlation } ({\color{red}T}): & \quad \mathbb{P}\left\{(X - \tilde{X})(Y - \tilde{Y}) > 0 | X, \tilde{X} > \eta_X; Y, \tilde{Y} > \eta_Y\right\} - \\ & \quad \mathbb{P}\left\{(X - \tilde{X})(Y - \tilde{Y}) < 0 | X, \tilde{X} > \eta_X; Y, \tilde{Y} > \eta_Y\right\}.\end{aligned}$$

- Background vs signal:

$$F(x, y | x > \eta_x, y > \eta_y) = F_{\eta_x, \eta_y}(x, y) \qquad \leftarrow \ {\color{red}PQD}$$

Colocalization via Statistical Lens



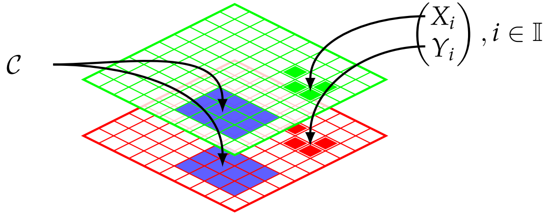
- Assume each $(X_i, Y_i)^\top$ is drawn from a bivariate distribution.
- Without colocalization

$$(X_i, Y_i) \sim \underbrace{F_0(x, y)}_{\text{no } \textit{PQD}}$$

- With colocalization

$$(X_i, Y_i) \sim \underbrace{F_1(x, y)}_{\text{exhibit } \textit{PQD}}$$

A Hypothesis Testing Approach to Colocalization



- Assume each $(X_i, Y_i)^\top$ is drawn from a bivariate distribution.
- Without colocalization

$$(X_i, Y_i) \sim \underbrace{F_0(x, y)}_{\text{no } PQD}, \qquad \forall i$$

- With colocalization located at an *unknown* set $\mathcal{C}$ of pixels

$$(X_i, Y_i) \sim \underbrace{F_1(x, y)}_{\text{exhibit } PQD}, \qquad \forall i \in \mathcal{C}$$

IF WE KNOW $\mathcal{C}$, ...

$$H_0 : (X_i, Y_i) \sim F_0 \quad \forall i \in \mathcal{C} \qquad \text{vs} \qquad H_1 : (X_i, Y_i) \sim F_1 \quad \forall i \in \mathcal{C}$$

- Positively quadrant dependent property implies

$$\tau_H := \mathbb{E}(\text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j)) > 0.$$

  Here τ_H is called *Kendall tau correlation*.
- Empirical version Kendall tau correlation is a good indicator of correlation of $H(x, y)$

$$\widehat{\tau}_H := \frac{1}{n_{\mathcal{C}}(n_{\mathcal{C}} - 1)} \sum_{i \neq j \in \mathcal{C}} \text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j)$$

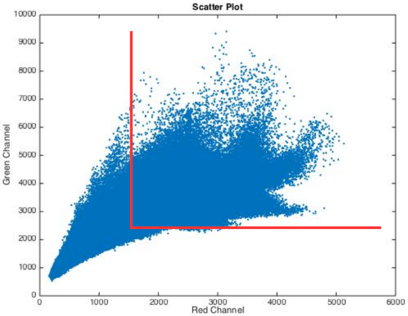
But we do *not* know $\mathcal{C}$ (or equivalently η_x and η_y)!

Test for Conditional PQD

- Known η_x and η_y – *conditioned and normalized* Kendall's tau

$$\widehat{\tau}(\eta) = \begin{cases} \sqrt{\frac{18}{n_\eta(n_\eta-1)(2n_\eta+5)}} \sum_{i,j\in\mathcal{K}(\eta):i<j} \operatorname{sign}(X_i - X_j)(Y_i - Y_j) & n_\eta > 1 \\ -\infty & n_\eta \leq 1 \end{cases}$$

where $\mathcal{K}(\eta) = \{i : X_i \geq \eta_x, Y_i \geq \eta_y\}$ and $n_\eta = |\mathcal{K}(\eta)|$.



- Unknown η_x and η_y,

$$\tau^* := \max_{T_x \geq X_{(i)}, T_y \geq Y_{(i)} : i,j \geq \lfloor n/2 \rfloor} \widehat{\tau}(T)$$

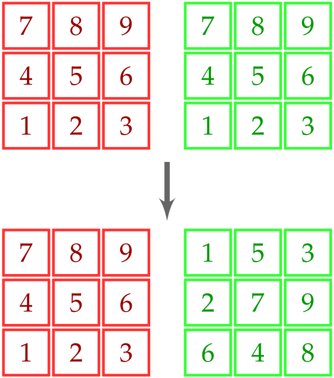
- Test

$$\psi_T = \begin{cases} \text{reject} \quad H_0 & \text{if } \tau^* > q_\alpha \\ \text{accept} \quad H_0 & \text{otherwise} \end{cases}$$

Sampling Distribution Estimation

Statistical significance by permutation test:

- Calculate τ^*_{app} and record it as E_0.
- For $j = 1 : B$, block-wise randomly shuffle $\{X_i\}_{i \in \mathbb{I}}$ with block size D. Calculate τ^*_{app} on shuffled data and recorded it as E_j.
- P-value: $\#\{E_j > E_0\}/B$



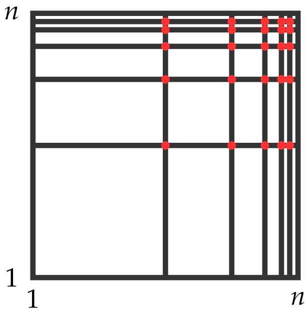
Computational Consideration

Fast computation:

$$\tau_f^* := \max_{T_x = X_{(j)}, T_y = Y_{(k)} : j,k \in \mathcal{R}_n} \widehat{\tau}(T)$$

where

$$\mathcal{R}_n := \left\{ s : s = \left\lfloor n - \left(1 + \frac{1}{\log\log n}\right)^j \right\rfloor j = 1, 2, \ldots \quad \text{and} \quad s \geq \lfloor n/2 \rfloor \right\}.$$



- τ_f^* is faster to compute

	τ^*	τ_f^*
$\# \, \widehat{\tau}(T)$	$O(n^2)$	$O(\log^3 n)$

- And just as powerful

Does it work?

- Assume that $\{(X_i, Y_i) : i \in \mathbb{I}\}$ ($n := |\mathbb{I}|$) are *independently* sampled from F obeying

$$\sup_{\eta_X, \eta_Y} V(\eta_X, \eta_Y) \cdot T^2(\eta_X, \eta_Y) \gg \frac{\log\log n}{n}.$$

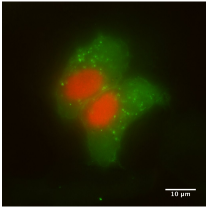
  Then Δ is a consistent test in that we reject H_0 in favor of H_1 with probability tending to one.

- Conversely, there exists a constant $c > 0$ such that for any α-level test Δ based on sample $\{(X_i, Y_i) : i \in \mathbb{I}\}$, there is an instance where joint distribution function F obeying

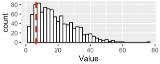
$$\sup_{\eta_X, \eta_Y} V(\eta_X, \eta_Y) \cdot T^2(\eta_X, \eta_Y) \geq c\frac{\log\log n}{n}$$

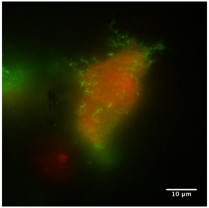
  and yet, we accept H_0 with probability tending to $1 - \alpha$ as if H_0 holds.

Does it REALLY work?

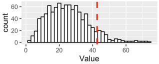


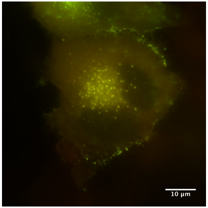
P-value: 0.866



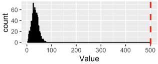


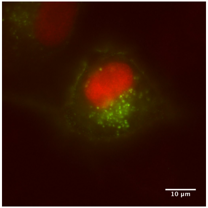
P-value: 0.092



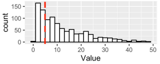


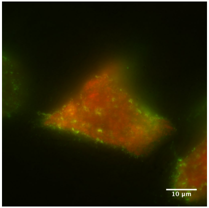
P-value: 0



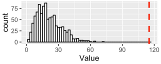


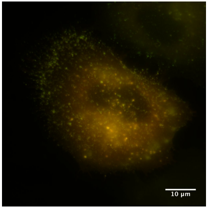
P-value: 0.702



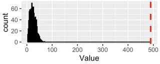


P-value: 0



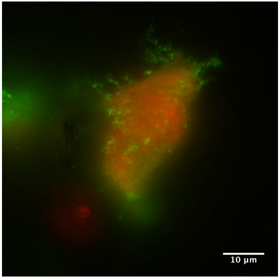


P-value: 0

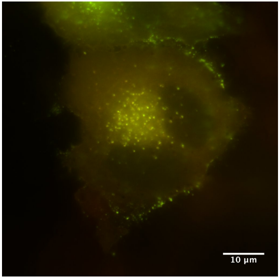


Overview

Where is Colocalization?



$p = 0.092$

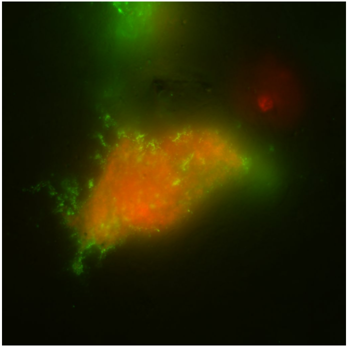


$p = 0$

- Pixel-wise hypothesis:

$$H_{k,0} : F_k \in \mathcal{F}_0 \qquad \text{v.s.} \qquad H_{k,1} : F_k \in \mathcal{F}_1, \qquad k \in \mathbb{I}$$

where F_k is the distribution of (X_k, Y_k).



- Colocalization as tail dependence:

$$H_{k,0} : Q(F_k; \eta_X, \eta_Y) = 0$$

and

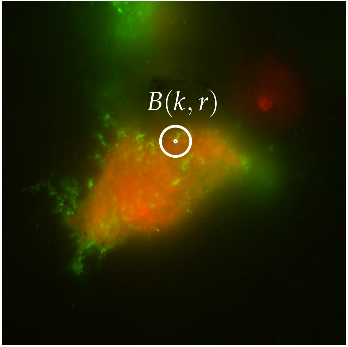
$$H_{k,1} : Q(F_k; \eta_X, \eta_Y) > 0.$$

for pre-specified η_X and η_Y.
- Only one pair (X_k, Y_k) available for each pixel k.

Local Quantification of Colocalization

Weighted Kendall tau's correlation in a neighborhood $B(k, r)$

$$\tau_w(k;r) := \frac{\sum_{i\neq j} w_i(k;r) w_j(k;r) \mathrm{sign}(X_i - X_j)\mathrm{sign}(Y_i - Y_j)}{\sum_{i\neq j} w_i(k;r) w_j(k;r)}$$



- Weight $w_i(k;r)$ is decomposed as

$$w_i(k;r) = \mathbf{K}_l\left(\frac{d(i,k)}{r}\right)\mathbf{K}_b(X_i, Y_i)$$

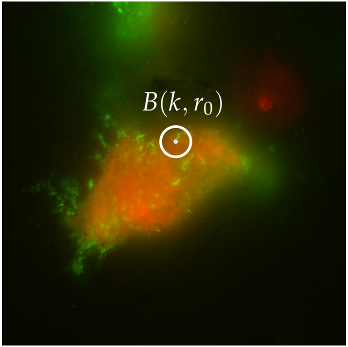
- $\mathbf{K}_l$ gives less weight to the pixel i whose location is far from k.
- $\mathbf{K}_b(X_i, Y_i) = \mathbf{1}_{(X_i > \eta_X)}\mathbf{1}_{(Y_i > \eta_Y)}$ deals with background.

Propagation-Separation

$$w_i(k; r_0)$$

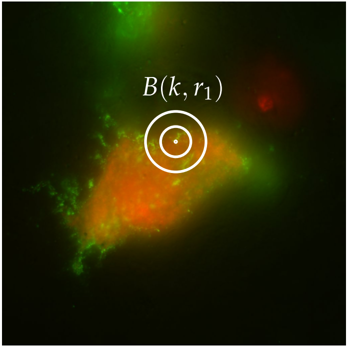
$$\downarrow$$

$$\tau_w(k; r_0)$$



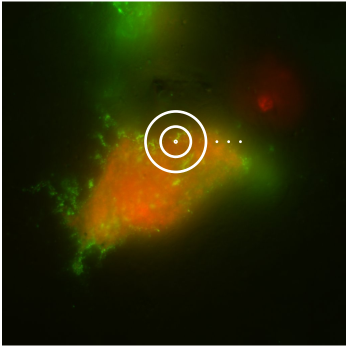
Propagation-Separation

$$\begin{tikzcd}
w_i(k;r_0) \arrow[d] & w_i(k;r_1) \arrow[d] \\
\tau_w(k;r_0) \arrow[ur] & \tau_w(k;r_1)
\end{tikzcd}$$



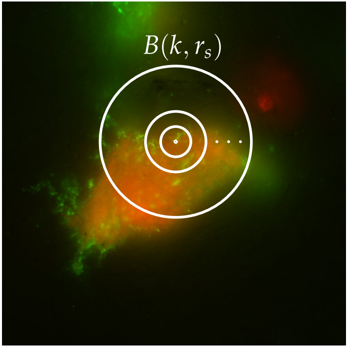
Propagation-Separation

$$\begin{tikzcd} w_i(k;r_0) \arrow[d] & w_i(k;r_1) \arrow[d] & \cdots \\ \tau_w(k;r_0) \arrow[ur] & \tau_w(k;r_1) \arrow[ur] & \end{tikzcd}$$



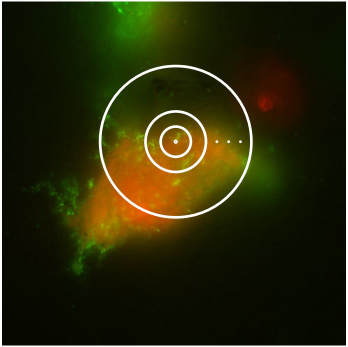
Propagation-Separation

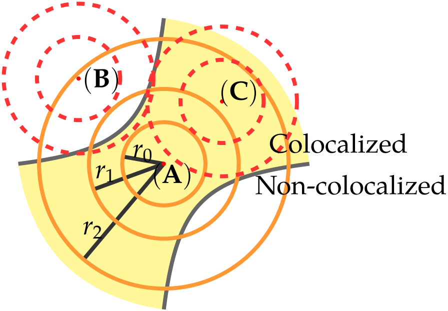
$$
\begin{tikzcd}
w_i(k;r_0) \arrow[d] & w_i(k;r_1) \arrow[d] & \cdots \arrow[dr, phantom, ""] & w_i(k;r_s) \arrow[d] \\
\tau_w(k;r_0) \arrow[ur] & \tau_w(k;r_1) \arrow[ur] & \arrow[ur] & \tau_w(k;r_s)
\end{tikzcd}
$$



Propagation-Separation

$$
\begin{array}{ccccccc}
w_i(k;r_0) & & w_i(k;r_1) & & & & w_i(k;r_s) \\
\downarrow & \nearrow & \downarrow & \nearrow & \cdots & \nearrow & \downarrow \\
\tau_w(k;r_0) & & \tau_w(k;r_1) & & & & \tau_w(k;r_s)
\end{array}
$$





Multiscale Adaptive Test

- Test statistics for $H_{k,0}$ against $H_{k,1}$ is

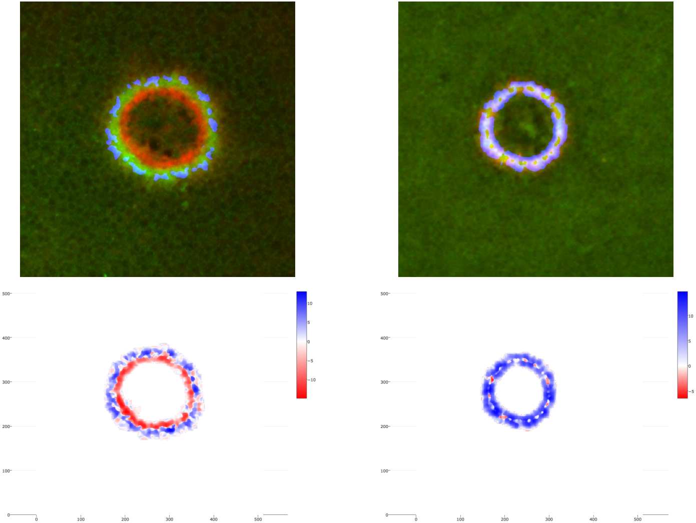
$$Z(k; r_T) = \frac{3}{2}\sqrt{\tilde{N}_k^{(T)}} \cdot \tau_w(k; r_T)$$

  where

$$\tilde{N}_k^{(T)} = \left(\sum_i w_i(k; r_T)\right)^2 \Big/ \sum_i w_i^2(k; r_T).$$

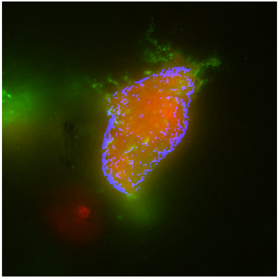
- Under $H_{k,0}$s, $Z(k; r_T)$s behave like standard normal distribution.
- Correct for multiple testing issue by either Bonferroni method, false discovery rate method or random field theory.

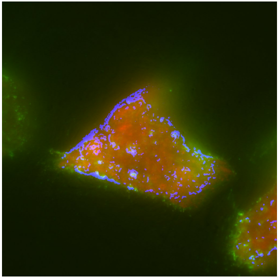
Example

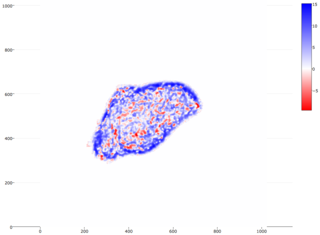


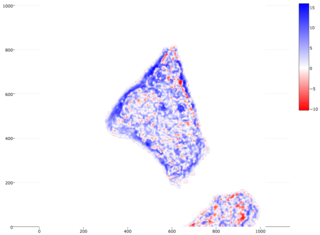
500
400
300
200
100
0
0
100
200
300
400
500
10
5
0
−5
−10
500
400
300
200
100
0
0
100
200
300
400
500
10
5
0
−5

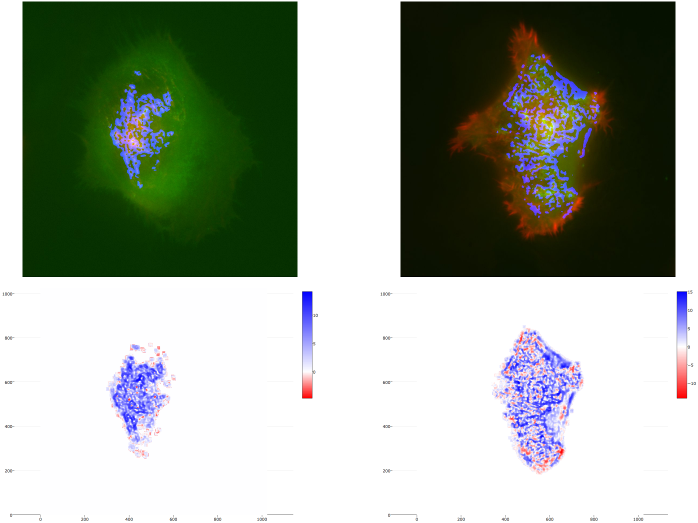
Example



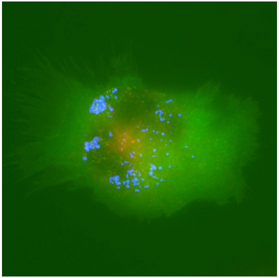


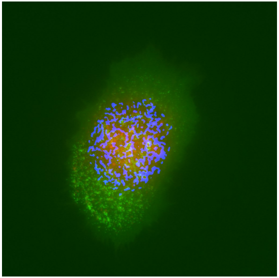

1000
800
600
400
200
0
0
200
400
600
800
1000
15
10
5
0
−5

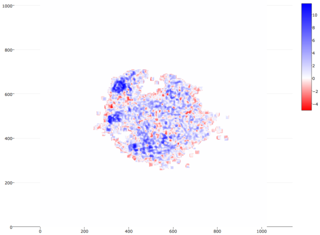

1000
800
600
400
200
0
0
200
400
600
800
1000
15
10
5
0
−5
−10

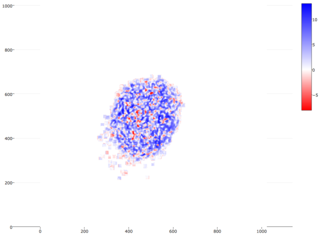

1000
800
600
400
200
0
0
200
400
600
800
1000
10
5
0
15
10
5
0
−5
−10

Example



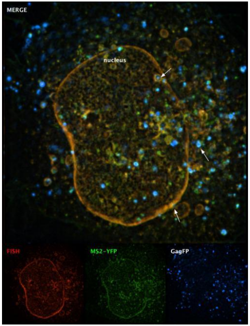


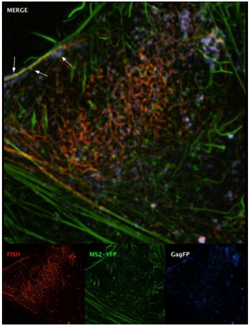

1000
800
600
400
200
0
0
200
400
600
800
1000
10
8
6
4
2
0
−2
−4


1000
800
600
400
200
0
0
200
400
600
800
1000
10
5
0
−5

Overview

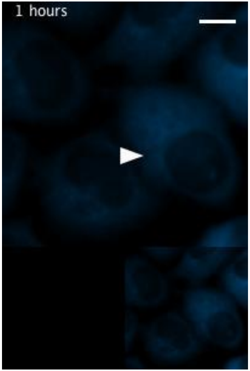
Multiple Channels

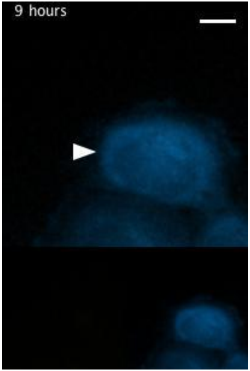




(Becker and Sherer, 2017)

Dynamical Colocalization

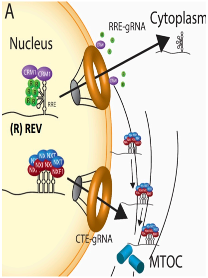




(Becker and Sherer, 2017)

Collaborative Team

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS



Summary

- Colocalization analysis is wide used
- Quantitation in colocalization analysis
- Challenges in quantitative imaging

Thank you!