Analyse statistique de colocalisation spatiale en microscopie à fluorescence

Thibault Lagache[1], Vannary Meas-Yedid[1], Nathalie Sauvonnet[2], Jean-Christophe Olivo-Marin[1]

[1]Unité d'Analyse d'Images Quantitative, Institut Pasteur

[2]Unité de Biologie des Interactions Cellulaires, Institut Pasteur
25 rue du Docteur Roux, 75724 Paris Cedex 15, France
thibault.lagache@pasteur.fr, vmeasyed@pasteur.fr, nathalie.sauvonnet@pasteur.fr, jcolivo@pasteur.fr

Résumé − La mesure de la colocalisation spatiale entre molécules fluorescentes permet de mesurer les interactions possibles entre les différents acteurs d'un processus cellulaire. Cependant, la plupart des études actuelles de colocalisation reposent sur le calcul de coefficients de corrélation entre les pixels significatifs des différents canaux d'acquisition microscopique et peuvent conduire soit à une sur-estimation de la corrélation si la fonction d'étalement du point (PSF) est importante, soit à une sous-estimation quand des spots fluorescents sont proches mais ne se chevauchent pas. De plus, la mesure du niveau de confiance de ces estimateurs de corrélation est le plus souvent basée sur des simulations de Monte-Carlo coûteuses en temps de calcul et dépendant de la géométrie de la zone d'étude. Pour ces deux raisons, nous développons ici une nouvelle méthode statistique basée objet rapide et facile à implémenter, qui repose sur l'estimation analytique des quantiles de la fonction K de Ripley en fonction de la géométrie de la zone d'étude. Elle permet d'estimer si deux populations colocalisent à partir des positions des différents spots obtenues *via* des algorithmes de détection robustes, et sans avoir recours à des simulations de Monte-Carlo. Nous confirmons ensuite la spécificité de notre test grâce à des simulations Monte-Carlo, avant d'en tester la sensibilité sur des données synthétiques. Nous illustrons enfin notre méthode sur des images biologiques.

Abstract − Proteins colocalization in fluorescence microscopy is a key quantitative tool to decipher cellular processes at a molecular level. Most colocalization analysis are based on intensity correlation between different colour channels, which can either lead to colocalization overestimates when proteins point spread functions (PSF) are large, or mis-colocalization when signals of spatially close proteins do not strictly overlap. In addition, the level of significance of correlation coefficients is mostly computed with Monte-Carlo simulations, which are time consuming and depend on the geometry of the study area. In this paper, we present a new object-based method that is both fast and easy to implement. Our statistical method is based on the analytical estimation of the critical quantiles of the Ripley's K function as function of the geometry of the study area. This allows to estimate whether two molecules' populations colocalize directly from the spots positions obtained with robust detection algorithms, circumventing the use of Monte Carlo simulations. Tests against Monte-Carlo simulations and synthetic data show that our method is both sensitive and specific. Finally, we illustrate our method on biologyical images.

1 Introduction

L'analyse de la colocalisation spatiale de spots fluorescents, correspondant par exemple à des molécules marquées, en microscopie permet de mieux comprendre l'organisation spatio-dynamique de processus cellulaires importants. Par exemple, l'analyse de la colocalisation entre des particules virales et des marqueurs spécifiques de certains compartiments cellulaires comme les molécules Rabs a permis d'élucider la dynamique intra-cellulaire de nombreux virus au cours de l'infection [1]. De même, l'analyse de la colocalisation dynamique de certaines molécules à la membrane cellulaire a permis de mieux comprendre l'orchestration moléculaire de l'endocytose [2]. Il est donc essentiel de développer des outils mathématiquement robustes et rapides afin de tester statistiquement et de quantifier la colocalisation entre différentes molécules fluorescentes. Ces méthodes statistiques sont d'autant plus importantes dans des environnement denses en particules où la proximité spatiale de certaines molécules peut être dû au hasard, les deux types de molécules étant distribuées aléatoirement dans la zone d'étude (Fig. 1).

Afin d'analyser statistiquement et de quantifier la colocalisation entre spots, différentes méthodes ont été développées ces dernières années, la plupart mesurant un indice de corrélation à partir du chevauchement spatial des signaux débruités (pixels significatifs) issus des différents canaux d'acquisition microscopique. Ces méthodes **basées sur la corrélation d'intensité** proposent en un coefficient global de corrélation entre les deux canaux, les plus

utilisés étant les coefficients de Pearson [3] et de Manders [4]. Cependant ces méthodes dépendent fortement de la forme de la PSF. En effet, le chevauchement de PSF étendues peut conduire à une sur-estimation de la colocalisation entre spots, notamment dans un environnement dense en particules, alors qu'au contraire, la réduction de la PSF par des méthodes super-résolutives entraîne une sous-estimation des interactions entre spots, certains pouvant être très proches sans se chevaucher.
Par conséquent des méthodes **basées objet** ont été développées en parallèle : celles-ci reposent sur la détection préalable des différents spots avec des algorithmes automatiques robustes comme les méthodes en ondelettes [5] ou en patches [6], puis sur la construction d'un estimateur statistique basé sur les distances entre les centres de masse des différents objets détectés [7, 8]. Cependant, même avec une méthode basée objet, la principale difficulté est alors de savoir si le niveau de colocalisation obtenu est statistiquement significatif ou dû au hasard, les deux populations de molécules étant distribuées aléatoirement dans la zone d'étude (hypothèse nulle, Fig. 1). Les études traitant de cet aspect statistique ([7, 8] entre autres) utilisent des simulations de Monte-Carlo coûteuses en temps de calcul et nécessitant une procédure de calibration en fonction de la géométrie de la zone d'étude et du nombre de molécules.
Afin de répondre à ce problème, nous présentons ici une méthode basée objet analytique où nous calculons des expressions fermées pour les niveaux de confiance statistiques associés à notre estimateur de colocalisation. Ces expressions prennent en compte la géométrie de la zone d'étude ainsi que le nombre de molécules et permettent donc d'éviter les simulations de Monte-Carlo. Plus précisément, nous construisons tout d'abord dans la sous section 2.1 une statistique de test $\tilde{K}_{12}$ basée sur la fonction K de Ripley, qui permet de prendre en compte les relations spatiales entre molécules à différentes échelles. Nous calculons analytiquement les quantiles de $\tilde{K}_{12}$ sous l'hypothèse nulle de non-interaction entre les molécules qui tient de la compte la géométrie de la zone d'étude ainsi que du nombre de molécules. Nous illustrons ensuite dans la sous-section 2.2 la puissance de notre test sur des données synthétiques et *in vivo*.

2 Résultats

2.1 Construction d'un test statistique de colocalisation

La plupart des méthodes statistiques développées pour étudier la colocalisation d'objets reposent sur la fonction K de Ripley, qui s'écrit de façon générique pour une zone d'étude Ω contenant n_1 objets A_1 et n_2 objets A_2 [9] :

$$K_{12}(r) = \frac{|\Omega|}{n_1 n_2} \sum_{\mathbf{x}\in A_1} \sum_{\mathbf{y}\in A_2} \mathbf{1}_{\{|\mathbf{x}-\mathbf{y}|\leq r\}} f(\mathbf{x},\mathbf{y}), \qquad (1)$$

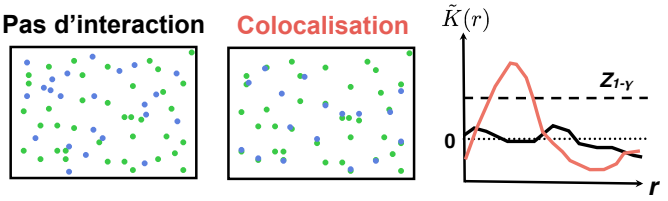

FIGURE 1 – **Analyse statistique de la colocalisation spatiale de molécules.** La colocalisation spatiale de molécules révèle leur interaction au sein d'un processus cellulaire. Cependant certaines molécules pouvant être proches par hasard en étant distribuées aléatoirement dans la zone d'étude, il est essentiel de pouvoir tester statistiquement la colocalisation à un niveau de confiance donné γ en comparant une statistique de test $\tilde{K}(r)$ par rapport au quantile $z_{1-\gamma}$ de la distribution de $\tilde{K}(r)$ sous l'hypothèse nulle de non-interaction entre les molécules. Nous utilisons ici une statistique basée sur la fonction K de Ripley qui permet d'analyser la colocalisation des molécules à différentes échelles grâce au paramètre de distance r.

où $f(\mathbf{x},\mathbf{y})$ est une correction de bord permettant de prendre en compte la sous-estimation possible d'événements de colocalisation à cause de la zone limitée d'étude. Une correction de bord classique est la correction de Ripley [9] $f(\mathbf{x},\mathbf{y}) = \frac{|\partial b(\mathbf{x},|\mathbf{x}-\mathbf{y}|)|}{|\partial b(\mathbf{x},|\mathbf{x}-\mathbf{y}|)\cap\Omega|}$, qui est inversement proportionnelle à la circonférence $\partial b(\mathbf{x},|\mathbf{x}-\mathbf{y}|)$ du disque $b(\mathbf{x},|\mathbf{x}-\mathbf{y}|)$ centré en $\mathbf{x}$ et de rayon $\mathbf{x}-\mathbf{y}$ incluse dans Ω. Avec la correction de Ripley, $\mathbb{E}\{K_{12}(r)\} = \pi r^2$ [9], et approximant localement $\partial\Omega$ au premier ordre par sa tangente nous avons calculé que [10]

$$\mathrm{var}\{K_{12}(r)\} = \frac{|\Omega|}{n_1^2 n_2}\left(\sum_{\mathbf{x}_1\in A_1}\beta(\mathbf{x}_1) + \sum_{\mathbf{x}_2\neq\mathbf{x}_1} A_{12}\right) - \frac{\pi^2 r^4}{n_2}. \qquad (2)$$

où $\beta(\mathbf{x}_1)$ est fonction de la distance $|\mathbf{x}_1 - \partial\Omega|$ de chacun des points $\mathbf{x}_1$ à la frontière $\partial\Omega$ de la zone d'étude [10], et A_{12} est égal à l'aire de l'intersection $|b(\mathbf{x}_1,r)\cap b(\mathbf{x}_2,r)|$, c'est à dire pour $d_{12} = |\mathbf{x}_1 - \mathbf{x}_2|$:

$$A_{12} = \mathbf{1}_{\{d_{12}<2r\}}\left(2r^2\arccos\left(\frac{d_{12}}{2r}\right) - \frac{d_{12}}{2}\sqrt{4r^2-d_{12}^2}\right). \qquad (3)$$

En utilisant les expressions de $\mathbb{E}\{K_{12}(r)\}$ et $\mathrm{var}\{K_{12}(r)\}$, nous avons alors construit la statistique de colocalisation centrée réduite ($\mathbb{E}=0$, var $=1$)

$$\tilde{K}_{12}(r) = \frac{K_{12}(r) - \pi r^2}{\sqrt{\mathrm{var}\{K_{12}(r)\}}}. \qquad (4)$$

et avons démontré [10] que sous l'hypothèse où A_2 est aléatoirement distribué dans Ω selon un processus de Poisson uniforme et ne colocalise pas avec A_1, $\tilde{K}_{12}(r)$ est asymptotiquement normal pour $n_2 \gg 1$: $\tilde{K}_{12}(r) \xrightarrow[n_2\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$. Cela nous a permis de construire un test statistique de colocalisation. En effet, si $z_{1-\gamma}$ est le quantile de

TABLE 1 – Test against Monte Carlo simulations

	$q_{0.99}$	$\frac{\lvert q_{0.99} - z_{0.99}\rvert}{z_{0.99}}$	$q_{0.999}$	$\frac{\lvert q_{0.999} - z_{0.999}\rvert}{z_{0.999}}$
$n = 1$	2.40	3%	3.13	1.5%
$n = 10$, uniform	2.43	4.5%	3.32	7.4%
$n = 10$, cluster	2.37	2%	3.19	3%

niveau $1-\gamma$ de la loi normale $\mathcal{N}(0,1)$, alors

$$\tilde{K}_{12}(r) > z_{1-\gamma} \tag{5}$$

indique que A_1 et A_2 colocalisent dans Ω avec un niveau de confiance au moins égal à γ (Fig. 1).

2.2 Test contre des données synthétiques et *in vivo*

Afin de vérifier la spécificité de notre test statistique, nous avons tout d'abord vérifié que le quantile de la loi normale $z_{1-\gamma}$ était une bonne approximation du quantile $q_{1-\gamma}$ au niveau $1-\gamma$ de $\tilde{K}_{12}(r)$ sous l'hypothèse de distribution aléatoire de A_2. Pour cela nous avons considéré trois distributions spatiales pour les points de A_1 (voir Fig. 2) : $n_1 = 1$ et $n_1 = 10$ distribués aléatoirement dans la zone d'étude Ω (carré 10×10) (Fig. 2 a-b), ou $n_1 = 10$ points agrégés suivant un processus Gaussien bi-dimensionnel $\mathcal{N}(\mathbf{P}, \sigma = 1)$ où $\mathbf{P}$ choisi aléatoirement dans Ω (Fig. 2 c). Nous avons ensuite calculé empiriquement $q_{1-\gamma}$ à l'aide de $N = 10^6$ simulations de Monte-Carlo (assurant ainsi la convergence de $q_{1-\gamma}$) : pour chaque simulation $n_2 = 100$ points de A_2 sont distribués aléatoirement dans Ω et la fonction de Ripley correspondante $\tilde{K}^j_{12}(r)$, pour $1 \leq j \leq N$ est calculée. Le quantile $q_{1-\gamma}$ de $\tilde{K}^j_{12}(r)$ au niveau $1-\gamma = 0.99$ et $1-\gamma = 0.999$ est ensuite obtenu en triant dans l'ordre croissant la liste des valeurs $\tilde{K}^j_{12}(r)$ et en choisissant

$$q_{1-\gamma} = \tilde{K}^{\lfloor (1-\gamma)N \rfloor}_{12}(r) \tag{6}$$

où $\lfloor (1-\gamma)N \rfloor$ est la partie entière de $(1-\gamma)N$. Dans la table 1, nous avons pu vérifier que l'erreur relative $\frac{\lvert q_{1-\gamma} - z_{1-\gamma}\rvert}{q_{1-\gamma}}$ était inférieure à 5%, excepté pour $n_1 = 10$ points distribués aléatoirement et $1-\gamma = 0.999$ ($\frac{\lvert q_{1-\gamma} - z_{1-\gamma}\rvert}{q_{1-\gamma}} = 7.4\%$). Dans ce cas, $\Phi(q_{1-\gamma}) = \Pr\{\mathcal{N}(0,1) < q_{1-\gamma}\} = 0.9995$, qui reste très proche de 0.999.

Nous avons ensuite testé la sensibilité de notre test statistique contre des données synthétiques où les points de A_2 sont soit distribués aléatoirement dans Ω, soit colocalisent partiellement avec les ponts de A_1 : une proportion $\alpha_2 = 0$, $\alpha_2 = 0.2$ ou $\alpha_2 = 0.5$ des points de A_2 ($n_2 = 100$) sont distribués suivant un processus Gaussien $\mathcal{N}(0, \sigma)$ autour des points de A_1 ($n_1 = 100$), les autres $((1-\alpha_2)n_2)$ étant distribués aléatoirement dans Ω. L'écart type σ permet de simuler une distance d'interaction l entre les molécules de l'ordre de $l \approx 2\sigma$. En effet, pour une molécule $\mathbf{x}_2 \in A_2$ interagissant avec une molécule $\mathbf{x}_1 \in A_1$, nous avons $\Pr\{\lvert \mathbf{x}_1 - \mathbf{x}_2 \rvert < 2\sigma\} \approx 99\%$. Dans la table 2 nous avons calculé $\tilde{K}_{12}(r)$ (Eq. 4) pour $r = 0.3, 0.5$ et $r = 1$, et en avons déduit les p-values correspondantes : p-value $= \Phi(\tilde{K}_{12}(r))$, avec Φ la fonction de répartition de la loi Normale standard $\mathcal{N}(0,1)$. Nous démontrons ainsi la spécificité et la sensibilité de notre test $\tilde{K}_{12}$, celui-ci ne pouvant rejeter l'hypothèse nulle de non-interaction entre molécules pour $\alpha_2 = 0$, et détectant significativement la colocalisation même pour $\alpha_2 = 0.2$ ($p_{value} < 1\%$). De plus, nous observons que notre statistique de test est maximale pour $r \approx l$, ce qui permet de déduire la distance d'interaction entre les objets.

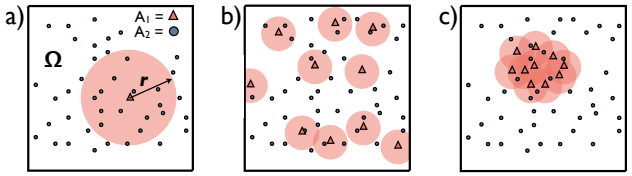

FIGURE 2 – **Distributions spatiales deA_1 considérées pour le test de spécificité.** Nous testons la spécificité de notre test statistique $\tilde{K}_{12}(r)$ contre l'hypothèse nulle où les points de A_2 sont aléatoirement distribués dans la zone d'étude Ω pour trois types de distribution spatiales de A_1 : a) un unique point de A_1 est distribué aléatoirement dans Ω, b) 10 points deA_1 sont distribués aléatoirement dans Ω et c) 10 points de A_1 sont agrégés suivant un processus Gaussien bi-dimensionnel.

Nous avons enfin testé notre méthode sur des données *in vivo* obtenues par l'imagerie en fluorescence multi-canaux de différentes molécules impliquées dans l'endocytose. Dans un premier temps (Fig. 3-A), nous avons analysé statistiquement la colocalisation des molécules A et B formant un complexe A-B impliqué dans la réorganisation du réseau d'actine au cours de l'endocytose, et servant donc ici de contrôle positif. Nous avons observé une colocalisation très significative $\tilde{K}_{12}(r) \gg z_{0.99} = 2.32$, avec un maximum atteint pour $r \approx 2-3$ pixels, révélant une distance d'interaction d'environ 100 nm en accord avec la taille du complexe A-B. Dans un second temps (Fig. 3-B), nous avons étudié le degré de colocalisation de la clathrine et de la cavéoline, qui sont des molécules engagées dans des voies d'endocytose différentes. Celles-ci servent au contraire de contrôle négatif, ce qui a été là aussi confirmé par notre test statistique ($\tilde{K}_{12}(r) < z_{0.99}$).

3 Conclusion

L'analyse statistique de la colocalisation d'objets en microscopie à fluorescence permet de mieux comprendre l'orchestration moléculaire de processus cellulaires complexes comme l'endocytose. Dans ce cadre, nous proposons une approche analytique basée objet permettant de tester statistiquement la colocalisation spatiale de deux populations de molécules marquées. Cette méthode analytique, calculée à partir de la fonction K de Ripley, prend en compte

TABLE 2 – Statistics on synthetic data, $n_2 = 100$

	$\alpha = 0$		$\alpha = 0.2$, $\sigma = 0.1$		$\alpha = 0.2$, $\sigma = 0.3$		$\alpha = 0.5$, $\sigma = 0.1$		$\alpha = 0.5$, $\sigma = 0.3$	
	$\tilde{K}_{12}(r)$	p-value	$\tilde{K}_{12}(r)$	p-value	$\tilde{K}_{12}(r)$	p-value	$\tilde{K}_{12}(r)$	p-value	$\tilde{K}_{12}(r)$	p-value
$r = 0.3$	-0.57	0.72	**4.34**	$\mathbf{7 \times 10^{-6}}$	1.49	0.07	**8.46**	$\mathbf{< 10^{-16}}$	**4.44**	$\mathbf{4.5 \times 10^{-6}}$
$r = 0.5$	-1.33	0.90	1.97	2.5×10^{-2}	**2.42**	$\mathbf{7.8 \times 10^{-3}}$	**4.54**	$\mathbf{2.78 \times 10^{-6}}$	**5.24**	$\mathbf{7.9 \times 10^{-8}}$
$r = 1.0$	-1.56	0.94	0.96	0.17	**2.52**	$\mathbf{5.8 \times 10^{-3}}$	**3.16**	$\mathbf{7.88 \times 10^{-4}}$	**2.96**	$\mathbf{1.5 \times 10^{-3}}$

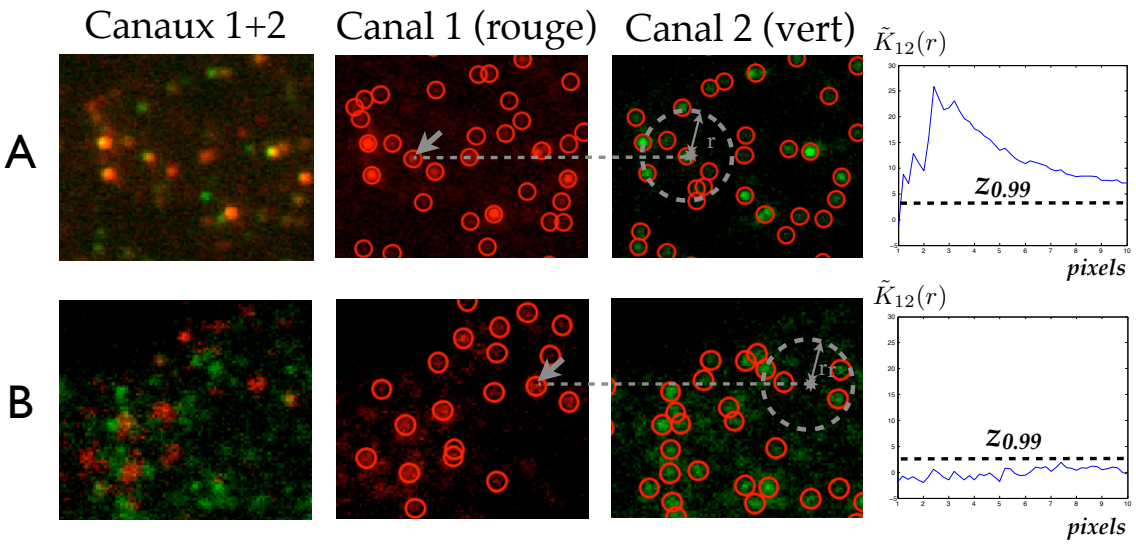

FIGURE 3 – **Analyse *in vivo* de la colocalisation entre molécules impliquées dans l'endocytose**. Après marquage fluorescent, les différentes molécules sont observées avec un microscope de fluorescence par réflexion totale interne (TIRF). Les positions des spots (cercles rouges) sont extraites par un algorithme de détection en ondelettes [5] et leur colocalisation statistique est testée en comparant la fonction modifiée de Ripley $\tilde{K}_{12}(r)$ (Eq.4) par rapport au quantile $z_{0.99} = 2.32$ de la loi normale standard. $\tilde{K}_{12}(r)$ est proportionnelle au nombre moyen de molécules du canal 2 dans un voisinage (rayon r) d'une molécule du canal 1 (schéma en gris). **A-** Les molécules A (rouge) et B (vert) forment un complexe A-B et colocalisent donc de façon très significative : $\tilde{K}_{12}(r) \gg z_{0.99} = 2.32$. **B-** Au contraire les molécules clathrine (rouge) et cavéoline (vert) sont engagées dans des voies d'endocytose différentes et ne colocalisent pas : $\tilde{K}_{12}(r) < z_{0.99}$.

la géométrie du domaine d'intérêt permettant ainsi de s'affranchir des simulations de Monte-Carlo. Notre test en plus de sa rapidité, s'est révélé spécifique et sensible sur des données synthétiques, ainsi que sur des données biologiques *in vivo*.

Références

[1] M. Lakadamyali, M. J. Rust, H. P. Babcock, and X. Zhuang, "Visualizing infection of individual influenza viruses," *Proc Natl Acad Sci U S A*, vol. 100, no. 16, pp. 9280–5, Aug 2003.

[2] M. J. Taylor, D. Perrais, and C. J. Merrifield, "A high precision survey of the molecular dynamics of mammalian clathrin-mediated endocytosis," *PLoS Biol*, vol. 9, no. 3, p. e1000604, Mar 2011.

[3] S. V. Costes *et al.*, "Automatic and quantitative measurement of protein-protein colocalization in live cells," *Biophys. J.*, vol. 86, pp. 3993–4003, 2004.

[4] E. Manders *et al.*, "Dynamics of three-dimensional replication patterns during the S-phase, analysed by double labelling of DNA and confocal microscopy," *J. Cell Sci.*, vol. 103, pp. 857–862, 1992.

[5] J. C. Olivo-Marin, "Extraction of spots in biological images using multiscale products," *Pattern Recognition*, vol. 35, no. 9, pp. 1989–1996, 2002.

[6] J. Boulanger, A. Gidon, C. Kervran, and J. Salamero, "A patch-based method for repetitive and transient event detection in fluorescence imaging," *PLoS ONE*, vol. 5, no. 10, p. e13190, 10 2010.

[7] B. Zhang, N. Chenouard, J.-C. Olivo-Marin, and V. Meas-Yedid, "Statistical colocalization in biological imaging with false discovery control," in *ISBI*, 2008, pp. 1327–1330.

[8] E. Diaz *et al.*, "Measuring spatiotemporal dependencies in bivariate temporal random sets with applications to cell biology," *IEEE Trans. on PAMI*, vol. 30, no. 9, pp. 1659 –1671, sept. 2008.

[9] B. Ripley, *Statistical inference for spatial processes*. Cambridge University Press, 1988.

[10] T. Lagache, V. Meas-Yedid, and J.-C. Olivo-Marin, "A statistical analysis of spatial colocalization using ripley's k function," in *IEEE International Symposium on Biological Imaging (ISBI)*, 2013.