

A Bayesian Framework for Cell-Level Protein Network Analysis for Multivariate Proteomics Image Data

Violeta N. Kovacheva^a, Korsuk Sirinukunwattana^b and Nasir M. Rajpoot^{b, c}

^aDepartment of Systems Biology, The University of Warwick, Coventry CV4 7AL, UK;

^bDepartment of Computer Science, The University of Warwick, Coventry CV4 7AL, UK;

^cDepartment of Computer Science and Engineering, Qatar University, Qatar

ABSTRACT

The recent development of multivariate imaging techniques, such as the Toponome Imaging System (TIS), has facilitated the analysis of multiple co-localisation of proteins. This could hold the key to understanding complex phenomena such as protein-protein interaction in cancer. In this paper, we propose a Bayesian framework for cell-level network analysis allowing the identification of several protein pairs having significantly higher co-expression levels in cancerous tissue samples when compared to normal colon tissue. It involves segmenting the DAPI-labeled image into cells and determining the cell phenotypes according to their protein-protein dependence profile. The cells are phenotyped using Gaussian Bayesian hierarchical clustering (GBHC) after feature selection is performed. The phenotypes are then analysed using Difference in Sums of Weighted cO-dependence Profiles (DiSWOP), which detects differences in the co-expression patterns of protein pairs. We demonstrate that the pairs highlighted by the proposed framework have high concordance with recent results using a different phenotyping method. This demonstrates that the results are independent of the clustering method used. In addition, the highlighted protein pairs are further analysed via protein interaction pathway databases and by considering the localisation of high protein-protein dependence within individual samples. This suggests that the proposed approach could identify potentially functional protein complexes active in cancer progression and cell differentiation.

Keywords: Bayesian hierarchical clustering, colon cancer, protein interaction, biomarkers, multi-tag imaging, TIS.

1. INTRODUCTION

Over the last few years several multivariate imaging techniques have been developed. These include the Toponome Imaging System (TIS),¹ MxIF,² Matrix-assisted laser desorption/ionization (MALDI) imaging,³ Raman microscopy⁴ and multi-spectral imaging methods.⁵ TIS is an automated high-throughput technique able to co-map up to a hundred different proteins or other tag-recognisable bio-molecules in the same pixel on a single tissue section without damaging it.⁶ It runs cycles of fluorescence tagging, imaging and soft bleaching *in situ*. These techniques present a new challenge for the development of analytical tools that can extract useful, quantitative information from the large amounts of data obtained.

Previous work using TIS has demonstrated the importance of colocalisation of proteins rather than abundance on its own. Despite the spherical and the exploratory cell states of rhabdomyosarcoma cells having identical average protein profiles, striking differences were found between the two states at the sub-cellular protein cluster level.⁷ Hence, rearrangement, rather than up- or down-regulation of proteins is (or can be) key to generating new cell functionalities.⁶ Furthermore, when comparing different samples, the differences in image intensities could be a result of differences in the staining or image acquisition processes, rather than due to different protein expression. Using the protein-protein dependence (PPD) substantially diminishes this problem. While colocalisation does not necessarily imply interaction, co-dependence between two proteins is an indication of a possible interaction, which may be indirect. Therefore, any protein interaction suggested using TIS would need to be further investigated using other experimental techniques.

Further author information: (Send correspondence to V.N.K.)

V.N.K.: E-mail: v.n.kovacheva@warwick.ac.uk

N.M.R.: E-mail: Nasir.Rajpoot@ieee.org

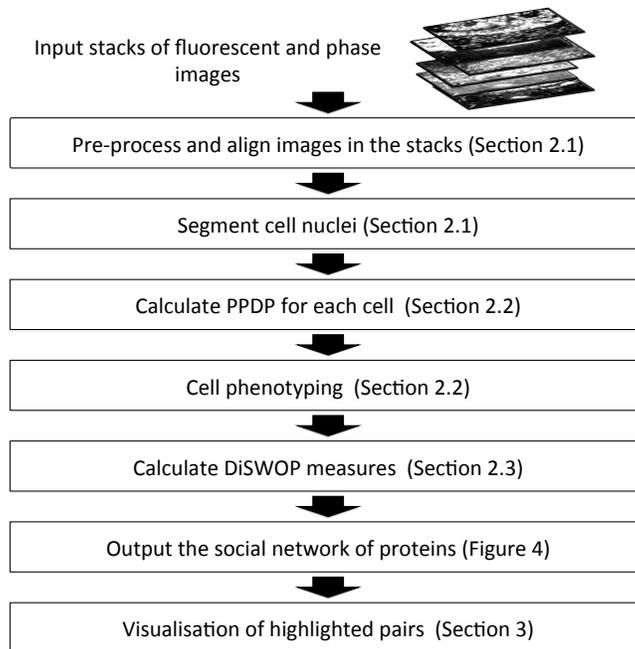


Figure 1. Outline of the method presented.

Kovacheva *et al.* (2013)⁸ present a novel measure, DiSWOP, for detecting protein pairs with different degree of co-localisation in cancerous and normal tissue. Once thoroughly validated, the pairs highlighted could be used as multiplex biomarkers for colon cancer. By analysing the expression of multiple proteins in relation to each other, this measure has the potential to improve the low prognostic value of the simple biomarkers currently used in clinical practice.^{9,10}

The analytical framework presented involves segmenting the DAPI-labeled image into cells and determining the cell phenotypes according to their protein-protein dependence profile (PPDP). An outline of the methodology is shown in Figure 1. In the earlier paper, cells are phenotyped using Affinity Propagation (AP)¹¹ clustering. Here, we present a novel Bayesian framework for phenotyping using Gaussian Bayesian hierarchical clustering (GBHC)¹² and analyse the phenotypes with the DiSWOP measure. Although we expect the DiSWOP scores to differ slightly, the protein pairs highlighted by the measure should be independent of the phenotyping method used. We then further verify the significance of the highlighted protein pairs using protein interaction pathways and by considering differences in localisation of high PPD between cancerous and normal samples.

2. MATERIALS AND METHODS

2.1 Data Acquisition and Pre-processing

The results presented here were obtained by considering a total of 11 samples of colon tissue 6 healthy and 5 cancerous. A library of 12 antibody tags, some of which are known tumour markers or cancer stem cell markers, were used based on the findings by Bhattacharya *et al.*¹³ CD133, CK19, Cyclin A, Muc2, CEA, CD166, CD36, CD44, CD57, CK20, Cyclin D1 and EpCAM were used in the analysis and a DAPI tag was included to identify the cell nuclei. Background autofluorescence is digitally subtracted at an early stage and hence any remaining fluorescence should be true protein expression. Sample images are shown in Figure 2. In each of the stacks, the images were aligned using the RAMTaB (Robust Alignment of Multi-Tag Bioimages) algorithm.¹⁴ This is done in order to prevent possible noise in the protein interactions that would result from the slight mis-alignment of the multi-tag images obtained using TIS. Each image stack is then segmented into individual cells so that analysis is restricted to cellular areas only. This is done using a modified form of the graph cut method¹⁵ applied to a DAPI channel.¹⁶ The nuclei segmented are used as a rough approximation of the cells (Figure 3). A total of 2945 cells were obtained. This is the same data as used in Kovacheva *et al.*⁸

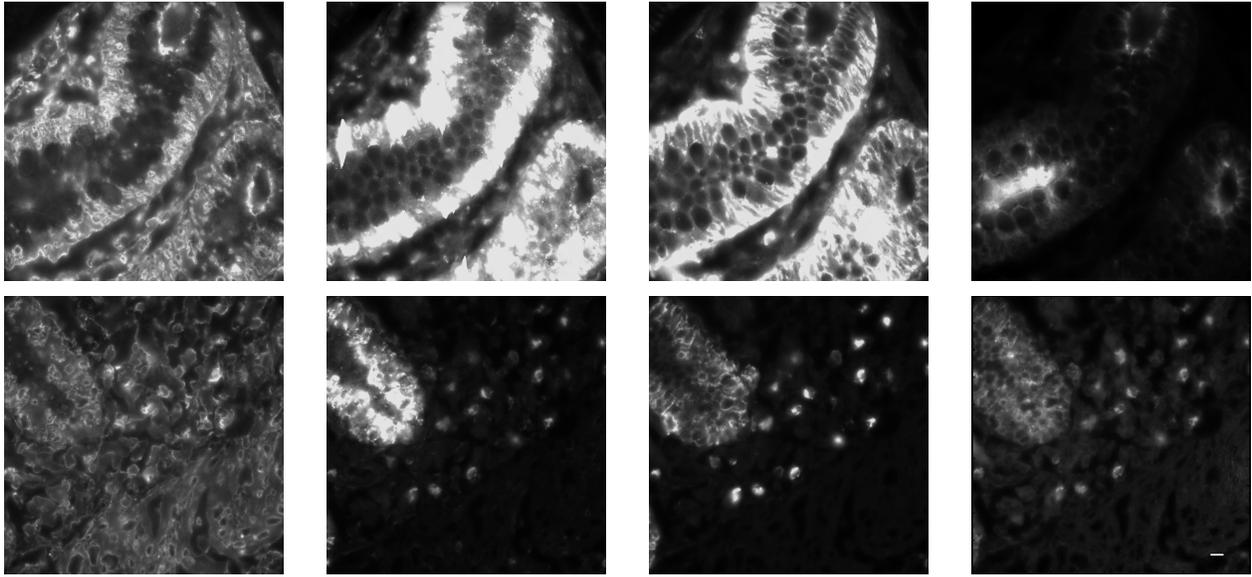


Figure 2. Sample data images. The columns show CD133, Muc2, EpCAM and CEA respectively. The first row shows images from a healthy sample and the second row shows images from a cancerous sample. The length of the scale bar is $10 \mu m$

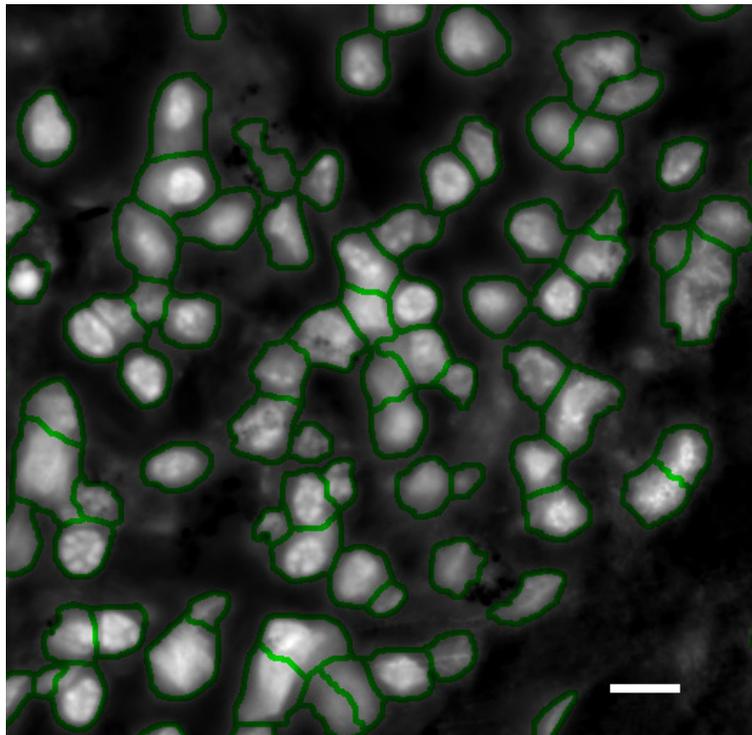


Figure 3. Segmentation of the cell nuclei on a part of a colon cancer sample. The outline of each identified nucleus is shown in green. The length of the scale bar is $10\mu m$.⁸

2.2 Cell Phenotyping

For each pair of proteins, localised to each individual cell, the maximal information coefficient (MIC)¹⁷ is calculated to obtain the protein-protein dependence profile (PPDP) of the cell. This statistic has been used since it has been shown to capture a wide range of associations, both functional and not, and it gives similar scores to equally noisy relationships of different types.¹⁷ The protein pairs that best discriminate between cancer and normal samples were selected using the Wilcoxon rank sum test.¹⁸ For a protein pair, this was done by calculating the p-value that the PPD values of the cancer cells and of the normal cells come from distributions with a different median. Then, out of the 66 protein pairs, the 33 with lowest p-values were selected for clustering to be performed on. This drastically speeds up performance of the algorithm. Cells with similar phenotype are expected to have PPDPs with similar nature. In terms of probability, we can hypothesise that different phenotypes are explained by different probability distribution, and cells with similar phenotype should come from the same distribution. Cell phenotyping can therefore be achieved through GBHC,¹² which models data as a mixture of probability distributions.

Let the PPDP of the i^{th} cell be denoted by $x^{(i)} = (x_1^{(i)}, \dots, x_d^{(i)})$, where $x_j^{(i)}$ is a PPD value of the j^{th} protein pair for the i^{th} cell, and $d = 33$ is the number of protein pairs after feature selection. Without loss of generality, we assume that the whole PPD data have zero mean and unit variance. Let D_k denote a set of PPD data for n_k cells belonging to the k^{th} phenotype. According to the assumptions of GBHC, for each protein pair j , $x_j^{(i)}$ are independent and identically normal distributed with unknown mean μ_j and variance σ_j^2 , i.e.

$$x_j^{(i)} \sim \mathcal{N}(x|\mu_j, \sigma_j^2) \quad \forall x^{(i)} \in D_k. \quad (1)$$

Furthermore, μ_j and σ_j^2 are assumed to be normal-gamma distributed with hyper-parameters λ_0 , β_0 , and κ_0 . The marginal likelihood of D_k based on this hierarchical probabilistic model can be expressed as

$$P(D_k|\lambda_0, \beta_0, \kappa_0) = \prod_{j=1}^d \left[\frac{\Gamma(\lambda_{n_k})}{\Gamma(\lambda_0)} \frac{\beta_0^{\lambda_0}}{\beta_{n_k,j}^{\lambda_{n_k}}} \left(\frac{\kappa_0}{\kappa_{n_k}} \right)^{\frac{1}{2}} (2\pi)^{\frac{-n_k}{2}} \right], \quad (2)$$

where

$$\lambda_0, \beta_0, \kappa_0 > 0, \quad (3)$$

$$\kappa_{n_k} = \kappa_0 + n_k, \quad (4)$$

$$\lambda_{n_k} = \lambda_0 + \frac{n_k}{2}, \quad (5)$$

$$\bar{x}_j = \frac{1}{n_k} \sum_{i=1}^{n_k} x_j^{(i)}, \quad (6)$$

$$\beta_{n_k,j} = \beta_0 + \frac{1}{2} \left[\sum_{i=1}^{n_k} \left(x_j^{(i)} - \bar{x}_j \right)^2 + \frac{\kappa_0 n_k \bar{x}_j^2}{\kappa_{n_k}} \right], \quad (7)$$

and $\Gamma(\cdot)$ denotes a gamma function. This likelihood term indicates how likely it is that cells in D_k have the same phenotype, and it will be used as an alternative to a distance-based dissimilarity measure, which is normally used in agglomerative hierarchical clustering methods.

GBHC uses Bayesian model selection to decide which pair of small data sets D_k and D_l is the most probable to belong to the same distribution, and should be merged together to form a larger data set D_m . This is done through Bayes rule:

$$r_m = \frac{\pi_m P(D_m|\lambda_0, \beta_0, \kappa_0)}{\pi_m P(D_m|\lambda_0, \beta_0, \kappa_0) + (1 - \pi_m) P(D_k|\lambda_0, \beta_0, \kappa_0) P(D_l|\lambda_0, \beta_0, \kappa_0)}, \quad (8)$$

in which $P(D_k|\lambda_0, \beta_0, \kappa_0)$ is the marginal likelihood of a cluster D_k as defined in Equation 2, $\pi_m = \alpha \Gamma(n_m) / \rho_m$, $\rho_m = \alpha \Gamma(n_m) + \rho_k \rho_l$, we set $\pi_k = 1, \rho_k = \alpha$ for every initial cluster set and α is a concentration parameter related

to the expectation of the number of clusters in the data. As we climb up a hierarchical tree, the probability that two clusters being merged come from the same distribution gets lower. Using this information, GBHC does not consider merges with probability less than 0.5 as valid merges. This in turn results in the algorithm automatically giving the final number of clusters, here found to equal 25.

Since there is no ground truth available for the number and distribution of cell phenotypes in these samples, evaluating the accuracy of the clustering methods is challenging. Hence, this clustering method was selected mainly due to its contrasting approach from AP clustering. This allows us best to demonstrate the robustness of the DiSWOP results.

2.3 Calculating DiSWOP

Once the phenotypes are obtained, the DiSWOP⁸ measure is calculated on the average PPDPs of the cell phenotypes. This is done as follows. For each cluster D we obtain a mean PPDP, $\bar{x}_D = (\bar{x}_1, \dots, \bar{x}_{\hat{d}})$, where $\hat{d} = 66$ is the total number of protein pairs. Suppose we only want to consider the top N PPD values of the M most frequent phenotypes in each sample. Let \hat{x}_D be the vector with the elements of \bar{x}_D (lying in $[0, 1]$) sorted in descending order, p_D^r is the probability of phenotype D in sample r , $D_{\alpha,r}$ is the α^{th} most frequent phenotype in sample r , and

$$X_D^j = \begin{cases} \bar{x}_D^j, & \text{if } \bar{x}_D^j \text{ is one of the first } N \text{ elements of } \hat{x}_D, \\ 0, & \text{otherwise} \end{cases}, \quad (9)$$

then the DiSWOP value for a protein pair j , w_j , is given by

$$w_j = \frac{1}{|\psi|} \sum_{r \in \psi} \sum_{\alpha=1}^M p_{D_{\alpha,r}}^r X_{D_{\alpha,r}}^j - \frac{1}{|\nu|} \sum_{r \in \nu} \sum_{\alpha=1}^M p_{D_{\alpha,r}}^r X_{D_{\alpha,r}}^j. \quad (10)$$

where ψ and ν are the sets of cancerous and normal samples, respectively. Hence, the measure weights the dependency score of a protein pair with the phenotype probability in the sample, and sums all occurrences of the protein pair in all the cancerous samples and of all the normal samples. The two sums are normalised by the number of samples that they were obtained from. The score for the normal samples is subtracted from the score for the cancer samples. Subsequently, we obtain a positive score if a pair appears more frequently and with higher dependency scores in the cancerous samples than in normal samples.

3. RESULTS AND DISCUSSION

The top and bottom 10% of the DiSWOP results are shown in Figure 4. It is encouraging to see that most of the protein pairs highlighted are the same as the ones found when Affinity Propagation (AP) is used for phenotyping.⁸ In particular the CEA and EpCAM protein pair comes out as one of the pairs that is more co-localised in cancer tissue and CD36 and CD57 as more co-localised in normal tissue. In fact, when we compare the full two sets of results there is very high agreement as to which pairs have high positive or negative DiSWOP values. In order to quantitatively evaluate the similarity between the networks we calculate distance measures between the vectors containing the DiSWOP values. The L_1 norm between all the edges in the graphs shown is 0.636 and the mean of the relative absolute difference between the edge weights, as defined by

$$mean \left(\frac{|w^{(1)} - w^{(2)}|}{\max(|w^{(1)}|, |w^{(2)}|)} \right) \quad (11)$$

is found to be 0.683, where $w^{(1)}$ and $w^{(2)}$ are the weights of each of the two graphs shown in Figure 4, respectively. The later measure can take values between 0 and 2, with 0 meaning that all the edges are the same, 1 meaning that non of the edges co-occur and 2 showing that all $w^{(1)} = -w^{(2)}$. On the other hand, when all of the edges (thresholded and non-thresholded) are considered, the L_1 norm is 0.89 and the mean of the relative absolute difference between the edge weights is found to be 0.561. In both cases, the maximum absolute difference between the edge weights is 0.0624.

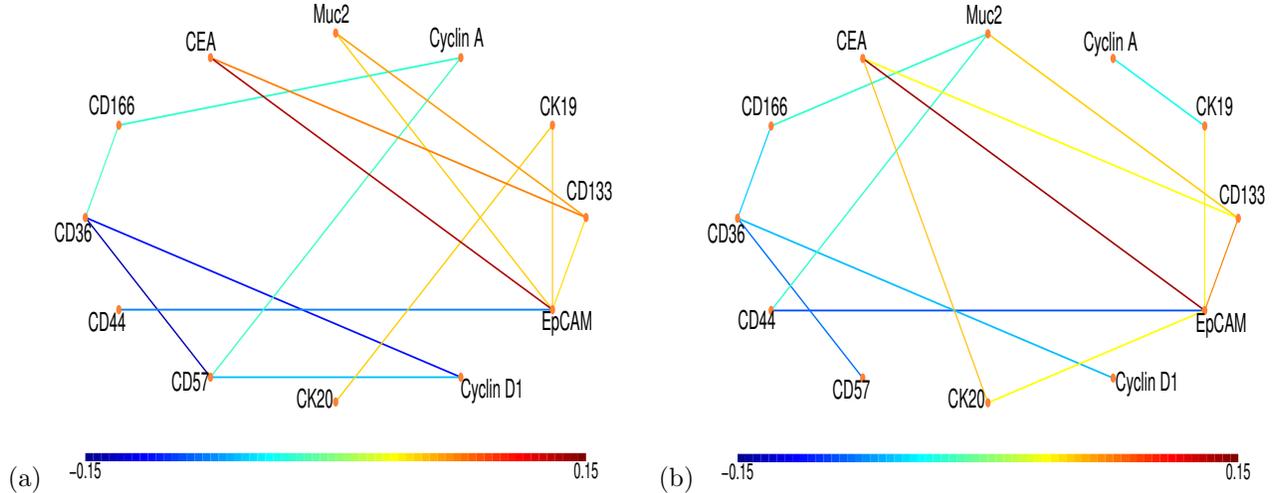


Figure 4. The interaction networks of proteins. Each node represents a protein and each edge colour shows a protein pair with different level of co-expression in the normal and cancer samples. The range of values in the colour bar has been slightly extended so that colours between the two figures correspond. Here, a large positive value (shown in red) indicates that the protein pair is more co-dependent in cancer samples, whereas a large negative value (shown in blue) means that the protein pair is more active in normal tissue. Only edges with the top and bottom 10% of the DiSWOP values are shown. Figure (a) shows the results obtained using GBHC after feature selection and figure (b) shows the results found using AP⁸

In addition, some of these protein pairs have been experimentally found to interact or to be part of a pathway involved in colorectal cancer. For example, several studies have established that CEA and EpCAM interact through the pathway CEA – SOX9 – Claudin7 – EpCAM^{19–22} (Figure 5), which plays an important role in determining the morphology of the colon epithelium and promotes colorectal cancer progression.²²

Further analysis of the results have been performed using an interactive tool for localisation of high PPD within the different samples, as shown in Figure 6. It enables the user to consider two protein pairs simultaneously and see where their PPD is above manually set thresholds. Alternatively, there is the option to see all cells in the samples coloured corresponding to the dependence between a selected protein pair. In this case, the PPDs are binned in intervals of size 0.2 and each cell is displayed in a corresponding colour. A screenshot of the tool has been shown in Figure 6, which shows the cells expressing high PPD (above 0.7) of the two pairs CEA and EpCAM, and CD36 and CD57. We can easily see that normal and cancerous samples show differences in the distribution of high PPD for these two protein pairs. This tool confirms the heterogeneity of protein co-localisation both of neighbouring cells within the same tissue specimen and between different cancerous and normal samples. It could help identify complex biomarkers for cancer stem cells or cancer prognosis.

4. CONCLUSIONS

We have presented a Bayesian framework for phenotyping cells according to their protein-protein dependence profiles. Analysing the results using the DiSWOP measure highlights protein pairs that could play an important role in carcinogenesis. Our results are in concordance with the results of recently published study, which suggests that the DiSWOP measure could help unravel the protein interaction networks involved in cancer. The highlighted protein pairs have been further analysed via protein interaction pathway databases and by considering the localisation of high protein-protein dependence within individual samples. Such analysis could aid our understanding of interactions between neighbouring cells and of the heterogeneity within and between cancer samples.

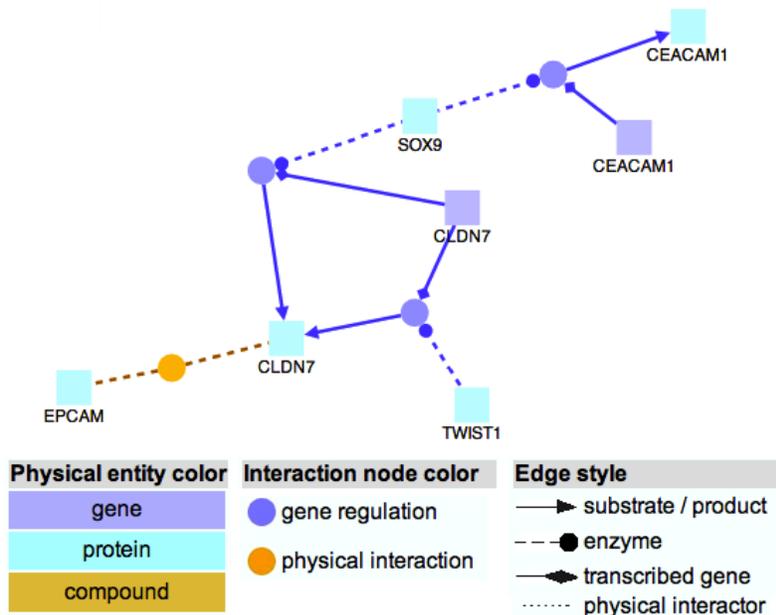


Figure 5. CEA and EpCAM interaction pathway.¹⁹ Sox 9 has been found to activate expression of CEA²⁰ and mediate repression of Claudin-7 by Tcf-4.²¹ Claudin-7 and EpCAM have been found to co-express in colon tissue and possibly be part of a complex.²²

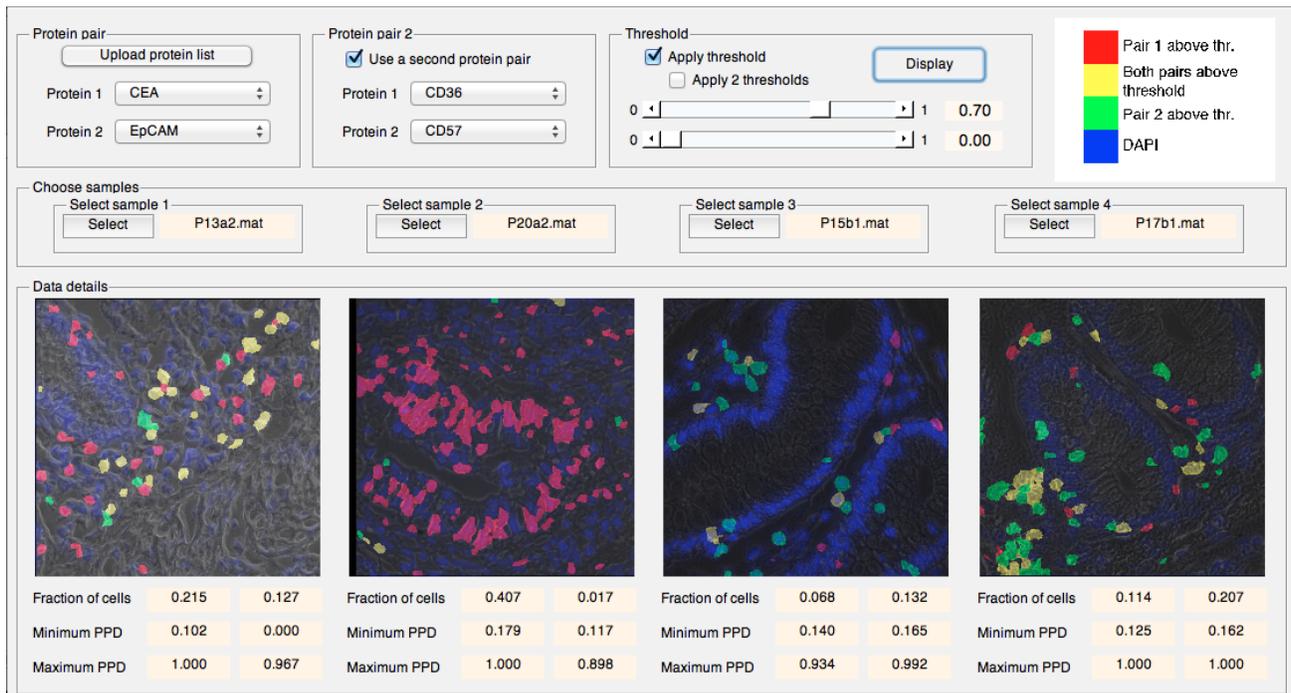


Figure 6. Screenshot of the interactive tool for high PPD localisation. The tool displays the location of PPD above a threshold of 0.7 between CEA and EpCAM (in red) and between CD57 and CD36 (in green). Overlap between the two is shown in yellow and other nuclear regions are shown in blue. Colours are overlaid on top of a phase image. The two samples on the left are cancerous and the two on the right are healthy tissue. Below each sample is information about the fraction of cells above the threshold, the minimum and maximum PPD between each of the two protein pairs.

REFERENCES

- [1] Schubert, W., Bonnekoh, B., Pommer, A. J., Philipsen, L., Böckelmann, R., Malykh, Y., Gollnick, H., Friedenberger, M., Bode, M., and Dress, A. W., “Analyzing proteome topology and function by automated multidimensional fluorescence microscopy,” *Nature biotechnology* **24**(10), 1270–1278 (2006).
- [2] Gerdes, M. J., Sevinsky, C. J., Sood, A., Adak, S., Bello, M. O., Bordwell, A., Can, A., Corwin, A., Dinn, S., Filkins, R. J., et al., “Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue,” *Proceedings of the National Academy of Sciences* **110**(29), 11982–11987 (2013).
- [3] Cornett, D. S., Reyzer, M. L., Chaurand, P., and Caprioli, R. M., “Maldi imaging mass spectrometry: molecular snapshots of biochemical systems,” *Nature Methods* **4**(10), 828–833 (2007).
- [4] van Manen, H.-J., Kraan, Y. M., Roos, D., and Otto, C., “Single-cell raman and fluorescence microscopy reveal the association of lipid bodies with phagosomes in leukocytes,” *Proceedings of the National Academy of Sciences of the United States of America* **102**(29), 10159–10164 (2005).
- [5] Barash, E., Dinn, S., Sevinsky, C., and Ginty, F., “Multiplexed analysis of proteins in tissue using multi-spectral fluorescence imaging,” *Medical Imaging, IEEE Transactions on* **29**(8), 1457–1462 (2010).
- [6] Schubert, W., Gieseler, A., Krusche, A., Serocka, P., and Hillert, R., “Next-generation biomarkers based on 100-parameter functional super-resolution microscopy tis,” *New biotechnology* **29**(5), 599–610 (2012).
- [7] Schubert, W., “On the origin of cell functions encoded in the toponome,” *Journal of biotechnology* **149**(4), 252–259 (2010).
- [8] Kovacheva, V. N., Khan, A. M., Khan, M., Epstein, D., and Rajpoot, N. M., “Diswop: A novel measure for cell-level protein network analysis in localised proteomics image data,” *Bioinformatics* , btt676 (2013).
- [9] Vucic, E. A., Thu, K. L., Robison, K., Rybaczyk, L. A., Chari, R., Alvarez, C. E., and Lam, W. L., “Translating cancer omics to improved outcomes,” *Genome research* **22**(2), 188–195 (2012).
- [10] Evans, R. G., Naidu, B., Rajpoot, N. M., Epstein, D., and Khan, M., “Toponome imaging system: multiplex biomarkers in oncology,” *Trends in molecular medicine* **18**(12), 723–731 (2012).
- [11] Frey, B. J. and Dueck, D., “Clustering by passing messages between data points,” *Science* **315**(5814), 972–976 (2007).
- [12] Sirinukunwattana, K., Savage, R. S., Bari, M. F., Snead, D. R., and Rajpoot, N. M., “Bayesian hierarchical clustering for studying cancer gene expression data with unknown statistics,” *PloS one* **8**(10) (2013).
- [13] Bhattacharya, S., Mathew, G., Ruban, E., Epstein, D. B., Krusche, A., Hillert, R., Schubert, W., and Khan, M., “Toponome imaging system: in situ protein network mapping in normal and cancerous colon from the same patient reveals more than five-thousand cancer specific protein clusters and their subcellular annotation by using a three symbol code,” *Journal of Proteome Research* **9**(12), 6112–6125 (2010).
- [14] Raza, S. e. A., Humayun, A., Abouna, S., Nattkemper, T. W., Epstein, D. B., Khan, M., and Rajpoot, N. M., “Ramtab: robust alignment of multi-tag bioimages,” *PloS one* **7**(2), e30894 (2012).
- [15] Al-Kofahi, Y., Lassoued, W., Lee, W., and Roysam, B., “Improved automatic detection and segmentation of cell nuclei in histopathology images,” *Biomedical Engineering, IEEE Transactions on* **57**(4), 841–852 (2010).
- [16] Khan, A. M., Humayun, A., Raza, S. e. A., Khan, M., and Rajpoot, N. M., “A novel paradigm for mining cell phenotypes in multi-tag bioimages using a locality preserving nonlinear embedding,” *Neural Information Processing* **7666**, 575–583 (2012).
- [17] Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C., “Detecting novel associations in large data sets,” *Science* **334**(6062), 1518–1524 (2011).
- [18] Wilcoxon, F., “Individual comparisons by ranking methods,” *Biometrics bulletin* **1**(6), 80–83 (1945).
- [19] Kamburov, A., Wierling, C., Lehrach, H., and Herwig, R., “Consensuspathdba database for integrating human functional interaction networks,” *Nucleic Acids Research* **37**(suppl 1), D623–D628 (2009).
- [20] Zalzali, H., Naudin, C., Bastide, P., Quittau-Prevostel, C., Yaghi, C., Poulat, F., Jay, P., and Blache, P., “Ceacam1, a sox9 direct transcriptional target identified in the colon epithelium,” *Oncogene* **27**(56), 7131–7138 (2008).

- [21] Darido, C., Buchert, M., Pannequin, J., Bastide, P., Zalzal, H., Mantamadiotis, T., Bourgaux, J.-F., Garambois, V., Jay, P., Blache, P., et al., "Defective claudin-7 regulation by tcf-4 and sox-9 disrupts the polarity and increases the tumorigenicity of colorectal cancer cells," Cancer Research **68**(11), 4258–4268 (2008).
- [22] Kuhn, S., Koch, M., Nübel, T., Ladwein, M., Antolovic, D., Klingbeil, P., Hildebrand, D., Moldenhauer, G., Langbein, L., Franke, W. W., et al., "A complex of epcam, claudin-7, cd44 variant isoforms, and tetraspanins promotes colorectal cancer progression," Molecular Cancer Research **5**(6), 553–567 (2007).