Advances in Discovery of Complex Biomarkers for Colorectal Cancer Using Multiplexed Proteomics Imaging

Violeta N Kovacheva[1]

David BA Epstein[2]

Nasir M Rajpoot[3,4]

[1] Department of Systems Biology,
The University of Warwick,
Coventry CV4 7AL, UK.
[2] Mathematics Institute,
The University of Warwick,
Coventry CV4 7AL, UK.
[3] Department of Computer Science,
The University of Warwick,
Coventry CV4 7AL, UK.
[4] Department of Computer Science and Engineering,
Qatar University, Qatar.

Correspondence:
Violeta N. Kovacheva,
E: v.n.kovacheva@warwick.ac.uk

Nasir M. Rajpoot,
E: Nasir.Rajpoot@ieee.org

Abstract

Multiplexed proteomics imaging techniques such as the Toponome Imaging System (TIS) can yield high-resolution images of multiple proteins co-localised within individual cells. This enables the study of protein interactions and tumour heterogeneity both within and between cancer samples. Our group has recently developed methods for cell-level analysis of the multiplexed proteomics image data obtained from colorectal cancer samples. These methods together with the highly informative multiplexed proteomics image data hold great promise for discovering complex biomarkers that can aid the development of personalised medicine.

Introduction

Cancer is continuously revealed to be ever more complex than previously thought. Recent studies using a variety of new proteomics technologies have revealed a surprisingly large degree of variation between individual cancer cells from the same tumour [1,2]. Current clinical practice uses biomarkers limited to the simultaneous analysis of only a handful of proteins. These biomarkers fail to assess the true complexity of cancer, and the resulting biomarkers have a low prognostic value [3]. This may be because tumour cell heterogeneity strongly influences treatment-unresponsiveness and treatment-resistance [4]. Hence, there is a need for novel multiplex markers that simultaneously capture the expression of numerous proteins with relation to each other. Therefore, an imaging system capable of studying heterogeneity *in situ* is potentially a very useful clinical research tool.

New bioimaging techniques have recently been proposed to visualise the co-location or interaction of several proteins in cells in intact tissue specimen. These include MALDI imaging [5], Raman microscopy [6], Toponome Imaging System (TIS) [7], MxIF [8] and multi-spectral imaging methods [9]. 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 spatial location on a single tissue section [10]. It runs cycles of fluorescence tagging, imaging and soft bleaching *in situ* and generates multi-channel image data, where each individual channel provides information about the abundance level of a specific protein molecule. While colocation does not necessarily imply interaction, it is a potential indication for an interaction that is not necessarily direct. The images were acquired using a TIS microscope (ToposNomos Ltd., Munich, Germany), which has a sub-cellular maximum lateral resolution of 206 × 206 nm/pixel [11]. This allows the determination of sub-cellular protein network architectures.

Developing quantitative methods to analyse the large amounts of data generated can aid the development of new sensitive and specific multiplex biomarkers – also termed as complex biomarkers – for risk stratification and diagnosis, as well as anti-cancer drug discovery by identifying 'hub' proteins that are essential regulators of protein networks [12].

Methods

Recently, a departure has been made from the pixel-level analysis, commonly employed in the analysis of combinatorial molecular phenotypes (CMPs) [7] or molecular co-expression patterns (MCEPs) [13-15] in TIS imaging data. Examining the image data at a cell level allows us to study the heterogeneity of cell phenotypes within cancer samples. Briefly, the samples used for the recent cell-level analysis frameworks developed by our group at Warwick had been surgically removed from colon cancer patients. One sample was taken from the surface of the tumour mass, and another one was selected from apparently healthy colonic mucosa at least 10cm away from the

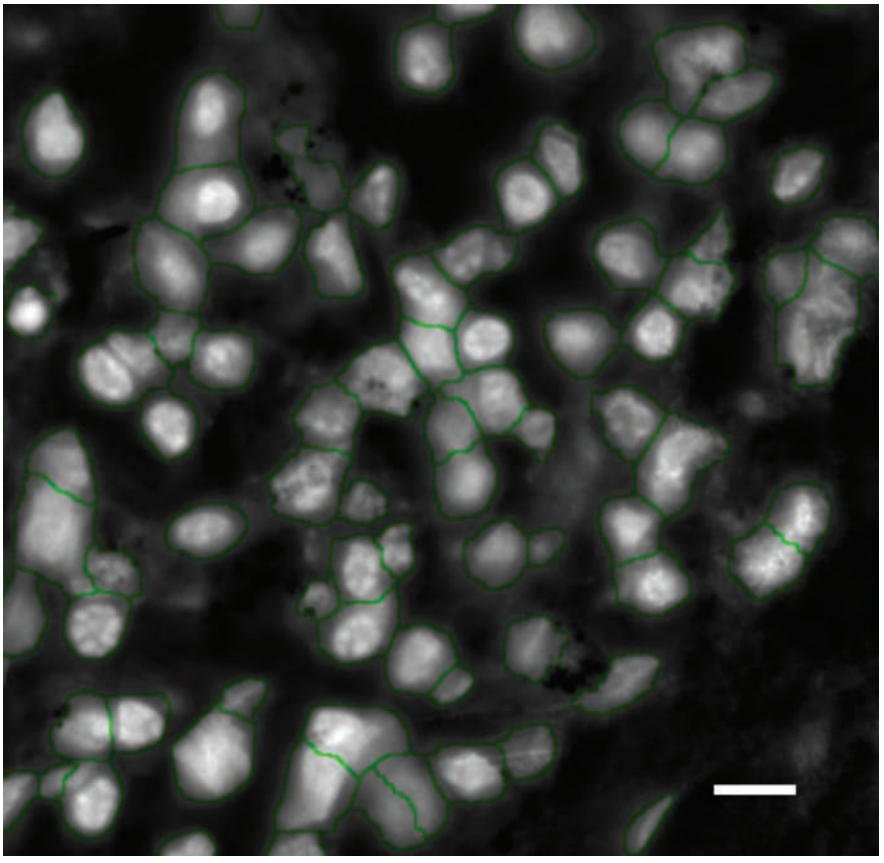

Figure 1: Segmentation of the cell nuclei on a section of a colon cancer sample as described in [19]. The outline of each identified nucleus is shown in green. The length of the scale bar is 10μm.

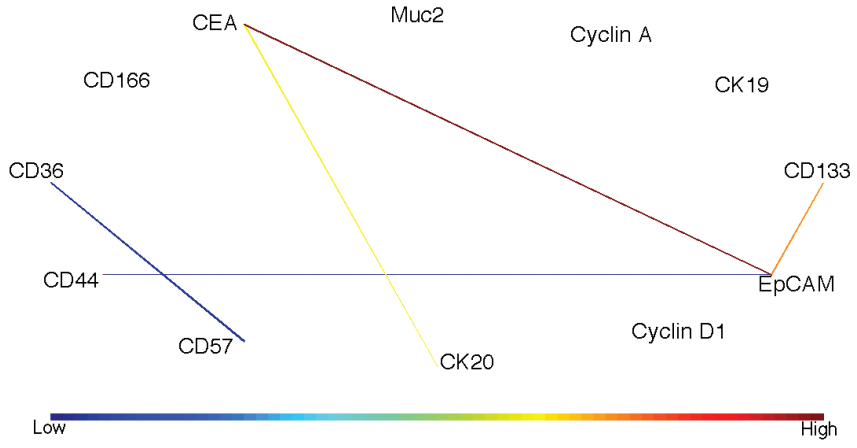

Figure 2: The social network of proteins. Each node represents a protein and each edge colour (as shown on the colour bar) shows a protein pair with different level of co-expression in the normal and cancer samples. Here, a large positive value (shown in red) indicates that the protein pair is more co-dependent and more frequently occurring 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 10% and the bottom 10% of the DiSWOP values are shown [19].

visible margin of the tumour. Two visual fields were manually selected in each tissue sample, resulting in four TIS data sets from a single patient. A library of 12 antibody tags, some of which are known tumour markers or cancer stem cell markers, was used. These were CD133, CK19, Cyclin A, Muc2, CEA, CD166, CD36, CD44, CD57, CK20, Cyclin D1 and EpCAM. The stacks also included a DAPI tag used to identify the cell nuclei. The image stacks obtained are then pre-processed using the following steps. Firstly, the images are aligned using the RAMTaB (Robust Alignment of Multi-Tag Bioimages) algorithm [16]. This is done in order to prevent potentially erroneous co-mapping resulting from the slight mis-alignment of the multi-tag images obtained using TIS. Background autofluorescence is digitally removed so any remaining signal should be true protein expression. The images are then segmented using a modified form of a graph cut based method [17] applied to the DAPI channel [18] (Figure 1). This is necessary in order to extract pixel locations of the nuclei and their immediate neighbourhood only, as the DAPI tag stains the DNA. This provides a rough approximation of the cells and removes possible noise from the stromal and lumen areas.

One of the cell-level analysis frameworks was presented by Khan et al. [20]. This work is an extension of a previous study [18] mining for cell phenotypes based on their high-dimensional protein co-expression profiles. Khan's study showed that the raw protein expression vectors have a non-linear high-dimensional structure, which can be effectively visualised using a symmetric neighbourhood embedding approach. The three-dimentional vectors obtained were found to outperform the original high-dimensional raw protein expression vectors in terms of their ability to discriminate between normal and cancer tissue samples on the basis of their phenotypic distributions [20].

Another related study proposed a way to analyse phenotypes obtained according to protein-protein dependence profiles of the cells [19]. Here, we proposed a new measure called Difference in Sum of Weighted cO-dependence Profiles (DiSWOP) that can highlight protein pairs that are more co-dependent and more frequently occurring in cancer than in normal samples, or vice versa. 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 in all the normal samples. The sums are normalised by the number of samples. It then subtracts the score for the normal from the score for the cancer samples, hence giving a positive score if a pair appears more frequently and with higher dependency scores in the cancerous samples [19]. Applying this measure to cell phenotypes obtained using Affinity Propagation clustering [21], several protein pairs were highlighted in terms of a small social network of proteins, as shown in Figure 2 [19]. One of these pairs is CEA and EpCAM, which came out as more co-dependent in the cancer samples. This protein pair has also been found experimentally to interact via the pathway CEA–SOX9–Claudin7–EpCAM, which is involved in determining the morphology of the colon epithelium [22-25]. The protein pairs highlighted by the measure are very similar when other phenotyping methods are used. This has been demonstrated by using Bayesian hierarchical clustering with a Gaussian prior [26] on half of the protein pairs, which discriminate best between cancer and normal tissue [25]. In addition, further validation of the measure has been done using synthetically generated image data [19].

Discussion and conclusions

The methods presented above have been applied only on a small number of samples – 3 [20] and 11 [19, 25]. In an on going

project, we are in the process of validating these methods on larger datasets. From the biological point of view, TIS provides an indication, sometimes a strong indication, of protein interactions and so direct interaction will need to be further tested by other experimental techniques such as immunoprecipitation. Other interactions may be less direct, and will need study via protein pathways. Despite this, the methods presented here allow us to consider protein expression and interactions localised within individual cells. This could aid the understanding of tumour heterogeneity and function of single cells within the cancerous tissue, and could be the key to discovering multiplex biomarkers that can help with diagnosis and prognosis of cancer patients. ■

References

1. Dalerba P, Kalisky T, Sahoo D, Rajendran PS, Rothenberg ME, Leyrat AA, Sim S, Okamoto J, Johnston DM, Qian D, Zabala M, Bueno J, Neff NF, Wang J, Shelton AA, Visser B, Hisamori S, Shimono Y, van de Wetering M, Clevers H, Clarke MF, Quake SR. *Single-cell dissection of transcriptional heterogeneity in human colon tumors.* Nat. Biotechnol. 2011;29:1120–7.
2. Shi Q, Qin L, Wei W, Geng F, Fan R, Shik Shin Y, Guo D, Hood L, Mischel PS, Heath JR. *Single-cell proteomic chip for profiling intracellular signalling pathways in single tumor cells.* Proc. Natl. Acad. Sci. U.S.A. 2012;109(2):419–24.
3. Vucic EA, Thu KL, Robison K, Rybaczyk LA, Chari R, Alvarez CE, Lam WL. *Translating cancer 'omics' to improved outcomes.* Genome Res. 2012;22:188-95.
4. Lee AJ, Swanton C. *Tumour heterogeneity and drug resistance: personalising cancer medicine through functional genomics.* Biochem. Pharmacol. 2011;83:1013–20.
5. Cornett D, Reyzer ML, Chaurand P, Caprioli RM. *MALDI imaging mass spectrometry: molecular snapshots of biochemical systems.* Nature Methods. 2007;4:828–33.
6. van Manen H, Kraan YM, Roos D, Otto C. *Single-cell Raman and fluorescence microscopy reveal the association of lipid bodies with phagosomes in leukocytes.* PNAS. 2005;102(29):10159–64.
7. Schubert W, Bonnekoh B, Pommer A, Philipsen L, Bockelmann R, Malykh Y, Gollnick H, Friedenberger M, Bode M, Dress A. *Analyzing proteome topology and function by automated multidimensional fluorescence microscopy.* N. Biotechnol. 2006;24:1270–8.
8. Gerdes MJ, Sevinsky CJ, Sood A, Adak S, Bello MO, Bordwell A, Can A, Corwin A, Dinn S, Filkins RJ, Hollman D, Kamath V, Kaanumalle S, Kenny K, Larsen M, Lazare M, Li Q, Lowes C, McCulloch CC, McDonough E, Montalto MC, Pang Z, Rittscher J, Santamaria-Pang A, Sarachan BD, Seel ML, Seppo A, Shaikh K, Sui Y, Zhang J, Ginty F. *Highly multiplexed single-cell analysis of formalin- fixed, paraffin-embedded cancer tissue.* PNAS. 2013;110(29):11982–7.
9. Barash E, Dinn S, Sevinsky C, Ginty F. *Multiplexed Analysis of Proteins in Tissue Using Multispectral Fluorescence Imaging.* IEEE Transactions on Medical Imaging. 2010;29(8):1457–62.
10. Schubert W, Gieseler A, Krusche A, Serocka P, Hillert R. *Next-generation biomarkers based on 100-parameter functional super-resolution microscopy TIS.* N. Biotechnol. 2012;29:599–610.
11. Bhattacharya S, Mathew G, Ruban E, Epstein DBA, Krusche A, Hillert R, Schubert W, 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.* J. Proteome Res. 2010;9(12):6112–25.
12. Evans RG, Naidu B, Rajpoot NM, Epstein DBA, Khan M. *Toponome imaging system: multiplex biomarkers in oncology.* Trends in Molecular Medicine. 2012;18(12):723–31.
13. Langenkamper D, Jan K, Humayun A, Abouna S, Epstein D, Khan M, Rajpoot NM, Nattkemper TW. *Towards protein network analysis using TIS imaging and exploratory data in Proceedings.* Workshop on Computational Systems Biology (WCSB), Zurich. 2011.
14. Humayun A, Raza SeA, Waddington C, Abouna S, Khan M, Rajpoot NM. *A Novel Framework for Molecular Co-Expression Pattern Analysis in Multi- Channel Toponome Fluorescence Images.* Microscopy Image Analysis with Applications in Biology. 2011;109–12.
15. Kolling J, Langenkamper D, Abouna S, Khan M, Nattkemper T. *WHIDE-A web tool for visual data mining colocation patterns in multivariate bioimages.* Bioinformatics. 2012;28(8):1143–50.
16. Raza SeA, Humayun A, Abouna S, Nattkemper TW, Epstein DBA, Khan M, Rajpoot NM. *RAMTaB: Robust Alignment of Multi-Tag Bioimages.* PLoS ONE. 2012; 7: e30894.
17. Al-Kofahi Y, Lassoued W, Lee W, Roysam B. *Improved automatic detection and segmentation of cell nuclei in histopathology images.* IEEE Trans Biomed Eng. 2010;57(4):841–52.
18. Khan AM, Humayun A, Raza SeA, Khan M, Rajpoot NM. *A Novel Paradigm for Mining Cell Phenotypes in Multi-Tag Bioimages using a Locality Preserving Nonlinear Embedding.* Lecture Notes in Computer Science, Neural Information Processing, Springer Berlin Heidelberg. 2012;7666:575– 83.
19. Kovacheva VN, Khan AM, Khan M, Epstein D, Rajpoot NM. *DiSWOP: A Novel Measure for Cell-Level Protein Network Analysis in Localised Proteomics Image Data.* Bioinformatics. 2013; 10.1093/bioinformatics/btt676.
20. Khan AM, Raza SeA, Khan M, Rajpoot NM. *Cell Phenotyping in Multi-Tag Fluorescent Bioimages.* Accepted for publication in Neurocomputing (Elsevier). 2013.
21. Frey BJ and Dueck D. *Clustering by Passing Messages Between Data Points.* Science. 2007;315:972–7.
22. Darido C, Buchert M, Pannequin J, Bastide P, Zalzali H, Mantamadiotis T, Bourgaux JF, Garambois V, Jay P, Blache P, Joubert D, Hollande F. *Defective claudin-7 regulation by Tcf-4 and Sox-9 disrupts the polarity and increases the tumorigenicity of colorectal cancer cells.* Cancer Res. 2008;68(11):4258–68.
23. Kuhn S, Koch M, Nubel T, Ladwein M, Antolovic D, Klingbeil P, Hilde-brand D, Moldenhauer G, Langbein L, Franke WW, Weitz J, Zoller M. *A complex of EpCAM, claudin-7, CD44 variant isoforms, and tetraspanins pro- motes colorectal cancer progression.* Mol. Cancer Res. 2007;5(6):553–67.
24. Zalzali H, Naudin C, Bastide P, Quittau-Prevostel C, Yaghi C, Poulat F, Jay P, Blache P. *CEACAM1, a SOX9 direct transcriptional target identified in the colon epithelium.* Oncogene. 2008;27(56):7131–8.
25. Kovacheva VN, Sirinukunwattana K, Rajpoot NM. *A Bayesian Framework for Cell-Level Protein Network Analysis for Multivariate Proteomics Image Data.* Accepted for publication in Proceedings of SPIE Digital Pathology, San Diego. 2014.
26. Sirinukunwattana K, Savage RS, Bari MF, Snead DRJ, Rajpoot NM. *Bayesian Hierarchical Clustering for Studying Cancer Gene Expression Data with Unknown Statistics.* PLoS ONE. 2013;8(10):e75748.