Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study

Patterns, 2021, 100399

Highlights of this publication

Accurate prediction of a sample's tumor purity from histopathology slides can be determined using a novel Multiple Instance Learning (MIL) model. The tumor purity varies spatially within a sample and can be spatially resolved from sample-level labels using MIL model. This study also revealed that the pathologist's region selection of the histopathology slides plays a role in determining the correct percentage tumor nuclei estimation (Oner et al., 2021).

Introduction

High-throughput genomic analysis has become a useful tool for cancer research (Schuster, 2008; Xuan et al., 2013). The proportion of cancer cells in the samples (out of other cell types such as normal epithelial cells, stromal cells and infiltrating immune cells) is one of the major factors affecting the quality of genomic analysis. The percentage of cancer cells within a tissue section, also known as tumor purities, are routinely evaluated by pathologists through hematoxylin and eosin (H&E)-stained histopathology slides. However, this procedure is typically tedious and time-consuming and exists inter-observer variability between pathologists' estimates (Mikubo et al., 2020; Smits et al., 2014). Tumor purity can also be inferred from different types of genomic data, such as somatic copy number (Carter et al., 2012) and mutations (Yuan et al., 2020), gene expression data (Yoshihara et al., 2013) and DNA methylation data (Zhang et al., 2015).

Research in Context

In this publication, the authors have developed a novel MIL model that could assist the pathologists in streamlining the tumor purity quantitation process. The MIL model consists of three modules: feature extractor module, MIL pooling filter (a novel distribution pooling filter that is more superior than standard pooling filters (like mean and maximum pooling), and bag-level representation transformation module. Given a bag of patches, the feature extractor module extracts a feature vector for each patch inside the bag. Then the distribution pooling filter obtains a strong bag-level representation by estimating the marginal distributions of the extracted features. Next, the bag level representation transformation module predicts tumor purity. Finally, the authors use neural networks to implement the feature extractor module and the bag-level representation transformation module to parameterize the learning process fully, training end-to-end using samples' genomic tumor purity values as labels (Oner et al., 2021).

To verify their MIL model, the authors first conducted a sample level tumor purity prediction using cancer datasets from different cohorts of The Cancer Genome Atlas (TCGA) and a local Singapore cohort. By using correlation analyses between genomic tumor purity values obtained from ABSOLUTE (Carter et al., 2012) and the MIL models' predictions, which are assessed by the performance metric, Spearman's rank correlation coefficient, the authors obtained significant

correlations in eight cohorts (breast invasive carcinoma (BRCA), glioblastoma multiforme (GBM), brain lower grade glioma (LGG), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), prostate adenocarcinoma (PRAD), and uterine corpus endometrial carcinoma (UCEC). This demonstrates that their novel MIL model correlates significantly with genomic tumor purity values.

The authors then repeated the correlation analyses between genomic tumor purity values and pathologists' percentage tumor nuclei estimates using Spearman's rank correlation coefficient and mean absolute error. They found that their MIL model's tumor purity predictions are more consistent with genomic tumor purity values than the pathologists' percentage tumor nuclei estimates, supported by statistical tests such as Fisher's z transformation-based method (Meng et al., 1992). Moreover, their MIL model has higher Spearman's rank correlation coefficient and lower mean absolute error than pathologists' percentage tumor nuclei, confirming this MIL model as the superior method.

Consequently, their MIL model could successfully predict tumor purity from H&E-stained digital histopathology slides of fresh-frozen sections in different TCGA cohorts, as well as slides of formalin-fixed paraffin-embedded (FFPE) sections in a local Singapore cohort.

The authors also discovered that tumor purity varies spatially within a sample: Top and bottom slides of a sample are statistically different in tumor purity using the Wilcoxon signed-rank test. They also found that the degree of spatial variation in tumor purity is different for different cancer types, UCEC, LGG, and GBM cohorts being the most spatially homogenous cancers among all cohorts, PRAD cohort being the most spatially heterogeneous cancer in tumor purity, based on the mean absolute differences between top and bottom slides' predictions. Moreover, predicting a sample's tumor purity using both top and bottom slides is more accurate than using only one slide.

Spatial tumor purity map analysis using the MIL model also revealed that the pathologist's region selection of the histopathology slides plays a role in determining the correct percentage tumor nuclei estimation.

In addition, the authors could qualitatively validate that their MIL model learned discriminant features for cancerous versus normal tissue histology and successfully classify them into tumor versus normal tissues from sample-level genomic tumor purity labels without requiring pixel-level annotations from pathologists.

All in all, the authors have demonstrated that their MIL model can be used for high-throughput genomic analysis, which will help reduce pathologists' workload and decrease inter-observer variability. They have also showed that their model is cost-effective compared with genomics methods. Moreover, because of the tumor purity maps showing spatial variation within sections, they can now better understand the tumor microenvironment in each slide. Lastly, the MIL model made it possible to stratify patients based on their models' predictions.

REFERENCES

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., Beroukhim, R., Pellman, D., Levine, D. A., Lander, E. S., Meyerson, M., & Getz, G. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, *30*, 413–421.

Meng, X.-L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*, 172–175.

Mikubo, M., Seto, K., Kitamura, A., Nakaguro, M., Hattori, Y., Maeda, N., Miyazaki, T., Watanabe, K., Murakami, H., Tsukamoto, T., Yamada, T., Fujita, S., Masago, K., Ramkissoon, S., Ross, J. S., Elvin, J., & Yatabe, Y. (2020). Calculating the Tumor Nuclei Content for Comprehensive Cancer Panel Testing. *Journal of Thoracic Oncology : Official Publication of the International Association for the Study of Lung Cancer*, *15*, 130–137.

Oner, M. U., Chen, J., Revkov, E., James, A., Heng, S. Y., Kaya, A. N., Alvarez, J. J. S., Takano, A., Cheng, X. M., Lim, T. K. H., Tan, D. S. W., Zhai, W., Skanderup, A. J., Sung, W.-K., & Lee, H. K. (2021). Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study. *Patterns*, 100399.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, *5*, 16–18.

Smits, A. J. J., Kummer, J. A., de Bruin, P. C., Bol, M., van den Tweel, J. G., Seldenrijk, K. A., Willems, S. M., Offerhaus, G. J. A., de Weger, R. A., van Diest, P. J., & Vink, A. (2014). The estimation of tumor cell percentage for molecular testing by pathologists is not accurate. *Modern Pathology : An Official Journal of the United States and Canadian Academy of Pathology, Inc*, *27*(2), 168–174.

Xuan, J., Yu, Y., Qing, T., Guo, L., & Shi, L. (2013). Next-generation sequencing in the clinic: promises and challenges. *Cancer Letters*, *340*, 284–295.

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P. W., Levine, D. A., Carter, S. L., Getz, G., Stemke-Hale, K., Mills, G. B., & Verhaak, R. G. W. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature Communications*, *4*, 2612.

Yuan, X., Li, Z., Zhao, H., Bai, J., & Zhang, J. (2020). Accurate Inference of Tumor Purity and Absolute Copy Numbers From High-Throughput Sequencing Data. *Frontiers in Genetics*, *11*, 1–8.

Zhang, N., Wu, H.-J., Zhang, W., Wang, J., Wu, H., & Zheng, X. (2015). Predicting tumor purity from methylation microarray data. *Bioinformatics (Oxford, England)*, *31*, 3401–3405.