Learning associations between gene defects and structural variant signatures in colorectal cancer

Renske de Wit, Carlos Manuel García Fernández, Soufyan Lakbir, Remond J.A. Fijneman, Sanne Abeln

Introduction

Cancer is a complex and multifaceted disease. Machine learning approaches are often employed to uncover intricate relationships within datasets from hundreds or thousands of cancer patients. A key focus for these projects is often not only model performance, but also explainability. Understanding model decisions is crucial to improve our understanding of the biological processes which enable tumor development. An additional challenge is that suitable datasets are often small, limiting the potential of deep-learning approaches. In this work, we aimed to learn the relations between gene defects and structural changes in the genome, based on a set of observed DNA mutations. We explored both simple machine learning models and deep learning-based strategies, comparing performance as well as explainability.

Biological background

Colorectal cancer (CRC) is the third most commonly diagnosed cancer world-wide and second leading cause of cancer-related deaths (Hossain et al., 2022). There is an urgent need to better understand the underlying biological mechanisms to improve patient treatment. One of the key features in (colorectal) cancer is defects in DNA repair, leading to an excessive accumulation of mutations in the tumor genome (Hanahan and Weinberg, 2000). A well-known gene that plays a crucial role in many cancers is *TP53*. Recently, two other genes have emerged that are potential key players in CRC development and are likely involved in DNA repair: *MACROD2* and *PRKN*. Defects in these genes are often observed in CRC and associated with an increase in DNA mutations, more specifically *structural variants* (SVs). We hypothesized that the role of *MACROD2*, *PRKN*, and *TP53* in CRC could be elucidated by identifying and analyzing characteristic patterns of SVs - *mutational signatures* - that are associated with defects in these three genes. To enable capture of non-linear relationships in the data, we chose to adopt a machine learning approach instead of existing methods for mutational signature analysis such as non-negative matrix factorization.

Methods

Data: We used a dataset of SVs from 745 CRC metastatic tumors, obtained from the Hartwig Medical Foundation (Priestley et al., 2019). Patients were labeled according to the gene defects observed in the tumor (in *MACROD2, PRKN,* and *TP53).* We explored 2 labeling strategies: in the *Soft labeling* approach, each gene was considered separately and tumors were labeled either *wild-type* or *mutant*. In contrast, in the *Strict labeling* approach, tumors were labeled based on the *combination* of observed gene defects. In addition, we computed both *global features,* reflecting characteristics of the tumor itself, as well as *event-based features*, characterizing properties of individual SVs.

Models: We performed binary classification for each pair of unique labels (e.g. *MACROD2* ∩ *PRKN* vs *MACROD2-only*), resulting in 28 tasks using the *Strict labeling* strategy and 3 tasks using

the *Soft labeling* approach. We explored two models for these classification tasks: a *SV set + transformer-based* approach and a *Dirichlet Process Gaussian Mixture Model (DPGMM) + logistic regression* strategy. The input for the *transformer-based* model consisted of the global features and individual structural variants observed in a tumor encoded as a set. In contrast, the *logistic regression* model was trained only on the global features, in addition to a set of features derived from the *event-based* features using a DPGMM approach, in which the information from individual SVs was aggregated across the tumor. Due to limited dataset size, the *Strict labeling* strategy was only employed for the *DPGMM + logistic regression* approach, whereas *Soft labeling* was also explored for the *SV set + transformer-based* model.

Results

Based on the results of the *Soft* and *Strict labeling* strategies, we found a strong co-occurrence between defects in the three genes: tumors with deficient *MACROD2* often also showed defects in *PRKN* and *TP53* (277/745; 37% of tumors). In contrast, our dataset contained only 38 patients with defects in *MACROD2* only. This highlights the importance of models that can be trained with small datasets.

Using the *Strict labeling* strategy, the *DPGMM + logistic regression* approach showed good performance for several classification tasks, in particular *MACROD2-only* vs *wildtype* and *TP53-only* vs *wildtype* (0.77 and 0.87 AUC; balanced training set of 60 and 92 samples, respectively). In addition, the *MACROD2* ∩ *TP53* ∩ *PRKN* tumors could be distinguished from *MACROD2-only* with good performance (0.83 AUC; balanced training set of 60 samples).

The performance of the *DPGMM + logistic regression* and *transformer-based* models was compared using the *Soft labeling* approach. Although *DPGMM + logistic regression* showed superior performance in 2/3 classification tasks, the *transformer-based* approach improved when more data was available, suggesting this as the main limiting factor.

Finally, analysis of the *logistic regression* model coefficients at different regularization strengths showed that *MACROD2-only*, *PRKN-only,* and *TP53-only* tumors were mainly distinguished from *wild-type* by their *high number* of SVs. In addition, the presence of deletions with size 1-10,000kb were important to distinguish *MACROD2-only* and *PRKN-only* tumors from *wild-type.*

Conclusion

From this work, we conclude that defects in genes responsible for DNA repair are associated with a characteristic SV signature that can be detected via machine learning approaches. Although the *DPGMM + logistic regression* approach still outperformed *transformer-based* strategies, this was likely due to the small dataset and shows potential if more training data was available.

Finally, the desire for explainability to understand underlying biological mechanisms creates a strong incentive to favor simple, transparent models over more complex architectures, such as those based on deep learning. In future work, we aim to expand our dataset by incorporating unlabeled data from additional sources, as well as validating our current results with gene expression data.

References

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. Cell, 100(1), 57–70. https://doi.org/10.1016/s0092-8674(00)81683-9.

Hossain, M. S., Karuniawati, H., Jairoun, A. A., Urbi, Z., Ooi, J., John, A., Lim, Y. C., Kibria, K. M. K., Mohiuddin, A. K. M., Ming, L. C., Goh, K. W., & Hadi, M. A. (2022). Colorectal Cancer: A Review of Carcinogenesis, Global Epidemiology, Current Challenges, Risk Factors, Preventive and Treatment Strategies. Cancers, 14(7), 1732. https://doi.org/10.3390/cancers14071732.

Priestley, P., Baber, J., Lolkema, M. P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., Roepman, P., Voda, M., Bloemendal, H. J., Tjan-Heijnen, V. C. G., van Herpen, C. M. L., Labots, M., Witteveen, P. O., Smit, E. F., Sleijfer, S., … Cuppen, E. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. Nature, 575(7781), 210–216. https://doi.org/10.1038/s41586-019-1689-y.