EBioMedicine 57 (2020) 102851

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.elsevier.com/locate/ebiom

Commentary

Somatic copy number aberrations detected in circulating tumor DNA can hold diagnostic value for early detection of hepatocellular carcinoma

Ekaterini Chatzaki[1,2,*], Ioannis Tsamardinos[3,4]

[1] Laboratory of Pharmacology, Medical School, Democritus University of Thrace, Alexandroupolis
[2] Institute of Agri-food and Life Sciences, University Research Centre, Hellenic Mediterranean University, Heraklion 71410, Crete, Greece
[3] Gnosis Data Analysis PC, Heraklion, Greece
[4] Department of Computer Science, University of Crete, Heraklion, Greece

ARTICLE INFO

Article History:
Received 24 May 2020
Revised 5 June 2020
Accepted 6 June 2020
Available online xxx

The study reported by Tao et al., 2020 appearing in the recent issue of EBioMedicine [1] targets the significant unmet clinical need for a blood-based minimally invasive accurate biomarker for the early detection of hepatocellular carcinoma (HCC). Circulating tumor DNA (ctDNA), a most promising liquid biopsy biomaterial holding clinically relevant information is analyzed. ctDNA had been shown to carry genetic and epigenetic aberrations that mirror the growing tumor [2,3], therefore could be used to monitor its lifespan. The present study focuses on the detection of somatic copy number aberrations (SCNAs) in ctDNA, not previously assessed in HCC. As a liquid biopsy parameter, SCNAs present some advantages over others, as they contribute larger number of ctDNA fragments to the overall cfDNA pool, span larger genomic regions, whereas compared to methylation, they remain much less affected by confounders such as age, diet, and life style.

Original data are produced from low-depth whole-genome sequencing (WGS) readings in a relatively large study group of 384 hepatitis B virus (HBV) infected patients, some with cancer. One training and 2 smaller validation cohorts were used, the last from a different clinical setting than the other two, adding credibility to the validation. In contrast to the training group, the 2 validation groups contained only early-stage cancer patients and a small percentage of unknown. For data analysis, authors employ machine learning, the current trend for high-dimensional dataset extrapolation to produce specific classifiers [4]. A 3-stage design was adopted, aiming to develop and evaluate a model for early stage HCC.

In the first step, tumor burden was estimated as the fraction of ctDNA representation in total ccfDNA, addressed by quantifying tumor fraction (TFx) in ccfDNA via low-coverage WGS using ichorCNA algorithm [5]. To explore the utility of the estimated tumor burden, TFx statistics using as cutoff TFx>0 showed that although ctDNA burden had remarkable specificity in classifying HCC regardless of tumor stage, it suffered from markedly reduced sensitivity in early-stage patients. These results agree with multiple other studies showing the ctDNA levels increase with stage, and therefore abundance is a restricting factor in its applicability for clinically relevant end-points in early stages.

In the second step, in order to consider the sequencing depth SCNA profile information across the whole genome in the classifier development, an RF-based machine learning algorithm was employed to distinguish HCC from non-cancer HBV patients, leading to significantly increased performance (AUC = 0.893) over TFx. The model performed better in the late rather than the early-stage patient group in terms of accuracy.

In a last attempt and in order to improve accuracy, as the early-stage detection is the most clinically challenging, archived external data (publicly available in The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC)) were also leveraged, succeeding to incorporate prior knowledge to further augment detection performance. Machine learning was applied to develop a novel weighted Random Forest model (wRFdriver) incorporating driver evidence derived from external data, ie GISTIC2 and iDriver scores. Although performance was not dramatically improved, especially in the validation cohort 2 deriving from a distinct clinical setting, this model seemed to overcome the limitation of low biomaterial abundance. The issue of cancer heterogeneity within the study group was also addressed.

The authors choose to adopt their own machine learning approach, bearing some risk for information leakage and overfitting. This reflects in the reduced predictive strength seen in the validation steps in relation to the training and is probably the result of incorporating the important features directly obtained from the SCNA profiling in ccfDNA of the discovery cohort in the final models, rather that considering all data equally in a naive way while constructing the

[*] Corresponding author.
E-mail address: achatzak@med.duth.gr (E. Chatzaki).

https://doi.org/10.1016/j.ebiom.2020.102851

model. Finally, the cut-off of predictive probability of 0.5 is not calibrated, resulting in reduced and not reproducible sensitivity. Uncalibrated reported probabilities is a general issue with most machine learning models. For clinical use, one should calibrate and optimize this threshold for the optimal trade-off between sensitivity and specificity that can be achieved by the model.

Overall, we found this work of interest, both as a methodological approach and as produced results, with three major strengths: 1. the focus of the limited studied SCNAs in liquid biopsies and in particular detected in ctDNA, 2. The integration of primary data and archived data for machine learning towards the development of an accurate model. 3. Validation in independent groups, adding credibility to the performance and clinical potential of the classifiers. However, although significant from a scientific point of view, the translational value of these findings is questionable. The building of a blood-based assay over the reported models would require WGS for SCNAs in adequate ctDNA and x5 readings, as lowering reading coverage reduced performance in early detection. This reduces clinical applicability to time- and cost-effective solutions. Furthermore, it is possible that the same original laboratory data could lead to better verified models upon automated machine learning. Processing through feature selection or even dimensionality reduction methods would be likely to lead to reduced-size simplified classifiers with the potential to be translated to fast and cost-effective assays, readily available to clinical practice. Automated machine learning tools are now available to biomedical scientists for exploring such −omics datasets to produce best performing and best interpretable models respective to classifying, biologically relevant signatures and their performing metrics. Some of them may also provide options for automated validation. These tools facilitate greatly machine learning approaches for non-experts and allow increased extrapolation of primary data [6].

Funding sources

The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the “First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant” (Project Number: 1955) .

Declaration of Competing Interests

We declare no financial or personal relationships with people or organizations that could inappropriately influence (bias) this work.

References

[1] Tao K, Bian Z, Zhang Q, Guo X, Yin C, Wang Y, et al. Machine learning-based genome-wide interrogation of somatic copy number aberrations in circulating tumor DNA for early detection of hepatocellular carcinoma. EBioMedicine 2020. https://doi.org/10.1016/j.ebiom.2020.102811.

[2] Panagopoulou M, Karaglani M, Balgkouranidou I, Pantazi C, Kolios G, Kakolyris S, et al. Circulating cell-free DNA release in vitro: kinetics, size profiling, and cancer-related gene methylation. J Cell Physiol 2019;234(8):14079–89.

[3] Fece de la Cruz F, Corcoran RB. Methylation in cell-free DNA for early cancer detection. Ann Oncol 2018;29(6):1351–3.

[4] Panagopoulou M, Karaglani M, Balgkouranidou I, Biziota E, Koukaki T, Karamitrousis E, et al. Circulating cell-free DNA in breast cancer: size profiling, levels, and methylation patterns lead to prognostic and predictive classifiers. Oncogene 2019;38(18):3387–401.

[5] Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun 2017;8(1):1324.

[6] Tsamardinos I, Charonyktakis P, Lakiotaki K, Borboudakis G, Zenklusen JC, Juhl H, et al. Just add data: automated predictive modeling and biosignature discovery. BioRχiv 2010. https://doi.org/10.1101/2020.05.04.075747.