A data-driven approach employing automated Machine Learning reveals novel liquid biopsy methylation biomarkers for breast cancer.

Maria Panagopoulou[1]
Makrina Karaglani[1], Eirini Biziota[1], Kyriakos Amarantidis[1], Triantafillia Koukaki[1], Konstantina Tzitzikou[1], Nikoletta Kessari[1], Theodosis Theodosiou[1], Stylianos Kakolyris[1] and Ekaterini Alexiou Chatzaki[1]
[1] Democritus University of Thrace

Background & objectives

DNA methylation plays an important role in Breast Cancer (BrCa) pathogenesis and could contribute to its personalized management via specific methylation biomarkers. We pursued a data-driven AI-aided pipeline for biomarker discovery to reveal novel clinically relevant methylation biomarkers for BrCa to be implemented in liquid biopsies.

Methods

Innovative automated machine learning (AutoML platform JADBio) was employed to analyze publicly available methylomes (28591 genes) from 520 BrCa and 185 healthy breast tissues. Features selected in the biosignatures built were further studied via text mining bioinformatic tools (UniRED) against proteins with an established role in BrCa pathology. Laboratory validation followed, via Methylation Specific qPCR in circulating cell-free DNA (ccfDNA) from 119 BrCa patients and 77 age-matched healthy women. Sequential ccfDNA samples from the same patient before and after adjuvant therapy (6 cases) and 1st line therapy (3 cases) were also examined.

Results

In the *in silico* part of the study, a biosignature was developed demonstrating high performance in distinguishing BrCa from healthy breast tissues (AUC 0.994 (0.985-0.999), precision 0.989 (0.972-0.999)) via Ridge Logistic Regression algorithm. The biosignature contained 3 features novel for BrCa, i.e. the methylation of *CLDN15* (the coding gene of Claudin-15, a protein related to Blood-Brain Barrier, immune cell transmigration and Tight junction), *MRGPRD* (the coding gene of MAS Related GPR Family Member D protein related to G protein-coupled receptor activity) and *ZNF430* (the coding gene of Zinc Finger Protein 430 involved in transcriptional regulation and

nucleic acid binding*)*. Literature mining revealed nearly no previous report of these genes in relation to BrCa. When investigated in patient ccfDNA, methylation levels of *CLDN15* and *ZNF430* were found significantly increased in BrCa patients as compared to controls whereas *MRGPRD* was more often methylated after therapy. Most importantly, further autoML multiparametric analysis of our ccfDNA methylation measurements built an optimised biosignature via Classification Random Forest, which could efficiently discriminate BrCa from health (AUC 0.870(0.826-0.913), precision 0.886).

Conclusion

Our data-driven AI-aided approach unraveled 3 novel gene methylation biomarkers, building a highly specific BrCa biosignature to be implemented in liquid biopsy with potential clinical value. The role of these genes in BrCa pathophysiology awaits further investigation.