Boosting perspectives for breast cancer intrinsic subtyping on RNA-sequencing data

Silvia Cascianelli,[1] Ivan Molineris,[2] Claudio Isella,[2] Marco Masseroli,[1] and Enzo Medico[2,3]

[1] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy
[2] Candiolo Cancer Institute, FPO-IRCCS, S.P. 142, km 3.95, 10060 Candiolo, TO Italy
[3] Department of Oncology, University of Torino, S.P. 142, km 3.95, 10060 Candiolo, TO Italy

Identification of breast cancer (BC) intrinsic subtypes by the PAM50 test has set a cornerstone in cancer genomics, linking unsupervised class discovery to biological insights and clinically relevant stratification. Yet, PAM50 classifier needs a reference sample arbitrarily built from the dataset under study; this profoundly affects subtyping, limiting robustness and reproducibility, as we clearly proved. Therefore, we propose our 'AWCA' procedure for reference construction, robust to the initial sample selection and improving PAM50 reproducibility: the concordance achieved by AWCA-classifications among them is beyond 98%, remarkably enhancing also the stability of samples comparably correlated to more than one subtype. Additionally, on wide controlled BC RNA-seq datasets we constructed AWCA-references that when used on independent RNA-seq data, subjected to the same normalization, bring over 94% of concordance with inner-AWCA-classifications. Thus, using predefined AWCA-references could make PAM50 a fully reproducible single-sample test. Furthermore, risk of recurrence scores computed from AWCA-PAM50 results show a more statistically significant p-value in discriminating good and poor prognoses in 10-year overall survival analysis.

But intrinsic subtypes are linked also to the expression of other genes than the PAM50 panel and using extended classifiers could strengthen subtyping reliability by exploiting discriminative information from genome-wide RNA-sequencing. Hence, we investigate several classifiers trained supervisely to recognize BC subtypes. Logistic Regression (R) appears the most effective and robust in this task, particularly using a proper feature selection. Besides PAM50, we trace two promising signatures of differentially expressed genes and our best LRs reach valuable and robust performances in cross-validation and on both internal and external testing with beyond 90% of accuracy, recalls and concordance with AWCA-PAM50-classifications. Furthermore, all LRs show an improved prognostic ability with respect to PAM50 calls.

Besides the results of our study, recently published in the scientific journal Scientific Reports, we make publicly available all our single-sample classifiers and the R-code to build AWCA-references on any expression data, as we successfully prove also for Affymetrix microarray data. On the other hand, the subtyping robustness and the key prognostic ability of the RNA sequencing-based proposed approaches encourage other studies on heterogeneous datasets, but mostly a prominent use of RNA-seq in BC clinical practice.