BITS :: Call for Abstracts 2021 - Oral communication

Type	Oral communication
Session	Gene regulation, transcriptomics and epigenomics
Title	Investigating transcript isoform RNA-seqdata and machine learning techniques for breast cancer subtyping
All Authors	Cascianelli S(1),Sanatdoost SN(1), Marco Masseroli(1)
Affiliation	

(1) Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano

Motivation

Identifying breast cancer (BRCA) intrinsic subtypes, five classes of confirmed prognostic value, is crucial in the clinical handling of BRCA patients. Beyond reflecting tumor molecular traits, intrinsic subtypes discriminate good-expected prognoses from poor ones. Yet, patient cohorts are still commonly stratified through the PAM50 test, a dataset-oriented approach that focuses on the expression of 50 genes and cannot assign each patient separately with a subtype. Furthermore, despite the improvements of BRCA subtyping currently aim at taking advantage of the accurate quantifications provided by the RNA-sequencing technology, an in-depth evaluation of the contribution of gene isoforms to intrinsic subtyping has not been addressed yet, particularly not through broad, well-annotated data of transcript isoforms and robust machine learning techniques. Different gene isoforms can alter and discriminate functions and products, and isoform diversity is known to be tissue- and disease-specific; such specificity has also been related to BRCA heterogeneity, in particular to hormonal-status-based classes. Yet, prognostically relevant classification of BRCA patients based on isoform expression level has barely been touched to date. We believe that isoform differentiation, hidden in a gene-level investigation, could be a precious resource to better characterize and distinguish BRCA subtypes.

Methods

We thoroughly investigated the isoform role in distinguishing the BRCA intrinsic subtypes on RNA-seq expression data of transcript isoforms, made available by The Cancer Genome Atlas project. We developed several computational approaches using isoform expression and different classifiers to predict the subtype of every sample separately. All computational models were trained supervisely using published subtypes as targets, 10-fold cross-validation to optimize accuracy or balanced accuracy up to find the most promising methods, and grid search to get the best hyperparameter tuning. We evaluated Logistic Regression and Support Vector Machine (SVM) algorithms to preserve the biological interpretability while carefully selecting the isoforms of interest as features. Notably, we explored both the isoforms generated from the 50 genes of the PAM50 test and those underlying another task-related signature, called LIMMA50, which we recently found valuable for gene-level BRCA subtyping. Additionally, we applied preprocessing and further data-driven feature selection steps to reduce the noise of the lowly expressed isoforms and the redundancy of the highly correlated features, and to highlight the subtype-specific differences in isoform occurrence. Filtering and embedded regularizations were used to face the curse of dimensionality and the overfitting risk due to the highly unbalanced dimensionalities of the isoform-level expression data, characterized by 1) relatively small number of available samples (few hundreds); 2) uneven sample distribution among subtypes; and 3) huge cardinality of quantified isoforms, with each sample including potentially over 60,000 isoforms. Isoforms showed also extremely varied expression values within each sample and can be much more correlated than genes; particularly, those originated from the same locus revealed strong co-expression or significant differences, a precious aspect in the case of subtype-specificity. This steers the use of similarity analyses, both globally and within-subtypes, and comparative assessments to further improve the investigation of isoform role.

Results

Our computational results were obtained by comparing the performances of all computed models both in cross-validation and testing via accuracy, balanced accuracy and macro-average metrics. In addition, precision and recall of each class allowed to evaluate the capabilities of each model comprehensively. Subtyping calls were also compared with disease-free survival annotations in a 10-year horizon to assess the goodness of the approaches in distinguishing disease-free patients from recurred/progressed ones. We found that the best Logistic Regression reaches an overall accuracy beyond 80%, while the best SVM provides slightly higher precision in discriminating bad-prognosis cases. Both show balanced performances across classes, but mainly work effectively at the isoform expression level, finding subtype-related isoforms useful to provide BRCA patients with accurate and prognostically reliable single-sample classifications. Similarly to new gene expression-based BRCA classifiers, our computational models

overcome the dataset-based approach of the PAM50 test. Most importantly, they offer an innovative perspective on the involvement of isoforms in characterizing and differentiating the intrinsic subtypes. In particular, feature importance analysis sheds light on the contribution of different isoforms in BRCA stratification, but also points out new subtype-specific peculiarities at the isoform level that should be investigated with further studies.

Info
-
Figure
-

Availability	-
Corresponding Author	
Name, Surname	Silvia, Cascianelli
Email	silvia.cascianelli@polimi.it
Submitted on	08.05.2021