BITS :: Call for Abstracts 2022 - Oral communication

Type	Oral communication
Session	Machine Learning in Bioinformatics
Title	Machine learning to discover genes predictive of RAS-mutated cases in mutational profiles of colorectal cancer patients
All Authors	Bellomo M (1), Cascianelli S (1), Medico E (2,3), Masseroli M(1)

Affiliation

(1) Department of Electronics, Information and Bioengineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milan, Italy
(2) Department of Oncology, University of Turin, Candiolo (TO), S.P. 142, km 3.95, 10060, Italy
(3) Candiolo Cancer Institute, FPO-IRCCS, Candiolo (TO), S.P. 142, km 3.95, 10060, Italy

Motivation

Colorectal cancer (CRC) is profoundly heterogeneous in expected prognosis and drug sensitivity, and predictive genes and models have been so far largely based on gene expression profiling. Although mutations like those of the RAS family are known to play a critical role in CRC development and clinical outcomes, the contribution of many other co-occurrent mutations is still mostly unclear. Next-Generation Sequencing (NGS) technologies are offering an increasing availability of mutational data, which can be explored using Machine Learning techniques to investigate CRC mutational complexity and acquire new knowledge about disease characteristics. Specifically, identifying somatic mutations predictive of prognosis and drug-response is one of the key goals of the GERSOM project, implemented by the Alliance Against Cancer Network and promoted by the Italian Ministry of Health. In the context of this multicentric project and in collaboration with the Candiolo Cancer Center, we focused on investigating CRC tumours with mutations in the RAS gene family (KRAS, HRAS, NRAS), which are particularly critical for clinical treatment. RAS mutations are associated with primary and acquired resistance to anti-EGFR blockade; consequently, affected patients have poor response to canonical treatments and worse expected prognosis due to the lack of a valid treatment option. Developing predictive models based on somatic mutations can contribute to identifying genes therapeutically relevant (actionable), or meaningfully associated with drug sensitivity or expected outcome: these aspects are all crucial for improving patient clinical handling.

Methods

We thoroughly investigated CRC mutational profiles using Machine Learning methods to identify gene variants characterizing or co-occurring with those of the RAS family. To this aim we first collected whole-exome NGS mutational data from 3 large publicly available repositories and discarded hypermutated patients, as not to bias our following analyses. We followed 3 key steps: 1) Mutational feature encoding; 2) Development of mutation-based predictive models; 3) Feature role analysis. In 1), we applied alternative encoding strategies to optimize the information content while reducing the high dimensionality and complexity of the mutational features extracted from the somatic profiles. In 2), we set up a supervised setting and trained classifiers to recognize RAS-mutated patients. CRC patients, stratified based on RAS mutations, were split in training and testing sets, and a bootstrapping approach using Lasso Logistic Regression models was adopted to find more stable task-related features. Both Logistic Regression and Random Forest classifiers were trained on the selected features and tested to assess their predictive performance. In 3), a final step of Feature Importance analysis was used to interpret feature roles and find most relevant variants that should be worthy of further validations through biological and clinical assessments.

Results

Results of our work are both methodological and applicative. We developed a reliable pipeline to identify noteworthy variants characterizing RAS-mutated cases. Statistical evaluations with the MutSig2CV algorithm were used to identify significantly mutated genes within both RAS-mutated and not-RAS-mutated patients, obtaining a total of 164 relevant genes. Mutations occurring in such genes were considered and extended with information about consequence and position in the obtained protein. MutClustSW, a cluster-based method for detecting hotspots, enabled us to trace 186 mutational hotspots within about one half of the MutSig2CV significant genes. Hotspot mutations were considered as individual features for the subsequent Machine Learning step; all other variants of significant genes were kept but aggregating in a single feature all mutations of a specific type for a gene, leading to a wide space of 739 features. Bootstrapping was used to train 100 Lasso Logistic Regression models with 10-fold stratified cross-validation; eventually, only the most preserved and significant features were kept and used also to develop a Random Forest model, able to capture feature interactions. After testing its performances, importance and interaction strength analyses were performed to interpret and prioritize feature roles. This step confirmed some key genes, but also brought new promising data-driven associations and indications of variants to be investigated for future clinical use. Also, almost one half of the relevant mutated genes belong to the GERSOM panel (that will be soon used for mutational profiling of CRC patients enrolled in the project) and are thus driver, actionable or associated with drug-sensitivity. The implemented pipeline led to interesting results and may be employed to perform similar tasks in other cancer types, using ad hoc stratifications and extracting clinically relevant variants to possibly improve handling of critical patient groups.

Info

-	
filename	-
Figure	
-	
Availability	-

Corresponding Author	
Name, Surname	Silvia, Cascianelli
Email	silvia.cascianelli@polimi.it
Submitted on	11.05.2022