SUPPLEMENTAL METHODS (Online Only)

Genomic DNA from peripheral blood mononuclear cells, whole blood, or saliva was analyzed for single nucleotide polymorphisms (SNPs) disrupting miRNA binding sites, promoter regions or coding sequences as previously identified[1]. Biomarkers in binding sites in genes involved in the immune system and DNA damage response, as well as promoters and coding sequences of miRNAs that regulated key genes known to be critical in the DNA damage or immune response were enriched in our analysis. As a final step in our evaluation for candidate variants to test, we limited ourselves to variants predicted to be found in between 0.5% to 25% of the population, as we were looking for biomarkers that are likely to be detected in reasonably small cohorts of patients. We reduced to a final list of ~116 variants by fitting the above defined priority parameters for miRNA pathway variants. Panels were run using the Sequenome platform. Each panel was run with internal controls that used Taqman Genotyping as the gold standard. Any biomarker with less than a 90% call rate or more than 1% error found by controls was excluded from further analysis. To insure sufficient marginal variation in the final panel, any biomarker with an observed rate of mutation less than 12.5% in the training sample was excluded from the analysis.

We evaluated the relationship of this set of 116 SNPs with the incidence of major wound complications. Each SNP was defined as a categorical variable. We also included lower extremity tumor site as a categorical variable as it was the only clinical variable associated with major wound complications. The association between this panel of potential germ-line biomarkers and tumor site with wound toxicity was assessed using four classifiers on the set of 50 sarcoma patients, with wound toxicity rate 32%. Trained classifiers, whose hyperparameters were selected to optimize the F1 score with leave-one-out cross-validation (LOOCV), included classification trees (CT)[2], random forests (RF)[3], boosted trees (BT)[4], and LASSO-regularized logistic regression (LASSO-LR)[5]. The CT were tuned on minimum split and minimum observations in any terminal node, RF were tuned on number of trees and variables considered at each split, BT were tuned on the learning parameter eta, tree depth, and the number of rounds, and LASSO-LR models were tuned on the regularization parameter

lambda. The subjects with toxicity were up-weighted through oversampling method. The final performance measures, accuracy, specificity, sensitivity, negative predictive value, positive predictive value, area under the curve (AUC), and F1 score were reported using stratified 10-fold cross-validation. The threshold of number of predictors to include in our model was determined as the use of top *k* predictors allowed with the highest AUC among *k*=5, 10, 15,…, 50. Importance measures via filter method with R package *FSelector*[6] were then used to select top *k* predictors to train our classifiers and this was determined by mean rank from 1000 sample sets of their respective value. The four important measures are entropy-based information gain between predictors and response, variable importance based on ranger impurity importance, the entropy-based gain ratio between predictors and response and the univariate model score. Via 1000 over sampling sets, the order of significance of these k predictors was determined according to the obtained variable importance measure of mean decrease in the Gini impurity from the trained random forest classifier for its best prediction performance among 4 classifiers. These top *k* predictors to wound toxicity with their order of significance were reported. CT, RF, BT, and LASSO-LR classifiers were fit in R (version 3.6.0)[7] with *mlr*[8] calling *rpart*[9], *ranger*[10], *xgboost*[11], and *glmnet*[12] respectively.

The final cross validated tuning parameters for the reported classifiers are as follows: minimum split of 5 and minimum observations of 5 in any terminal node for classification trees (CT), three variables considered at each split with 15 trees for random forests (RF), learning parameter eta of 0.367, max depth of 2, and 13 rounds for boosted trees (BT) and regularization parameter lambda equal to 0.001 for LASSO-regularized logistic regression (LASSO-LR). For each classifier, all remaining hyperparameters were assigned their default values as defined through their associated R packages.

REFERENCES

1. Chen X, Paranjape T, Stahlhut C, et al: Targeted resequencing of the microRNAome and 3'UTRome reveals functional germline DNA variants with altered prevalence in epithelial ovarian cancer. Oncogene 34:2125-37, 2015
2. Breiman L, Friedman J, Olshen R, et al: Classification and Regression Trees. Boca Raton, FL, CRC Press, 2017
3. Breiman L: Random Forests. Machine Learning 45:5-32, 2001
4. Chen T, Guestrin C: XGBoost: A Scalable Tree Boosting System, KDD. San Francisco, CA, 2016
5. Tibshirani R: Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological):267-88, 1996
6. Romanski P, Kotthoff L: FSelector: Selecting Attributes, (ed R package versioon 0.31), 2018
7. R Development Core Team: Stats package (power.prop.test() function) in R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2018
8. Bernd Bischl ML, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, Zachary M. Jones: mlr: Machine Learning in R, (ed R package version 2.13), 2018
9. Therneau T AB, and Ripley B. : rpart: Recursive Partitioning and Regression Trees, (ed R package version 4.1-13), 2018
10. Marvin N. Wright SW, Philipp Probst: ranger: A Fast Implementation of Random Forests, (ed R package version 0.11.2), 2019
11. Chen T, He T: xgboost: eXtreme Gradient Boosting, (ed R package version 0.82.1), 2019
12. Jerome Friedman TH, Rob Tibshirani, Noah Simon, Balasubramanian Narasimhan, Junyang Qian: glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models, (ed R package version 2.0-16), 2018