Supplemental Information for *Consensus Machine Learning for Gene Target Selection in Pediatric AML Risk*

Transparent Methods

Pediatric AML Dataset

We accessed TARGET pediatric cancer assay and clinical data from the Genomic Data Commons (GDC, website) on February 4th, 2018. The TARGET pediatric AML cohort consists of samples from 156 patients, with tissues including primary peripheral blood (N = 26), recurrent bone marrow samples (N= 40), primary bone marrow (N = 119), and recurrent peripheral blood (N = 2). For the following analyses, we combined primary blood and bone tissues from 145 patients, retaining one sample per patient.

Gene Expression Data

RNA-seq data is from pediatric AML patients (N = 137 samples) with clinical and assay data from pediatric cancer patients from the Therapeutically Applicable Research To Generate Effective Treatments (TARGET) initiative, a collaboration between the National Cancer Institute (NCI) and Children's Oncology Group (COG) clinical trials (website). We obtained RNA-seq expression data as raw gene counts, produced using the Illumina Hi-Seq platform from Genomic Data Commons repository (https://gdc.cancer.gov/). In brief, raw reads were aligned to GRCh38 using STAR aligned in 2-pass mode and gene counts were produced using the HTSeq-counts analysis workflow with Gencode v22 annotations. Full details of the data processing pipeline can be found at the GDC ('https://docs.gdc.cancer.gov/Data/'). The GDC file manifest are included in (Supplemental Table 4). Gene counts were then normalized using trimmed mean of M (TMM) values method and converted to log2 counts per million (CPM, [Robinson and Oshlack, 2010]).

Pediatric AML Clinical Risk and Binary Risk Classifier

We defined a binary version of the clinical risk group classifier (low vs. standard or high). AML clinical risk groups are defined based on patient cytogenetics, gene sequence mutations, and other molecular data, and which pertains broadly to patient outlook and outcome in terms of risk of relapse, recurrence, and/or disease progression.

We focused on the "Risk Group" variable from the patient clinical data table. This variable is an aggregate pertaining to a combination of risk of recurrence, progression, and relapse [Arber et al., 2016]. Patients were categorized as either low or not-low (e.g. standard or high) risk, and this categorization, called binarized risk group (BRG), was used in the machine learning investigation. Patients missing data for risk group were excluded from the analysis. BRG sample groups were approximately balanced according to important demographic variables, including age at first diagnosis and gender (Table 1).

Differentially Expressed Genes (DEGs)

To reduce noise and false positive rate, we opted to exclude genes with low expression levels and which demonstrated significant differential expression in a contrast between the binarized risk groups in the training data subset using the voom function from the limma Bioconductor package ([Friedman et al., 2010]). With this pre-filter, we

identified N = 1,998 (9.33% retained) differentially expressed genes (DEGs) showing substantial mean differences between risk groups (absolute log2 fold-change >1, adj. p-value <0.05).

Machine Learning Algorithms and Hyperparameter Optimizations

We trained and tested gene expression-basd models for predicting BRG using a variety of algorithms, including two types of ensemble approaches (random forest and XGBoost), a kernel-based classifier (Support Vector Machines or SVM), and penalized regression (lasso). These algorithms quantify feature importance in the following ways: 1. Lasso assigns beta-value coefficient (positive, negative, or null/0) for use in penalized regression; 2. SVM assigns a feature weight (positive or negative) for inclusion in kernel-based estimator; 3. XGBoost assigns importance (positive or null/0) from gain across splits; 4. Random forest assigns importance using mean decrease in Gini index (positive value).

With each algorithm type, we fitted models by varying algorithm hyperparameters (Table 1, Figure 2, and Results). For Random Forest, we varied the number of trees (ntrees) from 2,000 to 10,000. For XGBoost, we varied training depth and repetitions. For SVM, we varied the kernel type to be linear or radial, and the weight filter to be none or 50%. For lasso, we varied the alpha value to be from 0.8-1.2 (Table 1 column 3). These runs informed hyperparameters used in each of the 4 algorithms with bootstraps of Boruta permutations (Supplemental Material, Figure 4).

Permutations of Sample Label Switching

To test accuracy of sample labels and quantify possible miss-classification, we performed permutation tests with risk label reassignment. For each algorithm, the training dataset class labels were randomly permuted (switched) 5000 times, such that each patient in the training set was randomly assigned to, the class label switching allows one to infer that the feature contribution for correct classification is not likely due to chance.

Ablation Tests

To characterize predictive gene sets and networks, we performed ablation tests with penalized algorithms (lasso and XGBoost). In each ablation iteration, we excluded selected gene features from all prior iterations before re-fitting and assessing fitted models with remaining DEGs. We repeated this for 15 and 70 iterations for lasso and XGBoost, respectively (Figure 3, Supplemental Figures 1 and 2, Supplemental Materials). We assessed the expression correlation (whole sample dataset) between first iteration selected genes and the next successive 2 and 3 iterations for lasso and XGBoost, respectively (Figure 3B and 3C, Supplemental Figures 1 and 2).

Analysis Code and Data Availability

Analysis was conducted on the publicly available TARGET pediatric AML cohort (Supplemental Table 4 for download manifest). The majority of analysis was conducted using the R programming language with packages from Bioconductor and CRAN repositories ([Friedman et al., 2010, Meyer et al., 2018, Chen et al., 2019, Liaw and Wiener, 2002, Kursa and Rudnicki, 2010], Supplemental Methods). Pediatric AML RNA-seq and clinical data were bundled into SummarizedExperiment objects for convenience (Supplemental Materials). Scripts, notebooks, code, and comprehensive supplemental material are available online (website link).

References

Arber et al., 2016. Arber, D. A., Orazi, A., Hasserjian, R., Thiele, J., Borowitz, M. J., Le Beau, M. M., Bloomfield, C. D., Cazzola, M., and Vardiman, J. W. (2016). The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*, 127(20):2391–2405.

Chen et al., 2019. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., and implementation), X. c. b. X. (2019). xgboost: Extreme Gradient Boosting.

Friedman et al., 2010. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22.

Kursa and Rudnicki, 2010. Kursa, M. B. and Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, 36(1):1–13.

Liaw and Wiener, 2002. Liaw, A. and Wiener, M. (2002). Classication and Regression by randomForest. 2:6.

Meyer et al., 2018. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., C++-code), C.-C. C. l., and C++-code), C.-C. L. l. (2018). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.

Robinson and Oshlack, 2010. Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25.