=

Biomarkers of Acute Myeloid Leukemia From RNA-seq Expressions and Feature Selection with Machine Learning

Sean Maden
maden@ohsu.edu
Ph.D. Candidate,
Computational Biology program
Dept. of Biomedical Engineering
Oregon Health and Science University,

David Lee
davidlee327@gmail.com

Jenny L. Smith
jlsmith3@fredhutch.org
Research Bioinformatician
Clinical Research Division - Meshinchi Lab

Ronald Buie
buierw@uw.edu
Assistant Director, NNLM Evaluation Office
PhD Candidate, University of Washington, Biomedical and Health Informatics

Vikas Peddu
Vpeddu@uw.edu
M.S. Candidate,
University of Washington Department of Laboratory Medicine, Virology

Ryan Shean
rcs333@uw.edu
University of Washington Virology
1616 Eastlake Ave E, Seattle, WA 98105

=

Abstract

Acute myeloid leukemia (AML) is a cancer of the blood and bone marrow myeloid stem cells that poses substantial population burden, especially for pediatric populations. AML is a highly heterogeneous cancer. Prior literature has characterized dozens of molecular subtypes based primarily on cytogenetics, sequence variants, structural variants, and aberrant gene expression. This body of prior work represents substantial time, energy, and resources, and we were interested in the utility of automated feature selection methods for molecular characterization of AML. We used an ensemble machine learning approach, wherein we evaluated top important features across models fitted using AML RNA-seq data and a variety of different algorithm types. We assessed consensus of feature selection results across fitted models, and whether prior literature validates these genes' functional roles in AML development, risk, and progression.

Keywords

Pediatric, leukemia, myeloid, RNA-seq, computation, machine learning

Introduction

Leukemia is a cancer of the blood arising in white blood cells of the bone marrow. It poses a substantial population burden as the most common pediatric cancer ([1,2]). Leukemia can show aggressive progression if untreated at an early stage, and depending on whether it manifests as acute or chronic.

Acute Myelogenous Leukemia (AML) is a type of leukemia impacting the myeloblast stem cells. It arises at a current rate of approximately 20,000 cases per year with 27.4% 5-year survival [2]. AML shows high molecular heterogeneity, with several clinically relevant subtypes, including the Acute promyelocytic leukemia (APL) subtype caused by a gene fusion event, and perhaps dozens of subtypes defined by various factors ranging from cell differentiation state to cytogenic and sequencing assays ([2–4]). Pediatric AML is characterized at a molecular level by rare somatic mutations, absence of common adult AML mutations, and relatively frequent structural variants ([5]). These findings indicate the importance of age-based targeted therapies and the utility of molecular assays in enabling a better understanding of clinically relevant molecular variation underlying AML cases.

Here, we apply several machine learning approaches for feature selection of RNA-seq data from both pediatric and adult AML cases. This investigation aimed to help illuminate gene

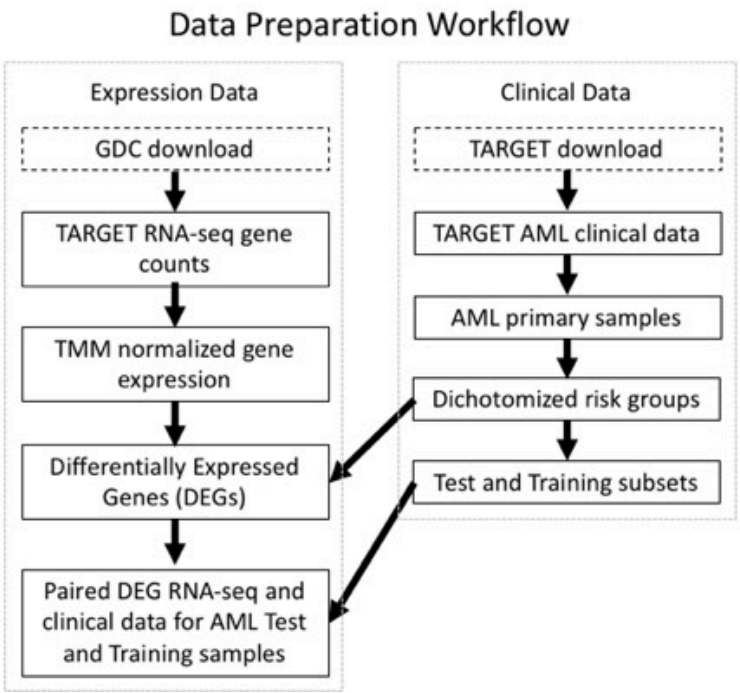
=

expression-based heterogeneity underlying AML cases, as well as age-related and -unrelated dysregulation patterns. We used clinical and assay data from pediatric cancer patients from the Therapeutically Applicable Research To Generate Effective Treatments (TARGET) initiative (https://ocg.cancer.gov/programs/target/).

In exploring the potential for machine learning to aid in molecular characterization of pediatric AML, we were interested in demonstrating the potential and efficacy of automated feature selection. Extensive prior analyses have identified and characterized molecular subtypes in AML, but these efforts require considerable time and expertise. In the era of big data science, we were interested in the potential for machine learning to reduce barriers to insight of cancer molecular subtypes.

Methods

We accessed TARGET pediatric cancer assay and clinical data, and TCGA adult cancer case data, through the project website and Genomic Data Commons (https://gdc.cancer.gov/) on February 4th, 2018. Information on patient enrollment, assay collection, and other related protocols are available in the online methods at the consortia websites ([6]; https://cancergenome.nih.gov/).



The TARGET pediatric AML cohort consists of samples from N = 156 patients, with tissues including primary peripheral blood (N = 26 samples), recurrent bone marrow samples (40 samples), primary bone marrow (119 samples), and recurrent peripheral blood (2 samples). For

=

the following analyses, we combined primary blood and bone tissues from 145 patients. TARGET clinical data includes a risk group variable that reflects patient risk of disease progression, based on an aggregate of histological and molecular evidence. **We recoded risk group to be Low (0) or Not-low (1), where the latter category included standard and high risk group samples**. We excluded samples of unknown risk. Primary pediatric AML samples were then randomly divided in training and validation subsets at a 1:1 sample size ratio, preserving frequencies of classifier categories across the data subsets.

The first iteration of feature selection involved choosing features based on variance. **For both the Low-Risk and Not-Low-Risk group, 1000 features with the most variance were selected from both groups and used to fit and test a Random Forest, XGBoost, and Gradient Boosting model.** The step was repeated selecting features with the least variance. Both methods of feature selection outperformed randomly choosing 2000 features.

The second iteration of modeling involved using all possible features. A logistic regression with Lasso regularization was fit on all features, reducing the number of important features to 3212. A Random Forest and XGBoost model were also fit and tested on all possible features, resulting in 1147 and 332 important features, respectively. The intersection of these features was then analyzed.

A second logistic regression with Lasso regularization was fit until complete convergence, resulting in a reduction to 174 features. The intersection of these 174 features and the 332 features of the XBoost model were then analyzed.

Logistic regressions were also performed using the R Bioconductor packages MLseq and DEseq2. Data were randomly subsetted with 30% used for testing and 70% used for model training. Features were selected as either low risk group, or high risk group. **The models trained and fitted in this manner were VoomNSC, voomDLDA, PLDA, PLDA2, and nblda.**

Algorithm Type	Resource Name
Support vector machines (SVM)	e1071
Random Forest, Gradient Boosting, XGBoost	XGboost; sklearn
Lasso	sklearn
glmboost	mlr
voomNSC, voomDLDA, PLDA, PLDA2, nblda, deepboost, blackboost, logitboost	MLseq;DEseq2

=

Moreover we are badasses and therefore have come up with a novel ensembl clustering algorithm that is the best thing ever.
We used a variety of mSVM with RNA sequencuing dataethods for feature selection, including…
(show methods table)
Experiment Design Table
Deliveables
We obtained RNA-seq gene expression data, and preprocessed with TMM
We then identified top differentially expressed genes between low and not-low risk groups
We trained and validated N algorithms of various classes, including...

Any file with more than 2 sample IDs associated with it was filtered out.

Implementation

This section describes how the tool works and relevant technical details for implementation.
We cleaned lots of data and then did machine learning.
We compiled TARGET data as R objects
We logged our analysis steps and code in Jupyter and R notebooks

Operation

This section should include the minimal system requirements needed to run the software and an overview of the workflow

Run it on the cloud
Run it on various linux iterations

Results

To reduce noise and false positive rate, we opted to exclude genes with low expression contrast between the classifier groups for patient risk. With this pre-filter, we identified N = 1,998 (9.33% retained) differentially expressed genes (DEGs) showing substantial mean differences between risk groups (t-test adj. p-value < 0.05). This increased the mean of expression differences

=

(means) from 0.50 to 1.71 (median increase from 0.32 to 1.51). Mean of variance differences also increased from 0.76 to 2.19 (median increase from 0.31 to 2.19).

We next fitted models to predict AML risk group using the filtered expression data. In total we used N algorithms. To determine whether a model showed acceptable predictive performance, we compared sensitivity, specificity, recall, false positive rate, loss, accuracy, and various other measures.

Gene ID	Gene Name	AML Context	Algorithms Selected
ENS#### (gene ensembl ID)	GeneA (common name)	(relevant lit if applicable)	(which algos)
...	...	...	…

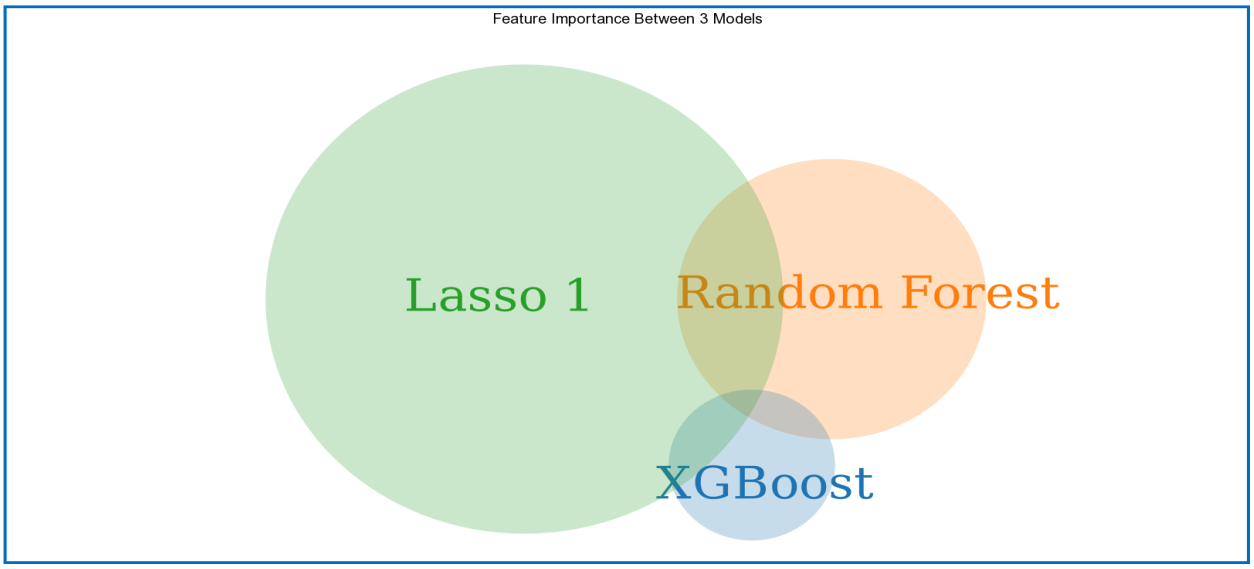
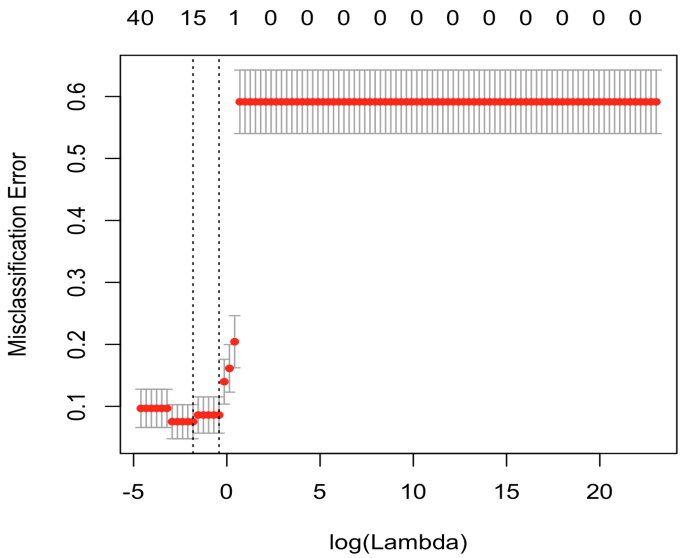


Figure X: Three models were trained and tested on all possible features. The Lasso, Random Forest, and XGBoost model deemed 3212, 1147 and 332 features important, respectively. Eleven features were deemed important by all three.

=



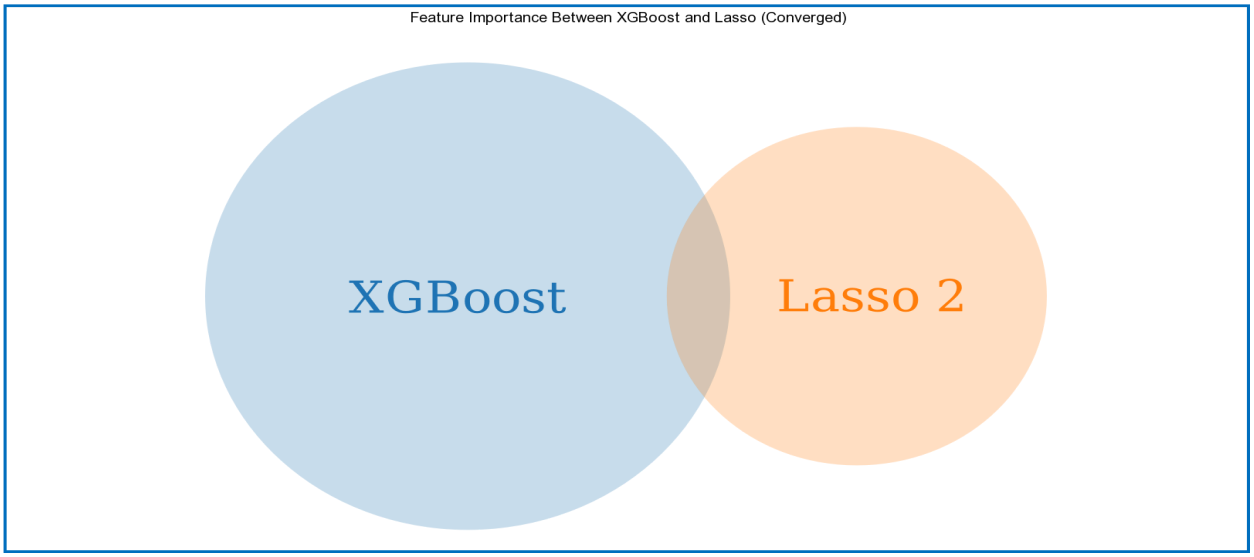
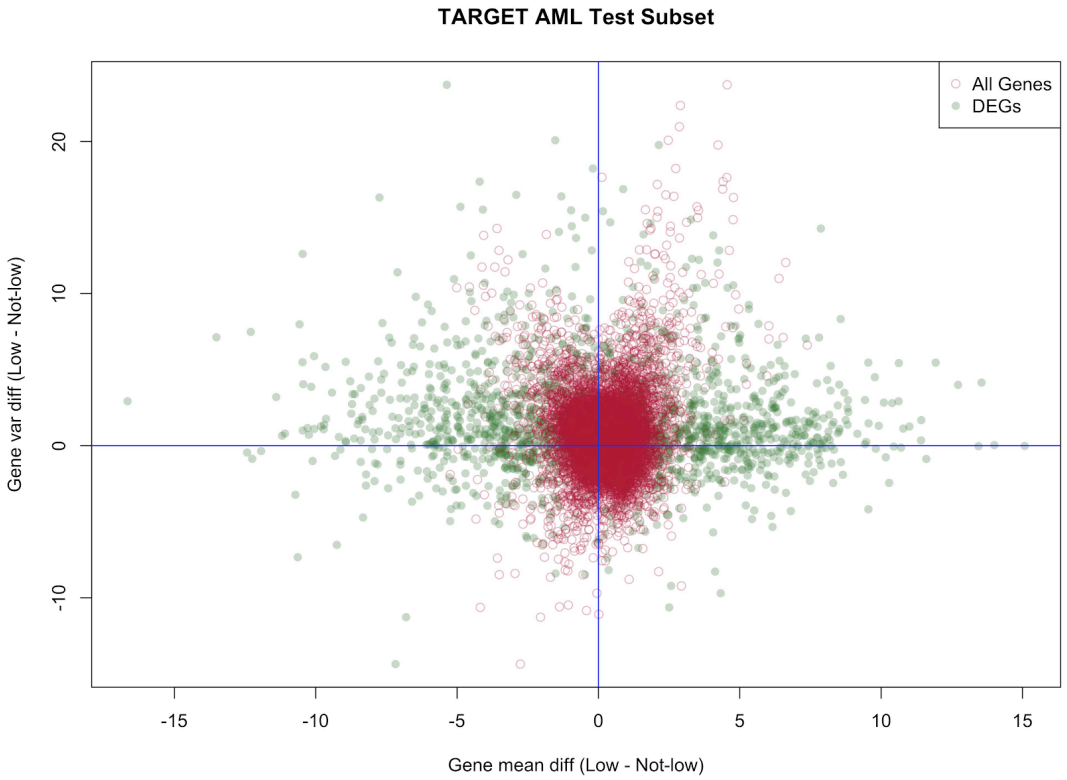
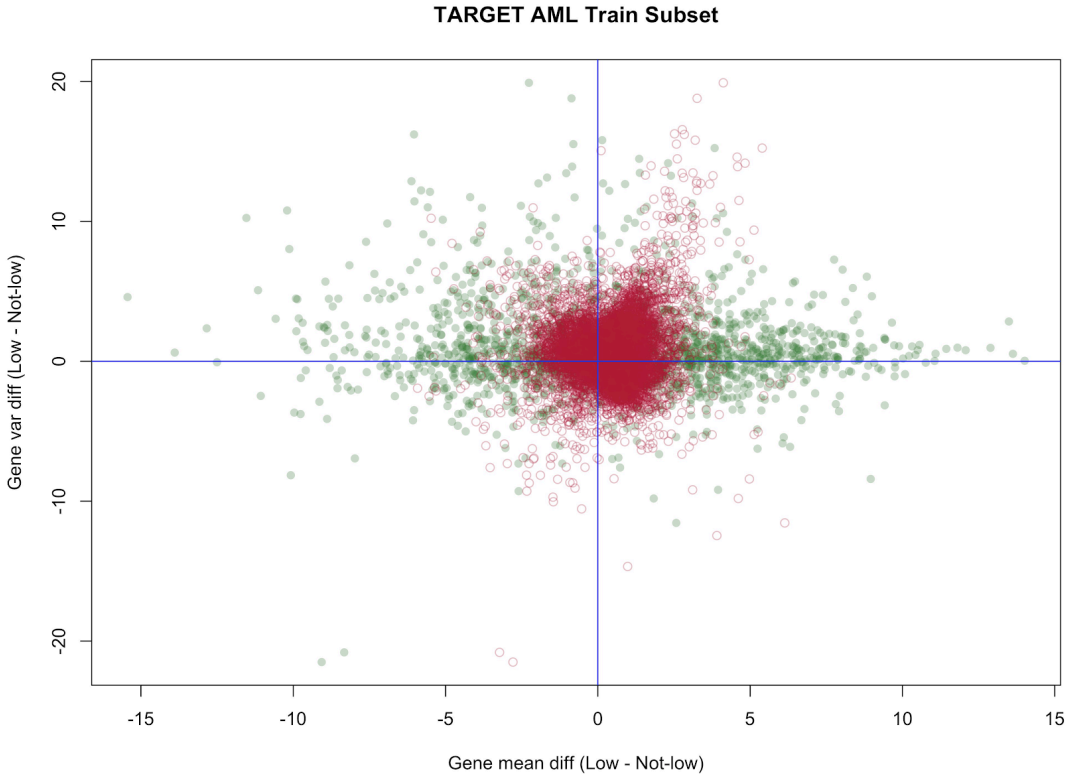


Figure X: The Lasso model was refit until convergence. The second Lasso model now deemed 174 model important. The XGBoost and Lasso model shared 15 features. Of those 15, six were in the original 11 above.

=


TARGET AML Test Subset
All Genes
DEGs
Gene var diff (Low - Not-low)
Gene mean diff (Low - Not-low)
20
10
0
-10
-15
-10
-5
0
5
10
15


TARGET AML Train Subset
Gene var diff (Low - Not-low)
Gene mean diff (Low - Not-low)
20
10
0
-10
-20
-15
-10
-5
0
5
10
15

=

	Condition Positive 34	Condition Negative 32
Predicted Positive 46	True positive 31	False positive 15
Predicted Negative 20	False Negative 3	True Negative 17

TPR = 31/34 = 91%
Precision = 31/46 = 67%
Specificity = 17/32 = 53%
False Discovery Rate = 32%

Logistic Regression with High/Standard Risk Patients

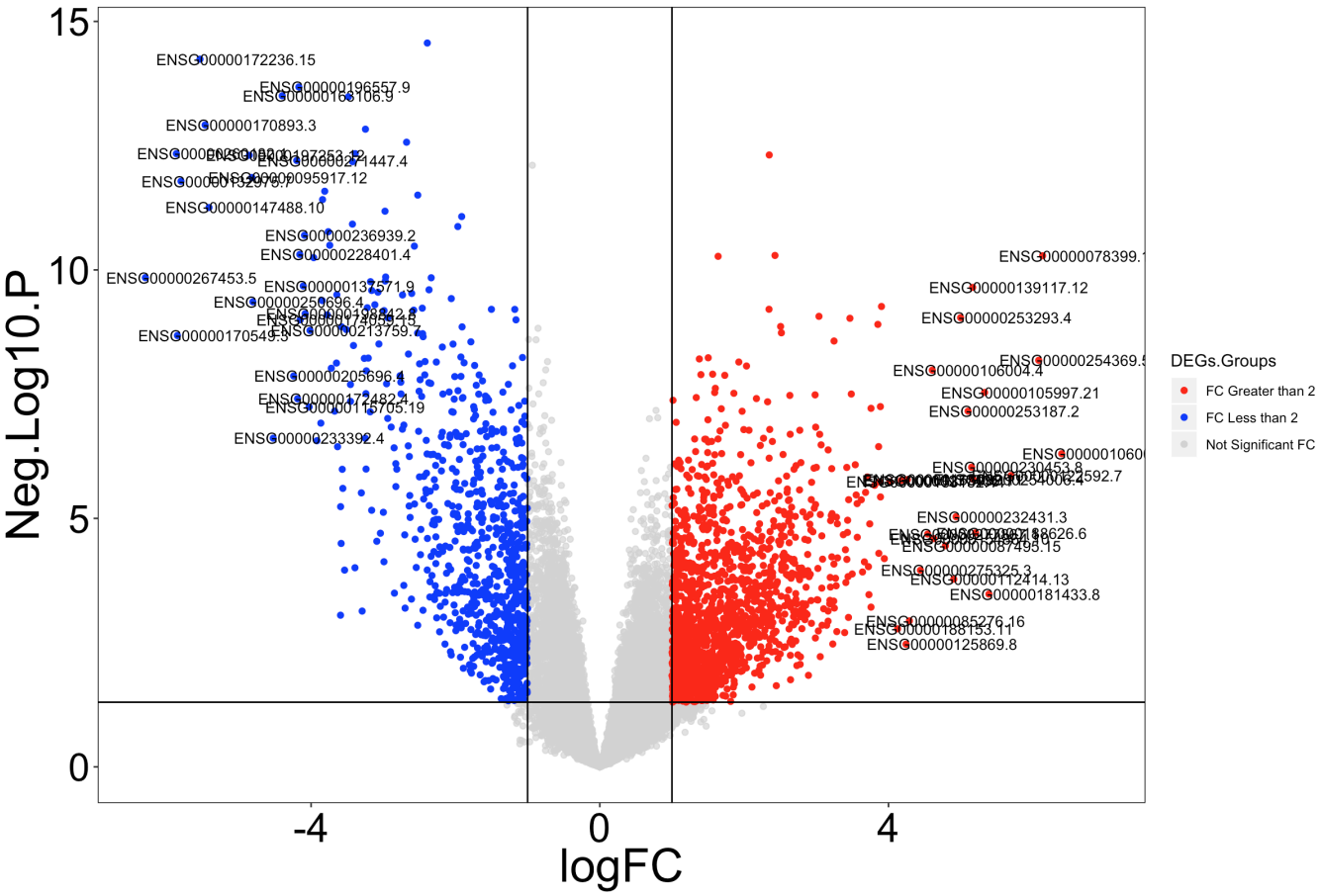
The TARGET AML RNA-seq dataset was split into training set (N=96) and test set (N=49) by randomly sampling the patient identifiers. High and Standard risk AML patients (N=55) defined by classic cytogenetic and molecular mutations were compared to low-risk AML (N=38) in the training dataset using differential expression analysis. Patients without risk classification were removed prior to analysis. There were 1,998 differentially expressed genes (absolute log2 FC > 1, adjusted p-value < 0.05), which then employed in logistic regression with the LASSO feature selection algorithm (**Fig2 A,B**). This resulted in 13 genes which were found to be associated with high/standard risk, including HOXA9, SLTRK5 and ITGB3 (**Table 2**). HOXA9 has been implicated in MLL (KMT2A)-rearranged AMLs and MLL-HOXA9 fusion has been shown to induce leukemogenesis in xenograph and mouse models (CITATION). However, the identification of SLTRK5 and ITGB3 provide powerful opportunities for targeted therapy as both are highly and aberrantly expressed on the cell-surface, allowing potential for antibody targeting or even CAR-T cell therapy technology. SLTRK5 has been shown to be aberrantly expressed in nearly 80% of AML and coincides with high/standard risk clinical features, allowing one the potential to improve outcomes for poor prognosis AML. The model performed well in predicting the risk classification of the test set with 6% error rate, and XX true positive rate and accuracy/specificity, suggesting these genes can also be used for the prediction of risk quickly using only diagnostic biological specimens, but will need to be further examined in a validation dataset.

=

Table 2. Genes associated with High/Standard molecular risk-groups.

Gene_ID	Gene name	Coefficient
ENSG00000078399	HOXA9	0.2184967372
ENSG00000101333	PLCB4	0.0931645222
ENSG00000188626	GOLGA8M	0.0868262987
ENSG00000259207	ITGB3	0.0757604883
ENSG00000165300	SLITRK5	0.0546755323
ENSG00000198722	UNC13B	0.0280847887
ENSG00000164659	KIAA1324L	0.0002750672
ENSG00000226321	CROCC2	-0.0257766559
ENSG00000267453	LINC01835	-0.0365784818
ENSG00000132514	CLEC10A	-0.0430241193
ENSG00000132975	GPR12	-0.0742916453
ENSG00000250696	AC111000.4	-0.085920489
ENSG00000260182	AC120498.2	-0.0956569772

=



Conclusion and next steps

This section should include a brief discussion of allowances made (if any) for controlling bias or unwanted sources of variability, and the limitations of any novel datasets. Also include any next steps for future development (whether your group actually plans to do this or these steps are just included a guidance for potential future development).

Our results indicate the potential for automation of feature selection to characterize subtypes of cancers showing molecular heterogeneity. Owing to these promising findings, we propose to extend this analysis to a pan-cancer experiment comparing comparable cancers from pediatric and adult samples in the TARGET and TCGA consortia, respectively. We were interested in comparing our aggregated results with various algorithms to automated meta machine learning packages.

For this investigation we focused primarily on RNA-seq data, but the data handling workflow described could be readily extended to numerous other assay types. Complementary data such as copy number variation could be further leveraged for quality control improvements to our analyses of expression data. This integration of additional data types would further increase

=

confidence in our results with the potential to further illuminate the functional roles and dynamics that explain gene dysregulation that is predictive for AML risk group.

Data and software availability

Code, including data tables, R and Python scripts and notebooks, is available at the GitHub repository (GitHub - NCBI-Hackathons/ConsensusML: Machine Learning to Detect Cancer Biomarkers from RNAseq Data). All data, code and analysis is provided under the MIT license.

Suggested Reviewers

1. **Hamid Bolouri**
2. **Timothy Triche Jr.**
3. **Soheil meshinchi**
4. **Sean Davis**

Author contributions

Sean Maden: Conceptualization, Methodology, Investigation, Visualization, Writing - Original Draft Preparation

David Lee: Methods, Results

Competing interests

No competing interests were disclosed.

Grant information

NCBI/NIH hackathons grant money
Other grant contributions

=

Acknowledgements

The authors would like to acknowledge the support of Amazon Web Services for provision of compute time and resources, the national center for biotechnology information (NCBI) and NCBI hackathons program, and Fred Hutch for hosting the hackathon where much of this work was produced.

References

Instructions on using the F1000R Google docs plug in for reference management: http://f1000.com/work/faq/google-docs-add-on/1

Instructions on using the F1000R Word plug in for reference management: http://f1000.com/work/faq/word-plugin

1. Steliarova-Foucher, E. *et al.* International incidence of childhood cancer, 2001-10: a population-based registry study. *Lancet Oncol.* **18**, 719–731 (2017).
2. Acute Myeloid Leukemia - Cancer Stat Facts. *SEER* Available at: https://seer.cancer.gov/statfacts/html/amyl.html. (Accessed: 5th February 2019)
3. Yi, G. *et al.* Chromatin-Based Classification of Genetically Heterogeneous AMLs into Two Distinct Subtypes with Diverse Stemness Phenotypes. *Cell Rep.* **26**, 1059-1069.e6 (2019).
4. Tyner, J. W. *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526 (2018).
5. Bolouri, H. *et al.* The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nat. Med.* **24**, 103–112 (2018).
6. Cancer Genome Atlas Research Network *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).

Figures and Tables

Figure 1. Workflow and methods for discovery of AML biomarkers.

Table 1. Machine learning algorithms applied for biomarker discovery.

Table 2. Genes associated with High/Standard molecular risk-groups.

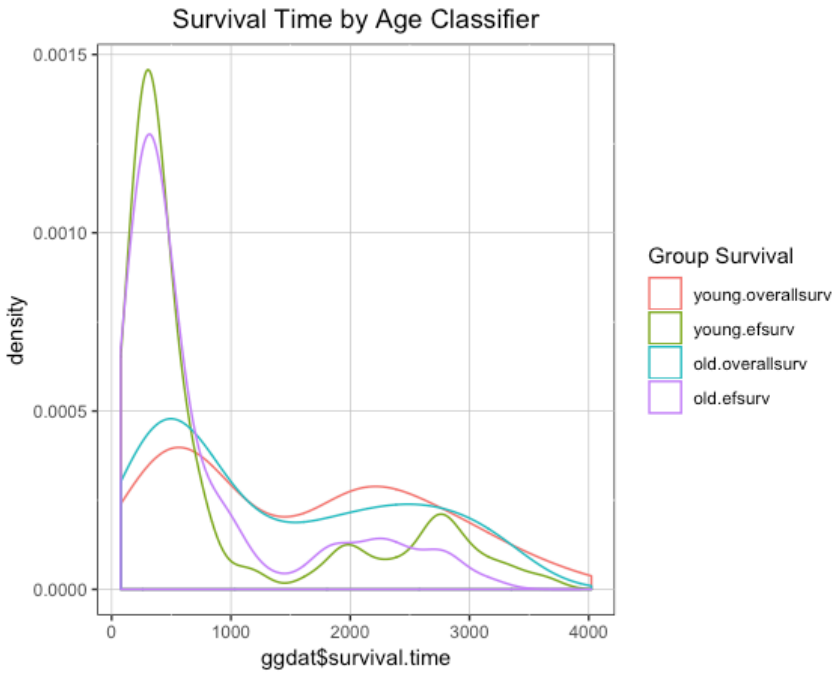
=

Figure 2. Scatter plots of gene expression contrasts in test (top) and training (bottom) AML data subsets. X-axis is difference in expression means between dichotomized risk groups, y-axis is difference in expression variances between these groups. Green dots are differentially expressed genes (DEGs, N = 1,998 loci), red circles are all genes (N = 21,407 loci).

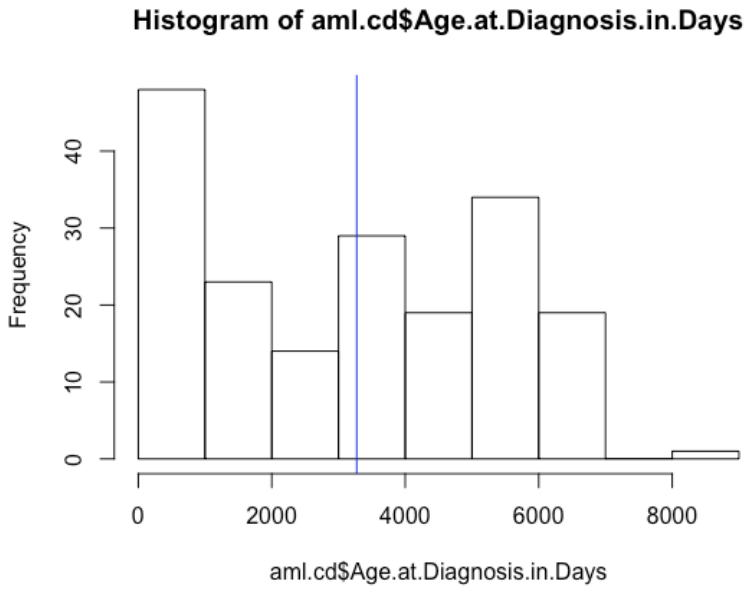
Figure X: Three models were trained and tested on all possible features. The Lasso, Random Forest, and XGBoost model deemed 3212, 1147 and 332 features important, respectively. Eleven features were deemed important by all three.

Figure X: The Lasso model was refit until convergence. The second Lasso model now deemed 174 model important. The XGBoost and Lasso model shared 15 features. Of those 15, six were in the original 11 above.

Supplemental Figures and Tables



=


Histogram of aml.cd$Age.at.Diagnosis.in.Days
Frequency
0
10
20
30
40
0
2000
4000
6000
8000
aml.cd$Age.at.Diagnosis.in.Days