Gene Expression Assay: A New Panel for Early Metastatic Risk Estimation for Breast Cancer

Melih Agraz (melih_agraz@brown.edu)
Brown University
Umut Agyuz
Genz Biotechnology
E. Celeste Welch
Brown University
Kaymaz Yasin
Harvard University
Kuyumcu Birol
Sefamerve R&D Center

Research Article

Keywords: cancer, machine learning, microarray, variable importance

Posted Date: March 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-279461/v1

RESEARCH

Gene Expression Assay: A New Panel for Early Metastatic Risk Estimation for Breast Cancer

Melih Agraz[1*], Umut Agyuz[2], E. Celeste Welch[3], Kaymaz Yasin[4], Kuyumcu Birol[5]

*Correspondence:
melih_agraz@brown.edu
[1]Brown University, Department of Applied Math, 170 Hope Street, Providence, US
Full list of author information is available at the end of the article

Background

Metastasis is one of the most challenging problems in cancer diagnosis and treatment, as its causes have not been yet well characterized. Prediction of the metastatic status of breast cancer is important in cancer research because it has the potential to save lives. However, the systems biology behind metastasis is complex and driven by a variety of factors beyond those that have already been characterized for various cancer types. Furthermore, prediction of cancer metastasis is a challenging task due to the variation in parameters and conditions specific to individual patients and mutation of the sub-types.

Results

In this paper, we apply tree-based machine learning algorithms for gene expression data analysis in the estimation of metastatic potentials within a group of 490 breast cancer patients. Hence, we utilize tree-based machine learning algorithms, decision trees, gradient boosting, and extremely randomized trees to assess the variable importance.

Conclusions

We obtained highly accurate values from all three algorithms, we observed the highest accuracy from the Gradient Boost method which is 0.8901. Finally, we were able to determine the 10 most important genetic variables used in the boosted algorithms, as well as their respective importance scores and biological importance. Common important genes for our algorithms are found as CD8, PB1, THP-1. CD8, also known as CD8A is a receptor for the TCR, or T-cell receptor, which facilitates cytotoxic T-cell activity and its association with cancer is defined in the paper. PB1, PBRM1 or polybromo 1 is a tumor suppressor gene. THP-1 or GLI2 is a zinc finger protein referred to as "Glioma-Associated Oncogene Family Zinc Finger 2". This gene encodes a protein for the zinc finger, which binds DNA and mediate Sonic hedgehog signaling (SHH). Disruption in the SHH pathway have long been associated with cancer and cellular proliferation.

Keywords: cancer; machine learning; microarray; variable importance

1 Background

Metastasis begins with the displacement of tumor cells from the primary tumor. Circulating tumor cells (CTCs) move through the vascular system to a distant organ. There, they colonize the new environment, forming a new tumor.

Metastasis is one of the most complex and challenging problems in cancer, as its main causes are multifaceted and not yet well understood. Additionally, it is strongly correlated with patient death, making it the most critical problem to anticipate and treat within the field of cancer diagnostics (Dillekas et al., 2019). Metastasis begins with the loss of cell-cell and cell-matrix adhesion. This facilitates local infiltration of tumor cells into adjacent tissues as well as transendothelial migration into vessels via the process of intravasation. Cancer cells must transform themselves from endothelial cells into mesenchymal cells, known as epithelial mesenchymal transition (EMT). This process is characterized by the loss of cellular adhesive properties and polarity with simultaneous gain of other properties that enable CTCs to migrate to distant organs, extravasate, proliferate and colonize a discrete competent organ. The other major factors for metastasis are cell adhesion defects, angiogenesis and disrupted cell signaling. Disrupted cell signaling interrupts foreign recognition response, allowing cancer cells to pass through the blood without being recognized by the immune system. However, CTCs are able to evade immune recognition by mimicking peripheral immune tolerance, as recently detailed by Gonzalez et al. (Gonzalez et al., 2018). The relations among these biological systems have most frequently been characterized in late stage metastatic tumors. The CTCs have abnormal gene expression characteristics (Wang et al., 2018). The gene expression characteristics are different than primary tumor and that helps them to survive in blood. The survivin is the major inhibitor of the apoptosis and assists the escape of tumor cells from immune recognition by blocking the cytotoxicity of NK cell and PD-L1 can mediate the regulatory T-cells (Tregs) to play a role in immunosuppression. This is a fact that gene expression analysis of CTCs revealed a genetic variation as compared to the primary tumor but it is still ambiguous how metastatic mechanisms begin in the primary tumor in early stages and how expression changes over time (Bertucci et al., 2019). This is important not only from the basic science perspective, but from the diagnostic and predictive perspective as well. Many lives can be saved when the anti-metastatic treatments are started earlier. Until then, inability to reliably characterize metastasis continues to drive cancer's reputation as the most unpredictable and challenging illness to treat, resulting in lower-than-predicted survival times (Glare et al., 2003).

Machine learning is a combination of statistics and computer science which has become popular in recent years due to increases in both data availability and quantity. Hereby, this approach has recently become popularized in specifically bioinformatics and cancer research (Cruz and Wishart, 2006). As machine learning capabilities grow, predictive models have become more and more accurate at determining cancer metastasis. For example, Huang et.al. (Huang et al. 2017) used support vector machine (SVM) and SVM Ensembles to predict breast cancer in 2017, Behravan et al. (Behravan et al.) predicted the breast cancer risk by machine learning algorithm for genetic and demographic datasets, (Xiaoa et.al 2018) used deep learning in cancer prediction, Kadir and Gleeson (Kadir et al.2018) implemented machine learning methods in prediction of image dataset for lung cancer and Azzawi et.al (Azzawi, 2016) made a lung cancer prediction from microarray data. Decision trees are one of the most popular non-parametric

supervised classification machine learning algorithms. They are used to classify the data in the form of an inverted tree that consists of a leaf node, root node and internal node (Sonf and Lu, 2015). The Extremely-Randomized Trees model is a tree based ensemble model which was first introduced by Geurts et.al (Geurts et.al, 2006) in 2006. This algorithm is similar to the random forest (RF) model which selects the subset of the K features when deciding to split on each node. However, the difference between the random forest and ERT model is that ERT creates the trees from the learning samples. The GB tree model is an ensemble model technique thought to originate from the study of Breiman (Breiman, 1997), that was actually developed by Friedman (Friedman, 1999).
Due to this success of this approach in distinct biological data, in this study, we perform it to the analysis of the metastasic gene expression data from breast cancer patients. These gene expression datasets are publicly available. The datasets have 23397 genes and 490 individuals. Because the starting datasets contain significant amounts of information i.e cancer type tumor grade and age for our analysis, t-statistics and a Bayesian approach were applied to select only significant ones out of the 23397 genes. We thereby determined the differentially expressed genes between 2 groups: metastatic and non-metastatic phenotypic profiles. A Differential Gene Expression (DGE) Analysis were performed between 2 groups (metastatic and non-metastatic patients' gene expression profiles), using R. By this way, 133 significant transcripts were detected (fold change > 1.5) under both groups were obtained by reducing the dimension of the gathered data considerably. In the subsequent framework, we compared metastatic and non-metastatic tumor transcripts' expressions. To do that, we first applied tree based machine learning algorithms to our reduced data and then used variable importance to see the variable response, and finally, we interpreted the variables which affect the metastatic potential of breast cancer.

2 Materials and Methods

There are two main aims of this study. The first is to show which of the tree-based algorithms is the most efficient in array analysis, and the second is to demonstrate which transcript outputs of these algorithms are the most important both biologically and for future modeling approaches. Therefore, we first determined the model with the highest accuracy by applying these data to machine learning algorithms including decision trees (DT), gradient boosting (GB) and extremely randomized trees (ERT). From this, we are able to determine the variables with the highest value of metastatic prediction as given by these models.
In this study two different publicly available datasets are used. The two datasets are publicly available on NCBI GEO Databank. The datasets, called GSE102484 [20] and GSE20685 [19], are merged together to create a single dataset. GSE102484 includes microarray data from tissue of I-III stage breast cancer patients who underwnent primary surgery in a freestanding cancer center in Taiwan and GSE20685 includes gene expression profiling data was conducted on fresh frozen breast cancer tissue collected from 327 patients. Both datasets were obtained from different experimental studies using the same platform chip, which is GPL570 (HG U133P lus2) Affymetrix Human Genome U133 Plus 2.0 Array.

Due to the standard collection methods, the datasets were able to be combined into one to create a more robust model. These two dataset were combined using R-programming language. R script was then subsequently used for data normalization. Background signals are removed, outliers are omitted, fitted to normal distribution and quality controls are checked. The Bioconductor RMA package was used for inter arrays data normalization. After the combination of the two underlying datasets, there were 23397 transcripts in total. 44 transcripts that had missing (NA) values were then extracted from the analyses. Later, the biological samples were categorized as belonging to either the metastatic or non metastatic group by using data feature of patients. The feature patient data included factors such as T-stage (tumor size classification), N-stage (lenf node positive/negative), M- stage (metastasis classification), age and metastasis information.

In the application, we split the data as train (%80) and test (%20) sets as 5-fold cross validation, then we train the model with the Decision Tree, Extremely Randomized Tree and Gradient Boosting approaches. These models were selected because tree structures are powerful in modeling and these particular approaches represented variable importance of genes within the model of the tree structures. Then the precision, recall, F1-score and accuracy are calculated. The formula of all accuracy measures can be seen in the supplementary document.

3 Results and Discussion

3.1 Conclusion and Discussion

Table 1, it is seen that the Gradient Boosting approach is the algorithm that has the most accurate results when compared to the others for all metrics of precision, recall, F1-score and accuracy. It is clearly seen in the table that Gradient Boosting is very successful for predicting the true metastasis classes which can be formalized with precision, recall, F1-score and calculated as 0.8901, 0.8550, 0.8666, respectively.

Table 1: Machine learning application results applied for cancer data for 133 features.

	Precision	Recall	F1-score	Accuracy
Decision Tree	0.6969	0.7037	0.6985	0.7073
Extremely Randomized Trees	0.8545	0.7925	0.8067	0.8293
Gradient Boosting Tree	**0.8901**	**0.8550**	**0.8666**	**0.8780**

Tree models give us the most important variables of the algorithms and we illustrate the outputs of the each variable importance of each model in Figure 1-3. The figures demonstrates the most significant 10 variables as determined via the use of each algorithm, as well as their respective contributions to the model as Table 2 in same order. Figure 1-3 represent the variable importance of decision trees, and they show the variable importance randomized trees of array IDs. We also represent the gene translated version of array IDs of Figure 1-3. Table 3 illustrates the variable importance of all algorithms as in the same order in Table 2. In this table, each variable is listed from high to low importance and common

Figure 1: Variable importance of Decision Trees for array IDs.

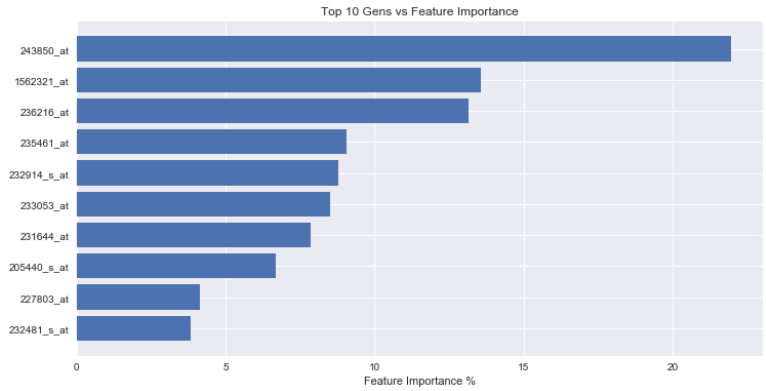


Figure 2: Variable importance of Extremely Randomized Trees for array IDs.

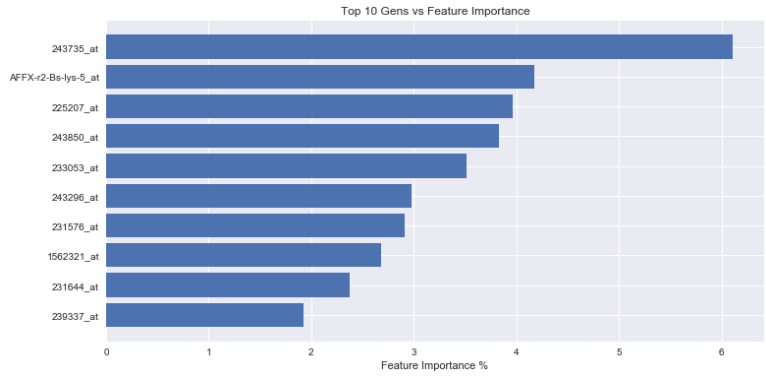


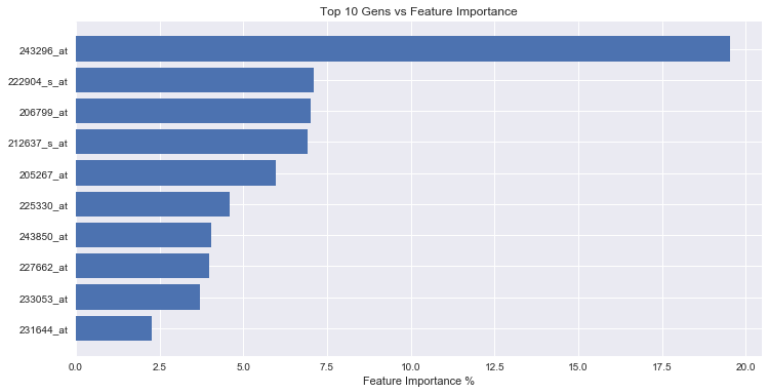
Table 2: Variable importance of corresponding genes as assessed by different algorithms

Decision Tree	Extremely Randomized Tree	Gradient Boosting
CD8	ELP2	NAMPT
PDK4	N/A	N/A
TCF7L2	PDK4	SCGBID2
TET2	**CD8**	E3 Ubiquitin
AL577781	**PB1**	POU2AF1
PB1	NAMPT	IGF1R
THP-1	**ETNK1**	**THP-1**
N/A	PDK4	SYNPO2
ENPP5	THP-1	**PB1**
SLITRK6	AL577781	**THP-1**

important genes for our algorithms are found as CD8, PB1, THP-1 as shown in bold lettering.

This analysis enabled a variety of metastatic biomarkers to be pinpointed. The genes with the most significant differential expression are discussed herein. CD8, also known as CD8A and cluster of differentiation 8, is a transmembrane glycoprotein that was found to be of key significance in this analysis. It is a receptor for the TCR, or T-cell receptor, which facilitates cytotoxic T-cell activity. Up-regulation has been associated with poor cancer prognosis in recent work by

Figure 3: Variable importance of Gradient Boosting for array IDs.



Saleh et al. (Saleh et al., 2020). PB1, PBRM1 or polybromo 1 is a tumor suppressor gene. Mutations in this gene are ubiquitous across multiple cancer subtypes. This gene encodes an ATP dependent chromatin-remodeling complex. THP-1 or GLI2 is a zinc finger protein referred to as "Glioma-Associated Oncogene Family Zinc Finger 2". This gene encodes a protein for the zinc finger, which binds DNA and mediate Sonic hedgehog signaling (SHH). Disruption in the SHH pathway have long been associated with cancer and cellular proliferation. The pathway has also been implicated in evolving treatment resistance (Carballo et al., 2018). Lastly, the ETNK1 ethanolamine kinase 1 gene encodes the EKI1 kinase protein. This protein is involved in the phosphatidylethanolamine synthesis pathway. Mutations thus affect glycerophospholipid biosynthesis and metabolism. Other significant genes were found as well. Two transcripts of interest (1562321_at and 225207_at)were found to correspond to PDK4, or pyruvate dehydrogenase kinase 4. PDK4 is a PDK-BCKDK protein kinase which encodes a mitochondrial histidine kinase protein. When mutated, pyruvate dehydrogenase is no longer regulated, leading to corresponding disregulation of glycolysis. PDK4 mutations are ubiquitous in fast-growing cancer cells. ELP2, or elongator acetyltransferase complex subunit 2, is another gene of interest. ELP2 encodes a core subunit of the histone acetyltransferrase of RNApolII, and is necessary for chromatin remodeling, which is disregulated in cancer. IGF1R encodes the insulin like growth factor 1 receptor, responsible for binding IGF and exhibiting tyrosine kinase activity. IGF1R is overexpressed in cancers, which confers mutated cells with anti-apoptotic properties. TET2, or Tet methylcytosine dioxygenase 2, encodes a gene which catalyzes conversion of methylcytosine to 5-hydroxymethylcytosine.Gene defects can cause myeloproliferation. POU2AF1 encodes a Class 2 Homeobox Associating Factor that associated with OCT1 and OCT2. Defects have been associated with lymphoma. SCGB1D2, which encodes Secretoglobin Family 1D Member 2, or Prostatein-like Lipophillin B. As a prostatatein analog, the protein encoded by this gene can bind steroid hormones and similar chemotherapeutic agents such as estramustine. SYNPO2 produces the protein Synaptopodin 2, which functions in actin bindling and bundling into F-actin. This is necessary for the formation of Z disks and stable autophagocytitic

function. ENPP5 is a member of the Ectonucleotide Pyrophosphatase and Phosphodiesterase Family. ENPP5 encodes a type-1 transmembrane glycoprotein which is a prognostic marker in a variety of cancer types.

Table 3: Variable cancer specificity of genes according to the Human Protein Atlas

Decision Tree	Extremely Randomized Tree	Gradient Boosting
CD8: All cancers	ELP2: All cancers	NAMPT: All cancers
PDK4: Many cancers	N/A	N/A
TCF7L2: All cancers	PDK4: Many cancers	SCGBID2: Some cancers
TET2: Many cancers	**CD8: All cancers**	E3 Ubiquitin: All cancers
AL577781: Highly specific	**PB1: All cancers**	POU2AF1: Many cancers
PB1: All cancers	NAMPT: All cancers	IGF1R: All cancers
THP-1: Many cancers	**ETNK1: All cancers**	**THP-1: Many cancers**
N/A	PDK4: Many cancers	SYNPO2: All cancers
ENPP5: All cancers	THP-1: Many cancers	**PB1: All cancers**
SLITRK6: Many cancers	AL577781: Highly specific	**THP-1: Many cancers**

Lastly, we created a network analysis of the output of Gradient Boosting results that were the most successful algorithms within the machine learning analysis. We used the online Genemani bioinformatics tool [10] that searches for information on the internet and performs network analysis to determine key interactions in results.

In this tool, a link showing the interaction between each pair of genes within the target pool is created by analyzing the relationships within the data. We analyzed the co-expression of transcripts and selected the interaction links defined based on previously categorized relationships from data presented in GeneMania Online Tool (https://genemania.org/). These findings were used to refine a target network for downstream network analysis. Thus, we have created a network by looking at the expressions of the differentially expressed genes that we found in different studies. Although it would be possible to create an entirely novel network based solely on the data presented here, it would be very limited and biased towards our data and prexisting results. Because of this, a well-categorized method already accepted in the literature was used for the foundation of the network analysis.

In this study, a genetic expression dataset was modeled via a decision tree, extremely randomized tree and gradient boosting tree. These three algorithms are similar in that they are all tree algorithms and are powerful in their ability to predict variable importance. After the training of the model, it was observed that the gradient boosting tree was the most powerful algorithm for predicting metastatic potential within the breast cancer dataset. These three algorithms were able to determine the most effective genetic predictors of metastatic potential and we choose 10 out 23397 possible predictors. It is seen that 243850_at, 233053_at, 231644_at and 231576_at, are common effective transcripts for predicting breast cancer metastasis, indicating that CD8, PB2, THP-1 and ETNK1 are important genes of interest.

In the future, we want to extend the study by adding more available next generation sequencing (NGS) data and using causal inference methods.

Author details

[1]Brown University, Department of Applied Math, 170 Hope Street, Providence, US. [2]Genz Biotechnology, 5699 Sk No:7/4 Çankaya, Ankara, Turkey. [3]Brown University, Center for Biomedical Engineering, 184 Hope Street, Providence, US. [4]Informatics Group, Harvard University, 38 Oxford Street, Cambridge, US. [5]Sefa Merve RD Center, Cevizli, Akasya Sk. No:67, 34846 Maltepe, Istanbul, Turkey.

References

1. Azzawi H, Hou J, Xiang Y, Alanni R. Lung cancer prediction from microarray data by gene expression programming. 2016: IET Systems Biology 10:168–178.
2. Behravan, H., Hartikainen, J.M., Tengström, M. Kosma, V.A. and Mannermaa, A. Predicting breast cancer risk using interacting genetic and demographic factors and machine learning, Nature Scientific Report, 2020. 10(1),1-16.
3. Bertucci, F., Ng, C.K.Y., Patsouris, A. et al. Genomic characterization of metastatic breast cancers. Nature 2019 (569), 560–564
4. Breiman, L. . "Arcing The Edge". Technical Report 486. Statistics Department, University of California, Berkeley.
5. Carballo, G.B., Honorato, J.R., de Lopes, G.P. et al. A highlight on Sonic hedgehog pathway. Cell Commun Signal 16, 11 (2018).
6. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2007 2:59-77.
7. Glare P, Virik K, Jones M, Hudson M, Eychmuller S, Simes J, Christakis N. A systematic review of physicians' survival predictions in terminally ill cancer patients. BMJ. 2003, 327(7408):195-8.
8. Dillekas, H, Rogers, MS, Straume, O. Are 90% of deaths from cancer caused by metastases? Cancer Med. 2019; 8: 5574– 5576. https://doi.org/10.1002/cam4.2474
9. Friedman, J. H. (February 1999). Greedy Function Approximation: A Gradient Boosting Machine (PDF).
10. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright G, Bader GD, Morris Q. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W214-20. doi: 10.1093/nar/gkq537. PMID: 20576703; PMCID: PMC2896186.
11. Geurts, P., Ernst, D. and Wehenkel, L. Extremely randomized trees, Machine Learning, 2006; 63: 3-42.
12. Gonzalez H., Hagerling C., Werb Z. Roles of the immune system in cancer: from tumor initiation to metastatic progression. Genes Dev. 2018(32), 1267–1284.
13. Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF. SVM and SVM Ensembles in Breast Cancer Prediction. PLoS One. 2017;12(1):e0161501. Published 2017 Jan 6. doi:10.1371/journal.pone.0161501
14. Kadir T, Gleeson F. Lung cancer prediction using machine learning and advanced imaging techniques. Transl Lung Cancer Res. 2018;7(3):304–312. doi:10.21037/tlcr.2018.05.15
15. Saleh R, Sasidharan Nair V, Toor SM, et al. Differential gene expression of tumor-infiltrating CD8+ T cells in advanced versus early-stage colorectal cancer and identification of a gene signature of poor prognosis, Journal for Immuno Therapy of Cancer, 2020, e001294.
16. Song, Y. and Lu, Y. Decision tree methods: applications for classification and prediction, Shanghai Archives of Psychiatry, 2015; 27(2): 130-135.
17. Wang WC, Zhang XF, Peng J, Li XF, Wang AL, Bie YQ, Shi LH, Lin MB, Zhang XF. Survival Mechanisms and Influence Factors of Circulating Tumor Cells. BioMed Research International. 2018, 6304701.
18. Xiaoa, Y., Wu, J., Linc, Z. and Zhao, X. A deep learning-based multi-model ensemble method for cancer prediction, Computer Methods and Programs in Biomedicine, 2018: 153, 1-9
19. Cheng, Skye Hung-Chun, et al. "Validation of the 18-gene classifier as a prognostic biomarker of distant metastasis in breast cancer." PloS one 12.9 (2017): e0184372.
20. Kao, Kuo-Jang, et al. "Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization." BMC cancer 11.1 (2011): 1-15.

Ethics declarations

Ethics approval and consent to participate

No ethics approval was required for the study.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Availability of data and materials

In this study, two different publicly available datasets are used. The two datasets are publicly available on NCBI GEO Databank (https://www.ncbi.nlm.nih.gov/geo/). The datasets, called GSE102484 and GSE20685, are merged together to create a single dataset.

Figures

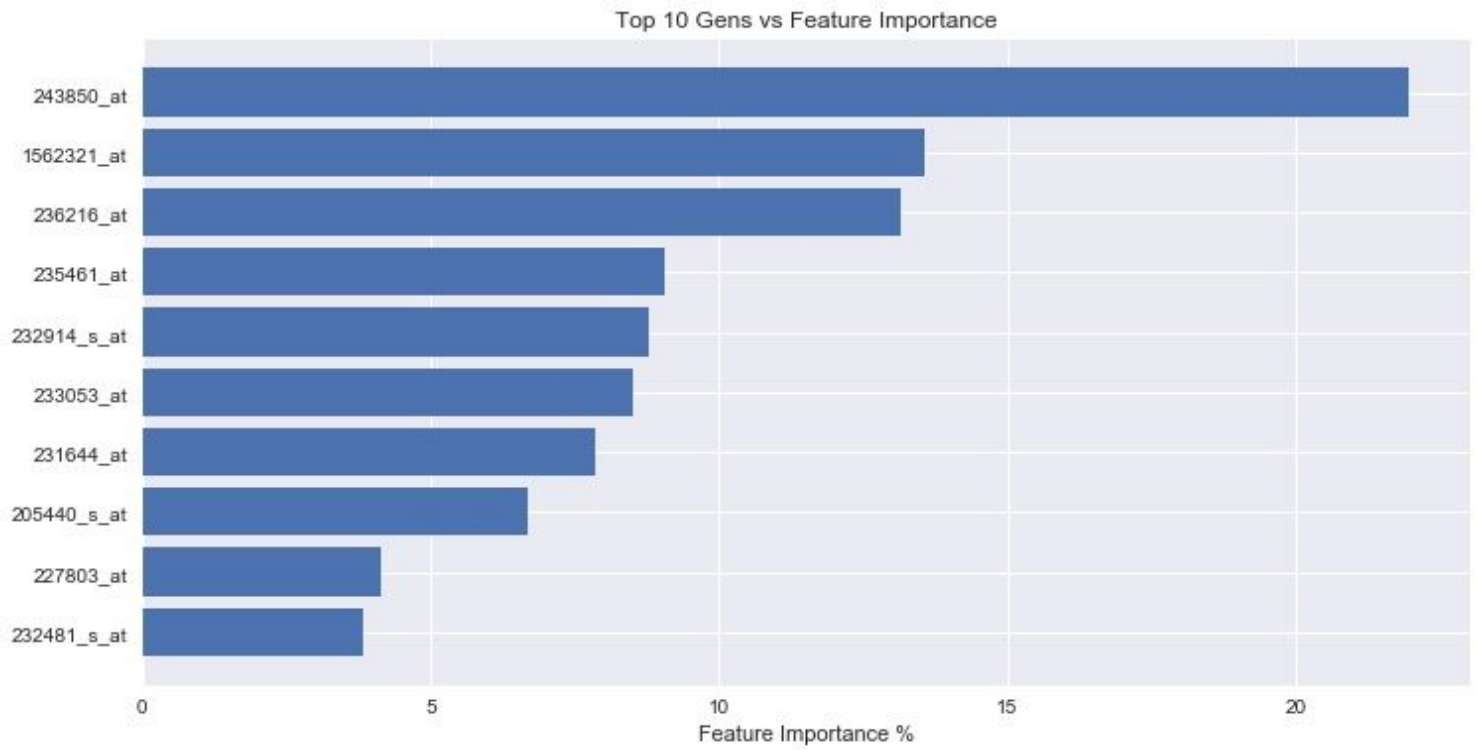


Figure 1

Variable importance of Decision Trees for array IDs.

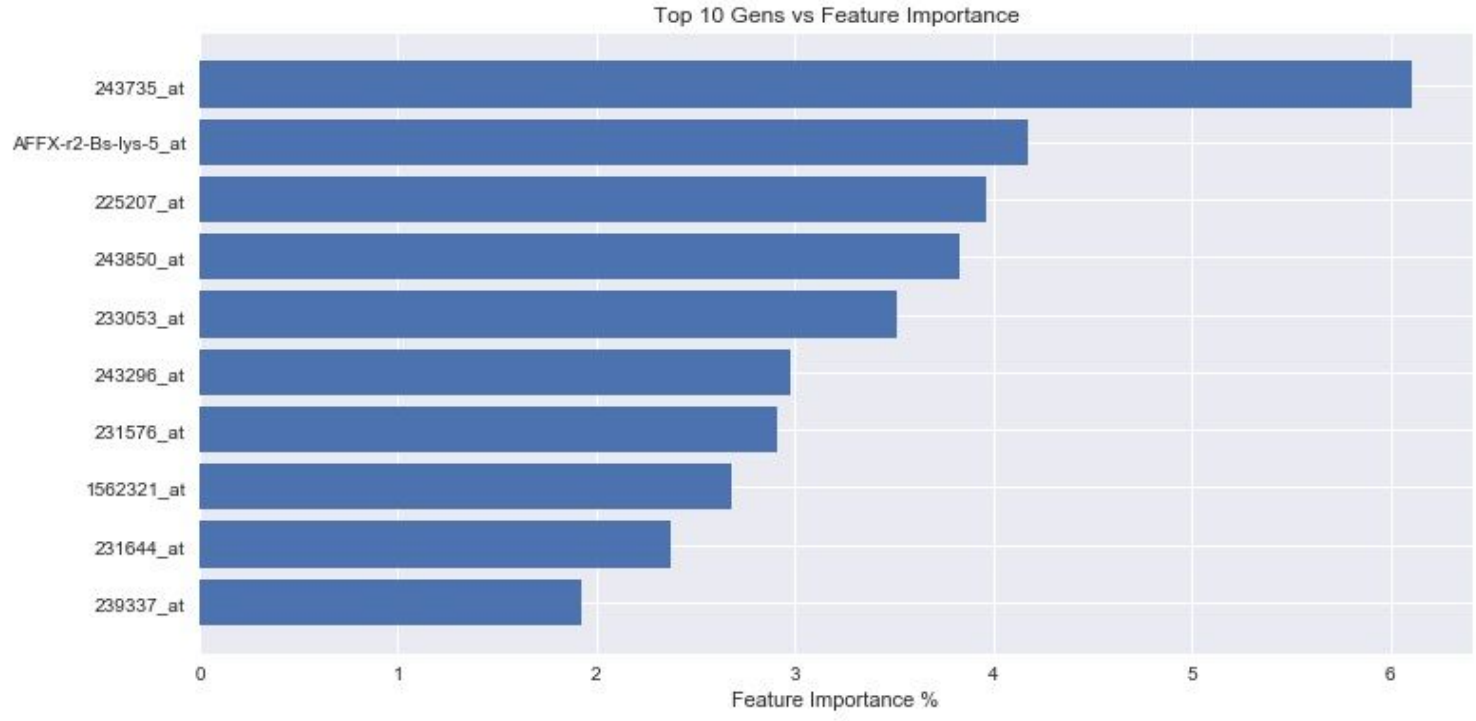


Figure 2

Variable importance of Extremely Randomized Trees for array IDs.

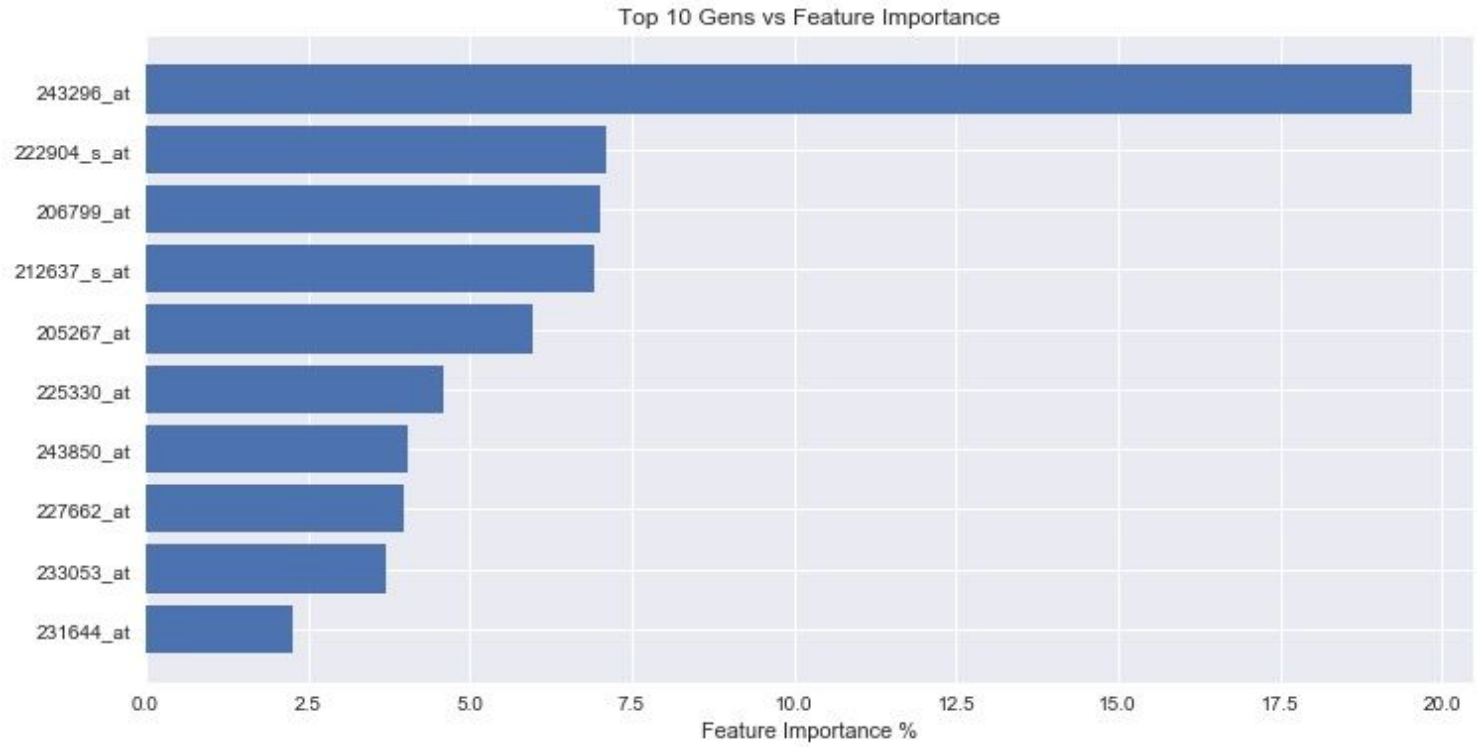


Figure 3

Variable importance of Gradient Boosting for array IDs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Tables.pdf
- supplementary.pdf