4

Machine Learning Approaches for Biomarker Discovery Using Gene Expression Data

Xiaokang Zhang[1,2] • Inge Jonassen[2,3] • Anders Goksøyr[4]

[1,2]Department of Molecular Oncology, Institute for Cancer Research, Oslo University Hospital-*Radiumhospitalet*, Oslo, Norway; [2]Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway; [3]Center for Cancer Biomarkers, Department of Informatics, University of Bergen, Bergen, Norway; [4]Department of Biological Sciences, University of Bergen, Bergen, Norway

Author for correspondence: Inge Jonassen, Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway. Email: inge.jonassen@uib.no

Doi: https://doi.org/10.36255/exonpublications.bioinformatics.2021.ch4

Abstract: Biomarkers are of great importance in many fields, such as cancer research, toxicology, diagnosis and treatment of diseases, and to better understand biological response mechanisms to internal or external intervention. High-throughput gene expression profiling technologies, such as DNA microarrays and RNA sequencing, provide large gene expression data sets which enable data-driven biomarker discovery. Traditional statistical tests have been the mainstream for identifying differentially expressed genes as biomarkers. In recent years, machine learning techniques such as feature selection have gained more popularity. Given many options, picking the most appropriate method for a particular data becomes essential. Different evaluation metrics have therefore been proposed. Being evaluated on different aspects, a method's varied performance across different datasets leads to the idea of integrating multiple methods. Many integration strategies are proposed and have shown great potential. This chapter gives an overview of the current research advances and existing issues in biomarker discovery using machine learning approaches on gene expression data.

In: *Bioinformatics*. Nakaya HI (Editor). Exon Publications, Brisbane, Australia.
ISBN: 978-0-6450017-1-6; Doi: https://doi.org/10.36255/exonpublications.bioinformatics.2021

Keywords: biomarker discovery; feature selection; gene expression; machine learning; statistical tests

INTRODUCTION

A biomarker is an indicator of a biological state, often in response to an intervention or the stage of a disease. Although biomarkers mostly refer to physiological or physical phenotypes, at the molecular level, a biomarker can indicate disease-associated molecular changes and may be useful in disease diagnosis (1, 2), various infections (3), neurological diseases (4), and for defining therapeutic targets (3). In toxicological studies, biomarkers are often used to define a set of differentially expressed genes or proteins in a toxic exposure or chemical risk assessment study (5–11). Data from various omics techniques, including transcriptomics, proteomics, and metabolomics, as well as epigenomics, are useful starting points for a biomarker discovery study (10, 12–15). In this chapter, we focus on the informative genes that can generally be used to distinguish samples from different groups, which can be normal or tumor tissues from human patients or tissues of animals that are exposed to toxic chemicals and their solvent controls, using gene expression data. Among the technologies for whole transcriptome gene expression profiling, DNA microarray and RNA sequencing (RNA-Seq) are the most popular (16).

On the methodology aspect, differential gene expression analysis has been the mainstream for its simplicity and interpretability. By comparing the mean expression values of different groups, we can measure the magnitude of difference between the groups, expressed as a fold change (FC), but it is important not to ignore the variance within each group. The genes of highly reproducible but comparably low difference in expression values are missed by looking solely at the FC (17). A statistical hypothesis test is usually applied, such as Student's *t*-test, which considers both the difference between two groups' mean values and the variability within each group. A *p*-value, which is the probability of obtaining an experimental result at least as extreme as the one observed under the null hypothesis, can be obtained from this kind of statistical tests. But such statistical tests usually require specific distributional assumptions; for example, the Student's *t*-test is applicable if the values are normally distributed, which is rarely the case for gene expression data (17). In recent years, more and more concerns and debates about misuse of *p*-value have arisen (18–23). The choice of thresholds for FC and *p*-value can also significantly alter the interpretation of results (24).

In recent years, machine learning has been widely applied in biomarker discovery (3, 25–28). Machine learning applies mathematical approaches to train a model to learn from data for a particular task (29). The relevant machine learning techniques for biomarker discovery are classification and feature selection. Classification is a form of supervised learning where the algorithm is fed with labeled samples each represented by a set of features. The task is to learn a function that can predict the label of a sample from its features. In our case, the labels correspond to the different groups, and the features are the gene expression profiles. As in the case of gene expression data, the number of genes can be tens of

thousands (5). Feature selection is usually applied prior to classification or during classification, to remove noise or non-informative features to train a more precise and robust classifier (30, 31). Feature selection methods can generally be divided into three groups: (i) filter methods that select the features based on their correlation with the sample labels and are therefore independent of the classification procedure; (ii) wrapper methods which use an objective function (usually classification accuracy) to assess the importance of features, and (iii) embedded methods which are incorporated in the classifiers (32, 33). Since the selected features are informative in distinguishing samples from different groups, they can therefore also be regarded as biomarkers.

EVALUATION OF A BIOMARKER DISCOVERY METHOD

Several biomarker discovery methods have been proposed in the fast-developing machine learning field. A reasonable evaluation metric is necessary to choose the most appropriate biomarker discovery method. Two aspects have been addressed when talking about the performance of a biomarker discovery method: its stability and its ability to improve a classifier's prediction accuracy (33–35). Another more direct way to assess performance is to look at the selected gene list given *a priori* knowledge of well-known biomarker sets which can be regarded as "gold standard" (36).

"Gold standard" gene sets

If *a priori* knowledge is available, such as the common gene mutations for breast cancer (37) or the common gene fusions for prostate cancer (38), at least conceptually, the relevant genes can be regarded as the true biomarker genes. In this case, evaluation of a biomarker discovery method becomes quite straightforward by simply comparing the selected gene set to the established "gold standard". But establishing a high-quality "gold standard" becomes crucial to obtain both high precision (as many genes as possible are true biomarkers in the selected gene list) and sensitivity (as many true biomarkers as possible are selected from the whole gene list). To evaluate multiple RNA-Seq analysis workflows (including differential expression analysis), Williams *et al*. prepared a reference gene set based on results from four previous independent microarray and BeadChip studies (39). To reduce bias from one single statistical method, they employed both significance analysis of microarrays (SAM) (40) and limma (41, 42) and used the genes at the intersection of the two methods as the final reference. The resulting reference set was later used as "gold standard" in other studies to assess the performance of RNA-Seq analysis workflows or differential expression analysis methods (43, 44).

Stability

Ideally, the biomarkers should reflect the characteristics of the disease or exposure and be applicable to any sample in the data set. Thus, the biomarker discovery method should select a consistent set of genes disregarding minor changes in

the samples. However, in reality, due to differences between the samples, a biomarker discovery method will select different genes. The robustness of selecting similar gene sets even when the input data varies is called the stability of a method. The similarity of the selected gene lists can be used to define an evaluation metric reflecting the stability of the method.

Starting with two gene sets, Kalousis *et al*. (45) proposed to use the ratio between the number of genes contained in both sets (intersection) and the number of the set of genes contained in either (union) as the similarity index. Kuncheva *et al*. (46) pointed out that this index has a tendency to increase when there are more genes included in the list, which can encourage false positive results. They proposed to take into account the expected number of genes to be shared between the two sets as a modified index to solve that problem.

When it comes to a collection of gene sets, the similarity between them can be calculated by averaging all pairwise similarity indices (46). However, those similarity indices require that gene numbers in all gene sets are the same. Davis *et al*. (47) proposed a more flexible way to calculate similarity which allows various gene set sizes and can also directly calculate the similarity among more than two sets instead of in a pairwise fashion.

Prediction accuracy

A biomarker is an indicator of a biological state in response to an intervention, meaning that it can represent the characteristics of the samples in the intervened group compared with the control group. Compared with using the whole gene list to train a classifier that can distinguish the samples from different groups, training a classifier using biomarkers that already include the most distinctive information should give a comparative prediction performance or even a better one, since non-related and noisy genes can reduce the predictive ability of a classifier. Using several selected gene sets (potential biomarkers) to train classifiers, the prediction accuracy can reflect the quality of the corresponding gene set. A confusion matrix (48) is often used to evaluate the prediction performance of a classifier. Based on that, some evaluation measures such as Recall, Precision, area under a receiver operating characteristics curve, and so on, have been proposed to measure different performance aspects of a classifier (49).

COMPARISON OF BIOMARKER DISCOVERY METHODS

In the case where a well-established "gold standard" gene set is available, a simple comparison of the selected gene list to the reference list can assess the biomarker discovery method in question. But in most cases, such a true biomarker list is not available.

Before looking at the stability and prediction accuracy, which requires greater effort, a simple look at the gene list can still give some hints on the performance of the methods. Comparing the selected gene sets from multiple methods can shed some light on the exploration of the candidate methods, when the absolute performance is not of the highest concern. Blanco *et al*. (50) compared the genes identified as most relevant for discriminating sick and healthy patients as

produced by two different machine learning methods, random forest (51) and generalized linear models (52), and one classical gene expression analysis approach, edgeR (53). They found that random forest and edgeR tend to select similar gene sets compared with generalized linear models.

When the "gold standard" biomarker list is not available, and one still wants to assess the performance of a biomarker discovery method or compare multiple to select the best one for their study, stability and prediction accuracy can be used as evaluation metrics.

For a long time, improving prediction accuracy has been the focus of biomarker discovery methods. Lyons-Weiler *et al*. combined statistical tests with classification (17). They chose the threshold for FC and *p*-value which could help to achieve the highest classification accuracy. Comparing the *F*-score algorithm (from Support Vector Machines (SVM) (54)) with three popular differential expression analysis methods (limma, edgeR, DESeq (55)), Liang *et al*. (56) found that *F*-score algorithm obtained the best predictive performance when training an SVM classifier to predict stages of human embryonic development using single-cell RNA-Seq data. Schirra *et al*. evaluated the feature selection/classifier combinations that lead to an improved classification performance, and preferred filter methods when comparable prediction accuracy can be obtained for their higher interpretability (57).

Stability of biomarker discovery has gained more and more attention in recent years (32, 58, 59). A more complete evaluation of a biomarker discovery method should address both prediction accuracy and stability (33–35). In a previous study (33), on those two aspects, we compared the performance of both traditional statistical tests and machine learning methods: SAM, minimum redundancy maximum relevance (mRMR) (60), and characteristic direction (GeoDE) (36) on multiple datasets. We found that no single method outperforms the others on these two aspects across all tested datasets.

ENSEMBLE OF MULTIPLE METHODS

Since it is hard to tell which is the best one, another solution is to combine the potential methods. There are already studies showing that an ensemble of multiple feature selection methods can obtain a very satisfactory performance regarding both stability and prediction accuracy. The ensemble gene set can therefore be regarded as the final biomarker gene set (Figure 1).

Van IJzendoorn *et al*. combined statistical tests with machine learning techniques (61). On top of the significantly differentially expressed genes (adjusted *p*-value < 0.05), they applied random forest to select the most informative genes. By employing the ensemble feature selection concept, multiple biomarker discovery methods can be combined to take advantage of the strengths and overcome the weaknesses of the individual methods (62, 63). This approach is called function perturbation (32, 62). Similar to this logic, data perturbation refers to approaches applying one method on several data subsets generated from the original data set (for example using bootstrap (64)), and combining the results (58, 63, 65), an approach that has been shown to be able to improve the stability of the biomarker discovery method.

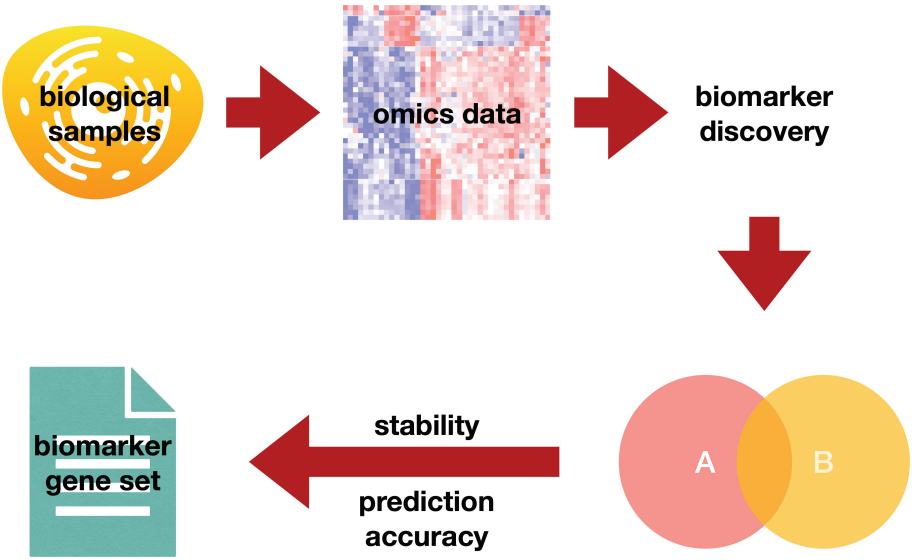

Figure 1. An illustration of using ensemble gene sets from multiple methods as the biomarker gene set. Omics data collected from biological samples are fed into multiple biomarker discovery methods which results in several gene sets (for example, A and B). Based on stability and prediction accuracy, the results from satisfactory methods are integrated into the final biomarker gene set.

To take advantage of both data perturbation and function perturbation, we proposed to combine both of them (66) (Figure 2). In the phase of data perturbation, the stability of each method is calculated, and in the phase of function perturbation, when combining the results from multiple methods, their stabilities are used as their weights, so as to achieve the most robust final result. Testing on six microarray data sets from cancer studies, we found that the proposed framework achieved both high stability and prediction accuracy compared with the individual methods and the pure function perturbation.

CONCLUSION

In this chapter, we discussed biomarker discovery using gene expression data of the samples from different groups, usually a control group under normal biological status and a treated group with intervention or disease. The biomarker genes are therefore the responders to the intervention. Traditional statistical tests have been widely used to identify the differentially expressed genes as biomarkers for their simplicity and high interpretability. Such statistical tests are based on a hypothesis that the genes are independent of each other. This is not the case in a normal biological setting, since genes usually work together composing pathways

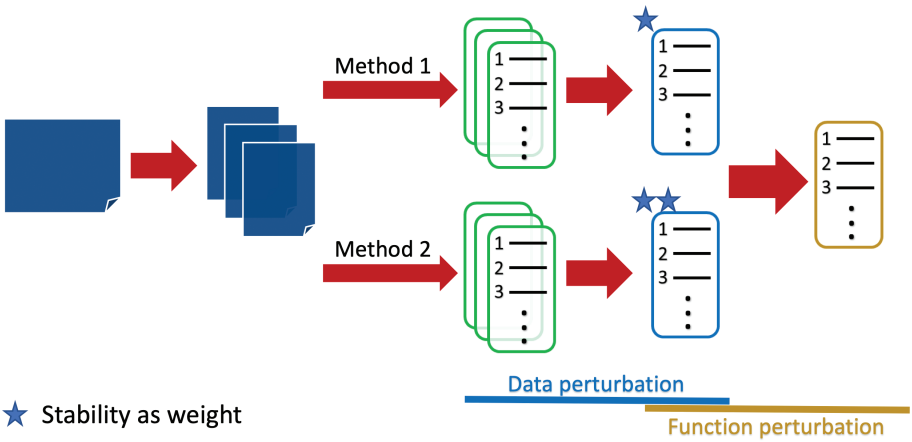

Figure 2. Combination of both data perturbation and function perturbation. The original dataset is subsampled into several sub-datasets. The genes are ranked based on each of them using different methods. In the data perturbation phase, the ranked gene lists are integrated into one ranked list and meanwhile, the stability of each method is calculated. In the phase of function perturbation, the results from different methods are combined using methods' stabilities as weights.

and networks (3), resulting in a highly correlated data set. Most of the statistical tests also require some specific distributional assumptions which cannot always be satisfied, especially when the biological replicates are quite limited. The misuse of FC and *p*-value and the choice of their threshold have also been debated in recent years.

Machine learning techniques, such as feature selection, have been applied with increasing frequency in biomarker discovery. Feature selection usually has fewer required assumptions compared with statistical tests. Many of them can take the interaction between genes and their joint power into consideration. The genes that are weak biomarkers by themselves but have a strong joint power can therefore be identified.

Another machine learning technique, classification, is also useful in biomarker discovery. Classification is not directly used to identify biomarkers but can be used to assess potential biomarkers selected by feature selection methods or statistical tests, since true biomarkers carry the characteristics of samples from the treated group compared with control group or vice versa and should therefore be informative in classifying the samples from different groups. The ability to improve a classifier's prediction accuracy of a biomarker discovery method is widely used as an evaluation metric of candidate methods. We have seen that the choice of classification algorithm can highly affect the evaluation conclusion of the biomarker discovery methods (33), and using SVM to assess the performance of a feature selection method implemented in its own package together with other methods is unfair (56).

Besides prediction accuracy, a biomarker discovery method's stability has gained more attention in recent years. A good biomarker discovery method should provide a consistent biomarker list with some variance in the training samples, since the true biomarkers are intervention dependent (such as a disease or a toxicant exposure) and should be independent of the samples. There are many ways to calculate stability, but some of them tend to give a higher stability when more genes are included in the lists (46) and that is unfair for the methods that are stricter with redundant genes. Instead of looking only at the original gene list, Dessì *et al*. proposed to compare the lists in functional terms based on the molecular function Gene Ontology annotations, which has greater biological significance (35).

Many alternative approaches for improving a method's performance based on the aforementioned aspects have been proposed. One of them is feature selection ensemble, which combines the results of multiple biomarker discovery methods to take advantage of their strengths. It also solves the problem of having to choose the most appropriate method for a particular dataset since the performance of a method usually varies a lot across different datasets.

Besides assessing a biomarker discovery method on prediction accuracy and stability, one can also simply compare the candidate marker genes to a reference biomarker list, if such a "gold standard" exists. It is however difficult to be sure that the *a priori* knowledge is adequate, and that the list is complete and clear of false positives. Establishing such a reference list becomes extremely critical. Williams *et al*. applied two well-recognized methods (SAM and limma) on four independent datasets and used the intersected genes as reference (39). Biological *a priori* knowledge can also help in constructing such a reference list. Clark *et al*. made use of the relationship between differential STAT3 binding and differential gene expression in two subtypes of diffuse large B-cell lymphoma (DLBCL): germinal center B-cell-like (GCB) and activated B-cell-like (ABC) (36).

Biomarker discovery is a fast-growing field with many new ideas continuously being proposed. So far none are perfect, considering that the method is data dependent and no universal agreement on the evaluation of a method's performance has been established. However, devoted efforts are obviously enhancing progress in this field, which has a huge potential for providing a better understanding of disease diagnosis, prevention, and therapy, and for risk assessment of chemical toxicity.

Acknowledgement: This work was supported by the Research Council of Norway to the Digital Life Norway project dCod 1.0: decoding the systems toxicology of Atlantic cod (project no. 248840).

Conflict of interest: The authors declare no potential conflict of interest with respect to research, authorship and/or publication of this chapter.

REFERENCES

1. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747–52. https://doi.org/10.1038/35021093
2. Løvf M, Zhao S, Axcrona U, Johannessen B, Bakken AC, Carm KT, et al. Multifocal primary prostate cancer exhibits high degree of genomic heterogeneity. Eur Urol. 2019;75(3):498–505. https://doi.org/10.1016/j.eururo.2018.08.009
3. Dix A, Vlaic S, Guthke R, Linde J. Use of systems biology to decipher host-pathogen interaction networks and predict biomarkers. Clin Microbiol Infect. 2016;22(7):600–6. https://doi.org/10.1016/j.cmi.2016.04.014
4. Dunckley T, Coon KD, Stephan DA. Discovery and development of biomarkers of neurological disease. Drug Discov Today. 2005;10(5):326–34. https://doi.org/10.1016/S1359-6446(04)03353-7
5. Yadetie F, Zhang X, Hanna EM, Aranguren-Abadía L, Eide M, Blaser N, et al. RNA-Seq analysis of transcriptome responses in Atlantic cod (Gadus morhua) precision-cut liver slices exposed to benzo[a]pyrene and 17α-ethynylestradiol. Aquat Toxicol. 2018;201:174–86. https://doi.org/10.1016/j.aquatox.2018.06.003
6. Khan EA, Zhang X, Hanna EM, Bartosova Z, Yadetie F, Jonassen I, et al. Quantitative transcriptomics, and lipidomics in evaluating ovarian developmental effects in Atlantic cod (Gadus morhua) caged at a capped marine waste disposal site. Environ Res. 2020;189:109906. https://doi.org/10.1016/j.envres.2020.109906
7. Khan EA, Zhang X, Hanna EM, Yadetie F, Jonassen I, Goksøyr A, et al. Application of quantitative transcriptomics in evaluating the ex vivo effects of per- and polyfluoroalkyl substances on Atlantic cod (Gadus morhua) ovarian physiology. Sci Total Environ. 2020;142904. https://doi.org/10.1016/j.scitotenv.2020.142904
8. Goksøyr A, Beyer J, Egaas E, Grøsvik BE, Hylland K, Sandvik M, et al. Biomarker responses in flounder (Platichthys flesus) and their use in pollution monitoring. Mar Pollut Bull. 1996;33(1–6):36–45. https://doi.org/10.1016/S0025-326X(96)00131-2
9. van der Oost R, Beyer J, Vermeulen NPE. Fish bioaccumulation and biomarkers in environmental risk assessment: a review. Environ Toxicol Pharmacol. 2003;13(2):57–149. https://doi.org/10.1016/S1382-6689(02)00126-6
10. Brooks BW, Sabo-Attwood T, Choi K, Kim S, Kostal J, LaLone CA, et al. Toxicology advances for 21st century chemical pollution. One Earth. 2020;2(4):312–6. https://doi.org/10.1016/j.oneear.2020.04.007
11. Hanna EM, Zhang X, Eide M, Fallahi S, Furmanek T, Yadetie F, et al. ReCodLiver0.9: Overcoming Challenges in Genome-Scale Metabolic Reconstruction of a Non-model Species. Front Mol Biosci. 2020;7. https://doi.org/10.3389/fmolb.2020.591406
12. Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. Nat Biotechnol. 2006;24(8):971–83. https://doi.org/10.1038/nbt1235
13. Hu Z-Z, Huang H, Wu CH, Jung M, Dritschilo A, Riegel AT, et al. Omics-based molecular target and biomarker identification. Methods Mol Biol. 2011;719:547–71. https://doi.org/10.1007/978-1-61779-027-0_26
14. Martins C, Dreij K, Costa PM. The State-of-the Art of Environmental Toxicogenomics: Challenges and Perspectives of "Omics" Approaches Directed to Toxicant Mixtures. Int J Environ Res Public Health. 2019;16(23). https://doi.org/10.3390/ijerph16234718
15. Dirks RAM, Stunnenberg HG, Marks H. Genome-wide epigenomic profiling for biomarker discovery. Clin Epigenetics. 2016;8:122. https://doi.org/10.1186/s13148-016-0284-4
16. Yang X, Kui L, Tang M, Li D, Wei K, Chen W, et al. High-Throughput Transcriptome Profiling in Drug and Biomarker Discovery. Front Genet. 2020;11:19. https://doi.org/10.3389/fgene.2020.00019
17. Lyons-Weiler J, Patel S, Bhattacharya S. A classification-based machine learning approach for the analysis of genome-wide expression data. Genome Res. 2003;13(3):503–12. https://doi.org/10.1101/gr.104003

18. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods. 2015;12(3):179–85. https://doi.org/10.1038/nmeth.3288
19. Ioannidis JPA. The proposal to lower P value thresholds to .005. JAMA. 2018;319(14):1429–30. https://doi.org/10.1001/jama.2018.1536
20. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567(7748):305–7. https://doi.org/10.1038/d41586-019-00857-9
21. Kraemer HC. Is it time to ban the P value? JAMA Psychiatry. 2019;76(12):1219–20. https://doi.org/10.1001/jamapsychiatry.2019.1965
22. Krueger JI, Heck PR. Putting the P-Value in its Place. The American Statistician. 2019;73(sup1):122–8. https://doi.org/10.1080/00031305.2018.1470033
23. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p < 0.05". Am Stat. 2019;73(sup1):1–19. https://doi.org/10.1080/00031305.2019.1583913
24. Dalman MR, Deeter A, Nimishakavi G, Duan Z-H. Fold change and p-value cutoffs significantly alter microarray interpretations. BMC Bioinformatics. 2012;13 Suppl 2:S11. https://doi.org/10.1186/1471-2105-13-S2-S11
25. Hou Q, Bing Z-T, Hu C, Li M-Y, Yang K-H, Mo Z, et al. RankProd Combined with Genetic Algorithm Optimized Artificial Neural Network Establishes a Diagnostic and Prognostic Prediction Model that Revealed C1QTNF3 as a Biomarker for Prostate Cancer. EBioMedicine. 2018;32:234–44. https://doi.org/10.1016/j.ebiom.2018.05.010
26. Mamoshina P, Volosnikova M, Ozerov IV, Putin E, Skibina E, Cortese F, et al. Machine Learning on Human Muscle Transcriptomic Data for Biomarker Discovery and Tissue-Specific Drug Target Identification. Front Genet. 2018;9:242. https://doi.org/10.3389/fgene.2018.00242
27. Torres R, Judson-Torres RL. Research techniques made simple: feature selection for biomarker discovery. J Invest Dermatol. 2019;139(10):2068-2074.e1. https://doi.org/10.1016/j.jid.2019.07.682
28. Xie Y, Meng W-Y, Li R-Z, Wang Y-W, Qian X, Chan C, et al. Early lung cancer diagnostic biomarker discovery by machine learning methods. Transl Oncol. 2020;14(1):100907. https://doi.org/10.1016/j.tranon.2020.100907
29. Mitchell TM. Machine Learning. New York: McGraw-Hill; 1997. 414 p.
30. Tang J, Alelyani S, Liu H. Data classification: Algorithms and applications. New York: CRC Press; 2014. Chapter 2, Feature selection for classification: A review; p.37–64. https://doi.org/10.1201/b17320
31. Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine learning research. 2003;3:1157–82.
32. He Z, Yu W. Stable feature selection for biomarker discovery. Comput Biol Chem. 2010;34(4):215–25. https://doi.org/10.1016/j.compbiolchem.2010.07.002
33. Zhang X, Jonassen I. A Comparative Analysis of Feature Selection Methods for Biomarker Discovery in Study of Toxicant-Treated Atlantic Cod (Gadus Morhua) Liver. In: Bach K, Ruocco M, editors. Nordic artificial intelligence research and development: third symposium of the Norwegian AI society, NAIS 2019, Trondheim, Norway, May 27-28, 2019, proceedings. Cham: Springer International Publishing; 2019. p. 114–23. https://doi.org/10.1007/978-3-030-35664-4_11
34. Yang F, Mao KZ. Robust feature selection for microarray data based on multicriterion fusion. IEEE/ACM Trans Comput Biol Bioinform. 2011;8(4):1080–92. https://doi.org/10.1109/TCBB.2010.103
35. Dessì N, Pascariello E, Pes B. A comparative analysis of biomarker selection techniques. Biomed Res Int. 2013;2013:387673. https://doi.org/10.1155/2013/387673
36. Clark NR, Hu KS, Feldmann AS, Kou Y, Chen EY, Duan Q, et al. The characteristic direction: a geometrical approach to identify differentially expressed genes. BMC Bioinformatics. 2014;15:79. https://doi.org/10.1186/1471-2105-15-79
37. Rebbeck TR, Couch FJ, Kant J, Calzone K, DeShano M, Peng Y, et al. Genetic heterogeneity in hereditary breast cancer: role of BRCA1 and BRCA2. Am J Hum Genet. 1996;59(3):547–53.
38. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun X-W, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005;310(5748):644–8. https://doi.org/10.1126/science.1117679

39. Williams CR, Baccarella A, Parrish JZ, Kim CC. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. BMC Bioinformatics. 2017;18(1):38. https://doi.org/10.1186/s12859-016-1457-z
40. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA. 2001;98(9):5116–21. https://doi.org/10.1073/pnas.091062498
41. Kim CC, Falkow S. Significance analysis of lexical bias in microarray data. BMC Bioinformatics. 2003;4:12. https://doi.org/10.1186/1471-2105-4-12
42. Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, editors. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. New York: Springer; 2005. p. 397–420. https://doi.org/10.1007/0-387-29362-0_23
43. Zhang X, Jonassen I. RASflow: an RNA-Seq analysis workflow with Snakemake. BMC Bioinformatics. 2020;21(1):110. https://doi.org/10.1186/s12859-020-3433-x
44. Quinn TP, Crowley TM, Richardson MF. Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. BMC Bioinformatics. 2018;19(1):274. https://doi.org/10.1186/s12859-018-2261-8
45. Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms. Fifth IEEE International Conference on Data Mining (ICDM'05). IEEE; 2005. p. 218–25. https://doi.org/10.1109/ICDM.2005.135
46. Kuncheva LI. A stability index for feature selection. 25th IASTED International Multi-Conference: Artificial Intelligence and Applications. ACTA Press; 2007. p. 390–395.
47. Davis CA, Gerick F, Hintermair V, Friedel CC, Fundel K, Küffner R, et al. Reliable gene signatures for microarray classification: assessment of stability and performance. Bioinformatics. 2006;22(19):2356–63. https://doi.org/10.1093/bioinformatics/btl400
48. Stehman SV. Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment. 1997;62(1):77–89. https://doi.org/10.1016/S0034-4257(97)00083-7
49. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861–74. https://doi.org/10.1016/j.patrec.2005.10.010
50. Liñares Blanco J, Gestal M, Dorado J, Fernandez-Lozano C. Differential Gene Expression Analysis of RNA-seq Data Using Machine Learning for Cancer Research. In: Tsihrintzis GA, Virvou M, Sakkopoulos E, Jain LC, editors. Machine learning paradigms: applications of learning and analytics in intelligent systems. Cham: Springer International Publishing; 2019. p.27–65. https://doi.org/10.1007/978-3-030-15628-2_3
51. Breiman L. Random forests. Machine learning. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324
52. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1):1–22. https://doi.org/10.18637/jss.v033.i01
53. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616
54. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97. https://doi.org/10.1007/BF00994018
55. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106. https://doi.org/10.1186/gb-2010-11-10-r106
56. Liang P, Yang W, Chen X, Long C, Zheng L, Li H, et al. Machine Learning of Single-Cell Transcriptome Highly Identifies mRNA Signature by Comparing F-Score Selection with DGE Analysis. Mol Ther Nucleic Acids. 2020;20:155–63. https://doi.org/10.1016/j.omtn.2020.02.004
57. Schirra L-R, Lausser L, Kestler HA. Selection stability as a means of biomarker discovery in classification. In: Wilhelm AFX, Kestler HA, editors. Analysis of large and complex data. Cham: Springer International Publishing; 2016. p.79–89. https://doi.org/10.1007/978-3-319-25226-1_7
58. Pes B, Dessì N, Angioni M. Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data. Information Fusion. 2017;35:132–47. https://doi.org/10.1016/j.inffus.2016.10.001

59. Bolón-Canedo V, Alonso-Betanzos A. Ensembles for feature selection: A review and future trends. Information Fusion. 2019;52:1–12. https://doi.org/10.1016/j.inffus.2018.11.008
60. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27(8):1226–38. https://doi.org/10.1109/TPAMI.2005.159
61. van IJzendoorn DGP, Szuhai K, Briaire-de Bruijn IH, Kostine M, Kuijjer ML, Bovée JVMG. Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. PLoS Comput Biol. 2019;15(2):e1006826. https://doi.org/10.1371/journal.pcbi.1006826
62. Ben Brahim A, Limam M. Robust ensemble feature selection for high dimensional data sets. 2013 International Conference on High Performance Computing & Simulation (HPCS). IEEE; 2013. p. 151–7. https://doi.org/10.1109/HPCSim.2013.6641406
63. Seijo-Pardo B, Porto-Díaz I, Bolón-Canedo V, Alonso-Betanzos A. Ensemble feature selection: Homogeneous and heterogeneous approaches. Knowledge-Based Systems. 2017;118:124–39. https://doi.org/10.1016/j.knosys.2016.11.017
64. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman and Hall; 1994. https://doi.org/10.1201/9780429246593
65. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics. 2010;26(3):392–8. https://doi.org/10.1093/bioinformatics/btp630
66. Zhang X, Jonassen I. An ensemble feature selection framework integrating stability. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2019. p. 2792–8. https://doi.org/10.1109/BIBM47256.2019.8983310