Informatics in Medicine Unlocked 24 (2021) 100595

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Identification of cancer related genes using feature selection and association rule mining

Consolata Gakii [a,b,*], Richard Rimiru [a]

[a] School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology, P.O Box 62000, Nairobi, Kenya
[b] Department of Mathematics, Computing and Information Technology, University of Embu, Embu, Kenya

ARTICLE INFO

Keywords:
Feature selection
Discretization
Association rule mining
Coexpression network

ABSTRACT

High throughput sequencing generates large volumes of high dimensional data. Identifying informative features from the generated big data is always a challenge. Feature selection reduces complex data into a smaller number of variables while preserving the information as much as possible. In this study, we used DaMiRseq, DESeq2, edgeR and Limma + voom to identify differentially expressed genes in 79 small cell lung cancer (sclc) and 7 normal controls. A gene network was used to identify any coexpressed genes. Association rule mining was used to identify any association between connected genes in the network. Limma + voom identified the highest number of differentially expressed genes. However, 81 genes were common in the four differential gene expression analysis methods used. After filtering out all nodes with a degree less than 5, the final network had 43 nodes and 63 edges. Association rule mining on the coexpressed genes generated 263 rules. Genes that were common in the rules were: SLC34A2, CAV2, EPAS1, CTSH, AQP1 and LRRK2. These genes have been associated with various types of cancer. Therefore, feature selection using differential gene expression analysis, co-expression networks and association rule mining could help infer relationships among genes and their possibility of having a shared biological function.

1. Introduction

High throughput sequencing technologies generate large volumes of data and this effectively ushers life sciences into the big data realm [1,2]. However, data generated using these technologies is oftentimes noisy or high-dimensional and therefore requires several preprocessing steps for its computational analysis [3–5]. As the dimensionality increases, the volume of data required to provide meaningful insights also grows exponentially. Bellman [6] described this phenomenon as a curse of dimensionality. Therefore, identification of relevant information from big data is always a challenge. Feature selection [7] is an approach that can be used to reduce complex data into a smaller number of variables (or features). Redundant as well as non-informative features are discarded while relevant information is preserved as much as possible [8]. This approach helps avoid the curse of dimensionality [9]. Methods applied in feature selection can be categorized as filters, wrappers or embedded [10,11]. provide comprehensive reviews on the tools and techniques used in feature selection. Most important is that a suitable feature selection algorithm should be able to identify only few informative features that exbibit some form of grouping/correlation [12].

Different feature selection (FS) techniques have been used in various biological investigations [13]. used different feature selection methods to identify genetic heritability of three traits (height, high density lipoprotein cholesterol and obesity) that are of human health importance [14]. applied Random Forest classification in combination with incremental feature selection to identify 29 novel biomarkers for pancreatic cancer from a salivary transcriptome. A novel method of feature selection whereby medical images were converted to quantitative imaging features was used by Ref. [15] to identify informative features in nasopharyngeal carcinoma images [16]. developed an R package containing several methods for performing complex-based feature selection from proteomics expression data. In the analysis of whole-genome expression data [17], used support vector machines to classify and identify 128 genes that can be used to predict the outcome of imatinib resistance in chronic myeloid leukemia. In an earlier study [18], used recursive feature elimination in combination with support vector machines to compare gene expression in myeloid leukemia. Their study

* Corresponding author. School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology, P.O Box 62000, Nairobi, Kenya.
E-mail address: gakii.consolata@embuni.ac.ke (C. Gakii).

https://doi.org/10.1016/j.imu.2021.100595
Received 14 February 2021; Received in revised form 1 May 2021; Accepted 1 May 2021
Available online 9 May 2021

revealed 114 differentially expressed genes whereby 34 of them had significant transcriptional changes between the two conditions under study.

Feature selection has also been applied on whole-genome sequencing data to understand heritable genomic methylation patterns in mammals using SVM with recursive feature elimination [19]. In cancer studies [20], selected important tubulointerstitial fibrosis features from renal fibrosis microarray data using iterative Random Forest [21]. used SVM and LR model to study deleterious mutations and neutral mutations using data downloaded from Uniprot database. In proteomics [22], proposed gene fuzzy scoring as a way of removing noise from low-ranking features as a way of boosting sensitivity in feature selection [23]. used the random forest approach to identify tissue-specific biomarkers that can be used to predict origin of tumors.

RNA-Seq experiments are designed to identify changes in gene expression patterns when comparing two or more conditions such as diseased versus healthy state [24–26]. From a data mining perspective, discovery of genes showing differential expression is a form of feature selection and dimensionality reduction because it helps identify the most significant subset of features (genes). Therefore, differential gene expression analysis tries to answer the question whether observed differences in gene expression is statistically significant [27,28]. Genes of interest are identified and assigned a p-value in relation to a statistical test of choice [25]. Methods such as DESeq and edgeR that are used in performing differential expression analysis model read counts for each gene using the negative binomial (NB) distribution [27,28]. Genes that are differentially expressed are then used for downstream analysis. Any patterns in gene expression can be visualized using coexpression networks.

Weighted Gene Co-expression Network Analysis (WGCNA) is one of the algorithms used in analyzing gene co-expression networks. It constructs a network based on the pairwise correlations between genes expression levels [29]. The algorithm assumes a scale-free network following a power law distribution and this reflects several biological phenomena [30]. Another dimensionality reduction technique used in big data analytics is data discretization, a process that transforms continuous data into discrete values [31,32]. Although the technique is mostly used in computer science and statistics, it has been adopted as a preprocessing step in biological data analysis. Thereafter, association rule mining (ARM) is used to discover possible associations in the discretized data. Agrawal et al. [33], introduced this algorithm as a market basket analysis tool to mine frequent itemsets. Datasets are organized in the form of a transaction t $\boldsymbol{T} \in \boldsymbol{D}$ that contains an itemset $X \subseteq I$ if $x \subseteq T$. In bioinformatics this concept has been used in mining gene expression data [34], inferring interactions among transcription factors in user-selected genomic regions [35], identification of malignant mesothelioma risk factors [36], classification of cancer gene expression data [37], extraction of complex markers from genomic, epigenomic and proteomic data [38] as well as gene ontology analysis [39]. However, interpreting gene expression data requires an understanding of the grouping structure since most biological features tend to group in a certain way based on e.g., shared metabolic pathway, methylation profiles or proximity in the genome [40]. Therefore, feature selection, network analysis or association rule mining applied individually may miss some of the underlying biological patterns or associations.

The aim of this study was to apply differential gene expression analysis and gene coexpression network analysis to reduce the dimensionality of RNAseq data and thereby facilitating the identification of informative features. Since biological features tend to associate in a certain way, association rule mining was used to generate rules and identify meaningful associations between various features (genes). Using this novel approach on the cancer dataset, genes that can be used in cancer diagnosis were identified.

The main contribution of this work is:

- A framework that combines differential expression analysis, gene coexpression network analysis and association rule mining in identification of informative features from complex data.
- The proposed approach eliminates any form of bias attributable to any of the tools and also identifies features that would be missed when using any of the tools as a standalone.

The next sections of the manuscript are organized as follows: section 2 describes the methodology used to address the objectives, Section 3 presents the outputs from each task while Section 4 puts the findings in context of related research. Lastly, section 5 provides a conclusion and recommendation.

2. Methodology

The proposed methodology has six phases as shown in the workflow (Fig. 1). These are data mining and preprocessing, differential expression analysis, construction of gene coexpression network to filter unconnected genes, discretization to convert continuous values to discrete, generation of frequent itemset and gene enrichment analysis.

2.1. Preprocessing of RNASeq data

Raw data was downloaded from Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/). The data set (accession ID GSE60052) comprised of 79 small cell lung cancer (sclc) and 7 normal controls [41]. This data was generated by sequencing the whole exome and transcriptome of primary tumors from Chinese SCLC patients. Attributes describing the dataset are summarized in supplementary Table S1. The raw data was pre-processed using GEO RNA-seq Experiments Interactive Navigator (GREIN) web platform as described by Ref. [42]. Briefly, GREIN uses NCBI SRA toolkit to generate FASTQ files using the run files downloaded from the sequence reads archive (SRA) as input. Quality control was done, and QC reports are generated using FastQC [43]. Any adapter sequences on the reads were removed using Trimmomatic [44]. Transcript abundances were then quantified by

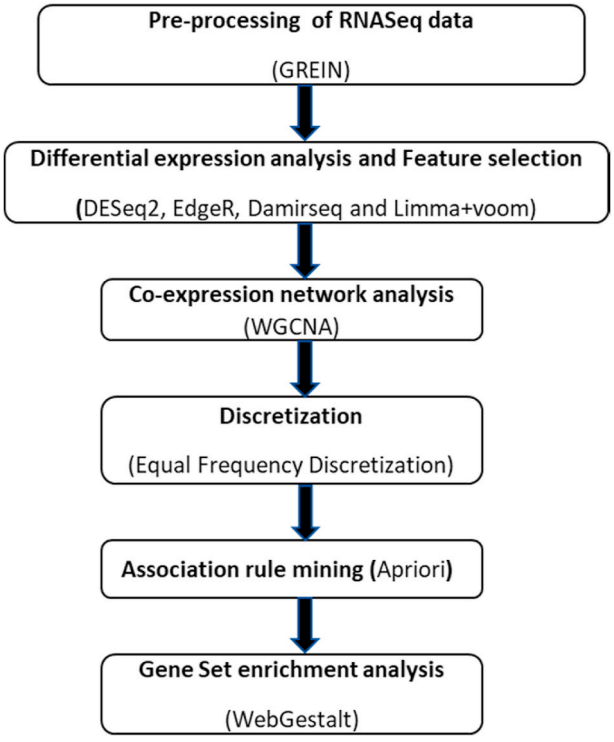

Fig. 1. Workflow of the steps used in data analysis.

mapping reads to a reference transcriptome using Salmon [45]. This allowed estimation of the TPM for each transcript in each of the data sets. The gene annotation used was for *Homo sapiens* (GRCh38). The final output from the pipeline was a count table of genes and metadata for each sample that we used for downstream analysis.

2.2. Differential expression analysis and feature selection

To perform differential expression analysis and feature selection, we used DaMiRseq [46], DESeq2 [47], edgeR [48] and Limma + voom [49] to identify any genes that could be differentially expressed. DaMiRseq [46] and DESeq2 [47] use negative binomial distribution and a shrinkage estimator for the distribution's variance, $K_{ij} \sim NB(\mu_{ij}, \sigma_{ij}^2$. Limma uses the voom function to convert the mean variance to precision weights and using a linear model, $E(y_{gi}) = \mu_{gi} = X_i^T \beta_g$. Finally, edgeR [48] uses a Poisson super dispersion model to account for technical and biological variation and then applies the Bayesian empirical method to moderate the degree of over-dispersion against transcripts. For each of the four methods, we used a cutoff p-value of 0.05 for the four methods.

2.3. Gene Co-expression network construction

Differentially expressed genes selected by Limma + voom were used to construct a co-expression networks using WGCNA v 1.69, a package in R [50]. We chose the gene list from Limma + voom because it returned the highest number of differentially expressed genes as compared to other methods. WGCNA calculates a similarity matrix using correlation of all genes $cor(i,j)$. The coexpression matrix is then transformed to an adjacency matrix by raising the coexpression similarity by use of soft thresholding power beta (β). A $\beta = 16$ satisfied the criterion of approximating a scale-free topology of the network. Modules consisting of co-expressed genes were detected based on TOM values using hierarchical clustering and any similar modules were merged. WGCNA identified a final list of 14 distinct putatively co-expressed gene modules. Network visualization and filtering were done using Cytoscape v. 3.8.1 [51].

2.4. Discretization and association rule mining

Genes that did not have any connection in the network were filtered out and the remaining geneset discretized using Equal Frequency Discretization (EFD). EFD is unsupervised discretization used in the absence of any knowledge of the class memberships of the instances. This method works by dividing a continuous attribute A = $\{a_1, a_2, \ldots, a_{n-1}, a_n\}$ into k intervals which include the same number of values. Each interval contains n/kbins where n is the number of values. We defined two bins as described in Gallo et al., [31]. The discretized data was used in identification of frequent itemsets using R package arules ver. v1.6–4 [52]. Three metrics were applied when generating the rules: rule support, confidence, and length. A confidence value of 0.9 and support of value of 0.8 were used with minimum and maximum length of three and five, respectively. Redundant rules were filtered out by selecting only the rules with a lift >2.

2.5. Gene enrichment analysis

Vital biological gene functions were identified by performing enrichment analysis using WEB-based GEneSeTAnaLysis Toolkit (WebGestalt) described by Ref. [53]. Default parameters (FDR < 0.05, 1000 permutations and 20 categories with the outputted leading-edge genes) as described by Ref. [54].

3. Results

3.1. Differential expression analysis

The data used in this study comprised RNAseq from 79 small cell lung cancer (sclc) and 7 normal controls. The highest number of differentially expressed genes (3,920) between tumor and normal samples was identified using Limma + voom. Damirseq on the other hand identified the lowest number of differentially expressed genes (197). EdgeR and Deseq2 identified 3423 and 386 differentially expressed genes respectively (Fig. 2). A set of 81 genes was common in the four differential gene expression analysis methods used (Fig. 2).

3.2. Coexpression network

The 3920 differentially expressed genes identified by Limma + voom were used for WGCNA analysis. At a minimum power value of 17, the squared correlation coefficients reached a value of 0.9 (Fig. 3a). This implied a scale-free network distribution and at that point, the average degree of the coexpression network conforms to small-world network properties [55]. We obtained a hierarchical cluster tree with 14 modules (Fig. 3b) while the resulting gene cluster dendrogram is shown in Fig. 3c. Each color in the cluster diagram represents an independent module of highly correlated genes. The final network had 350 nodes and 580 edges. We then filtered out all nodes with a degree value of less than 5 and generated a final network with 43 nodes and 63 edges (Fig. 3d).

3.3. Discretization and rules generation

A set of 43 genes from the coexpression network were discretized and converted from continuous to discrete values. In the resulting matrix, a value of 1 represents upregulation while 0 is for downregulation. Association rule mining allowed us to identify any underlying association between genes in the normal and tumor samples. After executing the algorithm using a minimum support of 0.8, a confidence of 0.9 and a lift >2, we got a set of 263 rules. A subset of the top rules is shown in Table 1 while the entire set is presented as Supplementary Table S2.

3.4. Enrichment analysis

Gene set enrichment analysis showed that genes involved in negative regulation of proteolysis (TIMP3, AQP1, LRRK2, COL4A3 and LAMP3) were highly enriched (Normalized Enrichment Score of 0.96) (Fig. 4).

The most enriched gene ontology terms were Caveola assembly and nuclear division. Gene CAV1 and CAV2 are involved in several

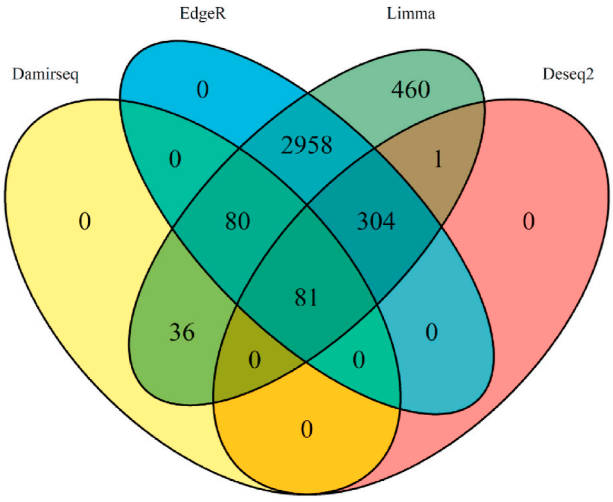

Fig. 2. Genes identified using edgeR, DaMiRseq, Limma + voom and Deseq2.

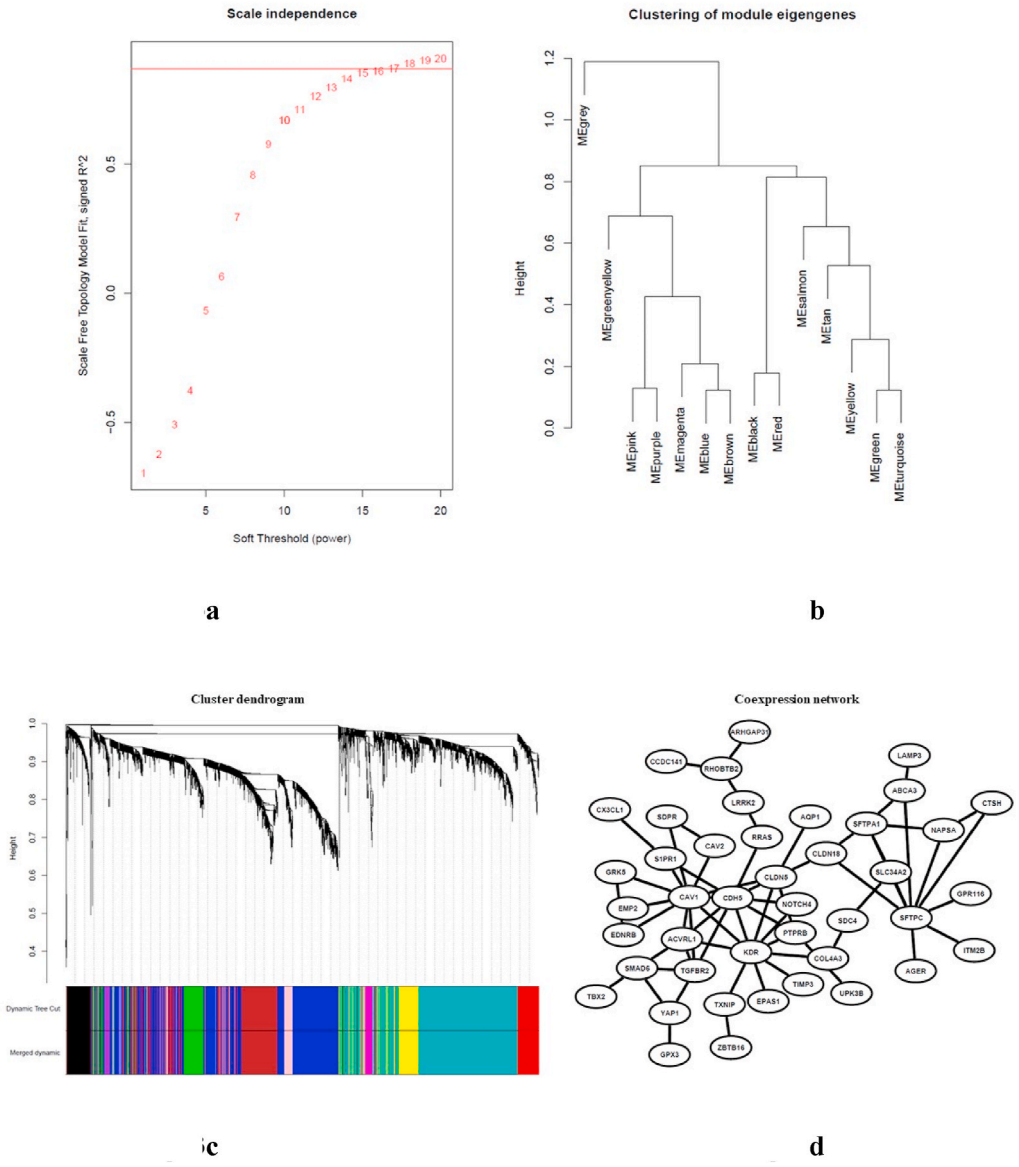

Fig. 3. a) network topology analysis at different soft-thresholding powers; b) Cluster dendrogram of genes based on topological overlap; c) Gene dendrogram with the corresponding module colors; d) Final gene coexpression network after filtering less connected genes. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 1
A set of rules generated using association rule mining.

	Rules	Support	Confidence	Lift
1	{SLC34A2} => {CAV1, SFTPA1, NAPSA, TGFBR2}	0.8	0.9	2.3
2	{CAV2} => {RHOBTB2, CAV1, S1PR1}	0.8	0.9	2.1
3	{EPAS1} => {YAP1, LRRK2, GPX3, AQP1}	0.8	0.9	2.1
4	{EPAS1} => {SMAD6, TGFBR2, LRRK2, AQP1}	0.8	0.9	2.1
5	{CAV2} => {TIMP3, COL4A3, CLDN5, LAMP3}	0.8	0.9	2.1

biological processes (Supplementary Table S3, S4).

4. Discussion

In this study we used a feature selection and association rule mining on lung cancer dataset. Four differential gene expression methods were used and Limma + voom selected the highest number of genes. It has been reported that Limma + voom gives consistent results [56]. In a recent study by Ref. [57], they compared 3 datasets (GSE40275, GSE99316, and GSE60052) and identified a set of 208 common genes that were differentially expressed. However, they only used Limma + voom and this may have been the reason for the high number of identified genes. In our study we combined four different differential expression methods to eliminate any bias that may be a result of any single method. Using this approach, 81 genes were picked by all the methods and these genes are very crucial because whichever differential expression method used, they would be selected. Mining gene expression data for association rule is useful in uncovering gene relationships. For example, rule 1 showed that in 90% of the cases where the gene SLC34A2 was up (highly expressed), all the genes on the right-hand side of the rule were also up. We also identified several genes that were common in the rules and with a support of greater that 80%, which explains their occurrence in the rules. These were the genes SLC34A2, CAV2, EPAS1, CTSH, AQP1 and LRRK2. Caveolin-1 (CAV1) and caveolin 2 (CAV2) are two principal structural proteins involved in vesicular trafficking and signal transduction [58]. However, there are reports that

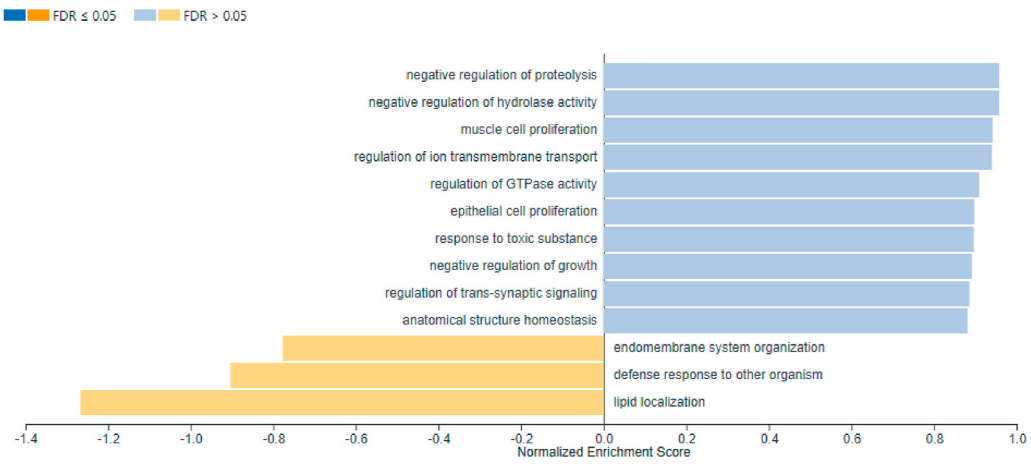

Fig. 4. Genest Enrichment results.

the two genes are involved in various types of cancer as summarized in Supplementary Table S2. The other genes play various regulatory roles as summarized in Supplementary Table S3. High expression of SLC34A2 has been reported in papillary thyroid carcinoma (PTC) tissues [59]. Mutations in the genes SFTPA1 or SFTPA2 have been associated with interstitial lung diseases and lung cancer [60] while Napsin-A (NAPSA) and anterior gradient protein 2 homolog (AGR2) might be candidates' proteins involved in stage IA and IIIA lung adenocarcinoma [61]. Over-expression of ADCY4, VIPR1 and TGFBR2 have been found to be associated with clinical stages of lung adenocarcinoma (LUAD).

5. Conclusion

In this study, we have presented a feature selection and association rule mining approach that can be used in identification of phenotypic characteristics from gene expression data. We used differential gene expression and coexpression network analysis to select significantly expressed genes thereby extracting informative features from high dimensional RNAseq data. Association rule mining was used to find itemsets that were always associated. This approach could help infer if there exists a relationship among the genes and their possibility of having a shared biological function. The main advantage of association rule mining is that unlike clustering, a gene is not limited to a single rule since it can belong to several rules. Based on our findings we recommend a combination of differential gene expression analysis, network analysis and association rule mining in order to find meaningful biological association between genes of interest when analysing RNAseq data. This is because of the complex nature of biological patterns or associations. Genes with unknown function can be assigned potential biological function based on their appearance in one or more rules together with features with an assigned function.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We acknowledge Prof. Romano Mwirichia of university of Embu for allowing us to access the computational resources used in data analysis.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imu.2021.100595.

References

[1] Marx V. The big challenges of big data. Nature 2013;498(7453):255–60.
[2] Mattmann CA. A vision for data science. Nature 2013;493(7433):473–5.
[3] Uma SM, Kirubakaran E. A hybrid heuristic dimensionality reduction technique for microarray gene expression data classification: a blending of GA, PSO and ACO. Int J Data Min Model Manag 2016;8(2):160–79.
[4] Zhou LT, Cao YH, Lv LL, Ma KL, Chen PS, Ni HF, Liu BC. Feature selection and classification of urinary mRNA microarray data by iterative random forest to diagnose renal fibrosis: a two-stage study. Sci Rep 2017;7(1):1–9.
[5] Guo X, Zhang Y, Zheng L, Zheng C, Song J, Zhang Q, Kang B, Liu Z, Jin L, Xing R, Gao R. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. Nat Med 2018;24(7):978–85.
[6] Bellman R, Kalaba R. Dynamic programming and statistical communication theory. Proc Natl Acad Sci U S A 1957;43(8):749.
[7] Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artif Intell 1997;97(1–2):245–71.
[8] Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinform 2015;2015:13 pages. https://doi.org/10.1155/2015/198363. Article ID 198363.
[9] Chen RC, Dewi C, Huang SW, Caraka RE. Selecting critical features for data classification based on machine learning methods. Journal of Big Data 2020;7(1):1–26.
[10] Vergara JR, Estévez PA. A review of feature selection methods based on mutual information. Neural Comput Appl 2014;24(1):175–86.
[11] Vanjimalar S, Ramyachitra D, Manikandan P. A review on feature selection techniques for gene expression data. In: *2018* IEEE international conference on computational intelligence and computing research (ICCIC). IEEE; 2018, December. p. 1–4.
[12] Jiang L, Greenwood CM, Yao W, Li L. Bayesian hyper-LASSO classification for feature selection with application to endometrial cancer RNA-seq data. Sci Rep 2020;10(1):1–16.
[13] Bermingham ML, Pong-Wong R, Spiliopoulou A, Hayward C, Rudan I, Campbell H, Haley CS. Application of high-dimensional feature selection: evaluation for genomic prediction in man. Sci Rep 2015;5(1):1–12.
[14] Liu HJ, Guo YY, Li DJ. Predicting novel salivary biomarkers for the detection of pancreatic cancer using biological feature-based classification. Pathol Res Pract 2017;213(4):394–9.
[15] Zhang B, He X, Ouyang F, Gu D, Dong Y, Zhang L, Zhang S. Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma. Canc Lett 2017;403:21–7.
[16] Goh WWB, Wong L. NetProt: complex-based feature selection. J Proteome Res 2017;16(8):3102–12.
[17] Frank O, Brors B, Fabarius A, Li L, Haak M, Merk S, Seifarth W. Gene expression signature of primary imatinib-resistant chronic myeloid leukemia patients. Leukemia 2006;20(8):1400–7.
[18] Zheng C, Li L, Haak M, Brors B, Frank O, Giehl M, Seifarth W. Gene expression profiling of CD34+ cells identifies a molecular signature of chronic myeloid leukemia blast crisis. Leukemia 2006;20(6):1028–34.
[19] Das R, Dimitrova N, Xuan Z, Rollins RA, Haghighi F, Edwards JR, Zhang MQ. Computational prediction of methylation status in human genomic sequences. Proc Natl Acad Sci Unit States Am 2006;103(28):10713–6.
[20] He J, Zhou M, Li X, Gu S, Cao Y, Xing T, Zou Q. SLC34A2 simultaneously promotes papillary thyroid carcinoma growth and invasion through distinct mechanisms. Oncogene 2020;39(13):2658–75.
[21] Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet 2015;24(8):2125–37.

[22] Wang W, Sue ACH, Goh WW. Feature selection in clinical proteomics: with great power comes great reproducibility. Drug Discov Today 2017;22(6):912–8.
[23] Tang W, Wan S, Yang Z, Teschendorff AE, Zou Q. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. Bioinformatics 2018;34(3):398–406.
[24] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):1–21.
[25] Wenric S, ElGuendi S, Caberg JH, Bezzaou W, Fasquelle C, Charloteaux B, Bours V. Transcriptome-wide analysis of natural antisense transcripts shows their potential role in breast cancer. Sci Rep 2017;7(1):1–12.
[26] Williams CR, Baccarella A, Parrish JZ, Kim CC. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. BMC Bioinf 2017;18(1):1–12.
[27] Anders S, Huber W. Differential expression analysis for sequence count data. Nature Precedings 2010. https://doi.org/10.1038/npre.2010.4282.2.
[28] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010;11(3):1–9.
[29] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 2005;4(1).
[30] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinf 2008;9(1):1–13.
[31] Gallo CA, Cecchini RL, Carballido JA, Micheletto S, Ponzoni I. Discretization of gene expression data revised. Briefings Bioinf 2016;17(5):758–70.
[32] Alagukumar S, Lawrance R. A selective analysis of microarray data using association rule mining. Procedia Computer Science 2015;47:3–12.
[33] Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD international conference on Management of data; 1993, June. p. 207–16.
[34] Creighton C, Hanash S. Mining gene expression databases for association rules. Bioinformatics 2003;19(1):79–86.
[35] Ceddia G, Martino LN, Parodi A, Secchi P, Campaner S, Masseroli M. Association rule mining to identify transcription factor interactions in genomic regions. Bioinformatics 2020;36(4):1007–13.
[36] Alam TM. Identification of malignant mesothelioma risk factors through association rule mining. 2019.
[37] Alagukumar S, Lawrance R. January). Classification of microarray gene expression data using associative classification. 2016 international conference on computing technologies and intelligent data engineering (ICCTIDE, vol. 16. IEEE; 2016. p. 1–8.
[38] Mallik S, Zhao Z. Distance based knowledge retrieval through rule mining for complex biomarker recognition from tri-omics profiles. Int J Comput Biol Drug Des 2019;12(2):105–27.
[39] Carmona-Saez P, Chagoyen M, Rodriguez A, Trelles O, Carazo JM, Pascual-Montano A. Integrated analysis of gene expression by association rules discovery. BMC Bioinf 2006;7(1):1–16.
[40] Chen RC, Dewi C, Huang SW, Caraka RE. Selecting critical features for data classification based on machine learning methods. Journal of Big Data 2020;7(1):1–26.
[41] Jiang L, Huang J, Higgs BW, Hu Z, Xiao Z, Yao X, Yao Y. Genomic landscape survey identifies SRSF1 as a key oncodriver in small cell lung cancer. PLoS Genet 2016;12(4):e1005895.
[42] Al Mahi N, Najafabadi MF, Pilarczyk M, Kouril M, Medvedovic M. GREIN: an interactive web platform for re-analyzing GEO RNA-seq data. Sci Rep 2019;9(1):1–9.
[43] Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. 2010, http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
[44] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30(15):2114–20.
[45] Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 2017;14(4):417–9.
[46] Chiesa M, Colombo GI, Piacentini L. DaMiRseq—an R/Bioconductor package for data mining of RNA-Seq data: normalization, feature selection and classification. Bioinformatics 2018;34(8):1416–8.
[47] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):1–21.
[48] 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.
[49] Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 2014;15(2):1–17.
[50] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 2005;4(1).
[51] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13(11):2498–504.
[52] Hahsler M, Buchta C, Gruen B, Hornik K. Arules: mining association rules and frequent itemsets. R package version 1.3-1. Michael Hahsler; 2014.
[53] Wang W, Sue ACH, Goh WW. Feature selection in clinical proteomics: with great power comes great reproducibility. Drug Discov Today 2017;22(6):912–8.
[54] Dębski KJ, Pitkanen A, Puhakka N, Bot AM, Khurana I, Harikrishnan KN, Kobow K. Etiology matters–genomic DNA methylation patterns in three rat models of acquired epilepsy. Sci Rep 2016;6(1):1–14.
[55] Watts DJ, Strogatz SH. Collective dynamics of 'small-world'networks. Nature 1998;393(6684):440–2.
[56] Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: an extended review and a software tool. PloS One 2017;12(12):e0190152.
[57] Chen X, Wang L, Su X, Luo S-y, Tang X, Huang Y. Identification of potential target genes and crucial pathways in small cell lung cancer based on bioinformatic strategy and human samples. PloS One 2020;15(11):e0242194.
[58] Elsheikh SE, Green AR, Rakha EA, Samaka RM, Ammar AA, Powe D, Ellis IO. Caveolin 1 and Caveolin 2 are associated with breast cancer basal-like and triple-negative immunophenotype. Br J Canc 2008;99(2):327–34.
[59] He J, Zhou M, Li X, Gu S, Cao Y, Xing T, Zou Q. SLC34A2 simultaneously promotes papillary thyroid carcinoma growth and invasion through distinct mechanisms. Oncogene 2020;39(13):2658–75.
[60] Legendre M, Butt A, Borie R, Debray MP, Bouvry D, Filhol-Blin E, Nathan N. Functional assessment and phenotypic heterogeneity of SFTPA1 and SFTPA2 mutations in interstitial lung diseases and lung cancer. Eur Respir J 2020;56(6).
[61] Kawamura T, Nomura M, Tojo H, Fujii K, Hamasaki H, Mikami S, Nishimura T. Proteomic analysis of laser-microdissected paraffin-embedded tissues:(1) Stage-related protein candidates upon non-metastatic lung adenocarcinoma. Journal of proteomics 2010;73(6):1089–99.