Journal of Biomedical Informatics 96 (2019) 103247

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Cancer classification and pathway discovery using non-negative matrix factorization

Zexian Zeng[a], Andy H. Vo[b], Chengsheng Mao[a], Susan E. Clare[c,*], Seema A. Khan[c,*], Yuan Luo[a,*]

[a] Department of Preventive Medicine, Northwestern University, Feinberg School of Medicine, Chicago, IL, USA
[b] Committee on Developmental Biology and Regenerative Medicine, The University of Chicago, Chicago, IL, USA
[c] Department of Surgery, Northwestern University, Feinberg School of Medicine, Chicago, IL, USA

ARTICLE INFO

Keywords:
Non-negative matrix factorization
Cancer
Classification
Whole-exome sequencing
Somatic mutation
Pathway

ABSTRACT

Objectives: Extracting genetic information from a full range of sequencing data is important for understanding disease. We propose a novel method to effectively explore the landscape of genetic mutations and aggregate them to predict cancer type.

Design: We applied non-smooth non-negative matrix factorization (nsNMF) and support vector machine (SVM) to utilize the full range of sequencing data, aiming to better aggregate genetic mutations and improve their power to predict disease type. More specifically, we introduce a novel classifier to distinguish cancer types using somatic mutations obtained from whole-exome sequencing data. Mutations were identified from multiple cancers and scored using SIFT, PP2, and CADD, and collapsed at the individual gene level. nsNMF was then applied to reduce dimensionality and obtain coefficient and basis matrices. A feature matrix was derived from the obtained matrices to train a classifier for cancer type classification with the SVM model.

Results: We have demonstrated that the classifier was able to distinguish four cancer types with reasonable accuracy. In five-fold cross-validations using mutation counts as features, the average prediction accuracy was 80% (SEM = 0.1%), significantly outperforming baselines and outperforming models using mutation scores as features.

Conclusion: Using the factor matrices derived from the nsNMF, we identified multiple genes and pathways that are significantly associated with each cancer type. This study presents a generic and complete pipeline to study the associations between somatic mutations and cancers. The proposed method can be adapted to other studies for disease status classification and pathway discovery.

1. Background and significance

Personalized medicine is becoming increasingly popular in cancer where genetic profiles of tumors can be used to guide clinical decisions such as treatment options and preventive measures [1]. The development of massively parallel, high throughput DNA sequencing technology has enabled the cataloging of somatic mutations in cancer, making genomic data increasingly accessible. Understanding the association between genetics and disease is important for understanding the underlying pathophysiology. In cancer, many molecular and genomic studies have identified somatic mutations within genes associated with cancer initiation, progression, and treatment responses [2–4].

The majority of sequencing studies have focused on the identification of individual driver genes [5]. However, driver mutations are often highly heterogeneous between cancer genomes, even within the same type of cancer [6]. Furthermore, studies have observed cancer to be highly complex, often resulting from multiple interacting mutations and related pathways [7,8]. While many methods attempt to address the complex mutational heterogeneity in cancer, it still remains a challenge due to limited study-power and lack of complete knowledge regarding gene and pathway interaction [9–13]. Despite the fact that mutations in many genes have been identified in cancer, it is not yet understood how these genes cumulatively interact in the development and progression of cancer. It has been a challenge to study these mutations and their interactions together due to large-scale complexity.

It is important to consider methods that can encompass the full

[*] Corresponding authors at: Department of Surgery, Feinberg School of Medicine, Northwestern University, Robert H Lurie Medical Research Center, Room 4-113 250 E Superior, Chicago, IL 60611, USA (S.E. Clare). Department of Surgery, Feinberg School of Medicine, Northwestern University, NMH/Prentice Women's Hospital, Room 4-420 250 E Superior, Chicago, IL 60611, USA (S.A. Khan). Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, 750 N Lake Shore Drive Room 11-189, Chicago, IL 60611, USA (Y. Luo).

E-mail addresses: susan.clare@northwestern.edu (S.E. Clare), s-khan2@northwestern.edu (S.A. Khan), yuan.luo@northwestern.edu (Y. Luo).

https://doi.org/10.1016/j.jbi.2019.103247
Received 6 February 2019; Received in revised form 23 April 2019; Accepted 1 July 2019
Available online 02 July 2019
1532-0464/

scope of genes. When genes and mutations are studied together, novel biological interactions and pathways can be identified to further provide biological and clinical insights. Many groups have previously utilized feature selection methods for removing irrelevant and redundant information to deal with complexity problems. Vector Quantization (VQ) [14] and Principle Component Analysis (PCA) [15] have been widely used for feature selection. More recently, attention has been drawn to non-negative matrix factorization (NMF). In a face recognition study, Lee et al. suggested NMF could outperform VQ and PCA for feature recognition [16]. In addition, the non-negative constraint of NMF is important because non-negativity is more realistic, easier to interpret, and prevalent in real world applications. In particular, NMF has been applied to disease subtype studies using gene expression data [17,18] and sequencing data [19–21]. With the aim to uncover the genetic complexity behind cancer development, and to identify mutations that directly affect processes involved with oncogenesis, we propose a framework utilizing NMF.

In our proposed framework, NMF was applied to discover latent factors from somatic mutations. The discovered latent factors were used to train an SVM model for cancer type classification. The NMF-SVM combination was rigorously evaluated and compared to different baselines. Association studies were performed between the factor matrices derived from NMF and cancer type using penalized logistical regressions. Major factors associated with each cancer type were investigated, and significant genes were identified for investigation in pathway discovery analysis. In addition to this proposed framework serving as a disease type classifier, it can also be utilized to elucidate novel biological interactions and pathways for disease. The details of the study are reported below.

2. Material and methods

2.1. Mutation profiles

As a pilot study, four prevalent cancers were retrieved from The Cancer Genome Atlas (TCGA), including Glioblastoma Multiforme (GBM), Breast invasive carcinoma (BRCA), Lung Squamous Cell Carcinoma (LUSC), and Prostate Adenocarcinoma (PRAD). Somatic mutations were identified from 2431 tumors (Table 1). SnpEFF [22] and ANNOVAR [23] were used to annotate 24,588 missense mutations and 57,319 nonsense mutations in the study cohort. Each mutation was functionally scored for being potentially deleterious using SIFT [25], PolyPhen2 (PP2) [26], and CADD [24] scores. In genes containing multiple mutations, SIFT, PP2, and CADD scores, as well as mutational frequency were collapsed and studied as a single variable separately, known as gene burden [24]. Predicted pathogenicity scores (SIFT, PP2, and CADD) were calculated for each mutation within a gene and collapsed as a sum to calculate the gene burden for a specific gene. Namely, gene burden represents a gene's total predictive pathogenicity based on mutation data. Thus, gene burden was used to represent the damage level of a gene from multiple perspectives. A workflow is illustrated to show the methods used in this study (Fig. 1).

Table 1
The number of samples in each cancer type and the corresponding number of somatic mutations. Mutations annotated with moderate effects are missense mutations or in-frame shift mutations. Mutations annotated as high effects are nonsense mutations. Numbers in parenthesis are standard deviations.

	Cancer	Sample size	Somatic moderate	Somatic high
BRCA	Breast Invasive Carcinoma	1044	68 (11)	20 (5)
LUSC	Lung Squamous Cell Carcinoma	497	214 (15)	44 (3)
PRAD	Prostate Adenocarcinoma	497	34 (19)	8 (3)
GBM	Glioblastoma Multiforme	393	133 (55)	27 (9)

2.2. Gene pre-selection

Prior to modeling, we evaluated whether a subset of representative genes could be derived without information lost to achieve a more balanced sample feature ratio and reduce noise. The collapsed score in each gene was used as an input variable while the cancer type was used as the output variable. Multinomial logistic regression was fit, and a P-value that yields the null hypothesis of corresponding coefficient being zero was used as an indicator for the pre-selection. The selection criterion for this initial screening was set with a P-value less than or equal to a cutoff. In order to reduce noise and to prevent model overfitting, we tested the model using multiple cutoff thresholds of 0.05, 0.1, 0.2, 0.5, and 1. We compared the results derived from each threshold and selected the most reasonable cutoff based on prediction accuracy and number of features.

2.3. Applying NMF to discover latent factors of somantic mutations

Genes passing the selection threshold were used as inputs for the NMF study. Assume there were N subjects and M selected genes. The data were represented by a matrix A_{Score} of size $M \times N$. The columns of A_{Score} represents the collapsed score of the M genes in the N subjects. The matrix A_{Score} was then decomposed using NMF. The purpose of this study was to find a set of intrinsic patterns that are likely to distinguish cancer types. To perform NMF, the matrix A_{Score} was factored into two low-rank matrices W and H. Mathematically, A_{Score} is approximated by $A_{Score} \approx WH$.Matrix A_{Score} is the approximate linear combinations of the column vectors in matrix W and the coefficients supplied by columns in matrix H. Matrix W has size $M \times K$, with each of the K columns representing a group of weighted genes and w_{ij} corresponding to the weight of gene i in group j. K denotes the number of factors and is a given input. Matrix H has size $K \times N$, where each of the N columns denotes the feature coefficients for each subject. Entry h_{ij} is the value of feature i in sample j. The decomposition is achieved by iteratively updating the matrix W and H to minimize a divergence objective [16,25]. Specifically, for the purpose of sparseness, we used non-smooth Nonnegative Matrix Factorization (nsNMF) [26]. More specifically, the application of nsNMF led to a high degree of sparseness by adding a positive symmetric matrix in the objective function. We achieved modest to high degree of sparseness in both the W (average 51% sparseness) and H matrix (average 85% sparseness). Each analysis was repeated ten times to address the local optima problem.

2.4. Classifier training

Matrix H has size $K \times N$, where each of the N columns denotes the feature coefficients for the corresponding subject. To retain the information from both W and H matrices, a new matrix F was generated by multiplying the transposed matrix A_{Score} with matrix W. Specifically, an entry in matrix F was computed as$f_{ij} = \sum_{x=1}^{m} A_{ix}^{\mathrm{T}} * W_{xj}$. Matrix F has size $N \times K$, with each of the K columns representing the coefficient for each subject. Since $A_{Score} \approx WH$, matrix F can be approximated by a kernel matrix $(W \times H)^T \times W$. Subsequently, columns in matrix F were utilized as features to train the classifier. Support vector machine (SVM) was used for the training, where each column corresponds to one predictor in the model. The Radial Basis Function (RBF) kernel was used, with parameters of gamma and C set to default. This trained SVM model was a cancer type classifier.

2.5. Factor number selection

Note that before factorization, the number of factors K need to be pre-defined. Typically, the number of factors K is chosen so that $(N + M) \times K < N \times M$ [16]. Selection of K is critical because it determines the number of patterns to be found. Numerous studies have presented different methods for factor number selection: The factor

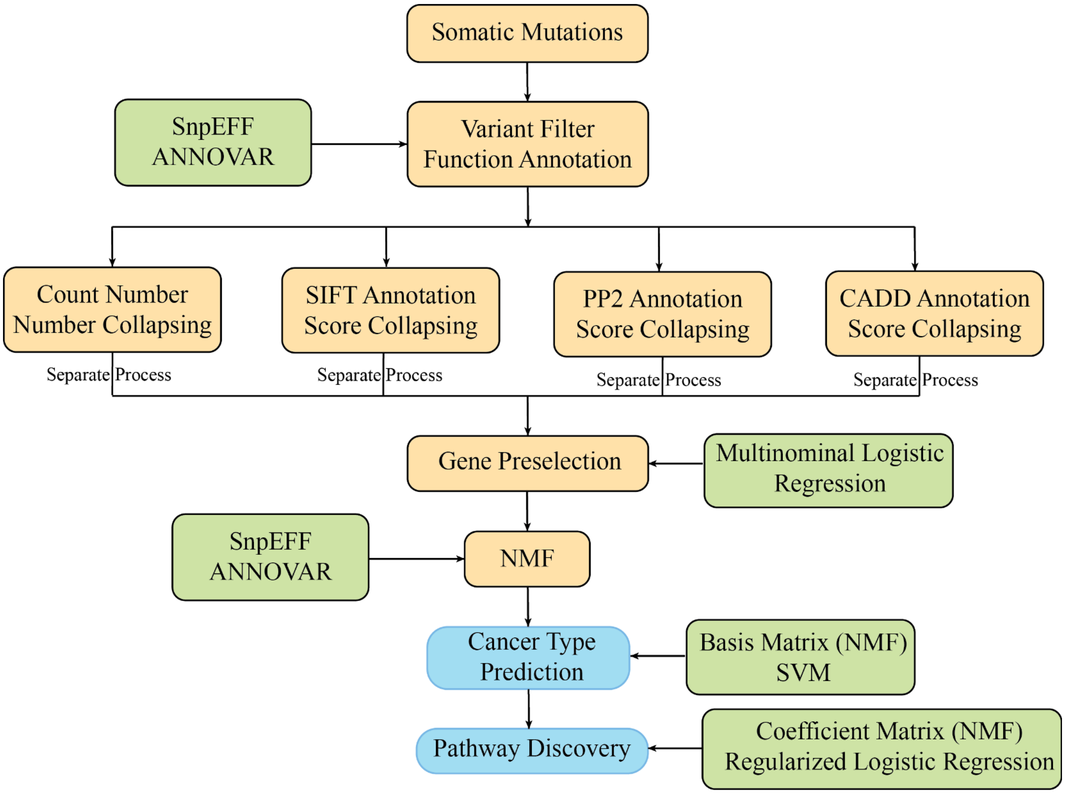

Fig. 1. Workflow of the study. Orange boxes are the data or processes; green boxes are the tools used; blue boxes are the results of the study. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

number K can be determined based on different metrics composing of a cophenetic correlation coefficient [18,26], variation of sum of squares [27], or maximum information reservation [28]. In our study, the most important feature for the classifier is the ability to identify intrinsic patterns that best distinguish cancer types accurately. To achieve this, a numerical screening test was conducted to screen through the different number of factors for best prediction performance. The screened factor numbers ranged from 2 to 15. Multi-class prediction accuracy in each classification was obtained as a performance measurement. Five-fold cross-validation was conducted using multiple different factor numbers. The factor number with the best performance in the cross-validation was chosen. Each experiment was replicated ten times with different initial seeds. Prediction accuracy, precision, recall, and f-measure were used as performance evaluation matrices.

2.6. Evaluation

To set up baselines for comparison, the mutation frequency and the collapsed scores were used as independent predictors to fit SVM models and penalized logistic regression models were used to predict cancer types. All somatic mutations were also used as predictive variables for cancer classification using the SVM and penalized logistical regression model. A number of models have been developed for cancer classification utilizing somatic mutation profiles, including variations of CADD scores [29], logistical regression on L1-regularised terms [30], and SVM-RFE [30]. In this study, we compared the performances between our proposed model and these reported methods for cancer classification. All studies were replicated ten times with different initial seeds and significance tests were performed for evaluations. P-values were obtained for the evaluations.

2.7. Pathway study

The feature matrix F was obtained by multiplying a transposed matrix A_{Score} with matrix W, with size $N \times K$, where each K columns represents the coefficient for each subject. To determine the association level of each factor with each cancer type, elastic net regression models were fitted using the cancer type as the output variable and each of the columns in the F matrix as input variables. To differentially discover pathways for a cancer type, subjects with the cancer type of interest were treated as cases and the remaining subjects as controls. The regulation parameter λ was selected using ten-fold cross-validation. Association effects for each feature were represented by the beta value and the vector was denoted as $\widehat{\boldsymbol{\beta}}$. We selected the factor corresponding to the largest $\widehat{\boldsymbol{\beta}}$, and denote the factor as a cancer-relevant factor. After selecting the factor, we utilized the matrix W to rank genes. The weights in the matrix W are composed of genes with different weights. We assume genes with the largest coefficients cumulatively and linearly interact with each other and are associated with cancer development. For each cancer, we repeated the experiment and selected the top genes in each factor for enrichment analysis and presented the top significant pathways (adjusted P-value < 0.05). To ensure stable results, we analyzed the top 100, 200, 300, 400, 500, and 600 genes for pathway analysis and determined 300 genes started yielding stable results as illustrated with BRCA (Table S2).

3. Experiment results

After collapsing mutations' numbers and scores in each gene, a matrix A_{score} was formed with the 2431 subjects as rows. Entry A_{ij} denotes the j_{th} gene's collapsed score for the i_{th} subject. For gene preselection, we screened the p_{value} of 0.05, 0.1, 0.2, 0.5, and 1 as cutoffs. For each cutoff, prediction accuracy was used as the performance measurement. The experiment was repeated using the Number (collapsed number of mutations), SIFT (collapsed sift scores), PP2 (collapsed PP2 score), and CADD (collapsed CADD scores) matrices. Factor numbers ranged from 2 to 15 (Fig. 2). The Number matrix was found to result in better performance than the other matrices. Using the Number

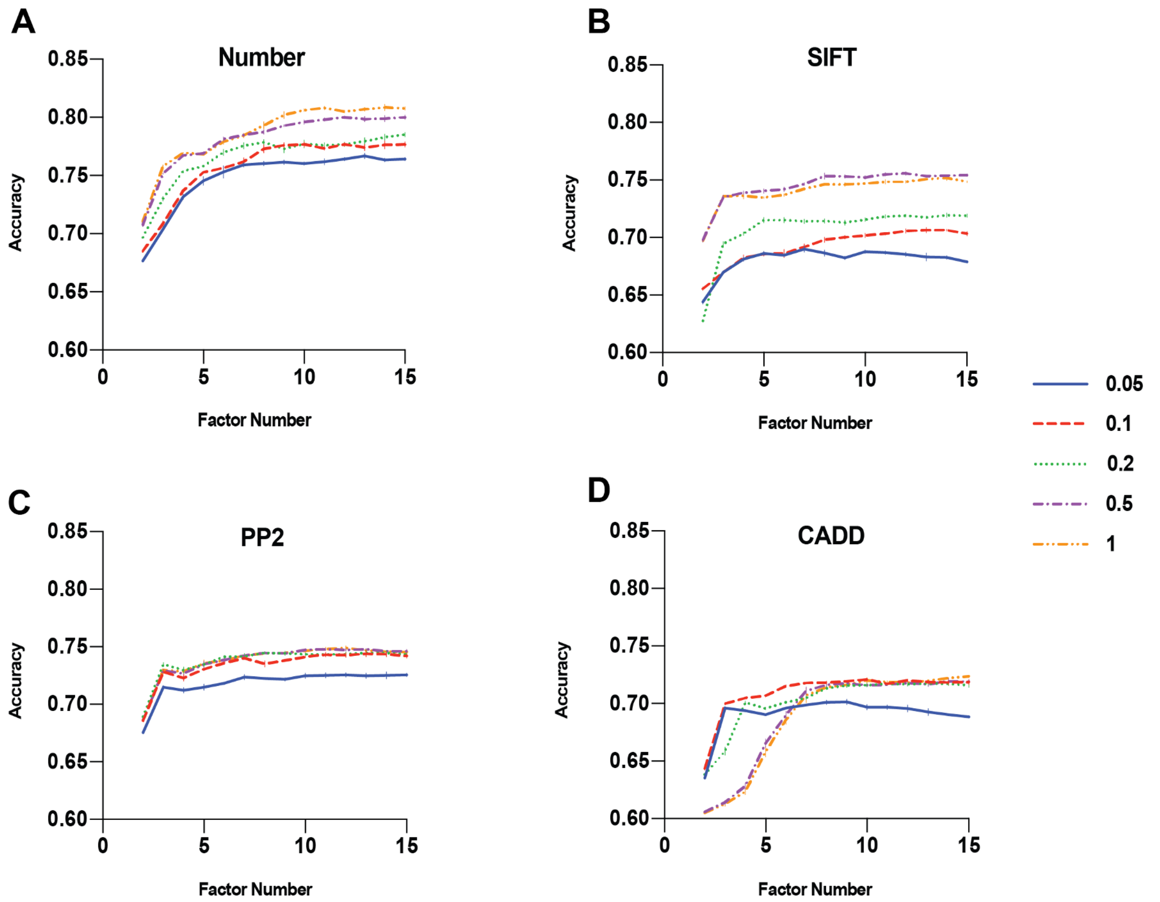

Fig. 2. The accuracy of cancer type classification using different P-value cutoffs. (A) Sum of the count of mutations (B) Sum of the SIFT scores (C) Sum of the PP2 scores (D) Sum of the CADD scores.

matrix, the performance derived from the cutoff of 1 is significantly better than the cutoff of 0.05 (P-value = 0.001), 0.1 (P-value = 0.01), and 0.2 (P-value = 0.04), but not significantly different from the cutoff of 0.5 (P-value = 0.61). Balancing the number of features to be included for computation and accuracy, we selected the cutoff of 0.5 for gene pre-selection. Following gene pre-selection, 11,949, 12,753, 17,702, and 11,734 genes were retained for subsequent analysis for the Number, SIFT, PP2, and CADD matrix, respectively.

We then compared these four matrices: Number, SIFT, PP2, and CADD. The number of factors K ranged from 2 to 15, a range within the constraint of the rule $(N + M)K < NM$. For each factor number K, nsNMF was applied to the matrices, and a corresponding classifier was trained. The performances derived from the Number matrix outperformed the other matrices significantly ($p < 0.01$ for all comparisons) (Fig. 3). The precision, recall, and f-measures were derived and similar patterns and trends were observed. Based on performance, the matrix Number was used for subsequent analyses. Using Number matrix, the maximum accuracy was 80.0% (Standard Error of the Mean SEM = 0.1%) when the factor number equaled 12 (Fig. 3). The accuracy was found to become stable when factor number was larger than 12. To prevent potential overfitting, we chose a factor of 12 for our analysis.

The performance of our proposed model (80.0%, SEM = 0.1%) significantly outperformed the other four baselines (Fig. 4). The P-value using the Student's *t*-test was 0.0001 comparing our proposed model to the second-ranked model (73.9% (SEM = 0.8%), which applies penalized logistical regression with the aggregated Number matrix. In the baselines, aggregating the mutations in a gene has improved the performance significantly as well ($p < 0.01$ in both comparisons). A comparison of our proposed model with previously applied methods for cancer classification was conducted. We found that our method achieved significantly improved performance (Fig. 5) compared to methods that utilize variations of CADD scores (71.6%) [29], logistical regression on L1-regularised term (74.0%) [30], and SVM-RFE (55.9%) [30].

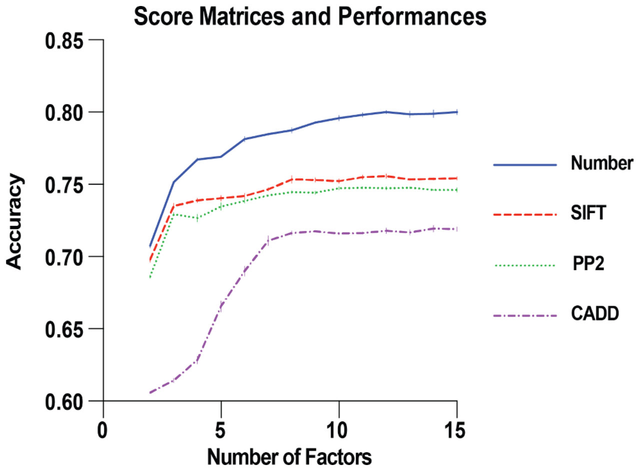

Fig. 3. The accuracy of cancer type predictions using different numbers of factors from the matrices of Number (blue), SIFT (red), PP2 (green), and CADD (purple) scores. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Using regularized logistic regression, we assessed each gene's association effect with each cancer type. The association score was defined as the sum of feature weights multiplied by the coefficient of each

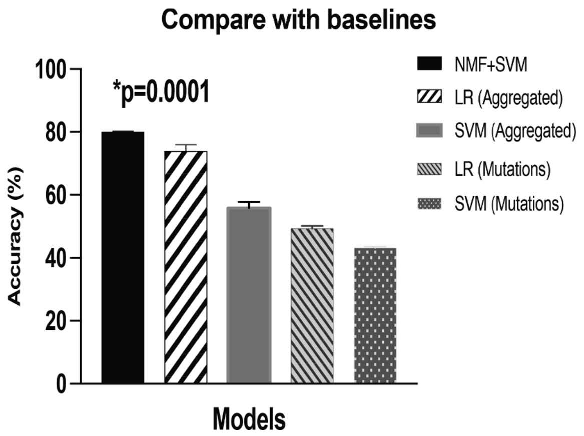

Fig. 4. Comparison of our proposed model (nsNMF + SVM) with baselines. LR is penalized logistical regression. SVM is support vector machine. Aggregated are the matrices to sum mutations together in the same gene. Mutations are the model that utilizes every single mutation as an input variable.

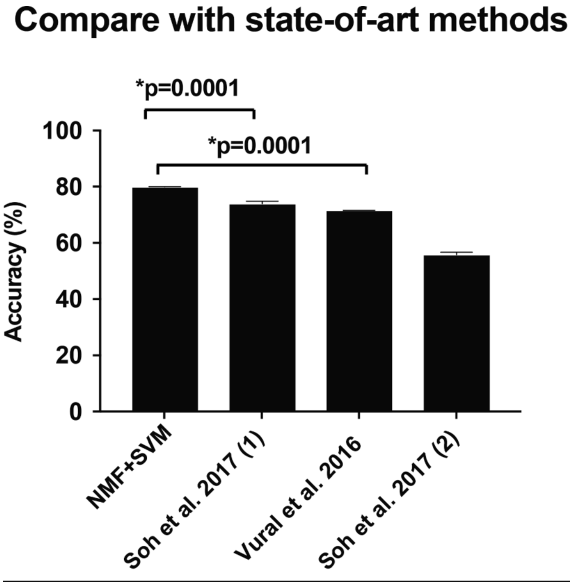

Fig. 5. Comparison of our proposed model (nsNMF + SVM) with the state-of-the-art methods.

Table 2
Biological processes most associated with each cancer type.

GO term	Cancer
Microtubule-based process	BRCA
Axon guidance	GBM
Morphogenesis of a polarized epithelium	PRAD
Chemical synaptic transmission	LUSC

gene. A high score indicates the significant role of a mutated gene in the disease. The genes for each cancer type were identified and sorted by association score (Table S1). The top 300 genes associated with each cancer type were derived and analyzed for pathway enrichment. Interestingly, distinct biological processes were found to be significantly associated ($p < 0.05$) with each type of cancer. Microtubule processes, axon guidance, cell morphogenesis, and synaptic transmission were found to be associated with BRCA, GBM, PRAD, and LUSC, respectively (Table 2). Additional pathways associated with each cancer type can be found in Table S2. Many of these pathways are known to be associated with their corresponding cancer type. For instance, a majority of breast cancer drugs involve targeting microtubules [31]. In glioblastoma, axon guidance is known to play a role in glioma progression [32]. To confirm the robustness of our findings, we performed and evaluated pathway enrichment and discovery using multiple methods [33–37]. We also discovered axon guidance from the Reactome pathway database [36] and synaptic transmission in the KEGG pathway database [37]. Together, these results corroborated the relevant pathway discovered by our method.

4. Discussion

We have proposed a novel method to fully use and understand somatic mutations to classify the cancer type and derive relevant genes and pathways. In this study, we applied nsNMF and SVM to train a classifier to distinguish and classify a tumor type as Glioblastoma Multiforme (GBM), Breast Invasive Carcinoma (BRCA), Lung Squamous Cell Carcinoma (LUSC), and Prostate Adenocarcinoma (PRAD). Products of the basis matrix and coefficient matrix derived from nsNMF were both retained to construct the feature matrix. Subsequently, the constructed features were used as input variables to train the classifier. We compared functional scores using CADD, SIFT, and PP2, and counted mutation number and found that counted mutation number yielded the best performance (accuracy = 80.0% with SEM = 0.1%). Finally, regularized logistic regression was applied to study each gene's association effect with cancer type. Using the associated features, we derived relevant genes and pathways for each cancer.

When training the classifier, we used an alternative method by multiplying the matrix A_{score} with matrix W to obtain the feature matrix F. Information from the basis component W was retained, providing information about weights in each gene group. This information was then used as features to train the classifier. Another benefit of this alternative method is the ease of us at the testing stage. With the trained W matrix, we only need to multiply the testing A_{score} matrix in order to get the test feature matrix. In addition to improving cancer type classification, each gene's association effect with the cancers was of interest and also studied. The p-value for gene pre-selection was to limit the number of features to be included. One of the challenges for genomics studies are the large number of genes accompanied by the small sample sizes, resulting in a wide and flat matrix, henceforth impact the performance of matrix decomposition. In this study, we utilized a p-value cutoff to pre-select genes but try to only introduce a minimum amount of influence on model performance. Therefore, we have tested multiple p-values and selected the cutoff of 0.5, in which we observed non-significant differences in model performances compared to the no-selection scenario. Genes that were filtered are those with only one or two mutations and only appeared in one or two subjects in the cohort. Removing these genes has a minimum amount of influence and has the potential to remove noises for model training. In our study, if we tune NMF + SVM (our model), we can get even better results. But our purpose is to focus on assessing improvements from NMF. In addition, with the default parameters for NMF + SVM, our model still outperformed the state-of-the-art methods that are parameter-tuned.

The development of high throughput sequencing technology has enabled the cataloging of large-scale mutation information. Somatic mutations are relatively stable and lead to the initiation and progression of many sporadic cancers. Hence in this study, we utilized mutations in protein-coding genes as input data. We acknowledge that non-protein-coding genes, including mutations in intronic areas [38,39], long non-coding RNAs [40], mi-RNAs [41] are also important for cancer development. In future work, we will incorporate these multiple dimensions of genetics data to increase the model performance. Traditionally, mutations derived from sequence data were examined as a single variable using the regression models [30,42]. Unfortunately, the large number of variables limit the power of such studies. To reduce the number of variables, studies have proposed to aggregate mutations at the gene level as an input in a regression model [24,43,44]. In other studies, mutations in a gene have also been proposed to be studied in a matrix as an input for a kernel test [45,46]. In this study, we proposed using a framework which utilizes a regression model to pre-select deleterious genes, nsNMF to decompose the matrix, SVM to train a classifier, and then penalized regression to derive relevant genes. Following the careful tuning of parameters and models, we have proved that this is an effective model to classify cancers, derive relevant genes, and identify associated pathways.

5. Conclusion

To fully understand a disease, studying mutations using a full range of genes together is of critical importance. Complex traits are modified by multiple genes and multiple mutations together [47]. Traditionally, NMF has been applied to study gene expression [18,28]. In this study, we proposed using somatic mutations for cancer classification. Furthermore, we proposed generating the feature matrix by integrating both the basis matrix W and the coefficient matrix H. Moreover, we developed a novel method to derive effect scores from the feature matrix. Using this method, we obtained the association score of each gene with a particular cancer type enabling relevant pathway discovery. The discovered effect scores have a high potential to help us better understand the genetic pathophysiology behind cancer.

In this study, we proposed a novel strategy to study the genetic landscape of multiple cancers. In the future, we will use tensor factorization to integrate known pathways to guide the grouping of mutational variants [48] and use external cohorts to validate the proposed model. Furthermore, this generic process only requires the input of somatic mutations and a disease type of interest, without much domain specific knowledge. This strategy has the potential to be easily adapted and applied to other diseases as well.

Funding statement

This study was supported in part by grant R21LM012618-01 from the National Institutes of Health, Breast Cancer Research Foundation, and the Lynn Sage Cancer Research Foundation.

Contributorship statement

ZZ, SC, SK, and YL originated the study. ZZ and YL performed analyses and wrote the first draft of the manuscript. ZZ, CM, and AV annotated the dataset. SC and SK reviewed and helped analyze the findings. All authors discussed the results and revised the manuscript.

Declaration of Competing Interest

The authors have no competing interests to declare.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2019.103247.

References

[1] I.S. Kohane, Ten things we have to do to achieve precision medicine, Science 349 (2015) 37–38.

[2] S. Misale, R. Yaeger, S. Hobor, E. Scala, M. Janakiraman, D. Liska, et al., Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer, Nature 486 (2012) 532.

[3] M. Peifer, L. Fernández-Cuesta, M.L. Sos, J. George, D. Seidel, L.H. Kasper, et al., Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer, Nat. Genet. 44 (2012) 1104.

[4] C. Greenman, P. Stephens, R. Smith, G.L. Dalgliesh, C. Hunter, G. Bignell, et al., Patterns of somatic mutation in human cancer genomes, Nature 446 (2007) 153.

[5] P.A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, et al., A census of human cancer genes, Nat. Rev. Cancer 4 (2004) 177.

[6] M.R. Stratton, P.J. Campbell, P.A. Futreal, The cancer genome, Nature 458 (2009) 719.

[7] C.-H. Yeang, F. McCormick, A. Levine, Combinatorial patterns of somatic gene mutations in cancer, FASEB J. 22 (2008) 2605–2622.

[8] Q. Cui, Y. Ma, M. Jaramillo, H. Bari, A. Awan, S. Yang, et al., A map of human cancer signaling, Mol. Syst. Biol. 3 (2007) 152.

[9] N.J. Risch, Searching for genetic determinants in the new millennium, Nature 405 (2000) 847–856.

[10] E.S. Lander, N.J. Schork, Genetic dissection of complex traits, Science- New York Then Washington 2037 (1994).

[11] M.D. Leiserson, D. Blokh, R. Sharan, B.J. Raphael, Simultaneous identification of multiple driver pathways in cancer, PLoS Comput. Biol. 9 (2013) e1003054.

[12] R.D. Melamed, J. Wang, A. Iavarone, R. Rabadan, An information theoretic method to identify combinations of genomic alterations that promote glioblastoma, J. Mol. Cell. Biol. 7 (2015) 203–213.

[13] Y. Luo, G. Riedlinger, P. Szolovits, Text mining in cancer gene and pathway prioritization, Cancer Inform. 69 (2014).

[14] R. Gray, Vector quantization, IEEE Assp Magazine. 1 (1984) 4–29.

[15] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometr. Intell. Lab. Syst. 2 (1987) 37–52.

[16] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–791.

[17] A. Frigyesi, M. Hoglund, Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes, Cancer Inform. 6 (2008) 275–292.

[18] J.-P. Brunet, P. Tamayo, T.R. Golub, J.P. Mesirov, Metagenes and molecular pattern discovery using matrix factorization, Proc. Natl. Acad. Sci. 101 (2004) 4164–4169.

[19] L.B. Alexandrov, S. Nik-Zainal, D.C. Wedge, S.A. Aparicio, S. Behjati, A.V. Biankin, et al., Signatures of mutational processes in human cancer, Nature 500 (2013) 415–421.

[20] M. Hofree, J.P. Shen, H. Carter, A. Gross, T. Ideker, Network-based stratification of tumor mutations, Nat. Methods 10 (2013) 1108–1115.

[21] Y. Luo, C. Mao, Y. Yang, F. Wang, F.S. Ahmad, D. Arnett, et al., Integrating hypertension phenotype and genotype with hybrid non-negative matrix factorization, Bioinformatics (2018).

[22] P. Cingolani, A. Platts, L. Wang le, M. Coon, T. Nguyen, L. Wang, et al., A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3, Fly 6 (2012) 80–92.

[23] K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, Nucl. Acids Res. 38 (2010) e164-e.

[24] B. Li, S.M. Leal, Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data, Am. J. Human Genet. 83 (2008) 311–321.

[25] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, Adv. Neural Inform. Process. Syst. (2001) 556–562.

[26] A. Pascual-Montano, P. Carmona-Saez, M. Chagoyen, F. Tirado, J.M. Carazo, R.D. Pascual-Marqui, bioNMF: a versatile tool for non-negative matrix factorization in biology, BMC Bioinf. 7 (2006) 1.

[27] L.N. Hutchins, S.M. Murphy, P. Singh, J.H. Graber, Position-dependent motif characterization using non-negative matrix factorization, Bioinformatics (Oxford, England). 24 (2008) 2684–2690.

[28] A. Frigyesi, M. Höglund, Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes, Cancer Inform. 6 (2008).

[29] S. Vural, X. Wang, C. Guda, Classification of breast cancer patients using somatic mutation profiles and machine learning approaches, BMC Syst. Biol. 10 (Suppl 3) (2016) 62.

[30] K.P. Soh, E. Szczurek, T. Sakoparnig, N. Beerenwinkel, Predicting cancer type from tumour DNA signatures, Genome Med. 9 (2017) 104.

[31] G.M. Higa, The microtubule as a breast cancer target, Breast cancer (Tokyo, Japan) 18 (2011) 103–119.

[32] J.W.S. Law, A.Y.W. Lee, The role of semaphorins and their receptors in gliomas, J. Signal Trans. 2012 (2012).

[33] J.H. Lee, X.M. Zhao, I. Yoon, J.Y. Lee, N.H. Kwon, Y.Y. Wang, et al., Integrative analysis of mutational and transcriptional profiles reveals driver mutations of metastatic breast cancers, Cell Discovery 2 (2016) 16025.

[34] K.Q. Liu, Z.P. Liu, J.K. Hao, L. Chen, X.M. Zhao, Identifying dysregulated pathways in cancers from pathway interaction networks, BMC Bioinf. 13 (2012) 126.

[35] E.Y. Chen, C.M. Tan, Y. Kou, Q. Duan, Z. Wang, G.V. Meirelles, et al., Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool, BMC Bioinf. 14 (2013) 128.

[36] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, et al., The reactome pathway knowledgebase, Nucl. Acids Res. 46 (2018) D649–D655.

[37] M. Kanehisa, Y. Sato, M. Furumichi, K. Morishima, M. Tanabe, New approach for understanding genome variations in KEGG, Nucl. Acids Res. 47 (2019) D590–D595.

[38] T.A. Lehman, B.G. Haffty, C.J. Carbone, L.R. Bishop, A.A. Gumbs, S. Krishnan, et al., Elevated frequency and functional activity of a specific germ-line p53 intron mutation in familial breast cancer, Cancer Res. 60 (2000) 1062–1069.

[39] S. Wang-Gohrke, H. Becher, R. Kreienberg, I.B. Runnebaum, J. Chang-Claude, Intron 3 16 bp duplication polymorphism of p53 is associated with an increased risk for breast cancer by the age of 50 years, Pharmacogenetics 12 (2002) 269–272.

[40] T. Gutschner, S. Diederichs, The hallmarks of cancer: a long non-coding RNA point of view, RNA Biol. 9 (2012) 703–719.

[41] X.M. Zhao, K.Q. Liu, G. Zhu, F. He, B. Duval, J.M. Richer, et al., Identifying cancer-related microRNAs based on gene expression data, Bioinformatics (Oxford,

England) 31 (2015) 1226–1234.

[42] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, et al., PLINK: a tool set for whole-genome association and population-based linkage analyses, Am. J. Human Genet. 81 (2007) 559–575.

[43] B.E. Madsen, S.R. Browning, A groupwise association test for rare mutations using a weighted sum statistic, PLoS Genet. 5 (2009) e1000384.

[44] F. Han, W. Pan, A data-adaptive sum test for disease association with multiple common or rare variants, Hum. Hered. 70 (2010) 42–54.

[45] M.C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, X. Lin, Rare-variant association testing for sequencing data with the sequence kernel association test, Am. J. Human Genet. 89 (2011) 82–93.

[46] S. Lee, M.C. Wu, X. Lin, Optimal tests for rare variant effects in sequencing association studies, Biostatistics (Oxford, England) 13 (2012) 762–775.

[47] J.N. Hirschhorn, M.J. Daly, Genome-wide association studies for common diseases and complex traits, Nat. Rev. Genet. 6 (2005) 95–108.

[48] Y. Luo, F.S. Ahmad, S.J. Shah, Tensor factorization for precision medicine in heart failure with preserved ejection fraction, J. Cardiovasc. Transl. Res. (2017) 1–8.