Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 55-20 (2022) 641–646

Genome-scale Metabolic Model Guided Subtyping Lung Cancer towards Personalized Diagnosis

Ezgi Tanıl, Nehir Kızılilsoley, Emrah Nikerel*

**Department of Genetics and Bioengineering, Yeditepe University, İstanbul, Turkey*
(Tel: +90216 5780619; e-mail: emrah.nikerel@yeditepe.edu.tr).

Abstract: Mathematical modeling of biological systems are useful in (i) gaining better understanding about the physiological processes in an organism, (ii) simulating alternative scenarios, (iii) finding targets for improved performance within metabolic engineering context (iv) performing several functional analyses, e.g. identify drug targets (v) process scheduling within the context of industrial biotechnology etc. Increasing the predictive capability of these models is of common interest within systems biology studies which allows identification of more effective and personalized treatment strategies for complex metabolic diseases such as cancer by investigation of disease metabolism and providing correct subtyping and staging. By transforming gene-level information to flux/metabolite level information, current disease state can be analyzed and diagnosis of cancer subtype can be performed using a less invasive methods.

In this study, subtyping and staging of lung cancer, that is one of the main causes of cancer related deaths, was performed by integrating publicly available RNAseq data of normal, lung adenoma and adenocarcinomas and lung squamous cell neoplasms to human genome scale metabolic model and classification of obtained flux distributions using linear support vector machine (SVM) classifications. Differential flux analysis and pathway enrichment methods showed that model adequately represented tumour metabolism. SVM classification accuracies were calculated as more than 99% for normal and cancer cells and 94% for adenomas and adenocarcinomas and squamous cell neoplasms, indicating high predictive capability of flux distributions.

Keywords: Subtyping and staging, lung cancer, modeling, bioinformatics, genome scale metabolic model, personalized medicine, machine learning, statistical analysis, biological network models

1. INTRODUCTION

Mathematical models are key tools to investigate underlying mechanisms of complex metabolic diseases including cancer. Previous efforts on construction of human genome scale models and integration of "-omic" data allows building patient/context/tissue specific constraint-based models which are very useful in predicting flux distributions at steady state. Flux distributions at a certain condition can be obtained by Flux Balance Analysis (FBA), which optimizes flux rates for given objective (Ghaffari, Mardinoglu and Nielsen, 2015). Typically, the solution space of FBA is not unique because of the high degrees of freedom. To address this challenge, different solving methods such as Flux Variability Analysis (FVA) which allows analysis of all solutions in the solution space (Gudmundsson and Thiele, 2010) and parsimonious FBA (pFBA) which calculates most efficient flux distribution through the objective (Lewis *et al.*, 2010). Integration of omic data to genome scale metabolic models is another method to decrease solution space. Due to the ease of collecting cancer related high throughput omic data, fluxome analysis methods based on integration of omic data to genome scale metabolic models (GSMMs) gained more importance in studying complex diseases. Tumour cells shown to have altered metabolism and the most commonly defined example is the Warburg effect. In this phenomenon, tumour cells shown to use lactic acid fermentation instead of oxidative phosphorylation due to the increased energy demand caused by fast proliferation rate (Warburg, 1956). GSMMs became an important tool for simulation and analysis of such metabolic alterations. Using this method, tissue/disease/condition specific models can be obtained and flux distributions at different conditions can be analysed (Reed, 2012; Schmidt *et al.*, 2013), even, personal fluxome profile of the patient can be determined and reporter reactions specific to cancer subtype can be identified (Ghaffari, Mardinoglu and Nielsen, 2015). Transcriptome data, assuming that gene expression levels are related to enzyme levels, can be used to constrain GSMMs. In this context, most commonly used methods to obtain flux distributions correlated to transcriptome data were Lee12 (Chung and Lee, 2012) or E-flux (Colijn *et al.*, 2009) Additionally, context/tissue/patient specific models can be generated by removing unused pathways or reactions using methods such as GIMME (Becker and Palsson, 2008), INIT (Agren *et al.*, 2012) or iMAT (Zur, Ruppin and Shlomi, 2010).

Machine learning became a popular tool for cancer researchers for accurate classification of complex data sets. They simply "learn" a model from provided data to predict the future observations. Support Vector Machines (SVM) is one of these commonly used classification methods. It is a

10.1016/j.ifacol.2022.09.168

supervised machine learning algorithm and designed to determine a decision boundary (optimal hyper plane) between two or more data categories that separates the data points linearly which maximizes the margin to decrease the classification error (Huang *et al.*, 2018). Numerous applications of SVM in cancer classification were present in the literature for different types of cancer using such as transcriptome (Golub *et al.*, 1999; Furey *et al.*, 2000; Moler, Chow and Mian, 2000; Li, Zhang and Ogihara, 2004), DNA methylation (Model *et al.*, 2001; Kim, 2016; Yang *et al.*, 2017), proteome (Tyanova *et al.*, 2016), copy number variations (Rapaport, Barillot and Vert, 2008), single nucleotide polymorphisms (Vural, Wang and Guda, 2016) and even multi-omic data sets (Kim *et al.*, 2017; Lin and Lane, 2017). Recently, Xie et al. published a study which analyses different ML algorithms including SVM from for early diagnosis of lung cancer using metabolome data (Xie *et al.*, 2021).

Lung cancer is the main cause of cancer related deaths, with more than 1.3 million deaths per year, especially for men (17% of the total new cancer diagnosis and 23% of the cancer related deaths) (Jemal *et al.*, 2011). 5 year survival rate of lung cancer is around 15% due to the difficulties in differentiation of lung cancer from other pulmonary diseases. Survival rate may increase to 85% if early diagnosis is possible (Callejon-Leblic *et al.*, 2016). Therefore, diagnosis of lung cancer is crucial for application of effective, personalized treatment. Additionally, accurately determining subtype and stage of the cancer using less invasive methods may allow development of effective and patient specific treatment strategies, even allowing screening programs nation or population wide. Numerous subtyping and staging reports were available in literature which are based on differential gene expression, mutations, copy number variation and metabolite excretion (Cai *et al.*, 2015; Chan and Hughes, 2015; Wikoff *et al.*, 2015; Callejon-Leblic *et al.*, 2016; Podolsky *et al.*, 2016; Moreno *et al.*, 2018; Xie *et al.*, 2021). Although genetic biomarkers such as PIK3CA, TP53, PTEN, KRAS and EGFR were associated with lung cancer, the expression profiles of the genes are different for each subtype and stage.

In this study, early stage lung cancer transcriptome data for normal and 2 lung cancer subtypes (adenoma and adenocarcinomas (AD) and squamous cell neoplasms (SC)) from literature are integrated to human genome scale metabolic model (Recon3D). Differential flux and expression analyses were performed to identify reporter genes and fluxes for each subtype. SVM algorithm was used to classify normal, AD and SC samples with minimum number of predictors and highest accuracy. Finally, functional analysis such as Gene Ontology analysis, pathway enrichment for hunting hallmarks of cancer were performed.

2. METHOD

The workflow of the study is presented in Fig 1. Differential expression and Gene Ontology (GO) analysis were performed in R. Calculations were performed in MATLAB r2018a equipped with COBRA Toolbox (Heirendt *et al.*, 2019) and Gurobi 9.5 for flux balance analysis.

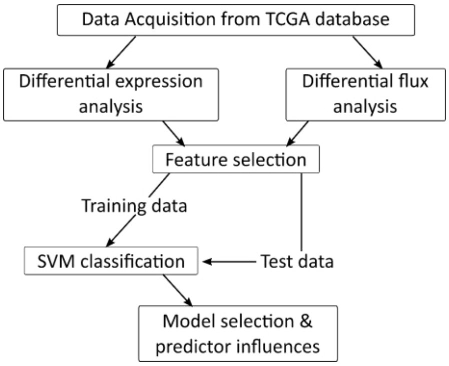
Data Acquisition from TCGA database
Differential expression analysis
Differential flux analysis
Feature selection
Training data
SVM classification
Test data
Model selection & predictor influences

Fig. 1. Workflow of the study.

2.1 Data Acquisition and Differential Expression Analysis

Normal (108 samples) and lung cancer high-throughput sequencing (htseq-count) data, for both adenoma and adenocarcinoma (535 samples) and squamous cell neoplasms (502 samples) were collected from The Cancer Genome Atlas Program (TCGA) Research Network (https://www.cancer.gov/tcga). Raw counts were filtered and normalized, differential expression analysis was performed to identify genes that are expressed in significantly different levels between normal and tumor tissues (N-C) and adenoma and adenocarcinomas and squamous cell neoplasms (A-S) using DESeq2 (Love, Huber and Anders, 2014). Genes with absolute log-2 ratio of gene expression values between 2 groups higher than 2 ($|\log_2\text{FoldChange}| > 2$) with p-value lower than 0.01 were considered as significantly different between two conditions. Gene Ontology (GO) analysis (Yu *et al.*, 2012) was performed to identify the biological processes involving significantly expressed genes.

2.2 Integration of Transcriptome Data to GSMM and Flux Balance Analysis (FBA)

Normalized transcriptome data using DESeq2, were integrated to Human Recon3D (Brunk *et al.*, 2018) GSMM version published by Lewis *et al.* (Lewis *et al.*, 2021) containing 13634 reactions and 8457 metabolites (S in (1) (13634x8457)) which has corrected energy generating cycles, using E-flux (Colijn *et al.*, 2009) method, based on the assumption that fluxes are correlated to the expression of corresponding genes. According to the transcriptome data, reaction expression values (a_j, b_j in (1)) that will be used as flux bounds that were mapped benefiting Gene-Protein-Reaction (GPR) association rules. GPR rules are defined in the model describing reactions catalysed by enzymes encoded by a single gene and multiple genes encoding enzymatic subunits and isozymes. Overall expression levels for fluxes were calculated using 'minSum' option which uses the minimum expression value for enzymes containing subunits and sum of expression values of all isozymes and then set as upper and lower bounds. FBA was performed to obtain flux distributions (v in (1)) by maximizing growth, which can be used as objective for cancer cells because their main aim is proliferate and invade other tissues (Damiani *et al.*, 2017; Nilsson and Nielsen, 2017).

$$\begin{aligned} &\max \quad c^T v \\ &\text{subject to} \begin{cases} S \cdot v = 0 \\ a_j \leq v_j \leq b_j \end{cases} \end{aligned} \qquad (1)$$

2.3 Differential Flux Analysis and Pathway Enrichment

Differential flux analysis was performed using non-parametric Wilcoxon rank sum test and FDR correction, by comparing distributions and expression of each flux for normal and tumour samples (N-C) and adenomas and adenocarcinomas and squamous cell neoplasms (A-S) (Nanda and Ghosh, 2020). Fluxes that have p-values lower than 0.05 were determined to be significantly different between two conditions. Pathway enrichment was performed using hypergeometric test employing subsystems defined in model. This test is commonly used to determine whether a property is overly represented or not in selected sample from population. Subsystems that have p-values lower than 0.1 were determined to be overrepresented.

2.4 Linear SVM Classification for N-C and A-S

A cascaded-tree like subtyping approach was performed by first classifying normal (N) and tumour cells (C). Then tumour cells were classified into subtypes. Using fluxome data randomly selected 80 normal samples and 850 cancer samples (420 samples for adenoma and adenocarcinoma (AD) and 430 samples for squamous cell neoplasms (SC)) were selected to train SVM model with 10 fold cross-validation. Remaining data were used to test the model. Reactions shown to carry significant fluxes between C/N and SC/AD ($p<0.05$, $|\log_2\mathrm{FoldChange}|>1$) were chosen as potential predictors for classification. Then, to decrease the number of predictors, effect of each predictor on validation accuracy were calculated. This process is applied for both N-C and A-S classification to obtain models that give maximum accuracy with minimum number of predictors.

3. RESULTS AND DISCUSSION

3.1 Differential Expression and GO Analysis

Differential expression analysis was performed to determine the genes that have significant difference between normal and cancer cells and adenomas and adenocarcinomas and squamous cell neoplasms. Out of 60483 genes, 5998 and 4074 genes were found to have significantly different expression levels between N-C and AD-SC, respectively. GO analysis results for the gene sets were shown in Fig 2 which indicates that, mostly genes that are significantly different in each case belong to processes of either cellular structure formation, cell cycle (DNA replication, packaging, chromosome assembly etc.) or signalling which is parallel to the hallmarks of cancer (Hanahan and Weinberg, 2011).

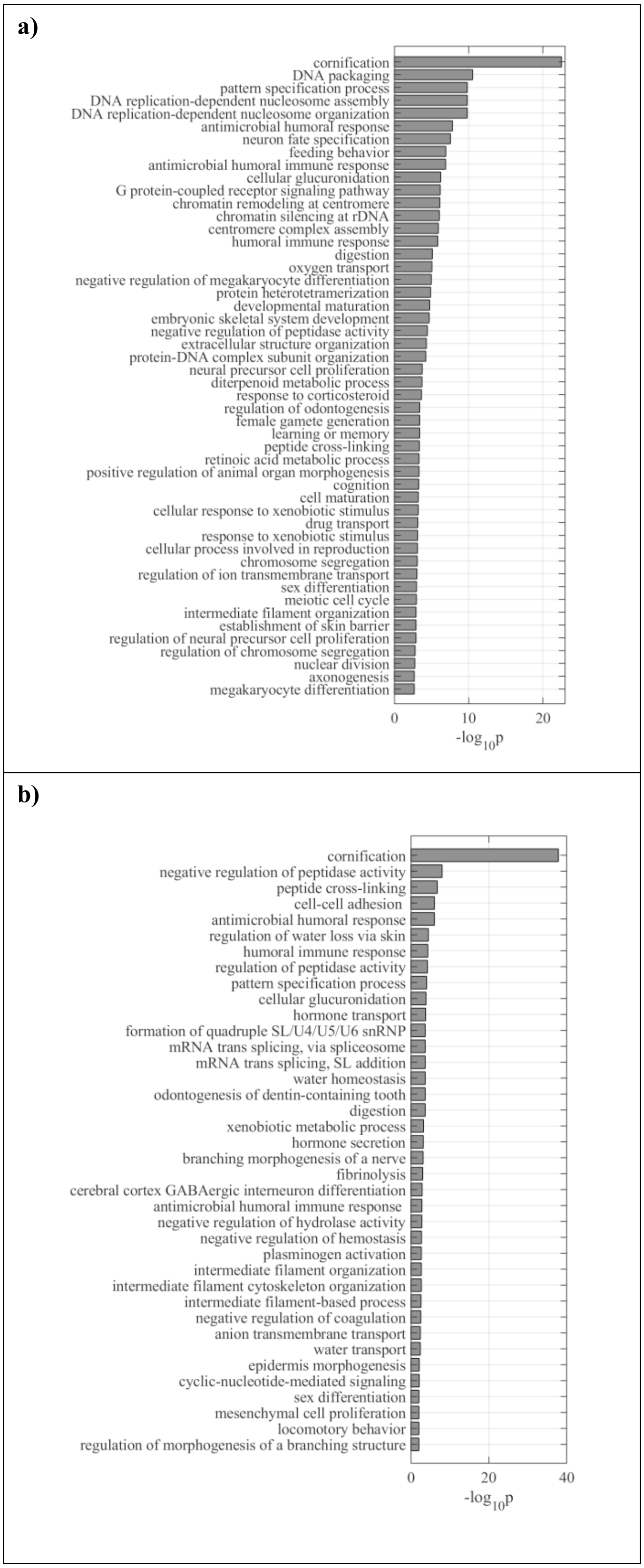

Fig. 2. GO analysis results for **a)** N-C and **b)** AD-SC.

Additionally, since these profiles do not represent the disease state only, unrelated mechanisms such as sex differentiation, regulation of water loss via skin were found to be significant. In conclusion, although transcriptomic analysis gives potential genetic biomarkers, these are not lung cancer specific and they cannot be measured easily.

3.2 Differential Flux Analysis and Pathway Enrichment

Wilcoxon rank sum test and hypergeometric test were performed to analyse statistical significance of the differences flux distributions and altered subsystems between different cell types. Pathway enrichment results for normal and cancer cells and subtypes and the corresponding p-values were shown in Fig 3. 292 reactions were found to have significantly different fluxes between normal and cancer cells, 103 reactions were found to have significantly different fluxes between adenoma and adenocarcinomas and squamous cell neoplasms.

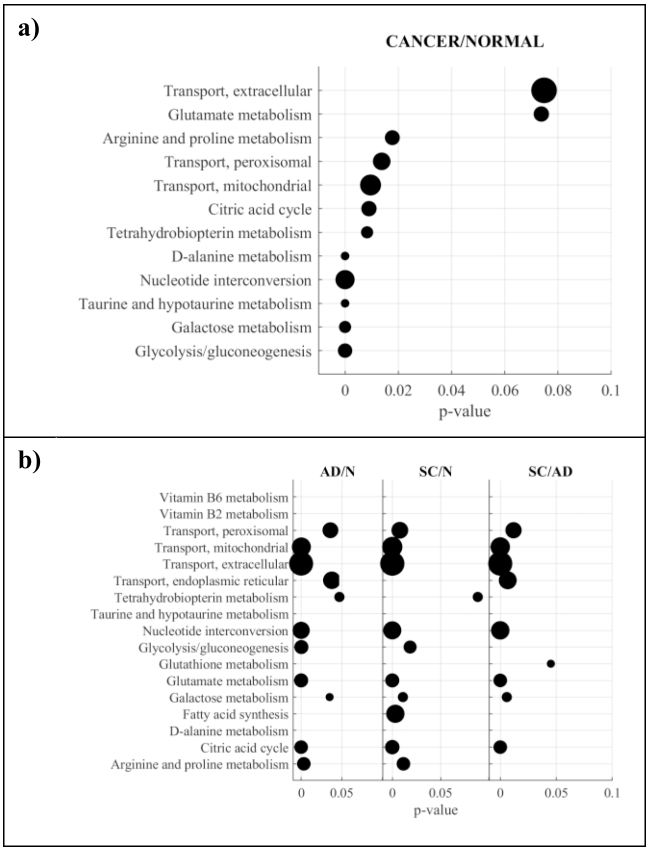

Fig. 3. Pathway enrichment results for **a)** cancer-normal groups **b)** subtypes. Dot sizes are proportional to number of fluxes involved in the subsystem.

Model currently represents tumour microenvironment as a single cell since transcriptome data is obtained from a tumour tissue sample, not the tumour cell itself. Cancer progress can be characterized by shifting towards oxidative stress due to uncontrolled growth. In this context, elevated flux through reduced glutathione exchange can represent the protection mechanisms of cancerous cells from reactive oxygen species. Changes in glycolysis and citric acid cycle are another mechanism reported to be affected during reprogramming of metabolism in cancer cells (Warburg effect). Additionally, amino acid metabolism were reported to be altered since they are used as growth factors such as glutamate (Stepulak *et al.*, 2014) and providing additional energy for uncontrolled proliferation. Oxidative stress in cancerous cells also affects lipid metabolism and oxidation pathways which are mainly involved in signalling and proliferation and metastasis. Inositol and fatty acid metabolisms were reported to be altered in highly metastatic cancer cell lines which is in line with our results.

Another hallmark of cancer is tissue inflammation. Tissue inflammation in lung cancer can be mediated by leukotrienes which are biologically active eicosanoid lipid mediators secreted by immune cells in tumour microenvironment. Elevated levels leukotriene B4 (LTB4) was reported to be observed in lung cancer. Also LTB4 enhances angiogenesis and blood vessel permeability (Salvador *et al.*, 2017; Tian *et al.*, 2020).

Nucleotide and cofactor metabolisms are other altered metabolisms in cancer cells. They have a key role in DNA synthesis and repair and increasing proliferation rate increases the nucleotide and NAD+ demand. In this context, inosine, adenosine, nicotinate, riboflavin concentrations were reported to be increased in various cancer cells (Hsu and Sabatini, 2008).

Overall, considering 12 hallmarks of cancer and previous transcriptome and metabolome markers published in literature (Hsu and Sabatini, 2008), it was shown that model reflected the biochemistry of the tumour environment qualitatively.

3.3 Linear SVM classification

Classification capability of significantly different fluxes were tested with linear SVM classification. In classification of normal and cancer cells, accuracy of 0.991 was obtained using all 292 reactions as predictors. When number of predictors increased one by one choosing the ones that have highest validation accuracy, final accuracy of 0.995 was obtained for 7 predictors as shown in Fig 4.a. For adenomas and adenocarcinomas and squamous cell neoplasms classification, 0.931 accuracy was obtained when all 103 fluxes were used as predictors. On the other hand, 0.94 accuracy was obtained using 13 predictors as shown in Fig 4.b. Overall, classification results shown that model flux predictions enables adequate classification. Also, flux predictions were shown to have potential to be used for enhancing accuracy and precision of classifications performed using transcriptome data.

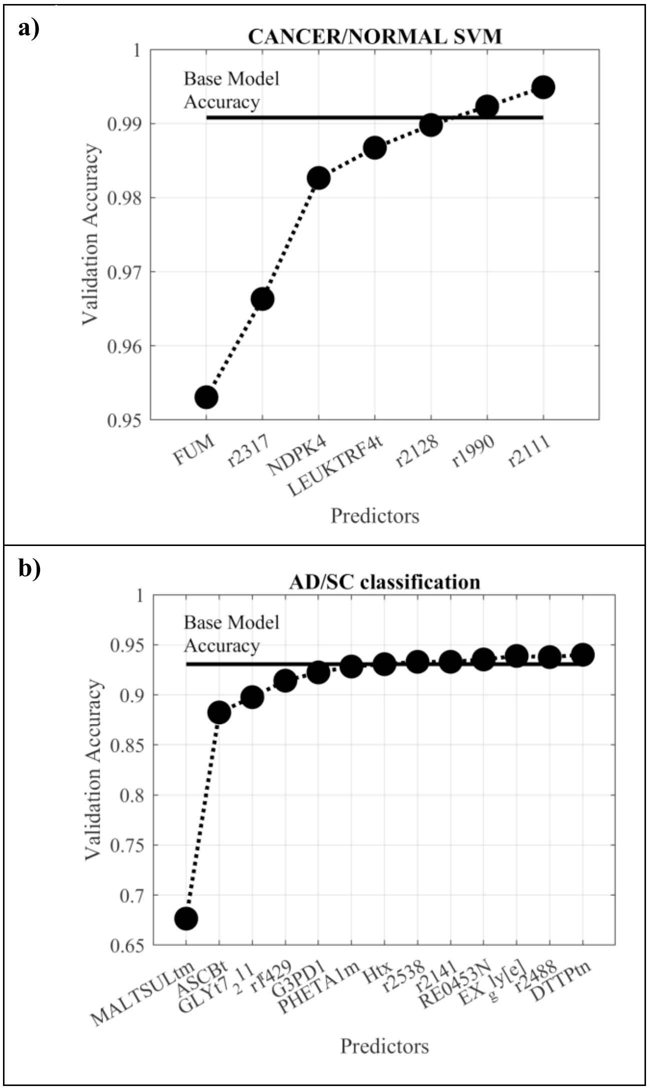

Fig. 4. Classification accuracies calculated cumulatively for **a)** cancer-normal groups **b)** subtypes.

4. CONCLUSIONS

Personalized medicine requires correct subtyping of complex diseases such as lung cancer using methods less invasive as possible. This application illustrates genome scale metabolic model-based subtyping among different cell types or disease conditions. By incorporating models of metabolism allows diagnosis with non-invasive methods, by transforming transcriptome level information to more accessible metabolite level information.

5. ACKNOWLEDGEMENTS

This work is funded by Health Institutes of Turkey (TÜSEB) project no: 2019-TA-01-4589.

REFERENCES

Agren, R. *et al.* (2012) ‘Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT’, *PLoS Comput Biol*, 8(5), p. e1002518.

Becker, S. A. and Palsson, B. O. (2008) ‘Context-specific metabolic networks are consistent with experiments’, *PLoS Comput Biol*, 4(5), p. e1000082.

Brunk, E. *et al.* (2018) ‘Recon3D enables a three-dimensional view of gene variation in human metabolism’, *Nature biotechnology*, 36(3), p. 272.

Cai, Z. *et al.* (2015) ‘Classification of lung cancer using ensemble-based feature selection and machine learning methods’, *Molecular BioSystems*, 11(3), pp. 791–800.

Callejon-Leblic, B. *et al.* (2016) ‘Metabolic profiling of potential lung cancer biomarkers using bronchoalveolar lavage fluid and the integrated direct infusion/gas chromatography mass spectrometry platform’, *Journal of proteomics*, 145, pp. 197–206.

Chan, B. A. and Hughes, B. G. M. (2015) ‘Targeted therapy for non-small cell lung cancer: current standards and the promise of the future’, *Translational lung cancer research*, 4(1), p. 36.

Chung, B. K. S. and Lee, D. Y. (2012) ‘Computational codon optimization of synthetic gene for protein expression’, *BMC Systems Biology*, 6(1), p. 1.

Colijn, C. *et al.* (2009) ‘Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production’, *PLoS Comput Biol*, 5(8), p. e1000489.

Damiani, C. *et al.* (2017) ‘A metabolic core model elucidates how enhanced utilization of glucose and glutamine, with enhanced glutamine-dependent lactate production, promotes cancer cell growth: The Warburg effect’, *PLoS computational biology*, 13(9), p. e1005758.

Furey, T. S. *et al.* (2000) ‘Support vector machine classification and validation of cancer tissue samples using microarray expression data’, *Bioinformatics*, 16(10), pp. 906–914.

Ghaffari, P., Mardinoglu, A. and Nielsen, J. (2015) ‘Cancer metabolism: a modeling perspective’, *Frontiers in physiology*, 6, p. 382.

Golub, T. R. *et al.* (1999) ‘Molecular classification of cancer: class discovery and class prediction by gene expression monitoring’, *science*, 286(5439), pp. 531–537.

Gudmundsson, S. and Thiele, I. (2010) ‘Computationally efficient flux variability analysis’, *BMC Bioinformatics*, 11(2), pp. 2–4. doi: 10.1186/1471-2105-11-489.

Hanahan, D. and Weinberg, R. A. (2011) ‘Hallmarks of cancer: the next generation’, *cell*, 144(5), pp. 646–674.

Heirendt, L. *et al.* (2019) ‘Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v. 3.0’, *Nature protocols*, 14(3), pp. 639–702.

Hsu, P. P. and Sabatini, D. M. (2008) ‘Cancer cell metabolism: Warburg and beyond’, *Cell*, 134(5), pp. 703–707.

Huang, S. *et al.* (2018) ‘Applications of support vector machine (SVM) learning in cancer genomics’, *Cancer Genomics-Proteomics*, 15(1), pp. 41–51.

Jemal, A. *et al.* (2011) ‘Global cancer statistics.’, *CA: a cancer journal for clinicians*, 61(2), pp. 69–90.

Kim, S. (2016) ‘Weighted K-means support vector machine for cancer prediction', *Springerplus*, 5(1), pp. 1–11.

Kim, S. *et al.* (2017) 'Meta-analytic support vector machine for integrating multiple omics data', *BioData mining*, 10(1), pp. 1–14.

Kuepfer, L., Sauer, U. and Blank, L. M. (2005) 'Metabolic functions of duplicate genes in Saccharomyces cerevisiae', *Genome research*, 15(10), pp. 1421–1430.

Lewis, J. E. *et al.* (2021) 'Personalized genome-scale metabolic models identify targets of redox metabolism in radiation-resistant tumors', *Cell Systems*, 12(1), pp. 68–81.

Lewis, N. E. *et al.* (2010) 'Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models', *Molecular systems biology*, 6(1), p. 390.

Li, T., Zhang, C. and Ogihara, M. (2004) 'A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression', *Bioinformatics*, 20(15), pp. 2429–2437.

Lin, E. and Lane, H.-Y. (2017) 'Machine learning and systems genomics approaches for multi-omics data', *Biomarker research*, 5(1), pp. 1–6.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome biology*, 15(12), pp. 1–21.

Model, F. *et al.* (2001) 'Feature selection for DNA methylation based cancer classification', *Bioinformatics*, 17(suppl_1), pp. S157–S164.

Moler, E. J., Chow, M. L. and Mian, I. S. (2000) 'Analysis of molecular profile data using generative and discriminative methods', *Physiological genomics*, 4(2), pp. 109–126.

Moreno, P. *et al.* (2018) 'Metabolomic profiling of human lung tumor tissues–nucleotide metabolism as a candidate for therapeutic interventions and biomarkers', *Molecular oncology*, 12(10), pp. 1778–1796.

Nanda, P. and Ghosh, A. (2020) 'Genome Scale-Differential Flux Analysis reveals deregulation of lung cell metabolism on SARS Cov2 infection', *bioRxiv*.

Nilsson, A. and Nielsen, J. (2017) 'Genome scale metabolic modeling of cancer', *Metabolic engineering*, 43, pp. 103–112.

Podolsky, M. D. *et al.* (2016) 'Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels', *Asian Pacific Journal of Cancer Prevention*, 17(2), pp. 835–838.

Rapaport, F., Barillot, E. and Vert, J.-P. (2008) 'Classification of arrayCGH data using fused SVM', *Bioinformatics*, 24(13), pp. i375–i382.

Reed, J. L. (2012) 'Shrinking the metabolic solution space using experimental datasets', *PLoS Comput Biol*, 8(8), p. e1002662.

Salvador, M. M. *et al.* (2017) 'Lipid metabolism and lung cancer', *Critical Reviews in Oncology/Hematology*, 112, pp. 31–40.

Schmidt, B. J. *et al.* (2013) 'GIM3E: condition-specific models of cellular metabolism developed from metabolomics and expression data', *Bioinformatics*, 29(22), pp. 2900–2908.

Stepulak, A. *et al.* (2014) 'Glutamate and its receptors in cancer', *Journal of neural transmission*, 121(8), pp. 933–944.

Tian, W. *et al.* (2020) 'Leukotrienes in Tumor-Associated Inflammation', *Frontiers in Pharmacology*, 11, p. 1289.

Tyanova, S. *et al.* (2016) 'Proteomic maps of breast cancer subtypes', *Nature communications*, 7(1), pp. 1–11.

Vural, S., Wang, X. and Guda, C. (2016) 'Classification of breast cancer patients using somatic mutation profiles and machine learning approaches', *BMC systems biology*, 10(3), pp. 263–276.

Warburg, O. (1956) 'On the origin of cancer cells', *Science*, 123(3191), pp. 309–314.

Wikoff, W. R. *et al.* (2015) 'Metabolomic markers of altered nucleotide metabolism in early stage adenocarcinoma', *Cancer prevention research*, 8(5), pp. 410–418.

Xie, Y. *et al.* (2021) 'Early lung cancer diagnostic biomarker discovery by machine learning methods', *Translational oncology*, 14(1), p. 100907.

Yang, Z. *et al.* (2017) 'Classification based on feature extraction for hepatocellular carcinoma diagnosis using high-throughput dna methylation sequencing data', *Procedia Computer Science*, 107, pp. 412–417.

Yu, G. *et al.* (2012) 'clusterProfiler: an R package for comparing biological themes among gene clusters', *Omics: a journal of integrative biology*, 16(5), pp. 284–287.

Zur, H., Ruppin, E. and Shlomi, T. (2010) 'iMAT: an integrative metabolic analysis tool', *Bioinformatics*, 26(24), pp. 3140–3142.