Differential Gene Expression Analysis of Non-Small Cell Lung Cancer Samples to Classify Candidate Genes

Neelambika B. Hiremath
Department of Computer Science and Engineering, JSS Academy of Technical Education, Bengaluru, India
neelambika@ieee.org

Pruthviraja Dayananda
Department of Information Science and Engineering, JSS Academy of Technical Education, Bengaluru, India
dayanandap@gmail.com
(corresponding author)

Received: 9 February 2023 | Revised: 25 February 2023 | Accepted: 4 March 2023

ABSTRACT

Differential gene expression is an analysis of gene data, in which the RNA sequence data after next-generation sequencing are to be visualized for any quantitative changes in the levels of the experimental data set. This work aims to derive the transcript statistics on a gene transcript file with a fold change of genes on a normalized scale, in order to identify quantitative changes in gene expression of the difference between the reference genome and Non-Small Cell Lung Cancer (NSCLC) samples. This insight makes a clinical impact in assessing and characterizing candidate genes. The pipeline comprises tuxedo protocol and programming language R with the standard ballgown package. The resultant data set and the plot displays depict the candidate genes in their respective location which are significant in expressing their changes in NSCLC samples. The samples are compared with prominent gene labels of NSCLC samples. The results explain the differential expression of particular samples across samples from both genders.

Keywords-differential gene expression; next-generation sequencing; RNA sequence; machine learning; non-small cell lung cancer; classification

I. INTRODUCTION

Lung cancer is the leading type of cancer worldwide. About 1.76 million deaths were caused by lung cancer during 2019. The disease is caused by lung malignancies. There are a lot of advances in lung cancer therapies, but, despite the advancement the prognosis of the disease is unfavourable [1], because after the diagnosis the survival rate is less than 15% for the next five years. The understanding of molecular characteristics leading to identifying cancer will speed up the diagnosis. Cancer is one of the most challenging diseases [2]. The recent advancement of next-generation sequencing of the human gene has created opportunities to use gene expression information for decision making [3]. Lung cancer can be categorised into NSCLC, which are about the 85% of the cases, and as small cell lung cancer. The study of molecular signature shows that carcinogenesis is caused by oncogenic drivers. This has led the research to oncogenic driver identification rather than clinical parameter studies [4]. The most common types of NSCLC are squamous cell carcinoma, large-cell carcinoma, and adenocarcinoma, with several other types occurring less frequently. All types can occur in unusual histologic variants and as mixed cell-type combinations [5]. The lack of early detection of NSCLC leads to an increased mortality rate, specifically in developing countries. Cancer is globally considered as a leading cause of death [6]. It was reported that 1 in 6 deaths in 2018 happened due to cancer. The lung cancer accounts for about 2.09 million cases worldwide. In India, about 70 thousand cases of lung cancer are reported every year [7]. Depending on the stage of cancer, the general condition of the person, age, reaction to chemotherapy, and other considerations, such as the possible side effects of the procedure, more than one methods of treatment are used. Usually, NSCLC patients are categorized as patients with early, non-metastatic disease (stages I and II, and pick type III tumors) and patients with locally advanced thoracic cavity-bound disease (e.g. large tumors). Several facets of cancer science are now being reshaped by the many implementations and uses of Next Generation Sequencing (NGS) technologies. As a result of aberrant mutagenesis and its contribution to

carcinogenesis, these technologies have enabled researchers to explain improvements in any degree of regulation carried out in a cell [8].

Usage of RNA sequencing to understand and characterize the genes involved in cancer diagnostics and therapy has become a well-used tool. The reduced cost of sequencing and its advantages over microarrays have made it accessible to everyone, which is improving the time taken to detect and treat cancer patients. RNA sequencing provides a view of the transcriptome of gene incomprehensive structure [9]. It is not dependent on any prior sequence knowledge and it can detect structural variations such as alternative splicing events and gene fusion, but the data storage and the analysis are more complex and it does not use any standard protocol, while it is also an expensive process. The most significant factor in NSCLC is that its identification is delayed, causing NSCLCs to have the lowest survival rate. Researchers have concluded that faster diagnosis happens by understanding gene expression at a molecular level, as gene sequence expression plays a very significant role in NSCLCs [10]. As the latest NGS sequencing provides us with a comprehensive genome structure with high throughput, identifying the candidate genes by analyzing the RNA sequence data set allows focusing on those sets of genes. The candidate genes mark a subset of biomarkers [11] or signatures of this biological condition. The present study aims to analyze and map the significant mutations in the genes responsible for NSCLC, something that might help biomarker identification, leading to early detection and helping adjuvant therapies in personalised medicine which increases survival rate. The DNA analysis at nucleotide level is a new research area [12], and the research on the identification of biomarkers not only supports the medication treatment, but also predicts the proteins which create and activate post gene expression [13].

II. MATERIALS AND METHODS

A. Datasets

This work is based on real biological samples' data, the dataset is available on the Sequence Read Archive database maintained by the National Centre for Biotechnology Information (NCBI). Data from the project SRP117020 were selected for analyzing representative samples on both genders with age bigger than 50 years. The data consisted of RNA sequences with the distribution of poor to well-differentiated adenocarcinomas and squamous cell cancers which were sequenced using Illumina Hiseq2500 [14]. The details of the sequence run according to the accession number are shown in Table I.

TABLE I. DETAILS ABOUT SEQUENCE ARCHIVE RUN (SRR) ACCESSION, WITH ACTUAL SEQUENCING DATA FROM THE PARTICULAR SEQUENCING EXPERIMENT [15]

Run	Average spot length	Bases – giga basepairs	Size	Histology	Date published	Access type	Gender (age > 50 years)
SRR6013475	199	6.3Gbp	3.86Gb	Squamous cell carcinoma	2018-08-03	Public	Male
SRR6013476	199	6.16Gbp	3.84Gb	Squamous cell carcinoma	2018-08-03	Public	Male
SRR6013477	199	5.46Gbp	3.33Gb	Adenocarcinoma	2018-08-03	Public	Male
SRR6013479	199	6.00Gbp	3.65Gb	Adenocarcinoma	2018-08-03	Public	Male
SRR6013492	199	5.56Gbp	3.38Gb	Adenocarcinoma	2018-08-03	Public	Female
SRR6013502	199	6.82Gbp	4.20Gb	Adenocarcinoma	2018-08-03	Public	Female
SRR6013508	199	5.56Gbp	3.45Gb	Adenocarcinoma	2018-08-03	Public	Female
SRR6013509	197	7.97Gbp	4.86Gb	Adenocarcinoma	2018-08-03	Public	Female

B. Data Analysis

The analysis of the dataset was carried out by raw sequencing data in fastq format. The reference genome was obtained from the assembly resources of the NCBI.

1) Quality Check

The quality of the RNAseq reads was validated with the FastQC software [16].

2) Read Alignment and Assembly of Transcripts

RNA-seq reads were aligned to the human reference genome using a fast and sensitive alignment programme called HISAT2 [17]. The visual exploration and differences in the expressions were obtained using the Ballgown R package. The experiment is carried out using a standard protocol of tuxedo suite tools which is useful in the analysis of RNA-seq data. The protocol is described by different processes which are more convenient in analyzing raw sequences [18] of large data. The detailed flow chart of the processes is depicted in Figure 1. The hardware environment utilized to run the tuxedo protocol was a 64-bit computer run in Linux environment with 8Gb RAM.

3) Differential Gene Expression and Pathway Analysis

The transcripts and expression levels obtained from Stringtie were subjected to the differentially expressed genes which were performed using the Ballgown package [18]. The package uses statistical methods to get the differentially expressed genes. The obtained list of all the differentially expressed genes was subjected to pathway and gene ontology analysis. This was carried out using KOBAS 3.0 annotation module [19]. The annotation module accepts gene sequences in FASTA format or Gene ID as input and presently covers 5944 different species to run the annotations.

4) Enrichment Analysis

From the pathway analysis results, only genes involved in NSCLC disease pathway were taken and functional enrichment analysis was performed using KOBAS 3.0 enrichment module [19]. Enrichment gives gene ontologies that are significant statistically. This gives an output based on the hypergeometric distribution.

5) Obtaining the Candidate Genes

To obtain the main candidate genes responsible for NSCLC, the mutations responsible for NSCLC were obtained using the ClinVar database [20]. The genes involved in mutations were taken into account and their corresponding expression values obtained from the ballgown package were evaluated to shortlist the candidate genes.

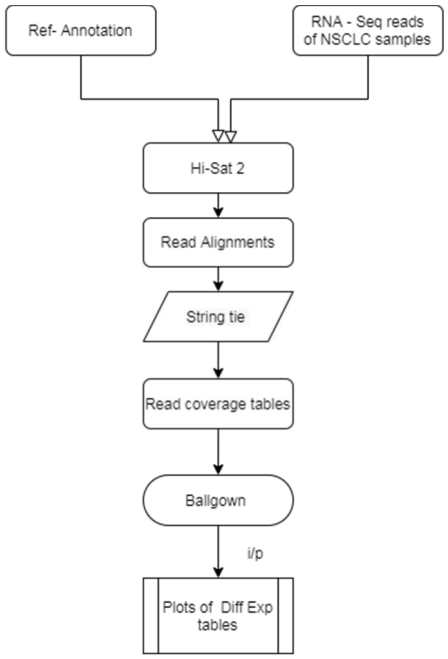

Fig. 1. The protocol to produce tables and plots and obtain differentially expressed genes.

III. RESULTS

A. Data Analysis

The experimental data were analyzed as shown in Table II.

1) Quality Check

Analysis of the raw reads in FastQC format gave a quick impression of the dataset with various features including sequence quality, GC content, N content, and statistics. A brief overview of the obtained results is shown in Table II.

2) Read Alignment and Assembly of Transcripts

Read alignment of the reference genome is a very important step. When the alignment does not occur properly, transcriptome reconstruction becomes difficult, especially for genes expressed at lower levels. The alignment of the reads to the human reference genome gave SAM files whose alignment rates were above 70% which were later converted to BAM, which is the compressed binary version of SAM, files. When the alignment files were subjected to transcriptome reconstruction using StringTie, annotation files were obtained with all the expression levels of all genes and transcripts.

3) Differential Expression and Pathway Analysis

The relevant packages under Ballgown were loaded, and the phenotype data were loaded in .csv format which contained sample ID and gender of the sample. Ballgown gives two tables with differentially expressed genes and transcripts between genders, giving 3 types of values. One, the fold change value referring to the ratio between expression levels in male and female. The values that are <1 indicate that it is expressed at a lower level and values >1 indicate that it is expressed at a higher level. Second, the p-value to get the idea of spotting data if no difference existed. Third, the q-value which is the adjusted p-value obtained by applying the false discovery rate. The gene abundance, measured in terms of FPKM (fragments per kilobase of the model per million reads mapped) distribution across all samples is visualised in Figure 2.

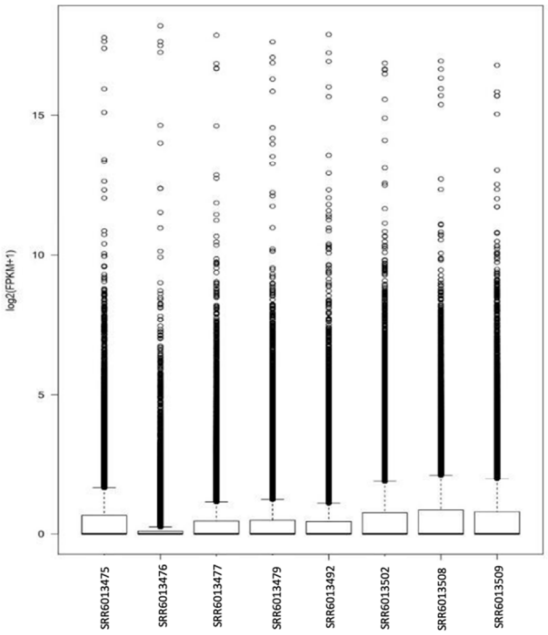

Fig. 2. Distribution of FPKM values across 8 samples.

TABLE II. OVERVIEW OF THE FASTQC RESULTS [15]

	SRR6013475	**SRR6013476**	**SRR6013477**	**SRR6013479**	**SRR6013492**	**SRR6013502**	**SRR6013509**	**SRR6013508**
Basic stats (total sequences)	3,15,05,786	3,08,90,798	2,74,86,767	3,00,87,832	2,79,04,343	3,42,10,392	4,02,79,600	2,78,67,756
Per base sequence quality average score for all bases	32-40	34-40	32-40	34-40	34-40	34-40	34-40	34-39
Sequence quality score (Phred mean)	Ave 2-32	Ave 2-30	Ave 2-32	Ave 2-30	Ave 2-30	Ave 2-30	Ave 2-28	Ave 2-28
Sequence GC content	46	48	47	45	44	45	43	45
Base N content	0-2	Nil	Nil	Nil	Nil	Nil	Nil	Nil

The y-axis shows the log2 transformation of FPKM and the x-axis shows the accession numbers of the data. Additionally, structure and expression levels of isoforms of KRAS were obtained from Ballgown which is shown in Figures 3-4. The KRAS gene belongs to a set of genes known as oncogenes. When mutated, normal cells with these oncogenes have the potential to cause cancer [21]. Figure 3 shows the structure and isoforms of the KRAS gene in sample SRR6013502. The expression levels are shown in varying colors (from yellow to red). The isoform expressed in a higher level is shown in a darker shade. The x-axis indicates the genomic position of the KRAS gene. Figure 4 describes the expressions and structure comparison of KRAS gene in male and female samples. The y-axis indicates the genomic position. The darker color indicates higher expression level which means the KRAS gene is expressed at a higher rate in females than in males. When all the differentially expressed genes were subjected to pathway analysis setting the organism as homo sapiens and the method as gene symbol ID mapping, a total of 5071 genes in different pathways were involved. The results showed different sections of output for a particular gene: the pathways the gene was involved in, the diseases the genes were involved in, and the gene ontology indicating the biological functions the genes were related to.

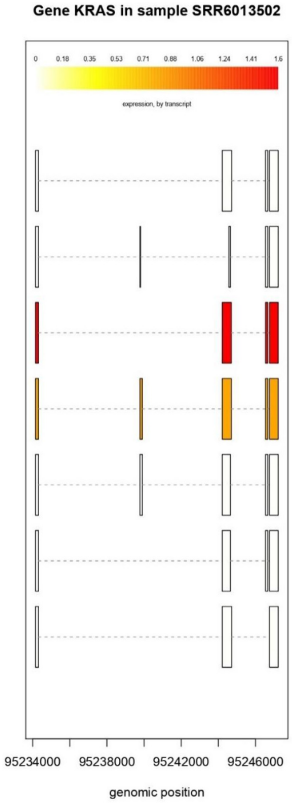

Fig. 3. Structure and expression level of KRAS gene in SRR6013502.

4) Enrichment Analysis

Out of the 5702 hits from the pathway analysis, the genes specifically related to NSCLC were chosen and were again subjected to pathway analysis to see the commonality between breast cancer and small-cell lung cancer. The complete pathway analysis of these genes is provided as a supplementary document. The genes involved in NSCLC and their involvement in breast and small cell lung cancer are shown in Table III.

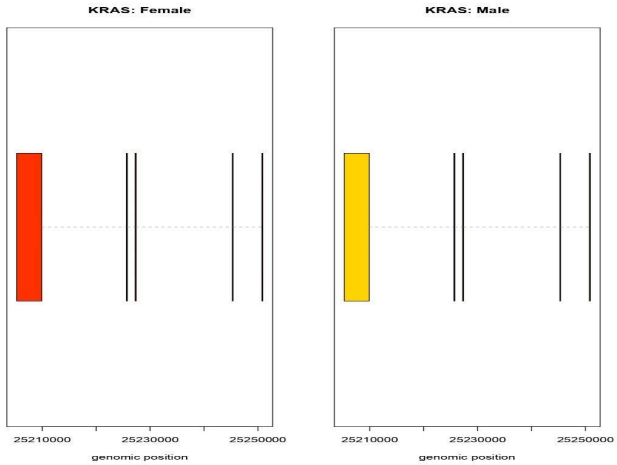

Fig. 4. The structure and expression of KRAS in male and female samples.

The genes responsible for NSCLC and their involvement in breast cancer and SCLC are indicated in Figure 5 and Table III. Figure 5 indicates that 20% of the genes are involved in both NSCLC and breast cancer, 22% are involved in NSCLC and SCLC, and 58% are involved only in NSCLC. The genes given in Table III were subjected to enrichment analysis. The detailed table of the functional enrichment is computed in the functional enrichment of the genes involved in NSCLC.

The enrichment analysis and the corrected p-values are conducted using a hypergeometric distribution which is used for over-representation analysis of genes. The barplot representation of functional enrichment is shown in Figure 6. Each row represents an enriched function, and the length of the bar represents the enrichment ratio, which is calculated as input gene number/background gene number. The indicated color codes are based on the module number which has been assigned based on the enrichment ratio. The functional annotation with the same module number is grouped and indicated by the same color.

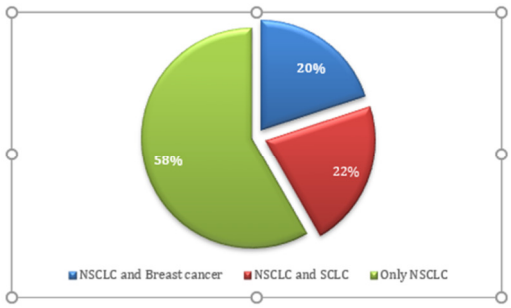

Fig. 5. Pie chart for genes involved in different cancers.

TABLE III. GENES INVOLVED IN NSCLC AND THEIR COMMONALITY WITH BREAST CANCER AND SCLC. THE HIGHLIGHTED GENES ARE THE ONES INVOLVED ONLY IN NSCLC

GENE name	Disease1	Disease2	Disease3
ABCC4	Breast Cancer	NSCLC	
AGER		NSCLC	
AKT1	Breast Cancer	NSCLC	
AKT2		NSCLC	
ARAF	Breast Cancer	NSCLC	
BAD		NSCLC	
BAK1		NSCLC	SCLC
BAX		NSCLC	SCLC
BRAF		NSCLC	SCLC
CASP9		NSCLC	SCLC
CCND1		NSCLC	SCLC
CDK4		NSCLC	SCLC
CDK6		NSCLC	SCLC
CDKN1A		NSCLC	SCLC
DCBLD1		NSCLC	SCLC
DDB2		NSCLC	SCLC
DLST		NSCLC	SCLC
E2F3		NSCLC	SCLC
EGFR	Breast cancer	NSCLC	SCLC
EIF4E2		NSCLC	SCLC
EML4	Breast cancer	NSCLC	SCLC
ERBB2	Breast cancer	NSCLC	SCLC
ETS2	Breast cancer	NSCLC	SCLC
FHIT		NSCLC	SCLC
FOXO3		NSCLC	SCLC
KRAS		NSCLC	SCLC
MAP2K1		NSCLC	SCLC
NRAS		NSCLC	SCLC
GADD45A		NSCLC	SCLC

The circular enrich network of the functional annotation is shown in Figure 7. Each node represents an enriched term, and the edges represent the connections between two enriched terms that have a gene-overlapped ratio more than a specific cut-off (default > 0.5). Different modules in different colors represent nodes belonging to specific topologic communities in a structured network, which were defined using the Infomap algorithm. The network is in a circular layout according to the gravity of the two nodes. The color of the bar in the bar plot is the same as the color in the circular network, which represents different modules and their interactions [19].

5) Obtaining the Candidate Genes

The results obtained from ClinVar database considered only single nucleotide polymorphisms and the corresponding expression analysis from Ballgown. It is seen that the genes BRAF, NRAS, KRAS, EGFR, and MAP2K1, involved in multiple mutations, are computed in the single nucleotide polymorphisms involved in NSCLC, which gives a vast idea about the SNPs. Table III provides detailed information about the nucleotide change and the corresponding amino acid change, chromosome number and the position in the human genome (GRCh38). The expression values of genes involved in mutations are highlighted in Table IV.

TABLE IV. GENES INVOLVED IN MULTIPLE SNPS (AMINO ACID CHANGES)

Gene	Fold change	p-value	q-value
BRAF	0.8108669	0.5386988	0.9979959
MAP2K1	1.004441	0.9728909	0.9993876
KRAS	0.7801372	0.2109766	0.9979959
EGFR	0.4406753	0.1661845	0.9979959
NRAS	1.0061834	0.9705951	0.9993876

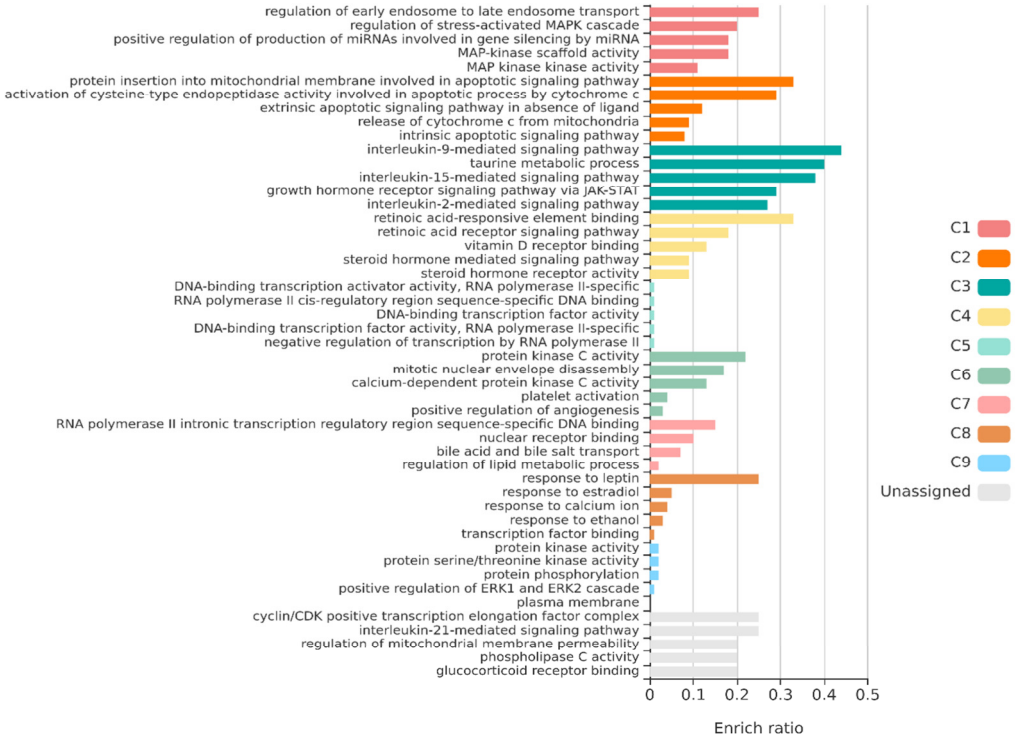

Fig. 6. Barplot representation of functional enrichment.

Fig. 7. Circular enrichment network.

IV. DISCUSSION

The epidermal growth factor receptor was the first oncogenic target uncovered in NSCLC (EGFR). 40% of Asian patients have EGFR mutations, compared to 11-17% of Caucasian patients. Almost all EGFR mutations involve exons 18 to 21. Around 40-50% of EGFR mutations are represented by small in-frame deletions in exon 19 (del 19), whereas p.Leu858Arg amino acid substitutions in exon 21 account for 30-40% [22]. BRAF mutations are found in 2%-8% of people with NSCLC. The BRAF exon 15 p.Val600Glu activating mutation is the cause of 50% of all BRAF mutations. Additional mutations are discovered in exons 11 and 15 and are classified as either triggering (p.Gly469X, p.Leu597Arg, or p.Lys601Glu) or faulty (p.Gly466Val, p.Asp594X, p.Gly596Cys). As predicted by melanoma results [23], single BRAF inhibitors (i.e. vemurafenib or dabrafenib) elicit cell cycle arrest and death in p.Val600Glu-mutated-NSCLC. KRAS-activating mutations are detected in around 30% of the cancer population and are being employed as an exclusion biomarker. In smokers, KRAS-mutated tumours are more prevalent and host other drug-related drivers less frequently. The precise type of KRAS mutation may provide information regarding the aggressiveness of a disease or its sensitivity to certain drugs [24]. For instance, the G12D mutation in NSCLC has been linked to a more favourable prognosis than the G12V or G12R variants [25]. MAP2K1 mutations are infrequent in NSCLC and are assumed to be mutually exclusive with known driver mutations. MEK1 cascade activation is believed to play a major role in the resistance to selective treatment regimens, and the MAP2K1 K57 N mutation has been associated with resistance in preclinical animals [26]. The fold change value for MAP2K1 and NRAS is greater than 1, indicating that these two genes are much more expressed than BRAF, KRAS, and EGFR. BRAF and KRAS are also expressed with values close to 1, although EGFR has a relatively low expression value. PIK3CA, ROS1, and FGFR1 are additional prevalent genes whose mutations are associated with NSCLC, with PIK3CA and FGFR1 being much more expressed than ROS1.

The functional enrichment employed in the research focuses on a novel classification strategy for biomarker genes across 3 major illnesses, such as breast cancer, NSCLC, and SCLC.

V. CONCLUSION

The differential gene expression of the human genome (GRCh38) identifies the most frequently mutated genes as candidates. Mutated genes BRAF, NRAS, KRAS, EGFR, MAP2K1, PIK3CA, MET, ROS1, and FGFR1 were found to be expressed more in the selected data of lung adenoma cancer patients. The knowledge gained from genomic profiling can be utilized to clinically correlate and target medicine for diseases caused by genomic abnormalities. By examining the gene mapping for the specific activity, it is possible to classify these medicines as tumour suppressors or chemotherapy resistant.

To comprehend and categorize the common genes associated with breast cancer and other malignancies, the employed enrichment technique has identified mutant genes as candidates. After a comprehensive validation of mutant genes, potential genes were identified. The discovery of a drug's interactions is prompted by a process of validation that must be carried out to match clinical therapy. The process of chemotherapy and tumor suppression can be revisited with a decreased dose to make the therapies less hazardous.

REFERENCES

[1] D. Wang *et al.*, "The predictive effect of the systemic immune-inflammation index for patients with small-cell lung cancer," *Future Oncology*, vol. 15, no. 29, pp. 3367–3379, Oct. 2019, https://doi.org/10.2217/fon-2019-0288.

[2] N. Behar and M. Shrivastava, "A Novel Model for Breast Cancer Detection and Classification," *Engineering, Technology & Applied Science Research*, vol. 12, no. 6, pp. 9496–9502, Dec. 2022, https://doi.org/10.48084/etasr.5115.

[3] S. Garinet, P. Laurent-Puig, H. Blons, and J.-B. Oudart, "Current and Future Molecular Testing in NSCLC, What Can We Expect from New Sequencing Technologies?," *Journal of Clinical Medicine*, vol. 7, no. 6, Jun. 2018, Art. no. 144, https://doi.org/10.3390/jcm7060144.

[4] A. El-Telbany and P. C. Ma, "Cancer Genes in Lung Cancer: Racial Disparities: Are There Any?," *Genes & Cancer*, vol. 3, no. 7–8, pp. 467–480, Jul. 2012, https://doi.org/10.1177/1947601912465177.

[5] S. A. Kenfield, E. K. Wei, M. J. Stampfer, B. A. Rosner, and G. A. Colditz, "Comparison of aspects of smoking among the four histological types of lung cancer," *Tobacco Control*, vol. 17, no. 3, pp. 198–204, Jun. 2008, https://doi.org/10.1136/tc.2007.022582.

[6] J. R. Molina, P. Yang, S. D. Cassivi, S. E. Schild, and A. A. Adjei, "Non-Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship," *Mayo Clinic Proceedings*, vol. 83, no. 5, pp. 584–594, May 2008, https://doi.org/10.4065/83.5.584.

[7] "Fact sheets," *WHO*. https://www.who.int/news-room/fact-sheets.

[8] D. E. Dupuy and M. Shulman, "Current Status of Thermal Ablation Treatments for Lung Malignancies," *Seminars in Interventional Radiology*, vol. 27, no. 3, pp. 268–275, Sep. 2010, https://doi.org/10.1055/s-0030-1261785.

[9] S.-S. Han *et al.*, "RNA sequencing identifies novel markers of non-small cell lung cancer," *Lung Cancer*, vol. 84, no. 3, pp. 229–235, Jun. 2014, https://doi.org/10.1016/j.lungcan.2014.03.018.

[10] S. Coco *et al.*, "Next generation sequencing in non-small cell lung cancer: new avenues toward the personalized medicine," *Current Drug Targets*, vol. 16, no. 1, pp. 47–59, 2015, https://doi.org/10.2174/1389450116666141210094640.

[11] S. Cheng *et al.*, "Predicting the regrowth of clinically non-functioning pituitary adenoma with a statistical model," *Journal of Translational Medicine*, vol. 17, no. 1, May 2019, https://doi.org/10.1186/s12967-019-1915-2, Art. no. 164.

[12] V. Mero and D. Machuve, "The Usability Testing of SSAAT, a Bioinformatic Web Application for DNA Analysis at a Nucleotide

Level," *Engineering, Technology & Applied Science Research*, vol. 11, no. 3, pp. 7075–7078, Jun. 2021, https://doi.org/10.48084/etasr.4107.

[13] S. Tahzeeb and S. Hasan, "A Neural Network-Based Multi-Label Classifier for Protein Function Prediction," *Engineering, Technology & Applied Science Research*, vol. 12, no. 1, pp. 7974–7981, Feb. 2022, https://doi.org/10.48084/etasr.4597.

[14] S. Bakr *et al.*, "A radiogenomic dataset of non-small cell lung cancer," *Scientific Data*, vol. 5, no. 1, Oct. 2018, Art. no. 180202, https://doi.org/10.1038/sdata.2018.202.

[15] N. B. Hiremath and P. Dayananda, "Identification and Characterization of SNP Mutation in Genes Related to Non-small Cell Lung Cancer," *Current Signal Transduction Therapy*, vol. 16, no. 3, pp. 253–261, http://doi.org/10.2174/1574362415999200819202218.

[16] E. Frenkel, "Gauge theory and Langlands duality," *Astérisque*, vol. 332, no. 1010, pp. 369–403, 2010.

[17] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype," *Nature Biotechnology*, vol. 37, no. 8, pp. 907–915, Aug. 2019, https://doi.org/10.1038/s41587-019-0201-4.

[18] M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, and S. L. Salzberg, "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown," *Nature Protocols*, vol. 11, no. 9, pp. 1650–1667, Sep. 2016, https://doi.org/10.1038/nprot.2016.095.

[19] C. Xie *et al.*, "KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases," *Nucleic Acids Research*, vol. 39, no. suppl_2, pp. W316–W322, Jul. 2011, https://doi.org/10.1093/nar/gkr483.

[20] M. J. Landrum *et al.*, "ClinVar: public archive of relationships among sequence variation and human phenotype," *Nucleic Acids Research*, vol. 42, no. D1, pp. D980–D985, Jan. 2014, https://doi.org/10.1093/nar/gkt1113.

[21] S. Jančík, J. Drábek, D. Radzioch, and M. Hajdúch, "Clinical Relevance of KRAS in Human Cancers," *BioMed Research International*, vol. 2010, Jun. 2010, Art. no. e150960, https://doi.org/10.1155/2010/150960.

[22] F. Barlesi *et al.*, "Routine molecular profiling of patients with advanced non-small-cell lung cancer: results of a 1-year nationwide programme of the French Cooperative Thoracic Intergroup (IFCT)," *The Lancet*, vol. 387, no. 10026, pp. 1415–1426, Apr. 2016, https://doi.org/10.1016/S0140-6736(16)00004-0.

[23] C. S. Baik, N. J. Myall, and H. A. Wakelee, "Targeting BRAF-Mutant Non-Small Cell Lung Cancer: From Molecular Profiling to Rationally Designed Therapy," *The Oncologist*, vol. 22, no. 7, pp. 786–796, Jul. 2017, https://doi.org/10.1634/theoncologist.2016-0458.

[24] P. A. Jänne *et al.*, "Selumetinib Plus Docetaxel Compared With Docetaxel Alone and Progression-Free Survival in Patients With KRAS-Mutant Advanced Non–Small Cell Lung Cancer: The SELECT-1 Randomized Clinical Trial," *JAMA*, vol. 317, no. 18, pp. 1844–1853, May 2017, https://doi.org/10.1001/jama.2017.3438.

[25] M. Román *et al.*, "KRAS oncogene in non-small cell lung cancer: clinical perspectives on the treatment of an old target," *Molecular Cancer*, vol. 17, no. 1, Feb. 2018, https://doi.org/10.1186/s12943-018-0789-x, Art. no. 33.

[26] M. Scheffler *et al.*, "Co-occurrence of targetable mutations in Non-small cell lung cancer (NSCLC) patients harboring MAP2K1 mutations," *Lung Cancer*, vol. 144, pp. 40–48, Jun. 2020, https://doi.org/10.1016/j.lungcan.2020.04.020.