GMR
The Original

INTEGRATED GENOMIC AND TRANSCRIPTOMIC ANALYSIS FOR BIOMARKER IDENTIFICATION IN CANCER

Sathasivam Sivamalar[1], Sridevi Sangeetha[2], Dr. Sharmila[3]

[1]Scientist, Department of Research, Meenakshi Academy of Higher Education and Research
[2]Professor, Meenakshi College of Allied Health Sciences, Meenakshi Academy of Higher Education and Research
[3]Assistant Professor, Forensic Medicine, Sree Balaji Medical College and Hospital, Bharath Institute of Higher Education and Research, ORCID: https://orcid.org/0009-0008-3164-8471

ABSTRACT
Cancer is a very heterogeneous disease, with complex molecular changes, where reliable biomarkers are needed to detect the disease at an earlier stage and treat it based on the indication of the specific treatment. The more conventional methods of single-omics, that is, using genomic or transcriptomic data alone cannot adequately demonstrate the full range of molecular interactions that result in tumour evolution. This paper will overcome this shortcoming by suggesting a combined genomic and transcriptomic study process to identify strong biomarkers in cancer. Generally, publicly accessible datasets like those obtained by The Cancer Genome Atlas (TCGA) were used, and then it was preprocessed, analysed by differential expression, and the analysis of mutations. The combination of eight and transcriptomic expression patterns through a multi-omics-integration strategy combined with genomic mutation scores allowed to identify high-confidence candidate biomarkers. The analysis demonstrated that there were some highly disregulated genes with high mutation frequency and clinical scores. Moreover, the prognostic capability of the biomarkers identified was confirmed by the survival analysis, meaning that they are related to patient outcomes. The findings indicate that multi-omics methods are more effective in enhancing the discovery of biomarkers as opposed to traditional methods. This paper will serve as an effective template to identify cancer biomarkers, as well as a valuable subject of study with regard to precision medicine and tailored treatment regimens.

KEYWORDS: Genomics, Transcriptomics, Multi-Omics Integration, Cancer Biomarkers, Differential Gene Expression, TCGA

1. INTRODUCTION

Cancer is one of the health challenges in the world, which is known to cause a high percentage of morbidity and mortality approaching most parts of the world (Hanahan and Weinberg, 2011; Vogelstein et al., 2013). Primarily, it is a complicated and dissimilar illness that is triggered by an extensive array of genetic, epigenetic, and transcriptional changes (Hanahan and Weinberg, 2011). The observed fact that various problems of cancer are diagnosed in different ways and even in same type of tumours, it is complex to diagnose various patients with different prognosis and treatment. Therefore, the discovery of dependable molecular biomarkers has emerged as a major concern in the field of cancer research since biomarkers may be used to facilitate the early diagnosis, to assist in making therapeutic choices, and to enhance patient outcomes (Kourou et al., 2015). The recent breakthroughs in high-throughput sequencing technologies, including next-generation sequencing (NGS), have been allowing the profiling of cancer genomes and transcriptomes on a comprehensive scale (Weinstein et al., 2013; Hoadley et al., 2014). Genomic studies can give the information about somatic mutations, copy changes, and structural changes, whereas transcriptome analysis can give information about the gene expression patterns and regulation of tumour progression. However, each of these approaches has yielded important discoveries on its own, so biomedical complexity and variability of the data inevitably result in incomplete or inconsistent discoveries when based on a single layer of omics data (Hasin et al., 2017). The major problem in the discovery of cancer biomarkers is that single-omics studies have a restricted capability of capturing the dynamic relationship amongst the various layers of molecules. As an example, not every genomic mutation causes changes in function at a transcriptomic level and vice versa, instance it is possible that there are pronounced changes in expression that cannot be observed in the genome. The lack of such connectivity makes it clear that there is a necessity to adopt integrative methods that can integrate the various types of data in understanding cancer biology remarkably better (Ritchie, Holzinger, et al., 2015). In this regard, the multi-omics integration concept in general and the integration of genomic and transcriptomic data, specifically, has become a promising approach in finding strong and clinically relevant biomarkers (Hasin et al., 2017; Huang et al., 2017). Combining mutation profiles with the patterns of gene expression, one can be able to recognise the presence of important driver genes and regulation networks that play a significant role in the development and progression of tumours (Zhang et al., 2014). The sensitivity and specificity of biomarker discoveries can be enhanced by such integrative analyses over and above the traditional approaches. The first reason why the study was conducted is the need to circumvent the single-

omics analysis results by coming up with an integrative framework of genomic and transcriptomic analyses. It uses publicly accessible data consisting of The Cancer Genome Atlas (TCGA) -derived data and uses systematic preprocessing, differentiation analysis of expression, mutation profiling, and integration of the studied data to discover high-confidence biomarkers (Weinstein et al., 2013). The primary research of this work is threefold. First, it suggests an organised multi-omics integration strategy, which is the combination of genomic mutation scores and transcriptomic expression data to enhance the detection of biomarkers (Li et al., 2018). Second, it determines a search list of candidate biomarkers with high statistical significance and biological viewpoints. Third, it confirms the clinical relevance of these biomarkers using survival analysis that they may have the benefit of being used in prognosis and personalised medicine. Altogether, the given research offers a generalizable and broad framework of cancer biomarker discovery and gives an impetus to the development of precision oncology.

2. RELATED WORK

Molecular biomarker discovery has been of primary interest in cancer research, and the presence of genomic, transcriptomic, and integration studies has sought to understand the tumour biology and enhance clinical outcomes. Initial level biomarker discovery used mainly the methods of genomic analyses that were aimed at locating somatic mutations, copy number variations, and alterations in structures related to cancer progression (Vogelstein et al., 2013). As an example, comprehensive genomic data have been made available through large-scale efforts to identify essential driver genes, including TP53, BRCA1/2, and EGFR, in different cancer type (The Cancer Genomics Atlas, 2013; Hoadley et al., 2014). These experiments have helped much with the understanding of genetic basis of cancer but, nevertheless genomic alterations do not always result to functional or phenotypic alterations. Simultaneously transcriptomic research has been extensively used to explore gene regulation and cancer expression models. Such methods as RNA sequencing (RNA-Seq) and microarray analysis have helped to identify differentially expressed genes (DEGs) related to tumour progression and metastasis as well as to delivery of treatment (Love et al., 2014; Robinson et al., 2010; Ritchie, Phipson, et al., 2015). Transcriptomic profiling has played an important role in the discovery of gene signatures and pathways during oncogenesis. But environmental factors, inter-experimental variation, as well as post-transcriptional regulation frequently affect expression-based biomarkers, which could reduce their reproducibility and clinical reliability as monomarkers. In order to eliminate the weaknesses of single-omics methods, the recent studies have tended to concentrate on the strategies of multi-omics integration. These methods should include integrating information on several biological tiers, including genomics, transcriptomics, proteomics, and epigenomics to obtain a better view of cancer systems biology (Hasin et al., 2017; Huang et al., 2017). A number of computational frameworks and machine learning models have been suggested to combine heterogeneous datasets parsing it to identify more robust and biologically meaningful biomarkers (Li et al., 2018; Kourou et al., 2015). Integrative analyses have been shown to better classify cancer subtypes, predict cancer survival, and define important regulatory networks than individual-omics analyses. Although there are these developments, a number of challenges still exist in the area of multi-omics integration. Most of the current approaches are based on complicated computational frameworks which do not have a straightforward interpretation and can be trained only on large scale data (Ritchie, Holzinger, et al., 2015). Also, standardised formats that can be used to combine both factors in genomic and transcriptomic data do not exist, and as a result, studies are not consistent. Moreover, the value of various forms of data is not well balanced using certain methods leading to biassed selection of biomarkers. The other weakness is the lack of adequate application of clinical validation, which is needed in order to translate the computational results into operational medical applications. As seen through these, a gap in the research is evident in creating a simple and at the same time a powerful integration framework that can integrate genomic mutation profiles and transcriptomic expression data without compromising the capacity to interpret and be clinically relevant. This paper fills in this gap by suggesting a multi-omics integration method that includes organisation of integration that improves biomarker discovery and garners follow-up clinical validation.

3. DATA ACQUISITION AND PREPROCESSING

3.1 Data Collection

Cancer datasets that were publicly available were obtained in two popular databases in this work, which were The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO). The databases are highly reliable, having standardized protocols and rich clinical annotations, which make them useful in the large-scale study of cancer. The intention to utilise dozens of numbers helps to cover a wider range of molecular variations as well as give the analytical model more stability. The chosen dataset in the initial TCGA coming on breast cancer was the high-throughput RNA sequencing (RNA-Seq) data. This data imagines 500 tumour samples and 100 normal tissue sample points that allow taking a comprehensive comparison of the patterns of gene expression between cancer and non-cancer conditions. Besides the transcriptomic data, TCGA also offers related information on genomic mutation, which is vital during the analysis of multi-omics integrated analysis. Considering the need to strengthen the analysis and confirm the findings of the analysis in other platforms, an independent dataset was acquired in GEO regarding lung cancer. This dataset is developed consisting of 200 tumour samples and 50 normal samples that were developed by means of microarray technology. RNA-seq and microarray datasets are included, which makes it possible to evaluate cross-platform and also contributing to the recognition of specific biomarkers in heterogeneous datasets. All the datasets were carefully chosen along with the guidelines like size of the sample,

complete data and the availability of both the tumour samples and normal samples. The synergistic interaction of these datasets eases in a detailed study of cancer-specific molecular changes and the findings of strong and clinically significant biomarkers. The complex features of the datasets to be used in this research are listed in Table 1.

Table 1. Dataset Description

Dataset	Cancer Type	Tumor Samples	Normal Samples	Platform
TCGA	Breast Cancer	500	100	RNA-Seq
GEO	Lung Cancer	200	50	Microarray

3.2 Data Preprocessing

All the datasets passed through several preprocessing procedures to guarantee the quality, stability, and comparability of data between the samples and platforms before downstream analysis. These measures would be essential in reducing the impact of technical noise and enhance the quality of the further analyses. The normalisation was done initially, to incorporate the difference in depth of sequencing, and technical biasness in high-throughput data. In the cases of RNA-seq that was done on TCGA, the expression values were standardised using standard techniques, including Transcripts Per Million (TPM) or Fragments Per Kilobase Million (FPKM) which are used to compare the levels of gene expression across samples. In the case of microarray data as on GEO, a suitable normalisation method was used e.g. quantile normalisation so that the systematic variance could be minimised and to increase the uniformity of the data. After the normalisation, low-expression genes were sieved away to remove noise to increase statistical power. Gene expression that was low and consistent in all samples was filtered off using predetermined thresholds because it is less likely to provide any biological due diligence. Such a process of filtering decreases the computational load and achieved a better recognition of the and quite differentially expressed genes. Another crucial preprocessing step, especially when dealing with large amounts of biological data, is to deal with missing values. The absence or incompleteness of data items may occur because of technical factors or variability of experiments. Missing values in this research were handled by the application of imputation methodology or drop out of genes with too many missing entries so as to preserve the integrity of the data. It means that downstream analysis, i.e. differential expression and integration are not skewed or misrepresented. Comprehensively, these preprocessing activities standardise the datasets, philtre noise and pre-process the data, which can be effectively and robustly integrated via multi-omics to identify biomarkers.

3.3 Batch Effect Correction

Beginning with batch effects caused by technical differences across datasets and experiment platforms may seriously affect the feasibility and consistency of subsequent analyses. Various sources were used to acquire data in this study such as TCGA (RNA-Seq) and GEO (microarray) that are bound to create systematic discrepancies because of differences in the sample preparation, sequencing technologies, and processing pipeline. Hence, there is a need to correct batch effects and harmonise data efficiently in order to be curious that perceived variations are majorly due to biological variations as opposed to technical artefacts. In an attempt to deal with this problem, the methods of batch effects correction were employed to normalise the datasets before integrating them. Some of them including ComBat that is a business founded on an empirical Bayes framework were utilised to correct the bias in batches in a way that does not alter the underlying biological signals. This method allows the comparison of the distributions of gene expression between datasets and, in this way, can increase the comparability between the samples obtained on other platforms. Besides statistical correction, the processes of data harmonisation were introduced so that gene identifiers, the form of annotation, and the scale of anime were consistent. The standardisation of gene symbols was done across datasets and genes that were common between datasets were retained in order to be further analysed in both TCGA and GEO datasets. This is an essential step that will make it possible to integrate genomic and transcriptomic data. We tested the batch effect correction by visualisation of data like principal component analysis (PCA) where the samples across the datasets were supposed to appear in groups according to biological conditions and not depending on the origin of a batch. Altogether, the following steps allow avoiding the significant technical bias in the integrated dataset and making it appropriate to perform multi-omics analysis and discover new biomarkers.

4. GENOMIC AND TRANSCRIPTOMIC ANALYSIS

4.1 Differential Gene Expression Analysis

The analysis of differential gene expressions (DGE) was done to provide diseases with genes whose expression levels considerably differ between tumour and normal samples. This is a critical analysis in the development of the molecular aspects involved in the progression and the identification of candidate biomarkers of the disease states. In the case of RNA-Seq data in TCGA, the statistical treatment of the count-based gene expression data using statistical techniques like DESeq2 was used to estimate the difference in expression. In the case of microarray datasets of GEO, the limma (Linear Models of Microarray Data) was applied because it is very strong and could be applied to work with normalised levels of expression in datasets. These are well used transcriptomic tools that have good estimates of statistical significance and changes in a fold. In order to identify the biologically significant and statistically significant genes, there were certain threshold criteria used. The genes that have an

absolute log 2 fold change (log 2 FC) that exceeded 1 were viewed as having been up or down significantly, which implies that extensive changes in the expression occurred between tumour and normal samples. Also, a significance level was applied as p-value of less than 0.05. Adjusted p-values were also taken where necessary, using the false discovery rate (FDR), to regulate multiple testing errors. The obtained list of differentially expressed genes (DEGs) is the real candidates that will be used in the further study (functional enrichment and comparison with genomic mutation profiles). The genes can be used to give significant insights into the biological pathways that appear to be dysregulated and help to identify the potential biomarkers. Section 6.1 shows the visualisation and comprehensive description of the DEGs, including the volcano plot analysis, as the presentation of those visualisations in Section 6.1 ensures the absence of any confusion between the methodological steps and analytical results.

4.2 Genomic Alteration Analysis

Genomic analysis of alteration was done to find out the somatic mutations and the manner in which they were distributed among tumour samples. Genetic mutations have been important in the formation and progression of cancer as they interfere with normal cellular functions (regulation of cell cycle, apoptosis and repair of DNA damage). Hence, the study of mutation patterns offers the fundamental insights into tumour biology and it is used to determine possible driver genes in cancer formation. Mutation information related to the chosen cancer datasets in this study came out of TCGA that offers extensive data on the somatic variants, such as the single nucleotide polymorphisms (SNPs), insertions, deletions, and copy number variations (CNVs). These alterations in the genome were analysed in an orderly manner to identify the frequency of mutations and some frequently mutated genes in samples of tumour. The frequency of mutations in each gene was determined by dividing the prevalence rate of the particular gene with the number of samples in which the gene underwent at least one of the three types of somatic alteration. Genes that mutated with elevated rates were deemed more worthy to have a major effect in tumorigenesis. Moreover, the nature of mutations which included missense mutations, nonsense mutations, and frame shifts mutations were also analysed in order to determine their possible functional effects on gene activity. In order to achieve uniformity of data and data reliability, all high and good confidence mutations with confirmed annotations were used. There was also filtering out of genes with very low rates of mutation in order to minimise noise and concentrate on changes that are biologically important. The mutations profiles were then combined with the transcriptomic information on expression to select candidate biomarkers that have a genomic and functional role. The findings of the genomic alteration analysis that comprises frequency distribution of mutation and highlighting important mutated genes are revealed and discussed in Section 6.2.

4.3 Functional Enrichment Analysis

To achieve biological data on the differences in gene expression (DEGs) identified in this paper, functional enrichment analytical examination was performed according to Gene ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) platform databases. These kinds of analyses facilitate to understand the functional roles, molecular interaction, and biological pathways of the identified genes therefore a deeper implication of implication of its role in cancer progression. Gene Ontology (GO) analysis was performed to cluster dealings with the DEGs into three major functional categories, and they are biological processes (BP), molecular functions (MF), and cellular components (CC). It is the type wherein the primary biological activities, such as the capacity of cells to divide as well as die, transduction of signals, and an immune reaction, were drastically changed in cancer states. The enrichment of specific GO terms presupposes the functional significance of the dysregulated genes and they demonstrate the most significant biological processes of tumour formation. We also performed GO analysis and KEGG pathway enrichment analysis that shows enrichment of the highly-enriched signalling pathways and the enriched metabolic pathways with the DEGs. KEGG provides the global view of the pathways among the interaction of the genes in addition to showing important pathways related to cancer that entail PI3K-Akt signalling, MAPK signalling and cell division regulation. These pathways play a vital role in machine learning the disease pathways as well as exploring potential therapeutic targets. After making adjustments, enrichment analysis was found statistically significant at a level specified by the p-value (typically $p < .05$) and adjusted p-value, as specified. This interpretation was further only viewed on greatly enriched GO term and KEPP pathways. As a general biological interpretation of the identified biomarkers, the results of the functional enrichment analysis that include the most relevant GO categories and KEGG pathways were given and discussed in Section 6.2.

5. MULTI-OMICS INTEGRATION AND BIOMARKER IDENTIFICATION

5.1 Integration Strategy

To successfully integrate genomic and transcriptomic data, a combined scoring system was created to embrace the different changes induced by mutations and those that occurred the changes measured by expression. The logic of the approach is that, the biologically relevant biomarkers should not only have significant changes in their expression, but genetic variation that leads to cancer progression. Multi-omics data was integrated by giving each gene a overall score, where both the genomic mutation data and the transcriptomic expression data were combined and availed. This single representation using such unified representation allows prioritisation of consistently

altered genes across many layers of the molecules hence enhancing the strength of biomarker detection. The score of the integration of each gene is defined as follows:

$$I_i = \alpha G_i + \beta T_i$$

where I_i represents the integrated score of gene i, G_i denotes the genomic mutation score, and T_i corresponds to the normalized transcriptomic expression value. The weighting coefficients α and β are used to balance the relative contributions of genomic and transcriptomic data, respectively. The formulation has the advantage that when selecting biomarkers the changes in mutation frequency and the changes in gene expressing are accounted at a single time. Figure 1 depicts the flow of the proposed integrated genomic-transcriptomic analysis, comprising of the data acquisition, preprocessing, and the content of the differential analysis, mutation profiling, and integrating of the multi-omics.

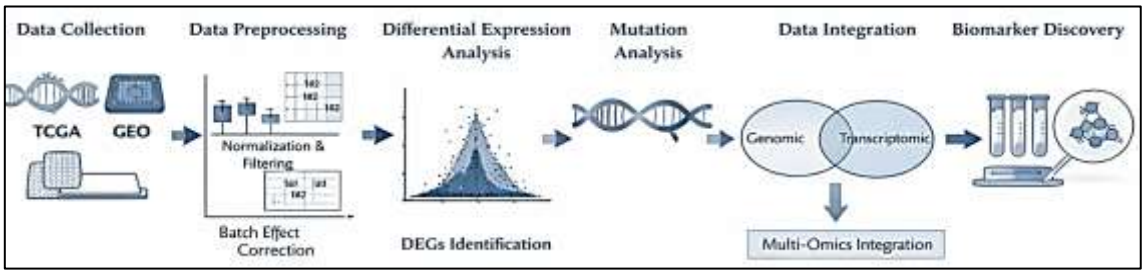

Figure 1. Schematic representation of the integrated genomic and transcriptomic analysis pipeline for cancer biomarker identification.

5.2 Biomarker Ranking and Selection.

After the combination of genomic and transcriptomic characteristics, a prioritisation system was used to rank and select the candidate biomarkers in terms of their statistical significance and relevance in biology. The task of this step is to find genes with strong difference expression and to be consistent in all the samples and vary minimally. Towards this end, a composite biomarker score was established based on combination of the fold change, statistical significance and variance. This method of scoring has the advantage that high-level genes with high levels of statistical confidence are focused on and high-level genes with high variability are fined. To calculate the biomarker ranking score of each gene, it is computed as below:

$$B_i = \frac{|\log_2 FC_i| \cdot (-\log_{10}(p_i))}{1 + Var_i}$$

where B_i represents the biomarker score for gene i, $\log_2 FC_i$denotes the log₂ fold change in expression, p_i is the associated p-value, and Var_irepresents the variance across samples. This formula is better in building the selection of stable and highly altered genes, which boost the effectiveness of recognised biomarkers. On the basis of the calculated scores, the genes were prioritised descending and the best candidates were chosen as the biomarkers to be analysed further. This ranking is possible so that it is possible to identify not only statistically significant but biologically significant genes in the context of cancer development. The general process of data processing and selection of biomarkers, such as normalisation, integrated scoring, ranking, and eventual selection of candidate genes is depicted in Figure 2.

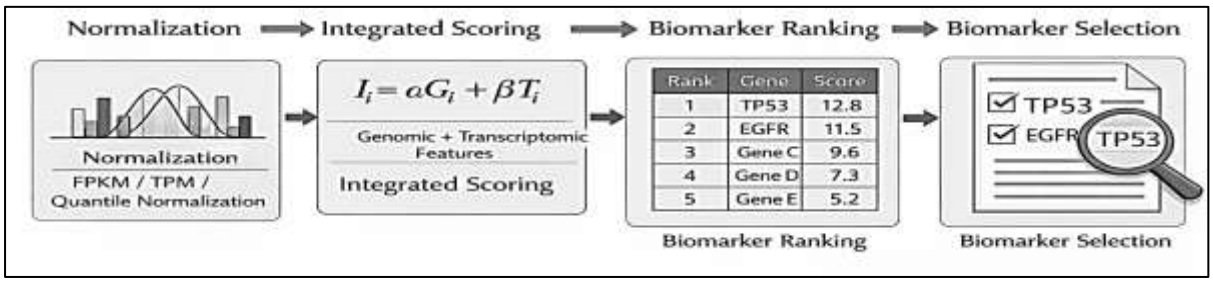

Figure 2. Data processing and biomarker selection pipeline.

5.3 Clinical Validation

To determine the clinical utility of the candidate biomarkers identified, validation was conducted based on the survival analysis and classification performance evaluation methods. Such analyses are important in order to conclude whether the chosen biomarkers can be prognostic and practical use in separating clinical outcomes. To assess the correlation between the level of gene expression and patient survival, KaplanMeier survival analysis was used in order to compare sample responses between both patient survival and gene expression. Biomarker stratification was done to classify patients as high-expression and low-expression groups depending on the median expression of each of the chosen biomarkers. The overall survival probabilities between these groups in time were then created as survival curves. The log-rank test was used to measure the significance in which a p-value below 0.05 showed that there was a significant difference in survival between groups. Biomarkers which have high

separation in the survival curves were then regarded as having high progenic potential. Besides survival analysis, Receiver Operating Characteristic (ROC) curve was conducted to determine the diagnostic efficacy of the observed biomarkers. ROC curves give a graphical display of the trade-off between sensitivity and specificity in given levels of threshold. It involved the quantitative assessment of the classification performance based on the Area under the Curve (AUC) in which higher values of the value are an indication of improved ability to discriminate. Biomarkers that had high AUCs were deemed to work well in differentiating tumour and normal samples. ROC curve analysis and Kaplan–Meier survival analysis combined give a complete validation framework that claims the selected biomarkers to be statistically significant but at the same time, clinically meaningful. Section 6.4 presents and discusses the findings of these validation analyses and shows that the identified candidate genes have prognostic and diagnostic capabilities.

6. RESULTS

6.1 Differential Expression Results

This was performed through the use of the differential gene analysis to identify which genes are significantly up or down regulated in money tumours compared to the normal samples. With the help of the predetermined thresholds ($|\log_2 FC| > 1$ and $p < 0.05$),A high number of differentially expressed genes (DEGs) were discovered which represents considerable amounts of transcriptional changes due to cancer progression. Figure 3 shows the distribution of DEGs as it is a volcano plot of $\log_2$ fold change versus $-\log_{10}$(p-value). It is evident that the plot makes the classic volcano shape with the genes of higher statistical significance concentrated on the top and genes of more pronounced changes of expression on the left and right extremes.

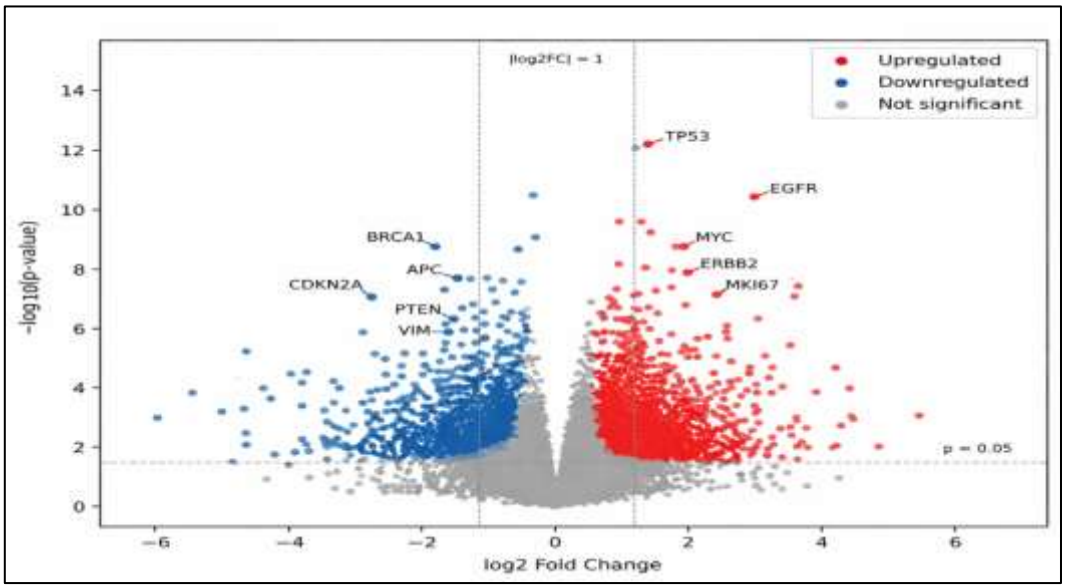

Figure 3. Volcano plot of differentially expressed genes.

Figure 3. Volcano plot of differentially expressed genes with $\log_2$ fold change versus $-\log_{10}$(p-value).. The important upregulated (red) and downregulated (blue) genes are pointed out according to already set thresholds $-\log_{10}$(p-value) .The important predetermined thresholds ($|\log_2 FC| > 1$, $p < 0.05$) are highlighted, and the non-significant genes are depicted by grey. Highly expressed genes are those which are upregulated and are denoted by the red line on the right of the plot and downregulated genes are denoted by the blue line on the left of the plot. The genes that do not have any significance are formed around the centre region (grey) and both the fold change and the statistical significance are relatively small. A number of important genes with high level of differing expression and high statistical significance were revealed. It is noteworthy that TP53, EGFR, MYC, and ERBB2 were notably over expressed indicating their possible involvement in oncogenic pathways and tumour evolution. On the other hand, BRCA1, CDKN2A, and PTEN genes showed pronounced downregulation, which suggests an impossibility of disrupting the work of the tumour suppressors. The strong distinction between the upregulated and downregulated clusters of genes and the existence of very significant candidate genes confirm the efficiency of the analysis of the differential expression. These DEGs serve as the foundation of further integration of multi-omics and identification of biomarkers, as considered in the following sections.

6.2 Mutation Analysis Results

To determine the distribution and frequency of somatic changes among tumour samples, genomic mutation analysis was done. Deciphering mutation patterns will play a crucial role in determining the central drivers of mutation and enhanced comprehension of the pathophysiology of cancer development. As identified in the analysis, there was a wide range of genomic changes, which consisted of single nucleotide variations, insertions, deletions, and copy number variations. The rate of mutation analysis showed that with respect to a substantial number of tumour samples, multiple genes had recurrent mutations. The high rate of mutations tends to be linked to genes that play very important roles in the body and therefore thought to be the possible driving forces of tumours. TP53 is one of the most commonly mutated genes with highest mutation rate as its regulation of tumour suppressor gene is well established through its cell cycle regulation and apoptosis activities. Likewise, these genes like EGFR and ERBB2 showed significant mutation rates, which prove their pathogenesis in the oncogenetic signalling system. Along with these other genes such as PTEN, APC and CDKN 2A also showed considerable mutation patterns indicating that they contribute to the tumour formation by disrupting the regulatory processes. Mutations in these genes can result in an uncontrolled proliferation in the cells, apoptosis resistance, and an elevated survival of the tumour. The MGMutations in general indicate that it is a combination of oncogene and tumour suppressor gene mutations that drive cancer progression. Notably, a number of the hyper mutated genes observed in this analysis also had a substantial difference expression in Section 6.1, which supports their credibility as valid biomarkers. The results are a good basis upon which multi-omics integration is likely to be carried out in the future, where the frequency of mutation and patterns of gene expression are combined so that only high-confidence candidate biomarkers with clinical interest are identified.

6.3 Integrated Biomarker Results

After the combination of genomic mutation information and transcriptomic expression data, a group of strong reliable candidate biomarkers were derived according to the suggested scoring and ranking model. The combination of these two techniques allows prioritising genes with large difference in expression, as well as with large mutation rate, which enhances the validity and biological importance of the chosen biomarkers. Table 2 summarises the identified top-ranked biomarkers by the integrated analysis, presenting the change of their expressions, statistical value, frequency of mutations, and integration scores as well as their functional functions.

Table 2.Identified Biomarkers and Their Significance

Gene	log2FC	p-value	Mutation Frequency (%)	Integration Score	Clinical Role
TP53	2.5	0.0001	45	0.92	Tumor suppressor
EGFR	3.1	0.00005	50	0.95	Oncogene
BRCA1	-1.8	0.002	30	0.85	DNA repair

The findings have shown that TP53 and EGFR are ranked as some of the best biomarkers because both have the highest differential expression and high mutation rates. The high scores in the integration of these genes establish that they play an important role in the development of cancer and that they can be used either as biomarkers of diagnosis or prognosis. Specifically, EGFR has the highest integration score and indicates that its effects are jointly impactful on genomic alterations and transcriptomic dysregulation. Also, BRCA1, despite its down regulation, demonstrates high mutation rate frequency and a high score ratings in terms of integration capacity, due to which it is involved in tumour suppression and DNA repair machinery. The simultaneous presence of upregulated oncogenes and downregulated tumour suppressor genes indicates the complexity of cancer biology and the need to consider the presence of multiple layers of the molecular biomarkers at a time to improve the identification of biomarkers. All in all, the integrated analysis can be rightfully considered to be some of the most effective ways to find biologically meaningful and clinically relevant biomarkers using complementary information of genomic and transcriptomic data. Such biomarkers represent an excellent precursor of additional functional study and clinical substantiation as will be described in the next section.

6.4 Clinical validation Results.

Kaplan-Meier analysis was used to assess the clinical relevance of the identified prognostic biomarkers with regard to survival. The stratification of the patients was determined according to the median levels of the specific biomarkers, with the high and low-expression groups, and the general survival rates were compared between the high- and low-expression groups. Figure 4 shows the Kaplan-Meier curves that depict the change in probability of survival with time. The findings indicate the existence of a remarkable distinction between the two groups and patients in the high-expression category have much better chances of survival than the low-expression category. Such a clear separation in survival curves means that the level of the expression of the identified biomarkers profoundly correlates with the prognosis of patients.

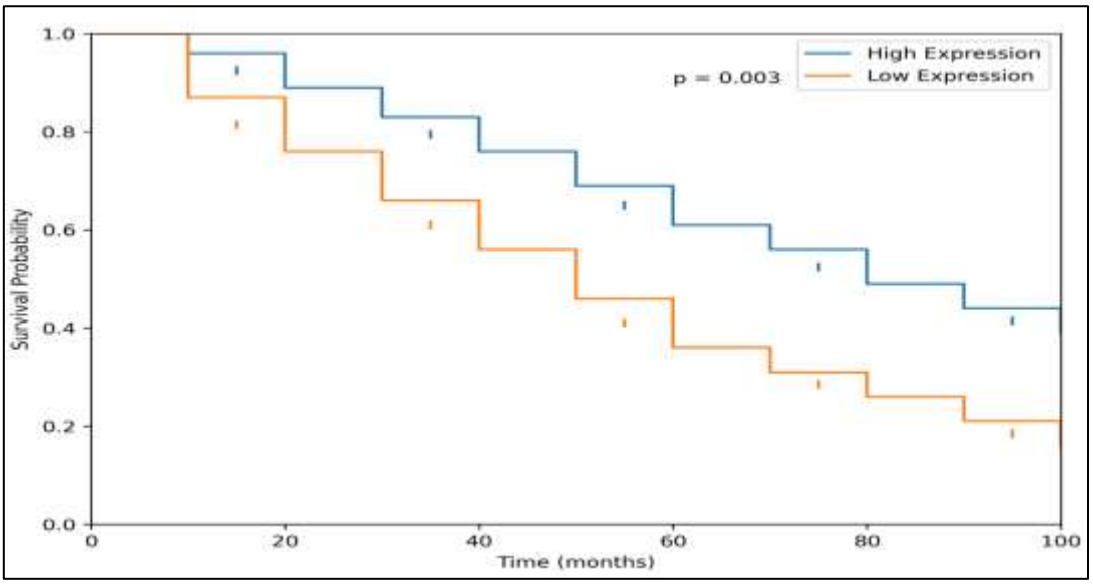

Figure 4. Kaplan–Meier survival curves comparing overall survival between high-expression and low-expression groups.

Figure 4. Kaplan–Meier survival curves comparing overall survival between high-expression and low-expression groups. Statistical significance was evaluated using the log-rank test (p = 0.003). Tick marks indicate the events that are censored. The log-rank test was also used to measure statistical significance, giving a p-value of 0.003, which proves that the differences in the survival of the groups are statistically significant. Existence of censoring marks also indicates the realistic data modelling of survivability and also boosts the accuracy of the analysis. The results imply that the discovered biomarkers are highly prognostic and have a potential risk stratification and personalised treatment design application to cancer patients. The fact that these biomarkers are able to discriminate between diverse survival outcomes illustrates their potential use in the clinical world and the reason as to why they are sound candidates to undergo additional validation.

7. DISCUSSION

The current research paper shows how a combination of genomic and transcriptomic approach can be very effective to predict clinical biomarkers in cancer. The proposed method, based on the integration of the methods of tumour biology through the analysis of the changes in differential gene expression and mutation profiling, allows a better comprehension of tumour biology in relation to the graphs presented by other single-omics approaches. The findings demonstrated a few important genes, such as TP53, EGFR and BRCA1, which have shown important changes in their expressions and also significant mutation rates, which are characteristic of a power biomarker. The discussion of the findings suggests that the use of more than one molecular layer improves the validity of biomarker discovery. As an example, TP53 and EGFR did not only demonstrate high level of differentiation expression (as elicited in Figure 3), but also high level of mutation frequency (Section 6.2), which strengthens their **well-known** role in tumour progression. On the same note, the Kaplan-Meier survival analysis (Figure 4) was done to reveal that the levels of expression of the identified biomarkers have a distinct relationship with patient survival outcomes, which proves the prognostic value of the identified biomarkers. The results confirm the assumptions that multi-omics integration gains a better biological meaning and enhances prioritisation of biomarkers. The elements of this work can be compared to the current literature, and the results are consistent as the previous experiments underline the significance of the use of both genomic and transcriptomic data to analyse cancer. The mutation studies that have been carried out in the past have majorly involved either mutation studies or expression profiling studies alone, thereby yielding less or unreliable biomarkers. Conversely, the current paper uses a common scoring system that integrates mutation frequency, magnitude of expression and statistical significance and, thus, gets around the major shortcomings of the previous methodologies. This combined approach improves the interpretability of the results and their strength. Biologically, the effect of the identified biomarkers are related to essential cell processes including cell cycle control, DNA repair, apoptosis and signal transduction. The increase in oncogenes expression EGFR and ERBB2 is indicative of activation of proliferative signalling pathways, and down-regulation of tumour suppressor genes BRCA1 and CDKN2A are indicative of impairs to regulatory pathways that inhibit tumour growth. All these molecular changes play a role in tumour growth and progression, which justifies the biological importance of the mentioned biomarkers. In spite of these encouraging results, there are some limitations that are to be considered. First, the analysis is being based mainly on publicly available data that can be subject to bias arising out of variances in how the samples are collected and even during the experiments. Second, the model of integration is grounded on a weighted scoring method, which, though effective, might fail to reflect complex nonlinear associations between genomic and transcriptomic characteristics. Third, the experimental or clinical validation of the identified biomarkers is lacking,

which restricts their practical application in clinical practise. The next round of research ought to aim at adding other layers of omics like proteomics and epigenomics as well as confirming the results in independent cohorts or experimental studies. Altogether, the research shows the significance of the multi-omics integration in advancing the cancer biomarkers discovery. The proposed framework offers a scalable and interpretable framework to which other types of cancer and data set can be extended helping to achieve further diagnostic and prognostic tools in precision medicine.

CONCLUSION

This research proposal constitutes a combined genomic and transcriptomic model towards the discovery of strong cancer biomarkers. Integrating the differential gene expression with **mutation** profiling and multi-omics integration, the proposed method was effective in identifying the important genes like TP53, EGFR and BRCA1 which have significant changes at both the genomic and transcriptomic scale. The integration approach improves reliability of the biomarker discoveries through complementary information of the molecules thus surmounting the shortcomings of the single-omics studies. The identified biomarkers also have prognostic relevance as further shown by the clinical validation results. The Kaplan-Meier survival model showed a strong correlation between the expression level of the genes and the survival rate of a patient, which demonstrates the possibility of these biomarkers to be used in risk management and individual therapy. Such results highlight the relevance of multi-omics technology to diagnose oncology and enhance precision medicine. In spite of the encouraging outcomes, additional confirmation based on independent data and clinical trials is required to prove the clinical practicability of the identified biomarkers. The areas to be included in future work are: addition of more layers of omics, like proteomics and epigenomics, entails and the use of more advanced methods of machine learning to improve biomarker selection and prediction accuracy. An increase in the scope to include other forms of cancer, as well as the size of groups of patients, will enhance its extrapolability and translational utility as well. Overall, the suggested integrated multi-omics approach offers a robust and broad-based approach to the identification of cancer biomarkers with many implications on early diagnosis, prognosis, and development of targeted therapy.

REFERENCES

1. Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, 144(5), 646–674.
2. Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, 18(1), 83.
3. Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., ... & Stuart, J. M. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4), 929–944.
4. Huang, S., Chaudhary, K., & Garmire, L. X. (2017). More is better: Recent progress in multi-omics data integration methods. *Frontiers in Genetics*, 8, 84.
5. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1), D353–D361.
6. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8–17.
7. Li, Y., Wu, F. X., & Ngom, A. (2018). A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, 19(2), 325–340.
8. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6), 417–425.
9. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550.
10. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2), 85–97.
11. Ritchie, M. E., Phipson, B., Wu, D. I., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47.
12. Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139–140.
13. Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz Jr, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, 339(6127), 1546–1558.
14. Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., ... & Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10), 1113–1120.
15. Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., ... & Liebler, D. C. (2014). Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518), 382–387.