Available online at www.IJournalSE.org

Emerging Science Journal

Vol. 3, No. 3, June, 2019

Developing an Integrated Genomic Profile for Cancer Patients with the Use of NGS Data

A. Kosvyra [a*], C. Maramis [a], I. Chouvarda [a]

[a] *Laboratory of Computing, Medical Informatics and Biomedical Imaging Technologies, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece*

Abstract

Next Generation Sequencing (NGS) technologies has revolutionized genomics data research by facilitating high-throughput sequencing of genetic material that comes from different sources, such as Whole Exome Sequencing (WES) and RNA Sequencing (RNAseq). The exploitation and integration of this wealth of heterogeneous sequencing data remains a major challenge. There is a clear need for approaches that attempt to process and combine the aforementioned sources in order to create an *integrated profile* of a patient that will allow us to build the complete picture of a disease. This work introduces such an integrated profile using Chronic Lymphocytic Leukemia (CLL) as the exemplary cancer type. The approach described in this paper links the various NGS sources with the patients' clinical data. The resulting profile efficiently summarizes the large-scale datasets, links the results with the clinical profile of the patient and correlates indicators arising from different data types. With the use of state-of-the-art machine learning techniques and the association of the clinical information with these indicators, which served as the feature pool for the classification, it has been possible to build efficient predictive models. To ensure reproducibility of the results, open data were exclusively used in the classification assessment. The final goal is to design a complete genomic profile of a cancer patient. The profile includes summarization and visualization of the results of WES and RNAseq analysis (specific variants and significantly expressed genes, respectively) and the clinical profile, integration/comparison of these results and a prediction regarding the disease trajectory. Concluding, this work has managed to produce a comprehensive clinico-genetic profile of a patient by successfully integrating heterogeneous data sources. The proposed profile can contribute to the medical research providing new possibilities in personalized medicine and prognostic views.

Keywords:

Bioinformatics;
Sequencing Analysis;
High-Throughput Sequencing;
Data Mining.

Article History:

Received:	08	March	2019
Accepted:	26	May	2019

1- Introduction

With the completion of the 1000 Genome project [1] and the rapid evolution of sequencing techniques, genomic information is now massively produced. The collection and process of the genome of living things is called genomic data. The availability of multiple types of genomics data has considerably affected the medical sciences [2]. It has now become possible to extract information from a great variety of genomic data. A result of this evolution is the emerging need for developing robust biotechnologies for storage, management, analysis and interpretation of this information. Taking into consideration additional types of health related data accompanying this genomic information for a person, such as the clinical information, other laboratory exams or behavioral data collected by smart devices, the data collected are of such volume that are difficult to manage and process. For this reason, biomedical informatics has become an essential component in medical research.

One of the most popular sequencing techniques that has emerged during the last decade is the Next Generation Sequencing (NGS) [3]. NGS is a technique that is based on massive, parallel, small reads of the genome and is a radical approach that changed the potential of sequencing. NGS opened new possibilities in exploration of diseases in a genomic

* **CONTACT**: Aekosvyra@auth.gr

DOI: http://dx.doi.org/10.28991/esj-2019-01178

level by making easy to detect mutations or gene behavior [4]. Some examples of NGS data is Whole Exome (WES) [5] and RNA (RNAseq) [6] Sequencing. WES is a sequencing technique focusing on the coding protein genes' regions. These regions cover about 1% of the whole genome and provide information about the mutations in the lower possible cost. RNAseq is a transcriptomic sequencing technique and provides information about the count level of transcripts and their isoforms and the gene structure along with their expression level.

In this paper we propose an end-to-end methodology that extends the state-of-the-art in the analysis and integration of two different sources of genomic data, WES and RNAseq, in descriptive and predictive level (see Section 1-1 below). The methodology starts with the raw data analysis, continues with further analysis of the results in order to summarize the large-scale datasets, visualize them and detect biomarkers that arise from this analysis to be used for the deployment of the predictive models. This further analysis is conducted not only individually for each sample but also for groups of patients so the results can be easily compared. Finally the integrated profile is created with the use of these methods and by summarizing the variety of information.

1-1- Literature Review

The raw data resulting from these sequencing techniques are bulky and complicated, thus difficult to process and interpret. To that end, a variety of tools have been developed for analysis of the sequences including different methods, and functionalities that vary from general purpose to more targeted analysis. Some of the most popular tools for WES analysis are: *SeqMule* [7], which is a tool that performs alignment and variant calling and is used in this study, *Impact* [8], which performs computation of variants and Copy Number Variations in order to perform drug response predictions, and, *WEP* [9] which performs a complete analysis for data cleaning and alignment and variant detection. For further exploration of these pipelines, variant annotation tools are available with the most popular being ANNOVAR [10] which offers a great variety of annotations. Respectively, many tools for RNAseq raw data analysis have been developed with the most popular being: *Tuxedo Protocol* [11] which performs alignment, gene and transcript count and expression level computation and differential analysis between patients or groups of patients and is used in this study, and *Tuxedo 2* [12], a more recent version of the protocol with the same outcome, *Viper* [13] which performs RNASeq data analysis and provides visualizations of the results, and, *IRAP* [14] which is also a workflow for alignment, gene expression computation and differential expression analysis.

Although these methods for raw data analysis are very useful, they do not provide a *comprehensive view* of the datasets so that the information can easily be used by the experts. So, there is a need for developing *integrated tools for analysis, presentation and summarization of the results*. There are some available tools for simultaneous presentation of the results of different analysis, such as the *Integrative genomics viewer* [15], which is a visualization tool for interactive exploration of large, integrated genomic data and supports a great variety of NGS data types and formats, *Epiviz* [16] which provides a variety of visualization methods based on the region of interest or characteristics as the expression, and, *Timiner* [17] which integrates WES and RNAseq analysis in order to detect neoantigens in a sample.

Moreover, there is an interest in research towards the integration of data in ways that their combination increases the individual data's value, such as the Codina-Solà *et al.* study, in which WES and RNAseq analysis' results are used to detect genes that cause the disease (Autism Spectrum Disorder), transcript mutations and mutations sensitivity [18]. Another interesting approach is the one described from Wilkerson *et al.* that use an integrated and novel approach to detect mutations, based on the assumption that WES analysis has minor sensitivity in low purity tumors [19]. In addition to that, Landesfeind *et al.* propose a method for mutation detection in both WES and RNAseq data and comparison of these results for a more punctual and comprehensive molecular characterization of the samples [20]. More recent studies focus on integrating gene expression with DNA methylation data, as of Cappelli *et al.* and Li *et al.*, for knowledge extraction beneficial to prognosis [21, 22]. Moving towards prognosis approaches, Fleck *et al.* propose a method based on the integration of mutations and gene expression to detect how mutations can lead to changes in gene expression, and, consequently, cancer progression [23]. In the same direction, Yu *et al.* and Zafeiris *et al.* propose the use of artificial neural networks for disease classification and biomarker discovery respectively [24, 25].

2- Methods

For the purposes of this study, the disease chosen is Chronic Lymphocytic Leukemia (CLL) which is a neoplastic blood disease with a strong genetic influence. This disease was selected as an example for the implementation of the methodology. CLL is a chronic disease, currently untreated and is the most frequent type of adult leukemia. Moreover, the knowledge of the nature, progression and treatment of the disease is in an early stage. The complete procedure is described in the block diagram of Figure 1.

2-1- Data

All data used for this analysis are open data acquired from the National Center for Biotechnology Information (NCBI). Two genomic data sources were selected for this study, RNASeq and WES data. Both cases consist of raw data

of early stage CLL patients in fastq format. RNAseq raw data belong to IGR_U985_RNASeq* study and consist of 12 cases, each one corresponding to a single subject, 4 cases with mutated and 8 cases with unmutated EGR2 gene. WES raw data belong to IGR_U985_CLL_Exome† study and consist of 24 cases, 17 with mutated and 7 with unmutated IGVH gene. Each WES case corresponds to cancer (B lymphocytes) and healthy (T lymphocytes) tissue sample of the patients. Healthy tissue is used as control sample for this analysis. A summary of the datasets is provided in Table 1.

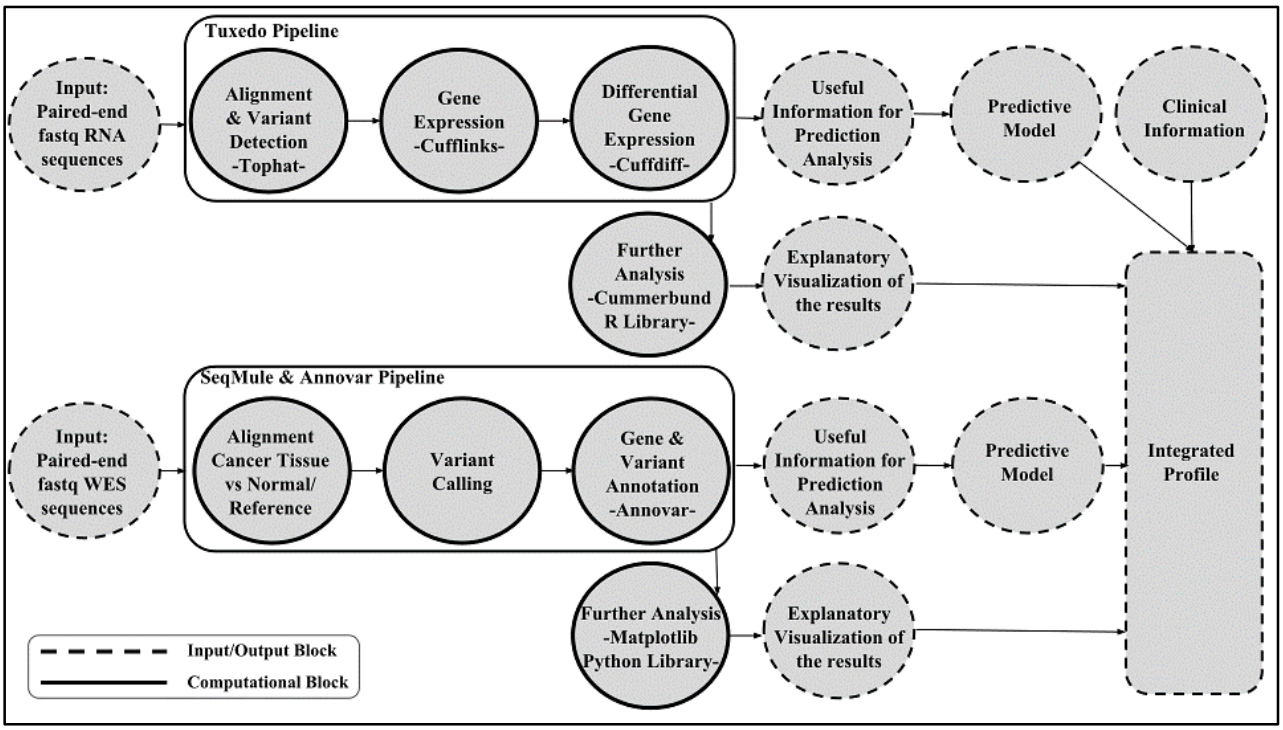

Figure 1. Integrated profile creation process.

This division between groups was used to perform the inter-group comparisons and the classification for the predictive analysis. The separation for the predictive analysis was made based on the assumption that, for the RNAseq analysis, patients with mutated EGR2 gene have a good prognosis [26] and, for the WES analysis, the patients with unmutated the IGVH gene are characterized as stable [27].

Table 1. Data Summary.

Case group description	Nr. of cases	Size in Gb	Disease Outcome Ground Truth
WES			
Group A - mutated IGVH	17	~340	Aggressive
Group B - unmutated IGVH	7	~140	Stable
RNAseq			
Group C - mutated EGR2	4	~40	Stable
Group D -un mutated EGR2	8	~80	Aggressive

2-2- Descriptive Analysis

The descriptive analysis includes as a first step the raw sequencing data analysis and proceeds with further exploration of the results with the methods described in detail below. All kinds of analysis were performed on one of the two high-performance computational clusters of AUTH, Afroditi‡, part of the National Grid Infrastructure, Hellasgrid. This infrastructure was selected because these tools demand high computational resources and time to be executed. After constant communication with the experts of this infrastructure maintenance and support we concluded in the optimal amount of resources needed for the fastest and most efficient execution.

* https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?study=ERP003791
† https://www.ncbi.nlm.nih.gov/bioproject/272497
‡ https://it.auth.gr/en/hpc

2-2-1- WES Analysis

Whole Exome Sequencing analysis was conducted using the Seqmule pipeline which performs alignment and variant calling, providing the opportunity of using a desirable combination of different aligners and variant callers. Two distinct configurations of this pipeline were used for this analysis, namely: (a) the normal vs. cancer tissue analysis for the descriptive analysis - this configuration uses BWA and Gatk aligners and Samtools and Bayes variant callers; (b) cancer tissue vs reference genome for the predictive analysis - the default configuration. The results of this analysis consist of information regarding the mutations detected in each sample and they need further processing.

Regarding variant annotation, different gene- and filter-based annotations were used in order to filter the detected Single Nucleotide Polymorphisms, SNPs. First, we perform a gene based annotation from which the type of each detected variant is derived. In this level, the variants are categorized as {synonymous, non-synonymous, other}. Secondly, in a filter-based approach, the variants are annotated based on the 1000 Genome Project findings and each one is given a MAF (Minor Allele Frequency [28]) value and then is filtered based on this value in order to distinguish frequent from infrequent ones. Finally, another filter-based approach is used, using the scores of dbNSFP [29] in order to categorize the variants as {tolerated, deleterious}.

These tools produce a great amount of information regarding the mutations detected in the samples that needs to be further analyzed and explored in order to be used for the deployment of predictive models. Furthermore, powerful visualizations can highlight the importance of these results. This visual exploration analysis was conducted using well known and easy to use Python Libraries, as Pandas and Matplotlib.

2-2-2- RNAseq Analysis

RNA sequencing raw data analysis was conducted using the Tuxedo pipeline. This pipeline performs (a) sequence alignment with the tool Tophat [30], which provides a list of successful alignments as well as information for mutations and junctions, (b) gene & transcript expression computation with the Cufflinks [31], and (c) differential gene expression analysis between two groups of patients with Cuffdiff [32]. For the purposes of this research, we focused on gene expression and differential gene expression data. In this study, we used this pipeline to calculate gene and transcript expression for all the patients and differential expression analysis between two groups.

Further analysis of these results was conducted using the Cummerbund R library [33], which provides a whole set of tools for exploring RNAseq analysis results, from statistics to dimensionality reduction tools. In addition, algorithms built from scratch in python programming language were employed to feature selection and model training process. Powerful visualizations indicate the importance of these results as described in Section 3-2. We further explored the outcomes of the descriptive analysis towards the design and deployment of predictive models.

2-3- Predictive Analysis

In any disease, there are a lot of questions that need to be answered so a clinician can proceed directly to a treatment or a personalized choice for the patients. Such questions are:

- Response to treatment, e.g. is the patient going to respond to a specific medication?
- Relapse, e.g. is the patient going to relapse after the treatment?
- Disease Outcome, e.g. is the patient going to be in a stable condition or the disease is aggressive and will lead to lower expected survival?

In this study we chose one of these questions as the target of the predictive analysis. The aim of this analysis is to predict the Disease Outcome, meaning the prediction if the disease is going to be aggressive, with devastating consequences for the patient, or stable. For this purpose each group of patients was divided in two classes, stable and aggressive. Two models were developed, one based on RNAseq analysis results and one based on WES analysis results. For every case, a dimensionality reduction algorithm was used to primarily explore the potential of these datasets. Particularly, Multidimensional Scaling (MDS) analysis was used and the results depicted in Figure 2. The figure shows that in the RNASeq (a) case the samples of different classes were grouped but in WES (b) case they weren't.

For this analysis, the scikit-learn [34] Python Library was used. This tool provides implementations of all the well-known machine learning algorithms in an easy-to-use environment. The specific classifiers tested is a simple 1 hidden layer neural network, decision tree and random forest, as well as a linear regression and a Bayesian one. The first step of the deployment of the predictive models is the feature selection and it is described for both cases below.

2-3-1- RNAseq-based Feature Selection

In this case, gene expression for all patients was calculated. Then, the results were merged on each gene, the mean expression of each gene was calculated for every group and, finally, the difference of mean expression between groups. The 20 most differentially expressed genes was selected as the features defining the model.

2-3-2- WES-based Feature Selection

In this case, the SNPs detected in every sample were merged for all patients in every group and after the aforementioned annotations they were filtered and only the SNPs that were non-synonymous, heterozygous, deleterious and with MAF>0.5 were finally selected. The number of SNPs per gene was calculated for each patient and these genes were selected as the features defining the model.

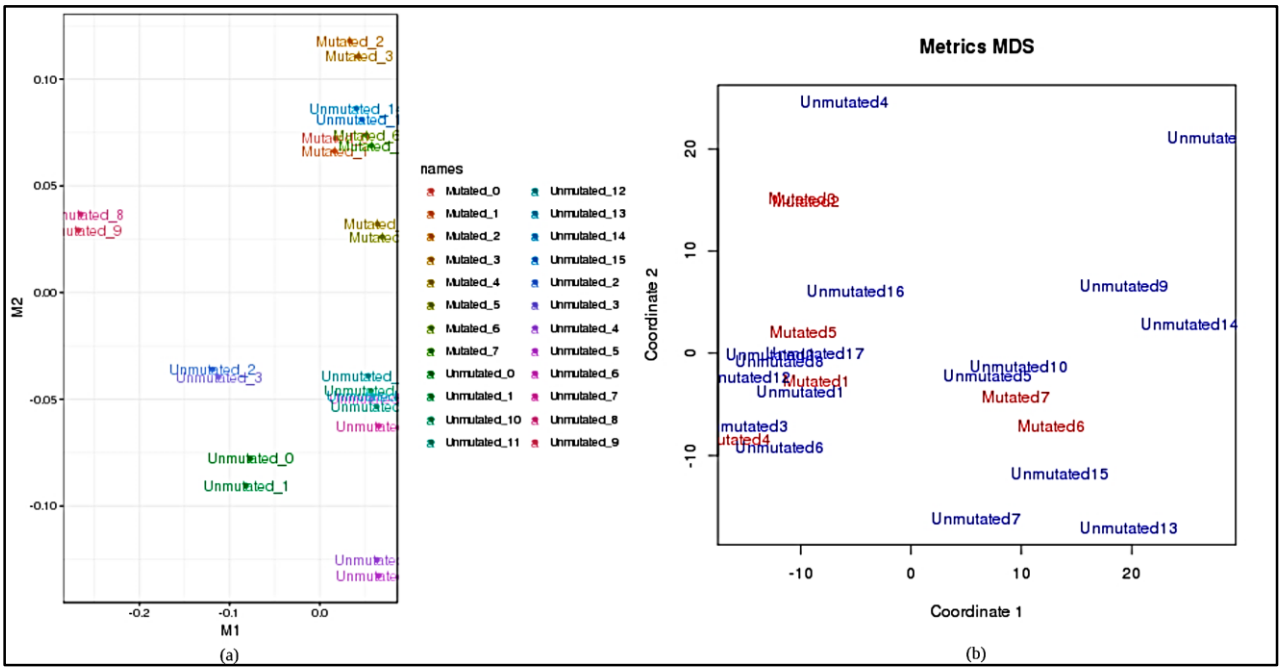

Figure 2. MDS Analysis Plots: (a) RNAseq; (b) WES.

2-3-3- Training and Evaluation

Both models ware trained with a number of well-known classifiers in order to select the one that performs the better in each case. For the validation of the models, the Leave-one-subject-out cross validation method was used. In cases were a classifier imports randomness the reported results are the mean value of 100 runs. The evaluation metrics used are accuracy, sensitivity and specificity.

2-4- Integrated Profile

The construction of the cancer patient' integrated profile is a complete methodology which starts with the raw data analysis, continues with the descriptive and predictive analyses and concludes with the summarization, visualization and integration of this information.

2-4-1- RNAseq-WES Data Correlation

The first approach, the first image of the *Combined Visualizations* section in Figure 7, attempts to explore correlation between gene expression and mutation frequency and depicts a combination of the results in gene level. The information used is the gene significance and expression, computed via RNASeq Analysis, and variant frequency per gene, computed via WES analysis. Each bubble in the scatter plot represents a gene and the size of the bubble represents the expression level of this gene. The position on the y-axis represents the number of non-synonymous/heterogeneous SNPs detected in the specific gene and the position on the x-axis the p-value of this gene.

The second approach, the second image of the *Combined Visualizations* section in Figure 7, explores the difference in the detection of SNPs from the different analyses and depicts a combination of the results in SNP level. The information used in this case is the number of SNPs detected in every chromosome via RNAseq vs WES analysis. From literature, it is known that RNAseq analysis is used in many cases to detect SNPs as an aid to the WES analysis as in some cancers is more accurate and detailed.

In addition to that, the results of the predictive analysis need to be combined for achieving higher accuracy. Although for this study it was not possible to achieve this integration due to lack of multi-omics data for the same patients.

2-4-2- Integrated Profile

After discussion with experts of the biology field, specializing in immunogenetic and blood cancer research, and after reviewing relevant literature on the important outcomes and integration of each analysis, this attempt tries to present the results of individual analyses and combine them in an easy readable way in order to facilitate a quick view of the condition of the patient. The information of the two previous analyses need to be combined in order to have a more comprehensive view of the disease, in genomic and transcriptomic level. In more detail, from the available clinical information, only a small and representative sample was chosen. The first panel of the profile provides personal information of the patient, containing demographics (e.g. gender) and clinical information (e.g. treatment schema). The second panel provides a summary of RNAseq analysis. The information selected for presentation, was considered to be the most representative of thin analysis and depicts the 20 most expressed genes of the patient. Regarding WES analysis, as important information for summarizing the results in the third panel, we chose the distribution of all variants detected in the patient categorized in SNPs/Indels. Furthermore, in the fourth panel, it provides a prediction for the aggressiveness or not of the disease for this patient presenting the results of both predictive models deployed in this study, and, finally, in the last panel, a presentation of the integrated RNAseq and WES analysis in gene and chromosome level as described in detail in Section 2-4-1.

3- Results

3-1- Execution Performance

As already discussed there has been a study on the optimal resources needed for every tool in order to achieve the desired execution time. Table 2 depicts the performance time for each of the external tools and the resources' demands. It was observed that, in some cases, increasing the resources after a certain amount didn't cause a reduction in execution time and, in some cases, it caused an increment. The execution time presented is the average time of all runs of the tools, #24 for seqmule pipelines, #12 for tophat, cufflinks and #1 for cuffmerge, cuffdiff.

Table 2. Performance and resources for the external tools

Case	Tool	Input/Output Size	Resources' constraints	Execution Time (hours)
WES analysis	Seqmule Pipeline (Default Normal configuration)	10/25 Gb	Max RAM: 40G, Max #CPU: 12	~8
	Seqmule Pipeline (Somatic configuration)	20/25 Gb	Max RAM: 40G, Max #CPU: 12	~7
RNAseq analysis	Tophat	10/5 Gb	Max RAM: 40G, Max #CPU: 12	~10
	Cufflinks	5/0,5 Gb	Max RAM: 40G, Max #CPU: 12	~0,5
	Cuffmerge	-/0,5 Gb	Max RAM: 40G, Max #CPU: 8	~0,5
	Cuffdiff	-/5 Gb	Max RAM: 80G, Max #CPU: 32	~36

3-2- Descriptive Analysis Results

The results of the descriptive analysis are a summarization-visualization of the pipelines' results. In both cases, WES and RNASeq analysis, the presentation takes place in two levels. The first level concerns the intra-person comparison. It contains visualizations from the two different analyses outcomes for one person in order to extract the combined information needed for integration in the last step. The second level concerns the inter-group comparison and contains visualizations that compare the results for two groups or two individual members of two groups. In any case, there are several graphs to be used to depict all the available information. Some examples are chosen to be displayed for explanatory purposes.

Examples of the intra-person and inter-group comparisons for WES analysis are depicted in Figures 3 and 4 respectively. In this case, the distribution of variants in the sample is depicted. Examples of the intra-person and inter-group comparison for RNAseq analysis are depicted in Figures 5 and 6 respectively. In this case, the differential gene expression between the two groups is depicted.

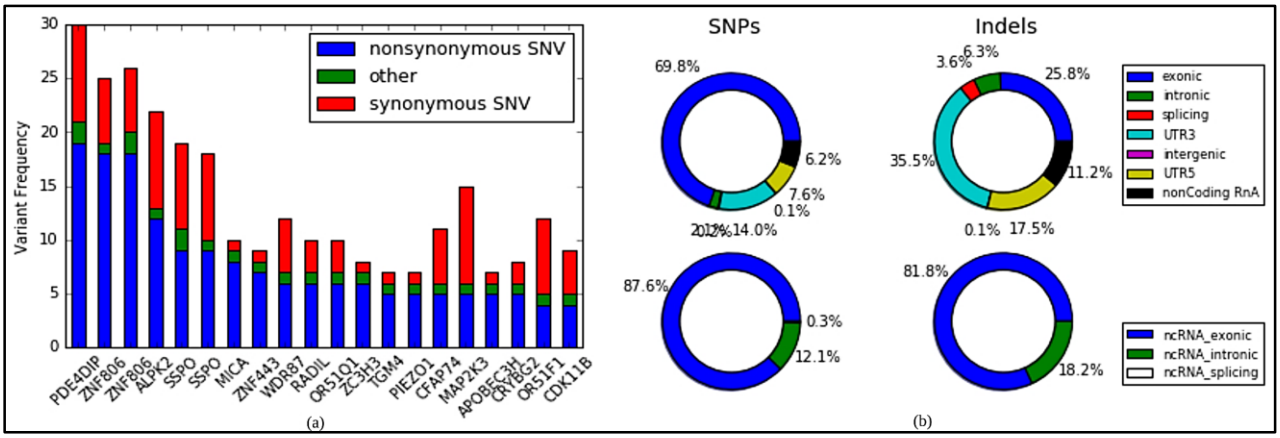

Figure 3. **WES analysis result intra-person visualization: (a) SNP frequency types (synonymous/non-synonymous/other) per gene for the 20 genes with the most variants and the pie charts (b) frequency of detected variants in the same sample, divided in SNPs and indels and per variant type.**

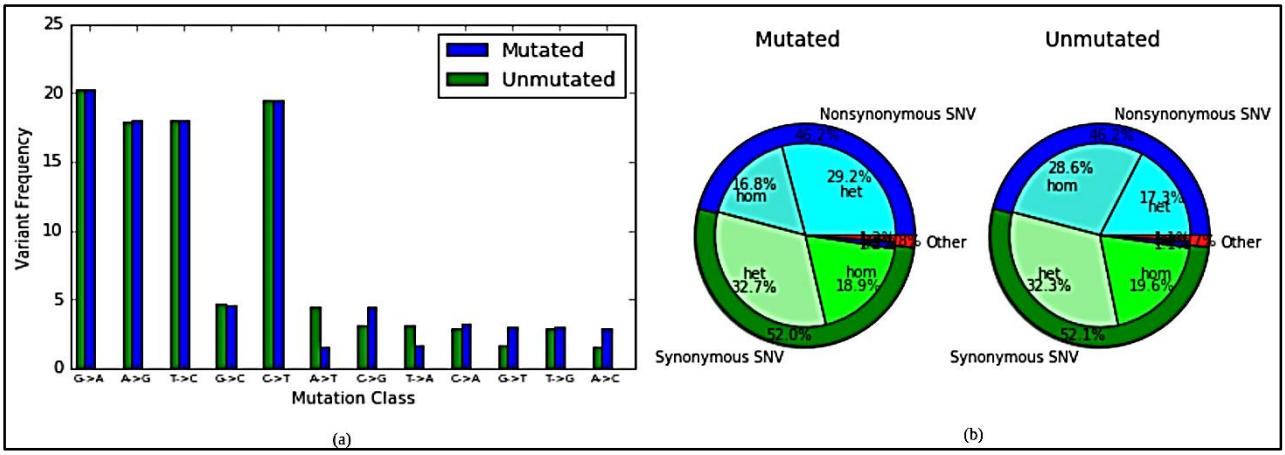

Figure 4. **WES analysis result inter-person visualization: (a) frequency of transitions and transversions in the two group (b) the distribution of synonymous/non-synonymous, homozygous/heterozygous SNPs for each group.**

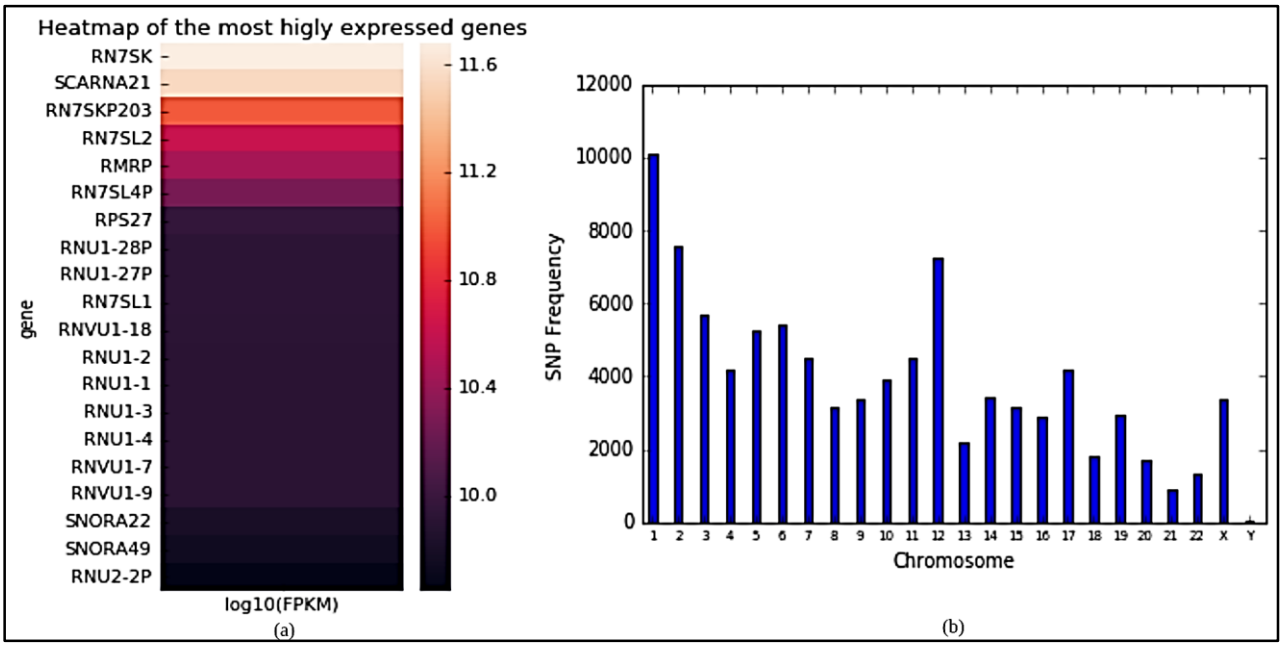

Figure 5. **RNAseq analysis results intra-person visualization: (a) the 20 most highly expressed genes in a patient (b) the distribution of SNPs per chromosome as they were computed via this analysis.**

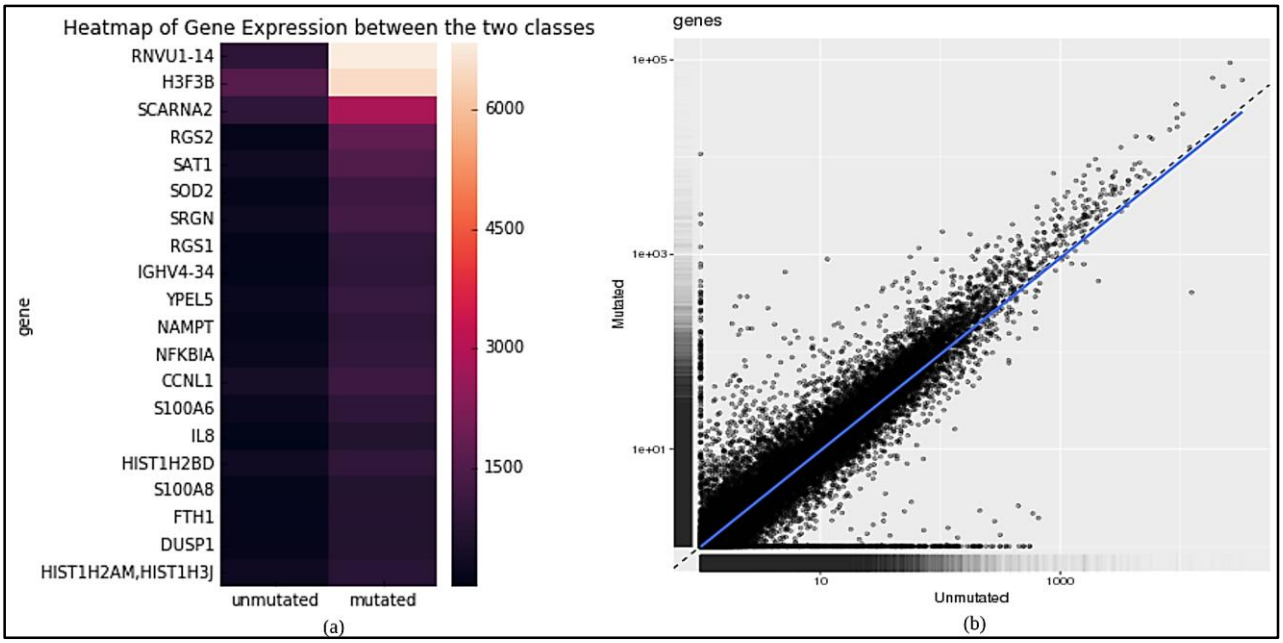

Figure 6. RNAseq analysis results intra-person visualization: (a) 20 most differentially expressed genes between these groups (b) distribution of the expression of all genes between the two groups, as they were divided based on the EGR2 gene mutation.

3-3- Predictive Analysis Results

The features selected for the RNA-seq based model were the 20 genes: {RNVU1-14, H3F3B, SCARNA2, RGS2, SAT1, SOD2, SRGN, RGS1, IGHV4-34, YPEL5, NAMPT, NFKBIA, CCNL1, S100A6, IL8, HIST1H2BD, S100A8, FTH1, DUSP1, HIST1H2AM}. The features selected for the WES based model were the 10 genes for the one class: {EXO1, GPRIN2, BCLAF1, OR13F1, OR1B1, OR6C68, ELOA2, KNG1, CYP4V2, SLC17A1} and 26 for the second: {CLCNKB, TRAF3IP3, FAM177B, LEXM, OR51S1, TMEM132C, CCDC175, CACTIN, TPO, RIF1, GRB14, HJURP, COL6A3, PLB1, DTX3L, LARP1B}.

As depicted in Table 3, for both cases the classifier with the better performance is Random Forest. As clearly Table 3 shows, and in accordance to what it was expected from the MDS analysis, the performance of the classifiers in the RNAseq-based model is better.

Table 3. Validation Results.

Case	Classifier	Accuracy	Sensitivity	Specificity
RNAseq -based model	Neural Network	0,43	0,37	0,50
	Logistic Regression	0,62	0,50	0,75
	Naive Bayes	0,81	0,75	0,87
	Decision Tree	0,68	0,75	0,62
	Random Forest	**0,87**	**0,74**	**0,87**
WES -based model	Neural Network	0,44	0,02	0,60
	Logistic Regression	0,45	0,00	0,61
	Naive Bayes	0,54	0,00	0,65
	Decision Tree	0,55	0,24	0,67
	Random Forest	**0,75**	**0,42**	**0,88**

3-3- Integrated Profile

Finally, with the combination of the results of the two analyses, Descriptive and Predictive, it has been able to create an integrated profile of a CLL patient that summarizes the clinical information, WES & RNAseq data and predictions. An example of the view of this profile is depicted in Figure 7.

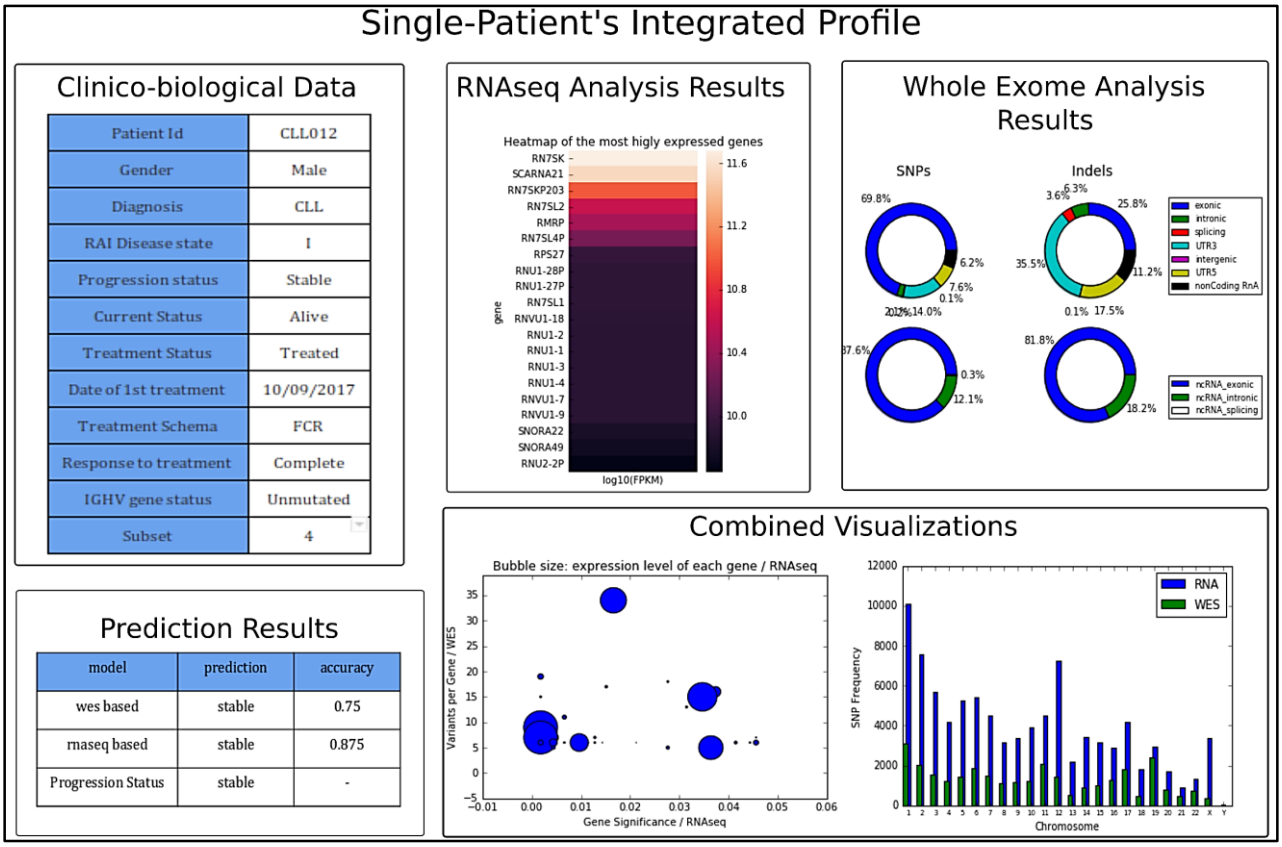

Figure 7. Integrated Genomic Profile.

4- *Discussion*

As already discussed, with the rapid growth of NGS technologies, the variety and volume of genomic data has increased. So, it is of great importance to integrate this wealth of data in order to build an integrated profile that characterizes each patient and provide a comprehensive view of the nature and origin of a disease. This work has made some first yet important step to this direction. However, there are some issues to be addressed in future work.

The predictive analysis conducted in this study is in preliminary level and needs to be optimized in order to achieve higher performance and accuracy. The potentially complementary findings at structural and functional level (WES & RNAseq, respectively), could be combined to achieve added value. This raises the need for combining the two different models, either on feature level, by performing the feature selection procedure simultaneously and create one integrated prediction model, or, simpler, on result level by combining the classifiers. Evidently, multi-omics data availability is necessary requirement to achieve the integration.

With multi-omics data accompanied with complete clinical meta-data, it will become possible to create more integrated predictive models that will answer to questions as the response to a treatment or overall survival of the patient. Moreover, progressing beyond the proposed presentation of information, a further step includes the development of a complete user interface that will incorporate all the methods described and will provide the information in an efficient way as an aid for the clinicians. This interface will serve as a decision support system by providing immediate answers to the clinicians.

Finally, the type of cancer selected for this study, CLL, was selected as a showcase and is expected that, after the optimization of the methodology and the completion of the user interface, this approach will be applicable in all types of cancer diseases.

During this research, some limitations have been encountered regarding the initial aim. First and most important is that the study was conducted using open data. We decided to use open data in order to ensure reproducibility of the results. In this case, it was difficult to have an open dataset that is accompanied with many clinical information for the patient, such as the nature of the disease (stable/aggressive). To that end, we had to infer the necessary clinical information (to label the patient for the purposes of ML) using specific biomarkers such as gene mutations indicating disease aggressiveness. Moreover, in public databases, there is a lack of multi-omics data. It was not possible to find a dataset which contains more than one genomic raw data for a patient along with proper clinical data. For this reasons, the integration performed is not totally validated and the methodology proposed is a collection of steps that can be used in any cancer case.

4- Conclusion

The proposed integrated profile, although not thoroughly validated, seems like a promising approach as it is able to convey useful and complementary information. After demonstrating the outcome of this study to the experts, we had a positive feedback about the usefulness and importance of the integrated profile. With its detailed and meticulous design, this profile can be established as useful and meaningful tool for clinical decisions.

Concluding, this innovative, exploratory, data-driven approach attempts to make use of the big genomic data by summarizing and presenting them in a way that renders them easily usable and interpretable by health professionals. It focuses on integrating different analyses, descriptive and predictive, creating an end-to-end service that begins raw data input and concludes with a complete summary of the patient status.

5- Acknowledgements

The authors would like to thank the CLL experts of the Institute of Applied Biosciences, Centre for Research and Technology, Thessaloniki, Greece, and especially Dr. Andreas Agathagelidis, for their consulting contribution in this study.

6- Conflict of Interest

The authors declare no conflict of interest.

7- Ethical Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

8- References

[1] Via, Marc, Christopher Gignoux, and Esteban Burchard. “The 1000 Genomes Project: New Opportunities for Research and Social Challenges.” Genome Medicine 2, no. 1 (2010): 3. doi:10.1186/gm124.

[2] Kahn, S. D. “On the Future of Genomic Data.” Science 331, no. 6018 (February 2011): 728-729. doi: 10.1126/science.1197891.

[3] Behjati, S., and P. S. Tarpey. “What is Next Generation Sequencing.”, Research in Practice 98, no. 6 (August 2013): 236–238. doi: 10.1136/archdischild-2013-304340.

[4] Nagymihály, Marianna, Attila Szűcs, and Attila Kereszt. "Next-Generation Sequencing and its new possibilities in medicine." Acta Biologica Szegediensis 59, no. suppl. 2. (2015): 323-339.

[5] Warr, A., C. Robert, D. Hume, A. Archibald, N. Deeb, and M. Watson. “Exome Sequencing: Current and Future Perspectives.” G3 5, no. 8 (August 2015): 1543–1550. doi: 10.1534/g3.115.018564.

[6] Wang, Z., M. Gerstein, and M. Snyder. “RNA-Seq: a revolutionary tool for transcriptomics.” Nature Reviews Genetics 10, no. 1 (January 2009): 57–63. doi: 10.1038/nrg2484.

[7] Guo, Y., X Ding., Y. Shen, G. J. Lyon, and K. Wang. “SeqMule: automated pipeline for analysis of human exome/genome sequencing data.” Scientific Reports 5 (September 2015). doi: 10.1038/srep14283.

[8] Hintzsche, J., J. Kim, V. Yadav, C. Amato, S. E. Robinson, E. Seelenfreund, Y. Shellman, J. Wisell, A. Applegate, M. McCarter, N. Box, J. Tentler, S. De, W.A. Robinson, and A. C. Tan. “IMPACT: a whole-exome sequencing analysis pipeline for integrating molecular profiles with actionable therapeutics in clinical samples.” Journal of the American Medical Informatics Association 23, no. 4 (July 2016): 721–730. doi: 10.1093/jamia/ocw022.

[9] D'Antonio, M., P. D'Onorio De Meo, D. Paoletti, B. Elmi, M. Pallocca, N. Sanna, E. Picardi, G. Pesole, and T. Castrignanò. “WEP: a high-performance analysis pipeline for whole-exome data.” BMC Bioinformatics 14, no. 7 (April 2013). doi: 10.1186/1471-2105-14-S7-S11.

[10] Wang, K., M. Li, and H. Hakonarson. “ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.” Nucleic Acids Research 38, no. 16 (September 2010): 1-7. doi: 10.1093/nar/gkq603.

[11] Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. Kelley, H. Pimentel, S. Salzberg, J. Rinn, and L. Pachter. “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.” Nature Protocol 7, no 3 (March 2012): 562-578. doi:10.1038/nprot.2012.016.

[12] Pertea, M., D. Kim, G. Pertea, J. T. Leek, and S. L. Salzberg. “Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown.” Nature Protocol 11, no. 9 (September 2016): 1650–1667. doi: 10.1038/nprot.2016.095.

[13] Cornwell, M., M. Vangala, L. Taing, Z. Herbert, J. Köster, B. Li, H. Sun, T. Li, J. Zhang, X. Qiu, M. Pun, R. Jeselsohn, M. Brown, S. Liu, and H. Long. "VIPER: Visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis." BMC Bioinformatics 19 (April 2018): 135-148. doi: 10.1186/s12859-018-2139-9.

[14] Fonseca, N., R. Petryszak, J. Marioni, and A. Brazma. "iRAP - an integrated RNA-seq Analysis Pipeline." bioRxiv (June 2014). doi: 10.1101/005991.

[15] Robinson, J. T., H. Thorvaldsdóttir,W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. "Integrative genomics viewer." Nature Biotechnology 29, no.1 (January 2011): 24–26. doi: 10.1038/nbt.1754.

[16] Chelaru, F., L. Smith, N. Goldstein, and H. Bravo. "Epiviz: interactive visual analytics for functional genomics data." Nature Methods 11, no. 9 (September 2014): 938–940. doi: 10.1038/nmeth.3038.

[17] http://icbi.i-med.ac.at/software/timiner/doc/index.html.

[18] Codina-Solà, M., B. Rodríguez-Santiago, A. Homs, J. Santoyo, M. Rigau, G. Aznar-Laín, M. del Campo, B. Gener, E. Gabau, M. P. Botella, A. Gutiérrez-Arumí, G. Antiñolo, L. A. Pérez-Jurado. "Integrated analysis of whole-exome sequencing and transcriptome profiling in males with autism spectrum disorders." Molecular Autism 6 (April 2015): 21:36. doi: 10.1186/s13229-015-0017-0.

[19] Wilkerson, M. D., C. R. Cabanski, W. Sun, K. A. Hoadley, V. Walter, L. E. Mose, M. A. Troester, P. S. Hammerman, J. S. Parker, C. M. Perou, and D. N. Hayes. "Integrated RNA and DNA sequencing improves mutation detection in low purity tumors." Nucleic Acids Research 42, no. 13 (July 2014): e107. doi: 10.1093/nar/gku489.

[20] Landesfeind, M., B. Zeitouni, A. Peille, and V. Vuaroqueaux. "Combining whole-exome and RNA-Seq data improves the quality of PDX mutation profiles." Cancer Research 76, no. 14 (July 2016): 2701-2701. doi: 10.1158/1538-7445.AM2016-2701.

[21] Cappelli, E., G. Felici, and E. Weitschek. "Combining DNA methylation and RNA sequencing data of cancer for supervised knowledge extraction." BioData Mining 11, no. 22 (October 2018). doi: 10.1186/s13040-018-0184-6.

[22] Li, C., J. Lee, J. Ding, and S. Sun. "Integrative analysis of gene expression and methylation data for breast cancer cell lines." BioData Mining 11, no. 13 (June 2018). doi: 10.1186/s13040-018-0174-8.

[23] Fleck, J. L., A. B. Pavel, and C. G. Cassandras, "Integrating mutation and gene expression cross-sectional data to infer cancer progression." BMC Systems Biology 10, no. 12 (January 2016). doi: 10.1186/s12918-016-0255-6.

[24] Yu, H., D. C. Samuels, Y. Zhao, and Y. Guo. "Architectures and accuracy of artificial neural network for disease classification from omics data." BMC Genomics 20, no. 167 (March 2019). doi: 10.1186/s12864-019-5546-z.

[25] Zafeiris, D., S. Rutella, and G. R. Ball. "An Artificial Neural Network Integrated Pipeline for Biomarker Discovery Using Alzheimer's Disease as a Case Study." Computational and Structural Biotechnology Journal 16 (February 2018): 77-87. doi: 10.1016/j.csbj.2018.02.001.

[26] Young, E, D Noerenberg, L Mansouri, V Ljungström, M Frick, L-A Sutton, S J Blakemore, et al. "EGR2 Mutations Define a New Clinically Aggressive Subgroup of Chronic Lymphocytic Leukemia." Leukemia 31, no. 7 (November 28, 2016): 1547–1554. doi:10.1038/leu.2016.359.

[27] Moreno, C., and E. Montserrat. "Genetic lesions in chronic lymphocytic leukemia: what's ready for prime time use?" Haematologica 95, no. 1 (January 2010): 12–15. doi: 10.3324/haematol.2009.016873.

[28] https://en.wikipedia.org/wiki/Minor_allele_frequency.

[29] Liu, X., X. Jian, and E. Boerwinkle. "dbNSFP: A Lightweight Database of Human Nonsynonymous SNPs and Their Functional Predictions." Human Mutation 32, no. 8 (August 2011): 894–899. doi: 10.1002/humu.21517.

[30] Trapnell, C., L. Pachter, and S. L. Salzberg. "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics 25, no. 9 (May 2009): 1105-1111. doi: 10.1093/bioinformatics/btp120.

[31] Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nature Biotechnology 28, no. 5 (May 2010): 511-515. doi: 10.1038/nbt.1621.

[32] Trapnell, C., D.G. Hendrickson, M. Sauvageau, L. Goff, J.L. Rinn, and L. Pachter. "Differential analysis of gene regulation at transcript resolution with RNA-seq." Nature Biotechnology 31, no. 1 (January 2013): 46-53. doi: 10.1038/nbt.2450.

[33] Goff, L., C. Trapnell, and D. Kelley. "cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data. R package version 2.8.2" (2013).

[34] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research 12 (October 2011): 2825−2830. doi: hal-00650905v1.