A REVIEW ON COMPUTATIONAL APPROACHES FOR IDENTIFYING NON-SYNONYMOUS MUTATION AND DNA METHYLATION IN CANCER

Binisha.R.B [1], Jeyakumar Natarajan [2], Shanmughavel Piramanayagam [1,*].
Computational Biology lab, Department of Bioinformatics, Bharathiar University, Pin:641046

ABSTRACT

Cancer is a complex disease where, genetic and epigenetic changes plays a major role. Non-synonymous mutations and DNA methylation are important factors in cancer starts and development. Non-synonymous mutations are genetic changes that cause amino acid substitutions within proteins, which frequently result in defective or carcinogenic products. DNA methylation, on the other hand, is the addition of a methyl group to cytosine residues, which affects gene expression and contributes to cancer. Understanding the underlying mechanisms and designing targeted therapeutics require identifying these molecular events in cancer genomes. The traditional methods of identifying the mutations and DNA methylation has made a great challenge in the research fields. Computational approaches have shown to be important in this attempt. They cover a wide range of approaches, from mutation variant calling procedures to methylation pattern analysis tools. These methods will make use of high-throughput sequencing data to detect and characterize non-synonymous mutations and DNA methylation changes on a genome-wide scale. In this review, we will discuss about the multi omics dataset retrievals, tools used for the identification of mutated non- synonymous mutation and the algorithms and techniques used for DNA Methylation for identifying mutated regions. The critical values for the specific results also been noted in this study.

Keywords: DNA Methylation, Non-synonymous genomic mutation, cancer, bioinformatics tools, algorithms

INTRODUCTION

By combining or integrating various omics approaches would provide more accurate results in disease biology. Dr.Hood established the notion of integrated omics in 2003, proposing a systems biology method for the integration of various omics data [1]. Hence, integrated omics or multiomics is described as a biological analysis strategy in which more than two omics measurements are done in the same cell, organ/tissue, or body at the same time [2]. This integrative approach is world widely accepted for identifying the real cause of human diseases, involving highly dynamic and interactive system of molecular layers such as genetics, epigenetics, mRNA transcripts, proteins, and metabolites that are influenced by a variety of environmental stimuli. The next crucial step is to gain a comprehensive knowledge of the molecular information flow and the interactive molecular system, which can only be accomplished by investigating numerous layers of omics data at the same time. Incorporating multi omics measurements from population samples into multidimensional network and system analyses will fill gaps in our existing understanding of molecular mediation processes, gene-environment interactions, and longitudinal impacts throughout chronic illness development [3].

This integrated approach of multi-omics data may improve understanding of the molecular dynamics underlying disease pathophysiology and lead to novel ways for illness detection, prevention, and therapy in humans[4]. They aid in analyzing the flow of information from one omics level to the next, thereby bridging the genotype-to-phenotype gap [5]. For example,an integrative analysis of ChIP-Seq and RNA-Seq data from HNSCC cell lines revealed that cancer-specific histone marks, H3K4me3 and H3K27ac, are related with transcriptional alterations in HNSCC driver genes, epidermal growth factor receptor (EGFR), FGFR1 and FOXA1 [6]. Also, the creation of dedicated repositories such as GEO (Gene Expression Omnibus), TCGA (The Cancer Genome Atlas), or cBioPortal, that contain different datasets covering diverse diseases and allow users to readily access and study them. Several algorithmic approaches for doing multi-omics analysis have been presented in recent years which includes iCluster+, Jive, SNF [7]. Single omics approach includes Genomics, transcriptomics, Epigenomics, Metabolomics. Each omics serves an important role in the field of molecular research.

Genomics is a field of biology that focuses on the study of an organism's entire genome, which is the complete set of its genetic material, including all of its genes and non-coding sequences of DNA. Genome data are collected using high throughput sequencing techniques which includes illumina sequencing, ion torrent sequencing, Roche sequencing [8]. There are different types of genomics which includes structural, functional, comparative and mutational genomics. The Structural genomics is concerned with determining the structure of proteins and other molecules encoded by the genome. This information can be utilized to better understand how these molecules function and interact with one another. The Functional genomics is concerned with determining the function of genes and proteins. This data can be used to find novel medication targets, improve diagnostic procedures, and better understand the molecular underpinnings of diseases. In Comparative genomics analyses the genomes of different species in order to uncover similarities and differences. This information can be utilized to better understand the evolution of life and to identify genes involved in specific features. The Mutational genomics is a branch of biology that studies the finding and analysis of DNA mutations. Mutations can occur naturally or as a result of environmental causes such as radiation or chemical exposure. Mutations can have a wide range of impacts on the organism, ranging from insignificant to fatal.

Epigenomics is defined as the identification of chemical changes to DNA or DNA-binding histone proteins across the genome [28]. Epigenetic changes to DNA and histone proteins are a major regulatory mechanism that regulates gene expression and cellular phenotypes [8]. Epigenetic alterations regulate proteins inside the cells and helps to determine whether the genes are switched on or off. DNA methylation, histone alterations, and chromatin remodeling are the epigenetic mechanisms [9] . From this mechanisms, DNA Methylation is considered to be the prevalent cause of cancer.

DNA Methylation In eukaryotes, includes the insertion of a methyl group to the carbon 5th position of the cytosine ring. The continuous sequence of 5′-CG-3′, also known as a CpG dinucleotide [10]. That is, the DNA-MTase target site in DNA is the dinucleotide palindrome CG (also known as CpG, with p signifying the phosphate group). The process of methylation is catalysed by an enzyme called DNA methyltransferase. Dnmts are enzymes that transfer a methyl group from S-adenyl methionine (SAM) to the fifth carbon of a cytosine residue to generate 5mC [11].

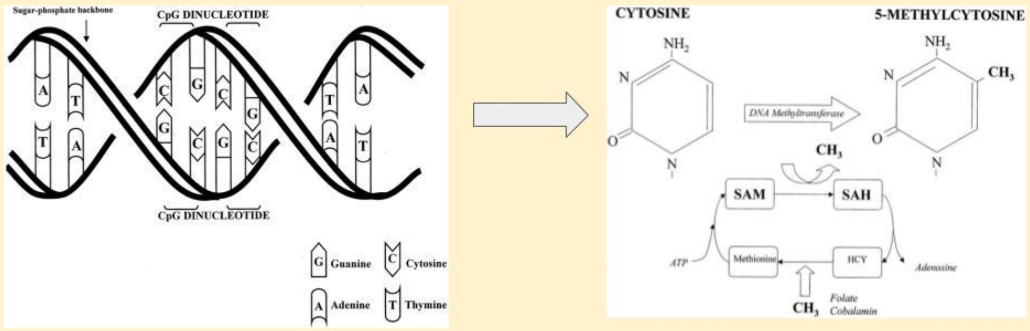
This review will provide a detailed understanding about the DNA Methylation and non-Synonymous genomic mutation which leads to cancer. Also, the computational techniques which are implemented for better data analysis of these mutations.

DNA METHYLATION IN CANCER

DNA methylation is a key regulator of gene transcription, and its function in carcinogenesis has received a lot of attention in recent years. Changes in the DNA methylation patterns are common in most of the cancers. Hypermethylation is the most common epigenetic change, and it suppresses transcription of tumor suppressor gene (TSG) promoter regions, which results in gene silencing. Global hypomethylation, on the other hand, has been identified as a cause of oncogenesis. Many regulatory proteins and enzymes have been discovered as a result of new information about the mechanism of methylation and its control [12]. One of the main enzymes called DNMTs are responsible for methyl group formation, whereas the recently discovered ten-eleven translocation (TET) family of dioxygenases provides a model for DNA demethylation. Through a series of enzymatic processes, these enzymes can convert 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC) and 5-formylcytosine (5fC) to 5-carboxylcytosine (5caC).5mC oxidation promotes the passive loss of DNA methylation during cell replication. Furthermore, via recurrent oxidation followed by base excision repair mediated by thymine DNA glycosylase (TDG), the oxidized intermediates can be returned to cytosine [24].

In normal cell, CpG islands are 1000-base-pair-long sections of DNA that have a higher CpG density than the rest of the genome but are frequently unmethylated. CpG islands appear to have evolved to increase gene expression by controlling chromatin shape and transcription factor binding [11]. When the enzyme DNMT binds to the CpG Islands of promoter regions, the genes get methylated causing increased methylation or hypermethylation and thus no proteins could able bind to the gene and leads to “Turning OFF” of a gene. But the tumor suppressor gene shows a decreased methylation which leads to the “Turning ON” of a gene. Also, S-adenosyl methionine (SAM) provides the methyl group, which is transformed to S-adenosyl homocysteine (SAH) during the process. Also, S-adenosyl methionine (SAM) provides the methyl group, which is transformed to S-adenosyl homocysteine (SAH) during the process [25]. SAH is a competitive inhibitor of methyltransferases and a key indicator and regulator of cellular methylation state. Both the increase in SAH levels and a drop in the SAM: SAH ratio are known to hinder transmethylation processes. The pathogenicity of intracellular SAH depends on its, high affinity binding to the catalytic area of most SAM-dependent methyltransferases [26].

The methylated DNA can be identified by various approaches such as bisulfite conversion-based approach which converts unmethylated cytosine to uracil whereas the methylated remains as cytosines. The second approach is Methylation-sensitive enzyme restriction (MSRE) which is a molecular biology technique used to investigate DNA methylation patterns at specific recognition sites of DNA by utilizing enzymes that are sensitive to the methylation status of cytosine residues and the third approach is affinity enrichment-based where the active binding site contains methylated cytosine [17]. Figure 1. Showing the normal CpG site and the Methylated CpG sites.



(a)Normal DNA CpG site before methylation

(b) Conversion of SAM to SAH, making methylation process more liable and which leads to cancer

COMPUTATIONAL TECHNIQUES RELATED TO DNA METHYLATION IN CANCER

The epigenetic data are mostly retrieved from GEO, TCGA, ArrayExpress, and ENCODE which offers DNA methylation datasets from Infinium HumanMethylation450 or Methylation EPIC arrays. Another database called Epigenetic Wide Association Studies (EWAS) Data Hub collects DNA methylation data as well as associated metadata from 75 344 samples (Figure 2), which include 470 tissue/cell types, 306 illnesses, and other situations like age, sex of the person [15]. MethDB is a database which contain around 20236 methylated data from 6312 individual patterns/ profiles. The retrieved raw datasets need to be filtered by taking only the CpG sites [16]. Primary raw data quality control step includes Trimming of undesirable bases, like as sequencing adapters or undesirable bases resulting from enzymatic end repair, from the reads follows [17]. For comparing the CpG sites, a normal dataset has to be retrieved. The differentially expressed genes (DEGs) and differentially methylated CpG sites (DMCs) between tumor and normal datasets are identified using the R package limma or COHCSAP [18]. Limma is an R/Bioconductor software package that provides a unified approach for analyzing gene expression data. It has many features for dealing with complex experimental designs and borrowing information to overcome the problem of limited sample numbers [26]. Each CpG site's and gene's adjusted p value was computed using the Benjamini-Hochberg (BH) false discovery rate (FDR) method. The cut-off point for determining DEGs and DMCs was a p value with an FDR adjustment of less than 0.05. Then the hypermethylated-low-expression DEGs from the overlapping hypermethylated and downregulated genes, as well as hypomethylated-high-expression DEGs from the overlapping hypomethylated and upregulated genes are to be selected, to test for DNA methylation-derived DEGs [18]. Another method to Weighted gene comethylation network analysis (WGCNA) was used to identify hub methylated-CpG sites and linked genes. Comethylation networks were built using the WGCNA algorithm. The connection of CpG site methylation patterns with clinical status is represented by module membership (MM). The average gene significance (GS) in each module was displayed in a bar graph, which also reflected the association between the module and clinical state (module significance is the absolute average value of the correlation between all the CpGs and the characteristic. [27]. The DAVID database was used to perform a GO function enrichment analysis on DEGs in order to elucidate the role of the identified DEGs in cancer carcinogenesis and progression. It is a set of data-mining tools that systematically combine functionally descriptive data with intuitive graphical displays. It provides exploratory visualization tools that facilitate discovery through functional classification, biochemical route maps, and conserved protein domain designs, while keeping linked to rich biological annotation sources [28]. For analysis, the default settings were used. Functions enriched at p 0.05 were deemed relevant in the context of the entire human genome. Differentially expressed mRNA was transformed from Gene Symbol to Entrez ID using the R package "org.Hs.eg.db," [30]. Furthermore, for KEGG pathway enrichment analysis on Methylated DEGs can be the KEBAS database was used.The KEGG pathway was tested for significance value of p 0.05 [19]. Pathway-centric approaches can be used to interpreting and -omics data. This approach will provide different statistical enrichment analyses and will provide a clearcut model of a protein. GSEA (gene set enrichment analysis) is another popular method for analyzing transcriptome and epigenome datasets. Genes in GSEA are often ranked by phenotypic parameters, such as gene expression levels. This will calculate the enrichment score for a pathway by first scanning the ranked gene list from top to bottom, then calculating the distances between the center of the rank and all genes tagged to this pathway [29]. The acquired gene network information was loaded into Cytospace software, and the CytoHubba plug-in determined the connection score of each protein node.[30]. The figure 2 shows the steps and tools that are employed for identifying differentially methylated regions.

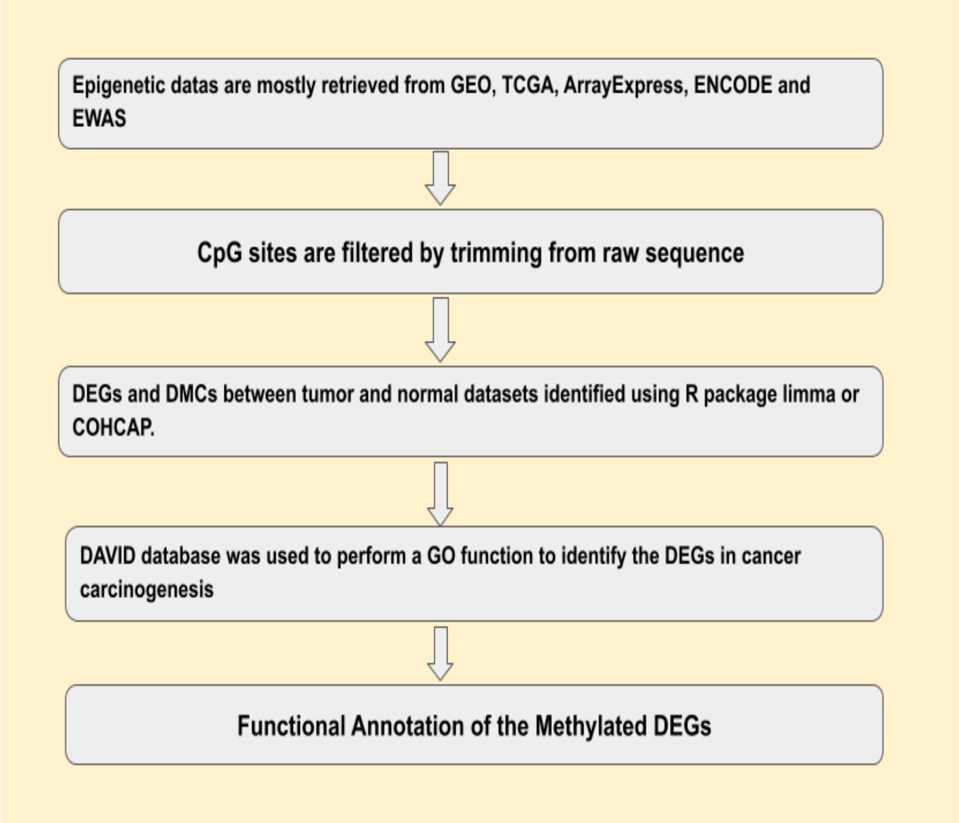

Epigenetic datas are mostly retrieved from GEO, TCGA, ArrayExpress, ENCODE and EWAS
CpG sites are filtered by trimming from raw sequence
DEGs and DMCs between tumor and normal datasets identified using R package limma or COHCAP.
DAVID database was used to perform a GO function to identify the DEGs in cancer carcinogenesis
Functional Annotation of the Methylated DEGs

Figure 2: Flow chat for the bioinformatics approach in DNA Methylation in cancer.

NON SYNONYMOUS GENOMIC MUTATION IN CANCER

A nonsynonymous mutation usually involves the addition or deletion of a single nucleotide in the sequence. Also, frameshift mutation may leads to NSM when a single nucleotide is deleted or inserted, causing the entire reading frame of the amino acid sequence to be altered and the codons will be mixed. This usually has an effect on the amino acids that are coded for and changes the protein that is produced. Non synonymous mutations alter protein sequences and are frequently subjected to gene alteration. Non synonymous mutations that modify protein sequences in the coding region of DNA could result in cancer. Nonsynonymous mutations are regarded to be mostly harmful due to their ability to change amino acids [21]. Oncogenes and tumor suppressors are two main categories of genes associated with oncogenesis. In the context of gain-of-function mutations i.e., those resulting to enhanced expression, oncogenes initiate and accelerate the tumorigenic process, whereas tumor suppressor genes give a growth benefit to cells when they acquire loss-of-function mutations [20].

COMPUTATIONAL TECHNIQUES TO IDENTIFY NON SYNONYMOUS GENOMIC MUTATION FOR CANCER STUDIES

The non synonymous datasets for a targeted SNP gene sequence is retrieved from the dbSNP in NCBI (National Centre for Biotechnology Information Website) database and Human Genome Mutation Database. To determine the influence of single amino acid polymorphisms, or SNPs, on a protein, a large variety of computational bioinformatics webservers have recently been built. PROVEAN, SIFT, SNAP2, FATHMM, PON-P2, and Predict SNP are some of the tools which are used to predict SNPs from a particular gene sequence. From this, SIFT, SNAP-2, and PROVEAN and PON-P2 are some of the tools which produce results that fall into either the tolerated (non-pathogenic) or harmful category. SIFT techniques, as the primary choice in SNP characterization, are frequently utilized in all computational study to predict the harmful nsSNPs. SIFT uses protein homology sequences with orthologous and paralogous protein sequences to identify nsSNPs as cancerous. A SIFT value of >0.05 indicates tolerance, but a score of 0.05 indicates negative consequences of nsSNPs on protein function or structure. The Protein Variation Effect Analyzer (PROVEAN) algorithm that uses delta alignment scores based on the variant version and reference of the protein sequence to forecast malignant variations. Malignant nsSNPs were identified in PROVEAN by comparing the score below the threshold value of 2.5. SNAP2 focuses on advanced machine-learning and neural network-based approaches to classify harmful nsSNPs and their functional impacts (effect or neutral) in protein. It uses functional and structural annotations, sequence,and evolutionary features of the query protein to predict changes in protein function (gain or loss). SNAP2 determines if a mutation is neutral (ranges from -100 to 0) or effect (ranges from 0 to +100) with good accuracy [22].

PMut is a tool which is used to confirm the nsSNPs which were identified before. The forecast has a range of 0-1, with 0-0.5 being considered neutral and 0.5-1 being deemed disease. Protein stability was determined using a tool called I-Mutant3.0. It employs on the principle of ProTherm, which is now the most complete library of thermodynamic experimental data on protein stability when modified (measured as free energy change value DDG). Its predictions are based on two support vector machines (SVM). It also predicts the reliability index (RI) of the results, which ranges from 0 to 10, with 10 being the most reliable. Aside from entering the protein sequence and mutation locations, the temperature and pH settings remained constant (25 °C, pH 7) [23]. ConSurf is a database which is used to evaluate the evolutionary conservation of amino acid residues in a protein using phylogenetic relationships between homologous sequences [25]. Finally, the structure of the particular non synonymous protein is created by I-TESSAR tool. The figure 3 shows the steps and the tools used for the identification of Non-synonymous SNP mutation using bioinformatics tools and databases.

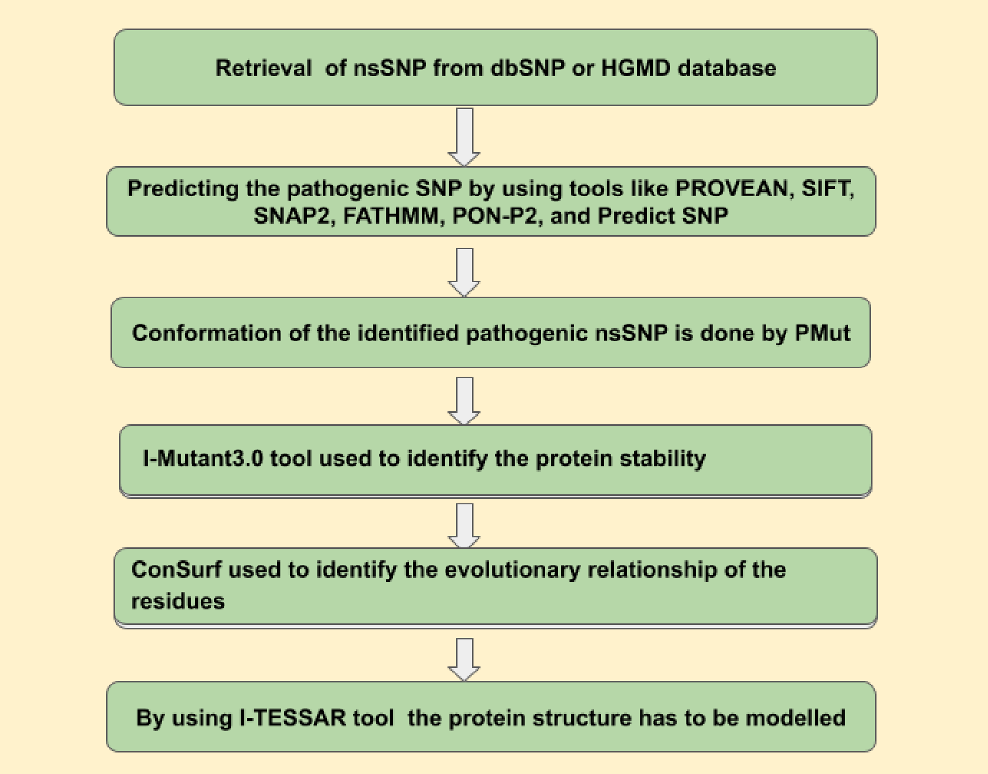

Retrieval of nsSNP from dbSNP or HGMD database
Predicting the pathogenic SNP by using tools like PROVEAN, SIFT, SNAP2, FATHMM, PON-P2, and Predict SNP
Conformation of the identified pathogenic nsSNP is done by PMut
I-Mutant3.0 tool used to identify the protein stability
ConSurf used to identify the evolutionary relationship of the residues
By using I-TESSAR tool the protein structure has to be modelled

Figure 3: Flow chart of Computational techniques in ns-protein identification

CONCLUSION

This review study explored into the complex world of non-synonymous mutations and DNA methylation in cancer. The bioinformatics approaches which are implemented rather than the traditional wet lab techniques, which is considered to be time consuming. This study provides a better understanding of tools and techniques used for genomic mutation and DNA Methylation identification in the field of cancer. In Future, by combing these approaches we can able to identify the cause of deadly cancer and can provide a way for proper drug discovery.

REFERENCES

1. Shin, T. H., Nithiyanandam, S., Lee, D. Y., Kwon, D. H., Hwang, J. S., Kim, S. G., ... & Lee, G. (2021). Analysis of nanotoxicity with integrated omics and mechanobiology. *Nanomaterials*, *11*(9), 2385.
2. Wang, X. (2018). Clinical trans-omics: an integration of clinical phenomes with molecular multiomics. *Cell Biology and Toxicology*, *34*, 163-166.
3. McShane, L. M., Cavenagh, M. M., Lively, T. G., Eberhard, D. A., Bigbee, W. L., Williams, P. M., ... & Conley, B. A. (2013). Criteria for the use of omics-based predictors in clinical trials: explanation and elaboration. *BMC medicine*, *11*(1), 1-22.
4. Sun, Y. V., & Hu, Y. J. (2016). Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Advances in genetics*, *93*, 147-190.
5. Nguyen, N., Jennen, D., & Kleinjans, J. (2022). Omics technologies to understand drug toxicity mechanisms. *Drug Discovery Today*, 103348.
6. Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, *14*, 1177932219899051.
7. Fiorentino, G., Visintainer, R., Domenici, E., Lauria, M., & Marchetti, L. (2021). Mousse: Multi-omics using subject-specific SignaturEs. *Cancers*, *13*(14), 3423.
8. Bhati, A., Garg, H., Gupta, A., Chhabra, H., Kumari, A., & Patel, T. (2012). Omics of cancer. *Asian Pacific Journal of Cancer Prevention*, *13*(9), 4229-4233.
9. Plass, C., Pfister, S. M., Lindroth, A. M., Bogatyrova, O., Claus, R., & Lichter, P. (2013). Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. *Nature reviews genetics*, *14*(11), 765-780.
10. Singal, R., & Ginder, G. D. (1999). DNA methylation. *Blood, The Journal of the American Society of Hematology*, *93*(12), 4059-4070.
11. Moore, L. D., Le, T., & Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology*, *38*(1), 23-38.
12. Das, P. M., & Singal, R. (2004). DNA methylation and cancer. *Journal of clinical oncology*, *22*(22), 4632-4642.
13. Merkel, A., & Esteller, M. (2022). Experimental and bioinformatic approaches to studying DNA methylation in cancer. *Cancers*, *14*(2), 349.

14. Xiong, Z., Li, M., Yang, F., Ma, Y., Sang, J., Li, R., ... & Bao, Y. (2020). EWAS Data Hub: a resource of DNA methylation array data and metadata. *Nucleic Acids Research*, *48*(D1), D890-D895.
15. Merkel, A., & Esteller, M. (2022). Experimental and bioinformatic approaches to studying DNA methylation in cancer. *Cancers*, *14*(2), 349.
16. Yang, Y., Chu, F. H., Xu, W. R., Sun, J. Q., Sun, X., Ma, X. M., ... & Wang, X. M. (2017). Identification of regulatory role of DNA methylation in colon cancer gene expression via systematic bioinformatics analysis. *Medicine*, *96*(47).
17. Zhao, L., Jia, Y., Liu, Y., Han, B., Wang, J., & Jiang, X. (2022). Integrated bioinformatics analysis of DNA methylation biomarkers in thyroid cancer based on TCGA database. *Biochemical Genetics*, *60*(2), 629-639.
18. Sun, H., Xin, R., Zheng, C., & Huang, G. (2021). Aberrantly DNA methylated-differentially expressed genes in pancreatic cancer through an integrated bioinformatics approach. *Frontiers in genetics*, *12*, 583568.
19. Gotea, V., Gartner, J. J., Qutob, N., Elnitski, L., & Samuels, Y. (2015). The functional relevance of somatic synonymous mutations in melanoma and other cancers. *Pigment cell & melanoma research*, *28*(6), 673-684.
20. Chu, D., & Wei, L. (2019). Nonsynonymous, synonymous and nonsense mutations in human cancer-related genes undergo stronger purifying selections than expectation. *BMC cancer*, *19*(1), 1-12.
21. Islam, R., Rahaman, M., Hoque, H., Hasan, N., Prodhan, S. H., Ruhama, A., & Jewel, N. A. (2021). Computational and structural based approach to identify malignant nonsynonymous single nucleotide polymorphisms associated with CDK4 gene. *Plos one*, *16*(11), e0259691.
22. Chai, C. Y., Maran, S., Thew, H. Y., Tan, Y. C., Rahman, N. M. A. N. A., Cheng, W. H., ... & Yap, W. S. (2022). Predicting deleterious non-synonymous single nucleotide polymorphisms (nsSNPs) of HRAS gene and in silico evaluation of their structural and functional consequences towards diagnosis and prognosis of cancer. *Biology*, *11*(11), 1604.
23. Lakshminarasimhan, R., & Liang, G. (2016). The role of DNA methylation in cancer. *DNA Methyltransferases-Role and Function*, 151-172.
24. Wajed, S. A., Laird, P. W., & DeMeester, T. R. (2001). DNA methylation: an alternative pathway to cancer. *Annals of surgery*, *234*(1), 10.

25. Kerins, D. M., Koury, M. J., Capdevila, A., Rana, S., & Wagner, C. (2001). Plasma S-adenosylhomocysteine is a more sensitive indicator of cardiovascular disease than plasma homocysteine. *The American journal of clinical nutrition*, *74*(6), 723-729.
26. 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-e47.
27. Dai, Y., Lv, Q., Qi, T., Qu, J., Ni, H., Liao, Y., ... & Qu, Q. (2020). Identification of hub methylated-CpG sites and associated genes in oral squamous cell carcinoma. *Cancer Medicine*, *9*(9), 3174-3187.
28. Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome biology*, *4*(9), 1-11.
29. Garcia-Moreno, A., López-Domínguez, R., Villatoro-García, J. A., Ramirez-Mena, A., Aparicio-Puerta, E., Hackenberg, M., ... & Carmona-Saez, P. (2022). Functional enrichment analysis of regulatory elements. *Biomedicines*, *10*(3), 590.
30. Zhu, K., Wang, L., Chen, Z., Ding, J., Dong, J., & Chen, J. (2022). Analysis and construction of an exosome derived competing endogenous RNA network for small cell lung cancer in the exoRbase database.