Current Genomics, 2024, *25*, 41-64

RESEARCH ARTICLE

Exploring the Role of Non-synonymous and Deleterious Variants Identified in Colorectal Cancer: A Multi-dimensional Computational Scrutiny of Exomes

Chandrashekar Karunakaran[1], Vidya Niranjan[1,*], Anagha S. Setlur[1], Dhanya Pradeep[2] and Jitendra Kumar[3,*]

[1]*Department of Biotechnology, R V College of Engineering, Bangalore, 560059, affiliated to Visveswaraya Technological University, Belagavi, 590018, India;* [2]*Department of Biotechnology, BMS College of Engineering, Bangalore, 560019, India;* [3]*Biotechnology Industry Research Assistance Council (BIRAC), CGO complex Lodhi Road, New Delhi, India*

ARTICLE HISTORY

Received: October 03, 2023
Revised: November 25, 2023
Accepted: December 13, 2023

DOI:
10.2174/0113892029285310231227105503

Abstract: ***Introduction*:** Colorectal cancers are the world's third most commonly diagnosed type of cancer. Currently, there are several diagnostic and treatment options to combat it. However, a delay in detection of the disease is life-threatening. Additionally, a thorough analysis of the exomes of cancers reveals potential variation data that can be used for early disease prognosis.

***Methods*:** By utilizing a comprehensive computational investigation, the present study aimed to reveal mutations that could potentially predispose to colorectal cancer. Ten colorectal cancer exomes were retrieved. Quality control assessments were performed using FastQC and MultiQC, gapped alignment to the human reference genome (hg19) using Bowtie2 and calling the germline variants using Haplotype caller in the GATK pipeline. The variants were filtered and annotated using SIFT and PolyPhen2 successfully categorized the mutations into synonymous, non-synonymous, start loss and stop gain mutations as well as marked them as possibly damaging, probably damaging and benign. This mutational profile helped in shortlisting frequently occurring mutations and associated genes, for which the downstream multi-dimensional expression analyses were carried out.

***Results*:** Our work involved prioritizing the non-synonymous, deleterious SNPs since these polymorphisms bring about a functional alteration to the phenotype. The top variations associated with their genes with the highest frequency of occurrence included *LGALS8*, *CTSB*, *RAD17*, *CPNE1*, OPRM1, *SEMA4D*, *MUC4*, *PDE4DIP*, *ELN* and *ADRA1*A. An in-depth multi-dimensional downstream analysis of all these genes in terms of gene expression profiling and analysis and differential gene expression with regard to various cancer types revealed *CTSB* and *CPNE1* as highly expressed and overregulated genes in colorectal cancer.

***Conclusion*:** Our work provides insights into the various alterations that might possibly lead to colorectal cancer and suggests the possibility of utilizing the most important genes identified for wet-lab experimentation.

Keywords: Colorectal cancer, exome analysis, non-synonymous, deleterious mutations, mutational profiling, multi-dimensional genomics, CTSB, CPNE1.

1. INTRODUCTION

With increasing incidences of cancer globally, colorectal cancer currently stands as the third most commonly diagnosed type of cancer. As of 2022, the American Cancer Society evaluations for the number of colorectal cancers in the United States were 1,06,180 fresh cases of colon cancer and 44,850 new cases of rectal cancer [1], with deaths estimated to be about 52,580 including both men and women. Despite there being several screening techniques for the prevention of colorectal cancer and treatment strategies such as surgery, chemotherapy and radiation for colorectal cancer, it still lacks a suitable prognosis. Exomes are the coding region of a gene. A thorough analysis of the exomes of cancers reveals potential variation data that can be used for early disease prognosis. Mutations in genes such as driver genes, tumor suppressor genes, proto-once and oncogenes trigger cancers by causing DNA damage and genetic instability [2].

Currently, pursuing recurring variations depending on the frequency by which a gene is altered is the current strate-

* Address correspondence to these authors at the Department of Biotechnology, R V College of Engineering, Bangalore, 560059, India;
E-mail: vidya.n@rvce.edu.in; and Biotechnology Industry Research Assistance Council (BIRAC), CGO complex Lodhi Road, New Delhi, India;
E-mail: md.birac@nic.in

1875-5488/24 $65.00+.00 © 2024 Bentham Science Publishers

Current Genomics

gy for the analysis of cancer exome sequences. Some other techniques include analyzing the predicted score for the impact the mutations have on the structure or function of a protein and identifying the spatial clusters of variations with respect to each residue [3]. Moreover, the identification of all cancer-causing genes from the exomes of cancers is a huge task suffused with numerous challenges. There still remain uncertainties as to the constitution of specific cancer-causing genes. Therefore, the lacunae existing currently in cancer exome research can be filled by a thorough investigation of all genes in the exome, including different mutations such as somatic and germline alterations. Thus, high-throughput analyses of cancer exomes using *in-silico* strategies are now paving the way for swift mutation identification and analysis.

The demerits of Sanger sequencing [4] allowed for the development of next-generation sequencing technology (NGS) that sequences several thousand samples simultaneously at high accuracy and precision [5]. Advancements in this technology have now paved the way for the identification of rare somatic and germline mutations [6]. With whole exome sequencing (WES) being utilized for the detection of cancers, the identification of prognostic and diagnostic biomarkers also becomes an essential part of this process. It is crucial in the design of suitable treatment strategies as well as early recognition of the disease. Since the detection of cancer and its prognosis is dependent on the understanding and analysis of the molecular pathways involved, there is a great need to unravel this data through the examination of existing cancer exomes to primarily broaden the prospect of early detection and early treatment.

At present, for colorectal cancer, genes such as *KRAS, BRAF* and *APC* are considered to be reliable markers that point towards its detection [7]. Additionally, *MSI* (microsatellite instability) is a prognostic marker for colorectal cancer that corresponds to a phenotype observed in colorectal cancer when genes of the mismatch repair pathway are altered [8]. This shows that the identification of mutations is essential for the detection of any potential markers. Moreover, non-synonymous germline variations that bring about a significant alteration to the function of a protein can be pivotal in determining genes that could predispose to colorectal cancer. These alterations could also serve as potential susceptibility markers, particularly while screening for a panel of variants that might incrementally contribute to the increased risk of colorectal cancer, suggesting the cumulative burden of these variants. Determination of SNPs that regulate gene expression in driver genes is crucial to understanding the complex mechanisms underlying colorectal cancer, thereby enabling early diagnosis and improved treatment outcomes. Therefore, the present study focused on extensively investigating 10 colorectal cancer exomes for the identification of different variations that could potentially point towards the disease and aid in early detection.

2. MATERIALS AND METHODS

2.1. Dataset Retrieval

Colorectal cancer exome sample exomes for ten NGS sequenced samples were retrieved from the publicly available NCBI SRA (National Centre for Biotechnology Information, Sequence Read Archive) database [9]. For comparative analysis, the human reference genome hg19 (Genome Reference Consortium Human Reference 19) was also retrieved using the Genome Reference Consortium database (https://www.ncbi.nlm.nih.gov/grc/human). All the sequence data files were downloaded using sra-toolkit in.SRA format and then converted to fastq. These were then split into forward and reverse reads. Table **1** details the various colorectal cancer exomes that were used in the study.

2.2. Pre-processing of Raw Data

All ten colorectal cancer exomes were pre-processed prior to the calling of variants.

Table 1. Details of colorectal cancer exomes used in the study.

Sl. No.	Colorectal Cancer Dataset	Design	Tissue	Isolate	Sex	Age	BioProject	BioSample
1.	SRR14684620	xGen System	FFPE*	CC_9	Male	76	PRJNA733593	SAMN19416430
2.	SRR15987777	Illumina TruSeq Exome	FFPE	P9	NA	-	PRJNA764756	SAMN21527883
3.	SRR14463450	SureSelectXT reagent kit; all exon v5 probeset	-	Colon cancer cell	Male	62	PRJNA726023	SAMN18928492
4.	SRR15987799	Illumina TruSeq Exome	FFPE	P13	NA	-	PRJNA764756	SAMN21527863
5.	SRR15987796	Illumina TruSeq Exome	FFPE	P16	NA	-	PRJNA764756	SAMN21527866
6.	SRR15987795	Illumina TruSeq Exome	FFPE	P17	NA	-	PRJNA764756	SAMN21527867
7.	SRR15987790	Illumina TruSeq Exome	FFPE	P20	NA	-	PRJNA764756	SAMN21527871
8.	SRR15987786	Illumina TruSeq Exome	FFPE	P24	NA	-	PRJNA764756	SAMN21527875
9.	SRR15987785	Illumina TruSeq Exome	FFPE	P25	NA	-	PRJNA764756	SAMN21527876
10.	SRR15987792	Illumina TruSeq Exome	FFPE	P19	NA	-	PRJNA764756	SAMN21527869

Abbreviation: *FFPE: Formalin-Fixed Paraffin-Embedded

2.2.1. Quality Control Checks for Raw Data

FastQC (https://github.com/sadrews/FastQC) and MutiQC (https://github.com/ewels/MultiQC) [8, 9] offer a straightforward yet effective approach to conducting quality assessments on raw sequence data. In a comprehensive review of tools for examining and verifying the quality of cancer genome sequencing data previously, it was identified that the preferred tools for evaluating the quality of these genomes, including exome sequencing data for comprehensive somatic variant calling in human cancer genomes, are FastQC and MultiQC [10]. Consequently, an analysis of mean sequence quality per read and per base, nucleotide content per base position, GC distribution, *etc*., was carried out using FastQC on the raw data sequences in fastq format. Therefore, the mean sequence quality per reading and per base, GC content distribution, nucleotide content per position of the base, adaptor content, *etc*., were all checked using FastQC with input as the raw sequences in fastq format. The quality checks output obtained as HTML reports were scrutinized and analyzed. MultiQC was then employed to obtain cumulative results for quality checks for all ten sequences, which also allowed for better visualization. The HTML reports from FastQC were uploaded, and MultiQc was run parallelly for all 10 raw data sequences. The final output generated showed the cumulative summary output for all the sequences, and the quality for each selected sequence was analyzed in terms of three categories: pass, warning and fail, using the log files obtained from FastQC quality runs for each of the forward and reverse reads.

2.2.2. Gapped Alignment and File Conversion

Due to its sensitivity, greater output accuracy and higher speed of operation, Bowtie 2 (https://github.com/BenLangmead/bowtie2)was utilized for the gapped alignment of the sequences [11]. This tool employs the Burrow-Wheeler Transformation (BWT) algorithm, which was used for gapped alignment with the human reference genome hg19. Bowtie 2 uses a combination of BWT and the Smith-Waterman algorithm for performing gapped alignment. Thus, gapped alignment was performed for all ten cancer exome sequences to map the pre-processed reads to the reference. The bowtie2 index files for the genome were first built followed by which the forward and reverse reads were aligned to the reference genome. The output in SAM format was then converted to BAM format using SAMtools (https://github.com/samtools). SAMtools stands out as a widely employed tool for processing data derived from high-throughput sequencing. With enhanced speed and an improved capacity to index files, it facilitates the swift sorting and creation of BAM files [12]. Consequently, in the current investigation, the reads underwent sorting, recalibration of quality scores, realignment of indels, and filtering of the reads using SAMtools.

The alignment qualities were further improved to reduce the occurrence of false variant calls by taking the aligned sequences through several refinement steps. In the current study, conversion of the file from SAM to BAM, sorting of the BAM files and merging were performed for all ten aligned sequences. BAM files act as the binary, compressed files to SAM that are easier to retrieve and use [13]. This conversion was performed since BAM has a compact size and allows for quicker retrieval of the aligned sequences. The outputs obtained were analyzed prior to variant calling.

2.3. Processing and Variant Calling

Processing and calling of variants were carried out to identify the mutations from the sequenced data. Processing of the variants was performed using PICARD, and the variants were called using Haplotype caller from the gold standard method of the GATK pipeline (https://github.com/broadinstitute/gatk, The Genome Analysis Toolkit) [14, 15]. The GATK pipeline, established as the gold standard method since its initial publication in 2010, is renowned for its reliability, with an impressive F-score of 0.978, representing the harmonic mean of precision and recall [16]. Moreover, GATK is recognized for its ability to identify potential variants across diverse sequencing platforms and experimental designs [17], excelling particularly in the accurate discovery of true SNPs in exome datasets. The PCR duplicates were marked using PICARD, and GATK was employed for recalibration of the base quality and local realignment with BAM files as the input. Analysis of the co-variates was carried out by building the BAM indexes, SortSam and base recalibration. Once this was completed, the mutants were called. Outputs were obtained as VCF (Variant calling files) files, and these were scrutinized.

The SNPs that were detected were then filtered and annotated using the snpEFF (http://pcingola.github.io/SnpEff/) open-source tool that is platform-independent, flexible and multi-organism compatible [18]. The output VCF files were obtained after running snpEFF for all ten exomes, and the HTML summary files were then analyzed.

2.4. Post-processing of Variants and Mutational Profiling

SIFT (https://sift.bii.a-star.edu.sg/) was utilised to further process the annotated variants [19]. Results from SIFT tool were also cross-verified using PolyPhen2 (https://github.com/hammerlab/vcf-annotate-polyphen), which also works on the same principle, but classifies the mutations as “benign,” “possibly damaging,” and “probably damaging” [20]. The SIFT protocol assesses whether the amino acid substitution responsible for a variant in cancer exome data affects the protein's function. Utilizing sequence homology, the SIFT algorithm performed predictions on the potential effects of all substitutions at each position in the protein sequence [21]. Batch query files containing information about the chromosome, position of the variant and the specific nucleotide alteration were prepared for each dataset and used as input files for PolyPhen2. Verification was conducted by cross-referencing the results using PolyPhen2 (https://github.com/hammerlab/vcf-annotate-polyphen) [20]. PolyPhen2 operates on a similar principle, classifying identified variants into categories such as “possibly damaging,”

"probably damaging," and "benign" [22]. The results obtained were examined thoroughly, and a mutational profile for the same was developed.

A study published in 2018 highlighted the sensitivity and specificity of tools such as SIFT, PolyPhen and MutationTaster2 [23]. This study identified SIFT to have the highest accuracy and sensitivity (1.0), while Polyphen2 was slightly on the lower sensitivity side. However, another study noted that PolyPhen-2 and SIFT both have a high median sensitivity of 0.90 and 0.85, respectively and have similar median specificity values [24]. Therefore, in the current study, since PolyPhen2 was used for cross-referencing and verification of SIFT results, both these tools were employed together to help produce more reliable and reproducible results.

The exome analysis pipeline was followed by our previous study, Padmavathi *et al.*, 2021 [25].

2.5. Downstream Expression Analysis of Important Mutations

A thorough analysis of all the processed mutations from each of the ten datasets was performed to uncover common and most frequently occurring alterations. The mutations that occurred >= 10 times in each dataset and the overlapping variants across all exomes were examined and screened. A frequency filter was used to screen these and identify the variants. Studies have previously shown that when Haplotype caller from GATK is used for calling variants, germline and somatic can be distinguished from each other by choosing the frequency of occurrence of the variants. A study used 80% or greater allele frequency for classifying as germline and taking it forward for further analysis [26]. In the present work, the authors chose to filter out the variants occurring in most of the exome samples selected since, in the future, using them to point towards a cancer risk would be much easier.

It was later found after mutational profiling that all these mutations belonged to two major categories: non-synonymous and deleterious, which were then shortlisted for downstream gene expression profile analysis using various tools to comprehend the possibility of these genes acting as indicators that predispose to colorectal cancer.

The differential gene expression and comparisons of the expression levels of each of these identified genes were performed to obtain better insights into the potential downstream activities that could reveal those genes with potential cancer risk probabilities that can be explored further from wet lab studies.

2.5.1. Gene Expression Profiling Using GEPIA

A gene expression profiling tool called GEPIA (Gene Expression Profiling Interactive Analysis, http://gepia.cancer-pku.cn/index.html) [27] was utilized for the top identified genes that occurred most frequently throughout all ten colorectal cancer exomes. This tool provides functions that are customizable, such as normal/tumor differential expression analysis, patient survival scrutiny, profiling as per types of cancers, correlation analysis, *etc*. In the present work, the gene "symbol" was provided in the webserver, and a single gene analysis was carried out in the form of box plots. The $|Log_2FC|$ threshold value was maintained at 1, the jitter size at 0.4 and the *p*-value cut-off to 0.01. The expression profile was studied for all the potential genes with respect to colon adenocarcinoma and rectum adenocarcinoma present in the GEPIA database. Once the box plots were obtained, these were analyzed for all the selected genes. The overall survival plots were also generated with a median group cut-off and at a 95% confidence interval. All other parameters, such as hazard ratios and axis units, were set to default. Moreover, a comparative multiple gene expression analysis was also performed to compare the expression of all frequently occurring genes in all ten colorectal cancer exomes. Colon adenocarcinoma and rectum adenocarcinoma were selected as the exomes for this differential gene expression. A heat chart was obtained that revealed which of the genes were predominantly expressed in each cancer type. These results were analyzed.

2.5.2. Differential Expression of Genes in Various Cancer Types

OncoMX (https://www.oncomx.org/) [28] was utilized To understand and compare the expression of each of these genes in colorectal cancer and in other cancer types,. This comparative study will help provide a potential possibility of these genes being overexpressed in other cancer types as well, thereby giving a wider possibility for the detection of a cancer marker. The gene symbols are keyed in as the input for each identified gene, for which upregulation and downregulation plots are generated with regard to different cancer types. This differential expression was studied thoroughly, and the results obtained helped in deciphering important colorectal cancer indicators.

2.5.3. Multi-dimensional Gene Analysis

To further obtain an understanding of the potential genes that incline towards causing colorectal cancer, the most frequent genes were then subjected to a multi-dimensional gene analysis. For this purpose, cBio Cancer Genomics Portal was utilized (http://cbioportal.org)_to interactively explore the various genetic alterations and link it to the clinical outcomes [29, 30]. Graphical representations, network analysis, and visualizations were analyzed. All identified genes that showed potential for pointing towards colorectal cancer were queried as the input against all available colorectal adenocarcinoma studies in the cBioPortal database. From the 13 exomes available on colorectal cancer in cBioPortal, all 4535 patient samples were analyzed and run against our queried genes. From the results obtained, cancer type summary for all genes together, plots of mRNA expression against mutations and the various types of mutations that these genes have caused in different colorectal cancer samples.

With all this data, the potential influencers of colorectal cancer that require further analysis may be identified. This entire protocol used in the study is illustrated in Fig. (**1**).

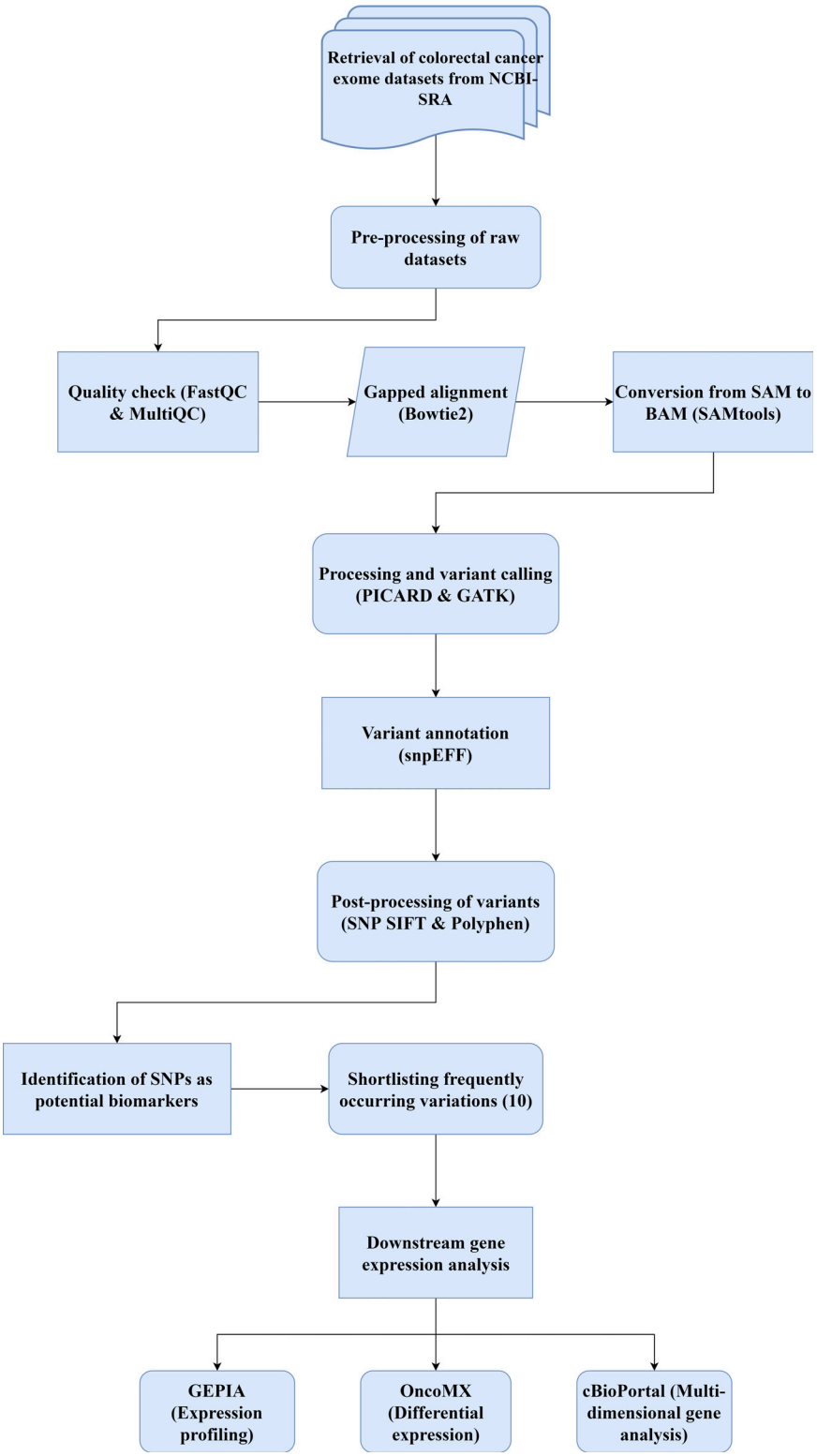


Fig. (1). Flowchart illustrating the protocol used for the entire study. All tools used in the figure have been cited in the reference section. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

2.6. Multi-variate Analysis Using Linear Regression

To establish the correlation between the gene frequencies and the gene expression data identified in the present study, a multi-variate analysis was performed. Linear regression using the least squares method [31] was employed, with 4 estimated parameters. These were:

β0- intercept, β1- gene expression, β2- gene frequency, and β3- correlation between gene expression and frequency.

The analysis was carried out using PRISM 9.0 [32]. The results obtained were scrutinized.

3. RESULTS

3.1. Pre-processing of Raw Data

All 10 colorectal cancer exomes were downloaded successfully and pre-processed to obtain enhanced quality data.

3.1.1. Quality Control Checks

FastQC checks for the ten colorectal cancer exomes showed that all the exomes used for this study were of good quality and could be used for further processing. None of the exomes required adapter trimming. The GC content of all exomes fell in the acceptable range between 40-60%, with SRR15987799 having a maximum GC content of 49%. In addition, all the exomes cleared the duplication levels and did not fall in the warning category. It was also found that the per sequence quality scores were all in the permissible range for all 10 exomes, and none of the exomes were marked with a warning for this parameter. The MultiQC summary report of all the ten exomes, including the individual forward and reverse read files, indicated that all the sequences had a Phred score greater than 20 with an error rate of 0.1% to 1% and an accuracy of ~99%. Further, all the 20 sequence files (duplicates of all 10 ten exomes) passed the per sequence quality scores check. The per sequence GC content of the 20 sequence files resulted in 13 files falling in the warning category while 7 failed the test. All the sequences formed a normal distribution with the peak of the curve at the mean GC content for *Homo sapiens*, with a deviation, causing FastQC to fail for the seven files at this step. Four samples passed the per base N content test, while a warning was raised for the remaining. It was further noticed that all the samples cleared the duplication levels, with most sequences falling into the far left of the plot. All 20 samples cleared the adapter content test, requiring no further trimming of the specific adapters (Fig. **2**).

Further, typically, to take forward the sequences for further analysis, basic statistics, per base sequence, per base quality, GC content and N content are important. To add to this, MultiQC runs duplicates of the sequences to make sure the results are accurately predicted. In our study, most of the sequences passed these quality checks. The red regions observed in the last column of Fig. (**2**) indicate the kmer content, which is generally not as important as the rest. Since an overall result showed that the sequences were of good quality, they were analyzed further. The raw MultiQC HTML files are provided as supplementary to corroborate our outcomes and provide more clarity on the same. These can be accessed in Supplementary file **S1**.

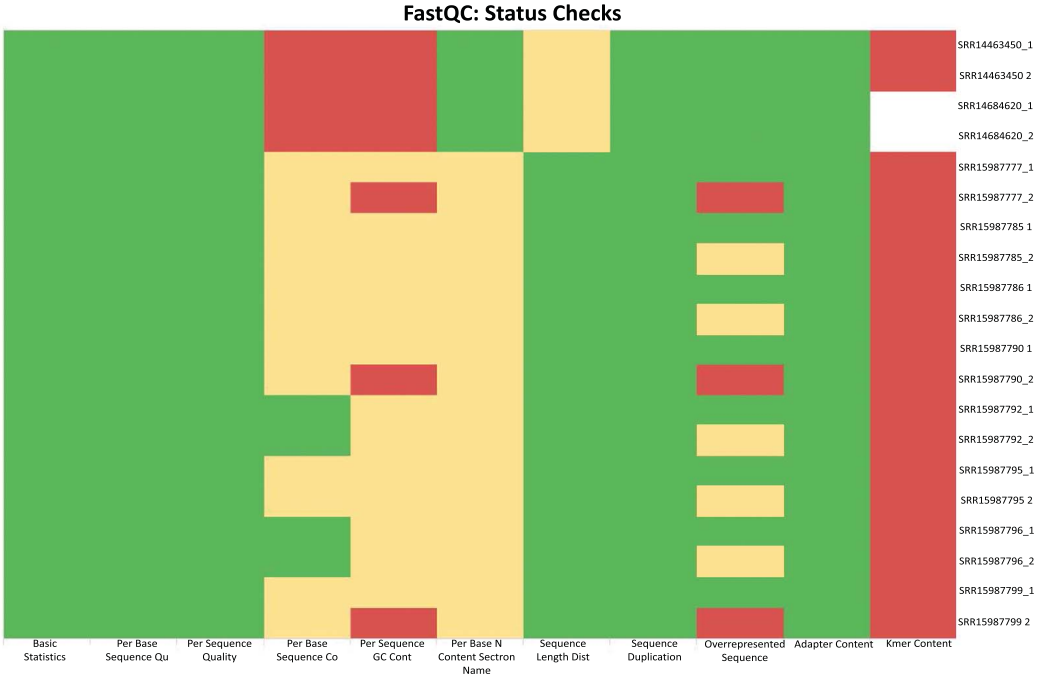


Fig. (2). MultiQC status checks summarizing all FastQC results for 10 colorectal cancer exomes. Green represents very good quality calls, orange represents reasonable quality and red indicates poor quality and failure of the test. All 20 sequence files (duplicates of all 10 ten exomes) passed the per sequence quality scores check. The per sequence GC content of the 20 sequence files resulted in 13 files falling in the warning category while 7 failed the test. All the sequences formed a normal distribution with the peak of the curve at the mean GC content for *Homo sapiens*, with a deviation, causing FastQC to fail for the seven files at this step. Four samples passed the per base N content test, while a warning was raised for the remaining. Moreover, all the samples cleared the duplication levels, with most sequences falling into the far left of the plot. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

3.2. Variant Calling

The VCF files generated for each of the exomes using the GATK pipeline revealed the total number of SNPs and indels, as observed in the bar graph (Fig. **3A**). Each dataset was subjected to variant filtration, followed by which the final number of SNPs and indels were recorded. Dataset SRR15987799 contained the highest number of SNPs (188,833) and indels (27,331). Overall, from all 10 colorectal cancer exomes, 1,220,885 SNPs were identified, 169,477 indels, all totaling up to 1,390,362 SNPs. Additionally, information regarding the number of variants processed, variant rate details, number of effects by type, region and functional class (missense, nonsense and silent mutations), and the Ts/Tv (transitions/ transversions) ratio were also obtained, represented in the bar graph (Fig. **3B**). SnpEff annotation revealed a total of 1,376,158 transitions, 619,870 transversions, 255,387 missense variations, 2083 nonsense mutations and 319,188 silent mutations. From this, dataset SRR15987799 had a maximum number of transitions (221,059), and transversions (108,426). From both these results, it is understood that dataset SRR15987799 mutated more than the rest, warranting further analysis into it for scouting possible potential genes that may get overexpressed.

Based on the functional classes, 44.3% of the variants were missense mutations, 55.4% were silent, and only 0.3% were found to be nonsense (Fig. **3C**). With this variation data, finding the frequently occurring mutations with implications for colorectal cancer was much simpler. More details are provided in Supplementary file **S2**. SIFT annotations summary 1 and 2, PolyPhen2 annotations summary, SNPs and Indels called during GATK processing, SnpEff annotations summary.

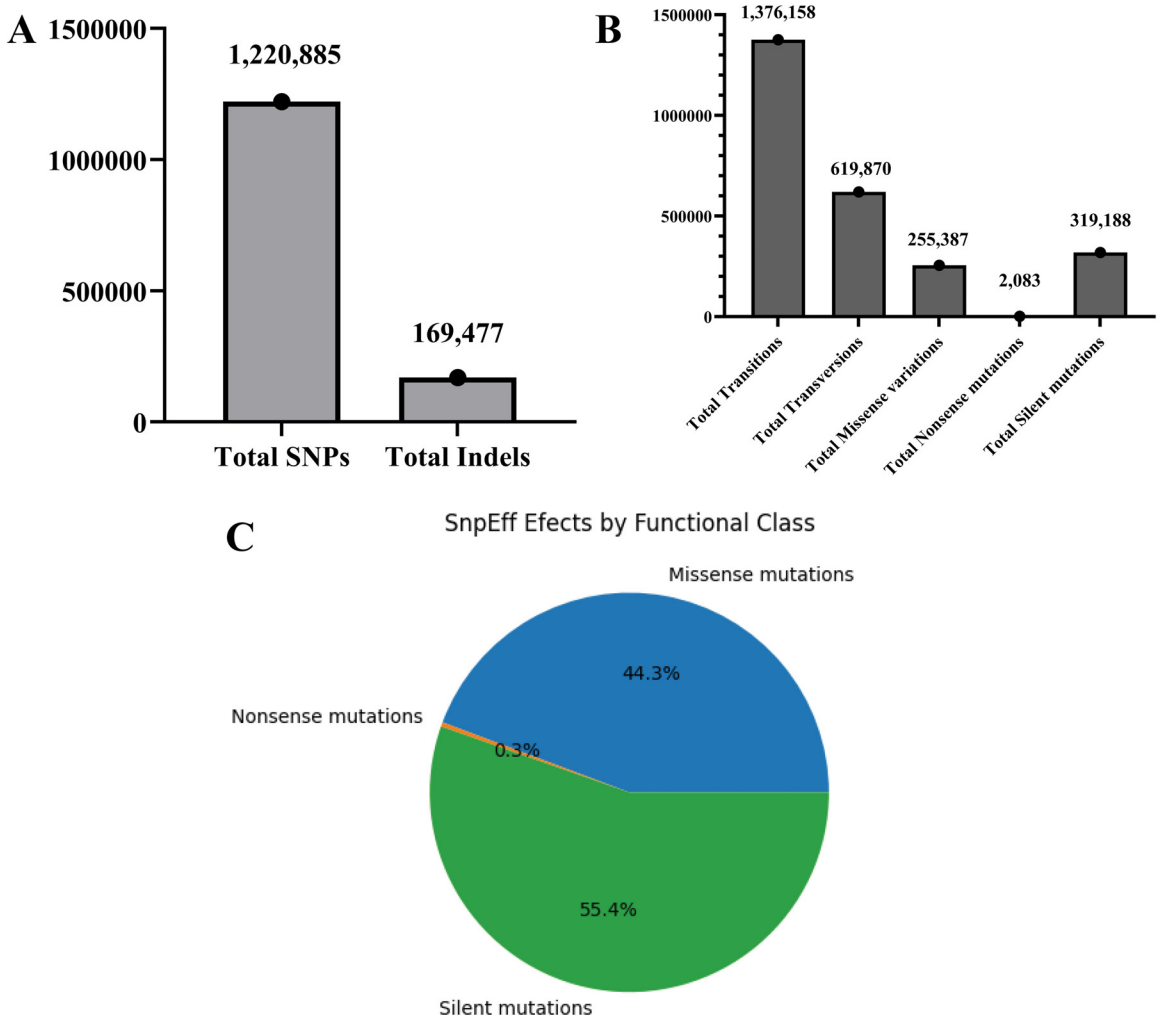


Fig. (3). Variant calling data showing the GATK processing and annotation *via* snpEFF. **3A**) the total number of SNPs and indels that were obtained after variant calling *via* the GATK pipeline. Totally, 1,220,885 SNPs were detected, with 169,477 indels. **3B**) Annotated variants showing the total number of transitions (1,376,158), transversions (619,870), missense variations (255,387), nonsense mutations (2083) and silent mutations (319,188). **3C**) Based on the functional classes, 44.3% of the variants called were missense mutations, 55.4% were silent, and only 0.3% were found to be nonsense mutations. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

3.3. Variant Post-processing and Mutational Profile Analysis

The SIFT annotation summary revealed a total of 515,762 tolerated variants, 30,980 deleterious variants and 10,787 deleterious low-confidence mutations (Fig. **4A**). The tool predicts the maximum number of mutations in the conserved regions and classifies it as deleterious or tolerated. Thus, dataset SRR15987790 had a maximum number of tolerated, deleterious and deleterious low-confidence mutations, different from the dataset that previously showed a higher number of SNPs and indels. This implied that annotation of the identified variants plays a major role in the analysis of mutational profiles. Literature implies that the amino acid substitutions that are deleterious as per the SIFT algorithm point towards the phenotype that is affected [33]. Thus, the results obtained through this analysis can be utilized for the identification of various plausible disease-causing genes. Furthermore, 407,952 non-coding variants were observed, along with 317,623 synonymous, 254,675 non-synonymous, 595 start lost, 2105 stop gain and 512 stop loss alterations (Fig. **4B**). Rare non-synonymous mutations in several genes are known to be accountable for colorectal tumors as per certain epidemiologic studies [34]. In the present study, most non-synonymous mutations were identified in SRR15987790 (38,093). Thus, further analysis of the non-synonymous variants identified can help understand the major genes that can predispose to colorectal cancer.

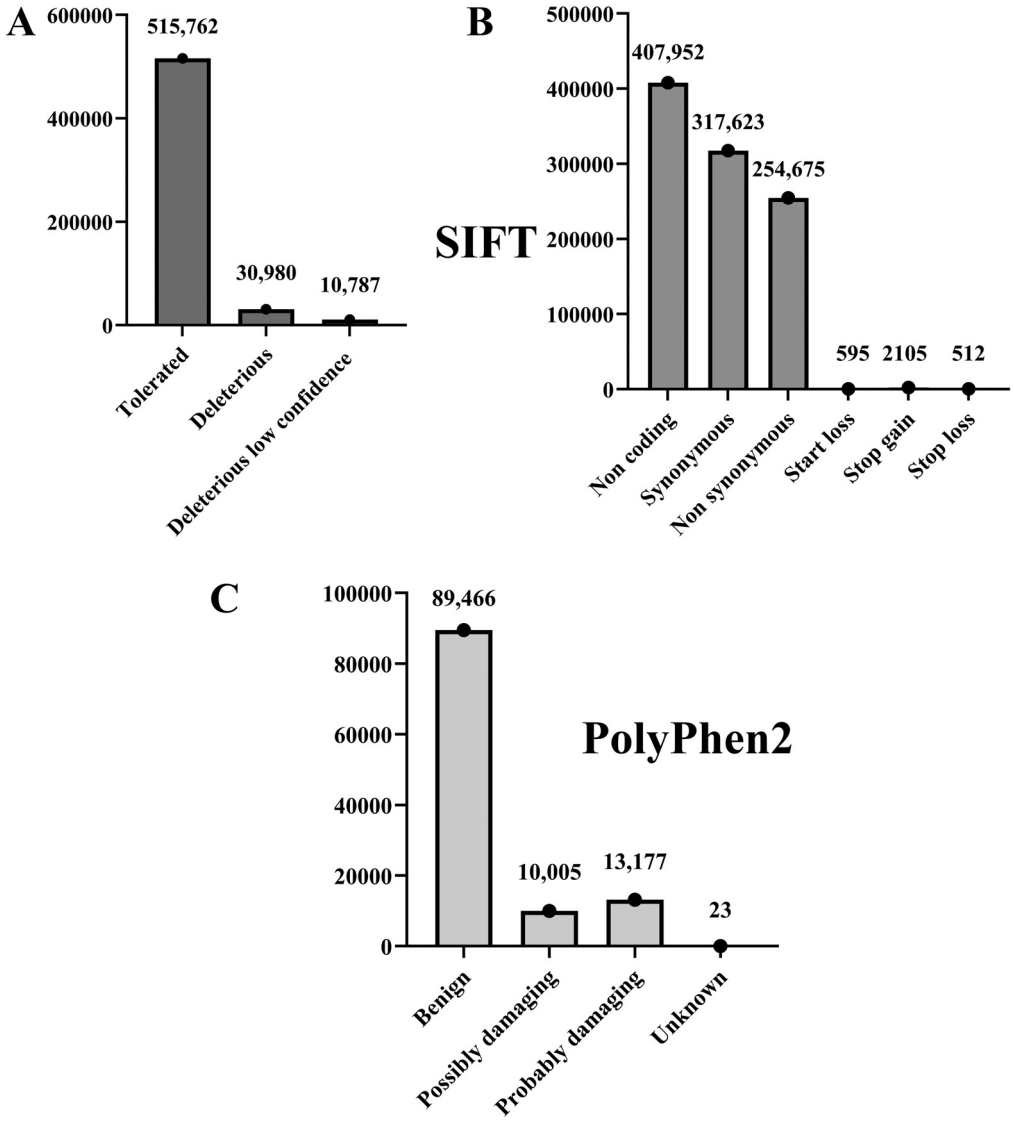


Fig. (4). Variant post-processing and mutational profile analysis. **4A**) The SIFT annotation summary revealed a total of 515,762 tolerated variants, 30,980 deleterious variants and 10,787 deleterious low-confidence mutations. **4B**) 407,952 non-coding variants were observed, along with 317,623 synonymous, 254,675 non-synonymous, 595 starts lost, 2105 stop gain and 512 stop loss alterations. **4C**) A total of 89,466 variants were benign, 10,005 were possibly damaging, and 13,177 were found to be probably damaging, according to PolyPhen 2.

Likewise, to cross-verify and build the mutational profile further, summary files generated from PolyPhen2 were also analyzed. A total of 89,466 variants were benign, 10,005 were possibly damaging, and 13,177 were found to be probably damaging (Fig. **4C**). Moreover, 1901 possible damaging mutations and 2461 probably damaging alterations were noted in SRR15987795, most among all other exomes. Further comprehension into these might reveal the possible and probable genes that could be involved in colorectal tumors. This mutational profile sheds light on the types of mutations and exomes that must be looked at in much more depth and can help in unearthing rare or commonly overexpressed genes for better cancer prognosis. A more detailed analysis for each exome sample is provided in Supplementary file **S2**. SIFT annotations summary 1 and 2, PolyPhen2 annotations summary, SNPs and Indels called during GATK processing, SnpEff annotations summary.

3.4. Analysis of Filtered Mutations and Gene Expression Profiling

The major SNPs and their associated genes with a frequency greater than 10 across all the exomes analyzed included rs1041935 (*LGALS8*), rs16604022 (*PDE4DIP*), rs12338 (*CTSB*), rs1045051 (*RAD17*), rs2071307 (*ELN*), rs11543244 (*CPNE1*), rs1799971 (*OPRM1*), rs11526468 (*SEMA4D*), rs729593 (*MUC4*), rs1061308 (*PDE4DIP*), rs17855988 (*ELN*) and rs2229125 (*ADRA1A*). It was observed that these mutations, as per SIFT and PolyPhen2 analysis, were categorized into non-synonymous and deleterious variants. Furthermore, the variant rs1664022 in *PDE4DIP* was identified in all the exomes analyzed in this study. Table **2** shows the percentage frequency of occurrence of each of the shortlisted variants, their gene names and functions, gene symbols, UniProt IDs, dbSNP IDs and HGNC IDs.

Table 2. Mutations and their associated genes and their functions, HGNC IDs, UniProt IDs, and dbSNP IDs with the highest frequency of occurrence across the ten colorectal cancer exomes analyzed.

Gene Symbol	HGNC ID	Gene name	UniProt ID	Function	dbSNP ID	Frequency of Occurrence
LGALS8	6569	Galectin 8/lectin, galactoside-binding, soluble, 8	O00214	A lectin with beta-galactoside-binding properties serves as a detector of membrane damage induced by infections. It hinders the growth of invading pathogens by directing them toward autophagy (Thurston *et al.*, 2012; Staring *et al.*, 2017) [35, 36].	rs1041935	80%
PDE4DIP	15580	Phosphodiesterase 4D interacting protein	Q5VU43	Serves as a tether, capturing elements of the cAMP-dependent pathway and localizing them to the Golgi and/or centrosomes (Mani *et al*., 2022) [37].	rs16604022, rs1061308	100%, 20%
CTSB	2527	Cathepsin B	P07858	A thiol protease is thought to play a role in the intracellular breakdown and renewal of proteins (Guo *et al.*, 2002) [38].	rs12388	60%
RAD17	9807	*RAD17* checkpoint clamp loader component/*RAD17* homolog (*S. pombe*)	O75943	Crucial for continual cell growth, preservation of chromosomal stability, and the activation of ATR-dependent checkpoints in response to DNA damage. Exhibits a modest ATPase activity necessary for chromatin binding. Plays a role in recruiting the RAD1-RAD9-HUS1 complex and RHNO1 to chromatin, and contributes to the activation of CHEK1. Additionally, it may function as a detector of DNA replication progression and be implicated in homologous recombination (Li *et al.*, 1999) [39].	rs1045051	40%
ELN	3327	Elastin	P15502	Primary structural protein is found in tissues like the aorta and nuchal ligament, where rapid expansion and complete recovery are essential. Functions as a molecular factor in the final stages of arterial morphogenesis, contributing to the stabilization of arterial structure by modulating the proliferation and organization of vascular smooth muscle (Keeley *et al.*, 2002) [40].	rs2071307, rs17855988	50%, 20%
CPNE1	2314	Copine 1	Q99829	A phospholipid-binding protein activated by calcium is involved in regulating intracellular processes mediated by calcium (Tomsig *et al.*, 2004) [41].	rs11543244	30%
OPRM1	8156	Opioid receptor mu 1	P35372	A receptor responsive to endogenous opioids like beta-endorphin and endomorphin (Pan *et al.,* 2003) [42].	rs1799971	50%
SEMA4D	10732	Semaphorin 4D	Q92854	A receptor located on the cell surface for PLXNB1 and PLXNB2 plays a crucial role in mediating cell-cell signaling (Janssen *et al.,* 2010) [43].	rs11526468	30%
MUC4	7514	Mucin 4, cell surface associated	Q99102	A mucin anchored to the membrane, belonging to a family of extensively glycosylated proteins that form the predominant constituent of mucus. Mucus is the slippery and thick secretion that covers epithelial surfaces (Moniaux *et al.,* 2000) [44].	rs729593	40%
ADRA1A	277	Adrenoceptor alpha 1A	P35348	This alpha-adrenergic receptor exerts its effects through interaction with G proteins, which in turn activate a phosphatidylinositol--calcium second messenger system. G(q) and G11 proteins mediate its effects. Nuclear ADRA1A-ADRA1B heterooligomers play a regulatory role in phenylephrine (PE)-stimulated ERK signaling in cardiac myocytes (Wright *et al.,* 2008) [45].	rs2229125	20%

3.4.1. Gene Expression Profiling Using GEPIA

The box plots obtained for these top ten shortlisted cancer genes associated with their SNPs showed that out of 10 genes, 7 genes were found to have slightly more expression in tumor samples than in normal tissues, with grey boxes representing expression of the genes in normal tissues and red in tumor samples (Fig. **5**). Additionally, the overall survival plots demonstrated that with high expression of specific genes, the survival rates of patients begin to decrease (given in terms of months and percentage survival). Gene *ADRA1* was found to have a higher expression in normal tissues than in tumor samples. The overall survival plot also showed not much difference in survival when expressed in high or low quantities (Fig. **5A**). However, mutations in genes *CPNE1*, *CTSB* and *ELN* (Figs **5B-D**), indicated that their expression was slightly higher in tumor samples than in normal ones. Their survival plots also pointed out that higher expression of these genes may also reduce the percentage of survival, especially for mutations in gene *CTSB*. *LGALS8* was found to be expressed almost equally in both tumor and normal tissues (Fig. **5E**). Similar to *ADRA1*, gene *MUC4* was predicted to be expressed more in normal samples than in tumors, with survival predictions being less in patients with low rates of *MUC4* expression (Fig. **5F**). Genes *PDE4DIP*, *RAD17*, and *SEMA4D* had slightly more expression in tumors than normal tissues, with patient survival rates decreasing with increased expression of these genes with the mutations (Figs. **5G-I**). *OPRM1* was found not to be expressed in tumor samples but only slightly in normal ones (Fig. **5J**). No survival analysis was possible for this gene as expression was nil in tumor samples.

Since the expression levels were studied in two exomes-colon adenocarcinoma and rectum adenocarcinoma, a differential gene expression of these ten genes against the two cancer types showed that the mutated *CTSB* gene had the highest expression in both cancer types, followed by *CPNE1*. Genes *LGALS8*, *PDE4DIP*, *RAD17*, *ELN*, *MUC4* and *SEMA4D* had a mid-range expression in both cancer types, while *OPRM1* and *ADRA1* were the least expressed, thereby corroborating the results from our box plot and survival analysis (Fig. **6**). These results indicate that two major genes-*CTSB* and *CPNE1* may point towards causing colorectal cancers.

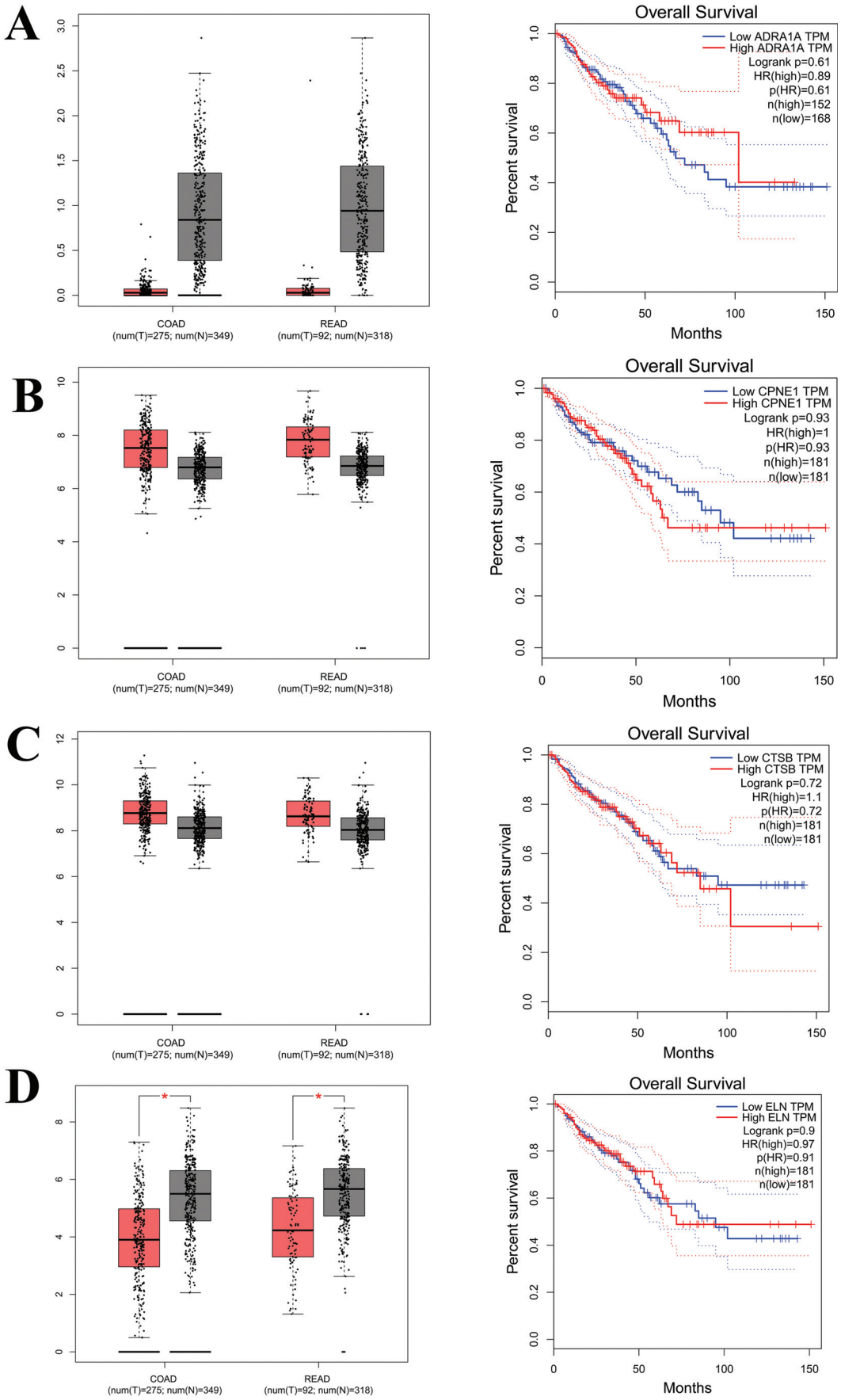
3.4.2. Differential Gene Expression Against Various Cancers

To understand the possibility of these SNP-associated genes being over or under-expressed in not just colorectal cancer but in other types, a differential gene expression against various cancers revealed that *ADRA1* was completely down-regulated in colorectal cancer and in several other cancer types as well (Fig. **7A**), rendering this gene less likely to be involved in cancers. Corroborating the result we obtained from GEPIA gene expression studies, genes *CTSB* and *CPNE1* are highly overexpressed in colorectal cancer and most of the other cancer types in the OncoMX database, such as head and neck cancer, liver cancer, stomach cancer, kidney cancer, esophageal, lung and thyroid (Figs. **7B** and **C**). *ELN* was predicted to have overexpression in colorectal cancer but was under-expressed in the majority of the other types, while *LGALS8* was found to be downregulated in colorectal cancer and overexpressed in all other types (Figs. **7D** and **E**). Likewise, *MUC4* expression was down-regulated in all cancer types except for thyroid and lung cancer, indicating that it does not play a major role in causing colorectal cancer (Fig. **7F**). Surprisingly, *OPRM1* was downregulated in colorectal and most other cancers, but completely expressed in uterine cancers (Fig. **7G**). *PDE4DIP* and *RAD17* both do not play a major role in causing colorectal cancer, as observed and corroborated by gene expression profiling studies. However, it still may be upregulated in thyroid and uterine cancer (*PDE4DIP*) and lung cancer (*RAD17*) (Figs. **7H** and **I**). Contrary to the predictions obtained from GEPIA, *SEMA4D* was found to be upregulated highly in colorectal cancer and in several other types (Fig. **7J**). These results show that the mutations identified in these genes may play a role in colorectal cancer and, if not, point towards their presence in implicating other cancer types, thereby allowing researchers and clinicians to focus more on these genes that predispose to specific cancer types.

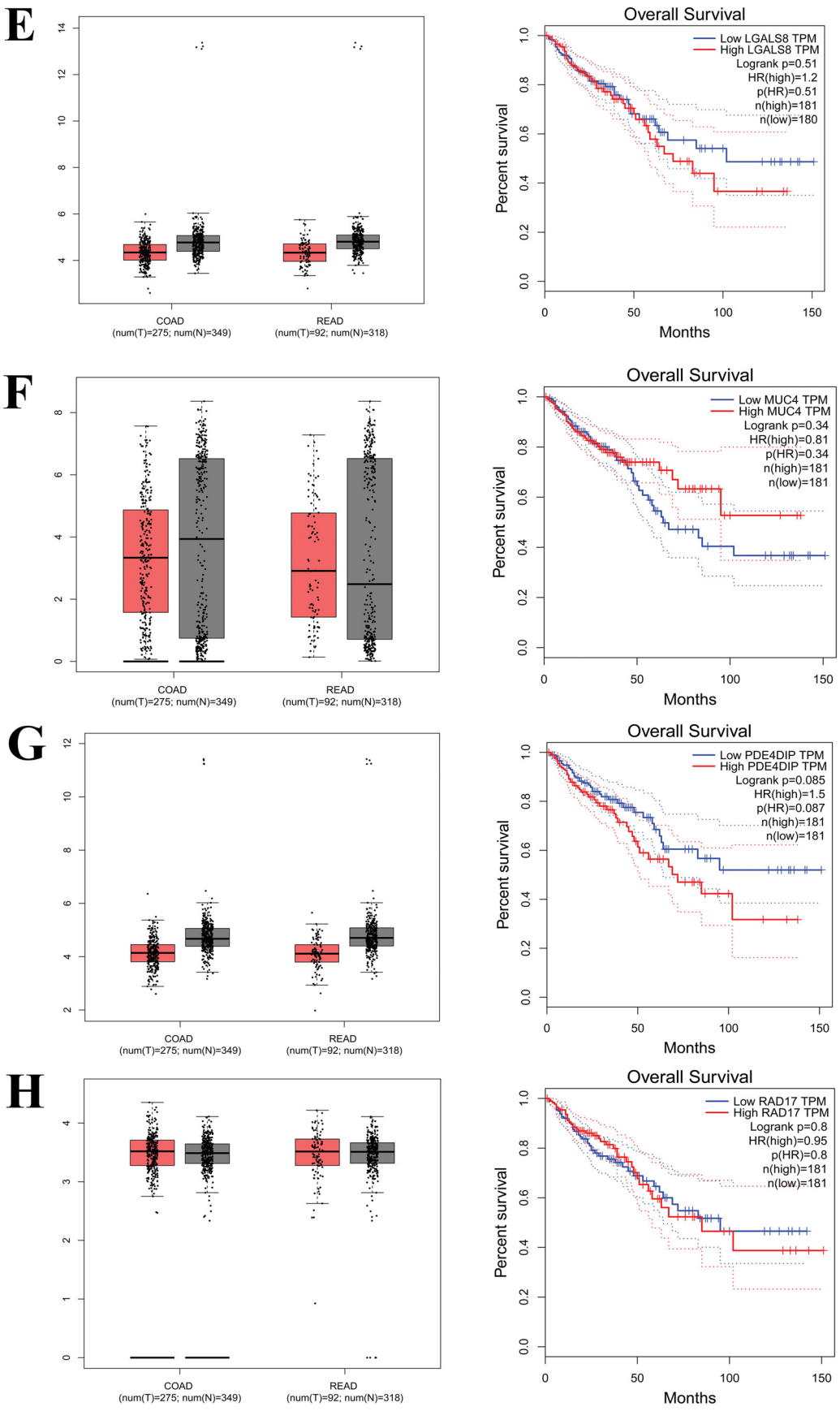
3.4.3. Multi-dimensional Gene Analysis

When all 4535 colorectal cancer samples were run against our ten genes associated with mutations, plots of mRNA expression (RNA seq data) v/s the mutations in each gene were obtained. Fig. (**8**) illustrates in detail the different types of available and possible mutations in each gene against the available colorectal patient database. A few missense mutations were predicted for *ADRA1*, *CPNE1*, *CTSB*, *ELN*, *LGALS8*, *OPRM1*, *RAD17*, *PDE4DIP*, and *SEMA4D*. The maximum number of missense variations were noted in the *MUC4* gene, and most of these genes had shallow and deep deletions. There were very few amplifications in *CTSB*; however, a large number of deep and shallow deletions, as in *PDE4DIP*. *CPNE1* showed a very high number of amplifications. From this analysis, it is understood that two genes, *CTSB* and *CPNE1,* may tend to be more active in colorectal cancer cases than any other, implying further analysis into its use as a cancer indicator.

Additionally, a summary of all alterations per sample available in cBioPortal has been portrayed in Fig. (**9**), with different genetic alterations highlighted in various colors. All the samples were sorted by gene and type of the genetic event detected. Each query gene was represented as a row, and the samples as columns. The study of origin provided the list of all 13 colorectal cancer exomes. It was noted that in *LGALS8*, 1.1% of the samples that the gene was run against were altered, 5% of samples in *PDE4DIP*, and *CPNE1*, 3% in *CTSB*, *MUC4*, and *ADRA1*, 1.7% in *RAD17*, 1.5% in *ELN*, 1.2% in *OPRM1* and 1.9% in *SEMA4D*. These results corroborate the mRNA v/s mutation plots understood previously in Fig. (**8**), with most of the *PDE4DIP* mutations being missense, deep deletions in *CTSB* and amplifications in *CPNE1*. This multi-dimensional analysis provides a different perspective and validation of the major genes that are essential in colorectal cancer detection.


A
COAD
(num(T)=275; num(N)=349)
READ
(num(T)=92; num(N)=318)
Overall Survival
Low ADRA1A TPM
High ADRA1A TPM
Logrank p=0.61
HR(high)=0.89
p(HR)=0.61
n(high)=152
n(low)=168
Percent survival
Months
B
COAD
(num(T)=275; num(N)=349)
READ
(num(T)=92; num(N)=318)
Overall Survival
Low CPNE1 TPM
High CPNE1 TPM
Logrank p=0.93
HR(high)=1
p(HR)=0.93
n(high)=181
n(low)=181
Percent survival
Months
C
COAD
(num(T)=275; num(N)=349)
READ
(num(T)=92; num(N)=318)
Overall Survival
Low CTSB TPM
High CTSB TPM
Logrank p=0.72
HR(high)=1.1
p(HR)=0.72
n(high)=181
n(low)=181
Percent survival
Months
D
COAD
(num(T)=275; num(N)=349)
READ
(num(T)=92; num(N)=318)
Overall Survival
Low ELN TPM
High ELN TPM
Logrank p=0.9
HR(high)=0.97
p(HR)=0.91
n(high)=181
n(low)=181
Percent survival
Months

(Fig. 5). contd….


E
COAD
(num(T)=275; num(N)=349)
READ
(num(T)=92; num(N)=318)
Overall Survival
Low LGALS8 TPM
High LGALS8 TPM
Logrank p=0.51
HR(high)=1.2
p(HR)=0.51
n(high)=181
n(low)=180
Percent survival
Months
F
COAD
(num(T)=275; num(N)=349)
READ
(num(T)=92; num(N)=318)
Overall Survival
Low MUC4 TPM
High MUC4 TPM
Logrank p=0.34
HR(high)=0.81
p(HR)=0.34
n(high)=181
n(low)=181
Percent survival
Months
G
COAD
(num(T)=275; num(N)=349)
READ
(num(T)=92; num(N)=318)
Overall Survival
Low PDE4DIP TPM
High PDE4DIP TPM
Logrank p=0.085
HR(high)=1.5
p(HR)=0.087
n(high)=181
n(low)=181
Percent survival
Months
H
COAD
(num(T)=275; num(N)=349)
READ
(num(T)=92; num(N)=318)
Overall Survival
Low RAD17 TPM
High RAD17 TPM
Logrank p=0.8
HR(high)=0.95
p(HR)=0.8
n(high)=181
n(low)=181
Percent survival
Months

(Fig. 5). contd….

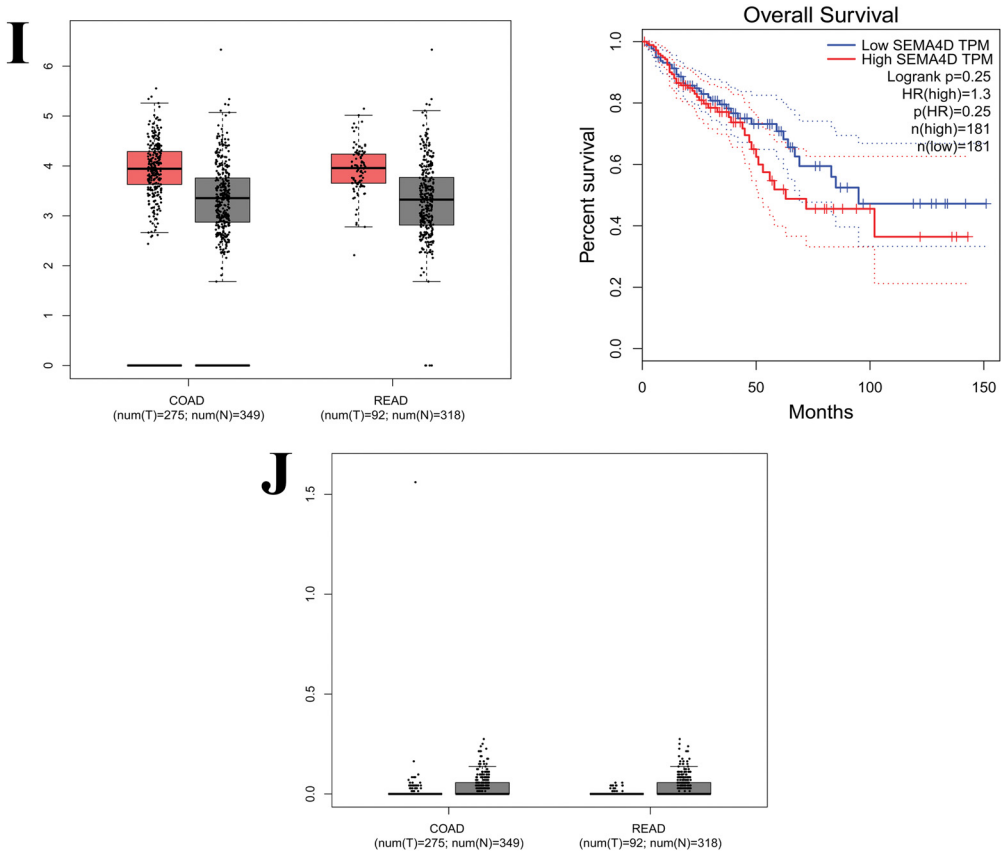


Fig. (5). Box plots and predicted survival rates for the top ten shortlisted cancer genes with highly occurring mutations. The grey boxes represent the expression of the genes in normal tissues, and the red in tumor samples. The overall survival plots demonstrate that with high expression of specific genes, survival rates of patients begin to decrease. The survival rates are represented as months v/s percentage survival. **5A**) The expression and overall survival plot for *ADRA1*. **5B**) The expression and overall survival rate for *CPNE1*. **5C**) The expression and overall survival plot for *CTSB*. **5D**) The expression and survival rate for *ELN*. **5E**) The expression and overall survival study for *LGALS8*. **5F**) The expression and overall survival plot for *MUC4*. **5G**) The expression and overall survival rate for *PDE4DIP*. **5H**) The expression and overall survival plot for *RAD17*. **5I**) The expression and survival rate for *SEMA4D*. **5J**) The expression and overall survival study for *OPRM1*. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

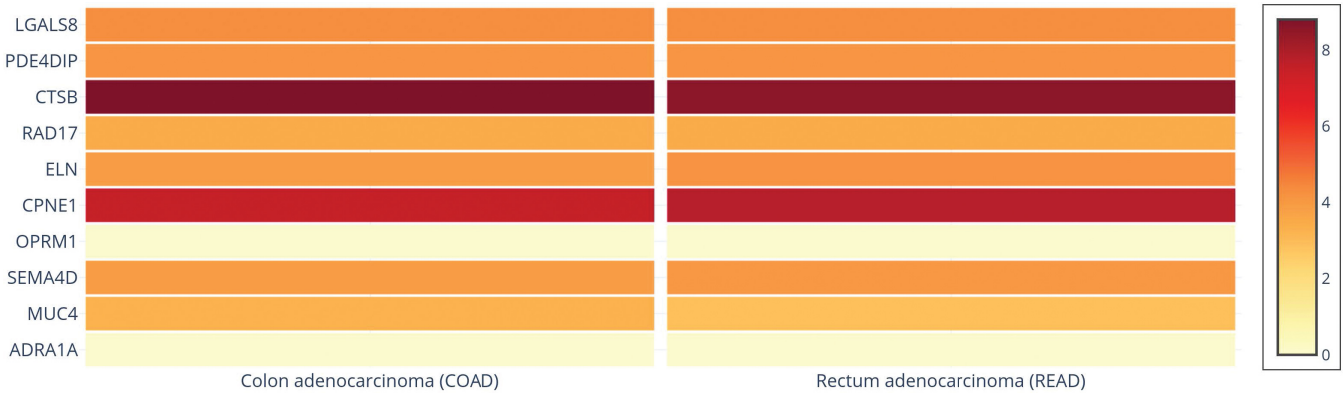
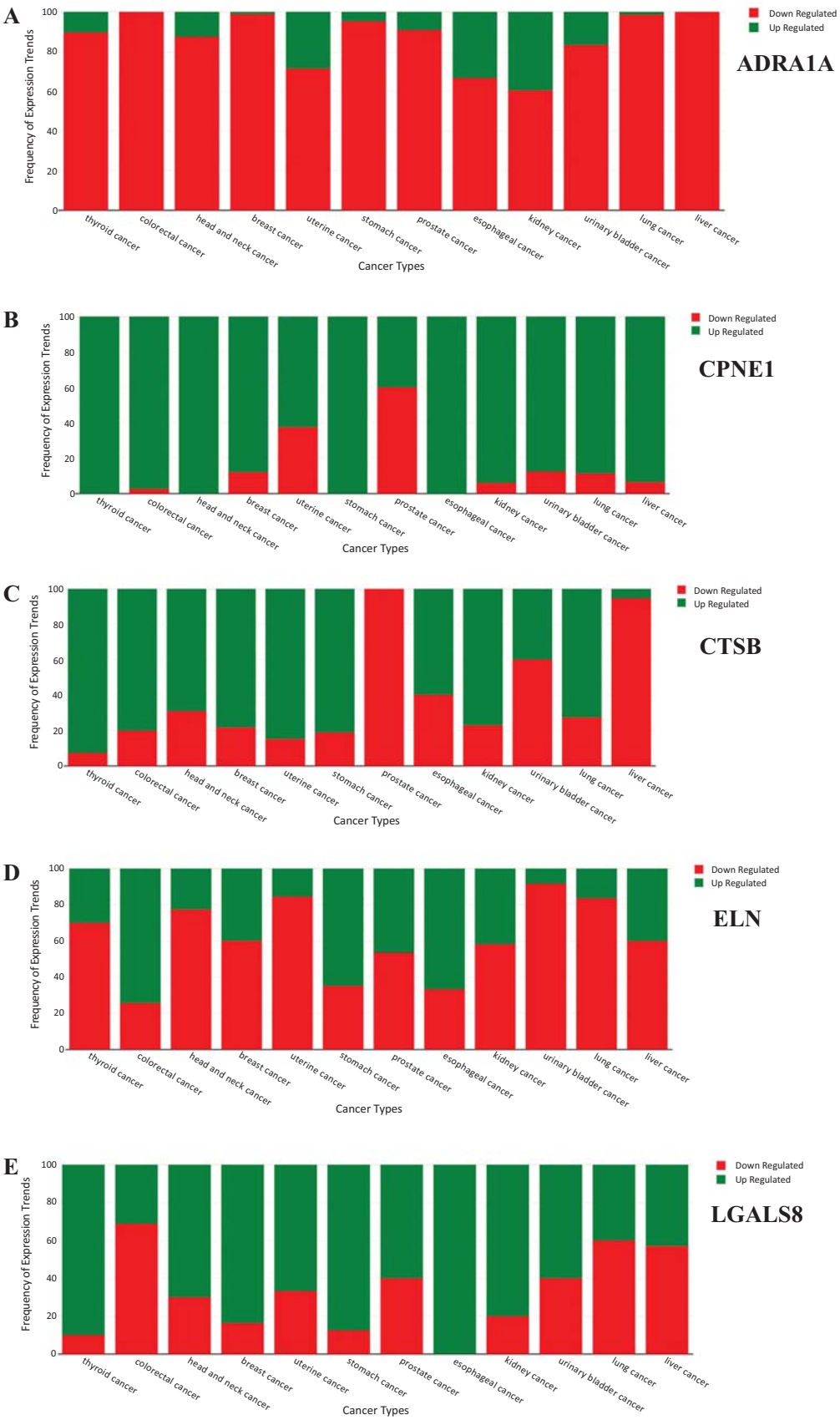


Fig. (6). Differential gene expression of top ten identified genes against colon adenocarcinoma and rectum adenocarcinoma using GEPIA server. The mutated *CTSB* gene had the highest expression in both cancer types, followed by *CPNE1*. Genes *LGALS8*, *PDE4DIP*, *RAD17*, *ELN*, *MUC4* and *SEMA4D* had a mid-range expression in both cancer types, while *OPRM1* and *ADRA1* were the least expressed, thereby corroborating the results from our box plot and survival analysis. These results indicate that two major genes- *CTSB* and *CPNE1* may point towards causing colorectal cancers. Light brown indicates lower expression levels and dark brown to higher. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).


A
Down Regulated
Up Regulated
ADRA1A
Frequency of Expression Trends
0
20
40
60
80
100
thyroid cancer
colorectal cancer
head and neck cancer
breast cancer
uterine cancer
stomach cancer
prostate cancer
esophageal cancer
kidney cancer
urinary bladder cancer
lung cancer
liver cancer
Cancer Types
B
Down Regulated
Up Regulated
CPNE1
Frequency of Expression Trends
Cancer Types
C
Down Regulated
Up Regulated
CTSB
Frequency of Expression Trends
Cancer Types
D
Down Regulated
Up Regulated
ELN
Frequency of Expression Trends
Cancer Types
E
Down Regulated
Up Regulated
LGALS8
Frequency of Expression Trends
Cancer Types

(Fig. 7). contd….

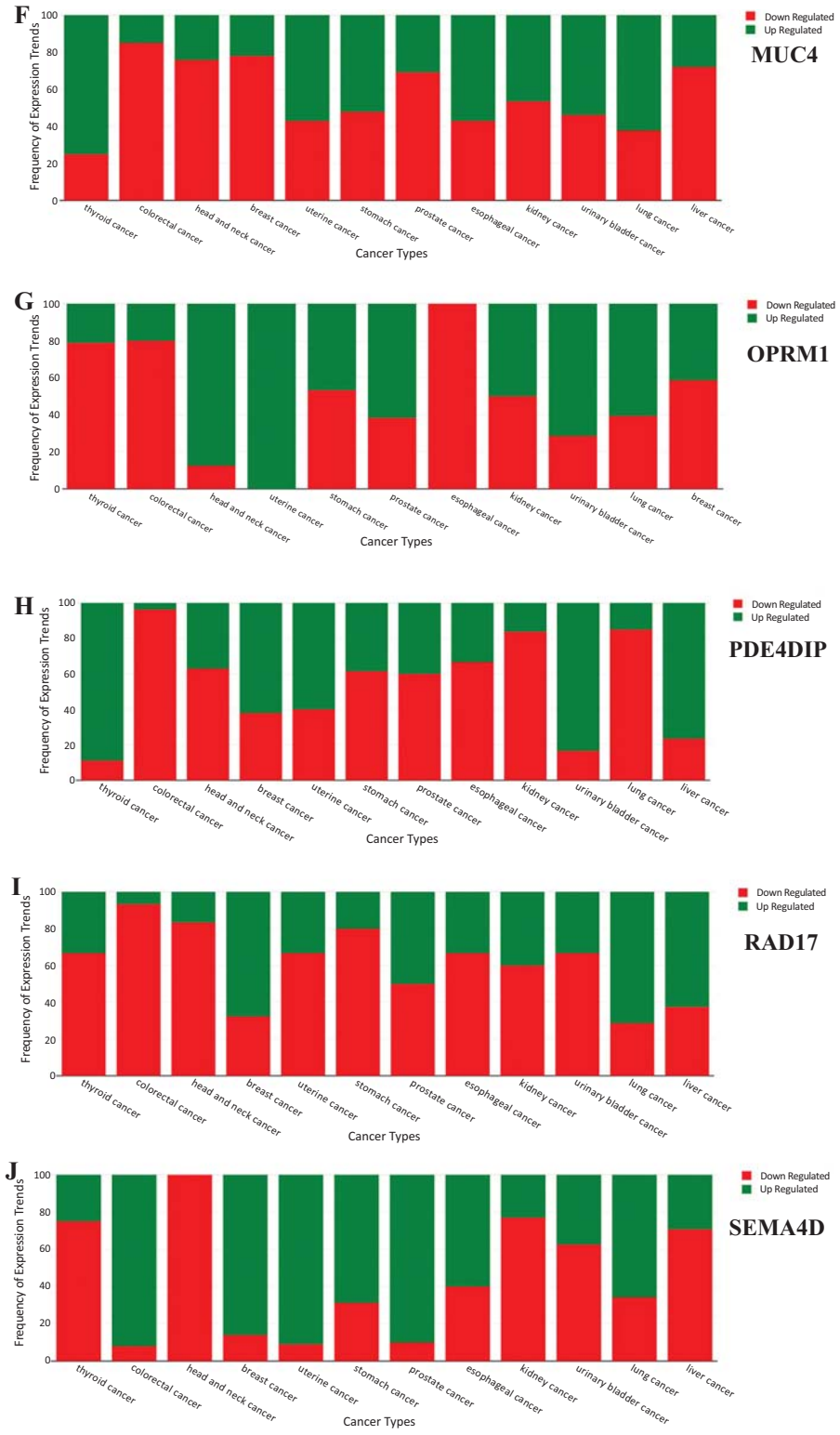
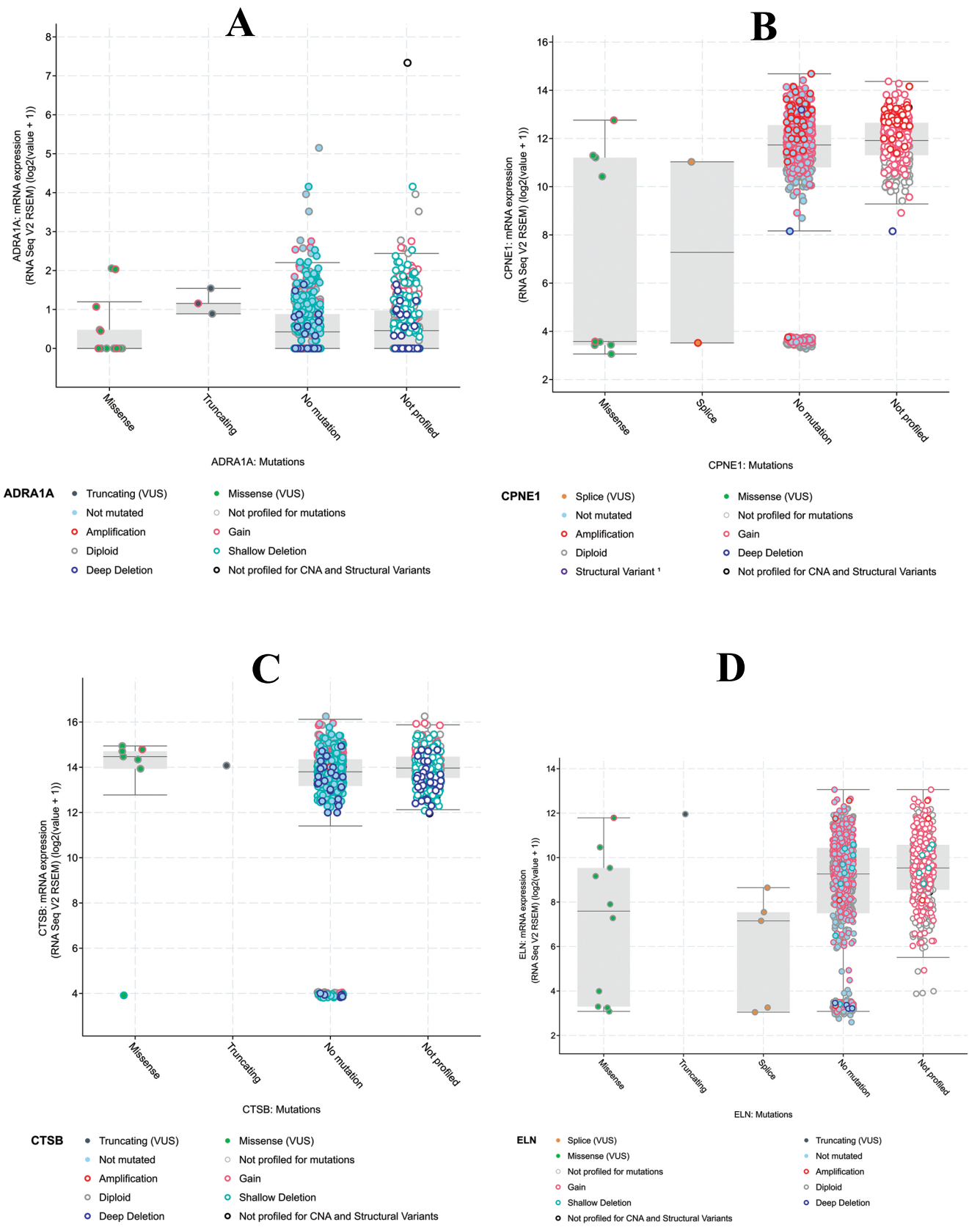
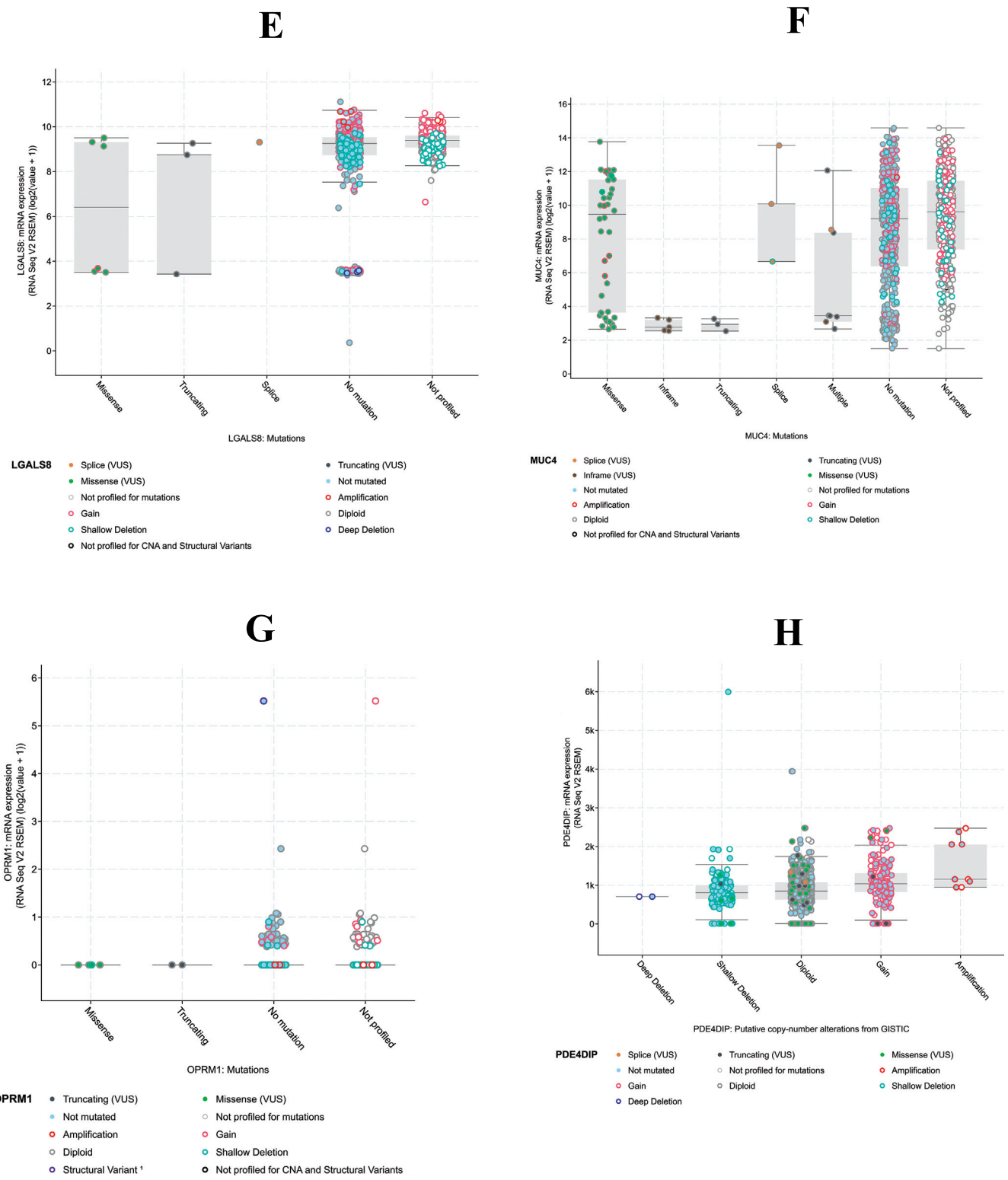


Fig. (7). Differential gene expression of top ten genes with respect to various cancer types. This plot shows that the mutations identified in these genes may play a role in colorectal cancer and, if not, point towards their presence in implicating other cancer types, thereby allowing researchers and clinicians to focus more on these genes that predispose to specific cancer types. **7A**) Expression of *ADRA1*. **7B**) Expression of *CPNE1*. **7C**) Expression of *CTSB*. **7D**) Levels of *ELN* expression. **7E**) *LGALS8* expression levels. **7F**) Expression of*MUC4*. **7G**) Expression of *OPRM1*. **7H**) Expression of *PDE4DIP*. **7I**) Levels of *RAD17* expression. **7J**) *SEMA4D* expression levels. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).


A
ADRA1A: mRNA expression (RNA Seq V2 RSEM) (log2(value + 1))
Missense
Truncating
No mutation
Not profiled
ADRA1A: Mutations
ADRA1A
Truncating (VUS)
Missense (VUS)
Not mutated
Not profiled for mutations
Amplification
Gain
Diploid
Shallow Deletion
Deep Deletion
Not profiled for CNA and Structural Variants
B
CPNE1: mRNA expression (RNA Seq V2 RSEM) (log2(value + 1))
Missense
Splice
No mutation
Not profiled
CPNE1: Mutations
CPNE1
Splice (VUS)
Missense (VUS)
Not mutated
Not profiled for mutations
Amplification
Gain
Diploid
Deep Deletion
Structural Variant ¹
Not profiled for CNA and Structural Variants
C
CTSB: mRNA expression (RNA Seq V2 RSEM) (log2(value + 1))
Missense
Truncating
No mutation
Not profiled
CTSB: Mutations
CTSB
Truncating (VUS)
Missense (VUS)
Not mutated
Not profiled for mutations
Amplification
Gain
Diploid
Shallow Deletion
Deep Deletion
Not profiled for CNA and Structural Variants
D
ELN: mRNA expression (RNA Seq V2 RSEM) (log2(value + 1))
Missense
Truncating
Splice
No mutation
Not profiled
ELN: Mutations
ELN
Splice (VUS)
Truncating (VUS)
Missense (VUS)
Not mutated
Not profiled for mutations
Amplification
Gain
Diploid
Shallow Deletion
Deep Deletion
Not profiled for CNA and Structural Variants

(Fig. 8). contd….


E
LGALS8: mRNA expression (RNA Seq V2 RSEM) (log2(value + 1))
Missense
Truncating
Splice
No mutation
Not profiled
LGALS8: Mutations
LGALS8
Splice (VUS)
Missense (VUS)
Not profiled for mutations
Gain
Shallow Deletion
Not profiled for CNA and Structural Variants
Truncating (VUS)
Not mutated
Amplification
Diploid
Deep Deletion
F
MUC4: mRNA expression (RNA Seq V2 RSEM) (log2(value + 1))
Missense
Inframe
Truncating
Splice
Multiple
No mutation
Not profiled
MUC4: Mutations
MUC4
Splice (VUS)
Inframe (VUS)
Not mutated
Amplification
Diploid
Not profiled for CNA and Structural Variants
Truncating (VUS)
Missense (VUS)
Not profiled for mutations
Gain
Shallow Deletion
G
OPRM1: mRNA expression (RNA Seq V2 RSEM) (log2(value + 1))
Missense
Truncating
No mutation
Not profiled
OPRM1: Mutations
OPRM1
Truncating (VUS)
Not mutated
Amplification
Diploid
Structural Variant ¹
Missense (VUS)
Not profiled for mutations
Gain
Shallow Deletion
Not profiled for CNA and Structural Variants
H
PDE4DIP: mRNA expression (RNA Seq V2 RSEM)
Deep Deletion
Shallow Deletion
Diploid
Gain
Amplification
PDE4DIP: Putative copy-number alterations from GISTIC
PDE4DIP
Splice (VUS)
Not mutated
Gain
Deep Deletion
Truncating (VUS)
Not profiled for mutations
Diploid
Missense (VUS)
Amplification
Shallow Deletion

(Fig. 8). contd….

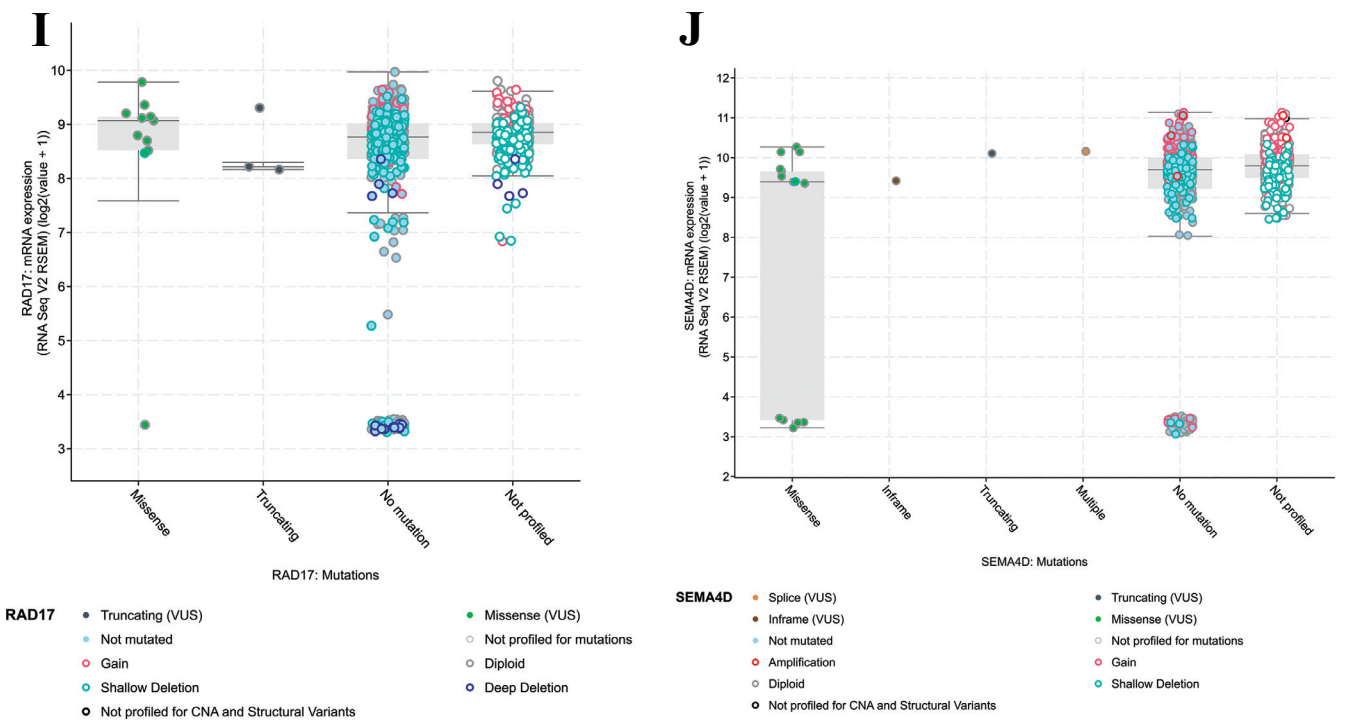


Fig. (8). Different types of available and possible mutations in each gene against 13 available colorectal patient exomes in cBioPortal, with 4535 samples. Plots of mRNA expression (RNA seq data) v/s mutations in each gene revealed few missense mutations for *ADRA1*, *CPNE1*, *CTSB*, *ELN*, *LGALS8*, *OPRM1*, *RAD17*, *PDE4DIP*, and *SEMA4D*. The maximum number of missense variations were noted in the *MUC4* gene, and most of these genes had shallow and deep deletions. There were very few amplifications in *CTSB*. However, a large number of deep and shallow deletions, as in *PDE4DIP*. *CPNE1* showed a very high number of amplifications. Two genes, *CTSB* and *CPNE1,* tend to be more active in colorectal cancer cases than any other, implying further analysis into its use as a cancer indicator. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

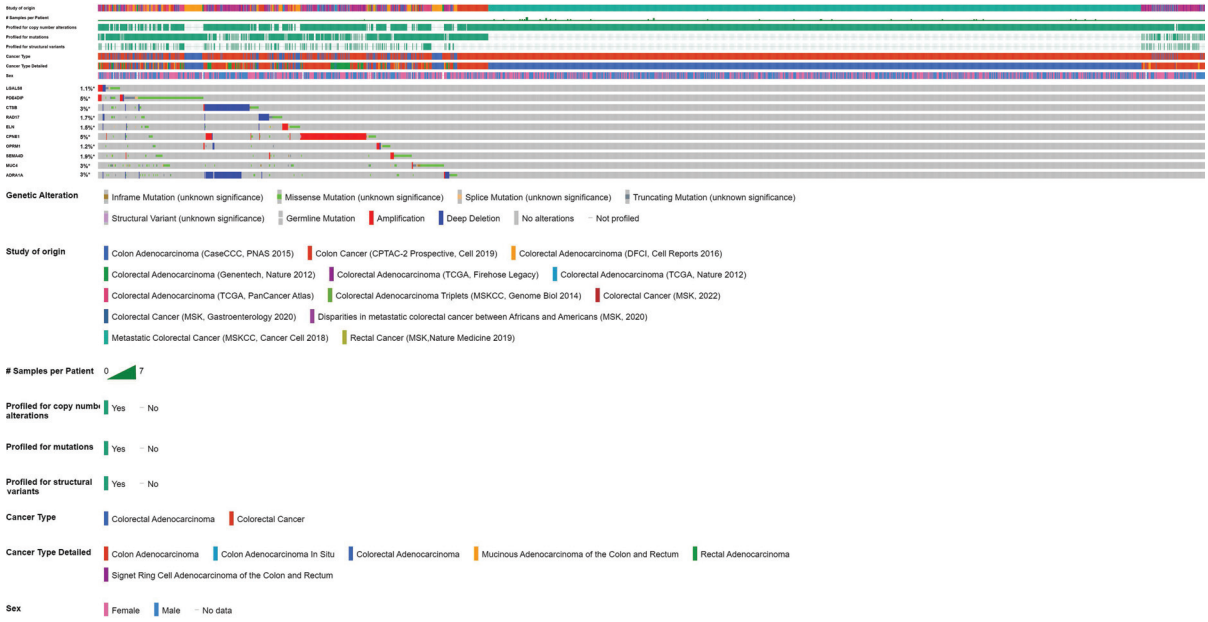


Fig. (9). Summary of all alterations per sample available in cBioPortal. Different genetic alterations are highlighted in various colors. All the samples were sorted by gene and type of the genetic event detected. Each query gene was represented as a row, and the samples as columns. The study of origin provided the list of all 13 colorectal cancer exomes. It was noted that in *LGALS8*, 1.1% of the samples that the gene was run against were altered, 5% of samples in *PDE4DIP* and *CPNE1*, 3% in *CTSB*, *MUC4*, and *ADRA1*, 1.7% in *RAD17*, 1.5% in *ELN*, 1.2% in *OPRM1* and 1.9% in *SEMA4D*. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Table 3. Multivariate analysis for correlation between gene expression data and gene frequency.

Regression Type	**Least Squares**		**-**
Dependent Variable	**Colorectal Cancer Label**		
Regression type	**Least squares**	**R^2 with other variables**	
β0	Intercept	-	-
β 1	Gene expression	0.8809	-
β 2	Gene frequency	0.8462	-
β 3	Gene expression: Gene frequency	0.9394	-
Normality of Residuals			
Normality of Residuals	**Statistics**	***P* value**	**Passed normality test (alpha=0.05)**
D'Agostino-Pearson omnibus (K2)	0.7998	0.6704	Yes
Shapiro-Wilk (W)	0.9431	0.5883	Yes

Further, Mutation data of each gene from cBioPortal is given as a different sheet in supplementary file **S3**. The study of origin, copy number, variant type, chromosome number, position of mutation, annotation, *etc*., are elucidated in detail. Supplementary file **S4** details the mutual exclusivity data obtained from cBioPortal on all 10 identified genes with the associated mutations, with its *p*-value, q-value, log2 odds ratio and tendency. The tendency of the genes to co-occur in the same samples (positive values) or occur in different samples (negative values) is also detailed.

3.5. Multi-variate Analysis Using Linear Regression

The least squares method showed that the R^2 values (the correlation between gene expression and gene frequency) were observed to be 0.8809 (for β1), 0.8462 (for β2) and 0.9394 (for β3). Although the R^2 values could have been better, it could be due to the smaller sample size since only 10 genes were shortlisted and studied for their expression and frequency. Additionally, the *p* -values for the D'Agostino-Pearson omnibus (K2) test and Shapiro-Wilk (W) test were found to be 0.6704 and 0.5883, respectively, where both passed the normality test with an alpha of 0.05.

Fig. (**10**) shows the correlation matrix for gene expression and frequency and the violin plot for the normality of residuals v/s the predicted *p* values. Table **3** highlights the statistical analysis results.

4. DISCUSSION

The present study provides a comprehensive computational perspective for obtaining important genes that predispose to colorectal cancer. The quality check results showed that the selected samples passed the tests in major criteria. Since all samples passed the adaptor content test, no adaptor trimming was needed. A recent study also examined the cancer exomes for breast cancer sequences and employed quality checks such as FastQC and MultiQC for quality assessment. Their results showed that 33 samples passed the tests out of 54 reads, with 21 falling into the 'warning' category [46]. The present study demonstrated slightly different results; however, the overall outcomes fell into the 'pass' category. Additionally, previously a study was also carried out for hepatocellular carcinoma cDNA end sequencing read, wherein adaptor clipping was carried out followed by alignment using Bowtie2 to map with the hg38 reference genome [47]. The Burrows-Wheeler Aligner (BWA) aligns all the short reads against the reference [12, 48]. Since Bowtie2 works rapidly and has better sensitivity and accuracy due to the presence of full-text minute index and dynamic programming algorithms, the present study employed this tool. Our study proved to be slightly better as none of the sequences required adaptor trimming. Moreover, the reference genome used was hg19 in the present study, and valid results were obtained from the previous works. However, the exome sequences selected were different in the present study, as was the cancer type, providing our study an edge over the previous works. Furthermore, SAMtools is considered a widely used software for the analysis of high-throughput sequenced data as it has a higher performance and enhanced ability for file indexing and sorting and writing the BAM files from SAM easily [12, 13]. Thus, our study employed this tool for SAM to BAM conversion.

A study by Xu *et al.*, 2020 [49], stated that germline genomic patterns were associated with the risk of cancers. This study employed the use of the GATK pipeline, with Haplotype caller for calling the variants for exome sequences. A comparison with Mutect2 also revealed that there were no differences in the outcome reproducibility. Additionally, another study employed low-input whole exome sequences and variants and INDELs were called using Haplotype caller from the GATK pipeline. The variants associated with cancers were then identified. In this study, both somatic and germline variants were called [26]. Another recent work employed Haplotype caller from GATK to call variants and select germline variants that were predisposed to colorectal cancer [50]. These works corroborate the work done in the present study. Moreover, our study identifies the germline mutations that also tend to have an effect on cancer risk. Studies have previously shown that germline mutations also affect tumor progression, thereby increasing their risk. Germline variations affect the expression of the genes and contribute vastly to disease progression [51].

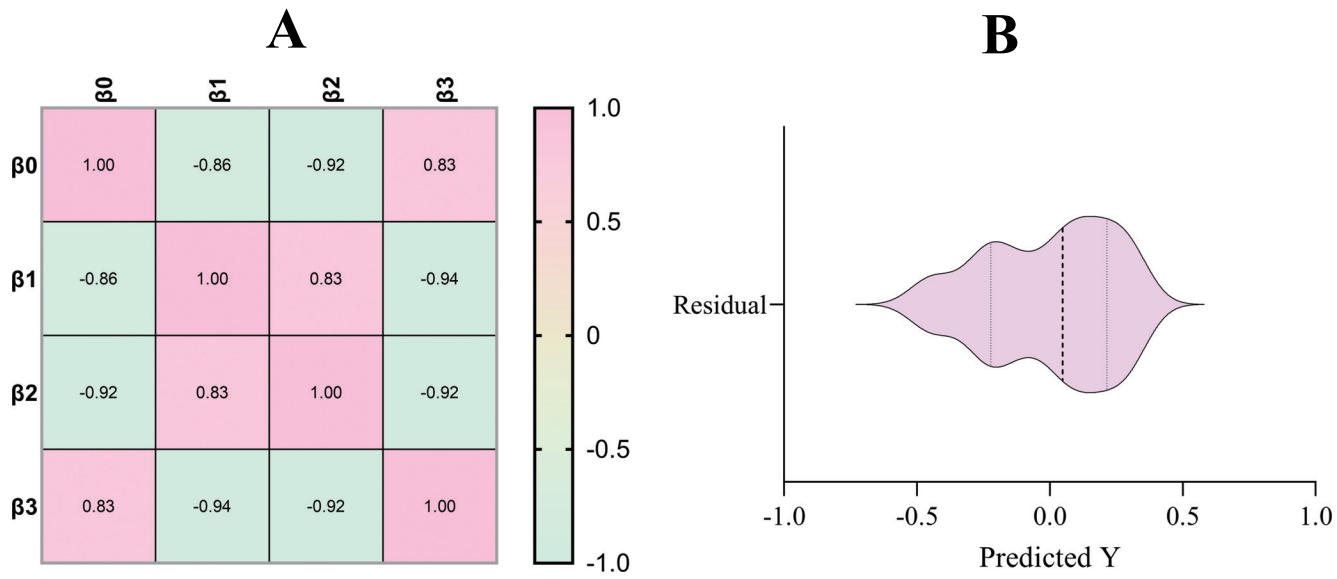


Fig. (10). The correlation matrix for gene expression and frequency and violin plot for the normality of residuals v/s the predicted *p* values. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Studies have stated that GATK has an F-score of 0.978, which is the harmonic mean of recall and precision, thereby making this toolkit a very reliable one for variant calling [16]. Moreover, GATK identifies all possible mutations and performs remarkably well in recognizing the true SNPs in the cancer exomes, making it the preferable choice over other somatic variant callers [17]. A recent study followed an NGS pipeline on the exonic and intronic sequences of colorectal cancer and identified in a single step, all different gene variations [16]. Different genes and variants are associated with them, such as POLE, POLD1, MSH3 and NTHL1. The clinical relevance of these different genes is yet to be understood, however, from among several identified mutations. The present study identified thousands of variants in terms of SNPs, indels, transitions, transversions, missense, nonsense and silent mutations. However, the exomes used were different in our study and exome sequence analysis was performed. Furthermore, the number of variants identified prior to annotation was large, which helped build a mutational profile.

A study involving the comparison of nine different *in silico* tools for post-processing of variants concluded that SIFT (Sorting intolerant from tolerant) stood as the most effective tool when taking into consideration the parameters of accuracy, sensitivity and specificity [52, 53]. SIFT classifies the variants into various classes such as “deleterious,” “tolerated,” “deleterious low confidence,” and “tolerated low confidence” [22, 33, 54]. This tool also predicts if the amino acid substitutions that have led to the formation of variants have an impact on the structure and function of a protein. SIFT performs the prediction of all the effects of possible amino acid substitutions at every position in the sequence [21]. The annotation of variants in the present study revealed tolerated, deleterious, non-coding, synonymous, non-synonymous, benign, possible and probably damaging mutations, from which the top ten genes with associated SNP changes were identified as non-synonymous and deleterious. Studies have shown that synonymous mutations are also responsible for acting as drivers in colorectal cancers [55, 56]. However, the present study focused primarily on understanding and identifying the non-synonymous mutations that could possibly be involved in colorectal cancer. A previous study detected somatic non-synonymous mutations associated with colorectal cancer in regions of liver metastases, 76.7% of the total somatic mutations detected, suggesting the importance of non-synonymous genes [57]. Another recent research noted a maximum number of non-synonymous mutations from WES-generated data of colorectal cancers, which were particular to metastatic tumors [58]. These studies point towards the importance of non-synonymous mutations that could be used as an indicator for colorectal cancer detection.

Due to their property of completely altering the amino acids, the non-synonymous variations are considered to be highly deleterious in nature, be it in germline or somatic variants. Additionally, non-synonymous SNPs can have adverse effects on proteins, such as altering the phenotype and genotype that may be the cause of a disease as dangerous as cancer [59, 60]. Moreover, the association of non-synonymous mutations with colorectal cancers has had some evidences in the past, with mutations being identified from exons [61]. A study has previously identified an E403K mutation in the *MCAK* gene, a non-synonymous variation to be linked with triggering colorectal cancer [62]. More recently, a study suggested that rare non-synonymous variations were associated with an increased risk of colorectal cancers, with several mutations identified in various genes [63]. These studies implicate non-synonymous variations as important colorectal cancer triggers and thus, more detailed studies are

needed to confirm this. However, in the present multi-dimensional study, we identified ten highly occurring non-synonymous, deleterious mutations in genes associated with these mutations, with two genes highly expressed every stage of the downstream expression analysis, implicating their usefulness in predisposing to colorectal cancer.

Additionally, Fig. (**4**) primarily shows the relative gene expression data in normal and tumor samples, and the predicted survival plots. As pointed out, for ELN, CTSB and CPEN1, the expression was higher than in normal samples and high expression reduced the rate of survival. The same was found to be true for RAD17 gene. However, upon further differential gene expression comparative analysis, it was noted that although all genes except OPRM1 and ADRA1A showed expression in colon and rectal adenocarcinoma, their levels of expression were not as high as CTSB and CPNE1, which demonstrated the maximum expression. These results were to further confirm, shortlist and identify those that showed high expression predictions from all conducted studies so as to have more reliability and reproducibility of the outcomes, and to evidence their involvement in colorectal cancers. Therefore, further *in-vitro* studies are indispensable, yet these outcomes pave the way for taking CTSB and CPNE1 forward for prospective studies. Computational analysis from both differential expression and survival plots highlighted CTSB and CPNE1, and therefore, these were implicated in colon and rectal cancers.

Previously, Yasuda *et al*., 2016 identified the mutation rs1045051 (on *RAD17*) to have an effect on colorectal cancer in a Japanese cohort [64]. Furthermore, expression of *LGALS8* in lung, prostate and colorectal cancer is considered to have potential prognostic capabilities, particularly in advanced stages in patients with distant metastases [65]. *CTSB* has also been identified as a potential target for colorectal cancer therapy, owing to its ability to contribute to tumor development and invasion [66]. The overexpression of *CPNE1* in tumors has been implicated in promoting the progression of colorectal cancer and metastasis [67]. Similarly, *SEMA4D* has also been studied to understand its role in various human malignancies, including breast, colon and pancreatic cancer [68]. The missense variant, rs1664022, in *PDE4DIP* was identified in all the exomes analyzed in the present study, suggesting the possibility of utilizing it as a potential indicator for colorectal cancer. However, the role of *PDE4DIP* is poorly understood in the context of cancer, and further studies are required to validate the role of this particular SNP in *PDE4DIP* in colorectal cancer. With some of these results corroborating the outcomes of the present study, our study implicates *CTSB* and *CPNE1* as important colorectal cancer indicators and suggests further experimental validation and comprehensive analysis for prospective studies. Our aim was to focus on the identification of variants that could possibly point towards the risk of colorectal cancer. This was thus achieved with supporting results of the high expression of germline variants in *CTSB,* and *CPNE1* predicted in colorectal cancer. The present work points towards results that will also have to be tested further *in vitro* to conclude the predictions.

CONCLUSION

Utilizing a comprehensive computational investigative approach to identify germline mutations in colorectal cancer exomes, this study revealed essential non-synonymous and deleterious SNPs that could potentially predispose to colorectal cancer. The present work utilized quality control tools such as FastQC and MultiQC,and filtered the variants generated using SIFT and PolyPhen2 that successfully categorized the mutations into synonymous, non-synonymous, start loss and stop gain mutations as well as marking them as possibly damaging, probably damaging and benign. Our work involved prioritizing the non-synonymous, deleterious SNPs since these polymorphisms bring about a functional alteration to the phenotype. The top variations associated with their genes with the highest frequency of occurrence included *LGALS8*, *CTSB*, *RAD17*, *CPNE1*, *OPRM1*, *SEMA4D*, *MUC4*, *PDE4DIP*, *ELN* and *ADRA1*A. An in-depth multi-dimensional downstream analysis of all these genes in terms of gene expression profiling and analysis and differential gene expression with regard to various cancer types revealed *CTSB* and *CPNE1* as highly expressed and overregulated genes in colorectal cancer. Several mutations were found to be predicted in these two genes in 4535 colorectal cancer *in-silico* patient analysis, which further validated our findings, warranting further experimental analysis as prospects for future studies. Our work sheds detailed light on the various alterations that might possibly lead to colorectal cancer and suggests further analysis into these genes to conclude the same.

AUTHORS' CONTRIBUTIONS

Chandrashekar K: Performed the experiments, collected all data, analysed the data and prepared figures for the manuscript.
Anagha S Setlur: Performed the experiments, collected all data, analysed the data and wrote the manuscript.
Dhanya Pradeep: Performed the experiments, analyzed the data and wrote the manuscript.
Jitendra Kumar: Reviewed the manuscript and analyzed the data.
Vidya Niranjan: Conceived the idea, analyzed all data and reviewed the manuscript.

LIST OF ABBREVIATIONS

BWA	=	Burrows-Wheeler Aligner
BWT	=	Burrow-Wheeler Transformation
MSI	=	Microsatellite Instability
NCBI SRA	=	National Centre for Biotechnology Information, Sequence Read Archive
NGS	=	Next-generation Sequencing Technology
SIFT	=	Sorting Intolerant from Tolerant
VCF	=	Variant Calling Files
WES	=	Whole Exome Sequencing

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

Not applicable.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The authors confirm that the data supporting the findings of this research are available within the article.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

We would like to acknowledge Dr. Shobha G, Professor, Department of Computer Science and Engineering, RV College of Engineering, Bangalore, for providing us with A100 GPU for performing computational analysis. We would also like to acknowledge Bayer's Medha for their support.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

[1] Colorectal Cancer Early Detection, Diagnosis, and Staging. **2022**. Available from: https://www.cancer.org/cancer/colon-rectal - cancer/detection-diagnosis-staging.html (Accessed on: 5th, August, **2022**).

[2] Yang, Y.; Sun, M.; Wang, L.; Jiao, B. HIFs, angiogenesis, and cancer. *J. Cell. Biochem.,* **2013**, *114*(5), 967-974. http://dx.doi.org/10.1002/jcb.24438 PMID: 23225225

[3] Hofree, M.; Carter, H.; Kreisberg, J.F.; Bandyopadhyay, S.; Mischel, P.S.; Friend, S.; Ideker, T. Challenges in identifying cancer genes by analysis of exome sequencing data. *Nat. Commun.,* **2016**, *7*(1), 12096-12096. http://dx.doi.org/10.1038/ncomms12096 PMID: 27417679

[4] Sanger, F.; Nicklen, S.; Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.,* **1977**, *74*(12), 5463-5467. http://dx.doi.org/10.1073/pnas.74.12.5463 PMID: 271968

[5] Guan, Y.F.; Li, G.R.; Wang, R.J.; Yi, Y.T.; Yang, L.; Jiang, D.; Zhang, X.P.; Peng, Y. Application of next-generation sequencing in clinical oncology to advance personalized treatment of cancer. *Chin. J. Cancer,* **2012**, *31*(10), 463-470. http://dx.doi.org/10.5732/cjc.012.10216 PMID: 22980418

[6] Cibulskis, K.; Lawrence, M.S.; Carter, S.L.; Sivachenko, A.; Jaffe, D.; Sougnez, C.; Gabriel, S.; Meyerson, M.; Lander, E.S.; Getz, G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.,* **2013**, *31*(3), 213-219. http://dx.doi.org/10.1038/nbt.2514 PMID: 23396013

[7] Vacante, M.; Borzì, A.M.; Basile, F.; Biondi, A. Biomarkers in colorectal cancer: Current clinical utility and future perspectives. *World J. Clin. Cases,* **2018**, *6*(15), 869-881. http://dx.doi.org/10.12998/wjcc.v6.i15.869 PMID: 30568941

[8] Andrews, S. Babraham bioinformatics-FastQC a quality control tool for high throughput sequence data. **2010**. Available from: https://www. bioinformatics. babraham. ac

[9] Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics,* **2016**, *32*(19), 3047-3048. http://dx.doi.org/10.1093/bioinformatics/btw354 PMID: 27312411

[10] He, X.; Chen, S.; Li, R.; Han, X.; He, Z.; Yuan, D.; Zhang, S.; Duan, X.; Niu, B. Comprehensive fundamental somatic variant calling and quality management strategies for human cancer genomes. *Brief. Bioinform.,* **2021**, *22*(3), bbaa083. http://dx.doi.org/10.1093/bib/bbaa083 PMID: 32510555

[11] Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods,* **2012**, *9*(4), 357-359. http://dx.doi.org/10.1038/nmeth.1923 PMID: 22388286

[12] Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M.; Li, H. Twelve years of SAMtools and BCFtools. *Gigascience,* **2021**, *10*(2), giab008. http://dx.doi.org/10.1093/gigascience/giab008 PMID: 33590861

[13] Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The sequence alignment/map format and samtools. *Bioinformatics,* **2009**, *25*(16), 2078-2079. http://dx.doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

[14] Patel, R.K.; Jain, M. NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS One,* **2012**, *7*(2), e30619-e30619. http://dx.doi.org/10.1371/journal.pone.0030619 PMID: 22312429

[15] McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; DePristo, M.A. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.,* **2010**, *20*(9), 1297-1303. http://dx.doi.org/10.1101/gr.107524.110 PMID: 20644199

[16] Supernat, A.; Vidarsson, O.V.; Steen, V.M.; Stokowy, T. Comparison of three variant callers for human whole genome sequencing. *Sci. Rep.,* **2018**, *8*(1), 17851. http://dx.doi.org/10.1038/s41598-018-36177-7 PMID: 30552369

[17] Hsu, Y.C.; Hsiao, Y.T.; Kao, T.Y.; Chang, J.G.; Shieh, G.S. Detection of somatic mutations in exome sequencing of tumor-only samples. *Sci. Rep.,* **2017**, *7*(1), 15959. http://dx.doi.org/10.1038/s41598-017-14896-7 PMID: 29162841

[18] Cingolani, P.; Platts, A.; Wang, L.L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S.J.; Lu, X.; Ruden, D.M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly,* **2012**, *6*(2), 80-92. http://dx.doi.org/10.4161/fly.19695 PMID: 22728672

[19] Cingolani, P.; Patel, V.M.; Coon, M.; Nguyen, T.; Land, S.J.; Ruden, D.M.; Lu, X. Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front. Genet.,* **2012**, *3*, 35-35. http://dx.doi.org/10.3389/fgene.2012.00035 PMID: 22435069

[20] Adzhubei, I.; Jordan, D.M.; Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.,* **2013**, *76*(1), 20. http://dx.doi.org/10.1002/0471142905.hg0720s76 PMID: 23315928

[21] Hu, J.; Ng, P.C. SIFT Indel: Predictions for the functional effects of amino acid insertions/deletions in proteins. *PLoS One,* **2013**, *8*(10), e77940. http://dx.doi.org/10.1371/journal.pone.0077940 PMID: 24194902

[22] LaFramboise, W.A.; Pai, R.K.; Petrosko, P.; Belsky, M.A.; Dhir, A.; Howard, P.G.; Becich, M.J.; Holtzman, M.P.; Ahrendt, S.A.; Pingpank, J.F.; Zeh, H.J.; Dhir, R.; Bartlett, D.L.; Choudry, H.A. Discrimination of low- and high-grade appendiceal mucinous neoplasms by targeted sequencing of cancer-related variants. *Mod. Pathol.,* **2019**, *32*(8), 1197-1209.

http://dx.doi.org/10.1038/s41379-019-0256-2 PMID: 30962504

[23] Ernst, C.; Hahnen, E.; Engel, C.; Nothnagel, M.; Weber, J.; Schmutzler, R.K.; Hauke, J. Performance of *in silico* prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. *BMC Med. Genomics,* **2018**, *11*(1), 35. http://dx.doi.org/10.1186/s12920-018-0353-y PMID: 29580235

[24] Hicks, S.; Wheeler, D.A.; Plon, S.E.; Kimmel, M. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Hum. Mutat.,* **2011**, *32*(6), 661-668. http://dx.doi.org/10.1002/humu.21490 PMID: 21480434

[25] Padmavathi, P.; Setlur, A.S.; Chandrashekar, K.; Niranjan, V. A comprehensive *in-silico* computational analysis of twenty cancer exome datasets and identification of associated somatic variants reveals potential molecular markers for detection of varied cancer types. *Inform. Med. Unlocked.,* **2021**, *26*, 100762. http://dx.doi.org/10.1016/j.imu.2021.100762

[26] Dietz, S.; Schirmer, U.; Mercé, C.; von Bubnoff, N.; Dahl, E.; Meister, M.; Muley, T.; Thomas, M.; Sültmann, H. Low input whole-exome sequencing to determine the representation of the tumor exome in circulating dna of non-small cell lung cancer patients. *PLoS One,* **2016**, *11*(8), e0161012-e0161012. http://dx.doi.org/10.1371/journal.pone.0161012 PMID: 27529345

[27] Tang, Z.; Li, C.; Kang, B.; Gao, G.; Li, C.; Zhang, Z. GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.,* **2017**, *45*(W1), W98-W102. http://dx.doi.org/10.1093/nar/gkx247 PMID: 28407145

[28] Dingerdissen, H.M.; Bastian, F.; Vijay-Shanker, K.; Robinson-Rechavi, M.; Bell, A.; Gogate, N.; Gupta, S.; Holmes, E.; Kahsay, R.; Keeney, J.; Kincaid, H.; King, C.H.; Liu, D.; Crichton, D.J.; Mazumder, R. OncoMX: A knowledgebase for exploring cancer biomarkers in the context of related cancer and healthy data. *JCO Clin. Cancer Inform.,* **2020**, *4*(4), 210-220. http://dx.doi.org/10.1200/CCI.19.00117 PMID: 32142370

[29] Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B.E.; Sumer, S.O.; Aksoy, B.A.; Jacobsen, A.; Byrne, C.J.; Heuer, M.L.; Larsson, E.; Antipin, Y.; Reva, B.; Goldberg, A.P.; Sander, C.; Schultz, N. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.,* **2012**, *2*(5), 401-404. http://dx.doi.org/10.1158/2159-8290.CD-12-0095 PMID: 22588877

[30] Gao, J.; Aksoy, B.A.; Dogrusoz, U.; Dresdner, G.; Gross, B.; Sumer, S.O.; Sun, Y.; Jacobsen, A.; Sinha, R.; Larsson, E.; Cerami, E.; Sander, C.; Schultz, N. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.,* **2013**, *6*(269), pl1-pl1. http://dx.doi.org/10.1126/scisignal.2004088 PMID: 23550210

[31] Boyko, A.A.; Kukartsev, V.V.; Tynchenko, V.S.; Korpacheva, L.N.; Dzhioeva, N.N.; Rozhkova, A.V.; Aponasenko, S.V. Using linear regression with the least squares method to determine the parameters of the Solow model. *J. Phys.: Conf. Ser.,* **2020**, *1582*, 012016.

[32] Swift, M.L. GraphPad prism, data analysis, and scientific graphing. *J. Chem. Inf. Comput. Sci.,* **1997**, *37*(2), 411-412. http://dx.doi.org/10.1021/ci960402j

[33] Ng, P.C.; Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.,* **2003**, *31*(13), 3812-3814. http://dx.doi.org/10.1093/nar/gkg509 PMID: 12824425

[34] Azzopardi, D.; Dallosso, A.R.; Eliason, K.; Hendrickson, B.C.; Jones, N.; Rawstorne, E.; Colley, J.; Moskvina, V.; Frye, C.; Sampson, J.R.; Wenstrup, R.; Scholl, T.; Cheadle, J.P. Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas. *Cancer Res.,* **2008**, *68*(2), 358-363. http://dx.doi.org/10.1158/0008-5472.CAN-07-5733 PMID: 18199528

[35] Thurston, T.L.M.; Wandel, M.P.; von Muhlinen, N.; Foeglein, Á.; Randow, F. Galectin 8 targets damaged vesicles for autophagy to defend cells against bacterial invasion. *Nature,* **2012**, *482*(7385), 414-418. http://dx.doi.org/10.1038/nature10744 PMID: 22246324

[36] Staring, J.; von Castelmur, E.; Blomen, V.A.; van den Hengel, L.G.; Brockmann, M.; Baggen, J.; Thibaut, H.J.; Nieuwenhuis, J.; Janssen, H.; van Kuppeveld, F.J.M.; Perrakis, A.; Carette, J.E.; Brummelkamp, T.R. PLA2G16 represents a switch between entry and clearance of Picornaviridae. *Nature,* **2017**, *541*(7637), 412-416. http://dx.doi.org/10.1038/nature21032 PMID: 28077878

[37] Mani, A. PDE4DIP in health and diseases. *Cell. Signal.,* **2022**, *94*, 110322. http://dx.doi.org/10.1016/j.cellsig.2022.110322 PMID: 35346821

[38] Guo, R.; Rowe, P.S.N.; Liu, S.; Simpson, L.G.; Xiao, Z.S.; Darryl, Q.L. Inhibition of MEPE cleavage by Phex. *Biochem. Biophys. Res. Commun.,* **2002**, *297*(1), 38-45. http://dx.doi.org/10.1016/S0006-291X(02)02125-3 PMID: 12220505

[39] Li, L.; Peterson, C.A.; Kanter-Smoler, G.; Wei, Y.F.; Ramagli, L.S.; Sunnerhagen, P.; Siciliano, M.J.; Legerski, R.J. hRAD17, a structural homolog of the Schizosaccharomyces pombe RAD17 cell cycle checkpoint gene, stimulates p53 accumulation. *Oncogene,* **1999**, *18*(9), 1689-1699. http://dx.doi.org/10.1038/sj.onc.1202469 PMID: 10208430

[40] Keeley, F.W.; Bellingham, C.M.; Woodhouse, K.A. Elastin as a self–organizing biomaterial: Use of recombinantly expressed human elastin polypeptides as a model for investigations of structure and self–assembly of elastin. *Philos. Trans. R. Soc. Lond. B Biol. Sci.,* **2002**, *357*(1418), 185-189. http://dx.doi.org/10.1098/rstb.2001.1027 PMID: 11911775

[41] Tomsig, J.L.; Sohma, H.; Creutz, C.E. Calcium-dependent regulation of tumour necrosis factor-alpha receptor signalling by copine. *Biochem. J.,* **2004**, *378*(3), 1089-1094. http://dx.doi.org/10.1042/bj20031654 PMID: 14674885

[42] Pan, Y.X.; Xu, J.; Mahurter, L.; Xu, M.; Gilbert, A.K.; Pasternak, G.W. Identification and characterization of two new human mu opioid receptor splice variants, hMOR-1O and hMOR-1X. *Biochem. Biophys. Res. Commun.,* **2003**, *301*(4), 1057-1061. http://dx.doi.org/10.1016/S0006-291X(03)00089-5 PMID: 12589820

[43] Janssen, B.J.C.; Robinson, R.A.; Pérez-Brangulí, F.; Bell, C.H.; Mitchell, K.J.; Siebold, C.; Jones, E.Y. Structural basis of semaphorin–plexin signalling. *Nature,* **2010**, *467*(7319), 1118-1122. http://dx.doi.org/10.1038/nature09468 PMID: 20877282

[44] Moniaux, N.; Escande, F.; Batra, S.K.; Porchet, N.; Laine, A.; Aubert, J.P. Alternative splicing generates a family of putative secreted and membrane-associated MUC4 mucins. *Eur. J. Biochem.,* **2000**, *267*(14), 4536-4544. http://dx.doi.org/10.1046/j.1432-1327.2000.01504.x PMID: 10880978

[45] Wright, C.D.; Chen, Q.; Baye, N.L.; Huang, Y.; Healy, C.L.; Kasinathan, S.; O'Connell, T.D. Nuclear α1-adrenergic receptors signal activated ERK localization to caveolae in adult cardiac myocytes. *Circ. Res.,* **2008**, *103*(9), 992-1000. http://dx.doi.org/10.1161/CIRCRESAHA.108.176024 PMID: 18802028

[46] Jaswanth Jenny, P.; Dhamotharan, R. Exome data analysis in the discovery of variants associated with breast cancer metastasis and their implications on protein structure. *Ann. Rom. Soc. Cell Biol.,* **2021**, *2021*, 1663-1682.

[47] Agarwal, R.; Cao, Y.; Hoffmeier, K.; Krezdorn, N.; Jost, L.; Meisel, A.R.; Jüngling, R.; Dituri, F.; Mancarella, S.; Rotter, B.; Winter, P.; Giannelli, G. Precision medicine for hepatocelluar carcinoma using molecular pattern diagnostics: results from a preclinical pilot study. *Cell Death Dis.,* **2017**, *8*(6), e2867-e2867. http://dx.doi.org/10.1038/cddis.2017.229 PMID: 28594404

[48] Li, H.; Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics,* **2010**, *26*(5), 589-595. http://dx.doi.org/10.1093/bioinformatics/btp698 PMID: 20080505

[49] Xu, X.; Zhou, Y.; Feng, X.; Li, X.; Asad, M.; Li, D.; Liao, B.; Li, J.; Cui, Q.; Wang, E. Germline genomic patterns are associated with cancer risk, oncogenic pathways, and clinical outcomes. *Sci. Adv.,* **2020**, *6*(48), eaba4905. http://dx.doi.org/10.1126/sciadv.aba4905 PMID: 33246949

[50] Toma, C.; Díaz-Gay, M.; Franch-Expósito, S.; Arnau-Collell, C.;

Overs, B.; Muñoz, J.; Bonjoch, L.; Soares de Lima, Y.; Ocaña, T.; Cuatrecasas, M.; Castells, A.; Bujanda, L.; Balaguer, F.; Cubiella, J.; Caldés, T.; Fullerton, J.M.; Castellví-Bel, S. Using linkage studies combined with whole exome sequencing to identify novel candidate genes for familial colorectal cancer. *Int. J. Cancer,* **2020**, *146*(6), 1568-1577. http://dx.doi.org/10.1002/ijc.32683 PMID: 31525256

[51] Chatrath, A.; Ratan, A.; Dutta, A. Germline variants that affect tumor progression. *Trends Genet.,* **2021**, *37*(5), 433-443. http://dx.doi.org/10.1016/j.tig.2020.10.005 PMID: 33203571

[52] Baert-Desurmont, S.; Coutant, S.; Charbonnier, F.; Macquere, P.; Lecoquierre, F.; Schwartz, M.; Blanluet, M.; Vezain, M.; Lanos, R.; Quenez, O.; Bou, J.; Bouvignies, E.; Fourneaux, S.; Manase, S.; Vasseur, S.; Mauillon, J.; Gerard, M.; Marlin, R.; Bougeard, G.; Tinat, J.; Frebourg, T.; Tournier, I. Optimization of the diagnosis of inherited colorectal cancer using NGS and capture of exonic and intronic sequences of panel genes. *Eur. J. Hum. Genet.,* **2018**, *26*(11), 1597-1602. http://dx.doi.org/10.1038/s41431-018-0207-2 PMID: 29967336

[53] Pshennikova, V.G.; Barashkov, N.A.; Romanov, G.P.; Teryutin, F.M.; Solov'ev, A.V.; Gotovtsev, N.N.; Nikanorova, A.A.; Nakhodkin, S.S.; Sazonov, N.N.; Morozov, I.V.; Bondar, A.A.; Dzhemileva, L.U.; Khusnutdinova, E.K.; Posukh, O.L.; Fedorova, S.A. Comparison of predictive *in silico* tools on missense variants in *GJB2*, *GJB6*, and *GJB3* genes associated with autosomal recessive deafness 1A (DFNB1A). *ScientificWorldJournal,* **2019**, *2019*, 1-9. http://dx.doi.org/10.1155/2019/5198931 PMID: 31015822

[54] Adzhubei, I.A.; Schmidt, S.; Peshkin, L.; Ramensky, V.E.; Gerasimova, A.; Bork, P.; Kondrashov, A.S.; Sunyaev, S.R. A method and server for predicting damaging missense mutations. *Nat. Methods,* **2010**, *7*(4), 248-249. http://dx.doi.org/10.1038/nmeth0410-248 PMID: 20354512

[55] Supek, F.; Miñana, B.; Valcárcel, J.; Gabaldón, T.; Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell,* **2014**, *156*(6), 1324-1335. http://dx.doi.org/10.1016/j.cell.2014.01.051 PMID: 24630730

[56] Bin, Y.; Wang, X.; Zhao, L.; Wen, P.; Xia, J. An analysis of mutational signatures of synonymous mutations across 15 cancer types. *BMC Med. Genet.,* **2019**, *20*(S2), 190. http://dx.doi.org/10.1186/s12881-019-0926-4 PMID: 31815613

[57] Oga, T.; Yamashita, Y.; Soda, M.; Kojima, S.; Ueno, T.; Kawazu, M.; Suzuki, N.; Nagano, H.; Hazama, S.; Izumiya, M.; Koike, K.; Mano, H. Genomic profiles of colorectal carcinoma with liver metastases and newly identified fusion genes. *Cancer Sci.,* **2019**, *110*(9), 2973-2981. http://dx.doi.org/10.1111/cas.14127 PMID: 31293054

[58] Tang, J.; Tu, K.; Lu, K.; Zhang, J.; Luo, K.; Jin, H.; Wang, L.; Yang, L.; Xiao, W.; Zhang, Q.; Liu, X.; Ge, X.; Li, G.; Zhou, Z.; Xie, D. Single-cell exome sequencing reveals multiple subclones in metastatic colorectal carcinoma. *Genome Med.,* **2021**, *13*(1), 148-148. http://dx.doi.org/10.1186/s13073-021-00962-3 PMID: 34507604

[59] Kulshreshtha, S.; Chaudhary, V.; Goswami, G.K.; Mathur, N. Computational approaches for predicting mutant protein stability. *J. Comput. Aided Mol. Des.,* **2016**, *30*(5), 401-412. http://dx.doi.org/10.1007/s10822-016-9914-3 PMID: 27160393

[60] Hassan, M.S.; Shaalan, A.A.; Dessouky, M.I.; Abdelnaiem, A.E.; ElHefnawi, M. A review study: Computational techniques for expecting the impact of non-synonymous single nucleotide variants in human diseases. *Gene,* **2019**, *680*, 20-33. http://dx.doi.org/10.1016/j.gene.2018.09.028 PMID: 30240882

[61] Prasad, V.V.T.S.; Padma, K. Non-synonymous polymorphism (Gln261Arg) of 12-lipoxygenase in colorectal and thyroid cancers. *Fam. Cancer,* **2012**, *11*(4), 615-621. http://dx.doi.org/10.1007/s10689-012-9559-x PMID: 22864639

[62] Kumar, A.; Rajendran, V.; Sethumadhavan, R.; Purohit, R. Evidence of colorectal cancer-associated mutation in MCAK: A computational report. *Cell Biochem. Biophys.,* **2013**, *67*(3), 837-851. http://dx.doi.org/10.1007/s12013-013-9572-1 PMID: 23564489

[63] Yu, L.; Yin, B.; Qu, K.; Li, J.; Jin, Q.; Liu, L.; Liu, C.; Zhu, Y.; Wang, Q.; Peng, X.; Zhou, J.; Cao, P.; Cao, K. Screening for susceptibility genes in hereditary non-polyposis colorectal cancer. *Oncol. Lett.,* **2018**, *15*(6), 9413-9419. http://dx.doi.org/10.3892/ol.2018.8504 PMID: 29844832

[64] Yasuda, Y.; Sakai, A.; Ito, S.; Sasai, K.; Ishizaki, A.; Okano, Y.; Kawahara, S.; Jitsumori, Y.; Yamamoto, H.; Matsubara, N.; Shimizu, K.; Katayama, H. Human NINEIN polymorphism at codon 1111 is associated with the risk of colorectal cancer. *Biomed. Rep.,* **2020**, *13*(5), 1. http://dx.doi.org/10.3892/br.2020.1352 PMID: 32934817

[65] Elola, MT; Ferragut, F; Cardenas, DVM; Nugnes, LG; Gentilini, L; Laderach, D; Troncoso, MF; Compagno, D; Wolfenstein-Tode, C; Rabinovich, GA Expression, localization and function of galectin-8, a tandem-repeat lectin, in human tumors. *Histol Histopathol,* **2014**, *29*(9), 1093-1105.

[66] Bian, Z.; Jin, L.; Zhang, J.; Yin, Y.; Quan, C.; Hu, Y.; Feng, Y.; Liu, H.; Fei, B.; Mao, Y.; Zhou, L.; Qi, X.; Huang, S.; Hua, D.; Xing, C.; Huang, Z. LncRNA—UCA1 enhances cell proliferation and 5-fluorouracil resistance in colorectal cancer by inhibiting miR-204-5p. *Sci. Rep.,* **2016**, *6*(1), 23892-23892. http://dx.doi.org/10.1038/srep23892 PMID: 27046651

[67] Wang, Y.; Pan, S.; He, X.; Wang, Y.; Huang, H.; Chen, J.; Zhang, Y.; Zhang, Z.; Qin, X. CPNE1 Enhances colorectal cancer cell growth, glycolysis, and drug resistance through regulating the AKT-GLUT1/HK2 pathway. *OncoTargets Ther.,* **2021**, *14*, 699-710. http://dx.doi.org/10.2147/OTT.S284211 PMID: 33536762

[68] Rezaeepoor, M.; Rashidi, G.; Pourjafar, M.; Mohammadi, C.; Solgi, G.; Najafi, R. SEMA4D knockdown attenuates *β*-catenin-dependent tumor progression in colorectal cancer. *BioMed Res. Int.,* **2021**, *2021*, 1-12. http://dx.doi.org/10.1155/2021/8507373 PMID: 34337054