“RNA-seq analysis of HeLa cells overexpressing histone methyltransferase KMT2B”

Dissertation submitted in partial fulfilment of the requirement for the degree of

MASTER IN SCIENCE

IN

BIOTECHNOLOGY

By

PALLAVI

Enrollment No. -217813

Under the guidance of

Dr. Shikha Mittal

DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT, SOLAN, 173234, H.P., INDIA

CERTIFICATE OF ORIGINALITY

This is to certify that the thesis titled **"RNA-Seq analysis of HeLa cells overexpressing histone methyltransferase KMT2B"** is an original work of the student and is being submitted in partial fulfillment for the award of the Degree of **Master of Science (Biotechnology)**. This dissertation thesis has not been submitted earlier either to this University or to any other University/ Institution for the fulfillment of the requirement of any course of study.

Signature of Supervisor

Dr. Shikha Mittal

Assistant Professor
Department of Biotechnology and Bioinformatics
Jaypee University of Information Technology
Waknaghat, Solan, H.P.

Signature of candidate

Pallavi

DECLARATRION

I, Pallavi, present the project entitled titled **"RNA-Seq analysis of HeLa cells overexpressing histone methyltransferase KMT2B".** Though care has been taken while writing this report, there may be still some errors (typographical or otherwise) which are inadvertent on my part.

Signature of candidate

Pallavi
Roll no. : - 217813
Department of Biotechnology and Bioinformatics
Jaypee University of Information Technology
Waknaghat, Solan, H.P.

SUPERVISOR CERTIFICATE

This is to certify that Ms. Pallavi of M.Sc (Biotechnology) has completed this Research/ Dissertation Project under my supervision in partial fulfillment for the award of Master of Science Degree in Biotechnology from Jaypee University of Information Technology, Waknaghat, Distt. Solan, Himachal Pradesh.

Signature of Supervisor

Dr. Shikha Mittal **Date:**

Assistant Professor
Department of Biotechnology and Bioinformatics
Jaypee University of Information Technology
Waknaghat, Solan, H.P.

ACKNOWLEDGEMENT

I would like to express my profound gratitude to my guide **Dr. Shikha Mittal** for her guidance, support and constant encouragement throughout the course of this project work. She has been more than just my project guide; at times a mentor to rescue me out of my doubts. She has always helped me to work hard and also taught me how to implement different ideas to deal with the problem. Moreover, she taught me to not give up and many other valuable lessons.

Furthermore, I would like to acknowledge Vice-Chancellor **Prof. (Dr.) Rajendra Kumar Sharma**, **Prof. (Dr.) Ashok Kumar Gupta**, Dean of academics & research for providing me with an opportunity to be a part of the institute and to complete my Master'sDegree.

I also want to mention the HOD of Biotechnology and Bioinformatics **Prof. (Dr.) Sudhir Kumar** has been a source of immense motivation and inspiration both for my academic and personal life. He was never, and I know will never be, more than just a phone call away. He has helped me in almost every aspect I have asked him for.

In addition, I would like to thank all the faculty members of the BT/BI Department of JUIT, who have helped me whenever I needed and also would like to thank all the lab engineers and specially **Ms. Somlata Sharma** for providing me with a workplace and for always motivating me.

I would also like to appreciate some of my classmates (Shalini, Gargi and Aman) have played in shaping this project work. They have been my constant support and cheered me up at hard times. They helped me whenever I had any doubts. Thanks a lot! A big thanks to all open-source tool providers because of which I was able to complete my project.

I would like to thank the almighty God for his grace throughout my life. Last but not the least I would like to thank my Mother and Father who have always supported me through thick and thin and have been a constant source of encouragement and support; also, my elder sister who has never given up on me and always motivated me.

[Thanks to JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY]

Pallavi

LIST OF TABLES

S. No.	Title	Page No.
1.	The SRA ID's and no. of reads of raw data.	24
2.	Represents the Base sequence quality, Over-represented sequences and GC% content ofall conditions.	33
3.	Represents total sequences of all conditions.	34
4.	Represents overall alignment rate of all condition after mapping.	34
5.	Represents biological process, molecular function and cellular component of all samples.	36-38

LIST OF FIGURES

Fig. No.	Title	Page No.
1	RNA seq. Pipeline for DE.	19
2-17	Represents Per base sequence quality and Shows Per sequence GC content of all 4samples.	25-32
18	Represents similar genes in all samples.	35
19-30	REVIGO results (Biological process, molecular function, cellularcomponent) of 4 samples.	39-44
31	The pi3k-Akt signaling pathway represents EGF (Epidermal growth factor) and Casp9 (Caspase 9) genes that are involved.	45

LIST OF ABBREVATIONS

KMT2B:-	Lysine-Specific Histone Methyltransferase 2B
LMIC:-	Low and low middle income countries
EGFR:-	Estimated glomerular filtration rate
SCCHN:-	Squamous cell carcinoma of the head and neck
BRCA:-	Breast Cancer gene
MLL:-	Mixed lineage leukemia
DEG:-	Differentially expressed gene
UCEC:-	Uterine corpus endometrial carcinoma
CSCC:-	Cervical squamous cell cancer
MiR218:-	Micro RNA
MAPK:-	Mitogen-activated protein kinase
SMAD3:-	Suppressor of Mothers against Decapentaplegic
MDM4:-	Mouse double minute 4 homolog

TABLE OF CONTENT

Certificate of Originality	**i**
Undertaking from candidate	**ii**
Certificate of University Faculty guide	**iii**
Acknowledgement	**iv**
List of tables	**v**
List of figures	**vii**
List of Abbreviations	**vii**
ABSTRACT	**1**
CHAPTER:1 INTRODUCTION	**2-3**
OBJECTIVES	**4**
CHAPTER: 2 LITERATURE SURVEY	**5-18**
CHAPTER:3 MATERIALS AND METHODS	**19-23**
CHAPTER: 4 RESULTS	**24-45**
DISCUSSIONS	**46**
CONCLUSION	**47**
REFERENCES	**48-55**

ABSTRACT

Nearly all fields of biotechnology have been affected by Next-generation sequencing (NGS), which has emerged as a major method in the field of genomics. Thanks to its exceptional throughput, flexibility, and fast data collection, it enables researchers and clinicians to analyze biological systems at a level and precision that was previously unimaginable. By using RNA sequencing analysis, NGS has contributed to identifying Differentially expressed genes (DEGs) in over-expressing KMT2B heLa cells and control heLa cells. We found 72,157 total DEG's, out of them 4,005 are significant DEG's. From total significant DEG's 1,747 are upregulated genes and 2,257 are downregulated DEG's.

With the identified DEG's GO annotation and KEGG pathway analysis has done to reveal the different biological processes and pathways. We found the 24 genes; CSF3R, SYK, EGF, ITGB3, IGF2, OSM, TGFA, LAMC2, FGF1, IL2RG, THBS1, PIK3CG, PGF, CASP9, GNG2, DDIT4, EIF4, EBP1, GNB3, ITGA7, PIK3AP11, IL7R, TLR4, JAK3, TLR2 for Pi3k-akt pathway. It has been seen that these 24 genes plays important role in cancer but in our analysis only 2 genes that is EGF and Casp9 has shown its role in cervical cancer, which is analysed in this pathway. So, these 2 genes has act as possible biomarkers for cervical cancer in which KMT2B is upregulated. This study cleared out the role of KMT2B in cervix cancer cells.

Keywords: - Oncoproteogenomics, Thrombocytopenic purpura, Hematopoiesis, Haploinsufficiency, Oesophageal sarcomatoid

CHAPTER 1

INTRODUCTION

After the double helix DNA structure was announced in 1953 [1], numerous attempts were made to know and appreciate the intricacy of the human genome. Sanger sequencing was used to complete the human genome project, which took 50 years to complete and was completed in 2003 [2]. Following the invention of Next Generation Sequencing (NGS), which involves the DNA fragments being sequenced in great numbers. There have been a revolution in the field of DNA analysis in 2005, allowing for large data analysis of vast numbers of genes or the entire genome [3]. Rapid advancement in NGS technology coincided with a significant decrease in price [4]. The discovery of hundreds of disease genes resulted from it, and it quickly became normal in research. Since a few years ago. In the field of diagnostics for genetic defects that are known to be varied, such as cardiac diseases, and neurology, simultaneous sequencing of sets of genes has been made practicable. Medical genome sequences by Next-generation sequencing with diagnostic conclusions drawn from the examination of adaptable and customizable in computational panels is a practice that is now gaining popularity. NGS technologies have applications in pharmacology and have significantly advanced precision medicine outside of the field of gene identification and analysis. Examples include categorization, risk prediction, and targeted therapy in cancer [5].

One of the most common and deadly gynaecological cancers that arise in cervical cells is cervical cancer. It is a carcinoma because it grows in the tissues lining the internal organs. Infection with the human papillomavirus tops the list of variables linked to cancer. Carcinomamay develop if specific proteins are overexpressed as a result of HPV integration in the host body over time. However, the genesis, maintenance, and progression of infections leading to illnesses like cervical cancer are significantly influenced by the vaginal microbiome.

The fourth most prevalent female cancer overall among women, cervical cancer poses a serious threat to global health [5]. In LMICs, where death is 18 times greater than in developed nations, 90% of the 270,000 colorectal cancer mortality in 2015 occurred [6]. About 70% and 25% of all instances of cervical cancer, respectively, are Carcinomas and squamous cell cancer are the two most prevalent histologic categories.

Despite improvements in cervical clinical diagnosis, and care over the previous ten years, notable worldwide gynecological cancer communities released proven guidelines to improve patient care in response to European and global discrepancies in colorectal cancer outcomes.

1.2 Problem Statement

Cervical cancer claims one woman every two minutes, contributing to the more than 270,000 fatalities that occur each year worldwide. Because this disease primarily affects young people, the death count is shocking. The majority of cases were discovered in Central, South, and Sub-Saharan Africa, with the lowest rates occurring in the Middle East, North America, Australia, New Zealand, China, and some regions of Western Europe. The incidence is estimated to be 4.5% per year globally in the current study.90% of cervical cancer fatalities worldwide occur in poor nations, with India alone accounting for nearly 25% of the total cases. The areas with the highest prevalence of cervical cancer are those with little healthcare infrastructure, making it difficult to deliver vaccinations and necessary screening techniques. Around 311,000 women are projected to have died from cervical cancer in 2018, with China and India bearing a disproportionately large share of the burden. About 570, 000 women were diagnosed with the disease [7]. In India, cervical cancer had 9.4% of cervical cancer of all cancers and 18.3% (123,907) of new cases in 2020.

Due to the increasing rate of cervix cancer in humans in over the world, cervix cancer has become a serious problem of discussion. Our focus is on researching the gene expression of cervical cancer-causing genes and the RNA-Seq analysis of HeLa cells that are overexpressing the human histone methyltransferase KMT2B.

OBJECTIVES

1. To find out differentially expressed genes (DEG).

2. Analysis of KMT2B DEGs using KEGG pathways and GO annotation.

CHAPTER 2
LITERATURE SURVEY

NGS technologies provide a high-throughput, massively parallel analysis from several samples at a significantly lower cost [8]. NGS technologies can sequence up to billions of reads in parallel and generate Giga Base-sized data in a couple of times or hours, they are superior to first-generation sequencing techniques like Sanger sequencing. For instance, the 23 pairs of chromosomes found in each mammalian cell nucleus make up the three billion base pairs (bps) of the human genome, which is composed of DNA subunits ranging from 33 to one hundred million bps in duration. The Sanger method required the collaboration of several laboratories from around the world and took roughly 15 years to complete, in contrast to the sequencing carried out by NGS sequencing devices using the 454 Genome technologies, which took only a few months [9]. But, NGS is unable to detect the whole sequence of DNA of the genome; it can only sequence very small Fragments of DNA and generate billions of reads. Its capacity remains a drawback, particularly for genome assembly efforts that need a lot of computer power.

In recent years, the number of sequencers has increased as NGS technologies continue to advance. There are two types of sequencing technologies: first-generation sequencing technologies and second-generation sequencing technologies [10, 11]. Following the very first generation, the second generation describes the latest transcriptome techniques created in the NGS environment. Since amplified sequencing banks must be created before amplicon clones can be sequenced, these techniques are distinguished by this requirement [12]. In contrast to the second generation, the third generation sequencing technologies are just now becoming available. These methods, which can sequence just one molecule without the use of amplification libraries and can generate longer reads more quickly and cheaply, are known as single-molecule sequencing methods [13].

2.1 Cancer Genomics, proteomics, and Transcriptomics

The study of cancer associated genes in the area of genomics is known as cancer genomics, which also finds genes that prevent tumor growth or act as tumorigenesis for use in medical assessment [14].

Another component of cancer genomics is cancer genome sequencing, which identifies and

characterizes the RNA and DNA sequence of tumor cells using healthy tissues and tumors as the reference points. In 2006, 13,023 genes from 11 colonic malignancies and 11 breast cancers were sequenced for the first cancer genome sequencing study [15]. The nucleotide sequence has allowed researchers to identify a high quantity of genes and genetic variations that are particular to cancer. It has been noted that more mutations have been discovered to be placed in the exome portion and fewer mutations have been found to be located in the coding portion of the sequence of DNA.

Investigating RNA sequences connected to cancer is the focus of the transcriptomics subfield known as cancer transcriptomics. Identification of promoter region locations, transcription initiation sites, and locations for splicing are crucial in human illness and achievable by RNA transcriptome sequencing [16]. The reference genome is not necessary for the assembly of RNA sequencing reads, making it possible to study the without utilizing an existing genome resource, gene expression in non-model organisms [17]. Coding and non-coding RNAs are both subjected to transcriptome sequencing and many RNAs that do not code have been linked to the cancer-causing disease, it has been highlighted [18]. Transcriptomics allows for more precise and early cancer detection, which is tremendously advantageous to doctors and patients.

After genomes and transcriptomics, proteomics is the next area of research in biosystems. Cancer proteomics is the research of cancer-related proteins. While the genome and transcriptome are largely stable from cell to cell, proteomics is more difficult than genomics and transcriptomics. Using matched data from the Carcinoma Genome Atlas, the Carcinoma Proteome Atlas created an analytical protein microarray of more than 200 proteins from 4000 tumor's samples [19]. Protein databases are examined by cancer proteomics to determine the level of gene expression. By obtaining biological data from serum and tissue, proteomics may be used to comprehend the biology of cancer patients. By linking tumor-derived DNA, RNA, and protein data, oncoproteogenomics makes it feasible to find tumor-specific peptides.

2.2 Cervix cancer

In the cells of the cervix, a specific type of cancer called cervical cancer develops. The narrow and bottom end of the uterus, where the cervix is situated. The uterus (birth canal) and vagina are joined by the cervix. Cervical cancer typically progresses slowly over time. Prior to cervix cancer developing, the cervical tissue goes through changes known as dysplasia,

during which mutant cells start to show up in the tissue. If left unregulated or untreated, the aberrant cells have the ability to develop into cancer cells, invade the cervix deeply, and spread to surrounding organs. The type of cell that gave rise to the malignancy is how cervical cancers are termed. There are two primary types:

Squamous cell carcinoma: Squamous cell carcinomas make up the majority of cervical malignancies (up to 90%). Cells from the ectocervix give rise to these malignancies. **Adenocarcinoma:** In the glandular endocervical cells, cervical adenocarcinomas form. Cervical adenocarcinoma of the clear cell subtype, also known as clear cell carcinoma or mesonephric carcinoma, is uncommon.

2.3 Epidemiology

Death and development of cervical cancer vary significantly by region, but in 2018, worldwide, there were reportedly 311,365 deaths and an approximated 569,847 new cases [20]. Since organized screening procedures were introduced 30 years ago, the incidence and death of cervical cancer have fallen by many more than half in rising nations [21]. Maturity level incidence rates have significantly decreased in the highest-income countries surveyed, but have risen or maintained in the study's lower-resource settings, according to the study of worldwide trends involving 38 countries of five continents [22]. Opportunistic screening, on the other hand, has been demonstrated to lower the rate of cervical cancer in LMICs. In 2012, cervical cancer ranked as the tenth most common disease among women and the ninth most common cancer-related cause of death in wealthy countries (3.3/100000 women). [23].

Contrarily, among LMICs, cervical cancer ranked second in terms of cancer prevalence (15.7%/100000 women) and third in terms of cancer-related deaths (8.3%/100000) [23]. The most common cancer-specific cause of death for females in Africa and America is cervical cancer. Women in high-income nations had an 0.9% lifetime chance of acquiring cervix cancer & a 16% lifetime chance of passing away from the illness, compared to an 0.9% lifetime riskin LMICs (up to the age of 74 Nearly 50% of cases are identified well before age of 35 in theUSA, while 47 years old is the typical diagnostic age. [24].

Cervix cancer is the most frequent reason for cancer mortality for women in South Africa, accounting for more than 25% of cases between 2004 2012, between the ages of 40 and 49

[25]. Age-specific mortality increased over this time, with women older than 50 years old accounting for 70% of fatalities [25]. In an analysis of populations that looked at more than 70,000 cervical cancer fatalities over the course of seven years, older women were more likely to have the disease in an advanced stage. [26].

2.4 Impact of HIV Infection

Most HIV infections occur in Sub-Saharan Africa, where this percentage is close to 70%. In addition to having a higher risk of cervical cancer, HIV-positive women are more likely to contract HPV when they are young (13 to 18).

Cervical cancer in HIV-positive women is discovered earlier than in non-infected women (15– 49 years. In South Africa, between 2001 and 2009, there was an increase in the prevalence of cervical cancer. This increase may be related to the country's rising usage of anti-retroviral drugs, which helped people with HIV and AIDS survive longer in the country. Cervical cancer incidence remained stable despite the fact that extended inflammation s a risk factor for cancers linked to viruses, in contrast to other AIDS-defining disorders. Treatment for female patients with cervical cancer is more challenging when they have HIV. To give just a few examples, the interaction between the tumor and the Human immunodeficiency virus can cause T-cell dysfunction, raise the risk of leukopenia and the again activation of infections during systemic therapy, make staging difficult regarding the non-associated Adenopathy, and cause autoimmune thrombocytopenic purpura that could increase the risks of surgical and chemotherapeutic complications.

2.5 Clinical presentation

Early cervical cancer typically has no symptoms, so regular screenings or pelvic exams are the only ways to detect it. Post-coital or unusual vaginal bleeding is a sign of the disorder. On the other hand, a profuse, foul-smelling vaginal discharge is rarely seen in isolation as a symptom. Invasion of the rectum is demonstrated by passing fecal matter through the vagina, a sign of a Rectovaginal fistula, whereas invasion of the bladder is demonstrated by passing urine through the vagina, a sign of a Vesicovaginal fistula.

2.6 Genetic risk assessment using NGS and therapeutic effects

The last ten years have seen the development of NGS panels of genes linked to gynecologic malignancies as crucial tools in the counseling of patients and their families regarding the risk of cancer, risk-reduction strategies, and available treatments. Genetic testing for inherited BRCA2 and BRCA1 mutations are linked to an high risk of endometrial, ovarian, colorectal, and other carcinomas. Lynch syndrome-related genes (MLH1, MSH2, MSH6, andPMS2) are also linked to an increased risk of these cancers included in these panels, which are essential for gynecologic cancers.

With no false-positive genetic changes, such as nonsense and frame-shift mutations, Walsh et al. successfully identified single-nucleotide substitutions, deletion and insertion mutations, duplications, and deletions for tumor suppressor genes linked to breast and ovarian cancer [27].

The majority of cervical cancers are caused by the dangerous HPV (HR-HPV) subtypes 18 and 16, which are distinguished from other non-breast gynecologic malignancies by the existence of a recognized screening procedure and proven DNA biomarkers [28,29]. The FDA approved the Cobas HPV test in April 2014 for use in the initial cervical cancer monitoring ofwomen between the ages of 25 and 29.

2.7 KMT2 Family

Methylation of H3K4 is largely catalyzed by the six-membered KMT2 family (Set1a, Set1b, KMT2A, KMT2B, KMT2C, and KMT2D), sometimes referred to as the mixed lineage leukemia (MLL) family. All three complexes— KMT2A-KMT2B, Set1a-Set1b and KMT2C-KMT2D—contain the catalytically functioning component of the KMT2 family of proteins. These complexes perform H3K4 mono-, di-, and tri-methylation by di- and tri-methylating promoters.

Several biological processes that happen while a mammal develops are impacted by the KMT2 family. When both copies of KMT2A were targeted for deletion in mouse embryonic stem cells, the consequence was embryonic death. Early embryonic development was delayed and neural tube abnormalities were produced by KMT2B knockout in the germ line. Moreover, KMT2 proteins participate in several pathogenic processes, and their abnormal expression is linked to a number of illnesses, including malignancies [30].

2.8 KMT2B gene

The family of mammalian histone H3 lysine 4 (H3K4) methyltransferases includes the protein MLL2, often referred to as KMT2B.With 2715 amino acids, it is a sibling of the MLL1 protein and is a sizable protein that is extensively expressed in adult human tissues.Both MLL2/KMT2B and MLL1/KMT2A are MLL subgroups and has two paralogs. The result of genome duplication throughout mammalian evolution. Comparable proteins belonging to the Subgroup for Trx known as the MLX clans are MLL2 and KMT2B [31]. The MLL2 (KMT2B) gene, which is also referred to as OMIM 606834, occupies a 20 kb region.

The transcript for it is about 8.5 and 9 kb long. It is expressed in most human organs and is present on chromosome 11q23 [32]. The 2715 amino acid long MLL2 protein is structurally organized to contain a SET domain that is catalytically active, a CXXC domain, an AT-hook, many PHDs, and the histone H3 protein's N-terminal tail which is bound by the SET domain, which creates a pocket and the S-adenosylmethionine cofactor for methyltransferase, which catalyzes the methylation process [33].

The MLL2 protein exhibits additional structurally distinct features before the SET domain at the C-terminus that define its non-redundant function as well as the inherent biological properties and molecular activities. The zinc-finger (ZF)-CXXC domain is another name for the ZF-CXXC domain, is made up of two Zn ions and four cysteine residues (Cys4). In order for MLL2 to link with chromatin, it can identify unmethylated CpG DNA and bind to it. The CpG islands that are found in the majority of active promoters are identified by the MLL2 CXXC domain and MLL1, which acts as a navigation mechanism.

Contrary to SETD1A/B, which are stabilized at promoters by the protein CFP1, which is a component of a complex with that protein CXXC domain. The CXXC domain is absent in MLL3. PHD1 to PHD4 of the MLL2 protein each contain two zinc ions that coordinate a Cys4- His-Cys3 motif that facilitates binding to methylation histone H3 [34].

ZF-CXXC is one such domain of these PHD fingers. Every member of the MLL family has PHD fingertips but only the PHD3 of MLL2 is primarily in charge of binding to H3K4me3 tails. Despite this, all MLL family members have different interactional properties. Between PHD3 and PHD4, a bromodomain (BRD) promotes PHD3 activity as opposed to serving as

the typical reader of acetylated lysine [35]. A second PHD, an FY-rich N-terminal (FYRN), and an FY-rich C-terminal (FYRC) domain follow the BRD that allows the C- and N-terminal fragments to dimerize non-covalently after proteolytic cleavage.

2.9 MLL2 Role in Transcription Control

Numerous studies have demonstrated that MLL2 may produce narrow H3K4me3 peaks close to active gene promoters and coexist with bivalent genes of embryonic stem cells include H3K27me3 using chromatin immunoprecipitation (ChIP) and next-generation sequencing (NGS). A well-known in vitro transcription using chromatin templates technique has been used to confirm the MLL2-related H3K4me's stimulatory effect on transcription.

Particularly, the MLL2 COMPASS complex has been demonstrated to mediate H3K4me3 at bivalent genes, pointing to a critical role for MLL2 throughout development. Unlike MLL1, MLL2 has been shown to successfully establish H3K4me3 on bivalent in nature promoters in mouse ESCs (mESCs) genes required for stem cell differentiation [36, 37, 38, 39]. Bivalent genes are often only minimally expressed in mESCs and frequently carry both H3K4me3 and H3K27me3 marks at the promoter. However, during differentiation, they either become H3K27me3- or H3K4me3-marked and, thus, silenced or activated, respectively. With the use ofparticular antibodies that recognized a significant number of MLL2-binding sites were discovered in the C-terminal region of MLL2, which contains two distinct epitopes [40-42]. Using antibody CT1 and more C- terminal antibody CT2 and the ChIP-seq method, it is shown that 70% of these regions are promoters, 16% to intergenic regions, and 14% to genebodies. The same study also found that 39% of them had markers containing H3K27ac and p300, which are active enhancers.

2.10 Function of MLL2 Physiology of the Human

Due to the MLL2 gene's similarity to MLL1, it was initially discovered, and it was later discovered that human tissues expressed it extensively. It has been established that it is essential for both paralogs to bind Menin/LEDGF in order to carry out their normal functions [43]. It is important to understand that increased apoptosis, which results in embryonic death before E11.5, and early growth retardation have all been linked to Mll2 germ-line deletions [44]. The mesodermal markers Mox1 and Hoxb1 were retained, while the HoxB group genes were dysregulated by Mll2, which is linked to the development of germ cells, as well as the

enhance promoter regions of spermatogonial stem cells, have H3K4me3 markers [44].

Loss of spermatogenesis as a result of its ablating. MLL2 also mediates the global H3K4me3 in oocytes, and anovulation and death were brought on by deletions in MLL2 [45]. The decrease of global H3K4me2/3 and the increased transcription of p53 and apoptotic factors were also discovered. It was also demonstrated that MLL2 engages in epigenetic reprogramming during fertilization and is autonomously necessary for fertility. Mid-gestational Mll2 deletion did not, however, have the same effects on hematopoiesis and global H3K4 methylation as Mll1 deletion [46]. Therefore, the Facilitator of H3K4me3 abundance as well as the opposition to encroaching repression complexes are two possible mechanisms by which MLL2 maintains the target genes' expression. The fact that H3K4 hypomethylated gene expression levels remained relatively stable in Mll2/macrophages shows that some genes are more susceptible to H3K4me3 promoter reduction than others. MLL2 also controls cell growth by influencing the MYC oncogene's activity.

2.11 Disease Implication of MLL2

MLL family members play a critical role in human health through controlling transcription, and genetic variations in the KMT2B gene have been found in a number of diseases. MLL2 has been shown to have a significant physiological function under volitional movement.

Particularly, A link between MLL2 haploinsufficiency and early onset-generalized pediatric dystonia, the most hyperkinetic movement condition that is severely characterized by twisting poses brought on by continuous or irregular jerks muscles, both agonist and antagonist [47,48,49]. The patients had typical brain magnetic resonance imaging results, heterozygous MLL2 gene alterations, and typical facial features [49]. It is likely that the patients would experience laryngeal and cranial dystonia over time. Reduced levels of several dystonia-related proteins, including TOR1A, THAP1, as well as D2 dopamine receptors (D2R) were seen in cerebral MLL2 mutant individuals' fluid and fibroblasts [47]. These data point to a function for MLL2 in the disease's pathophysiology, which needs to be further investigated. Additionally learning deficits were seen in adult mice with conditional MLL2 deletion due to the hyperactive forebrain neurons' increased activation. List genes involved in H3K4me2/3-mediated hippocampal plasticity histone modification. Regarding cell proliferation augmentation and carcinogenesis, MLL2 has also been linked to the promotion of illness [50].

Somatic MLL1 mutations were initially linked to the development of cancer. Due to the numerous chromosomal double-strand DNA breaks, which are not always repaired by growing hematopoietic cells, the gene MLL1 shows a significant number of rearrangements with a number of different translocation partner genes [50]. Exons 8–13, which make up the KMT2A-fusion proteins are encoded by the C-terminal region of the fusion protein and an arbitrary number of the fused partner exons that make up the N-terminal. There is a loss in H3K 4 methyltransferase activity. When MLL1 is translocated to its fusion partners because the SET domain is removed. More than 135 MLL1 rearrangements have already been identified, the bulk of which are in-frame translocations that produce abundance of function cancer-causing proteins with altered activity [51]. The SET and PHD domains are most affected by nonsense, missense, or frameshift mutations of MLL2, which are more frequently seen in malignancies. In stomach cancer, oesophagealsarcomatoid carcinoma, and UCEC, mutation rates are greater.

Additionally, MLL2 somatic mutations have been linked to the beginning of gliomagenesis and have been seen in neurofibromatosis 1-glioblastoma [52]. Other translocations in glioblastomas and MLL2 overexpression in cancerous cells of the pancreas also discovered. In colorectal cancer, physical interaction between MLL2 and -catenin has been found to boost cell proliferation, attracting MLL2 to the c-MYC enhancer site and inducing transcription. The study of MLL2 target genes in both wild-type and MLL2 null human colon cancer cells showed that MLL2 increases the transcription of retinoic acid-responsive genes like ASB2, which were before turned on in leukemia cells (HCT116 cells).

The discovery that MLL2 also regulated p53 and NR3C1 explains the probable mechanical role of MLL2 in the genesis of cancer [53]. Furthermore, MLL2 has been discovered in genomics to be a frequent target for the assimilation of hepatocellular carcinoma (HCC) tissues by oncogenic viruses, such as the hepatitis B virus and the adeno-associated virus 2. This finding raises the possibility of a link between increased MLL2 expression and the development of liver cancer, which warrants further study.

Additionally, MLL2 mutations were commonly found in follicular lymphoma (FL) at rates comparable to the t translocation, the disease's genetic signature, showing that MLL2 plays a crucial role in carcinogenesis.

Last but not least, a 17.9% mutation rate for somatic MLL2 mutations was often seen in SCCHN [54]. Due to the inactivating nature of these mutations, which may have an effect on

the expression of numerous gene sets, it is predicted that MLL2 functions as a tumor suppressor in neck and head cancer.

2.12 KMT2B mutations in human cancers

The analysis of the data that is now available indicates that human malignancies have fewer KMT2B distinct mutations. 72% of all known mutations in the KMT2B coding area are found in the liver, large intestine, lung, and carcinomas of the lung and glioma. Around 90% of KMT2B variants with known zygosity are heterozygous, whereas the remaining 10% are homozygous. Research on mice with germline or conditional Kmt2b haploinsufficiency has not demonstrated an oncogenic tendency. According to several reports, KMT2B is a generally beneficial regulator of cell development. In ESCs and in germline knockout mice models, homozygous inactivation of Kmt2b led to abnormalities in proliferation and embryonic mortality as a result of elevated apoptosis. In a manner similar to KMT2A, KMT2B is drawn to the MYC enhancer by a catenin-dependent process. This allows H3K4me3 methylation, which consequently encourages MYC transcription.

2.13 microRNA profiling in cervical squamous cell cancer at an early stage

Many biological processes, such as cell cycle control, differentiation, development, proliferation, metabolism, and apoptosis, are regulated by miRNAs in essential ways [55, 56]. Many diseases, including cancer, autoimmune diseases, schizophrenia, and cardiac abnormalities have been linked to changes in miRNA expression [57]. Nearly all forms of examined benign and malignant tumors, as well as non-tumor tissues, including CSCC, have been found to exhibit varying miRNA expression [58]. Using a miRNA microarray, Wilting et al. discovered that the expression of 46 miRNAs significantly varied between the healthy cervical squamous epithelium and CSCC [59]. Cervical cancer cell proliferation was found to be increased by miR-19a and miR-19b, while miR-125b reduced cervical cancer cell death [60]. Both of these pathways are involved in cervical carcinogenesis.

There haven't been any studies done yet to investigate the miRNA profile in the initial stages of CSCC. Therefore, analyzing the miRNA profile in earlier-stage CSCC using NGS may make it simpler to uncover novel miRNAs, find possible markers for treating and diagnosing CSCC at an early stage, and analyze the expression of all annotated miRNAs. By comparing

early-stage CSCC samples with paired normal samples, the NGS was used to describe miRNA changes broadly and systematically. They also examined miRNA target genes, and performed pathway analysis and GO annotation, as well as qRT-PCR validation of these miRNAs. In CSCC compared to healthy tissue, they discovered 37 known miRNAs that were significantly differentially regulated, and they validated 9 samples using qRT-PCR. Deep sequencing was used by Juan et al. [61] to find potential new miRNAs in the blood samples of healthy controls and cervical cancer patients. Only 1 of the 17 popular new miRNA candidates, they discovered, may be able to differentiate CC sufferers from wholesome controls.

Wang et al. produced profiles of miRNA expression in the blood samples of CSCC patients using a microarray as well. According to their findings, 291 of the 338 circulating miRNAs that were discovered in samples of serum from CSCC patients had levels that were >2-fold different from controls. They opt for Next generation sequencing as a more precise and targeted method to evaluate miRNA expression since it allows for rapid analysis and novel miRNA discoveries. In the current study, 9 miRNAs that displayed different levels of expression in CSCC were found using extensive sequencing & qRT-PCR analysis. Six genes, including miR-204-5p/3p, miR-211-5p, miR-202-5p and miR-218-1-3p/5p, were downregulated, whereas three genes—miR-21-3p, miR-34c-5p, and miR-34b-5p—were increased. These miRNAs include the previously identified miR-21, miR-218, and miR-34b-5p, whose functions are associated with cervical cancer.

Their findings were supported by the discovery that MiR-218 expression was lower in cervical cancerous tissues than in healthy cervical tissues [62]. According to their sequencing findings, In CSCC, miR-21 levels were raised and may act as an cancer-causing gene, in agreement with other studies. In CSCC, miR34b-5p was reported to be upregulated. NGS technology enables the quick discovery of new miRNAs, among them whose function in biology is unknown. They found that the expression of 5 common new miRNA candidates differed between CSCC and healthy tissues. Their team identified novel miR 7 as a putative new miRNA, and its existence was confirmed. According to functional annotation, its potential target genes belong to the MAPK signaling pathway, which is related to cell growth, cell death, and the equilibrium between them.

The association between MAPK activation and apoptosis in the development of the disease was shown by Engelbrecht et al. at the beginning of 2006, and they hypothesized that MAPK

is implicated in cervical cancer [63]. By inhibiting MAPK signaling, the human papillomavirus 16 E2 protein may induce apoptosis in CSCC [64]. Yet, because MAPK plays a part in cervical cancer, its expression level act as a marker for the development of the disease, and its inhibitors have demonstrated the potential to have a positive impact on the management of cervical cancer [65, 66]. Hence, their findings offer a strong foundation for describing the role of new miR 7 in the development, progression, and therapy of CSCC. Therefore, using high-throughput NGS, they essentially found both known and novel miRNAs in CSCC at an early stage. Furthermore, they used qRT-PCR on these miRNAs which should be verified in other separate samples. Their examination of differentially expressed miRNAs not only confirms earlier findings but also identifies additional differentially expressed miRNAs that had not previously been reported in CSCC, providing fresh molecular fingerprints of the disease. A potential diagnostic for CSCC might exist, according to new miR 7's prediction. Such findings offer new directions for CSCC molecular mechanistic study and have implications for understanding the regulation of a network of miRNAs. They also have implications for upcoming studies on the prognosis assessment, targeted treatment, and early diagnosis of CSCC patients.

2.14 Transcriptomic evaluation of miR-214 expression of genes in cervix cancer cells

Besides cancer, CRISPR has currently used as a genome-editing method in other diseases. RNA-sequencing (RNA-seq) has become a potent method for studying gene expression profiles through transcriptome analysis. Transcriptome analysis gives greater sensitivity to find unusual sequences and excellent single nucleotide resolution, allowing the distinction between closely related sequences. Moreover, RNA-seq offers a special benefit for assessing RNA expression levels quantitatively. They conducted a transcriptome analysis in their work to evaluation of the degree of gene expression various genes following the elevation and miR-214is downregulated in cervix cancer cells. The 2-ΔΔCT approach was used to compute fold changes for the pertinent genes that had shown substantial alterations. For qRT-PCR, the primers used were (β-actin Forward: 5′-TCACCCACACTGTGCCCATCTACGA-3′; Reverse: 5′-CAGCGGAACCGCTCATTGCCAATGG-3′ and miR-214 Forward: 5′-TGCGGACAG CAGGCACAGAC-3′; and Reverse: 5′-CCAGTGCAGGGTCCGAGGT-3′).

Using NGS, it was determined how miR-214 affected the networks that control gene expression in cervical cancer cells. In cervical cancer cells, the analysis is the first to describe how miR-214 affects the whole coding transcriptome. MiR-214 is often

downregulated in malignancies, including cervical carcinoma. Moreover, it has been shown that it is elevated in certain different malignancies. MiR-214 is often downregulated in malignancies, including cervical carcinoma. Moreover, it has been discovered that it is elevated in a few additional tumors [67]. As a result, in our investigation, they either overexpressed miR-214 or CRISPR-mediated deleted it. NGS was used to do the RNA sequencing, and the results showed that overall, 108 genes had increased expression, while 178 genes had decreased expression between each sample. In comparison to C33A cells that had not been transfected, miR-214 overexpression resulted in 103 genes which shows downregulation whereas 50 genes show upregulation. Similar with this, 58 genes were elevated and 75 genes were downregulated in the case of miR-214 knockdown compared to normal control. 10 genes that showed notable alterations and had a key role in cancer were chosen out of a total of 286 genes. The levels of expression were verified by Rt-PCR. The genes involved were IFI27, COX11, TP53INP1, NRG1, TNFAIP3, FGF8, MDM4, and SP3.

2.14.1 Expression level of genes

A number of malignancies, including ovarian cancer, have been recognized as being capable of producing the Interferon-inducible protein 27 (IFI27), which is an outcome of Interferon. IFI27 deletion in oral cancer decreased cell proliferation and invasion [68]. It was found that COX11 is upregulated by miR-214, but COX11 is decreased when miR-214 is deleted. Higher levels of COX11 in cancer cells may increase the cytochrome c oxidase's metabolism, which is necessary to sustain the growth of tumor cells. Neuregulin 1 (NRG1) regulates NRF2 in thyroid carcinoma to maintain redox equilibrium. [69].

Although miR-214 deletion did not significantly alter NRG1 levels, miR-214 overexpression caused a two-fold rise in NRG1. Since MiR-214 is frequently decreased in cervical cancer, this may result in an increase in NRG1 levels and cell proliferation. They found that miR-214 overexpression slightly increased the level of SMAD3 whereas miR-214 deletion dramatically decreased it. SMAD3 expressions are consequently positively regulated by miR-214. SP3 is frequently elevated in cancer cells. When miR-214 was overexpressed, SP3 expression was drastically reduced, but in mutant cells, the expression had recovered to levels that were up to 50% higher than in ordinary cells. This demonstrates that despite SP3 being linked to cancer, the tumor suppressor miR-214 could decrease SP3 expression in genes.

They found that whereas miR-214 overexpression only slightly decreased MDM4 levels, miR-214 deletion led to an elevation of MDM4 that was significantly more than that of normal cells by >5 fold. For maintaining normal levels of MDM4 as a result, miR-214 is essential.

TP53INP1 expression was not significantly affected by miR-214 overexpression, while it was significantly increased by miR-214 deletion. They noticed that when miR-214 was overexpressed, the levels of ABL2 were unaffected, whereas miR-214 was knocked out, and ABL2 was significantly elevated. The carcinogen ABL2 and the tumor suppressor miR-214 have a direct connection. A-20 is another name for TNF-inducible protein 3 (TNFAIP3), which mediates the various effects of tumor necrosis factor in cancer cells. TNF-α promotes apoptosis in other breast cancer cells when A20 is overexpressed, whereas triple-negative breast cancer has an aggressive character [70]. Hence, TNFAIP3's actions differ depending on the kind of cell. They found that both miR-214 overexpression and miR-214 deletion result in a considerable downregulation of TNFAIP3 expression. This suggests that miR-214 may exert both positive as well as negative control over TNFAIP3 levels. The development and proliferation of cells are stimulated by FGF8 (fibroblast growth factor 8), which is increased in several malignancies. FGF8 causes resistance to epidermal growth factor receptor (EGFR) inhibitors in hepatocellular cancer [71]. The levels of FGF8 expression did not appear to be significantly impacted by miR-214 in our investigations, though.

CHAPTER -3

MATERIALS AND METHODS

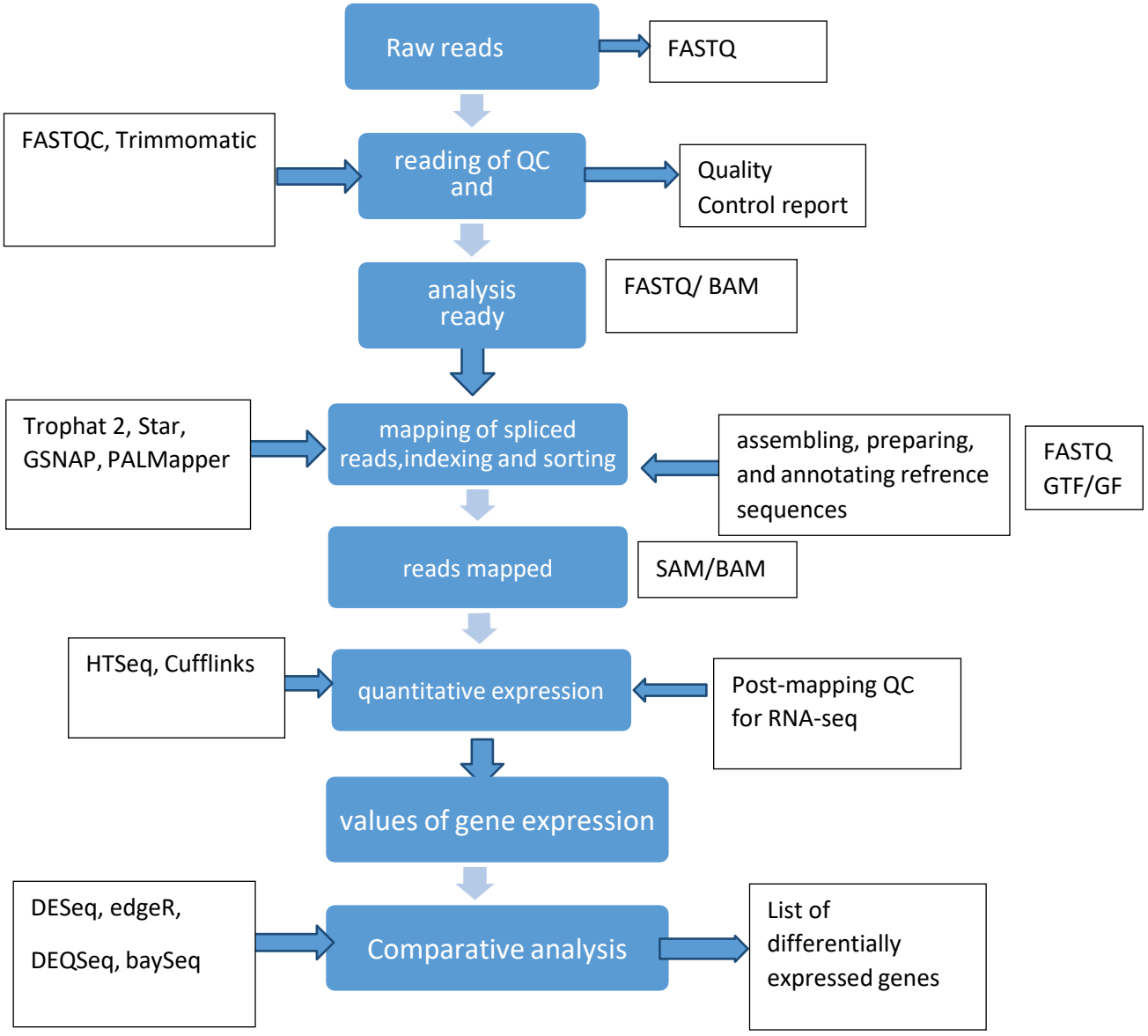

Fig.1 RNAseq Pipeline for DE

1.DATA COLLECTION

"RNA-seq analysis of HeLa cells overexpressing histone methyltransferase KMT2B" data was obtained from NCBI (https://www.ncbi.nlm.nih.gov/). Used the available transcriptomic data on NCBI [**Accession: PRJNA862962**]. Our data had 4 replicates and each replicate have its control.

SRA tool kit (sratoolkit.current-centos_linux64.tar.gz)

In order to create new runs and access those that have already been created, the NCBI SRA SDK provides loading and dumping tools with their corresponding libraries. The NCBI has specified the Sequence Read Archive that is SRA format for NGS data. Every piece of information sent to NCBI which should be in SRA format. The SRA Toolkit provides tools for downloading data, transforming other data formats into SRA format and vice versa, and extracting SRA data from SRA files in other formats.[72].

In our analysis, we dumped all 8 SRA experiments (SRR20677775, SRR20677776, SRR20677777, SRR20677778, SRR20677779, SRR20677780, SRR20677781, and SRR20677782).

1.PRE-PROCESSING OF DATA

Quality check and trimming

FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.12.1.zip)

FastQC is a quality checking tool for massively parallel sequencing. FastQC intends to offer a quick and easy solution to perform some quality control tests on the raw sequence data generated by the high-throughput sequencing procedures. It provides a modular collection of analysis that we may use to quickly determine whether our data has any issues that we should be informed of before conducting any additional investigation. Sequence reads having a Phred score of Q30 were selected for additional examination. Reads have to be removed from the data depending on their length and base call quality (Phred score). It is necessary to eliminate low-confidence base calls since they can result in the discovery of false-positive variations.

We have done FastQC on our 8 samples to check out some quality-based parameters. FastQC generates a FastQC report summary which contains basic statistics, per sequence quality, per base quality score, per base N content, per base sequence content, sequence length distribution, sequence duplication levels, overrepresented sequences, and adapter content [73].

Trimmomatic

(https://github.com/usadellab/Trimmomatic/archive/refs/tags/v0.39.tar.gz)

Trimmomatic carries out a number of beneficial trimming operations for Illumina paired-end and single-ended data. The selection of trimming steps and the associated parameters are specified via the command line. With the use of this tool adaptor sequences that were ligated to the ends of libraries during the library preparation procedure need to be taken out of the sequencing reads since they may obstruct mapping and assembly [74].

In our data, we were having two inputs that is input-1. Fastq (SRR/1) and input -2 fastq (SRR/2). SRR1/1 has paired and unpaired fastq data and similarly SRR/2 have paired and unpaired fastq data. Paired data of both inputs were matched and unpaired were not. Additionally, reads were trimmed to eliminate low-quality bases from their ends.

3. ALIGNMENT AND MAPPING

Bowtie(https://github.com/BenLangmead/bowtie2/releases/download/v2.5.1/bowtie2-2.5.1-linux-aarch64.zip)

When comparing sequencing reads to large reference sequences, Bowtie 2 is a memory- and time-effective tool which is especially effective at aligning reads that range in length from 50 to 100 or 1,000 characters, as well as moderately lengthy genomes (such as those of mammals). To minimize memory usage, Bowtie 2 uses an FM Index to index the genome; for the human genome, this results in a memory footprint of about 3.2 GB [75].

After trimming, **alignment** was done by **indexing** of reference genome (human GRch38.p14) and reads were mapped against the indexed genome to get the overall alignment rate.

Samtools(http://www.htslib.org/download**)**

Samtools is a collection of tools for working with alignments in the formats of SAM (Sequence Alignment/Map), BAM, and CRAM. It can extract reads quickly from any region and converts across formats while sorting, combining, and indexing. Samtools is made to function on a stream. An output file ending in '-' is regarded as standard output (stdout), while an input file ending in '-' is regarded as standard input (stdin). In our work, we have used samtools view for converting SAM file to BAM file and for sorting also [76].

4. TRANSCRIPT EXPRESSION ANALYSIS

Cufflink(http://cole-trapnell-lab.github.io/cufflinks/assets/downloads/cufflinks-2.2.1.tar.gz)

In RNA-Seq samples, Cufflinks assembles transcripts and calculates their abundances. It accepts RNA-Seq read alignments and compiles them into a compact set of transcripts. Based on the number of reads that support each of these transcripts, Cufflinks then calculates their relative abundances. We used this tool after using samtools on our 8 conditions to calculate their transcript abundances [77].

Trinity (https://github.com/trinityrnaseq/trinityrnaseq/releases**)**

Trinity allows for the effective and reliable de novo reconstruction of transcriptomes. We used trinityrnaseq-v2.14.0 version **for differential expression.**

5. SELECTION OF SIGNIFICANT DIFFERENTIALLY EXPRESSED GENES

Selection of significant differentially expressed genes has analysed in four comparisons (control vs. replicate) i.e SRR20677775 vs. SRR20677779 (C1 vs. R1), SRR20677776 vs. SRR20677780 (C2 vs. R2), SRR20677777 vs. SRR206777781 (C3 vs. R3), SRR20677778 vs. SRR20677782 (C4 vs. R4). In all samples log 2-fold change which is greater than 2 or less than -2 and <0.05 FDR value was set for the significant Differential expression genes from the total DEG's.

6. GO ANNOTATION

The Gene Ontology (GO) is an important project that aims to organise the illustration of gene and gene characteristics of products in all species. To be more precise, the project's objectives are to: 1) preserves and grows its regulated language of gene and gene product features; 2) annotate genes and its products as well as absorb and distribute annotated information and 3) Offer tools that make it simple to access all of the project's data and that allow for the operative perception of experimental results using the gene Ontology, for instance through enrichment investivation.

For GO annotation we have used GO net web-based programme (available at http://tools.dice- database.org/GOnet/),which performs GO word annotation analysis on a collection of gene orprotein entries obtained from human or mouse data.

For GO annotation figure analysis we used REVIGO, a web server called uses a straightforward clustering method that is based on semantic similarity measurements to identify a subset of the GO terms in lengthy, nonsensical lists. We have used a scatterplot of REVIGO for GO annotation.

7. KEGG PATHWAY ANALYSIS

For KEGG pathway analysis, we have used KEGG pathway database which is used to study the collection of manually drawn pathway maps which represents the molecular interactions, reactions and relations. We used this for human diseases for the analysis of endometrial cancer.

CHAPTER 4

RESULTS

1.Raw data results

This table shows the available transcriptomic raw data details which has taken from NCBI.

Table: 1. The SRA ID's and no. of reads of raw data.

Sr.no.	Condition	SRA ID	No. of reads
1.	KMT2B-overexpressing Hela cell 1	SRR20677775(1.9gb)	21,523,877
2.	KMT2B-overexpressing Hela cell 2	SRR20677776(2gb)	23,536,721
3.	KMT2B-overexpressing Hela cell 3	SRR20677777(2.1gb)	24,116,164
4.	KMT2B-overexpressing Hela cell 4	SRR20677778(2.3gb)	26,392,955
5.	Control HeLa cell 1	SRR20677779(2.1gb)	23,668,276
6.	Control HeLa cell 2	SRR20677780 (2.3gb)	26,215,913
7.	Control HeLa cell 3	SRR20677781(2.4gb)	27,046,596
8.	Control HeLa cell 4	SRR20677782(2.3gb)	26,630,730

1.FastQC ResultS

1. KMT2B -overexpressing HeLa cell 1(SRR20677775)

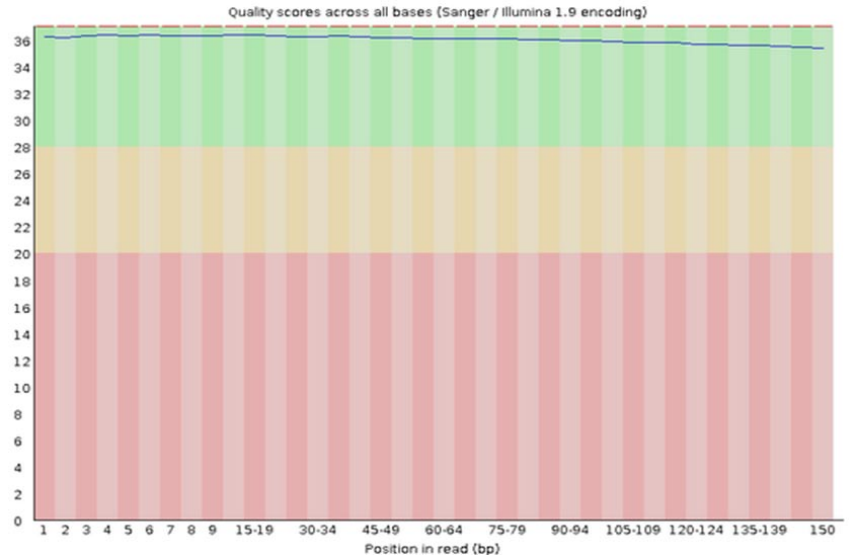

Fig.2.Represents Per base sequence quality. The y-axis on the graph represents quality/phred scores and x-axis represents position in read (bp). The background of the graph divides the y-axis into three parts. The green color represents good quality call, orange color represents call of reasonable quality, and call with poor quality represents with red color. This graph shows calls with good quality.

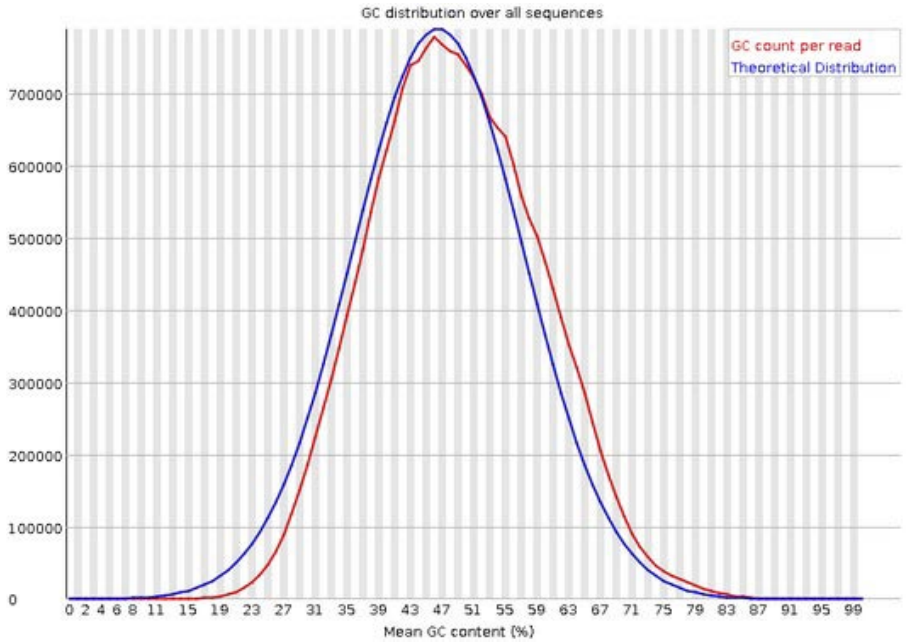

Fig.3 Represents Per sequence GC content. The x-axis represents mean GC content (%) and y-axis represents no. of sequences. Blue line represents theoretical distribution and red line represents GC count per read. This graph shows GC count per read corresponds with Theoretical distribution which shows less deviation between these two lines and shows 48% GC content.

2. KMT2B -overexpressing HeLa cell 1(SRR20677776)

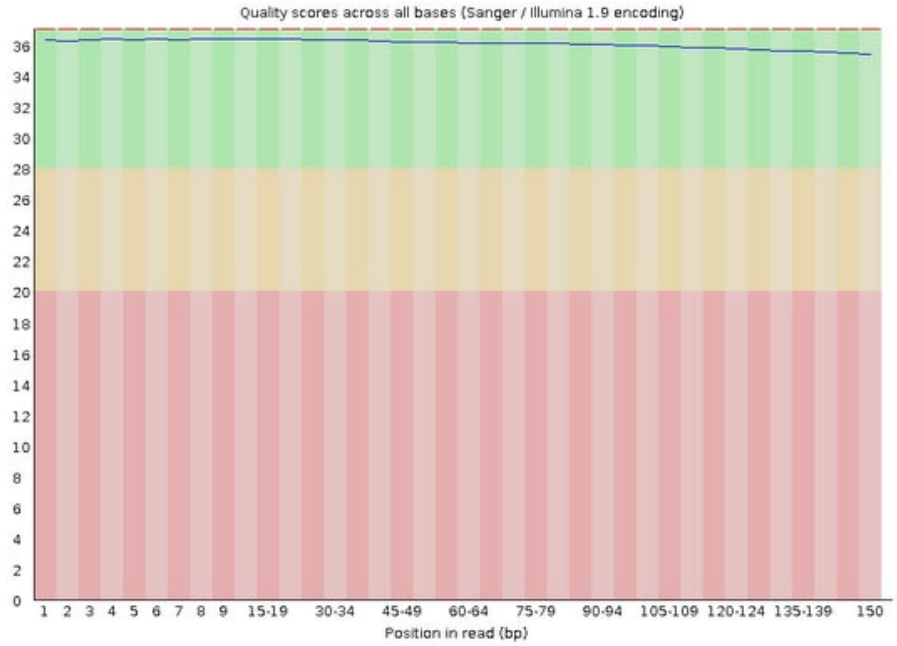

Fig.4 Represents Per base sequence quality. This graph shows good quality score as quality ofbase calls lying in green region.

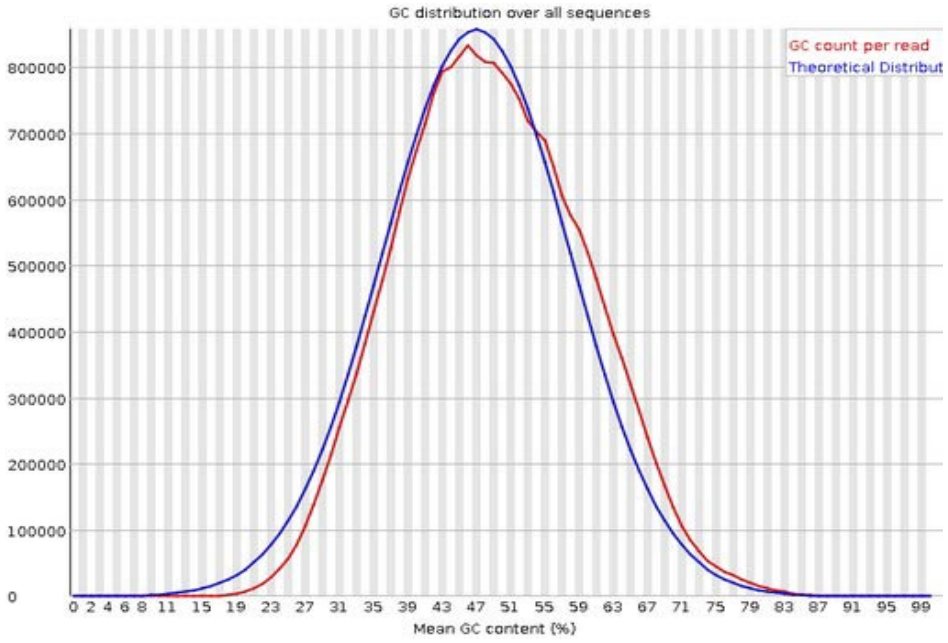

Fig.5. Represents Per sequence GC content. This graph shows GC count per read corresponds withTheoretical distribution which shows less deviation between these two lines and shows 48%GC content.

3. KMT2B -overexpressing HeLa cell 3(SRR20677777)

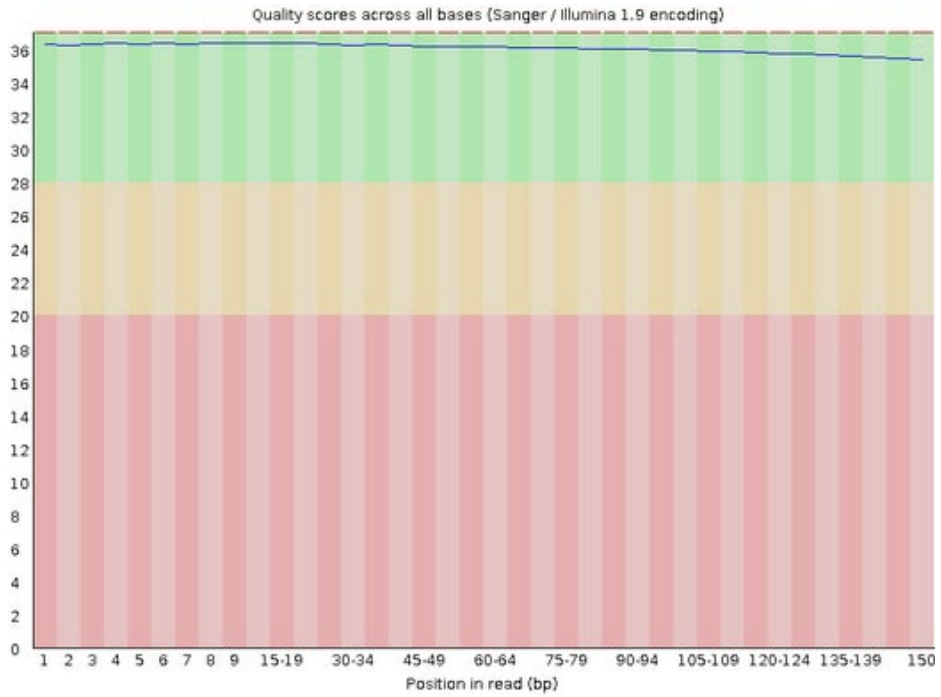

Fig.6. Represents Per base sequence quality. This graph shows good quality score as quality of base calls lying in green region.

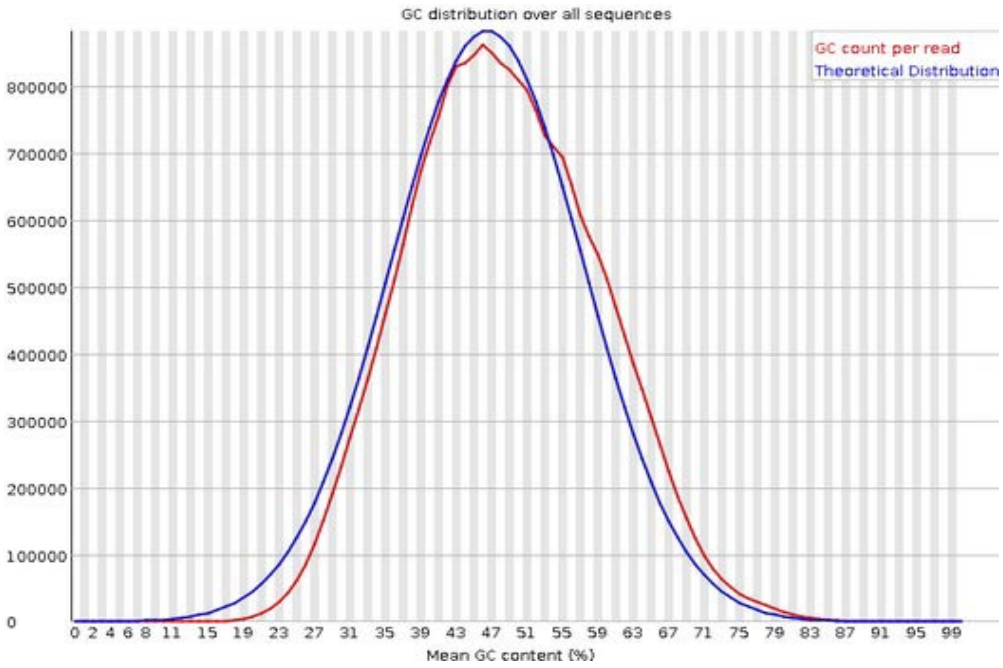

Fig.7 Represents Per sequence GC content. This graph shows GC count per read corresponds with Theoretical distribution which shows less deviation between these two lines and shows 48% GC content.

4. KMT2B -overexpressing HeLa cell 4(SRR20677778)

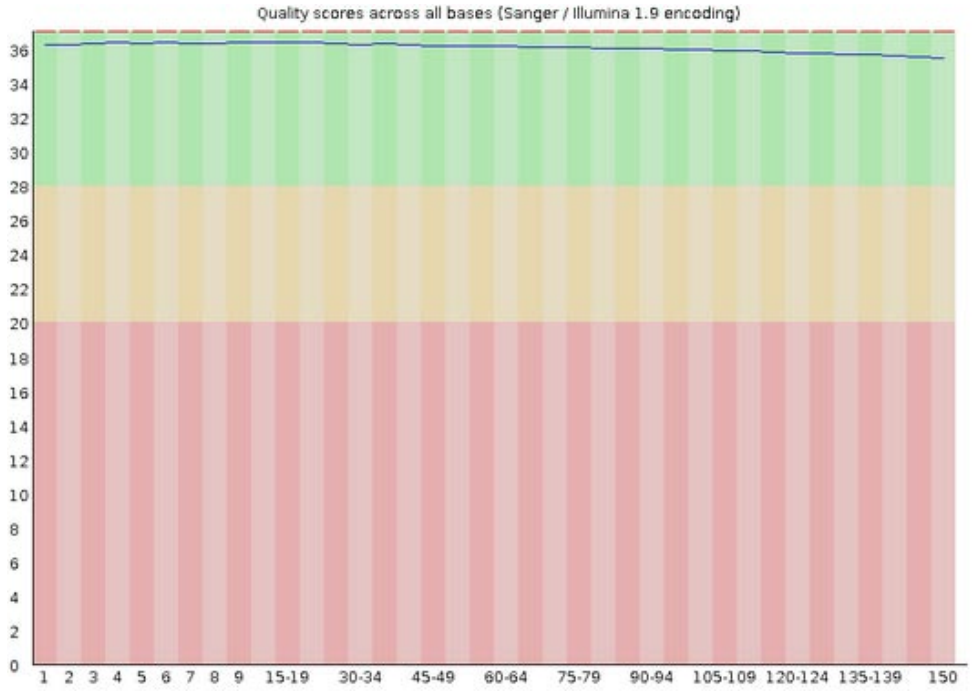

Fig.8. Represents Per base sequence quality. This graph shows good quality score as quality of base calls lying in green region.

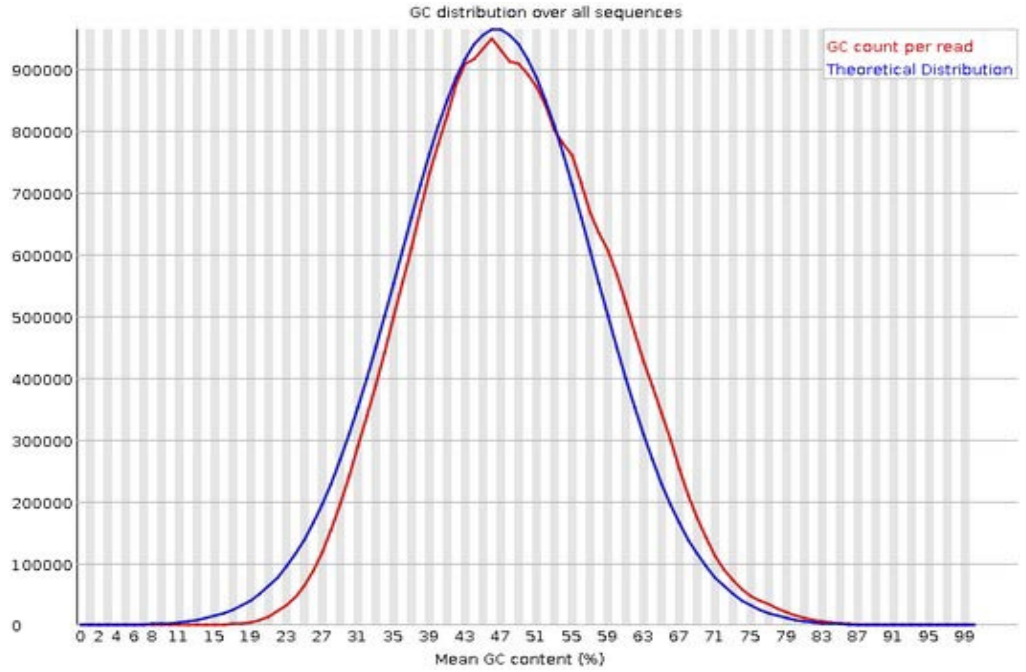

Fig.9. Represents Per sequence GC content. This graph shows GC count per read corresponds with Theoretical distribution which shows less deviation between these two lines and shows 48% GC content.

5. Control HeLa cell 1 (SRR20677779)

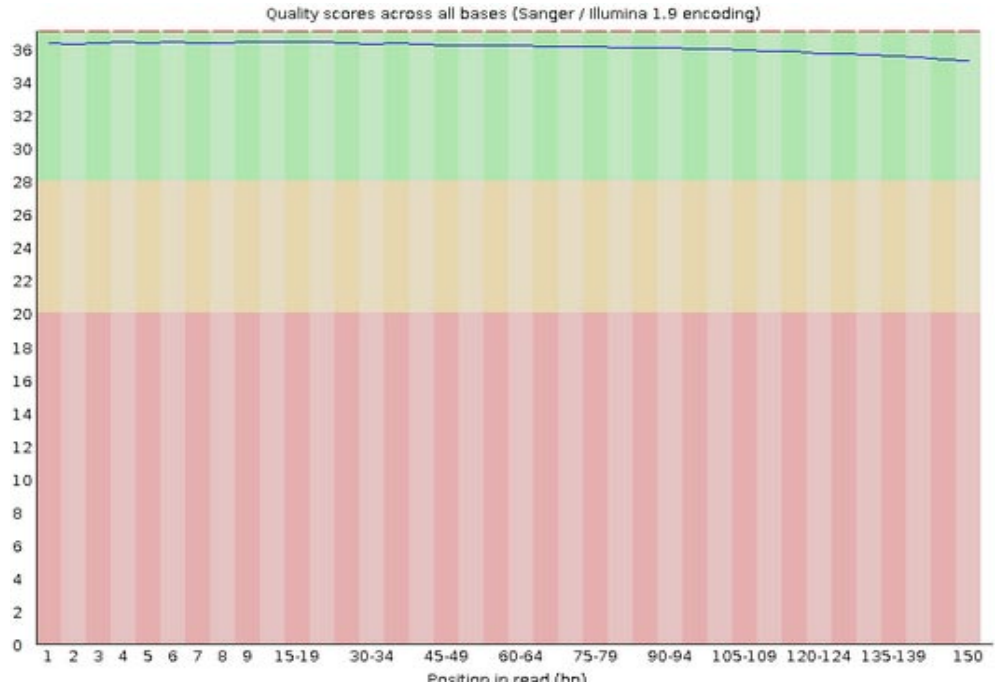

Fig.10. Represents Per base sequence quality. This graph shows good quality score as quality of base calls lying in green region.

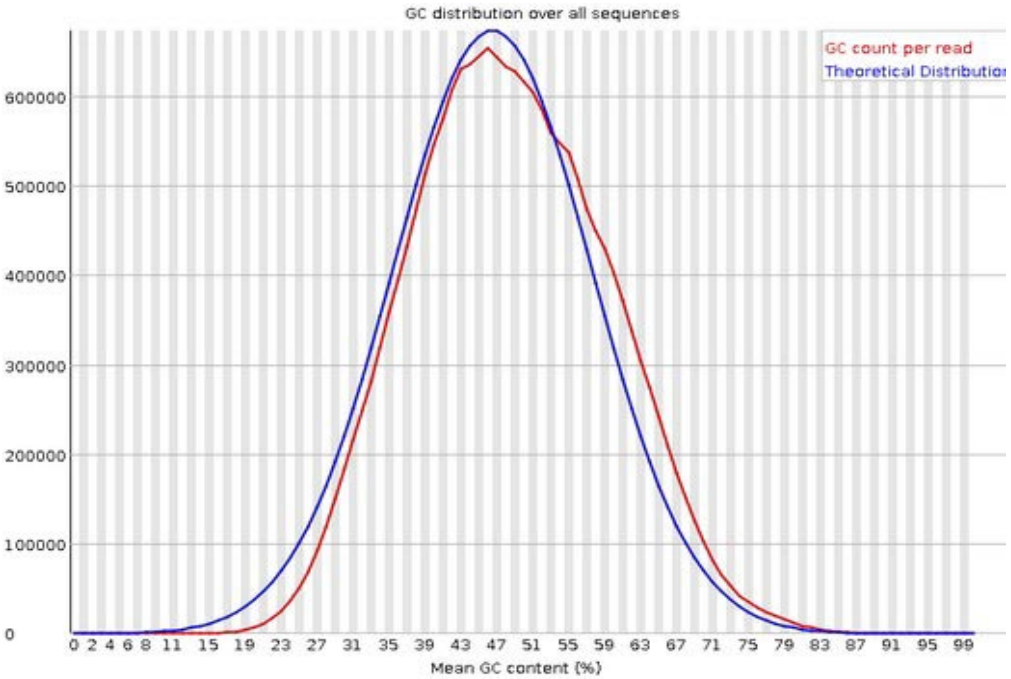

Fig.11 Represents Per sequence GC content. This graph shows GC count per read corresponds withTheoretical distribution which shows less deviation between these two lines and shows 48% GC content.

6. Control HeLa cell 2 (SRR20677780)

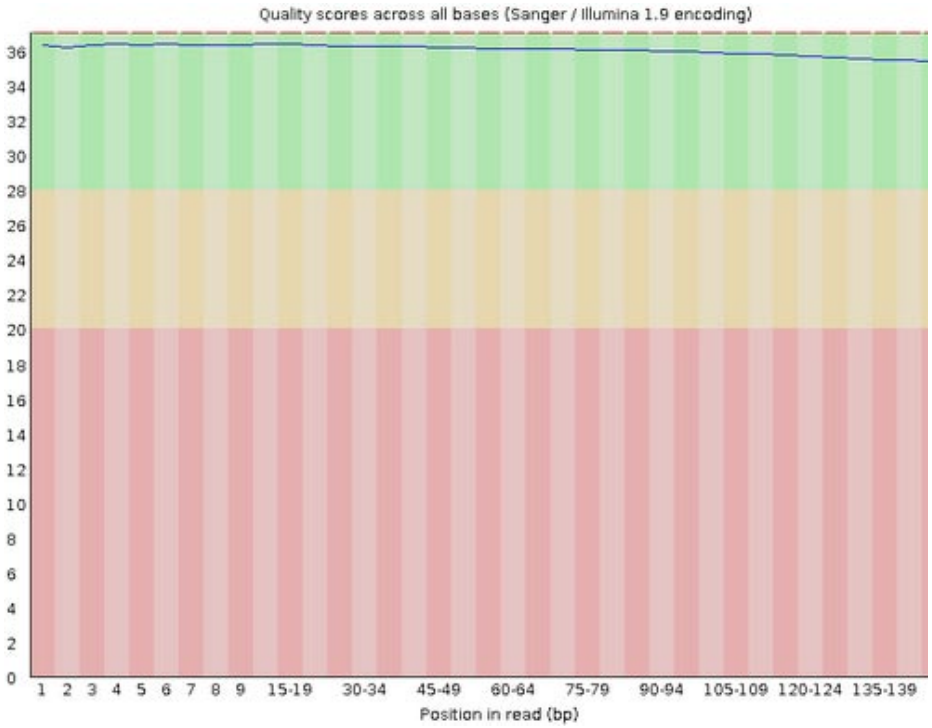

Fig.12 Represents Per base sequence quality. This graph shows good quality score as quality of base calls lying in green region.

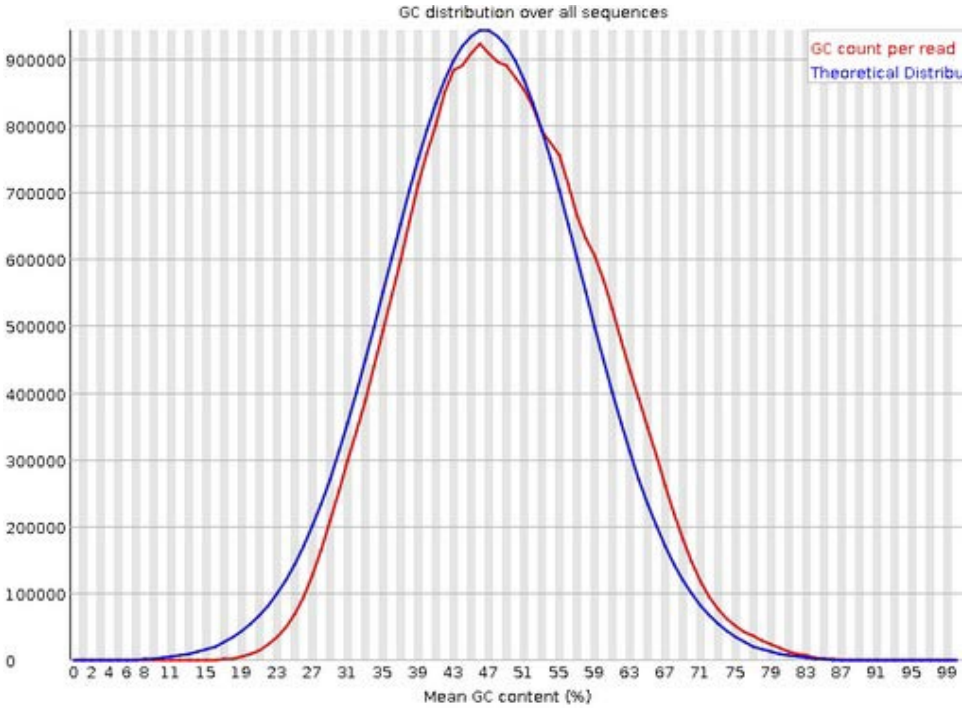

Fig.13 Represents Per sequence GC content. This graph shows GC count per read corresponds with Theoretical distribution which shows less deviation between these two lines and shows 48% GC content.

7. Control HeLa cell 3 (SRR20677781)

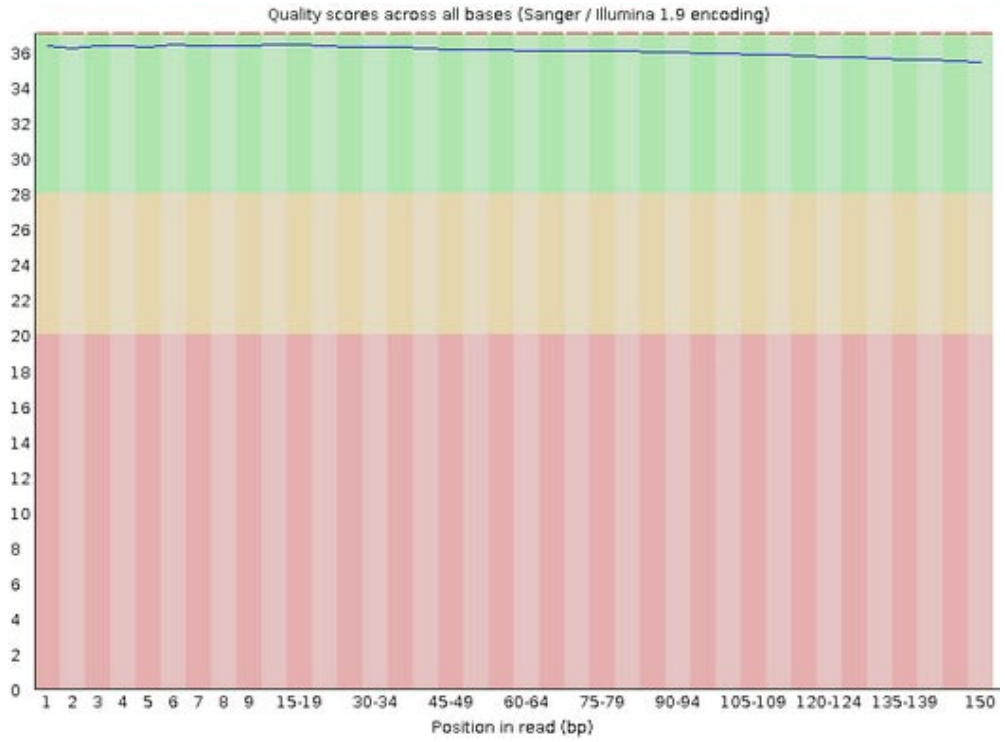

Fig.14. Represents Per base sequence quality. This graph shows good quality score as quality of base calls lying in green region.

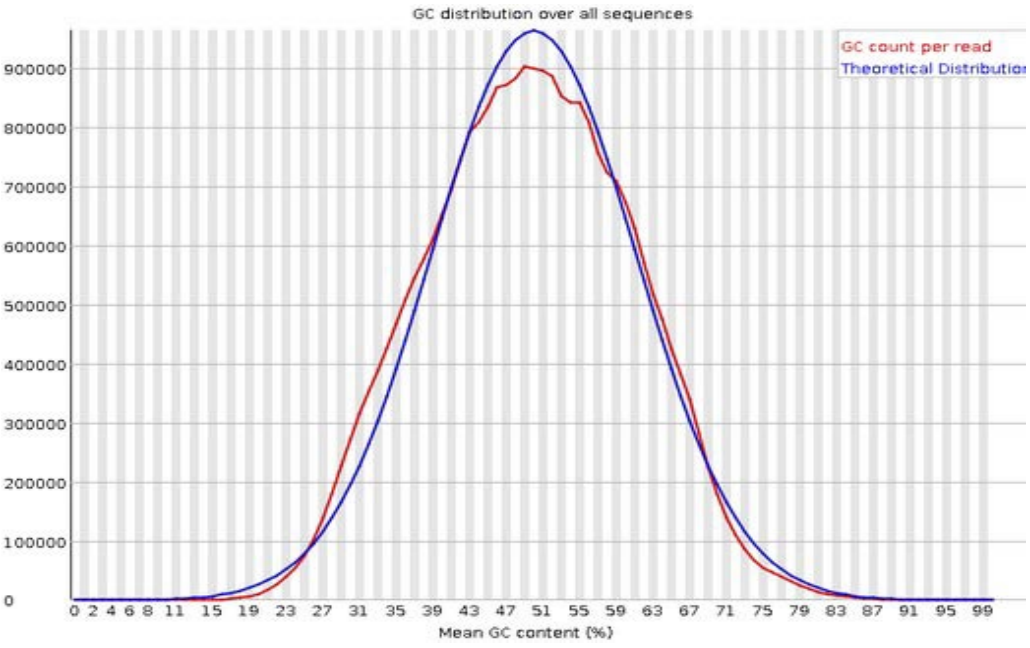

Fig.15. Represents Per sequence GC content. This graph shows GC count per read corresponds withTheoretical distribution which shows less deviation between these two lines and shows 48% GC content.

8. Control HeLa cell 4 (SRR20677782)

Fig.16. Represents Per base sequence quality. This graph shows good quality score as quality of base calls lying in green region.

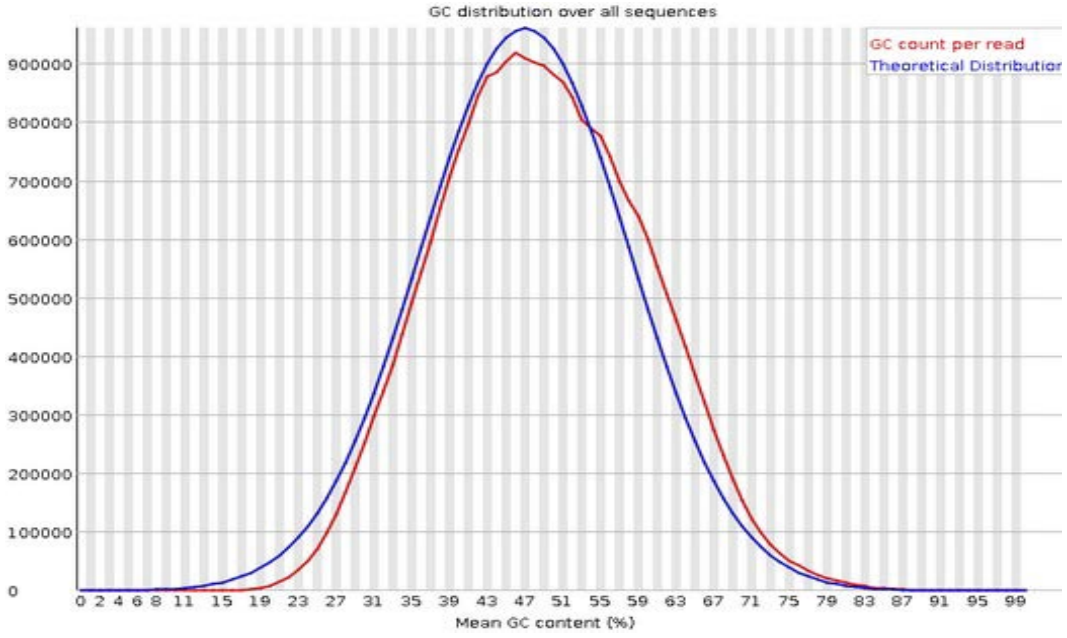

Fig.17. Represents Per sequence GC content. This graph shows GC count per read corresponds withTheoretical distribution which shows less deviation between these two lines and shows 48% GC content.

Table 2. Represents the Base sequence quality, Over-represented sequences and GC% content of all conditions. All conditions have good quality score as their calls are lying on green region of graph. For over-represented sequences it is shown that there is no over-represented sequences in all conditions. GC% content is 48% for all conditions but except for control heLa cell 3 is 49%.

Condition	**Base sequence quality**	**Over-represented sequences**	**GC% content**
KMT2B - overexpressing HeLa cell 1	PASS	PASS	48%
KMT2B - overexpressing HeLa cell 2	PASS	PASS	48%
KMT2B - overexpressing HeLa cell 3	PASS	PASS	48%
KMT2B - overexpressing HeLa cell 4	PASS	PASS	48%
Control heLa cell 1	PASS	PASS	48%
Control heLa cell 2	PASS	PASS	48%
Control heLa cell 3	PASS	PASS	49%
Control heLa cell 4	PASS	PASS	48%

3. Clean Data Results

These results are obtained after trimming the all data by removing the adapters with the help of trimmomatic tool.

Table 3. Shows total sequences of all conditions

Condition	Total sequence
KMT2B -overexpressing HeLa cell 1	21523877
KMT2B -overexpressing HeLa cell 2	23536721
KMT2B -overexpressing HeLa cell 3	24116164
KMT2B -overexpressing HeLa cell 4	23392955
Control heLa cell 1	18634381
Control heLa cell 2	26215913
Control heLa cell 3	27046596
Control heLa cell 4	26630730

4. Alignment Results

When we got trimmed reads after trimming, we aligned our reads of all conditions to reference genome to get overall alignment rate. This results shows that how much of our data matches with our reference human genome (GRCh38.p14).

Table 4. Represents the overall alignment rate of all condition after mapping

Condition	Overall Alignment rate
KMT2B -overexpressing HeLa cell 1	73.25%
KMT2B -overexpressing HeLa cell 2	74.61%
KMT2B -overexpressing HeLa cell 3	72.66%
KMT2B -overexpressing HeLa cell 4	74.04%
Control HeLa cell 1	72.94%
Control HeLa cell 2	71.99%
Control HeLa cell 3	77.96%
Control HeLa cell 4	70.85%

5. Differentially Expressed Genes Results

We found 72,157 total DEG's, out of them 4,005 are significant DEG's. From total significant DEG's 1,747 are upregulated genes and 2,257 are downregulated DEG's. We selected these genes on the basis of logFC value which is greater than 2 or less than -2 and FDR value which is <0.05.

6. VENNY Results

These results shows the number and percentage (%) of similar genes which are present in our 4 samples i.e control vs. replicate (SRR20677779 vs. SRR20677775, SRR20677780 vs. SRR20677776, SRR20677781 vs. SRR20677777, SRR20677782 vs. SRR20677778).

Sample 1 (SRR20677779 vs. SRR20677775)

Sample 2 (SRR20677780 vs. SRR20677776)

Sample 3 (SRR20677781 vs. SRR20677777)

Sample 4 (SRR20677782 vs. SRR20677778)

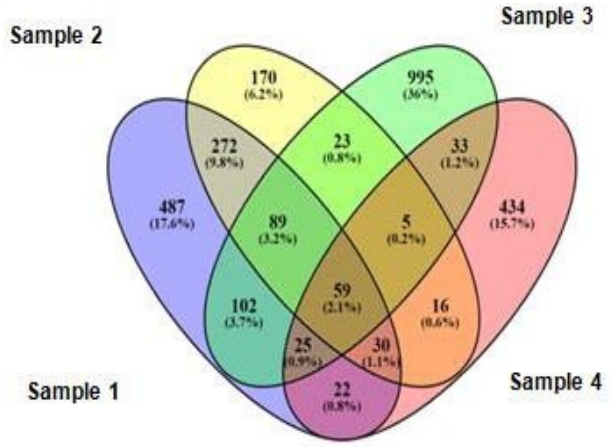

Fig. 18. Represents similar genes in all samples

7.GO Annotation Results

Sample 1(SRR20677779 vs. SRR20677775)

Table 5a. Top 5 biological process in which maximum genes are involved

Biological process	**No. of genes**
Anatomical structure development	232
Signal transduction	201
Cell differentiation	161
Transport	151
Response to stress	115

Table 5b. Top 5 Molecular function in which maximum genes are involved

Molecular function	**No. of genes**
ion binding	223
enzyme binding	80
DNA-binding	74
Binding of transcription factor activity	63
transmembrane transporter action	47

Table 5c. Top 5 Cellular component in which maximum genes are involved

Cellular component	**No. of genes**
Cell	539
Intracellular	450
Organelle	408
Cytoplasm	389
Plasma membrane	244

Sample 2 (SRR20677780 vs. SRR20677776)

Table 5d. Top 5 Biological process in which maximum genes are involved

Biological process	**No. of genes**
anatomical structure development	137
signal transduction	125
Cell differentiation	100
Transport	85
Response to stress	73

Table 5e. Top 5 Molecular function in which maximum genes are involved

Molecular Function	**No. of genes**
ion binding	128
enzyme binding	44
DNA binding	40
Binding of DNA transcription factor activity	35
transmembrane transporter action	30

Table 5f. Top 5 Cellular component in which maximum genes are involved

Cellular component	**No. of genes**
Cell	306
Intracellular	247
Organelle	225
Cytoplasm	210
Plasma membrane	157

Sample 3(SRR20677781 vs. SRR20677777)

Table 5g. Top 5 Biological process in which maximum genes are involved

Biological process	**No. of genes**
Anatomical structure development	149
Signal transduction	148
Cell differentiation	111
Transport	104
cellular protein modification process	89

Table 5h. Top 5 Molecular function in which maximum genes are involved

Molecular Function	**No. of genes**
ion binding	182
DNA binding	71
DNA binding transcription factor activity	59
Enzyme binding	53
Cytoskeletal protein binding	30

Table 5i. Top 5 Cellular component in which maximum genes are involved

Cellular component	**No. of genes**
Cell	427
Intracellular	361
Organelle	330
Cytoplasm	284
Nucleus	183

Sample 4(SRR20677782 vs. SRR20677778)

Table 5j. Top 5 Biological process in which maximum genes are involved

Biological process	**No. of genes**
Signal transduction	196
Immune system process	189
Transport	164
Response to stress	133
Anatomical structure develop	126

Table 5k. Top 5 Molecular function in which maximum genes are involved

Molecular Function	**No. of genes**
ion binding	138
enzyme binding	61
Enzyme regulator activity	36
Lipid binding	34
Cytoskeleton protein binding	31

Table 5l. Top 5 Cellular component in which maximum genes are involved

Cellular component	**No. of genes**
Cell	378
Intracellular	308
Cytoplasm	287
Organelle	281
Plasma membrane	222

By using **Revigo biotool**, where the input was GO_term_ID (eg. GO: 0005623, GO:0005622, GO:0043226) with the default parameters, this input gave GO annotation results as an output in three different typesBiological process, Molecular function, Cellular component This biotool was run for 4 different samples. The output was obtained in the form of scatterplot where the different sizes of black dots are seen, whereas these black dots represents the different genes of sample. Where it is seen that the different size of black dots are on the basis of log size. Each sample had different logsizes. It has been analysed the majorly involved genes have the larger log size those are represented in following figures.

Sample 1

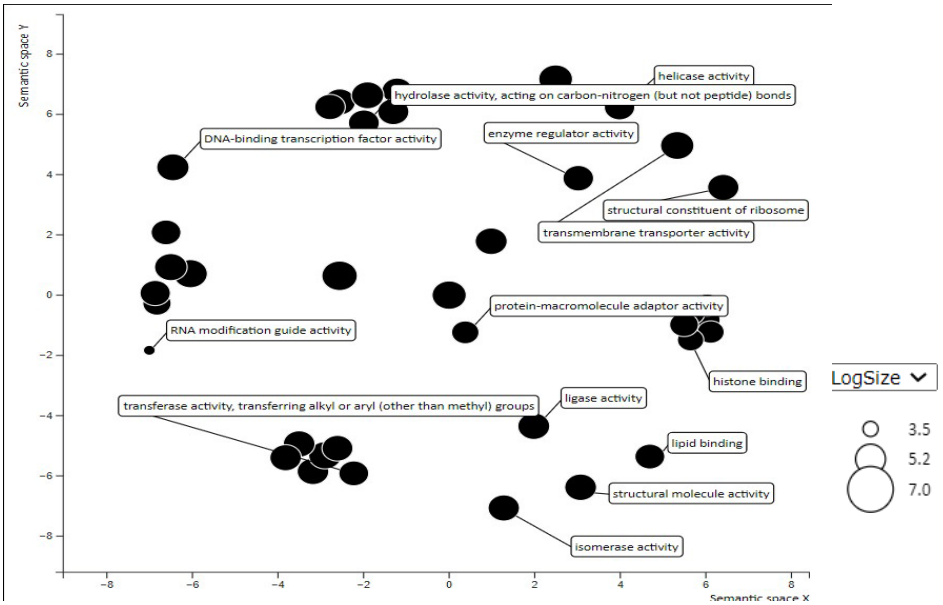

Fig. 19. Represents Biological process

.

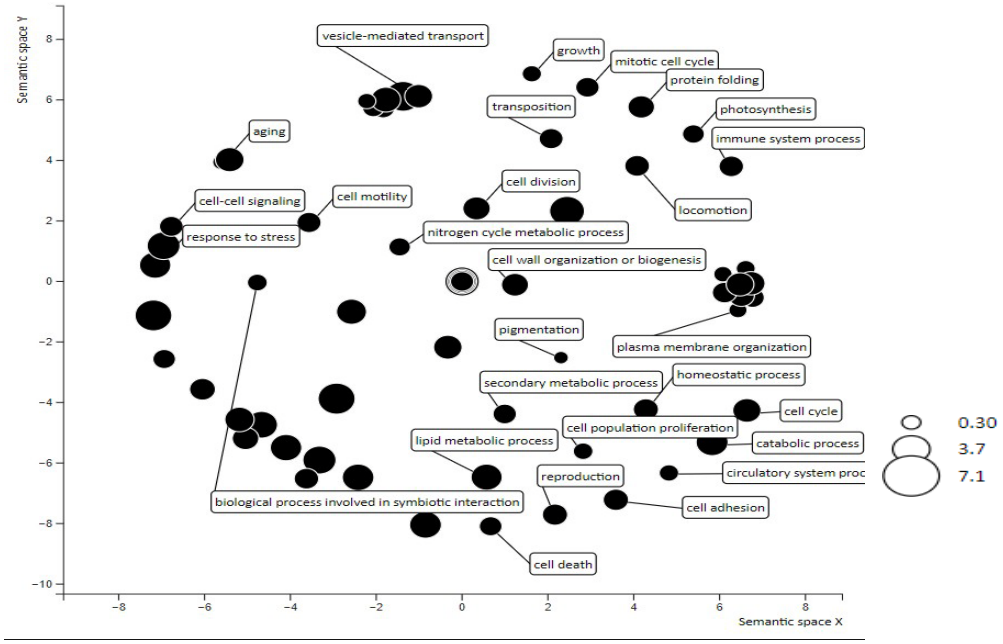

Fig.20. Represents Molecular Function

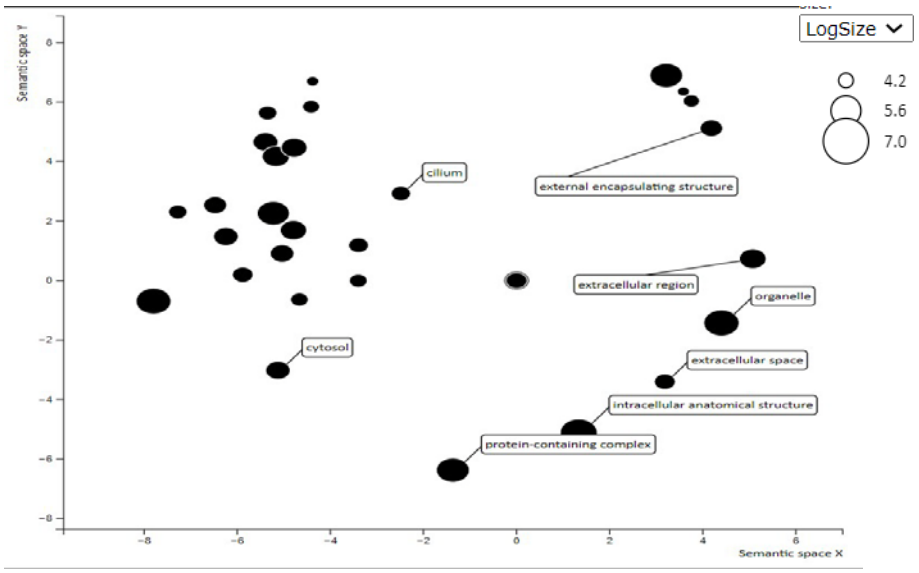

Fig.21 Represents cellular components

Sample 2

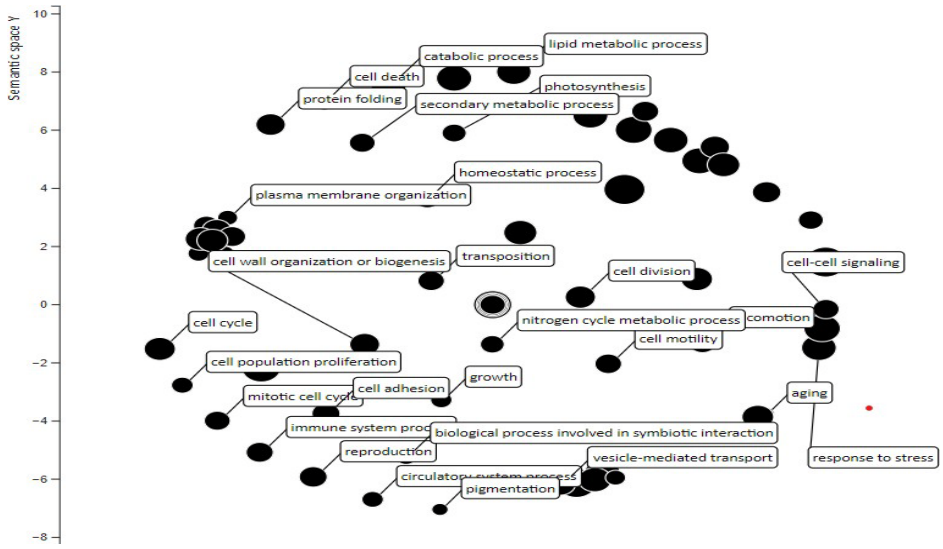

Fig. 22. Represents biological process

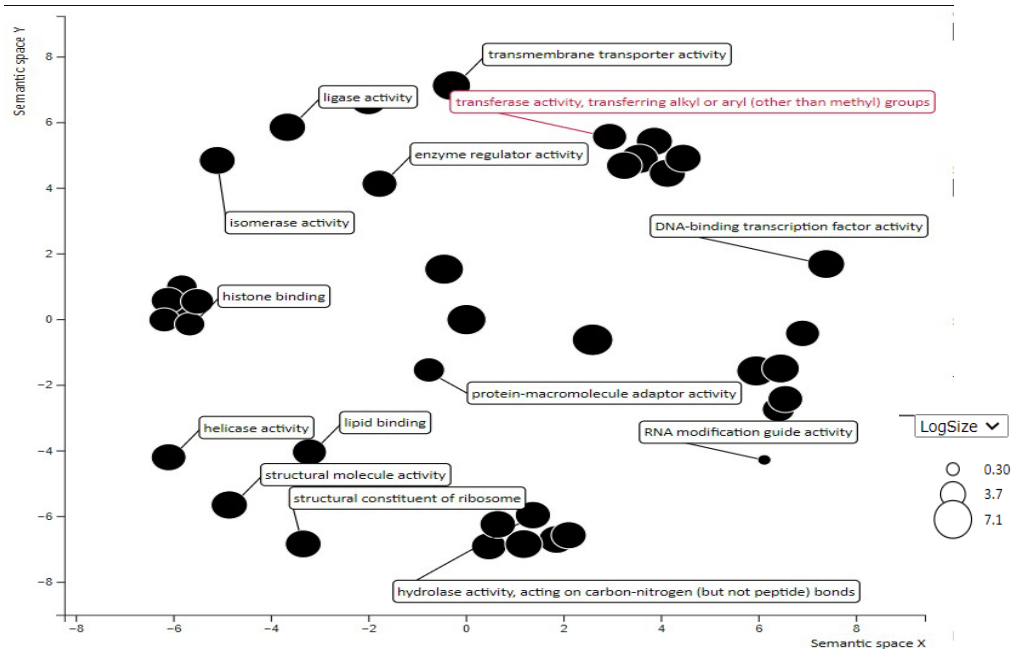

Fig.23. Represents Molecular Function

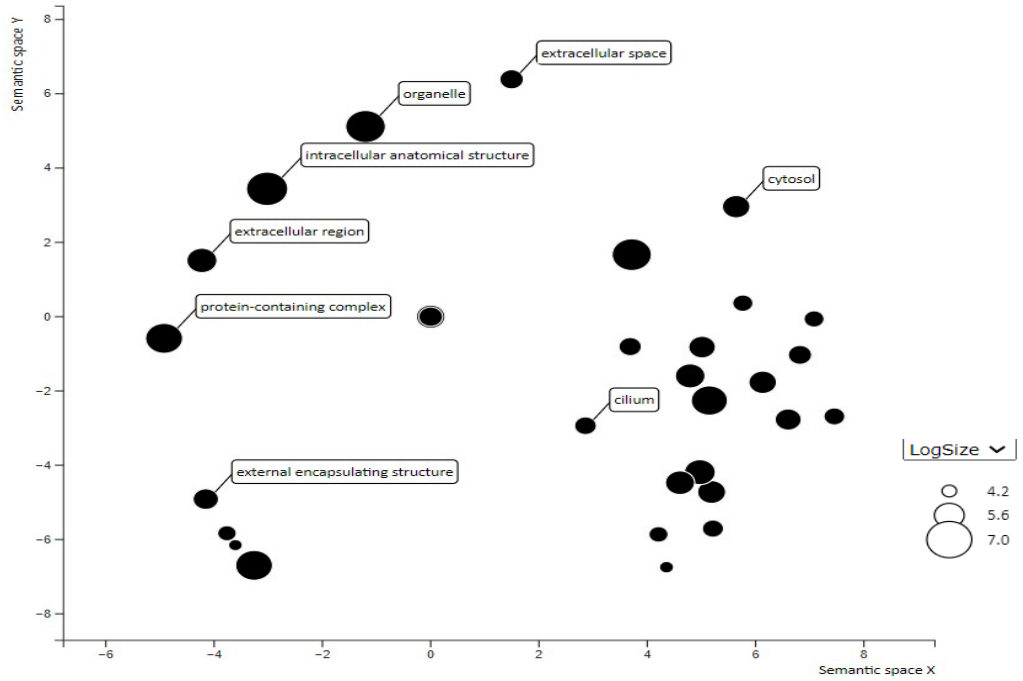

Fig. 24. Represents cellular components

Sample 3

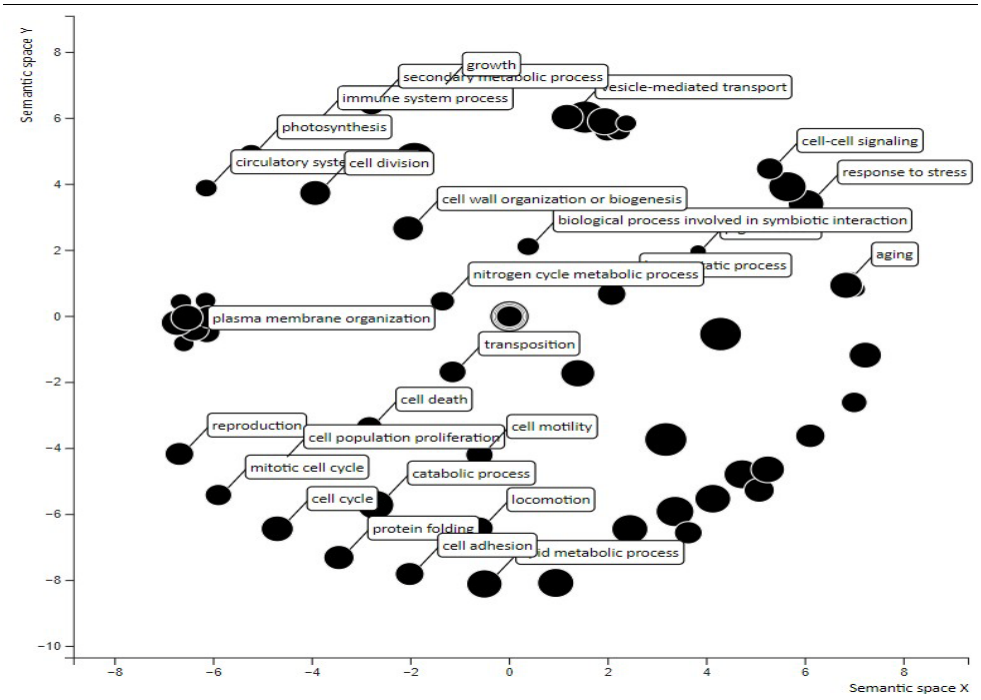

Fig.25. Represents biological process

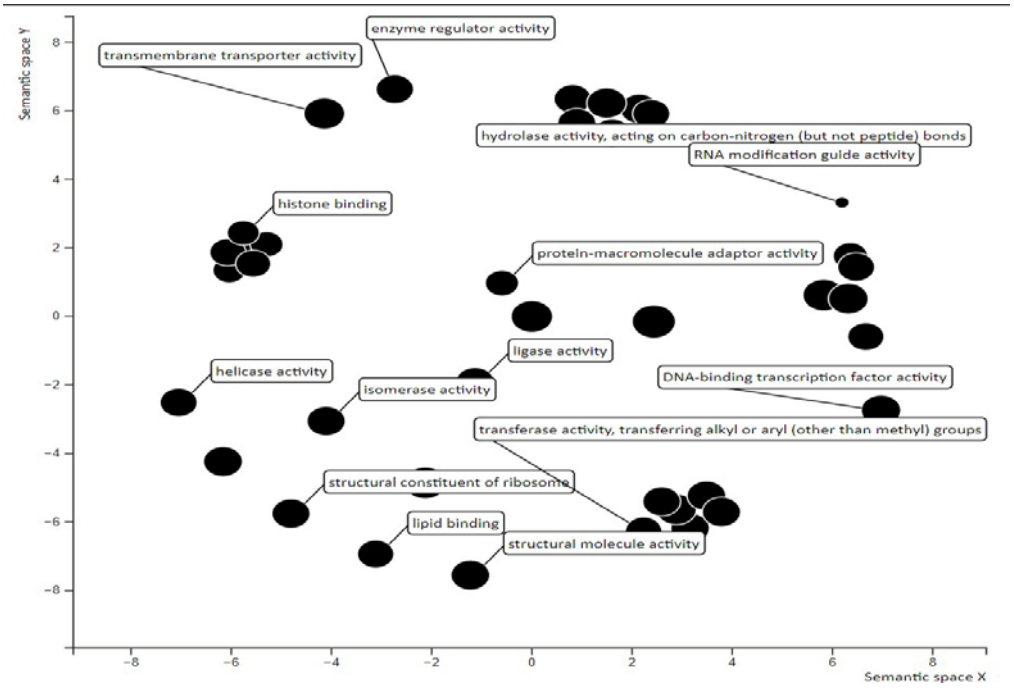

Fig.26. Represents molecular functions

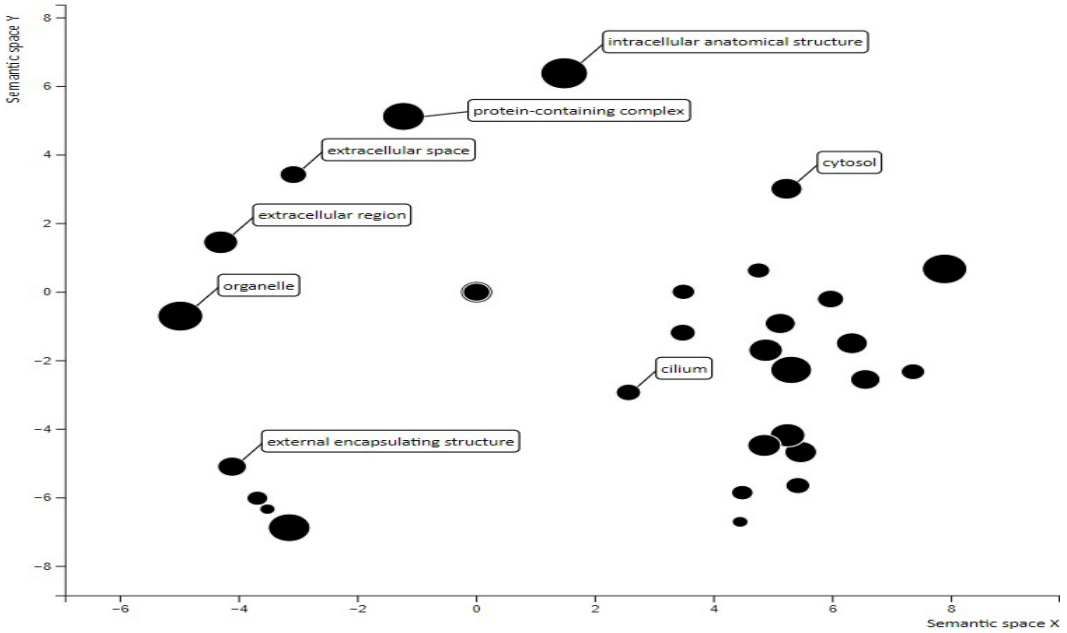

Fig.27. Represents cellular component

Sample 4

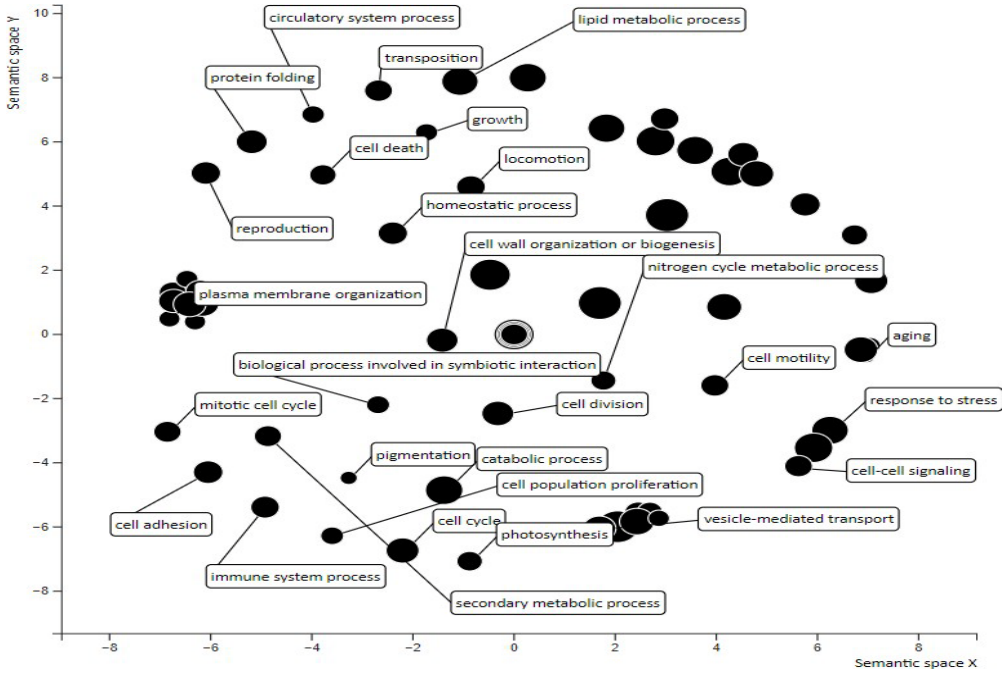

Fig. 28. Represents Biological process

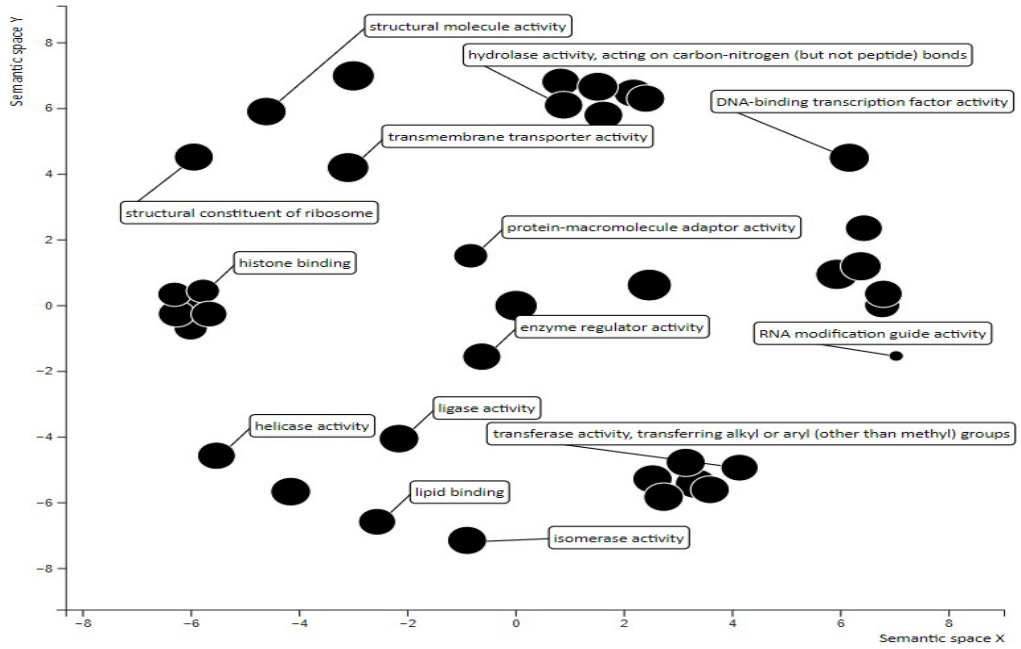

Fig.29. Represents molecular functions

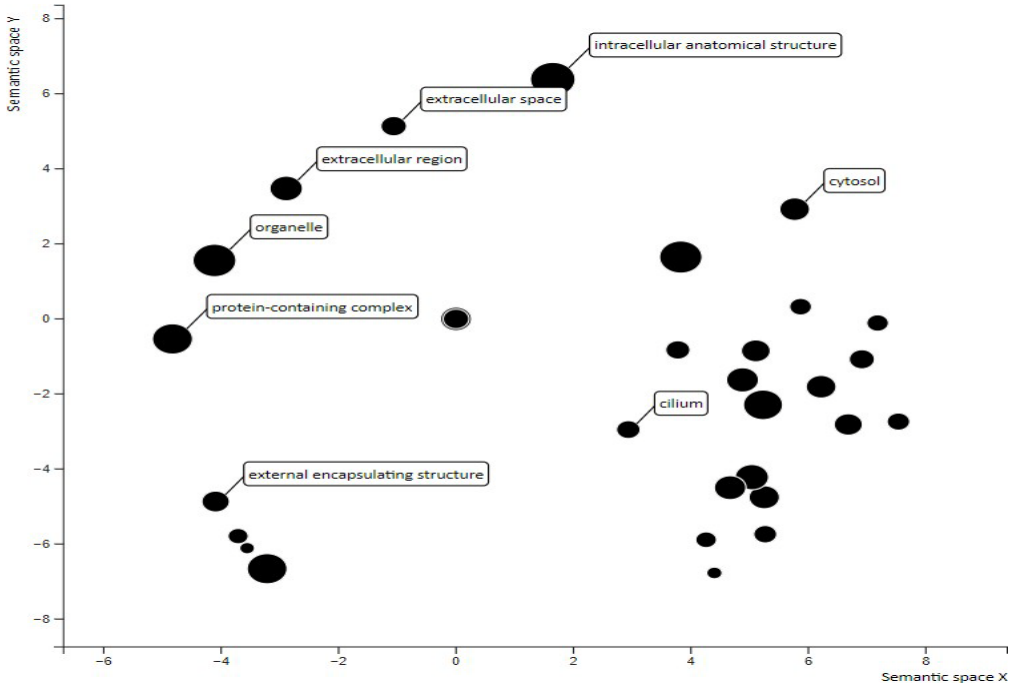

Fig. 30. Represents the cellular functions

7.KEGG PATHWAY ANALYSIS

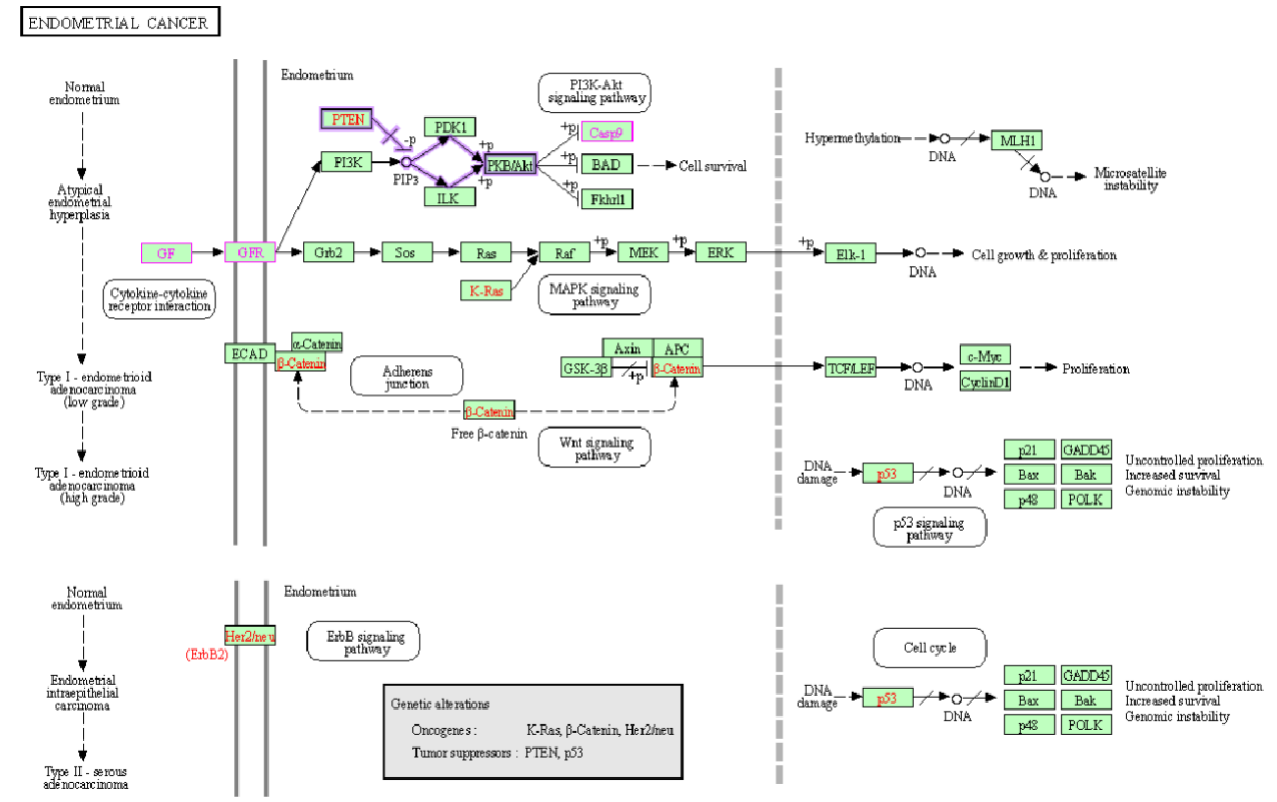

Fig.31. The pi3k-Akt signaling pathway represents EGF (Epidermal growth factor) and Casp9 (Caspase 9) genes that are involved. These genes are upregulated in this pathway. A number of signaling pathways, including the PI3K-Akt process, are activated by the epidermal growth factor (EGF). This pathway is crucial in cell survival and proliferation, its deregulation can support the development of cancer. The role of Casp9 is to kill cancerous cell by apoptosis. If it gets fails to get activated then this leads to some disorders and cancer.

DISCUSSION

KMT2B has received more attention in recent years due to its critical function in accumulating H3K4me3 at promoters that activate expression of gene, which justifies its extensive participation in numerous disease processes. The development and spread of several malignancies, including breast cancer, liver cancer, and colon cancer, have been linked to the KMT2B-H3K4me3 axis. A prior investigation in the setting of CC revealed that H3K4me3 expression was present in nearly all Cervical cancer tissues (96.8%), and that higher H3K4me3 expression was associated with a worse outlook for cervical cancer patients, suggesting a significant role for H3K4me3-mediated epigenomic control in cervical cancer. The function of KMT2B in CC, however, has not been examined. For the first time, the current study concentrated on KMT2B's function in CC. We have done RNA- sequencing analysis of KMT2B overexpressing heLa cellswith heLa cells as control by NGS.

In FastQC, all condition's phred score was good as base calls were lying on green region of graph. The phred score was above 30 which showed that all conditions had good quality. FastQC checks another parameters also for further analysis. If any parameter fails to give good quality then these parameters are corrected by trimming of data. In our further analysis, we got overall alignment rate with respect to reference that was 70 % above which determines that our data matches with the human reference genome. We found significant differentially expressed genes with four comparisons (control vs. replicate) viz., SRR20677779 vs. SRR20677775 (C1 vs. R1), SRR20677780 vs. SRR20677776 (C2 vs. R2), SRR20677781 vs. SRR20677777 (C3 vs. R3), SRR20677782 vs. SRR20677778 (C4 vs. R4). We have done GO annotation for all significant DEG's and out of them it was decided to focus on 24 genes that have a major impact on cancer and that exhibit considerable alterations in for PI3K-Akt signaling pathway. The genes were (CSF3R, SYK, EGF, ITGB3, IGF2, OSM, TGFA, LAMC2, FGF1, IL2RG, THBS1, PIK3CG, PGF, CASP9, GNG2, DDIT4, EIF4, EBP1, GNB3, ITGA7, PIK3AP11, IL7R, TLR4, JAK3, TLR2). From these genes, only 2 genes were upregulated and involved in the PI3K-Akt signaling pathway. The genes were EGF (Epidermal growth factor), and Casp9 (caspase 9). It has been seen studied that when EGF activates PI3K then it do not binds to PIP3 which dysregulates the PI3k/AKT pathway and causes cancer. When EGF regulates the PTEN (Phosphatase TENsin Homolog) then it converts PIP3 to PIP2. This will suppress the effect of cancer and regulates the caspase9 which further able to do its function by removing cancerous with the help of the apoptosis process. In this pathway, PTEN acted as a tumor suppressor gene.

CONCLUSION

The current analysis revealed that KMT2B is increased in Cervical cancer cells and connected to a bad prognosis. Additionally, it demonstrated that KMT2B increased EGF, and Cas9 levels to promote CC cells' angiogenesis and metastasis. This also suggested that KMT2B, EGF, and casp9 has a therapeutic value to treat cervical cancer by acting as possible biomarkers for cervical cancer propensity as well as grading the prognosis of the disease at different stages. Designing new therapeutic targets for these genes will benefit from further research on the mechanism of their interaction in the context of the KMT2B background. From this study, the role of KMT2B has been cleared in cervical cancer.

REFERENCES

1. Watson JD, Crick FH. The structure of DNA. InCold Spring Harbor symposia on quantitative biology 1953 Jan 1 (Vol. 18, pp. 123-131). Cold Spring Harbor Laboratory Press.
2. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH. DNA sequencing at 40: past, present and future. Nature. 2017 Oct 19;550(7676):345-53.
3. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016 Jun;17(6):333-51.
4. Ries LA, Eisner MP, Kosary CL, Hankey BF, Miller BA, Clegg L, Mariotto A, Feuer EJ, Edwards BK. SEER cancer statistics review. National Cancer Institute. 1975;2004.
5. Cibula, D., Pötter, R., Planchamp, F., Avall-Lundqvist, E., Fischerova, D., Haie-Meder, C., Köhler, C., Landoni, F., Lax, S., Lindegaard, J.C. and Mahantshetty, U., 2018. The European Society of Gynaecological Oncology/European Society for Radiotherapy and Oncology/European Society of Pathology guidelines for the management of patients with cervical cancer. *VirchowsArchiv*, *472*, pp.919-936.
6. Cibula D, Pötter R, Planchamp F, Avall-Lundqvist E, Fischerova D, Haie-Meder C, Köhler C, Landoni F, Lax S, Lindegaard JC, Mahantshetty U. The European Society of Gynaecological Oncology/European Society for Radiotherapy and Oncology/European Society of Pathology guidelines for the management of patients with cervical cancer. VirchowsArchiv. 2018 Jun;472:919-36.
7. Mailinh Vu MD , Jim Yu DO , Olutosin A. Awolude MBBS, MSc, FWACS , Linus Chuang MD, MPH, MS. Cervical Cancer Worldwide, Current Problems in Cancer (2018), doi: 10.1016/j.currproblcancer.2018.06.00
8. Mardis ER. A decade's perspective on DNA sequencing technology. Nature. 2011 Feb 10;470(7333):198-203.
9. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X. The complete genome of an individual by massively parallel DNA sequencing. nature. 2008 Apr 17;452(7189):872-6.

10. Shendure J, Ji H. Next-generation DNA sequencing. Nature biotechnology. 2008 Oct;26(10):1135-45.
11. Thudi M, Li Y, Jackson SA, May GD, Varshney RK. Current state-of-art of sequencing technologies for plant genomics research. Briefings in functional genomics. 2012 Jan 1;11(1):3-11
12. Kulski JK. Next-generation sequencing—an overview of the history, tools, and "Omic" applications. Next generation sequencing-advances, applications and challenges. 2016 Jan 14;10:61964.
13. Vezzi F. Next generation sequencing revolution challenges: Search, assemble, and validate genomes.
14. Strausberg RL, Simpson AJ, Old LJ, Riggins GJ. Oncogenomics and the development of new cancer therapies. Nature. 2004 May 27;429(6990):469-74.
15. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S. The consensus coding sequences of human breast and colorectal cancers. science. 2006 Oct 13;314(5797):268-74.
16. Costa V, Aprile M, Esposito R, Ciccodicola A. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. European Journal of Human Genetics. 2013 Feb;21(2):134-42.
17. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology. 2011 Jul;29(7):644-52.
18. Esteller M. Non-coding RNAs in human disease. Nature reviews genetics. 2011 Dec;12(12):861-74.
19. Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, Yang JY, Broom BM, Verhaak RG, Kane DW, Wakefield C. TCPA: a resource for cancer functional proteomics data. Nature methods. 2013 Nov;10(11):1046-7.
20. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018 Nov;68(6):394-424.
21. Smith M, Hammond I, Saville M. Lessons from the renewal of the National Cervical Screening Program in Australia. Public Health Research & Practice. 2019 Jul 31;29(2).
22. Vaccarella S, Lortet-Tieulent J, Plummer M, Franceschi S, Bray F. Worldwide trends in cervical cancer incidence: impact of screening against changes in disease risk factors. European journal of cancer. 2013 Oct 1;49(15):3262-73.

23. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA: a cancer journal for clinicians. 2015 Mar;65(2):87-108.
24. Waggoner SE. Cervical cancer. The lancet. 2003 Jun 28;361(9376):2217-25.
25. Olorunfemi G, Ndlovu N, Masukume G, Chikandiwa A, Pisa PT, Singh E. Temporal trends in the epidemiology of cervical cancer in South Africa (1994–2012). International journal of cancer. 2018 Nov 1;143(9):2238-49.
26. Fedewa SA, Cokkinides V, Virgo KS, Bandi P, Saslow D, Ward EM. Association of insurance status and age with cervical cancer stage at diagnosis: National Cancer Database, 2000–2007. American journal of public health. 2012 Sep;102(9):1782-90.
27. Walsh T, Casadei S, Lee MK, Pennil CC, Nord AS, Thornton AM, Roeb W, Agnew KJ, Stray SM, Wickramanayake A, Norquist B. Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing. Proceedings of the National Academy of Sciences. 2011 Nov 1;108(44):18032-7.
28. Khan MJ, Castle PE, Lorincz AT, Wacholder S, Sherman M, Scott DR, Rush BB, Glass AG, Schiffman M. The elevated 10-year risk of cervical precancer and cancer in women with human papillomavirus (HPV) type 16 or 18 and the possible utility of type-specific HPV testing in clinical practice. Journal of the National cancer Institute. 2005 Jul 20;97(14):1072-9.
29. de Sanjose S, Quint WG, Alemany L, et al. Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study. Lancet Oncol 2010;11:1048–1056.
30. Zhao D, Yuan H, Fang Y, Gao J, Li H, Li M, Cong H, Zhang C, Liang Y, Li J, Yang H. Histone Methyltransferase KMT2B Promotes Metastasis and Angiogenesis of Cervical Cancer by Upregulating EGF Expression. International Journal of Biological Sciences. 2023;19(1):34.
31. Rao RC, Dou Y. Hijacked in cancer: the KMT2 (MLL) family of methyltransferases. Nature Reviews Cancer. 2015 Jun;15(6):334-46.
32. FitzGerald KT, Diaz MO. MLL2: A new mammalian member of the trx/MLL family of genes. Genomics. 1999 Jul 15;59(2):187-92.
33. Li Y, Han J, Zhang Y, Cao F, Liu Z, Li S, Wu J, Hu C, Wang Y, Shuai J, Chen J. Structural basis for activity regulation of MLL family methyltransferases. Nature. 2016 Feb 25;530(7591):447-52.
34. Sanchez R, Zhou MM. The PHD finger: a versatile epigenome reader. Trends in biochemical sciences. 2011 Jul 1;36(7):364-72.

35. Wang Z, Song J, Milne TA, Wang GG, Li H, Allis CD, Patel DJ. Pro isomerization in MLL1 PHD3-bromo cassette connects H3K4me readout to CyP33 and HDAC-mediated repression. Cell. 2010 Jun 25;141(7):1183-94.
36. Bach C, Mueller D, Buhl S, Garcia-Cuellar MP, Slany RK. Alterations of the CxxC domain preclude oncogenic activation of mixed-lineage leukemia 2. Oncogene. 2009 Feb;28(6):815-23.
37. Milne TA, Kim J, Wang GG, Stadler SC, Basrur V, Whitcomb SJ, Wang Z, Ruthenburg AJ, Elenitoba-Johnson KS, Roeder RG, Allis CD. Multiple interactions recruit MLL1 and MLL1 fusion proteins to the HOXA9 locus in leukemogenesis. Molecular cell. 2010 Jun 25;38(6):853-63.
38. Xu C, Liu K, Lei M, Yang A, Li Y, Hughes TR, Min J. DNA sequence recognition of human CXXC domains and their structural determinants. Structure. 2018 Jan 2;26(1):85-95.
39. Tomizawa SI, Kobayashi Y, Shirakawa T, Watanabe K, Mizoguchi K, Hoshi I, Nakajima K, Nakabayashi J, Singh S, Dahl A, Alexopoulou D. Kmt2b conveys monovalent and bivalent H3K4me3 in mouse spermatogonial stem cells at germline and embryonic promoters. Development. 2018 Dec 1;145(23):dev169102.
40. Denissov S, Hofemeister H, Marks H, Kranz A, Ciotta G, Singh S, Anastassiadis K, Stunnenberg HG, Stewart AF. Mll2 is required for H3K4 trimethylation on bivalent promoters in embryonic stem cells, whereas Mll1 is redundant. Development. 2014 Feb 1;141(3):526-37.
41. Sze CC, Cao K, Collings CK, Marshall SA, Rendleman EJ, Ozark PA, Chen FX, Morgan MA, Wang L, Shilatifard A. Histone H3K4 methylation-dependent and-independent functions of Set1A/COMPASS in embryonic stem cell self-renewal and differentiation. Genes & development. 2017 Sep 1;31(17):1732-7.
42. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006 Apr 21;125(2):315-26.
43. Crump NT, Milne TA. Why are so many MLL lysine methyltransferases required for normal mammalian development?. Cellular and Molecular Life Sciences. 2019 Aug 1;76:2885-98.
44. Glaser S, Schaft J, Lubitz S, Vintersten K, van der Hoeven F, Tufteland KR, Aasland R, Anastassiadis K, Ang SL, Stewart AF. Multiple epigenetic maintenance factors implicated by the loss of Mll2 in mouse development.

45. Glaser S, Lubitz S, Loveland KL, Ohbo K, Robb L, Schwenk F, Seibler J, Roellig D, Kranz A, Anastassiadis K, Stewart AF. The histone 3 lysine 4 methyltransferase, Mll2, is only required briefly in development and spermatogenesis. Epigenetics & chromatin. 2009 Dec;2:1-6.
46. Andreu-Vieyra CV, Chen R, Agno JE, Glaser S, Anastassiadis K, Stewart AF, Matzuk MM. MLL2 is required in oocytes for bulk histone 3 lysine 4 trimethylation and transcriptional silencing. PLoS biology. 2010 Aug 17;8(8):e1000453.
47. Park K, Kim JA, Kim J. Transcriptional regulation by the KMT2 histone H3K4 methyltransferases. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms. 2020 Jul 1;1863(7):194545.
48. Zech M, Boesch S, Maier EM, Borggraefe I, Vill K, Laccone F, Pilshofer V, Ceballos-Baumann A, Alhaddad B, Berutti R, Poewe W. Haploinsufficiency of KMT2B, encoding the lysine-specific histone methyltransferase 2B, results in early-onset generalized dystonia. The American Journal of Human Genetics. 2016 Dec 1;99(6):1377-87.
49. Meyer E, Carss KJ, Rankin J, Nichols JM, Grozeva D, Joseph AP, Mencacci NE, Papandreou A, Ng J, Barral S, Ngoh A. Mutations in the histone methyltransferase gene KMT2B cause complex early-onset dystonia. Nature genetics. 2017 Feb;49(2):223-37.
50. Poreba E, Lesniewicz K, Durzynska J. Aberrant activity of histone–lysine n-methyltransferase 2 (Kmt2) complexes in oncogenesis. International journal of molecular sciences. 2020 Dec 8;21(24):9340.
51. Takahashi S, Yokoyama A. The molecular functions of common and atypical MLL fusion protein complexes. Biochimica Et Biophysica Acta (BBA)-Gene Regulatory Mechanisms. 2020 Jul 1;1863(7):194548.
52. Wong WH, Junck L, Druley TE, Gutmann DH. NF1 glioblastoma clonal profiling reveals KMT2B mutations as potential somatic oncogenic events. Neurology. 2019 Dec 10;93(24):1067-9.
53. Li Y, Zhao L, Tian X, Peng C, Gong F, Chen Y. Crystal structure of MLL2 complex guides the identification of a methylation site on P53 catalyzed by KMT2 family methyltransferases. Structure. 2020 Oct 6;28(10):1141-8.
54. Mountzios G, Rampias T, Psyrri A. The mutational spectrum of squamous-cell carcinoma of the head and neck: targetable genetic events and clinical impact. Annals of oncology. 2014 Oct 1;25(10):1889-900.
55. Gregory RI, Chendrimada TP, Cooch N, Shiekhattar R. Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. Cell. 2005 Nov 18;123(4):631-40.

56. Rossi JJ. New hope for a microRNA therapy for liver cancer. Cell. 2009 Jun 12;137(6):990-2.
57. Mendell JT, Olson EN. MicroRNAs in stress signaling and human disease. Cell. 2012 Mar 16;148(6):1172-87.
58. Calin GA, Croce CM. MicroRNA signatures in human cancers. Nature reviews cancer. 2006 Nov 1;6(11):857-66.
59. Wilting SM, Snijders PJ, Verlaat W, Jaspers AV, Van De Wiel MA, Van Wieringen WN, Meijer GA, Kenter GG, Yi Y, Le Sage C, Agami R. Altered microRNA expression associated with chromosomal changes contributes to cervical carcinogenesis. Oncogene. 2013 Jan;32(1):106-16.
60. Xu XM, Wang XB, Chen MM, Liu T, Li YX, Jia WH, Liu M, Li X, Tang H. MicroRNA-19a and-19b regulate cervical carcinoma cell proliferation and invasion by targeting CUL5. Cancer letters. 2012;2(322):148-58.
61. Juan L, Tong HL, Zhang P, Guo G, Wang Z, Wen X, Dong Z, Tian YP. Identification and characterization of novel serum microRNA candidates from deep sequencing in cervical cancer patients. Scientific Reports. 2014 Sep 3;4(1):6277.
62. Kogo R, How C, Chaudary N, Bruce J, Shi W, Hill RP, Zahedi P, Yip KW, Liu FF. The microRNA-218~ Survivin axis regulates migration, invasion, and lymph node metastasis in cervical cancer. Oncotarget. 2015 Jan;6(2):1090.
63. Engelbrecht AM, Gebhardt S, Louw L. Ex vivo study of MAPK profiles correlated with parameters of apoptosis during cervical carcinogenesis. Cancer letters. 2006 Apr 8;235(1):93-9.
64. Gao LJ, Gu PQ, Zhao W, Ding WY, Zhao XQ, Guo SY, Zhong TY. The role of globular heads of the C1q receptor in HPV 16 E2-induced human cervical squamous carcinoma cell apoptosis is associated with p38 MAPK/JNK activation. Journal of translational medicine. 2013 Dec;11(1):1-1.
65. Chen TP, Chen CM, Chang HW, Wang JS, Chang WC, Hsu SI, Cho CL. Increased expression of SKP2 and phospho-MAPK/ERK1/2 and decreased expression of p27 during tumor progression of cervical neoplasms. Gynecologic oncology. 2007 Mar 1;104(3):516-23.
66. Su PH, Lin YW, Huang RL, Liao YP, Lee HY, Wang HC, Chao TK, Chen CK, Chan MW, Chu TY, Yu MH. Epigenetic silencing of PTPRR activates MAPK signaling, promotes metastasis and serves as a biomarker of invasive cervical cancer. Oncogene. 2013 Jan;32(1):15-26.

67. Sharma G, Dua P, Mohan Agarwal S. A comprehensive review of dysregulated miRNAs involved in cervical cancer. Current genomics. 2014 Aug 1;15(4):310-23.
68. Mekhilef S, Saidur R, Kamalisarvestani M. Effect of dust, humidity and air velocity on efficiency of photovoltaic cells. Renewable and sustainable energy reviews. 2012 Jun 1;16(5):2920-5.
69. Zhang TT, Qu N, Sun GH, Zhang L, Wang YJ, Mu XM, Wei WJ, Wang YL, Wang Y, Ji QH, Zhu YX. NRG1 regulates redox homeostasis via NRF2 in papillary thyroid cancer. International journal of oncology. 2018 Aug 1;53(2):685-93.
70. Lee E, Ouzounova M, Piranlioglu R, Ma MT, Guzel M, Marasco D, Chadli A, Gestwicki JE, Cowell JK, Wicha MS, Hassan KA. The pleiotropic effects of TNFα in breast cancer subtypes is regulated by TNFAIP3/A20. Oncogene. 2019 Jan 24;38(4):469-82.
71. Yang Y, Zhang J, Xia T, Li G, Tian T, Wang M, Wang R, Zhao L, Yang Y, Lan K, Zhou W. MicroRNA-210 promotes cancer angiogenesis by targeting fibroblast growth factor receptor-like 1 in hepatocellular carcinoma. Oncology Reports. 2016 Nov 1;36(5):2553-62.
72. Team ST. SRA Toolkit. Nation of National Center for Biotechnology. 2020.
73. Andrews S. FastQC: a quality control tool for high throughput sequence data.
74. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20.
75. Langmead B. Aligning short sequencing reads with Bowtie. Current protocols in bioinformatics. 2010 Dec;32(1):11-7.
76. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. bioinformatics. 2009 Aug 15;25(16):2078-9.
77. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols. 2012 Mar;7(3):562-78.