Communication

Exome Sequencing Reveals Characteristic KMT2C Mutations in The Indian Phenotype Of Cervical Cancer

Santosh Kumari Duppala[2,9], Bhumandeep Kour[2,9], Nidhi Shukla[3,9], Maruti Dhakane[4], Ashwani Kumar Mishra[5], Raghunadharao Digumarti[6], Smitha C Pawar[7#], Prashanth Suravajhala[8, 9#] and Sugunakar Vuree[1, 2 9#]

1. MNR Foundation for Research & Innovation, Fasalwadi Village, Sangareddy (District), Telangana, India-502294.
2. Department of Microbiology and Biotechnology, Lovely Professional University, Jalandhar, Punjab, India
3. Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research, Jaipur, Rajasthan, India
4. Department of Pathology & Laboratory services Apollo Hospitals, Bilaspur Chhattisgarh India
5. DNA Xperts Private Limited, E-227, Sector 63, Noida, UP-201301.
6. KIMS-ICON Hospital, Visakhapatnam, India.
7. Department of Genetics and Biotechnology, Osmania University, Hyderabad, Telangana, India
8. Amrita School of Biotechnology, Amrita Vishwa Vidyapeetham, Amritapuri, Clappana PO 690525, Kerala
9. Bioclues.org, India

Correspondence: smita.prof@gmail.com; prash@bioclues.com, and sugunakarvure@gmail.com

Abstract:

We attempted to understand the cervical cancer patient samples through Whole Exome Sequencing. We derived the variants from raw reads via our in-house benchmarked pipeline and validated the variants by IGV. This is the first cervical cancer exome data from the Indian cohort.

Background: Cervical cancer (CC) is caused mainly by persistent infections of high-risk HPV, reduced parity, and factors like a decrease in average socioeconomic levels. We keep this because no cervical cancer exome data is available from the Indian cohort.

Methods: The CC patient clinical samples were initially subjected to preparation using Qiagen DNA extraction, quantified, and the library preparation using the Agilent target enrichment system. Further, exome capture by Illumina platform (100X), quality check, alignment, and variant calling followed by the downstream analysis and finally visualized by IGV.

Results: We observed a large number of SNVs or mutations from an Indian perspective, such as KMT2C, OR4M1, PDPR2P, EPHB1, FAS, OPCML, MGST1, C1QTNF9, HS6ST3, OR4K2, PRPSAP2, KCNJ12, FIGNL1, SFXN1, BAGE2, ARVCF, NAMPT variants of significance, unknown significance and possibly significant are reported.

Conclusion: For the first time, KMT2C is observed as a novel potent mutation and pathogenic, showing the variant position at (7:152265091, T>A, SNV 62478356) from the Indian context in CC. Further, we visualized

 doi:10.20944/preprints202211.0440.v1

and validated the mutations using the Integrative Genomic Viewer (IGV) browser. Finally, we discuss the inherent challenges through KMT2C mutations.

Keywords: Cervical cancer; Functional genomics; Bioinformatics; Next Generation Sequencing; Exome

Introduction:

In 2020, the world health organization (WHO) claimed cancer as the first or the second main reason for mortality earlier than 65-70 years of age in 185 countries. According to GLOBOCAN 2020 cervical cancer (CC) diagnosis and prevalence was seen in 26 countries and leading mortality cancer among 36 countries [1]. The CC is the world's 4th largest cancer-causing incidence and 4th largest mortality among women in India and other low-income countries. It is the second largest cancer among women with incidence cases worldwide count for 604, 127 and mortality cases 341,831 cases worldwide [1]. GLOBOCAN 2020 studies suggest that India and China contribute more than a third of CC cases. In India the incidence rates are 96,922 (17%) and death rates are 60,078(19.3%) per year [2,3]. While CC is a major health problem in young women and middle-aged in developing and underdeveloped countries, over the past few years, there is a decline in age standardised mortality in developed countries like the US, Europe, and the UK due to the increased awareness in screening CC diagnosis, and HPV vaccination [1, 4, 5]. There is much need for awareness and vaccination programs in underdeveloped and developing countries like Africa, India, and China. About half of the cases of CC are caused due to persistent infections with high-risk Human papillomavirus, and it turns mild dysplasia and later to TNM stages I, II, III, and invasive cancer [6,7]. Whereas the screening is performed through liquid cytology, pap smear, HPV co-testing to prevent CC and HPV. A late-stage diagnosis of CC may show the cervix as normal, and it spreads to the endocervix or can metastasis through lymphatic vessels to lymph nodes [6]. The first step to prevent CC is to avoid HPV infection through HPV vaccination Gardasil Quadrivalent associated with HPV 6, 11, 16, and 18 types [8].

Next generation sequencing (NGS) technologies like whole exome sequencing (WES) are used to find genetic variants or mutations that can cause an infectious disease or cancer. The WES approaches use various library preparation steps to capture exons for enrichment. As there is poor prognosis in CC caused for advanced and recurrent patients. So, using whole exome sequencing data, we can identify multiple genes responsible for CC and recurrent CC, SNVs and mutations responsible for both cervical carcinoma and adenocarcinoma. A large

 doi:10.20944/preprints202211.0440.v1

number of studies and biomarker discovery was observed in Chinese, USA, Europe and very few studies are known from Indian Perspective. Though there are several studies in CC relevant to WES in vivid populations, there are hardly any studies to screen mutations relevant to India. With poor prognosis in CC caused for advanced and recurrent patients, we in this study attempted to explore mutations inherent to Indian phenotype in 10 samples. A detailed pipeline and description of mutations across 5 inherent TNM grades of cancers are discussed.

Methods and Materials:

Clinical Samples:

The samples were collected from MNJ cancer hospital, Hyderabad, India. All ethics clearances were accorded as per the institutional ethics committee of MNJ cancer hospital Hyderabad, and informed consent was judiciously taken. Five normal and five malignant CC specimens with formalin fixed samples with stages TNM II and III were considered for the study. For DNA WES, formalin-fixed fresh cervical tumor samples were taken and sent to DNA Xperts, Ghaziabad, New Delhi. The CC samples are labelled for identification, as shown in Table 1

briefly, the management of five cervical cancer samples and five surrounding tissue samples of moderately differentiated squamous cell carcinoma and adenocarcinoma having TNM stages I, II, and III were included after surgical hysterectomy procedures. All the samples were subjected to DNA WES on the Illumina platform.

Sample Preparation:

Cervical tumor formalin dipped DNA samples are washed and cleaned with PBS (Phosphate Buffer Saline), and later isolation is performed using a Qiagen FFPE DNA extraction kit. Then (0.8%) agarose gel is prepared to run the genomic DNA at 110 V for 40 minutes (Figure.1). To determine concentration and quantification, each sample is estimated using Qubit Fluorometer. Further, library preparation was performed using the Agilent SureselectXT target enrichment system. Moreover, 200ng of each DNA sample was used for fragmentation using Covaris S. For FFPE-derived DNA samples with significantly degraded DNA, used the concentration of amplifiable DNA as determined by qPCR and use the maximum amount of DNA available in the range of 100–200 ng. After the fragmentation of DNA, it is put through to repair the ends. It is followed

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2022 doi:10.20944/preprints202211.0440.v1

by sample purification using AMPureXP beads and dA-tail the 3 of DNA fragments and later ligates the paired-end adaptor to capture DNA-based sequences (Figure 2). Finally, the adaptor-ligated exome library was used to capture the DNA.

Exome Capture:

The ten samples were subjected to exome sequencing using the Illumina platform. The exon capture of the reads with a mean target coverage depth of 100X per sample followed by in-solution capture of genomic DNA by Sureselect Human All exon V5 UTR kit (Agilent Technologies, USA) targeting 40Mb from exons. The captured DNA for each sample is sequenced by multiplexed paired-end sequencing using two pools of samples on the Illumina platform.

Quality check, Alignment, and Variant calling:

The sequences were passed for quality check and aligned to human genome hg38 using the bowtie2 command line for the alignment of the sequences. The files were cross-checked for contamination, and variant analysis and annotation of variants were filtered by setting a false discovery rate.

To check the quality and GC content of all the samples, raw readings are run via our in-house benchmarked pipeline [9] (Figure 2) The samples were processed at three levels: pre-processing, the discovery of variants, and prioritisation of variants. A crosscheck for contamination was undertaken to see the heterogeneity and filter the variants. BAM files are created by converting Sam files to BAM files. To detect heterozygous variations, all sequential processes such as sorting indels are conducted using the variant calling tool Varscan. The heterozygous variations are enumerated with awk/bash command one-liners, and the average depth of >5 is taken into account.

Downstream analysis:

The filtered variants and samples' VCF files go through downstream analysis using NCBI and various bioinformatics tools. The VCF files are uploaded in SNP nexus first and downloaded all the results of samples. Then, the variants are confirmed from the NCBI database using Clinvar to check the pathogenicity and missense mutations further, the CADD scores >10 and GREP scores values >2 are filtered. Those variants

 doi:10.20944/preprints202211.0440.v1

are seen in Venny 2.1 for common genomic variants (Figure 3a,3b,3c, 3d,3e,3f. Moreover, to confirm through dbSNP the SNP with rs ID whose value of minor allele frequency (MAF) of GnomAD is taken less than 0.05. While running the pipeline for variant identification, we have taken the depth of the variants that mainly fall between ≥DP 5 and ≤20.

Integrative Genome Viewer Browser(IGV): IGV tool representing Sashimi plots were used for quantitative alignment of RNA-seq reads for visualization and also to compare the exon coverage across the samples [10,11].

RESULTS & DISCUSSION:

Characteristics of CC WES from an Indian Perspective:

We observed that the total number of annotated variants of all samples is more than 1,40,062 after normalization from the SNP nexus. To remove false positives, we observed variation between the controls and inclusion of tumour samples to find the heterozygous variants and exclude the low coverage sites and deletions. We found 62,781 heterozygous variants after filtering, as per CADD scores >10, and Genomic evolutionary regulatory profiling GERP scores >2 were considered for pathogenicity and mutations. We observed a total of 54 SNPs, out of which 17 SNPs are confirmed that exhibited significant association with CC.

Variants showing mutations:

Among the variants showing mutations and pathogenicity were KMT2C, OR4M1, PDPR2P, EPHB1, FAS, OPCML, MGST1, C1QTNF9, HS6ST3, OR4K2, PRPSAP2, KCNJ12, FIGNL1, SFXN1, BAGE2, ARVCF, NAMPT (Table 2).

Gene enrichment pathway analysis: The mutations encoded genes were then subjected to the interaction network analysis with all the 17 genes (Figure 4a) showing coexpression (75.79%) physical interactions (14.49%), co-localisation (8.26%) and genetic association (1.46%) from GeneMania. Among them, associations show the majority of the interactions with uterus, CC, endometrial, oral, leukemia. The genes from GeneMania interactions show integrated biological pathway associations.

 doi:10.20944/preprints202211.0440.v1

Interaction network of KMT2C:

The interaction network of KMT2C (also known as MML3) (Figure 4b) includes Physical interactions- 77.64%, Co-expression- 8.01%, Genetic interaction: 2.87% from GeneMania. KMT2C is a chromatin remodelling gene, where histone acetylation and modulation of genes happen, KMT2C along with its lysine methyltransferase sub family KMT2D, A, E cause leukemia or Mixed lineage leukemia. The other genes that interact with KMT2C include WDR5 causes Kubuki syndrome and its pathways PKN1 stimulates transcription factor androgen receptors the same as KMT2C while ASCL2 is mostly seen in breast cancer cell lines and frequently most of the genes are involved in transcription factors.

IGV Browser:

KMT2C is a lysine methyltransferase 2C gene and is associated with Myeloid or lymphoid lineage leukemia. It is one of the 10 ten mutated genes among the CC in our study. In one of the IB 1 to IV FIGO stage population study progression-free survival rates at median follow-up by chemoradiation (87%) in CC were observed, While they also show dominant alterations in KMT2C 16%, KMT2D 15% and PIK3CA 40% [12], It is mainly observed as a Kleefstra Syndrome-2 seen as heterozygous deletion or transversion in KMT2C gene (https://www.omim.org/entry/606833#2). KMT2C mediates the mono and trimethylation of H3 histone lysine 4 commonly called MLL 3. Faundes et al 2018, [13] observed and reported in WES of 3 patient samples (Figure 5a-5e) deciphering developmental disorders showing de novo heterozygous mutations with severe intellectual disability, early infancy, poor speech and walk. KMT2C is implicated in tumorigenesis as oncogenes and tumour suppressors in various neoplasias. It is linked to tumor sequencing studies in various cancers ascertaining lower KMT2C activity led to a deficiency in homologous recombination-mediated double-strand break DNA repair and the cells suffer from substantially higher endogenous DNA damage and genomic instability. Such cells rely heavily on PARP1/2 for DNA repair. Studies in bladder cancer have earlier reported that individuals with KM2TC variants respond better to PARP inhibitors like Olaparib than Cisplatin. (Joshi, Asim, et al. 2019; Mann, Minakshi, et al. 2019),[14],[15]. Cancer cells acquire Cisplatin resistance through the DNA repair mechanism; hence, cancer cells with low KMT2C expression are attractive targets for therapies with PARP1/2 inhibitors. Therefore, studies focused on how KMT2C mutations alter gene expression and ways to therapeutically target these mutations are very important in the current scenario.

 doi:10.20944/preprints202211.0440.v1

COSMIC and cBioportal Database:

Through the Genome Browser COSMIC database the SNVs having rs ids were annotated and 138908625, 62478356 respectively (Figure 6). Cosmic database shows KMT2C mutations which were missense, frameshift, deletions and substitutions. The WES SNVs of CC were compared with available data of PAN cancer studies of 278 samples of Cervical squamous cell carcinoma from the cBioportal database, KMT2C plot of overall survival against the probability of overall survival that denotes the altered and unaltered groups (Figure.7). Our work has certain limitations as well with lack of Sanger validation, which otherwise we have used an Integrative genomic browser to visualise the genomic data.

Conclusions

CCs potentially have an increased rate of oncogenic mutations. The WES is used to find out the clinical phenotypes, genetic variants and causes of disease and to understand the molecular mechanisms of the genes to find the new diagnostic and therapeutic targets of CC. Our comprehensive WES analysis through bioinformatics lays the foundation for understanding novel pathogenic mutations from the Indian perspective. In our pilot study of 5 tumor and 5 surrounding tissue samples, we attempted to identify potent pathogenic and novel mutations. Among them, KMT2C is observed as one of the 10 ten mutated genes in CC specific to the Indian phenotype of CC

Acknowledgements:

We gratefully acknowledge the support from MNJ Hospital, Hyderabad for the ethical committee towards kind assistance of tissue sampling. We deeply acknowledge SCP for her support. We further acknowledge Surender Tiwari, Technician, Department of Pathology, Apollo hospitals, Bilaspur for their assistance in knowledge sharing, especially with regard to the selection of tumor types. We are greatly indebted to Bioclues.org and its founder PS for his valuable time and energy in helping with data analysis and editing the complete manuscript preparations. We gratefully acknowledge the resources provided by BISR, Jaipur, Rajasthan. Thank you to DNA Xperts Ashwani for performing Whole Exome Sequencing. SGV is grateful to the Department of Genetics and Biotechnology, Osmania University, Hyderabad and the Department of

Biotechnology, School of Bioengineering and Biosciences, Lovely Professional University, Jalandhar Punjab for their support and encouragement.

Author contributions:

SV, SCP and PS conceived the idea. SKD wrote the first draft of the manuscript. SKD and PS analysed the data and interpreted the whole manuscript. BK, MD, AK, RD and NS performed lateral analyses. PS and SV critically reviewed the manuscript and proofread it. All the authors have read and approved the final version of the manuscript.

Funding: None

Ethics clearance: All the institutional guidelines and ethical clearance are carried out from MNJ hospital Hyderabad Telangana (Regd No: ECR/227/Inst/AP/2013/RR-19) with an approval date of 20 Jan 2020.

Data Availability: All raw reads are available through NCBI-SRA- PRJNA

Bioproject submission of: SUB10812780

References:

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2021 May;71(3):209-49.DOI: 10.3322/caac.21660

[2] Poondla N, Madduru D, Duppala SK, Velpula S, Nunia V, Kharb S, Ghatak S, Mishra AK, Vuree S, Neyaz MK, Suravajhala P. Cervical cancer in the era of precision medicine: A perspective from developing countries. Advances in Cancer Biology-Metastasis. 2021 Dec 1;3:100015.https://doi.org/10.1016/j.adcanc.2021.100015

[3] Piñeros M, Saraiya M, Baussano I, Bonjour M, Chao A, Bray F. The role and utility of population-based cancer registries in cervical cancer surveillance and control. Preventive medicine. 2021 Mar 1;144:106237.DOI: 10.1016/j.ypmed.2020.106237

[4] Arbyn M, Weiderpass E, Bruni L, de Sanjosé S, Saraiya M, Ferlay J, Bray F. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. The Lancet Global Health. 2020 Feb 1;8(2):e191-203.DOI: 10.1016/S2214-109X(19)30482-6

[5] Chan CK, Aimagambetova G, Ukybassova T, Kongrtay K, Azizan A. Human papillomavirus infection and cervical cancer: epidemiology, screening, and vaccination—review of current perspectives. Journal of oncology. 2019 Oct 10;2019.DOI: 10.1155/2019/3257939

[6] Cohen PA, Jhingran A, Oaknin A, Denny L. Cervical cancer. The Lancet. 2019 Jan 12;393(10167):169-82. DOI: 10.1016/S0140-6736(18)32470-X

[7] Wang YX, Arvizu M, Rich-Edwards JW, Stuart JJ, Manson JE, Missmer SA, Pan A, Chavarro JE. Menstrual cycle regularity and length across the reproductive lifespan and risk of premature mortality: prospective cohort study. bmj. 2020 Sep 30;371. DOI: 10.1136/bmj.m3464

[8] Fischer AK, Reuter-Jessen K, Schildhaus HU, Hugo T, Scheel AH, Merkelbach-Bruse S, Heinmöller E, Buettner R, Jasani B, Walbeck J, Rüschoff J. Development of high-risk HPV associated cervical dysplasia despite HPV-vaccination: a regional dysplasia center cohort study. Eur. J. Gynaecol. Oncol. 2020 Feb 15;41(1). DOI: 10.31083/j.ejgo.2020.01.5093

[9] Meena N, Mathur P, Medicherla KM, Suravajhala P. A bioinformatics pipeline for whole exome sequencing: overview of the processing and steps from raw data to downstream analysis. Bio-protocol. 2018 Apr 20:e2805-. DOI: 10.21769/BioProtoc.2805

[10] Katz Y, Wang ET, Silterra J, Schwartz S, Wong B, Thorvaldsdóttir H, Robinson JT, Mesirov JP, Airoldi EM, Burge CB. Quantitative visualization of alternative exon expression from RNA-seq data. Bioinformatics. 2015 Jul 15;31(14):2400-2. DOI: 10.1093/bioinformatics/btv034

[11] Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics. 2013 Mar 1;14(2):178-92.DOI: 10.1093/bib/bbs017

[12] Scholl S, Popovic M, de la Rochefordiere A, Girard E, Dureau S, Mandic A, Koprivsek K, Samet N, Craina M, Margan M, Samuels S. Clinical and genetic landscape of treatment naive cervical cancer: Alterations in PIK3CA and in epigenetic modulators associated with sub-optimal outcome. EBioMedicine. 2019 May 1;43:253-60.DOI: 10.1016/j.ebiom.2019.03.069

[13] Faundes V, Newman WG, Bernardini L, Canham N, Clayton-Smith J, Dallapiccola B, Davies SJ, Demos MK, Goldman A, Gill H, Horton R. Histone lysine methylases and demethylases in the landscape of human developmental disorders. The American Journal of Human Genetics. 2018 Jan 4;102(1):175-87. DOI: 10.1016/j.ajhg.2017.11.013

[14] Joshi A, Mishra R, Desai S, Chandrani P, Kore H, Sunder R, Hait S, Iyer P, Trivedi V, Choughule A, Noronha V. Molecular characterization of lung squamous cell carcinoma tumors reveals therapeutically relevant alterations. Oncotarget. 2021 Mar 16;12(6):578. DOI: 10.18632/oncotarget.27905

[15] Mann M, Kumar S, Chauhan SS, Bhatla N, Kumar S, Bakhshi S, Gupta R, Sharma A, Kumar L. Correction: PARP-1 inhibitor modulate β-catenin signaling to enhance cisplatin sensitivity in cancer cervix. Oncotarget. 2019 Jul 7;10(46):4802. DOI: 10.18632/oncotarget.27101