Genome-wide approaches to characterize novel genetic elements causing Cancer

By

Trupti Ajay Togar

LIFE09201404002

Tata Memorial Centre, Navi Mumbai

A thesis submitted to the

Board of Studies Life Sciences

in partial fulfillment of requirements

for the Degree of

DOCTOR OF PHILOSOPHY

of

HOMI BHABHA NATIONAL INSTITUTE

January, 2021

Homi Bhabha National Institute

Recommendations of the Viva Voce Committee

As members of the Viva Voce Committee, we certify that we have read the dissertation prepared by Ms Trupti Ajay Togar entitled "Genome-wide approaches to characterize novel genetic elements causing cancer" and recommend that it may be accepted as fulfilling the thesis requirement for the award of Degree of Doctor of Philosophy.

S. N. Dalal — 7/1/21

Chairperson–Dr. Sorab N Dalal — Date:

A. Dutt — 7/1/21

Guide/Convener –Dr. Amit Dutt — Date:

A. Sharma — 07/01/2021

External Examiner–Dr. Ashok Sharma — Date:

Dr. Ashok Sharma Ph.D
Associate Professor/DST-Ramanujan Awardee
Department of Biochemistry
All India Institute of Medical Sciences
Ansari Nagar, New Delhi-110029, India

Prasanna — 7/1/21

Member –Dr. Prasanna Venkatraman — Date:

T. Teni — 7/1/21

Member -Dr. Tanuja Teni — Date:

Final approval and acceptance of this thesis is contingent upon the candidate's submission of the final copies of the thesis to HBNI.

I hereby certify that I have read this thesis prepared under my direction and recommend that it may be accepted as fulfilling the thesis requirement.

Date: 8/1/21
Place: Navi Mumbai

A. Dutt — 8/1/21

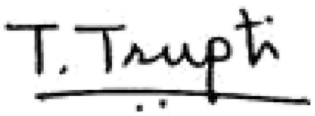
Dr. Amit Dutt
Guide

CONTENTS

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at Homi Bhabha National Institute (HBNI) and is deposited in the Library to be made available to borrowers under rules of the HBNI.

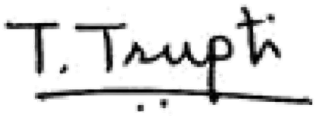
Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the Competent Authority of HBNI when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.



Trupti Ajay Togar

DECLARATION

I, hereby declare that the investigation presented in the thesis has been carried out by me. The work is original and has not been submitted earlier as a whole or in part for a degree / diploma at this or any other Institution / University.



Trupti Ajay Togar

List of Publications arising from the thesis

Journal

1. **Togar, T**., Desai, S., Mishra, R., Terwadkar, P., Ramteke, M., Ranjan, M., . . . Dutt, A. (2020). Identifying cancer driver genes from functional genomics screens. *Swiss Med Wkly, 150,* w20195. doi:10.4414/smw.2020.20195 (Thesis work)

Chapters in books and lectures notes

N/A

Conferences

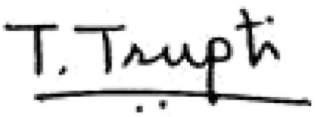
1. **T. Togar**, S Chopra, S Desai, N Gardi, B Dharavath, B Sahoo, H Kore, S Gupta, A Dutt: Genomic Characterization of Somatic Alterations in Cervical Adenocarcinoma, 4th EACR conference- Cancer Genomics (2019), Cambridge, UK (Poster presentation)
2. **T. Togar**, S Chopra, S Desai, N Gardi, B Dharavath, B Sahoo, H Kore, S Gupta, A Dutt: Genomic Characterization of Somatic Alterations in Cervical Cancer, NGBT conference (2018), Jaipur, India. (Poster presentation)
3. **T. Togar**, S Chopra, N Gardi, B Dharavath, B Sahoo, H Kore, S Gupta, A Dutt: Genomic Characterization of Somatic Alterations in Cervical Cancer,37th IACR Convention (2018), Kolkata, India. (Poster presentation)
4. **T. Togar**, N Gardi, B Sahoo, P Upadhyay, S Chopra, A Dutt: Genomic Characterization of Somatic Alterations in Cervical Adenocarcinoma, NGBT conference (2016), Cochin, India. (Poster presentation)
5. **T. Togar**, N Gardi, P Upadhyay, S Chopra, A Dutt: Genomic Characterization of Somatic Alterations in Cervical Adenocarcinoma, Tata Memorial Centre Platinum

Jubilee Conference (2016) in Mumbai, India. (Poster presentation)

6. **T. Togar**, N Gardi, P Upadhyay, S Chopra, A Dutt: Genomic Characterization of Somatic Alterations in Cervical Adenocarcinoma ,1st MOSCON conference (2016), Pune, India. (Poster presentation)

Others

1. Godbole, M., **Togar, T**., Patel, K., Dharavath, B., Yadav, N., Janjuha, S., . . . Dutt, A. (2018). Up-regulation of the kinase gene SGK1 by progesterone activates the AP-1-NDRG1 axis in both PR-positive and -negative breast cancer cells. J Biol Chem, 293(50), 19263-19276. doi:10.1074/jbc.RA118.002894
2. Upadhyay, P., Gardi, N., Desai, S., Sahoo, B., Singh, A., **Togar, T**., . . . Dutt, A. (2016). TMC-SNPdb: an Indian germline variant database derived from whole exome sequences. Database (Oxford), 2016. doi:10.1093/database/baw104



Trupti Ajay Togar

ACKNOWLEDGEMENTS

My Ph.D journey would not have been possible without the strong support from mentors, colleagues, friends and lab-mates. I take this opportunity to express my heartfelt gratitude to everyone who was a part of this journey.

First and foremost, I would like to express my sincere gratitude to my research advisor, Dr. Amit Dutt for giving me an opportunity to work on interesting projects during the course of my Ph.D tenure. His continuous guidance and constructive suggestions has helped me to improve my research and scientific writing.

I would sincerely like to thank my doctoral committee members Dr. Sorab N. Dalal (chairperson), Dr. Prasanna Venkatraman and Dr. Tanuja Teni for their critical advice, insightful suggestions and for extending their continuous support. The detailed discussions during DC meetings has helped me shape the project and progress in the right direction. I'm thankful to our clinical collaborator, Dr. Supriya Chopra for providing patient samples and help with clinical analysis. Her insights from clinical perspective were useful for data interpretation and have widened my understanding.

I would also like to thank Dr. Sudeep Gupta (Director, ACTREC), Dr. Rajendra Badwe (Director, TMH), Dr. Subhada Chiplunkar (Ex-director, ACTREC), Dr. Prasanna Venkatraman (Deputy Director, ACTREC) for providing infrastructure and facilities at ACTREC and TMH to conduct research. I thank Department of Biotechnology (DBT) for providing me fellowship during Ph.D, Terry fox foundation and DBT for funding the projects. I also appreciate the financial help provided by Homi Baba National Institute (HBNI) and Sam-Mistry foundation (TMC) for attending international conference.

I was fortunate to collaborate with several lab members on different projects and got an opportunity to learn new things from each one of them. I would like to thank my current lab members Sanket, Asim, Bhaskar, Neelima, Suhail, Supriya, Aniket, Dr.Ashwin, Shailesh sir, Deepak, Rohit, Aishwarya, Sonal and Dr. Archana for their invaluable assistance, continuous support and encouragement. I express deep sense of gratitude to my previous lab seniors Dr. Manoj, Dr. Hemant, Dr. Jyoti, Dr. Pawan, Dr. Pratik, Dr. Prajish, Dr. Mukul, Nilesh, and Rohan for establishing a strong lab foundation and Sunil, Ankita, Vidya and Mayur for being supportive friends. Special thanks to Dr. Manoj, Dr. Pawan and Dr. Prajish for personally guiding me on my projects in lab during the start of my Ph.D journey. I also appreciate the help rendered to me by the bioinformatics team members Sanket, Bikram, Hitesh, Aniket and Rohit for several project related analysis. I'm grateful to my trainees Roma, Apoorva and Ankita, for their tremendous help with the functional work and previous lab JRFs Prachi and Malika for their contribution to my projects. I thank Mr. Dhananjay and Ratnam for taking care of the IRB related work and for assuring smooth functioning of the lab. I also take this opportunity to thank Dr. Shilpee and her lab members for their valuable suggestions during common lab meetings.

I thank my batchmates and friends Jyothi, Usha, Shalini, Harish, Akash, Ajit, Saim, Mukund, Maitreyi and Arijit for being a continuous source of support. They were always available for any help with respect to sharing of reagents, protocols or for detailed discussions. Organizing NRSM conference with the batch mates was indeed a wonderful and learning experience. Special thanks to my friends Jyothi and Usha for providing me emotional support in times of difficulty. Thanks to ACTREC student community for being understanding, co-operative and helpful.

I wish to acknowledge the help provided by ACTREC staff. First, I would like to thank Mr. Dandekar and other staff members Mr. Chauhan, Mr. Burate, Mr. Kulkarni for maintaining and assuring smooth functioning of all the equipments and instruments of common facility. Thanks are due to Mrs. Sharda and Mr. Naren from genomics facility, Mrs. Tanuja and Mr. Jairaj of Microscopy facility and Mrs. Shamal, Mr. Ravindra, and Mr. Prasanth from FACS facility for their assistance. I would like to thank Dr. Rahul Thorat for help with animal experiments and Bhabhani for *in-vivo* imaging.

My Ph.D journey would be incomplete with thanking my previous research advisors Dr. D.V. Amla, Dr. P.B. Khare, Dr. T. Ramakrishna Murti and Dr. Shailly Tomar who introduced me to basic research and encouraged me to continue in this field. These short-term research projects enabled me to conduct research in labs working on different topics and thus, helped me develop scientific temperament.

Last but not the least, I thank my parents and my brother who supported me in all my decisions of life and continue to do so. It was their immense moral support, understanding, patience and encouragement that gave me strength to face every obstacle in life. They were besides me during my ups and downs in these six years of Ph.D

Once again, I thank all for making my Ph.D journey a memorable one.

Chapter VI

Summary and Conclusion

Chapter VI- Summary and Conclusions

Cancer is a genetic disease defined by several genomic alterations like mutations, gene expression changes, copy number alterations, epigenetic changes and structural variations. However, cancer is driven by only a few genetic alterations termed as driver genes whereas several other alterations that do not contribute to disease progression are termed as passenger alterations. Targeting of driver genes in cancer cells results in decreased cellular proliferation and viability, a phenomenon described as oncogene addiction. This is the basis of targeted therapy or precision medicine, wherein a patient's unique genomic profile is considered for deciding treatment and outcome. Targeted therapy has been successfully implemented in the clinical setting for several cancer types and yielded beneficial results in controlling the disease. Based on the concept of identifying gene targets for precision medicine or targeted therapy, two study approaches drive this thesis- one conceptual and other technical. The first approach focuses on integrated genomic approaches to identify driver alterations from cancer genomes and the second approach deals with functional genomics using pooled RNAi screen to predict therapeutically relevant driver alterations for targeted therapy. The studies were performed in two different cancer types.

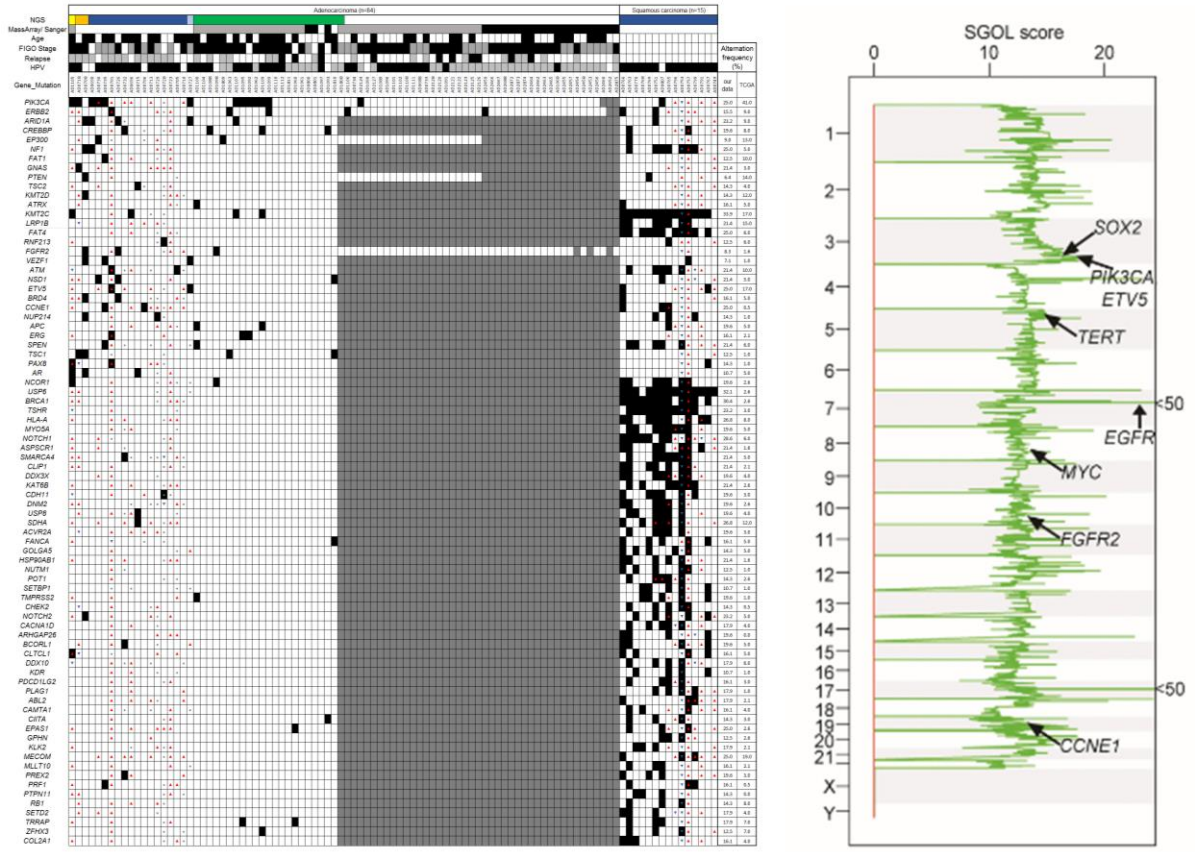
Comprehensive genomics efforts were undertaken to characterize the significantly altered mutations, expressed transcripts and structural variants underlying the cervical cancer genome. Several known and other cancer-associated genes were identified in this study.

Firstly, extensive genomic profiling for mutations was performed in 84 samples of cervical adenocarcinoma and 15 samples of squamous carcinoma using NGS approach and other genotyping methods to provide a landscape of somatic mutations in cervical cancer from the Indian population. Here, we report mutations in known hallmark gene - *PIK3CA*, *ERBB2*, *ARID1A*, *CREBBP*, *EP300*, *NF1*, *FAT1*, *PTEN* and *TSC2* and novel cancer-associated genes

FGFR2 and *AKT1*. In addition, mutations in epigenetic gene include *KMT2C*, *KMT2D*, *EP300*, *BRD3*, *BRD4*, *NSD1* and *PBRM1*. However, we did not observe mutations in *KRAS*, *STK11*, *FBXW7* and *TP53* genes, which are commonly mutated in cervical cancers as reported by TCGA group [94].

Secondly, copy number variation analysis from WES and WGS samples show recurrent copy gains in genes in *PIK3CA* (37%), *SOX2* (37%), *TERT* (33%), *ERBB2* (30%), *KRAS* (26%), *MYC* (22%) and *BRCA1* (22%), consistent with literature reports. Amplifications are also observed in other cancer-associated tyrosine kinases like *ERBB3* (22%), *ERBB4* (15%), *EGFR* (15%), *FGFR2* (15%), *FGFR3* (7%). In addition, we note copy gain and loss in known oncogenic fusion gene partners *FGFR3-TACC3*, *TMPRSS2-ERG* and *EML4-ALK*, also observed in the TCGA dataset. From WGS dataset, 14 broad-arm level amplifications, 5 broad arm deletions, 221 focal amplification and 31 focal deletions were predicted. Recurrent amplifications are observed at chromosome 1q, 3q, 8q, 11p, 17q, 19q, 20q, 5p, 9q, 1p, 11q, 20p and 9p and recurrent deletions at chromosome 3p, 4q, 11p, 11q, 18q, 19p, 2q and 5q.

The integrated mutation and copy number alterations in cervical cancer hallmark genes and other cancer genes along with CNV plot is shown in the figure below. Black box indicates mutation, red and blue triangle indicates copy gain and loss respectively.



Third, gene expression analysis was performed within tumor samples. Gene expressed in top 10% quartile and recurrent in at least 30% of the samples were considered further. We observe over-expression of *EGFR* (57%), *ERBB2* (81%), *ERBB3* (90%), *MET* (38%), *AKT1* (38%) and *AKT2* (90%). Increased expression *MMP2*, *MMP12* and *MMP14* has been reported in cervical cancer [179-181] and also observed in our dataset. Next, we identified expressed gene fusions with one of the genes with oncogenic function - *IDH3G-PPP2R1A*, *U2AF1-CASP2*, *RAP2A-MECOM*, *PPP6C-CASC3* and *ANKRD27-MYC* from fusion analysis. In addition, we report in-frame fusions with kinase gene partner such as *PKM-FUT2*, *PKM-CBX4*, *STK24-ZNF585A* and *CDK16-CAP1* with conserved domains. All the fusions observed are novel and not reported in the literature.

Fourth, we report several structural variants identified from WGS data. *ARHGAP11B-ARHGAP11A* and *CDK11B-SLC35E2B* were recurrent in two samples. Two samples show

structural re-arrangements *CAPN8-MUC3B* and *PNKD-MUC3B* have MUC3B as one of the gene partners. Gene translocation pairs *FAM172A-**ESR, DUX4**-ROCK1P1**, **MLLT4**-KIF25**, **MSH2**-TAF4**, **PAX7**-EEF1A2**, **PBX1**-SIK3* and ***PDGFRA**-MAN2A1* involves one of the partners known to be a cancer gene. All the genomic rearrangements reported from the study are unique and not reported in the literature for any cancer type.

Overall from the genomic studies, we identify several therapeutically relevant alterations in cervical cancer. We observe that most of the mutations in genes converge onto PI3K/AKT and MAPK pathway. Recurrent mutations of *PIK3CA* in the helical domain E545K and E542K are targetable using alpelisib and fulvestrant [144], *ERBB2* D769Y, S310F/Y by trastuzumab, lapatinib or neratinib [145], *FGFR2* K659E, S320F and C382R by Ponatinib and BGJ398 [137, 147].

Most of the mutations, amplification and over-expression of genes were common among the ERBB family members. Therefore, the role of ERBB signalling in cervical cancer was investigated using *in-vitro* and *in-vivo* approaches. Cervical cells subjected to Afatinib treatment revealed that C33A cells were sensitive to treatment as compared to other cells. This observation was consistent in the *in-vivo* studies, wherein mice with C33A tumors showed a delay in tumor growth on Afatinib treatment as compared to control group. Next, to identify Afatinib targets- EGFR, ERBB2 and ERBB4 conferring oncogenic dependency in C33A cells, individual gene knockdown by shRNA was performed in C33A and SiHa cells. Although a slight decrease in the p-MAPK was observed upon depletion of EGFR and ERBB2, the knockdown cells did not show reduced cellular proliferation, migration and anchorage-independent growth suggesting that cells are not dependent on ERBB2 or EGFR for growth and survival. These results indicate that there is a possible role of change in receptor heterodimerization upon depletion of one ERBB member or cross-talk with other pathways such as PI3K/AKT [113, 114], which is facilitating the continuation of signalling

in the cells. This needs to be further investigated. Our results also point to the fact that it is essential to target all ERBB receptors simultaneously as done by afatinib inhibitor to reduce cell growth since redundant gene functions are carried out by other receptors upon inhibition of one receptor. These findings are consistent with earlier reports in cervical cancer which shows that pan-ERBB inhibition by inhibitors Lapatinib and AST1306 display effectiveness in reducing proliferation in C33A cervical cells [110].

This study overall validates the current understanding of cervical cancer genomics and also extends our understanding of cervical cancer, especially the adenocarcinoma subtype and provides a detailed comprehensive landscape of somatic alterations from the Indian ethnicity for the first time to identify suitable molecular targets for precision medicine.

In addition, we have taken a complementary approach to establish the significance of a functional genomics approach to identify therapeutically relevant driver alterations using cells derived from HNSCC as a model system. We performed a pooled RNAi shRNA screen against human kinases in HNSCC cell line AW13516. To predict potential driver genes with high confidence, the RNAi screen data was integrated with genomics data of gene expression and copy number changes. Such an approach has been previously used by several groups. However, currently available data integration tools which combine RNAi data with genomics data, require intense computational processing and expertise and hence, of restricted use to a functional biologist. Here, we develop a simplified scoring system 'DepRanker' which integrates genomic data like gene expression, copy number and RNAi output data like depleted gene list and individual shRNA depletion list to assign scores for calculating a Rank Impact Score (RIS) [237]. Genes with high RIS are predicted to be potential cancer drivers. An input of RNAi data along with genomics data fed to DepRanker was able to predict AURKB and TK1 as drivers. To validate findings, TK1 knockdown was performed in AW13516 cells and it was observed that cell proliferation was inhibited in

knockdown clones as compared to control cells. In addition, AW13516 cells also exhibited sensitivity to AURKB inhibitor AZD1152-HQPA. Both the genes play an important role in regulating cell cycle and cell division and can act as attractive therapeutic targets for targeted therapy in clinics for head and neck cancer type.

DepRanker has a wider application in predicting cancer essential genes for other RNAi and CRISPR screen datasets as well. We provide a user-friendly GUI which can be used by a functional biologist by providing input data in the required format to identify genes showing oncogenic dependency in cancer cells.

Although we performed a pooled shRNA screen in AW13516 cells and used DepRanker to predict cancer essential genes by integrating genomics data, this study suffers from several limitations. The pooled screen was restricted to human kinases, therefore other non-kinase driver genes remain undetected. The screen was performed in triplicates, of which data from two screens were captured at lower coverage. With recent advances, pooled CRISPR screens offer better advantage than pooled RNAi screens and it is more specific and sensitive in predicting cancer essential genes [238]. With CRISPR screens, fewer variations are observed across the replicates and complete gene function perturbation is seen due to knockout [239]. Nevertheless, this study presents a proof-of-principle approach for validation of functional genomics using pooled RNAi screen against human kinases to identify therapeutic gene targets. We identified *AURKB* and *TK1* as gene targets with therapeutic relevance in the treatment of head and neck cancer patients. In addition, we present a simplified scoring system 'DepRanker' which can be readily used by a functional biologist to analyze pooled screen data and obtain useful insights in predicting essentiality genes in cancer cells [237].

Thesis Abstract:

Name : Ms. Trupti Togar

Enrollment Number: LIFE09201404002

Thesis Title: Genome-wide approaches to characterize novel genetic elements causing cancer

The thesis focuses of describing the genomic landscape of somatic alterations in cervical cancer, particularly adenocarcinoma subtype using NGS and other genotyping based approaches. The study identified recurrent somatic mutations in *PIK3CA, ERBB2, FGFR2* and *ARID1A*, copy number gain in *PIK3CA, KRAS, ERBB2, EGFR, AKT2* and over-expression of MMP, AKT, ERBB family members providing known and novel therapeutic targets which can be considered for treatment of cervical cancer. Moreover, we report novel gene fusion transcripts, gene translocation events and information of HPV integration sites in adenocarcinomas samples. Based on recurrent alterations observed in the ERBB-MAPK pathway members, the role of ERBB signaling in cervical carcinogenesis was investigated using *in-vitro* and *in-vivo* approaches. Pan-ERBB inhibitor afatinib treated C33A cells show decrease cell proliferation and a delay in tumor growth. But, individual knockdown of *EGFR* and *ERBB2* has no effect in impeding cell proliferation, migration or anchorage independent growth suggesting that the cells are not dependent on single receptor activity for survival and pan inhibition with afatinib inhibitor is required to attenuate MAPK signaling and promote cell death.

Identification of cancer essential genes is possible by means of a pooled RNAi screen. We performed a pooled shRNA kinome screen in AW13516 head and neck cancer cells. Data analysis was performed using edgeR package and data was further integrated with available genomics data of gene expression and copy number using an in house developed scoring system 'DepRanker'. It is an easy to tool for a functional biologist, who needs to provide raw data files and genomic data files in the required format and DepRanker predicts cancer essential genes conferring cell survival advantage. The predicted genes *AURKB* and *TK1*, upon depletion by inhibitor or knockdown approach has resulted in decreased cellular proliferation, thus confirming the essentiality of the genes in cell survival.

The thesis is based on utilizing different approaches to identify therapeutic genes in cancer that can be targeted using small molecule inhibitor to improve patient treatment outcome.

Thesis highlights

Name of the student: Trupti Togar

Name of the CI/OCC: Cancer Research Institute (CRI) **Enrolment no.:** LIFE09201404002

Thesis title: Genome-wide approaches to characterize novel genetic elements causing cancer

Discipline: Life Sciences **Sub area of discipline:** Cancer genomics

Date of Viva-Voce: 06/01/2021

Pooled RNAi screen is a robust approach to identify cancer essential genes in a cell line. Several data analysis tools are also available for processing the RNAi data to provide a list of potential oncogenes predicted from the analysis. However, prioritization of these several genes to identify few cancer essential cells showing oncogenic dependency in the cell lines is a difficulty. Few studies report the integration of data for different genomic features such as mutation, gene expression and copy number changes along with pooled RNAi data output to predict the cancer dependent genes. But, these integration tools require installation of several bioinformatics analysis packages and require computational expertise to analyze the data, which is a limitation for a functional biologist. To overcome this problem, we developed a simple scoring system 'DepRanker' which integrates the genomic data and RNAi data analysed using edgeR package to predict cancer essential genes and is available as a GUI. DepRanker operates in two modules. Module 1 analyses pooled RNAi data where the user provides input files in the required format for edgeR analysis and module 2 takes the output of module 1 and integrates with the genomic data for the same cell line provided by the user to predict cancer essential genes by calculating a Rank Impact Score (RIS) for each gene which is a cumulative score of all the genomic features and RNAi data score. Genes with high RIS values are potential oncogenes conferring cell survival. This is an easy to use tool for a functional biologist to draw insights into their own RNAi data to predict cancer dependent genes.

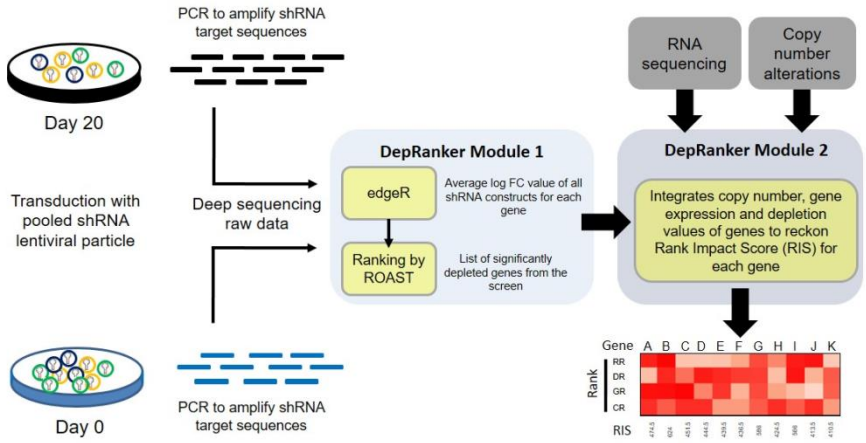


Figure: *Schematic outline depicting work flow of pooled shRNA data processing and gene prioritization in DepRanker. RR- Roast Rank, DR: Depletion Rank, GR: Gene expression Rank, CR: Copy number alteration Rank, FC: Fold change.*

Chapter I

Introduction and Review of Literature

Chapter I: Introduction and review of literature

1.1 Cancer a genomic disease

Cancer is a deadly disease globally with an estimated incidence of 18.1 million cases and 9.6 million deaths occurring annually. Cancer can be defined as a genetic disease arising from the alterations in genome and epigenome which promote abnormal cellular proliferation [1]. Genetic changes comprise point mutations, deletion, amplification and chromosomal translocations that alter the gene function allowing cancer cells to acquire certain phenotypic changes to sustain cellular proliferation and survival. Epigenetic alterations include hypomethylation, hypermethylation and deacetylation that alter the chromatin structure resulting in abnormal gene expression [2]. In simple terms, cancer can be explained as interplay of oncogenes and tumor suppressor gene functions. A gain of function mutations in proto-oncogenes promote cell proliferation and survival whereas as loss of function mutations in tumor suppressor genes result in abnormal cellular proliferation, thus contributing to activation of signalling pathway leading to transformation and cancer progression [3]. Not all mutations or alterations cause malignant transformation of cells. Genetic changes in genes that give a selective growth advantage are positively selected in the tissue micro-environment and can induce transformation are termed as ‘drivers’ whereas mutations or alterations that do not contribute to clonal growth advantage and have no role in disease progression are referred to as ‘Passengers’ [4, 5]. Distinguishing drivers from passengers is one of the challenges in cancer genomics. There are different ways to discern between driver and passenger alterations as reported by several studies. Driver genes are observed to be mutated in a large number of cancer samples than expected background mutations. Next, the oncogenic role of driver gene is predicted based on whether a change in gene activity is observed which is likely to impact the protein function. As per the studies,

about 5 to 8 driver mutations operate for cancer progression whereas the number of passenger mutations exceeds far more than drivers [6]. Although passenger mutations have been assumed to be neutral, growing supporting evidence is coming up that suggest passenger to be deleterious to cancer cells and has a role in predicting clinical outcomes [7]. Passengers were found to reduce cancer fitness by slowing tumor growth and metastasis, thus overcoming benefit induced by drivers [8].

Conventional cancer treatment involves surgical resection of tumor tissue, radiation and chemotherapy which targets fast-growing dividing cells of the tumor. In this case, the tumor surrounding normal non-cancerous cells and also other fast dividing normal cells are affected [9]. Conventional therapy is accompanied by several side effects [10]. Therefore, a more specific treatment approach such as precision medicine is required. Cancer therapy targeting specific driver genes cripple the oncogenic activity and control cancer growth forms the rationale of precision medicine. Precision medicine often involves fewer side effects as precise genetic changes of a patient's tumor are targeted to treat cancer [11]. Such genetic changes also enable patients to be classified into subpopulations that are more likely to respond to the targeted treatment designed against the genetic alteration [12]. Targeted therapy targets cell surface receptors, growth factors and signal transduction pathways using small molecule inhibitors, monoclonal antibodies, pro-drug and nanoparticulate antibody conjugates [13] to inhibit cellular proliferation, metastasis and disease progression [14]. Large scale sequencing of tumors has enabled the identification of clinically actionable gene alterations for therapy [15]. Few examples of successful clinical application of gene-targeted therapy are mentioned in the table below.

Gene target	FDA approved inhibitors/antibodies	Cancer types
HER2	Trastuzumab	Breast cancer [10]
BRAF	Vemurafenib	Melanoma [16]
EGFR	erlotinib	Non-small cell lung cancer [17]
ALK	crizotinib	Non-small cell lung cancer [17]
BCR-ABL	Imatinib	Chronic myeloid leukaemia [18].
EGFR/HER2	lapatinib	Breast cancer [19, 20]
VEGFR	Sorafenib	Renal cancer, kidney, liver, thyroid [19]
EGFR, HER2	Afatinib	Lung cancer [19, 21]
PARP	Olaparib	Ovarian cancer [22]
ESR1, ESR2	Tamoxifen	Breast cancer [19]
VEGF	Bevacizumab	Colorectal cancer; Lung cancer; Brain cancer; Kidney cancer [19], cervical cancer [23]
EGFR	Cetuximab	Head and neck cancer; Colorectal cancer [19]
mTOR	Everolimus	Breast cancer; Brain cancer; Kidney cancer; Pancreatic cancer [19]

I-Table I: Examples of targeted therapy used in clinics

Targeted therapies work on the principles of oncogene addiction. Oncogene addiction is defined as a dependency of cancer cells on a single oncogene to sustain malignant phenotype. Inactivation of a single oncogene result in inhibition of growth and survival of cancer cells [24]. Acute inactivation of an activated oncogenic protein and thereby repression of activated signalling pathway by targeted therapy has shown significant response in tumor shrinkage [25]. The important molecular targets involve cell surface receptors, gene transcription factors, proteasomes, hormonal factors, angiogenesis and immune cell targets [21, 26]. Among these, protein kinases are important molecular targets for targeted therapy and are often considered as Achilles' heel in several cancer types.

One of the growing advances in the field of precision medicine is immunotherapy. Immunotherapy strategies employ antibodies, checkpoint inhibitors and recombinant

proteins to activate the host immune system to fight against cancer [27]. Adoptive T-cell transfer therapy employs patients own T lymphocyte cells with anti-tumor activity by expanding T-cells *in-vitro* and infusing back into the patients. This therapy has proven successful in metastatic melanoma patients where 50% of the patients showed tumor regression [28]. Chimeric antigen receptor T cells (CART) therapy is an upcoming promising therapeutic option in which patient T cells are modified to recognize cancer-specific antigen along with stimulation with intracellular signals to enhance T cell responses when T cells are infused back into the patients [29].

1.2 Next-generation sequencing to capture somatic alterations in Cancer

Various next-generation sequencing approaches are available to identify different somatic alterations in cancer. For capturing somatic mutations, whole-exome sequencing (WES) is a preferred method if variants from the coding region are to be assessed. Low sequencing cost and ease of data processing are the advantages of using WES [30]. However, WES has problems associated with the capture of the entire coding region of the genome, which is better captured by WES [31]. Also, non-coding genomic alterations often go undetected in WES due to the nature of capturing only the exonic regions. However, because of the high cost of sequencing and intense computational processing for data analysis in WGS [30], WES remains a preferred choice for identifying somatic variants. WGS enables reliable detection of genome-wide copy number alterations, gene translocation events, pathogen detection and variant detection in coding and non-coding region [32] as compared to other NGS method. Whole transcriptome sequencing (WTS) basically provide information of expressed transcripts and expressed genes fusions in a tumor [33] and was rarely used for variant calling. However, recent advances in the computational analysis enable variant calling from transcriptome data by performing certain filtrations to get rid of errors introduced during library preparation method from RNA [34]. Transcriptome sequencing

facilitates the identification of coding and non-coding variants that are expressed in tumors [35]. Driver mutations tend to occur in expressed genes and are usually conserved [36]. Variants detected from transcriptome data can also be used to validate and confirm the findings of WES and WGS.

None of the NGS approaches is adequate to capture all the genomic alterations in a cancer type, thus multi-platform NGS analysis with WES, WGS and WTS data would help in providing comprehensive information and allow classification of variants accordingly for deciding treatment [37]. Therefore, it is wise to extract information from an integrated analysis to discover underlying somatic alterations driving carcinogenesis and identify therapeutically relevant gene alterations.

1.3 Genomics and functional genomics approach to identify novel therapeutic gene targets in cancer

With the advent of NGS technologies, our understanding of cancer genomes has improved drastically. Sequencing of thousands of patient tumor samples from different cancer types has led to the identification of driver mutations and other alterations driving the disease progression [38]. With an in-depth knowledge of somatic alterations and data integration at multi-omics levels such as genomics, transcriptomics, epigenomics and proteomics along with clinical data, tumor heterogeneity can be studied to predict patient's response to targeted therapy [39]. Molecular profiling of genomic alterations is essential for the application of precision medicine in clinical settings [40]. Genome-based assays are now implemented in clinics so that cancer's genomic dependencies can be targeted by small-molecule inhibitors or antibody-based therapies [39].

Functional genomics is defined as the use of genomics data to decipher the function of gene and protein expression on a genome-wide scale utilizing high-throughput methods [41]. It involves the extraction of data from multi-omic platforms such as genomics,

transcriptomics, proteomics, epigenomics and metabolomics to identify interactions regulating biological processes [41, 42]. Thus, genomics and functional genomics studies provide useful insights on discovering potential driver genes involved in promoting carcinogenesis.

For the application of functional genomics to study phenotypic changes in cancer cells, two approaches are available. RNA interference (RNAi) screen uses siRNA or shRNA that cause gene knockdown affecting gene expression at a transcriptional level whereas CRISPR-Cas9 screen alters the genome at the intended site to generate knockout to alter gene function permanently [43]. RNAi and CRISPR-Cas9 screens can be set in an arrayed or pooled manner. In arrayed format, there is stable knockdown or knockout of individual genes whereas, in pooled format, large scale construct libraries are used to transduce the cells and then phenotypic changes are assessed [44]. With the advancement of NGS, pooled RNAi or CRISPR-Cas9 screens are robust approaches to identify potential driver genes conferring cancer cell dependency and survival. In RNAi screens, a library targeting global genes or a specific set of genes is applied to large cell population, which upon selection helps to identify the depleted and enriched cell populations over time by high throughput sequencing of constructs [45]. This loss of function genetic screens, where the loss of individual constructs (shRNA) due to cell death helps to identify potential oncogenic driver genes which could serve as therapeutic targets [46]. An exhaustive pooled RNAi screen study was performed at BROAD institute on 216 cancer cell lines to discover genes with role supporting cell proliferation and survival [47]. Similarly, pooled CRISPR-Cas9 based screens that cause genetic perturbation resulting in loss of gene function are now being regularly used to study malignant cell phenotypes. CRISPR-Cas9 dropout screen are useful in identification of driver genes in cancer cells [48] and genes conferring resistance to treatment [49]. Use of different screens depends upon question being addressed.

Although pooled screens offer a useful tool, there are several limitations. Reproducibility of heterogeneous data sets from a pooled screen is often a problem with pooled data due to noise produced by non-reproducible hits, contributed by factors such as random integration for stable expression, processing of shRNA hairpin and common off-targets effects [46, 50]. One way to overcome these limitations is to perform a secondary screen with top hits obtained from the primary screen and to access reproducible hits obtained with both screens. Another way to deal with these limitations, several robust computational approaches is employed [51, 52]. One such tool is DepMap which has analyzed 501 genome-wide loss of function screens in cancer cell lines to predict genetic vulnerabilities considering several genomic features [53]. The genomics data of mutation, copy number alteration and gene expression are taken into consideration to predict cancer essential genes by several groups using bioinformatics tools [53, 54]. However, the usage of all these tools is computationally intense and requires bioinformatics expertise. Therefore, a biologist finds it difficult to predict the cancer essential genes using these computational approaches and therefore, there is a need for a simplified scoring algorithm that can be readily utilized to predict cancer dependency genes. Once the genes are identified, targeted therapy with available small molecule inhibitors can be employed.

1.4 Introduction to cervical cancer

Cervical cancer is a common gynaecological cancer among women worldwide with an incidence of 5, 70,000 cases and about 311,000 deaths occurring annually. India contributes to 17% of the world incidence of 96,322 new cases of cervical cancer diagnosed in India each year and 60,078 deaths occur annually (Globocan, 2018). The five-year survival rate is 46% [55]. It is observed that metastasis develops in 15-61% of the women within 1-2 years on treatment completion [56]. The five-year survival rate is 16.5% for metastatic disease with a median survival of only 8- 10 months and 91.5% for localized cancer [57]. The

common metastatic site includes lung, para-aortic lymph node region, supraclavicular lymph node region, mediastinal lymph node region, bone and liver [58].

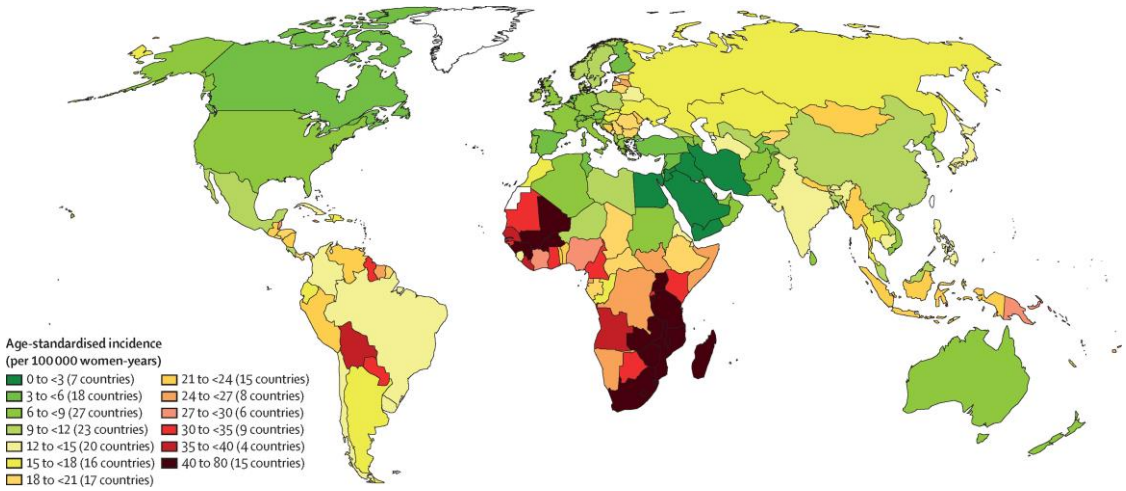
Cervical cancer can be classified into three histological subtypes namely squamous carcinoma, adenocarcinoma and adenosquamous carcinoma. Squamous carcinoma arises in the flat cells covering exocervix and accounts for 70-80% of the incidence rate. Adenocarcinoma which arises in the glandular cells of endocervix constitutes 10-20% of the incidence rate. Adenosquamous, a rarer subtype with an incidence rate of less than 1%, comprises both squamous and glandular cells [59, 60]. It has been observed that in the past 40 years, there has been a reduction in the incidence of squamous carcinoma due to regular cytological screening but increase in the incidences of adenocarcinoma [61] and is more commonly observed in younger women [62]. The most common etiological factor response for cervical carcinogenesis is infection with Human Papilloma Virus (HPV). Other factors include exogenous and environmental factors such as the use of oral contraceptives, tobacco smoking, diet, and infection with HIV and other sexually transmitted agents [63], host cofactors like hormones and immune response [64].

The clinical treatment includes brachytherapy and chemotherapy with agents like cisplatin, paclitaxel, and gemcitabine [65]. A difference in the treatment outcome has been observed where early-stage adenocarcinoma patients are 39% more likely to die from the disease than early-stage squamous counterparts [62] and adenocarcinoma patients had poor overall survival (OS) and disease-free survival (DFS) as compared to squamous carcinoma patients [66, 67] regardless of treatment modality.

1.4.1 Epidemiology of cervical cancer

Cervical cancer incidence and mortality rates are lower in high resource countries as compared to low resource countries. The high burden of disease with age-standardized incidence rate (ASIR) greater than 15 per 1,00,000 cases is reported in Africa, Melanesia,

Micronesia, southeastern Asia, eastern Europe, the Caribbean, and South America whereas ASIR> 40 per 100000 were observed in Zimbabwe, Tanzania, Burundi, Uganda, Lesotho, Madagascar, Comoros, Guinea, Burkina Faso, Mali, South Africa, and Mozambique, Malawi and Zambia with China and India's contributing accounting for 35% to the global incidence and mortality rates [68].



I-Figure 1: Global incidence of cervical cancer (Adopted from Arbyn, M., *et al.*, Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. Lancet Glob Health, 2020)

In low-income countries like India, factors such as inadequate screening programs, lack of awareness, poor hygiene and multiple pregnancies are additional contributors [69]. The screening in developing countries is only 19% as compared to developed countries wherein 63% of the population is screened [70]. According to National Cancer Registry Programme (NCRP), age-adjusted incidence rate is highest in Mizoram state (23.07/10000), Pasighat (22.54/10000) and lowest in Dibrugarh (4.91/10000). Population based cancer registries (PBCR) from Bangalore, Delhi, Bhopal Chennai and Barshi Rural has an age-adjusted incidence rate of 13-16 per 10000. 85% of the patients belong to the age group population 40 and above [55].

1.4.2 Etiology and risk factors associated with cervical cancer

1.4.2.1 Infection with oncogenic HPV viruses

The link between HPV infection and cervical cancer was first established by Zur Hausen in 1970s when the team isolated HPV16 and HPV18 from the cervical lesions [71]. Globally HPV infection is observed in 90-95% of the cervical cancer cases with a high frequency of HPV-16 (50%) followed by HPV-18 (12%), HPV-45 (8%), and HPV-31 (5%) [72]. Although more than 200 types of HPV are detected, HPV types can be grouped as high risk and low risk. High-risk HPV types comprises of HPV types 16, 18, 31, 33, 34, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68, and 70 whereas low-risk HPV types include types 6, 11, 42, 43, and 44 [73]. HPV types 16 and 18 are responsible for 76.6% cervical cases in India. HPV infects basal epithelium giving rise to histological changes that cause normal cervix tissue to progress to cancer. Pre-invasive lesions comprise of cervical intraepithelial neoplasia (CINs) namely CIN 1, CIN 2 and CIN 3 which eventually develops into carcinoma as cells invade the basement membrane. HPV mediated carcinogenesis is brought about by over-expression of E6 and E7 oncoproteins that are responsible for the degradation of tumor suppressor proteins p53 and Rb which result in uncontrolled cellular proliferation [74, 75]. Expression of E6 and E7 oncoproteins are transcriptionally repressed by E2 gene in the episomal form. However, upon the integration of HPV in the human genome at E2 site, E6 and E7 are oncoproteins are expressed [76] which is responsible for the progression of the disease. Although HPV infection is essential, it is not sufficient to drive tumorigenesis and additional factors such as genomic alterations are required to induce transformation [73]. This observation is supported by the fact that certain cervical cancer cases are HPV negative where other factors are contributing to carcinogenesis. Vaccines protecting against high-risk HPV types as Gardasil, Gardasil9 and Cervarix are available to prevent HPV infection [77].

But inadequate or lack of awareness programs, high cost and accessibility to the vaccine has been a major hindrance for conducting a mass vaccination program in India [78].

1.4.2.2 Tobacco Smoking

Several epidemiological studies suggest smoking is a risk factor associated with cervical cancer. Twelve studies in cervical carcinoma comprising of 8097 squamous carcinomas, 1374 adenocarcinoma and 26445 women without cancer suggested a strong association of tobacco smoking with increased risk of squamous carcinoma but not adenocarcinoma [79]. Increased risk of prevalent HPV infection was observed in heavy smokers as compared to non-smokers [80]. Another study on 7129 subjects also supported this association, high-risk HPV positive women who are smokers are at increased risk for developing CIN 3 and CIN3+ [81]. Women, who quit smoking at least 10 years before, have half the risk of developing CIN3/CIS and Invasive Cervical Cancer (ICC) [82]. There are several explanations of how smoking impacts carcinogenesis. Carcinogens like polycyclic aromatic hydrocarbons potential cause suppression of the immune system. Nicotine and cotinine have been detected in cervical mucus [80]. Exposure of carcinogens with HPV infected cells contribute to DNA damage in addition to damage induced by HPV oncoproteins resulting in cell cycle arrest and apoptosis inhibition [83].

1.4.2.3 Other factors

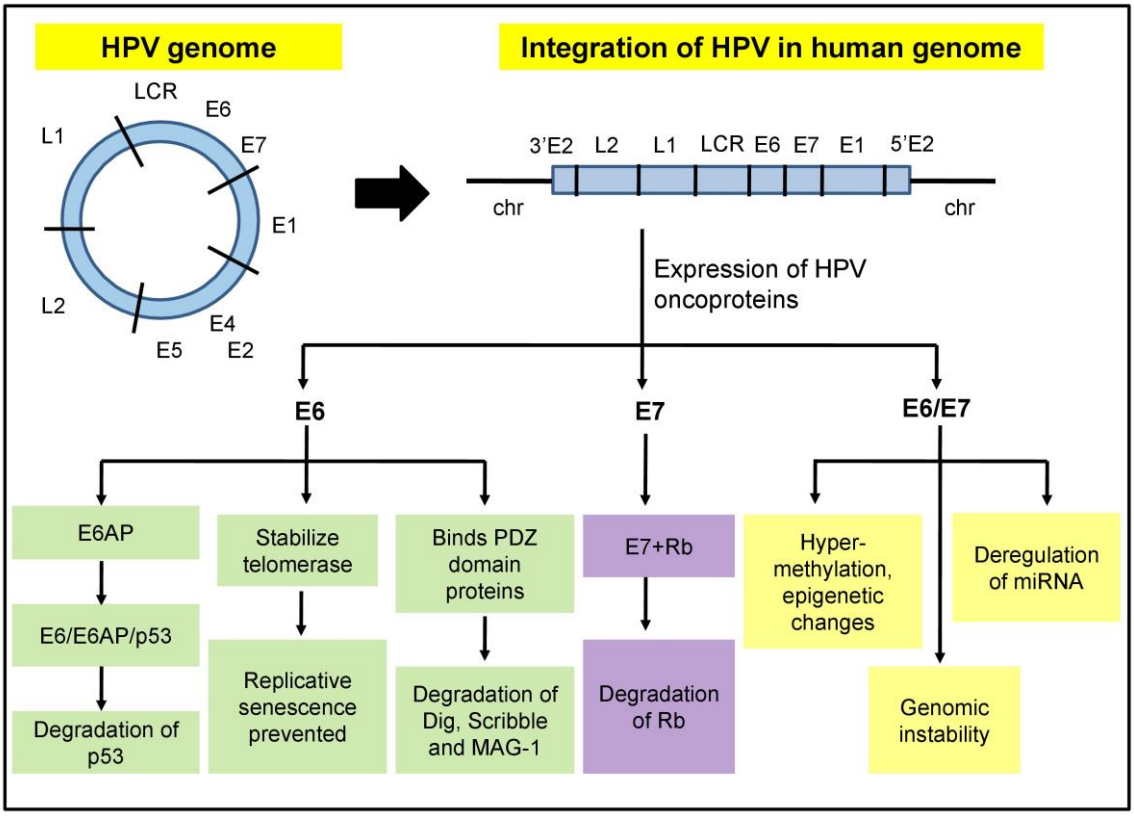
Early marriage, poor sanitary conditions, high parity, abortions, use of oral contraceptives, low socio-economic status and low education are reported to be significant risk factors responsible for cervical carcinogenesis [63]. Lack of knowledge about routine screening by Pap-smear or VIA test, lack of trained cytologists and limited infrastructure for screening often results in patients with disease presentation at late stages especially in rural areas [84].

1.4.3 Disease progression and staging

Premalignant transformation occurs in the squamocolumnar junction of the cervix which is brought about HPV infection [85]. Cervical Intraepithelial Neoplasia (CIN) refers to dysplasia of squamous cells in the cervical epithelium [86]. The pre-cancerous CIN can be graded as follows. CIN1 refers to mild or low-grade dysplasia in the lower one-third of the epithelium, which can be cleared with help of the immune system. High-grade CIN2 stage is moderate dysplasia affecting two-third of the epithelial cells and CIN3 is severe dysplasia with more than two-third of the epithelium affected. High-grade CIN further proceed to develop Carcinoma In situ (CIS) and invasive cervical cancer. CIN2/CIN3 develops within 2-3 years of HPV infection and it takes 10-12 years for developing invasive cancer [86, 87]. Therefore, regular screening and early detection will enable early treatment and the disease progression can be controlled.

Several molecular changes through stage progression take place upon HPV infection. Upon integration of the HPV genome into the host genome, E2 disruption enables the expression of E6 and E7 viral oncogenes [88]. E6 binds to E6 associated binding protein (E6AP) which causes structural changes in E6 and promote binding to p53 tumor suppressor protein resulting in the p53 degradation. Also, E6 also stimulates telomerase (TERT) activity that results in cell immortalization. E6 has also been reported to bind to proteins with PDZ sites such as Dig, MAGI-1 and Scribble and facilitate their degradation of these tumor suppressors [89]. E7 binds to pRB and promotes degradation. pRB is known to down-regulate E2F, a transcription factor promoting cell proliferation [90]. E6/E7 together brings about several epigenetic changes which include methylation of the viral genome, promote DNA methyltransferases to hypermethylate CpG promoter sites of tumor suppressor genes. Moreover, E6 and E7 expression modulates host miRNA which enables proliferation, migration and invasion of cancer cells, and promotes HPV amplification [89]. Increase in

methylation of genes such as *DAPK1*, *RARB*, *TIMP3*, *CCNA*, and *FHIT* has been reported to be associated with cervical cancer [91]. Most of these molecular changes and cellular transformation are brought about by high-risk HPV E6 and E7 proteins.



I-Figure 2: Molecular changes involved in cervical carcinogenesis upon HPV genome integration in the human host genome. (Adapted from Chan, C.K., *et al*., Human Papillomavirus Infection and Cervical Cancer: Epidemiology, Screening, and Vaccination-Review of Current Perspectives. J Oncol,2019) Upon disruption of the E2 gene E6 and E7 oncoproteins are expressed, which degrade tumor suppressor p53 and Rb and also cause several changes at epigenetic and miRNA level to promote cell proliferation, invasion and migration of cancer cells.

Cervical cancer patients are classified using the International Federation of Gynecology and Obstetrics (FIGO) classification system which was again revised in 2019 [92]. FIGO stage from IA, IB and IIA are early-stage disease whereas FIGO IIB, III and IV fall under advanced-stage disease category. Early-stage disease can be cured by surgery and

chemotherapy in 80-95% of the patients whereas late-stage III tumors have a lower cure rate of 60% [93].

Cancer spread	Stage	Substage	Description
Stage I confined to cervix uteri	IA	IA1	stromal invasion <3 mm in depth
		IA2	stromal invasion ≥3 mm and <5 mm in depth
	IB	IB1	≥5 mm depth of stromal invasion and <2 cm in greatest dimension
		IB2	≥2 cm and <4 cm in greatest dimension
		IB3	≥2 cm and <4 cm in greatest dimension
Stage II cancer spreads to the uterus but not into lower one-third of the cervix	IIA	IIA1	<4 cm in greatest dimension
		IIA2	≥4 cm in greatest dimension
	IIB	IIB	parametrial involvement but not up to the pelvic wall
Stage III involves lower one-third of the cervix and pelvic wall	III	IIIA	lower third of the vagina, with no extension to the pelvic wall
		IIIB	Extension to the pelvic wall and/or hydronephrosis or non-functioning kidney
		IIIC	IIIC1: Pelvic lymph node metastasis; IIIC2: Paraaortic lymph node metastasis
Stage IV cancer spreads beyond pelvis or mucosa of the bladder or rectum	IV	IVA	spread to adjacent organs
		IVB	spread to distant organs

I-Table 2: Classification of cervical cancer into FIGO stages based on morphological changes and disease spread.

1.5 Genomic Landscape of cervical cancer

Comprehensive genomic characterization of cervical cancer has been provided by TCGA and Ojesina *et al*., group [94, 95]. In total, the TCGA study analyzed 228 cervical cancer samples, describing mutation, copy number alteration, gene expression changes and structural variations using NGS approach. From this study, somatic mutations were recurrently observed in genes *PIK3CA, EP300, FBXW7, HLA-B, PTEN, NFE2L2, ARID1A, KRAS* and *MAPK1* along with mutations in novel genes *SHKBP1, ERBB3, CASP8, HLA-A* and *TGFBR2*. Recurrent amplification was observed at *EGFR, CD274, PDCD1LG2, KLF, BCAR4, TERC, MECOM, TP63, MYC, PVT1, YAP1,*

BIRC2, *BIRC3* and *ERBB2*. Recurrent deletions were observed in genes *TGFBR2*, *SMAD4*, *PTEN* and *EGFR*. In addition, integrating data from copy number, methylation, gene expression and miRNA, this study could classify cervical cancer samples as keratin high comprising of mostly squamous samples and keratin low consisting of adenocarcinoma samples. As the incidence of squamous carcinoma is high, this histological subtype is extensively characterized. But, very fewer reports are available for the low incidence of Adenocarcinoma subtype. One such study focusing on cervical adenocarcinoma subtype was first reported from a HongKong Chinese population. Whole-exome sequencing performed on 15 paired samples identified point mutations in genes *ARID1A* (20%), *FAT1* (4%), *PIK3CA* (20%) and *ERBB2* (2%). Recurrent copy number gain was identified in chromosome 1q, 3q, 8q, 11p, 17q, 19q, 20q and deletions in chromosome 11 and 16. Amplification of *ERBB2*, *PIK3CA* and deletion of *ARID1A* is reported [96].

Several of the other studies are based on targeted sequencing of few candidate genes. In such study, sixteen oncogenic genes were sequenced in 285 Chinese patients comprising of 179 squamous carcinoma samples, 62 adenocarcinomas, 34 adenosquamous and 10 with other histology. Recurrent mutations in *PIK3CA* (12.3%), *KRAS* (5.3%), *ERBB2* (3.5%), *FGFR2* (1.8%), *NRAS* (0.70%), *FGFR3* (0.7%) along with *FGFR3-TACC3* (3.9%) fusions were observed [97]. Another study conducted in 40 squamous carcinomas and 40 adenocarcinomas probing for 1250 mutations in 139 cancer-associated genes reveals recurrent mutations in oncogenes *PIK3CA* (31.3%), *KRAS* (8.8%), *EGFR* (3.8%) and tumor suppressor genes *STK11* (5%) and *PTEN* (1.2%). *KRAS* mutations were detected in adenocarcinoma subtype only [98]. Exome sequencing and targeted sequencing was performed by Luo. *et al.*, group in a Mexican population of cervical cancer in three histological subtypes comprising of 499 squamous, 67 adenocarcinoma and 21 adenosquamous samples. Frequent mutations were observed in *PIK3CA* (30%), *TP53* (5%),

HRAS (1.1%), *KRAS* (2.3%), *PTEN* (5.9%) and *STK11* (2.9%) and other genes *TSC1, BRCA1, BRCA2, BAP1, ATM, MAPK1*and *RB1* [99].

In comparison, fewer reports are describing the genomic landscape of cervical cancer from the Indian population. One study published in 2016, performed exome sequencing on 10 paired samples of cervical squamous carcinoma [100]. But no studies have been performed defining genomic profiling of rarer subtype adenocarcinoma from the Indian population.

1.6 Role of ERBB signalling in cervical cancer

ERBB family comprises of four receptors- *EGFR, ERBB2, ERBB3* and *ERBB4,* which are membrane-anchored proteins having extracellular ligand-binding domain, an intracellular tyrosine kinase domain and transmembrane domain. 11 known ligands can bind to ERBB receptors, except for ERBB2 [101]. The receptors are activated by binding of ligands which results in homo and heterodimerization leading to auto and transphosphorylation of tyrosine residues [102] leading to activation of downstream MAPK signalling, PI3K/AKT signalling, PLCγ/PKC, and JAK/STAT that play a role in promoting cellular proliferation, invasion, migration, differentiation and angiogenesis [101, 103]. ERBB2 lacks a known ligand; however, dimerization of ERBB2 with other members can activate signalling. ERBB3 has ligand binding site but lacks a functional kinase domain and hence can activate signalling upon hetero-dimerization with other ERBB receptor family members. Signalling cross-talk is also observed through ERBB3 which has 6 docking sites for p85 adaptor subunit of PI3K [104] resulting in activation of PI3K/AKT signalling. Somatic genomic alterations are commonly observed in *EGFR* and *ERBB2* genes in several cancer types. Mutations in *EGFR* are common in non-small cell lung cancer. EGFR exon 19 deletions are sensitive to targeted therapy like erlotinib and gefitinib and EGFR T790M to osimertinib [105]. Mutations in *ERBB2* S310F/Y, D769Y, and V777L are targeted by trastuzumab, neratinib and lapatinib [106] in breast cancer. *ERBB2* gene is commonly amplified in breast cancers [107].

Targeting of *EGFR* and *ERBB2* by targeted therapy is successfully carried out in clinics for several cancer types.

In cervical cancers also, mutations in the *ERBB2* and *ERBB3* gene are commonly observed [94, 95]. Mutations in the extracellular domain of *ERBB2* S310F/Y are commonly reported in cervical adenocarcinoma subtypes [106]. *EGFR* amplification is detected in cervical squamous carcinoma patients and is associated with reduced overall survival and *EGFR* over-expressing cervical cancer cells were found to be sensitive to inhibitor AG1478 [108]. In one study, PDX models developed using patient tumors overexpressing *HER2* showed a better response to treatment of trastuzumab and lapatinib [109]. A Phase 2 SUMMIT basket trial conducted with patients with metastatic cervical cancer harbouring *ERBB2* mutations S310F/Y, R678Q, D769N received neratinib orally showed better response. An exhaustive inhibitor study was performed on cervical cancer cell lines using EGFR, ERBB2, and pan-ERBB inhibitors like erlotinib, lapatinib and allitinib. C33A cervical carcinoma cells exhibited sensitivity to the inhibitors and SiHa cervical carcinoma cells were resistant to treatment and did not show any effect in cell phenotype-based assays [110]. Thus, *EGFR* and *ERBB2* serve as attractive targets for targeted therapy in cervical cancer.

All these studies are indicative of the emerging role of ERBB family in promoting tumorigenesis in cervical cancer. However, the role of individual members of the ERBB family and detailed study of the signaling pathway in carcinogenesis of cervical cancer has not been explored elaborately.

1.6.1 Failure of targeted therapy against single ERBB receptor

Sometimes targeting a single ERBB member by small molecule inhibitor does not attenuate MAPK signalling.. This is often observed because of functional redundancy as ERBB members share common downstream pathways, share common ligands, scaffold and adaptor proteins (SOS, GRB2, Shc) with multiple connectivities [111]. Literature reports have

suggested that depletion of one ERBB receptor causes increased expression of other receptor family members, thereby promoting cell survival [112] or a change in dimerization partners or interaction with novel dimerization partners, ligands or genes is observed to continue the signalling pathway [113, 114]. Also, a complex signalling cross-talk is observed between ERBB signalling with other signalling pathways such as PI3K/AKT pathway, NOTCH and NFκB pathways, which also needs to be targeted to sensitize cells to ERBB inhibitors [115, 116].

This is the reason probably why targeted therapy against a single gene is not successful in certain clinical cases. To overcome this problem, agents that disrupt the ERBB receptor interaction should be employed for treatment or the inhibitors that target the downstream signalling pathways should be used [117].

1.7 Research objectives

1.7.1 Rationale of the study

Based on the concept of identifying targets for precision medicine or targeted therapy, two study approaches drive this thesis- one conceptual and other technical. The first one is to discover therapeutic genomic alterations in cervical cancer using an integrated genomics approach and investigate signalling pathways involved in cervical carcinogenesis using *in-vitro* and *in-vivo* approaches. The second part involves functional genomics approach by performing a pooled RNAi screen to obtain cancer-specific gene targets for precision medicine and developing of a simplified scoring based system for use by a functional biologist to allow integration of available genomics data. This strategy has helped to identify cancer dependency genes which were further validated by functional assays.

1.7.2 Thesis objectives

Objective 1: Identification of oncogenic mutations and gene expression changes in cervical adenocarcinoma patients

Objective 2: Pooled shRNA screen to determine novel vulnerabilities using cancer cell lines

Chapter II

Genomic analysis to identify somatic mutations in cervical cancer

Chapter II: Genomic analysis to identify somatic mutations in cervical cancer

2.1 Abstract

Background: Cervical cancer is the second most cancer among women in India. Histologically, this cancer type can be classified as squamous carcinoma and adenocarcinoma. Squamous carcinoma has been extensively genomically characterized owing to the high incidence rate as compared to rarer adenocarcinoma subtype. Here, we describe the somatic mutation landscape in adenocarcinoma subtype from the Indian population for the first time and also compare with the somatic mutations identified from squamous carcinoma. This study aims to identify therapeutically relevant targets for precision medicine.

Material and methods: We performed whole-exome sequencing in 18 samples, RNA-sequencing in 24 tumor samples and whole-genome sequencing in 3 paired samples of cervical adenocarcinoma. Variant calling was performed in these samples using GATK workflow for WES, WGS and SNPiR in RNA-sequencing data. Mutations were further validated in additional samples using Mass Array genotyping and Sanger sequencing. Mutation profiling was performed in total of 84 cervical adenocarcinoma samples. To provide an overview of mutations observed in cervical adenocarcinoma and squamous carcinoma, WES data of cervical squamous carcinoma from previously published data were re-analyzed for variant calling in 15 samples.

For functionally validating the role of ERBB family receptors in cervical carcinogenesis, individual shRNA knockdown was performed for *EGFR*, *ERBB2* and *ERBB4* followed by western blotting to assess depletion. Further, cell-based phenotypic changes were assessed by migration, cell proliferation and anchorage-independent growth assay. To complement

in-vitro studies, *in-vivo* assay were performed in female NOD-SCID using pan-ERBB afatinib inhibitor.

Results: In our dataset of cervical cancer samples comprising of adenocarcinoma and squamous carcinoma subtypes, recurrent mutations are observed in cervical cancer hallmark genes *PIK3CA, ERBB2, ARID1A, EP300, CREBBP* and *PTEN*. Novel cancer-associated mutations of *FGFR2* and *AKT1* gene were specific to adenocarcinoma, whereas *KMT2C, LRP1B* and *FAT4* mutations were common to squamous carcinoma. Several of the epigenetic genes like *BRD4, KMT2D* and *ATRX* were also mutated. Clinical correlation analysis in 84 adenocarcinoma samples revealed that patients harbouring mutations in *PIK3CA, ERBB2* and *FGFR2* have better relapse-free survival compared to patients lacking these mutations.

From the genomic analysis, members of the ERBB family were found to be frequently altered. Therefore, the role of ERBB receptors in cervical carcinogenesis was investigated using inhibitor and knockdown approaches. C33A cells were sensitive to treatment with pan –ERBB inhibitor Afatinib whereas all other cervical cells were resistant. The findings were further validated in *in-vivo* using Afatinib inhibitor wherein C33A tumors showed a delay in tumor growth upon treatment whereas SiHa tumors were unresponsive to treatment. To identify ERBB family gene conferring survival advantage, individual knockdown of ERBB members was performed in SiHa and C33A cells. Although knockdown of *EGFR* and *ERBB2* was efficient in cells and there was a slight decrease in phosphorylation of MAPK, none of the cells of SiHa and C33A showed a decrease in cellular proliferation, migration or anchorage-independent growth upon depletion of ERBB receptors. *ERBB4* knockdown was not potent enough to cause depletion at the protein level.

Conclusion: We present somatic landscape of mutations in cervical cancer for the first time from the Indian population. The study was able to capture known cervical cancer hallmark mutations of *PIK3CA, ERBB2, ARID1A* and *EP300* and novel mutations in *FGFR2* and *AKT1* which could serve as therapeutically relevant gene targets for targeted therapy. Our functional validation results suggest that proliferation and tumor growth can be impeded in C33A cells with pan-ERBB inhibitor Afatinib but not by depletion of individual ERBB receptor. This point towards a possible complex role of re-arrangement of receptors heterodimerization partners among other ERBB receptors upon depletion of one ERBB member, which further needs to be investigated.

2.2 Introduction

Mutations in genes conferring selective growth advantage to cells and contributing to cancer growth are termed as driver mutations. The driver genes comprise of both oncogenes and tumor suppressor genes [6]. A gain of function mutations in proto-oncogenes and inactivating or loss of function mutations in tumor suppressor genes form the basis of cancer initiation and progression [3]. Advances in NGS technology has enabled identification of mutations and other somatic alterations in genes, which act as drivers. Targeting of specific driver genes has resulted in impeding cellular growth and proliferation of malignant cells. This is the underlying concept of targeted therapy. Targeted therapy has proven to be effective as compared to conventional therapy in several cancer types [11]. Few routine examples of the application of targeted therapy in clinics include use of Imatinib to inhibit *BCR-ABL* gene translocation product in Leukaemia, Trastuzumab for *ERBB2* over-expressing breast cancer cells and Lapatinib for *EGFR* and *ERBB2* mutated cancer cells [19]. Identification of driver genes for targeted therapy requires in-depth genomic characterization of a large number of samples for several cancer types. The genomic characterization or identification of driver mutations can be achieved through several NGS approaches [37]. Whole exome sequencing (WES), capturing exonic regions of the genome is the most common method for variant calling and finding mutations in driver genes [118]. Whole-genome sequencing (WGS) can also be used for genome-wide somatic variant identification. Although RNA-sequencing is specifically designed to identify expressed gene transcripts; advances in bioinformatics analysis has also allowed performing variant calling on transcriptome data to identify expressed gene mutations [34].

Cervical cancer is one of the common gynaecological cancers among women worldwide [119]. Of the two histological subtypes, squamous carcinoma accounts for 85-90% of the incidence rate whereas adenocarcinoma represents 10-15% [59]. Extensive genomic

characterization of cervical squamous has been done by the Caucasian population, with sparse representation of adenocarcinoma samples. No major genome-wide studies are reported for cervical cancer from India. Hence, there is an unmet need for systematic characterization of genomic alterations in cervical cancer from the Indian population to identify suitable gene targets for bringing targeted therapy into clinical application. Till date, only bevacizumab, an anti-angiogenesis agent has been approved by the FDA for cervical cancer treatment [23]. There is a necessity to identify driver oncogenic alterations in cervical cancer and more specifically in cervical adenocarcinoma subtype, which remains largely unexplored in Indian as well as Caucasian population.

Extensive genomic profiling has been done by TCGA studies (n=228) and Ojesina *et.al*., group (n=115) by whole-exome, whole genome and whole transcriptome sequencing [95, 120]. These studies identified somatic mutations, copy number alterations, structural aberrations and gene expression changes to report known and novel alterations which can serve as therapeutic or prognostic markers. Here, in this study, we are following similar genomic approaches to profile cervical cancer alterations in the Indian population to discover the ethnic-specific differences from the Caucasian population. Although exome data of cervical squamous carcinoma samples (n=15) is re-analyzed from previously published data [100], our study is dominated by a large number of adenocarcinoma samples (n=84) which provides a unique dataset to characterize the rarer subtype. To achieve the same, we performed somatic variant calling from WES, WGS, WTS and other methods such as MassArray genotyping and Sanger sequencing in cervical adenocarcinoma and provide an integrated genomic mutation analysis for both cervical subtypes.

2.3 Material and Methods

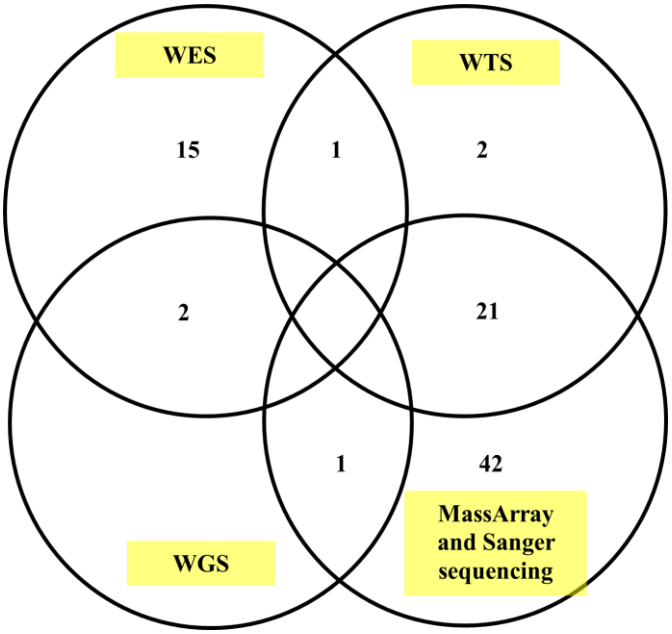
2.3.1 Sample collection and Patient information

Sixty-eight tumor samples and the matched blood samples were collected from cervical adenocarcinoma patients at Advanced Centre for Treatment, Research and Education in Cancer (ACTREC), Mumbai after obtaining informed consent from the patients. Tumor tissues were stored in RNALater and frozen at -80^{0}C until further processing. In addition, 16 tumor samples along with matched tissue normal of the patients were obtained from Tumor Tissue Repository (TTR) at Tata Memorial Hospital (TMH), Mumbai. The tissue samples were snap-frozen and stored at -80^{0}C until further processing.

The study is approved by ACTREC-TMC institutional review board (IRB- Project 116) and patients were recruited for the study from the year 2014-2019. Normal tissue samples were verified by a pathologist for the absence of tumor content.

2.3.2 Sample information used for the NGS and mutation genotyping study

In total, eighty four cervical adenocarcinoma samples were used for mutation profiling. Exome sequencing was performed on 17 paired samples and 1 orphan tumor sample. Whole-genome sequencing was performed on 3 paired samples and whole transcriptome sequencing on 24 tumors and 5 normal samples. 64 samples were used for massArray genotyping and Sanger sequencing. An overlap of tumor samples used for different studies is shown in Venn diagram below.



Information of WES and WGS patient samples is provided in II-Table1 and II-Table 2 respectively. II-Table 3 indicates sample used for variant calling from transcriptome-sequencing.

Sample ID	Adeno carcinoma Sample	Coverage (X)	Tissue type
AD0697	1N	93	Tissue normal
AD0728	1T	75	Tumor tissue
AD0722	2N	84	Tissue normal
AD0727	2T	73	Tumor tissue
AD0703	3N	77	Tissue normal
AD0723	3T	76	Tumor tissue
AD0704	4N	75	Tissue normal
AD0705	4T	94	Tumor tissue
AD0711	5N	73	Blood normal
AD0710	5T	80	Tumor tissue
AD0719	6N	140	Blood normal
AD0726	6T	138	Tumor tissue
AD0700	7N	181	Tissue normal
AD0701	7T	103	Tumor tissue
AD0689	8N	180	Tissue normal
AD0690	8T	181	Tumor tissue
AD0695	10N	107	Tissue normal
AD0696	10T	73	Tumor tissue

Sample ID	Adeno carcinoma Sample	Coverage (X)	Tissue type
AD0707	11N	165	Blood normal
AD0706	11T	183	Tumor tissue
AD0709	12N	66	Blood normal
AD0708	12T	182	Tumor tissue
AD0714	13N	133	Blood normal
AD0713	13T	142	Tumor tissue
AD0716	14N	169	Blood normal
AD0715	14T	146	Tumor tissue
AD0717	15N	179	Blood normal
AD0718	15T	155	Tumor tissue
AD0691	16N	175	Tissue normal
AD0732	16T	161	Tumor tissue
AD0698	17N	155	Tissue normal
AD0730	17T	159	Tumor tissue
AD0699	18N	115	Tissue normal
AD0729	18T	137	Tumor tissue
AD0735	19T	136	Tumor tissue

II-Table 1: Sample information along with sequencing coverage for cervical adenocarcinoma samples used for exome sequencing

Sample ID	Adenocarcinoma sample	Coverage (X)	Tissue type
AD0708	1T	41.83	Tumor tissue
AD0709	1N	48.42	Normal adjacent tissue
AD0718	2T	44.8	Tumor tissue
AD0717	2N	45.28	Normal adjacent tissue
AD1105	3T	43.36	Tumor tissue
AD1120	3N	47.95	Normal adjacent tissue

II-Table 2: Sample and coverage information of whole genome sequenced samples of cervical adenocarcinoma.

No.	Adeno carcinoma Sample	Tissue type
1	AD0722	Normal adjacent tissue
2	AD0724	Normal adjacent tissue
3	AD0685	Normal adjacent tissue
4	AD0702	Normal adjacent tissue
5	AD0703	Normal adjacent tissue
6	AD0800	Tumor tissue
7	AD0801	Tumor tissue
8	AD1097	Tumor tissue
9	AD0727	Tumor tissue
10	AD1811	Tumor tissue
11	AD1088	Tumor tissue
12	AD1095	Tumor tissue
13	AD1093	Tumor tissue
14	AD1104	Tumor tissue
15	AD1099	Tumor tissue

No.	Adeno carcinoma Sample	Tissue type
16	AD1808	Tumor tissue
17	AD1110	Tumor tissue
18	AD1092	Tumor tissue
19	AD1109	Tumor tissue
20	AD1098	Tumor tissue
21	AD1112	Tumor tissue
22	AD1810	Tumor tissue
23	AD1100	Tumor tissue
24	AD1107	Tumor tissue
25	AD1809	Tumor tissue
26	AD1960	Tumor tissue
27	AD1961	Tumor tissue
28	AD1962	Tumor tissue
29	AD1963	Tumor tissue

II-Table 3: Cervical adenocarcinoma samples used for RNA sequencing.

2.3.3 Extraction of DNA and sample QC

DNA was extracted from tumor tissue and the matched normal tissue or blood using DNeasy tissue extraction kit (Qiagen) and QIAamp DNA blood mini kit (Qiagen) following the manufacturer's instruction. Briefly, tissue samples were minced into smaller pieces using a

sterile surgical blade and collected in lysing matrix D tube (MP Biomedicals) containing lysis buffer and homogenized using MP Fast-prep 24 instrument and processed further following kit's protocol. DNA quantification was done using Nanodrop 2000c Spectrophotometer (Thermo Fischer Scientific) and the intactness of DNA was checked by separating the genomic DNA on 0.8% agarose gel. Samples with sufficient DNA amount of 2ug were quantified using ds DNA BR assay kit (Life Technologies, USA) and were used for library preparation and sequencing.

2.3.4 Exome capture, library preparation and sequencing

For sequencing of 17 paired and 1 tumor sample (n=35), two different capture kits were employed and the capture and sequencing were done as described previously [121].

In brief, SureselectXT Target enrichment Kit (Agilent Technologies, Santa Clara, CA, USA) capturing 50 Mb of the genome was used for 13 samples. Briefly, 200 ng of genomic DNA was sheared using covaris to generate 150-500 bp fragment size. The fragment ends were repaired followed by adenylation at 3'end and sample was purified using AMPure XP beads. The fragments were ligated to the adaptor and amplified by PCR. The generated library is then hybridized with SureselectTarget Enrichment system kit and hybrids are separated using streptavidin-coated magnetic beads. Then the sample was PCR amplified using indexing primers and purified. The quality of prepared libraries was assessed on a BioAnalyzer, quantified by qPCR and then loaded on Illumina flowcell to generate clusters. Libraries were sequenced for 301 cycles on the NextSeq 500 Illumina platform to generate 150 bp paired-end reads to obtain 100X coverage.

For remaining 22 samples, Sureselect Human All Exon Kit, v5 (Agilent Technologies, Santa Clara, CA, USA) comprising of probes that capture 3,57,999 exons of 21,522 genes was used to capture 50 Mb of the human genome. Library preparation was done using 1ug of

genomic DNA and the similar protocol was followed as mentioned above. The prepared libraries were loaded on Illumina flow cell and sequenced for 201 cycles on the HiSeq2500 platform to generate 100 bp paired-end reads to obtain 100X sequencing depth.

The sequencing coverage data is shown in II-Table 1.

2.3.5 Library preparation for Whole Genome Sequencing

Whole-genome sequencing was performed on 3 paired samples of cervical adenocarcinoma at a coverage of 40X and above.

In brief, the genomic DNA was fragmented by Covaris and fragments of 350bp were selected and A tailing was performed. Adapters were then added to the ends of DNA and amplified using ligation-mediated PCR approach which was further subjected to single end separation and cyclization. DNA Nanoballs were produced by rolling circle amplification and DNA nanoballs were loaded on patterned nanoarrays on the BGISEQ-500 sequencing platform and paired-end sequencing was done. The raw data was obtained in form of FASTQ files.

2.3.6 Variant analysis from NGS data

2.3.6.1 Variant analysis from Exome sequencing data

Raw data was available as Fastq files. Initial data QC was done using FastQC to assess data quality. The data was analyzed using the in-house optimized pipeline as described previously [121, 122]. In brief, paired-end reads were aligned to human genome hg19 using BWA v.0.6.2. Post alignment, PCR duplicates were removed using Picard tools v.1.74 and InDel realignment and Base quality score recalibration (BQSR) was done using GATK v.2.5-2. Variants were called from GATK Unified Genotyper and Mutect. Variants present in the tumor samples were filtered against the matched normal sample to obtain tumor-specific variants. These variants were further filtered against normal SNP databases like

TMC-SNPdb [121] and dbSNP [123]. Variants identified from both GATK and Mutect and also supported by >=5 reads were retained. Variants were then annotated using Oncotator v1.1.6.0 and deleterious nature of non-synonymous mutations was predicted by using 9 different prediction tools- Mutation Taster, Mutation Accessor, SIFT, Polyphen2_HDIV, Polyphen2_HVAR, LRT, CanDRA, FATHMM, and Provean. Variants which are called deleterious by at least 4 softwares were considered for further analysis. Few deleterious variants were visualized using the Integrated Genomics Viewer (IGV).

2.3.6.2 Variant analysis from Whole Genome sequencing data

The sequencing data of low-quality reads were eliminated and reads were mapped to the human reference genome (GRCh37/hg19) using BWA software. PCR duplicate reads were removed using Picard tools. Variant calling was performed using the HaplotypeCaller of GATK. Further, local realignment around InDels and BQSR were performed using GATK. List of variants in both tumor and normal samples were obtained. A depth filter of 5X was applied and then filtering out of variants common to both tumor and normal of the same sample was done. A unique variant list obtained was further subjected to depletion of SNP reported in dbSNP [123] and TMC-SNPdb [121]. Functional prediction to determine the deleterious nature of the variant was performed using 7 prediction softwares as described earlier. Variants predicted to be deleterious by at least 4 different softwares was considered for further analysis. Few deleterious variants were visualized using the Integrated Genomics Viewer (IGV).

2.3.6.3 Variant analysis from whole transcriptome sequencing data

Somatic variant calling from RNA-sequencing data was done using SNPiR [34]. In brief, reads were mapped to hg19 reference genome and across the splice junctions. To remove duplicate reads, Picard MarkDuplicates was used. Unmapped reads and reads with mapping

quality < 20 were filtered out. Then, indel re-arrangement and BQSR was done using IndelRealigner, CountCovariates and TableRecalibration of GATK. Variant calling was done using GATK UnifiedGenotyper. Next, reads with mismatches at 5' end and those in sites with repetitive regions were excluded. In the following steps, reads with intronic sites with 4 bp of splice junction, having homopolymer runs of >=5 bp were filtered out. In addition, BLAT was used for mismatch reads for aligning against a reference genome to ensure unique mapping. Reads with RNA-editing sites were filtered out. Variants obtained in tumor sample were filtered against normal samples (exome sequenced) to obtain tumor-specific variants. Variants were annotated using Oncotater. Downstream processing was done as described earlier for variant calling from exome sequencing.

2.3.7 Validation of somatic variants obtained in exome sequenced samples

Few candidate genes identified from exome sequencing were selected for validation by Sanger sequencing. PCR was done on genomic DNA to amplify the desired region harbouring mutation. In brief, a 20ul PCR reaction was set up using 2X KAPA Taq Readymix PCR kit (Kapa Biosystems), 0.5 ul of 10 uM Forward primer and reverse primer each and 20-40 ng of genomic DNA of tumor or matched normal sample. PCR was performed at following conditions in a thermocycler- initial denaturation at 98^0C for 5 min, 30 cycles of 95^0C for 30 sec, 50^0C for 30 sec and 72^0C for 30 sec and a final extension at 72^0C for 5 min. 5 ul of PCR product was run on a 1.5% agarose gel for visualization of a single band. Remaining 15 ul PCR product was purified using MN gel and PCR purification kit (Macherey Nagel) and quantitated using Nanodrop 2000c spectrophotometer (Thermo Fischer Scientific). 4-5 ng of PCR purified product was used for Sanger sequencing. Sanger traces were analysed for mutations using Mutation Surveyor DNA variant analysis software (Softgenetics LLC). The primer information is shown in II-Table 4.

Primer	Sequence	Amplicon size
OAD1519_*ERBB2*_S310F/Y_F	CACGAAGGGCCAGGGTATG	260 bp
OAD1520_*ERBB2*_S310F/Y_R	GGGTCTGAGGAAGGATAGGAC	
OAD1097_*ERBB2*_D769Y_F	ATCCCTGATGGGGAGAATGT	134 bp
OAD1098_*ERBB2*_D769Y_R	GGGTCCTTCCTGTCCTCCTA	
OAD1242_*ARID1A*_p.Q538_F	AGTCTCAACCACCACAGCTC	171 bp
OAD1243_*ARID1A*_p.Q538_R	GCTGGTAAGGAGACTGAGCC	
OAD1244_*ARID1A*_p.Q555_F	CAGCCTCCACATCAGCAGTC	180 bp
OAD1245_*ARID1A*_p.Q555_R	CTGGGGCTGAGGATACGC	
OAD1246_*ARID1A*_p.Q780H_F	TTATATGCAGAGGAACCCCCA	179 bp
OAD1247_*ARID1A*_p.Q780H_R	TAGTATACTGACCTTGTGGGCCAT	
OAD1248_*FGFR2*_p.K659E_F	TTGACGGCCTTTCTTCCTGG	180 bp
OAD1249_*FGFR2*_p.K659E_R	GCAGCCAGAAATGTTTTGGTA	
OAD1250_*FGFR2*_p.S320F_F	TGGCCTGCCCTATATAATTGGA	179 bp
OAD1251_*FGFR2*_p.S320F_R	TGGTGGGACCATAGACAATGC	
OAD1617_*FGFR2*_C382R_F	AGTCTGGCTTCTTGGTCGTG	254 bp
OAD1618_*FGFR2*_C382R_R	CTTGAGAATGGTCGTCGCCT	
OAD1252_*ATM*_p.W579_F	TCCAGGAACGGTAAAAATGGGA	177 bp
OAD1253_*ATM*_p.W579_R	AGCAGCATGCTAATGAACTTAAA	
OAD1254_*TSC2*_p.L1216I_F	CCAGAGATGGGTAAGGGGAGGT	157 bp
OAD1255_*TSC2*_p.L1216I_R	CCATGAGGGCGTTAGACAGCTC	
OAD1256_*EP300*_p.E1365K_F	GGTTCCCCCACCATCTCAAT	180 bp
OAD1257_*EP300*_p.E1365K_R	TCATACCTCTGGTTGGGTGGA	
OAD1258_*PIK3CA*_p.IL112fs_F	GACGACTTTGTGACCTTCGG	159 bp
OAD1259_*PIK3CA*_p.IL112fs_R	TTTAGAAAGGGACAACAGTTAAGC	
OAD963_*PIK3CA* _E542/E545_ F	CAATGAATTAAGGGAAAATGA	177 bp
OAD1227_*PIK3CA*_E542/E545_ R	AGATCAGCCAAATTCAGTTA	
OAD2112_*PIK3CA*_H1047R_F	ATTTGAGCAAAGACCTGAAGG	521 bp
OAD2113_*PIK3CA*_H1047R_R	CAAACCCTGTTTGCGTTTACA	
OAD1260_*FGFR3*_p.H350Y_F	TCTCTCCTTGCACAACGTCAC	176 bp
OAD1261_*FGFR3*_p.H350Y_R	AGCTTTGGCGTGTCCCGAG	
OAD1262_*TSC1*_p.L203fs_F	GTAGCAAACAAACAAGCAGTTTCA	178 bp
OAD1263_*TSC1*_p.L203fs_R	GGCGGAAGTCTATCTCGTCC	
OAD2110_*AKT1*_E17K_F	GAATCCCGAGAGGCCAAGG	401 bp
OAD2111_*AKT1*_E17K_R	TTTCAGACACAGCTCGGGGT	

II-Table 4: Primers used for validation of mutations

2.3.8 Validation of somatic variants in validation cohort samples by orthologous methods

2.3.8.1 Sanger sequencing for validation of mutations in additional samples

Mutations in *ERBB2* S310F/Y, *AKT* E17K, *PIK3CA* E545K/E542K and H1047R, *FGFR2* S320F, K659E, C382R were assessed in the validation cohort samples.

PCR was performed for amplification of genomic regions harbouring above mentioned mutations. Briefly, 20 ul PCR reaction was set up using 2X KAPA Taq Readymix PCR kit (Kapa Biosystems), 0.5 ul of 10 uM Forward primer and reverse primer each and 20-40 ng tumor genomic DNA or matched blood normal genomic DNA as a template. PCR was performed at following conditions in a thermocycler- initial denaturation at 98^0C for 5 min, 30 cycles of 95^0C for 30 sec, 57^0C/55^0C for 30 sec and 72^0C for 30 sec and a final extension at 72^0C for 5 min. 5 ul of PCR product was separated on a 1.5% agarose gel for visualization of the single amplicon. The PCR product was diluted and further cleaned using Exosap IT PCR product Cleanup reagent (Thermo Fischer scientific) and submitted for Sanger sequencing. Sanger traces were analyzed for mutations in the above-mentioned genes using Mutation Surveyor DNA variant analysis software (Softgenetics LLC). Primer information is provided in II-Table 4.

2.3.8.2 Mutation profiling using MassARRAY based genotyping

Mutations were screened in 43 tumor samples of cervical adenocarcinoma using a panel of 53 mutations across 17 oncogenic genes by MassARRAY® System (Agena Bioscience) using iPLEX Pro chemistry. The mutation panel of 53 oncogenic mutations was custom designed for this study using Assay Design 3.0.0 software (II-Table 5). The assay was performed following standard optimized protocol at Imperial Life Sciences (ILS, India). Briefly, PCR was performed to amplify the mutation region, followed by SAP treatment to

dephosphorylate incorporated nucleotides. Then, iPLEX reaction using single base extension PCR was performed using mass modified nucleotide terminators resulting in extension at the target site of mutation. The iPLEX reaction product was desalted using clean resin and loaded on SpectroCHIP bioarray and then products were analyzed by MALDI-TOF technology using the MassARRAY platform. Data analysis for mutation calling was performed using MassARRAY Typer Analyzer 3.3.0. All the mutations called by the software were manually reviewed to confirm the presence of a mutation in samples. Mutations observed in tumor samples were screened in the matched normal samples by Sanger sequencing. Mutations that were observed only in tumor samples are reported.

Gene	Mutation	Gene	Mutation	Gene	Mutation
AKT1	E17K	*ERBB2*	I767M	*FGFR3*	Y373C
ERBB2	S310F	*ERBB2*	S653C	*KRAS*	G12C
PIK3CA	E542K	*ERBB2*	V777L	*KRAS*	G12V
PIK3CA	E545K	*ERBB3*	A130T	*KRAS*	G13D
PIK3CA	H1047R	*ERBB3*	V104M	*NRAS*	Q61K
CDKN2A	D108G	*FGFR2*	S252W	*PTEN*	G165E
CDKN2A	D108Y	*FGFR2*	C382R	*PTEN*	R130G
CDKN2A	D84N	*FGFR2*	K659E	*PTEN*	R130Q
CTNNB1	S33C	*FGFR2*	K659N	*RB1*	C706F
CTNNB1	S37C	*FGFR2*	N549K	*RB1*	E748
DDR2	S768R	*FGFR2*	S320C	*RET*	A664D
EGFR	746-750del	*FGFR2*	S320F	*RET*	M918T
EGFR	G719A	*FGFR2*	W290C	*TP53*	R158L
EGFR	G719S	*FGFR2*	Y375C	*TP53*	R175H
EGFR	L858R	*FGFR3*	G691R	*TP53*	R248L
EP300	D1399N	*FGFR3*	K650Q	*TP53*	R273H
EP300	E1365K	*FGFR3*	R248C	*TP53*	R273L
ERBB2	G660R	*FGFR3*	S249C		

II-Table 5: List of mutations genotyped by MassARRAY in validation cohort samples

2.3.9 MTT Assay

MTT Assay was performed in cervical cells using inhibitor Afatinib. In brief, 3000 cells per well of the 96 well plate were seeded for SiHa and HeLa, 2000 cells for C33A and ME180 cells, 3000 cells of BT474 and 1000 cells of A549 cells. Cells were treated with Afatinib at different concentrations in 6 replicates. After 72 hours, MTT reagent (0.5 mg/ml) was added and cells were incubated for 4 hours or until formazan crystals appeared. Formazan crystals were dissolved in DMSO and absorbance was taken at 570nm using a microplate reader (Biorad). Cell viability was assessed by comparing with non-treated cells and IC50 value was inferred for each cell line. BT474 and A549 were used as sensitive and resistant controls cell lines for Afatinib treatment.

2.3.10 Virus production and transduction to generate knockdown clones

pZIP-hCMV shRNA construct targeting *ERBB2*, *EGFR* and *ERBB4* and scramble vector (TransOmics, Technologies, USA) were used for lentivirus production in 293FT cells using Lipofectamine 3000 reagent (Invitrogen). The virus was collected at 48 and 72 hours and filtered through 0.4uM filter and SiHa and C33A cells were transduced with a virus in presence of 8ug/ml of polybrene. Transduced cells were observed for GFP expression and selected using puromycin (1ug/ml) for 3-4 days.

2.3.11 Western blotting

Cells were lysed in RIPA or NP40 lysis buffer containing 1mM DTT and protease inhibitor cocktail (Calbiochem, Merck) and sonicated for 5 cycles with conditions of 30 sec on and 30 sec off. Protein was estimated using BCA reagent at 562 nm. 40 ug of protein was loaded on 8% SDS-PAGE gel, transferred on a nitrocellulose membrane (Amersham Hybond, GE healthcare) by electroblotting at 50V overnight. After verifying the transfer using Ponceau stain, the membrane was blocked in 5% BSA for 1 hour at room temperature and then, the

blots were incubated with primary antibody overnight at 4^0C and appropriate secondary HRP conjugated antibody for 1 hour at room temperature. Blots were washed with 1X TBST and developed using Pierce ECL Western blotting substrate (Thermo Fischer Scientific) or Western blot chemiluminescence HRP substrate (Takara) and luminescence was capture on Chemidoc System (Biorad). Primary antibody for Phospho-HER2 (Tyr1248)(1:500 dilution, #AP0152) from Abclonal, T-HER2 (1:500, #2168S) from cell signaling (CST), pEGFR (Y1068)(1:500, #2234S) from CST, T-EGFR (1:500, #sc03) from Santacruz, T-ERBB4 (1:500, #sc8050) from Santacruz, pMAPK p42/44 (Thr202/Tyr204) (1:1000, #9101S) from CST, T-MAPK (1:500, sc-154) from Santacruz, pAKT(Ser473)(1:500, #4060S) from CST, T-AKT (1:500, #4685S) from CST, ARID1A (1:500, #301-041A) from Bethyl labs and GAPDH (1:2000, #sc32233) from Santacruz were used. Secondary antibodies are HRP linked goat Anti-rabbit IgG (1:2000, sc2004) and goat anti-mouse IgG (1:2000, sc2005) from Santacruz.

2.3.12 Cell proliferation assay

20,000 cells were seeded per well of 24 well plates. Cell number was counted after 24 hours and 96 hours respectively in both scramble/parent and knockdown clones. Percent cell proliferation was calculated in knockdown clones with respect to scramble/parent. The experiment was performed thrice.

2.3.13 Migration assay

For assessing migration, scratch wound assay was performed. Cells were seeded in a 6 well plate at 95-99% confluency and treated with Mitomycin C (10 ng/ml) for two hours. Scratch was made with a sterile 10ul tip and detached cells were washed with PBS and fresh media was added to each well. Cell migration was monitored using time-lapse microscope and

images were obtained every 30 min for duration of 22-48 hours. Percent wound migration was estimated using ImageJ software.

Migration assay was also performed using 8 uM pore size insert. Briefly, $5X10^4$ cells of SiHa and 2 $X10^5$ cells of C33A were suspended in 200ul of serum-free media and added to the upper side on insert whereas media containing 10% FBS was below the insert in a companion plate. The plate was incubated for 12-16 hours at 37^0C in the incubator. Then the migrated cells were fixed with 3.7% formaldehyde and permeabilized with 100% ethanol, followed by staining with 0.4% crystal violet. The non-migrated cells were removed from the inner side of the insert with a cotton swab. Images of migrated cells were captured at 10X magnification and counted using ImageJ software. Each experiment was performed in triplicates.

2.3.14 Soft agar assay

SiHa and C33A cells were seeded at a density of 5000 cells per well in a 6 well plate along with DMEM containing 0.4% agar onto 0.8% bottom agar with DMEM. Cells were incubated at 37^0C in a CO_2 incubator for 10-14 days till the appearance of colonies. Ten images per well were taken at 10X magnification using Phase contrast inverted microscope (Zeiss axiovert 200m) and colonies were counted manually with ImageJ software. The average number of colonies in knockdown clones and control were plotted in GraphPad Prism and unpaired T-test was performed on individual knockdown clones as compared to control to calculate statistical significance.

2.3.15 In-vivo inhibitor studies

Female NOD-SCID mice of 6-8 weeks were subcutaneously injected with 7.5 X 10^6 cells of C33A and SiHa each. C33A and SiHa cells formed tumors within 1.5- 2 months. Then, each of C33A and SiHa tumors was further grafted in 12 mice each. The graft was allowed to establish for 7-10 days till tumor volume reached 100-150 mm^3. Of the 12, 6 mice each were

randomized and placed in the control group and treatment group. Afatinib/ BIBW-2992 was administered at 20mg/kg of body weight along with vehicle control using oral gavage following methodology as previously described [124]. Treatment was continued daily for 24 days in case of C33A and 5 days for SiHa and tumor volume was measured every 3 days. Micro-PET using 18F-FDG and CT scan was done before the treatment and at the end of the treatment. Mice were sacrificed after 25 days or whenever tumor volume reached 2000 mm^3. Tumors were excised, part of tumor was fixed in 10% formaldehyde for histological analysis and part of it was stored in RNALater for molecular analysis.

2.3.16 Detection of HPV virus and integration in the human genome

HPV infection was detected in cervical adenocarcinoma samples by two methods. In the case of patient samples which were subjected to NGS, the detection was done using in-house pathogen detection tool- Cancer Pathogen Detector (CPD). The raw reads are mapped to the pathogen genome and different HPV strains are identified.

For the rest of the samples, PCR using My09/11 primers was performed. In brief, PCR was done at thermocycler conditions- 95^0C for 5 min, 35 cycles of 94^0C for 30 seconds, 55^0C for 30 sec and 72^0C for 45 sec and final extension of 72^0C for 5 minutes. An amplicon size of 450 bp was visualised on 1.5% agarose gel. Primer sequences are provided in II-Table6.

Primer	Sequence	Amplicon Size
OAD450_*MY09*_HPV_F	CGTCCMARRGGAWACTGATC	450 bp
OAD451_*MY11*_HPV_R	GCMCAGGGWCATAAYAATGG	

II-Table 6: Primers used for HPV detection

The HPV integration site was identified using HPVDetector [125], a tool developed in our lab. The integration sites were compared to literature reports with information of known and novel fragile sites.

2.3.17 Knockdown of *ARID1A* by siRNA in cervical cancer cells

ON-Targetplus pooled siRNA against *ARID1A* were ordered from Dharmacon. siGLO was used as transfection control. In brief, cells were seeded at 50-60% confluency in a 6 well plate. siRNA transfection was performed with 25nM concentration using Lipofectamine RNAiMax transfection reagent (Invitrogen). Cells were harvested at 48 hours to check ARID1A depletion on mRNA level and 96 hours on protein level by western blotting.

2.3.18 Generation of *ARID1A* knockout clones using CRISPR-Cas9

sgRNA sequences were designed using the sgRNA design tool by Broad Institute. sgRNA targeting exon 2 of *ARID1A* were cloned in LentiCRISPR V2 plasmid by digesting with restriction enzyme BsmB1. Cloning was confirmed with Sanger sequencing. Lentivirus was produced in 293FT and transduced in cervical cells SiHa and CaSki following protocol as described previously. Post puromycin selection, single-cell dilution was performed in 96 well plate and single clones were expanded further for the screening of knockout clones using T7 endonuclease assay and PCR based approach. Two sets of PCR were performed. First PCR amplifies exon 2 of *ARID1A* and in second PCR, forward primer binds in the intended perturbed region. Presence of band in both PCR suggests that no genomic perturbation has occurred in the expected region. In case of a knockout clone, first PCR gives a band of a smaller size than expected and absence of a band in second PCR. Knockout clones which were positive by PCR were further sequenced to confirm.

2.3.19 Cloning of shRNA sequences in pEGFP vector

pEGFP vector was digested with restriction enzymes AgeI and EcoR1 and shRNA sequences (obtained from TRC library) of *ARID1A* and *ERBB4* were cloned. Positive clones were screened using EcoRV and KpnI digestion. A 1.5 Kb release is seen in empty vector upon digestion whereas cloned vector show absence of 1.5Kb band. Sanger sequencing was

performed for confirming the cloned sequence. *ARID1A* shRNA sequences are as follows: sh1- CCTCTCTTATACACAGCAGAT and sh2-CCGTTGATGAACTCATTGGTT and for *ERBB4,* sh1-CCTGTGGCTATTAAGATTCTT and sh2-GCGCAGGAAACATCTATATTA.

2.3.20 Survival analysis

Survival analysis was performed using statistical package SPSS statistics 21 (IBM). Relapse free survival was assessed using Kaplan-Meier survival analysis for patients harbouring mutations in *PIK3CA, ERBB2* and *FGFR2* and patients lacking these mutations.

2.4 Results:

2.4.1 Patient sample information

Eighty-four paired samples of cervical adenocarcinoma were analysed for somatic mutations. Treatment naive patient samples were collected at ACTREC-TMC and TMH-TMC for the study. The primary line of treatment comprised of radiation and chemotherapy. The adenocarcinoma histology was confirmed by the pathologist.

The clinical features of patient samples are described in II-Table 7. In brief, patients belonged to a median age of 51.5 years (Range 29-72 years) in our cohort. Considering FIGO staging for 84 samples, 28.6% patients belonged to stage I, 47.6% of stage II stage and 14.3 % of stage III. 11.9% of the patients showed relapse, whereas relapse did not occur in 32% of the patients. For 50% of the patients, data is not available. Since the follow-up data was available for 60 months for 84 patients; relapse-free survival analysis was done.

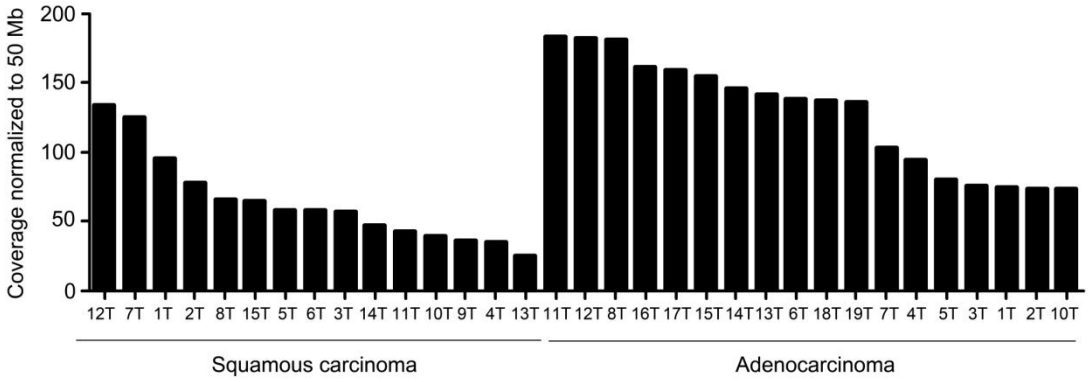
FIGO Stage	No. of patients (n=84)	Percent
Stage I	24	28.6
Stage II	40	47.6
Stage III	12	14.3
Information not available	8	9.5

II Table-7: FIGO stage distribution of cervical adenocarcinoma patient samples

2.4.2 Data QC

2.4.2.1 Data QC analysis of Whole Exome Sequencing

Whole exome sequencing was performed on 17 paired samples and 1 orphan tumor sample of cervical adenocarcinoma subtype and whole-exome sequencing data for squamous carcinoma of 15 samples (10 paired and 5 orphan tumors) was reanalysed from previously published data [100] to perform an integrated analysis to identify mutation profiles between two histological types. About 50 Mb of the genome was captured. Tumor samples for cervical adenocarcinoma were sequenced at an average coverage of 138X (range: 73-183X) and normal samples at 134X (range: 66-181X) whereas tumors for squamous carcinoma were sequenced at 57X (Range: 25-134X) and normal samples at 50X (Range: 30-94X) coverage. Sequencing depth or coverage (X) for tumor samples of squamous and adenocarcinoma subtype is shown in II-Figure 1. Somatic mutations were called using GATK [126] and Mutect [127] and variants present in normal samples were depleted from tumors to obtain tumor-specific variants.



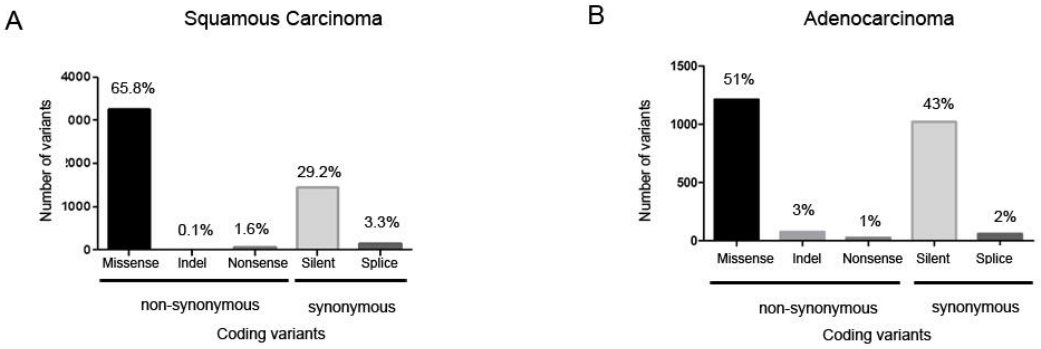
II-Figure 1: Sequencing depth or coverage (X) calculated for both squamous (n=15) and adenocarcinoma (n=18) tumor samples. Squamous carcinoma samples were sequenced on Illumina GAIIx platform whereas adenocarcinoma samples were sequenced on Illumina NextSeq and HiSeq 2500 platform.

2.4.2.2 Data QC analysis of Whole Genome Sequencing

Three paired samples of cervical adenocarcinoma were subjected to whole-genome sequencing. Tumor samples were sequenced at an average coverage of 42X (Range: 41X-44X) and normal samples at 46X coverage (Range: 45X-48X).

2.4.3 Somatic variants analysis in cervical adenocarcinoma and squamous subtype.

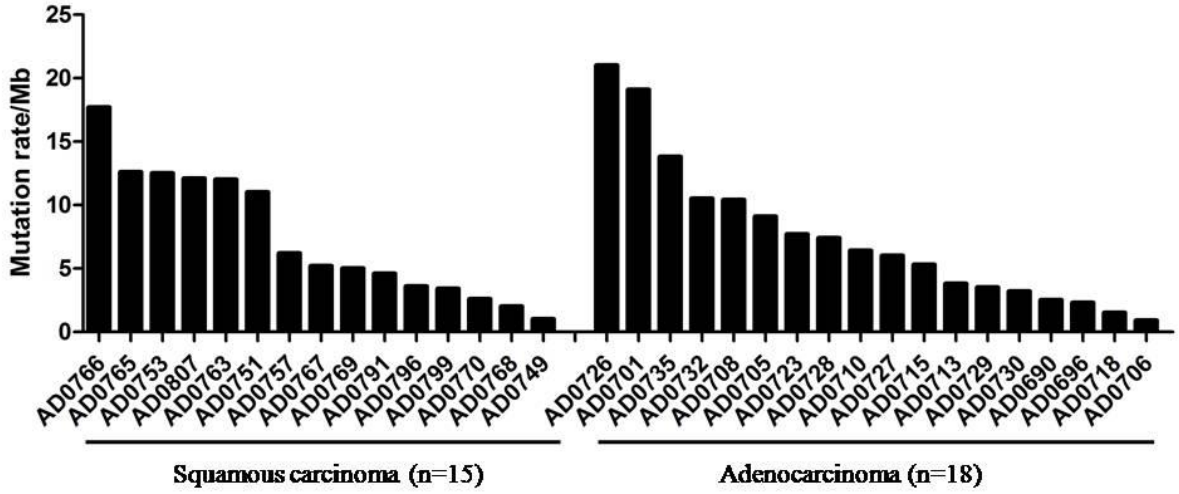
From the variants identified from exome sequencing following filtration criteria, we identified a total of 1178 missense, 234 non-sense, 315 indels, 1691 silent and 21 splice site mutations in cervical adenocarcinoma samples and 3269 missense, 79 non-sense, 4 indels, 1452 silent and 162 splice site mutations in squamous carcinoma samples among the coding variants. II-Figure 2- A and B show distribution of coding variants in squamous and adenocarcinoma samples.



II-Figure 2: The percent variant classification for synonymous and non-synonymous mutations belonging to the coding region is displayed for A) squamous and B) adenocarcinoma samples of cervical cancer.

Excluding hypermutated samples (mutation> 10/Mb), the aggregate non-synonymous mutation rate is 6 mutations per Mb and 4.8 mutations per Mb for adenocarcinoma and squamous carcinoma respectively. The mutation rate observed is consistent with the literature for cervical cancer [128, 129]. Mutation rate per Mb of the genome is shown in II-Figure 3 for tumor samples of cervical squamous and adenocarcinoma subtypes. The non-

silent mutation rate observed in Caucasian population reported for adenocarcinoma and squamous carcinoma is 1.6 and 4.2 mutations per Mb respectively [94, 95].



II-Figure 3: Mutation rate per Mb calculation for each sample of exome sequenced samples-squamous (n=15) and adenocarcinoma (n=18). The exome capture is 37Mb and 50 Mb for squamous and adenocarcinoma respectively.

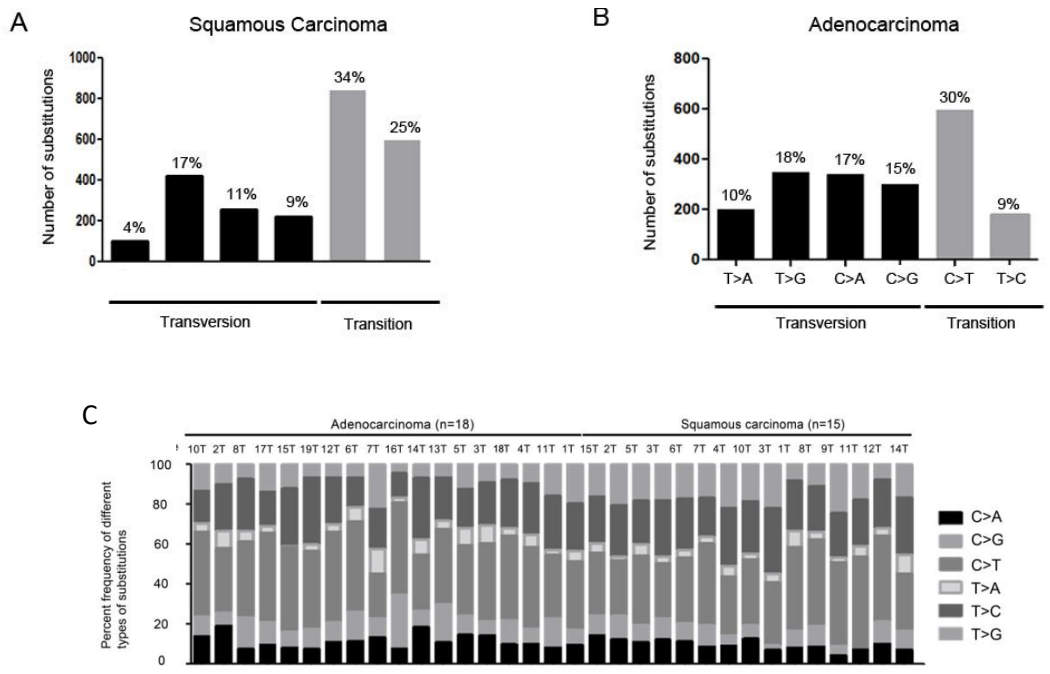
Sample	Total variants	Coding variants						Non-coding variants							Mutation rate
		Nonstop	Splice	Indel	Nonsense	Missense	Silent	3UTR	5Flank	5UTR	IGR	Intron	lincRNA	RNA	
10T	588	0	3	5	2	66	37	145	1	31	94	193	3	8	1.46
11T	1322	0	7	6	1	26	21	160	0	28	464	521	39	49	0.66
12T	8294	0	55	181	11	301	135	2053	0	183	1261	3877	84	153	9.86
13T	1504	1	6	14	5	108	44	181	0	37	408	614	31	56	2.54
14T	1159	0	9	6	2	157	72	192	0	81	164	424	18	34	3.3
15T	847	0	2	4	6	39	27	126	0	25	249	314	25	30	0.98
16T	2352	0	13	10	29	286	118	386	2	97	393	938	25	55	6.5
17T	1171	0	1	13	3	94	54	131	0	32	274	513	21	35	2.2
18T	1362	0	9	15	5	99	56	165	0	35	318	596	21	43	2.38
1T	1162	1	9	6	7	213	78	52	0	19	127	613	10	28	4.52
2T	903	0	14	2	18	163	78	34	0	28	93	446	13	14	3.66
3T	1341	0	11	7	17	213	85	47	0	24	112	763	9	53	4.74
4T	1380	0	12	7	16	256	102	47	0	22	126	740	13	39	5.58
5T	884	1	14	8	14	177	57	27	0	22	74	452	9	30	3.98
8T	1470	0	7	9	5	70	25	197	0	55	386	645	26	45	1.68
6T	3139	2	26	11	52	576	246	123	1	51	259	1675	29	90	12.78
7T	3211	2	26	12	26	545	211	119	1	40	341	1739	42	109	11.66
19T	3366	1	24	21	21	391	209	549	1	131	567	1300	49	103	8.66

II-Table 8: Distribution of coding and non-coding variants obtained from exome sequencing data of cervical adenocarcinoma tumors.

Sample		Coding variants						Non-coding variants							Mutation rate
	Total variants	Nonstop	Splice	Indels	Nonsense	Missense	Silent	3UTR	5Flank	5UTR	IGR	Intron	lincRNA	RNA	
1T	81	0	1	0	0	30	8	0	0	0	0	38	0	4	0.8
2T	507	0	17	0	11	320	116	3	2	2	8	21	0	7	8.9
3T	593	0	22	0	8	368	132	9	2	2	12	23	1	14	10.2
4T	333	0	7	0	7	180	76	2	3	2	5	45	0	5	5.1
5T	575	1	14	0	7	352	138	8	9	6	9	20	1	10	9.7
6T	575	0	22	0	8	369	124	11	3	5	6	19	0	8	10.2
7T	990	0	29	4	14	518	273	12	9	11	24	53	12	29	14.5
8T	324	1	10	0	1	154	110	5	0	0	5	24	2	12	4.2
9T	106	0	2	0	2	59	31	0	1	0	1	10	0	0	1.6
10T	266	0	5	0	4	147	76	2	0	2	2	24	0	3	4.1
11T	138	0	2	0	5	73	42	0	1	1	0	10	0	3	2.1
12T	365	0	3	0	2	137	93	5	0	0	9	99	1	16	3.8
13T	185	0	1	0	0	101	45	0	1	6	2	29	0	0	2.7
14T	227	0	3	0	0	108	60	1	0	2	2	46	0	4	2.9
15T	587	1	24	0	9	353	128	11	5	5	10	25	1	14	9.8

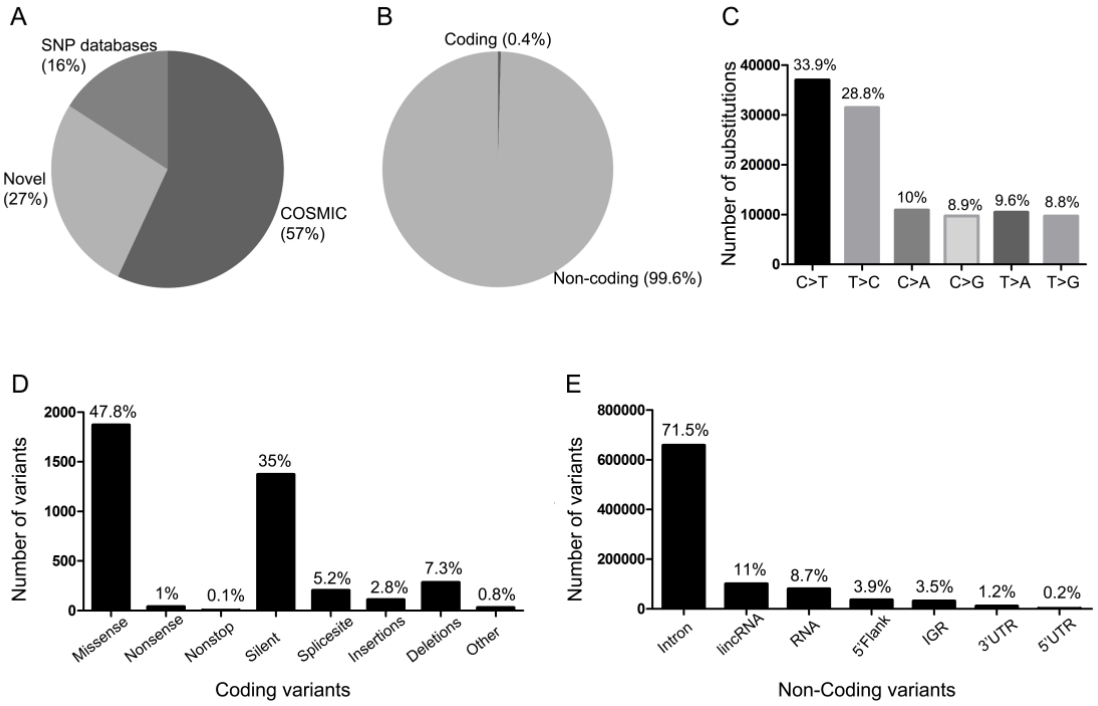
II-Table 9: Distribution of coding and non-coding variants obtained from exome sequencing data of cervical squamous tumors.

Considering both the histological subtypes, a mutational signature dominant in both subtypes was C>T transition, followed by T>C and T>G as shown in II-Figure 4, which is consistent with mutational signature reported for cervical cancer [94, 95, 130]. C>T mutational signature is associated with APOBEC deaminase enzyme activity [131]. However, we have not checked the expression of APOBEC in our dataset.



II-Figure 4: Mutational signatures for A) squamous and B) adenocarcinoma. In both histological subtypes, C to T transition is the dominant signature pattern. C) Transition-transversion distribution in individual patient samples of cervical cancer.

Similarly, variant calling from whole-genome sequencing data performed using standard GATK pipeline in 3 paired adenocarcinoma subtype identified mutations in coding region- 1870 missense, 39 nonsense, 4 non-stop, 1372 silent, 204 splice-site, 111 insertions, 284 deletions and 31 mutations belonging to other categories. Among the non-coding mutations, 11134 3'UTR, 36364 5'Flank, 1944 5'UTR, 31955 IGR, 658183 Intron, 101008 lincRNA and 80132 RNA were obtained. Only 0.4% of the variants belonged to coding region whereas 99.6% variants were in the non-coding region. Among the coding variants, about 57% variants were reported in the COSMIC database, 27% were novel and 16% belonged to the SNP database (II-Figure 5). Applying filtration criteria of missense mutations predicted to be deleterious by 4 or more tools, removing SNPs reported in dbSNP and TMC-SNP database, a list of 767 variants were identified.



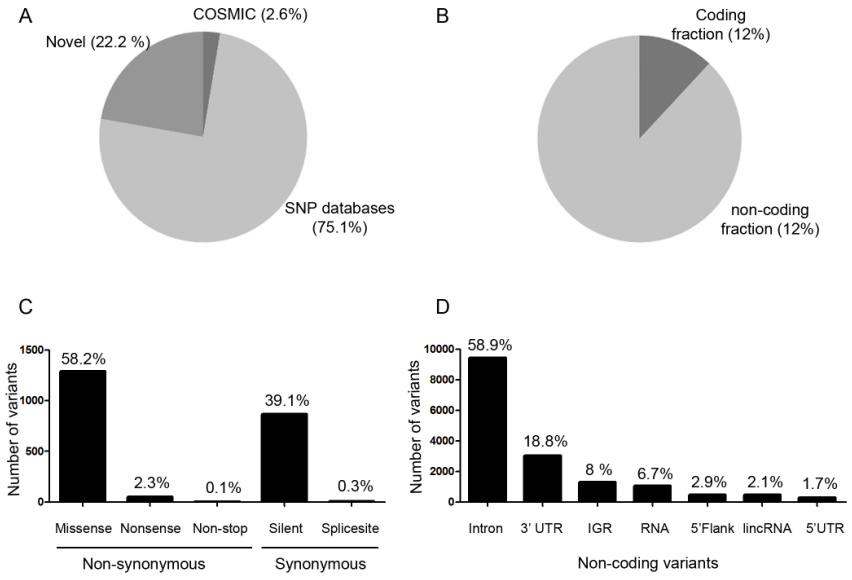
II-Figure 5: Variant features identified from whole-genome sequencing.

Mutational signatures for three samples show C>T and T>C transitions at 34% and 29% frequency respectively. Consistent with the literature reports and our exome sequencing dataset, C>T mutational signature is common in cervical cancer [132] and it corresponds to signature 1B which suggests over-expression of APOBEC enzymes [130]. Literature reports suggest that HPV activates APOBEC3 activity which damages the host genome and therefore enrichment of mutational signature is observed [133].

Next, SNPiR was used for variant calling from RNA-sequencing data of 23 tumors and 4 normal samples. Two samples AD0703 and AD1808 were sequenced at 270 million and 140 million reads respectively, for which the SNPiR run could not be performed. The number of variants identified in each sample is shown in appendix 6. A total number of 362883 variants in tumor sample and 23568 variants in 4 normal samples were obtained. Since all 4 normal samples were of poor RIN, all the expressed transcripts may not have been uniformly captured. Therefore, in addition to the variant list obtained from RNA-sequenced

normal tissue samples (n=4), the variant list from normal cervix tissue samples (n=17) from whole-exome sequencing was also used for depletion from tumor variants to identify tumor-specific variants. Oncotator was used for further annotation.

Depletion of RNA-sequenced normal variants from tumor samples yielded a total of 33654 variants whereas depletion of normal variants from exome and transcriptome sequenced samples (n=21), a total of 18265 tumor-specific variants were obtained. Among the coding variants, in total 1292 missense, 52 nonsense, 3 non-stop, 868 silent and 6 splice-site mutations were found whereas, for non-coding variants, there were 9448 introns, 3018 3'UTR, 1291 IGR, 1073 RNA, 472 5'Flank, 465 lincRNA and 277 5'UTR. The variant classification is shown in II-Figure 6. About 50% of the variants are observed in the intronic region whereas missense mutations comprise 7% of the variant fraction. Further, missense variants predicted to be deleterious by 4 or more prediction tools were considered along with other variants. Next, we depleted germline variants reported in the dbSNP database and TMC-SNP databases to obtain a list of 383 tumor-specific variants.



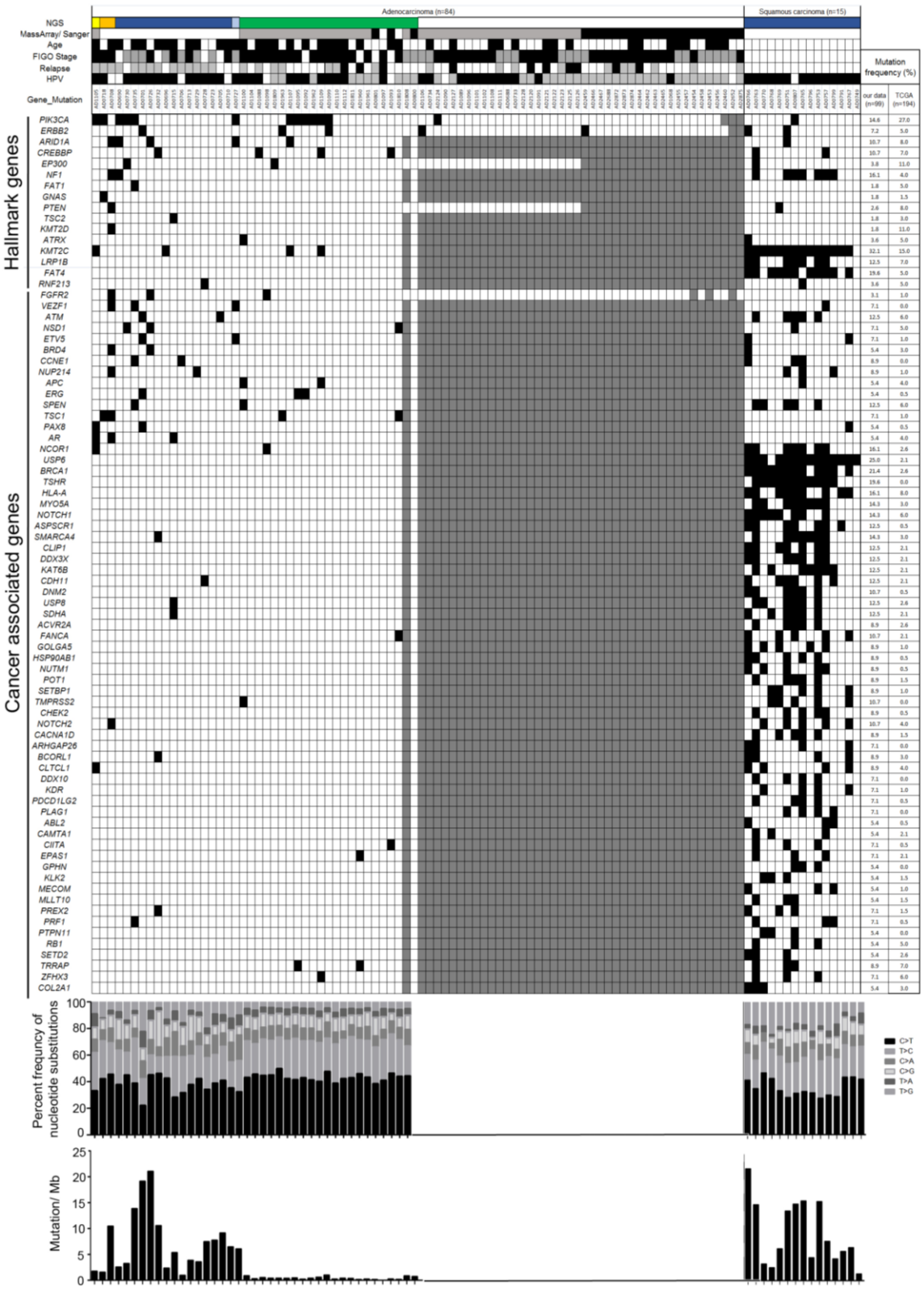
II-Figure 6: Distribution of variants identified from Transcriptome sequencing data using SNPiR variant calling method

The integrated variant analysis was performed for both subtypes of cervical cancer. Mutations in the COSMIC cervical cancer hallmark genes were assessed in 84 cervical adenocarcinoma samples and 15 squamous carcinoma samples. As seen in II-Figure 7 and 8 below, *PIK3CA* is the most frequently mutated gene (14.6%). Majority of the mutations are in helical domain E545K and E542K with one in kinase domain H1047R. Mutation in tumor suppressor gene *ARID1A* (10.7%) was queried in 60 samples of which 6 samples showed loss of function mutations introducing stop codons and frameshift mutations in adenocarcinoma subtype and missense mutation P1860H in squamous carcinoma. Mutations in *ERBB2* (7.2%) belonged to extracellular domain S310F/Y and one belonged to kinase domain D769Y in case of adenocarcinoma and S310Y and R130W mutations in squamous carcinoma. Mutations were observed in other hallmark cancer genes *CREBBP* (10.7%), *GNAS* (1.8 %) and *ATRX* (3.6 %). Among the tumor suppressor genes, mutations in *PTEN* (2.6%), *TSC2* (1.8 %), *FAT1* (1.8%), *FAT4* (19.6%), *LRP1B* (12.5%) and *NF1* (16.1%) were common. In addition, mutations were recurrent in other chromatin remodelling genes *EP300* (3.8%), *KMT2D* (1.8%) and *KMT2C* (32.1%). Considering the mutated hallmark genes, *PIK3CA*, *ARID1A* and *ERBB2* mutations are dominant in adenocarcinoma subtype whereas mutation in *KMT2C*, *LRP1B*, *FAT4* and *NF1* are common in squamous carcinoma (II-Figure 7 and II-Figure 8).

Our data is consistent with literature reports and TCGA studies wherein mutations are recurrently observed in the *PIK3CA* gene. Mutation in *ERBB2* and *ARID1A* are more common in cervical adenocarcinoma as also observed in our dataset. However, we do not observe mutations in genes like *KRAS* which have very high mutation frequency of 12.3% as per COSMICdb and 13.9% in TCGA for adenocarcinoma subtype. Absence of mutations in other tumor suppressors *TP53* and *STK11* was also noted.

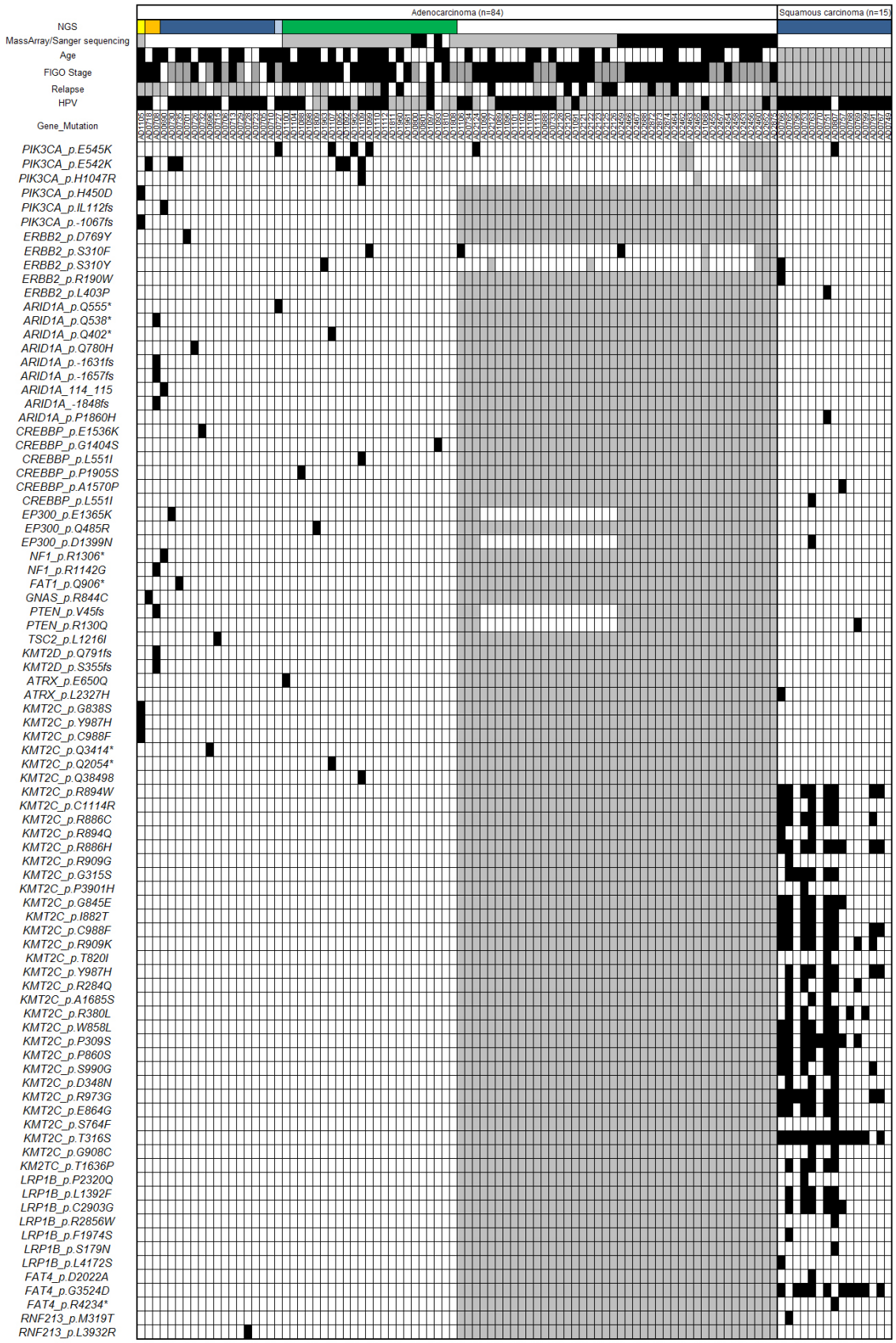
Most of the *PIK3CA* mutations are observed with HPV positive samples and also associated with late FIGO stages (IIB and above) whereas *ERBB2* mutated samples are associated with both early and late FIGO stages as well as with HPV presence and absence. None of the samples with hallmark mutations showed relapse.

Mutations in other cancer-associated genes are also reported here. Genes mutated in at least 3 samples and above were considered for heatmap generation shown in II-Figure 7. Recurrent oncogenic mutations were observed in *FGFR2* at 3.1% frequency. *FGFR2* mutated samples show association with late FIGO stages. In addition, additional mutations in tumor suppressors include genes *APC* (5.4%), *TSC1* (7.1%) and *ATM* (12.5%). *ERG* altered at 5.4% frequency, is a known fusion partner in prostate cancer [134]. Mutations dominant in squamous carcinoma include *BRCA1, TMPRSS2, NOTCH1*, *NOTCH2* and *RB1* along with other cancer-associated genes as shown in II-Figure 7.



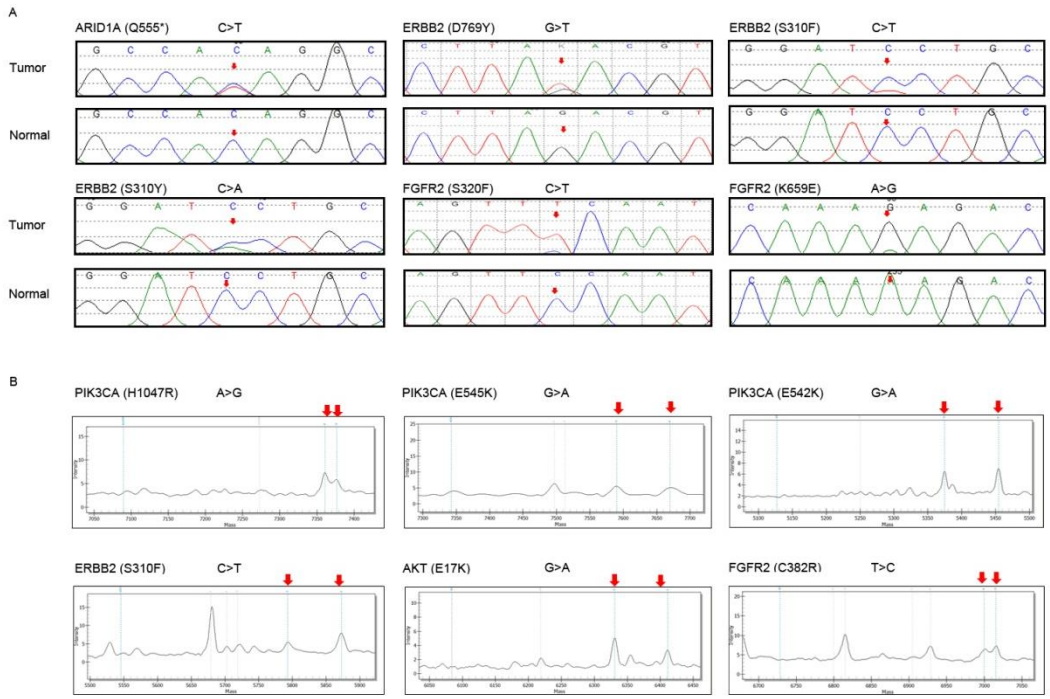
II-Figure 7: Heatmap showing mutations in cervical cancer hallmark genes and other cancer-associated genes for both histological subtypes. WES samples are represented by blue, WTS samples by green and WGS samples by a yellow box. An overlap of 2 samples

with both WES and WGS data is indicated in orange and overlap of one sample with WES and WTS by light blue. Samples with MassArray data is shown in grey whereas Sanger sequenced samples are indicated by black colour. For adenocarcinoma subtype, age > 51 are indicated by black boxes, age < 51 are indicated by white boxes. FIGO stages IIB and above are late stages shown by black boxes, early tumor stage is shown in the grey box and white box refers to information not available. Disease relapse is indicated by a black box, grey box refers to no relapse and white box refers to data not available. HPV positive sample is denoted by a black box, HPV negative by white box and HPV infection data not available is indicated by a grey box. In the case of mutation data, black box refers to mutation, white box for wildtype and grey box refers to samples with no mutation data available. The clinical data for squamous samples is not available and is indicated by the white box. Genetic alteration frequency of this study along with and TCGA for both subtypes of cervical cancer is shown towards the right. Percent frequency of nucleotide substitutions and mutation rate (mutations/Mb) for individual samples is shown at the bottom.


Adenocarcinoma (n=84)
Squamous carcinoma (n=15)
NGS
MassArray/Sanger sequencing
Age
FIGO Stage
Relapse
HPV
Gene_Mutation
PIK3CA_p.E545K
PIK3CA_p.E542K
PIK3CA_p.H1047R
PIK3CA_p.H450D
PIK3CA_p.IL112fs
PIK3CA_p.-1067fs
ERBB2_p.D769Y
ERBB2_p.S310F
ERBB2_p.S310Y
ERBB2_p.R190W
ERBB2_p.L403P
ARID1A_p.Q555*
ARID1A_p.Q538*
ARID1A_p.Q402*
ARID1A_p.Q780H
ARID1A_p.-1631fs
ARID1A_p.-1657fs
ARID1A_114_115
ARID1A_-1848fs
ARID1A_p.P1860H
CREBBP_p.E1536K
CREBBP_p.G1404S
CREBBP_p.L551I
CREBBP_p.P1905S
CREBBP_p.A1570P
CREBBP_p.L551I
EP300_p.E1365K
EP300_p.Q485R
EP300_p.D1399N
NF1_p.R1306*
NF1_p.R1142G
FAT1_p.Q906*
GNAS_p.R844C
PTEN_p.V45fs
PTEN_p.R130Q
TSC2_p.L1216I
KMT2D_p.Q791fs
KMT2D_p.S355fs
ATRX_p.E650Q
ATRX_p.L2327H
KMT2C_p.G838S
KMT2C_p.Y987H
KMT2C_p.C988F
KMT2C_p.Q3414*
KMT2C_p.Q2054*
KMT2C_p.Q38498
KMT2C_p.R894W
KMT2C_p.C1114R
KMT2C_p.R886C
KMT2C_p.R894Q
KMT2C_p.R886H
KMT2C_p.R909G
KMT2C_p.G315S
KMT2C_p.P3901H
KMT2C_p.G845E
KMT2C_p.I882T
KMT2C_p.C988F
KMT2C_p.R909K
KMT2C_p.T820I
KMT2C_p.Y987H
KMT2C_p.R284Q
KMT2C_p.A1685S
KMT2C_p.R380L
KMT2C_p.W858L
KMT2C_p.P309S
KMT2C_p.P860S
KMT2C_p.S990G
KMT2C_p.D348N
KMT2C_p.R973G
KMT2C_p.E864G
KMT2C_p.S764F
KMT2C_p.T316S
KMT2C_p.G908C
KM2TC_p.T1636P
LRP1B_p.P2320Q
LRP1B_p.L1392F
LRP1B_p.C2903G
LRP1B_p.R2856W
LRP1B_p.F1974S
LRP1B_p.S179N
LRP1B_p.L4172S
FAT4_p.D2022A
FAT4_p.G3524D
FAT4_p.R4234*
RNF213_p.M319T
RNF213_p.L3932R

II-Figure 8: Heatmap showing individual mutations of cervical cancer hallmark genes. WES samples are represented by blue, WTS samples by green and WGS samples by a yellow box. An overlap of 2 samples with both WES and WGS data is indicated in orange and overlap of one sample with WES and WTS by light blue. Samples with MassArray data is shown in grey whereas Sanger sequenced samples are indicated by black colour. For adenocarcinoma subtype, age > 51 are indicated by black boxes, age < 51 are indicated by white boxes. FIGO stages IIB and above are late stages shown by black boxes, early tumor stage is shown in the grey box and white box refers to information not available. Disease relapse is indicated by a black box, grey box refers to no relapse and white box refers to data not available. HPV positive sample is denoted by a black box, HPV negative by white box and HPV infection data not available is indicated by a grey box. In the case of mutation data, black box refers to mutation, white box for wildtype and grey box refers to samples with no mutation data available. The clinical data for squamous samples is not available and is indicated in white.

Somatic mutations in *PIK3CA* genes are commonly reported for cervical cancers [94-98] and also identified in our study. Most of the mutations belonged to extracellular domain-E545K and E542K. In addition, we find an oncogenic *AKT1* E17K mutation known to activate PI3K/AKT pathway [135]. Mutations in the extracellular domain of *ERBB2* at S310F and S310Y observed in adenocarcinoma subtype were also reported by several other studies for cervical adenocarcinoma [95-97]. However, mutations in *FGFR2* are not common as no mutations are reported for adenocarcinoma subtype in the TCGA dataset. We report three oncogenic mutations C382R, S310F and K659E in *FGFR2* from this study. Overall, the data suggests recurrent mutation in actionable targets belonging to PI3K/AKT and MAPK pathway. Few of the mutations detected by Sanger sequencing and MassArray approach is shown in II-Figure 9



II-Figure 9: Validation of mutations by A) Sanger sequencing and B) Mass array genotyping in cervical adenocarcinoma samples.

In cervical adenocarcinoma, several genes were mutated in a single sample and were non-recurrent. However, these genes belonged to a specific signalling pathway. Shown below, is a tabular compilation of genes mutated in our dataset belong to several cancer signalling pathways. Six genes *PIK3CA, PTEN, TSC1, TSC2, mTOR* and *AKT1* are members of the PI3K-AKT signalling pathway. Somatic alterations in *PIK3CA, mTOR* and *AKT* are oncogenic activating the signalling pathway which results in increased cellular proliferation and survival whereas tumor suppressor *PTEN, TSC1* and *TSC2* negatively regulate the signalling of PI3K-AKT pathway [136]. Recurrent mutations in cervical adenocarcinoma are observed in Receptor tyrosine kinases such as *ERBB2, FGFR2* whereas singleton mutations in ERBB3 and FGFR3. Other mutations are seen in MAP2K2 and MAPK1 which activates the MAPK/ERK pathway. Mutations in *ERBB2* (D769Y, S310F/Y) and *FGFR2* (K659E, S320F, C382R) are activating oncogenic mutations known to activate MAPK signalling [137-139] whereas mutations in *FGFR3* (H350Y) and *ERBB3* (A130T, T1254R) have not been reported earlier but are predicted to be deleterious by our analysis. Another class of

genes showing recurrent somatic mutations are members of chromatin remodelling complex *ARID1A, KMT2C, KMT2D, EP300, BRD3, BRD4, NSD1* and *PBRM1* genes have also been reported to be mutated in several solid tumors [140]. Inactivating *ARID1A* mutations or loss of expression is common in gynaecological cancers like ovarian, endometrial and cervical [141, 142]. We observed loss of function mutations in *ARID1A* from our analysis.

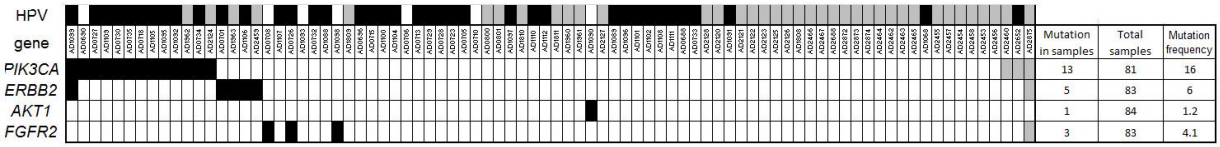
Members of the WNT signalling pathway also harbour somatic mutations. Canonical WNT signalling pathway LRP receptors- *LRP5, LRP6* are mutated. Four mutations observed in genes *APC, YAP* and *GSK3β* which belong to disruption complex that phosphorylates *β*-catenin for proteasomal destruction and inhibit WNT signalling. Expect for *APC* (R1640Q), all other mutations are novel, predicted to be deleterious. Wnt ligands- *WNT7B* (2.4%), reported in COSMICdb and novel *WNT9B* are also mutated in 2 samples. Another gene *CREBBP* is mutated across 4 samples. *CREBBP* associates with *β*-catenin and promotes transcription of genes promoting cellular proliferation [143]. Mutated genes belonging to different signalling pathway is shown in II-Table 10.

Pathways deregulated	**Mutated genes in the dataset**
PI3K/AKT signalling	*PIK3CA, PTEN, TSC1, TSC2, mTOR, AKT1*
Receptor tyrosine kinase and MAPK signalling	*ERBB2, ERBB3, FGFR2, FGFR3, MAP2K2, MAPK1*
WNT signalling pathway	*CREBBP, APC, LRP10, WNT7B,WNT9B, LRP6, GSK3B, YAP1, LRP5*
Epigenetic regulators	*ARID1A, KMT2C, EP300, BRD3, NSD1, KMT2D, PBRM1, BRD4, CREBBP*

II-Table 10: Mutation in genes belonging to different cancer pathways

Overall, from the mutation profiling studies, several therapeutically relevant mutations were identified in genes *PIK3CA, ERBB2, AKT1* and *FGFR* (II-Table 11) for which small-molecule inhibitors and antibodies are available. Alpelisib and fulvestrant have been used for the treatment of advanced breast cancer patients with *PIK3CA* mutations- E545K, E542K and H1047R has resulted in improved progression-free survival [144]. For *ERBB2* S310F/Y and D769Y, Trastuzumab, lapatinib and neratinib have shown clinical efficacy in breast cancer [145]. AZD5363 inhibitor is in use for *AKT1* mutation-positive breast and cervical patients [146]. BGJ398 and Ponatinib have been effective in cells expressing oncogenic K659E and C382R mutation of *FGFR2* gene [137, 147]. However, no patient clinical data is available.

Shown in II-Table 11 below, one sample shows mutations in both *PIK3CA* and *ERBB2* gene whereas all other mutations are mutually exclusive.

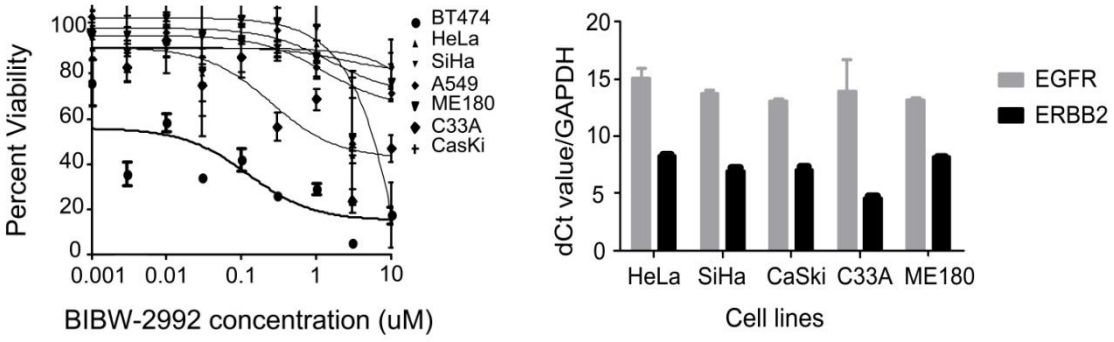


II-Table 11: Compilation of mutations in therapeutically relevant genes of 84 samples of cervical adenocarcinoma. Black box indicates the presence of the mutation, the white box indicates the wildtype gene and grey refers to information not available. Several samples sequenced for mutation detection for each gene is shown towards the right-hand side, followed by mutation frequency. For HPV, the black box indicates the presence of infection, white refers to HPV negative samples and grey box refers to no information available.

2.4.5 Role of ERBB signalling in cervical carcinogenesis

2.4.5.1 C33A cervical carcinoma cells are sensitive to treatment with Afatinib inhibitor but not dependent on ERBB2 signalling for cell survival.

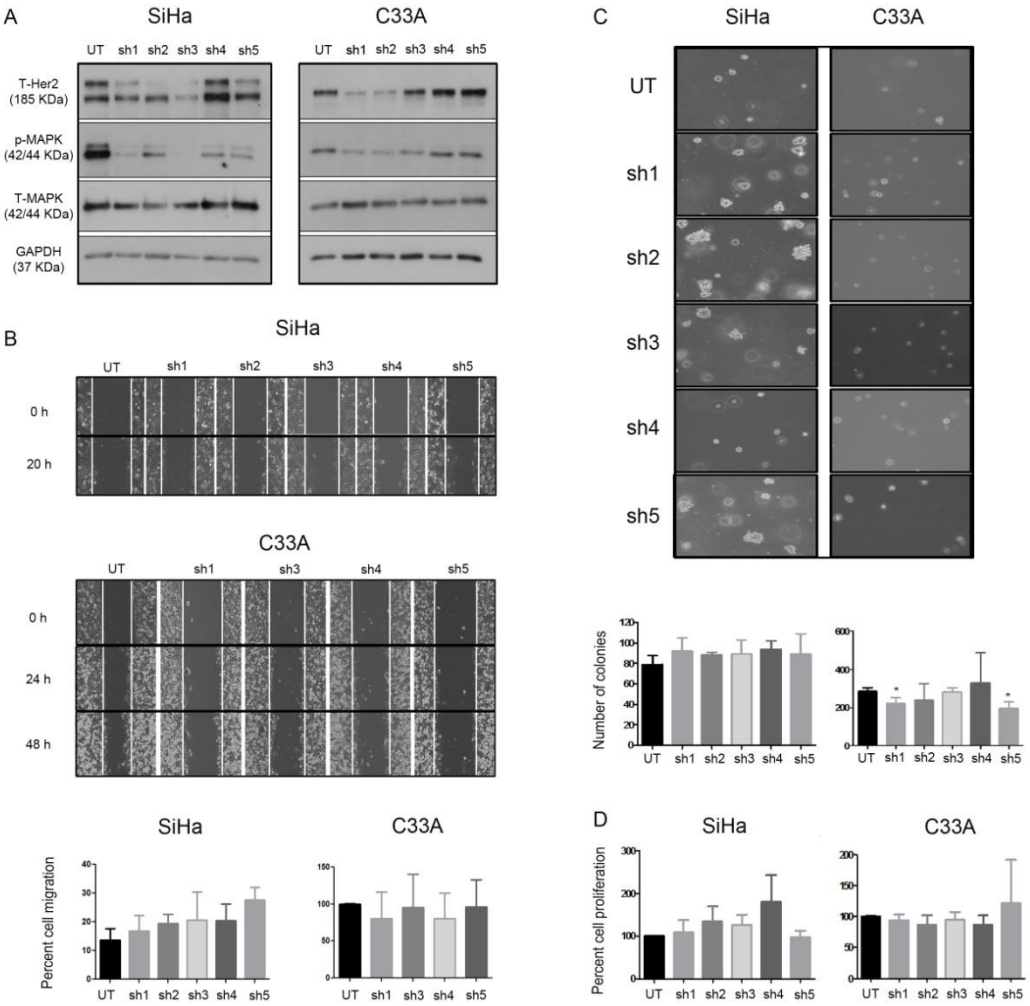
Cervical cancer cells HeLa are derived from adenocarcinoma of cervix whereas other available cancer cells SiHa, ME180, CaSki and C33A belonged to cervical squamous carcinoma. These five cervical cancer cells were treated with Afatinib, an inhibitor that targets three ERBB family members *EGFR*, *ERBB4* and *ERBB2*. Cell viability was assessed by MTT Assay for the cervical cells along with A549 and BT474 as resistant and sensitive control cells. C33A was the most sensitive cell line similar to BT474 as compared to other cervical cell lines. In addition, *EGFR* and *ERBB2* expression was assessed in cervical cells by real-time PCR. The over-expression of *ERBB2* was detected in C33A cells as compared to other cervical cells. Refer II-Figure 10.



II-Figure 10: MTT assay of cervical cancer cells with afatinib inhibitor and mRNA expression of *EGFR* and *ERBB2* in cervical cancer cells.

Since C33A cells were sensitive to Afatinib treatment; *ERBB2* expression was checked in cervical cells at both RNA and protein level. Indeed C33A cells showed increased expression of the *ERBB2* receptor as compared to other cells (II-Figure 10). To investigate the dependency of C33A cells on ERBB2 signalling for sustaining cellular proliferation, *ERBB2* knockdown was performed using shRNA approach. ERBB2 depletion in C33A and afatinib resistant SiHa cells resulted in decreased phosphorylation of MAPK as compared to control cells, however, no significant difference was observed between the proliferation, migration and anchorage-independent growth of parent and knockdown clones (II-Figure

11). These results collectively suggest that cervical cells C33A are not dependent on ERBB2 signalling for survival and the sensitivity to afatinib inhibitor is probably imparted due to inhibition of MAPK signalling mediated by *EGFR* and *ERBB4.*

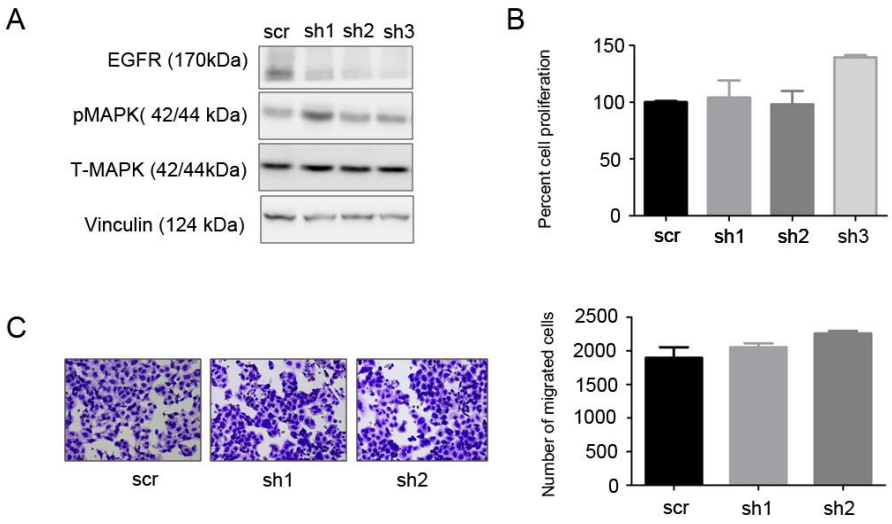


II-Figure 11: Effect on *ERBB2* knockdown in cervical cancer cells. A) Western blotting confirmation of ERBB2 knockdown in C33A and SiHa cells. B) Scratch wound assay C) Soft agar assay D) cell proliferation assay of control and knockdown clones of C33A and SiHa.

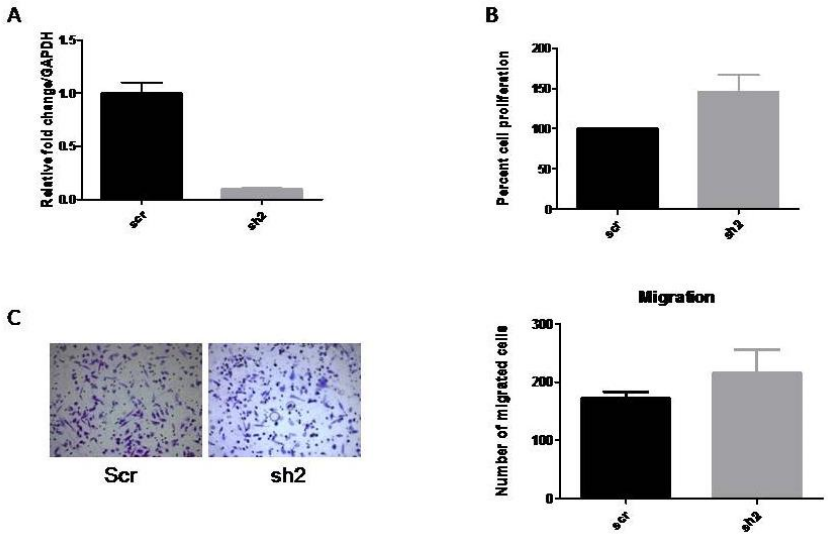
2.4.5.2 Depletion of EGFR and ERBB4 individually in C33A and SiHa cells has no effect on cell survival

Next, we investigated if *EGFR* or *ERBB4* are playing a role in conferring cell survival in C33A and SiHa cells which were observed to be sensitive and resistant respectively by

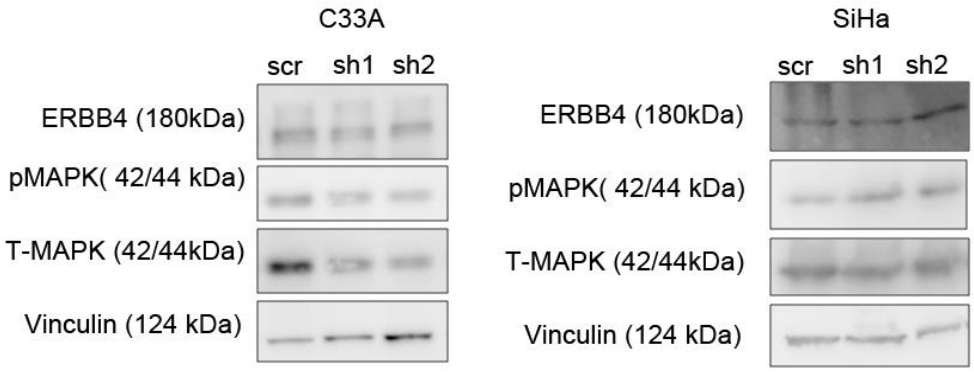
Afatinib treatment. Knockdown of *EGFR* and *ERBB4* was done in C33A and SiHa cells. Here, three shRNA constructs were used for knockdown of *EGFR* in C33A cells and one shRNA construct against *EGFR* in SiHa cells. The shRNAs were effective in depleting EGFR expression in C33A and SiHa cells. However, there was no decrease in phosphorylation of MAPK in C33A cells upon knockdown (II-Figure 12). Similarly, upon knockdown of *EGFR* in SiHa, there was no attenuation of MAPK signalling and no reduction in cellular proliferation or migration observed (II-Figure 13). Moreover, proliferation and migration potential of control and knockdown clones remained unaffected. Next, three shRNA used for knockdown of *ERBB4* in C33A and SiHa cells were inefficient in depleting the protein expression and therefore no change was observed in phosphorylation of MAPK in control and knockdown clones (II-Figure 14).



II-Figure 12: Knockdown of *EGFR* in C33A cervical cells. A) Western blotting confirmation of knockdown in 3 shRNA constructs B) Cell proliferation assay C) Migration assay of control and knockdown clones.



II-Figure 13: Knockdown of *EGFR* in SiHa cervical cells. A) Real-time PCR confirmation of *EGFR* knockdown B) Cell proliferation assay C) Migration assay of control and knockdown clones.



II-Figure 14: Western blotting to assess knockdown of *ERBB4* in C33A and SiHa cells

Collectively, these results suggest that individual knockdown of *EGFR, ERBB4* or *ERBB2* has no effect in reducing cellular viability. However, targeting all this receptor together does show an impact on survival as observed from afatinib inhibitor studies. This could probably occur because of several reasons. First, knockdown of the individual member of ERBB family might have resulted in homodimerization and heterodimerization of other ERBB family receptors and novel partners [113, 114] and therefore the ERBB signalling is unaffected. Second, some studies have shown that knockdown of one ERBB receptor results

in increased expression of other receptors, thus taking over the function [112]. Therefore, to attenuate ERBB signalling, receptors forming homo and heterodimerization must be blocked.

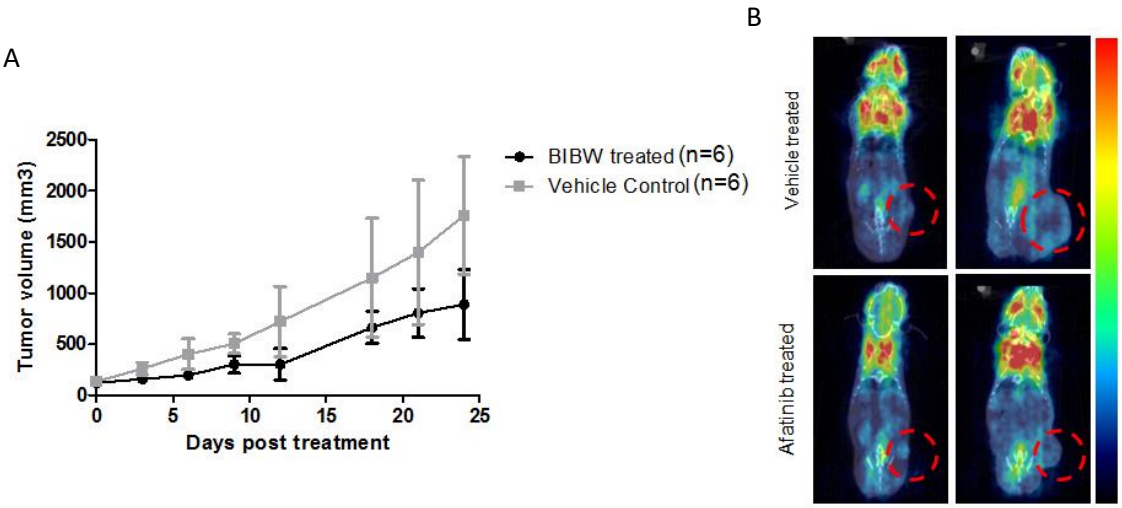
2.4.5.3 C33A tumors show a delayed growth on Afatinib treatment in *in-vivo* studies

To validate the findings of the *in-vitro* inhibitor studies, *in-vivo* studies were performed in C33A and SiHa tumor grafted NOD-SCID mice. Mice were treated with Afatinib/ BIBW-2992 and vehicle control for 24 days. We observed that C33A tumors displayed sensitivity to treatment with Afatinib. This observation supports our in-vitro inhibitor studies. C33A tumor of the treatment group showed an initial increase in tumor volume and then the growth was flattened post 15 days as compared to the control group.18F-FDG PET and CT scan imaging is shown below provide a visual observation of tumor growth in the control and treatment group (II-Figure 15). Standard Update Value (SUV) is a semi-quantitative measure of tracer uptake in the tumor region normalized with injected activity. SUV value for both the groups is presented in II-Table 12.

Sample (C33A tumors)	Standard Uptake Value (SUV) base/end
BIBW treated	114/502
Vehicle control	128/2385

II-Table 12: Standard uptake value (SUV) of base (day = 0) and end time point (day = 24) for vehicle control and treatment group mouse is shown. SUV value of control group mouse is 4.7 times more than BIBW treated mice.

C33A tumors



II-Figure 15: Female NOD-SCID mice with C33A tumors are sensitive to Afatinib treatment. A- 6 mice each bearing C33A tumors were subjected to treated with vehicle control and afatinib (20 mg/kg) for 24 days. The graph shows tumor growth trend in control and treatment growth. B- PET/CT scan image of the control and treatment group mouse at the beginning and end of the treatment. Dotted circle refer to tumor growth. The gradient intensity scale for the uptake of 18F-FDG is shown.

In the case of SiHa tumors, no difference was observed in the treatment and control group. SiHa grafted tumors displayed abrupt growth pattern wherein tumor volumes reach humane endpoints (>2000 mm^3) within 5 days of treatment for most of the mice in each group suggesting that inhibitor does not suppress tumor growth in SiHa cells. Therefore, the treatment was terminated and tumor tissues were collected and stored. This observation was consistent for three sets of experiments performed and also supports our *in-vitro* findings. The data is shown in II-Table 13.

SiHa tumors

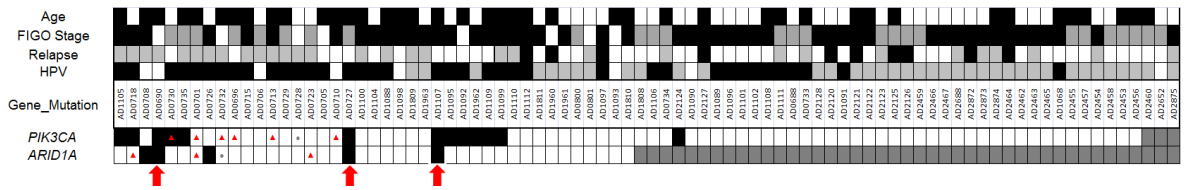
Mice no	Vehicle control (n=5) tumor volume (mm^3)		BIBW treated (n=4) tumor volume (mm^3)	
	Day 0	Day 4	Day 0	Day 4
1	130.977	1382.4	141.538	3611.63
2	88.4835	2051.43	134.019	1748.39
3	207.831	3216.7	171.088	2167.63
4	178.596	3023.79	159.528	3670.09
5	102.69	1717.72		
Average	141.715	2278.41	151.543	2799.43

II-Table 13: Tumor volume (mm^3) of female NOD-SCID mice bearing SiHa tumors at the beginning and after 4 days of treatment in Afatinib and vehicle control group is shown. Tumor growth on Day 4 has reached humane endpoints (>2000 mm^3) in several mice of both groups.

2.4.6 Role of co-occurring *ARID1A* and *PIK3CA* mutations in the oncogenesis of cervical cancer

Literature reports suggest that loss of ARID1A expression in *PIK3CA* mutation background confers sensitivity to inhibition to PI3K and AKT inhibitor [148] by downregulating PI3K-AKT signalling pathway. ARID1A negatively regulates the PI3K-AKT pathway and loss of ARID1A enhances phosphorylation of AKT.

In cervical cancer TCGA data, 7.5% of samples exhibit co-occurring *PIK3CA* and loss of function *ARID1A* mutations. Consistent with this observation, we note three samples having *PIK3CA* and *ARID1A* alterations from variant analysis (II-Figure16).

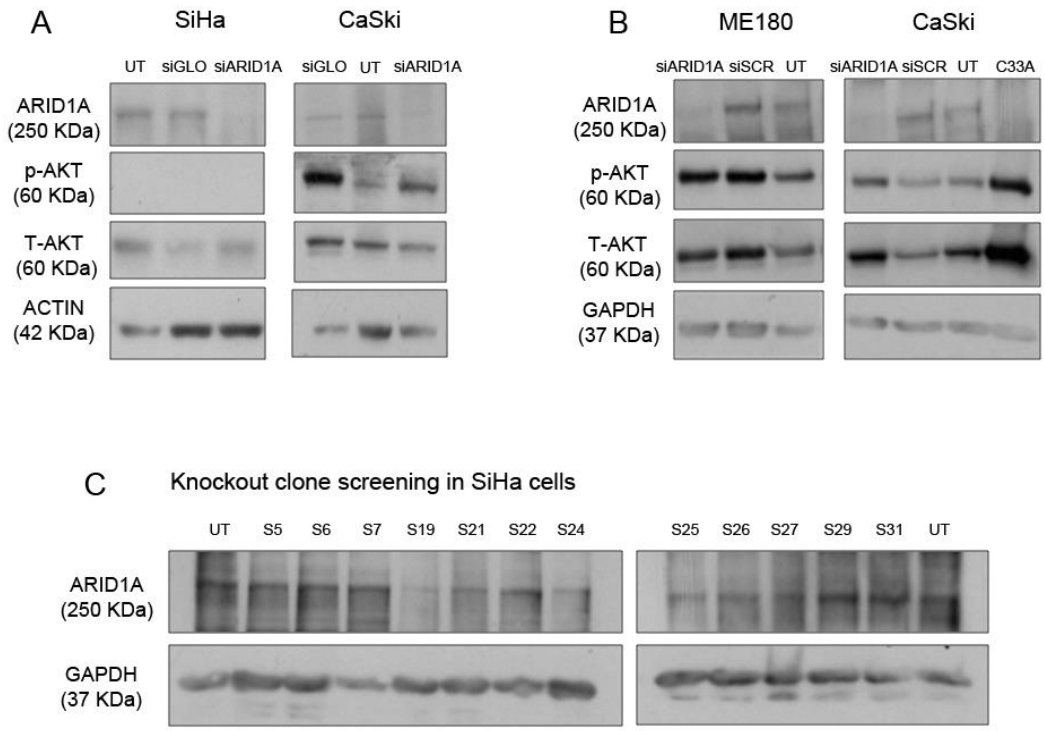


II-Figure 16: Heatmap representation of co-occurring *PIK3CA* and *ARID1A* mutations in cervical adenocarcinoma samples. The red triangle refers to copy gain, black box for

mutation and grey box to information not available. Samples indicated with red arrows show co-occurring *PIK3CA* and *ARID1A* mutations.

Since the role of co-occurring *PIK3CA* and *ARID1A* mutations in cervical cancer has not been explored previously, we performed the investigation. SiHa cervical cancer cells are wildtype for both the gene *PIK3CA* and *ARID1A* whereas ME180 and Caski are wildtype for *ARID1A* and mutant for *PIK3CA* (p.E545K). Knockdown was performed in all 4 cervical cancer cells using siRNA against *ARID1A* using siGLO as transfection control. Knockdown of *ARID1A* in SiHa (with wild type *PIK3CA* gene) did not show any effect in PI3K/AKT pathway activation. But upon depletion of ARID1A in *PIK3CA* mutant cell line CaSki, an increase in phosphorylation of AKT was observed as expected. However, siGLO transfected cells also displayed an increase in pAKT levels showing non-specific effect. Therefore, it was difficult to conclude whether increased phosphorylation of AKT is induced due to ARID1A knockdown or non-specific effect of siRNA on some other gene (II-Figure 17A). Next, siGLO was replaced with siSCR (from OriGene) and knockdown was performed in two *PIK3CA* mutant cell lines ME180 and CaSki. We observe that loss of ARID1A induced increased pAKT levels in CaSki cell line suggesting activation of the PI3K-AKT pathway but not in ME180 cells (II-Figure 17B).

To obtain stable loss of ARID1A expression, CRISPR-Cas9 mediated knockout was performed in SiHa and CaSki cells. In SiHa cells (wildtype for *PIK3CA* and *ARID1A*) screening of 14 single cell clonal populations by T7 endonuclease assay identified 10 potential knockout clones (data not shown). Western blotting of 12 clones reveals two clones with decreased ARID1A expression whereas all other clones show ARID1A expression similar to parental SiHa cells (II-Figure 17C).



II-Figure 17: Depletion of ARID1A expression in the cervical cancer cell lines. A- Knockdown of *ARID1A* in *PIK3CA* wild type SiHa and *PIK3CA* mutant CasKi cells by siRNA using siGLO as transfection control. B- Knockdown of *ARID1A* in *PIK3CA* mutant ME180 and CasKi cells by siRNA using siSCR as transfection control. C- Screening of CRISPR knockout clones in SiHa cells for ARID1A expression loss. Clone S19 and S24 show decrease in ARID1A expression as compared to parental cells (UT).

ARID1A knockout clones could not be generated in Caski cell lines as cell did not survive after puromycin selection after repeated attempts after transduction with LentiCRISPR virus. The ARID1A shRNA sequences are now cloned in pEGFP vectors and shRNA mediated knockdown has to be performed.

2.4.7 Identification of HPV infection and integration in the human genome in cervical cancer samples

17 paired samples and 1 orphan tumor (n=35) sample of cervical adenocarcinoma were subjected to Cancer Pathogen Detector (CPD) analysis. CPD output detects HPV type, read counts supporting HPV genome and ppm values. FeatureCounts represent the number of

reads supporting different genes of HPV Genome. To detect HPV presence with high confidence, a filter of 1 ppm was applied.

Out of 35 samples, HPV was detected in 9 samples. HPV16 and HPV18 strains were detected in 4 and 1 samples respectively. To further identify the integration of HPV in the human genome, HPVDetector tool was used. HPVDetector was able to identify integration in 4 samples- 1 HPV18 and 3 HPV16 positive samples. Two samples are displaying HPV integration in *SUPT7L* and *ASXL2* gene lying in cytoband 2p23.2. All the other integrations in different cytobands have been reported in the literature [96]. *CNTNAP2* gene showing HPV16 integration is located in highly unstable common fragile site (CFS) region of the human genome [76]. II-Table 14 shows a combined output of CPD and HPVDetector integration. HPV genome is integrated into the exonic region for 8 samples and intronic region remaining genes 6 genes (II-Table 15).

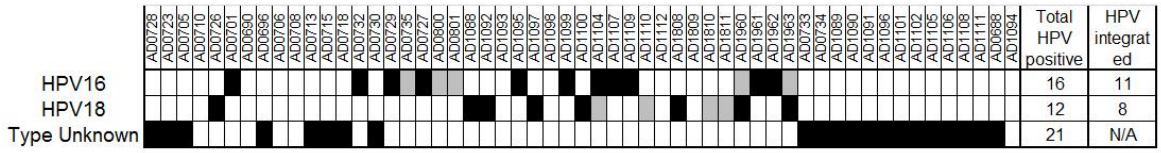
CPD output							HPVDetector: Integration output						
No	Sample	Pathogen	Genome Length	Read Count	PPM	FeatureCounts	HPV genome	Human chr	genomic coordinate	HPV gene	Human gene	Cytoband	Integration in known cytoband
1	AD0697_1N	NO											
	AD0728_1T	NO											
2	AD0722_2N	NO											
	AD0727_2T	NO											
3	AD0703_3N	NO											
	AD0723_3T	NO											
4	AD0704_4N	NO											
	AD0705_4T	NO											
5	AD0711_5N	NO											
	AD0710_5T	NO											
6	AD0719_6N	NO											
	AD0726_6T	HPV18	7857	14	0.22739	E2:4;E7:2;L1:6;L2:4;	3159	chr14	57686079	E2	*EXOC5*	q22.3	YES
7	AD0700_7N	NO											
	AD0701_7T	HPV16	7904	97	1.85713	E1:28;E2:18;E4:8;E5:2;E6:6;E7:6;L1:23;	5508	chr11	5344659	L2	*OR51B2*	p15.4	YES
8	AD0689_8N	NO											
	AD0690_8T	NO											
9	AD0695_10N	NO											
	AD0696_10T	NO											
10	AD0707_11N	NO											
	AD0706_11T	NO											
11	AD0709_12N	NO											
	AD0708_12T	NO											
12	AD0714_13N	NO											
	AD0713_13T	NO											
13	AD0716_14N	NO											
	AD0715_14T	NO											
14	AD0717_15N	NO											
	AD0718_15T	NO											
15	AD0691_16N	NO											
	AD0732_16T	HPV16	7904	439	5.44042	E1:94;E2:73;E4:41;E5:16;E6:33;E7:27;L1:107;L2:1;	1075	chr2	26033078	E1	*ASXL2*	p23.3	YES
							116	chr12	52188282	E6	*SCN8A*	q13.13	YES
							2491	chr22	35713794	E1	*TOM1*	q12.3	YES
							3527	chrX	1493447	E2	*IL3RA*	p22.33	YES
							3527	chrX	1493447	E4	*IL3RA*	p22.33	YES
							3842	chr7	146516510	E2	*CNTNAP2*	q35	YES
16	AD0698_17N	NO											
	AD0730_17T	NO											
17	AD0699_18N	NO											
	AD0729_18T	HPV16	7904	607	8.34161	E1:155;E2:79;E4:15;E5:30;E6:61;E7:60;L1:135;	1276	chr4	3352056	E1	*RGS12*	p16.3	YES
							1515	chr5	140565412	E1	*PCDHB16*	q31.3	YES
							1775	chr4	146059215	E1	*OTUD4*	q31.21	YES
							2841	chr22	43272196	E2	*PACSIN2*	q13.2	YES
							3551	chr8	8888033	E2	*ERI1*	p23.1	YES
							3551	chr8	8888033	E4	*ERI1*	p23.1	YES
							4875	chr3	52412598	L2	*DNAH1*	p21.1	YES
							4974	chr2	27883795	L2	*SUPT7L*	p23.3	YES
							5897	chr12	49522544	L1	*TUBA1B*	q13.12	YES
18	AD0735_19T	HPV16	7904	70	1.02795	E1:12;E2:16;E4:11;E5:3;E6:2;E7:1;L1:17;							

II-Table 14: Detection of HPV infection and integration in human genome from exome sequenced cervical adenocarcinoma samples

No	Cytoband	Recurrence (N=4)	Gene	Integration site
1	11p15.4	1	*OR51B2*	Exon
2	14q22.3	1	*EXOC5*	Exon
3	4p16.3	1	*RGS12*	Intron
4	5q31.3	1	*PCDHB16*	Exon
5	4q31.21	1	*OTUD4*	Exon
6	22q13.2	1	*PACSIN2*	Exon
7	8p23.1	1	*ERI1*	Exon
8	3p21.1	1	*DNAH1*	Intron
9	2p23.3	2	*SUPT7L, ASXL2*	Intron
10	12q13.12	1	*TUBA1B*	Exon
11	12q13.13	1	*SCN8A*	Exon
12	22q12.3	1	*TOM1*	Intron
13	Xp22.33	1	*IL3RA*	Intron
14	7q35	1	*CNTNAP2*	Intron

II-Table 15: HPV integration sites in the intronic and exonic region of genes.

A detailed heatmap representation of HPV infection from the NGS data comprising of the exome, transcriptome and whole-genome for 55 samples is shown in II-Table16.

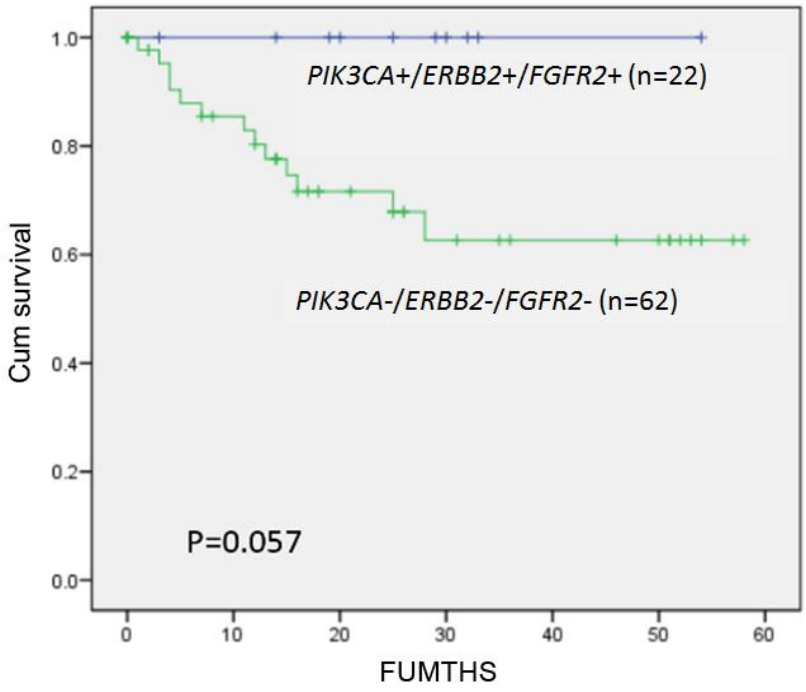


II-Table 16: Presence of HPV in integrated forms in 41 samples subjected to NGS sequencing and in 14 samples by PCR. Black box indicates the presence of HPV in integrated form; grey box refer to HPV in episomal form and the white box indicates the absence of HPV infection in samples.

2.4.8 Clinical correlation analysis with patient survival data

Kaplan- Meier survival analysis was performed on 84 patients with clinical follow-up for up to 5 years. As shown in II-Figure 18, patient tumors with a mutation in *PIK3CA*, *ERBB2* and *FGFR2* show a better trend of relapse-free survival (n=22, cumulative survival=100%) compared to patients lacking mutations in these genes (n=62, cumulative survival=60%).

This is a contrasting observation compared to literature reports. TCGA cervical patients with alterations in *ERBB2* display poor overall and disease-free survival. Another independent study suggests that patients with *ERBB2* over-expression have a poor prognosis [110]. TCGA data suggests no significant difference in overall survival in patients with and without *PIK3CA* mutations. However, a liquid biopsy study in Hong Kong Chinese women for *PIK3CA* mutations suggest than mutations in *PIK3CA* are associated with a significant decrease in disease-free and overall survival [149]. However, one study from the Chinese population is consistent with our finding wherein *PIK3CA* mutations conferred better treatment outcomes in the patients [150]. Since *FGFR2* mutations have not been reported in cervical adenocarcinoma subtype previously, we report for the first time that *FGFR2* mutations confer better relapse-free survival in patients.



II-Figure 18: Kaplan-Meier survival analysis of 84 cervical adenocarcinoma samples. Relapse free survival in patients was assessed for up to 60 months. Patients harbouring an oncogenic mutation in *PIK3CA, ERBB2* and *FGFR2* genes show better relapse-free survival than the patients with wildtype genotype.

2.5 Discussion

Cervical adenocarcinoma is poorly characterized subtype owing to low incidence rate. Cervical adenocarcinoma patients are less responsive to treatment by radiation and chemotherapy as compared to squamous counterparts [151]. Adenocarcinoma patients often present with recurrence and distant metastasis, resulting in poor overall survival as opposed to squamous carcinoma patients [152]. Therefore, there is a need to perform in-depth genomic characterization of this subtype to understand underlying molecular alterations that are distinct from the squamous carcinoma and fish for potential targets for precision medicine for improving treatment.

Here, we describe in detail the profiling of the somatic mutations in 84 patient samples of cervical adenocarcinoma using NGS approach and other genotyping methods. In addition, we re-analyzed WES data of squamous carcinoma for 15 samples, published previously. Adenocarcinoma subtype has a mutation rate of 6 mutations/ Mb whereas, for squamous carcinoma, it is 4.8 mutations/ Mb. In both subtypes, C>T transition is dominant and corresponds to mutational signature 1B, consistent with the literature report [130].

By integrating variant analysis from exome, transcriptome and whole-genome along with validation in additional samples by orthologous methods, we report recurrent mutations in oncogenes *ERBB2*, *FGFR2* and *PIK3CA* in cervical adenocarcinoma whereas mutations in *KMT2C, LRP1B* and *FAT4* were common in squamous carcinoma. Mutations in other cervical cancer COSMIC hallmark genes include *ARID1A*, *CREBBP*, *EP300*, *NF1*, *FAT1*, *PTEN* and *TSC2*. In addition, recurrent mutations in other epigenetic genes *KMT2C*, *KMT2D*, *EP300*, *BRD3*, *BRD4*, *NSD1* and *PBRM1* were observed but patients exhibited relapse-free survival. In contrast, a study mentioned that mutations in *KMT2C, KMT2D, KMT2A, KDM5C, EP300, CREBBP, ARID1A, ARID2* and *ATRX* show association with poor progression-free survival as opposed to patients lacking alterations in these genes [153].

Interestingly, we do not find mutations in the *KRAS* gene, which is found to be recurrently mutated in cervical adenocarcinomas from the Caucasian population [97, 98]. Moreover, our dataset report no mutations in commonly altered genes of cervical cancer such as *TP53*, *STK11* and *FBXW7* [95]. Mutations in *FGFR2* are reported at <1% frequency in cervical cancers and no mutations are reported in cervical adenocarcinoma till date. We report 3 oncogenic *FGFR2* mutations- K659E, S320F, C382R in cervical adenocarcinoma subtype for the first time in the Indian population. Several genes belonging to PI3K/AKT, Wnt/β-catenin and MAPK signalling pathway are mutated suggesting the potential role in promoting cervical carcinogenesis. PI3K/AKT pathway activation is common in cervical cancers as reported by several studies [154]. Mutation in genes upstream to MAPK such as *ERBB2, ERBB3* and *FGFR2* in our dataset suggests MAPK pathway activation. In addition, several epigenetic and Wnt/β-catenin pathway genes are mutated. Activation of MAPK and Wnt/β-catenin signalling has been reported in cervical cancer [155]. Loss of function mutation in epigenetic genes, mostly tumor suppressor *ARID1A* is common is adenocarcinoma subtype and is associated with poor prognosis [156]. Use of EZH2 inhibitor or inhibition of ARID1B in *ARID1A* deficient cells is known to inhibit cancer cell proliferation [157]. Moreover, we also observe co-occurring *PIK3CA* and *ARID1A* mutations in 3 samples of adenocarcinoma subtype, which in ovarian cancer has shown to render cancer cells sensitive to AKT and PI3K inhibitors MK2206 and buparlisib respectively [158]. Similar sensitivity of inhibitors in cervical cancer is speculated. Overall, mutation data analysis suggests several potential therapeutic targets in cervical adenocarcinoma which can be explored to utilize targeted therapy and improve patient survival. However, from clinical correlation analysis, an unanticipated result was noted. Patients with a mutation in *PIK3CA*, *ERBB2* and *FGFR2* show better relapse-free survival ($p<0.059$) as compared to patients lacking these mutations.

Since several somatic alterations in ERBB family members were noticed, the role of ERBB family members in promoting cervical carcinogenesis was investigated further using cell lines. Here, cervical cells subjected to Afatinib inhibitor treatment revealed that C33A and SiHa cells were sensitive and resistant to treatment, which was consistent with the *in vivo* findings as well. Since, afatinib targets EGFR, ERBB2 and ERBB4, each of the gene expression was inhibited using shRNA mediated knockdown to identify dependency of cells on a single ERBB member gene. However, single-gene knockdown of *ERBB2* and *EGFR* did not result in inhibition of cellular proliferation, migration or anchorage-independent growth of C33A cells. These results indicate that there is a potential role of other ERBB receptor interaction upon depletion of one member to continue signalling [113] or possible cross-talk with other signalling pathways like PI3K/ AKT [115], which remains to be validated further.

In brief, we describe the landscape of somatic alterations in cervical cancer using NGS and other orthologous approaches to identify known and novel cancer-associated genetic alterations. Further, the role of ERBB members in cervical carcinogenesis was explored using *in-vitro* and *in-vivo* approaches. We note that ERBB signalling in cervical cancer is much complicated and involves a complex interplay of several factors. Therefore, we speculate that inhibitors that disrupt receptor interaction or ERBB downstream pathways might be ideal candidates for inhibiting signalling and thereby, impeding cell proliferation [117].

Chapter III

Whole transcriptome sequencing to identify differentially expressed genes and fusion transcripts

Chapter III- Whole transcriptome sequencing to identify differentially expressed genes and fusion transcripts

3.1 Abstract

Background: Stratification of a patient based on gene expression profiles serves as a useful approach to predict relapse and response to treatment. Gene expression profiles also vary according to tumor histological subtypes and help in yielding useful information in identifying gene clusters. In this study, differential gene expression analysis has been performed among normal tissue and tumor tissues and among tumor tissues alone of cervical adenocarcinoma subtype to identify specific gene clusters to correlate with clinical information.

Material and methods: We performed transcriptome sequencing of 24 tumor samples and 5 normal cervix tissues to identify differential gene expression using different analysis methods- Tuxedo suite analysis, Salmon workflow and DESeq. Further, transcript fusions were predicted using STAR-fusion.

Results: Differential gene expression analysis between normal and tumor samples was done by both the methods. Tuxedo suite analysis identified only 10 significantly expressed genes between normal (n=4) and tumor samples (n=24). Salmon and DEseq analysis were able to cluster normal samples (n=4) together; however, no distinct gene clusters were identified in tumor samples (n=24). It was observed by real-time PCR that samples with bad quality RIN value displayed erroneous gene expression of house-keeping genes and therefore, these samples were excluded from further analysis. Next, we tried to identify differentially expressed genes between early-stage tumors (n=4) and late-stage tumors (n=18) of good RIN samples. Three out of 4 early-stage tumors were clustered together whereas one early-stage tumor showed an expression pattern similar to late-stage tumor. Among late-stage

tumors we could not find any distinct gene clusters. Analysis with an unequal number of samples in each group for comparison was inappropriate and therefore, yielded variable results. Eventually, gene expression analysis was performed between tumor samples considering only the genes expressed in the top 10% quartile from each sample and recurrent across 30% of the samples. We report up-regulation of genes belonging to MMP family and ERBB family members.

Fusion analysis using STAR-Fusion identified unique transcript fusion gene pairs, not reported earlier. The fusion pairs with at least one oncogenic partner are *IDH3G-PPP2R1A*, *U2AF1-CASP2*, *RAP2A-MECOM*, *PPP6C-CASC3* and *ANKRD27-MYC*. The functional role of these fusions in cervical cancer is unknown.

Conclusion: We describe gene expression pattern among tumor samples and report frequent upregulation of genes belonging to the MMP family and EGFR family. We also mention unique transcript fusion identified from the analysis.

3.2 Introduction

Cervical cancer is one of the common gynaecological cancers in India. The common treatment method consists of External beam radiation therapy (EBRT), brachytherapy, cisplatin based concurrent chemotherapy and radical hysterectomy. Although the treatment benefits initially, patients develop metastatic disease in 15-61% within two years [159]. The possible reason for the development of relapse and metastatic disease can be attributed to the fact that the above-mentioned treatments are targeting proliferative cells and do not consider the genetic alterations underlying tumor progression. The treatment is common for both the histological subtypes of cervical cancer. Recent studies suggest that adenocarcinoma subtype patients have overall poor survival and disease-free survival (DFS) as compared to squamous counterparts when treated with radiotherapy [66]. Hence, there is a need to classify the patients based on genetic alteration or gene expression profile to bring into practice targeted therapy. Tumor heterogeneity is one other factor that needs to be taken into account as it plays a vital role in predicting clinical outcomes [160]. The sub-clonal populations within the tumor show different degree of response to treatment and thereby confer resistance leading to relapse. Therefore, it is of utmost importance to have a multi-omics analysis of patient tumor samples to take an informed decision on treatment that can yield a positive outcome.

Gene expression analysis can provide useful insights into the stratification of patients based on unique expression profiles. One exhaustive study on larger sample size was done by the TCGA group wherein three mRNA clusters were identified from integrated analysis-squamous samples with high keratin expression, squamous samples with low keratin expression and adenocarcinoma specific cluster providing insights into potential therapeutic options that can be used for cervical cancer treatment [94]. Gene expression patterns also differ according to different stages of cervical cancer and these can serve as useful

prognostic markers to predict treatment outcomes [161]. A microarray study of cervical tumors and normal tissues, were able to identify differential gene expression pattern between the tumor stages. Tumors were classified based on response to radiotherapy and a set of genes favouring positive response to treatment were identified [162].

Gene fusion events are other useful therapeutic targets that have shown good clinical response to treatment. Treatment of *BCR-ABL* fusion protein with Imatinib and *EML4-ALK* rearrangement with Crizotinib is regular treatment in clinical settings for leukaemia and non-small cell lung cancer. Fusion genes occurring due to chromosomal rearrangements and involving oncogene or tumor suppressor gene may alter the expression of partner genes. Fusion with a known oncogene having functional kinase domain conserved may trigger constitutive expression of the gene and activation of the signalling pathway. Targeting fusion proteins is one of the potential therapeutic options for treatment. Fusion in some cancer types are common and are elaborately characterized whereas gene fusions in some cancer types such as cervical cancer need to study in additional samples to identify recurrent fusions with a potential role in cancer.

To address the above questions, we performed Transcriptome sequencing on cervical adenocarcinoma samples to identify gene expression patterns among tumor samples, gene fusion events, and the influence of expression of HPV oncogenes. Analysing RNA-sequenced samples from multiple methodologies can aid in identifying the potential driver that can be targeted.

3.3 Material and methods

3.3.1 Patient information

For transcriptome sequencing, 24 tumor tissues and 5 normal cervix tissue were processed. The sample information is provided in III-Table 1.

3.3.2 Extraction of RNA and sample QC

Total RNA was extracted from tumor tissue and available normal tissue samples using TRIzol reagent (Thermo Fisher Scientific), as previously described [163]. Total RNA concentration was quantified using Qubit Fluorometer (Thermo Fisher Scientific). The quality of RNA was assessed by performing a Tape station analysis using Agilent 2200 Tape station system (Agilent Technologies). Good and poor quality RNA samples with RNA integrity number (RIN) value shown in the table below were selected for further transcriptome sequencing. The QC information is shown in III-Table 1.

No.	Adeno carcinoma Sample	RIN	Tissue type
1	AD0722	1	Normal adjacent tissue
2	AD0724	1	Normal adjacent tissue
3	AD0685	3.2	Normal adjacent tissue
4	AD0800	3.3	Tumor tissue
5	AD1097	3.6	Tumor tissue
6	AD0702	4.2	Normal adjacent tissue
7	AD0801	4.4	Tumor tissue
8	AD0703	4.7	Normal adjacent tissue
9	AD0727	6.8	Tumor tissue
10	AD1811	7.1	Tumor tissue
11	AD1088	7.3	Tumor tissue
12	AD1095	7.5	Tumor tissue
13	AD1093	8	Tumor tissue
14	AD1104	8.1	Tumor tissue
15	AD1099	8.3	Tumor tissue
16	AD1808	8.4	Tumor tissue
17	AD1110	8.5	Tumor tissue
18	AD1092	8.6	Tumor tissue
19	AD1109	8.6	Tumor tissue
20	AD1098	8.7	Tumor tissue
21	AD1112	8.7	Tumor tissue
22	AD1810	8.7	Tumor tissue
23	AD1100	8.8	Tumor tissue
24	AD1107	8.8	Tumor tissue
25	AD1809	8.9	Tumor tissue
26	AD1960	6.5	Tumor tissue
27	AD1961	8.3	Tumor tissue
28	AD1962	8.1	Tumor tissue
29	AD1963	8.7	Tumor tissue

III Table 1: RIN values of samples used for transcriptome sequencing

Transcriptome sequencing was performed for 24 tumor and 5 normal samples of cervical adenocarcinoma by Medgenome Labs Ltd. Library preparation was performed using SENSE Total RNA-Seq Library Prep Kit from Lexogen (Lexogen Inc), a specialized kit designed for library preparation from poor quality RNA. The ribosomal RNA (rRNA) was depleted from total RNA using RiboCop rRNA Depletion Kit V1.2 (Lexogen Inc) as per the manufacturer's protocol. In brief, probes hybridize to different rRNA species- 28S, 18S, 5.8S, 45S, 5S, mt16S, mt12S which are then separated from the solution by magnetic bead separation method. Then the samples are run on Tape-station to verify depletion of rRNA, particularly for 18s and 28s rRNA peaks. RNA samples are then subjected to reverse transcription and ligation in a single tube employing starter-stopper heterodimer primers. Next, the enrichment of cDNA and the addition of barcode sequence by PCR amplification was done to ensure pooling of several samples during sequencing. The prepared libraries are run on using Agilent D1000 ScreenTape (Agilent Technologies) to ensure that the library consists of expected fragment size. Libraries were sequenced in Illumina Hiseq 4000 platform to yield 100 bp paired-end reads. The total output per sample was 60 million reads and above.

3.3.3 Identification of differential expressed transcripts among tumor sample

The preliminary data analysis was performed using the Tuxedo suite package [164]. Briefly, raw reads were aligned to human reference genome hg19 using Tophat and the alignment results were used to assemble aligned sequences into transcripts using Cufflinks. Assemblies were merged using Cuffmerge. Further, FPKM values were calculated and differential gene expression analysis between normal and tumor samples was performed using Cuffdiff.

Data obtained as Fastq files was subjected to data QC using FASTQC and Qualimap V2.2.1. Data analysis was done using Salmon workflow [165]. Reads were mapped to GRCh38, release 87 reference by using Quasi mapping based mode, specifying the library type

parameter as ‘auto’. Quantification files were generated for raw counts and TPM values. Transcript level quantification was done using Tximport and differentially expressed transcripts was obtained using DEseq [166]. Transcripts with adjusted p-value <0.05 and fold change >2 for up-regulated genes and <-2 for down-regulated genes were selected for further analysis. For generating a heatmap of differentially expressed transcripts, pheatmap R-package was used. Raw counts of transcripts across all the samples were obtained and transcripts having low expression (with median value < 1) across all the samples were excluded from further analysis. The values were log2 transformed and median centered for each transcript to provide input for heatmap generation. This methodology was followed to study differential gene expression analysis between normal and tumor samples and among tumor samples with different FIGO stages.

Next, data analysis was also done using Salmon workflow [165]. Samples with RIN value greater than 6.8 were considered for further analysis. Reads were mapped to GRCh38 reference genome using Quasi mapping based mode and quantification files were generated for raw counts and Transcript per million (TPM) values. Raw counts of transcripts across all the samples were obtained and transcripts having low expression (with TPM < 1) across all the samples were excluded from further analysis. Gene expression was log-transformed and sample-wise perform sorting of genes based on the expression values. Top 10% quartile genes per sample were taken further. Coding genes expressed in at least 30% of the samples were considered. Gene expressed in normal cervix tissue (GTEx Portal) were depleted from the list.

3.3.4 Real-time PCR for gene expression analysis

cDNA was prepared using High-Capacity cDNA Reverse Transcription kit from 2ug of RNA (Thermo Fischer Scientific) as per manufacturer’s protocol. Real-time PCR was performed using KAPA cDNA master mix (KAPA SYBR FAST Universal qPCR kit), real-

time primers and diluted cDNA in 6ul reaction volume. Triplicate reactions were set and amplification was performed using the Light cycler 480 (Roche, Mannheim, Germany) instrument. *Tubulin* was used as a reference gene. Data was analyzed using $2^{-\Delta\Delta ct}$ method. The gene expression was assessed in all tumor and normal samples. We consider gene expression with a fold change of greater than or equal to 2 as up-regulated. The real-time was performed twice independently.

3.3.5 Identification of fusion transcripts by Starfusion tool

Transcript fusions were detected using Starfusion in both tumor and normal samples. Paired-end Fastq files were mapped to hg 19 human reference genome and discordant reads pairs were processed further to detect transcript fusions. Fusion transcript detected in both normal and tumor samples were excluded from further analysis. Fusion transcripts identified in normal fusion transcript databases [167, 168] was also eliminated. Filtered fusion transcripts were annotated using AGFusion and information of fusion frame, chimeric sequences were obtained. Fusion transcripts were further subjected to fusion inspector to obtain information of spanning and junction reads.

3.3.6 Expression of oncogenic HPV transcripts identified from the Cancer Pathogen Detector

As described previously, the Cancer Pathogen Detector (CPD) was used to identify HPV presence from NGS data. Counts of different HPV strains were represented as parts per million (ppm). Samples with ppm values greater than 5 were considered positive for infection with HPV pathogen. The integration of HPV in the human genome was identified using a tool HPVDetector.

3.4 Results:

3.4.1. Patient sample information

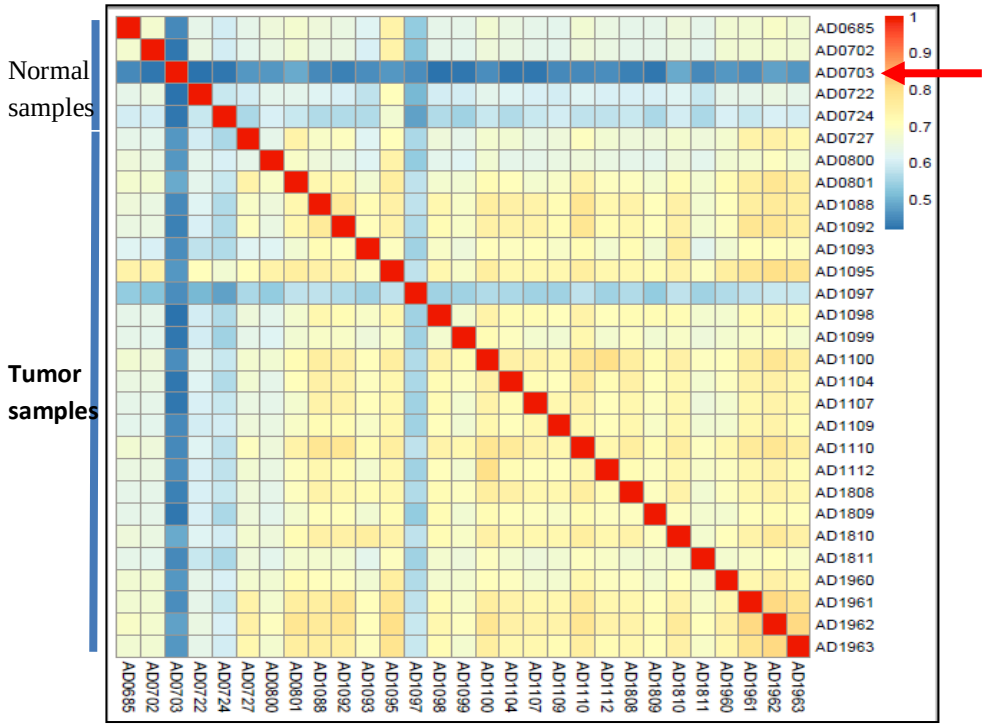
Twenty-four tumor samples of cervical adenocarcinoma and 5 normal tissue samples were analysed for differential gene expression, fusion transcripts and expression of E6 and E7 oncoproteins. In addition, we performed a variant calling from RNA-seq data. Our cohort of 24 tumor samples consist of patients with a median age of 53 years (range: 36-72 years), with 54 % of the patients belonging to Figo stage IIB, followed by FIGO stages IIIB at 13%, IB2 at 8% frequency and FIGO stages IIA2, IIIA at 4% frequency. FIGO stage information is not available for 17% samples. Tumor samples were collected from treatment naïve patients. Then, the patients are treated with radiation/ brachytherapy and chemotherapy.

3.4.2 Identification of differentially expressed transcripts among the tumor samples

RNA-sequencing was performed on 24 cervical adenocarcinoma tumor and 5 normal samples (n=29) with an average of 84 million reads per samples. The QC data performed on trimmed reads is shown in III-Table 2. Reads were mapped to reference genome (hg38) and genes. All the samples showed mapping of 85% and above (except for 1 sample showing 75% mapping). Percent alignment to genes was poor for 12 samples. In addition, 9 samples are showing intronic capture in the range of 27%-50%, which is in the acceptable range and has also been reported previously by other studies. The possible reason for finding the intronic region is due to the capture of nascent mRNA transcripts [169]. It is also observed that the Transcriptome sequencing performed after ribosomal depletion tends to show more of the intronic region capture [170-172]. Refer to III-Table 2. Among the 5 normal tissue samples sequenced, one sample AD0703 showed aberrant gene expression as compared to 4 other samples and was, therefore, excluded from further analysis. III-Figure 1 shows a correlation matrix for all RNA-sequenced samples.

No.	Sample ID	RIN number	Tissue type	No of reads (million)	Mapped reads	Percent mapped reads	Percent reads mapped to genes	Percent Exonic	Percent Intronic	Percent Inter-genic	Percent Intronic/ Intergenic overlapping exon
1	AD0722	1	Adjacent normal	68	60	88	63	73	19	8	1
2	AD0724	1	Adjacent normal	66	59	89	64	73	19	8	1
3	AD0685	3.2	Adjacent normal	59	53	90	59	68	23	9	1
4	AD0703	4.7	Adjacent normal	284	247	87	20	36	48	16	4
5	AD0702	4.2	Adjacent normal	68	63	93	63	72	19	8	1
6	AD0800	3.3	Tumor tissue	70	61	87	38	64	27	9	1
7	AD1097	3.6	Tumor tissue	112	88	79	46	62	28	9	2
8	AD0801	4.4	Tumor tissue	64	58	91	40	50	38	13	2
9	AD0727	6.8	Tumor tissue	70	65	93	35	39	50	11	2
10	AD1811	7.1	Tumor tissue	63	56	89	53	65	24	11	1
11	AD1088	7.3	Tumor tissue	72	66	92	70	79	14	7	1
12	AD1095	7.5	Tumor tissue	64	59	92	63	73	19	8	1
13	AD1093	8	Tumor tissue	69	65	94	77	86	9	5	1
14	AD1104	8.1	Tumor tissue	63	57	90	60	68	11	21	1
15	AD1099	8.3	Tumor tissue	65	60	92	71	78	13	9	1
16	AD1808	8.4	Tumor tissue	166	143	86	59	68	10	22	1
17	AD1110	8.5	Tumor tissue	74	69	93	69	77	16	8	1
18	AD1092	8.6	Tumor tissue	66	62	94	65	76	16	8	1
19	AD1109	8.6	Tumor tissue	64	59	92	65	77	13	11	1
20	AD1098	8.7	Tumor tissue	73	68	93	78	85	8	7	1
21	AD1112	8.7	Tumor tissue	107	100	93	63	82	10	8	1
22	AD1810	8.7	Tumor tissue	62	57	92	71	81	13	6	1
23	AD1100	8.8	Tumor tissue	107	98	92	68	81	8	12	1
24	AD1107	8.8	Tumor tissue	73	69	95	81	88	7	5	1
25	AD1809	8.9	Tumor tissue	63	59	94	67	78	12	10	1
26	AD1960	6.5	Tumor tissue	91	81	89	47	60	30	10	1
27	AD1961	8.3	Tumor tissue	99	92	93	49	61	28	11	1
28	AD1962	8.1	Tumor tissue	76	71	93	48	56	33	11	1
29	AD1963	8.7	Tumor tissue	78	73	94	48	59	31	10	1

III-Table 2: QC data of Transcriptome sequenced samples



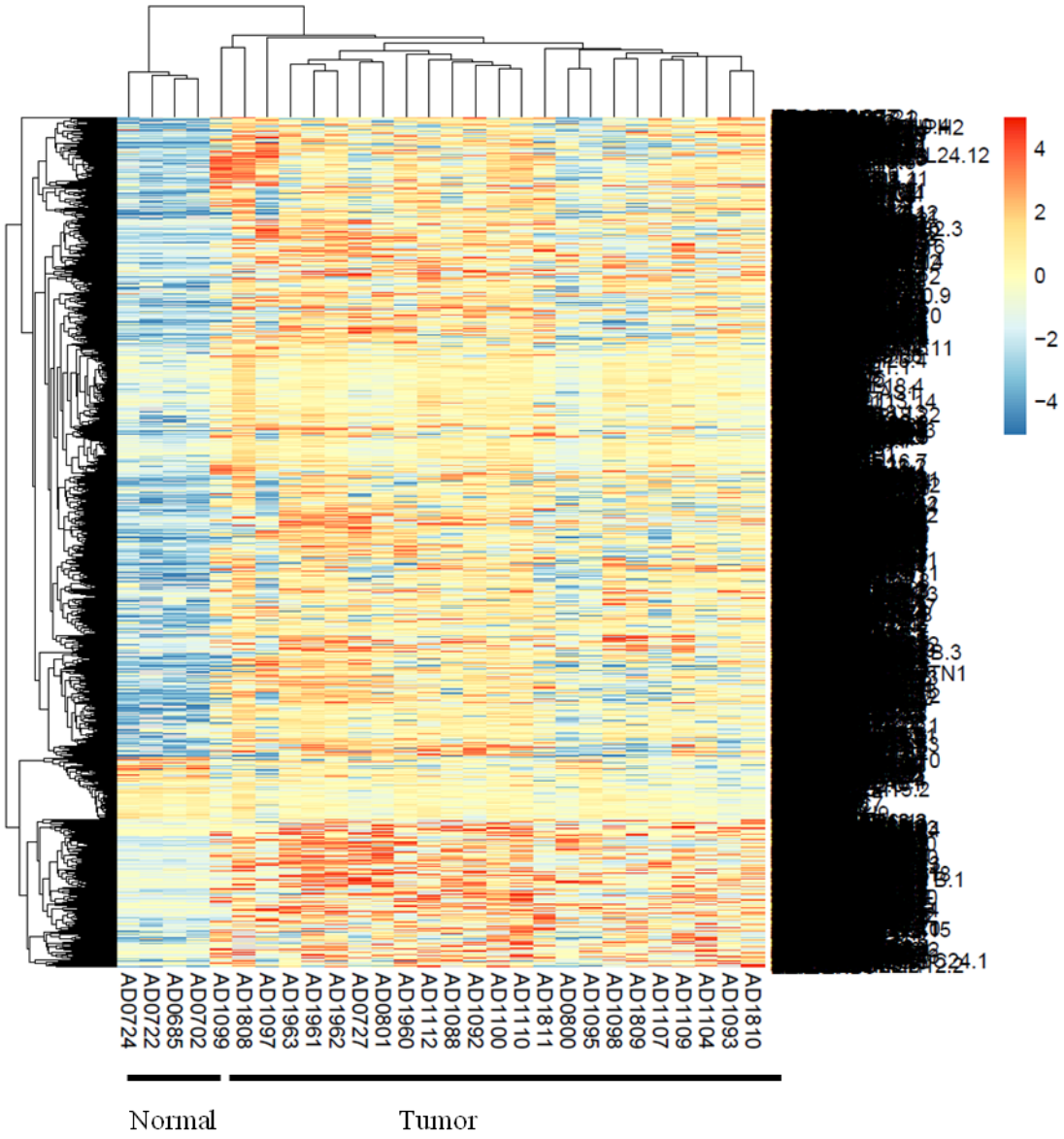
III-Figure 1: A correlation matrix of gene expression in all the RNA-sequenced samples. Matrix informs about the correlation of each sample with other samples with respect to gene expression. The colour scale denotes correlation co-efficient value wherein red indicates high correlation as per gene expression whereas blue denotes low correlation.

Preliminary differential gene (DE) expression analysis done using Tuxedo suite protocol among the normal (n=4) and tumor (n=24) samples of cervical adenocarcinoma identified only 10 significantly ($p<0.05$) differentially expressed genes. This type of analysis was not appropriate as results were not reliable when comparing unequal samples in each group. The DE expressed genes in tumor samples are shown in III-Table 3 below.

gene	locus	Log2 (fold_change)	p_value	q_value
GCNT3	chr15:59903981-59912210	7.49024	0.02146	0.99999
DAPK1	chr9:90112755-90323549	4.03434	0.02672	0.99999
MYO7B	chr2:128293377-128395303	5.63358	0.03496	0.99999
HHLA2	chr3:108021331-108097126	7.37381	0.03804	0.99999
BCL6B	chr17:6926368-6932961	7.09806	0.04232	0.99999
MAP2K6	chr17:67410837-67538470	7.20451	0.0424	0.99999
MEI1	chr22:42095517-42195459	6.67206	0.04273	0.99999
LYPD2	chr8:143831627-143833952	7.71724	0.04692	0.99999
ENPP5	chr6:46127761-46138717	6.67485	0.04948	0.99999
TSPAN8	chr12:71518876-71551779	6.48697	0.05091	0.99999

III-Table 3: RNA-sequencing analysis for differential gene expression using Tuxedo Suite. Significant differentially expressed genes comparing normal with tumor samples along with fold change value is shown.

Next, using Salmon workflow, differential gene expression analysis was performed between 4 normal samples (excluding AD0703) and 24 tumor samples. In total, 3201 genes were observed to be up-regulated and 61 genes were down-regulated. A heatmap representation is shown in III-Figure 2. Although all the normal samples were clustered together suggesting similar gene expression, variable gene expression was observed in tumor samples with no tight gene clusters.



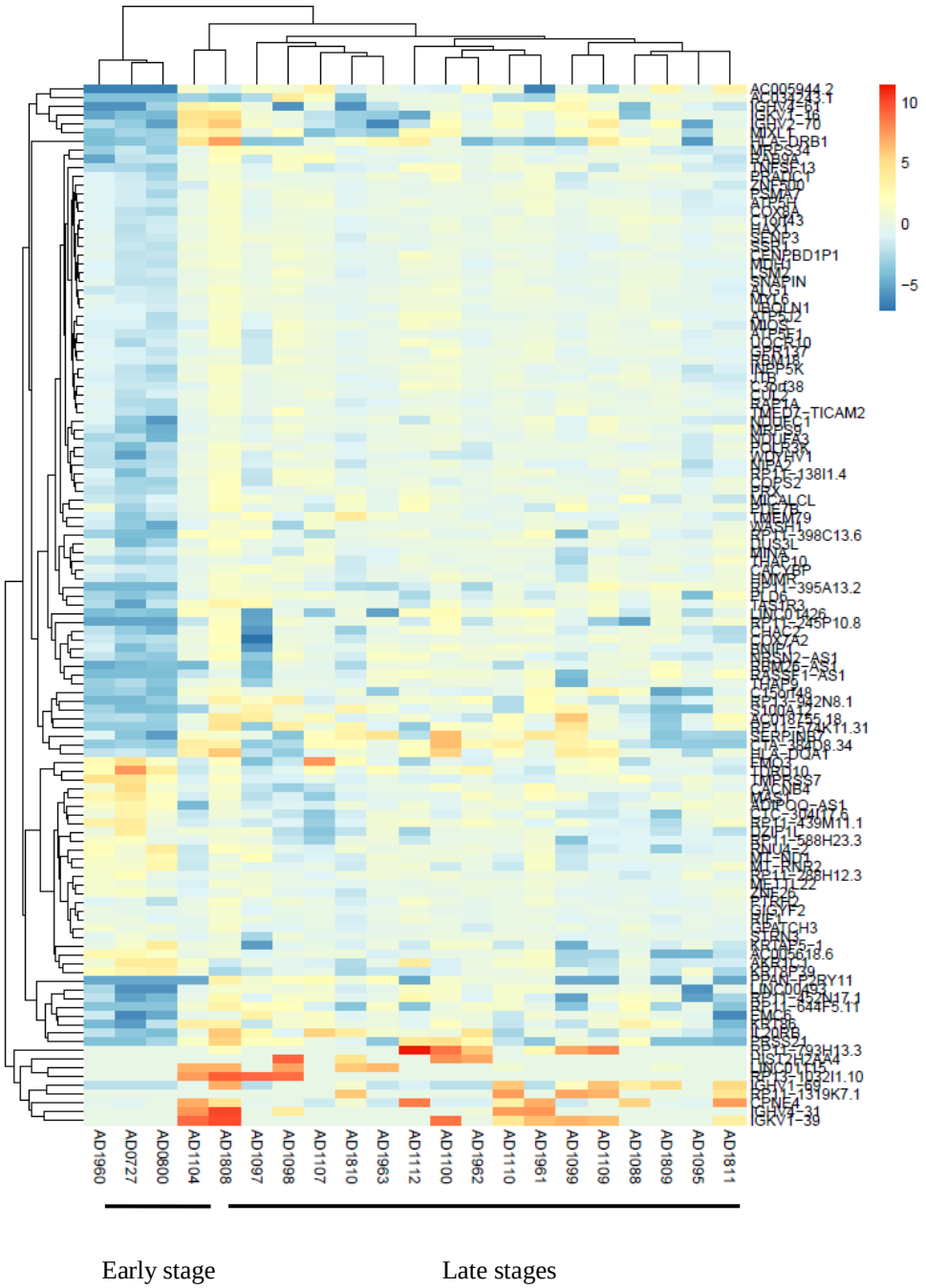
III-Figure 2: Differential gene expression analysis among normal and tumor samples using Salmon workflow. Normal samples clustered in one group together suggesting similar gene expression whereas no distinct gene clusters were observed among tumor samples. Heatmap gradient colour scale indicates a gradient for gene expression values where red indicates high gene expression and blue indicates low expression.

Next, we performed DE analysis between tumors with early and late FIGO stages (IIB and above). Tumor samples belonging to each of the FIGO stages are shown in III-Table 4. 63 genes were up-regulated and 18 genes were down-regulated in late-stage tumors. Of the 4 early-stage tumors, 3 clustered together whereas 1 early-stage tumor clustered with late-stage tumors (n=18). No specific gene clusters were identified in late-stage tumor samples

(III-Figure 3). Further Gene Set Enrichment Analysis (GSEA) was able to identify only 3 genes – *TNFSF12* (growth factor), *COPS2* and *MIXL1* (transcription factors) with no oncogenes or protein kinases. None of the down-regulated genes belonged to any of the gene categories of GSEA. An unequal number of tumor samples belonging to early (n=4) and late stages (n=18) was not able to distinguish and identify differentially expressed genes. Therefore, the analysis between different tumor stages could not give us reliable data.

Figo stage	**Number of samples**
Stage I (early)	3
Stage II (late)	14
Stage III and IV (late)	4

III-Table 4: Number of samples belonging to different FIGO stages of cervical adenocarcinoma samples.



III-Figure 3: Differential gene expression analysis between early and late FIGO stages of cervical adenocarcinoma samples. The first four samples from left belong to early-stage and the remaining samples belong to late stage.

It was observed that out of 24 tumor samples, 3 tumor samples and the 5 normal samples displaying poor RIN value; also showed aberrant expression pattern of the housekeeping genes (data not shown). These low RIN samples may be also incorrectly representing the expression of other genes as well and were probably influencing the analysis. Therefore,

these 3 tumor samples and 5 normal samples were excluded from further analysis. The limited and unequal number of tumor samples in each FIGO stage and smaller representation of normal samples disabled us to perform DE analysis involving tumor stages and normal samples as data reliability is a problem in this case. Next, excluding the bad RIN samples, gene expression analysis was performed among 21 tumor samples with good RIN values. Counts with low expression (<1 TPM) were excluded and log-transformed. We considered genes in the top 10% quartile region and expressed in at least 30% of the samples. Among the tumor samples, genes *EGFR* (57%), *ERBB2* (81%), *ERBB3* (90%), *MET* (38%), *AKT1* (38%) and *AKT2* (90%) were over-expressed in greater than 30% of the samples. Moreover, several of the genes belonging to the MMP family- *MMP2* (47%), *MMP12* (33%) and *MMP14* (100%) also showed over-expression in cervical adenocarcinoma tumor samples. Upregulation of other cancer-associated genes recurrent in at least 30% of the samples are shown in the III-Table 5 below:

Gene	Percent recurrence	Gene	Percent recurrence	Gene	Percent recurrence	Gene	Percent recurrence
DNAJB1	100	*ARHGAP26*	100	*UBR5*	86	*CDKN2A*	52
TPM4	100	*LPP*	100	*MYD88*	86	*PAX8*	52
NDRG1	100	*ATP1A1*	100	*TRAF7*	86	*CTNNB1*	52
MSN	100	*HIST1H3B*	100	*QKI*	81	*MYB*	52
NUMA1	100	*GNAS*	100	*SDC4*	81	*SETD2*	48
HSP90AA1	100	*SMARCE1*	100	*EPAS1*	81	*SMAD3*	48
FOXO3	100	*NONO*	100	*PSIP1*	81	*ZMYM2*	48
NPM1	100	*TMEM127*	95	*PBX1*	76	*MUC1*	48
ABI1	100	*AFF4*	95	*KDM5C*	76	*CCND1*	48
MYH9	100	*STAT6*	95	*HMGA1*	76	*CCDC6*	48
NCOA4	100	*NCOR1*	95	*NUP98*	76	*ELF4*	43
LMNA	100	*PPP2R1A*	95	*SDHB*	76	*TBL1XR1*	43
DICER1	100	*SRSF3*	95	*MECOM*	71	*ATF1*	43
DDX5	100	*EML4*	95	*FOXA1*	71	*MAX*	43
WWTR1	100	*CDH1*	95	*ARNT*	71	*SMAD4*	43
DCTN1	100	*FAT1*	95	*CREB3L2*	71	*MET*	38
PICALM	100	*SS18*	95	*BCL6*	71	*SRC*	38
HERPUD1	100	*YWHAE*	95	*AFDN*	67	*CBLB*	38
RAC1	100	*HIF1A*	90	*SRSF2*	67	*MLLT1*	38
FUBP1	100	*SH3GL1*	90	*TFEB*	67	*CDH11*	38
RPN1	100	*CLTC*	90	*ELK4*	62	*CDK12*	33
NUP214	100	*SET*	90	*EIF4A2*	62	*NF1*	33
LASP1	100	*H3F3A*	90	*AFF1*	62	*MSI2*	33
CALR	100	*FOXP1*	90	*SLC34A2*	62	*CRTC3*	33
EZR	100	*PIK3R1*	90	*DDX3X*	62	*NFE2L2*	33
RHOA	100	*KMT2C*	90	*EXT2*	57	*NCOA1*	33
PRCC	100	*SDHC*	90	*TRIM24*	57	*ETV5*	33
TPM3	100	*ETV6*	86	*JAK1*	52	*SFRP4*	33
EWSR1	100	*CDK4*	86	*STAT3*	52		

III-Table 5: Expression of cancer associated genes that are recurrent in 30% of the cervical adenocarcinoma samples

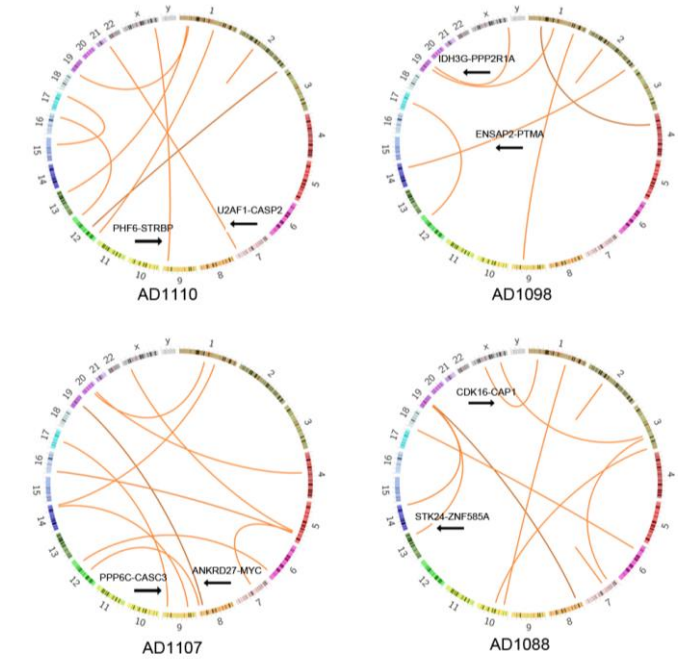
Consistent with our findings, TCGA also reports upregulation of *ERBB2* (12%), *EGFR* (8%) and *ERBB3* (6%) in cervical cancer patient samples. In a study comprising of adenocarcinoma and adenosquamous carcinoma patient samples, over-expression of EGFR, ERBB2 and ERBB3 was observed at 18.8%, 53.5% and 74.7% respectively by IHC [110]. Members of the MMP family found to be over-expressed in our dataset are also found to be upregulated in the TCGA dataset. *MMP2*, *MMP12* and *MMP14* are also upregulated in 5%, 6% and 5% respectively in TCGA patient samples of cervical cancer.

3.4.3 Identification of novel gene fusion transcripts in cervical cancer

We identified fusion transcripts from 24 tumor samples and 4 normal samples of cervical adenocarcinoma using Starfusion. Starfusion filters fusion with paralog gene pairs, fusions with a pseudogene partner and genes with multiple fusion partners within the same sample. Total 157 fusions were identified in tumor samples after depleting fusions present in normal samples; of which 132 were unique fusions (refer appendices 5a). The fusion transcripts were annotated using AGFusion to determine fusion type and also identify domains retained after fusion. 44 intra-chromosomal (28%) and 113 inter-chromosomal (72%) fusions were identified. Out of 132 unique fusions, only 5.4% fusions were identified in two or more samples and 94.7% fusions occurred as singletons. Further, the fusion transcripts were depleted from fusions reported in normal fusion database [167, 168]. Only 1 fusion *ITCH-ASIP* reported in fusion database of normal samples was identified. Next, the tumor-specific fusions from our data were compared with TCGA-Pancancer Fusion database (https://www.tumorfusions.org/) to identify previously reported fusions in our sample. However, there was no overlap with tumor-specific fusions of cervical cancer or any other cancer type.

Out of 157 total fusion transcripts, 5 fusions such as *IDH3G-PPP2R1A*, *U2AF1-CASP2*, *RAP2A-MECOM*, *PPP6C-CASC3* and *ANKRD27-MYC* involved an oncogenic gene partner. Whereas, there were 5 fusions involving tumor suppressor gene as one of the fusion partners are *ENSAP2-PTMA*, *RAB5C-PTMA*, *PHF6-STRBP*, *TCF7L2-AC068898.1* and *ENO1-ZFP36L2*. In addition, in-frame fusions with kinase gene partner such as *PKM-FUT2* and *PKM-CBX4* was observed with conserved PK domain. *STK24-ZNF585A* and *CDK16-CAP1* are other inframe fusions with conserved kinase domain. The information of spanning and junction reads for each fusion is shown in appendix 5b. Few candidate gene fusions were selected for validation. PCR was performed using forward primer binding to gene A and reverse primer binding to gene B. However, none of the fusions got validated by PCR.

From our analysis, none of the fusions identified overlapped with Pancancer fusion database. All the gene-fusion events we report are unique. A circus plot representing gene fusions using Circos Tools (v. 0.66) is shown in III-Figure 4.



III-Figure 4: Circos plot showing inter-chromosomal and intra-chromosomal gene fusion in four samples. Few of the gene fusions are indicated by arrows.

3.4.4 Expression of HPV transcripts and HPV integration from Transcriptome data of cervical adenocarcinoma

Cancer pathogen detector was able to identify different strains of HPV from transcriptome data of 24 cervical adenocarcinoma tumor samples and 5 normal samples. All the 5 normal samples comprising of adjacent cervical tissue were positive for HPV infection; 2 samples having HPV types 16 and 18. Among the 24 tumor samples, six samples were infected with one of the high-risk HPV types whereas all other samples showed the presence of infection with multiple high-risk HPV types HPV16 and HPV18. The output of CPD is shown in III-Table 6. The HPV genome length of different HPV types is mentioned, followed by the read counts supporting HPV genome and PPM (part per million) values. The feature counts represent the read counts for early and late genes of HPV genome that are expressed. We were able to detect the expression of HPV genes even in the adjacent normal tissue samples. HPV genome present in the episomal form is unable to activate expression of E6 and E7 oncoproteins as these genes are transcriptionally repressed by E2 gene. When HPV is integrated into the genome, E2 site is disrupted, thus resulting in increased expression of E6 and E7 oncoproteins [76] which is responsible for the progression of the disease. We assessed the integration of HPV in the human genome of cervical adenocarcinoma samples using the integration mode of our tool HPVDetector (III-Table 6). Although HPV was detected in all 29 samples by Cancer pathogen Detector (CPD), HPV integration was observed in 18 samples (including 3 normal samples) suggesting that in 11 samples, HPV might be present in the episomal form.

			CPD output				HPVDetector- Integration ouput						
NO	**Sample**	**Pathogen**	**Genome Length**	**Read Count**	**PPM**	**FeatureCounts**	**HPV genome**	**Human chr**	**genomic coordinate**	**HPV gene**	**Human gene**	**Cyto-band**	**Integration in known cytoband**
1	AD0685	HPV16	7904	12817	217.277	E1:8;E2:10783;E4:9445;E5:691;E6:668;E7:195;L1:30;	3520	chr2	33141544	E2	*LINC00486*	p22.3	YES
							3520	chr2	33141544	E4	*LINC00486*	p22.3	YES
2	AD0702	HPV16	7904	18894	278.291	E1:527;E2:14152;E4:11758;E5:1041;E6:1694;E7:1044;L1:218;	NO						
3	AD0703	HPV18	7857	107	0.377423	E1:1;E2:14;E4:11;E5:2;E6:7;E7:87;	2787	chr13	73636771	E1	*KLF5*	q22.1	YES
							2818	chr13	73636767	E1	*KLF5*	q22.1	YES
							2818	chr13	73636767	E2	*KLF5*	q22.1	YES
							3514	chr13	73636766	E2	*KLF5*	q22.1	YES
							3514	chr13	73636766	E4	*KLF5*	q22.1	YES
4	AD0722	HPV16	7904	18638	273.053	E1:767;E2:14105;E4:10947;E5:1261;E6:1515;E7:1171;L1:122;							
5	AD0724	HPV16	7904	18424	280.518	E1:374;E2:15475;E4:13432;E5:462;E6:1208;E7:913;L1:49;	3520	chr2	33141545	E2	*LINC00486*	p22.3	YES
							3520	chr2	33141545	E4	*LINC00486*	p22.3	YES
							3521	chr2	33141297	E2	*LINC00486*	p22.3	YES
							3521	chr2	33141297	E4	*LINC00486*	p22.3	YES
							707	chr6	69376712	E7	*BAI3*	q12	YES
6	AD0727	HPV16	7904	2715	38.8268	E1:2;E2:42;E4:38;E5:2;E6:1610;E7:218;L1:106;	112	chr1	234743457	E6	*IRF2BP2*	q42.3	YES
							124	chr1	234743400	E6	*IRF2BP2*	q42.3	YES
							125	chr1	234743473	E6	*IRF2BP2*	q42.3	YES
							678	chr1	234743467	E7	*IRF2BP2*	q42.3	YES
							707	chr1	234743464	E7	*IRF2BP2*	q42.3	YES
7	AD0800	HPV16	7904	10184	146.449	E1:486;E2:7271;E4:6700;E5:291;E6:561;E7:822;L1:308;	NO						
8	AD0801	HPV16	7904	5372	84.5296	E1:50;E2:4423;E4:3949;E5:191;E6:488;E7:256;L1:60;	NO						
9	AD1088	HPV18	7857	34074	473.53	E1:10701;E2:4388;E4:3312;E5:1948;E6:1295;E7:5763;L1:8214;L2:7831;	1420	chr3	42237068	E1	*TRAK1*	p22.1	YES
							1422	chr3	42237069	E1	*TRAK1*	p22.1	YES
							1431	chr3	42237068	E1	*TRAK1*	p22.1	YES
							1432	chr3	42236960	E1	*TRAK1*	p22.1	YES
							1433	chr3	42237069	E1	*TRAK1*	p22.1	YES
							1541	chr3	42237067	E1	*TRAK1*	p22.1	YES
							1574	chr3	42237069	E1	*TRAK1*	p22.1	YES
							1588	chr3	42237056	E1	*TRAK1*	p22.1	YES
							1598	chr3	42237068	E1	*TRAK1*	p22.1	YES
							3992	chr19	49469892	E5	*FTL*	q13.33	NOVEL
							3992	chr2	48605061	E5	*FOXN2*	p16.3	YES
							5374	chr3	42237179	L2	*TRAK1*	p22.1	YES
							5378	chr3	42237195	L2	*TRAK1*	p22.1	YES
							5506	chr2	33141549	L2	*LINC00486*	p22.3	YES
							738	chr3	42256199	E7	*TRAK1*	p22.1	YES
							752	chr3	42236466	E7	*TRAK1*	p22.1	YES
10	AD1092	HPV18	7857	3087	46.7582	E1:236;E2:90;E4:4;E5:4;E6:825;E7:2577;L1:56;L2:20;	739	chr3	50294943	E7	*GNAI2*	p21.31	YES
							772	chr2	33141524	E7	*LINC00486*	p22.3	YES
11	AD1093	NO					NO						
	AD1093	NO					NO						
12	AD1095	HPV16	7904	18483	286.996	E1:446;E2:14715;E4:12438;E5:997;E6:1070;E7:843;L1:270;	3465	chr2	33141424	E2	*LINC00486*	p22.3	YES
							3465	chr2	33141424	E4	*LINC00486*	p22.3	YES
13	AD1097	HPV18	7857	19939	178.487	E1:162;E2:14453;E4:10568;E5:4194;E6:93;E7:854;L1:102;L2:1643;	2787	chr13	73636771	E1	*KLF5*	q22.1	YES
							2796	chr13	73636701	E1	*KLF5*	q22.1	YES
							2818	chr13	73636767	E1	*KLF5*	q22.1	YES
							2818	chr13	73636767	E2	*KLF5*	q22.1	YES
							2819	chr13	73636771	E1	*KLF5*	q22.1	YES
							2819	chr13	73636771	E2	*KLF5*	q22.1	YES
							2900	chr13	73636771	E2	*KLF5*	q22.1	YES
							3177	chr13	73636764	E2	*KLF5*	q22.1	YES
							3484	chr13	73636106	E2	*KLF5*	q22.1	YES
							3484	chr13	73636106	E4	*KLF5*	q22.1	YES
							3484	chr13	73636528	E2	*KLF5*	q22.1	YES
							3484	chr13	73636528	E4	*KLF5*	q22.1	YES
							3484	chr13	73636615	E2	*KLF5*	q22.1	YES
							3484	chr13	73636615	E4	*KLF5*	q22.1	YES
							3486	chr13	73636547	E2	*KLF5*	q22.1	YES
							3486	chr13	73636547	E4	*KLF5*	q22.1	YES
							3514	chr13	73636766	E2	*KLF5*	q22.1	YES
							3514	chr13	73636766	E4	*KLF5*	q22.1	YES
							3725	chr2	33141425	E2	*LINC00486*	p22.3	YES
							3900	chr13	73636771	E2	*KLF5*	q22.1	YES
							649	chr13	73636155	E7	*KLF5*	q22.1	YES
							738	chr13	73636266	E7	*KLF5*	q22.1	YES

			CPD output				HPVDetector- Integration ouput						
NO	**Sample**	**Pathogen**	**Genome Length**	**Read Count**	**PPM**	**FeatureCounts**	**HPV genome**	**Human chr**	**genomic coordinate**	**HPV gene**	**Human gene**	**Cyto-band**	**Integration in known cytoband**
14	AD1098	NO					NO						
	AD1098	NO					NO						
15	AD1099	HPV16	7904	420	6.44848	E1:2;E2:66;E4:52;E5:4;E6:312;E7:93;	707	chr1	234743464	E7	*IRF2BP2*	q42.3	YES
16	AD1100	HPV18	7857	73293	685.71	E1:17572;E2:48;E5:5;E6:4101;E7:67729;L1:491;L2:233;	508	chr2	33091930	E6	*LINC00486*	p22.3	YES
							635	chrX	85492376	E7	*DACH2*	q21.2	NOVEL
							649	chr14	31598132	E7	*HECTD1*	q12	NOVEL
							749	chr2	33141543	E7	*LINC00486*	p22.3	YES
17	AD1104	HPV16	7904	68	1.07259	E2:56;E4:51;E5:4;E6:4;E7:1;L1:4;	124	chr1	234743400	E6	*IRF2BP2*	q42.3	YES
	AD1104	HPV18	7857	2317	36.547	E1:137;E2:6;E4:4;E5:186;E6:511;E7:1933;L1:26;L2:57;	NO						
18	AD1107	HPV16	7904	20268	275.95	E1:462;E2:15611;E4:14438;E5:715;E6:2555;E7:2481;L1:91;	125	chr10	32322769	E6	*KIF5B*	p11.22	YES
19	AD1109	HPV16	7904	37518	587.824	E1:535;E2:29570;E4:27673;E5:1407;E6:3499;E7:2824;L1:540;	3521	chr2	33141429	E2	*LINC00486*	p22.3	YES
							3521	chr2	33141429	E4	*LINC00486*	p22.3	YES
							3523	chr8	95444791	E2	*FSBP*	q22.1	YES
							3523	chr8	95444791	E2	*RAD54B*	q22.1	YES
							3523	chr8	95444791	E4	*FSBP*	q22.1	YES
							3523	chr8	95444791	E4	*RAD54B*	q22.1	YES
							3665	chr2	33141522	E2	*LINC00486*	p22.3	YES
							3936	chr6	73713535	E5	*KCNQ5*	q13	YES
							3936	chr7	87470905	E5	*SLC25A40*	q21.12	YES
20	AD1110	HPV18	7857	3953	53.4836	E1:494;E2:496;E4:13;E5:2;E6:895;E7:2913;L1:12;L2:12;	NO						
21	AD1112	NO											
22	AD1808	HPV18	7857	17818	107.604	E1:380;E2:8017;E4:7997;E6:2488;E7:8690;	3498	chr17	70595596	E2	*LINC00511*	q24.3	YES
							3498	chr17	70595596	E4	*LINC00511*	q24.3	YES
							3597	chr17	70597419	E2	*LINC00511*	q24.3	YES
							3597	chr17	70597419	E4	*LINC00511*	q24.3	YES
							648	chr7	151945301	E7	*MLL3*	q36.1	YES
23	AD1809	NO					NO						
	AD1809	NO					NO						
24	AD1810	HPV18	7857	3895	63.0619	E1:121;E2:1690;E4:1590;E6:401;E7:1921;L1:50;L2:80;	NO						
25	AD1811	HPV18	7857	629	9.9396	E1:55;E2:10;E4:6;E5:5;E6:94;E7:541;L1:6;L2:8;	NO						
26	AD1960	HPV16	7904	8544	93.3716	E1:371;E2:6403;E4:5588;E5:136;E6:835;E7:654;L1:200;	NO						
	AD1960	HPV18	7857	11781	128.747	E1:283;E2:2276;E4:1842;E5:400;E6:414;E7:1573;L1:3787;L2:5937;	508	chr3	65766509	E6	*MAGI1*	p14.1	YES
27	AD1961	HPV16	7904	43550	438.21	E1:1347;E2:25704;E4:23244;E5:1177;E6:5375;E7:3472;L1:907;	3268	chr11	28260379	E2	*METTL15*	p14.1	YES
							3521	chr2	33141545	E2	*LINC00486*	p22.3	YES
							3521	chr2	33141545	E4	*LINC00486*	p22.3	YES
							4549	chr2	97924006	L2	*ANKRD36*	q11.2	YES
28	AD1962	HPV16	7904	32384	422.85	E1:893;E2:21239;E4:20148;E5:623;E6:2975;E7:2655;L1:456;	3411	chrX	100615235	E2	*BTK*	q22.1	YES
							3411	chrX	100615235	E4	*BTK*	q22.1	YES
29	AD1963	HPV16	7904	3085	39.1947	E1:36;E2:2572;E4:2302;E5:51;E6:162;E7:36;L1:85;	NO						
	AD1963	HPV18	7857	4660	59.2049	E1:555;E2:2;E4:1;E6:544;E7:4328;L1:18;L2:14;	7129	chr3	185161247	L1	*MAP3K13*	q27.2	YES

III-Table 6: Detection of HPV infection and integration in the human genome in transcriptome sequenced cervical adenocarcinoma samples

11 samples show HPV integration in the known cytoband region- 2p22.3 and 13q22.1 as reported in the literature [173]. 9 samples harbouring HPV16 and HPV18 integration in 2p22.3 cytoband show integration at the intronic region of *LINC000468* gene. Integration in this long intergenic non-protein coding RNA (LINC000468) has been reported by our previous study [125]. 2 samples show HPV18 integration in the exonic region of *KLF5* gene belonging to 13q22.1 cytoband. Integration of HPV genome in the Kruppel like factor 5 (*KLF5*) gene has been reported in cervical cancer previously [96, 174, 175]. In addition, 3

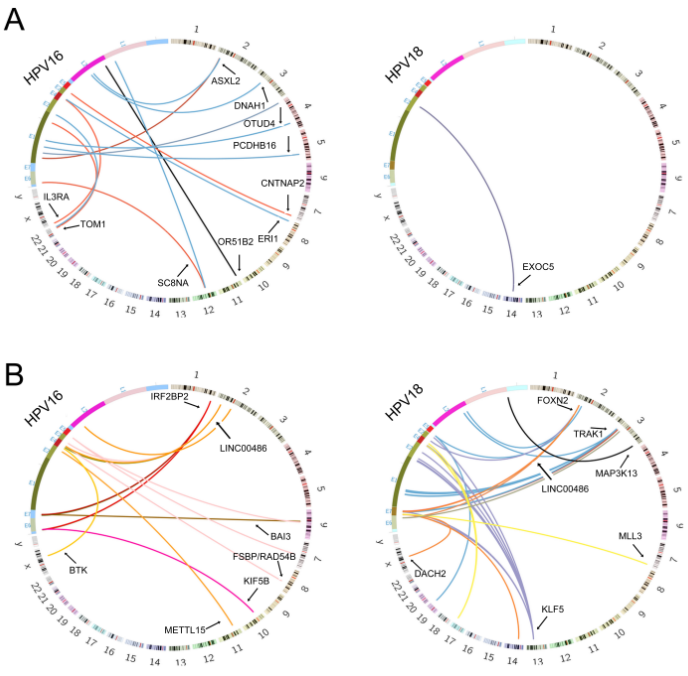
samples are showing HPV16 integration in the exonic region of *IRF2BP2* in 1q42.3 cytoband. Integration of HPV16 strain and *IRF2BP2* has been previously reported in one of the studies [96]. Integration of HPV in other genes such as *BAI3*, *FOXN2*, *DACH2*, *RAD54B*, *KCN5Q* and *MAP3K13* has also been reported by Hu *et. al*.,[96]. Refer to III-Table 7

No.	Cytoband	Recurrence (N=18)	Gene	Integration site
1	2p22.3	9	*LINC00486*	Intron
2	1q42.3	3	*IRF2BP2*	Exon
3	13q22.1	2	*KLF5*	Exon
4	8q22.1	1	*FSBP, RAD54B*	*FSBP*-Exon*; RAD54B*-Intron
5	6q12	1	*BAI3*	Intron
6	3p22.1	1	*TRAK1*	Intron
7	19q13.33	1	*FTL*	Exon
8	2p16.3	1	*FOXN2*	Exon
9	3p21.31	1	*GNAI2*	Exon
10	Xq21.2	1	*DACH2*	Intron
11	14q12	1	*HECTD1*	Exon
12	10p11.22	1	*KIF5B*	Intron
13	6q13	1	*KCNQ5*	Intron
14	7q21.12	1	*SLC25A40*	Intron
15	17q24.3	1	*LINC00511*	N/A
16	7q36.1	1	*MLL3*	Exon
17	3p14.1	1	*MAGI1*	Intron
18	11p14.1	1	*METTL15*	Intron
19	2q11.2	1	*ANKRD36*	Intron
20	Xq22.1	1	*BTK*	Intron
21	3q27.2	1	*MAP3K13*	Intron

III-Table 7: HPV integration sites in the intronic and exonic region of genes for RNA-sequenced samples

Three samples are displaying novel HPV integrations in the cytoband 19q13.33, Xq21.2 and 14q12. HPV integrations in the remaining samples are also reported in cervical cancer from other studies [176, 177]. III-Table 7 shows a list of genes showing HPV integration. There are 21 genes in which the HPV genome integration is detected. From our data, most of the integrations (n=13) are in the intronic region of genes whereas 8 genes show integration in the exonic region.

Circos plot representation of HPV integration from both exome and RNA-sequenced samples is shown in III-Figure 5.



III-Figure 5: Circos plots showing HPV integration in the human genome. A- HPV16 and HPV18 integration in exome sequenced samples, B- HPV16 and HPV18 integration in Transcriptome sequenced samples. Different colours correspond to HPV integration in different samples. Chromosome integration is marked by an arrow, followed by gene name.

A detailed heatmap representation of HPV infection from the NGS data is shown in II-Table 16.

3.5 Discussion

Cervical cancer treatment largely consists of radiation and chemotherapy. Targeted therapy is not commonly employed in case of cervical cancer as only bevacizumab has been approved for treatment to date. According to Cancer India statistics, the five-year survival average is 48.7%. There is an unmet need to utilize the known inhibitors for targeting different driver genes or identify new prognostic markers to improve the overall survival of patients. For the same, patients need to be stratified into groups based on the unique genetic makeup of the tumors to design the treatment strategies best suited for the individual.

Here, we have performed Whole Transcriptome Sequencing (WTS) on 24 tumor samples of cervical adenocarcinoma along with 5 normal cervix tissue samples to identify gene expression patterns and gene fusion events.

Differential gene expression analysis among the tumor samples was performed and we identified genes *EGFR, ERBB2, ERBB3, AKT1, AKT2, MMP2, MMP14* and *MMP12* to be overexpressed in at least 30% of the tumor samples. Expression of matrix metalloproteinases is common to some of the tumor types where MMPs are used as prognostic markers to predict disease outcome. We observed expression of several MMP genes *MMP12 (33%), MMP2* (47%) and *MMP14* (100%) recurrent in at least 30% of the samples. Expression of *MMP2* has been reported in cervical cancer wherein 42% of samples expressed *MMP2*. Over-expression of *MMP2* was associated with unfavourable survival [178]. MMP2 is involved in promoting invasion and migration of cervical cells [179]. *MMP12* expression in cervical dysplasia has been reported in the literature and has a potential role to play in the early stages of cervical transformation and invasion [180]. MMP14 promotes invasion of cervical cells. Downregulation of *MMP14* in HeLa cells resulted in inhibition of malignant phenotype [181]. Over-expression of *ERBB3* is associated with poor survival rate in several cancer types including cervical cancer [182]. In cervical adenocarcinoma, samples positive for *ERBB2, EGFR* and c-*MET* co-expression were significantly associated with lymph node metastasis and shorter relapse-free survival [183]. Our study along with the literature suggests that patients can be stratified based on the expression of MMP genes since several inhibitors are available and currently in pre-clinical trials [184, 185] serving as potential targets for treatment.

Recurrent gene fusions events are common in some of the tumor types and few fusions genes display oncogenic potential and act as drivers and hence, are attractive therapeutic

targets. Targeting the oncogenic fusions like *BCR-ABL*, *FGFR3-TACC3* and *EML4-ALK* have shown good clinical efficacy in several tumor types [186-188].

Gene fusions in cervical cancer have been reported by TCGA study and others [94, 187, 189]. Here, we identified novel gene fusions from our study. We report 5 fusions having one of the gene partners as oncogene-*IDH3G-PPP2R1A*, *U2AF1-CASP2*, *RAP2A-MECOM*, *PPP6C-CASC3* and *ANKRD27-MYC* and 5 fusions having one of the partners with tumor suppressor role- *ENSAP2-PTMA, RAB5C-PTMA, PHF6-STRBP, TCF7L2-AC068898.1* and *ENO1-ZFP36L2*. There are 4 inframe fusion genes with kinase partners wherein the kinase domain is conserved- *PKM-FUT2*, *PKM-CBX4, STK24-ZNF585A* and *CDK16-CAP1*. These fusion genes may contribute to the oncogenic effect in cervical cancer. We do not find overlap of fusion genes with the fusions reported in cervical cancer and other cancer types by TCGA.

HPV is the major etiological factor responsible for carcinogenesis of cervical cancer. HPV oncogenes E6 and E7 suppress the activity of p53 and Rb respectively. HPV genome in the episomal form does not express E6 and E7 oncogenes during to repressive activity of E2 gene. Upon integration of HPV in the human genome, mostly E2 gene is disrupted leading to expression of the HPV oncoproteins.

Therefore, it is interesting to study the integration pattern of HPV in the human genome to predict the role of HPV in carcinogenesis. HPV integration was identified using tool HPVDetector. We observe that 11 samples show recurrent integration in known cytoband region 2p22.3 and 13q22.1 with integration in genes *LINC000468* and *KLF5* gene respectively whereas other HPV integrations are common to integration observed in cytoband region as reported by Hu. *et al.*, [96]. Overall, 8 samples show HPV16 integration

and 4 samples are likely to have HPV16 in episomal form whereas 7 samples have HPV18 integration and 4 might be HPV18 in episomes. 4 samples were HPV negative.

In conclusion, we report differentially expressed genes among tumor samples that could help stratify patients into groups. Patients stratified based on unique gene expression pattern can be treated differently to improve treatment response. In addition, we identified novel gene fusion and HPV integration information from RNA-sequencing data which could play a role in promoting carcinogenesis. However, functional validation of gene fusion events is essential to confirm the role in promoting tumorigenesis of cervical cancer.

Chapter IV

Describing structural alterations in cervical cancer by performing whole genome and whole exome sequencing

Chapter IV: Describing structural alterations in cervical cancer by performing whole genome and whole exome sequencing

4.1 Abstract

Background: Cancer is driven by multiple alterations such as mutation, copy number changes, gene expression, epigenetic changes and structural variations. Exome sequencing performed on cancer samples identify disease-causing variants in coding region confidently; however, structural variations are better captured by whole genome sequencing. In this study, copy number alterations were predicted from WGS and WES data and other structural variations (SV) from WGS data to discover actionable gene targets.

Material and Methods: Whole-genome sequencing (WGS) has been performed on 3 paired samples and whole exome sequencing (WES) on 17 paired and 1 orphan tumor sample of cervical adenocarcinoma. Two samples overlap between WES and WGS. WES data available for 9 paired squamous samples was utilized for this study. Copy number alterations were predicted using control-FREEC in WES and WGS paired samples whereas other structural variations in WGS were detected using BreakDancer tool.

Result: WES and WGS copy number analysis of cervical adenocarcinoma (n=18) show recurrent copy gain in genes *PIK3CA SOX2 TERT ERBB2, ERBB3, KRAS, MYC* and *BRCA1*, consistent with the literature. Moreover, copy gain was observed in several of the cancer-associated tyrosine kinase genes such as *AKT1, AKT 2*, *ERBB4*, *EGFR, FGFR* and, *FGFR3*. *LRP1B* deletion, a commonly deleted gene in cervical cancer was observed in 1 sample. From WGS data, broad and focal level alterations were detected. On the chromosomal level, 14 broad-arm level amplification and 5 broad arm deletions, 221 focal amplification and 31 focal deletions were predicted. Recurrent amplification at chromosome 1q, 3q, 8q, 11p, 17q, 19q, 20q, 5p, 9q, 1p, 11q, 20p and 9p and deletion at 3p, 4q, 11p, 11q,

18q, 19p, 2q and 5q chromosomal regions were obtained, consistent with TCGA and other literature reports. In squamous carcinoma samples (n=9), recurrent amplification was observed in *PIK3CA, AKT1* and *ERBB2* genes.

In addition, we report all unique structural variations from our data. *DUX4-ROCK1P1, MLLT4-KIF25, MSH2-TAF4, PAX7-EEF1A2, PBX1-SIK3* and *PDGFRA-MAN2A1* SV pairs harbour one of the genes reported in different cancer types. The functional implication of these SV in cervical cancer is unknown.

Conclusion: Here, we report copy number alterations and structural variations identified in cervical adenocarcinoma and squamous samples. We capture known and novel copy number alterations in cervical cancer and all unique structural variations from our data-set.

4.2 Introduction

Cancer is driven by multiple genomic alterations such as driver mutations, copy number alterations and structural variations. Whole-genome sequencing (WGS) is an approach that can capture all these types of genomic alterations and provide a complete genomic landscape contributing to disease progression [32]. In addition, WGS can identify genetic aberrations in the non-coding region which include point mutations in promoter and enhancer regions that bring about change at the epigenetic level, affect transcriptional and post-transcriptional gene regulation [190, 191] which is often missed by whole-exome sequencing (WES) or targeted sequencing. Although WES is cost-effective as compared to WGS when querying the coding region of the genome, a rapid decrease in the sequencing cost of WGS will enable to extract information at both coding and non-coding regions thereby providing an overall landscape of alterations relevant from a clinical point of view.

Several cancer types are sequenced at large numbers by TCGA/ ICGC on whole-genome level, however, the representation of cervical cancer especially adenocarcinoma subtype is very poor [120]. In studies characterizing genomic landscape of cervical cancer published by the Ojesina *et al.,* group, whole-genome sequencing has been performed on 14 paired samples, of which 4 samples belonged to adenocarcinoma subtype [95]. Copy number analysis of exome and genome sequenced samples of adenocarcinoma subtype (n=24) revealed 4 broad level gains and 8 broad level losses; 8 focal amplifications overlapping with 3 broad arm gains with no significant focal deletions. Both subtypes showed focal amplification at 17q12 harbouring *MYC* and *ERBB2* genes. Other genes with copy gain in adenocarcinoma include *MCL1*, *PIK3CA* and *SOX2*. In squamous carcinoma (n=79), 9 broad level amplification and 11 broad level deletions were detected along with 16 focal gains and 25 focal deletions with no overlap. The significantly relevant focal amplification

was observed at 11q12 harbouring genes *BIRC3* and *YAP1*. Several genes listed in the cancer gene census show gain and loss in squamous carcinoma subtype. The general observation reveals that copy number alteration events are low in adenocarcinoma subtype. This observation can be partly due to the reason that the representation of adenocarcinoma samples was very less as compared to squamous carcinoma. Another study by TCGA on 144 squamous carcinomas and 31 adenocarcinoma cervical samples reveal 26 focal gain and 37 focal loss with 23 whole arm recurrent alterations. Recurrent copy gains are observed in genes *EGFR, CD274, PDCD1LG2, KLF5, BCAR4, TERC, MECOM, TP63 MYC, PVT1, YAP1, BIRC2/3* and *ERBB2* and recurrent deletions in *TGFBR2, SMAD4, FAT1* and *PTEN* [94]. In addition, the authors were able to cluster samples as high copy number comprising mostly of squamous carcinoma subtype and low copy number cluster dominated by adenocarcinoma samples.

WGS is a robust NGS approach to detect structural variations (SV). SV refer to large genomic rearrangements which are often accompanied by DNA copy number alterations [192]. Several cancer types such as high grade serous ovarian, neuroblastoma, small cell lung, triple-negative breast and oesophageal cancers are known to be driven by SV [193]. Pan-cancer analyses of 3299 tumors samples reveal SV driven tumors tend to have few point mutations as drivers [194]. WES of cancer samples fails to catch clinically relevant SV in samples harbouring low single nucleotide variations (SNV) drivers. Therefore, structural variant profiling with WGS analysis is essential to extend clinically possible therapeutic options.

Cervical cancer genomes and transcriptomes sequenced by TCGA were able to identify 22 structural re-arrangements with recurrent *ZC3H7A-BCAR4*. *BCAR4* is known to induce cellular proliferation in breast cancer via ERBB2/ERBB3 pathway activation and treatment

with Lapatinib could prove beneficial as in breast tumors [94]. Ojesina *et al.,* group did not find any recurrent SV events. However, several SVs were recurrent with genes *NTRK2, ARHGEF* and *NIPBL* [95]. Moreover, not only gene re-arrangements but a reshuffling of the genome can also be brought about by extranuclear DNA sequences from bacteria, viruses and mitochondria. Most of the cervical cancers show the integration of HPV viral genome leading to structural variations at the insertion site which causes the adjacent regions to be amplified [192, 195].

With this background and aim to identify genome-wide copy number alterations and novel structural variations, we performed WGS on 3 paired samples and WES of 18 samples of cervical adenocarcinoma and re-analyzed WES data of 9 paired samples of squamous carcinoma from the Indian population to gain an in-depth understanding of the genome-wide somatic alterations playing a role in cancer development and with a hope to identify actionable targets.

4.3 Material and methods

4.3.1 Sample information

About 3 ug of genomic DNA of three tumor samples and matched normal samples were submitted for Whole-genome sequencing (WGS) to BGI, China. To assess DNA integrity, 100 ng of genomic DNA was resolved on 0.8% agarose gel. Whole exome sequencing (WES) was performed on 17 paired and 1 orphan tumor of cervical adenocarcinoma sample. WES and WGS have an overlap of two samples. In addition, previously published WES data for cervical squamous carcinoma was used for this study. The sample information is already described in Chapter II.

4.3.2 Copy number analysis using Control-FREEC

Copy number analysis was performed using Control-FREEC (v11.5)[196] on three paired WGS samples, 17 paired WES samples of cervical adenocarcinoma and 9 paired samples of squamous carcinoma. Control-FRECC computes read counts in each region for both paired samples, normalizes as per mappability and GC content, and segments the copy number data according to LASSO-based algorithm. For WGS data analysis, the window size was set to 50000. In brief, Fastq reads were aligned to hg19 reference genome using bwa (v0.7.16). The Sam Files were converted to bam files using Samtools (v1.6) and then sorted. Next, mate-pair information was verified and duplicate removal was done, followed by conversion to mpile-up. Further, mpileup files were provided to FREEC to infer copy number alterations. A list of significant copy number altered genes is obtained and WilcoxonP test ranks significant copy change alterations and false positives are eliminated by Kolmogorov-Smirnov test. For WES, copy number analysis using Control-FREEC was performed as described previously [163]. Genes that show frequent copy number alterations as reported in the literature were chosen for further validation.

4.3.3 **Validation of copy number changes in adenocarcinoma samples by real-time PCR**

Copy number primers were designed for candidate genes *EGFR, ERBB2, PIK3CA, MYC, TERT* and *CCND1* that are frequently reported to show copy gains as per literature reports. 10 ng of genomic DNA was used per 6 ul reaction volume in triplicates and real-time PCR was done on Light cycler 480 (Roche, Mannheim, Germany). Relative copy number change was inferred in tumor sample with respect to its matched normal sample and the fold change was calculated. Samples with a fold change of ≥ 2.5 were considered as amplified and ≤ 1.5 as deleted. The range of 1.99 to 1.4 was considered as diploid. Primer information is shown in IV-Table 1.

Primer	Sequence	Amplicon Size
OAD1069_*CCND1*_F	GAACTACCTGGACCGCTTCC	89 bp
OAD1070_*CCND1*_R	TAGAGGCCACGAACATGCAA	
OAD1324_*TERT*_F	CTACGGGGTGCTCCTCAAGA	120 bp
OAD1325_*TERT*_R	TCTGTGTCCTCCTCCTCGG	
OAD1071_*MYC*_F	AGAGTTTCATCTGCGACCCG	76 bp
OAD1072_*MYC*_R	AAGCCGCTCCACATACAGTC	
OAD506_*GAPDH_F*	GAGGCTCCCACCTTTCTCATC	96 bp
OAD507_*GAPDH*_R	ATTATGGGAAAGCCAGTCCCC	
OAD1133_SOX2_F	TACAGCATGTCCTACTCGCAG	110 bp
OAD1134_SOX2_R	GAGGAAGAGGTAACCACAGGG	
OAD963_*PIK3CA*_F	CAATGAATTAAGGGAAAATGA	177 bp
OAD1227_*PIK3CA*_R	AGATCAGCCAAATTCAGTTA	
OAD1099_*ERBB2*_F	GAGGCTGTGTGGTGTTTGG	136 bp
OAD1100_*ERBB2*_R	CGTGGATGTCAGGCAGATG	
OAD1208_*EGFR*_F	TGCTGTGACCCACTCTGTCT	169 bp
OAD1209_*EGFR*_R	AACCTCCTACCCCTCCAGAA	

IV-Table 1: Primer information for copy number validation

4.3.4 Identification of structural alterations using Breakdancer

For calling structural variants from whole-genome sequencing data, Breakdancer was used with default settings. This package predicts five types of structural variations such as inversion (INV), insertion (INS), deletion (DEL), intra-chromosomal (ITX) and inter-chromosomal translocations (CTX) from paired-end sequencing reads which are mapped based on the separation distance and alignment orientation. Once the SV genomic co-ordinates were identified for each gene in all the samples, gene-based annotation was performed for both breakpoints of SV. Next, the SV common to both tumor and normal samples were filtered. SV represented by at least 5 supporting reads were retained.

4.4 Results

4.4.1 Copy number alterations identified from WGS data of cervical adenocarcinoma

Copy number analysis was performed using control-FREEC [196] on 3 paired cervical adenocarcinoma samples. Control-FREEC performs depletion of normal Copy Number Alterations (CNA) from corresponding tumor samples. A list of significant copy number alterations (CNA) was obtained for each sample. Overall, in three samples, 271 regions showed amplification with copy gain of 3 and above and 35 regions showed copy loss. Further, each of the regions was annotated to identify genes lying within the region. A heatmap was generated for all genes showing copy amplification and deletion. A total of 9092 genes were obtained with copy gain in at least 1 sample (appendix 4). Genes *PAK2, SPPL2B, TMPRSS9 and TRPM2* were amplified in all three samples. Amplification was observed in several genes such as *ERBB2, KLF5, SOX1, LPP* and *SPACA7, also* observed to be recurrent in TCGA and other meta-analysis studies [94, 197]. Amplification in other cancer-associated genes includes kinases genes *STK11, ERBB3, FGFR4, FGFR1* and *FGFR3*, epigenetic genes like *BRD4, KMT2C, KMT2D, EP300* and *BRD3*, AKT pathway genes *AKT1* and *AKT2* and WNT pathway genes such as *WNT1, WNT10B, WNT2B, WNT9B, WNT3* and *WNT7B*. Amplification of other oncogenes includes *NRAS, HRAS, MMP9, AURKA* and *AURKB*. Interestingly, amplification of genes belonging to oncogenic fusion partners such as *FGFR3-TACC3, TMPRSS2-ERG* and *EML4-ALK* was observed in the same sample. About 1314 genes were deleted in three samples with no overlap. Copy number alterations in driver genes are shown in IV-Table 2. Deletion of *LRP1B*, observed in data is also known to show recurrent deleted in cervical tumors. Deletion is also detected in tumor suppressor genes includes *ATM* and *NF2*. Integration of copy number changes from exome and genome sequencing is shown in appendix 3.

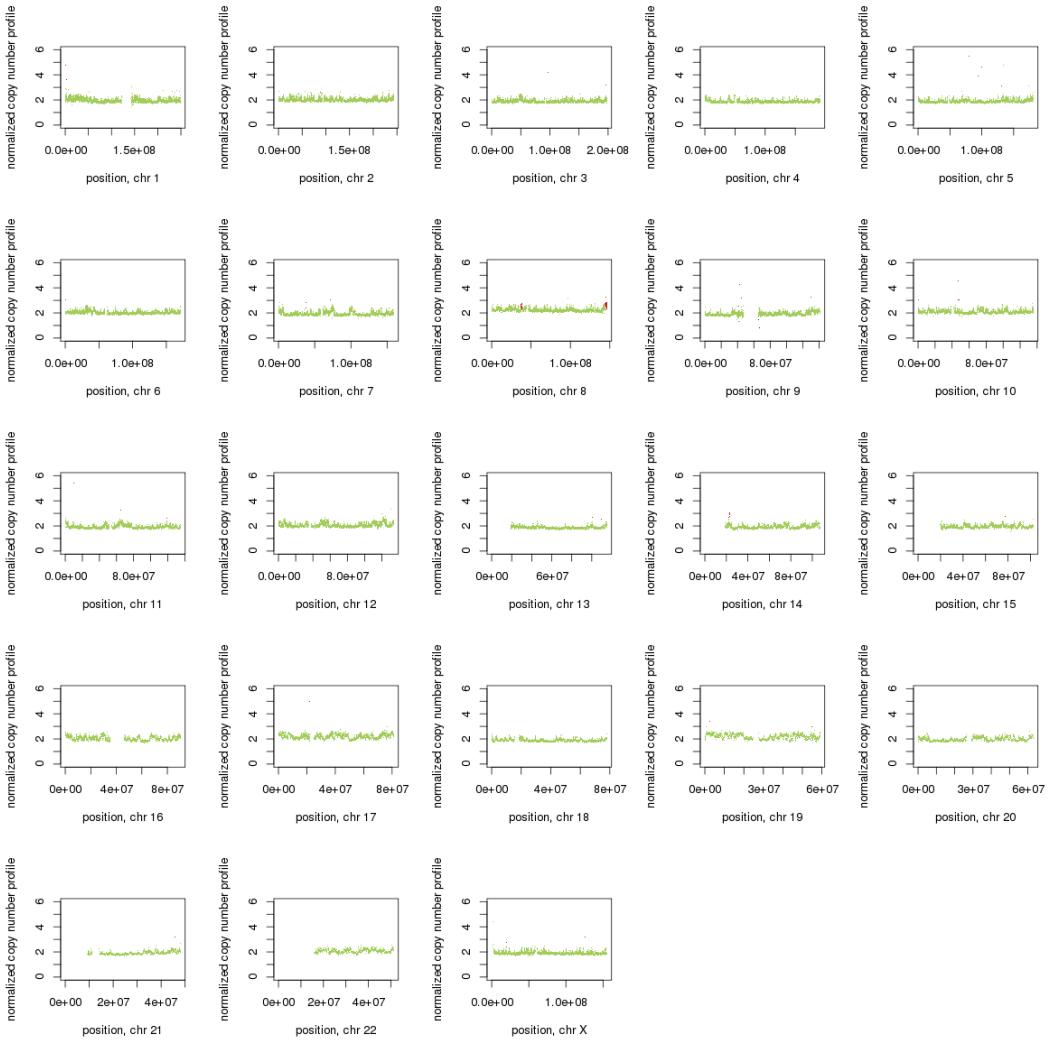
Gene	Cytoband region	Gain/ Loss	AD0708	AD0718	AD1105	Function
AKT1	14q32.33	Gain				Oncogene
AKT2	19q13.11-q13.2	Gain				Oncogene
AURKA	20q13.2-q13.31	Gain				Oncogene; Over-expression in cervical cancer
AURKB	17p13.3-p13.1	Gain				
CCNE1	19q12-q13.43	Gain				Cyclin E1 amplification has oncogenic role in cancer
ERBB2	17q12	Gain				Oncogene; Amplification reported in cervical cancer
ERBB3	12q13.2-q14.1	Gain				Oncogene
FGFR1	8p11.23-p11.22	Gain				Oncogene
FGFR4	5q35.1-q35.3	Gain				Oncogene
MMP9	20p12.1-q13.2	Gain				Over-expression reported in cervical cancer
FGFR3	4p16.3	Gain				FGFR3-TACC3 are oncogenic gene partners
TACC3	4p16.3	Gain				
EML4	2p21-p16.3	Gain				EML4-ALK are oncogenic gene partners
ALK	2p24.1-p23.1	Gain				
TMPRSS2	21q22.11-q22.3	Gain				TMPRSS2-ERG are oncogenic gene partners
ERG	21q22.11-q22.3	Gain				
ATM	11q14.1-q23.1	Loss				Tumor suppressor gene
LRP1B	2q21.2-q31.1	Loss				Tumor suppressor gene; deletion reported in cervical cancer
NF2	22q11.21-q12.3	Loss				Tumor suppressor gene

IV Table 2: Potential driver genes with copy number alterations

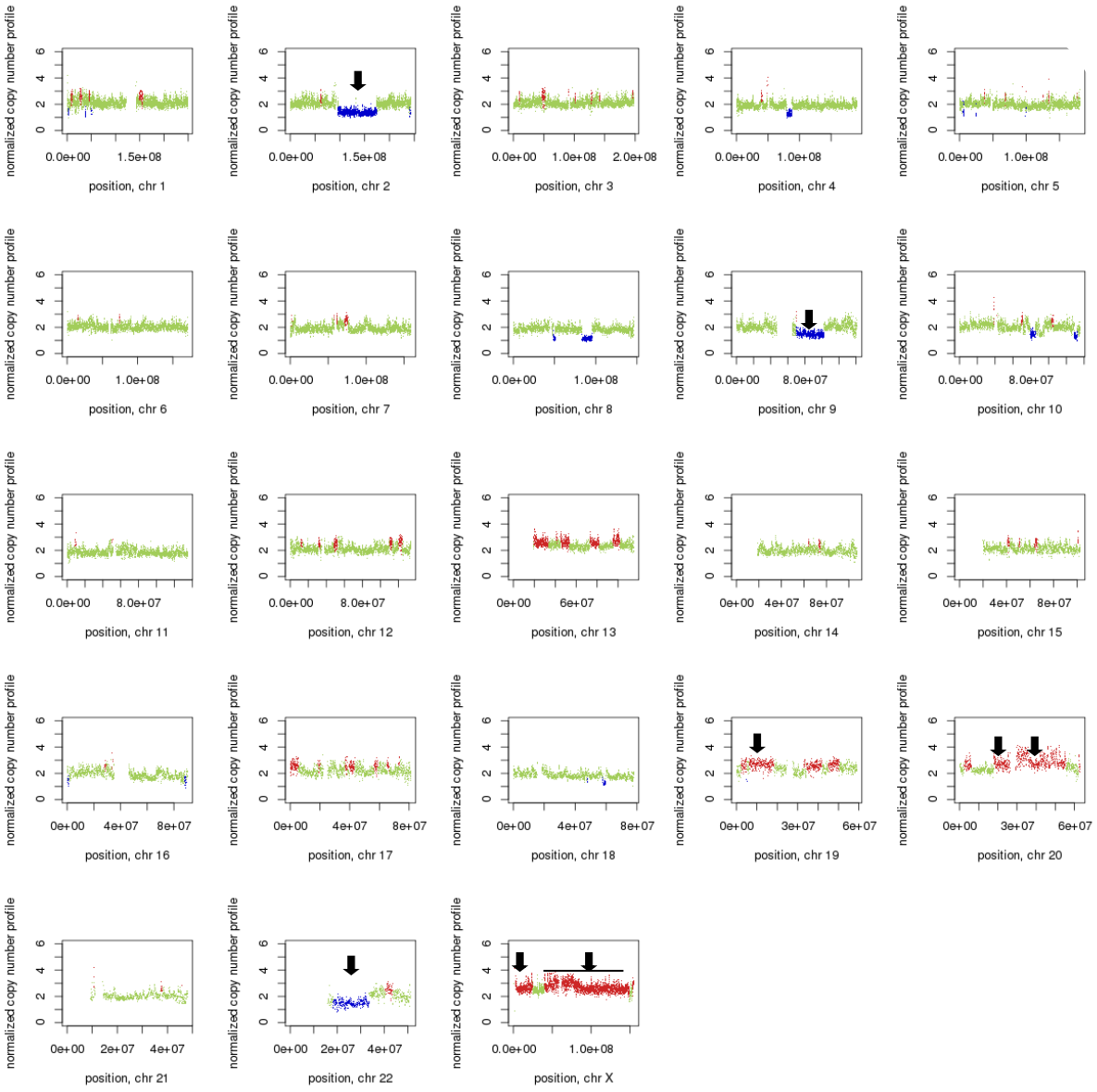
To predict focal and broad arm level amplification, the amplified and deleted regions in each of the sample were converted to the corresponding cytoband of a chromosome using CytobandIt tool. Considering the criteria of CNAs spanning 25% of the chromosome arm as large scale or broad arm level and those below 25% as focal-scale CNAs [198], focal and broad-level CNAs from each sample were determined. Overall, 14 broad-arm level amplification and 5 broad arm deletions were observed in addition to 221 focal amplification and 31 focal deletions. The size distribution for focal-scale CNA for amplification is 0.05-23.7 Mb and deletion is 0.1-36.6 Mb whereas broad arm amplification size ranges from 10.1-109.4Mb and for deletion 15.45-46.75 Mb.

Twelve cytobands- 12p13.33, 13q22.1, 13q34, 14q11.2, 17q12, 17q25.1, 1q21.3, 20q11.21, 22q13.31, 5p15.33, 7p11.2 and Xq28 showed overlap with the TCGA cervical data for amplification and 3 regions- 11p15.1, 16p13.3, 19p13.3 for deletion [94]. The recurrent amplifications in three samples were observed in 5 regions- 11p15.4, 16p11.2, 19p13.3, 21q22.3 and 3q29 from our data. Consistent with literature reports, recurrent amplifications are observed at chromosome 1q, 3q, 8q, 11p, 17q, 19q, 20q, 5p, 9q, 1p, 11q, 20p and 9p and recurrent deletions at chromosome 3p, 4q, 11p, 11q, 18q, 19p, 2q and 5q [96, 132, 199].

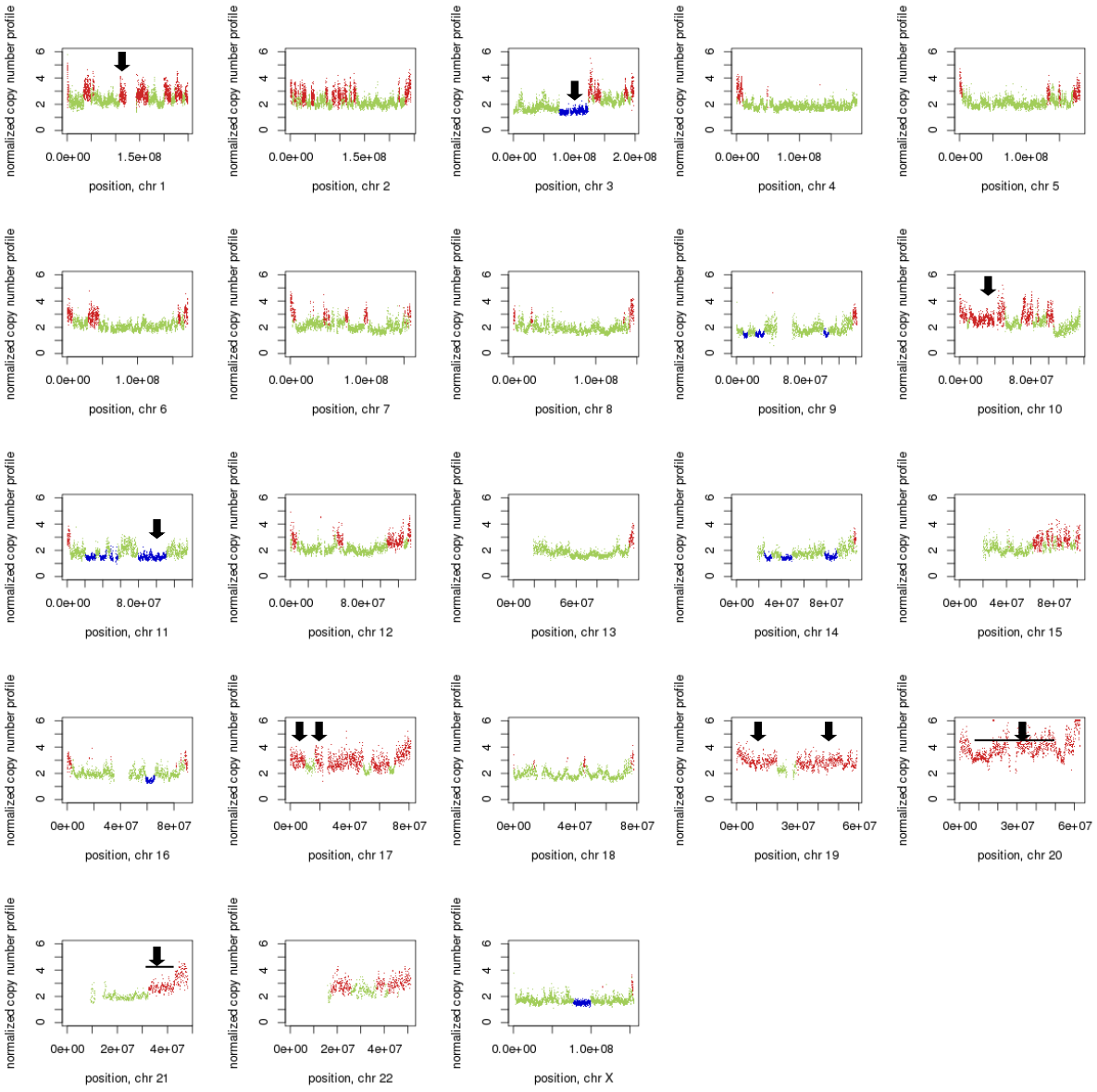
Control-FREEC provides visualization plots which show normalized copy number profile for each chromosome in a sample. Shown below are visualization plots for three samples. The X-axis refers to normalized ploidy level and Y-axis represents chromosomal genomic coordinates.



IV-Figure 1: Focal arm level copy number alterations in AD0708. AD0708 sample shows focal amplification of few regions indicated by red and one region with focal deletion.



IV-Figure 2: Broad and focal arm copy number changes in AD0718. AD0718 samples show amplified and deleted regions indicated by red and blue colour respectively. Broad-arm level copy number alterations are shown by arrows. 5 broad arm level amplification and 3 broad arm level deletions are predicted from the analysis.



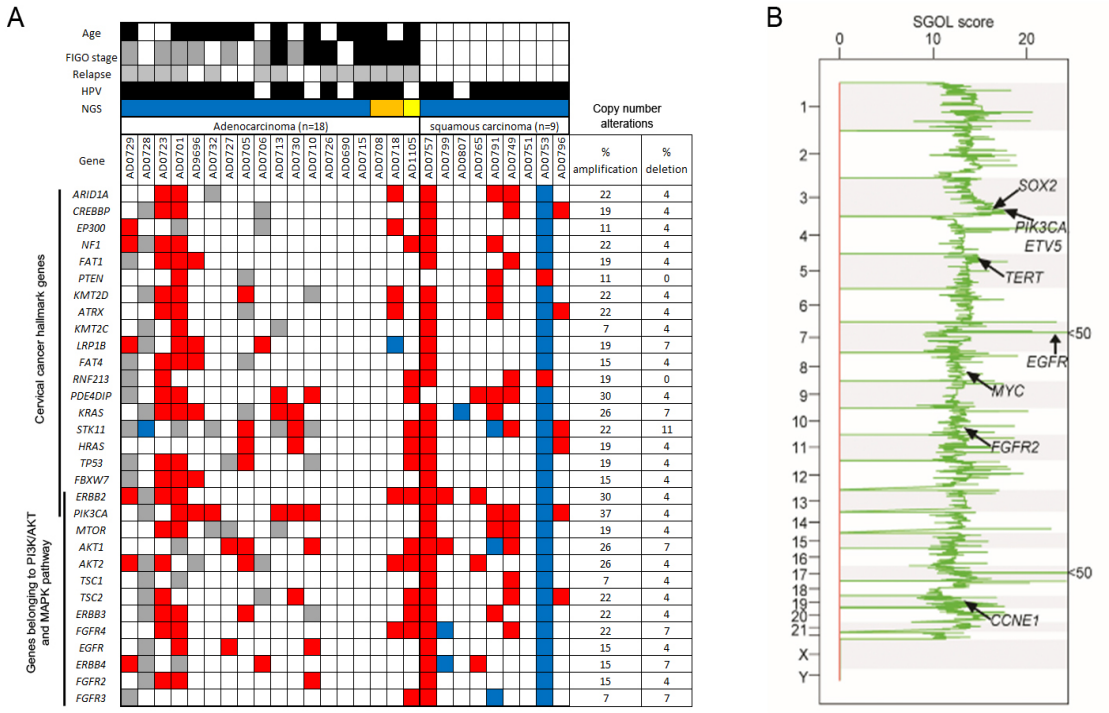
IV-Figure 3: Broad and focal arm copy number changes in AD1105. AD1105 samples show amplified and deleted regions indicated by red and blue colour respectively. 9 broad-arm level amplification and 2 broad arm level deletions are predicted from the analysis are indicated by arrows.

4.4.2 Copy number alterations identified from both histological subtypes of cervical cancer

Although whole-exome sequencing is designed to capture somatic variants, with recent advances in data analysis copy number alterations can be predicted from WES data with

several CNV detection tools [200]. CNV analysis in WES samples was performed similarly to WGS analysis using Control-FREEC.

From WES data, 214 genes with copy gain greater than 4 were identified across all samples with no significant copy deletions in cervical adenocarcinoma samples whereas 84 genes with copy gain 4 and above were found in squamous carcinoma. Combining copy number alterations identified from all the NGS analysis for both histological subtypes (n=27 samples) as shown in IV-Figure 4A, we report recurrent amplification in cervical cancer hallmark genes and genes belonging to PI3K/AKT and MAPK pathway. Recurrent amplification was observed in oncogenes *KRAS* (26%), *ERBB2* (30%), *PIK3CA* (37%) and epigenetic genes *EP300* (11%), *KMT2D* (23%), *KMT2C* (7%) among the hallmark genes. Amplifications are also observed in other cancer-associated tyrosine kinases like *ERBB3* (22%), *ERBB4* (15%), *EGFR* (15%), *FGFR2* (15%), *FGFR3* (7%) at a lower frequency. Amplification of PI3K/AKT pathway genes includes *mTOR* (19%), *AKT1/2* (26%), *TSC1* (7%) and *TSC2* (22%). Consistent with our mutation data observation, copy gain is recurrent in PI3K/AKT and MAPK associated genes, suggesting a potential role of these signalling pathways in cervical carcinogenesis. No KRAS mutations are detected in our dataset but an amplification of this gene is recurrent from copy number analysis. The copy number alteration data for all cancer-associated genes of both cervical subtypes in mentioned in appendix 3. The Segment gain or loss (SGOL) plot for adenocarcinoma subtype is shown in IV-Figure-4B.



IV-Figure 4: Copy number alterations in cervical cancer. A) Heatmap representation of copy number changes in cervical squamous and adenocarcinoma samples of hallmark genes and genes belonging to the PI3K-AKT and MAPK pathway. The red box indicates copy gain, blue for copy loss, grey for diploid samples and white box refers to information not available. On top, WES samples are represented by blue, WGS samples in orange and sample with WGS and WES data is shown in yellow. B) SGOL plot for cervical adenocarcinoma samples. X-axis indicates SGOL score and Y-axis indicate chromosome. Amplification is shown as green peaks on each chromosome. No significant deletions were observed from exome CNV analysis.

Next, copy amplification of a few candidate genes was validated in cervical adenocarcinoma samples using real-time PCR. A 60% concordance rate is observed between copy number gains predicted from bioinformatics analysis and real-time validation.

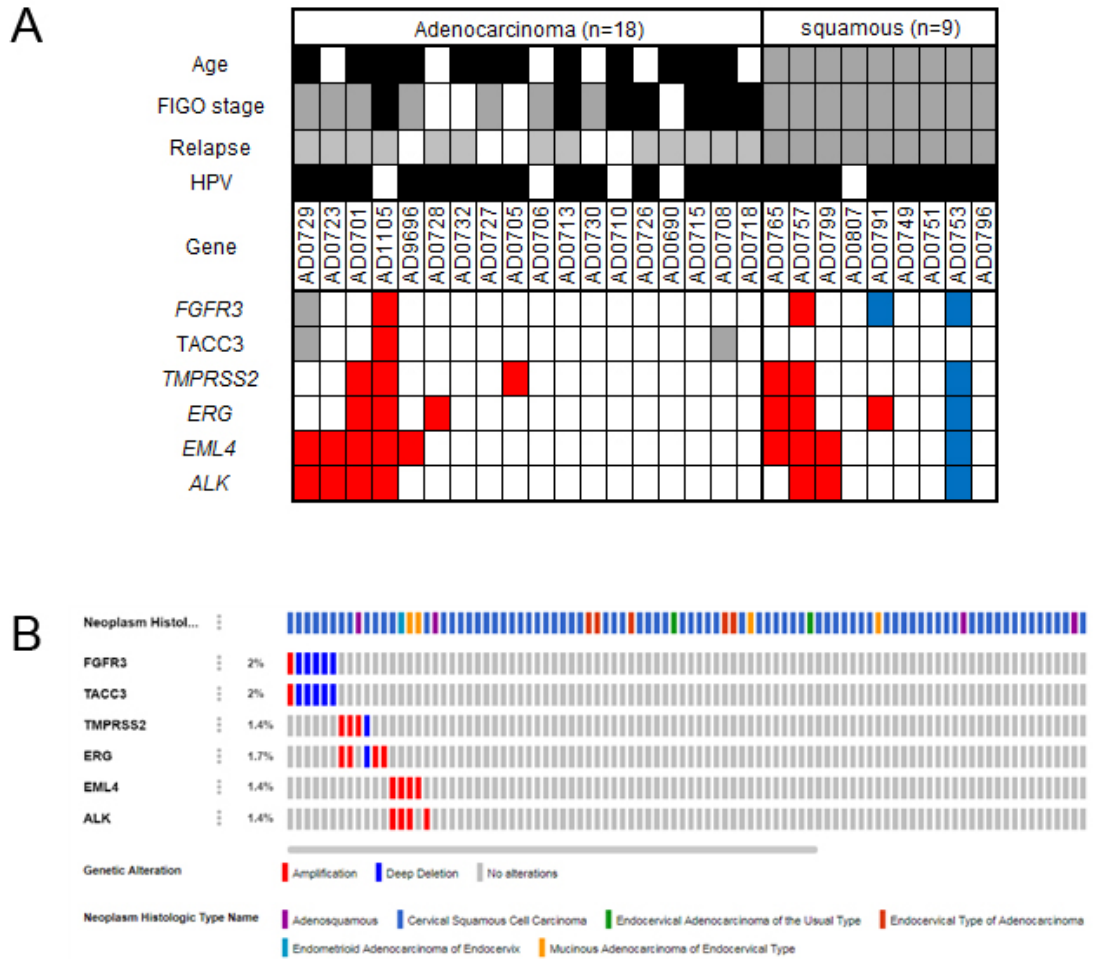
HPV																	
genes	1T	2T	3T	4T	5T	6T	7T	8T	10T	11T	12T	13T	14T	15T	16T	17T	% GAIN
CCND1																	25
EGFR																	44
ERBB2																	38
MYC																	13
PIK3CA																	38
SOX2																	50
TERT1																	44

IV-Figure 5: Copy number validation of candidate genes using real-time PCR. The copy gain or loss predicted from the bioinformatic analysis was verified in 16 paired samples of exome sequenced cervical adenocarcinoma by real-time PCR. The red box indicates copy amplification; white box indicates diploid copy number; black box refers to mutation whereas grey box represents both mutation and copy gain in samples.

Amplification of genes *SOX2* (63%), *PIK3CA* (50%), *TERT1* (50%), *CCND1* (44%) and *MYC* (37%) along with amplification of ERBB family members- *ERBB2* (63%) and *EGFR* (56%) was recurrent as detected by real-time PCR. Mutation and copy number alterations seem to be mutually exclusive except for 1 sample in which *PIK3CA* is mutated along with copy amplification.

As per TCGA data of cervical cancer on cBioPortal, *ERBB2* amplification is observed in 5% of the TCGA cervical cancer samples (n=295) and copy gain is observed in adenocarcinoma subtype only whereas deletion of tumor suppressor genes- *ATM* and *LRP1B* are observed at 5% and 10% frequency respectively in the TCGA dataset and the copy number alterations are common to the squamous carcinoma subtype.

Moreover, we note an interesting observation. Co-occurring copy gain and loss in oncogenic fusion partners *FGFR3-TACC3, TMPRSS2-ERG* and *EML4-ALK* were observed within the same sample in our dataset as well as the TCGA dataset as shown in the Figure below (IV-Figure 6A, 6B). The functional significance of such observation needs further validation.



IV-Figure 6: Co-occurring copy gain and loss in 3 gene pairs belonging to oncogenic fusions observed in our dataset (A) and TCGA dataset (B).

4.4.3 Structural variant identification from WGS data

Structural variations were called from 3 paired samples using BreakDancer tool. The gene translocation pairs were identified as genomic coordinates which were then annotated to obtain gene pair information. Excluding entries with gene translocation in the un-annotated regions and non-coding regions, only coding genes were considered. Depleting gene translocations identified in normal samples, tumor-specific events were identified. Overall, 67 gene translocations were identified in all three samples (IV-Table 3). Two structural rearrangements *ARHGAP11B-ARHGAP11A* and *CDK11B-SLC35E2B* were recurrent in two samples. Of these 67, 16 events are intra-chromosomal (ITX), 22 are inter-chromosomal

(CTX), 23 inversions (INV), 5 deletions (DEL) and 1 insertion (INV). Two samples show structural variations with *MUC3B* as one of the gene partner- *CAPN8-MUC3B* and *PNKD-MUC3B*. *FAM172A-ESR1* has ESR1 gene as a partner. *ESR1* role in breast cancer is well established and *ESR1-YAP1* translocation confers resistance to endocrine therapy [201]. *ESR1* gene translocation has not been reported in cervical cancer previously. Other gene translocations involving a known partner (highlighted in bold) reported in a pan-cancer analysis of chromosomal rearrangement [202] include ***DUX4****-ROCK1P1****, MLLT4****-KIF25****, **MSH2**-TAF4****, PAX7****-EEF1A2****, PBX1****-SIK3* and ***PDGFRA****-MAN2A1*.

DUX4, *MLLT4* and *PBX1* are common gene partners in the translocations observed in Leukaemia [203-205]. *PAX7* is a gene translocation partner in rhabdomyosarcoma [206] and *PDGFRA* rearrangements are observed in gliomas [207]. The role of above gene translocation events in cervical adenocarcinoma samples needs further validation to comment on the potential oncogenic role.

None of these gene translocation events are common to the structural variations and gene fusions reported in TCGA cervical, ChimerDB3.0, cancer gene census gene fusions, COSMICdb, TICdb [208] or cancer genome interpreter database [209].

Structural Variant	ChrA	PosA	ChrB	PosB	SV type	AD0708	AD0718	AD1105
CAPN8-MUC3B	chr1	2.24E+08	chr7	1.01E+08	CTX			
FAM172A-ESR1	chr5	93117678	chr6	1.52E+08	INV			
GNG12-AS1	chr1	68427727	chr10	1.33E+08	CTX			
HPN-AS1	chr19	35596405	chr19	35596854	ITX			
MSH2-TAF4	chr2	47863548	chr20	60557557	CTX			
PAX7-EEF1A2	chr1	19071184	chr20	62124411	CTX			
PBX1-SIK3	chr1	1.65E+08	chr11	1.17E+08	INS			
PDGFRA-MAN2A1	chr4	55008677	chr5	1.09E+08	INV			
PNKD-MUC3B	chr2	2.19E+08	chr7	1.01E+08	CTX			
RFWD2-PRIM2	chr1	1.76E+08	chr6	57399354	INV			
ST6GALNAC5-PRIM2	chr1	77357035	chr6	57257259	INV			
TTC28-AS1	chr22	28380829	chr22	28381046	DEL			
AP2A2-MUC6	chr11	991103	chr11	1023369	ITX			
ARHGAP11B-ARHGAP11A	chr15	30923677	chr15	32914666	ITX			
CCZ1-CCZ1B	chr7	5948904	chr7	6861157	INV			
CDK11B-SLC35E2B	chr1	1584241	chr1	1655890	ITX			
EVPLL-APP	chr17	18291769	chr21	27375197	INV			
GRID1-DOCK1	chr10	87817816	chr10	1.29E+08	INV			
HHAT-KCNH1	chr1	2.11E+08	chr1	2.11E+08	ITX			
LOC100507387-FAM153A	chr5	1.76E+08	chr5	1.77E+08	INV			
MZT2B-MZT2A	chr2	1.31E+08	chr2	1.32E+08	INV			
PCDHB8-PCDHB13	chr5	1.41E+08	chr5	1.41E+08	ITX			
PTP4A3-MROH5	chr8	1.42E+08	chr8	1.42E+08	ITX			
SLC2A14-NANOGP1	chr12	7972003	chr12	8051836	DEL			
SPATA21-NOB1	chr1	16735633	chr16	69778874	INV			
TMEM61-BSND	chr1	55453896	chr1	55473190	ITX			
BX647938-OVOS2	chr12	9719565	chr12	31278291	INV			
CATSPER2-STRC	chr15	43931774	chr15	44018271	DEL			
CERS3-PRKXP1	chr15	1.01E+08	chr15	1.01E+08	ITX			
CLEC18C-GLG1	chr16	70119756	chr16	74583981	INV			
DTX2P1-UPK3BP1	chr7	76626641	chr7	76628321	ITX			
DUX4-ROCK1P1	chr18	103886	chr18	113081	ITX			
MLLT4-KIF25	chr6	1.68E+08	chr6	1.68E+08	ITX			
UIMC1-CLK4	chr5	1.76E+08	chr5	1.78E+08	ITX			
AP3S2-YBEY	chr15	90427768	chr21	47709980	CTX			
ARPC1A-TRPV3	chr7	98965755	chr17	3447788	CTX			
AUTS2-PXDNL	chr7	70090713	chr8	52610021	CTX			
BSDC1-ARHGAP11B	chr1	32858755	chr15	30967300	CTX			
CDC14A-AK093107	chr1	1.01E+08	chr22	48206258	CTX			
COL6A5-THSD4	chr3	1.3E+08	chr15	71492561	CTX			
CSMD2-HPSE2	chr1	34530268	chr10	1E+08	INV			
CYP11B1-CYP11B2	chr8	1.44E+08	chr8	1.44E+08	DEL			
FCGR3A-FCGR3B	chr1	1.62E+08	chr1	1.62E+08	ITX			
GFPT1-C15orf32	chr2	69601021	chr15	93028438	INV			
GTF2B-IMMP2L	chr1	89352982	chr7	1.11E+08	INV			
HERC2-HERC2P10	chr15	28419737	chr15	31110412	INV			
INSIG2-TSHZ2	chr2	1.19E+08	chr20	51802383	INV			
ITPR2-SCARB1	chr12	26958424	chr12	1.25E+08	ITX			
LYST-TASP1	chr1	2.36E+08	chr20	13309224	INV			
MCOLN2-CERS6	chr1	85446795	chr2	1.7E+08	INV			
MGLL-PTPRN2	chr3	1.27E+08	chr7	1.58E+08	CTX			
NFASC-TCERG1L	chr1	2.05E+08	chr10	1.33E+08	CTX			
P2RX5-TAX1BP3	chr17	3583471	chr17	3583532	ITX			
PPP1R12B-ZNF681	chr1	2.02E+08	chr19	23936477	CTX			
PSG3-PSG6	chr19	43321110	chr19	43489688	DEL			
PTPRN2-RNLS	chr7	1.58E+08	chr10	89894464	CTX			
SLC41A3-SBNO2	chr3	1.26E+08	chr19	1142313	CTX			
SLC4A8-RYR1	chr12	51905016	chr19	39032878	CTX			
SNURF-SNRPN	chr12	51905016	chr19	39032878	CTX			
STXBP5L-EAF2	chr3	1.21E+08	chr3	1.22E+08	INV			
SYCP1-MOB2	chr1	1.15E+08	chr11	1692356	CTX			
TENM3-JUP	chr4	1.84E+08	chr17	39789566	CTX			
TMEM232-MACROD2	chr5	1.1E+08	chr20	14140941	INV			
TMEM56-KIRREL3	chr1	95575545	chr11	1.27E+08	INV			
TYW1-TYW1B	chr7	66513106	chr7	72244571	INV			
UTRN-MYO7A	chr6	1.45E+08	chr11	76904047	CTX			
ZMAT4-ATP11A	chr8	40427893	chr13	1.14E+08	CTX			

IV-Table 3: List of structural variations of the coding region that are identified from 3 paired samples. CTX: inter-chromosomal; ITX: intra-chromosomal; INV: inversion, INS: insertion and DEL: deletion

4.5 Discussion

Cervical adenocarcinoma is a rarer subtype of cervical cancer with low incidence rate and therefore, less genomically characterized as compared to squamous counterpart. Whole-genome sequencing data is available for very few numbers of samples. Not much information on structural variations in cervical cancer is reported. Therefore, structural re-arrangements acting as drivers in disease progression often goes undetected. Genome-wide copy number alterations contributing to increased gene expression and mutations serving as a driver can also be identified from WGS analysis. Thus, this approach can help us identify multiple alterations within a single sample to identify potential therapeutic targets. In this study, we performed WGS on 3 paired samples of cervical adenocarcinoma to learn about copy number alterations and structural re-arrangements. Further, we combined copy number data from WES for both cervical histological subtypes to represent the overall landscape of Copy Number Variations in cervical cancer

Here, we report recurrent gains in known genes *PIK3CA, ERBB2, MYC, BRCA1, TERT* and *SOX2* and novel genes *KRAS, FGFR2, FGFR3, ERBB3* and *ERBB4*. Among the PI3K/AKT pathway genes, *AKT1* and *AKT2* are amplified in the same sample whereas several of the MAPK pathway upstream genes belonging to ERBB and FGFR family were amplified. We note an interesting observation that amplification of both genes belonging to oncogenic fusion *TMPRRSS2-ERG, EML4-ALK* and *FGFR3-TACC3* are detected in the same samples, even though these fusions are not detected by SV analysis in WGS samples. Moreover, broad and focal level amplifications were predicted in WGS samples based on the criteria that alteration range more than 25% of the chromosomal arm. Overall, 14 broad-arm level amplification and 5 broad arm deletions, 221 focal amplification and 31 focal deletions were

present in total. The recurrent amplification at chromosome 1q, 3q, 8q, 11p, 17q, 19q, 20q, 5p, 9q, 1p, 11q, 20p and 9p and deletion at 3p, 4q, 11p, 11q, 18q, 19p, 2q and 5q chromosomal regions were consistent with TCGA and other literature reports [96, 132, 199].

Structural variation analysis detected 5 genes with known fusion gene pair reported in cancers namely, ***DUX4***-*ROCK1P1*, ***MLLT4***-*KIF25*, ***MSH2***-*TAF4*, ***PAX7***-*EEF1A2*, ***PBX1***-*SIK3* and ***PDGFRA***-*MAN2A1*. *DUX4*, *MLLT4* and *PBX1* are common gene partners in the translocations observed in Leukaemia [203-205] whereas *PAX7* is a known gene translocation partner in rhabdomyosarcoma [206] and *PDGFRA* rearrangements are commonly observed in gliomas [207]. We speculate that these genes might also be playing an oncogenic role in cervical cancer. All the structural variation events we report are unique and not reported in any of the literature or databases.

In conclusion, we describe the first landscape of structural variations in cervical cancer from the Indian population.

Chapter V

Identifying Cancer Driver Genes from Functional Genomics Screens

(As published in Swiss Medical Weekly (2020))

Chapter V: Identifying Cancer Driver Genes From Functional Genomics Screens

(As published in Swiss Medical Weekly (2020))

5.1 Abstract

With emerging advances in genomics and functional genomics approaches, there is a critical unmet need to integrate plural datasets to identify driver genes in cancer. An integrative approach, with the convergence of multiple genetic evidences, can limit false positives by adopting a posterior filtering strategy and reduce the burden for multiple hypotheses testing to identify true cancer vulnerabilities. We performed a pooled shRNA screen against 906 human kinase genes in an oral cancer cell line AW13516 independently by two different approaches. The genes depleted in the screen were ranked based on ROAST analysis and integrated with copy number alteration and gene expression data using an integrative scoring system 'DepRanker' to compute an 'Rank Impact Score (RIS)' for each gene. The RIS based ranking of candidate driver genes identified known and putative oncogenes such as AURKB and TK1 are essential for oral cancer cell proliferation and also altered in human cancers. We further validate these findings showing that shRNA mediated genetic knockdown of TK1 or pharmacological inhibition of AURKB by AZD-1152 HQPA in AW13516 cells could significantly impede the proliferation of the cells. Next, we analyzed the alteration in AURKB and TK1 genes in head and neck cancer and their association with prognosis using data obtained 528 patients from the TCGA, wherein patients harbouring alteration in AURKB and TK1 genes were associated with poor survival. Thus, we present DepRanker as a simple yet robust package to identify potential driver genes from a pooled shRNA functional genomic screen by integrating results from RNAi screens with the gene expression and copy number data. Using the DepRanker we identify AURKB and TK1 as

potential therapeutic targets in oral cancer. DepRanker is available in public domain for download at http://www.actrec.gov.in/pi-webpages/AmitDutt/DepRanker/DepRanker.html.

Keywords: Pooled RNAi screen, kinase, genomics, DepRanker, AURKB, TK1

Abbreviations: AURKB = aurora kinase B; CR = copy number alteration rank; DepRanker = dependency ranker; DR = depletion rank; FC = foldchange; GR = gene expression rank; GUI = graphic user interface; MOI = multiplicity of infection; RIS = Rank Impact Score; RNAi = RNAinterference; RR = ROAST rank; TCGA = The Cancer Genome Atlas; TK1 = thymidine kinase 1; W = Weight

5.2 Introduction

Cancer is a disease defined by several genetic alterations like mutation, gene expression and copy number changes in addition to epigenomic alterations [210]. While most of the alterations are passenger alterations with no significant effect on cellular phenotype, cancer cells are dependent on few driver genes for constitutive activation of signalling pathway which aids in cellular proliferation, a phenomenon described as oncogene addiction [24]. Targeting of oncogenic dependent genes has resulted in success as demonstrated in several cancer types [186, 211]. Often, the discovery or identification of a cancer-associated driver oncogene based on genomics approach necessitates screening for significant genetic alterations using stringent statistical methods followed by functional validation. On the other hand, a complementary functional genomics approach using RNAi or CRISPR effectively converts this structural knowledge of the cancer genome to define the functional consequences of the alterations, in an unbiased manner that may be performed in a pooled or arrayed format [212]. Methods to perform genome-wide RNAi screens with pooled human shRNA library on human cancer cell lines as experimental models offer a powerful methodology to identify genes essential for the survival of the cells. These efforts represent a new opportunity to fundamentally alter the scale and manner by which we are able to understand and validate molecules that when targeted lead to therapeutic benefit in cancer patients.

Typically, a pooled RNAi screen analysis involves quality assessment and normalization of the data followed by differential shRNA/ sgRNA representation. The differential analysis is either performed by custom scripts or packages like edgeR [213]. The “tags” (shRNA) are ranked according to their differential effects among classes of samples and further summarized into the top list of genes by packages like RIGER [214], RSA [215], ROAST [216], camera [217] and others. Moreover, there are specialized algorithms like DEMETER2

[218], to measure the on/off-target effect and also estimate gene-dependency by assigning 'essentiality scores' from the RNAi experiments. The genes obtained from these experiments may further be validated either by performing specific knock-down experiments or by extended secondary screens.

An alternative approach applied to define dependency from pooled screen experiments is the integration of genomic data along with the gene essentiality results. A classic example of this approach is the cancer dependency map [53], which integrates other genomic features such as expression, copy number and mutation information along with the gene dependencies obtained from screens performed on cancer cell lines representing various tumor types. Few computational methods also incorporate such genomic features in predicting driver or essential genes for pooled RNAi screen experiments [54]. Building on this integrative approach we have developed a gene ranking or scoring method, DepRanker which incorporates other genomic datasets like gene expression and copy number information of the same cell line, to prioritize genes from pooled screen results for their essentiality. The DepRanker consists of two modules that can be executed using a single user friendly GUI. Module I performs analysis of the pooled screen data to calculate the depletion of the tags and prioritize the genes, respectively. Module II integrates the results obtained from Module I with the genome wide datasets to compute the 'Rank Impact Score (RIS)' for individual genes.

Here, we performed a functional kinome screen using pooled shRNA comprising of 5419 constructs targeting 906 human kinases in AW13516 cells in two independent screens. The genes depleted in the screen were integrated with copy number alteration data and gene expression data for the AW13516 cells using 'DepRanker' to identify *AURKB* and *TK1* as potential therapeutic targets in oral cancer.

5.3 Material and Methods

5.3.1 Cell lines and cell culture

Indian patient derived head and neck cancer cell lines- AW13516 cells and other cells used in the study- 293FT, HCT116 and SiHa cells were maintained in Dulbecco's Modified Eagle's Medium (Gibco) supplemented with 10% FBS (Gibco) and 1% Penicillin-streptomycin solution (Sigma). Cells were grown at 37^0C in a 5% CO_2 incubator. Cells were treated with Mycoplasma elimination kit (EZKill solution, Himedia) prior to use.

5.3.2 Lentivirus production and transduction in HNSCC cell line

Lentivirus comprising of pZIP-SFFV pooled shRNA constructs (8.1 Kb) comprising of 5419 shRNA targeting 906 human kinases were obtained from TransOMIC Technologies, USA. For pooled shRNA screen, 18 million AW13516 cells were seeded in T-150 flasks at 60-70% confluency and lentivirus was transduced at a M.O.I of 0.3 in presence of 8 ug/ ml polybrene (Sigma) at 1000X fold representation of each shRNA in the screen. Cells were grown at 37^0C for 16 hours post virus addition and media was replaced. Cells were selected in presence of 1 ug/ml puromycin (Sigma) and half of the cells were harvested within 3-4 days after selection and this sample was termed as Day 0 (control) sample. Remaining cells were further expanded and maintained at 37^0C and further collected at Day 10 and Day 20 time point (test samples).

5.3.3 PCR amplification of shRNA and barcode sequencing by NGS

Genomic DNA was extracted from Day 0, Day 10 and Day 20 samples of AW13516 cells using QIAamp DNA blood kit (Qiagen). DNA concentration estimation was done using Nanodrop 2000c spectrophotometer (Thermo Fischer Scientific). Instructions provided in the TransOmics manual were followed for performing PCR for shRNA amplification with some modifications. For representing a fold of 1000 per shRNA, 36 ug of genomic DNA

was used to amplify shRNA cassette as per the calculation and primary PCR was performed (sequence information in V-Table 1) as follows – 10 ul of 5X HF buffer, 1.5 ul of 10 uM of each forward and reverse primary PCR primers, 1 ul of 10 mM dNTP mix, 5% DMSO, 3 mM $MgCl_2$, 0.5 ul of Phusion high-Fidelity polymerase enzyme (Thermo Fischer Scientific) and 850 ng genomic DNA in a total reaction volume of 50 ul. Primary PCR was performed at thermo-cycler conditions- 98^0C for 5 min, 25 cycles of 95^0C for 30 sec, 57^0C for 30 sec and 72^0C for 30 sec and final extension at 72^0C for 5 min. The PCR product was separated on 1.5% agarose gel to visualize an amplicon of 406 bp. Next, primary PCR product was pooled and purified using Nucleospin Gel and PCR clean-up kit (Macherey-Nagel) and quantified using Nanodrop 2000c. 2 ug of purified primary PCR was used for setting up nested secondary PCR (primer sequence information in V-Table 1) with indexed reverse primers that adds unique barcode sequence to each sample to facilitate sample pooling during NGS sequencing. Secondary PCR reaction comprised of 10 ul of 5X HF buffer, 1.5 ul of 10 uM for each of forward and indexed reverse secondary PCR primers, 1 ul of 10 mM dNTP mix, 5% DMSO, 0.5 ul of Phusion high-Fidelity polymerase enzyme (Thermo Fischer Scientific) and 500 ng of primary PCR product in a total reaction volume of 50 ul. Secondary PCR was performed at thermo-cycler conditions- 98^0C for 5 min, 15 cycles of 94^0C for 30 sec, 52^0C for 30 sec and 72^0C at 30 sec and final extension at 72^0C for 5 min. Secondary PCR product was separated on 1.5% agarose gel to visualize a band of 408 bp. Then, secondary PCR product was pooled and subjected to purification using AgencourtAmpure XP beads (NEB) and quantitated using QubitFluorometer (Thermo Fischer Scientific). About 8-20pM of secondary PCR purified product (indexed library) was loaded on IlluminaHiSeq 2500 platform and 50 bp single end sequencing was done.

Primers	Sequence
OAD1710_Primary PCR_F	CAGAATCGTTGCCTGCACATCTTGGAAAC
OAD1711_Primary PCR_R	CGTATCCACATAGCGTAAAAGGAGCAAC
OAD948_Secondary PCR_F	AATGATACGGCGACCACCGAGATCTACACACACT CTTTCCCTACACGACGCTCTTCCGATCTTAGTGAA GCCACAGATGTA
OAD1305_Secondary PCR_R_index1 (Day 0)	CAAGCAGAAGACGGCATACGAGAT**CGTGAT**GTG ACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGTA TCCACATAGCGTAAAAGG
OAD1306_Secondary PCR_R_index2 (Day 10)	CAAGCAGAAGACGGCATACGAGAT**ACATCG**GTG ACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGTA TCCACATAGCGTAAAAGG
OAD1307_Secondary PCR_R_index3 (Day 20)	CAAGCAGAAGACGGCATACGAGAT**GCCTAA**GTG ACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGTA TCCACATAGCGTAAAAGG
OAD947_NGS Read 1 loop	ACGACGCTCTTCCGATCTTAGTGAAGCCACAGAT GTA
OAD1788_AURKB_sh1_F	GCCACGATCATGGAGGAGT
OAD1789_AURKB_sh2_F	CCGAGAAGAAAAGCCATTTCAT
OAD1790_AURKB_sh3_F	TGCCCAGAAGGAGAACTCCT
OAD1791_AURKB_sh4_F	ACCATGGGAAGAAGGTGATTC
OAD1792_AURKB_sh5_F	CTGCAGAAGAGCTGCACAT
OAD1793_NTRK2_sh1_F	GGCCGAACAGAAGTAATGAAAT
OAD1794_NTRK2_sh2_F	GCCAGACACTCAGGATTTGTAC
OAD1795_NTRK2_sh3_F	CCCTGAGAACATCACCGAAAT
OAD1796_pooled_all_R	CCGGCAAGGTATTCAGTTTTAG
OAD1797_TBK1_sh1_F	GCCAGAGTTAGGTGAAATTTCA
OAD1798_TBK1_sh2_F	GGCGAAGACATAAGAAAACTGGT
OAD1799_TIE1_sh1_F	GCCAGAACTGGAGTTCAACTTA
OAD1800_TIE1_sh2_F	AGAGGAGACAAGCACCATCAT
OAD1801_TIE1_sh3_F	CCAAGGTCACACACACTGTGA
OAD1802_TIE1_sh4_F	AGGCATCTACAGTGCCACTTA
OAD1803_TIE1_sh5_F	TGAGCAGTGCCCAGGCAT
OAD1804_TK1_sh1_F	AGGTGATTGGGGGAGCAG
OAD1805_TK1_sh2_F	GAAAAAAGCACAGAGTTGATGA
OAD1806_TK1_sh3_F	CCAGTACAAGTGCCTGGTGAT
OAD1807_TK1_sh4_F	TGGTGATTCTCGGGCCGA

V-Table 1: Primer information

5.3.4 Data analysis of pooled shRNA using edgeR pipeline

Raw data was obtained as Fastq files, which was further processed using edgeR package [213] for analysis of pooled shRNA data. Counts for each shRNA were obtained per sample by mapping reads with the kinase shRNA sequence library. For screen 1 data, shRNA with counts less than 1000 in control sample (Day 0) were excluded since experiment was

performed at 1000X fold representation. For screen 2 and 3, a cut-off of 100 shRNA in control sample (Day 0) was considered for further analysis. Data normalization was performed within and across control and test samples. Screen data was analyzed using classical method of two group comparisons. Statistical analysis was done to estimate significance of the changes observed in shRNA abundance. edgeR provided a list of depleted shRNA belonging to kinases by calculating log fold change (Log FC). Based on these results, top enriched and depleted shRNA from the screen were identified and further converted to gene-level ranking using gene set analysis tool 'ROAST' [216]. Kinases represented by at least 2 shRNA were considered for further analysis. A list of kinases that were depleted in cells over Day 20 compared to Day 0 was obtained. Data from screen 1, 2 and 3 were not considered as data in triplicates because screen 1 data output was enormous and captured existing shRNA uniformly whereas screen 2 and screen 3 data output was comparatively lower, suggesting that few of the shRNAs were not captured. Therefore, screen 2 and screen 3 data was used as replicates. Hence, the combined results of screen 2 and screen 3 are referred to as Screen 2 data here after.

5.3.5 DepRanker assigns impact score for identification of potential kinase using genomic alteration data

To further prioritize the candidate kinases obtained from the RNAi screen analysis, we developed a scoring method named DepRanker (Dependency Ranker).DepRanker calculates 'Rank Impact Score' (RIS) for individual kinases which are derived from the kinome screen, calculated by integrating gene expression and copy number data from the same sample.RIS is derived as follows (equation):

$$RIS_{(Kinase\ A)} = DR_{(Kinase\ A)} + RR_{(Kinase\ A)} + GR_{(Kinase\ A)} + CR_{(Kinase\ A)}$$

where DR= Depletion rank, RR= ROASTrank, GR= Gene expression rank and CR= Copy number alteration rank

We use mean-rank method to calculate scores for each feature as described below. The DR is derived by converting the logFC values obtained from the edger depletion analysis into ranks. The kinase showing the highest depletion in the screen is assigned the highest rank and the one showing lowest depletion is assigned a rank '1'. The RR is based on the ranking given by the ROAST algorithm, in which the genes which are represented by at least 3 shRNA is considered and the kinases are sorted based on the *p-value*. The gene which is least prioritized by ROAST is given a rank of '1' and the top-most gene is assigned the highest rank. Further to calculate GR and CR, for all the kinases showing significant depletion in the pooled screen analysis, gene expression and copy number alteration data is extracted for the cell line. In this analysis, for all the kinases showing significant depletion in the pooled screen were extracted for AW13516 cells (as previously described [219]). The log transformed FPKM gene expression levels were extracted for the subset of kinases (obtained from the pooled screen result). Among this subset, the one showing lowest gene expression was assigned a rank of 1 and the gene with highest was assigned the highest rank in the ascending order. Similar ranking was assigned to the copy number levels for individual kinases from the AW13516 cells, to derive CR. Addition of all the four scores (DR, RR, GR, CR) was used to compute the RIS. This scoring approach enabled us to identify potential kinases with biological role from the list. Further to combine the results obtained from two screens performed on the same cell line we converted the RIS for individual kinase into *weight* (ranged from 0-1) based on its relevance in a particular screen. Further to combine the results from both the screens, we assigned weighting to each of the kinase by considering RIS for both screens. The weight was calculated using the formulae, W= RIS (Kinase A)/total of RIS for all the kinases. The results from both the screens were combined and sorted based on the weightings assigned. In case of kinases with overlap in both the screens, kinase with higher weight was retained.

5.3.6 Implementation of DepRanker and graphical user interface

This scoring system is implemented as a python-based package. DepRanker takes the output from edgeR pooled shRNA screen and result provided by ROAST, along with gene expression data and copy number variation data for individual gene belonging to the cell line and outputs the list of candidate kinases with Rank Impact Score. The package is available athttp://www.actrec.gov.in/pi-webpages/AmitDutt/DepRanker/DepRanker.html with complete installation instructions and user manual. The GUI was designed using Tkinter (https://wiki.python.org/moin/TkInter) python package. A detailed user manual for the GUI is available at http://www.actrec.gov.in/pi-webpages/AmitDutt/DepRanker/DepRanker.html. The GUI provides two modules for analysis. The first module is the pooled shRNA screen analysis module which takes in the fastq, hairpin and sample information file to perform the depletion analysis. The depletion analysis can be performed either by generalised linear model (GLM) or exact-test based method. The users are advised to refer Zuber *et al*., (http://bioinf.wehi.edu.au/shRNAseq/pooledScreenAnalysis.pdf) screen analysis manual for selection of suitable method for their screen data analysis. Internally the GUI calls the Bioconductor packages, edgeR and ROAST, for performing the depletion analysis and gene prioritization respectively. The results from this module (edger toptags result and ROAST result file) along with the copy number and gene expression data for the cell line analysed should be provided to the DepRanker module. This module provides the rank-based scores for individual kinases identified from the pooled screen. The DepRanker GUI package freely available for download at http://www.actrec.gov.in/pi-webpages/AmitDutt/DepRanker/DepRanker.html.

5.3.7 Survival analysis of HNSCC datasets

Genomic alteration data from TCGA provisional HNSCC datasets was assessed from cBioPortal [220] consisting of 528 samples with gene expression, copy number and mutation information. Kaplan- Meier survival plots were generated for patients having alterations in *AURKB* and *TK1* genes respectively.

5.3.8 Real-time PCR for amplification of shRNA

Real time primers were designed for each shRNA of *AURKB* and *TK1* wherein the forward primer sequence was complementary to kinase shRNA sequence and the reverse primer was common for all, binding to the 3' miR vector sequence. PCR was performed using purified primary PCR product as a template. An amplicon of 100 bp size is expected. Primer sequences are provided in V-Table 1.

5.3.9 MTT assay for functional validation of hit obtained from screen

MTT assay was performed using *AURKB* inhibitor AZD1152-HQPA (Sigma). Colon cancer cell line HCT116 (sensitive) and cervical cell line SiHa (resistant) were used as control cells for MTT assay. In brief, 1000 cells of AW13516, 1500 cells of HCT116 and 2000 cells of SiHa were seeded in 96 well plates respectively. Cells were treated with AZD1152-HQPA inhibitor for 72 hours following which MTT (0.5 mg/ml) reagent was added and cells were incubated for 3 hours at 37^0C in CO_2 incubator. DMSO was used for developing and reading was obtained at 570 nm using microplate reader (iMarkmicroplate reader, Biorad). Percent cell viability was calculated with respect to control untreated cells. The assay was performed thrice.

5.3.10 Generation of TK1 knockdown clones of AW13516:

pZIP-hCMV shRNA constructs targeting TK1 genes and scramble control (TransOmics Technologies, USA) was used for lentiviral production in 293FT cells using Lipofectamine

3000 transfection reagent (Invitrogen). Lentivirus was harvested at 48 and 72 hours respectively and filtered using 0.4uM filter. AW13516 cells were transduced with virus in presence of 8ug/ ml concentration of polybrene and selection was done using 1ug/ml puromycin for 4-5 days. Cells selected were positive for GFP expression. The shRNA sequences are as follows: TK1 sh1- AAGCAGACAAGTACCACTCCG and TK1 sh2 – CCCAGGTGATTCTCGGGCCGA.

5.3.11 Western blotting

Cells were lysed in RIPA lysis buffer (Sigma) supplemented with 1 mMdithrothreitol (DTT) and protease inhibitor cocktail (Calbiochem, Merck) and quantitated using BCA protein estimation method. 40 ug of protein was loaded on 12% SDS-PAGE gel, transferred onto PVDF membrane (AmershamHybond, GE healthcare) by electro blotting. Membrane was stained with Ponceau to confirm protein transfer. Blocking was done in 5% BSA (prepared in 1X Tris Buffered Saline buffer with Tween-20) and blots were incubated with primary antibody overnight at 4^0C and secondary HRP conjugated antibody for an hour at room temperature. Blots were then washed in 1X TBST buffer and developed using Pierce ECL Western blotting substrate (Thermo Fischer Scientific) and luminescence was capture on Chemidoc System (Biorad). Primary antibody for *TK1* (cell signaling) was used at a dilution of 1:1000 and secondary HRP conjugated goat anti-rabbit antibody (Santa Cruz Biotechnologies) was used at 1:2000 dilution.

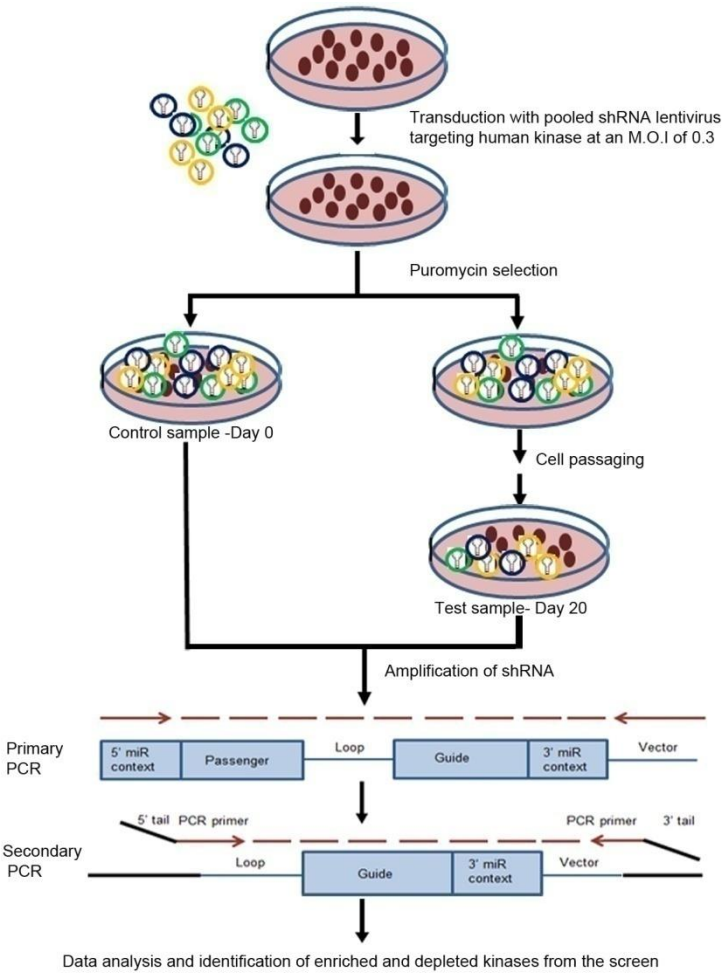
5.3.12 Cell proliferation assay

20,000 cells/ well were seeded in a 24 well plate. Cell growth was assessed at 24 and 96 hours respectively and cells were counted using hemocytometer. Percent cell proliferation was calculated with respect to scramble control cells. The experiments were repeated in triplicates.

5.4 Results

5.4.1 A pooled kinome shRNA screen to identify oncogenic dependency in head and neck cancer cells

In order to identify essential genes in head and neck cancer and contributing, we performed a pooled kinome shRNA screen in a head and neck cancer cell line- AW13516, derived from a tongue cancer patient from India, using 5419 pooled shRNA lentivirus targeting 906 human kinases. About 14 million cells were transduced with lentiviral particles harbouring shRNA against kinases at an M.O.I of 0.3. Following transduction, cells were subjected to puromycin selection (1ug/ml) and half of the cells were harvested at day 3 or 4 post selection, termed as Day 0 sample which served as control and remaining cells were passaged for 20 days in culture and collected as Day 10 and Day 20 respectively. Genomic DNA was extracted, shRNA amplification was performed, and barcode sequences were added by PCR (V-Figure 1). Each sample was tagged with a unique barcode to allow identification of shRNAs belonging to each sample to enable sample multiplexing during sequencing.



V-Figure 1: Schematic representation of pooled shRNA screen in AW13516 cells. AW13516 cells were transduced with pooled shRNA lentivirus targeting 906 human kinases at an M.O.I of 0.3. Cells were selected in puromycin and half of the cells were collected as Day 0 sample which is used as reference or control sample as it represents all the shRNA after transduction. The remaining half cells are passaged up to 20 days in culture and collected as Day 20 sample. Genomic DNA was extracted from both the samples and shRNA sequences are amplified by Primary and Secondary PCR. Then the indexed library, secondary PCR product is subjected to sequencing and shRNA counts was obtained for both samples. shRNA sequences depleted in Day 20 compared to Day 0 were considered. Depleted kinases refer to kinases having oncogenic role in these cells.

Data deconvolution was performed using the edgeR package. Briefly, reads with shRNA sequences were mapped to human kinome library and percent mapping was estimated. Data QC revealed that about 75% reads mapped to kinome reference in AW13516 (V-Table 2). shRNA hairpins with low counts (less than 0.5 counts per million) in Day 0 were excluded from the analysis since the screen was performed at 1000X fold representation. The relative

shRNA abundance was estimated in Day 0, Day 10 and Day 20 samples after performing within and across sample normalization. A list of enriched and depleted shRNA was obtained by comparing Day 20 with respect to control Day 0 sample. For screen 1, a time series analysis of enriched and depleted kinase in Day 10 and Day 20 was done using control as Day 0 sample. Data of screen 2 and screen 3 was used as replicates to identify enriched and depleted shRNA in Day 20 compared to Day 0.Gene- level information was derived for these shRNAs using 'ROAST' module and kinases that were de-regulated were ranked with respect to depletion (appendix 8). Kinases that are lost over the time from the screen have potential role as oncogene since depletion of this kinase by shRNA in cells is inducing a cell death phenotype whereas kinases that get enriched may be acting as tumor-suppressors because knocking down of these kinases tend to promote cell proliferation and therefore, enrichment of shRNA is observed over the time.

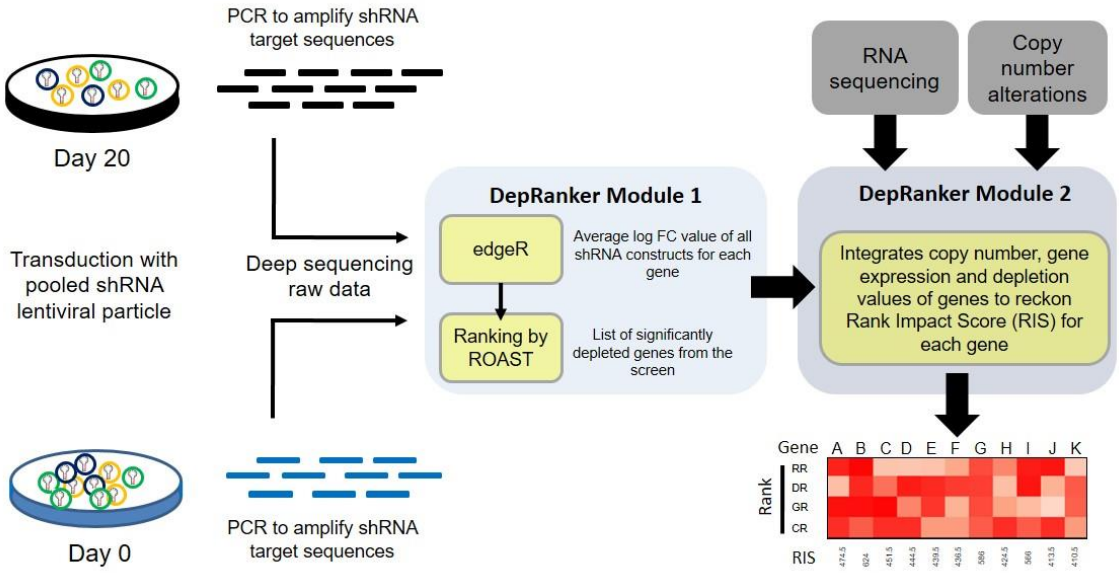
	AW13516 Screen1			**AW13516 Screen2**			**AW13516 Screen3**		
Sample	Day 0	Day 10	Day 20	Day 0	Day 10	Day 20	Day 0	Day 10	Day 20
Total Reads	7306986	24734650	12806948	768023	3116132	1932070	932563	1626831	929280
Total reads mapping to kinome	6885742	23167685	11999717	574936	1940642	1522511	635114	996063	466771
Percent mapping to Kinome	94.23	93.66	93.69	79.18	65.23	81.89	78.65	81.49	78.58

V-Table 2: QC data from sequencing showing percent of the reads mapping to kinome library for all the three samples for each of the three screens of AW13516 cell line.

5.4.2 An integrated scoring system and analytical package DepRanker to rank biologically relevant genes

The GUI based pooled shRNA screen analysis and gene prioritization package, DepRanker was used to rank and identify biologically relevant genes. In total, 127 kinases were

identified in screen1 and 146 kinases in screen 2 that were depleted in AW13516 cells with available gene expression and copy number data (appendix 9). Gene expression and copy number alteration data for all the kinases showing significant depletion in the pooled screen were analyzed for AW13516 cells [219]. Next, we used DepRanker to integrate genomics data such as gene expression, copy number, ranking given by ROAST analysis and average log FC value of all shRNA per gene to calculate Rank Impact Score (RIS) for each of the kinase in the screen (V-Figure 2), as described in the methodology. The result from both the screens was pooled together by considering the mean weight assigned for each kinase as described in the methodology (appendix10). The kinase ranking for both screens is shown in appendix 11.



V-Figure 2: Schematic outline depicting work flow of pooled shRNA data processing and gene prioritization in DepRanker. RR-ROAST Rank, DR: Depletion Rank, GR: Gene expression Rank, CR: copy number alteration Rank, FC: Fold change.

DepRanker ranked *AURKB* and *TK1* as the top genes after combining the results from the two screens, based on assigned weights (V-Figure 3). Due to the non-inclusion of the normal immortalized oral cells, the essential role of *AURKB* and *TK1* in oral cancer cells couldn't be exclusively established based on the screens performed. However, given that *AURKB* and

TK1 are overexpressed along with high copy gain in AW13516 oral cancer cells, the data along with the functional screen suggests their potential oncogenic role in oral cancer (appendix 10).To confirm the reproducibility of results obtained from our bioinformatics analysis, the counts for each shRNA in Day 0, Day 10 and Day 20 samples were validated using the real time PCR for the selected candidate kinases. The shRNA counts targeting kinases *AURKB* and *TK1* were observed to be depleted in Day 10 and Day 20 as compared to Day 0 control sample suggesting that these kinases are conferring oncogenic dependence in head and neck cancer cell line and are essential for cell survival as knockdown of these kinases resulted in the elimination of the corresponding shRNA from the population over the time (data not shown). These results were consistent with our bioinformatics analysis wherein we observed a depletion of shRNA constructs targeting kinase *AURKB* and *TK1* in Day 10 and Day 20 as compared to Day 0 sample. Here, we considered mean CPM (counts per million) counts of each shRNA construct for both genes across all the three screens and percent shRNA counts at each time point are plotted (V-Figure 4). All three shRNA of *AURKB* are showing consistent depletion in Day 10 and Day 20.

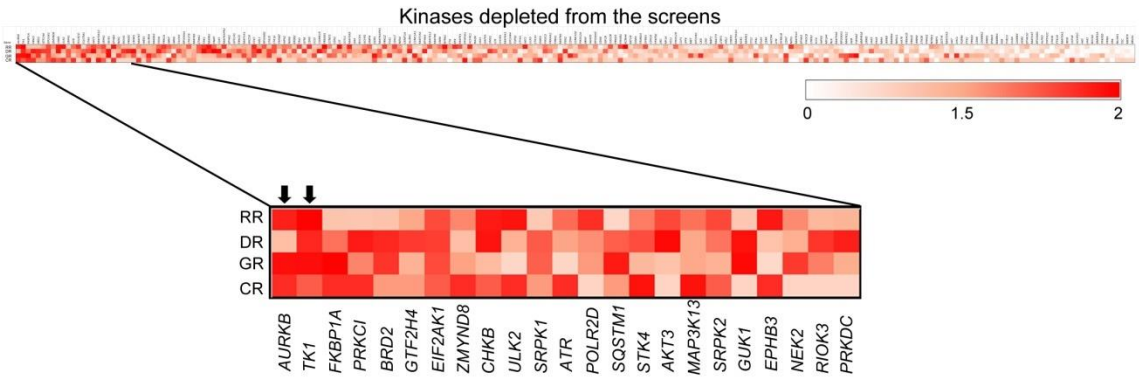


Figure 2: Heatmap representation of depleted kinases in the screen considering overall Rank Impact score (RIS)

V-Figure 3: Heatmap representation of depleted kinases in the screen considering overall Rank Impact Score (RIS). Heatmap representation of kinases depleted from the screen having a high impact score considering ranking assigned by ROAST, gene expression, copy number and average log FC of depleted shRNA for a kinase. The enlarged

view shows top 10 kinase depleted from screen with a high impact score. AURKB and TK1 kinases top the list

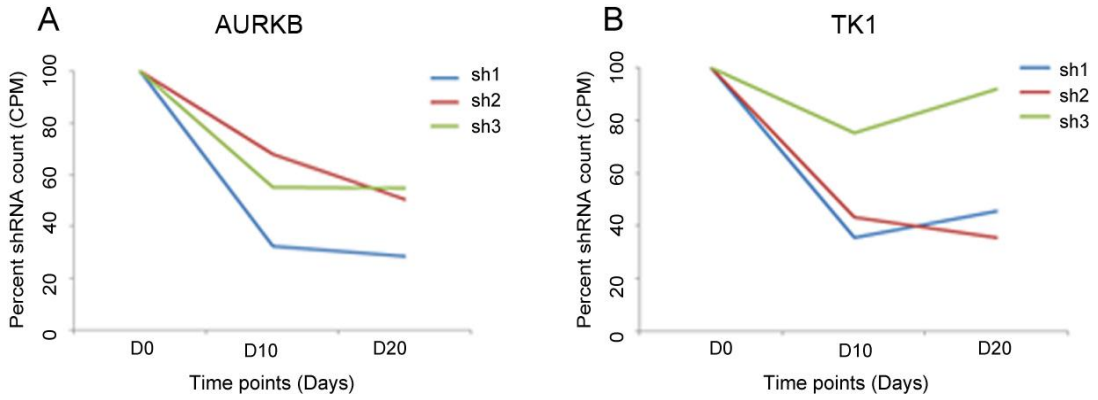


Figure 3: Graph showing percent shRNA count at Day 0, Day 10 and Day 20 for three shRNA of AURKB and TK1

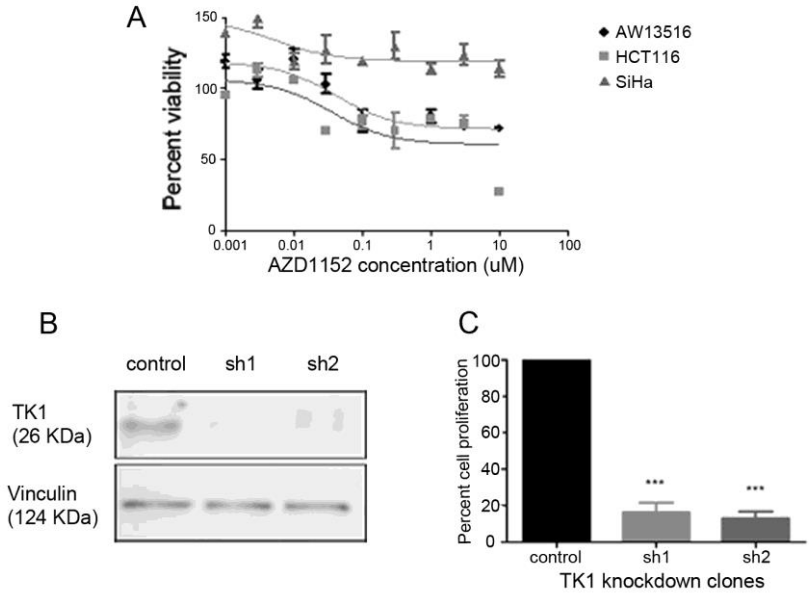
V-Figure 4: Graph showing percent shRNA count of Day 0, Day 10 and Day 20 for three shRNA of AURKB and TK1. Mean CPM counts of each shRNA construct for both genes across all the three screens were obtained and percent shRNA counts at each time point are plotted.

5.4.3 *AURKB* and *TK1* kinases confer oncogenic dependency in AW13516 cells

AURKB is a chromosomal passenger protein which is critical for the accurate segregation of chromosomes accurately during cell division[221]. However, in several cancers over-expression of *AURKB* is often associated with poor prognosis [222]. *AURKB* mediated phosphorylation suppresses the activity of p53 by several mechanism [223, 224]. However, several studies have also reported that the inhibitors of *AURKB* are effective in inhibiting cell growth in p53 mutant cell lines [225, 226]. AW13516 cells harbors p53 mutation p.R273H and p.R72fs*51.

To confirm *AURKB* as a potential oncogenic kinase conferring cell survival of AW13516 cells, we performed a MTT assay on AW13516 cells using AZD1152-HQPA inhibitor. We observed that AW13516 cells were sensitive to the inhibitor with IC50 value as 40 nM. HCT116 colon cells were used as a sensitive cell line for the assay whereas cervical cancer

cells SiHa served as resistant cells (V-Figure 5A). These results suggest that *AURKB* specific inhibitor AZD1152-HQPA could inhibit the cell viability of p53 mutant AW13156 cells. The results are consistent with the sensitivity of this inhibitor to other cells like HT29 having similar p53 mutation p.R273H [225].



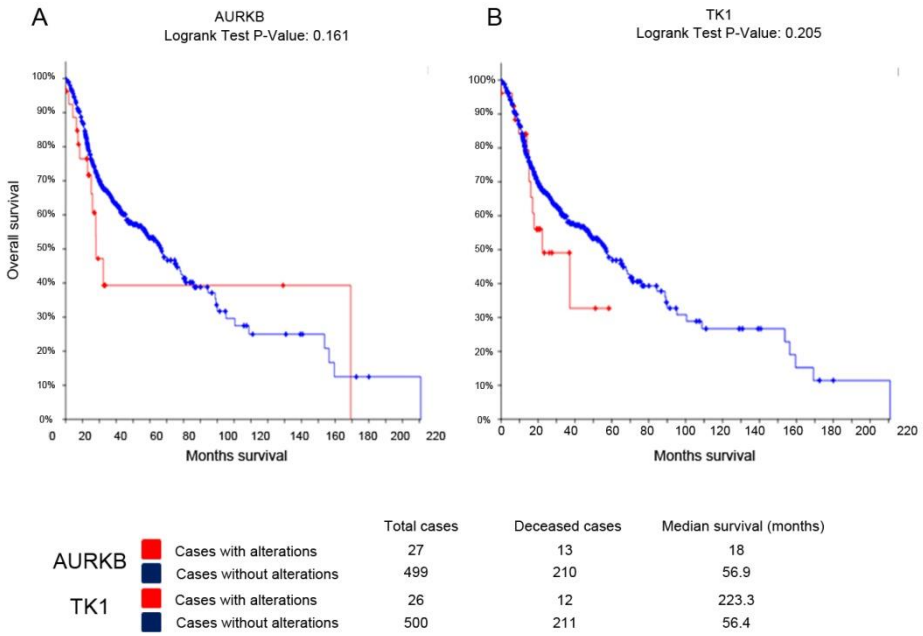
V-Figure 5: AURKB and TK1 show onocgenic dependency in AW13516 cells- A) MTT assay with AZD1152-HQPA inhibitor in AW13156, HCT116 and SiHa cells. B) Knockdown confirmation of TK1 in AW13516 cells by western blotting. C) Cell proliferation assay in control and TK1 knockdown clones of AW13516 cells.

Thymidine kinase 1 (*TK1*) was identified as another potential target from the screen. *TK1* is an enzyme that plays a role in the first step of the biosynthesis of dTTP during DNA synthesis in cells [227]. High expression of *TK1* in cancer tissues is associated with disease progression and poor prognosis [228]. Serum *TK1* levels are used as a prognostic biomarker in several cancers including head and neck cancer to predict the outcome of treatment [229], thus making *TK1* an attractive target. To functionally characterize the role of *TK1*, we performed knockdown of *TK1* in AW13156 cells and confirmed the knockdown by western blotting. We performed cell proliferation assay and observed that the proliferation was

significantly ($p<0.0001$) affected in knockdown clones as compared to scramble control cells (V-Figure 5B and 5C).

5.4.4 Patients with *AURKB* alterations show a poor overall survival

To assess the impact of *AURKB* alterations on the survival of patients, we accessed gene alteration data for *AURKB* and *TK1* from cBioPortal [220]. TCGA provisional HNSCC data sets comprising of mutations, copy number changes and mRNA up-regulation across 528 samples were analyzed. Survival analysis using Kaplan-Meier plots suggests that patients with *AURKB* alteration displayed poor survival of 18 months as compared to survival of 56months in non-altered group (V-Figure 6A). The survival of*TK1* altered and non-altered cohorts were 22 and 56 months respectively (V-Figure 6B), suggesting poor survival in TK1 genetic alteration group.



V-Figure 6: Kaplan-Meier survival analysis of TCGA HNSCC dataset- Survival plots for A) AURKB and B) TK1 gene. Red indicates altered cases and blue refer to non-altered cases.

5.5 Discussion

Pooled shRNA screen is a powerful tool to identify specific gene targets that are essential for the survival of cancer cells. However, heterogeneous data sets often have limited reproducibility as indicated by multiple studies and several approaches are adopted to minimize the noise generated by non-reproducible hits [46]. Other factors that contribute to the variability and complexity of screen data is effective delivery of shRNA, random integration for stable expression of shRNA , processing of shRNA hairpin into silencing complex and off-target effects [50]. Therefore, to overcome these limitations due to variability in the reproducibility of the data, several robust computational approaches have emerged [51, 52]. Some analysis method integratesgenomic data such as gene expression and copy number information to draw insights in predicting cancer essential genes [53, 54].

Although several data integration tools and packages are available to analyze the dataset from the screen, most have their specific third-party needs and necessitate intense computational infrastructure that cannot be run without specialized and advanced computational expertise of the researchers. Thus, a simplified scoring system for a functional biologist to rank genes from the screen data by integrating the genomics data remains a bottleneck. To address this, we have developed a scoring system 'DepRanker' which calculates a Rank Impact Score for each gene identified in the screen considering the gene expression and copy number data.

Two different screens in AW13516 cells have been analyzed by different approaches as well as sequenced at different depth. Because of the major difference in the overall capture of the libraries, we expected the results from the screens to be different. Since neither of the screens was performed at enough saturation we analyzed the data following separate protocols. The candidate genes *AURKB* and *TK1* identified from both the screen by

integrated genomics approach were validated by inhibitor and knockdown assays. DepRanker is an effort in a direction to reduce noise due to differences in capture of libraries, sequencing depth and analysis methods. This approach can be specifically useful in identifying dependencies in cell lines.

AURKB and *TK1* are reported to have oncogenic functions in several cancer types including HNSCC. A previous study on p53 mutant HNSCC cell lines using a kinome screen was also able to identify members of Aurora kinase and Thymidine kinase as therapeutic targets [230], which is consistent with our findings.

AW13516 cells display high copy amplification and gene expression of *AURKB* gene. Overexpression or amplification of *AURKB* has been reported in several cancer types [221, 223]. *AURKB* is a chromosomal protein involved in the segregation of chromosome and cytokinesis [231] and its overexpression leads to aneuploidy in the cells. It is also associated with aggressive tumor progression [232]. There are several pieces of evidence that point towards the oncogenic role of *AURKB* in head and neck cancer. High *AURKB* expression was observed to be associated with increased cell proliferation and lymph node metastasis [233],involved in activation of the RAS-MAPK pathway and contributing to cetuximab resistance [234]. Also, *AURKB* is one of the common essential genes identified from most of pooled RNAi and CRISPR screen on cancer cell lines, as identified by a search for this gene in DepMap portal [53]. We observed that AW13516 cells were sensitive to *AURKB* inhibitor AZD1152-HQPA. In addition, survival analysis of TCGA HNSCC data indicated that patients with alterations in *AURKB* genes display poor overall survival suggesting its role in carcinogenesis in HNSCC. Since, several *AURKB* inhibitors are in clinical trials [221, 235]; *AURKB* can serve as a potential therapeutic target for the treatment of HNSCC.

Similarly, *TK1* target identified in the screen also exhibited high copy gain and increased gene expression in AW13516 cells. Thymidine kinase 1 (*TK1*) has a role in regulating cell cycle [227]. Serum *TK1* levels are used to determine disease prognosis and predict treatment outcome [228]. A study on head and neck cancer shows that patients treated with chemotherapy and surgery showed decreased serum *TK1* levels whereas patients with stable disease displayed elevated *TK1* levels and hence, *TK1* can be used as a biomarker to evaluate disease outcome [229, 236]. We functionally validated another target *TK1* using a knockdown approach. A significant difference in the proliferation rate was observed in *TK1* knockdown clones as compared to control cells suggesting essentiality of *TK1* in the survival of cells. Also, a previous study from our lab identified significant up-regulation of *TK1* expression in tongue tumors [122].

In conclusion, we developed a data integration and scoring system 'DepRanker' which uses the output of shRNA screen analysis packages (like ROAST, RIGER and Chimera) and integrates with other genomics datasets to compute an integration score known as Rank Impact Score (RIS) for each gene. We performed pooled RNAi screen against 906 kinase genes and using the DepRanker, integrated the outcome with gene expression and copy number data for AW13516 cells to identify *AURKB* and *TK1* as essential genes in oral cancer.

Chapter VI

Summary and Conclusion

Chapter VI- Summary and Conclusions

Cancer is a genetic disease defined by several genomic alterations like mutations, gene expression changes, copy number alterations, epigenetic changes and structural variations. However, cancer is driven by only a few genetic alterations termed as driver genes whereas several other alterations that do not contribute to disease progression are termed as passenger alterations. Targeting of driver genes in cancer cells results in decreased cellular proliferation and viability, a phenomenon described as oncogene addiction. This is the basis of targeted therapy or precision medicine, wherein a patient's unique genomic profile is considered for deciding treatment and outcome. Targeted therapy has been successfully implemented in the clinical setting for several cancer types and yielded beneficial results in controlling the disease. Based on the concept of identifying gene targets for precision medicine or targeted therapy, two study approaches drive this thesis- one conceptual and other technical. The first approach focuses on integrated genomic approaches to identify driver alterations from cancer genomes and the second approach deals with functional genomics using pooled RNAi screen to predict therapeutically relevant driver alterations for targeted therapy. The studies were performed in two different cancer types.

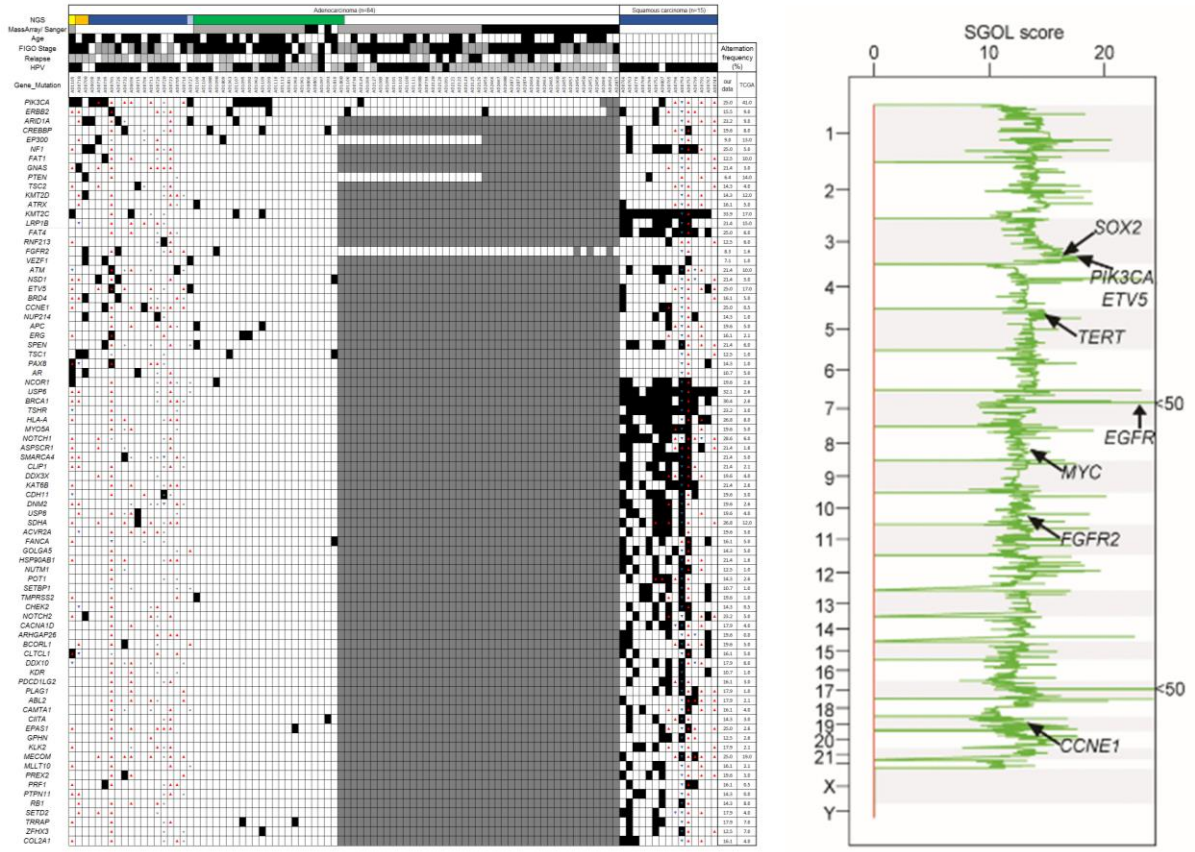
Comprehensive genomics efforts were undertaken to characterize the significantly altered mutations, expressed transcripts and structural variants underlying the cervical cancer genome. Several known and other cancer-associated genes were identified in this study.

Firstly, extensive genomic profiling for mutations was performed in 84 samples of cervical adenocarcinoma and 15 samples of squamous carcinoma using NGS approach and other genotyping methods to provide a landscape of somatic mutations in cervical cancer from the Indian population. Here, we report mutations in known hallmark gene - *PIK3CA*, *ERBB2*, *ARID1A*, *CREBBP*, *EP300*, *NF1*, *FAT1*, *PTEN* and *TSC2* and novel cancer-associated genes

FGFR2 and *AKT1*. In addition, mutations in epigenetic gene include *KMT2C*, *KMT2D*, *EP300*, *BRD3*, *BRD4*, *NSD1* and *PBRM1*. However, we did not observe mutations in *KRAS*, *STK11*, *FBXW7* and *TP53* genes, which are commonly mutated in cervical cancers as reported by TCGA group [94].

Secondly, copy number variation analysis from WES and WGS samples show recurrent copy gains in genes in *PIK3CA* (37%), *SOX2* (37%), *TERT* (33%), *ERBB2* (30%), *KRAS* (26%), *MYC* (22%) and *BRCA1* (22%), consistent with literature reports. Amplifications are also observed in other cancer-associated tyrosine kinases like *ERBB3* (22%), *ERBB4* (15%), *EGFR* (15%), *FGFR2* (15%), *FGFR3* (7%). In addition, we note copy gain and loss in known oncogenic fusion gene partners *FGFR3-TACC3*, *TMPRSS2-ERG* and *EML4-ALK*, also observed in the TCGA dataset. From WGS dataset, 14 broad-arm level amplifications, 5 broad arm deletions, 221 focal amplification and 31 focal deletions were predicted. Recurrent amplifications are observed at chromosome 1q, 3q, 8q, 11p, 17q, 19q, 20q, 5p, 9q, 1p, 11q, 20p and 9p and recurrent deletions at chromosome 3p, 4q, 11p, 11q, 18q, 19p, 2q and 5q.

The integrated mutation and copy number alterations in cervical cancer hallmark genes and other cancer genes along with CNV plot is shown in the figure below. Black box indicates mutation, red and blue triangle indicates copy gain and loss respectively.



Third, gene expression analysis was performed within tumor samples. Gene expressed in top 10% quartile and recurrent in at least 30% of the samples were considered further. We observe over-expression of *EGFR* (57%), *ERBB2* (81%), *ERBB3* (90%), *MET* (38%), *AKT1* (38%) and *AKT2* (90%). Increased expression *MMP2*, *MMP12* and *MMP14* has been reported in cervical cancer [179-181] and also observed in our dataset. Next, we identified expressed gene fusions with one of the genes with oncogenic function - *IDH3G-PPP2R1A*, *U2AF1-CASP2*, *RAP2A-MECOM*, *PPP6C-CASC3* and *ANKRD27-MYC* from fusion analysis. In addition, we report in-frame fusions with kinase gene partner such as *PKM-FUT2*, *PKM-CBX4*, *STK24-ZNF585A* and *CDK16-CAP1* with conserved domains. All the fusions observed are novel and not reported in the literature.

Fourth, we report several structural variants identified from WGS data. *ARHGAP11B-ARHGAP11A* and *CDK11B-SLC35E2B* were recurrent in two samples. Two samples show

structural re-arrangements *CAPN8-MUC3B* and *PNKD-MUC3B* have MUC3B as one of the gene partners. Gene translocation pairs *FAM172A-**ESR, DUX4**-ROCK1P1**, **MLLT4**-KIF25**, **MSH2**-TAF4**, **PAX7**-EEF1A2**, **PBX1**-SIK3* and ***PDGFRA**-MAN2A1* involves one of the partners known to be a cancer gene. All the genomic rearrangements reported from the study are unique and not reported in the literature for any cancer type.

Overall from the genomic studies, we identify several therapeutically relevant alterations in cervical cancer. We observe that most of the mutations in genes converge onto PI3K/AKT and MAPK pathway. Recurrent mutations of *PIK3CA* in the helical domain E545K and E542K are targetable using alpelisib and fulvestrant [144], *ERBB2* D769Y, S310F/Y by trastuzumab, lapatinib or neratinib [145], *FGFR2* K659E, S320F and C382R by Ponatinib and BGJ398 [137, 147].

Most of the mutations, amplification and over-expression of genes were common among the ERBB family members. Therefore, the role of ERBB signalling in cervical cancer was investigated using *in-vitro* and *in-vivo* approaches. Cervical cells subjected to Afatinib treatment revealed that C33A cells were sensitive to treatment as compared to other cells. This observation was consistent in the *in-vivo* studies, wherein mice with C33A tumors showed a delay in tumor growth on Afatinib treatment as compared to control group. Next, to identify Afatinib targets- EGFR, ERBB2 and ERBB4 conferring oncogenic dependency in C33A cells, individual gene knockdown by shRNA was performed in C33A and SiHa cells. Although a slight decrease in the p-MAPK was observed upon depletion of EGFR and ERBB2, the knockdown cells did not show reduced cellular proliferation, migration and anchorage-independent growth suggesting that cells are not dependent on ERBB2 or EGFR for growth and survival. These results indicate that there is a possible role of change in receptor heterodimerization upon depletion of one ERBB member or cross-talk with other pathways such as PI3K/AKT [113, 114], which is facilitating the continuation of signalling

in the cells. This needs to be further investigated. Our results also point to the fact that it is essential to target all ERBB receptors simultaneously as done by afatinib inhibitor to reduce cell growth since redundant gene functions are carried out by other receptors upon inhibition of one receptor. These findings are consistent with earlier reports in cervical cancer which shows that pan-ERBB inhibition by inhibitors Lapatinib and AST1306 display effectiveness in reducing proliferation in C33A cervical cells [110].

This study overall validates the current understanding of cervical cancer genomics and also extends our understanding of cervical cancer, especially the adenocarcinoma subtype and provides a detailed comprehensive landscape of somatic alterations from the Indian ethnicity for the first time to identify suitable molecular targets for precision medicine.

In addition, we have taken a complementary approach to establish the significance of a functional genomics approach to identify therapeutically relevant driver alterations using cells derived from HNSCC as a model system. We performed a pooled RNAi shRNA screen against human kinases in HNSCC cell line AW13516. To predict potential driver genes with high confidence, the RNAi screen data was integrated with genomics data of gene expression and copy number changes. Such an approach has been previously used by several groups. However, currently available data integration tools which combine RNAi data with genomics data, require intense computational processing and expertise and hence, of restricted use to a functional biologist. Here, we develop a simplified scoring system 'DepRanker' which integrates genomic data like gene expression, copy number and RNAi output data like depleted gene list and individual shRNA depletion list to assign scores for calculating a Rank Impact Score (RIS) [237]. Genes with high RIS are predicted to be potential cancer drivers. An input of RNAi data along with genomics data fed to DepRanker was able to predict AURKB and TK1 as drivers. To validate findings, TK1 knockdown was performed in AW13516 cells and it was observed that cell proliferation was inhibited in

knockdown clones as compared to control cells. In addition, AW13516 cells also exhibited sensitivity to AURKB inhibitor AZD1152-HQPA. Both the genes play an important role in regulating cell cycle and cell division and can act as attractive therapeutic targets for targeted therapy in clinics for head and neck cancer type.

DepRanker has a wider application in predicting cancer essential genes for other RNAi and CRISPR screen datasets as well. We provide a user-friendly GUI which can be used by a functional biologist by providing input data in the required format to identify genes showing oncogenic dependency in cancer cells.

Although we performed a pooled shRNA screen in AW13516 cells and used DepRanker to predict cancer essential genes by integrating genomics data, this study suffers from several limitations. The pooled screen was restricted to human kinases, therefore other non-kinase driver genes remain undetected. The screen was performed in triplicates, of which data from two screens were captured at lower coverage. With recent advances, pooled CRISPR screens offer better advantage than pooled RNAi screens and it is more specific and sensitive in predicting cancer essential genes [238]. With CRISPR screens, fewer variations are observed across the replicates and complete gene function perturbation is seen due to knockout [239]. Nevertheless, this study presents a proof-of-principle approach for validation of functional genomics using pooled RNAi screen against human kinases to identify therapeutic gene targets. We identified *AURKB* and *TK1* as gene targets with therapeutic relevance in the treatment of head and neck cancer patients. In addition, we present a simplified scoring system 'DepRanker' which can be readily used by a functional biologist to analyze pooled screen data and obtain useful insights in predicting essentiality genes in cancer cells [237].

Chapter VII

Bibliography

Chapter VII- Bibliography

1. Wang, Q., *Cancer predisposition genes: molecular mechanisms and clinical impact on personalized cancer care: examples of Lynch and HBOC syndromes.* Acta Pharmacol Sin, 2016. **37**(2): p. 143-9.
2. Imran, A., et al., *Role of Molecular Biology in Cancer Treatment: A Review Article.* Iran J Public Health, 2017. **46**(11): p. 1475-1485.
3. Lee, E.Y. and W.J. Muller, *Oncogenes and tumor suppressor genes.* Cold Spring Harb Perspect Biol, 2010. **2**(10): p. a003236.
4. Wodarz, D., A.C. Newell, and N.L. Komarova, *Passenger mutations can accelerate tumour suppressor gene inactivation in cancer evolution.* J R Soc Interface, 2018. **15**(143).
5. Stratton, M.R., P.J. Campbell, and P.A. Futreal, *The cancer genome.* Nature, 2009. **458**(7239): p. 719-24.
6. Pon, J.R. and M.A. Marra, *Driver and passenger mutations in cancer.* Annu Rev Pathol, 2015. **10**: p. 25-50.
7. McFarland, C.D., L.A. Mirny, and K.S. Korolev, *Tug-of-war between driver and passenger mutations in cancer and other adaptive processes.* Proc Natl Acad Sci U S A, 2014. **111**(42): p. 15138-43.
8. McFarland, C.D., et al., *The Damaging Effect of Passenger Mutations on Cancer Progression.* Cancer Res, 2017. **77**(18): p. 4763-4772.
9. Falzone, L., S. Salomone, and M. Libra, *Evolution of Cancer Pharmacological Treatments at the Turn of the Third Millennium.* Front Pharmacol, 2018. **9**: p. 1300.
10. Krzyszczyk, P., et al., *The growing role of precision and personalized medicine for cancer treatment.* Technology (Singap World Sci), 2018. **6**(3-4): p. 79-100.
11. Burney, I.A. and R. Lakhtakia, *Precision Medicine: Where have we reached and where are we headed?* Sultan Qaboos Univ Med J, 2017. **17**(3): p. e255-e258.
12. Ginsburg, G.S. and K.A. Phillips, *Precision Medicine: From Science To Value.* Health Aff (Millwood), 2018. **37**(5): p. 694-701.
13. Padma, V.V., *An overview of targeted cancer therapy.* Biomedicine (Taipei), 2015. **5**(4): p. 19.
14. Lee, Y.T., Y.J. Tan, and C.E. Oon, *Molecular targeted therapy: Treating cancer with specificity.* Eur J Pharmacol, 2018. **834**: p. 188-196.
15. McCutcheon, J.N. and G. Giaccone, *Next-Generation Sequencing: Targeting Targeted Therapies.* Clin Cancer Res, 2015. **21**(16): p. 3584-5.
16. Del Bufalo, F., et al., *BRAF V600E Inhibitor (Vemurafenib) for BRAF V600E Mutated Low Grade Gliomas.* Front Oncol, 2018. **8**: p. 526.
17. Chan, B.A. and B.G. Hughes, *Targeted therapy for non-small cell lung cancer: current standards and the promise of the future.* Transl Lung Cancer Res, 2015. **4**(1): p. 36-54.
18. Schwaederle, M., et al., *Association of Biomarker-Based Treatment Strategies With Response Rates and Progression-Free Survival in Refractory Malignant Neoplasms: A Meta-analysis.* JAMA Oncol, 2016. **2**(11): p. 1452-1459.
19. Sun, J., et al., *A systematic analysis of FDA-approved anticancer drugs.* BMC Syst Biol, 2017. **11**(Suppl 5): p. 87.
20. Opdam, F.L., et al., *Lapatinib for advanced or metastatic breast cancer.* Oncologist, 2012. **17**(4): p. 536-42.
21. Seebacher, N.A., et al., *Clinical development of targeted and immune based anti-cancer therapies.* J Exp Clin Cancer Res, 2019. **38**(1): p. 156.
22. Ledermann, J.A., *PARP inhibitors in ovarian cancer.* Ann Oncol, 2016. **27 Suppl 1**: p. i40-i44.
23. Tewari, K.S., et al., *Improved survival with bevacizumab in advanced cervical cancer.* N Engl J Med, 2014. **370**(8): p. 734-43.
24. Weinstein, I.B. and A. Joe, *Oncogene addiction.* Cancer Res, 2008. **68**(9): p. 3077-80; discussion 3080.

25. Sharma, S.V. and J. Settleman, *Oncogene addiction: setting the stage for molecularly targeted cancer therapy.* Genes Dev, 2007. **21**(24): p. 3214-31.
26. Rosa, D.D., et al., *Molecular-targeted therapies: lessons from years of clinical development.* Cancer Treat Rev, 2008. **34**(1): p. 61-80.
27. Lee, J.K. and S.J. Priceman, *Precision Medicine-Enabled Cancer Immunotherapy.* Cancer Treat Res, 2019. **178**: p. 189-205.
28. Rosenberg, S.A., et al., *Adoptive cell transfer: a clinical path to effective cancer immunotherapy.* Nat Rev Cancer, 2008. **8**(4): p. 299-308.
29. Cruz-Ramos, M. and J. Garcia-Foncillas, *CAR-T cell and Personalized Medicine.* Adv Exp Med Biol, 2019. **1168**: p. 131-145.
30. Majewski, J., et al., *What can exome sequencing do for you?* J Med Genet, 2011. **48**(9): p. 580-9.
31. Meienberg, J., et al., *Clinical sequencing: is WGS the better WES?* Hum Genet, 2016. **135**(3): p. 359-62.
32. Nakagawa, H. and M. Fujita, *Whole genome sequencing analysis for cancer genomics and precision medicine.* Cancer Sci, 2018. **109**(3): p. 513-522.
33. Heyer, E.E., et al., *Diagnosis of fusion genes using targeted RNA sequencing.* Nat Commun, 2019. **10**(1): p. 1388.
34. Piskol, R., G. Ramaswami, and J.B. Li, *Reliable identification of genomic variants from RNA-seq data.* Am J Hum Genet, 2013. **93**(4): p. 641-51.
35. Ku, C.S., et al., *Exome versus transcriptome sequencing in identifying coding region variants.* Expert Rev Mol Diagn, 2012. **12**(3): p. 241-51.
36. Prodduturi, N., et al., *Indel sensitive and comprehensive variant/mutation detection from RNA sequencing data for precision medicine.* BMC Med Genomics, 2018. **11**(Suppl 3): p. 67.
37. Rusch, M., et al., *Clinical cancer genomic profiling by three-platform sequencing of whole genome, whole exome and transcriptome.* Nat Commun, 2018. **9**(1): p. 3962.
38. Tomczak, K., P. Czerwinska, and M. Wiznerowicz, *The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.* Contemp Oncol (Pozn), 2015. **19**(1A): p. A68-77.
39. Berger, M.F. and E.R. Mardis, *The emerging clinical relevance of genomics in cancer medicine.* Nat Rev Clin Oncol, 2018. **15**(6): p. 353-365.
40. Malone, E.R., et al., *Molecular profiling for precision cancer therapies.* Genome Med, 2020. **12**(1): p. 8.
41. Bunnik, E.M. and K.G. Le Roch, *An Introduction to Functional Genomics and Systems Biology.* Adv Wound Care (New Rochelle), 2013. **2**(9): p. 490-498.
42. O'Loughlin, T.A. and L.A. Gilbert, *Functional Genomics for Cancer Research: Applications In Vivo and In Vitro.* 2019. **3**(1): p. 345-363.
43. So, R.W.L., et al., *Application of CRISPR genetic screens to investigate neurological diseases.* Mol Neurodegener, 2019. **14**(1): p. 41.
44. Boettcher, M. and J.D. Hoheisel, *Pooled RNAi Screens - Technical and Biological Aspects.* Curr Genomics, 2010. **11**(3): p. 162-7.
45. Burgess, D.J., *Functional genomics: Shining a light on genetic screen strategies.* Nat Rev Genet, 2018. **19**(1): p. 6-7.
46. Schaefer, C., et al., *Target discovery screens using pooled shRNA libraries and next-generation sequencing: A model workflow and analytical algorithm.* PLoS One, 2018. **13**(1): p. e0191570.
47. Cowley, G.S., et al., *Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies.* Sci Data, 2014. **1**: p. 140035.
48. Kiessling, M.K., et al., *Identification of oncogenic driver mutations by genome-wide CRISPR-Cas9 dropout screening.* BMC Genomics, 2016. **17**(1): p. 723.
49. Wei, L., et al., *Genome-wide CRISPR/Cas9 library screening identified PHGDH as a critical driver for Sorafenib resistance in HCC.* Nat Commun, 2019. **10**(1): p. 4681.

50. Fellmann, C. and S.W. Lowe, *Stable RNA interference rules for silencing.* Nat Cell Biol, 2014. **16**(1): p. 10-8.
51. Dempster, J.M., et al., *Agreement between two large pan-cancer CRISPR-Cas9 gene dependency data sets.* Nat Commun, 2019. **10**(1): p. 5817.
52. Shao, D.D., et al., *ATARiS: computational quantification of gene suppression phenotypes from multisample RNAi screens.* Genome Res, 2013. **23**(4): p. 665-78.
53. Tsherniak, A., et al., *Defining a Cancer Dependency Map.* Cell, 2017. **170**(3): p. 564-576 e16.
54. Guan, Y., et al., *Prioritizing predictive biomarkers for gene essentiality in cancer cells with mRNA expression data and DNA copy number profile.* Bioinformatics, 2018. **34**(23): p. 3975-3982.
55. Bobdey, S., et al., *Burden of cervical cancer and role of screening in India.* Indian J Med Paediatr Oncol, 2016. **37**(4): p. 278-285.
56. Peiretti, M., et al., *Management of recurrent cervical cancer: a review of the literature.* Surg Oncol, 2012. **21**(2): p. e59-66.
57. Li, H., X. Wu, and X. Cheng, *Advances in diagnosis and treatment of metastatic cervical cancer.* J Gynecol Oncol, 2016. **27**(4): p. e43.
58. Liu, X., et al., *Predictors of Distant Metastasis in Patients with Cervical Cancer Treated with Definitive Radiotherapy.* J Cancer, 2019. **10**(17): p. 3967-3974.
59. Berrington de Gonzalez, A., S. Sweetland, and J. Green, *Comparison of risk factors for squamous cell and adenocarcinomas of the cervix: a meta-analysis.* Br J Cancer, 2004. **90**(9): p. 1787-91.
60. Colombo, N., et al., *Cervical cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up.* Ann Oncol, 2012. **23 Suppl 7**: p. vii27-32.
61. Yokoi, E., et al., *Impact of histological subtype on survival in patients with locally advanced cervical cancer that were treated with definitive radiotherapy: adenocarcinoma/adenosquamous carcinoma versus squamous cell carcinoma.* J Gynecol Oncol, 2017. **28**(2): p. e19.
62. Rose, P.G., *Are the differences in treatment outcome for adenocarcinoma of the cervix different enough to change the treatment paradigm?* Gynecol Oncol, 2012. **125**(2): p. 285-6.
63. Castellsague, X. and N. Munoz, *Chapter 3: Cofactors in human papillomavirus carcinogenesis--role of parity, oral contraceptives, and tobacco smoking.* J Natl Cancer Inst Monogr, 2003(31): p. 20-8.
64. Fonseca-Moutinho, J.A., *Smoking and cervical cancer.* ISRN Obstet Gynecol, 2011. **2011**: p. 847684.
65. Zhao, H., et al., *Concurrent paclitaxel/cisplatin chemoradiotherapy with or without consolidation chemotherapy in high-risk early-stage cervical cancer patients following radical hysterectomy: preliminary results of a phase III randomized study.* Oncotarget, 2016. **7**(43): p. 70969-70978.
66. Hu, K., et al., *Comparison of treatment outcomes between squamous cell carcinoma and adenocarcinoma of cervix after definitive radiotherapy or concurrent chemoradiotherapy.* Radiat Oncol, 2018. **13**(1): p. 249.
67. Lee, J.Y., et al., *Prognosis of Cervical Cancer in the Era of Concurrent Chemoradiation from National Database in Korea: A Comparison between Squamous Cell Carcinoma and Adenocarcinoma.* PLoS One, 2015. **10**(12): p. e0144887.
68. Arbyn, M., et al., *Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis.* Lancet Glob Health, 2020. **8**(2): p. e191-e203.
69. Srivastava, A.N., et al., *Cervical cancer screening in rural India: Status & current concepts.* Indian J Med Res, 2018. **148**(6): p. 687-696.
70. Sreedevi, A., R. Javed, and A. Dinesh, *Epidemiology of cervical cancer with special focus on India.* Int J Womens Health, 2015. **7**: p. 405-14.

71. Haverkos, H.W., *Multifactorial etiology of cervical cancer: a hypothesis.* MedGenMed, 2005. **7**(4): p. 57.
72. Bosch, F.X., et al., *Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. International biological study on cervical cancer (IBSCC) Study Group.* J Natl Cancer Inst, 1995. **87**(11): p. 796-802.
73. Burd, E.M., *Human papillomavirus and cervical cancer.* Clin Microbiol Rev, 2003. **16**(1): p. 1-17.
74. Werness, B.A., A.J. Levine, and P.M. Howley, *Association of human papillomavirus types 16 and 18 E6 proteins with p53.* Science, 1990. **248**(4951): p. 76-9.
75. Dyson, N., et al., *Homologous sequences in adenovirus E1A and human papillomavirus E7 proteins mediate interaction with the same set of cellular proteins.* J Virol, 1992. **66**(12): p. 6893-902.
76. Gao, G. and D.I. Smith, *Human Papillomavirus and the Development of Different Cancers.* Cytogenet Genome Res, 2016. **150**(3-4): p. 185-193.
77. Monie, A., et al., *Cervarix: a vaccine for the prevention of HPV 16, 18-associated cervical cancer.* Biologics, 2008. **2**(1): p. 97-105.
78. Kaarthigeyan, K., *Cervical cancer in India and HPV vaccination.* Indian J Med Paediatr Oncol, 2012. **33**(1): p. 7-12.
79. International Collaboration of Epidemiological Studies of Cervical, C., *Comparison of risk factors for invasive squamous cell carcinoma and adenocarcinoma of the cervix: collaborative reanalysis of individual data on 8,097 women with squamous cell carcinoma and 1,374 women with adenocarcinoma from 12 epidemiological studies.* Int J Cancer, 2007. **120**(4): p. 885-91.
80. Vaccarella, S., et al., *Smoking and human papillomavirus infection: pooled analysis of the International Agency for Research on Cancer HPV Prevalence Surveys.* Int J Epidemiol, 2008. **37**(3): p. 536-46.
81. Fang, J.H., et al., *Effect of smoking on high-grade cervical cancer in women on the basis of human papillomavirus infection studies.* J Cancer Res Ther, 2018. **14**(Supplement): p. S184-S189.
82. Roura, E., et al., *Smoking as a major risk factor for cervical cancer and pre-cancer: results from the EPIC cohort.* Int J Cancer, 2014. **135**(2): p. 453-66.
83. Castle, P.E., *How does tobacco smoke contribute to cervical carcinogenesis?* J Virol, 2008. **82**(12): p. 6084-5; author reply 6085-6.
84. !!! INVALID CITATION !!! {Srivastava, 2018 #78;Castellsague, 2003 #77}.
85. Yang, E.J., et al., *Microanatomy of the cervical and anorectal squamocolumnar junctions: a proposed model for anatomical differences in HPV-related cancer risk.* Mod Pathol, 2015. **28**(7): p. 994-1000.
86. Balasubramaniam, S.D., et al., *Key Molecular Events in Cervical Cancer Development.* Medicina (Kaunas), 2019. **55**(7).
87. Vink, M.A., et al., *Clinical progression of high-grade cervical intraepithelial neoplasia: estimating the time to preclinical cervical cancer from doubly censored national registry data.* Am J Epidemiol, 2013. **178**(7): p. 1161-9.
88. Nishimura, A., et al., *Mechanisms of human papillomavirus E2-mediated repression of viral oncogene expression and cervical cancer cell growth inhibition.* J Virol, 2000. **74**(8): p. 3752-60.
89. Chan, C.K., et al., *Human Papillomavirus Infection and Cervical Cancer: Epidemiology, Screening, and Vaccination-Review of Current Perspectives.* J Oncol, 2019. **2019**: p. 3257939.
90. Yim, E.K. and J.S. Park, *The role of HPV E6 and E7 oncoproteins in HPV-associated cervical carcinogenesis.* Cancer Res Treat, 2005. **37**(6): p. 319-24.
91. Senapati, R., N.N. Senapati, and B. Dwibedi, *Molecular mechanisms of HPV mediated neoplastic progression.* Infect Agent Cancer, 2016. **11**: p. 59.

92. Bhatla, N., et al., *Revised FIGO staging for carcinoma of the cervix uteri.* Int J Gynaecol Obstet, 2019. **145**(1): p. 129-135.
93. Petignat, P. and M. Roy, *Diagnosis and management of cervical cancer.* BMJ, 2007. **335**(7623): p. 765-8.
94. Cancer Genome Atlas Research, N., et al., *Integrated genomic and molecular characterization of cervical cancer.* Nature, 2017. **543**(7645): p. 378-384.
95. Ojesina, A.I., et al., *Landscape of genomic alterations in cervical carcinomas.* Nature, 2014. **506**(7488): p. 371-5.
96. Chung, T.K., et al., *Genomic aberrations in cervical adenocarcinomas in Hong Kong Chinese women.* Int J Cancer, 2015. **137**(4): p. 776-83.
97. Xiang, L., et al., *Comprehensive analysis of targetable oncogenic mutations in chinese cervical cancers.* Oncotarget, 2015. **6**(7): p. 4968-75.
98. Wright, A.A., et al., *Oncogenic mutations in cervical cancer: genomic differences between adenocarcinomas and squamous cell carcinomas of the cervix.* Cancer, 2013. **119**(21): p. 3776-83.
99. Lou, H., et al., *Genome Analysis of Latin American Cervical Cancer: Frequent Activation of the PIK3CA Pathway.* Clin Cancer Res, 2015. **21**(23): p. 5360-70.
100. Das, P., et al., *Somatic Variations in Cervical Cancers in Indian Patients.* PLoS One, 2016. **11**(11): p. e0165878.
101. Hsu, J.L. and M.C. Hung, *The role of HER2, EGFR, and other receptor tyrosine kinases in breast cancer.* Cancer Metastasis Rev, 2016. **35**(4): p. 575-588.
102. Oda, K., et al., *A comprehensive pathway map of epidermal growth factor receptor signaling.* Mol Syst Biol, 2005. **1**: p. 2005 0010.
103. Chen, W.W., et al., *Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data.* Mol Syst Biol, 2009. **5**: p. 239.
104. Hynes, N.E. and G. MacDonald, *ErbB receptors and signaling pathways in cancer.* Curr Opin Cell Biol, 2009. **21**(2): p. 177-84.
105. Cao, Y., et al., *Effectiveness and safety of osimertinib in patients with metastatic EGFR T790M-positive NSCLC: An observational real-world study.* PLoS One, 2019. **14**(8): p. e0221575.
106. Mishra, R., A.B. Hanker, and J.T. Garrett, *Genomic alterations of ERBB receptors in cancer: clinical implications.* Oncotarget, 2017. **8**(69): p. 114371-114392.
107. Sircoulomb, F., et al., *Genome profiling of ERBB2-amplified breast cancers.* BMC Cancer, 2010. **10**: p. 539.
108. Iida, K., et al., *EGFR gene amplification is related to adverse clinical outcomes in cervical squamous cell carcinoma, making the EGFR pathway a novel therapeutic target.* Br J Cancer, 2011. **105**(3): p. 420-7.
109. Oh, D.Y., et al., *HER2 as a novel therapeutic target for cervical cancer.* Oncotarget, 2015. **6**(34): p. 36219-30.
110. Martinho, O., et al., *HER Family Receptors are Important Theranostic Biomarkers for Cervical Cancer: Blocking Glucose Metabolism Enhances the Therapeutic Effect of HER Inhibitors.* Theranostics, 2017. **7**(3): p. 717-732.
111. Citri, A. and Y. Yarden, *EGF-ERBB signalling: towards the systems level.* Nat Rev Mol Cell Biol, 2006. **7**(7): p. 505-16.
112. Hu, Y.P., et al., *Reorganization of ErbB family and cell survival signaling after Knock-down of ErbB2 in colon cancer cells.* J Biol Chem, 2005. **280**(29): p. 27383-92.
113. Vermeer, P.D., et al., *Targeting ERBB receptors shifts their partners and triggers persistent ERK signaling through a novel ERBB/EFNB1 complex.* Cancer Res, 2013. **73**(18): p. 5787-97.
114. Lakshmanan, I., et al., *Novel HER3/MUC4 oncogenic signaling aggravates the tumorigenic phenotypes of pancreatic cancer cells.* Oncotarget, 2015. **6**(25): p. 21085-99.

115. Yamaguchi, H., et al., *Signaling cross-talk in the resistance to HER family receptor targeted therapy.* Oncogene, 2014. **33**(9): p. 1073-81.
116. Ma, J., et al., *Targeting of erbB3 receptor to overcome resistance in cancer treatment.* Mol Cancer, 2014. **13**: p. 105.
117. Britten, C.D., *Targeting ErbB receptor signaling: a pan-ErbB approach to cancer.* Mol Cancer Ther, 2004. **3**(10): p. 1335-42.
118. Warr, A., et al., *Exome Sequencing: Current and Future Perspectives.* G3 (Bethesda), 2015. **5**(8): p. 1543-50.
119. Waldmann, A., N. Eisemann, and A. Katalinic, *Epidemiology of Malignant Cervical, Corpus Uteri and Ovarian Tumours - Current Data and Epidemiological Trends.* Geburtshilfe Frauenheilkd, 2013. **73**(2): p. 123-129.
120. Consortium, I.T.P.-C.A.o.W.G., *Pan-cancer analysis of whole genomes.* Nature, 2020. **578**(7793): p. 82-93.
121. Upadhyay, P., et al., *TMC-SNPdb: an Indian germline variant database derived from whole exome sequences.* Database (Oxford), 2016. **2016**.
122. Upadhyay, P., et al., *Genomic characterization of tobacco/nut chewing HPV-negative early stage tongue tumors identify MMP10 asa candidate to predict metastases.* Oral Oncol, 2017. **73**: p. 56-64.
123. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.
124. Iyer, P., et al., *ERBB2 and KRAS alterations mediate response to EGFR inhibitors in early stage gallbladder cancer.* Int J Cancer, 2019. **144**(8): p. 2008-2019.
125. Chandrani, P., et al., *NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome.* Br J Cancer, 2015. **112**(12): p. 1958-65.
126. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.
127. Cibulskis, K., et al., *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.* Nat Biotechnol, 2013. **31**(3): p. 213-9.
128. Tokheim, C.J., et al., *Evaluating the evaluation of cancer driver genes.* Proc Natl Acad Sci U S A, 2016. **113**(50): p. 14330-14335.
129. Martincorena, I. and P.J. Campbell, *Somatic mutation in cancer and normal cells.* Science, 2015. **349**(6255): p. 1483-9.
130. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-21.
131. Roberts, S.A., et al., *An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers.* Nat Genet, 2013. **45**(9): p. 970-6.
132. Huang, J., et al., *Comprehensive genomic variation profiling of cervical intraepithelial neoplasia and cervical cancer identifies potential targets for cervical cancer early warning.* J Med Genet, 2019. **56**(3): p. 186-194.
133. Kuong, K.J. and L.A. Loeb, *APOBEC3B mutagenesis in cancer.* Nat Genet, 2013. **45**(9): p. 964-5.
134. Pflueger, D., et al., *N-myc downstream regulated gene 1 (NDRG1) is fused to ERG in prostate cancer.* Neoplasia, 2009. **11**(8): p. 804-11.
135. Yu, Y., et al., *Targeting AKT1-E17K and the PI3K/AKT Pathway with an Allosteric AKT Inhibitor, ARQ 092.* PLoS One, 2015. **10**(10): p. e0140479.
136. Porta, C., C. Paglino, and A. Mosca, *Targeting PI3K/Akt/mTOR Signaling in Cancer.* Front Oncol, 2014. **4**: p. 64.
137. Liao, R.G., et al., *Inhibitor-sensitive FGFR2 and FGFR3 mutations in lung squamous cell carcinoma.* Cancer Res, 2013. **73**(16): p. 5195-205.
138. Li, Y., et al., *Activation of FGF receptors by mutations in the transmembrane domain.* Oncogene, 1997. **14**(12): p. 1397-406.

139. Wen, W., et al., *Mutations in the Kinase Domain of the HER2/ERBB2 Gene Identified in a Wide Variety of Human Cancers.* J Mol Diagn, 2015. **17**(5): p. 487-95.
140. Tyagi, M., et al., *Chromatin remodelers: We are the drivers!!* Nucleus, 2016. **7**(4): p. 388-404.
141. De, P. and N. Dey, *Mutation-Driven Signals of ARID1A and PI3K Pathways in Ovarian Carcinomas: Alteration Is An Opportunity.* Int J Mol Sci, 2019. **20**(22).
142. Wu, J.N. and C.W. Roberts, *ARID1A mutations in cancer: another epigenetic tumor suppressor?* Cancer Discov, 2013. **3**(1): p. 35-43.
143. Zhan, T., N. Rindtorff, and M. Boutros, *Wnt signaling in cancer.* Oncogene, 2017. **36**(11): p. 1461-1473.
144. Stirrups, R., *Alpelisib plus fulvestrant for PIK3CA-mutated breast cancer.* Lancet Oncol, 2019. **20**(7): p. e347.
145. Gaibar, M., et al., *Somatic Mutations in HER2 and Implications for Current Treatment Paradigms in HER2-Positive Breast Cancer.* J Oncol, 2020. **2020**: p. 6375956.
146. Hyman, D.M., et al., *AKT Inhibition in Solid Tumors With AKT1 Mutations.* J Clin Oncol, 2017. **35**(20): p. 2251-2259.
147. Byron, S.A., et al., *The N550K/H mutations in FGFR2 confer differential resistance to PD173074, dovitinib, and ponatinib ATP-competitive inhibitors.* Neoplasia, 2013. **15**(8): p. 975-88.
148. Chandler, R.L., et al., *Coexistent ARID1A-PIK3CA mutations promote ovarian clear-cell tumorigenesis through pro-tumorigenic inflammatory cytokine signalling.* Nat Commun, 2015. **6**: p. 6118.
149. Chung, T.K.H., et al., *Liquid biopsy of PIK3CA mutations in cervical cancer in Hong Kong Chinese women.* Gynecol Oncol, 2017. **146**(2): p. 334-339.
150. Xiang, L., et al., *PIK3CA mutation analysis in Chinese patients with surgically resected cervical cancer.* Sci Rep, 2015. **5**: p. 14035.
151. Kaidar-Person, O., S. Yosefia, and R. Abdah-Bortnyak, *Response of adenocarcinoma of the uterine cervix to chemoradiotherapy.* Oncol Lett, 2015. **9**(6): p. 2791-2794.
152. Jung, E.J., et al., *Cervical Adenocarcinoma Has a Poorer Prognosis and a Higher Propensity for Distant Recurrence Than Squamous Cell Carcinoma.* Int J Gynecol Cancer, 2017. **27**(6): p. 1228-1236.
153. Scholl, S., et al., *Clinical and genetic landscape of treatment naive cervical cancer: Alterations in PIK3CA and in epigenetic modulators associated with sub-optimal outcome.* EBioMedicine, 2019. **43**: p. 253-260.
154. Bossler, F., K. Hoppe-Seyler, and F. Hoppe-Seyler, *PI3K/AKT/mTOR Signaling Regulates the Virus/Host Cell Crosstalk in HPV-Positive Cervical Cancer Cells.* Int J Mol Sci, 2019. **20**(9).
155. Manzo-Merino, J., et al., *The role of signaling pathways in cervical cancer and molecular therapeutic targets.* Arch Med Res, 2014. **45**(7): p. 525-39.
156. Cho, H., et al., *Loss of ARID1A/BAF250a expression is linked to tumor progression and adverse prognosis in cervical cancer.* Hum Pathol, 2013. **44**(7): p. 1365-74.
157. Okawa, R., et al., *Aberrant chromatin remodeling in gynecological cancer.* Oncol Lett, 2017. **14**(5): p. 5107-5113.
158. Samartzis, E.P., et al., *Loss of ARID1A expression sensitizes cancer cells to PI3K- and AKT-inhibition.* Oncotarget, 2014. **5**(14): p. 5295-303.
159. Vora, C. and S. Gupta, *Targeted therapy in cervical cancer.* ESMO Open, 2018. **3**(Suppl 1): p. e000462.
160. Dagogo-Jack, I. and A.T. Shaw, *Tumour heterogeneity and resistance to cancer therapies.* Nat Rev Clin Oncol, 2018. **15**(2): p. 81-94.
161. Thomas, A., et al., *Expression profiling of cervical cancers in Indian women at different stages to identify gene signatures during progression of the disease.* Cancer Med, 2013. **2**(6): p. 836-48.

162. Wong, Y.F., et al., *Expression genomics of cervical cancer: molecular classification and prediction of radiotherapy response by DNA microarray.* Clin Cancer Res, 2003. **9**(15): p. 5486-92.
163. Upadhyay, P., et al., *Notch pathway activation is essential for maintenance of stem-like cells in early tongue cancer.* Oncotarget, 2016. **7**(31): p. 50437-50449.
164. Ghosh, S. and C.K. Chan, *Analysis of RNA-Seq Data Using TopHat and Cufflinks.* Methods Mol Biol, 2016. **1374**: p. 339-61.
165. Patro, R., et al., *Salmon provides fast and bias-aware quantification of transcript expression.* Nat Methods, 2017. **14**(4): p. 417-419.
166. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.
167. Babiceanu, M., et al., *Recurrent chimeric fusion RNAs in non-cancer tissues and cells.* Nucleic Acids Res, 2016. **44**(6): p. 2859-72.
168. Hu, X., et al., *TumorFusions: an integrative resource for cancer-associated transcript fusions.* Nucleic Acids Res, 2018. **46**(D1): p. D1144-D1149.
169. Gaidatzis, D., et al., *Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation.* Nat Biotechnol, 2015. **33**(7): p. 722-9.
170. Kapranov, P., et al., *The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA.* BMC Biol, 2010. **8**: p. 149.
171. Zhao, W., et al., *Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling.* BMC Genomics, 2014. **15**: p. 419.
172. Zhao, S., et al., *Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion.* Sci Rep, 2018. **8**(1): p. 4781.
173. Bodelon, C., et al., *Genomic characterization of viral integration sites in HPV-related cancers.* Int J Cancer, 2016. **139**(9): p. 2001-11.
174. Li, W., et al., *Characteristic of HPV Integration in the Genome and Transcriptome of Cervical Cancer Tissues.* Biomed Res Int, 2018. **2018**: p. 6242173.
175. Zhang, R., et al., *Dysregulation of host cellular genes targeted by human papillomavirus (HPV) integration contributes to HPV-related cervical carcinogenesis.* Int J Cancer, 2016. **138**(5): p. 1163-74.
176. Xu, B., et al., *Multiplex Identification of Human Papillomavirus 16 DNA Integration Sites in Cervical Carcinomas.* PLoS One, 2013. **8**(6): p. e66693.
177. Brant, A.C., et al., *Characterization of HPV integration, viral gene expression and E6E7 alternative transcripts by RNA-Seq: A descriptive study in invasive cervical cancer.* Genomics, 2019. **111**(6): p. 1853-1861.
178. Rauvala, M., et al., *Matrix metalloproteinases-2 and -9 in cervical cancer: different roles in tumor progression.* Int J Gynecol Cancer, 2006. **16**(3): p. 1297-302.
179. Feng, M., et al., *IL-17A promotes the migration and invasiveness of cervical cancer cells by coordinately activating MMPs expression via the p38/NF-kappaB signal pathway.* PLoS One, 2014. **9**(9): p. e108502.
180. Vazquez-Ortiz, G., et al., *Overexpression of cathepsin F, matrix metalloproteinases 11 and 12 in cervical cancer.* BMC Cancer, 2005. **5**: p. 68.
181. Zhang, Y.H., et al., *Matrix Metallopeptidase 14 Plays an Important Role in Regulating Tumorigenic Gene Expression and Invasion Ability of HeLa Cells.* Int J Gynecol Cancer, 2016. **26**(3): p. 600-6.
182. Liu, X., et al., *Development of Effective Therapeutics Targeting HER3 for Cancer Treatment.* Biol Proced Online, 2019. **21**: p. 5.
183. Ueda, A., et al., *Prognostic significance of the co-expression of EGFR and HER2 in adenocarcinoma of the uterine cervix.* PLoS One, 2017. **12**(8): p. e0184123.
184. Fields, G.B., *Mechanisms of Action of Novel Drugs Targeting Angiogenesis-Promoting Matrix Metalloproteinases.* Front Immunol, 2019. **10**: p. 1278.

185. Krishnamurthy, N. and R. Kurzrock, *Targeting the Wnt/beta-catenin pathway in cancer: Update on effectors and inhibitors.* Cancer Treat Rev, 2018. **62**: p. 50-60.
186. Iqbal, N. and N. Iqbal, *Imatinib: a breakthrough of targeted therapy in cancer.* Chemother Res Pract, 2014. **2014**: p. 357027.
187. Carneiro, B.A., et al., *FGFR3-TACC3: A novel gene fusion in cervical cancer.* Gynecol Oncol Rep, 2015. **13**: p. 53-6.
188. Liao, B.C., et al., *Treating patients with ALK-positive non-small cell lung cancer: latest evidence and management strategy.* Ther Adv Med Oncol, 2015. **7**(5): p. 274-90.
189. Wu, P., et al., *The Landscape and Implications of Chimeric RNAs in Cervical Cancer.* EBioMedicine, 2018. **37**: p. 158-167.
190. Cuykendall, T.N., M.A. Rubin, and E. Khurana, *Non-coding genetic variation in cancer.* Curr Opin Syst Biol, 2017. **1**: p. 9-15.
191. Gan, K.A., et al., *Identification of Single Nucleotide Non-coding Driver Mutations in Cancer.* Front Genet, 2018. **9**: p. 16.
192. Yi, K. and Y.S. Ju, *Patterns and mechanisms of structural variations in human cancer.* Exp Mol Med, 2018. **50**(8): p. 98.
193. Macintyre, G., B. Ylstra, and J.D. Brenton, *Sequencing Structural Variants in Cancer for Precision Therapeutics.* Trends Genet, 2016. **32**(9): p. 530-542.
194. Ciriello, G., et al., *Emerging landscape of oncogenic signatures across human cancers.* Nat Genet, 2013. **45**(10): p. 1127-33.
195. Peter, M., et al., *Frequent genomic structural alterations at HPV insertion sites in cervical carcinoma.* J Pathol, 2010. **221**(3): p. 320-30.
196. Boeva, V., et al., *Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data.* Bioinformatics, 2012. **28**(3): p. 423-5.
197. Luo, H., et al., *Genome-wide somatic copy number alteration analysis and database construction for cervical cancer.* Mol Genet Genomics, 2020.
198. Jang, H., Y. Hur, and H. Lee, *Identification of cancer-driver genes in focal genomic alterations from whole genome sequencing data.* Sci Rep, 2016. **6**: p. 25582.
199. Ren, T., et al., *Using low-coverage whole genome sequencing technique to analyze the chromosomal copy number alterations in the exfoliative cells of cervical cancer.* J Gynecol Oncol, 2018. **29**(5): p. e78.
200. Zare, F., et al., *An evaluation of copy number variation detection tools for cancer using whole exome sequencing data.* BMC Bioinformatics, 2017. **18**(1): p. 286.
201. Reinert, T., et al., *Clinical Implications of ESR1 Mutations in Hormone Receptor-Positive Advanced Breast Cancer.* Front Oncol, 2017. **7**: p. 26.
202. Li, Y., et al., *Patterns of somatic structural variation in human cancer genomes.* Nature, 2020. **578**(7793): p. 112-121.
203. Tian, L., et al., *Long-read sequencing unveils IGH-DUX4 translocation into the silenced IGH allele in B-cell acute lymphoblastic leukemia.* Nat Commun, 2019. **10**(1): p. 2789.
204. De Braekeleer, E., et al., *Identification of a MLL-MLLT4 fusion gene resulting from a t(6;11)(q27;q23) presenting as a del(11q) in a child with T-cell acute lymphoblastic leukemia.* Leuk Lymphoma, 2010. **51**(8): p. 1570-3.
205. Wiemels, J.L., et al., *Site-specific translocation and evidence of postnatal origin of the t(1;19) E2A-PBX1 fusion in childhood acute lymphoblastic leukemia.* Proc Natl Acad Sci U S A, 2002. **99**(23): p. 15101-6.
206. Davis, R.J., et al., *Fusion of PAX7 to FKHR by the variant t(1;13)(p36;q14) translocation in alveolar rhabdomyosarcoma.* Cancer Res, 1994. **54**(11): p. 2869-72.
207. Ozawa, T., et al., *PDGFRA gene rearrangements are frequent genetic events in PDGFRA-amplified glioblastomas.* Genes Dev, 2010. **24**(19): p. 2205-18.
208. Novo, F.J., I.O. de Mendibil, and J.L. Vizmanos, *TICdb: a collection of gene-mapped translocation breakpoints in cancer.* BMC Genomics, 2007. **8**: p. 33.

209. Tamborero, D., et al., *Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations.* Genome Med, 2018. **10**(1): p. 25.
210. Sadikovic, B., et al., *Cause and consequences of genetic and epigenetic alterations in human cancer.* Curr Genomics, 2008. **9**(6): p. 394-408.
211. Smith, J., *Erlotinib: small-molecule targeted therapy in the treatment of non-small-cell lung cancer.* Clin Ther, 2005. **27**(10): p. 1513-34.
212. Campeau, E. and S. Gobeil, *RNA interference in mammals: behind the screen.* Brief Funct Genomics, 2011. **10**(4): p. 215-26.
213. Dai, Z., et al., *edgeR: a versatile tool for the analysis of shRNA-seq and CRISPR-Cas9 genetic screens.* F1000Res, 2014. **3**: p. 95.
214. Luo, B., et al., *Highly parallel identification of essential genes in cancer cells.* Proc Natl Acad Sci U S A, 2008. **105**(51): p. 20380-5.
215. Konig, R., et al., *A probability-based approach for the analysis of large-scale RNAi screens.* Nat Methods, 2007. **4**(10): p. 847-9.
216. Wu, D., et al., *ROAST: rotation gene set tests for complex microarray experiments.* Bioinformatics, 2010. **26**(17): p. 2176-82.
217. Wu, D. and G.K. Smyth, *Camera: a competitive gene set test accounting for inter-gene correlation.* Nucleic Acids Res, 2012. **40**(17): p. e133.
218. McFarland, J.M., et al., *Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration.* Nat Commun, 2018. **9**(1): p. 4610.
219. Chandrani, P., et al., *Integrated genomics approach to identify biologically relevant alterations in fewer samples.* BMC Genomics, 2015. **16**: p. 936.
220. Gao, J., et al., *Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.* Sci Signal, 2013. **6**(269): p. pl1.
221. Dar, A.A., et al., *Aurora kinase inhibitors--rising stars in cancer therapeutics?* Mol Cancer Ther, 2010. **9**(2): p. 268-78.
222. Gonzalez-Loyola, A., et al., *Aurora B Overexpression Causes Aneuploidy and p21Cip1 Repression during Tumor Development.* Mol Cell Biol, 2015. **35**(20): p. 3566-78.
223. Tang, A., et al., *Aurora kinases: novel therapy targets in cancers.* Oncotarget, 2017. **8**(14): p. 23937-23954.
224. Gully, C.P., et al., *Aurora B kinase phosphorylates and instigates degradation of p53.* Proc Natl Acad Sci U S A, 2012. **109**(24): p. E1513-22.
225. Tao, Y., et al., *Enhancement of radiation response in p53-deficient cancer cells by the Aurora-B kinase inhibitor AZD1152.* Oncogene, 2008. **27**(23): p. 3244-55.
226. Ha, G.H. and E.K. Breuer, *Mitotic Kinases and p53 Signaling.* Biochem Res Int, 2012. **2012**: p. 195903.
227. Bagegni, N., et al., *Serum thymidine kinase 1 activity as a pharmacodynamic marker of cyclin-dependent kinase 4/6 inhibition in patients with early-stage breast cancer receiving neoadjuvant palbociclib.* Breast Cancer Res, 2017. **19**(1): p. 123.
228. Ning, S., et al., *Clinical significance and diagnostic capacity of serum TK1, CEA, CA 19-9 and CA 72-4 levels in gastric and colorectal cancer patients.* J Cancer, 2018. **9**(3): p. 494-501.
229. Zhou, J., E. He, and S. Skog, *The proliferation marker thymidine kinase 1 in clinical use.* Mol Clin Oncol, 2013. **1**(1): p. 18-28.
230. Moser, R., et al., *Functional kinomics identifies candidate therapeutic targets in head and neck cancer.* Clin Cancer Res, 2014. **20**(16): p. 4274-88.
231. Smith, S.L., et al., *Overexpression of aurora B kinase (AURKB) in primary non-small cell lung carcinoma is frequent, generally driven from one allele, and correlates with the level of genetic instability.* Br J Cancer, 2005. **93**(6): p. 719-29.
232. Hegyi, K., et al., *Aurora kinase B expression in breast carcinoma: cell kinetic and genetic aspects.* Pathobiology, 2012. **79**(6): p. 314-22.

233. Mehra, R., et al., *Aurora kinases in head and neck cancer.* Lancet Oncol, 2013. **14**(10): p. e425-35.
234. Boeckx, C., et al., *Overcoming cetuximab resistance in HNSCC: the role of AURKB and DUSP proteins.* Cancer Lett, 2014. **354**(2): p. 365-77.
235. Falchook, G.S., C.C. Bastida, and R. Kurzrock, *Aurora Kinase Inhibitors in Oncology Clinical Trials: Current State of the Progress.* Semin Oncol, 2015. **42**(6): p. 832-48.
236. Chen, Y., et al., *Serum thymidine kinase 1 correlates to clinical stages and clinical reactions and monitors the outcome of therapy of 1,247 cancer patients in routine clinical settings.* Int J Clin Oncol, 2010. **15**(4): p. 359-68.
237. Togar, T., et al., *Identifying cancer driver genes from functional genomics screens.* Swiss Med Wkly, 2020. **150**: p. w20195.
238. Schuster, A., et al., *RNAi/CRISPR Screens: from a Pool to a Valid Hit.* Trends Biotechnol, 2019. **37**(1): p. 38-55.
239. Luo, J., *CRISPR/Cas9: From Genome Engineering to Cancer Drug Discovery.* Trends Cancer, 2016. **2**(6): p. 313-324.

Chapter VIII

Appendices

Chapter VIII: Appendices

Please find these appendices linked to google drive for assessing.

8.1 Appendix 1: List of somatic variants identified from whole exome sequencing of 18 samples

8.2 Appendix 2: List of transcripts identified from whole transcriptome sequencing of 21 samples

8.3 Appendix 3: List of copy number alterations in cervical adenocarcinoma and squamous carcinoma

8.4 Appendix 4: List of copy number alterations from whole genome sequencing

8.5 Appendix 5a: List of gene fusions identified from whole transcriptome sequencing

8.6 Appendix 5b: List of gene fusions identified from whole transcriptome sequencing with information of spanning and junction reads.

8.7 Appendix 6: List of somatic variants identified from SNPiR

8.8 Appendix 7: List of somatic mutations from whole genome sequencing

8.9 Appendix 8: ROAST ranking for all depleted kinases identified from screen 1 and screen 2

8.10 Appendix 9: Gene expression, copy number and shRNA log FC values for screen 1 and screen 2

8.11 Appendix 10: List of top depleted kinases from the screen identified by considering the cumulative effect of four parameter- gene rank, copy number change, gene expression and logFC value of shRNA depletion for the kinase and calculating impact score.

8.12 Appendix 11: The Rank Impact Score (IS) and weights (W) assigned for all kinases in each of the two screens is shown. All values of all the four parameters- rank (RR), copy number (CR), gene expression (GR) and logFC value (DR) is shown.

Chapter IX

Reprints of publication

Established in 1871

Swiss Medical Weekly

Formerly: Schweizerische Medizinische Wochenschrift

An open access, online journal • www.smw.ch

Original article | Published 21 February 2020 | doi:10.4414/smw.2020.20195
Cite this as: Swiss Med Wkly. 2020;150:w20195

Identifying cancer driver genes from functional genomics screens

Togar Trupti[ab], Desai Sanket[ab], Mishra Rohit[a], Terwadkar Prachi[a], Ramteke Manoj[a], Ranjan Malika[a], Kawle Dhananjay[a], Sahoo Bikram[a], Pal Ankita[a], Upadhyay Pawan[ab], Dutt Amit[ab]

[a] Integrated Genomics Laboratory, Advanced Centre for Treatment, Research and Education in Cancer, Tata Memorial Centre, Navi Mumbai, India
[b] Training School Complex, Homi Bhabha National Institute, Anushakti Nagar, Mumbai, India

Summary

With the emerging advances made in genomics and functional genomics approaches, there is a critical and growing unmet need to integrate plural datasets in order to identify driver genes in cancer. An integrative approach, with the convergence of multiple types of genetic evidence, can limit false positives through a posterior filtering strategy and reduce the need for multiple hypothesis testing to identify true cancer vulnerabilities. We performed a pooled shRNA screen against 906 human genes in the oral cancer cell line AW13516 in triplicate. The genes that were depleted in the screen were integrated with copy number alteration and gene expression data and ranked based on ROAST analysis, using an integrative scoring system, DepRanker, to compute a Rank Impact Score (RIS) for each gene. The RIS-based ranking of candidate driver genes was used to identify the putative oncogenes *AURKB* and *TK1* as essential for oral cancer cell proliferation. We validated the findings, showing that shRNA mediated genetic knockdown of *TK1* or pharmacological inhibition of AURKB by AZD-1152 HQPA in AW13516 cells could significantly impede their proliferation. Next we analysed alterations in *AURKB* and *TK1* genes in head and neck cancer and their association with prognosis using data on 528 patients obtained from TCGA. Patients harbouring alterations in *AURKB* and *TK1* genes were associated with poor survival. To summarise, we present DepRanker as a simple yet robust package with no third-party dependencies for the identification of potential driver genes from a pooled shRNA functional genomic screen by integrating results from RNAi screens with gene expression and copy number data. Using DepRanker, we identify *AURKB* and *TK1* as potential therapeutic targets in oral cancer. DepRanker is in the public domain and available for download at http://www.actrec.gov.in/pi-webpages/AmitDutt/DepRanker/DepRanker.html.

Keywords: *pooled RNAi screen, kinase, genomics, DepRanker,* AURKB, *TK1*

Introduction

Cancer is a disease defined by several genetic alterations, such as mutations, gene expression changes and copy number changes, in addition to epigenomic alterations [1]. While most of the alterations are passenger alterations with no significant effect on cellular phenotype, cancer cells are dependent on a few driver genes for the constitutive activation of the signalling pathways which aid cellular proliferation, a phenomenon described as oncogene addiction [2]. Targeting oncogenic-dependent genes has resulted in success, as demonstrated in several cancer types [3, 4]. Often, the discovery or identification of a cancer-associated driver oncogene based on a genomics approach requires screening for significant genetic alterations using stringent statistical methods, followed by functional validation. On the other hand, a complementary functional genomics approach using RNAi or CRISPR effectively uses this structural knowledge of the cancer genome to define the functional consequences of the alterations in an unbiased manner, and may be performed in a pooled or arrayed format [5]. Methods which perform genome-wide RNAi screens on human cancer cell lines using a pooled human shRNA library as experimental models offer a powerful methodology for the identification of those genes essential for the survival of the cells. These efforts provide a new opportunity to fundamentally alter the extent to which we are able to understand and validate molecules that, when targeted, lead to therapeutic benefits in cancer patients.

ABBREVIATIONS:

AURKB	aurora kinase B
CR	copy number alteration rank
DepRanker	dependency ranker
DR	depletion rank
FC	fold change
GR	gene expression rank
GUI	graphic user interface
MOI	multiplicity of infection
RIS	Rank Impact Score
RNAi	RNA interference
RR	ROAST rank
TCGA	The Cancer Genome Atlas
TK1	thymidine kinase 1
W	weight

Author contributions
TT, PT, M Ramteke, M Ranjan, DM, AP and PU contributed to the functional work; TT, SD and AD designed the research; SD, BS and RM performed the bioinformatics analysis; TT, SD and AD analysed the data; and TT, SD and AD wrote the paper. All authors have read and approved the manuscript.

Correspondence:
Dr Amit Dutt, PhD, Advanced Centre for Treatment, Research and Education in Cancer, Tata Memorial Centre, Navi Mumbai, Maharashtra, India-410210, adutt[at]actrec.gov.in

Typically, a pooled RNAi screen analysis involves a quality assessment and normalisation of the data, followed by differential shRNA/sgRNA representation. The differential analysis is performed either by custom scripts or by packages like edgeR [6]. The "tags" (shRNA) are ranked according to their differential effects among classes of samples, and are further organised into a ranked list of genes by packages like RIGER [7], RSA [8], ROAST [9], camera [10] and others. Moreover, there are specialised algorithms like DEMETER2 [11] which measure the on/off-target effect and also estimate gene-dependency by deriving 'essentiality scores' from the RNAi experiments. The genes obtained from these experiments may be further validated, either by performing specific knock-down experiments or by extended secondary screens.

An alternative approach used to define dependency from pooled screen experiments is the integration of genomic data with the gene essentiality results. A classic example of this approach is the cancer dependency map [12], which integrates genomic features such as expression, copy number and mutation information with the gene dependencies obtained from screens performed on cancer cell lines representing various tumour types. Few computational methods incorporate such genomic features when predicting driver or essential genes for pooled RNAi screen experiments [13]. Building on this integrative approach, we have developed a gene ranking or scoring method, DepRanker, which incorporates other genomic datasets like gene expression and copy number information of the same cell line to prioritise genes from pooled screen results for their essentiality. DepRanker consists of two modules that can be executed using a single, user-friendly GUI. Module I analyses the pooled screen data to calculate the depletion of the tags and prioritise the genes. Module II integrates the results obtained from Module I with the genome-wide datasets to compute the Rank Impact Score (RIS) for individual genes.

We performed a functional kinome screen using pooled shRNA, comprised of 5419 constructs targeting 906 human kinases in AW13516 cells, in two independent screens. The genes depleted in the screen were integrated with copy number alteration data and gene expression data for the AW13516 cells using DepRanker, allowing us to identify *AURKB* and *TK1* as potential therapeutic targets in oral cancer.

Materials and methods

Cell lines and cell culture

Indian patient-derived head and neck cancer cell lines – AW13516 cells and other cells used in the study, namely 293FT, HCT116 and SiHa cells – were maintained in Dulbecco's Modified Eagle Medium (Gibco) supplemented with 10% FBS (Gibco) and 1% Penicillin-Streptomycin solution (Sigma). Cells were grown at 37°C in a 5% CO_2 incubator. Cells were treated with Mycoplasma elimination kit (EZKill solution, Himedia) prior to use.

Lentivirus production and transduction in HNSCC cell line

Lentivirus comprised of 5419 pZIP-SFFV pooled shRNA constructs (8.1 Kb) targeting 906 human kinases were obtained from TransOMIC Technologies, USA. For the pooled shRNA screen, 18 million AW13516 cells were seeded in T-150 flasks at 60-70% confluency. Lentivirus was transduced at an MOI of 0.3 in the presence of 8 µg/ml Polybrene (Sigma) at 1000-fold representation of each shRNA in the screen. Cells were grown at 37°C for 16 hours post virus addition, and the medium was replaced. Cells were selected in the presence of 1 µg/ml puromycin (Sigma). Half the cells were harvested within 3-4 days after selection and this sample was termed the day 0 (control) sample. The remaining cells were further expanded and maintained at 37°C, and collected as test samples at the day 10 and day 20 time points.

PCR amplification of shRNA and barcode sequencing by NGS

Genomic DNA was extracted from the day 0, day 10 and day 20 samples of the AW13516 cells using a QIAamp DNA blood kit (Qiagen). DNA concentration estimation was done using a Nanodrop 2000c spectrophotometer (Thermo Fischer Scientific). Instructions provided in the TransOmics manual for performing PCR for shRNA amplification were followed, with some modifications. To provide a 1000-fold representation of shRNA, 36 µg of genomic DNA was used to amplify the shRNA cassette as per the calculation, and primary PCR was performed (sequence information in supplementary table S1 in appendix 1) as follows: 10 µl of 5X HF buffer, 1.5 µl of each of the forward and reverse primary PCR primers at concentrations of 10 µM, 1 µl of 10 mM dNTP mix, 5% DMSO, 3 mM $MgCl_2$, 0.5 µl of Phusion High-Fidelity Polymerase Enzyme (Thermo Fischer Scientific) and 850 ng of genomic DNA in a total reaction volume of 50 µl. Primary PCR was performed at thermocycler conditions: 98°C for 5 min, 25 cycles of 95°C for 30 sec, 57°C for 30 sec and 72°C for 30 sec, and a final extension at 72°C for 5 min. The PCR product was separated on 1.5% agarose gel to visualise an amplicon of 406 bp. Next, the primary PCR product was pooled and purified using Nucleospin Gel and a PCR clean-up kit (Macherey-Nagel) and quantified using the Nanodrop 2000c spectrophotometer. 2 µg of purified primary PCR was used for setting up nested secondary PCR (primer sequence information in supplementary (table S1) with indexed reverse primers that add a unique barcode sequence to each sample to facilitate sample pooling during NGS sequencing. The secondary PCR reaction mixture was comprised of 10 µl of 5X HF buffer, 1.5 µl of each of the forward and indexed reverse secondary PCR primers at concentrations of 10 µM, 1 µl of 10 mM dNTP mix, 5% DMSO, 0.5 µl of Phusion High-Fidelity Polymerase Enzyme (Thermo Fischer Scientific) and 500 ng of primary PCR product in a total reaction volume of 50 µl. Secondary PCR was performed at thermocycler conditions: 98°C for 5 min, 15 cycles of 94°C for 30 sec, 52°C for 30 sec and 72°C at 30 sec, and a final extension at 72°C for 5 min. The secondary PCR product was separated on 1.5% agarose gel to visualise a band of 408 bp. It was then pooled and subjected to purification using Agencourt Ampure XP beads (NEB) and quantitated using a Qubit Fluorometer (Thermo Fischer Scientific). About 8-20 pM of purified secondary PCR product (indexed library) was loaded on an Illumina HiSeq 2500 platform and 50 bp single-end sequencing was done.

Data analysis of pooled shRNA using the edgeR pipeline

Raw data was obtained as fastq files and further processed using the edgeR package [6] for analysis of pooled shRNA data. Counts per sample were obtained for each shRNA by mapping reads with the kinase shRNA sequence library. For screen 1 data, shRNA with control sample (day 0) counts less than 1000 were excluded, since the experiment was performed at 1000-fold representation. For screens 2 and 3, a cut-off of 100 shRNA in the control sample (day 0) was used for further analysis. Data normalisation was performed within and across the control and test samples. The screen data was analysed using the classical method of two group comparisons. Statistical analysis was done to estimate the significance of the observed changes in shRNA abundance. The edgeR package provided a list of depleted shRNAs by calculating the log fold change (logFC). Based on these results, the top enriched and depleted shRNAs from the screen were identified and further converted to a gene-level ranking using the gene set analysis tool 'ROAST' [9]. Kinases represented by at least two shRNAs were considered for further analysis. A list of the kinases that were depleted in cells at day 20 compared to day 0 was obtained. Data from screens 1, 2 and 3 were not considered as data in triplicate because the screen 1 data output was enormous and captured existing shRNA uniformly, whereas the screen 2 and screen 3 data outputs were comparatively lower, suggesting that some of the shRNAs were not captured (table 1). Therefore, the screen 2 and screen 3 data were used as replicates. Hence, the combined results of screen 2 and screen 3 are referred to as screen 2 data hereafter.

DepRanker assigned impact score for the identification of potential kinases using genomic alteration data

To further prioritise the candidate kinases obtained from the RNAi screen analysis, we developed a scoring method named DepRanker (Dependency Ranker). DepRanker calculates a Rank Impact Score (RIS) for the individual kinases, which are derived from the kinome screen, by integrating gene expression and copy number data from the same sample. The RIS is derived using the following equation:

$$RIS_{(Kinase\ A)} = DR_{(Kinase\ A)} + RR_{(Kinase\ A)} + GR_{(Kinase\ A)} + CR_{(Kinase\ A)}$$

(where DR = depletion rank, RR = ROAST rank, GR = gene expression rank and CR = copy number alteration rank).

We used a mean-rank method to calculate the scores for each feature as described below. The DR is derived by converting the logFC values obtained from the edgeR depletion analysis into rankings. The kinase showing the highest depletion in the screen is assigned the highest rank and the one showing the lowest depletion is assigned a rank of '1'. The RR is based on the ranking given by the ROAST algorithm, in which those genes which are represented by at least three shRNAs are considered, and the kinases are sorted based on their p-value. The gene which is least prioritised by ROAST is given a rank of '1' and the most prioritised gene is assigned the highest rank. To calculate GR and CR, gene expression and copy number alteration data for all the kinases showing significant depletion in the pooled screen analysis is extracted for the relevant cell line, AW13516 cells in this analysis (as described previously [14]). The log transformed FPKM gene expression levels were extracted for this subset of kinases (all those showing significant depletion in the pooled screen), and the kinase showing the lowest gene expression was assigned a rank of 1 while the gene with the highest expression was assigned the highest rank. Similar rankings were assigned to the copy number levels for individual kinases from the AW13516 cells in order to derive CR. All four scores (DR, RR, GR, CR) were added together to compute the RIS. This scoring approach enabled us to identify potential kinases with biological roles from the list. To combine the results obtained from two screens performed on the same cell line, we converted the RIS for an individual kinase into a *weight* (range between 0 and 1) based on its relevance in a particular screen.

Furthermore, to combine the results from both screens, we assigned a weighting to each of the kinases by considering their RIS for both screens. The weights were calculated using the formula W = (RIS for kinase A) / (sum of RIS for all the kinases). The results from both screens were combined and sorted based on the assigned weightings. In the case of kinases with overlap in both screens, the kinase with the higher weight was retained.

Implementation of DepRanker and graphical user interface

This scoring system is implemented as a python-based package. DepRanker takes the output from edgeR analysis of pooled shRNA screens and the results provided by ROAST, along with gene expression data and copy number variation data for individual genes belonging to the cell line, and outputs the list of candidate kinases with their Rank Impact Scores. The package is available at http://www.actrec.gov.in/pi-webpages/AmitDutt/DepRanker/DepRanker.html, along with complete installation instructions and a user manual. The GUI was designed using the Tkinter python package. A detailed user manual for the GUI is available. The GUI provides two modules for analysis. The first module is the pooled shRNA screen analysis module, which takes in the fastq, hairpin and sample information file to perform the depletion analysis. The depletion analysis can be performed using either a generalised linear model (GLM) or an exact-test based method. The users are advised to refer to the screen analysis manual of Zuber et al. (http://bioinf.wehi.edu.au/shRNAseq/pooledScreenAnalysis.pdf) for guidance on selecting a suitable method for their screen data analysis. Internally, the GUI calls the Bioconductor packages edgeR and ROAST to perform the depletion analysis and the gene prioritisation respectively. The results from this module (edgeR toptags result and ROAST result file), along with the copy number and gene expression data for the cell line analysed, should be provided to the DepRanker module. This module provides the rank-based scores for the individual kinases identified from the pooled screen. The DepRanker GUI package is freely available for download.

Survival analysis of HNSCC datasets

Genomic alteration data from TCGA provisional HNSCC datasets from cBioPortal [15], consisting of 528 samples with gene expression, copy number and mutation information, was assessed. Kaplan-Meier survival plots were

generated for patients with alterations in *AURKB* and *TK1* genes.

Real time PCR for amplification of shRNA

Real time primers were designed for each shRNA of *AURKB* and *TK1* wherein the forward primer sequence was complementary to the kinase shRNA sequence and the reverse primer was common for all, binding to the 3′ miR vector sequence. PCR was performed using purified primary PCR product as a template. An amplicon of 100 bp was expected. Primer sequences are provided in supplementary table S1 (appendix 1)

MTT assay for functional validation of hit obtained from screen

An MTT assay was performed using the *AURKB* inhibitor AZD1152-HQPA (Sigma). The colon cancer cell line HCT116 (sensitive) and the cervical cell line SiHa (resistant) were used as control cells for the MTT assay. In brief, 1000 cells of AW13516, 1500 cells of HCT116 and 2000 cells of SiHa were seeded in 96 well plates. The cells were treated with AZD1152-HQPA inhibitor for 72 hours before the MTT (0.5 mg/ml) reagent was added and the cells were incubated for 3 hours at 37°C in a CO_2 incubator. DMSO was used for developing and a reading was obtained at 570 nm using a microplate reader (iMark microplate reader, Biorad). The percentage cell viability was calculated with respect to the untreated control cells. The assay was performed three times.

Generation of TK1 knockdown clones of AW13516

pZIP-hCMV shRNA constructs targeting TK1 genes and a scrambled control (TransOmics Technologies, USA) were used along with Lipofectamine 3000 transfection reagent (Invitrogen) for lentiviral production in 293FT cells. Lentivirus was harvested at 48 and 72 hours and filtered using a 0.4 µM filter. AW13516 cells were transduced with virus in the presence of 8 µg/ml concentration of polybrene and selection was done using 1 µg/ml puromycin for 4-5 days. The cells selected were positive for GFP expression. The shRNA sequences are as follows: TK1 sh1 – AAGCAGACAAGTACCACTCCG and TK1 sh2 – CCCAGGTGATTCTCGGGCCGA.

Western blotting

The cells were lysed in RIPA lysis buffer (Sigma) supplemented with 1 mM dithrothreitol (DTT) and protease inhibitor cocktail (Calbiochem, Merck), and quantitated using the BCA protein estimation method. 40 µg of protein was loaded on 12% SDS-PAGE gel and transferred onto PVDF membrane (Amersham Hybond, GE healthcare) by electro blotting. The membrane was stained with Ponceau to confirm protein transfer. Blocking was done in 5% BSA (prepared in 1X Tris Buffered Saline buffer with Tween-20) and blots were incubated with primary antibody overnight at 4°C, and then with secondary HRP conjugated antibody for one hour at room temperature. Blots were then washed in 1X TBST buffer and developed using Pierce ECL western blotting substrate (Thermo Fischer Scientific). Luminescence was captured on a Chemidoc System (Biorad). Primary antibody for *TK1* (cell signalling) was used at a dilution of 1:1000 and secondary HRP conjugated goat anti-rabbit antibody (Santa Cruz Biotechnologies) was used at 1:2000 dilution.

Cell proliferation assay

Twenty thousand cells/well were seeded in a 24 well plate. Cell growth was assessed at 24 and 96 hours and the cells were counted using a haemocytometer. The percentage cell proliferation was calculated with respect to the scrambled control cells. The experiments were repeated in triplicate.

Results

A pooled kinome shRNA screen to identify oncogenic dependency in head and neck cancer cells

In order to identify essential genes in head and neck cancer, we performed a pooled kinome shRNA screen in the head and neck cancer cell line AW13516, derived from a tongue cancer patient from India, using 5419 pooled shRNA constructs targeting 906 human kinases. About 14 million cells were transduced with lentiviral particles harbouring shRNA against kinases at an MOI of 0.3. Following transduction, the cells were subjected to puromycin selection (1 µg/ml) and half the cells were harvested 3 or 4 days post selection. These cells were called the day 0 sample and served as a control. The remaining cells were passaged for 20 days in culture and collected at day 10 and day 20. Genomic DNA was extracted, shRNA amplification was performed, and barcode sequences were added by PCR (fig. 1). Each sample was tagged with a unique barcode to allow identification of the shRNAs belonging to each sample in order to enable sample multiplexing during sequencing.

Data deconvolution was performed using the edgeR package. Briefly, reads with shRNA sequences were mapped to the human kinome library and the percent mapping was estimated. Data QC revealed that about 75% of reads mapped to kinome references in AW13516 (table 1). shRNA hairpins with low counts (less than 0.5 counts per million) at day 0 were excluded from the analysis since the screen was performed at 1000-fold representation. The relative shRNA abundances in the day 0, day 10 and day 20 samples were estimated after performing within- and across-sample normalisation. A list of enriched and depleted shRNA hairpins was obtained by comparing the day 20 samples with the day 0 control samples. For screen 1, a time series analysis of the kinases enriched and depleted at day 10 and day 20 was done using the day 0 sample as a control. Data from screen 2 and screen 3 were used as replicates to identify shRNA hairpins that were enriched and depleted at day 20 compared to day 0. Gene-level information was derived for these shRNAs using the 'ROAST' module, and kinases that were de-regulated were ranked according to their depletion (supplementary table S2 in appendix 2). Kinases that are lost from the screen over time have potential roles as oncogenes, since depletion of these kinases by shRNA in cells is inducing a cell death phenotype, whereas kinases that get enriched may be acting as tumour-suppressors. Knockdown of these enriched kinases tends to promote cell proliferation, and therefore enrichment of shRNA is observed over the time.

Figure 1: Schematic representation of pooled shRNA screen in AW13516 cells. AW13516 cells were transduced with pooled shRNA lentivirus targeting 906 human kinases at an MOI of 0.3. Cells were selected in puromycin and half the cells were collected as the day 0 sample, which was used as the reference or control sample, as it represents all the shRNA after transduction. The remaining half of the cells were passaged for up to 20 days in culture and collected as the day 20 sample. Genomic DNA was extracted from both samples and shRNA sequences were amplified by primary and secondary PCR. Then the indexed library, secondary PCR product was sequenced and shRNA counts were obtained for both samples. The shRNA sequences which were depleted in the day 20 compared to the day 0 ones were considered. Depleted kinases are those which have an oncogenic role in these cells.

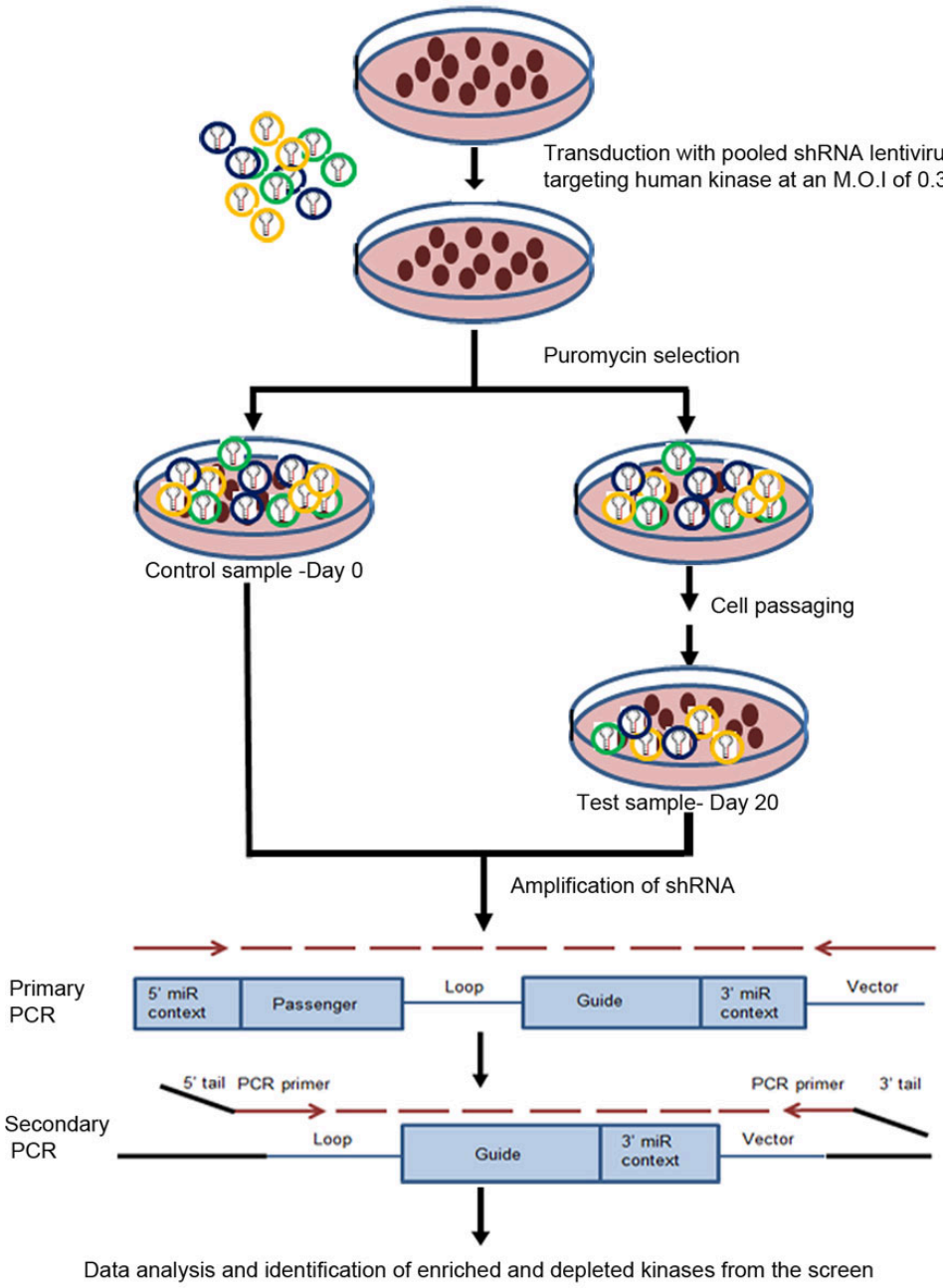


Table 1: QC data from sequencing showing the percentage of the reads mapping to the kinome library for all three samples for each of the three screens of the AW13516 cell line.

	AW13516 Screen1			AW13516 Screen2			AW13516 Screen3		
Sample	Day 0	Day 10	Day 20	Day 0	Day 10	Day 20	Day 0	Day 10	Day 20
Total reads	7,306,986	24,734,650	12,806,948	768,023	3,116,132	1,932,070	932,563	1,626,831	929,280
Total reads mapping to kinome	6,885,742	23,167,685	11,999,717	574,936	1,940,642	1,522,511	635,114	996,063	466,771
Percent mapping to Kinome	94.23	93.66	93.69	79.18	65.23	81.89	78.65	81.49	78.58

An integrated scoring system and analytical package, DepRanker, to rank biologically relevant genes

The GUI based pooled shRNA screen analysis and gene prioritisation package DepRanker was used to rank and identify biologically relevant genes. In screen 1, 127 kinases that were depleted in AW13516 cells and had available gene expression and copy number data were identified, while 146 such kinases were identified in screen 2 (table S3 in appendix 2). Gene expression and copy number alteration data for all the kinases showing significant depletion in the pooled screen were analysed for AW13516 cells [14]. Next, we used DepRanker to integrate genomics data such as the gene expression data, copy number data, ranking given by ROAST analysis and average logFC value of all the shRNAs associated with a gene to calculate the Rank Impact Score (RIS) for each kinase in the screen (fig. S1), as described in the methodology. The results from both screens were pooled together by considering the mean weight assigned to each kinase as described in the methodology (table S4). The kinase rankings for both screens are shown in table S5.

DepRanker ranked *AURKB* and *TK1* as the top genes after combining the results from the two screens using the assigned weights (fig. 2). Due to the non-inclusion of the normal immortalised oral cells, the essential role of *AURKB* and *TK1* in oral cancer cells couldn't be established exclusively based on the screens performed. However, given that *AURKB* and *TK1* are overexpressed and show high copy gain in AW13516 oral cancer cells, the data, along with the functional screen, suggest their potential oncogenic role in oral cancer (table S4). To confirm the reproducibility of the results obtained from our bioinformatics analysis, the counts for each shRNA in the day 0, day 10 and day 20 samples were validated using real time PCR for the selected candidate kinases. The shRNA counts targeting *AURKB* and *TK1* were observed to be depleted in the day 10 and day 20 samples compared to the day 0 control sample, suggesting that these kinases confer oncogenic dependence in head and neck cancer cell lines and are essential for cell survival, as knockdown of these kinases resulted in the elimination of the corresponding shRNAs from the population over time (data not shown). These results were consistent with our bioinformatics analysis, wherein we observed a depletion of the shRNA constructs targeting *AURKB* and *TK1* in the day 10 and day 20 samples compared to the day 0 sample. Here, we considered the mean CPM (counts per million) of each shRNA construct for both genes across all three screens. The percent shRNA counts at each time point are plotted in figure 3. All three shRNAs of *AURKB* show consistent depletion at day 10 and day 20.

Figure 2: Heatmap representation of depleted kinases in the screen considering overall Rank Impact Score (RIS). Heatmap representation of the kinases depleted in the screen which have a high impact score according to the ranking assigned by considering ROAST, gene expression data, copy number data and the average logFC of the depleted shRNA for that kinase. The enlarged view shows the top 10 kinases with the highest impact scores. *AURKB* and *TK1* kinases top the list.

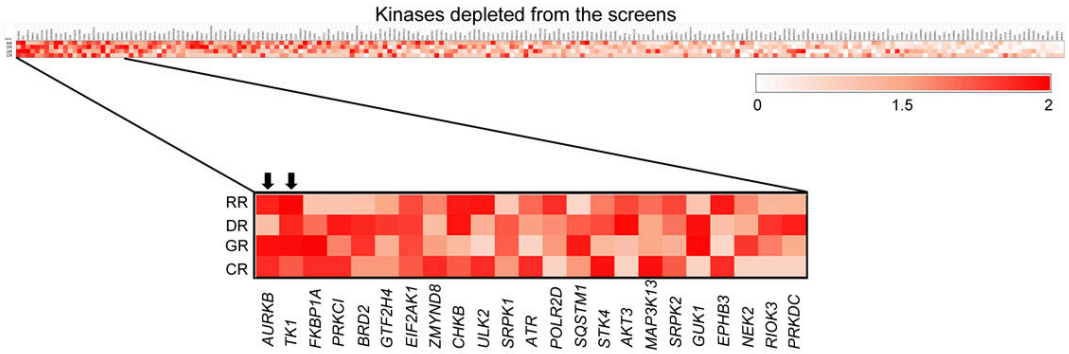
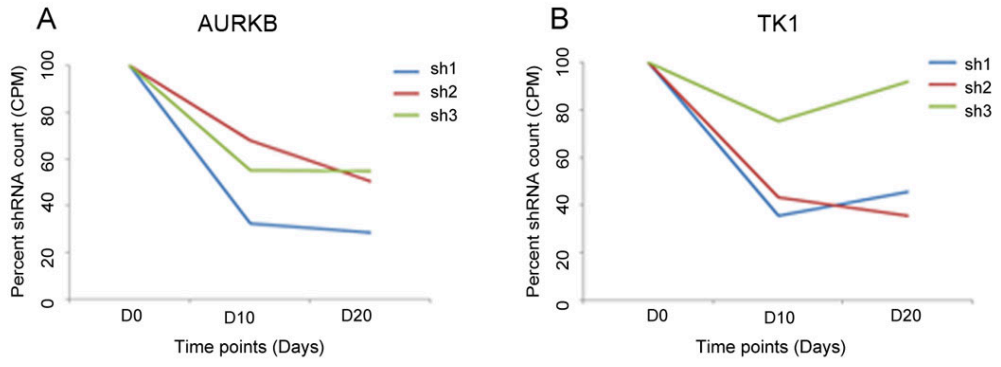


Figure 3: Graph showing percent shRNA counts at day 0, day 10 and day 20 for three shRNAs of *AURKB* and *TK1*. The mean CPM counts of each shRNA construct for both genes across all three screens were obtained and the percent shRNA counts at each time point are plotted.



AURKB and *TK1* kinases confer oncogenic dependency in AW13516 cells

AURKB is a chromosomal passenger protein which is critical for the accurate segregation of chromosomes during cell division [16]. However, in several cancers over-expression of *AURKB* is often associated with poor prognosis [17]. *AURKB*-mediated phosphorylation suppresses the activity of p53 through several mechanisms [18, 19]. However, several studies have also reported that inhibitors of *AURKB* are effective at inhibiting cell growth in p53 mutant cell lines [20, 21]. AW13516 cells harbour the p53 mutations p.R273H and p.R72fs*51.

To confirm *AURKB* as a potential oncogenic kinase conferring cell survival of AW13516 cells, we performed an MTT assay on AW13516 cells using AZD1152-HQPA inhibitor. We observed that the AW13516 cells were sensitive to the inhibitor, with an IC50 value of 40 nM. HCT116 colon cells were used as a sensitive cell line for the assay, and cervical cancer SiHa cells were used as resistant cells (fig. 4A). These results suggest that the *AURKB*-specific inhibitor AZD1152-HQPA could inhibit the cell viability of p53 mutant AW13156 cells. The results are consistent with the sensitivity of this inhibitor to other cells, such as HT29 cells with the similar p53 mutation p.R273H [20].

Thymidine kinase 1 (*TK1*) was identified from the screen as another potential target. *TK1* is an enzyme that plays a role in the first step of the biosynthesis of dTTP during DNA synthesis in cells [22]. High expression of *TK1* in cancer tissues is associated with disease progression and poor prognosis [23]. Serum *TK1* levels are used as a prognostic biomarker in several cancers, including head and neck cancer, to predict the outcome of treatment [24]. *TK1* is thus an attractive target. To functionally characterise the role of *TK1*, we performed knockdown of *TK1* in AW13156 cells and confirmed the knockdown by western blotting. We performed a cell proliferation assay and observed that proliferation was significantly ($p < 0.0001$) affected in the knockdown clones compared to the scrambled control cells (figs 4B and 4C).

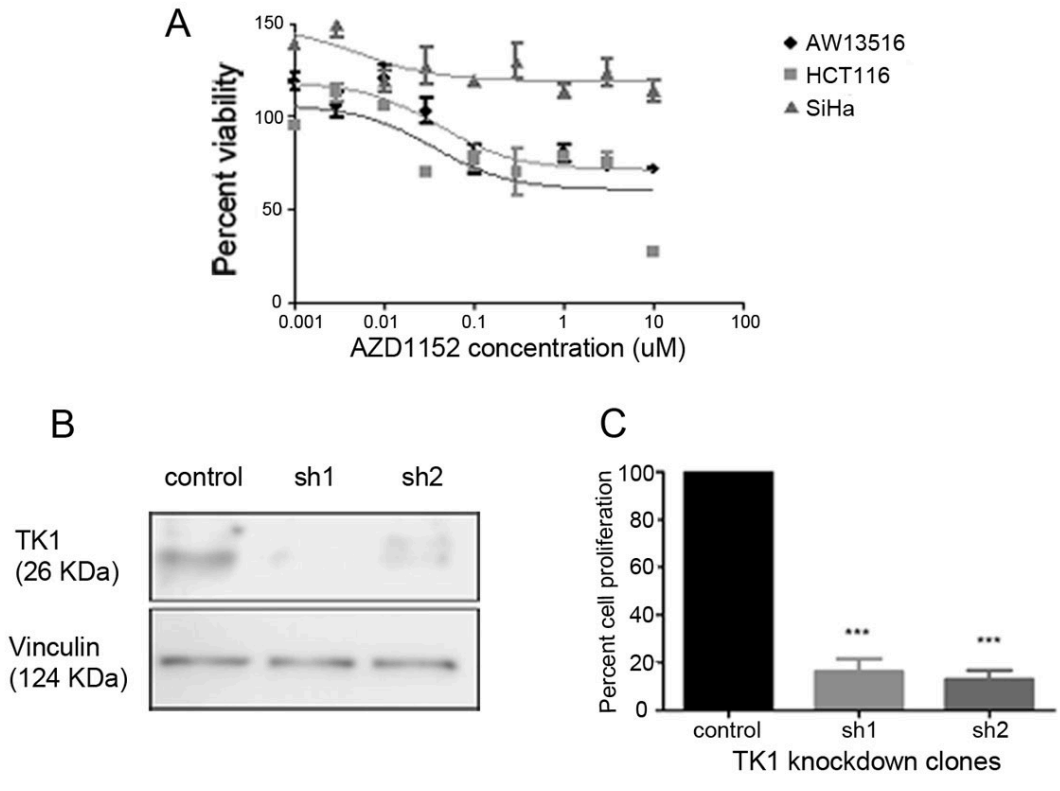
Patients with *AURKB* alterations show a poor overall survival

To assess the impact of *AURKB* alterations on the survival of patients, we accessed gene alteration data for *AURKB* and *TK1* from cBioPortal [15]. TCGA provisional HNSCC data sets comprising mutations, copy number changes and mRNA upregulation across 528 samples were analysed. Survival analysis using Kaplan-Meier plots suggests that patients with *AURKB* alteration display a poor survival of 18 months, compared to a survival of 56 months in the non-altered group (fig. 5A). The survivals of the *TK1*-altered and the non-altered cohorts were 22 and 56 months respectively (fig. 5B), suggesting poor survival in the TK1 genetic alteration group.

Discussion

Pooled shRNA screens are a powerful tool for the identification of specific gene targets that are essential for the survival of cancer cells. However, heterogeneous data sets often have limited reproducibility, as indicated by multiple

Figure 4: *AURKB* and *TK1* show oncogenic dependency in AW13516 cells. (A) MTT assay with AZD1152-HQPA inhibitor in AW13156, HCT116 and SiHa cells. (B) Knockdown confirmation of *TK1* in AW13516 cells by western blotting. (C) Cell proliferation assay in control and *TK1* knockdown clones of AW13516 cells.



studies, and several approaches are adopted to minimise the noise generated by non-reproducible hits [25]. Other factors that contribute to the variability and complexity of screen data are the effective delivery of shRNA, random integration for the stable expression of shRNA, processing of shRNA hairpins into silencing complexes, and off-target effects [26]. Therefore, to overcome these limitations due to variability in the reproducibility of the data, several robust computational approaches have emerged [27, 28]. Some analysis methods integrate genomic data such as gene expression and copy number information to provide insights into and predict essential genes in cancer [12, 13].

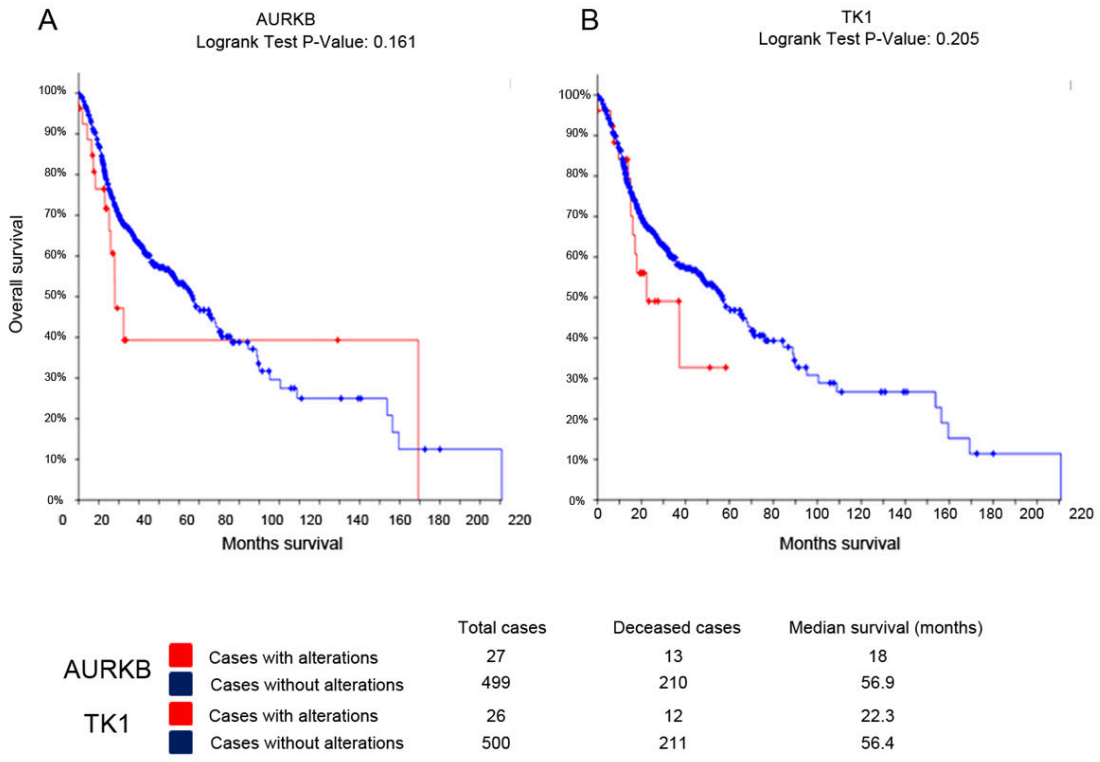
Although several data integration tools and packages for analysing the dataset from the screen are available, most have their specific third-party needs and necessitate intense computational infrastructure that cannot be run by researchers without specialised and advanced computational expertise. Thus, the lack of a simplified scoring system allowing a functional biologist to rank genes from the screen data by integrating the genomics data remains a limitation. To address this, we have developed a scoring system, DepRanker, which calculates a Rank Impact Score for each gene identified in the screen by considering the gene expression and copy number data.

Two different screens in AW13516 cells were analysed using different approaches, and were also sequenced at different depths. Because of the major differences in the overall capture of the libraries, we expected the results from the screens to be different. Since neither of the screens were performed at a high enough saturation, we analysed the data following separate protocols. The candidate genes *AURKB* and *TK1,* identified from both the screens using an integrated genomics approach, were validated by inhibitor and knockdown assays. DepRanker is a step towards reducing noise due to differences in the capture of libraries, sequencing depth and analysis methods. This approach can be useful specifically for identifying dependencies in cell lines.

AURKB and *TK1* are reported to have oncogenic functions in several cancer types, including HNSCC. A previous study on p53 mutant HNSCC cell lines using a kinome screen was also able to identify certain aurora kinases and thymidine kinases as therapeutic targets [29], which is consistent with our findings.

AW13516 cells display high copy amplification and gene expression of the *AURKB* gene. Overexpression or amplification of *AURKB* has been reported in several cancer types [16, 18]. *AURKB* is a chromosomal protein involved in the segregation of chromosomes and cytokinesis [30], and its overexpression leads to aneuploidy in the cells. It is also associated with aggressive tumour progression [31]. There are several pieces of evidence that point towards the oncogenic role of *AURKB* in head and neck cancer. High *AURKB* expression has been observed to be associated with increased cell proliferation and lymph node metastasis [32], involved in the activation of the RAS-MAPK pathway, and contributing to cetuximab resistance [33]. Also, *AURKB* is one of the essential genes most commonly identified from pooled RNAi and CRISPR screens on cancer cell lines, as identified by a search for this gene in the DepMap portal [12]. We observed that AW13516 cells were sensitive to the *AURKB* inhibitor AZD1152-HQPA.

Figure 5: Kaplan-Meier survival analysis in TCGA-HNSCC dataset.



In addition, survival analysis of TCGA HNSCC data indicated that patients with *AURKB* genes alterations display poor overall survival, which suggests it plays a role in carcinogenesis in HNSCC. *AURKB* is a potential therapeutic target for the treatment of HNSCC, and several *AURKB* inhibitors are in clinical trials [16, 34].

Similarly, the *TK1* target identified in the screen also exhibited high copy gain and increased gene expression in AW13516 cells. Thymidine kinase 1 (*TK1*) has a role in regulating the cell cycle [22]. Serum *TK1* levels are used to determine disease prognoses and to predict treatment outcomes [23]. A study of head and neck cancer showed that patients treated with chemotherapy and surgery showed decreased serum *TK1* levels, whereas patients with stable disease displayed elevated *TK1* levels. Hence, *TK1* can be used as a biomarker to evaluate disease outcomes [24, 35]. We functionally validated another target, *TK1*, using a knockdown approach. A significant difference in the proliferation rate was observed in *TK1* knockdown clones compared to control cells, suggesting that *TK1* is essential for the survival of cells. Also, a previous study from our lab identified significant up-regulation of *TK1* expression in tongue tumours [36].

In conclusion, we developed a data integration and scoring system, DepRanker, which uses the output of shRNA screen analysis packages (like ROAST, RIGER and Chimera) and integrates this with other genomics datasets to compute an integration score, known as a Rank Impact Score (RIS), for each gene. We performed a pooled RNAi screen against 906 kinase genes and, using the DepRanker, integrated the outcome with gene expression and copy number data for AW13516 cells to identify *AURKB* and *TK1* as essential genes in oral cancer.

Acknowledgements

We would like to thank all members of the Dutt laboratory for reviewing the manuscript, Medgenome Labs Ltd. for providing sequencing services and Dr Dulal Panda from IIT Mumbai for providing AZD1152-HQPA inhibitor.

Financial disclosure

The project was funded by the Department of Biotechnology (DBT), Govt of India (BT/PR2372/AGR/36/696/2011) to AD. TT is supported by a senior research fellowship from DBT. The study sponsors had no role in the study design, data analysis, decision to publish or preparation of the manuscript.

Potential competing interests

The authors declare that they have no competing interests.

References

1 Sadikovic B, Al-Romaih K, Squire JA, Zielenska M. Cause and consequences of genetic and epigenetic alterations in human cancer. Curr Genomics. 2008;9(6):394–408. doi: http://dx.doi.org/10.2174/138920208785699580. PubMed.

2 Weinstein IB, Joe A, Felsher D. Oncogene addiction. Cancer Res. 2008;68(9):3077–80, discussion 3080. doi: http://dx.doi.org/10.1158/0008-5472.CAN-07-3293. PubMed.

3 Iqbal N, Iqbal N. Imatinib: a breakthrough of targeted therapy in cancer. Chemother Res Pract. 2014;2014:. doi: http://dx.doi.org/10.1155/2014/357027. PubMed.

4 Smith J. Erlotinib: small-molecule targeted therapy in the treatment of non-small-cell lung cancer. Clin Ther. 2005;27(10):1513–34. doi: http://dx.doi.org/10.1016/j.clinthera.2005.10.014. PubMed.

5 Campeau E, Gobeil S. RNA interference in mammals: behind the screen. Brief Funct Genomics. 2011;10(4):215–26. doi: http://dx.doi.org/10.1093/bfgp/elr018. PubMed.

6 Dai Z, Sheridan JM, Gearing LJ, Moore DL, Su S, Wormald S, et al. edgeR: a versatile tool for the analysis of shRNA-seq and CRISPR-Cas9 genetic screens. F1000 Res. 2014;3:95. doi: http://dx.doi.org/10.12688/f1000research.3928.2. PubMed.

7 Luo B, Cheung HW, Subramanian A, Sharifnia T, Okamoto M, Yang X, et al. Highly parallel identification of essential genes in cancer cells. Proc Natl Acad Sci USA. 2008;105(51):20380–5. doi: http://dx.doi.org/10.1073/pnas.0810485105. PubMed.

8 König R, Chiang CY, Tu BP, Yan SF, DeJesus PD, Romero A, et al. A probability-based approach for the analysis of large-scale RNAi screens. Nat Methods. 2007;4(10):847–9. doi: http://dx.doi.org/10.1038/nmeth1089. PubMed.

9 Wu D, Lim E, Vaillant F, Asselin-Labat ML, Visvader JE, Smyth GK. ROAST: rotation gene set tests for complex microarray experiments. Bioinformatics. 2010;26(17):2176–82. doi: http://dx.doi.org/10.1093/bioinformatics/btq401. PubMed.

10 Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res. 2012;40(17):. doi: http://dx.doi.org/10.1093/nar/gks461. PubMed.

11 McFarland JM, Ho ZV, Kugener G, Dempster JM, Montgomery PG, Bryan JG, et al. Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. Nat Commun. 2018;9(1):4610. doi: http://dx.doi.org/10.1038/s41467-018-06916-5. PubMed.

12 Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, et al. Defining a Cancer Dependency Map. Cell. 2017;170(3):564–576.e16. doi: http://dx.doi.org/10.1016/j.cell.2017.06.010. PubMed.

13 Guan Y, Li T, Zhang H, Zhu F, Omenn GS. Prioritizing predictive biomarkers for gene essentiality in cancer cells with mRNA expression data and DNA copy number profile. Bioinformatics. 2018;34(23):3975–82. doi: http://dx.doi.org/10.1093/bioinformatics/bty467. PubMed.

14 Chandrani P, Upadhyay P, Iyer P, Tanna M, Shetty M, Raghuram GV, et al. Integrated genomics approach to identify biologically relevant alterations in fewer samples. BMC Genomics. 2015;16(1):936. doi: http://dx.doi.org/10.1186/s12864-015-2138-4. PubMed.

15 Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6(269):pl1. doi: http://dx.doi.org/10.1126/scisignal.2004088. PubMed.

16 Dar AA, Goff LW, Majid S, Berlin J, El-Rifai W. Aurora kinase inhibitors--rising stars in cancer therapeutics? Mol Cancer Ther. 2010;9(2):268–78. doi: http://dx.doi.org/10.1158/1535-7163.MCT-09-0765. PubMed.

17 González-Loyola A, Fernández-Miranda G, Trakala M, Partida D, Samejima K, Ogawa H, et al. Aurora B Overexpression Causes Aneuploidy and p21Cip1 Repression during Tumor Development. Mol Cell Biol. 2015;35(20):3566–78. doi: http://dx.doi.org/10.1128/MCB.01286-14. PubMed.

18 Tang A, Gao K, Chu L, Zhang R, Yang J, Zheng J. Aurora kinases: novel therapy targets in cancers. Oncotarget. 2017;8(14):23937–54. doi: http://dx.doi.org/10.18632/oncotarget.14893. PubMed.

19 Gully CP, Velazquez-Torres G, Shin JH, Fuentes-Mattei E, Wang E, Carlock C, et al. Aurora B kinase phosphorylates and instigates degradation of p53. Proc Natl Acad Sci USA. 2012;109(24):E1513–22. doi: http://dx.doi.org/10.1073/pnas.1110287109. PubMed.

20 Tao Y, Zhang P, Girdler F, Frascogna V, Castedo M, Bourhis J, et al. Enhancement of radiation response in p53-deficient cancer cells by the Aurora-B kinase inhibitor AZD1152. Oncogene. 2008;27(23):3244–55. doi: http://dx.doi.org/10.1038/sj.onc.1210990. PubMed.

21 Ha GH, Breuer EK. Mitotic Kinases and p53 Signaling. Biochem Res Int. 2012;2012:. doi: http://dx.doi.org/10.1155/2012/195903. PubMed.

22 Bagegni N, Thomas S, Liu N, Luo J, Hoog J, Northfelt DW, et al. Serum thymidine kinase 1 activity as a pharmacodynamic marker of cyclin-dependent kinase 4/6 inhibition in patients with early-stage breast cancer receiving neoadjuvant palbociclib. Breast Cancer Res. 2017;19(1):123. doi: http://dx.doi.org/10.1186/s13058-017-0913-7. PubMed.

23 Ning S, Wei W, Li J, Hou B, Zhong J, Xie Y, et al. Clinical significance and diagnostic capacity of serum TK1, CEA, CA 19-9 and CA 72-4 levels in gastric and colorectal cancer patients. J Cancer. 2018;9(3):494–501. doi: http://dx.doi.org/10.7150/jca.21562. PubMed.

24 Zhou J, He E, Skog S. The proliferation marker thymidine kinase 1 in clinical use. Mol Clin Oncol. 2013;1(1):18–28. doi: http://dx.doi.org/10.3892/mco.2012.19. PubMed.

25 Schaefer C, Mallela N, Seggewiß J, Lechtape B, Omran H, Dirksen U, et al. Target discovery screens using pooled shRNA libraries and next-generation sequencing: A model workflow and analytical algorithm. PLoS One. 2018;13(1):. doi: http://dx.doi.org/10.1371/journal.pone.0191570. PubMed.

26 Fellmann C, Lowe SW. Stable RNA interference rules for silencing. Nat Cell Biol. 2014;16(1):10–8. doi: http://dx.doi.org/10.1038/ncb2895. PubMed.

27 Dempster JM, Pacini C, Pantel S, Behan FM, Green T, Krill-Burger J, et al. Agreement between two large pan-cancer CRISPR-Cas9 gene dependency data sets. Nat Commun. 2019;10(1):5817. doi: http://dx.doi.org/10.1038/s41467-019-13805-y. PubMed.

28 Shao DD, Tsherniak A, Gopal S, Weir BA, Tamayo P, Stransky N, et al. ATARiS: computational quantification of gene suppression phenotypes from multisample RNAi screens. Genome Res. 2013;23(4):665–78. doi: http://dx.doi.org/10.1101/gr.143586.112. PubMed.

29 Moser R, Xu C, Kao M, Annis J, Lerma LA, Schaupp CM, et al. Functional kinomics identifies candidate therapeutic targets in head and neck cancer. Clin Cancer Res. 2014;20(16):4274–88. doi: http://dx.doi.org/10.1158/1078-0432.CCR-13-2858. PubMed.

30 Smith SL, Bowers NL, Betticher DC, Gautschi O, Ratschiller D, Hoban PR, et al. Overexpression of aurora B kinase (AURKB) in primary non-small cell lung carcinoma is frequent, generally driven from one allele, and correlates with the level of genetic instability. Br J Cancer. 2005;93(6):719–29. doi: http://dx.doi.org/10.1038/sj.bjc.6602779. PubMed.

31 Hegyi K, Egervári K, Sándor Z, Méhes G. Aurora kinase B expression in breast carcinoma: cell kinetic and genetic aspects. Pathobiology. 2012;79(6):314–22. doi: http://dx.doi.org/10.1159/000338082. PubMed.

32 Mehra R, Serebriiskii IG, Burtness B, Astsaturov I, Golemis EA. Aurora kinases in head and neck cancer. Lancet Oncol. 2013;14(10):e425–35. doi: http://dx.doi.org/10.1016/S1470-2045(13)70128-1. PubMed.

33 Boeckx C, Op de Beeck K, Wouters A, Deschoolmeester V, Limame R, Zwaenepoel K, et al. Overcoming cetuximab resistance in HNSCC: the role of AURKB and DUSP proteins. Cancer Lett. 2014;354(2):365–77. doi: http://dx.doi.org/10.1016/j.canlet.2014.08.039. PubMed.

34 Falchook GS, Bastida CC, Kurzrock R. Aurora Kinase Inhibitors in Oncology Clinical Trials: Current State of the Progress. Semin Oncol. 2015;42(6):832–48. doi: http://dx.doi.org/10.1053/j.seminoncol.2015.09.022. PubMed.

35 Chen Y, Ying M, Chen Y, Hu M, Lin Y, Chen D, et al. Serum thymidine kinase 1 correlates to clinical stages and clinical reactions and monitors the outcome of therapy of 1,247 cancer patients in routine clinical settings. Int J Clin Oncol. 2010;15(4):359–68. doi: http://dx.doi.org/10.1007/s10147-010-0067-4. PubMed.

36 Upadhyay P, Gardi N, Desai S, Chandrani P, Joshi A, Dharavath B, et al. Genomic characterization of tobacco/nut chewing HPV-negative early stage tongue tumors identify MMP10 asa candidate to predict metastases. Oral Oncol. 2017;73:56–64. doi: http://dx.doi.org/10.1016/j.oraloncology.2017.08.003. PubMed.

Appendix 1

The DepRanker scoring system

Appendix 2

Supplementary tables

Table S1: Primary and secondary PCR primer sequences. Highlighted in bold are the unique 6-base index sequences of the secondary PCR primers.

Table S2: ROAST ranking of all the depleted kinases in both screens.

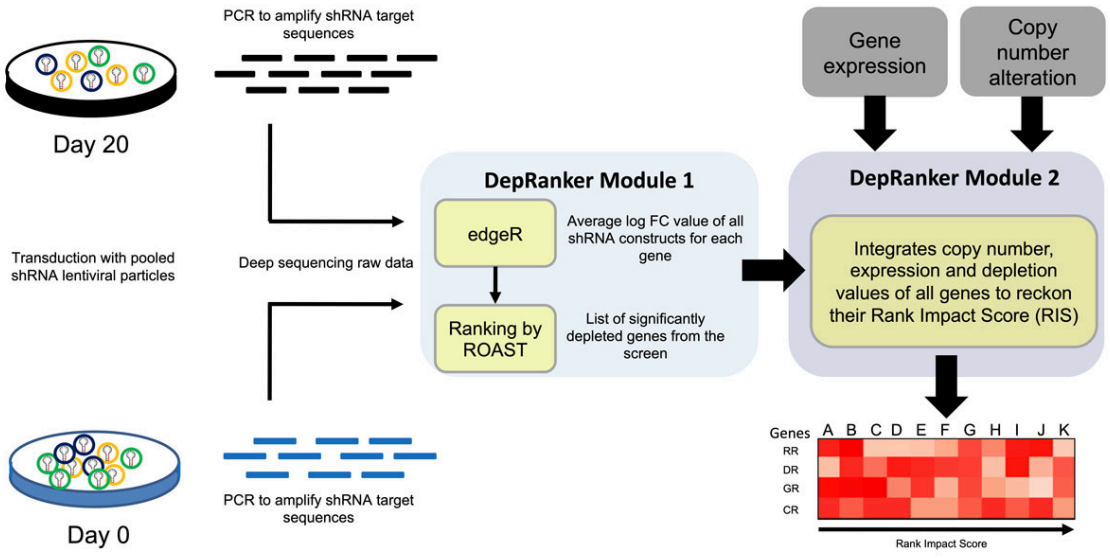
Table S3: Integrated copy number, gene expression and logFC value data for each kinase for both screens are shown.

Table S4: A list of the top depleted kinases from the screen, identified by considering the cumulative effect of four parameters: gene rank (RS), copy number alteration (CS), gene expression (GS) and logFC value of shRNA depletion (DS). The cumulative effect is represented by the Rank Impact Score (RIS) and a weighting.

Table S5: The Rank Impact Scores (RIS) and weightings (W) of all the kinases in each of the two screens are shown. The values of all four parameters, rank (RS), copy number (CS), gene expression (GS) and logFC value (DS) are shown.

This appendix is available in a separate file at https://smw.ch/article/doi/smw.2020.20195.

Figure S1: Schematic outline depicting the DepRanker, an automated tool to identify driver genes from pooled functional genomic screens. CR = copy number alteration rank; DR = depletion rank; FC = fold change; GR = gene expression rank; RR = ROAST rank



LIST OF FIGURES

LIST OF TABLES