Integrated Analysis of Head and Neck Squamous Cell Carcinoma: A Genomics Approach

By

Pawan Kumar Upadhyay

[LIFE09201104001]

Tata Memorial Centre

Mumbai

A thesis submitted to the
Board of Studies in Life Sciences
In partial fulfilment of requirements
for the Degree of

DOCTOR OF PHILOSOPHY

of

HOMI BHABHA NATIONAL INSTITUTE

May 2017

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 Mr. Pawan Kumar Upadhyay entitled "Integrated analysis of Head and Neck Squamous Cell Carcinoma: A Genomics Approach" and recommend that it may be accepted as fulfilling the thesis requirement for the award of Degree of Doctor of Philosophy.

Chairman – Dr. Sohab N. Dalal S. N Dalal Date: 24/5/17

Guide/Convener – Dr. Amit Dutt A Dutt Date: 24/5/17

External Examiner – Prof. Tapas K. Kundu Date: 24/5/17

Member 1 – Dr. Pritha Ray Pritha Ray Date: 24/5/17

Member 2 – Dr. Shilpee Dutt Shilpee Dutt Date: 24/5/17

Invitee – Dr. Nagaraj Balasubramanian Date: 25/5/17

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: 24th May 2017

Place: Navi Mumbai

A Dutt 24/5/17

Dr. Amit Dutt

Guide

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfilment 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.

Navi Mumbai Pawan Kumar Upadhyay

Date: 24th May 2017

DECLARATION

I, hereby declare that the investigation presented in the thesis has been carried out by me. This 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.

Navi Mumbai Pawan Kumar Upadhyay

Date: 24th May 2017

LIST OF PUBLICATIONS ARISING FROM THE THESIS

Journals:

Peer-reviewed papers:

1. "Notch pathway activation is essential for maintenance of stem-like cells in early tongue cancer". **Upadhyay P***, Nair S*, Kaur E, Aich J, Dani P, Sethunath V, Gardi N, Chandrani P, Godbole M, Sonawane K, Prasad R, Kannan S, Agarwal B, Kane S, Gupta S, Dutt S, Dutt A. *Oncotarget*. 2016 Jul 6. doi: 10.18632/oncotarget.10419. [Epub ahead of print] PubMed PMID: 27391340.
2. "TMC-SNPdb: an Indian germline variant database derived from whole exome sequences". **Upadhyay P**, Gardi N*, Desai S*, Sahoo B, Singh A, Togar T, Iyer P, Prasad R, Chandrani P, Gupta S, Dutt A. *Database (Oxford)*. 2016 Jul 9; 2016. pii: baw104. doi: 10.1093/database/baw104. Print 2016. PubMed PMID: 27402678.
3. "Integrated genomics approach to identify biologically relevant alterations in fewer samples". Chandrani, P*., **Upadhyay P*.,** Iyer P., Tanna M., Shetty M., Raghuram G. V., Oak N., Singh A., Chaubal R., Ramteke M., Gupta S., Dutt A. *BMC Genomics* (2015), 16, 936. PubMed PMID: 26572163; PubMed Central PMCID: PMC4647579.
4. "Distinct genomic alterations underlie tobacco/ nut chewing HPV-negative early stage tongue tumors". **Pawan Upadhyay,** Nilesh Gardi, Asim Joshi, Sanket Desai, Pratik Chandrani, Prachi Terwadkar, Bhasker Dharavath, Manoj Ramteke, Priyancka Arora, Rohan Chaubal, Munita Bal, Sudhir Nair, Amit Dutt; 2017.

Chapters in books and lecture notes: NA

Conferences:

1) "TMC-SNPdb: an Indian germline variant database derived from whole exome sequences" India International Science Festival, New Delhi (Dec 2016). **Pawan Upadhyay,** Nilesh Gardi, Sanket Desai, Bikram Sahoo, Ankita Singh, Trupti Togar, Prajish Iyer, Ratnam Prasad, Pratik Chandrani, Sudeep Gupta, Amit Dutt. **(Travel Award)**
2) "OF50-003 Pro-oncogenic role of NOTCH1 in early tongue squamous cell carcinoma", New Ideas in Cancer – Challenging Dogmas (Feb 2016) Mumbai, India. **P. Upadhyay** , S. Nair , E. Kaur , J. Aich , P. Dani , V. Sethunath , N. Gardi , P. Chandrani , M. Godbole , K. Sonawane , S. Kannan , B. Agarwal , S. Kane , S. Dutt A., Europian Journal of Cancer,10.1016/S0959-8049(16)31953-0 (**top abstracts for oral presentation**).
3) "Abstract 5161: Discovery of a matrix metalloproteinase MMP10 as a clinically relevant biomarker to predict lymph node metastasis in tongue squamous cell carcinoma". **Pawan Upadhyay**, Nilesh Laxman Gardi, Hemant Ramesh Dhamne, Kavitha Sonawane, Anil D'Cruz, Sudhir Nair, Amit Dutt; Proceedings of the 106th Annual Meeting of the American Association for Cancer Research, Philadelphia, PA. USA; 04/2015.

4) "Abstract 4811: Pro-oncogenic role of NOTCH1 in early tongue squamous cell carcinoma". **Pawan Upadhyay,** Jyotirmoi Aich, Vidyalakshmi Sethunath, Prachi Dani, Sadhana Kannan, Pratik Chandrani, Kavitha Sonawane, Beamon Agarwal, Shubhada Kane, Sudhir Nair, Amit Dutt; AACR 106th Annual Meeting 2015, Philadelphia, USA; 08/2015.
5) "P220-K020: Integrated genomics analysis of early stage tongue squamous cell carcinoma patients", New Ideas in Cancer – Challenging Dogmas (Feb 2016), Mumbai, India. **P. Upadhyay ,** N. Gardi , P. Chandrani , S. Desai , A. Joshi, S. Nair, A. Dutt Europian Journal of Cancer,10.1016/S0959-8049(16)31953-0.
6) "Abstract C154: Integrated genomics approach to identify driver alterations". Pratik Chandrani*, **Pawan Upadhyay*,** Prajish Iyer, Mayur Tanna, Madhur Shetty, Raghuram Venkata Gorantala, Ninad Oak, Ankita Singh, Rohan Chaubal, Manoj Ramteke, Sudeep Gupta, Amit Dutt; AACR-NCI-EORTC International Conference: Molecular Targets and Cancer Therapeutics, Boston, MA; 11/2015, USA.
7) "Pro-oncogenic Role of NOTCH1 in Early Tongue Squamous Cell Carcinoma" at 1ST MOSCON (Molecular Oncology Society) conference (January 29th–30th 2016). **Pawan Upadhyay,** Sudhir Nair, Ekjot Kaur, Jyotirmoi Aich, Prachi Dani, Vidyalakshmi Sethunath, Nilesh Gardi, Pratik Chandrani, Mukul Godbole, Kavita Sonawane, Sadhana Kannan, Beamon Agarwal, Shubhada Kane, Shilpee Dutt, Amit Dutt.
8) "Aberrant activation of NOTCH pathway components in tongue squamous cell carcinoma". **Pawan Upadhyay,** Jyoti Aich, Prachi Dani, Vidya Sethunath, Kavitha Sonawane, Sadhana Kannan, Beamon Agrawal, Sudhir Nair and Amit Dutt 34th Annual Convention of Indian Association for Cancer Research (IACR-2015), 19th-21st, February, JAIPUR, India.
9) "Development of gene signature profile predicting nodal metastasis in squamous cell carcinoma of tongue" 2015 NextGen Genomics, Biology, Bioinformatics and Technologies (NGBT) Conference 1st-3rd October 2015, HICC, Hyderabad, India **Pawan Upadhyay,** Nilesh Laxman Gardi, Kavitha Sonawane, Anil D'Cruz, Sudhir Nair, Amit Dutt. .
10) "Integrated Genomic Characterization of Head & Neck Cancer Cell Lines Derived From Indian Patients" in 34th Convention of Indian Association for Cancer Research Emerging Trends in Cancer research: Road to prevention & cure & International Symposium on: Infection and Cancer (IACR-2013 - February 13 -16 2013) New Delhi, India. **(top abstracts for oral presentation under IACR Sitaram Joglekar award and Mangala Bamne award category)**
    Pawan Upadhyay, Pratik Chandrani, Rohan Chaubal, Ninad Oak, Vaibhav Kulkarni,Madhur Shetty, Maulik Vyas, Raghuram Gorantla, and Amit Dutt
11) "Integrated Genomic Characterization of Head & Neck Cancer Cell Lines Derived from Indian Patients" at The Global Cancer Genomics Consortium Second Annual Symposium: Genomics Medicine in Cancer Research, 2012, ACTREC, India. **(Recipient of best poster presentation award). Pawan Upadhyay,** Pratik

Chandrani, Rohan Chaubal, Ninad Oak, Madhur Shetty, Kunal Karve, Ratnam Prasad, Prajish Iyer and Amit Dutt.

Other:

1. "Mutations Are Associated with Spindle cell and Sclerosing Rhabdomyosarcomas with Aggressive Clinical Outcomes". Bharat Rekhi*, **Pawan Upadhyay*,** Manoj P. Ramteke Dutt A. MYOD1 (L122R)*; Modern Pathology, 2016*; DOI: 10.1038/modpathol.2016.144. PubMed PMID: 27562493.
2. "Applications of next-generation sequencing in cancer. Special Section: Cancer". **Upadhyay P**, Dwivedi R, Dutt A. *Current Science*, September 2014, vol. 107, no. 5, 10.

3. "NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome". Chandrani P*, Kulkarni V*, Iyer P, **Upadhyay P**, Chaubal R, Das P, Mulherkar R, Singh R, Dutt A. *Br J Cancer*; 2015 Jun 9;112(12):1958-65. doi: 10.1038/bjc.2015.121. Epub 2015 May 14. PubMed PMID: 25973533; PubMed Central PMCID: PMC4580395.
4. "Circulating nucleic acids damage DNA of healthy cells by integrating in to their genomes". Indraneel Mitra, Naveen Khare, Gorantla Venkata Raghuram, Rohan Chaubal, Fatema Khambatti, Deepika Gupta ,Ashwini Gaikwad ,Preeti Prasannan, Akshita Singh, Aishwarya Iyer, Ankita Singh ,**Pawan Upadhyay**, Naveen Kumar Nair, Pradyumna Kumar Mishra, Amit Dutt. *J. Biosci.* March 2015, 40(1), PubMed PMID: 25740145

5. "Frequency of EGFR mutations in 907 lung adenocarcioma patients of Indian ethnicity". Chougule A, Prabhash K, Noronha V, Joshi A, Thavamani A, Chandrani P, **Upadhyay P**, Utture S, Desai S, Jambhekar N, Dutt A. *PLoS One*. 2013 Oct 4;8(10):e76164. doi: 10.1371/journal.pone.0076164. eCollection 2013. PubMed PMID: 24124538; PubMed Central PMCID: PMC3790706.
6. "Drug-sensitive FGFR3 mutations in lung adenocarcinoma". P. Chandrani, K. Prabhash, A. Chaughule, R. Prasad, V. Sethunath ,M. Ranjan ,J. Aich , P. Iyer, H. Dhamne, D. N. Iyer, **P. Upadhyay**, B. Mohanti, P. Chandna, R. Kumar, A. Joshi, V. Noronha, V. Patil, A. Ramaswamy, A. Karpe, R. Thorat, P. Chaudhari, A. Ingle & Amit Dutt. *Ann Oncol* (2017) 28 (3): 597-603, doi: 10.1093/annonc/mdw636, PubMed PMID: 27998968; PubMed Central PMCID: PMC5391708

7. "Cell-free chromatin from dead cancer cells integrate into genomes of bystander healthy cells to induce DNA damage and inflammation". Urmila Samant, Suvarna Sharma, Raghuram GV, Deepika Gupta, Preeti Prasannan, Ashwini Gaikwad, Pritishkumar Tidke, Nilesh Gardi, Rohan Chaubal, **Pawan Upadhyay,** Tannistha Saha, Pradyumna Mishra, Naveen Khare, Naveen Nair, Amit Dutt. **(*Cell Death Discovery, In press*).**

8. "Proline 152 mutation in p53 abolishes cognate DNA binding, induces gain of function tumorigenesis". Manoj Kumar*, Siddharth Singh*, Vivek Singh Tomar, **Pawan Upadhyay**, Amit Dutt and Tapas K. Kundu**.**

ACKNOWLEDGEMENT

Pursuing Ph.D was one of very rare easy decision I made. It rendered me an opportunity to associate with many great people who boarded with me on this journey and helped to face the fascinating challenges and together transforming them into cherishable one. Today, I would like to thank all of you on this occasion.

I would like to express my deepest gratitude for my mentor, Dr. Amit Dutt, for his constant guidance, motivation and support in all aspects of my career throughout my tenure. It was his incredible passion and infectious enthusiasm for the science has motivated me to produce new scientific knowledge, allowed me to develop my own research interest and fuelling my personal as well as professional development. His constructive, sharp and extremely critical supervision has always encouraged me to perform interdisciplinary research in cancer and develop deep scientific temperament. He constantly encouraged me to explore novel ideas with full support and helped me during difficult time in my tenure. I am highly privileged to you for allowing me to actively participating in such an endeavour.

I am thankful to doctoral committee members, Dr. Sorab N. Dalal (chairperson), Dr. Shilpee Dutt, Dr. Pritha Ray and Dr. Nagaraj balasubramanian for their continuous critical advice and constructive suggestion during doctoral committee meetings. Advices made during DC meetings were invaluable to my research work and has allowed me to progress in focused manner. I am also thankful to Dr. Girish Maru (ex-chairperson), Venu Bommireddy (ex-member) for their advice in my thesis work.

It was privileged to utilize the excellent infrastructure and facilities at ACTREC and Tata Memorial Centre for which I would like to thank Dr. Rajendra Badwe (Director, TMC), Dr. Shubhada Chiplunkar (Director, ACTREC), Dr. Sudeep Gupta (Deputy Director, ACTREC), Dr. Rajiv Sarin (Ex-Director, ACTREC), Dr. Surekha Zingde (Ex-Deputy Director, ACTREC). I would also like to thank CSIR for providing me regular fellowship, Wellcome Trust/DBT India Alliance, Department of Biotechnology, India and Tata Memorial Centre for funding my research work. I am appreciative to Homi Bhabha National Institute (HBNI) and Tata Memorial Centre (TMC)- Sam Mistry foundation for proving financial support to present my work at an international meeting.

I am grateful to my clinical collaborators Dr. Sudhir Nair, Dr. Kumar Prabhash, Dr. Sudeep Gupta and Dr. Bharat Rekhi. I am thankful to Dr. Nair and his surgical oncology team (Kavita, Priyancka and Swapnil) for their dedication and active support for providing the tongue tissue samples and well annotated clinical information of patients. I am grateful to get an opportunity to me conduct molecular research in early tongue cancer. I am also thankful to Dr. Kumar for agreeing to be external reviewer of my thesis work and providing me an opportunity to associate with his group to learn new things. Dr.

Gupta has taught me different approaches to analyse and interpret the clinical correlation data. I learned from his critical way of analysing and interpreting the survival data. Also, he was always open to provide the valuable insight to our manuscripts. Working with Dr. Rekhi on rare cancer (rhabdomyosarcoma) has been one of amazing learning experience in my tenure. I am also thankful to Dr. Shubhada Kane, Dr. Beamon Agarwal, Dr. Munita Menon for training and helping me with histopathological and IHC analysis. I am also thankful to Dr. Kishor Amin (Tissue Biorepository) for helping me during initial days and providing open access to genomics related facility. I also thank Ms. Sadhana Kannan for helping me the biostatical analysis and interpretation of clinical data. Working with her was a great learning experience.

I would like to express my deepest appreciation to Dr. Shilpee Dutt for her valuable advice during lab meetings as well as doctoral committee meetings, which helped me widen my research interests. Her open criticism and constructive suggestion in several aspects of my work allowed me to embark on cancer stem cell related work in my thesis and widen my knowledge.

I have also got opportunity work with several esteemed collaborators during my tenure; Prof. Indraneel Mitra (ACTREC-TMC), Prof. Annapoorni Rangarajan (IISc), Prof. Sudhir Krishnan (JNCASR). I am thankful to Prof. Mitra for allowing me to learn and contribute in area of chromatin biology. He his zeal and passion towards science is really inspiring for me and interaction with him has always been thought provoking and stimulating. I would like to thank members of Prof. Mitra Laboratory (Dr. Raghuram G.V., Naveen, Ashwani) for always being supportive and work in collaborative manner. Prof. Rangarajan and Prof. Krishnan has always been supportive and open to share various reagents related NOTCH1 functional analysis project.

I am exceptionally grateful to all the patients for agreeing to be a part of my thesis work and helping me to conduct research.

I would like to thank the various ACTREC facilities staffs for their assistance. I am thankful to Mr. Dandekar sir for his dexterous assistance in common instrument facility. Additionally help from staff in sequencing, microscopy, flow cytometry, sequencing, IT, library, steno-pool, administration, account, canteen, and engineering and security personals. It’s due to their efforts I could get all the facilities running smoothly and organized manner. A special thank you to Mr. Naresh, Sharda Mam, Annand sir, Tanuja, Jayraj, M.A. Sharma, Chitra, Aruna Mam, Karan sir, Arun Shetty for always being ready to help in administrative work.

I express my humble thank you to all the past and present members of Shilpee Laboratory for providing such a nurturing atmosphere and contributing to my personal and professional growth. A special thanks to Ekjot, Sameer, Jacinth, Priya, Sailesh, Tejal, Swati, Samreen, Neha, Aloka for being such a nice

friend, supportive at several occasions and for thoughtful discussions about scientific and non-scientific stuffs. A special thanks to Smita Mam and Ganesh sir for various technical support during my tenure.

The Dutt laboratory members have immensely contributed to my personal and professional growth by providing an amazing fertile ground and creating an interdisciplinary research ecosystem in the lab. I am thankful to Pratik, Prajish, Trupti, Mukul, Ratnam, Malika, Asim, Sanket, Bikram, Rohan, Nilesh, Prachi, Vidya, Mayur for their constant support in my work and motivation during difficult times. I am especially thankful to Pratik, Prajish, Rohan, Nilesh, Sanket for thoughtful discussion about the computational work and serving as catalyst for honing my computational skills and working together at several challenging projects. Special thanks to Ratnam and Renu Mam for being supportive backbone of lab by managing the IRB and administrative stuffs. I am also thankful to Prajish and Trupti for being always supportive in ups and downs of my life and equally criticizing me to improvise at various level during this tenure. I would also like to thank Kanishka, Ankita, Sunil, Reshma, Sharan, Madhur, Ninad, Rashmi, Abhishek, Maulik, Renu, Vaibhav, Deepak Iyer, Naren, Kunal for being great friends and assisting me at various occasions during my tenure. I really indebted to your constant support and meaningful interactions in and out of the lab. I am thankful to Post-doctoral fellows at Dutt Laboratory, Dr. Kuldeep, Dr. Jyoti, Dr.Hemant, Dr. Manoj for training me in several experimental stuffs, constant advice, feedback in my research work and providing lively ecosystem for my growth as a researcher. I am thankful to members of lab, Bhashkar and Neelima for their timely help and through reading for my manuscripts. I am thankful to Deepak Amburle, Deepak Chauhan, Mr. Rane, Sayee Mam, Mr. Ganesh for being providing endless technical support and being exceedingly friendly.

I am thankful to my Ph.D batch mates, Rasika, Prasanna, Rahul, Sonali, Qadir, Divya, Sajad, Moquit for being supportive in all phases of tenure. We all shared several memorable moments together in an out of ACTREC campus which I will remember lifetime. I also would like to thank my roommates Dr. Surya, Rahul, Swapnil for being extremely helpful and motivating during difficult times and providing a nice atmosphere at home. ACTREC student community including my supportive seniors (Lalit, Ajit Chande, Ram, Snehal, Asif, Rajan, Polomi, Amir Ali, Kedar, Ajit sharma) and lovely juniors.

I would like to express deepest gratitude to the building block of my career i.e. my previous teachers (Dr. Satayaram Chaudhari, Dr. Manoj Sing, Dr. F.D. Yadav, Dr. N.N. Tiwai and Dr. V.K. Pandey, Dr. Farrukh Jamal, Dr. Vandana Ranjan, Dr. Sudhir Rai), who inspired me, introduced me to the basics philosophy of science, motivated me to pursue research career in science. I am also thankful to several senior and friends outside the lab, Dr. Prabhash, Late Sri Prakash, Pravesh, Vinay, Mithilesh, Raj Kumar, Ravi, Richa, Sandeep, Anktrish and Praveen, who motivated and provided strong support. I will always be indebted them for caring guidance and support in my life.

ACKNOWLEDGEMENT

I am immensely grateful to my parents for sacrificing all their life and supported me unconditionally in every choices I made in my career. Their love, support and encouragement has allowed me to follow my dreams in an uninterrupted manner with full passion and dedication. I am thankful to my brother (Deepak) for his constant background support.

Lastly, I would like to thank almighty (Krishna) for his absolute blessing, unlimited energy when things seemed impossible in my life. I am always feel grateful to Krishna for everything in my life.

Thank you.

24th May, 2016 Pawan Kumar Upadhyay

Homi Bhabha National Institute

SYNOPSIS OF Ph.D. THESIS

1. **Name of the Student:** Pawan Kumar Upadhyay
2. **Name of the Constituent Institution:** Tata Memorial Centre, ACTREC
3. **Enrolment No. and Date of Enrolment:** LIFE09201104001, 25th July 2011
4. **Title of the Thesis:** Integrated analysis of Head and Neck Squamous Cell Carcinoma: A Genomics Approach
5. **Board of Studies:** Life Science

SYNOPSIS

1. INTRODUCTION

Cancer is largely categorized as a somatic genetic disease. The cancer cell genome acquire a set of variations in the DNA sequences which is different from the normal cells counterpart, collectively termed as 'somatic mutations' to distinguish them from germline variations. These somatic changes includes single nucleotide changes, insertions, and deletion, copy number alterations or large structural rearrangements [1]. Genome sequencing has revolutionized our understanding of somatic genomic

alterations in cancer. Over the past decades, comprehensive sequencing efforts across the globe such as The Cancer Genome Atlas (TCGA: http://cancergenome.nih.gov) and International Cancer Genome Consortium (ICGC: https://icgc.org) have revealed the landscape of somatic genomic alterations of the common human cancers types in large number of samples [2].

Despite advances in sequencing technologies, it is a challenging task to catalog all bona fide somatically variants in cancer genome [3]. To obtain the complete catalog of somatic mutation, each human tumor need to be analyzed at higher coverage depending on normal tissue contamination and known genetic heterogeneity has been suggested [4, 5]. Thus, it is essential to subtract the germline variants from the ones obtained during the analysis to get the tumor specific bona fide somatic variants. Although, most of the germline variants common in human population (>5% allele frequency) have been cataloged in the public databases, there are myriad of rare inherited single nucleotide polymorphisms (SNPs) that are not and outnumber the number of somatic mutation present in the individual tumor [6]. Thus, to deplete the false positive somatic variants in cancer genome analysis demands development and application of ethnic-specific germline variant databases to filter out low allele frequency variants for the identification of bona fide somatic mutations.

Head and neck squamous cell carcinoma (HNSCC) is the sixth-most-common cancer worldwide, with about 600,000 new cases diagnosed every year, and includes cancer of the nose cavity, sinuses, lips, tongue, mouth, salivary glands, upper aerodigestive tract and voice box, and is often lethal with a five-year survival rate of 40-50% [7, 8]. The major risk factors known to be associated include use of tobacco related products, alcohol, and infection with high-risk human papillomaviruses (HPV). The HPV-positive and HPV-negative HNSCCs have been established as clinically and

molecularly separate entities and later exhibiting an increased resistance to therapy and poor prognosis [9, 10]. Recently several genome-wide characterization studies including TCGA, ICGC and others have described the common and unique molecular alternations associated with HPV-positive and negative HNSCC tumors [11-13].

A subsite in HNSCC, tongue squamous cell carcinoma (TSCC) accounts for two-thirds of all cancer cases of the HNSCC[14]. Several reports suggest an increase in incidence over recent years whereas five-year overall survival rates for TSCC remain low regardless of the advances in detection and treatment modalities worldwide, including in India[15, 16] and remains the poorest in terms of prognosis in HNSCC [17]. The genome sequencing studies carried out on TSCC have been restricted to single genes or gene panels with few genome-wide studies suggesting difference in specific genomic profile [18, 19]. Notably, studies till focused on HNSCC genomic characterization considered it as a single entity disease despite known existence of biological and genetic differences and clinical distinctiveness in different subsites, molecular changes associated with each site could vary[20, 21]. The presence of HPV infection in TSCC is almost absent, as shown by several studies from Asian population including Indian population [22-25].

Importantly, studies carried out till now in tongue cancer include the advanced stage tumor samples (pT3-pT4) and understanding of early Tongue (pT1-pT2) cancer genomic alterations is still dismal. Additionally, comprehensive efforts to analyze and catalogue genomic alterations such as mutations, copy number alterations and gene fusions in targetable genes in early TSCC is lacking.

2. RATIONALE

The identification of bona fide somatic mutation requires the subtraction of variants against matched normal sample followed by depletion with publically available

database (1000 genomes and dbSNP). However, due to inadequate representation of individuals from Indian populations in public SNP databases and unique genetic features, low allele frequency variants could not be filtered during somatic mutation analysis, posing huge challenge to identify the bona fide somatic variants.

Two large efforts has been taken to genomically characterize the HNSCC by TCGA and ICGC which defined the molecular alteration landscape of different sub-sites. The efforts by ICGC-India for genomic characterization of HNSCC had been mainly restricted to single sub-site i.e. gingiva-buccal and revealed distinct molecular alteration in these tumors as compared to Caucasian populations[12]. The HNSCC tumors has been shown to be highly heterogeneous and harbor distinct genomic alterations [21, 26]. Most of these studies have characterized the HNSCC tumors derived from Caucasian and Asian population as a single entity and site-specific genomic alterations remain to be explored. Despite being such high incidences in Indian subcontinent, lack of understanding of molecular alteration that govern HNSCC tumorigenesis hampers the ability to develop effective targeted therapeutics for head and neck cancer.

Similarly, the functional analysis of genomic alterations identified from the population-specific tumor genome characterization would require the genomically defined tumor-derived cancer cell lines from Indian origin. One of the objective of my work is to genomic characterization and functionally analyze biologically relevant alteration in Indian patient's derived HNSCC cell lines using various functional assays.

3. KEY QUESTIONS

1. Can we identify ethnic specific SNPs and develop a SNP database to deplete ethnic specific germline variants during cancer genome analysis?
2. What are the underlying somatic genomic alterations in human early tongue cancers?

3. How to elucidate the role of novel genes in head and neck squamous cell carcinoma identified using integrated genomic analysis?

4. OBJECTIVES

- **Objective-1: Integrated genomic characterization of HNSCC cell lines.**

The availability of genomically defined pre-clinical model systems such as patient tumor-derived HNSCC cell line has been shown to serve as an elegant experimental resource for systematic functional analysis of genomic alterations and provides an improved understanding of biological features and a more rational approach to develop therapy in HNSCC.

(I) Genomic analysis of tumor derived HNSCC cell lines from Indian patients

We opted for an integrated analysis approach by applying the widely used posterior filtering strategy to effectively reduce the amount of data obtained from individual platforms. Furthermore, I performed validation of copy number alterations, mutations and gene expression changes using orthologous methods. The novel gene with biologically relevant alterations were functionally characterized using loss-of-function and gain-of-function approach in HNSCC cell line and NIH/3T3 cells.

- **Validation of copy number alterations in HNSCC cell lines**

SNP array was performed to define the copy number alterations of HNSCC cell lines (AW13516, AW8507, NT8e, and OT9 cells). Several hallmark genes were found to be amplified in cell lines such as *CCND1*, *NOTCH1*, and *HES1* in all four cells; *PIK3CA* in AW13516 and AW8507 cells; hallmark genes with copy number deletion includes *CDKN2A* in AW13516; *FBXW7* in NT8e, AW13516 and OT9 cells. In contrast, we

found amplification of *NOTCH1* and its target gene *HES1* across HNSCC cell line, as opposed to previously reported frequent inactivation of Notch receptors [27, 28].

- **Validation of gene expression changes in HNSCC cell lines**

I validated the expression data using quantitative reverse transcriptase PCR (qRT-PCR). We observed expression changes in of hallmarks of HNSCC such as *CCND1, MYC, MET, CTNNB1, JAK1, HRAS, JAG1, and HES1* and down regulation of *FBXW7, SMAD4* in at least 3 cell line were observed. I performed correlation analysis and observed strong positive correlation (R^2~0.7) between transcriptome log10 (FPKM+1) and qRT-PCR C_T values across all four cell lines.

- **Validation of mutations in HNSCC cell lines**

Of 20 HNSCC hallmark variants predicted as deleterious by two of three algorithms used for functional prediction [29-31], 17 variants could be validated by Sanger sequencing, including: *TP53* (R273H and P72R), *PTEN* (H141Y), *EGFR* (R521K), *HRAS* (G12S and R78W), and *CASP8* (G328E). We have also validated mutation in novel gene in HNSCC, which includes; *NRBP1* (Q73*), *DCC* (R677H), *NETO1* (G402R) and *ZNF594* (K758N).

- **Integrated analysis to identify biologically relevant genes in HNSCC cell lines**

As a proof of principle, we performed an integrated characterization of copy number analysis, whole transcriptome and exome sequencing of 4 HNSCC cell lines established from Indian patients to identify biologically relevant alterations from fewer number of samples. In summary, we identified 38 genes having two or more type of alterations. Of the known genes in HNSCC includes *TP53*, *HRAS*, *MET*, *PTEN* and *CASP8*[12]. Among the novel genes identified in HNSCC cells, *Nuclear receptor binding protein,*

NRBP1, harboring heterozygous truncating mutation (Q73*) in NT8e cells was interesting, as identical mutation has previously been reported in lung cancer [32].

(II) Functional analysis of *NRBP1* in HNSCC cells and NIH/3T3 cells

The studies have shown poor clinical outcomes are associated with *NRBP1* overexpression in prostate cancer and activating role in drosophila [33-35]. Additionally, it has been also act as tumor suppressive role in the mouse model and human cancers [36, 37]

I demonstrate that even partial knockdown of mutant *NRBP1* expression in the NT8e cells, but not WT *NRBP1* expression in the OT9, significantly inhibited anchorage-independent growth and cell survival. I next tested the oncogenic role of *NRBP1*. This work provides the first functional analysis of mutant *NRBP1* and establish that NIH/3T3 cells expressing the mutant *NRBP1* enhance their survival and anchorage-independent growth with concomitant phosphorylation of the MAPK, while its knockdown diminishes survival and anchorage-independent growth by oral cancer cells expressing activating *NRBP1* mutations. Thus, NT8e cells harboring mutant *NRBP1* were found to be consistent with its suggestive role in prostate cancer biology and other model organisms.

- **Objective-2: Integrated genomic characterization of head and neck squamous cell carcinoma tumors**

• Development of first Indian germline SNP database "TMC-SNPdb"

To identify the bona fide somatic genomic alterations I developed the Tata Memorial Centre-SNP database "TMC-SNPdb" as the first freely available open access Indian population specific germline variant database which comprises of 114,309 unique novel

variants obtained from whole exome data of 62 ‘normal’ samples from tongue, gallbladder, and cervical cancer patients of Indian origin.

Application of TMC-SNPdb in cancer somatic mutation analysis significantly depletes 42%, 33% and 28% false positive somatic events post dbSNP depletion in Indian origin tongue, gallbladder, and cervical cancer samples, respectively deplete additional Indian population-specific SNPs over and above dbSNP and 1000 Genomes databases. Beyond cancer somatic analyses, we anticipate utility of the TMC-SNPdb in several Mendelian germline diseases.

(I) Exome sequencing analysis of early tongue squamous cell carcinomas

The most of the studies in tongue cancer include the advanced stage samples (pT3-pT4) and understanding of early tongue cancer genomic alterations are largely un-explored [18, 19]. To identify underlying genomic alterations in HPV-negative early stage TSCC tumors I carried out whole-exome sequencing of twenty-two paired and one orphan tumor (n=47).

• Somatic variants analysis in early tongue squamous cell carcinoma tumors

I observed recurrent mutations in *TP53* (42%), *NOTCH1* (23%), *CDKN2*A (12%), *HRAS*, (8%), *USP6* (8%) and *FANCA, HLA-A, PIK3CA, KMT2D* and *PDE4DIP* in single patients. Overall the frequency of mutations observed in the hallmark genes was consistent with the COSMIC and TCGA HNSCC data with altered frequency for *TP53* and *NOTCH1*, but consistent with reports from ICGC-India (Gingivobuccal) (8), tongue from India and Asia (16, 17).

• Somatic copy number alterations analysis using whole-exome sequencing data

Several known genes in HNSCC with copy number alterations were observed in early TSCC other than previously described for *NOTCH1* [38] were amplifications of

CCND1, *EGFR*, *FADD*, *PIK3CA*, *FGF19*, *ORAOV1*, *MET*, *SOX2* and *MYC* and others with deletions of *CDKN2A*, *DDX3X*, *APC* and *CECR6*, as previously described in HNSCC [11, 12, 18] and candidate genes with copy number amplification and deletions were validated using qPCR.

(II) Genomic analysis of Notch pathway genes in early tongue squamous cell carcinoma

Here, I studied clinical and functional significance of Notch pathway alterations in early stage tongue squamous cell carcinoma (TSCC) by using whole exome and transcriptome sequencing along with real-time PCR-based analysis of copy number changes, transcript expression and immunohistochemical analysis of Notch pathway components, followed by cell-based biochemical and functional assays in HNSCC cells.

• Notch pathway is activated in early stage TSCC patients and correlates with node positive and non-smoking habit

The inactivating *NOTCH1* mutation (4%) were found at a lower frequency in our sample set than that reported from the Caucasian population [27, 28, 39] but consistent with similar finding from a recent Asian study [18, 40]. We found somatic amplification at *NOTCH1* in 12 of 38 tumors; over-expression of *NOTCH1* transcripts was observed in 16 of 45 samples-- consistent with our analysis of the TCGA-TSCC data set (N=126), which is not reported earlier [41]. Consistent with amplification and overexpression of Notch pathway components, Immunohistochemical analysis for activated NOTCH1 intracellular domain (NICD) in a set of 50 patients indicated strong immune-reactivity for active Notch signaling present in 40% tumor samples.

Clinically, NOTCH1 transcript expression significantly correlate with non-smoking habit (χ^2=7.325, P=0.026) of patients, consistent with previous reports in other pathological conditions including lung adenocarcinoma [31-33]; and activated NOTCH1 intracellular domain NICD correlates with lymph node metastasis (χ^2=7.10, *P*=0.029).

- **Functional analysis of *NOTCH1* in HNSCC cells and its role in cancer stem-like cells maintenance**

A significant reduction in cancer stem-like cells features, soft-agar colony formation, survival and migration of the HNSCC cells was observed post shRNA-mediated knockdown or inhibition of Notch pathway activation by gamma secretase inhibitor (GSI-XXI). The expression of *NOTCH1* and its pathway genes has been shown to maintain cancer stem-like cells (CSCs) in various tumors, as determined by their ability to form spheroids and expression of molecular markers ALDH1, CD133 and CD44 [12, 24]. These findings are consistent with reports where Notch signaling has been shown to be required for stem cell-like features in several cancer types, including HNSCC [42, 43].

(III) Transcriptome sequencing to define transcriptional landscape of HPV-negative early tongue squamous cell carcinoma tumors

- **Discovery and validation of transcript fusions in early tongue cancer patients**

Post stringent bioinformatics filtering 242 unique cancer specific transcript fusions were identified across 10 tumor and 4 cell lines using Chimerascan [44]. The forward and the reverse primers were designed to encompass the junction point of the fusion transcript and validated twelve fusions in cDNA and PCR amplified amplicons were

sequenced using Sanger sequencing. I further screened twelve validated fusions in large cohort (N=48) and identified several known recurrent transcript fusions (*CLN6-CALML4* (9/48), *RRM2-C2orf48* (7/48), *POLA2-CDC42EP2* (4/48), and *CTSC-RAB38* (2/48)) and novel transcript fusion(*LRP5-UBE3C* (7/48), *YIF1A-RCOR2* (6/48), *SLC39A1-CRTC2* (2/48), *BACH1-GRIK1* and *EXT1-MED30* (2/48) in HPV-negative early tongue cancer patients tumors. Of the 12 transcript fusions validated, previously known fusions include *CLN6-CALML4*, *RRM2-C2orf48, POLA2-CDC42EP2,* and *CTSC-RAB38,* described in various cancer types such as breast, pancreatic, leukemia, renal cell carcinoma and gastrointestinal stromal tumors and non-tumors samples [45-50]. Interestingly, *LRP5-UBE3C* and *FTSJD2-BTBD9* transcript fusions were also confirmed at the genomic DNA level in NT8e and OT9 cell line, respectively, suggesting their origin by genomic level rearrangements.

- **Integrated analysis of genes and miRNAs in early TSCC tumors and functional validation**

Integrated analysis of genes and miRNAs in nodal based analysis revealed *MMP10* (matrix metalloproteinase 10) predicted to be target of two novel miRNAs (*miR-944*, *miR-183-5p*). I further validated the transcript and protein expression of MMP10 in extended cohort of early tongue tumors using qRT-PCR (N=34) and immunohistochemical analysis (N=50) and observed significant up-regulation in proportion of early tongue tumors. Furthermore, *miR-944*, *miR-183-5p* expression was confirmed using qRT-PCR and observed negative correlation for *MMP10* transcript and miRNAs (*miR-944*, *miR-183-5p*) expression. To validate *MMP10* as a target of microRNAs (*miR-944* and *miR-183-5p*), I performed cloning of the microRNA (*miR-944* and *miR-183-5p*) in pcDNA vector and 3'UTR of *MMP10* in pGL3 vector. I

performed the luciferase assay of protein lysate obtained post transfection in 293FT and observed that mir-944 significantly reduced luciferase activity and it was recovered post treatment with microRNA inhibitor compared to mir-183-5p, indicating *MMP10* as target of *miR-944*.

5. CONCLUSIONS AND FUTURE PROSPECTS

This study presents the integrated genomic characterization of Indian patient-derived HNSCC cell lines. This study identifies mutant *NRBP1* to play an oncogenic role in head and neck cancer. However, in depth systematic sequencing of *NRBP1* in a wide variety of tumor types may help indicate utility of *NRBP1* inhibition in human cancer.

This study also present the Tata Memorial Centre-SNP database (TMC-SNPdb), as the first open source, flexible, upgradable, and freely available Indian SNP database (accessible through dbSNP build 149 and ANNOVAR. The application of TMC-SNPdb in somatic mutation analysis significantly depletes false positive somatic events post dbSNP depletion in Indian cancer samples. Beyond cancer somatic analyses, we anticipate utility of the TMC-SNPdb in several Mendelian germline diseases.

Using genetic and chemical perturbation of Notch pathway using shRNA and gamma secretase inhibitors revealed significant decrease in cell survival, anchorage-independent colony formation, migration, and cancer-stem-like features indicating oncogenic role of NOTCH1 in early TSCC. Clinically, Notch pathway activation was significantly correlated with greater lymph node metastasis and non-smoking habit of the early TSCC patients. Taken together, this study suggests that NOTCH1 could be a potential therapeutic target in early TSCC patients.

Importantly, this study for the first time provides the comprehensive the landscape of genomic and transcriptomic alterations in HPV-negative early TSCC patients. Of the

several novel recurrent validated transcript fusions identified here in HPV-negative early TSCC could be implicated in the initiation and progression of these tumors and functional analysis of candidate fusion in cells would provide novel therapeutic target in early TSCC.

Finally, the analysis of deregulated genes and miRNAs revealed common upregulation of matrix metalloproteases in early tongue tumor and novel microRNA. This study validates a novel microRNA (miR-944) targeting 3'UTR of *MMP10* using luciferase assay, however, detailed functional analysis in HNSCC cell lines in needed to understand the implication in cellular system.

6. REFERENCES:

1. Watson, I.R., et al., *Emerging patterns of somatic mutations in cancer.* Nat Rev Genet, 2013. **14**(10): p. 703-18.
2. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.
3. Garraway, L.A. and E.S. Lander, *Lessons from the cancer genome.* Cell, 2013. **153**(1): p. 17-37.
4. Tan, D.S., T.S. Mok, and T.R. Rebbeck, *Cancer Genomics: Diversity and Disparity Across Ethnicity and Geography.* J Clin Oncol, 2016. **34**(1): p. 91-101.
5. Jones, S., et al., *Personalized genomic analyses for cancer mutation discovery and interpretation.* Sci Transl Med, 2015. **7**(283): p. 283ra53.
6. Stratton, M.R., P.J. Campbell, and P.A. Futreal, *The cancer genome.* Nature, 2009. **458**(7239): p. 719-24.
7. Rothenberg, S.M. and L.W. Ellisen, *The molecular pathogenesis of head and neck squamous cell carcinoma.* J Clin Invest, 2012. **122**(6): p. 1951-7.
8. Leemans, C.R., B.J. Braakhuis, and R.H. Brakenhoff, *The molecular biology of head and neck cancer.* Nat Rev Cancer, 2011. **11**(1): p. 9-22.
9. Ang, K.K., et al., *Human papillomavirus and survival of patients with oropharyngeal cancer.* N Engl J Med, 2010. **363**(1): p. 24-35.

10. O'Rorke, M.A., et al., *Human papillomavirus related head and neck cancer survival: a systematic review and meta-analysis.* Oral Oncol, 2012. **48**(12): p. 1191-201.

11. Cancer Genome Atlas, N., *Comprehensive genomic characterization of head and neck squamous cell carcinomas.* Nature, 2015. **517**(7536): p. 576-82.

12. India Project Team of the International Cancer Genome, C., *Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups.* Nat Commun, 2013. **4**: p. 2873.

13. Seiwert, T.Y., et al., *Integrative and comparative genomic analysis of HPV-positive and HPV-negative head and neck squamous cell carcinomas.* Clin Cancer Res, 2015. **21**(3): p. 632-41.

14. Moore, S.R., et al., *The epidemiology of tongue cancer: a review of global incidence.* Oral Dis, 2000. **6**(2): p. 75-84.

15. Chaturvedi, A.K., et al., *Worldwide trends in incidence rates for oral cavity and oropharyngeal cancers.* J Clin Oncol, 2013. **31**(36): p. 4550-9.

16. Krishnamurthy, A. and V. Ramshankar, *Early stage oral tongue cancer among non-tobacco users--an increasing trend observed in a South Indian patient population presenting at a single centre.* Asian Pac J Cancer Prev, 2013. **14**(9): p. 5061-5.

17. Sano, D. and J.N. Myers, *Metastasis of squamous cell carcinoma of the oral tongue.* Cancer Metastasis Rev, 2007. **26**(3-4): p. 645-62.

18. Vettore, A.L., et al., *Mutational landscapes of tongue carcinoma reveal recurrent mutations in genes of therapeutic and prognostic relevance.* Genome Med, 2015. **7**(1): p. 98.

19. Krishnan, N., et al., *Integrated analysis of oral tongue squamous cell carcinoma identifies key variants and pathways linked to risk habits, HPV, clinical parameters and tumor recurrence.* F1000Res, 2015. **4**: p. 1215.

20. Krishna Rao, S.V., et al., *Epidemiology of oral cancer in Asia in the past decade--an update (2000-2012).* Asian Pac J Cancer Prev, 2013. **14**(10): p. 5567-77.

21. Ledgerwood, L.G., et al., *The degree of intratumor mutational heterogeneity varies by primary tumor sub-site.* Oncotarget, 2016. **7**(19): p. 27185-98.

22. 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.

23. Laprise, C., et al., *No role for human papillomavirus infection in oral cancers in a region in southern India.* Int J Cancer, 2016. **138**(4): p. 912-7.

24. Patel, K.R., et al., *Prevalence of high-risk human papillomavirus type 16 and 18 in oral and cervical cancers in population from Gujarat, West India.* J Oral Pathol Med, 2014. **43**(4): p. 293-7.

25. Priyanka G. Bhosale, M.P., Rajiv S. Desai, Asawari Patil, Shubhada and K.P. Kane, Manoj B. Mahimkar, *Low prevalence of transcriptionally active human papilloma virus in Indian patients with HNSCC and leukoplakia.* Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology, 2016.

26. Zhang, X.C., et al., *Tumor evolution and intratumor heterogeneity of an oropharyngeal squamous cell carcinoma revealed by whole-genome sequencing.* Neoplasia, 2013. **15**(12): p. 1371-8.

27. Stransky, N., et al., *The mutational landscape of head and neck squamous cell carcinoma.* Science, 2011. **333**(6046): p. 1157-60.

28. Agrawal, N., et al., *Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1.* Science, 2011. **333**(6046): p. 1154-7.

29. Adzhubei, I., D.M. Jordan, and S.R. Sunyaev, *Predicting functional effect of human missense mutations using PolyPhen-2.* Curr Protoc Hum Genet, 2013. **Chapter 7**: p. Unit7 20.

30. Choi, Y. and A.P. Chan, *PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels.* Bioinformatics, 2015.

31. Reva, B., Y. Antipin, and C. Sander, *Predicting the functional impact of protein mutations: application to cancer genomics.* Nucleic Acids Res, 2011. **39**(17): p. e118.

32. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603-7.

33. Hooper, J.D., et al., *Cloning of the cDNA and localization of the gene encoding human NRBP, a ubiquitously expressed, multidomain putative adapter protein.* Genomics, 2000. **66**(1): p. 113-8.

34. Gluderer, S., et al., *Bunched, the Drosophila homolog of the mammalian tumor suppressor TSC-22, promotes cellular growth.* BMC Dev Biol, 2008. **8**: p. 10.

35. Ruiz, C., et al., *High NRBP1 expression in prostate cancer is linked with poor clinical outcomes and increased cancer cell growth.* Prostate, 2012. **72**(15): p. 1678-87.

36. Wei, H., et al., *NRBP1 is downregulated in breast cancer and NRBP1 overexpression inhibits cancer cell proliferation through Wnt/beta-catenin signaling pathway.* Onco Targets Ther, 2015. **8**: p. 3721-30.

37. Wilson, C.H., et al., *Nuclear receptor binding protein 1 regulates intestinal progenitor cell homeostasis and tumour formation.* EMBO J, 2012. **31**(11): p. 2486-97.

38. Upadhyay P*, N.S., Kaur E, Aich J, Dani P, Sethunath V, Gardi N, Chandrani P, Godbole M, Sonawane K, Prasad R, Kannan S, Agarwal A, Kane S, Gupta S, Dutt S, Dutt A, *Notch pathway activation is essential for maintenance of stem-like cells in early tongue cancer.* Oncotarget, 2016.

39. *Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups.* Nat Commun, 2013. **4**: p. 2873.

40. Izumchenko, E., et al., *Notch1 mutations are drivers of oral tumorigenesis.* Cancer Prev Res (Phila), 2015. **8**(4): p. 277-86.

41. Sun, W., et al., *Activation of the NOTCH pathway in head and neck cancer.* Cancer Res, 2014. **74**(4): p. 1091-104.

42. Zhao, Z.L., et al., *NOTCH1 inhibition enhances the efficacy of conventional chemotherapeutic agents by targeting head neck cancer stem cell.* Sci Rep, 2016. **6**: p. 24704.

43. Lee, S.H., et al., *Notch1 signaling contributes to stemness in head and neck squamous cell carcinoma.* Lab Invest, 2016. **96**(5): p. 508-16.

44. Iyer, M.K., A.M. Chinnaiyan, and C.A. Maher, *ChimeraScan: a tool for identifying chimeric transcription in sequencing data.* Bioinformatics, 2011. **27**(20): p. 2903-4.

45. Qin, F., et al., *Discovery of CTCF-sensitive Cis-spliced fusion RNAs between adjacent genes in human prostate cells.* PLoS Genet, 2015. **11**(2): p. e1005001.

46. Grosso, A.R., et al., *Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma.* Elife, 2015. **4**.

47. Wen, H., et al., *New fusion transcripts identified in normal karyotype acute myeloid leukemia.* PLoS One, 2012. **7**(12): p. e51203.

48. Kang, G., et al., *Integrated genomic analyses identify frequent gene fusion events and VHL inactivation in gastrointestinal stromal tumors.* Oncotarget, 2016. **7**(6): p. 6538-51.

49. Kalyana-Sundaram, S., et al., *Gene fusions associated with recurrent amplicons represent a class of passenger aberrations in breast cancer.* Neoplasia, 2012. **14**(8): p. 702-8.

50. Babiceanu, M., et al., *Recurrent chimeric fusion RNAs in non-cancer tissues and cells.* Nucleic Acids Res, 2016. **44**(6): p. 2859-72.

Publications in Refereed Journal:

a. Published:

- **Upadhyay P**, Gardi N*, Desai S*, Sahoo B, Singh A,Togar T, Iyer P, Prasad R, Chandrani P, Gupta S, Dutt A. TMC-SNPdb: an Indian germline variant database derived from whole exome sequences. ***Database, 2016***, 1–8. doi: 10.1093/database/baw104 **(thesis work)**

- **Upadhyay P***, Nair S*, Kaur E, Aich J, Dani P, Sethunath V, Gardi N, Chandrani P, Godbole M, Sonawane K, Prasad R, Kannan S, Agarwal A, Kane S, Gupta S, Dutt S,

Dutt A. Notch pathway activation is essential for maintenance of stem-like cells in early tongue cancer. ***Oncotarget, 2016***. doi: 10.18632/oncotarget.10419 **(thesis work)**

- Chandrani, P*., **Upadhyay P*.,** Iyer P., Tanna M., Shetty M., Raghuram G. V., Oak N., Singh A., Chaubal R., Ramteke M., Gupta S., Dutt A. Integrated genomics approach to identify biologically relevant alterations in fewer samples. ***BMC Genomics*** 16, 936 (2015). **(thesis work)**

- Bharat Rekhi*, **Pawan Upadhyay*,** Manoj P. Ramteke Dutt A. MYOD1 (L122R) Mutations Are Associated with Spindle cell and Sclerosing Rhabdomyosarcomas with Aggressive Clinical Outcomes. ***Modern Pathology, 2016*** 201634

- P. Chandrani, V. Kulkarni, P. Iyer, **P. Upadhyay**, R. Chaubal, P. Das, R. Mulherkar, R. Singh, A. Dutt: NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome. ***BJC*** **05/2015**; 112(12). DOI:10.1038/bjc.2015.121

- Indraneel Mitra, Naveen Khare, Gorantla Venkata Raghuram, Rohan Chaubal, Fatema Khambatti, Deepika Gupta ,Ashwini Gaikwad ,Preeti Prasannan, Akshita Singh, Aishwarya Iyer, Ankita Singh ,**Pawan Upadhyay**, Naveen Kumar Nair, Pradyumna Kumar Mishra, Amit Dutt. Circulating nucleic acids damage DNA of healthy cells by integrating in to their genomes. ***J. Biosci.*** **40(1), March 2015**.

- **Upadhyay P**, Dwivedi R, Dutt A. Applications of next-generation sequencing in cancer. Special Section: Cancer. ***Current Science***, vol. 107, no. 5, 10 September 2014.

- Chougule A, Prabhash K, Noronha V, Joshi A, Thavamani A, Chandrani P, **Upadhyay P**, Utture S, Desai S, Jambhekar N, Dutt A .Frequency of EGFR Mutations in 907 Lung Adenocarcioma Patients of Indian Ethnicity. ***PLoS One. 2013*** Oct 4;8(10):e76164. doi: 10.1371/journal.pone.0076164.

- P. Chandrani, K. Prabhash, A. Chaughule, R. Prasad, V. Sethunath ,M. Ranjan ,J. Aich , P. Iyer, H. Dhamne, D. N. Iyer, **P. Upadhyay**, B. Mohanti, P. Chandna, R. Kumar, A. Joshi, V. Noronha, V. Patil, A. Ramaswamy, A. Karpe, R. Thorat, P. Chaudhari, A. Ingle & Amit Dutt. Drug-sensitive FGFR3 mutations in lung adenocarcinoma, Annals of Oncology **(In press, Annals of Oncology).**

b. Accepted: NA

c. In Communication

- **Pawan Upadhyay,** Nilesh Gardi, Sanket Desai, Pratik Chandrani, Asim Joshi, Prachi Terwadkar, Rohan Chaubal, Sudhir Nair & Amit Dutt Integrated analysis of tobacco/nut chewing HPV-negative early tongue cancer tumors identifies recurrent transcript fusions, 2016, **(Oncotarget Journal, under review).**
- Urmila Samant, Suvarna Sharma, Raghuram GV, Deepika Gupta, Preeti Prasannan, Ashwini Gaikwad, Pritishkumar Tidke, Nilesh Gardi, Rohan Chaubal, **Pawan Upadhyay,** Tannistha Saha, Pradyumna Mishra, Naveen Khare, Naveen Nair, Amit Dutt. Cell-free chromatin from dead cancer cells integrate into genomes of bystander

healthy cells to induce DNA damage and inflammation **(Cell Death Discovery, under review)**

- Manoj Kumar*, Siddharth Singh*, Vivek Singh Tomar, **Pawan Upadhyay**, Amit Dutt and Tapas K. Kundu., Proline 152 mutation in p53 abolishes cognate DNA binding, induces gain of function tumorigenesis, **(Journal of Biological Chemistry, under Review).**
- **Pawan Upadhyay,** Nilesh Gardi, Asim Joshi,, Manoj Ramteke, Priynacka Arora, Hemant Dhamne, Munita Menon, Sudhir Nair, Amit Dutt Gene expression meta-analysis identify MMP10-miR-944 axis in early tongue cancer tumors.

d. Other publications: NA

e. Conference abstracts:

- **P. Upadhyay** , S. Nair , E. Kaur , J. Aich , P. Dani , V. Sethunath , N. Gardi , P. Chandrani , M. Godbole , K. Sonawane , S. Kannan , B. Agarwal , S. Kane , S. Dutt A. OF50-003 Pro-oncogenic role of NOTCH1 in early tongue squamous cell carcinoma, New Ideas in Cancer – Challenging Dogmas (Feb 2016), Europian Journal of Cancer,10.1016/S0959-8049(16)31953-0 (ORAL PRESENTATION).
- **Pawan Upadhyay**, Nilesh Laxman Gardi, Hemant Ramesh Dhamne, Kavitha Sonawane, Anil D'Cruz, Sudhir Nair, Amit Dutt: Abstract 5161: Discovery of a matrix metalloproteinase MMP10 as a clinically relevant biomarker to predict lymph node metastasis in tongue squamous cell carcinoma. Proceedings of the 106th Annual Meeting of the American Association for Cancer Research, Philadelphia, PA. USA; 04/2015.
- **Pawan Upadhyay,** Jyotirmoi Aich, Vidyalakshmi Sethunath, Prachi Dani, Sadhana Kannan, Pratik Chandrani, Kavitha Sonawane, Beamon Agarwal, Shubhada Kane, Sudhir Nair, Amit Dutt: Abstract 4811: Pro-oncogenic role of NOTCH1 in early tongue squamous cell carcinoma. AACR 106th Annual Meeting 2015, Philadelphia, USA; 08/2015.
- **P. Upadhyay ,** N. Gardi , P. Chandrani , S. Desai , A. Joshi, S. Nair, A. Dutt P220-K020: Integrated genomics analysis of early stage tongue squamous cell carcinoma patients, New Ideas in Cancer – Challenging Dogmas (Feb 2016), Europian Journal of Cancer,10.1016/S0959-8049(16)31953-0.
- **Pawan Upadhyay,** Pratik Chandrani, Rohan Chaubal, Ninad Oak, Madhur Shetty, Kunal Karve, Ratnam Prasad, Prajish Iyer and Amit Dutt titled "Integrated Genomic Characterization of Head & Neck Cancer Cell Lines Derived from Indian Patients" at The Global Cancer Genomics Consortium Second Annual Symposium: Genomics Medicine in Cancer Research, 2012 , ACTREC, India. **(Best poster award)**
- **Pawan Upadhyay,** Nilesh Laxman Gardi, Kavitha Sonawane, Anil D'Cruz, Sudhir Nair, Amit Dutt. "Development of gene signature profile predicting nodal metastasis in squamous cell carcinoma of tongue" 2015 NextGen Genomics, Biology, Bioinformatics and Technologies (NGBT) Conference 1st-3rd October 2015, HICC, Hyderabad, INDIA.
- **Pawan Upadhyay**, Pratik Chandrani, Rohan Chaubal, Ninad Oak, Vaibhav Kulkarni ,Madhur Shetty, Maulik Vyas, Raghuram Gorantla, and Amit Dutt titled "Integrated

Genomic Characterization of Head & Neck Cancer Cell Lines Derived From Indian Patients" in 34th Convention of Indian Association for Cancer Research Emerging Trends in Cancer research: Road to prevention & cure & International Symposium on: Infection and Cancer (IACR-2013 - February 13 -16 2013) under Shri Rambhau Kulkarni and Shri Rajnikant Baxi award category, New Delhi, India. Kumar Prabhash, Pratik Chandrani, Jyotirmoi Aich**, Pawan Upadhyay,** Anuradha Chougule, Vanita Noronha, Amit Joshi, Saral Desai, Nirmala Jambekar, Abhishek Thavamani, Tony Jose, Puneet Chandna, Amit Dutt Profiling of actionable aletartions.

in lung adenocarcinoma, JOURNAL OF THORACIC ONCOLOGY (S1261-S1261) 2013.

- Sudhir V. Nair, **Pawan Upadhyay,** Anil D'Cruz, Amit Dutt: Abstract C125: Development of gene signature profile predicting nodal metastasis in squamous cell carcinoma of tongue.. AACR-NCI-EORTC International Conference: Molecular Targets and Cancer Therapeutics, Boston, MA. Philadelphia; 10/2013
- P. Chandrani, J. Aich, **P. Upadhyay**, A. Chougule, T. Jose, P. Chandna, K. Prabhash, A. Profiling and Discovery of actionable alterations in lung adenocarcinoma" poster at symposium Worldwide Innovative Networking (WIN) in personalized cancer medicine 2013 titled "held between July 10-12; 2013, Paris, France.

Signature of Student:

Date: 30th August 2016

Doctoral Committee:

S. No.	Name	Designation	Signature	Date
1.	Dr. Sorab Dalal	Chairman		30.8.16
2.	Dr. Amit Dutt	Convener		30/8/16
3.	Dr. Shilpee Dutt	Member		30/8/16
4.	Dr. Pritha Ray	Member		30/8/16
5.	Dr. Nagaraj Balasubramanayam	Invitee		30/08/16

Forwarded through:

Dr. S.V. Chiplunkar
Director, ACTREC
Chairperson, Academic & Training Programme
ACTREC

Dr. S. V. Chiplunkar
Director
Advanced Centre for Treatment, Research & Education in Cancer (ACTREC)
Tata Memorial Centre
Kharghar, Navi Mumbai - 410 210.

To
Dean-Academic
Dr. K. Sharma,
Director, Academics,
T.M.C.

PROF. K. S. SHARMA
DIRECTOR (ACADEMICS)
TATA MEMORIAL CENTRE,
PAREL, MUMBAI

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACTREC	**Advanced Centre for Treatment Research and Education in Cancer**
CCAMP	Centre for Cellular and Molecular Platforms
CCLE	Cancer Cell Line Encyclopedia
CLI	Command Line Interface
COSMIC	Catalogue of Somatic Mutations in Cancer
CPU	Central Processing Unit
dbSNP	Single Nucleotide Polymorphism Database
DNA	Deoxyribonucleic acid
ExAC	Exome Aggregation Consortium
FDA	Food and Drug Administration
FFPE	Formalin-Fixed, Paraffin-Embedded
MMP10	Matrix metalloproteinase 10
GUI	Graphical User Interface
HBNI	Homi Bhabha National Institute
HNSCC	Head and Neck Squamous Cell Carcinoma
HPV	Human Papillomavirus
ICGC	International Cancer Genome Consortium
IGV	Integrative Genomics Viewer
IRB	Institutional Review Board
mL	Millilitre
mm	Millimetre
NCI	National Cancer Institute
ng	Nanogram
NGGF	Next-Generation Genomics Facility

NGS	Next-Generation Sequencing
NRBP1	Nuclear Receptor Binding Protein 1
RNAi	RNA interference
SNP	Single Nucleotide Polymorphism
SRA	Sequence Read Archive
TCGA	The Cancer Genome Atlas
TMC	Tata Memorial Centre
TMH	Tata Memorial Hospital
TMC-SNPdb	Tata Memorial Centre - Single Nucleotide Polymorphism Database
TSCC	Tongue Squamous Cell Carcinoma

Chapter 1

Introduction and Review of Literature

1. Chapter 1. Introduction and Review of Literature

1.1. An introduction to head and neck cancer biology

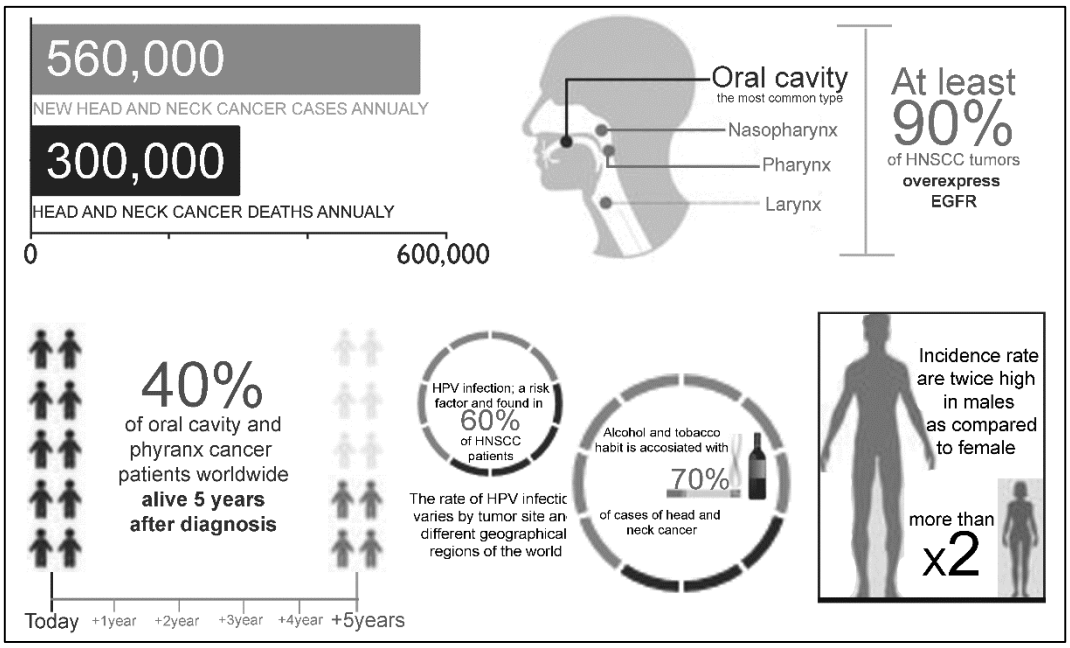
Head and Neck cancer is sixth most common malignancies worldwide, with about 600,000 new cases diagnosed annually, of which 62% arise in developing countries [1, 2]. The most common head and neck cancer type is head and neck squamous cell carcinoma (HNSCC), representing >90% of cases [1]. HNSCC is a highly heterogeneous disease because of the different subsites associated with the variable range of etiological factor, patients outcome, treatment regimens, and molecular profile [3]. The anatomic subsites include oral cavity (tongue, lip, and law), pharynx (nasal cavity, oropharynx, and hypopharynx), and larynx. HNSCC is known to arise due to the interplay of an environmental and genetic factor and is highly complex in nature [1]. The major environmental chemical risk factors include smoking, tobacco/betel quid chewing, excessive alcohol consumption, and infection by certain viruses such as by high-risk human papillomaviruses (HPV) (Figure 1.1) [4-7]. The role of genetic risk factors include genetic susceptibility to a mutagen and inherited disorders such as Fanconi Anemia, Li-Fraumeni syndrome, Bloom syndrome [8-11].

In developed countries, an overall decrease in the prevalence of smoking has been associated with a reduction in the incidence of hypopharyngeal and laryngeal squamous cell carcinomas, whereas an increased incidence of oral cavity cancers are attributed to increased prevalence of HPV infection [2, 6, 12, 13]. In the developing countries, an increasing rate of HNSCC, particularly oral cavity squamous cell carcinoma, in southern and southeastern Asian countries are often attributed to betel quid exposure [14]. Moreover, an increasing incidence of HNSCC is seen in males with >50 years age; with an increasing trend in the diagnosis of HNSCC in younger and females observed in recent years [15, 16]. Regardless of the advances in detection

and treatment modalities incidence over recent years and five-year overall survival rates for HNSCC has seen only partial improvement over the past three decades [17, 18]. The possible explanation for this includes clinical challenges in early detection, loco-regional lymph node metastasis, and distant metastasis, due to which the five-year overall survival rate remains 25-65% depending upon tumor stage and primary site (Figure 1.1) [19].

1.1.1 Aetiology of HNSCC: International and National status

A number of etiological factors could be attributed to increase number of incidences in head and neck cancer. The major risk factors of HNSCC include tobacco, alcohol consumption and viral infection. The association of tobacco chewing, smoking and alcohol consumption with HNSCC development has been well established. Among western population a strong association of cigarette smoking and alcohol usage is known for HNSCC, whereas smoking and chewing tobacco has been associated with oral cancer in South East Asia, including India. In India, it is estimated that about 65% of men and 33% of women use some or the other form of tobacco [20].

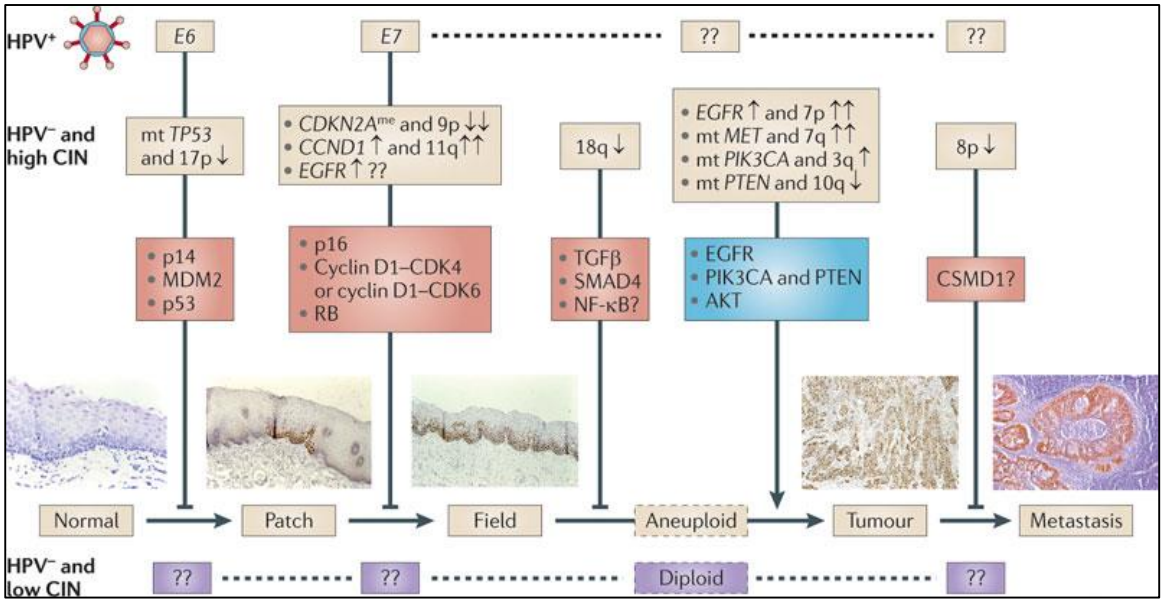


I Figure 1.1: An overview of Head and Neck Squamous Cell Carcinoma (HNSCC)

1.1.2 Genetic progression of head and neck cancer

The multi-step clonal evolution molecular carcinogenesis model of HNSCC describes the origin from a single genetically altered clone or progenitor stem cell in the normal epithelium which progresses to a patch, a pre-cancerous field, invasive carcinoma, and finally metastatic disease (Figure 1.2) [2, 21, 22]. The majority of these genetic alterations involves activation of pro-oncogenes and/or inactivation of tumor suppressor genes which confers selective growth advantage and cell proliferation, evasion of cell apoptosis, and increased metastatic capability [23].

The earliest and the most common alteration in HNSCC is the inactivation of the *TP53* (tumor suppressor protein) by multiple mechanisms such as loss of heterozygosity (LOH) at chromosome 17p, somatic mutations, which is present in the 40-80% HNSCC cases, degradation by the E6 HPV oncoprotein in HPV-positive tumors [24-27]. The dysplastic oral lesions harboring the mutation in *TP53* has been shown to be associated with a significantly higher risk of malignant progression and poor prognosis in HNSCC [28-30]. The subsequent event in HPV-negative tumors includes the upregulation of oncogene *CCND1*, by copy number amplification at locus 11q13 and down-regulation of tumor suppressor cyclin-dependent kinase inhibitor 2A (*CDKN2A*) [31, 32]. However, concomitant inactivation of p53 and upregulation of cyclin D1 leads to immortalization of oral keratinocytes suggesting interplay of additional alterations for the malignant transformation [33].



II Figure 1.2: The integrated model of multi-step progression of molecular carcinogenesis in head and neck cancer squamous cell carcinoma

(Reprinted by permission from Macmillan Publishers Ltd:[Nature Reviews Cancer] [2], copyright (2011). Most information has been deciphered from oral carcinogenesis, there are fewer data on the other subsites of HNSCC. A progenitor or adult stem cell acquires one (or more) genetic alterations, including a mutation in TP53, and forms a patch containing genetically altered daughter cells that can be detected by immunostaining for mutant p53. By escaping normal growth control and/or gaining growth advantage, this patch or clonal unit develops into an expanding field, laterally replacing the normal mucosal epithelium. Eventually, a subclone in the field evolves into an invasive cancer, and progresses to metastasis. Three critical steps can be discriminated in this model: the conversion of a single mutated stem cell in a patch into a group of stem cells without proper growth control (field); the eventual transforming event, which turns a field into an overt carcinoma showing invasive growth and metastasis; and the development of metastasis. Both aneuploidy and the accumulation of cancer-associated genetic changes in fields are linked to the risk of malignant progression. The normal epithelium is promoted through the several successive genetic alterations via several

stages that progress from a patch of genetically altered epithelial cells to a precancerous field, to a carcinoma in-situ, and ultimately invasive, angiogenic and metastatic carcinomas. Genetic and chromosome alterations are indicated in yellow boxes, oncogenic pathways are depicted in the blue box, tumour-suppressive pathways are shown in the orange boxes. ↑ indicates overexpression or gain; ↑↑ indicates high-level amplification; ↓ indicates loss; and ↓↓ indicates homozygous loss. *CCND1*, cyclin D1; *CDK*, cyclin-dependent kinase; *CDKN2A*, cyclin-dependent kinase inhibitor 2A; me, methylated; mt, mutated; NF-κB, nuclear factor-κB; *PIK3CA*, phosphoinositide-3 kinase subunit-α; TGFβ, transforming growth factor-β, high chromosome instability (high CIN), with few genetic changes (low CIN).

The late stage events include upregulation of oncogenes *EGFR*, *MET*, and *PIK3CA* and downregulation of tumor suppressors *PTEN via* copy number alterations and somatic mutations which promote enhanced tumor cell growth, proliferation, mortality, angiogenesis, and are associated with conventional treatment resistance and worse clinical outcomes [34-38]. Investigation of the molecular cross-talk of receptor tyrosine kinases and its downstream signaling network (*MET, EGFR* and *PIK3CA*) is still an area of active investigation and are a potential therapeutic target in HNSCC [1, 39]. The detailed understanding of the way these pathways interact would catalyze the development of targeted treatment approaches and discovery of reliable biomarker of treatment response in HNSCC [39].

1.2. Clinical staging

For the assessment of tumor burden, clinical staging of HNSCC plays a crucial role by providing the universal criteria and facilitating the treatment of choice for the patients [3, 40]. The most frequently applied and widely acceptable classification system is the TNM

classification systems of the Union Internationale Contre le Cancer (UICC) and the American Joint Committee on Cancer (AJCC) tumor staging system [3, 40]. Currently, HNSCCs are grouped into four clinical stage based on the TNM classification which describes the extent of the primary tumor (T), presence and extent of regional lymph node metastasis (N), and presence or absence of distant metastasis (M) [40]. When this classification is made before any treatment, it is called as clinical TNM (cTNM) and when made post histopathological examination of resected tissue, it is known as pathological TNM (pTNM). All the possible combination of T, N, and M categories could generate a large number of TNM stage categories (Table 1.1).

I Table 1.1: Clinical staging criteria for oral cavity carcinomas. (The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging) [41].

Primary tumor	
TX	Primary tumor cannot be assessed
T0	There is no evidence of primary tumor
Tis	Carcinoma is in situ
T1	Tumor is 2 cm or less in greatest dimension
T2	Tumor is more than 2 cm but not greater than 4 cm in greatest dimension
T3	Tumor is more than 4 cm in greatest dimension
T4a	Moderately advanced local disease.
	Tumor invades adjacent structures only (e.g. through cortical bone, [mandible or maxilla] into deep [extrinsic] muscle of tongue [Genioglossus, hyoglossus, palatoglossus, and styloglossus], maxillary sinus, the skin of face).
T4b	Very advanced local disease
	Tumor invades masticator space, pterygoid plates, or skull base and/or encases internal carotid artery (ICA).
Regional lymph nodes	
Nx	Regional lymph nodes cannot be assessed
N0	No regional lymph nodes metastasis
N1	Ipsilateral single lymph node 3 cm or less in greatest dimension
N2a	Ipsilateral single lymph node more than 3 cm, not more than 6 cm in greatest dimension
N2b	Ipsilateral multiple lymph nodes, none more than 6cm in greatest dimension
N2c	Bilateral/contralateral lymph nodes, none more than 6 cm in greatest dimension
N3	Lymph node more than 6 cm in greatest dimension
Metastasis	
M0	No metastasis
M1	Distant metastasis

Moreover, the different T, N, and M combination are grouped together in 1 of 4 unique stage categories (stage I, II, III and IV) (Table 1.2). Stage I and II are classified as 'early stage' while III and IV as 'late stage' disease. These staging systems would facilitate two most important aspect of disease i.e., in the prediction of the clinical course of the tumor, prognosis for the patient, and choice of the most appropriate treatment. While the clinical staging system of HNSCCs is unsatisfactory as it stands and is continuously evolving with the deeper understanding and advances in the ability in accessing the tumor [3, 40]. Due to heterogeneous clinical behavior and outcome associated with the various HNSCC subsites, the staging criteria slightly varies for each subsites, yet regional nodal metastasis unequivocally remains the most important prognostic factor in predicting outcome in HNSCC [40].

II Table 1.2: Staging grouping of HNSCC.

	N0	N1	N2	N3	Status
T1	Stage I				Early
T2	Stage II				
T3	Stage III				Late
T4a	Stage IV A (Resectable)				
T4b	Stage IV B (Nonresectable)				

1.3. Unmet need to treat head and neck cancer

Mostly the treatment decision of HNSCC is based on the TNM staging system, anatomical site of the tumor, age of the patient, and lifestyle factors (*e.g.* tobacco usages and alcohol

consumption) [42]. The most important concern while making the treatment decision is to preserve organ nearby of the tumor in order to provide a better quality of life while obtaining high cure rate and reducing the probabilities of disease relapse. The early stage disease is treated with single modality therapy either surgery or radiotherapy (RT) and has favorable prognosis [42, 43]. Surgery is favored because of its simplicity, low cost, no significant functional or cosmetic deficit. A randomized controlled trial at Tata Memorial hospital, India, confirmed that elective neck dissection was significantly better than therapeutic neck dissection in node negative early oral cancer patients [44]. Locally advanced operable HNSCC tumors are treated with combined modality therapy, surgery followed by postoperative radiotherapy or chemoradiation therapy (CRT) [45]. Moreover, the non-surgical approaches such as chemotherapy (CT) have been shown to offer improved loco-regional control and better organ preservation in randomized clinical trials worldwide including India [46-48].

The molecularly targeted agents are highly attractive therapeutic option in HNSCC due to their inherent specificity and are expected to have fewer side effects. A classic example is the targeting epidermal growth factor receptor (EGFR)-targeting antibody (Cetuximab), which is the single FDA-approved treatment option for the loco-regionally advanced HNSCC with a response rate of 15% [49]. The addition of cetuximab with radiotherapy was found to be associated with significantly higher 5-year survival rate (46%) as compared to the radiotherapy alone (36%) in a clinical trial [49]. In India, efficacy of the cetuximab in the palliative settings along with conventional chemotherapy is currently being investigated. The results from the recent Phase IIb randomized clinical trial reported the clinical utility of nimotuzumab, a monoclonal anti-EGFR antibody, used concurrently with RT and CRT indicates acceptable toxicity and confers long-term survival benefit [50]. The application of additional promising molecular targeted agents such as VEGF, mTOR, and PI3K-inhibitors are being actively

investigated in pre-clinical and clinical settings with an anticipation that some would serve as mono-therapies or in combination with RT/CRT to further provide the long-term survival advantage and better quality of life with minimal toxicity in HNSCC patients [39].

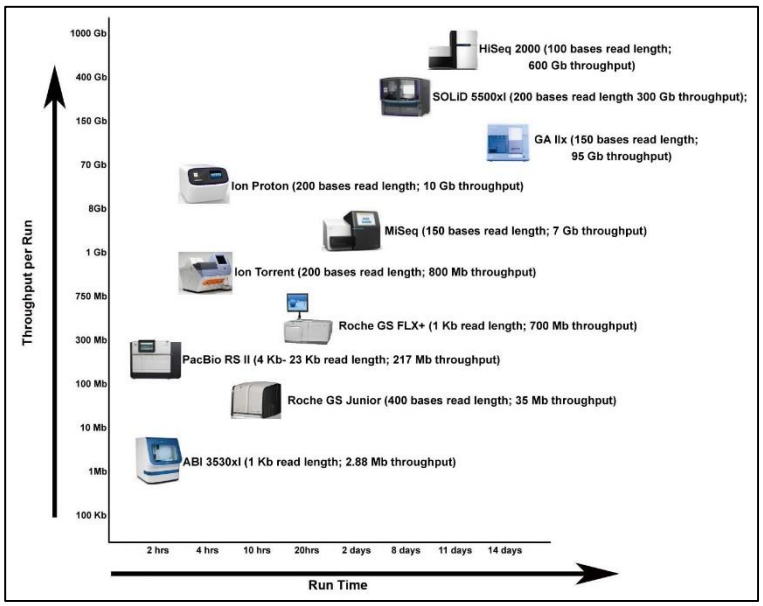
1.4. Application of next generation sequencing in cancer genome analysis

The rapid advancement in biological research in last three decades revealed cancer to be a disease, involving the dynamic changes in the cellular genome. The major foundation was set in past three decades by the discovery of mutations which produces oncogene with a gain of function and tumor suppressor with the recessive loss of function mutation [51].

The availability of the first draft of the human genome sequence in 2000 and near complete sequence by 2004 gave rise to a new field of molecular biology as 'genomics', which involves investigation of the biological phenomenon in a comprehensive, unbiased and hypothesis-free manner to reveal the surprises in biology [52]. In medicine, genomics has provided the first systematic approach to discovering genes and cellular pathways, especially in cancer research; it emerged as new revolutionary field 'cancer genomics', which involves the systematic study of the genome to find the sites of recurrent derangements in specific cancer type[53]. Initially, mutations were identified by traditional capillary-based sequencing for a specific region of the few genes and copy number alterations using DNA microarrays due to high cost and infrastructure. Unfortunately, it is was not technically feasible to interrogate the complete set of these genomic alterations in a tumor in a systematic and comprehensive manner using the classical Sanger sequencing or Microarray-based approaches due to their limiting resolution. However, with the recent possibility of using a disruptive next-generation DNA sequencing technology allows reading billions of nucleotides in a single run enabling complete genome characterization of cancer (Figure 1.3) [54]. The emergence of massively parallel sequencing

revolutionized the entire genomics community [55, 56]. This tectonic shift in the sequencing technology in past two decades led to massive increase in sequencing data throughput and a decrease in per base cost of DNA sequencing, far outpacing Moore's law of technological advance in the semiconductor industry (Figure 1.4) [52, 54, 57].

This technological innovation is referred to as "next-generation sequencing" (NGS), or massively parallel sequencing [58-60]. The most commonly available platforms currently include Illumina's GAIIx and HiSeq machines (www.illumina.com), Roche's 454 sequencer (www.454.com), Applied Biosystem's Ion Torrent or Ion Proton machines (www.appliedbiosystems.com), and SMRT sequencing system introduced by Pacific Biosciences. IN 2013, a human genome with 30X average coverage costs 5-10 thousands of dollars [61], and in late 2015 that figure had fallen below 1,500$ (https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/) . Moreover, it is expected to drop further to allow its routine application in the clinical settings [62] (Figure 1.4).

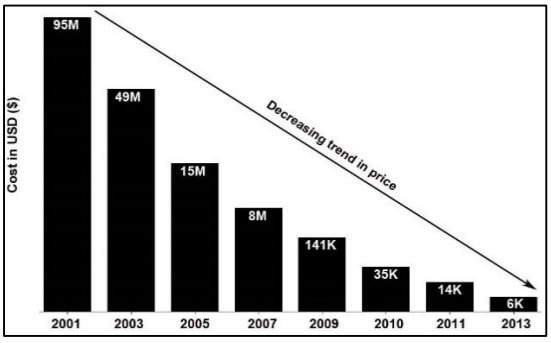


III Figure 1.3: Advancement in Next Generation DNA Sequencing Technologies.

Figure showing the advancement in next generation DNA sequencing technologies in term of per run time, throughput and read length. The axis is not as per scale.

Till now, the most of the application of next generation sequencing has been re-sequencing of whole genome or only protein coding region (whole-exome) and mRNA sequencing of human

samples to understand the inherited variations or somatic mutations and gene expression [63]. The parallel advancement in the computational tools and methodology along with next generation sequencing (NGS) empowered the identification of full range of genetic alterations in cancer, including point mutation, insertion, and deletion, copy number changes, genomic rearrangements as well as transcriptome sequencing to identify the gene expression profile related changes in an unbiased manner [54, 64]. Whole-genome (WGS) and whole-exome technologies (WES), has been extremely successful in cataloging all the different kinds of alterations in the DNA of the cancer genome. The Whole transcriptome approach (RNA-Seq) allows quantifying gene expression profiles and detecting the variant forms of alternative splicing, RNA editing and detection of novel fusion transcripts. Additionally, epigenetic modifications of the cancer genome could be studied, viz: DNA changes, histone methylation patterns using Bisulfite-Seq and ChIP-sequencing.



IV Figure 1.4: Descending trends of sequencing cost per human genome.

Figure showing the decreasing cost per human genome sequencing over the years. Cost is represented in US dollars (USD). The bar graphs are for representation purpose, not as per scale.

1.5. Cataloguing somatic alterations in cancer genome: current prospect and future challenges

The sequencing of human genome and explosion in genomic research has opened a new era in genetic research. The field of cancer genetics have applied the advances in the sequencing technologies because genomic analysis of tumors and affected individuals is not a prohibitive endeavor anymore, but instead a cost effective methodology to study the underpinning of the cancer genome. Interestingly, before the advent of large-scale sequencing efforts using next generation sequencing technologies, there has been several single group based effort which laid down the foundation of cancer-specific somatic alterations analysis using low-throughput sequencing approaches in an unbiased manner. One of the first efforts was published by Sjoblom *et al*. 2006, aimed at investigating the mutation in 13,203 genes in the CCDS Database (consensus coding sequence database) of two most aggressive cancer types; breast and colorectal cancer in 11 cell line and 11 breast cancer and two matched normal samples and further validated in 24 colorectal or breast cancer tissue samples [65]. The analysis identified 1307 confirmed somatic mutation across 1149 genes. This study provided several key messages; the difference in mutational spectrum between cancer types, gene and context specific presence of a mutation in cancer genome, the discovery of known gene which previously linked to cancer via some other genomic alterations as well as the discovery of novel candidates which would have missed based on candidate gene-based approach.

After this study, several efforts have followed with increased experience in approaches and analytical tools led to the first report of the whole genome sequenced tumor in 2008 by Ley *et al*. investigating an euploid acute myeloid leukemia (AML) and matched non-affected skin samples and identified eight heterozygous variants, all in genes not previously linked to AML: *CDH24, PCDH24, GPR123, EBI2, PTPRT, KNDC1, SLC15A1, GRINL1B*. They also identified two known somatic insertions in *FLT3* and *NPM* genes [66].

In recent years, many genome-wide efforts have underlined significant understanding of breast cancer [67-74], ovarian cancer [75], colorectal cancer [76, 77], lung cancer [68], liver cancer [78], kidney cancer [79], head and neck cancer [80], melanoma [81], acute myeloid leukemia (AML) [82, 83]. As an interesting example, six studies by different groups reported their findings on large breast cancer dataset: TCGA reported sequencing on 510 samples from 507 patients [68], As a compilation of these works, Banerji et al. carried out exome sequencing on a set of 103 samples and whole genome sequencing on 17 samples [67], Ellis et al. did exome sequencing on 31 samples and whole genome sequencing on 46 samples [69], Stephens et al. sequenced exome of 100 samples, Shah et al. performed whole genome/exome and RNA sequencing on 65 and 80 samples of triple-negative breast cancers [72], and Nik-Zainal et al. performed whole genome sequencing on 21 tumor/normal pairs [70]. Besides confirming recurrent somatic hallmark mutations in *TP53*, *GATA3* and *PIK3CA*, taken together these studies identified several other novel cancer-related mutations- mainly mutations of specific genes enriched in subtypes of breast cancers.

1.5.1 Driver versus Passenger alterations

The cancer genome analysis also involves distinguishing tumorigenic "driver" mutations from their neutral "passenger" counterparts, which occur as a result of decreased genomic stability but are not pathogenic. Several methods have already been applied to predict which missense mutations might be drivers, including CHASM [84], CanPredict [85], MutPred [86], KinaseSVM [87], SIFT [88], PolyPhen [89], MutationTaster [90] and MutationAssesor [91]. A recent study detailing exome sequencing of 72 colon tumor-normal pairs identified a magnanimous 36,303 missense somatic mutations. However, statistical analysis for significantly mutated genes using tools mentioned above led to only 23 candidates that included expected cancer genes such as *KRAS*, *TP53* and *PIK3CA* and novel genes such as *ATM*, which

regulates the cell cycle checkpoint [77]. In an exome-based study, 224 lung tumor and normal pairs sequencing identified 15 highly mutated genes in the hypermutated cancers and 17 in the non-hypermutated cancers. Among the non-hypermutated cancers, novel frequent mutations in *SOX9, ARID1A, ATM, and FAM123B* were detected besides the known *APC, TP53, and KRAS* hallmark mutations. The downstream analysis of the mutations and functional roles of *SOX9, ARID1A, ATM and FAM123B* suggested they are highly potential colorectal cancer-related genes. Non-hypermutated colon and rectum cancers were found to have similar patterns in genomic alteration.

1.5.2 Comprehensive cancer genome analysis initiatives

The potential of a comprehensive analysis of human cancer genome was first recognized by the Welcome Trust Sanger institute in Cambridge, UK, where the Cancer Genome Project was originally proposed and started (Dickson, 1999). Few years later, in 2006, the National Human Genome Research Institute (NHGRI) and the National Cancer Institute (NCI) agencies in the U.S.A launched The Cancer Genome Atlas (TCGA), an initiative comprising of the multiple research institutions and research groups to study cancer-specific somatic alterations and their relationship to cancer phenotype and clinical outcomes (http://cancergenome.nih.gov/). Two years later, the International Cancer Genome Consortium (ICGC) was established as an umbrella to majority of cancer genome sequencing efforts worldwide to provide to the research community a comprehensive set of somatic genomic abnormalities in cancer and to foster the discovery of molecular target of cancer, drug and therapeutic strategies, clinical tests and biomarker development [92, 93]. Both TCGA and ICGC also aims to make the generated data available to the public and also to establish guidelines, standards and common approaches for the myriad of studies of different cancer types. The ICGC comprises of Cancer Genome Project from the Wellcome Trust Sanger Institute, the Tumor Sequencing Project from the Broad

Institute, Baylor College of Medicine, and the TCGA effort (Table 1.3).

III Table 1.3: Current statistics of ICGC consortium.

https://dcc.icgc.org/. Last accessed: 21.11.2016

Cancer projects	70
Cancer primary sites	21
Donors with molecular data in DCC	16,236
Total Donors	19,290
Simple somatic mutations	46,429,997
Mutated Genes	57,658

As the cancer genome analysis field has evolved massively several databases, tool, and analytical workflows have been developed and matured over recent year and have been compiled and discussed in detail in several reviews [94-96]. For the identification of somatic alterations from the whole exome sequencing data of tumor and matched normal sample, a typical cancer-specific somatic mutation analysis workflow involves following five steps (Figure 1.6).

1. Quality control

Checking quality of raw sequencing data.

2. Preprocessing of raw sequencing data

Adaptor trimming and coverage analysis

3. Alignment to human reference genome.

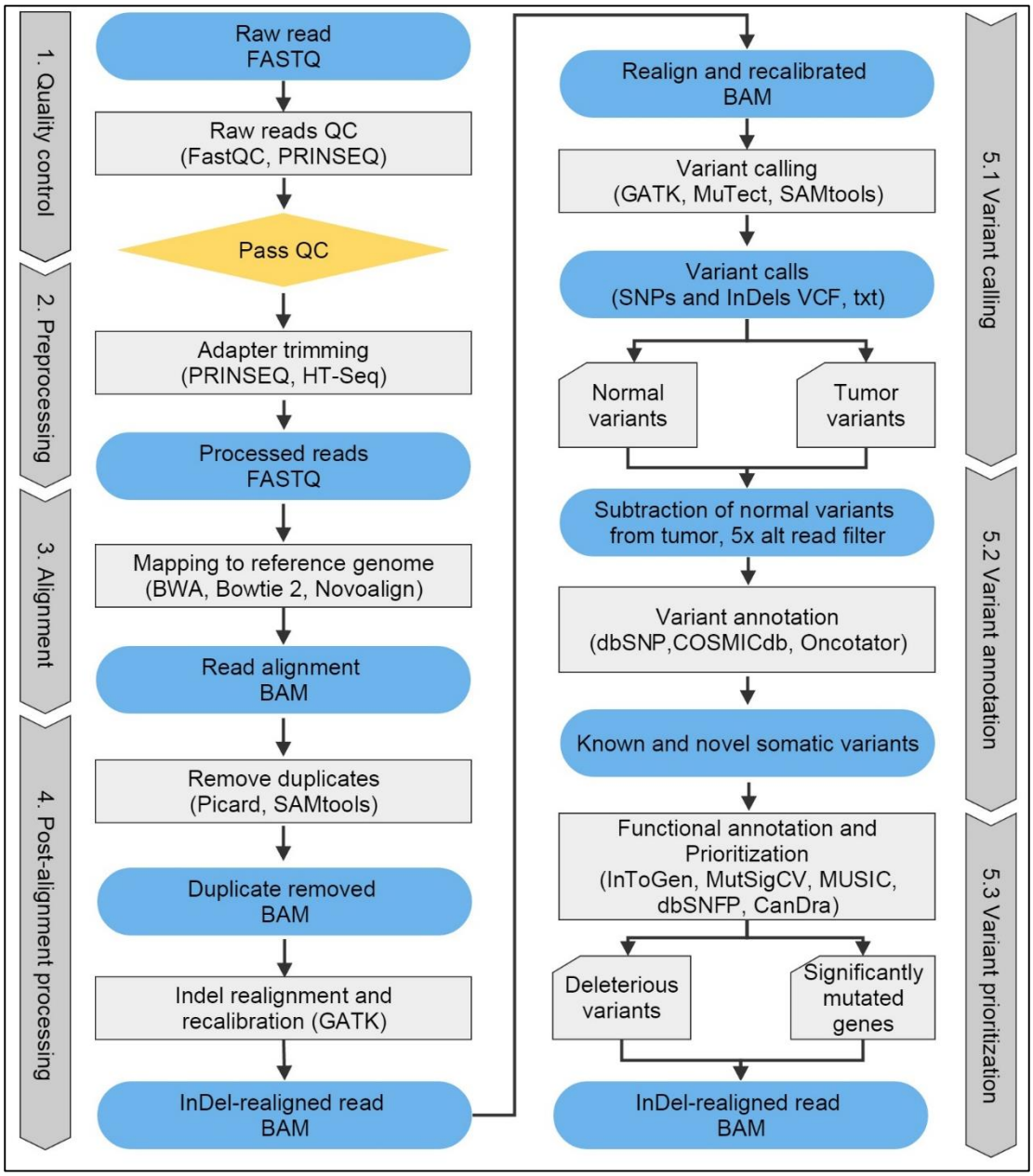
Alignment to hg19 human reference genome and generation of .sam and .bam file.

4. Variant calling

Variant calling from sequence alignment file using GATK and MuTect.

5. Variant downstream analysis

Identification of somatic variants, functional annotation, and prioritization.



V Figure 1.5: A general framework of whole exome sequencing data analysis.

1.5.3 Challenges in cancer genome analysis

Even with the advances made in sequencing technologies it's a challenging task to catalog all bonafide somatically acquired variants for following stated reasons:

First: Tumors are enormously genetically heterogeneous and complex that imposes huge

technical restrictions to discern between cancer-causing genomic aberrations ("drivers") and innocent bystander mutations ("passengers") that have no oncogenic potential in the cells [97-99]. Several factors lead to the biological complexity of primary tumors: the inter-tumor heterogeneity arises due to different subtypes with distinct morphological phenotype, expression profiles, mutation and copy number variation patterns-- for example, 73 different combination possibilities of mutated cancer genes were recently found among 100 breast cancers [100-103]; intra-tumor subclonal genetic heterogeneity are driven by multiple distinct driver genetic events [79]; in addition to the tumor heterogeneity in solid cancers, normal DNA contamination further confounds the determination of allele fraction of individual driver genetic events. To overcome these biological complexities of tumor samples, the analysis for causal genetic event needs to be adjusted for the underlying ploidy, purity and copy number alterations at chromosomal regions harboring these events.

Second: A comprehensive understanding of the genetic events during tumorigenesis can, however, be gained only by integrating the mutational analyses at the nucleotide level with analyses of copy number alterations, methylation status, and translocations. For instance, a tumor suppressor genes could be inactivated by point mutation or deleted and haploinsufficient or subject to promoter methylation- it might be deleted in 10% of patients, mutated in another 3%, promoter-hypermethylated in another 12%, and out of frame fused with some other chromosomal region in 2%. Combining this information would reveal that the gene is altered in 27% of patients, elevating its relevance threshold. Therefore, an integrated analysis provides a high-resolution and a global view of the alterations underlying the cancer genome.

Third: Between the two humans, 99.9% of the bases in the genome are similar and 0.1% of the bases that makes a person unique. Somatic mutations are known to sequentially accumulate in tumor cell genome. A typical cancer genome known to contain several polymorphic 'normal'

germline [64, 104, 105] and to identify the bona fide somatic mutation in each tumor variant list has to be subtracted from matched normal sample to deplete the germline variants [106]. Additionally, a critical aspect of somatic variant analysis involves to deplete the residual variants from the public databases of common single nucleotide polymorphism (SNPs) such as dbSNP [107] and 1000 Genomes Project [108]. The 1000 Genomes project has catalogued the variants at minor allele frequency (MAF) >1% from multiple ethnic groups, however, low frequency (0.5% <MAF ≤5%), rare (0.1% <MAF ≤0.5%) or very rare (MAF ≤0.1%) variants have not been extensively catalogued as compared to common variants (MAF>5%) because lower-frequency variants are population specific [108]. This, ethnic-specific germline variant databases are required to filter such low allele frequency false positive somatic variants in cancer genome analysis. Adopting such an analytical approach ensures filtering of paired-germline and population-specific polymorphic variants from dbSNP and 1000 Genomes Project for Caucasian population [109]. Two exhaustive initiatives addressing this issue are the publicly available exome variation datasets: NHLBI Exome Sequencing Project (ESP) (https://esp.gs.washington.edu/EVS/) and Exome Aggregation Consortium (ExAC) (http://exac.broadinstitute.org/) [110]. Information gathered from these studies is an integral part of variant annotation tools like Annovar [111]. Multiple studies such as the Indian Genome Variation Consortium [112, 113] and HUGO Pan- Asian SNP Consortium [114] have described the genomic distinctiveness of Indian population based on varying allele frequency of known SNPs, complex origin, genetic diversity [115-118], and high variation of male lineages (Y-chromosome) within the population [119, 120]. However, a concerted effort to comprehensively identify and catalog novel SNPs present exclusively in Indian population is yet to be undertaken. Lack of Indian specific SNP database has been an important impediment in cancer research, especially in efforts to discover bona fide novel somatic mutations in Indian cancer genome analysis.

1.6. Genomic analysis of head and neck using NGS: International and National status

Several genomic studies have investigated the molecular alteration landscape of HNSCCs as individual group efforts or international consortia which involve genomic, epigenetic, transcriptomic and proteomic profile with broadly three expectations. First; to attain a better understanding of the molecular alteration underlying progression and biological basis of clinical behavior, second; to discover potential novel prognostic and predictive molecular signature, and third; to identify therapeutically relevant vulnerabilities as a molecular target. Several groups have investigated the genomic landscape of HNSCCs and out of 12 studies published, 10 performed NGS-based WES on primary tumors and normal samples to identify the somatic mutations. These studies are listed in Table 1.4 (modified from Tabatabaeifar, S. *et al.*[121]).

The results from these studies revealed that HNSCC is a heterogeneous disease in nature with tumors having different mutations in several different genes. Genomic analysis not only revealed the previously known hallmark mutated genes in HNSCC, i.e. *TP53*, *CDKNA2*, *PTEN*, *PIK3CA*, and *HRAS*, but several novel mutated genes [2]. *NOTCH1* is one of an often mutated gene known to be involved in the regulating the squamous differentiation [122] and mutated about 11-19% of HNSCC tumors across studies. Moreover, the studies from Chinese population using deep sequencing showed that it was mutated in 43% (22 of 51) of OSCC tumors [123]. *FBXW7* was another novel gene found to be mutated in HNSCC in about 4% of patients by Agarwal *et al*. [80], wherein mutations in *FBXW7* were previously reported in ovarian, breast and colorectal cancer cell lines [124]. The report from ICGC-India investigating gingiva-buccal oral squamous cell carcinoma suggested heterogeneity once again and mutation profile was quite unique. There is small overlap in a number of mutation in genes *NOTCH1, TP53, HRAS* with TCGA data, however, novel mutation have been reported in *USP9X*, *MLL4*,

ARID2, UNC13C and *TRPM3* with mutation rates ranging from 10% to 16% OSCC-GB [125]. The frequent genomic alterations in several oncogenic and tumor suppressor pathways such as EGFR, FGFR, Notch and NF-kB were observed via multiple mechanisms involving copy number alterations, somatic mutations [126]. As expected, mutational signature related to tobacco exposure were also revealed, where smokers displayed high proportion of C:G > A:T transversions than non-smokers [125, 127]. Moreover, the TCGA group analysis revealed overall higher mutation rate in HNSCC [128].

IV Table 1.4: High-throughput genomic studies of HNSCC to identify somatic genomic alterations.

Author, Year	NGS platform	Cohort	Median coverage	Approach
Stransky et. al., 2011 [127]	Illumina GAII/HiSeq	74 HNSCC	150×	WES
Agrawal et al., 2011 [80]	Illumina GAIIx/HiSeq or SOLiD V3/V4	32 HNSCC	77×	WES
Fanjul-Fernández et al., 2013 [129]	Illumina GAIIx	4 LSCC	106×	WES
Nichols et al., 2012 [130]	Applied Biosystems SOLiD platform	1 HPV + and 1 HPV-	28.1×	WES
Lui et al., 2013 [131]	Illumina Platform	45 HNSCC	N/A	WES
ICGC-India, 2013 [125]	Illumina HiSeq 2000 and Roche GS-FLX	50 OSCC-GB	37×	WES
TCGA/Kandoth et al., 2013 [132]	N/a	306 HNSCC	⩾8×	WES
Pickering et al. , 2013 [133]	Illumina HiSeq 2000	38 OSCC	N/A	WES
Pickering et al. , 2014 [134]	Illumina HiSeq 2000	40 OSCC	N/A	WES
TCGA, 2016 [135]	Illumina HiSeq 2000	279 HNSCC	95x	WES
Nichols et al., 2012 [136]	Illumina HiSeq 2000	6 HNSCC cell lines	129×	Targeted(535genes)
Lechner et al., 2013 [137]	Illumina HiSeq 2000	20 HPV+ and 20 HPV− FFPE OPSCC	N/A	Targeted(182genes)
Seiwert et al., 2015 [138]	Illumina HiSeq 2000	69 HNSCC	~231x	Targeted(617genes)

LSCC: laryngeal squamous cell carcinoma. OSSC-GB: gingivobuccal oral squamous cell carcinoma. FFPE: formalin-fixed paraffin embedded. OPSCC: oropharyngeal squamous cell

carcinoma. ICGC-India: International Cancer Genome Consortium. WES: whole exome sequencing. N/A: Not available.

Most of these studies performed the genomic analysis of HNSCC tumors and cell lines using samples from the multiple subsites and advanced stage (III and IV) primary tumors leaving a void space which need to filled by future studies involving site-specific analysis in large number of samples and analysis of early stage (I & II) primary tumors with their matched normal samples. Furthermore, sequencing of healthy tissues and pre-malignant lesions to determine if premalignant could be observed and analysis of lymph node metastasises to uncover the possible genetic alterations which could promote the process of metastasis in HNSCC. Genomic analysis from different sites of primary tumors of HNSCCs would provide an important insight about the intra-tumor heterogeneity as recently reported by Mroz *et al*. [139, 140] where patients with highly heterogeneous tumors had a worse clinical outcome.

1.7. Clinical and molecular classification of HPV-positive and HPV-negative head and neck cancer

HPVs are epitheliotropic circular double-stranded DNA viruses which infect epithelial cells and the genome constitutes eight reading frame which encodes for the "early" (*E1, E2, E4-E7*) and "late" (*L1* and *L2*) genes [141]. It is well established now that two oncogenes *E6* and *E7* possess the ability to induce the malignant transformation upon infection with high-risk HPV types which leads to inactivation of tumor suppressor protein p53 and pRb, respectively, causing cell cycle deregulation and inhibition of p53-mediated apoptosis. Moreover, E7 binds pRb and targeting it towards proteasomal degradation, in turn activating E2F transcription factor to bring overexpression of *CDKN2A* resulting in cell cycle progression [141-143].

The status of HPV in HNSCC patient's samples mostly determined using molecular testing based on polymerase chain reaction (PCR) or *in situ* hybridization (ISH) to detect HPV infection at the genomic level or by immunohistochemical (IHC) staining for the tumor suppressor p16 (CDKN2A) [144]. The expression of *E7* mediates the degradation of pRb leading to overexpression of cell senescence pathway via p16 overexpression, so pRb inactivation by any other mechanism such as mutation and deletion can also result in the false assumption that *E7* is expressed and false positive detection of HPV positivity. Since the prevalence of HPV infection is high in the oropharynx, p16 expression pretest probability is also high, however, for the other anatomic sites such as oral cavity where HPV infection rate is relatively lower, the true positive detection rates falls below 50 %, rendering p16 as an ineffective diagnostic tool in other subsites of HNSCC [145, 146]. High expression of p16 also occurs in 5% of HPV-negative cases and the reason for this is still unclear [147]. Moreover, the importance of HPV infection in HNSCC is more controversial because no unanimously accepted detection method exits and reported prevalence of HPV infection in HNSCC varies from 0-100%. For these reasons, there is common census to ensure proper HPV classification of tumors by performing detection using multiple methods such as PCR, ISH, p16 staining, transcript detection and also unbiased NGS approaches has been could be adopted in future studies [148-150].

1.7.1. Clinical classification

Besides distinction by anatomic sites, HNSCC can be broadly classified into two classes: HPV-positive (HPV+) and HPV-negative (HPV-) disease and associated with the distinct etiology, clinical behavior, treatment outcomes, imaging, and pathology appearance and molecular profile (Table 1.5) [2, 151]. An oropharyngeal subsite is a primarily HPV-positive subsite whereas the only small fraction of other HNSCC subsites is associated with infection with

high-risk HPVs. The HPV-positive oropharyngeal is associated with younger age at diagnosis and more favorable outcome in spite they are typically presented with the locally advanced stage of disease [2, 4, 152]. The HPV-positive HNSCC cases also associated a significantly better prognosis for locally advanced recurrent/metastatic HNSCC and some treatment options could be applied [152, 153].

V Table 1.5: Clinical and biological characteristics of HPV-positive and HPV-negative HNSCC.

Features	HPV-positive	HPV-negative
Proportion (%)	20	80
Etiology	Oral sex and hygiene	Tobacco and excessive alcohol use
Predilection site	Oropharynx	None
Incidence	Increasing	Decreasing
Age	< 60 years	> 60 years
Prognosis	Favourable	Poor
Field cancerization	Unknown	Yes
TP53 mutation	Low	High

1.7.2. Molecular classification

Several efforts have been taken to define the underlying distinct genomic landscape of HPV-positive and HPV-negative HNSCC to attain the better understanding of the molecular and biological basis of the clinical behaviors, discover novel pathogonomic molecular signatures and identify the candidate therapeutic targets. The genomic alterations in HPV+ and HPV-HNSCC tumors identified using high-throughput genomic studies is provided in Table 1.6.

VI Table 1.6: Genomic alterations in HPV+ and HPV- tumor identified using high-throughput genomic studies in HNSCC.

TCGA (2015)[135]	**Seiwert et al. (2015) [138]**	**Stransky et al. (2011) [127]**	**Agrawal et al. (2011) [80]**
HPV-negative			
***n* = 243**	***n* = 69**	***n* = 63**	***n* = 28**
TP53 (84 %, M)	*TP53* (81 %, M)	*TP53* (73 %, M)	*TP53* (79 %, M)
CDKN2A (57 %, M/D)	*CDKN2A* (33, M/D)	*CDKN2A* (25 %, M/D)	*NOTCH1* (14 %, M)
let-7c (40 %, miRNA)	*MDM2* (16 %, A)	*SYNE1* (22 %, M)	*RELN* (14 %, M)
PIK3CA (34 %, M/A)	*MLL2* (16 %, M)	*CCND1* (22 %, Ac)	*SYNE1* (14 %, M)
FADD (32 %, A)	*NOTCH1* (16 %, M)	*MUC16* (19 %, M)	*EPHA7* (11 %, M)
FAT1 (32 %, M/D)	*CCND1* (13 %, A)	*USH2A* (18 %, M)	*FLG* (11 %, M)
CCND1 (31 %, A)	*PIK3CA* (13 %, M)	*FAT1* (14 %, M)	*HRAS* (11 %, M)
NOTCH1/2/3 (29 %, M/D)	*PIK3CB* (13 %, M/A)	*LRP1B* (14 %, M)	*PIK3AP1* (11 %, M)
TP63 (19 %, A)	*UBR5* (13 %, M/D)	*ZFHX4* (14 %, M)	*RIMBP2* (11 %, M)
EGFR (15 %, M/A)	*EGFR* (12 %, A)	*NOTCH1* (13 %, M)	*SI* (11 %, M)
HPV-positive			
***n* = 36**	***n* = 51**	***n* = 11**	***n* = 4**
E6/7 (100 %)	*E6/7* (100 %)	*E6/7* (100 %)	*E6/7* (100 %)
PIK3CA (56 %, M/A)	*PIK3CA* (22 %,M)	*PIK3CA* (27 %, M)	*EPHB3* (25 %, M)
TP63 (28 %, A)	*TP63* (16 %,M/A)	*RUFY1* (18 %, M)	*UNC5D* (25 %, M)
TRAF3 (22 %, M/D)	*PIK3CB* (13 %,M/A)	*EZH2* (18 %, M)	*NLRP12* (25 %, M)
E2F1 (19 %, A)	*FGFR3* (14 %, M)	*CDH10* (18 %, M)	*PIK3CA* (25 %, M)
let-7c (17 %, miRNA)	*NF1/2* (12 %, M)	*THSD7A* (18 %, M)	*TM7SF3* (25 %, M)
NOTCH1/3 (17 %, M)	*SOX2* (12 %, A)	*FAT4* (18 %, M)	*ENPP1* (25 %, M)
FGFR3 (11 %, F/M)	*ATM* (10 %, D)	*KMT2D* (18 %, M)	*NRXN3* (25 %, M)
HLA-A/B (11 %, M/D)	*FLG* (12 %, M)	*ZNF676* (18 %, M)	*MICAL2* (25 %, M)
EGFR (6 %, M)	*MLL3* (10 %, M)	*MUC16* (18 %, M)	

N/A HPV status not available, M; mutation, A; amplification, D; deletion, F; fusion

1.7.2.1. Landscape of known structural alterations in HNSCC

Overall, HPV+ tumors display significantly fewer chromosomal alterations overall than HPV- HNSCC tumors, possibly relatively lower number of oncogenic events are required to induce the malignant transformation in presence of high-risk HPV infection [154]. Several studies have reported the copy number landscape of HPV+ versus HPV- HNSCC tumors [32, 80, 126,

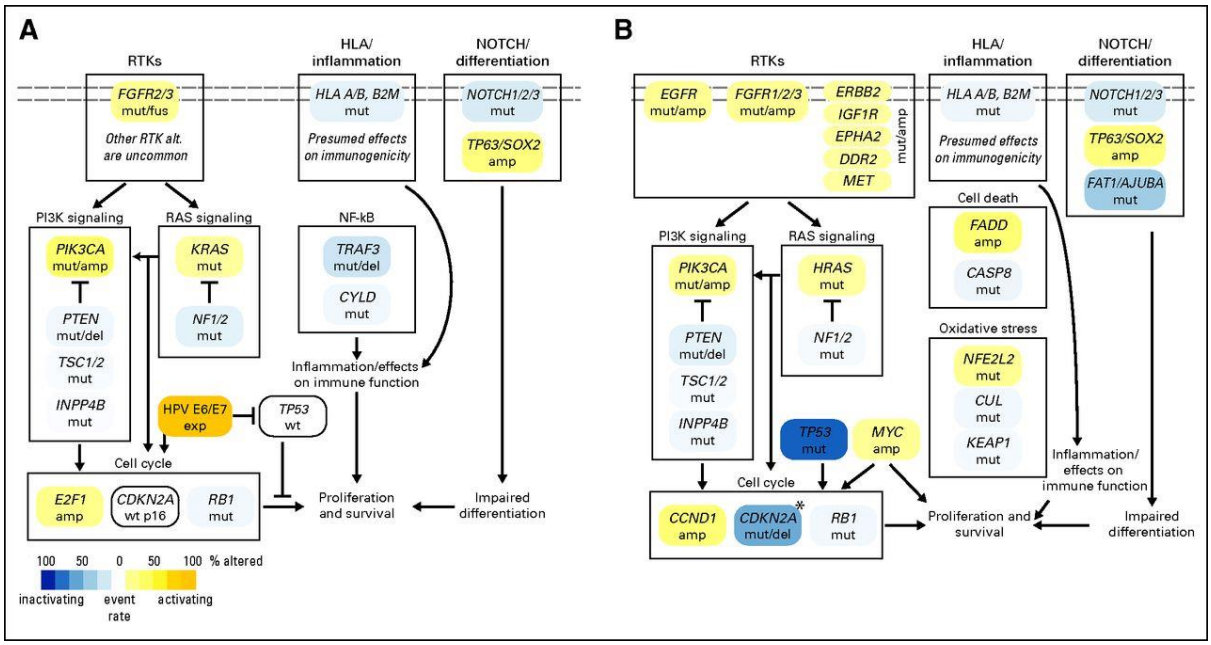
127, 133, 135, 138, 155]. There has been high concordance in the identification of copy number alterations across different studies for HPV+ and HPV- tumors, where shared amplifications are 1q, 3q, 5p, 8q, and others and deletion, includes 3p, 5q, 11q others. Recent study by TCGA noted several novel alterations in HPV-positive tumors, such as 14q32, a deep deletion [135]. Since focal deletions often include homozygous deletions which are relatively rare event in cancer and are often known to harbor key tumor suppressor genes. TCGA study identified homozygous deletion of novel gene *TRAF3* in about 20% of HPV+, which was previously reported in nasopharyngeal carcinomas [135, 156]. *TRAF3* is known to be implicated in innate interferon and acquired antiviral response for Epstein-Barr virus, HPV, and HIV and also serve as a ubiquitin ligase and negative regulator NF-kB pathway which is implicated in cell survival and other features of malignant transformation [156, 157]. Another novel deletion in HPV+ tumors occurs on chromosome 11q, where the broader deletion was observed making difficult to identify the specific gene from the several prominent tumor suppressor genes including *ATM [135]*. The focal amplification was seen at chromosome 7 (*EGFR*) in HPV+ tumors suggesting an avenue for the anti-*EGFR* targeted therapy [138]. Moreover, the HPV+ tumors also display clear focal amplification at 3q harboring squamous lineage transcription factor and an oncogene (*SOX2, TP63*, and *PIK3CA*), suggesting activation of the cell cycle, metabolism transcription factor which are known to promote the malignant phenotype [135].

Looking at cell cycle genes, a statistically significant copy number changes is observed in three genes locus, which includes; *CDKN2A* (9p), *CCND1* (11q13) and *E2F* (20q11). The *CDKN2A* (9p) is frequently lost in HPV- tumors versus HPV+ tumors. Similarly, HPV- tumors showed focal amplification at *CCND1* (11q13), whereas HPV+ tumors displayed strikingly focal amplification at *E2F* (20q11). The E2F encodes for the key transcription factor required for cell cycle initiation from G0 to G1 and cell proliferation, is amplified in 20% of HPV+ tumors

but only in 2% of HPV- tumors [135, 138]. Other alterations such as deletion 10q harboring *PTEN* found to be more frequent in HPV- tumors [135]. Additionally, *FGFR1* amplification was found to be restricted to HPV- tumors [135]. The detection of activated oncogene fusions has been a rare instance in HNSCC besides minority of HPV+ samples harbored *FGFR3-TACC3* fusions that seem promising therapeutic target [135]. In summary, global chromosomal alteration analysis supports the presence of at least two distinct subtypes of HNSCC and appears to follow the unique route of genetic progression which await functional characterization.

1.7.2.2. Landscape of known mutational alterations in HNSCC

The mutational landscape of HNSCC has been greatly contributed by exome sequencing studies in recent years revealing an untapped opportunity for the personalized therapy (reviewed in details in [126, 154]). Initial studies suggested that HNSCC has a relatively high mutational load, ranking 9th highest among 27 tumor types [158], consistent with the copy number alteration patterns. The HPV- or tobacco-associated HNSCCs were contained a significantly higher number of mutation versus HPV+ HNSCC and the five-fold difference in mutational burden between two groups [80, 127, 138]. Moreover, the mutational profile of HPV- HNSCC tumors similar to those of the smoking associated tumors types such as lung and esophageal SCCs, where frequent transitions and transversions at CpG sites was seen [127, 135, 158].



VI Figure 1.6: Frequent genetic aberrations in selected pathways altered in HPV-positive (A) and HPV-negative (B) head and neck squamous cell carcinoma.
(Reprinted with permission. © (2015) American Society of Clinical Oncology). Overview of key genetic aberrations for HPV-positive and HPV-negative head and neck cancers. Shades of gold indicate the frequency of activating changes in presumed oncogenes, and shades of blue indicate the frequency of inactivating changes in presumed tumor suppressor genes. amp, amplification; del, deletion; HLA, human leukocyte antigen; but, mutation; fun, fusion/translocation; Wt, wild type

Furthermore, mutation profile of HPV+ HNSCC was similar to cervical cancer where the number of Tp*Cp(A/C/T) caused by the HPV-induced APOBEC3B cytosine deaminase activity [80, 127, 135, 159]. The exome sequencing studies also confirmed the underlying mutational complexity and validated the role of p53 and Rb tumor suppressor pathway as driver event in HNSCC. In the TCGA study, the *TP53* was mutated in 70% of HPV- HNSCCs and rare in HPV+ tumors [80, 127, 133-138, 155]. In addition, *CDKN2A* is mutated in 20% of HNSCC and overall inactivation rate was about 80% via various other mechanism including chromosomal deletions and promoter hypermethylation leading to disruption of the Rb

pathway and cell cycle progression [135]. Interestingly, HPV+ tumors have high mutation frequencies than HPV- HNSCC tumors in a number of clinically relevant pathways PI3K, receptor tyrosine kinases, MAPK pathway and DNA repair genes (Figure 1.7) [126]. The mutation in *PIK3CA* was found in 37% of HPV+ versus 18% of HPV- tumors [131]. This is the most commonly mutated oncogene in HNSCC and mostly altered in some HPV+ HNSCCs suggesting a highly actionable target in HNSCC. The mutation in *EGFR* is not similar to those observed in kinase domain in lung cancer rather often in the juxtamembrane region of the gene which has been documented to be functional in other cancer types, including glioblastoma. Several driver mutations in the immune and cell death pathway are revealed in HNSCC which includes mutations in *HLA-A/HLA-B* and *B2M*, previously known in lung squamous cell carcinoma and recently approval of nivolumab in lung squamous cell carcinoma [135, 160]. These immune alterations are not clearly been functionally validated in HNSCC. The loss of function mutation in *CASP8* highlights the dependence on the NF-kB pathway in HNSCC. Three additional pathway which is frequently altered, but difficult to target with current therapies includes Notch, Wnt, and Oxidative stress pathway. Several studies demonstrated loss-of-function mutation in *NOTCH1* and amplification of 3q (*TP63*) involved in squamous cell differentiation. The less appreciated pathway and increasing recognized in another squamous cell differentiation pathway includes two genes; *FAT1* and *AJUBA* [135, 161]. The gene involved in the oxidative stress and frequently mutated in HNSCC includes *NEF2L2* if validated with *KEAP1* and *CUL3*, possibly represent the potential therapeutic biomarker for the clinical use. In summary, both common and unique actionable targets are present in HPV+ and HPV- HNSCC tumors and preclinical studies validating their therapeutic relevance are highly recommended.

1.8. Tongue squamous cell carcinoma: A lethal and poorly genomically characterized subsite of HNSCC

A subsite in HNSCC, Tongue squamous cell carcinoma (TSCC) comprises of two-thirds of all cancer cases in HNSCC [162, 163]. Several reports suggest an increase in incidence over recent years whereas five-year overall survival rates for TSCC remains low regardless of advances in detection and treatment modalities worldwide including India [12, 164-166]. Of all cancers arising in HNSCC, tongue cancer remains the poorest in terms of prognosis [164]. In India, tongue cancer is second most common subsite after buccal mucosa in HNSCC and affects individuals with relatively younger age group [167]. While there is gradual decrease in the incidence of other oral cavity cancers, the age-adjusted incidence rate of tongue cancer is increasing especially in urban population [168]. Previously, tongue cancer thought to be cancer of males arising due to long term exposure tobacco and alcohol but recently an epidemiological shift is seen towards never-smokers, women, never-tobacco users, and young patients [15, 166, 169-171]. The role of HPV in TSCC development is still unclear and earlier reports suggested overall prevalance rate in about 20-50% cases of TSCC, lower than buccal mucosa subsite [172, 173]. However, recent studies from western countries and India investigated HPV infection using more sophisticated manner on large number of samples and suggested overall low prevalence of HPV in TSCC tumors [197, 240-242, 244, 247]., about 27-40% of patients even at early stage (pT1 or pT2) have nodal metastasis Moreover, majority of patient at early stage TSCC undergo surgery in the form of wide local excision of the primary tumor and dissection of ipsilateral neck nodes followed by radiation and chemotherapy [172, 174]. Most important unique feature of TSCC from other subsites in HNSCC that, about 27-40% of patients even at early stage (pT1 or pT2) have nodal metastasis and may be undergoing a neck

dissection which further adds to morbidity and worse survival due to disease recurrence [175-177].

Since a number of evidence suggests that the different subsites of HNSCC displays significant dissimilarity in clinical behavior which is not attributable to its pathogenesis [178, 179]. Notably, previous studies have focused on HNSCC as a single entity and given the existence of biological differences and clinical distinctiveness in different subsites, molecular changes associated with each site could vary[180]. For example, *PIK3CA* mutations that are more predominant in human papillomavirus (HPV)-induced oropharyngeal cancer, alterations in *USP9X, MLL4, ARID2, UNC13C* and *TRPM3* in gingiva-buccal cancers [125, 135]. Due to this, several investigators have been focusing on subsite-specific studies, including large-scale efforts initiated by the International Cancer Genome Consortium (ICGC-India) on gingiva-buccal and other individual group on tongue cancer [125, 181-183]. Moreover, precancerous lesion of OSCC is also shown to have a distinct molecular profile as compared to OSCC [184]. Moreover, most of the studies carried out in TSCC tumors were restricted to candidate genes or gene panels along with few whole exome studies suggested difference in genomic profile indicating unique molecular features associated with TSCC tumors [134, 181-183, 185, 186]. Interestingly, tongue tumors has been described to have relatively lower mutation burden compared to other sub-sites in HNSCC [187].

Li *et al*, 2014 investigated 6 non-smokers TSCC exome sequencing and 20 TSCC tumors transcriptome sequencing to identify if there is any significant difference in somatic mutation and involvement of pathogenic viruses [183]. Smokers TSCC patients were having significantly higher number of *TP53* mutation as compared to non-smokers and no tumor-associated viruses were identified [183]. Pickering *et al*. 2014 performed whole exome sequencing of 38 advanced stage TSCC samples to investigate the mutational difference

between younger versus older TSCC patients and identified similar mutation [134].Another study led by Vettore *et al*. 2015 to investigate the mutational landscape of TSCC using whole exome sequencing (n=19) and targeted sequencing (n=60) of advanced stage (pT3-pT4) TSCC samples and identified somatic mutation in hallmark genes *TP53*, while mutations in *CDKN2A* and *NOTCH1* were less frequent [181]. The frequent mutation in two novel genes *DST* and *RNF213* were also identified [181]. Overall the mutation in Notch pathway and chromatin remodeling genes was prevalent [181]. Another study recently published study from India by Krishnan et al. 2016 performed integrated analysis using whole exome sequencing, SNP array, and cDNA microarray of 50 paired normal advanced stage (pT3-pT4) TSCC samples to identify the somatic mutations, copy number alterations and gene expression changes and reconfirmed the findings from previous studies [182].

In summary, these reports raise the crucial need to study HNSCC as a variety of different disease, thus necessitates genomic analysis of specific cohorts that are well-defined with regards to disease subsite, HPV-status, and tumor stage [134, 188]. Importantly, most of the genomic analysis studies in HNSCC and TSCC include the advanced stage samples (pT3-pT4) and understanding of early stage (pT1-pT2) tongue tumor genomic alterations are largely unexplored.

1.9. Research objective

1.13.1 Rationale

A major treatment modality for early stage TSCC patients is surgery; about 60% of patients at early stage (pT1 or pT2) TSCC who do not have nodal metastasis still need to undergo a neck dissection that further adds to morbidity and worse survival. Thus, there has been a huge unmet need to identify a molecular biomarker in early TSCC tumors to accurately distinguish from node negative to node positive to avoid unnecessary morbid neck dissection in clinical settings. Recent genomic characterization studies in HNSCC have been restricted to buccal mucosa (OSCC-GB) subsite, an effort led by the ICGC-IPT. On the other hand, studies carried out in TSCC of Indian origin have largely described advanced stage tumors (pT3 & pT4). However, systematic efforts have been lacking to investigate the underlying genomic alteration of tongue squamous cell carcinoma patients of early stage (pT1 & pT2). The key question and specific aims of the thesis is listed below.

1.13.2 Key Questions

What are the underlying somatic genomic alterations of human tongue cancers of Indian origin?

1.13.3 Specific Aims

1. Integrated genomic characterization of cancer cell lines model system established from the primary tumors derived from HNSCC patients of Indian origin.

2. Development of ethnic-specific SNP database to deplete the false positive somatic variants in cancer genome analysis.

3. Integrated genomic characterization of early-stage tongue squamous cell carcinoma primary tumors followed by functional validation and clinical correlation.

1.10. Summary

In this thesis, I characterize the genomic and transcriptomic alterations underlying early stage tongue primary tumors. To begin with, I describe a major impediment in the field to identify somatic mutations in TSCC patients of Indian origin due lack of ethnic specific Indian SNP database. To resolve this deficiency, in collaboration with computational biologists, I developed the TMC-SNPdb: the first Indian SNP db based on whole exome sequencing of 62 normal samples. Next, I undertook a comprehensive approach to integrate and characterize genomic and transcriptomic alterations of 4 cancer cell lines derived from primary tumors derived and 68 early stage (pT1 & pT2) tongue primary tumors. Two novel genomic alterations found in *NOTCH1* and *MMP10* were systematically functionally followed to understand their role in the head and neck cancer biology.

Chapter 2

Integrated genomics approach to identify biologically relevant alterations in fewer samples

(an excerpt; as published in BMC Genomics (2015) 16:936-949)

2. Chapter 2. Integrated genomics approach to identify biologically relevant alterations in fewer samples

(An excerpt; as published in BMC Genomics (2015) 16:936-949)

2.1. Abstract

Background: Several statistical tools have been developed to identify genes mutated at rates significantly higher than background, indicative of positive selection, involving large sample cohort studies. However, studies involving smaller sample sizes are inherently restrictive due to their limited statistical power to identify low frequency genetic variations.

Results: We performed an integrated characterization of copy number, mutation and expression analyses of four head and neck cancer cell lines - NT8e, OT9, AW13516 and AW8507-- by applying a filtering strategy to prioritize for genes affected by two or more alterations within or across the cell lines. Besides identifying *TP53*, *PTEN*, *HRAS* and *MET* as major altered HNSCC hallmark genes, this analysis uncovered 34 novel candidate genes altered. Of these, we find a heterozygous truncating mutation in Nuclear receptor binding protein, *NRBP1* pseudokinase gene, identical to as reported in other cancers, is oncogenic when ectopically expressed in NIH-3T3 cells. Knockdown of *NRBP1* in an oral carcinoma cell line bearing *NRBP1* mutation inhibit transformation and survival of the cells.

Conclusions: In overall, we present the first comprehensive genomic characterization of four head and neck cancer cell lines established from Indian patients. We also demonstrate the ability of integrated analysis to uncover biologically important genetic variation in studies involving fewer or rare clinical specimens.

2.2. Introduction

Head and neck squamous cell carcinoma (HNSCC) is the sixth-most-common cancer worldwide, with about 600,000 new cases every year, and includes cancer of the nose cavity, sinuses, lips, tongue, mouth, salivary glands, upper aerodigestive tract and voice box [1]. Recent large scale cancer genome sequencing projects have identified spectrum of driver genomic alterations in HNSCC including *CDKN2A*, *TP53*, *PIK3CA*, *NOTCH1*, HRAS, *FBXW7*, *PTEN*, *NFE2L2*, *FAT1*, and *CASP8* [80, 127, 135]. These landmark studies apply elegant statistical methodologies like MutSig [189], Genome MuSiC [190], Intogen [191], InVEx [192], ActiveDrive [193] and GISTIC [194] in identifying significantly altered genes across large sample cohorts by comparing rate of mutations of each gene with background mutation rate to determine an unbiased enrichment-- a minimum ~150 patients or higher is required for identification of somatic mutations of 10% population frequency in HNSCC [195]. This genome-wide analysis may not be directly applicable for studies involving fewer or rare clinical specimen that are inherently restrictive due to the limited statistical power to detect alterations existing at lower frequency.

On the other hand, given that a cancer gene could be selectively inactivated or activated by multiple alterations, an integrative study design performed by combining multiple data types can potentially be helpful to achieve the threshold for statistical significance for studies involving fewer or rare clinical specimen. For example, a tumor suppressor gene-- deleted in 1% of patients, mutated in another 3%, promoter-hypermethylated in another 2% and out of frame fused with some other chromosomal region in 2%-- may be considered to be altered with a cumulative effect of 8% based on integrative analysis [63, 196]. Combinatorial sources of genetic evidence converging at same gene or signalling pathway can also limit false positives by filtering strategy and potentially reducing the multiple hypothesis testing burden for

identification of causal genotype-phenotype associations [197]. Using similar approaches for posterior refinement to indicate positive selection, Pickering *et al.* identified four key pathways in oral cancer by integrating methylation to copy number variation and expression [139]; and, more recently, Wilkerson *et al.* proposed superior prioritisation of mutations based on integrated analysis of the genome and transcriptome sequencing than filtering based on conventional quality filters [198]. These and several other reports all together emphasize integration of multi-platform genomic data for identification of cancer related genes [199].

Here, we perform characterization of four head and neck cancer cell lines, established from Indian head and neck cancer patients, using classical cytogenetic approach, SNP arrays, whole exome and whole transcriptome sequencing. Adopting an integrative approach using posterior filtering will allow us to identify biological relevant alterations affected by two or more events even from fewer samples.

2.3. Materials & Methods

2.3.1 Cell culturing and single cell dilution for establishing clonal cells

Four HNSCC tumor cell lines established within Tata Memorial Center from Indian patients and described before were acquired: NT8e, OT9, AW13516, AW8507 [200],[201]. Cells were maintained in DMEM media (Gibco, USA) supplemented with 10% FBS and 2% penicillin & streptomycin antibiotics and maintained in a humidified incubator supplied with 5% CO_2 at 37oC.

2.3.2 Integrated analysis

Genes identified to be altered by SNP array, transcriptome sequencing and exome sequencing were then used for integrative analysis to prioritize the genes which are harbouring multiple

types of alteration in same or different cell line. Gene level converging of genomic data were emphasized in identification of biologically relevant alterations across platform and samples. Taking this into consideration, we designed gene prioritization based on three steps: 1) selection of genes harbouring positive correlation of focal copy number and gene expression; 2) selection of genes harbouring point mutations with detectable transcript and or altered copy number, and 3) selection of genes harbouring multiple type of alterations identified from above two gene lists. Circos plot representation of integrated genomics data was generated using Circos tool (v. 0.66) [202].

2.3.3 Sanger Sequencing validation

PCR products were purified using NucleoSpin Gel and PCR Clean-up kit (MACHEREY-NAGEL) as per manufacture's protocol and quantified using Nano-Drop 2000c Spectrophotometer (Thermo Fisher Scientific Inc.) and submitted for sequencing in capillary electrophoresis 3500 Genetic Analyzer (Life Technologies). Sanger sequencing traces were analysed for mutation using Mutation Surveyor [203]. The details of all the primers used for mutation analysis have been provided in Table 2.1.

VII Table 2.1: Primer sequences used for Sanger sequencing based validation of mutations.

S.No.	Primer ID	Sequence(5'---------->3')
1	OAD1119_EGFR_F	CCTCCACTGTCAGGCACATTTC
2	OAD1120_EGFR_R	GGATTATATTAGGCAATAATAC
3	OAD1121_MET_F	ACACAAGAATAATCAGGTTCTG
4	OAD1122_MET_R	CTTTTAGTTACCATTTACATTTTA
5	OAD1123_PTEN_F	CTCTGGAATCCAGTGTTTCTTT
6	OAD1124_PTEN_R	ATAATTATGTGAGGTGATGAAT
7	OAD1125_TP53_R273H_F	TTCCACTTGATAAGAGGTCCC
8	OAD1126_TP53_R273E_R	CTTACCGATTTCTTCCATACT
9	OAD1127_TP53_P72R_F	CACTGACAGGAAGCCAAAGGG
10	OAD1128_TP53_P72R_R	CTGTGGGAAGCGAAAATTCCA
11	OAD1129_CASP8_F	GGTCAAATTCTTATCTATCAAT
12	OAD1130_CASP8_R	CTCTGGCAAAGTGACTGGATG
13	OAD24_DCC_F	GCAAAGGAGAGGTGATAATGCATTC
14	OAD25_DCC_R	CCAATAGATGTCACCTGCCTGG
15	OAD26_PAK6_F	CCTGCTCCATAGGGGACTTGGC
16	OAD27_PAK6_R	CCAGGATGTTCTGCCATTGTGG
17	OAD30_UBE2O_F	CAGCTGGGAGACGGACAATG
18	OAD31_UBE2O_R	GTTGGCTTCTCAGGCTCCAC
19	OAD36_FLG_F	GCTCAGGAGCAGTCAAGAGATG
20	OAD37_FLG_R	GTCCAGACCTTCCTGCTGAC
21	OAD55_UNC5C_F	CTGGGTTGTTGATCTATGGCATC
22	OAD56_UNC5C_R	GGATTACAGGCGTGAGTCACCG
23	OAD676_HRAS_G12S_F	ATGACGGAATATAAGCTGGTG
24	OAD678_HRAS_G12S_R	CTGTACTGGTGGATGTCCTC
25	OAD689_HRAS_R68W_F	CATCCAGGACATGCGCAGA
26	OAD690_HRAS_R68W_R	CCCTGTCTCCTGCTTCCTCT
27	OAD79_ZNF594_F	GAATGTGGGAATGCCTTCAGGCG
28	OAD80_ZNF594_R	GTGTGTGACAAGGTGGGACCTC
29	OAD85_NETO1_F	CCTGCTGGACCAGCTGACCAAC
30	OAD86_NETO1_R	GGCTGTGTGGGCATCTCTGTC
31	OAD96_CAPN2_F	GGGCTTCCTGTGTTGCCCAG
32	OAD97_CAPN2_R	GGCCGAGGAACAGACCAGTG
33	OAD98_NRBP1_Q73*_F	GCTGGCAGAAGAGGCGAGAAG
34	OAD99_NRBP1_Q73*_R	GAAGCCATTCCCTATCCCTCC

VIII Table 2.2: Primers used for copy number and gene expression study using qPCR.

Application	S.No.	Primer ID	Sequence(5'---------->3')
COPY NUMBER AND EXPRESSION ANALYSIS	1	OAD1061_MET_F	TTCTGACCGAGGGAATCATCA
	2	OAD1062_MET_R	CCTTCACTTCGCAGGCAGAT
	3	OAD1051_JAK1_F	CTTTGCCCTGTATGACGAGAAC
	4	OAD1052_JAK1_R	ACCTCATCCGGTAGTGGAGC
	5	OAD1059_CDKN2A_F	CAAGATCACGCAAAAACCTCTG
	6	OAD1060_CDKN2A_R	CGACCCTATACACGTTGAACTG
	7	OAD1067_FBXW7_F	CCACTGGGCTTGTACCATGTT
	8	OAD1068_FBXW7_R	CAGATGTAATTCGGCGTCGTT
	9	OAD1049_NSD1_F	TCCTGAGTCAGAACATGACCTG
	10	OAD1050_NSD1_R	CGAGATTTAGCGCAAGGCTTTT
EXPRESSION ANALYSIS	11	OAD67_GAPDH_F	AATCCCATCACCATCTTCCA
	12	OAD68_GAPDH_R	TGGACTCCACGACGTACTCA
	13	OAD1065_SMAD4_F	GCTGCTGGAATTGGTGTTGATG
	14	OAD1066_SMAD4_R	AGGTGTTTCTTTGATGCTCTGTCT
	15	OAD1131_HRAS_F	TTTGAGGACATCCACCAGTACA
	16	OAD1132_HRAS_R	GCCGAGATTCCACAGTGC
COPY NUMBER ANALYSIS	17	OAD1057_NOTCH1_F	GTGACTGCTCCCTCAACTTCAAT
	18	OAD1058_NOTCH1_R	CTGTCACAGTGGCCGTCACT
	19	OAD1063_HRAS_F	CGGCAGGGAGTGGAGGAT
	20	OAD1064_HRAS_R	TTCAGCTTCCGCAGCTTGT
	21	OAD1069_CCND1_F	GAACTACCTGGACCGCTTCC
	22	OAD1070_CCND1_R	TAGAGGCCACGAACATGCAA
	23	OAD1071_MYC_F	AGAGTTTCATCTGCGACCCG
	24	OAD1072_MYC_R	AAGCCGCTCCACATACAGTC
	25	OAD 939_PIK3CA_F	TATTTGCTTTTTCTGTAAATCATC
	26	OAD 940_PIK3CA_R	TATCTAGACCAACTAAATCAC
	27	OAD502_HES1_F	AGGGCGTTAATACCGAGGTG
	28	OAD503_HES1_R	AGGTCATGGCATTGATCTGGG
	29	OAD282_DLL3_F	CCCTACCCTTCCTCGATTCTG
	30	OAD283_DLL3_R	GAACTGAAAATGGGCTTAAAACCTT

2.3.4 DNA copy number validation

Quantitative-real time PCR and data analysis was performed using Type-it® CNV SYBR® Green PCR (cat. No. 206674) as per manufacturer's instructions on 7900HT Fast Real-Time PCR System. The details of all the primers used for DNA copy number analysis have been provided in Table 2.2.

2.3.5 RNA extraction, cDNA synthesis, quantitative real time PCR

Total RNA was extracted from cell lines using RNeasy RNA isolation kit (Qiagen) and Trizol reagent (Invitrogen) based methods and later resolved on 1.2% Agarose gel to confirm the RNA integrity. RNA samples were DNase treated followed (Ambion) by first strand cDNA synthesis using Superscript III kit (Invitrogen) and semi-quantitative evaluative PCR for GAPDH was performed to check the cDNA integrity. cDNA was diluted 1:10 and reaction was performed in 10µl volume in triplicate. The melt curve analysis was performed to check the primer dimer or non-specific amplifications. Real-time PCR was carried out using KAPA master mix (KAPA SYBR® FAST Universal q PCR kit) as per manufacturer's instructions in triplicate on 7900HT Fast Real-Time PCR System. All the experiments were repeated at least twice independently. The data was normalized with internal reference *GAPDH*, and analysed by using delta-delta Ct method described previously [204]. The details of all the primers used for expression analysis have been provided in Table 2.2.

2.3.6 Generation of pBABE-*NRBP1*-PURO constructs

The cDNA of Human *NRBP1* was amplified from AW13516 cell line using Superscript III (Invitrogen, cat no 18080-093) in a TA cloning vector pTZ57R/T(InsTAclone PCR cloning kit, K1214, ThermoScientific), later site-directed mutagenesis was done using QuikChange II

Site-Directed Mutagenesis Kit (cat.no. 200523) as per manufacturer's instructions. Later both wild type and mutant *NRBP1* cDNA sequenced confirmed using Sanger sequencing and were sub-cloned in to retroviral vector p BABE-puro using restriction digestion based cloning (SalI and BamHI).

2.3.7 Generation of stable clone of NIH-3T3 overexpressing *NRBP1* cDNA

293T cells were seeded in 6 well plates one day before transfection and each constructs (pBABE-puro) along with pCL-ECO helper vector were transfected using Lipofectamine LTX reagent (Invitrogen) as per manufacturer's protocol. Viral soup was collected 48 and 72 hours post transfection, passed through 0.45µM filter and stored at 4^{O}C. Respective cells for transduction were seeded one day before infection in six well plate and allowed to grow to reach 50-60% confluency. One ml of the virus soup (1:5 dilution) and 8ug/ml of polybrene (Sigma) was added to cells and incubated for six hours. Cells were maintained under puromycin (Sigma) selection.

2.3.8 shRNA mediated knockdown of *NRBP1* in HNSCC cells

We retrieved shRNA sequences targeting human *NRBP1* from TRC (The RNAi Consortium) library database located in sh1 (3' UTR) and sh2 (CDS). Target sequences of NRBP1 shRNA constructs: sh1 (TRCN0000001437), 5'-CCCTCTGCACTTTGTTTACTTCT-3'; sh2 (TRCN0000001439), 5'-TGTCGAGAAGAGCAGAAGAATCT-3'. shGFP target sequences is 5'-GCAAGCTGACCCTGAAGTTCAT-3'. p-LKO.1 GFP shRNA was a gift from David Sabatini (Addgene plasmid # 30323) [205]. Cloning of shRNA oligos were done using AgeI and EcoRI restriction site in p-LKO.1 puro constructs. Bacterial colonies obtained screened using PCR and positive clone were sequence verified using Sanger sequencing. Lentiviral production and stable cell line generation performed as described earlier [206]. In brief,

Lentivirus were produced by transfection of shRNA constructs and two helper vector in 293T cells as described [207]. Transduction was performed in HNSCC cells by incubating for 6 hours in presence of 10µg/ml polybrene and post infection media was replaced with fresh media. Puromycin selection was performed two days post infection in presence of 1µg/ml. Puromycin selected cells were harvested and total cell lysate prepared and expression of NRBP1 was analysed using anti-NRBP1 antibody (Santa Cruz Biotechnology; sc-390087) and GAPDH (Santa Cruz Biotechnology; sc-32233).

2.3.9 Soft Agar colony formation Assay

The cells were harvested 48 hours after transfection, and an equal number of viable cells were plated onto soft agar after respective treatments for determination of anchorage-independent growth. For analysis of growth in soft agar, 5×10^3cells were seeded in triplicate onto a six well dish (Falcon) in 3 ml of complete medium containing 0.33% agar solution at 37 °C. Cells were fed with 500µl of medium every 2 days. From each well randomly 10 field images were taken using Phase contrast Inverted microscope (Zeiss axiovert 200m) and colonies were counted manually.

2.3.10 Growth Curve analysis

The 25000 cells/well were seeded in 24 well plates and growth was assessed post day 2, 4 and 6 by counting the cells using a haemocytometer. Percent survival were plotted at day 4 relative to day2 and later normalized against scrambled or empty vector control.

2.3.11 Western blot analysis

Cells were lysed in RIPA buffer and protein concentration was estimated using BCA method [208]. 50 and 100 µg protein was used for NIH-3T3 and HNSCC cell lines western analysis.

The protein was separated on 10% SDS-PAGE gel, transfer was verified using Ponceau S (Sigma), transferred on nitrocellulose membrane and blocked in Tris-buffered saline containing 5% BSA (Sigma) and 0.05% Tween-20(Sigma). Later, blots were probed with anti-NRBP1 (Santa Cruz Biotechnology; sc-390087), anti-total ERK1/2 (Cell signaling; 4372S), anti-Phospho ERK1/2 (Cell signaling; 4370S) and anti- GAPDH antibody (Santa Cruz Biotechnology; SC-32233). The membranes were then incubated with corresponding secondary HRP-conjugated antibodies (Santa Cruz Biotechnology, USA) and the immune complexes were visualized by Pierce ECL (Thermo Scientific, USA) according to manufacturer's protocol. Western blot experiments were performed as independent replicates.

2.3.12 Statistical analysis

Chi-square and t-test were performed using R programming language and GraphPad Prism. A p-value cut-off of 0.05 was used for gene expression, copy number and variant analysis.

2.4 Results

We characterized genetic alterations underlying four head and neck cancer cell lines followed by TCGA dataset to identify cumulative significance of biologically relevant alterations by integrating copy number, expression and point mutation data. The genomic analysis part of this work was in parallel and independently carried out in the lab by another graduate student (Mr. Pratik Chandrani) as thesis work so it's not discussed here in my thesis. My thesis part mainly involves biological interpretation, validation and functional characterization of biologically relevant alterations in HNSCC cells.

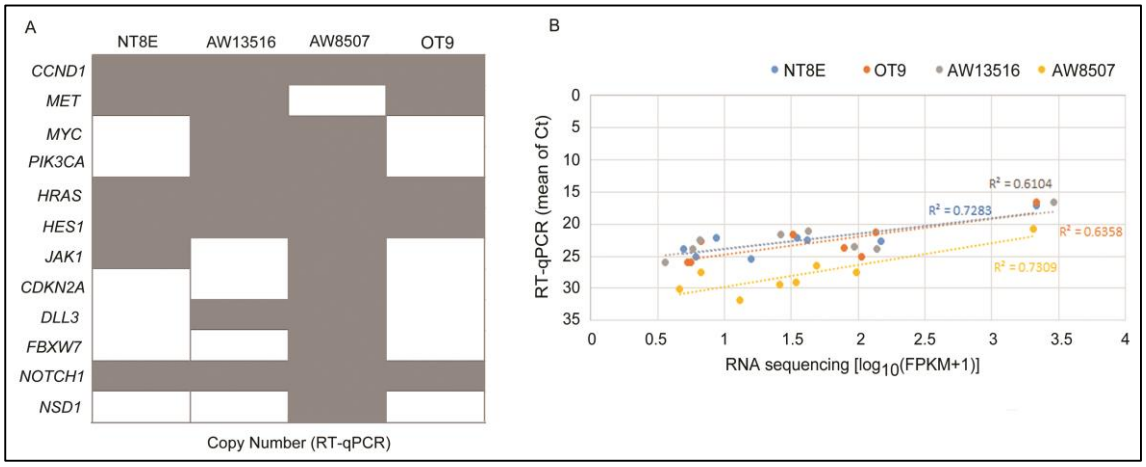
2.4.1 Characterization of four HNSCC cell lines established from Indian patients

Given that higher accumulative effect of individual genes can be reckoned by integrative analysis, we argue that these alterations can possibly be determined even with fewer samples. As a proof of principle, we performed an integrated characterization of karyotype analysis, copy number analysis, whole transcriptome and exome sequencing of 4 HNSCC cell lines (AW13516, AW8507, NT8e, and OT9) previously established from Indian patients. In brief, significantly altered chromosomal segments based on copy number analysis were filtered based on nucleotide variant information and aberrant expression of transcripts to allow prioritization of regions harboring either deleterious mutation or expressing the transcript at significantly high levels, in addition to the stringent intrinsic statistical mining performed for each sample.

2.4.1.1 Copy number validation in HNSCC cell lines

SNP array was performed to define the copy number alterations of HNSCC cell lines (AW13516, AW8507, NT8e, and OT9 cells). We identified copy number aberrations previously described in HNSCC including loss of copy number and loss of heterozygosity

(LOH) at 3p, 8p, and 9p which are known to be associated with advanced stage of tumors [209-211]. Copy number gain on 11q is known to be frequently altered in advanced stage HNSCC tumors [212]. The copy number alterations were validated using blood as normal and Type-it CNV Reference Primer Assay. We observed amplification of known oncogenes (such as *EGFR* in AW13516 and OT9; *MYC* in AW13516 and AW8507 cells; *JAK1* in NT8e, AW8507; *NSD1* in AW8507; and *MET* in AW13516 and OT9). Several hallmark genes were found to be amplified in cell lines such as *CCND1*, *NOTCH1*, and *HES1* in all four cells; *PIK3CA* in AW13516, AW8507; deletion of *CDKN2A* in AW13516; *FBXW7* in NT8E, AW13516 and OT9 cells were detected and validated by real time PCR (Figure 2.1A) in each cell line. In contrast, we found amplification of *NOTCH1* and its target gene *HES1* across HNSCC cell line, as opposed to previously reported frequent inactivation of Notch receptors [80, 127].



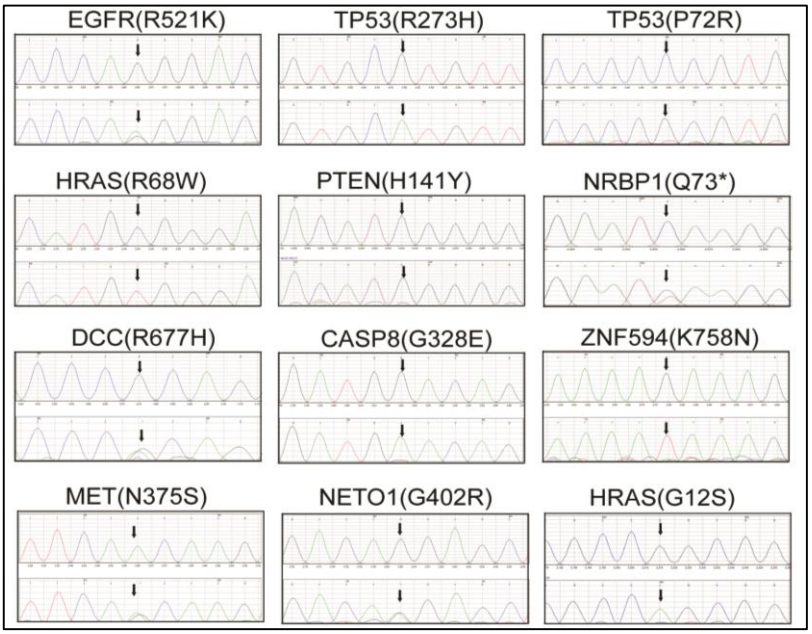
VII Figure 2.1: DNA copy number and gene expression validation using qPCR in HNSCC cells
(A) Schematic representation of copy number changes in key hallmark genes identified by SNP array and validated by qPCR. Grey box indicates validated and white box indicates invalidated copy number change in respective cell line. **(B)** Scatter plot representation of hallmark gene's expression. X-axis shows gene expression quantified by RNA sequencing and Y-axis shows gene expression quantified by RT-qPCR. Genes showed positive correlation between RNA sequencing and RT-qPCR.

2.4.1.2 Gene expression validation in HNSCC cell lines

Whole transcriptome sequencing revealed 17,067, 19,374, 16,866 and 17,022 genes expressed in AW13516, AW8507, NT8e and OT9 respectively. Over expression of hallmark of HNSCC such as *CCND1, MYC, MET, CTNNB1, JAK1, HRAS, JAG1, and HES1* and down regulation of *FBXW7, SMAD4* in at least 3 cell line were observed and validated by quantitative real time PCR. A positive correlation was observed between transcriptome FPKM and qPCR Ct values (Figure 2.1B).

2.4.1.3 Mutation validation in HNSCC cell lines

All the cell lines were sequenced for whole exome at about 80X coverage using Illumina HiSeq. The coding part of the four cell line genome consist 28813, 47892, 20864 and 25029 variants in AW13516, AW8507, NT8e and OT9 cell line, respectively. Filtering of known germline variants (SNPs) and low quality variants left 5623, 4498, 2775, 5139 non-synonymous variants in AW13516, AW8507, NT8e and OT9 cell line, respectively. Of 20 HNSCC hallmark variants predicted as deleterious by two of three algorithms used for functional prediction [213-215], 17 variants could be validated by Sanger sequencing including: *TP53* (R273H), *TP53* (P72R), *PTEN* (H141Y), *EGFR* (R521K), *HRAS* (G12S and R78W), and *CASP8* (G328E) (Figure 2.2C).



VIII Figure 2.2 Sanger sequencing based validation of mutations in HNSCC cells

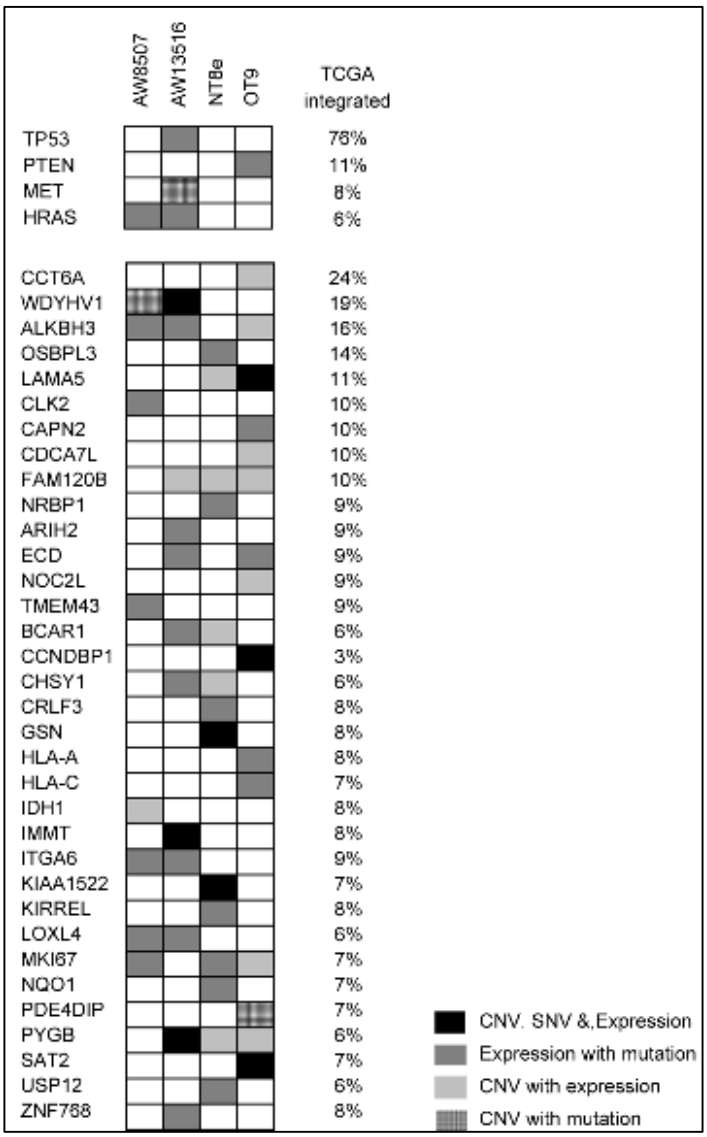
Several high-quality point mutations identified by exome sequencing were validated by Sanger sequencing. Sanger sequencing trace were visualized using Mutation Surveyor software showing reference sequence trace in upper panel and mutant sequencing trace in below panel. Arrows represent mutation position with reference and mutated base indicated above the arrow.

2.4.2 Integrated analysis identifies hallmark alterations in HNSCC cell lines

Briefly, genes harbouring two or more type of alterations were selected: harbouring positive correlation of focal copy number and gene expression; or those harbouring point mutations with detectable transcript harbouring the variant—based on which, we identified 38 genes having multiple types of alterations (Figure 2.3). These include genes known to have somatic incidences in HNSCC: *TP53*, *HRAS*, *MET* and *PTEN*. We also identified *CASP8* in AW13516 cell line which was recently identified as very significantly altered by ICGC-India team in ~50 Indian HNSCC patients [125]. Among the novel genes identified, of genes with at least one identical mutation previously reported include a pseudokinase Nuclear receptor binding protein (*NRBP1*) harboring heterozygous truncating mutation (Q73*) in NT8e cells, identical to as reported in lung cancer and altered in other cancers [216, 217].

IX Figure 2.3: Integrative genomic landscape of HNSCC.

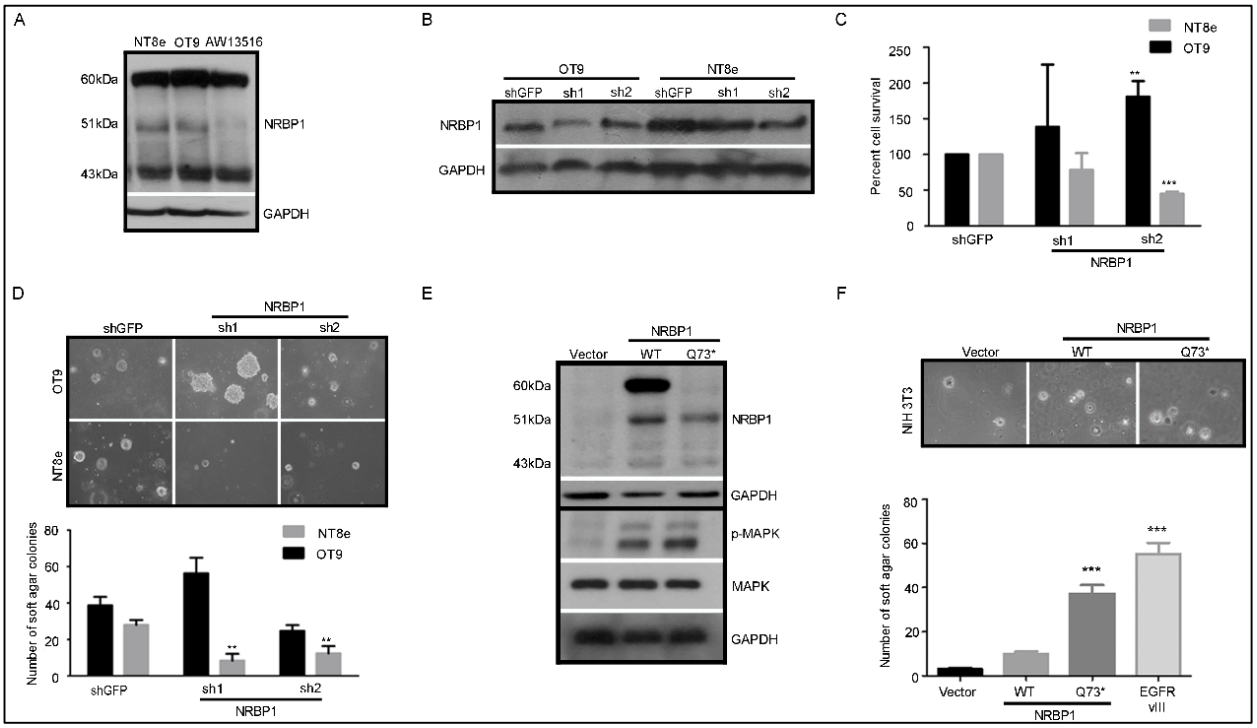
Schematic representation of 38 genes identified by integrated analysis of four HNSCC cell lines and their incidence in 279 HNSCC samples from TCGA study.



2.4.3 Mutant NRBP1 is required for tumor cell survival and is oncogenic in NIH-3T3 cells

NRBP1 encodes for three different nuclear receptor binding protein isoform using three alternative translational initiation sites of 60kDa, 51kDa and 43kDa [218], as were observed in 2 of 3 HNSCC cells (Figure 2.4A). To determine whether expression of mutant *NRBP1* is required for tumor cell survival, we tested shRNA constructs in two HNSCC cells expressing all three forms of WT *NRBP1* (OT9 cells) and mutant *NRBP1* (NT8e cells). We demonstrate that even partial knockdown of mutant *NRBP1* expression in the NT8e cells, but not WT *NRBP1* expression in the OT9, significantly inhibited anchorage-independent growth and cell survival (Fig 2.4B-D). We next tested the oncogenic role of *NRBP1*. mRNAs harboring premature termination (nonsense) codons degraded by Nonsense-mediated mRNA decay (NMD) [219]. However, mRNAs with nonsense mutations in the first exon with alternative translation initiation site are

selectively are known to bypass NMD [220]. When ectopically expressed in NIH-3T3 cells, mutant *NRBP1* transcript escape non-sense mediated degradation as determined by real time PCR (Figure 2.5). All three isoform of NRBP1 were detected in NIH-3T3 cells expressing wild type *NRBP1* cDNA. However, only two isoform of 51kDa and 43kDa were detected in cells transfected with mutant *NRBP1* cDNA (Figure 2.4E upper panel). The over expression of the mutant *NRBP1* in NIH-3T3 cells conferred anchorage-independent growth, forming significantly higher colonies in soft agar than cells expressing wild type *NRBP1* (Figure 2.4F). Transformation of NIH-3T3 cells by *NRBP1* over expression was accompanied by elevated phosphorylation of the MAPK (Figure 2.4E lower panel).

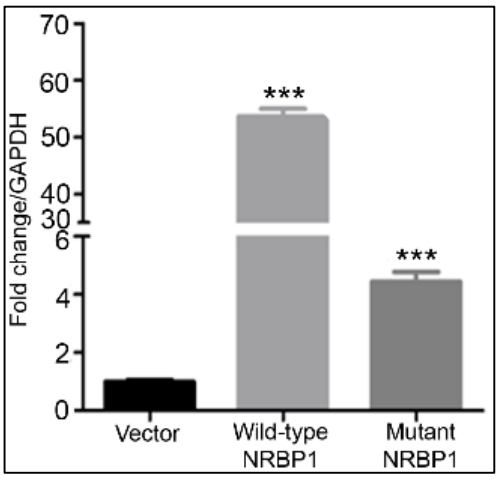


X Figure 2.4: Mutant *NRBP1* is required for tumor cell survival and is oncogenic
Knockdown of mutant *NRBP1* expression with shRNA inhibits transformation and survival of HNSCC cell lines. **(A)** Western blot analysis of total NRBP1 expression level in HNSCC cell lines. NRBP1 encodes for three different nuclear receptor binding protein isoform: 60kDa, 51kDa and 43kDa in NT8e cells (lane 1), OT9 cells (lane 2). AW13516 cells express only two isoforms (lane 3). **(B)** Western blot representation of NRBP1 partial knockdown by two independent hairpins (sh*NRBP1*#1 and sh*NRBP1*#2) in NT8e and OT9 cells. The hairpin constructs inhibit cell survival as assessed by cell counting as described in methods. **(C)** and anchorage-independent growth as assessed by colony formation in soft agar **(D)** in the NT8e cells harboring an *NRBP1* Q73* mutation, but not the OT9 cells, which express WT *NRBP1*.

shGFP, a hairpin specific for green fluorescent protein, was used as a negative control. All results are normalized to survival or colony formation by cells infected with empty vector. Images were taken at 10x magnification. **(E)** Western blot analysis of NRBP1 wild type and mutant (Q73*) in NIH-3T3 cells. Vector control, NIH-3T3 clones overexpressing wild-type and mutant *NRBP1* is shown. *NRBP1* Q73* mutation expresses 51 kDa isoform, while two isoform of 60kDa and 51kDa can be seen in 3T3 clones overexpressing wild-type NRBP1. Western blot analysis of total and phosphorylated MAPK in NIH-3T3 clones expressing *NRBP1* wild type (lane 2) and mutant *NRBP1* (Q73*). **(F)** Representative images (taken at 20x) of soft agar colony formed by NIH-3T3 cells expressing *NRBP1* wild type and Q73*. Bar graph representation of number of soft agar colonies formed by stable NIH-3T3 clones. NIH-3T3 cells expressing EGFR VIII was used as positive control and relative comparison of transforming ability of mutant *NRBP1*. *** P-value <0.0001, ** P-value <0.001.

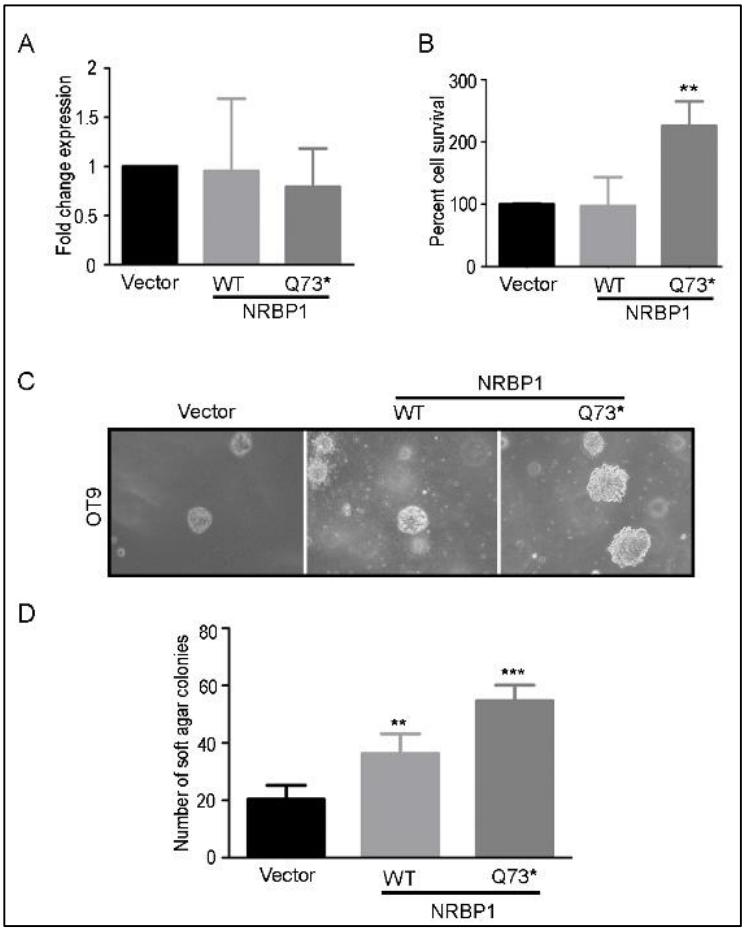
2.4.4 Stable overexpression of mutant *NRBP1* (Q73*) cDNA in OT9 cell line leads to increased cell survival and anchorage-independent growth

To test whether mutant *NRBP1* (Q73*) is acting in dominant negative manner, we have performed experiments with OT9 cells (wild type *NRBP1*) transfected with mutant *NRBP1* cDNA. OT9 cells transfected with mutant *NRBP1* cDNA consistently display proliferation advantage and enhanced anchorage-independent growth over OT9 cells transfected with wild-type *NRBP1* (Figure 2.6B-D). However, we could not experimentally validate the overexpression of mutant or wild-type NRBP1 by western or qRT-PCR analysis, possibly because all the four HNSCC cells analyzed show high endogenous expression of NRBP1 protein (Figure 2.6A). The wild type and mutant *NRBP1* constructs overexpression was validated in 293FT cells by qRT-PCR analysis (Figure 2.5).



XI Figure 2.5: *NRBP1* cDNA overexpression in NIH-3T3 cells.

qPCR analysis of *NRBP1* gene expression in NIH-3T3 stably expressing wild and mutant. Data was normalized against *GAPDH* and fold change plotted. P value <0.0001 is denoted as ***.



XII Figure 2.6: Ectopic expression mutant NRBP1 leads to increased cell survival and anchorage-independent growth of OT9 head and neck cancer cells.

Ectopic expression of mutant *NRBP1* increases cell survival and soft agar colony formation potential of OT9 cells. **(A)** qRT-PCR confirmation of *NRBP1* expression stably expressing NRBP1 (wild type and mutant) in OT9 cells. **(B)** Ratio of cell survival of day4 over day2 is plotted. Ectopic expression of mutant *NRBP1* (Q73*) leads to increase in cell survival. **(C)** Representative images (taken at 10x) of soft agar colony formed by OT9 cells expressing *NRBP1* wild type and mutant Q73*. **(D)** Bar graph representation of number of soft agar colonies formed by stable OT9 cells expressing *NRBP1* (wild type and mutant). Significantly higher number of soft agar colonies were observed in cells expressing mutant (Q73*) *NRBP1* in OT9 cells. Student t-test was performed to access statistical significance and P-value <0.05 considered as threshold for significance. *** P-value <0.0001, ** P-value <0.001.

2.5 Discussion

We have characterized genetic alterations of unknown somatic status underlying four head and neck cancer cell lines of Indian origin patient by subjecting them to a thorough karyotype based characterization, SNP array based analysis, whole exome capture sequencing, and mRNA sequencing.

Integrated analysis of the cell lines establish their resemblance to primary tumors. Consistent with literature, most frequent copy number gains in head and neck cancer cells in this study were observed at 2q, 3q, 5p and 7p, and deletions at 3p, 9p, 10p, 11q, 14q, 17q and 19p, as reported earlier [32, 221]. Integration of multiple platform with the copy number variation, allowed us to identify the functionally relevant alterations including several hall marks genes known to be involved in HNSCC, viz. *PIK3CA, EGFR, HRAS, MYC, CDKN2A, MET, TRAF2, PTK2* and *CASP8*. Of the novel genes, *JAK1* was found to be amplified in two of the cell lines and overexpressed in all 4 HNSCC cells; *NOTCH1* known to harbor inactivating mutations in HNSCC [125, 127] was found to be amplified in all 4 and overexpressed in 2 of 4 HNSCC cells, known to be play dual role in a context dependent manner [222].

We also observed mutations in several novel genes such as *CLK2, NRBP1, CCNDBP1, IDH1, LAMA5, BCAR1, ZNF678,* and *CLK2*. Of these, genes with at least one identical mutation previously reported include *NRBP1* (Q73*), a pseudo kinase, found in NT8e cells, earlier reported in lung and other cancers [216, 217], with an overall 9% cumulative frequency alteration in TCGA HNSCC dataset (Supplementary Figure S8). Of 48 pseudo kinases known in human genome, several have been shown to retain their biochemical catalytic activities despite lack of one or more of the three catalytic residues essential for its kinase activity, with their established roles in cancer [223-225].

Interestingly, several activating mutant alleles of *NRBP1 homolog Drosophila Madm* (Mlf1 adapter molecule) *3T4* (Q46*); *2U3* (C500*); *3G5* (Q530*); *7L2* and *3Y2* (that disrupts splice donor site of first exon) are known, wherein alternative translation start codons is similarly suggestive for a varying degree of pinhead phenotype severity associated with the mutant alleles [218, 226]. Studies in the fruit fly have provided important insights into mechanisms underlying the biology of growth promoting *NRBP1 homolog Drosophila Madm*. A recent study suggests *Drosophila Madm* interacts with *Drosophila bunA* that encodes a gene homologous to human *Transforming Growth Factor-β1 stimulated clone-22 TSC-22* [226]; that were later shown to interact even in mammalian system[227]. Interestingly, mammalian tumor suppressor *TSC-22* is known to play an important role in maintaining differentiated phenotype in salivary gland tumors [228], a subtype of head and neck cancer. More recently, studies have shown poor clinical outcomes are associated with *NRBP1* over expression in prostate cancer [227]. We provide the first functional analysis of mutant *NRBP1* and establish that NIH-3T3 cells expressing the mutant *NRBP1* enhance their survival and anchorage independent growth, while its knock down diminishes survival and anchorage-independent growth by oral cancer cells expressing activating *NRBP1* mutations. Thus, NT8e cells harboring mutant *NRBP1* was found to be consistent with its suggestive role in prostate cancer biology and other model organisms. Interestingly, *NRBP1* has also been shown to be involved in intestinal progenitor cell homeostasis with tumor suppressive function [229], suggesting its role is specific to the cellular context. This study identifies *NRBP1* mutant to play an oncogenic role in head and neck cancer. However, in depth systematic sequencing of *NRBP1* in a wide variety of tumor types may help indicate utility of NRBP1 inhibition in human cancer.

In overall, this study underscores integrative approaches through a filtering strategy to help reckon higher cumulative frequency for individual genes affected by two or more alterations

to achieve the threshold for statistical significance even from fewer samples. The integrative analysis as described here, in essence, is based on a linear simplified assumption of disease etiology that variation at DNA level lead to changes in gene expression causal to transformation of the cell. As a major deficiency, only genes that are subject to multiple levels of biological regulation are likely to be determined by this approach than genes that are primarily altered by single alteration like amplification or over expression. Beyond validation the genomic alteration using orthologous method, this we demonstrated that *NRBP1*, a novel gene identified this analysis is playing oncogenic role in NIH-3T3 and required for the survival of HNSCC cells.

Chapter 3

TMC-SNPdb: An Indian germline variant database derived from whole exome sequences

(as published in *Database (Oxford)* Jul 9;201; 10.1093/database/baw104)

3. Chapter 3. TMC-SNPdb: An Indian germline variant database derived from whole exome sequences

(as published in Database (Oxford) Jul 9;201; 10.1093/database/baw104)

3.1 Abstract

Cancer is predominantly a somatic disease. A mutant allele present in a cancer cell genome is considered somatic when it's absent in the paired normal genome along with public SNP databases. The current build of dbSNP, the most comprehensive public SNP database, however inadequately represents several non-European Caucasian populations, posing a limitation in cancer genomic analyses of data from these populations. We present the **T**ata **M**emorial **C**entre-SNP **d**ata**b**ase (TMC-SNPdb), as the first open source, flexible, upgradable, and freely available SNP database (accessible through dbSNP build 149 and ANNOVAR)--- representing 114,309 unique germline variants--- generated from whole exome data of 62 normal samples derived from cancer patients of Indian origin. The TMC-SNPdb is presented with a companion subtraction tool that can be executed with command line option or using an easy-to-use graphical user interface (GUI) with the ability to deplete additional Indian population specific SNPs over and above dbSNP and 1000 Genomes databases. Using an institutional generated whole exome data set of 132 samples of Indian origin, we demonstrate that TMC-SNPdb could deplete 42%, 33% and 28% false positive somatic events post dbSNP depletion in Indian origin tongue, gallbladder, and cervical cancer samples, respectively. Beyond cancer somatic analyses, we anticipate utility of the TMC-SNPdb in several Mendelian germline diseases. In addition to dbSNP build 149 and ANNOVAR, the TMC-SNPdb along with the subtraction tool is also available for download in the public domain at http://www.actrec.gov.in/pi-webpages/AmitDutt/TMCSNP/TMCSNPdp.html.

3.2 Introduction

Somatic mutations sequentially accumulate in cancer cell genomes. In addition, a typical cancer genome contains several polymorphic 'normal' germline variants [64, 104, 105]. Subtracting the tumor DNA variants against matched normal DNA derived from the same individual and those polymorphic in the population is, therefore, essential to identify an exclusive somatic event [106]. Apropos, a critical aspect of any tumor genome sequence analysis involves depletion of paired normal variants followed by depletion of residual variants from public databases of common single nucleotide polymorphism (SNPs) such as dbSNP [107] and 1000 Genomes Project [108]. A sequence variant not observed in matched normal derived genome sequence and absent from public SNP database is considered somatic in origin. Adopting such an analytical approach ensures filtering of paired-germline and population-specific polymorphic variants from dbSNP and 1000 Genomes Project for Caucasian population [109].

However, despite depletion against dbSNP, unknown SNPs especially those with lower minor allele frequency not represented in dbSNP, are likely to confound somatic mutation analyses in studies involving non- Caucasian and non-European Caucasian populations (5). Two exhaustive initiatives addressing this issue are the publicly available exome variation datasets: NHLBI Exome Sequencing Project (ESP) (https://esp.gs.washington.edu/EVS/) and Exome Aggregation Consortium (ExAC) (http://exac.broadinstitute.org/) [110]. Information gathered from these studies is an integral part of variant annotation tools like Annovar [111].

Multiple studies such as the Indian Genome Variation Consortium [112, 113] and HUGO Pan-Asian SNP Consortium [114] have described the genomic distinctiveness of Indian population based on varying allele frequency of known SNPs, complex origin, genetic diversity [115-118],

and high variation of male lineages (Y-chromosome) within the population [119, 120]. However, a concerted effort to comprehensively identify and catalogue novel SNPs present exclusively in Indian population is yet to be undertaken. Lack of Indian specific SNP database has been an important impediment in cancer research, especially in efforts to discover bona fide novel somatic mutations.

Here, we describe **T**ata **M**emorial **C**entre-SNP **d**ata**b**ase "TMC-SNPdb" as the first, open source, freely available database of unique germline variants obtained from whole exome data of 62 'normal' samples from tongue, gallbladder, and cervical cancer patients of Indian origin. "TMC-SNPdb" is presented with an easy-to-use graphic user interface feature to enable researchers to call true somatic mutations by depleting against Indian population specific SNPs, in addition to those already catalogued in dbSNP and 1000 Genomes databases. We demonstrate that "TMC-SNPdb" effectively filters false positive somatic events across 75 tumor whole exome data.

3.3 Materials & Methods

3.3.1 Ethical approval and informed consent

The sample set and study protocol was approved by Institutional Review Board (project # 116 for cervical adenocarcinoma samples; project # 88 for head and neck cancer samples, project 104 for gallbladder cancer samples). Cervical squamous carcinoma whole exome data has been described earlier [150]. Written informed consent was obtained from all patients.

3.3.2 Extraction of DNA

All 'normal' tissue samples under study were verified by an onco-pathologist to not harbour any cancer. A total of 62 samples 'normal' samples (16 peripheral venous blood and 46

adjacent normal tissue) were obtained for analysis: peripheral venous blood from patients with cervical squamous cell carcinoma (n=10), cervical adenocarcinoma (n=18) (adjacent normal tissue; n=12 and peripheral venous blood; n=6) and adjacent normal tissue from patient with tongue squamous cell carcinoma (n=23) and gallbladder (n=11) were obtained from Tata Memorial Hospital (TMH). Genomic DNA from tissues was extracted using DNeasy blood and tissue DNA extraction kit (Qiagen) according to manufacturer's instructions. Quantification of DNA was assessed using Nanodrop 2000c Spectrophotometer (Thermo Fisher Scientific Inc.) and DNA integrity was determined by resolving on 0.8% Agarose gel. DNA was also quantified using Qubit ds DNA BR assay kit (Life Technologies, USA). DNA samples showing both DNA concentration >50 ng/µl and intact DNA visualized on agarose gel were used for whole exome sequencing.

3.3.3 Exome capture, library preparation and sequencing

Three different library preparation kits were used to prepare libraries for different tumor types (Appendix I). First, TruSeq Exome Enrichment kit (v2 and v3, Illumina) was used to capture 62Mb region (>3,40,000 probes) of human genome comprising 201,121 exons representing 20,974 gene sequences, including 5'UTR, 3'UTR, microRNAs and other non-coding RNA. For exome library preparation, two microgram genomic DNA was sheared using Covaris (Covaris Inc) for generating fragment sizes of 200-300bp. Fragments end repairing, purification, A- tailing, adaptor ligation and quality control steps were carried out using TruSeq DNA Sample Prep Kit (Illumina) following manufacturer's instructions. Qualitative and quantitative analysis of genomic DNA libraries were performed using High Sensitivity DNA chip on 2100 Bioanalyzer (Agilent) and qPCR with KAPA Library Quant Kit (Kapa Biosystems). Exome enrichment was done by incubation at 93ºC for 1 min (decreasing 2ºC per

cycle for 18 cycles) followed by 58ºC for 19 hours in ABI 9700 PCR system (Life Technologies) using 500 ng of genomic libraries.

Second, NimbleGen SeqCap EZ Exome Library (v3.0, Roche) targeting 64Mb of the human genome was also used for library preparation. The protocol was adopted from the manufacture's application note (http://www.nimblegen.com/products/lit/NimbleGen_SeqCap_EZ_SR_Pre-Captured_Multiplexing.pdf). Sequencing libraries were exome captured and later quality-controlled using a bioanalyzer (Agilent 2100) and libraries were qPCR quantified using KAPA Library Quant Kit (Kapa Biosystems) prior to cluster generation on an Illumina cBOT.

Third, SureSelect Human All Exon Kit, v5 (Agilent Technologies, Santa Clara, CA, United States) was also used to capture 50Mb of the human genome using > 5,50,000 probes. One microgram of genomic DNA was utilized for library preparation and a similar protocol was followed as previously stated. Eluted exome-enriched library fragments were PCR amplified and purified.

qPCR quantified 7 pmol of 6-plex DNA library pool was loaded per lane on flow cell (Flow Cell v3) to generate clusters using TruSeq PE Cluster Kit v3-cBot-HS kit and clustered flow was sequenced for 201 and 301 cycles on HiSeq-1500 and NextSeq System (Illumina) using TruSeq SBS Kit v3 (Illumina), respectively.

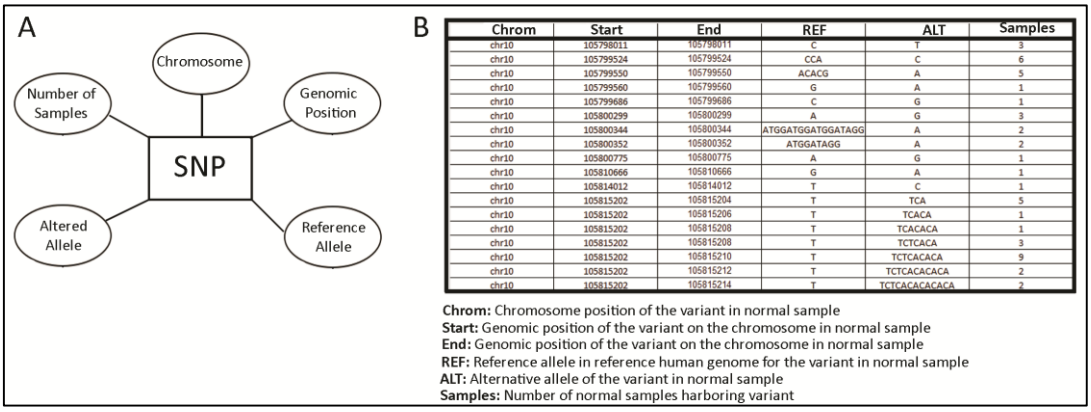
3.3.4 Exome sequencing variant analysis for TMC-SNP database

Paired-end raw sequence reads were mapped to human reference genome (build hg19) using BWA v. 0.6.2 [230]. Quality control analysis of bam files was carried out using qualimap (v0.7.1) [231]. Mapped reads were then used to identify and remove PCR duplicates using

Picard tools v.1.74 (http:broadinstitute.github.io/picard/). Base quality score recalibration and indel re-alignment were performed and variants were called from each sample separately using GATK Unified Genotyper (version 2.5-2) [232].

3.3.5 Development of TMC-SNP database

To restrict our analysis to high quality germline variants we applied filters of minimal base coverage and recurrence in cohort. In house developed scripts (Awk and Perl) were used to merge all 62 VCF files from normal tissues and mutational recurrence was calculated. We applied a standard filter of coverage ≥5 reads for altered alleles. Additionally, we included variants with coverage ≤5 but recurrent in ≥4 normal samples. Using these filters, we identified high quality variants in the dataset. High quality variants were further annotated using COSMICdb (version 68) [233] and dbSNP (version 142) [107]. Remaining variants were further depleted against dbSNP and COSMICdb to remove all known somatic and germline variants. Finally, all remaining variants constitute the TMC-SNP database. A detailed schema of resource and data representation is provided in Figure 3.1.



Chrom	Start	End	REF	ALT	Samples
chr10	105798011	105798011	C	T	3
chr10	105799524	105799524	CCA	C	6
chr10	105799550	105799550	ACACG	A	5
chr10	105799560	105799560	G	A	1
chr10	105799686	105799686	C	G	1
chr10	105800299	105800299	A	G	3
chr10	105800344	105800344	ATGGATGGATGGATAGG	A	2
chr10	105800352	105800352	ATGGATAGG	A	2
chr10	105800775	105800775	A	G	1
chr10	105810666	105810666	G	A	1
chr10	105814012	105814012	T	C	1
chr10	105815202	105815204	T	TCA	5
chr10	105815202	105815206	T	TCACA	1
chr10	105815202	105815208	T	TCACACA	1
chr10	105815202	105815208	T	TCTCACA	3
chr10	105815202	105815210	T	TCTCACACA	9
chr10	105815202	105815212	T	TCTCACACACA	2
chr10	105815202	105815214	T	TCTCACACACACA	2

Chrom: Chromosome position of the variant in normal sample
Start: Genomic position of the variant on the chromosome in normal sample
End: Genomic position of the variant on the chromosome in normal sample
REF: Reference allele in reference human genome for the variant in normal sample
ALT: Alternative allele of the variant in normal sample
Samples: Number of normal samples harboring variant

XIII Figure 3.1: Schema of resource and data representation of TMC-SNP database.

(A) Schematic representation of information provided for each SNP in TMC-SNPdb. Schema of TMC-SNPdb for each SNP in the file. Attribute and data type format in the TMC-SNP database in the file. (**B)** A snap view of the table of TMC-SNPdb variant file showing different columns in the file. Description of each column in the table is given at the bottom of the table.

3.3.6 Application of TMC-SNP database in analyzing tumor samples

GATK (version 2.5-2) and MuTect (version 1.0.2) [189] were utilized to generate raw variants of tumor samples and filtered against its matched normal . Variants obtained from GATK and MuTect were merged and variants having ≥5 reads for altered allele were kept for further downstream analysis. Similar analysis was carried out for three cancer types. Comparison with dbSNP(version142) and COSMICdb(version 68) was performed using in-house developed scripts in Perl and Awk which were later used to calculate the percentage changes in variants in different cancer type post filtration with dbSNP and TMC-SNP database. Functional annotation of variants was performed using Oncotator (variant annotation tool) [234].

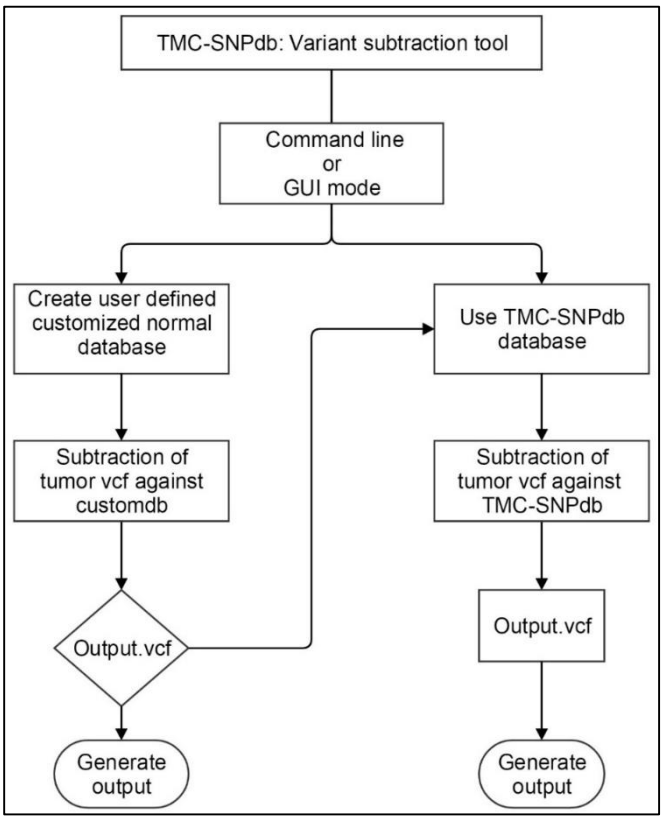
3.3.7 Germline variant subtraction program

TMC-SNPdb is distributed as a SQLite file containing variant information table. A companion tool for subtraction of germline variants from tumor sample has been developed in python (version 3.4). It depends on PyVCF (version ≥ 1.6) and sqlite3 python packages. The variants in TMC-SNPdb are characterized by a unique combination of chromosome number, genomic position, reference allele, altered allele for each variant and subtraction was carried out based on these unique fields for each variant in VCF file. The tool is an executable compatible with Linux operating system and has been tested on several Linux platform such Red Hat (version 6.5), Fedora (version 22) and Ubuntu (version 14.04). It can be executed using a command line interface ('TMC-SNP') or a graphical user interface (GUI) ('TMC-SNP_GUI'). The GUI mode additionally depends on TKinter python library (version ≥ 2.4). Moreover, the tool has a feature which lets users create their own germline variant database from VCF format files of normal samples. The output obtained from the tool is in VCF format. Detailed user manual with

snapshots of the GUI and schematic representation of overall usages are provided in Appendix III.

3.3.8 Availability of supporting data

The raw sequence data has been deposited at the ArrayExpress (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-4618), hosted by the European Bioinformatics Institute (EBI). "TMC-SNPdb" have been submitted to Annovar (http://annovar.openbioinformatics.org/en/latest/user-guide/download/) and dbSNP (http://www.ncbi.nlm.nih.gov/SNP/snp_viewTable.cgi?handle=TMC_SNPDB) for public access.



XIV Figure 3.2: Schematic representation of TMC-SNPdb variant subtraction tool usages.
Subtraction tool could be run via command line mode or graphical user interface (GUI) mode for subtraction. The user can create custom database by providing a list of normal VCF file and later subtract tumor VCF against it or can directly use TMC-SNPdb for subtracting the variant from given tumor sample VCF file. It generates output in VCF format.

3.4 Results

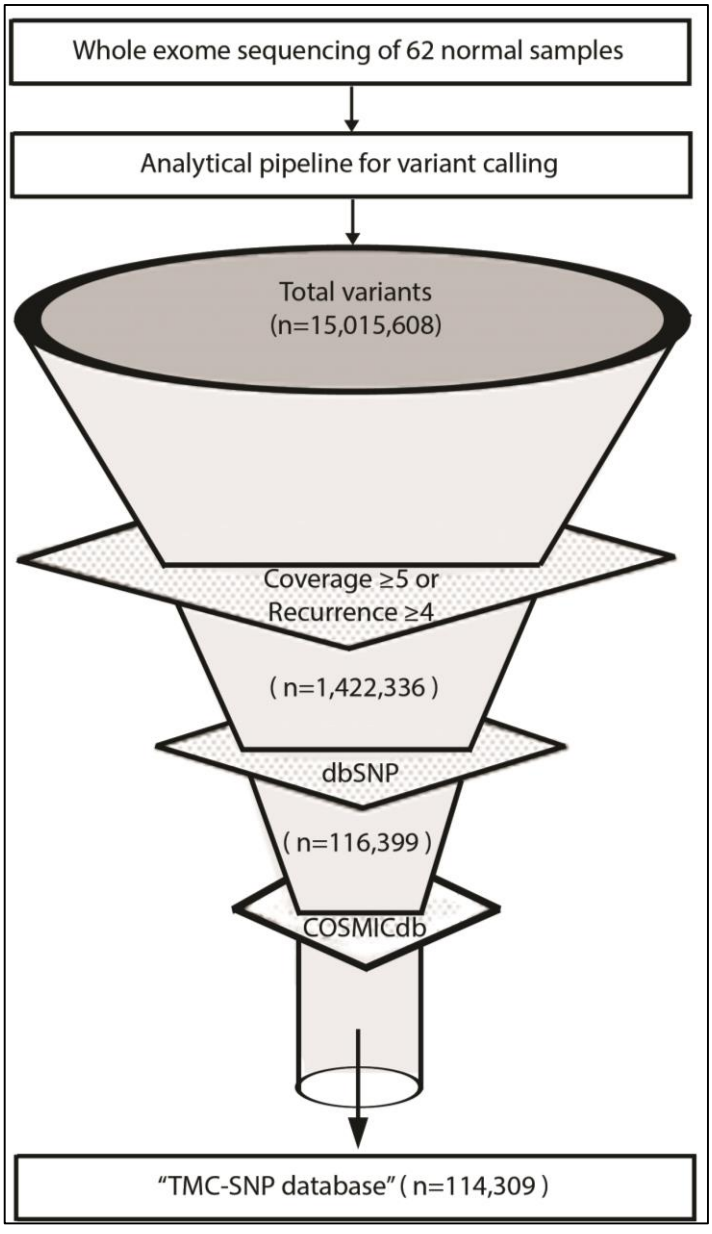
3.4.1 Development of TMC-SNP database

We analyzed whole exome sequencing at a median of 88x coverage for 62 normal samples derived from cancer patients, comparable with similar reports [235] as detailed in Appendix II. Of note, coverage among 4 of 62 samples were < 30x due to high duplication reads and low yield in these samples. Germline mutations were called using GATK [232]: a total of 15,015,608 germline variants were identified across the complete dataset. As shown in Figure 3.3, standard quality filters of minimal 5X coverage or recurrence in at least 4 samples for each variant led to about 90% reduction in raw variants (see methods for details). The remaining 1,422,336 variants of higher confidence were further depleted against dbSNP v142. 1,305,937 of 1,422,336 variants, constituting 92% SNPs were depleted. To remove variants known to be somatically associated with cancer in literature but figured as a germline event in our study (most likely due to inadequate or non-uniform coverage of their paired normal samples), we further depleted 2090 variants (2%) overlapping with COSMICdb with an assumption of these variants to be false somatic events in our data set. Finally, a total of 114,309 variants were identified after filtering with dbSNP and COSMICdb as a pool of previously unknown germline variants of high confidence recurring in the Indian population to constitute the "TMC-SNPdb".

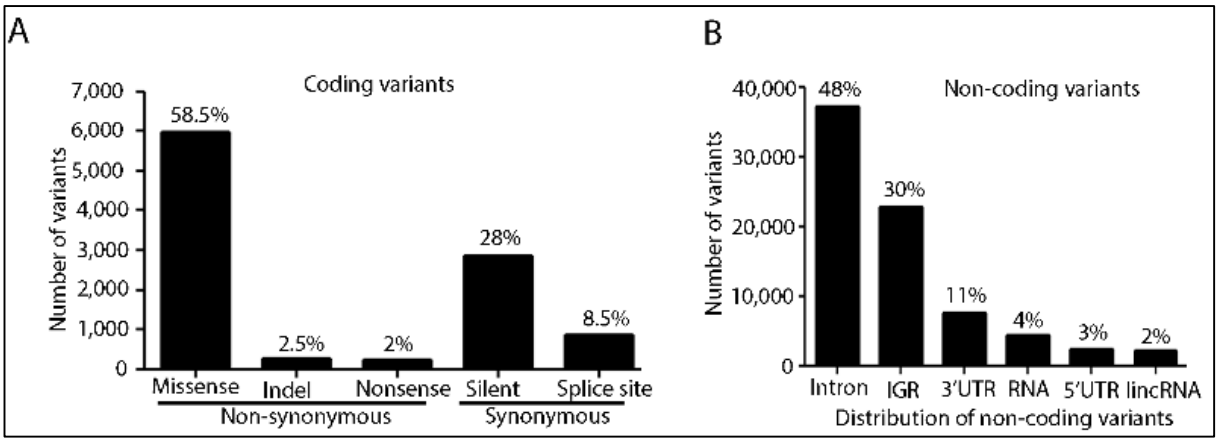
3.4.2 Characteristic features of TMC-SNP database

A total 114,309 variants were annotated using Oncotator for functional features [234]. A distribution pattern of coding (~17,973) and non-coding variants germline variants (~96,336) is shown in Figure 3.5. Of 17,973 coding variants, 11,466 were of non-synonymous (NS) (~63%) and 6,507 were synonymous variants (S) (~36%) with NS/S ratio 1.76, consistent with

previous reports for exome data from normal samples [236, 237]. Furthermore, we observed a high frequency of missense (~58%) and silent variants (~30%) as compared to indel (~3%), nonsense (~2%) and splice site (~6%) region Figure 3.4A.



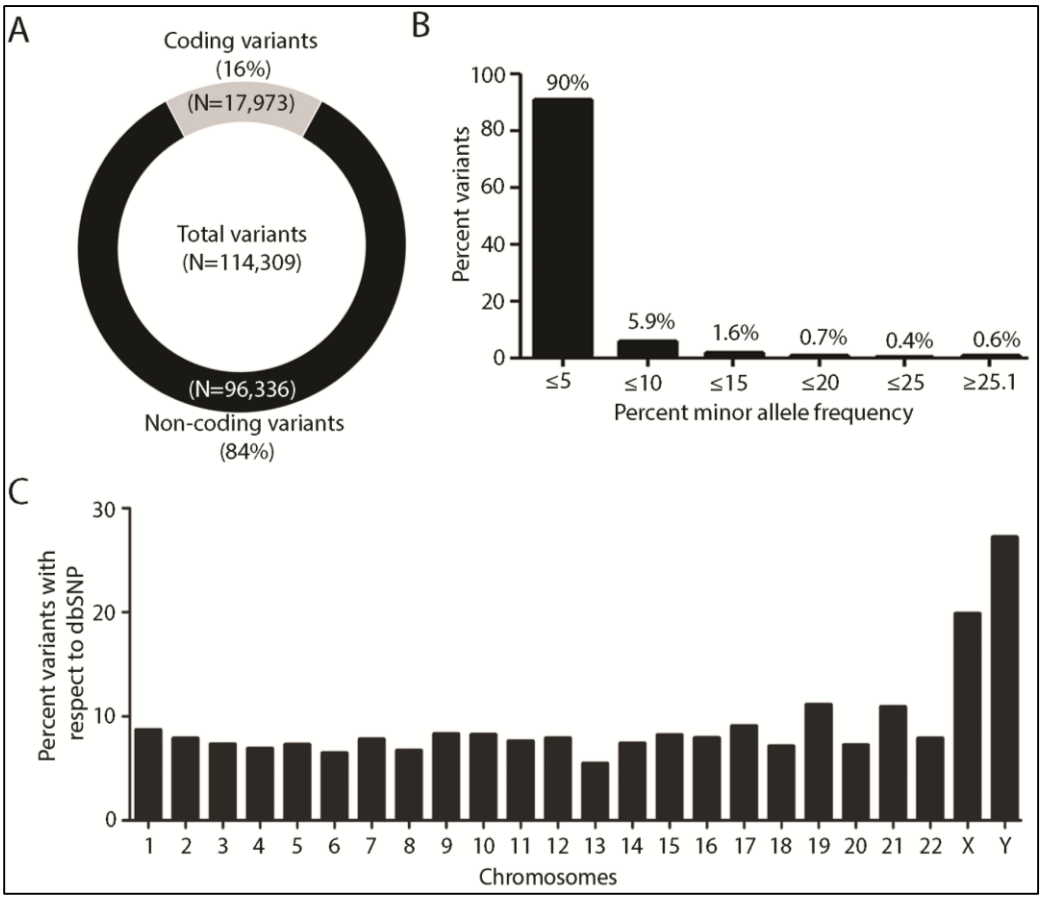
XV Figure 3.3: Development of TMC-SNPdb using whole exome sequencing.
Schematic flow representation of steps followed during development of TMC-SNP database. The whole exome sequencing of 62 normal tissue obtained from three different tissues of cancer patients was performed and analyzed using GATK (Genome Analysis Tool Kit) to generate VCF files. Raw variants obtained were further filtered using mentioned criteria to find a list of variants absent in dbSNP v142 and COSMICdb v68. Remaining variants constitutes the "TMC-SNPdb" shown at the end of the funnel.



XVI Figure 3.4: Distribution of variants obtained in TMC-SNPdb at coding and non-coding region of the genome.

(A) Coding variants obtained in TMC-SNPdb and bar graph representation of number and percent frequencies for various type of functional class. Percentage frequencies are denoted on the top of each bar. **(B)** Bar graph representation of a number of various types of non-coding variants obtained in TMC-SNPdb. Percentage frequencies are denoted on the top of each bar.

Of all the SNPs present in TMC-SNPdb, distribution varied across the genome as follows: protein coding exon (15.7%), intron (40%), IGR (25.8%), 3'UTR (9.5%), 5'UTR (2.37%), RNA (3.74%), and lincRNA (1.7%), consistent with earlier report from exome sequencing data Figure 3.4B. [238, 239]. Next, we computed the allele frequency of all 114,309 variants present in the TMC-SNPdb, across 62 samples. Given that TMC-SNPdb predominantly enlists low frequency germline variants prevalent among Indian population, similar to 1000 genomes and ExAC wherein about 99% of SNPs are estimated to have a minor allele frequency over 1% [110, 235], Similarly, in TMC-SNPdb more than 90% of variants present exist at a minor allele frequency ≤5% (Figure 3.5B).



XVII Figure 3.5: Overall overview of characteristic features of TMC-SNP database.

(A) Circle plot of coding and non-coding variants obtained in the dataset. **(B)** Percent minor allele frequency distribution of variants in "TMC-SNPdb" across 62 normal samples. Percentage frequencies are presented on the top of each bar. **(C)** Genome-wide distribution of percent frequency of variants obtained in each chromosome as compared to dbSNP database.

Furthermore, a comparative measure of variability added by the TMC-SNPdb variants to the known pool of SNPs per chromosome was reckoned following comparison with dbSNP variants across the genome. Interestingly, we found maximal variability at the Y-chromosome among 2418 of 8,885 SNPs (27%), while the distribution of the variants across the autosomal chromosomes was found to be uniformly distributed among 106,184 of 1,346,256 SNPs (7.6%) similar to the dbSNP (Figure 3.5C). Of note, variants at Y-chromosome tend to be more localized geographically than those of mitochondrial DNA (mtDNA) and autosomes, which is

reflective of the degree of inter-population genetic differences [240-242]. Y-chromosomes have been shown to harbor population specific unique haplotype in Indian population and have frequently been used as a marker for studying human demographic history [243, 244]. The higher variability at the Y-chromosome found in TMC-SNPdb is thus consistent with several earlier reports describing a high variation of male lineages within Indian population [119, 120] that further emphasizes the Indian specific characteristics of the TMC-SNPdb germline variants, and a need for distinct Indian specific germline database .

Finally, a significant characteristic feature of TMC-SNPdb is the companion subtraction tool with command line and GUI based interface. The user can deplete their data set against TMC-SNPdb or create a customized normal variant database. The program has been tested to run on various Linux platforms such Fedora, Ubuntu and Red Hat operating systems. (Detailed user manual and snapshot of different steps have been provided in Appendix III. Using the companion tool on an 8GB machine, it takes 56 and 72 min to filter standard VCFs containing 115,884 and 227,779 raw variants (provided as example file with tool) against the TMC-SNPdb variants, respectively.

3.4.3 Application of "TMC-SNPdb" in depleting germline variants predominant among Indian population

With the flexibility of using GUI interface or through the command line (Appendix III), we tested the robustness and practical utility of "TMC-SNPdb" across various cancer types to infer the extent of depletion of population specific variants over and above the dbSNP. We analyzed 132 samples of three cancer types: head and neck cancer (n=43), cervical cancer (n=62) and gallbladder cancer (n=27). Significant fold reduction of variants was observed following TMC-SNPdb subtraction in addition to depletion by dbSNP in all cancer types studied. Of 613,055

variants found across 24 head and neck cancer tumor samples about 92% SNPs were depleted post dbSNP subtraction with 84,001 candidate somatic variants. Subsequent depletion using TMC-SNPdb identified 35,819 additional variants as Indian specific germline variants existing at varying frequency in normal Indian population. In overall, TMC-SNPdb allowed us to filter an additional 42.6% of post dbSNPs depleted SNPs. in 24 tongue cancer samples (Table 3.1). Similarly, TMC-SNPdb significantly reduced about 33.3% and 27.7% SNPs in 17 gallbladder and 34 cervical tumor whole exome data, as tabulated in detail in Table 3.1.

IX Table 3.1: Application of TMC-SNPdb across cancer types to filter germline variants in Indian population.

S.No.	Cancer type	Total variants	Number of samples	Number and percentage of novel variants		Overall reduction by TMC-SNPdb post dbSNP depletion
				Post dbSNP depletion	Post TMC-SNPdb depletion	
1	Tongue cancer	613,055	24	84,001 (13.7%)	48,182 (7.8%)	42.6%
2	Cervical cancer	923,547	34	99,032 (10.7%)	71,594 (7.7%)	27.7 %
3	Gallbladder cancer	328,245	17	26,530 (8%)	17,682 (5.3%)	33.3%

Total number of variants observed for each cancer types and reduction in number and percent variants post dbSNP and post TMC-SNPdb subtraction is tabulated for three cancer types. Number of samples analysed across tumor is also denoted.

3.5 Discussion

TMC-SNPdb is a freely available open access Indian population specific germline variant database consisting of 114,309 germline variants using whole exome sequencing of 62 normal tissues from patients with different types of cancer. Its usage is analogous to depletion against pooled normal variants from unrelated normal samples of Indian origin for paired or orphan tumor samples. The utility of subtraction against pooled normal variants has been well described as a reference for depletion, especially for orphan tumor samples wherein paired normal variant data for the tumor samples are not available [245-248]. Our dataset and companion tool can be used, along with other public databases, as 'normal' counterpart to identify disease specific somatic mutations, especially in cancer exome studies. Using TMC-SNPdb across 132 whole exome data of 3 tumor types, we show that it can significantly deplete false positive somatic variants.

TMC-SNPdb is presented with a companion program with command line or user-friendly GUI interface for non-computational biologists. It has two built-in features: first, a user can input tumor VCF to subtract against TMC-SNPdb and second, create a custom database of germ line mutation with the availability of multiple normal VCF files and then subtract with tumor VCF to deplete germ line variants. The subtraction program has been tested on several Linux platforms such as Fedora, Ubuntu and Red Hat system. Because it is an open source tool, it could be further modified to alter filtering parameters for analysis indicative of its expandability and universal applicability on Linux platforms.

There are two major limitations of TMC-SNPdb database. First, it is presumed that a sample derived from cancer patients represents 'normal' genome variation. However, because of their diseased status, a fraction of such individuals are likely to harbour cancer predisposing variants

in their germline. Any such germline variant that is novel in Indian population (not yet included in Caucasian databases) and which predisposes to cancer (for e.g. in *BRCA 1* gene) would be characterized as 'normal' population variation in TMC-SNPdb. Thus, this database will be limited in application to analyses that seek to evaluate germline predisposition to cancer. Second, majority of 'normal' samples were obtained from sites adjacent to a tumor with histopathological based inspection for the absence of tumor cells. However, it is possible that these tissues harbours some bona fide somatic mutations due to effect of field cancerization [249, 250]. Thus, depleting against TMC-SNPdb could potentially 'over-subtract' mutations that are bona fide somatic. To minimize this possibility, we have filtered TMC-SNPdb variants against COSMIC database to remove any known cancer related somatic variants. However, there remains a residual potential for missing 'somatic' mutations that are novel in tumors of Indian patients and present in adjacent 'normal' tissue. With these caveats, we believe that TMC-SNPdb with its companion tool is a step towards fulfilling a significant unmet need for an Indian population 'normal' variant database, especially in somatic mutation analyses in tumors from Indian patients.

In summary, TMC-SNPdb is an open source database of 'normal' germline variants derived from Indian-- non-European Caucasian-- population, not yet included in the public databases with predominant Caucasian representations. It comes along with a companion tool that can apply this information for somatic cancer genome analyses by depleting against the TMC-SNPdb. This database is flexible to accommodate the need for customization by allowing inclusion of similar datasets from additional individuals.

Chapter 4

Integrated analysis of tobacco/nut chewing HPV-negative early tongue cancer tumors identifies recurrent transcript fusions

(Detailed form of manuscript submitted to Oncotarget Journal, Under Review)

4. Chapter 4. Integrated analysis of tobacco/ nut chewing HPV-negative early tongue cancer tumors identifies recurrent transcript fusions

4.1 Abstract

Background: Of multiple anatomical sites represented in oral cancer, the tongue squamous cell carcinomas (TSCC) are distinct based on their aggressive behavior, poor prognosis, higher incidence among relatively younger age groups, and divergent traditional habits as etiological agents across populations. Despite advancement in detection and treatment modalities, the five year survival of tongue cancer patients remains abysmally low.

Methods: We performed high throughput sequencing of fifty four samples derived from HPV negative early stage tongue cancer patients habitual of chewing betel nuts, areca nuts, lime or tobacco using whole exome (n=47) and transcriptome (n=17) sequencing. Somatic mutation analysis, copy number alterations, differential expression and fusion transcript analysis were performed using in house developed bioinformatics pipelines.

Results: We report presence of a classical tobacco mutational signature C:G>A:T transversion in 53% TSCC patients of Indian origin, and describe a landscape of somatic alterations. These include previously described *TP53, NOTCH1, CDKN2A, HRAS, USP6, PIK3CA, CASP8, FAT1, APC,* and *JAK1* mutations. Based on the copy number analysis, we describe significant gains at genomic locus 11q13.3 (*CCND1, FGF19, ORAOV1, FADD*), 5p15.33 (*TERT*), 8q21.3 (*MMP16*), and 8q24.3 (*BOP1*); and, losses at 5q22.2 (*APC*), 6q25.3 (*GTF2H2*) and 5q13.2 (*SMN1*). Whole transcriptome analysis identified up-regulated gene-sets involved in EMT processes that may be crucial to predict nodal metastases in early tongue cancer patients. Furthermore, we identify 58 recurrent somatic fusion transcripts, including 5 novel fusion transcripts: *LRP5-UBE3C* (15%), *YIF1A-RCOR2* (13%), *POLA2-CDC42EP2* (8%), *SLC39A1-CRTC2* (4%), and *BACH1-GRIK1* (2%).

Conclusion: We present the first portrait of somatic alterations and mutational signature underlying the genome of tobacco/ nut chewing HPV-negative early tongue cancer patients of Indian origin.

4.2 Introduction

Tongue cancer is the most predominant form of oral cancer in developed countries with varying incidence in developing countries [16]. The major etiological factors associated with tongue cancer includes tobacco related products, alcohol and human papilloma virus (HPV) infections [251]. These factors lend to variability across populations, particularly in the Indian subcontinent wherein chewing betel-quid comprising betel leaf (Piper betel), areca nut (Areca catechu) and slaked lime (predominantly calcium hydroxide) is a part of the tradition [252]. While tobacco usage show a 5- to 25-fold increased risk of cancer [253], HPV infection defines clinical and molecularly distinct subgroups of head and neck squamous cell carcinoma (HNSCC) patients [254]. Such as, HPV-negative tumors are driven by amplification at 11q13, *EGFR* and *FGFR* loci*;* focal deletions at *NSD1, FAT1, NOTCH1, SMAD4* and *CDKN2A* loci; and, point mutations in *TP53, CDKN2A, FAT1, PIK3CA, NOTCH1, KMT2D,* and *NSD1* [135, 138]. On the other hand, HPV-positive tumors harbor *TRAF3*, *ATM* deletion, *E2F1* amplification, *FGFR2/3* and *KRAS* mutations. In addition, a specific targetable *FGFR3-TACC3* activated fusion oncogene aberration has been described to occur exclusively in HPV-positive HNSCC tumors [135, 255, 256].

Most of the genomic analysis studies involving TSCC have been restricted to advanced stage samples (pT3-pT4), while genomic alterations underlying HPV negative early tongue tumor genome has largely been unexplored. Furthermore, among early stage tongue squamous cell cancers, nodal metastases status plays a decisive role for choice of treatment [44]. About 70%

patients may be spared from surgery with accurate prediction of negative pathological lymph node status [44]. However, there is an unmet need for prognostic biomarkers to stratify the patients who could be spared unnecessary surgery lessening morbidity and cost of treatment.

In the present study, we present a portrait of somatic alterations in HPV negative early tongue cancer (pT1-pT2) using integrative genomic approaches to identify marker to stratify those likely to develop metastases.

4.3 Material and Methods

4.3.1 Ethical approval and informed consent

The sample set and study protocol were approved by (ACTREC-TMC) institutional Internal Review Board. All the tissue samples used under study have been taken after obtaining informed consent from patients. Samples were duly verified by two independent reviewers for histological examinations such as normal sample verification, percent tumor nuclei, and lymph node metastasis. Tumor sample with concordant histopathological diagnosis by both reviewers was included in the study. Tumor with at least >50% tumor nuclei was used for data analysis. Clinical, histological and etiological features was collected along with follow-up data for reoccurrence and alive status (Supplementary Table 1). All tumor were of early stages (pT1/pT2), 60% were from early age group (<40 years) and 60% were male. 80% of patients were exposed to some form of tobacco or alcohol. Clinico-histopathological examination during surgery confirmed the lymph node involvement in 65% of the cases analyzed. None of the samples showed the presence of HPV using HPVDetector and PCR-based validation using the MY09/11 method as described previously [150, 257].

X Table 4.1: The demographic and clinical characteristics of 54 tongue cancer patients in the study.

Clinicopathologic features	Variable	Frequency (N=57), N, % along the row
Age	Age, median (range)	45(25-72)
	>45 year	24(44%)
	<45 year	33(61%)
Gender	Male	42(74%)
	Female	15(26%)
Tumor stage	pT1-T2	57(100%)
	pT3-T4	0(0%)
AJCC TNM Stage	I-II	22(39%)
	III-IVA	34(61%)
Nodal Status	Node positive	34(61%)
	Node negative	22(39%)
Smoking	Smoker	16(27%)
	Non-smoker	35(61%)
	NA	6(12%)
Alcohol	Yes	12(21%)
	No	37(65%)
	NA	8(14%)
Tobacco	Yes	37(65%)
	No	20(35%)
Recurrences	Yes	15(26%)
	No	34(60%)
	NA	8(14%)
Metastasis	Yes	6(11%)
	No	43(75%)
	NA	6(11%)
Outcome	AWD or Died	15(26%)
	FOD	34(60%)
	NA	8(14%)

4.3.2 Extraction of DNA and RNA

Genomic DNA from tissues were extracted using DNeasy Blood and tissue DNA extraction kit (Qiagen) according to manufacturer's instructions. Quantification of DNA was assessed using Nanodrop 2000c Spectrophotometer (Thermo Fisher Scientific Inc.) and DNA integrity was determined by resolving on 0.8% Agarose gel. Later DNA was also quantified using Qubit ds DNA BR assay kit (Life Technologies, USA). DNA samples showing DNA concentration >50ng/µl and intact DNA at agarose gel were used for whole exome capture and sequencing studies.

For the extraction of total RNA from tissue samples using Trizol (Life Technologies) method as per manufacture's protocol with slight modification in homogenization process. In brief, 40-50mg tissue was homogenized in 1ml of trizol reagent using bead-based homogenizer using MP FastPrep-24 Instrument (MP Biomedicals Inc.) using lysing matrix tubes "D". Post homogenization homogenate was taken to a new eppendorf tube and standard Trizol method for RNA extraction was pursued. RNA was quantified using Nanodrop 2000c Spectrophotometer (Thermo Fisher Scientific Inc.) and RNA integrity determined by resolving on 1.2% agarose gel in TBE buffer. Later, total RNA was treated using DNase (DNA-free removal kit, Ambion, Life technologies) for removing possible DNA contamination. RNA samples having OD^{260}/OD^{280} ratio $\geq$ 1.8, OD^{260}/OD^{230} ratio $\geq$ 1.8 have been considered for further experiments. For whole transcriptome sequencing study RNA quality was further accessed Agilent bioanalyzer (Agilent RNA 6000 Nano Kit) and samples having RNA integrity number (RIN) $\geq$ 7.0 was used for mRNA capture and sequencing.

4.3.3 Exome capture and NGS DNA sequencing

Exome capture and sequencing were performed as described previously [257, 258]. Briefly, TruSeq Exome Enrichment kit (Illumina) and NimbleGen SeqCap EZ Exome Library v3.0 were used to capture ~62Mb region of human genome comprising of ~201,121 exons representing ~20,974 gene sequences, including 5'UTR, 3'UTR, microRNAs and other non-coding RNA. For the exome library preparation, 2µg genomic DNA was sheared using Covaris (Covaris Inc) for generating the fragment size of 200-300bp size. Libraries were prepared by following TruSeq Exome Enrichment Kit (Illumina) and NimbleGen SeqCap EZ Exome kit protocol. Seven picomol of 6-plex DNA library pool was loaded per lane on flow cell (Flow Cell v3) to generate clusters using TruSeq PE (Paired-End) Cluster Kit v3-cBot-HS kit and clustered flow was sequenced for 201 cycles on HiSeq-1500 System (Illumina) using TruSeq PE Cluster Kit v3 and TruSeq SBS Kit v3 (Illumina).

4.3.4 Transcriptome sequencing and data analysis to identify expressed genes

Transcriptome libraries for sequencing were constructed according to the TruSeq RNA library protocol (Illumina) outlined in TruSeq RNA Sample Prep (Illumina) performed as described previously [257]. Briefly, mRNA was purified from 4µg of intact total RNA using oligodT beads (TruSeq RNA Sample Preparation Kit, Illumina). Qualitative and quantitative analysis of cDNA libraries were performed using High Sensitivity DNA chip on 2100 Bioanalyzer (Agilent) and qPCR with KAPA Library Quant Kit (Kapa Biosystems) in ABI 7900HT system (Life Technologies), respectively. cDNA libraries were loaded on Illumina flow cell (version 3) to generate clusters using TruSeq PE (Paired-End) Cluster Kit v3-cBot-HS kit and clustered

flow was sequenced for 201 cycles on HiSeq-1500 System (Illumina) using TruSeq PE Cluster Kit v3 and TruSeq SBS Kit v3 (Illumina) to generate at least 30 million reads per sample.

Post sequencing, De-multiplexing was carried out on raw sequencing data was performed on basis of index sequences using CASAVA (version 1.8.4, Illumina). Transcriptome data analysis was performed using previously published a protocol for transcriptome sequencing data analysis [259]. To estimate the transcript expression abundance from transcriptome, RSEM software package was applied using hg19 as reference genome and counts were obtained for each tumor sample. First, to identify the bona fide expressed transcripts, we filtered all the transcripts which were lowly expressed (≤0.1 log10 (RSEM+1)) for each sample; second, transcript expressed in 10% of samples was considered as a candidate expressed gene in tongue tumor tissue. We identified 16,525 to be expressed in TSCC tumors (Appendix XII) and used to filter mutation and DNA copy number changes in this study. The transcriptome sequencing data generated in this chapter has been used for analysis in chapter 5 and 6.

4.3.5 Identification of somatic variants

The variant analysis was performed as described previously [257]. In brief, paired-end Raw sequence reads generated were mapped to the human reference genome (build hg19) using BWA v. 0.6.2 [230] and PCR duplicates were removed using Picard tools v.1.74 (http:broadinstitute.github.io/picard/). Base quality score recalibration and indel realignment were performed and variants were called from each sample separately using GATK (version 2.5-2) [232, 260] Unified Genotyper and MuTect (v1.0.27783)[189]. Variants called from GATK for each patient tumor were further filtered against its matched normal sample. Variants

obtained from MuTect and GATK were later merged and variants having ≥5 reads supporting altered base were considered for further analysis. To reduce the possibility of misclassification of germline variations as somatic mutation all the variants obtained from tumor were further filtered against Indian specific germline database (TMC-SNPdb) [258] to remove ethnic specific germline variants. Later, all the mutation were annotated using Oncotator (v1.1.6.0)[234]. Using in-house developed script, we compared and identified variants present in dbSNP (v142)[107] and COSMIC database (v68)[261]. Post filtration of variations present in the dbSNP database, only somatic COSMIC mutation, and novel mutations were obtained.

4.3.6 Identification of significantly mutated genes and deleterious mutations

MutSigCV v2.0 [158] and IntOgen [191] was used for identification of the significantly mutated gene in our cohort and p value ≤0.05 was considered for the threshold for significance. Since our dataset was inherently not suitable for above tools due to limited number of sample (n=26) so we have also performed extensive functional prediction tool based analysis for non-synonymous variants using nine different such as tools: dbNSFP v2.0 (includes SIFT, Polyphen2_HDIV, Polyphen2_HVAR, LRT, Mutation Taster, Mutation Accessor and FATHMM)[262], CanDRA (v1.0)[263], and PROVEAN [264][264][264][272][272][272][273][292][273][292][264]. Mutations which were predicted to be potentially deleterious by ≥4 prediction algorithm were further used for analysis. We also filtered mutation in those genes which are not expressed in our TSCC transcriptome data. Comparative analysis of genes harboring potentially deleterious mutations was performed with COSMIC Cancer gene census database[261], PanCancer gene list (significantly mutated genes across 12 cancer type)[132], and TCGA HNSCC mutation dataset using cBioPortal [135]. Total number of identified somatic substitutions in exome sequencing were extracted from

MutSigCV output and were processed to calculate the number and frequency distribution of various transitions and transversions.

4.3.7 Somatic copy number analysis from Exome sequencing data

DNA copy number analysis of exome sequencing data was performed as described previously [257]. In brief, normal and tumor BAM files were processed to get raw read counts for each target region defined in the bed file using Control-FREEC [265]. Conversion of counts to copy number ratio was done and segmentation was performed using lasso method. Annotation and post-processing were carried out using R programming and we also combined the segmentation results for 23 early TSCC patient samples. For identification of genes with statistically significant copy number changes, we used cghMCR (http://www.bioconductor.org/packages/2.3/bioc/html/cghMCR.html), an R implementation of a modified version of GISTIC analysis [266]. The cghMCR package usages a Segments-of-Gain-Or-Loss (SGOL) score to specify the frequency and degree of copy number changes in each gene or segment. Three standard deviations (SDs) from the mean SGOL score was selected because 97.17% of genes had SGOL scores within this range. So, a gene with SGOL score ≥ 4 defined as amplified genes and ≤ -2 as deleted genes.

4.3.8 Validation of somatic copy number changes

Validation of somatic copy number changes was performed as described previously [257]. All primers used have been tested for their specificity by performing evaluative PCR and resolving on agarose gel as well as melt curve analysis during quantitative real-time PCR. Ten nanograms of genomic DNA per 6 µl reaction volume in triplicates were amplified on Light cycler 480

(Roche, Mannheim, Germany) twice independently and relative copy number analysis was performed. The threshold for calling high and low copy number was ≥2 and ≤1.5, respectively and ≤1.99 and 1.49 ≤; diploid. Primers details used for copy number study has been provided in Table 4.2.

4.3.9 Differential gene expression analysis

To identify the differentially expressed genes between normal and tumors samples we followed previously described protocol of transcriptome sequencing data analysis [267]. Briefly, raw reads were aligned to human reference genome (hg19) using TOPHAT, and Cufflinks was used to perform the reference guided transcript assembly for each sample individually. To identify the differentially expressed genes between normal and tongue tumors we applied Cuffmerge and Cuffdiff method. We applied fold change up-regulation ≥2, downregulation ≤-2 and a p-value <0.05 to identify the significantly differentially expressed genes. Gene set enrichment analysis (GSEA) was performed using MutSigDB [268].

XI Table 4.2: Details of primer sequences used for validation of DNA copy number alterations.

S.No	Primer Name	5'----------->3'
1	OAD1069_CCND1_5'	GAACTACCTGGACCGCTTCC
	OAD1070_CCND1_3'	TAGAGGCCACGAACATGCAA
2	OAD1328_FGF19_5'	CGGAGGAAGACTGTGCTTTCG
	OAD1329_FGF19_3'	CTCGGATCGGTACACATTGTAG
3	OAD1314_ORAOV1_5'	CATGGGGAAGGGTATCGGG
	OAD1315_ORAOV1_3'	GCTCCATGCAGCGTGCCAT
4	OAD1316_SHANK2_5'	TCTCCTTCTCCCGCTCTCTC
	OAD1317_SHANK2_3'	CTGTTGCAGTATCGAGGGGG
5	OAD1318_FADD_5'	GGCTGACCTGGTACAAGAGG
	OAD1319_FADD_3'	GGACGCTTCGGAGGTAGATG
6	OAD1320_BOP1_5'	CTTCCTGTTGGTGGCGTCC
	OAD1321_BOP1_3'	CACTTGCAGTTGGGCATCAG
7	OAD_1027_MMP16_5'	AGCACTGGAAGACGGTTGG
	OAD_1028_MMP16_3'	CTCCGTTCCGCAGACTGTA
8	OAD1324_TERT_5'	CTACGGGGTGCTCCTCAAGA
	OAD1325_TERT_3'	TCTGTGTCCTCCTCCTCGG
9	OAD1071_MYC_5'	AGAGTTTCATCTGCGACCCG
	OAD1072_MYC_3'	AAGCCGCTCCACATACAGTC
10	OAD1306_CECR6_5'	ACCTGCTCGACAGCTTCAC
	OAD1307_CECR6_3'	AGACGGCGATAAGCAGGTAG
11	OAD1312_APC_5'	TCCATACAGGTCACGGGGAG
	OAD1313_APC_3'	TCTTCTTGACACAAAGACTGGCT
12	OAD506_GAPDH_5'	GAGGCTCCCACCTTTCTCATC
	OAD507_GAPDH_3'	ATTATGGGAAAGCCAGTCCCC

4.3.10 qRT-PCR validation of gene expression

Two microgram of total RNA extracted using Trizol method (Invitrogen) was subjected to cDNA synthesis using High capacity cDNA reverse transcription kit (Applied Biosystems) and quantitative PCR (qPCR) was performed using KAPA master mix (KAPA SYBR® FAST Universal qPCR kit). Quantitative PCR analysis was carried out by dilution of cDNA (1:10) in a 6µl volume in triplicate on Light cycler 480 (Roche, Mannheim, Germany) machine. The data analysis was performed using the 2–deltadeltaCt method where *GAPDH* was used as reference gene [204]. The primer sequences for genes are provided in Table 4.3 and specificity of primer was verified by performing semi-quantitative RT-PCR and melt curve analysis by dilution cDNA.

XII **Table 4.3: Primer sequences of genes for expression analysis using qRT-PCR analysis**

Primer ID	Primer sequence
OAD1035_CCNB1_5'	AATAAGGCGAAGATCAACATGGC
OAD1036_CCNB1_3'	TTTGTTACCAATGTCCCCAAGAG
OAD1037_SNAI2_5'	CGAACTGGACACACATACAGTG
OAD1038_SNAI2_3'	CTGAGGATCTCTGGTTGTGGT
OAD1033_CXCL13_5'	GCTTGAGGTGTAGATGTGTCC
OAD1034_CXCL13_3'	CCCACGGGGCAAGATTTGAA
OAD_1021_MMP14_5'	GGCTACAGCAATATGGCTACC
OAD_1022_MMP14_3'	GATGGCCGCTGAGAGTGAC
OAD_1023_MMP12_5'	GATCCAAAGGCCGTAATGTTCC
OAD_1024_MMP12_3'	TGAATGCCACGTATGTCATCAG
OAD_1031_MMP13_5'	ACTGAGAGGCTCCGAGAAATG
OAD_1032_MMP13_3'	GAACCCCGCATCTTGGCTT
OAD_521_GAPDH_5'	AATCCCATCACCATCTTCCA
OAD_522_GAPDH_3’	TGGACTCCACGACGTACTCA
OAD1131_ALDH1A1_5'	CCGTGGCGTACTATGGATGC
OAD1132_ALDH1A1_3'	GCAGCAGACGATCTCTTTCGAT
OAD1133_SOX2_5'	TACAGCATGTCCTACTCGCAG
OAD1134_SOX2_3'	GAGGAAGAGGTAACCACAGGG
OAD1135_NANOG_5'	CCCCAGCCTTTACTCTTCCTA
OAD1136_NANOG_3'	CCAGGTTGAATTGTTCCAGGTC
OAD1137_OCT4_5'	GGGAGATTGATAACTGGTGTGTT
OAD1138_OCT4_3'	GTGTATATCCCAGGGTGATCCTC
OAD1139_CD133_5'	AGAGCTTGCACCAACAAAGTACAC
OAD1140_CD133_3'	AAGCACAGAGGGTCATTGAGAGA
OAD1141_CD44_5'	CTGCCGCTTTGCAGGTGTA
OAD1142_CD44_3'	CATTGTGGGCAAGGTGCTATT

4.3.11 Transcript fusion detection

Chimerscan [269] was used to detect transcript fusion following the default parameters in the tumor, normal and cell lines. For fusion mapping, paired-end raw read

sequences were mapped to human reference genome sequences (hg19). Using in-house developed scripts in python we filtered fusions pairs without spanning read support, transcript allele fraction (TAF) <0.01 for both the partner, pseudogenes and homologous sequences spanning read, as described previously [270]. Fusion specific to tumors were further processed for annotation using Oncofuse [271] and frame of fusions was determined.

4.3.12 Validation of transcript fusions

Selected candidate fusions were validated using RT-PCR coupled with Sanger sequencing. Upstream and downstream sequences longer than spanning read sequences supporting fusions within exon were retrieved and primers were designed using Primer-BLAST[272]. Two micrograms total RNA from tumor and normal tissue was converted to cDNA using random hexamer primer and Reverse transcriptase kit (Thermo Scientific) as per manufactures instructions. Two microliter of 1:10 diluted cDNA were used for PCR amplification in 20 µl volume containing fusion specific forward and reverse primers and *GAPDH* primers (2pmol) using KAPA Taq PCR kit (KAPA Biosystems, KK1024) in an ABI verti thermal cycler under cycling conditions of 95^{o}C 3 minutes, 35 cycle of 94^{o}C 30 seconds, 55^{o}C 30 seconds, and 72^{o}C 45 seconds, and final extension at 72^{o}C 7 minutes. The PCR reactions were loaded on 1.2% agarose gel and expected band was excised from the gel and purified with NucleoSpin Gel and PCR product clean-up kit (MACHERET-NAGEL). If the fusions PCR product band was observed in the corresponding normal sample then it was not taken for further analysis. Sanger sequencing of purified products was performed using capillary electrophoresis 3500 Genetic Analyzer (Life Technologies) and sequencing traces were analyzed using BLAST [273]. The details of all the primers used for fusion transcript validation have been provided in Table 4.4.

***XIII* Table 4.4: Details of primer sequences used for validation of fusion transcripts.**

S.No.	Primer Name	Sequence (5'----------->3')
1	OAD418_EXT1-MED30_5'	CTTGGCCTGACTACACCGAG
	OAD419_EXT1-MED30_5'	TAGTTCCTCATTGCCAACATGG
2	OAD422_LRP5-UBE3C_5'	AAGCTCTACTGGGCTGACGC
	OAD423_LRP5-UBE3C_3'	AGCGGCACATTCAATCGCATA
3	OAD426_YIFA1-RCOR2_5'	AGTATTCGTACCCACGAGGCG
	OAD427_YIFA1-RCOR2_3'	GTACCTGGCGCTTGAGGGAG
4	OAD432_PSMD5-VAV2_5'	CCCCACTATGATTGGTGTAGC
	OAD433_PSMD5-VAV2_3'	CGGGACTTGTAGGGGTACTTG
5	OAD436_FTSJD2-BTBD9_5'	GAATGCACTGCCCCCATCAA
	OAD437_FTSJD2-BTBD9_3'	CCTGTCCCCGATTCCTCACT
6	OAD438_NAIP-GTF2H2B_5'	CTCTGTCACCAAGTGCTCCATA
	OAD439_NAIP-GTF2H2B_3'	TAAAGGTGGCGCATCATTCC
7	OAD1366_CLN6-CALML4_5'	AGGCATGGCTCTGTGAGCG
	OAD1367_CLN6-CALML4_3'	CCGTCCAGGAAGGGTGATCTT
8	OAD1368_RRM2-C2orf48_5'	GAGGGTTTACACTGTGATTTTGC
	OAD1369_RRM2-C2orf48_3'	AGACCTTCCCGAGCTTCTCAT
9	OAD1370_SLC39A1-CRTC2_5'	CAGAAAGCCCTGAGCCTAGT
	OAD1371_SLC39A1-CRTC2_3'	AGGGAGAGCTGTCAATGTGG
10	OAD1372_POLA2-CDC42EP2_5'	CGGAGTGATCTTCGGCTTGA
	OAD1373_POLA2-CDC42EP2_3'	CACGTTCTCCCCAAAGGACG
11	OAD1374_CTSC-RAB38_5'	ACTTTGTGAAAGCTATCAATGCCA
	OAD1375_CTSC-RAB38_3'	TCCACTTTGCCACTGCTTCA
12	OAD1376_BACH1-GRIK1-AS2_5'	TGAGACGGACACCGAAGGAG
	OAD1377_BACH1-GRIK1-AS2_3'	TCTCTCCCATCTTCTGGCTTC

4.3.13 Statistical analysis

The clinicopathologic association analysis was performed as described previously using IBM SPSS statistics software version 2. Test for overlap significance for a number of genes overlap for copy number changes and transcript fusions with different studies and databases were carried out using previously described method (http://nemates.org/MA/progs/representation.stats.html). The mutual exclusivity and co-occurrence analysis were performed using Gitools [274]. Data in the bar graph is represented

as mean ± standard deviation (SD) and plotted using Graph Pad prism version 5. The significant differences between selected two groups were estimated using unpaired Student t-test and P-value <0.05 was considered as a threshold for statistical significance.

4.3.14 Availability of supporting data

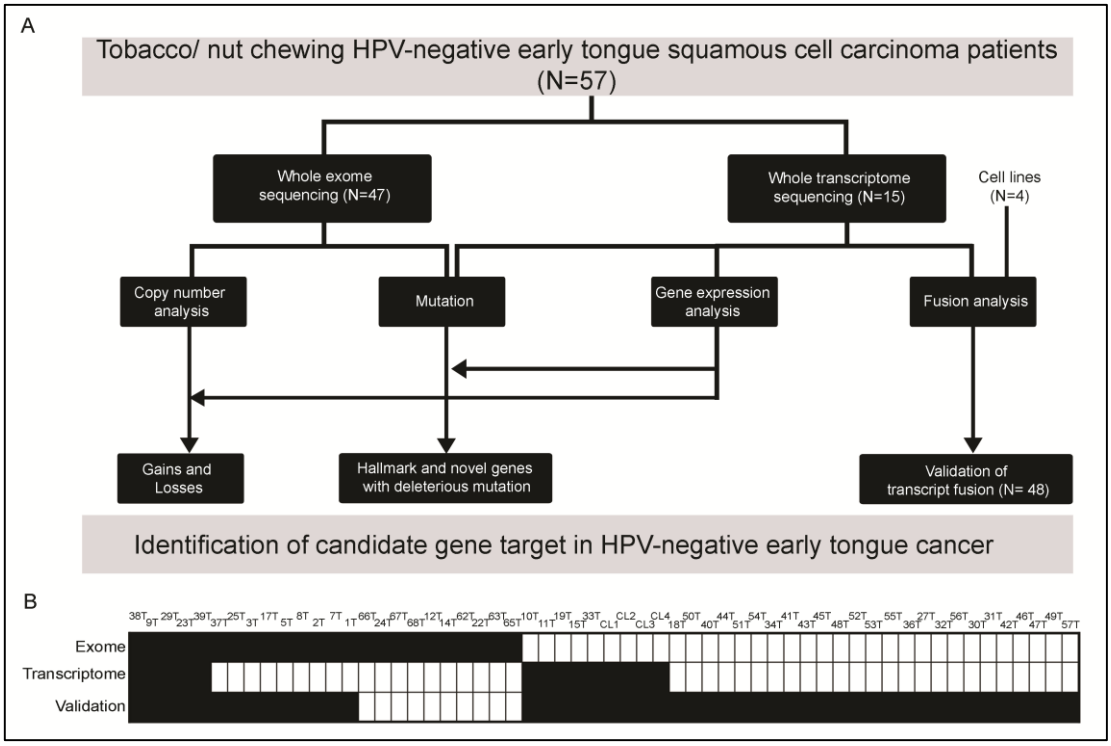
The raw sequence data has been deposited at the ArrayExpress (http://www.ebi.ac.uk/arrayexpress/), hosted by the European Bioinformatics Institute (EBI), under the following accession number: E-MTAB-3958: Whole transcriptome cell lines, E-MTAB-4654: Whole transcriptome tissue samples, E-MTAB-4653: Whole exome sequencing tumors tissue, E-MTAB-4618: Whole exome sequencing normal tissu

4.4 Result

4.4.1 Patient details

Fifty seven patient-matched normal early tongue cancer patient tumor were analyzed for somatic mutations, copy number changes, differential expression and transcript fusions by whole exome and transcriptome sequencing approach (Figure 4.1A,B). The clinicopathological details for fifty-seven the cohort is detailed in Table 4.1. In brief, our cohort comprised of 72% were male; 61% tobacco users; 80% chewers of either betel, tobacco, areca, or lime; and, 28% smokers, with a median age of 45 years (range 25-72 years). 56% of all patients with pT1 and pT2 staged tongue tumors were lymph node positive (n=32) at the time of registration. Primary treatment modality for all the patients was surgery followed by radiation and chemotherapy alone or in combination with chemo-radiation therapy. None of the patients were positive for HPV infection as reported previously [150]. Forty patient follow-up data was available and

median survival duration for the cohort was 29 months (range 2-42 months). During this time period, there were 9 recurrence, 6 metastasis, and 8 fatal events.



XVIII Figure 4.1: Schematic representation of Study overview of Tongue cancer integrated analysis.

(A) Integrated analysis involves somatic point mutations and DNA copy number changes and gene expression as well as transcript fusion identified from transcriptome sequencing. Somatic mutations and DNA copy number changes were filtered against gene expressed in tongue tumors to identify expressed alterations. (**B)** Sample distribution in different study such as exome, transcriptome sequencing, and validation of copy number changes and transcript fusions. The black and white filled boxes denotes patient sample included and excluded for respective analysis, respectively. In the exome sequencing track, asterisk * denotes tumors sample which were not included for variant analysis due to low coverage and/or poor correlation with their matched normal based on SNP profiler analysis. The asterisk ** denotes unpaired tumor sample.

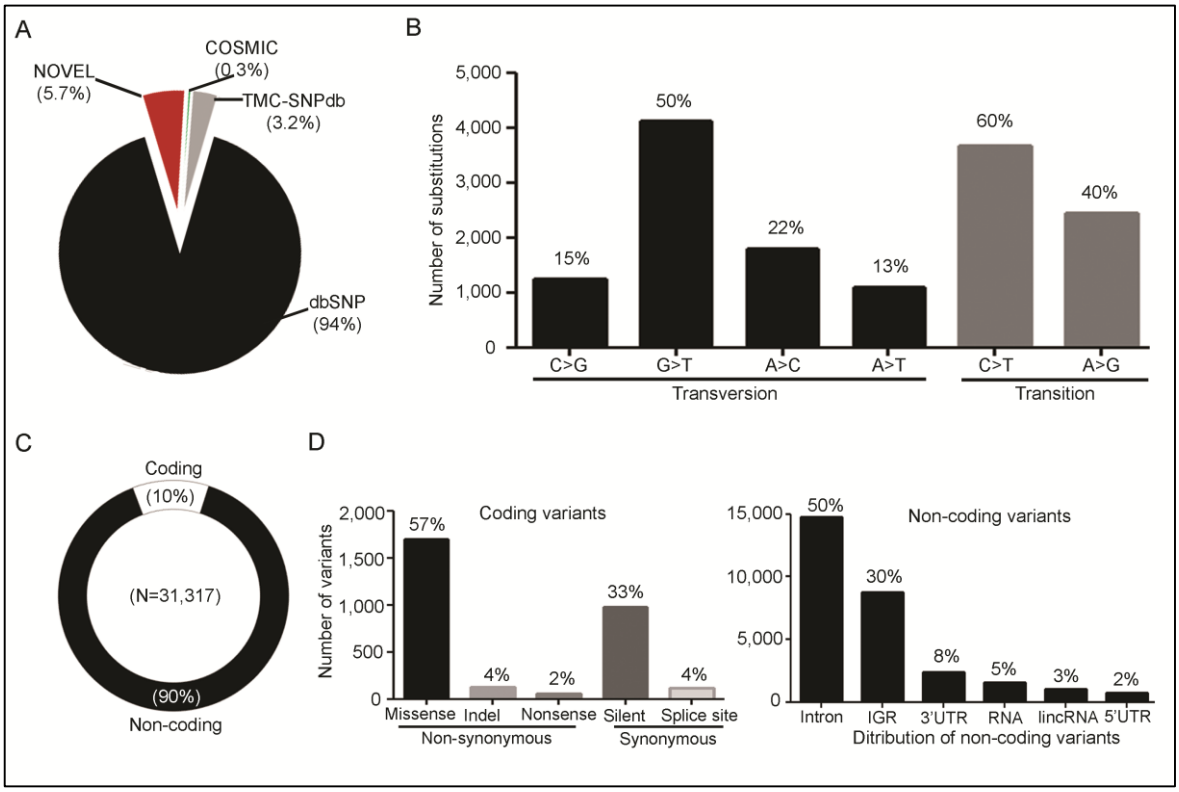
4.4.2 Somatic variants in HPV-negative early tongue squamous cell carcinoma

We performed whole exome sequencing of forty seven samples including early TSCC tumor (n=24) and matched normal (n=23) samples (n=47). We captured ~62Mb of coding genome at

a median coverage of 97x for tumor samples and 103x coverage for normal samples. Somatic variants were called using Mutect [189] and GATK algorithm [232]. We identified a total of 2,969 somatic mutations across 19 TSCC patients (5 patients were excluded from further analysis due to low coverage), which included 1,693 missesnse, 60 nonsense, 972 silent, 124 splice site as well as 120 indels. The aggregated non-silent mutation rate across the dataset was 4.12 per Mb, consistent with literature [125, 182]. The sequencing statistics and somatic mutational features are provided in Table 4.5, Appendix II and Figure 4.2A-E.

XIV Table 4.5: Descriptive statistics of various types and number of mutations identified for individuals.

S.No.	Sample ID	Total mutations	Total nonsilent	Coding mutations						Non-coding mutations							Mutation rate
				Missense	Nonsense	Nonstop	Silent	Splice	Indel	5'UTR	3'UTR	5'Flank	IGR	Intron	lincRNA	RNA	
1	12T	1583	118	98	1	0	65	11	8	74	141	0	408	646	49	82	3.9
2	14T	2105	148	127	3	0	55	11	7	98	193	0	619	801	84	107	4.9
3	63T	1764	120	106	2	0	34	3	9	38	87	0	516	847	43	79	4.0
4	22T	1880	116	99	3	0	52	6	8	42	138	0	415	1007	42	68	3.9
5	71T	662	73	59	6	0	17	3	5	13	61	0	116	334	8	40	2.4
6	23T	2121	112	99	2	0	50	5	6	45	135	0	493	1154	54	78	3.7
7	24T	1415	164	143	5	0	52	8	8	33	109	0	236	725	24	72	5.5
8	25T	2574	70	60	3	0	35	5	2	27	118	0	1106	973	133	112	2.3
9	68T	1437	106	86	8	0	33	6	6	31	94	0	343	728	34	68	3.5
10	29T	1088	72	63	0	0	30	6	3	21	91	0	268	540	24	42	2.4
11	2T	1561	230	200	5	0	114	15	10	29	136	0	500	398	68	86	7.7
12	62T	3062	282	249	3	0	137	13	17	50	216	0	708	1443	82	144	9.4
13	69T	849	94	74	4	0	25	10	6	14	60	0	190	412	15	39	3.1
14	37T	1592	99	86	4	0	43	2	7	32	98	0	405	791	39	85	3.3
15	38T	1694	108	91	4	0	35	7	6	31	102	0	371	933	44	70	3.6
16	39T	1370	93	82	0	1	36	5	5	26	74	0	407	634	34	66	3.1
17	7T	977	129	96	3	0	43	28	2	22	118	0	289	274	38	64	4.3
18	8T	1885	103	92	0	1	51	8	2	49	133	0	660	726	77	86	3.4
19	9T*	1045	99	83	4	0	44	5	7	36	125	0	278	348	46	69	3.3
20	66T	2010	141	123	4	1	64	9	4	47	131	0	457	1012	53	105	4.7
21	10T	5288	451	228	5	1	123	59	35	101	1099	186	1239	2022	206	400	7.6
22	11T	4956	395	184	6	1	110	57	37	85	1014	143	1044	1663	243	727	6.1
23	19T*	4221	488	255	6	0	154	38	35	73	901	107	795	1559	142	609	8.5
24	29T*	3450	543	259	4	1	141	55	83	47	1057	120	731	913	131	368	8.6
25	33T*	9630	669	342	12	0	204	96	15	116	2031	309	2455	3452	291	961	11.4
26	36T*	8969	710	382	11	1	197	100	19	106	1961	294	2480	2424	371	1314	12.7



XIX Figure 4.2: Various characteristic features of variants identified from whole exome sequencing data.

(A) Pie-chart representation percent frequency of dbSNP, COSMIC, TMC-SNPdb, and novel of variants identified in exome sequencing of early tongue tumors. (**B)** Distribution of number and frequency of various substitutions in exome sequencing variants. Bar graph representation of percentage frequency of transition and transversions in early tongue tumors. (**C)** Doughnut plots representation of somatic coding and non-coding variants. (**D)** Variant classification of somatic coding and non-coding variants. Bar plot representation of coding variants (left panel) and non-coding variants (right panel) percentage frequency distribution in various types of categories.

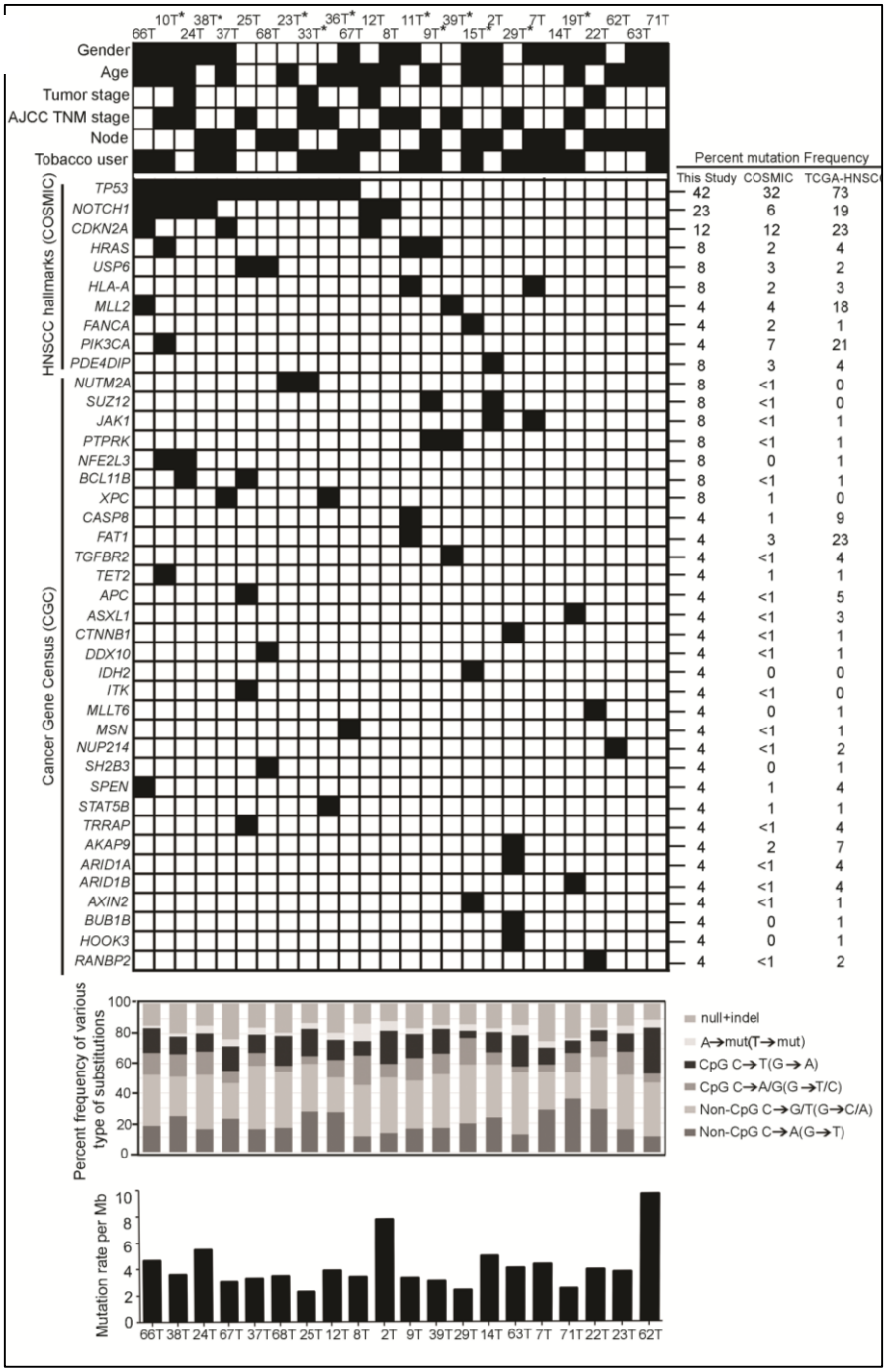
863 of 1,693 non-silent somatic variants across 767 genes were predicted to be deleterious (Appendixes V). Further posterior filtering of these variants were performed by restricting to 33 genes found to be significantly mutated using IntOgen (Q-value ≤ 0.05) as potential driver variants (Table 4.6) [191]. Among the HNSCC hallmark genes reported in COSMIC database, we observed recurrent mutations *in TP53* (42%), *NOTCH1* (20%), *CDKN2A* (12%), *HRAS*, (8%), *USP6* (8%); while, mutations in *FANCA, HLA-A, PIK3CA, KMT2D* and *PDE4DIP* were

observed as non-recurrent (Figure 4.3A). Overall the frequency of mutations observed in the hallmark genes were consistent with COSMIC and TCGA HNSCC data with altered frequency for *TP53* and *NOTCH1*, but consistent with reports from ICGC-India (Gingivo-buccal) [125], tongue from India and Asia [181, 182].

XV Table 4.6: List of significantly mutated genes observed in TSCC patients.

Hugo Symbol	Number of samples	Percent frequency	IntOgen (FM Q-value)
TP53	7	35	1E-11
ATP6V1A	6	30	7E-09
RASSF1	6	30	3E-06
P4HTM	6	30	5E-05
BZRAP1	6	30	5E-03
MCM8	5	25	3E-07
HECTD4	4	20	1E-13
NOTCH1	5	20	7E-05
IDUA	4	20	7E-05
TRIM39	4	20	4E-04
BRD8	4	20	1E-03
INTS2	4	20	2E-02
USP49	4	20	2E-02
LRRC37A3	4	20	4E-02
CCNL2	4	15	2E-04
SLC38A10	3	15	4E-04
FBXL14	3	15	4E-04
CDKN2A	3	15	2E-03
MST1	3	15	2E-03
SLC9A3R1	3	15	3E-03
PKD1	3	15	7E-03
CAMKK2	3	15	2E-02
PAPSS1	2	10	2E-03
CRIPAK	2	10	2E-02
PLEKHH1	2	10	2E-02
SRGAP2	2	10	2E-02
BMS1	2	10	3E-02
FKRP	2	10	3E-02
COQ9	2	10	4E-02

DOCK4	2	10	4E-02
RWDD1	2	10	4E-02
TTK	2	10	5E-02
XPC	2	10	5E-02



XX Figure 4.3: Identification of somatic mutations and DNA copy number changes in HPV-negative early TSCC.

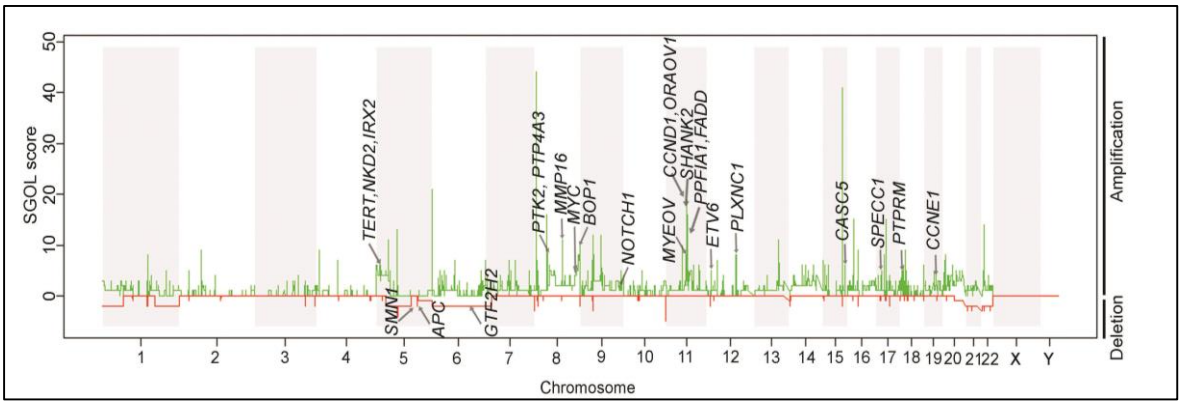
Mutational features of early tongue squamous carcinoma samples (n=25) using whole exome and transcriptome sequencing. Different clinicopathological factors such as; gender, age, tumor stage, AJCC TNM stage and lymph node metastasis status and etiological factors such as tobacco users are shown for each patient. The black solid boxes denotes gender: male, age: >45 years, tumor stage: pT1, AJCC-TNM stage: Stage I-II, Nodal status: positive and tobacco habit. The white boxes denote gender: female, age: <45 years, tumor stage: pT2, AJCC-TNM stage: Stage III-IV, Nodal status: negative and without tobacco habit. Grey filled boxes denotes no information available. Samples ID's with asterisk (*) denotes samples with transcriptome sequencing. The ten HNSCC hallmark genes and cancer gene census (COSMIC) found to be mutated in exome and transcriptome sequencing data, is arranged in decreasing order of percent frequency. Black filled box denotes presence of a somatic mutation in the patient. Mutation frequencies for the hallmark and cancer census genes observed in this study (n=25), COSMIC-HNSCC (n=>500) and TCGA-HNSCC (n=279) samples. The substitution frequencies spectrum for each patient for whole exome sequencing data is shown. Percent frequency of various types of SNVs and indels are shown. Different types of substitutions shown by different shades. Somatic non-silent mutation rate/30Mb derived from whole exome sequencing data for each tumor is shown.

The known mutational signature feature induced by tobacco is C:G > A:T transversion was found to be represented in high fraction (53%) (Fig 4.2B), which is much higher than observed in various cancers (15-26%) not associated with tobacco [275] and consistent with reports in HNSCC [125, 135]. Two patients which were hypermutated showed high proportion of C>T transition in CpG island, a known mutational signature due to overactivity of APOBEC family genes induced by deamination of 5-methyl-cytosine to uracil described previously [128, 181].

4.4.3 Somatic Copy number alterations derived from whole exome sequencing data

We used Control-FREEC [265] and cghMCR package to identify genomic regions harboring statistically significant copy number gains and losses relative to normal tissues. 440 amplified and 2275 deleted regions genes were identified across 23 TSCC patients. 18 genes exhibited copy number greater than three (Figure 4.4, Appendix VI, Appendix VII). Among most frequently amplified regions include 11q13.3 (*CCND1*, *FGF19*, *ORAOV1*, *FADD*); 5p15.33 (*SHANK2*, *MMP16*, *TERT*) and *BOP1* (8q24.3). We have also observed large fraction of TSCC

samples harboring amplification of key hallmark genes; *EGFR* (33%), *RICTOR* (33%), *PLEC* (33%), *TERT* (26%), *TNK2* (26%), *PIK3CA* and *SOX2* (22%), *MYC* (14%) and *NOTCH1* (14%). Comparative analysis of amplified and deleted gene with previous HNSCC including tongue cancer studies [125], TCGA-HNSCC [135], Vettore *et al*. [181]) and PanCancer [276] revealed statistically significant overlap in the number of genes (Table 5.7).



XXI Figure 4.4: Somatic copy number alterations in early TSCC patients.

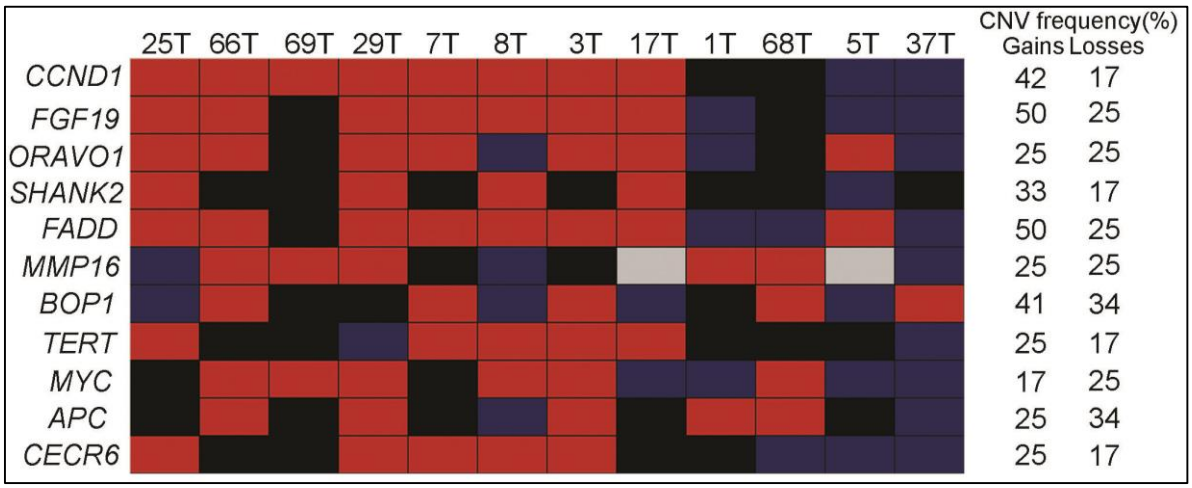
Overall view of Somatic DNA copy number changes identified using exome sequencing data. Somatic DNA copy number gains and losses were generated using Segments-of-Gain-Or-Loss (SGOL) scores across 23 TSCC patients. SGOL score is plotted (horizontal axis) for DNA copy number gains (green) and losses (red) are plotted as a function of distance along with human genome (vertical axis). Representative amplified and deleted regions are annotated for HNSCC-associated genes and denoted by an arrow.

XVI Table 4.7: Comparison of amplified and deleted gene overlap for this study with ICGC-India, TCGA-HNSCC and PanCancer study.

NA; not available. Asterisk (*); where only amplified genes were reported. Asterisk (**); Representation factor was less than one so not relevant.

Additionally, we validated copy number changes in hallmark genes using qPCR (Figure 4.5). We observed significant co-amplification of *CCND1*, *FGF19*, *ORAOV1*, *FADD* (P-value <0.01); *PIK3CA* and *SOX2* (P-value <0.001) in this study and TCGA-tongue tumors, which contains genes implicated in cell cycle, cell death/NF-kB pathway and, consistent with

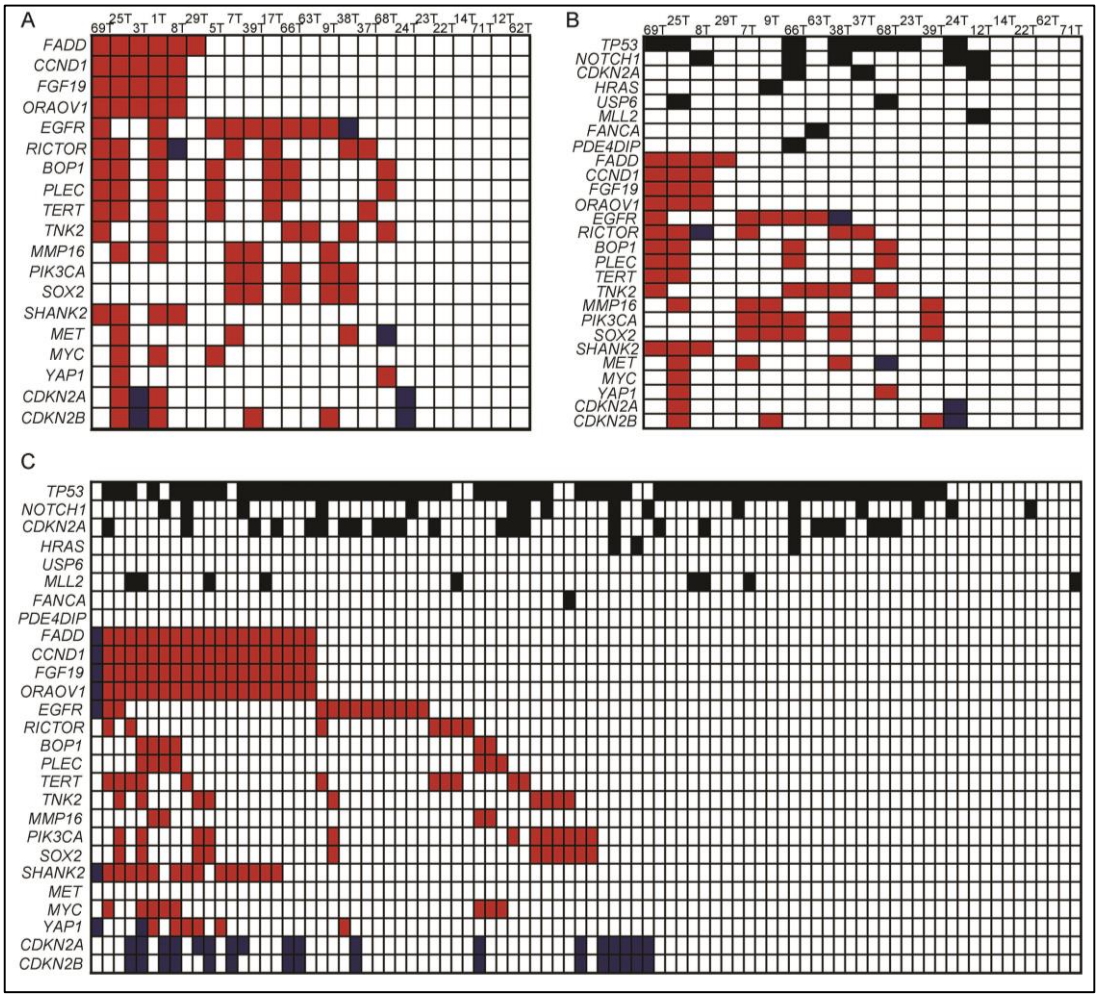
previously described in HPV-negative HNSCC tumors [135, 138] (Figure 4.6A). Interestingly, *EGFR* amplification was significantly mutually exclusive to *CCND1*, *FGF19*, *ORAOV1* and *FADD* amplification (P-value <0.01) including TCGA-tongue cohort (Figure 4.6A,C)[135], suggesting unique molecular features associated in our study cohort. These novel genetic associations could serve as pathognomonic alterations in HPV-negative early TSCC tumors.



XXII Figure 4.5: DNA copy number validation using qPCR.

The DNA copy number alterations were validated using qPCR. *GAPDH* gene was used as reference to normalize the data in qPCR.The solid red filled; amplification, blue; deletion and black; diploid and gray, could not be determined. Percentage frequencies for each gene are denoted.

	Amplified genes (n=440)				Deleted genes (n=2275)		
	ICGC-India	TCGA-HNSCC	Pan Cancer *	Vettore et al (2015)	ICGC-India	TCGA-HNSCC	Vettore et al (2015)
Total genes	2997	598	461	17	2146	4547	50
Number of gene overlapped	80	131	49	0	116	437	8
P-value	0.036	<0.0001	<0.0001	NA	<0.0001**	< 0.377**	< 0.203

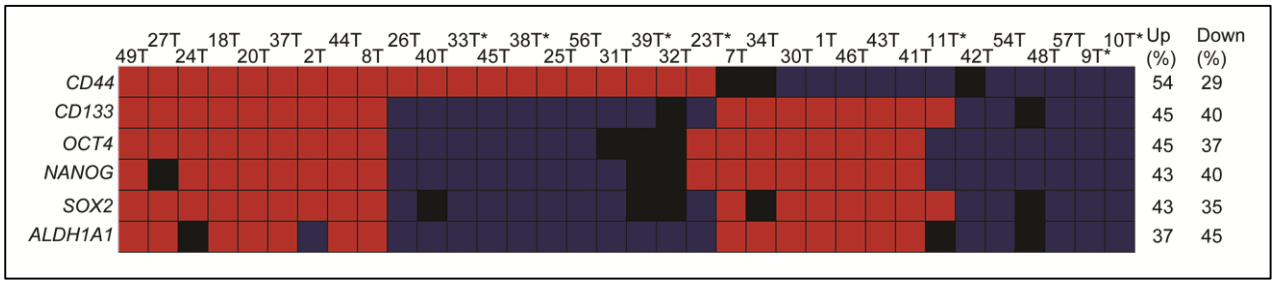


XXIII Figure 4.6: Genetic association analysis of hallmark genes copy number alterations and somatic variants in tongue tumors.
(A & B) Schematic representation of copy number alterations and mutation in hallmark genes in tongue tumors in this study (n=23). **(C)** Schematic representation of copy number alterations and mutation in hallmark genes in TCGA-TSCC cohort (n=79). The red filled box denotes amplification, blue; deletion and black; mutations.

4.4.4 Differentially expressed genes derived from whole transcriptome sequencing data

We performed whole transcriptome sequencing of five adjacent normal and ten tumour tissue samples to generate an average of 25 and 34 million paired end reads, respectively. Cufflinks [267], a transcript assembler, was used to perform reference guided assembly of the transcripts with an average expression of 11824 (SD±606) genes per sample ≥1 FPKM. The expression of these genes were validated by performing real time PCR for 6 genes across a set of 34

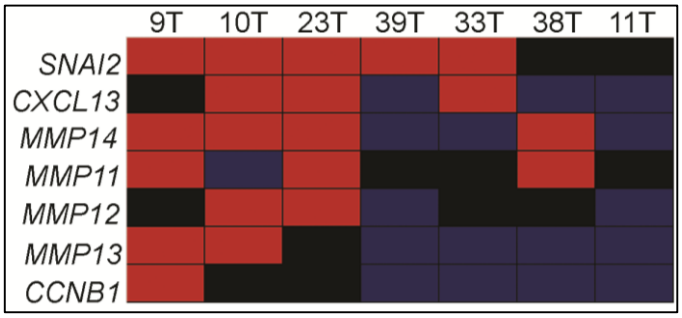
additional tongue cancer tumor samples and plotted against their expression in normal samples (Figure 4.7).



XXIV Figure 4.7: Gene expression verification of expressed genes in transcriptome using qRT-PCR.

Schematic representation of gene expression changes of candidate genes using qRT-PCR. The relative fold change of gene expression in each tumor sample was calculated by normalizing with respect to *GAPDH* and comparing with respective normal tissue sample. Asterisk (*) denoting tumor for which transcriptome sequencing was performed. The genes found to be expressed in transcriptome data were also identified in qRT-PCR analysis. The solid red filled; up-regulated (≥2 fold change), blue; down regulation (≤0.5 fold change), and black; basal (<1.99 - >1 fold change) and gray, could not be determined.

To identify the differentially expressed genes (DEGs) in whole transcriptome dataset, we used Cuffmerge and Cuffdiff method [267] and applied P-value ≤0.05 and log fold change 2 as a cut-off to identify 739 significantly differentially expressed genes (Appendix VIII). Of the 739 DEGs identified, 561 genes were up-regulated and 178 genes were down-regulated across the tumors. Gene set enrichment analysis (GESA) using MutSigDB [268] revealed an upregulation of gene-sets primarily involved in epithelial to mesenchymal transition (EMT) processes (Appendix IX). Additionally, gene sets involved in cardiac muscle contraction, calcium signaling pathway, and vascular smooth muscle contraction were found to be down regulated, consistent with earlier findings [105, 277-280]. The expression of 7 differentially expressed genes identified based on whole transcriptome analysis were validated by real time PCR in the same tumor samples (Figure 4.8).



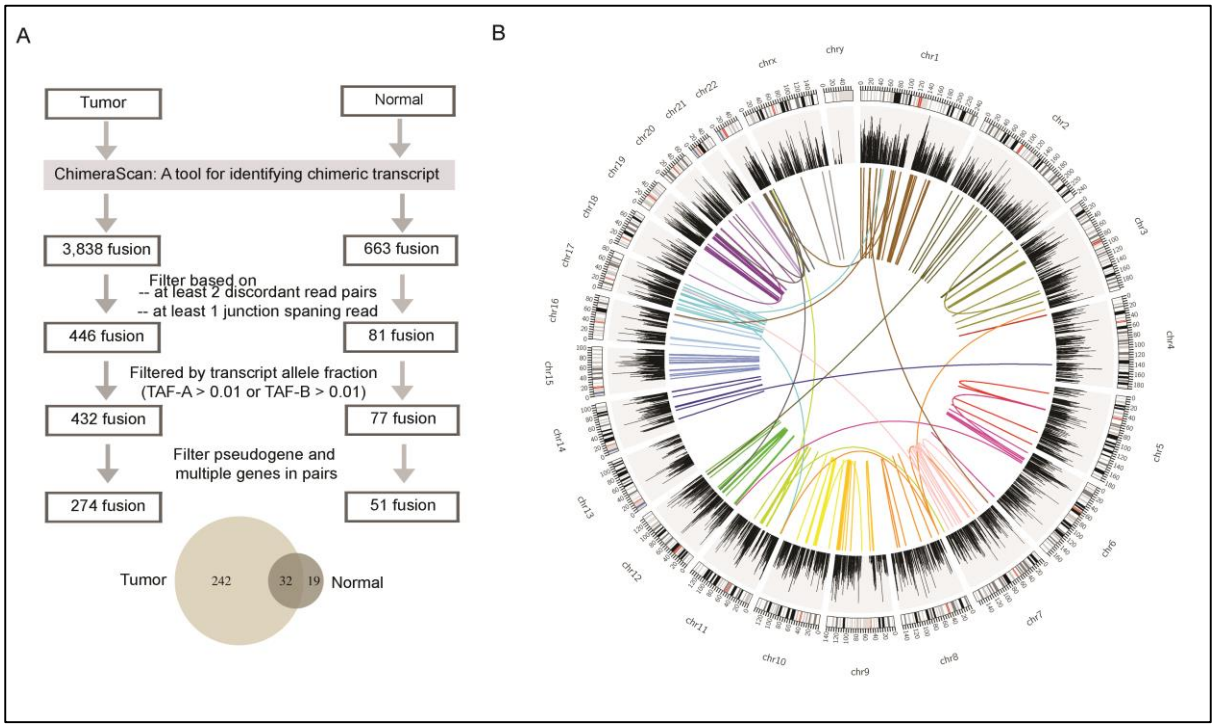
XXV Figure 4.8: Gene expression verification of significantly differentially expressed genes using qRT-PCR.

Schematic representation of gene expression changes in up-regulated genes. The relative fold changes in each sample was calculated by normalizing with reference to *GAPDH* expression and comparing with respective normal tissue sample. Transcriptome sequencing was carried out for these tumors samples which are denoted in figure. The solid red filled; up-regulated (≥2 fold change), blue; down regulation (≤ 0.5 fold change), and black; basal (<1.99 - >1 fold change) and gray, could not be determined.

4.4.5 Transcript fusion in HPV-negative early tongue squamous cell carcinoma

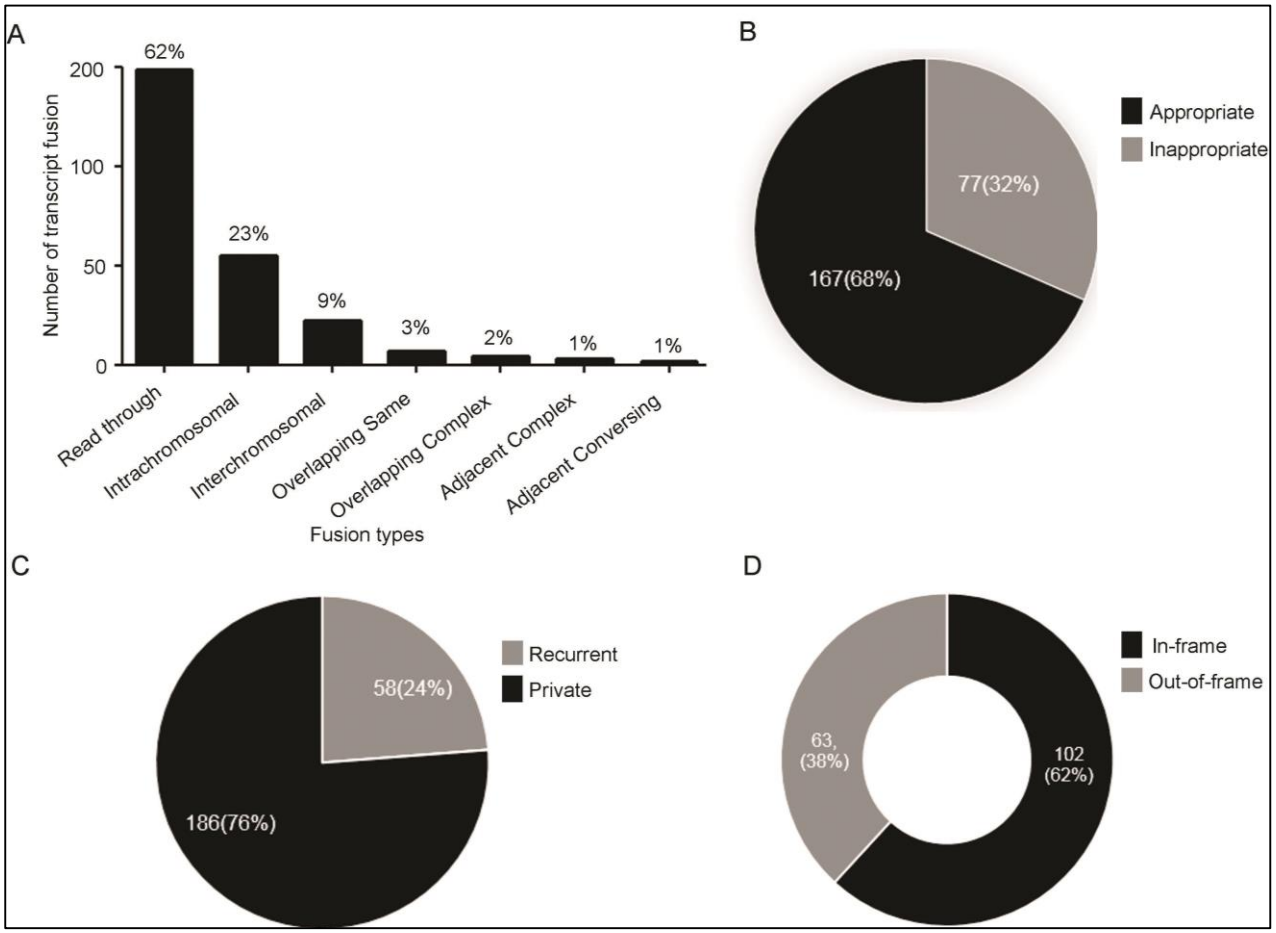
Next, we performed fusion transcript analysis of the whole transcriptome sequence data of five normal, ten tumors and four primary tumor derived cell lines of Indian ethnicity (n=19)[281] using Chimerascan [269] and observed a total of about 274 raw fusions per sample. Applying three rounds of filters: discordant read pairs with spanning reads; fusions with <0.01 transcript allele fraction (TAF) for both gene partners [270, 282]; and, depleting fusions occurring in normal tissue (n=32) [283] (Figure 4.9A), a total of 242 unique transcript fusion events affecting 416 genes were identified across the fourteen samples with 25±19 per sample, consistent with TCGA- HNSCC (Appendix X) [135]. Of 242 unique transcript fusion events, 9% (22/242) were inter-chromosomal fusion transcripts, while 91% (220/242) were intra-chromosomal fusion transcripts (Figure 4.10A); 68% transcript fusions had appropriate donor (5')-acceptor (3') sites (Figure 4.10B); About 24% of potential fusion transcript were expressed in two or more samples; while, 76% were identified in a single sample (Figure 4.10C); and,

62% of potential fusion transcripts were in-frame fusions (Figure 4.10D). Next, we compared our data with fusion databases; ChimeraDB [284], ChiTaRS [285], FusionCancer [286], TICdb [287], COSMIC[261], TCGA-Pan[270], TCGA-HNSCC[135], and recent study in HPV-positive HNSCC tumors by Guo *et al* [255]. We observed 48 unique transcript fusions overlap, showing identical fusion transcript pair (Table 4.8).



XXVI Figure 4.9: The landscape of transcript fusion identified in early tongue tumors and HNSCC cell lines.

(A) Flow chart representation of various steps and filters applied such as spanning read support, transcript allele fraction (TAF) and pseudogene and homology to identify and prioritize the putative high confidence transcript fusions in the study. (**B)** Circos plot representation of high-quality candidate fusions transcripts identified in tongue tumors. From outside to inside: karyotype, Gene expression (TPM) and transcript fusions. Black lines track for gene expression, fusion transcripts arc colored by their chromosome of origin.



XXVII Figure 4.10: Characteristics features of transcript fusions.

A) The bar graph representation of different types of fusion transcripts identified in early tongue tumors. Percentage frequency of each type of fusion is denoted at the top of bar graph. **B)** Pie-chart representation of number and percent frequency of appropriate (donor (5')-acceptor (3') relationships) and inappropriate (donor-donor or acceptor –acceptor) transcript fusions. **C)** Pie-chart representation of number and percent frequency of private and recurrent transcript fusions. **D)** Doughnut plot representation of number and percent frequency of reading frame of fusion transcripts. The in-frame (continuous reading frame for translation) and out-of-frame (reading frame with stop codon or truncation).

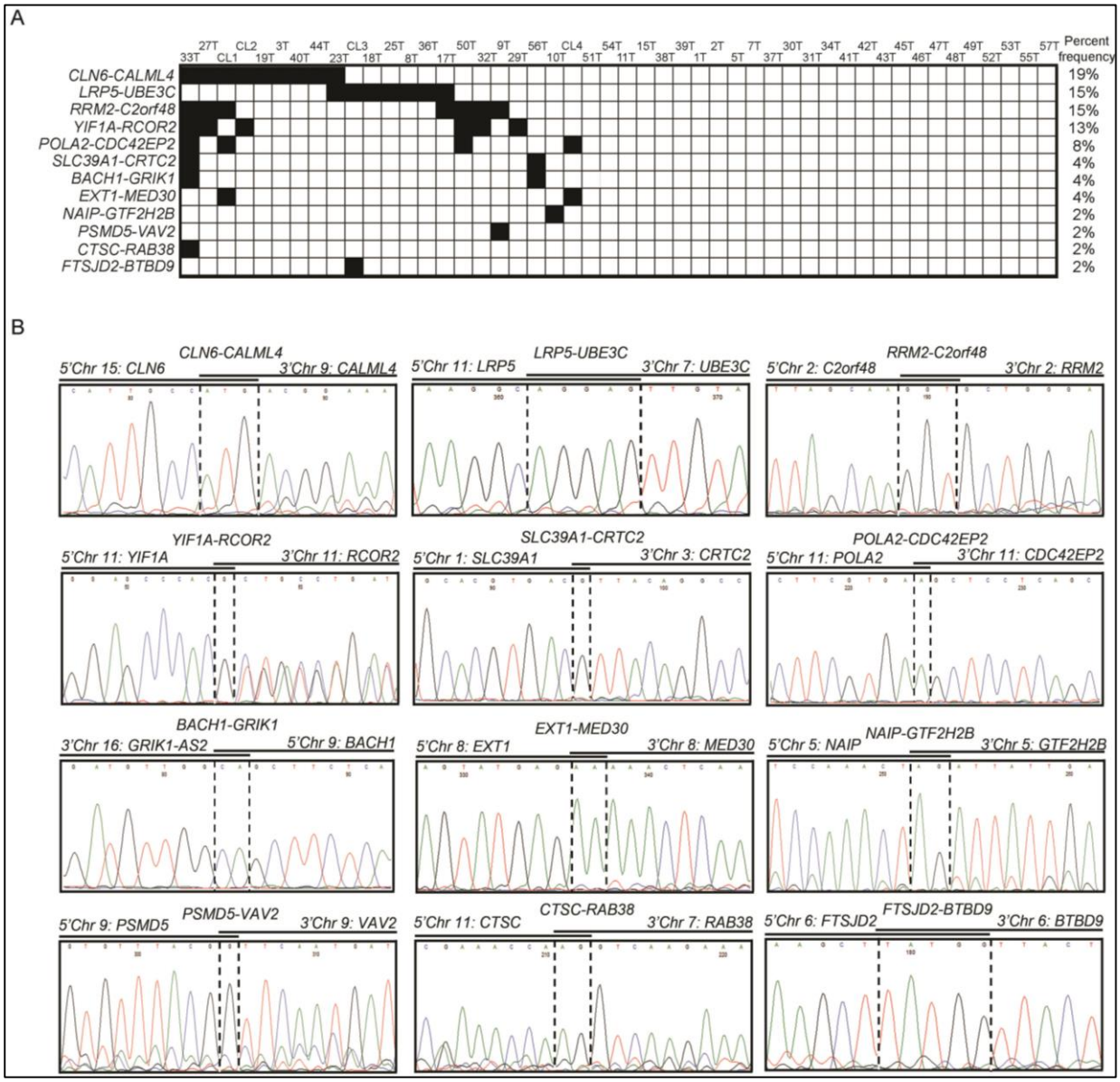
Among the 242 potential fusion transcripts, twelve high confidence fusion transcripts were selected based on following criteria: presence of appropriate donor (5') and acceptor (3') relationship; presence of reads spanning the junction region; and, those recurrent across multiple samples or expressed at very high level. Next, we performed Sanger sequencing for high confidence fusion transcript candidate using cDNA extracted from the samples to validate

twelve fusion transcripts (including 5 novel fusion transcripts) across 44 paired primary tumors and 4 primary tumor derived cell lines of Indian ethnicity [288] and identified recurrent fusion transcripts such as *CLN6-CALML4* (9/48), *LRP5-UBE3C* (7/48), *RRM2-C2orf48* (7/48), *YIF1A-RCOR2* (6/48), *POLA2-CDC42EP2* (4/48), *SLC39A1-CRTC2* (2/48), *BACH1-GRIK1* (2/48), *EXT1-MED30* (2/48), while fusions *NAIP-GTF2H2B, PSMD5-VAV2, CTSC-RAB38, FTSJD2-BTBD9* were observed as non-recurrent (Figure 4.11A). Interestingly, *LRP5-UBE3C* and *FTSJD2-BTBD9* fusion transcripts could be confirmed even at the genomic DNA level in NT8e and OT9 cell line, respectively, suggesting the origin of the fusion transcripts by genomic level rearrangements (Figure 4.11B).

XVII Table 4.8: Detail list of transcript fusions pair overlap with fusion databases

Fusion	FusionCancer	TCGA-PanCancer	TCGA-HNSCC	Number of samples
CLTC-VMP1				1
PMS2P5-CCDC146				3
PMS2P5-RASA4				1
SAV1-GYPE				1
SMA4-GTF2H2				1
BMS1-AQP7P1				1
BPTF-AMZ2				1
C15orf26-IL16				1
C17orf99-SYNGR2				1
CECR7-IL17RA				1
CIRBP-C19orf24				1
CLN6-CALML4				7
CNPY2-CS				1
CRIP2-CRIP1				1
CTSC-RAB38				2
CTSD-IFITM10				3
HERC2-HERC2P3				1
HOXB6-HOXB3				1
IGSF3-GGT1				3
IL17RC-CRELD1				2
KIAA0889-CCDC165				6
KIAA1267-ARL17A				7
LMAN2-MXD3				3
MAEA-CTBP1				1
MAPKAPK5-ALDH2				1
MBD1-CCDC11				7
NBPF24-NBPF15				1
NCKIPSD-CELSR3				1
NF1-AK4				1
NHP2L1-LLPH				1
NSUN4-FAAH				1
OSBPL8-TSPAN8				1
PLEKHO2-ANKDD1A				1
POLA2-CDC42EP2				5
PRIM1-NACA				3
RRM2-C2orf48				4
RRN3P3-CDR2				3
SAR1A-TYSND1				1
SCNN1A-TNFRSF1A				1
SIRPB2-NSFL1C				1
SLC35A3-HIAT1				1
TBC1D23-NIT2				2
TCRBV5S1A1T-TCRVB				1
TLK2-AL137655				5
TMEM141-KIAA1984				4
TPD52L2-DNAJC5				4
WRB-SH3BGR				1

Bold highlighted transcript fusions denotes, which were also validated by Sanger sequencing



XXVIII Figure 4.11: **Validation of transcript fusions across 92 tongue samples**

(A) Schematic representation of twelve validated transcript fusions by RT-PCR followed by Sanger sequencing confirmation in cohort 44 paired early TSCC tumors and 4 cancer cell lines (n=92). The black filled denotes positive by Sanger sequencing and white for no event. The percentage frequency is shown for each fusion transcript in the cohort. The sample name CL1, CL2, CL3, CL4 denotes AW13516, OT9, NT8e and AW8507 cell lines, respectively. (**B)** Representative Sanger sequencing chromatogram of twelve validated fusions using RT-PCR followed by Sanger sequencing. For each transcript fusion, chromosome number and gene name and the direction of gene involvement are shown as 5' and 3'. Sequence traces spanning the fusion breakpoint is shown and overlap of sequences between both the partner genes is denoted using the dotted black line.

4.4.6 Clinical correlation with hallmark genes mutated and transcript fusions in early tongue cancer

The follow-up data were available for 49 of 57 cases in the cohort did not reveal any significant association between clinical features such as age, gender, tumor stage, American Joint committee on Cancer (AJCC) TNM stage, nodal status, smoking, alcohol, tobacco usages with mutations in HNSCC hallmark gene; *TP53*, *NOTCH1*, *CDKN2A*, *CASP8*, *HRAS* and *PIK3CA* (Table 4.9). Of note, we observed 3 of 3 patients with *HRAS* mutation were tobacco chewers, which is consistent with previous reports in Indian oral cancer patients [289, 290]. Clinical correlation analysis with transcript fusions revealed significant correlation of *RRM2-C2orf48* and *YIF1A-RCOR2* transcript fusions tumor stage (P-value=0.006 and 0.05), where 3 of 4 and 2 of 2 pT1 tumor were positive, respectively (Appendix XI). In addition, *SLC39A1-CRTC2* fusions (3/3) were observed in stage I-II and node negative patients.

XVIII Table 4.9: Clinicopathologic correlation of clinical features with HNSCC hallmark gene mutation in early tongue cancer.

Clinicopathological features	Variable	Number (%), along the column	Hallmarks gene mutations (N=26) Number (%), along the row											
			TP53		P value	*NOTCH1*		P value	*HRAS*		P value	*CDKN2A*		P value
			Mutant	Wild type		Mutant	Wild type		Mutant	Wild type		Mutant	Wild type	
Age	>45	14 (54%)	6 (43%)	8 (57%)	1	5 (36%)	9 (64%)	0.17	2 (14%)	12 (86%)	1	3 (21%)	11 79%)	0.225
	<45	12 (46%)	5 (43%)	7 (58%)		1 (8)	11 (92%)		1 (8)	11 (92%)		0 (0%)	11 (100%)	
Gender	Male	16 (62%)	6 (38%)	10 (62%)	0.689	5 (31%)	11 (69%)	352	2 (13%)	12 (87%)	1	2 (13%)	14 (87%)	1
	Female	10 (38%)	5 (50%)	5 (50%)		1 (10%)	9 (90%)		1 (10%)	9 (90%)		1 (10%)	9 (90%)	
Tumor stage	p T1	4 (15%)	2 (50%)	2 (50%)	1	2 (50%)	2 (50%)	0.166	0 (0%)	4 (100%)	1	1 (25%)	3 (75%)	0.3
	p T2	21 (81%)	8 (38%)	13 (62%)		3 (14%)	18 (86%)		3 (14%)	18 (86%)		1 (5%)	20 (95%)	
	Information not available	1 (4%)												
AJCC TNM Stage	I-II	10 (38%)	5 (50%)	5 (50%)	0.442	3 (30%)	7 (70%)	0.358	2 (20%)	8 (80%)	0.543	0 (0%)	10 (100%)	0.5
	III-IVA	15 (58%)	5 (33%)	10 (67%)		2 (13%)	13 (87%)		1 (7%)	14 (93%)		2 (13%)	13 (87%)	
	Information not available	1 (4%)												
Nodal Status	Node positive	15 (58%)	5 (33%)	10 (67%)	0.442	2 (13%)	13 (87%)	0.358	1 (7%)	14 (93%)	543	2 (13%)	13 (87%)	0.5
	Node negative	10 (38%)	5 (50%)	5 (50%)		3 (30%)	7 (70%)		2 (20%)	8 (80%)		0 (0%)	10 (100%)	
	Information not available	14 (54%)	6 (43%)	8 (57%)	1	5 (36%)	9 (64%)	0.17	2 (14%)	12 (86%)	1	3 (21%)	11 79%)	0.225
Smoking	Smoker	5 (21%)	2 (40%)	3 (60%)	1	1 (20%)	4 (80%)	1	1 (20%)	4 (80%)	1	1 (20%)	4 (80%)	0.447
	Non-smoker	15 (63%)	6 (40%)	9 (60%)		4 (27%)	11 (73%)		2 (13%)	13 (87%)		1 (7%)	14 (93%)	
	Information not available	4 (17%)												
Alcohol	Yes	3 (13%)	1 (33%)	2 (67%)	1	1 (33%)	2 (67%)	1	1 (33%)	2 (67%)	0.404	0 (0%)	3 (100%)	1
	No	17 (71%)	7 (41%)	10 (59%)		4 (24%)	13 (76%)		2 (12%)	15 (88%)		2 (12%)	15 (88%)	
	Information not available	4 (17%)												
Tobacco	Yes	15 (58%)	7 (46%)	8 (54%)	0.701	3 (20%)	12 (80%)	1	3 (20%)	12 (80%)	0.238	2 (13%)	13 (87%)	1
	No	11 (42%)	4 (40%)	6 (60%)		3 (27%)	8 (72%)		0 (0%)	11 (100%)		1 (9%)	10 (91%)	

4.5 Discussion

A unique feature of TSCC from other subsites in oral cancer is that about 27-40% of patients even at an early stage (pT1 or pT2) have nodal metastasis and may be undergoing a neck dissection which further adds to morbidity and worse survival due to disease recurrence [44]. There is still an unmet need for reliable and robust prognostic biomarkers in early stage TSCC to stratify the patients who are likely to have an adverse clinical outcome [44]. Here, we describe the landscape of genomic alterations in a unique set of early staged HPV-negative tobacco or nut chewing tongue cancer patients, using whole exome sequencing and transcriptome sequencing. Lack of survival data however is a major limitation of the study that is currently underway. Since the cases were collected from 2010 to 2013, survival data of these

patients were far from maturity. Though several lines of distinct features underlie this study attributing to unique aetiology, subsite, and specific population, which have been previously described for HNSCC [138].

Firstly, the mutational profile of large fraction of patients display high frequency (53%) of C:G > A:T transversion in exome sequencing data—a hallmark of tobacco usage—reflecting tobacco as the most predominant etiological agent. We also observed enriched fraction of C>T transition and C>G transversion, consistent with previous report in gingiva-buccal (ICGC-India) and tongue tumors with tobacco chewing habit [125, 182]. The C>G transversions are known to be caused by tobacco due to reactive oxygen species (8-oxoguanine lesions) and or APOBEC family of cytidine deaminases genes overactivity induced by deamination of 5-methyl-cytosine to uracil in CpG island as described previously [181].

Secondly, recent reports suggest presence of low frequency (~5%) of RAS mutations in tongue tumors [181, 182]. However, we observed 12% *HRAS* mutation, which though were all tobacco chewers, consistent with previous reports from the Indian population[290]. Similarly, unlike previously reported, inactivating and low-frequency mutation in *NOTCH1* in HNSCC [80, 125, 127, 135], most of mutations were missense, consistent with recent report in Asian population and our report [257]. However, consistent with previous reports, frequent copy number alterations including gains at 5p, 8q, 20q, 22q and 11q and losses at 1p, 5p, 6q, 7p and 21q [125, 135, 181, 182] were significantly represented. Moreover, deleterious somatic variants in HNSCC hallmark genes: *TP53, NOTCH1, CDKN2A, CASP8, PIK3CA, USP6, MLL2, HLA-A, FANCA, PDE4DIP,* and *FAT1* were also identified [181, 182]. Furthermore, significantly co-occurring alterations in *FADD CCND1, FGF19,* and *ORAOV1* ($P<0.0001$) were found to occur mutually exclusive with *EGFR* amplification among HPV-negative early TSCC tumors, as previously described in other cancers [135, 138].

XIX Table 4.10: Detailed list of mutation frequencies across various previously published studies for the genes identified in this study.

Year	This study	2016	2015	2014	2013	2014	2016	2015	2011	2011
Study	Dutt lab	Krishnan *et al.*	Vettore *et al.*	Pickering *et al.*	Pickering *et al.*	ICGC-India	TCGA	Seiwert *et al.*	Stransky *et al.*	Agrawal *et al.*
Site	OT-26	OT-50	OT	34, OT-28	42, OT-30	GB	279, OT-70	All (HNSCC)	63, 38-OT	28, 25-OT
HPV (%)	0	44	NA	0	0	26	13	0	14	12
Number of samples	26	50	78	62	40	110	279	69	74	32
Country origin	India	India	Singapore	USA	USA	India	USA	USA	USA	USA
TP53	42	38	38	63	66 , OT-33	62	72, OT-39	81	76	79, OT-33
NOTCH1	23	4	5	25	25, OT-13	16	19, OT-3	16	13	14, OT-33
CDKN2A	12	6	5	5	0	2	22, OT-19	21	25, OT-16	OT-4
HRAS	12	0	1	7	OT-7	12	4, OT-4	0	OT-5	11, OT-12
USP6	8	8	3	2	2	2	2	0	0	0
HLA-A	8	2	0	14	OT-10	0	3, OT-1	0	OT-3	0
MLL2	8	2	10	5	OT-7	10	18, OT-7	18	OT-11	0
FANCA	4	0	5	2	2	2	1	0	0	0
PIK3CA	4	6	8	9	OT-10	6	21 , OT-11	13	OT-8	OT-8
PDE4DIP	4	0	8	5	4	0	4	0	6	0
JAK1	8	0	0	2	2	2	1	0	1.5	0
PTPRK	8	0	0	0	0	0	1	0	4	0
NFE2L3	8	0	0	0	0	2	1	0	0	0
BCL11B	8	0	0	0	0	0	1	0	1.5	0
CASP8	4	8	8	11	OT-10	36	9	0	OT-13	OT-4
FAT1	4	2	16	18	46, OT-20	44	23, OT-17	0	14, OT-13	0
TGFBR2	4	0	4	0	0	0	4	0	3	0
APC	4	0	9	0	2	0	5	0	6	0
CTNNB1	4	0	0	0	0	0	1	0	0	0

OT: Oral tongue, GB: Gingivo-buccal, OSCC: Oral squamous cell carcinoma

Thirdly, we identified and validated five novel somatic recurrent fusion transcripts: *LRP5-UBE3C*, *YIF1A-RCOR2*, *SLC39A1-CRTC2*, *BACH1-GRIK1*, and *EXT1-MED30* in HPV-negative early tongue cancer samples. Of these, *LRP5-UBE3C* fusion transcript that were found to be recurrent in 9 of 48 samples involve exon 1 of *LRP5* and exon 6 of *UBE3C*, affecting extra-cellular epidermal-growth factor-1 repeat (EGF-1) and HECT domain, respectively. In

parathyroid tumor and breast cancers, a truncated version of *LRP5* acts as a potential therapeutic target required for the active Wnt signalling for tumor growth [291]. Similarly, validated intra-chromosomal in-frame fusion *YIF1A-RCOR2* transcript connects *YIF1A* transcript to the sixth exon of REST corepressor 2 *RCOR2,* involving the SANT domain, known to form complex with Histone demethylase Lysine-specific demethylase *LSD1* and mediate hedgehog signalling pathway [292]. Interestingly, clinical analysis suggests that *RRM2-C2orf48*, *YIF1A-RCOR2* was significantly correlated with early stage (pT1) and *SLC39A1-CRTC2* exclusively occurred in node negative and early stage (pT1) tongue tumors. These fusions could potentially play a significant role to facilitate in diagnosis of early tongue tumors. Further characterization is warranted to elucidate the biological significance of these novel validated fusion transcripts. Moreover, consistent with previous reports, clinical correlation in the current study does not reveal association of hallmark mutated genes (*TP53, NOTCH1, HRAS, and CDKN2A*) with age or smoking habit younger [134, 293], however, there was also a trend of frequent mutation in *NOTCH1* and *CDKN2A* in older (>45 years) patients as compared to younger (<45 years) TSCC patients.

In conclusion, we describe the portrait of genomic and transcriptomic alterations using whole exome, transcriptome of HPV-negative early tongue cancer primary tumors to identified known and novel variants, copy number alterations, gene expression changes and transcript fusions that could play critical role potential driving role in early tongue tumors. We have identified known and novel genes with somatic mutations and copy number alterations along with recurrent fusion transcripts. However, insights about their specific role await validation followed by functional analysis of the alterations.

Chapter 5

Notch Pathway Activation is Essential for Maintenance of Stem-like Cells in Early Tongue Cancer

(as published in Oncotarget (2016),10.18632/oncotarget.10419)

5. Chapter 5. Notch Pathway Activation is Essential for Maintenance of Stem-like Cells in Early Tongue Cancer

(as published in *Oncotarget* (2016),10.18632/oncotarget.10419)

5.1 Abstract

Background: Notch pathway plays a complex role depending on cellular contexts: promotes stem cell maintenance or induces terminal differentiation in potential cancer-initiating cells; acts as an oncogene in lymphocytes and mammary tissue or plays a growth-suppressive role in leukemia, liver, skin, and head and neck cancer. Here, we present a novel clinical and functional significance of *NOTCH1* alterations in early stage tongue squamous cell carcinoma (TSCC).

Material and Methods: We analyzed the Notch signaling pathway in 68 early stage TSCC primary tumor samples by whole exome and transcriptome sequencing, real-time PCR based copy number, expression, immuno-histochemical, followed by cell based biochemical and functional assays.

Results: We show, unlike TCGA HNSCC data set, *NOTCH1* harbors significantly lower frequency of inactivating mutations (4%); is somatically amplified; and, overexpressed in 31% and 37% of early stage TSCC patients, respectively. HNSCC cell lines over expressing *NOTCH1*, when plated in the absence of attachment, are enriched in stem cell markers and form spheroids. Furthermore, we show that inhibition of NOTCH activation by gamma secretase inhibitor or shRNA mediated knockdown of *NOTCH1* inhibits spheroid forming capacity, transformation, survival and migration of the HNSCC cells suggesting an oncogenic role of *NOTCH1* in TSCC. Clinically, Notch pathway activation is higher in tumors of non-

smokers compared to smokers (50% Vs 18%, respectively, *P*=0.026) and is also associated with greater nodal positivity compared to its non-activation (93% Vs 64%, respectively, *P*=0.029).

Conclusion: We anticipate that these results could form the basis for therapeutic targeting of NOTCH1 in tongue cancer.

5.2 Introduction

Head and neck squamous cell carcinomas (HNSCC) are the sixth most frequent malignancy worldwide with more than 550,000 cases and around 300,000 deaths each year [1, 294]. Recent large-scale genome wide studies have underscored a complex role of *NOTCH1* as a candidate tumor suppressor harboring inactivating mutation and deletions [80, 127, 295], as well as a candidate driver of tumorigenesis harboring activating missense mutations and amplifications [133, 296-298] in a context dependent manner in head and neck and other cancers [222, 299]. In addition, Notch signaling pathway is known to play a significant role in maintaining cancer stem-like population of cells (CSCs) in several human cancers [295, 300-302]. Inhibition of Notch signaling has been shown to prevent the formation of secondary mammospheres from cell lines and primary patient samples [303]. However, pathogenesis and biological significance of CSCs in HNSCC has not been well characterized. Targeted elimination of these cells may provide a new insight in the etiology and treatment of head and neck cancer. Interestingly, cancer stem cells share features with their role in the development of tumorigenesis [304-306]. However, the biological significance of cancer stem-like cells (CSCs) in HNSCC has not been well characterized.

To understand the role of Notch signaling pathway in early-stage (pT1-pT2) tongue tumors, we examined the mutational landscape, copy number alterations and differential expression of

receptor, ligands, modifiers and target genes of the Notch pathway, along with effect of genetic and pharmacologic perturbation of Notch pathway on cancer stem-like cells (CSCs) features of HNSCC cells.

5.3 Materials & Methods

5.3.1 Patient Samples

Tumor-normal paired samples were collected at Tata Memorial Hospital and Advanced Centre for Treatment, Research and Education in Cancer (ACTREC), Mumbai. Sample set and study protocols were approved by (ACTREC-TMC) Internal Review Board (IRB) and most of the patients were recruited from 2010-2013 with a predefined inclusion criteria of early (pT1 and pT2) stage. Percent tumor content was determined using hematoxylin and eosin based staining by two independent pathologists which varied from 60 to 90%. Patient samples and characteristics are provided in the Table 5.1.

XX Table 5.1: The demographic and clinical characteristics of 68 tongue tumor samples

Variable	Frequency (N=68)
Age, median (range)	42(23-76)
Gender	
Male	48(71%)
Female	20(29%)
Sub-site	
Oral tongue	100%
Tumor Stage	
pT1	9(14%)
pT2	55(86%)
Nodal Stage	
Node Negative	26(41%)
Node Positive	37(58%)
Habit	
Yes	55(81%)
No	13(19%)
Smoking	
Non-Smoker	47(69%)
Smoker	16(22%)
Alcohol	
Yes	43(74%)
No	15(26%)

5.3.2 DNA and RNA extraction

DNA from tongue primary paired normal-tumor tissue samples were extracted using DNeasy Blood and tissue DNA extraction kit (Qiagen) according to manufacturer's instructions. DNA was quantified using Nanodrop 2000c Spectrophotometer (Thermo Fisher Scientific Inc.) and DNA quality was checked by resolving on 0.8% agarose gel. Total RNA was extracted from tongue primary paired normal-tumor samples and cell lines using RNeasy RNA isolation kit (Qiagen) and Trizol reagent (Invitrogen) based methods and later resolved on 1.2% Agarose gel to confirm the RNA integrity.

5.3.3 Exome capture and NGS DNA sequencing

Two different Exome capture kits were used to capture exome for different samples. The TruSeq Exome Enrichment kit (Illumina) was used to capture 62Mb region of human genome comprising of 201,121 exons representing 20,974 gene sequences, including 5'UTR, 3'UTR, microRNAs and other non-coding RNA and NimbleGen SeqCap EZ Exome Library v3.0 was also used to capture 64 Mb region of the human genome. Exome library preparation and sequencing was performed as per manufacturer's instructions. Briefly, 2 µg genomic DNA was sheared using Covaris (Covaris Inc) for generating the fragment size of 200-300bp size. DNA libraries were prepared from both the kits were quantified by qPCR using KAPA Library Quant Kit (Kapa Biosystems) in ABI 7,900HT system (Life Technologies). Seven pmol of 6-plex DNA library pool was loaded per lane on flow cell (Flow Cell v3) to generate clusters using TruSeq PE (Paired-End) Cluster Kit v3-cBot-HS kit and clustered flow was sequenced for 201 cycles on HiSeq-1,500 System (Illumina) using TruSeq SBS Kit v3 (Illumina) at in-house core NGS facility.

5.3.4 Identification of Somatic Mutations from Exome Sequencing

Paired-end raw sequence reads generated were mapped to the human reference genome (build hg19) using BWA v. 0.6.2 [230]. Mapped reads were then used to identify and remove PCR duplicates using Picard tools v1.100 (http:broadinstitute.github.io/picard/). Base quality score recalibration and indel re-alignment performed and variants were called from each sample separately using GATK 2.5-2 Unified Genotyper and MuTect v. 1.0.27783 [189]. Post subtraction of variants from its paired normal, remaining variants was taken for further analysis if they were having ≥5 altered reads. Furthermore, all samples variants were further filtered against pooled normal variants database (n=62) to reduce the possibility of the germline variation We further annotated variants using Oncotator v1.1.6.0 [234] and dbSNP v142 [107]and COSMIC database v68 [261] using an in-house developed script. Later, we performed functional prediction tool based analysis for somatic non-synonymous variants using nine different tools such as: dbNSFP v2.0 (includes SIFT, Polyphen2_HDIV, Polyphen2_HVAR, LRT, Mutation Taster, Mutation Accessor and FATHMM) [307], CanDRA v1.0 [263] and Provean v.1.1 [214]. Variants called deleterious in nature by at least one software was taken for further analysis. We confirmed the identity of mutations by manual visualization in IGV [308, 309].

5.3.5 Somatic Copy number analysis from Exome sequencing data

BAM files prepared for variant calling were used for copy number analysis using Control-FREEC [265]. Paired tumor-normal samples BAM files were fed into Control-FREEC along with target region for Illumina and Nimblegen exome kits as bed file. Read count were generated and normalized for GC content for each of the target region followed by computation of ratio of read count in a tumor to normal. Read count ratio was converted to copy numbers

followed by segmentation using lasso method. Segmented copy number data generated by control-FREEC was further used for annotation and post-processing using R programming.

5.3.6 Transcriptome sequencing and data analysis

Transcriptome libraries for sequencing were constructed according to the TruSeq RNA library protocol (Illumina). Briefly, mRNA was purified from 4 µg of intact total RNA using oligodT beads and library preparation was done as per manufacturer's instructions (TruSeq RNA Sample Preparation Kit, Illumina). 7pmol of quantified cDNA libraries were loaded on Illumina flow cell (v3) to generate clusters using TruSeq PE (Paired-End) Cluster Kit v3-cBot-HS kit and clustered flow was sequenced for 201 cycles on HiSeq-1500 System (Illumina) using TruSeq PE Cluster Kit v3 and TruSeq SBS Kit v3 (Illumina) to generate at least 30 million reads per sample. Post sequencing, de-multiplexing was carried out on the basis of index sequences using CASAVA (version 1.8.4, Illumina). Transcriptome data analysis was performed using Tuxedo-suite pipeline [267]. In brief, alignment of short reads was done against reference genome (hg19) using TOPHAT2 v. 2.0.8b [310] in which 95-99% of reads were mapped to the reference genome. Cufflinks v.2.0.2 was used to find the expressed transcripts in the data and quality control steps was performed using CummeRbund package v2.0. All the actively expressed transcripts per samples were then binned by $\log_{10}$ (FPKM+1) to differentiate the significantly expressed transcripts from the background noise and transcripts represented by <0.1 $\log_{10}$ (FPKM+1) were filtered out from further analysis. Since paired normal of these tumors cannot be obtained, we defined a significant change in expression for those genes whose expression is higher (>80%) or lower (<20%) than the median expression as suggested [311].

5.3.7 Analysis of The Cancer Genome Atlas Tongue cancer data

The Cancer Genome Atlas (TCGA) dataset of HNSCC including DNA copy number dataset (gistic2 threshold) from 452 HNSCC tumor tissue, RNA seq expression (Illumina HiSeq) dataset from 541 HNSCC was downloaded from UCSC Cancer genome browser on 20th June 2014. Later, tongue cancer patient data for DNA copy number (n=126) and gene expression (n=129 has been taken for further analysis. For expression and DNA copy number the median centered RSEM counts and gene-level copy number estimates have been used, respectively (n=126). Notch pathway genes (n=13) data has been retrieved and heatmaps were generated using MeV 4.9.0. The RNAseq gene expression data has been retrieved for Notch pathway genes (n=13) and raw data has been median centered using Cluster 3.0 software. The median centered RSEM counts for each gene has been used to generate heatmap using MeV 4.9.0. Fold change criteria was ≥1.5 fold change for upregulation, ≤1.499 fold change for down-regulation and in between -1.5 to 1.499 fold change was denoted as a basal expression. DNA copy number and Expression correlation analysis and clinical correlation analysis have been performed using SPSS. P=value <0.05 was criteria for statistical significance.

5.3.8 Tissue processing

Surgically resected oral tumor tissues and matched nonmalignant (cut margins) adjacent tissues were obtained from patients with informed consent after IRB approval from ACTREC. These tissues were processed for paraffin embedding and sectioned at 4µm for H/E staining for evaluation of tumor.

5.3.9 Immunohistochemistry

Immunohistochemistry was done following the standard protocol of DAKO Envision Flex. Briefly, the slides were microwaved by incubating them for 10 minutes in high pH antigen Retrieval Solution (DAKO;DM828), then allowed to cool to room temperature before rinsing with Tris-buffered saline wash buffer (DAKO;DM831). Endogenous peroxidase activity was blocked by incubating the slides for 20 minutes in 3% hydrogen peroxide (EnVision/HRP, Dako). After rinsing in wash buffer, the sections were incubated for 3hours at room temperature with the monoclonal human anti-activated Notch1 antibody (Cat.ab8925; dilution 1:50) in Tris-HCl buffer antibody diluent (Dako; K8016). Slides were rinsed in wash buffer (DAKO; DM831) and incubated for 90 minutes with peroxidase-labeled polymer conjugated to goat anti-rabbit immunoglobulins (EnVision/HRP, Dako; SM801). The chromogenic reaction was carried out with 3,3′-diaminobenzidine chromogen solution for 5 minutes, resulting in the expected brown-colored signal. Finally, after rinsing with deionized water, the slides were counterstained with hematoxylin, dehydrated, mounted with toluene-based mounting medium (Thermo Scientific Richard-Allan) and cover slip.

5.3.10 Immunohistochemical staining Analysis

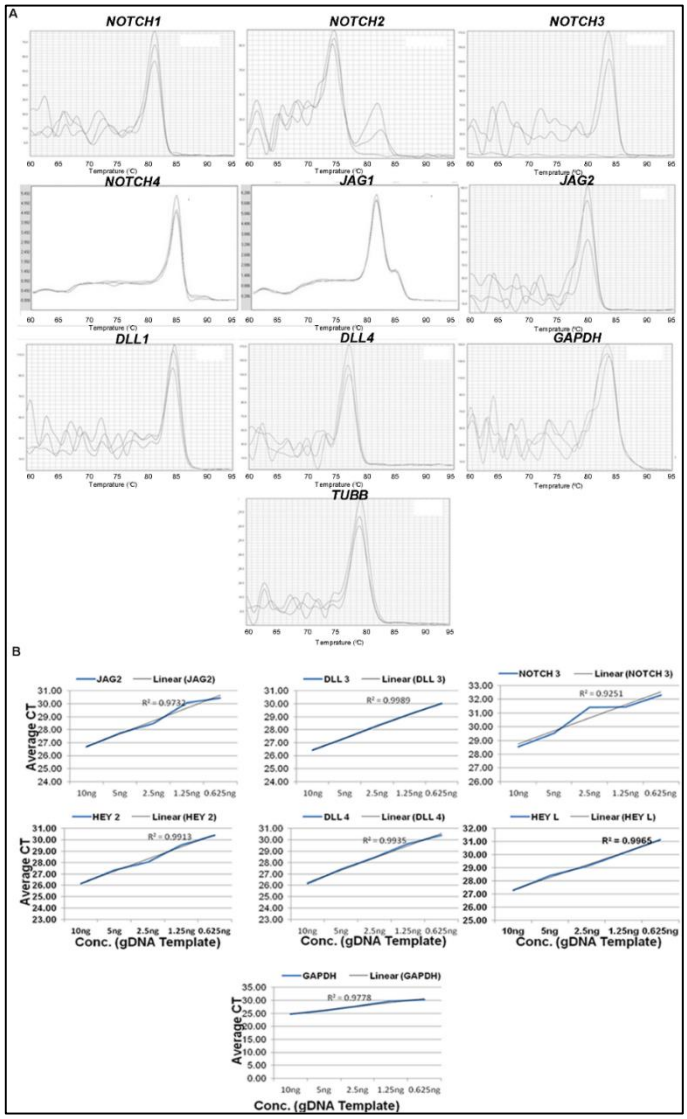
Evaluation of immunohistochemical staining of activated Notch1 expression was scored as 0, 1+, 2+ and 3+. The percentage of cells with positive staining was scored from 0 to 4 (0=0% positive cells; 1: <10% positive cells; 2: 10-49% positive cells; 3: 50-80% positive cells; 4: >80% positive cells) and staining intensity was scored from 0 to 3 (0, negative; 1, weak; 2, moderate; 3, strong). The two scores were then multiplied. Final scores of 0-2 were scored as 0, 3-5 as 1, 6-8 as 2 and 9-12 as 3.6

5.3.11 Quantitative real-time PCR for Copy number analysis

Primers details used for copy number study has been provided in Table 5.2**.** All primers used have been tested for their specificity by performing evaluative PCR as well as melt curve analysis during quantitative real-time PCR. Amplification efficiency for all primer was tested with series of dilutions (0.625 ng, 1.25 ng, 2.5 ng, 5 ng, 10 ng) of genomic DNA and PCR amplification efficiency was ~97%(~R^2=0.979) (Figure 4.1A,B). Based on above quality control, 10 ng of genomic DNA per 10 µl reaction volume in triplicates were run on Light cycler 480 (Roche, Mannheim, Germany) twice independently and relative copy number analysis was performed as described previously [204]. The threshold for calling high and low copy number was ≥2.5 and ≤1.5, respectively and ≤2.5 and 1.5≤ ; diploid.

XXI Table 5.2: Details of Primer sequences for Notch pathway gene used for DNA copy number (CNV) and expression (EXP). 5' and 3' denoted forward and reverse orientation of primer.

Primer	Sequence	Primer	Sequence
NOTCH1 5'_CNV_EXP	GTGACTGCTCCCTCAACTTCAAT	HES1 5'_EXP	TCAACACGACACCGGATAAA
NOTCH1 3'_CNV_EXP	CTGTCACAGTGGCCGTCACT	HES1 3'_EXP	TCAGCTGGCTCAGACTTTCA
NOTCH2 5'_CNV_EXP	GGCATTAATCGCTACAGTTGTGTCT	HES5 5'_EXP	CTCAGCCCCAAAGAGAAAAA
NOTCH2 3'_CNV_EXP	GGAGGCACACTCATCAATGTCA	HES5 3'_EXP	GACAGCCATCTCCAGGATGT
NOTCH3 5'_CNV_EXP	TGATCGGCTCGGTAGTAATGC	HEY2 5'_EXP	TCGCCTCTCCACAACTTCAGA
NOTCH3 3'_CNV_EXP	GACAACGCTCCCAGGTAGTCA	HEY2 3'_EXP	GAATCCGCATGGGCAAAC
TUBULIN 5'_CNV_EXP	CTACAATGCCACCCTCTCCG	HEYL 5'_EXP	GGCTGCTTACGTGGCTGTT
TUBULIN 3'_CNV_EXP	CTTCAGAGTGCGGAAGCAGA	HEYL 3'_EXP	GACCCAGGAGTGGTAGAGCAT
DLL4 5'_CNV_EXP	GTCTCCACGCCGGTATTGG	JAG1 5'_EXP	GCCGAGGTCCTATACGTTGC
DLL4 3'_CNV_EXP	CAGGTGAAATTGAAGGGCAGT	JAG1 3'_EXP	CCGAGTGAGAAGCCTTTTCAA
JAG2 5'_CNV_EXP	GGCACTCGCTGTATGAAAGGA	GAPDH 5'_EXP	AATCCCATCACCATCTTCCA
JAG2 3'_CNV_EXP	GCACAACCTCTGGTAACAAACG	GAPDH 3'_EXP	TGGACTCCACGACGTACTCA
DLL3 5'_CNV_EXP	CCCTACCCTTCCTCGATTCTG	NOTCH4 5'_CNV	AAGGTACCCCAGGTGTCAGT
DLL3 3'_CNV_EXP	GAACTGAAAATGGGCTTAAAACCTT	NOTCH4 3'_CNV	TTCAACCAGGTCTTCCACCG
NOTCH4 5'_EXP	AACTCCTCCCCAGGAATCTG	JAG1 5'_CNV	TGACCAAGCTCTGCTCAAGG
NOTCH4 3'_EXP	CCTCCATCCAGCAGAGGTT	JAG1 3'_CNV	GCAGACAAGTTGACGAGGGA
DLL1 5'_EXP	TGCAACCCTGGCTGGAAA	GAPDH 5'_CNV	GAGGCTCCCACCTTTCTCATC
DLL1 3'_EXP	AATCCATGCTGCTCATCACATC	GAPDH 3'_CNV	ATTATGGGAAAGCCAGTCCCC



XXIX Figure 5.1: Melt curve and amplification efficiency analysis of primers used for Copy number analysis.
(A) Melt curve analysis of Notch pathway genes. Distinct peak suggest that primers are amplifying single amplicon and less dimer formation. (**B)** Amplification efficiency was performed with series of dilutions of genomic DNA.

5.3.12 Quantitative real-time RT-PCR for expression analysis

Prepared cDNA was diluted 1:10 and reaction were performed in 10µl volume in triplicate. The melt curve analysis was performed to check the primer dimer or non-specific amplifications. Real-time PCR was carried out using KAPA master mix (KAPA SYBR® FAST Universal qPCR kit) in 10 µl volume in triplicate on Light cycler 480 (Roche, Mannheim, Germany) machine. All the experiments were repeated at least twice independently. The data was normalized with internal reference *GAPDH*, and analyzed by using delta-delta Ct method described previously The criteria were ≥2 fold change for upregulation, ≤0.5 fold change for down-regulation and in between 1.99-0.501 fold change as a basal expression. The details of all the primers used for expression analysis have been provided in Table 5.2.

5.3.13 Cell culture

Cell lines established from different sub-sites of head and neck cancer: AW13516 from tongue, NT8e from upper aero-digestive tract ,CAL27 cells from tongue [312] and partially transformed cell line DOK (tongue) [313] were used in this study. AW13516 and NT8e were acquired from Tata Memorial Hospital while CAL27 and DOK cells were procured from ATCC and Sigma, respectively. All cells were grown in Dulbecco's Modified Eagle Medium (Pan biotech, Germany). Culture media was supplemented with 10% FBS (Gibco, US), 1% Penicillin-Streptomycin solution (Sigma) and maintained at 37°C in an incubator with 5% CO2. DOK cells were grown with 5ug/ml hydrocortisone (Sigma) as a supplement. Trypsinization was performed using 0.25% Trypsin-EDTA (Invitrogen) and freezing of cells performed in 90 % FBS (Gibco, US) and 10% DMSO (Sigma) and were stored in liquid Nitrogen for long term storage. All the cell lines were authenticated using a short tandem repeat (STR) analysis kit (Gene Print v10, Promega, USA). The results are shown in Table 5.3.

5.3.14 Retrovirus Production, Infection and drug selection

Retroviral shRNA constructs were purchased from TransOMIC technologies, USA. Target sequences of *NOTCH1* shRNA constructs: sh1 5'-CAGTGAGCGATGACTGCACAGAGAGCTCCTAT-3', sh2 5'-CAGTGAGCGATGGACGGACCCAACACTTACAT-3', and sh3 5'-CAGTGAGCGAGACGAGGACCTGGAGACCAAAT-3'. 293T cells were seeded in 6 well plates one day before transfection and each construct (pMLP Retroviral-puro) along with pCL-ECO and pVSVG helper vector were transfected using Lipofectamine LTX reagent (Invitrogen). The viral soup was collected 48 and 72 hours post transfection, passed through 0.45µM filter and stored at 4OC. Respective cells for transduction were seeded one day before

infection in a six-well plate and allowed to grow to reach 50-60% confluency. One ml of the virus soup (1:5 dilution) and 8µg/ml of polybrene (Sigma) was added to cells and incubated for six hours. Cells were maintained under puromycin (Sigma) selection.

5.3.15 Overexpression of *NOTCH1* and selection

The human full-length *NOTCH1* (pcDNA-*NOTCH1*) [314] was obtained from Artavanis-Tsakonas Laboratory (Havard Medical School) and activated *NOTCH1* (pEGFP-NICD) [315] constructs was obtained from Annapoorni Rangarajan (Indian Institute of Sciences (IISc), Bangalore, India). Cells expressing pcDNA-*NOTCH1 or* pEGFP-NICD were generated by transfection with 10µg of DNA using Lipofectamine 3000 (Invitrogen) as per manufacturer's instructions. After 48hours, cells were cultures for 8-10 days in complete medium supplemented with 1mg/ml of G418 for antibiotics selection of transfected cells or cells were sorted based on GFP expression using BD FACSAria II. Pooled GFP sorted or antibiotics selected cells were later used for oralsphere assay. In case of 293T cells, post 48hours transfection cells were taken for RNA extraction and protein extraction for quantitative real-time PCR and western blot analysis, respectively.

5.3.16 Western blotting

Cells were lysed in RIPA buffer (Sigma) and protein concentration was estimated using BCA (MP biomedicals) method [208]. Forty microgram protein was separated on 10% SDS-PAGE gel, transferred to nitrocellulose membrane and transfer was verified using Ponceau S (Sigma). Later the blots were blocked in Tris-buffered saline containing 5% BSA (Sigma) and 0.01% Tween-20(Sigma) and were probed with full-length NOTCH1 (sc-6014-R, Santacruz biotechnology), anti-activated NOTCH1 antibody (Abcam; ab8925) and anti-actin (A5316, Sigma) antibody. The membranes were then incubated with corresponding secondary HRP-

conjugated antibodies (Santa Cruz Biotechnology, USA) and the immune complexes were visualized by Pierce ECL (Thermo Scientific, USA) according to manufacturer's protocol. Western blot experiments were performed in triplicate.

5.3.17 Anchorage-independent Growth Assay

For analysis of growth in soft agar, 5 × 103 cells were seeded in triplicate onto a six-well dish (Falcon) in 2 ml of complete medium containing 0.33% agar solution along with respective treatments of GSI-XXI at 37°C in CO2 incubator. Ten images per well were photographed after 21 days using inverted phase contrast microscope and colonies were counted manually.

5.3.18 MTT assay

A Thousand cells per well (six replicate per concentration) were seeded in 96-well plate followed by incubation with the drug for 72 hours and subsequently incubated with MTT (0.5 mg/ml) for 4 hours. Later, MTT assay was performed and data was acquired at 570nm using Microplate reader. Percentage cell viability was calculated against vehicle treated control.

5.3.19 Wound healing Assay

The cells were grown in 6 well plates to 95% confluency and were replaced with fresh medium containing 5/ml mitomycin C (Sigma). After 2 hours incubation, the medium was discarded and wounds were scratched with the help of sterile 10µl pipette tip. Cells were washed with PBS to remove the detached cells post creating a wound. The cells were fed with fresh medium and observed by time-lapse microscopy, and images were taken every 10 min for 20 hr. Migration was measured using Image J software.

5.3.20 Oralsphere formation assay

Ninety-six hundred cells were seeded in 1.2 % agar coated 6-well plates supplemented with stem cell media (recombinant EGF (20 ng/ml), human basic FGF (20 ng/ml), L-glutamine (2 mM), B-27 supplement and N2 supplement) and allowed to grow for 10 days. After every five days media, additional media was supplemented. Five hundred cells from NT8e, AW13516 and CAL27 shRNA clones were seeded on an ultra-low adherent 96-well plate in stem cell medium. Oralspheres were then cultured and maintained in low adherent 24-well plates. Additionally, the parent NT8e, AW13516, and CAL27 cells were also checked for the spheroid formation capacity upon 5 µM and 10 µM GSI-XXI administration using the same conditions.

5.3.21 ALDH activity and CD133 staining

The ALDH activity was checked using ALDEFLUOR™ detection kit (StemCell Technology, 01700) following the kit protocol and data was acquired on FACS Caliber and analysis was carried out using CellQuest software. For CD133 staining was performed using CD133 (AC133) antibody (MACS Miltenyi Biotech) in FACS buffer for 15 min in dark at 4 °C. The cells were then washed twice with staining buffer and acquired on FACS Caliber, BD Biosciences.

5.3.22 β-Galactosidase activity staining

Ten thousand cells were seeded in 12 well plates in triplicates and next day, AW13516 cells, vector control and overexpressing full length NOTCH1 were washed with 1X PBS and fixed with 0.5ml of fixative solution in the Abcam Senescence detection kit (Ab65351) for 10–15min at 25ºC. Fixed cells then washed twice with 1X PBS and incubated for 8 hours with 0.5ml of staining solution containing 20mg/ml of X-gal. Stained cells were microscopically analyzed

using Olympus IX-71. Images were analyzed using Image J and percentage β-Galactosidase positive cells were plotted.

5.3.23 Survival and Statistical analysis

The relative impact of Notch pathway alterations on disease free survival (DFS) of TSCC patients was analyzed using Kaplan-Meier method [316] and was compared using the log-rank test for statistical significance. Data are expressed as mean ± standard deviation (SD) or standard error (SE). Significant differences between selected two groups were estimated using unpaired Student t-test using Graph Pad prism version 5. Statistical significance was set at $p \leq 0.05$. Pearson correlation analysis and chi-square tests were performed in IBM SPSS statistics software version 21 for correlation analysis.

5.4 Results

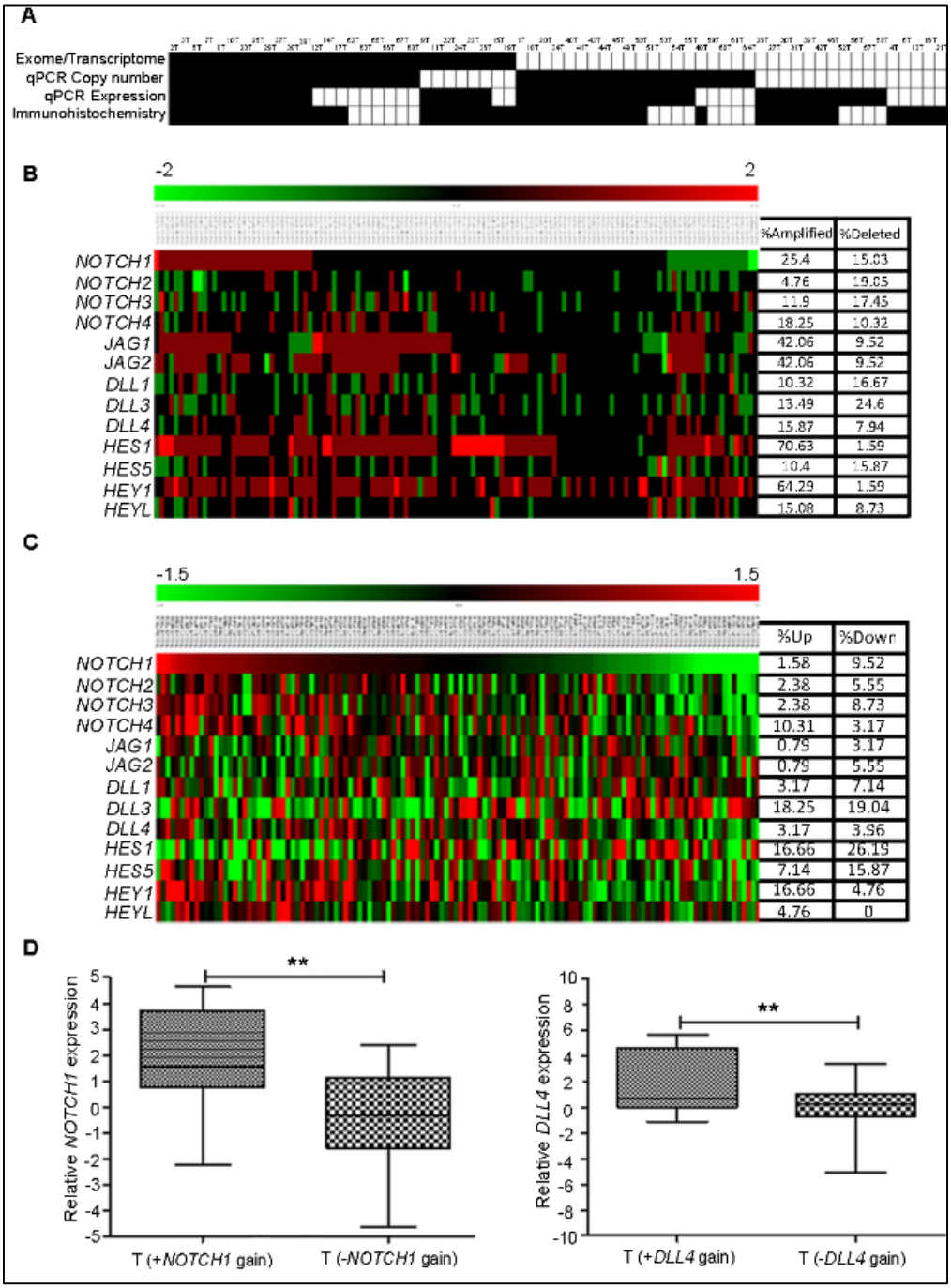
All the samples with available genomic DNA were tested for the presence of HPV using MY09/11 PCR and E6 transcript PCR primers. 40 of 71 samples analyzed, all were found to be HPV negative. Where exome sequence was available, the absence of HPV was re-confirmed using HPVDetector, as previously described [150]. TSCC samples of Indian origin to be HPV negative is consistent with other studies [317-319].

5.4.1 Notch pathway is activated in early TSCC patients

To characterize somatic alterations across 48 genes of Notch signaling pathway in 29 early-stage (T1-T2) tongue squamous cell carcinoma (TSCC) patient-derived tumors, we analyzed 23 paired whole exome and 10 whole transcriptome tongue cancer tumor sequencing data, as detailed in Table 5.1 and Figure 5.2A. Fourteen mutations were observed in 7 genes across 12 of 22 samples (Appendix IV). Of note, inactivating *NOTCH1* mutation (4%) were found at a lower frequency in our sample set than that reported from the Caucasian population [80, 127, 320] but consistent with similar finding from a recent Asian study [181, 296].

In further contrast to Caucasian population, we observed Notch family receptors, ligands, and downstream effector genes were amplified or over expressed in 59% samples (17 of 29 patients) based on copy number variations called from whole- exome and whole- transcriptome data. To extend and validate these findings, we performed real-time quantitative PCR to estimate DNA copy number and transcript levels, along with an immunohistochemical analysis of Notch pathway components in paired tumor-normal samples from tongue cancer patients. We found somatic amplification at *NOTCH1* in 12 of 38 tumors (Figure 5.3B); overexpression of *NOTCH1* transcripts was observed in 16 of 45 samples (Figure 5.3B, C)-- consistent with our analysis of the TCGA TSCC data set (n=126) (Figure 5.2B,C), not reported earlier. Also,

samples harboring amplification at *NOTCH1* (*P* value <0.001) and *DLL4* (*P* value <0.001) showed significantly higher expression of transcript as compared to no amplification. (Figure 5.2D and Figure 5.4).

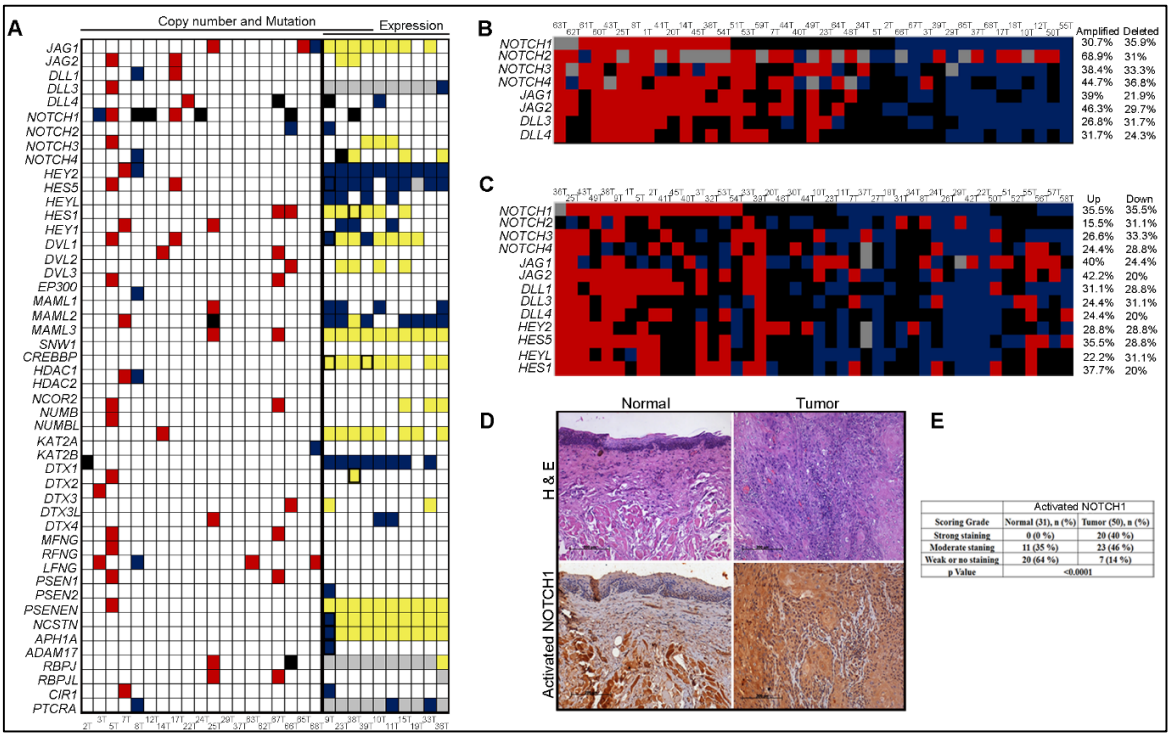


XXX Figure 5.2: Study overview and Notch pathway genes copy number and expression analysis in TCGA tongue tumors data.

(A) Study overview for Notch pathway genes in this study. Filled black box denotes sample was analyzed. (**B**) The TCGA DNA copy number data analysis of tongue cancer patients

(n=126) for Notch pathway genes. Data for Notch pathway genes (n=13) for tongue cancer patients has been retrieved. In the dataset gene-level copy number estimated values to -2,-1,0,1,2, representing homozygous deletion, single copy deletion, diploid normal copy, low-level copy number amplification, or high-level copy number amplification and heatmap were generated using MeV 4.9.0. Colors denotes: Red; high copy number, green; low copy number and black; diploid normal copy. **C)** The Expression data analysis of tongue cancer patients (n=126) for Notch pathway genes. Colors denotes: Red; higher expression, green; lower expression and black; no change. **D)** Box plot representation of *NOTCH1* and *DLL4* gene harboring DNA copy number gain in tongue tumor samples and their respective changes in gene expression was plotted. T, denotes Tumor and + and – denotes with and without DNA copy number gain or diploid. P value was calculated by unpaired student t-test, two sided considering P value ≤0.05 as threshold for statistical significance.

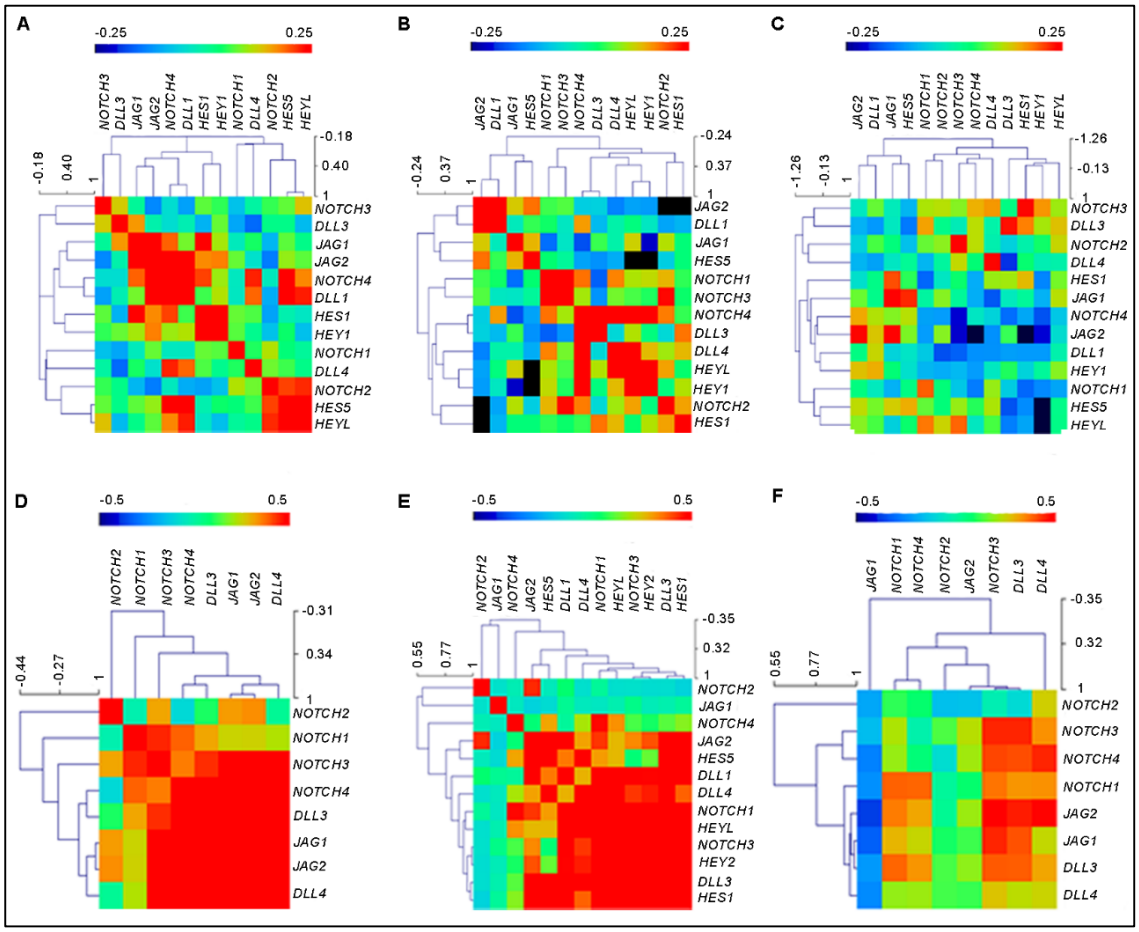
Consistent with amplification and over expression of Notch pathway components, Immunohistochemical analysis for activated NOTCH1 intracellular domain (NICD) in a set of 50 patients indicated strong immunoreactivity for active Notch signaling present in 40% tumor samples (Figure 5.3D-E, Figure 5.5A-C).



XXXI Figure 5.3: Activation of Notch pathway in early stage tongue squamous cell carcinoma.

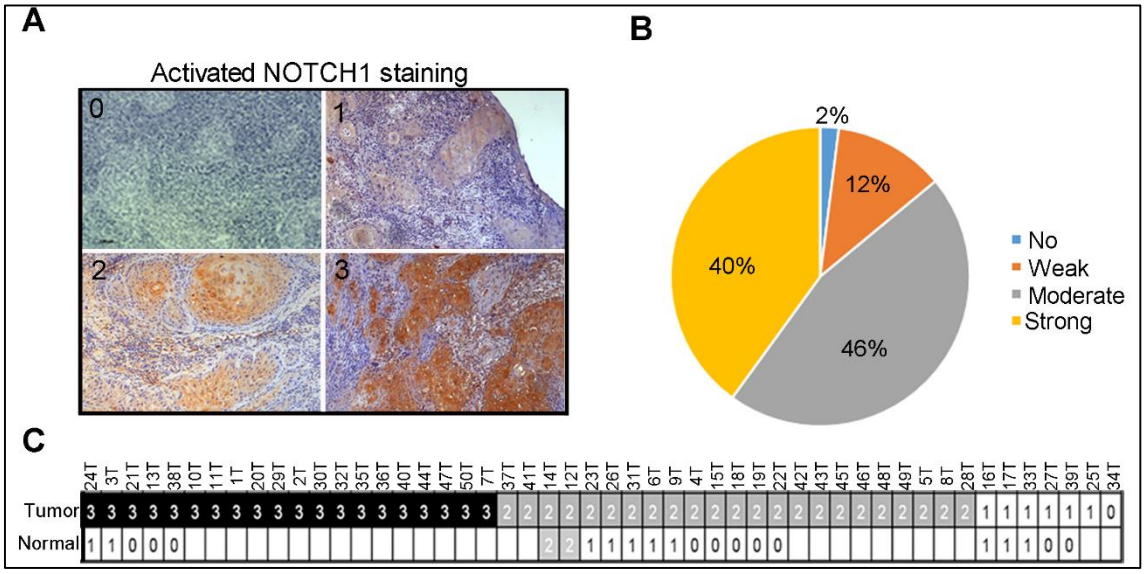
(A) Schematic representation of somatic mutation, copy number changes and expression changes identified in Notch pathway genes (*n*=48) using whole exome and transcriptome

sequencing. Red filled; copy number gains, Yellow; high transcript expression, blue; copy number loss and low transcript expression, black; mutation, white; no events, grey; transcript not detected and black borderline boxes ;any two events. Thick black line denoting separation of samples with exome and transcriptome sequencing. (**B)** Schematic representation of DNA copy number alteration of Notch pathway genes in a cohort of 41 paired tumor samples estimated by quantitative real-time PCR. Red blocks; high copy number, blue; low copy number, black; diploid and grey color; experiment could not be done or data could not be acquired. (**C)** Schematic representation of gene expression of Notch signaling pathway and its downstream targets in the cohort of 44 paired tongue tumor samples. Colors denotes: Red; upregulation, blue; down regulation, black; basal expression and grey color; experiment could not be done or results could not be acquired. (**D)** Immunohistochemistry (IHC) was performed for activated NOTCH1 in paired normal and tongue tumor samples (*n*=50). Brown color indicates positive expression. Representative IHC stained photomicrographs from normal and tongue tumor samples are shown. Scale bar, 100µM; corresponding H&E stained slides are shown in the upper panel. (**E)** Tabular representation for quantification of activated NOTCH1 immunostaining data. Significant differences of IHC staining scores between normal and tumor were estimated using the Chi-square test.



XXXII Figure 5.4: Pearson Correlation analysis of DNA copy number and expression changes.

Pearson Correlation analysis was performed in SPSS for DNA copy number and gene expression changes. Using Pearson correlation coefficient (R^2) obtained from SPSS analysis, the heatmap and clustering was performed in MeV software for three comparisons. (**A and D)** DNA copy number to DNA copy number, (**B and E)** Gene expression to gene expression and (**C and F)** DNA copy number to gene expression. A, B & C for TCGA cohort and D, E and F for our TSCC cohort.



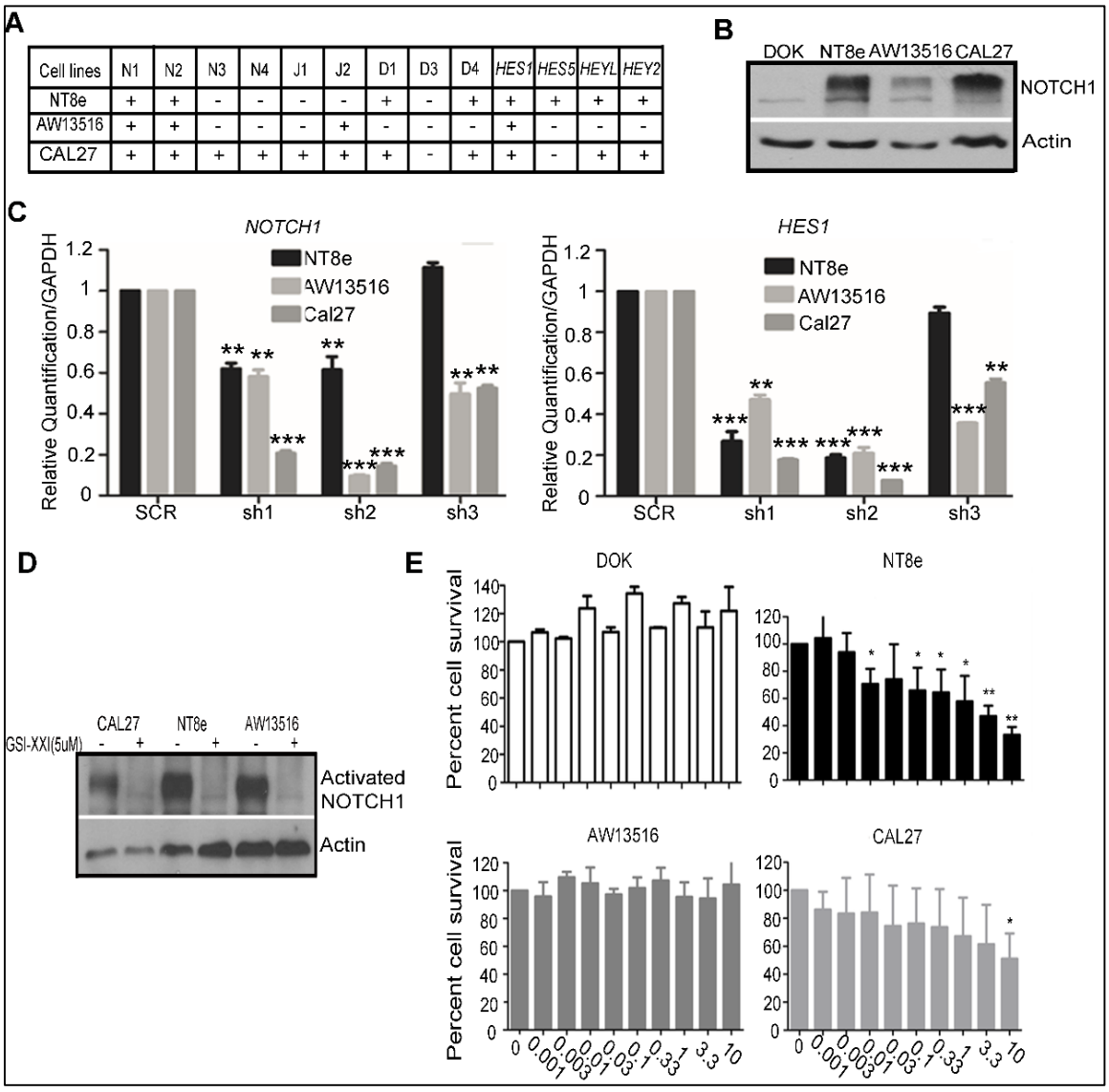
XXXIII Figure 5.5: Activated NOTCH1 immunohistochemistry in TSCC tumor samples.

(A) Representative images of different scoring pattern is shown: 0 or 1; no or weak, 2; moderate, 3; strong staining. **(B)** Pie chart representation of immunohistochemical score of activated NOTCH1 staining. **(C)** Schematic representation of immunohistochemical scores of IHC slides. Immunostaining scores obtained from each sample is indicated.

5.4.2 Expression of *NOTCH1* is required for survival, migration and stemness of TSCC tumor cells

To assess the functional significance of Notch pathway activation, we asked if the expression of *NOTCH1* is essential for survival, migration and stem-like feature of HNSCC cells *in vitro*. First, we checked for the presence of Notch pathway transcript expression by real-time PCR and western analysis of NOTCH1 protein using multiple head and neck cancers cell lines (NT8e, AW13516, CAL27 and DOK) [281]. NT8e and CAL27 cells showed higher expression of NOTCH1 as compared to AW13516 and DOK cells (Fig. 5.6A, B). Next, we tested a series

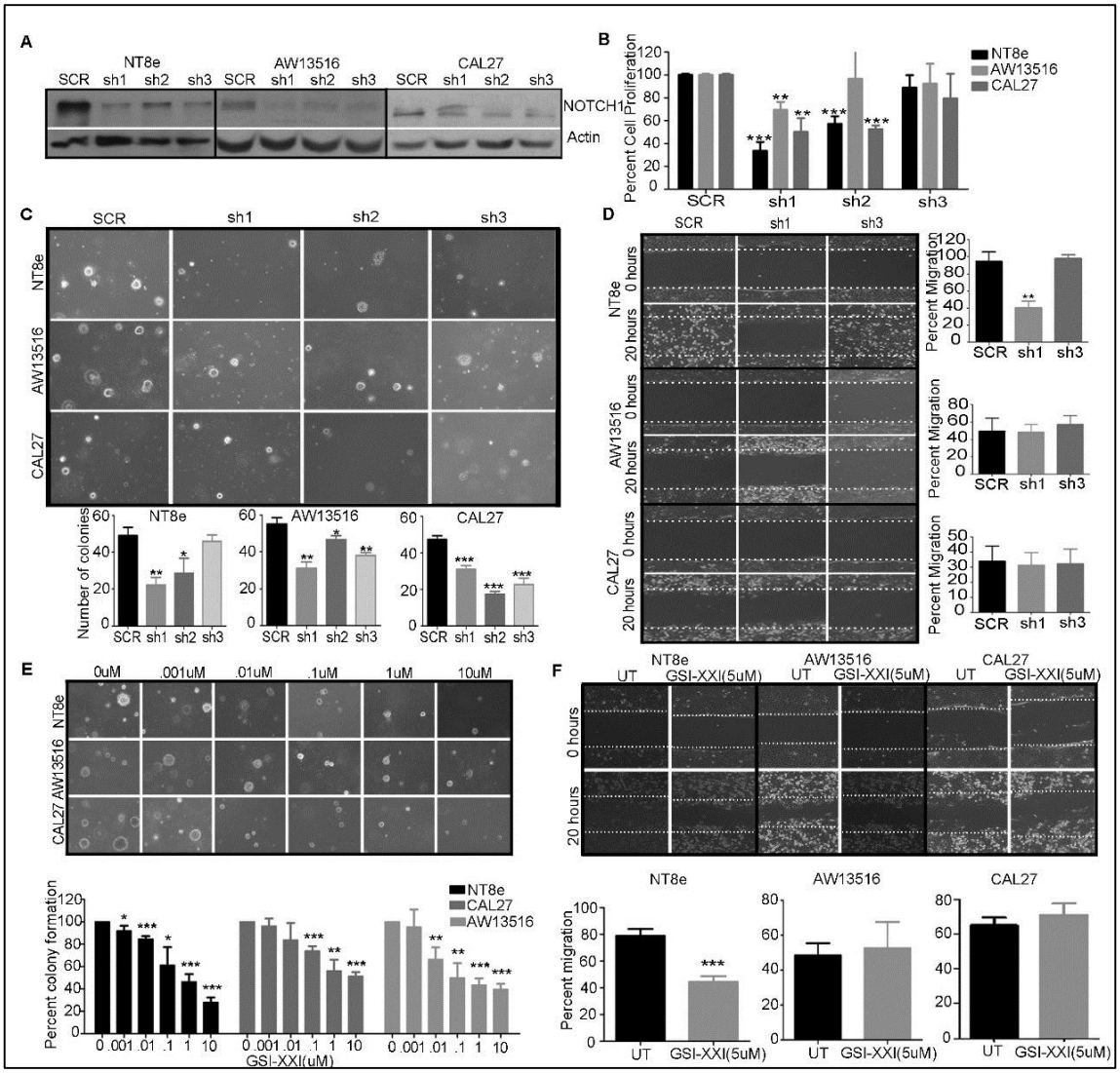
of shRNA constructs to knockdown *NOTCH1* in these cells. The knockdowns were confirmed by western blot analysis for NOTCH1 (Fig. 5.7A) and quantitative real-time PCR for *NOTCH1* and its target gene *HES1* (Supplementary Fig. 5.6C). We identified two shRNA clones sh1 and sh2 that efficiently knocked down expression of *NOTCH1* compared to scrambled (SCR). Knock down of *NOTCH1* inhibited cell survival (Fig. 5.7B), anchorage-independent growth (Fig. 5.7C), in NT8e and CAL27, and migration in NT8e (Fig. 5.7D).



XXXIV Figure 5.6: Western blot and quantitative real time PCR analysis based analysis of Notch pathway and effect of GSI-XXI on HNSCC cells.

(A) qRT-PCR based analysis of *NOTCH1* transcript expression in NT8e, CAL27 and AW13516 cells. + & - denotes presence and absence of transcript, respectively. N1, J1 and D1 denotes *NOTCH1*, *JAG1*, and *DLL1* transcripts (**B)** Western blot analysis of NOTCH1

expression in HNSCC cells. The NOTCH1 indicated in the upper panel. Lower panel denotes loading control, blotted for Actin. (**C)** Quantitative real time PCR analysis based knockdown confirmation of *NOTCH1* and *HES1* in NT8e and AW13516 cells. **(D)** Western blot analysis of activated Notch1 (NICD) after gamma secretase inhibitor (XXI) treatment post 48 hours and actin was used as loading control. (**E)** Cell lines were treated with different concentration of GSI-XXI for 48 hours and MTT assay was performed. Percent cell survival for individual cell lines was calculated and plotted using GraphPad prism version 5. Unpaired Student-t-test, two sided was used for calculating P value. Data is shown as mean ± SE. P-value is denoted as, ** ; P value < 0.001 versus untreated.; * denote P < 0.05 versus untreated.



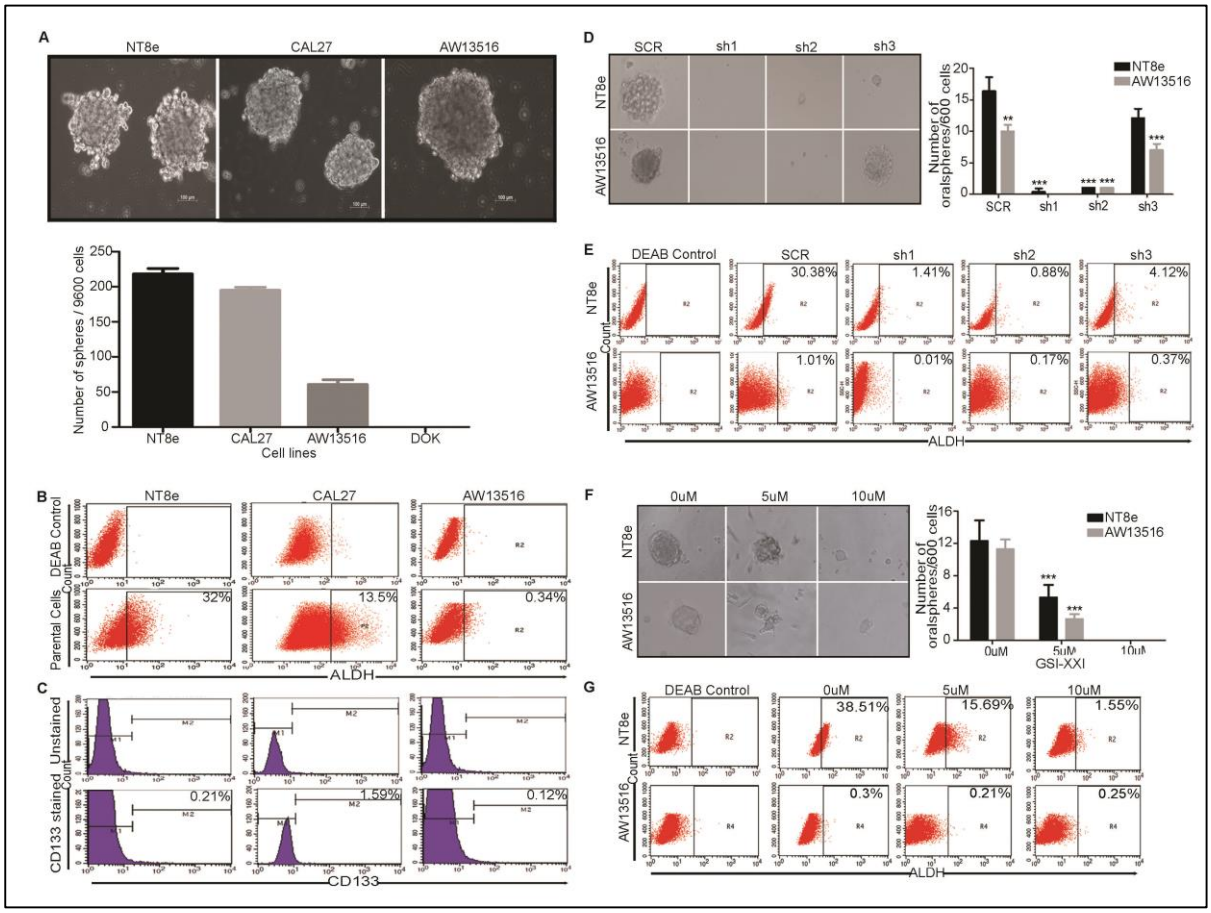
XXXV Figure 5.7: shRNA mediated knockdown and inhibition of *NOTCH1* inhibits transformation, survival and migration of HNSCC cells.

(A) shRNA constructs used to knock down *NOTCH1* expression in NT8e, AW13516, and CAL27 cells. Anti-NOTCH1 immunoblot shows that hairpins knock down to varying extents in different cells. Actin is included as a loading control. SCR, scrambled hairpin used as a negative control. (**B)** Infection with 2 of 3 independent hairpins (sh*NOTCH1*#1 and

sh*NOTCH1*#2) inhibited cell survival of NT8e and CAL27 cells expressing higher NOTCH1 levels as compared to the AW13516 cells-- as assessed by plotting total cell count on day 6 compared to day 2, normalized against cells infected with SCR. (**C)** Infection with independent hairpins inhibit soft agar colony formation by the NT8e and CAL27 cells expressing higher NOTCH1 levels compared to the AW13516 cells (upper panel). Colonies were photographed after 3 weeks (Magnification: ×10). Bar graph representation of soft agar colony formation (lower panel). (**D)** Wound healing assay of knockdown clones of NT8e, CAL27 and AW13516 cells. NT8e cells with highest migration potential was most significantly inhibited following infection with sh*NOTCH1* constructs. Percent inhibition of migration was calculated after 20 hours of wound incision. **(E)** Representative images of soft agar colony formation (upper panel) and bar graph representation of soft agar colony formation post gamma secretase inhibitor (GSI-XXI) treatment in HNSCC cell lines (**F)** Wound healing assay of NT8e, CAL27 and AW13516 cells were performed post GSI-XXI inhibitor as indicated concentration and % migration was calculated after 20 hours of wound healing. UT; Untreated. Experiment was performed in triplicate and colonies were counted and shown as mean ± SD and P value is denoted as *; $P < 0.01$, **; $P < 0.001$, ***; $P < 0.0001$ versus non-targeting shRNA. Experiments were repeated two times independently.

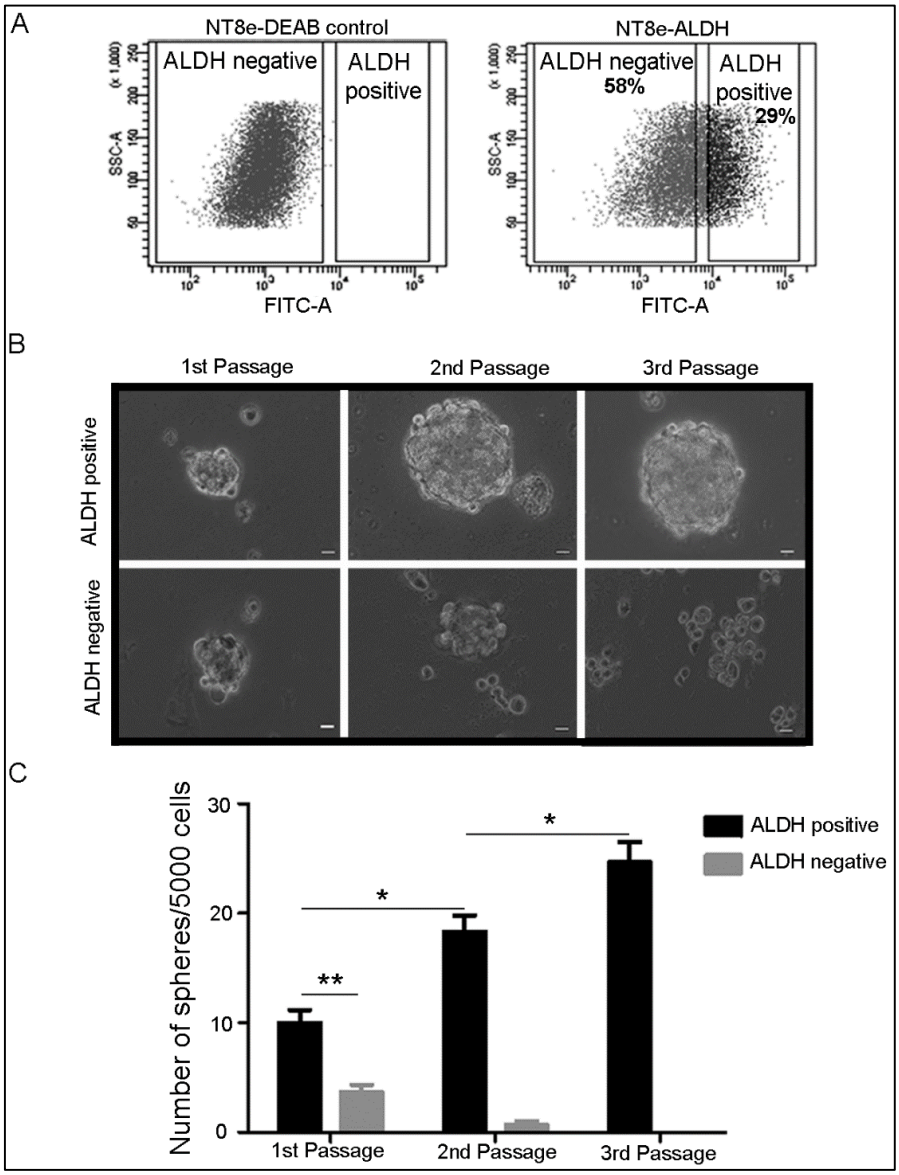
Expression of *NOTCH1* and its pathway genes maintains cancer stem-like cells (CSCs) in various tumors, as determined by their ability to form spheroids and expression of molecular markers ALDH1, CD133 and CD44 [321, 322]. An *in vitro* spheroid assay formation was performed to examine the cancer stem cell population (CSCs) in HNSCC cell lines (NT8e, CAL27, AW13516, and DOK) expressing a variable level of *NOTCH1* expression (Figure 5.6B). As shown in Fig. 3A, following 10 days of incubation in undifferentiating stem cell media, NOTCH1 over expressing NT8e cells showed a higher number of oralspheres with 32% and 0.21% NT8e cells for cancer stem-like cells molecular marker such as ALDH and CD133, respectively. Similarly, Cal27 cells also showed a significantly higher number of oralspheres with 13.5% and 1.59% CAL27 cells positive for ALDH and CD133 (Figure 5.8B,C). In contrast, AW13516 cells expressing comparatively lower NOTCH1 levels showed a reduced spheroid formation capacity with 0.34% ALDH positive and 0.12% CD133 positive while DOK cells did not show any oralsphere formation.

To test whether a high fraction of the NT8e population constitutes the stem-like cells, we sorted NT8e cells in ALDH positive and ALDH negative fraction and assessed the sphere-forming efficiency (Figure 5.9A). With subsequent passaging, the cells form ALDH negative population could not maintain their spheroid formation capacity while the ALDH positive population retained their self-renewal capacity demonstrating that indeed NT8e possess high ALDH positive cells are showing cancer stem-like cells features (Figure 5.9B, C). *NOTCH1* knockdown clones showed significant reduction in oralsphere formation ability with concomitant decrease in ALDH positive cells in NT8e and AW13516 cells as compared to scrambled (SCR) cells (Figure 5.8D,E), highlighting their dependency on NOTCH1 expression with concomitant decrease in ALDH positive population of cells, thus regulating and promoting the survival of HNSCC CSCs.



XXXVI Figure 5.8: Notch pathway is essential for cancer stem-like property of HNSCC cells.

(A) Oralsphere formation capacity of HNSCC cells. Representative images of oralsphere are shown in HNSCC cells. Oralsphere (>75µm size) were counted manually in triplicate via visualization under microscope and data was represented as Mean± SD. (**B) and (C)** Analysis of cancer stem-like cells (CSCs) marker ALDH and CD133 in HNSCC cells. Percentage ALDH positive cells were calculated against DEAB control**. (D)** Representative images of oralsphere are shown in scrambled (SCR) and different shRNA clones (sh1, sh2 and sh3) of NT8e and AW13516 cells. Oralsphere formation assay was performed in triplicate and counting was done by observing under phase contrast microscope and data was represented as Mean± SD. (**E)** ALDH staining for shRNA mediated knockdown clones and GSI-XXI treatment in HNSCC cells, respectively. Percentage ALDH positive cells were calculated against DEAB control**. (F & G)** Representative images of oralsphere post GSI-XXI treatment. NT8e and AW13516 cells post respective concentration treatment and ALDH positive cells. Number of oralsphere were counted and represented as Mean± SD. ALH staining for shRNA mediated knockdown clones and GSI-XXI treatment in HNSCC cells, respectively. Percentage ALDH positive cells were calculated against DEAB control *P-value* *; ≥ 0.05 was considered as threshold for significance. All the above experiment were performed by at least two times independently by separate individuals.



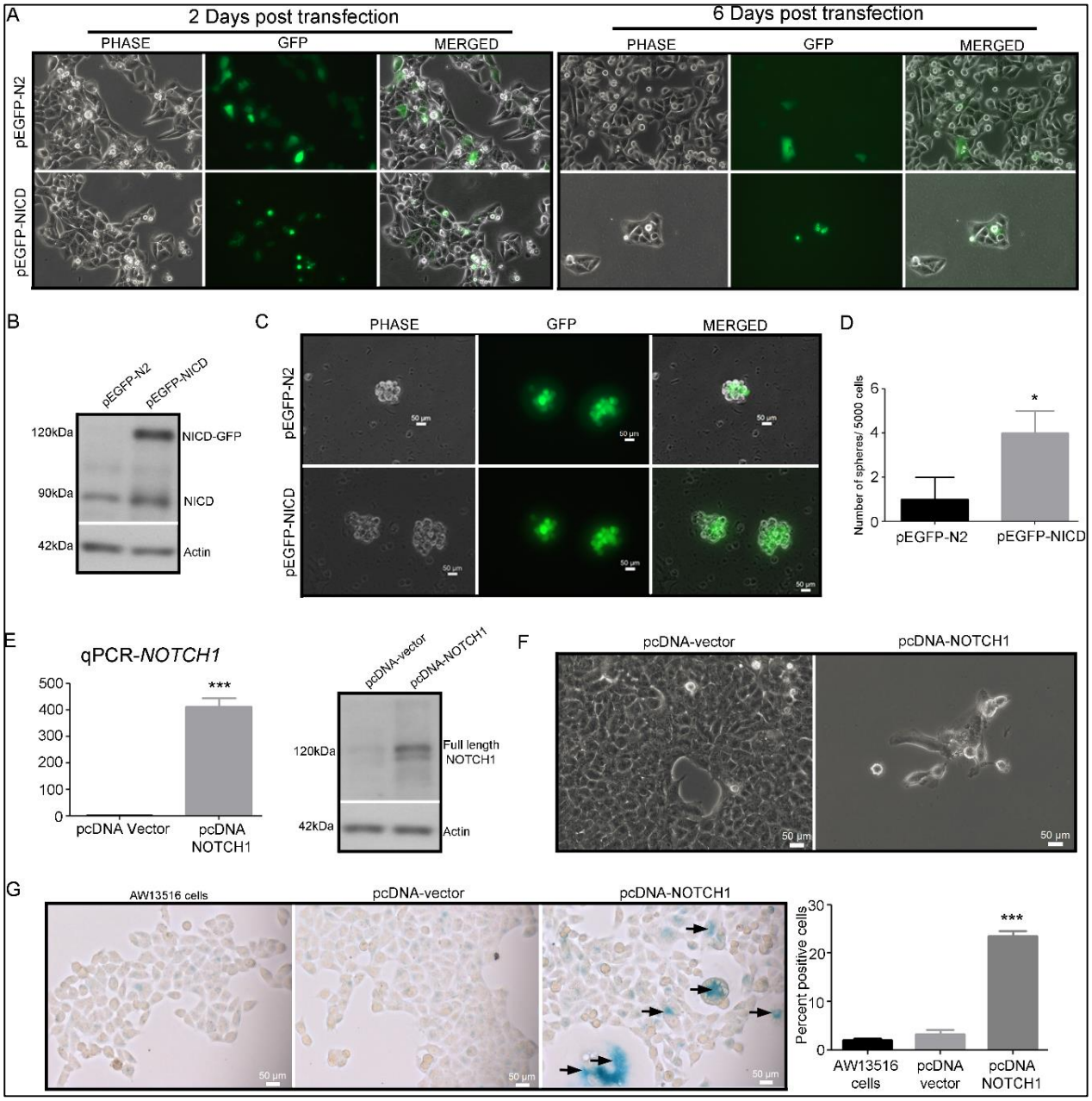
XXXVII Figure 5.9: Oralsphere formation assay of ALDH positive NT8e cells.

(A) Representative FACS analysis of ALDH activity. The ALDH positive and ALDH negative fractions in NT8e cells was determined by Aldefluor assay. Percentage ALDH positive and ALDH negative fractions are indicated at gated area. (**B)** Representative images of the oralsphere formation of ALDH sorted fractions of NT8e cells taken at every passage at 20X magnifications. The sphere-forming efficiency was assessed for ALDH positive and ALDH negative fractions. Cells were grown under anchorage-independent cell culture media and sphere were counted on day 6 as indicated in material and methods section. (**C)** Quantitative representation of sphere-forming ability of ALDH-sorted fractions of NT8e cells taken at every passage. Bar-graph reflects the means and SDs of three independent experiments. Results were considered statistically significant if, *P*- value <0.05 (* p<0.05, ** p <0.01).

Next, we attempted to overexpress activated *NOTCH1* and full-length *NOTCH1* in AW13516 cells and checked for the sphere forming efficiency. Activated *NOTCH1* form more number of spheres as compared to vector control cells post 5 days (Figure 5.10A-E). However, given that AW13516 cells are HPV negative [150] and harbor wild-type *p16INK4A* and mutant *Tp53* [281], ectopic expression of full-length *NOTCH1* or *NICD* led to continuous cell death and senescence mediated growth arrest (Figure 5.10, F,G), as described earlier [323-325].

4.4.3 Notch pathway inhibitors block stem-like feature, proliferation, and survival of HNSCC cells over expressing *NOTCH1*

Finally, we investigated whether pharmacological inhibition of Notch pathway activation would be effective against HNSCC cell lines over expressing NOTCH1. Treatment of the NT8e, CAL27, AW13516 and DOK HNSCC cell lines with gamma secretase inhibitor (GSI-XXI) that abolished the presence of activated NOTCH1 (Figure 5.6D) that resulted in significant reduction in soft agar colony formation (Figure 5.7E) and cell survival (Figure 5.6E) as compared to vehicle treated in NT8e and CAL27 cells but not AW13516 cells. Additionally, the migration potential of NT8e cells was significantly inhibited by GSI-XXI, consistent with our observation using shRNA knockdown based approach (Figure 5.7F). Similarly, marked decrease in spheroid forming ability and ALDH expression in NT8e and AW13516 cells was observed post GSI-XXI treatment (Figure 5.8F, G).



XXXVIII Figure 5.10: Effect of *NOTCH1* overexpression on AW13516 cells.

(A) Representative images of AW13516 cells transfected with vector control (pEGFP-N2) and NICD (pEGFP-NICD) and images were acquired at 20X magnification post 2 and 6 days of transfection. Cell growth arrest was observed in NICD-GFP transfected cells as compared to vector control. (**B)** Western blot analysis of NICD-GFP expression in 293T cells post 48 hours of transfection. Endogenous NICD is shown at ~90 kDa and exogenous (NICD-GFP) at ~120 kDa. Actin was used as loading control. Overexpression was observed in NICD-GFP transfected cells. (**C)** Representative images of oralsphere formation ability of AW13516 cells overexpressing NICD-GFP. AW13516 cells sorted using GFP were seeded for oralsphere assay for growing under anchorage-independent cell culture media and images were acquired post 6 days at 20X magnification. (**D)** Quantitative bar graph representation of sphere-forming ability in AW13516 cells overexpressing NICD-GFP cells. (**E)** Full length NOTCH1 overexpression

confirmation in 293T cells. **Left panel:** Quantitative real-time PCR analysis of full length NOTCH1 transcript expression in 293T cells. **Right panel:** Western blot analysis of full length NOTCH1 overexpression in 293T cells post 48 hours of transfection. (**F)** Representative images of AW13516 cells transfected with vector control (pcDNA-vector) and full-length *NOTCH1* (pcDNA-NOTCH1) and images were acquired at 20X magnification two days post neomycin selection. Cell growth arrest was observed in full-length *NOTCH1* transfected cells as compared to vector control. (**G)** NOTCH1 induced senescence observed in AW13516 cells using β-Galactosidase activity staining assay. **Left panel:** Representative image of β-Galactosidase positive (indicated by black arrow, in blue color) AW13516 cells overexpressing full length NOTCH1 (day 15 post neomycin selection). **Right panel:** Quantitative bar graph representation of percentage β-Galactosidase positive cells in AW13516. Results were considered statistically significant if, *P*- value <0.05 (* $p<0.05$; *** $p<0.0001$).

5.4.4 Activation of Notch pathway correlates with node positive and non-smoker TSCC patients

Of particular significance is the correlation between clinicopathological characteristics and overall Notch pathway activation: immuno-histochemical based expression of activated NOTCH1 intracellular domain NICD (χ^2=7.10, *P*=0.029), amplification at *DLL4* (χ^2=7.5, *P*=0.023), and transcript over expression of Notch pathway effector genes *HEY2* (χ^2=9.8, *P*=0.007) and *HES5* (χ^2=5.71 *P*=0.057) significantly correlated with lymph node metastases (Table 5.3) and poor prognosis (Figure 5.11A). Interestingly, this was consistent also with our analysis of the TCGA tongue cancer patient dataset (Figure 5.11B), and with other cancers [322, 326, 327].

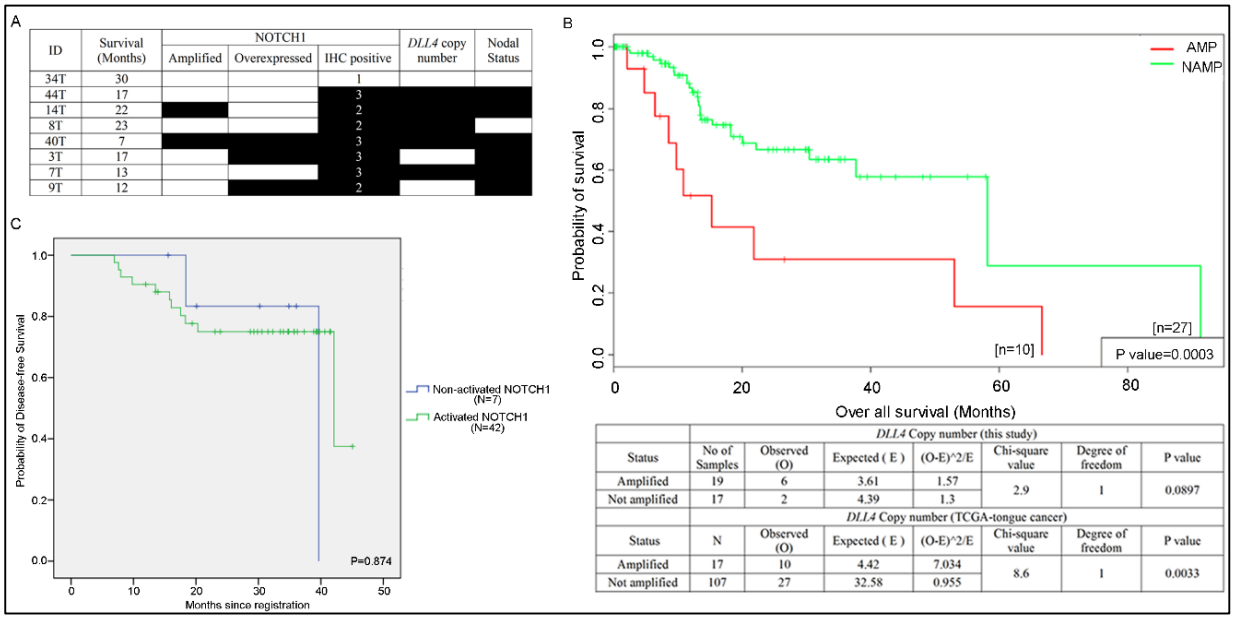
XXII Table 5.4: Clinical correlation analysis of Notch pathway alterations.

Clinicopathologic features	Variable	N (% along column)	*DLL3 Copy Number* (N=26) N (% along row)			P- value	N, % along column	*HEY2 Expression* (N=34) N (% along row)			P- value
			Gain (N=8)	Diploid (N=16)	Deleted (N=2)			Up (N=15)	Basal (N=12)	Down (N=7)	
Nodal Status	Node positive	19 (73%)	5 (26%)	14 (74%)	0 (0%)	**0.023**	22 (65%)	12 (55%)	9 (41%)	1 (5%)	**0.007**
	Node negative	7 (27%)	3 (43%)	2 (29%)	2 (29%)		12 (35%)	3 (25%)	3 (25%)	6 (50%)	

Clinicopathologic features	Variable	N (% along column)	*HES5* Expression (N=34) N (% along row)			P- value	N, % along column	Activated NOTCH1 IHC (N=49) N (% along row)			P- value
			Up (N=15)	Basal (N=12)	Down (N=7)			Strong (N=15)	Moderate (N=12)	Weak or No (N=7)	
Nodal Status	Node positive	23 (66%)	9 (39%)	12 (52%)	2 (9%)	0.057+	23 (66%)	13 (43%)	15 (50%)	2 (7%)	**0.029**
	Node negative	12 (34%)	4 (33%)	3 (25%)	5 (42%)		12 (34%)	6 (32%)	6 (32%)	7 (37%)	

Clinicopathologic features	Variable	N (% along column)	*NOTCH1* transcript expression (N=35), % along row			P- value
			Up (N=14)	Basal (N=9)	Down (N=12)	
Smoking	Smoker	11 (32%)	2 (18%)	6 (55%)	3 (27%)	**0.026**
	Non-smoker	24 (68%)	12 (50%)	3 (13%)	9 (38%)	

All clinical correlation analysis were performed in SPSS and significant alterations has been presented in the table. Patient's samples showing strong and moderate staining of activated NOTCH1 was considered as having activated Notch signaling. N; Number of samples, Up; upregulation, Down; downregulation. Chi-square test was used to calculate statistical significance. Significant *P-value* are highlighted in bold font. *P-value* ≤ 0.05 was considered as threshold for significance. + denotes; marginally significant.



ID	Survival (Months)	NOTCH1 Amplified	NOTCH1 Overexpressed	NOTCH1 IHC positive	*DLL4* copy number	Nodal Status
34T	30			1		
44T	17			3		
14T	22			2		
8T	23			2		
40T	7			3		
3T	17			3		
7T	13			3		
9T	12			2		

DLL4 Copy number (this study)							
Status	No of Samples	Observed (O)	Expected (E)	(O-E)^2/E	Chi-square value	Degree of freedom	P value
Amplified	19	6	3.61	1.57	2.9	1	0.0897
Not amplified	17	2	4.39	1.3			
DLL4 Copy number (TCGA-tongue cancer)							
Status	N	Observed (O)	Expected (E)	(O-E)^2/E	Chi-square value	Degree of freedom	P value
Amplified	17	10	4.42	7.034	8.6	1	0.0033
Not amplified	107	27	32.58	0.955			

XXXIX Figure 5.11: Survival data of patient harboring *NOTCH1* and *DLL4* alterations.

(**A)** Survival duration and clinical features representation of patients which met fatal event and their status for NOCH1 and *DLL4*. The black filled box denotes positive and while for negative. Immunohistochemistry staining score are denoted in number, 1; weak, 2; moderate, 3; strong staining of activated NOTCH1. (**B)** Disease-free survival analysis by immunohistochemistry defined NOTCH1 activation in tongue cancer patients. Patients were followed up and disease-free survival (DFS) analyzed by Kaplan–Meier survival analysis and survival difference was compared using log- rank test for statistical significance. There was no statistically significant difference in DFS between patients with activated vs non-activated NOTCH1 status; however the sample size is underpowered to detect significant difference. Numerically, the patients with activated NOTCH1 tumors has an inferior DFS as compared to those with NOTCH1 non-activated tumors. (**C)** Kaplan–Meier survival analysis of TSCC patients with or without *DLL4* amplification. Survival analysis of tongue cancer patients with and without *DLL4* amplification in TCGA- tongue cancer. AMP; amplified and NAMP; not amplified. Death of patients was taken as end point of analysis and P value ≤0.05 was considered as threshold for statistical significance.

XXIII Table 5.5: Details of correlation between clinicopathologic features of tongue cancer patients by IHC defined activated NOTCH1 status.

Clinicopathologic features	Variable	N (% along the column)	Activated NOTCH1 staining (n=49) N (% along row)		P-value*
			Strong and Moderate (n=42)	Weak (n=7)	
Age	<45 years	22 (45%)	18(82%)	4(18%)	0.68
	>45 years	27(55%)	24(89%)	3(11%)	
Sex	Male	33(67%)	30(91%)	3(9%)	0.19
	Female	16(33%)	12(75%)	4(25%)	
AJCC Stage	I-II	19 (38%)	13(68%)	6(32%)	**0.01**
	III-IVA	30 (61%)	29(97%)	1(3%)	
Alcohol	No	35 (71%)	29(83%)	6(17%)	0.65
	Yes	14 (29%)	13(93%)	1(7%)	
Tobacco	No	20 (41%)	16(80%)	4(20%)	0.42
	Yes	29 (59%)	26(90%)	3(10%)	

Highlighted P-value are statistically significant or marginal significance. * Fisher-exact test

In addition, *NOTCH1* expression significantly correlated with a non-smoking habit of patients (χ^2=7.325, *P*=0.026), where 12 of 24 non-smokers patients derived tumors showed upregulation of *NOTCH1* transcript, consistent with previously described *NOTCH1* upregulation in non-smokers in other diseases including lung adenocarcinoma [328-330]. We also observed a significant correlation with AJCC (American Joint committee on Cancer) TNM tumor staging wherein stage III-IVA showed increases activation of NOTCH pathway (χ^2=7.84, P=0.02). However, no statistically significant correlation was observed between the activated NOTCH1 expression with the sex, age, alcohol, and tobacco consumption in the cohort, as represented in Table 5.5. Next, we performed an interim analysis and assessed

disease-free survival (DFS) by IHC defined activated NOTCH1 (strong and moderate staining) status vs non-activated NOTCH1 (weak or no staining) status. DFS was defined as time interval between the date of registration and the date of first documented evidence of relapse at any site (local, regional, metastatic, or secondary primary) or death from any cause, whichever earlier. There was no statistically significant difference was observed in tumors with activated NOTCH1 compared to those with non-activated NOTCH1 tumors, as shown in Supplementary Figure 5.11A-C.

Taken together, we present a novel clinicopathological correlation such that expression of Notch pathway components and activated NOTCH1 levels predispose TSCC patients to lymph node metastasis, and that non-smokers TSCC patients tend to have higher *NOTCH1* levels as compared to smokers. Clinically, determination of NICD by immuno-histochemistry could be a good predictor of nodal status. This could be a biomarker to predict lymph node metastasis for therapeutic utility among early stage tongue cancer patient to help patient stratification for treatment [44, 110].

4.5 Discussion

We demonstrate that 40% TSCC tumors have strong Notch pathway activation and that this property may be important in the maintenance of stem cell component in these tumors. Genetic or chemical perturbation of NOTCH pathway using shRNA and GSI-XXI showed decrease soft agar colony formation, migration potential and cancer stem-like features of HNSCC cells, highlighting their dependency on *NOTCH1* expression. Thus, targeted elimination of these cells may provide a new lead in treatment of head and neck cancer. These findings are consistent with reports where Notch signaling has been shown to be required for stem cell-like features in several cancer types [133, 295, 297, 331, 332]. Interestingly, genetic determinants

of cancer stem cells share features with their role in the development of tumorigenesis [306]. These findings are consistent with reports where Notch signaling has been shown to be required for stem cell-like features in several cancer types, including HNSCC [333, 334].

Clinically, *NOTCH1* transcript expression significantly correlate with non-smoking habit of patients, consistent with previous reports in other pathological conditions including lung adenocarcinoma [328-330]; lymph node metastasis in tongue cancer correlate with poor prognosis and survival of the patients, thus activated NOTCH1 could serve as a reliable marker to predict lymph node metastasis [44]. Moreover, AJCC TNM tumor stage III-IVA significantly correlates with activation of Notch pathway as compared to stage I-II, consistent with reports in HNSCC [331]. The sample size in this study, however, is underpowered to reach the statistical significance for survival data. No significant difference was observed in disease-free survival of the patients with IHC defined activated NOTCH1 tumors as compared to non-activated NOTCH1.

In conclusion, we demonstrate that a considerable fraction of TSCC tumors has upregulated Notch pathway and that this property may be important for the maintenance of stem cell component in these tumors. And that, *NOTCH1* could be a potential therapeutic target in these patients.

Chapter 6

Gene expression meta-analysis identify *MMP10-miR-944* axis in early tongue cancer tumors

(as submitted to European Journal of Cancer)

6.1 Gene expression meta-analysis identify *MMP10-miR-944* axis in early tongue cancer tumors

6.2 Abstract

Background: Nodal metastases status plays a decisive role for choice of treatment in early stage tongue squamous cell cancers; about 70% patients may be spared from surgery with accurate prediction of negative pathological lymph node status. However, there is an unmet need for prognostic biomarkers to stratify the patients who are likely to develop metastases.

Material and Methods: We performed whole transcriptome sequencing of 18 early primary tongue samples. Gene expression meta-analysis was carried out for 253 tongue cancer samples for transcriptomic alterations, integrating 4 published with our datasets. Candidate genes and miRNAs were validated using qPCR and immuno- histochemical analysis in an extended set of 50 early primary tongue cancer samples.

Results: Gene expression meta-analysis of 5 data sets derived from 253 tongue cancer samples including our whole transcriptome data of 18 samples identified metastatses-related pathways to be significantly upregulated, involving 9 matrix metalloproteases. Our qRT-PCR and immuno- histochemical analysis confirmed the overexpression of *MMP10* in an additional set of 50 early primary tongue tumor samples. Further, we analyzed for negative expression correlation of 7 miRNAs that were predicted to target *MMP10* across 21 tumor samples. Of these, we identify *miR-944*, a novel miRNA, to regulate *MMP10* expression by targeting *MMP10* 3'UTR using luciferase assay. In overall, we present the first systematic expression profiling of *MMP10 - miR-944* axis in early primary tongue cancer samples.

Conclusion: The experimental validation of *MMP10- miR-944* interaction and their negative expression correlation suggests *MMP10* as a potential prognostic biomarker in early stage tongue cancer patients.

6.3 Introduction

Oral cavity cancer is the sixth leading cause of cancer worldwide and in India, it is a major health problem and accounts for over 30% of all cancers [153, 335]. Tongue squamous cell carcinoma (TSCC) cases comprises of two-thirds of all cancer cases in HNSCC and regarded as a biologically unique entity compared to other sub-sites affected with cancer [162, 163].

There has been few studies describing the gene expression profile of advanced stage TSCC tumors using microarray approach and identified *CDK1*, *NDRG1*, *ADAM15*, *CDC7*, *MMP9*, *TNFRSF8*, *NPM* and *CHES1* as biomarker for oral cancer progression [336, 337]. Additionally, there have been several studies describing the identification of microRNAs (a class of non-coding RNA acts as posttranscriptional regulators) as promising molecular biomarkers in diagnosis and prognosis of various cancers including oral cancer [338-340]. Several oncogenic (*miR-21, miR-155, miR-106b-25, miR-130b* and others) and tumor suppressive (*miR-1, Let -7, miR-29s, miR-375, miR-99a, miR-99b, miR-100* and others) microRNAs have been described in oral cancer [338, 339, 341, 342]. However, the reproducibility of the gene expression profiles have been poor possibility due to biological, technical and experimental differences leading to difficulty in translation of these biomarkers for clinical benefit [343, 344]. Growing high-throughput data archives such as Gene Expression Omnibus, TCGA and ICGC repositories allows to apply the meta-analysis approaches in the gene expression studies previously done independently across different platforms to identify high confidence markers which would have clinical applicability [345]. Meta-analysis studies in HNSCC and few in TSCC has led to identification of robust molecularly defined sub-types which could improve the patient selection and design better therapeutic strategies [280, 346-349].

Here, we integrate our whole transcriptome data with the Cancer Genome Atlas and published sources to analyse 255 tongue cancer samples for transcriptomic alterations. We identify a

robust gene expression profile of differentially expressed genes involving matrix metalloproteinases (MMPs) that are known to mediate tumor invasion and metastasis as underlying events leading to tumor dissemination. Of 9 MMPs identified, we show *MMP10* is regulated by *miR-944* and overexpressed in early stage tongue cancer patients.

6.4 Material and Methods

6.3.1 Patient's details

All selected patients have undergone surgery as the initial modality of treatment and the tumor samples were collected and frozen for molecular analysis at Memorial Hospital and Advanced Centre for Treatment, Research and Education in Cancer (ACTREC), Mumbai. The sample set and study protocols were approved by (ACTREC-TMC) Internal Review Board (IRB) and most of the patients were recruited from 2010-2014 with predefined inclusion criteria of early (pT1 and pT2) stage. The percent tumor content was evaluated using hematoxylin and eosin based staining by two independent pathologists and tumor content varied from 60-90%.

6.3.2 Tissue processing and RNA extraction

The tissue samples were cut into small pieces immediately after surgical resection and stored in RNA later (Ambion) at -80oC freezer for long term storage. The homogenization was performed using Fastprep homogenizer (MP Biomedicals, USA) as per manufacturer's instructions in 1ml Trizol reagent (Life Technologies) according to manufacturers for total RNA extraction. Later, RNA was subjected to DNase (RNase free) treatment to remove genomic DNA contamination. The RNA integrity was assessed using RNA 6000 Nano Lab Chip on the 2100 Bioanalyzer (Agilent), following the manufacturer's instructions. The quality

of the RNA which was considered to transcriptome sequencing which was having rRNA 28S/18S ratios was greater than or equal to 1.5.

6.3.3. Transcriptome Sequencing

Transcriptome sequencing libraries were prepared following standard Illumina TruSeq RNA library protocol as described previously [257]. Briefly, mRNA was purified from 4µg of intact total RNA using oligodT beads. Purified mRNA was processed for cDNA synthesis using Superscript II Reverse transcriptase by priming with Random Hexamers oligos. Later second strand cDNA was synthesized in the presence of DNA Polymerase I and RNase H. Cleaning of cDNA were done using Agencourt Ampure XP SPRI beads (Beckman-Coulter). Cleaned cDNA was ligated with Illumina Adapters after end repair and the addition of A base. The second cleanup was performed after ligation step. PCR amplification of libraries was done to enrich them. The prepared libraries were quantified using Nanodrop and validated for quality by running an aliquot on Bioanalyzer Chip (Agilent). The cluster generation was performed using TruSeq PE Cluster Kit protocol (Illumina) and 7 pmol of each library will be loaded on Illumina flow cell (version 3) for cluster generation on cBot cluster generation system (Illumina) and clustered flow cell and sequencing using Illumina paired-end reagents TruSeq SBS Kit (Illumina) for 200 cycle.

6.3.4 Transcriptome sequencing data analysis

Post de-multiplexing using CASAVA (Illumina), the alignment of raw reads performed against human reference genome (hg19) using TOPHAT package [267, 310]. Cufflinks, a transcript assembler was used to perform reference guided transcript assembly of aligned reads for individual samples individually. Further differential expression analysis has been performed using cuffmerge and cuffdiff package. The downstream analysis and quality control check were

carried out using CummeRbund package [267]. The genes showing gene expression fold change up-regulation ≥2, downregulation ≤-2 and a p-value <0.05 were considered differentially expressed in two groups. Multi-experiment Viewer (MeV) software was used to generate the heat maps [350].

6.3.5 Gene expression dataset meta-analysis and GSEA analysis

TSCC gene expression profiling studies were identified searching Gene Expression Omnibus (GEO) database [351] and all the identified datasets were downloaded. The studies having both normal and tumors samples, greater than 10 tumor samples and having tongue subsites were included into the study. Studies involving cell lines, few number of samples, and non-human tissue samples were excluded from the analysis. The TCGA-HNSCC dataset was downloaded from Cancer Genome Browser [352] and tongue sub-site data was extracted for the analysis. Gene set enrichment analysis (GSEA) was performed selecting KEGG gene set in MutSigDB[268] to identify underlying biological processes and pathways.

6.3.6 Gene miRNA target prediction, Primer designing

The miRNAs targeting *MMP10*- 3' UTR were predicted using miWalk v2.0 [353]. The primers for miRNAs qRT-PCR were designed by retrieving the sequences of miRNAs from miRBase [354] and converted to cDNA using web tool (http://www.attotron.com/cybertory/analysis/trans.htm). For increasing the Tm of primers to 60oC, 2-3 bases 'GC' was manually added at the 5' end of primers. The primers for genes qRT-PCR were designed as described previously [257]. The melt curve analysis was carried out to determine the specificity of primers. The primer sequences for genes are provided in Table 6.1.

XXIV Table 6.1: Primer sequences of miRNA and Genes used in qRT-PCR analysis.

6.3.6 Gene qRT-PCR analysis

Primer ID	Primer sequence
OAD933_MMP10_5'	CATACCCTGGGTTTTCCTCCAA
OAD934_MMP10_3'	GTCCGCTGCAAAGAAGTATGTTTTC
OAD971_MMP10_3'UTR_F	GCTCTAGAGCGAGATAGGGGGAAGAC
OAD972_MMP10_3'UTR_R	GCTCTAGAAGAAAGTAAGGAACAGGCC
OAD_1021_MMP14_5'	GGCTACAGCAATATGGCTACC
OAD_1022_MMP14_3'	GATGGCCGCTGAGAGTGAC
OAD_1023_MMP12_5'	GATCCAAAGGCCGTAATGTTCC
OAD_1024_MMP12_3'	TGAATGCCACGTATGTCATCAG
OAD_1029_MMP11_5'	CCGCAACCGACAGAAGAGG
OAD_1030_MMP11_3'	ATCGCTCCATACCTTTAGGGC
OAD_1031_MMP13_5'	ACTGAGAGGCTCCGAGAAATG
OAD_1032_MMP13_3'	GAACCCCGCATCTTGGCTT
OAD_521_GAPDH_5'	AATCCCATCACCATCTTCCA
OAD_522_GAPDH_3'	TGGACTCCACGACGTACTCA
OAD1481_miR-148a-3p	GCGCTCAGTGCACTACAGAACTTTGT
OAD1482_miR-152-3p	GGCTCAGTGCATGACAGAACTTGG
OAD1483_miR-148b-3p	GCGCTCAGTGCATCACAGAACTTTGT
OAD1484_miR-454-3p	GCGCTAGTGCAATATTGCTTATAGGGT
OAD1485_miR-130a-3p	GCCAGTGCAATGTTAAAAGGGCAT
OAD1486_miR-496	GCGCTGAGTATTACATGGCCAATCTC
OAD943_mir-944	ccgAAATTATTGTACATCGGATGAG
OAD_1264_miR-944_F cloning	GTAGGATCCCACCAACTAACAAATTCAG
OAD_1265_miR-944_R cloning	GTACTCGAGCACTAGACAGATTCTCC
U6_5'	GCTTCGGCAGCACATATACTAAAAT
U6_3'	CGCTTCACGAATTTGCGTGTCAT

A 2µg total RNA was used for the cDNA synthesis using High capacity cDNA reverse transcription kit (Applied Biosystems) and quantitative PCR (qPCR) was carried out using KAPA master mix (KAPA SYBR® FAST Universal qPCR kit). The qPCR performed by dilution cDNA 1:10 and in 6 µl volume in triplicate on Light cycler 480 (Roche, Mannheim, Germany) machine where *GAPDH* was used as internal controls for normalization. The data analysis was performed using the $2^{-deltadeltaCt}$ method where *GAPDH* was used as reference gene. The primer sequences for genes are provided in Supplementary Table 6.1.

6.3.7 miRNA qRT-PCR analysis

The transcript level of miRNAs was analyzed using qRT-PCR. In brief, 1 µg RNA was used for the reverse transcription using Mir-X miRNA First-Strand Synthesis Kit (Clontech Takara) and PCR (qPCR) was carried out using the Mir-X miRNA qRT-PCR SYBR Kit (2X) Master Mix (Clontech Takara). The data analysis was using the $2^{-deltadeltaCt}$ method and *U6* miRNA was used as internal reference [204]. The average C_T values of each samples reference gene

were normalized to C_T values of candidate miRNAs and further to calculate fold change. The primer sequences for miRNAs are provided in Supplementary Table 6.1.

6.3. Immunohistochemical analysis

The tissue processing was performed as described previously [257]. The paraffin section slides were microwaved by incubating them at 58oC for 30 minutes, followed by deparaffinization, rehydration, and quenching. The antigen retrieval was performed in a pressure cooker (1 whistle) with citrate buffer at pH 6.0 followed by cooling at room temperature (RT). 1:50 diluted horse serum from Vectashield Kit (Vectashield, Vector laboratories, USA) was used for the blocking. Tissue sections were incubated overnight at 4oC with the anti-MMP10 antibody (Abcam; ab59437). Next, slides were rinsed in wash buffer and incubated for the 30 minutes with biotinylated secondary antibody provided with Vectashield Kit, followed by tertiary antibody (Reagent A & B) incubation for 1 hour. The chromogenic reaction was carried out with 3,3′- diaminobenzidine chromogen solution for 5 minutes leading to the expected brown color signal. Later, slides were rinsed in deionized water and counterstained with hematoxylin, dehydrated and mounted with the DPX mounting reagent and cover slip. The evaluation of immunohistochemical staining as described previously [257].

6.3.9 3'UTR cloning and Luciferase assay

The primers were designed to amplify the 400bp flanking the miR-944 seed sequence from the genomic DNA of A549 cells. The amplicons were sequencing verified for mutations using Sanger sequencing and cloned in a (Fermentas, USA) followed by subcloning in BamHI and HindIII sites of pCDNA 3.1 (-), a CNV promoter based expression vector (Invitrogen). The expression of mir-944 was verified by doing transfection followed by qPCR in 293FT cells. Similarly, complete *MMP10*-3'UTR was amplified by designing the primers flanking the

3'UTR with XbaI sites for cloning in pGL3-promoter vector (Luciferase Expressing vector, Promega). Sequencing of the construct was done to ensure the absence of any mutation. The luciferase assay was performed using Dual Luciferase Reporter assay kit (Promega, USA) following manufacture's instructions. In brief, 293FT cells (50,000/ well) were seeded in a 24 well plate (Nunc plates) 12 hours prior to transfection. The transfection was performed using lipofectamine 3000 reagent (Life Technologies, USA) along with Renilla luciferase vector (for normalizing transfection efficiency) and 15 pmol mirVana miRNA-inhibitors (Ambion, USA) per well. The cells were lyzed post 48 hours of transfection and luciferase assay was performed to measure luminescence using luminometer (Berthold Luminometer, Germany). The data was plotted as the ratio of firefly to rennila luciferase as described previously [355].

6.3.10 Statistical and Clinical correlation analysis

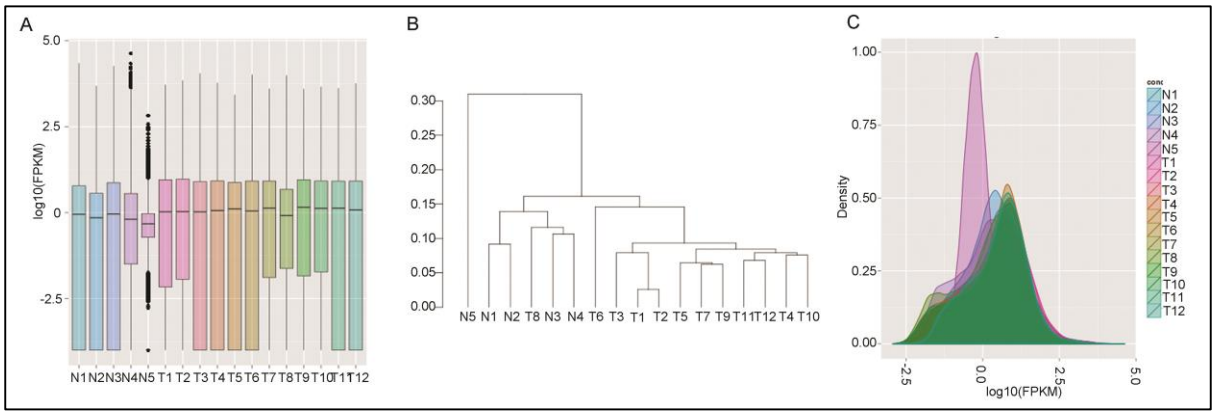
The clinical correlation and survival analysis was performed as described previously using SPSS package [257]. The data are expressed as mean± SD or SE. The unpaired Student-t-test was used to determine the significance between two using Graph Pad prism version 5. The threshold for statistical significance was set at P value ≤ 0.05.

6.4 Results

6.4.1 Gene expression analysis of tongue squamous cell carcinoma identified recurrently deregulated genes and pathways

To identify the global gene expression changes in tongue cancer tissues we performed massively parallel paired-end cDNA sequencing of primary tumors (n=12) and adjacent normal (n=5) tissue samples and generated an average of 25 and 34 million paired end reads per sample for normal and tumor samples, respectively. We observed an average of 11824 (SD±606)

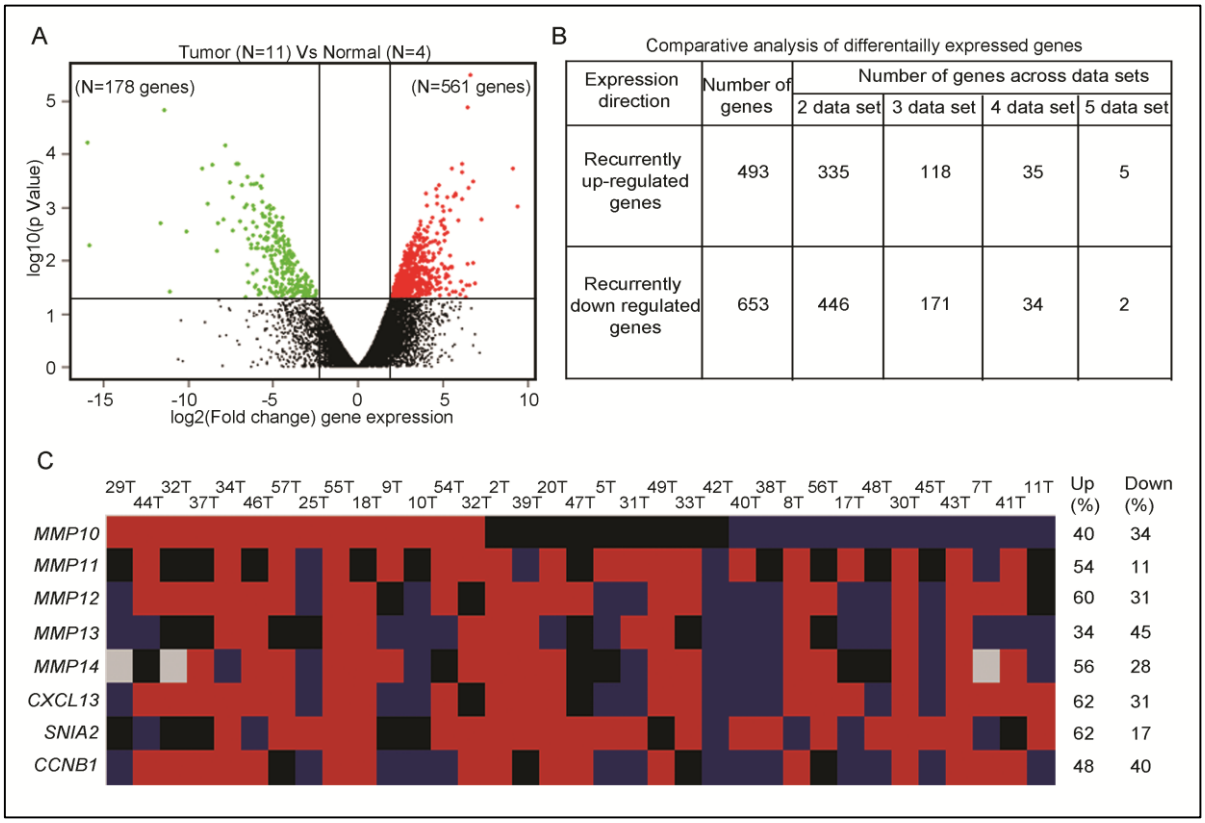
transcripts in each samples having expression level ≥1 FPKM, which included the majority of annotated human reference genes. We further performed the unsupervised hierarchical clustering and observed robust classification of normal and tumors samples in distinct clusters (Figure 6.1A-C). We also observed single normal and tumor samples were having abnormal expression profile and were misclassified, which were further excluded from the differential expression analysis.



XL Figure 6.1: Quality control analysis of transcriptome sequencing data.

(A) Box pot representation of log10 (FPKM) values for each samples to accesses the overall distribution. **(B)** Dendrogram representation using unsupervised hierarchical clustering of each samples based on global gene expression profile. Normal and tumor samples are distinctly clusterd besides two samples showing mixed behaviour. **(C)** Density plot showing the bell shaped curve suggesting the uniform distribution of transcriptome sequencing across samples for each gene. Sample N5 and T8 were found to be outlier and showing unusual pattern.

We identified 739 significantly differentially expressed by applying *P-value* ≤0.05 and log2 fold change 2 as a cutoff (Figure 6.2A). Of the 739 DEGs identified, 561 genes were up-regulated and 178 genes were down-regulated in tongue tumor samples as compared to adjacent normal samples (Appendix V).



Expression direction	Number of genes	Number of genes across data sets			
		2 data set	3 data set	4 data set	5 data set
Recurrently up-regulated genes	493	335	118	35	5
Recurrently down regulated genes	653	446	171	34	2

XLI Figure 6.2: Differential expression profile of tongue squamous cell carcinoma using mRNA sequencing and meta-analysis.

Differential expression analysis to identify the distinct gene expression profile of tongue tumors. **(A)** Volcano plot representation of differentially expressed in between early tongue tumors and adjacent normal tongue tissues. The red and blue dots denote the up-regulated and down-regulated differentially expressed genes with P value < 0.05 and fold changes ≥ 2 or ≤ 2 for, respectively. **(B)** The tabular representation of a number of genes overlapped in tongue cancer across different studies. **(C)** Schematic representation of commonly up-regulated qRT-PCR analysis data in a cohort of 35 paired tongue tumor samples. The Red denotes up-regulation, blue as downregulation, black as basal expression and gray color; experiment could not be done or results could not be acquired. The ≥2 mean fold change is for up-regulation, ≤0.5 mean fold change for down-regulation and in between 1.99-0.501 mean fold change as a basal level expression compared to the adjacent normal tissue sample.

Several studies previously reported the key genes and pathways deregulated in tongue squamous cell carcinoma tumors. To determine whether gene expression profile in this study was in agreement with the previous studies, we performed a systematic meta-analysis of 4 GEO (microarray) and TCGA (transcriptome sequencing) datasets comprising of 243 tongue tumors and 79 adjacent normal tissue samples expression profile using BRB array toolkit [356]

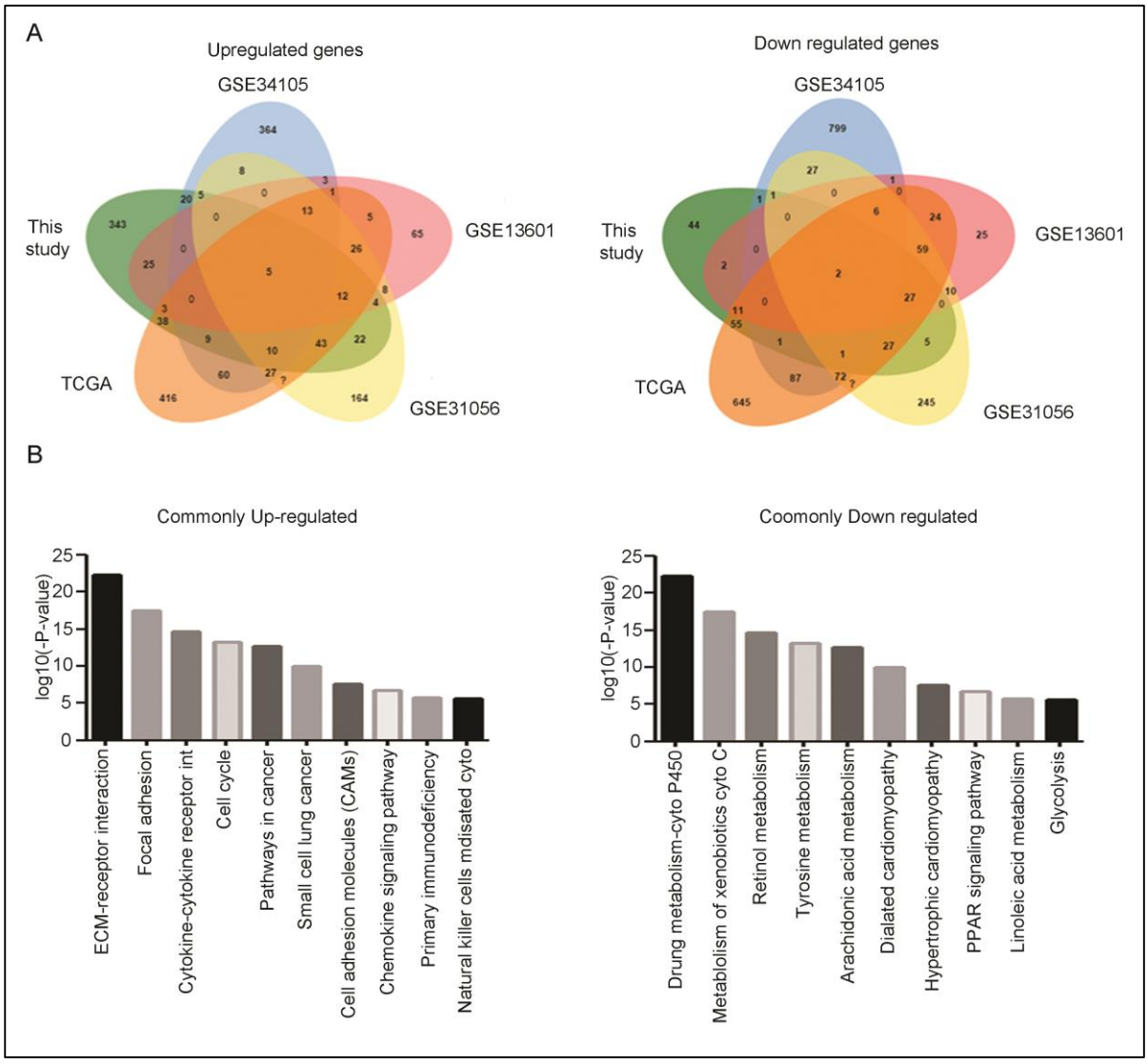
(detailed in the methods section) and fold change 1.5 and *P-value* ≤0.05 and fold change 1.5 was applied as a cutoff to identify the DEGs in each dataset. We identified an average 1281 (SD±719) genes to be significantly differentially expressed, where average 619 (SD±364) and 662 (SD± 447) genes showing up and down regulation, respectively (Table 6.2), after correcting for the variable number of samples used for each study.

XXV Table 6.2: Gene expression data sets used for the meta-analysis and statistics of differentially expressed genes in each data set.

S.No.	GEO ID	Samples (T/N)	Location	Platform	Total # differentially expressed genes	Up	Down	References
1	GSE34106	78 (62/16)	Sweden	Illumina HumanHT-12 WG-DASL V4.0 R2 expression beadchip	1523	998	525	Rentoft et al., 2012
2	GSE31056	47 (23/24)	USA	Affymetrix, HG-U133_Plus_2	1209	493	716	Reis et al., 2011
3	GSE13601	57 (31/26)	USA	Affymetrix, HG_U95Av2	338	171	167	Estilo et al., 2009
4	TCGA	139 (13/126)	USA	Illumina HiSeq1000	2055	814	1241	TCGA-HNSC, 2016
Mean number of genes					1281	619	662	

In overall, the average number of up-regulated genes were comparable with study based on our cohort (Table 6.1). To identify the commonly deregulated genes across dataset we performed recurrence based comparative analysis across dataset and observed 1146 genes to be deregulated in two or more number of datasets (Appendix V). Among the 1146 deregulated genes, 493 and 653 were showing common upregulation and down-regulation in ≥2 datasets (including this study) in the meta-analysis (Figure 6.2B, C). Interestingly, observed significant overlap i.e. 39% (196/493) up-regulated genes (*P-value*<0.0001); and 20% (133/653) down

regulated genes overlap (*P-value*<0.0001) with recurrently up-regulated genes in previous datasets.



XLII Figure 6.3: Commonly deregulated gene and pathways in tongue cancer.

(A) Venn diagram to illustrate the overlapped of significantly differentially expressed and commonly up (left panel) and down (right panel) regulated genes in at least two gene expression studies. **(B)** Bar plot representation of significantly deregulated gene sets in tongue cancer identified from meta-analysis.

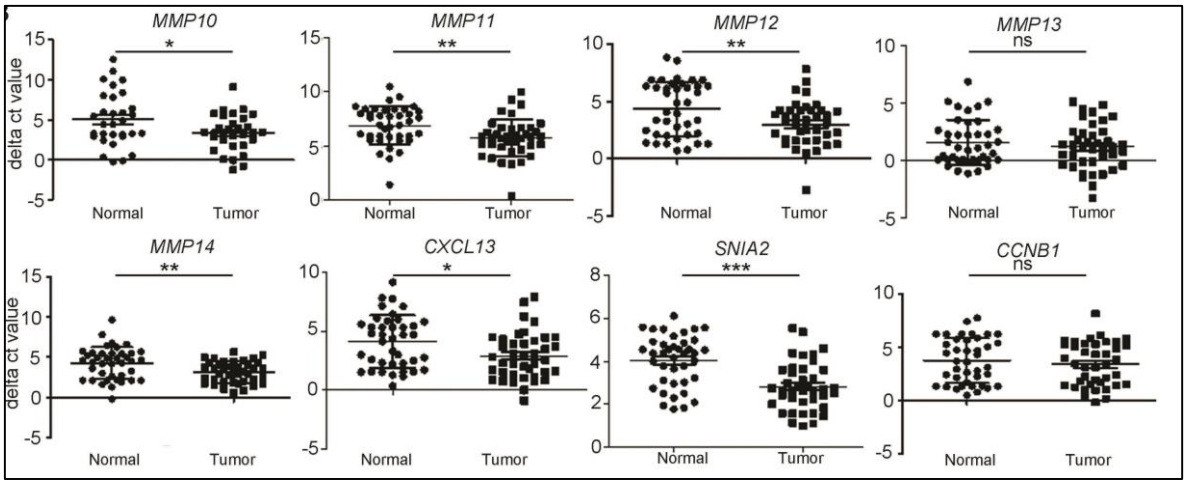
To gain the broader insight about biological processed related to the commonly DEGs in tongue cancer, we performed gene set enrichment analysis against KEGG gene sets using MSigDB [268]. For the up-regulated genes, significantly enriched KEGG gene sets includes several pathway involved in tumor cells metastasis process such as ECM-receptor interaction (adj P

value=8.21E-^{21}), Focal adhesion (adj P value=3.41E-^{18}), Cytokine-cytokine receptor interaction (adj P value=2.34E-^{15}), Cell adhesion molecules (CAMs) (adj P value=2.46E-^{08}) Chemokine signaling pathway (adj P value=1.58E-^{07}) and cell cycle, pathways in cancer, immune system related pathways, consistent with previous reports in HNSCC and tongue cancer [280, 340, 346] (Figure 6.3B). The down-regulated genes, significantly enriched KEGG gene sets includes pathways implicated in detoxification of carcinogenic compounds and environmental toxins such as Drug metabolism - cytochrome P450 (adj P value=3.9E-^{22}), Metabolism of xenobiotics by cytochrome P450 (adj P value=4.5E-^{18}), Retinal and tyrosine metabolism (adj P value=2.68E-^{13} and 1.16E-^{16}), Arachidonic acid metabolism (adj P value=2.76E-^{11}) consistent with previous reports in HNSCC, including tongue tumors [280, 340, 346] (Figure 6.3B). Interestingly, Arachidonic acid metabolism pathway was previously shown to be downregulated and inactivated via somatic mutations in Indian Gingivobuccal cancer patients, suggesting its possible tumor suppressive role via downregulation in tongue cancer patients in this study [357].

6.4.2 Upregulation of MMP10 and other MMPs in early stage tongue primary tumors

Several matrix metalloproteinase (MMPs) family genes were among the highly up-regulated genes across ≥3 dataset (Appendix V). We performed the qRT-PCR-based validation of eight genes which includes *MMP10, MMP11, MMP12, MMP13, MMP14, CXCL13, CCNB1, SNIA2* in 35 primary paired normal tongue tumors samples (Figure 6.4A). Overall, the qRT-PCR analysis revealed significant up-regulation of *MMP10, MMP11, MMP12, MMP14, CXCL13, SNIA2* in tongue tumor samples as compared to adjacent normal tissue samples and *MMP13, CCNB1* showed a trend towards up-regulation (Figure 6.4). Most of the genes such as *MMP10, MMP11, MMP12, MMP14, CXCL13, CCNB1, SNIA2* showed up-regulation in >40% of tongue cancer patients (Figure 6.2B). We also observed the considerable proportion of tongue tumors

showing the down regulation of *MMP13* (45%), *CCNB1* (40%), *MMP10* (34%) and *MMP12* (31%) (Figure 6.2B).



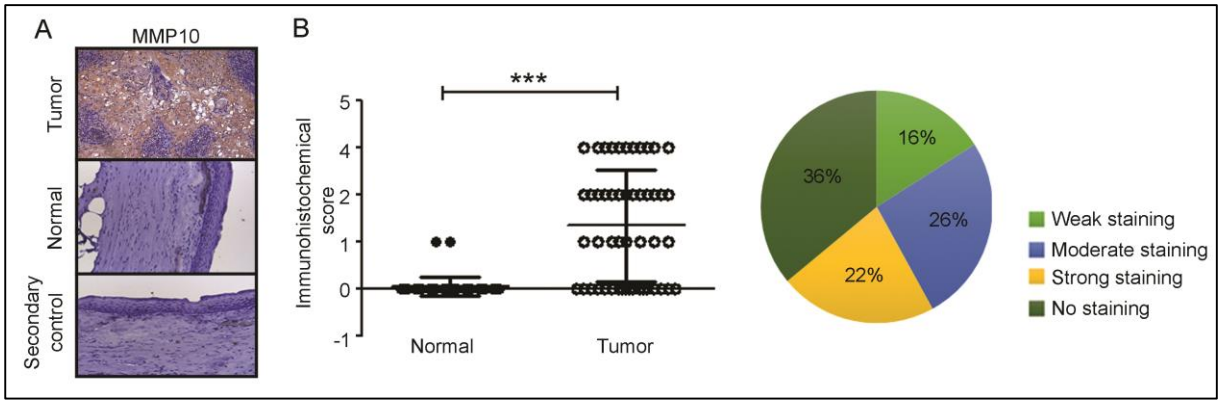
XLIII Figure 6.4: qRT-PCR validation of up-regulated gene in early stage tongue cancer.

qRT-PCR analysis of *MMP11, MMP12, MMP13, MMP14, CXCL13, CCNB1,* and *SNAI2* transcript expression in paired normal early tongue tumors (n=35). Dot plot representation of ΔCt value distribution and its significance between normal and tumors tongue tissue samples for *MMP10*. Each dot represents the average normalized ΔCt value of a gene in a single sample. Median with interquartile range is shown for each gene for normal and tumor samples. The P-value was calculated by Mann-Whitney *U test* using GraphPad Prism 5 program and p value ≤0.05 was considered as a threshold for statistical significance. P-value is denoted as *; P < 0.01, **; P < 0.001, ***; P < 0.0001.

Overall, the qRT-PCR analysis revealed significant up-regulation of *MMP10, MMP11, MMP12, MMP14, CXCL13, SNIA2* in tongue tumor samples as compared to adjacent normal tissue samples and *MMP13, CCNB1* showed a trend towards up-regulation (Figure 2A). Most of the genes such as *MMP10, MMP11, MMP12, MMP14, CXCL13, CCNB1, SNIA2* showed up-regulation in >40% of tongue cancer patients (Figure 6.4).

6.4.3 MMP10 protein overexpression in early stage primary TSCC tumors

Of the several MMPs found to be deregulated in this study, we performed immunohistochemical validation of MMP10 in 50 primary paired normal tongue tumors using analysis. In the adjacent normal samples, positive MMP10 staining was not observed, whereas positive cytoplasmic staining of MMP10 was detected in 32/50 (64%) of tongue cancer patients tumors, consistent with previously reported in HNSCC (Figure 6.5A) [346]. Overall, a statistically significant difference in immunohistochemical scores was observed in tumors as compared to adjacent normal tissues (P-value<0.0001, Unpaired student-t-test) (Figure 6.5B). About 48% of primary tongue tumors displayed strong or moderate immunostaining of MMP10 protein, whereas 62% tongue tumors showed weak or no staining (Figure 6.5C). Overall, MMP10 protein was significantly up-regulated in a large proportion of primary tongue tumors.



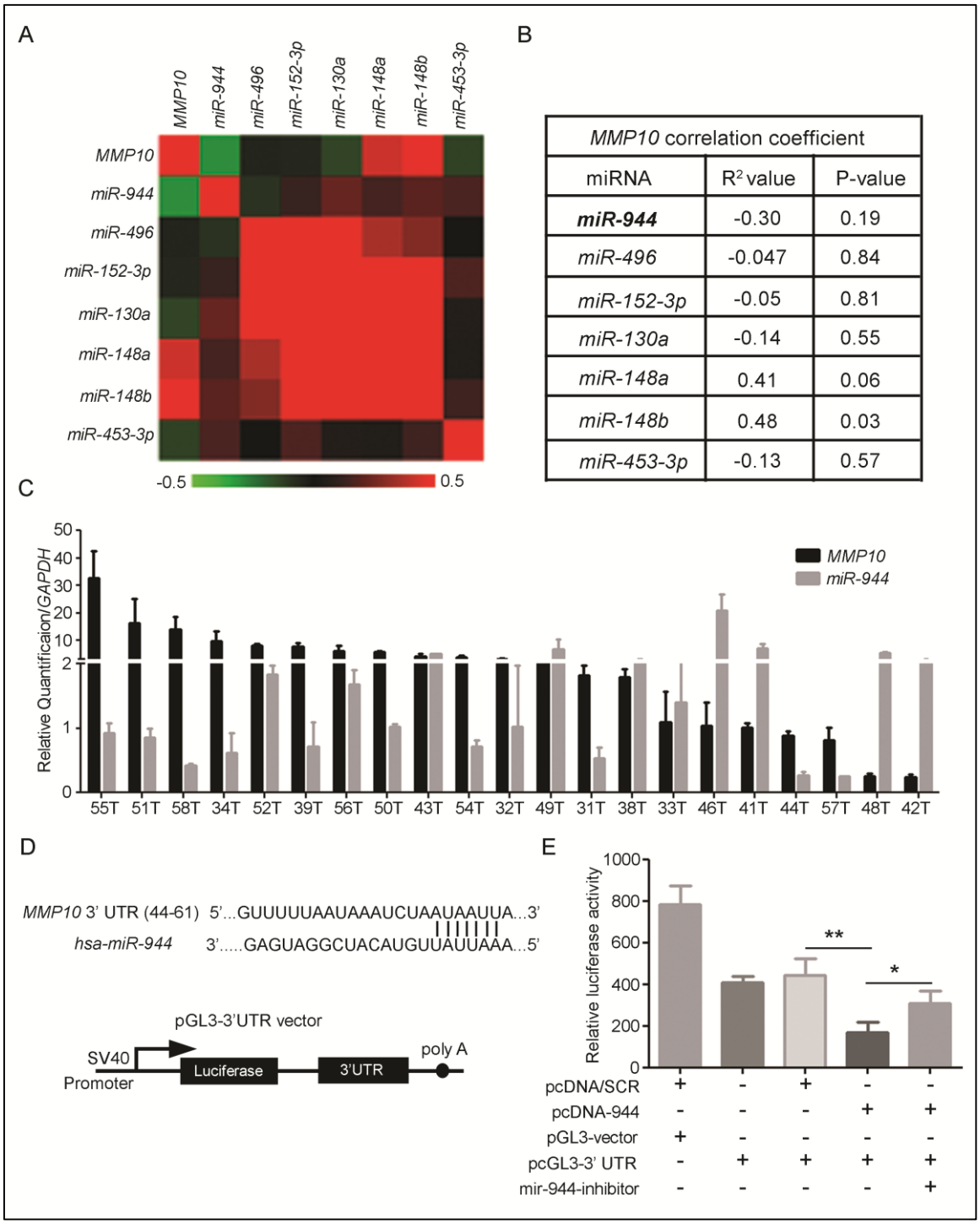
XLIV Figure 6.5: Analysis of MMP10 protein expression in early tongue squamous cell carcinoma patient samples (n=50).
(A) Representative IHC stained photomicrographs tongue tumors and paired normal samples are shown. The brown color indicates positive expression of MMP10 protein. **(B)** Dot plot representation of immunohistochemical score of MMP10 expression in tongue tumors and adjacent normal tissues (n=50). Each dot represents that final IHC score for each sample and median with interquartile range is shown. Median with interquartile range is shown for MMP10 protein expression in normal and tumor samples. **(C)** Pie chart representation of percent frequency distribution of various immunohistochemical scores of MMP10 protein such as strong, moderate, weak and no staining in early stage primary tongue tumors (n=50). The P-value was calculated by Mann-Whitney U test using GraphPad Prism 5 program and p value ≤0.05 was considered as a threshold for statistical significance. P value is denoted as ***; P < 0.0001.

6.4.4 *miR-944* targets 3'UTR of *MMP10* to regulate expression in tongue primary tumors

Since MMP10 is found to be frequently overexpressed in primary tumors at both transcript and protein level, we investigated miRNAs predicted to bind to 3'UTR of *MMP10* gene using 10 different miRNA binding site prediction tools such as MirWalk [353], miRanda [358], mirbridge [359], miRDB [360], miRMap [361], miRNAMap [362], Pictar2 [363], PITA [364], RNAhybrid [365] and Targetscan [366]. Bioinformatics prediction analysis revealed seven miRNAs (*miR-944, miR-496, miR-152-3p, miR-130a, miR-148a, miR-148b, miR-453-3p*) putatively targeting the 3'UTR of *MMP10* by at least 7 of 10 prediction algorithms and found to be down regulated in human cancers using miRCancer (Supplementary Table 4) [367]. Next, we performed qRT-PCR-based expression analysis of these 7 miRNA in 21 primary tongue tumor and adjacent normal samples. qRT-PCR analysis followed by correlation analysis between the relative fold change expression values revealed negative correlation between *MMP10* transcript and *miR-944* (R^2= -0.30), *miR-130a* (R^2= -0.14), and *miR-453-3p* (R^2= -0.13) (Figure 6.6A,B Figure 6.7).

To confirm whether *miR-944* directly targets the 3'UTR of *MMP10*, 3' UTR of *MMP10* gene was cloned downstream of the luciferase open reading frame (ORF) to construct a reporter plasmid pGL3-3'UTR-*MMP10* (Figure 6.6C&D) and *miR-944* was cloned in pcDNA-3.1 (-) downstream to CMV promoter. Transient co-transfection of 293FT cells with pc-DNA-miR-944, pGL3-3'UTR-*MMP10* reporter led to a significant decrease (*P-value* <0.01, Unpaired student-t-test) in the luciferase reporter activity as compared to control (Figure 6.7E), whereas a significant increase (*P-value* <0.05, Unpaired student-t-test) in luciferase activity was observed post co-transfection with anti-*miR-944* oligo (Figure 6.7E). These results support the

bioinformatics prediction and qRT-PCR data that the 3'UTR of *MMP10* might be a target for *miR-944*.

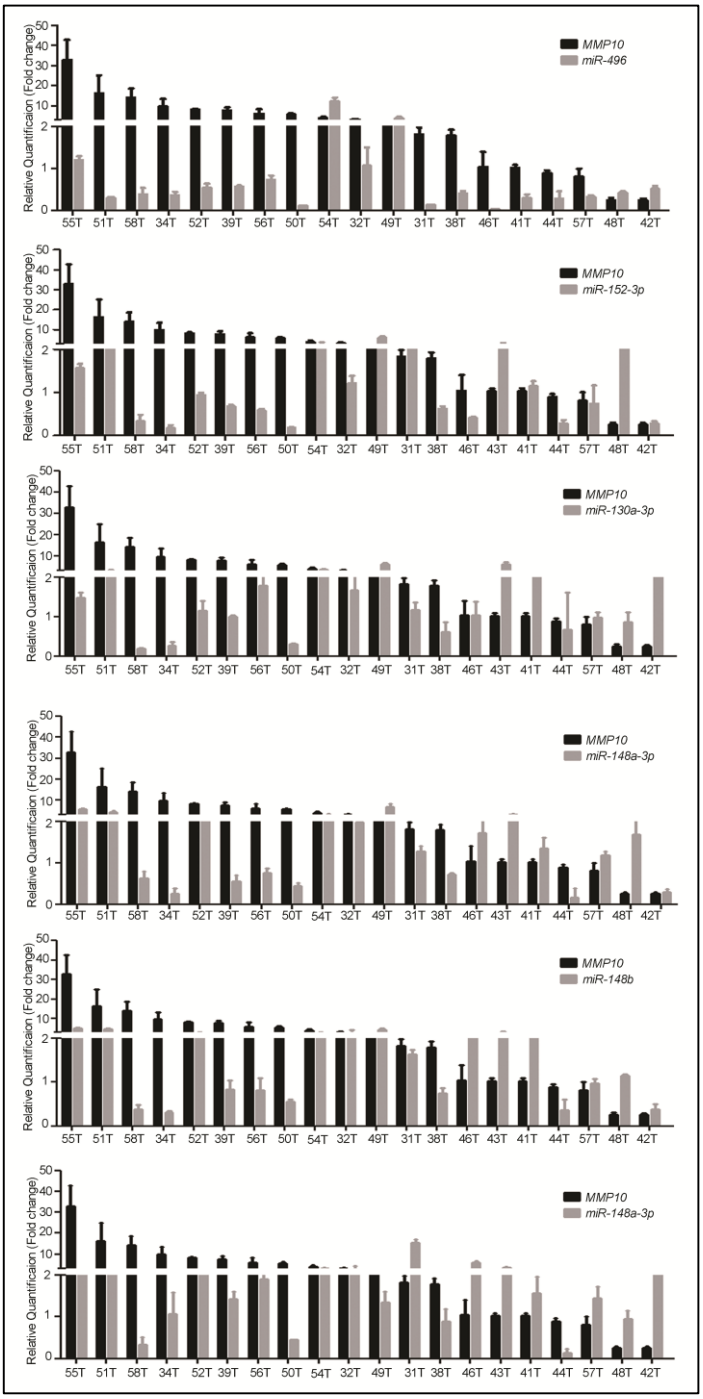


XLV Figure 6.6: Functional validation of miRNAs targeting *MMP10* 3' UTR using quantitative RT-PCR-based expression analysis and luciferase assay.

(A) Correlation analysis of *MMP10* and miRNAs targeting its 3' UTR from qRT-PCR data. **Left Panel:** Heatmap representation of the correlation values for *MMP10* gene and seven

miRNAs expression values which are predicted to be target the 3'UTR of *MMP10*. The qRT-PCR analysis was performed in paired normal tongue tumor tissue samples (n=20) and relative fold change was calculated post-normalization to an internal reference. Relative fold change values were used to perform correlation analysis using Pearson method in SPSS program. The color red and green denotes positive and negative correlation, respectively. **Right Panel:** A tabular representation of Pearson correlation values for miRNAs with *MMP10* transcript expression. The *miR-944* showing highest negative correlation is highlighted in bold. **(B)** The bar plot representation of *MMP10* transcript and *miR-944* expression levels in tongue tumor samples (n=21). **(C)** The regulation of *MMP10* gene by *miR-944* through the 3'-UTR. Schematic diagram of the *miR-944* predicted site on the 3'UTR region of *MMP10* gene as predicted by TargetScan which is an exact match to positions 2-8 of the mature miRNA followed by an 'A'. **(D)** The 3'UTR region of *MMP10* gene was cloned downstream of the luciferase open reading frame of the pGL3-control plasmid and *miR-944* was cloned into pcDNA3.1 (-) plasmid in the downstream of CMV promoter. **(E)** Bar plot representation of relative luciferase activity. Luciferase assay was performed in 293FT cells by co-transfecting with pre-*miR-944*/scrambled/ anti-*miR944* inhibitor and pGL3-*MMP10*-3'UTR with the putative *miR-944* binding site. The relative luciferase activity was measured by normalizing with the Renilla luciferase and results are presented as mean± SD. The P-value was calculated by Unpaired student-t-test using GraphPad Prism 5 program and p value ≤ 0.05 was considered as a threshold for statistical significance. P value is denoted as *; $P < 0.01$, ***; $P < 0.0001$.

XLVI Figure 6.7: qRT-PCR analysis of other miRNAs targeting *MMP10* gene in tongue tumors.



The bar plot representation of *MMP10* transcript and six miRNA expression relative fold change expression in tongue tumor samples (n=20). The data was normalized against internal reference control for genes and miRNA *GAPDH* and *U6* gene, respectively and relative fold change was obtained by comparing the expression in corresponding normal sample for each patient.

6.5 Discussion

We show that integrated meta-analysis approach lead to an increased reliability of gene expression signatures across different expression analysis platform and precise estimation of recurrently expressed genes across data set. We find, 8 MMP family members (*MMP1, MMP7, MMP9, MMP10, MMP11, MMP12, MMP13, MMP14*) were among the up-regulated genes across ≥3 of 5 data sets analyzed suggesting the prevalence of biological process underlying the tumor progression in tongue cancer [368, 369]. The overexpression of MMPs has been known to be involved in ECM degradation thereby facilitating the process of

tumor invasion and metastasis leading to an aggressive course of disease in HNSCC patients [369]. qRT-PCR validation across 35 paired normal early stage primary tumors for up-regulated genes (*MMP10, MMP11, MMP12, MMP14, CXCL13, CCNB1, SNIA2*) showed significant up-regulation in tumors suggesting reliability of genes identified from this study. Of the multiple MMPs found to be up-regulated, immunohistochemical analysis of MMP10 in 50 paired normal early stage primary tumors showed significant up-regulation of protein expression in primary tumors owing its possible role in early stage progression as described in other cancer types [370-373].

Further, *in-silico* prediction of the miRNAs binding site in 3' UTR of *MMP10* revealed seven miRNAs binding sites and qRT-PCR based analysis of seven miRNAs across 21 paired normal early stage primary tumors indicated *miR-944* as potential candidate miRNAs showing highest negative correlation with *MMP10* transcript expression. We further confirmed *MMP10* as a direct target of *miR-944* by performing luciferase reporter assay in presence of *miR-944* and *miR-944* inhibitors. We, for the first time provide evidence of novel miRNA (*miR-944*) regulating the 3' UTR of *MMP10* gene.

A unique feature of TSCC from other subsites in oral cancer is that about 27-40% of patients even at an early stage (pT1 or pT2) have nodal metastasis and may be undergoing a neck dissection which further adds to morbidity and worse survival due to disease recurrence [44, 175-177]. Although, clinically, the poor prognostic indicates for TSCC such as occult node positivity, tumor depth, lymphovascular invasion and perineural invasion is well defined, there is still an unmet need for reliable and robust prognostic biomarkers in early stage TSCC to stratify the patients who are likely to have an adverse clinical outcome [44, 374]. The identification of *MMP10* 3' UTR regulation via *miR-944* and its underlying role in metastases provides an opportunity for the future investigation of its specific role in TSCC development.

The detailed functional analysis and validation of *MMP10* and *miR-944* in early tongue cancer could provide an avenue for the development of molecular predictor of pathological lymph node status, whereby a considerable fraction can be spared unnecessary surgery lessening morbidity and cost of treatment.

Chapter 7

Summary and Conclusions

7. Chapter 7. Summary and Conclusions

HNSCC represents a highly heterogeneous group of cancers and has a most complex underlying molecular profile. In the current genome sequencing era, large-scale genome profiling efforts such as The Cancer Genome Atlas (TCGA) and International Cancer Genomics Consortium (ICGC) has presented the first comprehensive glance of the underlying genomic complexity of advanced stage HNSCCs genome using massively parallel genome sequencing technologies in an unbiased manner. Findings from ICGC, TCGA, and other individual efforts not only uncovered several attractive molecular alterations which currently is being tested for translation in the clinics, but also revealed emerging key challenges in translation due to diversity among the different populations in terms of clinical, etiological and biological features of tumors in various ethnicity. Due to the difference in associated etiological factors such as tobacco and betel nut chewing in the Indian population, HNSCC tumors molecular features are expected to differ between ethnicity. This led to the origin of my thesis work. The major focus of my thesis involves the understanding the portrait of genomic alteration landscape of tongue squamous cell carcinoma patient's tumors from Indian origin.

As a first step, I profiled the underlying genomic alterations of primary tumor derived oral cancer cell lines from Indian patients. The major goal of this part of the thesis to uncover the landscape of molecular aberration's in cancer cell line genome and to optimize the integrated genomic analysis workflow to identify the biologically relevant alterations from a fewer number of samples by using posterior filtering approach. The results from functional analysis indicate mutant *NRBP1* as a novel oncogene in HNSCC cells.

Secondly, I developed the first Indian specific novel germline SNP database: TMC-SNPdb derived from 62 whole exome sequencing data of normal samples from Indian origin cancer

patients. The application TMC-SNPdb in somatic cancer genome analysis efficiently filters ethnic specific low allele frequency germline variants over and above dbSNP filtering.

In the third section of my thesis, genomic characterization of landscape of somatic alterations underlying tobacco/ nut chewing HPV-negative early tongue primary tumors using whole exome and transcriptome sequencing. Here I identified and validated several known and novel recurrent fusion transcript in early stage HPV-negative primary tumors along with the identification of previously known molecular alterations in known hallmark genes. Additionally the alterations in Notch pathway genes and deregulation of EMT processes related genes were identified and were further validated in my fourth and fifth part of the thesis.

Forth, I undertook a genomic and functional genomic approach and present a pro-oncogenic role of NOTCH1 in early stage tongue squamous cells carcinoma and required for the maintenance of cancer stem-like cell populations in oral cancer cells, unlike previously known tumor suppressive role in HNSCC. I anticipate that these findings could form the basis for the therapeutic targeting of NOTCH1 in early tongue cancer.

Last part of my thesis describes identification and validation of *MMP10* expression in early TSCC using transcriptome sequencing and meta-analysis of previous gene expression profile data sets. My work in this part also identifies the first miRNA (*miR-944*) targeting 3' UTR of *MMP10*.

7.1 Integrated genomic characterization of HNSCC cell lines derived from Indian patients tumors

The genomically defined patient-derived cancer cell line model system serve as a pre-clinical model system and provide an improved understanding of biological features and a more rational approach to the development of therapy. There have been several consortium based

efforts such as Cancer cell line encyclopedia (CCLE: http://www.broadinstitute.org/ccle), which systematically profiled the spectrum of driver genomic alteration in cancer cell lines using genomic and functional approaches using genome-wide RNAi screen and pharmacological compound libraries across several cancer types [216]. The studies revealed that cancer cell lines represents much of the tissue-type and displayed genetic diversity known in human cancers [375]. Additionally, genomically defined cancer cell lines derived from HNSCC tumor of Indian patients is lacking and a thorough understanding of underlying genomic alterations of previously established cell lines allows the researcher to uncover new biologically and therapeutically relevant alterations in HNSCC.

Here I performed the whole exome (mutation), transcriptome (gene expression), SNP array (copy number alterations) and classical karyotyping of four HNSCC cell lines (AW13516, AW8507, NT8e, and OT9) previously established from the primary tumors derived from Indian HNSCC patients. The overall genomic profile of cell lines was similar to primary tumors. The application posterior filtering approach and integration of multiple platform with the copy number variation, allowed us identify alterations in HNSCC hallmark genes (*PIK3CA*, *EGFR*, *HRAS*, *MYC*, *CDKN2A*, *MET*, *TRAF2*, *PTK2* and *CASP8)*. Importantly, I also noted amplification of *NOTCH1* in 3/4 HNSCC cell lines, as oppose to previously described frequent deletion of *NOTCH1* in HNSCC. I also identified several genomic alterations in biologically relevant novel gene (*CLK2*, *NRBP1*, *CCNDBP1*, *IDH1*, *LAMA5*, *BCAR1*, and *ZNF678)* in HNSCC cells. Of these , *NRBP1*(Q73*) (NT8e cells) was a potentially interesting gene due to its previous report in lung and other cancers and 9% cumulative alteration frequency in TCGA-HNSCC dataset [216, 217].

Next, I performed the first functional analysis of mutant *NRBP1* (Q73*) in NIH-3T3 and HNSCC cancer cell lines. Overexpression of mutant *NRBP1* (Q73*) in NIH-3T3, but not wild

type, leads to increased anchorage-independent growth via MAP kinase activation, while knockdown in HNSCC cells (NT8e cells) having mutant *NRBP1* (Q73*) leads to diminishing survival and anchorage-independent growth consistent with its known role in prostate cancer and another model organism [218, 226, 227]. Moreover, NRBP1 has been shown to be involved in the maintenance of cellular homeostasis in mouse intestinal progenitor cells and have tumor suppressive function [229], indicating its context-dependent role in various cancer. Results from this study indicate the oncogenic role of mutant *NRBP1* (Q73*) in HNSCC cells, however, in-depth systematic functional analysis of *NRBP1* in HNSCC cells and primary tumors is required to establish its specific role in HNSCC.

Additionally, we observed large number of novel variants in HNSCC cells despite depletion with dbSNP which is possibly due to lack paired normal blood samples or abundance of low allele frequency ethnic specific SNP.

7.2 Development of Indian germline variant database from whole exome sequences

As I observed very high number of novel variants in HNSCC cells variant analysis, I wanted to develop the ethnic specific SNP database to deplete these low allele frequency SNPs in order to identify the bona fide somatic variants while analysing the tongue cancer patients tumors exome data.

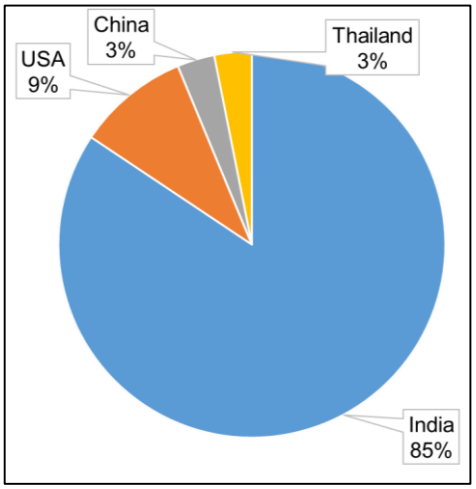
For the identification of somatic variants in tumors, a typical cancer analysis involves subtraction of matched normal DNA derived variants from tumor derived variants from the same individual, followed by depletion of residual tumor-specific variants from the public SNP databases such as dbSNP and 1000 Genomes. Post this step, the remaining variants are considered as somatic in nature. Moreover, adopting such filtering approach depletes high-

frequency variants, while unknown SNPs those with minor allele frequency from the population which is not adequately respected in dbSNP and 1000 Genomes are likely to confound the somatic mutation analysis in studies from non-Caucasian and non-European Caucasian populations (5). To fill this gap of low allele frequency SNPs in public databases there have been several global (NHLBI Exome Sequencing Project (ESP) and Exome Aggregation Consortium (ExAC) [110]) and ethnic specific (Indian Genome Variation Consortium [112, 113] and HUGO Pan- Asian SNP Consortium [114]) initiatives worldwide. A concerted effort to identify and catalog the novel SNPs present in Indian population was lacked, posing a challenge in the identification of bona fide somatic mutation.

I developed the Tata Memorial Centre-SNP database "TMC-SNPdb" as the first, open source, freely available database of 114,309 unique germline variants obtained from whole exome data of 62 'normal' samples from cancers patients of Indian origin. TMC-SNPdb is also provided with a companion tool with command line and graphical user interface (GUI) for non-computational biologists. The application of TMC-SNPdb in cancer somatic variant analysis significantly depletes in low allele frequency false positive somatic variants over and above dbSNP and 1000 Genomes across 132 whole exome sequencing data of 3 tumors types. The availability of companion tool provides easy expandability of TMC-SNPdb in future by adding more number of normal samples. There are two limitations of TMC-SNPdb. First; the presumption of samples derived from cancer patients are normal and over subtraction of cancer predisposing variants due to depletion with COSMICdb. These two limitations limit its application in cancer predisposition studies.

In spite of these caveats, I believe that TMC-SNPdb is a step towards fulfilling the significant unmet need for an Indian population specific 'normal' variant database and would be helpful in depleting false positive somatic variants in cancer studies involving paired normal as well as orphan tumor sequencing. Since TMC-SNPdb is a pilot initiative and expected to grow via

usages of easy to use companion tool in future, as more number of normal sequence data is available from Indian population. As on 15th Nov 2016, TMC-SNPdb has been downloaded and used in 18 institutes across 4 different countries (Figure 7.1).



XLVII Figure 7.1: TMC-SNPdb usage statistics (as on 15th Nov 2016)

7.3 Integrated analysis of tobacco/ nut chewing HPV-negative early tongue cancer tumors identifies recurrent transcript fusions

Recent large-scale genomic studies defined the landscape of HNSCC includes International Cancer Genome Consortium (ICGC-India) for gingiva-buccal tumors and individual group efforts for tongue cancer revealed a frequent mutation in major hallmark genes; *TP53*, *CDKN2A*, *FAT1*, *PIK3CA*, *NOTCH1*, *KMT2D*, and *NSD1*. While in TSCC genome characterization most of the studies were limited to candidate genes or gene panels along with few whole exome studies from Asia and India suggested a difference in genomic profile, indicating unique molecular features associated with TSCC tumors among oral cancer subsites [181, 182]. Most of the genomic analysis studies in TSCC have been restricted to advanced stage samples (pT3-pT4) while underlying genomic alterations of the HPV-negative early stage (pT1-pT2) tongue tumor genome remained unexplored.

Here, I present the portrait of somatic alterations in HPV-negative early tongue cancer using high throughput sequencing of fifty-four samples derived from HPV-negative early stage

tongue cancer patients habitual of chewing betel nuts, areca nuts, lime or tobacco using whole exome (n=47) and transcriptome (n=17) sequencing. The mutational profile of 53% TSCC patient's tumors displayed tobacco-associated signatures, consistent with predominant tobacco chewing habit in our TSCC cohort. Somatic mutation analysis revealed the mutations in hallmark genes such as *TP53, NOTCH1, CDKN2A, HRAS, USP6, PIK3CA, CASP8, FAT1, APC,* and *JAK1*. The unique genetic association was observed in early stage TSCC tumors, where, *EGFR* amplification was mutually exclusive to 11q13.3 (*CCND1, FGF19, ORAOV1, FADD*) amplification, as reported in other HNSCC subtypes. These observations were also verified in TCGA-tongue tumor cohort (n=79).

Importantly, this study presents the first glance of a portrait of 242 tumor specific transcript fusions, followed by exhaustive validation of 12 candidate fusion transcripts (including the discovery of 5 novel fusion transcripts) across 44 paired HPV-negative early TSCC tumor samples and 4 HNSCC cell lines. Comparative analysis our data with various fusion database revealed 48 previously described transcript fusion in various cancer types. Moreover, I identified and validated five novel somatic recurrent fusion transcripts: *LRP5-UBE3C (15%), YIF1A-RCOR2 (13%), POLA2-CDC42EP2 (8%), SLC39A1-CRTC2 (4%), and BACH1-GRIK1 (2%)* in tumor samples. *LRP5-UBE3C* fusion transcript was recurrent in 9/48 samples and fusion involves the exon 1 of *LRP5* and exon 6 of *UBE3C* which affects extracellular epidermal growth factor-1 repeat (EGF-1) and HECT domain, respectively. A truncated version of *LRP5* responsible for the constitutive active Wnt signaling in parathyroid tumor and breast cancers and shown to a potential therapeutic target in those tumors [291, 376]. Second potentially interesting validated recurrent transcript fusions is *YIF1A-RCOR2*, involving the first exon of *YIP1A* and sixth exon of *RCOR2* affecting the SANT domain. The SANT domain *RCOR2* has been shown to require for complex formation with Histone demethylase Lysine-

specific demethylase *LSD1* and mediate hedgehog signaling pathway [292]. Further characterization is warranted to elucidate the biological significance of this novel validated fusion transcripts. Recurrent transcript fusions described here could serve as an attractive candidate to facilitate in diagnosis of HPV-negative early stage TSCC patients.

Overall, this study provides first systematic analysis of genomic alterations such as somatic mutation, copy number alteration and discovery of novel transcript fusion in tobacco/ nut chewing HPV-negative early TSCC patients of Indian origin.

Further, two very important observation were made from integrated genomic analysis, First; unlike previous report inactivating mutations in HNSCC, we observed Notch pathway genes harboring missense mutations, similar to those reported in leukemia and recent report from Asian population, Second; EMT-related processes related genes were upregulated in early stage tongue tumors. Hence I further explored for these two observations.

7.4 Pro-oncogenic role of *NOTCH1* in early stage tongue cancer is required for the maintenance of cancer stem-like population

Recent large-scale genomic characterization studies in HNSCC revealed inactivating alterations such as mutation and deletions in *NOTCH1* receptor suggesting a complex role of *NOTCH1* as candidate tumor suppressor. Few studies also noted the activating role of *NOTCH1* via amplification, overexpression and prevalence of missense mutation in HNSCC highlighting the context-dependent role of *NOTCH1* in HNSCC tumors. Moreover, during HNSCC cell line characterization work, *NOTCH1* was found to be amplified in 3/4 HNSCC cells, suggesting a possible oncogenic role in India origin tumors.

So to further verify the observation made in HNSCC cell lines, here, I undertook an integrated genomic approach to investigate the underlying genomic alterations in early tongue tumors by analyzing the whole exome and transcriptome sequencing data in Notch pathway (n=48 gene) including Notch receptor, ligands, target genes, and regulators. Unlike previously reported an inactivating mutation in Notch pathway gene, including *NOTCH1*, I observed somatic amplification, overexpression in early stage tongue tumors and further verified in large data set i.e., TCGA-tongue cancer (n=126). Immunohistochemically, 40% of primary tumors displayed activation of NOTCH1 (NICD) as compared to adjacent normal counterpart suggesting active signaling in early stage tongue tumor cells.

Next, functional analysis of *NOTCH1* in HNSCC cells harboring activation of Notch pathway was performed using genetic and chemical perturbation approach. Interestingly, shRNA hairpin-mediated knockdown or chemical inhibition of NOTCH1 signaling led to significant reduction in soft agar colony formation, migration potential and cancer stem-like features of HNSCC cells, highlighting their dependency on *NOTCH1* expression. Notch signaling has been previously shown to be implicated in the maintenance of cancer stem-like cell in several cancer types [133, 295, 297, 331, 332]. Findings from this study are consistent with a recent report in HNSCC, suggesting reproducibility of results [333, 334]. Thus, the targeted elimination of these cells would be an attractive area of research for developing targeted therapy in TSCC.

Clinically, activated NOTCH1 overexpression was associated with lymph node metastasis, the non-smoking habit of early TSCC patients highlighting activated NOTCH1 as a reliable marker to predict lymph node metastasis. There was a trend towards poor disease-free survival was observed in patients with activated NOTCH1, however, this observation could not reach statistical significance due to small samples size and awaits for verification in a large number of samples.

In summary, results from this part of thesis work strongly indicate the involvement of NOCTH1 signaling in TSCC tumors and might be required for the maintenance of cancer stem-like the population in these tumors, and that, *NOTCH1* could be a potential therapeutic target in early stage TSCC patients.

7.5 *MMP10* as a novel prognostic marker and target of *miR-944* in early stage tongue cancer

Several studies have been carried out on advanced stage TSCC primary tumor revealed the global gene expression profile using microarray approach. Among the key candidate genes highlighted from these studies includes *CDK1, NDRG1, ADAM15, CDC7, MMP9, TNFRSF8, NPM* and *CHES1* as a biomarker of OSCC progression [336, 337]. One of the major highlights from these reports includes up-regulation of gene related to epithelial to mesenchymal (EMT) processes such as matrix metalloproteinase (MMPs) family proteins and suggested as a robust biomarker in HNSCC and tongue cancer [346, 348, 349, 369]. However, a major challenge in the identification of reliable and robust gene expression based biomarkers requires the increased reproducibility of gene signatures across different studies in tongue cancer in the different population. The possible reasons behind this variability are due to biological, technical and experimental differences across different studies posing a great difficulty in translation of these biomarkers for the clinical benefit [343, 344].

To overcome these challenges the meta-analysis of multiple gene expression datasets seems to be a valid approach. Application of such meta-analysis has been shown to have a potential to deliver a robust and reliable candidate biomarkers in the prostate, breast cancer, ovarian and lung cancer [349, 377-380]. To perform the similar studies in TSCC, I performed whole transcriptome sequencing of early tongue primary tumors (n=11) and normal samples (n=4) was carried out which identified 739 differentially expressed genes (DEGs) (561 up-regulated

and 178 down-regulated genes) in tongue tumor samples. A meta-analysis of previously published 4 TSCC expression data set including TCGA-tongue (total 243 tumors and 79 adjacent normal samples) was performed to identify the differentially expressed genes in individual studies. Analysis of commonly deregulated gene revealed several previously known biological processes in tongue cancer including up-regulation of epithelial to mesenchymal transition (EMT) processes, extracellular remodeling pathways and downregulation of metabolism related pathways. Moreover, eight members of matrix metalloproteinase family were showing consistent up-regulation across 3/5 gene expression data set. The qRT-PCR based validation of 8 genes *(MMP10, MMP11, MMP12, MMP14, CXCL13, CCNB1, SNAI2*) was performed across 35 paired normal early stage TSCC tumors revealed significant up-regulation in primary tumors as compared to normal tissue. I further explored the expression of MMP10 gene because the role of MMP10 was not explored in TSCC. Immunohistochemical analysis across 50 paired normal early stage TSCC tumors indicated significant overexpression of MMP10 protein in tumors. Since there was no report describing the miRNAs regulating the MMP10 expression, I performed the investigation and present *miR-944* as first validated miRNA controlling the expression of *MMP10* gene via binding to its 3' UTR using qRT-PCR analysis in paired normal primary TSCC tumors and luciferase assay in cells.

I hope that the validation in large number patients samples coupled with functional analysis would provide more detailed information regarding the specific role of the *miR-944-MMP10* axis in early stage TSCC for designing the reliable molecular biomarkers in future.

Taken together, this thesis work presents a comprehensive landscape of known and several novel genomic and transcriptomic alterations in early stage tongue squamous cell carcinoma using integrated genomics approach. Additionally, I present the first normal SNP database from Indian patients and shown its significant applicability in somatic variant analysis, and

expected to grow further by addition of more number of normal samples. Alterations in *NOTCH1*, *MMP10* and discovery of novel transcript fusions from this thesis could be explored as a pathognomonic candidate in early TSCC patients in future.

8. Chapter 8. Bibliography

1. Rothenberg, S.M. and L.W. Ellisen, *The molecular pathogenesis of head and neck squamous cell carcinoma.* J Clin Invest, 2012. **122**(6): p. 1951-7.

2. Leemans, C.R., B.J. Braakhuis, and R.H. Brakenhoff, *The molecular biology of head and neck cancer.* Nat Rev Cancer, 2011. **11**(1): p. 9-22.

3. Patel, S.G. and J.P. Shah, *TNM staging of cancers of the head and neck: striving for uniformity among diversity.* CA Cancer J Clin, 2005. **55**(4): p. 242-58; quiz 261-2, 264.

4. Gillison, M.L., et al., *Evidence for a causal association between human papillomavirus and a subset of head and neck cancers.* J Natl Cancer Inst, 2000. **92**(9): p. 709-20.

5. Hashibe, M., et al., *Alcohol drinking in never users of tobacco, cigarette smoking in never drinkers, and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium.* J Natl Cancer Inst, 2007. **99**(10): p. 777-89.

6. Hashibe, M., et al., *Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium.* Cancer Epidemiol Biomarkers Prev, 2009. **18**(2): p. 541-50.

7. Wynder, E.L., I.J. Bross, and R.M. Feldman, *A study of the etiological factors in cancer of the mouth.* Cancer, 1957. **10**(6): p. 1300-23.

8. Kutler, D.I., et al., *High incidence of head and neck squamous cell carcinoma in patients with Fanconi anemia.* Arch Otolaryngol Head Neck Surg, 2003. **129**(1): p. 106-12.

9. Baez, A., *Genetic and environmental factors in head and neck cancer genesis.* J Environ Sci Health C Environ Carcinog Ecotoxicol Rev, 2008. **26**(2): p. 174-200.

10. Lacko, M., et al., *Genetic susceptibility to head and neck squamous cell carcinoma.* Int J Radiat Oncol Biol Phys, 2014. **89**(1): p. 38-48.

11. Jefferies, S., et al., *The role of genetic factors in predisposition to squamous cell cancer of the head and neck.* Br J Cancer, 1999. **79**(5-6): p. 865-7.

12. Chaturvedi, A.K., et al., *Incidence trends for human papillomavirus-related and -unrelated oral squamous cell carcinomas in the United States.* J Clin Oncol, 2008. **26**(4): p. 612-9.

13. Michmerhuizen, N.L., et al., *Genetic determinants in head and neck squamous cell carcinoma and their influence on global personalized medicine.* Genes Cancer, 2016. **7**(5-6): p. 182-200.

14. Gupta, B. and N.W. Johnson, *Systematic review and meta-analysis of association of smokeless tobacco and of betel quid without tobacco with incidence of oral cancer in South Asia and the Pacific.* PLoS One, 2014. **9**(11): p. e113385.

15. Patel, S.C., et al., *Increasing incidence of oral tongue squamous cell carcinoma in young white women, age 18 to 44 years.* J Clin Oncol, 2011. **29**(11): p. 1488-94.

16. Majchrzak, E., et al., *Oral cavity and oropharyngeal squamous cell carcinoma in young adults: a review of the literature.* Radiol Oncol, 2014. **48**(1): p. 1-10.

17. Pulte, D. and H. Brenner, *Changes in survival in head and neck cancers in the late 20th and early 21st century: a period analysis.* Oncologist, 2010. **15**(9): p. 994-1001.

18. Das, L.C., et al., *Comparison of outcomes of locoregionally advanced oropharyngeal and non-oropharyngeal squamous cell carcinoma over two decades.* Ann Oncol, 2015. **26**(1): p. 198-205.

19. Conley, B.A., *Treatment of advanced head and neck cancer: what lessons have we learned?* J Clin Oncol, 2006. **24**(7): p. 1023-5.

20. Shimkhada, R. and J.W. Peabody, *Tobacco control in India.* Bull World Health Organ, 2003. **81**(1): p. 48-52.

21. Braakhuis, B.J., C.R. Leemans, and R.H. Brakenhoff, *Expanding fields of genetically altered cells in head and neck squamous carcinogenesis.* Semin Cancer Biol, 2005. **15**(2): p. 113-20.

22. Braakhuis, B.J., et al., *A genetic explanation of Slaughter's concept of field cancerization: evidence and clinical implications.* Cancer Res, 2003. **63**(8): p. 1727-30.

23. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.

24. Vogelstein, B., D. Lane, and A.J. Levine, *Surfing the p53 network.* Nature, 2000. **408**(6810): p. 307-10.

25. Gasco, M. and T. Crook, *The p53 network in head and neck cancer.* Oral Oncol, 2003. **39**(3): p. 222-31.

26. Olshan, A.F., et al., *p53 mutations in head and neck cancer: new data and evaluation of mutational spectra.* Cancer Epidemiol Biomarkers Prev, 1997. **6**(7): p. 499-504.

27. Scheffner, M., et al., *The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53.* Cell, 1990. **63**(6): p. 1129-36.

28. Kusama, K., et al., *p53 gene alterations and p53 protein in oral epithelial dysplasia and squamous cell carcinoma.* J Pathol, 1996. **178**(4): p. 415-21.

29. Qin, G.Z., et al., *A high prevalence of p53 mutations in pre-malignant oral erythroplakia.* Int J Cancer, 1999. **80**(3): p. 345-8.

30. Poeta, M.L., et al., *TP53 mutations and survival in squamous-cell carcinoma of the head and neck.* N Engl J Med, 2007. **357**(25): p. 2552-61.

31. Perez-Sayans, M., et al., *p16(INK4a)/CDKN2 expression and its relationship with oral squamous cell carcinoma is our current knowledge enough?* Cancer Lett, 2011. **306**(2): p. 134-41.

32. Smeets, S.J., et al., *Genome-wide DNA copy number alterations in head and neck squamous cell carcinomas with or without oncogene-expressing human papillomavirus.* Oncogene, 2006. **25**(17): p. 2558-64.

33. Smeets, S.J., et al., *Immortalization of oral keratinocytes by functional inactivation of the p53 and pRb pathways.* Int J Cancer, 2011. **128**(7): p. 1596-605.

34. Sheu, J.J., et al., *Functional genomic analysis identified epidermal growth factor receptor activation as the most common genetic event in oral squamous cell carcinoma.* Cancer Res, 2009. **69**(6): p. 2568-76.

35. Temam, S., et al., *Epidermal growth factor receptor copy number alterations correlate with poor clinical outcome in patients with head and neck squamous cancer.* J Clin Oncol, 2007. **25**(16): p. 2164-70.

36. Qiu, W., et al., *PIK3CA mutations in head and neck squamous cell carcinoma.* Clin Cancer Res, 2006. **12**(5): p. 1441-6.

37. Murugan, A.K., et al., *Oncogenic mutations of the PIK3CA gene in head and neck squamous cell carcinomas.* Int J Oncol, 2008. **32**(1): p. 101-11.

38. Okami, K., et al., *Analysis of PTEN/MMAC1 alterations in aerodigestive tract tumors.* Cancer Res, 1998. **58**(3): p. 509-11.

39. Suh, Y., et al., *Clinical update on cancer: molecular oncology of head and neck cancer.* Cell Death Dis, 2014. **5**: p. e1018.

40. Takes, R.P., et al., *Future of the TNM classification and staging system in head and neck cancer.* Head Neck, 2010. **32**(12): p. 1693-711.

41. Edge, S.B. and C.C. Compton, *The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM.* Ann Surg Oncol, 2010. **17**(6): p. 1471-4.

42. Gleich, L.L., et al., *Therapeutic decision making in stages III and IV head and neck squamous cell carcinoma.* Arch Otolaryngol Head Neck Surg, 2003. **129**(1): p. 26-35.

43. Jones, A.S., et al., *The treatment of early laryngeal cancers (T1-T2 N0): surgery or irradiation?* Head Neck, 2004. **26**(2): p. 127-35.

44. D'Cruz, A.K., et al., *Elective versus Therapeutic Neck Dissection in Node-Negative Oral Cancer.* N Engl J Med, 2015. **373**(6): p. 521-9.

45. Bernier, J., et al., *Postoperative irradiation with or without concomitant chemotherapy for locally advanced head and neck cancer.* N Engl J Med, 2004. **350**(19): p. 1945-52.

46. Tuljapurkar, V., et al., *The Indian scenario of head and neck oncology - Challenging the dogmas.* South Asian J Cancer, 2016. **5**(3): p. 105-10.

47. Forastiere, A.A., et al., *Concurrent chemotherapy and radiotherapy for organ preservation in advanced laryngeal cancer.* N Engl J Med, 2003. **349**(22): p. 2091-8.

48. Calais, G., et al., *Randomized trial of radiation therapy versus concomitant chemotherapy and radiation therapy for advanced-stage oropharynx carcinoma.* J Natl Cancer Inst, 1999. **91**(24): p. 2081-6.

49. Bonner, J.A., et al., *Radiotherapy plus cetuximab for locoregionally advanced head and neck cancer: 5-year survival data from a phase 3 randomised trial, and relation between cetuximab-induced rash and survival.* Lancet Oncol, 2010. **11**(1): p. 21-8.

50. Reddy, B.K., et al., *Nimotuzumab provides survival benefit to patients with inoperable advanced squamous cell carcinoma of the head and neck: a randomized, open-label, phase IIb, 5-year study in Indian patients.* Oral Oncol, 2014. **50**(5): p. 498-505.

51. Weinberg, R.A., *How cancer arises.* Sci Am, 1996. **275**(3): p. 62-70.

52. Lander, E.S., *Initial impact of the sequencing of the human genome.* Nature, 2011. **470**(7333): p. 187-97.

53. Garraway, L.A. and E.S. Lander, *Lessons from the cancer genome.* Cell, 2013. **153**(1): p. 17-37.

54. Meyerson, M., S. Gabriel, and G. Getz, *Advances in understanding cancer genomes through second-generation sequencing.* Nat Rev Genet, 2010. **11**(10): p. 685-96.

55. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.

56. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.

57. Lunshof, J.E., et al., *Personal genomes in progress: from the human genome project to the personal genome project.* Dialogues Clin Neurosci, 2010. **12**(1): p. 47-60.

58. Hutchison, C.A., 3rd, *DNA sequencing: bench to bedside and beyond.* Nucleic Acids Res, 2007. **35**(18): p. 6227-37.

59. Mardis, E.R., *The impact of next-generation sequencing technology on genetics.* Trends Genet, 2008. **24**(3): p. 133-41.

60. Reis-Filho, J.S., *Next-generation sequencing.* Breast Cancer Res, 2009. **11 Suppl 3**: p. S12.

61. Mamanova, L., et al., *Target-enrichment strategies for next-generation sequencing.* Nat Methods, 2010. **7**(2): p. 111-8.

62. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.

63. Upadhyay, P., R. Dwivedi, and A. Dutt, *Applications of next-generation sequencing in cancer.* CURRENT SCIENCE, 2014. **107**(5): p. 795.

64. Stratton, M.R., P.J. Campbell, and P.A. Futreal, *The cancer genome.* Nature, 2009. **458**(7239): p. 719-24.

65. Sjoblom, T., et al., *The consensus coding sequences of human breast and colorectal cancers.* Science, 2006. **314**(5797): p. 268-74.

66. Ley, T.J., et al., *DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome.* Nature, 2008. **456**(7218): p. 66-72.

67. Banerji, S., et al., *Sequence analysis of mutations and translocations across breast cancer subtypes.* Nature, 2012. **486**(7403): p. 405-9.

68. Cancer Genome Atlas Research, N., *Comprehensive genomic characterization of squamous cell lung cancers.* Nature, 2012. **489**(7417): p. 519-25.

69. Ellis, M.J., et al., *Whole-genome analysis informs breast cancer response to aromatase inhibition.* Nature, 2012. **486**(7403): p. 353-60.

70. Nik-Zainal, S., et al., *Mutational processes molding the genomes of 21 breast cancers.* Cell, 2012. **149**(5): p. 979-93.

71. Nik-Zainal, S., et al., *The life history of 21 breast cancers.* Cell, 2012. **149**(5): p. 994-1007.

72. Shah, S.P., et al., *The clonal and mutational evolution spectrum of primary triple-negative breast cancers.* Nature, 2012. **486**(7403): p. 395-9.

73. Stephens, P.J., et al., *Complex landscapes of somatic rearrangement in human breast cancer genomes.* Nature, 2009. **462**(7276): p. 1005-10.

74. Stephens, P.J., et al., *The landscape of cancer genes and mutational processes in breast cancer.* Nature, 2012. **486**(7403): p. 400-4.

75. Cancer Genome Atlas Research, N., *Integrated genomic analyses of ovarian carcinoma.* Nature, 2011. **474**(7353): p. 609-15.

76. Cancer Genome Atlas, N., *Comprehensive molecular characterization of human colon and rectal cancer.* Nature, 2012. **487**(7407): p. 330-7.

77. Seshagiri, S., et al., *Recurrent R-spondin fusions in colon cancer.* Nature, 2012. **488**(7413): p. 660-4.

78. Totoki, Y., et al., *High-resolution characterization of a hepatocellular carcinoma genome.* Nat Genet, 2011. **43**(5): p. 464-9.

79. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing.* N Engl J Med, 2012. **366**(10): p. 883-92.

80. Agrawal, N., et al., *Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1.* Science, 2011. **333**(6046): p. 1154-7.

81. Berger, M.F., et al., *Melanoma genome sequencing reveals frequent PREX2 mutations.* Nature, 2012. **485**(7399): p. 502-6.

82. Ding, L., et al., *Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing.* Nature, 2012. **481**(7382): p. 506-10.

83. Welch, J.S., et al., *The origin and evolution of mutations in acute myeloid leukemia.* Cell, 2012. **150**(2): p. 264-78.

84. Wong, W.C., et al., *CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer.* Bioinformatics, 2011. **27**(15): p. 2147-8.

85. Kaminker, J.S., et al., *CanPredict: a computational tool for predicting cancer-associated missense mutations.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W595-8.

86. Li, B., et al., *Automated inference of molecular mechanisms of disease from amino acid substitutions.* Bioinformatics, 2009. **25**(21): p. 2744-50.

87. Torkamani, A. and N.J. Schork, *Prediction of cancer driver mutations in protein kinases.* Cancer Res, 2008. **68**(6): p. 1675-82.

88. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic Acids Res, 2003. **31**(13): p. 3812-4.

89. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.

90. Schwarz, J.M., et al., *MutationTaster evaluates disease-causing potential of sequence alterations.* Nat Methods, 2010. **7**(8): p. 575-6.

91. Reva, B., Y. Antipin, and C. Sander, *Determinants of protein function revealed by combinatorial entropy optimization.* Genome Biol, 2007. **8**(11): p. R232.

92. International Cancer Genome, C., et al., *International network of cancer genome projects.* Nature, 2010. **464**(7291): p. 993-8.

93. Jennings, J. and T.J. Hudson, *Reflections on the founding of the International Cancer Genome Consortium.* Clin Chem, 2013. **59**(1): p. 18-21.

94. Vazquez, M., V. de la Torre, and A. Valencia, *Chapter 14: Cancer genome analysis.* PLoS Comput Biol, 2012. **8**(12): p. e1002824.

95. Yang, Y., et al., *Databases and web tools for cancer genomics study.* Genomics Proteomics Bioinformatics, 2015. **13**(1): p. 46-50.

96. Chin, L., et al., *Making sense of cancer genomic data.* Genes Dev, 2011. **25**(6): p. 534-55.

97. Marusyk, A. and K. Polyak, *Tumor heterogeneity: causes and consequences.* Biochim Biophys Acta, 2010. **1805**(1): p. 105-17.

98. Merlo, L.M., et al., *Cancer as an evolutionary and ecological process.* Nat Rev Cancer, 2006. **6**(12): p. 924-35.

99. Liu, E.T., *Functional genomics of cancer.* Curr Opin Genet Dev, 2008. **18**(3): p. 251-6.

100. Almendro, V. and G. Fuster, *Heterogeneity of breast cancer: etiology and clinical relevance.* Clin Transl Oncol, 2011. **13**(11): p. 767-73.

101. Curtis, C., et al., *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.* Nature, 2012. **486**(7403): p. 346-52.

102. Russnes, H.G., et al., *Insight into the heterogeneity of breast cancer through next-generation sequencing.* J Clin Invest, 2011. **121**(10): p. 3810-8.

103. Yancovitz, M., et al., *Intra- and inter-tumor heterogeneity of BRAF(V600E))mutations in primary and metastatic melanoma.* PLoS One, 2012. **7**(1): p. e29336.

104. Martincorena, I. and P.J. Campbell, *Somatic mutation in cancer and normal cells.* Science, 2015. **349**(6255): p. 1483-9.

105. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.

106. Jones, S., et al., *Personalized genomic analyses for cancer mutation discovery and interpretation.* Sci Transl Med, 2015. **7**(283): p. 283ra53.

107. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.

108. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.

109. Wang, L. and D.A. Wheeler, *Genomic sequencing for cancer diagnosis and therapy.* Annu Rev Med, 2014. **65**: p. 33-48.

110. Consortium, E.A., et al., *Analysis of protein-coding genetic variation in 60,706 humans.* Biorxiv, 2015.

111. Yang, H. and K. Wang, *Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR.* Nat Protoc, 2015. **10**(10): p. 1556-66.

112. Narang, A., et al., *IGVBrowser--a genomic variation resource from diverse Indian populations.* Database (Oxford), 2010. **2010**: p. baq022.

113. Indian Genome Variation, C., *The Indian Genome Variation database (IGVdb): a project overview.* Hum Genet, 2005. **118**(1): p. 1-11.

114. Consortium, H.P.-A.S., et al., *Mapping human genetic diversity in Asia.* Science, 2009. **326**(5959): p. 1541-5.

115. Tamang, R., L. Singh, and K. Thangaraj, *Complex genetic origin of Indian populations and its implications.* J Biosci, 2012. **37**(5): p. 911-9.

116. Tamang, R. and K. Thangaraj, *Genomic view on the peopling of India.* Investig Genet, 2012. **3**(1): p. 20.

117. Reich, D., et al., *Reconstructing Indian population history.* Nature, 2009. **461**(7263): p. 489-94.

118. Majumder, P.P. and A. Basu, *A genomic view of the peopling and population structure of India.* Cold Spring Harb Perspect Biol, 2015. **7**(4): p. a008540.

119. Basu, A., et al., *Ethnic India: a genomic view, with special reference to peopling and structure.* Genome Res, 2003. **13**(10): p. 2277-90.

120. Sengupta, S., et al., *Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists.* Am J Hum Genet, 2006. **78**(2): p. 202-21.

121. Tabatabaeifar, S., et al., *Use of next generation sequencing in head and neck squamous cell carcinomas: a review.* Oral Oncol, 2014. **50**(11): p. 1035-40.

122. Nguyen, B.C., et al., *Cross-regulation between Notch and p63 in keratinocyte commitment to differentiation.* Genes Dev, 2006. **20**(8): p. 1028-42.

123. Song, X., et al., *Common and complex Notch1 mutations in Chinese oral squamous cell carcinoma.* Clin Cancer Res, 2014. **20**(3): p. 701-10.

124. Tan, Y., O. Sangfelt, and C. Spruck, *The Fbxw7/hCdc4 tumor suppressor in human cancer.* Cancer Lett, 2008. **271**(1): p. 1-12.

125. India Project Team of the International Cancer Genome, C., *Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups.* Nat Commun, 2013. **4**: p. 2873.

126. Hayes, D.N., C. Van Waes, and T.Y. Seiwert, *Genetic Landscape of Human Papillomavirus-Associated Head and Neck Cancer and Comparison to Tobacco-Related Tumors.* J Clin Oncol, 2015. **33**(29): p. 3227-34.

127. Stransky, N., et al., *The mutational landscape of head and neck squamous cell carcinoma.* Science, 2011. **333**(6046): p. 1157-60.

128. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-21.

129. Fanjul-Fernandez, M., et al., *Cell-cell adhesion genes CTNNA2 and CTNNA3 are tumour suppressors frequently mutated in laryngeal carcinomas.* Nat Commun, 2013. **4**: p. 2531.

130. Nichols, A.C., et al., *A Pilot Study Comparing HPV-Positive and HPV-Negative Head and Neck Squamous Cell Carcinomas by Whole Exome Sequencing.* ISRN Oncol, 2012. **2012**: p. 809370.

131. Lui, V.W., et al., *Frequent mutation of the PI3K pathway in head and neck cancer defines predictive biomarkers.* Cancer Discov, 2013. **3**(7): p. 761-9.

132. Kandoth, C., et al., *Mutational landscape and significance across 12 major cancer types.* Nature, 2013. **502**(7471): p. 333-9.

133. Pickering, C.R., et al., *Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers.* Cancer Discov, 2013. **3**(7): p. 770-81.

134. Pickering, C.R., et al., *Squamous cell carcinoma of the oral tongue in young non-smokers is genomically similar to tumors in older smokers.* Clin Cancer Res, 2014. **20**(14): p. 3842-8.

135. Cancer Genome Atlas, N., *Comprehensive genomic characterization of head and neck squamous cell carcinomas.* Nature, 2015. **517**(7536): p. 576-82.

136. Nichols, A.C., et al., *Frequent mutations in TP53 and CDKN2A found by next-generation sequencing of head and neck cancer cell lines.* Arch Otolaryngol Head Neck Surg, 2012. **138**(8): p. 732-9.

137. Lechner, M., et al., *Targeted next-generation sequencing of head and neck squamous cell carcinoma identifies novel genetic alterations in HPV+ and HPV- tumors.* Genome Med, 2013. **5**(5): p. 49.

138. Seiwert, T.Y., et al., *Integrative and comparative genomic analysis of HPV-positive and HPV-negative head and neck squamous cell carcinomas.* Clin Cancer Res, 2015. **21**(3): p. 632-41.

139. Mroz, E.A., et al., *High intratumor genetic heterogeneity is related to worse outcome in patients with head and neck squamous cell carcinoma.* Cancer, 2013. **119**(16): p. 3034-42.

140. Mroz, E.A. and J.W. Rocco, *MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma.* Oral Oncol, 2013. **49**(3): p. 211-5.

141. zur Hausen, H., *Papillomaviruses and cancer: from basic studies to clinical application.* Nat Rev Cancer, 2002. **2**(5): p. 342-50.

142. 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.

143. Dyson, N., et al., *The human papilloma virus-16 E7 oncoprotein is able to bind to the retinoblastoma gene product.* Science, 1989. **243**(4893): p. 934-7.

144. Venuti, A. and F. Paolini, *HPV detection methods in head and neck cancer.* Head Neck Pathol, 2012. **6 Suppl 1**: p. S63-74.

145. Liang, C., et al., *Biomarkers of HPV in head and neck squamous cell carcinoma.* Cancer Res, 2012. **72**(19): p. 5004-13.

146. Schlecht, N.F., et al., *A comparison of clinically utilized human papillomavirus detection methods in head and neck cancer.* Mod Pathol, 2011. **24**(10): p. 1295-305.

147. Vokes, E.E., N. Agrawal, and T.Y. Seiwert, *HPV-Associated Head and Neck Cancer.* J Natl Cancer Inst, 2015. **107**(12): p. djv344.

148. Smeets, S.J., et al., *A novel algorithm for reliable detection of human papillomavirus in paraffin embedded head and neck cancer specimen.* Int J Cancer, 2007. **121**(11): p. 2465-72.

149. Kreimer, A.R., et al., *Human papillomavirus types in head and neck squamous cell carcinomas worldwide: a systematic review.* Cancer Epidemiol Biomarkers Prev, 2005. **14**(2): p. 467-75.

150. 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.

151. Slebos, R.J., et al., *Gene expression differences associated with human papillomavirus status in head and neck squamous cell carcinoma.* Clin Cancer Res, 2006. **12**(3 Pt 1): p. 701-9.

152. Psyrri, A., T. Rampias, and J.B. Vermorken, *The current and future impact of human papillomavirus on treatment of squamous cell carcinoma of the head and neck.* Ann Oncol, 2014. **25**(11): p. 2101-15.

153. Ferlay, J., et al., *Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012.* Int J Cancer, 2015. **136**(5): p. E359-86.

154. Mountzios, G., T. Rampias, and A. Psyrri, *The mutational spectrum of squamous-cell carcinoma of the head and neck: targetable genetic events and clinical impact.* Ann Oncol, 2014. **25**(10): p. 1889-900.

155. Chung, C.H., et al., *Genomic alterations in head and neck squamous cell carcinoma determined by cancer gene-targeted sequencing.* Ann Oncol, 2015. **26**(6): p. 1216-23.

156. Chung, G.T., et al., *Constitutive activation of distinct NF-kappaB signals in EBV-associated nasopharyngeal carcinoma.* J Pathol, 2013. **231**(3): p. 311-22.

157. Hacker, H., P.H. Tseng, and M. Karin, *Expanding TRAF function: TRAF3 as a tri-faced immune regulator.* Nat Rev Immunol, 2011. **11**(7): p. 457-68.

158. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-8.

159. Henderson, S., et al., *APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development.* Cell Rep, 2014. **7**(6): p. 1833-41.

160. Rizvi, N.A., et al., *Activity and safety of nivolumab, an anti-PD-1 immune checkpoint inhibitor, for patients with advanced, refractory squamous non-small-cell lung cancer (CheckMate 063): a phase 2, single-arm trial.* Lancet Oncol, 2015. **16**(3): p. 257-65.

161. Morris, L.G., et al., *Recurrent somatic mutation of FAT1 in multiple human cancers leads to aberrant Wnt activation.* Nat Genet, 2013. **45**(3): p. 253-61.

162. Moore, S.R., et al., *The epidemiology of tongue cancer: a review of global incidence.* Oral Dis, 2000. **6**(2): p. 75-84.

163. Warnakulasuriya, S., *Global epidemiology of oral and oropharyngeal cancer.* Oral Oncol, 2009. **45**(4-5): p. 309-16.

164. Sano, D. and J.N. Myers, *Metastasis of squamous cell carcinoma of the oral tongue.* Cancer Metastasis Rev, 2007. **26**(3-4): p. 645-62.

165. Chaturvedi, A.K., et al., *Worldwide trends in incidence rates for oral cavity and oropharyngeal cancers.* J Clin Oncol, 2013. **31**(36): p. 4550-9.

166. Krishnamurthy, A. and V. Ramshankar, *Early stage oral tongue cancer among non-tobacco users--an increasing trend observed in a South Indian patient population presenting at a single centre.* Asian Pac J Cancer Prev, 2013. **14**(9): p. 5061-5.

167. Sherin, N., et al., *Changing trends in oral cancer.* Indian J Cancer, 2008. **45**(3): p. 93-6.

168. Elango, J.K., et al., *Trends of head and neck cancers in urban and rural India.* Asian Pac J Cancer Prev, 2006. **7**(1): p. 108-12.

169. Bodner, L., et al., *Oral squamous cell carcinoma in patients twenty years of age or younger--review and analysis of 186 reported cases.* Oral Oncol, 2014. **50**(2): p. 84-9.

170. Montero, P.H., et al., *Changing trends in smoking and alcohol consumption in patients with oral cancer treated at Memorial Sloan-Kettering Cancer Center from 1985 to 2009.* Arch Otolaryngol Head Neck Surg, 2012. **138**(9): p. 817-22.

171. Mackenzie, J., et al., *Increasing incidence of oral cancer amongst young persons: what is the aetiology?* Oral Oncol, 2000. **36**(4): p. 387-9.

172. Indian Council of Medical Research, I., *Tongue Cancer - Indian Council of Medical Research.* 2015.

173. Chocolatewala, N.M. and P. Chaturvedi, *Role of human papilloma virus in the oral carcinogenesis: an Indian perspective.* J Cancer Res Ther, 2009. **5**(2): p. 71-7.

174. Patil, V.M., et al., *Induction chemotherapy in technically unresectable locally advanced oral cavity cancers: does it make a difference?* Indian J Cancer, 2013. **50**(1): p. 1-8.

175. Bhattacharyya, N. and M.P. Fried, *Benchmarks for mortality, morbidity, and length of stay for head and neck surgical procedures.* Arch Otolaryngol Head Neck Surg, 2001. **127**(2): p. 127-32.

176. Kapoor, C., et al., *Lymph node metastasis: A bearing on prognosis in squamous cell carcinoma.* Indian J Cancer, 2015. **52**(3): p. 417-24.

177. Thiagarajan, S., et al., *Predictors of prognosis for squamous cell carcinoma of oral tongue.* J Surg Oncol, 2014. **109**(7): p. 639-44.

178. Byers, R.M., et al., *Frequency and therapeutic implications of "skip metastases" in the neck from squamous carcinoma of the oral tongue.* Head Neck, 1997. **19**(1): p. 14-9.

179. Zelefsky, M.J., et al., *Postoperative radiotherapy for oral cavity cancers: impact of anatomic subsite on treatment outcome.* Head Neck, 1990. **12**(6): p. 470-5.

180. Krishna Rao, S.V., et al., *Epidemiology of oral cancer in Asia in the past decade--an update (2000-2012).* Asian Pac J Cancer Prev, 2013. **14**(10): p. 5567-77.

181. Vettore, A.L., et al., *Mutational landscapes of tongue carcinoma reveal recurrent mutations in genes of therapeutic and prognostic relevance.* Genome Med, 2015. **7**(1): p. 98.

182. Krishnan, N., et al., *Integrated analysis of oral tongue squamous cell carcinoma identifies key variants and pathways linked to risk habits, HPV, clinical parameters and tumor recurrence.* F1000Res, 2015. **4**: p. 1215.

183. Li, R., et al., *Clinical, genomic, and metagenomic characterization of oral tongue squamous cell carcinoma in patients who do not smoke.* Head Neck, 2014.

184. Samman, M., et al., *A novel genomic signature reclassifies an oral cancer subtype.* Int J Cancer, 2015. **137**(10): p. 2364-73.

185. Heaton, C.M., et al., *TP53 and CDKN2a mutations in never-smoker oral tongue squamous cell carcinoma.* Laryngoscope, 2014. **124**(7): p. E267-73.

186. Lim, A.M., et al., *Differential mechanisms of CDKN2A (p16) alteration in oral tongue squamous cell carcinomas and correlation with patient outcome.* Int J Cancer, 2014. **135**(4): p. 887-95.

187. Ledgerwood, L.G., et al., *The degree of intratumor mutational heterogeneity varies by primary tumor sub-site.* Oncotarget, 2016. **7**(19): p. 27185-98.

188. Lee, H., P. Flaherty, and H.P. Ji, *Systematic genomic identification of colorectal cancer genes delineating advanced from early clinical stage and metastasis.* BMC Med Genomics, 2013. **6**: p. 54.

189. Cibulskis, K., et al., *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.* Nat Biotechnol, 2013. **31**(3): p. 213-9.

190. Dees, N.D., et al., *MuSiC: identifying mutational significance in cancer genomes.* Genome Res, 2012. **22**(8): p. 1589-98.

191. Gonzalez-Perez, A., et al., *IntOGen-mutations identifies cancer drivers across tumor types.* Nat Methods, 2013. **10**(11): p. 1081-2.

192. Hodis, E., et al., *A landscape of driver mutations in melanoma.* Cell, 2012. **150**(2): p. 251-63.

193. Reimand, J. and G.D. Bader, *Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers.* Mol Syst Biol, 2013. **9**: p. 637.

194. Mermel, C.H., et al., *GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.* Genome Biol, 2011. **12**(4): p. R41.

195. Lawrence, M.S., et al., *Discovery and saturation analysis of cancer genes across 21 tumour types.* Nature, 2014. **505**(7484): p. 495-501.

196. Akavia, U.D., et al., *An integrated approach to uncover drivers of cancer.* Cell, 2010. **143**(6): p. 1005-17.

197. Natrajan, R. and P. Wilkerson, *From integrative genomics to therapeutic targets.* Cancer Res, 2013. **73**(12): p. 3483-8.

198. Wilkerson, M.D., et al., *Integrated RNA and DNA sequencing improves mutation detection in low purity tumors.* Nucleic Acids Res, 2014. **42**(13): p. e107.

199. Kristensen, V.N., et al., *Principles and methods of integrative genomic analyses in cancer.* Nat Rev Cancer, 2014. **14**(5): p. 299-313.

200. Mulherkar, R., et al., *Establishment of a human squamous cell carcinoma cell line of the upper aero-digestive tract.* Cancer Lett, 1997. **118**(1): p. 115-21.

201. Tatake, R.J., et al., *Establishment and characterization of four new squamous cell carcinoma cell lines derived from oral tumors.* J Cancer Res Clin Oncol, 1990. **116**(2): p. 179-86.

202. Krzywinski, M., et al., *Circos: an information aesthetic for comparative genomics.* Genome Res, 2009. **19**(9): p. 1639-45.

203. Minton, J.A., S.E. Flanagan, and S. Ellard, *Mutation surveyor: software for DNA sequence analysis.* Methods Mol Biol, 2011. **688**: p. 143-53.

204. Livak, K.J. and T.D. Schmittgen, *Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.* Methods, 2001. **25**(4): p. 402-8.

205. Sancak, Y., et al., *The Rag GTPases bind raptor and mediate amino acid signaling to mTORC1.* Science, 2008. **320**(5882): p. 1496-501.

206. Dutt, A., et al., *Drug-sensitive FGFR2 mutations in endometrial carcinoma.* Proc Natl Acad Sci U S A, 2008. **105**(25): p. 8713-7.

207. Moffat, J., et al., *A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen.* Cell, 2006. **124**(6): p. 1283-98.

208. Walker, J.M., *The bicinchoninic acid (BCA) assay for protein quantitation.* Methods Mol Biol, 1994. **32**: p. 5-8.

209. Yamamoto, N., et al., *Allelic loss on chromosomes 2q, 3p and 21q: possibly a poor prognostic factor in oral squamous cell carcinoma.* Oral Oncol, 2003. **39**(8): p. 796-805.

210. Partridge, M., G. Emilion, and J.D. Langdon, *LOH at 3p correlates with a poor survival in oral squamous cell carcinoma.* Br J Cancer, 1996. **73**(3): p. 366-71.

211. Chen, Y. and C. Chen, *DNA copy number variation and loss of heterozygosity in relation to recurrence of and survival from head and neck squamous cell carcinoma: a review.* Head Neck, 2008. **30**(10): p. 1361-83.

212. Meredith, S.D., et al., *Chromosome 11q13 amplification in head and neck squamous cell carcinoma. Association with poor prognosis.* Arch Otolaryngol Head Neck Surg, 1995. **121**(7): p. 790-4.

213. Adzhubei, I., D.M. Jordan, and S.R. Sunyaev, *Predicting functional effect of human missense mutations using PolyPhen-2.* Curr Protoc Hum Genet, 2013. **Chapter 7**: p. Unit7 20.

214. Choi, Y. and A.P. Chan, *PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels.* Bioinformatics, 2015. **31**(16): p. 2745-7.

215. Reva, B., Y. Antipin, and C. Sander, *Predicting the functional impact of protein mutations: application to cancer genomics.* Nucleic Acids Res, 2011. **39**(17): p. e118.

216. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603-7.

217. Davies, H., et al., *Somatic mutations of the protein kinase gene family in human lung cancer.* Cancer Res, 2005. **65**(17): p. 7591-5.

218. Hooper, J.D., et al., *Cloning of the cDNA and localization of the gene encoding human NRBP, a ubiquitously expressed, multidomain putative adapter protein.* Genomics, 2000. **66**(1): p. 113-8.

219. Schweingruber, C., et al., *Nonsense-mediated mRNA decay - mechanisms of substrate mRNA recognition and degradation in mammalian cells.* Biochim Biophys Acta, 2013. **1829**(6-7): p. 612-23.

220. Neu-Yilik, G., et al., *Mechanism of escape from nonsense-mediated mRNA decay of human beta-globin transcripts with nonsense mutations in the first exon.* RNA, 2011. **17**(5): p. 843-54.

221. Ambatipudi, S., et al., *Genomic Profiling of Advanced-Stage Oral Cancers Reveals Chromosome 11q Alterations as Markers of Poor Clinical Outcome.* PLoS ONE, 2011. **6**(2): p. e17250.

222. Ntziachristos, P., et al., *From fly wings to targeted cancer therapies: a centennial for notch signaling.* Cancer Cell, 2014. **25**(3): p. 318-34.

223. Hua, F., et al., *TRB3 interacts with SMAD3 promoting tumor cell migration and invasion.* J Cell Sci, 2011. **124**(Pt 19): p. 3235-46.

224. Manning, G., et al., *The protein kinase complement of the human genome.* Science, 2002. **298**(5600): p. 1912-34.

225. Zeqiraj, E., et al., *Structure of the LKB1-STRAD-MO25 complex reveals an allosteric mechanism of kinase activation.* Science, 2009. **326**(5960): p. 1707-11.

226. Gluderer, S., et al., *Bunched, the Drosophila homolog of the mammalian tumor suppressor TSC-22, promotes cellular growth.* BMC Dev Biol, 2008. **8**: p. 10.

227. Ruiz, C., et al., *High NRBP1 expression in prostate cancer is linked with poor clinical outcomes and increased cancer cell growth.* Prostate, 2012. **72**(15): p. 1678-87.

228. Doi, Y., et al., *Expression and cellular localization of TSC-22 in normal salivary glands and salivary gland tumors: implications for tumor cell differentiation.* Oncol Rep, 2008. **19**(3): p. 609-16.

229. Wilson, C.H., et al., *Nuclear receptor binding protein 1 regulates intestinal progenitor cell homeostasis and tumour formation.* EMBO J, 2012. **31**(11): p. 2486-97.

230. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

231. Garcia-Alcalde, F., et al., *Qualimap: evaluating next-generation sequencing alignment data.* Bioinformatics, 2012. **28**(20): p. 2678-9.

232. 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.

233. Bamford, S., et al., *The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website.* Br J Cancer, 2004. **91**(2): p. 355-8.

234. Ramos, A.H., et al., *Oncotator: cancer variant annotation tool.* Hum Mutat, 2015. **36**(4): p. E2423-9.

235. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

236. Xu, B., et al., *Exome sequencing supports a de novo mutational paradigm for schizophrenia.* Nat Genet, 2011. **43**(9): p. 864-8.

237. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.

238. Guo, Y., et al., *Exome sequencing generates high quality data in non-target regions.* BMC Genomics, 2012. **13**: p. 194.

239. Samuels, D.C., et al., *Finding the lost treasures in exome sequencing data.* Trends Genet, 2013. **29**(10): p. 593-9.

240. Perez-Lezaun, A., et al., *Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA.* Am J Hum Genet, 1999. **65**(1): p. 208-19.

241. Oota, H., et al., *Extreme mtDNA homogeneity in continental Asian populations.* Am J Phys Anthropol, 2002. **118**(2): p. 146-53.

242. Kumar, V., et al., *Global patterns in human mitochondrial DNA and Y-chromosome variation caused by spatial instability of the local cultural processes.* PLoS Genet, 2006. **2**(4): p. e53.

243. Jobling, M.A. and C. Tyler-Smith, *The human Y chromosome: an evolutionary marker comes of age.* Nat Rev Genet, 2003. **4**(8): p. 598-612.

244. Shrivastava, P., et al., *Y STR haplotype diversity in central Indian population.* Ann Hum Biol, 2015: p. 1-8.

245. Kumar, A., et al., *Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers.* Proc Natl Acad Sci U S A, 2011. **108**(41): p. 17087-92.

246. Suzuki, A., et al., *Identification and characterization of cancer mutations in Japanese lung adenocarcinoma without sequencing of normal tissue counterparts.* PLoS One, 2013. **8**(9): p. e73484.

247. Raymond, V.M., et al., *Germline Findings in Tumor-Only Sequencing: Points to Consider for Clinicians and Laboratories.* J Natl Cancer Inst, 2016. **108**(4).

248. McCarthy, M., *Genomic sequencing of only tumor tissue could be misleading in nearly half of patients, study shows.* BMJ, 2015. **350**: p. h2036.

249. Dakubo, G.D., et al., *Clinical implications and utility of field cancerization.* Cancer Cell Int, 2007. **7**: p. 2.

250. Mohan, M. and N. Jagannathan, *Oral field cancerization: an update on current concepts.* Oncol Rev, 2014. **8**(1): p. 244.

251. Garnaes, E., et al., *Increasing incidence of base of tongue cancers from 2000 to 2010 due to HPV: the largest demographic study of 210 Danish patients.* Br J Cancer, 2015. **113**(1): p. 131-4.

252. Datta, S., et al., *A review of Indian literature for association of smokeless tobacco with malignant and premalignant diseases of head and neck region.* Indian J Cancer, 2014. **51**(3): p. 200-208.

253. Wyss, A., et al., *Cigarette, cigar, and pipe smoking and the risk of head and neck cancers: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium.* Am J Epidemiol, 2013. **178**(5): p. 679-90.

254. O'Rorke, M.A., et al., *Human papillomavirus related head and neck cancer survival: a systematic review and meta-analysis.* Oral Oncol, 2012. **48**(12): p. 1191-201.

255. Guo, T., et al., *Characterization of functionally active gene fusions in human papillomavirus related oropharyngeal squamous cell carcinoma.* Int J Cancer, 2016. **139**(2): p. 373-82.

256. Daly, C., et al., *FGFR3-TACC3 fusion proteins act as naturally occurring drivers of tumor resistance by functionally substituting for EGFR/ERK signaling.* Oncogene, 2016.

257. Upadhyay, P., et al., *Notch pathway activation is essential for maintenance of stem-like cells in early tongue cancer.* Oncotarget, 2016.

258. Upadhyay P, G.N., Desai S*, Sahoo B, Singh A,Togar T, Iyer P, Prasad R, Chandrani P, Gupta S, Dutt A, *TMC-SNPdb: an Indian germline variant database derived from whole exome sequences.* Database (Oxford), 2016.

259. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**: p. 323.

260. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-8.

261. Forbes, S.A., et al., *The Catalogue of Somatic Mutations in Cancer (COSMIC).* Curr Protoc Hum Genet, 2008. **Chapter 10**: p. Unit 10 11.

262. Liu, X., X. Jian, and E. Boerwinkle, *dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations.* Hum Mutat, 2013. **34**(9): p. E2393-402.

263. Mao, Y., et al., *CanDrA: cancer-specific driver missense mutation annotation with optimized features.* PLoS One, 2013. **8**(10): p. e77945.

264. Choi, Y., et al., *Predicting the functional effect of amino acid substitutions and indels.* PLoS One, 2012. **7**(10): p. e46688.

265. 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.

266. Beroukhim, R., et al., *Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma.* Proc Natl Acad Sci U S A, 2007. **104**(50): p. 20007-12.

267. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nat Protoc, 2012. **7**(3): p. 562-78.

268. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

269. Iyer, M.K., A.M. Chinnaiyan, and C.A. Maher, *ChimeraScan: a tool for identifying chimeric transcription in sequencing data.* Bioinformatics, 2011. **27**(20): p. 2903-4.

270. Yoshihara, K., et al., *The landscape and therapeutic relevance of cancer-associated transcript fusions.* Oncogene, 2015. **34**(37): p. 4845-54.

271. Shugay, M., et al., *Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions.* Bioinformatics, 2013. **29**(20): p. 2539-46.

272. Ye, J., et al., *Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction.* BMC Bioinformatics, 2012. **13**: p. 134.

273. Johnson, M., et al., *NCBI BLAST: a better web interface.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W5-9.

274. Gagliardi, A.R., et al., *A framework of the desirable features of guideline implementation tools (GItools): Delphi survey and assessment of GItools.* Implement Sci, 2014. **9**: p. 98.

275. Schwartzentruber, J., et al., *Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma.* Nature, 2012. **482**(7384): p. 226-31.

276. Chen, Y., et al., *Identification of druggable cancer driver genes amplified across TCGA datasets.* PLoS One, 2014. **9**(5): p. e98293.

277. Ye, H., et al., *Transcriptomic dissection of tongue squamous cell carcinoma.* BMC Genomics, 2008. **9**: p. 69.

278. Rickman, D.S., et al., *Prediction of future metastasis and molecular characterization of head and neck squamous-cell carcinoma based on transcriptome and genome analysis by microarrays.* Oncogene, 2008. **27**(51): p. 6607-22.

279. O'Donnell, R.K., et al., *Gene expression signature predicts lymphatic metastasis in squamous cell carcinoma of the oral cavity.* Oncogene, 2005. **24**(7): p. 1244-51.

280. Yu, Y.H., H.K. Kuo, and K.W. Chang, *The evolving transcriptome of head and neck squamous cell carcinoma: a systematic review.* PLoS One, 2008. **3**(9): p. e3215.

281. Pratik Chandrani, P.U., Prajish Iyer, Mayur Tanna, Madhur Shetty, Gorantala Venkata Raghuram, Ninad Oak, Ankita Singh, Rohan Chaubal and S.G. Manoj Ramteke, Amit Dutt, *Integrated genomics approach to identify biologically relevant alterations in fewer samples.* BMC Genomics, 2015. **16**: p. 936.

282. Brennan, C.W., et al., *The somatic genomic landscape of glioblastoma.* Cell, 2013. **155**(2): p. 462-77.

283. Babiceanu, M., et al., *Recurrent chimeric fusion RNAs in non-cancer tissues and cells.* Nucleic Acids Res, 2016. **44**(6): p. 2859-72.

284. Kim, P., et al., *ChimerDB 2.0--a knowledgebase for fusion genes updated.* Nucleic Acids Res, 2010. **38**(Database issue): p. D81-5.

285. Frenkel-Morgenstern, M., et al., *ChiTaRS 2.1--an improved database of the chimeric transcripts and RNA-seq data with novel sense-antisense chimeric RNA transcripts.* Nucleic Acids Res, 2015. **43**(Database issue): p. D68-75.

286. Wang, Y., et al., *FusionCancer: a database of cancer fusion genes derived from RNA-seq data.* Diagn Pathol, 2015. **10**: p. 131.

287. 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.

288. Chandrani, P., et al., *Integrated genomics approach to identify biologically relevant alterations in fewer samples.* BMC Genomics, 2015. **16**: p. 936.

289. Saranath, D., et al., *High frequency mutation in codons 12 and 61 of H-ras oncogene in chewing tobacco-related human oral carcinoma in India.* Br J Cancer, 1991. **63**(4): p. 573-8.

290. Sathyan, K.M., K.R. Nalinakumari, and S. Kannan, *H-Ras mutation modulates the expression of major cell cycle regulatory proteins and disease prognosis in oral carcinoma.* Mod Pathol, 2007. **20**(11): p. 1141-8.

291. Bjorklund, P., et al., *The internally truncated LRP5 receptor presents a therapeutic target in breast cancer.* PLoS One, 2009. **4**(1): p. e4243.

292. Wang, Y., et al., *LSD1 co-repressor Rcor2 orchestrates neurogenesis in the developing mouse brain.* Nat Commun, 2016. **7**: p. 10481.

293. Vettore, A.L., et al., *Mutational landscapes of tongue carcinoma reveal recurrent mutations in genes of therapeutic and prognostic relevance.* Genome Med, 2015. **7**: p. 98.

294. Jemal, A., et al., *Global cancer statistics.* CA Cancer J Clin, 2011. **61**(2): p. 69-90.

295. Shrivastava, S., et al., *Identification of molecular signature of head and neck cancer stem-like cells.* Sci Rep, 2015. **5**: p. 7819.

296. Izumchenko, E., et al., *Notch1 mutations are drivers of oral tumorigenesis.* Cancer Prev Res (Phila), 2015. **8**(4): p. 277-86.

297. Sun, W., et al., *Activation of the NOTCH pathway in head and neck cancer.* Cancer Res, 2014. **74**(4): p. 1091-104.

298. Rettig, E.M., et al., *Cleaved NOTCH1 Expression Pattern in Head and Neck Squamous Cell Carcinoma Is Associated with NOTCH1 Mutation, HPV Status, and High-Risk Features.* Cancer Prev Res (Phila), 2015. **8**(4): p. 287-95.

299. Egloff, A.M. and J.R. Grandis, *Molecular pathways: context-dependent approaches to Notch targeting as cancer therapy.* Clin Cancer Res, 2012. **18**(19): p. 5188-95.

300. Suman, S., T.P. Das, and C. Damodaran, *Silencing NOTCH signaling causes growth arrest in both breast cancer stem cells and breast cancer cells.* Br J Cancer. **109**(10): p. 2587-96.

301. Wang, Z., et al., *Targeting notch to eradicate pancreatic cancer stem cells for cancer therapy.* Anticancer Res. **31**(4): p. 1105-13.

302. Prince, M.E., et al., *Identification of a subpopulation of cells with cancer stem cell properties in head and neck squamous cell carcinoma.* Proc Natl Acad Sci U S A, 2007. **104**(3): p. 973-8.

303. Grudzien, P., et al., *Inhibition of Notch signaling reduces the stem-like population of breast cancer cells and prevents mammosphere formation.* Anticancer Res, 2010. **30**(10): p. 3853-67.

304. Lapidot, T., et al., *A cell initiating human acute myeloid leukaemia after transplantation into SCID mice.* Nature, 1994. **367**(6464): p. 645-8.

305. Al-Hajj, M., et al., *Prospective identification of tumorigenic breast cancer cells.* Proc Natl Acad Sci U S A, 2003. **100**(7): p. 3983-8.

306. Singh, S.K., et al., *Identification of a cancer stem cell in human brain tumors.* Cancer Res, 2003. **63**(18): p. 5821-8.

307. Liu, A., X. Yu, and S. Liu, *Pluripotency transcription factors and cancer stem cells: small genes make a big difference.* Chin J Cancer, 2013. **32**(9): p. 483-7.

308. Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.* Brief Bioinform, 2013. **14**(2): p. 178-92.

309. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.

310. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* Genome Biol, 2013. **14**(4): p. R36.

311. Barbieri, C.E., et al., *Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer.* Nat Genet, 2012. **44**(6): p. 685-9.

312. Gioanni, J., et al., *Two new human tumor cell lines derived from squamous cell carcinomas of the tongue: establishment, characterization and response to cytotoxic treatment.* Eur J Cancer Clin Oncol, 1988. **24**(9): p. 1445-55.

313. Chang, S.E., et al., *DOK, a cell line established from human dysplastic oral mucosa, shows a partially transformed non-malignant phenotype.* Int J Cancer, 1992. **52**(6): p. 896-902.

314. Jin, G., et al., *MT1-MMP cleaves Dll1 to negatively regulate Notch signalling to maintain normal B-cell development.* EMBO J, 2011. **30**(11): p. 2281-93.

315. Castel, D., et al., *Dynamic binding of RBPJ is determined by Notch signaling status.* Genes Dev, 2013. **27**(9): p. 1059-71.

316. Dinse, G.E. and S.W. Lagakos, *Nonparametric estimation of lifetime and disease onset distributions from incomplete observations.* Biometrics, 1982. **38**(4): p. 921-32.

317. Laprise, C., et al., *No role for human papillomavirus infection in oral cancers in a region in southern India.* Int J Cancer, 2015.

318. Patel, K.R., et al., *Prevalence of high-risk human papillomavirus type 16 and 18 in oral and cervical cancers in population from Gujarat, West India.* J Oral Pathol Med, 2014. **43**(4): p. 293-7.

319. Pathare, S.M., et al., *Clinicopathological and prognostic implications of genetic alterations in oral cancers.* Oncol Lett, 2011. **2**(3): p. 445-451.

320. *Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups.* Nat Commun, 2013. **4**: p. 2873.

321. Wang, J., B.A. Sullenger, and J.N. Rich, *Notch signaling in cancer stem cells.* Adv Exp Med Biol, 2012. **727**: p. 174-85.

322. Takebe, N., et al., *Targeting cancer stem cells by inhibiting Wnt, Notch, and Hedgehog pathways.* Nat Rev Clin Oncol, 2011. **8**(2): p. 97-106.

323. Kagawa, S., et al., *Cellular senescence checkpoint function determines differential Notch1-dependent oncogenic and tumor-suppressor activities.* Oncogene, 2015. **34**(18): p. 2347-59.

324. Lefort, K., et al., *Notch1 is a p53 target gene involved in human keratinocyte tumor suppression through negative regulation of ROCK1/2 and MRCKalpha kinases.* Genes Dev, 2007. **21**(5): p. 562-77.

325. Kunnimalaiyaan, M., et al., *Overexpression of the NOTCH1 intracellular domain inhibits cell proliferation and alters the neuroendocrine phenotype of medullary thyroid cancer cells.* J Biol Chem, 2006. **281**(52): p. 39819-30.

326. Alniaimi, A.N., et al., *Increased Notch1 expression is associated with poor overall survival in patients with ovarian cancer.* Int J Gynecol Cancer, 2015. **25**(2): p. 208-13.

327. Zhou, L., et al., *Overexpressions of DLL4 and CD105 are Associated with Poor Prognosis of Patients with Pancreatic Ductal Adenocarcinoma.* Pathol Oncol Res, 2015. **21**(4): p. 1141-7.

328. Huang, J., et al., *Expression of Notch-1 and its clinical significance in different histological subtypes of human lung adenocarcinoma.* J Exp Clin Cancer Res, 2013. **32**: p. 84.

329. Tilley, A.E., et al., *Down-regulation of the notch pathway in human airway epithelium in association with smoking and chronic obstructive pulmonary disease.* Am J Respir Crit Care Med, 2009. **179**(6): p. 457-66.

330. Wang, G., et al., *Genes associated with MUC5AC expression in small airway epithelium of human smokers and non-smokers.* BMC Med Genomics, 2012. **5**: p. 21.

331. Yoshida, R., et al., *The pathological significance of Notch1 in oral squamous cell carcinoma.* Lab Invest, 2013. **93**(10): p. 1068-81.

332. Hijioka, H., et al., *Upregulation of Notch pathway molecules in oral squamous cell carcinoma.* Int J Oncol, 2010. **36**(4): p. 817-22.

333. Zhao, Z.L., et al., *NOTCH1 inhibition enhances the efficacy of conventional chemotherapeutic agents by targeting head neck cancer stem cell.* Sci Rep, 2016. **6**: p. 24704.

334. Lee, S.H., et al., *Notch1 signaling contributes to stemness in head and neck squamous cell carcinoma.* Lab Invest, 2016. **96**(5): p. 508-16.

335. Dikshit, R., et al., *Cancer mortality in India: a nationally representative survey.* Lancet, 2012. **379**(9828): p. 1807-16.

336. Yong-Deok, K., et al., *Molecular genetic study of novel biomarkers for early diagnosis of oral squamous cell carcinoma.* Med Oral Patol Oral Cir Bucal, 2015. **20**(2): p. e167-79.

337. Chang, J.T., et al., *Identification of differentially expressed genes in oral squamous cell carcinoma (OSCC): overexpression of NPM, CDK1 and NDRG1 and underexpression of CHES1.* Int J Cancer, 2005. **114**(6): p. 942-9.

338. Sethi, N., et al., *MicroRNAs and head and neck cancer: reviewing the first decade of research.* Eur J Cancer, 2014. **50**(15): p. 2619-35.

339. Babu, J.M., et al., *A miR-centric view of head and neck cancers.* Biochim Biophys Acta, 2011. **1816**(1): p. 67-72.

340. Zhou, X.L., et al., *Integrated microRNA-mRNA analysis revealing the potential roles of microRNAs in tongue squamous cell cancer.* Mol Med Rep, 2015. **12**(1): p. 885-94.

341. Yu, X. and Z. Li, *MicroRNA expression and its implications for diagnosis and therapy of tongue squamous cell carcinoma.* J Cell Mol Med, 2016. **20**(1): p. 10-6.

342. Manikandan, M., et al., *Oral squamous cell carcinoma: microRNA expression profiling and integrative analyses for elucidation of tumourigenesis mechanism.* Mol Cancer, 2016. **15**: p. 28.

343. Siddiqui, A.S., et al., *Sequence biases in large scale gene expression profiling data.* Nucleic Acids Res, 2006. **34**(12): p. e83.

344. Ramasamy, A., et al., *Key issues in conducting a meta-analysis of gene expression microarray datasets.* PLoS Med, 2008. **5**(9): p. e184.

345. Rung, J. and A. Brazma, *Reuse of public genome-wide gene expression data.* Nat Rev Genet, 2013. **14**(2): p. 89-99.

346. Thangaraj, S.V., et al., *Molecular Portrait of Oral Tongue Squamous Cell Carcinoma Shown by Integrative Meta-Analysis of Expression Profiles with Validations.* PLoS One, 2016. **11**(6): p. e0156582.

347. De Cecco, L., et al., *Comprehensive gene expression meta-analysis of head and neck squamous cell carcinoma microarray data defines a robust survival predictor.* Ann Oncol, 2014. **25**(8): p. 1628-35.

348. De Cecco, L., et al., *Head and neck cancer subtypes with biological and clinical relevance: Meta-analysis of gene-expression data.* Oncotarget, 2015. **6**(11): p. 9627-42.

349. Sun, Y., et al., *Transcriptomic characterization of differential gene expression in oral squamous cell carcinoma: a meta-analysis of publicly available microarray data sets.* Tumour Biol, 2016.

350. Chu, V.T., et al., *MeV+R: using MeV as a graphical user interface for Bioconductor applications in microarray analysis.* Genome Biol, 2008. **9**(7): p. R118.

351. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets--update.* Nucleic Acids Res, 2013. **41**(Database issue): p. D991-5.

352. Goldman, M., et al., *The UCSC Cancer Genomics Browser: update 2015.* Nucleic Acids Res, 2015. **43**(Database issue): p. D812-7.

353. Dweep, H., N. Gretz, and C. Sticht, *miRWalk database for miRNA-target interactions.* Methods Mol Biol, 2014. **1182**: p. 289-305.

354. Kozomara, A. and S. Griffiths-Jones, *miRBase: annotating high confidence microRNAs using deep sequencing data.* Nucleic Acids Res, 2014. **42**(Database issue): p. D68-73.

355. Jin, Y., et al., *Evaluating the microRNA targeting sites by luciferase reporter gene assay.* Methods Mol Biol, 2013. **936**: p. 117-27.

356. Simon, R., et al., *Analysis of gene expression data using BRB-ArrayTools.* Cancer Inform, 2007. **3**: p. 11-7.

357. Biswas, N.K., et al., *Somatic mutations in arachidonic acid metabolism pathway genes enhance oral cancer post-treatment disease-free survival.* Nat Commun, 2014. **5**: p. 5835.

358. Enright, A.J., et al., *MicroRNA targets in Drosophila.* Genome Biol, 2003. **5**(1): p. R1.

359. Tsang, J.S., M.S. Ebert, and A. van Oudenaarden, *Genome-wide dissection of microRNA functions and cotargeting networks using gene set signatures.* Mol Cell, 2010. **38**(1): p. 140-53.

360. Wong, N. and X. Wang, *miRDB: an online resource for microRNA target prediction and functional annotations.* Nucleic Acids Res, 2015. **43**(Database issue): p. D146-52.

361. Vejnar, C.E., M. Blum, and E.M. Zdobnov, *miRmap web: Comprehensive microRNA target prediction online.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W165-8.

362. Hsu, S.D., et al., *miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes.* Nucleic Acids Res, 2008. **36**(Database issue): p. D165-9.

363. Krek, A., et al., *Combinatorial microRNA target predictions.* Nat Genet, 2005. **37**(5): p. 495-500.

364. Kertesz, M., et al., *The role of site accessibility in microRNA target recognition.* Nat Genet, 2007. **39**(10): p. 1278-84.

365. Kruger, J. and M. Rehmsmeier, *RNAhybrid: microRNA target prediction easy, fast and flexible.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W451-4.

366. Lewis, B.P., C.B. Burge, and D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* Cell, 2005. **120**(1): p. 15-20.

367. Xie, B., et al., *miRCancer: a microRNA-cancer association database constructed by text mining on literature.* Bioinformatics, 2013. **29**(5): p. 638-44.

368. Kessenbrock, K., V. Plaks, and Z. Werb, *Matrix metalloproteinases: regulators of the tumor microenvironment.* Cell, 2010. **141**(1): p. 52-67.

369. Iizuka, S., N. Ishimaru, and Y. Kudo, *Matrix metalloproteinases: the gene expression signatures of head and neck cancer progression.* Cancers (Basel), 2014. **6**(1): p. 396-415.

370. Zhang, X., et al., *Expression of MMP-10 in lung cancer.* Anticancer Res, 2007. **27**(4C): p. 2791-5.

371. Mathew, R., et al., *Stromelysin-2 overexpression in human esophageal squamous cell carcinoma: potential clinical implications.* Cancer Detect Prev, 2002. **26**(3): p. 222-8.

372. Liu, H., et al., *Overexpression of matrix metalloproteinase 10 is associated with poor survival in patients with early stage of esophageal squamous cell carcinoma.* Dis Esophagus, 2012. **25**(7): p. 656-63.

373. Zhang, G., et al., *Matrix metalloproteinase-10 promotes tumor progression through regulation of angiogenic and apoptotic pathways in cervical tumors.* BMC Cancer, 2014. **14**: p. 310.

374. Ziober, A.F., L. D'Alessandro, and B.L. Ziober, *Is gene expression profiling of head and neck cancers ready for the clinic?* Biomark Med, 2010. **4**(4): p. 571-80.

375. Garnett, M.J., et al., *Systematic identification of genomic markers of drug sensitivity in cancer cells.* Nature, 2012. **483**(7391): p. 570-5.

376. Bjorklund, P., G. Akerstrom, and G. Westin, *An LRP5 receptor with internal deletion in hyperparathyroid tumors with implications for deregulated WNT/beta-catenin signaling.* PLoS Med, 2007. **4**(11): p. e328.

377. Goonesekere, N.C., et al., *A meta analysis of pancreatic microarray datasets yields new targets as cancer genes and biomarkers.* PLoS One, 2014. **9**(4): p. e93046.

378. Wang, X.Y., et al., *Meta-analysis of gene expression data identifies causal genes for prostate cancer.* Asian Pac J Cancer Prev, 2013. **14**(1): p. 457-61.

379. Gyorffy, B. and R. Schafer, *Meta-analysis of gene expression profiles related to relapse-free survival in 1,079 breast cancer patients.* Breast Cancer Res Treat, 2009. **118**(3): p. 433-41.

380. Chen, R., et al., *A meta-analysis of lung cancer gene expression identifies PTK7 as a survival gene in lung adenocarcinoma.* Cancer Res, 2014. **74**(10): p. 2892-902.

9. Chapter 9. Appendixes

9.1 Appendix I

Whole exome sequencing data statistics of samples included in TMC-SNPdb.

S.No	ADLAB ID	Tissue Type	Sample type	Library Kit	Number of reads	Mapped reads	Percent Mapped	Mean Coverage
1	AD0697	Cervical	Tissue	SureSelect v5	46358172	30871424	67	93
2	AD0722	Cervical	Tissue	SureSelect v5	37346018	28005048	74	85
3	AD0703	Cervical	Tissue	SureSelect v5	40668628	25570273	63	77
4	AD0704	Cervical	Tissue	SureSelect v5	38010492	24847592	65	75
5	AD0711	Cervical	Blood	SureSelect v5	42307838	29243577	69	88
6	AD0719	Cervical	Blood	SureSelect v5	67952764	46526964	68	141
7	AD0700	Cervical	Tissue	SureSelect v5	90430850	89418505	99	179
8	AD0689	Cervical	Tissue	SureSelect v5	90177536	89229356	99	178
9	AD0693	Cervical	Tissue	SureSelect v5	96782488	95824313	99	192
10	AD0695	Cervical	Tissue	SureSelect v5	54056398	53213454	98	106
11	AD0707	Cervical	Blood	SureSelect v5	82318198	81601296	99	163
12	AD0709	Cervical	Tissue	SureSelect v5	32992728	32733386	99	65
13	AD0714	Cervical	Blood	SureSelect v5	66391960	65873552	99	132
14	AD0716	Cervical	Blood	SureSelect v5	84342838	83428663	99	167
15	AD0717	Cervical	Blood	SureSelect v5	89481162	88588499	99	177
16	AD0691	Cervical	Tissue	SureSelect v5	87667854	86708681	99	173
17	AD0698	Cervical	Tissue	SureSelect v5	77698290	76869536	99	154
18	AD0699	Cervical	Tissue	SureSelect v5	57855550	57156699	99	114
19	**AD0746	Cervical	Blood	TruSeq v2	4229630	3728161	88	6
20	**AD0792	Cervical	Blood	TruSeq v2	2434825	1803554	74	3
21	AD0793	Cervical	Blood	TruSeq v2	49920373	49482640	99	57
22	AD0794	Cervical	Blood	TruSeq v2	44251517	43841324	99	51
23	AD0795	Cervical	Blood	TruSeq v2	25106181	22167990	88	26
24	AD0797	Cervical	Blood	TruSeq v2	51917164	49171428	95	29
25	AD0798	Cervical	Blood	TruSeq v2	23676945	22211029	94	13
26	AD0806	Cervical	Blood	TruSeq v2	35855610	35015679	98	56
27	AD0764	Cervical	Blood	TruSeq v2	37472890	36799409	98	59
28	AD0762	Cervical	Blood	TruSeq v2	31341238	30807228	98	50
29	AD0752	Gallbladder	Tissue	SureSelect v5	63858844	35885515	56	108
30	AD0759	Gallbladder	Tissue	SureSelect v5	43455624	27369663	63	83
31	AD0754	Gallbladder	Tissue	SureSelect v5	54998184	37809365	69	114
32	AD0755	Gallbladder	Tissue	SureSelect v5	66115458	47741170	72	144
33	AD0788	Gallbladder	Tissue	SureSelect v5	39594984	23065796	58	70
34	AD0757	Gallbladder	Tissue	SureSelect v5	51254660	34364520	67	104
35	AD0437	Gallbladder	Tissue	SureSelect v5	44676668	29362140	66	89
36	AD0439	Gallbladder	Tissue	SureSelect v5	49994928	29952529	59	90
37	AD0746	Gallbladder	Tissue	SureSelect v5	37563776	21449942	57	65

38	AD0761	Gallbladder	Tissue	SureSelect v5	43581276	28263519	64	85
39	AD0763	Gallbladder	Tissue	SureSelect v5	37563776	21449942	57	65
40	AD0725	tongue	Tissue	TruSeq v3	33260084	32260230	97	52
41	AD0775	tongue	Tissue	TruSeq v3	28727389	27914222	97	45
42	AD0744	tongue	Tissue	NimbleGen	66015439	63720122	97	101
43	AD0771	tongue	Tissue	TruSeq v3	7380086	5736294	78	9
44	AD0782	tongue	Tissue	NimbleGen	74188836	72453644	98	115
45	AD0777	tongue	Tissue	NimbleGen	46637332	45169670	97	72
46	AD0783	tongue	Tissue	NimbleGen	65866424	64340950	98	102
47	AD0778	tongue	Tissue	NimbleGen	50733602	49388815	97	78
48	AD0779	tongue	Tissue	NimbleGen	34344688	33508795	98	53
49	AD0780	tongue	Tissue	NimbleGen	70964617	69346504	98	110
50	AD0784	tongue	Tissue	NimbleGen	82425599	80259390	97	127
51	AD0731	tongue	Tissue	NimbleGen	69614600	67126889	96	107
52	AD0786	tongue	Tissue	NimbleGen	75480562	73581363	97	117
53	AD0787	tongue	Tissue	NimbleGen	67960054	66555223	98	106
54	AD0781	tongue	Tissue	NimbleGen	80188961	78116302	97	124
55	AD0733	tongue	Tissue	NimbleGen	64055460	61981799	97	98
56	AD0734	tongue	Tissue	NimbleGen	75054746	72899349	97	116
57	AD0720	tongue	Tissue	TruSeq v3	31016587	13746645	44	22
58	**AD0721	tongue	Tissue	TruSeq v3	7161446	6259124	87	10
59	**AD0772	tongue	Tissue	TruSeq v3	12684423	11584458	91	19
60	AD0773	tongue	Tissue	TruSeq v3	76225980	66292936	87	107
61	AD0774	tongue	Tissue	TruSeq v3	32504474	30146066	93	49
62	AD0789	tongue	Tissue	NimbleGen	34939297	34104607	98	54

In the ADLAB_ID column ** asterisk denotes sample with low coverage due to high duplication rate, * due to low data yield by sequencing using ` Illumina GAIIx in separate run.

9.2 Appendix II: Summary of sequencing data QC and variants statistics for each patient.

Sample name	Library Kit	Total bases	Mapped bases	Percent mapped	Coverage
12N	TruSeq v3	3359268484	3258283230	97	27
14N	TruSeq v3	2901466289	2819336422	97	24
17N	NimbleGen v3	6667559339	6435732322	97	53
63N	NimbleGen v3	7493072436	7317818044	98	60
22N	NimbleGen v3	4710370532	4562136670	97	38
71N	NimbleGen v3	6652508824	6498435950	98	53
23N	NimbleGen v3	5124093802	4988270315	97	41
24N	NimbleGen v3	3468813488	3384388295	98	28
25N	NimbleGen v3	7167426317	7003996904	98	57
68N	NimbleGen v3	8324985499	8106198390	97	66
29N	NimbleGen v3	7031074600	6779815789	96	56
62N	NimbleGen v3	7623536762	7431717663	97	61
69N	NimbleGen v3	6863965454	6722077523	98	55
37N	NimbleGen v3	8099085061	7889746502	97	65
38N	NimbleGen v3	6469601460	6260161699	97	52
39N	NimbleGen v3	7580529346	7362834249	97	60
8N	TruSeq v3	7698823980	6695586536	87	62
9N	TruSeq v3	3282951874	3044752666	93	27
66N	NimbleGen v3	3528868997	3444565307	98	28
3N	TruSeq v3	3132675287	1388411145	44	26
5N	TruSeq v3	723306046	632171524	87	6
7N	TruSeq v3	1281126723	1170030258	91	11
1N	TruSeq v3	745388686	579365694	78	6
2T	TruSeq v3	4608158734	4400958951	95	71
12T	TruSeq v3	2410031094	2331574496	96	19
14T	NimbleGen v3	3411610017	3328251384	97	28
17T	NimbleGen v3	10606842442	10354353855	97	86
63T	NimbleGen v3	7052560330	6869567520	97	57
22T	NimbleGen v3	5653930611	5245857887	92	44
71T	NimbleGen v3	3137799320	3078690989	98	26
23T	NimbleGen v3	9708933555	9479697289	97	78
24T	NimbleGen v3	4014134304	3883930558	96	32
25T	NimbleGen v3	8558331556	8331992980	97	69
68T	NimbleGen v3	6933522134	6781843061	97	56
29T	NimbleGen v3	5894061848	5711216498	96	47
62T	NimbleGen v3	6702323943	6469267352	96	54
69T	NimbleGen v3	4424259045	4335696387	97	36
37T	NimbleGen v3	9019949733	8739416072	96	72
38T	NimbleGen v3	7261534380	7010166994	96	58
39T	TruSeq v3	7900900437	7712882978	97	64
8T	TruSeq v3	5747453278	4789609476	80	40
9T	NimbleGen v3	2560353030	2490016832	97	21
66T	TruSeq v3	7283377044	7126042476	97	59
1T	TruSeq v3	457863603	384326210	83	3
3T	TruSeq v3	741016396	631637638	84	5
5T	TruSeq v3	1055828245	886770506	83	8
7T	TruSeq v3	1480092885	1051300920	66	9

9.3 Appendix III: TMC-SNPdb –Variant subtraction tool user manual

Dependencies:

TMC-SNPdb subtraction tool depends on three python libraries sqlite3, Tkinter (for GUI version) and pyvcf (version >= 0.6.7). These libraries can be installed using following commands:

Get and Install sqlite

sqlite3 (http://www.sqlite.org/download.html)

wget http://www.sqlite.org/sqlite-autoconf-3070603.tar.gz

tar xvfz sqlite-autoconf-3071502.tar.gz

cd sqlite-autoconf-3071502

./configure --prefix=/usr/local

make

sudo make install

Get and Install PyVcf

pyvcf (http://pyvcf.readthedocs.org/en/latest/)

wget https://pypi.python.org/packages/source/P/PyVCF/PyVCF-0.6.7.tar.gz

tar xvfz PyVCF-0.6.7.tar.gz

sudo python setup.py install

Get and Install Python TKinter

python TKinter library for executing the GUI mode

sudo apt-get install python-tk

Installation:

To install 'tmc-snpdb' on your Linux system, please follow step 1 & 2.

1) Untar TMC-SNPdb package using following command:

> tar xvf tmcsnpdb1.0.tar.gz

2) Run the INSTALL script (as administrator) on your system

> chmod +x INSTALL

> sudo sh INSTALL

3) Subtraction tool runs in two modes, Graphical user interface (GUI) and command line (CMD). Use 'tmc-snpdb' command on the terminal to execute it in CMD mode; Use 'tmc-snpdb-gui' to run the subtraction tool in GUI mode.

1) Running Subtraction program in Command line mode:

For CMD mode use the following command:

$ tmc-snpdb

Usage: tmc-snpdb [options]

Options:

-h, --help show this help message and exit

-i Input tumor VCF file (required)

-o Output file after TMC-SNPdb subtraction (Optional)

-l Load a custom normal variations database (with tmc-snpdb schema - refer README/ SCHEMA document in the package). Give SQLITE file as input (Optional)

--vcf_dir Directory path containing VCF files containing germline variants. Required for creating custom database.

--sql Output germline variant database. Required for creating custom database.

Example:

*use the help option

tmc-snpdb -h

or

tmc-snpdb --help

Subtraction of user tumor vcf against TMC-SNPdb.

*To subtract germline variants from tumor VCF using TMC-SNPdb, use the command (without specifying output file name).

tmc-snpdb –i input.vcf –o output.vcf (Optional)

Create a custom germline database:

User can create their own germline database with a set of normal/germline variant VCF files. The following command can be used;

$ tmc-snpdb --vcf-dir directory/path/to/vcf --sql output.sqlite

The output SQLite file is created with the following schema:

chr - holds chromosome number data, e.g. 'chr2'

pos - integer storing the position on the chromosome, e.g. 190023

orig – reference base as found in the reference genome, e.g. 'A'

change - altered base found in the samples, e.g. 'T'

reccur - recurrence of a particular change (A->T as in the above example) in across samples

Schema to create a custom database

Following is sql syntax could be used to create custom sqlite database file format

```
CREATE TABLE tmcsnpdb
( "chr" TEXT,
        "pos" INTEGER,
        "orig" TEXT,
        "change" TEXT,
        "reccur" INTEGER
);
```

Output from this program can be loaded for subtraction from tumor samples using the '-l' option of the program.

For example:

$ tmc-snpdb -i tumor.vcf -l custom_database.sql

Test Run:

Test VCF files are provided in the "data/test_vcf" directory. This directory contains two files; test1.vcf (227779 variants) and test2.vcf (115884 variants)

Command line mode:

To subtract germline variants from tumor VCF using TMC-SNPdb, use the command

$ tmc-snpdb –i data/test_vcf/test1.vcf –o test1_output.vcf

'-o' - output is an optional argument

Output file after subtraction from test1_output.vcf and test2_output.vcf will contains 224330 and

114388 variants, respectively.

In GUI mode:

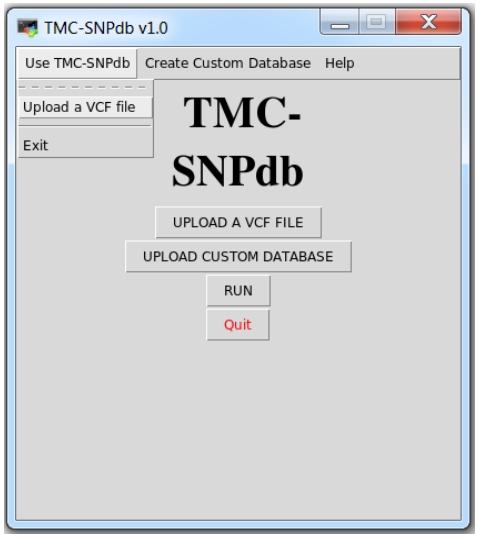
Click on "UPLOAD A VCF FILE", select a file in /data/test_vcf

Directory in the file dialogue box. Then click on "RUN" to subtract variants.

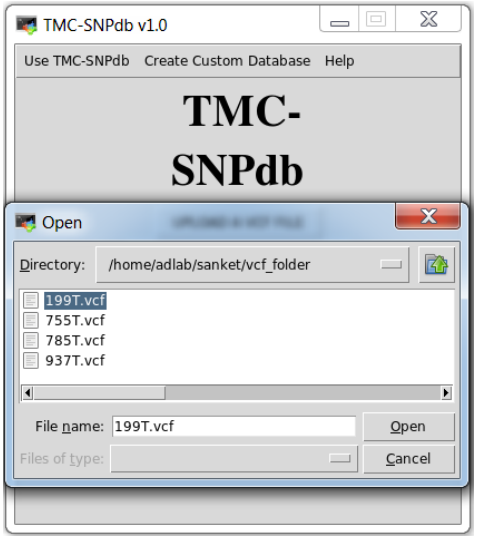
On an Intel-i5-3210M CPU @2.5GHz x 4-32 bit Ubuntu (14.04) system with 8 GB RAM takes 72 minutes to process test1.vcf and 56 minutes to process test2.vcf.

2) Running Subtraction program in Graphical User Interface (GUI) mode:

Menu to upload VCF file

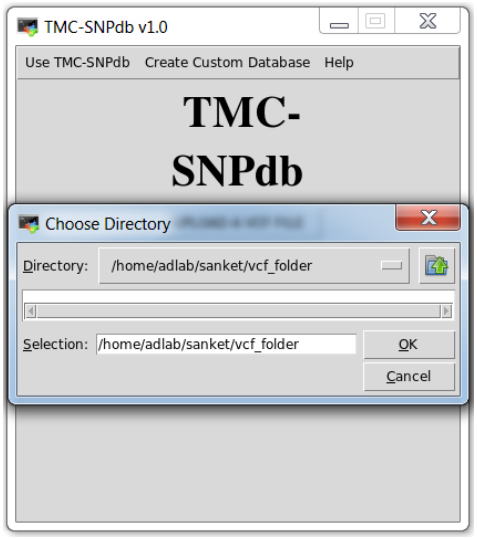


Select VCF file to upload

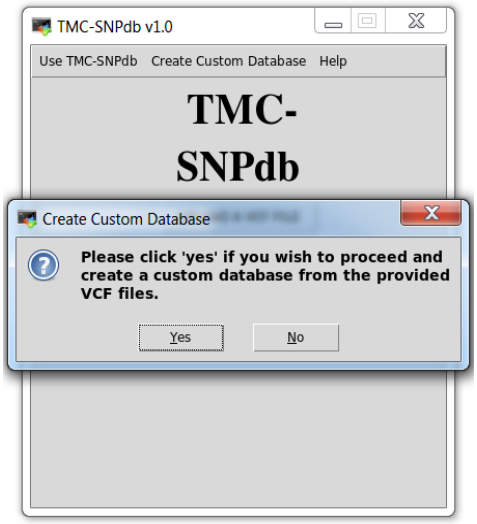


Menu to upload VCF directory

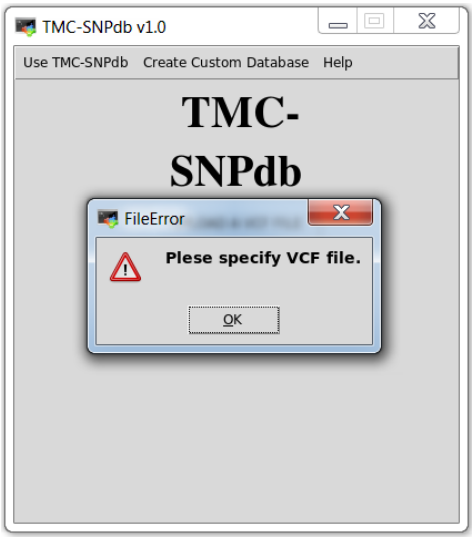
Select a directory for creating custom germline database



Confirm to create database



Error message if vcf files not selected



9.4 Appendix IV: List of Notch pathway gene (n=48) and mutations identified in the study.

S.no.	Gene Name	Mutation	Number of samples mutated	Functional Class
1	*JAG1*	None	0	Ligand
2	*JAG2*	None	0	Ligand
3	*DLL1*	None	0	Ligand
4	*DLL3*	None	0	Ligand
5	*DLL4*	p.R100G	69T,9T	Ligand
6	*NOTCH1*	p.R780Q	70T	receptor
		p.D573A	8T	
		p.T859P	12T	
		p.A465T	38T	
		p.C554*	24T	
7	*NOTCH2*	None	0	receptor
8	*NOTCH3*	None	0	receptor
9	*NOTCH4*	p.LLLLL12fs	23T	receptor
10	*DVL1*	None	0	Negative regulator
11	*DVL2*	None	0	Negative regulator
12	*DVL3*	None	0	Negative regulator
13	*EP300*	None	0	Coactivator
14	*MAML1*	None	0	Coactivator
15	*MAML2*	None	0	Coactivator
16	*MAML3*	p.QQ508fs	25T	Coactivator
17	*SNW1*	None	0	Coactivator
18	*CREBBP*	None	0	Coactivator
19	*HDAC1*	None	0	Co-repressor
20	*HDAC2*	None	0	Co-repressor
21	*NCOR2*	None	0	Co-repressor
22	*NUMB*	None	0	NOTCH1 negative regulator
23	*NUMBL*	None	0	NOTCH1 negative regulator
24	*KAT2A*	None	0	Transcription Regulator
25	*KAT2B*	None	0	Transcription Regulator
26	*DTX1*	p.D122E	2T	NOTCH1 Positive regulator
27	*DTX2*	None	0	NOTCH1 Positive regulator
28	*DTX3*	None	0	NOTCH1 Positive regulator
29	*DTX3L*	None	0	NOTCH1 Positive regulator
30	*DTX4*	None	0	NOTCH1 Positive regulator
31	*LFNG*	None	0	Modifier (likely positive regulator)
32	*MFNG*	None	0	Modifier (likely positive regulator)

33	*RFNG*	None	0	Modifier (likely positive regulator)
34	*PSEN1*	None	0	Secretase complex(receptor proteolysis)
35	*PSEN2*	None	0	Secretase complex(receptor proteolysis)
36	*PSENEN*	None	0	Secretase complex(receptor proteolysis)
37	*NCSTN*	None	0	Secretase complex(receptor proteolysis)
38	*APH1A*	None	0	Secretase complex(receptor proteolysis)
39	*ADAM17*	None	0	receptor proteolysis
40	*RBPJ*	p.M1R	70T	Transcription factor
41	*RBPJL*	None	0	Transcription factor
42	*CIR1*	None	0	Transcription factor
43	*HES1*	None	0	Downstream effector
44	*HES5*	None	0	Downstream effector
45	*HEY1*	None	0	Downstream effector
46	*PTCRA*	None	0	Others
47	*HEYL*	None	0	Downstream effector
48	*HEY2*	None	0	Downstream effector

9.5 Appendix V: List of commonly up regulated genes in at least two datasets identified from meta-analysis.

S.No.	Dataset ID	Number of data sets	Hugo Symbol	Number of genes
1	This study,GSE34105	2 data sets	FCRL5,CD19,SELL,CD177,TIGIT,PIM2,CD180,SPN,ZC3H12D,POU2AF1,PIK3CG,PTPRC,LCP2,SELPLG,LGALS3BP,COTL1,KIRREL,IL2RG,HLA-DQA1,CLSTN3	335
2	This study,GSE13601		DEFB4A,PI3,LGALS7,SERPINB3,GJB3,GZMB,PDZK1IP1,IGSF3,LAD1,GJA1,CLCA2,GRN,SERPINB5,ELF4,TUBB2A,AHNAK2,ANXA8,EREG,PHLDA2,PNP,DUSP14,PLCB3,STK17B,BAK1,CTPS	
3	This study,GSE31056		S100A7A,CNTNAP2,GINS4,IL36G,SLC38A5,KYNU,FERMT1,FAT1,F2RL1,FOSL1,TLR2,EXT1,FJX1,DSG2,FADD,ECT2,TNFAIP3,MARCKSL1,LMNB2,ATAD2,PCYT1A,MCM5	
4	This study,TCGA		MAGEA3,MAGEB2,AQP9,DLGAP3,CXCR1,NXPH4,LILRB4,GRIN2D,KIF26B,TSPAN10,CPZ,TUBB3,CNGB1,PLA2G2F,IL11,WNT7B,GPRIN1,LRRC15,EPHB2,C6orf141,RAC2,HES2,SDK2,KIAA1644,WFDC5,GPR176,ODZ2,KIFC1,LAMA1,C19orf21,ADAMTS15,SLC11A1,CDT1,THBS2,PRSS23,CD276,GNA12,GLT25D1	
5	GSE34105,GSE13601		GBP1,TNFSF10,GPR183	
6	GSE34105,GSE31056		POPDC3,GALNT10,MDK,OLR1,MUCL1,SERPINA1,PSMB9,HIST1H2BF	
7	GSE34105,TCGA		NLRP10,ORC1L,CENPI,GBX2,LTBP1,SLC6A2,FOXC2,HTRA4,HTR7,ADAM8,CNTD2,NAGS,STAG3,HOXA11AS,HIST1H3H,KREMEN2,FSD1,NCRNA00152,VSIG1,TRIML2,TMEM92,DHRS2,NFKBIL2,HOXC8,FIBCD1,GPR158,LOC100216001,HOXC10,AMTN,STX1A,RTP3,MGC45800,USP18,NFE2,CDKN2A,LHX1,HOXD13,RAG1,ZIC5,PAEP,FCRL3,CA9,ZBP1,CLEC12A,CSAG1,GFI1,ATP6V0D2,COL22A1,LILRA3,SPOCD1,FCGR3A,FADS2,MAGED4,FAM176A,LAIR2,CDX1,ZBTB32,DDX60L,DNAH17,HOXC9	
8	GSE13601,GSE31056		CHEK1,LAMB3,ACTN1,TP63,NUP155,CCL20,RGS20,COL17A1	
9	GSE13601,TCGA		CDC25B,CTSC,CTSL2,COL3A1,KIAA0101	
10	GSE31056,TCGA		CDCA2,FADS1,CDH11,APCDD1L,HAS2,PLAC1,FCGR3B,NCAPG,SPRY4,PSRC1,CSPG4,HERC5,RCN3,OLFML2B,VEGFC,KDELC1,E2F7,FEZ1,CDK1,TPBG,GPX8,BEND6,HAPLN1,IFIT1,IKBIP,TPST1,EMILIN1,SLAMF8,NEK2,DEPDC1B,FSTL3,NEIL3,TTK,CHIT1,PLOD1,RFC4,C4orf48,BUB1B,NTM,NUF2,APOE,MICAL2,PHLDB2,MKI67,PLK1,KDELR3,RAD54B,DDX58,STC2,CD300LF,RAD51AP1,C16orf74,SH2D5,IFI44L,FZD2,IL12RB2,ARL14,ASPRV1,PCDH17,SPC25,EN1,FAM83A,HSF2BP,FNDC1,NID2,WDR54,SLC16A1,EGFL6,LEPRE1,PLOD2,GINS1,TMEFF1,EIF5A2,IFI44,IFI35,FAM64A,HMMR,KIF18A,FST,CPXM1,MELK,MCM10,MTHFD1L,DTL,BCAT1,DEPDC1,CENPA,OAS2,DLGAP5,CMPK2,KIF20A,AMIGO2,CCNA2,ELAVL2,RGS4,SERPINH1,CENPE,NRIP3,ANO1,COL5A3,COL7A1,FN1,COL12A1,AURKA,BIRC5,CDC45,EXO1,SHCBP1,COL1A1,C1QTNF6,NOX4,NFE2L3,HOMER3,HJURP,CEP55,CKAP2L,TRIP13,KIF14,DNMT3B,CYP27B1,ESM1,KANK4,GBP5,FOXM1,TNFSF4,SKA1,PDPN,WNT2,IGF2BP3,PMEPA1,SULF1,HOXA1,AURKB,HOXB7,EPSTI1,COL5A1,MFAP2,DFNA5,SCG5,DCBLD1,IGFL2,PPAPDC1A,WDR66,HOXC6,CTHRC1,COL10A1	
11	This study,GSE34105,GSE31056	3 data sets	CST1,CXCL6,IL7R,UBD,CXCL1	118
12	This study,GSE34105,TCGA		SDS,TNFRSF9,IL2RA,ITGAX,CD80,CTLA4,ICOS,CASP14,TMEM132A	
13	This study,GSE13601,GSE31056		KRT75,LAMP3,PYGL,MYO10	
14	This study,GSE13601,TCGA		CCL18,PTGS2,BASP1	
15	This study,GSE31056,TCGA		CALB1,CDSN,PCSK9,GALNT6,KHDC1L,GREM1,ULBP2,SLC2A1,MYBL2,HOXC13,ASF1B,OAS3,TTYH3,CYP27C1,ACP5,NETO2,KIF4A,CHST11,LEPREL1,THY1,CDCA8,TNFAIP6,TGFB1,COL4A2,PLEK2,**MMP14**,COL4A1,LOXL2,NCAPH,ZNF469,TK1,ANLN,TOP2A,CENPF,CDCA5,FBN2,CDC6,TNFRSF12A,KIF2C,PXDN,IFIT2,PARP12,LPCAT1	
16	GSE34105,GSE13601,TCGA		LAMA3	
17	GSE34105,GSE31056,TCGA		AIM2,RTP4,ZNF114,ZIC2,WDHD1,HSD17B6,COL4A6,SNX10,APOC1,GNLY,SLCO1B3,ITGA3,PPP4R4,HOXD10,ADAM12,IL24,TREM1,ADAMDEC1,GRP,IFIT3,NRG1,SOX11,OASL,HOXC4,INHBA,TDO2,CSF2	
18	GSE13601,GSE31056,TCGA		PTK7,ITGA6,MICB,IFI6,KRT17,XAF1,SERPINE2,FAP,COL1A2,COL5A2,NDC80,APOBEC3B,SERPINE1,CXCL9,PLAU,CDKN3,MYO1B,UBE2C,FSCN1,PTHLH,LAMC2,TGFBI,ISG15,IFI27,RBP1,**MMP3**	
19	This study,GSE34105,GSE31056,TCGA	4 data sets	CCL11,**MMP7**,FOLR3,**MMP11**,TREM2,PI15,CDCA3,IDO1,DDX60,SPAG5	35
20	This study,GSE13601,GSE31056,TCGA		TNC,IFI30,CHST2,CDH3,BNC1,BUB1,CDC20,NEFL,CCNB2,TPX2,CCNB1,SNAI2,	
21	GSE34105,GSE13601,GSE31056,TCGA		COL11A1,KIF23,BST2,TYMP,**MMP1**,APOL1,POSTN,DKK1,CXCL10,CXCL11,**MMP10**,PLA2G7,SPP1	
22	This study,GSE34105,GSE13601,GSE31056,TCGA	5 data sets	CXCL13,**MMP12**,**MMP9**,**MMP13**,RSAD2	5

Because below appendixes were large in size so they have been hyperlinked to google drive link and can be assessed from there.

9.6 Appendixes V: List of deleterious non-silent mutations identified from exome sequencing across 19 TSCC tumors.

9.7 **Appendix VI:** List of genes with DNA copy number gains in TSCC patients.

9.8 **Appendix VII:** List of genes with DNA copy number losses in TSCC patients.

9.9 Appendix VIII: List of significantly differentially expressed genes identified in HPV-negative early tongue tumors.

9.10 Appendix IX: List of gene sets identified for deregulated genes in HPV-negative early tongue tumors.

9.11 **Appendix X:** Detailed list of transcript fusions identified in tongue tumors.

9.12 Appendix XI: Clinicopathologic features correlation analysis of fusion transcripts.

9.13 **Appendix XII:** List of expressed genes in TSCC tumors.

Chandrani et al. BMC Genomics (2015) 16:936
DOI 10.1186/s12864-015-2138-4

BMC Genomics

RESEARCH ARTICLE Open Access

Integrated genomics approach to identify biologically relevant alterations in fewer samples

Pratik Chandrani[1], Pawan Upadhyay[1], Prajish Iyer[1], Mayur Tanna[1], Madhur Shetty[1], Gorantala Venkata Raghuram[1], Ninad Oak[1], Ankita Singh[1], Rohan Chaubal[1], Manoj Ramteke[1], Sudeep Gupta[2] and Amit Dutt[1*]

Abstract

Background: Several statistical tools have been developed to identify genes mutated at rates significantly higher than background, indicative of positive selection, involving large sample cohort studies. However, studies involving smaller sample sizes are inherently restrictive due to their limited statistical power to identify low frequency genetic variations.

Results: We performed an integrated characterization of copy number, mutation and expression analyses of four head and neck cancer cell lines - NT8e, OT9, AW13516 and AW8507– by applying a filtering strategy to prioritize for genes affected by two or more alterations within or across the cell lines. Besides identifying *TP53*, *PTEN*, *HRAS* and *MET* as major altered HNSCC hallmark genes, this analysis uncovered 34 novel candidate genes altered. Of these, we find a heterozygous truncating mutation in Nuclear receptor binding protein, *NRBP1* pseudokinase gene, identical to as reported in other cancers, is oncogenic when ectopically expressed in NIH-3 T3 cells. Knockdown of *NRBP1* in an oral carcinoma cell line bearing *NRBP1* mutation inhibit transformation and survival of the cells.

Conclusions: In overall, we present the first comprehensive genomic characterization of four head and neck cancer cell lines established from Indian patients. We also demonstrate the ability of integrated analysis to uncover biologically important genetic variation in studies involving fewer or rare clinical specimens.

Background

Head and neck squamous cell carcinoma (HNSCC) is the sixth-most-common cancer worldwide, with about 600,000 new cases every year, and includes cancer of the nose cavity, sinuses, lips, tongue, mouth, salivary glands, upper aerodigestive tract and voice box [1]. Recent large scale cancer genome sequencing projects have identified spectrum of driver genomic alterations in HNSCC including *CDKN2A*, *TP53*, *PIK3CA*, *NOTCH1*, *HRAS*, *FBXW7*, *PTEN*, *NFE2L2*, *FAT1*, and *CASP8* [2–4]. These landmark studies apply elegant statistical methodologies like MutSig [5], Genome MuSiC [6], Intogen [7], InVEx [8], ActiveDrive [9] and GISTIC [10] in identifying significantly altered genes across large sample cohorts by comparing rate of mutations of each gene with background mutation rate to determine an unbiased enrichment– a minimum ~150 patients or higher is required for identification of somatic mutations of 10 % population frequency in HNSCC [11]. These genome-wide analysis may not be directly applicable for studies involving fewer or rare clinical specimen that are inherently restrictive due to the limited statistical power to detect alterations existing at lower frequency.

On the other hand, given that a cancer gene could be selectively inactivated or activated by multiple alterations, an integrative study design performed by combining multiple data types can potentially be helpful to achieve the threshold for statistical significance for studies involving fewer or rare clinical specimen. For example, a tumor suppressor gene– deleted in 1 % of patients, mutated in another 3 %, promoter-hypermethylated in another 2 % and out of frame fused with some other chromosomal region in 2 %– may be considered to be altered with a cumulative effect of 8 %

* Correspondence: adutt@actrec.gov.in
[1]Advanced Centre for Treatment, Research and Education in Cancer, Tata Memorial Center, Navi Mumbai, Maharashtra 410210, India
Full list of author information is available at the end of the article

based on integrative analysis [12, 13]. Combinatorial sources of genetic evidence converging at same gene or signalling pathway can also limit false positives by filtering strategy and potentially reducing the multiple hypothesis testing burden for identification of causal genotype-phenotype associations [14]. Using similar approaches for posterior refinement to indicate positive selection, Pickering et al. identified four key pathways in oral cancer by integrating methylation to copy number variation and expression [15]; and, more recently, Wilkerson et al. proposed superior prioritisation of mutations based on integrated analysis of the genome and transcriptome sequencing than filtering based on conventional quality filters [16]. These and several other reports all together emphasize integration of multi-platform genomic data for identification of cancer related genes [17].

Here, we perform characterization of four head and neck cancer cell lines, established from Indian head and neck cancer patients, using classical cytogenetic approach, SNP arrays, whole exome and whole transcriptome sequencing. Next, we apply the widely used posterior filtering strategy of results obtained from genome wide studies to effectively reduce the amount of data obtained from individual platforms. Adopting such an integrative approach allow us to identify biological relevant alterations affected by two or more events even from fewer samples.

Methods

Cell culturing and single cell dilution for establishing clonal cells

Four HNSCC tumor cell lines established within Tata Memorial Center from Indian patients and described before were acquired: NT8e, OT9, AW13516, AW8507 [18, 19]. All the cell lines were maintained in DMEM media (Gibco, USA). For clonal selection, growing culture was trypsinized and diluted as 1 cell per 100 ml of media and dispensed in a 96 well plate with follow up subculture of clones that survived.

SNP array analysis

Genomic DNA was extracted from pre-clonal and clonal cell lines using PAXgene Tissue DNA Kit (Qiagen, USA). 200 ng of good quality DNA from each sample was submitted to Sandor Proteomics (Hyderabad, India) for sample preparation and genome wide SNP array using Illumina Infinium assay (Human660W-quad BeadArray chip) following manufacturer's standard protocol. Array data was pre-processed using GenomeStudio (Illumina Inc., USA) for quality control check. To retain only good quality genotyping calls, a threshold GenCall score of 0.25 was used across all samples. A total of 396, 266 SNPs were retained after this filtering. These SNPs were then used for copy number analysis using Genome Studio plug-in cnvPartition 3.2 and an R package Genome Alteration Print (GAP) [20]. Inferred copy numbers were then annotated with genomic features using BedTools (v. 2.17.0) [21]. Copy number segments of more than 10 Mb in size were classified as arm-level amplifications and were identified as non-significant alterations. Focal amplifications (less than 10 Mb) were used for further analysis.

Cytogenetic karyotyping

Cells grown in complete media (60–70 % confluent) were treated with colcemid (0.1 ng/ul, Sigma, USA) to arrest them in metaphase. After incubation of 6 h at 37oC and trypsinisation, cells were washed with pre-warmed KCl (0.075 M) (Sigma, USA) and incubated with KCl at 37 °C in water bath for 60 min. After the incubation is over, cells were fixed with Carnoy's fixative solution on pre chilled microscopic glass slides (chilled in alcohol) by pipette around 70 μl of cell suspension, drop by drop from height (50 cm). Slides were kept on the water bath at 70 °C for few seconds followed by drying on heating block (set at 80 °C). Metaphase of cells was confirmed by observing chromosomes using a phase contrast inverted microscope (Zeiss, USA). Confirmed metaphase captured cells were aged by keeping the slides at 60 °C for 3 h followed by trypsin digest (Trypsin/EDTA - concentration of 0.025 %, Sigma, USA). Giemsa stain (Sigma, USA) (3 %) was applied using coplin Jar for 15 min on slides followed by washing with distilled water.

Exome sequencing

Exome enrichment was performed using manufacturer's protocol for Illumina TruSeq exome enrichment kit in which 500 ng of DNA libraries from six samples were pooled to make total 3 μg DNA mass from which 62 MB of targeted exonic region covering 20,976 genes was captured. Exome enriched library was quantified and validated by real-time PCR using Kappa quantification kit at the Next-Generation Genomics Facility (NGGF) at Center for Cellular and Molecul ar Platforms (CCAMP, India). Whole exome libraries of AW13516, AW8507 and OT9 were loaded onto Illumina HiSeq 1000 for 2 X 100 bp paired-end sequencing with expected coverage of ~ 100 X. NT8e cell line was sequenced with 2 X 54 bp paired-end and 2 X 100 bp paired-end sequencing. Raw sequence reads generated were mapped to NCBI human reference genome (build GRCh37) using BWA v. 0.6.2 [22]. Mapped reads were then used to identify and remove PCR duplicates using Picard tools v. 1.74 (http:broadinstitute.github.io/picard/). Base quality score recalibration and indel re-alignment were performed and variants were called from each cell line separately using GATK v. 1.6-9 [23, 24] and MuTect

v. 1.0.27783 [25]. All the variants were merged and dumped into local MySQL database for advanced analysis and filtering. We used hard filter for removing variants having below 5X coverage to reduce false positives. For cell lines we use dbSNP (v. 134) [26] as standard known germline variants database and COSMIC (v. 62) [27] as standard known somatic variants database. Variants identified in cell lines, which are also there in dbSNP but not in COSMIC were subtracted from the database. Remaining variants were annotated using Oncotator (v. 1.0.0.0rc7) [28], and three functional prediction tools PolyPhen2 (build r394) [29], Provean (v. 1.1) [30] and MutationAccessor (release 2) [31]. Variants found deleterious by any two out of three tools were prioritized. Variants having recurrent prediction of deleterious function were prioritized. Variants from exome sequencing were compared to variants identified from transcriptome sequencing for cross-validation using in-house computer program.

Transcriptome sequencing

Transcriptome libraries for sequencing were constructed according to the manufacturer's protocol. Briefly, mRNA was purified from 4 μg of intact total RNA using oligodT beads (TruSeq RNA Sample Preparation Kit, Illumina). 7 pmol of each library was loaded on Illumina flow cell (version 3) for cluster generation on cBot cluster generation system (Illumina) and clustered flow cell was transferred to Illumina HiSeq1500 for paired end sequencing using Illumina paired end reagents TruSeq SBS Kit v3 (Illumina) for 200 cycle. De-multiplexing was done using CASAVA (version 1.8.4, Illumina). Actively expressed transcripts were identified from sequencing data by aligning them to the reference genome hg19 using Tophat (v. 2.0.8b) [32] and quantifying number of reads per known gene using cufflinks (v.2.1.1) pipeline [33]. All the transcripts were then binned by $\log_{10}(\mathrm{FPKM} + 1)$ to differentiate the significantly expressed transcripts from the background noise. Since paired normal of these cell lines cannot be obtained, we defined significant change in expression for those genes whose expression is higher (>60 %) or lower (<40 %) than the median expression as suggested in [34]. Gene set enrichment was performed by submitting actively expressed transcripts lists to MSigDB V4 [35] and filtering resulting gene lists by *p*-value of enrichment. Variants were identified from transcriptome sequencing using GATK [23, 24]. Only variants having overlap with exome sequencing were considered as true genomic variants. Fusion transcripts were identified using ChimeraScan (v.0.4.5) [36]. Candidate fusion events supported by minimum 10 read pairs were used for integration and visualization in Circos plot.

Integrated analysis

Genes identified to be altered by SNP array, transcriptome sequencing and exome sequencing were then used for integrative analysis to prioritize the genes which are harbouring multiple types of alteration in same or different cell line. Gene level converging of genomic data were emphasized in identification of biologically relevant alterations across platform and samples. Taking this into consideration, we designed gene prioritization based on three steps: 1) selection of genes harbouring positive correlation of focal copy number and gene expression; 2) selection of genes harbouring point mutations with detectable transcript and or altered copy number, and 3) selection of genes harbouring multiple type of alterations identified from above two gene lists (Additional file 1: Figure S7). Circos plot representation of integrated genomics data was generated using Circos tool (v. 0.66) [37].

Sanger sequencing validation

PCR products were purified using NucleoSpin Gel and PCR Clean-up kit (MACHEREY-NAGEL) as per manufacture's protocol and quantified using Nano-Drop 2000c Spectrophotometer (Thermo Fisher Scientific Inc.) and submitted for sequencing in capillary electrophoresis 3500 Genetic Analyzer (Life Technologies). Sanger sequencing traces were analysed for mutation using Mutation Surveyor [38]. The details of all the primers used for mutation analysis have been provided in Additional file 2: Table S7.

DNA copy number validation

Quantitative-real time PCR and data analysis was performed using Type-it® CNV SYBR® Green PCR (cat. No. 206674) as per manufacturer's instructions on 7900HT Fast Real-Time PCR System. The details of all the primers used for DNA copy number analysis have been provided in Additional file 2: Table S8.

RNA extraction, cDNA synthesis, quantitative real time PCR

Total RNA was extracted from cell lines using RNeasy RNA isolation kit (Qiagen) and Trizol reagent (Invitrogen) based methods and later resolved on 1.2 % Agarose gel to confirm the RNA integrity. RNA samples were DNase treated followed (Ambion) by first strand cDNA synthesis using Superscript III kit (Invitrogen) and semi-quantitative evaluative PCR for GAPDH was performed to check the cDNA integrity. cDNA was diluted 1:10 and reaction was performed in 10 μl volume in triplicate. The melt curve analysis was performed to check the primer dimer or non-specific amplifications. Real-time PCR was carried out using KAPA master mix (KAPA SYBR® FAST Universal q PCR kit) as per manufacturer's instructions in triplicate on

7900HT Fast Real-Time PCR System. All the experiments were repeated at least twice independently. The data was normalized with internal reference *GAPDH*, and analysed by using delta-delta Ct method described previously [39]. The details of all the primers used for expression analysis have been provided in Additional file 2: Table S8.

Generation of p BABE-*NRBP1*-PURO constructs

The cDNA of Human *NRBP1* was amplified from AW13516 cell line using Superscript III (Invitrogen, cat no 18080–093) in a TA cloning vector pTZ57R/T(InsTAclone PCR cloning kit, K1214, ThermoScientific), later site-directed mutagenesis was done using QuikChange II Site-Directed Mutagenesis Kit (cat.no. 200523) as per manufacturer's instructions. Later both wild type and mutant *NRBP1* cDNA sequenced confirmed using Sanger sequencing and were sub-cloned in to retroviral vector *p* BABE-puro using restriction digestion based cloning (SalI and BamHI).

Generation of stable clone of NIH-3 T3 overexpressing *NRBP1*

Two hundred ninety-three T cells were seeded in 6 well plates one day before transfection and each constructs (pBABE-puro) along with pCL-ECO helper vector were transfected using Lipofectamine LTX reagent (Invitrogen) as per manufacturer's protocol. Viral soup was collected 48 and 72 h post transfection, passed through 0.45 μM filter and stored at 4 °C. Respective cells for transduction were seeded one day before infection in six well plate and allowed to grow to reach 50–60 % confluency. One ml of the virus soup (1:5 dilution) and 8ug/ml of polybrene (Sigma) was added to cells and incubated for six hours. Cells were maintained under puromycin (Sigma) selection.

shRNA mediated knockdown of *NRBP1* in HNSCC cells

We retrieved shRNA sequences targeting human NRBP1 from TRC (The RNAi Consortium) library database located in sh1 (3′ UTR) and sh2 (CDS). Target sequences of NRBP1 shRNA constructs: sh1 (TRCN0000001437), 5′-CCCTCTGCACTTTGTTTACTTCT-3′; sh2 (TRCN0000001439), 5′-TGTCGAGAAGAGCAGAAGAATCT-3′. shGFP target sequences is 5′-GCAAGCTGACCCTGAAGTTCAT-3′. p-LKO.1 GFP shRNA was a gift from David Sabatini (Addgene plasmid # 30323) [40]. Cloning of shRNA oligos were done using AgeI and EcoRI restriction site in p-LKO.1 puro constructs. Bacterial colonies obtained screened using PCR and positive clone were sequence verified using Sanger sequencing. Lentiviral production and stable cell line generation performed as described earlier [41]. In brief, Lentivirus were produced by transfection of shRNA constructs and two helper vector in 293 T cells as described [42]. Transduction was performed in HNSCC cells by incubating for 6 h in presence of 10 μg/ml polybrene and post infection media was replaced with fresh media. Puromycin selection was performed two days post infection in presence of 1 μg/ml. Puromycin selected cells were harvested and total cell lysate prepared and expression of NRBP1 was analysed using anti-NRBP1 antibody (Santa Cruz Biotechnology; sc-390087) and GAPDH (Santa Cruz Biotechnology; sc-32233).

Soft Agar colony formation assay

The cells were harvested 48 h after transfection, and an equal number of viable cells were plated onto soft agar after respective treatments for determination of anchorage-independent growth. For analysis of growth in soft agar, 5×10^3cells were seeded in triplicate onto a six well dish (Falcon) in 3 ml of complete medium containing 0.33 % agar solution at 37 °C. Cells were fed with 500 μl of medium every 2 days. From each well randomly 10 field images were taken using Phase contrast Inverted microscope (Zeiss axiovert 200 m) and colonies were counted manually.

Growth curve analysis - 25,000 cells/well were seeded in 24 well plates and growth was assessed post day 2, 4 and 6 by counting the cells using a haemocytometer. Percent survival were plotted at day 4 relative to day 2 and later normalized against scrambled or empty vector control.

Western blot analysis

Cells were lysed in RIPA buffer and protein concentration was estimated using BCA method [43]. 50 and 100 μg protein was used for NIH-3 T3 and HNSCC cell lines western analysis. The protein was separated on 10 % SDS-PAGE gel, transfer was verified using Ponceau S (Sigma), transferred on nitrocellulose membrane and blocked in Tris-buffered saline containing 5 % BSA (Sigma) and 0.05 % Tween-20(Sigma). Later, blots were probed with anti-NRBP1 (Santa Cruz Biotechnology; sc-390087), anti-total ERK1/2 (Cell signaling; 4372S), anti-Phospho ERK1/2 (Cell signaling; 4370S) and anti- GAPDH antibody (Santa Cruz Biotechnology; SC-32233). The membranes were then incubated with corresponding secondary HRP-conjugated antibodies (Santa Cruz Biotechnology, USA) and the immune complexes were visualized by Pierce ECL (Thermo Scientific, USA) according to manufacturer's protocol. Western blot experiments were performed as independent replicates.

Statistical analysis

Chi-square and t-test were performed using R programming language and GraphPad Prism. A *p*-value cut-off of 0.05 was used for gene expression, copy number and variant analysis.

Availability of supporting data

All genomics data have been deposited at the ArrayExpress (http://www.ebi.ac.uk/arrayexpress/), hosted by the European Bioinformatics Institute (EBI), under following accession numbers: E-MTAB-3958 : Whole transcriptome data; E-MTAB-3961 : Whole exome data; and, E-MTAB-3960 : SNP array data

Results

We characterized genetic alterations underlying four head and neck cancer cell lines followed by TCGA dataset to identify cumulative significance of biologically relevant alterations by integrating copy number, expression and point mutation data.

Characterization of four HNSCC cell lines established from Indian patients

Given that higher accumulative effect of individual genes can be reckoned by integrative analysis, we argue that these alteration can possibly be determined even with fewer samples. As a proof of principle, we performed an integrated characterization of karyotype analysis, copy number analysis, whole transcriptome and exome sequencing of 4 HNSCC cell lines established from Indian patients. In brief, significantly altered chromosomal segments based on copy number analysis were filtered based on nucleotide variant information and aberrant expression of transcripts to allow prioritization of regions harboring either deleterious mutation or expressing the transcript at significantly high levels, in addition to the stringent intrinsic statistical mining performed for each sample.

Karyotype analysis

The hyperploidy status of AW13516, AW8507, NT8E and OT9 cell lines were inferred by classical karyotyping with an average ploidy of 62, 62, 66 and 64, respectively that were largely consistent with ploidy as inferred form SNP array analysis (Fig. 1a; Additional file 1: Figure S1) and as reported for tumor cells lines [18, 19]. We specifically observed dicentric and ring chromosomes at elevated frequency indicating higher chromosomal instability (CIN) [44]. Overall distribution of chromosomal aberrations in each HNSCC cell lines showed similar pattern, representing an overall similar genomic structure of all HNSCC cell lines.

Copy number analysis

We performed genotyping microarray using Illumina 660 W quad SNP array chips of all the cell lines (Additional file 1: Figure S2). After stringent filtering of initial genotyping calls, on an average, 253 genomic segments of copy number changes were obtained per cell line. By limiting segment size at 10 Mb an average 166 focal segments were identified, including loss of copy number and LOH at 3p which is known to have correlation to advanced stage of tumor progression and poor clinical outcome [45, 46]; copy number gain on 11q known to be associated with advanced stage, recurrence and poor clinical outcome [47]; LOH at 8p and 9p which are known to be associated with advanced stage and survival [48] (Additional file 2: Table S1); and amplification of known oncogenes *EGFR* in AW13516, OT9; *MYC* in AW13516 and AW8507 cells; *JAK1* in NT8E, AW8507; *NSD1* in AW8507; and *MET* in AW13516 and OT9 (Additional file 2: Table S2). Several hallmark genes were found to be amplified in cell lines such as *CCND1, NOTCH1,* and *HES1* in all four cells; *PIK3CA* in AW13516, AW8507; deletion of *CDKN2A* in AW13516; *FBXW7* in NT8E, AW13516 and OT9 cells were detected and validated by real time PCR (Fig. 1b, Additional file 2: Table S2) in each cell line.

Whole transcriptome analysis

Whole transcriptome sequencing revealed 17,067, 19,374, 16,866 and 17,022 genes expressed in AW13516, AW8507, NT8e and OT9 respectively. Total ~5000 transcripts having less than 0.1 $\log_{10}$(FPKM + 1) were filtered out because of biologically non-significant expression level (Additional file 1: Figure S3). The upper quartile (>60 %) was considered as highly expressed genes and lower quartile (<40 %) was considered as lowly expressed genes. Gene set enrichment analysis of upper quartile showed enrichment of genes (data not shown) known to be up regulated in nasopharyngeal carcinoma [49]. All the transcripts showed 75 % overlap of expression profile with each other (Additional file 1: Figure S4) indicating overall similar nature of cell lines. Over expression of hallmark of HNSCC such as *CCND1, MYC, MET, CTNNB1, JAK1, HRAS, JAG1, and HES1* and down regulation of *FBXW7, SMAD4* in at least 3 cell line were observed and validated by quantitative real time PCR (Additional file 2: Table S3). A positive correlation was observed between transcriptome FPKM and qPCR Ct values (Fig. 1c).

Analysis of mutational landscape

All the cell lines were sequenced for whole exome at about 80X coverage using Illumina HiSeq. The relative coverage of each coding region was comparable across all four cell lines (Additional file 2: Table S4; Additional file 1: Figure S5). The coding part of the four cell line genome consist 28813, 47892, 20864 and 25029 variants in AW13516, AW8507, NT8e and OT9 cell line, respectively. Filtering of known germline variants (SNPs) and low quality variants left 5623, 4498, 2775, 5139 non-synonymous variants in AW13516, AW8507, NT8e and OT9 cell line, respectively (Additional file 2: Table S4). Of 20 HNSCC hallmark variants predicted as deleterious

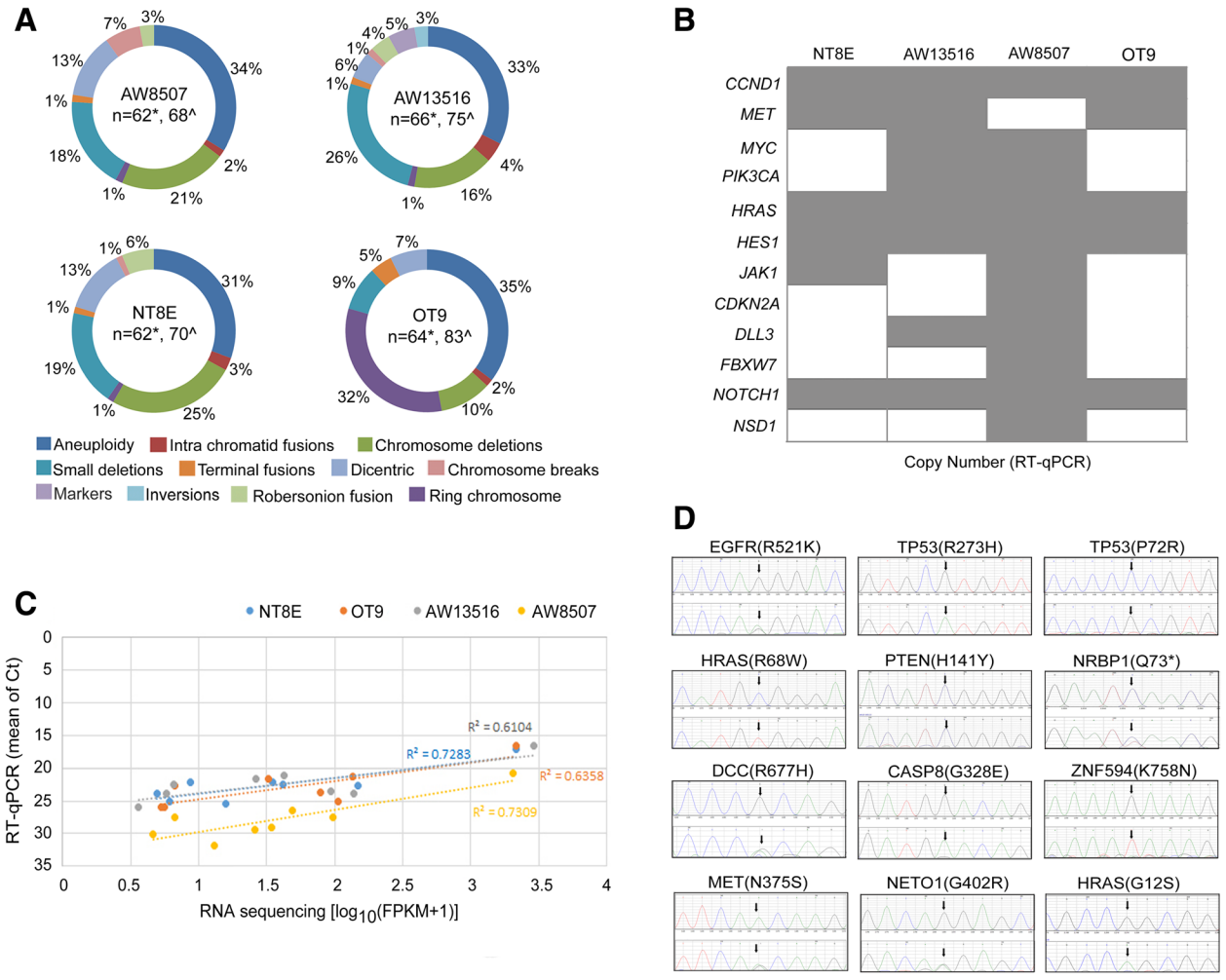


Fig. 1 Genomic alterations identified in HNSCC cell lines. **a** Chromosomal aberrations identified by 25 independent karyotype of each cell line is represented in circular form. Chromosome numbers in each cell line are indicated by n, as observed from karyotype (*) and predicted by SNP array (^). **b** Heatmap representation of copy number changes in key hallmark genes identified by SNP array and validated by qPCR. Grey box indicates validated and white box indicates invalidated copy number change in respective cell line. **c** Scatter plot representation of hallmark gene's expression. X-axis shows gene expression quantified by RNA sequencing and Y-axis shows gene expression quantified by RT-qPCR. Genes showed positive correlation between RNA sequencing and RT-qPCR. **d** Several high-quality point mutations identified by exome sequencing were validated by Sanger sequencing. Sanger sequencing trace were visualized using Mutation Surveyor software showing reference sequence trace in upper panel and mutant sequencing trace in below panel. Arrows represent mutation position with reference and mutated base indicated above the arrow

by two of three algorithms used for functional prediction [29–31], 17 variants could be validated by Sanger sequencing (Fig. 1d; Additional file 2: Table S5) including: *TP53* (R273H), *TP53* (P72R), *PTEN* (H141Y), *EGFR* (R521K), *HRAS* (G12S and R78W), and *CASP8* (G328E).

Integrated analysis identifies hallmark alterations in HNSCC cell lines

The first step of integration analysis involved identification of genes with positively correlated copy number and expression data. While no significant correlation was observed among expression and arm-level copy number segments (Additional file 1: Figure S6a), median expression of focally amplified and deleted genes positively correlated to their expression (Fig. 2a and Additional file 1: Figure S6b). About 1000 genes with focal copy number changes with consistent expression pattern were identified from four cell lines. The second step of integration analysis involved identification of mutated genes that were expressed. Number of missense mutations identified from transcriptome sequencing (67,641 variants) were much higher than from exome sequencing (30,649 variants). Filtering of exome variants against transcriptome variants reduced total number of 9253 unique missesne variants in all four cell lines (Fig. 2b). Two thousand four hundred seventy-nine missense mutations of 9523 total mutations found across all cells were used for further integration with copy number and expression data

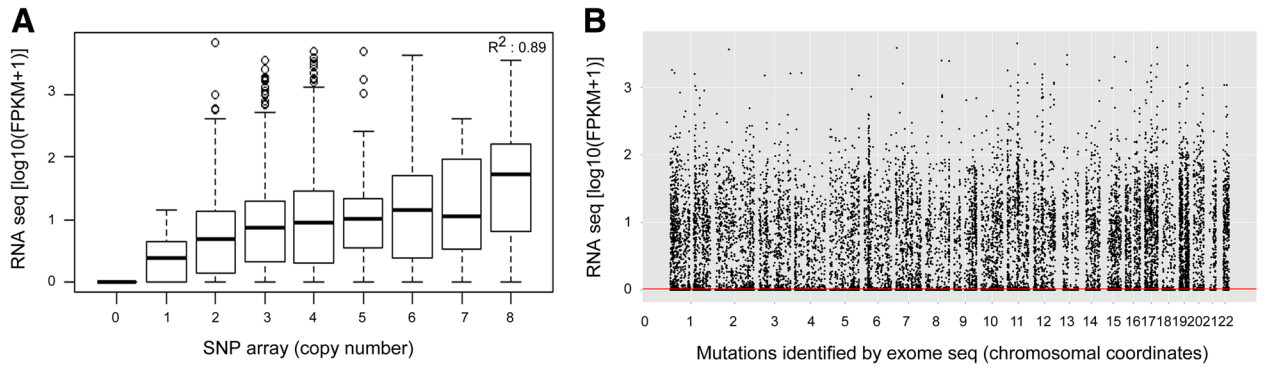


Fig. 2 Integration of copy-number, gene expression and single nucleotide variants. **a** Box plot representations showing focal copy number variations identified by SNP array on X-axis and gene expression quantified by RNA sequencing on Y-axis. Focally amplified or deleted genes from all four cell lines showed a positive trend with over expression or under expression respectively. **b** Scatter plot representation of genes harboring single nucleotide variants and it's expression from all four cell lines. Variants identified by whole exome sequencing are displayed by genomic coordinates on X-axis and normalized quantity of gene expression (by transcriptome sequencing) is displayed on Y-axis. Red line denotes 0.01 log10(FPKM + 1) filter which removes ~50 % of the total variants

(Additional file 1: Figure S7). Next, as third step of integration, we sorted genes with altered copy number, expression levels and harboring non-synonymous mutations for integrated analysis based on criterion as described in methodology in four cell lines (Additional file 1: Figure S7). Briefly, genes harbouring two or more type of alterations were selected: harbouring positive correlation of focal copy number and gene expression; or those harbouring point mutations with detectable transcript harbouring the variant—based on which, we identified 38 genes having multiple types of alterations (Additional file 2: Table S6). These include genes known to have somatic incidences in HNSCC: *TP53*, *HRAS*, *MET* and *PTEN*. We also identified *CASP8* in AW13516 cell line which was recently identified as very significantly altered by ICGC-India team in ~50 Indian HNSCC patients [50]. We additionally identified novel genes like *CCNDBP1, GSN, IMMT, LAMA5, SAT2* and *WDYHV1* to be altered by all three analysis i.e. CNV, expression and mutation. These all genes were also found to be altered in TCGA dataset with minimum 3 % cumulative frequency (Additional file 1: Figure S8). The overall convergence of copy number, expression and mutation data in each cell line is represented as circos plot (Fig. 3a; Additional file 1: Figure S9). Among the novel genes identified, of genes with at least one identical mutation previously reported include a pseudokinase *Nuclear receptor binding protein NRBP1* harboring heterozygous truncating mutation (Q73*) in NT8e cells, identical to as reported in lung cancer and altered in other cancers [51, 52].

Mutant *NRBP1* is required for tumor cell survival and is oncogenic in NIH3T3 cells

NRBP1 encodes for three different nuclear receptor binding protein isoform using three alternative translational initiation sites of 60 kDa, 51 kDa and 43 kDa [53], as were observed in 2 of 3 HNSCC cells (Fig. 4a). To determine whether expression of mutant *NRBP1* is required for tumor cell survival, we tested shRNA constructs in two HNSCC cells expressing all three forms of WT *NRBP1* (OT9 cells) and mutant *NRBP1* (NT8e cells). We demonstrate that even partial knockdown of mutant *NRBP1* expression in the NT8e cells, but not WT *NRBP1* expression in the OT9, significantly inhibited anchorage-independent growth and cell survival (Fig. 4b–d). We next tested the oncogenic role of *NRBP1*. mRNAs harboring premature termination (nonsense) codons are selectively degraded by Nonsense-mediated mRNA decay (NMD) [54]. However, mRNAs with nonsense mutations in the first exon are known to bypass NMD [55]. When ectopically expressed in NIH-3 T3 cells, mutant *NRBP1* transcript escape non-sense mediated degradation as determined by real time PCR (Additional file 1: Figure S10). All three isoform of NRBP1 were detected in NIH-3 T3 cells expressing wild type *NRBP1* cDNA. However, only two isoform of 51 kDa and 43 kDa were detected in cells transfected with mutant *NRBP1* cDNA (Fig. 4e upper panel). The over expression of the mutant *NRBP1* in NIH3T3 cells conferred anchorage-independent growth, forming significantly higher colonies in soft agar than cells expressing wild type *NRBP1* (Fig. 4f). Transformation of NIH-3 T3 cells by *NRBP1* over expression was accompanied by elevated phosphorylation of the MAPK (Fig. 4e lower panel).

Integrated analysis of TCGA dataset for HNSCC hallmark genes

Next, as a proof of principle, we computated cumulative frequency of copy number variations, expression changes and point mutations across 43 genes with ~3 %

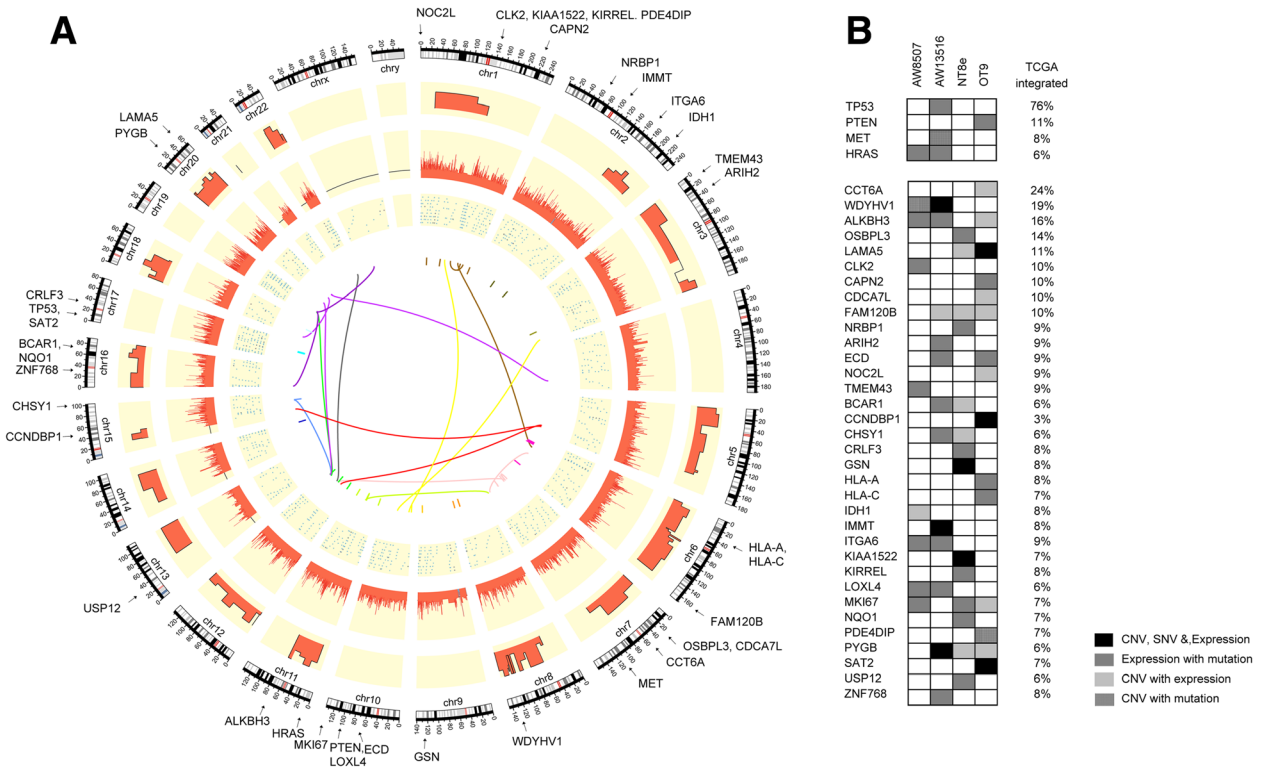


Fig. 3 Integrative genomic landscape of HNSCC. **a** Circos plot representations of genes identified by integrated genomic analysis of HNSCC cell lines. From outside to inside: karyotype, CNVs, Gene expression (FPKM), SNVs and translocations. Red colour indicates copy number gain or higher gene expression and blue colour indicates copy number loss or lower gene expression in CNV and FPKM tracks respectively. Non-synonymous mutations are indicated as blue triangles and grey circles represents non-sense mutations in SNV track. Fusion transcripts identified by transcriptome sequencing are shown as arc coloured by their chromosome of origin identified by ChimeraScan. **b** Heatmap representation of novel genes identified by integrated analysis of four HNSCC cell lines and their incidence in 279 HNSCC samples from TCGA study. Amplification (red) and deletions (blue) are indicated by filled box, over expression (red) and under expression (blue) are indicated by border line to the box, mis-sense (green), non-sense (black) and in-frame (brown) mutations are indicated by smaller square box

and higher mutation frequency in HNSCC TCGA dataset. As expected and described for few genes [4, 56], most of the genes were found to be altered at higher cumulative incidence than as reckoned by individual alterations (Fig. 5). Interestingly, three class of hallmark genes involved in HNSCC could be distinctly identified: genes that are primarily altered by mutations like *TP53* and *SYNE1*; genes that are sparsely altered by amplification or overexpression in addition to mutations like *FAT1*, *NOTCH1*, *KMT2D*, and *FLG*; and, genes that are preferentially altered by amplification or over expression over point mutations with higher cumulative effect than known before. Of these, previously described genes like *PIK3CA*, *CDKN2A*, *TP63*, *EGFR*, *CASP8*, *NFE2L2*, and *KRAS* show more than twice cumulative effect of alteration while rest of the genes are altered at several folds higher cumulative frequency based on integrated analysis. Furthermore, three genes– *UBR5*, *ZNF384* and *TERT* were found to be altered with cumulative frequency of 32, 19, and 16 %, respectively that has not been previously described in HNSCC.

Discussion

We have characterized genetic alterations of unknown somatic status underlying four head and neck cancer cell lines of Indian origin patient by subjecting them to a thorough karyotype based characterization, SNP array based analysis, whole exome capture sequencing, and mRNA sequencing.

Integrated analysis of the cell lines establish their resemblance to primary tumors. Consistent with literature, most frequent copy number gains in head and neck cancer cells in this study were observed at 2q, 3q, 5p and 7p, and deletions at 3p, 9p, 10p, 11q, 14q, 17q and 19p, as reported earlier [57, 58]. Integration of multiple platform with the copy number variation, allowed us to identify the functionally relevant alterations including several hall marks genes known to be involved in HNSCC, viz. *PIK3CA, EGFR, HRAS, MYC, CDKN2A, MET, TRAF2, PTK2* and *CASP8*. Of the novel genes, *JAK1* was found to be amplified in two of the cell lines and overexpressed in all 4 HNSCC cells; *NOTCH1* known to harbor inactivating mutations in HNSCC [3, 50] was

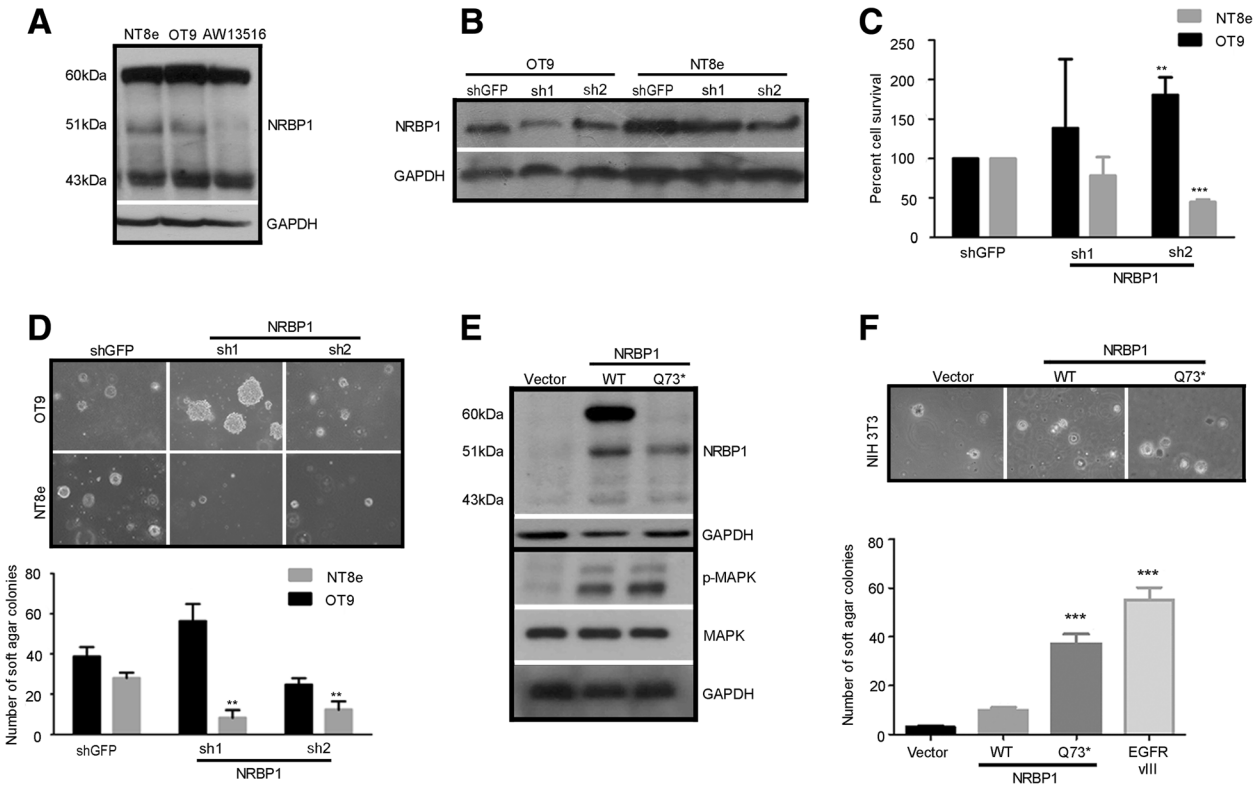


Fig. 4 Mutant *NRBP1* is required for tumor cell survival and is oncogenic in NIH3T3 cells. Knockdown of mutant *NRBP1* expression with shRNA inhibits transformation and survival of HNSCC cell lines. **a** Western blot analysis of total NRBP1 expression level in HNSCC cell lines. *NRBP1* encodes for three different nuclear receptor binding protein isoform: 60 kDa, 51 kDa and 43 kDa in NT8e cells (lane 1), OT9 cells (lane 2). AW13516 cells express only two isoforms (lane 3). **b** Western blot representation of *NRBP1* partial knockdown by two independent hairpins (shNRBP1#1 and shNRBP1#2) in NT8e and OT9 cells. The hairpin constructs inhibit cell survival as assessed by cell counting as described in methods. **c** and anchorage-independent growth as assessed by colony formation in soft agar **d** in the NT8e cells harboring an *NRBP1 Q73** mutation, but not the OT9 cells, which express WT *NRBP1*. shGFP, a hairpin specific for green fluorescent protein, was used as a negative control. All results are normalized to survival or colony formation by cells infected with empty vector. Images were taken at 10X magnification. **e** Western blot analysis of *NRBP1* wild type and mutant (Q73*) in NIH-3 T3 cells. Vector control, NIH-3 T3 clones overexpressing wild-type and mutant *NRBP1* is shown. *NRBP1* Q73* mutation expresses 51 kDa isoform, while two isoform of 60 kDa and 51 kDa can be seen in 3 T3 clones overexpressing wild-type *NRBP1*. Western blot analysis of total and phosphorylated MAPK in NIH-3 T3 clones expressing *NRBP1* wild type (lane 2) and mutant NRBP1 (Q73*). **f** Representative images (taken at 20x) of soft agar colony formed by NIH-3 T3 cells expressing *NRBP1* wild type and Q73*. Bar graph representation of number of soft agar colonies formed by stable NIH-3 T3 clones. NIH-3 T3 cells expressing *EGFR* VIII was used as positive control and relative comparison of transforming ability of mutant *NRBP1*. *** *P*-value <0.0001, ** *P*-value <0.001

found to be amplified in all 4 and overexpressed in 2 of 4 HNSCC cells, known to be play dual role in a context dependent manner [59].

We also observed missense mutations in several novel genes such as *CLK2, NRBP1, CCNDBP1, IDH1, LAMA5, BCAR1, ZNF678,* and *CLK2*. Of these, genes with at least one identical mutation previously reported include *NRBP1* (Q73*), a pseudo kinase, found in NT8e cells, earlier reported in lung and other cancers [51, 52], with an overall 9 % cumulative frequency alteration in TCGA HNSCC dataset (Additional file 1: Figure S8). Of 48 pseudo kinases known in human genome, several have been shown to retain their biochemical catalytic activities despite lack of one or more of the three catalytic residues essential for its kinase activity, with their established roles in cancer [60–62].

Interestingly, several activating mutant alleles of *NRBP1 homolog Drosophila Madm* (Mlf1 adapter molecule) *3 T4* (Q46*); *2U3* (C500*); *3G5* (Q530*); *7 L2* and *3Y2* (that disrupts splice donor site of first exon) are known, wherein alternative translation start codons is similarly suggestive for a varying degree of pinhead phenotype severity associated with the mutant alleles [53, 63]. Studies in the fruit fly have provided important insights into mechanisms underlying the biology of growth promoting *NRBP1 homolog Drosophila Madm.* A recent study suggests *Drosophila Madm* interacts with *Drosophila bunA* that encodes a gene homologous to human *Transforming Growth Factor-β1 stimulated clone-22 TSC-22* [63]; that were later shown to interact even in mammalian system [64]. Interestingly, mammalian tumor suppressor *TSC-22* is

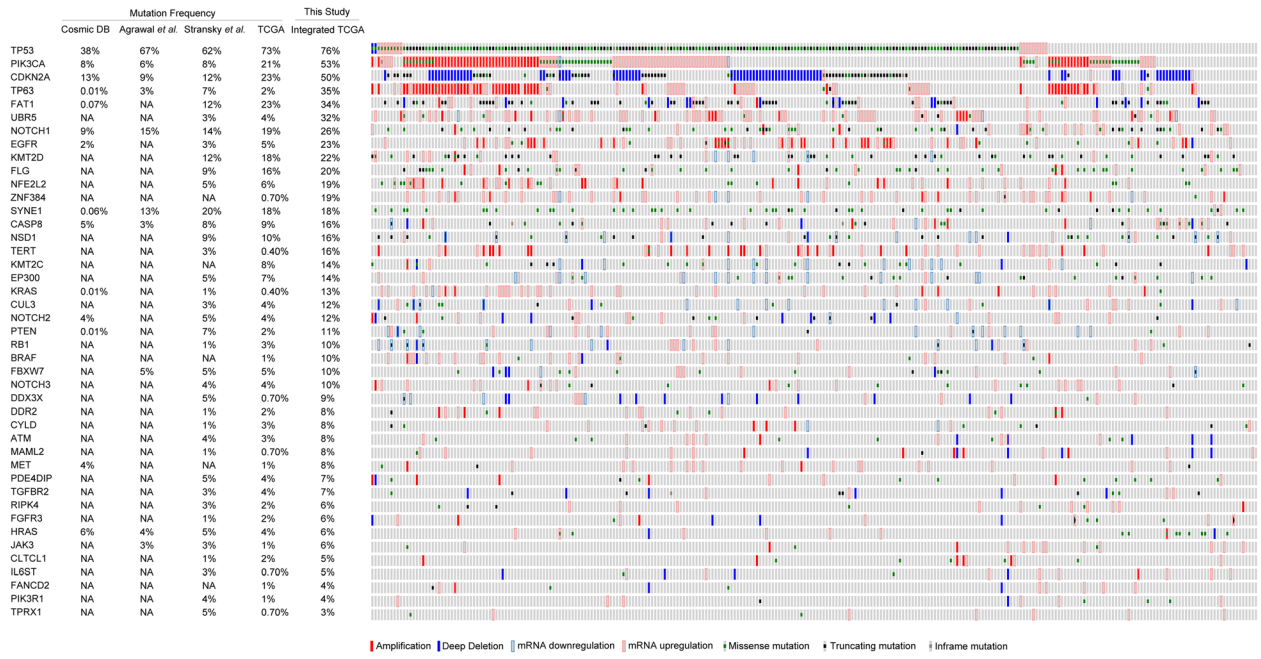


Fig. 5 Integrative genomic landscape of HNSCC tumors in TCGA dataset. Heatmap representation of hallmark genes in 279 HNSCC samples from TCGA study with frequency of alterations based on integrated CNVs, gene expression and SNVs. Frequency of alterations in the gene based on Cosmic database, Agrawal et al., Stransky et al. and TCGA study are also shown for comparison with integrated analysis. Amplification (red) and deletions (blue) are indicated by filled box, over expression (red) and under expression (blue) are indicated by border line to the box, mis-sense (green), non-sense (black) and in-frame (brown) mutations are indicated by smaller square box

known to play an important role in maintaining differentiated phenotype in salivary gland tumors [65], a subtype of head and neck cancer. More recently, studies have shown poor clinical outcomes are associated with *NRBP1* over expression in prostate cancer [64]. We provide the first functional analysis of mutant *NRBP1* and establish that NIH-3 T3 cells expressing the mutant *NRBP1* enhance their survival and anchorage independent growth, while its knock down diminishes survival and anchorage-independent growth by oral cancer cells expressing activating *NRBP1* mutations. Thus, NT8e cells harboring mutant *NRBP1* was found to be consistent with its suggestive role in prostate cancer biology and other model organisms. Interestingly, *NRBP1* has also been shown to be involved in intestinal progenitor cell homeostasis with tumor suppressive function [66], suggesting its role is specific to the cellular context. This study identifies *NRBP1* mutant to play an oncogenic role in head and neck cancer. However, in depth systematic sequencing of *NRBP1* in a wide variety of tumor types may help indicate utility of NRBP1 inhibition in human cancer.

Furthermore, based on TCGA data integrated analysis, cumulative alteration frequency of *TP63* (35 %), *EGFR* (23 %) and *NFEL2* (19 %) were found to be higher than reported in COSMIC and cBioPortal, consistent with as described in other reports [4, 56]. Of alterations not defined before, *UBR5*, *ZNF384* and *TERT* were found to be altered at higher frequency at 32, 19, 16 %, respectively. Interestingly, recurrent *UBR5-ZNF384* fusion has been shown to be oncogenic in EBV-associated nasopharyngeal subtype of HNSCC [67]; amplification of *TERT* has been shown to be higher in lung squamous [68], suggesting these alterations as potential squamous specific event, though that warrants detailed systematic assessment.

In overall, this study underscores integrative approaches through a filtering strategy to help reckon higher cumulative frequency for individual genes affected by two or more alterations to achieve the threshold for statistical significance even from fewer samples. The integrative analysis as described here, in essence, is based on a linear simplified assumption of disease aetiology that variation at DNA level lead to changes in gene expression causal to transformation of the cell. As a major deficiency, only genes that are subject to multiple levels of biological regulation are likely to be determined by this approach than genes that are primarily altered by single alteration like amplification or over expression.

Conclusion

As a proof of principle, integrated analysis of copy number variation, exome and transcriptome of 4 head

and neck cancer cell lines and TCGA HNSCC dataset identify *NRBP1, UBR5*, *ZNF384* and *TERT* as novel candidate oncogenes in HNSCC. However, systematic functional experimental validation is required to further guide and identify true driver events of these alterations. Additionally, the genetically- defined cellular systems characterized by integrated genomics analysis in this study (NT8e, OT9, AW13516, AW8507), together with the identification of novel actionable molecular targets, may help further facilitate the pre-clinical evaluation of emerging therapeutic agents in head and neck cancer.

Additional files

Additional file 1: Additional Figures S1-S10. Chromosomal aberration in HNSCC patient derived cell lines AW8507, AW13516, NT8e and OT9. (A) Representative karyotype of AW8507, AW13516, NT8e and OT9 cells is shown from total 25 karyotypes obtained per cell line. (B) Chromosomal aberrations identified by 25 independent karyotype of each cell line is represented in circular form. Chromosome numbers in each cell line are indicated by n, as observed from karyotype (*) and predicted by SNP array (^). Copy number changes in HNSCC cell lines identified by SNP array. Genome alteration print (GAP) of (A) AW8507, (B) AW13516, (C) NT8e AND (D) OT9 cell lines obtained by SNP array. First horizontal block represents B-allele frequency, second block represents absolute copy number, third block is log R ratio. Frequency of transcripts per binned log transformed FPKM + 1. Raw RNA sequencing data was binned to obtain frequency of genes per log10(FPKM + 1) in (A) AW8507, (B) AW13516, (C) NT8e and (D) OT9. Horizontal dotted lines indicates percentile of transcripts in the quadrant. Similarity of gene expression in HNSCC cell lines. Number of genes commonly expressed between AW8507, AW13516, NT8e and OT9 cell lines. Relative depth in exome sequencing. Relative depth of sequencing for various genomic regions in (A) AW8507, (B) AW13516, (C) NT8e and (D) OT9. Correlation of copy number with gene expression. (A) Arm level and (B) focal copy number changes and gene expression (y-axis) are shown for AW8507, AW13516, NT8e and OT9 cell lines. Correlation of focal copy number with gene expression was 1.5 fold higher in AW8507, 5.2 fold higher in AW13516, 2.4 fold higher in NT8e and 1.6 fold higher in OT9 cell line compare to arm level copy number changes. *P*-value cut-off of 0.05 was used as threshold for statistical significance.* denotes *P*-value <0.05, ** <0.005, *** <0.0005. Schematic view of data reduction in integrated genomic analysis. Flow chart depicts the reduction of data at each stage of integration. First row indicates number of genes identified from each platform as raw calls. Second and third row indicates number of genes left after each step of integration with main selection parameter indicated outside the box. Integrative genomic alterations of genes in TCGA dataset of HNSCC tumors. Heatmap representation of 38 genes in 279 HNSCC samples from TCGA study with frequency of alterations based on integrated CNVs, gene expression and SNVs. Amplification (red) and deletions (blue) are indicated by filled box, over expression (red) and under expression (blue) are indicated by border line to the box, mis-sense (green), non-sense (black) and in-frame (brown) mutations are indicated by smaller square box. Circos plot representation of HNSCC cell lines. Circos plot representations of integrated genomic data of (A) AW8507, (B) AW13516, (C) NT8e and (D) OT9 cell lines. From outside to inside: karyotype, CNVs, Gene expression (FPKM), SNVs and translocations. Red colour indicates copy number gain or higher gene expression and blue colour indicates copy number loss or lower gene expression in CNV and FPKM tracks, respectively. Non-synonymous mutations are indicated as blue triangles and grey circles represents non-sense mutations in SNV track. Fusion transcripts identified by transcriptome sequencing are shown as arc coloured by their chromosome of origin identified by ChimeraScan. *NRBP1* expression in NIH-3 T3 cells. qPCR analysis of *NRBP1* gene expression in NIH-3 T3 stably expressing wild and mutant. Data was normalized against *GAPDH* and fold change plotted. *P* value <0.0001 is denoted as ***. (PDF 9017 kb)

Additional file 2: Additional Tables S1-S8: Copy number alterations of known genomic locations identified in HNSCC cell lines. Copy number alterations in halmark genes identified in HNSCC cell lines. Gene expression of hallmark genes by RNA sequencing and qPCR. Features of whole exome and transcriptome sequencing. Validation of mutations in hallmark and novel genes. Details of mutations identified by integrated analysis in HNSCC cell lines. Primer sequences used for Sanger sequenicng based validation of mutations. Primers used for copy number and gene expression study using qPCR. (ZIP 249 kb)

Abbreviations

HNSCC: Head and neck squamous cell carcinoma; COSMIC: Catalogue of Somatic Mutation In Cancer; TCGA: The Cancer Genome Atlas; ICGC: International Cancer Genome Consortium; CIN: Chrmosomal Instability; CNV: Copy Number Variation.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PC and PU contributed equally to this work; PC, PU, SG, and AD designed research; PC, PU, PI, MT, MS, GVR, NO, AK, RC, and MR performed research; PC, PU, MS, and NO contributed new reagents/analytic tools; PC, PU, PI, MT, MS, GVR, NO, AK, RC, MR and AD analyzed data; and PC, PU, SG and AD wrote the paper. All authors have read and approved the manuscript.

Acknowledgements

All members of the Dutt laboratory for critically reviewing the manuscript. Rita Mulherkar for establishing and sharing OT9 and NT8e cells. Anti Cancer Drug Screening Facility at ACTREC, Tata Memorial Center for AW13516 and AW8507 cells. Genotypic Inc., Sandor Proteomics Pvt. Ltd. and Centre for Cellular and Molecular Platforms (C-CAMP), for providing sequencing and SNP array genotyping services. A.D. is supported by an Intermediate Fellowship from the Wellcome Trust/DBT India Alliance (IA/I/11/2500278), by a grant from DBT (BT/PR2372/AGR/36/696/2011), and intramural grants (IRB project 92 and 55). P.C. and P.I are supported by senior research fellowship from ACTREC. P.U. is supported by senior research fellowship from CSIR. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author details

[1]Advanced Centre for Treatment, Research and Education in Cancer, Tata Memorial Center, Navi Mumbai, Maharashtra 410210, India. [2]Department of Medical Oncology, Tata Memorial Hospital, Tata Memorial Center, Mumbai, Maharashtra, India.

Received: 13 August 2015 Accepted: 23 October 2015
Published online: 14 November 2015

References

1. Rothenberg SM, Ellisen LW. The molecular pathogenesis of head and neck squamous cell carcinoma. J Clin Invest. 2012;122(6):1951–7.
2. Agrawal N, Frederick MJ, Pickering CR, Bettegowda C, Chang K, Li RJ, et al. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. Science. 2011;333(6046):1154–7.
3. Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, et al. The mutational landscape of head and neck squamous cell carcinoma. Science. 2011;333(6046):1157–60.
4. Cancer Genome Atlas N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015;517(7536):576–82.
5. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214–8.
6. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC: identifying mutational significance in cancer genomes. Genome Res. 2012;22(8):1589–98.
7. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, et al. IntOGen-mutations identifies cancer drivers across tumor types. Nat Methods. 2013;10(11):1081–2.
8. Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP, et al. A landscape of driver mutations in melanoma. Cell. 2012;150(2):251–63.

9. Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol Syst Biol. 2013;9:637.
10. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12(4):R41.
11. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014;505(7484):495–501.
12. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, et al. An integrated approach to uncover drivers of cancer. Cell. 2010;143(6):1005–17.
13. Upadhyay P, Dwivedi R, Dutt A. Applications of next-generation sequencing in cancer. Curr Sci. 2014;107(5):795.
14. Natrajan R, Wilkerson P. From integrative genomics to therapeutic targets. Cancer Res. 2013;73(12):3483–8.
15. Pickering CR, Zhang J, Yoo SY, Bengtsson L, Moorthy S, Neskey DM, et al. Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. Cancer Discov. 2013;3(7):770–81.
16. Wilkerson MD, Cabanski CR, Sun W, Hoadley KA, Walter V, Mose LE, et al. Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. Nucleic Acids Res. 2014;42(13):e107.
17. Kristensen VN, Lingjaerde OC, Russnes HG, Vollan HK, Frigessi A, Borresen-Dale AL. Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer. 2014;14(5):299–313.
18. Mulherkar R, Goud AP, Wagle AS, Naresh KN, Mahimkar MB, Thomas SM, et al. Establishment of a human squamous cell carcinoma cell line of the upper aero-digestive tract. Cancer Lett. 1997;118(1):115–21.
19. Tatake RJ, Rajaram N, Damle RN, Balsara B, Bhisey AN, Gangal SG. Establishment and characterization of four new squamous cell carcinoma cell lines derived from oral tumors. J Cancer Res Clin Oncol. 1990;116(2):179–86.
20. Popova T, Manie E, Stoppa-Lyonnet D, Rigaill G, Barillot E, Stern MH. Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. Genome Biol. 2009;10(11):R128.
21. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
22. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
23. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
24. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8.
25. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31(3):213–9.
26. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.
27. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet. 2008;Chapter 10:Unit 10 11.
28. Ramos AH, Lichtenstein L, Gupta M, Lawrence MS, Pugh TJ, Saksena G, et al. Oncotator: cancer variant annotation tool. Hum Mutat. 2015;36(4):E2423–9.
29. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7 20.
30. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015;31(16):2745-2747.
31. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118.
32. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36.
33. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562–78.
34. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet. 2012;44(6):685–9.
35. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
36. Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 2011;27(20):2903–4.
37. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639–45.
38. Minton JA, Flanagan SE, Ellard S. Mutation surveyor: software for DNA sequence analysis. Methods Mol Biol. 2011;688:143–53.
39. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(−Delta Delta C(T)) Method. Methods. 2001;25(4):402–8.
40. Sancak Y, Peterson TR, Shaul YD, Lindquist RA, Thoreen CC, Bar-Peled L, et al. The Rag GTPases bind raptor and mediate amino acid signaling to mTORC1. Science. 2008;320(5882):1496–501.
41. Dutt A, Salvesen HB, Chen TH, Ramos AH, Onofrio RC, Hatton C, et al. Drug-sensitive FGFR2 mutations in endometrial carcinoma. Proc Natl Acad Sci U S A. 2008;105(25):8713–7.
42. Moffat J, Grueneberg DA, Yang X, Kim SY, Kloepfer AM, Hinkle G, et al. A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. Cell. 2006;124(6):1283–98.
43. Walker JM. The bicinchoninic acid (BCA) assay for protein quantitation. Methods Mol Biol. 1994;32:5–8.
44. Roschke AV, Tonon G, Gehlhaus KS, McTyre N, Bussey KJ, Lababidi S, et al. Karyotypic complexity of the NCI-60 drug-screening panel. Cancer Res. 2003;63(24):8634–47.
45. Yamamoto N, Mizoe J, Numasawa H, Tsujii H, Shibahara T, Noma H. Allelic loss on chromosomes 2q, 3p and 21q: possibly a poor prognostic factor in oral squamous cell carcinoma. Oral Oncol. 2003;39(8):796–805.
46. Partridge M, Emilion G, Langdon JD. LOH at 3p correlates with a poor survival in oral squamous cell carcinoma. Br J Cancer. 1996;73(3):366–71.
47. Meredith SD, Levine PA, Burns JA, Gaffey MJ, Boyd JC, Weiss LM, et al. Chromosome 11q13 amplification in head and neck squamous cell carcinoma. Association with poor prognosis. Arch Otolaryngol Head Neck Surg. 1995;121(7):790–4.
48. Chen Y, Chen C. DNA copy number variation and loss of heterozygosity in relation to recurrence of and survival from head and neck squamous cell carcinoma: a review. Head Neck. 2008;30(10):1361–83.
49. Dodd LE, Sengupta S, Chen IH, den Boon JA, Cheng YJ, Westra W, et al. Genes involved in DNA repair and nitrosamine metabolism and those located on chromosome 14q32 are dysregulated in nasopharyngeal carcinoma. Cancer Epidemiol Biomarkers Prev. 2006;15(11):2216–25.
50. India Project Team of the International Cancer Genome C. Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups. Nat Commun. 2013;4:2873.
51. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603–7.
52. Davies H, Hunter C, Smith R, Stephens P, Greenman C, Bignell G, et al. Somatic mutations of the protein kinase gene family in human lung cancer. Cancer Res. 2005;65(17):7591–5.
53. Hooper JD, Baker E, Ogbourne SM, Sutherland GR, Antalis TM. Cloning of the cDNA and localization of the gene encoding human NRBP, a ubiquitously expressed, multidomain putative adapter protein. Genomics. 2000;66(1):113–8.
54. Schweingruber C, Rufener SC, Zund D, Yamashita A, Muhlemann O. Nonsense-mediated mRNA decay - mechanisms of substrate mRNA recognition and degradation in mammalian cells. Biochim Biophys Acta. 2013;1829(6–7):612–23.
55. Neu-Yilik G, Amthor B, Gehring NH, Bahri S, Paidassi H, Hentze MW, et al. Mechanism of escape from nonsense-mediated mRNA decay of human beta-globin transcripts with nonsense mutations in the first exon. RNA. 2011;17(5):843–54.
56. Rusan M, Li YY, Hammerman PS. Genomic landscape of human papillomavirus-associated cancers. Clin Cancer Res. 2015;21(9):2009–19.

57. Smeets SJ, Braakhuis BJ, Abbas S, Snijders PJ, Ylstra B, van de Wiel MA, et al. Genome-wide DNA copy number alterations in head and neck squamous cell carcinomas with or without oncogene-expressing human papillomavirus. Oncogene. 2006;25(17):2558–64.
58. Ambatipudi S, Gerstung M, Gowda R, Pai P, Borges AM, Schäffer AA, et al. Genomic profiling of advanced-stage oral cancers reveals chromosome 11q alterations as markers of poor clinical outcome. PLoS ONE. 2011;6(2):e17250.
59. Ntziachristos P, Lim JS, Sage J, Aifantis I. From fly wings to targeted cancer therapies: a centennial for notch signaling. Cancer Cell. 2014;25(3):318–34.
60. Hua F, Mu R, Liu J, Xue J, Wang Z, Lin H, et al. TRB3 interacts with SMAD3 promoting tumor cell migration and invasion. J Cell Sci. 2011;124(Pt 19):3235–46.
61. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science. 2002;298(5600):1912–34.
62. Zeqiraj E, Filippi BM, Deak M, Alessi DR, van Aalten DM. Structure of the LKB1-STRAD-MO25 complex reveals an allosteric mechanism of kinase activation. Science. 2009;326(5960):1707–11.
63. Gluderer S, Oldham S, Rintelen F, Sulzer A, Schutt C, Wu X, et al. Bunched, the Drosophila homolog of the mammalian tumor suppressor TSC-22, promotes cellular growth. BMC Dev Biol. 2008;8:10.
64. Ruiz C, Oeggerli M, Germann M, Gluderer S, Stocker H, Andreozzi M, et al. High NRBP1 expression in prostate cancer is linked with poor clinical outcomes and increased cancer cell growth. Prostate. 2012;72(15):1678–87.
65. Doi Y, Kawamata H, Ono Y, Fujimori T, Imai Y. Expression and cellular localization of TSC-22 in normal salivary glands and salivary gland tumors: implications for tumor cell differentiation. Oncol Rep. 2008;19(3):609–16.
66. Wilson CH, Crombie C, van der Weyden L, Poulogiannis G, Rust AG, Pardo M, et al. Nuclear receptor binding protein 1 regulates intestinal progenitor cell homeostasis and tumour formation. EMBO J. 2012;31(11):2486–97.
67. Chung GT, Lung RW, Hui AB, Yip KY, Woo JK, Chow C, et al. Identification of a recurrent transforming UBR5-ZNF423 fusion gene in EBV-associated nasopharyngeal carcinoma. J Pathol. 2013;231(2):158–67.
68. Zhu CQ, Cutz JC, Liu N, Lau D, Shepherd FA, Squire JA, et al. Amplification of telomerase (hTERT) gene is a poor prognostic marker in non-small-cell lung cancer. Br J Cancer. 2006;94(10):1452–9.

Notch pathway activation is essential for maintenance of stem-like cells in early tongue cancer

Pawan Upadhyay[1,*], Sudhir Nair[2,*], Ekjot Kaur[3], Jyotirmoi Aich[1], Prachi Dani[1], Vidyalakshmi Sethunath[1], Nilesh Gardi[1], Pratik Chandrani[1], Mukul Godbole[1], Kavita Sonawane[2], Ratnam Prasad[1], Sadhana Kannan[4], Beamon Agarwal[5], Shubhada Kane[6], Sudeep Gupta[7], Shilpee Dutt[3], Amit Dutt[1]

[1]Integrated Genomics Laboratory, Advanced Centre for Treatment, Research and Education In Cancer, Tata Memorial Centre, Navi Mumbai- 410210, India

[2]Division of Head and Neck Oncology, Department of Surgical Oncology, Tata Memorial Hospital, Tata Memorial Centre,Mumbai- 4100012, India

[3]Shilpee Laboratory, Advanced Centre for Treatment, Research and Education In Cancer, Tata Memorial Centre, Navi Mumbai- 410210, India

[4]Advanced Centre for Treatment, Research and Education In Cancer, Tata Memorial Centre, Navi Mumbai- 410210, India

[5]Department of Pathology, Advanced Centre for Treatment, Research and Education In Cancer, Tata Memorial Centre, Navi Mumbai- 410210, India

[6]Department of Pathology, Tata Memorial Hospital, Tata Memorial Centre, Mumbai- 400012, India

[7]Department of Medical Oncology, Advanced Centre for Treatment, Research and Education In Cancer, Tata Memorial Centre, Mumbai- 400012, India

[*]These authors have contributed equally to this work

Correspondence to: *Amit Dutt,* ***email:*** *adutt@actrec.gov.in*

Keywords: *early stage tongue cancer, exome and transcriptome sequencing, IHC based expression analysis, cancer stem cell-like feature, Notch pathway inhibitors*

Received: *January 25, 2016* ***Accepted:*** *June 07, 2016* ***Published:*** *July 06, 2016*

ABSTRACT

Background: Notch pathway plays a complex role depending on cellular contexts: promotes stem cell maintenance or induces terminal differentiation in potential cancer-initiating cells; acts as an oncogene in lymphocytes and mammary tissue or plays a growth-suppressive role in leukemia, liver, skin, and head and neck cancer. Here, we present a novel clinical and functional significance of *NOTCH1* alterations in early stage tongue squamous cell carcinoma (TSCC).

Patients and Methods: We analyzed the Notch signaling pathway in 68 early stage TSCC primary tumor samples by whole exome and transcriptome sequencing, real-time PCR based copy number, expression, immuno-histochemical, followed by cell based biochemical and functional assays.

Results: We show, unlike TCGA HNSCC data set, *NOTCH1* harbors significantly lower frequency of inactivating mutations (4%); is somatically amplified; and, overexpressed in 31% and 37% of early stage TSCC patients, respectively. HNSCC cell lines over expressing *NOTCH1*, when plated in the absence of attachment, are enriched in stem cell markers and form spheroids. Furthermore, we show that inhibition of NOTCH activation by gamma secretase inhibitor or shRNA mediated knockdown of *NOTCH1* inhibits spheroid forming capacity, transformation, survival and migration of the HNSCC cells suggesting an oncogenic role of *NOTCH1* in TSCC. Clinically, Notch pathway activation is higher in tumors of non-smokers compared to smokers (50% Vs 18%, respectively, *P*=0.026) and is also associated with greater nodal positivity compared to its non-activation (93% Vs 64%, respectively, *P*=0.029).

Conclusion: We anticipate that these results could form the basis for therapeutic targeting of NOTCH1 in tongue cancer.

INTRODUCTION

Recent large-scale genome wide studies have underscored a complex role of *NOTCH1* as a candidate tumor suppressor harboring inactivating mutation and deletions , as well as a driver of tumorigenesis harboring activating missense mutations and amplifications in a context dependent manner in HNSCC, and other cancers [1-10]. In addition, Notch signaling pathway plays a significant role in the maintenance of cancer stem-like population of cells (CSCs) in several human cancers [11-15] . Inhibition of Notch signaling prevents the formation of secondary mammospheres by cell lines derived from primary breast cancer patient samples. However, the biological significance of cancer stem-like cells (CSCs) in HNSCC has not been well characterized.

To understand the role of Notch signaling pathway in early-stage (T1-T2) tongue tumors, we examined the mutational landscape, copy number alterations and differential expression of receptor, ligands, modifiers and target genes of the Notch pathway, along with effect of genetic and pharmacologic perturbation of Notch pathway on cancer stem-like cells (CSCs) features of HNSCC cells.

RESULTS

All the samples with available genomic DNA were tested for the presence of HPV using MY09/11 PCR and E6 transcript PCR primers. 40 of 71 samples analyzed, all were found to be HPV negative. Where exome sequence was available, the absence of HPV was re-confirmed using HPVDetector, as previously described [16]. TSCC samples of Indian origin to be HPV negative is consistent with other studies [17-19].

Notch pathway is activated in early TSCC patients

To characterize somatic alterations across 48 genes of Notch signaling pathway in 29 early-stage (T1-T2) tongue squamous cell carcinoma (TSCC) patient-derived tumors, we analyzed 23 paired whole exome and 10 whole transcriptome tongue cancer tumor sequencing data (unpublished data), as detailed in Supplementary Table S1 and Supplementary Figure S1A. Fourteen mutations were observed in 7 genes across 12 of 22 samples (Supplementary Table S1). Of note, inactivating *NOTCH1* mutation (4%) were found at a lower frequency in our sample set than that reported from the Caucasian population [7, 8, 20] but consistent with similar finding from a recent Asian study [21, 22]. In further contrast to Caucasian population, we observed Notch family receptors, ligands, and downstream effector genes were amplified or over expressed in 59% samples (17 of 29 patients) based on copy number variations called from whole- exome and whole- transcriptome data. To extend and validate these findings, we performed real-time quantitative PCR to estimate DNA copy number and transcript levels, along with an immunohistochemical analysis of Notch pathway components in paired tumor-normal samples from tongue cancer patients. We found somatic amplification at *NOTCH1* in 12 of 38 tumors (Figure 1A); overexpression of *NOTCH1* transcripts was observed in 16 of 45 samples (Figure 1B, 1C)-- consistent with our analysis of the TCGA TSCC data set (n=126) (Supplementary Figure 1B, 1C), not reported earlier. Also, samples harboring amplification at *NOTCH1* (*P* value <0.001) and *DLL4* (*P* value <0.001) showed significantly higher expression of transcript as compared to no amplification. (Supplementary Figure S1D and Supplementary Figure S2). Consistent with amplification and over expression of Notch pathway components, Immunohistochemical analysis for activated NOTCH1 intracellular domain (NICD) in a set of 50 patients indicated strong immunoreactivity for active Notch signaling present in 40% tumor samples (Figure 1D-1E, Supplementary Figure S3A-S3C).

Expression of *NOTCH1* is required for survival, migration and stemness of TSCC tumor cells

To assess the functional significance of Notch pathway activation, we asked if the expression of *NOTCH1* is essential for survival, migration and stem-like feature of HNSCC cells *in vitro*. First, we checked for the presence of Notch pathway transcript expression by real-time PCR and western analysis of NOTCH1 protein using multiple head and neck cancers cell lines (NT8e, AW13516, CAL27 and DOK) [23]. NT8e and CAL27 cells showed higher expression of NOTCH1 as compared to AW13516 and DOK cells (Supplementary Figure S4A, S4B). Next, we tested a series of shRNA constructs to knockdown *NOTCH1* in these cells. The knockdowns were confirmed by western blot analysis for NOTCH1 (Figure 2A) and quantitative real-time PCR for *NOTCH1* and its target gene *HES1* (Supplementary Figure S4C). We identified two shRNA clones sh1 and sh2 that efficiently knocked down expression of *NOTCH1* compared to scrambled (SCR). Knock down of *NOTCH1* inhibited cell survival (Figure 2B), anchorage-independent growth (Figure 2C), in NT8e and CAL27, and migration in NT8e (Figure 2D).

Expression of *NOTCH1* and its pathway genes maintains cancer stem-like cells (CSCs) in various tumors, as determined by their ability to form spheroids and expression of molecular markers ALDH1, CD133 and CD44 [12, 24]. An *in vitro* spheroid formation assay was performed to examine the cancer stem cell population (CSCs) in HNSCC cell lines (NT8e, CAL27, AW13516, and DOK) expressing a variable level of *NOTCH1* expression (Supplementary Figure S4B). As shown in Figure 3A, following 10 days of incubation in undifferentiating stem cell media, NOTCH1 over expressing NT8e cells showed a higher number of

oralspheres with 32% and 0.21% NT8e cells for cancer stem-like cells molecular marker such as ALDH and CD133, respectively. Similarly, CAL27 cells also showed a significantly higher number of oralspheres with 13.5% and 1.59% CAL27 cells positive for ALDH and CD133 (Figure 3B, 3C). In contrast, AW13516 cells expressing comparatively lower NOTCH1 levels showed a reduced spheroid formation capacity with 0.34% ALDH positive and 0.12% CD133 positive while DOK cells did not show any oralsphere formation. To test whether a high fraction of the NT8e population constitutes the stem-like cells, we sorted NT8e cells in ALDH positive and ALDH negative fraction and assessed the sphere-forming efficiency (Supplementary Figure S5A). With subsequent passaging, the cells form ALDH negative population could not maintain their spheroid formation capacity while the ALDH positive population retained their self-renewal capacity demonstrating that indeed NT8e possess high ALDH positive cells are showing cancer stem-like cells features (Supplementary Figure S5B, S5C). *NOTCH1* knockdown clones showed significant reduction in oralsphere formation ability with concomitant decrease in ALDH positive cells in NT8e and AW13516 cells as compared to scrambled (SCR) cells (Figure 3D, 3E), highlighting their dependency on NOTCH1 expression with concomitant decrease in ALDH positive population of cells, thus regulating and promoting the survival of HNSCC CSCs. Next, we attempted to overexpress activated *NOTCH1* and full-length *NOTCH1* in AW13516 cells and checked for the sphere forming efficiency. Activated *NOTCH1* form more number of spheres as compared to vector control cells

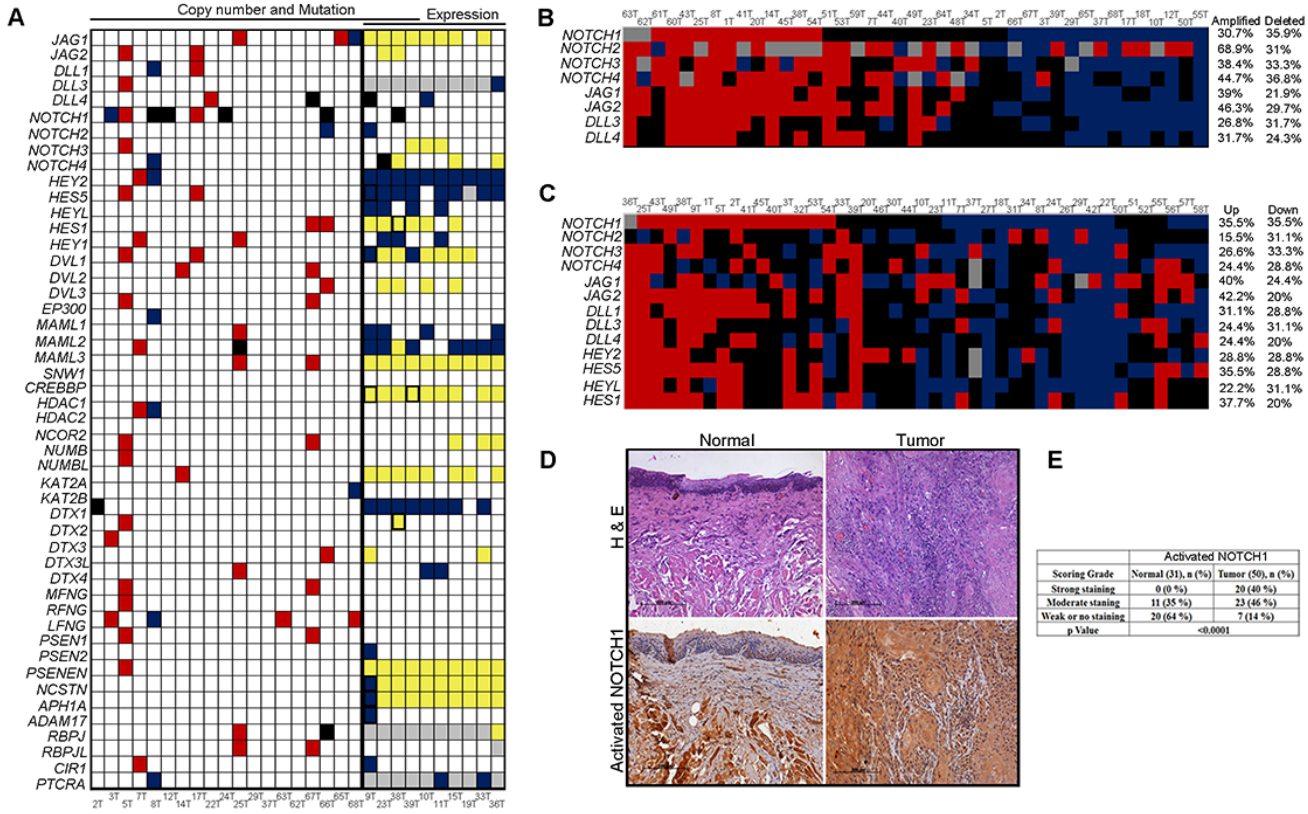


Figure 1: Activation of Notch pathway in early stage tongue squamous cell carcinoma. A. Schematic representation of somatic mutation, copy number changes and expression changes identified in Notch pathway genes (*N*=48) using whole exome and transcriptome sequencing. Red filled; copy number gains, Yellow; high transcript expression, blue; copy number loss and low transcript expression, black; mutation, white; no events, grey; transcript not detected and black borderline boxes ;any two events. Thick black line denoting separation of samples with exome and transcriptome sequencing. **B.** Schematic representation of DNA copy number alteration of Notch pathway genes in a cohort of 41 paired tumor samples estimated by quantitative real-time PCR. Red blocks; high copy number, blue; low copy number, black; diploid and grey color; experiment could not be done or data could not be acquired. **C.** Schematic representation of gene expression of Notch signaling pathway and its downstream targets in the cohort of 44 paired tongue tumor samples. Colors denotes: Red; upregulation, blue; down regulation, black; basal expression and grey color; experiment could not be done or results could not be acquired **D.** Immunohistochemistry (IHC) was performed for activated NOTCH1 in paired normal and tongue tumor samples (*N*=50). Brown color indicates positive expression. Representative IHC stained photomicrographs from normal and tongue tumor samples are shown. Scale bar, 100µM; corresponding H&E stained slides are shown in the upper panel. **E.** Tabular representation for quantification of activated NOTCH1 immunostaining data. Significant differences of IHC staining scores between normal and tumor were estimated using the Chi-square test and p value ≤0.05 was considered as threshold for statistical significance.

post 5 days (Supplementary Figure S6, S6C). However, given that AW13516 cells are HPV negative [25] and harbor wild-type *p16INK4A* and mutant *Tp53* [23], ectopic expression of full-length *NOTCH1* or *NICD* led to continuous cell death and senescence mediated growth arrest (Supplementary Figure S6, S6G), as described earlier [26-28].

Notch pathway inhibitors block stem-like feature, proliferation, and survival of HNSCC cells over expressing NOTCH1

Finally, we investigated whether pharmacological inhibition of Notch pathway activation would be effective against HNSCC cell lines over expressing NOTCH1.

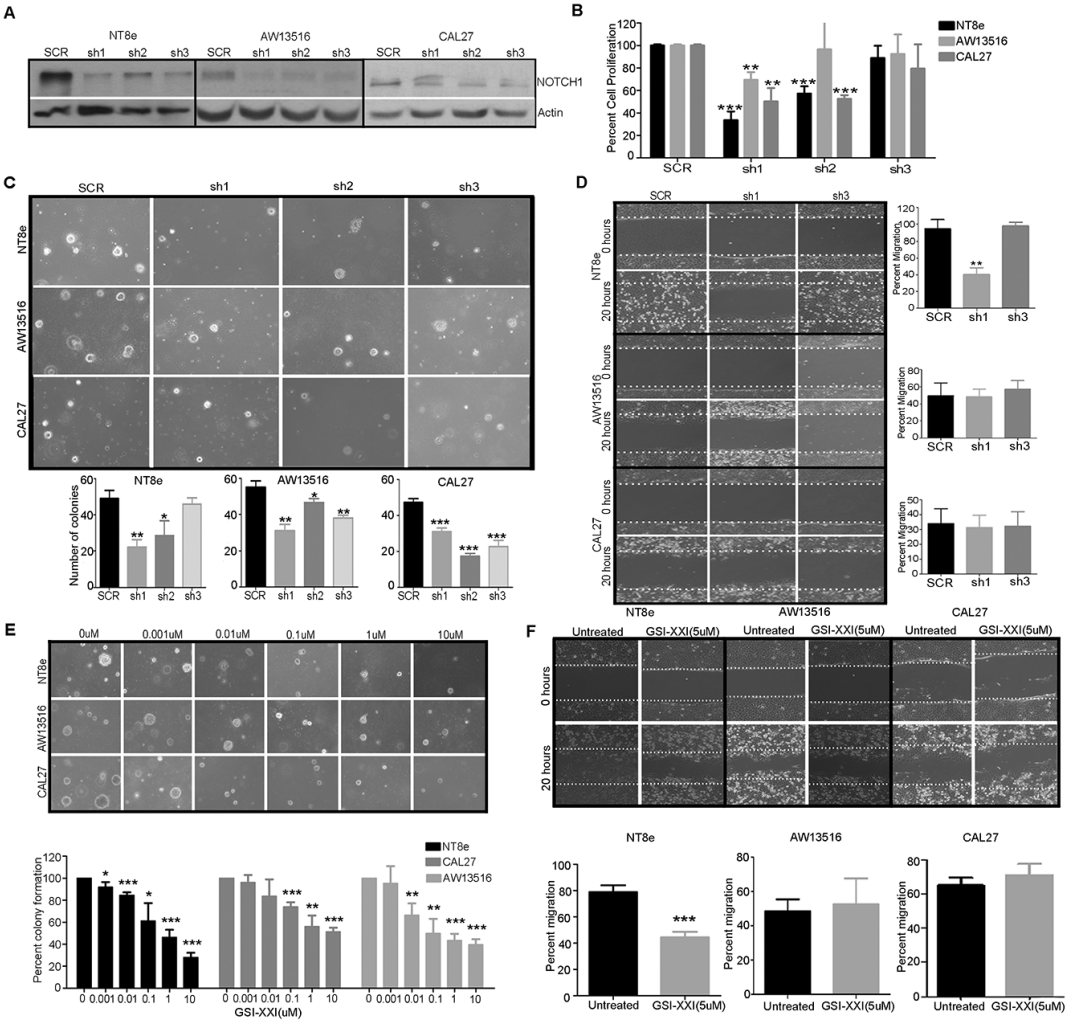


Figure 2: shRNA mediated knockdown and inhibition of NOTCH1 inhibits transformation, survival and migration of HNSCC cells. A. shRNA constructs used to knock down *NOTCH1* expression in NT8e, AW13516, and CAL27 cells. Anti-NOTCH1 immunoblot shows that hairpins knock down to varying extents in different cells. Actin is included as a loading control. SCR, scrambled hairpin used as a negative control. **B.** Infection with 2 of 3 independent hairpins (sh*NOTCH1*#1 and sh*NOTCH1*#2) inhibited cell survival of NT8e and CAL27 cells expressing higher NOTCH1 levels as compared to the AW13516 cells-- as assessed by plotting total cell count on day 6 compared to day 2, normalized against cells infected with SCR. **C.** Infection with independent hairpins inhibit soft agar colony formation by the NT8e and CAL27 cells expressing higher NOTCH1 levels compared to the AW13516 cells (upper panel). Colonies were photographed after 3 weeks (Magnification: ×10). Bar graph representation of soft agar colony formation (lower panel). **D.** Wound healing assay of knockdown clones of NT8e, CAL27 and AW13516 cells. NT8e cells with highest migration potential was most significantly inhibited following infection with sh*NOTCH1* constructs. Percent inhibition of migration was calculated after 20 hours of wound incision. **E.** Representative images of soft agar colony formation (upper panel) and bar graph representation of soft agar colony formation post gamma secretase inhibitor (GSI-XXI) treatment in HNSCC cell lines **F.** Wound healing assay of NT8e, CAL27 and AW13516 cells were performed post GSI-XXI inhibitor as indicated concentration and % migration was calculated after 20 hours of wound healing. Experiment was performed in triplicate and colonies were counted and shown as mean ± SD and P value is denoted as *; $P < 0.01$, **; $P < 0.001$, ***; $P < 0.0001$ versus non-targeting shRNA. Experiments were repeated two times independently.

Treatment of the NT8e, CAL27, AW13516 and DOK HNSCC cell lines with gamma secretase inhibitor (GSI-XXI) that abolished the presence of activated NOTCH1 (Supplementary Figure S4D) that resulted in significant reduction in soft agar colony formation (Figure 2E) and cell survival (Supplementary Figure S4E) as compared to vehicle treated in NT8e and CAL27 cells but not AW13516 cells. Additionally, the migration potential of NT8e cells was significantly inhibited by GSI-XXI, consistent with our observation using shRNA knockdown based approach (Figure 2F). Similarly, marked decrease in spheroid forming ability and ALDH expression in NT8e and AW13516 cells was observed post GSI-XXI treatment (Figure 3F, 3G).

Activation of notch pathway correlates with node positive and non-smoker TSCC patients

Of particular significance is the correlation between clinicopathological characteristics and overall Notch pathway activation: immuno-histochemical based expression of activated NOTCH1 intracellular domain NICD (χ^2=7.10, *P*=0.029), amplification at *DLL4* (χ^2=7.5, *P*=0.023), and transcript over expression of Notch pathway effector genes

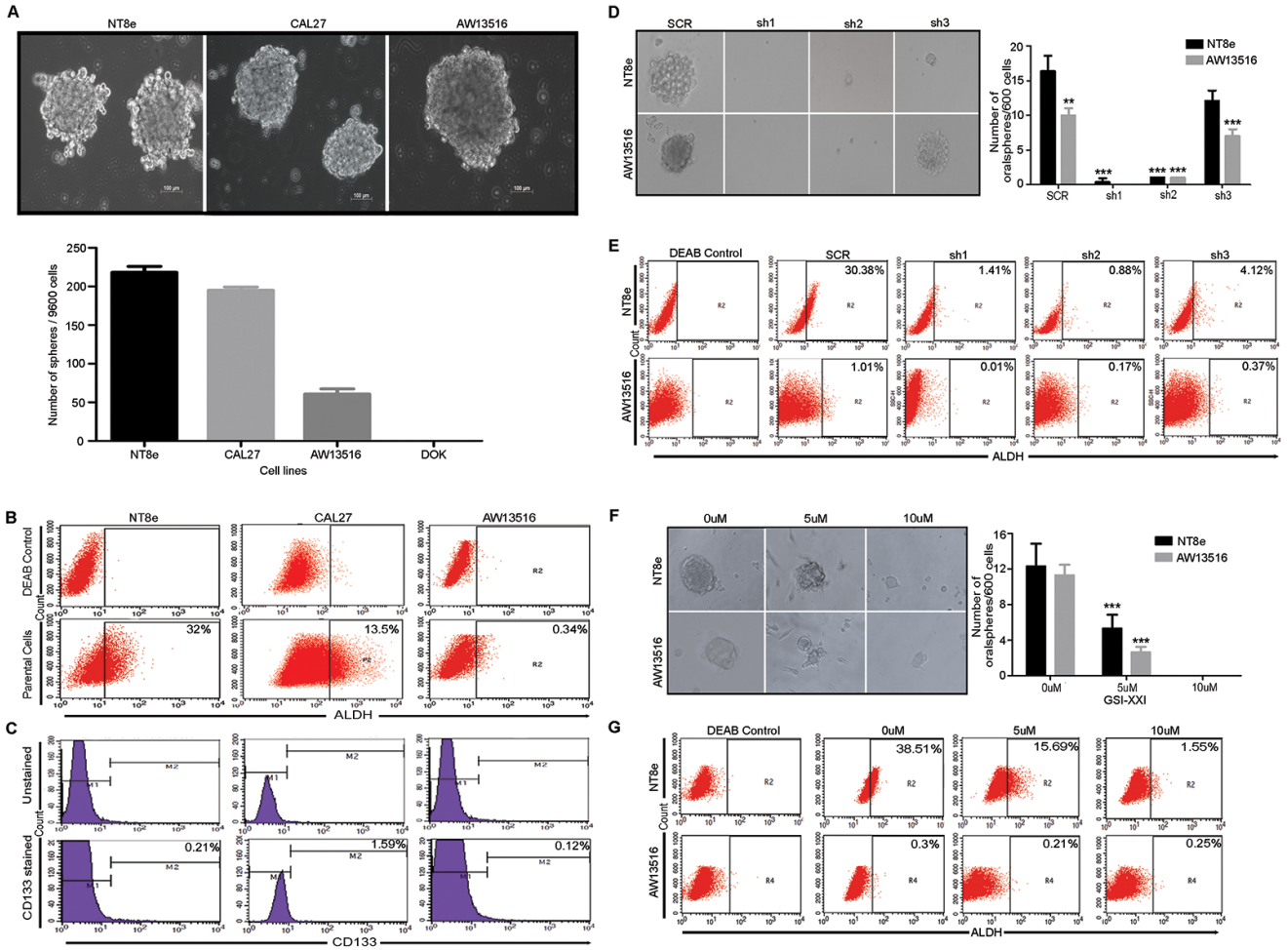


Figure 3: Notch pathway is essential for cancer stem-like property of HNSCC cells. A. Oralsphere formation capacity of HNSCC cells. Representative images of oralsphere are shown in HNSCC cells. Oralsphere (>75µm size) were counted manually in triplicate via visualization under microscope and data was represented as Mean± SD. **B.** and **C.** Analysis of cancer stem-like cells (CSCs) marker ALDH and CD133 in HNSCC cells. Percentage ALDH positive cells were calculated against DEAB control. **D.** Representative images of oralsphere are shown in scrambled (SCR) and different shRNA clones (sh1, sh2 and sh3) of NT8e and AW13516 cells. Oralsphere formation assay was performed in triplicate and counting was done by observing under phase contrast microscope and data was represented as Mean± SD. **E.** ALDH staining for shRNA mediated knockdown clones and GSI-XXI treatment in HNSCC cells, respectively. Percentage ALDH positive cells were calculated against DEAB control. **F.** and **G.** Representative images of oralsphere post GSI-XXI treatment. NT8e and AW13516 cells post respective concentration treatment and ALDH positive cells. Number of oralsphere were counted and represented as Mean± SD. ALDH staining for shRNA mediated knockdown clones and GSI-XXI treatment in HNSCC cells, respectively. Percentage ALDH positive cells were calculated against DEAB control *P-value* *; ≥ 0.05 was considered as threshold for significance. All the above experiment were performed by at least two times independently by separate individuals.

Table 1: Clinical correlation analysis of Notch pathway alterations

Clinicopathologic features	Variable	N (% along column)	*DLL3 Copy Number* (N=26) N (% along row)			P- value	N, % along column	*HEY2 Expression* (N=34) N (% along row)			P- value
			Gain (N=8)	Diploid (N=16)	Deleted (N=2)			Up (N=15)	Basal (N=12)	Down (N=7)	
Nodal Status	Node positive	19 (73%)	5 (26%)	14 (74%)	0 (0%)	**0.023**	22 (65%)	12 (55%)	9 (41%)	1 (5%)	**0.007**
	Node negative	7 (27%)	3 (43%)	2 (29%)	2 (29%)		12 (35%)	3 (25%)	3 (25%)	6 (50%)	

Clinicopathologic features	Variable	N (% along column)	*HES5* Expression (N=34) N (% along row)			P- value	N, % along column	Activated NOTCH1 IHC (N=49) N (% along row)			P- value
			Up (N=15)	Basal (N=12)	Down (N=7)			Strong (N=15)	Moderate (N=12)	Weak or No (N=7)	
Nodal Status	Node positive	23 (66%)	9 (39%)	12 (52%)	2 (9%)	0.057+	23 (66%)	13 (43%)	15 (50%)	2 (7%)	**0.029**
	Node negative	12 (34%)	4 (33%)	3 (25%)	5 (42%)		12 (34%)	6 (32%)	6 (32%)	7 (37%)	

Clinicopathologic features	Variable	N (% along column)	*NOTCH1* transcript expression (N=35), % along row			P- value
			Up (N=14)	Basal (N=9)	Down (N=12)	
Smoking	Smoker	11 (32%)	2 (18%)	6 (55%)	3 (27%)	**0.026**
	Non-smoker	24 (68%)	12 (50%)	3 (13%)	9 (38%)	

All clinical correlation analysis were performed in SPSS and significant alterations has been presented in the table. Patient's samples showing strong and moderate staining of activated NOTCH1 was considered as having activated Notch signaling. N; Number of samples, Up; upregulation, Down; downregulation. Chi-square test was used to calculate statistical significance. Significant *P-value* are highlighted in bold font. *P-value* ≤ 0.05 was considered as threshold for significance. + denotes; marginally significant.

HEY2 (χ^2=9.8, *P*=0.007) and *HES5* (χ^2=5.71 *P*=0.057) significantly correlated with lymph node metastases (Table 1) and poor prognosis (Supplementary Figure S7A). Interestingly, this was consistent also with our analysis of the TCGA tongue cancer patient dataset (Supplementary Figure S7B), and with other cancers [24, 29, 30].

In addition, *NOTCH1* expression significantly correlated with a non-smoking habit of patients (χ^2=7.325, *P*=0.026), where 12 of 24 non-smokers patients derived tumors showed upregulation of *NOTCH1* transcript, consistent with previously described *NOTCH1* upregulation in non-smokers in other diseases including lung adenocarcinoma [31-33]. We also observed a significant correlation with AJCC (American Joint committee on Cancer) TNM tumor staging wherein stage III-IVA showed increases activation of NOTCH pathway (χ^2=7.84, P=0.02). However, no statistically significant correlation was observed between the activated NOTCH1 expression with the sex, age, alcohol and tobacco consumption in the cohort, as represented in Supplementary Table S2. Next, we performed an interim analysis and assessed disease-free survival (DFS) by IHC defined activated NOTCH1 (strong and moderate staining) status vs non-activated NOTCH1 (weak or no staining) status. DFS was defined as time interval between the date of registration and the date of first documented evidence of relapse at any site (local, regional, metastatic, or secondary primary) or death from any cause, whichever earlier. There was no statistically significant difference was observed in tumors with activated NOTCH1 compared to those with non-activated NOTCH1 tumors, as shown in Supplementary Figure S7C.

Taken together, we present a novel clinicopathological correlation such that expression of Notch pathway components and activated NOTCH1 levels predispose TSCC patients to lymph node metastasis, and that non-smokers TSCC patients tend to have higher *NOTCH1* levels as compared to smokers. Clinically, determination of NICD by immuno-histochemistry could be a good predictor of nodal status. This could be a biomarker to predict lymph node metastasis for therapeutic utility among early stage tongue cancer patient to help patient stratification for treatment [34, 35].

DISCUSSION

We demonstrate that 40% TSCC tumors have strong Notch pathway activation and that this property may be important in the maintenance of stem cell component in these tumors. Genetic or chemical perturbation of NOTCH pathway using shRNA and GSI-XXI showed decrease soft agar colony formation, migration potential and cancer stem-like features of HNSCC cells, highlighting their dependency on *NOTCH1* expression. Thus, targeted elimination of these cells may provide a new lead in treatment of head and neck cancer. These findings are consistent with reports where Notch signaling has been shown to be required for stem cell-like features in several cancer types [4, 13, 37]. Interestingly, genetic determinants of cancer stem cells share features with their role in the development of tumorigenesis [38]. These findings are consistent with reports where Notch signaling has been shown to be required for stem cell-like features in several cancer types, including HNSCC [14, 15].

Clinically, NOTCH1 transcript expression significantly correlate with non-smoking habit of patients, consistent with previous reports in other pathological conditions including lung adenocarcinoma [31-33]; lymph node metastasis in tongue cancer correlate with poor prognosis and survival of the patients, thus activated NOTCH1 could serve as a reliable marker to predict lymph node metastasis [35]. Moreover, AJCC TNM tumor stage III-IVA significantly correlates with activation of NOTCH pathway as compared to stage I-II, consistent with reports in HNSCC [39]. The sample size in this study, however, is underpowered to reach the statistical significance for survival data. No significant difference was observed in disease-free survival of the patients with IHC defined activated NOTCH1 tumors as compared to non-activated NOTCH1, as shown in Supplementary Figure S7C.

In conclusion, we demonstrate that a considerable fraction of TSCC tumors has upregulated Notch pathway and that this property may be important for the maintenance of stem cell component in these tumors. And that, NOTCH1 could be a potential therapeutic target in these patients.

MATERIALS AND METHODS

Patient Samples

Tumor-normal paired samples were collected at Tata Memorial Hospital and Advanced Centre for Treatment, Research and Education in Cancer (ACTREC), Mumbai. Sample set and study protocols were approved by (ACTREC-TMC) Internal Review Board (IRB) and most of the patients were recruited from 2010-2013 with a predefined inclusion criteria of early (pT1 and pT2) stage. Percent tumor content was determined using hematoxylin and eosin based staining by two independent pathologists which varied from 60 to 90%. Patient samples and characteristics are provided in the Supplementary Table S3.

DNA and RNA extraction

DNA from tongue primary paired normal-tumor tissue samples were extracted using DNeasy Blood and tissue DNA extraction kit (Qiagen) according to manufacturer's instructions. DNA was quantified using Nanodrop 2000c Spectrophotometer (Thermo Fisher Scientific Inc.) and DNA quality was checked by resolving on 0.8% agarose gel. Total RNA was extracted from tongue primary paired normal-tumor samples and cell lines using RNeasy RNA isolation kit (Qiagen) and Trizol reagent (Invitrogen) based methods and later resolved on 1.2% Agarose gel to confirm the RNA integrity.

Exome capture and NGS DNA sequencing

Two different Exome capture kits were used to capture exome for different samples. The TruSeq Exome Enrichment kit (Illumina) was used to capture 62Mb region of human genome comprising of 201,121 exons representing 20,974 gene sequences, including 5'UTR, 3'UTR, microRNAs and other non-coding RNA and NimbleGen SeqCap EZ Exome Library v3.0 was also used to capture 64 Mb region of the human genome. Exome library preparation and sequencing was performed as per manufacturer's instructions. Briefly, 2 µg genomic DNA was sheared using Covaris (Covaris Inc) for generating the fragment size of 200-300bp size. DNA libraries were prepared from both the kits were quantified by qPCR using KAPA Library Quant Kit (Kapa Biosystems) in ABI 7,900HT system (Life Technologies). Seven pmol of 6-plex DNA library pool was loaded per lane on flow cell (Flow Cell v3) to generate clusters using TruSeq PE (Paired-End) Cluster Kit v3-cBot-HS kit and clustered flow was sequenced for 201 cycles on HiSeq-1,500 System (Illumina) using TruSeq SBS Kit v3 (Illumina) at in-house core NGS facility.

Identification of somatic mutations from exome Sequencing

Paired-end raw sequence reads generated were mapped to the human reference genome (build hg19) using BWA v. 0.6.2 [40]. Mapped reads were then used to identify and remove PCR duplicates using Picard tools v1.100 (http:broadinstitute.github.io/picard/). Base quality score recalibration and indel re-alignment performed and variants were called from each sample separately using GATK 2.5-2 Unified Genotyper and MuTect v. 1.0.27783 [41]. Post subtraction of variants from its paired normal, remaining variants was taken for further analysis if they were having ≥5 altered reads. Furthermore, all samples variants were further filtered against pooled normal variants database (N=62) to reduce the possibility of the germline variation We further annotated variants using Oncotator v1.1.6.0 [42] and dbSNP v142 [43] and COSMIC database v68 [44] using an in-house developed script. Later, we performed functional prediction tool based analysis for somatic non-synonymous variants using nine different tools such as: dbNSFP v2.0 (includes SIFT, Polyphen2_HDIV, Polyphen2_HVAR, LRT, Mutation Taster, Mutation Accessor and FATHMM) [45], CanDRA v1.0 [46] and Provean v.1.1 [47]. Variants called deleterious in nature by at least one software was taken for further analysis. We confirmed the identity of mutations by manual visualization in IGV [48, 49].

Somatic copy number analysis from exome sequencing data

BAM files prepared for variant calling were used for copy number analysis using Control-FREEC [50]. Paired tumor-normal samples BAM files were fed into Control-FREEC along with target region for Illumina and Nimblegen exome kits as bed file. Read count were generated and normalized for GC content for each of the target region followed by computation of ratio of read count in a tumor to normal. Read count ratio was converted to copy numbers followed by segmentation using lasso method. Segmented copy number data generated by control-FREEC was further used for annotation and post-processing using R programming.

Transcriptome sequencing and data analysis

Transcriptome libraries for sequencing were constructed according to the TruSeq RNA library protocol (Illumina). Briefly, mRNA was purified from 4 µg of intact total RNA using oligodT beads and library preparation was done as per manufacturer's instructions (TruSeq RNA Sample Preparation Kit, Illumina). 7pmol of quantified cDNA libraries were loaded on Illumina flow cell (v3) to generate clusters using TruSeq PE (Paired-End) Cluster Kit v3-cBot-HS kit and clustered flow was sequenced for 201 cycles on HiSeq-1500 System (Illumina) using TruSeq PE Cluster Kit v3 and TruSeq SBS Kit v3 (Illumina) to generate at least 30 million reads per sample. Post sequencing, de-multiplexing was carried out on the basis of index sequences using CASAVA (version 1.8.4, Illumina). Transcriptome data analysis was performed using Tuxedo-suite pipeline [51]. In brief, alignment of short reads was done against reference genome (hg19) using TOPHAT2 v. 2.0.8b [52] in which 95-99% of reads were mapped to the reference genome. Cufflinks v.2.0.2 was used to find the expressed transcripts in the data and quality control steps was performed using CummeRbund package v2.0. All the actively expressed transcripts per samples were then binned by $\log_{10}$(FPKM+1) to differentiate the significantly expressed transcripts from the background noise and transcripts represented by <0.1 $\log_{10}$(FPKM+1) were filtered out from further analysis. Since paired normal of these tumors cannot be obtained, we defined a significant change in expression for those genes whose expression is higher (>80%) or lower (<20%) than the median expression as suggested [53].

Analysis of the cancer genome atlas tongue cancer data

The Cancer Genome Atlas (TCGA) dataset of HNSCC including DNA copy number dataset (gistic2 threshold) from 452 HNSCC tumor tissue, RNA seq expression (Illumina HiSeq) dataset from 541 HNSCC was downloaded from UCSC Cancer genome browser on 20th June 2014. Later, tongue cancer patient data for DNA copy number (n=126) and gene expression (n=129 has been taken for further analysis. For expression and DNA copy number the median centered RSEM counts and gene-level copy number estimates have been used, respectively (n=126). Notch pathway genes (n=13) data has been retrieved and heatmaps were generated using MeV 4.9.0. The RNAseq gene expression data has been retrieved for Notch pathway genes (n=13) and raw data has been median centered using Cluster 3.0 software. The median centered RSEM counts for each gene has been used to generate heatmap using MeV 4.9.0. Fold change criteria was ≥1.5 fold change for upregulation, ≤1.499 fold change for down-regulation and in between -1.5 to 1.499 fold change was denoted as a basal expression. DNA copy number and Expression correlation analysis and clinical correlation analysis have been performed using SPSS. P=value <0.05 was criteria for statistical significance.

Tissue processing

Surgically resected oral tumor tissues and matched nonmalignant (cut margins) adjacent tissues were obtained from patients with informed consent after IRB approval

from ACTREC. These tissues were processed for paraffin embedding and sectioned at 4µm for H/E staining for evaluation of tumor.

Immunohistochemistry

Immunohistochemistry was done following the standard protocol of DAKO Envision Flex. Briefly, the slides were microwaved by incubating them for 10 minutes in high pH antigen Retrieval Solution (DAKO;DM828), then allowed to cool to room temperature before rinsing with Tris-buffered saline wash buffer (DAKO;DM831). Endogenous peroxidase activity was blocked by incubating the slides for 20 minutes in 3% hydrogen peroxide (EnVision/HRP, Dako). After rinsing in wash buffer, the sections were incubated for 3hours at room temperature with the monoclonal human anti-activated Notch1 antibody (Cat.ab8925; dilution 1:50) in Tris-HCl buffer antibody diluent (Dako; K8016). Slides were rinsed in wash buffer (DAKO; DM831) and incubated for 90 minutes with peroxidase-labeled polymer conjugated to goat anti-rabbit immunoglobulins (EnVision/HRP, Dako; SM801). The chromogenic reaction was carried out with 3,3′-diaminobenzidine chromogen solution for 5 minutes, resulting in the expected brown-colored signal. Finally, after rinsing with deionized water, the slides were counterstained with hematoxylin, dehydrated, mounted with toluene-based mounting medium (Thermo Scientific Richard-Allan) and cover slip.

Immunohistochemical staining analysis

Evaluation of immunohistochemical staining of activated Notch1 expression was scored as 0, 1+, 2+ and 3+. The percentage of cells with positive staining was scored from 0 to 4 (0=0% positive cells; 1: <10% positive cells; 2: 10-49% positive cells; 3: 50-80% positive cells; 4: >80% positive cells) and staining intensity was scored from 0 to 3 (0, negative; 1, weak; 2, moderate; 3, strong). The two scores were then multiplied. Final scores of 0-2 were scored as 0, 3-5 as 1, 6-8 as 2 and 9-12 as 3.6

Quantitative real-time PCR for copy number analysis

Primers details used for copy number study has been provided in Supplementary Table S4. All primers used have been tested for their specificity by performing evaluative PCR as well as melt curve analysis during quantitative real-time PCR. Amplification efficiency for all primer was tested with series of dilutions (0.625 ng, 1.25 ng, 2.5 ng, 5 ng, 10 ng) of genomic DNA and PCR amplification efficiency was ~97%(~R^2=0.979) (Supplementary Figure S8A, S8B). Based on above quality control, 10 ng of genomic DNA per 10 µl reaction volume in triplicates were run on Light cycler 480 (Roche, Mannheim, Germany) twice independently and relative copy number analysis was performed as described previously [54]. The threshold for calling high and low copy number was ≥2.5 and ≤1.5, respectively and ≤2.5 and 1.5≤ ; diploid.

Quantitative real-time RT-PCR for expression analysis

Prepared cDNA was diluted 1:10 and reaction were performed in 10µl volume in triplicate. The melt curve analysis was performed to check the primer dimer or non-specific amplifications. Real-time PCR was carried out using KAPA master mix (KAPA SYBR® FAST Universal qPCR kit) in 10 µl volume in triplicate on Light cycler 480 (Roche, Mannheim, Germany) machine. All the experiments were repeated at least twice independently. The data was normalized with internal reference *GAPDH*, and analyzed by using delta-delta Ct method described previously The criteria were ≥2 fold change for upregulation, ≤0.5 fold change for down-regulation and in between 1.99-0.501 fold change as a basal expression. The details of all the primers used for expression analysis have been provided in Supplementary table S4.

Cell culture

Cell lines established from different sub-sites of head and neck cancer: AW13516 from tongue, NT8e from upper aero-digestive tract, CAL27 cells from tongue [55] and partially transformed cell line DOK (tongue) [56] were used in this study. AW13516 and NT8e were acquired from Tata Memorial Hospital while CAL27 and DOK cells were procured from ATCC and Sigma, respectively. All cells were grown in Dulbecco's Modified Eagle Medium (Pan biotech, Germany). Culture media was supplemented with 10% FBS (Gibco, US), 1% Penicillin-Streptomycin solution (Sigma) and maintained at 37°C in an incubator with 5% CO2. DOK cells were grown with 5ug/ml hydrocortisone (Sigma) as a supplement. Trypsinization was performed using 0.25% Trypsin-EDTA (Invitrogen) and freezing of cells performed in 90 % FBS (Gibco, US) and 10% DMSO (Sigma) and were stored in liquid Nitrogen for long term storage. All the cell lines were authenticated using a short tandem repeat (STR) analysis kit (Gene Print v10, Promega, USA). The results are shown in Supplementary Table S5.

Retrovirus production, infection and drug selection

Retroviral shRNA constructs were purchased from TransOMIC technologies, USA. Target sequences of *NOTCH1* shRNA constructs: sh1 5'-CAGTGAGCGA TGACTGCACAGAGAGCTCCTAT-3', sh2 5'-CAGT GAGCGATGGACGGACCCAACACTTACAT-3', and sh3 5'-CAGTGAGCGAGACGAGGACCTGGAGA CCAAAT-3'. 293T cells were seeded in 6 well plates

one day before transfection and each construct (pMLP Retroviral-puro) along with pCL-ECO and pVSVG helper vector were transfected using Lipofectamine LTX reagent (Invitrogen). The viral soup was collected 48 and 72 hours post transfection, passed through 0.45µM filter and stored at 4OC. Respective cells for transduction were seeded one day before infection in a six-well plate and allowed to grow to reach 50-60% confluency. One ml of the virus soup (1:5 dilution) and 8µg/ml of polybrene (Sigma) was added to cells and incubated for six hours. Cells were maintained under puromycin (Sigma) selection.

Overexpression of NOTCH1 and selection

The human full-length *NOTCH1* (pcDNA-*NOTCH1*) [57] was obtained from Artavanis-Tsakonas Laboratory (Havard Medical School) and activated *NOTCH1* (pEGFP-NICD) [58] constructs was obtained from Annapoorni Rangarajan (Indian Institute of Sciences (IISc), Bangalore, India). Cells expressing pcDNA-*NOTCH1 or* pEGFP-NICD were generated by transfection with 10µg of DNA using Lipofectamine 3000 (Invitrogen) as per manufacturer's instructions. After 48hours, cells were cultures for 8-10 days in complete medium supplemented with 1mg/ml of G418 for antibiotics selection of transfected cells or cells were sorted based on GFP expression using BD FACSAria II. Pooled GFP sorted or antibiotics selected cells were later used for oralsphere assay. In case of 293T cells, post 48hours transfection cells were taken for RNA extraction and protein extraction for quantitative real-time PCR and western blot analysis, respectively.

Western blotting

Cells were lysed in RIPA buffer (Sigma) and protein concentration was estimated using BCA (MP biomedicals) method [59]. Forty microgram protein was separated on 10% SDS-PAGE gel, transferred to nitrocellulose membrane and transfer was verified using Ponceau S (Sigma). Later the blots were blocked in Tris-buffered saline containing 5% BSA (Sigma) and 0.01% Tween-20(Sigma) and were probed with full-length NOTCH1 (sc-6014-R, Santacruz biotechnology), anti-activated NOTCH1 antibody (Abcam; ab8925) and anti-actin (A5316, Sigma) antibody. The membranes were then incubated with corresponding secondary HRP-conjugated antibodies (Santa Cruz Biotechnology, USA) and the immune complexes were visualized by Pierce ECL (Thermo Scientific, USA) according to manufacturer's protocol. Western blot experiments were performed in triplicate.

Anchorage-independent growth assay

For analysis of growth in soft agar, 5 × 103 cells were seeded in triplicate onto a six-well dish (Falcon) in 4 ml of complete medium containing 0.33% agar solution along with respective treatments of GSI-XXI at 37°C in CO2 incubator. Ten images per well were photographed after 21 days using inverted phase contrast microscope and colonies were counted manually.

MTT assay

A Thousand cells per well (six replicate per concentration) were seeded in 96-well plate followed by incubation with the drug for 72 hours and subsequently incubated with MTT (0.5 mg/ml) for 4 hours. Later, MTT assay was performed and data was acquired at 570nm using Microplate reader. Percentage cell viability was calculated against vehicle treated control.

Wound healing assay

The cells were grown in 6 well plates to 95% confluency and were replaced with fresh medium containing 5/ml mitomycin C (Sigma). After 2 hours incubation, the medium was discarded and wounds were scratched with the help of sterile 10µl pipette tip. Cells were washed with PBS to remove the detached cells during creating a wound. The cells were fed with fresh medium and observed by time-lapse microscopy, and images were taken every 10 min for 20 hr. Migration was measured using Image J software.

Oralsphere formation assay

Ninety-six hundred cells were seeded in 1.2 % agar coated 6-well plates supplemented with stem cell media (recombinant EGF (20 ng/ml), human basic FGF (20 ng/ml), L-glutamine (2 mM), B-27 supplement and N2 supplement) and allowed to grow for 10 days. After every five days media, additional media was supplemented. Five hundred cells from NT8e, AW13516 and CAL27 shRNA clones were seeded on an ultra-low adherent 96-well plate in stem cell medium. Oralspheres were then cultured and maintained in low adherent 24-well plates. Additionally, the parent NT8e, AW13516, and CAL27 cells were also checked for the spheroid formation capacity upon 5 µM and 10 µM GSI-XXI administration using the same conditions.

ALDH activity and CD133 staining

The ALDH activity was checked using ALDEFLUOR™ detection kit (StemCell Technology, 01700) following the kit protocol and data was acquired on FACS Caliber and analysis was carried out using CellQuest software. For CD133 staining was performed using CD133 (AC133) antibody (MACS Miltenyi Biotech) in FACS buffer for 15 min in dark at 4 °C. The cells were then washed twice with staining buffer and acquired on FACS Caliber, BD Biosciences.

β-Galactosidase activity staining

Ten thousand cells were seeded in 12 well plates in triplicates and next day, AW13516 cells, vector control and overexpressing full length NOTCH1 were washed with 1X PBS and fixed with 0.5ml of fixative solution in the Abcam Senescence detection kit (Ab65351) for 10–15min at 25°C. Fixed cells then washed twice with 1X PBS and incubated for 8 hours with 0.5ml of staining solution containing 20mg/ml of X-gal. Stained cells were microscopically analyzed using Olympus IX-71. Images were analyzed using Image J and percentage β-Galactosidase positive cells were plotted.

Survival and Statistical analysis

The relative impact of Notch pathway alterations on disease free survival (DFS) of TSCC patients was analyzed using Kaplan-Meier method [60] and was compared using the log-rank test for statistical significance. Data are expressed as mean ± standard deviation (SD) or standard error (SE). Significant differences between selected two groups were estimated using unpaired Student t-test using Graph Pad prism version 5. Statistical significance was set at $p \leq 0.05$. Pearson correlation analysis and chi-square tests were performed in IBM SPSS statistics software version 21 for correlation analysis.

ACKNOWLEDGMENTS

All members of the Dutt laboratory for critically reviewing the manuscript. Sandor Proteomics Pvt. Ltd. and Scigenome Labs, for providing Exome and Transcriptome library preparation services. A.D. is supported by an Intermediate Fellowship from the Wellcome Trust/DBT India Alliance (IA/I/11/2500278), by a grant from DBT (BT/PR2372/AGR/36/696/2011), and intramural grants (IRB project 92 and 55). P.U. is supported by a senior research fellowship from CSIR. E.K. is supported by a senior research fellowship from UGC. P.C. and M.G are supported by a senior research fellowship from ACTREC. N.G is supported by a junior research fellowship from Tata Memorial hospital. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

CONFLICTS OF INTEREST

The authors declare no competing financial interests.

FUNDING

This work was supported by Wellcome Trust/DBT India Alliance (IA/I/11/2500278) and intramural grants (IRB project 92 and 55).

REFERENCES

1. Ntziachristos P, Lim JS, Sage J and Aifantis I. From fly wings to targeted cancer therapies: a centennial for notch signaling. Cancer cell. 2014; 25:318-334.
2. Egloff AM and Grandis JR. Molecular pathways: context-dependent approaches to Notch targeting as cancer therapy. Clin Cancer Res. 2012; 18:5188-5195.
3. Izumchenko E, Sun K, Jones S, Brait M, Agrawal N, Koch W, McCord CL, Riley DR, Angiuoli SV, Velculescu VE, Jiang WW and Sidransky D. Notch1 mutations are drivers of oral tumorigenesis. Cancer Prev Res (Phila). 2015; 8:277-286.
4. Sun W, Gaykalova DA, Ochs MF, Mambo E, Arnaoutakis D, Liu Y, Loyo M, Agrawal N, Howard J, Li R, Ahn S, Fertig E, Sidransky D, Houghton J, Buddavarapu K, Sanford T, et al. Activation of the NOTCH pathway in head and neck cancer. Cancer research. 2014; 74:1091-1104.
5. Pickering CR, Zhang J, Yoo SY, Bengtsson L, Moorthy S, Neskey DM, Zhao M, Ortega Alves MV, Chang K, Drummond J, Cortez E, Xie TX, Zhang D, Chung W, Issa JP, Zweidler-McKay PA, et al. Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. Cancer Discov. 2013; 3:770-781.
6. Rettig EM, Chung CH, Bishop JA, Howard JD, Sharma R, Li RJ, Douville C, Karchin R, Izumchenko E, Sidransky D, Koch W, Califano J, Agrawal N and Fakhry C. Cleaved NOTCH1 Expression Pattern in Head and Neck Squamous Cell Carcinoma Is Associated with NOTCH1 Mutation, HPV Status, and High-Risk Features. Cancer Prev Res (Phila). 2015; 8:287-295.
7. Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, Kryukov GV, Lawrence MS, Sougnez C, McKenna A, Shefler E, Ramos AH, Stojanov P, Carter SL, Voet D, Cortes ML, et al. The mutational landscape of head and neck squamous cell carcinoma. Science. 2011; 333:1157-1160.
8. Agrawal N, Frederick MJ, Pickering CR, Bettegowda C, Chang K, Li RJ, Fakhry C, Xie TX, Zhang J, Wang J, Zhang N, El-Naggar AK, Jasser SA, Weinstein JN, Trevino L, Drummond JA, et al. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. Science. 2011; 333:1154-1157.
9. Jemal A, Bray F, Center MM, Ferlay J, Ward E and Forman D. Global cancer statistics. CA: a cancer journal for clinicians. 2011; 61:69-90.
10. Rothenberg SM and Ellisen LW. The molecular pathogenesis of head and neck squamous cell carcinoma. The Journal of clinical investigation. 2012; 122:1951-1957.
11. Suman S, Das TP and Damodaran C. Silencing NOTCH signaling causes growth arrest in both breast cancer stem cells and breast cancer cells. Br J Cancer. 109:2587-2596.
12. Wang J, Sullenger BA and Rich JN. Notch signaling in cancer stem cells. Advances in experimental medicine and biology. 2012; 727:174-185.

13. Shrivastava S, Steele R, Sowadski M, Crawford SE, Varvares M and Ray RB. Identification of molecular signature of head and neck cancer stem-like cells. Scientific reports. 2015; 5:7819.
14. Zhao ZL, Zhang L, Huang CF, Ma SR, Bu LL, Liu JF, Yu GT, Liu B, Gutkind JS, Kulkarni AB, Zhang WF and Sun ZJ. NOTCH1 inhibition enhances the efficacy of conventional chemotherapeutic agents by targeting head neck cancer stem cell. Scientific reports. 2016; 6:24704.
15. Lee SH, Do SI, Lee HJ, Kang HJ, Koo BS and Lim YC. Notch1 signaling contributes to stemness in head and neck squamous cell carcinoma. Lab Invest. 2016; 96:508-516.
16. Chandrani P, Kulkarni V, Iyer P, Upadhyay P, Chaubal R, Das P, Mulherkar R, Singh R and Dutt A. NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome. British journal of cancer. 2015; 112:1958-1965.
17. Laprise C, Madathil SA, Allison P, Abraham P, Raghavendran A, Shahul HP, ThekkePurakkal AS, Castonguay G, Coutlee F, Schlecht NF, Rousseau MC, Franco EL and Nicolau B. No role for human papillomavirus infection in oral cancers in a region in southern India. Int J Cancer. 2015.
18. Patel KR, Vajaria BN, Begum R, Desai A, Patel JB, Shah FD, Shukla SN and Patel PS. Prevalence of high-risk human papillomavirus type 16 and 18 in oral and cervical cancers in population from Gujarat, West India. Journal of oral pathology & medicine : official publication of the International Association of Oral Pathologists and the American Academy of Oral Pathology. 2014; 43:293-297.
19. Pathare SM, Gerstung M, Beerenwinkel N, Schaffer AA, Kannan S, Pai P, Pathak KA, Borges AM and Mahimkar MB. Clinicopathological and prognostic implications of genetic alterations in oral cancers. Oncology letters. 2011; 2:445-451.
20. Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups. Nat Commun. 2013; 4:2873.
21. Vettore AL, Ramnarayanan K, Poore G, Lim K, Ong CK, Huang KK, Leong HS, Chong FT, Lim TK, Lim WK, Cutcutache I, McPherson JR, Suzuki Y, Zhang S, Skanthakumar T, Wang W, et al. Mutational landscapes of tongue carcinoma reveal recurrent mutations in genes of therapeutic and prognostic relevance. Genome medicine. 2015; 7:98.
22. Izumchenko E, Sun K, Jones S, Brait M, Agrawal N, Koch W, McCord CL, Riley DR, Angiuoli SV, Velculescu VE, Jiang WW and Sidransky D. Notch1 mutations are drivers of oral tumorigenesis. Cancer prevention research. 2015; 8:277-286.
23. Pratik Chandrani PU, Prajish Iyer, Mayur Tanna, Madhur Shetty, Gorantala Venkata Raghuram, Ninad Oak, Ankita Singh, Rohan Chaubal and Manoj Ramteke SG, Amit Dutt. Integrated genomics approach to identify biologically relevant alterations in fewer samples. BMC genomics. 2015; 16:936.
24. Takebe N, Harris PJ, Warren RQ and Ivy SP. Targeting cancer stem cells by inhibiting Wnt, Notch, and Hedgehog pathways. Nat Rev Clin Oncol. 2011; 8:97-106.
25. Chandrani P, Kulkarni V, Iyer P, Upadhyay P, Chaubal R, Das P, Mulherkar R, Singh R and Dutt A. NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome. British journal of cancer. 2015; 112:1958-1965.
26. Kagawa S, Natsuizaka M, Whelan KA, Facompre N, Naganuma S, Ohashi S, Kinugasa H, Egloff AM, Basu D, Gimotty PA, Klein-Szanto AJ, Bass AJ, Wong KK, Diehl JA, Rustgi AK and Nakagawa H. Cellular senescence checkpoint function determines differential Notch1-dependent oncogenic and tumor-suppressor activities. Oncogene. 2015; 34:2347-2359.
27. Lefort K, Mandinova A, Ostano P, Kolev V, Calpini V, Kolfschoten I, Devgan V, Lieb J, Raffoul W, Hohl D, Neel V, Garlick J, Chiorino G and Dotto GP. Notch1 is a p53 target gene involved in human keratinocyte tumor suppression through negative regulation of ROCK1/2 and MRCKalpha kinases. Genes Dev. 2007; 21:562-577.
28. Kunnimalaiyaan M, Vaccaro AM, Ndiaye MA and Chen H. Overexpression of the NOTCH1 intracellular domain inhibits cell proliferation and alters the neuroendocrine phenotype of medullary thyroid cancer cells. The Journal of biological chemistry. 2006; 281:39819-39830.
29. Alniaimi AN, Demorest-Hayes K, Alexander VM, Seo S, Yang D and Rose S. Increased Notch1 expression is associated with poor overall survival in patients with ovarian cancer. International journal of gynecological cancer : official journal of the International Gynecological Cancer Society. 2015; 25:208-213.
30. Zhou L, Yu L, Ding G, Chen W, Zheng S and Cao L. Overexpressions of DLL4 and CD105 are Associated with Poor Prognosis of Patients with Pancreatic Ductal Adenocarcinoma. Pathology oncology research : POR. 2015; 21:1141-1147.
31. Huang J, Song H, Liu B, Yu B, Wang R and Chen L. Expression of Notch-1 and its clinical significance in different histological subtypes of human lung adenocarcinoma. Journal of experimental & clinical cancer research : CR. 2013; 32:84.
32. Tilley AE, Harvey BG, Heguy A, Hackett NR, Wang R, O'Connor TP and Crystal RG. Down-regulation of the notch pathway in human airway epithelium in association with smoking and chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine. 2009; 179:457-466.
33. Wang G, Xu Z, Wang R, Al-Hijji M, Salit J, Strulovici-Barel Y, Tilley AE, Mezey JG and Crystal RG. Genes associated with MUC5AC expression in small airway epithelium of human smokers and non-smokers. BMC medical genomics. 2012; 5:21.
34. Ren ZH, Xu JL, Li B, Fan TF, Ji T and Zhang CP. Elective versus therapeutic neck dissection in node-negative oral

cancer: Evidence from five randomized controlled trials. Oral Oncol. 2015; 51:976-981.

35. D'Cruz AK, Vaish R, Kapre N, Dandekar M, Gupta S, Hawaldar R, Agarwal JP, Pantvaidya G, Chaukar D, Deshmukh A, Kane S, Arya S, Ghosh-Laskar S, Chaturvedi P, Pai P, Nair S, et al. Elective versus Therapeutic Neck Dissection in Node-Negative Oral Cancer. The New England journal of medicine. 2015; 373:521-529.
36. Hijioka H, Setoguchi T, Miyawaki A, Gao H, Ishida T, Komiya S and Nakamura N. Upregulation of Notch pathway molecules in oral squamous cell carcinoma. Int J Oncol. 2010; 36:817-822.
37. Zhang T, Liu H, Liang Y, Liang L, Liao G, Wu J and Huang H. [The expression and significance of the Notch signaling pathway molecules in tongue squamous cell carcinoma]. Hua Xi Kou Qiang Yi Xue Za Zhi. 2013; 31:303-309.
38. Singh SK, Clarke ID, Terasaki M, Bonn VE, Hawkins C, Squire J and Dirks PB. Identification of a cancer stem cell in human brain tumors. Cancer Res. 2003; 63:5821-5828.
39. Yoshida R, Nagata M, Nakayama H, Niimori-Kita K, Hassan W, Tanaka T, Shinohara M and Ito T. The pathological significance of Notch1 in oral squamous cell carcinoma. Lab Invest. 2013; 93:1068-1081.
40. Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754-1760.
41. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES and Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology. 2013; 31:213-219.
42. Ramos AH, Lichtenstein L, Gupta M, Lawrence MS, Pugh TJ, Saksena G, Meyerson M and Getz G. Oncotator: cancer variant annotation tool. Human mutation. 2015; 36:E2423-2429.
43. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM and Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic acids research. 2001; 29:308-311.
44. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA and Stratton MR. The Catalogue of Somatic Mutations in Cancer (COSMIC). Current protocols in human genetics / editorial board, Jonathan L Haines [et al]. 2008; Chapter 10:Unit 10 11.
45. Liu A, Yu X and Liu S. Pluripotency transcription factors and cancer stem cells: small genes make a big difference. Chinese journal of cancer. 2013; 32:483-487.
46. Mao Y, Chen H, Liang H, Meric-Bernstam F, Mills GB and Chen K. CanDrA: cancer-specific driver missense mutation annotation with optimized features. PloS one. 2013; 8:e77945.
47. Choi Y and Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015; 31:2745-2747.
48. Thorvaldsdottir H, Robinson JT and Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013; 14:178-192.
49. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G and Mesirov JP. Integrative genomics viewer. Nature biotechnology. 2011; 29:24-26.
50. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O and Barillot E. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012; 28:423-425.
51. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL and Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012; 7:562-578.
52. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R and Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013; 14:R36.
53. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, White TA, Stojanov P, Van Allen E, Stransky N, Nickerson E, Chae SS, Boysen G, Auclair D, Onofrio RC, Park K, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nature genetics. 2012; 44:685-689.
54. Livak KJ and Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 2001; 25:402-408.
55. Gioanni J, Fischel JL, Lambert JC, Demard F, Mazeau C, Zanghellini E, Ettore F, Formento P, Chauvel P, Lalanne CM and et al. Two new human tumor cell lines derived from squamous cell carcinomas of the tongue: establishment, characterization and response to cytotoxic treatment. European journal of cancer & clinical oncology. 1988; 24:1445-1455.
56. Chang SE, Foster S, Betts D and Marnock WE. DOK, a cell line established from human dysplastic oral mucosa, shows a partially transformed non-malignant phenotype. Int J Cancer. 1992; 52:896-902.
57. Jin G, Zhang F, Chan KM, Xavier Wong HL, Liu B, Cheah KS, Liu X, Mauch C, Liu D and Zhou Z. MT1-MMP cleaves Dll1 to negatively regulate Notch signalling to maintain normal B-cell development. The EMBO journal. 2011; 30:2281-2293.

58. Castel D, Mourikis P, Bartels SJ, Brinkman AB, Tajbakhsh S and Stunnenberg HG. Dynamic binding of RBPJ is determined by Notch signaling status. Genes Dev. 2013; 27:1059-1071.

59. Walker JM. The bicinchoninic acid (BCA) assay for protein quantitation. Methods Mol Biol. 1994; 32:5-8.

60. Dinse GE and Lagakos SW. Nonparametric estimation of lifetime and disease onset distributions from incomplete observations. Biometrics. 1982; 38:921-932.

Database, 2016, 1–8
doi: 10.1093/database/baw104
Original article

Original article

TMC-SNPdb: an Indian germline variant database derived from whole exome sequences

Pawan Upadhyay,[1] Nilesh Gardi,[1,†] Sanket Desai,[1,†] Bikram Sahoo,[1] Ankita Singh,[1] Trupti Togar,[1] Prajish Iyer,[1] Ratnam Prasad,[1] Pratik Chandrani,[1] Sudeep Gupta[2] and Amit Dutt[1,*]

[1]Integrated Genomics Laboratory, Advanced Centre for Treatment Research Education in Cancer (ACTREC), [2]Department of Medical Oncology, Tata Memorial Centre, Mumbai, Maharashtra 410012, India

***Corresponding author:** Tel: +91-22-27405056; Email: adutt@actrec.gov.in

[†]These authors contributed equally.

Citation details: Upadhyay, P., Gardi, N., Desai, S. *et al.* TMC-SNPdb: an indian germline variant dataset derived from whole exome sequence. *Database* (2016) Vol. 2016: article ID baw103; doi:10.1093/database/baw104

Received 14 April 2016; Revised 7 June 2016; Accepted 8 June 2016

Abstract

Cancer is predominantly a somatic disease. A mutant allele present in a cancer cell genome is considered somatic when it's absent in the paired normal genome along with public SNP databases. The current build of dbSNP, the most comprehensive public SNP database, however inadequately represents several non-European Caucasian populations, posing a limitation in cancer genomic analyses of data from these populations. We present the **T**ata **M**emorial **C**entre-SNP **d**ata**b**ase (TMC-SNPdb), as the first open source, flexible, upgradable, and freely available SNP database (accessible through dbSNP build 149 and ANNOVAR)—representing 114 309 unique germline variants—generated from whole exome data of 62 normal samples derived from cancer patients of Indian origin. The TMC-SNPdb is presented with a companion subtraction tool that can be executed with command line option or using an easy-to-use graphical user interface with the ability to deplete additional Indian population specific SNPs over and above dbSNP and 1000 Genomes databases. Using an institutional generated whole exome data set of 132 samples of Indian origin, we demonstrate that TMC-SNPdb could deplete 42, 33 and 28% false positive somatic events post dbSNP depletion in Indian origin tongue, gallbladder, and cervical cancer samples, respectively. Beyond cancer somatic analyses, we anticipate utility of the TMC-SNPdb in several Mendelian germline diseases. In addition to dbSNP build 149 and ANNOVAR, the TMC-SNPdb along with the subtraction tool is available for download in the public domain at the following:

Database URL: http://www.actrec.gov.in/pi-webpages/AmitDutt/TMCSNP/TMCSNPdp.html

Introduction

Somatic mutations sequentially accumulate in cancer cell genomes. In addition, a typical cancer genome contains several polymorphic 'normal' germline variants (1–3). Subtracting the tumor DNA variants against matched normal DNA derived from the same individual and those polymorphic in the population is, therefore, essential to identify an exclusive somatic event (4). Apropos, a critical aspect of any tumor genome sequence analysis involves depletion of paired normal variants followed by depletion of residual variants from public databases of common single nucleotide polymorphism (SNP) such as dbSNP (5) and 1000 Genomes Project (6). A sequence variant not observed in matched normal derived genome sequence and absent from public SNP database is considered somatic in origin. Adopting such an analytical approach ensures filtering of paired-germline and population-specific polymorphic variants from dbSNP and 1000 Genomes Project for Caucasian population (7).

However, despite depletion against dbSNP, unknown SNPs especially those with lower minor allele frequency not represented in dbSNP, are likely to confound somatic mutation analyses in studies involving non- Caucasian and non-European Caucasian populations (5). Two exhaustive initiatives addressing this issue are the publicly available exome variation datasets: NHLBI Exome Sequencing Project (https://esp.gs.washington.edu/EVS/) and Exome Aggregation Consortium (ExAC) (http://exac.broadinstitute.org/) (8). Information gathered from these studies is an integral part of variant annotation tools like Annovar (9).

Multiple studies such as the Indian Genome Variation Consortium (10, 11) and HUGO Pan- Asian SNP Consortium (12) have described the genomic distinctiveness of Indian population based on varying allele frequency of known SNPs, complex origin, genetic diversity (13–16), and high variation of male lineages (Y-chromosome) within the population (17, 18). However, a concerted effort to comprehensively identify and catalogue novel SNPs present exclusively in Indian population is yet to be undertaken. Lack of Indian specific SNP database has been an important impediment in cancer research, especially in efforts to discover bona fide novel somatic mutations.

Here, we describe **T**ata **M**emorial **C**entre-SNP **d**ata**b**ase 'TMC-SNPdb' as the first, open source, freely available database of unique germline variants obtained from whole exome data of 62 'normal' samples from tongue, gallbladder, and cervical cancer patients of Indian origin. 'TMC-SNPdb' is presented with an easy-to-use graphic user interface feature to enable researchers to call true somatic mutations by depleting against Indian population specific SNPs, in addition to those already catalogued in dbSNP and 1000 Genomes databases. We demonstrate that 'TMC-SNPdb' effectively filters false positive somatic events across 75 tumor whole exome data.

Materials and methods

Ethical approval and informed consent

The sample set and study protocol was approved by Institutional Review Board (project no. 116 for cervical adenocarcinoma samples; project no. 88 for head and neck cancer samples, project 104 for gallbladder cancer samples). Cervical squamous carcinoma whole exome data have been described earlier in (19). Written informed consent was obtained from all patients.

Extraction of DNA

All 'normal' tissue samples under study were verified by an onco-pathologist to not harbor any cancer. A total of 62 samples 'normal' samples (16 peripheral venous blood and 46 adjacent normal tissue) were obtained for analysis: peripheral venous blood from patients with cervical squamous cell carcinoma ($n = 10$), cervical adenocarcinoma ($n = 18$) (adjacent normal tissue; $n = 12$ and peripheral venous blood; $n = 6$) and adjacent normal tissue from patient with tongue squamous cell carcinoma ($n = 23$) and gallbladder ($n = 11$) were obtained from Tata Memorial Hospital (TMH). Genomic DNA from tissues was extracted using DNeasy blood and tissue DNA extraction kit (Qiagen) according to manufacturer's instructions. Quantification of DNA was assessed using Nanodrop 2000c Spectrophotometer (Thermo Fisher Scientific Inc.) and DNA integrity was determined by resolving on 0.8% Agarose gel. DNA was also quantified using Qubit ds DNA BR assay kit (Life Technologies, USA). DNA samples showing DNA concentration >50 ng/μl and intact DNA visualized on agarose gel were used for whole exome sequencing.

Exome capture, library preparation and sequencing

Three different library preparation kits were used to prepare libraries for different tumor types (Supplementary Table S1). First, TruSeq Exome Enrichment kit (v2 and v3, Illumina) was used to capture 62 Mb region (>3 40 000 probes) of human genome comprising 201 121 exons representing 20 974 gene sequences, including 5′UTR, 3′UTR, microRNAs and other non-coding RNA. For exome library preparation, two microgram genomic DNA was sheared using Covaris (Covaris Inc) for generating fragment sizes of 200–300 bp. Fragments end repairing,

purification, A-tailing, adaptor ligation and quality control steps were carried out using TruSeq DNA Sample Prep Kit (Illumina) following manufacturer's instructions. Qualitative and quantitative analysis of genomic DNA libraries were performed using High Sensitivity DNA chip on 2100 Bioanalyzer (Agilent) and qPCR with KAPA Library Quant Kit (Kapa Biosystems). Exome enrichment was done by incubation at 93 °C for 1 min (decreasing 2 °C per cycle for 18 cycles) followed by 58 °C for 19 h in ABI 9700 PCR system (Life Technologies) using 500 ng of genomic libraries.

Second, NimbleGen SeqCap EZ Exome Library (v3.0, Roche) targeting 64 Mb of the human genome was also used for library preparation. The protocol was adopted from the manufacture's application note (http://www.nimblegen.com/products/lit/NimbleGen_SeqCap_EZ_SR_Pre-Captured_Multiplexing.pdf). Sequencing libraries were exome captured and later quality-controlled using a bioanalyzer (Agilent 2100) and libraries were qPCR quantified using KAPA Library Quant Kit (Kapa Biosystems) prior to cluster generation on an Illumina cBOT.

Third, SureSelect Human All Exon Kit, v5 (Agilent Technologies, Santa Clara, CA, USA) was also used to capture 50 Mb of the human genome using > 5 50 000 probes. One microgram of genomic DNA was utilized for library preparation and a similar protocol was followed as previously stated. Eluted exome-enriched library fragments were PCR amplified and purified.

qPCR quantified 7 pmol of 6-plex DNA library pool was loaded per lane on flow cell (Flow Cell v3) to generate clusters using TruSeq PE Cluster Kit v3-cBot-HS kit and clustered flow was sequenced for 201 and 301 cycles on HiSeq-1500 and NextSeq System (Illumina) using TruSeq SBS Kit v3 (Illumina), respectively.

Exome sequencing variant analysis for TMC-SNP database

Paired-end raw sequence reads were mapped to human reference genome (build hg19) using BWA v. 0.6.2 (20). Quality control analysis of bam files was carried out using qualimap (v0.7.1) (21). Mapped reads were then used to identify and remove PCR duplicates using Picard tools v.1.74 (http:broadinstitute.github.io/picard/). Base quality score recalibration and indel re-alignment were performed and variants were called from each sample separately using GATK Unified Genotyper (version 2.5-2) (22).

Development of TMC-SNP database

To restrict our analysis to high quality germline variants we applied filters of minimal base coverage and recurrence in cohort. In house developed scripts (Awk and Perl) were used to merge all 62 VCF files from normal tissues and mutational recurrence was calculated. We applied a standard filter of coverage ≥ 5 reads for altered alleles. Additionally, we included variants with coverage ≤ 5 but recurrent in ≥ 4 normal samples. Using these filters, we identified high quality variants in the dataset. High quality variants were further annotated using COSMICdb (version 68) (23) and dbSNP (version 142) (5). Remaining variants were further depleted against dbSNP and COSMICdb to remove all known somatic and germline variants. Finally, all remaining variants constitute the TMC-SNP database. A detailed schema of resource and data representation is provided in Supplementary Figure 3.

Application of TMC-SNP database in analyzing tumor samples

GATK (version 2.5-2) and MuTect (version 1.0.2) (24) were utilized to generate raw variants of tumor samples and filtered against its matched normal . Variants obtained from GATK and MuTect were merged and variants having ≥ 5 reads for altered allele were kept for further downstream analysis. Similar analysis was carried out for three cancer types. Comparison with dbSNP(version142) and COSMICdb(version 68) was performed using in-house developed scripts in Perl and Awk which were later used to calculate the percentage changes in variants in different cancer type post filtration with dbSNP and TMC-SNP database. Functional annotation of variants was performed using Oncotator (variant annotation tool) (25).

Germline variant subtraction program

TMC-SNPdb is distributed as a SQLite file containing variant information table. A companion tool for subtraction of germline variants from tumor sample has been developed in python (version 3.4). It depends on PyVCF (version ≥ 1.6) and sqlite3 python packages. The variants in TMC-SNPdb are characterized by a unique combination of chromosome number, genomic position, reference allele, altered allele for each variant and subtraction was carried out based on these unique fields for each variant in VCF file. The tool is an executable compatible with Linux operating system and has been tested on several Linux platform such Red Hat (version 6.5), Fedora (version 22) and Ubuntu (version 14.04). It can be executed using a command line interface ('TMC-SNP') or a graphical user interface (GUI) ('TMC-SNP_GUI'). The GUI mode additionally depends on TKinter python library (version ≥ 2.4). Moreover, the tool has a feature which lets users create their own germline variant database from VCF format files

of normal samples. The output obtained from the tool is in VCF format. Detailed user manual with snapshots of the GUI and schematic representation of overall usages are provided in Supplementary file 1 and Supplementary Figure S2.

Availability of supporting data

The raw sequence data has been deposited at the ArrayExpress (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-4618), hosted by the European Bioinformatics Institute (EBI). The 'TMC-SNPdb' has been submitted to Annovar (http://annovar.openbioinformatics.org/en/latest/user-guide/download/) and dbSNP (http://www.ncbi.nlm.nih.gov/SNP/snp_viewTable.cgi?handle=TMC_SNPDB) for public access.

Results

Development of TMC-SNP database

We analyzed whole exome sequencing at a median of 88x coverage for 62 normal samples derived from cancer patients, comparable with similar reports (26) as detailed in Supplementary Table S1. Of note, coverage among 4 of 62 samples were <30× due to high duplication reads and low yield in these samples. Germline mutations were called using GATK (22): a total of 15 015 608 germline variants were identified across the complete dataset. As shown in Figure 1, standard quality filters of minimal 5× coverage or recurrence in at least four samples for each variant led to about 90% reduction in raw variants (see Materials and Methods section for details). The remaining 1 422 336 variants of higher confidence were further depleted against dbSNP v142. 1 305 937 of 1 422 336 variants, constituting 92% SNPs were depleted. To remove variants known to be somatically associated with cancer in literature but figured as a germline event in our study (most likely due to inadequate or non-uniform coverage of their paired normal samples), we further depleted 2090 variants (2%) overlapping with COSMICdb with an assumption of these variants to be false somatic events in our data set. Finally, a total of 114 309 variants were identified after filtering with dbSNP and COSMICdb as a pool of previously unknown germline variants of high confidence recurring in the Indian population to constitute the 'TMC-SNPdb'.

Characteristic features of TMC-SNP database

A total 114 309 variants were annotated using Oncotator for functional features (25). A distribution pattern of

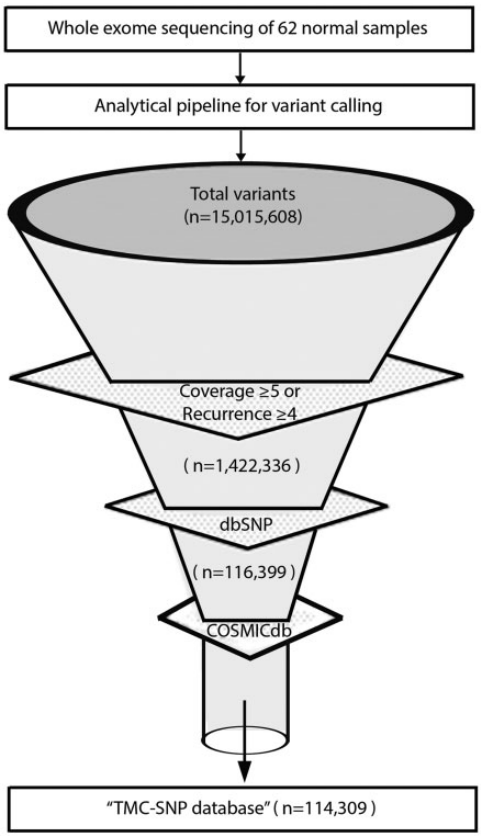


Figure 1. Development of TMC-SNPdb using whole exome sequencing. Schematic flow representation of steps followed during development of TMC-SNP database. The whole exome sequencing of 62 normal tissue obtained from three different tissues of cancer patients was performed and analysed using GATK (Genome Analysis Tool Kit) to generate VCF files. Raw variants obtained were further filtered using mentioned criteria to find a list of variants absent in dbSNP v142 and COSMICdb v68. Remaining variants constitutes the 'TMC-SNPdb' shown at the end of the funnel.

coding (~17 973) and non-coding variants germline variants (~96 336) is shown in Figure 2A. Of 17 973 coding variants, 11 466 were of non-synonymous (NS) (~63%) and 6507 were synonymous variants (S) (~36%) with NS/S ratio 1.76, consistent with previous reports for exome data from normal samples (27, 28). Furthermore, we observed a high frequency of missense (~58%) and silent variants (~30%) as compared with indel (~3%), nonsense (~2%) and splice site (~6%) region (Supplementary Figure 1A). Of all the SNPs present in TMC-SNPdb, distribution varied across the genome as follows: protein-coding exon (15.7%), intron (40%), IGR (25.8%), 3′UTR (9.5%), 5′UTR (2.37%), RNA (3.74%) and lincRNA (1.7%), consistent with earlier report from exome sequencing data (Supplementary Figure S1B) (29, 30). Next, we computed the allele frequency of all 114 309 variants present in the TMC-SNPdb, across 62 samples. Given that TMC-SNPdb predominantly enlists low frequency germline

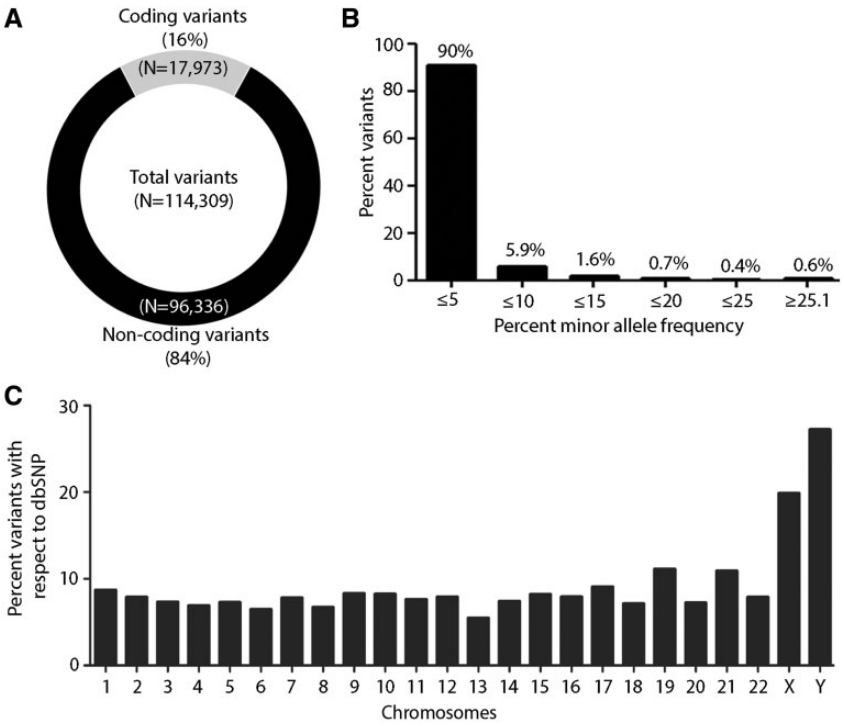


Figure 2. Overall overview of characteristic features of TMC-SNP database. **(A)** Circle plot of coding and non-coding variants obtained in the dataset. **(B)** Percent minor allele frequency distribution of variants in 'TMC-SNPdb' across 62 normal samples. Percentage frequencies are presented on the top of each bar. **(C)** Genome-wide distribution of percent frequency of variants obtained in each chromosome as compared with dbSNP database.

variants prevalent among Indian population, similar to 1000 genomes and ExAC wherein about 99% of SNPs are estimated to have a minor allele frequency over 1% (8, 26), Similarly, in TMC-SNPdb >90% of variants present exist at a minor allele frequency≤5% (Figure 2B).

Furthermore, a comparative measure of variability added by the TMC-SNPdb variants to the known pool of SNPs per chromosome was reckoned following comparison with dbSNP variants across the genome. Interestingly, we found maximal variability at the Y-chromosome among 2418 of 8885 SNPs (27%), while the distribution of the variants across the autosomal chromosomes was found to be uniformly distributed among 106 184 of 1 346 256 SNPs (7.6%) similar to the dbSNP (Figure 2C). Of note, variants at Y-chromosome tend to be more localized geographically than those of mitochondrial DNA (mtDNA) and autosomes, which is reflective of the degree of inter-population genetic differences (31–33). Y-chromosomes have been shown to harbor population specific unique haplotype in Indian population and have frequently been used as a marker for studying human demographic history (34, 35). The higher variability at the Y-chromosome found in TMC-SNPdb is thus consistent with several earlier reports describing a high variation of male lineages within Indian population (17, 18) that further emphasizes the Indian specific characteristics of the TMC-SNPdb germline variants, and a need for distinct Indian specific germline database .

Finally, a significant characteristic feature of TMC-SNPdb is the companion subtraction tool with command line and GUI based interface. The user can deplete their data set against TMC-SNPdb or create a customized normal variant database. The program has been tested to run on various Linux platforms such Fedora, Ubuntu and Red Hat operating systems. (Detailed user manual and snapshot of different steps have been provided in Supplementary Materials S1, S2 and Supplementary Figure S2). Using the companion tool on an 8GB machine, it takes 56 and 72 min to filter standard VCFs containing 115 884 and 227 779 raw variants (provided as example file with tool) against the TMC-SNPdb variants, respectively.

Application of 'TMC-SNPdb' in depleting germline variants predominant among indian population

With the flexibility of using GUI interface or through the command line (refer to Supplementary Material S1), we tested the robustness and practical utility of 'TMC-SNPdb' across various cancer types to infer the extent of depletion of population specific variants over and above the dbSNP. We analyzed 132 samples of three cancer types: head and neck cancer ($n = 43$), cervical cancer ($n = 62$) and gallbladder cancer ($n = 27$). Significant fold reduction of variants was observed following TMC-SNPdb subtraction in addition to depletion by dbSNP in all cancer types studied. Of

613 055 variants found across 24 head and neck cancer tumor samples about 92% SNPs were depleted post dbSNP subtraction with 84 001 candidate somatic variants. Subsequent depletion using TMC-SNPdb identified 35 819 additional variants as Indian specific germline variants existing at varying frequency in normal Indian population. In overall, TMC-SNPdb allowed us to filter an additional 42.6% of post dbSNPs depleted SNPs. in 24 tongue cancer samples (Table 1). Similarly, TMC-SNPdb significantly reduced about 33.3% and 27.7% SNPs in 17 gallbladder and 34 cervical tumor whole exome data, as tabulated in detail in Table 1.

Discussion

TMC-SNPdb is a freely available open access Indian population specific germline variant database consisting of 114 309 germline variants using whole exome sequencing of 62 normal tissues from patients with different types of cancer. Its usage is analogous to depletion against pooled normal variants from unrelated normal samples of Indian origin for paired or orphan tumor samples. The utility of subtraction against pooled normal variants has been well described as a reference for depletion, especially for orphan tumor samples wherein paired normal variant data for the tumor samples are not available (36–39). Our dataset and companion tool can be used, along with other public databases, as 'normal' counterpart to identify disease specific somatic mutations, especially in cancer exome studies. Using TMC-SNPdb across 132 whole exome data of 3 tumor types, we show that it can significantly deplete false positive somatic variants.

TMC-SNPdb is presented with a companion program with command line or user-friendly GUI interface for non-computational biologists. It has two built-in features: first, a user can input tumor VCF to subtract against TMC-SNPdb and second, create a custom database of germ line mutation with the availability of multiple normal VCF files and then subtract with tumor VCF to deplete germ line variants. The subtraction program has been tested on several Linux platforms such as Fedora, Ubuntu and Red Hat system. Because it is an open source tool, it could be further modified to alter filtering parameters for analysis indicative of its expandability and universal applicability on Linux platforms.

There are two major limitations of TMC-SNPdb database. First, it is presumed that a sample derived from cancer patients represents 'normal' genome variation. However, because of their diseased status, a fraction of such individuals are likely to harbor cancer predisposing variants in their germline. Any such germline variant that is novel in Indian population (not yet included in Caucasian databases) and which predisposes to cancer (e.g. in *BRCA 1* gene) would be characterized as 'normal' population variation in TMC-SNPdb. Thus, this database will be limited in application to analyses that seek to evaluate germline predisposition to cancer. Second, majority of 'normal' samples were obtained from sites adjacent to a tumor with histopathological based inspection for the absence of tumor cells. However, it is possible that these tissues harbors some bona fide somatic mutations due to effect of field cancerization (40, 41). Thus, depleting against TMC-SNPdb could potentially 'over-subtract' mutations that are bona fide somatic. To minimize this possibility, we have filtered TMC-SNPdb variants against COSMIC database to remove any known cancer related somatic variants. However, there remains a residual potential for missing 'somatic' mutations that are novel in tumors of Indian patients and present in adjacent 'normal' tissue. With these caveats, we believe that TMC-SNPdb with its companion tool is a step towards fulfilling a significant unmet need for an Indian population 'normal' variant database, especially in somatic mutation analyses in tumors from Indian patients.

In summary, TMC-SNPdb is an open source database of 'normal' germline variants derived from Indian—non-European Caucasian—population, not yet included in the public databases with predominant Caucasian representations. It comes along with a companion tool that can apply this information for somatic cancer genome analyses by

Table 1 Application of TMC-SNPdb across cancer types to filter germline variants in Indian population

S.No.	Cancer type	Total variants	Number of samples	Number and percentage (along row) of novel variants		Overall reduction by TMC-SNPdb post dbSNP depletion
				Post dbSNP depletion	Post TMC-SNPdb depletion	
1	Tongue cancer	613 055	24	84 001 (13.7%)	48 182 (7.8%)	42.6%
2	Cervical cancer	923 547	34	99 032 (10.7%)	71 594 (7.7%)	27.7%
3	Gall-bladder	328 245	17	26 530 (8%)	17 682 (5.3%)	33.3%

Total number of variants observed for each cancer types and reduction in number and percent variants post dbSNP and post TMC-SNPdb subtraction is tabulated for three cancer types. Number of samples analysed across tumor is also denoted.

depleting against the TMC-SNPdb. This database is flexible to accommodate the need for customization by allowing inclusion of similar datasets from additional individuals.

Supplementary data

Supplementary data are available at *Database* Online.

Acknowledgements

Sandor Lifesciences Pvt. Ltd., Genotypic Pvt. Ltd, and Scigenome Labs, as commercial vendors for generating the whole exome library preparation and sequencing services. Rita Mulherkar for cervical squamous sample exome data.

Funding

A.D. is supported by an Intermediate Fellowship from the Wellcome Trust/DBT India Alliance (IA/I/11/2500278), by a grant from DBT (BT/PR2372/AGR/36/696/2011), and intramural grants (IRB project 92 and 55). P.U is supported by Senior Research Fellowship from CSIR, N.G is supported by Junior Research Fellowship from TMH, P.C and P.I. are supported by Senior Research Fellowship from ACTREC and T.T. and S.D. is supported by Junior Research Fellowship from DBT and ACTREC, respectively. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the article.

Conflict of interest. None declared.

References

1. Martincorena,I. and Campbell,P.J. (2015) Somatic mutation in cancer and normal cells. *Science*, 349, 1483–1489.
2. Stratton,M.R., Campbell,P.J., and Futreal,P.A. (2009) The cancer genome. *Nature*, 458, 719–724.
3. Vogelstein,B., Papadopoulos,N., Velculescu,V.E. *et al.* (2013) Cancer genome landscapes. *Science*, 339, 1546–1558.
4. Jones,S., Anagnostou,V., Lytle,K. *et al.* (2015) Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci. Transl. Med.*, 7, 283ra253.
5. Sherry,S.T., Ward,M.H., Kholodov,M. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, 29, 308–311.
6. Genomes Project,C., Abecasis,G.R., Auton,A. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491, 56–65.
7. Wang,L. and Wheeler,D.A. (2014) Genomic sequencing for cancer diagnosis and therapy. *Annu. Rev. Med.*, 65, 33–48.
8. Lek,M., Karczewski,K., Minikel,E., *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv*.
9. Yang,H. and Wang,K. (2015) Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.*, 10, 1556–1566.
10. Narang,A., Roy,R.D., Chaurasia,A. *et al.* (2010) IGVBrowser–a genomic variation resource from diverse Indian populations. *Database (Oxford)*, 2010, baq022.
11. Indian Genome Variation,C. (2005) The Indian Genome Variation database (IGVdb): a project overview. *Hum. Genet.*, 118, 1–11.
12. Consortium,H.P.A.S., Abdulla,M.A., Ahmed,I. *et al.* (2009) Mapping human genetic diversity in Asia. *Science*, 326, 1541–1545.
13. Tamang,R., Singh,L., and Thangaraj,K. (2012) Complex genetic origin of Indian populations and its implications. *J. Biosci.*, 37, 911–919.
14. Tamang,R. and Thangaraj,K. (2012) Genomic view on the peopling of India. *Investig. Genet.*, 3, 20.
15. Reich,D., Thangaraj,K., Patterson,N. *et al.* (2009) Reconstructing Indian population history. *Nature*, 461, 489–494.
16. Majumder,P.P. and Basu,A. (2015) A genomic view of the peopling and population structure of India. *Cold Spring Harb Perspect Biol*, 7, a008540.
17. Basu,A., Mukherjee,N., Roy,S. *et al.* (2003) Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Res.*, 13, 2277–2290.
18. Sengupta,S., Zhivotovsky,L.A., King,R. *et al.* (2006) Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *Am. J. Hum. Genet.*, 78, 202–221.
19. Chandrani,P., Kulkarni,V., Iyer,P. *et al.* (2015) NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome. *Br. J. Cancer*, 112, 1958–1965.
20. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25, 1754–1760.
21. Garcia-Alcalde,F., Okonechnikov,K., Carbonell,J. *et al.* (2012) Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics*, 28, 2678–2679.
22. McKenna,A., Hanna,M., Banks,E. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20, 1297–1303.
23. Bamford,S., Dawson,E., Forbes,S. *et al.* (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer*, 91, 355–358.
24. Cibulskis,K., Lawrence,M.S., Carter,S.L. *et al.* (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, 31, 213–219.
25. Ramos,A.H., Lichtenstein,L., Gupta,M. *et al.* (2015) Oncotator: cancer variant annotation tool. *Hum. Mutat.*, 36, E2423. (2429)
26. Genomes Project,C., Auton,A., Brooks,L.D. *et al.* (2015) A global reference for human genetic variation. *Nature*, 526, 68–74.
27. Xu,B., Roos,J.L., Dexheimer,P. *et al.* (2011) Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat. Genet.*, 43, 864–868.
28. Genomes Project,C., Abecasis,G.R., Altshuler,D. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, 467, 1061–1073.
29. Guo,Y., Long,J., He,J. *et al.* (2012) Exome sequencing generates high quality data in non-target regions. *BMC Genomics*, 13, 194.
30. Samuels,D.C., Han,L., Li,J. *et al.* (2013) Finding the lost treasures in exome sequencing data. *Trends Genet.*, 29, 593–599.
31. Perez-Lezaun,A., Calafell,F., Comas,D. *et al.* (1999) Sex-specific migration patterns in Central Asian populations, revealed by

analysis of Y-chromosome short tandem repeats and mtDNA. *Am. J. Hum. Genet.*, 65, 208–219.
32. Oota,H., Kitano,T., Jin,F. *et al.* (2002) Extreme mtDNA homogeneity in continental Asian populations. *Am. J. Phys. Anthropol.*, 118, 146–153.
33. Kumar,V., Langstieh,B.T., Madhavi,K.V. *et al.* (2006) Global patterns in human mitochondrial DNA and Y-chromosome variation caused by spatial instability of the local cultural processes. *PLoS Genet.*, 2, e53.
34. Jobling,M.A. and Tyler-Smith,C. (2003) The human Y chromosome: an evolutionary marker comes of age. *Nat. Rev. Genet.*, 4, 598–612.
35. Shrivastava,P., Gupta,U., Jain,T. *et al.* (2015) Y STR haplotype diversity in central Indian population. *Ann. Hum. Biol.*, 1–8.
36. Kumar,A., White,T.A., MacKenzie,A.P. *et al.* (2011) Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. *Proc. Natl. Acad. Sci. U S A*, 108, 17087–17092.
37. Suzuki,A., Mimaki,S., Yamane,Y. *et al.* (2013) Identification and characterization of cancer mutations in Japanese lung adenocarcinoma without sequencing of normal tissue counterparts. *PLoS One*, 8, e73484.
38. Raymond,V.M., Gray,S.W., Roychowdhury,S. *et al.* (2016) Germline findings in tumor-only sequencing: points to consider for clinicians and laboratories. *J. Natl. Cancer Inst.*, 108.
39. McCarthy,M. (2015) Genomic sequencing of only tumor tissue could be misleading in nearly half of patients, study shows. *Bmj*, 350, h2036.
40. Dakubo,G.D., Jakupciak,J.P., Birch-Machin,M.A. *et al.* (2007) Clinical implications and utility of field cancerization. *Cancer Cell Int.*, 7, 2.
41. Mohan,M. and Jagannathan,N. (2014) Oral field cancerization: an update on current concepts. *Oncol. Rev.*, 8, 244.