Discovery of Potential Therapeutic Targets in Human Cancer: A Functional Genomics Approach

By

Pratik Chandrani

[LIFE09201104003]

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

February, 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. Pratik Chandrani entitled "Discovery of Potential Therapeutic Targets in Human Cancer: A Functional Genomics Approach" and recommend that it may be accepted as fulfilling the thesis requirement for the award of Degree of Doctor of Philosophy.

S Chiplunkar 7/2/17

Chairman – Dr. S. Chiplunkar Date:

A Dutt 7/2/17

Guide/Convener – Dr. Amit Dutt Date:

V. Prasanna. 7/2/17.

Member 1 – Dr. V. Prasanna Date:

Mahimkar 7/2/2017

Member-2 – Dr. M. Mahimkar Date:

13/2/2017

Invitee – Dr. Santosh Noronha Date:

F Habib 9/2/2017

Technology Advisor – Dr. Farhat Habib Date:

7/2/17

External Examiner – Dr. Shantanu Chowdhury Date:

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

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

A Dutt 7/2/17

Date: 7/2/17
Place: Navi Mumbai

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 Pratik Chandrani

Date:

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 Pratik Chandrani

21th July, 2016

LIST OF PUBLICATIONS ARISING FROM THE THESIS

Peer-reviewed papers from this thesis work:

1. **Chandrani P***, Prabhash K.*, Choughule A, Prasad R, Sethunath V, Ranjan M, Aich J, Dhamne H, Iyer D, Upadhyay P, Sundaram P, Mohanty B, Chandna P, Kumar R, Joshi A, Noronha V, Patil V, Ramaswamy A, Karpe A, Thorat R, Chaudhary P, Ingle A, Dutt A. Drug-sensitive FGFR3 mutations in lung adenocarcinoma. *Annals of Oncology*. 2016 Dec 19. pii: mdw636. doi: 10.1093/annonc/mdw636. PubMed PMID: 27998968 [Epub ahead of print].
2. **Chandrani P***, Upadhyay P*, Iyer P, Tanna M, Shetty M, Raghuram GV, 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*. 2015 Nov 14;16(1):936. doi: 10.1186/s12864-015-2138-4. PubMed PMID: 26572163; PubMed Central PMCID: PMC4647579.
3. **Chandrani P***, Kulkarni V*, Iyer P, Upadhyay P, Chaubal R, Das P, Mulherkar R, Singh R, Dutt A. NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome. *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.

Conference abstracts:

1. **P Chandrani**, V Sethunath, A Chougule, J Aich, H Dhamne, P Upadhyay, D N Iyer, A Thavamani, R Sindhi, P Chandna, K Prabhash and A Dutt. Discovery of Actionable Alterations in Lung Adenocarcinoma poster presentation at "Conference - Decoding the Genetics of Common Cancers in India", 19-21 February, 2016 in Pune, India. **(recipient of first prize)**

2. **P Chandrani**, V Sethunath, A Chougule, J Aich, H Dhamne, P Upadhyay, D N Iyer, A Thavamani, R Sindhi, P Chandna, K Prabhash and A Dutt. Discovery of Actionable Alterations in Lung Adenocarcinoma poster presentation at "A Conference of New Ideas in Cancer – Challenging Dogmas", 26-28 February, 2016 in Mumbai, India. **(recipient of third prize)**
3. **P Chandrani**, Pawan Upadhyay, Prajish Iyer, Mayur Tanna, Madhur Shetty, Raghuram Venkata Gorantala, Ninad Oak, Ankita Singh, Rohan Chaubal, Manoj Ramteke, Sudeep Gupta, Amit Dutt. Integrated genomics approach to identify driver alterations. [abstract]. In: *Proceedings of the AACR-NCI-EORTC International Conference: Molecular Targets and Cancer Therapeutics*; 2015 Nov 5-9; Boston, MA. Philadelphia (PA): AACR; Mol Cancer Ther 2015;14(12 Suppl 2): Abstract nr C154.
4. **P Chandrani**, V Sethunath, A Chougule, J Aich, H Dhamne, P Upadhyay, D N Iyer, A Thavamani, R Sindhi, P Chandna, K Prabhash and A Dutt. Discovery of Actionable Alterations in Lung Adenocarcinoma oral presentation at "34th Annual Convention of Indian Association for Cancer Research", 19-21 February, 2015 in Jaipur, India. **(selected in top abstracts for oral presentation under IACR Sitaram Joglekar award and Mangala Bamne award category)**
5. **P Chandrani**, P Iyer, V Kulkarni, P Upadhyay, R Chaubal, P Das, R Mulherkar, R Singh, A Dutt. HPVDetector: A Tool to Detect HPV and Their Integration Sites Using Next Generation Sequencing Data poster presentation by P Iyer at "NextGen Genomics & Bioinformatics Technologies (NGBT) Conference", 17-19 November, 2014 in Bangalore, India. **(recipient of first prize)**
6. **P Chandrani**, Aich J, Upadhyay P, Chougule A, Jose T, Chandna P, Prabhash K, Dutt A. Profiling and Discovery of Actionable Alterations in Lung Adenocarcinoma

poster presentation at "Worldwide Innovative Networking (WIN-2013)", 10-12 July, 2013 in Paris, France.

7. **P Chandrani**, Iyer D, Aich J, Upadhyay P, Chougule A, Jose T, Chandna P, Prabhash K, Dutt A. Profiling and Discovery of Actionable Alterations in Lung Adenocarcinoma poster presentation at "Integrating Basic and Translational Research in Modern Biology", 27-28 December, 2013 in Vadodara, India.
8. **P Chandrani**, Prasad R, Upadhyay P, Thavamani A, Trivedi V, Sharma R, Chougule A, Jambhekar N, Noronha V, Jose T, Prabhash K, Dutt A. Mutational Profiling of Actionable Alterations in Lung Adenocarcinoma poster presentation at "2nd Global Cancer Genomics Consortium Symposium: Genomics Medicine in Cancer Research", 19-20 November, 2012 in Mumbai, India. **(recipient of best poster presentation award)**

Other journal papers and book chapter:

1. 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 (Oxford)*. 2016 Jul 9;2016. pii: baw104. doi: 10.1093/database/baw104. Print 2016. PubMed PMID: 27402678.
2. 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. Notch pathway activation is essential for maintenance of stem-like cells in early tongue cancer. *Oncotarget*. 2016 Jul 6. doi: 10.18632/oncotarget.10419. [Epub ahead of print] PubMed PMID: 27391340.
3. Iyer P, Barreto SG, Sahoo B, **Chandrani P**, Ramadwar MR, Shrikhande SV, Dutt A. Non-typhoidal Salmonella DNA traces in gallbladder cancer. *Infect Agent*

Cancer. 2016 Mar 3;11:12. doi: 10.1186/s13027-016-0057-x. eCollection 2016. PubMed PMID: 26941832; PubMed Central PMCID: PMC4776363.

4. Choughule A*, Sharma R*, Trivedi V*, Thavamani A, Noronha V, Joshi A, Desai S, **Chandrani P**, Sundaram P, Utture S, Jambhekar N, Gupta S, Aich J, Prabhash K, Dutt A. Coexistence of KRAS mutation with mutant but not wild-type EGFR predicts response to tyrosine-kinase inhibitors in human lung cancer. *Br J Cancer*. 2014 Nov 25;111(11):2203-4. doi: 10.1038/bjc.2014.401. Epub 2014 Aug 12. PubMed PMID: 25117816; PubMed Central PMCID: PMC4260019.
5. **Chandrani P** and Dutt A. Domain Specific Targeting of Cancer. in Nuclear Signalling Pathways and Targeting Transcription in Cancer 299-310 (Springer, 2014).
6. 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. eCollection 2013. PubMed PMID: 24124538; PubMed Central PMCID: PMC3790706.

ACKNOWLEDGEMENT

It has been five years since I embarked on my dream journey to pursue doctorate of philosophy at Dutt laboratory, ACTREC-Tata Memorial Centre. A great number of people have helped in making this journey possible and enjoyable for me. Today, I take this opportunity to thank them all in my humble acknowledgement.

Firstly, I would like to thank my thesis supervisor Dr. Amit Dutt. Doctoral research under the guidance of Dr. Amit Dutt has been a single greatest learning experience of my life. His scholastic guidance, dynamic supervision and congruent encouragement and inspiration in planning and execution of the research projects are really appreciable. He has been a rock-solid support and a good friend while I have been through highs and lows in my life. I would like to thank him with my deepest respect and gratitude.

I would like to thank my doctoral committee members Dr. Girish Maru (ex-chairperson), Dr. Narayana Rao (ex-chairperson), Dr. Shubhada Chiplunkar (chairperson), Dr. Prasanna Venkatraman, Dr. Manoj Mahimkar, Dr. Santosh Noronha, Dr. Farhat Habib for all their guidance and invaluable suggestions throughout the work. I always enjoyed my presentations and discussions with doctoral committee members and have learnt a lot from all of them.

I would like to express my special gratitude for our clinical collaborator Dr. Kumar Prabhash for valuable advice and support throughout the work. I am thankful to him for being inspirational and providing clinical touch to my work. I would also like to extend my sincere gratitude towards Dr. Anuradha Chaughule and her entire team for providing precious tumor samples. I am equally thankful to the entire team of thoracic OPD at Tata Memorial Hospital for aiding me in clinical data collection. I especially thank Dr. Vijay Patil for teaching me clinical data analysis and Dr. Rajiv Kumar for histo-pathological analysis.

I have been fortunate to benefit from excellent infrastructure and facilities of ACTREC, inspirational environment and great people of 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 ACTREC for providing PhD fellowship, Wellcome Trust/DBT India Alliance, Department of Biotechnology, India and Tata Memorial Centre for funding the research work. I am grateful to Homi Bhabha National Institute (HBNI) and Tata

Memorial Centre (TMC) for proving financial assistance to present my work at an international platform.

I would be humbly thankful to Dr. Randeep Singh, for sparkling discussions regarding different aspects of next-generation sequencing. I would also like to thank Prof. Indraneel Mitra for providing access and Dr. Raghuram G.V. for helping in karyotyping analysis. I am thankful to Dr. Rita Mulherkar and Anti-Cancer Drug Screening Facility at ACTREC for providing HNSCC cell lines. I would also like to thank Mr. Prasad Kanvinde, Mr. Anand Jadhav, Mr. Padmakar Nagle, Mr. Manoj Chavan and entire IT department for their proficient IT/computer related help, Mr. Uday Dandekar for his efforts to maintain the common instrument facility. I also thank Dr. Rahul Thorat, Dr. Arvind Ingle and Dr. Pradip Chaudhari of animal house for their help during my work. My special thanks to Mr. Mahesh Pawar, Mr. Shetty, Mr. Bhabani Mohanty and other animal house staff members who have helped me in animal experimentation. I am also thankful to the staff in microscopy, genomics, library, steno-pool, administration, account and security departments of ACTREC for their constant help and making all the processes organized and smooth during this time.

I would like to thank Dr. Shilpee Dutt and her entire team with special respect. The wonderful, encouraging and caring friends and colleagues of Dutt lab and Shilpee lab have immensely contributed to my scientific work and personal life. I thank Pawan, Ratnam, Malika, Vidya, Vaibhav, Bikram, Deepak, Madhur, Ninad, Dr. Jyoti, Dr. Hemant, Ankita, Dr. Manoj and Mayur for their valuable friendship and contribution to my work. I am equally thankful to Sameer, Ekjot, Alokananda, Mukul, Kanishka, Jyothi, Nilesh, Dr. Kuldeep, Kunal, Rohan, Abhishek, Maulik, Jacinth, Trupti, Prajish, Sharan, Sanket, Asim, Anagha, Shailesh, Prachi, Rachayata, Smita mam and Renu mam for great support and making life enjoyable throughout this time. Together we have passed through ups and downs and sure have bonded in a unique way. I would like to express a friendly thanks to all of these guys.

I would also like to thank all of my friends Rasika, Prasanna, Rahul, Sonali, Qadir, Divya, Sajad, Moquit, Priyanka, Amit, Sanket, Dhwanit, Alok, Ram, Gauri, Rupa, Abira, Dhanashree, Burhan, Tanmay, Aditi, Swati, Sushmita, Amir, Sujath and all other seniors/juniors for all the memorable time we spent together.

My deepest gratitude will always remain for my previous teachers Dr. Sumeet Kumar, Dr. Archana Gattupalli, Dr. G. Naresh Kumar, Dr. Pranav Vyas, Dr. Shashikant Acharya, Mr. Jatin Kotak and Mr. Ilesh Parmar. Their thought-provoking teaching have

constantly inspired me to indulge into scientific research. The philosophical and scientific learning at the early stage of career has been helpful for embarking onto the doctoral research with great motivation and energy.

The complete journey of doctoral research would not have been possible without strong support and understanding of my family. I would like to thank my whole family for nurturing me, making me capable of fulfilling my dreams, forgiving me for my mistakes and ignoring my absence in many family events. I thank you all for your love, caring and your trust in me.

Finally, I proudly dedicate this thesis to my parents, brother Pallav and wife Vaishakhi. Dad and Mom, I can't ever thank you enough for your love, care and nurturing this young scientist from an early age. Pallav and Vaishakhi, thank you for loving and caring companionship, love you.

Dedicated to my dear

Dad – Mom

Pallav and

Vaishakhi

21th July, 2016 Pratik Chandrani

TABLE OF CONTENTS

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
CTD2	Cancer Target Discovery and Development
CTRP	Cancer Therapeutics Response Portal
dbSNP	Single Nucleotide Polymorphism Database
DNA	Deoxyribonucleic acid
EGFR	Epidermal Growth Factor Receptor
ExAC	Exome Aggregation Consortium
FDA	Food and Drug Administration
FFPE	Formalin-Fixed, Paraffin-Embedded
FGFR3	Fibroblast Growth Factor Receptor 3
GDSC	Genomics of Drug Sensitivity in Cancer
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
IPT-ICGC	India Project Team - International Cancer Genome Consortium
IRB	Institutional Review Board
JFCR	Japanese Foundation for Cancer Research
KRAS	Kirsten rat sarcoma viral oncogene homolog
mL	Millilitre
MLEM	Maximum Likelihood Expectation Maximization
mm	Millimetre
NCI	National Cancer Institute
ng	Nanogram
NGGF	Next-Generation Genomics Facility
NGS	Next-Generation Sequencing
NRBP1	Nuclear Receptor Binding Protein 1
OPD	Out Patients Department
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

SUMMARY

Several human diseases, most notably cancer research has seen unprecedented progress in past decade due to identification of various molecular alterations driving tumor growth and development of highly specific therapeutic solutions targeting these molecular alterations. Several global consortia based projects have been initiated to profile large number of tumor genomes to identify therapeutically exploitable genomic alterations. These findings have revealed promising preliminary results in clinical practice such as Herceptin for breast cancer, Imatinib for leukemia, Erlotinib for lung cancer etc. that are unfortunately primarily restricted to developed nations. On the other hand, to generate and utilize these resources has been deterrent for the developing world due to the lack of appropriate infrastructure, resources and technical expertise requirement. As an attempt to circumvent the problem and help the wider community to tap the knowledge present in vast genomic data, my thesis focuses on following three functional aspects of the cancer genome as detailed below:

1) Generate a landscape of therapeutically relevant alterations using high-throughput sequencing platform in human lung adenocarcinoma samples of Indian origin. In brief, I split my study design to sequence a smaller set of 45 samples by next-generation sequencing as the discovery set. Next, the mutations found were specifically genotyped in a larger validation set of 363 samples using mass-spectrometry based genotyping. This study led to establishment of the first therapeutically relevant landscape of alterations in lung adenocarcinoma of Indian origin. This study also led to the identification of novel recurrent mutations in fibroblast growth factor receptor 3 (*FGFR3*) which were validated to be oncogenic and sensitive to pharmacological inhibition using *in-vitro* and *in-vivo* approaches.

2) I present a proof-of-principle strategy for deriving biological interpretation from studies involving fewer samples by performing integrated genomics analysis. In brief, I set to establish an integrated genomics analytical workflow by systematically integrating data from multi-platform omics. Four rare cancer cell lines derived from head & neck cancer patients of Indian origin were characterized by adopting a posterior filtering approach. This allowed us to identify most of the major hallmark alterations in head & neck cancer, otherwise possible to be determined only using larger sample cohort. In addition, the study also led to the discovery of a novel oncogene Nuclear Receptor Binding Protein (*NRBP1*) in head and neck cancer.
3) I developed a user friendly computational tool – HPVDetector, specifically for non-computational researchers to analyze the high-throughput dataset. The tools is easy to install, comes along with graphical user interface (GUI) to help biologists utilize the huge cancer genomic dataset already available as free resource in repositories or generated by themselves using next-generation sequencing to detect presence of Human Papillomavirus (HPV) with minimal third party dependencies.

Homi Bhabha National Institute

SYNOPSIS OF PhD THESIS

1. **Name of the Student:** Pratik Chandrani
2. **Name of the Constituent Institution:** Tata Memorial Centre, ACTREC
3. **Enrolment No. and Date of Enrolment:** LIFE09201104003, 25th July, 2011
4. **Title of the Thesis:** Discovery of Potential Therapeutic Targets in Human Cancer: A Functional Genomics Approach
5. **Board of Studies:** Life Science

SYNOPSIS

1. INTRODUCTION

Cancer, causal of second most disease associated mortality worldwide, is a complex and dreadful disease [1]. Despite improvement of cancer patient survival in general, there is an unmet need of better therapeutic solutions due to non-specificity, excessive side effects and limited gain in survival by conventional therapies [2, 3]. Extensive biological research in past three decades has redefined cancer as genetic disease arising from a stepwise accumulation of genetic and epigenetic alterations that deregulate complex regulatory pathways of genes, proteins and biochemical components affecting cellular growth, division, migration and survival [4, 5]. While some of the genomic alterations are inherited through germline, majority of them are somatically acquired due to excessive instability of the cancer genome [6]. Majority of acquired alterations are random events and are not sufficient to transform normal cell, hence called passenger alterations, while few alterations called driver events are actually driving the cancer. Experimental characterization of driver genomic alterations in oncogenes and tumor-

suppressor genes has not only fostered our understanding of molecular and biochemical pathway of carcinogenesis, but also to establishment of oncogene addiction theory [7]. When tumor cells become addicted to oncogenic signaling, reversal of only one or a few of the signaling pathways can inhibit the growth of cancer cells, thus providing a rationale for molecularly targeted therapy. This novel approach of targeting driver alterations has shown impressive results in clinics, for example chronic myeloid leukemia harboring BRC-ABL fusion are treated with selective inhibitor Imatinib [8], lung adenocarcinoma harboring EGFR alterations are treated with Gefitinib or Erlotinib [9, 10], Breast cancer patients harboring HER2 amplification are treated with Trastuzumab [11], BRAF mutant melanoma treated with Vemurafenib [12] and many others have shown improvement of clinical outcome in practice [13]. Identification of above mentioned classical genomic alterations using conventional genomic approach were tedious and time consuming processes. For example, BCR-ABL1 fusion discovery to Imatinib drug development took ~40 years, and HER2 amplification discovery to approval of antibody Herceptin took ~15 years [14]. Recent technological advancement of high-throughput genomic analysis techniques, such as massively parallel next-generation sequencing, has enabled faster discovery of therapeutic targets in cancer genome, for example, EML4-ALK fusion discovery to Crizotinib drug approval took only 3 years [14]. Next-generation sequencing (NGS) of human genome can be performed now in under 14 days at ~ 3000 US $ [6] which has been drastically reduced compared to first human genome sequencing taking ~13 years and costing ~1.2 billion US $ (https://www.genome.gov/sequencingcosts/). This, in turn, has resulted into large scale tumor sample profiling consortiums such as The Cancer Genome Atlas (TCGA: http://cancergenome.nih.gov) and International Cancer Genome Consortium (ICGC: https://icgc.org) to systematically study cancer genome and discover therapeutically

relevant alterations. These and other individual studies have revealed hundreds of novel molecular alterations in tumor genome for which several pharmaceutical compounds are under development or clinical trials [15, 16]. Despite these global efforts, there are ethnicity specific alterations and differences in therapeutic response rates which are not yet explored and understood completely [17, 18]. Several research groups across the globe are taking similar genomic approach to profile genomic alterations in each population to identify ethnicity specific molecular alterations followed by therapeutic applications.

Diversity of genomic alterations and their therapeutic response in different population demands systematic biological and functional analysis of underlying molecular mechanisms [19, 20]. Functional studies of molecular somatic alterations require a tractable model system for *in-vitro* and *in-vivo* experimentations. The goal of functional genomics is to integrates information from various molecular analysis to gain an understanding of how complex biological function is governed in a cell. Patient derived immortalized cell lines have been utilized for several molecular biological and functional genomic assays [21]. Systematic genomic characterization and functional genomic analysis of cancer cell lines have been undertaken by global consortium such as Cancer Cell Line Encyclopedia (CCLE: http://www.broadinstitute.org/ccle) [22] and individual research groups [23], generating a well annotated cell line library for further functional and mechanistic study. Several functional genomic screening assays have also been adapted for high-throughput screening in cancer cells such as genome wide knock-down using RNA interference (RNAi) and pharmacological compound libraries [24, 25]. Together, these developments have opened up the way for biological feature based identification of novel cellular dependencies which can be subsequently used as therapeutic targets in cancer. Further development and systematic characterization of

model cell lines representing vast diversity of tumor cells can further improvise on current understanding of the disease and their therapeutic solutions.

Vast majority of the high-throughput genomics and functional genomics data of primary tumors and cell lines have been generated and deposited into open-source databases freely available to the community [15, 26]. Efficient and optimal utilization of this information requires computational expertise to work with complex command-line operated tools in Linux/Unix based operating systems, limiting this field to bioinformaticians and computational experts. Many of the bioinformatics algorithms have now matured enough to make automated computational tools with minimal technical input required form the user [27-29]. Development of online web-interface based cancer genomic analysis tools such as cBioPortal [30] and Galaxy [31] and offline computer based tools such as Integrative Genomics Viewer (IGV) [32] and others [29] have provided a user-friendly interface for efficient utilization of high-throughput genomic data. Despite this development, there has been unmet need for various computational tools with easy to use user-interface and abstracted computational complexity.

2. RATIONALE

Clinically actionable alterations have been characterized in diversified human cancers by TCGA, ICGC [33, 34] and other individual research groups [35-37]. While majority of these studies have characterized Caucasian population, ethnicity specific diversity has not been well explored and understood. Many of the recent studies have focused on genomic profiling of specific populations and have reported several significant differences in frequency of alterations and therapeutic response rate [17, 38]. India has been lacking back in utilization of targeted therapeutics because of deficit in systematic

genome wide profiling of cancer patients of Indian origin. Thus, an objective of this study focuses on systematic profiling of clinically actionable genomic alterations in cancer patients of Indian origin.

Similar to systematic genomic analysis of primary tumors of each specific population, functional genomic and mechanistic analysis of therapeutic solutions requires well characterized cell lines derived from patients of Indian origin. While only few patient derived cell lines of Indian origin are available (Anti-Cancer Drug Screening Facility, ACTREC: http://www.actrec.gov.in/anticancer_main.htm) [39], but lack of genome wide characterization and annotations restrict the usage of these cell lines from vast majority of high-throughput functional genomics experiments. Hence, one of objective of this study also aims to characterize cell lines of Indian origin using genome wide approach.

Analysis of primary tumors and cell lines using genome wide high-throughput technologies require complicated computations setup and expertise, restricting biologists and clinicians from efficiently utilizing genomics data. Therefore, one of the objective of this study focuses on development of user-friendly bioinformatics tool for genomic data analysis.

3. KEY QUESTIONS

A. What are the clinically actionable genomic alterations in human cancers of Indian origin?

B. How adequate is limited number of cell line's evaluation for discovery of novel alterations?

C. Develop a friendly informatics tools for biologists to analyze high throughput data.

4. OBJECTIVES

- **Objective-1: Analysis of clinically actionable alterations in Non-Small Cell Lung Carcinoma (NSCLC).**

  Lung cancer is one of the most predominant cancer type worldwide and majority of cancer related deaths are attributed to lung cancer. Non-small cell lung cancer (NSCLC), a histopathological subtype of lung cancer accounting for ~80-90% of total cases is characteristically distinct from small cell lung cancer and is also a major target for therapeutic development [40] [41]. NSCLC is further divided into three sub-classes of adenocarcinoma (~50% of cases), squamous cell carcinoma (~30%), large cell carcinoma (~10%) and unclassified carcinoma (~10%). To identify clinically targetable genomic alterations in lung adenocarcinoma, most predominant sub-type of lung cancer, I take genomic approach followed by functional characterization of molecular alterations *in-vitro* and *in-vivo*.

  Discovery of novel *FGFR3* mutations using NGS

  NGS of 125 FFPE NSCLC samples was performed with target coverage of more than 1500x per base. Systematic computational analysis of NGS data was carried out to identify previously known cancer related genes such as *TP53*, *EGFR*, *KRAS*, *RB1*, *ERBB2* etc. along with novel mutations in *FGFR3*. Selected clinically relevant mutations were further validated using Sequenom MassArray platform in ~400 samples confirming mutation frequency of ~29% in *EGFR*, 15% in *KRAS*, 5% in *FGFR3*, 3% in *AKT1*, and 1% each in *PIK3CA*, *FGFR4* and *HER2*. Novel mutations of FGFR3 were further selected for functional characterization using in-vitro and in-vivo assay systems.

***In-vitro* & *in-vivo* assays to determine the oncogenic potential of *FGFR3* mutants**

While the focus of the study is to identify clinically actionable genomic alterations, we determined the oncogenic potential and drug sensitivity of *FGFR3* mutations observed in lung adenocarcinoma by cloning *FGFR3* wild-type, R248C, S249C, and G691R in NIH-3T3 cells. NIH-3T3 cells harbouring mutant *FGFR3* formed colonies in significantly higher quantities then wild-type *FGFR3* harbouring cells. Similarly, Tumour formation was observed in 14 out of 14 mice for FGFR3-S249C followed by 6 out of 14 mice for FGFR3-G691R and 3 out of 14 mice for FGFR3-WT overexpressing NIH-3T3 cells. Tumor size was stably maintained during 11 days' BGJ398 drug treatment in the xenografts of mutants FGFR3-G691R and FGFR3-S249C. NIH-3T3 cells expressing *FGFR3* mutants showed hyper phosphorylation and ligand independent activation of downstream proteins ERK1 and AKT1 which was blocked in the presence of inhibitor, as confirmed by western-blotting and immunohistochemistry of xenografts.

***FGFR3* – a clinically relevant oncogene for lung adenocarcinoma**

I accessed the clinical outcome of patients involved in current study to understand the overall survival in presence of different genetic alterations. Consistent with previous reports [38], *EGFR* positive patients receiving TKI showed mean survival of 24 months (95% CI: 18.4-29.6 months; HR: 0.3) versus 12 months (95% CI: 10.3-14.2) survival of patients not harboring known oncogenic mutations. *KRAS* positive mutations had worst survival being mean survival of 12 months (95% CI: 6-18.8 months; HR: 1.1), while *FGFR3* positive patients had mean survival of 15 months (95% CI: 9.2-22.7 months; HR: 0.69) being intermediate to no mutant and *EGFR* positive patients receiving TKI.

These observations indicate that *FGFR3* mutations can act a positive predictor of clinical outcome of chemotherapy for lung adenocarcinomas. Nevertheless, a phase I clinical trial of BGJ398, a selective FGFR inhibitor, tends to improve the overall survival of *FGFR3* mutants [42]. Several other selective inhibitors of *FGFR* family are under clinical trials (trial no. NCT01928459, NCT02160041, NCT02278978, NCT01697605) which can further provide better therapeutic options for *FGFR* positive patients.

- **Objective-2: Integrated genomic characterization of Head and Neck Squamous Cell Carcinoma (HNSCC) cell lines derived from Indian patients.**

Tumours arising from oral cavity, tongue, pharynx and larynx are all together classified as HNSCC. HNSCC is one of the most prevalent cancer is India, being 3rd most common with 33% mortality rate [43]. HNSCC patients derived cell lines have been used as good model system for functional study of genomic alterations and effective therapeutic development. Here, I take integrated genomic approach to characterise and identify biologically relevant genomic alterations from HNSCC cell lines derived from Indian patients using high-throughput SNP array and NGS platforms.

Integrated genomic analysis of cell lines

Karyotype analysis: The hyperploidy status of four tumour-derived cell lines (NT8e [44], AW13516[45], AW8507[45] and OT9) were inferred by classical karyotyping with an average ploidy of 62, 62, 66 and 64, respectively [44, 45].

*Copy number analysis***:** We performed genotyping microarray using Illumina 660W quad SNP array chips of all the cell lines. Several known alterations such as loss of copy number and LOH at 3p [46, 47]; copy number gain on 11q [48] 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.

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. 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.

Analysis of mutational landscape: All the cell lines were sequenced for whole exome at about 80X coverage using Illumina HiSeq. 20 HNSCC hallmark variants including *TP53* (R273H), *TP53* (P72R), *PTEN* (H141Y), *EGFR* (R521K), *HRAS* (G12S and R78W), and *CASP8* (G328E) were identified and predicted as deleterious by two of three algorithms used for functional prediction.

Integrated analysis of copy-number, gene expression and mutation data: 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, median expression of focally amplified and deleted genes positively correlated to their expression. The second step of integration analysis involved identification of mutated genes that were expressed. Next, as third step of integration, genes harboring two or more type of alterations were selected: harboring positive correlation of focal copy number and gene expression; or those harboring point mutations with detectable transcript harboring the variant—based on which, we identified 38 genes having multiple types of alterations. 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 [49].

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 [50, 51]. Validation and biological characterization of this novel variant studied independently in our laboratory 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.

- **Objective-3: HPVDetector: NGS Based Approach to Determine the Presence of HPV and Their Sites of Integration in Human Cancer Genome**

Human papilloma viral (HPV) infections has been associated with various types of cancer. Epidemiological studies indicate that about 90% of cervical cancers, 90-93% of anal canal cancers, 12-63% of oropharyngeal cancers, 36-40% of penile cancers, 40-64% of vaginal cancers and 40-51% of vulvar cancers are attributable to HPV infection [52, 53]. Currently HPV detections are primarily carried out using PCR based MY09/11 and CPI/II systems [54], hybridization based SPF LiPA method, signal amplification assays (Hybrid Capture 2 and Cervista) etc. [54-56]. These technologies come with limitations to detect minor, low-abundance HPV genotypes and complex mixture of co-infections that can be a negative determinant of the clinical outcome [57, 58]. Next generation sequencing (NGS) technologies overcomes such limitations, as evident from the recently described TEN16 methodology [59], and few other studies [60, 61]. However, there's an unmet need for a simplified bioinformatics tool for biologists and clinicians with no previous experience or knowledge of informatics to analyze the data generated by whole exome, transcriptome or genome

sequencing using NGS technology to detect the presence of HPV sequences along with their integration sites.

HPVDetector is a tool to quickly detect hundreds of Human Papilloma Virus types from next generation sequence data without any prerequisite knowledge about virus types. It runs on paired end sequenced samples. It is composed of two modes or sub pipelines as quick detect & integration detect mode.

Quick detect and integration detection mode

Quick detect mode is to quickly determine the HPV type or types to check if multiple HPV co-infections are existing or not in a given sample. This mode output result file which enlists one or more HPV type(s) and number of HPV Reads. Integrations detection mode of HPVDetector determines genomic location of HPV integrant, annotate with HPV gene, human chromosomal loci, human gene and cytobands.

Detection of HPV type integrated in the host genome

In order to test the precision of HPVDetector we analyzed 22 cervical cancer exome sequencing data, 23 paired and one orphan tongue squamous cell carcinoma (TSCC) sample and 7 head and neck squamous cell carcinoma cell lines, 13 gall bladder cancer whole exome, 1 gall bladder cancer whole transcriptome and 1 liposarcoma whole genome sequence data. Among the 22 cervical samples analyzed, HPV was detected in 18 cervical samples, with maximum number of reads supporting HPV16 sequence. We also detected the presence of additional HPV types such as HPV71, HPV82 and HPV31 with some of the HPV types co-infecting the patients. None of the TSCC primary tumors were found to be HPV positive, as reported earlier [62, 63]. At the same time, among the cell lines, NT8e cells [44] of 7 cell lines analyzed were found to be positive for HPV71. No trace of HPV sequence was detected in gallbladder and liposarcoma samples.

Determination of the HPV integration sites in the host genome

I identified 55 integration sites in 7 cervical cancer tumor samples and 1 head and neck tumor sample using the HPVDetector. In this study, chromosomal loci 17q21, 3q27, 7q35, Xq28 were observed with higher frequency compared to other loci for HPV integration, as reported earlier [64].

In total, I analyzed 116 whole exome, 23 whole transcriptome and 2 whole genome sequencing data sets, out of which I have detected presence of HPV in 20 exome and 4 transcriptome data.

5. CONCLUSIONS AND FUTURE PROSPECTS

This study represents the first landscape of clinically actionable alterations in non-small cell lung cancer of Indian origin. In addition to novel genomic alterations in *FGFR3*, other known alterations in *EGFR*, *KRAS*, *EML4-ALK*, *AKT1*, *PIK3CA*, *FGFR4* and *HER2* genes were identified in ~400 NSCLC patients. Further, novel *FGFR3* mutations were confirmed to be activating and sensitive to small-molecule inhibitor using *in-vitro* and *in-vivo* approach. Additionally, patient survival data shows that *FGFR3* mutations can act as positive prognosis predictor, however, a large sample study should be required to confirm *FGFR3* as prognosis marker. Taken together, this study suggests that tumors driven by *FGFR3* in murine xenografts are sensitive to pharmacological inhibition and systematic clinical trials should be designed to test *FGFR3* targeted therapy and prognosis potential in patients.

As a proof of principle, integrated analysis of copy number variation, exome and transcriptome of 4 cell lines derived from Indian HNSCC patients and TCGA HNSCC dataset identify *NRBP1*, *UBR5*, *ZNF384* and *TERT* as novel candidate oncogenes in HNSCC. However, systematic 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.

Additionally, user-friendly genomic analysis tool called HPVDetector has been developed for HPV detection from tumor samples using variety of NGS platforms including whole genome, whole exome and transcriptome. Two different modes (quick detection and integration mode) along with a GUI widen the usability of HPVDetector for biologists and clinicians with minimal computational knowledge.

6. REFERENCES

1. Mortality, G.B.D. and C. Causes of Death, *Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013.* Lancet, 2015. **385**(9963): p. 117-71.
2. Zeng, C., et al., *Disparities by Race, Age, and Sex in the Improvement of Survival for Major Cancers: Results From the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) Program in the United States, 1990 to 2010.* JAMA Oncol, 2015. **1**(1): p. 88-96.
3. Siegel, R., et al., *Cancer statistics, 2014.* CA Cancer J Clin, 2014. **64**(1): p. 9-29.
4. Tomasetti, C. and B. Vogelstein, *Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions.* Science, 2015. **347**(6217): p. 78-81.
5. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
6. Upadhyay, P., R. Dwivedi, and A. Dutt, *Applications of next-generation sequencing in cancer.* CURRENT SCIENCE, 2014. **107**(5): p. 795.
7. Weinstein, I.B. and A.K. Joe, *Mechanisms of disease: Oncogene addiction--a rationale for molecular targeting in cancer therapy.* Nat Clin Pract Oncol, 2006. **3**(8): p. 448-57.
8. Druker, B.J., et al., *Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome.* N Engl J Med, 2001. **344**(14): p. 1038-42.
9. Lynch, T.J., et al., *Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib.* N Engl J Med, 2004. **350**(21): p. 2129-39.
10. Thatcher, N., et al., *Gefitinib plus best supportive care in previously treated patients with refractory advanced non-small-cell lung cancer: results from a randomised,*

placebo-controlled, multicentre study (Iressa Survival Evaluation in Lung Cancer). Lancet, 2005. **366**(9496): p. 1527-37.
11. Slamon, D.J., et al., *Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2.* N Engl J Med, 2001. **344**(11): p. 783-92.
12. Flaherty, K.T., et al., *Inhibition of mutated, activated BRAF in metastatic melanoma.* N Engl J Med, 2010. **363**(9): p. 809-19.
13. Sudhakar, A., *History of Cancer, Ancient and Modern Treatment Methods.* J Cancer Sci Ther, 2009. **1**(2): p. 1-4.
14. Chin, L., J.N. Andersen, and P.A. Futreal, *Cancer genomics: from discovery science to personalized medicine.* Nat Med, 2011. **17**(3): p. 297-303.
15. Cancer Genome Atlas Research, N., et al., *The Cancer Genome Atlas Pan-Cancer analysis project.* Nat Genet, 2013. **45**(10): p. 1113-20.
16. Yap, T.A. and P. Workman, *Exploiting the Cancer Genome: Strategies for the Discovery and Clinical Development of Targeted Molecular Therapeutics.* Annual Review of Pharmacology and Toxicology, 2012. **52**(1): p. 549-573.
17. Bollig-Fischer, A., et al., *Racial diversity of actionable mutations in non-small cell lung cancer.* J Thorac Oncol, 2015. **10**(2): p. 250-5.
18. O'Donnell, P.H. and M.E. Dolan, *Cancer pharmacoethnicity: ethnic differences in susceptibility to the effects of chemotherapy.* Clin Cancer Res, 2009. **15**(15): p. 4806-14.
19. Chin, L., et al., *Making sense of cancer genomic data.* Genes Dev, 2011. **25**(6): p. 534-55.
20. Boehm, J.S. and W.C. Hahn, *Towards systematic functional characterization of cancer genomes.* Nat Rev Genet, 2011. **12**(7): p. 487-98.
21. Domcke, S., et al., *Evaluating cell lines as tumour models by comparison of genomic profiles.* Nat Commun, 2013. **4**: p. 2126.
22. Yang, W., et al., *Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells.* Nucleic Acids Res, 2013. **41**(Database issue): p. D955-61.
23. Sharma, S.V., D.A. Haber, and J. Settleman, *Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents.* Nat Rev Cancer, 2010. **10**(4): p. 241-53.
24. Schlabach, M.R., et al., *Cancer proliferation gene discovery through functional genomics.* Science, 2008. **319**(5863): p. 620-4.
25. Johannessen, C.M., P.A. Clemons, and B.K. Wagner, *Integrating phenotypic small-molecule profiling and human genetics: the next phase in drug discovery.* Trends Genet, 2015. **31**(1): p. 16-23.
26. Zhang, J., et al., *International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data.* Database (Oxford), 2011. **2011**: p. bar026.
27. Alioto, T.S., et al., *A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing.* Nat Commun, 2015. **6**: p. 10001.
28. Yang, Y., et al., *Databases and web tools for cancer genomics study.* Genomics Proteomics Bioinformatics, 2015. **13**(1): p. 46-50.
29. Ding, L., et al., *Expanding the computational toolbox for mining cancer genomes.* Nat Rev Genet, 2014. **15**(8): p. 556-70.
30. Gao, J., et al., *Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.* Sci Signal, 2013. **6**(269): p. pl1.
31. Blankenberg, D., et al., *Galaxy: a web-based genome analysis tool for experimentalists.* Curr Protoc Mol Biol, 2010. **Chapter 19**: p. Unit 19 10 1-21.

32. 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.
33. Zack, T.I., et al., *Pan-cancer patterns of somatic copy number alteration.* Nat Genet, 2013. **45**(10): p. 1134-40.
34. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-8.
35. Rusan, M., Y.Y. Li, and P.S. Hammerman, *Genomic landscape of human papillomavirus-associated cancers.* Clin Cancer Res, 2015. **21**(9): p. 2009-19.
36. Wen, Y.S., et al., *Concurrent oncogene mutation profile in Chinese patients with stage Ib lung adenocarcinoma.* Medicine (Baltimore), 2014. **93**(29): p. e296.
37. Horvath, A., et al., *Novel Insights into Breast Cancer Genetic Variance through RNA Sequencing.* Scientific Reports, 2013. **3**.
38. Noronha, V., et al., *EGFR mutations in Indian lung cancer patients: clinical correlation and outcome to EGFR targeted therapy.* PLoS One, 2013. **8**(4): p. e61561.
39. Pandrangi, S.L., et al., *Establishment and characterization of two primary breast cancer cell lines from young Indian breast cancer patients: mutation analysis.* Cancer Cell Int, 2014. **14**(1): p. 14.
40. Travis, W.D., *Classification of lung cancer.* Semin Roentgenol, 2011. **46**(3): p. 178-86.
41. Molina, J.R., et al., *Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship.* Mayo Clin Proc, 2008. **83**(5): p. 584-94.
42. Sequist, L.V., et al., *Phase I study of BGJ398, a selective pan-FGFR inhibitor in genetically preselected advanced solid tumors*, in *American Association for Cancer Research 2014 Congress*. 2014: San Diego, CA, USA.
43. Dikshit, R., et al., *Cancer mortality in India: a nationally representative survey.* Lancet, 2012. **379**(9828): p. 1807-1816.
44. 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.
45. 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.
46. 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.
47. 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.
48. 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.
49. 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.
50. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603-7.
51. Davies, H., et al., *Somatic mutations of the protein kinase gene family in human lung cancer.* Cancer Res, 2005. **65**(17): p. 7591-5.
52. Munoz, N., et al., *Epidemiologic classification of human papillomavirus types associated with cervical cancer.* N Engl J Med, 2003. **348**(6): p. 518-27.
53. Shukla, S., *Infection of human papillomaviruses in cancers of different human organ sites* Indian J Med Res, 2009. **130**: p. 222-233.

54. Kleter, B., et al., *Novel short-fragment PCR assay for highly sensitive broad-spectrum detection of anogenital human papillomaviruses.* Am J Pathol, 1998. **153**(6): p. 1731-9.
55. Abreu, A.L., et al., *A review of methods for detect human Papillomavirus infection.* Virol J, 2012. **9**: p. 262.
56. Brink, A.A., P.J. Snijders, and C.J. Meijer, *HPV detection methods.* Dis Markers, 2007. **23**(4): p. 273-81.
57. Mendez, F., et al., *Cervical coinfection with human papillomavirus (HPV) types and possible implications for the prevention of cervical cancer by HPV vaccines.* J Infect Dis, 2005. **192**(7): p. 1158-65.
58. Trottier, H., et al., *Human papillomavirus infections with multiple types and risk of cervical neoplasia.* Cancer Epidemiol Biomarkers Prev, 2006. **15**(7): p. 1274-80.
59. Xu, B., et al., *Multiplex Identification of Human Papillomavirus 16 DNA Integration Sites in Cervical Carcinomas.* PLoS One, 2013. **8**(6): p. e66693.
60. Johansson, H., et al., *Metagenomic sequencing of "HPV-negative" condylomas detects novel putative HPV types.* Virology, 2013. **440**(1): p. 1-7.
61. Ameur, A., et al., *Comprehensive profiling of the vaginal microbiome in HIV positive women using massive parallel semiconductor sequencing.* Sci Rep, 2014. **4**: p. 4398.
62. 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.
63. Tsimplaki, E., et al., *Prevalence and expression of human papillomavirus in 53 patients with oral tongue squamous cell carcinoma.* Anticancer Res, 2014. **34**(2): p. 1021-5.
64. Thorland, E.C., et al., *Common fragile sites are preferential targets for HPV16 integrations in cervical tumors.* Oncogene, 2003. **22**(8): p. 1225-37.

Signature of Student:

Date: 28th March, 2016

Doctoral Committee:

S. No.	Name	Designation	Signature	Date
1.	Dr. AVSS Narayana Rao	Chairman		28·3·2016
2.	Dr. Amit Dutt	Guide/Convener		28/3/16
3.	Dr. Prasanna Venkatraman	Member		28/3/16
4.	Dr. Manoj Mahimkar	Member		28/3/2016
5.	Dr. Santosh Noronha	Invitee		28/3/16
6.	Dr. Farhat Habib	Invitee		28/3/16

Forwarded through:

S. Chiplunkar 22/4/16

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 410210.

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

I. GENERAL INTRODUCTION

There is a proverb in Gujarati culture (as nicely phrased by Indian poet Kalapi) "જે પોષતું તે મારતું, એ ક્રમ નિશ્ચ છે કુદરતી" (pronounced as: je poshatu te maratu, a kram nisch chhe kudarati) which means the one who nurtures also destroys, so is the principle of mother nature. According to this philosophy, the genes/pathways responsible for development and proliferation of tumor often becomes the vulnerability of the same tumor. As now demonstrated by modern cancer genomics, the genomic alterations driving tumor become most suitable therapeutic targets in themselves.

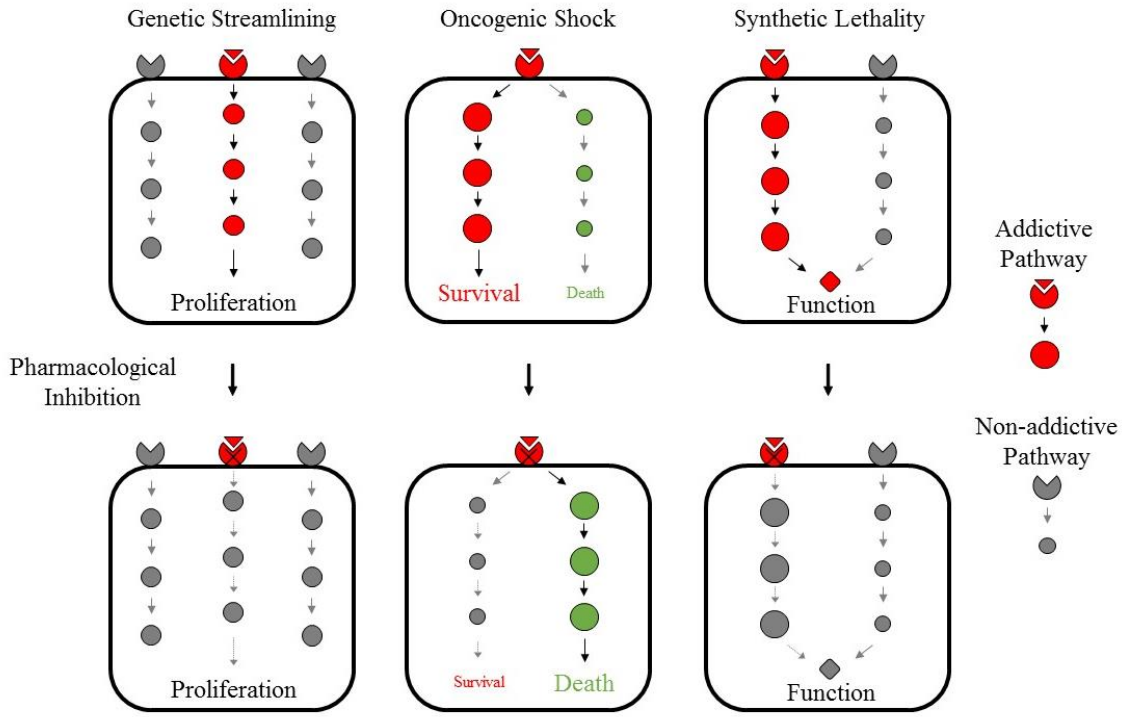
1.1 Human Cancers and Therapeutics

Cancer, causal of the second most disease associated mortality worldwide, is a complex and dreadful disease. Out of all known malignancies in humans, cancers of five organs (lung, stomach, head & neck, liver and colon) constitute 60% of cancer related deaths worldwide [1]. While incidence of cancer types in India largely matches the global pattern, few are more predominant than others such as cervical and head & neck cancer (GLOBOCAN, 2012; http://globocan.iarc.fr). Extensive research for understanding and treating cancer has resulted into overall improvement of life expectancy of patients of certain cancer types but still majority of cancer patients have median survival of less than 5 years after diagnosis [2, 3]. Conventional therapeutics such as surgery, radio- and chemo- therapy are successful in many cases but are mostly accompanied by excessive side effects resulting into overall poor quality of life [4-6]. Better therapeutic options are much needed not only for better survival of the patients but also to minimize side effects and improve the quality of life of the patients.

1.2 Genomics of Cancer

Given the perplexing diversity of cancer, investigation of commonalities in the genesis and progression of cancer is a daunting task. Extensive research and development in past five decades has resulted into identification of commonality of cancers, as defined by Hanahan and Weinberg "the hallmarks of cancer" that govern the transformation of normal cell to malignant cell [7]. These commonalities provide a hope that it can be exploited in combating the disease. One of the commonality in cancers is that the genesis and progression of tumor is governed by a handful of genes corrupted by genomic alteration. While few of the genomic alterations are inherited through germline, majority of them are stochastically acquired during the course of tumor genesis and progression [8]. Majority of acquired alterations are stochastic events and are not sufficient to provide growth advantage to the cell, hence are called passenger mutations while few mutations, called driver mutations, are able to provide selective growth advantage to the cell. Driver genes providing selective growth advantage through activating mutations are called oncogenes while those providing selective growth advantage through inactivating mutations are termed tumor suppressor genes. Identification and characterization of driver genomic alterations in oncogenes and tumor-suppressor genes has not only fostered our understanding of molecular and biochemical pathways of carcinogenesis, but has also lead to establishment of oncogene addiction theory [9-11]. Oncogene addiction, in its simplest incarnation, refers to the dependency of tumor cell on a single oncogenic protein or pathway for sustained proliferation and survival. Three models have been put forward (I. Figure 1) to elucidate the mechanism of oncogene addiction at molecular level, (i) genetic streamlining, (ii) oncogenic shock and (iii) synthetic lethality [12]. Theory of genetic streamlining postulates that addictive pathways (maintained in "on" state) are dominant over other non-essential pathways (maintained in "off" state), providing cellular growth

advantage. Cellular fitness collapse upon abrogation of addictive pathway, leading to cell-cycle arrest or apoptosis. Oncogenic shock theory postulates that addictive oncoprotein triggers pro-survival and pro-apoptotic signals at the same time and under normal



I. Figure 1: Mechanism of oncogenic addiction.

circumstances a pro-survival signal dominates. Abrogation of addictive oncoprotein leads to rapid decline of pro-survival pathway, subverting the balance in favor of pro-apoptotic signaling.

On the other hand, synthetic lethality considers two genes to be in synthetic lethal relationship when loss of either is still compatible with survival but loss of both is fatal. Altogether, these three mechanisms of oncogenic addiction reveal a possible "Achilles' heel" within the cancer cell that can be exploited therapeutically. When tumor cells become addicted to oncogenic signaling, reversal of only one protein or signaling pathway can inhibit the growth of cancer cells, thus providing a rationale for molecularly targeted therapeutics.

1.3 Genomically Guided Personalized Cancer Therapeutics

The novel approach of targeting addicted oncogene requires daunting task of drug development but has shown impressive results in clinics (I. Figure 2). For example, nearly 60 years after the discovery of Philadelphia chromosome (*BRC-ABL* fusion) in chronic myeloid leukemia, oncogene BRC-ABL targeting inhibitor Imatinib came into practice [13]. Trastuzumab, an antibody targeting *HER2* amplified breast cancer came into practice nearly after 15 years of its discovery [14]. Similarly, Erlotinib or Gefitinib inhibitors targeting *EGFR* mutations in lung adenocarcinoma [15, 16], Vemurafenib inhibitor targeting *BRAF* mutant in melanoma [17], Imatinib for *c-KIT* mutant gastrointestinal stromal cancers, and many others have successfully reached to clinical practice after several years of painstaking investigations [18, 19]. While these early genomically guided therapeutic solutions have shown positive results, the total applicability of them altogether has been limited to only a subgroup of patients. Identification of newer target oncogenes has always been a hurdle due to technological limitations. Low throughput of conventional genomics platforms coupled with limited rate of functional biological characterization of newly identified oncogenic alterations altogether restricts the overall development of more therapeutic solutions.

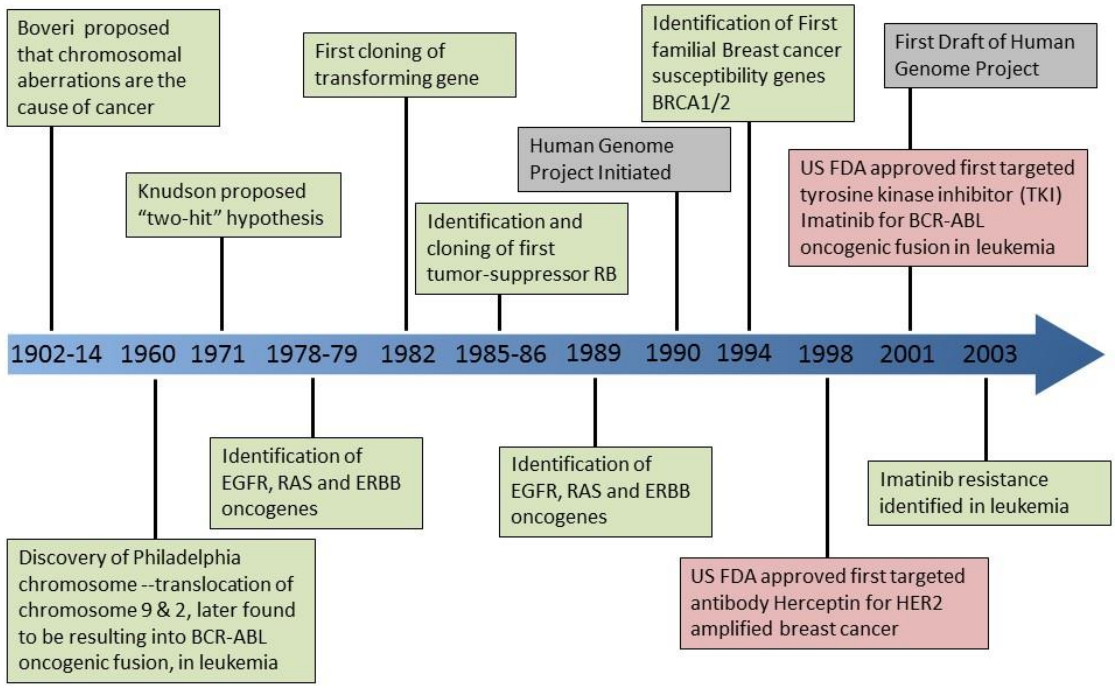
1.4 Modern Cancer Genomics is Redefining Cancer Therapeutics

Recent technological advancement in genomic analysis techniques, referred as massively parallel or next-generation sequencing, has enabled faster discovery of therapeutic targets in cancer genome at much cheaper cost (I. Table 1). As of 2016, next-generation sequencing (NGS) of human genome can be performed in under 14 days at ~ 3000 US $ [8] which has been drastically reduced compared to first human genome sequencing taking ~13 years at cost of ~1.2 billion US $ (https://www.genome.gov/sequencingcosts).

I. Table 1: Commonly available sequencing platforms.

Method (Generation)	Commonly Available Platforms	Sequencing cost per million base	Company
Chain termination - Sanger sequencing (1st)	Genetic Analyzer	$2400	ThermoFisher
Pyrosequencing (2nd)	GS Junior, GS FLX+	~ $10	Roche/454
Sequencing by synthesis (2nd)	Illumina HiSeq, NextSeq, MiSeq and MiniSeq	$0.05 - $0.15	Illumina
Sequencing by ligation (2nd)	SOLiD 5500 Genetic Analyzer	$0.13	ThermoFisher
Semiconductor sequencing (2nd)	Ion Proton, Ion PGM, Ion Chef	$0.6 - $1	ThermoFisher
DNA nanoball sequencing (2nd)	Sequencing as a service	--	Complete Genomics/BGI
Heliscope single molecule sequencing (2nd)	Sequencing as a service	--	SeqLL
Single molecule real time sequencing - SMRT (3rd)	PacBio RS II, Sequel	$0.13 - $0.6	Pacific Biosciences
Nanopore DNA sequencing (3rd)	MinION, PromethION	~$0.15	Oxford Nanopore

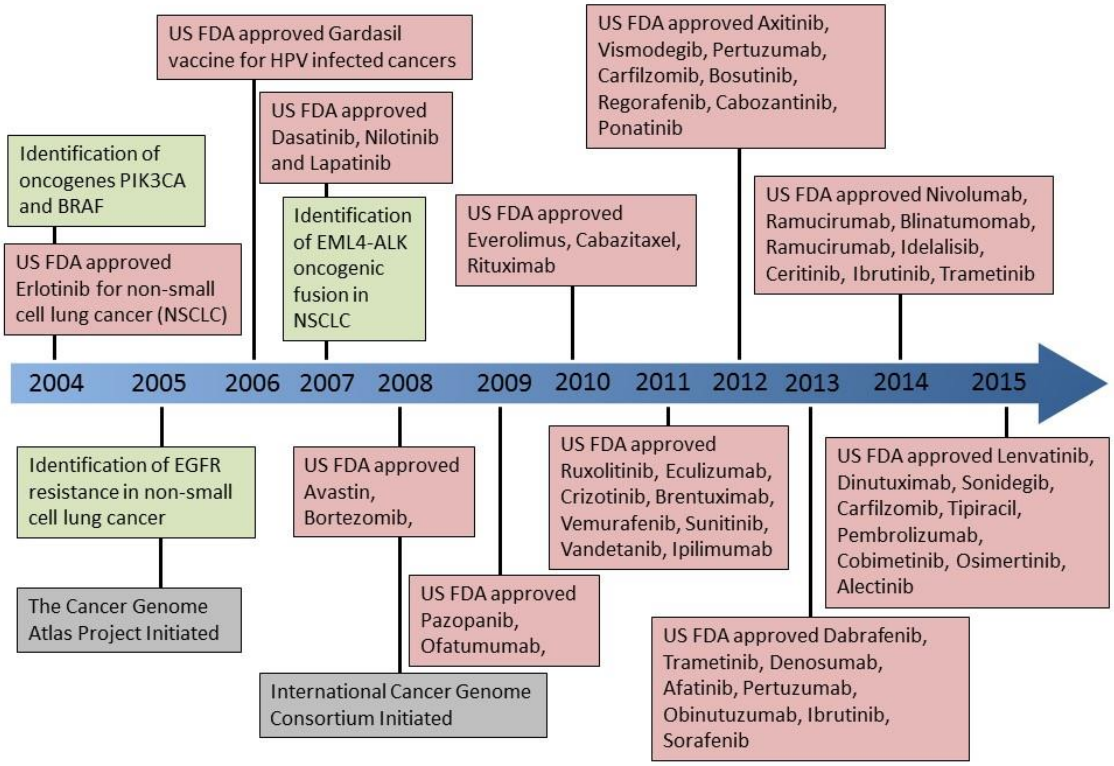
This, in turn, has resulted into large scale tumor sample profiling consortiums such as The Cancer Genome Atlas (TCGA: http://cancergenome.nih.gov) and International Cancer Genome Consortium (ICGC: https://icgc.org) to systematically study cancer genome and discover therapeutically relevant alterations. These and other individual studies have revealed landscape of novel molecular alterations in diverse tumor genomes for which several pharmaceutical compounds are under various stages of development or clinical trials [20, 21]. Despite these global efforts, there are ethnic specific alterations and



I. Figure 2: Selected key discoveries and approval of cancer therapeutics.

(continued on next page)

differences in oncogenic landscape and therapeutic responses which are not yet well explored and understood completely [22, 23]. Several research groups across the globe are taking similar genomic approach to profile genomic alterations in each population to identify ethnicity specific molecular alterations followed by therapeutic applications. On the other hand, several technological advancements for functional biological characterization of newly identified targets using tractable model system has further enhanced the rate of development of newer therapeutic solutions. Altogether, these technological developments has led to faster development of therapeutic solutions (I. Figure 2), for example, oncogenic EML4-ALK fusion discovery to Crizotinib drug approval by Food and Drug Administration (FDA) of United States of America took only 3 years [19].



I. Figure 2: Selected key discoveries and approval of cancer therapeutics.

(continued from previous page)

Furthermore, the precedential capacity of genomics platforms has helped us gain substantial insights into therapy resistance, tumor heterogeneity and evolution. According to the "trunk-branch model" [24] of tumor growth, somatic mutations during the early stage of tumor development represents the trunk of the tree. Over the time, additional somatic alterations in the subset of tumor cells result into branching in tumors as well as in metastatic sites. Later on, some of the heterogeneous tumor cells evolve further and become more isolated from the main tree and form branches. The branching can result into a "bottleneck effect" which further enhances chromosomal instability and expansion of tumor heterogeneity [24]. This ongoing and parallel evolution of cancer cells may contribute to faster adaptation of cancer cells to develop resistance against given therapy. In case of personalized targeted therapeutics, a proportion of patients are

reported to develop resistance following initial response to the therapy [25]. The investigation of resistant patients has revealed a range of devious molecular mechanisms to elude targeted therapeutics, including secondary alterations in the target gene, alterations in the downstream canonical pathways and activation of other parallel signaling pathways [25, 26]. Advent of technological advancement has helped researchers gain substantial insights into the rapidly evolving tumor genome, which was difficult to study otherwise. The new information obtained from tumor genome may further help in designing better diagnostic and therapeutic solutions to improve cancer treatment in the practice.

1.5 Functional Genomics Complements Modern Genomics

Discovery of new cancer drugs targeting specific genomic alterations are limited by the fact that underlying mechanism of oncogenic addiction and molecular pathway needs to be understood for successful drug designing. Therefore, the goal of functional genomics is to integrate information from various molecular analysis to gain an understanding of how complex biological function is governed in a cell. Historically, laboratory models such as patient derived cell lines and their xenografts in mice have been used for functional studies and drug efficacy testing. Although the cell line based drug efficacy tests have been used from very early time, only recently it has been realized that tumor derived cell lines represent genomic diversity similar to that of parent tumor [27, 28]. This realization of underlying genotype being responsible for diversity in clinical response to treatments has bolstered efforts to exploit cell lines for capturing genotype to clinical response relationships. The first of systematic genotyping and their correlation with pharmacological screen in panel of cell lines was performed as a part of National Cancer Institute - 60 (NCI-60) project (I. Table 2) [29]. NCI-60 panel represents tumor derived cell lines from nine cancer types which covers many diverse genomic alterations observed

in these tissue types. However, to recapitulate the diversity of genomic alterations in tumor tissues, especially those occurring at lower frequency, much larger number of cell lines are required. Later, several other consortium based pharmacogenomics studies such as the Cancer Cell Line Encyclopedia (CCLE) [27], Genomics of Drug Sensitivity in Cancer (GDSC) [30], Connectivity Map [31], the Cancer Therapeutics Response Portal (CTRP) and several other individual research groups have expanded the numbers of cell lines, drugs, and cancer types. Several functional genomic screening assays have also been adapted for high-throughput screening in array of cell lines such as genome wide knock-down using RNA interference (RNAi) and large pharmacological compound libraries screening [32, 33]. This has led to development of several newer therapeutic solutions which are currently at various stages of development and clinical trials. However, majority of these cell lines have been derived from patients in developed nations, hence lack the genomic diversity of different ethnic groups. Disparities in molecular alterations and clinical response in different population demands cell lines development and characterization from different populations [34, 35]. One of the similar approach is taken by Japanese Foundation for Cancer Research (JFCR) to establish a set of 39 cell lines, called JFCR-39, representing diversity of tumor types prevalent in Japan [36]. Several other research groups across the globe have taken similar approach [37] to characterize cell lines derived from patients of diverse ethnicity but largely these attempts have been restricted by limited number of cell lines poorly representing genomic diversity of each population.

I. Table 2: Cell lines in large pharmacogenomics databases.

Tumor tissue type	**NCI-60**	**JFCR-39**	**CCLE**	**GDSC**	**CTRP**
Bladder	0	0	28	18	5
Bone	0	0	29	31	1
Breast	5	5	60	43	1

Cervix	0	0	0	12	0
Colon	7	5	63	35	37
Endometrium	0	0	28	10	11
Head & neck	0	0	35	23	2
Hematopoietic and lymphoid	6	0	181	113	24
Kidney	8	2	36	22	2
Liver	0	0	36	14	4
Lung	9	7	187	141	91
Nervous system	6	6	86	79	3
Esophagus	0	0	27	23	3
Ovary	7	5	52	20	26
Pancreas	0	0	46	18	10
Prostate	2	2	8	5	1
Skin	10	1	62	45	9
Soft tissues	0	0	21	17	3
Stomach	0	6	38	18	6
Testis	0	0	0	2	0
Thyroid	0	0	12	12	0
Others	0	0	11	7	3
Total	60	39	1046	707	242

Limited availability of tumor derived cell lines can be compensated in some ways by two of the widely used murine cell lines NIH-3T3 and Ba/F3. Early in 1970s, Robert Weinberg and his group identified that NIH-3T3 cells, a line of immortalized mouse fibroblasts, were particularly adept at taking up transfected DNA followed by phenotypic changes induced by expression of transfected DNA. Early work of Weinberg group using NIH-3T3 model system led to discovery of first human oncogene Ras in 1982 [38]. Since then, NIH-3T3 cell line has become a *de facto* for testing transfection mediated oncogenic transformation of cells by various oncogenes. Similarly, another murine cell line Ba/F3 has also evolved as model system for oncogenic transformation testing. Ba/F3 is

interleukin (IL)-3 dependent hematopoietic cell line which was developed in 1980s. Ba/F3 cells which are dependent on IL-3 were first rendered to be IL-3 independent upon expression of Bcr-Abl tyrosine kinase [39], a well-known oncogene in chronic myeloid leukemia (CML). Since then, Ba/F3 cells have been used for various functional characterization of oncogenes, their molecular pathways and testing drug efficacies [40].

Altogether, development of various cell line based model systems coupled with advancement of functional genomic techniques has opened up the way for biological feature based identification of novel cellular dependencies which can be subsequently used as therapeutic targets in cancer. Furthermore, development and systematic characterization of model cell lines representing vast diversity of tumor cells of different ethnic groups can further improve our current understanding of the ethnic specific variation to clinical response in cancer therapeutics.

1.6 Computational Tools for Effective Utilization of Vast Genomics and Functional Genomics Data Sets

With the increase in throughput of vast majority of the genomics and functional genomics techniques, the volume and complexity of analysis and interpretation of these data also increases. [20, 41]. Vast majority of the data are made available for the cancer genome sequences including point mutations and structural alternations enabling the study of cancer at molecular level in an unprecedented global view. However, efficient and optimal utilization of massive amount of data remains a challenge due to limitations of computational methodologies, insufficient collaborations and sharing between biologists, clinicians and computational experts. Computational hardware has evolved to efficiently store, process and analyze large scale genomics data sets in compute clusters

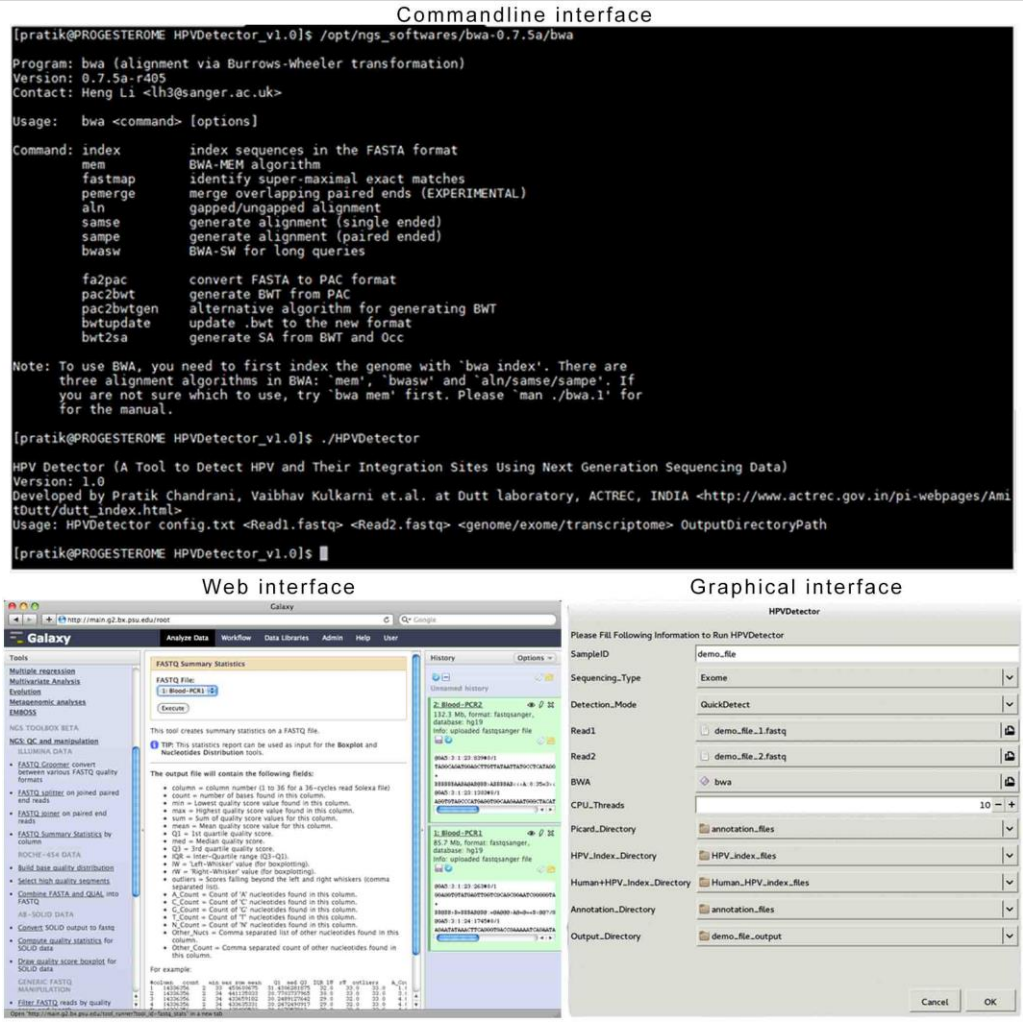
or cloud platforms. Similarly, several open source and commercial bioinformatics software has also evolved (I. Table 3) for quality control and filtration of large scale data, reproducible analysis, distinguish "true signals" from background noise with enough statistical power, and produce reusable output in standard file structure. [42-44].

I. Table 3: Selected commonly used cancer genomics data repositories, databases, tools and file formats (numbers as of 1st June, 2016).

Data repositories			
Database/Repository	Number of Cancer Types	Number of Specimens	Availability
TCGA	33	14531	Raw data: controlled access Processed data: publically available
ICGC	68	15613	
SRA/GEO	--	--	Raw data: publically available
Databases/Resources			
Name	Description		
dbSNP	A database of single nucleotide polymorphisms in germline of human genome.		
COSMIC	A database of mutations observed in cancer genome.		
TMC-SNPdb	A database of single nucleotide polymorphisms identified by whole exome sequencing of matched normal tissues of cancer patients of Indian origin.		
ExAC	A database of single nucleotide polymorphisms identified by whole exome sequencing.		
Quality control			
FASTQC	General raw data quality control checking tool.		
Qualimap	General purpose alignment quality checking tool.		
RNAseqQC	General purpose transcriptome data quality checking tool.		
Alignment tools			
BWA	General DNA/RNA sequencing alignment tool.		
Bowtie	General DNA/RNA sequencing alignment tool.		

TopHat	Mostly used for spliced transcriptome alignment.
STAR	Mostly used for spliced transcriptome alignment.
SAM/BAM processing	
Samtools	General SAM/BAM processing tool.
Picard tools	General SAM/BAM processing tool.
Variant calling	
Samtools	General variant calling tool.
GATK	General variant calling tool.
MuTect	Variant caller specific for tumor samples.
VarScan	General variant calling tool.
Gene expression	
Cufflinks/Cuffdiff	A set of transcriptome quantification and differential expression analysis tools.
edgeR	Differential expression analysis tool.
Annotation	
Annovar	General genomic annotation tools.
Oncotator	Cancer specific genomic annotation too.
Visualization	
IGV	General purpose NGS data visualization.
Circos	General purpose NGS data visualization in circular format.

While most of bioinformatics tools available are mature enough for the complex cancer genome analysis, they still need computational expertise for installation, running analysis and interpretation of results because of computational complexity involved in command-line operation in UNIX/Linux environment. As an alternative to command-line interface (CLI), easy to use user interface can be designed for software (I. Figure 3) to make it easily accessible by non-computational experts. Online web-interface based cancer genomic analysis tools such as Galaxy [45] provide easy to learn user interface with point and click operation.



I. Figure 3: Types of user interface for bioinformatics tools.

Web-interface based bioinformatics tools run on a remote computer where user needs to first upload and store data in the remote server via internet. Unequal availability of internet bandwidth to all the users globally results into limitations of web-interface based

bioinformatics servers for analysis of large scale data sets. On the other hand, some offline computer based tools running on local computer have been used to develop graphical user interface (GUI) for easy operations using mouse point and click (I Figure 3). While several of tools for visualization of processed genomics data such as Integrative Genomics Viewer (IGV) [46], IGB and others [44] have GUI, majority of the core genomics analysis tools lacks such easy to use GUI. Given the availability of large scale open source genomics data sets through global genomics consortiums and individual research groups, there has been unmet need for various computational tools with easy to use user-interface and abstracted computational complexity for wider utilization of these data sets by biologists, clinicians and non-computational experts.

II. OBJECTIVES OF THE STUDY

2.1 Analysis of primary tumors for identification of therapeutically relevant alterations.

In this objective, I aim to identify therapeutically relevant alterations using genomic analysis of primary tumor samples followed by functional characterization of novel alterations using *in-vitro* and *in-vivo* methods.

2.2 Integrated genomic analysis and characterization of patient derived cell lines.

In the second part of this work, I focus on establishment of integrated genomics analysis workflow for identification of biologically relevant alterations using small set of patient derived cell lines.

2.3 Development of bioinformatics tool to analyze high throughput genomic data.

Followed by primary tumor and cell line analysis, I focus on development of bioinformatics tool for easy utilization of high-throughput next-generation sequencing data sets by biologists.

III. MUTATIONAL SPECTRUM OF ACTIONABLE ALTERATIONS IN LUNG ADENOCARCINOMA REVEALS NOVEL RECURRENT DRUG SENSITIVE *FGFR3* MUTATIONS

(Accepted for publication in *Annals of Oncology*)

3.1 INTRODUCTION

Lung cancer is the leading cause of cancer-related deaths worldwide, accounting for over a million deaths annually [47]. Non-small cell lung cancer (NSCLC) is the most common type of lung cancer. Of the three major histological subgroups of NSCLC, lung adenocarcinoma occurs most frequently. Despite advances in conventional therapies, the 5-year survival rate of lung cancer remain significantly low (16%) [48]. Molecularly targeted therapies instead indicate an improved outcome for lung adenocarcinoma patients whose tumors harbor mutant *EGFR* or translocated *ALK, RET,* or *ROS1* [15, 49-53], with an encouraging response for those with mutated *BRAF*, and *ERBB2* [54, 55]. However, lung adenocarcinomas patients those who do not harbor known somatic actionable driver oncogenic alterations are treated with conventional chemotherapy, underscoring an unmet need for additional therapeutic targets in lung adenocarcinoma patients.

Interestingly, the occurrence of oncogenic somatic alterations varies across populations/ethnic groups. For e.g., *EGFR* mutations are present in over 30% of East Asian lung adenocarcinoma patients, however, they are only found in about 23-25% of Indian and 10% of Western lung adenocarcinoma patients [15, 49, 56-58]. Similarly, *KRAS* mutations are present at 60% lower frequency in Indian lung adenocarcinoma patients than compared to the Caucasian population [50, 55, 59]. Such diversity in

somatic alterations lends similarity to the known plurality in clinical response based on ethnicity and the divergent genetic and environmental factors [60], Thus, besides the unmet need for additional therapeutic targets in lung adenocarcinoma patients, it is equally pertinent to profile known oncogenic somatic alterations across different populations to understand their landscape of variability.

Here, in an attempt to profile for activating alterations, we have generated a comprehensive mutational spectrum of activating alterations prevalent among lung adenocarcinoma patients of Indian origin, considered to be an ad-mixture of at-least five ancestral sub-populations [61-63]. We also report the first incidence of activating and drug sensitive *FGFR3* mutations in lung adenocarcinoma. *FGFR3* mutated samples, with ~5% population frequency, form a distinct subclass apart from *EGFR*, *KRAS*, and *EML4-ALK*.

3.2 MATERIALS & METHODS

3.2.1 Sample processing

FFPE blocks for 45 consecutive lung adenocarcinoma patients tumor sample for sequencing and an additional set of 363 consecutive lung adenocarcinoma patients tumor sample for mass spectrometry were retrospectively collected from Tata Memorial Hospital, where an adequate amount of acceptable quality of genomic DNA was available. The histological diagnosis of adenocarcinoma or squamous cell carcinomas were made based on immunohistochemistry staining using antibodies against TTF1, TP53, Napsin A, and CK 5/6. As a routine practice, 2 or more antibodies were used to distinguish adenocarcinoma from squamous carcinoma. Smoking history was recorded by directly asking a specific question to all the patients at the Medical Oncology

Department, Tata Memorial Hospital. The Institutional Review Board (IRB) and the Ethics Committee (EC) of Tata Memorial Center (TMC)- Advanced Centre for Treatment, Research and Education in Cancer (ACTREC) (Mumbai, India) approved the project (# 55 and 108) during the 21st TMC-ACTREC IRB meeting. Since this was a retrospective analysis, the IRB and the EC waived the need for an informed consent. The patient characteristics including the age, gender, smoking/tobacco use, and histopathology were recorded. The EML4-ALK fusion test, performed for molecular diagnosis of the patients at TMH, was recorded for the patients wherever available.

3.2.2 Pooling of samples, target gene-capturing and next generation sequencing

A set of 45 lung adenocarcinoma samples with known *EGFR* mutation status were divided into duplicate pools of different population size i.e. 2 pools of 5 individuals (5XA and 5XB), pools of 10 individuals (10XA and 10XB), and, 1 pool of 15 individuals (15X). 400 ng of each sample in the 5X pool; 200 ng of each sample in the 10X pool and 133 ng of each sample in the 15X to make the respective pools (Additional Figure S1). All 5 pools were submitted to RainDance Technologies Inc., USA to capture 676 genomic regions of 158 genes (127 KB of DNA) using RainDance Cancer panel, as described earlier [64]. Briefly, the genomic DNA (2 µg from each pool) was emulsified along with PCR mix, DNA Polymerase, dNTPs and reaction buffer, using RDT 1000 system. Amplification products were recovered by breaking the emulsion followed by secondary PCR amplification to incorporate Illumina adaptor sequences and multiplex sequences. These amplicons were then submitted to Sandor proteomics (Hyderabad, India) for sequencing using paired-end chemistry using two lanes of lllumina flow cells on GA-IIx giving expected sequencing coverage of more than 1,500X per base.

3.2.3 Discovery of genomic variants using computational analysis

A computer program was written in Perl for de-multiplexing and conversion from QSEQ to FASTQ formatted raw sequencing data. Undetermined reads due to degenerate barcode were put together and labeled unknown pool. The resulting FASTQ files for each pool (5XA, 5XB, 10XA, 10XB, 15XB and Unknown pool) were fed into mapping/alignment program BWA [65] for mapping onto reference human genome sequence GRCh37. On average, 6.3 million reads mapped to the genome, and of the uniquely mapping reads, 94% (range 75-95%, Additional Table S2) mapped to the amplified region, demonstrating the high specificity of this approach corresponding to an enrichment factor of 19,508 [range: 9172 – 22197; (sequenced base-pairs in ROI/total sequenced base-pairs)/(size of ROI/size of the genome)] (Additional Table S2). With a strict coding-sequence target definition, 94% of the reads mapped on target. Furthermore, 66 % of the targeted bases were covered at >10% of the mean reads per million (range: 32-78%, mean reads per million = 1500, Additional Figure S2). Further SAM/BAM files were prepared for variant calling using Picard toolkit (https://broadinstitute.github.io/picard/). Resulting BAM files were fed into GATK [66] and Mutect [67] for variant calling to generate median 837 coding variants per pool (range: 756 – 2145) representing total 3349 unique variants (Additional Table S2).

3.2.4 Filtering of genomic variants

We combined variant calls from Mutect and GATK and loaded into MySQL for further analysis. Mutational signature of FFPE tissues [68], as described earlier, were observed by 60% representation of C:G/T:A mutations (2340 out of 3349 total; Figure S3), majority of which were identified at coverage lower than 15x and allele frequency lower than 0.05 (2184 out of 2340 total; Figure S3), hence were removed from the dataset. Pool 5XB harbored two times the variants than median (2145 in 5XB; median 973) which

might be due to the presence of one or more sample with hyper mutator phenotype and/or poor quality FFPE blocs because more than half of the C:G/T:A variants were identified in pool 5XB (1423 out of total 2340). Removal of these artifacts resulted into 1288 unique variants across all pools. To further prioritize cancer-related variants, we subtracted known SNPs overlapping with dbSNP database [69] (v.142) and TMC-SNPdb database, an in-house developed database representing Indian ethnicity-specific SNPs identified using whole exome sequencing [70] (Figure S4). We also used 9 functional prediction tools SIFT, Polyphen2-HDIV, Polyphen2-HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, RadialSVM, and LR retrieved through dbNSFP [71] to further prioritize cancer related variants called as deleterious by at-least 7 tools. This prioritization resulted into identification of total 99 variants in 26 genes (Figure S4 and S5).

3.2.5 Single base extension based mass spectrometry based genotyping

DNA from 363 samples were submitted to AceProbe technologies for mass spectrometry following company's standard protocol. Briefly, PCR and extension primers for 49 mutations in 23 genes were designed using single base extension based mass spectrometry assay design 3.1 software (Appendix 2). Mutation calls were analyzed using Typer 4 (Sequenom Inc., USA) and were reviewed by manually observing mass-spectra.

3.2.6 Cell culture, reagents, transfection, and infection

Pre-authenticated NIH/3T3 cells were obtained from ATCC (CRL-1658) and used within 6 months of thawing. The cells were maintained in DMEM (Gibco/ Invitrogen) supplemented with 10% fetal calf serum (Gibco/ Invitrogen) and 2% antibiotics penicillin/streptomycin (Gibco/Invitrogen). *FGFR3* was cloned into pBABE-puro from pDONR223 (was a gift from William Hahn & David Root addgene plasmid # 23933) for

retroviral production [72]. *FGFR3* mutants were generated by site-directed mutagenesis using the Quikchange II kit (cat.no. 200523) and confirmed by Sanger sequencing. Replication-incompetent retroviruses were produced from 5µg pBABE-puro vectors (with pCL-ECO as helper vector) by transfection into HEK 293T packaging cell line by using Lipofectamine 2000 (Invitrogen). NIH 3T3 cells were infected with these retroviruses in the presence of 8 µg/ml polybrene. Infected cells were puromycin (Sigma) selected (2 µg/ml) after two days of infection. To induce cells with FGF1, cells were placed in media containing 0.5% calf serum 12 hours before 50 ng/ml FGF1 (Abcam ab91374) stimulation for 20 min at 37°C. PD173074 was purchased from Calbiochem and diluted in DMSO to the indicated concentrations. BGJ-398 was purchased from Santa Cruz biotechnology and diluted in 10% tween-80.

3.2.7 Anchorage-independent growth assay

Anchorage independent growth assay was performed as described earlier [73]. Briefly, independent set of 5 x 10^3 and 20 x 10^3 cells were suspended in a layer of DMEM containing 10% fetal calf serum and 0.4% select agar (Gibco/Invitrogen) and plated on a bottom layer of DMEM containing 10% fetal calf serum and 0.8% select agar. Each *FGFR3* clones were analyzed in triplicates. PD173074 was added at described concentration to the top agar. Agar plates were photographed after 3 weeks of incubation. Colonies were counted in triplicate wells from 9 fields photographed with 10x objective using phase contrast inverted microscope (Zeiss axiovert 200 m). IC_{50} was determined by nonlinear regression with Prism GraphPad software and Dr Fit tool [74].

3.2.8 Immunoblotting

Cells were lysed in RIPA buffer and protein estimation was done by BCA method [75]. Lysates were boiled in sample buffer, separated by SDS-PAGE on 8% or 10% polyacrylamide gels, transferred to PVDF membrane, and probed as described previously

[73]. Primary antibodies used for immunoblotting were: anti-FGFR3 (Santa Cruz Biotechnology; 1:500), anti- total-ERK1/2 (Santacruz Biotechnology; 1:200), phospho-ERK1/2 (Cell signaling Technology; 1:1000). Secondary antibodies used were- Goat anti-rabbit (Santacruz Biotechnology; 1:2000) and Bovine anti-mouse (Santacruz Biotechnology; 1:2000). Pierce ECL (Thermo Scientific, USA) substrate was used for visualizing the blots.

3.2.9 Xenograft development

The study was approved by the Institutional Animal Ethics Committee of ACTREC, Navi Mumbai which is endorsed by the 'Committee for the Purpose of Control and Supervision of Experiments on Animals', Government of India for the use of animals vide approval no. 06/2014. For this study 8-weeks old NOD SCID mice were procured from the Animal Facility of ACTREC, Navi Mumbai. A cohort of 8 NOD-SCID mice per clone were subcutaneously injected with 5 million cells. All mice were observed for tumor formation in 2-3 months. Tumors were collected to subcutaneously graft ~2-3 mm tumor piece in further expanded set of mice. Inhibitor BGJ-398 was given at 15 and 30 mg/kg along with vehicle control (10% tween-80) independently to randomized xenograft groups after tumor size reaching ~150 mm^3. Tumor size was measured every alternate day using Vernier caliper during 14 days' drug treatment. At the end of the treatment, mice were starved for 6-8 hours before microPET-CT scan for measuring ^{18}F-FDG (fluorodeoxyglucose) tumor uptake. Micro PET images were acquired one hour post intravenous administration of ~ 18 to 20 MBq ^{18}F-FDG via tail vein using Triumph trimodality imaging system (TriFoil Imaging Inc., Northridge, CA, USA). The PET data were reconstructed using a maximum likelihood expectation maximization (MLEM) two-dimensional algorithm (30 iterations) and CT data were reconstructed using filter

back projection. The reconstructed PET image data was analyzed, co-registered with CT and visualized using PMOD 3.310. (Figure 3C)

3.2.10 Tissue processing and Immunohistochemistry

Tumors obtained from mouse were formalin fixed and paraffin embedded for long term storage. FFPE blocks were sectioned at 4μm for H and E staining for evaluation of tumor. Paraffin sections were baked at 58 ^{o}C for 30 minutes before staining, followed by deparaffinization, rehydration, and quenching. Heat-induced antigen retrieval was performed in a pressure cooker (1 whistle) with Citrate Buffer (pH 5.9-6.1) followed by cooling at room temperature. Blocking performed with 1:50 horse serum provided in Vectashield Kit (Vectashield, Vector laboratories). Tissue sections were incubated overnight at room temperature with anti- total-ERK1/2 (Santacruz Biotechnology) and phospho-ERK1/2 (Cell signaling Technology) in PBS. Slides were rinsed in wash buffer and incubated for 30 minutes with biotinylated secondary antibody provided in the kit, followed by tertiary antibody (Reagent A & B in the kit) incubation for 1 hour. The chromogenic reaction was developed with 3,3′-diaminobenzidine chromogen solution for 5 minutes, resulting in the expected brown-colored signal. Further, sections were counter stained with haematoxylin and covered with DPX mounting reagent and cover slip. Imaging was performed at 10X & 20X magnification using an upright microscope.

3.2.11 Overall survival analysis

Overall survival of patients was assessed using Kaplan-Meier method via R packages survival [76], survMisc [77] and IBM SPSS software package. The end point was taken as death. Patients with unknown date of death were censored at the date of the last contact. Overall survival of patients from TCGA cohort was estimated by querying cBioPortal [78] for cancer types indicated.

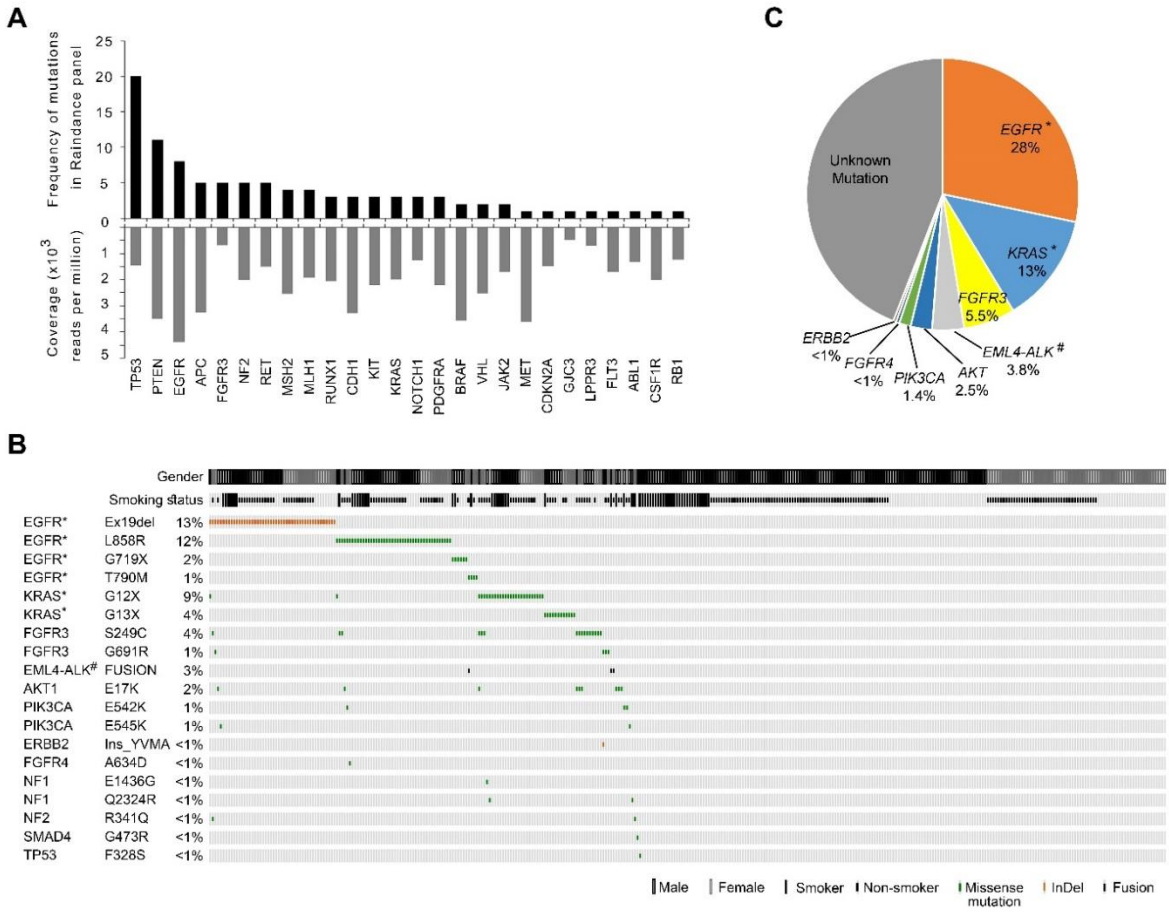
3.3 RESULTS

3.3.1 Recurrent *FGFR3* mutations in lung adenocarcinoma patient of Indian origin

To profile for therapeutically relevant genome alterations in lung adenocarcinoma of Indian origin (an ad-mixture of at-least five ancestral sub-populations), we analyzed 45 formalin fixed paraffin embedded (FFPE) primary lung adenocarcinoma stage IV tumors (Table S1) using RainDance cancer panel for 676 amplicons derived from 158 genes in a pooled format, as described in Figure S1. The library generated were sequenced at an average coverage of ~1500 reads per million mapped reads (Table S2). Ninety-nine deleterious mutations representing 26 genes were found to be recurrently mutated: *TP53*, *EGFR*, *KRAS*, *PTEN*, *CDKN2A*, *BRAF*, *NF2*, *JAK2*, etc. along with those previously not known to be significantly mutated in lung adenocarcinomas such as *PDGFRA*, *KIT*, *FGFR3*, and *RET* among the Caucasian population [53, 59, 79-82] (see Figure 1A; S2-S5; Appendix 1), after correcting for FFPE artifacts [68] as detailed in materials and methods. The confounding germline variants were depleted against the dbSNP (v.146) database and the Indian germline SNP database TMC-SNPdb as reported earlier by our group [70]. Among the genes not known to be significantly implicated in lung adenocarcinoma, 3 recurrent mutations predicted as deleterious were observed in *FGFR3* gene-- R248C, S249C, and G691R [59, 82, 83]. Next, to validate these findings using an orthologous technology, 49 mutations occurring across 23 genes as described in Appendix 2 were genotyped in an additional set of 363 primary lung adenocarcinoma stage IV tumors (Appendix 3) using single base extension (SBE) mass spectrometry.

Based on the mutation profiling across 363 lung adenocarcinoma patients, we present the first spectrum of activating mutations present in the Indian lung cancer genome (Figure 1B), wherein 160 of 363 patients were found to harbor activating mutations across 8

genes at following frequency: *EGFR* (28.4%), *KRAS* (13%), *ALK* (3.8%), *AKT1* (2.5%), *PIK3CA* (1.4%), *FGFR4* (0.4%), and *ERBB2* (0.3%) as shown in Figure 1C. Ethnic-specific variability was observed mainly in *AKT1*, *PIK3CA*, *FGFR4*, and *ERBB2* that were altered at frequencies distinct from Caucasian lung cancer patients, as reported



III. Figure 1: Recurrently mutated genes in lung adenocarcinoma.

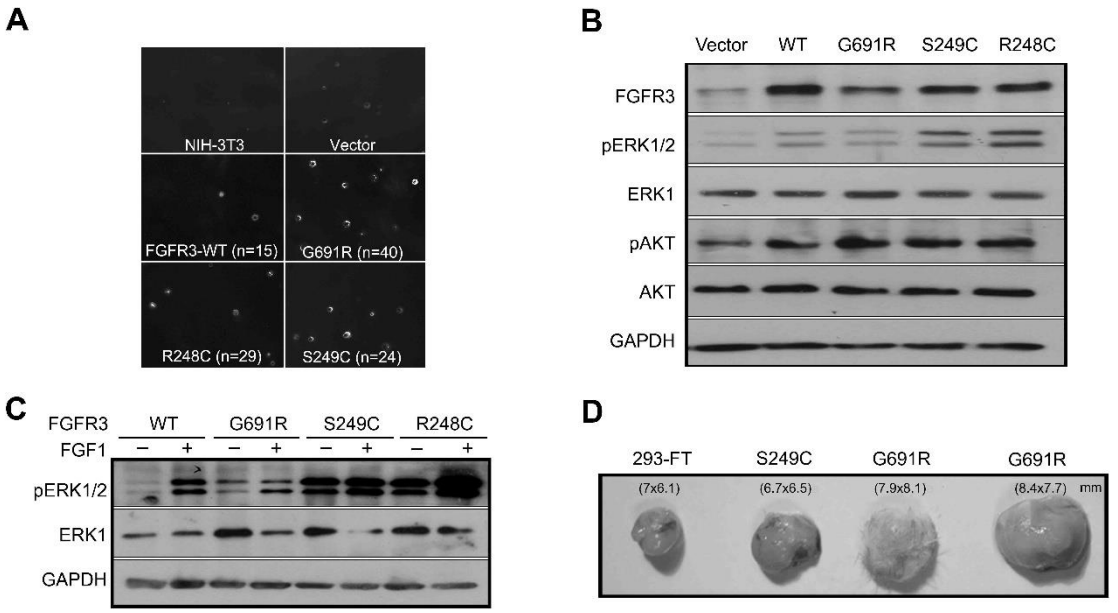
(A) Recurrently mutated genes identified in 45 lung adenocarcinomas using microfluidics-based high-throughput targeted sequencing of 158 cancer-related genes are shown. Top bar plot represents a percent frequency of mutations in RainDance panel sequencing and bottom bar plot represents sequencing coverage on Illumina platform as reads per million mapped reads. (B) Validated mutations in 363 samples identified by single base extension based mass spectrometry are visualized using OncoPrinter tool available at cBioPortal. Grey bar indicates each patient negative for respective mutation and or annotation in the rows. Missense mutations (green), small deletions (brown), and

fusion events (black) are indicated for respective genes/patients. Smoking status (smokers: black box, non-smokers: half black box, grey: information not available), gender (male: black outline, female: dark grey outline) are indicated in top annotation track wherein light grey box represents unavailability of data. The asterisk (*) denotes that genes genotyped using TaqMan and SNaPShot assays in addition to single base extension based mass spectrometry. # Fusion frequency was determined using fluorescent in-situ hybridization in only 79 patients out of 363 total. (C) Pie-chart representation of the frequency of clinically relevant genes observed in 363 Indian lung adenocarcinomas.

earlier for other mutations [56, 84]. In addition to the mass spectrometry-based genotyping, *ALK* fusions were determined for 79 samples by FISH. Three patients were found to harbor *EML4-ALK* translocation (3.8%). Among the other most significantly mutated genes, we found recurrent *FGFR3* mutations in 20 of 363 tumors (5.5%). Sixteen patients harbored *FGFR3* (S249C); and, a novel *FGFR3* (G691R) mutation in 4 patients (Figure 1C; Figure S6). Interestingly, identical germline *FGFR3* (R248C) and *FGFR3* (S249C) mutations are known to be associated with autosomal dominant Thanatophoric dysplasia syndrome [85, 86], while the somatic occurrences of these mutations have been shown to be involved in premalignant conditions of Seborrheic keratosis of skin and in cervical, urothelial and lung squamous carcinoma [86-90]. The somatic status of *FGFR3* mutations in this study couldn't be confirmed due to the non-availability of paired normal samples, though no records of congenital disorder among patients were found.

3.3.2 *FGFR3* mutations in lung adenocarcinoma are activating *in-vitro* & *in-vivo*

To test whether the novel and recurrent *FGFR3* (G691R) mutations found in this study are activating we transduced NIH 3T3 fibroblast cells with retroviruses encoding the



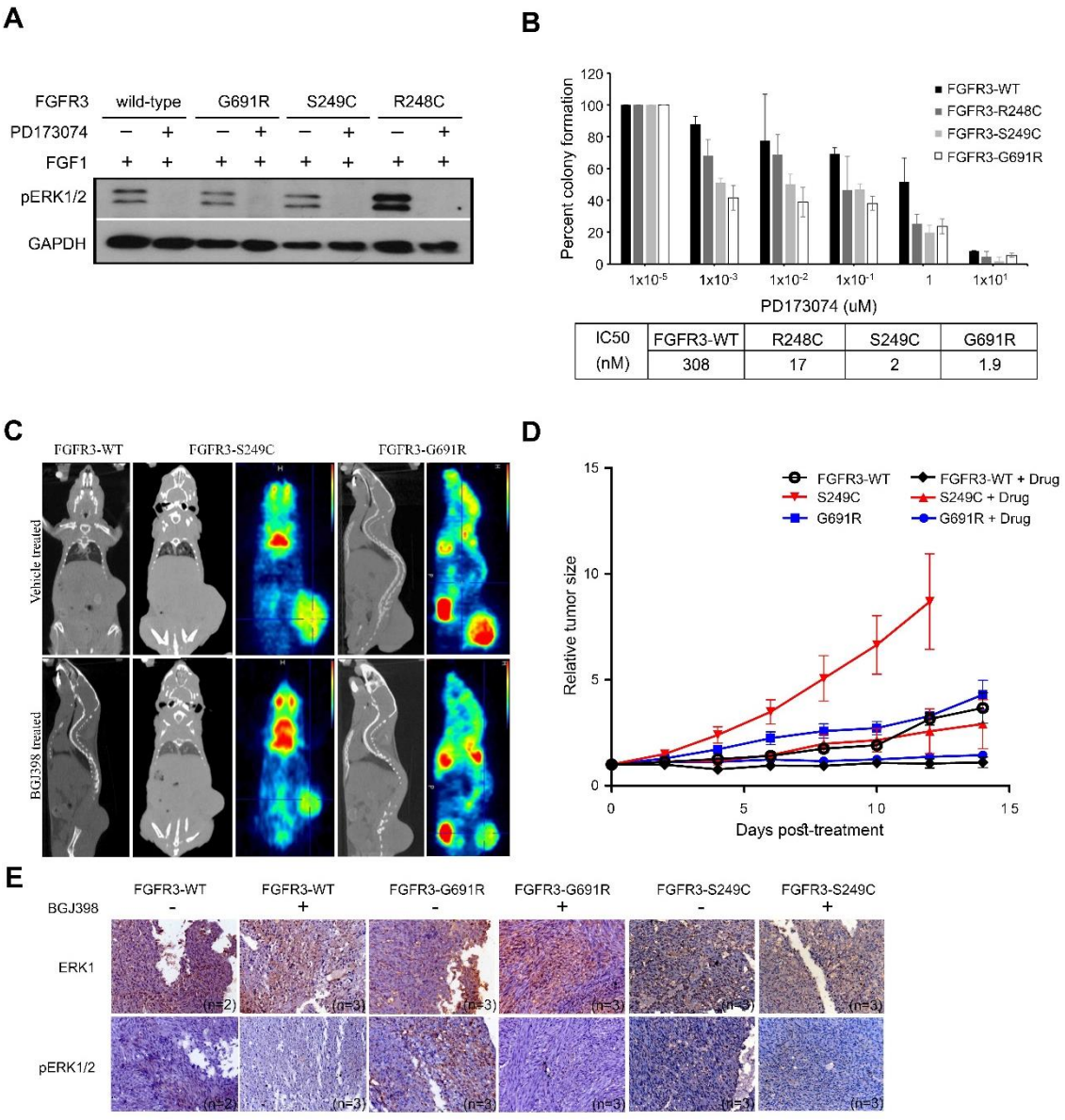
III. Figure 2: *FGFR3* mutations transforms NIH 3T3 and forms tumors in xenografts.

(A) NIH 3T3 cells stably expressing wild-type or mutant *FGFR3*, as indicated, were suspended in soft agar for colony formation assay and photographed after 3 weeks of incubation at 37^{O}C. Representative pictures and colony count (averaged from triplicate) are shown for NIH 3T3 clones as indicated. (B) Immunoblot analysis of NIH 3T3 clones for anti- FGFR3, total- and phospho- ERK1/2 and AKT. GAPDH was used as loading control. Clones expressing mutant *FGFR3* show over activation of the downstream pathway by hyper phosphorylation of ERK1/2 and AKT compared to wild-type. (C) Immunoblot analysis of NIH 3T3 clones with and without ligand (50 ng/ml FGF1) treatment for total- and phospho- ERK1/2 demonstrating ligand-independent activation of signaling pathway. GAPDH was used as loading control. (D) NIH 3T3 cells were subcutaneously injected into NOD-SCID mice for tumor formation to be observed in 2 months. The relative size of tumor xenografts is shown along with tumor size (mm) indicated in brackets.

FGFR3 G691R mutation along with WT *FGFR3* and the previously characterized *FGFR3* (R248C) and (S249C) mutations [89]. Similar to *FGFR3* R248C and S249C, the ectopic expression of the novel G691R mutant clone in pooled NIH 3T3 cells conferred anchorage-independent growth, forming 3-fold more colonies in soft agar than cells expressing WT *FGFR3* (Figure 2A), despite higher expression levels of WT *FGFR3* (Figure 2B). The transformation was accompanied by elevated phosphorylation of the downstream ERK1/2 and AKT1 in a constitutive manner (Figure 2C). Further, consistent with the *in vitro* data, NIH 3T3 cells expressing transforming *FGFR3* mutations or WT when injected subcutaneously into NOD/SCID mice formed tumors within 2 months post injection of cells (Figure 2d). While 3 of 11 mice injected with cells expressing *FGFR3* WT formed tumors, 12 of 12 mice injected with cells expressing *FGFR3* S249C; and, 6 of 12 mice injected with cells expressing *FGFR3* G691R formed tumors (Table S3). The tumor size doubling time was ~7 days for cells expressing *FGFR3* G691R, ~5 days for cells expressing *FGFR3* S249C; the *FGFR3*-WT tumors doubled in size in ~9-10 days.

3.3.3 *FGFR3* mutations in lung adenocarcinoma are sensitive to inhibitors *in-vitro* & *in-vivo*

After confirming that the lung adenocarcinoma patient-derived *FGFR3* mutations drive anchorage-independent growth in NIH 3T3 cells, we then investigated whether inhibition of FGFR3 kinase activity using pan FGFR inhibitor could block the transformation. NIH 3T3 cells were seeded into soft agar in the presence or absence of pan FGFR inhibitor PD173074. Treatment with the pan FGFR inhibitor PD173074 abrogated the phosphorylation of ERK1/2, which was constitutively phosphorylated in the NIH 3T3 cells expressing activating *FGFR3* mutations (Figure 3A). Similarly, treatment of cells harboring activating *FGFR3* mutations with a pan FGFR inhibitor PD173074 in cells



III. Figure 3: Transformed NIH 3T3 cells and xenografts are sensitive to FGFR inhibitor.

(A) Immunoblot analysis of NIH 3T3 clones treated with FGF1 (50 ng//ml) followed by FGFR inhibitor PD173074 (2μM) is shown indicating blockage of signaling pathway (phospho-ERK1/2) in drug-treated cells. GAPDH was used as loading control. (B) NIH 3T3 clones expressing wild-type or mutant *FGFR3*, as indicated, were suspended in soft-agar plates with increasing concentration of PD173074. Quantification was averaged from 3 replicates. IC-50 calculations were carried out by using non-linear regression in GraphPad software. (C) NIH 3T3 clones were subcutaneously injected into NOD-SCID

for tumor formation in ~2 months. Selective FGFR inhibitor BGJ-398 or vehicle treatment was administered orally in mice after tumor size reaching ~100-200 mm^3. ^{18}F-FDG uptake was studied in these tumors after 21 days of drug treatment. A readout for relative ^{18}F-FDG uptake is shown by a gradient color code with red indicating as maximum uptake. (D) Tumor size was also measured every alternate day using Vernier caliper in each xenograft of treatment and control arm. The plot shows tumor size (normalized to the size at day 0 of drug treatment) during the course of drug treatment indicating a reduced tumor size in drug-treated mice. (E) Immuno-histochemical staining of total- and phospho- ERK1/2 is shown in xenografts treated with drug and vehicles.

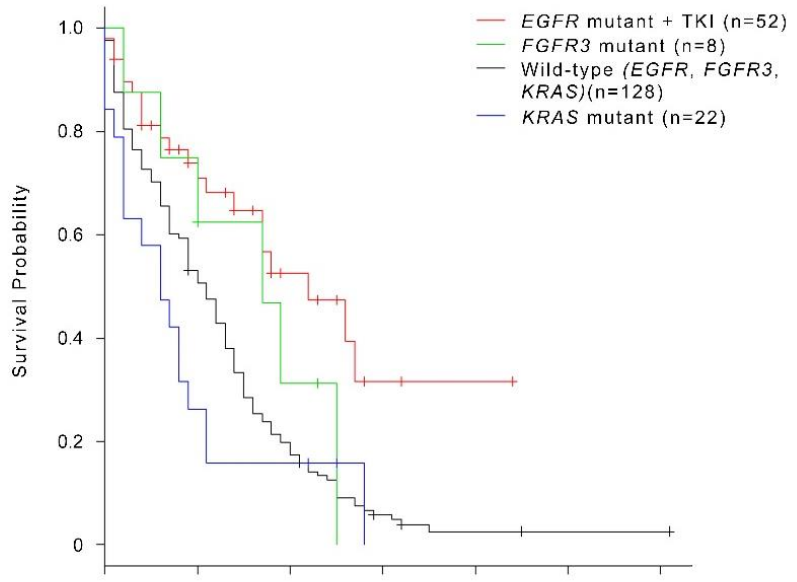
resulted in a marked decrease in colony formation in soft agar and cell survival in liquid culture (Figure 3B). *FGFR3* R248C and S249C mutations lost colony-forming potential when exposed to 17 and 2 nM of drug respectively, consistent with the previous reports [89], whereas cells expressing the novel kinase domain *FGFR3* G91R mutation lost colony-forming potential at 1.9 nM of the drug. Extending the effect *in vivo* studies, when tumors reached approximately 100-200 mm^3 in all mice injected with NIH 3T3 cells began treatment with 15 or 30 mg/kg pan-FGFR inhibitor, BGJ398 [91], or vehicle for 14 days. Tumors treated with BGJ398 slowed or reversed their growth compared with vehicle (Figure 3C), so that by the end of the study, the effect on tumor burden in vehicle-treated versus BGJ398-treated mice were noticeably distinct, 3.3-folds in *FGFR3* (S249C), 3-folds *FGFR3* (G691R) and 2.25-folds in FGFR3 (WT) (Figure 3D). This reduction in tumor size was paralleled by a reduction in the amounts of phospho-ERK1/2 in immuno-histochemical analyses (Figure 3E) of explanted tumors, validating our *in vitro* findings that MAPK signaling is the key pathway engaged by mutated *FGFR3*.

3.3.4 Correlation of *FGFR3* mutations with clinicopathological features of lung cancer patients

FGFR3 mutations were observed to be significantly higher in patients < 45 years (9 of 95) than in patients > 45 years (11 of 269) (P = 0.048, Table S4), consistent with previous report of higher oncogenic mutations in younger non-small cell lung cancer patients [92] including *EGFR* and *KRAS* mutations [92, 93] . Other features like gender and smoking status were not significantly associated with patient survival (Table S4).

Additionally, survival analyses were performed and compared between different groups with data available from 205 of 363 patients using a log-rank test and visualized with Kaplan-Meier plots. As shown in Figure 4, median overall survival (OS) of patients with *FGFR3* mutant NSCLC was 17 months (n= 8; 95% CI: 6.4-27.5; HR: 0.6), as compared to 14 months (n= 197; 95% CI: 8.7-13.2) in patients with wild-type *FGFR3*. Among the patients with wild-type *FGFR3*, median OS of patients with *EGFR* mutation having received TKI was 22 months (n= 53; 95% CI: 18.4-30.3; HR: 0.4), consistent with literature [49]; and, with *KRAS* mutation was 6 months (n= 22; 95% CI: 5.1-18.9; HR: 1.3). In a nutshell, though statistically underpowered, based on the Kaplan-Meier analysis, the patients with *FGFR3* mutations show a trend towards better overall survival than those with wild-type *FGFR3* (n= 8; log rank P = 0.53).

We also queried the cBioPortal [78] for survival data of patients harboring activating *FGFR3* mutations in different cancers [78, 94] (Figure S7). Based on the Kaplan–Meier analysis of TCGA dataset, of 279 head & neck cancer and 178 lung squamous patients with *FGFR3* mutations had significantly shorter overall survival than those with wild-type *FGFR3*, as shown in Figure S7B and D). However, of the 130 bladder urothelial carcinomas and 343 skin cutaneous melanoma patients harboring *FGFR3* mutations had better survival than wild-type patients (Figure S7A and C), consistent with our study.



Months	0	10	20	30	40	50	60	Median survival
EGFR+TKI	52	27	11	3	2	0	0	22
FGFR3	8	5	3	1	0	0	0	17
Wild-type	128	64	24	6	3	1	1	11
KRAS	22	6	3	0	0	0	0	6

III. Figure 4: Kaplan-Meier overall survival plot of lung adenocarcinoma patient with different mutation status.

Clinical follow-up of total 205 patients for up to 62 months was used for Kaplan-Meier analysis. *EGFR* positive patients received Gefitinib as a regular therapeutic regimen while rest of the patients were received conventional chemotherapy. The table below the plot indicates patients at risk during the course of 60 months and median survival for each mutant group. Median OS of patients with *FGFR3* mutant NSCLC was 17 months (green; n= 8; 95% CI: 6.4-27.5; HR: 0.6), *EGFR* mutation having received TKI was 22 months (red; n= 53; 95% CI: 18.4-30.3; HR: 0.4), *KRAS* mutation was 6 months (blue; n= 22; 95% CI: 5.1-18.9; HR: 1.3) while *EGFR*, *KRAS* and *FGFR3* mutation negative patients was 11 months (black; 95% CI: 8.7-13.2).

3.4 DISCUSSION

We present the first spectrum of clinically actionable alterations in lung adenocarcinoma of Indian origin which includes *EGFR, KRAS, EML4-ALK, AKT1, PIK3CA, FGFR4* and *ERBB2*, similar to that identified in other ethnic groups [83, 95, 96], and an additional subset of patients with *FGFR3* mutations. Ethnic-specific variations have been well known in lung cancer [22, 97, 98] across different populations. We observed 28.4% *EGFR* mutations and 13% *KRAS* mutations in lung adenocarcinoma patients, consistent with our previous report [50, 56]. Similarly, variation in frequency of other molecular alterations is also observed such as 3% *EML4-ALK* alteration in our study compared to 8% in Caucasian population [55] and in 5% Chinese population [83]. *ERBB2* mutation found at <1% frequency in our cohort exists at ~2-3% among the Caucasian [55] and Chinese populations [83]. Similarly, *AKT1* mutations were found at higher than the reported <1% in both Caucasian [55] and Chinese populations [83] indicating the higher therapeutic relevance of *AKT1* targeted compounds in Indian population.

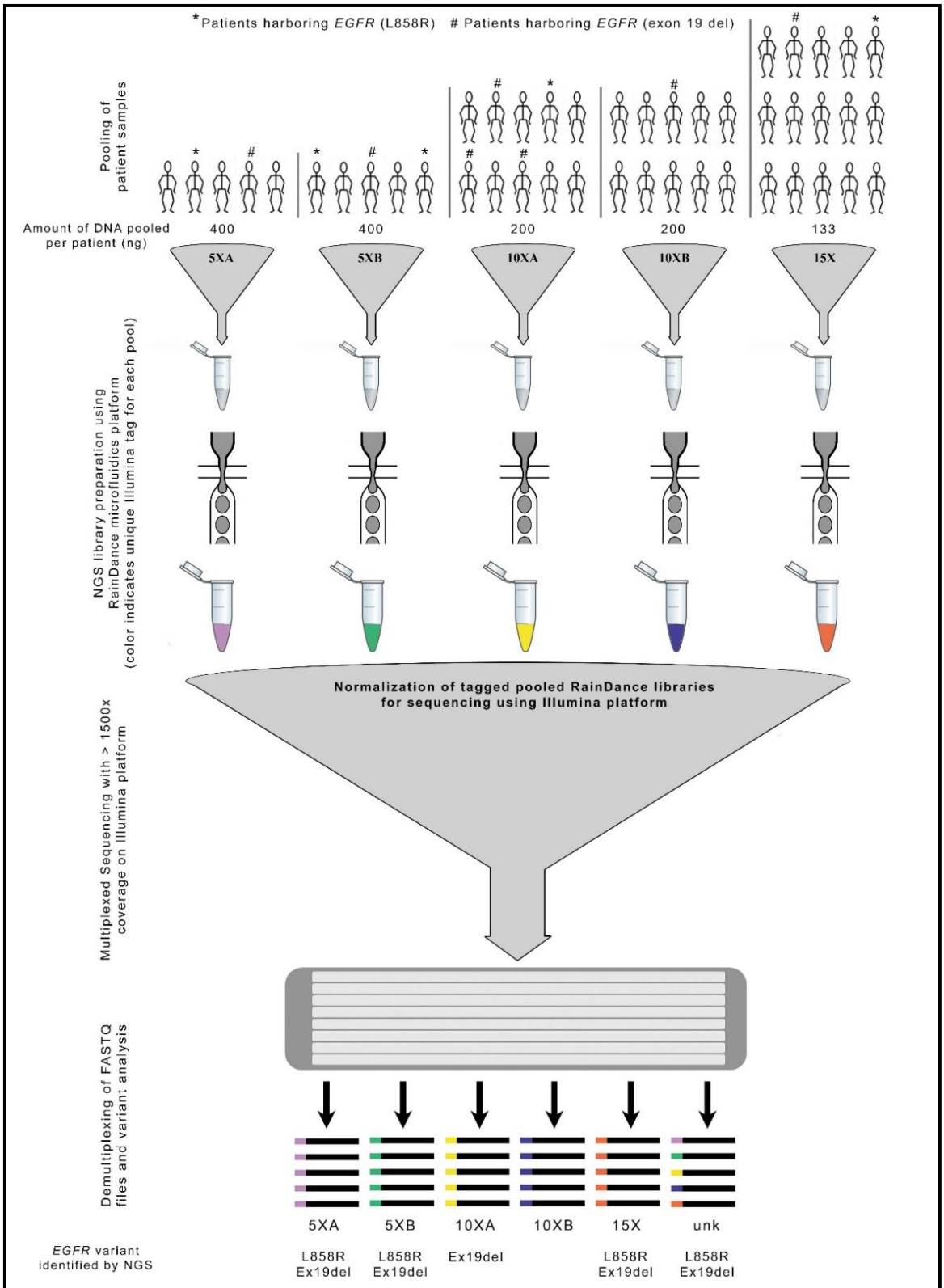
We have also identified frequent and recurrent drug sensitive *FGFR3* mutations in lung adenocarcinoma patients of Indian origin. Our study establishes that *FGFR3* extracellular and kinase domain mutations found in lung adenocarcinoma patients are oncogenic and sensitive to small molecule inhibitors *in vitro* and *in vivo* systems, using mouse xenograft assays.

Among the Caucasians, activating mutations in *FGFR3* have been earlier reported in bladder carcinoma [90], lung squamous cell carcinomas [89], and, cervical cancer [88], but were found to be largely absent in lung adenocarcinomas [81, 83, 99], except for Imielinski *et al.* who reported non-recurrent somatic *FGFR3* mutations of unknown

functional significance in 3 of 183 lung adenocarcinoma patients [59]. On the other hand, the presence of frequent *FGFR3* mutations is tangentially referred to in the literature among Korean lung adenocarcinomas patients [100]. Along with these few reports, our finding of activating *FGFR3* mutations in lung adenocarcinoma patients further provides an interesting convergence with several mouse genetic experiments from literature: mice expressing an activating kinase domain mutant of *FGFR3* (K644E) develop either skin or lung cancer [101]; activated FGF9-FGFR3 signal acts as the primary oncogenic pathway involved in initiation of lung adenocarcinoma [102, 103]; and, Inhibition of fibroblast growth factor receptor 3-dependent lung adenocarcinoma with a human monoclonal antibody as an effective treatment in mouse lung adenocarcinoma [102, 104]. Taken together, this suggests an emerging trend for a role *FGFR3* mutations in lung adenocarcinoma, more likely to be predominant among the Asian population.

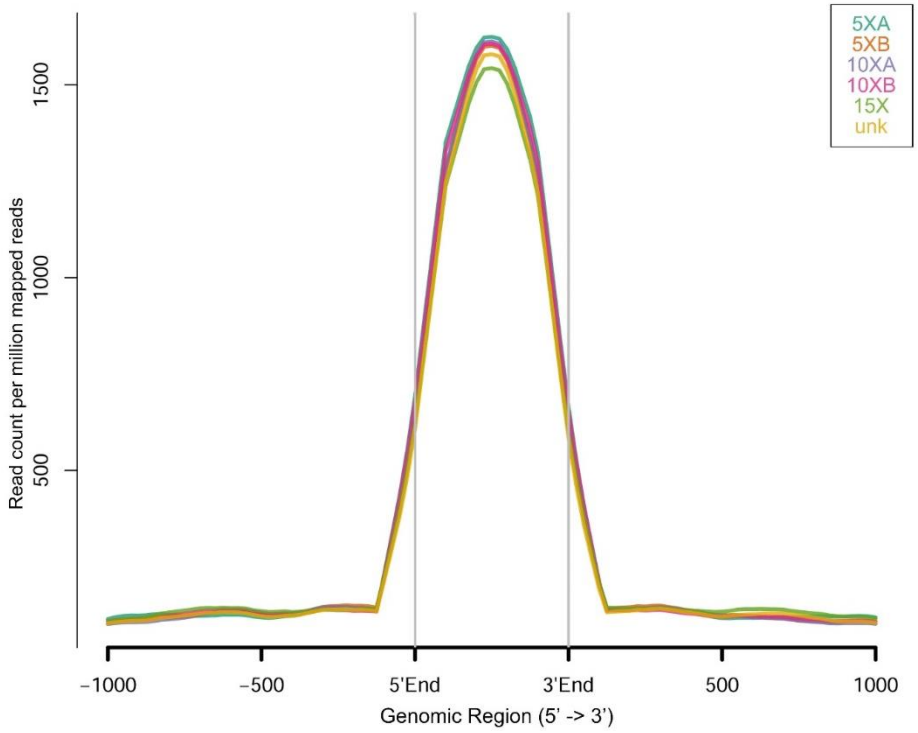
Analyzing the potential effect of *FGFR3* driver mutations on survival lung cancer patients, we observed a trend towards better survival for *FGFR3* mutations in lung adenocarcinoma, compared to lung adenocarcinoma patients with wild-type *FGFR3* and those harboring *KRAS* mutation, similar to as reported in the bladder and skin cancer [105]. Thus, *FGFR3* mutation represents an opportunity for targeted therapy in lung adenocarcinoma. FGFR inhibitors, which are currently in clinical testing in tumor types bearing genetic alterations in FGFR genes [106, 107], may be extended to evaluated in patients with *FGFR3*-mutated lung adenocarcinoma. Finally, with a broader emerging role across different cancers [89, 108-110], this study further underscores that *FGFR* family may potentially join the *EGFR* family as a widespread target for therapeutic intervention in several human cancers.

3.5 Additional Supporting Data



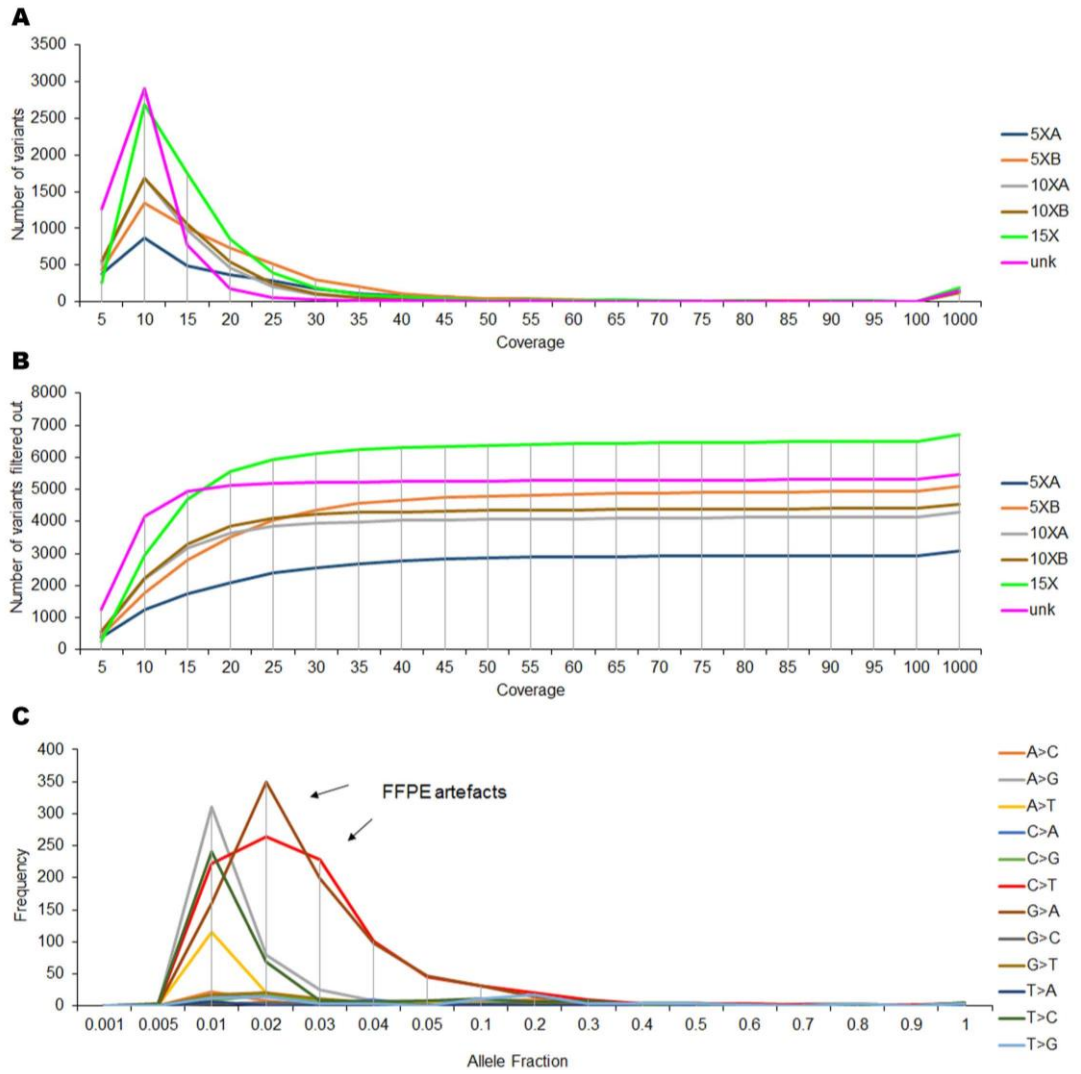
III. Additional Figure S1: Schematic diagram of pooled next-generation sequencing of 45 lung adenocarcinomas.

(A) Set of 45 lung adenocarcinoma sample with known EGFR mutation status were divided into duplicate pools of different population size i.e. 2 pools of 5 individuals (5XA and 5XB), 2 pools of 10 individuals (10XA and 10XB), and, 1 pool of 15 individuals (15X). 400 ng of each sample in 5X pool; 200 ng of each samples in the 10X pool and 133 ng of each sample in 15X to make the respective pools. All 5 pools were used to capture 676 genomic regions of 158 genes using RainDance Cancer panel that was unique tagged as shown. The tagged libraries were normalized and sequenced on Illumina GAIIx. The multiplexed sequencing data was de-multiplexed per the unique tags into five fastq files, each corresponding to five pools. Reads with unidentifiable or degenerate barcode sequences were put together in pool of unknown size labelled as "unk". The ability to make variant calls for EGFR within each pool served as the positive control that was found to be largely concordant.



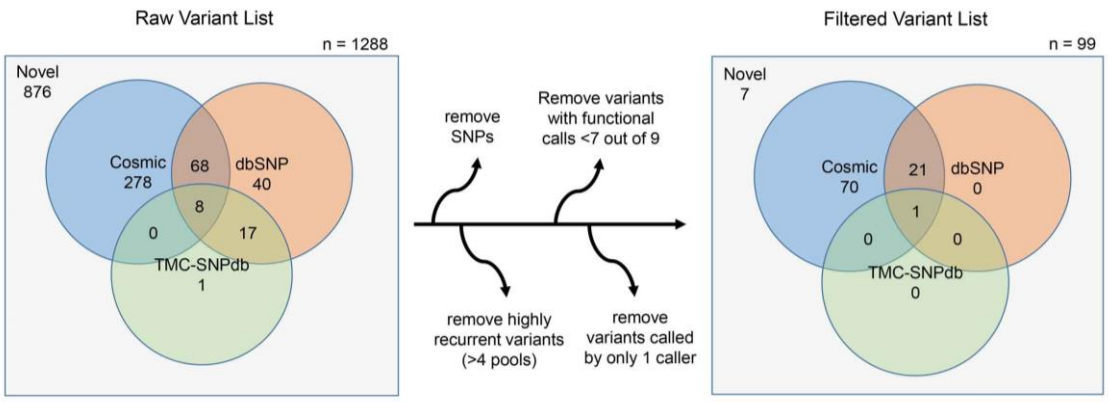
III. Additional Figure S2: Sequencing coverage of 158 target genes in RainDance panel.

Sequencing coverage of 676 target regions (± 1 KB) representing 158 genes in 45 lung adenocarcinoma samples pooled at variant sample size is represented as smoothened line plot indicating homogenous sequencing of each target region in each pool. 5XA & 5XB: pools of 5 individuals; 10XA & 10XB: a pool of 10 individuals; 15XB: a pool of 15 individuals; unk: sequencing reads not assigned to any of the above pools were kept in a pool called unknown (unk).



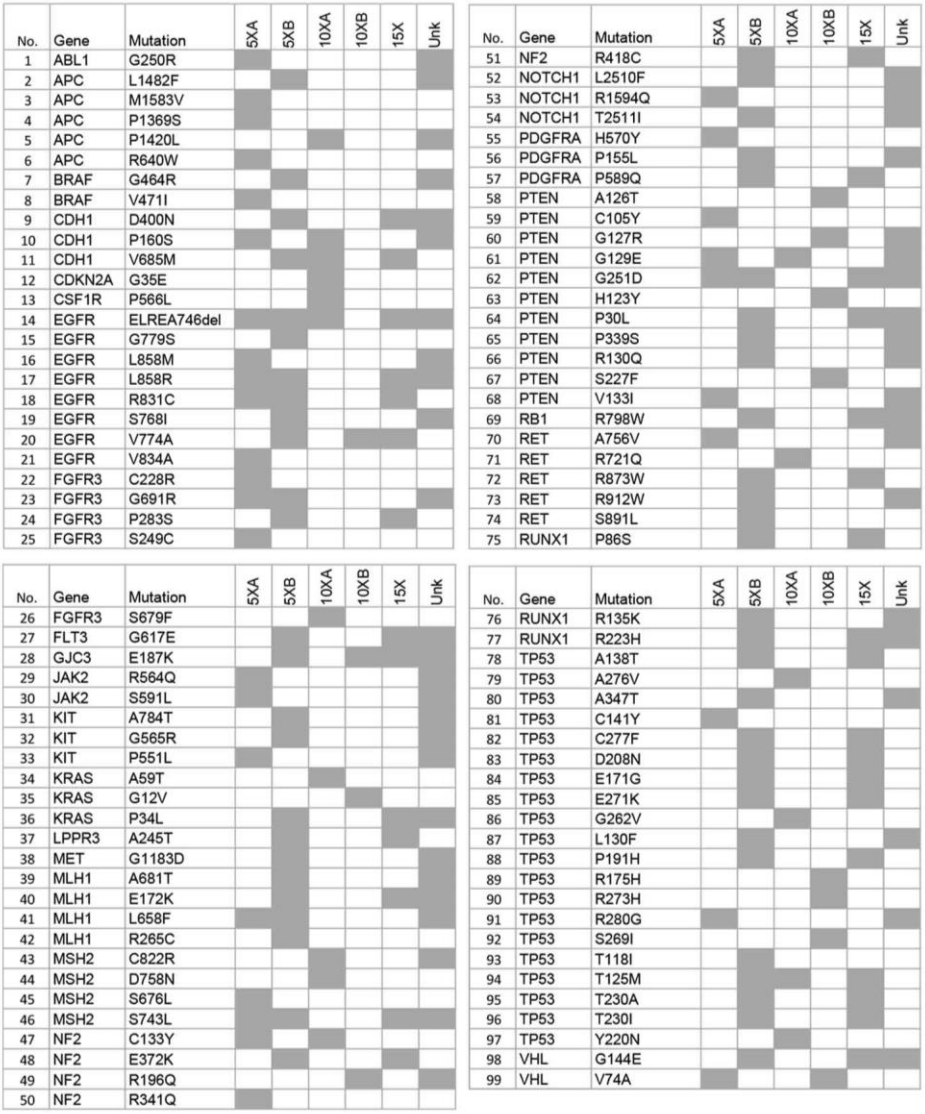
III. Additional Figure S3: Coverage and allele fraction distribution in high-throughput sequencing data.

(A) Variants called at various sequencing coverage in all pools indicating accumulation of probable false positive variant calls at coverage lower than 15x while average normalized coverage of target region was ~1500x. (B) A number of variants filtered out from the raw list at different coverage cut-off reaching near saturation at a minimum of 15x coverage. (C) All the variants are shown by a base-pair change at various allele fractions. Arrow points to C:G/T:A variants, known as a product of FFPE artifacts, to be enriched below 0.05 allele fraction which was filtered out from the study.



III. Additional Figure S4: Variant prioritization strategy to enrich cancer-related variants.

The box on left shows total coding variants called in the cohort. The box on the right shows filtered variants after removal of known SNPs, highly recurrent variants (>4 pools), variants failing to be called deleterious by at least 7 functional prediction tools and variants not called by both of the variant caller GATK & Mutect, left with total 99 variants in 26 genes.



No.	Gene	Mutation	5XA	5XB	10XA	10XB	15X	Unk
1	ABL1	G250R	■					■
2	APC	L1482F		■				■
3	APC	M1583V	■					
4	APC	P1369S	■					
5	APC	P1420L			■			■
6	APC	R640W	■					
7	BRAF	G464R		■				■
8	BRAF	V471I	■					
9	CDH1	D400N		■			■	■
10	CDH1	P160S	■		■			■
11	CDH1	V685M		■	■		■	
12	CDKN2A	G35E			■			
13	CSF1R	P566L			■			
14	EGFR	ELREA746del	■	■	■		■	■
15	EGFR	G779S		■				
16	EGFR	L858M	■					■
17	EGFR	L858R	■	■			■	■
18	EGFR	R831C	■	■			■	
19	EGFR	S768I		■				■
20	EGFR	V774A		■		■	■	
21	EGFR	V834A	■					
22	FGFR3	C228R	■					
23	FGFR3	G691R	■	■				■
24	FGFR3	P283S		■			■	
25	FGFR3	S249C	■					

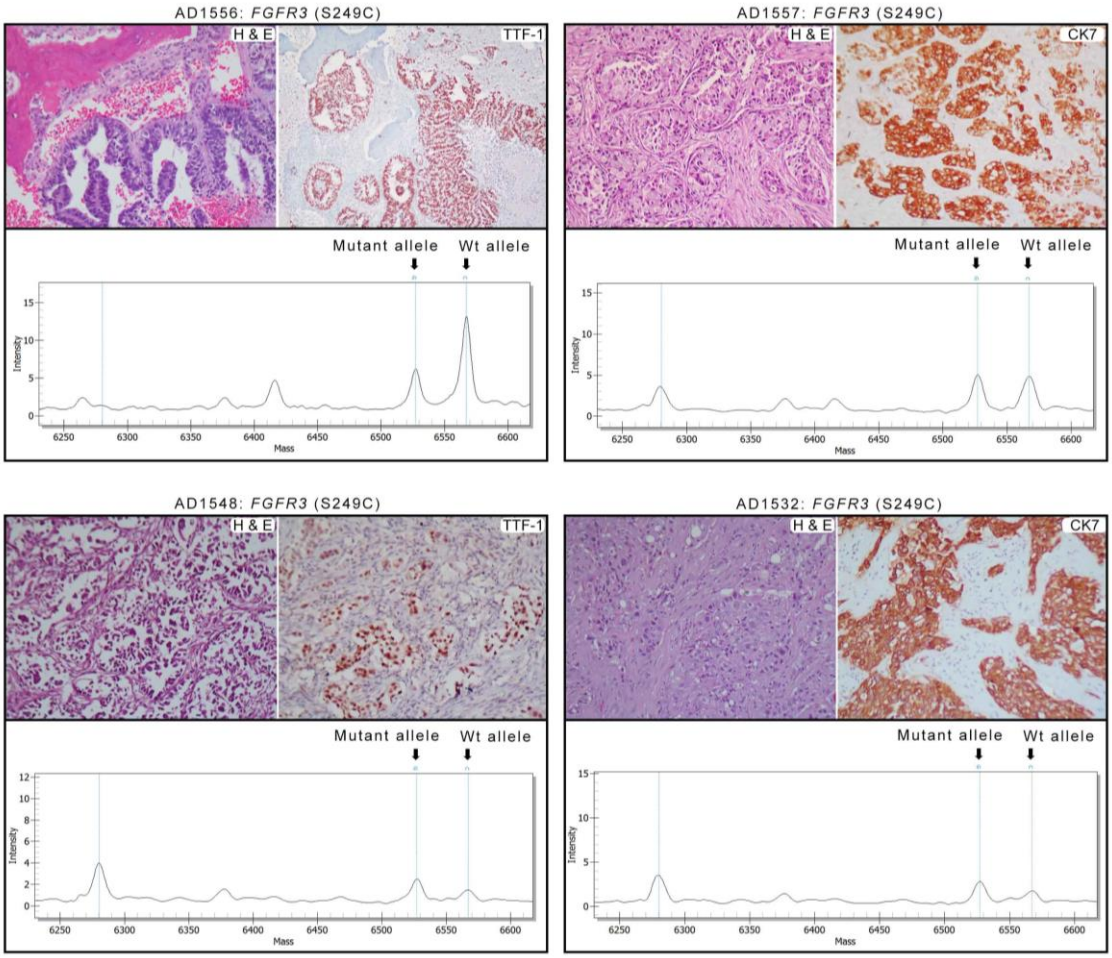
No.	Gene	Mutation	5XA	5XB	10XA	10XB	15X	Unk
51	NF2	R418C		■			■	
52	NOTCH1	L2510F		■				■
53	NOTCH1	R1594Q	■					■
54	NOTCH1	T2511I		■				■
55	PDGFRA	H570Y	■					
56	PDGFRA	P155L		■				■
57	PDGFRA	P589Q		■			■	
58	PTEN	A126T				■		
59	PTEN	C105Y	■					
60	PTEN	G127R				■		■
61	PTEN	G129E	■		■			■
62	PTEN	G251D	■	■			■	■
63	PTEN	H123Y				■		
64	PTEN	P30L		■			■	■
65	PTEN	P339S		■				■
66	PTEN	R130Q		■				■
67	PTEN	S227F				■		
68	PTEN	V133I	■					■
69	RB1	R798W		■			■	■
70	RET	A756V	■					■
71	RET	R721Q			■			
72	RET	R873W		■			■	
73	RET	R912W		■				■
74	RET	S891L		■				
75	RUNX1	P86S		■			■	

No.	Gene	Mutation	5XA	5XB	10XA	10XB	15X	Unk
26	FGFR3	S679F			■			
27	FLT3	G617E		■			■	■
28	GJC3	E187K		■		■	■	■
29	JAK2	R564Q	■					■
30	JAK2	S591L	■					■
31	KIT	A784T		■				■
32	KIT	G565R		■				■
33	KIT	P551L	■					■
34	KRAS	A59T			■			
35	KRAS	G12V				■		
36	KRAS	P34L		■			■	■
37	LPPR3	A245T		■			■	
38	MET	G1183D		■				■
39	MLH1	A681T		■				■
40	MLH1	E172K		■			■	■
41	MLH1	L658F	■	■				■
42	MLH1	R265C		■				
43	MSH2	C822R			■			■
44	MSH2	D758N			■			
45	MSH2	S676L	■					
46	MSH2	S743L	■	■			■	■
47	NF2	C133Y	■		■			
48	NF2	E372K	■	■			■	■
49	NF2	R196Q				■		■
50	NF2	R341Q	■					

No.	Gene	Mutation	5XA	5XB	10XA	10XB	15X	Unk
76	RUNX1	R135K		■				■
77	RUNX1	R223H		■			■	
78	TP53	A138T		■			■	
79	TP53	A276V			■			
80	TP53	A347T		■				■
81	TP53	C141Y	■					
82	TP53	C277F		■			■	
83	TP53	D208N		■			■	
84	TP53	E171G		■			■	
85	TP53	E271K		■			■	
86	TP53	G262V			■			
87	TP53	L130F		■				■
88	TP53	P191H		■			■	
89	TP53	R175H				■		
90	TP53	R273H				■		
91	TP53	R280G	■					■
92	TP53	S269I				■		
93	TP53	T118I		■				
94	TP53	T125M		■	■			
95	TP53	T230A		■			■	
96	TP53	T230I		■			■	
97	TP53	Y220N			■			
98	VHL	G144E		■			■	■
99	VHL	V74A	■			■		

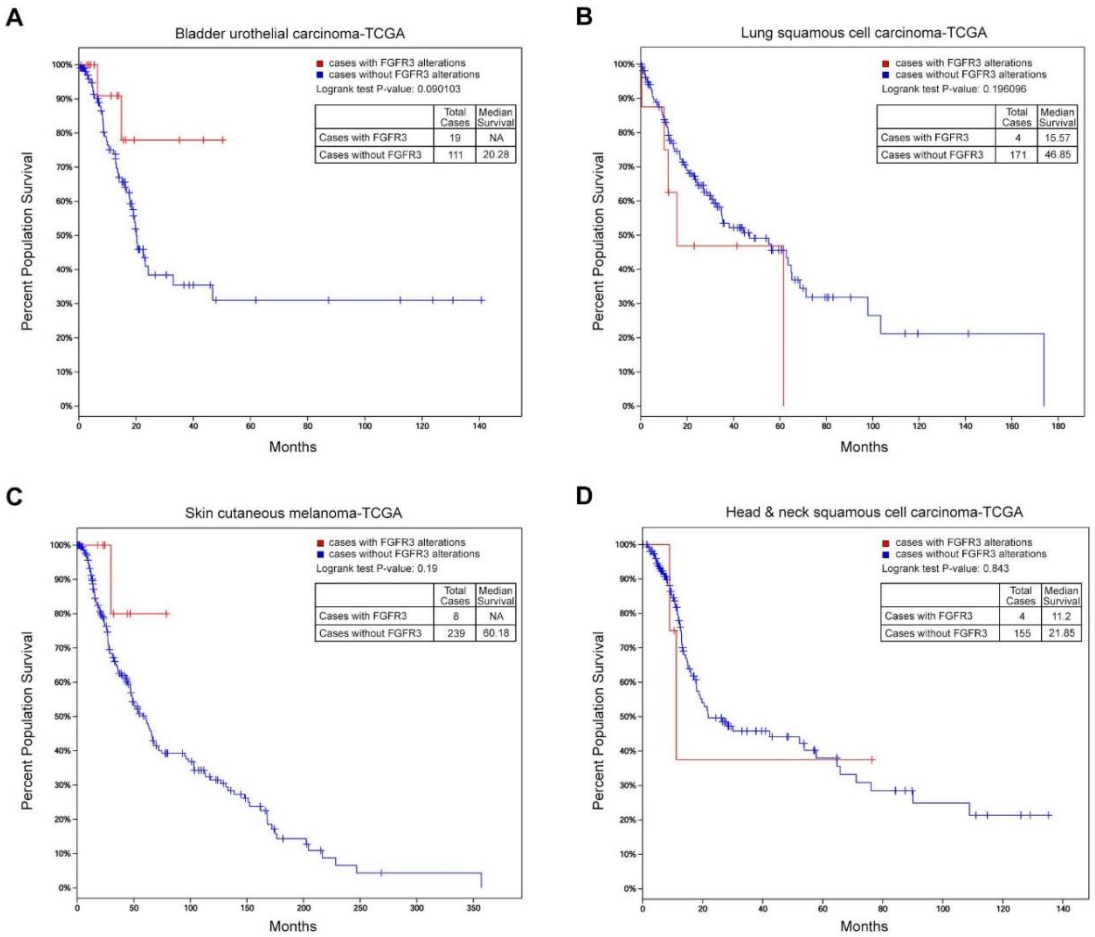
III. Additional Figure S5: List of mutations identified by high-throughput sequencing.

List of 99 mutations in 23 genes qualifying after FFPE signature, dbSNP and TMC-SNPdb and functional prioritization filters is shown. Grey box indicates the presence of a mutation in the given pool.



III. Additional Figure S6: Immunohistochemical and mutational analysis of samples harboring *FGFR3* mutations.

Representative images of immunohistochemical (IHC) analysis and spectra of *FGFR3* mutation in lung adenocarcinoma are shown. Upper panel: Hematoxylin and Eosin stain (H & E; on left) and thyroid transcription factor-1 (TTF-1) or cytokeratin 7 (CK7) positive immunostaining are shown. Lower panel: Spectra of FGFR3 mutations detected by mass-spectrometry are shown. Wild-type allele (Wt) peak and mutant allele peak are indicated with arrows.



III. Additional Figure S7: Kaplan-Meier overall survival analysis of patients harboring *FGFR3* mutations using cBioPortal.

We queried cBioPortal for *FGFR3* mutant patients in TCGA cohort to compare survival with *FGFR3* wild-type population. Four cancer types in cBioPortal were identified with *FGFR3* mutations and availability of survival information, (a) bladder urothelial carcinoma (b) lung squamous cell carcinoma, (c) skin cutaneous melanoma and (d) head & neck squamous cell carcinoma. Survival curve of *FGFR3* mutant patients (red) and *FGFR3* wild-type patients (blue) is shown.

III. Additional Table S1: Demographic characteristics of lung adenocarcinoma patients.

Discovery set: total number of patients for next-generation sequencing	**45**
Sex	
Male	20
Female	23
Habits	
Smokers	3
Non-smokers	23
Not available	19
Validation set: total number of patients for mass-spectrometry based validation	**363**
Sex	
Male	226
Female	137
Habits	
Smokers	73
Non-smokers	268
Not available	22

III. Additional Table S2: Quantitative analysis of variants per pool identified by NGS.

Pool	Read pairs	Reads mapped on target	Percent reads mapped on target	Bases covered	Bases covered on target	Bases covered on target coverage > 1650	Enrichment factor	Number of coding variants	Number of known SNPs
5XA	51,18,296	4713208	92	1,49,294	1,08,911	93,863	18802	837	111
5XB	63,77,317	5987503	94	1,44,443	1,09,328	96,467	19508	2145	214
10XA	68,92,629	6462346	94	1,29,117	1,09,067	95,659	21772	756	94
10XB	52,00,746	4950561	95	1,26,939	1,09,323	99,887	22197	795	115
15X	83,30,714	6217784	75	3,16,354	1,12,575	1,02,817	9172	1110	125
Unk	72,95,420	1433631	20	2,29,488	1,09,559	62,236	12305	1607	191

III. Additional Table S3: In-vivo tumorigenicity of NIH 3T3 cells expressing FGFR3 mutants and wild-type.

NIH 3T3 cells transfected with	Total number of mice injected	Total number of mice showing tumor formation
FGFR3 (wild-type)	11	3
*FGFR*3 (S249C)	12	12
FGFR3 (G691R)	12	6
Control	3	0

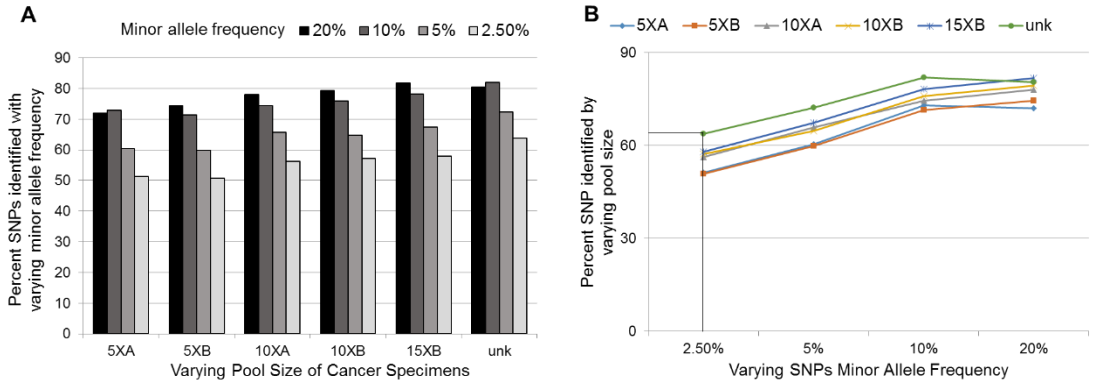
III. Additional Table S4: Details of correlation between clinicopathological features of lung cancer patients and *FGFR3* mutation status.

Clinicopathological features	Variables	N (% along column)	*FGFR3* status (N=363) (% along row)		P-value*
			Mutant (N=20)	Wild-type (N=323)	
Age	< 45 years	95 (26%)	9 (9.5%)	86 (90.5%)	**0.048**
	> 45 years	268 (74%)	11 (4%)	257 (96%)	
Gender	Male	226 (62%)	9 (4%)	217 (96%)	0.102
	Female	137 (38%)	11 (8%)	126 (92%)	
Smoking	Smoker	73 (20%)	2 (3%)	71 (97%)	0.199
	Non-smoker	268 (80%)	18 (7%)	250 (93%)	

3.6 Additional experiment to validate the pooled sequencing assay

Efficiency of pooled sequencing by taking SNPs as control

To determine the efficiency of variant calling and filtering procedure from pooled DNA sequencing, we used *EGFR* mutations as positive control (SI Table-2), independently identified in all the samples using TaqMan assay [49]. All of the expected mutations were identified in each pool except *EGFR* exon 19 deletion in pool 10XB, which was called in pools of undetermined reads. Next, we quantitatively analyzed known SNPs to estimate statistical power for determining mutant alleles of variant frequency. First, a reference list of known SNPs (III. Additional Table 2) was generated from the abbreviated genome size of 127 Kb representing 676 exons and stratified based on their minor allele frequency (MAF) as reported in dbSNP version 132. Next, SNPs at varying MAF observed in each pool against those expected based on the reference list was analyzed as a function of increasing pool size. A histogram plot representation (III. Additional Figure 8) for observed SNPs with variable minor allele frequency against pool size revealed 15XB pool was able to capture ~ 70% of expected variants having 5% MAF and ~60% of the expected variants at 2.5% MAF-- suggesting 15X pooling size or higher retains adequate statistical power to detect variants accusing at 2.5% or higher in the given sample set for the given genome size and depth coverage. Pool of undetermined reads (unk pool) retained more statistical power then 15X pool indicating that still higher pool sizes can further enhance the statistical power.



III. Additional Figure S8: Efficiency of pools to capture known SNPs at varying minor allele frequency.

(A) Histogram representing percent of SNPs identified with varying minor allele frequency in various pools of 5 individuals (5XA and 5XB); pools of 10 individuals (10XA and 10XB); 15 individuals (15X). (B) Line plot shows cumulative percent of SNPs identified at varying minor allele frequency in each pool.

IV. INTEGRATED GENOMICS APPROACH TO IDENTIFY BIOLOGICALLY RELEVANT ALTERATIONS IN FEWER SAMPLES

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

4.1 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 aero-digestive tract and voice box [111]. 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* [112-114]. These landmark studies apply elegant statistical methodologies like MutSig [115], Genome MuSiC [116], Intogen [117], InVEx [118], ActiveDrive [119] and GISTIC [120] 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 [121]. 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 hyper-methylated 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 [8, 122]. Combinatorial sources of genetic evidence converging at same gene or signaling pathway can also limit false positives by filtering strategy and potentially reducing the multiple hypothesis testing burden for identification of causal genotype-phenotype associations [123]. 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 [124]; and, more recently, Wilkerson *et al.* proposed superior prioritization of mutations based on integrated analysis of the genome and transcriptome sequencing than filtering based on conventional quality filters [125]. These and several other reports all together emphasize integration of multi-platform genomic data for identification of cancer related genes [126].

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.

4.2 MATERIALS & METHODS

4.2.1 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 [127],[128]. 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.

4.2.2 SNP array analysis

Genomic DNA was extracted from pre-clonal and clonal cell lines using PAXgene Tissue DNA Kit (Qiagen, USA). 200ng 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) [129]. Inferred copy numbers were then annotated with genomic features using BedTools (v. 2.17.0) [130]. Copy number segments of more than 10Mb 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.

4.2.3 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 hours at 37oC and tyrosination, cells were washed with pre-warmed KCl (0.075M) (Sigma, USA) and

incubated with KCl at 37ºC in water bath for 60 minutes. 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 (50cm). Slides were kept on the water bath at 70ºC for few seconds followed by drying on heating block (set at 80ºC). 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 60ºC for 3 hours followed by trypsin digest (Trypsin/EDTA - concentration of 0.025%, Sigma, USA). Giemsa stain (Sigma, USA) (3%) was applied using coplin Jar for 15 minutes on slides followed by washing with distilled water.

4.2.4 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 Molecular 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 [65]. 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 [66, 131] and MuTect v.

1.0.27783 [67]. 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) [69] as standard known germline variants database and COSMIC (v. 62) [132] 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) [133], and three functional prediction tools PolyPhen2 (build r394) [134], Provean (v. 1.1) [135] and MutationAssessor (release 2) [136]. 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.

4.2.5 Transcriptome sequencing

Transcriptome libraries for sequencing were constructed according to the manufacturer's protocol. Briefly, mRNA was purified from 4µg 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) [137] and quantifying number of reads per known gene using cufflinks (v.2.1.1) pipeline [138]. All the transcripts were then binned by $\log_{10}$(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 [139]. Gene set enrichment was performed by submitting actively expressed transcripts lists to MSigDB V4 [140] and filtering resulting gene lists by p-value of enrichment. Variants were identified from transcriptome sequencing using GATK [66, 131]. Only variants having overlap with exome sequencing were considered as true genomic variants. Fusion transcripts were identified using ChimeraScan (v.0.4.5) [141]. Raw fusion reads were filtered using following criteria: (1) Minimum 10 paired-end supporting reads or minimum 2 chimeric reads and (2) complex alterations. Minimum number of reads helps to reduce false positives due to low quality reads while complex rearrangements involving more than three different genomic locations in single rearrangement event were filtered by manual analysis. Candidate fusion events were used for integration and visualization in Circos plot.

4.2.6 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 harboring 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 harboring positive correlation of focal copy number and gene expression; 2) selection of genes harboring point mutations with detectable transcript and or altered copy number, and 3) selection of genes harboring multiple type of alterations identified from above two gene lists (Additional Figure S7). Circos plot representation of integrated genomics data was generated using Circos tool (v. 0.66) [142].

4.3 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.

4.3.1 Characterization of four HNSCC cell lines established from Indian patients

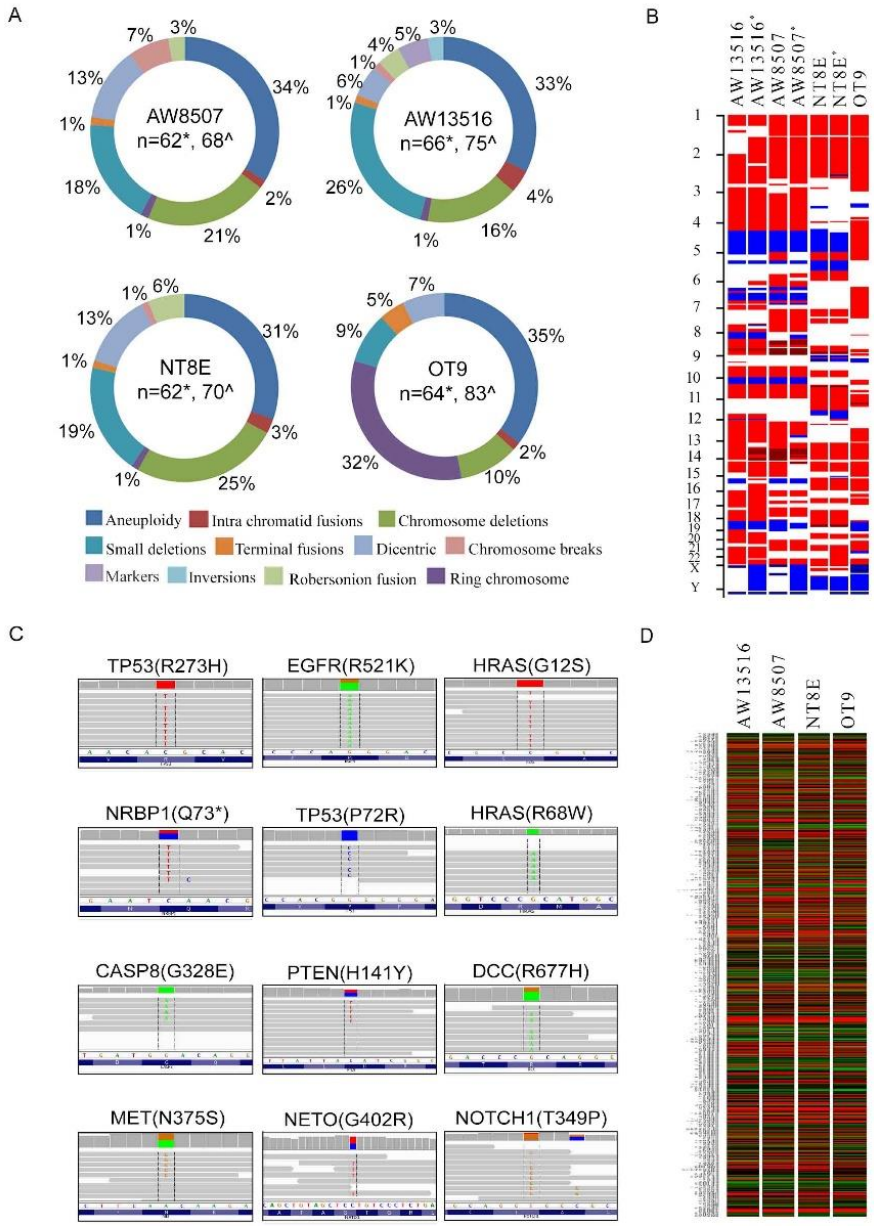
Given that higher accumulative effect of individual genes can be reckoned by integrative analysis, we argue that these alterations can possibly be determined even with fewer samples. As a proof of principle, we performed an integrated characterization of karyotype analysis, copy number analysis, whole transcriptome and exome sequencing of 4 HNSCC cell lines 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.

4.3.1.1 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 (Figure 1A; Additional Figure S1) and as reported for tumor cells lines [127, 128]. We specifically observed dicentric and ring chromosomes at elevated frequency indicating higher chromosomal instability (CIN) [143]. Overall distribution of chromosomal aberrations in each HNSCC cell lines showed similar pattern, representing an overall similar genomic structure of all HNSCC cell lines.

4.3.1.2 Copy number analysis

We performed genotyping microarray using Illumina 660W quad SNP array chips of all the cell lines (Additional Figure S2). After stringent filtering of initial genotyping calls, on an average, 253 genomic segments of copy number changes were obtained per cell



IV. Figure 1: Genomic alterations identified in HNSCC cell lines (B, C and D are modified from the manuscript).

(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

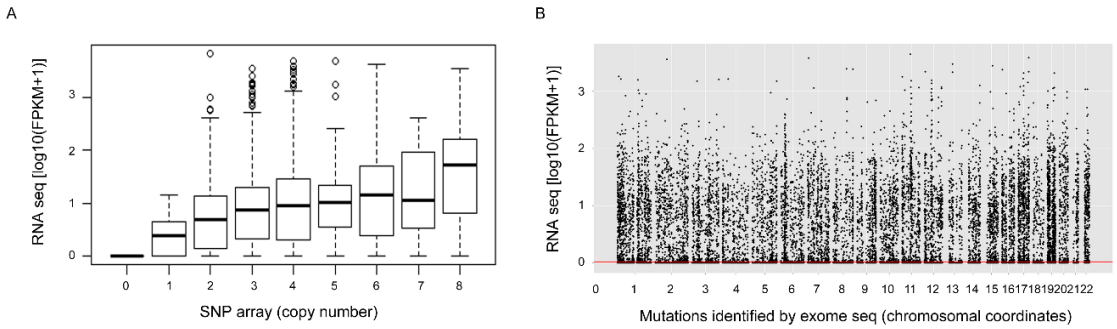
representation of copy number changes in each cell line identified by SNP array. Red indicates copy number gain and blue indicates copy number loss. * denotes clonal cells. (C) Representation of point mutations in hallmark genes identified by whole exome sequencing. (D) Heatmap representation of gene expression in each cell line identified by whole transcriptome sequencing. Red indicates overexpression and green indicates under expression.

line. By limiting segment size at 10Mb 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 [144, 145]; copy number gain on 11q known to be associated with advanced stage, recurrence and poor clinical outcome [146]; LOH at 8p and 9p which are known to be associated with advanced stage and survival [147] (Additional Table 1); 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 Table 2). Several hallmark genes were found to be amplified in cell lines such as *CCND1, NOTCH1, HES1* and *PIK3CA*; deletion of *CDKN2A* and *FBXW7* (Figure 1B, Additional Table 2).

4.3.1.3 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 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 [148]. All the transcripts showed 75% overlap of expression profile with each other (Additional

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 (Additional Table 3).



IV. Figure 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.

4.3.1.4 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 Table 4; Figure S5). The coding part of the four cell-lines 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 Table 4). HNSCC hallmark variants including: *TP53* (R273H and P72R), *PTEN* (H141Y), *EGFR* (R521K), *HRAS* (G12S and R78W), and *CASP8* (G328E) were identified (Appendix 4).

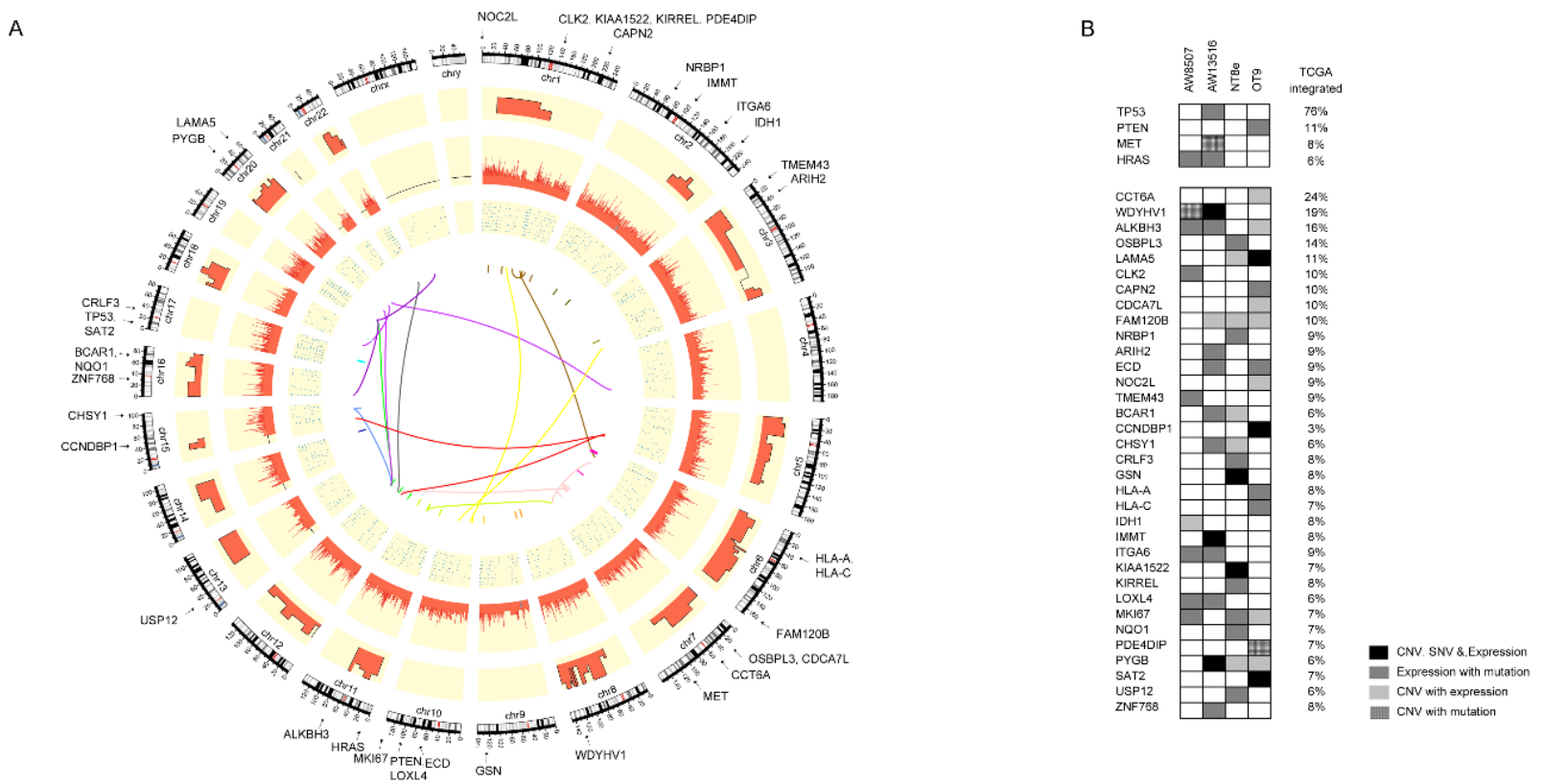
4.3.2 Integrated analysis identifies hallmark alterations in HNSCC cell lines

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 Figure S6A), median expression of focally amplified and deleted genes positively correlated to their expression (Figure 2A and Additional Figure S6B). About 1,000 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 9,253 unique missense variants in all four cell lines (Figure 2B). 2,479 missense mutations of 9,523 total mutations found across all cells were used for further integration with copy number and expression data (Additional 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 Figure S7). Briefly, genes harboring two or more type of alterations were selected: harboring positive correlation of focal copy number and gene expression; or those harboring point mutations with detectable transcript harboring the variant—based on which, we identified 38 genes having multiple

types of alterations (Appendix 4). 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 [149]. We additionally identified novel genes like *CCNDBP1, GSN, IMMT, LAMA5, SAT2* and *WDYHV1* to be altered by all three analyses i.e. CNV, expression and mutation. These all genes were also found to be altered in TCGA dataset with minimum 3% cumulative frequency (Additional Figure S8). The overall convergence of copy number, expression and mutation data in each cell line is represented as Circos plot (Figure 3A; Additional Figure S9). Among the novel genes identified, of genes with at least one identical mutation previously reported include a pseudo-kinase 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 [27, 150].

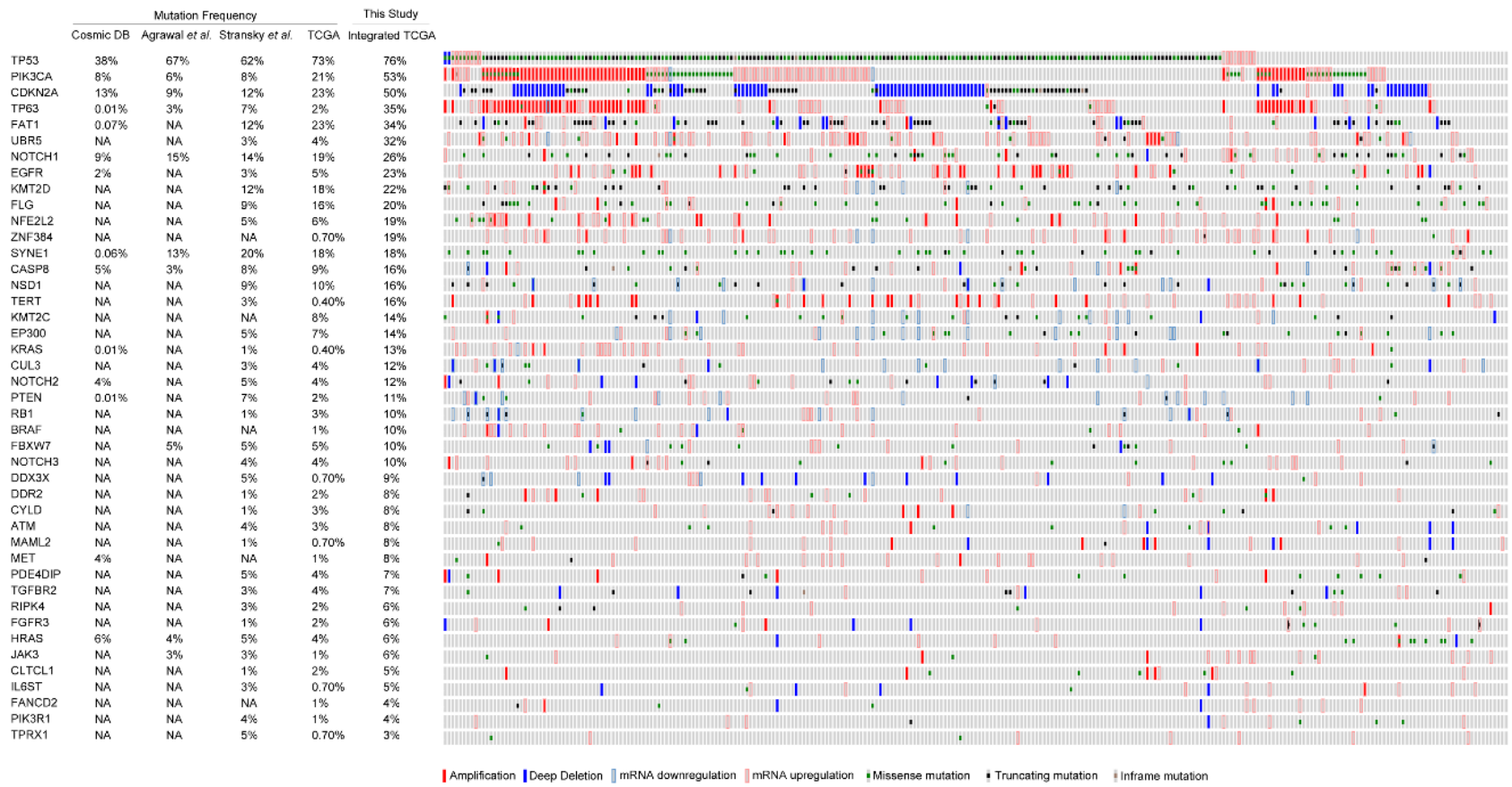
4.3.3 Integrated analysis of TCGA dataset for HNSCC hallmark genes

Next, as a proof of principle, we computed cumulative frequency of copy number variations, expression changes and point mutations across 43 genes with ~3% and higher mutation frequency in HNSCC TCGA dataset. As expected and described for few genes [114, 151], most of the genes were found to be altered at higher cumulative incidence than as reckoned by individual alterations (Figure 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.



IV. Figure 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 color indicates copy number gain or higher gene expression and blue color 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 colored 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.



IV. Figure 4: 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.

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.

4.4 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.

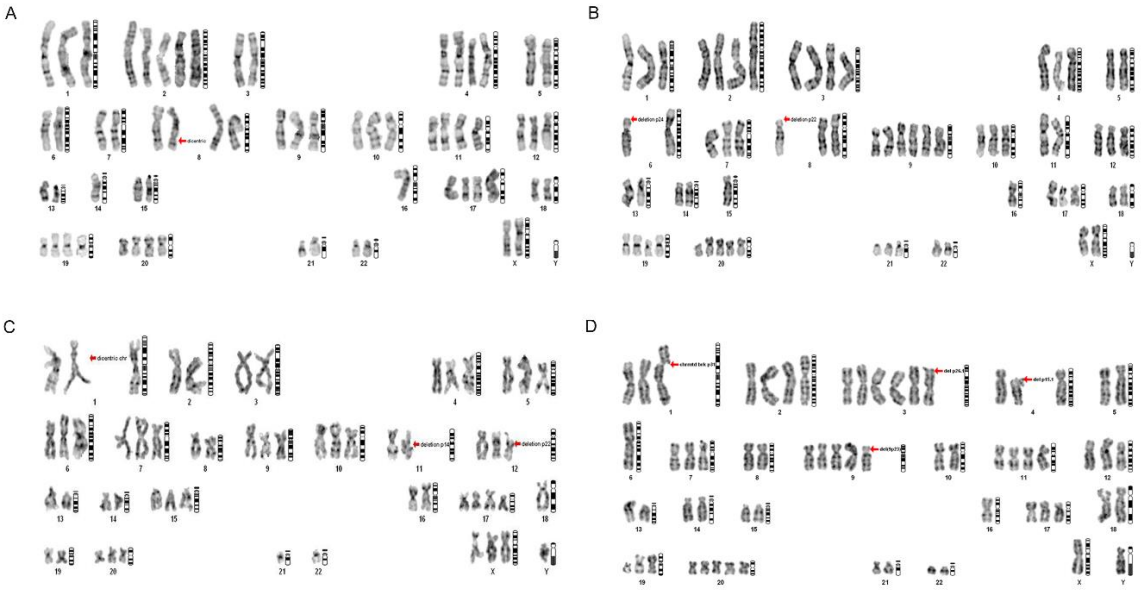
Integrated analysis of the cell lines establishes their resemblance to primary tumors. Consistent with literature, most frequent copy number gains in head and neck cancer cells in this study were observed at 2q, 3q, 5p and 7p, and deletions at 3p, 9p, 10p, 11q, 14q, 17q and 19p, as reported earlier [152, 153]. Integration of multiple platform with the copy number variation, allowed us to identify the functionally relevant alterations including several hallmark 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 [113, 149] 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 [154]. 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 [27, 150], with an overall 9% cumulative frequency alteration in TCGA HNSCC dataset (Additional 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 [155-157].

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 [114, 151]. 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 [158]; amplification of *TERT* has been shown to be higher in lung squamous [159], suggesting these alterations as potential squamous specific event, though that warrants detailed systematic assessment.

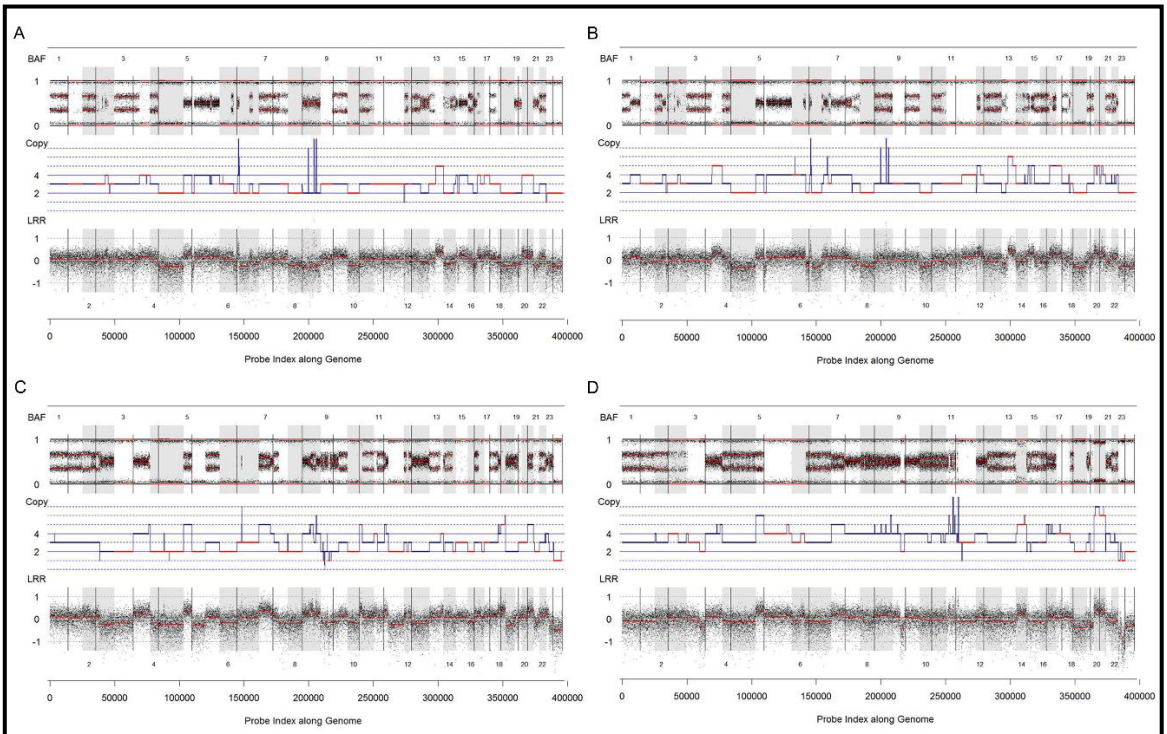
In overall, this study underscores integrative approaches through a filtering strategy to help reckon higher cumulative frequency for individual genes affected by two or more alterations to achieve the threshold for statistical significance even from fewer samples. The integrative analysis as described here, in essence, is based on a linear simplified assumption of disease 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.

4.5 Additional Supporting Data



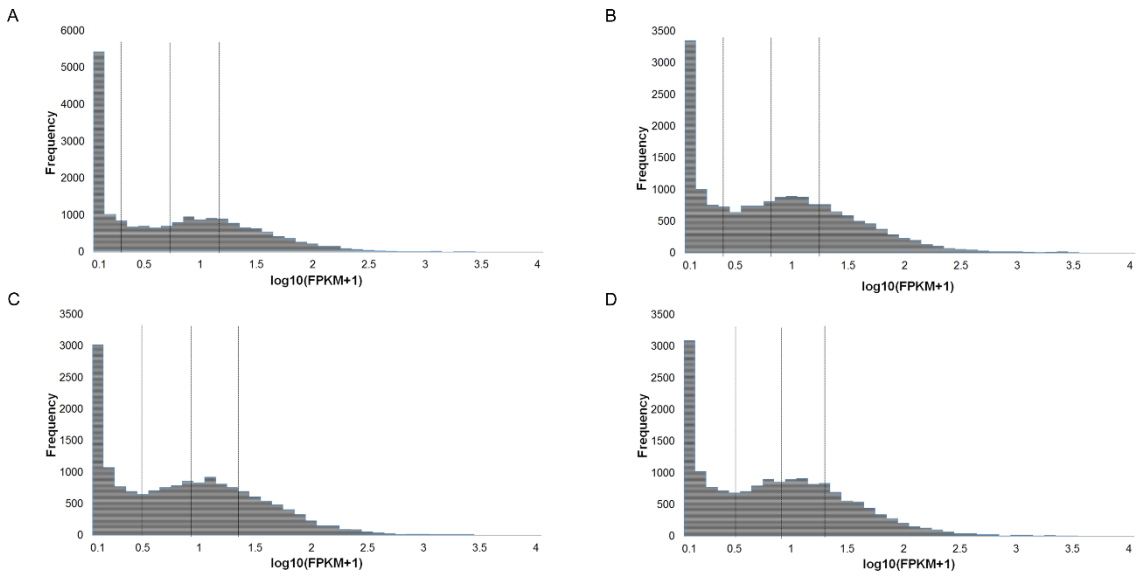
IV. Additional Figure S1: Chromosomal aberration in HNSCC patient derived cell lines AW8507, AW13516, NT8e and OT9.

Representative karyotype of (A) AW8507, (B) AW13516, (C) NT8e and (D) OT9 cells is shown from total 25 karyotypes obtained per cell line.



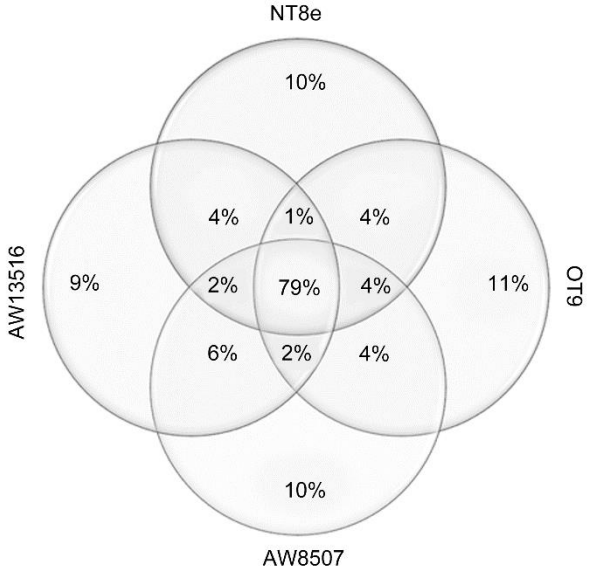
IV. Additional Figure S2: Copy number changes in HNSCC cell lines identified by SNP array.

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 (BAF), second block represents absolute copy number, third block is log R ratio (LRR).



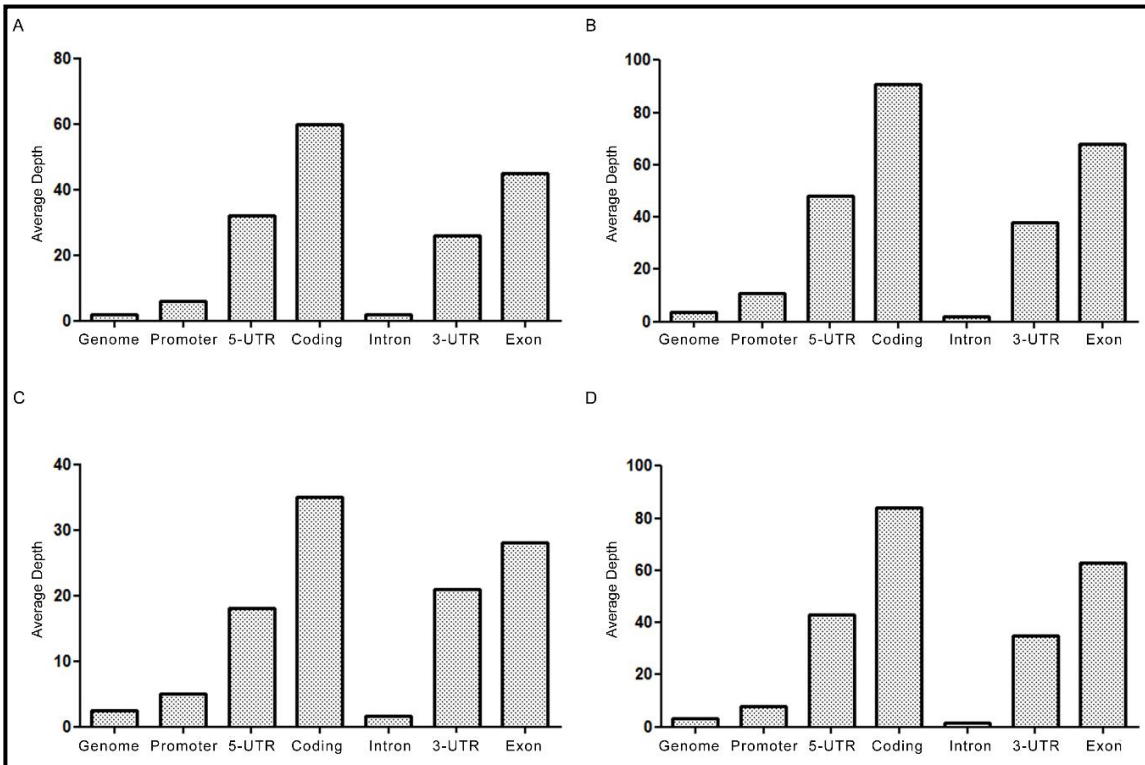
IV. Additional Figure S4: Frequency of transcripts per binned log transformed FPKM+1.

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 indicate percentile of transcripts in the quadrant.



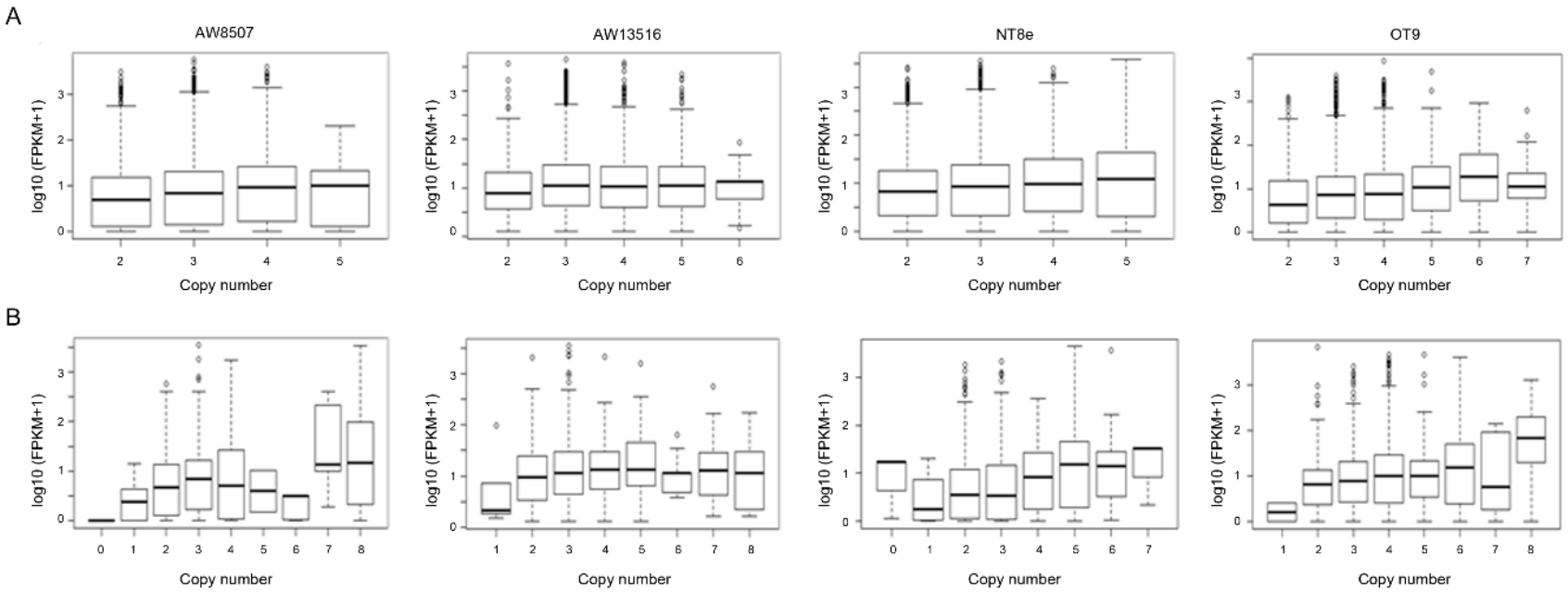
IV. Additional Figure S3: Similarity of gene expression in HNSCC cell lines.

Number of genes commonly expressed between AW8507, AW13516, NT8e and OT9 cell lines.



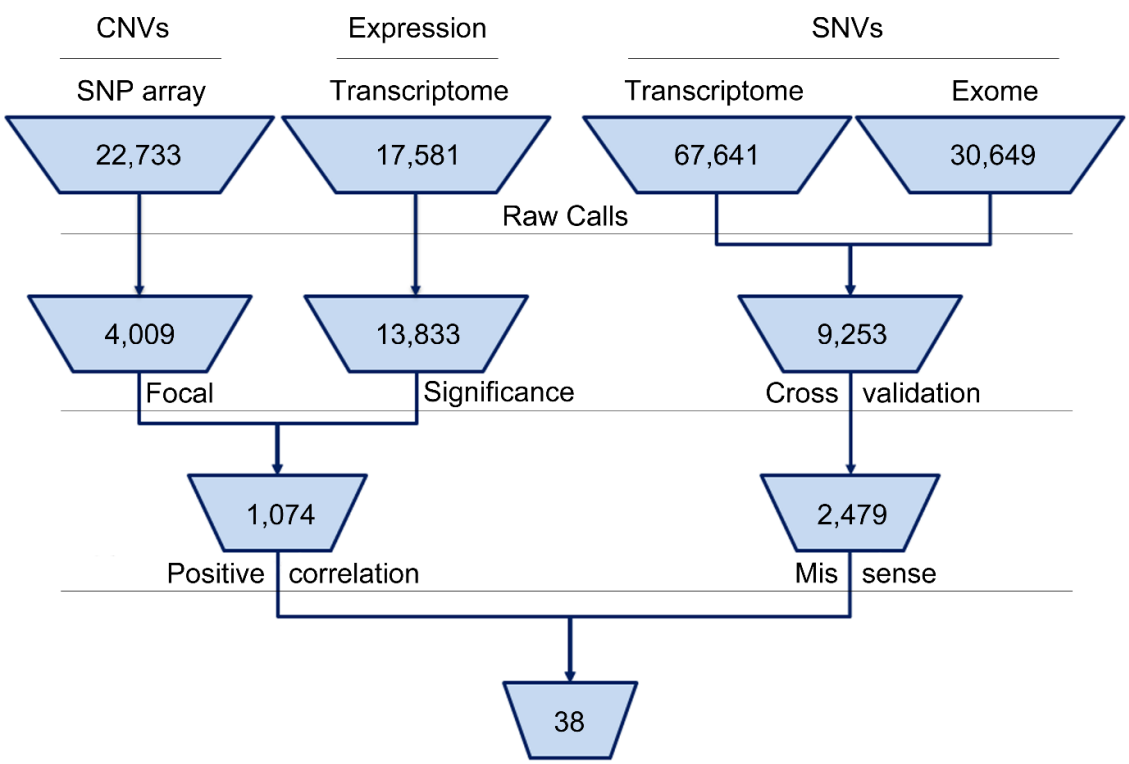
IV. Additional Figure S5: Relative depth in exome sequencing.

Relative depth of sequencing for various genomic regions in (A) AW8507, (B) AW13516, (C) NT8e and (D) OT9.



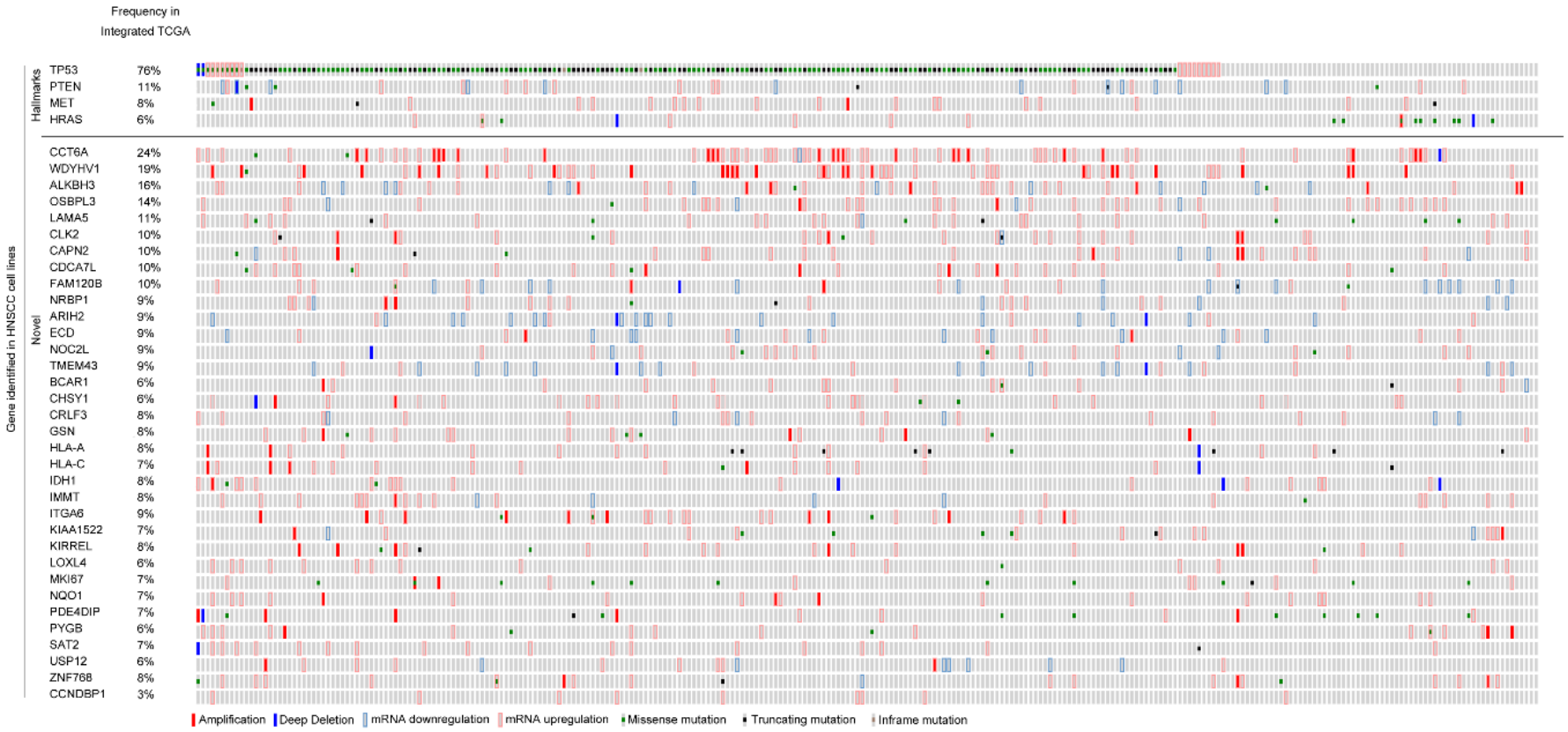
IV. Additional Figure S6: Correlation of copy number with gene expression.

(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



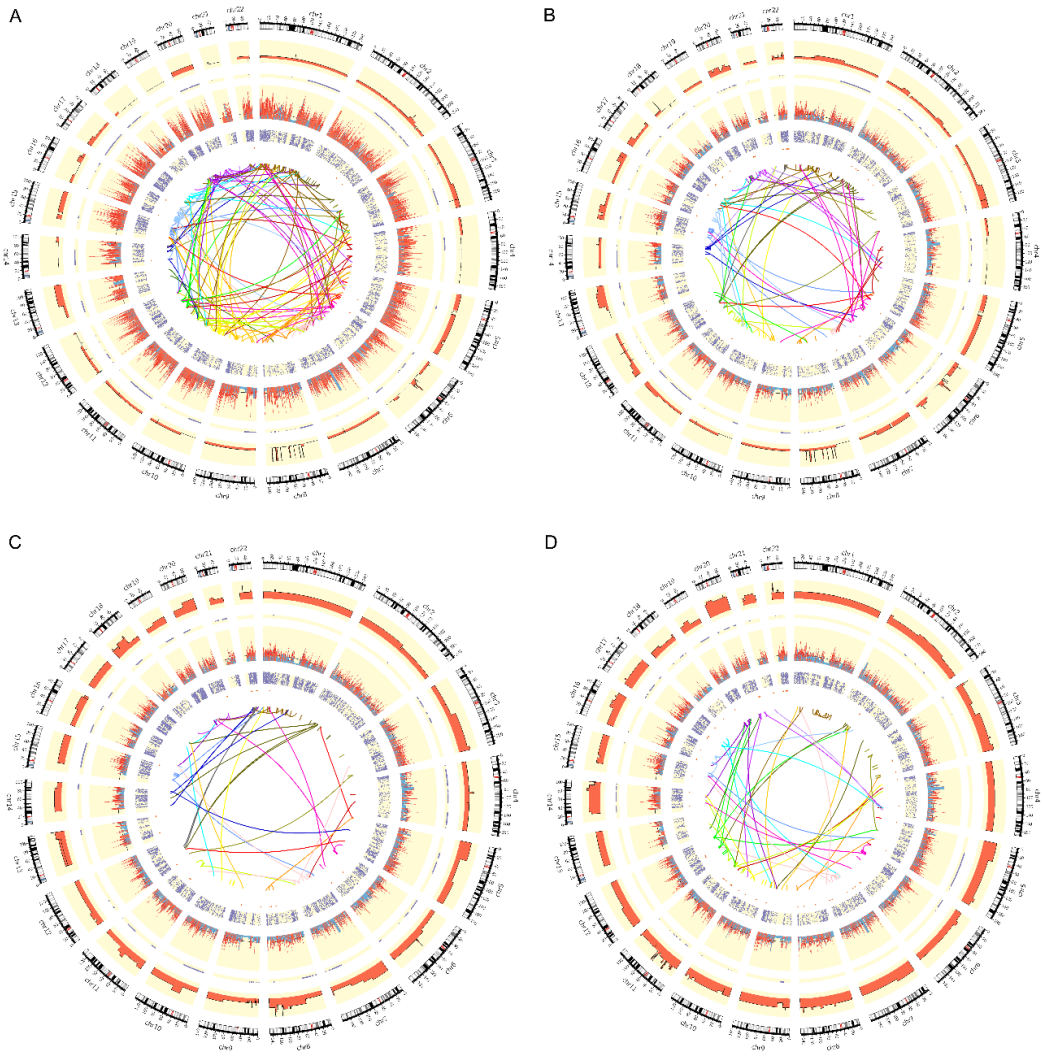
IV. Additional Figure S7: Schematic view of data reduction in integrated genomic analysis.

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.



IV. Additional Figure S8: 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, missense (green), non-sense (black) and in-frame (brown) mutations are indicated by smaller square box.



IV. Additional Figure S9: 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 color indicates copy number gain or higher gene expression and blue color 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 colored by their chromosome of origin identified by ChimeraScan.

IV. Additional Table S1: Copy number alterations of known genomic locations identified in HNSCC cell lines.

Alteration		Observed in our analysis
Copy Number Gains	2q	AW8507, AW13516, OT9
	3q	AW8507, AW13516, NT8e
	5p	AW8507, NT8e, OT9
	7p	AW13516, NT8e, OT9
Copy Number Loss	3p	NT8e, OT9
	9p	NT8e, OT9
	10p	AW8507, AW13516, NT8e
	11q	OT9
	14q	NT8e
	17q	AW13516
	19p	AW8507, AW13516, NT8e

IV. Additional Table S2: Copy number alterations in hallmark genes identified in HNSCC cell lines.

	SNP array copy number			
	NT8e	AW13516	AW8507	OT9
CCND1	4	3	3	8
MET	2	3	3	4
MYC	6	7	8	4
PIK3CA	4	5	4	4
HRAS	4	3	3	4
HES1	5	5	4	5
JAK1	3	4	3	3
CDKN2A	1	3	3	2
DLL3	3	3	2	2
FBXW7	2	2	2	3
NOTCH1	3	3	3	4
NSD1	3	4	4	4

IV. Additional Table S3: Gene expression of hallmark genes by RNA sequencing

	RNA sequencing (log10[FPKM+1])			
Gene	NT8e	OT9	AW13516	AW8507
GAPDH	3.3	3.3	3.5	3.3
HRAS	2.2	1.9	2.0	2.0
MET	1.6	2.1	1.6	1.7
JAK1	1.5	1.5	1.4	1.4
CDKN2A	1.2	2.0	2.1	1.5
NSD1	0.9	0.8	0.8	1.1
FBXW7	0.8	0.7	0.6	0.7
SMAD4	0.7	0.7	0.8	0.8

IV. Additional Table S4: Features of whole exome and transcriptome sequencing.

Cell Line	AW8507	AW13516	NT8e	OT9
Number of mapped reads (% of total reads)	66,251,943 (97.66%)	78,224,226 (93.69%)	49,728,777 (93.95%)	68,206,969 (94.54%)
Average coverage of coding region	50	88	84	81
Total variants in exome sequencing (Ti/Tv)	47892 (2.21)	28813 (2.19)	20864 (2.04)	25029 (2.17)
Total variants in transcriptome sequencing (Ti/Tv)	53330 (1.03)	44168 (1.19)	109609 (1.07)	63458 (1.03)
Non-synonymous variants	5623	4498	2775	5139
Mean number of non-synonymous variants per Mb coding DNA	90	72	44	82

Ti = Transition; Tv = transversion

V. NGS BASED APPROACH TO DETERMINE THE PRESENCE OF HPV AND THEIR SITES OF INTEGRATION IN HUMAN CANCER GENOME.

(as published in *British Journal of Cancer* (2015) 112:1958-1965)

5.1 INTRODUCTION

Human papilloma viral (HPV) infections has been associated with various types of cancer. Epidemiological studies indicate that about 90% of cervical cancers, 90-93% of anal canal cancers, 12-63% of oropharyngeal cancers, 36-40% of penile cancers, 40-64% of vaginal cancers and 40-51% of vulvar cancers are attributable to HPV infection [160, 161]. Currently HPV detections are primarily carried out using PCR based MY09/11 and CPI/II systems [162]. Other techniques used include hybridization based SPF LiPA method, signal amplification assays (Hybrid Capture 2 and Cervista) and nucleic acid based amplification like microarray, real time PCR based methods (COBAS 4800 real time test) [162-164]. These technologies come with limitations to detect minor, low-abundance HPV genotypes and complex mixture of co-infections that can be a negative determinant of the clinical outcome [165, 166]. Next generation sequencing (NGS) technologies overcomes such limitations, as evident from the recently described high-risk HPV genotyping assay for primary cervical cancer screening based on self-collection [167], using TEN16 or HIVID methodology, and to determine co-infection among the HPV types probed along with their sites on integration [168-172]. However, there's an unmet need for a simplified tool for biologists with no previous experience or knowledge of informatics to analyze the data generated by whole exome, transcriptome or genome sequencing using NGS technology to detect the presence of HPV sequences along with their integration sites. There are a variety of gene integration finding tools available that can detect different pathogen insertions in the human genome such as ViralFusionSeq

[173], VirusSeq [174], VirusFinder [175], Path-Seq [176], RINS [177], and ReadSCAN [178]. These tools have their specific third party needs, and are not specific for HPV detection. They can detect presence of HPV sequence along with other viruses, but lacks information to annotate the region of HPV genome detected. Here, we describe "HPVDetector" as a specific *in-silico* automated tool that is capable of multi HPV type detection, their annotation and determination of site of HPV integration utilizing raw exome, transcriptome, or whole genome data as input with minimal requirement for third party tools.

5.2 MATERIALS & METHODS

HPV detection involves computational subtraction based approach, where next generation sequencing data is used for alignment against custom made HPV multi-reference genome sequences to detect the traces of multiple HPV types using an automated pipeline (Figure 1).

5.2.1 HPV reference sequences and annotation

As a first step of the pipeline HPV genomes in FASTA format is required. We have acquired GenBank (.gb) files of 143 types of HPVs from a web resource Papillomavirus Episteme (PAVE) [179]. We converted these GenBank (.gb) files into FASTA files. These all reference sequences were concatenated to compose a multi-FASTA sequences using bio-perl modules [180]. Apart from this we also parsed the GenBank (.gb) files to generate HPV gene reference having nucleotide intervals for each gene of each HPV type. This gene reference file was used to annotate the HPV gene.

5.2.2 HPV type & HPV aligned reads detection

Evaluating the HPV type and HPV aligned reads is crucial to find HPV in the respective sample. For HPV type detection, we indexed the multi-FASTA HPV reference file using

BWA aligner followed by alignment of reads to indexed genome [65]. The aligned reads were extracted from the SAM file using a utility ViewSam from Picard Tools package (http://broadinstitute.github.io/picard/). The alignment files were parsed using UNIX shell program to detect the type of HPV as well as number of reads that align to a particular HPV type. Number of HPV reads were normalized to the total depth of coverage per sample and with respect to different HPV gene size.

5.2.3 Assessment of specificity and sensitivity of HPVDetector

We downloaded SiHa whole genome sequence from Sequence Read Archive database of DDBJ (https://trace.ddbj.nig.ac.jp/DRASearch/; study: SRP048769). The data was converted from SRA to FASTQ using SRAtoolkit. The resulting FASTQ files represents >36x genome coverage which was further down-sampled to 1x, 2x, 3x, 4x, 5x, 10x, 15x, 20x, 25x and 30x using Picard Toolkit's DownsampleSam function (http://broadinstitute.github.io/picard/). The resulting FASTQ files were used for testing HPV detection using HPVDetector.

5.2.4 Human-HPV integration loci detection

To detect integration sites, we created a custom reference genome comprised of human chromosomes and HPV FASTA sequences as pseudo-chromosomes. HPV genomes were appended to human chromosomes to compose a multi-FASTA reference genome. This custom Human-HPV reference genome was then used for aligning reads with short-read aligner BWA. The alignment files were parsed for the reads where one mate is aligned to human chromosome and another to HPV. The Human chromosomal positions, HPV type and HPV reference position were parsed and annotated with gene reference annotation file acquired from UCSC table browser [181] to get a list of integration sites.

5.2.5 RNA extraction, cDNA synthesis and E6 specific PCR:

Total RNA extraction was performed from tumor and cell lines using Trizol reagent (Invitrogen) as per manufacture's instruction and later resolved on 1.2% Agarose gel to confirm the RNA integrity. DNase treatment was done using DNase Free kit (Ambion, cat AM1906) followed by first-strand cDNA synthesis taking 2ug of total RNA using Superscript III kit (Invitrogen, 18080-051). E6 (HPV-16) and GAPDH expression checked as described previously [182].

5.2.6 HPV detection using MY09/11 and PCR primers

MY09/11 primer sequences were taken from previously reported literature [183]. All samples were screened by PCR first using MY09/11 primer. GAPDH was used as internal control for each sample. SiHa cell line [184] was used as positive control for HPV and AW13516 cell line [128] as negative control. The PCR reaction was performed in 20µl volume containing 10µl (2x) Biomix-Red master mix (Bio25005), 5µM each primer, 50ng gDNA. PCR condition was: initial denaturation: 95ºC, denaturation: 94ºC for 1min, annealing: 55ºC, (MY09/11, GAPDH), extension: 72ºC for 1 min and final extension at 72ºC for 5min on PCR machine (Veriti 96-Well Fast Thermal Cycler). The 10µl of PCR reactions were resolved on 1.8 % Agarose gel with EtBr with 100bp ladder for 1hr at 80 volts. To validate the HPV presence detected in SiHa cells, PCR was performed using SiHa cell line genomic DNA (100ng) as a template under the following conditions (Initial denaturation: 95ºC for 30s, denaturation: 95ºC for 30s, annealing: 58ºC for 30s extension:72ºC for 45sand final extension for 5 minutes at 72 ºC) using KAPA 2X ready-mix Taq Polymerase (KK1024). Primers were designed to amplify 122bp, 126bp and 120bp read sequences of HPV16 identified by HPVDetector in SiHa cell line. Primers flanking the human reads were designed to amplify 119bp and 290bp respectively. These sequences were further validated by Sanger sequencing. All the

experiments were repeated at least twice independently. The details of the primers used in the study are provided in Additional Table 2.

5.3 RESULTS

HPVDetector is a tool to quickly detect hundreds of Human Papilloma Virus types from next generation sequence data without any prerequisite knowledge about virus types. It runs on paired end sequenced samples. It is composed of two modes or sub pipelines as quick detect & integration detect mode (Appendix 5).

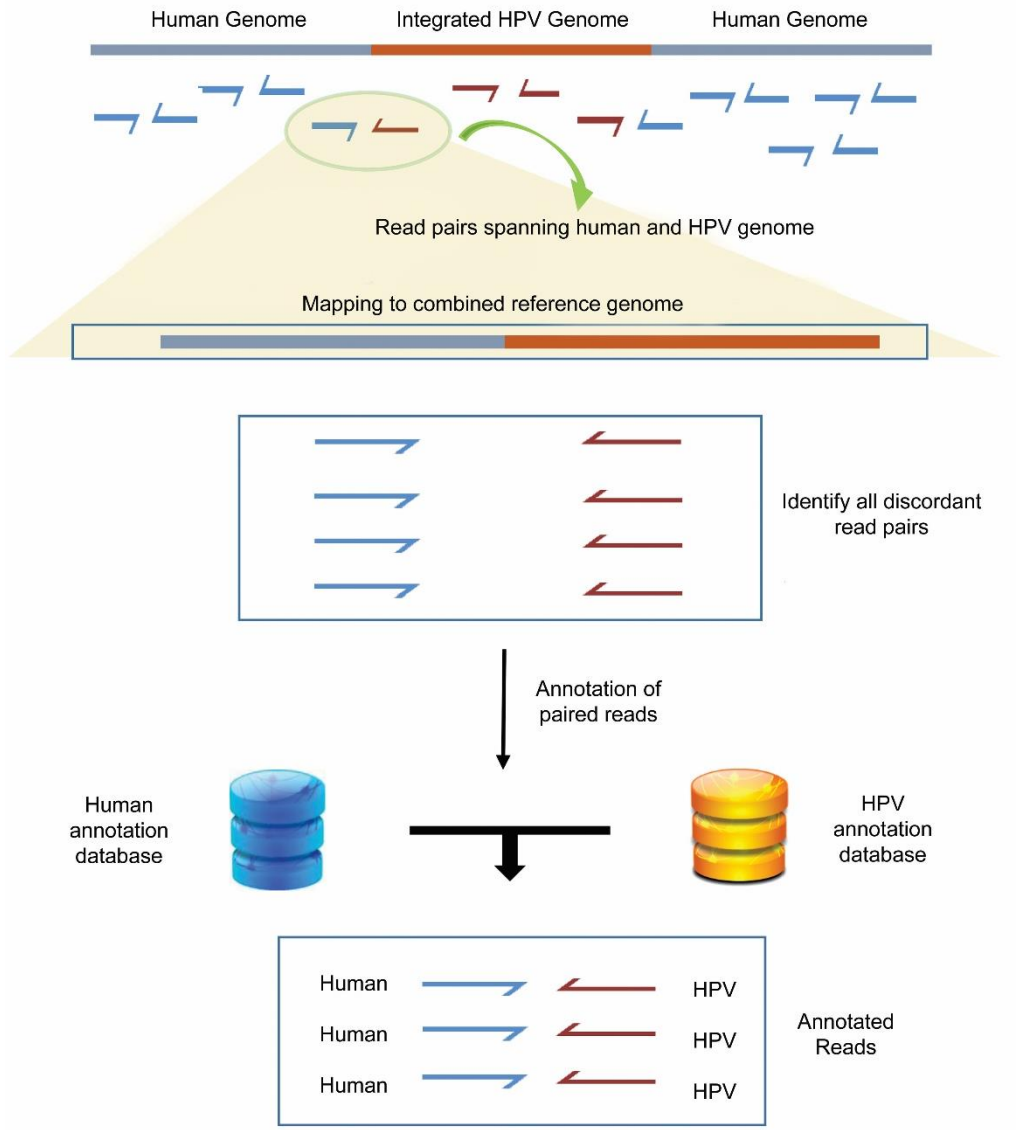
5.3.1 Quick detect mode

This mode is to quickly determine the HPV type or types to check if multiple HPV co-infections are existing or not in a given sample. Quick detect mode starts with alignment of raw paired end sequencing reads against custom made multi HPV genome using BWA aligner. Computational subtraction of the reads is then carried in which HPV aligned reads are retained using Picard Tools and further processed using UNIX shell program to distinguish reads mapping to different HPV types. Finally, HPVDetector outputs result file which enlists one or more HPV type(s) and number of HPV Reads.

5.3.2 Integration detect mode

This mode of HPVDetector determines genomic location of HPV integrant, annotate with HPV gene, human chromosomal loci and human gene. This mode of HPVDetector pipeline starts with alignment of raw reads against custom made reference including pseudo chromosome like multi-FASTA reference genome containing 143 Human Papilloma Virus reference sequences and Hg19 human reference genome. Computational subtraction is carried out to retain discordant read pairs where the sequences are aligned to both human as well as HPV genomes. Finally, HPVDetector outputs result file which

enlists HPV integration loci on human genome, annotation of HPV genes, human genes and human genome cytobands.



V. Figure 1: Conceptual workflow of the HPVDetector.

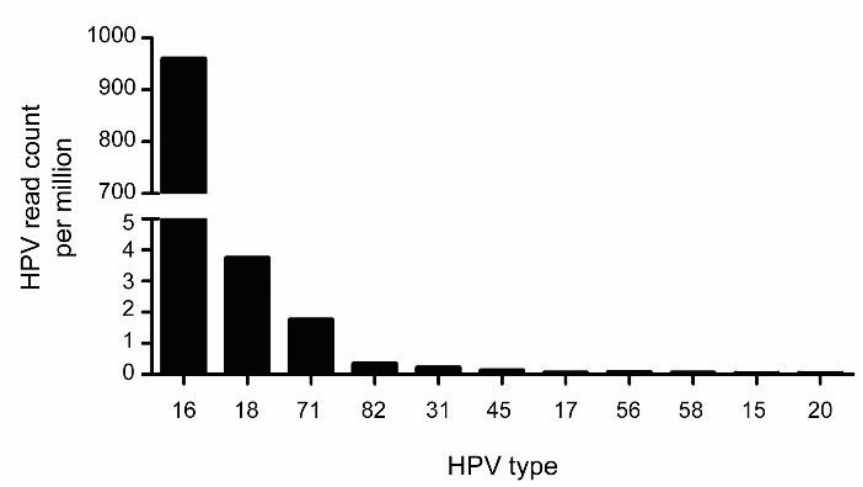
The flowchart represents workflow for HPVDetector. Paired end reads obtained from next generation sequencing data are aligned to a combined Human-HPV reference database. All discordant read pairs with one read aligning to human and other to the HPV genome are identified and annotated utilizing human and HPV database using an inbuilt annotator module.

5.3.3 Detection of HPV type integrated in the host genome

Cervical cancer exome sequencing data: In order to test the precision of HPVDetector we analyzed 22 cervical cancer exome sequencing data (generated in house at ACTREC, unpublished data) to detect the presence of HPV. Among the 22 samples analyzed, HPV was detected in 18 cervical samples, with maximum number of reads supporting HPV16 sequence (Figure 2) [185]. We also detected the presence of additional HPV types such as HPV71 (in 6 samples), HPV82 (in 5 samples) and HPV31 (in 2 samples) with variable number of supporting reads as shown in (Figure 3). Co-infection with more than one HPV type is known to be associated with significantly increased risk of cervical intraepithelial neoplasia 2+ and found in 43.2% of HPV-positive women [165, 186-188]. 6 of 22 cervical cancer patients (43%) were found to be co-infected with one or more HPV subtypes in this study using HPVDetector (Figure 3). Interesting to note, based of phylogenetic analysis of HPV types, HPV16 and HPV31 of the virulent alpha 7 group infection occurred in a mutually exclusive manner (in 13 of 22 samples), while HPV71 of alpha 15 subgroup known to be involved in commensal infections that infected 6 of 14 cervical tumor samples invariably co-occurred with other HPV subtypes [189, 190]. The HPV sequence detected in primary cervical tumor sample were independently validated by directed sequencing in T1094, the only sample with sufficient quality DNA (as shown in Additional Figure 1).

Tongue squamous cell carcinoma exome and transcriptome data: HPV is an independent risk factor in head and neck squamous cell carcinoma (HNSCC), in particular for oral and oropharyngeal carcinomas [191, 192]. We analyzed whole exome data from 23 paired and one orphan tongue squamous cell carcinoma (TSCC) sample and 7 head and neck squamous cell carcinoma cell lines (generated in house at ACTREC, unpublished data). None of the TSCC primary tumors were found to be HPV positive, as reported

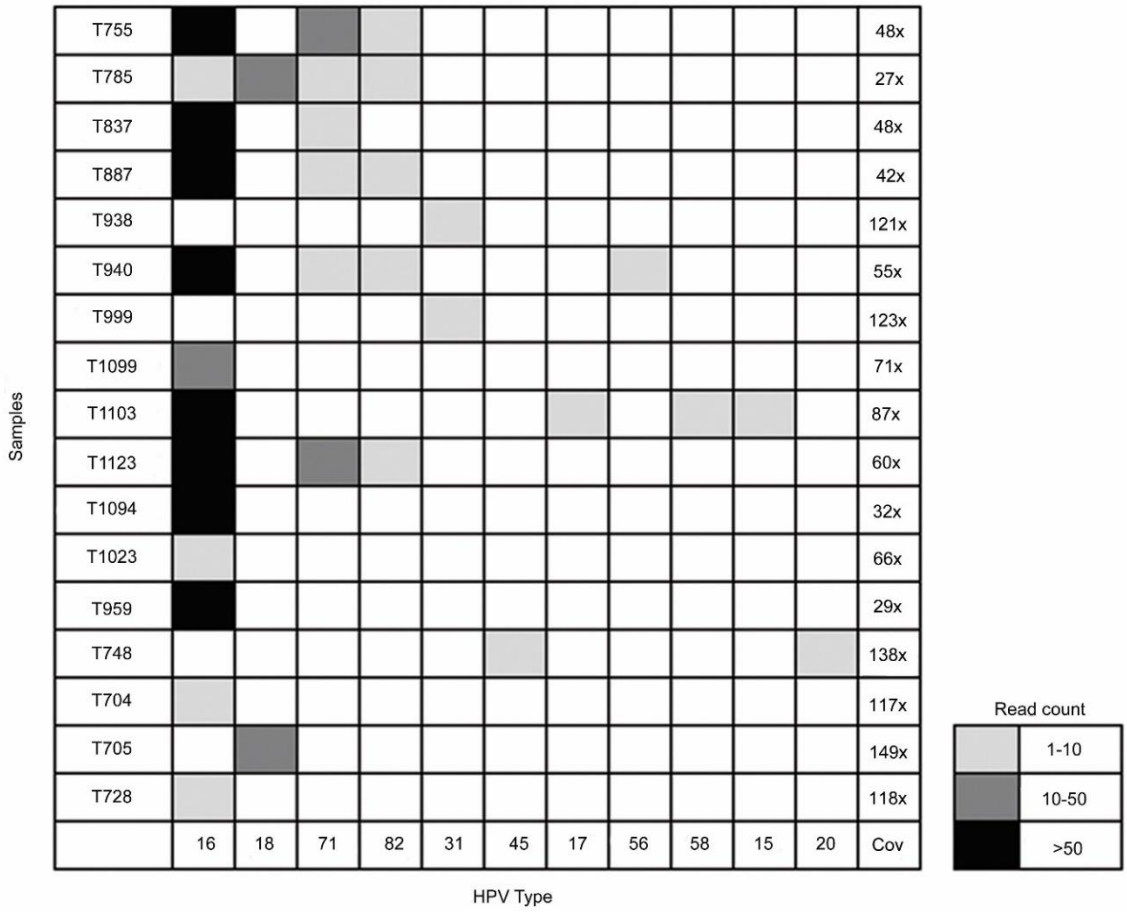
earlier [193-195]. The absence of HPV infection was further validated by PCR using MY09/11 and E6 primers (Additional Figure 2) on genomic DNA and cDNA respectively, suggesting low false negative feature of the HPVDetector. At the same time, among the cell lines, NT8e cells [127] of 7 cell lines analyzed was found to be positive for HPV71. Next, we analyzed whole transcriptome data of 17 tongue squamous cell carcinoma and 6 TSCC cell lines (generated in house at ACTREC, unpublished data) using the HPVDetector. 3 of 17 primary tumors were found to be HPV18 positive. In addition, HPV18 reads were found in HEP-2 cell line, consistent with earlier reports in literature [196]. The HPV 18 genes (E1, E6, and E7) were validated in Hep2 cell line by PCR and Sanger sequencing (as shown in Additional Figure 1 and Table 2).



V. Figure 2: Quantitative representation by number of reads of HPV types detected in cervical tumors.

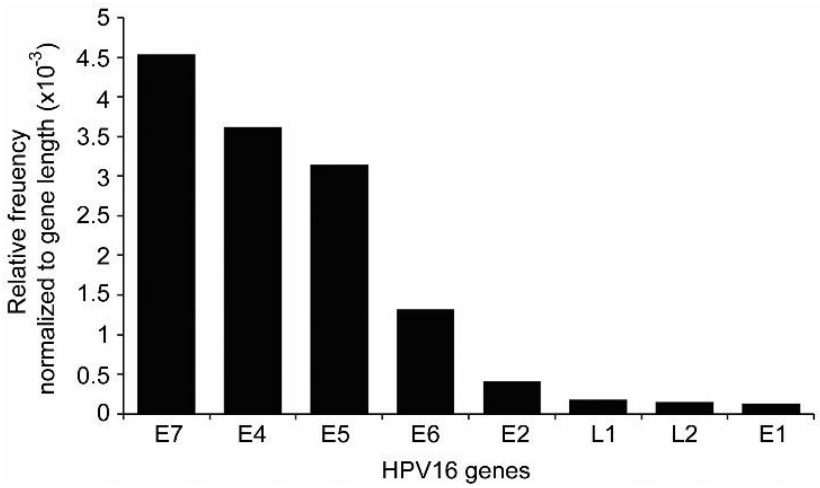
The graph represents distribution of total number of HPV reads per million of total reads for all HPV types detected across 17 cervical samples. HPV16 has highest number of reads across 17 samples followed by HPV (18,71,45,31,82,17,56,58,15,20) in the decreasing order of their read-counts.

Gall bladder and liposarcoma exome and whole genome data: Additionally, we analyzed 13 gall bladder cancer whole exome, 1 gall bladder cancer whole transcriptome and 1 liposarcoma whole genome sequence data (generated in house at ACTREC, unpublished data). No traces of HPV sequences were detected in these samples.



V. Figure 3: HPV gene integration frequency across different cervical cancer samples.

Heat map representation of HPV types detected across 17 cervical cancer samples. HPV 16 in 13 samples, HPV18 in 2 sample, HPV71 in 6 samples, HPV82 in 5 samples, HPV31 in 2 samples, HPV45 in 5 samples, and HPV17, 56, 58, 15, 20 were detected in one sample, each. Total coverage of the exome sequencing is indicated in last column "cov". Based on read count the abundance of HPV is graded in different samples: read count 1 to 10 as light grey; read count 10 to 50 as dark grey; and read count >50 as black.



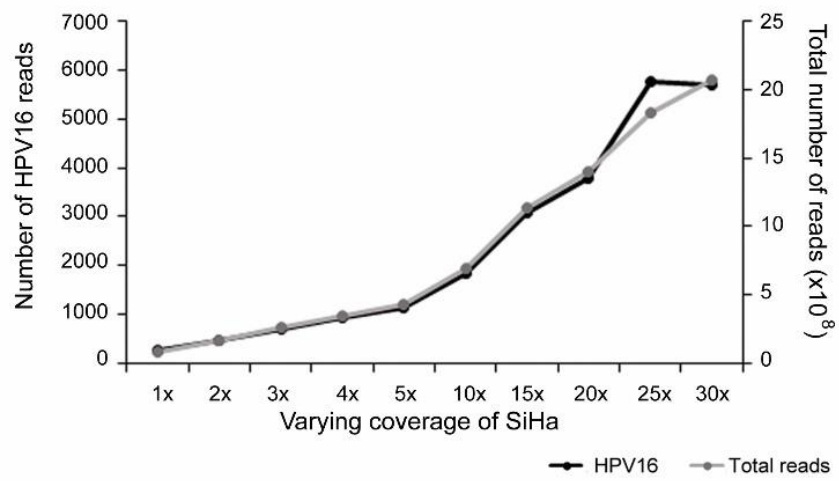
V. Figure 4: Relative frequency of integration of HPV genes in cervical carcinoma. HPV16 reads were annotated using inbuilt annotation module of the HPVDetector to identify the viral genes. Number of reads per viral gene were normalized to the gene length, and frequency reads for individual genes are plotted, as shown.

5.3.4 Assessment of specificity and sensitivity of HPVDetector

SiHa cell line developed from cervical squamous cell carcinoma patient represents single copy integration of HPV16 [197] We analyzed SiHa whole genome sequence using HPVDetector. Consistent with published report [170], HPVDetector could detect integration at chr13 intragenic location of KLF5 – KLF12 genes and other regions (Additional Table 1). The integration was validated by PCR followed by sequencing (Additional Figure 3).

Sensitivity: To determine the sensitivity of HPVDetector, we down-sampled the SiHa genome using a "down-sampling" method, a Picard Toolkit's DownsampleSam function (http://broadinstitute.github.io/picard/) [198] to generate varying coverage of the SiHa whole genome data ranging from 1x to 30x coverage, and analyzed using HPVDetector. Reads supporting presence of HPV reads linearly increased as a function of increasing coverage from 1X to 25X coverage. Beyond 25X, no significant increase in HPV reads were found beyond, suggesting saturation of the genome coverage (Figure 5).

Additionally, among primary tumors, two pairs of HPV56 reads detected by the HPVDetector in T9440 as described in Additional Table 1 were validated earlier by Luminex array and SPF1/2 [185]. Taken together, this suggests HPVDetector could detect reads with as low as 1X genome coverage with reads supported by as low as just two paired reads.



V. Figure 5: Detection of HPV 16 at varying coverage of SiHa whole genome sequencing data.

SiHa WGS data was down-sampled to the coverage of 1x, 2x, 3x, 4x, 5x, 10x, 15x, 20x, 25x, and 30x. HPV16 reads (black) were counted using HPVDetector at varying coverage and plotted along with total number of reads (grey).

Specificity: Having benchmarked the HPVDetector against SiHa for sensitivity, next we tweaked the SiHa whole genome sequence data to test specificity of the tool by taking reverse (not complement) of the SiHa genome to simulate as a random sequence but retaining composition of nucleotides and genome complexity, using an in-house perl script. We found no spurious HPV reads when the SiHa whole genome sequence was reversed, suggesting the HPVDetector is specific to detect true HPV traces. Further, to address issue of specificity among primary tumors we performed another round of

functional validation on tongue squamous tumors that were found HPV negative based on HPVDetector (Additional Table 1) and validated by My09/11 primers using genomic DNA. We analyzed the expression of HPV E6 (Additional Figure 2b) in these samples. All samples were found negative for HPV presence. This suggests that the tool has low false negative.

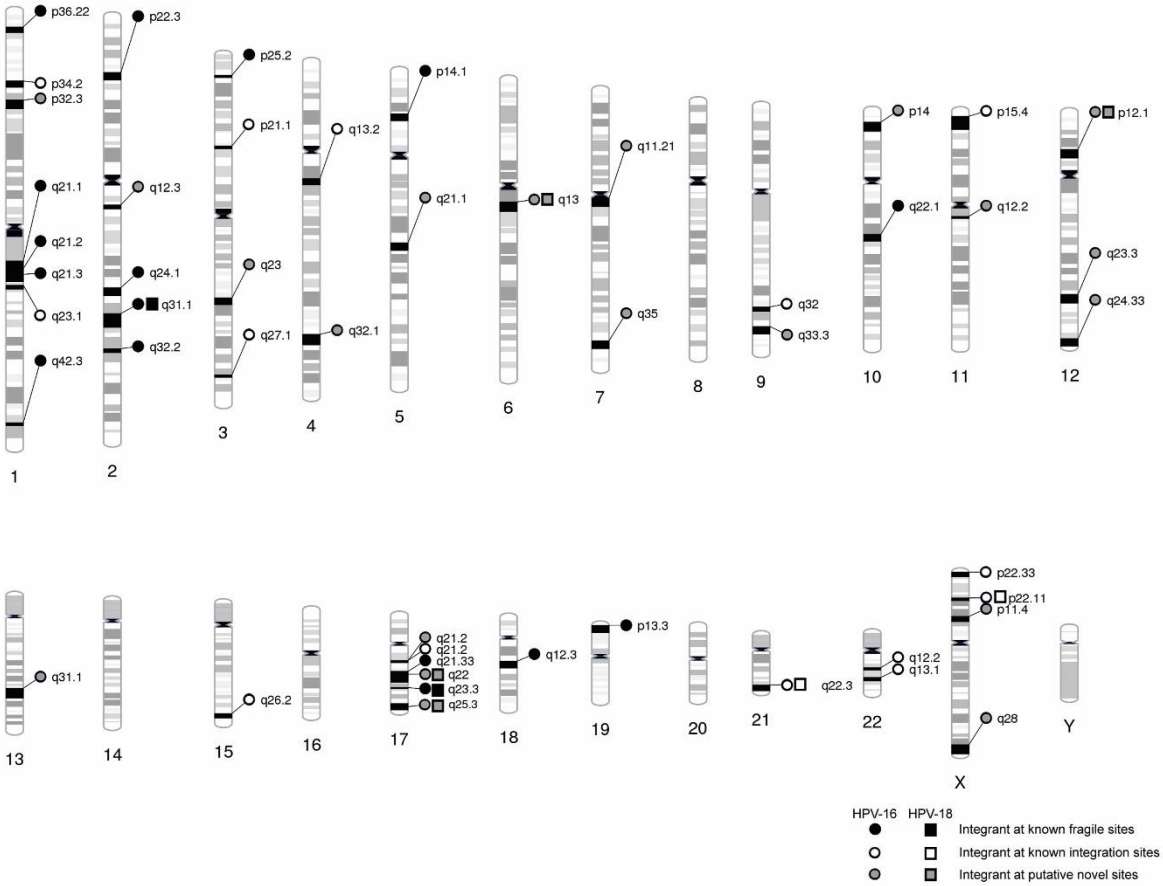
5.3.5 Annotation of the HPV genome integrated in the host genome

To enable accurate gene annotation of the HPV genome sequenced we prepared gene annotation database of 143 HPV types from PAVE database [179]. 32 reads of viral ORFs were found in 5 of 11 cervical tumors positive for HPV 16. Following the normalization for the total number of reads against the length of individual genes, the viral gene E7 was found be predominantly represented among the cervical tumors infected with HPV 16, followed by E4, E5 and E6, in decreasing order (Fig 4). Of these genes found to be enriched among all the integrant, it's interesting to note that the viral proteins E6 and E7 function as oncogenes by regulating the known human tumor suppressors, p53 and pRb, respectively [199, 200].

5.3.6 Determination of the HPV integration sites in the host genome

We identified 55 integration sites in 7 cervical cancer tumor samples T1099, T1123, T755, T887, T938, T1094, and T959 and 1 head and neck tumor sample using the HPVDetector (Additional Table 1). In this study, chromosomal loci 17q21, 3q27, 7q35, Xq28 were observed with higher frequency compared to other loci for HPV integration, as reported earlier [201]. Interesting to note, we found HPV integration in the following fragile regions- (1p, 1q, 2p, 2q, 3p, 3q, 4p, 4q, 5p, 5q, 7q, 9q, 10q, 11p, 11q, 12p, 12q, 13q, 15q, 17q, 18q, 22q, Xp and Xq) that are prone to chromosome breaks to facilitate

foreign DNA integration (Figure 6) [202]. More significantly the HPVDetector detected integration sites in T1123 and T755 samples identical integration site of HPV16 (chr1q42.3), HPV16 (chr3q23) respectively as reported in literature [203, 204]. Additionally, in T755 integration of HPV16 were found within the coding region at SLC25A36, a pyrimidine nucleotide carrier. These sites of integration were also determined in T755 and T1123 samples using APOT assay, as described earlier [185] (Additional Table 1).



V. Figure 6: Schematic representation of all HPV 16 and 18 integration sites in human genome detected across cervical cancer samples using HPVDetector.

Site of integration as determined by HPVDetector in cervical cancer samples is shown. HPV16 integration sites are depicted by circles and HPV18 by rectangles. Black, open and grey circles or rectangles represent integrations at known fragile sites, at known integration sites and at novel sites, respectively.

In total, we analyzed 116 exomes, 23 transcriptomes and 2 whole genome sequencing data, out of which we have detected presence of HPV in 20 exomes, 4 transcriptomes data (Table 1).

5.4 DISCUSSION

Human papilloma virus (HPV) accounts for the most common cause of all virus-associated human cancers. However, despite large-scale genome wide DNA sequencing efforts of the cancer genome there is no dedicated informatics tool to rapidly detect the presence of HPV in these genomes, in an exclusive manner. There are indeed a variety of gene integration finding tools available that can detect different pathogen insertions in the human genome such as ViralFusionSeq, VirusSeq, VirusFinder, Path-Seq, RINS, and ReadSCAN. These sophisticated tools though have their specific third party needs, necessitate intense computational infrastructure, cannot be run without specialized and advanced computational expertise of the researcher, and more importantly are not specific for HPV detection, *per se*—for e.g., lacks information to annotate the region of HPV genome to predict the integrated viral gene, of which some are known to function as oncogenes.

We present a new user-friendly in-silico tool "HPVDetector" as a unique tool to analyze NGS data to detect HPV sequences for non-computational biologists (Appendix 5). Using the HPVDetector tool we have detected 55 integration sites from cervical exome and Head & Neck transcriptome data set. The tool allowed us to perform a comprehensive analysis to generate the information for co-occurrence of HPV subtypes across cervical cancer patients that is known to affect the clinical outcome of the disease. Additionally, our finding of significant enrichment of viral gene E7 > E4 > E5 > E6 reads

among the cervical tumor samples using the inbuilt annotation module of the HPVDetector, is consistent with the known biology of HPV genes and their role in carcinogenesis, as E6 and E7 are known viral oncogenes. This unique feature of the HPVDetector with an inbuilt HPV annotation module could potentially be helpful to understand the function of other HPV ORFs with unknown function by studying their incidence against varying tumor stage and types. While the analysis of cervical tumors was restricted to its exome data set, a complete spectrum of the load of viral genes present in a sample can similarly be determined using the whole genome data as input to the HPVDetector.

HPVDetector demonstrated a low false negative and false positive rate that could detect reads with as low as 1X genome coverage with reads supported by as low as just two paired reads in this study undertaken. No viral reads were detected across 54 head and neck cancer samples of Indian origin, as reported earlier [193-195], but detected a low-risk HPV71 in a cell line that could be validated by performing MY09/11 PCR on the primary tumors as shown in Additional Figure 2. On the other hand, all the 4 HPV reads detected across different tumor types using HPVDetector could be validated by directed PCR followed by Sanger sequencing. One interesting utility of the HPV-Detector would be to explore for HPV reads in NGS data from different cancer types. We analyzed 13 gall bladder exomes, 1 gall bladder transcriptome and 1 liposarcoma whole-genome sequencing data using HPVDetector. No HPV reads were found in these samples in this study.

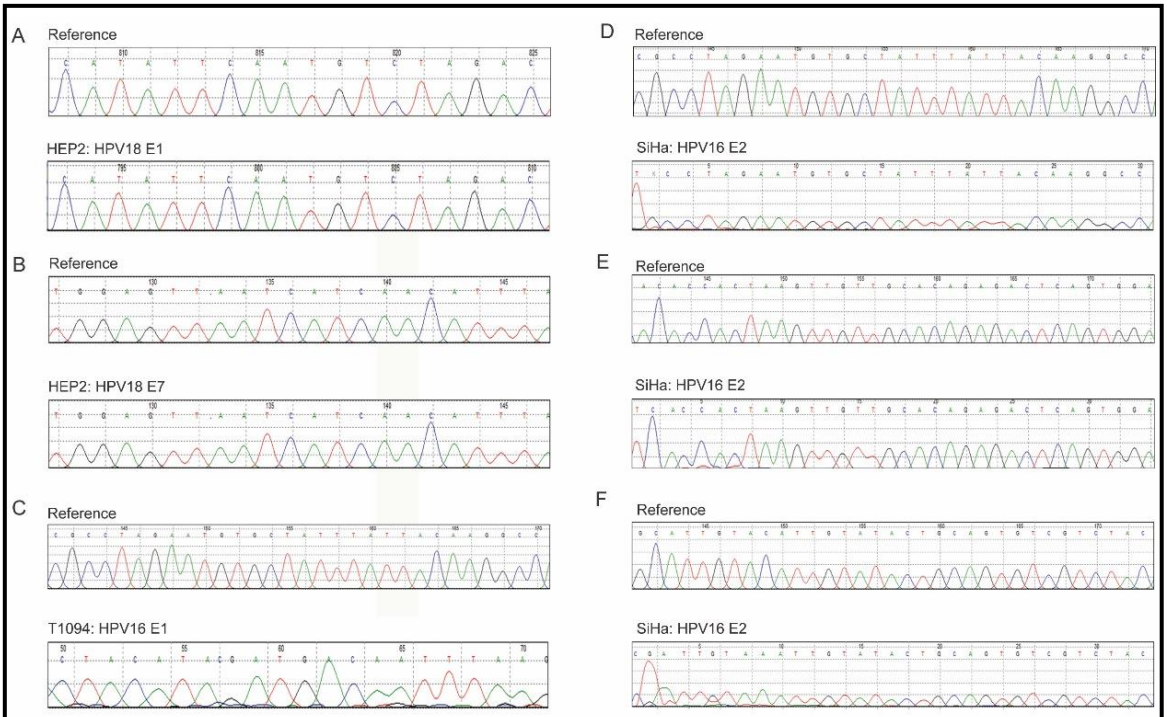
Another critical feature of the HPVDetector is determination of HPV integration sites at the host genome. These integrations are known to occur at preferred regions of the

genome [201, 205]. Using the integration site detection feature of the HPVDetector, we detected integration at various chromosomal locations (for eg- 1p, 2p, 2q, 3p, 3q), some with significant overlap to the known fragile sites in literature and at several novel sites as summarized in Figure 6 and Additional Table 1. In summary, HPVDetector is a simple yet precise and robust tool for detecting HPV from tumor samples using variety of next generation sequencing platforms including whole genome, whole exome and transcriptome. Two different modes (quick detection and integration mode) along with a GUI widen the usability of HPVDetector for biologists with minimal computational knowledge.

V. Table 1: Summary of HPV Detection in all samples.

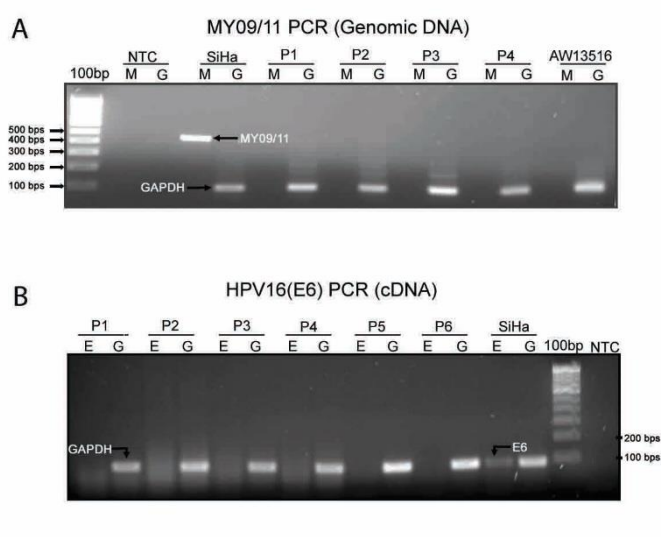
Study Type	**Samples Tested**	**Presence of virus in Samples**
Cervical cancer exome	36 (22 tumor, 14 paired normal)	18
SiHa cell line WGS	1	1
HNSCC cell lines exome	7	1
TSCC exome	47 (24 tumor, 23 paired normal)	0
Gall bladder exome	26 (13 tumor, 13 paired normal)	1
TSCC transcriptome	17 (11 tumor, 6 paired normal)	3
HNSCC cell lines transcriptome	5	1
Gall bladder transcriptome	1	0
Liposarcoma WGS	1	0
Total number of samples	**141**	**25**

5.5 Additional Supporting Data



V. Additional Figure S1: Sanger sequencing based validation of HPV integrant

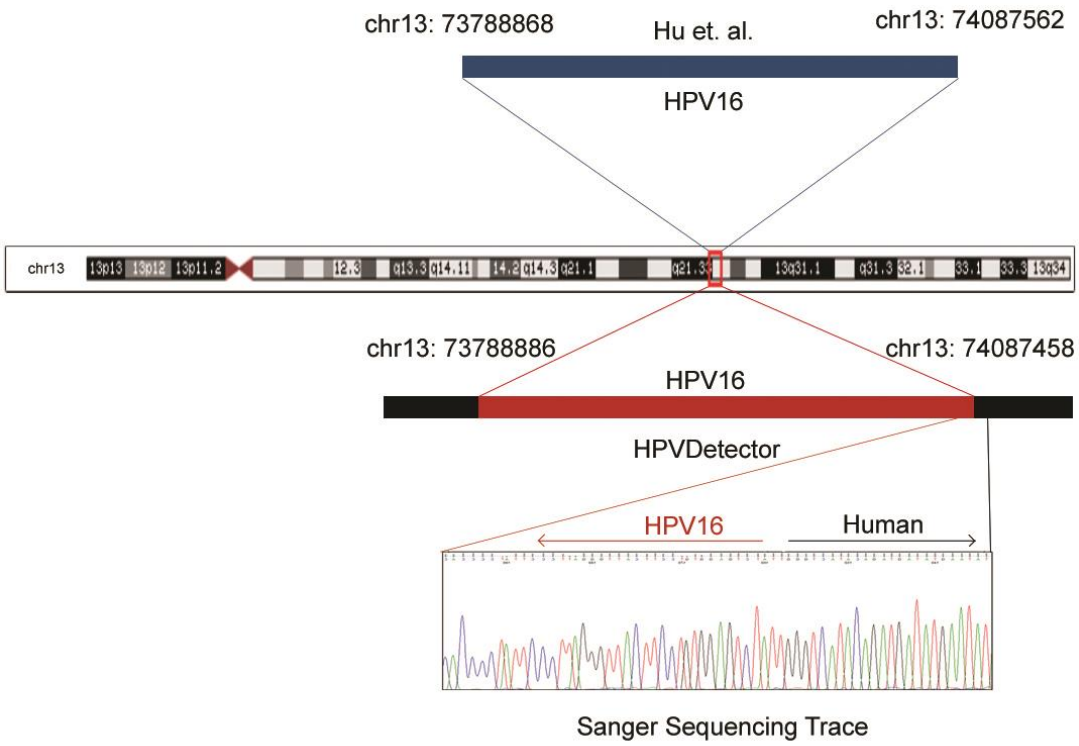
Presence of HPV was validated by sequencing HPV genes in the cervical and HNSCC cell lines using Sanger sequencing. Sequencing traces were analyzed and compared to GenBank reference database using mutation surveyor. Each of the sequence trace is shown with alignment to corresponding reference sequence: (A) E1 gene and (B) E7 gene from cDNA of Hep2 cell line, (C) E1 gene from genomic DNA of cervical tumor sample (D-F) HPV16 E2 reads in SiHa cells.



V. Additional Figure S2: Validation of HPV negative TSCC samples by PCR.

(A) PCR was performed on genomic DNA using MY09/11 primers. SiHa cell line was used as positive control for HPV and AW1356 cell line as negative control and GAPDH as internal control for genomic DNA. NTC - no template control; M - MY09/11; G - GAPDH; P1 to P4 - TSCC patient samples. MY09/11 amplifies 450bp and GAPDH 100bp indicated using black arrow.

(B) PCR of E6 gene of HPV16 was performed in cDNA form TSCC tumor samples along with SiHa cell line as positive control. P1 to P6 - TSCC patient samples; E - E6 band; G – GAPDH band, NTC - no template control. E6 amplifies 86bp and GAPDH 100bp indicated using black arrow.



V. Additional Figure S3: Integration of HPV16 on chr13 of human genome.

HPV16 integration in human chromosome 13 q21.3 near KLF5-KLF12 genes has been shown as reported in literature and detected by HPVDetector. Sanger sequencing trace is shown from the right junction of HPV16 integration.

V. Additional Table S1: List of integration sites in 7 cervical cancer exomes, SiHa cell line and 1 head & neck transcriptome samples.

Sample Name & Sequencing Type	HPV Type	HPV Position	HPV Gene	Human Gene	Genomic Loci
T1099- Exome	HPV16	1198	E1	PDE4DIP	1q21.1
	HPV16	1460	E1	NBPF8	1q21.2
	HPV16	3050	E2	COL5A2	2q32.2
	HPV16	4998	L2	STAB1	3p21.1
	HPV16	3551	E2	MCCC1	3q27.1
	HPV16	3551	E4	MCCC1	3q27.1
	HPV16	5506	L2	CDH9	5p14.1
	HPV16	2140	E1	SBDS	7q11.21
	HPV16	3555	E2	Null	7q35
	HPV16	3555	E4	Null	7q35
	HPV16	1022	E1	DFNB31	9q32
	HPV16	6700	L1	PPA1	10q22.1
	HPV16	7008	L1	PGA4	11q12.2
	HPV16	5283	L2	Null	11p15.4
	HPV16	2808	E1	ITGA3	17q21.33
	HPV16	2808	E2	ITGA3	17q21.33
	HPV16	4246	L2	AP1B1	22q12.2
	HPV16	4036	E5	SMCR7L	22q13.1
	HPV16	1047	E1	Null	Xp11.4
T1123- Exome	HPV16	6522	L1	PODN	1p32.3
	HPV16	4244	L2	LYST	1q42.3
	HPV16	5672	L1	Null	9q33.3
	HPV16	1597	E1	Null	12q23.3
T755- Exome	HPV16	7257	Null	LINC00486	2p22.3
	HPV16	4975	L2	SLC25A36	3q23
	HPV16	4839	L2	Null	12q24.33
	HPV16	7525	Null	Null	15q26.2
	HPV16	4861	L2	ARSF	Xp22.33
	HPV16	3542	E2	Null	Xq28
	HPV16	3542	E4	Null	Xq28
T887- Exome	HPV16	3059	E2	CD5L	1q23.1
	HPV16	4271	L2	KIF1B	1p36.22
	HPV16	2258	E1	IQSEC1	3p25.2
	HPV16	1984	E1	Null	4q13.2
	HPV16	973	E1	PAM	5q21.1
	HPV16	6087	L1	Null	13q31.1
	HPV16	7179	Null	RIT2	18q12.3
T938- Exome	HPV31	6629	L1	HIVEP3	1p34.2
	HPV31	239	E6	RANBP2	2q12.3

T1094- Exome	HPV16	2416	E1	S100A9	1q21.3
	HPV16	4193	Null	NR4A2	2q24.1
	HPV16	478	E6	MLLT1	19p13.3
T959- Exome	HPV16	223	E6	Null	4q32.1
HEP2-Transcriptome	HPV18	429	E6	CYBRD1	2q31.1
	HPV18	637	E7	SP3	2q31.1
	HPV18	1099	E1	FAM135A	6q13
	HPV18	57	Null	EEF1A1	6q13
	HPV18	1482	E1	C10orf47	10p14
	HPV18	644	E7	GOLT1B	12p12.1
	HPV18	2404	E1	SMARCE1	17q21.2
	HPV18	129	E6	PRR11	17q22
	HPV18	1309	E1	DDX5	17q23.3
	HPV18	1945	E1	LGALS3BP	17q25.3
	HPV18	1719	E1	ABCG1	21q22.3
	HPV18	920	E1	SAT1	Xp22.11
SiHa	HPV16	616	E7	FHIT	3p14.2
	HPV16	2529	E1	FHIT	3p14.2
	HPV16	2027	E1	PRIM2	6p11.2
	HPV16	198	E6	DPY19L4	8q22.1
	HPV16	3694	E2	Null	13q22.1
	HPV16	2777	E2	Null	13q22.1
	HPV16	454	E6	TM9SF2	13q32.3
	HPV16	298	E6	ELL	19p13.11

VI. GENERAL DISCUSSION

Global cancer genomics profiling efforts such as The Cancer Genome Atlas (TCGA) and International Cancer Genomics Consortium (ICGC) has led to the identification of several hundred therapeutic targets in human cancers. While these findings are being translated to the clinics, an emerging challenge posed due to ethnic diversity – disparity of therapeutic targets and their clinical response lends itself to the need for further research. My thesis is a step towards genomic characterization of tumors among patients of Indian origin known to be an admixture of Ancestral North Indians (closer to non-European Caucasian) and Ancestral South Indians [61, 62]. As a first step, I profiled primary lung adenocarcinoma tumors using high-throughput next-generation sequencing followed by mass-spectrometry based assay to present the first spectrum of therapeutically relevant alterations and *FGFR3* as a novel therapeutically relevant target in lung adenocarcinoma of Indian origin. Secondly, I set to perform integrated genomics analysis to present a proof-of-principle strategy to identify biologically relevant genomic alterations from a limited set of samples. This section of my thesis describes an alternative but effective way to enrich genomic dataset by performing posterior filtering strategy as adopted and described for few cell lines. In the third section of thesis, I describe the making of a computational tool HPVDetector to help biologists and clinicians analyze unlimited set of data freely available at various public resources or their own data generated using high-throughput NGS platform for identification and analysis of Human Papillomavirus (HPV) in tumor samples. Minimal third party dependency for installation of the tool and easy to use graphical user interface (GUI) widen the usability of HPVDetector for non-computational experts.

Primary lung tumor profiling of Indian origin

Recent population genomics studies have revealed much genomic diversity among the Indian population [61, 62]. The underlying genomic diversity poses a considerable challenge for developing population specific therapeutics. As a significant initiative in this direction, The India Project Team – International Cancer Genomic Consortium (IPT-ICGC) has undertaken one of such major effort to understand genome of head and neck cancer patients of Indian origin. IPT-ICGC team has profiled 50 gingiva-buccal cancers – a subtype of head & neck cancer using whole exome sequencing and presented the first landscape of somatic alterations underlying the disease [149]. In a similar approach, I characterize lung cancer, another major cancer type prevalent in India. I performed targeted sequencing of actionable alterations using next-generation sequencing of 45 lung adenocarcinoma samples, the most common lung cancer subtype. This endeavor led to identification of novel therapeutically relevant alterations in fibroblast growth factor receptor-3 (*FGFR3)* and those known to be altered in lung cancer: *EGFR, KRAS, ALK*, *AKT1*, *PIK3CA, NOTCH4* and *ERBB2* [50, 56, 59, 99]. These mutations were validated by using mass-spectrometry based approach in a larger and additional set of 363 lung adenocarcinoma samples. Recurrent mutations of *FGFR3* observed among 5.5% of samples were transforming and sensitive to small molecule inhibitors using *in-vitro* and *in-vivo* methods. Interestingly, the *FGFR3* mutations were found to be significantly higher in a proportion of younger patients and show a trend towards better overall survival, compared to patients lacking actionable alterations or those harboring *KRAS* mutations. This work establishes *FGFR3* as potential therapeutic targets in lung adenocarcinoma and provides a rational for designing clinical trials to establish the clinical utility of FGFR inhibitors in lung cancer of Indian origin.

Aberrant FGFR signaling has been identified with prominent alterations frequency in several cancer types. In a recent study of total 4853 tumors of 17 different types, average 7% were harboring alterations in FGFR family [81]. Notably, 20-50 % of bladder cancer and 5 % of lung squamous cell carcinoma patients are reported with *FGFR3* alterations [89, 206], 10-20 % of lung squamous cell carcinoma, 10% of breast cancer and head & neck cancer patients harbors *FGFR1* alterations [108, 207, 208], 10% of endometrial carcinoma and 5-10 % gastric cancer patients harbors *FGFR2* alterations [209, 210]. More than a dozen FGFR inhibitors have been developed and reached to various stages of clinical trials. Most importantly, four tyrosine kinase inhibitors targeting FGFRs have been recently approved by Food and Drug Administration of USA: Pazopanib is approved for renal cell carcinoma and sarcoma; Ponatinib is approved for drug-resistant chronic myelogenous leukemia and Philadelphia chromosome-positive acute lymphocytic leukemia; Regorafenib is approved for advanced colorectal carcinoma and drug-resistant gastrointestinal stromal tumors; and Lenvatinib which is approved for iodine-refractory, well-differentiated thyroid carcinoma [211, 212]. Additionally, selective FGFR inhibitors AZD4547, BGJ398 (Infigratinib) and several others have shown promising preliminary results with manageable toxicities and significant antitumor activity in molecularly selected patients of bladder cancer, lung squamous cell carcinoma, glioblastoma and cholangiocarcinoma [91, 211, 213]. In continuation to the efforts of bringing FGFR targeted therapeutics to clinics, this study provides a rational for testing FGFR inhibitors in lung adenocarcinoma.

In summary, this study establishes the first spectrum of therapeutically relevant genomic alterations consisting eight oncogenes in lung adenocarcinoma of Indian origin. About 43% of the lung adenocarcinoma patients can benefit from targeted therapy, which is

readily available or is in active development. The major limitation of this study is that this study was designed to profile actionable oncogenic alterations using targeted sequencing. As evident form this study, actionable oncogenic spectrum of lung adenocarcinoma of Indian origin harbors ethnicity specific variations. This study lays the logical foundation to systematically analyze lung cancer of Indian origin using deeper genome wide approaches in an unbiased manner to further our understanding of ethnicity specific genomic diversity.

Integrated genomics of head & neck cancer cell lines

Research methodologies for cancer genomics analysis have evolved from conventional single patient-single gene based studies to large cohort-multi-omics based genome wide studies, largely due to reduced sequencing cost and time by next-generation sequencing platforms. Moreover, the identification of an entire landscape of relatively low frequency genomic alterations in cancer genome has further necessitated large cohort based studies. According to a recent analysis of more than 5000 tumor-normal pairs representing 21 cancer types, detection of genomic alterations of 10% population frequency in head & neck cancer requires more than 150 cancer sample set [121]. The larger sample size based studies (in the order of several hundred to thousands) have been undertaken through global consortium based projects such as TCGA, ICGC, CCLE etc. While a wealth of data describing underlying alterations of Caucasian population is available as public resource, data describing such alterations of non-Caucasian or non-European Caucasian has been sparsely addressed. This is partially due to the lack of computational and technical resources to analyze similar number of samples in such population. Limitations of horizontal expansion of sample set can be compensated, in some ways, by integrated analysis using multiple genomics platforms on fewer samples to enhance the analytical

power to discover biologically relevant alteration. The principle of integrated genomics analysis can be simplified as follows: same causal phenotypic effect can be achieved using multiple routes of genomic alteration such as inactivating point mutation in a tumor suppressor, copy number loss of the gene or differential methylation of promoter leading to under-expression. Integrating genomic information at gene or pathway level could help identification of these multiple routes of causal genomic alterations. Additionally, different genomic information could be used to filter against each other resulting into lesser false positives.

Here, I used integrated genomics approach for analysis of NT8E, AW13516, AW8507 and OT9 cell lines derived from HNSCC patients of Indian origin. I integrated copy number, gene expression and mutations information to identify genes which are under biological regulation via multiple routes. Interestingly, I identified *TP53*, *PTEN*, *HRAS*, *MET* and *CASP8* as major altered HNSCC hallmark genes [114, 149] and additional set of 34 novel candidate genes altered in HNSCC. To further expand this analysis, I used TCGA HNSCC data set of 279 patient samples in which three different classes of genes could be observed. First, genes which are primarily altered by point mutations; second, genes which are sparsely altered by amplification or over-expression in addition to mutations; and third, genes which are primarily altered by multiple routes (amplification or over-expression and mutations). The genes altered by single route were observed with frequency similar to those reported by conventional single platform analysis while genes altered by multiple routes were identified with higher cumulative alteration frequency than previously reported single platform based studies, in turn establishing more profound role of the gene in disease etiology than previously reported.

In summary, this study provides a proof-of-principle for identification of biologically relevant genes from a smaller sample set using integration of multi-omics datasets. The study also reveals the genes which are preferentially altered by multiple routes demanding systematic analysis of these genes to enhance our understanding of their biological role. The major limitation of this approach is that, genes with multiple level of biological regulation could be identified, while genes altered by only single type of regulation could be missed out. Thus, identified gene set in cell lines and TCGA cohort could be an underrepresentation of all biologically relevant genes in HNSCC. Nevertheless, this study emphasizes filtering strategy based on biologically guided integrative genomics approach for identification of relevant genes from fewer samples. Further, expansion of cell line panel with development of additional cell lines is recommended to aid in identification and testing of diversified therapeutic solutions.

Computational tool development for HPV analysis using NGS data

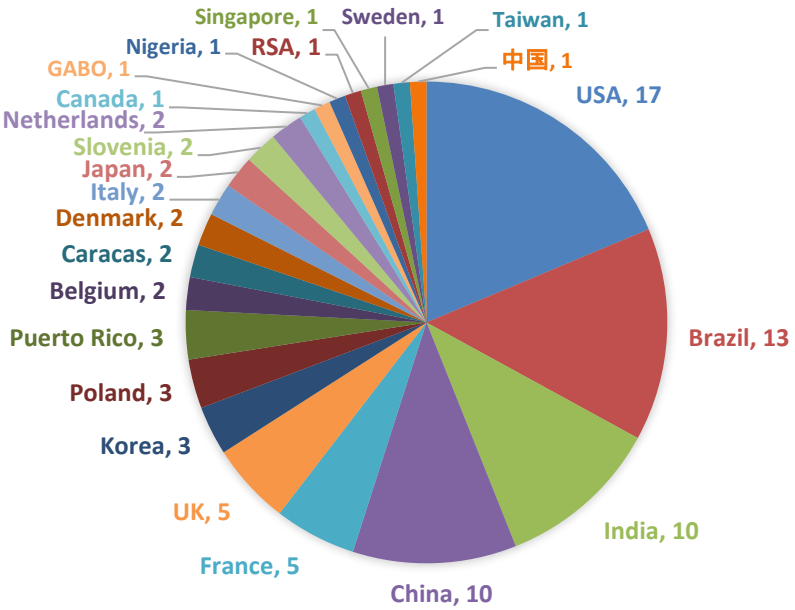
While high-throughput analysis of primary tumors and cell lines have been delivering promising results, it has been deterrent to biologists and clinicians. In my continued pursuit to simplify genomics analysis using few samples, I describe here a user-friendly tool to help non-computational biologist to parse genome wide dataset and derive biological insight. I developed an open source computational tool 'HPVDetector' powered by graphical user interface (GUI) to aid biologists and clinicians in HPV study utilizing high-throughput NGS data generated from cancer tissues or cell lines without going through computational complexity involved in the analysis. Having minimal third party needs, HPVDetector can be easily installed on any modern Linux/UNIX computer by non-experienced user. User can launch the tool in either graphical user interface mode, for ease of use or in command line mode for advanced use. In either mode, HPVDetector

provides a quick detect module to quickly identify presence of HPV types in the given data set (generally taking a few minutes to an hour for an exome sequencing data set of ~100x coverage on a modern personal computer) and another integration module taking more time and computational resources to identify integration sites of HPV in human genome. In both the modes, HPVDetector presents the output with detailed annotations of HPV and human genome regions for downstream biological analysis and interpretation.

HPV associated cancers are common all across the world but the incidence rate, specifically of head & neck and cervical cancer is higher in India [214]. This necessitates detailed analysis of HPV associated cancers in India to understand underlying biology. We analyzed 116 whole exomes, 23 whole transcriptomes and 2 whole genome data sets of various cancer types using HPVDetector, out of which we identified HPV in 20 exomes and 4 transcriptomes of cervical and head and neck cancer tumor samples of Indian origin. We observed viral oncogene E7 frequently integrated at known 17q21, 3q27, 7q35, Xq28 [185, 204] and novel sites of integration in the human genome. Interestingly, high-risk HPV 16 and 31 were observed to be mutually exclusive compared to low-risk HPV 71.

This analysis demonstrate utility of HPVDetector for various biological analysis. As on 1st February 2017, HPVDetector has been downloaded and used by at-least 83 researchers across 23 different countries (VI. Figure 1). To provide user support, we have developed user group (https://groups.google.com/forum/#!forum/hpvdetector) through which users can post and discuss about the tool. We have received feedback from users regarding their experience and to cite one of the response: Ms. Aditi Kulkarni from

University of Michigan posted "…I would also like to tell you that the tool developed by your group is way better than the existing tools (VirusFinder etc). It is very user friendly and the outputs are also easy to interpret. The Readme file is very well written and easy to follow!! (Unlike most tools that do not have a good tutorial). Our group has really benefited from this tool and we plan on using it a lot more for our future data sets…".



VI. Figure 1: HPVDetector usage statistics (as on 1st February, 2017)

Number of HPVDetector downloads are shown by the countries (total = 91).

Overall, this thesis work provides first therapeutically relevant spectrum of genomic alterations for lung adenocarcinoma, a proof-of-principle integrated genomic analysis workflow for identification of biologically relevant alterations form fewer samples, and an open source bioinformatics tool HPVDetector coupled with user friendly graphical user interface for effective utilization of next-generation sequencing platforms for HPV analysis by non-computational experts. I anticipate the increment in knowledge made by this thesis would have significant impact on cancer research, therapeutics and beyond.

VII. BIBLIOGRAPHY

1. Mortality, G.B.D. and C. Causes of Death, *Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013.* Lancet, 2015. **385**(9963): p. 117-71.
2. Zeng, C., et al., *Disparities by Race, Age, and Sex in the Improvement of Survival for Major Cancers: Results From the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) Program in the United States, 1990 to 2010.* JAMA Oncol, 2015. **1**(1): p. 88-96.
3. Siegel, R., et al., *Cancer statistics, 2014.* CA Cancer J Clin, 2014. **64**(1): p. 9-29.
4. Stone, J.B. and L.M. DeAngelis, *Cancer-treatment-induced neurotoxicity--focus on newer treatments.* Nat Rev Clin Oncol, 2016. **13**(2): p. 92-105.
5. Argyriou, A.A., et al., *Chemotherapy-induced peripheral neurotoxicity (CIPN): an update.* Crit Rev Oncol Hematol, 2012. **82**(1): p. 51-77.
6. Bentzen, S.M., *Preventing or reducing late side effects of radiation therapy: radiobiology meets molecular pathology.* Nat Rev Cancer, 2006. **6**(9): p. 702-13.
7. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
8. Upadhyay, P., R. Dwivedi, and A. Dutt, *Applications of next-generation sequencing in cancer.* CURRENT SCIENCE, 2014. **107**(5): p. 795.
9. Weinstein, I.B. and A.K. Joe, *Mechanisms of disease: Oncogene addiction--a rationale for molecular targeting in cancer therapy.* Nat Clin Pract Oncol, 2006. **3**(8): p. 448-57.
10. Weinstein, I.B., *Disorders in cell circuitry during multistage carcinogenesis: the role of homeostasis.* Carcinogenesis, 2000. **21**(5): p. 857-64.
11. Weinstein, I.B., *Cancer. Addiction to oncogenes--the Achilles heal of cancer.* Science, 2002. **297**(5578): p. 63-4.
12. Torti, D. and L. Trusolino, *Oncogene addiction as a foundational rationale for targeted anti-cancer therapy: promises and perils.* EMBO Mol Med, 2011. **3**(11): p. 623-36.
13. Druker, B.J., et al., *Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome.* N Engl J Med, 2001. **344**(14): p. 1038-42.
14. Slamon, D.J., et al., *Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2.* N Engl J Med, 2001. **344**(11): p. 783-92.
15. Lynch, T.J., et al., *Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib.* N Engl J Med, 2004. **350**(21): p. 2129-39.
16. Thatcher, N., et al., *Gefitinib plus best supportive care in previously treated patients with refractory advanced non-small-cell lung cancer: results from a randomised, placebo-controlled, multicentre study (Iressa Survival Evaluation in Lung Cancer).* Lancet, 2005. **366**(9496): p. 1527-37.
17. Flaherty, K.T., et al., *Inhibition of mutated, activated BRAF in metastatic melanoma.* N Engl J Med, 2010. **363**(9): p. 809-19.
18. Sudhakar, A., *History of Cancer, Ancient and Modern Treatment Methods.* J Cancer Sci Ther, 2009. **1**(2): p. 1-4.
19. Chin, L., J.N. Andersen, and P.A. Futreal, *Cancer genomics: from discovery science to personalized medicine.* Nat Med, 2011. **17**(3): p. 297-303.
20. Cancer Genome Atlas Research, N., et al., *The Cancer Genome Atlas Pan-Cancer analysis project.* Nat Genet, 2013. **45**(10): p. 1113-20.
21. Yap, T.A. and P. Workman, *Exploiting the Cancer Genome: Strategies for the Discovery and Clinical Development of Targeted Molecular Therapeutics.* Annual Review of Pharmacology and Toxicology, 2012. **52**(1): p. 549-573.
22. Bollig-Fischer, A., et al., *Racial diversity of actionable mutations in non-small cell lung cancer.* J Thorac Oncol, 2015. **10**(2): p. 250-5.

23. O'Donnell, P.H. and M.E. Dolan, *Cancer pharmacoethnicity: ethnic differences in susceptibility to the effects of chemotherapy.* Clin Cancer Res, 2009. **15**(15): p. 4806-14.
24. Swanton, C., *Intratumor heterogeneity: evolution through space and time.* Cancer Res, 2012. **72**(19): p. 4875-82.
25. Gonzalez de Castro, D., et al., *Personalized cancer medicine: molecular diagnostics, predictive biomarkers, and drug resistance.* Clin Pharmacol Ther, 2013. **93**(3): p. 252-9.
26. Chandrani, P. and A. Dutt, *Domain Specific Targeting of Cancer*, in *Nuclear Signaling Pathways and Targeting Transcription in Cancer*. 2014, Springer. p. 299-310.
27. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603-7.
28. Garnett, M.J., et al., *Systematic identification of genomic markers of drug sensitivity in cancer cells.* Nature, 2012. **483**(7391): p. 570-5.
29. Shoemaker, R.H., *The NCI60 human tumour cell line anticancer drug screen.* Nat Rev Cancer, 2006. **6**(10): p. 813-23.
30. Yang, W., et al., *Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells.* Nucleic Acids Res, 2013. **41**(Database issue): p. D955-61.
31. Lamb, J., *The Connectivity Map: a new tool for biomedical research.* Nat Rev Cancer, 2007. **7**(1): p. 54-60.
32. Schlabach, M.R., et al., *Cancer proliferation gene discovery through functional genomics.* Science, 2008. **319**(5863): p. 620-4.
33. Johannessen, C.M., P.A. Clemons, and B.K. Wagner, *Integrating phenotypic small-molecule profiling and human genetics: the next phase in drug discovery.* Trends Genet, 2015. **31**(1): p. 16-23.
34. Chin, L., et al., *Making sense of cancer genomic data.* Genes Dev, 2011. **25**(6): p. 534-55.
35. Boehm, J.S. and W.C. Hahn, *Towards systematic functional characterization of cancer genomes.* Nat Rev Genet, 2011. **12**(7): p. 487-98.
36. Yamori, T., *Panel of human cancer cell lines provides valuable database for drug discovery and bioinformatics.* Cancer Chemother Pharmacol, 2003. **52 Suppl 1**: p. S74-9.
37. Sharma, S.V., D.A. Haber, and J. Settleman, *Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents.* Nat Rev Cancer, 2010. **10**(4): p. 241-53.
38. Shih, C. and R.A. Weinberg, *Isolation of a transforming sequence from a human bladder carcinoma cell line.* Cell, 1982. **29**(1): p. 161-9.
39. Daley, G.Q. and D. Baltimore, *Transformation of an interleukin 3-dependent hematopoietic cell line by the chronic myelogenous leukemia-specific P210bcr/abl protein.* Proc Natl Acad Sci U S A, 1988. **85**(23): p. 9312-6.
40. Warmuth, M., et al., *Ba/F3 cells and their use in kinase drug discovery.* Curr Opin Oncol, 2007. **19**(1): p. 55-60.
41. Zhang, J., et al., *International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data.* Database (Oxford), 2011. **2011**: p. bar026.
42. Alioto, T.S., et al., *A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing.* Nat Commun, 2015. **6**: p. 10001.
43. Yang, Y., et al., *Databases and web tools for cancer genomics study.* Genomics Proteomics Bioinformatics, 2015. **13**(1): p. 46-50.
44. Ding, L., et al., *Expanding the computational toolbox for mining cancer genomes.* Nat Rev Genet, 2014. **15**(8): p. 556-70.
45. Blankenberg, D., et al., *Galaxy: a web-based genome analysis tool for experimentalists.* Curr Protoc Mol Biol, 2010. **Chapter 19**: p. Unit 19 10 1-21.
46. 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.
47. Society, A.C., *Cancer Facts & Figures 2012.* Atlanta: American Cancer Society, 2012.
48. Siegel, R., et al., *Cancer statistics, 2011: the impact of eliminating socioeconomic and racial disparities on premature cancer deaths.* CA Cancer J Clin, 2011. **61**(4): p. 212-36.

49. Noronha, V., et al., *EGFR mutations in Indian lung cancer patients: clinical correlation and outcome to EGFR targeted therapy.* PLoS One, 2013. **8**(4): p. e61561.
50. Choughule, A., et al., *Coexistence of KRAS mutation with mutant but not wild-type EGFR predicts response to tyrosine-kinase inhibitors in human lung cancer.* Br J Cancer, 2014. **111**(11): p. 2203-4.
51. Soda, M., et al., *Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer.* Nature, 2007. **448**(7153): p. 561-6.
52. Bergethon, K., et al., *ROS1 rearrangements define a unique molecular class of lung cancers.* J Clin Oncol, 2012. **30**(8): p. 863-70.
53. Ju, Y.S., et al., *Fusion of KIF5B and RET transforming gene in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing.* Genome Research, 2011.
54. Barlesi, F., et al., *Biomarkers (BM) France: Results of routine EGFR, HER2, KRAS, BRAF, PI3KCA mutations detection and EML4-ALK gene fusion assessment on the first 10,000 non-small cell lung cancer (NSCLC) patients (pts).* J Clin Oncol, 2013. **31**(15S): p. 486s.
55. Kris, M.G., et al., *Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs.* JAMA, 2014. **311**(19): p. 1998-2006.
56. Chougule, A., et al., *Frequency of EGFR mutations in 907 lung adenocarcioma patients of Indian ethnicity.* PLoS One, 2013. **8**(10): p. e76164.
57. Paez, J.G., et al., *EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy.* Science, 2004. **304**(5676): p. 1497-500.
58. Pao, W., et al., *EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib.* Proc Natl Acad Sci U S A, 2004. **101**(36): p. 13306-11.
59. Imielinski, M., et al., *Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing.* Cell, 2012. **150**(6): p. 1107-1120.
60. Patel, J.N., *Cancer pharmacogenomics: implications on ethnic diversity and drug response.* Pharmacogenet Genomics, 2015. **25**(5): p. 223-30.
61. Reich, D., et al., *Reconstructing Indian population history.* Nature, 2009. **461**(7263): p. 489-94.
62. Moorjani, P., et al., *Genetic evidence for recent population mixture in India.* Am J Hum Genet, 2013. **93**(3): p. 422-38.
63. Basu, A., N. Sarkar-Roy, and P.P. Majumder, *Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure.* Proc Natl Acad Sci U S A, 2016. **113**(6): p. 1594-9.
64. Harismendy, O., et al., *Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing.* Genome Biol, 2011. **12**(12): p. R124.
65. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
66. 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.
67. Cibulskis, K., et al., *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.* Nat Biotechnol, 2013. **31**(3): p. 213-9.
68. Wong, S.Q., et al., *Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing.* BMC Med Genomics, 2014. **7**: p. 23.
69. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.
70. Upadhyay, P., et al., *TMC-SNPdb: An Indian germline variant dataset derived from whole exome sequence.* Database (Oxford), In Press.
71. Liu, X., et al., *dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs.* Hum Mutat, 2016. **37**(3): p. 235-41.
72. Johannessen, C.M., et al., *COT drives resistance to RAF inhibition through MAP kinase pathway reactivation.* Nature, 2010. **468**(7326): p. 968-72.
73. Chandrani, P., et al., *Integrated genomics approach to identify biologically relevant alterations in fewer samples.* BMC Genomics, 2015. **16**(1): p. 936.

74. Di Veroli, G.Y., et al., *An automated fitting procedure and software for dose-response curves with multiphasic features.* Sci Rep, 2015. **5**: p. 14701.
75. Walker, J.M., *The bicinchoninic acid (BCA) assay for protein quantitation.* Methods Mol Biol, 1994. **32**: p. 5-8.
76. Therneau, T. *A Package for Survival Analysis in S*. 2016; Available from: http://cran.r-project.org/package=survival.
77. Dardis, C. *survMisc: Miscellaneous Functions for Survival Data*. 2016; Available from: https://cran.r-project.org/web/packages/survMisc/index.html.
78. Gao, J., et al., *Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.* Sci Signal, 2013. **6**(269): p. pl1.
79. Ramos, A.H., et al., *Amplification of chromosomal segment 4q12 in non-small cell lung cancer.* Cancer Biol Ther, 2009. **8**(21): p. 2042-50.
80. Looyenga, B.D., et al., *STAT3 is activated by JAK2 independent of key oncogenic driver mutations in non-small cell lung carcinoma.* PLoS One, 2012. **7**(2): p. e30820.
81. Helsten, T., et al., *The FGFR Landscape in Cancer: Analysis of 4,853 Tumors by Next-Generation Sequencing.* Clin Cancer Res, 2016. **22**(1): p. 259-67.
82. Govindan, R., et al., *Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers.* Cell, 2012. **150**(6): p. 1121-1134.
83. Wang, R., et al., *Comprehensive investigation of oncogenic driver mutations in Chinese non-small cell lung cancer patients.* Oncotarget, 2015. **6**(33): p. 34300-8.
84. Hunt, J.D., et al., *Differences in KRAS mutation spectrum in lung cancer cases between African Americans and Caucasians after occupational or environmental exposure to known carcinogens.* Cancer Epidemiol Biomarkers Prev, 2002. **11**(11): p. 1405-12.
85. Goriely, A., et al., *Activating mutations in FGFR3 and HRAS reveal a shared genetic origin for congenital disorders and testicular tumors.* Nat Genet, 2009. **41**(11): p. 1247-52.
86. Hafner, C., et al., *FGFR3 mutations in seborrheic keratoses are already present in flat lesions and associated with age and localization.* Mod Pathol, 2007. **20**(8): p. 895-903.
87. Kato, S., et al., *The Conundrum of Genetic "Drivers" in Benign Conditions.* J Natl Cancer Inst, 2016. **108**(8).
88. Rosty, C., et al., *Clinical and biological characteristics of cervical neoplasias with FGFR3 mutation.* Mol Cancer, 2005. **4**(1): p. 15.
89. Liao, R.G., et al., *Inhibitor-sensitive FGFR2 and FGFR3 mutations in lung squamous cell carcinoma.* Cancer Res, 2013. **73**(16): p. 5195-205.
90. Tomlinson, D.C., C.D. Hurst, and M.A. Knowles, *Knockdown by shRNA identifies S249C mutant FGFR3 as a potential therapeutic target in bladder cancer.* Oncogene, 2007. **26**(40): p. 5889-99.
91. Guagnano, V., et al., *Discovery of 3-(2,6-dichloro-3,5-dimethoxy-phenyl)-1-{6-[4-(4-ethyl-piperazin-1-yl)-phenylamin o]-pyrimidin-4-yl}-1-methyl-urea (NVP-BGJ398), a potent and selective inhibitor of the fibroblast growth factor receptor family of receptor tyrosine kinase.* J Med Chem, 2011. **54**(20): p. 7066-83.
92. Sacher, A.G., et al., *Association Between Younger Age and Targetable Genomic Alterations and Prognosis in Non-Small-Cell Lung Cancer.* JAMA Oncol, 2016. **2**(3): p. 313-20.
93. Nishii, T., et al., *Clinicopathological features and EGFR gene mutation status in elderly patients with resected non-small-cell lung cancer.* BMC Cancer, 2014. **14**: p. 610.
94. Cerami, E., et al., *The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.* Cancer Discov, 2012. **2**(5): p. 401-4.
95. Campbell, J.D., et al., *Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas.* Nat Genet, 2016. **48**(6): p. 607-16.
96. Barlesi, F., et al., *Routine molecular profiling of patients with advanced non-small-cell lung cancer: results of a 1-year nationwide programme of the French Cooperative Thoracic Intergroup (IFCT).* Lancet, 2016. **387**(10026): p. 1415-26.
97. Gadgeel, S.M., et al., *Impact of race in lung cancer: analysis of temporal trends from a surveillance, epidemiology, and end results database.* Chest, 2001. **120**(1): p. 55-63.
98. Hardy, D., et al., *Racial disparities and treatment trends in a large cohort of elderly black and white patients with nonsmall cell lung cancer.* Cancer, 2009. **115**(10): p. 2199-211.

99. Cancer Genome Atlas Research, N., *Comprehensive molecular profiling of lung adenocarcinoma.* Nature, 2014. **511**(7511): p. 543-50.
100. Lim, S.M., et al., *Targeted sequencing identifies genetic alterations that confer primary resistance to EGFR tyrosine kinase inhibitor (Korean Lung Cancer Consortium).* Oncotarget, 2016. **7**(24): p. 36311-36320.
101. Ahmad, I., et al., *K-Ras and beta-catenin mutations cooperate with Fgfr3 mutations in mice to promote tumorigenesis in the skin and lung, but not in the bladder.* Dis Model Mech, 2011. **4**(4): p. 548-55.
102. Arai, D., et al., *Characterization of the cell of origin and propagation potential of the fibroblast growth factor 9-induced mouse model of lung adenocarcinoma.* J Pathol, 2015. **235**(4): p. 593-605.
103. Yin, Y., et al., *Rapid induction of lung adenocarcinoma by fibroblast growth factor 9 signaling through FGF receptor 3.* Cancer Res, 2013. **73**(18): p. 5730-41.
104. Yin, Y., et al., *Inhibition of fibroblast growth factor receptor 3-dependent lung adenocarcinoma with a human monoclonal antibody.* Dis Model Mech, 2016. **9**(5): p. 563-71.
105. Liu, X., et al., *Clinical significance of fibroblast growth factor receptor-3 mutations in bladder cancer: a systematic review and meta-analysis.* Genet Mol Res, 2014. **13**(1): p. 1109-20.
106. Milowsky, M.I., et al., *Phase 2 trial of dovitinib in patients with progressive FGFR3-mutated or FGFR3 wild-type advanced urothelial carcinoma.* Eur J Cancer, 2014. **50**(18): p. 3145-52.
107. Sequist, L.V., et al., *Phase I study of BGJ398, a selective pan-FGFR inhibitor in genetically preselected advanced solid tumors*, in *American Association for Cancer Research 2014 Congress*. 2014: San Diego, CA, USA.
108. Dutt, A., et al., *Inhibitor-sensitive FGFR1 amplification in human non-small cell lung cancer.* PLoS One, 2011. **6**(6): p. e20351.
109. Dutt, A., et al., *Drug-sensitive FGFR2 mutations in endometrial carcinoma.* Proc Natl Acad Sci U S A, 2008. **105**(25): p. 8713-7.
110. Weiss, J., et al., *Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer.* Sci Transl Med, 2010. **2**(62): p. 62ra93.
111. 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.
112. 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.
113. Stransky, N., et al., *The mutational landscape of head and neck squamous cell carcinoma.* Science, 2011. **333**(6046): p. 1157-60.
114. Cancer Genome Atlas, N., *Comprehensive genomic characterization of head and neck squamous cell carcinomas.* Nature, 2015. **517**(7536): p. 576-82.
115. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-8.
116. Dees, N.D., et al., *MuSiC: identifying mutational significance in cancer genomes.* Genome Res, 2012. **22**(8): p. 1589-98.
117. Gonzalez-Perez, A., et al., *IntOGen-mutations identifies cancer drivers across tumor types.* Nat Methods, 2013. **10**(11): p. 1081-2.
118. Hodis, E., et al., *A landscape of driver mutations in melanoma.* Cell, 2012. **150**(2): p. 251-63.
119. 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.
120. 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.
121. Lawrence, M.S., et al., *Discovery and saturation analysis of cancer genes across 21 tumour types.* Nature, 2014. **505**(7484): p. 495-501.
122. Akavia, U.D., et al., *An integrated approach to uncover drivers of cancer.* Cell, 2010. **143**(6): p. 1005-17.
123. Natrajan, R. and P. Wilkerson, *From integrative genomics to therapeutic targets.* Cancer Res, 2013. **73**(12): p. 3483-8.

124. 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.
125. 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.
126. Kristensen, V.N., et al., *Principles and methods of integrative genomic analyses in cancer.* Nat Rev Cancer, 2014. **14**(5): p. 299-313.
127. 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.
128. 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.
129. Popova, T., et al., *Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays.* Genome Biol, 2009. **10**(11): p. R128.
130. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.
131. 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.
132. Forbes, S.A., et al., *The Catalogue of Somatic Mutations in Cancer (COSMIC).* Curr Protoc Hum Genet, 2008. **Chapter 10**: p. Unit 10 11.
133. Ramos, A.H., et al., *Oncotator: cancer variant annotation tool.* Hum Mutat, 2015. **36**(4): p. E2423-9.
134. 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.
135. Choi, Y. and A.P. Chan, *PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels.* Bioinformatics, 2015.
136. 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.
137. 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.
138. 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.
139. 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.
140. 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.
141. 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.
142. Krzywinski, M., et al., *Circos: an information aesthetic for comparative genomics.* Genome Res, 2009. **19**(9): p. 1639-45.
143. Roschke, A.V., et al., *Karyotypic complexity of the NCI-60 drug-screening panel.* Cancer Res, 2003. **63**(24): p. 8634-47.
144. 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.
145. 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.
146. 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.
147. 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.
148. Dodd, L.E., 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): p. 2216-25.

149. 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.
150. Davies, H., et al., *Somatic mutations of the protein kinase gene family in human lung cancer.* Cancer Res, 2005. **65**(17): p. 7591-5.
151. Rusan, M., Y.Y. Li, and P.S. Hammerman, *Genomic landscape of human papillomavirus-associated cancers.* Clin Cancer Res, 2015. **21**(9): p. 2009-19.
152. 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.
153. 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).
154. Ntziachristos, P., et al., *From fly wings to targeted cancer therapies: a centennial for notch signaling.* Cancer Cell, 2014. **25**(3): p. 318-34.
155. Hua, F., et al., *TRB3 interacts with SMAD3 promoting tumor cell migration and invasion.* J Cell Sci, 2011. **124**(Pt 19): p. 3235-46.
156. Manning, G., et al., *The protein kinase complement of the human genome.* Science, 2002. **298**(5600): p. 1912-34.
157. 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.
158. Chung, G.T., et al., *Identification of a recurrent transforming UBR5-ZNF423 fusion gene in EBV-associated nasopharyngeal carcinoma.* J Pathol, 2013. **231**(2): p. 158-67.
159. Zhu, C.Q., et al., *Amplification of telomerase (hTERT) gene is a poor prognostic marker in non-small-cell lung cancer.* Br J Cancer, 2006. **94**(10): p. 1452-9.
160. Munoz, N., et al., *Epidemiologic classification of human papillomavirus types associated with cervical cancer.* N Engl J Med, 2003. **348**(6): p. 518-27.
161. Shukla, S., *Infection of human papillomaviruses in cancers of different human organ sites* Indian J Med Res, 2009. **130**: p. 222-233.
162. Kleter, B., et al., *Novel short-fragment PCR assay for highly sensitive broad-spectrum detection of anogenital human papillomaviruses.* Am J Pathol, 1998. **153**(6): p. 1731-9.
163. Abreu, A.L., et al., *A review of methods for detect human Papillomavirus infection.* Virol J, 2012. **9**: p. 262.
164. Brink, A.A., P.J. Snijders, and C.J. Meijer, *HPV detection methods.* Dis Markers, 2007. **23**(4): p. 273-81.
165. Mendez, F., et al., *Cervical coinfection with human papillomavirus (HPV) types and possible implications for the prevention of cervical cancer by HPV vaccines.* J Infect Dis, 2005. **192**(7): p. 1158-65.
166. Trottier, H., et al., *Human papillomavirus infections with multiple types and risk of cervical neoplasia.* Cancer Epidemiol Biomarkers Prev, 2006. **15**(7): p. 1274-80.
167. Yi, X., et al., *Development and validation of a new HPV genotyping assay based on next-generation sequencing.* Am J Clin Pathol, 2014. **141**(6): p. 796-804.
168. Johansson, H., et al., *Metagenomic sequencing of "HPV-negative" condylomas detects novel putative HPV types.* Virology, 2013. **440**(1): p. 1-7.
169. Ameur, A., et al., *Comprehensive profiling of the vaginal microbiome in HIV positive women using massive parallel semiconductor sequencing.* Sci Rep, 2014. **4**: p. 4398.
170. Hu, Z., et al., *Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism.* Nat Genet, 2015. **47**(2): p. 158-63.
171. Li, W., et al., *HIVID: an efficient method to detect HBV integration using low coverage sequencing.* Genomics, 2013. **102**(4): p. 338-44.
172. Xu, B., et al., *Multiplex Identification of Human Papillomavirus 16 DNA Integration Sites in Cervical Carcinomas.* PLoS One, 2013. **8**(6): p. e66693.
173. Li, J.W., et al., *ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution.* Bioinformatics, 2013. **29**(5): p. 649-51.

174. Chen, Y., et al., *VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue.* Bioinformatics, 2013. **29**(2): p. 266-7.
175. Wang, Q., P. Jia, and Z. Zhao, *VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data.* PLoS One, 2013. **8**(5): p. e64465.
176. Kostic, A.D., et al., *PathSeq: software to identify or discover microbes by deep sequencing of human tissue.* Nat Biotechnol, 2011. **29**(5): p. 393-6.
177. Bhaduri, A., et al., *Rapid identification of non-human sequences in high-throughput sequencing datasets.* Bioinformatics, 2012. **28**(8): p. 1174-5.
178. Naeem, R., M. Rashid, and A. Pain, *READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation.* Bioinformatics, 2013. **29**(3): p. 391-2.
179. Van Doorslaer, K., et al., *The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis.* Nucleic Acids Res, 2013. **41**(Database issue): p. D571-8.
180. Stajich, J.E., et al., *The Bioperl toolkit: Perl modules for the life sciences.* Genome Res, 2002. **12**(10): p. 1611-8.
181. Karolchik, D., et al., *The UCSC Table Browser data retrieval tool.* Nucleic Acids Res, 2004. **32**(Database issue): p. D493-6.
182. 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.
183. Baay, M.F., et al., *Comprehensive study of several general and type-specific primer pairs for detection of human papillomavirus DNA by PCR in paraffin-embedded cervical carcinomas.* J Clin Microbiol, 1996. **34**(3): p. 745-7.
184. Adler, K., T. Erickson, and M. Bobrow, *High sensitivity detection of HPV-16 in SiHa and CaSki cells utilizing FISH enhanced by TSA.* Histochem Cell Biol, 1997. **108**(4-5): p. 321-4.
185. Das, P., et al., *HPV genotyping and site of viral integration in cervical cancers in Indian women.* PLoS One, 2012. **7**(7): p. e41012.
186. Liaw, K.L., et al., *A prospective study of human papillomavirus (HPV) type 16 DNA detection by polymerase chain reaction and its association with acquisition and persistence of other HPV types.* J Infect Dis, 2001. **183**(1): p. 8-15.
187. Chaturvedi, A.K., et al., *Human papillomavirus infection with multiple types: pattern of coinfection and risk of cervical disease.* J Infect Dis, 2011. **203**(7): p. 910-20.
188. Vaccarella, S., et al., *Concurrent infection with multiple human papillomavirus types: pooled analysis of the IARC HPV Prevalence Surveys.* Cancer Epidemiol Biomarkers Prev, 2010. **19**(2): p. 503-10.
189. Harari, A., Z. Chen, and R.D. Burk, *Human papillomavirus genomics: past, present and future.* Curr Probl Dermatol, 2014. **45**: p. 1-18.
190. Schiffman, M., G. Clifford, and F.M. Buonaguro, *Classification of weakly carcinogenic human papillomavirus types: addressing the limits of epidemiology at the borderline.* Infect Agent Cancer, 2009. **4**: p. 8.
191. Chaudhary, A.K., et al., *Role of human papillomavirus and its detection in potentially malignant and malignant head and neck lesions: updated review.* Head Neck Oncol, 2009. **1**: p. 22.
192. Pannone, G., et al., *The role of human papillomavirus in the pathogenesis of head & neck squamous cell carcinoma: an overview.* Infect Agent Cancer, 2011. **6**: p. 4.
193. 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.
194. Tsimplaki, E., et al., *Prevalence and expression of human papillomavirus in 53 patients with oral tongue squamous cell carcinoma.* Anticancer Res, 2014. **34**(2): p. 1021-5.
195. Siebers, T.J., et al., *No high-risk HPV detected in SCC of the oral tongue in the absolute absence of tobacco and alcohol--a case study of seven patients.* Oral Maxillofac Surg, 2008. **12**(4): p. 185-8.

196. Ogura, H., et al., *Human papillomavirus type 18 DNA in so-called HEP-2, KB and FL cells--further evidence that these cells are HeLa cell derivatives.* Cell Mol Biol (Noisy-le-grand), 1993. **39**(5): p. 463-7.
197. el Awady, M.K., et al., *Molecular analysis of integrated human papillomavirus 16 sequences in the cervical cancer cell line SiHa.* Virology, 1987. **159**(2): p. 389-98.
198. Meynert, A.M., et al., *Variant detection sensitivity and biases in whole genome and exome sequencing.* BMC Bioinformatics, 2014. **15**: p. 247.
199. Yim, E.K. and J.S. Park, *The role of HPV E6 and E7 oncoproteins in HPV-associated cervical carcinogenesis.* Cancer Res Treat, 2005. **37**(6): p. 319-24.
200. Lu, D.W., S.K. El-Mofty, and H.L. Wang, *Expression of p16, Rb, and p53 proteins in squamous cell carcinomas of the anorectal region harboring human papillomavirus DNA.* Mod Pathol, 2003. **16**(7): p. 692-9.
201. Thorland, E.C., et al., *Common fragile sites are preferential targets for HPV16 integrations in cervical tumors.* Oncogene, 2003. **22**(8): p. 1225-37.
202. Smith, D.I., et al., *Common fragile sites, extremely large genes, neural development and cancer.* Cancer Lett, 2006. **232**(1): p. 48-57.
203. Wentzensen, N., S. Vinokurova, and M. von Knebel Doeberitz, *Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract.* Cancer Res, 2004. **64**(11): p. 3878-84.
204. Schmitz, M., et al., *Non-random integration of the HPV genome in cervical cancer.* PLoS One, 2012. **7**(6): p. e39632.
205. Matovina, M., et al., *Identification of human papillomavirus type 16 integration sites in high-grade precancerous cervical lesions.* Gynecol Oncol, 2009. **113**(1): p. 120-7.
206. Gust, K.M., et al., *Fibroblast Growth Factor Receptor 3 Is a Rational Therapeutic Target in Bladder Cancer.* Molecular Cancer Therapeutics, 2013. **12**(7): p. 1245-1254.
207. Reis-Filho, J.S., et al., *FGFR1 emerges as a potential therapeutic target for lobular breast carcinomas.* Clin Cancer Res, 2006. **12**(22): p. 6652-62.
208. Freier, K., et al., *Recurrent FGFR1 amplification and high FGFR1 protein expression in oral squamous cell carcinoma (OSCC).* Oral Oncol, 2007. **43**(1): p. 60-6.
209. Dutt, A., et al., *Drug-sensitive FGFR2 mutations in endometrial carcinoma.* Proceedings of the National Academy of Sciences, 2008. **105**(25): p. 8713-8717.
210. Matsumoto, K., et al., *FGFR2 gene amplification and clinicopathological features in gastric cancer.* Br J Cancer, 2012. **106**(4): p. 727-32.
211. Touat, M., et al., *Targeting FGFR Signaling in Cancer.* Clin Cancer Res, 2015. **21**(12): p. 2684-94.
212. CenterWatch. *FDA Approved Drugs for Oncology*. 2016 [cited 2016 13-07-2016]; Available from: https://www.centerwatch.com/drug-information/fda-approved-drugs/therapeutic-area/12/oncology.
213. Gavine, P.R., et al., *AZD4547: an orally bioavailable, potent, and selective inhibitor of the fibroblast growth factor receptor tyrosine kinase family.* Cancer Res, 2012. **72**(8): p. 2045-56.
214. Dikshit, R., et al., *Cancer mortality in India: a nationally representative survey.* Lancet, 2012. **379**(9828): p. 1807-1816.

VIII. APPENDIX

Appendix 1: Details of variants identified by NGS in lung adenocarcinoma patients (N=45) by pooled sequencing strategy.

Sr. no.	Gene	Mutation	Chromosome	Start position	End position	Variant Classification	Reference base	Mutated base
1	ABL1	G250R	9	133738348	133738348	Missense Mutation	G	A
2	APC	R640W	5	112170822	112170822	Missense Mutation	C	T
3	APC	P1369S	5	112175396	112175396	Missense Mutation	C	T
4	APC	P1420L	5	112175550	112175550	Missense Mutation	C	T
5	APC	L1482F	5	112175735	112175735	Missense Mutation	C	T
6	APC	M1583V	5	112176038	112176038	Missense Mutation	A	G
7	BRAF	V471I	7	140481397	140481397	Missense Mutation	C	T
8	BRAF	G464R	7	140481418	140481418	Missense Mutation	C	T
9	CDH1	P160S	16	68842417	68842417	Missense Mutation	C	T
10	CDH1	D400N	16	68847276	68847276	Missense Mutation	G	A
11	CDH1	V685M	16	68857418	68857418	Missense Mutation	G	A
12	CDKN2A	G35E	9	21974723	21974723	Missense Mutation	C	T
13	CSF1R	P566L	5	149441342	149441342	Missense Mutation	G	A
14	EGFR	ELREA746del	7	55242464	55242479	In Frame Del	AGGAATTAAGAGAAGC	A
15	EGFR	S768I	7	55249005	55249005	Missense Mutation	G	T
16	EGFR	V774A	7	55249023	55249023	Missense Mutation	T	C
17	EGFR	G779S	7	55249037	55249037	Missense Mutation	G	A
18	EGFR	R831C	7	55259433	55259433	Missense Mutation	C	T
19	EGFR	V834A	7	55259443	55259443	Missense Mutation	T	C
20	EGFR	L858M	7	55259514	55259514	Missense Mutation	C	A
21	EGFR	L858R	7	55259515	55259515	Missense Mutation	T	G
22	FGFR3	C228R	4	1803413	1803413	Missense Mutation	T	C
23	FGFR3	S249C	4	1803568	1803568	Missense Mutation	C	G
24	FGFR3	P283S	4	1803669	1803669	Missense Mutation	C	T
25	FGFR3	S679F	4	1808278	1808278	Missense Mutation	C	T
26	FGFR3	G691R	4	1808313	1808313	Missense Mutation	G	A
27	FLT3	G617E	13	28608116	28608116	Missense Mutation	C	T
28	GJC3	E187K	7	99526685	99526685	Missense Mutation	C	T
29	JAK2	R564Q	9	5072541	5072541	Missense Mutation	G	A
30	JAK2	S591L	9	5072622	5072622	Missense Mutation	C	T
31	KIT	P551L	4	55593586	55593586	Missense Mutation	C	T
32	KIT	G565R	4	55593627	55593627	Missense Mutation	G	A
33	KIT	A784T	4	55598153	55598153	Missense Mutation	G	A
34	KRAS	A59T	12	25380283	25380283	Missense Mutation	C	T
35	KRAS	P34L	12	25398218	25398218	Missense Mutation	G	A
36	KRAS	G12V	12	25398284	25398284	Missense Mutation	C	A
37	LPPR3	A245T	19	814532	814532	Missense Mutation	C	T
38	MET	G1183D	7	116422067	116422067	Missense Mutation	G	A
39	MLH1	E172K	3	37050365	37050365	Missense Mutation	G	A
40	MLH1	R265C	3	37058999	37058999	Missense Mutation	C	T
41	MLH1	L658F	3	37090083	37090083	Missense Mutation	C	T
42	MLH1	A681T	3	37090446	37090446	Missense Mutation	G	A
43	MSH2	S676L	2	47703527	47703527	Missense Mutation	C	T
44	MSH2	S743L	2	47705428	47705428	Missense Mutation	C	T
45	MSH2	D758N	2	47705472	47705472	Missense Mutation	G	A
46	MSH2	C822R	2	47707840	47707840	Missense Mutation	T	C
47	NF2	C133Y	22	30038225	30038225	Missense Mutation	G	A

48	NF2	R196Q	22	30051653	30051653	Missense Mutation	G	A
49	NF2	R341Q	22	30067837	30067837	Missense Mutation	G	A
50	NF2	E372K	22	30067929	30067929	Missense Mutation	G	A
51	NF2	R418C	22	30069387	30069387	Missense Mutation	C	T
52	NOTCH1	T2511I	9	139390659	139390659	Missense Mutation	G	A
53	NOTCH1	L2510F	9	139390663	139390663	Missense Mutation	G	A
54	NOTCH1	R1594Q	9	139399362	139399362	Missense Mutation	C	T
55	PDGFRA	P155L	4	55129930	55129930	Missense Mutation	C	T
56	PDGFRA	H570Y	4	55141062	55141062	Missense Mutation	C	T
57	PDGFRA	P589Q	4	55141120	55141120	Missense Mutation	C	A
58	PTEN	P30L	10	89653791	89653791	Missense Mutation	C	T
59	PTEN	C105Y	10	89692830	89692830	Missense Mutation	G	A
60	PTEN	H123Y	10	89692883	89692883	Missense Mutation	C	T
61	PTEN	A126T	10	89692892	89692892	Missense Mutation	G	A
62	PTEN	G127R	10	89692895	89692895	Missense Mutation	G	A
63	PTEN	G129E	10	89692902	89692902	Missense Mutation	G	A
64	PTEN	R130Q	10	89692905	89692905	Missense Mutation	G	A
65	PTEN	V133I	10	89692913	89692913	Missense Mutation	G	A
66	PTEN	S227F	10	89717655	89717655	Missense Mutation	C	T
67	PTEN	G251D	10	89717727	89717727	Missense Mutation	G	A
68	PTEN	P339S	10	89720864	89720864	Missense Mutation	C	T
69	RB1	R798W	13	49039407	49039407	Missense Mutation	C	T
70	RET	R721Q	10	43612057	43612057	Missense Mutation	G	A
71	RET	A756V	10	43612162	43612162	Missense Mutation	C	T
72	RET	R873W	10	43615538	43615538	Missense Mutation	C	T
73	RET	S891L	10	43615593	43615593	Missense Mutation	C	T
74	RET	R912W	10	43617397	43617397	Missense Mutation	C	T
75	RUNX1	R223H	21	36206763	36206763	Missense Mutation	C	T
76	RUNX1	R135K	21	36252877	36252877	Missense Mutation	C	T
77	RUNX1	P86S	21	36259154	36259154	Missense Mutation	G	A
78	TP53	A347T	17	7573988	7573988	Missense Mutation	C	T
79	TP53	R280G	17	7577100	7577100	Missense Mutation	T	C
80	TP53	C277F	17	7577108	7577108	Missense Mutation	C	A
81	TP53	A276V	17	7577111	7577111	Missense Mutation	G	A
82	TP53	R273H	17	7577120	7577120	Missense Mutation	C	T
83	TP53	E271K	17	7577127	7577127	Missense Mutation	C	T
84	TP53	S269I	17	7577132	7577132	Missense Mutation	C	A
85	TP53	G262V	17	7577153	7577153	Missense Mutation	C	A
86	TP53	T230I	17	7577592	7577592	Missense Mutation	G	A
87	TP53	T230A	17	7577593	7577593	Missense Mutation	T	C
88	TP53	Y220N	17	7578191	7578191	Missense Mutation	A	T
89	TP53	D208N	17	7578227	7578227	Missense Mutation	C	T
90	TP53	P191H	17	7578277	7578277	Missense Mutation	G	T
91	TP53	R175H	17	7578406	7578406	Missense Mutation	C	T
92	TP53	E171G	17	7578418	7578418	Missense Mutation	T	C
93	TP53	C141Y	17	7578508	7578508	Missense Mutation	C	T
94	TP53	A138T	17	7578518	7578518	Missense Mutation	C	T
95	TP53	L130F	17	7578542	7578542	Missense Mutation	G	A
96	TP53	T125M	17	7579313	7579313	Missense Mutation	G	A
97	TP53	T118I	17	7579334	7579334	Missense Mutation	G	A
98	VHL	V74A	3	10183752	10183752	Missense Mutation	T	C
99	VHL	G144E	3	10188288	10188288	Missense Mutation	G	A

Appendix 2: List of mutations assayed using single base extension mass-spectrometry in lung adenocarcinoma (N=363).

Gene	Mutation
AKT1	E17K
APC	S237F
BRAF	K601E
BRAF	V600E
CDKN2A	E69*
CDKN2A	R58Stop
CDKN2A	R80Stop
EGFR	769-770insDSA
EGFR	770-771insSVD
EGFR	ELREA746del
EGFR	L858R
EGFR	T790M
ERBB2	780-781insGSP
ERBB2	S760F
FGFR1	Q309K
FGFR2	E116K
FGFR2	S252W
FGFR3	G691R
FGFR3	R248H
FGFR3	S249C
FGFR3	S679F
FGFR4	A634D
KRAS	G12C
KRAS	G12V
KRAS	G13D

Gene	Mutation
MSH2	S743L
NF1	A1417D
NF1	E1436G
NF1	P1421L
NF1	Q2324R
NF2	A193T
NF2	R341Q
NOTCH1	L2510F
NOTCH1	T2511I
NRAS	G12S
PDGFRA	V824I
PIK3CA	E542K
PIK3CA	E545K
PTEN	P30L
PTEN	Q245*
PTEN	R130G
PTEN	R233*
RB1	D479N
RB1	G248D
RB1	G840R
SMAD4	G473R
TP53	E171G
TP53	F328S
TP53	T125M

Appendix 3: Clinicopathological status of lung adenocarcinoma patients (N=363).

Sr. no.	Patient ID	Tumor Stage	Age	Sex	Smoking status	EGFR	KRAS	FGFR3	EML4-ALK	AKT1	PIK3CA	FGFR4	ERBB2
1	AD1588	IV	61	M	NA								
2	AD0549	IV	59	F	Non-smoker	E746							
3	AD0550	IV	60	F	NA	E746							
4	AD1690	IV	68	M	NA								
5	AD1572	IV	63	F	NA								
6	AD0552	IV	42	F	Non-smoker								
7	AD0290	IV	41	F	Non-smoker	E746		S249C					
8	AD0553	IV	50	M	Non-smoker	E746				E17K			
9	AD0264	IV	38	M	Non-smoker								
10	AD0262	IV	62	F	Non-smoker								
11	AD0554	IV	53	F	NA	E746							
12	AD0556	IV	64	M	NA	E746	G12C						
13	AD0557	IV	50	M	Smoker		G12C						
14	AD0559	IV	56	M	Non-smoker								
15	AD0288	IV	41	F	Non-smoker								
16	AD0560	IV	53	M	Smoker		G12C						
17	AD0561	IV	75	M	Non-smoker								
18	AD0304	IV	68	M	Smoker		G12V						
19	AD0564	IV	68	M	Smoker								
20	AD0311	IV	69	M	Smoker		G12C						
21	AD0565	IV	55	F	Non-smoker	E746							
22	AD0327	IV	45	M	Non-smoker		G12C						
23	AD0566	IV	63	F	Non-smoker	E746							
24	AD0567	IV	39	F	Non-smoker		G12C						
25	AD0568	IV	78	F	Non-smoker								
26	AD0569	IV	79	F	Non-smoker								
27	AD0347	IV	60	F	Non-smoker								
28	AD0573	IV	60	M	Non-smoker								
29	AD0575	IV	55	M	Non-smoker								
30	AD0350	IV	56	F	Non-smoker	E746	G12V	S249C					
31	AD0577	IV	42	F	Non-smoker								
32	AD0578	IV	57	F	Non-smoker		G12C						
33	AD0579	IV	72	M	Non-smoker		G12V						
34	AD0580	IV	40	F	Non-smoker		G12C						
35	AD1662	IV	30	F	NA								
36	AD0581	IV	73	M	Non-smoker	E746							
37	AD0582	IV	62	F	Non-smoker		G12C						
38	AD0365	IV	65	F	Non-smoker		G12V						
39	AD0375	IV	38	M	NA								
40	AD0586	IV	73	M	Non-smoker		G12V						
41	AD0322	IV	53	F	Non-smoker								
42	AD0587	IV	44	F	NA								
43	AD0363	IV	53	F	Smoker								
44	AD0001	IV	67	F	NA		G12V						
45	AD0046	IV	58	M	NA								
46	AD0044	IV	50	F	NA								
47	AD0049	IV	48	M	NA								
48	AD0047	IV	32	F	NA								
49	AD0067	IV	40	M	Non-smoker								
50	AD0025	IV	59	M	NA								
51	AD0031	IV	55	F	Non-smoker								

52	AD0033	IV	30	M	Non-smoker								
53	AD0034	IV	62	F	Non-smoker	E746		G691R					
54	AD0036	IV	31	F	Non-smoker	T790M							
55	AD0040	IV	52	F	Non-smoker		G13X						
56	AD0043	IV	63	M	Smoker	G719X							
57	AD0045	IV	57	F	Non-smoker								
58	AD0047	IV	30	F	Non-smoker								
59	AD0048	IV	60	M	Smoker								
60	AD0050	IV	50	F	NA								
61	AD0051	IV	48	F	Non-smoker								
62	AD0054	IV	62	M	Smoker								
63	AD0055	IV	46	M	Non-smoker								
64	AD0056	IV	39	F	Non-smoker								
65	AD0057	IV	51	F	Non-smoker								
66	AD0062	IV	45	F	Non-smoker	L858R		S249C					
67	AD0066	IV	57	M	Smoker								
68	AD0069	IV	40	F	Non-smoker								
69	AD0073	IV	75	F	Non-smoker	L858R							
70	AD0074	IV	46	F	Non-smoker								
71	AD0076	IV	52	M	Smoker								
72	AD0078	IV	29	M	NA								
73	AD0081	IV	36	M	Non-smoker			S249C					
74	AD0085	IV	68	M	Smoker	L858R	G12C						
75	AD0086	IV	67	M	Smoker								
76	AD0087	IV	70	M	Non-smoker	L858R							
77	AD0091	IV	65	M	Non-smoker								
78	AD0093	IV	75	M	Non-smoker	L858R							
79	AD0097	IV	49	M	Non-smoker								
80	AD0098	IV	40	F	Non-smoker								
81	AD0107	IV	57	M	Smoker								
82	AD0114	IV	38	M	Non-smoker								
83	AD0115	IV	66	M	Non-smoker	G719X							
84	AD0121	IV	42	F	Non-smoker	E746							
85	AD0122	IV	49	F	Non-smoker	G719X							
86	AD0125	IV	62	M	Non-smoker	E746					E542K		
87	AD0126	IV	64	M	NA								
88	AD0133	IV	54	M	Non-smoker	L858R							
89	AD0134	IV	68	M	Smoker								
90	AD0136	IV	31	M	Smoker								
91	AD0140	IV	56	M	Smoker								
92	AD0150	IV	47	F	Non-smoker								
93	AD0152	IV	57	F	Non-smoker								
94	AD0154	IV	59	F	Non-smoker		G13X						
95	AD0155	IV	50	F	Non-smoker	L858R							
96	AD0158	IV	56	F	Non-smoker								
97	AD0165	IV	34	F	Non-smoker								
98	AD0167	IV	49	F	Non-smoker								
99	AD0174	IV	50	M	Non-smoker								
100	AD0181	IV	62	M	Non-smoker								
101	AD0185	IV	55	M	Non-smoker								
102	AD0186	IV	38	M	Non-smoker								
103	AD0188	IV	50	M	Smoker								
104	AD1590	IV	43	M	NA								
105	AD0191	IV	57	M	Non-smoker								
106	AD0192	IV	59	F	Non-smoker	E746							
107	AD0194	IV	63	F	Non-smoker								

108	AD0196	IV	72	F	Non-smoker		G12C						
109	AD0197	IV	61	F	Non-smoker		G13X						
110	AD0198	IV	53	F	Non-smoker	E746							
111	AD0199	IV	33	M	Non-smoker	L858R							
112	AD0200	IV	36	M	Non-smoker								
113	AD0201	IV	68	M	Smoker	E746							
114	AD0202	IV	55	M	Non-smoker	E746							
115	AD0203	IV	52	M	Non-smoker								
116	AD0204	IV	58	M	Smoker		G13X						
117	AD0207	IV	56	M	Non-smoker								
118	AD0213	IV	66	M	Non-smoker								
119	AD0214	IV	56	M	Non-smoker		G13X						
120	AD0215	IV	58	F	Non-smoker			S249C					
121	AD0221	IV	43	F	Non-smoker	E746							
122	AD0224	IV	48	M	Smoker			G691R					780-781 insGSP
123	AD0226	IV	48	F	Non-smoker	E746							
124	AD0227	IV	60	F	Non-smoker	L858R							
125	AD0228	IV	70	F	Non-smoker	E746							
126	AD0229	IV	42	F	Non-smoker								
127	AD0231	IV	48	F	Non-smoker								
128	AD0233	IV	45	F	Non-smoker		G12C						
129	AD0235	IV	58	F	Non-smoker								
130	AD0237	IV	48	M	Smoker								
131	AD0240	IV	43	M	Non-smoker								
132	AD0242	IV	52	M	Non-smoker								
133	AD0244	IV	66	M	Smoker								
134	AD1534	IV	68	M	Smoker		G12C						
135	AD1591	IV	55	M	Smoker								
136	AD1501	IV	42	F	Non-smoker	L858R							
137	AD1592	IV	64	M	Smoker								
138	AD1474	IV	35	M	Non-smoker	E746							
139	AD1593	IV	55	F	Non-smoker								
140	AD1573	IV	65	F	Non-smoker								
141	AD1552	IV	60	F	Non-smoker			S249C					
142	AD1475	IV	32	M	Smoker	L858R							
143	AD1574	IV	55	F	Non-smoker								
144	AD1465	IV	75	F	Non-smoker	L858R					E545K		
145	AD1476	IV	71	F	Non-smoker	L858R							
146	AD1693	IV	55	F	Non-smoker								
147	AD1648	IV	47	M	Smoker								
148	AD1594	IV	65	F	Non-smoker								
149	AD1477	IV	50	M	Smoker	L858R							
150	AD1539	IV	47	F	Non-smoker		G12V						
151	AD1641	IV	62	M	NA								
152	AD1656	IV	34	M	Non-smoker								
153	AD1497	IV	52	M	Non-smoker	L858R							
154	AD1509	IV	60	F	Non-smoker	E746							
155	AD1685	IV	45	F	Non-smoker								
156	AD1642	IV	65	F	Non-smoker								
157	AD1478	IV	52	M	Smoker	E746							
158	AD1510	IV	43	M	Smoker	E746							
159	AD1664	IV	64	M	Non-smoker								
160	AD1595	IV	48	M	Non-smoker								
161	AD1570	IV	59	M	Non-smoker								
162	AD1498	IV	52	M	Smoker	E746							
163	AD1538	IV	40	F	Non-smoker		G12V	S249C					

164	AD1643	IV	45	F	NA								
165	AD1649	IV	57	F	Non-smoker								
166	AD1557	IV	41	M	Non-smoker			S249C					
167	AD1596	IV	41	M	Non-smoker								
168	AD1473	IV	54	M	Smoker	E746							
169	AD1479	IV	42	M	Non-smoker	E746							
170	AD1559	IV	78	M	Non-smoker					E17K			
171	AD1502	IV	36	M	Non-smoker	L858R							
172	AD1597	IV	54	M	Non-smoker								
173	AD1598	IV	55	M	Smoker								
174	AD1467	IV	68	M	Non-smoker	E746							
175	AD1480	IV	56	M	Non-smoker	E746							
176	AD1581	IV	64	F	Non-smoker								
177	AD1548	IV	37	M	Non-smoker			S249C		E17K			
178	AD1665	IV	77	M	Non-smoker								
179	AD1599	IV	53	M	Non-smoker								
180	AD1468	IV	31	M	Non-smoker	E746							
181	AD1511	IV	56	F	Non-smoker	L858R							
182	AD1666	IV	57	M	Non-smoker								
183	AD1600	IV	63	M	Smoker								
184	AD1553	IV	59	M	Non-smoker			G691R					
185	AD1543	IV	65	M	Non-smoker		G13X						
186	AD0425	IV	31	M	Non-smoker								
187	AD1667	IV	57	M	Non-smoker								
188	AD1568	IV	56	M	Non-smoker								
189	AD1560	IV	31	F	Non-smoker					E17K			
190	AD1469	IV	55	F	Non-smoker	E746							
191	AD1499	IV	59	M	Non-smoker	E746							
192	AD1686	IV	57	M	Smoker								
193	AD1526	IV	33	M	Non-smoker	G719X							
194	AD1601	IV	55	F	Non-smoker								
195	AD1481	IV	53	M	Smoker	L858R							
196	AD1602	IV	55	M	Non-smoker								
197	AD1694	IV	51	F	Non-smoker			G691R					
198	AD1482	IV	35	F	Non-smoker	T790M			EML4-ALK				
199	AD1603	IV	53	F	Non-smoker								
200	AD1604	IV	44	F	Non-smoker								
201	AD1605	IV	75	M	Smoker								
202	AD1606	IV	52	M	Non-smoker								
203	AD1531	IV	63	M	Smoker	T790M							
204	AD1564	IV	60	M	Smoker				EML4-ALK				
205	AD1607	IV	51	M	Non-smoker								
206	AD1608	IV	24	F	Non-smoker								
207	AD1644	IV	33	F	Non-smoker								
208	AD1609	IV	37	M	Non-smoker								
209	AD1610	IV	72	M	Smoker								
210	AD1611	IV	50	M	Non-smoker								
211	AD1612	IV	40	M	Non-smoker								
212	AD1532	IV	48	M	Non-smoker		G12C	S249C		E17K			
213	AD1613	IV	47	M	Non-smoker								
214	AD1614	IV	59	M	Non-smoker								
215	AD1512	IV	40	F	Non-smoker	E746							
216	AD1554	IV	50	M	Non-smoker								
217	AD1503	IV	60	F	Non-smoker	E746							
218	AD1615	IV	55	M	Smoker								
219	AD1616	IV	45	F	Non-smoker								

220	AD1569	IV	52	M	Non-smoker								
221	AD1650	IV	44	F	Non-smoker								
222	AD1483	IV	68	F	Non-smoker	T790M							
223	AD1617	IV	56	M	Non-smoker								
224	AD1618	IV	65	M	Smoker								
225	AD1658	IV	52	M	Non-smoker								
226	AD1645	IV	45	M	Non-smoker								
227	AD1646	IV	34	M	Non-smoker								
228	AD1619	IV	39	F	Non-smoker								
229	AD1549	IV	21	F	Non-smoker			S249C		E17K			
230	AD1620	IV	41	M	Non-smoker								
231	AD1484	IV	56	M	Non-smoker	E746							
232	AD1651	IV	62	M	Non-smoker								
233	AD1507	IV	65	F	Non-smoker	E746							
234	AD1621	IV	46	F	Non-smoker								
235	AD1485	IV	50	M	Non-smoker	E746							
236	AD1540	IV	32	F	Non-smoker		G12V						
237	AD1535	IV	63	M	Non-smoker		G12C						
238	AD1622	IV	48	M	Non-smoker								
239	AD1655	IV	61	F	Non-smoker								
240	AD1652	IV	55	M	Smoker								
241	AD1513	IV	63	M	Smoker	G719X							
242	AD1561	IV	55	M	Non-smoker						E542K		
243	AD1695	IV	27	F	Non-smoker								
244	AD1486	IV	60	M	Smoker	G719X							
245	AD1659	IV	71	M	Non-smoker								
246	AD1623	IV	47	M	Non-smoker								
247	AD1487	IV	61	M	Non-smoker	L858R							
248	AD1660	IV	58	M	Non-smoker								
249	AD1691	IV	44	M	Non-smoker								
250	AD1668	IV	45	M	Non-smoker								
251	AD1514	IV	62	M	Smoker	E746							
252	AD1565	IV	44	F	Non-smoker				EML4-ALK				
253	AD1537	IV	60	M	Non-smoker		G12C						
254	AD1669	IV	62	M	Non-smoker								
255	AD1575	IV	57	M	Smoker								
256	AD1670	IV	40	F	Non-smoker								
257	AD1624	IV	66	F	Non-smoker								
258	AD1504	IV	62	M	Non-smoker	L858R							
259	AD1625	IV	45	M	Non-smoker								
260	AD1626	IV	25	M	Non-smoker								
261	AD1627	IV	46	M	Non-smoker								
262	AD1653	IV	33	M	Non-smoker								
263	AD1671	IV	42	M	Non-smoker								
264	AD1488	IV	65	F	Non-smoker	L858R							
265	AD1489	IV	64	M	Non-smoker	E746							
266	AD1515	IV	62	F	Non-smoker	E746							
267	AD1547	IV	42	M	Non-smoker		G13X						
268	AD1672	IV	36	F	Non-smoker								
269	AD1490	IV	50	M	Non-smoker	E746							
270	AD1582	IV	65	M	Non-smoker								
271	AD1583	IV	61	M	Non-smoker								
272	AD1562	IV	75	M	Smoker						E545K		
273	AD1576	IV	55	M	Non-smoker								
274	AD1463	IV	63	M	Non-smoker	L858R							
275	AD1491	IV	55	M	Non-smoker	L858R							

276	AD1470	IV	40	M	Smoker	L858R							
277	AD1528	IV	46	M	Non-smoker	L858R							
278	AD1628	IV	60	M	Non-smoker								
279	AD1629	IV	38	M	Smoker								
280	AD1673	IV	65	M	Non-smoker								
281	AD1505	IV	56	F	Non-smoker	L858R							
282	AD1630	IV	49	F	Non-smoker								
283	AD1571	IV	54	M	Smoker								
284	AD1529	IV	73	M	Non-smoker	L858R							
285	AD1674	IV	50	F	Non-smoker								
286	AD1675	IV	52	M	Non-smoker								
287	AD1530	IV	39	F	Non-smoker	L858R							
288	AD1584	IV	51	M	Non-smoker								
289	AD1516	IV	55	M	Smoker	L858R							
290	AD1631	IV	66	M	Non-smoker								
291	AD1676	IV	71	M	Non-smoker								
292	AD1577	IV	28	M	Non-smoker								
293	AD1632	IV	36	M	Non-smoker								
294	AD1527	IV	66	F	Non-smoker	E746							
295	AD1523	IV	63	M	Smoker	L858R							
296	AD1492	IV	45	M	Smoker	E746							
297	AD1563	IV	76	F	Non-smoker						E545K		
298	AD1472	IV	65	M	Non-smoker	E746							
299	AD1677	IV	46	F	Non-smoker								
300	AD1633	IV	51	M	Non-smoker								
301	AD1493	IV	50	F	Non-smoker	L858R							
302	AD1587	IV	69	M	Non-smoker								
303	AD1494	IV	58	M	Smoker	L858R							
304	AD1678	IV	50	F	Non-smoker								
305	AD1689	IV	37	F	Non-smoker								
306	AD1634	IV	61	M	Non-smoker								
307	AD1578	IV	77	F	Non-smoker								
308	AD1555	IV	27	M	Non-smoker			S249C					
309	AD1679	IV	51	M	Non-smoker								
310	AD1680	IV	73	M	Non-smoker								
311	AD1654	IV	72	M	Non-smoker								
312	AD1635	IV	58	M	Smoker								
313	AD1506	IV	61	M	Non-smoker	L858R							
314	AD1545	IV	47	M	Non-smoker		G13X						
315	AD1579	IV	72	M	Smoker								
316	AD1580	IV	69	M	Smoker								
317	AD1495	IV	57	M	Non-smoker	L858R							
318	AD1687	IV	71	M	Non-smoker								
319	AD1558	IV	76	M	Smoker					E17K			
320	AD1585	IV	41	M	Non-smoker								
321	AD1517	IV	59	F	Non-smoker	L858R							
322	AD1466	IV	66	F	Non-smoker	L858R						A634D	
323	AD1688	IV	49	F	Non-smoker								
324	AD1692	IV	64	F	Non-smoker								
325	AD1541	IV	57	F	Non-smoker		G13X						
326	AD1636	IV	48	M	Non-smoker								
327	AD1536	IV	45	M	Smoker		G12C						
328	AD1518	IV	33	F	Non-smoker	E746							
329	AD1647	IV	39	F	Non-smoker								
330	AD1464	IV	46	M	Non-smoker	L858R				E17K			
331	AD1542	IV	57	M	Smoker		G13X						

332	AD1637	IV	53	M	Non-smoker								
333	AD1519	IV	48	M	Non-smoker	E746							
334	AD1462	IV	51	M	Smoker	L858R		S249C					
335	AD1500	IV	61	F	Non-smoker	L858R							
336	AD1638	IV	43	M	Smoker								
337	AD1586	IV	45	F	Non-smoker								
338	AD1556	IV	39	M	Non-smoker			S249C					
339	AD1520	IV	46	M	Non-smoker	L858R							
340	AD1521	IV	66	M	Non-smoker	L858R							
341	AD1496	IV	39	F	Non-smoker	E746							
342	AD1681	IV	63	M	Smoker								
343	AD1682	IV	53	M	Smoker								
344	AD1567	IV	63	M	Smoker								
345	AD1683	IV	39	F	Non-smoker								
346	AD1546	IV	58	F	Non-smoker		G13X						
347	AD1639	IV	45	M	Smoker								
348	AD1522	IV	61	M	Non-smoker	L858R							
349	AD1524	IV	72	M	Non-smoker	L858R							
350	AD1471	IV	49	M	Non-smoker	E746							
351	AD1661	IV	75	M	Non-smoker								
352	AD1640	IV	51	F	Non-smoker								
353	AD1551	IV	63	F	Non-smoker			S249C					
354	AD1657	IV	56	M	Smoker								
355	AD1566	IV	59	M	Smoker								
356	AD1544	IV	78	M	Non-smoker		G13X						
357	AD1589	IV	59	M	Smoker								
358	AD1508	IV	72	M	NA	L858R							
359	AD1663	IV	58	M	Non-smoker								
360	AD1533	IV	69	M	Smoker		G12C						
361	AD1550	IV	63	F	Non-smoker			S249C		E17K			
362	AD1684	IV	59	F	Smoker								
363	AD1525	IV	75	M	Non-smoker	E746							

Appendix 4: Details of mutations identified by integrated analysis in HNSCC cell lines.

Gene	Cell line	Copy number	Gene Expression [log10(FPKM+1)]	Mutation	Chromosome	Start position	End position	Reference base	Mutant base
ALKBH3	AW13516	3	1.100081	D228E	11	43940602	43940602	C	G
ALKBH3	AW8507	3	1.233849	D228E	11	43940602	43940602	C	G
ARIH2	AW13516	3	1.443341	E25K	3	48965064	48965064	G	A
BCAR1	AW13516	5	1.565994	G188E	16	75276438	75276438	C	T
CAPN2	OT9	3	2.186179	R461S	1	2.24E+08	2.24E+08	C	A
CCNDBP1	OT9	2	1.053946	R259G	15	43483788	43483788	C	G
CHSY1	AW13516	3	1.148402	R705Q	15	1.02E+08	1.02E+08	C	T
CHSY1	AW13516	3	1.148402	R588T	15	1.02E+08	1.02E+08	C	G
CLK2	AW8507	3	1.202428	R109H	1	1.55E+08	1.55E+08	C	T
CRLF3	NT8e	3	1.102042	L389P	17	29111368	29111368	A	G
ECD	AW13516	3	1.244495	D667G	10	74894375	74894375	T	C
ECD	OT9	3	1.202516	D667G	10	74894375	74894375	T	C
FAM120B	AW13516	3	1.031148	D370Y	6	1.71E+08	1.71E+08	G	T
FAM120B	OT9	3	1.052298	D370Y	6	1.71E+08	1.71E+08	G	T
FAT2	NT8e	3	1.245177	R574C	5	1.51E+08	1.51E+08	G	A
GSN	NT8e	3	1.831882	A129T	9	1.24E+08	1.24E+08	G	A
HLA-A	OT9	4	2.352047	I166T	6	29911198	29911198	T	C
HLA-C	AW13516	4	2.192754	E69K	6	31324603	31324603	C	T
HLA-C	AW13516	4	2.192754	D98Y	6	31324516	31324516	C	A
HLA-C	OT9	4	2.247367	E69K	6	31324603	31324603	C	T
HRAS	AW13516	3	1.974521	G12S	11	534289	534289	C	T
HRAS	AW8507	3	1.984534	G12S	11	534289	534289	C	T
IDH1	OT9	4	1.630559	V178I	2	2.09E+08	2.09E+08	C	T
IMMT	AW13516	3	1.695099	L228V	2	86393741	86393741	G	C
ITGA6	AW13516	3	2.055401	A380T	2	1.73E+08	1.73E+08	G	A
ITGA6	AW8507	3	1.832214	A380T	2	1.73E+08	1.73E+08	G	A
KIAA1522	NT8e	5	1.861432	S638C	1	33236693	33236693	C	G
KIRREL	NT8e	3	1.208807	T233M	1	1.58E+08	1.58E+08	C	T
LAMA5	OT9	6	1.859615	V3147F	20	60887294	60887294	C	A
LOXL4	NT8e	3	1.058828	R188W	10	1E+08	1E+08	G	A
MET	AW13516	3	1.631316	N375S	7	1.16E+08	1.16E+08	A	G
MKI67	AW13516	3	1.431414	V1559M	10	1.3E+08	1.3E+08	C	T
MKI67	AW13516	3	1.431414	T1247I	10	1.3E+08	1.3E+08	G	A

MKI67	AW8507	3	1.764421	V1559M	10	1.3E+08	1.3E+08	C	T
MKI67	AW8507	3	1.764421	T1247I	10	1.3E+08	1.3E+08	G	A
NOC2L	AW8507	3	2.009731	S556L	1	881918	881918	G	A
NQO1	NT8e	3	1.773881	P187S	16	69745145	69745145	G	A
NRBP1	NT8e	4	1.694574	Q73*	2	27656546	27656546	C	T
OSBPL3	NT8e	5	1.14373	S16L	7	24932045	24932045	G	A
PDE4DIP	OT9	3	1.04605	A1066T	1	1.45E+08	1.45E+08	C	T
PRKD3	NT8e	4	1.164998	N42D	2	37543544	37543544	T	C
PTEN	OT9	3	1.135136	H141Y	10	89692937	89692937	C	T
PYGB	AW13516	5	1.96517	A303S	20	25259006	25259006	G	T
SAT2	OT9	4	1.494584	R126C	17	7529902	7529902	G	A
TMEM43	AW8507	3	1.611693	R28W	3	14170981	14170981	C	T
TP53	AW13516	5	1.816318	R273H	17	7577120	7577120	C	T
TP53	AW8507	4	1.753342	R273H	17	7577120	7577120	C	T
USP12	NT8e	3	1.087217	C50Y	13	27680062	27680062	C	T
WDYHV1	AW13516	3	1.050256	R134C	8	1.24E+08	1.24E+08	C	T
ZNF768	AW13516	3	1.301685	E181D	16	30536918	30536918	C	G

Appendix 5: HPVDetector user guide.

I) Pre-requisites, installation, and execution of HPV Detector:

A) Pre-requisites:

1. Linux/Unix based Operating System
2. RAM: 6GB or more.
3. Burrows Wheeler Aligner (BWA) (minimum Ver. 0.6.*).
4. Awk scripting language (generally included in Linux/Unix system)
5. Yad for GUI (generally included in Linux/Unix system)

B) Installation:

Decompress the HPVDetector_v0.1.tar.gz file to a suitable location.

HPV Detector package bundle zip is composed of following files:

1) HPV Detector programme files.
2) Directory of indexed HPV reference genome.
3) Directory of indexed Human-HPV pseudo reference genome files.
4) Directory of HPV-Human Gene Annotation file.
5) In case YAD is required to be installed:

   Command for RHEL/Fedora/Cent OS: sudo yum install yad

   Command for Ubuntu: sudo apt-get install yad
6) In case BWA is required to be installed, run following commands:

   wget -o bwa-0.6.2.tar.bz2 http://sourceforge.net/projects/bio-bwa/files/bwa-0.6.2.tar.bz2/download

   tar -xvf bwa-0.6.2.tar.bz2

   cd bwa-0.6.2/

   make
7) In case Picard tool is required to be installed, run following commands:

wget -o picard-tools-1.100.zip http://sourceforge.net/projects/picard/files/picard-tools/1.100/picard-tools-1.100.zip/download

unzip picard-tools-1.100.zip

C) Tools & software's used while testing HPV Detector:

1. BWA version 0.6.2-r126
2. Picard Tools version 1.100
3. Sam Tools version 0.1.18 (r982:295)
4. Bio-perl for GenBank data parsing. Ver. 1.006901
5. Linux operating system Fedora 20 (x86_64)
6. Yad 0.25

II) Execution - GUI:

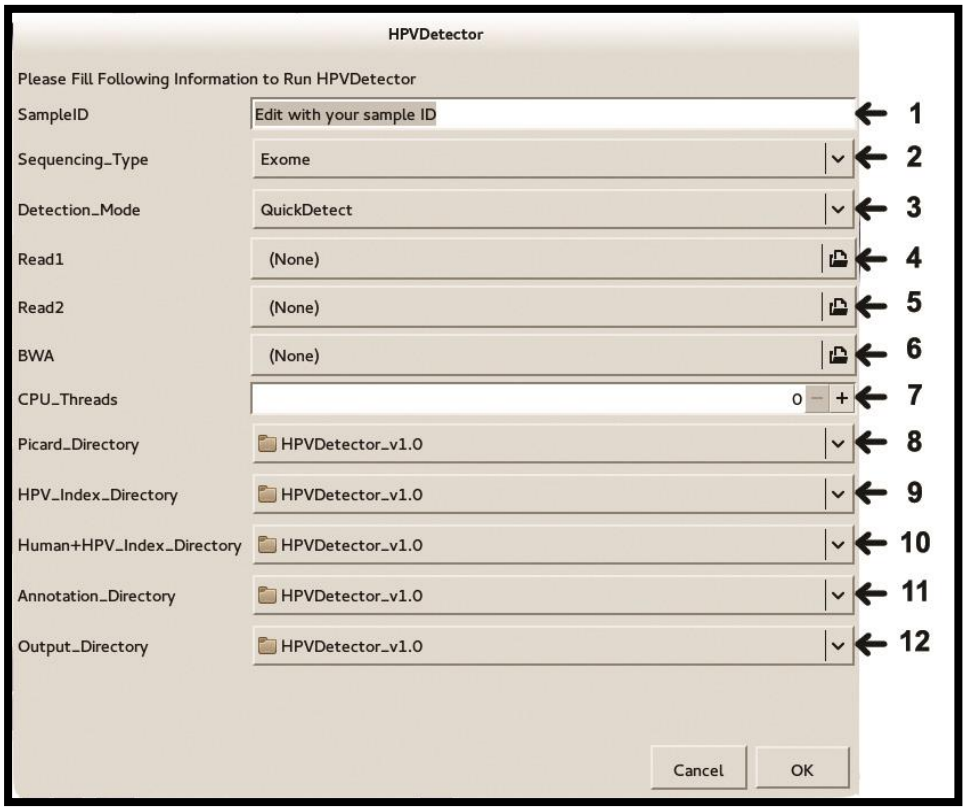
HPV Detector can be executed with GUI by running HPVdetector_GUI from Linux/UNIX shell:

> ./HPVdetector_GUI

Or

> bash HPVdetector_GUI

The GUI of HPVDetector.



Upon successful execution, HPVDetector GUI will pop-up with 12 options for user selection. User can easily select required options using point-and-click interface and clicking on OK button will start running HPVDetector.

Step-1: Provide the sample ID

Step-2: Select sequencing type from Exome, Genome or Transcriptome

Step-3: Select detection mode from QuickDetect or Integration

Step-4: Select input FASTQ file (forward reads FASTQ)

Step-5: Select input FASTQ file (reverse reads FASTQ)

Step-6: Select BWA executable

Step-7: Set the number of CPU threads as per your system (most modern CPUs have at least two cores/threads)

Step-8: Select directory where Picard Tools are installed

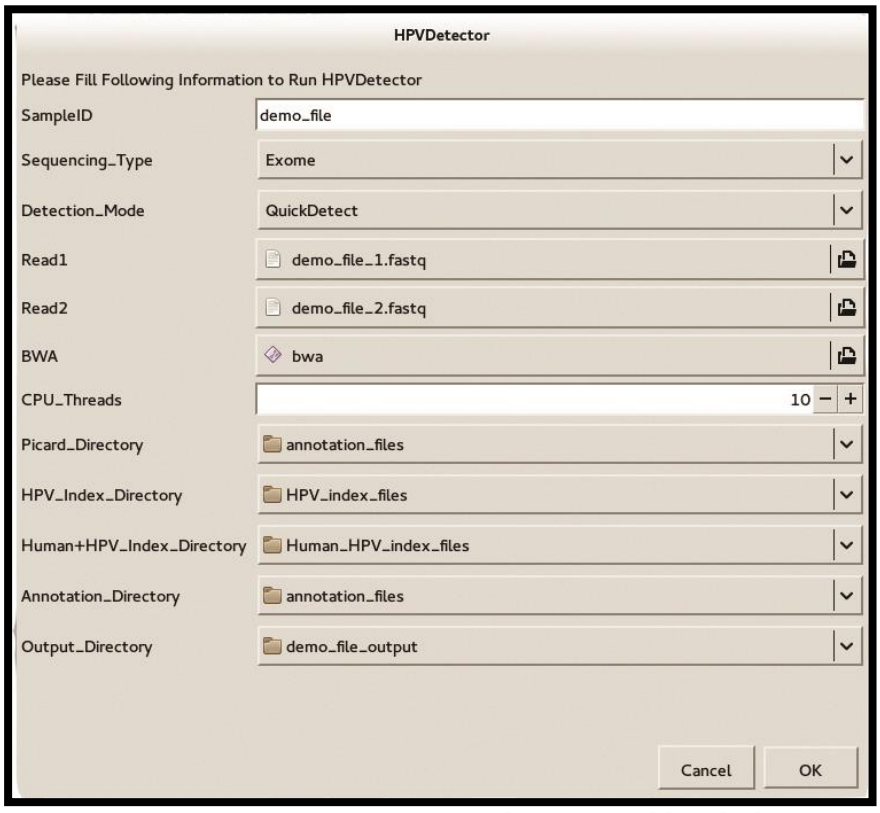
Step-9: Select directory where HPV_index is stored

Step-10: Select directory where Human-HPV_index is stored

Step-11: Select directory where Human-HPV Annotation files are stored

Step-12: Select a directory where output files should be stored

A typical selection of options with supplied supporting files.



After selecting appropriate options, program can be executed by clinking on OK button. Progress can be monitored on screen and after completion, program will point to respective output files.

III) Execution – command line:

HPVDetector can also be executed from command line of Linux/Unix systems. Edit the config.txt file in root directory of HPVDetector to make changes of the variables/path required for the tool. There should be no changes made other than following:

1) mode should be specified as “Integration” or “QuickDetect”

Ex. mode=QuickDetect

2) bwa should be specified as path to the BWA 0.6.2 executable

Ex. bwa=/home/bwa-0.6.2/bwa

3) picard should be specified as path to the Picard tool 1.100 directory

Ex. picard=/home/HPVDetector_v1.0/picard-tools-1.100/

4) hpv_bwa_index should be specified as path to HPV_index_files supplied with the tool

Ex.

hpv_bwa_index=/home/HPVDetector_v1.0/HPV_index_files/HPV_143types_EBVref

5) hpv_human_bwa_index should be specified as path to Human_HPV_index_files supplied with the tool

Ex.

hpv_human_bwa_index=/home/HPVDetector_v1.0/Human_HPV_index_files/human_HPV

6) annotation_dir should be specified as path to annotation_files supplied with the tool

Ex. annotation_dir=/home/HPVDetector_v1.0/annotation_files/

7) threads should be specified as whole number (should not exceed the actual number of threads available in the system)

Ex. threads=4

After making above changes, execute HPVDetector as:

./HPVDetector config.txt <read1.fastq> <read2.fastq> <analysis mode> <output path>

<analysis mode> could be WGS, Transcriptome or Exome

Ex.:

./HPVDetector config.txt /siha/read_1.fastq /siha/read_2.fastq WGS /siha_output

Original article *Annals of Oncology*. 2016 Dec 19. pii: mdw636

Drug-sensitive *FGFR3* mutations in lung adenocarcinoma

P. Chandrani[1,2*], K. Prabhash[2,3*], A. Choughule[3], R. Prasad[1], V. Sethunath[1], M. Ranjan[1], P. Iyer[1,2], J. Aich[1], H. Dhamne[1], D. N. Iyer[1], P. Upadhyay[1,2], B. Mohanty[4], P. Chandna[5], R. Kumar[3], A. Joshi[3], V. Noronha[3], V. Patil[3], A. Ramaswamy[3], A. Karpe[3], R. Thorat[4], P. Chaudhari[4], A. Ingle[4], A. Dutt[1,2]

1 Integrated Genomics Laboratory, Advanced Centre for Treatment, Research and Education in Cancer, Tata Memorial Centre, Navi Mumbai, India

2 Homi Bhabha National Institute, Training School Complex, Anushakti Nagar, Mumbai, India

3 Department of Medical Oncology, Tata Memorial Hospital, Tata Memorial Centre, Mumbai, India

4 Small Animal Imaging Facility, Animal Sciences, Advanced Centre for Treatment, Research and Education in Cancer, Tata Memorial Centre, Navi Mumbai, India

5 AceProbe Technologies Pvt. Ltd., New Delhi, India

* Both authors contributed equally

Correspondence should be addressed to:

Dr. Amit Dutt

Wellcome Trust/ DBT India Alliance Intermediate Fellow

Tata Memorial Centre, ACTREC

Navi Mumbai, 410 210 INDIA.

Phone: +91-22-27405056

Email address: adutt@actrec.gov.in

Abstract

Background: Lung cancer is the leading cause of cancer-related deaths across the world. In this study we present therapeutically relevant genetic alterations in lung adenocarcinoma of Indian origin.

Materials and methods: Forty-five primary lung adenocarcinoma tumors were sequenced for 676 amplicons using RainDance cancer panel at an average coverage of 1500X (reads per million mapped reads). To validate the findings, 49 mutations across 23 genes were genotyped in an additional set of 363 primary lung adenocarcinoma tumors using mass spectrometry. NIH/3T3 cells over expressing mutant and wild type *FGFR3* constructs were characterized for anchorage independent growth, constitutive activation, tumor formation and sensitivity to FGFR inhibitors using *in vitro* and xenograft mouse models

Results: We present the first spectrum of actionable alterations in lung adenocarcinoma tumors of Indian origin, and show that mutations of *FGFR3* are present in 20 of 363 (5.5%) patients. These *FGFR3* mutations are constitutively active and oncogenic when ectopically expressed in NIH/3T3 cells and using a xenograft model in NOD/SCID mice. Inhibition of *FGFR3* kinase activity inhibits transformation of NIH/3T3 overexpressing *FGFR3* constructs and growth of tumors driven by *FGFR3* in the xenograft models. The reduction in tumor size in the mouse is paralleled by a reduction in the amounts of phospho-ERK, validating the *in vitro* findings. Interestingly, the *FGFR3* mutations are significantly higher in a proportion of younger patients and show a trend towards better overall survival, compared to patients lacking actionable alterations or those harboring *KRAS* mutations.

Conclusion: We present the first actionable mutation spectrum in Indian lung cancer genome. These findings implicate *FGFR3* as a novel therapeutic in lung adenocarcinoma.

Keywords: lung adenocarcinoma; actionable mutations; fibroblast growth factor receptor 3; oncogene; FGFR inhibitors; mass spectrometry

Key Message

Diversity in cancer-specific alterations lends similarity to the known plurality in clinical response based on ethnicity and other divergent factors. This study establishes that *FGFR3* mutations found in lung adenocarcinoma patients of Indian origin are oncogenic, and forms a subclass of FGFR inhibitor-sensitive patients largely distinct from those harboring *EGFR*, *KRAS*, or *EML4-ALK* mutations.

Introduction

Lung cancer is the leading cause of cancer-related deaths worldwide, accounting for over a million deaths annually [1]. Molecularly targeted therapies improve outcome for lung adenocarcinoma patients whose tumors harbor mutant *EGFR* or translocated *ALK, RET,* or *ROS1*, with an encouraging response for those with mutated *BRAF*, *MET*, *NTRK-1 & 2* and *ERBB2* [2-5]. Such oncogenic somatic alterations though vary across populations/ethnic groups, e.g., *EGFR* mutations are present in over 30% of East Asian lung adenocarcinoma patients, however, they are only found in about 23-25% of Indian and 10% of Western lung adenocarcinoma patients [6-8]. Similarly, *KRAS* mutations are present at 60% lower frequency in Indian lung adenocarcinoma patients than compared to the Caucasian population [3, 9, 10]. The diversity in somatic alterations lends similarity to the known plurality in clinical response based on ethnicity and divergent genetic and environmental factors [11], Thus, besides the unmet need for additional therapeutic targets in lung adenocarcinoma patients, it is equally pertinent to profile known oncogenic somatic alterations across different populations to understand their landscape of variability.

Here, in an attempt to profile for activating alterations, we have generated a comprehensive mutational spectrum of activating alterations prevalent among lung adenocarcinoma patients of Indian origin, considered to be an admixture population of non-European Caucasian and Ancestral South Indians. We also report the first incidence of activating and drug sensitive *FGFR3* mutations in lung adenocarcinoma. *FGFR3* mutated samples, with ~5% population frequency, form a distinct subclass apart from *EGFR*, *KRAS*, and *EML4-ALK*.

Methods & Materials

Patients

To profile for therapeutically relevant genome alterations in lung adenocarcinoma of Indian origin, FFPE blocks with known *EGFR* mutation status for 45 consecutive histologically confirmed lung adenocarcinoma patients tumor samples (stage IV, 49% males and 51% non-smokers) for sequencing and an additional set of 363 consecutive lung adenocarcinoma patients tumor samples (stage IV, 62% males and 73% non-smokers) for mass spectrometry were retrospectively collected from Tata Memorial Hospital (Supplementary Table S1).

Pooling of samples, target gene-capturing and next generation sequencing

A set of 45 lung adenocarcinoma samples were sequenced using pooled sequencing approach to capture low frequency variants [12-14]. Briefly, 45 samples were divided into duplicate pools of different population size (Supplementary Figure-S1) i.e. 2 pools of 5 individuals (5XA and 5XB), 2 pools of 10 individuals (10XA and 10XB), and, 1 pool of 15 individuals (15X) for next-generation sequencing (NGS) of 676 genomic regions of 158 genes as described earlier [15].

Discovery of genomic variants using computational analysis

FASTQ files were analyzed using BWA-Picard-GATK/MuTect pipeline generating 3349 unique variants (Supplementary Table S2). Polymorphisms overlapping with dbSNP database (v.142) and Indian specific SNP database TMC-SNPdb derived from whole exome sequencing of 62 normal samples [16] were filtered (Supplementary Figure S2, S3). Stringent mutation analysis was performed as further detailed in supplementary methods to derive list of significant mutations for further validation (Supplementary Table S2 and S3).

Mass spectrometry based genotyping

Briefly, PCR and extension primers for 49 mutations in 23 genes were designed using single base extension based mass spectrometry assay design 3.1 software

(Supplementary Table S4). Mutation calls were analysed using Typer 4 (Sequenom Inc., USA) and were reviewed by manually observing mass-spectra.

Cell culture, anchorage-independent growth assay and immunoblotting

NIH/3T3 cells transduced with *FGFR3* wild-type and mutant construct were used for induction and drug inhibition study as detailed in supplementary methods. Anchorage independent growth assay and immunoblotting were performed as described earlier [17], and as detailed in the supplementary methods.

Xenograft development

A cohort of 8 NOD-SCID or Nude mice per clone were subcutaneously injected with 5 million cells for tumor formation in 2-3 months. Inhibitor BGJ-398 [18] was given at 15 and 30 mg/kg along with vehicle control (10% tween-80) independently to randomised xenograft groups after tumor size reaching ~150 mm^3. Tumor size was measured every alternate day using Vernier calliper for 14 days' drug treatment. microPET-CT scan was performed at the end of the drug treatment.

Immunohistochemistry

Immunohistochemistry analysis was performed as described earlier [19] and detailed in supplementary methods.

Overall survival analysis

Overall survival of patients was assessed using Kaplan-Meier method using R and IBM SPSS software package, as detailed in supplementary methods. The end point was taken as date of death with censoring implied at the date of the last contact.

Results

Recurrent *FGFR3* mutations in lung adenocarcinoma patient of Indian origin

To identify low frequency and ethnic specific therapeutically relevant genome alterations in lung adenocarcinoma of Indian origin, we sequenced 45 primary lung adenocarcinoma stage IV tumors (Supplementary Table S1) for 676 amplicons at an average coverage of 1500X (reads per million mapped reads), as described in Supplementary Figure S1-S5; Supplementary Table S2 & S3. To validate the findings, we selected 49 mutations occurring across 23 genes based on their recurrence and therapeutic significance (Supplementary Table S4), for genotyping in an additional set of 363 primary lung adenocarcinoma stage IV tumors (Figure 1A; Supplementary Table S5) using mass spectrometry.

Based on the mutation profiling of 363 lung adenocarcinoma patients, we present the first portrait of activating mutations present in the Indian lung cancer genome (Figure 1B), wherein 160 of 363 patients were found to harbor activating mutations across 8 genes at following frequency: *EGFR* (28.4%), *KRAS* (13%), *ALK* (3.8%), *AKT1* (2.5%), *PIK3CA* (1.4%), *FGFR4* (0.4%), and *ERBB2* (0.3%) as shown in Figure 1A, consistent with earlier reports [6, 8, 9]. In addition, 3 of 79 patients were found to harbor *EML4-ALK* translocation as determined by FISH. Among the other most significantly mutated genes, we found recurrent *FGFR3* mutations in 20 of 363 tumors (5.5%), of which 7 co-occurred in samples harboring *EGFR* (*n=5*) and *KRAS* (*n=2*) mutation. In total, sixteen patients harbored *FGFR3* (S249C) mutation; and, 4 patients harbored a novel *FGFR3* (G691R) mutation (Figure 1A; Figure 1C upper panel; Supplementary Figure S6; Supplementary Table S6). Interestingly, *FGFR3* (S249C) mutation has previously been described as activating and drug sensitive in lung squamous [20], while the novel *FGFR3* (G691R) mutation was predicted to be deleterious based on using 7 of 10 functional prediction tools (Supplementary Table S3).

FGFR3 mutations in lung adenocarcinoma are activating *in-vitro* & *in-vivo*

To test whether the recurrent *FGFR3* mutations found in this study are activating we transduced NIH/3T3 fibroblast cells with retroviruses encoding the *FGFR3* G691R

mutation along with WT *FGFR3* and the previously characterized *FGFR3* (R248C) and (S249C) mutations [20]. Similar to *FGFR3* R248C and S249C, the ectopic expression of the novel G691R mutant clone in pooled NIH/3T3 cells conferred anchorage-independent growth, forming 3-fold more colonies in soft agar than cells expressing WT *FGFR3* (Figure 1C left panel), despite higher expression levels of WT *FGFR3* (Figure 1C right panel). The transformation was accompanied by elevated phosphorylation of the downstream ERK1/2 and AKT1 in a constitutive manner (Figure 1D). Further, consistent with the *in vitro* data, NIH/3T3 cells expressing transforming *FGFR3* mutations or WT when injected subcutaneously into NOD/SCID mice formed tumors within 2 months post injection of cells. While 3 of 11 mice injected with cells expressing *FGFR3* WT formed tumors, 12 of 12 mice injected with cells expressing *FGFR3* S249C; and, 6 of 12 mice injected with cells expressing *FGFR3* G691R formed tumors (Figure 1E). The tumor size doubling time was ~7 days for cells expressing *FGFR3* (G691R), ~5 days for cells expressing *FGFR3* (S249C); the *FGFR3*-WT tumors doubled in size in ~9-10 days.

FGFR3* mutations in lung adenocarcinoma are sensitive to inhibitors *in-vitro* & *in-vivo

We further show that inhibition of FGFR3 kinase activity using pan FGFR inhibitor PD173074 block the constitutive phosphorylation of ERK1/2 (Figure 2A). Similarly, treatment of cells harboring activating *FGFR3* mutations with PD173074 result in a marked decrease in colony formation in soft agar and cell survival in liquid culture (Figure 2B). Extending the effect *in vivo* studies, when tumors reached approximately 100-200 mm^3 in all mice injected with NIH/3T3 cells began treatment with 15 or 30 mg/kg BGJ398 – a selective FGFR inhibitor currently under clinical trials for various cancer types (as PD173074 is incompatible with *in vivo* studies [21]), or vehicle for 14 days. Tumors treated with BGJ398 slowed or reversed their growth compared with vehicle (Figure 2C upper panel), so that by the end of the study, the effect on tumor burden in vehicle-treated versus BGJ398-treated mice were 3.3 folds in *FGFR3* (S249C), 3 folds in *FGFR3* (G691R) and 2.25 folds in *FGFR3*-WT xenografts (Figure 2D). This reduction in tumor size was paralleled by a reduction in the amounts of phospho-ERK1/2 in immuno-histochemical analyses

(Figure 2C lower panel) of explanted tumors, validating our *in vitro* findings that MAPK signaling is the key pathway engaged by mutated *FGFR3*.

Correlation of *FGFR3* mutations with clinicopathological features of lung cancer patients

Clinically, lung adenocarcinoma patients with *FGFR3* mutation positive tumors expressing higher activated MAPK levels (Supplementary Figure S7) show a better trend in overall survival (OS) with 17 months (*n= 8*; 95% CI: 6.4-27.5; HR: 0.6) compared to 14 months (*n= 197*; 95% CI: 8.7-13.2) in patients with wild-type *FGFR3* (Figure 2E). The OS trend in lung adenocarcinoma patients though is similar to bladder urothelial carcinomas and skin cutaneous melanoma patients, but not to head & neck cancer and lung squamous carcinoma patients, based on our analysis using cBioPortal for survival of patients harboring activating *FGFR3* mutations in different cancers (Supplementary Figure S8). Furthermore, the *FGFR3* mutations were observed to be significantly higher in patients < 45 years (9 of 95) than in patients > 45 years (11 of 269) (*P = 0.048*) but not with their gender and smoking status (Supplementary Table S7). The sample size in this study, however, is underpowered to reach statistical significance for survival data.

Discussion

We present the first portrait of clinically actionable alterations in lung adenocarcinoma of Indian origin which includes *EGFR*, *KRAS*, *EML4-ALK*, *AKT1*, *PIK3CA*, *FGFR4* and *ERBB2,* similar to that identified in other ethnic groups [5, 22, 23], and an additional subset of patients with *FGFR3* mutations. Ethnic-specific variations have been well known in lung cancer [24, 25] across different populations. We observed 28.4% *EGFR* mutations and 13% *KRAS* mutations in lung adenocarcinoma patients, consistent with our previous report [6, 9]. Similarly, variation in frequency of other molecular alterations is also observed such as 3% *EML4-ALK* alteration in our study compared to 8% in Caucasian population [3] and in 5% Chinese population [23]. *ERBB2* mutation found at <1% frequency in our cohort exists at ~2-3% among the Caucasian [3] and Chinese populations [23]. Similarly, *AKT1* mutations were found at higher than the reported <1% in both Caucasian [3] and Chinese populations [23] indicating the higher therapeutic relevance of *AKT1* targeted compounds in Indian population.

We have also identified frequent and recurrent drug sensitive *FGFR3* mutations in lung adenocarcinoma patients. Among the Caucasians, activating mutations in *FGFR3* have been earlier reported in bladder carcinoma [26], lung squamous cell carcinomas [20], and, cervical cancer [27], but were found to be largely absent in lung adenocarcinomas [23, 28, 29], except for Imielinski *et al.* who reported non-recurrent somatic *FGFR3* mutations of unknown functional significance in 3 of 183 lung adenocarcinoma patients [10]. On the other hand, the presence of frequent *FGFR3* mutations (with unknown driving protential) is tangentially referred to in the literature among Korean lung adenocarcinomas patients [30]. Along with these reports, our finding of activating *FGFR3* mutations in lung adenocarcinoma patients provides an interesting convergence with mouse genetic experiments wherein activated FGF9-FGFR3 signal acts as the primary oncogenic pathway involved in initiation of lung adenocarcinoma [31, 32].

Analyzing the potential effect of *FGFR3* driver mutations on survival lung cancer patients, we observed a trend towards better survival for *FGFR3* mutations in lung adenocarcinoma, compared to lung adenocarcinoma patients with wild-type *FGFR3* and those harboring *KRAS* mutation, similar to as reported in the bladder and skin

cancer [33]. Thus, *FGFR3* mutation represents an opportunity for targeted therapy in lung adenocarcinoma. FGFR inhibitors, which are currently in clinical testing in tumor types bearing genetic alterations in FGFR genes [34, 35], may be extended to evaluated in patients with *FGFR3*-mutated lung adenocarcinoma. Finally, with a broader emerging role across different cancers [20, 36-39], this study further underscores that *FGFR* family may potentially join the *EGFR* family as a widespread target for therapeutic intervention in several human cancers.

Acknowledgements:

All members of the Dutt laboratory for critically reviewing the manuscript. RainDance technologies Inc. for providing NGS library preparation services. Sandor LifeSciences Pvt. Ltd. for providing NGS 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), from Terry Fox Foundation through TMC- Research Administrative Council (TRAC; project 108); and intramural grants (IRB project 92 and 55). P.C. is supported by a senior research fellowship from ACTREC. P.U. is supported by a 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.

Funding

This work was supported by Wellcome Trust/Department of Biotechnology India Alliance [IA/I/11/2500278], Department of Biotechnology, Government of India [BT/PR2372/AGR/36/696/2011], and intramural grants [IRB project 92, 55 and 108].

Disclosure

The authors have declared no conflicts of interest.

References

1. Society AC. Cancer Facts & Figures 2012. Atlanta: American Cancer Society 2012.
2. Barlesi F, Blons H, Beau-Faller M et al. Biomarkers (BM) France: Results of routine EGFR, HER2, KRAS, BRAF, PI3KCA mutations detection and EML4-ALK gene fusion assessment on the first 10,000 non-small cell lung cancer (NSCLC) patients (pts). J Clin Oncol 2013; 31: 486s.
3. Kris MG, Johnson BE, Berry LD et al. Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs. JAMA 2014; 311: 1998-2006.
4. Tsao AS, Scagliotti GV, Bunn PA, Jr. et al. Scientific Advances in Lung Cancer 2015. J Thorac Oncol 2016; 11: 613-638.
5. Campbell JD, Alexandrov A, Kim J et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat Genet 2016; 48: 607-616.
6. Chougule A, Prabhash K, Noronha V et al. Frequency of EGFR mutations in 907 lung adenocarcioma patients of Indian ethnicity. PLoS One 2013; 8: e76164.
7. Lynch TJ, Bell DW, Sordella R et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med 2004; 350: 2129-2139.
8. Noronha V, Prabhash K, Thavamani A et al. EGFR mutations in Indian lung cancer patients: clinical correlation and outcome to EGFR targeted therapy. PLoS One 2013; 8: e61561.
9. Choughule A, Sharma R, Trivedi V et al. Coexistence of KRAS mutation with mutant but not wild-type EGFR predicts response to tyrosine-kinase inhibitors in human lung cancer. Br J Cancer 2014; 111: 2203-2204.
10. Imielinski M, Berger Alice H, Hammerman Peter S et al. Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing. Cell 2012; 150: 1107-1120.
11. Patel JN. Cancer pharmacogenomics: implications on ethnic diversity and drug response. Pharmacogenet Genomics 2015; 25: 223-230.
12. Tewhey R, Warner JB, Nakano M et al. Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat Biotechnol 2009; 27: 1025-1031.
13. Niranjan TS, Adamczyk A, Bravo HC et al. Effective detection of rare variants in pooled DNA samples using Cross-pool tailcurve analysis. Genome Biol 2011; 12: R93.
14. Kaartokallio T, Wang J, Heinonen S et al. Exome sequencing in pooled DNA samples to identify maternal pre-eclampsia risk variants. Sci Rep 2016; 6: 29085.
15. Harismendy O, Schwab RB, Bao L et al. Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. Genome Biol 2011; 12: R124.
16. Upadhyay P, Gardi N, Desai S et al. TMC-SNPdb: an Indian germline variant database derived from whole exome sequences. Database (Oxford) 2016; 2016.
17. Chandrani P, Upadhyay P, Iyer P et al. Integrated genomics approach to identify biologically relevant alterations in fewer samples. BMC Genomics 2015; 16: 936.
18. Guagnano V, Furet P, Spanka C et al. Discovery of 3-(2,6-dichloro-3,5-dimethoxy-phenyl)-1-{6-[4-(4-ethyl-piperazin-1-yl)-phenylamin o]-pyrimidin-4-yl}-1-methyl-urea (NVP-BGJ398), a potent and selective inhibitor of the fibroblast growth factor receptor family of receptor tyrosine kinase. J Med Chem 2011; 54: 7066-7083.
19. Upadhyay P, Nair S, Kaur E et al. Notch pathway activation is essential for maintenance of stem-like cells in early tongue cancer. Oncotarget 2016.
20. Liao RG, Jung J, Tchaicha JH et al. Inhibitor-sensitive FGFR2 and FGFR3 mutations in lung squamous cell carcinoma. Cancer Research 2013.
21. Touat M, Ileana E, Postel-Vinay S et al. Targeting FGFR Signaling in Cancer. Clin Cancer Res 2015; 21: 2684-2694.
22. Barlesi F, Mazieres J, Merlio JP et al. Routine molecular profiling of patients with advanced non-small-cell lung cancer: results of a 1-year nationwide programme of the French Cooperative Thoracic Intergroup (IFCT). Lancet 2016.

23. Wang R, Zhang Y, Pan Y et al. Comprehensive investigation of oncogenic driver mutations in Chinese non-small cell lung cancer patients. Oncotarget 2015; 6: 34300-34308.
24. Hardy D, Liu CC, Xia R et al. Racial disparities and treatment trends in a large cohort of elderly black and white patients with nonsmall cell lung cancer. Cancer 2009; 115: 2199-2211.
25. Bollig-Fischer A, Chen W, Gadgeel SM et al. Racial diversity of actionable mutations in non-small cell lung cancer. J Thorac Oncol 2015; 10: 250-255.
26. Tomlinson DC, Hurst CD, Knowles MA. Knockdown by shRNA identifies S249C mutant FGFR3 as a potential therapeutic target in bladder cancer. Oncogene 2007; 26: 5889-5899.
27. Rosty C, Aubriot MH, Cappellen D et al. Clinical and biological characteristics of cervical neoplasias with FGFR3 mutation. Mol Cancer 2005; 4: 15.
28. Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. Nature 2014; 511: 543-550.
29. Helsten T, Elkin S, Arthur E et al. The FGFR Landscape in Cancer: Analysis of 4,853 Tumors by Next-Generation Sequencing. Clin Cancer Res 2016; 22: 259-267.
30. Lim SM, Kim HR, Cho EK et al. Targeted sequencing identifies genetic alterations that confer Primary resistance to EGFR tyrosine kinase inhibitor (Korean Lung Cancer Consortium). Oncotarget 2016.
31. Arai D, Hegab AE, Soejima K et al. Characterization of the cell of origin and propagation potential of the fibroblast growth factor 9-induced mouse model of lung adenocarcinoma. J Pathol 2015; 235: 593-605.
32. Yin Y, Betsuyaku T, Garbow JR et al. Rapid induction of lung adenocarcinoma by fibroblast growth factor 9 signaling through FGF receptor 3. Cancer Res 2013; 73: 5730-5741.
33. Liu X, Zhang W, Geng D et al. Clinical significance of fibroblast growth factor receptor-3 mutations in bladder cancer: a systematic review and meta-analysis. Genet Mol Res 2014; 13: 1109-1120.
34. Milowsky MI, Dittrich C, Duran I et al. Phase 2 trial of dovitinib in patients with progressive FGFR3-mutated or FGFR3 wild-type advanced urothelial carcinoma. Eur J Cancer 2014; 50: 3145-3152.
35. Sequist LV, Cassier P, Varga A et al. Phase I study of BGJ398, a selective pan-FGFR inhibitor in genetically preselected advanced solid tumors. In American Association for Cancer Research 2014 Congress. San Diego, CA, USA: 2014.
36. Dutt A, Ramos AH, Hammerman PS et al. Inhibitor-sensitive FGFR1 amplification in human non-small cell lung cancer. PLoS One 2011; 6: e20351.
37. Dutt A, Salvesen HB, Chen TH et al. Drug-sensitive FGFR2 mutations in endometrial carcinoma. Proc Natl Acad Sci U S A 2008; 105: 8713-8717.
38. Weiss J, Sos ML, Seidel D et al. Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer. Sci Transl Med 2010; 2: 62ra93.
39. Dienstmann R, Rodon J, Prat A et al. Genomic aberrations in the FGFR pathway: opportunities for targeted therapies in solid tumors. Ann Oncol 2014; 25: 552-563.

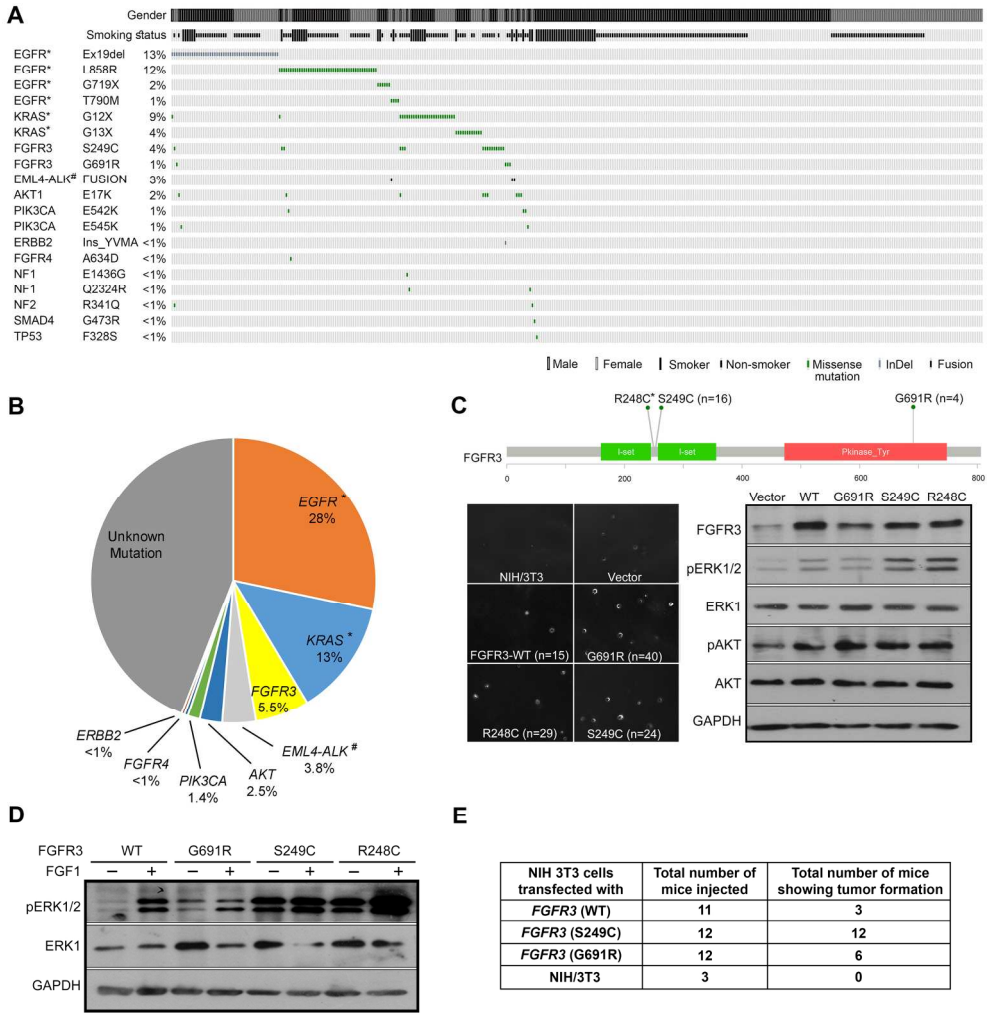
Figure legends

Figure-1: Recurrently mutated genes in lung adenocarcinoma. (A) Validated mutations in 363 samples identified by single base extension based mass spectrometry were visualized using OncoPrinter tool available at cBioPortal. The asterisk (*) denotes genes genotyped using TaqMan and SNaPShot assays in addition to the mass spectrometry. # Fusion frequency was determined using fluorescent in-situ hybridization in 79 of 363 patients. (B) Pie-chart representation of the frequency of clinically relevant genes observed in 363 lung adenocarcinoma patients of Indian origin. (C) Upper panel: Schematic diagram to represent position of point mutations identified in *FGFR3* using next-generation sequencing. Numbers of patients found to be mutated by mass-spectrometry based genotyping are denoted in brackets. Lower left panel: Representative pictures and soft agar colony count (averaged from triplicate) are shown for NIH/3T3 clones. Lower right panel: Immunoblot analysis of NIH/3T3 clones using anti- FGFR3, total- and phospho-ERK1/2 and AKT antibody. (D) Immunoblot analysis of NIH/3T3 clones with (50 ng/ml) and without FGF1 ligand treatment for total- and phospho- ERK1/2. GAPDH was used as loading control in immunoblots. (E) *In-vivo* tumorigenicity of NIH/3T3 cells expressing *FGFR3* mutants and wild-type in NOD-SCID mice is shown. Detailed figure legend can be found in Supplementary materials.

Figure-2: Transformed NIH/3T3 cells and xenografts are sensitive to FGFR inhibitor. (A) Immunoblot analysis of NIH/3T3 clones treated with FGF1 (50 ng//ml) followed by FGFR inhibitor PD173074 (2μM) is shown. GAPDH was used as loading control. (B) Soft agar colony count (averaged from 3 replicates) and IC-50 values of NIH/3T3 clones expressing wild-type or mutant *FGFR3*, treated with increasing concentration of PD173074 is shown. (C) Upper panel: NIH/3T3 xenografts developed into NOD-SCID mice were treated with FGFR inhibitor BGJ-398 or vehicle for 21 days. CT-scan and a readout for relative ^{18}F-FDG uptake is shown by a gradient color code with red indicating as maximum uptake. Lower panel: Immuno-histochemical staining of total- and phospho- ERK1/2 is shown in xenografts treated with drug and vehicles. (D) The plot shows tumor size (normalized to the size at day 0 of drug treatment) during the course of drug treatment indicating a reduced tumor size in drug-treated xenografts. (E) Clinical follow-up of total 205 patients for up to 62

months was used for Kaplan-Meier analysis. *EGFR* positive patients received Gefitinib as a regular therapeutic regimen while rest of the patients received conventional chemotherapy. The table below the plot indicates patients at risk during the course of 60 months and median survival for each mutant group. Detailed figure legend can be found in Supplementary materials.

Figure-1

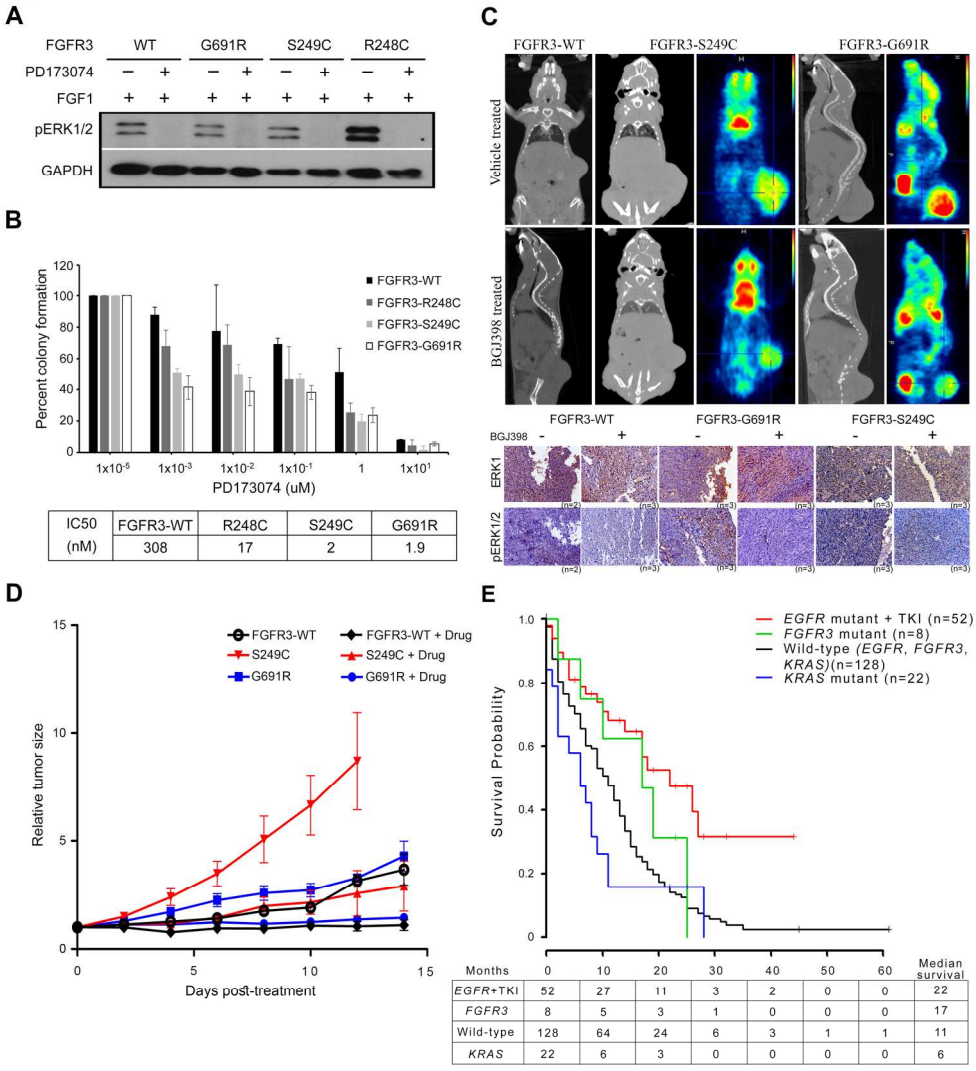


NIH 3T3 cells transfected with	Total number of mice injected	Total number of mice showing tumor formation
FGFR3 (WT)	11	3
FGFR3 (S249C)	12	12
FGFR3 (G691R)	12	6
NIH/3T3	3	0

Figure-1: Recurrently mutated genes in lung adenocarcinoma. (A) Validated mutations in 363 samples identified by single base extension based mass spectrometry were visualized using OncoPrinter tool available at cBioPortal. The asterisk (*) denotes genes genotyped using TaqMan and SNaPShot assays in addition to the mass spectrometry. # Fusion frequency was determined using fluorescent in-situ hybridization in 79 of 363 patients. (B) Pie-chart representation of the frequency of clinically relevant genes observed in 363 lung adenocarcinoma patients of Indian origin. (C) Upper panel: Schematic diagram to represent position of point mutations identified in FGFR3 using next-generation sequencing. Numbers of patients found to be mutated by mass-spectrometry based genotyping are denoted in brackets. Lower left panel: Representative pictures and soft agar colony count (averaged from triplicate) are shown for NIH/3T3 clones. Lower right panel: Immunoblot analysis of NIH/3T3 clones using anti- FGFR3, total- and phospho- ERK1/2 and AKT antibody. (D) Immunoblot analysis of NIH/3T3 clones with (50 ng/ml) and without FGF1 ligand treatment for total- and phospho- ERK1/2. GAPDH was used as loading control in immunoblots. (E) In-vivo tumorigenicity of NIH/3T3 cells expressing FGFR3 mutants and wild-type in NOD-SCID mice is shown. Detailed figure legend can be found in Supplementary materials.

Figure 1
189x200mm (300 x 300 DPI)

Figure-2



IC50 (nM)	FGFR3-WT	R248C	S249C	G691R
	308	17	2	1.9

Months	0	10	20	30	40	50	60	Median survival
EGFR+TKI	52	27	11	3	2	0	0	22
FGFR3	8	5	3	1	0	0	0	17
Wild-type	128	64	24	6	3	1	1	11
KRAS	22	6	3	0	0	0	0	6

Figure-2: Transformed NIH/3T3 cells and xenografts are sensitive to FGFR inhibitor. (A) Immunoblot analysis of NIH/3T3 clones treated with FGF1 (50 ng//ml) followed by FGFR inhibitor PD173074 (2μM) is shown. GAPDH was used as loading control. (B) Soft agar colony count (averaged from 3 replicates) and IC-50 values of NIH/3T3 clones expressing wild-type or mutant FGFR3, treated with increasing concentration of PD173074 is shown. (C) Upper panel: NIH/3T3 xenografts developed into NOD-SCID mice were treated with FGFR inhibitor BGJ-398 or vehicle for 21 days. CT-scan and a readout for relative 18F-FDG uptake is shown by a gradient color code with red indicating as maximum uptake. Lower panel: Immuno-histochemical staining of total- and phospho- ERK1/2 is shown in xenografts treated with drug and vehicles. (D) The plot shows tumor size (normalized to the size at day 0 of drug treatment) during the course of drug treatment indicating a reduced tumor size in drug-treated xenografts. (E) Clinical follow-up of total 205 patients for up to 62 months was used for Kaplan-Meier analysis. EGFR positive patients received Gefitinib as a regular therapeutic regimen while rest of the patients received conventional chemotherapy. The table below the plot indicates patients at risk during the course of 60 months and median survival for each mutant group. Detailed figure legend can be found in Supplementary materials.

Figure 2
189x212mm (300 x 300 DPI)

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

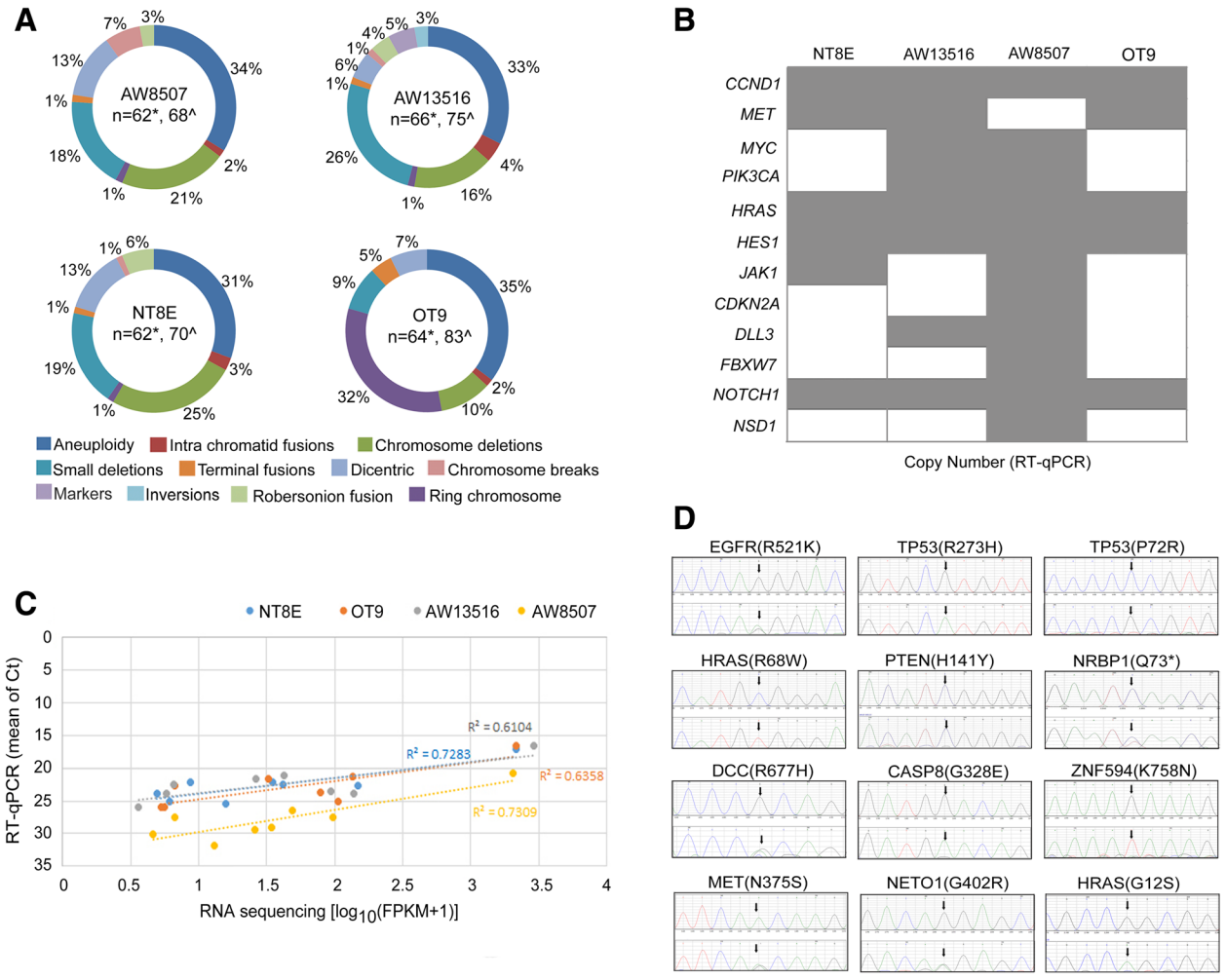


Fig. 1 Genomic alterations identified in HNSCC cell lines. **a** Chromosomal aberrations identified by 25 independent karyotype of each cell line is represented in circular form. Chromosome numbers in each cell line are indicated by n, as observed from karyotype (*) and predicted by SNP array (^). **b** 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

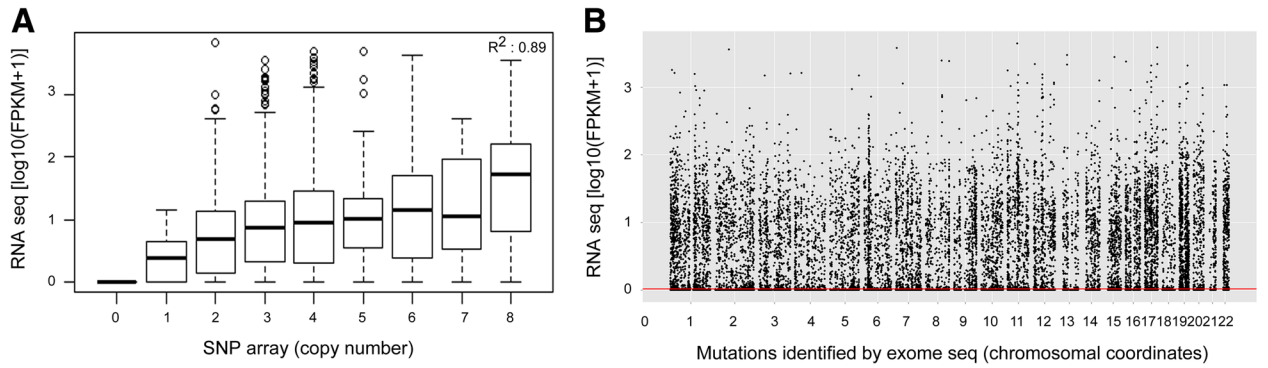


Fig. 2 Integration of copy-number, gene expression and single nucleotide variants. **a** Box plot representations showing focal copy number variations identified by SNP array on X-axis and gene expression quantified by RNA sequencing on Y-axis. Focally amplified or deleted genes from all four cell lines 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 %

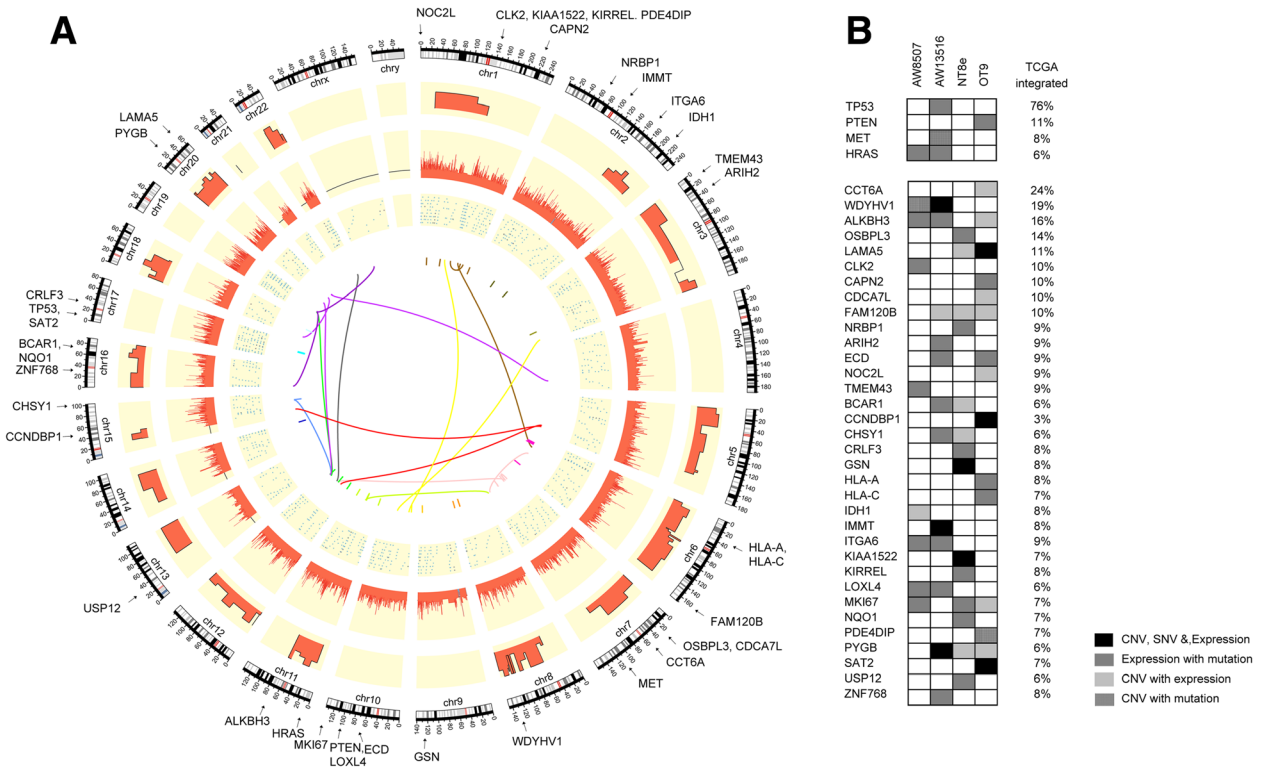


Fig. 3 Integrative genomic landscape of HNSCC. **a** Circos plot representations of genes identified by integrated genomic analysis of HNSCC cell lines. From outside to inside: karyotype, CNVs, Gene expression (FPKM), SNVs and translocations. Red colour indicates copy number gain or higher gene expression 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

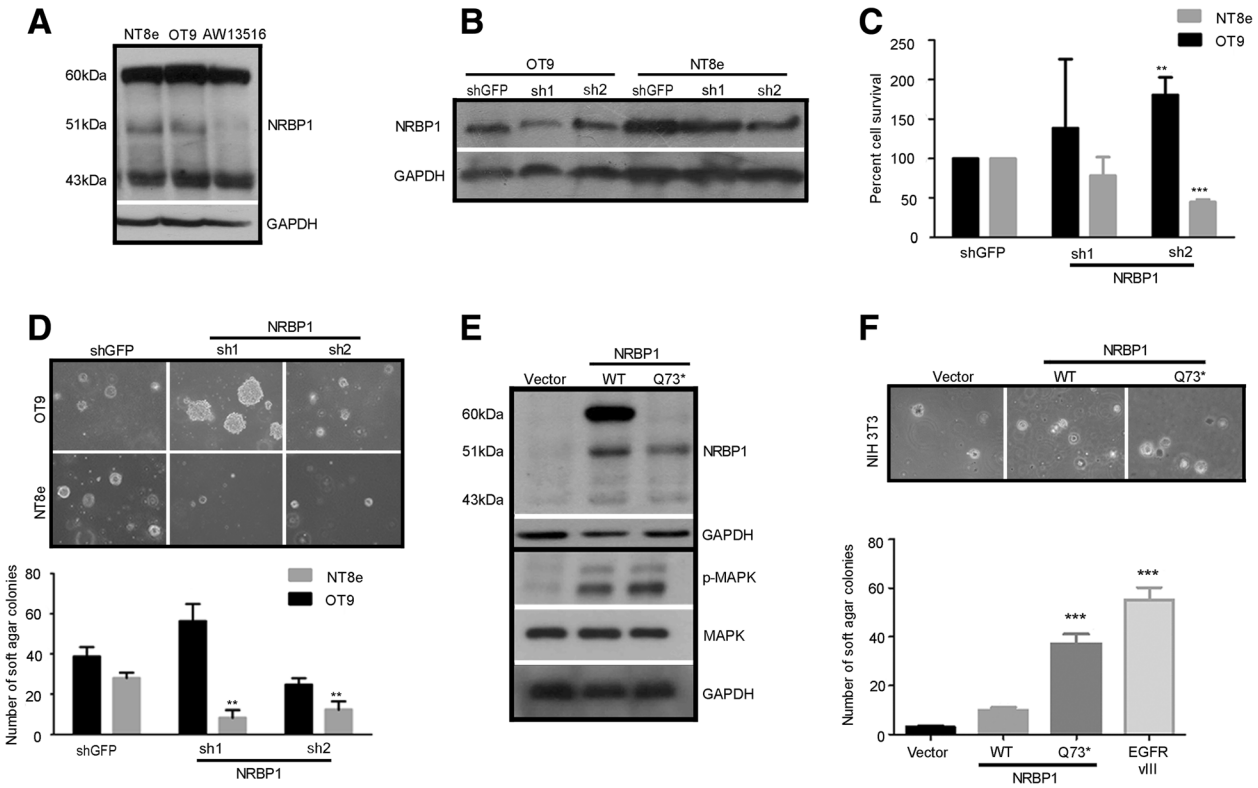


Fig. 4 Mutant *NRBP1* is required for tumor cell survival and is oncogenic in NIH3T3 cells. Knockdown of mutant *NRBP1* expression with shRNA inhibits transformation and survival of HNSCC cell lines. **a** Western blot analysis of total NRBP1 expression level in HNSCC cell lines. *NRBP1* encodes for 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

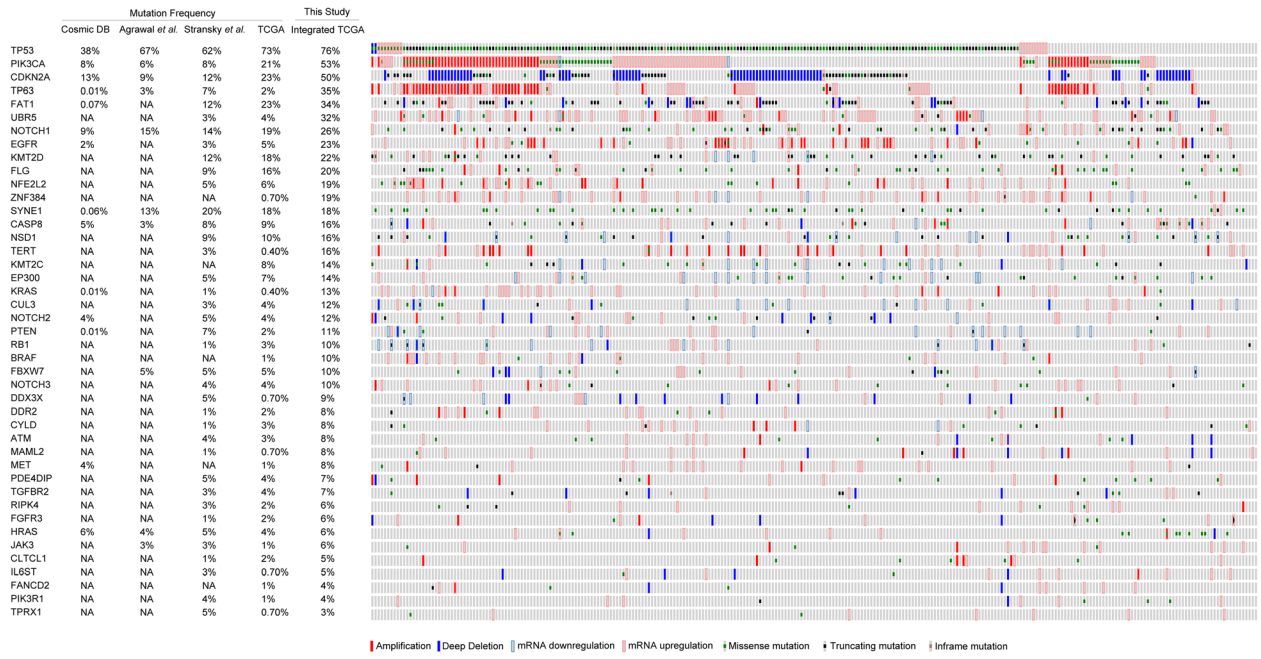


Fig. 5 Integrative genomic landscape of HNSCC tumors in TCGA dataset. Heatmap representation of hallmark genes in 279 HNSCC samples from TCGA study with frequency of alterations based on integrated CNVs, gene expression and SNVs. Frequency of alterations in the gene based on Cosmic database, Agrawal 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.

FULL PAPER

BJC

British Journal of Cancer (2015) 112, 1958–1965 | doi: 10.1038/bjc.2015.121

Keywords: HPV detection; human cancer; next-generation sequencing (NGS)

NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome

P Chandrani[1,2], V Kulkarni[1,2], P Iyer[1], P Upadhyay[1], R Chaubal[1], P Das[1], R Mulherkar[1], R Singh[1] and A Dutt*[,1]

[1]*Advanced Centre for Treatment, Research and Education in Cancer, Tata Memorial Centre, Kharghar, Navi Mumbai, Maharashtra 410210, India*

Background: Human papilloma virus (HPV) accounts for the most common cause of all virus-associated human cancers. Here, we describe the first graphic user interface (GUI)-based automated tool 'HPVDetector', for non-computational biologists, exclusively for detection and annotation of the HPV genome based on next-generation sequencing data sets.

Methods: We developed a custom-made reference genome that comprises of human chromosomes along with annotated genome of 143 HPV types as pseudochromosomes. The tool runs on a dual mode as defined by the user: a 'quick mode' to identify presence of HPV types and an 'integration mode' to determine genomic location for the site of integration. The input data can be a paired-end whole-exome, whole-genome or whole-transcriptome data set. The HPVDetector is available in public domain for download: http://www.actrec.gov.in/pi-webpages/AmitDutt/HPVdetector/HPVDetector.html.

Results: On the basis of our evaluation of 116 whole-exome, 23 whole-transcriptome and 2 whole-genome data, we were able to identify presence of HPV in 20 exomes and 4 transcriptomes of cervical and head and neck cancer tumour samples. Using the inbuilt annotation module of HPVDetector, we found predominant integration of viral gene *E7*, a known oncogene, at known 17q21, 3q27, 7q35, Xq28 and novel sites of integration in the human genome. Furthermore, co-infection with high-risk HPVs such as 16 and 31 were found to be mutually exclusive compared with low-risk HPV71.

Conclusions: HPVDetector is a simple yet precise and robust tool for detecting HPV from tumour samples using variety of next-generation sequencing platforms including whole genome, whole exome and transcriptome. Two different modes (quick detection and integration mode) along with a GUI widen the usability of HPVDetector for biologists and clinicians with minimal computational knowledge.

Human papilloma viral (HPV) infections has been associated with various types of cancer. Epidemiological studies indicate that about 90% of cervical cancers, 90–93% of anal canal cancers, 12–63% of oropharyngeal cancers, 36–40% of penile cancers, 40–64% of vaginal cancers and 40–51% of vulvar cancers are attributable to HPV infection (Munoz *et al*, 2003; Shukla, 2009). Currently, HPV detections are primarily carried out using PCR-based MY09/11 and CPI/II systems (Kleter *et al*, 1998). Other techniques used include the hybridisation-based SPF LiPA method, signal-amplification assays (Hybrid Capture 2 and Cervista) and nucleic-acid-based amplification-like microarray, real-time PCR-based methods (COBAS 4800 real-time test) (Kleter *et al*, 1998; Brink *et al*, 2007; Abreu *et al*, 2012). These technologies come with limitations to detect minor, low-abundance HPV genotypes and a complex mixture of co-infections that can be a negative determinant of the clinical outcome (Mendez *et al*, 2005; Trottier *et al*, 2006). Next-generation sequencing (NGS) technologies overcomes such limitations, as evident from the recently described

*Correspondence: Dr A Dutt; E-mail: adutt@actrec.gov.in
[2]These authors contributed equally to this work.

Received 31 July 2014; revised 3 March 2015; accepted 7 March 2015; published online 14 May 2015

high-risk HPV genotyping assay for primary cervical cancer screening based on self-collection (Yi *et al*, 2014), using TEN16 or HIVID methodology, and to determine co-infection among the HPV types probed along with their sites on intregration (Johansson *et al*, 2013; Xu *et al*, 2013; Li *et al*, 2013b; Ameur *et al*, 2014; Hu *et al*, 2015). However, there is an unmet need for a simplified tool for biologists with no previous experience or knowledge of informatics to analyse the data generated by whole-exome, transcriptome or genome sequencing using NGS technology to detect the presence of HPV sequences along with their integration sites. There are a variety of gene integration finding tools available that can detect different pathogen insertions in the human genome such as ViralFusionSeq (Li *et al*, 2013a), VirusSeq (Chen *et al*, 2013), VirusFinder (Wang *et al*, 2013), Path-Seq (Kostic *et al*, 2011), RINS (Bhaduri *et al*, 2012), and ReadSCAN (Naeem *et al*, 2013). These tools have their specific third-party needs, and are not specific for HPV detection. They can detect presence of a HPV sequence along with other viruses, but lack information to annotate the region of the HPV genome detected. Here, we describe 'HPVDetector' as a specific *in silico* automated tool that is capable of multi-HPV type detection, their annotation and determination of site of HPV integration utilising raw exome, transcriptome, or whole-genome data as input with minimal requirement for third-party tools.

MATERIALS AND METHODS

HPV detection involves a computational subtraction-based approach, where NGS data are used for alignment against custom-made HPV multi-reference genome sequences to detect the traces of multiple HPV types using an automated pipeline (Figure 1).

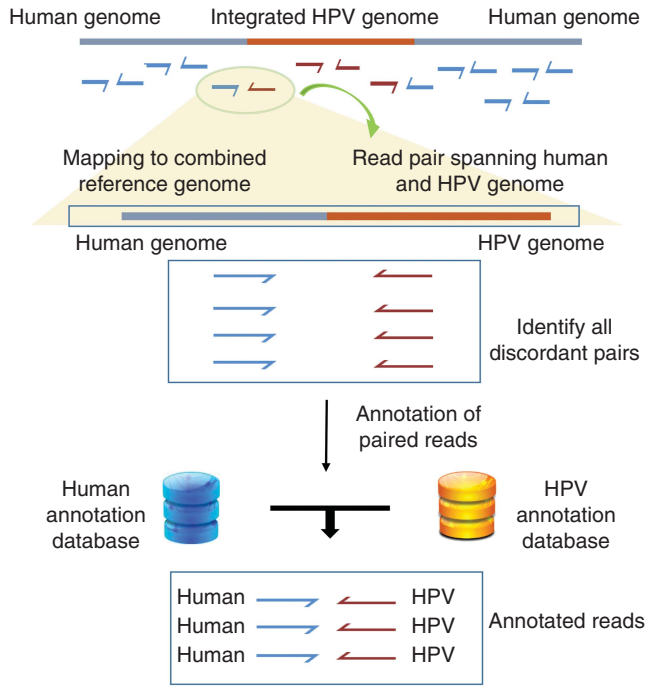


Figure 1. Conceptual workflow of the HPVDetector. The flowchart represents workflow for HPVDetector. Paired-end reads obtained from next-generation sequencing data are aligned to a combined Human–HPV reference database. All discordant read pairs with one read aligning to human and other to the HPV genome are identified and annotated utilising human and HPV database using an inbuilt annotator module.

HPV reference sequences and annotation. As a first step of the pipeline, HPV genomes in fasta format is required. We have acquired GenBank (.gb) files of 143 types of HPVs from a web resource Papillomavirus Episteme (PAVE) (Van Doorslaer *et al*, 2013). We converted these GenBank (.gb) files into fasta files. All these reference sequences were concatenated to compose a multi-fasta sequence using bio-perl modules (Stajich *et al*, 2002). Apart from this, we also parsed the GenBank (.gb) files to generate a HPV gene reference having nucleotide intervals for each gene of each HPV type. This gene reference file was used to annotate the HPV gene.

HPV type and HPV-aligned reads detection. Evaluating the HPV type and HPV-aligned reads is crucial to find HPV in the respective sample. For HPV type detection, we indexed the multi-fasta HPV reference file using BWA aligner followed by alignment of reads to indexed genome (Li and Durbin, 2009). The aligned reads were extracted from the same file using a utility ViewSam from Picard Tools package (http://broadinstitute.github.io/picard/). The alignment files were parsed using UNIX shell program to detect the type of HPV as well as number of reads that align to a particular HPV type. Number of HPV reads were normalised to the total depth of coverage per sample and with respect to different HPV gene sizes.

Assessment of specificity and sensitivity of HPVDetector. We downloaded SiHa whole-genome sequence from Sequence Read Archive database of DDBJ (https://trace.ddbj.nig.ac.jp/DRASearch/; study: SRP048769). The data were converted from SRA to FASTQ using SRAtoolkit. The resulting FASTQ files represents $>36\times$ genome coverage which was further downsampled to $1\times$, $2\times$, $3\times$, $4\times$, $5\times$, $10\times$, $15\times$, $20\times$, $25\times$ and $30\times$ using Picard Toolkit's DownsampleSam function (http://broadinstitute.github.io/picard/). The resulting FASTQ files were used for testing HPV detection using HPVDetector.

Human–HPV integration loci detection. To detect integration sites, we created a custom reference genome comprised of human chromosomes and HPV fasta sequences as pseudochromosomes. HPV genomes were appended to human chromosomes to compose a multi-fasta reference genome. This custom Human–HPV reference genome was then used for aligning reads with short-read aligner BWA. The alignment files were parsed for the reads where one mate is aligned to human chromosome and another to HPV. The Human chromosomal positions, HPV type and HPV reference position were parsed and annotated with a gene reference annotation file acquired from UCSC table browser (Karolchik *et al*, 2004) to get a list of integration sites.

RNA extraction, cDNA synthesis and E6-specific PCR. Total RNA extraction was performed from primary tumours and cell lines using Trizol reagent (Invitrogen, Grand Island, NY, USA) as per the manufacture's instruction and later resolved on 1.2% agarose gel to confirm the RNA integrity. DNase treatment was done using the DNase Free kit (Ambion, Foster City, CA, USA; cat AM1906) followed by first-strand cDNA synthesis taking 2 μg of total RNA using Superscript III kit (Invitrogen, 18080-051). E6 (HPV-16) and GAPDH expression were checked as described previously (Smeets *et al*, 2007).

HPV detection using MY09/11 and PCR primers. MY09/11 primer sequences were taken from previously reported literature (Baay *et al*, 1996). All samples were screened by PCR first using MY09/11 primer. GAPDH was used as internal control for each sample. SiHa cell line (Adler *et al*, 1997) was used as a positive control for HPV and AW13516 cell line (Tatake *et al*, 1990) as a negative control. The PCR reaction was performed in 20 μl volume

containing 10 μl (2 ×) Biomix-Red master mix (Biomix, Port Orange, FL, USA; cat Bio25005), 5 μM each primer, 50 ng gDNA. PCR condition was: initial denaturation: 95 °C, denaturation: 94 °C for 1 min, annealing: 55 °C, (MY09/11, GAPDH), extension: 72 °C for 1 min and final extension at 72 °C for 5 min on a PCR machine (Veriti 96-Well Fast Thermal Cycler, Applied Biosystems, Carlsbad, CA, USA). Primers were designed to amplify 122-, 126- and 120-bp read sequences of HPV16 identified by HPVDetector in SiHa cell line. Primers flanking the human reads were designed to amplify 119 and 290 bp, respectively. These sequences were further validated by Sanger sequencing. All the experiments were repeated at least twice independently. The details of the primers used in the study are provided in Supplementary Table 2.

RESULTS

HPVDetector is a tool to quickly detect hundreds of HPV types from next-generation sequence data without any prerequisite knowledge about virus types. It runs on paired-end sequenced samples. It is composed of two modes or sub pipelines as quick detect and integration detect mode.

Quick detect mode. This mode is to quickly determine the HPV type or types to check whether multiple HPV co-infections are existing or not in a given sample. Quick detect mode starts with alignment of raw paired-end sequencing reads against the custom-made multi-HPV genome using BWA aligner. Computational subtraction of the reads is then carried out, in which HPV-aligned reads are retained using Picard Tools and further processed using UNIX shell program to distinguish reads mapping to different HPV types. Finally, HPVDetector outputs a result file, which enlists one or more HPV type(s) and number of HPV reads.

Integration detect mode. This mode of HPVDetector determines the genomic location of HPV integrant, annotate with HPV gene, human chromosomal loci and human gene. This mode of HPVDetector pipeline starts with alignment of raw reads against a custom-made reference including a pseudochromosome such as the multi-fasta reference genome containing 143 HPV reference sequences and the HG19 human reference genome. Computational subtraction was carried out to retain discordant read pairs where the sequences are aligned to both human as well as HPV genomes. Finally, HPVDetector outputs a result file, which enlists HPV integration loci on the human genome, annotation of HPV genes, human genes and human genome cytobands.

Detection of HPV type integrated in the host genome

Cervical cancer exome sequencing data. We analysed 22 cervical cancer exome sequencing data (generated in-house at ACTREC, unpublished data) to detect the presence of HPV. Among the 22 samples analysed, HPV was detected in 18 cervical samples, with maximum number of reads supporting the HPV16 sequence (Figure 2) (Das *et al*, 2012). We also detected the presence of additional HPV types such as HPV71 (in six samples), HPV82 (in five samples) and HPV31 (in two samples) with variable number of supporting reads as shown in (Figure 3). Co-infection with more than one HPV type is known to be associated with significantly increased risk of cervical intraepithelial neoplasia 2 + and found in 43.2% of HPV-positive women (Liaw *et al*, 2001; Mendez *et al*, 2005; Vaccarella *et al*, 2010; Chaturvedi *et al*, 2011). Six of 22 cervical cancer patients (43%) were found to be co-infected with one or more HPV subtypes in this study using HPVDetector (Figure 3). Interesting to note, based on phylogenetic analysis of HPV types, HPV16 and HPV31 of the virulent alpha 7 group infection occurred in a mutually exclusive manner (in 13 of 22 samples), whereas HPV71 of the alpha 15 subgroup, known to be involved in commensal infections that infected 6 of 14 cervical tumour samples, invariably co-occurred with other HPV subtypes (Schiffman *et al*, 2009; Harari *et al*, 2014). The HPV sequence detected in primary cervical tumour sample were independently validated by directed sequencing in T1094, the only sample with sufficient quality DNA (as shown in Supplementary Figure 1).

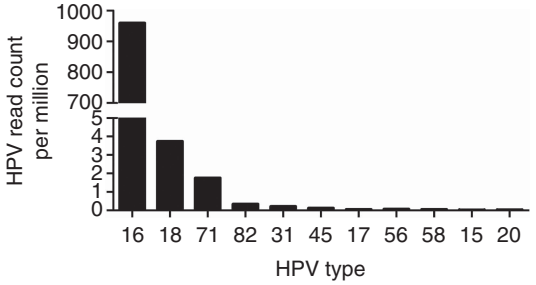


Figure 2. Quantitative representation, by number of reads, of HPV types detected in cervical tumours. The graph represents distribution of total number of HPV reads per million of total reads for all HPV types detected across 17 cervical samples. HPV16 has the highest number of reads across 17 samples followed by HPV(18, 71, 45, 31, 82, 17, 56, 58, 15, 20) in the decreasing order of their read counts.

Tongue squamous cell carcinoma exome and transcriptome data. HPV is an independent risk factor in head and neck squamous cell carcinoma (HNSCC), in particular for oral and oropharyngeal carcinomas (Chaudhary *et al*, 2009; Pannone *et al*, 2011). We analysed whole-exome data from 23 paired and one orphan tongue squamous cell carcinoma (TSCC) sample and 7 HNSCC cell lines (generated in-house at ACTREC, unpublished data). None of the TSCC primary tumours were found to be HPV positive, as reported earlier (Siebers *et al*, 2008; Patel *et al*, 2014; Tsimplaki *et al*, 2014). The absence of HPV infection were further validated by PCR using MY09/11 and E6 on genomic DNA and cDNA, respectively, suggesting a low false-negative feature of the HPVDetector primers (Supplementary Figure 2). At the same time, among the cell lines, NT8e cells (Mulherkar *et al*, 1997) of seven cell lines analysed was found to be positive for HPV71. Next, we analysed whole-transcriptome data of 17 TSCC and 6 TSCC cell lines (generated in-house at ACTREC, unpublished data) using the HPVDetector. Three of 17 primary tumours were found to be HPV18 positive. In addition, HPV18 reads were found in HEP2 cell line, consistent with earlier reports in literature (Ogura *et al*, 1993). The HPV18 genes (E1, E6, and E7) were validated in Hep2 cell line by PCR and Sanger sequencing (as shown in Supplementary Figure 1 and Table 2).

Gall bladder and liposarcoma exome and whole-genome data. We analysed 13 gall bladder cancer whole-exome, 1 gall bladder cancer whole-transcriptome and 1 liposarcoma whole-genome sequence data (generated in-house at ACTREC, unpublished data). No trace of the HPV sequence was detected in these samples.

Assessment of specificity and sensitivity of HPVDetector. SiHa cell line developed from a cervical squamous cell carcinoma patient represents single-copy integration of HPV16 (el Awady *et al*, 1987) We analysed SiHa whole-genome sequence using HPVDetector. Consistent with a published report (Hu *et al*, 2015), HPVDetector could detect integration at chr13 intragenic location of KLF5—KLF12 genes and other regions (Supplementary Table 1). The integration was validated by PCR followed by sequencing (Supplementary Figure 3).

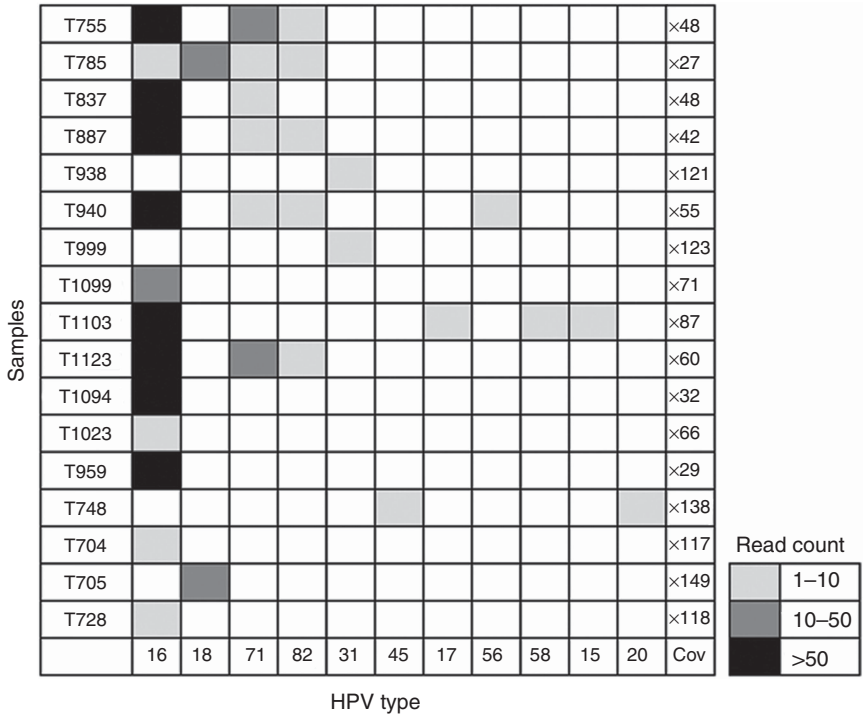


Figure 3. HPV gene integration frequency across different cervical cancer samples. Heatmap representation of HPV types detected across 17 cervical cancer samples. HPV16 in 13 samples, HPV18 in 2 samples, HPV71 in 6 samples, HPV82 in 5 samples, HPV31 in 2 samples, HPV45 in 5 samples, and HPV17, 56, 58, 15, 20 were detected in 1 sample, each. Total coverage of the exome sequencing is indicated in the last column 'cov'. On the basis of read count, the abundance of HPV is graded in different samples: read counts 1–10 as light grey; read counts 10–50 as dark grey; and read counts >50 as black.

Sesnsitivity. To determine the sensitivity of HPVDetector, we downsampled the SiHa genome using a 'downsampling' method, a Picard Toolkit's DownsampleSam function (http://broadinstitute.github.io/picard/) (Meynert *et al*, 2014) to generate varying coverage of the SiHa whole-genome data ranging from 1 × to 30 × coverage, and analysed using HPVDetector. Reads supporting presence of HPV reads linearly increased as a function of increasing coverage from 1 × to 25 × coverage. Beyond 25 ×, no significant increase in HPV reads were found, suggesting saturation of genome coverage (Figure 4). In addition, among primary tumours, two pairs of HPV56 reads detected by the HPVDetector in T9440 as described in Supplementary Table 1 were validated earlier by Luminex array and SPF1/2 (Das *et al*, 2012). Taken together, this suggests HPVDetector could detect reads with as low as 1 × genome coverage with reads supported by as low as just two paired reads.

Specificity. Having benchmarked the HPVDetector against SiHa for sensitivity, next we tweaked the SiHa whole-genome sequence data to test specificity of the tool by taking reverse (not complement) of the SiHa genome to simulate as a random sequence but retaining composition of nucleotides and genome complexity, using an in-house perl script. We found no spurious HPV reads when the SiHa whole-genome sequence was reversed, suggesting the HPVDetector is specific to detect true HPV traces. Further, to address the issue of specificity among primary tumours, we performed another round of functional validation on tongue squamous tumours that were found HPV negative based on HPVDetector (Supplementary Table 1) and validated by My09/11 primers using genomic DNA. We analysed the expression of HPV E6 (Supplementary Figure 2b) in these samples. All samples were found negative for HPV presence. This suggests that the tool has low false negative.

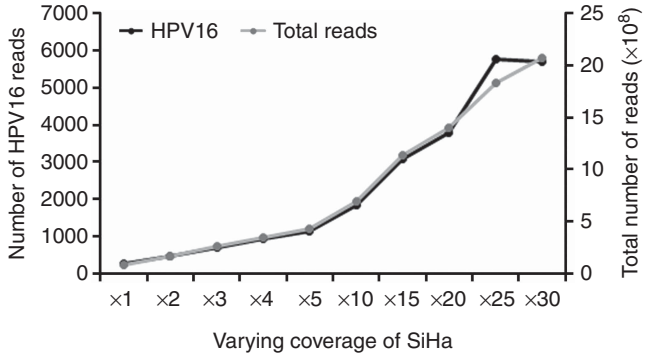


Figure 4. Sensitivity of HPVDetector as a function of increasing genome sequence coverage. SiHa WGS data were downsampled to the coverage of 1 ×, 2 ×, 3 ×, 4 ×, 5 ×, 10 ×, 15 ×, 20 ×, 25 ×, and 30 ×. HPV16 reads (black) were counted using HPVDetector at varying coverage and plotted along with total number of reads (grey).

Annotation of the HPV genome integrated in the host genome. To enable accurate gene annotation of the HPV genome sequenced, we prepared a gene annotation database of 143 HPV types from PAVE database (Van Doorslaer *et al*, 2013). Thirty-two reads of viral ORFs were found in 5 of 11 cervical tumours positive for HPV-16. Following the normalisation for the total number of reads against the length of individual genes, the viral gene *E7* was found to be predominantly represented among the cervical tumours infected with HPV16, followed by E4, E5 and E6, in decreasing order (Figure 5). Of these genes found to be enriched among all the integrants, it is interesting to note that the viral proteins E6 and E7 function as oncogenes by regulating the known

human tumour suppressors, p53 and pRb, respectively (Lu *et al*, 2003; Yim and Park, 2005).

Determination of the HPV integration sites in the host genome. We identified 55 integration sites in 7 cervical cancer tumour samples T1099, T1123, T755, T887, T938, T1094, and T959 and 1 head and neck tumour sample using the HPVDetector (Supplementary Table 1). In this study, chromosomal loci 17q21, 3q27, 7q35, and Xq28 were observed with higher frequency compared with other loci for HPV integration, as reported earlier (Thorland *et al*, 2003). Interesting to note, we found HPV integration in the following fragile regions—(1p, 1q, 2p, 2q, 3p, 3q, 4p, 4q, 5p, 5q, 7q, 9q, 10q, 11p, 11q, 12p, 12q, 13q, 15q, 17q, 18q, 22q, Xp and Xq) that are prone to chromosome breaks to facilitate foreign DNA integration (Figure 6) (Smith *et al*, 2006). In T1123 and T755 HPV16 integration sites were detected at chr1q42.3 and chr3q23, respectively, identical to as reported earlier (Wentzensen *et al*, 2004; Schmitz *et al*, 2012). In addition, in T755 integration of HPV16 were found within the coding region at SLC25A36, a pyrimidine nucleotide carrier. This site of integration were also determined in T755 and T1123 samples using the APOT assay, as described earlier (Das *et al*, 2012) (Supplementary Table 1).

In total, we analysed 116 exome, 23 transcriptome and 2 whole-genome sequencing data, out of which we have detected presence of HPV in 20 exome and 4 transcriptome data (Table 1).

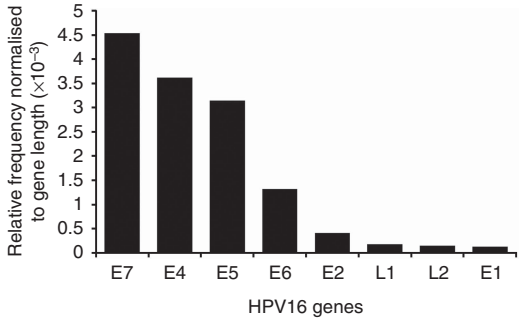


Figure 5. Relative frequency of integration of HPV genes in cervical carcinoma. HPV16 reads were annotated using an inbuilt annotation module of the HPVDetector to identify the viral genes. Number of reads per viral gene were normalised to the gene length, and frequency reads for individual genes are plotted, as shown.

DISCUSSION

HPV accounts for the most common cause of all virus-associated human cancers. However, despite large-scale genome-wide DNA sequencing efforts of the cancer genome, there is no dedicated

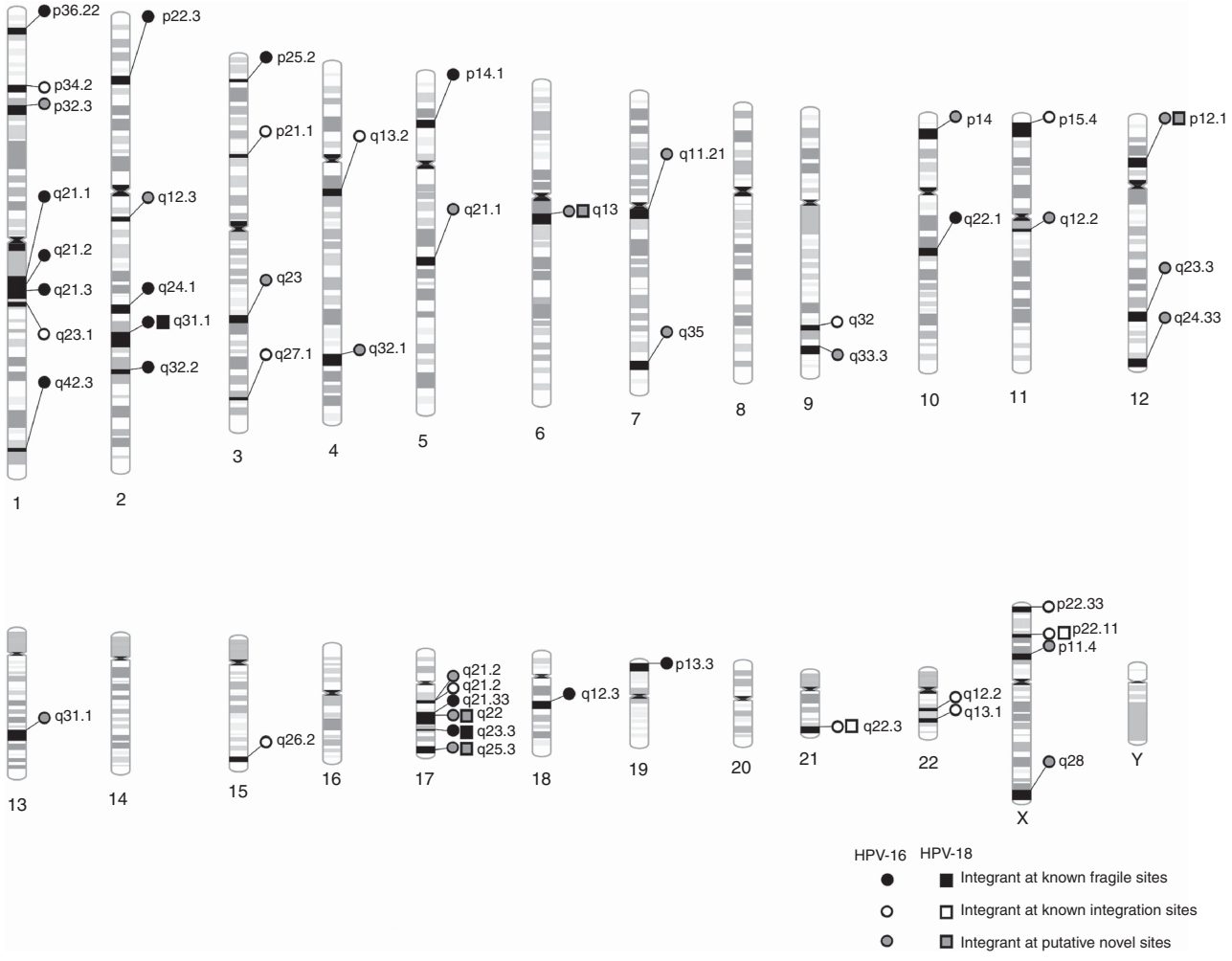


Figure 6. Schematic representation of all HPV16 and 18 integration sites in the human genome detected across cervical cancer samples using HPVDetector. Site of integration as determined by HPVDetector in cervical cancer samples is shown. HPV16 integration sites are depicted by circles and HPV18 by rectangles. Black, open and grey circles or rectangles represent integrations at known fragile sites, at known integration sites and at novel sites, respectively.

Table 1. Summary of HPV detection in all samples

Study type	Samples tested	Presence of virus in samples
Cervical cancer exome	36 (22 tumour, 14 paired normal)	18
SiHa cell line WGS	1	1
HNSCC cell line exome	7	1
TSCC exome	47 (24 tumour, 23 paired normal)	0
Gall bladder exome	26 (13 tumour, 13 paired normal)	1
TSCC transcriptome	17 (11 tumour, 6 paired normal)	3
HNSCC cell line transcriptome	5	1
Gall bladder transcriptome	1	0
Liposarcoma WGS	1	0
Total number of samples	141	25

Abbreviations: HNSCC = head and neck squamous cell carcinoma; HPV = human papilloma virus; TSCC = tongue squamous cell carcinoma; WGS = whole-genome sequencing.

informatics tool to rapidly detect the presence of HPV in these genomes, in an exclusive manner. There are indeed a variety of gene integration finding tools available that can detect different pathogen insertions in the human genome, such as ViralFusionSeq, VirusSeq, VirusFinder, Path-Seq, RINS, and ReadSCAN. These sophisticated tools although have their specific third-party needs, necessitate intense computational infrastructure, cannot be run without specialised and advanced computational expertise of the researcher, and more importantly are not specific for HPV detection, *per se*—for example, lacks information to annotate the region of the HPV genome to predict the integrated viral gene, of which some are known to function as oncogenes.

We present a new user-friendly *in silico* tool 'HPVDetector' as a unique tool to analyze NGS data to detect HPV sequences for non-computational biologists. Using the HPVDetector tool, we have detected 55 integration sites from the cervical exome and head and neck transcriptome data set. The tool allowed us to perform a comprehensive analysis to generate the information for co-occurrence of HPV subtypes across cervical cancer patients that is known to affect the clinical outcome of the disease. In addition, our finding of significant enrichment of viral gene *E7* > *E4* > *E5* > *E6* reads among the cervical tumour samples, using the inbuilt annotation module of the HPVDetector, is consistent with the known biology of HPV genes and their role in carcinogenesis, as *E6* and *E7* are known viral oncogenes. This unique feature of the HPVDetector with an inbuilt HPV annotation module could potentially be helpful to understand the function of other HPV ORFs with unknown function by studying their incidence against varying tumour stage and types. Although the analysis of cervical tumours were restricted to its exome data set, a complete spectrum of the load of viral genes present in a sample can similarly be determined using the whole-genome data as input to the HPVDetector.

HPVDetector demonstrate a low false-negative and false-positive rate that can detect HPV reads at as low as 1 × genome coverage. Reads supported by even two paired reads were found to be credible. No viral reads were detected across 54 head and neck primary tumour samples of Indian origin, as reported earlier (Siebers *et al*, 2008; Patel *et al*, 2014; Tsimplaki *et al*, 2014), but detected a low-risk HPV71 in a cell line that could be validated by performing MY09/11 PCR on the primary tumours as shown in Supplementary Figure 2. On the other hand, all the four HPV reads detected across different tumour types using HPVDetector could be validated by directed PCR followed by Sanger sequencing. One interesting utility of the HPVDetector would be to explore for HPV reads in NGS data from different cancer types. We analysed 13 gall bladder exome, 1 gall bladder transcriptome and 1 liposarcoma whole-genome sequencing data using HPVDetector. No HPV reads were found in these samples in this study.

Another critical feature of the HPVDetector is determination of HPV integration sites at the host genome. These integrations are known to occur at preferred regions of the genome (Thorland *et al*, 2003; Matovina *et al*, 2009). Using the integration site detection feature of the HPVDetector, we detected integration at various chromosomal locations (for e.g., 1p, 2p, 2q, 3p, 3q), some with significant overlap to the known fragile sites in literature and at several novel sites as summarised in Figure 6 and Supplementary Table 1. In summary, HPVDetector is a simple yet precise and robust tool for detecting HPV from tumour samples using variety of NGS platforms including whole genome, whole exome and transcriptome. Two different modes (quick detection and integration mode) along with a GUI widen the usability of HPVDetector for biologists with minimal computational knowledge (as described in the attached supplementary 'HPVDetector User Guide' including Supplementary Figures 4 and 5).

ACKNOWLEDGEMENTS

Prasad Kanvinde for the help with web hosting of HPVDetector and other IT-related support at ACTREC, Tata Memorial Centre, Mumbai. Dhwanit Shah for help with coding of the tool; all members of the Dutt laboratory for critically reviewing the manuscript. Genotypic Inc., SciGenome Pvt. Ltd and Sandor Proteomics for providing sequencing services. AD 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). PC and PI are supported by a senior research fellowship from ACTREC. PU is supported by a 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.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Abreu AL, Souza RP, Gimenes F, Consolaro ME (2012) A review of methods for detect human Papillomavirus infection. *Virol J* **9**: 262.

Adler K, Erickson T, Bobrow M (1997) High sensitivity detection of HPV-16 in SiHa and CaSki cells utilizing FISH enhanced by TSA. *Histochem Cell Biol* **108**(4-5): 321–324.

Ameur A, Meiring TL, Bunikis I, Haggqvist S, Lindau C, Lindberg JH, Gustavsson I, Mbulawa ZZ, Williamson AL, Gyllensten U (2014) Comprehensive profiling of the vaginal microbiome in HIV positive women using massive parallel semiconductor sequencing. *Sci Rep* **4**: 4398.

Baay MF, Quint WG, Koudstaal J, Hollema H, Duk JM, Burger MP, Stolz E, Herbrink P (1996) Comprehensive study of several general and type-specific primer pairs for detection of human papillomavirus DNA by PCR in paraffin-embedded cervical carcinomas. *J Clin Microbiol* **34**(3): 745–747.

Bhaduri A, Qu K, Lee CS, Ungewickell A, Khavari PA (2012) Rapid identification of non-human sequences in high-throughput sequencing datasets. *Bioinformatics* **28**(8): 1174–1175.

Brink AA, Snijders PJ, Meijer CJ (2007) HPV detection methods. *Dis Markers* **23**(4): 273–281.

Chaturvedi AK, Katki HA, Hildesheim A, Rodriguez AC, Quint W, Schiffman M, Van Doorn LJ, Porras C, Wacholder S, Gonzalez P, Sherman ME, Herrero R. CVT Group (2011) Human papillomavirus infection with multiple types: pattern of coinfection and risk of cervical disease. *J Infect Dis* **203**(7): 910–920.

Chaudhary AK, Singh M, Sundaram S, Mehrotra R (2009) Role of human papillomavirus and its detection in potentially malignant and malignant head and neck lesions: updated review. *Head Neck Oncol* **1**: 22.

Chen Y, Yao H, Thompson EJ, Tannir NM, Weinstein JN, Su X (2013) VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics* **29**(2): 266–267.

Das P, Thomas A, Mahantshetty U, Shrivastava SK, Deodhar K, Mulherkar R (2012) HPV genotyping and site of viral integration in cervical cancers in Indian women. *PLoS One* **7**(7): e41012.

el Awady MK, Kaplan JB, O'Brien SJ, Burk RD (1987) Molecular analysis of integrated human papillomavirus 16 sequences in the cervical cancer cell line SiHa. *Virology* **159**(2): 389–398.

Harari A, Chen Z, Burk RD (2014) Human papillomavirus genomics: past, present and future. *Curr Probl Dermatol* **45**: 1–18.

Hu Z, Zhu D, Wang W, Li W, Jia W, Zeng X, Ding W, Yu L, Wang X, Wang L, Shen H, Zhang C, Liu H, Liu X, Zhao Y, Fang X, Li S, Chen W, Tang T, Fu A, Wang Z, Chen G, Gao Q, Li S, Xi L, Wang C, Liao S, Ma X, Wu P, Li K, Wang S, Zhou J, Wang J, Xu X, Wang H, Ma D (2015) Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nat Genet* **47**(2): 158–163.

Johansson H, Bzhalava D, Ekstrom J, Hultin E, Dillner J, Forslund O (2013) Metagenomic sequencing of 'HPV-negative' condylomas detects novel putative HPV types. *Virology* **440**(1): 1–7.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**(Database issue): D493–D496.

Kleter B, van Doorn LJ, ter Schegget J, Schrauwen L, van Krimpen K, Burger M, ter Harmsel B, Quint W (1998) Novel short-fragment PCR assay for highly sensitive broad-spectrum detection of anogenital human papillomaviruses. *Am J Pathol* **153**(6): 1731–1739.

Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RG, Getz G, Meyerson M (2011) PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat Biotechnol* **29**(5): 393–396.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754–1760.

Li JW, Wan R, Yu CS, Co NN, Wong N, Chan TF (2013a) ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. *Bioinformatics* **29**(5): 649–651.

Li W, Zeng X, Lee NP, Liu X, Chen S, Guo B, Yi S, Zhuang X, Chen F, Wang G, Poon RT, Fan ST, Mao M, Li Y, Li S, Wang J, Jianwang, Xu X, Jiang H, Zhang X (2013b) HIVID: an efficient method to detect HBV integration using low coverage sequencing. *Genomics* **102**(4): 338–344.

Liaw KL, Hildesheim A, Burk RD, Gravitt P, Wacholder S, Manos MM, Scott DR, Sherman ME, Kurman RJ, Glass AG, Anderson SM, Schiffman M (2001) A prospective study of human papillomavirus (HPV) type 16 DNA detection by polymerase chain reaction and its association with acquisition and persistence of other HPV types. *J Infect Dis* **183**(1): 8–15.

Lu DW, El-Mofty SK, Wang HL (2003) Expression of p16, Rb, and p53 proteins in squamous cell carcinomas of the anorectal region harboring human papillomavirus DNA. *Mod Pathol* **16**(7): 692–699.

Matovina M, Sabol I, Grubisic G, Gasperov NM, Grce M (2009) Identification of human papillomavirus type 16 integration sites in high-grade precancerous cervical lesions. *Gynecol Oncol* **113**(1): 120–127.

Mendez F, Munoz N, Posso H, Molano M, Moreno V, van den Brule AJ, Ronderos M, Meijer C, Munoz A. Instituto Nacional de Cancerologia Human Papillomavirus Study Group (2005) Cervical coinfection with human papillomavirus (HPV) types and possible implications for the prevention of cervical cancer by HPV vaccines. *J Infect Dis* **192**(7): 1158–1165.

Meynert AM, Ansari M, FitzPatrick DR, Taylor MS (2014) Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics* **15**: 247.

Mulherkar R, Goud AP, Wagle AS, Naresh KN, Mahimkar MB, Thomas SM, Pradhan SA, Deo MG (1997) Establishment of a human squamous cell carcinoma cell line of the upper aero-digestive tract. *Cancer Lett* **118**(1): 115–121.

Munoz N, Bosch FX, de Sanjose S, Herrero R, Castellsague X, Shah KV, Shah KV, Snijders PJ, Meijer CJ. International Agency for Research on Cancer Multicenter Cervical Cancer Study Group (2003) Epidemiologic classification of human papillomavirus types associated with cervical cancer. *N Engl J Med* **348**(6): 518–527.

Naeem R, Rashid M, Pain A (2013) READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation. *Bioinformatics* **29**(3): 391–392.

Ogura H, Yoshinouchi M, Kudo T, Imura M, Fujiwara T, Yabe Y (1993) Human papillomavirus type 18 DNA in so-called HEP-2, KB and FL cells–further evidence that these cells are HeLa cell derivatives. *Cell Mol Biol (Noisy-le-grand)* **39**(5): 463–467.

Pannone G, Santoro A, Papagerakis S, Lo Muzio L, De Rosa G, Bufo P (2011) The role of human papillomavirus in the pathogenesis of head & neck squamous cell carcinoma: an overview. *Infect Agent Cancer* **6**: 4.

Patel KR, Vajaria BN, Begum R, Desai A, Patel JB, Shah FD, Shukla SN, Patel PS (2014) 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* **43**(4): 293–297.

Schiffman M, Clifford G, Buonaguro FM (2009) Classification of weakly carcinogenic human papillomavirus types: addressing the limits of epidemiology at the borderline. *Infect Agent Cancer* **4**: 8.

Schmitz M, Driesch C, Jansen L, Runnebaum IB, Durst M (2012) Non-random integration of the HPV genome in cervical cancer. *PLoS One* **7**(6): e39632.

Shukla S (2009) Infection of human papillomaviruses in cancers of different human organ sites. *Indian J Med Res* **130**: 222–233.

Siebers TJ, Merkx MA, Slootweg PJ, Melchers WJ, van Cleef P, de Wilde PC (2008) No high-risk HPV detected in SCC of the oral tongue in the absolute absence of tobacco and alcohol–a case study of seven patients. *Oral Maxillofac Surg* **12**(4): 185–188.

Smeets SJ, Hesselink AT, Speel EJ, Haesevoets A, Snijders PJ, Pawlita M, Meijer CJ, Braakhuis BJ, Leemans CR, Brakenhoff RH (2007) A novel algorithm for reliable detection of human papillomavirus in paraffin embedded head and neck cancer specimen. *Int J Cancer* **121**(11): 2465–2472.

Smith DI, Zhu Y, McAvoy S, Kuhn R (2006) Common fragile sites, extremely large genes, neural development and cancer. *Cancer Lett* **232**(1): 48–57.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12**(10): 1611–1618.

Tatake RJ, Rajaram N, Damle RN, Balsara B, Bhisey AN, Gangal SG (1990) Establishment and characterization of four new squamous cell carcinoma cell lines derived from oral tumors. *J Cancer Res Clin Oncol* **116**(2): 179–186.

Thorland EC, Myers SL, Gostout BS, Smith DI (2003) Common fragile sites are preferential targets for HPV16 integrations in cervical tumors. *Oncogene* **22**(8): 1225–1237.

Trottier H, Mahmud S, Costa MC, Sobrinho JP, Duarte-Franco E, Rohan TE, Ferenczy A, Villa LL, Franco EL (2006) Human papillomavirus infections with multiple types and risk of cervical neoplasia. *Cancer Epidemiol Biomarkers Prev* **15**(7): 1274–1280.

Tsimplaki E, Argyri E, Xesfyngi D, Daskalopoulou D, Stravopodis DJ, Panotopoulou E (2014) Prevalence and expression of human papillomavirus in 53 patients with oral tongue squamous cell carcinoma. *Anticancer Res* **34**(2): 1021–1025.

Vaccarella S, Franceschi S, Snijders PJ, Herrero R, Meijer CJ, Plummer M. IARC HPV Prevalence Surveys Study Group (2010) Concurrent infection with multiple human papillomavirus types: pooled analysis of the IARC HPV Prevalence Surveys. *Cancer Epidemiol Biomarkers Prev* **19**(2): 503–510.

Van Doorslaer K, Tan Q, Xirasagar S, Bandaru S, Gopalan V, Mohamoud Y, Huyen Y, McBride AA (2013) The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Res* **41**(Database issue): D571–D578.

Wang Q, Jia P, Zhao Z (2013) VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. *PLoS One* **8**(5): e64465.

Wentzensen N, Vinokurova S, von Knebel Doeberitz M (2004) Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. *Cancer Res* **64**(11): 3878–3884.

Xu B, Chotewutmontri S, Wolf S, Klos U, Schmitz M, Durst M, Schwarz E (2013) Multiplex identification of human papillomavirus 16 DNA integration sites in cervical carcinomas. *PLoS One* **8**(6): e66693.

Yi X, Zou J, Xu J, Liu T, Liu T, Hua S, Xi F, Nie X, Ye L, Luo Y, Xu L, Du H, Wu R, Yang L, Liu R, Yang B, Wang J, Belinson JL (2014) Development and validation of a new HPV genotyping assay based on next-generation sequencing. *Am J Clin Pathol* **141**(6): 796–804.

Yim EK, Park JS (2005) The role of HPV E6 and E7 oncoproteins in HPV-associated cervical carcinogenesis. *Cancer Res Treat* **37**(6): 319–324.

Supplementary Information accompanies this paper on British Journal of Cancer website (http://www.nature.com/bjc)