Identification of Molecular Biomarkers Associated with Non-Small-Cell Lung Carcinoma (NSCLC) Using Whole-Exome Sequencing Analysis

Varsha Singh
All India Institute of Medical Sciences
Amit Katiyar
All India Institute of Medical Sciences
Prabhat Malik
All India Institute of Medical Sciences
Sunil Kumar
All India Institute of Medical Sciences
Anant Mohan
All India Institute of Medical Sciences
Harpreet Singh
ICMR-AIIMS Computational Genomics Center, Indian Council of Medical Research
Deepali Jain (deepalijain76@gmail.com)
All India Institute of Medical Sciences

Research Article

Keywords: Non-small cell lung cancer, adenocarcinoma, squamous cell carcinoma, biomarker, whole-exome sequencing

Posted Date: December 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1100571/v1

Abstract

Significant advancement has been made in the treatment of patients with on the basis of the molecular profile. However, no such molecular target exists for squamous cell carcinoma (SQCC). Whole-exome sequencing (WES) has been in wide use for the discovery of new genetic pulmonary adenocarcinoma (ADCA) markers, which may offer more information for the development of personalized medicine for all subtypes of lung cancer. The aim of the current study is to find out novel genetic markers for non-small-cell lung carcinoma (NSCLC). WES of 19 advanced NSCLC patients (10 ADCA and 9 SQCC) was done on the Illumina HiSeq 2000 (Illumina Inc., USA). Variant calling was performed using GATK HaplotypeCaller and subsequent the impacts of variants on protein structure or function were predicted using SnpEff and ANNOVAR. Clinical impact of variants was evaluated using cancer-related archives. Somatic variants were further prioritized using knowledge-driven variant interpretation approach. Functionally important variants were validated by Sanger sequencing. We identified 24 rare single-nucleotide variants (SNVs) including 17 non-synonymous SNVs, and 7 INDELs in 18 genes possibly linked to lung carcinoma. Sanger sequencing of 10 high confidence somatic SNVs showed 100% concordance in 7 genes, whereas 80% in the remaining 3 genes. Our bioinformatics analysis identified *KCNJ18, GPRIN2, TEKT4, HRNR, FOLR3, ESSRA, CTBP2, MPRIP, TBP*, and *FBXO6* may contribute to progression in NSCLC and could be used as new biomarkers for the treatment. Although the mechanism of *GPRIN2, KCNJ12* and *TEKT4* in tumorigenesis is unclear; our results suggest that these may play a major role in NSCLC and it is worth investigating in future.

Introduction

Various studies have proven NSCLC to be a histologically and molecularly heterogeneous group of cancer. The two main histological NSCLC subtypes are adenocarcinoma (ADCA) and squamous cell carcinoma (SQCC). Although the incidence of ADCA is on the rise, SQCC is currently the second most frequent histologic subtype. Distinct subtypes of NSCLC are driven by a specific genetic alteration, the molecular mechanisms of which remain to be fully elucidated. The Cancer Genome Atlas (TCGA) has conducted comprehensive genome studies of NSCLC, displaying a great diversity of molecular variations. Some of the mutated genes were common in both the histology subtypes and some were group specific. ADCA shows more complex and heterogeneous molecular patterns than SQCC, with a greater number of associated genomic aberrations (1, 2). Tumor genotype analysis has identified driver alterations in 50–80% of NSCLC patients according to demographics, and particularly ethnicity. Asian people have unique clinical characteristics, tumor histology and show different prevalence of oncogenic mutations (3).

Significant advancement has been made in the treatment of patients with pulmonary ADCA on the basis of the molecular profile. The discovery of EGFR mutations and ALK rearrangement has opened a new era of targeted therapy in ADCA. However, no such molecular target exists for squamous cell carcinoma (SQCC). Whole exome sequencing (WES) has been in wide use for the discovery of new genetic markers which may offer more information for the development of personalized medicine for all subtypes of lung cancer (4). WES has been widely used in clinical research for the discovery of new genetic markers. The key objective of this study is to find out novel genetic markers for NSCLC which can be used as a universal biomarker for the treatment. This study also identifies and compares the genomic alterations of ADCA subtype with SQCC subtype.

Methods

Sample collection and diagnosis

A total of 19 NSCLC *cases* (*EGFR*, *ALK* and *ROS1* negative) with available clinical follow-up were retrieved from the Department of Pathology, A.I.I.M.S., New Delhi. The haematoxylin and eosin stained slides were analysed and histological type of the tumour was determined according to World Health Organization 2021 classification of thoracic tumours**.** Blocks showing more than 80% tumour component in their respective sections were used for WES. Treatment and follow up details were retrieved from case record files from the Department of Medical Oncology, AIIMS, New Delhi.

Formalin fixed paraffin embedded (FFPE)DNA isolation and repair

DNA extraction was performed using FFPE DNA tissue extraction kit (A2352, Promega, USA) according to the manufacturer's instructions.FFPE DNA was repaired and purified using Gene JET FFPE DNA Purification Kit (K0881, Thermo Scientific, USA) according to the manufacturer's instructions.Quantity and purity of gDNA were assessed by Qubit® 2.0 Fluorometer (Invitrogen, Carlsbad, CA, USA) and NanoDrop ND-1000 (Thermo Scientific, USA).

Data generation

Sample sequencing

Sequencing libraries were prepared using the SureSelect All Human Exon V5 Kit (USA) according to the manufacturer's instructions.The final enriched pooled were sequenced on Illumina HiSeq 2000 platform (Illumina Inc., USA) generating 2×150-bp paired-end reads. Image analysis and base calling were carried out by Illumina software (CASAVA) with default parameters. *Demultiplexing* and FASTQ *file generation from Illumina basecall (BCL) files was performed using Bcl2fastq conversion software.*

Variant calling and quality control

Quality of raw reads (FASTQ files) were examined using FastQC (6). Adaptor and low-quality sequences were trimmed using Trimmomatic software(5). Paired clean reads with longer than 50 bases were aligned against the Genome Reference Consortium Human Build 38 patch release 7 (GRCh38.p7) using BWA-*MEM algorithm* of Burrows-Wheeler Aligner (BWA *0.6.1)(7)*. SAM/BAM post-processing steps including *SAM*to *BAM conversion, sorting,* adding read group information, *mark duplicates*, and base quality score recalibration were performed using the Genome Analysis Toolkit (GATK 4.0.6.0)(8,9). The quality of the recalibrated BAM files was checked with QualiMap v2.0.2(10). Finally, a genomic variation, including single-nucleotide polymorphisms (SNPs) and small

INDELs (insertion and deletion) were detected for each sample individually using GATK HaplotypeCaller in GVCF mode (-ERC GVCF), and the results were combined using GenotypeGVCFs. Raw variant calls were soft filtered using GATK VariantFiltration based on the following parameters: LowCoverage (DP < 5), LowQual (Q < 50), StrandBias (FS *P*-value > 60), SNV cluster (three or more SNVs within 10 bp), Poor Mapping Quality (>10% of reads have nonunique alignments).

Variant annotations

The impacts of variants on protein structure or function were predicted using SnpEff (11) and ANNOVAR (12). It compiles prediction scores from multiple algorithms including*PhyloP, SIFT, LRT,* SiPhy*, Polyphen-2,* GERP++, *MutationAssessor, Fathmm, MutationTaster, CADD, and MetaSVM. In addition to these tools, variants were reannotated using the germline/population databases (dbSNP, 1000 Genomes and ExAC) (13,14), and cancer/somatic databases (COSMIC, TCGA and ICGC) (15,16). The clinical significance of each variant was determined using*ClinVar (17), and My Cancer Genome (http://www.mycancergenome.org), whereas the *drug databases (PharmGKB, OncoKB) (18,19) was utilized to gain the information about the treatment implications of specific cancer gene alterations, and how these mutations affect response to treatment.*

Additional filters to reduce false positives somatic variants

A high-confidence somatic variants for tumor samples without a matched normal control were selected based on the following criteria: 1) mutations were considered true positives if they have a) QUAL ≥ 20, b) genotype quality (GQ) ≥ 20, c) mapping quality (MQ) ≥ 20, d) coverage depth at candidate site (DP) ≥ 20, e) QualByDepth (QD)≥ 2.0, and g) frequency ≥ 25% in tumor samples (20), 2) all common variants with minor allele frequency (MAF) of >1% in the germline/population databases (ExAC, and 1000 Genomes) were filtered out since those variants are deemed polymorphic/benign rather than pathogenic somatic driver mutations, 3) known germline variants reported at dbSNP (version 151) were excluded, and alterations listed as known somatic variations in COSMIC, The International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) were retain (21), 4) MAF threshold of 0.0001 was used in the gnomAD or TopMed database to filter variants for the somatic mutation (22), 5) variants were considered if the variant allele frequency (VAF; also known as variant allele fraction) is deviate from germline polymorphisms (~0.5/1 for heterozygous/homozygous) (23,24), 6) variants were considered if they are truncating variants (nonsense mutations, frameshift deletions/insertions, mutations located at exon-flanking regions, and highly conserved intronic splice sites), or apparent missense mutations predicted to be pathogenic by in-silico prediction tools, 7) synonymous variants that were not previously reported as pathogenic and not predicted to alter splicing were filtered out.

Enrichment analysis and candidate gene prioritization

Known or predicted variants to be involved in the lung or in related cancers were predicted using DisGenNET (25). The Human Gene Damage Index server (http://lab.rockefeller.edu/casanova/GDI) was used to predict LoF-intolerant genes. The gene product physically interact with a protein encoded by a known disease gene was explored using NetworkAnalyst (26). Gene product in a pathway associated with the disease and gene expressed in the tissue or organ of interest was retrieved from the literature.

Sanger sequencing

PCR was carried out on ABI Palm thermal cycler (Applied Biosystem, California), using both forward and reverse primers for greater accuracy and the results were analyzed using SeqMan II software (DNASTAR 5.07). The mutations with both base count more than 10% and QV (Quality value) more than 20 were considered to be trusted mutations.

Ethical clearance

The study on 19 NSCLC patients retrieved from the Department of Pathology, A.I.I.M.S., New Delhi was conducted in accordance with the ethical guidelines and regulations of the AIIMS and after obtaining approval from the AIIMS ethics committee. The ethical approval number is IECPG No. 480/29.08.2016. Study participants were enrolled following their voluntary written informed consent.

Results

Patient characteristics and sequencing statistics

A case study of 19 patients was included in this study, where 10 were ADCA (*EGFR*, *ALK* and *ROS1* wild type), and 9 were SQCC histological subtypes of NSCLC. Patients were early-onset with the average diagnostic age of 56 years, where male: female ratio was 5.6: 1. Among, three patients were non-smoker, whereas one case was with unknown smoking history. Nearly 75% of patients were present with co-morbidities, where all patients were either in stage III or IV. A total of 13.48 GB raw data and 11.98 GB processed data were generated per exome for the tumor sample of ADCA, whereas on an average 13.76 GB raw data and 12.24 GB processed data were obtained for the tumor sample of SQCC using WES. A higher percentage of reads were aligned to the human reference genome (GRCh38) in the tumors of ADCA patients (98.87%; range 96.48-99.95%) compared to tumors of SQCC patients (97.40%; range 85.03-99.51%), indicating that the generated dataset was highly relevant with the reference genome. The average GC content (49.74%) in the tumors of ADCA patients ranged from 42.10 to 64.63%, whereas average GC content (48.50%) in the tumors of SQCC patients ranged from 41.81 to 52.28%. Clinical characteristics and sequencing summary of lung cancer patient's participants in this study are listed in **Table S1 in Supplementary File 1**.

Detection and characterization of SNVs

After initial variant filtering (as described in methods), a total of 1,157,921 (single-allelic 1,157,792 and multi-allelic 129) variants were retained in the tumors of ADCA patients (n=10) which was slightly higher than total variants (1,076,209; single-allelic 1,076,069 and multi-allelic 140) detected in SQCC patients (n=9). The number of variants per chromosomes ranged from 6,227 (chrY) to 98,788 (chr4) in the tumors of ADCA subtype, whereas ranged from 8,385 (chrY) to 104,645 (chr4) in the tumors of SQCC subtype. The variants rate per chromosomes varied from 1,925 (chr4) to 9,190 (chrY) and revealed on average 1 variants after every 2,667 bases in ADCA subtype, whereas it was after every 2,869 base in SQCC subtype. We observed higher known variants i.e. 616,999 (53.285%) in ADCA subtype as compared to 537,980 (49.988%) in SQCC subtype. The distribution of variants by their type disclosed 1,066,489 SNPs, 38,751 insertions and 52,681 deletions in ADCA subtype. However, the different distributions of insertions/deletions (35,799/49,010) and SNP (991,400) was observed in SQCC subtype. The identified SNPs from NSCLC were categorized into two clusters based on nucleotide substitutions i.e. transitions (A/G and C/T) and transversions (A/C, A/T, C/G, and G/T). The transition-to-transversion (Ts/Tv) ratio was slightly higher (1.698 for all SNPs and 2.214 for known SNPs) in ADCA subtype compared to Ts/Tv ratio of 1.4859 (all SNPs) and 2.137 (known SNPs) in SQCC subtype. Among detected SNPs, 23.05% were heterozygous, and 76.93% were homozygous in ADCA subtype, whereas it was 21.35% and 78.64% in SQCC subtype, respectively. The ratio of heterozygous SNVs to homozygous SNVs (Het/Hom ratio) was 0.29 and 0.27 in ADCA and ADCA, respectively where the lower value was associated with true positive variants. The ratio of nonsense to missense mutations (0.007), and missense to silent mutations (0.829) in ADCA subtype was nearly similar to 0.007 and 0.863, respectively in SQCC subtype. The ratio of nonsense to missense and missense to silent mutations in the human genome may reflect a role for natural selection, especially purifying selection. The 'GAT' codons have been replaced maximum times by 'GAC' codons in both ADCA (588) and SQCC (665) subtype. The characterization of SNVs in ADCA and SQCC subtype, revealed that ADCA was more genetically unstable compared to SQCC. Variant's summary identified by whole exome sequencing are listed in Figure 1 and Table S2 in Supplementary File 1.

Detection of somatic variants in tumor only samples

To detect somatic SNV, the present study focused on missense variants in the exonic regions or splice sites. The downstream filtering by genomic location revealed a total of 6,712 and 8,000 exonic variants in ADCA and SQCC subtype, respectively. Among exonic SNVs (ADCA subtype), 2,113 were synonymous, 1,985 were nonsynonymous, 28 were frame-shift indels, 51 were nonframe-shift indels, 15 were stop-gain, 2 were stop-loss, and 2,518 were non-coding SNVs. In SQCC-subtype, 3,249 were synonymous, 2,656 were nonsynonymous, 52 were frame-shift indels, 59 were nonframe-shift indels, 36 splicing-variant, 1 was gene_fusion, and 1,947 were non-coding SNVs. Exonic missense, nonsense, stop-loss, frameshift and splice site variants all have potential to affect protein function. Therefore, we excluded the synonymous variants that have no functional impact and retained 4,599 and 4,751 variants from ADCA and SQCC subtype, respectively. As rarity(27) is the key criterion to have a functional effect on the encoded protein, the filtered variants were used to eliminate the common germline mutations (minor allele frequency below 5% in population/germline databases) and as a outcome 1,642 and 2,141 variants were retained in ADCA and SQCC subtype, respectively. After excluding false positive mutations (based on additional filter *criteria's* $1_{a\text{-}g}$given in methodology), 500 and 734 variants in ADCA and SQCC subtype, respectively was observed. To exclude deemed polymorphic/benign variants, high quality rare variants (MAF $\leq$ 1% and QUAL $\geq$ 500) were excavated which revealed 94 variants in ADCA subtype, whereas 87 variants in SQCC subtype. To identify candidates likely to have deleterious effects, combination of multiple variant annotation tools were applied that revealed *the impact of amino acid changes on protein function* based on the combine scores. *The variants were classified as damaging (predict pathogenic by maximum number of tools), probably damaging (predict pathogenic or benign by an equal number of tools), benign (predict benign by maximum number of tools) and uncertain significance (unknown) as per the variant assessment guidelines by the American College of Medical Genetics. The alterations listed in COSMIC, ICGC and TCGA were considered as known somatic variations in this study. The final outcome, revealed a total of 24 high confidence somatic variants (ADCA=14, SQCC=10)* associated with 18 genes and were *classified as* known somatic variant (n=10), deleterious variant (n=8), and variant of *uncertain significance (VUS) (n=6)* ***(Table 1, Figure 2).*** The *gene (n=11) namely CTBP2, ESRRA, FDFT1, FOLR3, GPRIN2, HRNR, KCNJ18, KRT18, LILRA2, MTRNR2L8, and TEKT4 were observed to be mutated in both ADCA and SQCC* histologic *subtype* of lung cancer. In addition, mutated gene (n=4) namely *LRP2, MPRIP, NYX,* and *TBP* were observed in histologic subtype of ADCA only, whereas *FBXO6, MIR3689F,* and *UMPS* mutated genes were found in *SQCC subtype only* ***(Figure 3).*** *Out of 24 high confidence somatic variants,* 19 (79.17%) variants had a previously known dbSNP ID while the remaining 5 (20.83%) were unassigned, new variants. The 17 variants had *higher proportion of driver mutations (≥25% allele frequency) in tumor samples, whereas 21 variants were observed to be true somatic mutation based on* variant allele frequency which is used to infer whether a variant comes from somatic cells or inherited from parents when a matched normal sample is not provided.

Knowledge-driven variant prioritization

Though, we followed the guidelines suggested for experimental design and variant filtering, yet we obtained more candidates with likely functional effects than can be verified experimentally. High priority candidates based on the biological hypothesis defined for this study were selected. The analysis of known or predicted variants to be involved in the lung or in a related cancers revealed eight genes (*CTBP2, ESRRA, FBXO6, FDFT1, KRT18, MPRIP, TBP,* and *UMPS*) associated with the lung carcinoma. In addition, four genes i.e. *LRP2* (bone, breast, colorectal, pancreatic, prostate, and renal cell carcinoma), *HRNR* (breast, liver and pancreatic carcinoma), *FOLR3* (breast and ovarian carcinoma) and *TEKT4* (breast and thyroid carcinoma) were involved in other's cancer types. The genes namely *MIR3689F, MTRNR2L8, NYX, GPRIN2* and *KCNJ18* were observed to be not associated with any type of cancer and considered as low priority genes for the validation. Loss-of-function (LoF) variants in three genes i.e. *FOLR3, UMPS* and *LILRA2* were observed which damage or eliminate the function of the encoded protein. The LoF-intolerant genes (*CTBP2, GPRIN2, HRNR* and *TEKT4*) were classified as extremely loss of function intolerant (pLI $\geq$ 0.9), whereas gene *(MTRNR2L8*) with low pLI scores ($\leq$ 0.1) was considered as LoF-tolerant (common loss-of-function variants) and was not selected for the validation. We also prioritized genes based on the interactome of known disease-associated proteins. The genes, *FBXO6* (degree 153; betweenness centrality 77253.04), *TBP* (degree 152; betweenness centrality 78171.64), *KRT18* (degree 89; betweenness centrality 44603.52), *CTBP2* (degree 64; betweenness centrality 33732.54), *ESRRA* (degree 44; betweenness centrality 21895.15), *LRP2* (degree 38; betweenness centrality 20031.73), *MPRIP* (degree 24; betweenness centrality 10184.34), *FDFT1* (degree 17; betweenness centrality 7452.33), *UMPS* (degree 15; betweenness centrality 7266.93), and *HRNR* (degree 14; betweenness centrality 6258.26) were observed to be the most highly ranked hub genes in this study. Moreover, relevant information for the genes of interest was retrieved from the literature. Nine gene/proteins including *HRNR, KCNJ18, ESRRA, MPRIP, FBXO6, FOLR3, FDFT1, KRT18,* and *LILRA* were observed

to be overexpressed in lung cancer which might have a potential role in cancer development, proliferation, and metastasis. Remarkably, most of the genes were involved in important cancer-related pathways. Close observation showed that the variant in gene *MIR3689F* (miRNA) and *MTRNR2L8* (It is unclear if this is a transcribed protein-coding gene, or if it is a nuclear pseudogene of the mitochondrial *MT-RNR2* gene) was incorrect for this study and hence not selected for the validation. The characteristics of candidate variants from public resources and published literature are given in **Table 2.**

Somatic variants validation by Sanger sequencing

To eliminate false-positive rates of the identified high confidence somatic mutations from WES data, we selected 10 genes for Sanger sequencing validations. Interestingly we observed that 7 genes were mutated in more than 60% samples and 3 genes were mutated in either one or two samples. Further, Sanger sequencing results showed 100% concordance in 7 genes and the remaining 3 genes concordances were found only in 80% cases. The mutations observed along with WES and Sanger sequencing data have been depicted in the **Table 3 and Figure S1 in Supplementary File 1**.

The gene-wise results of Sanger sequencing are given below:

TEKT1, GPRIN2 and KCNJ18 point mutation: These three genes were found mutated in all the samples by WES. On further validation by Sanger sequencing, we also found *TEKT1* (exon6:c.G1213A:p.A405T) were positive in 18 cases, *GPRIN2* (exon3:c.G721A:p.V241M) in 16 cases and *KCNJ18* (c.C631T:p.L211F) point mutations in all samples.

Hornerin (HRNR) and FOLR3 mutation: These two genes were found mutated in 18 samples by WES. Validation by Sanger sequencing revealed 100% concordance. The Hornerin gene was present with the point mutation in exon 3 (c.C5050G:p.R1684G) and *FOLR3* gene with deletion in exon 3 (c.46_47del:p.Y16fs).

ESSRA and CTBP2 mutation: WES revealed *ESSRA* gene was mutated in 16 cases and *CTBP2* in 17 cases. Sanger sequencing revealed *ESRRA* gene exon 7 point mutation (c.G1127T:p.R376L) and deletion (c.1130_1132del:p.377_378del) in 12 and 5 cases whereas *CTBP2* (exon5:c.G2292T:p.Q764H) point mutations in 15 cases.

MPRIP, TBP and FBXO6 mutation: Exon 6 deletion of *MPRIP* gene (c.537_539del:p.179_180del) was found in 2 samples whereas *TBP* gene (exon3:c.222_223insCAG:p.Q74delinsQQ) deletion and *FBXO6* gene (exon2:c.A151G:p.M51V) point mutation were found in one case each by WES and Sanger sequencing.

Discussion

In this study, whole-exome sequencing was used to predict genomic alterations in ADCA and SQCC histological subtypes of NSCLC. Overall, we detected 24 high confidence somatic variants (ADCA=14, SQCC=10) in 18 genes. Many of the gene alterations were common in both subtypes whereas few were group specific, these findings will throw more light on personalized medicine. Of interest, 16 genes (≥50% mutation frequency) were observed to be mutated in lung cancer, where gene *GPRIN2, KCNJ18* and *TEKT4* was found mutated in all the patients (100% mutation frequency). Pathway enrichment analysis confirmed that the majority is involved in processes relevant for tumorigenesis such as cell differentiation and proliferation. In the end, 10 novel somatic variants (affecting 10 genes i.e. *CTBP2, ESRRA, FBXO6, FOLR3, GPRIN2, HRNR, KCNJ18, MPRIP, TBP*, and *TEKT4*) that were identified for the first time were validated by Sanger sequencing. Our data expand the mutation spectrum for NSCLC and will be a useful resource for the NSCLC research community. Each biomarker has been discussed in details in the following paragraphs.

Mutated genes present in all samples of ADCA and SQCC subtype

Of interest, the three genes, *KCNJ18, TEKT4*, and *GRIPN2* are mutated in all NSCLC samples and can serve as common diagnostic markers for both subtypes.

Potassium inwardly rectifying channel subfamily J member 18 (KCNJ18) gene encodes a member of the inwardly rectifying potassium channel family and plays a role in resting membrane potential maintenance.(28) The potassium channel involvement in tumour cell proliferation has been studied previously in colorectal carcinoma cell line DLD-1 and human prostate cancer cell line LNCaP by modulating calcium influx. (29, 30) The E139K (rs76265595), G145S (rs75029097) and A185V (rs73979896) mutations in *KCNJ12/KCNJ18* gene were identified in esophageal SQCC (31). Mutations were found in *KCNJ18* gene in all the NSCLC patients studied but the amino acid variations (c.C631T, p.L211F) were different from those reported earlier. So, it can be postulated that KCNJ18 might be involved in p53 pathway, and it may be investigated in larger cohort of patients.

G protein-regulated inducer of neurite outgrowth 2 (GPRIN2) gene is located *on* chromosome 16 and encodes glutamate NMDA receptor(32). Variations in this gene have been found in malignant as well as non-malignant diseases(31, 33). Rare damaging novel mutations in *GPRIN2* genes has been found in 33% melanoma patients (somatic)(34), familial human esophagheal SQCC (germline/somatic)(31) as well as 501Mel melanoma cell line.(34) Mutated *GPRIN2* might play a major role in tumorigenesis via glutamate pathway where excess release of glutamate showed more aggressive growth(35, 36). In the present study we found p.V241M (c.G721A) mutation in all the NSCLC cases, although mutation observed was different from those reported in the literature (p.A233S, rs11204659). The role of this mutation in tumorigenesis is unclear; however, high frequency observed in our study hints that it may be explored in other studies.

Tektin 4 gene (TEKT4) is present on chromosome 2, encodes tektin4, a constitutive protein of microtubules in cilia, flagella, basal bodies, and centrioles(37). The biological function of *TEKT4* has not been well explained in cancer initiation and development. Variations in the *TEKT4* gene play an important role in papillary thyroid cancer progression. *TEKT4* knockdown in papillary thyroid cancer cell lines inhibits tumorigenesis by impairing cell proliferation, colony formation, migration, and invasion via blocking the activity of PI3K/AKT pathway(38). We found *TEKT4* gene mutations in all 20 cases studied. However,

mutations (c.G1213A, A405T) were different from those reported in papillary thyroid cancer (c.1276_1279delinsACCC). Mutated *TEKT4* is associated with increased paclitaxel resistance and poor prognosis in breast cancer patients(39). This mutation might play a vital role in the pathogenesis of lung cancer however, the role of *TEKT4* gene in PI3K/AKT pathway signalling and treatment resistance require further investigations.

Mutated genes present in 80% samples of ADCA and SQCC subtype

In addition to the three aforementioned genes, *HRNR, FOLR3, CTBP2* and *ESSRA* are significantly mutated in more than 80% of the NSCLC samples. All these mutated genes are directly or indirectly play a role in tumorigenesis and can additionally serve as common pathogenetic link for subtypes of NSCLC.

C-terminal-binding protein 2 (CTBP2) is a member of the *CTBP* family protein located in the human chromosome 10. *CTBP2* is evolutionarily conserved transcriptional co-regulator that interact with DNA binding transcription factors and chromatin remodelers. *CTBP2* represses a number of tumour suppressor genes *(E-cadherin, PTEN, INK4*), induces the epithelial-to-mesenchymal transition and functions as an apoptosis antagonist. Aberrant expression of *CTBP2* has been found to be associated with tumorigenesis, cancer progression, and poor prognosis(40, 41). Accumulating evidences indicated that *CTBP2* expression is elevated in several types of malignancies which include gastric cancer, melanoma, breast cancer, esophageal SQCC, prostate cancer, hepatocellular carcinoma, and ovarian cancer. High expression of *CTBP2* results in progression of esophageal SQCC through negatively regulating *p16 (INK4A). CTBP2* is considered as a co-factor of TGF-β-signalling pathway in promoting cancer metastasis and also participates in the regulation of WNT signalling. *CTBP2* modulated the androgen receptor to promote prostate cancer cell proliferation through c-MYC signalling and also promoted its progression. *CTBP2* can be considered as driver oncogene in solid tumours and also as an emerging target in cancer as it encodes a druggable dehydrogenase domain for which first and second generation inhibitors have already been identified(42). *CTBP2* plays a crucial role in NSCLC progression, and its depletion can provide a new target for NSCLC treatment(43). *CTBP2* was mutated (c.G2292T, Q764H) in 17 cases in the present analysis. We believe that *CTBP2* has the potential to become a high-efficacy target however, it warrants further research.

Estrogen related receptor alpha (ESRRA) is evolutionarily related to estrogen receptor and can efficiently bind to estrogen receptor that are commonly shared by many target genes. Over-expression of *ESRRA* has been found in carcinoma of thyroid, ovary, breast, prostate, colon and endometrium (44, 45). It is correlated with poor prognosis. *ESRRA* suggested being a molecular target for treatment of endometrial cancer. Other investigators reported *ESRRA* as one of the negative prognostic factors in human prostate cancer. *ESRRA* is also over-expressed in lung cancer patients and cell line A549 while some studies report low or undetectable (46), estrogen receptor expression in NSCLC cells. *ESRRA* is up-regulated in NSCLC tissues and promotes the progression, proliferation and invasion via NF-κB mediated up-regulation of IL-6 (47). *ESRRA* knockdown xenografts sensitized cells to paclitaxel and reduce tumour growth and angiogenesis. Overall review of literature and our preliminary experience with *ESRRA* suggest that it can be studied in detail in NSCLC patients.

The hornerin gene *(HRNR)* is clustered on the chromosome region 1q21 and it is a member of the S100 protein family. The function of *HRNR* is poorly clarified in the development of human tumours. Altered expression of *HRNR* was reported to be involved in cancer development, malignant transformation and invasion. Elevated *HRNR* has been found in many tumours viz lung SQCC, hepatocellular carcinoma, colorectal cancer, prostate cancer, glioblastoma and cell lines, breast carcinoma and cell linesand acute myeloid leukemia (48). *HRNR* has been found to contribute to hepatocellular carcinoma progression via the regulation of the AKT pathway (49). In lung SQCC and colorectal carcinoma, altered *HRNR* expression has been associated with disease recurrence(50). In the current study we have also found recurrence occurred in 9 patients all of whom were mutated with the *HRNR* gene.

The gene for folate receptor gamma *(*FOLR3) is located on chromosome 11 and consists of five exons. The *FOLR3* receptor is a constitutively secreted form of the folate receptor. *FOLR3* is one of the key genes involved in the pemetrexed pathway. Variation in *FOLR3* gene affects pemetrexed uptake, metabolism, treatment tolerability, response and survival(51). In NSCLC and mesothelioma patients, variation in the *FOLR3* gene has been reported. *FOLR3* germline mutation (rs61734430, c.292C>T variant) has been associated with an increased rate of disease progression(51). Pemetrexed is a folate antimetabolite(52) approved for the treatment of advanced NSCLC in the first line, second line setting as well as for maintenance therapy. We found c.46_47del, Y16fs mutation which is different from reported mutation type. Future studies are required to know the role of *FOLR3* as predictive marker for personalised pemetrexed therapy (which can improve both efficacy and tolerability).

Conclusions

In this study, novel somatic mutations and subtype-specific mutations were found using WES and subsequently confirmed by Sanger sequencing. *TBP* and *MPRIP* mutated genes were solely associated with ADCA subtype whereas *FBOX6* with SQCC. In addition, *GPRIN2, KCNJ18* and *TEKT4* genes were found mutated in all the patients (70). Although the mechanism of *GPRIN2, KCNJ12* and *TEKT4* in tumorigenesis is unclear; our results suggest that these may play a major role in NSCLC and it is worth to be investigated in future. The identified target genes from our study can be used as biomarkers for the detection and diagnosis of NSCLC. This study demonstrates that WES can be applied to FFPE clinical samples for finding or validation of biomarkers in cancer research.

Abbreviations

NSCLC: non-small-cell lung carcinoma; **ADCA:** pulmonary adenocarcinoma; **SNVs:** single-nucleotide variants; **FFPE:** formalin fixed paraffin embedded; **BWA:** burrows-wheeler aligner; **GRCh:** genome reference consortium human; **COSMIC:** the catalogue of somatic mutations in cancer; **ICGC:** *the international cancer genome consortium;* ***TCGA:*** *the cancer genome atlas;* ***VAF:*** *variant allele frequency;* **PCR:** polymerase chain reaction.

Declarations

Data availability: Raw data files have been submitted to the Sequence Read Archive (SRA) of National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/sra) under BioProject accession number **PRJNA734015.**

Acknowledgments: We gratefully acknowledge the financial support from the Lady Tata Memorial Trust (N-1611). We thank the ICMR-AIIMS Genomics Center and CCRF: Bioinformatics Facility at All India Institute of Medical Sciences (A.I.I.M.S.), New Delhi for providing their facility for the data analysis.

Author contributions: DJ conceived, designed and supervised the study. AM, PM, SK managed the sample collection. VS and AK did the data analysis. VS and AK interpreted the results and drafted the manuscript. DJ, HS reviewed and modified the draft. All authors have read and approved the manuscript for publication.

Competing interests: All the authors declare that there is no conflict of interest related to this study.

Author information

Affiliations

Department of Pathology, All India Institute of Medical Sciences, Ansari Nagar, New Delhi-110029, India

Varsha Singh & Deepali Jain

Bioinformatics Facility, Centralized Core Research Facility, All India Institute of Medical Sciences, Ansari Nagar, New Delhi-110029, India

Amit Katiyar

Department of Medical Oncology, All India Institute of Medical Sciences, Ansari Nagar, New Delhi-110029, India

Prabhat Malik

Department of Surgical Oncology, All India Institute of Medical Sciences, Ansari Nagar, New Delhi-110029, India

Sunil Kumar

Department of Pulmonary Critical Care & Sleep Medicine, All India Institute of Medical Sciences, New Delhi-110029, Ansari Nagar, India

Anant Mohan

ICMR-AIIMS Computational Genomics Center, Division of Biomedical Informatics, Indian Council of Medical Research, Ansari Nagar, New Delhi-110029, India

Harpreet Singh

References

1. Hammerman P, Voet D, Lawrence M, Voet D, Jing R, Cibulskis K, et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489(7417):519–25. doi: 10.1038/nature11404 (2016).
2. Hwang DH, Sholl LM, Rojas-Rudilla V, Hall DL, Shivdasani P, Garcia EP, et al. KRAS and NKX2-1 Mutations in Invasive Mucinous Adenocarcinoma of the Lung. *J Thorac Oncol* 11(4):496–503. doi: 10.1016/j.jtho.2016.01.010 (2016).
3. Dearden S, Stevens J, Wu YL, Blowers D. Mutation incidence and coincidence in non small-cell lung cancer: meta-analyses by ethnicity and histology (mutMap). *Ann Oncol Off J Eur Soc Med Oncol* 24(9):2371-6. doi: 10.1093/annonc/mdt205 (2013).
4. Maggi E, Patterson NE, Montagna C. Technological advances in precision medicine and drug development. *Expert Rev Precis Med Drug Dev.* 1(3):331-343. doi: 10.1080/23808993.2016.1176527 (2016).
5. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114-20. doi: 10.1093/bioinformatics/btu170 (2014).
6. Andrews S. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. *Soil* 5(1) (2020).
7. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754-60. doi: 10.1093/bioinformatics/btp324 (2009).
8. 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* 20(9):1297-303. doi: 10.1101/gr.107524.110 (2010).
9. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43(1110):11.10.1-11.10.33. doi: 10.1002/0471250953.bi1110s43 (2013).
10. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 32(2):292-4. doi: 10.1093/bioinformatics/btv566 (2016).
11. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 6(2):80-92. doi: 10.4161/fly.19695 (2012).
12. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38(16):e164. doi: 10.1093/nar/gkq603 (2010).
13. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The Exome Aggregation Consortium, Daly MJ, MacArthur DG. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res* 45(D1):D840-D845. doi: 10.1093/nar/gkw971

(2017).

14. Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56-65. doi: 10.1038/nature11632 (2012).
15. 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* 10 (10.11). doi: 10.1002/0471142905.hg1011s57 (2008).
16. International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al. International network of cancer genome projects. International network of cancer genome projects. *Nature* 464(7291):993-8. doi: 10.1038/nature08987 (2010).
17. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 42(Database issue):D980-5. doi: 10.1093/nar/gkt1113 (2014).
18. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* 92(4):414-7. doi: 10.1038/clpt.2012.96 (2012).
19. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* PO.17.00011. doi: 10.1200/PO.17.00011 (2017).
20. Laginestra MA, Cascione L, Motta G, Fuligni F, Agostinelli C, Rossi M, et al. Whole exome sequencing reveals mutations in FAT1 tumor suppressor gene clinically impacting on peripheral T-cell lymphoma not otherwise specified. *Mod Pathol* 33:179–187. https://doi.org/10.1038/s41379-019-0279-8 (2020).
21. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* 47(D1):D941-D947. doi: 10.1093/nar/gky1015 (2019).
22. Koboldt DC. Best practices for variant calling in clinical sequencing. *Genome Med* 12(1):91. doi: 10.1186/s13073-020-00791-w (2020).
23. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. *Nature* 502(7471):333-339. doi: 10.1038/nature12634 (2013).
24. Heinrich V, Stange J, Dickhaus T, Imkeller P, Krüger U, Bauer S, et al. The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. *Nucleic Acids Res* 40(6):2426-31. doi: 10.1093/nar/gkr1073 (2012).
25. Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database (Oxford)* bav028. doi: 10.1093/database/bav028 (2015).
26. Zhou G, Soufan O, Ewald J, Hancock REW, Basu N, Xia J. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Res* 47(W1):W234-W241. doi: 10.1093/nar/gkz240 (2019).
27. Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, et al. Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* 173(2):305-320.e10. doi: 10.1016/j.cell.2018.03.033 (2018).
28. Hugnot JP, Pedeutour F, Le Calvez C, Grosgeorge J, Passage E, Fontes M, et al. The human inward rectifying K+ channel Kir 2.2 (KCNJ12) gene: gene structure, assignment to chromosome 17p11.1, and identification of a simple tandem repeat polymorphism. *Genomics* 39(1):113-6. doi: 10.1006/geno.1996.4450 (1997).
29. Skryma RN, Prevarskaya NB, Dufy-Barbe L, Odessa MF, Audin J, Dufy B. Potassium conductance in the androgen-sensitive prostate cancer cell line, LNCaP: involvement in cell proliferation. *Prostate* 33(2):112-22. doi: 10.1002/(sici)1097-0045(19971001)33:2<112::aid-pros5>3.0.co;2-m (1997).
30. Xiaoqiang Y, Kwan HY. Activity of voltage-gated K+ channels is associated with cell proliferaton and Ca2+ influx in carcinoma cells of colon cancer. *Life Sci* 65(1):55-62. doi: 10.1016/s0024-3205(99)00218-0 (1999).
31. Khalilipour N, Baranova A, Jebelli A, Heravi-Moussavi A, Bruskin S, Abbaszadegan MR. Familial Esophageal Squamous Cell Carcinoma with damaging rare/germline mutations in KCNJ12/KCNJ18 and GPRIN2 genes. *Cancer Genet* 221:46–52. doi: 10.1016/j.cancergen.2017.11.011 (2018).
32. Johnson JW, Ascher P. Glycine potentiates the NMDA response in cultured mouse brain neurons. *Nature* 325(6104):529-31. doi: 10.1038/325529a0 (1987).
33. Iida N, Kozasa T. Identification and biochemical analysis of GRIN1 and GRIN2. *Methods Enzymol* 390:475-83. doi: 10.1016/S0076-6879(04)90029-8 (2004).
34. Wei X, Walia V, Lin JC, Teer JK, Prickett TD, Gartner J, et al. Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat Genet* 43(5):442-6. doi: 10.1038/ng.810 (2011).
35. Takano T, Lin JH, Arcuino G, Gao Q, Yang J, Nedergaard M. Glutamate release promotes growth of malignant gliomas. *Nat Med* 7(9):1010-5. doi: 10.1038/nm0901-1010 (2001).
36. Dalva MB, Takasu MA, Lin MZ, Shamah SM, Hu L, Gale NW, et al. EphB receptors interact with NMDA receptors and regulate excitatory synapse formation. *Cell* 103(6):945-56. doi: 10.1016/s0092-8674(00)00197-5 (2000).
37. Amos LA. The tektin family of microtubule-stabilizing proteins. *Genome Biol* 9(7):229. doi: 10.1186/gb-2008-9-7-229 (2008).
38. Zheng Z, Zhou X, Cai Y, Chen E, Zhang X, Wang O, et al. TEKT4 Promotes Papillary Thyroid Cancer Cell Proliferation, Colony Formation, and Metastasis through Activating PI3K/Akt Pathway. *Endocr Pathol* 29(4):310–6. doi: 10.1007/s12022-018-9549-0 (2018).
39. Jiang YZ, Yu KD, Peng WT, Di GH, Wu J, Liu GY, et al. Enriched variations in TEKT4 and breast cancer resistance to paclitaxel. *Nat Commun* 5:3802. doi: 10.1038/ncomms4802 (2014).
40. Zhang C, Gao C, Xu Y, Zhang Z. CtBP2 could promote prostate cancer cell proliferation through c-Myc signaling. *Gene* 546(1):73-9. doi: 10.1016/j.gene.2014.05.032 (2014).

41. Dai F, Xuan Y, Jin JJ, Yu S, Long ZW, Cai H, et al. CtBP2 overexpression promotes tumor cell proliferation and invasion in gastric cancer and is associated with poor prognosis. *Oncotarget* 8(17):28736-28749. doi: 10.18632/oncotarget.15661 (2017).
42. Straza MW, Paliwal S, Kovi RC, Rajeshkumar B, Trenh P, Parker D, et al. Therapeutic targeting of C-terminal binding protein in human cancer. *Cell Cycle* 9(18):3740-50. doi: 10.4161/cc.9.18.12936 (2010).
43. Wang DP, Gu LL, Xue Q, Chen H, Mao GX. CtBP2 promotes proliferation and reduces drug sensitivity in non-small cell lung cancer via the Wnt/β-catenin pathway. *Neoplasma* 65(6):888–97. doi: 10.4149/neo_2018_171220N828 (2018).
44. Luo H, Rankin GO, Liu L, Daddysman MK, Jiang BH, Chen YC. Kaempferol inhibits angiogenesis and VEGF expression through both HIF dependent and independent pathways in human ovarian cancer cells. *Nutr Cancer* 61(4):554-63. doi: 10.1080/01635580802666281 (2009).
45. Sun P, Sehouli J, Denkert C, Mustea A, Könsgen D, Koch I, et al. Expression of estrogen receptor-related receptors, a subfamily of orphan nuclear receptors, as new tumor biomarkers in ovarian cancer cells. J Mol Med. 83(6):457–67 (2005).
46. Lai JC, Cheng YW, Chiou HL, Wu MF, Chen CY, Lee H. Gender difference in estrogen receptor alpha promoter hypermethylation and its prognostic value in non-small cell lung cancer. *Int J Cancer* 117(6):974-80. doi: 10.1002/ijc.21278 (2005).
47. Zhang J, Guan X, Liang N, Li S. Estrogen-related receptor alpha triggers the proliferation and migration of human non-small cell lung cancer via interleukin-6. *Cell Biochem Funct* 36(5):255-262. doi: 10.1002/cbf.3337 (2018).
48. Cho JH, Sun J, Lee S, Ahn JS, Park K, Park KU, et al. OA10.05 An Open-Label, Multicenter, Phase II Single Arm Trial of Osimertinib in NSCLC Patients with Uncommon EGFR Mutation(KCSG-LU15-09). *J Thorac Oncol* 13(10):S344. DOI:https://doi.org/10.1016/j.jtho.2018.08.291 (2018).
49. Fu SJ, Shen SL, Li SQ, Hua YP, Hu WJ, Guo BC, et al. Hornerin promotes tumor progression and is associated with poor prognosis in hepatocellular carcinoma. *BMC Cancer* 18(1). doi: 10.1186/s12885-018-4719-5 (2018).
50. Zhang H, Liu J, Yue D, Gao L, Wang D, Zhang H, et al. Clinical significance of E-cadherin, β-catenin, vimentin and S100A4 expression in completely resected squamous cell lung carcinoma. *J Clin Pathol* 66(11):937–45. doi: 10.1136/jclinpath-2013-201467 (2013).
51. Corrigan A, Walker JL, Wickramasinghe S, Hernandez MA, Newhouse SJ, Folarin AA, et al. Pharmacogenetics of pemetrexed combination therapy in lung cancer: Pathway analysis reveals novel toxicity associations. *Pharmacogenomics J* 14(5):411–7. doi: 10.1038/tpj.2014.13 (2014).
52. Scagliotti GV, Parikh P, Von Pawel J, Biesma B, Vansteenkiste J, Manegold C, et al. Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small-cell lung cancer. *J Clin Oncol* 26(21):3543–51. doi: 10.1200/JCO.2007.15.0375 (2008).
53. Vaishnavi A, Capelletti M, Le AT, Kako S, Butaney M, Ercan D, et al. Oncogenic and drug-sensitive NTRK1 rearrangements in lung cancer. *Nat Med* 19(11):1469–72. doi: 10.1038/nm.3352 (2013).
54. Peifer M, Fernández-Cuesta L, Sos ML, George J, Seidel D, Kasper LH, et al. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nat Genet* 44(10):1104–10. doi: 10.1038/ng.2396 (2012).
55. Drilon A, Li G, DOgan S, GOunder M, Shen R, Arcila M, et al. What Hides Behind the MASC: Clinical Response and Acquired Resistance to Entrectinib After ETV6-NTRK3 Identification in a Mammary Analogue Secretory Carcinoma (MASC). *Ann Oncol* 27(5):920-6. doi: 10.1093/annonc/mdw042 (2016).
56. Russo M, Misale S, Wei G, Siravegna G, Crisafulli G, Lazzari L, et al. Acquired resistance to the TRK inhibitor entrectinib in colorectal cancer. *Cancer Discov* 6(1):36-44. doi: 10.1158/2159-8290.CD-15-0940 (2016).
57. Zhong S, Fromm J, Johnson DL. TBP Is Differentially Regulated by c-Jun N-Terminal Kinase 1 (JNK1) and JNK2 through Elk-1, Controlling c-Jun Expression and Cell Proliferation. *Mol Cell Biol* 27(1):54-64. doi: 10.1128/MCB.01365-06 (2007).
58. Xu MJ, Johnson DE, Grandis JR. EGFR-targeted therapies in the post-genomic era. *Cancer Metastasis Rev* 36(3):463-473. doi: 10.1007/s10555-017-9687-8 (2017).
59. Goel HL, Mercurio AM. VEGF targets the tumour cell. *Nat Rev Cancer* 13(12):871-82. doi: 10.1038/nrc3627 (2013).
60. Randle SJ, Laman H. F-box protein interactions with the hallmark pathways in cancer. *Semin Cancer Biol* 36:3-17. doi: 10.1016/j.semcancer.2015.09.013 (2016).
61. Harms PW, Vats P, Verhaegen ME, Robinson DR, Wu YM, Dhanasekaran SM, et al. The Distinctive Mutational Spectra of Polyomavirus-Negative Merkel Cell Carcinoma. *Cancer Res* 75(18):3720-3727. doi: 10.1158/0008-5472 (2015).
62. Hong X, Huang H, Qiu X, Ding Z, Feng X, Zhu Y, et al. Targeting posttranslational modifications of RIOK1 inhibits the progression of colorectal and gastric cancers. *Elife* 7:e29511. doi: 10.7554/eLife.29511 (2018).
63. Xu HZ, Wang ZQ, Shan HZ, Zhou L, Yang L, Lei H, et al. Overexpression of Fbxo6 inactivates spindle checkpoint by interacting with Mad2 and BubR1. *Cell Cycle* 17(24):2779-2789. doi: 10.1080/15384101.2018.1557488 (2018).
64. Gong J, Cao J, Liu G, Huo JR. Function and mechanism of F-box proteins in gastric cancer (Review). *Int J Oncol* 47(1):43-50. doi: 10.3892/ijo.2015.2983 (2015).
65. Zhao Y, Liu J, Cai X, Pan Z, Liu J, Yin W, et al. Efficacy and safety of first line treatments for patients with advanced epidermal growth factor receptor mutated, non-small cell lung cancer: Systematic review and network meta-analysis. *BMJ* 367:l5460. doi: 10.1136/bmj.l5460 (2019).
66. Cai L, Li J, Zhao J, Guo Y, Xie M, Zhang X, et al. Fbxo6 confers drug-sensitization to cisplatin via inhibiting the activation of Chk1 in non-small cell lung cancer. *FEBS Lett* 593(14):1827-1836. doi: 10.1002/1873-3468.13461 (2019).
67. Schiller JH, Harrington D, Belani CP, Langer C, Sandler A, Krook J, et al. Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer. *N Engl J Med* 346(2):92-8. doi: 10.1056/NEJMoa011954 (2002).
68. Arriagada R, Bergman B, Dunant A, Le Chevalier T, Pignon JP, Vansteenkiste J. Cisplatin-Based Adjuvant Chemotherapy in Patients with Completely Resected Non-Small-Cell Lung Cancer. *N Engl J Med* 350(4):351–60. doi: 10.1056/NEJMoa031644 (2004).

69. Gong J, Zhou Y, Liu D, Huo J. F-box proteins involved in cancer-associated drug resistance. *Oncol Lett* 15(6):8891-8900. doi: 10.3892/ol.2018.8500 (2018).
70. Singh V, Katiyar A, Malik P, Mohan A, Singh H, Jain D. P1. 14-48 Whole Exome Sequencing (WES) in Non-Small Cell Lung Carcinoma (NSCLC): Identification of Novel Biomarkers. *Journal of Thoracic Oncology* 14 (10), S574. DOI: 10.1016/j.jtho.2019.08.1199 (2019).

Tables

Table 1: Summary of high-confidence somatic SNVs and indels observed in lung cancer patients

Gene	Nucleotide mutation	GRCh38 location	Mutation type	Amino acid alteration
TEKT4	G>A	Chr2: 94876674	SNV/nonsynonymous	NM_144705:exon6:c.G1213A:p.A405T
HRNR	G>C	Chr1:152216579	SNV/nonsynonymous	NM_001009931:exon3:c.C5050G:p.R1684
KCNJ18	C>T	Chr17:21703417	SNV/nonsynonymous	NM_001194958:exon3:c.C631T:p.L211F
	T>G	Chr17:21703571		NM_001194958:exon3:c.T785G:p.I262S
	G>A	Chr17:21703568		NM_001194958:exon3:c.G782A:p.R261H
ESRRA	G>T	Chr11:64315821	SNV/nonsynonymous	NM_001282450:exon7:c.G1127T:p.R376L
	CGGG>C	Chr11:64315823	Deletion/nonframeshift	NM_001282450:exon7:c.1130_1132del:p.3
CTBP2	C>CAAA	Chr10:124994577	SNV/nonsynonymous	NM_022802:exon5:c.G2292T:p.Q764H
	A>T	Chr10:124994505		NM_022802:exon5:c.T2364A:p.H788Q
	A>T	Chr10:124994563		NM_022802:———exon5:c.T2306A:p.L769
	C>T	Chr10:124994542		NM_022802:exon5:c.G2327A:p.S776N
MPRIP	CCAGCAG> CCAG,C	Chr17:17136247	Deletion/nonframeshift	NM_015134:exon6:c.537_539del:p.179_18
TBP	A>ACAG	Chr6:170561958	Insertion /nonframeshift	NM_003194:exon3:c.222_223insCAG:p.Q7
FBXO6	A>G	Chr1:11668809	SNV/nonsynonymous	exon2:c.A151G:p.M51V
FOLR3	CTA>C	Chr11:72139110	Deletion/nonframeshift	exon3:c.46_47del:p.Y16fs
GPRIN2	C>T	Chr10:46550016	SNV/nonsynonymous	NM_014696:exon3:c.G721A:p.V241M
LILRA2	G>A	Chr19:54574903	SNV/stopgain	NM_001290270:exon3:c.G489A:p.W163X
MTRNR2L8	T>TGTGTC	Chr11:10508153	Insertion/frameshift	NM_001190702:exon1:c.73_74insGACAC:
UMPS	CT>C	Chr3:124730565	Deletion/nonframeshift	NM_000373:exon1:c.95delT:p.L32fs
FDFT1	GTCCCAC>G	Chr8:11808709	Deletion/nonframeshift	exon1:c.193_198del:p.65_66del
KRT18	G>T	Chr12:52949285	SNV/stopgain	NM_000224:exon1:c.G112T:p.G38C
NYX	G>A	ChrX:41473563	SNV/nonsynonymous	NYX:NM_022567:exon2:c.G110A:p.C37Y
LRP2	C>T	Chr2:169256124	SNV/nonsynonymous	LRP2:NM_004525:exon19:c.G2752A:p.G9
MIR3689F	CGGGATCACACCTCCCAGGAAA	Chr9:134850674	SNV/nonsynonymous	-
	GCACGGGATCAGACCTCCCAGG			
	GAGCACGGGATCACACCTCCCAGCGAGTGT>C			

Table 2: Characteristics of candidate variants from public resources and published literature

Gene	Lung carcinoma	Other carcinoma	Hub genes	Protein overexpression in lung cancer	Pathway associated with lung cancer	Ongoing drug trials	LoF-intolerant genes
CTBP2	+	+	+		P16	+	
ESRRA	+	+	+	+	VEGF		
FBXO6	+		+	+	ERDA	+	
FDFT1	+	+	+	+	Mevalonate, WNT		
FOLR3		+		+			+
GPRIN2					Glutamate		
HRNR	+	+	+	+	AKT		
KCNJ18				+	P53		
LILRA2	+			+			+
LRP2		+	+		MAPK, JNK, Headong		
MIR3689F							
MPRIP	+	+	+	+	Fusion	+	
MTRNR2L8							+
NYX							
TBP	+	+	+		RAS		
TEKT4		+			P13K/AKT		
UMPS	+	+	+	+	Nucleotide metabolism		+
KRT18	+	+	+	+			

Table 3: Candidate variants identified through whole-exome sequencing which passed in the validation practice by Sanger sequencing

T.S.	KCNJ18	GPRIN2	TEKT4	HRNR	FOLR3	ESSRA	CTBP2	MPRIP	TBP	FBXO6
ADCA subtype										
ADCA01	B	B	B	B	B	B	B	-	-	-
ADCA02	B	B	B	B	B	B	B	-	-	-
ADCA03	B	B	B	B	B	W	B	-	-	-
ADCA04	B	B	B	B	B	B	W	-	B	-
ADCA05	B	B	B	B	B	W	B	-	-	-
ADCA06	B	W	B	B	B	B	B	-	-	-
ADCA07	B	B	B	B	B	B	W	-	-	-
ADCA08	B	B	B	B	-	B	B	-	-	-
ADCA09	B	B	B	B	B	B	B	B	-	-
ADCA10	B	W	B	B	B	W	B	B	-	-
SQCC subtype										
SQCC11	B	B	B	B	B	W	W	-	-	B
SQCC12	B	B	B	B	B	B	B	-	-	-
SQCC13	B	W	B	B	B	B	B	-	-	-
SQCC14	B	B	B	B	B	W	W	-	-	-
SQCC15	B	B	B	B	B	B	B	-	-	-
SQCC16	B	B	B	B	B	W	B	-	-	-
SQCC17	B	B	B	-	B	B	B	-	-	-
SQCC18	B	B	B	B	B	W	B	-	-	-
SQCC19	B	B	B	B	B	B	B	-	-	-

B: candidate variants confirmed by both WES and Sanger sequencing

W: candidate variants confirmed by WES only and missed by Sanger sequencing possibly due to low sensitivity

-: candidate variants absent in both WES and Sanger sequencing methods

T.S: tumor sample IDs

Figures

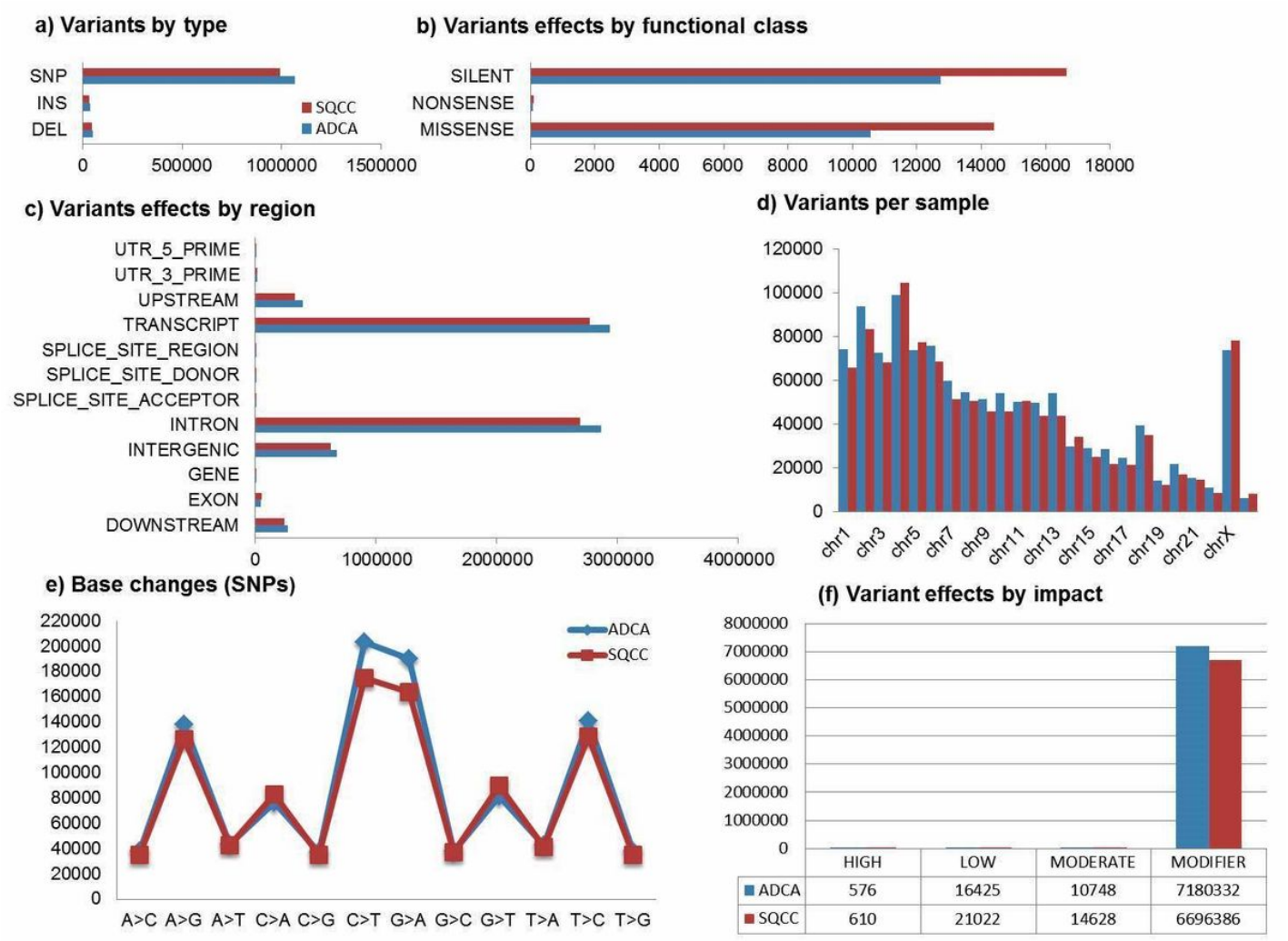

Figure 1

Summary of variants identified by whole exome sequencing. The bar graph represents the data as a) Number of variants by type, b) Variants effects by functional class, c) Variants effects by region, d) Variants per sample, e) Base changes in SNPs and f) Variants effect by impact.

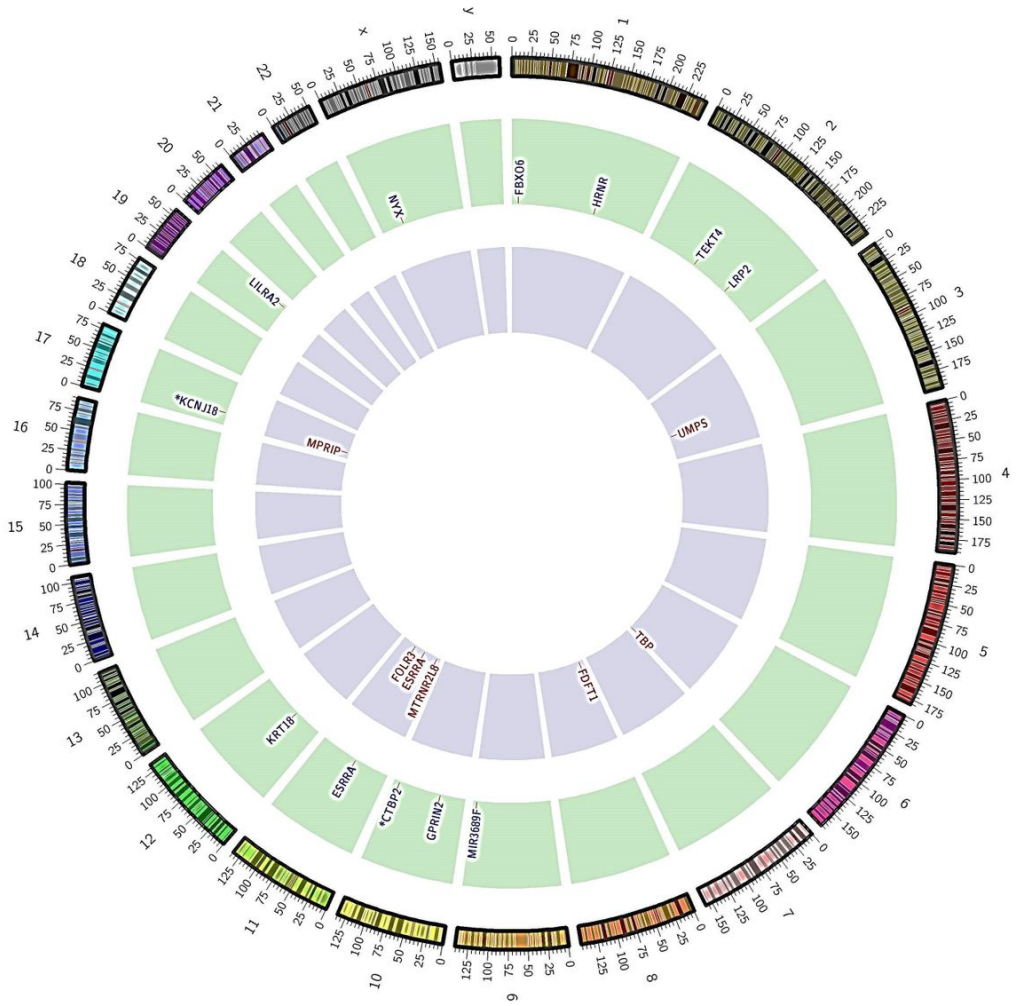

Figure 2

Non-synonymous somatic SNVs and INDELs identified in lung cancer patients by whole-exome sequencing. The outer colored ring and number indicate chromosome number and partition; the middle green ring and letters represent genes with non-synonymous SNVs and their corresponding chromosomes; the inner violet ring shows non-synonymous INDELs and their corresponding chromosomes. The star symbol signifies that genes associated with more than one mutation.

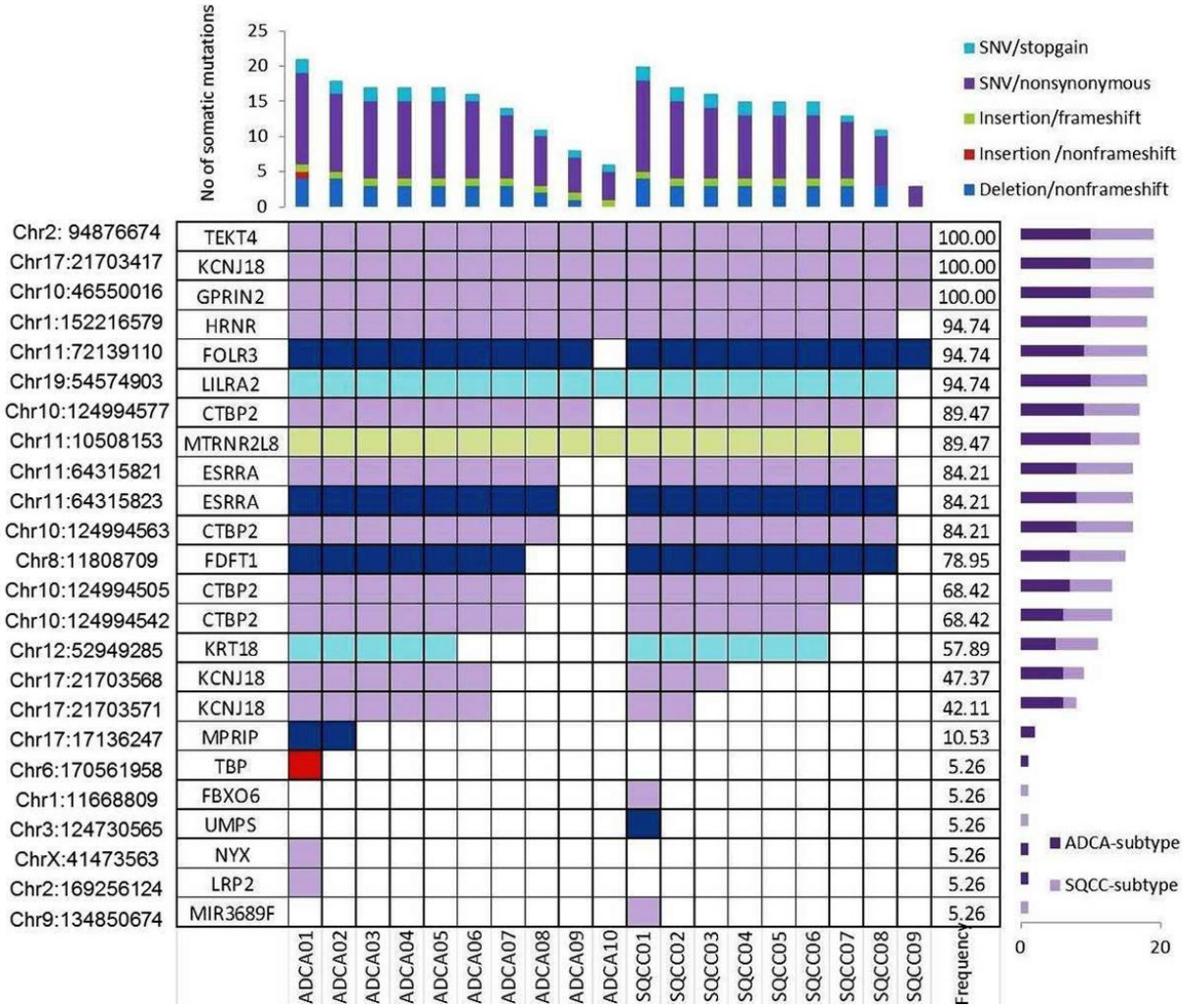

Figure 3

Oncoplot for the somatic variants in non-small cell lung cancer (NSCLC). The graph depicting top 18 mutated genes ordered by decreasing frequency. The right barplot shows overall frequency in ADCA and SQCC-subtype. The colour box represents the type of mutations including SNV/nonsynonymous (violate), SNV/stopgain (light blue), deletion/nonframeshift (dark blue), insertion/frameshift (light green) and insertion nonframeshift (red). The top stacked barplot shows a number of somatic mutations per sample.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile1SciRep.docx