Aus dem Charité Comprehensive Cancer Center
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Somatic genome alterations in relation to age in smoking related cancers

zur Erlangung des akademischen Grades
Doctor rerum medicinalium (Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Stefano Meucci

aus Fiesole (Italien)

Datum der Promotion: 05.03.2021

Table of contents

Abstract (English)

Head and Neck Squamous Cell Carcinoma (HNSCC), Lung Squamous Cell Carcinoma (LUSC) and Lung Adenocarcinoma (LUAD) are among the most common cause of global cancer-related mortality and the common major risk factor is smoking consumption. By analyzing 203 HNSCC, 480 LUSC and 486 LUAD samples from The Cancer Genome Atlas, we systematically studied the mutational load as well as mutational patterns in relation to patient age. Multiple mutational processes appear to be simultaneously operative with various dynamic changes due to the endogenous and exogenous environments, life style habits and physiological ageing. We found a proportional increase, independently of smoking consumption, of the HNSCC mutation frequency rate in relation to the patient age. Therefore, multiple factors might participate to the accumulation of genetic events in the elderly and the prolonged tobacco exposure might increase the ageing-related SNPs burden. On the contrary, LUSC and LUAD showed a higher mutational rate among younger patients. *TP53* mutations in younger LUAD patients might be a crucial factor enhancing the sensitivity to smoking related mutations leading to a burst of somatic alterations. Indeed, *TP53* mutations and patient age significantly affected the higher mutational rate of younger patients. *TP53* itself showed a higher sensitivity to smoking related C>A mutations in younger LUAD patients. *TP53* mutated and *TP53* wild type patient groups might represent phenotypes which endure ageing related mutational processes with different strength. LUSC was enriched of defective DNA mismatch repair (MMR) related signatures, in particular the Signature 6 (SI6) in younger and the Signature 26 (SI26) in older patients. Therefore, the two distinct age-related defective DNA MMR signatures SI6 and SI26 might be crucial mutational patterns in LUSC tumorigenesis, which may develop distinct phenotypes. The accumulation of SNPs may not follow distinct mutational patterns but rather an accumulation of mutations in specific pathways. Disruption of Axon Guidance and ECM-extracellular matrix pathways were enriched among the higher mutational rate samples of HNSCC and LUSC. We hypothesize that these pathways might have unknown crucial roles in genome stability maintenance. Further studies with larger numbers of individuals of different ages and diversity of normal tissues are essential to elucidate the intricate relationship between smoking consumption and mutational patterns in relation to intrinsic ageing processes. A better comprehension of tumorigenesis in relation to patient age might be relevant for cancer prevention and age adjusted treatment decisions and should therefore be taken under closer consideration in future studies.

Abstract (German)

Head and Neck Squamous Cell Carcinoma (HNSCC), Lung Squamous Cell Carcinoma (LUSC) und Lung Adenocarcinoma (LUAD) gehören zu den häufigsten Ursachen der weltweiten krebsbedingten Mortalität. Der häufigste gemeinsame Risikofaktor ist hierbei der Rauchkonsum. Durch die Analyse von 203 HNSCC, 480 LUSC und 486 LUAD Proben aus dem „The Cancer Genome Atlas", untersuchten wir systematisch die Mutationslast sowie Mutationsmuster in Bezug auf das Alter der Patienten. Mehrere Mutationsprozesse scheinen aufgrund der endogenen und exogenen Umgebung, der Lebensgewohnheiten und der physiologischen Alterung gleichzeitig mit verschiedenen dynamischen Veränderungen aufzutreten. Wir fanden einen Anstieg der HNSCC-Mutationsfrequenzrate unabhängig vom Rauchkonsum im Verhältnis zum Patientenalter. Daher könnten mehrere Faktoren zur Anhäufung genetischer Ereignisse bei älteren Menschen beitragen, und die verlängerte Tabakbelastung könnte die altersbedingte SNPs-Belastung erhöhen. Im Gegensatz dazu zeigten LUSC und LUAD eine höhere Mutationsrate bei jüngeren Patienten. TP53-Mutationen bei jüngeren LUAD-Patienten könnten ein entscheidender Faktor sein, der die Empfindlichkeit gegenüber rauchbedingten Mutationen erhöht, was zu einem Ausbruch von somatischen Veränderungen führt. Tatsächlich beeinflussten TP53-Mutationen und das Alter der Patienten signifikant die höhere Mutationsrate jüngerer Patienten. TP53 selbst zeigte eine höhere Empfindlichkeit gegenüber rauchbedingten C> A-Mutationen bei jüngeren LUAD-Patienten. TP53-mutierte und TP53-Wildtyp-Patientengruppen könnten Phänotypen darstellen, die altersbedingte Mutationsprozesse in unterschiedlichem Maße ertragen. LUSC war angereichert mit defekten DNA Mismatch Repair (MMR) verwandten Signaturen, insbesondere die Signatur 6 (SI6) bei jüngeren und die Signatur 26 (SI26) bei älteren Patienten. Daher könnten die zwei unterschiedlichen altersbedingten defekten DNA-MMR-Signaturen SI6 und SI26 entscheidende Mutationsmuster in der LUSC-Tumorgenese sein, die unterschiedliche Phänotypen entwickeln können. Die Akkumulation von SNPs folgt möglicherweise nicht bestimmten Mutationsmustern, sondern eher einer Akkumulation von Mutationen in spezifischen Signalwegen. Eine Unterbrechung der Axon-Führung und der extrazellulären ECM-Matrix-Wege traten bei den Proben mit höheren Mutationsraten von HNSCC und LUSC gehäuft auf. Wir stellen die Hypothese auf, dass diese Wege eine unbekannte und entscheidende Rolle bei der Aufrechterhaltung der Genomstabilität spielen könnten. Weitere Studien mit einer größeren Anzahl von Individuen unterschiedlichen Alters und unterschiedlicher Gewebeverteilung sind essentiell, um den komplizierten Zusammenhang zwischen Rauchkonsum und Mutationsmustern in Bezug auf intrinsische Alterungsprozesse

aufzuklären. Ein besseres Verständnis der Tumorgenese in Abhängigkeit vom Patientenalter könnte für die Krebsvorsorge und altersgerechte Behandlungsentscheidungen relevant sein und sollte daher in zukünftigen Studien näher betrachtet werden.

Introduction

Background research and wet lab experiences

One of the main research interest of our laboratory at the Charité Comprehensive Cancer Center is the establishment of predicting and monitoring non-invasive assays in order to assess the tumor mutational profile and personalized treatment strategies.

Among our past studies, we investigated on CTCs specific markers in non-small-cell lung carcinoma (NSCLC) through the surface staining of protein such as CD45, EpCam and CK-19 followed by flow citometry analysis. In a recent study we investigated the Guanylyl Cyclase C (GCC) expression in tumor and normal rectal tissues in comparison with metastasis, CTCs and circulating cell-free mRNA through immunohistochemistry, PCR-based methods and flow citometry [1]. Guanylyl cyclase C (GCC) is a transmembrane surface receptor restricted to intestinal epithelial cells, from the duodenum to the rectum. Our study revealed a higher GCC expression in tumor tissues than in normal tissues of the rectum and a significant correlation of high GCC mRNA in circulation with tumor emboli in vessels, distal organ metastasis, and poor survival, which may promote the clinical application of GCC as a survival predictor for assessing tumor burden and a valuable biomarker for guiding treatment strategies in the future. Furthermore, we conducted a systematic review and meta-analysis to compare KRAS and BRAF mutations in paired CTCs and primary tumors from 244 CRC patients, to detect any possible discordance [2]. As predictive markers for anti-EGFR therapy, KRAS and BRAF mutations are routinely detected in primary and metastatic colorectal cancer (CRC) cells, but seldom in circulating tumor cells (CTCs). The results indicated mutational discordance between CTCs and primary CRCs, particularly in the stage IV and KRAS subgroups. Detecting mutations in CTCs could help explain mutational differences between tumor cells at local sites and distant metastases, thereby improving treatment outcomes and liquid biopsies. In parallel, we are currently establishing a protocol for single cell targeted mRNA sequencing and Whole Exome Sequencing (WES) of CTCs isolated from various cancer entities.

Finally, the collaboration with the Focus Area DynAge project arose a new line of research focused on the relationships between smoking related cancers and human ageing. Due to the high mutational rate and the intra-heterogeneity of smoking related cancers, PCR-based and targeted sequencing methods were not suitable for a comprehensive study of the tumor mutational landscape in relation to patient age. Therefore, we established a protocol for isolation

of DNA from primary FFPE (Formalin Fixed Paraffin Embeded) tissue available at the Charité biobank (ZeBanC) in order to establish DNA WES analysis. The DNA was isolated from tumor as well as from the adjacent normal tissue of 100 HNSCC samples. Meanwhile the enlarging The Cancer Genome Atlas (TCGA) public data bank allowed us to establish a bioinformatics workflow in order to perform a broader comprehensive study for the evaluation of the somatic alterations landscape in relation to patient age in smoking related cancers.

Thesis project

The mutational landscape present in a cancer genome is the cumulative result of endogenous and/or exogenous mutational processes (e.g., smoking), constant or sporadic and with different strengths along patient ageing. Therefore, multiple mutational processes are operative resulting in jumbled composite signatures and tumor characteristics vary between patients of different ages [3–6]. In order to provide better insight into the underlying genetic and epigenetic patterns of smoking-related cancers in relation to patient ageing, we performed three studies using the publicly available TCGA dataset of Head and Neck Squamous Cell Carcinoma (HNSCC), Lung Squamous Cell Carcinoma (LUSC) and Lung Adenocarcinoma (LUAD).

HNSCCs affect 600,000 patients per year worldwide while lung cancer is the most common cause of global cancer-related mortality. The major and common risk factor is smoking consumption. However, HNSCC might be caused by human papillomaviruses (HPV) infection [7]. In previous studies the mutation rate of HPV-positive tumors was lower than that found in HPV-negative HNSCC, which is mostly occurring in smokers [8]. In general, patients with HPV-positive tumors have non-mutated TP53, however, HPV itself inhibits p53 function. Conversely, HPV-negative tumors frequently harbour TP53 mutations [7,9]. Thus, in order to select a homogenous patient cohort for investigating possible age-related differences, we only considered the 203 patients from the TCGA-cohort which were HPV negative and carried a TP53 mutation.

The two major lung cancer histological classes are non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC) [10]. NSCLCs mostly comprise lung adenocarcinomas (LUAD) and lung squamous carcinomas (LUSC) [11]. LUAD and LUSC are characterized by a largely distinct mutational landscape. Only six common significantly mutated genes (i.e., TP53, RB1, ARID1A, CDKN2A, PIK3CA, and NF1) have been found between these two cancer types [11,12]. LUSC shows a pattern of somatic genome alteration analogous to the head and neck

squamous cell carcinoma (HNSCC), suggesting that cancers arising from developmentally similar cells of origin across different tissues may be more similar than cancers arising from different cells of origin within an anatomically defined tissue [11,13]. As shown in lung cancer the onset of smoking is usually in adolescence or young adult age, and smoking cessation is associated with the diagnosis of the malignancy [14,15]. Although the age at diagnosis is very closely correlated with the duration of smoking [14,15], previous studies performed on 34 tumor types of the TCGA dataset [16,17], showed significant negative correlations between SNPs and patient age only in LUSC and LUAD. While 29 tumor types exhibited positive correlations, among which the HNSCC [16,18]. Therefore, in the case of lung cancers, the hypothesis is that a tumor with defective DNA polymerases and DNA repair genes (a.k.a. mutator phenotype), rapidly accumulate somatic mutations and might have concealed any age-related increase in mutation frequency [16].

In order to investigate the relation between age-related accumulation of mutations and tumor mutational patterns, firstly we evaluated the correlation between patient age and the average number of SNPs in HNSCC, LUSC and LUAD. The study was expanded in lung cancers to CNVs and methylation changes as well as to SNPs profiling and the respective correlation to the previously defined signatures from the Catalogue Of Somatic Mutations In Cancer (COSMIC) [19] (http://cancer.sanger.ac.uk/cosmic/signatures). Characteristic combinations of mutation types arising from specific mutagenesis processes such as DNA replication infidelity, exogenous and endogenous genotoxins exposures, defective DNA repair pathways and DNA enzymatic editing are reported on the COSMIC database.

Furthermore, we performed gene-specific correlation analysis in relation to patient age with a particular focus on the significantly mutated genes [11]. Gene set enrichment analysis was as well performed in order to explore functional effect of somatic alterations in relation to patient age in HNSCC and LUSC. While a special focus was placed on the TP53 mutational profile in LUAD.

The current study may pave the way for future studies of molecular tumorigenesis in relation to human ageing and underlines the need to consider age-adjusted treatments not only based on age and morbidity of older patients, but also on differences in tumor biology.

Material and Methods

TCGA data sets

Multiplatform genomic data sets were generated by TCGA Research Network (http://cancergenome.nih.gov/). Cancer molecular profiling data were generated through informed consent as part of previously published studies [12] and analyzed in accordance with each original study's data use guidelines and restrictions. The clinical data of the HNSCC, LUSC and LUAD normal paired exome sequences were derived via download from the publicly available TCGA data matrix (https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm).

Patients selection and whole exome analysis

Somatic mutations of HNSCCs from the TCGA study was derived by download from the cBio Portal [18]. To create a homogenous set of HNSCC tumors we selected a subset of the TCGA cohort (**Figure S1**). The original patient cohort was selected by two selection criteria, the first (i) excluded 36 HPV(+) patients. The second selection criteria (ii) excluded 40 TP53 wild type patients, which left the final selected cohort of 203 HPV-negative/TP53-mutated patients with a total amount of 29.860 single nucleotide polymorphisms (SNPs) distributed on 11.489 genes. Entries without official gene names were removed.

LUSC and LUAD somatic mutations were obtained from the open access MAFs available from the GDC Legacy Archive (2016). We considered three different exclusion criteria for mutation data entries. In the first exclusion criteria, we considered only once a mutation present in different samples belonging to the same patient. With the second exclusion criterion, we removed mutations that were associated to more than one gene. In order to prevent false positive variant calls due to repetitive sequences, only somatic mutations with "ref context" containing less than 6 continuous single repetitions, less than 4 continuous duplets, less than 3 continuous triplets, less than 3 continuous quadruplets, less than 3 continuous quintuplets were kept.

SNP array-based copy number analysis

High-level copy gain or copy loss events for individual genes of LUSC and LUAD patients were inferred using publicly available Firehose's data (http://gdac.broadinstitute.org/runs/analyses__2016_01_28/data/) (+2 values being indicative of gains greater than 1-2 copies, -2 values being indicative of near total copy loss). Global CNV load were calculated summing the absolute values from each patients.

Array-based DNA methylation assay

The level 3 beta value DNA methylation scores for individual genes of LUSC patients were inferred using publicly available data generated by Illumina Human Methylation 450 platform downloaded from the GDC Legacy Archive (https://portal.gdc.cancer.gov/legacy-archive). Methylation values were mean centered and scaled to unit variance. After the transformation, the rate of methylation changes was calculated summing the values of each gene.

Single nucleotide variants and COSMIC signatures

The signature profile of LUSC and LUAD were evaluated considering the 96 possible single nucleotide variants (6 types of substitution x 4 types of 5' base x 4 types of 3' base). The profile of these 96 single nucleotide variants was considered as the results of the combination of the 30 different COSMIC signatures. The profile of each tumor sample can be represented by a unique contribution of each COSMIC signature as the following expression: a_1 x *SI1* + a_2 x *SI2* + a_3 x *SI3* + ... + a_{30} x *SI30 (1)*, where a_i is the coefficient representing the contribution of the i^{th} COSMIC signature. The coefficients of each tumor samples were calculated minimizing the difference between the tumor profile and the expression *(1)*. This procedure was implemented using the function *optim* (method "L-BFGS-B" [21]) of the R software [22].

Molecular pathway and biological process analysis

Unsupervised hierarchical clustering based on gene mutation frequencies was performed for different age groups of the HNSCC patient cohort. Genes were clustered according to Euclidian distance measure using the method "complete". Genes from age group specific clusters were extracted and tested using the online David Gene Ontology tool [23]. The full list of identified genes was used as background for enrichment calculation. In reverse, all genes of an enriched KEGG pathway were mapped to our list of mutation frequencies to investigate the number and distribution of mutated genes in the respective pathway within all ages of our dataset. A minimum difference of 0.15 in mutation frequency between at least two age groups was considered for cluster analysis. This cut-off allowed for a comprehensive set of mutational patterns within the specific age groups including less frequently mutated genes.

In LUSC and LUAD, pathway analyses were performed by ssGSEA using the GenePattern module ssGSEA Projection (v4) (genepattern.broadinstitute.org). ssGSEA enrichment scores were calculated from SNPs and CNVs in LUSC and LUAD data sets, as well as methylation

only in LUSC data set. The result is a single score per patient per gene set, transforming the original data sets into a more interpretable higher-level description. For the use of ssGSEA software, annotated gene sets reference were obtained from the C2 KEGG sub-collection of the Molecular Signature database (MSigDB) [24]. Silent mutations (point mutations that would not result in a change in the amino acid sequence) were not included in the analysis.

Statistical analysis

The relations between total average mutation frequencies and HNSCC patient age groups were calculated by linear regression using F-statistics. A two-sided p-value of below 0.05 (Pr(>|t|)) was considered significant. The frequency of mutation of a specific gene was calculated using the sum of mutations in a specific age group divided by the number of patients in the respective age group. The Spearman's Rank Correlation Coefficient was used to identify correlation between LUSC and LUAD patient ages and genomic/epigenomic data (e.g., SNP, CNV, and methylation loads). For every Spearman's test performed in this study, p-values were computed using algorithm AS 89 included in the R function *cor.test* where the permutation distribution was estimated by an Edgeworth approximation [25]. The coefficient interval of rho value was calculated by bootstraping (with 1000 replicates) using the function *spearman.ci* of the R package *RVAideMemoire*. Fisher's exact test was used to examine the significance of the association between COSMIC signature related subgroups (i.e., low-SI6/high-SI26 and high-SI6/low-SI26) and clinical/demographic/molecular patient features, such as gender, tobacco smoking history indicator, and mutated / wild type genes. Fisher's exact test was computed using the R function *fisher.test*. Wilcoxon Rank-Sum test was performed to compare continuous variables between two patient subgroups using the R function *wilcox.test*. A p-value <0.05 was considered to be significant. To account for multiple testing, a FDR of $\leq 20\%$ was applied to reduce identification of false positives [26]. The FDR was calculated using the R function *p.adjust*. All calculations were made using R software [22].

Results

Somatic alterations and patient age

Genome-wide mutations and epigenomic changes are expected to varying among tumor subtypes showing a different distribution across age. To characterize these distinct distribution patterns, we firstly estimated the global number of SNPs across HNSCC (203 patients), LUSC (480 patients) and LUAD (486 patients) cancer patient cohorts available through The Cancer Genome Atlas (TCGA).

A significant rise in the average number of mutations, calculated by linear regression using F-statistics, was observed with increasing age in HNSCC (**Table 1, Figure 1**). While through the Spearman's rank correlation coefficient analysis we observed a negative correlation between patient age and the global SNPs load in LUSC and LUAD, which indicated a higher mutational rate among younger patients (**Table 1, Figure 2a and 2c**).

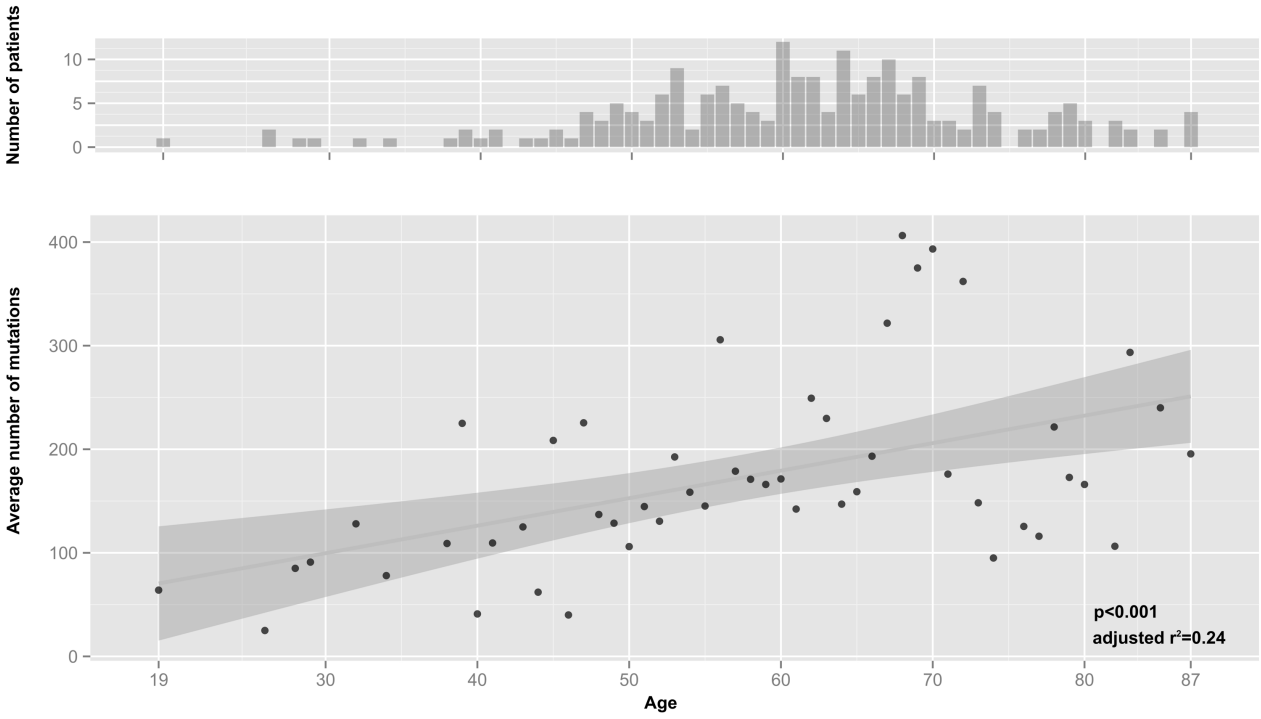

Figure 1: Correlation of average mutations and patient age in HNSCC – Upper graph number of patients in the respective age group. Lower graph average number of mutations for all patients in the respective age groups. Linear regression analysis was done using F-statistics and shows a significant increase in mutations in older patients ($p=0.000161$, adjusted $r^2=0.24$). Grey area around regression line indicates 95% confidence interval. Adapted from [18]. Copyright 2016 by Stefano Meucci. Adapted with permission.

Patient cohorts	Patients n.	rho [95%CI]	p-value	FDR
HNSCC (linear regression)	203	-	$1.61x10^{-4}$	-
LUSC (Spearman´s Rank Correlation)	480	-0.09 [-0.19 0]	$4.53x10^{-2}$	$1.81x10^{-1}$
LUAD (Spearman´s Rank Correlation)	486	-0.16 [-0.25 -0.07]	$3.93x10^{-4}$	$4.78x10^{-3}$

Table 1: SNPs loads correlations with patient age in HNSCC, LUSC and LUAD - Correlations between the SNPs loads and patient age for each patient cohort.

In order to evaluate only the disruptive mutations, we repeated the mutational load analysis excluding silent mutations (classified as low impact in **Table S1)** and multiple mutations in one gene**.** The same significant p-value was detected in HNSCC (**Figure S3**, **list of mutation frequencies in Table S2**). While we reported a slightly lower correlation in LUSC (*rho*=-0.08, *P*=0.077, FDR=0.26) and LUAD (*rho*=-0.16, *P*=0.0005, FDR=0.008) (**Table S3**).

The global CNVs load was as well investigated in LUSC and LUAD. No correlation with patient age was observed in LUSC (**Figure 2b**), while a significant negative correlation (*rho*=-0.16, *P*=0.0006, *FDR*=0.02) was identified in LUAD (**Figure 2d**). Methylation changes at CpG sites were evaluated in LUSC, showing a negative correlation with patient age (*rho*=-0.11, *P*=0.030, *FDR*=0.23), which indicated a higher a higher level of methylation at CpG sites among younger patients. We repeated the analysis on patient sub-cohorts established according to available clinical and molecular data for each tumor entity such as tumor localization, tobacco exposure data, tumor staging and mutational rate profile in order to explore the influence of patient features on the relation among somatic alterations and patient age. Both smokers and non-smokers subsets showed a significant correlation between the SNPs load and patient age in HNSCC, therefore the smoke history surprisingly seemed not to influence the age-related mutational load tendency (**Table S4**). While only the current smokers sub-cohort of LUAD showed a negative correlation (*rho*=-0.23, *P*=0.01, FDR=0.06) analog to the global cohort. LUSC current smokers group did not show any significant correlation. The analysis of lung cancer sub-cohorts with a high mutational rate (i.e., transversion-high status) showed a negative correlation between the SNPs load and patient age in both LUSC (*rho*=-0.11, *P*=0.03, FDR=0.16) and LUAD (*rho*=-0.23, *P*=0.00002, FDR=0.0007) while no correlations were detected in the low mutational rate sub-cohorts (i.e., transversion low status).

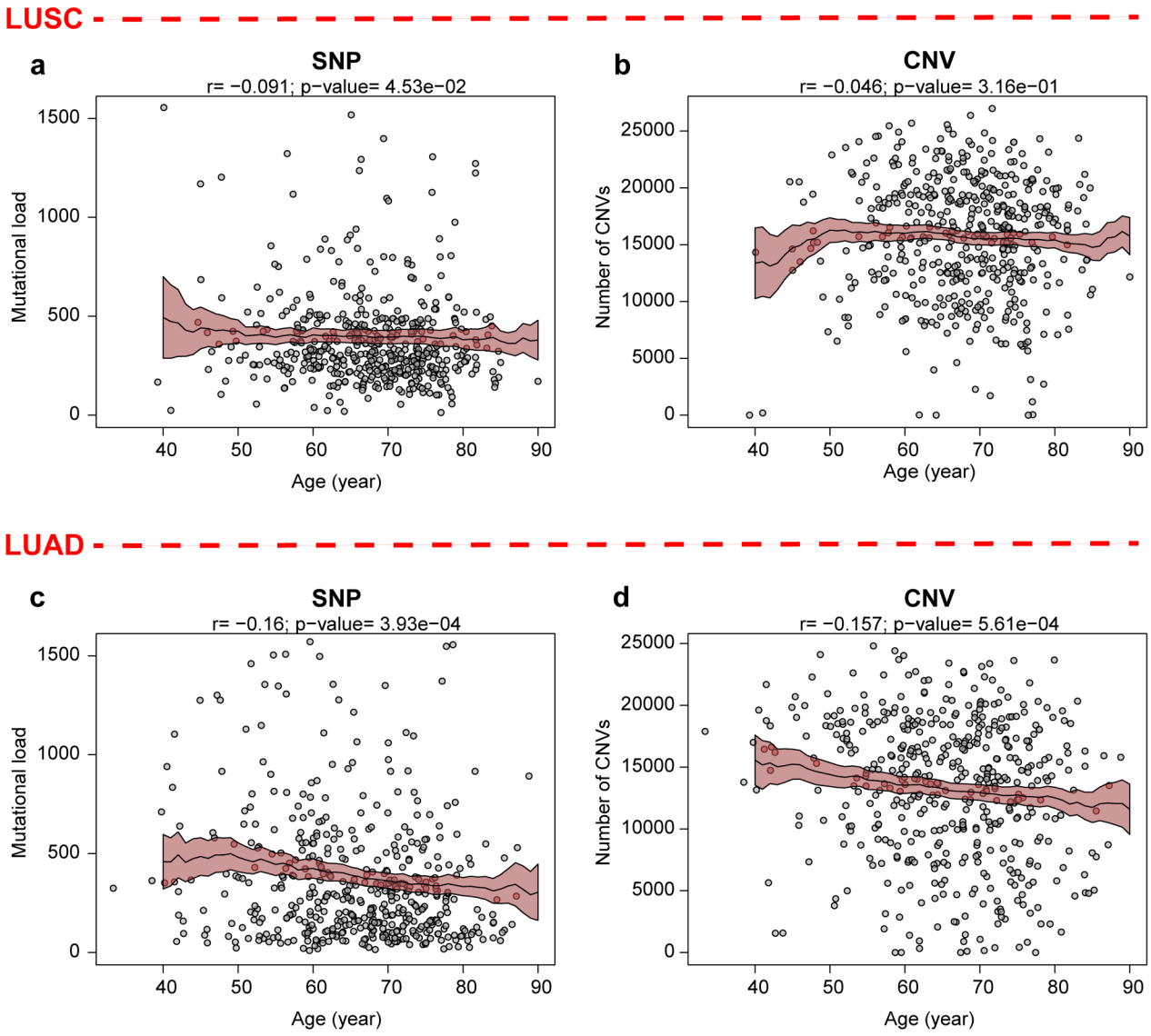

Figure 2: Correlation between genomic alterations and patient age in LUSC and LUAD – Number of **(a)** LUSC SNPs, **(b)** LUSC CNVs, **(c)** LUAD SNPs and **(d)** LUAD CNVs with their relative 95% confidence interval for each patient distributed along patient age. Medians (black line) and their relative 95% confidence interval (red area) were calculated locally in a range of ±10 years.

A special focus was placed on sub-cohorts established according to the *TP53* mutational profile in LUAD, in order to explore the influence of the most frequently mutated gene on the SNPs load. The *TP53* mutated sub-cohort showed a significant enrichment of SNPs among younger patients (*rho*=-0.21, *P*=5.74E-04, FDR=2.14E-03), while no correlation was detected in the *TP53* wild type cohort (**Figure S4, Table S3**). The *TP53* mutated sub-cohort showed a significantly higher percentage of current smokers (P=6.58 x10^{-6}, FDR=1.97x10^{-5}) and transversion-high profiles (P=2.13x10^{-4}, FDR=3.20x10^{-4}) (**Table S5**). The overall percentage of *TP53* mutated patient increased with an inverse proportion to time since smoking cessation, form

41.5% in Lifelong Non-Smokers to 70.7% in Current Smokers sub-group. Moreover, through the Wilcoxon's test, we detected a significant (P=2.44x10^{-3}) lower age mean in *TP53* mutated patients compared to *TP53* wild type patients. Overall the highest percentage of *TP53* mutated patients was detected in <50 (66.7%) and 50-60 (66%) age groups (**Figure S5, Table S6**). The CNVs load in *TP53* mutated cohort was overall higher than the wild-type counterpart and negatively correlated with patient age. No correlation was detected in *TP53* wild-type cohort (**Figure S4**, **Table S3).** Additionally, We used two-ways ANOVA to evaluate, independently, the effect of age and *TP53* mutations as well as their combination effect. We reported a higher mutational load in patients with *TP53* mutation (P<10^{-10}) and in younger patient (P=3.75x10^{-4}). Interestingly, we observed a statistically significant interaction between the patient age and the *TP53* status (P=4.04x10^{-2}) on the mutational load.

While no correlation between tumor staging and patient age was detected in lung cancers, HNSCCs showed a smaller size of the primary tumor, a decrease of lymph node metastasis and a higher percentage of “localized cancers” and “first stage of locally advanced cancers” in old patients (**Table S4**). The youngest patients group showed the same statistics as the oldest, however the results are not comparable due to the significant difference between the two age group ranges.

Gene Set Enrichment Analysis

As a next step, we wanted to investigate whether the mutational patterns in HNSCC and LUSC were a mere accumulation of random mutations or whether we could find age-specific patterns of mutated genes. Unsupervised hierarchical clustering based on gene mutation frequencies was performed for HNSCC patients pooled into age groups of decades aside two groups of the very young (pooled age 19-40) and very old (pooled ages 81 to 87). The results showed two relevant clusters of frequently mutated genes, in particular, 39 genes in very young patients and 108 genes in very old patients (**Table S7 with cut clusters A and B** from **Figure S6**). Both the young as well as the old group, each consisting of 11 patients, displayed prominent differences to middle age groups. The genes of the two specific clusters were tested for KEGG pathway enrichments using the online David Gene Ontology tool [23]. The old age gene cluster showed six statistically significant enriched pathways, two of which, the Axon-Guidance ($p < 0.003$) and ECM-Receptor Interaction ($p < 0.04$) pathways, stood out due to their role in angiogenesis processes and cell / tissue architecture maintenance respectively (**Table S8**). Further division of

the old and young age groups (**Figure S7** and **Figure S8**) allowed a highly detailed overview (**Table S8)**.

We mapped all genes of the Axon-Guidance pathway (according to the KEGG database) to our list of mutation frequencies to see if the pathway is affected in other age groups as well without being particularly enriched (**Figure 3, Table S8**). From the total 127 genes of the "Axon Guidance" pathway 99 mapped to our data. As the heatmap showed, mutations in this pathway were present in other ages as well, yet at a lower frequency than in the two old age groups. Again, the trend of increased pathway alterations towards old ages was visible. ECM-Receptor Interaction, NOTCH Signaling and Focal Adhesion pathways were as well mapped to our list of mutation frequencies (**Figure S9**, **Figure S10**, **Figure S11**). To study the molecular effects of somatic alterations in LUSC, we projected the SNPs, CNVs and DNA methylation values into the space of the 186 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways by means of single-sample gene set enrichment analysis (ssGSEA) (**Table S9**).

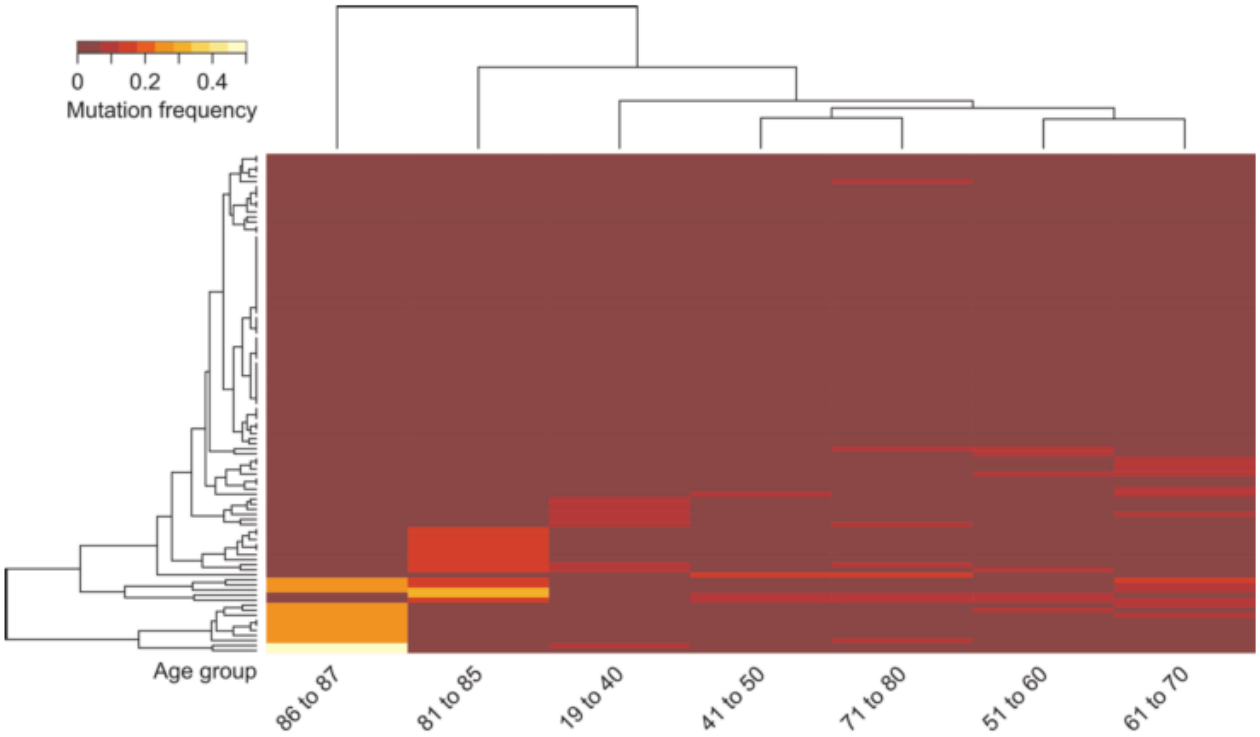

Figure 3: Unsupervised hierarchical clustering of mutation frequencies of genes involved in the "Axon Guidance" pathway (according to the KEGG database) - Patients were grouped into age groups of very young (ages 19-40), decades in-between and two separate old ages groups (ages 81-85 and 86 to 87), then clustered according to the mutation frequencies of all quantified genes of the "Axon Guidance" pathway. Adapted from [18]. Copyright 2016 by Stefano Meucci. Adapted with permission.

When evaluating the global cohort, we detected a significant negative correlation between patient age and SNPs harboring on Axon-Guidance (*rho*=-0.15, *p*=0.0007, *FDR*=0.14) and ECM Receptor Interaction (*rho*=-0.13, *p*=0.003, *FDR*=0.16) pathways, particularly in the 51-60 age group. Furthermore, the Axon-Guidance (*rho*=-0.16, *p*=0.001, *FDR*=0.12) pathway was the only negatively enriched pathway in transversion-high sub-cohort.

Age-related COSMIC signatures in lung cancer

Somatic mutation profile is the sum of multiple mutation processes, such as the intrinsic infidelity of the DNA replication machinery, exogenous or endogenous mutagen exposures, enzymatic modification of DNA, and defective DNA repair. In order to investigate on the differences in mutational profile, we categorized each single nucleotide variants incorporating information on the bases immediately 5' and 3' to each mutated base. We deconvoluted trinucleotide variants profiles into the 30 different signatures described in the COSMIC database [5,19]. So, we were able to characterize each patient by a different "intensity" combination of the 30 COSMIC signatures. Then, we performed the Spearman's rank correlation test between the intensities of each COSMIC signature and the patient age (**Table S10**).

The smoking-related Signature 4 (SI4) associated with C>A transversions was negatively correlated with patient age in both LUSC (*rho*=-0.11, *p*=0.02, *FDR*=0.21) (**Figure S12c)** and LUAD (*rho*=-0.18, *p*=0.000006, *FDR*=0.002). In particular, in LUAD we identified the same trend in TP53-mutated (*rho*=-0.27, *p*=0,000007, *FDR*=0.0002) (**Figure 4a)**, transversion-high (*rho*=-0.26, *p*=0.000002, *FDR*=0.00005) and current smokers (*rho*=-0.30, *p*=0.001, *FDR*=0.03) sub-cohorts. While no correlation was identified in TP53 wild-type sub-cohort (**Figure 4b**).

The age-related Signature 1 (SI1), mainly consisting of C>T transitions, was positively correlated in LUAD global cohort (*rho*=0.14, *p*=0,001, *FDR*=0.02) as well as TP53 mutated (*rho*=0.18, *p*=0.003, *FDR*=0.03) (**Figure 4a)** and transversion-high (*rho*=0.15, *p*=0.007, *FDR*=0.07) sub-cohorts, showing the simultaneous ongoing age-related accumulation of SNPs. While no correlation was identified in TP53 wild-type sub-cohort (**Figure 4b**).

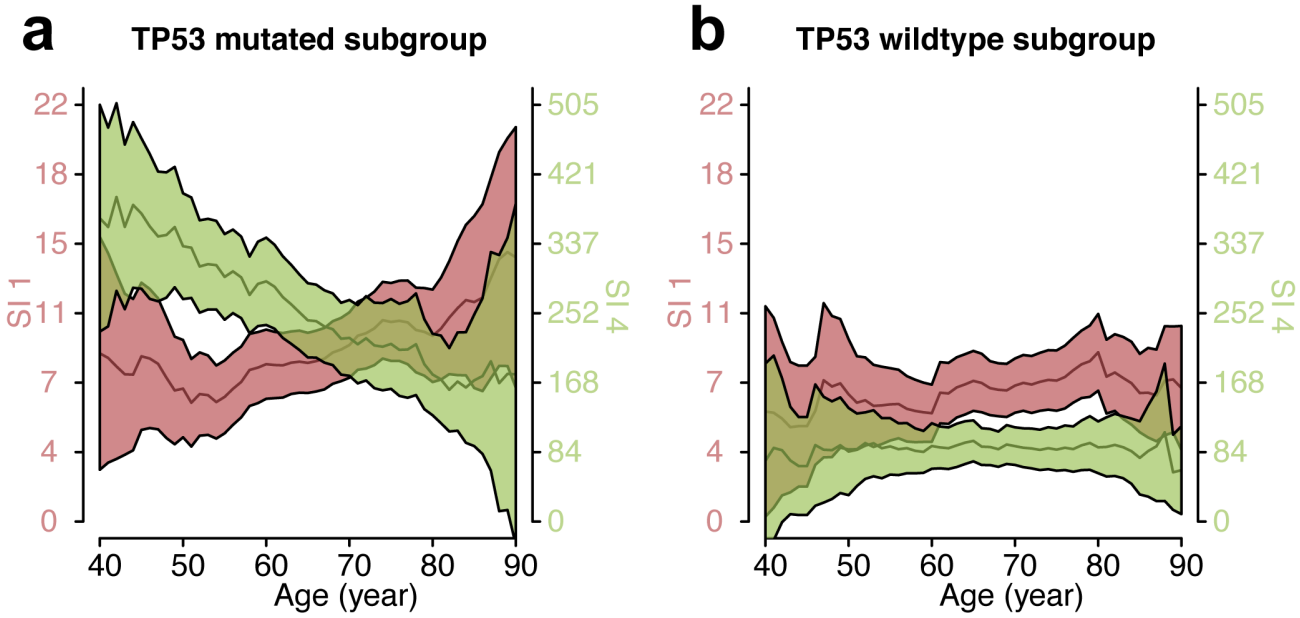

Figure 4: Correlation of SNPs profiling and patient age in TP53 mutated and TP53 wild type patient sub-cohorts. Correlation between the smoking related SI4 (green graph) and the age related SI1 (red graph) with patient age in **(a)** TP53 mutated and **(b)** TP53 wild type patient sub-cohorts. Medians (black line) and their relative 95% confidence interval (colored area) were calculated locally in a range of ±10 years. Adapted from "Somatic genome alterations in relation to age in lung adenocarcinoma" by S.Meucci (unpublished). Copyright 2016 by Stefano Meucci. Adapted with permission.

In LUSC, the defective DNA mismatch repair (MMR)-related signature 6 (SI6) was negatively correlated (*rho*=-0.13, *p*=0.004, *FDR*=0.12) with patient age (**Figure S12a**) while the signature 26 (SI26) as well associated with defective DNA MMR, was positively correlated (*rho*=0.11, *p*=0.013, *FDR*=0.20) with patient age (**Figure S12b**). In order to study the patient sub-cohorts, which predominantly exhibit SI26 and SI6, we divided the overall LUSC cohort into four subgroups using the mean values of SI6 and SI26 as threshold (**Figure S12d**): high-SI6/high-SI26 (77/480=16.0%), low-SI6/high-SI26 (55/480=11.0%), high-SI6/low-SI26 (223/480=45.8%), and low-SI6/low-SI26 (130/480=27.1%). We selected and characterized the low-SI6/high-SI26 and high-SI6/low-SI26 subgroups (**Table S11**). The patients age of the low-SI6/high-SI26 cohort was significantly higher than the high-SI6/low-SI26 cohort (Wilcoxon Rank-Sum test: *p*=0.005). The ssGSEA analysis [27] was repeated for the high-SI6/low-SI26 and low-SI6/high-SI26 sub-cohorts (**Table S12**). Using the Wilcoxon Rank- Sum test, we reported as major significant differences, that Extracellular Matrix (ECM)-Receptor Interaction pathway (*p*=0.0002, *FDR*=0.04) was significantly enriched of SNPs while the Nucleotide Excision Repair pathway was enriched in CNVs (*p*=0.0007, *FDR*=0.14) in high-SI6/low-SI26 sub-cohort. Using the Spearman's Rank Correlation Coefficient, we detected a negative correlation between SNPs

harboring on ECM Receptor Interaction pathway and patient age (*rho*=-0.16, *p*=0.016, *FDR*=0.73) in high-SI6/low-SI26 sub-cohort (**Figure S13)**.

Gene-specific alterations enrichment along patient ageing

In order to investigate on gene-specific driver mutations in relation to patient age, which might contribute to the higher mutational rate detected in younger patients, the Spearman's rank correlation was computed between patient age and gene specific SNPs load for each significantly mutated genes previously detected in LUSC and LUAD [11] (**Table S13, Figure S14**). CNVs and methylation changes were as well evaluated in LUSC. A negative correlation between LUSC patients age and both CNVs (*rho*=-0.13, *p*=0.005, *FDR*=0.16) and methylation changes (*rho*=-0.14, *p*=0.006, *FDR*=0.06) was detected on NOTCH1, while no SNPs correlation was displayed. In LUAD, a significant enrichment of SNPs on TP53 (rho=-0.13, P=5.25x10-3, FDR=9.98x10-2), as well as ATM (rho=-0.11, P=1.78x10-2, FDR=2.26x10-1) was detected in younger patients. While RBM10 disruptions were enriched among older patients (rho=0.13, P=4.81x10-3, FDR=9.98x10-2).

Finally, we calculated the frequencies of COSMIC signatures using the mutations identified in each of the LUAD significantly mutated genes (**Table S14**). TP53 and RMB10 were especially enriched of smoking related SI4 and the aflatoxin related SI24, both constituted of C>A transversions, indicating guanine damage that is being repaired by transcription-coupled nucleotide excision repair. The defective DNA mismatch repair related SI6, associated with high numbers of small (shorter than 3bp) insertions and deletions at mono/polynucleotide repeats was as well relatively enriched in TP53 (**Figure S15a**). RMB10 was also enriched of SI14, with unknown aetiology, mainly constituted by C>A and C>T mutations. Interestingly, ATM was enriched of SI3 and SI10, while the smoking related SI4 was entirely absent.

Unsupervised KODAMA algorithm was performed in order evaluate similarities among COSMIC signature profiles of the significant mutated genes in LUAD. TP53 and RMB10 shared the same cluster (**Figure S15b**), while ATM showed an independent profile.

Discussion

In the present study, we evaluated the genetic and epigenetic patterns of smoking-related cancers in relation to patient ageing. The average number of mutations in HNSCC for each patient showed a significant rise ($p < 0.01$) with increasing age in both smokers and non-smokers patients. Smoking consumption increased the overall mutational burden of HNSCC, although it did not influence the increasing SNPs burden in relation to patient age. As showed in previous studies [16,17], lung cancers displayed opposite correlations, indeed we confirmed a higher SNPs load in LUSC and LUAD younger patients. In particular, the correlation was higher in tumors with high mutational burden of smoking related C>A transversions. In LUAD, the analysis was repeated on sub-cohorts established according to the *TP53* mutational profile in order to explore the influence of the most frequently mutated gene on the SNPs load. The group of *TP53* mutated patients showed a higher percentage of current smokers and transversion-high profiles as well as a lower age mean compared to *TP53* wild type patients, which instead displayed a lower average number of SNPs with no correlation with patient age. Additionally, we identified that the effect of patient age and *TP53* mutations separately, as well as their interaction, significantly affected the higher mutational rate of younger *TP53*-mutated patients. Investigating on the underlined mutational patterns, we identified a significant enrichment of the smoking-related signature (i.e., SI4) among the overall LUSC younger patients, as well as among younger LUAD *TP53* mutated patients.

Therefore, in LUAD the cumulative effect of smoking consumption, *TP53* mutations and a younger age significantly affected the overall mutational load among younger patients. Although the correlation is subtle in LUSC, we observed that younger patients also developed a higher sensitivity to smoking-related mutations. Past studies described a similar scenario showing that despite maintained carcinogen exposure, tumors from smokers showed a relative decrease in smoking-related mutations over time [28,29].

Furthermore, the LUAD *TP53* mutated sub-cohort displayed the concurrent ongoing accumulation of SI1 along patient ageing. SI1 is largely made up of C>T substitutions at CpG dinucleotides, which are the results of an endogenous mutational process initiated by spontaneous deamination of 5-methylcytosine, enzymatic deamination of cytosine, or polymerase errors [3,4,30,31]. The SI3, associated with failure of DNA double-strand break-repair by homologous recombination was as well increasing along patient age in the *TP53* mutated sub-cohort. Recent studies showed that impaired DNA double-strand break repair

contributes to the age-associated rise of genomic instability in humans [32]. Meanwhile, although present and past smoking is reported in *TP53* wild type patients, no correlation between mutational signatures and patient age was detected. As shown in past studies, the mutational profile of cancer cells might reflect the mutational processes operative in aging in a given tissue [33]. Therefore, we hypothesize that *TP53* wild type patients might represent a phenotype with greater DNA stability, which may confine the ongoing age-related accumulation of genetic events as well as the increasing mutational burden due to smoking consumption. Although previous studies revealed that the number of *TP53* mutations are common in noncancerous tissue and accumulate with age and tobacco consumption [33–35], we detected an overall higher rate of *TP53* mutations in younger patients particularly in <50 and 50-60 age groups.

Besides the negative correlation of SI4, LUSC mutational profile showed the defective DNA MMR SI6 enriched in younger patients. SI6 is characterized predominantly by C<T at NpCpG sites (any nucleotide followed by C followed by G). While the SI26, mostly composed of T<C transitions, was enriched in older patients. Both SI6 and SI26 are found in microsatellite unstable tumors with high numbers of small (shorter than 3bp) insertions and deletions at mono/polynucleotide repeats [36]. The role of MMR system is to recognize and repair erroneous insertion, deletion, and mis-incorporation of bases arising during DNA replication and homologous recombination, as well as repairing some forms of DNA damage. Given the importance of these processes in the maintenance of genomic stability, DNA MMR deficiency might leads to hypermutation [37]. A recent study showed that out of a large number of DNA repair deficiencies analyzed, MMR deficiency leads to the by far highest mutation rate [36]. Our results suggest that different causing factors might contribute to MMR system aberrations along LUSC patient ageing.

To study the molecular effects of the mutational patterns, gene unsupervised clustering and gene ontology analysis was performed on HNSCC data, while single-sample gene set enrichment analysis (ssGSEA) was performed on the LUSC data set. HNSCC patient cohort showed distinct clusters of genes expressed with a higher mutation frequency for the very young and very old ages. While the young age group did not reveal pathway enrichment, the old age group showed enrichment of KEGG pathways, including ECM-Receptor Interaction and Axon Guidance [18]. Interestingly, the same pathways were found enriched of SNPs in younger LUSC patients, which have the higher mutational rate profile [38]. Therefore, although the inverse tendency, Axon Guidance and ECM-Receptor Interaction pathways disruptions seem to show a relation with

higher mutational rate squamous carcinomas. Several studies reported that Axon Guidance pathway is involved in lung cancer development and progression through interacting with cell survival, migration, and tumor angiogenic pathways [39]. While the ECM-Receptor Interaction pathway is structurally and functionally involved in interactions at the ECM which lead to a direct or indirect control of cellular activities such as cell migration, differentiation, proliferation, and apoptosis [40]. Aberrant ECM may promote genetic instability and might compromise DNA repair pathways necessary to prevent malignant transformation [41]. Further studies are needed to determine whether disruptions in these pathways are a correlative phenotype to higher mutational rate squamous carcinomas or a causative factor.

In conclusion, multiple mutational processes appear to be simultaneously operative with various dynamic changes due to the endogenous and exogenous environments, life style habits and physiological ageing. We found a proportional increase, independently of smoking consumption, of the HNSCC mutation frequency rate in relation to the patient age. Therefore, multiple factors might participate to the accumulation of genetic events in the elderly and the prolonged tobacco exposure might increase the ageing-related SNPs burden. On the contrary, LUSC and LUAD showed a higher mutational rate among younger patients. *TP53* mutations in younger LUAD patients might be a crucial factor enhancing the sensitivity to smoking related mutations leading to a burst of somatic alterations. Indeed, *TP53* mutations and patient age significantly affected the higher mutational rate of younger patients. *TP53* itself showed a higher sensitivity to smoking related C>A mutations in younger patients. *TP53* mutated and *TP53* wild type patient groups might represent phenotypes which endure ageing related mutational processes with different strength. LUSC was enriched of defective DNA MMR signatures, in particular the SI6 in younger and the SI26 in older patients. Therefore, the two distinct age-related defective DNA MMR signatures SI6 and SI26 might be crucial mutational patterns in LUSC tumorigenesis, which may develop distinct phenotypes.

The accumulation of SNPs may not follow distinct mutational patterns but rather an accumulation of mutations in specific pathways. Disruption of Axon Guidance and ECM-extracellular matrix pathways were enriched among the higher mutational rate samples of both HNSCC and LUSC. We hypothesize that these pathways might have unknown crucial roles in genome stability maintenance.

Further studies with larger numbers of individuals of different ages and diversity of normal tissues are essential to elucidate the intricate relationship between smoking consumption and

mutational patterns in relation to intrinsic ageing processes. A better comprehension of tumorigenesis in relation to patient age might be relevant for cancer prevention and age adjusted treatment decisions and should therefore be taken under closer consideration in future studies.

References

1. Liu Y, Cheng G, Qian J, Ju H, Zhu Y, Stefano M, Keilholz U, Li D. Expression of guanylyl cyclase C in tissue samples and the circulation of rectal cancer patients. Oncotarget. 2017;8:38841–9.

2. Liu Y, Meucci S, Sheng L, Keilholz U. Meta-analysis of the mutational status of circulation tumor cells and paired primary tumor tissues from colorectal cancer patients. Oncotarget. 2015;8:77928–41.

3. Alexandrov LB, Stratton MR. Mutational signatures: The patterns of somatic mutations hidden in cancer genomes. Curr Opin Genet Dev. 2014;24:52–60.

4. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, Stratton MR. Clock-like mutational processes in human somatic cells. Nat Genet. 2015;47:1402–7.

5. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio S a JR, Behjati S, Biankin A V, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjörd JE, Foekens J a, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Imielinsk M, Jäger N, Jones DTW, Jones D, Knappskog S, Kool M, Lakhani SR, López-Otín C, Martin S, Munshi NC, Nakamura H, Northcott P a, Pajic M, Papaemmanuil E, Paradiso A, Pearson J V, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt ANJ, Valdés-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR. Signatures of mutational processes in human cancer. Nature. 2013;500:415–21.

6. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Rep. 2013;3:246–59.

7. Leemans CR, Braakhuis BJM, Brakenhoff RH. The molecular biology of head and neck cancer. Nat Rev Cancer. 2011;11:9–22.

8. Lawrence MS, Sougnez C, Lichtenstein L, et al. Comprehensive genomic characterization

of head and neck squamous cell carcinomas. Nature. 2015;517:576–82.

9. Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, Kryukov G V, Lawrence MS, Sougnez C, McKenna A, Shefler E, Ramos AH, Stojanov P, Carter SL, Voet D, Cortés ML, Auclair D, Berger MF, Saksena G, Guiducci C, Onofrio RC, Parkin M, Romkes M, Weissfeld JL, Seethala RR, Wang L, Rangel-Escareño C, Fernandez-Lopez JC, Hidalgo-Miranda A, Melendez-Zajgla J, Winckler W, Ardlie K, Gabriel SB, Meyerson M, Lander ES, Getz G, Golub TR, Garraway L a, Grandis JR. The mutational landscape of head and neck squamous cell carcinoma. Science. 2011;333:1157–60.

10. Collisson EA, Campbell JD, Brooks AN, at al. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511:543–50.

11. Campbell JD, Alexandrov A, Kim J, Wala J, Berger AH, Pedamallu CS, Shukla SA, Guo G, Brooks AN, Murray BA, Imielinski M, Hu X, Ling S et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat Genet. 2016;48:607–16.

12. Hammerman P, Lawrence M, Voet D et al. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489:519–25.

13. Polo V, Pasello G, Frega S, Favaretto A, Conte P, Bonanno L. Squamous cell carcinomas of the lung and of the head and neck : new insights on molecular characterization. Oncotarget. 2016;7:25050–63.

14. Westmaas JL, Newton CC, Stevens VL, Flanders WD, Gapstur SM, Jacobs EJ. Does a recent cancer diagnosis predict smoking cessation? An analysis from a large prospective US cohort. J Clin Oncol. 2015;33:1647–52.

15. Baser S, Shannon VR, Eapen GA, Jimenez CA, Onn A, Lin E MR. Smoking cessation after diagnosis of lung cancer is associated with a beneficial effect on performance status. Chest. 2005;130:1784–90.

16. Milholland B, Auton A, Suh Y, Vijg J. Age-related somatic mutations in the cancer genome. Oncotarget. 2015;6:24627–35.

17. Zhang W, Flemington EK, Zhang K. Mutant TP53 disrupts age-related accumulation

patterns of somatic mutations in multiple cancer types. Cancer Genet. 2016;209:376–80.

18. Meucci S, Keilholz U, Tinhofer I, Ebner O. Mutational load and mutational patterns in relation to age in head and neck cancer. Oncotarget. 2016;7:69188–99.

19. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, Kok CY, Jia M, De T, Teague JW, Stratton MR, McDermott U, Campbell PJ. COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015;43:D805–11.

20. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6:pl1.

21. Byrd R, Lu P, Nocedal J ZC. A Limited Memory Algorithm for Bound Constrained Optimization. SIAM J Sci Comput. 1995;16:1190–208.

22. R Core Team. R: A language and environment for statistical computing. R Found Stat Comput. 2014;

23. Huang DW, Sherman BT, Lempicki R a. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44–57.

24. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.

25. Best DJ R DE. Algorithm AS 71 : The Upper Tail Probabilities of Spearman's Rho. J R Stat Soc. 1975;24:377–9.

26. Storey JD. A direct approach to false discovery rates. J R Stat Soc. 2002;64:479–98.

27. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, Schinzel AC, Sandy P, Meylan E, Scholl C, Fröhling S, Chan EM, Sos ML, Michel K, Mermel C, Silver SJ, Weir BA, Reiling JH, Sheng Q, Gupta PB, Wadlow RC, Le H, Hoersch S, Wittner BS, Ramaswamy S, Livingston DM, Sabatini DM, Meyerson M, Thomas RK, Lander ES, Mesirov JP, Root DE, Gilliland DG, Jacks T, Hahn WC. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009;462:108–12.

28. de Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L, Jamal-Hanjani M, Shafi S, Murugaesu N, Rowan AJ, Grönroos E, Muhammad MA, Horswell S, Gerlinger M, Varela I, Jones D, Marshall J, Voet T, Van Loo P, Rassl DM, Rintoul RC, Janes SM, Lee S-M, Forster M, Ahmad T, Lawrence D, Falzon M, Capitanio A, Harkins TT, Lee CC, Tom W, Teefe E, Chen S-C, Begum S, Rabinowitz A, Phillimore B, Spencer-Dene B, Stamp G, Szallasi Z, Matthews N, Stewart A, Campbell P, Swanton C. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science (80-). 2014;346:251–6.

29. Zhang J, Fujimoto J, Zhang J, Wedge DC, Song X, Zhang J, Seth S, Chow C-W, Cao Y, Gumbs C, Gold KA, Kalhor N, Little L, Mahadeshwar H, Moran C, Protopopov A, Sun H, Tang J, Wu X, Ye Y, William WN, Lee JJ, Heymach J V, Hong WK, Swisher S, Wistuba II, Futreal PA. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. Science. 2014;346:256–9.

30. Fox EJ, Salk JJ, Loeb LA. Exploring the implications of distinct mutational signatures and mutation rates in aging and cancer. Genome Med. 2016;8:30.

31. Tomasetti C, Vogelstein B, Parmigiani G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proc Natl Acad Sci U S A. 2013;110:1999–2004.

32. Li Z, Zhang W, Chen Y, Guo W, Zhang J, Tang H, Xu Z, Zhang H, Tao Y, Wang F, Jiang Y, Sun FL, Mao Z. Impaired DNA double-strand break repair contributes to the age-associated rise of genomic instability in humans. 2016;1765–77.

33. Risques RA, Kennedy SR. Aging and the rise of somatic cancer-associated mutations in normal tissues. PLoS Genet. 2018;14:1–12.

34. Halvorsen AR, Silwal-Pandit L, Meza-Zepeda LA, Vodak D, Vu P, Sagerup C, Hovig E, Myklebost O, Børresen-Dale AL, Brustugun OT, Helland Å. TP53 mutation spectrum in smokers and never smoking lung cancer patients. Front Genet. 2016;7:1–10.

35. Gibbons DL, Byers LA, Kurie JM. Smoking, p53 mutation, and lung cancer. Mol Cancer Res. 2014;12:3–13.

36. Pj C, Regulation G, Molecular E, Ebi E-, Biology C, Project CG, Regulation G, Dow D,

Dundee S, Uk EH. Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers. bioRxiv. 2017;44.

37. Jiricny J. The multifaceted mismatch-repair system. Nat Rev Mol Cell Biol. 2006;7:335–46.

38. Meucci S, Keilholz U, Heim D, Klauschen F. Somatic genome alterations in relation to age in lung squamous cell carcinoma. 2018;9:32161–72.

39. Nasarre P, Potiron V, Drabkin H, Roche J. Guidance molecules in lung cancer. Cell Adh Migr. 2009;4:130–45.

40. Sprenger CC, Plymate SR RM. Aging-related alterations in the extracellular matrix modulate the microenvironment and influence tumor progression. Int J Cancer. 2010;127:2739–48.

41. Pickup MW, Mouw JK, Weaver VM. The extracellular matrix modulates the hallmarks of cancer. EMBO Rep. 2014;15:1243–53.

Eidesstattliche Versicherung

„Ich, Stefano Meucci, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: „Somatic genome alterations in relation to age in smoking related cancers" selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.

Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche in korrekter Zitierung (siehe „Uniform Requirements for Manuscripts (URM)" des ICMJE -*www.**icmje**.org*) kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) entsprechen den URM (s.o) und werden von mir verantwortet.

Meine Anteile an den ausgewählten Publikationen entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit den Betreuern, angegeben sind. Sämtliche Publikationen, die aus dieser Dissertation hervorgegangen sind und bei denen ich Autor bin, entsprechen den URM (s.o) und werden von mir verantwortet.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§156,161 des Strafgesetzbuches) sind mir bekannt und bewusst."

_____

Datum Unterschrift

Anteilserklärung an den erfolgten Publikationen

Stefano Meucci hatte folgenden Anteil an den folgenden Publikationen:

Publication 1:

Stefano Meucci, Ulrich Keilholz, Ingeborg Tinhofer, Olivia Ebner.

Mutational load and mutational patterns in relation to age in head and neck cancer.

Oncotarget, 2016.

Contribution in detail:

-Study conception Ideas, formulation of research question and statement of hypothesis.

-Development and design of methodology and creation of models.

-Contribution on the statistical analysis: From my statistical evaluation and gene ontology analysis, tables 1, S1, S4, S9 and S10 have been developed.

-Preparation, creation and presentation of the published work, specifically writing the initial draft and the data presentation.

Publication 2:

Stefano Meucci, Ulrich Keilholz, Daniel Heim, Frederick Klauschen and Stefano Cacciatore. **Somatic genome alterations in relation to age in lung squamous cell carcinoma**.

Oncotarget, 2018.

Contribution in detail:

-Study conception Ideas, formulation of research question and statement of hypothesis.

-Development and design of methodology and creation of models.

-Contribution on the statistical analysis: From my statistical evaluation and gene set enrichment analysis, tables 1, S1, S2, S8, S9 and S10 have been developed.

-Preparation, creation and presentation of the published work, specifically writing the initial draft and the data presentation.

Publication 3:

Stefano Meucci, Ulrich Keilholz, Daniel Heim, Frederick Klauschen and Stefano Cacciatore. **Somatic genome alterations in relation to age in lung adenocarcinoma**.

International Journal of Cancer, 2019.

Contribution in detail:

-Study conception Ideas, formulation of research question and statement of hypothesis.

-Development and design of methodology and creation of models.

-Contribution on the statistical analysis: From my statistical evaluation, tables 1, S1, S2, S3 and S4 have been developed.

-Preparation, creation and presentation of the published work, specifically writing the initial draft and the data presentation.

_____

Unterschrift, Datum und Stempel der betreuenden Hochschullehrer/der betreuenden Hochschullehrerinnen

_____

Unterschrift des Doktoranden/der Doktorandin

Research Paper: Gerotarget (Focus on Aging)

Mutational load and mutational patterns in relation to age in head and neck cancer

Stefano Meucci[1], Ulrich Keilholz[1], Ingeborg Tinhofer[2] and Oliva A. Ebner[1]

[1] Charité Comprehensive Cancer Center, Charité University Hospital, Charitéplatz, Berlin, Germany

[2] Department of Radiation Oncology and Radiotherapy, Charité University Hospital Berlin, Translational Radiation Oncology Research Laboratory, Charitéplatz, Berlin, Germany

***Correspondence to**: Ulrich Keilholz, **email:** ulrich.keilholz@charite.de*

*Correspondence to: Stefano Meucci, **email:** stefano.meucci@charite.de*

Keywords: head and neck squamous cell carcinoma; aging; somatic mutations; sequencing; genomics; Gerotarget

Received: *December 21, 2015* ***Accepted***: *July 23, 2016* ***Published***: *August 16, 2016*

ABSTRACT

Head and neck squamous cell carcinoma (HNSCC) is a cancer with well-defined tumor causes such as HPV infection, smoking and drinking. Using The Cancer Genome Atlas (TCGA) HNSCC cohort we systematically studied the mutational load as well as patterns related to patient age in HNSCC. To obtain a homogenous set we excluded all patients with HPV infection as well as wild type TP53. We found that the overall mutational load is higher in patients of old age. Through unsupervised hierarchical clustering, we detected distinct mutational clusters in very young as well as very old patients. In the group of old patients, we identified four enriched pathways ("Axon Guidance", "ECM-Receptor Interaction", "Focal Adhesion" and "Notch Signaling") that are only sporadically mutated in the other age groups. Our findings indicate that the four pathways regulate cell motility, tumor invasion and angiogenesis supposedly leading to less aggressive tumors in older age patients. Importantly, we did not see a strict pattern of genes always mutated in older age but rather an accumulation of mutations in the same pathways. Our study provides indications of age-dependent differences in mutational backgrounds of tumors that might be relevant for treatment approaches of HNSCCs patients.

INTRODUCTION

Head and neck squamous cell carcinomas (HNSCCs) affect 600,000 patients per year worldwide [1]. HNSCCs are characterized by phenotypic, etiological, biological and clinical heterogeneity and can originate from the paranasal sinuses, nasal cavity, oral cavity, pharynx and larynx. The major known risk factors of HNSCC are consumption of tobacco and alcohol, as well as human papillomaviruses (HPV) infection [2]. Multiple studies have elucidated the specific genetic background of HNSCC, establishing subclasses of tumors alongside HPV infection and/ or TP53 mutations. Due to the heterogeneity of study cohorts, the estimated percentages are relatively high variable. Overall, approximately 20% of HNSCCs contain transcriptionally active human papillomavirus (HPV+), and mainly TP53 wild type, which however is inactivated by the viral E6 and E7 oncogenes [3]. The incidence in HPV positive tumors, in oropharyngeal tumors, is exceeding 50% in current cohorts [4-6], and these tumors have been associated with a favorable clinical outcome [7, 8]. Approximately 80% of HNSCCs are HPV-negative (HPV-), the majority of them contain a mutation in TP53 and are characterized by many numerical genetic changes (high chromosome instability). In the remaining cases, characterized by a lower number of numerical genetic changes, p53 seems not to be inactivated [3].

Tumor characteristics vary between patients of different ages. Elderly patients are mainly diagnosed with a lower incidence of regional lymph node metastasis at diagnosis, often associated with a less aggressive tumor phenotype [9, 10]. Yet, whether older patients have similar or shorter survival is up for debate and shows controversial results in different studies [11-13]. Ageing related physiological alterations and the duration of active smoking should be simultaneously taken into account. As shown in lung cancer the onset of smoking is usually in

adolescence or young adult age, and smoking cessation is associated with the diagnosis of the malignancy [14,15]. In addition, several studies report HPV-negative HNSCCs mostly occurring in smokers [16-18]. Therefore the age at diagnosis of HNSCC patients is very closely correlated to the duration of smoking.

While some differences in tumor behavior in older patients have been recorded, no study so far systematically explored the relationship between genetic tumor background and age in HNSCC. A recent study performed on the extensive data set available on The Cancer Genome Atlas (TCGA) portal showed the age-related accumulation of somatic mutations in diverse human tissues [19]. However, it is still an open question whether differences in mutations between the ages are random coincidence or follow distinct patterns. Therefore, this study aims to explore the specific age-mutation relationship in tumors of HNSCC patients to determine if age-related genetic parameters have to be considered in the disease prognosis and treatment decision.

To investigate a possible connection between patterns and frequency of genetic mutations and patient age, we used the recently published TCGA study on HNSCCs of 279 patients.

HPV infection shows an age bias as a relevant impact on the mutational background of the tumor. In previous studies the mutation rate of HPV-positive tumors was lower than that found in HPV-negative HNSCC, consistent with recent epidemiologic studies that establish biological differences between HPV-positive and HPV-negative disease [20]. The major biologic difference between HPV-positive and - negative tumors, however, concerns p53, which in its role as a guardian of the genome influences multiple genes. In general, patients with HPV-positive tumors have non-mutated TP53, however, HPV itself inhibits p53 function. Conversely, HPV-negative tumors frequently harbour TP53 mutations.

Thus, in order to select a homogenous patient cohort for investigating possible age-related differences, we only considered the 203 patients from the TCGA-cohort which were HPV negative and carried a TP53 mutation. This homogenous subset of patients allowed for a systematic study of the age influence on mutation load and spectrum without introducing a heterogeneity caused by HPV or p53. Of course, it would have been equally interesting to study the other subclasses of HNSCCs. However, since only few patients belonged into these subsets, a statistical sound investigation would not have been possible within the current TCGA cohort.

RESULTS

Patients selection and cohort features

To create a homogenous set of tumors we used a subset of the TCGA cohort (Figure 1). The original patient cohort was selected by two criteria, the first (i) excluded 36 HPV(+) patients (HPV classification according to the TCGA publication). As expected, the HPV-positive phenotype was strongly associated with the oropharyngeal site and the patient age mean was lower than for the HPV-negative patients. The second selection criteria (ii) excluded 40 TP53 wild type patients, which left the final selected cohort of 203 HPV-negative/TP53-mutated patients. The original TCGA data set showed a higher rate (86%) of TP53 mutations among HPV-negative samples than have been previously reported, while only 1 out of 36 HPV-positive cases had a non-synonymous TP53 mutation. Our selection rendered 203 patients with a total amount of 29.860 single nucleotide polymorphisms (SNPs) distributed on 11.489 genes.

Statistical hypothesis tests showed that the patient selection does not lead to a biased distribution of patients features such as age, smoking and alcohol consumption. Whereas, the hypergeometric test performed on the tumor localization distribution showed an enrichment in Larynx tumor ($p < 0.05$) and a depletion in Oropharynx tumor ($p < 0.05$) in the selected cohort (Table S1).

A comparison of the original TCGA cohort and our subset can be found in Table 1.

Furthermore, the results of the investigation of TNM and overall staging at diagnosis in relation to the age showed a smaller size of the primary tumor, a decrease of lymph node metastasis and a higher percentage of "localized cancers" and "first stage of locally advanced cancers" in old patients (Table S1). The youngest patients group showed the same statistics as the oldest, however the results are not comparable due to the significant difference between the two age group ranges.

Increase of mutation frequency with age

When looking at the mutations in each patient, a wide range in numbers along with a single extreme value in one patient (age 69) was observed (Figure S1).

A significant rise ($p < 0.01$) in the average number of mutations in different genes was observed with increasing age (Figure 2), which was also true if the patient with the extreme number of mutations was removed. Interestingly, we repeated the mutation load analysis excluding silent mutations and multiple mutations in one gene, to have an additional overview of disruptive mutations with increasing age. Although the number of mutations in each patient decrease drastically, the same significant p-value

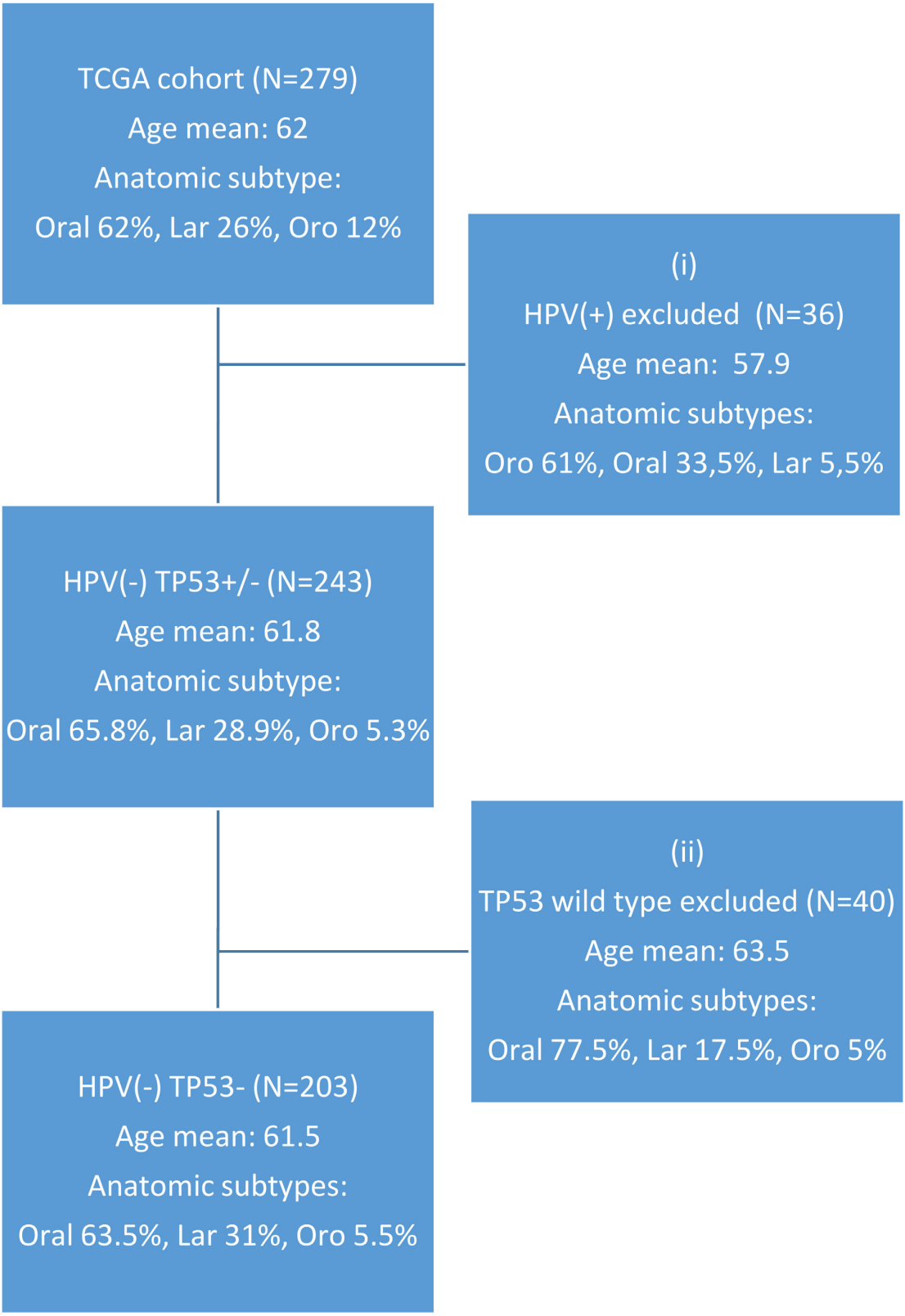

Figure 1: CONSORT diagram of original and selected patient cohort.

Table 1: Comparison of TCGA cohort and selected patients

Characteristics	TCGA (279 patients)	Selected cohort (203 Patients)
Age (years)		
Median	61	62
Range	19-90	19-87
Smoking history		
Yes	220	159
No	52	37
Unknown	7	7
Primary tumor location		
Oral Cavity	172	129
Larynx	72	64
Oropharynx	33	9
Hypopharynx	2	1
HPV/ p53 status		
HPV + / p53 +	35	0
HPV + / p53 -	1	0
HPV - / p53 +	40	0
HPV - / p53 -	203	203

was detected (Figure S2, list of mutation frequencies in Table S2). Thus, the number of mutated genes found in a tumor was higher in older patients than in younger patients, designating a quantitative difference in mutational load between patient ages irrespective of whether this signified a pure stochastic increase or the accumulation of disease relevant mutations. Interestingly, there were only five genes with recurrent mutations, three of which are well known players (TP53, CDKN2A, PIK3CA) and two are pseudogenes (RPSAP58, WASH3P).

In order to investigate the influence of patients features such the tumor localization and the smoking consumption on the increasing of the average number of mutations with age, we repeated the regression analysis for each subset. Although it´s necessary to consider the different number of patients in each subset and the difficulty on the acquisition of smoking habits (possible false-positive/-negative), both smokers and non-smokers subsets showed a significant correlation (linear regression as well as Spearman's rank correlation analysis), therefore the smoke history surprisingly seems not to influence the age-related mutational load of our sub-cohort. Whereas, the mutational load of our sub-cohort as well as the original cohort are mostly influenced by the location of the cancers, only oral cavity tumors showed a significant correlation with age (Table S1).

Relationship of mutational patterns and age

As a next step, we wanted to investigate whether this increase was a mere accumulation of random mutations or whether we could find age-specific patterns of mutated genes. Unsupervised hierarchical clustering based on gene mutation frequencies was performed for patients pooled into age groups of decades aside two groups of the very young (pooled age 19-40) and very old (pooled ages 81 to 87). A minimum difference of 0.15 in mutation frequency between at least two age groups was considered for cluster analysis (Figure S3). This cut-off allowed for a comprehensive set of mutational patterns within the specific age groups including less frequently mutated genes.

The results showed two relevant clusters of frequently mutated genes, in particular, 39 genes in very young patients and 108 genes in very old patients (Table S3 with cut clusters A and B from Figure 3, unmarked heatmap Figure S3). Both the young as well as the old group, each consisting of 11 patients, displayed prominent differences to middle age groups. The genes of the two specific clusters were tested for KEGG pathway enrichments using the online David Gene Ontology tool [21]. The old age gene cluster showed six statistically significant enriched pathways, two of which, the "Axon Guidance" ($p < 0.003$) and "ECM-Receptor Interaction" ($p < 0.04$) pathways, stood out due to their role in angiogenesis processes and cell / tissue architecture maintenance respectively (Table S4).

Six "Axon Guidance" genes (SEMA5A, DCC, PLXNB2, UNC5D, ITGB1, EPHA2) and four of the "ECM-Receptor Interaction" pathway (LAMA2, LAMA4, TNC, ITGB1) were significantly enriched. In contrast, no enriched pathways were found for the young age gene cluster. To see if the very old and very young patients were homogenous groups or comprised of smaller subgroups we looked at both in more detail. Further division of the

young age group into two fractions comprising ages 19 to 35 (7 patients) and 36 to 40 (4 patients) did yield an enrichment of the TGFB pathway for the latter age group (Figure S3, Table S4). This enrichment, however, was mainly based on the mutations of the enriched genes in one single patient (age 39). In addition, TGFβRII and TGFb1, two main players of the TGFB pathway and responsible for more aggressive tumors in HNSCCs were not mutated in the young ages at all [22]. We therefore did not interpret this finding as an age-specific enrichment.

Old age specific clusters and pathway enrichment

The old age group, on the other hand, showed distinct clusters and enrichment when split into smaller fractions of ages 81 to 85 (7 patients) and 86 to 87 (4 patients, all age 87).

Figure 4 shows a major cluster of 542 genes in the "87" age class (Cluster D) and two gene clusters in the "81-85" age group (Cluster A and B, unmarked heatmap Figure S5). The latter consisted of 47 specific mutations and 59 genes overlapping with the "87" age group, indicating a common genetic background of tumors in elderly patients (Table S3).

Note that we included six genes from the old ages clusters (CDKN2A, CSMD3, FAT1, NOTCH1, PIK3CA, TTN) that did make the 0.15 cutoff, but were in fact frequently mutated in most other ages as well. However, since the enrichment analysis showed significant p-values for both the set with as well as without the six genes we kept them in our gene set.

The "87" age group showed six significantly enriched pathways among which are again "ECM-Receptor Interaction" ($p < 0.006$, 11 genes) and "Axon Guidance" ($p < 0.006$, 13 genes), and in addition "Notch Signalling" ($p < 0.007$, 7 genes) as well as "Focal Adhesion" ($p < 0.05$, 15 genes) (Table S4). The 47 genes specifically mutated in the "81-85" group did not yield any enrichment. However, when combining all highly mutated genes in this group, "Axon Guidance" was enriched as well yet with a slightly elevated p-value ($p = 0.056$, 4 genes) (Table S4). The overlap of the two old age groups is only three genes (PLXNB2, SEMA5A, UNC5D), yet genes of the same families (such as Ephrins, Slits and Rho-associated protein kinases) were mutated in both age groups. Overall, the old age groups only revealed a certain degree of homogeneity, with the very old patients (age 87) showing distinct pathway enrichment. However, the high number of overlapping genes as well as the commonly

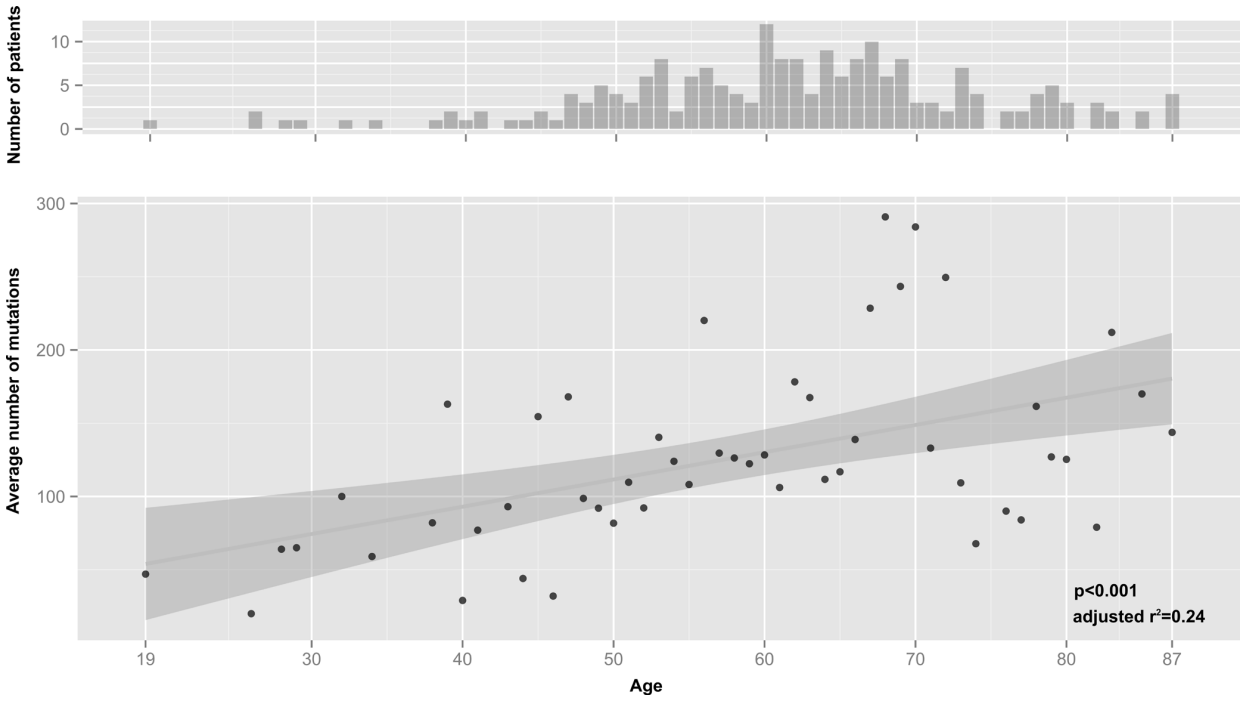

Figure 2: Correlation of average mutations and age considering one mutation per gene. Upper graph number of patients in the respective age group. Lower graph, average number of mutations for all patients in the respective age groups. Linear regression analysis was done using F-statistics and shows a significant increase in mutations in older patients ($p = 0.000161$, adjusted $r^2 = 0.24$). Grey area around regression line indicates 95% confidence interval. Regression without the patient having an extreme number of mutations (TCGA-D6-6516, see Figure S1) also yields a significant connection between average mutations and age $p = 0.000291$, $r^2 = 0.22$ (data not shown).

enriched pathways fostered the idea of specific mutational trends happening at the old age rather than one specific mutational pattern.

Mapping of "Axon Guidance" pathway genes to all age groups

We mapped all genes of the "Axon Guidance" pathway (according to the KEGG database) to our list of mutation frequencies to see if the pathway is affected in other age groups as well without being particularly enriched (Figure 5, Table S5). From the total 127 genes of the "Axon Guidance" pathway 99 mapped to our data. As the heatmap showed, mutations in this pathway were present in other ages as well, yet at a lower frequency than in the two old age groups. Again, the trend of increased pathway alterations towards old ages was visible. In addition to genes mutated at a high frequency we found a group of 10 genes mutated at lower frequencies in the "81-85" group (Figure 5).

Pathway enrichment upon fusion of clusters

As the mapping of all genes to the "Axon Guidance" pathway had shown, we were missing mutated genes of a pathway that did not make the 0.15 frequency cutoff. However, since our results suggested that mutations in old patients did not necessarily follow a strict common pattern and different genes were affected, we did not expect all mutations in a pathway to occur at high frequencies. At the same time we had seen a common trend for both old age groups with impact of mutations on similar gene families. Thus we decided to merge the highly mutated genes of both separate age groups for a more comprehensive picture of affected pathways.

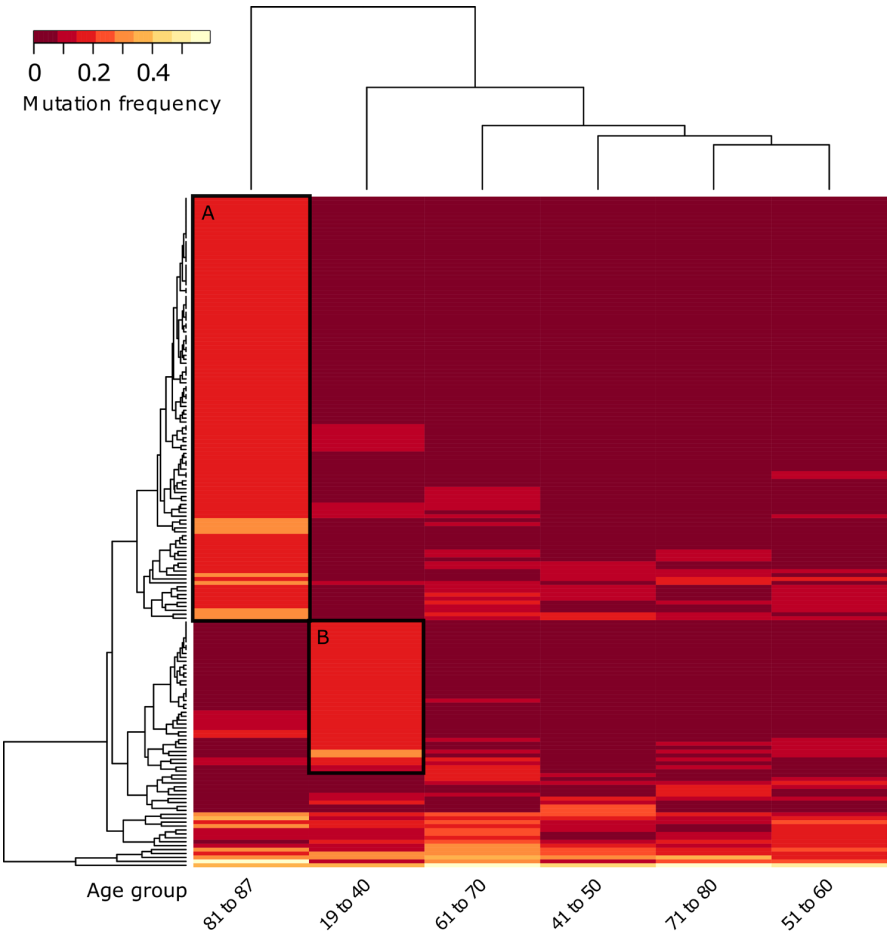

Figure 3: Unsupervised hierarchical clustering of gene mutation frequencies of specific age groups with pooled young and old ages. Patients were grouped into age groups of very young (ages 19-40) and very old (ages 81-87) with 10-year bins in between, then clustered according to the mutations frequencies of the quantified genes. Only genes with a minimum frequency difference of 0.15 between at least two of the age groups are displayed. Black boxes indicate extracted genes for **A.**, age group 81 to 87 and **B.**, age group 19 to 40 (for genes see Table S3).

Pathway enrichment analysis was performed on a consistent group of 589 mutated genes, many of which were hidden by the 0.15 clustering threshold when calculating the mutation frequency of the two old ages groups together, rather than separately.

The result showed 10 significantly enriched pathways (Table S4), two of which had been found enriched before, namely "Axon Guidance" ($p < 0.004$, 14 genes) and "ECM-Receptor Interaction" ($p < 0.001$, 13 genes). In addition, "Notch Signaling" ($p < 0.002$, 8 genes) and "Focal Adhesion" ($p < 0.009$, 18 genes) were selected for detailed analysis due to their important role on cell growth, cell / tissue architecture and cell motility. When mapping all pathway genes to our data, we saw in all three the same pattern as in the "Axon Guidance" pathway of sporadic mutations that occur in all ages, yet the frequency of mutated genes was much higher in the old age groups of 81 to 87 (Figure S4-S6, Table S6-S8).

Overall, we identified 24 genes mutated in the "Axon Guidance" pathway among the 81-87 group of patients. 14 genes resulted from the pathway enrichment analysis and 10 genes were additionally revealed by the pathway heatmap clusters (Table S9). Even though 79% (19 out of 24) carried a single mutation showing the tumor heterogeneity, in the end 82% of the old patient group "81-87" (9 out of 11) reported a disruption in the Axon Guidance pathway. Furthermore, 82% (9 out of 11) of the patients showed mutations in "ECM-Receptor Interaction" pathway and all patients reported mutations in "Focal Adhesion" pathway. In particular we identified 21 and 34 significantly mutated genes respectively, which overlapped in 20 genes (Table S9). Interestingly we found several mutated genes of the same protein families, six laminins (LAMA1, LAMA2, LAMA3, LAMA4, LAMC1, LAMB1), three integrins (ITGA11, ITGA2, ITGB1) and eight collagen mutate genes (COL1A1, COL11A1,

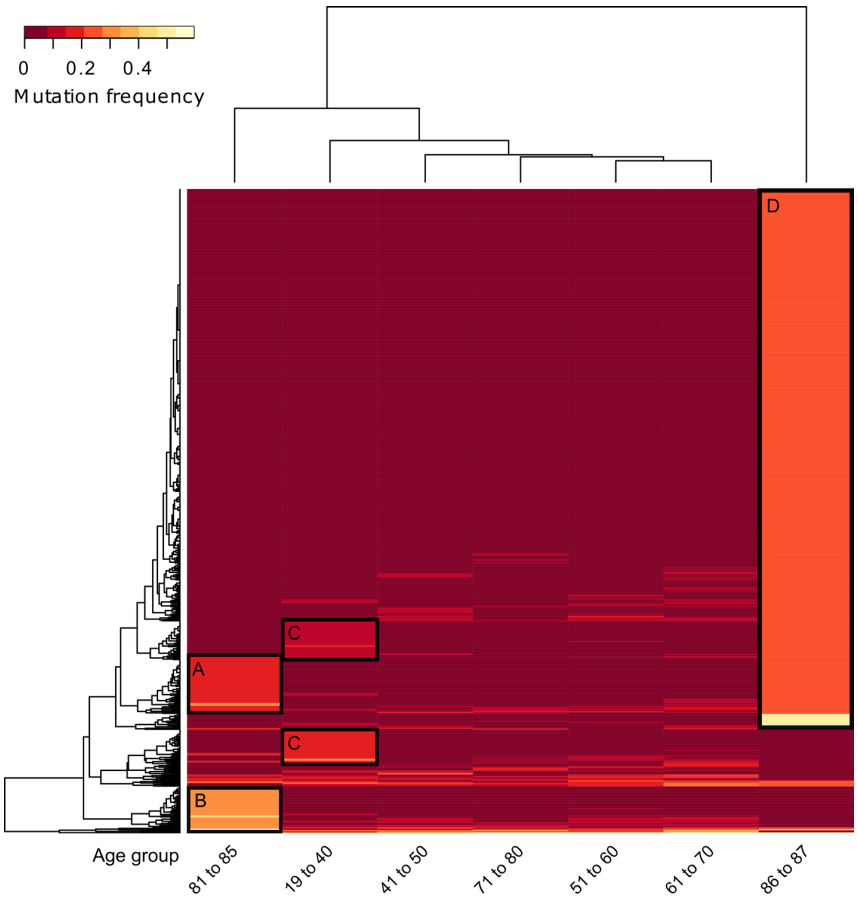

Figure 4: Unsupervised hierarchical clustering of gene mutation frequencies of specific age groups with separate old ages. Patients were grouped into age groups of very young (ages 19-40), decades in-between and two separate old ages groups (ages 81-85 and 87), then clustered according to the mutations frequencies of the quantified genes. Only genes with a minimum frequency difference of 0.15 between two of the age groups are displayed. Black boxes indicate extracted genes for **A.**, age group 81 to 85 upper cluster (shared with ages 86 to 87), **B.**, age group 81 to 85 lower cluster (separate genes from ages 86 to 87), **C.**, two clusters of ages 19 to 40 and **D.**, ages 86 to 87 (for genes see Table S3).

COL29A1, COL6A3, COL27A1, COL11A1, COL4A5, COL4A4), crucial in ECM-signaling processes and focal adhesions, as well as six ephrin genes (EPHA2, EPHA3, EPHA5, EPHA6, EPHA8, EPHB6), which have a central role in "Axon Guidance" signalling processes. Lastly, 64% of the patients (7 out of 11) displayed mutations in "Notch-Signaling" pathway. We found 10 mutated genes, three of which were NOTCH genes (NOTCH1, NOTCH3, NOTCH4).

Altogether, we only saw a few genes recurrently mutated and thus no age specific mutational pattern. However, even though not always the same genes were affected, the mutations accumulated in the same pathways, indicating a common trend in elderly patient to have similar functions of the cell changed by mutations.

DISCUSSION

In the present study, we evaluated the somatic DNA single nucleotide polymorphisms (SNPs) landscape of a selected HNSCC patient cohort in relation to ageing processes. The average number of mutations for each patient showed a significant rise ($p < 0.01$) with increasing age. Multiple factors participate to the accumulation of genetic events in the elderly such as prolonged tobacco exposure and ageing-related genomic instability. To provide better insight into the underlying mechanisms, we therefore investigated whether a mere accumulation of random mutations or distinctive mutational patterns are related to patient age.

Unsupervised clustering of the selected cohort showed distinct clusters of genes expressed with a higher mutation frequency for the very young and very old ages. While the young age group did not reveal pathway enrichment, the old age group showed enrichment of KEGG pathways, including "ECM-Receptor Interaction " and "Axon Guidance". Further division of the old age group into ages 81 to 85 and age 87 showed several genes of the same pathways shared by the two groups. Which when clustered together showed ten enriched pathways. Besides "Axon Guidance" and "ECM-Receptor Interaction", "Notch-Signaling" and "Focal Adhesion" were of special interest to us. When mapping all genes of the respective pathways to our data, we saw mutations in all patients, however no evident clusters were present other than for ages 81 to 87. These results indicate that while the specific mutational patterns might only exist in very subtle ways due to the heterogeneity of the tumors,

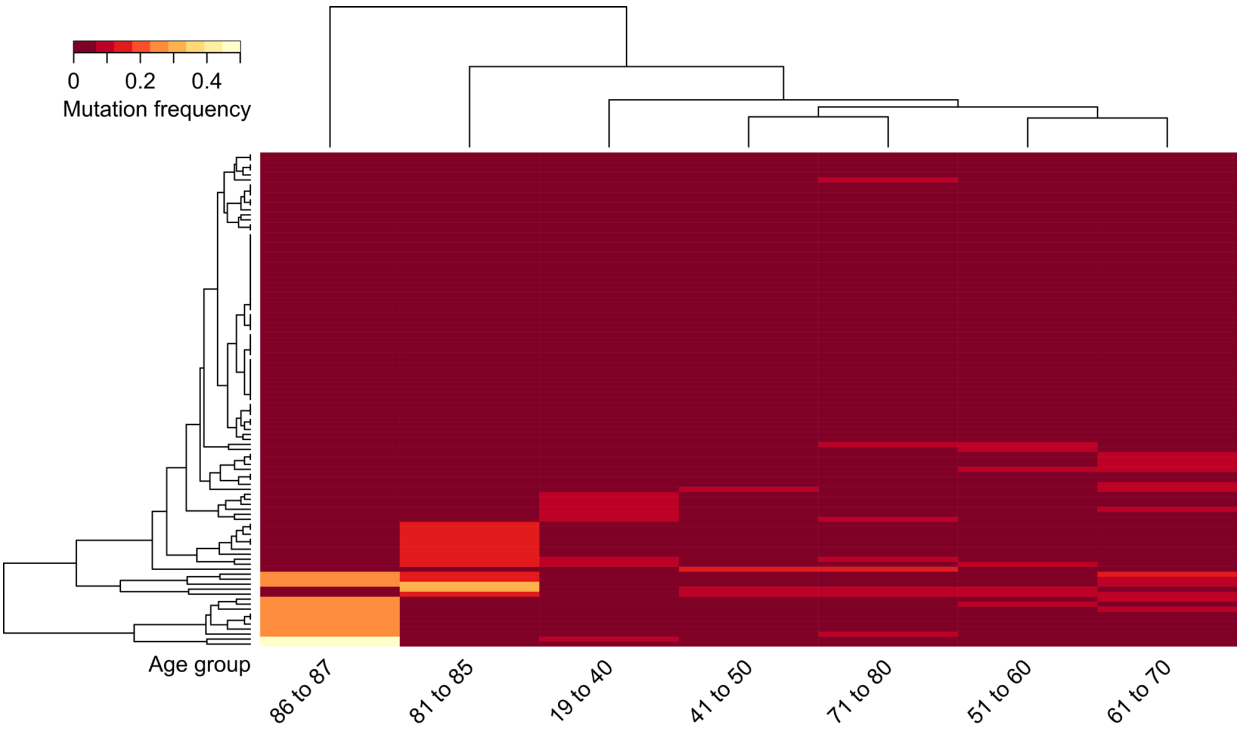

Figure 5: Unsupervised hierarchical clustering of mutation frequencies of genes involved in the "Axon Guidance" pathway (according to the KEGG database). Patients were grouped into age groups of very young (ages 19-40), decades in-between and two separate old ages groups (ages 81-85 and 86 to 87), then clustered according to the mutation frequencies of all quantified genes of the "Axon Guidance" pathway. Genes that did not make the original cut off of 0.15 difference in mutation frequency but are found mutated in this pathway in age group 81-85 are LRRC4C, DPYSL2, EPHA8, LIMK1, NCK1, NGEF, RAC1, ROBO2, ROCK1 and SLIT2.

we could see an increase in mutations in distinct pathways over the ages with a peak in the very old fraction. Next, we looked at the involved genes in more detail, starting with the "Axon Guidance" pathway.

Although "Axon Guidance" represents a key stage in the formation of neuronal networks, recent studies also linked this pathway to regulation of angiogenesis processes (endothelial cell migration, proliferation and vessel formation). Involved proteins are Netrins, Slit proteins, Semaphorins, Ephrins and their cognate receptors (e.g. UNC5, ROBO1-4), all of which are frequently mutated in the elderly groups [24, 25]. The expression of Eph receptors, five of which were mutated in our data, is frequently elevated in different types of malignant tumors possibly resulting in increased cellular motility, tumor cell invasion and metastasis [26, 27]. One of these, EPHA2, is overexpressed in squamous cell carcinoma of oral tongue [27] and was also protruding in our mutational analysis. Moreover, studies have proven that blocking EphA receptor signaling decreases tumor vascular density, volume and cell proliferation *in vivo* [28-31]. Since we found many of these genes to be mutated in the old age group, we postulated a correlation between "Axon Guidance" aberrations and the HNSCC features of elderly patients which may result in decreased cellular motility and tumor cell invasion.

For the "ECM-Receptor Interaction" pathway, we identified 21 frequently mutated genes, including six laminins, three integrins and eight collagens. These molecules are structurally and functionally involved in interactions at the extracellular matrix (ECM) which lead to a direct or indirect control of cellular activities such as cell migration, differentiation, proliferation, and apoptosis [32, 33]. A re-expression of the laminin α2 and α4 chains, which were significantly mutated in our old patients group, could be shown in adult hyperproliferative, dysplastic and carcinomatous lesions [31, 32]. Several studies showed Laminin 332 (LAMA3, LAMB3 and LAMC2) to be highly expressed in HNSCC and foster tumor invasiveness, an effect that is reversed when the laminins are repressed by microRNA-29s [36-40]. Several other integrins composed by the Integrin β1 chain, highly mutated in patients of old age, have been identified as crucial for tumor cell invasion and angiogenesis [38, 39]. Altogether, the many mutations in this pathway again suggested a decreased cell migration and tumor invasion in old patients.

The "Focal Adhesion" pathway was significantly enriched in the 81-87 age group as well. We identified 34 genes, 20 of which overlap with the "ECM-Receptor Interaction" pathway, in particular laminins and integrins. The latter regulate kinases, such as the focal adhesion kinase (FAK), which is crucial for the attachment and signal transduction between cells and the ECM [43]. Several studies demonstrated that FAK disruption caused decreased cell attachment and motility while FAK overexpression increased cell invasion in HNSCC [41, 42]. Therefore mutations on upstream proteins like laminins and integrins again suggested a decreased cellular motility in HNSCC.

NOTCH signalling is a highly conserved pathway that plays distinct roles during tissue homeostasis, proliferation and apoptosis [46]. Even though NOTCH1 is one of the most frequently mutated genes in HNSCC, there are contradictory studies about its influence on tumor development [20]. Loss-of-function mutations in the NOTCH1 gene have been detected in a significant proportion of patients [47]. On the other side it is known that NOTCH activation can enhance proliferation, inhibit apoptosis and promote angiogenesis [45, 46]. Thus, while we reported high mutation frequencies for 10 genes of the notch pathway (inter alia NOTCH1, NOTCH3, NOTCH4), we couldn´t make assumptions about the exact impact of this finding.

Concluding, we found a proportional increase of the mutation frequency rate in relation to the age of the HNSCC patients. This increase, however, did not follow distinct mutational patterns but rather an accumulation of mutations in specific pathways.

The results of this pathway analysis suggested a reduced tumor invasiveness and metastasis in older patients, which was underlined by the tumor staging at diagnosis. This distinct mutational background might be relevant for treatment approaches decisions and should therefore be taken under closer consideration in future studies.

MATERIALS AND METHODS

Acquisition and processing of data

The clinical data of the TCGA patient set was derived via download from the TCGA data matrix (https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm). Somatic mutations of HNSCCs from the TCGA study was derived by download from the cBio Portal [18, 19]. The somatic mutations of the 279 HNSCC patients detected by exome sequencing within the TCGA project [16] were merged with the clinical data from TCGA to combine genomic mutations with the age information of all patients. Entries without official gene names were removed. Since we were interested in patients with a HPV negative genomic background, we excluded all patients with a positive HPV status according to the standards of the TCGA publication. As HPV determination has certain limitations [16], it remained uncertain, whether all tumors classified as HPV negative were truly negative, or whether diagnostic sensitivity may have misclassified some cases. Therefore, we selected for patients with at least one mutation in TP53 for our study, as TP53 mutations are generally not found in HPV positive tumors. For a

first evaluation of the mutational rate related to age we considered all mutations, including silent mutations as well as multiple mutations in one gene. For all subsequent analysis, silent mutations were removed and multiple mutations in the same gene were only considered once per patient to prevent statistical overestimations.

Statistical analysis

Mutation frequency and age of patients

The primary statistical hypothesis was, that there would be an increase in the number of mutations with increasing age. The relations between total average mutation frequencies and patient age groups (in years) were calculated by linear regression using F-statistics. A two-sided p-value of below 0.05 (Pr(>|t|)) was considered significant. The frequency of mutation of a specific gene was calculated using the sum of mutations in a specific age group divided by the number of patients in the respective age group. Additional Spearman's rank correlation analysis was performed to identify the genes whose mutation frequency correlates with the age (Table S10).

All data analysis was done using R [52] unless stated otherwise.

Recurrent genes

To check for recurrent genes we only considered genes mutated in at least five patients with the same starting and end position of the mutation.

Clustering and pathway enrichment analysis

Unsupervised hierarchical clustering based on gene mutation frequencies was performed for different age groups. Genes were clustered according to Euclidian distance measure using the method "complete". Genes from age group specific clusters were extracted and tested using the online David Gene Ontology tool [21]. The full list of identified genes was used as background for enrichment calculation. In reverse, all genes of an enriched KEGG pathway were mapped to our list of mutation frequencies to investigate the number and distribution of mutated genes in the respective pathway within all ages of our dataset.

ACKNOWLEDGMENTS / FUNDING

This work was supported by the Focus Area Dynage (www.fu-berlin.de/dynage) and the Charité Comprehensive Cancer Center, Berlin.

CONFLICTS OF INTEREST

There is no conflict of interest that I should disclose.

REFERENCES

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer. 2010; 127(12): 2893–2917. doi:10.1002/ijc.25516.
2. Keck MK, Zuo Z, Khattri A, Stricker TP, Brown CD, Imanguli M, Rieke D, Endhardt K, Fang P, Bragelmann J, DeBoer R, El-Dinali M, Aktolga S, et al. Integrative analysis of head and neck cancer identifies two biologically distinct HPV and three non-HPV subtypes. Clin Cancer Res. 2014; 21(4): 870–881. doi:10.1158/1078-0432.CCR-14-2481.
3. Leemans CR, Braakhuis BJM, Brakenhoff RH. The molecular biology of head and neck cancer. Nat Rev Cancer. 2011; 11(1): 9–22. doi:10.1038/nrc2982.
4. Fakhry C, Cohen E. The rise of HPV-positive oropharyngeal cancers in the United States. Cancer Prev Res. 2015; 8(1): 9–11. doi:10.1158/1940-6207.CAPR-14-0425.
5. Field N, Lechner M. Exploring the implications of HPV infection for head and neck cancer. Sex Transm Infect. 2015; 0(0): 2014–2016. doi:10.1136/sextrans-2014-051808.
6. Chaturvedi AK, Engels EA, Pfeiffer RM, Hernandez BY, Xiao W, Kim E, Jiang B, Goodman MT, Sibug-Saber M, Cozen W, Liu L, Lynch CF, Wentzensen N, et al. Human papillomavirus and rising oropharyngeal cancer incidence in the United States. J Clin Oncol. 2011; 29(32): 4294–4301. doi:10.1200/JCO.2011.36.4596.
7. Gugić J, Strojan P. Squamous cell carcinoma of the head and neck in the elderly. Reports Pract Oncol Radiother. 2012; 8: 16–25. doi:10.1016/j.rpor.2012.07.014.
8. El-Mofty SK, Lu DW. Prevalence of human papillomavirus type 16 DNA in squamous cell carcinoma of the palatine tonsil, and not the oral cavity, in young patients: a distinct clinicopathologic and molecular disease entity. Am J Surg Pathol. 2003; 27(11): 1463–1470. doi:10.1097/00000478-200311000-00010.
9. Koch WM, Patel H, Brennan J, Boyle JO, Sidransky D. Squamous cell carcinoma of the head and neck in the elderly. Arch Otolaryngol Head Neck Surg. 1995; 121(3): 262–265. doi:10.1016/j.bjoms.2005.05.011.
10. Ershler WB. Why tumors grow more slowly in old people. J Natl Cancer Inst. 1986; 77(4): 837–839.
11. Mountzios G. Optimal management of the elderly patient with head and neck cancer: Issues regarding surgery, irradiation and chemotherapy. World J Clin Oncol. 2015; 6(1): 7–15. doi:10.5306/wjco.v6.i1.7.
12. Italiano A, Ortholan C, Dassonville O, Poissonnet G, Thariat J, Benezery K, Vallicioni J, Peyrade F, Marcy PY, Bensadoun RJ. Head and neck squamous cell carcinoma in patients aged > or = 80 years: patterns of care and survival. Cancer. 2008; 113(11): 3160–3168. doi:10.1002/cncr.23931.

13. Vanderwalde NA, Fleming M, Weiss J, Chera BS. Treatment of older patients with head and neck cancer: a review. Oncologist. 2013; 18(5): 568–78. doi:10.1634/theoncologist.2012-0427.
14. Westmaas JL, Newton CC, Stevens VL, Flanders WD, Gapstur SM, Jacobs EJ. Does a recent cancer diagnosis predict smoking cessation? An analysis from a large prospective US cohort. J Clin Oncol. 2015; doi:10.1200/JCO.2014.58.3088.
15. Baser S, Shannon VR, Eapen GA, Jimenez CA, Onn A, Lin E, Morice RC. Smoking cessation after diagnosis of lung cancer is associated with a beneficial effect on performance status. CHEST J. 2005; 130(6). doi:10.1378/chest.130.6.1784.
16. Lawrence MS, Sougnez C, Lichtenstein L, Cibulskis K, Lander E, Gabriel SB, Getz G, Ally A, Balasundaram M, Birol I, Bowlby R, Brooks D, Butterfield YSN, et al. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015; 517(7536): 576–582. doi:10.1038/nature14129.
17. Ang K, Harris J, Wheeler R. Human papillomavirus and survival of patients with oropharyngeal cancer. N Engl J Med 2010; 363(1): 24–35. doi:10.1056/NEJMoa0912217.
18. Maruyama H, Yasui T, Ishikawa-Fujiwara T, Morii E, Yamamoto Y, Yoshii T, Takenaka Y, Nakahara S, Todo T, Hongyo T, Inohara H. Human papillomavirus and p53 mutations in head and neck squamous cell carcinoma among Japanese population. Cancer Sci. 2014; 105(4): 409–417. doi:10.1111/cas.12369.
19. Milholland B, Auton A, Suh Y, Vijg J. Age-related somatic mutations in the cancer genome. Oncotarget. 2015; 6(28): 24627-35. doi: 10.18632/oncotarget.5685.
20. Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, Kryukov GV, Lawrence MS, Sougnez C, McKenna A, Shefler E, Ramos AH, Stojanov P, et al. The mutational landscape of head and neck squamous cell carcinoma. Science. 2011; 333(6046): 1157–1160. doi:10.1126/science.1208130.
21. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4(1): 44–57. doi:10.1038/nprot.2008.211.
22. White RA, Malkoski SP, Wang X-J. TGFβ signaling in head and neck squamous cell carcinoma. Oncogene. 2010; 29(40): 5437–5446. doi:10.1038/onc.2010.306.
23. Pircher A, Wellbrock J, Fiedler W, Heidegger I, Gunsilius E, Hilbe W. New antiangiogenic strategies beyond inhibition of vascular endothelial growth factor with special focus on axon guidance molecules. Oncol. 2014; 86(1): 46–52. doi:10.1159/000356871.
24. Kuijper S, Turner CJ, Adams RH. Regulation of angiogenesis by Eph-Ephrin interactions. Trends Cardiovasc Med. 2007; 17(5): 145–151. doi:10.1016/j.tcm.2007.03.003.
25. Chen J, Zhuang G, Frieden L, Debinski W. Eph receptors and Ephrins in cancer. Changes. 2012; 29(6): 997–1003. doi:10.1016/j.biotechadv.2011.08.021.
26. Dodelet VC, Pasquale EB. Eph receptors and Ephrin ligands: embryogenesis to tumorigenesis. Oncogene. 2000; 19(49): 5614–5619. doi:10.1038/sj.onc.1203856.
27. Shao Z, Zhang WF, Chen XM, Shang ZJ. Expression of EphA2 and VEGF in squamous cell carcinoma of the tongue: Correlation with the angiogenesis and clinical outcome. Oral Oncol. 2008; 44(12): 1110–1117. doi:10.1016/j.oraloncology.2008.01.018.
28. Dobrzanski P, Hunter K, Jones-Bolin S, Chang H, Robinson C, Pritchard S, Zhao H, Ruggeri B. Antiangiogenic and antitumor efficacy of EphA2 receptor antagonist. 2004;(14): 910–919. doi: 10.1158/0008-5472.CAN-3430-2
29. Brantley DM, Cheng N, Thompson EJ, Lin Q, Brekken RA, Thorpe PE, Muraoka RS, Cerretti DP, Pozzi A, Jackson D, Lin C, Chen J. Soluble Eph A receptors inhibit tumor angiogenesis and progression in vivo. Oncogene. 2002; 21(46): 7011–7026. doi:10.1038/sj.onc.1205679.
30. Cheng N, Brantley D, Fang W Bin, Liu H, Fanslow W, Cerretti DP, Bussell KN, Reith AD, Jackson D, Chen J. Inhibition of VEGF-dependent multistage carcinogenesis by soluble EphA receptors. Neoplasia. 2003; 5(5): 445–456. doi:10.1016/S1476-5586(03)80047-7.
31. Mosch B, Reissenweber B, Neuber C, Pietzsch J. Eph receptors and Ephrin ligands: Important players in angiogenesis and tumor angiogenesis. J Oncol. 2010; 2010:135285. doi:10.1155/2010/135285.
32. Frantz C, Stewart KM, Weaver VM. The extracellular matrix at a glance. J Cell Sci. 2010; 123(Pt 24): 4195–4200. doi:10.1242/jcs.023820.
33. Sprenger CC, Plymate SR, Reed MJ. Aging-related alterations in the extracellular matrix modulate the microenvironment and influence tumor progression. Changes. 2012; 29(6): 997–1003. doi:10.1016/j.biotechadv.2011.08.021.Secreted.
34. Kosmehl H, Berndt A, Strassburger S, Borsi L, Rousselle P, Mandel U, Hyckel P, Zardi L, Katenkamp D. Distribution of laminin and fibronectin isoforms in oral mucosa and oral squamous cell carcinoma. Br J Cancer. 1999; 81(6): 1071–1079. doi:10.1038/sj.bjc.6690809.
35. Wu RX, Bi CS, Yu Y, Zhang LL, Chen FM. Age-related decline in the matrix contents and functional properties of human periodontal ligament stem cell sheets. Acta Biomater. 2015; doi:10.1016/j.actbio.2015.04.024.
36. Kainulainen T, Autio-Harmainen H, Oikarinen A, Salo S, Tryggvason K, Salo T. Altered distribution and synthesis of laminin-5 (kalinin) in oral lichen planus, epithelial dysplasias and squamous cell carcinomas. Br J Dermatol. 1997; 136(3): 331–336. doi:10.1111/j.1365-2133.1997.tb14938.x.
37. Marinkovich MP. Tumour microenvironment: laminin 332 in squamous-cell carcinoma. Nat Rev Cancer. 2007; 7(5):

370–380. doi:10.1038/nrc2089.

38. Richter P, Umbreit C, Franz M, Berndt A, Grimm S, Uecker A, Böhmer FD, Kosmehl H, Berndt A. EGF/TGFβ1 co-stimulation of oral squamous cell carcinoma cells causes an epithelial-mesenchymal transition cell phenotype expressing laminin 332. J Oral Pathol Med. 2011; 40(1): 46–54. doi:10.1111/j.1600-0714.2010.00936.x.
39. Kinoshita T, Nohata N, Hanazawa T, Kikkawa N, Yamamoto N, Yoshino H, Itesako T, Enokida H, Nakagawa M, Okamoto Y, Seki N. Tumour-suppressive microRNA-29s inhibit cancer cell migration and invasion by targeting laminin-integrin signalling in head and neck squamous cell carcinoma. Br J Cancer. 2013; 109(10): 2636–45. doi:10.1038/bjc.2013.607.
40. Hunt S, Jones AV, Hinsley EE, Whawell SA, Lambert DW. MicroRNA-124 suppresses oral squamous cell carcinoma motility by targeting ITGB1. FEBS Lett. 2011; 585(1): 187–192. doi:10.1016/j.febslet.2010.11.038.
41. Foubert P, Varner JA. Integrins in tumor angiogenesis and lymphangiogenesis. Methods Mol Biol. 2011; 757: 471–486. doi:10.1007/978-1-61779-166-6_27.
42. Fransvea E, Mazzocca A, Antonaci S, Giannelli G. Targeting transforming growth factor (TGF)-betaRI inhibits activation of beta1 integrin and blocks vascular invasion in hepatocellular carcinoma. Hepatology. 2009; 49(3): 839–850. doi:10.1002/hep.22731.
43. Desgrosellier JS, Cheresh DA. Integrins in cancer: biological implications and therapeutic opportunities. Nat Rev Cancer. 2010; 10(1): 9–22. doi:10.1038/nrc2965.
44. Golubovskaya VM, Kweh F, Cance WG. Focal adhesion kinase and cancer. Histol Histopathol. 2008; 24(4):503-10.
45. Canel M, Secades P, Garzón-Arango M, Allonca E, Suarez C, Serrels A, Frame M, Brunton V, Chiara M-D. Involvement of focal adhesion kinase in cellular invasion of head and neck squamous cell carcinomas via regulation of MMP-2 expression. Br J Cancer. 2008; 98(7): 1274–1284. doi:10.1038/sj.bjc.6604286.
46. Bray SJ. Notch signalling: a simple pathway becomes complex. Nat Rev Mol Cell Biol. 2006; 7(9): 678–689. doi:10.1038/nrm2009.
47. Zhang H, Wang X, Xu J, Sun Y. Notch1 activation is a poor prognostic factor in patients with gastric cancer. Br J Cancer. 2014; 110(9): 2283-90. doi:10.1038/bjc.2014.135.
48. Ranganathan P, Weaver KL, Capobianco AJ. Notch signalling in solid tumours: a little bit of everything but not all the time. Nat Rev Cancer. 2011; 11(5): 338–351. doi:10.1038/nrc3035.
49. Yap L, Lee D, Khairuddin A, Pairan M, Puspita B, Siar C, Paterson I. The opposing roles of NOTCH signalling in head and neck cancer: a mini review. Oral Dis. 2015; 21(7): 850-7. doi:10.1111/odi.12309.
50. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013; 6(269): pl1. doi:10.1126/scisignal.2004088.
51. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, et al. The cBio cancer genomics portal : An open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012; 2(5): 401-4. doi:10.1158/2159-8290.CD-12-0095.
52. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2014.

www.oncotarget.com Oncotarget, 2018, Vol. 9, (No. 63), pp: 32161-32172

Research Paper

Somatic genome alterations in relation to age in lung squamous cell carcinoma

Stefano Meucci[1], Ulrich Keilholz[1], Daniel Heim[2], Frederick Klauschen[2] and Stefano Cacciatore[3,4]

[1]Charité Comprehensive Cancer Center, Charité University Hospital, Berlin, Germany

[2]Institut für Pathologie, Charité University Hospital, Berlin, Germany

[3]Imperial College Parturition Research Group, Division of the Institute of Reproductive and Developmental Biology, Imperial College London, London, England, UK

[4]International Centre for Genetic Engineering and Biotechnology, Cancer Genomics Group, Cape Town, South Africa

Correspondence to: *Ulrich Keilholz,* ***email:*** *Ulrich.Keilholz@charite.de*
Stefano Meucci, ***email:*** *Stefano.Meucci@charite.de*

Keywords: *lung squamous cell carcinoma; aging; somatic mutations; copy number variations; methylation*

Received: *January 19, 2018* ***Accepted:*** *July 12, 2018* ***Published:*** *August 14, 2018*

ABSTRACT

Lung squamous cell carcinoma (LUSC) is the most common cause of global cancer-related mortality and the major risk factors is smoking consumption. By analyzing ~500 LUSC samples from The Cancer Genome Atlas, we detected a higher mutational burden as well as a higher level of methylation changes in younger patients. The SNPs mutational profiling showed enrichments of smoking-related signature 4 and defective DNA mismatch repair (MMR)-related signature 6 in younger patients, while the defective DNA MMR signature 26 was enriched among older patients. Furthermore, gene set enrichment analysis was performed in order to explore functional effect of somatic alterations in relation to patient age. Extracellular Matrix-Receptor Interaction, Nucleotide Excision Repair and Axon Guidance seem crucial disrupted pathways in younger patients. We hypothesize that a higher sensitivity to smoking-related damages and the enrichment of defective DNA MMR related mutations may contribute to the higher mutational burden of younger patients. The two distinct age-related defective DNA MMR signatures 6 and 26 might be crucial mutational patterns in LUSC tumorigenesis which may develop distinct phenotypes. Our study provides indications of age-dependent differences in mutational backgrounds (SNPs and CNVs) as well as epigenetic patterns that might be relevant for age adjusted treatment approaches.

INTRODUCTION

Lung cancer is the most common cause of global cancer-related mortality and the major risk factors are smoking consumption and occupational exposure to carcinogens [1]. The two major histological classes are non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC). NSCLCs mostly comprise lung adenocarcinomas (LUAD) and lung squamous carcinomas (LUSC) [2], characterized by largely distinct mutational patterns [3].

The mutational landscape present in a cancer genome is the cumulative result of endogenous and/or exogenous mutational processes (e.g., smoking), constant or sporadic and with different strengths along patient ageing [4–7]. Therefore, multiple mutational processes are operative resulting in jumbled composite signatures and tumor characteristics vary between patients of different ages [7–9]. From the Catalogue Of Somatic Mutations In Cancer (COSMIC) which includes 10,952 exomes and 1,048 whole-genomes across 40 distinct types of human

cancer [10], 30 different mutational signatures were identified and publicly released (http://cancer.sanger.ac.uk/cosmic/signatures). Each signature is characterized by the contribution of different factor (e.g., smoking, age, sex). Signature 1 (SI1) characterized by C>T transitions at CpG sites due to the deamination of 5-methylcytosine was associated to mutational processes related to the ageing [4–6, 11]. While Signature 4 (SI4) associated with C>A transversions was found in cancers in which tobacco smoking increases risk and mainly in those derived from cells directly exposed to the tobacco smoke. According to the SI4 pattern, LUSC patients can be classified by the "transversion status" in order to study high and low mutational rate profiles [3]. Past studies hypothesized that chemicals of tobacco smoke increases the speed with which these mutations accumulate [12]. Although the age at diagnosis of lung tumors is very closely correlated with the duration of smoking [13, 14], a previous study performed on 34 tumor types of the TCGA dataset [15], showed significant negative correlations between SNPs and patient age only in LUSC and LUAD. While 29 tumor types exhibited positive correlations, among which the smoking-related tumors such as HNSCC [15, 16]. Therefore the hypothesis of the "mutator phenotype", which is a tumor harboring mutations in DNA polymerases and DNA repair genes [15, 17], has to be taken into account.

Furthermore, Copy Number Variations (CNVs) play also important roles in the development of cancer showing an association with ageing in terms of longevity, healthy aging, and aging-related pathologies [18–20]. Although the number of studies about CNVs and ageing are very limited, age-related CNVs increase observed in human blood cell genomes [21, 22] suggests that CNVs could play a key role even in LUSC.

Moreover, epigenomic alteration is now increasingly recognized as part of aging and its associated pathologic phenotypes as cancer [23]. There is ample evidence for changes in DNA methylation patterns at CpG sites during development and aging, driving essential somatic functions. A general demethylation is linked with aging which may reflects some deficiency in maintenance re-methylation. The epimutation rate appears to be almost 100,000 times the mutation rate and aberrant DNA methylation can predispose to malignancy [22, 24, 25].

This study aims to provide better insight into the underlying genetic and epigenetic patterns of LUSC in relation to patient age. To this end, we investigated the relationships between patient age and the average number of SNPs, CNVs and methylation changes as well as the SNPs profiling and the respective correlation to the previously defined signatures in COSMIC. Furthermore, we performed gene-specific correlation analysis in relation to patient age with a particular focus on the significantly mutated genes in LUSC [3] and the most frequently mutated DNA repair genes in lung cancer [26]. Finally, gene set enrichment analysis was performed in order to explore functional effect of somatic alterations in relation to patient age.

The current study may pave the way for future studies of molecular tumorigenesis in relation to human ageing and underlines the need to consider age-adjusted treatments not only based on age and morbidity of older patients, but also on differences in tumor biology.

RESULTS

Somatic alterations and patient age

Genome-wide mutations and epigenomic changes are expected to varying among tumor subtypes showing a different distribution across age. To characterize these distinct distribution patterns, we firstly estimated the global number of SNPs, CNVs, and methylation changes at CpG sites for 504 samples across LUSC cancer cohort available through The Cancer Genome Atlas (TCGA). We used the Spearman's rank correlation coefficient to explore the relation between the number of SNPs, CNVs and methylation changes with patient age.

The global SNPs load showed a slightly negative correlation with patient age (Table 1), which indicated a higher mutational rate among younger patients (Figure 1A). Then, we classified SNPs according to their expected biological effect as low, moderate, or severe (as shown in Supplementary Table 1) and we identified the genes with at least a severe or moderate mutation. We reported a lower correlation between the age and the number of genes with disruptive mutations (*rho*=-0.08, *p*=0.077, *FDR*=0.26). The global CNVs load showed no correlation with patient age (Figure 1B). While methylation changes were negatively correlated with patient age (*rho*=-0.11, *p*=0.030, *FDR*=0.23) displaying a higher level of methylation at CpG sites among younger patients (Figure 1C).

We repeated the analysis on patient sub-cohorts established according to the tobacco exposure data (i.e., tobacco smoking history indicator), tumor staging (i.e., ajcc pathologic tumor stage), and mutational rate profile (i.e., transversion status) in order to explore the influence of patient features on the relation among SNPs, CNVs, and methylation changes with patient age. The analysis of sub-cohort with a high mutational load (i.e., transversion-high status) showed a negative correlation between the SNPs load and patient age while no correlations were detected in the low mutational load sub-cohort (i.e., transversion low status) (Table 1). The results regarding CNVs and methylation changes were fully reported in Supplementary Table 2.

Gene-specific alterations enrichment along patient ageing

The Spearman's rank correlation was computed between SNPs, CNVs, and methylation changes in

Table 1: SNPs loads correlations with patient age

Classification	Patients n.	rho [95%CI]	p-value	FDR
Global	480	-0.09 [-0.19 0]	4.53×10^{-2}	1.81×10^{-1}
Transversion Status				
High	387	-0.11 [-0.22 -0.01]	2.60×10^{-2}	1.56×10^{-1}
Low	84	0.15 [-0.05 0.34]	1.87×10^{-1}	3.21×10^{-1}
Tobacco smoking history indicator				
Lifelong non-smokers	18	0.11 [-0.41 0.61]	6.54×10^{-1}	7.85×10^{-1}
Current smokers	131	-0.12 [-0.29 0.05]	1.66×10^{-1}	3.21×10^{-1}
Current reformed smokers for >15 yrs	78	-0.19 [-0.38 0.03]	9.88×10^{-2}	2.96×10^{-1}
Current reformed smokers for < or = 15 yrs	236	-0.09 [-0.22 0.05]	1.59×10^{-1}	3.21×10^{-1}
Current reformed smokers, duration not specified	5	-0.1 [-1 1]	9.50×10^{-1}	9.50×10^{-1}
Ajcc pathologic tumor stage				
1	233	-0.07 [-0.19 0.06]	3.13×10^{-1}	4.70×10^{-1}
2	153	0.02 [-0.13 0.19]	7.66×10^{-1}	8.36×10^{-1}
3	83	-0.35 [-0.53 -0.15]	1.12×10^{-3}	1.34×10^{-2}
4	7	-0.29 [-0.96 0.62]	5.56×10^{-1}	7.41×10^{-1}

Correlations between the SNPs loads and patient age for each patient sub-group established according to the patient characteristic evaluated in our study, such as tobacco exposure data (i.e., tobacco smoking history indicator), tumor staging (i.e., ajcc pathologic tumor stage), and mutational rate profile (i.e., transversion status).

each gene and patient age, we reported the results in Supplementary Table 3. A special focus was placed on the 20 significantly mutated genes previously found in LUSC [3] (Supplementary Table 4, Figure 1D–1F). A negative correlation between patient age and both CNVs (*rho*=-0.13, *p*=0.005, *FDR*=0.16) and methylation changes (*rho*=-0.14, *p*=0.006, *FDR*=0.06) was detected on NOTCH1, while no SNPs correlation was displayed. A significantly higher level of methylation at CpG sites in younger patients was as well exhibited in RASA1 (*rho*=-0.19, *p*=0.0002, *FDR*=0.01), ARID1A1 (*rho*=-0.22, *p*=0.00005, *FDR*=0.006), PASK (*rho*=-0.11, *p*=0.04, *FDR*=0.16) and NSD1 (*rho*=-0.13, *p*=0.02, *FDR*=0.09).

In order to explore the hypothesis of possible mutator phenotypes contributing to the high mutational rate detected among younger patients, we analyzed whether mutations harboring on the top 20 frequently mutated DNA repair genes in lung cancer [26] might have a significant impact on the SNPs load. For each of them, the Wilcoxon test was performed to compare the mutational load of the patient sub-cohorts exhibiting the somatic alterations against the wild-type patient groups (Supplementary Table 5). The percentage of patients which have at least one of the genes mutated was >83% in each age-group. The mutator phenotype had a significant impact on the mutational load in 60-70 and 70-80 age classes. Therefore the analysis was repeated grouping the patient global cohort in younger and older than 60 years old. While only 3 genes were significant in ≤60 years old patients, 14 out of 20 genes had a significant impact on the mutational load in >60 years old patients.

Age-related COSMIC signatures

Somatic mutation profile is the sum of multiple mutation processes, such as the intrinsic infidelity of the DNA replication machinery, exogenous or endogenous mutagen exposures, enzymatic modification of DNA, and defective DNA repair. In order to analyze each mutation process separately, we correlated the patient age with single nucleotide variants (Supplementary Table 6) and COSMIC signatures (Supplementary Table 7) using the Spearman's rank correlation. Additionally, the Wilcoxon Rank-Sum test was performed to evaluate the differences between each age group (i.e., <50, 50-60, 60-70, 70-80, >80) and the rest of the cohort.

The defective DNA mismatch repair (MMR)-related signature 6 (SI6) was negatively correlated (*rho*=-0.13, *p*=0.004, *FDR*=0.12) with the patient age (Figure 2A) while the signature 26 (SI26) as well associated with defective DNA MMR, was positively correlated (*rho*=0.11, *p*=0.013, *FDR*=0.20) with the patient age (Figure 2B). Both signatures showed similar trend in the transversion-high sub-cohort. The smoking-related SI4

was negatively correlated (*rho*=-0.11, *p*=0.02, *FDR*=0.21) with patient age (Figure 2C), showing higher values in the ≤50 and 51-60 age groups (Supplementary Table 7). No correlation was detected for the age-related SI1.

In order to study the patient sub-cohorts, which predominantly exhibit SI26 and SI6, we divided the overall LUSC cohort into four subgroups using the mean values of SI6 and SI26 as threshold (Figure 2D): high-SI6/high-SI26 (77/480=16.0%), low-SI6/high-SI26 (55/480=11.0%), high-SI6/low-SI26 (223/480=45.8%), and low-SI6/low-SI26 (130/480=27.1%). We selected and characterized the low-SI6/high-SI26 and high-SI6/low-SI26 subgroups (Supplementary Table 8). The patients age of the low-SI6/high-SI26 cohort was significantly higher than the high-SI6/low-SI26 cohort (Wilcoxon Rank-Sum test: *p*=0.005).

Gene set enrichment analysis

On the basis of the previous analysis, the LUSC mutation profile in relation to ageing is characterized by two major defective DNA MMR-related signatures (i.e., SI6 and SI26). To study the molecular effects of these signatures independently, we projected the SNPs, CNVs and DNA methylation values from the high-SI6/low-SI26 and low-SI6/high-SI26 subtypes into the space of the 186 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways by means of single-sample gene set enrichment analysis (ssGSEA) (Supplementary Table 9) [27].

Using the Wilcoxon Rank-Sum test, we reported as major significant differences, that Extracellular Matrix (ECM)-Receptor Interaction pathway (*p*=0.0002, *FDR*=0.04) was significantly enriched of SNPs while

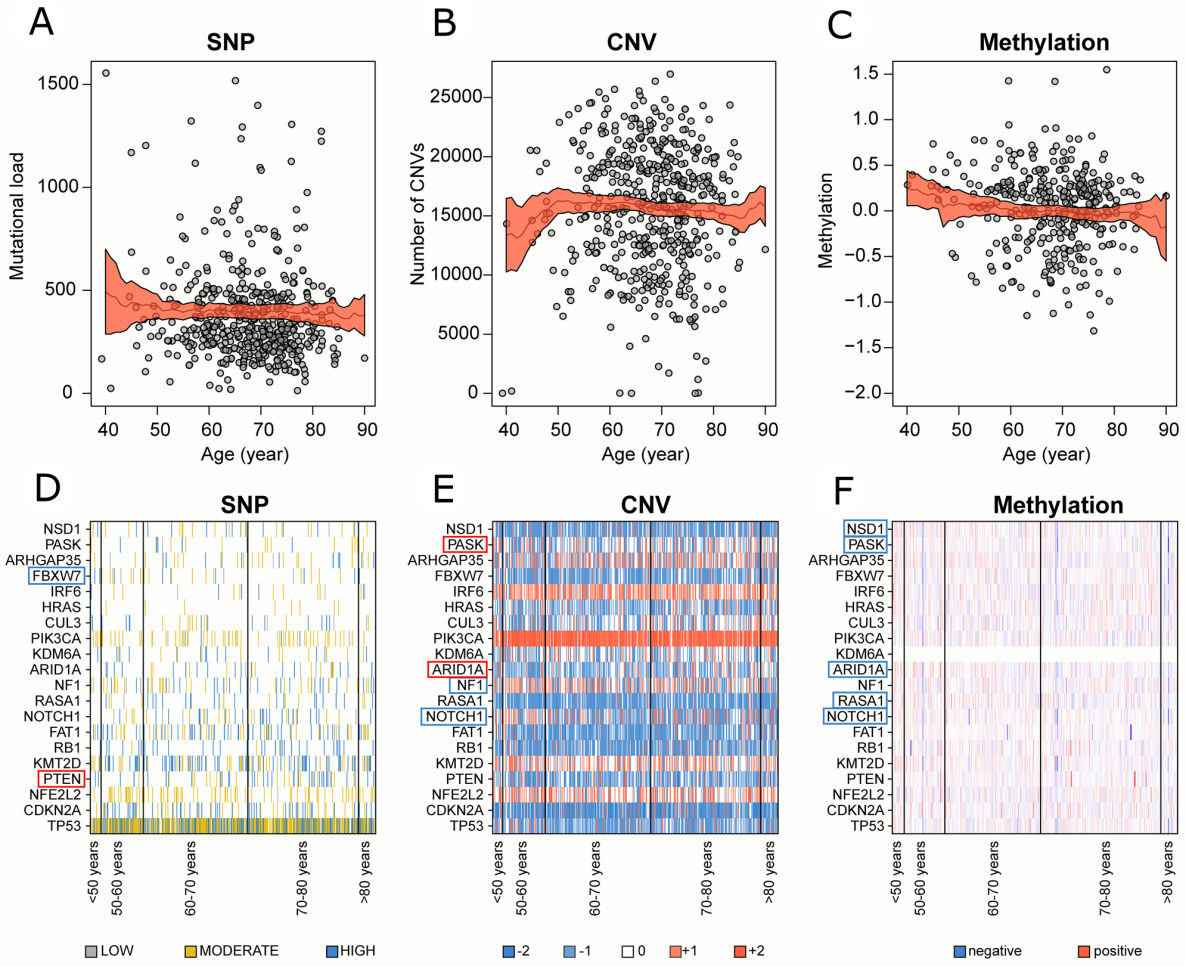

Figure 1: Correlation between genomic alterations and patient age in global cohort. Number of **(A)** SNPs, **(B)** CNVs and **(C)** methylation changes with their relative 95% confidence interval for each patient distributed along patient age. Medians (black line) and their relative 95% confidence interval (red area) were calculated locally in a range of ±10 years. **(D)** SNPs, **(E)** CNVs and **(F)** methylation changes profile of the 20 significantly mutated genes in LUSC. Significantly positive and negative correlated genes were highlighted in red and blue respectively.

the Nucleotide Excision Repair pathway was enriched in CNVs (p=0.0007, *FDR*=0.14) in high-SI6/low-SI26 sub-cohort (Figure 3). The Regulation of Autophagy pathway (p=0.0006, *FDR*=0.06) showed an enrichment of SNPs in low-SI6/high-SI26 patient sub-cohort. Using the Spearman's Rank Correlation Coefficient, we detected a negative correlation between SNPs harboring on ECM Receptor Interaction pathway and patient age (*rho*=-0.16, p=0.016, *FDR*=0.73) in high-SI6/low-SI26 sub-cohort. In Figure 3, the GSEA values of "ECM-Receptor Interaction" pathway were reported for both (Figure 3A) high-SI6/low-SI26 and (Figure 3B) low-SI6/high-SI26 patient sub-cohorts in order to visualize the different trends. Unsupervised hierarchical clustering of SNPs frequencies of genes involved in the "ECM Receptor Interaction" pathway (according to the KEGG database) was added in order to report the pathway mutation profile (Figure 3C–3D).

When evaluating the global cohort, we detected a significant negative correlation between patient age and SNPs harboring on "Axon-Guidance" (*rho*=-0.15, p=0.0007, *FDR*=0.14) and ECM Receptor Interaction (*rho*=-0.13, p=0.003, *FDR*=0.16) pathways, particularly in the 51-60 age group. Furthermore, the Axon-Guidance (*rho*=-0.16, p=0.001, *FDR*=0.12) pathway was the only negatively enriched pathway in transversion-high sub-cohort (Supplementary Table 10).

DISCUSSION

We identified a slightly higher SNPs load among younger patients of the TCGA LUSC patient cohort confirming a previous study [15]. In particular, the correlation was higher in tumors with high mutational burden. Since the correlation was not robust, we believe that our results must be evaluated in an independent cohort to confirm higher mutational rate in younger patients. Interestingly, a higher overall methylation rate at CpG sites was as well detected among younger patients. Although the knowledge is still limited, numerous studies showed that CpG methylation plays an important role in maintaining gene silencing. Several studies have revealed that tumor suppressor gene promoter hypermethylation is noted in tumor cells [28]. However, normal non-proliferative cells also showed gene promoter hypermethylation as age increases [29, 30]. Age-dependent hypermethylation at CpGs was observed to be enriched with DNA binding factors and transcription factors, therefore the dysregulation can simultaneously affect several biological processes [31, 32]. On the contrary Heyn *et al.* [32] revealed that centenarians exhibit lower DNA methylation levels compared with newborns. Therefore, the higher methylation level at CpG sites among younger patients detected in our study

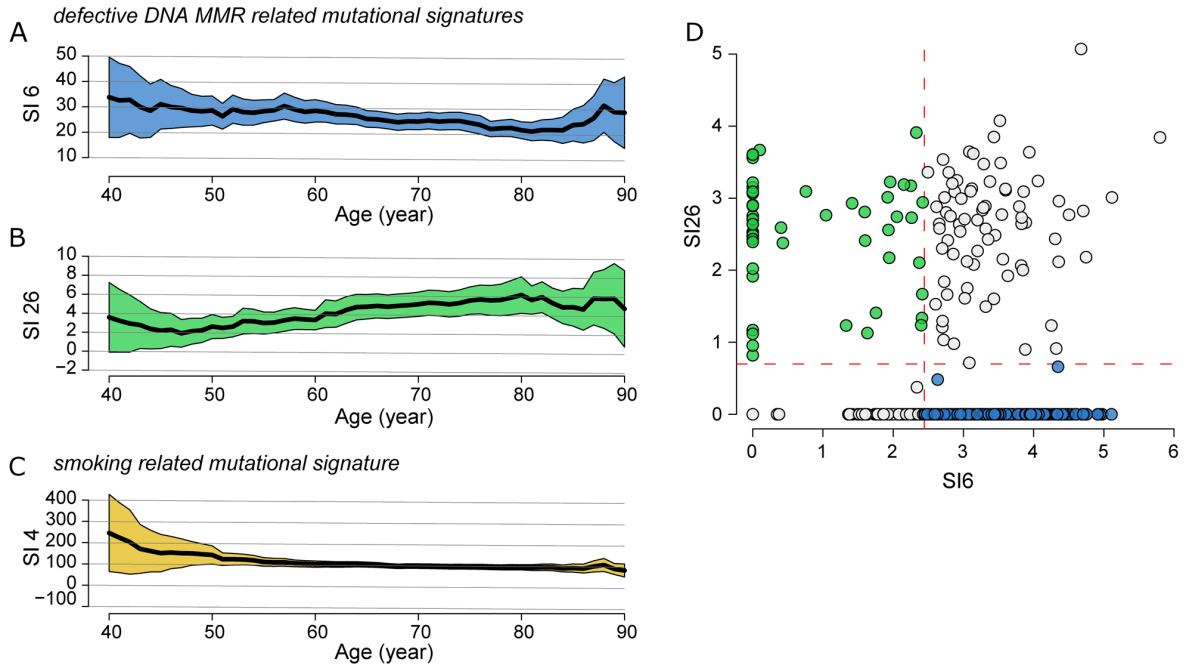

Figure 2: Correlation of SNPs profiling and patient age in global cohort. Correlation between defective DNA MMR **(A)** SI6 and **(B)** SI26, and smoking related **(C)** SI4 with patient age. Medians (black line) and their relative 95% confidence interval (colored area) were calculated locally in a range of ±10 years. **(D)** Classification of the overall LUSC cohort into four subgroups using the mean values (dashed red lines) of SI6 and SI26 as threshold: high-SI6/high-SI26, low-SI6/high-SI26 (green circle), high-SI6/low-SI26 (blue circle) and low-SI6/low-SI26. The values are converted as log(x+1).

might comprise both aberrations and normal age-related patterns. We detected 5 out of 20 significantly mutated genes in LUSC (NOTCH1, RASA1, ARID1A1, PASK, NSD1) exhibiting a significantly higher methylation levels in younger patients. CNVs enrichment was as well detected in NOTCH1 among younger patients. NOTCH1 is one of the highly significant mutated genes in Cancer. Cross-talking with many other critical cancer genes and pathways, NOTCH1 is involved in multifaceted regulation of cell survival, proliferation, tumor angiogenesis, and metastasis. A recent study observed that with long-term smoking exposure, the DNA sequence suffers persistent miscoding that triggers epigenetic changes in NOTCH1 [33]. Therefore NOTCH1 aberrations might be involved

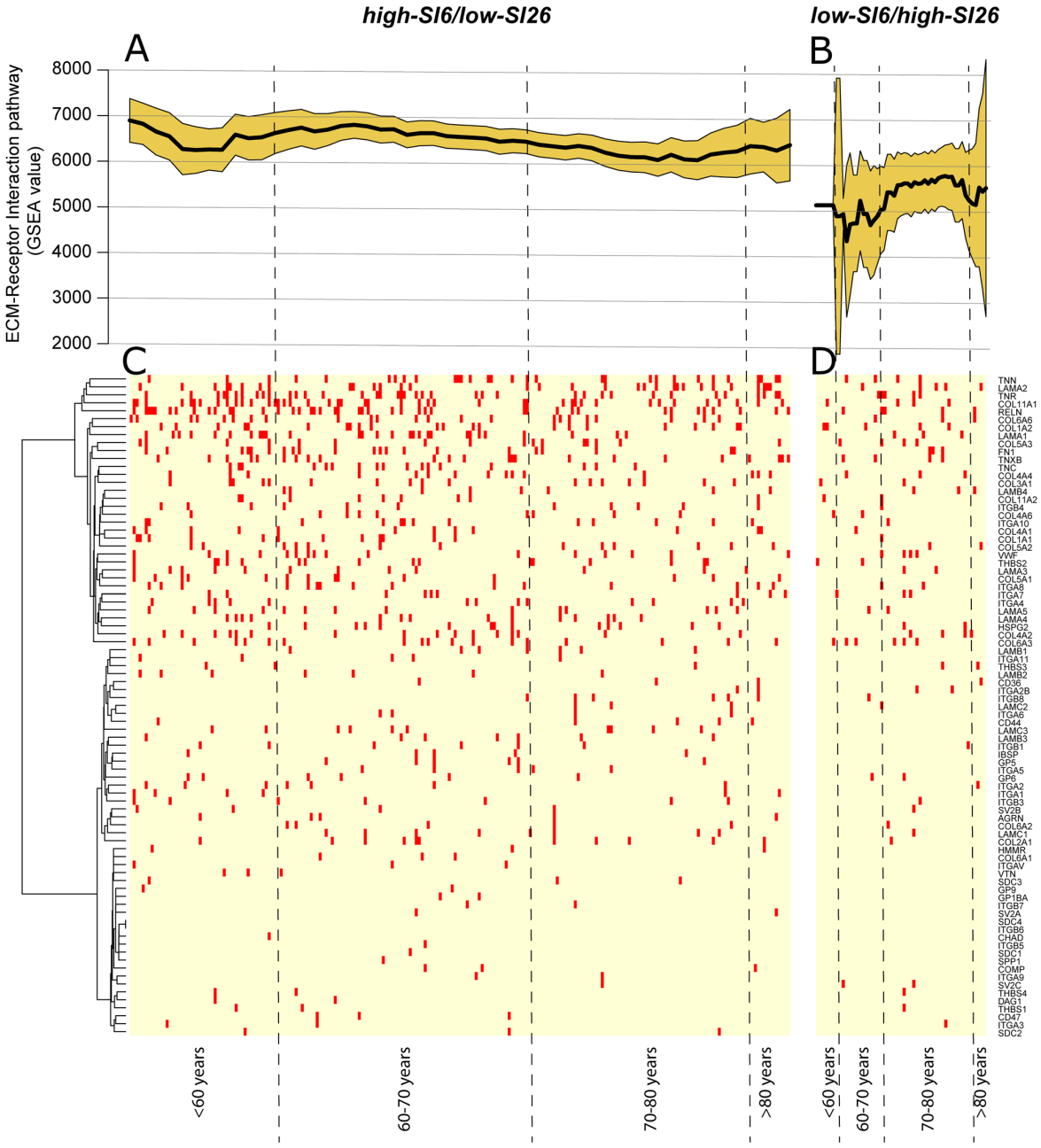

Figure 3: **(A)** GSEA value of "ECM-Receptor Interaction" pathway in high-SI6/low-SI26 and **(B)** low-SI6/high-SI26 patient sub-cohorts. Unsupervised hierarchical clustering of SNPs frequencies of genes involved in the "ECM Receptor Interaction" pathway (according to the KEGG database) in **(C)** high-SI6/low-SI26 and **(D)** low-SI6/high-SI26.

in the peculiar higher mutational burden of younger LUSC patients.

Mutator phenotypes might develop in LUSC tumorigenesis [15], therefore we evaluated the mutational profile of the top 20 frequently mutated DNA repair genes in lung cancer [26]. No significant differences in mutation frequencies were detected among the age classes. More than 83 % of the patients harbored at least one of the genes mutated in all age classes. Thus, mutator phenotypes seem evenly distributed along patient ageing, contributing to the overall high mutational burden in LUSC patients. On the contrary, the impact of these mutations on the mutational load was significantly higher in >60 years old patients. Therefore, mutator phenotypes might have different consequences in relation to ageing processes.

The overall SNPs mutational profiling and the corresponding correlations with COSMIC signatures showed an enrichment of the smoking-related signature (i.e., SI4) among younger patients. Past studies described a similar scenario showing that despite maintained carcinogen exposure, tumors from smokers showed a relative decrease in smoking-related mutations over time [34, 35]. Therefore, younger patients may develop higher sensitivity to smoking-related mutations. The defective DNA MMR SI6 and SI26 were as well significantly correlated with patient age. The SI6, characterized predominantly by C>T at NpCpG sites (any nucleotide followed by C followed by G), was enriched in younger patients. While the SI26, mostly composed of T>C transitions, was enriched in older patients. Both SI6 and SI26 are found in microsatellite unstable tumors with high numbers of small (shorter than 3bp) insertions and deletions at mono/polynucleotide repeats [36, 37]. The role of MMR system is to recognize and repair erroneous insertion, deletion, and mis-incorporation of bases arising during DNA replication and homologous recombination, as well as repairing some forms of DNA damage. Given the importance of these processes in the maintenance of genomic stability, DNA MMR deficiency might leads to hypermutation [38, 39]. A recent study showed that out of a large number of DNA repair deficiencies analyzed, MMR deficiency leads to the by far highest mutation rate [36]. Our results suggest that different causing factors might contribute to MMR system aberrations along patient ageing. Therefore we performed gene set enrichment analysis in patient sub-cohorts which predominantly exhibit SI6 or SI26. We identified the SNPs enrichment in ECM-Receptor Interaction pathway among younger patients of high-SI6/low-SI26 sub-cohort. The ECM-Receptor Interaction pathway is structurally and functionally involved in interactions at the ECM which lead to a direct or indirect control of cellular activities such as cell migration, differentiation, proliferation, and apoptosis [40–42]. Aberrant ECM may promote genetic instability and might compromise DNA repair pathways necessary to prevent malignant transformation [40]. Furthermore, we identified an enrichment of CNVs in Nucleotide Excision Repair (NER) pathway in high-SI6/low-SI26 sub-cohort. Since the NER system is primarily responsible for detecting and removing bulky DNA lesions induced by tobacco smoke in the respiratory tract [43], SNPs in NER protein-encoding genes may contribute to the higher sensitivity to smoking consumption detected in younger patients. Early studies identified associations with lung cancer risk in selected mutated NER genes (ERCC1-6, LIG1, POLE, XPA, and XPC genes) [44–47].

The low-SI6/high-SI26 sub-cohort was enriched in SNPs disruptions of Regulation of Autophagy pathway involved in lysosome-dependent degradation processes. On one hand, autophagy has been shown to regulate some of the DNA repair proteins after DNA damage by maintaining the balance between their synthesis, stabilization, and degradation. One the other hand, some evidence has demonstrated that some DNA repair molecules have a crucial role in the initiation of autophagy [48, 49]. Therefore, disruption of Regulation of Autophagy pathway might contribute to the defective DNA MMR system in low-SI6/high-SI26 patient sub-cohort.

Considering the "global" cohort, SNPs harboring on genes involved in ECM-Receptor Interaction and Axon Guidance pathways were enriched among younger patients. Intriguingly, in our previous study on HNSCC, we detected the same two pathways enriched among older patients, which were the higher mutational rate samples due to the proportional relation between the HNSCC global mutational load and patient age [16]. Therefore, although the inverse tendency, Axon Guidance and ECM-Receptor Interaction pathways seem to show a relation with higher mutational rate squamous carcinomas. Several studies reported that Axon Guidance pathway is involved in lung cancer development and progression through interacting with cell survival, migration, and tumor angiogenic pathways [50–54]. Further studies are needed to determine whether disruptions in these pathways are a correlative phenotype to higher mutational rate squamous carcinomas or a causative factor.

In conclusion, multiple mutational processes appear to be simultaneously operative with various dynamic changes due to the endogenous and exogenous environments, life style habits and physiological ageing. Previous hypothesis of a mutator phenotype concealing the effect of age-related accumulation of mutations might have different causing factors in relation to ageing processes. We hypothesize that a higher sensitivity to smoking-related damages and the enrichment of defective DNA MMR SI6 may contribute to the higher mutational burden of younger patients. A higher overall level of methylation was as well detected in younger patients. While the defective DNA MMR SI26 showed increasing tendency along patient ageing. Therefore, the two distinct age-related defective DNA MMR signatures SI6 and SI26 might be crucial mutational patterns

in LUSC tumorigenesis which may develop distinct phenotypes.

The evaluation of somatic genomic alterations along patients ageing might be relevant for a better comprehension of LUSC tumorigenesis and development of age-adjusted treatments.

MATERIALS AND METHODS

TCGA data sets

Multiplatform genomic data sets were generated by TCGA Research Network (http://cancergenome.nih.gov/). Cancer molecular profiling data were generated through informed consent as part of previously published studies [55] and analyzed in accordance with each original study's data use guidelines and restrictions. The clinical data of the 504 LUSC normal paired exome sequences was derived via download from the publicly available GDC Data Portal (https://portal.gdc.cancer.gov/).

Whole exome analysis

Somatic mutations were obtained from the open access MAFs available from the GDC Legacy Archive (https://portal.gdc.cancer.gov/legacy-archive). We considered three different exclusion criteria for mutation data entries. Samples belonging to the same patient share a very similar mutational profile. In the first exclusion criteria, we considered only once a mutation present in different samples belonging to the same patient. The mutations not included were equal to the 25.2% (282163 =>210948).

Some genes can share a similar sequence, such as paralogous genes. In presence of a mutation event on a sequence shared among different genes, it will not be possible to identify the mutated gene. With the second exclusion criterion, we decide to remove mutations that were associated to more than one gene. In this step we removed the 0.1% of mutations (210948 => 210700).

The challenges of repetitive sequence, which constitute 50–69 % of the human genome leads to false positive variant calls due to systematic sequencing errors and local alignment challenges [56]. Therefore, only somatic mutations with "ref context" containing less than 6 continuous single repetitions, less than 4 continuous duplets, less than 3 continuous triplets, less than 3 continuous quadruplets, less than 3 continuous quintuplets were kept. With the third exclusion criteria, the mutations were reduced from 210700 to 194170 (~8.8%).

The patient TGCA-66-2755 was excluded from the following analysis due to the unusual number of mutations.

SNP array-based copy number analysis

DNA from each tumor or germline-derived sample had been hybridized to Affymetrix SNP 6.0 arrays [57] and processed through GISTIC [58, 59] by the TCGA consortium.

High-level copy gain or copy loss events for individual genes were inferred using the publicly available Firehose's (Gistic2.Level4) data (http://gdac.broadinstitute.org/runs/analyses__2016_01_28/data/LUSC/20160128/) (+2 values being indicative of gains greater than 1-2 copies, -2 values being indicative of near total copy loss). Global CNV load were calculated summing the absolute values from each patients.

Array-based DNA methylation assay

DNA methylation profiles had been previously generated by TCGA using either the Infinium HM450 or HM27 assay probe. The level 3 beta value DNA methylation scores for individual genes were inferred using publicly available data generated by Illumina Human Methylation 450 platform downloaded from the GDC Legacy Archive (https://portal.gdc.cancer.gov/legacy-archive). Methylation values were mean centered and scaled to unit variance. After the transformation, the rate of methylation changes was calculated summing the values of each gene.

Single nucleotide variants and COSMIC signatures

The signature profile was evaluated using the six subtype: C>A, C>G, C>T, T>A, T>C, and T>G (all substitutions were referred to by the pyrimidine of the mutated Watson-Crick base pair). Further, each of the substitutions was examined by incorporating information on the bases immediately 5' and 3' to each mutated base generating 96 possible single nucleotide variants (6 types of substitution x 4 types of 5' base x 4 types of 3' base). The profile of these 96 single nucleotide variants was considered as the results of the combination of the 30 different COSMIC signatures. The profile of each tumor sample can be represented by a unique contribution of each COSMIC signature as the following expression:

$$a_1 \times SI1 + a_2 \times SI2 + a_3 \times SI3 + \ldots + a_{30} \times SI30 \qquad (1)$$

where a_i is the coefficient representing the contribution of the i_{th} COSMIC signature. The coefficients of each tumor samples were calculated minimizing the difference between the tumor profile and the expression *(1)*. This procedure was implemented using the function *optim* (method "L-BFGS-B" [60]) of the R software [61].

Molecular pathway and biological process analysis

Pathway analyses were performed by ssGSEA using the GenePattern module ssGSEA Projection (v4) (genepattern.broadinstitute.org). ssGSEA enrichment

scores were calculated from SNPs, CNV, and methylation LUSC data sets. The result is a single score per patient per gene set, transforming the original data sets into a more interpretable higher-level description. For the use of ssGSEA software, annotated gene sets reference were obtained from the C2 KEGG sub-collection of the Molecular Signature database (MSigDB) [62]. Silent mutations (point mutations that would not result in a change in the amino acid sequence) were not included in the analysis.

Statistical analysis

The Spearman's Rank Correlation Coefficient was used to identify correlation between patient age and genomic/epigenomic data (e.g., SNP, CNV, and methylation loads). For every Spearman's test performed in this study, p-values were computed using algorithm AS 89 included in the R function *cor.test* where the permutation distribution was estimated by an Edgeworth approximation [63]. The coefficient interval of rho value was calculated by bootstraping (with 1000 replicates) using the function *spearman.ci* of the R package *RVAideMemoire*. Fisher's exact test was used to examine the significance of the association between COSMIC signature related subgroups (i.e., low-SI6/high-SI26 and high-SI6/low-SI26) and clinical/demographic/molecular patient features, such as gender, tobacco smoking history indicator, and mutated / wild type genes. Fisher's exact test was computed using the R function *fisher.test*. Wilcoxon Rank-Sum test was performed to compare continuous variables between two patient subgroups using the R function *wilcox.test*. A p-value <0.05 was considered to be significant. To account for multiple testing, a FDR of $\leq 20\%$ was applied to reduce identification of false positives [64]. The FDR was calculated using the R function *p.adjust*. All calculations were made using R software [61].

ACKNOWLEDGMENTS AND FUNDING

This work was supported by the Focus Area Dynage (www.fu-berlin.de/dynage), German Cancer Research Center (DKTK) and the Charité Comprehensive Cancer Center, Berlin.

CONFLICTS OF INTEREST

There is no conflicts of interest that I should disclose.

REFERENCES

1. Collisson EA, Campbell JD, Brooks AN, Berger AH, Lee W, Chmielecki J, Beer DG, Cope L, Creighton CJ, Danilova L, Ding L, Getz G, Hammerman PS, et al, and Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014; 511:543–50. https://doi.org/10.1038/nature13385.
2. Polo V, Pasello G, Frega S, Favaretto A, Koussis H, Conte P, Bonanno L. Squamous cell carcinomas of the lung and of the head and neck: new insights on molecular characterization. Oncotarget. 2016; 7:25050–63. https://doi.org/10.18632/oncotarget.7732.
3. Campbell JD, Alexandrov A, Kim J, Wala J, Berger AH, Pedamallu CS, Shukla SA, Guo G, Brooks AN, Murray BA, Imielinski M, Hu X, Ling S, et al, and Cancer Genome Atlas Research Network. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat Genet. 2016; 48:607–16. https://doi.org/10.1038/ng.3564.
4. Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr Opin Genet Dev. 2014; 24: 52–60. https://doi.org/10.1016/j.gde.2013.11.014.
5. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, Stratton MR. Clock-like mutational processes in human somatic cells. Nat Genet. 2015; 47:1402–07. https://doi.org/10.1038/ng.3441.
6. Fox EJ, Salk JJ, Loeb LA. Exploring the implications of distinct mutational signatures and mutation rates in aging and cancer. Genome Med. 2016; 8:30. https://doi.org/10.1186/s13073-016-0286-z.
7. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale AL, Boyault S, Burkhardt B, Butler AP, et al, and Australian Pancreatic Cancer Genome Initiative, and ICGC Breast Cancer Consortium, and ICGC MMML-Seq Consortium, and ICGC PedBrain. Signatures of mutational processes in human cancer. Nature. 2013; 500:415–21. https://doi.org/10.1038/nature12477.
8. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. Cell Reports. 2013; 3:246–59. https://doi.org/10.1016/j.celrep.2012.12.008.
9. Gao Y, Gao F, Ma JL, Zhang XZ, Li Y, Song LP, Zhao DL. Analysis of the characteristics and prognosis of advanced non-small-cell lung cancer in older patients. Patient Prefer Adherence. 2015; 9:1189–94. https://doi.org/10.2147/PPA.S87069.
10. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, Kok CY, Jia M, De T, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015; 43:D805–11. https://doi.org/10.1093/nar/gku1075.
11. Tomasetti C, Vogelstein B, Parmigiani G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proc Natl Acad Sci USA. 2013; 110:1999–2004. https://doi.org/10.1073/pnas.1221068110.
12. Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, Totoki Y, Fujimoto A, Nakagawa H,

Shibata T, Campbell PJ, Vineis P, Phillips DH, Stratton MR. Mutational signatures associated with tobacco smoking in human cancer. Science. 2016; 354:618–22. https://doi.org/10.1126/science.aag0299.

13. Westmaas JL, Newton CC, Stevens VL, Flanders WD, Gapstur SM, Jacobs EJ. Does a recent cancer diagnosis predict smoking cessation? An analysis from a large prospective US cohort. J Clin Oncol. 2015; 33:1647–52. https://doi.org/10.1200/JCO.2014.58.3088.

14. Baser S, Shannon VR, Eapen GA, Jimenez CA, Onn A, Lin E, Morice RC. Smoking cessation after diagnosis of lung cancer is associated with a beneficial effect on performance status. Chest. 2006; 130:1784–90. https://doi.org/10.1016/S0012-3692(15)50902-1.

15. Milholland B, Auton A, Suh Y, Vijg J. Age-related somatic mutations in the cancer genome. Oncotarget. 2015; 6:24627–35. https://doi.org/10.18632/oncotarget.5685.

16. Meucci S, Keilholz U, Tinhofer I, Ebner OA. Mutational load and mutational patterns in relation to age in head and neck cancer. Oncotarget. 2016; 7:69188–99. https://doi.org/10.18632/oncotarget.11312.

17. Loeb LA. Human cancers express mutator phenotypes: origin, consequences and targeting. Nat Rev Cancer. 2011; 11:450–57. https://doi.org/10.1038/nrc3063.

18. Iakoubov L, Mossakowska M, Szwed M, Duan Z, Sesti F, Puzianowska-Kuznicka M. A common copy number variation (CNV) polymorphism in the CNTNAP4 gene: association with aging in females. PLoS One. 2013; 8:e79790. https://doi.org/10.1371/journal.pone.0079790.

19. Wang C, Su H, Yang L, Huang K. Integrative analysis for lung adenocarcinoma predicts morphological features associated with genetic variations. Pac Symp Biocomput. 2017; 22:82–93. https://doi.org/10.1142/9789813207813_0009.

20. Nygaard M, Debrabant B, Tan Q, Deelen J, Andersen-Ranberg K, de Craen AJ, Beekman M, Jeune B, Slagboom PE, Christensen K, Christiansen L. Copy number variation associates with mortality in long-lived individuals: a genome-wide assessment. Aging Cell. 2016; 15:49–55. https://doi.org/10.1111/acel.12407.

21. Forsberg LA, Rasi C, Razzaghian HR, Pakalapati G, Waite L, Thilbeault KS, Ronowicz A, Wineinger NE, Tiwari HK, Boomsma D, Westerman MP, Harris JR, Lyle R, et al. Age-related somatic structural changes in the nuclear genome of human blood cells. Am J Hum Genet. 2012; 90:217–28. https://doi.org/10.1016/j.ajhg.2011.12.009.

22. Vijg J, Suh Y. Genome instability and aging. Annu Rev Physiol. 2013; 75:645–68. https://doi.org/10.1146/annurev-physiol-030212-183715.

23. Yang Y, Zhao L, Huang B, Hou G, Zhou B, Qian J, Yuan S, Xiao H, Li M, Zhou W. A new approach to evaluating aberrant DNA methylation profiles in hepatocellular carcinoma as potential biomarkers. Sci Rep. 2017; 7:46533. https://doi.org/10.1038/srep46533.

24. Gravina S, Vijg J. Epigenetic factors in aging and longevity. Pflugers Arch. 2010; 459:247–58. https://doi.org/10.1007/s00424-009-0730-7.

25. Lin Q, Wagner W. Epigenetic aging signatures are coherently modified in cancer. PLoS Genet. 2015; 11:e1005334. https://doi.org/10.1371/journal.pgen.1005334.

26. Chae YK, Anker JF, Carneiro BA, Chandra S, Kaplan J, Kalyan A, Santa-Maria CA, Platanias LC, Giles FJ. Genomic landscape of DNA repair genes in cancer. Oncotarget. 2016; 7:23312–21. https://doi.org/10.18632/oncotarget.8196.

27. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, Schinzel AC, Sandy P, Meylan E, Scholl C, Fröhling S, Chan EM, Sos ML, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009; 462:108–12. https://doi.org/10.1038/nature08460.

28. Wang Y, Zhang J, Xiao X, Liu H, Wang F, Li S, Wen Y, Wei Y, Su J, Zhang Y, Zhang Y. The identification of age-associated cancer markers by an integrative analysis of dynamic DNA methylation changes. Sci Rep. 2016; 6:22722. https://doi.org/10.1038/srep22722.

29. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nat Rev Genet. 2002; 3:415–28. https://doi.org/10.1038/nrg816.

30. Alisch RS, Barwick BG, Chopra P, Myrick LK, Satten GA, Conneely KN, Warren ST. Age-associated DNA methylation in pediatric populations. Genome Res. 2012; 22:623–32. https://doi.org/10.1101/gr.125187.111.

31. Yuan T, Jiao Y, de Jong S, Ophoff RA, Beck S, Teschendorff AE. An integrative multi-scale analysis of the dynamic DNA methylation landscape in aging. PLoS Genet. 2015; 11:e1004996. https://doi.org/10.1371/journal.pgen.1004996.

32. Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, Diez J, Sanchez-Mut JV, Setien F, Carmona FJ, Puca AA, Sayols S, Pujana MA, et al. Distinct DNA methylomes of newborns and centenarians. Proc Natl Acad Sci USA. 2012; 109:10522–27. https://doi.org/10.1073/pnas.1120658109.

33. Ma Y, Li MD. Establishment of a strong link between smoking and cancer pathogenesis through DNA methylation analysis. Sci Rep. 2017; 7:1811. https://doi.org/10.1038/s41598-017-01856-4.

34. de Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L, Jamal-Hanjani M, Shafi S, Murugaesu N, Rowan AJ, Grönroos E, Muhammad MA, Horswell S, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science. 2014; 346:251–56. https://doi.org/10.1126/science.1253462.

35. Zhang J, Fujimoto J, Zhang J, Wedge DC, Song X, Zhang J, Seth S, Chow CW, Cao Y, Gumbs C, Gold KA, Kalhor N, Little L, et al. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. Science. 2014; 346:256–59. https://doi.org/10.1126/science.1256930.

36. Pj C, Regulation G, Molecular E, Ebi E. -, Biology C, Project CG, Regulation G, Dow D, Dundee S, Uk EH. Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers. bioRxiv. 2017; 44. https://doi.org/10.1101/149153..

37. Alexandrov LB, Nik-Zainal S, Siu HC, Leung SY, Stratton MR. A mutational signature in gastric cancer suggests therapeutic strategies. Nat Commun. 2015; 6:8683. https://doi.org/10.1038/ncomms9683.

38. Jiricny J. The multifaceted mismatch-repair system. Nat Rev Mol Cell Biol. 2006; 7:335–46. https://doi.org/10.1038/nrm1907.

39. Marinus MG. DNA Mismatch Repair. Ecosal Plus. 2012; 5:87–100. https://doi.org/10.1128/ecosalplus.7.2.5.

40. Pickup MW, Mouw JK, Weaver VM. The extracellular matrix modulates the hallmarks of cancer. EMBO Rep. 2014; 15:1243–53. https://doi.org/10.15252/embr.201439246.

41. Frantz C, Stewart KM, Weaver VM. The extracellular matrix at a glance. J Cell Sci. 2010; 123:4195–200. https://doi.org/10.1242/jcs.023820.

42. Sprenger CC, Plymate SR Sr, Reed MJ. Aging-related alterations in the extracellular matrix modulate the microenvironment and influence tumor progression. Int J Cancer. 2010; 127:2739–48. https://doi.org/10.1002/ijc.25615.

43. Sakoda LC, Loomis MM, Doherty JA, Julianto L, Barnett MJ, Neuhouser ML, Thornquist MD, Weiss NS, Goodman GE, Chen C. Germ line variation in nucleotide excision repair genes and lung cancer risk in smokers. Int J Mol Epidemiol Genet. 2012; 3:1–17.

44. Marteijn JA, Lans H, Vermeulen W, Hoeijmakers JH. Understanding nucleotide excision repair and its roles in cancer and ageing. Nat Rev Mol Cell Biol. 2014; 15:465–81. https://doi.org/10.1038/nrm3822.

45. Li X, Zhang J, Su C, Zhao X, Tang L, Zhou C. The association between polymorphisms in the DNA nucleotide excision repair genes and RRM1 gene and lung cancer risk. Thorac Cancer. 2012; 3:239–48. https://doi.org/10.1111/j.1759-7714.2012.00115.x.

46. Cheng L, Spitz MR, Hong WK, Wei Q. Reduced expression levels of nucleotide excision repair genes in lung cancer: a case-control analysis. Carcinogenesis. 2000; 21:1527–30. https://doi.org/10.1093/carcin/21.8.1527.

47. Kiyohara C, Yoshimasu K. Genetic polymorphisms in the nucleotide excision repair pathway and lung cancer risk: a meta-analysis. Int J Med Sci. 2007; 4:59–71. https://doi.org/10.7150/ijms.4.59.

48. Zhang D, Tang B, Xie X, Xiao YF, Yang SM, Zhang JW. The interplay between DNA repair and autophagy in cancer therapy. Cancer Biol Ther. 2015; 16:1005–13. https://doi.org/10.1080/15384047.2015.1046022.

49. Czarny P, Pawlowska E, Bialkowska-Warzecha J, Kaarniranta K, Blasiak J. Autophagy in DNA damage response. Int J Mol Sci. 2015; 16:2641–62. https://doi.org/10.3390/ijms16022641.

50. Nasarre P, Potiron V, Drabkin H, Roche J. Guidance molecules in lung cancer. Cell Adhes Migr. 2010; 4:130–45. https://doi.org/10.4161/cam.4.1.10882.

51. Kuijper S, Turner CJ, Adams RH. Regulation of angiogenesis by Eph-ephrin interactions. Trends Cardiovasc Med. 2007; 17:145–51. https://doi.org/10.1016/j.tcm.2007.03.003.

52. Chen J, Zhuang G, Frieden L, Debinski W. Eph receptors and Ephrins in cancer: common themes and controversies. Cancer Res. 2008; 68:10031–33. https://doi.org/10.1158/0008-5472.CAN-08-3010.

53. Zhang Y, Zhu C, Sun B, Lv J, Liu Z, Liu S, Li H. Integrated high throughput analysis identifies GSK3 as a crucial determinant of p53-mediated apoptosis in lung cancer cells. Cell Physiol Biochem. 2017; 42:1177–91. https://doi.org/10.1159/000478873.

54. Mehlen P, Delloye-Bourgeois C, Chédotal A. Novel roles for Slits and netrins: axon guidance cues as anticancer targets? Nat Rev Cancer. 2011; 11:188–97. https://doi.org/10.1038/nrc3005.

55. Hammerman P, Lawrence M, Voet D, Jing R, Cibulskis K, Sivachenko A, Stojanov P, McKenna A, Lander E, Gabriel S, Getz G, Sougnez C, Imielinski M, et al, and Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489:519–25. https://doi.org/10.1038/nature11404.

56. Goldfeder RL, Priest JR, Zook JM, Grove ME, Waggott D, Wheeler MT, Salit M, Ashley EA. Medical implications of technical accuracy in genome sequencing. Genome Med. 2016; 8:24. https://doi.org/10.1186/s13073-016-0269-0.

57. Carvalho B, Irizarry RA, Scharpf RB, Carey VJ. Processing and analyzing Affymetrix SNP chips with Bioconductor. Stat Biosci. 2009; 1:160–80. https://doi.org/10.1007/s12561-009-9015-0..

58. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet JP, Ahmann GJ, Adli M, Anderson KC, Ardlie KG, Auclair D, et al. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011; 471:467–72. https://doi.org/10.1038/nature09837.

59. 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:R41. https://doi.org/10.1186/gb-2011-12-4-r41.

60. Byrd R, Lu P, Nocedal JZ, Zhu C. A Limited memory algorithm for bound constrained optimization. SIAM J Sci Comput. 1995; 16:1190–208. https://doi.org/10.1137/0916069..

61. R Core Team. R: A language and environment for statistical computing. R Found Stat Comput. 2014. http://www.R-project.org/.

62. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011; 12:323. https://doi.org/10.1186/1471-2105-12-323.

63. Best DJ. Algorithm AS 71: the upper tail probabilities of spearman's rho. J R Stat Soc [Ser A]. 1975; 24:377–79.

64. Storey JD. A direct approach to false discovery rates. J R Stat Soc. 2002; 64:479–98. https://doi.org/10.1111/1467-9868.00346..

IJC
International Journal of Cancer

Somatic genome alterations in relation to age in lung adenocarcinoma

Stefano Meucci[1], Ulrich Keilholz[1], Daniel Heim[2], Frederick Klauschen[2] and Stefano Cacciatore[3,4]

[1]Charité Comprehensive Cancer Center, Charité University Hospital, Charitéplatz 1, 10117, Berlin, Germany
[2]Institut für Pathologie, Charité University Hospital, Charitéplatz 1, 10117, Berlin, Germany
[3]Imperial College Parturition Research Group, Division of the Institute of Reproductive and Developmental Biology, Imperial College London, London, W120NN, United Kingdom
[4]Cancer Genomics Group, International Centre for Genetic Engineering and Biotechnology, Cape Town, South Africa

Lung adenocarcinoma (LUAD) is the most common cause of global cancer-related mortality and the major risk factor is smoking consumption. By analyzing 486 LUAD samples from The Cancer Genome Atlas, we detected a higher mutational burden among younger patients in the global cohort as well as in the *TP53*-mutated subcohort. The interaction effect of patient age and *TP53* mutations significantly affected the mutational rate of younger *TP53*-mutated patients. Furthermore, we detected a significant enrichment of the smoking-related signature SI4 (SI4) among younger *TP53*-mutated patients, meanwhile the age-related Signature 1 (SI1) significantly increased in proportion to patient age. Although present and past smoking is reported in the *TP53* wild-type patients, we observed a lower average number of somatic mutations, with no correlation with patient age. Overall, *TP53* mutations were significantly higher in younger patients and mainly characterized by SI4 and Signature 24 (SI24). Therefore, *TP53* seemed to acquire a particular sensitivity to smoking related C>A mutations in younger patients. We hypothesize that *TP53* mutations at a younger age might be a crucial factor enhancing the sensitivity to smoking-related mutations leading to a burst of somatic alterations. The mutational profile of cancer cell might reflect the mutational processes operative in aging in a given tissue. Therefore, *TP53*-mutated and *TP53* wild-type patient groups might represent phenotypes which endure aging-related mutational processes with different strength. Our study provides indications of age-dependent differences in mutational backgrounds that might be relevant for cancer prevention and age-adjusted treatment approaches.

Introduction

Lung cancer is the most common cause of global cancer-related mortality. The two major histological classes are non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLCs mostly comprise lung adenocarcinomas (LUAD) and lung squamous carcinomas (LUSC).[1] Although the main common risk factor remain the consumption of tobacco,[2] 10–15% of patients with LUAD had never smoked.

Somatic mutations in a cancer genome are mutations accumulating during the human aging with different rates and strength due to endogenous and exogenous factors.[3] Mathematical modeling strongly suggests that half or more of somatic passengers mutations in tumors arise before initiation of the tumor, that is, during development and aging.[4] They are classified as either "driver" mutations, which initiate the tumor and confer selected cancer phenotypes, or the far more numerous "passenger" mutations which have minor impact on cell growth and proliferation but they might built strength to cancer progression or confer resistance to the treatments.[5–7] Indeed, several studies illustrated the ongoing aging-related process of accumulation of mutations, selection, and clonal expansion.[8] Furthermore, the age at diagnosis is very closely correlated with the duration of

Key words: mutational patterns, aging, *TP53*, somatic mutations, CNAs

Abbreviations: AJCC: American Joint Committee on Cancer; CNA: copy number alterations; COSMIC: Catalogue of Somatic Mutations in Cancer; FDR: false discovery rate; LUAD: lung adenocarcinoma; LUSC: lung squamous carcinomas; NSCLC: nonsmall cell lung cancer; SCLC: small cell lung cancer; SI1: signature 1; SI24: signature 24; SI4: signature 4; TCGA: The Cancer Genome Atlas

Additional Supporting Information may be found in the online version of this article.

Conflict of interest: The authors declare no potential conflicts of interest.

Grant sponsor: Charité Comprehensive Cancer Center, Berlin; **Grant sponsor:** German Cancer Research Center (DKTK); **Grant sponsor:** Focus Area Dynage

DOI: 10.1002/ijc.32265

History: Received 2 Oct 2018; Accepted 27 Feb 2019; Online 12 Mar 2019.

Correspondence to: Stefano Meucci, Charité Comprehensive Cancer Center, Charité University Hospital, Charitéplatz 1, 10117, Berlin, Germany, E-mail: stefano.meucci@charite.de

What's new?

In lung adenocarcinoma, previous analyses have reported a higher mutational rate among younger patients. These authors investigated the genetic mechanisms at work, drawing on data from The Cancer Genome Atlas (TCGA). They found that in patients whose tumors carried TP53 mutations, younger age correlated with higher mutational load. This association was not seen with wild-type TP53. Younger patients were also more likely to have the smoking-related signature 4 mutation profile. The authors suggest that TP53 mutations in younger patients may increase sensitivity to smoking-related somatic mutations.

smoking, therefore aging- and smoking-related mutations should be simultaneously taken into account.[9–11]

However, a higher mutational rate among younger patients with LUAD was highlighted from the analysis of patient data available on The Cancer Genome Atlas (TCGA) portal, hypothesizing that a tumor with defective DNA polymerases and DNA repair genes (i.e., mutator phenotype), rapidly accumulates somatic mutations and might have concealed any age-related increase in mutation frequency.[12,13] In particular, tumors harboring TP53-mutated gene showed a negative association between the mutational load and patient age, while this association was not observed for the wild-type counterparts.[14]

The tumor suppressor *TP53* is the most frequently altered gene in LUAD.[2] Considering the crucial roles of p53 in maintaining genome stability, the loss or disruption of p53 function can lead to uncontrolled cell proliferation and cancer.[7] Past studies demonstrated that the frequency of *TP53* mutations increased with tobacco consumption.[15–17] Therefore, the relation between *TP53* mutations, tobacco consumption and patient age remains an open question.

In order to explore the underlying genetic mechanisms of LUAD mutational patterns, we investigated on the relationships between patient age and the number of somatic mutations. Specific mutagenesis processes such as DNA replication infidelity, exogenous and endogenous genotoxins exposures, defective DNA repair pathways and DNA enzymatic editing occur along patient aging. Therefore, we analyzed the mutational profiling and the respective correlation with the previously defined mutational signatures described in the Catalogue of Somatic Mutations in Cancer (COSMIC).[3,18] Furthermore, we performed gene-specific correlation analysis in relation to patient age on each significantly mutated genes in LUAD[1] with a special focus on the tumor suppressor *TP53*.

Additionally, we correlated copy number alterations (CNAs) load with patient age. Several CNAs have been reported to be associated with aging and cancer.[19,20] CNAs are defined as DNA segments larger than 1 kb in size that vary in copy number between individuals due to insertion, deletion or duplication.[21] The mechanisms through which CNAs can lead to phenotypic effects include among others gene interruption, gene fusion and changes in gene expression.

The combination of these multiple mutational processes may compose jumbled signatures which develop different tumor characteristics in relation to patient age.[22–25] The results from the current study may pave the way for future studies of molecular tumorigenesis in relation to human aging and underlines the need to consider age-adjusted treatments not only based on age and morbidity of older patients but also on differences in tumor biology.

Materials and Methods

TCGA data sets

Multiplatform genomic data sets were generated by TCGA Research Network (http://cancergenome.nih.gov/). Cancer molecular profiling data were generated through informed consent as part of previously published studies[26] and analyzed in accordance with each original study's data use guidelines and restrictions. The clinical data of the LUAD data set was obtained via download from the publicly available TCGA data matrix (https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm). Tumor staging classification was established according to the American Joint Committee on Cancer (AJCC) pathologic tumor staging. Tobacco smoking history was defined as: (*i*) lifelong nonsmokers (a person who was not smoking at the time of the interview and has smoked less than 100 cigarettes in their life), (*ii*) current smokers (includes daily smokers and occasional smokers), (*iii*) current reformed smokers for >15 years and (*iv*) current reformed smokers for ≤15 years (a person who was not smoking at the time of the interview since at least 15 years or since less than 15 years, but has smoked at least 100 cigarettes in their life). Smoker patients were classified as heavy smokers (more than 50 packs per year) and mild smokers (less or equal to 50 packs per years). Patients with no information about age were excluded by the following analysis.

Somatic mutations

Somatic mutations were obtained from the open access MAFs available from the GDC Legacy Archive (2016)[27] and directly utilized. Three different exclusion criteria for mutation data entries were considered in our study. (*i*) Mutations present in different samples belonging to the same patient were excluded. The mutations not included were equal to the 32.5% (from 347,181 to 234,434 entries). (*ii*) In presence of a mutation event on a sequence shared among different genes (e.g., paralogous genes), it will not be possible to identify the mutated gene. Mutations associated with more than one gene were excluded. In this step, the 0.1% of mutations were removed (from 234,434 to 234,217 entries). (*iii*) The challenges of repetitive sequence, which constitute more than half of the human genome leads to false positive variant calls due to systematic sequencing errors and local alignment challenges.[28] Therefore, only somatic mutations with

"ref_context" containing <6 continuous single repetitions, <4 continuous duplets, <3 continuous triplets, <3 continuous quadruplets and <3 continuous quintuplets were kept. With the third exclusion criteria, 8.8% of the mutations were excluded (from 234,217 to 222,139 entries). Somatic mutations belonging to patients with no reported age were removed (from 222,139 to 190,598 entries). Finally, data were available for 486 patients. Patients were classified as *TP53* mutated according to the presence of one (or more) somatic mutation with moderate or high impact on the gene *TP53*.

Copy number alterations

DNA from each tumor or germline-derived sample had been hybridized to Affymetrix SNP 6.0 arrays[29] and processed through GISTIC[30,31] by the TCGA consortium.

High-level copy gain or copy loss events for individual genes were interfered from the publicly available Firehose's "thresholded by genes" results table (http://gdac.broadinstitute.org/runs/analyses__2016_01_28/data/LUAD/20160128/gdac.broadinstitute.org_LUAD-TP.CopyNumber_Gistic2.Level_4.2016012800.0.0.tar.gz; −2 values being indicative of near total copy loss, +2 values being indicative of gains greater than 1–2 copies). The global CNAs load was calculated summing the CNAs absolute values from each patient. Data were available for 482 patients and 24,776 CNAs entries.

COSMIC signatures

The six mutation subtypes were considered in order to evaluate the signature profile: C>A, C>G, C>T, T>A, T>C and T>G. Each of the substitutions was considered by incorporating information on the bases immediately 5′ and 3′ to each mutated base generating 96 possible single nucleotide variants (6 types of substitution × 4 types of 5′ base × 4 types of 3′ base). The 96 single nucleotide variants profile was evaluated as the results of the combination of the 30 different COSMIC signatures. Tumor sample profiles can be represented by a unique contribution of each COSMIC signature as the following expression:

$$a_1 \mathrm{x} SI1 + a_2 \mathrm{x} SI2 + a_3 \mathrm{x} SI3 + \ldots + a_{30} x SI30 \qquad (1)$$

where a_i is the coefficient representing the contribution of the ith COSMIC signature. The coefficients of each tumor samples were calculated minimizing the difference between the tumor profile and the expression (1). The function *optim* (method "L-BFGS-B"[32]) of the R software[33] was implemented in order to perform the above-mentioned procedure.

Statistical analysis

The Spearman's rank correlation coefficient was used to identify correlation between patient age and somatic mutations. For every Spearman's test performed in our study, p values were computed using algorithm AS 89 included in the R function *cor.test* where the permutation distribution was estimated by an Edgeworth approximation.[34] Fisher's exact test was performed to compare categorical variables between two patient subgroups using the R function *fisher.test*. Wilcoxon Rank-Sum test was performed to compare continuous variables between two patient subgroups using the R function *wilcox.test*.

Two-way ANOVA was used to investigate the interaction between patient age and smoking history on somatic mutations. A p-value <0.05 was considered to be significant. To account for multiple testing, a FDR of ≤20% was applied to reduce identification of false positives.[35] The FDR was calculated using the R function *p.adjust*. All calculations were made using R software.[33]

The KODAMA algorithm[36–38] was used to facilitate the identification of patterns among COSMIC signature profiles of the significant mutated genes in LUAD.

Results

Correlation analysis between somatic alterations and patient age

We performed our study on the 486 patient samples of the LUAD cancer cohort available through the TCGA data set (Supporting Information Table S1). Due to the counterintuitive negative correlation between the somatic mutation load and patient age observed in past studies, we wanted to investigate the distribution of genome-wide mutations across age in patient subgroups by means of the Spearman's rank correlation coefficient between the global number of somatic mutations and CNAs for each patient. Noteworthily, we reported a higher consumption of tobacco (i.e., smoking intensity) among the older "current smokers" and "current reformed smoker for ≤15 years" patients, while no correlation is observed when the global smokers were considered (Supporting Information Table S2).

The global somatic mutations load showed a significant negative correlation with patient age (Table 1), which indicated a higher mutational rate among younger patients. In order to evaluate only the disruptive mutations, we classified somatic mutations according to their expected biological effect as low, moderate or high (Supporting Information Table S3). We confirmed a significant negative correlation also when mutations classified as low and multiple mutations in one gene were excluded. Also, the correlation between the global CNAs load and patient age had a negative trend, showing a higher rate of CNAs among younger patients (Supporting Information Table S4). A significant negative correlation was also detected when either only amplifications or only deletion was considered, independently. We repeated this analysis on patient subcohorts established according to the tobacco smoking history indicator, the pathologic tumor stage and the transversion status defined by Campbell *et al.*[1] as high and low through the smoking related C>A transversions bimodal pattern identified in LUAD.[2]

The transversion-high subcohort showed a significant negative correlation of both somatic mutations (Table 1) and CNAs load (Supporting Information Table S4) with patient age, indicating an enrichment of somatic mutations and CNAs among younger high mutational rate samples. In particular,

Table 1. Correlation between somatic mutations and patient age

	Somatic mutation				
Classification	*n*	Age, median [95%CI]	*rho*	*p*	FDR
Global	486	66.8 [42.7–83.8]	−0.16	3.93×10^{-4}	2.14×10^{-3}
Transversion status					
High	337	66.4 [42.4–81.9]	−0.23	2.38×10^{-5}	3.09×10^{-4}
Low	133	68.0 [44.3–84.7]	0.03	7.69×10^{-1}	9.16×10^{-1}
Tobacco smoking history indicator					
Lifelong nonsmokers	65	66.3 [46.3–81.6]	−0.02	9.04×10^{-1}	9.16×10^{-1}
Current smokers	116	61.7 [41.4–79.8]	−0.23	1.20×10^{-2}	3.12×10^{-2}
Current reformed smokers for >15 years	127	72.0 [49.8–85.5]	0.01	9.16×10^{-1}	9.16×10^{-1}
Current reformed smokers for ≤15 years	161	64.3 [42.2–79.2]	0.03	6.65×10^{-1}	9.16×10^{-1}
Tumor staging					
Stage I	261	67.6 [42.4–84.0]	−0.21	6.59×10^{-4}	2.14×10^{-3}
Stage II	112	65.3 [46.8–79.9]	−0.15	1.10×10^{-1}	2.04×10^{-1}
Stage III	80	67.9 [47.2–82.5]	0.11	3.46×10^{-1}	5.62×10^{-1}
Stage IV	26	62.7 [40.4–78.8]	−0.34	9.23×10^{-2}	2.00×10^{-1}
TP53					
Mutated	264	64.9 [41.9–82.2]	−0.21	5.74×10^{-4}	2.14×10^{-3}
Wild-type	222	68.0 [44.4–84.0]	0.01	8.47×10^{-1}	9.16×10^{-1}

Spearman's rank correlations between the somatic mutations loads and patient age for each patient subgroup established according to the patient characteristic such as mutational rate profile (i.e., transversion status), tobacco exposure data (i.e., tobacco smoking history indicator), tumor staging (i.e., AJCC pathologic tumor stage) and *TP53* mutational profile.

deletions were negatively correlated in the transversion-high subcohort, while amplifications were negatively correlated in both transversion-high and transversion-low subcohorts. No somatic mutations load correlation was detected in the transversion-low subcohort. The patient subcohort of current smokers showed significant negative correlation of both somatic mutations and CNAs load with patient age.

Additionally, we investigated the differences between the subcohorts established according to the *TP53* mutational profile (i.e., mutated or wild-type) in order to explore the influence of the most frequently mutated gene in LUAD. The *TP53*-mutated subcohort showed a significant enrichment of somatic mutations among younger patients, while no correlation was detected in the *TP53* wild-type cohort (Table 1). The CNAs load in the *TP53*-mutated cohort was overall higher than the wild-type counterpart and negatively correlated with patient age. No correlation was detected in the *TP53* wild-type cohort. These differences are highlighted in Figure 1. Full results are reported in Supporting Information Table S4. Then, we repeated the analysis in the subgroup of heavy smokers (Supporting Information Table S5) and in the mild smoker (Supporting Information Table S6). Interestingly, only the mild smokers showed a negative correlation of both somatic mutations and CNAs with patient age in global cohort and transversion high subcohort.

We used two-way ANOVA to evaluate, independently, the effect of age and *TP53* mutations as well as the combination effect of both on the somatic mutations and CNAs loads (Supporting Information Table S7). We reported a higher mutational load in younger *TP53*-mutated patients (Supporting Information Table S4) and we detected a significant effect of patient age and *TP53* mutations separately, on both the somatic mutations and CNAs loads. Interestingly, the interaction of patient age and *TP53* mutations significantly affected the higher mutational rate of younger *TP53* mutated patients. However, we did not observe any statistically significant interactions on the CNAs load.

Age, smoking habits, transversion status and TP53 mutational profile

Next, we explored the relation among *TP53* mutation, transversion status and smoking habits with patient age. Using the Fisher's test, we noted in the *TP53*-mutated subcohort a significantly higher percentage of current smokers ($p = 6.58 \times 10^{-6}$, FDR = 1.97×10^{-5}) and transversion-high profiles ($p = 2.13 \times 10^{-4}$, FDR = 3.20×10^{-4}), while no significant difference was detected in tumor staging (Supporting Information Table S8). Moreover, we used the Fisher's test to compare the percentage of patients with *TP53* mutated in all subcohort. The overall percentage of *TP53*-mutated patients was significantly different ($p = 2.13 \times 10^{-4}$, FDR = 8.48×10^{-4}) between transversion-high (59.6%) and transversion-low (40.6%) subcohorts (Fig. 2, Supporting Information Table S9). After the classification of the patients based on their age, we noted that this difference was even bigger in younger and absent in older patients. Noteworthy the percentage of *TP53*-mutated patients in <50 years age class was 30% and 85% in transversion-low and transversion-high, respectively ($p = 4.84 \times 10^{-3}$, FDR = 9.69×10^{-9}).

The overall percentage of *TP53*-mutated patients in "lifelong nonsmokers" subgroup was 41.5%, analogous to the

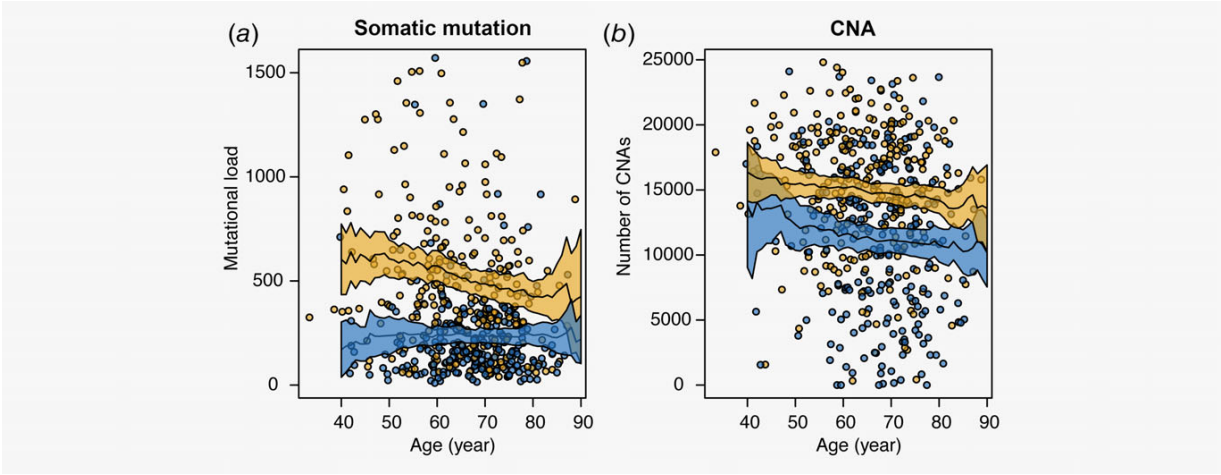

Figure 1. Correlation between genomic alterations and patient age in the *TP53*-mutated and *TP53* wild-type patient subcohorts. Number of (*a*) somatic mutations and (*b*) CNAs with their relative 95% confidence interval for each patient distributed along patient age in the TP53-mutated (yellow) and TP53 wild-type (blue) patient subcohorts. Medians (black line) and their relative 95% confidence interval (yellow and blue areas) were calculated locally in a range of ±10 years.

"current reformed smokers for >15 years" (40.9%). Although the comparable percentage of transversion-high profiles in "current reformed smokers for ≤15 years" (86.6%) and "current smokers" (86.4%) subgroups, they displayed 57.1% and 70.7% of *TP53*-mutated patients, respectively, showing an inverse proportional increase to time since smoking cessation. Overall, smoking consumption significantly increased ($p = 6.20 \times 10^{-7}$, FDR = 3.72×10^{-6}) the percentage of harboring *TP53* mutations independently of the age. Indeed, not only younger patients of <50 and 50–60 age classes but also those of the 70–80 age

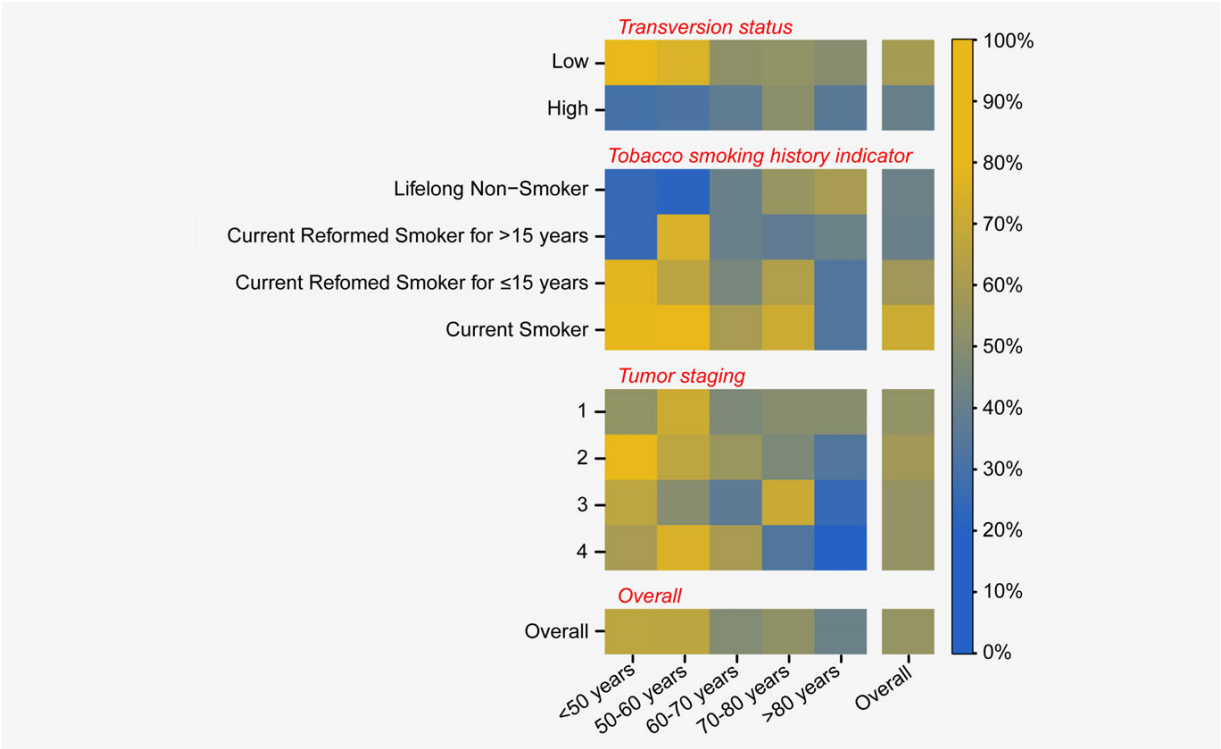

Figure 2. TP53-mutated distributions. Heatmap representing the percentage of patients with TP53 mutated across different subgroup of LUAD cohort.

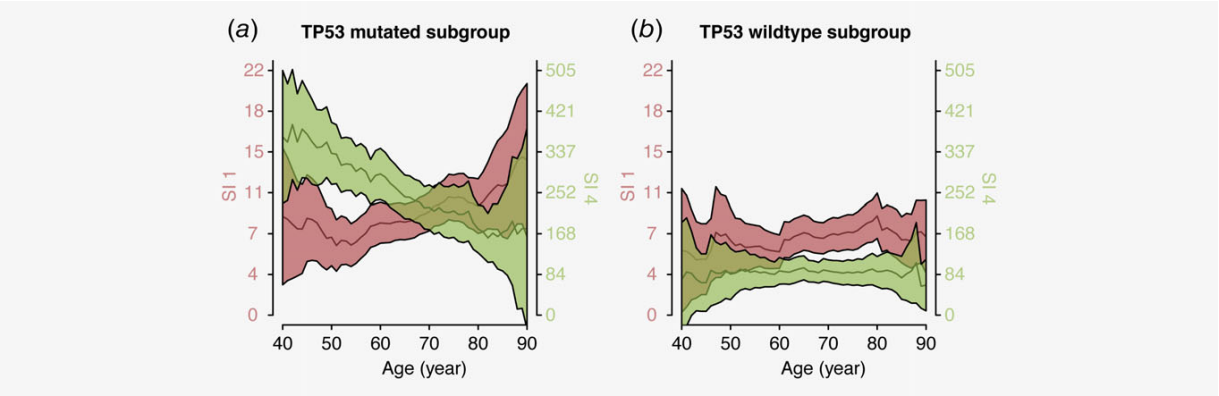

Figure 3. Correlation of somatic mutations profiling and patient age in the TP53-mutated and TP53 wild-type patient subcohorts. Correlation between the smoking related SI4 (green graph) and the age related SI1 (red graph) with patient age in (*a*) TP53-mutated and (*b*) TP53 wild-type patient subcohorts. Medians (black line) and their relative 95% confidence interval (colored area) were calculated locally in a range of ±10 years.

class showed significant correlations (Supporting Information Table S9).

Then, we performed the Wilcoxon's test to evaluate the difference in patient age between *TP53*-mutated and *TP53* wild-type patients for each subcohorts. The overall global cohort showed a significant ($p = 2.44 \times 10^{-3}$) lower age mean in the *TP53*-mutated patients compared to *TP53* wild-type patients as well as the transversion-high subcohort ($p = 3.14 \times 10^{-4}$). We detected the highest percentage of *TP53*-mutated patients in <50 (66.7%) and 50–60 (66%) age groups.

Mutational profiling

Multiple mutational processes such as exogenous or endogenous mutagen exposures, defective DNA repair and replication and enzymatic modification of DNA operate with different strength and nucleotide specificity along patient aging[3] showing unique mutational signatures.

In order to investigate on the differences in mutational profile between *TP53*-mutated and *TP53* wild-type subcohorts, we categorized each single nucleotide variants incorporating information on the bases immediately 5′ and 3′ to each mutated base. We deconvoluted trinucleotide variants profiles into the 30 different signatures described in the COSMIC database.[18,22] Therefore, we were able to characterize each patient by a different "intensity" combination of the 30 COSMIC signatures. Then, we performed the Spearman's rank correlation test between the intensities each COSMIC signature and the patient age in both *TP53*-mutated and *TP53* wild-type subcohorts (Supporting Information Table S10).

The *TP53*-mutated subcohorts showed a significant negative correlation between the smoking-related Signature 4 (SI4), associated with C>A transversions, and patient age (*rho* = −0.27, $p = 6.89 \times 10^{-6}$, FDR = 2.07×10^{-4}). While the age-related Signature 1 (SI1), mainly consisting of C>T transitions, was positively correlated with patient age (*rho* = 0.18, $p = 3.46 \times 10^{-3}$, FDR = 3.46×10^{-2}; Fig. 3*a*), showing the simultaneous ongoing age-related accumulation of somatic mutations. Moreover, the SI3 associated with failure of DNA double-strand break-repair by homologous recombination was positively correlated in the *TP53*-mutated subcohort (*rho* = 0.22, $p = 3.51 \times 10^{-4}$, FDR = 5.26×10^{-3}), while no significant correlations were identified in the *TP53* wild-type subcohort (Fig. 3*b*).

Focus on the significantly mutated genes in LUAD

In order to investigate on gene-specific driver mutations in relation to patient age, which might contribute to the higher mutational rate detected in younger patients, we computed the Spearman's rank correlation between patient age and somatic mutations load of each genes significantly mutated in LUAD as reported in a previous study by Campbell *et al.*[1] (Supporting Information Table S11). We report that somatic mutations on *TP53* were significantly enriched in younger patients (*rho* = −0.13, $p = 5.25 \times 10^{-3}$, FDR = 9.98×10^{-2}), as well as *ATM* (*rho* = −0.11, $p = 1.78 \times 10^{-2}$, FDR = 2.26×10^{-1}). While RBM10 disruptions were enriched among older patients (*rho* = 0.13, $p = 4.81 \times 10^{-3}$, FDR = 9.98×10^{-2}).

Finally, we hypothesized that multiple factors, such as the specific base sequence of each gene or the secondary DNA structure, could promote a mutational signature among the others. Therefore, we calculated the frequencies of COSMIC signatures using the mutations identified in each of these genes (Supporting Information Table S12). *TP53* and *RMB10* were especially enriched of smoking related SI4 and the aflatoxin related SI24, both constituted of C>A transversions, indicating guanine damage that is being repaired by transcription-coupled nucleotide excision repair. The defective DNA mismatch repair related SI6, associated with high numbers of small (shorter than 3 bp) insertions and deletions at mono/polynucleotide repeats was as well relatively enriched in TP53 (Fig. 4*a*). RMB10 was

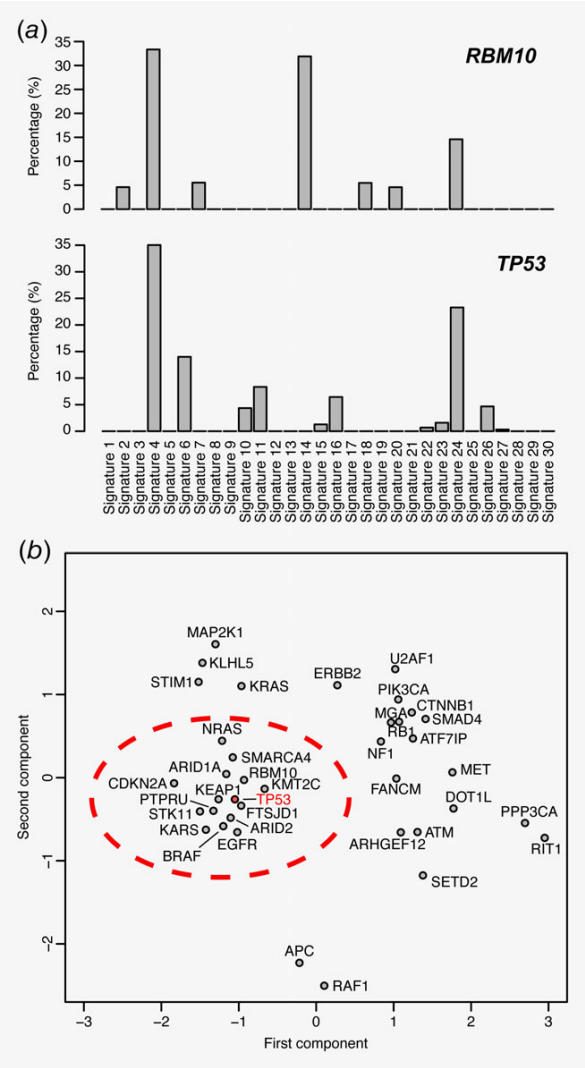

Figure 4. COSMIC signature profiling. (*a*) COSMIC signature profile of *RBM10* and *TP53* genes. (*b*) KODAMA plot of the COSMIC signature profiles of genes significantly mutated in LUAD as reported in a previous study by Campbell *et al.*[1]

also enriched of SI14, with unknown etiology, mainly constituted by C>A and C>T mutations. Interestingly, ATM was enriched of SI3 and SI10, while the smoking related SI4 was entirely absent.

To evaluate similarities among COSMIC signature profiles of these genes, we performed the unsupervised KODAMA analysis (Fig. 4*b*) showing that TP53 and RMB10 shared the same cluster, while ATM showed an independent profile.

Discussion

In the present study performed on the TCGA LUAD data set, we investigated genetic patterns (somatic mutations and CNAs) in relation to patient age. We confirmed the negative correlation between the somatic mutation load and patients age in the global cohort and *TP53*-mutated subcohort observed in previous studies.[12,14] The CNAs load was as well higher in younger patients, even when only amplification or deletion were considered. *TP53* mutated, transversion-high and current smokers subcohorts showed the same higher somatic mutations and CNAs burden among younger patients, displaying a relation among these factors. These results overlap with our previous investigation on LUSC,[39] which might be indicative of a tissue-specific higher sensitivity to smoking-related damages in younger patients. The group of *TP53*-mutated patients showed a higher percentage of current smokers and transversion-high profiles as well as a lower age mean compared to *TP53* wild-type patients, which instead displayed a lower average number of somatic mutations with no correlation with patient age. Most noteworthy, we identified that the effect of patient age and *TP53* mutations separately, as well as their interaction, significantly affected the mutational rate of younger *TP53* mutated patients. While only the separate effect of patient age and TP53 mutations was significant on the CNAs load correlation displayed by the *TP53*-mutated subcohort. Furthermore, we detected a significant enrichment of the smoking-related signature SI4 among younger *TP53*-mutated patients. Overall, smoking consumption significantly increase the percentage of harboring *TP53* mutation independently of the age. The percentage of *TP53*-mutated patients increased with an inverse proportion to the time since smoking cessation. Noteworthy is that 70.7% of current smokers were *TP53*-mutated patients.

Milholland *et al.*[12] hypothesized that smoking has a strong effect on both the frequency and the spectrum of somatic mutations. Our study enlarges this hypothesis by showing that the cumulative effect of smoking consumption, *TP53* mutations and a younger age significantly affected the overall mutational load among younger LUAD patients.

Alexandrov *et al*[11] displayed that the number of mutations (including *TP53*) increased with the number of pack-years. However, older current smokers, which had a lower mutational load, showed a higher consumption of tobacco (pack-years). The heavy smokers did not show correlation between genetic alteration and patient ages, while the mild smokers showed a higher mutational burden among younger patients. Therefore, smoking intensity might be not the only factor affecting the somatic alteration burst detected in younger patients.

The *TP53*-mutated subcohort also displayed the concurrent ongoing accumulation of SI1 along patient aging. SI1 is largely made up of C>T substitutions at CpG dinucleotides, which are the results of an endogenous mutational process initiated by spontaneous deamination of 5-methylcytosine, enzymatic deamination of cytosine or polymerase errors.[3–6] The SI3 associated with failure of DNA double-strand break-repair by homologous recombination was as well increasing along patient age in the *TP53*-mutated subcohort. Recent studies showed that impaired DNA double-strand break repair contributes to the age-associated

rise of genomic instability in humans.[40] Meanwhile, although present and past smoking is reported in the *TP53* wild-type patients, no correlation between mutational signatures and patient age was detected. As shown in past studies, the mutational profile of cancer cell might reflect the mutational processes operative in aging in a given tissue.[8,41] Therefore, we hypothesized that *TP53* wild-type patients might represent a phenotype with greater DNA stability, which may confine the ongoing age-related accumulation of genetic events as well as the increasing mutational burden due to smoking consumption.

Although previous studies revealed that the number of *TP53* mutations are common in noncancerous tissue and accumulate with age[8] and tobacco consumption,[16,17,42] we detected an overall higher rate of *TP53* mutations in younger patients particularly in <50 and 50–60 age groups. *TP53* mutations showed a strong enrichment of smoking related SI4 and aflatoxin related SI24, both constituted of C>A transversions, indicating guanine damages that are being repaired by transcription-coupled nucleotide excision repair. RBM10 was as well enriched of SI4 and SI24, but in older patients. Therefore, TP53 and RBM10 seemed to acquire a particular sensitivity to smoking-related mutations with contrary tendencies in relation to patient age. Whereas *ATM* mutations were enriched in younger patients and mainly constituted of SI3, while the smoking related SI4 was absent. *ATM* encodes a cell-cycle checkpoint kinase that function as a regulator of p53, and it acts as the apical regulators of the response to DNA double-strand breaks.[43] A recent study[44] suggested that *ATM* mutations may substitute functionally for TP53 mutations. Past studies[16,45,46] showed that purines seem to be the major target of carcinogens in tobacco smoke and that G:C>T:A transversions tended to cluster in the *TP53* hotspots, such as codons 157, 158 and 248. Therefore, we hypothesize that the nucleotide sequence might contribute to determine the different sensitivity displayed by *TP53* and *ATM* to smoking-related mutations and we speculate that the secondary DNA structure might as well have its influence. Furthermore, the association between chromatin structure and mutation rates showed striking heterogeneity along the genome. A past study[47] detected lower base substitution rates in open chromatin due to the higher accessibility of DNA repair mechanisms. Although the transcriptional activity of *TP53* might be higher in younger patients,[48] our results showed an increased *TP53* mutational burden. Furthermore, prolonged exposure to cigarette smoke and oxidants of lung epithelial cells *in vitro* resulted in marked temporal changes in histone acetylation and methylation patterns.[49,50] Therefore, the relation between the mutational pattern, which might damage different DNA repair systems, and the chromatin state might drastically change the sensitivity to smoking-related mutations.

In conclusion, *TP53* mutations at a younger age might be a crucial factor enhancing the sensitivity to smoking-related mutations leading to a burst of somatic alterations. *TP53* itself showed a higher sensitivity to smoking related C>A mutations in younger patients. *TP53*-mutated and *TP53* wild-type patient groups might represent phenotypes which endure aging-related mutational processes with different strength. Further studies with larger numbers of individuals of different ages and diversity of normal tissues are essential to elucidate the intricate relationship between smoking consumption and mutational patterns in relation to intrinsic aging processes. A better comprehension of LUAD tumorigenesis in relation to patient age might be relevant for cancer prevention and age-adjusted treatment decisions and should therefore be taken under closer consideration in future studies.

References

1. Campbell JD, Alexandrov A, Kim J, et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat Genet* 2016;48:607–16.
2. Collisson EA, Campbell JD, Brooks AN, et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 2014;511:543–50.
3. Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr Opin Genet Dev* 2014;24:52–60.
4. Tomasetti C, Vogelstein B, Parmigiani G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc Natl Acad Sci USA* 2013;110:1999–2004.
5. Alexandrov LB, Jones PH, Wedge DC, et al. Clock-like mutational processes in human somatic cells. *Nat Genet* 2015;47:1402–7.
6. Fox EJ, Salk JJ, Loeb LA. Exploring the implications of distinct mutational signatures and mutation rates in aging and cancer. *Genome Med* 2016;8:30.
7. Zhang W, Edwards A, Flemington EK. Significant prognostic features and patterns of somatic TP53 mutations in human cancers. 2017;16.
8. Risques RA, Kennedy SR. Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLoS Genet* 2018;14:1–12.
9. Westmaas JL, Newton CC, Stevens VL, et al. Does a recent cancer diagnosis predict smoking cessation? An analysis from a large prospective US cohort. *J Clin Oncol* 2015;33:1647–52.
10. Baser S, Shannon VR, Eapen GA, et al. Smoking cessation after diagnosis of lung cancer is associated with a beneficial effect on performance status. *Chest* 2005;130:1784–90.
11. Alexandrov LB, Ju YS, Haase K, et al. Mutational signatures associated with tobacco smoking in human cancer. *Science (80-)* 2016;354:618–22.
12. Milholland B, Auton A, Suh Y, et al. Age-related somatic mutations in the cancer genome. *Oncotarget* 2015;6:24627–35.
13. Loeb LA. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nat Rev Cancer* 2011;11:450–7.
14. Zhang W, Flemington EK, Zhang K. Mutant TP53 disrupts age-related accumulation patterns of somatic mutations in multiple cancer types. *Cancer Genet* 2016;209:376–80.
15. Sun S, Schiller JHGA. Lung cancer in never smokers—a different disease. *Nat Rev Cancer* 2007;7:778–90.
16. Halvorsen AR, Silwal-Pandit L, Meza-Zepeda LA, et al. TP53 mutation spectrum in smokers and never smoking lung cancer patients. *Front Genet* 2016;7:1–10.
17. Gibbons DL, Byers LA, Kurie JM. Smoking, p53 mutation, and lung cancer. *Mol Cancer Res* 2014;12:3–13.
18. Forbes SA, Beare D, Gunasekaran P, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 2015;43:D805–11.
19. Kuningas M, Estrada K, Hsu YH, et al. Large common deletions associate with mortality at old age. *Hum Mol Genet* 2011;20:4290–6.
20. Iakoubov L, Mossakowska M, Szwed M, et al. A common copy number variation (CNV) polymorphism in the CNTNAP4 gene: association with aging in females. *PLoS One* 2013;8:1–9.
21. Feuk L, Feuk L, Carson AR, et al. Structural variation in the human genome. *Nat Rev Genet* 2006;7:85–97.

22. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature* 2013;500:415–21.
23. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* 2013;3:246–59.
24. Meucci S, Keilholz U, Tinhofer I, et al. Mutational load and mutational patterns in relation to age in head and neck cancer. *Oncotarget* 2016;7: 69188–99.
25. Roberts SA, Gordenin DA. Hypermutation in human cancer genomes: footprints and mechanisms. *Nat Rev Cancer* 2014;14:786–800.
26. Hammerman P, Lawrence M, Voet D, et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 2012;489:519–25.
27. Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. *N Engl J Med* 2016;375:1109–12.
28. Goldfeder RL, Priest JR, Zook JM, et al. Medical implications of technical accuracy in genome sequencing. *Genome Med* 2016;8:24.
29. Carvalho B, Irizarry RA, Scharpf RB, et al. Processing and analyzing Affymetrix SNP chips with bioconductor. *Stat Biosci* 2009;1:160–80.
30. Chapman MA, Lawrence MS, Keats JJ, et al. Initial genome sequencing and analysis of multiple myeloma. *Nature* 2011;471:467–72.
31. Mermel CH, Schumacher SE, Hill B, 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:R41.
32. Byrd R, Lu P, Nocedal JZC. A limited memory algorithm for bound constrained optimization. *SIAM J Sci Comput* 1995;16:1190–208.
33. R Core Team. R: A language and environment for statistical computing. *R Found Stat Comput* 2014;
34. Best DJ, Roberts DE. Algorithm AS 71 : the upper tail probabilities of Spearman's rho. *J R Stat Soc* 1975;24:377–9.
35. Storey JD. A direct approach to false discovery rates. *J R Stat Soc* 2002;64:479–98.
36. Cacciatore S, Luchinat CTL. Knowledge discovery by accuracy maximization. *Proc Natl Acad Sci USA* 2014;111:5117–22.
37. Cacciatore S, Tenori L, Luchinat C, et al. KODAMA: an R package for knowledge discovery and data mining. *Bioinformatics* 2017;3:621–3.
38. Bray R, Cacciatore S, Jiménez B, et al. Urinary metabolic phenotyping of women with lower urinary tract symptoms. *J Proteome Res* 2017;16:4208–16.
39. Meucci S, Keilholz U, Heim D, et al. Somatic genome alterations in relation to age in lung squamous cell carcinoma. *Oncotarget* 2018;9:32161–72.
40. Li Z, Zhang W, Chen Y, et al. Impaired DNA double-strand break repair contributes to the age-associated rise of genomic instability in humans. *Cell Death Differ* 2016;23:1765–77.
41. Hoang ML, Kinde I, Tomasetti C, et al. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc Natl Acad Sci USA* 2016;113: 9846–51.
42. Rivlin N, Brosh R, Oren M, et al. Mutations in the p53 tumor suppressor gene: important milestones at the various steps of tumorigenesis. *Genes Cancer* 2011;2:466–74.
43. Weber AM, Ryan AJ. ATM and ATR as therapeutic targets in cancer. *Pharmacol Ther* 2015;149:124–38.
44. Ding L, Getz G, Wheeler DA, et al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 2008;455:1069–75.
45. Menzies GE, Reed SH, Brancale A, et al. Base damage, local sequence context and TP53 mutation hotspots: a molecular dynamics study of benzo[a]pyrene induced DNA distortion and mutability. *Nucleic Acids Res* 2015;43: 9133–46.
46. Pleasance ED, Stephens PJ, O'Meara S, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 2010;463: 184–90.
47. Makova KD, Hardison RC. The effects of chromatin organization on variation in mutation rates in the genome. *Nat Rev Genet* 2015;16: 213–23.
48. Feng Z, Hu W, Teresky AK, et al. Declining p53 function in the aging process: a possible mechanism for the increased tumor incidence in older populations. *Proc Natl Acad Sci USA* 2007;104: 16633–8.
49. Mortaz E, Masjedi MR, Barnes PJ, et al. Epigenetics and chromatin remodeling play a role in lung disease. *Tanaffos* 2011;10:7–16.
50. Sundar IK, Nevid MZ, Friedman AE, et al. Cigarette smoke induces distinct histone modifications in lung cells: implications for the pathogenesis of COPD and lung cancer 2014. 982-996

For reasons of data protection law, my curriculum vitae will not be published in the electronic version of my work.

Publication list

1
Meucci S, Keilholz U, Tinhofer I, Ebner OA.
Mutational load and mutational patterns in relation to age in head and neck cancer
Oncotarget. 2016 Oct 25;7(43):69188-69199. doi: 10.18632/oncotarget.11312.
(Impact factor 5.168)

2
Liu Y, Meucci S, Sheng L, Keilholz U.
Meta-analysis of the mutational status of circulation tumor cells and paired primary tumor tissues from colorectal cancer patients.
Oncotarget. 2017 May 26;8(44):77928-77941. doi: 10.18632/oncotarget.18272.
(Impact factor 5.168)

3
Liu Y, Cheng G, Qian J, Ju H, Zhu Y, Meucci S, Keilholz U, Li D.
Expression of guanylyl cyclase C in tissue samples and the circulation of rectal cancer patients
Oncotarget. 2017 Jun 13;8(24):38841-38849. doi: 10.18632/oncotarget.16406.
(Impact factor 5.168)

4
Meucci S, Keilholz U, Heim D, Klauschen F, Cacciatore S.
Somatic genome alterations in relation to age in lung squamous cell carcinoma
Oncotarget. 2018 Aug 14;9(63):32161-32172. doi: 10.18632/oncotarget.25848.
(Impact factor 5.168)

5
Meucci S, Keilholz U, Heim D, Klauschen F, Cacciatore S.
Somatic genome alterations in relation to age in lung adenocarcinoma
International Journal of Cancer. 2019 Mar 12. https://doi.org/10.1002/ijc.32265
(Impact factor 7.36)

Acknowledgments

This work was supported by the Focus Area Dynage (www.fu-berlin.de/dynage) and the Charité Comprehensive Cancer Center, Berlin.